

Big Data and AI Strategies

Machine Learning and Alternative Data Approach to Investing

May, 2017

Dear Investor,

Over the past few years, we have witnessed profound changes in the marketplace with participants increasingly adopting quantitative investing techniques. These include [Risk Premia](#) investing, algorithmic trading, merging of fundamental and quantitative investment styles, consumption of increasing amounts and differentiated types of data, and adoption of new methods of analysis such as those based on Machine Learning and Artificial Intelligence.

In fact, over the past year, the exponential increase of the amount and types of data available to investors prompted some to completely change their business strategy and adopt a 'Big Data' investment framework. Other investors may be unsure on how to assess the relevance of Big Data and Machine Learning, how much to invest in it, and many are still paralyzed in the face of what is also called the 'Fourth Industrial Revolution'.

In this report we aim to provide a framework for Machine Learning and Big Data investing. This includes an overview of types of alternative data, and Machine Learning methods to analyze them. Datasets are at the core of any trading strategy. For this reason, we first **classify and analyze the types of alternative datasets**. We assess the relevance of various datasets for different types of investors and illustrate the use of Big Data in trading strategies. Datasets covered include **data generated by individuals** (e.g. social media), **data generated by business processes** (e.g. commercial transactions) and **data generated by machines** (e.g. satellite image data). After focusing on Datasets, we explain and **evaluate different Machine Learning methods** which are necessary tools to analyze Big Data. These methods include **Supervised Machine Learning**: regressions, classifications; **Unsupervised Machine Learning**: clustering, factor analyses; as well as methods of **Deep and Reinforcement Learning**. We provide theoretical, practical (e.g. codes) and investment examples for different Machine Learning methods, and compare their relative performance. The last part of the report **is a handbook of over 500 alternative data and technology providers**, which can be used as a rough roadmap to the Big Data and Artificial Intelligence landscape.

We hope this guide will be educative for investors new to the concept of Big Data and Machine Learning, and provide new insights and perspectives to those who already practice it.



Marko Kolanovic, PhD
Global Head of Macro Quantitative and Derivatives Strategy
J.P.Morgan Securities LLC

Table of Contents

I: INTRODUCTION AND OVERVIEW	6
Summary	7
Introduction to Big Data and Machine Learning	9
Classification of Alternative Data Sets	12
Classification of Machine Learning Techniques.....	16
Positioning within the Big Data Landscape	21
II: BIG AND ALTERNATIVE DATA	25
Overview of Alternative Data	26
Data from Individual Activity	30
Data from Business Processes.....	38
Data from Sensors.....	42
III: MACHINE LEARNING METHODS.....	51
Overview of Machine Learning Methods	52
Supervised Learning: Regressions	57
Supervised Learning: Classifications.....	77
Unsupervised Learning: Clustering and Factor Analyses	93
Deep and Reinforcement Learning	102
Comparison of Machine Learning Algorithms	117
IV: HANDBOOK OF ALTERNATIVE DATA	135
Table of contents of data providers.....	136
A. Data from Individual Activity	137
B. Data from Business Processes.....	147
C. Data from Sensors	176
D. Data Aggregators	189
E. Technology Solutions.....	191
APPENDIX	214
Techniques for Data Collection from Websites	215
Packages and Codes for Machine Learning	226
Mathematical Appendices.....	231
References.....	254
Glossary	270

I: INTRODUCTION AND OVERVIEW

Summary

Big Data and Machine Learning ‘revolution’: Most records and observations nowadays are captured electronically by devices connected to the internet. This, in principle, allows investors to access a broad range of market relevant data in real time. For instance, online prices of millions of items can be used to assess inflation, the number of customers visiting a store and transacting can give real time sales estimates, and satellite imaging can assess agricultural yields or activity of oil rigs. Historically, similar data were only available at low frequency (e.g. monthly CPI, weekly rig counts, USDA crop reports, retail sales reports and quarterly earnings, etc.). Given the amount of data that is available, a skilled quantitative investor can nowadays in theory have near real time macro or company specific data not available from traditional data sources. In practice, useful data are not readily available and one needs to purchase, organize and analyze alternative datasets in order to extract tradeable signals. Analysis of large or unstructured datasets is often done with the use of Machine Learning. Successful application of Machine Learning techniques requires some theoretical knowledge and a lot of practical experience in designing quantitative strategies.

Datasets and Methodologies: There are two main components of a Big Data investment approach: acquiring and understanding the data, and using appropriate technologies and methods to analyze those data. **New datasets are often larger in volume, velocity and variability** as compared to traditional datasets such as daily stock prices. Alternative datasets include **data generated by individuals** (social media posts, product reviews, search trends, etc.), **data generated by business processes** (company exhaust data, commercial transaction, credit card data, etc.) and **data generated by sensors** (satellite image data, foot and car traffic, ship locations, etc.). In most cases these datasets need a level of analysis before they can be used in a trading strategy. We aim to provide a roadmap to different types of data and assess their relevance for different asset classes as well as their relevance for different investment styles (e.g. macro, equity long-short, etc.). Methods to analyze big and alternative datasets include traditional statistics but also methods of Machine Learning. Machine Learning techniques include **Supervised Learning** (regressions, classifications), **Unsupervised Learning** (factor analysis, clustering) as well as novel techniques of **Deep and Reinforcement Learning** that are often used to analyze unstructured data and show promise in identifying data patterns in structured data. In this report, we provide theory and practical examples of these Machine Learning methods and assess their efficacy.

Fear of Big Data and Artificial Intelligence: While many traditional investors don’t have a good understanding of the types of data available, and feel uneasy about adopting Machine Learning methods, we want to point out that **these are not new concepts**. On a limited basis, many investors already deal with alternative datasets and some form of Machine Learning. For instance – Sam Walton, founder of Walmart, in the 1950s used airplanes to fly over and count cars on parking lots to assess real estate investments. The current extensive use of satellite imaging is a more technologically advanced, scalable and broadly available extension of the same idea. Machine Learning methods are often simple extensions of well-known statistical methods. Supervised learning methods aim to establish a relationship between two datasets and use one dataset to forecast the other. These methods are often as simple as regression models improved to accommodate changing market regimes, data outliers, and correlated variables. Unsupervised learning methods try to understand the structure of data and identify the main drivers behind it. These models are often closely related to well-known statistical methods such as principal component analysis. However, there are significant differences between a simple regression between two financial time series and a Big Data, Machine Learning framework. Big Data requires new analytical skills and infrastructure in order to derive tradeable signals. Strategies based on Machine Learning and Big Data also require market intuition, understanding of economic drivers behind data, and experience in designing tradeable strategies.

How will Big Data and Machine Learning change the investment landscape? We think the change will be profound. As more investors adopt alternative datasets, the market will start reacting faster and will increasingly anticipate traditional or ‘old’ data sources (e.g. quarterly corporate earnings, low frequency macroeconomic data, etc.). This gives an edge to quant managers and those willing to adopt and learn about new datasets and methods. Eventually, ‘old’ datasets will lose most predictive value and new datasets that capture ‘Big Data’ will increasingly become standardized. There will be an ongoing effort to uncover new higher frequency datasets and refine/supplement old ones. Machine Learning techniques will become a standard tool for quantitative investors and perhaps some fundamental investors too. Systematic strategies such as risk premia, trend followers, equity long-short quants, etc., will increasingly adopt Machine Learning tools and methods. **The ‘Big Data ecosystem’** involves specialized firms that collect, aggregate and sell new datasets, and research teams on both

the buy side and sell side that evaluate data. As the Big Data ecosystem evolves, datasets that have high Sharpe ratio signals (viable as a standalone funds) will disappear. The bulk of Big Data signals will not be viable as stand-alone strategies, but will still be very valuable in the context of a quantitative portfolio.

Potential Pitfalls of Big Data and Machine Learning: The transition to a Big Data framework will not be without setbacks. Certain types of data may lead into blind alleys - datasets that don't contain alpha, signals that have too little investment capacity, decay quickly, or are simply too expensive to purchase. Managers may invest too much into unnecessary infrastructure e.g. build complex models and architecture that don't justify marginal performance improvements. Machine Learning algorithms cannot entirely replace human intuition. Sophisticated models, if not properly guided, can overfit or uncover spurious relationships and patterns. **Talent** will present another source of risk - employing data scientists who lack specific financial expertise or financial intuition may not lead to the desired investment results or lead to culture clashes. In implementing Big Data and Machine Learning in finance, it is more important to understand the economics behind data and signals, than to be able to develop complex technological solutions. Many Big Data and AI concepts may sound plausible but will not lead to viable trading strategies.

Roles of Humans and Machines: A question that we are often asked is what will be the respective roles of humans and machines in the finance industry after the Big Data/AI 'revolution'. First we note that for short term trading, such as high frequency market making, humans already play a very small role. On a medium term investment horizon, machines are becoming increasingly relevant. Machines have the ability to quickly analyze news feeds and tweets, process earnings statements, scrape websites, and trade on these instantaneously. These strategies are already eroding the advantage of fundamental analysts, equity long-short managers and macro investors. On a long term horizon, machines will likely not be able to compete with strong macro and fundamental human investors. The current stage of development of Artificial Intelligence is still modest. For instance, machines still have a hard time passing Winograd's test¹. Machines will likely not do well in assessing regime changes (market turning points) and forecasts which involve interpreting more complicated human responses such as those of politicians and central bankers, understanding client positioning, or anticipating crowding.

Regardless of the timeline and shape of the eventual investment landscape, we believe that **analysts, portfolio managers, traders and CIOs will eventually have to become familiar with Big Data and Machine Learning approaches to investing**. This applies to both fundamental and quantitative investors, and is true across asset classes.

In the first chapter of this report, we provide an initial framework to understand the Big Data and Machine Learning methods. The second chapter of the report **classifies big and alternative datasets according to their type and relevance for different investment styles.** The third chapter elaborates on **individual Machine Learning techniques, their implementation and practical examples.** The fourth chapter is a **handbook of over 500 alternative data providers** which can be used as a rough roadmap to the Big Data and Machine Learning landscape. Finally, in the Appendix we provide computer codes and Machine Learning libraries, additional theoretical considerations, references to relevant literature, and a glossary of terms.

¹ In the sentence "The councilmen refused the demonstrators a permit because *they* feared violence", the pronoun "*they*" refers to the councilmen. If we replace the verb "*feared*" with "*advocated*", the pronoun "*they*" refers to the demonstrators instead as in "The councilmen refused the demonstrators a permit because *they* advocated violence". Interpreting such ambiguous pronouns has remained a challenge, as of this writing. For details, see Levesque et al (2011).

Introduction to Big Data and Machine Learning

In the search for uncorrelated strategies and alpha, fund managers are increasingly adopting quantitative strategies. Beyond strategies based on [alternative risk premia](#)², a new source of competitive advantage is emerging with the availability of alternative data sources as well as the application of new quantitative techniques of Machine Learning to analyze these data. This ‘*industrial revolution of data*’ seeks to provide alpha through informational advantage and the ability to uncover new uncorrelated signals. The Big Data **informational advantage** comes from datasets created on the back of new technologies such as mobile phones, satellites, social media, etc. The informational advantage of Big Data is **not related to expert and industry networks, access to corporate management, etc., but rather the ability to collect large quantities of data and analyze them in real time**. In that respect, Big Data has the ability to profoundly change the investment landscape and further shift investment industry trends from a discretionary to quantitative investment style.

There have been three trends that enabled the Big Data revolution:

- 1) Exponential increase in amount of data available
- 2) Increase in computing power and data storage capacity, at reduced cost
- 3) Advancement in Machine Learning methods to analyze complex datasets

Exponential increase in amount of data: With the amount of published information and collected data rising exponentially, it is now estimated that 90% of the data in the world today has been created in the past two years alone³. This data flood is expected to increase the accumulated digital universe of data from 4.4 zettabytes (or trillion gigabytes) in late 2015 to 44 zettabytes by 2020⁴. The Internet of Things (IoT) phenomenon driven by the embedding of networked sensors into home appliances, collection of data through sensors embedded in smart phones (with ~1.4 billion units shipped in 2016⁵), and reduction of cost in satellite technologies lends support to further acceleration in the collection of large, new alternative data sources.

Increases in computing power and storage capacity: The benefit of parallel/distributed computing and increased storage capacity has been made available through remote, shared access to these resources. This development is also referred to as **Cloud Computing**. It is now estimated that by 2020, over one-third of all data will either live in or pass through the cloud⁶. A single web search on Google is said to be answered through coordination across ~1000 computers⁷. Open source frameworks for distributed cluster computing (i.e. splitting a complex task across multiple machines and aggregating the results) such as Apache Spark⁸ have become more popular, even as technology vendors provide remote access classified into Software-as-a-service (SaaS), Platform-as-a-service (PaaS) or Infrastructure-as-a-service (IaaS) categories⁹. Such shared access of remotely placed resources has dramatically diminished the barriers to entry for accomplishing large-scale data processing and analytics, thus opening up big/alternative data based strategies to a wide group of both fundamental and quantitative investors.

Machine Learning methods to analyze large and complex datasets: There have been significant developments in the field of pattern recognition and function approximation (*uncovering relationship between variables*). These analytical methods are known as ‘**Machine Learning**’ and are part of the broader disciplines of **Statistics** and **Computer Science**. Machine Learning techniques enable analysis of large and unstructured datasets and construction of trading strategies. In addition to methods of **Classical Machine Learning** (that can be thought of as advanced Statistics), there is an increased

² See our primers titled “[Systematic strategies across assets](#)”, “[Momentum strategies across asset classes](#)”, and “[Equity risk premia strategies](#)” for performance of value, momentum, carry and short volatility risk premia across equities, bonds, currencies and commodities.

³ Source: IBM, “[Bringing Big Data to the enterprise](#)”.

⁴ Source: EMC, “[The digital universe of opportunities: Rich data and the increasing value of the Internet of Things](#)”.

⁵ Source: Statista, the Statistics Portal, “[Statistics and facts about smartphones](#)”.

⁶ Source: CSC, “[Big data universe beginning to explode](#)”.

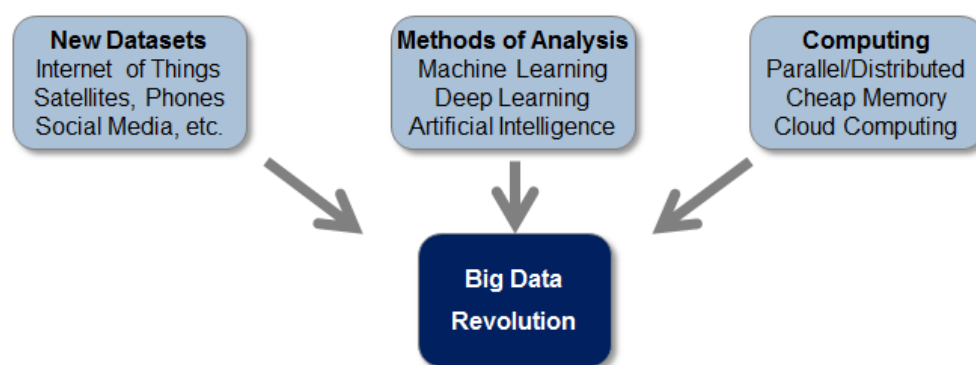
⁷ Source: Internet Live Stats, “[Google search statistics](#)”.

⁸ Source: Apache Spark, “[Main webpage](#)”.

⁹ Source: Rackspace, “[Understanding the cloud computing stack: SaaS, PaaS, IaaS](#)”.

focus on investment applications of **Deep Learning** (an analysis method that relies on multi-layer neural networks), as well as **Reinforcement learning** (a specific approach that is encouraging algorithms to explore and find the most profitable strategies). While neural networks have been around for decades¹⁰, it was only in recent years that they found a broad application across industries. The year 2016 saw the widespread adoption of smart home/mobile products like Amazon Echo¹¹, Google Home and Apple Siri, which relied heavily on Deep Learning algorithms. This success of advanced Machine Learning algorithms in solving complex problems is increasingly enticing investment managers to use the same algorithms.

Figure 1: Factors leading to Big Data Revolution



Source: J.P.Morgan Macro QDS.

While there is a lot of hype around Big Data and Machine Learning, researchers estimate that just 0.5% of the data produced is currently being analyzed [Regalado (2013)]. These developments provide a compelling reason for market participants to invest in learning about new datasets and Machine Learning toolkits.

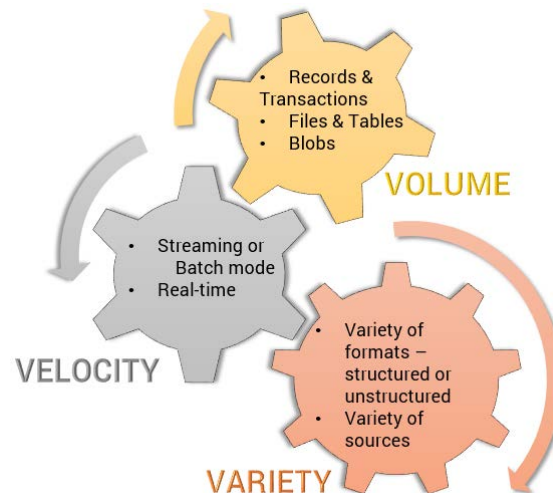
This primer provides a comprehensive overview of Big and Alternative Data sources, Machine Learning methods, and their relevance to both fundamental and quantitative strategies. As there are quite a lot of terms commonly used to describe Big Data, we provide brief descriptions of Big Data, Machine Learning and Artificial Intelligence below.

Big Data: The systematic collection of large amounts of novel data over the past decade followed by their organization and dissemination has led to the notion of Big Data; see Laney (2001). The moniker 'Big' stands in for three prominent characteristics: **Volume:** The size of data collected and stored through records, transactions, tables, files, etc. is very large; with the subjective lower bound for being called 'Big' being revised upward continually. **Velocity:** The speed with which data is sent or received often marks it as Big Data. Data can be streamed or received in batch mode; it can come in real-time or near-real-time. **Variety:** Data is often received in a variety of formats – be it structured (e.g. SQL tables or CSV files), semi-structured (e.g. JSON or HTML) or unstructured (e.g. blog post or video message).

¹⁰ Inspired originally from models of neurons in the human brain as in the work of McCulloch-Pitts (1945)

¹¹ Source: [Amazon Developer Blogs](#).

Figure 2: Features of Big Data



Source: J.P.Morgan Macro QDS.

Big and alternative datasets include **data generated by individuals** (social media posts, product reviews, internet search trends, etc.), **data generated by business processes** (company exhaust data, commercial transaction, credit card data, order book data, etc.) and **data generated by sensors** (satellite image data, foot and car traffic, ship locations, etc.). The definition of alternative data can also change with time. As a data source becomes widely available, it becomes part of the financial mainstream and is often not deemed 'alternative' (e.g. Baltic Dry Index – data from ~600 shipping companies measuring the demand/supply of dry bulk carriers). Alternative and Big Datasets are the topic of the second chapter in this report.

Machine Learning (ML): Machine Learning is a part of the broader fields of Computer Science and Statistics. The goal of Machine Learning is to enable computers to learn from their experience in certain tasks. Machine Learning also enables the machine to improve performance as their experience grows. A self-driving car, for example, *learns* from being initially driven around by a human driver; further, as it drives itself, it *reinforces* its own learning and gets better with experience. In finance, one can view Machine Learning as an attempt at *uncovering relationships between variables*, where given historical patterns (input and output), the machine forecasts outcomes out of sample. Machine Learning can also be seen as a *model-independent (or statistical or data-driven) way for recognizing patterns* in large data sets. Machine Learning techniques include **Supervised Learning** (methods such as regressions and classifications), **Unsupervised Learning** (factor analyses and regime identification) as well as fairly new techniques of **Deep and Reinforced Learning**. Deep learning is based on neural network algorithms, and is used in processing unstructured data (e.g. images, voice, sentiment, etc.) and pattern recognition in structured data. Methods of Machine Learning are the topic of the third part of this report.

Artificial Intelligence (AI): Artificial Intelligence is a broader scheme of enabling machines with human-like cognitive intelligence (note that in this report, we sparsely use this term due to its ambiguous interpretation). First attempts of achieving AI involved hardcoding a large number of rules and information into a computer memory. This approach was known as 'Symbolic AI', and did not yield good results. Machine Learning is another attempt to achieve AI. Machine Learning and specifically Deep Learning so far represent the most serious attempt at achieving AI. Deep Learning has already yielded some spectacular results in the fields of image and pattern recognition, understanding and translating languages, and automation of complex tasks such as driving a car. While Deep Learning based AI can excel and beat humans in many tasks, it cannot do so in all. For instance, it is still struggling with some basic tasks such as the Winograd Schema Challenge. More on the historical development of Big Data, Machine Learning and Artificial Intelligence can be found in the Appendix.

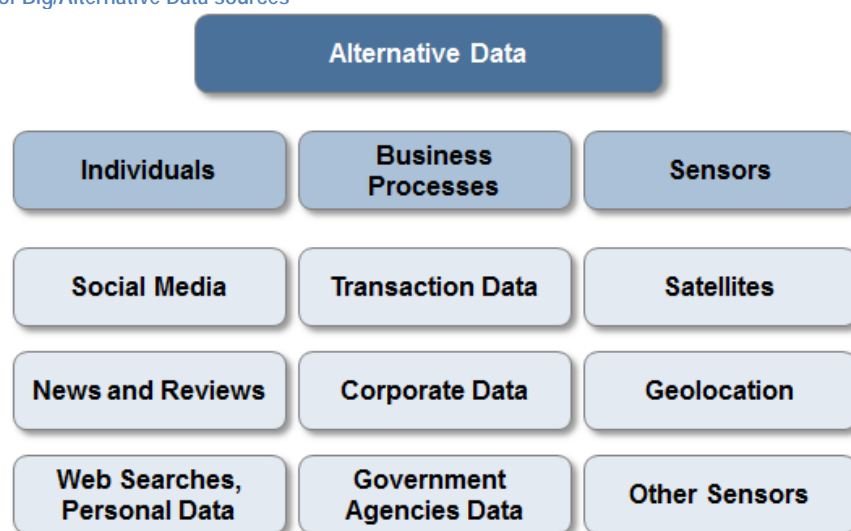
Classification of Alternative Data Sets

At the core of the Big Data transformation in investment management are the new sources of data that can provide informational advantages. The advantage given by data can be in the form of uncovering new information not contained in traditional sources, or uncovering the same information but at an earlier time. For instance, satellite imagery of mines or agricultural land can reveal supply disruptions before they are broadly reported in the news or official reports.

In this section we aim to provide a framework or classification of Big Data. First, we classify data based on the manner in which the data was generated. Then we consider attributes of datasets that are directly relevant for investment professionals such as mapping datasets to an asset class or investment style, alpha content, quality of data, technical specifications, etc. In the next chapter of this report, we will cover different types of data in greater detail, discuss specific datasets and providers, and provide specific examples and tests of Big Data trading strategies.

We first classify data sources at a high level by indicating whether they were **produced by individuals** (such as social media posts), **generated through business processes** (such as e-commerce or credit card transactions data), or **generated by sensors** (such as satellite imagery, radar, etc.). Figure 3 below shows this classification. This approach expands upon earlier attempts in a non-financial context by Kitchin (2015) and in a United Nations report (2015). While this taxonomy is somewhat theoretical, there are common features, common methods to analyze and common challenges for data in each of these three categories. For instance, data generated by individuals are often in the unstructured textual format, and commonly require natural language processing. Sensor generated data tends to be unstructured and may require analysis techniques such as counting objects, or removing the impact of weather/clouds from a satellite image, etc. Many business generated datasets, such as credit card transactions and company ‘exhaust’ data have common legal and privacy considerations.

Figure 3: Classification of Big/Alternative Data sources



Source: J.P.Morgan Macro QDS.

Data generated by Individuals: Mostly recorded through textual mediums, such data is often both unstructured and distributed across multiple platforms. One can further classify it into: 1) data from social media (websites like Twitter, Facebook, LinkedIn, etc.), 2) specialized sites such as business-reviewing websites like Yelp, E-commerce groups like Amazon, and Mobile app analytics companies like App Annie, 3) Web searches, and personalized data such as Google Search trends, data from personal inboxes, etc.


Data generated by Business Processes refer to data produced or collected by corporations and public entities. An important sub-category is transaction records such as credit card data. Corporate data can be a byproduct or ‘exhaust’ of corporate record-keeping such as banking records, supermarket scanner data, supply chain data, etc. Data generated by business processes is often highly structured (in contrast to human-generated data) and can act as a leading indicator for corporate metrics, which tend to be reported at a significantly lower frequency. Business process generated data can arise from public agencies; an example would be the large federal data pools made available online over the past few years by the US government.

Data generated by sensors: Mechanically collected data through sensors embedded in various devices (connected to computers, wireless and networking technologies). The data generated is typically unstructured and its size is often significantly larger than either human or process-generated data streams. The most well-known example is satellite imaging that can be used to monitor economic activities (construction, shipping, commodity production, etc.). Geolocation data can be used to track foot traffic in retail stores (volunteered smartphone data), ships in ports, etc. Other examples of sensors include cameras fixed at a location of interest, weather and pollution sensors, etc. Perhaps the most promising is the future concept of the Internet of Things (IoT) - the practice of embedding micro-processors and networking technology into all personal and commercial electronic devices.

After classifying data according to whether they were produced by individuals, business processes or sensors, we provide another descriptive classification that may be of more interest for investment professionals. For example, a Retail sector portfolio manager will care to identify store specific sales data, regardless of whether they were created by satellite imaging of a parking lot, volunteered customer geolocation data, or e-receipts in customers’ inboxes. High frequency quant traders will care about all signals that can be produced on an intraday basis such as tweets, news releases, etc. but will care less about e.g. credit card data that come with substantial delays and are less broadly followed. Data quants will prefer to classify data based on methods used to collect and analyze them, such as the technology to retrieve data, treat outliers and missing values, etc. In this ‘Investment Classification’ we assign various attributes to each big/alternative data set. These are the attributes most relevant for investment professionals such as CIOs, Portfolio Managers, Traders and Data quants. The figure below shows this classification:

Figure 4: Attributes of an alternative data set

Asset Class	Investment Style	Alpha (Net of Cost)	Known	Stage of Processing	Quality	Technical Aspects
Equity	Macro	Viable Stand alone	Public Free of cost	Raw	History	Frequency
Commodity	Sector Specific	Viable In a Portfolio	Well Known	Semi Processed	Outliers	Latency
Credit	Stock Specific	Not Viable	Lesser Known	Processed	Missing Values	Format
Rates	Risk Indicator	Capacity	Proprietary Not Known	Trading Signal	Methodology Transparency	Robust API
FX	Quant Signal	Orthogonality	Limited Sales Deals	Research Piece or Alert	Support Structure	Conflicts and Legal Risk



CIOs and
Portfolio Managers

Quants and
Data Scientists

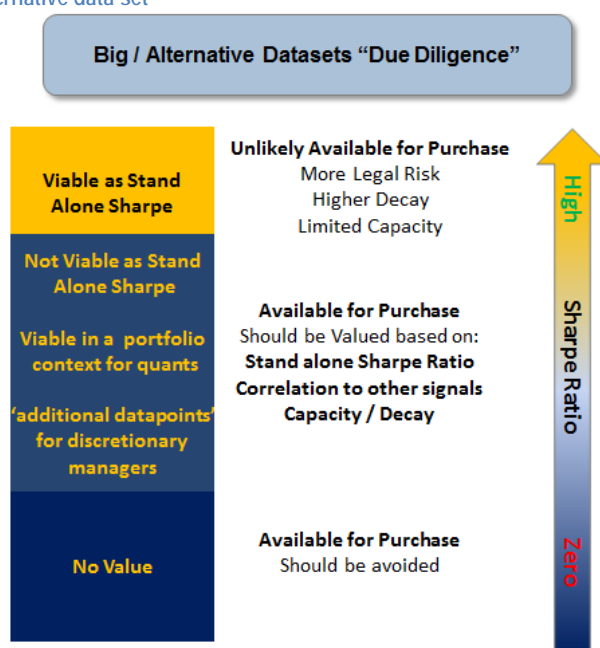
Source: J.P.Morgan Macro QDS.

The first attribute of a dataset is its relevance for **1) Asset class**. Most Big Data are still focused on equities and commodities. There is relatively little alternative data on interest rates and currencies, making such data sets more valuable

to select investors. In terms of **2) Investment style** – most data are sector and stock specific and relevant for equity long-short investors. There is also a significant amount of data relevant for macro investors (e.g. consumer credit, China economic activity, shipping data, etc.). Certain alternative datasets can be used to substitute traditional metrics of market risk, and some signals are relevant only for e.g. high frequency quant traders.

Perhaps the most important data attribute is its potential **3) Alpha content**. Alpha content has to be analyzed in the context of the price to purchase and implement the dataset. **Costs** of alternative datasets vary widely – sentiment analysis can be obtained for a few hundred or thousand dollars, while comprehensive credit card data can cost up to a few million USD a year. Trading strategies based on alternative data are tested, and Alpha is estimated from a backtest. These tests can find whether a dataset has enough alpha to make it a **viable standalone** trading strategy. These situations are rare. Most data have a small positive Sharpe ratio that is not sufficiently high for a standalone investment strategy. Despite this, these datasets are very valuable, as the signals can be combined with other signals to yield a **viable portfolio level** strategy. Investors should not be surprised to come across alternative datasets with no alpha content. In addition to alpha, one needs to **assess orthogonality** of the information contained in the dataset (is it unique to a dataset, or already captured by other data), as well as the potential **capacity** of a strategy based on the dataset. The figure below shows potential outcomes of an “alpha assessment” of a dataset.

Figure 5: Information content of an alternative data set



Source: J.P.Morgan Macro QDS

Closely related to the alpha content, is the question of **4) How well-known is a dataset**. The more broadly a dataset is known, the less likely it is to lead to a stand-alone strategy with a strong Sharpe ratio. **Well-known public datasets** such as financial ratios (P/E, P/B, etc.) likely have fairly low alpha content and are not viable as a standalone strategies (they may still be useful in a diversified risk premia portfolio). Most Big Datasets will be **less well-known** and new datasets emerge on a frequent basis. To assess how well a dataset is known, managers can ask the data provider about existing clients. Initial clients can influence the scope of data collection and curation affecting the subsequent customers. Initial clients can sometimes ask for **exclusive** or limited-sales deals, through which the provider commits to sell only to a pre-defined number of clients.

An important attribute of data is the **5) Stage of processing** of data when acquired. Fundamental investors prefer processed signals and insights instead of a large amount of raw data. The highest level of data processing happens when data is presented in the form of **research reports**, alerts or trade ideas. A lesser degree of processing comes when the provider sells

a **signal** that can be directly fed into a multi-signal trading model. Most big/alternative datasets come in a **semi-processed format**. Such data is semi-structured with data presented in tabular (such as CSV) or tagged (such as JSON/XML) format. Semi-processed data still has some outliers and gaps, and is not readily usable as input into a trading model. Its alpha content and mapping to tradable instruments needs to be assessed, and one needs to consider the economics driving seasonality, outliers, etc. Finally, **raw data** is likely to be of little use for most investors. For instance a satellite image file of oil tanks would be of little use for someone with no expertise in image processing, adjusting for seasonal and weather effects, or identifying types of storage, etc.

6) Quality of data is another important feature, especially so for data scientists and quants. Data with longer **history** is often more desirable for the purpose of testing (for satellite imagery typically > 3 years, sentiment data > 5 years, credit card data > 7 years; conversely, datasets with less 50 points are typically less useful). **Gaps or outliers** are an important consideration. If data has been backfilled, the method of imputation of missing values must be mentioned. It must be specified, if the missing data was missing at random or had patterns. Consider the case where we want to track oil storage inventory time-series at different locations. Missing data can be classified as: a) Missing completely at random (MCAR): missing value for storage has no relation to actual value of storage or location; b) Missing at random (MAR): certain locations tend to miss reporting storage levels regularly, but their omission of value does not stem from the actual value of storage at that location; or c) Missing not at random (MNAR): missing value of inventory has predictive connection to the level. **Transparent Methodology** is necessary to assess if the sample is representative of the total population, how it's adjusted for errors (e.g. credit card regional and demographic biases, satellite data corrected for weather effects, etc.). Alternative data has no standardized format; additionally, sampling methodology and understanding of the dataset often evolves. For this reason, data providers should have a **robust support structure** for clients.

There are a number of **7) Technical Aspects** of a big and alternative datasets. **Frequency** of data: can be intra-day, daily, weekly, or even lower frequency. **Latency**: Data providers often provide data in batches, and a delay is possible either due to collection, operational or legal constraints. **Format**: the data must be extracted in a suitable format, preferably CSV or JSON for static data. The **API** (Application Programming Interface) should be robust. It should not fail or result in additional latency, and it should be flexible to accommodate different programming languages. There could be potential **conflicts of interest** if the data provider is trading using the dataset. **Legal and reputational risk**: usage of most of the alternative data sets carries little legal or reputational risk. However, clients should understand and even anticipate legal risks of buying information that is not available broadly.

Classification of Machine Learning Techniques

Large and less structured datasets often can't be analyzed with simple spreadsheet work and scatterplot charts. New methods are needed to tackle the complexity and volume of new datasets. For instance, automated analysis of unstructured data such as images, social media, and press releases is not possible with the standard tools of financial analysts. Even with a large traditional dataset, using a simple linear regression often leads to overfitting or inconsistent results. Methods of Machine Learning can be used to analyze Big Data, as well as to more efficiently analyze traditional datasets.

There is no question that techniques of Machine Learning yielded some spectacular results when applied to problems of image and pattern recognition, natural language processing, and automation of complex tasks such as driving a car. What is the application of Machine Learning in finance, and how do these methods differ from each other?

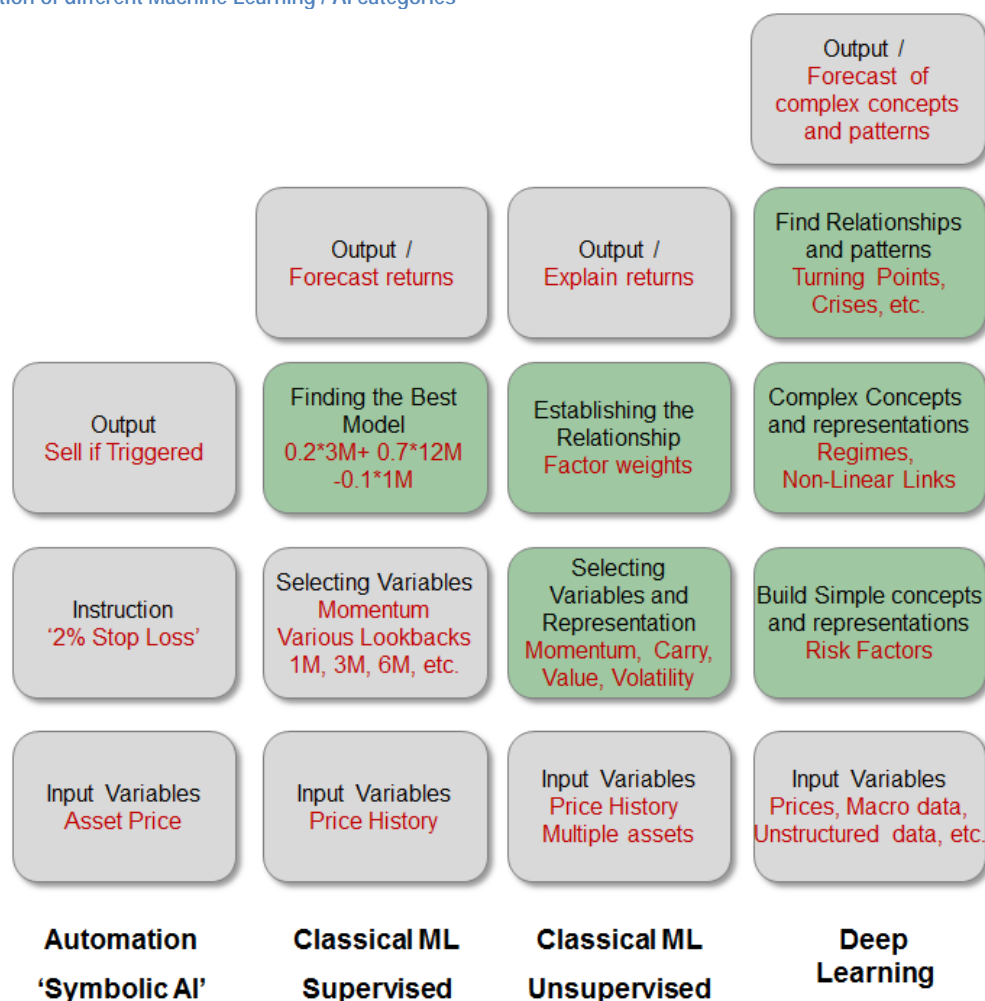
First let's note that automation of tasks is not considered Machine Learning. We can instruct a computer to perform certain operations based on a fixed set of rules. For instance we can instruct a computer to sell an asset if the asset price drops by a certain amount (stop loss). Even giving a large number of complex rules to a machine - also known as '**Symbolic Artificial Intelligence**' - does not represent Machine Learning, but rather the automation of tasks. With this 'Symbolic AI', the machine will freeze the first time it encounters a situation that does not exactly match a set of pre-programmed rules. In **Machine Learning**, the computer is given an input (set of variables and datasets) and output that is a consequence of the input variables. The machine then finds or 'learns' a rule that links the input and output. Ultimately the success of this learning task is tested 'out of sample' - its ability to gain useful knowledge of the relationship between variables and predict outcomes in yet unseen situations. Machine Learning can be supervised or unsupervised. In **Supervised Learning** we are trying to find a rule, an 'equation' that we can use to predict a variable. For instance, we may want to look for a momentum (trend following) signal that will have the best ability to predict future market performance. This may be accomplished by running advanced regression models to assess which one has higher predictive power, and is most stable to regime changes. In **Unsupervised Learning**, we are uncovering the structure of data. For instance, we may take market returns and try to identify the main drivers of the market. For instance, a successful model may find that at one point in time, the market is driven by the momentum factor, energy prices, level of USD, and a new factor that may be related to liquidity. **Deep Learning** is a Machine Learning method that analyzes data in multiple layers of learning (hence 'deep'). It may start doing so by learning about simpler concepts, and combining these simpler concepts to learn about more complex concepts and abstract notions. It is often said that the goal of automation (or 'Symbolic AI') is to perform tasks that are easy for people to define, but tedious to perform. On the other hand the goal of Deep Learning AI systems is to perform tasks that are difficult for people to define, but easy to perform. Deep Learning is in its essence more similar to how people learn, and hence is a genuine attempt to artificially recreate human intelligence.¹² For instance, a child will learn to recognize the concept of 'face' by looking at pictures and identifying some simple features like eyes, nose, mouth, etc. From simple features one builds a more complex concept such as relationship between these features (eyes are above nose, etc.) and their relative importance (e.g. is a face without one eye still identified as a 'face'). By the process of recognizing these features and linking them, a child will be able to recognize previously unseen examples of: animal faces, fictional characters, or completely stylized representations such as emoticons. Only a flexible learning system - as opposed to 'hardcoded' symbolic AI - can achieve this level of pattern recognition and generalization.

Deep Learning is used in pre-processing of unstructured Big Data sets (for instance, it is used to count cars in satellite images, identify sentiment in a press release, etc.). An illustration of Deep Learning in a hypothetical financial time series example would be to predict (or estimate) the probability of a market correction. For instance, we may feed a large number of datasets into a Deep Learning model. The model may first identify some simple features that negatively affect the market such as a momentum breakdown, an increase in volatility, a decline in liquidity, etc. Each of these factors may not lead on its own to a market correction. Further, the algorithm may identify patterns between these simple features and non-linear relationships between them. From those models it can build more complex features such as EM driven crises, financial stress, that eventually may lead to more significant market corrections or even recessions.

¹² It is likely not a coincidence that Deep Learning methods are based on neural networks - which are in turn inspired by the way neurons are connected in human brain.

The figure below illustrates various categories of Machine Learning / Artificial Intelligence and potential applications in trading strategies. The steps represented by the grey boxes are initially provided to the algorithm (as part of the training set), and green boxes are generated by the Machine Learning algorithm.

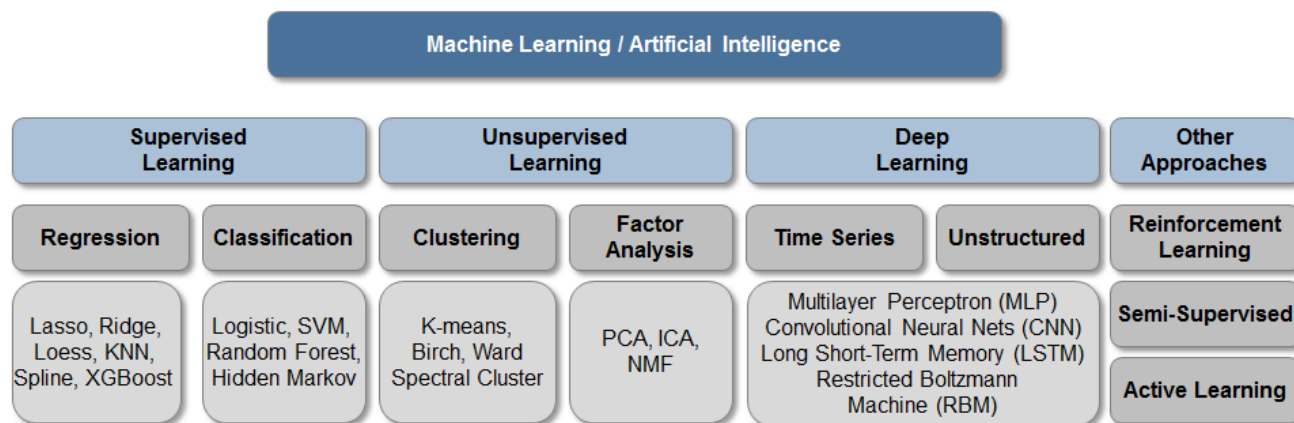
Figure 6: Illustration of different Machine Learning / AI categories



Source: J.P.Morgan Macro QDS

Now that we have broadly classified Machine Learning techniques into Supervised Machine Learning, Unsupervised Machine Learning, and Deep Learning, let's look in more detail into the tasks these methods aim to accomplish and some specific models. The figure below shows our classification of Machine Learning and examples of different methods.

Figure 7: Classification of Machine Learning techniques



Source: J.P.Morgan Macro QDS

Supervised and Unsupervised Learning are often called Classical Machine Learning. In **Supervised Learning**, an algorithm is provided historical data (both input and output variables), and is trying to find the relationship that has the best predictive power for out of sample data. Methods of supervised learning are further classified into regression and classification.

Regressions try to predict output variables based on a number of input variables. An example would be trying to predict how much the market will move if there is a sudden spike in inflation. **Classification** methods attempt to group or classify output into categories. For instance, we may want the output of a model to be a binary action such as ‘buy market’ or ‘sell market’ based on a number of macro, fundamental, or market input variables.

Even a simple linear regression can be thought of as a Supervised Machine Learning method. However, linear regressions may be too simplistic to uncover the true relationship between variables. For instance, a very large increase in inflation may be bad for market returns, and linear regression would fail to capture such a ‘non-linear’ relationship.¹³ One simple method of Machine Learning regression is called **Lasso regression**. Lasso regression tries to establish the relationship (a forecast) by choosing the smallest and most relevant set of input variables. **K-nearest neighbors** regressions forecast the data by simply looking at the historical samples and establishing what has happened in similar situations as a best forecast of the future. Another algorithm, **Logistic regression**, is a classification method tailored to handling data where the output is a binary decision, e.g. “buy” or “sell” (despite the name ‘logistic regression’ this method is actually classification). **Decision tree** classifications try to find the optimal rule to forecast an outcome based on a sequence of simple decision steps. For instance, by looking at past stock performance, we may find that a simple rule to choose a ‘winning stock’ is to buy if earnings have positive momentum, stock price has positive momentum, and if valuation is below the universe median. Such a simple ‘decision tree’ may explain past returns but not work well out of sample. **Random Forests** is a classification Machine Learning method that is based on ‘decision trees’. Random Forests are averaging simple decision tree models (e.g. calibrated over different historical episodes) and they often yield better and more reliable forecasts as compared to decision trees.

Unsupervised learning algorithms examine the dataset and identify relationships between variables and their common drivers. In unsupervised learning the machine is simply given the entire set of returns of assets and it does not have a notion of what are independent and what are the dependent variables. Methods of unsupervised learning can often be categorized as either Clustering or Factor analyses. **Clustering** involves splitting a dataset into smaller groups based on some notion of similarity. In finance that may involve identifying historical regimes such as high/low volatility regime, rising/falling rates regime, rising/falling inflation regime, etc. Correctly identifying the regime can in turn be of high importance for allocation between different assets and risk premia. One example of a clustering method is the **K-means** technique in Machine Learning. This method splits the data into K subsets of data in such a way to minimize the dispersion of points within each cluster. **Factor analyses** aim to identify the main drivers of the data or identify best representation of the data. For instance,

¹³ Linear regression also cannot reliably deal with large number of variables and data outliers

yield curve movements may be described by parallel shift of yields, steepening of the curve, convexity of the curve. In a multi asset portfolio, factor analysis will identify the main drivers such as momentum, value, carry, volatility, liquidity, etc. A very well-known method of factor analysis is **Principal Component Analysis** (PCA). The PCA method carries over from the field of statistics to unsupervised Machine Learning without any changes.

Neural network techniques are loosely inspired by the working of the human brain. In a network, each neuron receives inputs from other neurons, and ‘computes’ a weighted average of these inputs. If this weighted average exceeds a certain threshold, the neuron sends out an output to other neurons, ultimately leading to a final output (in this sense, neurons are simple mathematical functions). The relative weighting of different inputs is guided by the past experience, i.e. based on some training set of inputs and outputs. Computer scientists have found that they could mimic these structures in an approach called **Deep Learning**. Specifically, Deep Learning is a method to analyze data by passing it through multiple layers of non-linear processing units - neurons. Once when the signal weightings are calibrated from the sample dataset (training/learning dataset) these models have strong out of sample predictive power. Multiple layers of signal processing units (i.e. neurons) allow these models to progressively learn more complex concepts out of simpler ones. Certain types of Deep Learning architectures are more suitable for analysis of time series data, and others are more suitable for analyses of unstructured data such as images, text, etc. **Multilayer Perceptron** (MLP) is one of the first designs of multi-layer neural networks, designed in such a way that the input signal passes through each node of the network only once (also known as a ‘feed-forward’ network). **Convolutional Neural Networks** (CNN) are often used for classifying images. They extract data features by passing multiple filters over overlapping segments of the data (this is related to mathematical operation of convolution). **Long Short-term memory** (LSTM) is a neural network architecture that includes feedback loops between elements. This can also simulate memory, by passing the previous signals through the same nodes. LSTM neural networks are suitable for time series analysis because they can more effectively recognize patterns and regimes across different time scales.

Other Machine Learning approaches include reinforcement learning, semi-supervised learning and active learning. Especially promising is the approach of **Reinforcement learning**, where the goal is to choose a course of successive actions in order to maximize the final (or cumulative) reward. For instance, one may look for a set of trading rules that maximizes PnL after 100 trades. Unlike supervised learning (which is typically a one step process), the model doesn’t know the correct action at each step. At the core of reinforcement learning are two challenges that the algorithm needs to solve: 1) Explore vs. Exploit dilemma - should the algorithm explore new alternative actions that may maximize the final reward, or stick to the established one that maximizes the immediate reward. 2) Credit assignment problem - given that we know the final reward only at the last step, it is not straightforward to assess which step was critical for the final outcome. **Semi-supervised learning** combines elements of supervised and unsupervised learning, and **Active learning** is an approach that actively selects and analyzes datasets that are most beneficial to solving the task at hand.

Selecting the best model

While applying a Machine Learning method to a dataset is a science, choosing and calibrating a specific model has elements of art. Computer scientists often refer to the ‘No Free Lunch Theorem’ that states that there is no one Machine Learning algorithm that gives the best result when applied to different types of data. In particular some models may ‘over fit’ the data: they may look good on a backtest but perform poorly on out of sample data. Stability of out of sample forecast is a challenge often encountered with Risk Premia strategies. Big Data and Machine Learning strategies are not exempt from this challenge.

At the core of this issue is theory of **Variance-Bias Tradeoff**. Variance-Bias Tradeoff states that an out of sample forecast will deteriorate because of three factors: in-sample forecast error, model instability, and error due to our inability to forecast completely random events. More specifically:

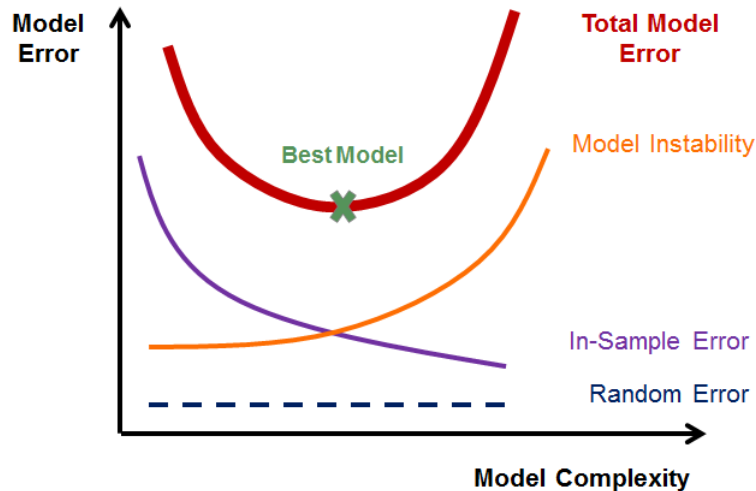
Forecast Error = In-Sample Error + Model Instability + Random Error

Two major factors impacting the quality of our forecasts are in-sample error, and model instability. If our model poorly fits the historical data – resulting in a large ‘**In-sample error**’ – most likely it will also poorly describe the future, leading to

error in our forecast. This is called '**model bias**'. We can always reduce this 'model bias' by increasing the **model complexity**. By adding more parameters (bells and whistles) to the model we can reduce in-sample error to better fit historical data (i.e. make the backtest look better). But this can lead to 'overfitting' and likely result in much larger errors out of sample. Increasing model complexity always leads to higher instability of the model (when applied to different historical datasets or out of sample data). This instability is also called '**model variance**' and contributes to higher forecast error. The 'art' part of Machine Learning is selecting a **model that will find the optimal balance between in-sample error and model instability** (tradeoff between 'model bias' and 'model variance'). In almost all cases of financial forecasts, we can model the future only to a certain degree. There are always random, idiosyncratic events that will add to the error of our forecasts.

The quality of our forecast will largely be a function of model complexity: more complex models will reduce in-sample error, but will also result in higher instability. According to mathematical theory, there is always an optimal selection of model and its complexity that will minimize the error of our forecast. This is illustrated in the figure below.

Figure 8: Tradeoff between 'model bias' and 'model variance'



Source: J.P.Morgan Macro QDS

In order to select the most appropriate (best possible) method to analyze data one needs to be familiar with different Machine Learning approaches, their pros and cons, and specifics of applying these models in financial forecasts. In addition to knowledge of models that are available, successful application requires a strong understanding of the underlying data that are being modelled, as well as strong market intuition. In the third chapter of this report we review in more detail various Machine Learning models and illustrate their application with financial data.

Positioning within the Big Data Landscape

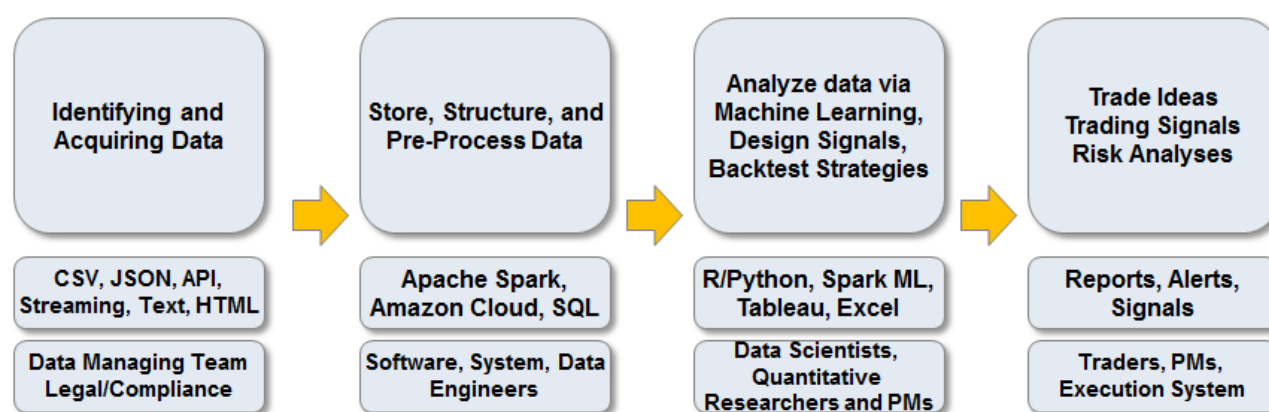
We think the Big Data and Machine Learning revolution will profoundly change the investment landscape. As more investors adopt Big Data, the market will start reacting faster and will increasingly anticipate traditional or 'old' data sources. This will give an edge to quant managers and those willing to adopt and learn about new datasets and methods of analysis. Those that don't learn and evolve will be at risk of becoming obsolete. Regardless of the timeline of these changes, we believe that **analysts, portfolio managers, traders and CIOs will eventually have to become familiar with Big Data and Machine Learning developments and related trading strategies**. This applies to both fundamental and quantitative investors, and is true across asset classes.

Potential Pitfalls of Big Data: The transition to a Big Data framework will not be without setbacks. Certain types of **data** may lead into blind alleys - datasets that don't contain alpha, signals that have too little investment capacity, decay quickly, or are simply too expensive to purchase relative to their benefit. Managers may invest too much into unnecessary infrastructure, e.g. build complex models and architecture that don't justify marginal performance improvements. Machine Learning algorithms cannot entirely replace human intuition. Sophisticated models, if not properly guided, can overfit or uncover spurious relationships and patterns. **Talent** will present another source of risk – employing data scientists who lack specific financial expertise or financial intuition may not lead to desired investment results. Many Big Data concepts and methods may sound plausible and appealing but will not lead to viable trading strategies.

Given the risks and uncertain rewards, many investors are wondering how far and how fast to go with changes when adopting a more quantitative, data driven investment style.

Before we discuss how to go about building the effort (e.g. how much to outsource, what talent is required to build a Big Data/ML effort in-house, typical technology setups, etc.) we want to briefly outline the main steps in implementing a Big Data strategy. This is shown in the Figure below.

Figure 9: Big Data workflow for investment managers



Source: J.P.Morgan Macro QDS

One first needs to **identify and acquire data**. A data-sourcing team licenses new data sources either directly through data owners/vendors, or specialized firms that aggregate third-party data (and match vendors and end users). Once the licensing agreement is in place, the data needs to be **stored and pre-processed**. Big Data is rarely available in the clean format that can be fed directly into Machine Learning algorithms. A team of quantitative researchers is devoted to the task of pre-processing the data (e.g. detecting outliers, missing values, etc.). The **data science team** consists of quantitative researchers who analyze the data with the aid of; Machine Learning, backtest strategies, and visualization techniques, with a goal of deriving tradable signals or insights based on the data. **Finally, the signals are implemented** by portfolio managers, or in some cases executed in an automated fashion (which would involve an additional layer of system and software experts linking signals to execution systems).

Outsourcing vs. Internalizing

In the 'Big Data' transformation, an institution can use different levels of outsourcing:

Fully Outsourcing: One can decide to not build any of the Big Data infrastructure, and rely on third-party research reports and trade recommendations and signals (based on Big Data). These reports and analyses can be provided by sell side research, independent research firms, and firms that specialize in data gathering, analysis and aggregation. An advantage of this approach is that no expertise in Big Data and Machine Learning is needed, so it can serve as a first step in learning about the field. A disadvantage is there's little exclusivity of these insights which have broader availability.

Partially Outsourcing: The manager can buy datasets and analyze them in-house to derive proprietary trading signals and insights. Purchased data can be raw, semi-processed or fully processed. Advantages include the flexibility in crafting signals most suitable for the manager's investment style and relatively modest investment in infrastructure and quantitative talent. We believe most fundamental and traditional quantitative investors will chose this route.

Fully Internalize: Collect, store, and process datasets internally. Internal datasets can be acquired by e.g. web scraping, twitter and internet searches, and crowd-sourcing. Acquire only raw data and maintain maximal control of data processing (de-trending, treatment of outliers, missing data, etc.). Creating proprietary signals from external raw data such as satellite images, bulk transaction data, etc. This approach involves building a Big Data research and technology team and requires a more substantial investment of resources. Many advanced quant investors will follow this route in our view.

We also note that managers can combine these approaches in different areas (e.g. outsource satellite data signals, but internalize company specific sentiment signals, etc.). Investors can also start with a high level of outsourcing, and once they gain traction and confidence, start internalizing some of the Big Data techniques and processes.

Big Data Talent

Based on the choice of which aspects of the Big Data program should be run internally and which externally, a fund manager may look for the appropriate talent to accomplish these tasks. Different skill-sets are needed to partially or fully internalize and integrate Big Data into the investment process.

Capturing, storing and sharing data is often supported by software engineers. The skillset for the role of data scientists is virtually the same as for any other quantitative researchers. These include various buy side and sell side quants with typical educational backgrounds in Computer Science, Statistics/Mathematics, Financial Engineering/Econometrics and Natural sciences. Some experience with programming and knowledge of advanced statistics is necessary. In the past few years, Big Data architectures have largely become programming language-agnostic so there is no strong preference for one programming language. However we do see a more common use of Python and R as compared to others. For analysts with good quantitative background, Machine Learning methods can be relatively easily learned from primers such as this report and other more specialized textbooks. Most of the Machine Learning methods are already coded (e.g. in R), so one just needs to learn how to use these packages rather than code models from scratch.

Perhaps more important than Big Data technical expertise, is experience in working with quantitative trading strategies. Designing and testing many tradable strategies builds intuition on assessing data quality, tradability, capacity, variance-bias tradeoff, and economics driving returns. It is much easier for a quant researcher to change the format/size of a dataset, and employ better statistical and Machine Learning tools, than for an IT expert, silicon valley entrepreneur, or academic to learn how to design a viable trading strategy.

We believe that many fund managers will get the problem of Big Data talent wrong, leading to culture clashes, and lack of progress as measured by PnL generated from Big Data. In addition to the example of skillset/talent mismatch, we have come across individuals misrepresenting their skills and abilities and investment hiring managers unable to distinguish soft skills (such as ability to talk about Artificial Intelligence) and hard skills (such as ability to design a tradeable strategy). This has led to occasional mocking of AI (Artificial Intelligence) by calling it 'Artificially Inflated'.

Finally, Legal and compliance teams are often involved to vet the contracts and to ensure that data is properly anonymized and does not contain material non-public information.

Technology Setup

Technology has a critical role in building a Big Data strategy. Technology involves the tools to perform Big Data analyses such as Machine Learning libraries, and visualization software, as well as computing and storage solutions, be they cloud computing resources or local databases and programing frameworks.

Machine Learning Tools: Many Machine Learning tools developed by technology firms and academia are freely available for use by finance professionals. For example, the language R has a comprehensive set of machine and statistical learning packages¹⁴. Less quantitative analysts might start exploring Machine Learning algorithms on small datasets through GUI-based software like Weka. The language Python also has extensive libraries on data analysis and Machine Learning¹⁵. Python is however more commonly used in Deep Learning research. Keras is a Python library that allows setting up a neural network with just a few lines of instructions. Other popular open source Machine Learning libraries are Google Tensorflow and Theano. Big Data is often multi-dimensional (e.g. might be a mix of different formats such as images, text, numbers) and requires robust **visualization tools**. Data gaps, outliers, trends, etc. are also easier to spot through visual inspection. These tools need to be robust for the data size, real time streaming, flexible and easy to use (e.g. build dashboards, shareable between multiple users, etc.). For example, Tableau has emerged as one of the most popular Big Data visualization tools¹⁶.

Computation and Storage: Many applications of Deep Learning require resources beyond that of an individual personal computer. Also it is not economical to provide each analyst with massive storage and computational power. **Cloud computing** involves centralized resources from which multiple users can provision computing power and storage on a temporary basis. The actual processing can happen in an internal system ("private cloud") or an external vendor system ("public cloud" like Amazon AWS). Most firms are gravitating towards a hybrid cloud system, whereby some critical data remains on a private cloud and others reside on a public one. Within Machine Learning, **graphics processing units** (GPUs) are commonly used to accelerate computational tasks. GPUs are optimized to run a single instruction over multiple segments of data. Google Tensorflow and Theano frameworks automatically handle dispatching computations across multiple GPUs. Big datasets often have to be stored across a number of computers, some of which may fail. Open source programming frameworks such as **Hadoop and Spark** can handle the distributed nature of storing and analyzing data. Hadoop includes a distributed file system called HDFS (storage part); and a system to allocate computational tasks to different processing units and aggregate results called MapReduce. One can think of Spark as an enhancement to Hadoop that is more efficient in processing data and comes with a Machine Learning library (Spark MLlib).

Big Data Compliance

When building a Big Data strategy and putting a dataset into production, one should ensure that all legal and compliance requirements are satisfied. These requirements include the following considerations: potential issues around **obtaining data and terms of use**; potential **Material Non-Public Information** (MNPI) in the dataset, and potential **Personally Identifiable Information** (PII) in a dataset. Data license agreements and terms of use need to be fully understood (e.g. some data may be free for academic use but not for commercial use, data licenses may limit the number of users/teams allowed to access data, obtaining alternative data through scraping of the web may in some cases faces certain [legal challenges](#), etc.). One needs to ensure that data does not contain MNPI. Most Big Data are in the public domain, and individual datasets in most cases do not contain material information. Only through the sophisticated analysis of different datasets, an analyst builds an investment thesis (mosaic theory). While most Big Datasets will be free of MNPI, one needs to stay vigilant and attentive to this risk. For instance geolocation foot traffic for a sample of retail stores that an analyst may use to model sales is free of MNPI. However, obtaining actual sales data from all stores that are otherwise not available

¹⁴ Statistical learning in R is, unlike Python, split across multiple packages; see caret. For visualization, see ggplot2, lattice and ggvis. For data munging, see dplyr, plyr, data.table, stringr and zoo. For visualization in Python, see matplotlib, seaborn, bokeh, networkx.

¹⁵ For data analysis in Python, see pandas (data manipulation), numpy/scipy (numerical computation and optimization), statsmodels/scikit-learn (time-series and machine learning) and pymc3 (Bayesian statistics).

¹⁶ Other visualization software includes older business intelligence tools like IBM Cognos, Tibco Spotfire, QlikView, Microsoft PowerBI as well as D3.js.

publicly would be problematic. Some datasets may include personally identifiable information, and are hence of risk of being discontinued, and the data provider or data user is at risk of being implicated in a lawsuit. In all examples that we could think of, investment firms are interested in aggregated data and not PII. Recent media reports¹⁷ suggest the Federal Trade Commission's division of privacy and identity protection has started scrutinizing alternative data sources and using a dataset with PII poses a significant risk. In the absence of any industry-wide standard, we refer to NIST's 'Guide to protecting the confidentiality of personally identifiable information' published as [NIST 800-122](#) (2010) for creating guidelines for appropriate use of alternative data.

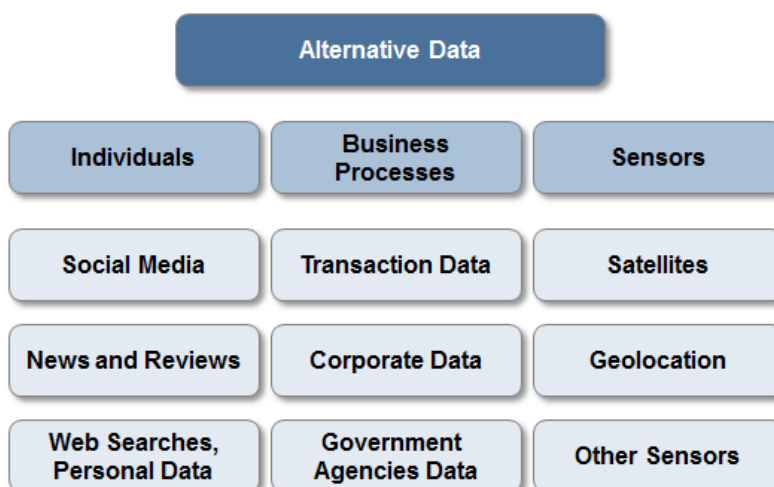
¹⁷ "Big data sellers fail to hide personal information, say funds", Financial Times [article](#) published on Dec 11, 2016.

II: BIG AND ALTERNATIVE DATA

Overview of Alternative Data

In this chapter, we analyze different alternative data sets and illustrate the use of alternative data in specific trading strategies. We follow the classification of alternative data outlined in the previous section (figure below), and provide strategy backtests for select datasets from each category. In the fourth chapter of this handbook (which can be viewed as an extension of this chapter), we provide an extensive directory of alternative data providers.

Figure 10: Alternative data categories



Source: J.P.Morgan Macro QDS

The first step in designing a Big Data trading strategy is identifying and acquiring appropriate datasets. Cost of a dataset is an important consideration. The cost of a dataset involves the direct cost of purchasing the data, and opportunity cost of time invested in analyzing a dataset that may not be put into production. It is not straightforward to assess the relevance and quality of the data and there is little standardization for most data offerings. Initially, one should gather anecdotal intelligence on how well-known and relevant is the dataset. One then needs to scrutinize the quality of the dataset (completeness, outliers, sampling methodology), understand the level of pre-processing, and various technical aspects of the dataset which were discussed in the previous section. Finally, a trading strategy based on the dataset needs to be designed and tested. The backtest should be performed over different time periods, and under different transaction cost assumptions. As with any quantitative strategy, special attention should be paid to avoid overfitting and in-sample biases. In these various steps (from acquiring data to trading implementation of a strategy), managers often partner with various participants in the Big Data market.

Big Data Market

The global market for Big Data, related technology and analytics is [currently estimated](#) at \$130bn, and is expected to grow to over \$200bn by 2020. The financial industry, with ~15% spending share, is one of the important drivers of this growth. Our estimate of the investment management industry's spend on Big Data is in the \$2-3bn range, and the number is expected to have double digit annual growth (e.g. 10-20%, in line with Big Data growth in other industries). This spend includes acquiring datasets, building Big Data technology, and hiring appropriate talent.

Currently, the market of alternative data providers is quite fragmented. Our directory for alternative data providers lists over 500 specialized data firms (see chapter 4). We expect some level of consolidation of data providers as the Big Data market matures. There are roughly three types of data providers in the marketplace: **Providers of raw data** collect and report alternative data with minimal aggregation or processing. Examples are companies collecting foot fall data from individual malls, satellite imagery for requested locations, flows from trading desks, etc. **Providers of semi-processed data** partially

process data by aggregating over geographic regions, industry sectors or map Big Data to specific securities. These firms often produce documentation with some visual proof of relevance for financial assets. **Providers of signals and reports** are focused on the investment industry alone. They can produce bespoke analysis for fundamental clients or sell quantitative signals. These include specialized firms, boutique research shops and quantitative teams and major sell-side firms.

As the market for alternative data expands, we see the emergence of three kinds of **data intermediaries**: **Consultants** advise buy-side clients on the entire process of onboarding Big Data, logistical issues around IT/Quant, legal aspects, information on data-sets/data-providers and insights into competition (e.g. firms such as Integrity Research). **Data Aggregators** specialize in collecting data from different alternative data providers. Investors can access hundreds of datasets often through a single portal by negotiating with the data aggregators (e.g. firms such as Eagle Alpha). Many IT firms offer **technology solutions** to Big Data clients. These solutions include public, private or hybrid cloud architecture enabling clients to onboard data quickly. Examples of such firms are IBM (with Big Insights product) and SAP (with HANA product). **Sell-side research** teams educate clients on Big Data and Machine Learning, consult on designing quantitative strategies based on big and alternative data and provide (aggregated and derived) internal and external market data.

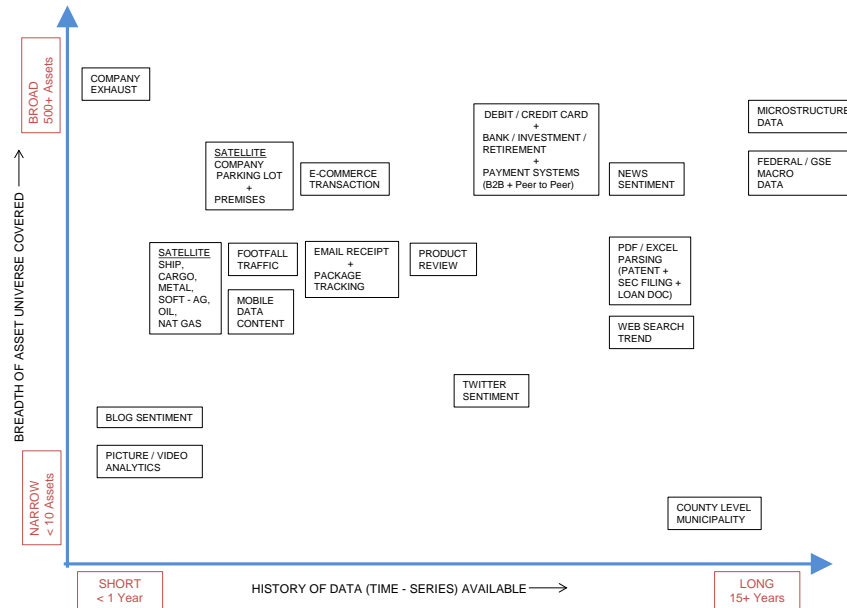
Data History

Once the dataset is identified and acquired, one needs to proceed with a backtest. In the previous section, we classified alternative data sets based on how they were generated: by online activity of individuals, business processes and sensors (Figure below left). Different types of alternative data are often available with limited history.¹⁸ Detecting and correcting sampling biases and seasonality is non-trivial as data sets typically have short history (Figure below).¹⁹

¹⁸ History depends on individual data provider. We interpret the figure as an approximate history available from a typical data provider in the specific category.

¹⁹ Having a short history makes seasonality detection harder. This affected bearish [calls](#) on Chipotle in late 2015, when analysts relied on lower foot traffic believing them to be a consequence of food-borne illnesses, instead of the cold season inducing customers to order meals at home.

Figure 11: Typical length of history for alternative data sets



Source: J.P.Morgan Macro QDS

Data based on the records of individual experiences (social media, product reviews, web searches, etc.) is often obtained through web scraping. For an overview of web scraping methods see the [Appendix](#).²⁰ Google Web Search trends are available since 2004 and can be used to construct signals with reasonably long time history. Reliable datasets based on social media activity (e.g. blogs, tweets, videos) are available usually with less than ~5 years of history. In this chapter we provide two examples of strategies based on data generated by individuals' activity:

- [Twitter sentiment data to trade the broad equity market](#) (iSentium)
- [News sentiment data to trade bonds, equities and currencies](#) (RavenPack)

The second category is data generated by business processes. The most reliable datasets in this category are based on credit card transactions and company exhaust data.²¹ Credit card data are usually available with approximately 10 years history. A large amount of historical data is made available by federal and state-level agencies and the data is usually available with history longer than 10 years. Market microstructure data (such as L-2 and L-3 order-book tick data) is also available with over 15 years of history. In the future we expect increased availability of trade-flow data from sell-side institutions. Sell-side flow data is typically available with less than 5 years of consistent history. In this chapter we provide an example of a strategy based on data generated by business processes:

- [Consumer transaction data to trade individual stocks](#) (Eagle Alpha)

The third category of data is collected by sensors embedded inside phones, drones, satellites, etc. A typical history of ~3-4 years is available for satellite imagery, mobile app data and cellular location tracking. As an illustration of the use of data from sensors we provide two example strategies:

- [Geolocation data to estimate retail activity and trade individual stocks](#) (Advan Research)

²⁰ Legal considerations include unauthorized download and use of information from website or placement of unwanted load on server through automated polling for information.

²¹ Company exhaust data refers to data collected by individual companies along their internal supply chain and sold to external firms after adequate aggregation and anonymization. Some of these data sets require legal and reputational risk assessment.

- [Satellite image data to estimate employee activity and trade individual stocks](#) (RS Metrics)

Backtests provided in this section should be used as illustrations of Big Data trading strategies, rather than endorsement of certain data types or providers. Performance of tested Big Data signals should also be taken as indicative as signals may weaken in time due to changes in transaction costs/capacity, etc. Finally, we stress that use of alternative data will not always lead to profitable strategies as some datasets and methodologies will simply add no value. Indeed, the media has highlighted some [prominent failures](#) of calls made using alternative data.

Data from Individual Activity

This type of alternative data is a result of the online activity of individuals. We further classify this data into the three sub-categories: **social media data** (e.g. Twitter, LinkedIn, blogs), data from **specialized sites** (e.g. news media, product reviews) and **web searches and volunteered personal data** (e.g. Google search, email receipts). Below, we highlight several prominent data providers in these categories, and a more comprehensive list is provided in the handbook of Big Data providers (Chapter 4).

Social media sentiment analysis is a fairly popular type of alternative data. Twitter is the most common source, followed by various news vendors and blog sites. Sentiment data is typically cheaper to acquire as compared to other alternative datasets (e.g. credit card data). We can conceptually understand sentiment analysis of written text by breaking it down in a four step process:

- **Named entity extraction:** The algorithm has to first extract the identity of the speaker. Other examples of entities include places/addresses, organizations, products/brands.
- **Theme and category extraction:** The algorithm establishes the topic being discussed. “Theme” is an industry term for important phrases that are decided by the Machine Learning algorithm; examples would be “Fed hike”, “Middle-Eastern crisis”, “iPhone7”, etc. “Category” refers to pre-defined buckets into which the sentiment score could be aggregated; examples would be a stock ticker. The algorithm may also isolate new trending topics from news articles.
- **Intention and Sentiment:** A sentiment score is assigned to the article using standard Natural Language Processing (NLP) techniques. Often a simple “bag-of-words” approach or a rule-based assignment might be used to assess sentiment.
- **Relevance and Influence:** A quantitative metric of relevance is assigned to the article based on its relation to the traded assets. A quantitative metric of influence is assigned to the article/tweet, which measures the projection of an opinion over social network connections. It could be based on popularity/following of author, links to other prominent authors, etc.

There are many challenges in integrating different social media feeds. Each data source may give a different interface and a different format for delivering data. Polling websites for updates is complicated by varying latency. Many of the data sources report the same activity multiple times; so there is a need to carefully eliminate duplicates. Further, if implementing the natural language processing (NLP) process by oneself, the analyst must note that the language styles differ across sources. For example, Twitter language differs in style from SEC documentation.

These technical and operational challenges provide an opportunity for intermediary firms to bridge the gap between investment managers and social media. Typically such firms have a partnership with the data source and they offer a single API through which the client can access multiple sources of data. For example, [Gnip](#) is Twitter’s enterprise API platform that offers a single interface to access Facebook, YouTube, Google+ and StackOverflow. In this case, *Gnip* claims to analyze 25 billion activities per day and to serve customers in >40 countries spanning 95% of the Fortune 500 list.

Sentiment analysis is applicable not only to individual stock names; it can also be used to trade broad market indices. For examples, [iSentium](#) provides a daily directional indicator which consolidates sentiment across many social media sources to produce a buy/sell signal on the S&P 500. We include a detailed analysis of *iSentium* data later in this section²².

Sentiment analysis companies largely focus on the equity market. For commodities, sentiment data is available from [Descartes Labs](#); the firm also provides satellite imagery tracking acreage and production of soft-agricultural assets. Some firms offer sentiment analysis across asset classes. Examples include [RavenPack](#), [Sentiment Trader](#), [InfoTrie](#) and [Knowsis](#). To track sentiment outside of single-name equities, firms use news media as an additional source to augment posts on social media. With the growth of passive, macro and thematic trading, it is important to track sentiment on prominent ETFs (such data is available from *Sentiment Trader*, [Social Alpha](#), *Knowsis*).

²² J.P.Morgan offers an investable index based on iSentium data

Some sentiment providers focus exclusively on fundamental clients; others seek to cater to both quant and fundamental funds. As an example, data from [DataMinr](#) is well suited for fundamental clients. The firm uses its access to the full fire-hose of Twitter feeds to distill upwards of 500 million tweets per day to a small limited number of alerts provided to its clients. Other firms like *Social Alpha* provide support for quantitative back-testing along with standardized alerts for fundamental clients. Fund managers should choose data providers depending on their comfort level with handling raw data. Advanced quant firms could use the NLP engine from [Lexalytics](#) to directly analyze text on their own. Less quantitative (e.g. fundamental) managers might prefer using support from [DataSift](#) (that e.g. uses *Lexalytics*'s NLP engine amongst other tools) to analyze data from media sources like LinkedIn.

Within social media, there is a sub-category of firms targeting just investment professionals. Some like [StockTwits](#) and [Scutify](#) offer a Twitter-like feed focused on trading opportunities and market developments. Others like [SumZero](#) and [TrustedInsight](#) aim to connect buy-side professionals (including research notes, job posts, etc.) and do not focus narrowly on reporting opinions. Apart from social media for professionals, there is also a sub-category of firms analyzing the sentiment of blog posts. Examples include [Accern](#), [Alphamatician](#) and [Datasift](#). In light of the inability of current Machine Learning algorithms to pass Winograd's test, we are less confident of sentiment analysis of long and elaborated blogs.

Some sentiment providers focus mainly on news media. The [GDELT](#) project includes >250 million data points tracking news articles published in >100 languages from 1979 to present. The entire dataset is publicly available over the Google Cloud Platform and consequently finds use in academic research. Some firms like [RelateTheNews](#) provide not only sentiment data, they also act as a "Platform as a Service" provider whereby clients can onboard their data and use the firm's proprietary sentiment analysis engine to extract signals (a similar solution in the satellite space is provided by *Descartes Labs*). While majority of the firms focus on US companies, firms like *Alphamatician* and [Inferess](#) track companies in Asia and Europe as well. Most sentiment providers focus on English language text. Multiple language support is increasingly available, as in offerings from *Lexalytics* and [Reputate](#). *Reputate* covers social media composed in 15 languages including French, German, Russian and Chinese. Sentiment analysis directed for VC deals and PE investments is available from [Heckyl](#). Data from Wordpress and Wikipedia is analyzed by *Datasift*, amongst other firms.

There is a separate set of firms that track **specialized websites**. Companies like [App Annie](#) and [Apptopia](#) track major mobile apps. They provide consolidated data on downloads, usage (times per day), and revenues. Such data is also commonly segmented across demographics and regions. [Yipit](#) Data focusses on companies' web pages to locate information on both public and private firms. A broader segment of websites is covered by a firm called [DataProvider.com](#) that tracks >280 million websites across 40 countries to provide data on job trends and other business intelligence metrics. [7Park](#) provides both business intelligence metrics from tracking apps as well as from tracking websites; its information can be used to evaluate both public and private companies.

Web search trend data is available from specialized internet traffic tracking firms like [Alexa](#) and also from other bespoke providers. Search trend data is directly available from [Google Trends](#) website, which also allows for contextual differences in search (i.e. the engine treats Apple the company as different from apple the fruit). For evaluating the search trend on a company, the analyst must form a list of nouns/phrases relevant to the company (anecdotally, 20-25 seems to suffice), get the data from Google Trends, and treat for seasonality correction after removing outliers. Analysis of other websites is suited to specific kinds of investors: risk-arb teams may scrape data from the SEC and FTC; consumer-discretionary sector analysts may look at data from *Yelp*, and technology sector analysts can look at postings on [Glassdoor](#) and [LinkedIn](#).

[Return Path](#) specializes in volunteered **personal email data** covering approximately 70% of the worldwide total email accounts. By collating this with purchase email receipt data for ~5000 retailers, it offers analytics around purchase behavior and consumer preferences. Similar transaction data is also available from [Slice Intelligence](#) and [Superfly Insights](#). Note that receipts are emailed not only for e-commerce transactions, but also increasingly for many transactions at brick-and-mortar stores.

Case Study: Using Twitter Sentiment to trade S&P 500 ([iSentium](#))

[iSentium](#) provides real-time sentiment time series based on Twitter messages. It is effectively a sentiment search engine which can provide investors with a way to judge the potential market impact of a tweet, a news article, or other social media activities. Backtests of iSentium Daily Directional Indicators (DDI) indicate that social media sentiment can be used as a predictor for short term market moves.

J.P. Morgan has constructed the JPUSISEN Index that takes intraday long or short positions in the S&P 500 Index, based on a daily directional indicator provided by iSentium. The steps used in the construction of the indicator are described below.

Construction of iSentium Daily Directional Indicator

1. The universe is limited to the 100 stocks which are most representative of the S&P 500, filtered using tweet volume and realized volatility measures.
2. Tweets are assigned a sentiment score using a patented NLP algorithm.
3. By aggregating tweet scores, a sentiment level is produced per minute between 8:30 AM and 4:30 PM every day. Sentiment for the day is aggregated using an exponentially weighted moving average over the past ten days.
4. S&P 500 returns are forecasted using a linear regression over the sentiment scores for the past two days, with betas evolved via a Kalman filter.

Historical backtests indicated that the JPUSISEN index would have returned 13.7% annually since January 2013 with an Information Ratio (IR) of 1.4 (after 5bps of transaction cost). This compares with S&P 500 return of 12.1% and an IR of 0.95 over the same period. The sentiment based index also had a smaller maximum drawdown than the S&P 500.

The table below shows backtested performance on JPM iSentium index separating performance of long signals, and short signals. This is also compared to the relevant S&P 500 performance. One can see that both long and short sentiment components contributed to the outperformance of the Long-Short index.

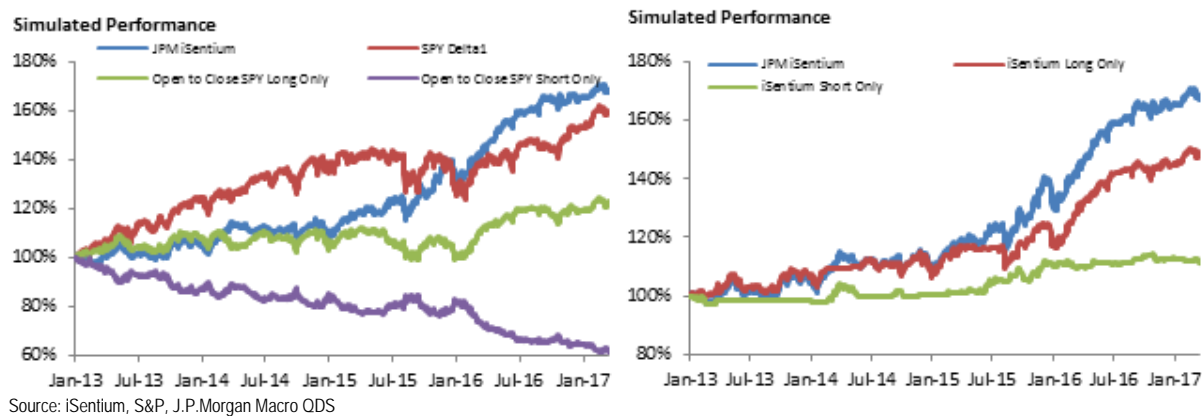
Figure 12: Comparison of iSentium risk and return to the S&P 500

Strategy/Index	Return Ann. (%)	Volatility (%)	IR	Max DD (%)
iSentium L/S (JPM iSentium Index)	13.74	9.79	1.40	-8.10
iSentium – Act only on Long signal	10.33	8.74	1.18	-7.29
iSentium – Act only on Short signal	2.83	4.46	0.63	-4.66
S&P 500 Index	12.08	12.76	0.95	-12.08
S&P 500 – Long from open to close only	5.25	9.83	0.95	-5.25
S&P 500 – Short from open to close only	-10.97	9.83	-1.12	-10.97

Source: iSentium, S&P, J.P.Morgan Macro QDS

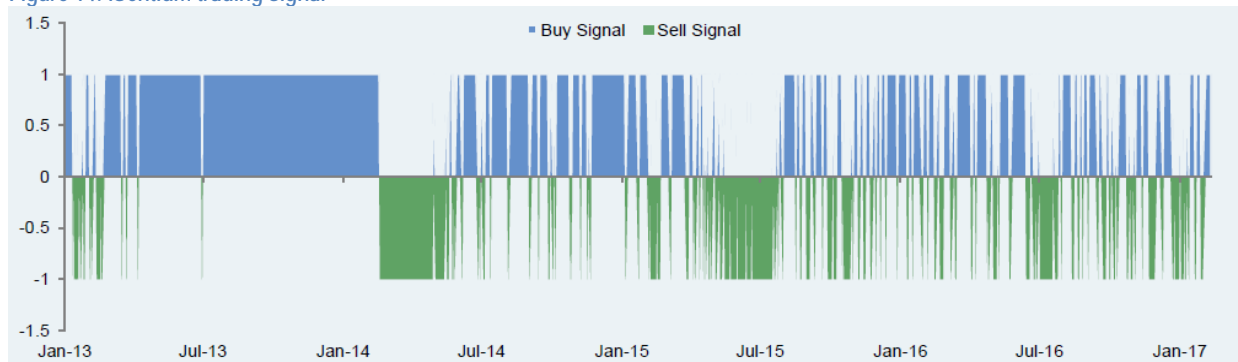
Performance of iSentium index components and S&P 500 index is shown in the figure below.

Figure 13: Simulated performance for iSentium and S&P 500 indices



The iSentium Signal is shown in the figure below. Since 2014, the signal has been changing frequently and had a success rate of 56%, with an average gain of 45bps. This is typical for a quantitative strategy to place a large number of low conviction bets.

Figure 14: iSentium trading signal



Overall the iSentium signal exhibited moderately positive correlation to the S&P 500 (46%) due to its direct exposure to the S&P 500 and overall positive sentiment/performance environment over the past 4 years. However, we do look at the iSentium strategy as an absolute long/short strategy and returns derived from the sentiment signal as a source of alternative risk premia. Comparing correlation properties of the sentiment signal to traditional equity risk premia (Momentum, Value, Quality, Low Volatility, etc.) uncovers very low correlation (less than 5%) as shown below. This confirms that including signals based on alternative datasets (in this case social media sentiment), can be an important building block of a broad risk premia portfolio.

Figure 15: Correlation of iSentium signal with traditional equity risk premia

		1	2	3	4	5	6	7
JPM iSentium	1	1.00						
Global Multi-Factor	2	0.05	1.00					
Value	3	-0.01	-0.05	1.00				
Low Size	4	-0.01	-0.07	0.37	1.00			
Momentum	5	0.03	0.42	-0.64	-0.42	1.00		
Quality	6	0.03	0.38	0.02	-0.33	0.11	1.00	
Low Volatility	7	0.03	0.75	-0.05	-0.23	0.22	0.50	1.00

Source: iSentium, J.P.Morgan QDS

Case Study: Using News Sentiment to trade Bonds, Currencies and Commodities ([Ravenpack](#))

[RavenPack](#) analyzes unstructured datasets to produce structured and granular indicators of relevance to investment professionals. Unstructured data include premium newswires^[1], regulatory news providers, press releases and over 19,000 web publications. In this section we illustrate the usage of RavenPack news sentiment data in trading currencies, sovereign bonds and commodities.

As a first step we calculated daily sentiment scores for different assets of interest from RavenPack's raw newsfeed. The exact process of translating the granular data (>150GB of CSV files) into a daily sentiment score is given in the grey box below.

Translating RavenPack News Feed into Daily Sentiment Score

RavenPack provides 50 data fields for each event. We analyzed data since 2005 for each asset of interest.

1. Step One: isolate all unique events on a given day specific to a certain "ENTITY_NAME". Entity name was set equal to the currency, commodity or country name. We set a cutoff time of 4 PM EST to reflect NY market close.
2. Step Two: RavenPack provides a field called "RELEVANCE", which is an integer score between 0 and 100. A higher value indicates that the mention of the entity is more integral to the underlying news story. We used RELEVANCE as a cut-off filter, ignoring stories with a value < 75.
3. Step Three: RavenPack provides a field called "EVENT_SENTIMENT_SCORE" or ESS. It is a granular score between -1.00 and +1.00 that represents the news sentiment for a given entity. The average of ESS for all filtered events was designated as the sentiment value for the day. We forward-filled the sentiment for days on which no news was received by the analytics engine.

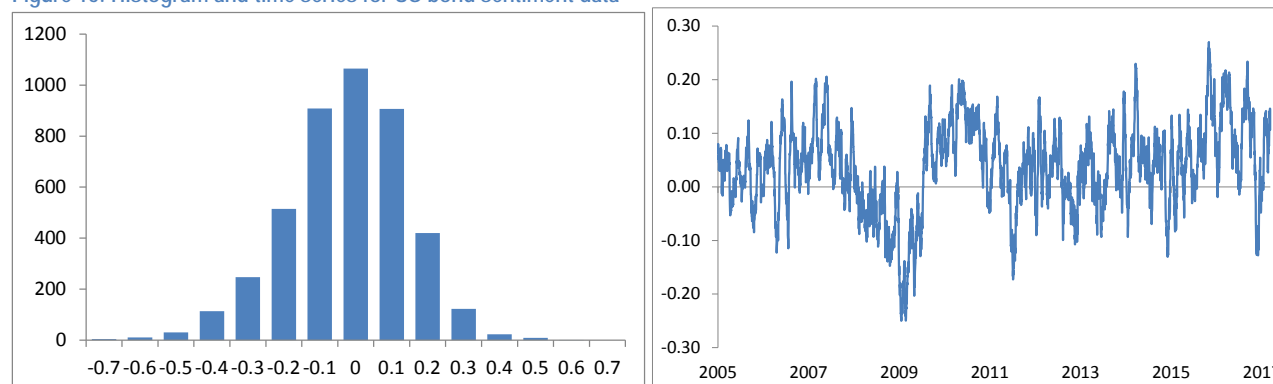
To illustrate potential use of sentiment data, we constructed simple long-short strategies: going long assets with the top 3 sentiment scores, and shorting assets with the bottom 3 sentiment scores.

Sovereign Bonds

A daily long-short strategy is constructed from currency-hedged returns of sovereign bonds from Australia, Canada, Denmark, Germany, United Kingdom, Japan, Sweden, New Zealand and the United States. RavenPack does not issue a 'bond sentiment' indicator; so we used economic sentiment as a contrarian indicator; we go long the 3 bonds with the most negative economic sentiment and short the 3 with the most positive sentiment.

^[1] News wires include Dow Jones, Benzinga, MT Newswires, Alliance News, FX Street and The Fly.

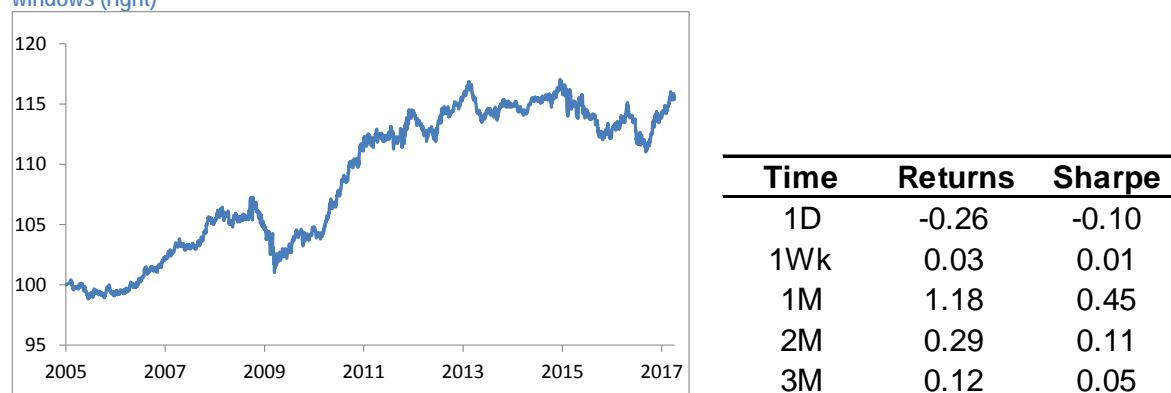
Figure 16: Histogram and time series for US bond sentiment data



Source: RavenPack, J.P.Morgan Macro QDS.

Using average sentiment value as signal, we obtained positive annualized returns and **Sharpe Ratios in a 0 to 0.5 range**. For instance the top 3 / bottom 3 portfolios where we take 1-month average of sentiment as a signal, yields a Sharpe ratio of 0.45. Results are somewhat sensitive to the lookback window for signal averaging as shown in the figure below.

Figure 17: Daily performance of bond sentiment, Top 3/ Bottom 3 long-short strategy (left), and performance of signals with different lookback windows (right)



Source: J.P.Morgan Macro QDS, RavenPack

Correlation of this bond sentiment strategy with traditional bond risk factors: Value, Momentum, Carry, and Volatility, was nearly zero (less than 4%) over the past 10 years. Despite relatively low Sharpe ratios, the sentiment signal may be useful in a portfolio context (i.e. when combined with other signals).

Figure 18: Correlation of RavenPack sentiment signal with traditional Bond Risk Premia

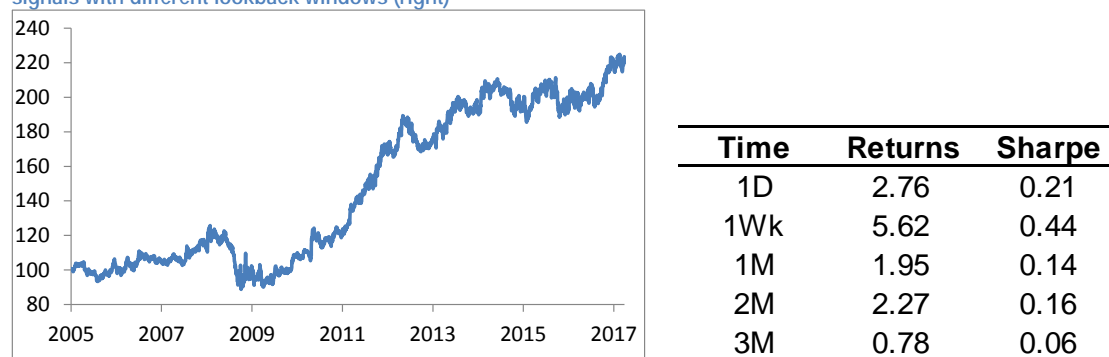
Risk Premia	1	2	3	4	5
Volatility - Bond	1				
Value - Bond	2	-0.04			
MoM - Bond	3	0.00	0.63		
Carry - Bond	4	-0.04	0.49	0.44	
Sentiment - Bond	5	-0.03	0.04	0.03	0.10

Source: J.P.Morgan Macro QDS, RavenPack.

Equity Indices

A daily long-short strategy is constructed from benchmark equity indices of Australia, Canada, Denmark, Germany, United Kingdom, Japan, Sweden, New Zealand and United States. We are using RavenPack's 'economic sentiment' signal as a confirming indicator to go long/short equity indices. We obtain the following results:

Figure 19: Daily performance of economic sentiment as a confirming indicator, Top 3/ Bottom 3 long-short strategy (left), and performance of signals with different lookback windows (right)



Source: J.P.Morgan Macro QDS, RavenPack

Figure 20: Correlation of RavenPack sentiment signal with traditional Equity Risk Premia

Risk Premia		1	2	3	4	5
Volatility - Equity	1					
Value - Equity	2	0.16				
MoM - Equity	3	0.02	0.14			
Carry - Equity	4	0.16	0.03	-0.16		
Sentiment - Equity	5	0.04	0.09	0.08	-0.04	

Source: J.P.Morgan Macro QDS, RavenPack.

Developed-Market Currencies

A daily long-short strategy was designed for the following currencies: Australian Dollar, Canadian Dollar, Swiss Franc, Danish Krone, Euro, British Pound, Japanese Yen, Norwegian Krone, New Zealand Dollar and Swedish Krona (carry adjusted returns vs. USD). We used a currency-specific sentiment indicator derived from RavenPack's raw newsfeed (all data points with ENTITY_TYPE marked as FX, were used). The strategy held long the 3 currencies with the most negative sentiment and held short the 3 currencies with the most positive sentiment.

Interestingly using average sentiment value as a contrarian signal and it resulted in moderately positive results (Sharpe ratios in a 0 to 0.4 range).

Figure 21: Daily performance of DM FX sentiment as a contrarian indicator, Top 3/ Bottom 3 long-short strategy (left), and performance of signals with different lookback windows (right)



Time	Returns	Sharpe
1D	0.48	0.07
1Wk	3.24	0.43
1M	2.53	0.32
2M	1.32	0.16
3M	-0.22	-0.03

Source: J.P.Morgan Macro QDS, RavenPack.

Correlation of this signal to other FX risk premia (Value, Momentum, Carry and Volatility) was very low. Similar to bond sentiment strategies, FX contrarian sentiment signal is relatively weak, and likely viable only in a portfolio context.

Figure 22: Correlation of RavenPack sentiment signal with traditional FX Risk Premia

Risk Premia		1	2	3	4	5
Volatility - FX	1					
Value - FX	2	-0.02				
MoM - FX	3	-0.06	-0.06			
Carry - FX	4	0.22	0.08	0.01		
Sentiment - FX	5	-0.03	-0.03	-0.02	-0.02	

Source: J.P.Morgan Macro QDS, RavenPack.

Data from Business Processes

Data generated by Business Processes includes data made available by **public agencies** (e.g. federal and state governments), **commercial transactions** (including e-commerce and credit card spending, exchange transaction data), and data from **other private agencies** (e.g. industry specific supply chain data).

The first sub-category within business process generated data is data collected by either the **government or government-affiliated agencies**. Massive data sets are now available at the international level ([IMF](#), [World Bank](#), [WTO](#)), federal level ([US](#) including Federal Reserve and Treasury, [China](#)) and even the city level (e.g. [San Francisco](#)). Such data tends to be lower frequency (say, monthly), and is useful for long-term investors and low frequency strategies.

A number of firms have emerged seeking to track **commercial transactions**. Firms like [Nielsen](#) track point of sale systems at retail chains; [BuildFax](#) tracks building permit data and collates over 23 billion data points on residential and commercial real estate structures. [Eagle Alpha](#) also provides bill of lading data that aggregates cargo receipts from ships across different ports. [Price Stats](#) produces a daily inflation series from tracking price information for 1000 retailers across 70 countries. Firms like [Cignifi](#) use mobile phone data to create a credit score on individuals, and are emerging as important players in the alternative credit industry. Companies like [Dun and Bradstreet](#) monitor inter-company payments (amount, timing, delays and failures) by analyzing company invoices. Historically, this was used by smaller firms seeking to evaluate the credit risk of a new customer; now, the same data when available across many companies can be used to create long-short portfolios – the idea being that a sudden increase in overdue amounts is an early indicator of credit stress at the firm. Firms like [Slice](#) also [track](#) order/shipment from online retailers (like Amazon); they also gather information from email accounts and social media information offered by users. This activity can aid investors to track sales at private companies like Uber.

A set of firms have released apps to help consumers manage their finances – such apps track spending patterns and advise their clients. These apps typically gain access to bank/investment/retirement accounts, loan/insurance details, bills/rewards data and even payment transactions. Such data is then anonymized/aggregated and then sold to third parties. Firms in this category prefer to remain out of the media spotlight. In 2015, [Investment Yodlee](#) (partnering with 12 of the 20 largest US banks and tracking ~6 million users) published a [rebuttal](#) seeking to refute a Wall Street Journal [article](#) (titled “Firm tracks cards, sells data”) that claimed that “the company sells some of the data it gathers from credit- and debit-card transactions to investors and research firms, which mine the information for clues about trends that can move stock prices” ([alternative link](#)).

[Second Measure](#) and Earnest report aggregated/anonymized data on consumer transactions (e.g. see WSJ [article](#); the same article also discusses a potential increase in volatility of retail stocks due to dependence on credit-card data). Second Measure also [claims](#) that its Big Data analysis allows it to show how many subscribers Netflix has or how Uber is doing relative to Lyft. Consumer transaction data is also available from data aggregators including [Eagle Alpha](#), [Quandl](#) and [1010 Data](#). In recent months we have seen large banks become aware of the value of their “data assets” and hence one can expect them to try to sell this data directly to hedge funds bypassing third-party aggregators.

The segment of firms tracking customers’ own finances is significant. These firms provide a variety of services to the end-user who is asked to provide his/her bank details to avail of the services. Examples include [Intuit](#) (for tax preparation), [Digit](#) (for analyzing personal spending behavior), [Xero](#) (for small business accounting), and [Square/PayPal](#) (for payment processing). Financial data also passes through account aggregators such as [FiServ](#) and [Quovo](#).

While evaluating a source of transaction data, investors will have to correct for sampling bias. Different banks target different market segments; many banks are concentrated on large states and mid-income demographics. Data typically has seasonality; correcting it necessitates a longer time history. In the Consumer Discretionary sector, analysts look at same-store sales; analogously in credit card transaction analysis, we should look at same-customers spend, i.e. one has to account for attrition and gain of new customers. Mapping tags recorded in point-of-sale systems to company tickers is non-trivial as tags change without notice. As a safety measure, we recommend outlier detection and Winsorization of data before its aggregation. Note that small business spend is a separate category by itself. Here credit and debit card data will not be sufficient. We have to include flows through Demand Deposit Accounts (DDA) – these include tracking checks, wire

transfer and ACH (automatic clearing house) transactions. This has a challenge in that the analyst has to remove loan/tax payments and even intra-account transfers.

Once this data is available in a clean format, it can be used in conjunction with analyst consensus and intra-quarter sales guidance by companies (~1 month into new quarter, when results for the previous quarter are announced) to forecast US retail sales. This appears to be the main use-case of such data currently. We suggest looking at monthly change in spend and expect that such data might detect inflection points in BLS non-farm payrolls and PCE, apart from retail sales. New indices on small business expenditures can emerge which can offer better insight into the economy than the Small Business Optimism Index released by [NFIB](#) (National Federation of Small Business).

The third category of business process data arises from **private agencies** that are involved in collecting data specific to industry sectors. We list in our directory over 80 companies that seek to provide such data through the Bloomberg terminal alone. Companies tracking the consumer discretionary and staples sector include [AROQ](#) (for automotive, beverage, food and apparel), [Edmunds](#) (cars), [SNL Financial](#) (cable/broadcast), [Smith Travel](#) (lodging), [Experian Footfall](#) (retail traffic) and [Redbook Research](#) (single-store-sales trends). [M Science](#) (formerly, ITG/Majestic Research) licenses databases on DMV registrations, internet traffic and prescription drug sales. *Eagle Alpha* offers a China Auto Insight (CAI) dataset based on data collected from a panel of dealerships across China. *Quandl* [tracks](#) auto-insurance (as a proxy for automobile sales) and construction permits across county municipal offices (as a proxy for construction activity). Auto industry data is also valuable for general equity investors as it serves as a leading indicator for broader economic recovery. Within the Energy sector, data is available for bespoke regions, e.g. [CDU-TEK](#) (for Russia).

Other private agency data includes intra-day tick-level market microstructure data. Such data is available from exchanges as well as from aggregators like [Tick Data](#). With the development of NLP techniques, text in pdf and Excel format is increasingly read by machines (especially, if they are in machine-readable XBRL format). [Edgar Online](#) provides a repository of documents on financial statements, annual/quarterly filings, disclosures and initial/secondary public offering details.

Case Study: Using Email Receipt Data to trade US Equities ([Eagle-Alpha](#))

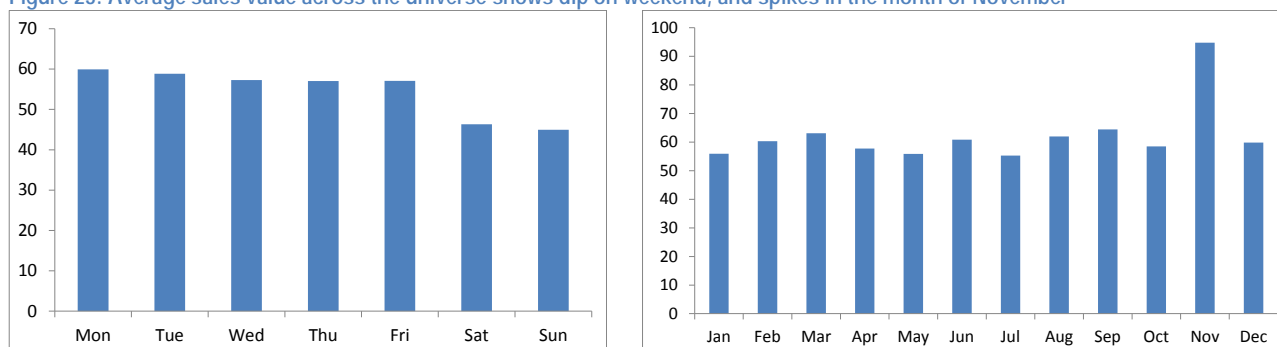
Aggregated consumer transaction data is a prominent category within business-process data. In the example below, we analyze a trading strategy based on email receipt data provided by Eagle Alpha. [Eagle Alpha](#) provides a variety of services across the spectrum of alternative data, including data sourcing, bespoke projects, teach-ins and analytical services. The firm provides ~500 data points spanning 20 categories, including online search trends, expert views and trade data.

The data was collected by an Eagle Alpha partner firm, which tracks transaction data representing 80% of all online purchases. The firm tracks such data for >5000 retailers and provides item and SKU-level transaction data filtered into 53 product categories. The email receipt data time series we analyzed below starts in 2013. This data included a constant group of users (that are still active and were active prior to Dec 1, 2013). The data included aggregated figures for total dollar spend, total number of orders, and the total number of unique buyers.

We have analyzed a dataset of email receipts for 97 companies. 36 of these were private companies, and 61 public, 31 of which were S&P 500 constituents. Taking liquidity into consideration, we decided to test trading signals for the S&P 500 companies only.

Given the disparity of actual spend for different companies, it is necessary to normalize dollar spend into relative score (e.g. percentage change, or normalize data for an average spend). There is also meaningful seasonality in the data. For instance spend data is slightly lower on weekends and is higher during month of November.

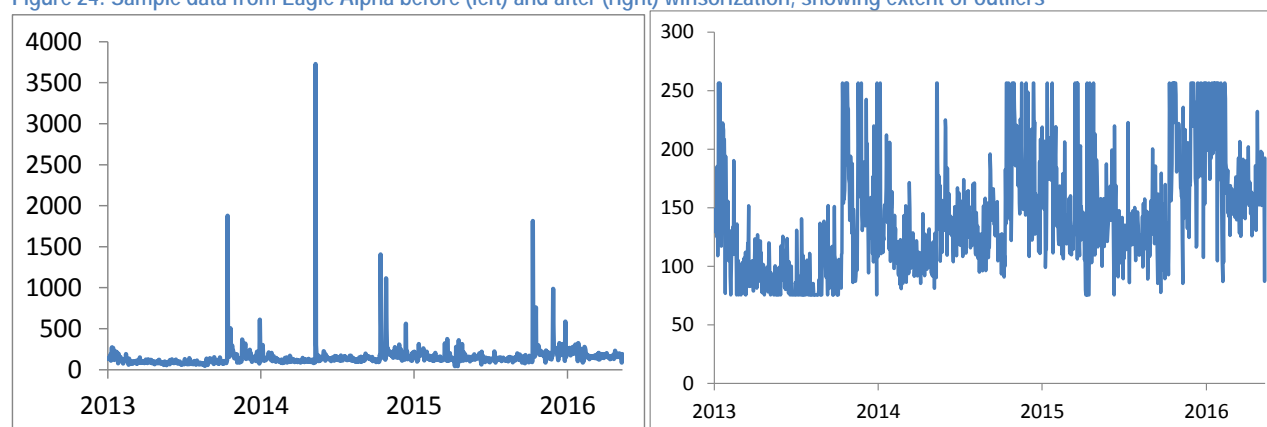
Figure 23: Average sales value across the universe shows dip on weekend, and spikes in the month of November



Source: J.P.Morgan Macro QDS, Eagle Alpha

The actual daily sales time series were volatile and we noticed data outliers. For this reason, we winsorized the time series. The Figure below shows a time series for one particularly noisy company without and with winsorization.

Figure 24: Sample data from Eagle Alpha before (left) and after (right) winsorization, showing extent of outliers



Source: J.P.Morgan Macro QDS, Eagle Alpha

We analyzed three time series: the dollar spend, number of orders and number of buyers. While number of orders and number of buyers are highly correlated (~99%), dollar spend is not highly correlated with number of buyers/orders (~25%).

We aggregated the daily spend/order/buyer data into a weekly score and calculated week-over-week percentage change for each. After winsorizing to 5th-95th percentile, we tested both the level and z-score as signals. Based on a cross-sectional comparison, we went long the top 5 stocks and short the bottom 5 stocks. The portfolio was rebalanced weekly.

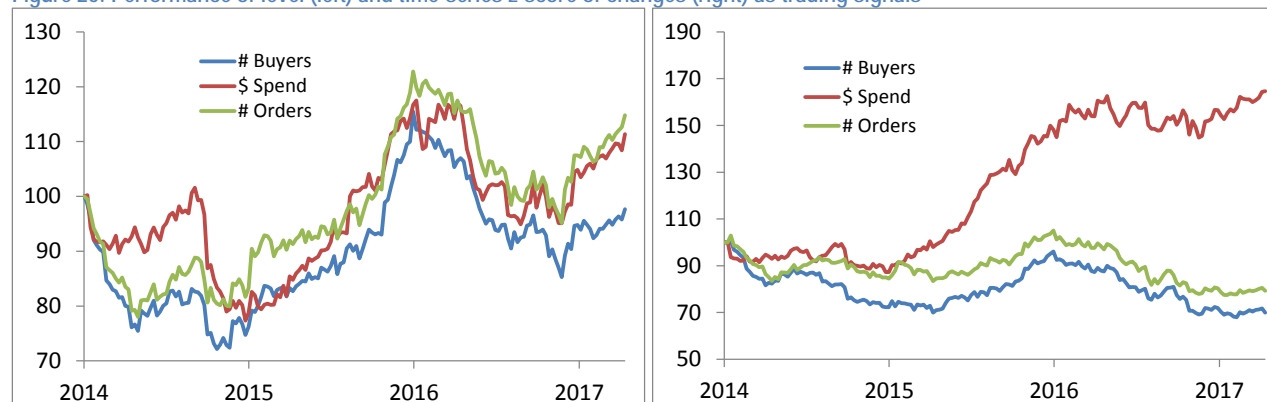
Figure 25: Sharpe ratios of various strategies using Dollar Spend, Buyer Count and Order Count

Dollar Spend	Top 6/ Bottom 6	Buyer Count	Top 6/ Bottom 6	Order Count	Top 6/ Bottom 6
Data		Data		Data	
Level	0.29	Level	0.02	Level	0.36
Z-score 4 week	1.13	Z-score 4 week	-0.71	Z-score 4 week	-0.49
Z-score 5 week	0.72	Z-score 5 week	-0.49	Z-score 5 week	-0.14
Z-score 6 week	0.67	Z-score 6 week	0.04	Z-score 6 week	0.11

Source: J.P.Morgan Macro QDS, Eagle Alpha

We also plot cumulative returns using the level (i.e. percentage of aggregated figure) and the 4-week z-score for all 3 data sets.

Figure 26: Performance of level (left) and time-series z-score of changes (right) as trading signals



Source: J.P.Morgan Macro QDS, Eagle Alpha

Data from Sensors

We categorized data generated by sensors into three groups: **satellite data**, **geolocation data**, and data generated by **other sensors**.

One of the most popular alternative data offering relates to **satellite imagery**. Twenty years back, launching a traditional geo-spatial imaging satellite cost of tens of millions of dollars and years of preparation. Nowadays, companies (such as [Planet Labs](#)) are able to launch a fleet of nano-satellites (many the size of a shoe-box) as secondary payloads into low-earth orbits. These nano-sats have significantly brought down the cost of satellite imagery. Image recognition is also standardized with pre-trained Deep Learning architectures (around convolutional neural networks or CNNs) such as LeNet (461k parameters), AlexNet (61M parameters) and VGG 16 (138M parameters) being used. Imagery is currently available with daily frequency with providers anticipating real-time capture within the next three years.

An impressive example of imagery analysis is Oil reserve analysis. Oil reserves are often stored in floating tanks around the world, whose lid rises and falls with increasing or decreasing storage. With the rising and falling height, the crescent-shaped shadows cast by the sun also change shape. By tracking these shape changes for 20000 tanks, [Orbital Insights](#) predicts oil inventory levels.

[Descartes Labs](#) tracks data from hundreds of satellites; in some cases, the delay between acquisition of the image and its presence on Descartes's platform is only a few hours. The firm also provides a platform for conducting Big Data analysis, where it makes available 5 peta bytes (5000 TB) of information. The firm uses this information to forecast production for cotton, rice, soy and wheat across different regions including the US, Brazil/Argentina, and Russia/Ukraine. Within the satellite ecosystem are companies like *Planet* which offer to design and launch a satellite for a customer on demand. To supplement information from satellites, there are firms like [AggData](#) that compile lists of all company locations (latitude and longitude).

Satellite firms offer data on every asset class: *Orbital Insights* tracks car counts in parking lots for trading equities; [Genscape](#) tracks oil tankers in Cushing for oil; [RezaTec](#) provides data on wheat, corn, coffee and timber; [Skywatch](#) for oil and [RS Metrics](#) for copper, zinc and aluminum storage areas. Maritime data on ships is provided by [Windward](#), [Vessel Finder](#) and [MarineTraffic](#) (note that 90% of trade is transported by sea). Aerial surveillance through drones is conducted by companies like [DroneDeploy](#) that provides data for agriculture, mining and construction industries.

While analyzing satellite imagery or using data derived from it, investors must be aware of several challenges. For instance, one must understand treatment of cloud cover, and seasonalities around e.g. holidays. Every location may not be covered with equal frequency or equal accuracy, hence error and predictive power may vary. There are other nuances specific to type of data or location (e.g. Europe has more multi-story covered car parks).

Firms like RS Metrics triangulate satellite data - obtained from other sources like France's Airbus and DigitalGlobe (acquired by Canadian firm MDA in Feb 2017) - with other data sources (like demographic and open-source data) to offer guidance on storage at commodity mines, car count at parking lots, etc.

The second category of sensor-generated data comprises **geolocation data** (also referred to as volunteered geographic information). By tracking the location of smartphones through either GPS, WiFi or CDMA signals, these firms determine footfall traffic in and around store locations. Firms in this category include [AirSage](#), [Placed](#) and [Advan Research](#). Many companies in this segment are currently focusing on tracking footfall traffic in retail stores; in future, we expect real-time tracking of trucks for analysis of B2B companies as well.

A scalable way to collect footfall data is by tracking the location of mobile phones. Close to 70% of the US population has smart phones. The direct way to collect the cellular location is, of course, by obtaining from data providers like Verizon (cellular towers triangulation). However, precision of this method does not suffice for identification of stores in urban locations. Another source for footfall data would be companies placing mobile advertisements (tracking location). A third way to collect data would be to strike deals with different cellphone apps. Each app installs a geo-location code on the cell

phone with its user's explicit consent. Following the install, the app tracks location using multiple channels (like WiFi, Bluetooth and cellular signal) to improve accuracy. We study this use-case in detail by analyzing a dataset provided by *Advan Research*.

Other types of sensors are also used in collecting alternative datasets. *Genscape* uses non-contact ground sensors to monitor power lines; such data is used by power grid owners and operators. Sensors have been installed in many malls and retail stores for tracking customers. [*Nomi*](#) (*Brickstream*) uses 3D stereo video using cameras for counting visitors at stores. A competing technology based on thermal imaging is used by [*Irisys*](#), which has installed >300k sensors worldwide. Using thermal images reduces privacy concerns (since no picture of the person is taken); but the technology works best only on moving objects. Fixed sensors on the ceiling are available from [*RetailNext*](#) which provides in-store analytics through its Aurora product. Sensors within pressure-sensitive mats are available from [*ShopperTrak*](#). Firms like [*Percolata*](#) combine data from multiple sensors - including vision, audio and cell-phone tracking - to provide data on shopper count, dwell time, conversion rate and frequency of visits. Yet another variety of sensors is embedded in drones used for agriculture by firms like [*Agribotix*](#). Use of such drones, which has to be cleared by FAA as '333 exempt', is based on the observation that only healthy leaves reflect near infra-red light.

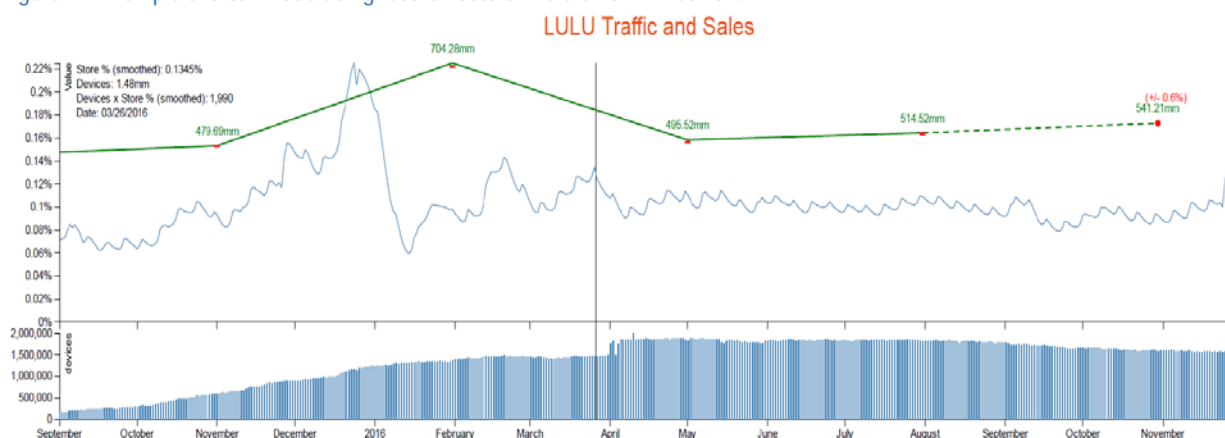
Case Study: Using Cellular Location to Estimate Retail Sales ([Advan](#))

Advan Research estimates customer traffic at physical store locations by tracking location of smart phones. Data is collected via apps that install geo-location codes on cell phones with the users' explicit consent. Following the installations, these apps track location using multiple channels (like WiFi, Bluetooth and cellular signal) to improve accuracy.

[Advan](#) geolocation data can be used to proxy the top-line revenue of publicly traded firms. Spanning 30% of the US population, the company sees approximately 25 million devices/per day, 60mm/per month, analyzes >3B data points per day and computes traffic for 1,033,050 locations of which 498,749 have been manually reviewed. The data can be mapped to revenues of 80 different S&P 500 stocks²³ (381 stocks in total) from retail, big box stores, supermarkets, hotels, hospitals, restaurants, movie theaters, amusement parks, dollar stores and fast food companies. In order to independently assess the predictive power of geolocation signals, we analyzed a sample dataset from Advan.

As a recent example, on Dec 2, 2016 Advan Research noted QTD traffic at Lululemon (LULU) locations were up 7%. This motivated them to predict LULU sales to beat the street consensus; a claim that was validated on Dec 7th, when the firm announced sales results. This sales surprise boosted the stock from \$58 to \$67 within a day.

Figure 27: Example of a call made using foot-fall data on Lululemon in Dec 2016



According to Advan's Location Data:

- Sales for Q3 2017 (08/01/2016 - 10/30/2016) are projected to be \$541.21mm.
- For the week ended 11/27/2016, LULU traffic increased 30% W/W and increased 6% Y/Y (vs 11/16/2015 - 11/22/2015).
- Fiscal MTD traffic increased 12% M/M (vs 10/03/2016 - 10/30/2016) and increased 7% Y/Y (vs 11/02/2015 - 11/29/2015).
- Fiscal QTD traffic increased 3% Q/Q (vs 08/01/2016 - 08/28/2016) and increased 7% Y/Y (vs 11/02/2015 - 11/29/2015).

Source: AdvanResearch, J.P.Morgan QDS

We have analyzed 6.5 GB of data provided by Advan. Raw data were aggregated by

- ticker level (traffic aggregated up to ticker level)
- store level (traffic count for each individual store location)
- device level (devices in each store and parking lot, with timestamp, inside/outside store hours, precise location)
- "device-vector" level (a vector of all locations traversed by each device during a day).

²³ Data as of April 2017.

The data set provided included a list of the top 1000 applications tracking the cellular location besides master files describing securities tracked and procedure used for generating the data. The procedure used for generating a signal is provided below.

Signal Generation From Raw Foot Fall Traffic Data of Advan Research

- To prevent double-counting of devices visiting the same store on the same day, we set app_id = device_app_id. App_id is a unique, anonymous identifier for the application collecting the data.
- We selected the sub-universe of the above applications which had the SDK (Software Development Kit) flag set to 1 in a file enlisting all the app categories. We further restricted the universe to applications, where we had consecutive data for at least 360 days.
- We computed the daily traffic of the combined applications using

$$\text{traffic} = \frac{\sum_{\text{across selected apps}} (\text{pct_devices_in_store} * \text{devices})}{\sum_{\text{across selected apps}} \text{devices}}$$

Here, “devices” refers to the total number of unique devices observed on a particular day by the devices_app_id (devices_app_id = app_id, hence it is total number of unique devices observed on a particular day by the app_id). “Pct_devices_in_store” refers to the ratio of unique devices inside the polygons marking the stores corresponding to this ticker divided by the “devices” field.

- We computed the prediction for the next quarter as

$$\text{Sales}_{i+1} = \text{Sales}_i * \frac{\sum_{\text{each day in Quarter } i+1} \text{traffic}}{\sum_{\text{each day in Quarter } i} \text{traffic}}$$

- Financial year varies for companies (it can be Jan to Dec, Feb to Jan or March to Feb) and Revenue numbers are not for sales from start of quarter to end of quarter so we have to keep note of such details while calculating predicted revenue.
- Reported sales revenue from Bloomberg are downloaded historically and mapped against predicted values from traffic. Historical analysts’ revenue estimates are obtained from FactSet for all quarters. Incorporate 2 day lag included in Advan data.
- If predicted sales value is greater than analysts’ estimate, we go long the stock (similarly if predicted value is smaller than analysts’ estimate we short the stock). We keep the stock in the basket till 5 days after the announcement date. Stocks are equal weighted.
- Our analysis has some limitations. We analyzed only the in-store foot fall data. We did not correct the data for seasonality. Users of Advan Research may benefit from independent analysis of the same data set using similar or different analysis techniques.

To understand if traffic by itself added value to the analyst estimates, we computed the percentage of times where a higher-than-consensus prediction by Advan led to an actual higher sales figure reported by the company. Our data covered 80 companies across five earning seasons; so prediction covers the latter four earning seasons²⁴.

Figure 28: Success rate in predicting sales using foot-fall data

Sector	Number of companies	Success rate for sales beats	Success rate for sales misses
C. Discretionary	43	58%	57%
Consumer Staples	6	25%	80%
Industrials	5	33%	30%
Health Care	3	63%	50%
Financials	11	91%	9%
Telecomms	2	25%	50%
Energy	3	40%	86%
Materials	1	100%	100%
Technology	2	25%	50%
Real Estate	4	50%	0%
Total	80	60%	52%

Source: JPM QDS Research, Advan Research, Bloomberg and Factset

From the table above, one can see that a higher-than-consensus call by Advan Research leads to future sales figure beating consensus 60% of the time. A lower-than-consensus call is right 52% of the time. As expected most of the companies covered are in Consumer Discretionary sector.

Quantitative Strategy that we tested is described below

We derived sales estimates for the upcoming quarter from Advan Research's raw traffic data using the methodology explained. We construct two baskets of stocks. In the "Outperform" basket, we place the stocks where we expect revenue to exceed analyst consensus. In the "Underperform" basket, we place the stocks where we expect revenue to be less than analyst consensus. Positions are taken on day data is received (typically, two days after the end of the financial quarter for the stock). Positions are held until five days after the announcement of earnings. On any given day, the portfolio is equal weighted across its constituents. The backtest is run for 80 stocks (only S&P 500 constituents) out of 381 stocks' data provided by Advan Research²⁵.

Results: The outperform basket (where traffic data indicated that sales will beat consensus) outperformed the S&P 500, and the underperform basket underperformed the S&P 500. **Correlation of the Long-Short strategy to the S&P 500 was 0.7%, i.e. virtually zero.**

²⁴ Sales for the retail sector also act as a proxy for the entire financial market, since consumer spending accounts for ~70% of US GDP. We did not explore that angle in this analysis.

²⁵ For each stock we start trading 2 days (lag from Advan) after the end of quarter and keep on trading till 5 days after the earnings announcements dates. So in this process there might be some days when there are no stocks in the basket which is usually after the end of earnings season and before the start of next earnings season. On such days we go long cash (JPCAUS3M Index) when there are no stocks in the basket.

Figure 29: Performance of long and short legs based on foot-fall data

Strategy	Mean	Volatility	Sharpe ratio
Outperform-Basket	24.31%	20.36%	1.19
Underperform-Basket	7.79%	17.08%	0.46

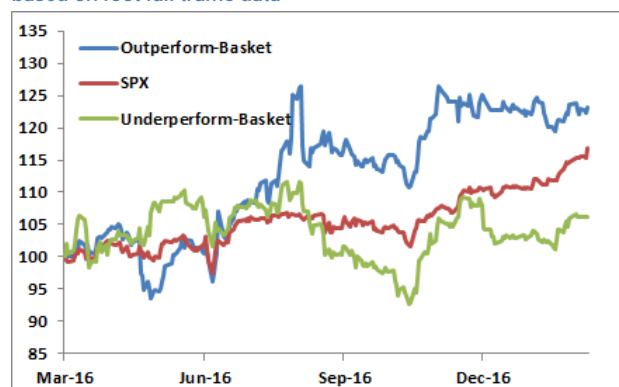
Source: J.P. Morgan QDS.

Figure 30: Performance of long/short strategy vs. S&P 500

Strategy	Mean	Volatility	Sharpe ratio
Long-Short Strategy	16.52%	18.81%	0.88
SPX Index	17.12%	10.19%	1.68

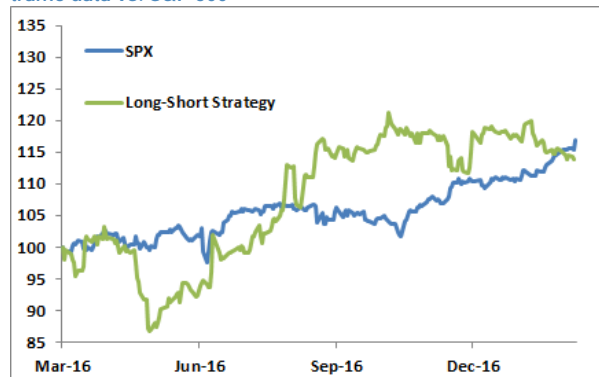
Source: J.P. Morgan QDS.

Figure 31: Performance of Outperform and Underperform baskets based on foot fall traffic data



Source: J.P. Morgan QDS, Advan Research, Bloomberg and FactSet.

Figure 32: Performance of Long-Short basket based on foot fall traffic data vs. S&P 500



Source: J.P. Morgan QDS, Advan Research, Bloomberg and FactSet.

Case Study: Satellite Imagery of Parking Lots and Trading Retail Stocks ([RS Metrics](#))

RS Metrics analyzes geospatial data drawn from various sources including satellites, drones and airplanes. RS Metrics provides signals, predictive analytics, alerts and end-user applications. The firm uses ~10 high-resolution satellites, orbiting the Earth and imaging between 11 AM and 1:30 PM. Such data can be used to estimate retail traffic (e.g. at some ~50 retail chains). Besides the retail data, the firm also estimates commercial real estate traffic activity, production and storage of base metals (including Aluminum, Copper and Zinc) and employment data at factory locations.

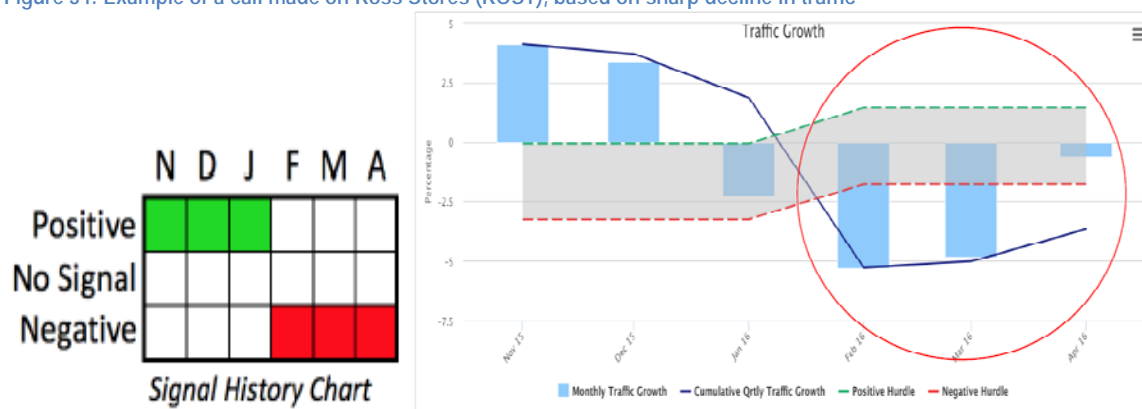
Figure 33: RS Metrics: sample satellite images of semi-trailers, employee cars and finished inventory



Source: RS Metrics LLC.

RS Metrics provided us with their buy-sell indicators for a set of retail stocks, alongside with detailed signal construction methodology. A sample anecdote is provided below. Based on traffic growth indicators, RS Metrics reported negative signals three times in 1Q 16 for Ross Stores. In line with predictions, ROST reported lower than expected revenues on May 19th.

Figure 34: Example of a call made on Ross Stores (ROST), based on sharp decline in traffic

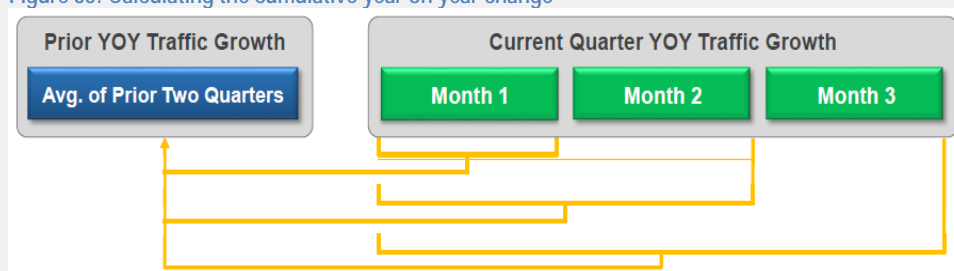


Source: RS Metrics LLC.

The methodology for generating the buy, sell and hold signals is described below. For current analysis, JPM Research relied on the final buy/sell/hold indicators provided by RS Metrics.

Every month the YoY change in traffic is calculated. Traffic is measured by the fill rate across all parking lots for a given company. Fill rate is defined as the ratio of the total number of cars by the total number of parking lot spaces. There is also a “cumulative year on year” change, which is a rolling weighted moving average of the monthly data. This is reset every quarter.

Figure 35: Calculating the cumulative year on year change

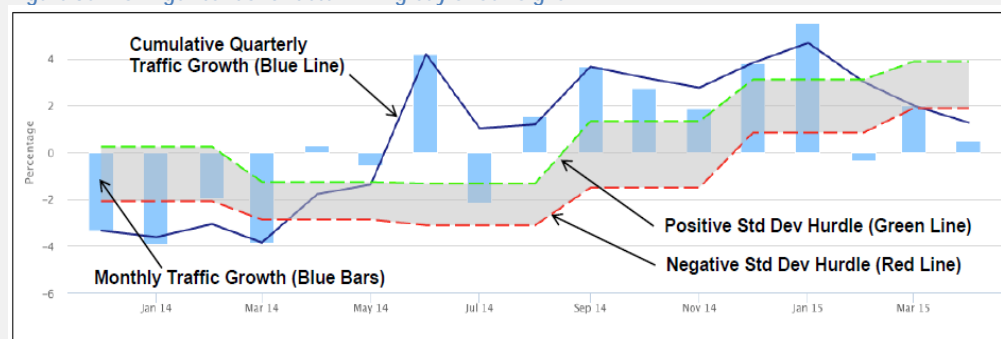


Source: RS Metrics LLC

Bollinger bands are calculated every month (± 0.5 standard-deviation), where mean and standard-deviation are calculated over the trailing 6 months.

If the “cumulative year-on-year” number is more than mean + 0.5 standard deviations, a buy signal is issued. If less than mean – 0.5 standard deviations, a sell signal is issued.

Figure 36: Bollinger bands for determining buy or sell signal

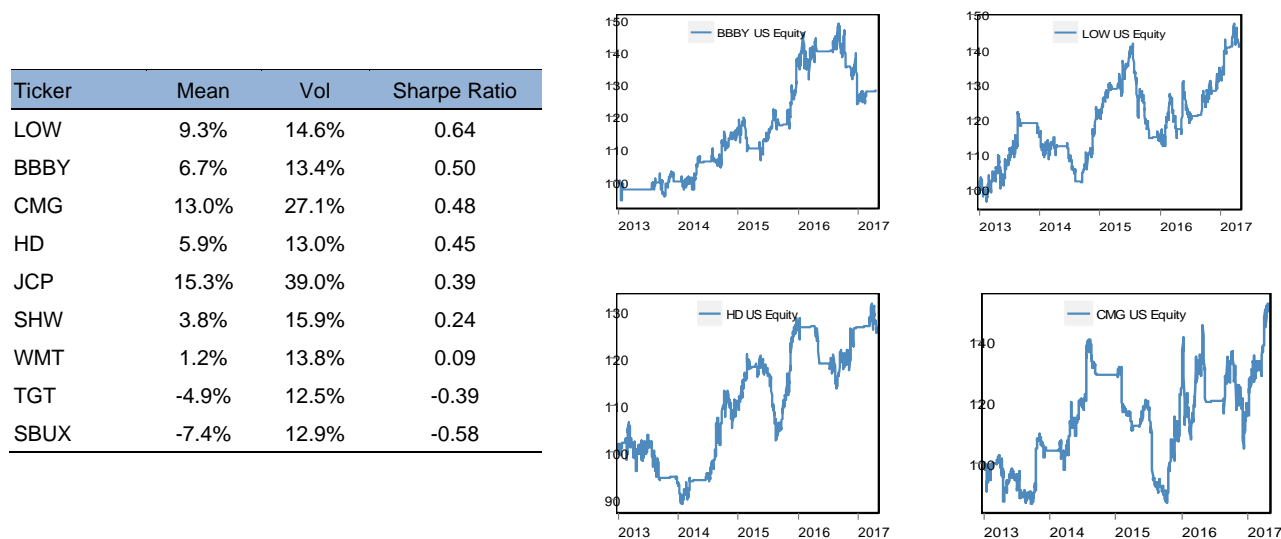


Source: RS Metrics LLC

Our strategy involved trading individual stocks based on the buy/sell indicators from traffic. Positions were taken and held until the next signal. We concentrated on names in the S&P 500 with sufficiently long history of the signal; results for other stocks in the basket are also provided for completeness. The strategy for each stock was as follows: a position starts from the date on which a signal (1, 0, -1) is produced by RS Metrics and the position is held until a change in the signal. A signal of 1 triggers a long stock / short US Retail index benchmark position; -1 triggers a short stock and long US Retail index position. For a 0 signal, we are invested in cash.

Results were positive for most of the S&P 500 stocks for which data was available since Jan 1, 2013. There are a total of 15 stocks which are in the S&P 500, and only 9 out of them have history since 2013.

Figure 37: Performance of strategy using RS Metrics signal (selected stocks in the S&P 500 with history since 2013)



Source: J.P.Morgan Macro QDS, RS Metrics LLC, Bloomberg.

We have also aggregated performance for all stocks with history since 2013. The performance is shown in the figure below. The strategy delivered a Sharpe ratio of 0.68, and had very low correlation to the S&P 500 (7.2%).

Figure 38: Performance of retail stock strategy using car count signal



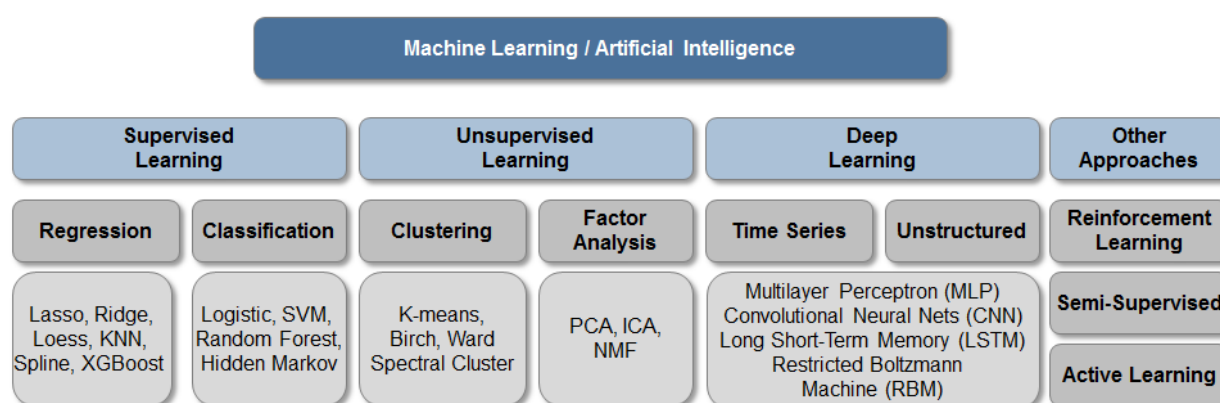
Source: J.P.Morgan Macro QDS, RS Metrics LLC, Bloomberg.

III: MACHINE LEARNING METHODS

Overview of Machine Learning Methods

In this chapter, we analyze different Machine Learning methods and illustrate their application in specific investment examples and trading strategies. One section of this chapter will be dedicated to each class of Machine Learning methods: Supervised Learning, Unsupervised Learning, Deep Learning, and Other Machine Learning approaches (e.g. Reinforcement learning). At the end of the chapter, we compare the performance of different Machine Learning techniques when applied on specific trading strategies or regime classification problems. The chapter will follow the classification of Machine Learning methods that we introduced in the first chapter (Figure below, please read the [Machine Learning introduction](#)).

Figure 39: Classification of Machine Learning techniques



Source: J.P.Morgan Macro QDS

In the description of individual methods (e.g. Regressions, Classifications, Clustering, etc.) we first provide a basic introduction and rationale for using the method. We then demonstrate the use of method on a specific financial example. Theoretical considerations as well as examples of codes and libraries (e.g. in R and Python) are also provided for each method. Mathematical considerations (e.g. formulas, derivations) are placed in text boxes, and readers more interested in practical application can skip these.

Machine Learning is a hybrid field combining concepts and research from two fields: Statistics (Mathematics) and Computer Science. Related to its origins, Machine Learning can roughly be divided into two categories: Classical Machine Learning and Deep Learning. Classical Machine Learning methods of supervised and unsupervised learning are a natural extension of statistics. Most of the applications by financial analysts fall in the category of classical Machine Learning. Methods of Deep Learning are critical in processing satellite images, natural language processing, and analysis of other unstructured data. In practice, front office investment professionals will likely acquire Big Data signals that are already processed with methods of Deep Learning. For example, an analyst will acquire car count as a time series, and is less likely to build a convolutional neural network framework to count cars directly from satellite image files. Methods of deep and reinforcement learning do offer great promise in time series analysis and portfolio construction. For instance, recognizing price patterns and forecasting returns can likely be significantly improved via the use of Deep Learning. Applying methods of reinforcement learning to automated trading offers a promise of building strategies that are many times faster and likely more successful than what can be achieved by an average discretionary trader.

Terminology: The diverse origin of Machine Learning has led to sometimes confusing terminology used to describe them. For example: inputs, features, attributes, predictors and independent variables are all synonyms. Similarly: outputs, outcomes, responses and dependent variables are also synonyms. There are also words which have different connotations when used by different practitioners. Data mining, in the quant trading context is a negative term that refers to overfitting, whereas in the context of computer science data mining refers to the insight obtained through a thorough analysis of large data sets. A factor, in the context of statistical learning literature is always qualitative, while in the context of Machine Learning and quantitative finance, a factor can be a FAMA-French risk factor or a PCA factor. Even within Machine

Learning, the terminology is not uniform. Logistic regression is a classification, not regression technique, etc. The table below links some terms that are interchangeably used in Statistics and Machine Learning.

Figure 40: Statistics vs Machine Learning terminology

Term in Statistics	Equivalent Term in Machine Learning
Statistical Learning	Classical Machine Learning
Independent Variable, X	Input Feature, attribute
Dependent Variable, Y	Output Feature, response
In-Sample	Training Set
Out-of-Sample	Test Set
Estimate, Fit a Model	Learn a Model
Model Parameters	Model Weights
Regression	Supervised Learning
Clustering and Dimensionality Reduction	Unsupervised Learning
Classifier	Hypothesis
Response (e.g. 0 or 1)	Label
Data Point	Example, Instance, Training Sample

Source: J.P.Morgan Macro QDS

Classical Machine Learning

Supervised and Unsupervised Learning are often called Classical Machine Learning²⁶. A Machine Learning algorithm calibrates its parameters by using data fed by an analyst i.e. the algorithm learns (fits) the model and improves this model as more and more data are collected. The term “supervised learning” arises from noting that the analyst guides (and hence supervises) the computer’s algorithm in its calibration of parameters by giving it a training set with clearly labelled input variables and clearly labelled output or predicted variables.

Take an example of predicting the relationship of market returns to different macro variables such as oil, dollar, volatility, consumer sentiment, rates, etc. In traditional statistics, an analyst would employ linear regressions to compute the beta of market returns to these variables. Through Machine Learning, an analyst could compute exposures with advanced regression models that will account for outliers, deal with a large number of variables in robust way, discriminate between correlated input variables, account for potential non-linear effects, etc. Such extensions to linear regressions are made possible through new algorithms developed by computer scientists. For example, one extension – called lasso regression – chooses the smallest, necessary subset of input variables. Another algorithm – called logistic regression – is tailored to handling data, where the resulting output is binary valued, as in “buy” or “sell”.

Another common task in financial analysis is to understand the factors that are driving the asset price. For example, an equity analyst might seek to attribute stock returns to returns of the broad market, sector, style, etc. In classical statistics, an analyst would employ a technique known as Principal Component Analysis (PCA). PCA (or similar methods such as Independent Component Analysis - ICA) carry over to Machine Learning without any changes. Algorithms such as PCA fall under the umbrella of “Unsupervised Learning” in the Machine Learning worldview. The term “unsupervised learning” arises from noting that the computer is simply given the entire set of returns of assets; the computer does not have any notion of independent variables or any notion of an output/dependent variable. Clustering is another method of unsupervised learning. It involves splitting a set of variables into smaller groups based on some notion of similarity.

Deep Learning

In recent years, a new set of analyses techniques have emerged which are loosely inspired by the working of the human brain. Inside our brain, we have a network of individual neurons. Each neuron receives an electrical current input from

²⁶ For a broad overview of statistical learning, see the canonical text from Hastie-Tibshirani-Friedman (2009). For other treatments, see Duda et al (2000), Bishop (1995, 2006), Cherkassy and Mulier (2007), Vapnik (1996) and Friedman (1994b).

many sources. The neuron roughly computes a weighted average of these inputs, where the relative weighting of different inputs is guided by its past experience. If this weighted average exceeds a certain in-built threshold, the neuron “fires” and sends out an output to other neurons. On the other hand, if the weighted average is below the threshold value, the neuron remains de-activated and simply ignores the inputs. Computer scientists have found that they could replicate these structures, and use them in both supervised as well as unsupervised learning tasks. The use of such multi-layered neural networks is called Deep Learning²⁷.

Deep Learning is behind almost all of the prominent accomplishments of Machine Learning in the past decade. Image recognition, speech recognition, language translation and even self-driving cars all rely on the new Deep Learning algorithms. Unlike supervised and unsupervised learning, relatively little is known about applications of Deep Learning to trading. Deep Learning is certainly used to process Big Data: satellite images, natural language processing, etc. Most investment managers will reap benefit of Deep Learning, without actually implementing the methods.

Selecting the ‘right’ Machine Learning Model

There is no one Machine Learning model that will be the best choice for all investment research projects. For this reason quantitative analysts should be familiar with a broad range of Machine Learning models and their application (in addition to understanding methods of analysis, one needs to understand the dataset and financial assets used to trade the data signal). The first step in tackling a dataset is to make an educated guess on which method of analysis is expected to yield the best results. The table below illustrates this by linking typical tasks and frequently used methods. The rest of this chapter will detail the use of various Machine Learning methods on specific financial examples.

Figure 41: Typical tasks and frequently used Machine Learning methods

Question	Data Analysis Technique
Given set of inputs, predict asset price direction	Support Vector Classifier, Logistic Regression, Lasso Regression, etc.
How will a sharp move in one asset affect other assets?	Impulse Response Function, Granger Causality
Is an asset diverging from other related assets?	One-vs-rest classification
Which assets move together?	Affinity Propagation, Manifold Embedding
What factors are driving asset price?	Principal Component Analysis, Independent
Is the asset move excessive, and will it revert?	Component Analysis
What is the current market regime?	Soft-max classification, Hidden Markov Model
What is the probability of an event?	Decision Tree, Random Forest
What are the most common signs of market stress?	K-means clustering
Find signals in noisy data	Low-pass filters, SVM
Predict volatility based on a large number of input variables	Restricted Boltzmann Machine, SVM
What is the sentiment of an article / text?	Bag of words
What is the topic of an article/text?	Term/InverseDocument Frequency
Counting objects in an image (satellite, drone, etc)	Convolutional Neural Nets
What should be optimal execution speed?	Reinforcement Learning using Partially Observed Markov Decision Process

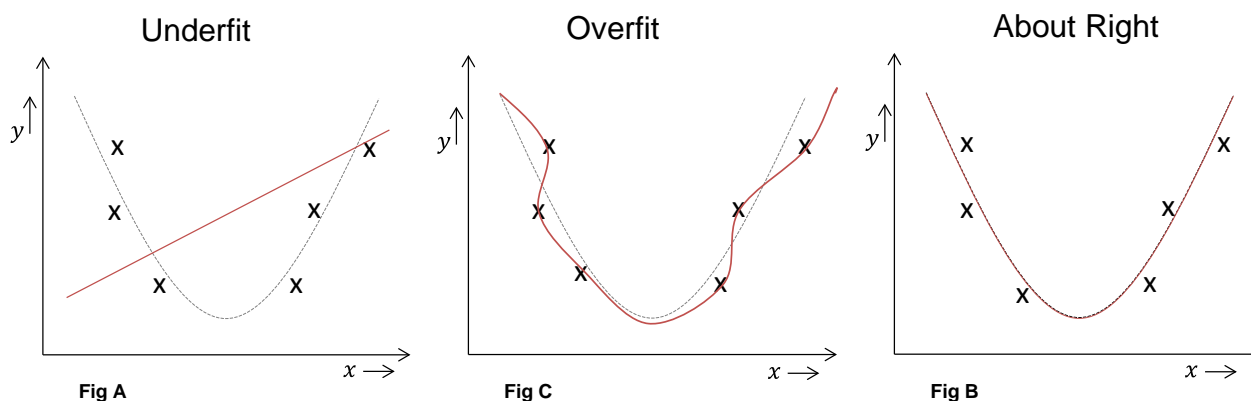
Source: J.P.Morgan Macro QDS

After selecting a specific Machine Learning method, one still needs to specify the number of model parameters and their values. As we [discussed in the first chapter](#), this is often related to the tradeoff between ‘variance and bias’, which we again summarize below.

²⁷ Deep learning is covered in a recent book by Bengio et al – Bengio [2016]. Deep learning algorithms are also reviewed at length in Schmidhuber (2014), LeCun (2015); older reviews include Bengio (2012), Bengio (2009) and Graves (2012). Reinforcement Learning is covered in the classic text by Sutton and Barto (); newer treatments are available from David Silver, John Schulman and documentation of Open AI’s gym framework

To illustrate the ‘variance and bias’ tradeoff, consider the three graphs below. For these graphs, we first generated data for a simple quadratic function plus some random noise (e.g. $y = x^2 + n$, $n \sim N(0, 0.5^2)$). In finance these quadratic functions may be encountered in e.g. response of USD to level of economic growth, or market volatility’s response to level of inflation. We can fit quadratic data with a linear, quadratic and a high order polynomial (e.g. 10th order).

Figure 42: Variance-bias trade-off (overfitting vs. underfitting)



Source: J.P.Morgan Macro QDS

It is clear that a linear fit ‘under-fit’, introducing large historical errors also known as ‘bias’ (of predicted versus observed data points). It is clear that a linear model is too simple and inadequate to explain the data points. Use of a High order polynomial (figure in the middle), resulted in very low error on historical data (low bias) but this model is clearly ‘over-fitting’ the data. Over-fitting implies that new data points will likely not be predicted with same degree of accuracy, and our historical estimation of error may be misleading. In learning theory terminology, we have higher *variance* of the model (but lower historical ‘bias’).²⁸

As model complexity increases, one expects the bias to decline and variance to increase. Total error of our forecast is the sum of both variance and bias. As the PnL of a trading strategy is negatively impacted by error in the forecast, the goal is always to minimize the forecast error. The model should not be too complex (to increase variance), and it should not be too simple (to increase bias). Selecting the optimal level of model complexity is illustrated in the diagram in the first chapter (see Figure 8).

Another important question is how we measure model complexity. Informally, we shall measure model complexity by the number of parameters. For optimal model selection one can use cross-validation which is described in the Appendix (a brief introduction to Learning Theory and use the notion of Vapnik-Chervonenkis dimension to measure model complexity).

Variance – Bias Tradeoff

In mathematical terms, we can formalize Variance-Bias tradeoff as follows. We are trying to estimate functional relationship between inputs (x) and output (y) in the presence of random noise
 $y = f(x) + n$, $n \sim N(0, \sigma^2)$.

Variables (x) are some set of signals, and predicted variable (y) is often future return of asset. Goal of Machine Learning is to come up with a model function \hat{f} , that mimic function y and is used to predict future asset returns (y). Error in our estimation (this error will be detrimental to our forecast and hence PnL of a strategy) is given as

²⁸ One can interpret bias and variance as follows. Bias is the portion of the generalization error attributable to the simplifying assumptions made in the model. Hence, it naturally declines as model complexity increases. Variance is the portion of generalization error generated through the fitting of idiosyncratic and spurious patterns found in any finite/limited sampling of a population. Variance increases as model complexity increases.

$$E \left[(Y - \hat{f}(X))^2 \right] = [Bias(\hat{f})]^2 + Variance(\hat{f}) + IrreducibleError$$

Here, bias is defined as

$$Bias(\hat{f}) = E_X [f(X) - \hat{f}(X)]$$

And variance is defined as

$$Variance(\hat{f}) = E_X \left([\hat{f}(X)]^2 \right) - (E_X[\hat{f}(X)])^2$$

And irreducible, random error is

$$IrreducibleError = \sigma^2$$

The rest of this chapter is organized as follows:

We first present methods of **Supervised Learning**. From the category of Regressions (and related) Models, and **Classification** models we show:

Regressions

- [Penalized Regression Techniques: Lasso, Ridge, and Elastic Net](#)
- [Non-Parametric Regression: Loess and K-Nearest Neighbor](#)
- [Dynamical Systems: Kalman Filtering](#)
- [Extreme Gradient Boosting](#)

Classification

- [Logistic Regression](#)
- [Support Vector Machines](#)
- [Decision Trees and Random Forests](#)
- [Hidden Markov Model](#)

In the category of **Unsupervised Learning**, we illustrate various methods of Clustering and Principal Components analysis:

Unsupervised learning

- [Clustering and Factor Analysis](#)

Deep and Reinforcement Learning is illustrated in the following methods:

- [Multi-Layer Perceptron](#)
- [Time-Series Analysis: Long Short-Term Memory](#)
- [Convolutional Neural Networks](#)
- [Restricted Boltzmann Machines](#)
- [Reinforcement Learning](#)

Finally, we present several case studies in which we compare different Machine Learning methods on both actual tradable strategies, and simulated sample data:

Comparison Case Studies

- [Supervised Learning - Regression](#)
- [Supervised Learning - Classification](#)
- [Unsupervised Learning - Clusterings](#)

Supervised Learning: Regressions

In **supervised learning**, an algorithm is provided historical data (both input and output variables), and is trying to find the relationship that has the best predictive power on out-of-sample data. Methods of supervised learning are further classified in methods of regression and methods of classification. **Regressions** try to predict output variables based on a number of input variables. **Classification** methods attempt to group or classify output into categories. For instance we may want the output of a model to be a binary action as ‘buy’ or ‘sell’ based on a number of variables. One can think of regression and classification as the same methods, with the forecast of a regression being a continuous number (e.g. market will go up 1%, 1.15%, 2%, etc.) and forecast of classification being a discrete number (e.g. buy=1, sell=0, or volatility regime will be: high=+1, medium=0, low=-1).

Even a simple linear regression can be thought of as a supervised Machine Learning method. However, linear regressions may not be suitable to deal with outliers, a large number of variables, variables that are correlated, or variables that exhibit non-linear behavior. To illustrate one example of the inadequacy of simple linear regression, consider a hypothetical regression forecast of a risky asset price:

$$\text{Asset Price} = 0.2 * \text{US Growth} - 0.05 * \text{EM Growth} + 0.1 * \text{US HY Bonds} + 22 * \text{US Equities} - 21 * \text{Global Equities}$$

Ordinary Linear regression²⁹ typically produces these types of results (i.e. large opposite sign coefficients) when one includes variables that are highly correlated (such as US and Global Equities). In Big Data strategies, one is expected to include a large number of variables, without necessarily knowing which ones are correlated and which ones are not. The simple regression above would suggest an irrational trading strategy with unnecessarily high leverage and idiosyncratic risk. One simple extension of linear regression is called a **Lasso regression**. Lasso regression tries to establish the relationship (a forecast) by choosing the smallest and most relevant set of input variables. In the example above, Lasso regression would ‘realize’ that US and Global Equities are highly correlated, and would penalize and eliminate the large long-short position.

K-nearest neighbors method forecasts data by simply looking at historical samples and establishing what has happened in similar situations as a best forecast of the future. While K-nearest neighbors is a simplistic method that overly relies on historical patterns, it has the advantage of including non-linear effects. In this section, we analyze the following regression methods: **Lasso, Ridge, Elastic Net, and K-nearest neighbors**.

In the Mathematical box below we provide some additional theory behind regressions in the context of supervised Machine Learning. This can be skipped by all but the most quantitatively inclined readers.

In supervised learning, we seek to determine an appropriate functional relationship between an output or target variable and a set of input or predictor variables. Given m training examples, denoted by $(\underline{x}^{(i)}, y^{(i)})$, we seek the function h such that $y = h(\underline{x})$. Defining \mathcal{X} as the input space for input variables (say, \mathbb{R}^n) and \mathcal{Y} as the space of output values (say, \mathbb{R}); it is conventional to refer to the function $h: \mathcal{X} \rightarrow \mathcal{Y}$ as the hypothesis function. The learning task is called either a regression or a classification, depending on whether the output space \mathcal{Y} is continuous (as in \mathbb{R}) or discrete (as in $\{0, 1\}$).

In the classical Ordinary Least Squares model for linear regression, one defines $h_{\underline{\theta}}(\underline{x}) = \underline{\theta}^T \underline{x}$, where $\underline{\theta}$ is the parameter or weight vector. Errors during model fitting are penalized using the cost function $J(\underline{\theta}) = \frac{1}{2} \sum_{i=1}^m (h_{\underline{\theta}}(\underline{x}^{(i)}) - y^{(i)})^2$. The

²⁹ For classical extensions and analysis of ordinary linear regression, see Seber (1984), Weisberg (1980) and Kennard (1970); for ridge, see Hoerl and Kennard (1970); for lasso, see Tibshirani (1996). Shrinkage methods are compared in Frank and Friedman (1993). Bayesian methods for estimating regression parameters are covered in George and McCulloch (1993), Madigan and Raftery (1994), Clyde, DeSimone and Parmigiani (1996), West (2003). Basis pursuit in signal processing was proposed in Chen et al (1998). For least angle regression and related homotopy methods, see Efron et al (2004), Osborne et al (2000a) and Osborne et al (2000b). Forward stage-wise criterion and a path algorithm for generalized regression models are covered in Hastie et al (2007) and Park and Hastie (2007). PLS (Partial least squares) was introduced in Wold (2007).

traditional approach was to stack the training examples to form $X = \begin{bmatrix} \underline{x}^{(1)T} \\ \vdots \\ \underline{x}^{(m)T} \end{bmatrix}$ and $\underline{y} = [y^{(1)} \dots y^{(m)}]^T$. Matrix

differentiation of the cost function $J(\underline{\theta}) = ||X\underline{\theta} - \underline{y}||^2$ yields the normal equations $X^T X \underline{\theta} = X^T \underline{y}$ and the estimator as $\underline{\theta} = (X^T X)^{-1} X^T \underline{y}$. Unfortunately, this traditional approach of theoretical derivation and practical use is not extensible to modern Machine Learning models. So we review this question using two other approaches, namely

- A purely numerical approach, which we shall use to derive perceptron models; and
- A purely probabilistic approach, which we shall use to derive ridge and lasso regressors.

The numerical approach is to note that $J(\underline{\theta})$ can be minimized using the gradient descent algorithm, where the j^{th} element of the vector $\underline{\theta}$ is updated as $\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\underline{\theta})$. Here α is the learning rate. Repeated decrease in the direction of steepest descent of $J(\underline{\theta})$ leads to convergence to a local (global in the case of convex functions) minima. Applied to our definition of $J(\underline{\theta})$ and denoting the j^{th} element of the vector $\underline{x}^{(i)}$ as $x_j^{(i)}$, the gradient descent rule yields $\theta_j \leftarrow \theta_j + \alpha (y^{(i)} - h_{\underline{\theta}}(\underline{x}^{(i)})) x_j^{(i)}$. This rule is called the Widrow-Hoff or Least Mean Squares (LMS) rule in Machine Learning literature.

When the LMS rule is applied – as shown below – in one shot to all training examples, it is called Batch Gradient Descent.

Repeat until convergence

$$\{ \forall j \in \{1, \dots, n\}, \theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^m (y^{(i)} - h_{\underline{\theta}}(\underline{x}^{(i)})) x_j^{(i)} \}$$

If the training set is very large, we can update all the weights using information from each individual training example, instead of iterating over the entire training set to update a single weight parameter. This idea leads to Incremental or Stochastic Gradient Descent (SGD).

Repeat until convergence

$$\{ \forall i \in \{1, \dots, m\} \{ \forall j \in \{1, \dots, n\}, \theta_j \leftarrow \theta_j + \alpha (y^{(i)} - h_{\underline{\theta}}(\underline{x}^{(i)})) x_j^{(i)} \} \}$$

While stochastic gradient descent lacks theoretical guarantees on convergence of $\underline{\theta}$ to the local minima, one finds its performance to be superior to batch gradient descent on large data sets. We shall use stochastic gradient descent to train many of the models employed in this primer. The above LMS rule shall also recur in identical form when we study logistic classifiers.

To motivate and understand many of the Machine Learning algorithms, researchers rely frequently on probabilistic interpretations. In the context of Ordinary Least Squares, one can model the output as $y^{(i)} = \underline{\theta}^T \underline{x}^{(i)} + \varepsilon^{(i)}$. If $\varepsilon^{(i)} \sim N(0, \sigma^2)$ represents independent and identically distributed (i.i.d.) noise, then the likelihood of observing the particular training set is given by the likelihood function

$$L(\underline{\theta}) = p(y|X; \underline{\theta}) = \prod_{i=1}^m p(y^{(i)} | \underline{x}^{(i)}; \underline{\theta}) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \underline{\theta}^T \underline{x}^{(i)})^2}{2\sigma^2}\right).$$

The principle of Maximum Likelihood mandates the choice of parameter $\underline{\theta}$ to maximize the above likelihood expression, i.e. render the appearance of the training set as likely as possible. Maximizing equivalently the log-likelihood $l(\underline{\theta})$, we recover

$$\hat{\theta}_{ML} = \arg \min_{\underline{\theta}} l(\underline{\theta}) = \arg \min_{\underline{\theta}} \log L(\underline{\theta}) = \arg \max_{\underline{\theta}} \frac{1}{2} ||X\underline{\theta} - \underline{y}||^2.$$

This is same as the expression obtained in the traditional analysis for OLS. The advantage of this probabilistic modeling is that we will extend it using Bayesian priors to derive more stable linear regression techniques like ridge and lasso.

Penalized Regression Techniques: Lasso, Ridge, and Elastic Net

Penalized regression techniques like Lasso (also, spelt as LASSO) and Ridge are simple modifications of ordinary linear regression aimed at creating a more robust output model in the presence of a large number of potentially correlated variables. When the number of input features is large or when the input features are correlated, classical linear regression has a tendency to overfit and yield spurious coefficients. LASSO, Ridge, and Elastic Net regressions are also examples of ‘regularization’ – a technique in Machine Learning that is expected to reduce the out-of-sample forecasting errors (but does not help with reducing in-sample backtest errors).

In ordinary linear regression, we forecast the value y to be a linear combination of the inputs x_1, x_2, \dots, x_n . In short we assume that variable y has ‘Betas’ to a number of variables x (plus some random noise).

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon.$$

To find the ‘betas’ $\beta_0, \beta_1, \dots, \beta_n$, linear regression minimizes the historical error (square of error) between actual observations of variable ‘ y ’, and predicted (or model) values of the variable. This is the reason the method is also called least-squares (since it minimizes the square of errors):

$$\text{OLS: Minimize Historical Sum of } \left(y - \left(\beta_0 + \sum_{i=1}^n \beta_i x_i \right) \right)^2.$$

As we saw in the numerical example above, this minimization is not stable and can yield spurious and/or large values of betas. One way to prevent that from occurring is to change the objective function in the minimization above. Instead of minimizing the least-squares objective, we can modify it by adding a penalty term that reflects our aversion towards complex models with large ‘betas’. If we change the objective to include a penalty term equal to the absolute value of the beta coefficients, i.e.

$$\text{Lasso: Minimize Historical Sum of } \left(y - \left(\beta_0 + \sum_{i=1}^n \beta_i x_i \right) \right)^2 + \alpha \sum_{i=1}^n |\beta_i|,$$

then the optimizer will set ‘unnecessary’ and very large betas to zero. The addition of a penalty term equal to the absolute value of the coefficients is called L_1 regularization and the modified linear regression procedure is called Lasso (or LASSO). By concentrating only on the most relevant predictors, Lasso performs an implicit feature selection for us³⁰.

The objective function for Lasso can be understood as follows:

- If we set $\alpha = 0$, then we recover the coefficients of ordinary linear regression.
- As α increases, we choose a smaller and smaller set of predictors, concentrating only on the most important ones

Here, α is called a parameter of the model. Similarly, if we tweak our objective to include the square of the ‘betas’ we arrive at Ridge regression

$$\text{Ridge: Minimize Historical Sum of } \left(y - \left(\beta_0 + \sum_{i=1}^n \beta_i x_i \right) \right)^2 + \alpha \sum_{i=1}^n \beta_i^2,$$

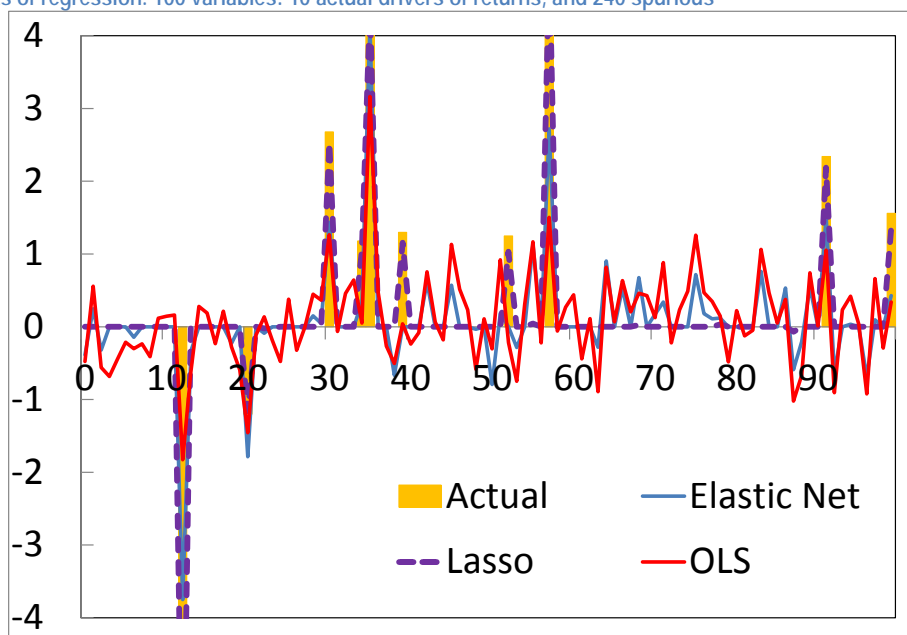
The additional penalty related to the square of the magnitude of the coefficients is called L_2 regularization and the modified linear regression procedure is called Ridge. An intermediate objective between the Ridge and Lasso objectives is used by Elastic Net regression, which finds

³⁰ A geometric intuition for sparsity comes from noting the level set of the L_1 norm to be a rhombus. Please see Hastie et al. (2013) for details.

$$\text{Elastic Net: Minimize Historical Sum of } \left(y - \left(\beta_0 + \sum_{i=1}^n \beta_i x_i \right) \right)^2 + \alpha_1 \sum_{i=1}^n |\beta_i| + \alpha_2 \sum_{i=1}^n |\beta_i|^2 .$$

We illustrate the 3 penalized regression examples with a hypothetical example (before considering an example of a realistic trading strategy). In the hypothetical example, we have chosen 10 variables that are actually driving the forecast for asset returns and added 90 spurious variables that are adding noise (throwing off the regression model). The algorithms were given all 100 variables (10 actual + 90 spurious) and asked to predict the weights for each feature. We plot the coefficients in graph below: horizontal axis ranges from 1 to 100 (reflecting the input variables) and vertical axis plots the values of beta coefficients as predicted by each algorithm. An ideal algorithm would select the 10 actual variables (light blue), and would discard the spurious ones.

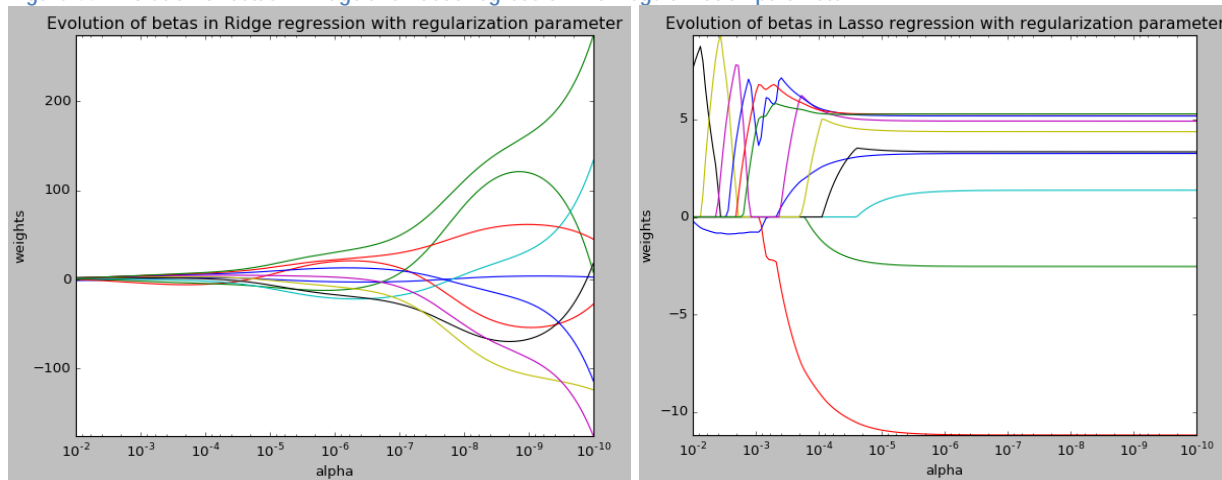
Figure 43: Coefficients of regression. 100 variables: 10 actual drivers of returns, and 240 spurious



Source: J.P.Morgan Macro QDS

Note that Lasso picked up a sub-set of actual coefficients. Ridge assigned weights to all coefficients, irrespective of whether they were spurious noise terms. Elastic net behaved in an intermediate way between Ridge and Lasso. As one increases the regularization parameter of the model (α), the models start suppressing spurious variables and focusing on actual ones. This is illustrated in the figures below showing the evolution of model 'betas' as we increase/decrease regularization parameter.

Figure 44: Evolution of betas in Ridge and Lasso regression with regularization parameter



Source: J.P.Morgan Macro QDS

Ridge, Lasso, and Elastic Net methods also have a Bayesian interpretation. While this is likely beyond the interest of practically oriented readers, we discuss it in the math box below.

Bayesian Interpretation of Penalized Regression

There is a natural Bayesian interpretation for both Ridge and Lasso regression³¹. We have shown earlier that for the linear model (assuming zero means) $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, $\underline{\varepsilon} \sim N(0, I)$, that the maximum likelihood estimate for $\underline{\beta}$ yields the ordinary least squares (OLS) answer. Suppose we assume a Laplacian prior on $\underline{\beta}$, i.e. $f(\underline{\beta}) \propto e^{-\lambda|\underline{\beta}|}$, the posterior distribution of $\underline{\beta}$ is given by

$$f(\underline{\beta} | \underline{y}) \propto f(\underline{\beta}) \cdot f(\underline{y} | \underline{\beta}) = e^{-\lambda|\underline{\beta}|} \cdot e^{-\frac{1}{2}\|\underline{y} - X\underline{\beta}\|^2}.$$

This implies that the maximum a posteriori (MAP) estimate for $\underline{\beta}$ is given as

$$\hat{\underline{\beta}}_{MAP} = \arg \max f(\underline{\beta} | \underline{y}) = \arg \min \|\underline{y} - X\underline{\beta}\|^2 + 2\lambda|\underline{\beta}|.$$

This is the same optimization as used in Lasso regression. Similarly, it is easy to see that assuming a Gaussian prior on $\underline{\beta}$, $f(\underline{\beta}) \propto e^{-\frac{1}{2}\|\underline{\beta}\|^2}$ will coerce the MAP estimate to obtain the Ridge regression estimate. It is known from Bayesian analysis, that assuming a prior avoids overfitting by using our prior knowledge (in this case, knowledge of the parsimonious nature of the model); unsurprisingly, using a prior distribution and deriving the MAP estimate leads to robust regressors³².

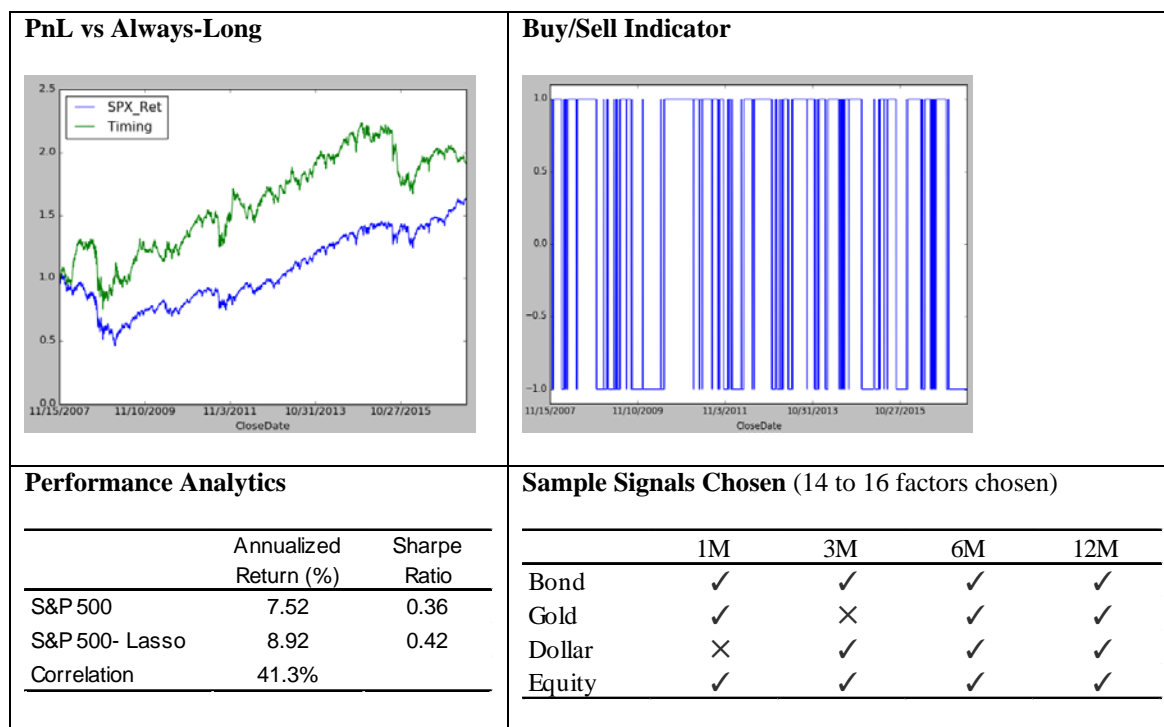
³¹ For statistical – including Bayesian – inference, see Gelman et al (1995), Neal (1996). For Markov Chain Monte Carlo methods, see Gelfand and Smith (1990) and Spiegelhalter et al (1996). Expectation-Maximization (EM) algorithm is covered by Neal and Hinton (1998).

³² The literature on Bayesian statistics is vast. Bayesian statistics for social sciences is covered in Hoff (2009), Gill (2002), Jackman (2009), Kruschke (2011), Christensen et al (2010). For reviewing application of Bayesian methods in different scenarios, see Breslow (1990) and the volumes by Gatsonis et al (1993-2002). The canonical result on exchangeability is from De Finetti (1974); see also Draper et al (1993). Historical development of Bayesian statistics is covered in Stigler (1986) – including the famous essays of Bayes (1763) and Laplace (1785, 1810). Non-informative priors are discussed in Jeffreys (1961), Kass and Wasserman (1996). Conjugate priors are covered in Box and Tiao (1973). Maximum entropy principle for constructing priors is discussed in Jaynes (1982, 1983), Donoho et al (1992). The debates between Bayesian and non-Bayesian approaches is given in Lindley (1958), Pratt (1965), Jaynes (1976), Berger and Wolpert (1984).

Example of Penalized Regression Approach in a Multi Asset Trend Following Strategy

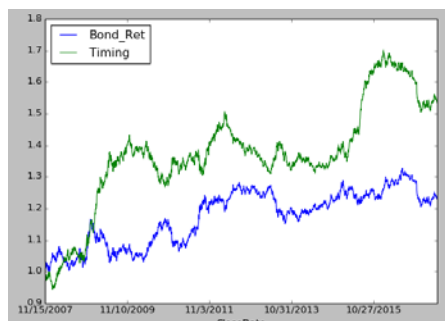
We illustrate an application of Lasso by estimating the 1-day returns of assets in a cross-asset momentum model. For more background on trend following strategies see our [CTA primer](#). We attempt to predict the returns of 4 assets: S&P 500, 7-10Y Treasury Bond Index, US dollar (DXY) and Gold. For predictor variables, we choose lagged 1M, 3M, 6M and 12M returns of these same 4 assets, yielding a total of 16 variables. To calibrate the model, we used a rolling window of 500 trading days (~2y); re-calibration was performed once every 3 months. The model was used to predict the next day's return. If the next day predicted return was positive, we went long the asset, otherwise we shorted it. Prior to regression, all inputs were standardized to avoid the problem of input features being of different scales. Performance of this momentum strategy is shown in tables below for each of the assets. Note that each of the momentum strategies outperformed a long only position in their respective asset.

S&P 500 (Lasso $\alpha = 0.001$) IR = 0.43

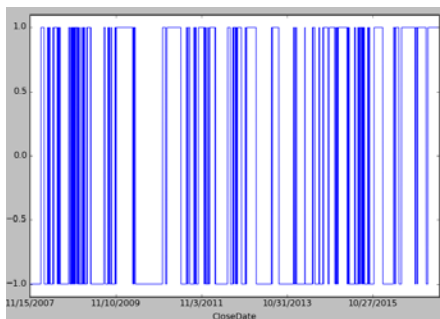


Result for IEF - Lasso ($\alpha = 0.001$) yields IR = 0.67

PnL vs Always-Long



Buy/Sell Indicator



Performance Analytics

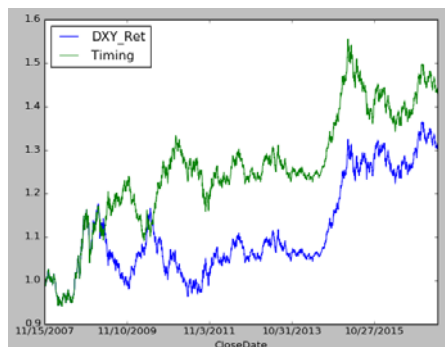
	Annualized Return (%)	Sharpe Ratio
IEF	2.30	0.32
IEF- Lasso	4.86	0.67
Correlation	-19.4%	

Sample Signals Chosen(12 to 16 factors chosen)

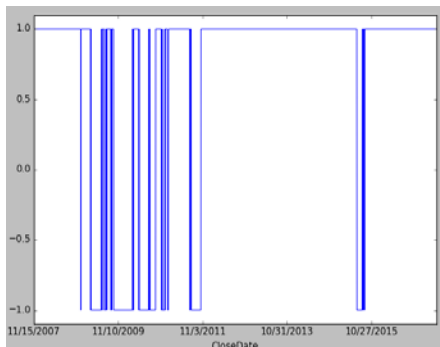
	1M	3M	6M	12M
Bond	✓	✓	✓	✓
Gold	✓	✓	✓	✓
Dollar	✓	✓	✓	✓
Equity	✓	×	✓	×

Result for DXY - Lasso ($\alpha = 0.05$) yields IR = 0.50

PnL vs Always-Long



Buy/Sell Indicator



Performance Analytics

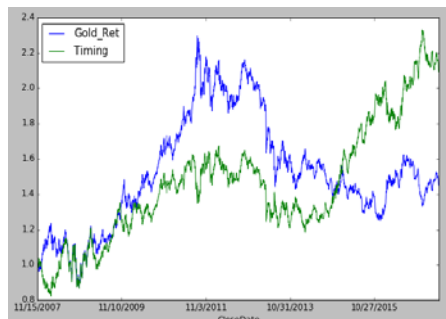
	Annualized Return (%)	Sharpe Ratio
DXY	3.22	0.38
DXY- Lasso	4.20	0.49
Correlation	58.9%	

Sample Signals Chosen(0 to 3 factors chosen)

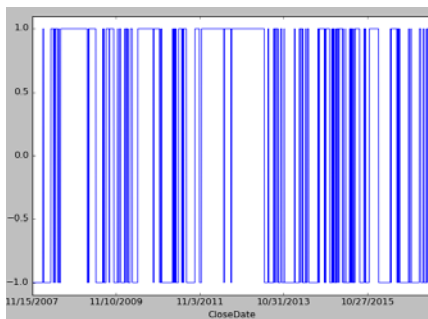
	1M	3M	6M	12M
Bond	×	×	×	×
Gold	×	×	✓	×
Dollar	×	×	×	×
Equity	✓	×	×	✓

Result for GLD - Lasso ($\alpha = 0.05$) yields IR = 0.50

PnL vs Always-Long



Buy/Sell Indicator



Performance Analytics

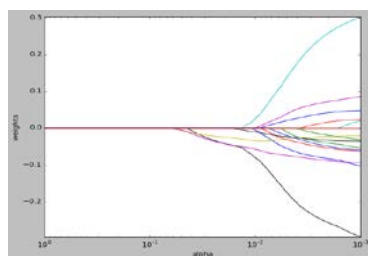
	Annualized Return (%)	Sharpe Ratio
Gold	6.12	0.31
GLD- Lasso	9.49	0.48
Correlation	20.3%	

Sample Signals Chosen(1 to 7 factors chosen)

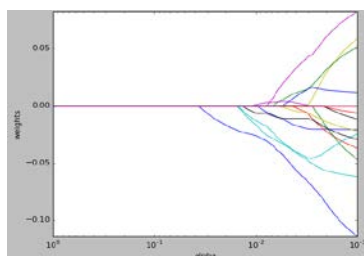
	1M	3M	6M	12M
Bond	×	✓	×	×
Gold	✓	×	×	✓
Dollar	✓	×	×	×
Equity	×	×	×	×

Evolution of betas as function of alpha in Lasso regression for 4 assets

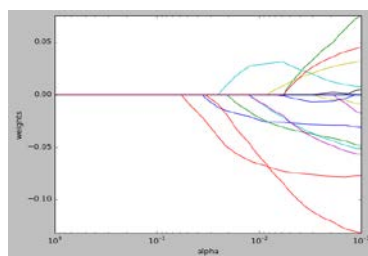
SPX



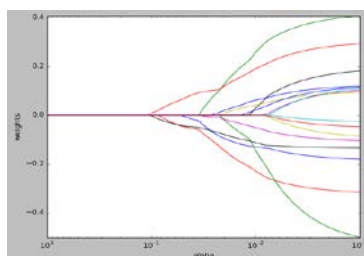
IEF



DXY

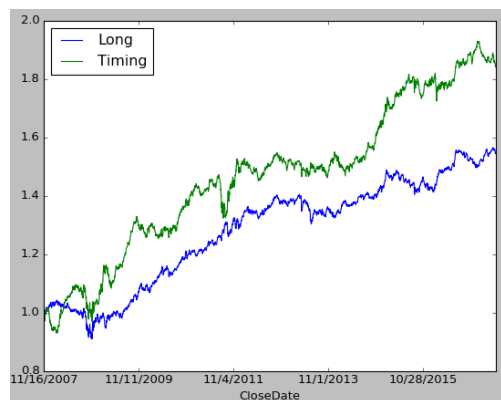


GLD



Combined Performance

PnL vs Long : Cross-Asset



Performance Analytics

	Annualized Return (%)	Sharpe Ratio
X-A Long	4.80	0.75
X-A Lasso	6.80	0.85

Non-Parametric Regression: K-Nearest Neighbor and LOESS

Supervised learning algorithms are sometimes classified as being either parametric or non-parametric techniques. In parametric techniques, the model is described by a set of parameters such as linear regression beta (that is estimated from historical data). In non-parametric techniques, we do calibrate parameters of a model, but directly identify similar historical instances, and assume the behavior will be the same. Given a new datapoint, we search through historical data and identify a number 'K' of similar instances that we call "nearest neighbors" 'NN' (hence name K-NN). Then we then make predictions by averaging historical outcomes for these "nearest neighbors". Two examples of non-parametric regressions are.

- K-Nearest Neighbor (KNN)³³ rule: Once we have located K nearest neighbors, we can average the output variable y for this subset and use that as our prediction.
- LOESS: Using data for the K nearest neighbors, for each new point we fit a linear regression based on the K nearest neighbors (subset of historical data), and predict the output using those coefficients. This is called LOESS or localized linear regression.

Non-parametric techniques offer a simple way to extrapolate analysis on past similar events. KNN is commonly used by financial analysts, who intuitively employ it without referring to it as such. The use of a historical sample makes the technique "supervised", while the absence of parameters or coefficient betas makes it "non-parametric". In many financial analyses, the output variable is often not linearly related to the inputs; this makes linear regression and its extensions like ridge and lasso unsuitable. In such cases, the K-nearest neighbor can capture those non-linear properties. A drawback of using the KNN rule lies in its extreme sensitivity to outliers.

One can view linear regression and k-nearest neighbor methods as two ends of the spectrum of classical Machine Learning³⁴. On one hand, linear regression 'under-fits' the data and hence suffers from higher 'bias' to achieve lower 'variance'. On the other hand, locating the K-most similar inputs and averaging their output makes a weak structural assumption. In formal terms, we say that the K-nearest neighbor method can 'over-fit' and hence suffer from higher 'variance' (while having a low 'bias').

We can gain further insight into the problem of under/over-fitting by varying the value of K in the KNN. At one extreme, we can set K to be equal to the number of samples itself, in which case we simply predict the quantized sample average as the output for all inputs. As we reduce K, we obtain a decision boundary that splits the training set into a number of parts. In the limiting case of K=1, we form a tiny region around each training sample. In this case, we have over-fit the data, which is likely to lead to a high error rate on unseen samples, while yielding a deceptively low (in this case, zero) error rate on the training set. One can informally think of a K-nearest neighbor method as dividing the N samples in historical data into N/K sets and computing their respective means. This implies that the effective number of parameters in K-nearest neighbor method is N/K, which increases as K decreases. The notion of effective number of parameters or degrees of freedom is a

³³ K-nearest neighbor rules and extensions are covered in Dasarthy (1991), Ripley (1996), MacQueen (1967), Hastie and Tibshirani (1996a), Simard et al (1993). For learning vector quantization, see Kohonen (1989). For discussion on Bayesian non-parametrics including Dirichlet Process Models, see Dunson (2009, 2010b), Shen and Ghosal (2013), Tokdar (2011), Wang and Dunson (2011a), Carvalho et al (2010), Ohlssen, Sharples and Spiegelhalter (2007), Ray and Mullick (2006), Rodriguez, Dunson and Gelfand (2009), Bigelow and Dunson (2009).

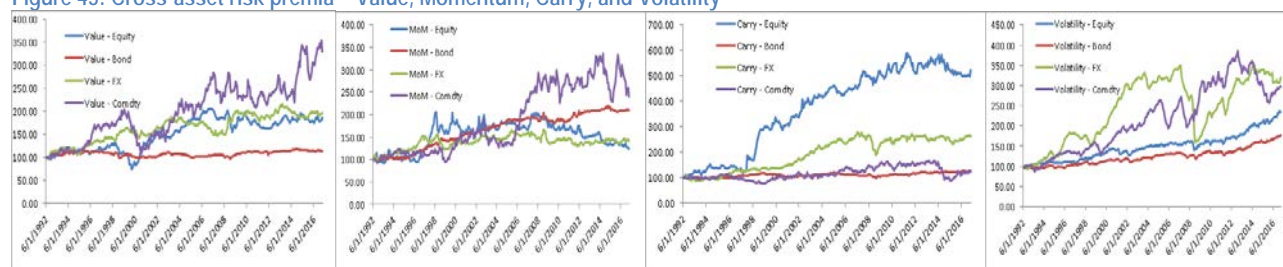
³⁴ Splines (including B-splines, thin-plate splines and reproducing kernel Hilbert spaces) are covered in Green and Silverman (1994), Wahba (1990), Girosi et al (1995) and Evgenion et al (2000). For wavelets, consult Daubechies (1992), Chni (1992), Wickerhauser (1994), Donoho and Johnstone (1994), Vidakovic (1999), Bruce and Gao (1996). Local regression and kernel methods are covered in Loader (1999), Fan and Gijbels (1996), Hastie and Tibshirani (1990). Statistical treatment of parametric non-linear models is illustrated in Reilly and Zeringue (2004), Gelman, Chew and Shnaidman (2004), Denison et al (2002), Chipman, George and McCulloch (1998, 2002), DiMatteo et al (2001) and Zhao (2000). For basis function models, see Bishop (2006), Biller (2000), DiMatteo, Genovese and Kass (2001), Barbieri and Berger (2004), Park and Casella (2008), Seeger (2008), Ramsay and Silverman (2005), Neelon and Dunson (2004), Dunson (2005), Hazelton and Turlach (2011), Hannah and Dunson (2011), Patti and Dunson (2011).

way to characterize the model complexity. As K decreases, and the number of parameters increases, the model complexity essentially increases, thereby leading to over-fitting.³⁵

Financial Example

In our previous work, we studied [cross-asset risk premia investing](#).³⁶ In particular, we had constructed simplified risk premia factors of value, momentum, carry, and volatility (across equities, bonds, currencies and commodities). Our analysis had shown that risk premia strategies delivered a positive Sharpe ratio over an extended period of ~40 years and exhibited low and stable correlation to traditional asset classes. We will attempt to use the K -nearest neighbor algorithm to illustrate potential for timing of these cross-asset risk premia based on macro regimes.

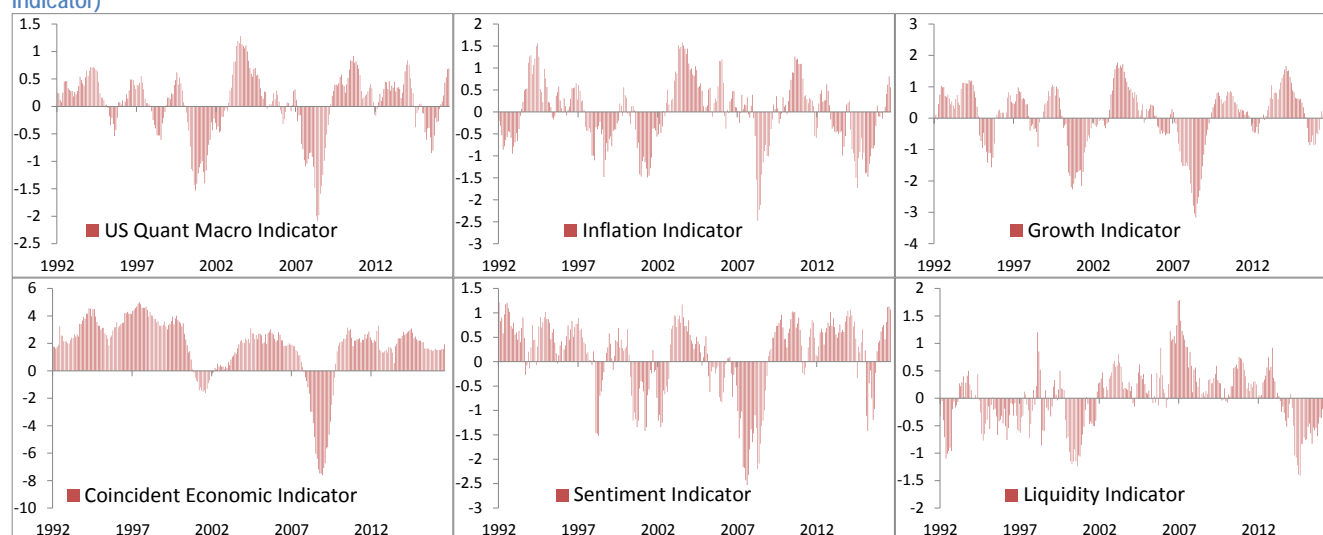
Figure 45: Cross-asset risk premia – Value, Momentum, Carry, and Volatility



Source: J.P.Morgan Macro QDS

To identify the current macro regime, we use 7 aggregated macro indicators: Quant Macro indicator, Growth, Liquidity, Sentiment, Inflation, Coincident economic indicator, as well as change in composite macro indicator. These aggregated indicators are based on 50 macro-economic signals; for details of construction, please see our earlier report titled [“Framework for style investing: Style rotation and the business cycle”](#).

Figure 46: Six aggregated macro indicators that are used to identify current economic regime (in addition, we use change in composite indicator)



Source: J.P.Morgan Macro QDS

³⁵ Formally, model complexity is characterized by Vapnik-Chervonenkis dimension of the set of classifiers or hypotheses. We describe that in a mathematical box in our discussion on model selection theory.

³⁶ Kolanovic, M and Wei, Z (2013), “Systematic strategies across asset classes: Risk factor approach to investing and portfolio management”, J.P.Morgan Research Publication.

Figure 47: List of 50 macro-economic signals used to construct the 7 macro indicators above (for more details see [here](#))

ISM Manufacturing PMI	ISM Non-manufacturing PMI
Initial Jobless Claims	US Capacity Utilization
US Leading Indicator	Yield Curve – 10y less 2y
Global Leading Indicator	Railroads Freight Index, 12m change
Leading-Lagging Economic Indicator Spread	US relative to World Stock Index
Baltic Dry Index	Global Economic Momentum
Manufacturing New Orders ex. Transportation	Retail Sales ex. Transportation
New Housing Permits	Credit Spread Moody's AAA-BAA
Michigan Consumer Sentiment	New Company Revisions, 3M Avg
NAPM Percentage Surprise, 3M Wt. Avg.	JULI Inv Grade USD Spread over Tsy
CESI – Citigroup Economic Surprise Index	VIX
Dow Transportation/Utilities, 12M Chg.	Small – Large Cap Outperformance, 12M Chg.
Barclays Investment Grade Spread	Barclays High Yield Spread
Term Structure of Momentum, Diffusion Index	Stock Market, 3Y Change
Margin Debt Level	Margin Debt as Percentage of Market Cap
Loan Officers Survey	M2 Money Stock, 12M Percentage Change
10Y Bond Yield, Real Rate	Correlation of MZM (Money) and SPX
Term Structure of Fed Fund Futures	USD Trade Weighted Index
Avg Treasury Trading Volume % of Mkt Debt	US Credit Manager Index
Free Liquidity	Earnings Yield Dispersion
Real Loan Growth	ISM Prices Index, 12M Change
PPI, YOY Percentage Change	Unit Labor Cost, 12M Change
US Import Price Inflation	Wage Trend Index, 12M Change
Breakeven Inflation. 10Y	Oil Price, WTI
Commodity Price Index	Median Home Price, 12M Change

Source: J.P.Morgan Macro QDS

We construct a monthly rebalancing strategy, where every month we look at a rolling window of the prior 10 years. We characterize the macro regime in a given month by the value of the 7 aggregated indicators. Then we look for K nearest neighbors, i.e.

- We compute the distance between the economic regime vector of the current month to the economic regime vector for all months in the past 10 year window.
- We rank them in ascending order of distance and choose the first (nearest) K neighbors

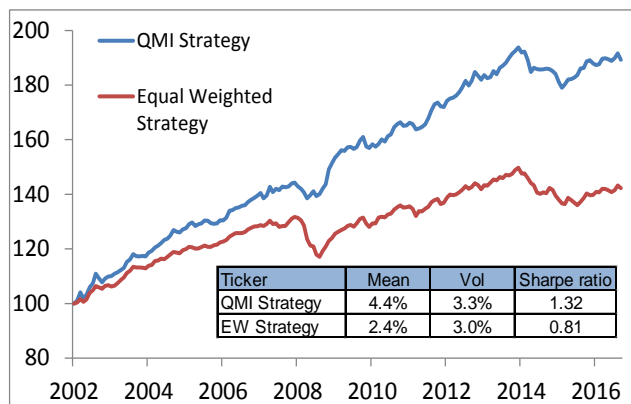
For each of these K nearest neighbors, we find the average return over a succeeding month for all 20 cross-asset risk premia factors. We then rank these factors according to their average returns, and choose a subset S that performed the best.

If we set S=1, it is equivalent to finding the best performing strategy historically in the same macro regime. If we set S=20, we will get an equal weighted strategy over the 20 risk premia. If we set K=1, we choose the nearest instance and replicate that for the current month. If we set K = 120, then we take all months in the input sample and average out the results, and invest in the best S strategies.

The Sharpe ratios of strategies based on K-nearest neighbors is tabulated below. We find that using K between 1 and 10, and number of risk factors between 10 and 19 generally outperforms simple equal weighted risk premia portfolio. For instance K=2 and S=14 yields the highest Sharpe ratio of 1.3, as compared to 0.8 shapre ratio of equal weighted strategy. We plot that case below alongside the equal weighted strategy that assigned equal weights to all 20 risk premia.

Figure 48: Sharpe ratio of portfolio chosen using K-Nearest Neighbor rule over cross-asset systematic strategies (left, K=1 to 25); Performance by timing cross-asset risk premia using JPM's Quant Macro Indicators (right)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.6	0.6	0.7	0.8	0.8	0.9	1.0	1.1	1.2	1.2	1.2	1.3	1.3	1.3	1.3	1.3	1.2	1.1	0.9	0.8
2	0.4	0.6	0.7	0.8	0.8	0.9	0.9	1.0	1.1	1.1	1.1	1.2	1.2	1.3	1.3	1.3	1.2	1.2	1.0	0.8
3	0.1	0.3	0.6	0.7	0.8	0.8	0.9	0.9	0.9	0.9	1.0	1.0	1.1	1.1	1.1	1.1	1.1	1.1	1.0	0.8
4	0.2	0.2	0.3	0.5	0.6	0.6	0.7	0.7	0.8	0.8	0.8	0.8	0.9	0.9	0.9	1.0	0.9	1.0	1.0	0.8
5	0.0	0.3	0.4	0.5	0.5	0.7	0.7	0.8	0.8	0.8	0.8	0.9	1.0	1.0	1.1	1.1	1.0	1.0	0.9	0.8
6	0.2	0.3	0.4	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.8	0.9	0.9	0.9	0.9	0.9	1.0	1.0	1.0	0.8
7	-0.1	0.3	0.4	0.4	0.4	0.3	0.4	0.5	0.5	0.6	0.7	0.8	0.9	0.9	0.9	1.0	1.0	0.9	0.9	0.8
8	0.2	0.2	0.2	0.4	0.6	0.5	0.6	0.7	0.7	0.8	0.8	0.9	1.0	0.9	0.9	1.0	1.0	1.0	0.9	0.8
9	0.1	0.2	0.2	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1.0	0.9	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.8
10	0.0	0.1	0.2	0.3	0.5	0.5	0.6	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.9	0.8	0.9	0.9	0.9	0.8
11	-0.1	0.2	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.7	0.8	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.8	0.8
12	0.1	0.0	0.1	0.3	0.5	0.4	0.5	0.6	0.7	0.7	0.8	0.9	0.9	0.9	0.9	1.0	0.9	0.9	0.8	0.8
13	0.0	0.1	0.2	0.3	0.4	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.9	0.9	1.0	1.0	1.0	1.0	0.9	0.8
14	0.1	0.3	0.4	0.3	0.4	0.5	0.6	0.7	0.7	0.8	0.7	0.7	0.8	0.8	0.9	0.9	0.9	1.0	0.9	0.8
15	0.0	0.3	0.3	0.4	0.5	0.7	0.7	0.6	0.6	0.6	0.7	0.8	0.9	1.0	1.0	1.0	0.9	0.9	0.9	0.8
16	-0.1	0.1	0.3	0.4	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.8
17	0.1	0.2	0.2	0.5	0.3	0.4	0.6	0.6	0.6	0.7	0.7	0.8	0.8	0.9	1.0	0.9	0.9	0.9	0.9	0.8
18	0.2	0.4	0.3	0.5	0.6	0.5	0.6	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8
19	0.2	0.2	0.1	0.4	0.5	0.6	0.7	0.7	0.7	0.6	0.7	0.8	0.9	0.8	0.8	0.9	0.9	0.9	0.9	0.8
20	0.3	0.1	0.3	0.4	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.9	0.9	0.8	0.8	0.9	0.9	1.0	0.8
21	0.1	0.2	0.4	0.6	0.7	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.9	0.8
22	0.3	0.1	0.3	0.5	0.6	0.5	0.5	0.5	0.6	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.8	0.8
23	0.3	0.4	0.3	0.6	0.7	0.5	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.8	0.9	0.8
24	-0.1	0.3	0.3	0.4	0.6	0.5	0.6	0.6	0.7	0.8	0.7	0.7	0.6	0.7	0.8	0.8	0.8	0.8	0.9	0.8
25	-0.1	0.3	0.4	0.5	0.6	0.6	0.6	0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8



Source: J.P.Morgan Macro QDS

Dynamical Systems: Kalman Filtering

We can extend the linear regression models by allowing the beta coefficients to evolve slowly with time. Development of this idea led to the concept of a *Kalman filter*. Kalman filtering is often used in statistical trading (evolving beta) and volatility estimations (evolving volatility regimes). In a Kalman filter, the beta coefficient varies continuously over a certain range and is estimated recursively. If we discretize this range to a finite set of values, then we can derive a *Hidden Markov Model* (HMM) which will be discussed in the section on Classification methods.

The Kalman Filter (Kalman, 1960) is a technique that combines a series of observations, in the presence of uncertainty, in order to estimate and forecast parameters of an evolving system. The algorithm usually works in two-steps. In the first step, one comes with an estimate of the current state and an estimated error. The next observation (and its error) is then incorporated to obtain a new forecast (by properly weighting the previous estimate and error, and new observation and error).

The dynamic system is described by a State Space model (or, Dynamic Linear Models, DLMs), where there are 2 components:

1) State evolution:

The unobserved variables of interest are labelled as the state of the system, and we know how the state evolves over time via a linear expression with Gaussian noise:

$$x_t = Fx_{t-1} + w_t, \quad w_t \sim N(0, Q)$$

This is the first piece of information: Given the previous state x_{t-1} , we have some knowledge of the current state x_t , but there are uncertainties due to external random influence.

2) Measurement:

Although we cannot directly observe the state x_t , we can make some measurements z_t that are related to the state x_t , but again, our measurements come with Gaussian noise:

$$z_t = Hx_t + v_t, \quad v_t \sim N(0, R)$$

This is the second piece of information: We have the measurements that can infer the state, but there are uncertainties. The Kalman Filter combines the above two pieces of information and gives the optimal state variable as a Gaussian distribution³⁷ $N(x_{t|t}, P_{t|t})$, where the mean and covariance are

$$\begin{aligned} x_{t|t} &= x_{t|t-1} + K_t(z_t - Hx_{t|t-1}) \\ P_{t|t} &= P_{t|t-1} - KHP_{t|t-1} \end{aligned}$$

Here K is the Kalman gain and is given by $K = P_{t|t-1}H^T(HP_{t|t-1}H^T + R)^{-1}$. Despite the complicated expressions, the derivation of the formulae relies only on conditional Gaussian distributions:

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_X \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$$X|Z=z \sim N(\hat{\mu}, \hat{\Sigma})$$

³⁷ This utilizes the property in which the product of two Gaussian distributions is also a Gaussian distribution.

Where

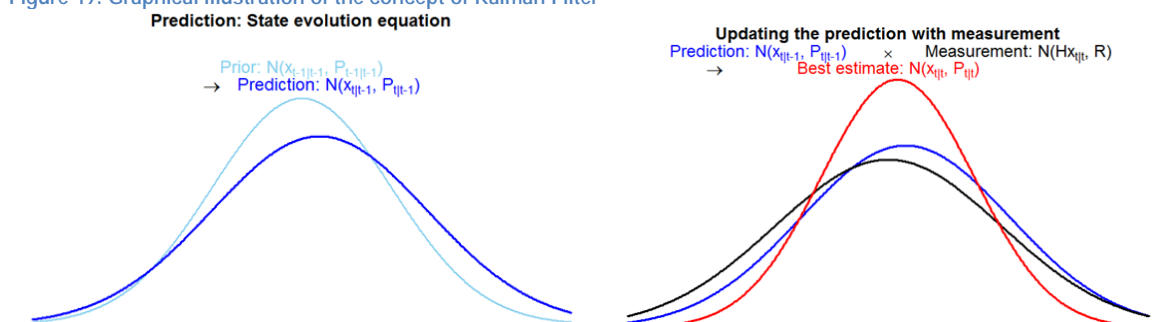
$$\hat{\mu} = \mu_X + \Sigma_{12}\Sigma_{22}^{-1}(z - \mu_Z)$$

$$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Figure 49 gives a graphical illustration of the workings of a Kalman Filter, which in essence is a Bayesian prediction conditioned on measurements.

Readers who want to have a step-by-step derivations of the formulae can refer to the tutorial by Faragher (2012), and "[How a Kalman filter works, in pictures](#)". The most famous application of the Kalman Filter was in spacecraft navigation system in the Apollo mission. Nowadays non-linear versions of the Kalman Filter are also applied in self-driving cars. In finance, Kalman Filters can be used to estimate trends and de-noise signals, to estimate unobserved economic activities, or to describe the dynamic relationships between assets and the market³⁸ (Petrakis et al (2009)).

Figure 49: Graphical illustration of the concept of Kalman Filter



Source: J.P.Morgan Macro QDS. Procedure: in the first step, we predict the state based on the evolution, and because of noise the variance increases. In the second step, we combine the predicted distribution and the measurement to update the state distribution, which is more precise than any single piece of information

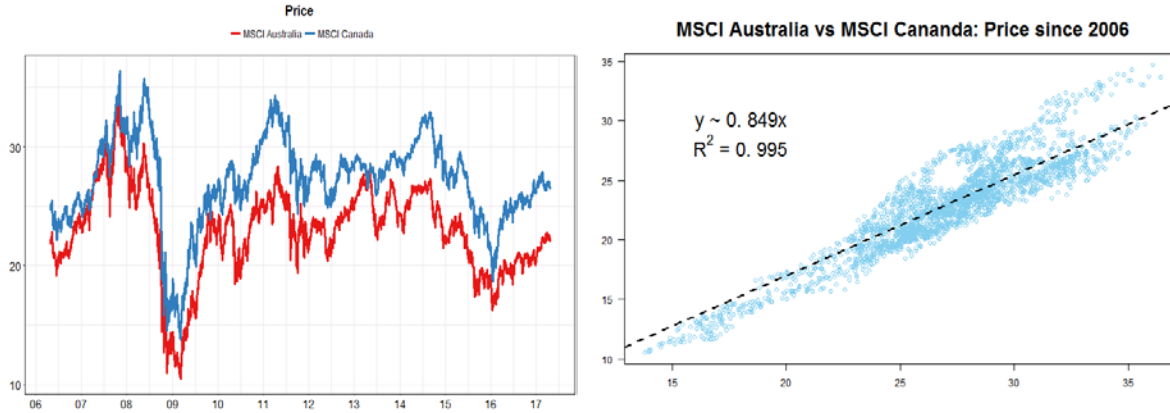
Financial Example:

To illustrate how one can apply Kalman Filter in finance, let us consider an example on pair trading. In general, one has to determine a prominent pair of assets before applying any pair trading strategies. Let us consider some widely known potential pairs of ETFs³⁹. We choose the iShares MSCI Australia ETF and Canada ETF as a pair, given strong correlation between the two.

³⁸ There is an example on the dynamic CAPM model in Petrakis et al (2009).

³⁹ Please refer to <http://etfdb.com/etf-trading-strategies/how-to-use-a-pairs-trading-strategy-with-etfs/>

Figure 50: MSCI Australia and MSCI Canada indices used in Kalman Filtering example



Source: J.P.Morgan Macro QDS

As such, let us determine the beta that relates the two asset prices S_t , in other words, the co-integration factor:

$$S_{t,AU} = \beta_t S_{t,CA} + v_t, \quad v_t \sim N(0, \sigma_v^2)$$

We may further assume that beta is time-dependent instead of being a constant. This has the advantage of capturing the dynamics of the system, and to make the residuals v_t to be stationary. For simplicity, let us assume that the evolution of beta simply follows a random walk:

$$\beta_t = \beta_{t-1} + w_t, \quad w_t \sim N(0, \sigma_w^2)$$

This is a dynamic linear regression, and in the State Space notations, β_t is the state variable. In State Space notations, our system is:

State equation:

$$\beta_t = F_t \beta_{t-1} + w_t, \quad w_t \sim N(0, Q_t)$$

where $F_t = 1$, $Q_t = \sigma_w^2$

Measurement equation:

$$z_t = H_t x_t + v_t, \quad v_t \sim N(0, R_t)$$

where $z_t = S_{t,AU}$, $H_t = S_{t,CA}$ and $R_t = \sigma_v^2$.

Hence, we can use the Kalman Filter to estimate the optimal values of beta when we update our observations (in this case, the prices of the ETFs). With some algebra, the Kalman gain in our univariate example is given by

$$K_t = \frac{S_{t,CA}}{S_{t,CA}^2 + \gamma^{-1}}$$

where

$$\gamma = \frac{\hat{P}_{t|t-1}}{\sigma_v^2}$$

is the Signal-to-Noise Ratio (SNR): The ratio of the state variance to the measurement error. If SNR is small, the measurement is noisy and not informative, hence a larger weight is put on the prior knowledge $\hat{\beta}_{t|t-1}$. If SNR is large, more weight is put on the observation. Substituting the variables into $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(z_t - H_t \hat{x}_{t|t-1})$, we have

$$\hat{\beta}_{t|t} = \hat{\beta}_{t|t-1} + \frac{S_{t,CA}}{S_{t,CA}^2 + \gamma^{-1}} (S_{t,AU} - \hat{\beta}_{t|t-1} S_{t,CA})$$

$$= \hat{\beta}_{t|t-1} \left(1 - \frac{S_{t,CA}^2}{S_{t,CA}^2 + \gamma^{-1}} \right) + \frac{S_{t,CA} S_{t,AU}}{S_{t,CA}^2 + \gamma^{-1}}$$

As a result, if SNR γ is very large, we basically just use the latest observation to estimate beta, without using any prior:

$$\hat{\beta}_{t|t} \approx \frac{S_{t,AU}}{S_{t,CA}}$$

If SNR γ is very small, we basically ignore the observation (because it is noisy), and we just use the information from the prior:

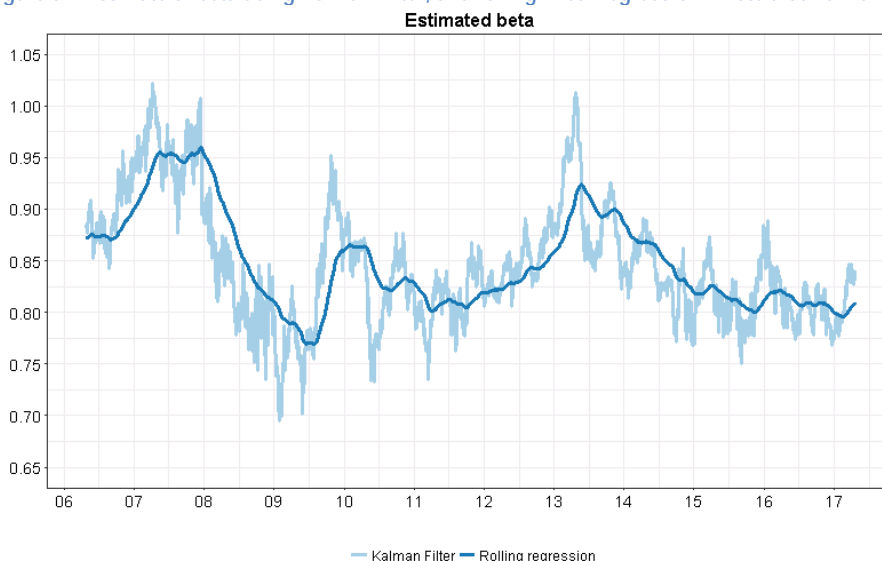
$$\hat{\beta}_{t|t} \approx \hat{\beta}_{t|t-1}$$

Instead of using a Kalman Filter, of course we can also estimate β using standard linear regression, without imposing any dynamics on it:

$$S_{t,AU} = \beta S_{t,CA} + v_t, \quad v_t \sim N(0, \sigma_v^2)$$

We compare the estimates of beta using the Kalman Filter, and the one using rolling linear regressions with a lookback window of 6 months. Note that Kalman Filter is more reactive. In fact, the Kalman Filter is closely related to exponential smoothing where it puts more weights on recent observations, and it can adjust the weights depending on how ‘noisy’ are the measurements⁴⁰.

Figure 51: Estimate of beta using Kalman Filter, and rolling linear regression – note that Kalman filter is more reactive to price movements



Source: J.P.Morgan Macro QDS

Our pairs trading signal simply depends on the residuals v_t , which is supposed to fluctuate around mean zero. At the end of each trading day, we use the closing price of the ETFs to update our estimate of the beta β_t , and then calculate the residuals:

$$v_t = S_{t,AU} - \beta_t S_{t,CA}$$

⁴⁰ Meinhold and Singpurwalla, Understanding the Kalman Filter. Available at <http://www-stat.wharton.upenn.edu/~steele/Resources/FTSResources/StateSpaceModels/KFExposition/MeinSing83.pdf>

We also record the uncertainty of the residuals σ_t , so that we can use it to determine whether the magnitude of the residual is large enough to trigger our strategy (otherwise no position is taken):

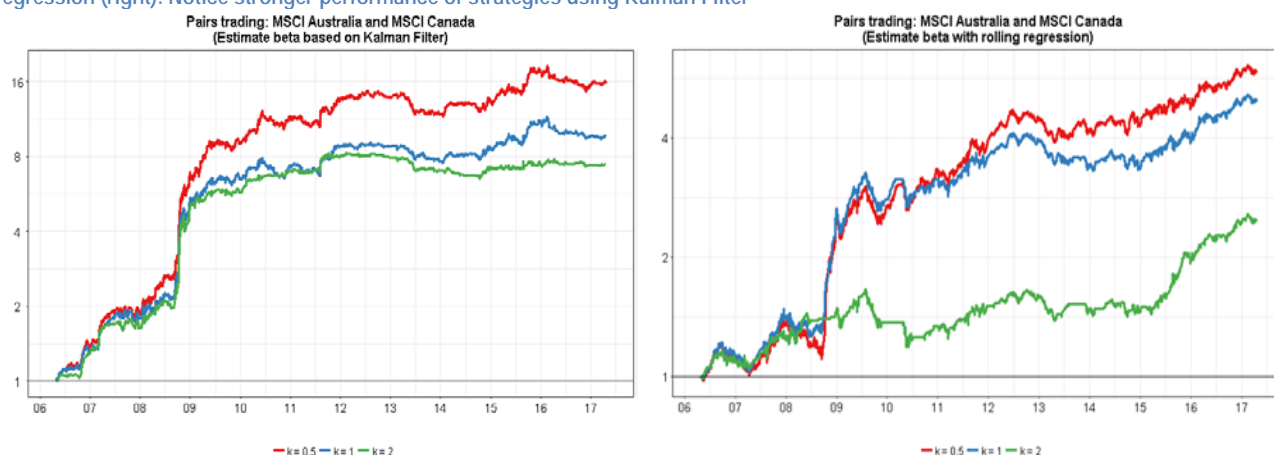
If $v_t \geq k\sigma_t$, we long β_t unit of MSCI Canada and short one unit of MSCI Australia

If $v_t \leq -k\sigma_t$, we short β_t unit of MSCI Canada and long one unit of MSCI Australia

As an example, we take $k = 0.5, 1, 2$. This scaling parameter can actually be tuned in a more optimal way by modelling the mean-reversion process, but we will not investigate further here.

The performance of the strategy since 2006 is given in Figure 52. Next to it, we show the performance of the same strategy when using a simple rolling regression beta

Figure 52: MSCI Australia vs. MSCI Canada pair trading strategy. Beta is estimated from the Kalman Filter (left), and from rolling 6 months regression (right). Notice stronger performance of strategies using Kalman Filter



Source: J.P.Morgan Macro QDS, Bloomberg

To compare the 2 strategies, we choose $k = 0.5$. We see that using a Kalman Filter that has a faster response helped deliver higher returns during time of market dislocations in 2008/2009.

Figure 53: Comparing the pairs trading strategy - Kalman Filter vs. rolling regressions

	CAGR (%)	Vol (%)	Sharpe	Max DD (%)
Since 2006				
Kalman Filter	28.6	18.1	1.58	21.1
Regression	17.4	17.9	0.97	22.5
Since 2009				
Kalman Filter	10.9	14.5	0.75	21.1
Regression	10.4	15.3	0.68	19.9

Source: J.P. Morgan Macro QDS, Bloomberg

Kalman filter gives good results (closed form solution) when applied to linear systems with Gaussian noise. For non linear system or systems with non-Gaussian noise, a numerical technique called Particle Filtering is used. For technical details and an example of a Particle Filtering application, see the [Appendix](#).

Extreme Gradient Boosting

The term ‘boosting’ refers to a technique of iteratively combining weak ‘learners’ (i.e. algorithms with weak predictive power) to form an algorithm with strong predictive power⁴¹. Boosting starts with a weak learner (typically, a regression tree algorithm, see below) and records the error between the learner’s predictions and the actual output. At each stage of the iteration, it uses the error to improve the weak learner from the previous iteration step. If the error term is calculated as a negative gradient of a loss function, the method is called ‘gradient boosting’. Extreme gradient boosting (or XGBoost as it is called) refers to an optimized implementation by Chen and Guestrin⁴² that has become a popular supervised learning algorithm in use for financial time series data. Theoretical details of this method are likely of little interest for most readers. In practice, one uses open source algorithms (e.g. in R) to implement this method on problems that would have otherwise been addressed by other regression methods.

XGBoost Algorithm

A regression tree is similar to a decision tree, except that at each leaf, we get a continuous score instead of a class label. If a regression tree has T leaves and accepts a vector of size m as input, then we can define a function $q: \mathbb{R}^m \rightarrow \{1, \dots, T\}$ that maps an input to a leaf index. If we denote the score at a leaf by the function w , then we can define the k -th tree (within the ensemble of trees considered) as a function $f_k(\underline{x}) = w_{q(\underline{x})}$, where $w \in \mathbb{R}^T$. For a training set of size n with samples given by $(\underline{x}_i, y_i), \underline{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$, a tree ensemble model will use K additive functions to predict the output as follows:

$$\hat{y}_i = \phi(\underline{x}_i) = \sum_{k=1}^K f_k(\underline{x}_i).$$

To learn the set of functions in the model, the regularized objective is defined as follows:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k),$$

$$\text{where } \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2.$$

The tree ensemble model is optimized in an additive manner. If $\hat{y}_i^{(t)}$ is the prediction for the i -th training example at the t -th stage of boosting iteration, then we seek to augment our ensemble collection of trees by a function f_t that minimizes the following objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\underline{x}_i)) + \Omega(f_t).$$

The objective is approximated by a second-order Taylor expansion and then optimized. For calculation steps, we refer the reader to Chen and Guestrin (2016) and the expository material on xgboost at [link](#). To prevent overfitting, xgboost uses shrinkage (to allow future trees to influence the final model) and feature sub-sampling (as in random forest). An option for row sub-sampling is also provided.

Example of macro trading strategy using XGBoost

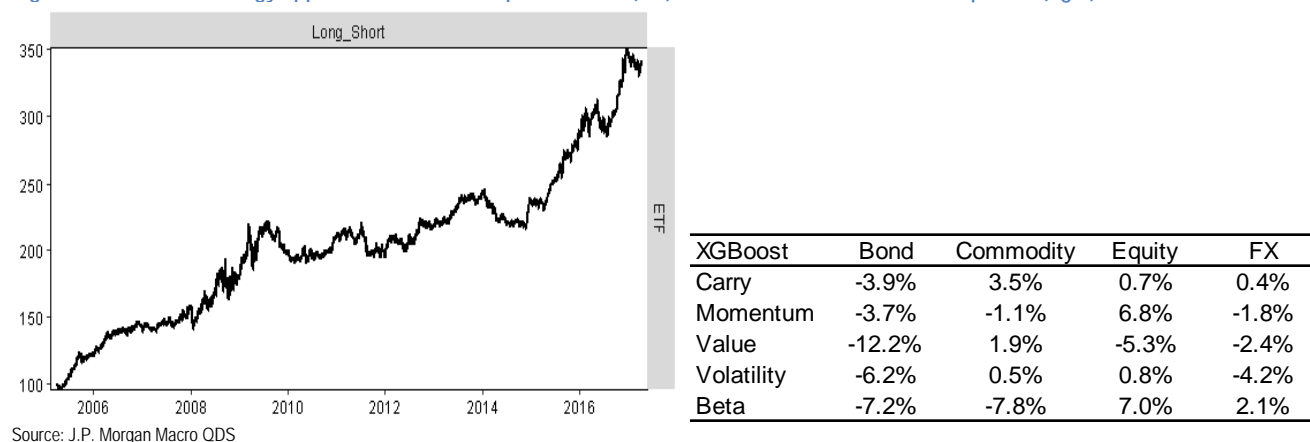
⁴¹ For bootstrap techniques, see Efron and Tibshirani (1993), Hall (1992), Breiman and Spector (1992), Breiman (1992), Shao (1996), Zhang (1993) and Kohavi (1995). For bagging, see Breiman (1996a); for stacking, see Wolpert (1992), Breiman (1996b) and Leblanc and Tibshirani (1996); for bumping, see Tibshirani and Knight (1999). For academic discussions and disputes around boosting, see Schapire (1990, 2002), Freund (1995), Schapire and Singer (1999), Freund and Schapire (1996b), Breiman (1998, 1999), Schapire et al (1998), Meir and Ratsch (2003), Friedman et al (2000), Friedman (1999, 2001), Mason et al (2000), Jiang (2004), Lugosi and Vayatis (2004), Zhang and Yu (2005), Bartlett and Traskin (2007), Mease and Wyner (2008), Friedman et al (2008a), Buhlman and Hothorn (2007). AdaBoost is covered in Freund and Schapire (1996a).

⁴² Chen, T and Guestrin, C (2016), “XGBoost: a scalable tree boosting system”, Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pg. 785-794. Conference held at San Francisco, CA from Aug13-17, 2016.

We construct a long-short strategy trading nine US sector ETFs: financials, energy, utilities, healthcare, industrials, technology, consumer staples, consumer discretionary and materials. We use XGBoost algorithm to predict next day returns based on 8 macro factors: Oil, Gold, Dollar, Bonds; economic surprise index (CESIUSD), 10Y-2Y spread, IG credit (CDX HG) and HY credit spreads (CDX HY).

After obtaining prediction returns for US sectors (based on previous day's macro factors), we rank sectors by expected return, and go long top 3 and short bottom 3 sectors. We used the open-source implementation of Xgboost available in R through the 'xgboost' package. For optimizing the parameters and hyper-parameters, we used 5-fold cross-validation, 30 boosting iterations and allowed decision trees to have a maximum depth of 7. The strategy used a rolling window of 252 days and rebalanced daily at the market close. This yielded an annualized return of 10.97% and an annualized volatility of 12.29%, for a Sharpe ratio of 0.89. Correlation of the strategy to the S&P 500 was ~7%. Correlation of the strategy to equity long-short styles, as well as other cross-asset risk premia were also low (see the right figure below).

Figure 54: XGBoost strategy applied to US sectors – performance (left), correlation to cross-asset risk premia (right)



Sample R Code

```
## To choose best parameters, iterate over segment below
param = list(
  objective = "reg:linear",
  max.depth = sample(c(1:7),1),
  num_class = 1
)

xgtrain = data.matrix(df[1:n,X_var])
num_rounds = 150
model2 = xgb.cv(params = param, xgtrain,
  label = data.matrix(df[1:n,'y_var']),
  num_rounds,nfold=5,
  verbose = FALSE,
  early_stopping_rounds=25,
  maximize=FALSE)

## Once optimal parameters are chosen and stored in best_param,
## model is run as follows.
model <- xgboost(xgtrain,label = data.matrix(df[1:n,'y_var']),
  params=best_param,
  nrounds=nround, nthread=10,verbose = FALSE,)
preds = predict(model, data.matrix(df[(n+1),X_var]))
```

Supervised Learning: Classifications

Classification methods of supervised learning⁴³ have a goal of classifying observations into distinct categories. For instance, we may want the output of a model to be a binary action such as 'buy market' or 'sell market' based on a number of macro, fundamental, or market input variables. Similarly, we may want to classify asset volatility regimes as "high", "medium", and "low volatility". In this section, we cover the following classification algorithms: logistic regression, Support Vector Machines (SVM), decision trees and random forests, and Hidden Markov Models.

Logistic regression is a classification algorithm that produces output as a binary decision, e.g. "buy" or "sell". It is the simplest adaptation of linear regression to a specific case when the output variable is binary (0 or 1). **Support Vector Machine** is one of the most commonly used off-the shelf classification algorithms. It separates data into classes via abstract mathematical processes (mapping a dataset into a higher dimensional space, and then fitting a linear classifier). **Decision trees** try to find the optimal rule to forecast an outcome based on a sequence of simple decision steps. **Random Forest** is a classification Machine Learning method that is based on 'decision trees'. Random Forests are averaging simple decision tree models (e.g. calibrated over different historical episodes) and they often yield better and more reliable forecasts as compared to decision trees. The fourth class of algorithms is **Hidden Markov Models (HMM)**. HMMs originate in signal processing theory, and are used in finance to classify asset regimes.

Logistic Regression

Logistic Regression: Logistic regression (or logit in econometrics) is used to forecast the probability of an event given historical sample data⁴⁴. For example, we may want to forecast the direction of next-day price movement given the observation of other market returns or value of quant signals today. By mapping the probability to either 0 or 1, logistic regression can also be used for classification problems, where we have to choose to either buy (mapped to 0) or sell (mapped to 1).

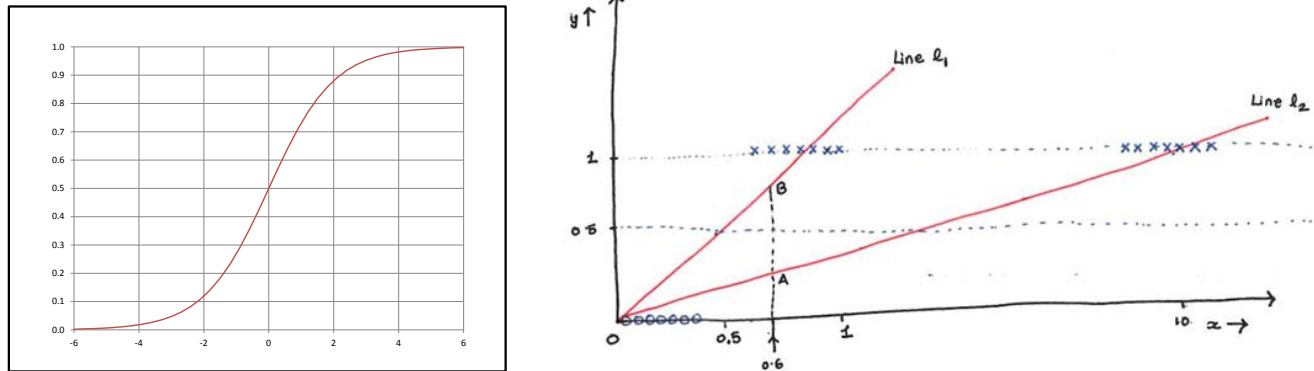
Logistic regression is derived via a simple change to ordinary linear regression.⁴⁵ We first form a linear combination of the input variables (as in conventional regression), and then apply a function that maps this number to a value between 0 and 1. The mapping function is called the logistic function, and has the simple shape shown below (left). An alternative to the logistic regression would have to simply 'quantize' the output of a linear regressor to either 0 or 1, i.e. if the value was less than 0.5, we declare the output to be 0; and if its greater than 0.5, we declare it to be 1. It turns out that such an approach is very sensitive to outliers. As shown in the figure below (right), the outliers cause the line to have an extremely low slope, which in turn leads to many points in the set being classified as 0, instead of being classified correctly as 1.

⁴³ For algorithms related to classification, see Duda et al (2000) for perceptrons, see Ripley (1996). Density estimation is covered in Scott (1992).

⁴⁴ For information on the generalized linear model, see Gelman and Hill (2007), Gelman, Jakulin et al (2008), Ormerod and Wand (2012), Cseke and Heskes (2011), Imai and van Dyk (2005), Bradlow and Fader (2001), Jackman (2001), Martin and Quinn (2002), Agresti (2002), Dobra, Tebaldi and West (2003). Hierarchical linear models are covered in Gelman and Hill (2007). Analysis of Variance or ANOVA is covered in Gelman (2005), Nelder (1994), Hodges and Sargent (2001).

⁴⁵ Statisticians place logistic regression and ordinary linear regression within a broader category of Generalized Linear Models (GLM). In all GLMs, one first forms a linear combination of the inputs and then maps this to the output via a link function. For the ordinary linear regression, the link function is an identity function, i.e. does nothing, but maps a number to itself. For logistic regression, the link function is a logistic function. There are other models in the GLM family, such as probit and Poisson regression models.

Figure 55: Logistic Function (left), linear regression fails as a classifier in the presence of outliers (Right)



Source: J.P.Morgan Macro QDS

Most financial analysts will use logistic regression algorithms provided by libraries like R, and there is little need to understand the derivation or formulas behind the method. Nevertheless, we provide some theoretical considerations in the mathematical box below.

Mathematical Model for Logistic Regression

Consider a binary classification problem, where $y \in \{0,1\}$. Elements of our training set of size m , can be referenced as $(\underline{x}^{(i)}, y^{(i)})$, $\forall i \in \{1, \dots, m\}$. In logistic regression, we use the sigmoid (or logit) function $g(z) = \frac{1}{1+e^{-z}}$ to define $h_{\underline{\theta}}(\underline{x}) = \frac{1}{1+e^{-\underline{\theta}^T \underline{x}}}$.

Here, we use the probabilistic approach to model $P(y = 1 | \underline{x}; \underline{\theta}) = h_{\underline{\theta}}(\underline{x}) \in [0,1]$, and $P(y = 0 | \underline{x}; \underline{\theta}) = 1 - h_{\underline{\theta}}(\underline{x})$. In other words, we model as $Y \sim \text{Binomial}(m, h_{\underline{\theta}}(\underline{x}))$.

This yields a likelihood function of $L(\underline{\theta}) = p(y|X; \underline{\theta}) = \prod_{i=1}^m (h_{\underline{\theta}}(\underline{x}^{(i)}))^{y^{(i)}} (1 - h_{\underline{\theta}}(\underline{x}^{(i)}))^{1-y^{(i)}}$. Optimizing the log-likelihood $l(\underline{\theta}) = \log L(\underline{\theta})$ leads to the LMS rule

$\underline{\theta}_j \leftarrow \underline{\theta}_j + \alpha (y^{(i)} - h_{\underline{\theta}}(\underline{x}^{(i)})) \underline{x}_j^{(i)}$. In practice, implementations try to solve for $l'(\underline{\theta}) = 0$ using the Newton-Raphson method as

$$\underline{\theta} \leftarrow \underline{\theta} - H^{-1} \nabla_{\underline{\theta}} l(\underline{\theta}),$$

where H is the Hessian matrix defined as

$$H_{ij} = \frac{\partial^2 l(\underline{\theta})}{\partial \theta_i \partial \theta_j}.$$

This implementation of logistic regression involving the Hessian matrix is called Fisher Scoring.

Other points to note are:

- Extension to $K \neq 2$ is available through soft-max regression.
- Probit regression is obtained by modifying the modeling assumption to $Y \sim \text{Binomial}(m, \Phi(\underline{\theta}^T \underline{x}))$, where $\Phi(\cdot)$ is the cumulative distribution function for the $N(0,1)$ distribution.
- In practice, it is common to use L1 or L2 regularization with logistic regression.

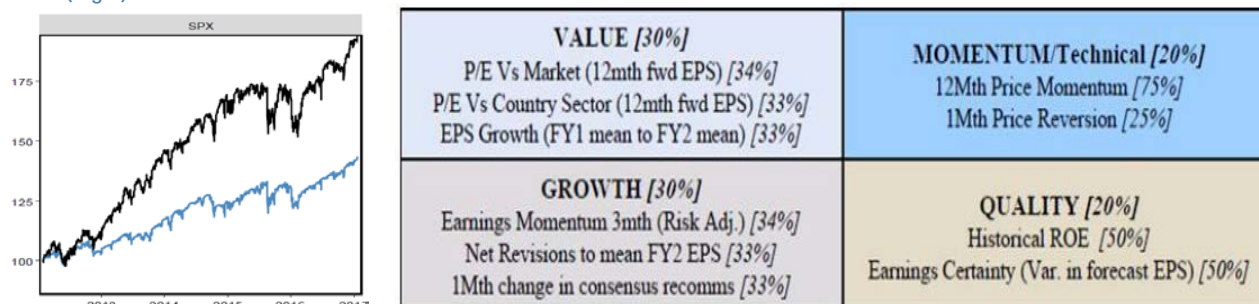
Logistic Regression: Financial Example

We illustrate an application of logistic regression by improving the performance of call overwriting (OW) strategies. This analysis was first reported [here](#) by our team in Jan 2017. Over the past 5 years, call overwriting strategies have underperformed underlying equity returns for both the S&P 500 (SPX) and EuroStroxx50 (SX5E) index. While the underperformance was largely due to lower equity exposure, the question clients are often asking us if it is possible to make

a quantitative selection of stocks that will enable one to improve performance of overwriting strategy. The figure below shows the PnL from selling a 1M ATM call on a monthly rolling basis with the underlying equity index.

The PnL from call overwriting is quasi-linear in the underlying equity returns. Stocks that perform very strongly will outperform an overwriting strategy (on the same stock). In other instances, overwriting tends to outperform a long only position. The relationship between stock performance and the binary outcome (stock vs overwrite outperformance) is nonlinear. This will make the use of logistic regression suitable to determine which stocks should be overwritten and which ones should be held long only. We can start with modeling stock returns with a multi-factor model. Just like the Fama-French model uses 3 factors in a linear fashion to estimate returns, we use a 10-factor model split into four buckets as shown in the figure. Once we forecast the stock performance, we can make a prediction of the probability of outperformance or underperformance of an overwrite strategy (probability will be in a 0 to 1 range)

Figure 56: Performance of call OW strategies vs underlying equity indices (left), 10-factor model along with weights as used in JPM's Q-Score Model (Right)



Source: J.P.Morgan Macro QDS

We considered all large-cap stocks in Europe which had volatility data available for at least 5 years and compare the PnL from selling 1M 103% options and rolling monthly with the underlying equity index returns. The output variable y was assigned the value 1 if the call OW strategy outperformed the equity index and 0 otherwise.

For fitting the logistic regression model, we used around 11,000 data points from Jan 2012 to Dec 2015 as our training set and the period from Jan 2016 to Dec 2016 as our test set. To reduce variance of the fit further, we adopted bootstrapping, i.e. for 1000 times, we sampled with replacement 25% of the training set and fit the logistic model to that. For each of the 1000 classifiers, the result on the test set was calculated and averaged out to produce the final decision.

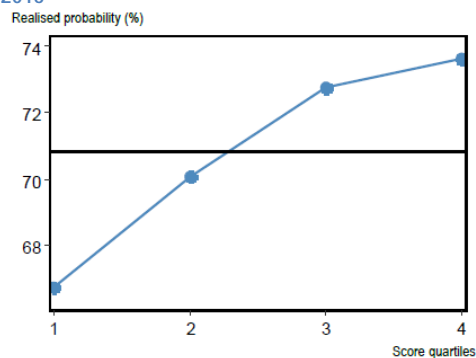
The analysis yielded the prediction coefficients shown in the table below left. To interpret the coefficient values in the second column above, we can follow the analogous intuition from linear regression. A larger positive value implies that the factor has a positive effect on the probability of call overwriting outperformance. From the results below, we conclude that we should choose stocks with high 12M price momentum and high earnings certainty (i.e., low earnings uncertainty). We should avoid stocks with low 3M realized volatility.

We test our model on out-of-sample examples from Jan 2016 to Dec 2016. We find that the actual probability of call OW outperformance increases monotonically with the call OW score (from logistic regression). For figure below (right), we split the results from logistic regression (which are probabilities ranging from 0 to 1) into four quartiles (PnL of the strategy also shows that 4th quartile clearly outperformed names in the first quartile).

Figure 57: Logistic Regression: Coefficients and Z-value (left), Model score from logistic regression (into quartiles) vs. realized probability of call OW outperforming the equity index; Out-of-sample results covering calendar year 2016

Factor	Coef Estimate	Z-value
3M Realised Volatility	-0.36	-6.1
Historical ROE	-0.06	-1.5
1M Price Momentum	-0.05	-1.2
1Y Earnings Yield vs. Country Sector	-0.08	-1.2
EPS Growth	-0.05	-1.0
1Y Earnings Yield	-0.05	-0.8
Earnings Momentum 3M	-0.03	-0.5
Net revisions to Mean FY2 EPS	-0.02	-0.4
1M Change in Consensus Recs	0.00	-0.1
Earnings Certainty	0.01	0.2
12M Price Momentum	0.11	2.3

Source: J.P.Morgan Macro QDS



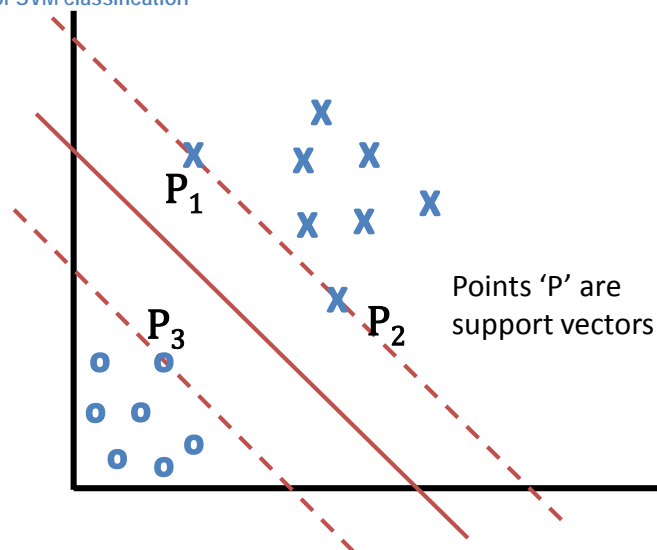
Support Vector Machines

Support Vector Machines (SVM): Support vector machines are one of the most popular classification algorithms⁴⁶. Their popularity stems from their ease of use and calibration. SVM are based on fitting a linear classifier to the dataset after enhancing it with new derived features (like squares/cubes of input columns and interaction terms). We provide an informal explanation of the method below. Most practitioners will use ready made codes (e.g. in R/Python) to implement the method and will not care about theoretical considerations.

Consider a scenario where we seek to relate the direction of an asset's price movement to a set of input features via a linear function. We would take a linear combination of the inputs, and if this was above a threshold, we would predict 'buy'; else we would predict 'sell'. One would be satisfied with such a classifier if historical examples of asset price increases were accompanied by *very* low values for the linear combination, and historical examples of asset price decreases were accompanied by *very* large values for the linear combination. We are seeking a linear combination such that its value on the two classes (buy and sell) is as different and as separated as possible. Such a linear combination is called a maximum margin classifier, and is an example of Support Vector Machine.

This is illustrated in the figure below where we try to predict sell signal ("O") and buy signal ("X") on a set of two input variables (shown on vertical and horizontal axis). Note that the solid line "l" is right in middle of the two sets; in formal terms, we have maximized the margin (or gap) between the classes (this is why the algorithm is also called a 'maximum margin classifier').

Figure 58: Simplified illustration of SVM classification



Source: J.P.Morgan Macro QDS

The line "l" can be determined just from the border points of the two sets. Points P1 and P2 create a border line for the crosses and we drew a parallel line through P3 as well. These border points P1, P2 and P3 which alone determine the optimal separating line "l" are called 'support vectors'. This was a very simple illustration of a linear classification when the input variables can be visualized on a chart. Mathematics and visualization become significantly more complex when we are dealing with a large number of input variables.

One can improve performance of a linear classifier by providing it more features; instead of just variable x , we could give additional inputs like x^2 and x^3 . Computer scientists found an efficient way – called the 'kernel trick' – of mapping the data

⁴⁶ Support Vector Machines are described in Vapnik (1996), Wahba et al (2000), Evgeniou et al (2000), and Burges (1998).

to higher-dimensional spaces (mathematical box below). The second idea to improve is to allow for a small number of errors to account for cases where the data is not linearly separable. This is accomplished by adding a regularization term which penalizes the optimizer for each misclassification. Putting these two ideas together, leads to Support Vector Machines.

Kernel Trick Within Support Vector Machine

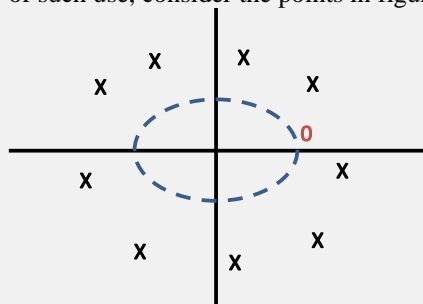
We describe the kernel trick in this section.

Given a set of input *attributes*, say $\underline{x} = [x_1, x_2, x_3]$, one often finds it useful to map these to a set of *features*, say

$$\phi(\underline{x}) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ \vdots \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}.$$

Such a mapping into higher dimensions can often improve the performance of algorithms for function approximation. In many algorithms, one can reduce the functional dependence on input attributes to just inner products, i.e. the optimization function depends on the training example $\underline{x}^{(i)}$ only through the form of $\langle \underline{x}^{(i)}, \underline{c} \rangle$, where \underline{c} is another vector. In this case, the mapping to the higher-dimensional feature space creates the inner product $\langle \phi(\underline{x}^{(i)}), \phi(\underline{c}) \rangle$. The kernel trick refers to the observation that the inner product $k(\underline{x}^{(i)}, \underline{c}) = \langle \phi(\underline{x}^{(i)}), \phi(\underline{c}) \rangle$ can often be computed very efficiently.

As an example, consider the mapping ϕ as defined above. A naïve computation of $\langle \phi(\underline{x}^{(i)}), \phi(\underline{c}) \rangle$ would be an $O(n^2)$ operation. But noting that $\langle \phi(\underline{x}^{(i)}), \phi(\underline{c}) \rangle = (\underline{x}^{(i)T} \underline{c})^2 = k(\underline{x}^{(i)}, \underline{c})$, we can compute the value in just $O(n)$ time. This implies that an algorithm designed for use in a low-dimensional space of attributes can function even in a high-dimensional space of features, merely by replacing all inner products of the type $\langle \underline{x}^{(i)}, \underline{c} \rangle$ by $k(\underline{x}^{(i)}, \underline{c})$. As an example of such use, consider the points in figure.



These points are not linearly separable in two dimensions, but mapping them to higher dimensions, say three, will make them so. In practice, this kernel trick is widely used to improve the performance of algorithms as diverse as the perceptron to support vector machines to Principal Component Analysis. The kernel trick enables linear learning algorithms to learn a non-linear boundary without specifying an explicit mapping. One need not stop at a finite-dimensional mapping. The

commonly used kernel function is the Gaussian Radial Basis Function, $k(\underline{x} - \underline{y}) = e^{-\frac{\|\underline{x} - \underline{y}\|^2}{2}}$, which corresponds to an infinite-dimensional mapping function ϕ and yet is computable in just $O(n)$ time.

To decide what forms are acceptable for the kernel function $k(\underline{x}, \underline{y})$, one defers to Mercer's theorem. Simply put, it states that for a function $k: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ to be a valid kernel function, it is necessary and sufficient that for any set of vectors $\{\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(n)}\}$, the corresponding kernel matrix $K_{ij} = k(\underline{x}^{(i)}, \underline{x}^{(j)})$ is symmetric and positive definite.

SVM Financial Example: Trading Currency volatility

We illustrate the use of SVM Classifiers in optimizing the PnL from an FX volatility strategy. This analysis was presented in a report "[Machine Learning approach to FX option trading](#)" in Jan 2017. The PnL from a rolling long position in 1M ATM EURUSD options was tracked and the output variable y was defined via 3 classes: a class for "vol sell", when PnL <

-20 bps, a class for “vol buy”, when PnL > 20 bps and a class for “neutral”, when neither of the above applied. 377 different indicators were chosen which included market prices on FX, rates, currencies, commodities, as well as macroeconomic indicators like the Economic Activity Surprise Index (from JPM) and IMM positioning. The full list is given in the table below.

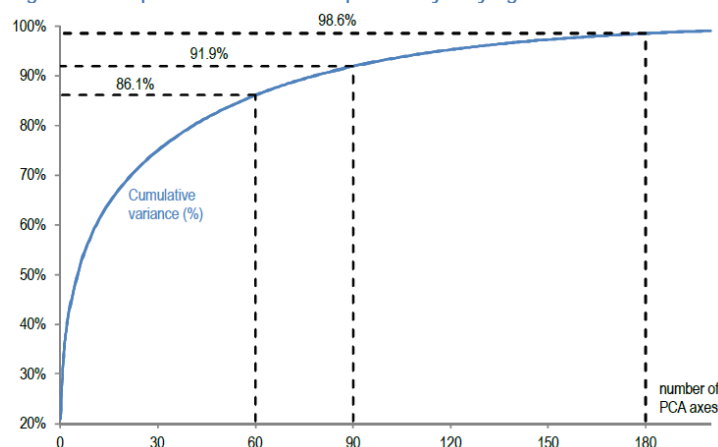
Figure 59: List of input variables for analysis of FX volatility strategy

Market data type	Market data	Level	1 week change	1M change	Count
FX realised vols	2M realised vols in USD vs G10, MXN, BRL, ZAR, TRY, NR and KRW	X	X	X	45
FX ATM vols	1M, 3M and 1Y ATM in same pairs	X	X	X	135
FX skews	3M 25D RRs	X	X	X	45
FX spots	FX Spots in 15 G10 and EM USD pairs		X	X	30
Depo rates / Basis	3M FX Forward drops		X	X	30
Interest Rates	10Y Gov yields: US, Japan, UK, Germany, France, Italy, Spain, Australia		X	X	16
Equity Indices	S&P, Nikkei, FTSE, E-Stoxx, ASX, Mexbol, Bovespa, KOSPI, Hang Seng		X	X	18
Commodities	Gold and Brent spot		X	X	4
Credit spreads	CDX IG and HY, iTraxx spread indices		X	X	6
EASI indices	Global, US, CAD, EU, UK, CHF, NOK, SEK, Japan, AU, NZ, China		X	X	24
IMM positions	USD, EUR, JPY, GBP, CHF, CAD, AUD, NZD, MXN, RUB, Gold		X	X	24

Source: J.P.Morgan Macro QDS

For evaluating different supervised learning models, we considered daily samples over the past 10 years, yielding 2609 samples. To avoid overfitting, it is necessary to reduce the dimensionality of the input feature space. Using Principal Component Analysis, it was determined that just 60 features could explain 86% of the variance. For our analysis, we considered separately the top 60, 90 and 180 features located through PCA.

Figure 60: Proportion of variance explained by varying the number of PCA factors



Source: J.P.Morgan Macro QDS

To validate the models, data was split into a training set spanning the first eight years and a test set spanning the latter two. 10-fold cross-validation was used to train the models. Supervised learning models were applied on raw data (just the 377 input features as such), normalized data (where inputs were standardized to mean zero and variance 1) and on 180/90/60 features extracted through PCA. Models tested included the K-nearest neighbor, Support Vector Classifier with linear

kernel, Support Vector Classifier with polynomial kernel, ridge regression, linear discriminant analysis, logistic regression, Classification and Regression Tree (CART), Passive-Aggressive Classifier ([PAC](#); where parameters are updated on wrong decisions and left unchanged on right decisions), and Stochastic Gradient Descent regression (ordinary regression, where the parameters are estimated using a stochastic version of gradient descent, that chooses a different sub-sample of the data in each iteration).

Out-of-sample results are tabulated below. **They showed that support vector classifiers – both linear kernel and polynomial kernel of degree 3 – and the K-nearest neighbor with K=5 performed best.**

Figure 61: Prediction accuracies of supervised learning models on out-of-sample data from 2014 to 2016

Algorithm	Raw Data	Normalised	PCA 180	PCA 90	PCA 60
kNN	69.0%	83.9%	83.7%	83.3%	84.1%
SVC (polynomial kernel - degree 3)	59.5%	74.1%	82.4%	84.1%	84.1%
SVC (linear kernel)	62.1%	80.8%	83.3%	83.3%	82.4%
Ridge Regression	81.0%	80.8%	73.9%	68.4%	69.3%
Gaussian NB	67.2%	68.4%	69.2%	69.5%	69.3%
Linear Discriminant Analysis	81.2%	81.2%	73.4%	68.6%	68.2%
Logistic Regression	79.3%	79.9%	74.1%	68.6%	68.0%
CART Decision Tree	75.7%	76.1%	61.9%	68.6%	67.6%
PAC Regression	48.5%	76.6%	63.2%	65.7%	60.3%
SGD Regression	41.2%	76.4%	69.3%	64.9%	57.5%

Source: J.P.Morgan Macro QDS

We can further investigate the output from the support vector classifier. One can define precision as the percentage of times the model was correct, given its output. For example, the model proposed that we go “long vol” 34 times. Out of these 34 occasions, 30 occasions yielded success (i.e. PnL > 20 bps) and 4 occasions yielded a neutral result (i.e. PnL lay between -20 bps and +20 bps). This implies the precision of the model when predicting long is 88.2%. Similarly, the precision of the model when predicting short is 85.7%.

One can define recall as the accuracy of the model, given the actual value. For example, there were 56 occasions when “long vol” would have yielded a positive PnL. The model caught 30 of those, to get a recall of 53.6%. Similarly, the model had a recall of 67.6% for short vol values. Results are tabulated below.

Figure 62: Performance of Support Vector classifier on test set

		Predicted			Recall
		Long	Neutral	Short	
Actual	Long	30	26	0	53.6%
	Neutral	4	304	16	93.8%
	Short	0	46	96	67.6%
Precision		88.2%	80.9%	85.7%	

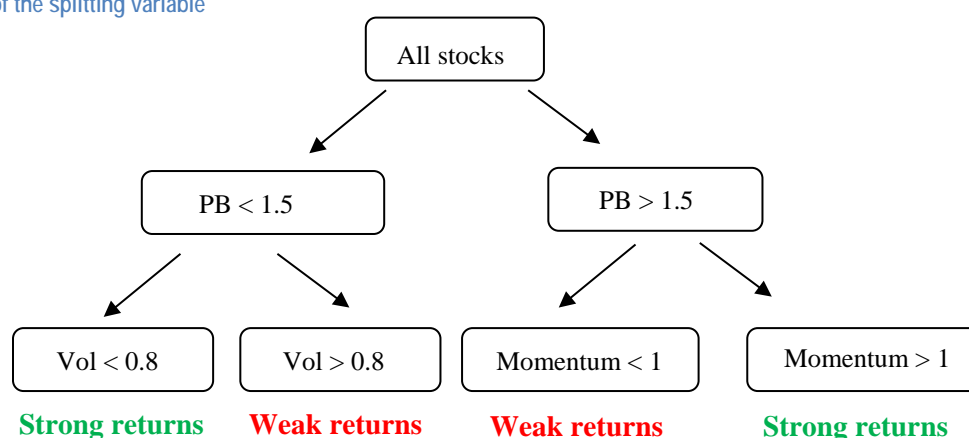
Source: J.P.Morgan Macro QDS

Decision Trees and Random Forests

Decision tree models are essentially flow charts used commonly in business management and financial analysis. To arrive at a "decision", the analyst asks a series of questions. At each step, a decision tree branches based on the answer. The questions are asked in order of importance; we ask a question on the most important factor first, and then subsequently refine the question. In the end, the combination of answers to different questions enables one to make a decision.

Decision trees are one of the simplest non-linear models that can be used to classify outcomes. For example, we can classify whether a stock has strong or weak returns by looking at a few factors such as momentum, volatility, and price-to-book ratio. An example of such a decision tree is shown in the figure below. Decision trees are able to consider interactions between variables, e.g. stock returns are strong if momentum is low, volatility is high, and price-to-book is high.

Figure 63: Example of decision tree to classify whether future stock returns are good or poor. In each node, we split the data into two subsets depending on the value of the splitting variable



Source: J.P.Morgan Macro QDS, FactSet

Random forests improve upon this idea by averaging the output of many decision trees⁴⁷. Each decision tree is then fit on a small subset of training examples or is constrained to use only a small subset of input features. Averaging the output of these trees reduces variance of the overall estimator. Alongside support vector machines, random forests were the most popular amongst Machine Learning methods until the recent invention of XGBoost and rise of Deep Learning.

To fit a decision tree, the algorithm usually looks for the best variable and the best splitting value among all possibilities, so that a particular loss function is minimized. The loss function can be defined as the impurities in the child nodes, which are measured by Gini index or entropy. Criteria can be used to ensure the tree is interpretable and prevent overfitting, e.g.

- **Max depth:** deciding a maximum depth of the tree
- **Node size:** at least N observations in each node

One can also build a large tree with many branches, and then prune the tree by combining subtrees with the lowest trade-off in the goodness of fit. Apart from classifications, decision trees can also be used for regressions that predict a continuous outcome. In that case, the model is simply a piecewise constant "surface" depending on the thresholds of the explanatory variables (Hastie et al 2013).

⁴⁷ For detailed discussion of random forests and related techniques, see Breiman (2001), Ho (1995), Kleinberg (1990, 1996), Amit and Geman (1997), Breiman (1996a), Dietterich (2000b), Friedman and Hall (2007). For an overview of ensemble methods, see Dietterich (2000a) and Kittler et al (1998). Graphical models are covered in Whittaker (1990), Lauritzen (1996), Cox and Wermuth (1996), Edwards (2000), Pearl (2000), Anderson (2003), Jordan (2004), Koller and Friedman (2007), Wasserman (2004), Bishop (2006) and Ripley (1996)

Although decision trees are intuitive and easy to understand, they tend to be very noisy. Small changes in the data can lead to different splits and completely different models. The instability of the tree renders it impractical as a prediction model by itself. Nevertheless, decision trees are useful to visualize interactions between variables

Random Forests

As we see above, a single decision tree is unstable and in general overfits the data, i.e. it can capture the structure of the in-sample data very well, but it tends to work poorly out-of-sample. In the jargon of statistics, decisions trees have low bias (as it can fit the data well) but high variances (the predictions are noisy).

Bagging (Breiman, 1996) is a useful technique that helps to reduce the variances of the predictions. The idea of bagging (a.k.a. bootstrap sampling) is to draw many random subsets of the original data, fit a decision tree on each subset, and average the predictions of all decision trees. This procedure is useful to decrease the variances in the predictions. Bagging is particularly useful for decision trees, since they have low biases and high variances.

Random Forests (Breiman, 2001) are essentially a collection of a large number of decision trees, each trained on a different bagged sample. Predictions are then obtained by averaging all the trees. One may ask: if the trees are correlated and give similar predictions, then they will look like the same tree and taking the average will not help much. Random Forests actually use a better way to generate (grow) the trees so as to ensure that they are not highly correlated. This is achieved by **select splitting variables at random**: before splitting each node, instead of considering all p possible variables, the algorithm selects only $m < p$ variables at random, and consider those variables for splitting. This can de-correlate the trees, and improve the predictions by reducing the variances. In general, choosing $m \leq \sqrt{p}$ is good enough.

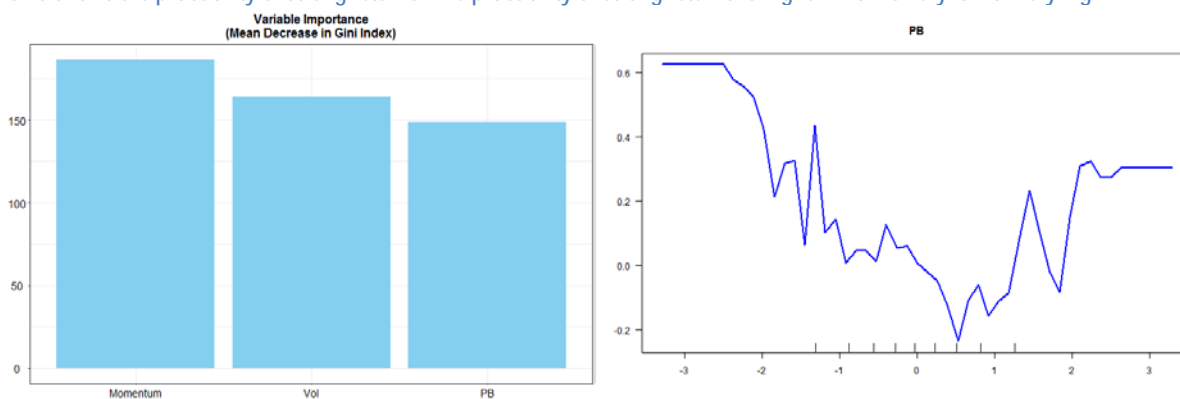
Since Random Forest is an ensemble model consisting of many decision trees, it is difficult to interpret the model. Nevertheless, we can still use some measures to gauge the importance of a variable to our model. When we split a node with a variable, we can measure the expected improvement (i.e. decrease in the impurity in the node) due to this particular split. For each variable, we can record this improvement over all nodes and all trees in the Random Forest model.

Another useful measure to understand how a variable varies with the output is the partial dependency plot. Consider our example in Figure 64 where we try to use 3 factors to classify whether a stock will have strong or weak future returns. In this case, we are modeling the log odds:

$$f(x) = \log \frac{\Pr(\text{Strong}|x)}{\Pr(\text{Weak}|x)}$$

where x is either Momentum, Volatility or P/B ratio. The partial dependency plots tell us the relationship of a variable with the output (e.g. returns). The Figure below shows an example of the partial dependency plot for the variable P/B.

Figure 64: The variable importance plot (left) shows that momentum is the most important variable. In the partial dependency plot (right), the y-axis shows the probability of strong returns. The probability of strong returns is higher if P/B is very low or very high



Source: J.P.Morgan Macro QDS, FactSet

Use of Random Forests in Global stock selection strategy

Let us consider a global stock selection model that predicts 1-month stock returns across the MSCI World universe (over 2400 stocks), using the 14 risk factors shown below. We consider a subset of this universe where we have observations for all 14 factors, which corresponds to about 1400 stocks. The table below shows the 14 risk factors used in this example.

Figure 65: 14 Factors used to predict 1-month ahead stock returns

Factors	
Price-to-book ratio	Earnings Certainty
Gross profit / Total assets	Cash flow yield
ROE	Dividend yield
Net Margin	Realized vol
Asset Turnover	1M momentum
Gearing	12M-1M momentum
Forward earnings yield	Market cap

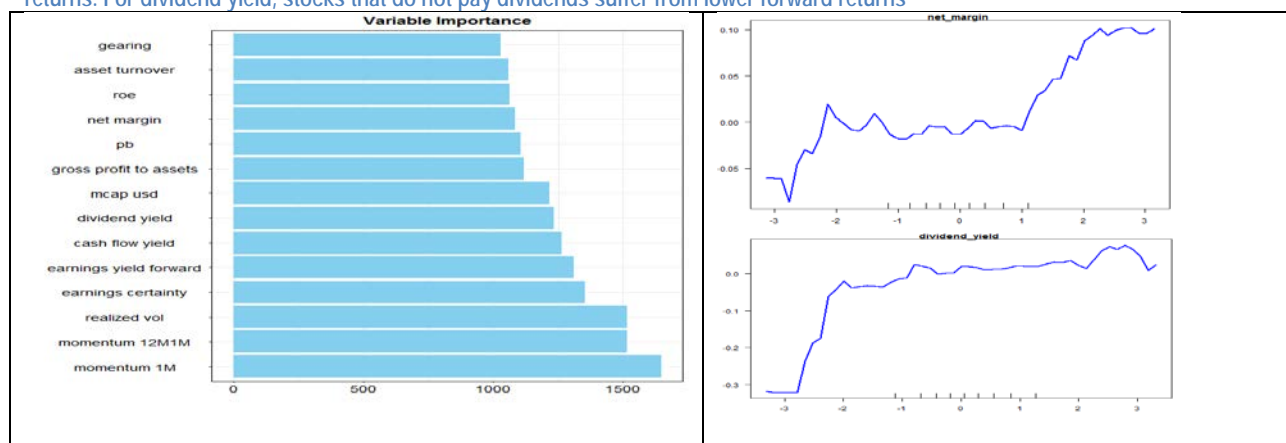
Source: J.P.Morgan Macro QDS

First, it is important to normalize the variables. At each month end, we normalize the forward returns as well as each factor by transforming them into Gaussian distributions. We then run the Random Forest algorithm for regression on the training data to estimate the model, using the R packages “randomForest” and “caret” (Kuhn 2008). We tried to use different numbers of randomly-selected variables at each node split. The optimal number is determined by the out-of-bag (OOB) prediction errors. Recall that in the Bagging procedure, we subsample the data to build decision trees, i.e. the trees are not trained on all data points, but only a subset of it. Hence, for an observation we can obtain a prediction only from the trees not trained on that data point, and calculate the OOB error. In general, we can use the OOB error to tune other parameters, e.g. number of trees, maximum depth of the trees, etc.

Since the Random Forest algorithm is computationally intensive and takes a relatively long time to run, we look at a relatively small training set that contains monthly observations between January 2014 and January 2015. Such a training set has about 18000 rows of distinct observations. We find that using 14 factors at each node split leads to the lowest OOB error. We grow 100 trees and take the average prediction as the output of the Random Forest model.

The Figure below shows the variable importance plot and the partial dependency plots for 2 of the variables: Net Margin and Dividend Yield.

Figure 66: Left: Variable importance plot for the Random Forest model that predicts 1-month ahead stock returns based on 14 factors. Right: Partial dependency plots for Net Margin and Dividend Yield. It shows that the higher the Net Margin of a stock, the higher the 1-month ahead returns. For dividend yield, stocks that do not pay dividends suffer from lower forward returns



Source: J.P.Morgan Macro QDS, FactSet

Using the Random Forest model fitted between January 2014 and January 2015, we obtain 1-month return predictions at each month end. We use these predictions as our stock selection signal, and consider 5 quintiles of equal-weighted basket,

each with about 280 stocks in MSCI World. Figure below shows the long/short performance (green), the long-only strategy (blue) and the MSCI World net return index, all in USD.

Figure 67: Performance of the long/short strategy (green), the long-only strategy (blue) and the MSCI World index (red)



Source: J.P.Morgan Macro QDS, FactSet

The table below shows the returns statistics of the baskets constructed from the model.

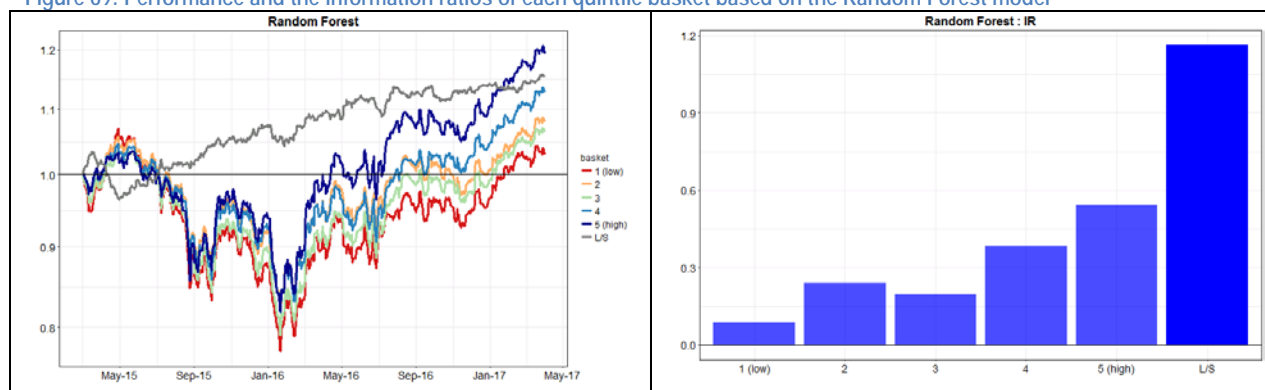
Figure 68: Performance statistics (Feb 2015 – Mar 2017) of the baskets based on the Random Forest model

Basket	Cum. Return	CAGR	Volatility	IR	Max Drawdown	Hit Ratio
1 (low)	3.0%	1.0%	11.3%	0.09	27.7%	37.3%
2	7.9%	2.5%	10.7%	0.24	22.7%	37.8%
3	6.4%	2.1%	10.7%	0.19	23.6%	38.3%
4	12.8%	4.1%	10.6%	0.38	21.9%	37.0%
5 (high)	19.2%	6.0%	11.1%	0.54	20.9%	39.5%
L/S	15.4%	4.8%	4.2%	1.16	6.9%	37.7%

Source: J.P.Morgan Macro QDS, FactSet

The Figure below shows the wealth curves, rolling 12-month returns and the information ratios of each basket.

Figure 69: Performance and the information ratios of each quintile basket based on the Random Forest model



Source: J.P.Morgan Macro QDS, FactSet

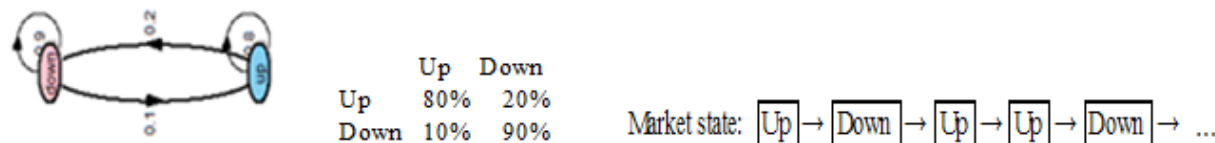
Hidden Markov Models

In the section on supervised learning we discussed estimating regression ‘betas’ in dynamic systems (e.g. Kalman filter). If these betas change discretely (rather than continuously), we refer to them as ‘regimes’. **Hidden Markov Models (HMM)** are similar to the Kalman filter (i.e. similar to other State Space models) where the probability of the next state only depends on the current state (i.e. hidden state follows a discrete Markov process). HMMs are useful statistical models because in many real world problems, we are interested in identifying some events which are not directly observable (e.g. are we in an up-trending or down-trending market?), but these events can be inferred from other variables that we can observe (e.g. market returns, market volatility, etc.).

Historically, HMMs have been applied extensively in speech recognition in the 1990s and more recently, in bioinformatics such as gene sequence analysis. In finance, it has been used to model market regimes. For inquisitive readers, we recommend the introduction to HMM by Rabiner and Juang (1986). We describe HMMs briefly here, motivated by our use-case of applying it to detect market regimes. The first question is to know the state of the market: Is it trending upwards? Unfortunately, we cannot observe it directly. Hence it is called a hidden state.

A Hidden Markov Model can be formulated as follows: A Markov process for the state of the market (suppose we only have 2 states – Up and Down):

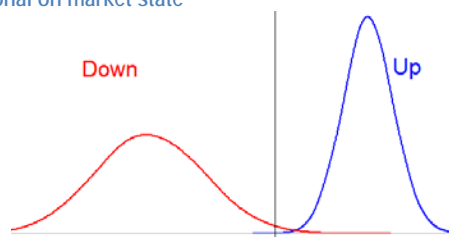
Figure 70: 2-state HMM



Source: J.P.Morgan Macro QDS

This means that if the market is currently in the upward state, then the probability of remaining in the upward state is 80%, and there is 20% of probability of transiting to a downward state. Conditional on the state of the market, assume a distribution for the returns as $(r|state) \sim N(\mu_{state}, \sigma_{state}^2)$.

Figure 71: Distributions of log returns conditional on market state



Source: J.P.Morgan Macro QDS

Since we observe the historical sequence of returns, we can infer the likelihood of being in a particular state at each time. Using the Expectation-Maximization algorithm⁴⁸ (i.e. the Baum-Welch algorithm, Bilmes (1998)), we can estimate the parameters in the Hidden Markov Model by maximizing the likelihood. The estimated parameters include: the initial probabilities in each state, transition probabilities of the state, the probabilities of being in each state, and mean and volatilities of the returns conditional on each state.

⁴⁸ The Expectation Maximization (EM) algorithm and extensions are discussed in Meng and Pedlow (1992), Little and Rubin (2002), Meng and Rubin (1991, 1993), Meng (1994a), van Dyk, Meng and Rubin (1995), Liu and Rubin (1994), Liu, Rubin and Wu (1998), Meng and van Dyk (1997), Liu and Wu (1999) and Liu (2003). Variational Bayes is described in Jordan et al (1999), Jaakkola and Jordan (2000), Blei, Ng and Jordan (2003), Gershman, Hoffman and Blei (2013).

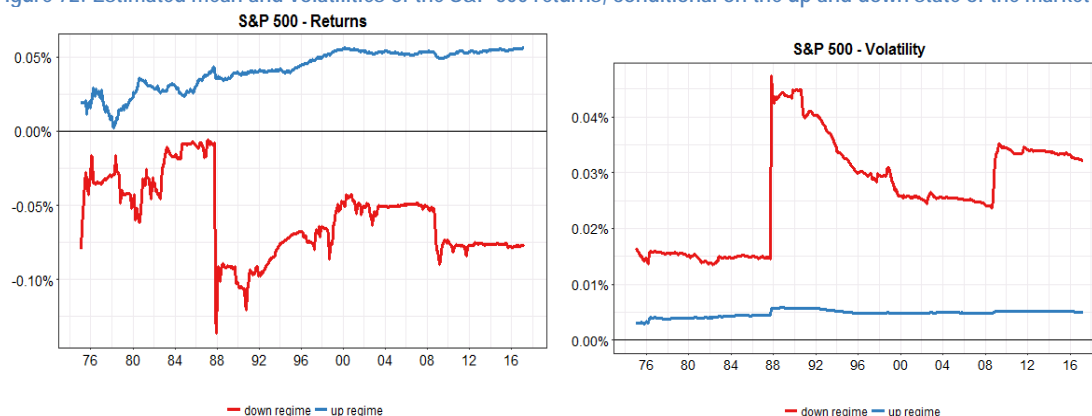
Use of Hidden Markov Models in Factor Timing:

As an illustrative example, we design a market timing trading strategy based on HMM. It is a simple trend following strategy in which we are long the S&P 500 if market is trending higher, and are in cash if the market is down-trending. Using daily returns of the S&P 500 from April 1971, we estimate the above HMM model using the R package "mhsmm" (O'Connell et al 2011). We start the estimation from January 1975 so that we have about 4 decades of daily returns. In general, it is preferable to have more observations in the estimation of HMM so that one can ensure different states (in this case, Up and Down market) have occurred with a significant frequency.

On the last trading day of each month, we re-estimate the HMM model. The Figure below shows the estimated mean and volatilities in the conditional Gaussian distributions. This gives intuitive sense of the meaning of the two market states:

- Up state: Periods of positive returns, with lower volatilities; and
- Down state: Periods of negative returns, with higher volatilities.

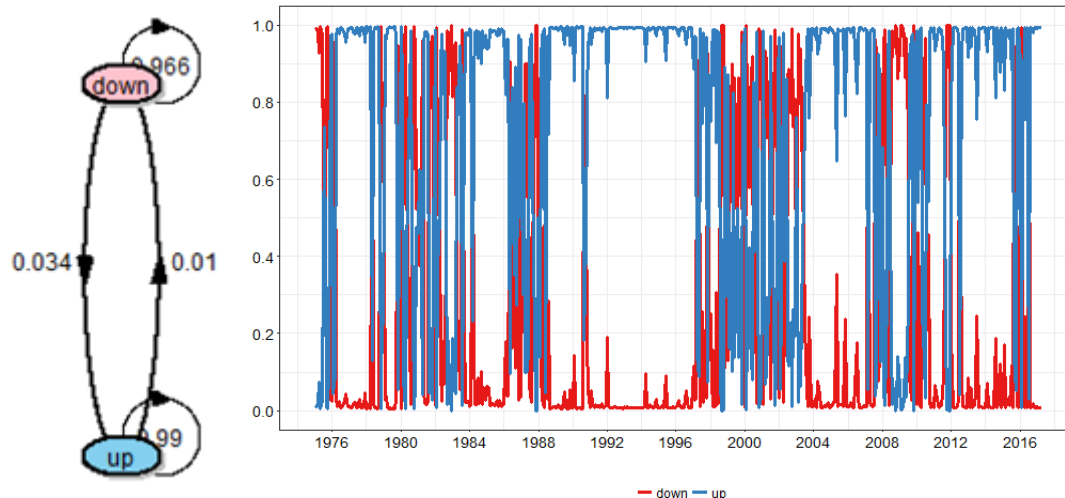
Figure 72: Estimated mean and volatilities of the S&P 500 returns, conditional on the up and down state of the market



Source: J.P.Morgan Macro QDS

The probabilities in each state are shown in the figure below. These probabilities will be used to infer the state of the market: if the probability in the upward state is greater than 50%, we take the current state as Up. We also show the latest transition probabilities on the left.

Figure 73: Estimated probabilities in each state (Up or Down market)



Source: J.P.Morgan Macro QDS

Using the HMM estimated on the last trading day of each month, we determine the latest state of the market. The simple trading strategy is as follows: if the market state is Up, we go long the S&P 500 index, and if the market state is Down, we are not invested (zero returns). HMM market timing strategy significantly reduced the drawdown (as compared to long S&P 500), and modestly improved the overall Sharpe ratio. Performance is shown in the figure below.

Figure 74: HMM monthly timing strategy and S&P 500



	Long-only	Timing
CAGR (%)	8.2	6.1
Volatility (%)	16.8	11
Information ratio	0.49	0.56
Max DD (%)	56.8	38.4

Source: J.P.Morgan Macro QDS

Unsupervised Learning: Clustering and Factor Analyses

Unsupervised learning algorithms examine the dataset and identify relationships between variables and their common drivers. In unsupervised learning the machine is simply given the entire set of returns of assets and it does not have a notion of what are independent and what are the dependent variables. Methods of unsupervised learning are often categorized as either **Clustering** or **Factor analyses**. Clustering⁴⁹ involves splitting a dataset into smaller groups based on some notion of similarity. In finance that may involve identifying historical regimes such as high/low volatility regime, rising/falling rates regime, rising/falling inflation regime, etc. Correctly identifying the regime can in turn be of high importance for allocation between different assets and risk premia. **Factor analyses**⁵⁰ aim to identify the main drivers of the data or identify best representation of the data. For instance, yield curve movements may be described by parallel shift of yields, steepening of the curve, and convexity of the curve. In a multi-asset portfolio, factor analysis will identify the main drivers such as momentum, value, carry, volatility, liquidity, etc. A very well-known method of factor analysis is **Principal Component Analysis** (PCA). The PCA method carries over from the field of statistics to unsupervised Machine Learning without any changes.

Clustering

In this section we evaluate and compare several clustering algorithms: K-Means, Ward, Birch, Affinity, Spectral, MiniBatch, Aggregate, HDSBscan, and ICA. We will not provide theoretical details behind each of the methods, and just provide a brief description of different methods below. Readers interested in more details can see the [sklearn](#) and [hdbscan](#) documentation. We have applied all clustering algorithms by using Python library “sklearn”.

K-means – Simplest clustering algorithm that starts by initially marking random points as exemplars. It iteratively does a two-step calculation: in the first step, it maps points to closest exemplar; in the second step it redefines the exemplar as the mean of the points mapped to it. It locates a fixed number of clusters (which is assigned to two in our code example that follows).

Ward – A hierarchical clustering technique that is similar to K-means, except that it uses a decision tree to cluster points.

Birch – A hierarchical clustering technique designed for very large databases. It can incrementally cluster streaming data; hence in many cases, it can cluster with a single pass over the data set.

Affinity Propagation – Algorithm involves passing of ‘soft’ information that treats every point as a possible exemplar and allows for the possibility of every point being included in a cluster around it. The algorithm is known to be good for finding a large number of small clusters. The number of clusters chosen can be indirectly influenced via a ‘preference’ parameter; we have used the default settings within the Python *sklearn* library.

Spectral Clustering – Like affinity propagation, it passes messages between points, but does not identify the exemplar of each cluster. It needs to be told the number of clusters to locate (we have specified two). This is understandable, since the algorithm computes an affinity matrix between samples, embeds into a low-dimensional space and then runs K-means to locate the clusters. This is known to work well for a small number of clusters.

⁴⁹ K-means clustering is covered in Gordon (1999), Kaufman and Rousseeuw (1990). Applications to image compression are in Gersho and Gray (1992). Theoretical analysis of unsupervised learning is covered in a series of papers by Krishnamachari and Varanasi (see bibliography). For association rules, see Agrawal et al (1995); for self-organizing map, see Kohonen (1990) and Kohonen et al (2000).

⁵⁰ For PCA (principal component analysis), see Mardia et al (1979); for principal curves and surfaces, see Hastie (1984) and Hastie and Stuetzle (1989); for principal points, see Flury (1990), Tarpey and Flury (1996); for spectral clustering, see von Luxburg (2007), Spielman and Teng (1996). For ICA (independent component analysis), see Comon (1994), Bell and Sejnowski (1995), Hyvarinen and Oja (2000).

Mini-Batch – To reduce processing time, mini-batches of data are processed with a k-means algorithm. However unlike k-means the cluster centroids are only updated with each new batch, rather than each new point. Mini-Batch typically converges faster than K-Means, but the quality can be slightly lower.

HDBSCAN – Hierarchical Density-Based Spatial Clustering of Applications with Noise is a ‘soft’ clustering algorithm where points are assigned to clusters with a probability and outliers are excluded. HDBSCAN runs multiple DBScans over the data seeking the highest cluster stability over various parameters.

ICA – Independent Component Analysis is similar to Principal Component Analysis (PCA) but is better at distinguishing non-Gaussian signals. While these techniques are typically used in factor extraction and dimension reduction, we have used the exposures to these common (hidden) factors to form clusters to explore the potential of these components to identify similarity between members’ returns.

Agglomerative Clustering – This is another hierarchical clustering technique that starts with each point as exemplar and then merges points to form clusters. Unlike Ward which looks at the sum of squared distance within each cluster, this looks at the average distance between all observations of pairs of clusters.

HDBScan – This forms clusters with a roughly similar density of points around them. Points in low-density regions are treated as outliers. It can potentially scan the data set multiple times before converging on the clusters.

In our comparison of clustering methods we analyze a portfolio of Global Equities, specifically the MSCI Global Developed Markets Index. We test clustering techniques and compare them to the default Country/Sector groupings. In this study we used 36 months of excess USD returns to calculate a correlation matrix and form clusters with each of the techniques of interest. The Mean-Shift required a 2D dataset to cluster, so we used a 2 component PCA model in this case. This market is broken down into 173 Country/Sector groupings according to MSCI/GICS. We use these Country/Sector groups as our ‘ground truth’ when calculating various metrics for the clusters. Note that the Real Estate sector separation occurred in Sep 2016.

To compare the commonality of our clusters we use a suite of similarity scores, rank each method and report the average rank. A description of similarity scores, as well as their value when applied to the group of clustering algorithms is shown in the table below.

Figure 75: List of metrics to analyze performance of clustering algorithms

Metric	Name	Notes: Most scores range -1 or 0 to+ 1
ARI	Adjusted Rand Index	See detailed breakout box below
AS	Accuracy Score	Subsets must exactly match
PS	Precision Score	Ratio of True Positives to TP + False Positives
F1	F1 Score (AKA F-measure)	Precision (TP/TP+FP) to Recall (TP/TP+FN) $F1=2P*R/P+R$
HS	Homogeneity Score	Each cluster contains only members of a single class
CS	Completeness Score	All members of a given class are assigned the same cluster
HCV	Homogeneity Completeness V-Measure	Average of HS and CS
HL^	Hamming Loss	Fraction of labels that are incorrectly predicted, 0 is best to 1
JS	Jaccard Similarity Score	Size of the intersection divided by the size of the union
MI*	Mutual Information	Not standardized, max value unbounded
A_MI	Adjusted Mutual Information Score	Adjusted for chance, perfect =1, chance = 0, can be negative
Z_MI	Normalized Mutual Information Score	Normalization of MI to be between 0 and 1
Avg^	Average Rank (Ascending)	All metrics are ranked, then averaged.

Source: [Scikit-learn: Machine Learning in Python](#), Pedregosa et al. (2011), JMLR 12, Pg. 2825-2830, J.P. Morgan QDS.

Adjusted Rand Index

The Adjusted Rand Index (ARI) can take values between -1 to 1 when there is perfect overlap of the two data clusters. The adjustment allows the index to be less than 0 when the overlap is less than expected. For each subset S of two clustered datasets X & Y, a is the number of pairs in the same subset (Si), b those in different subsets (i.e. not Si in either X or Y) while c & d capture the two difference sets.

$$R = \frac{a + b}{a + b + c + d}$$

A contingency table for the two sets can be used to show the number of elements formed by the union at each point. The ARI is then calculated by comparing the actual Rand Index to the Expected Maximum Index;

$$ARI = \frac{\text{Index} - \text{ExpectedIndex}}{\text{MaxIndex} - \text{ExpectedIndex}}$$

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

Where n_{ij} is the union of members in X_i and Y_j and a_i and b_j are the contingency table row and column sums over all the n_{ij} 's.

Figure 76: Comparing the output of clustering algorithms to grouping using MSCI Country & GICS sectors

Method \ Score	ARI	F1	Z_MI	A_MI	HS	CS	HCV	AS	PS	JS	MI*	HL^	Avg^
Birch (2D PCA)	0.142	0.008	0.656	0.277	0.712	0.604	0.654	0.008	0.028	0.008	3.054	0.992	1.8
Ward	0.120	0.003	0.648	0.260	0.704	0.596	0.645	0.005	0.004	0.005	3.018	0.995	3.7
Affinity	0.124	0.003	0.619	0.275	0.653	0.586	0.618	0.005	0.004	0.005	2.800	0.995	4.0
Spectral	0.031	0.011	0.584	0.217	0.599	0.569	0.584	0.011	0.044	0.011	2.569	0.989	4.4
MiniBatch	0.123	0.003	0.616	0.265	0.651	0.583	0.615	0.005	0.003	0.005	2.790	0.995	4.9
Agglomerative	0.129	0.001	0.645	0.259	0.700	0.594	0.643	0.002	0.000	0.002	3.002	0.998	5.0
ICA	0.050	0.004	0.539	0.134	0.569	0.510	0.538	0.003	0.057	0.003	2.441	0.997	5.9
HDBSCAN	0.013	0.004	0.508	0.043	0.503	0.513	0.508	0.002	0.012	0.002	2.157	0.998	6.9
HRP	0.003	0.000	0.314	0.017	0.175	0.562	0.267	0.002	0.000	0.002	0.749	0.998	8.4

Source: MSCI, FactSet, J.P. Morgan. * indicates score unbounded, all other scores/indices are bounded between -1 or 0 and +1. Hamming Loss is a loss function, so smaller values are preferred

We found the Birch, Ward and Affinity Propagation techniques return clusters that most closely match MSCI Country GICS Sectors, while HRP, HDBScan and ICA are the most different (note that we are not suggesting that these scores should be closer to 1; the goal of the clustering algorithms was not to match GICS country sectors but we are using GICS as a comparison).

Assignment Persistence

There is some concern about the stability of clusters in time-series (some of this concern can be attributed to instabilities in the correlation matrix which they are often based on). To test this we have created a series of clusters every year over rolling 3 year windows of excess USD returns.

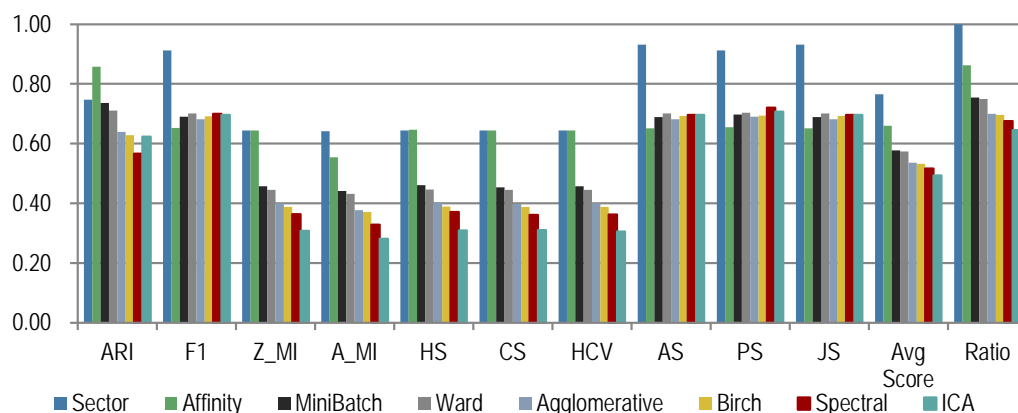
To measure the persistence of the clusters we use the prior year's assignment as our 'ground truth' for each subsequent year. (Note we cannot use a classic turnover measure because, while sector assignment is consistent from one period to the next, cluster assignment is arbitrary - Clusters labelled 1, 2, 3 one period might be labelled 3, 2, 1 the next.)

Figure 77: Comparison of cluster stability vs. GICS sectors

Clustering Algorithm	ARI	F1	Z_MI	A_MI	HS	CS	HCV	AS	PS	JS	Avg Score	Ratio
Sector	0.75	0.91	0.64	0.64	0.64	0.64	0.64	0.93	0.91	0.93	0.77	100%
Ward	0.71	0.70	0.45	0.43	0.45	0.44	0.45	0.70	0.70	0.70	0.57	75%
Agglomerative	0.64	0.68	0.40	0.38	0.40	0.40	0.40	0.68	0.69	0.68	0.54	70%
Birch	0.63	0.69	0.39	0.37	0.39	0.39	0.39	0.69	0.69	0.69	0.53	70%
Spectral	0.57	0.70	0.36	0.33	0.37	0.36	0.36	0.70	0.72	0.70	0.52	68%
ICA	0.62	0.70	0.31	0.28	0.31	0.31	0.31	0.70	0.71	0.70	0.49	64%
MiniBatch	0.74	0.69	0.46	0.44	0.46	0.45	0.46	0.69	0.70	0.69	0.58	75%
Affinity	0.86	0.65	0.64	0.55	0.65	0.64	0.64	0.65	0.65	0.65	0.66	86%

Source: J.P.Morgan Macro QDS, MSCI

Figure 78: Comparison of cluster stability vs. GICS sector stability

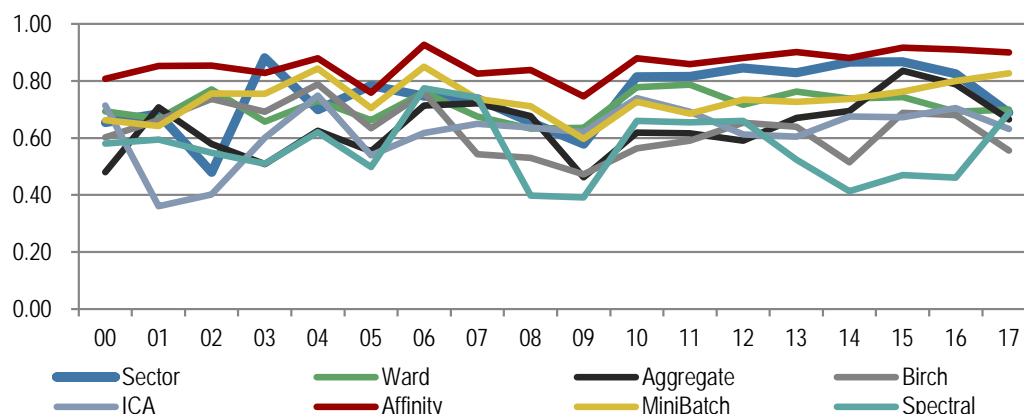


Source: J.P.Morgan Macro QDS, MSCI

We find that most clustering algorithms exhibit approximately 75% of the stability with GICS Sector assignment within the MSCI GDM index. Of the methods tested, using the MiniBatch method of clustering of a correlation matrix formed from excess returns over a rolling 3 year window is one of the strongest, with an Adjusted Rand Index of 0.74, which favorably compares to the GICS Sector ARI of 0.75.

Affinity clustering also looks promising, with cluster sets that can be 86% as stable as GICS Sectors. However the results are not directly comparable because affinity clustering forms an unbounded set of clusters, typically over 100 with this data, so direct comparisons to the 10/11 GICS Sectors should be made with care.

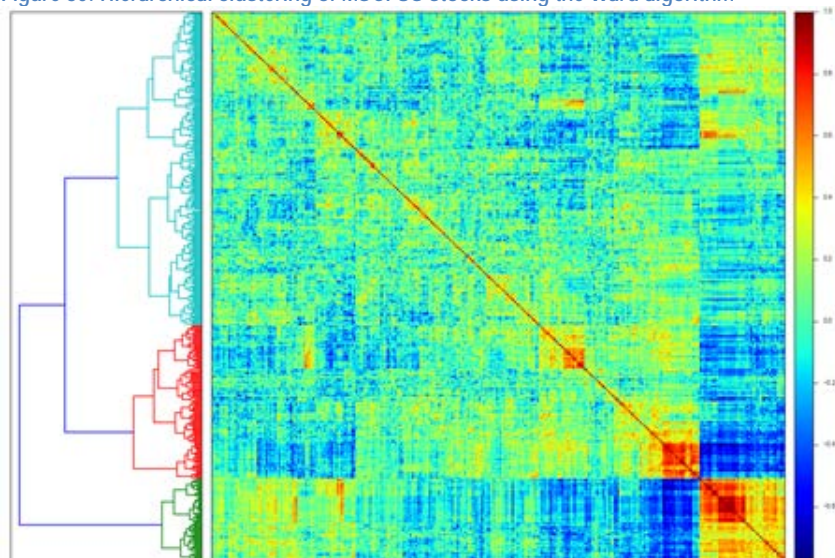
Figure 79: Comparison of cluster stability vs. GICS sector stability across time



Source: J.P.Morgan Macro QDS, MSCI

Visualization is an important part of clustering analysis. The example below shows the correlation matrix of monthly returns for MSCI US stocks (colors in the heat map indicating levels of pairwise correlation). Progressive grouping into larger and larger clusters (hierarchical clustering) is achieved via the Ward method and is shown on the left side of the chart. Such clusters can for instance be used to diversify portfolio risk, use in pair trading models, etc.

Figure 80: Hierarchical clustering of MSCI US stocks using the Ward algorithm



Source: J.P.Morgan Macro QDS

If the number of assets is small, one can visualize clustering by using minimum spanning trees (MST). MSTs are a graphical representation of correlation within a group of assets. Assets are represented by points, and MSTs connect all the assets (points) via lines. The length of a line between two assets is inversely proportional to correlation. In MSTs there are no loops, i.e. each asset is connected to at most two other assets. The lines connecting the assets are chosen in such a way to minimize the sum of all line distances. The Figure below shows an example of a minimum spanning tree for JPM cross-asset risk premia indices. For instance, one can find relatively similar behavior of short volatility strategies across asset classes. For more examples of MSTs see our report on [Hierarchical Risk Parity](#).

Figure 81: Minimum Spanning Tree for 31 JPM tradable risk premia indices



Source: J.P.Morgan Macro QDS

Factor Analysis through PCA

In addition to Clustering, Factor analyses are another important segment of unsupervised learning. Factor analyses aim to identify the main drivers of the data or identify the best representation of the data. For instance, yield curve movements may be described by parallel shift of yields, steepening of the curve, and convexity of the curve. In a multi-asset portfolio, factor analysis will identify the main drivers such as momentum, value, carry, volatility, liquidity, etc. A very well-known method of factor analysis is Principal Component Analysis (PCA). The PCA method carries over from the field of statistics to unsupervised Machine Learning without any changes.

Principal Component Analysis (PCA) is a statistical tool for dimensionality reduction that continues to find use in Big Data analysis. Given market data, PCA can decompose each time series into a linear combination of uncorrelated factors. Further, PCA can quantify the impact of each factor on the time series by calculating the proportion of the variance that is explained by any given factor. By plotting the PCA factors, we can visualize and estimate typical moves one can expect to see in a given market regime.

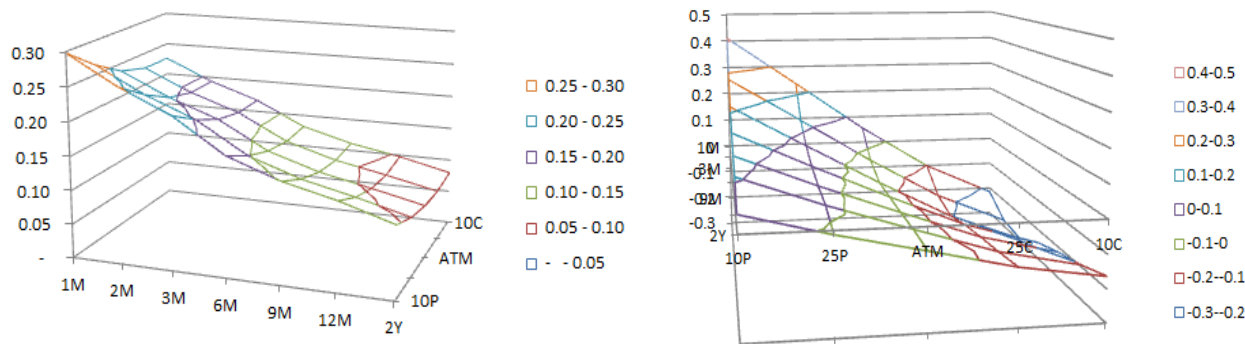
PCA stems directly from the singular value decomposition used widely in linear algebra. Given a set of p time series $R = (r_1, r_2, \dots, r_p)$ with mean zero, we can find the singular value decomposition as $R = UDV^T$ with the elements of D sorted in descending order. The columns of V are called the principal directions and the first k columns of UD are called the first k principal components. If we retain only the first k diagonal elements in D and denote it as $D_{(k)}$, then we can construct a rank- k approximation for R as $R_{(k)} = UD_{(k)}V^T$. Randomized algorithms are [used](#) for calculating PCA when the size of matrix is large.

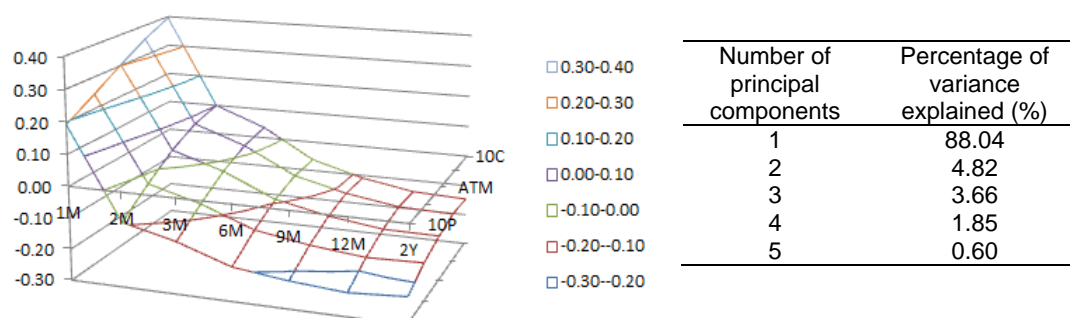
Example of PCA: Parametrizing JPY Volatility Surface

In the section on supervised learning, we have already demonstrated the use of PCA in conjunction with supervised learning techniques (e.g. SVM) to make return forecasts. In that example, PCA was used to reduce the number of considered variables from 377 to 60. In turn, this enabled a more stable implementation of the SVM algorithm.

In this section, we illustrate the use of PCA to explore the USDJPY implied volatility surface. If we apply PCA to daily changes in the USDJPY implied volatility data from 2001 to present, we find that first three principal components of the daily change can explain over 95% of surface variation. These three principal components of daily changes are shown below.

Figure 82: First three principal components of USDJPY volatility surface change and percentage of variance explained

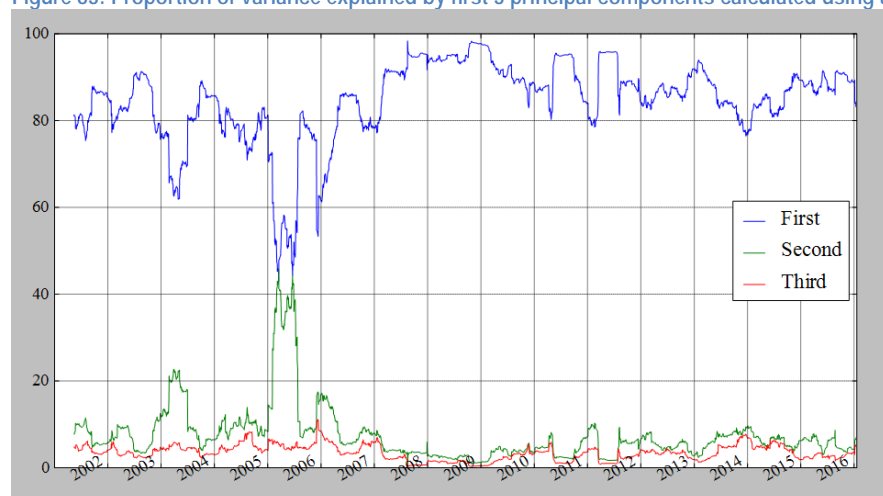




Source: J.P.Morgan Macro QDS

The figures above reveal that the main movement is a complicated 2-dimensional move, that is not flat across the vol surface (e.g. it is not a parallel shift, or shift in skew or term structure). Rather, the typical move in a day is skewed towards the left showing an enhanced perception of risk by JPY appreciation rather than depreciation. This bias further increases with maturity. The values of the first principal component are hence used to calculate hedge ratios. By calculating the residual with respect to a few principal components, analysts can also make relative value calls by assuming that the residual will revert to zero. Further, by plotting the percentage of variance explained by various principal components, we see that the variance explained by the first principal component has not changed dramatically in recent years.

Figure 83: Proportion of variance explained by first 3 principal components calculated using a rolling 90 day covariance matrix



Source: J.P.Morgan Macro QDS

The most common application of PCA is to decompose portfolio risk (finding the beta of a portfolio to different principal components). In our primer on [systematic cross asset strategies](#), we applied PCA to the set of cross-asset risk premia to locate the principal components. The performance of each principal component is reported below.

Figure 84: Performance of PCA factors as derived from cross-asset risk premia (value, momentum, carry, short-volatility across equity, bond, currency and commodity complex)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Average (%)	8.9	1.0	2.9	3.1	5.6	1.0	8.0	4.3	1.6	8.1
CAGR (%)	6.5	-0.8	1.4	1.8	5.1	0.4	7.8	3.9	1.1	8.0
STDev (%)	22.8	18.7	17.6	15.6	11.9	11.5	9.6	10.1	9.7	8.5
MaxDD (%)	-63.8	-72.1	-56.8	-57.2	-34.6	-37.3	-47.9	-39.0	-37.4	-26.8
MaxDDur (in yrs)	9.1	30.2	24.6	18.8	15.1	23.3	5.4	9.0	23.9	4.1
Sharpe Ratio	0.39	0.05	0.16	0.20	0.47	0.09	0.83	0.43	0.16	0.96
Sortino Ratio	0.61	0.07	0.24	0.28	0.92	0.13	1.37	0.66	0.24	1.67
Calmar Ratio	0.31	0.02	0.18	0.30	0.47	0.07	0.58	0.47	0.14	1.08
Pain Ratio	0.39	0.03	0.09	0.10	0.47	0.05	1.26	0.57	0.09	1.99
Correl with SPX	0.23	-0.54	-0.03	0.22	0.20	0.54	0.00	-0.20	-0.07	0.30
Correl with UST	0.09	0.14	0.01	0.68	0.11	-0.03	0.05	-0.01	0.27	-0.01
Skewness	0.05	0.20	0.00	-0.50	3.16	0.03	-0.35	-0.26	-0.18	-0.12
Kurtosis	2.28	2.98	2.00	2.10	32.65	1.11	2.69	2.77	0.57	0.97

	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
Average (%)	2.0	0.1	6.5	2.7	2.6	0.8	8.6	3.2	1.1	2.1
CAGR (%)	1.7	-0.2	6.4	2.5	2.4	0.6	8.8	3.1	1.0	2.1
STDev (%)	8.7	7.5	7.8	7.0	6.6	6.2	5.7	5.3	4.2	2.8
MaxDD (%)	-30.3	-46.3	-24.1	-41.8	-21.0	-28.4	-8.0	-17.0	-14.1	-6.3
MaxDDur (in yrs)	11.8	32.8	4.9	27.8	12.8	14.1	1.8	8.0	9.5	3.8
Sharpe Ratio	0.23	0.02	0.84	0.38	0.39	0.13	1.50	0.61	0.26	0.74
Sortino Ratio	0.35	0.02	1.81	0.64	0.59	0.19	3.05	0.97	0.38	1.25
Calmar Ratio	0.16	0.02	0.85	0.56	0.14	0.13	1.55	1.19	0.17	0.55
Pain Ratio	0.22	0.00	1.89	0.14	0.48	0.08	7.58	0.98	0.30	1.77
Correl with SPX	-0.12	0.08	0.04	-0.09	-0.18	0.07	-0.09	0.19	0.14	0.04
Correl with UST	-0.01	-0.07	0.38	0.05	0.05	0.48	-0.10	0.13	-0.08	-0.03
Skewness	-0.06	-0.26	3.05	0.83	-0.15	0.00	0.00	-0.34	-0.14	-0.08
Kurtosis	1.94	3.44	30.97	5.71	0.86	1.85	0.55	4.09	1.32	0.63

Source: J.P.Morgan Macro QDS

We caution that principal components are not independent factors. When the returns are not jointly Gaussian, uncorrelated density does not imply independence. “Uncorrelated” means the lack of a linear relationship; “independence” is a more comprehensive statistical notion. PCA yields uncorrelated factors, but not independent ones. Independent factors can be obtained through a different technique called Independent Component Analysis (ICA).

Deep and Reinforcement Learning

Deep Learning

Deep Learning is a Machine Learning method that analyzes data in multiple layers of learning. It may start doing so by learning about simpler concepts, and combining these simpler concepts to learn about more complex concepts and abstract notions. It is often said that the goal of automation is to perform tasks that are easy for people to define, but tedious to perform. On the other hand, the goal of Deep Learning AI systems is to perform tasks that are difficult for people to define, but easy to perform. Deep Learning is more similar to how people learn, and hence is a genuine attempt to artificially recreate human intelligence. It is likely not a coincidence that Deep Learning methods are based on neural networks – which are in turn inspired by the way neurons are connected in the human brain.

Neural network techniques are loosely inspired by the workings of the human brain. In a network, each neuron receives inputs from other neurons, and ‘computes’ a weighted average of these inputs. The relative weighting of different inputs is guided by the past experience, i.e. based on some training set of inputs and outputs. Computer scientists have found that they could mimic these structures in an approach called **Deep Learning**. Specifically, Deep Learning is a method to analyze data by passing it through multiple layers of non-linear processing units – neurons (‘Deep’ refers to multiple layers of neurons in the neural network). Once the signal weightings are calibrated from the sample dataset (training/learning dataset), these models have strong out of sample predictive power. Multiple layers of signal processing units (i.e. neurons) allow these models to progressively learn more complex concepts out of simpler ones. In this section we illustrate potential applications in finance for four different Deep Learning architectures: Multilayer Perceptron, Long Short-term memory, Convolutional Neural Networks and Restricted Boltzmann Machine. **Multilayer Perceptron (MLP)** is one of the first designs of multi-layer neural networks, designed in such a way that the input signal passes through each node of the network only once (also known as a ‘feed-forward’ network). **Long Short-term memory (LSTM)** is a neural network architecture that includes feedback loops between elements. This can also simulate memory, by passing the previous signals through the same nodes. LSTM neural networks are suitable for time series analysis, because they can more effectively recognize patterns and regimes across different time scales. **Convolutional Neural Networks (CNN)** are often used for classifying images. They extract data features by passing multiple filters over overlapping segments of the data (this is related to the mathematical operation of convolution). **Restricted Boltzmann Machine (RBM)** is a neural-network based dimensionality reduction (unsupervised learning) technique. The neurons in RBM form two layers called the visible units (reflecting returns of assets) and the hidden units (reflecting latent factors). Neurons within a layer – hidden or visible – are not connected; this is the ‘restriction’ in the restricted Boltzmann machine.

Deep Learning vs. Classical Machine Learning

In the sections before, we discussed classical Machine Learning tools like lasso regressions, support vector machines and decision trees. While they perform well on a variety of numerical prediction/analysis tasks, they are not designed to perform certain tasks that human beings can perform intuitively like recognizing objects in images or analyzing text. **Neural network based Deep Learning tools** are behind many of the newer developments in the technology sector, like smart home products (e.g. Amazon’s Alexa), automated chat-bots, automated personal assistants, language translators, sentiment analysis, drug discovery, recommendation systems (e.g. Netflix), image processors, self-driving cars and speech recognition. Despite such successes, Deep Learning tools are rarely used in time series analysis, where tools from ‘classical’ Machine Learning dominate.

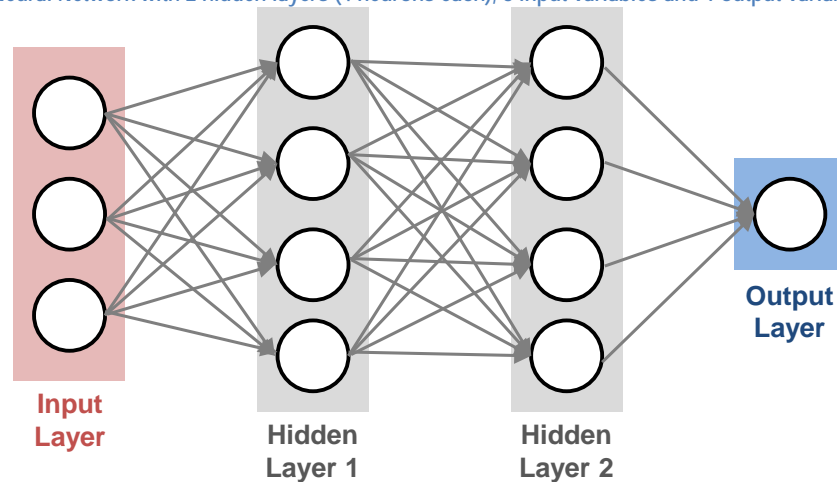
Anecdotal evidence from observing winning entries at data science competitions (like Kaggle) suggests that structured data is best analyzed by tools like XGBoost and Random Forests. Use of Deep Learning in winning entries is limited to analysis of images or text. Deep Learning tools still require a substantial amount of data to train. Training on small sample sizes (through so-called generative-adversarial models) is still at an incipient research stage. The necessity of having large sample data implies that one may see application of Deep Learning to intraday or high-frequency trading before we see its application in lower frequencies.

Deep Learning finds immediate use for portfolio managers in an indirect manner. Parking lot images are analyzed using Deep Learning architectures (like convolutional neural nets) to count cars. Text in social media is analyzed using Deep Learning architectures (like long short-term memory) to detect sentiment. Such traffic and sentiment signals can be

integrated directly into quantitative strategies, as shown in earlier sections of this report. Calculation of such signals themselves will be outsourced to specialized firms that will design bespoke neural network architecture for the task.

To understand Deep Learning, we first look at the idea of a neuron. A neuron can be viewed as a simple calculator that computes first the weighted average of inputs given to it and then outputs the result if the value exceeds a pre-determined threshold. Computer scientists could arrange layers of these neurons and connect them to form a neural network as shown below. The neurons at the first layer act upon the user's inputs to produce outputs that are fed to the second layer. The second layer in turn processes its input and feeds its output to the third layer and so on. The last layer produces an output for the user. Just as in classical Machine Learning, a data scientist can use a training set with input-output pairs and use these to calibrate the weights of all the neurons in the neural network.

Figure 85: Example of a Neural Network with 2 hidden layers (4 neurons each), 3 input variables and 1 output variable



Source: J.P.Morgan Macro QDS

The idea of a neural network itself is not new, but the most recent developments in computer science enable us to calibrate the weights of the interconnections efficiently. Computer scientists have found that neural networks not only approximate functions well on the training set, but also generalize well to unseen examples in certain chosen applications like image and text analysis. The reason behind such generalization and the lack of overfitting even in the presence of a large number of parameters remains less clear⁵¹.

In the neural network shown in the picture above, the left-most layer that processes the input is called the **input layer**. The right-most layer is correspondingly called the **output layer**. When a user provides a training set, it is essentially used to provide values for the input and the output layer. Other layers in the middle are called **hidden layers**. The architect of a neural network can decide the number of hidden layers as well as the number of neurons in each hidden layer. Deep Learning refers to the use of neural networks with ‘many’ hidden layers. The word ‘many’ is not fixed; in practice, **3 or more layers is often called “Deep Learning”**. The term “very Deep Learning” is sometimes used for systems with more than 10 hidden layers.

⁵¹ [“Understanding Deep Learning requires rethinking generalization”](#) by Zhang et al (2017) argues that conventional understanding of generalization cannot explain robust performance of neural networks. The paper written by researchers at Google Brain, U C Berkeley and MIT won the Best Paper Award at ICLR 2017 conference. In a mathematical box, we draw connections between a Restricted Boltzmann Machine and quantum field theory; such research might explain better the performance of Deep Learning in practice.

Within a single neuron, it is common to apply a non-linear **activation function** such as $\max(x,0)$ ⁵² to a weighted average of inputs⁵³. After selecting an activation function, one needs to specify the number of hidden layers and the **number of neurons in each layer**.

The number of neurons in the input and output layers are fixed by the dataset we are analyzing. For example, if we are predicting returns for an asset using 10 signals, then we will have 10 neurons in the first layer and one in the last. The number of neurons in each hidden layer can be chosen by the designer; the number is typically chosen to lie between the number of neurons in input and output layers. Prior to the advent of Deep Learning, it was common to use a single hidden layer. It can be shown that any continuous function can be approximated well by a neural network with a single hidden layer. New results on very large data sets show that addition of hidden layers enables the network to learn abstract concepts and features from input data.

Apart from the number of layers and number of neurons, there are other specifications that characterize a neural network. Some of the prominent ones are listed in table below.

Figure 86: Additional attributes that characterize a neural network

Feature of Neural Network	Role in Network Design and Performance	Most Common Example	Other Examples Used in Practice
Cost Function	Used to calculate penalty/error in prediction versus true output	Mean squared error (for regression), Binary cross-entropy (for classification)	Mean absolute error, Categorical cross-entropy, Kullback-Leibler divergence, Cosine proximity, Hinge/Squared-Hinge, log-cosh
Optimizer	Used to calibrate network weights based on error	Stochastic Gradient Descent or SGD	RMSprop ⁵⁴ , Adagrad , Adadelta , Adam /Adamax/ Nestorov-Adam
Initialization Scheme	Used to initialize network weights	Xavier (including Glorot-Normal and Glorot-Uniform)	Ones/Zeros/Constant, Random Normal/Uniform, Variance Scaling, Orthogonal , Le Cun – Uniform , He – Normal/Uniform
Activation Function	Used at the end of each neuron after the weighted linear combination to get non-linear effect	ReLU (for all intermediate layers), Linear (for final layer in regression), Sigmoid (for final layer in classification)	Softmax/Softplus/Softsign, Leaky/Parametrized ReLU, tanh, Hard Sigmoid
Regularization Scheme	Used to penalize large weights to avoid overfitting	Dropout	L1/L2 regularization for kernel, bias and activity

Source: J.P.Morgan Macro QDS

⁵² The $\max(x,0)$ function is called ReLU or rectified linear unit in neural network literature.

⁵³ Other neuron types are used only for bespoke purposes like image prediction and sequence analysis. For an example of image analysis, see the VGG [model](#) of Oxford University that won the ImageNet competition in 2014. Older examples for sequence analysis include the [original](#) LSTM design of Hochreiter and Schmidhuber (1997), [peephole](#) version of Gers and Schmidhuber (2000); newer designs include [Gated recurrent unit](#) of Cho et al (2014), [Depth Gated recurrent unit](#) of Yao et al (2015), [Clockwork recurrent](#) design of Koutnik et al (2014).

⁵⁴ Unpublished, yet popular, algorithm by Geoffrey Hinton of U Toronto/Google is described at [link](#).

Why does Deep Learning work: Insights from Theoretical Physics

The advances in image and speech recognition afforded by Deep Learning methods have motivated researchers to enquire into fundamental reasoning behind advantages of Deep Learning over traditional statistical analysis. In particular, one seeks to understand why and how the addition of hidden layers leads to the improvement in performance over alternative methods employing the same number of parameters.

Research is still nascent in this direction. In Zeiler and Fergus (2014), it was suggested that each hidden layer in a neural network learns a progressively more abstract feature from an image. For example, if the initial layers learn about just edges and colors, intermediate layers can learn progressively larger facial features till the final layer learn the complete human face itself. In formal terms, we are evaluating the object of interest – be it images or speech or (in future) financial time series – at various time/length scales. The initial layer concentrates on microscopic phenomena like boundaries and edges. Each successive hidden layer represents a coarse-graining operation, where we marginalize or integrate out the irrelevant features. This allows us to expand and look at macroscopic fluctuations of just the relevant operators. In Mehta and Schwab (2014), it is pointed out the operation above is identical to the Kadanoff's variational real-space renormalization scheme over spin systems [Kadanoff (1966)].

In this section, we offer a brief summary of the Mehta-Schwab result. While incipient and still inchoate, research in this direction is promising as any insight relating the structure of the universe around us to the working of Deep Learning can potentially lead to improvement in the design of Deep Learning algorithms.

In the Ising model, we have an ensemble of N binary spins $\{v_i\}$ taking values in the set $\{+1, -1\}^N$. Setting temperature to one and referring to the Hamiltonian as $H(\{v_i\})$, we assign probability to a spin configuration through the Boltzmann distribution as

$$P(\{v_i\}) = \frac{e^{-H(\{v_i\})}}{Z},$$

with the partition function

$$Z = \text{Tr}_{v_i} e^{-H(\{v_i\})}$$

serving as the normalizing factor. The Hamiltonian is characterized by couplings $K = \{K_s\}$ between the spin interactions (equivalent to correlations in Deep Learning) as

$$H(\{v_i\}) = -\sum_i K_i v_i - \sum_{i,j} K_{ij} v_i v_j - \dots$$

The free energy is defined as $F^v = -\log(Z)$. In Kadanoff's renormalization scheme, we typically double the lattice spacing. We also introduce hidden spins $\{h_j\}$ to represent the coarse-grained degrees of freedom (or features), where the finer microscopic fluctuations have been averaged out; this yields a new Hamiltonian

$$H_\lambda^{RG}(\{h_j\}) = -\sum_i \tilde{K}_i h_i - \sum_{i,j} \tilde{K}_{ij} h_i h_j - \dots$$

The mapping $\{K\} \rightarrow \{\tilde{K}\}$ is implemented through a function $T_\lambda(\{v_i\}, \{h_j\})$ depending on the variational parameter λ through the expression

$$e^{-H_\lambda^{RG}(\{h_j\})} = \text{Tr}_{v_i} e^{T_\lambda(\{v_i\}, \{h_j\}) - H(\{v_i\})}$$

yielding a corresponding free energy

$$F_\lambda^h = -\log(\text{Tr}_{h_i} e^{-H_\lambda^{RG}(\{h_j\})})$$

To ensure invariance of long-run observables to the coarse-graining procedure, one minimizes the free energy difference as $\min_\lambda F_\lambda^h - F^v$.

Mehta and Schwab relate this to a restricted Boltzmann machine (RBM) used in the context of unsupervised learning. Consider the set of all hand-written black and white images of digits (0 to 9) in the MNIST database. For each pixel i , we denote the spin variable v_i to be either +1 or -1 depending on whether the pixel is on or off. This creates a probability distribution $P(\{v_i\})$ over N variables, which we want our RBM to learn using a set of M hidden spins $\{h_j\}$. One can view the task as one of dimensionality reduction or alternatively as to learn a generative process that creates the data distribution using a set of fewer features/parameters. The interactions between the hidden spins $\{h_j\}$ and visible spins $\{v_i\}$ are modeled through a quadratic energy function

$$E(\{v_i\}, \{h_j\}) = \sum_i b_i v_i + \sum_{i,j} w_{ij} v_i h_j + \sum_i c_i v_i$$

The analogue of the parameter λ in the physics literature is now $\lambda = \{b_j, w_{ij}, c_i\}$. We use the energy function to define

the joint probability of the hidden and visible spins as

$$p_{\lambda}(\{v_i\}, \{h_j\}) = \frac{e^{-E(\{v_i\}, \{h_j\})}}{Z}.$$

We can marginalize the joint distribution to obtain

- For the visible spins, $p_{\lambda}(\{v_i\}) = \text{Tr}_{h_j} p_{\lambda}(\{v_i\}, \{h_j\})$ and set $p_{\lambda}(\{v_i\}) \triangleq \frac{e^{-H_{\lambda}^{RBM}(\{v_i\})}}{Z}$,
- For the hidden spins, $p_{\lambda}(\{h_j\}) = \text{Tr}_{v_i} p_{\lambda}(\{v_i\}, \{h_j\})$ and set $p_{\lambda}(\{h_j\}) \triangleq \frac{e^{-H_{\lambda}^{RBM}(\{h_j\})}}{Z}$.

The analogue of minimization of the difference in free energy in this case is minimization of the Kullback-Leibler divergence between the actual data distribution and the reconstructed distribution, or $\min_{\lambda} D_{KL}(P(\{v_i\}) || p_{\lambda}(\{v_i\}))$, where the Kullback-Leibler divergence (c.f. the notion of mutual information from information theory) is defined as

$$D_{KL}(P(\{v_i\}) || p_{\lambda}(\{v_i\})) = \sum_{\{v_i\}} P(\{v_i\}) \log \left(\frac{P(\{v_i\})}{p_{\lambda}(\{v_i\})} \right).$$

The main result in the Mehta-Schwab paper is to carefully delineate the correspondence between renormalization group and RBM, which leads them to demonstrate an exact mapping between the two by proving that

$$T_{\lambda}(\{v_i\}, \{h_j\}) = -E(\{v_i\}, \{h_j\}) + H(\{v_i\}).$$

Further, it is shown that the Hamiltonians obtained for hidden spins by Kadanoff are identical to the ones obtained in the RBM auto-encoding case, viz.

$$H_{\lambda}^{RG}(\{h_j\}) = H_{\lambda}^{RBM}(\{h_j\}).$$

Such a mapping between renormalization group – a fundamental construct across many disciplines of theoretical physics – and the Deep Learning algorithms opens the path for further investigation into the unexpected success of Deep Learning over the past decade in practical pattern recognition tasks.

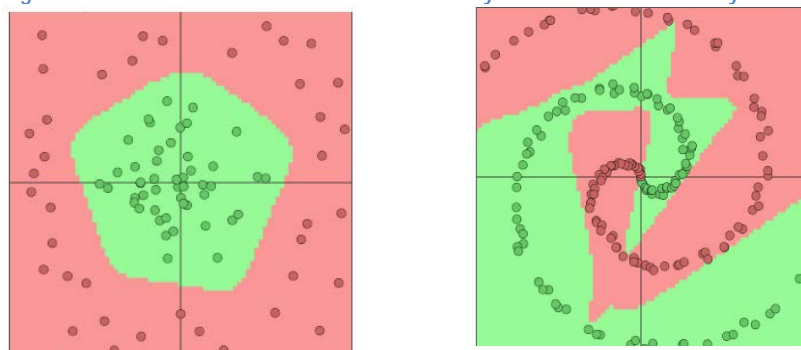
Multi-Layer Perceptron

Multi-layer perceptron models date back to the 1980s. It was made outdated by Support Vector Machines in the early 1990s, due to difficulty with model selection⁵⁵. Below we illustrate MLP on one illustrative classification example, and an example of an actual trading strategy.

Consider the toy model below where the 2-dimensional training set is marked by either green points (representing $y = 0$) or red points (representing $y = 1$). A neural network with 2 hidden layers was trained on the data and asked to predict on each point in the square grid. The neural network is able to partition the space for this toy example.

In the figure below we demonstrate the classification via a 2-layer neural network on a toy dataset (generated using [ConvnetJS](#) package from Stanford University). For the left figure, data was generated with clean separation between green and red; neural nets worked well. For the right figure, data was generated along a spiral path; neural networks could still reasonably well capture the pattern.

Figure 87: Demonstration of classification via a 2-layer neural network on 2 toy datasets



Source: JPM QDS.

MLP Neural Network and trading of Equity sectors

We construct a long-short strategy trading nine US sector ETFs: financials, energy, utilities, healthcare, industrials, technology, consumer staples, consumer discretionary and materials. We use the XGBoost algorithm to predict next day returns based on 8 macro factors: Oil, Gold, Dollar, Bonds, economic surprise index (CESIUSD), 10Y-2Y spread, IG credit (CDX HG), and HY credit spreads (CDX HY).

After obtaining prediction returns for US sectors (based on the previous day's macro factors), we rank sectors by expected return, and go long the top 3 and short the bottom 3 sectors. We used a look-back window of 9 years and rebalance daily. To implement the MLP algorithm, we have used the "Keras" Deep Learning library along with the "Theano" python library. Detailed specification of MLP design and sample code is shown in the box below.

⁵⁵ There are few differences between modern MLP and the MLP of the early 90s. Earlier, activation function was typically sigmoid, chosen due to the differentiability of the logistic function near zero. Nowadays, ReLU or $\max(x, 0)$ is used as the activation function. Earlier, weights on network interconnections were set to either constant or random values. Nowadays, weights are initialized via Xavier or other initialization schemes. These differences do not change the performance of neural nets significantly.

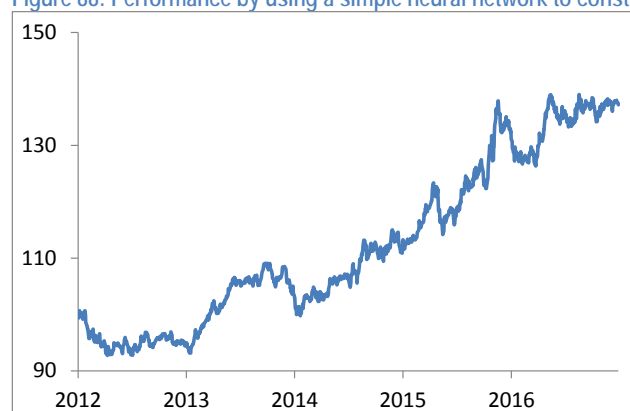
Explanation of Network Architecture Used

- We used a simple neural network with 8 neurons at input layer (for handling the 8 macro factors), 8 neurons in a hidden layer and 1 neuron at output layer (for handling the return predictors).
- Following a suggestion in [Krizhevsky](#) et al (2012), we used ReLU activation at the input and hidden layer; and used a linear activation for the output layer. The paper had found a 6-fold improvement in convergence by choosing ReLU over sigmoid/tanh. Output activation is linear since predicted returns can be either positive or negative.
- For each iteration of the optimizer, we compute the mean squared error between our predictions and actual next day returns and use the RMSProp algorithm to calibrate the parameters. As input to RMSProp, we can feed a batch of inputs instead of single training examples; so we chose a batch-size of 350 samples and repeated the process 50 times (called 'number of epochs' in Deep Learning jargon).
- Following a suggestion in [Srivastava](#) et al (2014), we exclude 20% of units (randomly chosen every time) from each iteration of the parameter calibration algorithm. This technique, called 'dropout regularization', has become popular in the past 3 years as a means to prevent overfitting in neural networks.

```
model = Sequential()
model.add(Dense(8, activation='relu', input_dim=8))
model.add(Dropout(dropout))
model.add(Dense(8, activation='relu'))
model.add(Dropout(dropout))
model.add(Dense(1))
```

Performance of the MLP based sector selection strategy is encouraging. An annualized return of 6.7% and volatility of 8.0%, yielded an information ratio of 0.83 (note that MLP in this scenario slightly underperformed the XGBoost technique, described earlier in this report). While we currently have little intuition about the trading strategy, we note that the correlation of strategy to the S&P 500 was 26.5%, and correlation was also low to other major risk factors such as Momentum, Value, Volatility, and Carry.

Figure 88: Performance by using a simple neural network to construct a dollar-neutral ETF basket



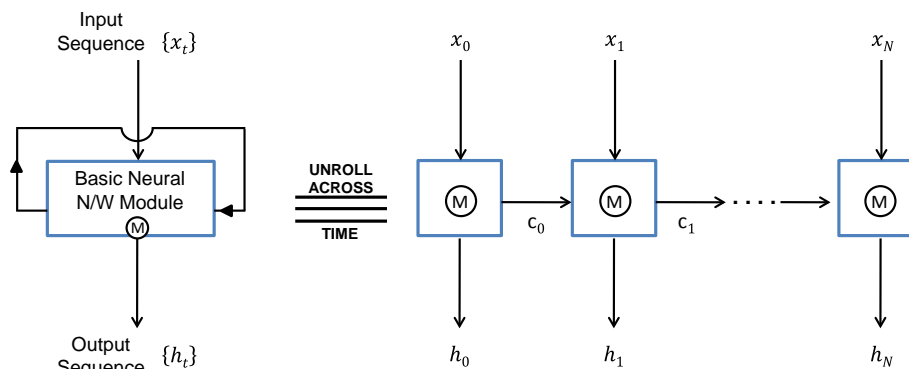
Source: JPM QDS, Bloomberg.

MLP	Bond	Commodity	Equity	FX
Carry	-17.8%	0.5%	5.5%	9.7%
Momentum	-10.4%	0.2%	0.7%	-1.4%
Value	-18.7%	-1.7%	-0.3%	-11.2%
Volatility	0.4%	0.9%	8.4%	8.8%
Beta	-17.0%	7.4%	25.6%	2.7%

Time-Series Analysis: Long Short-Term Memory

Recurrent neural networks (RNNs) were inspired by noting that human beings retain a context of recent events and react to a new input based both on their model of the world and the current context in which the input is received. Traditional neural networks as described in previous sections do not account for this (and are also called feedforward neural networks). One can visualize a simple Recurrent Neural Network as possessing a simple feedback loop. Alternatively, we can unroll the operation of feedback to see how the module retains state across multiple time steps. Both are shown in the figure below.

Figure 89: Recurrent Neural Network: unrolled across time



Source: J.P.Morgan Macro QDS

Econometricians rely on “Autoregressive integrated moving average” (ARIMA) models for forecasting time-series. These work well on small datasets and can accommodate time series memory effects such as persistency, mean reversion, seasonality, etc. Within Deep Learning, the analogue to ARIMA is Long Short-Term Memory (LSTM)⁵⁶. LSTM is a recurrent neural network, i.e. it retains a memory of inputs fed through the network previously. LSTM involves an architectural modification to regular RNN to enable it to remember events over longer horizons.

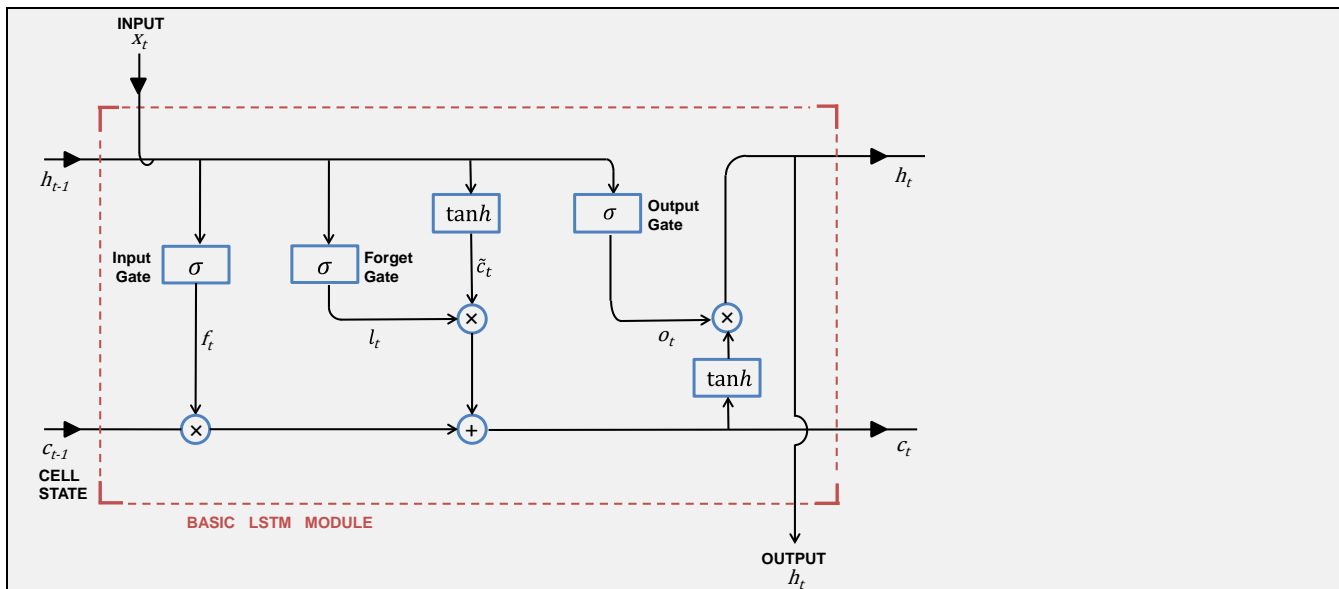
LSTMs have been used in many scenarios, where there are time lags of uncertain length between important events. This includes language translation (sequence of words), connected handwriting recognition (sequence of letters) and video captioning (sequence of images).

While LSTM is designed keeping in mind such time-series, there is little research available on its application to econometrics or financial time-series. After describing the basic architecture of an LSTM network below, we also provide a preliminary example of a potential use of LSTM in forecasting asset prices.

Mathematical Formulation of LSTM block

The basic module in a LSTM network is different from a conventional neuron. Architecture of a single LSTM neuron is illustrated in the figure below:

⁵⁶ Like other neural networks, the concept has been known for long, but the sharp rise in interest is new. The original concept is in Hochreiter, S and Schmidhuber, J (1997), “Long short-term memory”, Neural Computation, Vol 9(8). To gauge recent interest, see Karpathy, A (2015), “The unreasonable effectiveness of recurrent neural networks”, available at [link](#). For a tutorial introduction, see Olah, C (2015), “Understanding LSTM networks”, available at [link](#).



Here, h_t represents the output of the neuron and C_t represents the state of the neuron. The operation can be explained as follows:

- Calculate what needs to be forgotten at the cell state: $f_t = \sigma(W_f [h_{t-1}, x_t] + b_f)$.
- Calculate what needs to be added to cell state via $i_t = \sigma(W_i [h_{t-1}, x_t] + b_i)$ and $\tilde{C}_t = \tanh(W_c [h_{t-1}, x_t] + b_c)$.
- Update the cell state as $C_t = f_t * C_{t-1} + i_t * \tilde{C}_{t-1}$.
- Compute the output via $o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$ and $h_t = o_t * \tanh(C_t)$.

There are many other LSTM designs. Comparison of different architectures is available in [Greff et al \(2015\)](#) and [Jozefowicz et al \(2015\)](#).

Testing LSTM in an S&P 500 timing strategy:

We attempted to use an LSTM neural network to design an S&P 500 trading strategy. The dataset included S&P 500 returns since 2000, and the first 14 years were used for training the LSTM. The last 3 years worth of data were used to predict monthly returns. Some specifications of the LSTM network architecture used are described below.

Architecture of LSTM Network Used for Prediction

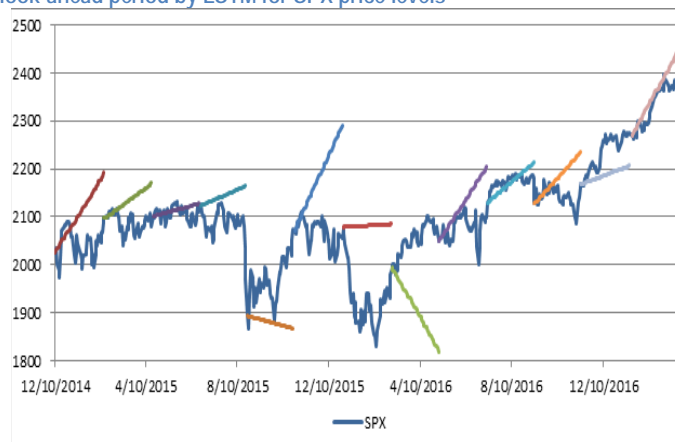
To make these predictions, we fed the scaled input to a 2-layer LSTM network with 50 and 100 neurons respectively in the hidden layers. We used the default LSTM implementation from Keras library and set dropout regularization to 0.2. Following is the code showing the LSTM architecture used for prediction:

```
model = Sequential()
model.add(LSTM(input_dim=1, output_dim=50,
               return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM(100, return_sequences=False))
model.add(Dropout(0.2))
model.add(Dense(output_dim=1))
model.add(Activation("linear"))
model.compile(loss="mse", optimizer="rmsprop")
```

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, None, 50)	10400
dropout_1 (Dropout)	(None, None, 50)	0
lstm_2 (LSTM)	(None, 100)	60400
dropout_2 (Dropout)	(None, 100)	0
dense_1 (Dense)	(None, 1)	101
activation_1 (Activation)	(None, 1)	0

The Figure below illustrates predictions made for the S&P 500 by the LSTM algorithm. The timing strategy yielded positive, but relatively low (insignificant) Sharpe ratio. Results of our initial LSTM designs are not conclusive about the merits of LSTM in forecasting asset prices. We will continue researching the use of LSTM and report more findings in another publication.

Figure 90: Prediction for 2-month look-ahead period by LSTM for SPX price levels



Source: J.P.Morgan Macro QDS

Convolutional Neural Nets

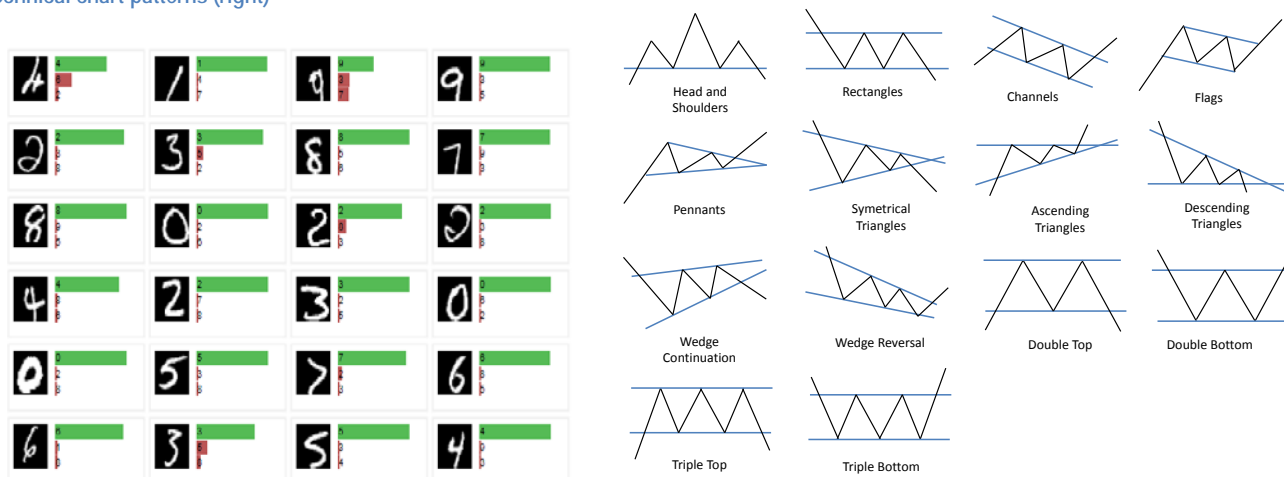
Convolutional neural networks (CNN) are a special type of MLP designed specifically for handling images. The interconnections between neurons in a CNN are loosely inspired by models of human/animal vision. Unlike regular MLPs, the neurons in a CNN are connected only to a small region of their input; moreover, neurons share their weights with other neurons near them. Another difference from regular MLPs is that the input to a CNN is presented as either a 2-dimensional array of numbers (for black/white pictures) or as a 3-dimensional array (for Red/Green/Blue color images). Some well-known CNN architectures have specific names, like LeNet, AlexNet, ZF net, GoogLeNet, VGGNet and ResNet. These network architectures are also used to analyze satellite images; for example, counting cars in parking lots or in estimating metal storage in mines. In signal processing terms, CNN uses many filters (a mechanism to detect edges or color change) which are applied to different parts of the image to identify the object. Unlike earlier signal processing attempts, the weights in the filters are calibrated (i.e. learned) during the training process. It was using CNN that Krizhevsky et al in 2012 argued that hidden layers in neural networks learn progressively more abstract features.

At present, there is little known application of CNN to trading. Below we lay out a hypothesis about a potential application of CNN to trading – specifically in the field of technical analysis. CNN excel and are currently used in detection of objects in images. An example is to identify numbers from hand-written pictures of images. The Figure on the left below shows an example of a CNN output in recognizing hand-written figures. The challenge in detecting hand-written digits was the minor variations that each writer makes in his/her hand-writing. We believe that the image recognition strength of CNN can be used to detect price chart patterns of Technical analysis (the Figure on the right below).

Similar to hand-written digits, there are subtle variations in technical analysts' patterns that are hard to define in a way suitable for time series testing. Patterns in technical analysis are also hard to define mathematically as they can have many variations based on time scale (e.g. head and shoulders can develop on daily, weekly, monthly time horizon, time between head/shoulder). Patterns can also vary based on relative dimensions (e.g. head can be 1.5, 2, 2.5 times the height of 'shoulder', etc.).

Our hypothesis is that various technical patterns (and perhaps specific calls from prominent technical analysts) can be used to train CNNs, and then be rigorously tested (i.e. how much predictive power is in the specific pattern, or even specific analyst). The patterns that show significant forecasting power, can then be industrialized (i.e. automated, applied over a broad range of assets in continuous time) at a scale that would be impossible to achieve by a 'human technical analyst'.

Figure 91: Predictions by Convolutional Neural Nets on a segment of hand-written digits (left); our hypothesis is that CNNs can be applied to technical chart patterns (right)



Source: J.P.Morgan Macro QDS

Restricted Boltzmann Machines

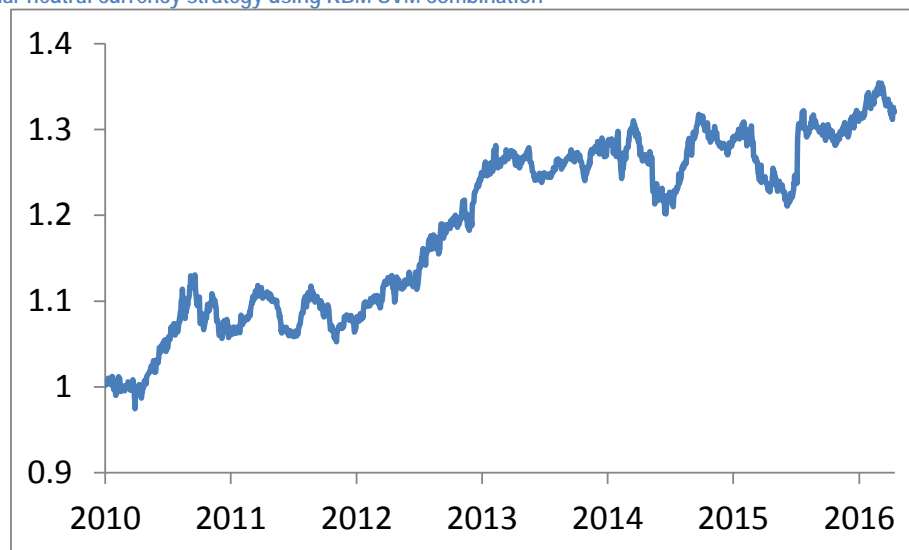
Restricted Boltzmann Machine (RBM) is a neural-network based dimensionality reduction technique (i.e. neural network based unsupervised learning algorithm). The neurons in RBM form two layers called the visible units (reflecting returns of assets) and the hidden units (reflecting latent factors). Neurons within a layer – hidden or visible – are not connected; this is the ‘restriction’ in the restricted Boltzman machine. The architecture is inspired by statistical physics; in particular, the Boltzman distribution (a probability density function) is utilized in the design of the algorithm, lending the name “Boltzmann Machine” to the technique.

FX Trading Strategy implemented with use of RBM

We tested a simple long/short strategy for trading 10 developed market currencies (Australia, Canada, Switzerland, Denmark, Euro, United Kingdom, Japan, Norway, New Zealand and Sweden). For inputs we used the lagged daily returns over the past 10 days for each of the currencies (i.e. we provided 100 input features). We used a rolling window of 252 days to calibrate our Machine Learning model and predicted the next day returns for each of the 10 currencies. We went long the three currencies with the best anticipated next-day returns and short the three currencies with the worst ones. We used a three-step model for prediction: we first scaled the rolling input (mean zero and variance one), then reduced the dimensionality via a Restricted Boltzmann Machine to 20, and then applied Support Vector Regressor setting Kernel as RBF (Radial basis function). For the support vector regressor and RBM, we used implementations in *sklearn*; setting $C=1$ (regularization in SVM) and $\gamma = 0.0001$ (for RBF kernel bandwidth).

The long-short strategy had an annualized mean of 4.5%, annualized volatility of 6.7% yielding an IR of 0.67. Correlation to S&P 500 returns over the same period was 13.8% and correlation to DXY was -6%.

Figure 92: PnL of dollar neutral currency strategy using RBM-SVM combination



Source: J.P.Morgan Macro QDS, Bloomberg.

Reinforcement Learning

An especially promising approach to Machine Learning is reinforcement learning. The goal of reinforcement learning is to choose a course of actions that maximizes some reward. For instance, one may look for a set of trading rules that maximizes PnL after 100 trades. Unlike supervised learning (which is typically a one-step process), the model doesn't know the correct action at each step, but learns over time which succession of steps led to the highest reward at the end of the process.

A fascinating illustration is a performance of the reinforcement learning algorithm in playing a simple [Atari videogame](#) (Google's DeepMind). After training the algorithm on this simple task, the machine can easily beat a human player. While most human players (and similarly traders) learn by maximizing rewards, humans tend to stop refining a strategy after a certain level of performance is reached. On the other hand, the machine keeps on refining, learning, and improving performance until it achieves perfection.

At the core of reinforcement learning are two challenges that the algorithm needs to solve: 1) Explore vs. Exploit dilemma – should the algorithm explore new alternative actions that may not be immediately optimal but may maximize the final reward (or stick to the established ones that maximize the immediate reward); 2) Credit assignment problem – given that we know the final reward only at the last step (e.g. end of game, final PnL), it is not straightforward to assess which step during the process was critical for the final success. Much of the reinforcement learning literature aims to answer the twin questions of the credit assignment problem and exploration-exploitation dilemma.

When used in combination with Deep Learning, reinforcement learning has yielded some of the most prominent successes in machine learning, such as self-driving cars. Within finance, reinforcement learning already found application in execution algorithms and higher-frequency systematic trading strategies.

Reinforcement learning has attributes of both supervised and unsupervised learning. In supervised learning, we have access to a training set, where the correct output “y” is known for each input. At the other end of the spectrum was unsupervised learning where we had no correct output “y” and we are learning the structure of data. In reinforcement learning we are given a series of inputs and we are expected to predict y at each step. However, instead of getting an instantaneous feedback at each step, we need to study different paths/sequences to understand which one gives the optimal final result.

Mathematical Details and Deep Q-Learning Algorithm

Earlier in the report, we studied Hidden Markov Models, where the underlying environment/system was partially observable, but autonomous. In the case of reinforcement learning, we retain the assumption of partial observability of the system, but now actively endeavor to control the outcome. In this section, we give a brief introduction to deep Q-learning, which is an approach to reinforcement learning using neural networks. We do not discuss mathematical details corresponding to (Partially Observed) Markov Decision Processes.

Let the sequence of states, actions and rewards be denoted by $s_0, a_0, r_0, \dots, s_n, a_n, r_n$. From any step t in the process, one can define a discounted future reward as

$$R_t = r_t + \gamma r_{t+1} + \dots + \gamma^{n-t} r_n.$$

In Deep Q-learning, we define a Q-function to denote the maximum future reward from performing a certain action now, i.e.

$$Q(s_t, a_t) = \max R_{t+1}.$$

In the case where the Q-function is known, the optimal action is trivial to determine using $\pi(s) = \arg \max_a Q(s, a)$. The catch is that the Q-function is not known a priori. The classical solution for the problem is to iteratively approximate the Q-function using Bellman's equation, which claims that

$$Q(s_t, a_t) = r_t + \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}).$$

The Bellman equation simply says that the maximum reward for an action in the current state is equal to the sum of the immediate reward and the maximum reward possible from the succeeding state.

If the number of states and actions were small, the above equation would work fine. In many cases, the number of states and actions are large. This motivates the use of a neural network to learn the Q-function. A neural network can process a large number of states and extract the best features from them. Often, it may be more efficient in learning the function than a mere tabular approach that matches values of state and action pairs with possible value output. Putting all this together leads to the deep Q-learning algorithm.

Algorithm 1 Q-Learning

```

1: 1. Initialisation:
   Load a simulation environment: price series, fill probability;
   Initialise the value function  $V_0$  and set the parameters:  $\alpha, \epsilon$ ;
2: 2. Optimisation:
3: for  $episode = 1, 2, 3, \dots$  do
4:   for  $t = 1, 2, 3, \dots T$  do
5:     Observe current state  $s_t$ ;
6:     Take an action  $a_t(Q_t, s_t, \epsilon)$ ;
7:     Observe new state  $s_{t+1}$ ;
8:     Receive reward  $r_t(s_t, a_t, s_{t+1})$ ;
9:     Update value function using  $r_t$  and current estimate  $Q_t$ :
       a) compute  $y_t = r_t + \max_a Q_t(s_{t+1}, a)$ 
       b) update the function  $Q_t$  with target  $y_t$ 
10:   end for
11: end for

```

Reinforcement Learning in Algorithmic Trading

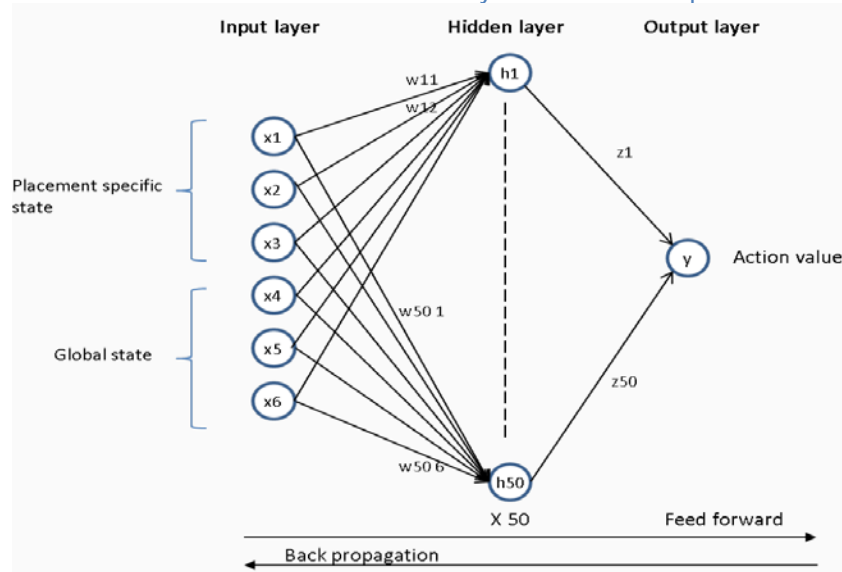
The J.P.Morgan Electronic Trading group developed algorithms that use reinforcement learning. An execution algorithm consists of a scheduler that makes macro decisions across a time horizon of hours, and a limit order placement module (LOPM) that makes micro decisions across a time horizon of seconds to minutes.

LOPM is based on a reinforcement learning model. Given the task of selling or buying a given quantity within a stated horizon, the LOPM module aims to minimize slippage and complete the task within some tolerance. The module is given market size/price data (e.g. spreads), execution data (e.g. filled quantity and realized p.o.v. rate), market signals (e.g. medium-frequency like momentum and high-frequency like order flow) and model estimates (e.g. volume/volatility prediction and fill probability estimate). The model operates under constraints on schedule (e.g. quantity and time horizon), order (e.g. limit price), and market conditions (e.g. tick size). Constraints can also arise from client (e.g. client aversion to risk) or from model parameters (e.g. evaluation frequency). Under these constraints, the LOPM decides to place either aggressive (i.e., cross the spread) or passive orders (with an optimal quantity placed at each price level of the orderbook).

In execution algorithms operating over multiple time-periods, the complete market impact of one's actions depends not only on individual transactions at the time of execution, but also on the propagation of latent terms across time. This implies that the rewards attributable to a certain step can be discerned only at the end of all transactions. Such delayed-reward knowledge calls for the use of reinforcement learning.

Further, in reinforcement learning, one does not aim to directly learn the optimal action for each state. Instead, one learns the value of each ("state", "action") pair through a "value" function. The precise functional mapping is learned via a neural network with one hidden layer. Such use of neural networks to learn the value function within the framework of reinforcement learning is called Deep Reinforcement Learning.

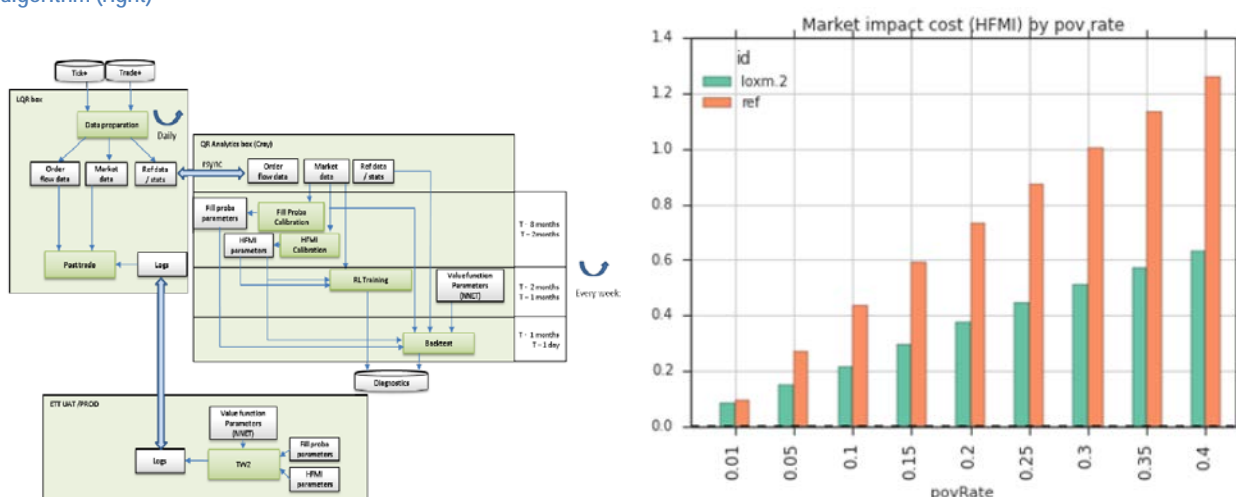
Figure 93: Neural Network used to learn the value function has one hidden layer with 50 nodes. Output node uses ReLU activation function



Source: JPM Linear Quantitative Research Team.

To enhance the performance of the system, one can derive signals from supervised learning models to predict price moves. Combining all these algorithms, the final architecture involved training over 640 CPU cores and 4 X 2880 GPU cores. Finally, backtest statistics shows the improvement in market impact accruing from the use of the reinforcement learning algorithm. In comparison to a reference strategy, the use of reinforcement learning has reduced market impact at every level of p.o.v. (percentage of volume) rate.

Figure 94: Architecture for execution algo trading using reinforcement learning (left); Outperformance of Reinforcement learning execution algorithm (right)



Source: JPM Linear Quantitative Research Team.

Comparison of Machine Learning Algorithms

In this section we compare the performance of different algorithms on 3 typical Machine Learning tasks. The first example demonstrates the performance of Machine Learning algorithms when applied on a typical **supervised learning regression** problem. In this example we study our equity sector trading strategy and attempt to forecast daily sector returns based on changes in various macro variables. This example illustrates supervised learning tasks most commonly tackled by methods of regression. The second example is a comparison of the performance of various Machine Learning algorithms when applied on the **supervised learning task of classification**. The last example demonstrates the performance of different Machine Learning algorithms when applied to the **unsupervised learning task of clustering**. In both classification and clustering examples we used computer simulated data for illustrative purposes (rather than actual financial data). Please note that in the sector trading example we used both methods of regression (to forecast returns) and classification (e.g. to forecast if the sector will go up or down).

Comparison of Algorithms: Supervised Learning - Regression

We use different supervised learning methods to model the next-day returns of US sectors using a series of macro indicators. We construct a long-short strategy trading nine sectors: financials, energy, utilities, healthcare, industrials, technology, consumer staples, consumer discretionary and materials. We use different Machine Learning algorithms to predict next day returns based on 8 macro factors: Oil, Gold, Dollar, Bonds; economic surprise index (CESIUSD), 10Y-2Y spread, IG credit and HY credit spreads (this is the same trading example that we used in the XGBOOST and MLP sections)

After fitting the model, we predicted either the returns for the next day (while using a regressor model) or the probability of a positive return (while using a classifier model), and used this as our signal. On each day we ranked the assets by the signal, and went long the top 3 and short the bottom 3 sector ETFs. For most models we use default values of the hyper-parameters and calibrated only the parameters through maximum likelihood. For some models, where the computational load was lower, we fit the hyper-parameters through 5-fold cross-validation. Note that the model fit focuses on return alone, so it is possible the strategy using a cross-validated model can have a lower Sharpe than a model without cross-validation.

Note that we use a rolling window to fit models each day, for each of the sectors. This is analogous to traditional quantitative backtesting. In machine learning, it is conventional to have a simple train-test split, where the model is trained on a fixed subset of initial samples and then tested over all the latter ones. Such a split may be sub-optimal in financial time-series where the relationship between variables changes (hence we use a rolling window to calibrate). We tested the strategies starting in 2003, and a comparison of Sharpe Ratios for various strategies is shown below:

Figure 95: Equity Sector Trading Strategy - performance of different supervised learning algorithms

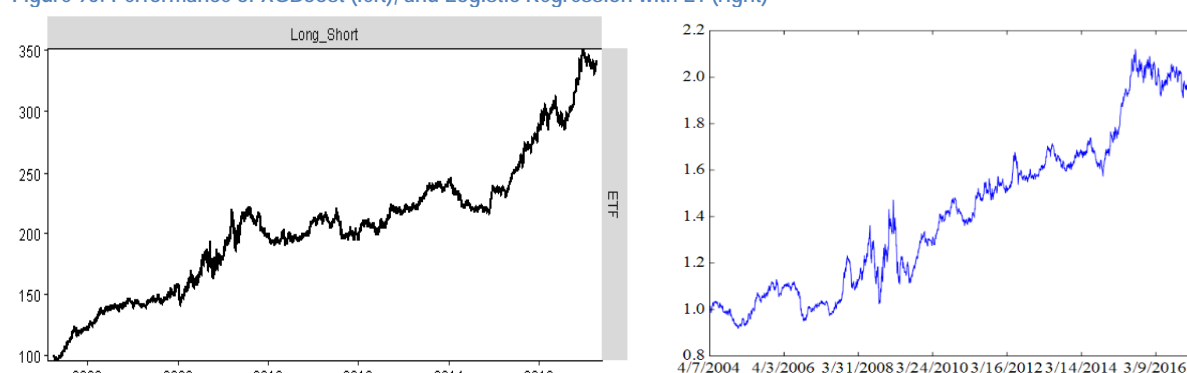
	Type	Hyper-parameter Fit	Model	Annualized Return (%)	Sharpe ratio
1	Regressor	Cross Validation	XGBoost	10.97	0.89
2	Classifier	Default	Logistic Regression – L1 regularization	5.90	0.52
3	Classifier	Default	Logistic Regression – L2 regularization	4.59	0.40
4	Classifier	Default	Linear Discriminant Analysis	4.34	0.38
5	Regressor	Default	Support Vector Regression – RBF kernel	4.64	0.36
6	Regressor	Cross Validation	Elastic Net	4.35	0.33
7	Regressor	Default	Decision Tree	3.14	0.27
8	Classifier	Default	Gaussian Naïve Bayes	3.19	0.27
9	Regressor	Cross Validation	Support Vector Regression – RBF kernel	3.39	0.27
10	Regressor	Cross Validation	Lasso	3.44	0.27
11	Classifier	Default	Random Forest -- Trees = 25)	2.59	0.24
12	Classifier	Default	Support Vector Classification – RBF kernel	2.50	0.21
13	Regressor	Default	Random Forest Regressor	1.95	0.17

14	Classifier	Default	Decision Tree	1.14	0.10
15	Regressor	Cross Validation	Logistic Regression – L2 regularization	0.55	0.05
16	Classifier	Default	Quadratic Discriminant Analysis	0.43	0.04
17	Regressor	Default	Ridge	-0.12	-0.01
18	Regressor	Cross Validation	Ridge	-0.68	-0.05
19	Classifier	Default	Decision Tree	-0.76	-0.07
20	Regressor	Default	Lasso	-1.17	-0.08
21	Classifier	Default	Support Vector Classification - Linear	-1.52	-0.13
22	Regressor	Default	Elastic Net	-2.46	-0.17

Source: J.P.Morgan Macro QDS

Results were best for XGBoost which yielded a Sharpe ratio of 0.89. XGBoost effectively averages over a number of different strategies (decision trees) so perhaps its strong risk-adjusted performance should not come as a surprise. In addition to XGBoost, strong performance is recorded by quasi-linear models like logistic regression and linear discriminant analysis. Performance of logistic regression with L1 regularization is plotted below (right).

Figure 96: Performance of XGBoost (left), and Logistic Regression with L1 (right)



Source: J.P.Morgan Macro QDS

To gain some more insights about different algorithms, we calculated the correlation of ML strategies to broad equity factors, as well as the correlation of strategies between themselves. The figure below shows the correlation of different Machine Learning algorithms to equities, as well as to main equity long-short styles: momentums, value, volatility and carry. One can see that strategies exhibited relatively low (often negative) correlation to equity styles.

Figure 97: Correlation of Machine Learning strategies to equity styles

Type	Hyper-parameter Fit	Model	SPX Index	Carry - Equity	MoM - Equity	Value - Equity	Volatility - Equity
1 Regressor	Cross Validation	XGBoost	6.0%	-4.0%	2.3%	-4.2%	-0.5%
2 Classifier	Default	Logistic Regression – L1 regularization	-9.7%	7.5%	-11.7%	-13.1%	-0.8%
3 Classifier	Default	Logistic Regression – L2 regularization	-8.7%	4.8%	-11.0%	-12.0%	-3.3%
4 Classifier	Default	Linear Discriminant Analysis	-11.1%	9.1%	-12.3%	-12.9%	-1.6%
5 Regressor	Default	Support Vector Regression – RBF kernel	-22.4%	7.6%	-12.7%	-33.7%	-2.9%
6 Regressor	Cross Validation	Elastic Net	-29.1%	-1.2%	-13.1%	-38.7%	-7.3%
7 Regressor	Default	Decision Tree	3.2%	4.5%	0.9%	3.3%	-1.6%
8 Classifier	Default	Gaussian Naïve Bayes	-13.2%	-6.0%	-6.6%	-13.0%	-8.0%
9 Regressor	Cross Validation	Support Vector Regression – RBF kernel	-17.2%	9.0%	-11.6%	-32.2%	1.1%
10 Regressor	Cross Validation	Lasso	-28.2%	0.1%	-11.7%	-37.0%	-6.5%
11 Classifier	Default	Random Forest -- Trees = 25)	-0.5%	-8.9%	-1.1%	-1.9%	-8.5%
12 Classifier	Default	Support Vector Classification – RBF kernel	-24.7%	3.3%	-6.8%	-28.2%	-0.1%
13 Regressor	Default	Random Forest Regressor	3.3%	0.6%	0.6%	-2.7%	-3.6%

14	Classifier	Default	Decision Tree	25.1%	-5.0%	3.5%	11.5%	5.2%
15	Regressor	Cross Validation	Logistic Regression – L2 regularization	-26.0%	6.5%	-14.7%	-30.9%	-3.4%
16	Classifier	Default	Quadratic Discriminant Analysis	-1.5%	-4.4%	-1.9%	-5.2%	1.0%
17	Regressor	Default	Ridge	2.9%	4.3%	5.0%	-6.0%	3.4%
18	Regressor	Cross Validation	Ridge	0.1%	5.1%	4.7%	-8.4%	1.3%
19	Regressor	Default	Lasso	-34.9%	5.5%	-13.3%	-48.3%	-0.8%
20	Classifier	Default	Support Vector Classification - Linear	-28.1%	-2.0%	-0.7%	-24.5%	-3.6%
21	Regressor	Default	Elastic Net	-32.6%	7.1%	-14.5%	-48.0%	0.5%

Source: J.P.Morgan Macro QDS

The figure below shows the correlation of different Machine Learning algorithms between themselves. One can notice that regression-based strategies tend to be more correlated (e.g. Lasso CV is strongly correlated to Elastic Net, etc.); however, on average the pairwise correlation of all strategies was only ~20%, which offers the potential to combine strategies at a portfolio level.

Figure 98: Correlation of Machine Learning strategies amongst themselves

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
XGBoost	1	--																			
Logistic – L1	2	20%	--																		
Logistic – L2	3	16%	93%	--																	
LDA	4	18%	92%	93%	--																
SVR – RBF kernel	5	26%	53%	51%	57%	--															
Elastic Net CV	6	19%	28%	25%	26%	45%	--														
Regressor Tree	7	14%	13%	8%	7%	-1%	3%	--													
Gaussian Naïve Bayes	8	27%	46%	47%	46%	36%	23%	15%	--												
SVR – RBF kernel CV	9	29%	48%	44%	51%	82%	52%	3%	31%	--											
Lasso CV	10	21%	31%	27%	29%	46%	96%	7%	25%	53%	--										
Random Forest Classifier	11	11%	29%	26%	24%	15%	11%	16%	23%	14%	10%	--									
SVC – RBF kernel	12	6%	23%	22%	26%	53%	34%	-10%	11%	49%	34%	1%	--								
Random Forest Regressor	13	18%	15%	13%	12%	2%	8%	28%	12%	9%	10%	18%	-4%	--							
Classification Tree	14	12%	16%	14%	11%	3%	-3%	10%	11%	5%	-3%	22%	-14%	9%	--						
Logistic – L2 CV	15	19%	57%	57%	62%	66%	43%	-3%	35%	65%	44%	18%	54%	4%	3%	--					
QDA	16	18%	34%	37%	36%	24%	13%	10%	50%	21%	16%	20%	10%	16%	12%	26%	--				
Ridge	17	18%	35%	34%	34%	26%	32%	12%	15%	27%	33%	12%	9%	13%	12%	22%	18%	--			
Ridge CV	18	17%	36%	34%	36%	28%	33%	12%	15%	28%	34%	13%	9%	14%	11%	23%	19%	98%	--		
Lasso	19	19%	25%	24%	30%	69%	64%	-3%	22%	62%	65%	3%	53%	1%	-10%	53%	20%	18%	20%	--	
SVC - Linear Kernel	20	11%	20%	19%	25%	48%	29%	-3%	21%	48%	30%	4%	50%	-2%	-4%	51%	23%	7%	8%	51%	--
Elastic Net	21	19%	27%	26%	32%	68%	63%	-3%	22%	62%	64%	3%	52%	2%	-10%	53%	20%	19%	22%	98%	48%

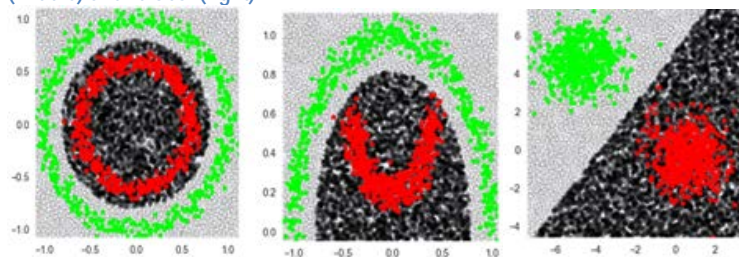
Source: J.P.Morgan Macro QDS

Comparison of Algorithms: Supervised Learning - Classification

In this section we compare the performance of different classification algorithms. In particular, we wanted to test how robust the methods are when we change the shape, sample size, and level of noise in the data. We applied algorithms to 3 different computer generated datasets shown in the figure below. While there is no direct financial interpretation to these datasets, it is not difficult to envision their applications. For instance 'green' points can be 'buy' signals, and 'red' points can be 'sell' signals. In the example with 'circles' (left) – a buy signal would be triggered when 2 asset prices take large (positive or negative) values symmetrically, and sell otherwise (e.g. we would buy VIX when jointly the USD is too low or too high, and an inflation reading is also too low or too high). The dataset on the right with two 'blobs' could be interpreted as a buy signal being triggered when the horizontal variable has a low reading, and the vertical variable a high reading (e.g. low equity volatility and high equity price momentum).

We then trained a classifier to learn the location of green and red points, i.e. the classifier learns a mapping between the (x,y) coordinates. For testing, we evaluated the classifier on all points within a rectangular grid surrounding the two circles. Test sample points that were classified by the algorithm as green were marked as white and those classified as red were marked as black. If our classifier was perfect, we would expect to see the pictures that appear below:

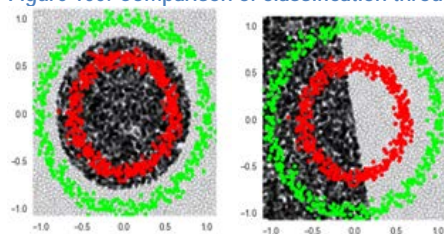
Figure 99: Computer generated patterns used for comparing performance of classification and clustering algorithms: 'circles' (left), 'curves' (middle) and 'blobs' (right)



Source: J.P.Morgan Macro QDS

We start with a simple training set which has two circles – large green circle containing a smaller red circle (with added random noise). Successful classification was obtained by a non-linear classifier (left, note 'black' prediction coincides with the 'red' test sample, and 'white' prediction coincides with the 'green' test sample). If we used a linear classifier instead we obtain a poor result in the sense that the classifier incorrectly attributes white to green, and black to red points (figure below right).

Figure 100: Comparison of classification through non-linear (left) vs. linear (right) classifiers



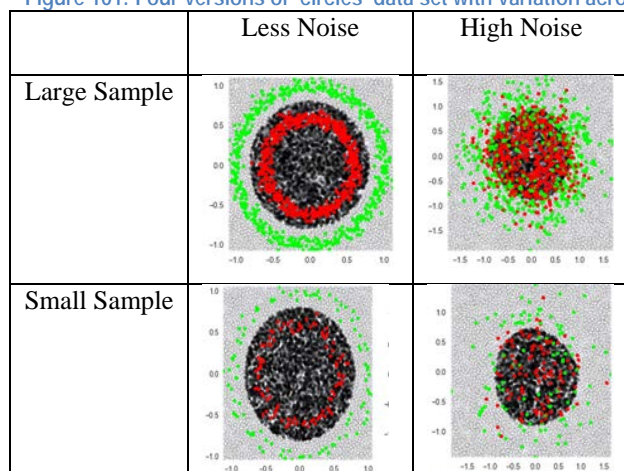
Source: J.P.Morgan Macro QDS

We also want to evaluate different classifiers to four versions of the dataset with:

- Large sample ($N = 1000$), low noise ($\sigma = 0.05$)
- Large sample ($N = 1000$), high noise ($\sigma = 0.30$);
- Small sample ($N = 200$), low noise ($\sigma = 0.05$);
- Small sample ($N = 200$), high noise ($\sigma = 0.30$)

These data samples are illustrated below (where they are analyzed with a non-linear classifier – SVM with a radial basis function kernel).

Figure 101: Four versions of 'circles' data set with variation across sample size and noise level



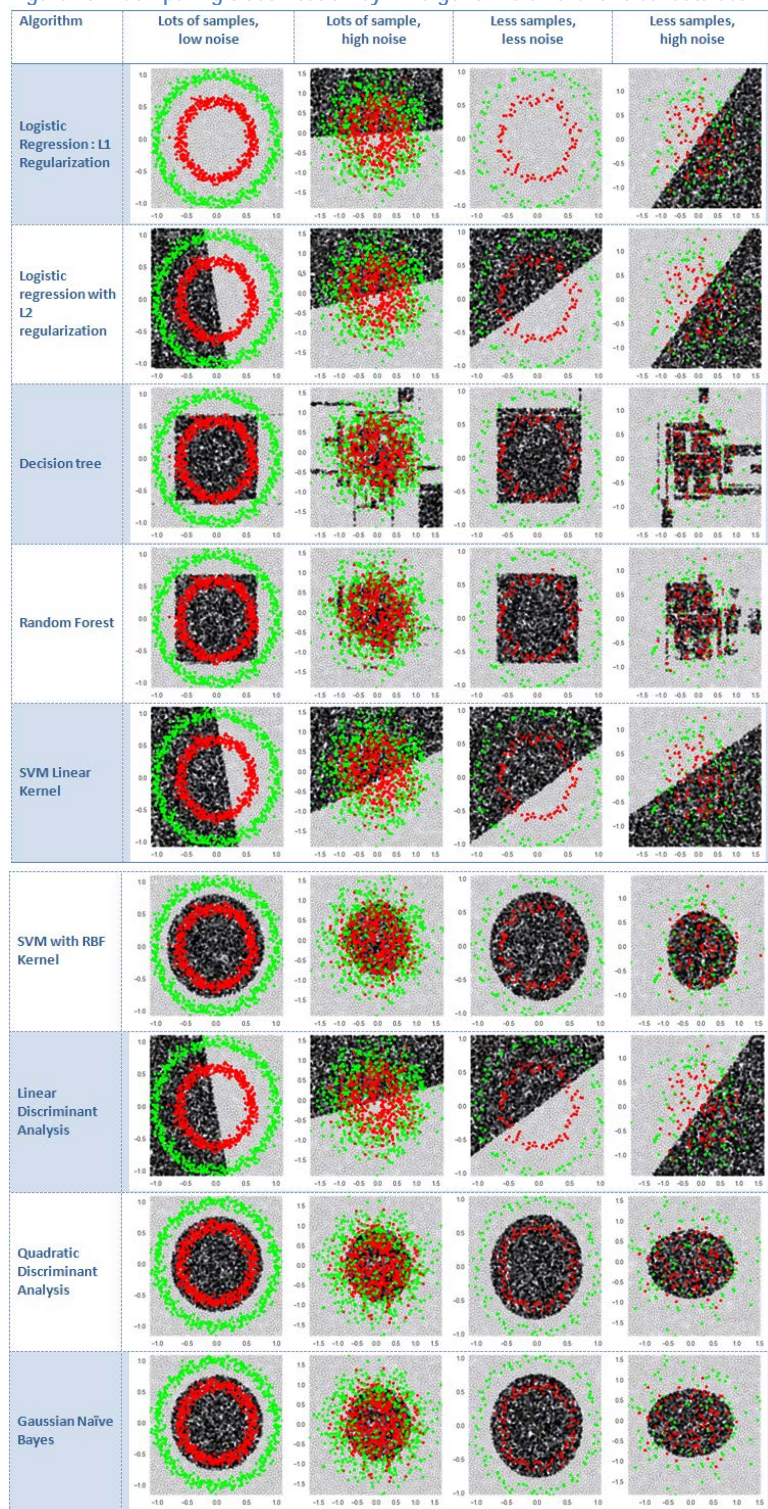
Source: J.P.Morgan Macro QDS

We tested the following classification algorithms on the 'circles' dataset as described above:

1. Logistic regression with L1 regularization – fits a linear classifier while trying to use as few input features as possible.
2. Logistic regression with L2 regularization – fits a linear classifier trying to keep magnitude of weights roughly similar.
3. Decision tree – fits a non-linear classifier (analogous to coding rules using if-else statements). This is similar to how human beings think about a decision – a decision is taken based on yes/no answers to a series of questions. Being non-linear, it would outperform linear classifiers in the presence of a higher number of samples; but conversely suffers from overfitting in the presence of noise.
4. Random forest – fits many decision trees and averages their output to predict the answer. This should be more robust to noise than decision trees.
5. Support Vector Machine with Linear Kernel – this is another linear classifier, not unlike logistic regression.
6. Support Vector Machine with Radial Basis Function Kernel – this maps the points into an infinite-dimensional space and fits a linear classifier in that space. In the original space, this leads to a non-linear fit. Alongside random forests, these are widely used for classification.
7. Linear Discriminant Analysis – this is a generative technique (like Quadratic Discriminant Analysis and Gaussian Naïve Bayes). Instead of just modeling the probability of output given the input, these methods model the p.d.f. of the input as well. In LDA, the input features are assumed to have a Gaussian p.d.f. with the same covariance matrix. This leads to a linear classifier. LDA makes a stringent assumption of Gaussianity of input, unlike logistic regression.
8. Quadratic Discriminant Analysis – this is similar to Linear Discriminant Analysis, but assumes different covariances for different input features. This is more general than LDA and can fit a quadratic curve as a decision boundary.
9. Gaussian Naïve Bayes – this is a simplification of Quadratic Discriminant Analysis, where we assume a diagonal covariance matrix. Given the output label, this assumes that input features are conditionally independent. The simplicity of its assumptions should make it robust to noise.

The results on applying the above nine classifiers on the circles dataset are tabulated in the figure below.

Figure 102: Comparing classification by ML algorithms on the 'circles' data set



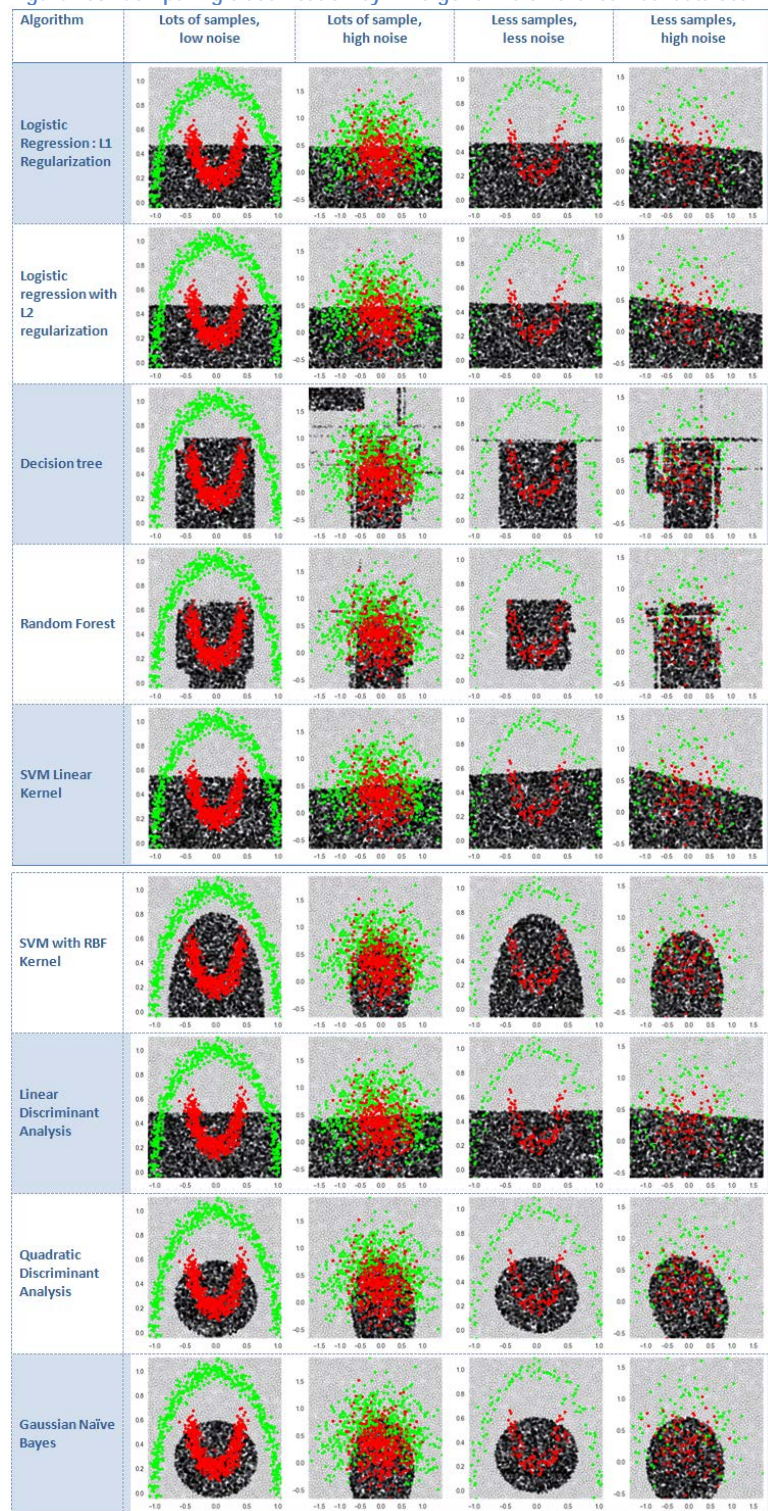
Source: J.P.Morgan Macro QDS

Some of the observations are listed below:

- Logistic regression with L1 regularization tries to enforce stringent sparsity. Sparsity induces the algorithm to choose only one feature out of the two inputs in the case of lots of samples and low noise; this results in the entire region ascribed to green and hence colored white.
- Logistic regression with L2 regularization: the answer is predictably incorrect in all four scenarios. The fit is unstable as noise is increased and the sample size is decreased.
- Decision tree: in the case of low noise (with both large and small sample set sizes), we see that the decision tree is able to classify points accurately. Once noise is added, the decision tree overfits significantly and performs worse than any other classifier.
- Random forest: as an ensemble classifier built over decision trees, it shares the benefits of decision trees and manages to avoid the extreme overfitting. Still, random forest also overfits in the presence of noise.
- SVM with linear kernel: as a linear classifier, its performance is similar to logistic regression with L2 regularization.
- SVM with RBF kernel: given sufficient samples and low noise, the classifier offers the best fit amongst all our classifiers. It is robust to noise as well as reduction in sample size.
- Linear Discriminant Analysis (LDA): behavior as a linear classifier is similar to that of logistic regression.
- Quadratic Discriminant Analysis (QDA): for this dataset, a quadratic decision boundary suffices to separate the data. The fit is remarkably robust to noise and sample size reduction.
- Gaussian Naïve Bayes: this performs as well as QDA. Further, it is robust to noise – a feature that can serve in financial analysis.

We next analyze a data set with two curves intertwined with one another (Figure 99 middle – ‘curves’). Interpretation in this case would be that we trigger a buy signal if the vertical variable is above average, or when the vertical variable is below average and horizontal variable at extreme (e.g. buy VIX when the cross-asset volatility is above average, or when the inflation reading is too low, or too high). As before, we color them as green and red. After training our classifier to locate them, we ask the classifier to map all points in the rectangular area to either the green or red zones. Running again the entire list of supervised learning algorithms on the curves dataset, we get the following results:

Figure 103: Comparing classification by ML algorithms on the 'curves' data set



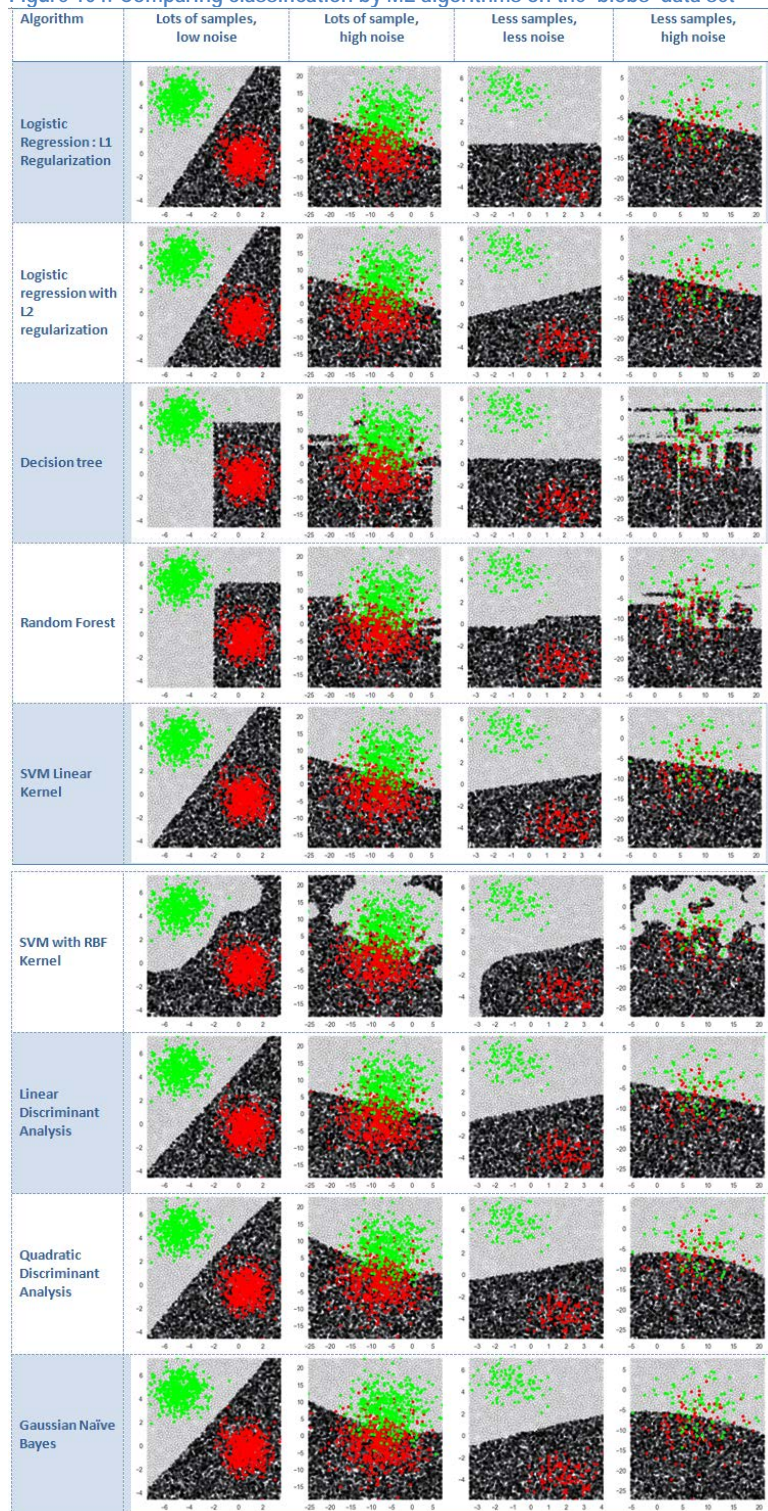
Source: J.P.Morgan Macro QDS

Some of the observations are listed below:

- Logistic regression with L1 regularization: a linear fit is inaccurate on this highly non-linear data sample. However, note that the fit is robust to both an increase in noise and reduction in sample size, and captures the main feature ('buy' if vertical variable is above average)
- Logistic regression with L2 regularization: the results are similar to the fit with L1 regularization. We see this similarity between L1 and L2 regularization across many data sets.
- Decision tree: the non-linear fit enables the algorithm to learn the decision boundary. It holds up well when either the sample size is reduced or if the noise is increased. However, when faced with both high noise and less samples, the algorithm overfits.
- Random forest: performance of random forest is an average across performances of many decision trees. It has less overfitting than decision trees, but its performance in the best case scenario (large sample size, low noise) is also less perfect.
- SVM with linear kernel: performance is similar to logistic regression with L2 regularization.
- SVM with RBF kernel: performance is again the best amongst all our classifiers when applied to non-linear datasets. It is remarkably robust to noise increase or sample size decrease.
- Linear Discriminant Analysis: performance is again similar to logistic regression.
- Quadratic Discriminant Analysis: a quadratic fit (a circle here) is arguably better than the linear fit. It is also reasonably robust to noise and sample size reduction.
- Gaussian Naïve Bayes: algorithm performance is similar to QDA in spite of using much fewer parameters. Given a choice, one would choose Naïve Bayes over LDA and QDA for such datasets.

The third and last of our datasets just comprises of two 'blobs' (Figure 99 right – 'blobs'). The data set could be interpreted as a buy signal being triggered when the horizontal variable has a low reading, and vertical variable a high reading (e.g. low equity volatility and high equity price momentum). The ideal answer in this case is a linear classifier, unlike in the previous two datasets. One expects an ideal classifier to cleanly separate the green and red blobs as shown below. Running our list of supervised learning algorithms on the blobs data set, we get the results depicted below.

Figure 104: Comparing classification by ML algorithms on the 'blobs' data set



Source: J.P.Morgan Macro QDS

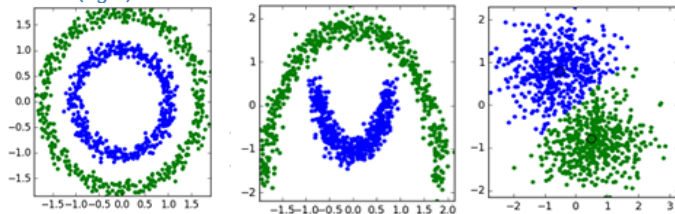
Some of the observations are listed below:

- Logistic regression with L1 regularization: the dataset was suitable for linear classifiers, and a logistic regression with sparsity constraint classifies correctly. It is not robust to noise, but that weakness is shared with all the other classifiers on our list. It is actually better than decision trees and SVM with RBF kernel in the case of a low sample size and high noise.
- Logistic regression with L2 regularization: performance is similar to logistic with L1 regularization.
- Decision tree: a rectangular decision region (caused by two if-conditions, one each on each input variable) is not the optimal answer. Further, it deteriorates considerably under noise.
- Random forest: for this dataset, it shares the weakness of decision tree models.
- SVM with linear kernel: for this dataset, the linear kernel outperforms the RBF kernel. Its robustness to noise is similar to that seen for logistic regression.
- SVM with RBF kernel: unlike the previous two datasets, SVM with RBF kernel overfits in all four scenarios. In data tailor made for linear predictions, complicated models invariably overfit.
- Linear Discriminant Analysis: performance is similar to logistic regression.
- Quadratic Discriminant Analysis: in a linearly separable dataset, fitting a quadratic boundary offers no additional benefit. So QDA reduces to LDA here.
- Gaussian Naïve Bayes: in this dataset, given the output variable, the input variables are actually conditionally independent. So predictably, Gaussian Naïve Bayes outperforms other algorithms and matches the simple logistic regression.

Comparison of Algorithms: Unsupervised Learning - Clustering

In our last comparison, we tested the performance of different clustering algorithms over the same computer generated hypothetical data sets: 'circles', 'curves', and 'blobs'. Unlike the case study of classification, the points in the training set are not marked with output labels (i.e. they are not colored 'red' and 'green' as in the classification test). The aim of clustering algorithms is to find a way to group the points (according to some measure of similarity) and assign them different colors. If we had ideal clustering algorithms, the algorithm would assign the 'colors' to the three datasets as below (here we used green and blue, to differentiate from the classification example).

Figure 105: Computer generated patterns used for comparing performance of clustering algorithms: 'circles' (left), 'curves' (middle) and 'blobs' (right)



Source: J.P.Morgan Macro QDS

As with the case study using supervised learning algorithms, we test each unsupervised learning algorithm over four versions of the datasets with high/low noise levels and large/small sample sizes.

In the jargon of unsupervised learning, an 'exemplar' is a point that is a good representation of its neighboring points. The aim of all clustering algorithms is to locate the exemplars and then group all points 'close' to it into a cluster. Algorithms differ in how they define closeness, in whether they aim to locate a fixed/variable of clusters, and in whether they return the precise location of 'exemplar' points or not.

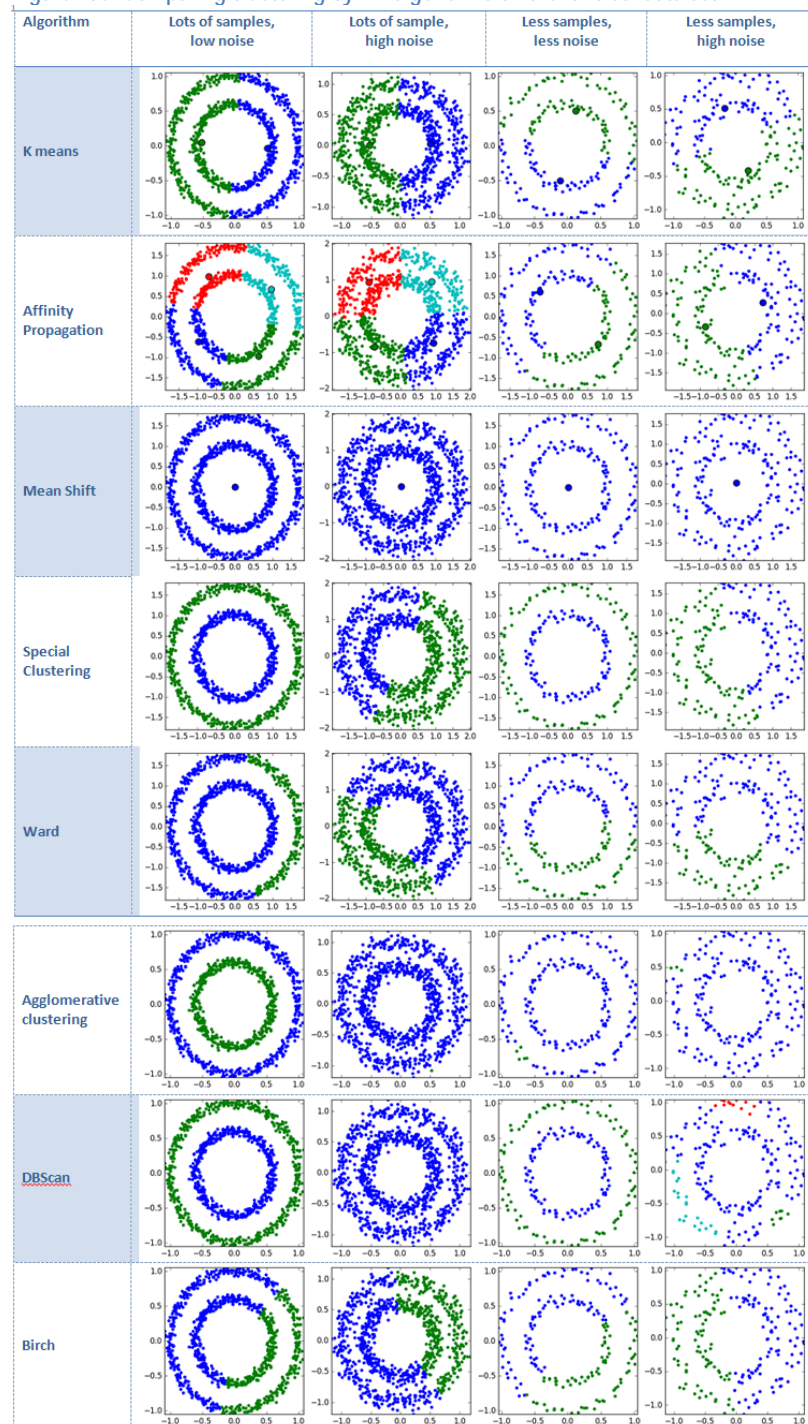
We try the following list of unsupervised learning algorithms:

1. K-means – the simplest clustering algorithm that starts by initially marking random points as exemplars. It iteratively does a two-step calculation: in the first step, it maps points to the closest exemplar, in the second step it redefines the exemplar as the mean of the points mapped to it. It locates a fixed number of clusters, which is assigned to two in our code.
2. Affinity Propagation – locates a dynamic number of exemplars. The algorithm involves the passing of 'soft' information that treats every point as a possible exemplar and allows for the possibility of every point being included in a cluster around it. The algorithm is known to be good for finding a large number of small clusters. The number of clusters chosen can be indirectly influenced via a 'preference' parameter; we have used the default settings within the *sklearn* library. For two clusters, it is probably better to use a simpler algorithm.
3. Mean Shift – based on iteratively finding centroids and aims to find blobs in a smooth density of samples. It automatically selects the number of clusters, influenced indirectly by the 'bandwidth' parameter; we have retained default settings. The algorithm prunes the candidate list in the end to avoid near-duplicates as exemplars.
4. Spectral clustering – like affinity propagation, it passes messages between points, but does not identify the exemplar of each cluster. It needs to be told the number of clusters to locate; we have specified two. This is understandable, since the algorithm computes an affinity matrix between samples, embeds into a low-dimensional space and then runs K-means to locate the clusters. This is known to work well for a small number of clusters.
5. Ward – a hierarchical clustering technique that is similar to K-means except that it uses a decision tree to cluster points.
6. Agglomerative Clustering – another hierarchical clustering technique that starts with each point as an exemplar and then merges points to form clusters. Unlike Ward which looks at the sum of squared distance within each cluster, this looks at the average distance between all observations of pairs of clusters.
7. DBScan – forms clusters with roughly similar density of points around them. Points in low-density regions are treated as outliers. It can potentially scan the data set multiple times before converging on the clusters.

8. Birch – A hierarchical clustering technique designed for very large databases. It can incrementally cluster streaming data; hence in many cases, it can cluster with a single pass over the data set.

The results on applying the above nine classifiers on the circles dataset are tabulated in the figure below.

Figure 106: Comparing clustering by ML algorithms on the 'circles' data set



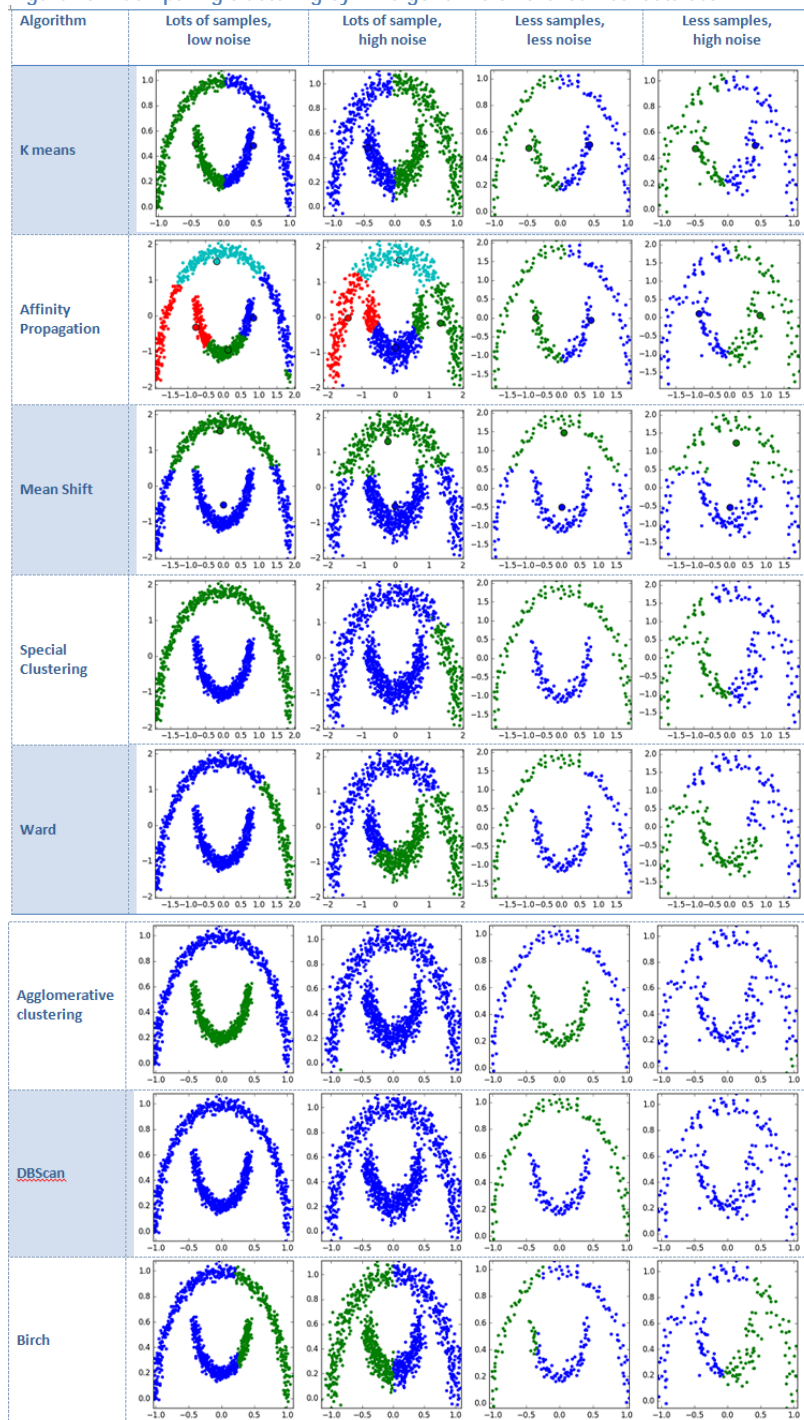
Source: J.P.Morgan Macro QDS

Some of the observations are listed below:

- K-means – a linear split based on Euclidean metrics is made, which is not optimal for the non-linear data set. Simplicity of model leads to under-fitting.
- Affinity Propagation – more samples induces formation of more clusters. In scenarios with a fewer number of samples, performance improves..
- Mean Shift – algorithm fails completely in this scenario. Having more samples or less noise does not help.
- Spectral clustering – works perfectly when noise is low. It is not sensitive to sample size. When noise is high, it reduces to a linear fit. Algorithm works better than other alternatives for this data set.
- Ward – is even worse than K-means when the number of samples is high.
- Agglomerative Clustering – decision tree enables efficient performance when the sample size is high and noise is low. In all other cases, it fails.
- DBScan – another decision tree algorithm that is sensitive to noise alone, and not so much to sample size.
- Birch – performance is similar to K-means, in spite of additional complexity of algorithm.

The results on applying the above nine classifiers on the ‘curves’ dataset are tabulated in the figure below.

Figure 107: Comparing clustering by ML algorithms on the 'curves' data set



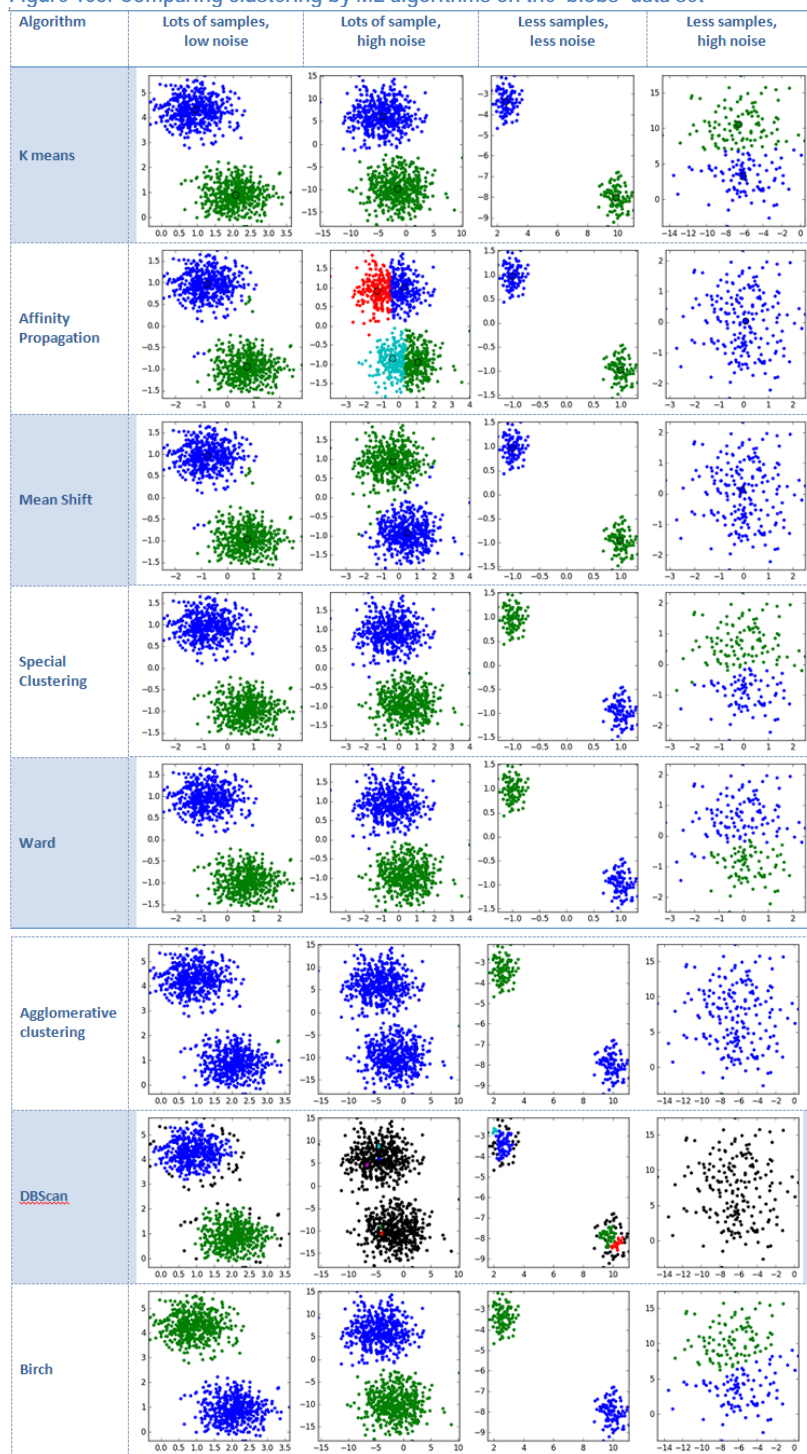
Source: J.P.Morgan Macro QDS

Some of the observations are listed below:

- K-means – the linear fit fails again in the non-linear data model. Again the fit is robust to sample size and noise level.
- Affinity Propagation – as in first example, the algorithm over-estimates the number of clusters when the sample size is large. In case of low sample size, its performance is no better than K-means.
- Mean Shift – in this case, the algorithm performs identical to K-means.
- Spectral clustering – as in the first example, this is the best algorithm for separating the two classes. Algorithm is sensitive to noise; presence of noise can deteriorate performance to a level similar to K-means.
- Ward – fails for the second consecutive dataset. Performance is worse than what K-means can offer.
- Agglomerative Clustering – the decision tree based algorithm works best when noise is low. In that case, it can cluster correctly. It fails completely under high noise.
- DBScan – like Ward, it fails on this dataset.
- Birch – performance is similar to K-means; one would prefer K-means due to its simplicity.

The results on applying the above nine classifiers on the blobs dataset are tabulated in the figure below.

Figure 108: Comparing clustering by ML algorithms on the 'blobs' data set



Source: J.P.Morgan Macro QDS

Some of the observations are listed below:

- K-means – algorithm works correctly. Data set is simple and linearly separable, and hence is ideally suited for K-means to work.

- Affinity Propagation – algorithm works correctly when noise is low. When noise is high, behavior depends on the number of samples: a high number of samples leads to an excessive number of clusters, a low number of samples leads to a very low number of clusters.
- Mean Shift – algorithm works well on this simple dataset, after failing to cluster the previous two examples.
- Spectral clustering – algorithm works correctly again. It is the only algorithm to work well on all three test cases.
- Ward – algorithm works well on this simple dataset, after failing to cluster the previous two examples. Performance is similar to Mean Shift.
- Agglomerative Clustering – surprisingly, the decision tree fit fails to cluster when the number of sample points is high. Algorithm fails on this simple data set.
- DBScan – algorithm fails on the data set.
- Birch – performance is similar to K-means in spite of more complicated structure.

IV: HANDBOOK OF ALTERNATIVE DATA

Table of Contents of Data Providers

In this section we provide a comprehensive list of alternative data and technology providers. In order to navigate this handbook, we provide a table of contents below.

- A. [Data from individual activity](#)
 - 1. [Social media](#)
 - i. [Investment professional social media](#)
 - ii. [Social network sentiment](#)
 - iii. [Blogs, picture and video analytics](#)
 - 2. [News and reviews](#)
 - i. [Mobile data content and reviews](#)
 - ii. [News sentiment](#)
 - 3. [Web searches and personal data](#)
 - i. [Email and purchase receipt](#)
 - ii. [Web search trends](#)
- B. [Data from business processes](#)
 - 1. [Transaction data](#)
 - i. [Other commercial transactions](#)
 - ii. [E-commerce and online transactions](#)
 - iii. [Credit card data](#)
 - iv. [Orderbook and flow data](#)
 - v. [Alternative credit](#)
 - 2. [Corporate data](#)
 - i. Sector data ([C.Discretionary](#), [Staples](#), [Energy/Utilities](#), [Financials](#), [Health Care](#), [Industrials](#), [Technology](#), [Materials](#), [Real Estate](#))
 - ii. [Text parsing](#)
 - iii. [Macroeconomic data](#)
 - iv. [Accounting data](#)
 - v. [China/Japan data](#)
 - 3. [Government agencies data](#)
 - i. [Federal or regional data](#)
 - ii. [GSE data](#)
- C. [Data from sensors](#)
 - 1. [Satellites](#)
 - i. [Satellite imagery for agriculture](#)
 - ii. [Satellite imagery for maritime](#)
 - iii. [Satellite imagery for metals and mining](#)
 - iv. [Satellite imagery for company parking](#)
 - v. [Satellite imagery for energy](#)
 - 2. [Geolocation](#)
 - 3. [Other sensors](#)
- D. [Data aggregators](#)
- E. [Technology solutions](#)
 - 1. [Data storage \(databases\)](#)
 - 2. [Data transformation](#)
 - 3. [Hadoop and Spark framework](#)
 - 4. [Data analysis infrastructure](#)
 - 5. [Data management and security](#)
 - 6. [Machine learning tools](#)
 - 7. [Technology consulting firms](#)

Company descriptions in this section are excerpted or adapted from the companies' websites

A. Data from Individual Activity

1) Social Media

Investment Professional's Social Media

S. No.	Company Name	Asset Class	Notes ⁵⁷
1.	StockTwits www.stocktwits.com	Equity	StockTwits is the leading financial communications platform for the investing community. More than 300,000 investors, market professional and public companies share information, ideas about the market and individual stocks using StockTwits, producing streams that are viewed by an audience of over 40 million across the financial web and social media platform.
2.	SumZero www.sumzero.com	Equity	SumZero is the world's largest community of investment professionals working at hedge funds, mutual funds, and private equity funds. With more than 12,000 pre-screened professionals collaborating on a fully-transparent platform, SumZero fosters the sharing of thousands of proprietary investment reports every year, and offers several ancillary services in support of that effort. These free services include capital introduction services, buy side career placement services, media placement, and more.
3.	Scutify www.scutify.com	Equity	Scutify is Social Network for Investors, Sentiment derived from post can be used by traders or anyone new to the markets. It can be used as an educational tool to learn more about the markets.
4.	TrustedInsight www.thetrustedinsight.com	Equity	Trusted Insight is the world's biggest network of institutional investors. Connect investment decision-makers at endowments, foundations, pensions, insurance companies, sovereign wealth funds, family offices, corporations and healthcare systems. Provide institutional investors with access to a global professional network, alternative investment opportunities and an informational advantage in private markets.
5.	GNIP www.gnip.com	Equity	Gnip has partnered with StockTwits to create a product in order to bring social finance conversation to market – in both real-time and historical tailored to Hedge Funds and Traders.

Social Networks Sentiment

S. No.	Company Name	Asset Class	Notes
1.	Accern www.accern.com	Equity	Trading alerts and analytics are derived from relevant finance-related news processed on over 300 million public news websites, blogs, social media websites such as Twitter, and public financial documents such as SEC filings. Accern primarily serves institutional investors and majority of current clients are quantitative hedge funds. They attract small firms because of flexible and affordable pricing options and large firms due to dedicated support, news source customization, and much more. Aside from hedge funds, existing clients include pension and endowment funds, banks, brokerage firms, and more.
2.	iSentium www.isentium.com	Equity	iSense App transforms over 50 million Twitter messages per hour into a real-time sentiment time series. It is effectively a sentiment search engine which provides investors with a quick way to judge the potential market impact of a tweet, a news article, or other buzz discussed on social media. Daily Directional Indicators (DDI) has shown that social media sentiment is a strong indicator of market conditions.
3.	Sentiment Trader www.sentimentrader.com	Equity, Bond, FX, ETF and Commodity	Sentiment trader has over 300 Sentiment Surveys, Social Sentiment and other indicator for Equities, Bond, Currencies, ETFs and commodities. They have created models, with time frames between intraday and weekly, that aggregate many of their sentiment indicators.
4.	Social Alpha www.social-alpha.com	Equity, ETF	Social Alpha offers social analytics factor feeds, dashboards, real time alerts and custom widgets to financial institutions, day and prop traders, swing and trend traders, event-driven and system traders. Active Traders get the same sophisticated analytics used by our hedge fund and broker-dealer clients, at a cost scaled to the size of watch list. Cover all U.S.-listed stocks and ETFs and can supply multiple years of historical data for analysis, system development and back-testing.
5.	Social Market Analytics www.socialmarketanalytics.com	Equity	Social Market Analytics harness unstructured yet valuable information embedded in social media streams and provide actionable intelligence in real time to clients. SMA's analytics generate high signal data streams based on the intentions of professional traders. Data is unique in that we have over four years of out-of-sample data which cannot be recreated given a user's ability to delete Tweets at a later time. Dashboard provides a real-time contextual view of the social media conversation. SMA identifies viral topics quickly, provides account credibility rankings, and offers instantaneous evaluation of Tweets.
6.	Descartes Labs www.descarteslabs.com	Commodity, Agriculture	Descartes Labs has full imagery archives (some including data only a few hours old) from hundreds of satellites. The Descartes Platform is built to ingest virtually any kind of data, including satellite, weather data, commodity price histories, web crawls, and sentiment analysis from social media networks.
7.	DataSift www.datasift.com	Equity	DataSift's platform connects to real-time feed of social data including Facebook topic data, uncovers insights with sophisticated data augmentation, filtering and classification engine, and provides the data for analysis with the appropriate privacy protocol required by the data sources. It does analysis on data sources like Bitly, Blogs, Boards, Daily Motion, Disqus, FB, Instagram, IMDB, Intense Debate, LexisNexis, NewsCred, Reddit, Topix, Tumblr, Videos, Wikipedia, Wordpress, Yammer and YouTube. They get data from Edgar Online, Wikipedia and Wordpress.
8.	Knowsis www.knowsis.com	Equity, ETF, Commodity, FX and Fixed Income	Knowsis is a web intelligence company using cutting edge natural language processing and data science to extract value from non-traditional online sources into quantifiable and actionable output for the capital markets. Enhance trading strategies and risk management with statistically significant insights from social media data. They cover cross assets (Equities, Equity Indices, ETFs, Commodities, FX and Fixed Income) with intraday to end-of-

			day data points available.
9.	Lexalytics www.lexalytics.com	Equity	Lexalytics processes billions of unstructured documents globally every day. It translates text into profitable decisions; makes state-of-the-art cloud and on-premise text and sentiment analysis technologies that transform customers' thoughts and conversations into actionable insights. The on-premise Salience® and SaaS Semantria® platforms are implemented in a variety of industries for social media monitoring, reputation management and voice of the customer programs.
10.	Dataminr www.dataminr.com	Commodity, Equity	Dataminr transforms real-time data from Twitter and other public sources into actionable alerts, identifying the most relevant information in real-time for clients in Finance. Dataminr's strategic partnership with Twitter includes real-time access to all public tweets and distills over 500 million tweets a day to the handful of signals. It classifies signals based on geo-location, topic relevancy and market moving implications and ranks them by level of urgency.
11.	Alphamatician www.alphamatician.com	Real Estate, Equity	Alphamatician collect data from dozens of sources that collectively build a clear picture of such things as company and brand health, product and product category pricing and demand trends, customer engagement, and company risk factors. And also track real estate markets and numerous other sources. Cover more than 5,000 publicly traded companies across North America, Europe and Asia. Alphamatician also covers thousands of privately held companies and individual brands. Alphamatician provides FaceBook, Instagram, E Commerce, Browser Statistics, Complaints, SEC filings, Twitter, Pinterest, Employment, Mobile Apps, and Promotional Activity data.
12.	TheySay www.theysay.io	Equity	TheySay Analytics tools provide Sentiment Analysis. Apart from positive, negative and neutral they have many classifications such as emotion, humour, gender, risk, speculation, and even sarcasm. Services include Social Media Monitoring, Digital media monitoring, Stock Market Sentiment Monitoring, Sentiment Analysis API, Secure Installations, Bespoke Classifiers, Emotion Analysis and Topic Detection.
13.	Trackur www.trackur.com	Equity	Trackur offers affordable social media monitoring tools for individuals, small companies, large corporations & agencies. Full monitoring of all social media and mainstream news, including Twitter, Facebook, Google+ and more. Provides executive insights including trends, keyword discovery, automated sentiment analysis & influence scoring.
14.	MatterMark www.mattermark.com	Equity	MatterMark is a data platform to search companies and investors to effortlessly create actionable lists of leads. MatterMark monitors trends in team size, web traffic, media mentions, and more.
15.	ShareThis www.sharethis.com	Equity	The ShareThis consumer engagement and sharing tools are used by publishers to drive engagement, traffic and personalization, capturing the widest and deepest sentiment of people across the internet. These sentiment flows into the ShareThis platform as terabytes of data that is processed daily to better understand people, making social data actionable for any business that requires a holistic view of people or customers.

Blogs, Picture and Video Analytics

S. No.	Company Name	Asset Class	Notes
1.	Accern www.accern.com	Equity	Trading alerts and analytics are derived from relevant finance-related news processed on over 300 million public news websites, blogs, social media websites such as Twitter, and public financial documents such as SEC filings. Accern primarily serves institutional investors and majority of current clients are quantitative hedge funds. They attract small firms because of flexible and affordable pricing options and large firms due to dedicated support, news source customization, and much more. Aside from hedge funds, existing clients include pension and endowment funds, banks, brokerage firms, and more.
2.	Alphamatician www.alphamatician.com	Real Estate, Equity	Alphamatician collect data from dozens of sources that collectively build a clear picture of such things as company and brand health, product and product category pricing and demand trends, customer engagement, and company risk factors. And also track real estate markets and numerous other sources. Cover more than 5,000 publicly traded companies across North America, Europe and Asia. Alphamatician also covers thousands of privately held companies and individual brands. Alphamatician provides FaceBook, Instagram, E Commerce, Browser Statistics, Complaints, SEC filings, Twitter, Pinterest, Employment, Mobile Apps, and Promotional Activity data.
3.	DataSift www.datasift.com	Equity	DataSift's platform connects to real-time feed of social data including Facebook topic data, uncovers insights with sophisticated data augmentation, filtering and classification engine, and provides the data for analysis with the appropriate privacy protocol required by the data sources. It does analysis on data sources like Bitly, Blogs, Boards, Daily Motion, Disqus, FB, Instagram, IMDB, Intense Debate, LexisNexis, NewsCred, Reddit, Topix, Tumblr, Videos, Wikipedia, Wordpress, Yammer and YouTube. They get data from Edgar Online, Wikipedia and Wordpress.
4.	MixRank www.mixrank.com	Equity	Data-driven sales and marketing. MixRank tracks your customers so you don't miss an opportunity. They accomplish this by collecting and analyzing vast amounts of data across the web and mobile.

2) News and Reviews

Mobile Data Content & Product Reviews

S. No.	Company Name	Asset Class	Notes
1.	App Annie www.appannie.com	Equity	App Annie's intelligence products provide detailed app download, revenue, demographic and usage estimates for every major mobile app. Track app performance around usage, downloads, revenue and spend across countries and app stores—for free. Optimize your app store strategy by analyzing market data metrics including historical rankings, ratings, reviews and keywords for any app across categories and countries.
2.	YipitData www.yipitdata.com	Equity	YipitData specializes in the collection and analysis of website data: publicly available information on company and government websites. When aggregated and analyzed correctly, website data can provide significant value to the investment research process given its unique characteristics relative to other datasets.
3.	7Park www.7parkdata.com	Equity	7Park provides clarity into your investments with 1)Traffic Intelligence - Tracks global panel of over 100+ million internet and mobile users and delivers detailed metrics including page views, conversion metrics and more 2) Web Intelligence - Key performance metrics on public and private companies, 3) Product Intelligence - Analyze billions of purchases made by millions of US consumers to isolate spending patterns and trends analysis, 4) App Intelligence : App-level insights from real-time, on-device monitoring of millions of mobile phones and tablets streaming billions of data points on key metrics such as App installs, "active" users, engagement and retention trends and 5) Merchant Intelligence - Quantify share-of-wallet, macro consumer trends and user level studies of merchant-level detail by tracking billions of purchases made by millions of US consumers.
4.	Alphamatician www.alphamatician.com	Real Estate, Equity	Alphamatician collect data from dozens of sources that collectively build a clear picture of such things as company and brand health, product and product category pricing and demand trends, customer engagement, and company risk factors. And also track real estate markets and numerous other sources. Cover more than 5,000 publicly traded companies across North America, Europe and Asia. Alphamatician also covers thousands of privately held companies and individual brands. Alphamatician provides FaceBook, Instagram, E Commerce, Browser Statistics, Complaints, SEC filings, Twitter, Pinterest, Employment, Mobile Apps, and Promotional Activity data.
5.	Dataprovider.com www.dataprovider.com	Equity	Dataprovider.com provides reports on information extracted from more than 100 million websites in 40 countries. We collect and structure thousands of variables, including contact details, from which you can create data sets, matching your exact criteria Provide all the vacancies straight from the websites of the business owner in the structured way. We index and structure websites and see what technologies are used and currently track 220+ different variables.
6.	Grandata www.grandata.com	Equity	Grandata integrates first-party and telco partner data to understand key market trends, predict customer behavior, and deliver impressive business results. It is a data monetization company.
7.	Bitly www.bitly.com	Equity	Optimize the link so marketers can own the customer experience. Bitly Enterprise - branded links, mobile deep linking, omnichannel campaign tracking, and audience intelligence - gives the world's leading brands a holistic, unbiased view into an increasingly complex digital landscape, and a powerful way to see, control and own their customer experience across the internet.

News Sentiment

S. No.	Company Name	Asset Class	Notes
1.	Accern www.accern.com	Equity	Accern has rapidly grown to become the market leader in the news and social analytics space in terms of content coverage. Trading alerts and analytics are derived from relevant finance-related news processed on over 300 million public news websites, blogs, social media websites such as Twitter, and public financial documents such as SEC filings. Accern primarily serves institutional investors. The majority of our current clients are quantitative hedge funds. Many small firms are coming to us because of our flexible and affordable pricing options. Large firms are coming to us due to our dedicated support, news source customization, and much more. Aside from hedge funds, our existing clients include pension and endowment funds, banks, brokerage firms, and more.
2.	RavenPack www.ravenpack.com	All Asset Classes	RavenPack Analytics transforms unstructured big data sets, such as traditional news and social media, into structured granular data and indicators to help financial services firms improve their performance. Clients with an intraday horizon value RavenPack's ability to detect relevant, novel and unexpected events - be they corporate, macroeconomic or geopolitical - so they can enter new positions, or protect existing ones. The product serves to overcome the challenges posed by the characteristics of Big Data - volume, variety, veracity and velocity - by converting unstructured content into a format that can be more effectively analyzed, manipulated and deployed in financial applications.
3.	RelateTheNews www.relatethenews.com	Equity	RelateTheNews was founded to meet the needs of individuals and trading firms of all sizes. RelateTheNews is built to provide actionable insight to drive alpha. We combine data science techniques, psychology, domain specific understanding and cutting edge technology to drive real-time analysis of text data (unstructured big data) to power effective, unique, and leading decision making in the global capital markets.
4.	Heckyl www.heckyl.com	Equity	Real time trending news from the web, government wires, news wires, blogs and Twitter. Get news sentiment and velocity to know the news points making an impact in the market. Heckyl big data engine mines Public, Private and Proprietary data sources to help analyze, connect and discover the next big opportunity. Update constantly on Government moves, hedge funds or PE investments, VC deals, economic announcements and more.
5.	DataSift www.datasift.com	Equity	DataSift's platform connects to real-time feed of social data including Facebook topic data, uncovers insights with sophisticated data augmentation, filtering and classification engine, and provides the data for analysis with the appropriate privacy protocol required by the data sources. It does analysis on data sources like Bitly, Blogs, Boards, Daily Motion, Disqus, FB, Instagram, IMDB, Intense Debate, LexisNexis, NewsCred, Reddit, Topix, Tumblr, Videos, Wikipedia, Wordpress, Yammer and YouTube. They get data from Edgar Online, Wikipedia and Wordpress
6.	GDELT Project www.gdeltproject.org	Equity	Creating a platform that monitors the world's news media from nearly every corner of every country in print, broadcast, and web formats, in over 100 languages, every moment of every day and that stretches back to January 1, 1979 through present day, with daily updates, required an unprecedented array of technical and methodological innovations, partnerships, and whole new mindsets to bring this all together and make it a reality. Creating a database of a quarter billion georeferenced records covering the entire world over 30 years, coupled with the massive networks that connect all of the people, organizations, locations, themes, and emotions underlying those events, required not only solving unparalleled challenges to create the database, but also a "reimagining" of how we interact and think about societal-scale data.
7.	InfoTrie www.infotrie.com	Equity, Commodity, FX	InfoTrie's FinSents scans and monitors millions of websites, blogs, and business news publications in real-time to analyze 50,000 + stocks, topics, people, commodities and other assets by advanced low-latency algorithms. It can offer premium stock sentiment analysis data, predictive analysis data, and other

			alternative data. I-Feed offers sentiment data for 50,000 + stocks, major FX, commodities or topics/ people, other assets tracked by InfoTrie sentiment engine. Up to 15 years of tick by tick and/ or daily time series can be provided for any asset. Th user can download historical data on demand. InfoTrie scans millions of sources in real-time: websites, blogs, social media, user-provided private data and processes premium such as Bloomberg, Reuters or Dow Jones.
8.	Repustate www.repustate.com	Equity	Fast, reliable & accurate sentiment analysis in 15 languages. Perform sentiment analysis and extract semantic insights from social media, news, surveys, blogs, forums or any of your company data. The world's fastest text analytics engine. Sentiment analysis, entity extraction and topic detection in Arabic, Chinese, Dutch, French, German, Hebrew, Italian, Polish, Portuguese, Russian, Spanish, Turkish, Thai, Vietnamese and of course, English.
9.	Alexandria www.alexability.com	Equity, Commodity, Fixed Income, FX	Alexandria's SentMap is for investment professionals who are overwhelmed with financial news, fear they may not be taking full advantage of it, and don't have a quick and efficient research environment for exploring its effects on their holdings. The ACTA engine of Alexandria is a set of proprietary algorithms that assess the sentiment of unstructured information – such as high-value financial news feeds. The output, scored for negative or positive sentiment, is then delivered as real-time signals or a set of indices to clients, who use such contextual-intelligence to predict the outcome of their actions and to assess the risks involved.
10.	Inferess www.inferess.com	Equity	Inferess News Analytics turns news feeds into event-driven Analytics feeds, rich of sentiment labels and identifiers for a broad range of types of events. Indicators are quantitative and qualitative measures that capture the essential meaning of textual data. Stock market news, commentary and analysis, Disclosure and corporate actions, Coverage of 14,000 companies in North and South America, Europe, Asia, the BRICs and other emerging markets, Industry and sector news, Blogs and opinions, Economic statistics and trends, Global Energy and Commodities, Historical analytics archive dating back to 2012.
11.	Alphasense www.alpha-sense.com	Equity	AlphaSense is a search engine that helps you to instantly cut through the noise and uncover critical data points to search through all major filings, earning calls and conference transcripts. Linguistic search algorithms allow you to search across millions of documents with a few clicks, dramatically reducing research times. Identify key data points, trends and themes on any 1 company or across 35,000+ global companies
12.	MarketPsych www.marketpsych.com	Equity	MarketPsych develops sentiment data across major news and social media outlets, quantitatively explores behavioral analytics in alpha generation and risk management and uncovers investors' innate to help their financial advisors better serve them.
13.	Sentifi www.sentifi.com	Equity	Sentifi is building the largest online ecosystem of crowd-experts and influencers in global financial markets (Sentifi Crowd) to generate Sentifi Signals, market intelligence that is not available via traditional news media using our proprietary technology Sentifi Engine.
14.	BrandWatch www.brandwatch.com	Equity	BrandWatch builds intelligent software solutions that meet the needs of forward-thinking social businesses. Over 1,200 brands and agencies, including Unilever, Whirlpool, British Airways, Asos, Walmart and Dell, use the products to fuel smarter decision making. It is a powerful social listening and analytics platform.
15.	MatterMark www.mattermark.com	Equity	Mattermark is a data platform to search companies and investors to effortlessly create actionable lists of leads. Mattermark monitors trends in team size, web traffic, media mentions, and more.
16.	Estimize www.estimize.com	Equity	An earnings data set with exclusive insight on over 2000 stocks. Estimize crowdsources earnings and economic estimates from 44,022 hedge fund, brokerage, independent and amateur analysts.
17.	LinkUp	Macro	LinkUp provides real-time labor market job data goes back nearly 10 years and

www.linkup.com	Economic	includes over 75 million job listings from 30,000 companies. It integrated additional 3 rd party data elements into the platform and developed a wide range of proprietary analytics to deliver unique, powerful, and predictive insights into labor markets in the U.S. and around the world
--	----------	--

3) Web Searches, Personal Data

Email and Purchase Receipt

S. No.	Company Name	Asset Class	Notes
1.	Return Path www.returnpath.com	Equity	Return Path Data Exchange has brought together the world's most comprehensive source of data from the email ecosystem. It partners with more than 70 providers of mailbox and security solutions, covering 2.5 billion inboxes—approximately 70 percent of the worldwide total. Also feeding into this data platform is our consumer network of more than 2 million consumers and purchase receipts from 5,000 retailers around the world, delivering unparalleled insight into purchase behavior, brand affinity, and consumer preferences. Return Path is the world's leading email data solutions provider. Our solutions provide the insight companies need to build better relationships, drive more response, and increase revenue.
2.	Slice Intelligence www.intelligence.slice.com	Equity	Slice Intelligent's market research products and custom analytics solutions offer brand-new insights on online shopping, including: share of wallet, loyalty and switching, pricing, shipping and payment method. All data is reported daily, at the item level and can be analyzed by the shopper's geographic location.
3.	Superfly insights www.superfly.com	Equity	Provides detailed consumer purchasing insights using live transaction data in-app, online and in-store. They also provide competitor's sales data, statistical data and insights on user's purchasing habits for leading brands.

Web Search Trends

S. No.	Company Name	Asset Class	Notes
1.	Alexa (Internet's traffic company) www.alexacom	Equity	Alexa's traffic estimates are based on data from our global traffic panel, which is a sample of millions of Internet users using one of over 25,000 different browser extensions. In addition, they gather much of traffic data from direct sources in the form of sites that have chosen to install the Alexa script on their site and certify their metrics. Alexa's global traffic rank is a measure of how a website is doing relative to all other sites on the web over the past 3 months. We provide a similar country-specific ranking, which is a measurement of how a website ranks in a particular country relative to other sites over the past month.
2.	AddThis www.addthiscom	Equity	On January 5, 2016, Oracle completed the acquisition of AddThis, of publisher personalization, audience insight and activation tools that powers 15 million websites and enables unmatched audience segment quality, scale and insight. AddThis offers unparalleled insight into the interests and behaviors of over 1.9 billion web visitors. This vast global footprint reaches 96% of the U.S. web — ranking above Google and Facebook in comScore for distributed content and ad reach.

B. Data from Business Processes

1) Transaction Data

Other Commercial transactions

S. No.	Company Name	Asset Class	Notes
1.	Brave New Coin www.bravenewcoin.com	FX	Brave New Coin (BNC) is a Data & Research company who has built a superior Market-Data Engine for the Blockchain & Digital Equities industry. They collect, process & index live data from over 100 trading platforms in real-time to produce a number of useful data tools for Developers, Traders & Enterprise. BNC also provides industry news and insights. The coverage spans every aspect of the Digital Currency and Blockchain Ecosystem, including its impact on the greater FinTech and payments space. Featuring a mix of regular tech wraps, weekly market updates, features, interviews and a steady stream of industry thought-leader guest posts.
2.	Quandl www.quandl.com	All Asset Classes	Quandl delivers financial, economic and alternative data to over 150,000 people worldwide. Quandl offers essential financial and economic data alongside a suite of unique, alpha-generating alternative datasets.
3.	Alphamatician www.alphamatician.com	Equity	Alphamatician collect data from dozens of sources that collectively build a clear picture of such things as company and brand health, product and product category pricing and demand trends, customer engagement, and company risk factors. And also track real estate markets and numerous other sources. Cover more than 5,000 publicly traded companies across North America, Europe and Asia. Alphamatician also covers thousands of privately held companies and individual brands. Alphamatician provides FaceBook, Instagram, E Commerce, Browser Statistics, Complaints, SEC filings, Twitter, Pinterest, Employment, Mobile Apps, and Promotional Activity data.
4.	Buildfax www.buildfax.com	Real Estate	Buildfax catalogue over 23 billion data points on residential and commercial structures and proprietary, providing predictive analytics about how building permit data predicts loss and tells the story of property history, improvements, structural risks, and changes over time. These insights can be a game changer for everyone involved in Real Estate. Buildfax count some of the world's largest insurance, property companies and lending companies as clients.
5.	Data Provider www.dataprovider.com	Equity	Dataprovider.com provides reports on information extracted from more than 280 million websites in 40 countries. They collect and structure hundreds of variables, including contact details, company profile, economic footprint and heartbeat (activity), from which users can create novel data sets.
6.	NIELSEN Data research.chicagobooth.edu/nielsen	Equity	The Retail Scanner data consist of weekly pricing, volume, and store environment information generated by point-of-sale systems from more than 90 participating retail chains across all US markets. The Consumer Panel Data captures 40-60,000 US households' purchases of fast-moving consumer goods.
7.	Factset Revere www.factset.com/news/revere	Equity	Revere has built a dynamic industry taxonomy that offers investors a unique way to classify companies. It offers a comprehensive database of supply chain relationships that helps investors identify companies' interrelationships and mutual dependencies, as well as geographic revenue exposure to manage geopolitical and macroeconomic risk. Factset acquired Revere in 2013 and Revere's database is available from Factset.

Ecommerce and Online Transaction

S. No.	Company Name	Asset Class	Notes
1.	Premise www.premise.com	Agriculture, FX, Macro Economic	Premise indexes and analyzes millions of observations captured daily by global network of contributors, unearthing connections that impact global decisions. Discerning the economic effect of droughts, new policies and poor cold storage Premise publishes its Argentina, Brazil, China, India and U.S. food staples indexes via the Bloomberg terminal, giving traders, forecasters, analysts and government officials an advantage to now cast expected market behaviors up to several weeks early.
2.	PriceStats (Billion Prices Project) www.pricestats.com	Equity	PriceStats collects price information from over 1,000 retailers across nearly 70 countries. They publish daily inflation series for over 20 countries and PPP indices for 8 economies. Most of their series start in 2007-2008. PriceStats gathers price information for key economic sectors: food & beverages; furnishing & household products; recreation & culture; clothing & footwear; housing, electricity & fuel; and health. They continue to update this as more data becomes available.
3.	7Park www.7parkdata.com	Equity	7Park provides clarity into your investments with 1) Web Traffic Intelligence - Tracks global panel of over 100+ million internet and mobile users and delivers detailed metrics 2) Web Intelligence - Key performance metrics on public and private companies, 3) Product Intelligence - Analyze billions of purchases made by millions of US consumers to isolate spending patterns and trends analysis, 4) App Intelligence : App-level insights from real-time, on-device monitoring of millions of mobile phones and tablets streaming billions of data points on key metrics such as App installs, "active" users, engagement and retention trends and 5) Merchant Intelligence - Quantify share-of-wallet, and macro consumer trends.
4.	Alphamatician www.alphamatician.com	Equity	Alphamatician collect data from dozens of sources that collectively build a clear picture of such things as company and brand health, product and product category pricing and demand trends, customer engagement, and company risk factors. And also track real estate markets and numerous other sources. Cover more than 5,000 publicly traded companies across North America, Europe and Asia. We also cover thousands of privately held companies and individual brands. Alphamatician provides FaceBook, Instagram, E Commerce, Browser Statistics, Complaints, SEC filings, Twitter, Pinterest, Employment, Mobile Apps and Promotional Activity data.
5.	DataStreamX www.datastreamx.com	Equity & All Asset Classes	DataStreamX helps unleash the power of the world's data. They strive to provide a delightful marketplace experience that connects suppliers of data and consumers of data. Logistics, Import/Export, Foot Traffic and Financial datasets are available.

Credit/Debit Card and Bank/Investment/Mortgage account data

S. No.	Company Name	Asset Class	Notes
1.	SuperFly www.superfly.com	Equity	SuperFly Insights is an analytics company that provides detailed consumer purchasing insights using live transactional in-app, online and in-store data. SuperFly collects data from a massive set of permission-based accounts, apply advanced data extraction, normalization, and predictive analytics to create elaborate anonymous profiles and aggregate the profiles to deliver an easy-to-consume, dynamic and actionable insights.
2.	Second Measure www.secondmeasure.com	Equity	Second measure transforms credit card transactions into actionable insights for investors and analyzes billions of consumer purchases to deliver unprecedented insight into public and private companies. Second Measure delivers effective insights by analyzing billions of consumer purchases (i.e. greater than 1 percent of all U.S. consumers spending) and presenting detailed insights to investors.
3.	Yodlee www.yodlee.com	Equity	With Envestnet Yodlee Data Analytics, businesses will be able to garner insights based upon millions of anonymized consumer debit and credit transactions, providing the ability to track and monitor retail sales performance by competitor and by geo-location, better understand purchase timing and seasonality, and benefit from enhanced customer profiling and income segmentation. Yodlee's APIs, FinApps, and analytical tools leverage data from over 15,000 data sources to increase insight, enhance and extend digital capabilities, and engage users.

Orderbook and Flow data

S. No.	Company Name	Asset Class	Notes
1.	Second Measure www.secondmeasure.com	Equity	Second measure transforms credit card transactions into actionable insights for investors and analyzes billions of consumer purchases to deliver unprecedented insight into public and private companies. Second Measure delivers effective insights by analyzing billions of consumer purchases (i.e. greater than 1 percent of all U.S. consumers spending) and presenting detailed insights to investors.
2.	Quantcube www.q3-technology.com	Equity , Commodity	QuantCube Technology is a Fintech startup specialized in real-time Big Data analytics to predict macro trends, success and failures of companies and individual behaviors, by crossing billions of heterogeneous data with advanced analytics. Products include real-time indicators, resulting from the aggregation of multiple micro-data sources to deliver looking forward macro-data. QuantCube Technology offers a wide range of real-time Smart Data at a global level: from Global Macro activity to specific equities and commodities.
3.	EIDOSearch www.eidosearch.com	All Asset Classes	EidoSearch uses patented pattern matching technology to project probable event outcomes and find relationships in Big Data. 100+ billion pattern comparisons daily to uncover opportunities and allow for better estimation of risks
4.	Yodlee www.yodlee.com	Equity	With Envestnet Yodlee Data Analytics, businesses will be able to garner insights based upon millions of anonymized consumer debit and credit transactions, providing the ability to track and monitor retail sales performance by competitor and by geo-location, better understand purchase timing and seasonality, and benefit from enhanced customer profiling and income segmentation. Yodlee's APIs, FinApps, and analytical tools leverage data from over 15,000 data sources to increase insight, enhance and extend digital capabilities, and engage users.
5.	Quant Connect www.quantconnect.com	Equity, FX, Fixed Income, Commodity	Quant Connect provides 1) US Equities tick data going back to January 1998 for every symbol traded, totaling over 29,000 stocks. 2) Morning Star Fundamental data for the most popular 8,000 symbols for 900+ indicators since 1998. 3) Forex (FXCM, ONADA brokerage) since Apr 2007. 4) Futures tick trade and quote data from January 2009 onwards for every contract traded in CME, COMEX and GLOBEX. 5) Option trades and quotes down to minute resolution, for every option traded on ORPA since 2007. Cloud based backtesting tools are also available.
6.	Tick Data www.tickdata.com	Equity, FX, Fixed Income, Commodity	Tick Data provides historical intraday stock, futures, options and forex data for back-testing trading strategies, develop risk & execution models, perform post-trade analysis, and conduct important academic research with data as far back as 1974. API, AWS-S3 & 'OneTick' in memory database solutions are available.

Alternative credit

S. No.	Company Name	Asset Class	Notes
1.	FirstAccess www.firstaccessmarket.com	Consumer Credit	FirstAccess is an award-winning FinTech company that offers a customizable credit scoring platform for lending institutions in emerging markets to credit score anyone.
2.	Cignifi www.cignifi.com	Consumer Credit	Cignifi is changing consumer finance for over 80% of the adults globally using or owning a mobile phone. Cignifi's big data platform is the first to develop credit risk and marketing scores using mobile phone data.
3.	Affirm www.affirm.com	Consumer Credit	Affirm is modernizing consumer credit and changing the way people shop. And enable customers low credit scores or thin/no credit files to buy what they want today and pay over time—boosting conversion, revenue, and customer loyalty.
4.	BillGuard www.billguard.com	Consumer Credit	BillGuard is now part of Prosper. It is a crowd-powered peer-to-peer lending company.
5.	Enova www.enova.com	Consumer Credit	Enova is a technology and data analytics driven online lending company operating 11 brands in 6 countries and providing analytics services to businesses.
6.	Elevate www.elevate.com	Consumer Credit	Elevate was founded on a legacy of data and innovation coupled with a customer-first approach Elevate is reinventing the non-prime lending industry by giving consumers access to responsible and transparent credit options. US & UK focus.
7.	Kabbage www.kabbage.com	Consumer & SME Credit	Kabbage's comprehensive lending platform can be configured for your organization, allowing you to use the power of data to rapidly underwrite and monitor millions of customers.
8.	VantageScore www.vantagescore.com	Consumer & SME Credit	VantageScore Solutions is committed to providing greater score accuracy and consistency so that lenders and consumers alike can make credit decisions with a higher level of confidence. VantageScore serves three industries 1) Bank Card 2) Auto 3) Mortgage. Many of the most sophisticated secondary market participants use the VantageScore model to help evaluate and monitor risk, and to price and benchmark deals more accurately. Credit rating agencies accept loans based on the VantageScore model, and it's the dominant model used in the valuation of previously issued, private-label mortgage-backed securities.
9.	VisualDNA www.visualdna.com	All Asset Classes	VisualDNA was started in order to change communication between people for the better. By combining the approaches of data scientists, psychologists, creatives and engineers we have crafted new ways of understanding the human personality, allowing people to understand themselves and businesses to serve their customers better. We want to enable a web where people have control of their own data, and where the organizations that they choose to share it with can use it as constructively as possible.
10.	FairLoan www.fairloans.org.au	Consumer Credit	FairLoan provide an alternative to pay-day lenders and loan providers that charge very high interest rates to people who have been refused access to mainstream credit and personal loans. The Fair Loans offering has proven to be a convenient and private process that is borrower friendly and enables people to cover sudden or emergency costs in a reasonable and manageable fashion.
11.	Earnest www.earnest.com	Consumer Credit	Earnest uses unrivaled service, technology and additional factors such as your education, employment history, and financial picture in parallel with your credit history to bring low-interest loans to high-potential people.

12.	Lenddo www.lenddo.com	Consumer Credit	Lenddo has developed its technology based on 4 years of actual online lending experience that included collection, analysis and processing of billions of data points. Lenddo is a technology company that forged the idea of using non-traditional data to compute people's credit scores. Lenddo now offers a simple and secure way to prove identity and establish a consumers character online to unlock loans, online shopping and improve chances of employment. Lenddo is open to third parties, such as banks, lending institutions, utilities and credit card companies to reduce risk, increase portfolio size, improve customer service and verify applicants.
13.	LendingClub www.lendingclub.com	Consumer Credit	LendingClub is the world's largest online credit marketplace, facilitating personal loans, business loans, and financing for elective medical procedures. Borrowers access lower interest rate loans through a fast and easy online or mobile interface. Investors provide the capital to enable many of the loans in exchange for earning interest. LendingClub is transforming the banking system into a frictionless, transparent and highly efficient online marketplace, helping people achieve their financial goals every day.
14.	LendUp www.lendup.com	Consumer Credit	LendUp's mission is to provide anyone with a path to better financial health. LendUp is a better alternative to Payday Loans and offers online loans and credit cards with free financial education and the opportunity to build credit knowledge.
15.	Funding Circle www.fundingcircle.com	Consumer Credit	Funding Circle is the world's leading marketplace exclusively focused on small businesses — more than \$3bn has been lent to 25,000 businesses in the UK, USA, Germany and the Netherlands. Today, businesses can borrow directly from a wide range of investors, including more than 60,000 people, the UK Government, local councils, a university and a number of financial organizations.
16.	FactorTrust ws.factortrust.com	Consumer Credit	FactorTrust, the Alternative Credit Bureau, is relentlessly dedicated to proven analytics and clean credit information that provides lenders opportunities to grow more revenue meet compliance regulations and serve more consumers with more credit options. At the core of FactorTrust is alternative credit data not available from the Big 3 bureaus and analytics and risk scoring information lenders need to make informed decisions about the consumers they want. The company's solutions enable financial service companies an opportunity to uncover creditworthy prospects that are not surfacing via traditional credit sources.
17.	Prosper Marketplace www.prosper.com	Consumer Credit	Prosper is America's first marketplace lending platform, with over \$8 billion in funded loans. Prosper allows people to invest in each other in a way that is financially and socially rewarding. Prosper handles the servicing of the loan on behalf of the matched borrowers and investors.
18.	Experian Micro Analytics www.e-microanalytics.com	Consumer Credit	A double bottom line business unit specialized in alternative data, mobile technology, distributed models and enabling technology, Experian Micro Analytics' mission is to create credit risk profiles for millions of people and offer them easy access to credit. With innovative capabilities, millions of credit facilities are extended each year to people who previously did not have access to credit.
19.	Cognical www.zibby.com	Consumer Credit	Zibby, a Cognical company, offers customers a lease purchase transaction with no long term obligation and options for ownership. Our lease-to-own solution fully integrates with e-retailers' storefront and checkout processes to provide the only financing option for 65Mn underbanked and nonprime consumers to acquire electronics, appliances, furniture, and other durable goods online.
20.	CoreLogic www.corelogic.com	All Asset Classes	CoreLogic provides information intelligence to identify and manage growth opportunities, improve business performance and manage risk. Whether in real estate, mortgage finance, insurance, or the public sector, our clients turn to us as a market leader for unique property-level insights. CoreLogic collect and maintain the largest, most comprehensive property and related financial

			databases in the United States, Australia and New Zealand, with a growing presence in the United Kingdom, Canada, Mexico and India.
21.	DemystData www.demyst.com	All Asset Classes	DemystData help organizations unlock financial services through the use of better data, smarter decisions, and best in class infrastructure. DemystData's proprietary Attribute Platform works with large number of data sources like Telco, social, ID, fraud, websites, text, news, logs, and much more, and allows for the creation, hosting and delivery of actionable attributes that are unavailable anywhere else. They have helped the world's largest banks, telecommunications providers, insurers, and lenders optimize their workflows and more accurately screen over 400m customers.
22.	Experian www.experian.com	Consumer Credit	Experian is a leading global information services company, providing data and analytical tools to their clients around the world. They help businesses manage credit risk, prevent fraud, target marketing offers and automate decision making. Experian also help people to check their credit report and credit score, and protect against identity theft.
23.	Happy Mango www.happymangocredit.com	Consumer Credit	Happy Mango is a free personal finance management tool that helps customers keep track of their money, forecast cash flows and achieve financial goals. Happy Mango boosts individuals credit standing with endorsements from people who know them well and strengthen their finances with the help of personalized advice. Happy Mango expands borrower base and reduce defaults using cash-flow-based credit assessment.
24.	LexisNexis www.lexisnexis.com	All Asset Classes	LexisNexis® has unique data to help businesses make more profitable lending decisions across all consumer segments, including consumers with no, little or extensive credit histories. With a unique data set that covers over 98% of the U.S. population and provides consumer behavior information not available in traditional credit bureau data—organizations get a more complete picture of their customers. Combining cutting-edge technology, unique data and advanced scoring analytics, we provide products and services that address evolving client needs in the risk sector while upholding the highest standards of security and privacy.
25.	Netspend www.netspend.com	Consumer Credit	Netspend is a leading provider of Visa® Prepaid debit cards, Prepaid Debit MasterCard® cards, and commercial prepaid card solutions in the U.S. And serve the estimated 68 million underbanked consumers who do not have a traditional bank account or who rely on alternative financial services, by offering products that are flexible with powerful features designed just for them.
26.	Kreditech www.kreditech.com	Consumer Credit	Kreditech's mission is to provide access to credit for people with little or no credit history: The underbanked. Mainstream financial institutions have neglected this group of customers for years due to their lack of solid credit rating. Excluded from mainstream access to credit – those customers are left either with no or only with poor credit options such as Pawn or Payday Lending. By the use of its proprietary credit decision technology, Kreditech is driven to bridge this gap and to provide access to credit at fair and sustainable conditions to the non-prime market segment.
27.	RevolutionCredit www.revolutioncredit.com	Consumer Credit	RevolutionCredit's unique engagement platform and database of unique consumer behavioral economic data transforms the credit decisioning process for both lenders and consumers. RevolutionCredit was founded in 2012 by Zaydoon Munir, a veteran of Experian and the credit and financial services industry.
28.	Think Finance www.thinkfinance.com	Consumer Credit	Think Finance provide software, analytics, and marketing service for the B2B online lending environment leveraging technology tools, resources, and decades of experience to help you meet the needs of your customers—and shareholders. Customized services allow you to create, develop, launch and manage your loan portfolio while effectively serving your customers.

29.	ZestFinance www.zestfinance.com	Consumer Credit	Credit's technology platform is transforming consumer lending using machine learning— it helps companies make smarter credit decisions to expand the availability of fair and transparent credit. They have partnered with China's largest online direct-sales company, JD.com, to expand consumer credit. There are more than half a billion people in China with no credit history. This lack of data makes it incredibly difficult to determine credit risk and Credit's platform turns shopping data into credit data, creating credit histories from scratch.
30.	Innovate UK www.gov.uk/government/organizations/innovate-uk	Macro Economic SME, Grants	Innovate UK is the UK's innovation agency, accelerating economic growth. They fund, support and connect innovative businesses through a unique mix of people and programmes to accelerate sustainable economic growth.

2) Corporate Data

Consumer Discretionary

S. No.	Company Name	Asset Class	Notes
1.	Crain Communications Inc. Data available from Bloomberg	Consumer Discretionary	Cairn provides US monthly unit inventory of car and light truck by model, make, and manufacturer.
2.	Edmunds Data available from Bloomberg	Consumer Discretionary	Edmunds provides monthly US car manufacturing distribution detail including True Market Value, % Leased, TCI (True Cost of incentives), and TEMSRP by manufacturer and by model.
3.	Boxoffice Media LLC Data available from Bloomberg	Consumer Discretionary	Boxoffice Media LLC provides box office receipts data.
4.	Comscore Inc. Data available from Bloomberg	Consumer Discretionary	Comscore database is updated monthly with comScore top web properties, U.S. core search engine rankings - Shares of searches in percentage, U.S. core search Engine rankings-queries, U.S. core search engine rankings - qSearch Top 5, top U.S. video content properties by videos viewed, top U.S. online video content properties by unique viewers, top U.S. video content properties, top 15 Ad Networks/ Ad Focus rankings, top global properties based on total hours, top 10 countries by number of searches conducted, worldwide internet breakdown: regional breakdown, ad matrix total display ads by publisher.
5.	iResearch Data available from Bloomberg	Consumer Discretionary	iResearch provides Asia website traffic statistics.
6.	Magna Global Research Data available from Bloomberg	Consumer Discretionary	US advertising data and distribution data is updated four times in a year and global advertising data twice in a year by Magna Global Research.
7.	SNL Kagan Data available from Bloomberg	Consumer Discretionary	SNL Kagan's database has annual net advertising revenue, affiliate revenue, operating revenue and cash flow data for Cable and HD networks, broadcast networks, Telco, Home video and wireless companies.
8.	Veronis Suhler Stevenson Data available from Bloomberg	Consumer Discretionary	VSS provides quarterly US Advertising spending data.
9.	Smith Travel Data available from Bloomberg	Consumer Discretionary	Smith Travel provides monthly lodging data and occupancy rates.
10.	Chain Store Guide Information Services Data available from Bloomberg	Consumer Discretionary	Chain Store Guide Information services provide monthly US Chain Store data such as Department Stores, Apparel Specialty Stores, Discount Department Stores, Drug Stores and HBC Chains with Pharmacies, Home Center Operators and Hardware Chains, Home Furnishing Headquarters, Supermarket, Grocery Stores and Convenience Stores, Chain Restaurants, Single Unit Supermarkets ,High Volume Independent Restaurants, Wholesale Grocers and Foodservice Distributors.
11.	Experian Footfall Data available from Bloomberg	Consumer Discretionary	Experian Footfall provides monthly retail foot traffic data compiled for Western Europe.
12.	Redbook Research Inc. Data available from Bloomberg	Consumer Discretionary	Redbook Research provides retail Redbook Index: percentage change from week/month/year ago.
13.	Shoppertrak Rct Corporation	Consumer	Shoppertrack Corporation provides monthly US Retail traffic data.

Data available from Bloomberg		Discretionary	
14.	Borrell www.borrellassociates.com	Consumer Discretionary	Borrell works with more than 700 media properties, Internet "pure-play" companies, investment analysts, and industry vendors. Borrell's work focuses on helping companies understand and capitalize on the evolving media landscape, and to grow revenues exponentially rather than incrementally.
15.	Discern www.discern.com	Consumer Discretionary	Discern delivers insights for better investment decision-making. They provide up-to-date data on companies, retail stores, restaurants, oil wells and real estate.
16.	Orbital Insights www.orbitalinsight.com	Consumer Discretionary	With remarkable advancements in computer vision and cloud computing, Orbital Insight is building a geo-analytics platform for ingesting, processing, classifying and analyzing all types of geospatial data at massive scale creating unprecedented transparency, and empowering global decision makers with a new source of market insights. By analyzing millions of satellite images at a time, equips innovative industry leaders with advanced, unbiased knowledge of socio-economic trends.
17.	Placemeter www.placemeter.com	Consumer Discretionary	Placemeter uses advanced computer vision technology to lift data points from video streams and is robust and built to scale. First, the system handles an ever-increasing amount of video streams. Second, thanks to machine learning, the algorithms process video and classify objects in a wide range of new contexts. They keep track of retail data like: Store door counts, Pedestrian Traffic in front of your store, Street to purchase conversion rate, Impact of black Friday, other sales holidays and seasons.
18.	Sky Watch www.skywatch.co	Consumer Discretionary	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
19.	RS Metrics www.rsmetrics.com	Consumer Discretionary	RS Metrics is the leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.

Consumer Staples

S. No.	Company Name	Asset Class	Notes
1.	Euromonitor International Ltd Data available from Bloomberg	Consumer Staples	Euromonitor provides Global, North American, Asia Pac, Eastern Europe, Western Europe Total industry retail sales: Total category retail sales, Top 10 brands in each category by share for the latest year, top 10 country markets for each category by per capita consumption for the latest year and preceding 4 years, top 10 companies in each category by share, Forecast CAGR %, Smoking prevalence by region and country for the latest 5 years.
2.	FHS - Swiss Watch Data Data available from Bloomberg	Consumer Staples	FHS provides monthly watch manufacturing data.
3.	Tobacco Merchants Assoc Inc. Data available from Bloomberg	Consumer Staples	Tobacco Merchants Association provides World Cigarette Guide - Top 10 Producers of Cigarettes by country, Top 10 Consumers of Cigarettes by country, Top 10 Importers of Cigarettes by country, Top 10 Exporters of Cigarettes by country, Top-Selling Cigarette Brands, Domestic Cigarette Sales by Manufacturer, Indicated Retail Price for Cigarettes.
4.	Chain Store Guide Information Services Data available from Bloomberg	Consumer Staples	Chain Store Guide Information services provide monthly US Chain Store data such as Department Stores, Apparel Specialty Stores, Discount Department Stores, Drug Stores and HBC Chains with Pharmacies, Home Center Operators and Hardware Chains, Home Furnishing Headquarters, Supermarket, Grocery Stores and Convenience Stores, Chain Restaurants, Single Unit Supermarkets, High Volume Independent Restaurants, Wholesale Grocers and Foodservice Distributors.
5.	Experian Footfall Data available from Bloomberg	Consumer Staples	Experian Footfall provides monthly retail foot traffic data compiled for Western Europe.
6.	Redbook Research Inc. Data available from Bloomberg	Consumer Staples	Redbook Research Inc provides retail Redbook Index: percentage change from week/month/year ago.
7.	Shoppertrak Rct Corporation Data available from Bloomberg	Consumer Staples	Shoppertrack Corporation provides monthly US Retail traffic data.
8.	Discern www.discern.com	Consumer Staples	Discern delivers insights for better investment decision-making.
9.	Orbital Insights www.orbitalinsight.com	Consumer Staples	With remarkable advancements in computer vision and cloud computing, Orbital Insight is building a geo analytics platform for ingesting, processing, classifying and analyzing all types of geospatial data at massive scale creating unprecedented transparency, and empowering global decision makers with a new source of market insights. By analyzing millions of satellite images at a time, equips innovative industry leaders with advanced, unbiased knowledge of socio-economic trends.
10.	Placemeter www.placemeter.com	Consumer Staples	Placemeter uses advanced computer vision technology to lift data points from video streams and is robust and built to scale. First, the system handles an ever-increasing amount of video streams. Second, thanks to machine learning, the algorithms process video and classify objects in a wide range of new contexts. They keep track of retail data like : Store door counts, Pedestrian Traffic in front of your store, Street to purchase conversion rate, Impact of black Friday, other sales holidays and seasons.
11.	Sky Watch www.skywatch.co	Consumer Staples	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use

			the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
12.	RS Metrics www.rsmetrics.com	Consumer Staples	RS Metrics is the leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.

Energy/Utilities

S. No.	Company Name	Asset Class	Notes
1.	CDU-TEK: Central Dispatching Department of Fuel Energy Complex of Russia Data available from Bloomberg	Energy	Russian energy metrics and refinery outage data are provided by CDU TEK.
2.	InfoTEK Publishing House Data available from Bloomberg	Energy	Monthly gasoline import, export and production data in the Balkans are aggregated by InfoTEK Publishing House.
3.	Quest Offshore Data available from Bloomberg	Energy	Quarterly data is provided by Quest Offshore on subsea activity like floating rig contracts, rig assignments, oil well status and subsea awards and subsea capital expenditure. Their database includes Subsea Database and global prospects report.
4.	RigLogix Data available from Bloomberg	Energy	RigLogix provides offshore rig data like rig specifications, equipment details and rig status.
5.	Discern www.discern.com	Energy	Discern delivers insights for better investment decision-making with a focus on data for; listed companies, retail stores, restaurants, oil wells and real estate
6.	Orbital Insights www.orbitalinsight.com	Energy	With remarkable advancements in computer vision and cloud computing, Orbital Insight is building a geo-analytics platform for ingesting, processing, classifying and analyzing all types of geospatial data at massive scale creating unprecedented transparency, and empowering global decision makers with a new source of market insights. By analyzing millions of satellite images at a time, equips innovative industry leaders with advanced, unbiased knowledge of socio-economic trends.
7.	Sky Watch www.skywatch.co	Energy	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
8.	Genscape www.genscape.com	Utilities	Genscape has added satellite reconnaissance, Artificial Intelligence, and maritime freight tracking to its data acquisition capabilities. Genscape Maritime offers exclusive Vessel Coverage of U.S. Inland Waterways and covers 90% of US inland waterways. Genscape measures market fundamentals using thousands of patented and proprietary land, sea, and satellite monitors strategically deployed worldwide, delivering exceptional insight and intelligence to clients. Genscape uses non-contact ground sensors to monitor power lines; such data is used by power grid owners and operators

Financials

S. No.	Company Name	Asset Class	Notes
1.	ICI Data available from Bloomberg	Financials	ICI provides US Institutional Money Flow data and is updated weekly.
2.	Inside Mortgage Finance Publication Data available from Bloomberg	Financials	Inside Mortgage Finance provides top 25 mortgage originators and top 25 mortgage servicers on a quarterly basis.
3.	P&I Research Data available from Bloomberg	Financials	P&I Research provide annual assets under management by investment company demographics.
4.	First Data Merchant Services Corporation Data available from Bloomberg	Financials	First Data Merchant Services give US credit card spending trends, year on year change, prelim, mid-month, and final monthly numbers.
5.	LIMRA (LL Global Inc.) Data available from Bloomberg	Financials	Monthly US Life Insurance and Annuity Sales data are provided by LIMRA.
6.	Marketscout Corporation Data available from Bloomberg	Financials	Marketscout provides insurance sector data, and is referred to as the barometer information by line of business.
7.	Discern www.discern.com	Financials	Discern delivers insights for better investment decision-making.

Health Care

S. No.	Company Name	Asset Class	Notes
1.	Health Forum (AHA) Data available from Bloomberg	Health Care	Quarterly Healthcare financial data is provided by AHA.
2.	Millennium Research Group Inc. Data available from Bloomberg	Health Care	Quarterly Medical Device Market and Procedure data is available through Millennium Research Group.
3.	IMS Quintiles www.quintilesIMS.com	Health Care	QuintilesIMS is a Health Care service company. They bring proprietary data, new technologies and expert services together to help clients launch successful products

Industrials

S. No.	Company Name	Asset Class	Notes
1.	Airports Council International Data available from Bloomberg	Industrials	Airports Council International provides cargo data from the top 30 US airports and major international airports.
2.	Ascend Worldwide Limited Data available from Bloomberg	Industrials	Ascend Worldwide has global and regional breakdowns of airline programs and models parked and in service for each program.
3.	Arab Air Carrier Organization Data available from Bloomberg	Industrials	AACO provides monthly Total Aggregate RPKs and ASKs of member airlines. Quarterly Information are broken out by carrier and include Passengers Carried (PAX), Freight Lifted (FL), Revenue Passenger Kilometers (RPKs), Available Seat Kilometers (ASKs), Revenue Ton Kilometers (RTK), and Available Ton Kilometers (ATK). Yearly information represents detailed operations data of individual AACO members and will include all information published in the appendix of AACO AATS publications.
4.	Manfredi & Associates Data available from Bloomberg	Industrials	Quarterly updated market share data by product category, consumption is divided into 3 categories by Manfredi & Associates (quarterly updates): historic (annual) shipments, imports, exports and consumption data.
5.	SJ Consulting Group Inc. Data available from Bloomberg	Industrials	Quarterly and Annual transportation data is provided by SJ Consulting Group.
6.	Wards Automotive Group Data available from Bloomberg	Industrials	Wards Automotive Group's database contains monthly car class info including retail sales, inventory, factory shipments and Inventory/Retail sales ratio.
7.	American Trucking Associations Data available from Bloomberg	Industrials	ATA provides monthly Truck Tonnage Report.
8.	Datamyne Data available from Bloomberg	Industrials	Datamyne offers bill of lading information for imports and exports. It allows tracking of specific cargos.
9.	Drewry Shipping Consultants Ltd Data available from Bloomberg	Industrials	Drewry Shipping delivers a monthly update of air freight index from Shanghai to 5 destinations - London (LHR), Moscow (DME), Prague (PRG), New York (JFK) and Los Angeles (LAX) & weekly update of Shanghai to Rotterdam average spot freight rate in US\$ per full 40ft container load, excluding terminal handling charge at origin port.
10.	FTR Freight Transport Research Associates Data available from Bloomberg	Industrials	FTR provides monthly car class information, including retail sales, inventory, factory shipments, and Inventory/Retail sales ratio.
11.	Index Marketing Solutions Limited (World Container Index) Data available from Bloomberg	Industrials	Weekly shipping container rates are provided by Index Marketing Solutions.
12.	Intermodal Association of North America (IANA) Data available from Bloomberg	Industrials	Monthly Intermodal market trends & statistics are delivered by IANA
13.	Internet Truckstop Data available from Bloomberg	Industrials	Internet Truckstop provides weekly newsletter with proprietary data about the health of the trucking market called ITS TRANS4CAST letter.

14.	Off-Highway Research Limited Data available from Bloomberg	Industrials	Off Highway Research Limited provides annual market share by region, market share by product by region, regional sales data- totals, regional unit sales data by product and production unit data by region by product.
15.	Transport Topics Publishing Group Data available from Bloomberg	Industrials	Transport Topics Publishing Group provides a list of Annual Transport Topics (TT) Top 100 For-Hire Carriers, TT Top 100 Private Carriers, TT Top 50 Logistics Companies.
16.	VesselsValue Data available from Bloomberg	Industrials	VesselValue offers vessel valuations data.
17.	RS Metrics www.rsmetrics.com	Industrial	RS Metrics is a leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.

Information Technology/Telecommunication Services

S. No.	Company Name	Asset Class	Notes
1.	Semiconductor Equipment & Materials International (SEMI) Data available from Bloomberg	Information Technology	SEMI delivers worldwide semiconductor equipment market statistics.
2.	Everest Group Data available from Bloomberg	Information Technology	Everest Group provides annual and quarterly data on Labor Pool (ITO, BPO), delivery center set up, pricing for IT-ADM Services and pricing for non-Voice BPO Services for US, India, Europe Africa, LATAM, Mexico Philippines & China.
3.	InSpectrum Data available from Bloomberg	Information Technology	Mobile Dynamic random-access memory data is offered by InSpectrum.
4.	International Data Corporation Inc. Data available from Bloomberg	Information Technology	International Data Corporation publishes quarterly and annual data sets for Semiconductors, IT services and Telecommunications companies.
5.	Business Monitor International Data available from Bloomberg	Telcos	Business Monitor International provides annual Telecom industry data by worldwide regions.
6.	Ovum Ltd Us Branch Data available from Bloomberg	Telcos	Ovum offers a data pack providing an historical analysis and forecast of 1) fixed line voice, data revenues and connections and wireless voice and data revenues, connections & ARPU for the period 2000 to 2015, 2) smartphone and large screen device shipments and penetration for the period 2008 to 2015, 3) consumer broadband revenue and connections for the period 2000 to 2015 (2005 onwards for Mobile Consumer broadband).

Materials

S. No.	Company Name	Asset Class	Notes
1.	American Chemistry Council Data available from Bloomberg	Materials	ACA is one of the best global long-term databases of chemical output, trade data and consumption data. Subsectors level data including widely cited "chemicals utilization" indices grouped by region or aggregated worldwide.
2.	Beijing Chuang Yi Fang Technology (CHINA CCM) Data available from Bloomberg	Materials	China CCM offers Basic and Agricultural Chemical data for localized region, AsiaPac concentrating on China data.
3.	ChemOrbis Data available from Bloomberg	Materials	ChemOrbis provides a weekly update of Polymer pricing indices (Polypropylene, Poly Vinyl Chloride, Polyethylene and Polystyrene) for South East Asia, Egypt, Turkey, Italy and China.
4.	China National Chemical Information Center Data available from Bloomberg	Materials	Daily and Monthly update of China Fertilizer Market data is published by China National Chemical Information Center.
5.	Cropnosis Data available from Bloomberg	Materials	Cropnosis, through its Agrochemical Markets Database, allows users to analyze herbicide, insecticide, fungicide and other agrochemical sales by crop for each country and region; identify the main competing companies in the major markets and compare the agrochemical sales split globally and regionally with the leading country markets. Current and historical data on the area and production of all commercially important field and specialist crops by country and region are also included.
6.	Doane Advisory Service Data available from Bloomberg	Materials	Doane Advisory Services provides commodity pricing and supply demand forecasts for commodities like; Barley, Corn, Dairy, Hogs, Sorghum, Soybeans. Wheat, Cattle, Cotton, Rice, Soy Complex and Specialty Oils.
7.	Nexant Inc. Data available from Bloomberg	Materials	Nexant Inc offers Basic chemicals supply, demand and margin data for Global coverage.
8.	Tecnon Orbichem Data available from Bloomberg	Materials	Monthly Quarterly and Annual pricing, Contract production margin, Capacity, Production, Percent offline (outages), Net Exports, Consumption and Consumption by End Use data is supplied by Tecnon Orbichem.
9.	The Fertilizer Institute Data available from Bloomberg	Materials	The Fertilizer Institute gives monthly Chemical Production, Imports, Exports, Total Ending Inventory and Producer Disappearance for North America.
10.	ISSB Ltd Data available from Bloomberg	Materials	Quarterly metals pricing production and distribution data is provided by ISSB Ltd.
11.	World Bureau of Metal Statistics Data available from Bloomberg	Materials	World Bureau of Metal Statistics provides quarterly metals (including gold) pricing data.
12.	Beijing UC Science & Technology (CU Steel) Data available from Bloomberg	Materials	CU Steel offers China Steel Production, prices, inventories, coke inventories and China Steel Mills Information.
13.	Shanghai Metals Market Data available from Bloomberg	Materials	Monthly metals production and distribution data are published by Shanghai Metals Market.

14.	Steel Orbis Data available from Bloomberg	Materials	Steel Orbis offers monthly steel pricing data.
15.	U.S. Census Bureau - U.S. Coal Export/Import Data Data available from Bloomberg	Materials	The U.S. Census Bureau publishes monthly US Coal import/export data.
16.	Sky Watch www.skywatch.co	Materials	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
17.	RS Metrics www.rsmetrics.com	Materials	RS Metrics is the leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.

Real Estate

S. No.	Company Name	Asset Class	Notes
1.	China Real Estate Information Corporation (CRIC) Data available from Bloomberg	Real Estate	Real Estate data such as regional rents and vacancies are provided by CRIC.
2.	CB Richard Ellis Inc. Data available from Bloomberg	Real Estate	CB Richard Ellis offers quarterly Real Estate data such as regional rents and vacancies.
3.	NIC Data available from Bloomberg	Real Estate	NIC gives quarterly elderly housing statistics.
4.	Real Capital Analytics Data available from Bloomberg	Real Estate	Real Capital Analytics delivers data for different US Property Types: Office, Flex/Industrial, Apartment, Hotel, Regional Mall, and Community Shopping Center. Data provided include: 12 month moving average cap rate and price per square foot; total investment sales volume (\$, # of units, # of transactions, # of transactions with cap rate observations)
5.	REIS Data available from Bloomberg	Real Estate	REIS provide monthly Real Estate data such as regional rents and vacancies data.
6.	Discern www.discern.com	Real Estate	Discern delivers insights for better investment decision-making.
7.	RS Metrics www.rsmetrics.com	Real Estate	RS Metrics is the leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.

Text parsing

S. No.	Company Name	Asset Class	Notes
1.	Accern www.accern.com	Equity	Accern has rapidly grown to become the market leader in the news and social analytics space in terms of content coverage. Trading alerts and analytics are derived from relevant finance-related news processed on over 300 million public news websites, blogs, social media websites such as Twitter, and public financial documents such as SEC filings. Accern primarily serves institutional investors. The majority of our current clients are quantitative hedge funds. Many small firms are coming to us because of our flexible and affordable pricing options. Large firms are coming to us due to our dedicated support, news source customization, and much more. Aside from hedge funds, our existing clients include pension and endowment funds, banks, brokerage firms, and more.
2.	Edgar Online www.edgar-online.com	Equity	EDGAR® Online serves as the reporting engine for financial and regulatory filings for companies worldwide. This puts us in a unique position of having a huge repository of financial and company data. Financial statements, annual and quarterly filings, disclosures, initial and secondary public offering details, money market holdings, XBRL-tagged and as-reported summaries in its raw form is just data, but through EDGAR Online content subscriptions the data becomes a wealth of insight to drive your business decisions.
3.	Alphamatician www.alphamatician.com	Equity, Real Estate	Alphamatician collect data from dozens of sources that collectively build a clear picture of such things as company and brand health, product and product category pricing and demand trends, customer engagement, and company risk factors. And also track real estate markets and numerous other sources. Cover more than 5,000 publicly traded companies across North America, Europe and Asia. We also cover thousands of privately held companies and individual brands. Alphamatician provides access to dataset and insights of FaceBook, Instagram, E Commerce, Browser Statistics, Complaints, SEC filings, Twitter, Pinterest, Employment, Mobile Apps, Promotional Activity data
4.	Enigma www.enigma.io	Oil, Health Care, Equity	Enigma is building data discovery and analytics tools that make it simple for organizations to liberate their own private data that is locked away in data silos and obscure formats, just waiting to be released, and for the wider community to explore and build upon Enigma's own integrated public data platform. Enigma has created Signals, a set of applications that leverage billions of public data points to allow to automate intelligent decision-making at scale.
5.	Selerity www.seleritycorp.com	Event, Macro, Equity, Oil	Selerity is the market leader in detecting and delivering machine-readable event data in real-time as events are breaking. The Selerity Intelligence Platform pulls market-moving, factual information from public-only sources using proprietary real-time search and extraction technology. The product offering is designed for automated investment professionals looking to incorporate events into their risk management and trading strategies. Selerity offers the following market-moving event data: Global Economic Indicators, U.S. Corporate Earnings, Same Store Sales and Guidance, Monthly Auto Sales, U.S. Earnings Date Announcements, U.S. Dividends, Genscape Cushing Oil Storage Data, EIA energy Statistics, US M&A events and U.S. Earnings Pre-announcements.

Macroeconomic data

S. No.	Company Name	Asset Class	Notes
1.	MAC Data jpmm.com	Macro Economic	Collection of various global economic indicators maintained by J.P. Morgan's in-house economic team.
2.	EconData www.inforum.umd.edu	Macro Economic	Several thousand economic time series, produced by a number of U.S. Government agencies and distributed in a variety of formats and media, can be found here. These series include national income and product accounts (NIPA), labor statistics, price indices, current business indicators, and industrial production. The EconData Archive page contains U.S., State and Local, and International databanks that have not been updated for several years. These series are all in the form of banks for the G regression and model building program.
3.	ThinkNum www.thinknum.com	Equity	Analysts from around the world are building the largest repository of financial models on ThinkNum. ThinkNum monitors companies' websites and can be used to access the web's financial knowledge.
4.	OECD Data stats.oecd.org	Macro Economic	OECD Stat includes data and metadata for OECD countries and selected non-member economies. GDP, FDI, Health, unemployment, income distribution, population, labor, education, trade, finance, prices, Economic Outlook, Government Debt, Social expenditure data is available via OECD Stat.
5.	UN Data data.un.org	Macro Economic	The United Nations Statistics Division (UNSD) of the Department of Economic and Social Affairs (DESA) launched a new internet based data service for the global user community. It brings UN statistical databases within easy reach of users through a single entry point. The numerous databases, tables and glossaries containing over 60 million data points cover a wide range of themes including Agriculture, Crime, Education, Employment, Energy, Environment, Health, HIV/AIDS, Human Development, Industry, Information and Communication Technology, National Accounts, Population, Refugees, Tourism, Trade, as well as the Millennium Development Goals indicators.
6.	IMF DATA data.imf.org/	Macro Economic	The IMF publishes a range of time series data on IMF lending, exchange rates and other economic and financial indicators. Manuals, guides, and other material on statistical practices at the IMF, in member countries, and of the statistical community at large are also available.
7.	Trading Economics www.tradingeconomics.com	Commodity	Trading Economics provides its users with accurate information for 196 countries including historical data for more than 300,000 economic indicators, exchange rates, stock market indexes, government bond yields and commodity prices. Our data is based on official sources, not third party data providers, and our facts are regularly checked for inconsistencies.
8.	CESSDA www.cessda.net	Macro Economic	CESSDA stands for Consortium of European Social Science Data Archives. CESSDA provides large scale, integrated and sustainable data services to the social sciences. It brings together social science data archives across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation.

Accounting data

S. No.	Company Name	Asset Class	Notes
1.	Audit Analytics www.auditanalytics.com	Equity	Audit Analytics tracks in detail the issues related to audit, compliance, governance, corporate actions, and federal litigation from a broad range of public disclosures, including approximately 20,000 SEC company registrants. It quantifies disclosures related to the quality of financial reporting for public companies, and provides information on events such as financial restatement, SEC comment letter, or merger and acquisition.
2.	AAER dataset accounting.haas.berkeley.edu	Equity	Since 1982, the SEC has issued Accounting and Auditing Enforcement Releases (AAERs) during or at the conclusion of an investigation against a company, an auditor, or an officer for alleged accounting or auditing misconduct. The dataset currently consists of 3,813 SEC AAERs (1,540 firm misstatement events) issued since May 17th 1982. It contains 1,019 firm misstatement events that affect at least one of the firms' quarterly or annual financial statements.
3.	Markit Securities Finance www.markit.com	All Asset Classes	Markit (previously DataExplorer) provides global securities financing data. The dataset covers more than \$15 trillion of global securities in the lending market spanning over 20,000 institutional funds. It includes more than 3 million intraday transactions dating back 10 years. Markit crowdsources the data directly from leading industry practitioners in the stock lending market, including prime brokers, custodians, asset managers and hedge funds.
4.	IRS 990 Filings on AWS www.aws.amazon.com/public-datasets/irs-990	Equity	Machine-readable data from certain electronic 990 forms filed with the IRS from 2011 to present.

China/Japan data

S. No.	Company Name	Asset Class	Notes
1.	Toyo Keizai dbs.toyokeizai.net	Equity	Toyo Keizai provides fundamentals, estimates and shareholder data covering listed companies in Japan, as well as data on affiliated Japanese companies overseas. The earnings estimate data are based on independent research by analyst in Toyo Keizai. Other information is collected from the questionnaire surveys conducted by Toyo Keizai as far back as 40 years ago. Part of the Toyo Keizai database can also be accessed from Factset.
2.	DataYes www.datayes.com	Equity	DataYes is China's leading edge financial technology company founded by a group of experts in the fields of finance and internet technology. They provide comprehensive financial data, reports and news, as well as proprietary alternative data for in-depth sector research
3.	Gildata www.gildata.com	Equity	Gildata is a data service provider based in Shanghai, and is a subsidiary of financial software firm Hundsun Technologies. It provides financial data terminal, products and services through its platform
4.	Wind www.wind.com.cn/en	All Asset Classes	Wind is the premier provider of Chinese financial information, covering stocks, bonds, funds, indices, warrants, commodity futures, foreign exchanges, and the macro industry. It has a data-feed service that provides historical reference data, real-time market data, and historical intraday market data.
5.	Credit Ease creditease.cn	Equity, Consumer Credit	CreditEase is a leading FinTech company in China, specializing in small business and consumer lending as well as wealth management for high net worth and mass affluent investors. It is a Standing Committee member of China's Internet Finance Industry Association and Chairman of Beijing Marketplace Lending Association. Its majority owned subsidiary Yirendai (NYSE: YRD), an online consumer finance marketplace, is listed on the New York Stock Exchange. They disclose loan data to a public Chinese Marketplace Lender website: www.wdji.com
6.	China Money Network www.chinamoneynetwork.com	All Asset Classes	China Money Network is the go-to source of information and intelligence for millions of global investors looking to invest in the Greater China region. China Money Network's services include real time coverage of China's venture capital, private equity and institutional investment industries; a top-rated audio/video podcast featuring prominent China-focused investment managers; an alternative investment manager's database; offline events connecting top investors, a Chinese language platform ZhongguoJinrongTouziWang.com, and other value-added services bridging global investors with China opportunities.

3) Government Agencies Data

Federal or Regional Data

S. No.	Company Name	Asset Class	Notes
1.	Enigma www.enigma.io	Equity, Real Estate, Macro Economic	Enigma is building data discovery and analytics tools that make it simple for organizations to liberate their own private data, and for the wider community to explore and build upon Enigma's own integrated public data platform. Enigma's public data platform unifies billions of data points from more than 5,000 local, state, federal, and international sources. Explore things as diverse as SEC filings, government spending contracts, liens, patents, asset ownership, bills of lading, and much more.
2.	Global Open Data Index okfn.org	Macro Economic	The first initiative of its kind, the Global Open Data Index provides the most comprehensive snapshot available of the global state of open data.
3.	Socrata www.socrata.com	Macro Economic	Socrata is the market leader in Cloud-based Data Democratization solutions for Government (CDDG). Socrata provides a software-as-a-service (SaaS) data platform and cloud applications exclusively for city, county, state and federal government organizations, allowing government data to be discoverable, usable, and actionable for government knowledge workers and the citizens they serve.
4.	Canada Open Data www.open.canada.ca	Macro Economic	Government of Canada offers Government of Canada services, financials, national demographic information, high resolution maps and more through our open data portal, your one-stop shop for Government of Canada open datasets.
5.	DataFerrett dataferrett.census.gov	Macro Economic	DataFerrett is a data analysis and extraction tool to customize federal, state, and local data to suit your requirements. Using DataFerrett, you can develop an unlimited array of customized spreadsheets that are as versatile and complex as your usage demands then turn those spreadsheets into graphs and maps without any additional software.
6.	World Bank data.worldbank.org	Macro Economic, Equity, Energy, Commodity	World Bank provides free access to global development data by country and indicators like agriculture and rural development , aid effectiveness, climate change, economy and growth, financial sector social development energy and mining etc.
7.	Australia Data www.data.gov.au	Macro Economic	Australia Data provides a central catalogue to discover public data. It also provides hosting for tabular, spatial and relational data with hosted APIs and the option for agencies to link data and services hosted by other government sources. Improving the quantity and quality of the government data and the data.gov.au stack will be an ongoing process. In addition to open datasets, the data.gov.au catalogue includes unpublished data and data available for purchase.
8.	New Zealand www.data.govt.nz	Macro Economic	New Zealand Data is a directory of publicly-available New Zealand government datasets. The site focuses on machine-readable (ie, well-structured and open) datasets, but recognizes that "grey" data (for example, web pages) can still be of use to developers and others. We're working with agencies to improve data formatting. Official Statistics are produced by government departments to provide information for government and government departments, local government, businesses and the general public.
9.	Singapore Data data.gov.sg	Macro Economic	It aims to make government data relevant and understandable to the public, through the active use of charts and articles. Economics, Education, Finance, Transport, Technology data is available.
10.	Hong Kong Data	Macro	This is the public sector information portal of the Government of the Hong

	data.gov.hk	Economic	Kong Special Administrative Region. Data Categories: Climate Weather, Commerce and Industry, Employment and Labor, Food, Housing.
11.	US Gov. Data www.data.gov	Macro Economic	Open government data is important because the more accessible, discoverable, and usable data is, the more impact it can have. These impacts include, but are not limited to: cost savings, efficiency, fuel for business, improved civic services, informed policy, performance planning, research and scientific discoveries, transparency and accountability, and increased public participation in the democratic dialogue. Below are just a few examples of citizens leveraging open data. Topics include: Agriculture, Climate, Consumer, Energy, Fin, Manufacturing, Maritime.
12.	France Gov. Data data.gouv.fr	Macro Economic	This is an open platform for French public data.
13.	Germany Gov. Data www.govdata.de	Macro Economic	In the information area, you find everything worth knowing about the topics of Open Data, Open Government and citizen participation. Information is targeted for citizens as well as people from; economic, science, administration, and civil society organizations and media.
14.	China Gov. data www.stats.gov.cn	Macro Economic	The National Bureau of Statistics of the People's Republic of China or NBS is charged with the collection and publication of statistics related to the economy, population and society of the People's Republic of China at national and local levels.
15.	European Union Open data data.europa.eu	Macro Economic	The European Union Open Data Portal is the single point of access to a growing range of data from the institutions and other bodies of the European Union (EU). Data are free for you to use and reuse for commercial or non-commercial purposes. The EU Open Data Portal is managed by the Publications Office of the European Union. Implementation of the EU's open data policy is the responsibility of the Directorate-General for Communications Networks, Content and Technology of the European Commission.
16.	Japan Gov. Data www.data.go.jp	Macro Economic	This Site is the Data Catalog Site of the Japanese Government. Its purposes are to provide a sphere for the use of data owned by different governmental ministries and agencies as open data and to present the image of open data to both data providers and data users. It will encourage the use of open data and help collect examples of use. From an international perspective, publicity will be increased as an initiative of the Japanese government for an integrated open data site.
17.	South Africa southafrica.opendataforafrica.org	Macro Economic	This is an open platform for South Africa public data.
18.	Korea www.data.go.kr	Macro Economic	It is an open platform for Korea public data.
19.	Taiwan data.gov.tw	Macro Economic	It is an open platform for Taiwan public data.
20.	Norway www.data.norge.no	Macro Economic	Data.norge.no is a registry of open data in Norway. We also offer guidance as well as your hotel for businesses that want to use our technical infrastructure to publish their own data in machine readable formats. Examples of open data include financial statements and budgets. Results from surveys and Addresses and contact information for businesses.
21.	India Open Government Data www.data.gov.in	Macro Economic	Open Government Data Platform India has 4 (four) major modules; 1) Data Management System, 2) Content Management System, 3) Visitor Relationship Management, 4) Communities-Module for community users to interact and share their zeal and views with others who share common

			interests.
22.	Russia www.data.gov.ru	Macro Economic	This is an open government data platform for Russia
23.	Italy www.dati.gov.it	Macro Economic	This is an open government data platform for Italy.
24.	Spain www.datos.gob.es	Macro Economic	Open data initiative of government of Spain offers Public sector, Environmental, Healthcare, transport, employment, energy, housing data.
25.	Sweden www.opnadata.se	Macro Economic	Open data initiative of government of Sweden offers data related to the Environment, transport etc.
26.	UK www.data.gov.uk	Macro Economic	Open data initiative of government of UK offers data related Environment, transport, government spending, business and economy and Health etc.
27.	Brazil www.dados.gov.br	Macro Economic	Brazilian open data portal provides data on health, transportation, traffic, housing and environment data.
28.	Mexico www.datos.gob.mx	Macro Economic	Open data initiative of government of Mexico.
29.	Data Catalogs www.opengovernmentdata.org	Macro Economic	This website is a project and home of the Working Group on Open Government Data at the Open Knowledge Foundation. Open Government Data means i) data produced or commissioned by government or government controlled entities. ii) Data which can be freely used, reused and redistributed by anyone.
30.	CIA Data www.cia.gov	Macro Economic	The World Factbook provides information on the history, people, government, economy, geography, communications, transportation, military, and transnational issues for 267 world entities. Reference tab includes: maps of the major world regions, as well as Flags of the World, a Physical Map of the World, a Political Map of the World, a World Oceans map, and Standard Time Zones of the World.
31.	US Healthcare Data www.healthdata.gov	Macro Economic	This site is dedicated to making high value health data more accessible to entrepreneurs, researchers, and policy makers in the hopes of better health outcomes for all. And it has over 3100 datasets and over 7000 resources.
32.	UK Healthcare Data www.content.digital.nhs.uk	Macro Economic	Trusted national provider of high-quality information, data and IT systems for health and social care.
33.	UNICEF Data www.data.unicef.org	Macro Economic	UNICEF maintains a series of global databases for tracking the situation of children and women globally. The databases include only statistically sound and nationally representative data from household surveys and other sources. They are updated annually through a process that draws on a wealth of data maintained by UNICEF's network of 140 country offices.
34.	Global Health Observatory www.who.int	Macro Economic	Global Health Observatory is a gateway to health related statistics for more than 1000 indicators for its 194 member states.
35.	Trade and Tariff Data www.wto.org	Macro Economic	Trade and tariff data provided from World Trade Organization.

Federal Government Sponsored Enterprise Data

S. No.	Company Name	Asset Class	Notes
1.	US Census Bureau www.census.gov	Macro Economic	The Census Bureau's mission is to serve as the leading source of quality data about the nation's people and economy. We collect Decennial Census of Population and Housing, Economic Census, Census of Government, American Community Survey, Economic Indicators.
2.	International Labour Organization www.ilo.org	Macro Economic	It is a collection of databases connected to Labour unemployment characteristics
3.	United Nations data.un.org	Macro Economic	A good collection of historical data points. That can serve as a useful backdrop for real time data from elsewhere
4.	OECD.Stat stats.oecd.org	Macro Economic	OECD.Stat includes data and metadata for OECD countries and selected non-member economies. GDP, FDI, Health, unemployment, income distribution, population, labor, education, trade, finance, prices, Economic Outlook, Government Debt, Social expenditure data is available via OECD.Stat.
5.	Asset Macro www.assetmacro.com	Macro Economic	Access to Historical & Live Data of 20,000 Financial Data & Macroeconomic Indicators of Global Markets & Economies.
6.	The BIS International Financial Statistics www.bis.org	Macro Economic	Bank for international settlement has data of 1) International Banking and Financial statistics a) Banking b) Debt securities c) Derivative statistics 2) Government Financial conditions a) Global liquidity indicators b) Credit to the non-financial sector c) Credit to GDP sector d) Debt service ratios e) External Debt statistics 3) Price and Exchange Rates a) Consumer prices b) Property prices c) Effective exchange rates 4) Payment Systems a) Statistics on payment, clearing and settlement systems.
7.	Qlik qlik.com	Macro Economic	Qlik recently acquired DataMarket. The DataMarket dataset allows users to visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers
8.	FedStats fedstats.sites.usa.gov	Macro Economic	A trusted source for federal statistical information since 1997. FedStats supports a community of practice for over 100 agencies engaged in the production and dissemination of official federal statistics, and provides the public with a showcase of information, tools and news related to official federal statistics.
9.	IMF Data data.imf.org	Macro Economic	IMF Data provide access to macroeconomic and financial data and has data related to external sector, fiscal sector, financial and real sector.
10.	Haver Analytics www.haver.com	Macro Economic	Haver Analytics updates and manages 200 plus databases from over 1350 government and private sources. The data offering ranges from daily market data to annual economic statistics and most time sensitive data are updated within minutes of release. Global economic and financial coverage includes country-sourced macro detail, financial indicators, daily & weekly, industry, international organizations, forecasts and as reported, and U.S. regional.

C. Data from Sensors

1) Satellites

Satellite imagery for agriculture

S. No.	Company Name	Asset Class	Notes
1.	Orbital Insights www.orbitalinsight.com	Agriculture, Oil, Equity	With remarkable advancements in computer vision and cloud computing, Orbital Insight is building a geo-analytics platform for ingesting, processing, classifying and analyzing all types of geospatial data at massive scale creating unprecedented transparency, and empowering global decision makers with a new source of market insights. By analyzing millions of satellite images at a time, equips innovative industry leaders with advanced, unbiased knowledge of socio-economic trends.
2.	Rezatec www.rezatec.com	Agriculture, Commodity	Rezatec is the specialist geospatial data analytics company providing valuable and actionable landscape intelligence as a critical decision-support tool to help drive the smart management of land-based assets and serves customers around the world, spread across the Agribusiness, Oil & Energy, Water, Forestry, Urban Infrastructure, Commodities and FMCG sectors. Satellite-derived geospatial analytical insights for Commodities including: Wheat, Maize, Rice, Coffee, Cotton, Sugar, Metals, Energy and Timber. Using the most appropriate Earth Observation techniques from a multitude of data sources including satellite optical and RADAR data, aerial and unmanned aerial vehicle (UAV) imagery and ground-based observations, Rezatec provide decision support tools to help agribusinesses.
3.	Genscape www.genscape.com	Commodity, Agriculture, Freight.	Genscape has added satellite reconnaissance, Artificial Intelligence, and maritime freight tracking to its data acquisition capabilities. Genscape Maritime offers exclusive Vessel Coverage of U.S. Inland Waterways and covers 90% of US inland waterways. Genscape measures market fundamentals using thousands of patented and proprietary land, sea, and satellite monitors strategically deployed worldwide, delivering exceptional insight and intelligence to clients. Genscape uses non-contact ground sensors to monitor power lines; such data is used by power grid owners and operators
4.	Sky Watch www.skywatch.co	Commodity, Agriculture, Oil	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
5.	Digital Globe www.digitalglobe.com	Energy, Commodity	Digital Globe's satellites collect over three million square kilometers a day, giving access to all parts of the globe and revisit at a high rate, so one can detect changes over extremely short periods of time. Sixteen year image library has unmatched spatial resolution and global coverage. Provides elevation and terrain information which is foundational to mapping and understanding the surface of our planet. With global coverage and the ability to order custom-built models, DigitalGlobe's Advanced Elevation Series provides the Digital Surface Models or Digital Terrain Models you need without the challenges of local access with options ranging from 2 m to 8 m in accuracy and resolution.
6.	TellusLabs www.telluslabs.com	Agriculture	TellusLabs combines decades of satellite imagery with a machine learning platform to answer critical, time-sensitive economic and environmental questions. They use terabyte-scale satellite and weather database that is updated every 24 hours with the latest plant health, local weather, and crop conditions to train across decades of historical data and all 175 million acres of US corn and soy cropland. Tellus Lab's game-changing technology enable efficient and intuitive access to very large and complex planetary-scale geospatial data sets, combined with state-of-the-art computing and machine learning tools.

7.	HySpecIQ www.hyspeciq.com	Commodity, Agriculture, Energy	HySpecIQ will unlock the power of high resolution hyperspectral imaging with latent power of data fusion, cloud computing and advanced big data analytical algorithms derived from our low earth orbiting small satellites to serve national security, civil and commercial clients. Our tailored hyperspectral-driven decision support tools, risk management solutions and monitoring capabilities will be used to develop important applications in natural resource industries such as Mining, Oil & Gas, Agriculture as well as Insurance and Government.
8.	DroneDeploy www.dronedeploy.com	Agriculture, Real Estate, Commodity	DroneDeploy allows automating drone flights and exploring map data directly from app. They provide Orthomosaics, terrain models, NDVI analysis and 3D models for Agriculture, Construction and Mining industries. Make real time measurements including distance, area and volume. View NDVI helps in detecting crop stress and variability

Satellite imagery for Maritime

S. No.	Company Name	Asset Class	Notes
1.	Windward www.windward.eu	Freight	The Windward Mind is the only platform in the world that aggregates, analyzes and vets all maritime data, globally 24/7. It takes the data and creates a unique Vessel Story on each ship worldwide, a complete history of each ship's activities over time. With over 90% of the world's trade transported over the oceans, data on ship activity is critical to decision makers across industries.
2.	Spire www.spire.com	Agriculture, Commodity	Spire uses satellite imagery to provide maritime, weather and aviation data that covers 90% of global trade. By utilizing a network of tens or hundreds of CubeSats with radio occultation payloads, the number of opportunities to receive radio occultation increases dramatically which results in better forecasts. Spire uses Satellite AIS which offers better coverage of remote places than other Terrestrial AIS (more effective in densely packed ports) providers.
3.	Genscape www.genscape.com	Commodity, Agriculture, Freight.	Genscape has added satellite reconnaissance, Artificial Intelligence, and maritime freight tracking to its data acquisition capabilities. Genscape Maritime offers exclusive Vessel Coverage of U.S. Inland Waterways and covers 90% of US inland waterways. Genscape measures market fundamentals using thousands of patented and proprietary land, sea, and satellite monitors strategically deployed worldwide, delivering exceptional insight and intelligence to clients. Genscape uses non-contact ground sensors to monitor power lines; such data is used by power grid owners and operators
4.	MarineTraffic www.marinetraffic.com	Freight	MarineTraffic maintains a database of real-time and historical ship positions sourced from the largest station network and satellite constellation. Historical AIS data from Terrestrial AIS receivers or Satellite AIS receiver is available since 2009. Maritime have been analyzing AIS positions data in relation to port boundaries and have extracted the time of arrival and departure of vessels in ports all around the world.
5.	Vessel Finder www.vesselfinder.com	Freight	Vessel Finder supply world-wide vessel position data via Real-Time XML/JSON data and Raw NMEA data stream which is offered to make timely decisions, manage costs and optimize performance. Historical AIS data is also used for analyzing the vessels movements on a global scale, potential trends in the shipping market or vessel behavior patterns for prosecution of illegal activities. The historical data include video simulation, vessel movement report, port calls, traffic density analysis and even tracking of thousands of ships around the world. The vessel movement report service is based on historical AIS data collected by terrestrial AIS stations since 2009.
6.	Sky Watch www.skywatch.co	Commodity, Agriculture, Oil	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
7.	Marinexplore www.marinexplore.org	Freight	Marinexplore is a platform to explore, discover, and share public ocean data. Our data hub catalog provides over 2,000 unique parameters and is updated weekly with new datasets.

Satellite imagery for Metals and mining

S. No.	Company Name	Asset Class	Notes
1.	Terra Bella (Google Skybox) www.terrabella.google.com	Real Estate, Equity, Commodity	Planet has signed an agreement to acquire the Terra Bella business from Google. Terra Bella provides commercial high-resolution Earth observation satellite imagery, high-definition video and analytics services.
2.	RS Metrics www.rsmetrics.com	Real Estate, Commodity, Energy	RS Metrics is the leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.
3.	Digital Globe www.digitalglobe.com	Energy, Commodity	Digital Globe's satellites collect over three million square kilometers a day, giving access to all parts of the globe and revisit at a high rate, so one can detect changes over extremely short periods of time. Sixteen year image library has unmatched spatial resolution and global coverage. Provides elevation and terrain information which is foundational to mapping and understanding the surface of our planet. With global coverage and the ability to order custom-built models, DigitalGlobe's Advanced Elevation Series provides the Digital Surface Models or Digital Terrain Models you need without the challenges of local access with options ranging from 2 m to 8 m in accuracy and resolution.
4.	HySpecIQ www.hyspeciq.com	Commodity, Agriculture, Energy	HySpecIQ will unlock the power of high resolution hyperspectral imaging with latent power of data fusion, cloud computing and advanced big data analytical algorithms derived from our low earth orbiting small satellites to serve national security, civil and commercial clients. Our tailored hyperspectral-driven decision support tools, risk management solutions and monitoring capabilities will be used to develop important applications in natural resource industries such as Mining, Oil & Gas, Agriculture as well as Insurance and Government.

Satellite imagery for Company parking

S. No.	Company Name	Asset Class	Notes
1.	Terra Bella (Google Skybox) www.terrabella.google.com	Real Estate, Equity, Commodity	Planet has signed an agreement to acquire the Terra Bella business from Google. From building satellites to writing code, they are changing the way the world looks at big problems. Terra Bella provides commercial high-resolution Earth observation satellite imagery, high-definition video and analytics services.
2.	Orbital Insights www.orbitalinsight.com	Agriculture, Oil, Macro Economic	With remarkable advancements in computer vision and cloud computing, Orbital Insight is building a geo-analytics platform for ingesting, processing, classifying and analyzing all types of geospatial data at massive scale creating unprecedented transparency, and empowering global decision makers with a new source of market insights. By analyzing millions of satellite images at a time, equips innovative industry leaders with advanced, unbiased knowledge of socio-economic trends.
3.	Placemeter www.placemeter.com	Equity, Real Estate	Placemeter uses advanced computer vision technology to lift data points from video streams and is robust and built to scale. First, the system handles an ever-increasing amount of video streams. Second, thanks to machine learning, the algorithms process video and classify objects in a wide range of new contexts. They keep track of retail data like : Store door counts, Pedestrian Traffic in front of your store, Street to purchase conversion rate, Impact of black Friday, other sales holidays and seasons.
4.	Sky Watch www.skywatch.co	Commodity, Agriculture and Oil	Skywatch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
5.	RS Metrics www.rsmetrics.com	Real Estate, Commodity Energy	RS Metrics is the leading provider of applications and data from large-scale analysis of satellite and aerial imagery, and other geospatial information sources.

Satellite imagery for Oil

S. No.	Company Name	Asset Class	Notes
1.	URSA www.ursaspace.com	Oil	URSA provides valuable information from satellite imagery acquired up to twice a day, anywhere in the world. Ursa's Oil Storage data feed uses radar satellite imagery to measure storage level changes of major oil locations around the world. They focus on critical locations that have little or no data available like China.
2.	Rezatec www.rezatec.com	Agriculture, Commodity	Rezatec is the specialist geospatial data analytics company providing valuable and actionable landscape intelligence as a critical decision-support tool to help drive the smart management of land-based assets and serves customers around the world, spread across the Agribusiness, Oil & Energy, Water, Forestry, Urban Infrastructure, Commodities and FMCG sectors. Satellite-derived geospatial analytical insights for Commodities including: Wheat, Maize, Rice, Coffee, Cotton, Sugar, Metals, Energy and Timber. Using the most appropriate Earth Observation techniques from a multitude of data sources including satellite optical and RADAR data, aerial and unmanned aerial vehicle (UAV) imagery and ground-based observations, Rezatec provide decision support tools to help agribusinesses.
3.	CargoMetrics www.cargometrics.com	Freight, Equity	CargoMetrics is a technology driven investment company which links satellite signals, historical shipping data and proprietary analytics for its own trading.
4.	Orbital Insights www.orbitalinsight.com	Agriculture , Oil, Macro Economic	With remarkable advancements in computer vision and cloud computing, Orbital Insight is building a geo-analytics platform for ingesting, processing, classifying and analyzing all types of geospatial data at massive scale creating unprecedented transparency, and empowering global decision makers with a new source of market insights. By analyzing millions of satellite images at a time, equips innovative industry leaders with advanced, unbiased knowledge of socio-economic trends.
5.	Genscape www.genscape.com	Commodity, agriculture, freight.	Genscape has added satellite reconnaissance, Artificial Intelligence, and maritime freight tracking to its data acquisition capabilities. Genscape Maritime offers exclusive Vessel Coverage of U.S. Inland Waterways and covers 90% of US inland waterways. Genscape measures market fundamentals using thousands of patented and proprietary land, sea, and satellite monitors strategically deployed worldwide, delivering exceptional insight and intelligence to clients. Genscape uses non-contact ground sensors to monitor power lines; such data is used by power grid owners and operators
6.	ClipperData www.clipperdata.com	Oil	ClipperData provides comprehensive data sets, analysis and commentary on global crude and refined product movements. Not only do they track movements on a vessel-by-vessel basis, but also cargo-by-cargo, and dock-to-dock. They track the tanker beyond the destination port all the way to the ultimate discharge facility which means they can draw a picture of refinery operations that others cannot. They use a combination of unique data collection techniques and exclusive licensing arrangements with best-in-the-business data providers. We also apply proprietary algorithms to build our data sets.
7.	Sky Watch www.skywatch.co	Commodity, Agriculture, Oil	Sky watch's satellite data is used by people in the mining industry to inspect areas of interest; by oil companies to monitor pipelines; and by farmers to monitor their crops. Market intelligence companies also use the data to count cars in parking lots and the number of construction cranes that are in use across the entire country. Sky watch wants to make it easy to monitor the progress of crops, predict the markets, track ship and airplanes, measure global warming, or create other game-changing applications.
8.	Satellite Imaging Corporation www.satimagingcorp.com	Agriculture, Commodity	Satellite Imaging Corporation (SIC) specializes in satellite imaging collections, image processing and producing seamless ortho rectified imaging mosaics, 3D Digital Elevation Models (DEM's) and 3D Digital Terrain Models for many industries including Engineering and Construction, Defense and Security, Environmental Monitoring, Media and Entertainment, Natural Resources, Tourism, Energy, Mining, Oil and Gas, Oil and Gas Development, Oil and Gas Exploration,

Oil and Gas Production, Pipeline and Transmission.

9. Digital Globe
www.digitalglobe.com Energy,
Commodity

Digital Globe's satellites collect over three million square kilometers a day, giving access to all parts of the globe and revisit at a high rate, so one can detect changes over extremely short periods of time. Sixteen year image library has unmatched spatial resolution and global coverage. Provides elevation and terrain information which is foundational to mapping and understanding the surface of our planet. With global coverage and the ability to order custom-built models, DigitalGlobe's Advanced Elevation Series provides the Digital Surface Models or Digital Terrain Models you need without the challenges of local access with options ranging from 2 m to 8 m in accuracy and resolution.

10. HySpecIQ
www.hyspeciq.com Commodity,
Agriculture,
Energy

HySpecIQ will unlock the power of high resolution hyperspectral imaging with latent power of data fusion, cloud computing and advanced big data analytical algorithms derived from our low earth orbiting small satellites to serve national security, civil and commercial clients. Our tailored hyperspectral-driven decision support tools, risk management solutions and monitoring capabilities will be used to develop important applications in natural resource industries such as Mining, Oil & Gas, Agriculture as well as Insurance and Government.

2) Geolocation

S. No.	Company Name	Asset Class	Notes
1.	Spire www.spire.com	Agriculture, Commodity	Spire uses satellite imagery to provide maritime, weather and aviation data that covers 90% of global trade. By utilizing a network of tens or hundreds of CubeSats with radio occultation payloads, the number of opportunities to receive radio occultation increases dramatically which results in better forecasts. Spire uses Satellite AIS which offers better coverage of remote places than other Terrestrial AIS (more effective in densely packed ports) providers.
2.	Understory www.understoryweather.com	Equity, Agriculture, Natural Gas, Commodity	Understory captures hyper-local weather data and provides analytics on the same. With sensors collecting 3000 measurements per second on rainfall, hail/wind speed, temperature, heat index, the company collates observations across distances ranging from kilometers to 100 m resolution.
3.	The Climate Corporation www.climate.com	Agriculture	The Climate Corporation's proprietary Climate FieldView™ platform combines hyper-local weather monitoring, agronomic modeling, and high-resolution weather simulations to deliver Climate FieldView products and mobile SaaS solutions. These provide an overview of the global \$3 trillion agriculture industry.
4.	Descartes Labs www.descarteslabs.com	Commodity, Agriculture	Descartes labs have full imagery archives from hundreds of satellites. The Descartes Platform is built to ingest virtually any kind of data, including satellite , weather data, commodity price histories, web crawls, and sentiment analysis from social media networks. Currently, the Descartes Platform ingests 5 terabytes (TB) of near real-time data per day, roughly equivalent to 5,000 hours of standard video. Our current corpus is over 3 petabytes of data (3,000 TB) with the ability to grow much larger. With sensor data growing exponentially, the Descartes Platform is designed to respond elastically to this data explosion and harness it for real-time forecasting. Descartes Labs provides forecasts for different commodities (including cotton, rice, soy and wheat) across different growing regions (including US, Brazil, Argentina, Russia and Ukraine).
5.	AirSage www.airsage.com	Equity	AirSage provides consumer location and movement data nationwide based on mobile device activity. AirSage anonymously collects and analyzes wireless signaling data – processing more than 15 billion mobile device locations everyday – turning them into meaningful and actionable information. They cover three industries: Transportation, Tourism and Market Research.
6.	Placed www.placed.com	Equity	Placed measures 1 billion locations a day, across more than 1 million opted-in smartphones. This data is then run through the Inference Pipeline which references a place database with nearly 300 million features for the US alone. The Inference Pipeline doesn't require panelists to check-in, complete surveys, or keep a journal; rather the models predict a place using directly measured device data. Placed Targeting leverages location (latitude / longitude), not individuals. Provides Top 100 monthly business list, ranked by highest in-store foot traffic.
7.	StreetLight Data www.streetlightdata.com	Real Estate	StreetLight's technology contextualizes anonymous location data from mobile devices to measure population mobility patterns. Their solution transforms messy and disparate spatial data into contextual metrics for the real world. StreetLight utilizes a multitude of data inputs in conjunction with proprietary Route Science® engine.
8.	Placemeter www.placemeter.com	Consumer Discretionary	Placemeter uses advanced computer vision technology to lift data points from video streams and is robust and built to scale. First, the system handles an ever-increasing amount of video streams. Second, thanks to machine learning, the algorithms process video and classify objects in a wide range of new contexts. They keep track of retail data like: Store door counts, Pedestrian Traffic in front of your store, Street to purchase conversion rate, Impact of black Friday, other sales holidays and seasons.
9.	Factual	Equity	Factual provides location data for enterprise solutions. Factual gather raw data from millions of different sources, clean and structure it, and then package and distribute it in multiple ways to make data easier for the world to use. Factual

	www.factual.com		focuses on location data — data about where places are and how to better understand people based on their geographic behavior. They have APIs, mobile drivers, and on-premise implementations to make the data easy to use and integrate.
1.	Advan www.advan.us	Equity	Advan was established to bridge the gap between Big Data and investment decisions. Advan's attention to detail brings turn-key products to the Financial Trader and Analyst that offer unique and timely insights into several hundred companies across multiple sectors.
2.	Sense360 www.sense360.com	Equity	Sense360 uses the mobile sensors built into smartphones to understand a user's location and activity. It has trained algorithms with hundreds of thousands of labeled data points and incorporate data from GPS, accelerometer, gyroscope, barometer, Wi-Fi, ambient light, many other sensors. This provides with an anonymous, but highly accurate understanding of where, how, and when people interact with physical locations and businesses. Sense360 covers various industries including Retail, C-Store, Supermarkets, Hospitality and Automotive.
3.	Google trends www.trends.google.com	All Asset Classes	Google has begun to provide access to aggregated information on the volume of queries for different search terms, and how these volumes change over time via the public service called Google trends. For instance, the Google Trends Uncertainty Index has been used to measure economic and political uncertainties
4.	GeoWiki www.geo-wiki.org	Equity	Geo-wiki aids in both the validation of existing geographical information and the collection of new geographical information through crowdsourcing. Provides data collected via the Geo-Wiki in raw format, with no filtering or error checking applied.
5.	Wikimapia www.wikimapia.org	Equity	Wikimapia is an open-content collaborative mapping project, aimed at marking all geographical objects in the world and providing a useful description of them. It combines an interactive web map with a wiki system. Both registered users and guests have already marked over 20,000,000 objects and this number grows every day.
6.	OpenstreetMap www.openstreetmap.org	Equity	OpenstreetMap is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world. NStreetMap is a map of the world, created by people in the community and is free to use under an open license.

3) Other Sensors

S. No.	Company Name	Asset Class	Notes
1.	Space Know www.spaceknow.com	Equity, Agriculture	Space Know's mission is to bring transparency to the global economy by tracking global economic trends from space through their Analytics-as-a-Service products. Complement traditional surveys and statistics with satellite data and social media data and measure economic growth with far greater accuracy than ever before and get the whole picture of agricultural activities and make objective estimations of seasonal yields of commodities to optimize land usage. Space know has developed two trading indices using unique and customizable satellite imagery data 1) China SMI which uses proprietary algorithms to monitor over 6,000 industrial facilities across China to measure levels of manufacturing activity and 2) Space know Africa Night Light Index is based on the nighttime light intensity measured by satellites and then aggregated according to individual countries.
2.	Black Sky www.blacksky.com	All Asset Classes	Raw unprocessed satellite images. Commentary pieces available Black Sky's planned constellation of 60 satellites – with six spacecraft on orbit by the end of 2017 will provide frequent revisit rates over 95% of the Earth's population. BlackSky provides color imagery at a resolution of one meter (1 square meter = 1 image pixel) which makes monitoring economic activity easier than ever – see ships in ports, earthquake damage, or herd migration. We combine satellite imagery, social media, news and other data feeds to create timely and relevant insights.
3.	Urthecast www.urthecast.com	All Asset Classes	UrtheDaily™ is a planned global coverage constellation aiming to acquire high-quality, multispectral imagery, at 5-m GSD, taken at the same time, from the same altitude every day. With its exceptional capabilities, it presents a disruptive and problem-solving disruptive technology that will transform the way we observe our planet. Urthecast provides Earth imagery and video at multiple resolutions to help solve complex problems. Deimos-1 provides 22m resolution imagery for precision agriculture and forestry monitoring: 3-band multispectral imagery with a wide swath of 650 km. In addition to its 6-year archive, Deimos-1 focuses on specific areas like the continental U.S., which is captured in full every month.
4.	Satellite Imaging Corporation www.satimagingcorp.com	Agriculture, Commodity, Oil	Satellite Imaging Corporation (SIC) specializes in satellite imaging collections, image processing and producing seamless ortho rectified imaging mosaics, 3D Digital Elevation Models (DEM's) and 3D Digital Terrain Models for many industries including Engineering and Construction, Defense and Security, Environmental Monitoring, Media and Entertainment, Natural Resources, Tourism, Energy, Mining, Oil and Gas, Oil and Gas Development, Oil and Gas Exploration, Oil and Gas Production, Pipeline and Transmission. In most instances, SIC can provide image data within 24 hours after the initial data has been acquired and delivered.
5.	Digital Globe www.digitalglobe.com	Energy, Commodity	Digital Globe's satellites collect over three million square kilometers a day, giving access to all parts of the globe and revisit at a high rate, so one can detect changes over extremely short periods of time. Sixteen year image library has unmatched spatial resolution and global coverage. Provides elevation and terrain information which is foundational to mapping and understanding the surface of our planet. With global coverage and the ability to order custom-built models, DigitalGlobe's Advanced Elevation Series provides the Digital Surface Models or Digital Terrain Models you need without the challenges of local access with options ranging from 2 m to 8 m in accuracy and resolution.
6.	Planet Labs www.planet.com	Agriculture, Energy	Planet's satellites image every location on earth, every day, at high resolution — providing a vital new data source. The Planet Platform is built to ingest 150 million km2 of imagery each day. Provide fresh insights from yesterday or look back a few years. The Planet Platform is the first fully automated image processing pipeline that seamlessly downloads, transfers to the cloud, orthorectifies and processes more than 5 terabytes of data per day. Their worldwide

			imagery archive dates back to 2009.
7.	HySpecIQ www.hyspeciq.com	Commodity, Agriculture, Energy	HySpecIQ will unlock the power of high resolution hyperspectral imaging with latent power of data fusion, cloud computing and advanced big data analytical algorithms derived from our low earth orbiting small satellites to serve national security, civil and commercial clients. Our tailored hyperspectral-driven decision support tools, risk management solutions and monitoring capabilities will be used to develop important applications in natural resource industries such as Mining, Oil & Gas, Agriculture as well as Insurance and Government.
8.	Descartes Labs www.descarteslabs.com	Commodity, Agriculture	Descartes Labs have full imagery archives (some including data only a few hours old) from hundreds of satellites. The Descartes Platform is built to ingest virtually any kind of data, including satellite, weather data, commodity price histories, web crawls, and sentiment analysis from social media networks. Currently, the Descartes Platform ingests 5 terabytes (TB) of near real-time data per day, roughly equivalent to 5,000 hours of standard video. Our current corpus is over 3 petabytes of data (3,000 TB) with the ability to grow much larger. With sensor data growing exponentially, the Descartes Platform is designed to respond elastically to this data explosion and harness it for real-time forecasting. Descartes Labs provides forecasts for different commodities (including cotton, rice, soy and wheat) across different growing regions (including US, Brazil, Argentina, Russia and Ukraine).
9.	Amazon Web Services www.aws.amazon.com/public-datasets	Equity, Real Estate	Ongoing collection of satellite imagery is available.
10.	Landsat on AWS: www.aws.amazon.com/public-datasets/landsat	All Asset Classes	An ongoing collection of moderate-resolution satellite imagery of all land on Earth produced by the Landsat 8 satellite.
11.	SpaceNet on AWS www.aws.amazon.com/public-datasets/spacenet	Equity, Real Estate	A corpus of commercial satellite imagery and labeled training data to foster innovation in the development of computer vision algorithms. The current SpaceNet corpus includes approximately 1,900 square kilometers full-resolution 50 cm imagery collected from DigitalGlobe's WorldView-2 commercial satellite and includes 8-band multispectral data.
12.	Terrain Tiles www.aws.amazon.com/public-datasets/terrain	Equity, Real Estate	A global dataset providing bare-earth terrain heights, tiled for easy usage and provided on S3.
13.	GDELT www.aws.amazon.com/public-datasets/gdelt	Equity, Real Estate	The Global Database of Events, Language and Tone (GDELT) Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day.
14.	NAIP www.aws.amazon.com/public-datasets/naip	Agriculture	The National Agriculture Imagery Program (NAIP) acquires aerial imagery during the agricultural growing seasons in the continental U.S. This "leaf-on" imagery is approximately 1 meter in resolution.
15.	NEXRAD on AWS www.aws.amazon.com/noaa-big-data/nexrad	Agriculture	Real-time and archival data from the Next Generation Weather Radar (NEXRAD) network. The real-time feed and full historical archive of original resolution (Level II) NEXRAD data, from June 1991 to present, is now freely available on Amazon S3 for anyone to use.
16.	Capella Space www.capellaspace.com	Equity, Real Estate	Capella Space is a data company that will provide persistent and reliable information from space with its planned constellation of (very small) satellites every hour. Capella's satellites can see through the clouds and at night time making them immune to bad weather or light conditions. They monitor changes constantly with no interruptions.

17.	DroneDeploy www.dronedeploy.com	Agriculture, Real Estate, Commodity	Dronedeploy allows automating drone flights and exploring map data directly from app. They provide Orthomosaics, terrain models, NDVI analysis and 3D models for Agriculture, Construction and Mining industries. Make real time measurements including distance, area and volume. View NDVI helps in detecting crop stress and variability
18.	Genscape www.genscape.com	Commodity, agriculture, freight	Genscape has added satellite reconnaissance, Artificial Intelligence, and maritime freight tracking to its data acquisition capabilities. Genscape Maritime offers exclusive Vessel Coverage of U.S. Inland Waterways and covers 90% of US inland waterways. Genscape measures market fundamentals using thousands of patented and proprietary land, sea, and satellite monitors strategically deployed worldwide, delivering exceptional insight and intelligence to clients. Genscape uses non-contact ground sensors to monitor power lines; such data is used by power grid owners and operators
19.	Brickstream www.brickstream.com	Equity	Brickstream (formerly Nomi) uses 3D Stereo video using bespoke cameras for counting visitors at stores. The Brickstream line of sensors provides highly accurate, anonymous information about how people move into, around, and out of physical places. These smart devices are installed overhead inside retail stores, malls, banks, stadiums, transportation terminals and other brick-and-mortar locations to measure people's behaviors within the space
20.	Irisys www.irisys.net	Equity	InfraRed Integrated Systems Ltd, or Irisys is a supplier of People Counting sensors. Their sensors are designed to detect the heat emitted by people passing underneath it as infrared radiation. If privacy is a concern, either for a specific application or because of local regulations, our thermal-only detectors provide an ideal way of anonymously counting people.
21.	RetailNext www.retailnext.net	Equity	Robust and scalable, the patented RetailNext platform uses video analytics, Wi-Fi detection of mobile devices (including Bluetooth), data from point-of-sale systems, and other sources to inform retailers about how customers engage with their stores. This comprehensive and highly versatile platform easily integrates with a number of store data sources, including promotional calendars, staffing systems, and even weather services to analyze how numerous factors impact shopping behavior.
22.	ShopperTrak www.shoppertrak.com	Equity	ShopperTrak helps retailers, malls and entertainment venues around the globe understand when shoppers are coming in their doors, where they're going, and how to make the most of that information.
23.	Percolata www.percolata.com	Equity	Percolata offers an all-in-one hardware and software solution that helps retailers predict in-store customer traffic using video, audio, and mobile fingerprinting and then staff employees accordingly.
24.	Agribotix www.agribotix.com	Agriculture	Agribotix processes and analyzes agricultural data gathered by drones. All Agribotix solutions include FarmLens™, the cloud-based data analysis and reporting solution for people using drones in agriculture.
25.	Flexport www.flexport.com	Commodity	Flexport, as the freight forwarder for the internet age, has data on air, ocean, truck, and rail carriers across 100 countries. They have data for a large number of commodities.
26.	AggData www.aggdata.com	Equity	AggData provides a complete list of locations for companies, grouped by Standard Industry Code (SIC).
27.	Foursquare Pinpoint www.pinpoint.foursquare.com	Equity	Identify, reach and measure audiences based on where they go in the world. Pinpoint by Foursquare, allows brands to understand and connect to targeted audiences as well as measure foot traffic and advertising success.
28.	Here www.here.com	Equity	The HERE Open Location Platform is a multi-sided platform that unlocks the power of location through its ability to ingest, process and analyze location data. Here transforms information from devices, vehicles, infrastructure and

			other sources into real-time location services that play a key role in how we move, live and interact with one another.
29.	PlaceIQ www.placeiq.com	Equity	PlaceIQ is a leading data and technology company that helps businesses leverage location-based insights to connect with and understand audiences.
30.	Mobiquity Networks www.mobiquitynetworks.com	Equity	Mobiquity Networks provides precise, unique, at-scale mobile location data and insights on consumer's real-world behavior and trends for use in marketing and research.

D. Data Aggregators

S. No.	Company Name	Asset Class	Notes
1.	GuidePoint (including Quanton Data) www.guidepoint.com	All Asset Classes	GuidePoint offers clients dedicated teams of researchers and project managers who understand your urgency and have the proven experience to find the most appropriate Advisors for even the most obscure questions. Guidepoint makes several different avenues to expert research available to our clients: phone consultations; surveys and quick polls; events; and Guidepoint Tracker; and a suite of data products. GuidePoint deliver boots-on-the-ground, practical insights from sources that live in the space you're investigating, enabling you to make critical decisions with informed confidence.
2.	Quandl www.quandl.com	All Asset Classes	Quandl delivers financial, economic and alternative data to over 150,000 people worldwide. Quandl offers essential financial and economic data alongside a suite of unique, alpha-generating alternative datasets. With our unrivaled consumption experience, we have cemented a reputation for understanding and delivering what professional quantitative analysts need and want. Quandl's customers include the world's top hedge funds, asset managers and investment banks.
3.	RavenPack www.ravenpack.com	All Asset Classes	RavenPack Analytics transforms unstructured big data sets, such as traditional news and social media, into structured granular data and indicators to help financial services firms improve their performance. Clients with an intraday horizon value RavenPack's ability to detect relevant, novel and unexpected events - be they corporate, macroeconomic or geopolitical - so they can enter new positions, or protect existing ones. The product serves to overcome the challenges posed by the characteristics of Big Data - volume, variety, veracity and velocity - by converting unstructured content into a format that can be more effectively analyzed, manipulated and deployed in financial applications.
4.	Eagle Alpha www.eaglealpha.com	All Asset Classes	Eagle Alpha enables asset managers to obtain alpha from alternative data. Eagle Alpha provides analytical Tools which enable clients to do proprietary analyses and Data sources includes a database of all the best alternative datasets worldwide, advisory services to link datasets to research questions. Each dataset is tagged by S&P sector, MSCI region and Eagle Alpha's 20 categories of alternative data e.g. consumer transactions, IOT and trade data.
5.	Dun and Bradstreet (formerly Aventon) www.dnb.com	All Asset Classes	DNB's ability to identify and connect business relationships is powered by 30,000 global data sources updated 5 million times per day, including: credit scores and rating, Over 1B trade experiences, 3M corporate family trees, Banking data, Firmographics, Social Signals, Business registrations. Not only do they aggregate and compile the most accurate and comprehensive repository of business-class data on the planet, they also employ a system of techniques and process while leveraging unmatched expertise to turn the data into insights.
6.	Discern www.discern.com	Energy, Equity, Real Estate	Discern delivers insights for better investment decision-making. They provide up-to-date data on companies, retail stores, restaurants, oil wells and real estate.
7.	Bloomberg www.bloomberg.com	All Asset Classes	Consolidated access to all asset classes, aggregated from 330+ exchanges and 5,000+ contributors, as well as Bloomberg composite tickers and market indices. Provides key reference data to describe instruments, comprehensive data for buy and sell orders or quotes across multiple markets and access to normalized, end-of-day reference prices for all exchange-traded securities and some exchange indices.
8.	Factset www.factset.com	All Asset Classes	Factset provides access to hundreds of Global Databases to manipulate, audit and export data. It gives data on benchmarks, real-time exchange data, economics, market aggregates, ETF analytics and Risk Metrics and Models.

9.	1010 Data www.1010data.com	All Asset Classes	1010 data provides access to a variety of insights like retailer, brand and product market share across brick and mortar, mobile and e-commerce sites and quickly access and analyze all the data that's relevant to one's investment decisions, unlocking insights that help drive alpha and give a competitive edge.
10.	Thomson Reuters www.thomsonreuters.com	All Asset Classes	Thomson Reuters Market Psych Indices analyze news and social media in real-time and indicators are updated every minute for individual global equities and indices, countries, commodities, and currencies. They convert the volume and variety of professional news and social media into manageable information flows that drive sharper decisions. They analyze 40,000 primary global news sources, 7000 blogs, stock message boards and social media sites and have historical data from 1998.
11.	Wharton Research Data Services (WRDS), https://wrds-web.wharton.upenn.edu/wrds	All Asset Classes	WRDS provides the user with one location to access over 250 terabytes of data across multiple disciplines including Accounting, Banking, Economics, ESG, Finance, Healthcare, Insurance, Marketing, and Statistics.
12.	MorningStar www.morningstar.com	All Asset Classes	Morningstar unites global data and institutional research with analytics on its platform to provide clients advanced portfolio optimization and reporting capabilities. Their management system streamlines investment research, portfolio accounting and performance reporting.
13.	Moody's Analytics www.moodyanalytics.com	All Asset Classes	Moody's analytics is a risk management company to help clients generate and evaluate investment ideas, monitor and manage risk/return profile of the portfolio, complete compliance reporting requirements, assist with underwriting, reinsurance and treasury services.
14.	S&P Global www.spglobal.com	All Asset Classes	S&P Global is a provider of ratings, benchmarks, analytics and data to the capital and commodity markets around the world.
15.	Airex Market www.airexmarket.com	All Asset Classes	Airex Market provides access to premium investment research, analysis, data and apps to investment managers, analysts, traders and self-directed investors. Airex is committed to democratizing the world's financial apps, information, and reports, enabling suppliers and customers to cost-effectively conduct business.
16.	Estimize www.estimize.com	Equity, Macro Economic	An earnings data set with exclusive insight on over 2000 stocks. Estimize crowdsources earnings and economic estimates from 44,022 hedge fund, brokerage, independent and amateur analysts.
17.	Quant Connect www.quantconnect.com	Equity, FX, Fixed Income, Commodity	Quant Connect provides 1) US Equities tick data going back to January 1998 for every symbol traded, totaling over 29,000 stocks. 2) Morning Star Fundamental data for the most popular 8,000 symbols for 900+ indicators since 1998. 3) Forex (FXCM, ONADA brokerage) since Apr 2007. 4) Futures tick trade and quote data from January 2009 onwards for every contract traded in CME, COMEX and GLOBEX. 5) Option trades and quotes down to minute resolution, for every option traded on ORPA since 2007.
18.	Tick Data www.tickdata.com	Equity, FX, Fixed Income, Commodity	Tick Data provides historical intraday stock, futures, options and forex data for back-testing trading strategies, develop risk & execution models, perform post-trade analysis, and conduct important academic research with data as far back as 1974.

E. Technology Solutions

1) Data Storage (Databases)

S. No.	Company Name	Notes
1.	Amazon Dynamo DB www.aws.amazon.com/dynamodb	Amazon DynamoDB is a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale. It is a fully managed cloud database and supports both document and key-value store models.
2.	MongoDB www.mongodb.com	MongoDB is free and open-source distributed database. It stores data in flexible, JSON-like documents, meaning fields can vary from document to document and data structure can be changed over time. The document model maps to the objects in your application code, making data easy to work with. Ad hoc queries, indexing, and real time aggregation provide powerful ways to access and analyze your data.
3.	Aerospike www.aerospike.com	Aerospike is an enterprise-class, NoSQL database solution for real-time operational applications, delivering predictable performance at scale, superior uptime, and high availability at the lowest TCO compared to first-generation NoSQL and relational databases.
4.	Sequoia DB www.sequoiadb.com	SequoiaDB, a document-oriented NewSQL database that supports JSON transaction processing and SQL query. It can either be a standalone product to interface with applications providing high performance and horizontally scalable data storage and processing functions, or serve as the frontend of Hadoop and Spark for both real-time query and data analysis. It is designed to integrate well with Spark, Hadoop/Cloudera.
5.	Redis Labs www.redislabs.com	Redis Labs is the home of open source Redis and commercial provider of Redis Enterprise – Redis, the world's most popular in-memory database platform for transactional, analytics and hybrid deployments. The high performance, true high availability and seamless scaling of Redis enhance popular Redis use cases including high speed transactions, job & queue management, user session stores, real time data ingest, notifications, content caching and time-series data.
6.	Oracle www.oracle.com/index.html	Oracle offers a comprehensive and fully integrated stack of cloud applications and platform services. Oracle NoSQL Database provides a powerful and flexible transaction model that greatly simplifies the process of developing a NoSQL-based application. It scales horizontally with high availability and transparent load balancing even when dynamically adding new capacity.
7.	MarkLogic www.marklogic.com	MarkLogic is an operational and transactional enterprise NoSQL database that integrates data better, faster, with less cost. Key features include scalability, time base audit trails, tiered storage, geospatial tagging, real time alerting and semantics for data relationships.
8.	Datastax www.datastax.com	DataStax Enterprise (DSE) accelerates your ability to deliver real-time value at a large scale by providing a comprehensive and operationally simple data management layer with a unique always-on architecture built on Apache Cassandra™. DataStax Enterprise provides the distributed, responsive and intelligent foundation to build and run cloud applications.
9.	Couchbase www.couchbase.com	Couchbase Server is a NoSQL database, which has become the de facto standard for building Systems of Engagement. It is designed with a distributed architecture for performance, scalability, and availability. It enables developers to build applications easier and faster by leveraging the power of SQL with the flexibility of JSON. Couchbase Server can be deployed as a document database, key-value store or distributed cache.
10.	SAP Hana www.sap.com/product/technology-platform/hana/features.html	Deployable on premise or in the cloud, SAP HANA is an in-memory data platform that lets you accelerate business processes, deliver more business intelligence, and simplify your IT environment. By providing the foundation for all your data needs, SAP HANA removes the burden of maintaining separate legacy systems and siloed data, so you can run live and make better business decisions in the new digital economy.
11.	Clustrix www.clustrix.com	ClustrixDB is the next generation MySQL-compatible, scale-out relational database for high-value, high-transaction workloads.

S. No.	Company Name	Notes
12.	Paradigm4 www.paradigm4.com	Paradigm4 is the creator of SciDB, a computational DBMS used to solve large-scale, complex analytics challenges on Big and Diverse Data. SciDB is a new computational database for mining insights from data sources as diverse as genomics, clinical, RWE, image, financial markets, instrument, and wearable devices.
13.	Memsql www.memsql.com	MemSQL is a scalable real-time data warehouse that ingests data continuously to perform analytics for the front lines of business. It can ingest and transform millions of events of data per day while simultaneously analyzing billions of rows of data using standard SQL. It queries traditional SQL, JSON and Geospatial data types in real time.
14.	NuoDB www.nuodb.com	NuoDB is a memory-centric ACID-compliant, SQL database that can be dynamically tuned to customer requirements while easily scaling in and out on commodity hardware with automated disaster recovery. In short, a technologically advanced, elastic SQL database for cloud- and container-based environments.
15.	Cockroach labs www.cockroachlabs.com	Cockroach Labs is the company building CockroachDB, an open source database for building global, scalable cloud services that survive disasters. It combines the rich functionality of SQL with the horizontal scalability common to NoSQL offerings, and provides enterprise-grade disaster recovery.
16.	DEEP DB www.npmjs.com/package/deep-db	DEEP Framework is a full-stack web framework for building cloud-native web applications. Using DEEP Framework, developers get streamlined "production-like" development environment, enterprise-level platform using microservices architecture, virtually infinite scalability with zero devops (aka serverless computing) and abstracted use of web services from cloud providers (e.g. AWS, GCP, etc.).
17.	Maria DB www.mariadb.org	MariaDB is open source and as a relational database it provides an SQL interface for data access. MariaDB turns data into structured information in a wide array of applications. It is an enhanced, drop-in replacement for MySQL. MariaDB is used because it is fast, scalable and robust, with a rich ecosystem of storage engines, plugins and many other tools make it very versatile for a wide variety of use cases.
18.	VoltDB www.voltDB.com	VoltDB is an enterprise-ready, NewSQL offering. It combines real-time analytics on flows of data, strong ACID guarantees, familiar SQL data interactivity, and native clustering. VoltDB is used for applications that require fast decisions on live streams of data with the speed of NoSQL and strong consistency.
19.	Splice Machine www.splicemachine.com	Splice Machine is the open-source dual-engine RDBMS for mixed operational and analytical workloads, powered by Hadoop and Spark. Splice Machine makes it easy to create modern, real-time, and scaleable applications or to offload operational and analytical workloads from expensive systems. The Splice Machine RDBMS has all of the key functionality of industrial SQL databases but on a scale-out architecture typically found in less functional NoSQL systems.
20.	Citus Data www.citusdata.com	Citus is a distributed database that extends PostgreSQL, allowing you to continue using all the powerful Postgres features while still scaling. Citus handles capturing billions of events per day, while providing real-time analytics across your entire dataset. It allows ingesting all the data, as it happens in real-time, scan, filter, and analyze billions of events in under a second and power application with single-digit millisecond latency operations at scale.
21.	Neo4j www.neo4j.com	Neo4j is a highly scalable, native graph database purpose-built to leverage not only data but also its relationships. Neo4j's native graph storage and processing engine deliver constant, real-time performance, helping enterprises build intelligent applications to meet today's evolving data challenges.
22.	OrientDB www.orientdb.com	OrientDB is a Distributed Multi-Model NoSQL Database with a Graph Database Engine. It provides the power of a Distributed Graph Database engine with the flexibility of a Document Database. OrientDB incorporates a search engine, Key-Value and Object-Oriented concepts along with a reactive model (with Live Queries) and geospatial awareness.

S. No.	Company Name	Notes
23.	TERADATA www.teradata.com	Designed to deliver high-performance, diverse queries, in-database analytics, and sophisticated workload management, the Teradata Database supports, enables all Teradata Data Warehouse solutions.

2) Data Transformation (Extract-Transform-Load)

S. No.	Company Name	Notes
1.	Alteryx www.alteryx.com	Alteryx is a self-service data analytics with a platform that can prep, blend, and analyze all of your data, then deploy and share analytics at scale for deeper insights in hours, not the weeks that you may be used to.
2.	Talend www.talend.com	Talend is open source integration software provider to data-driven enterprises. Talend's customers connect anywhere, at any speed. From ground to cloud and batch to streaming, data or application integration, Talend connects at big data scale, 5x faster and at 1/5th the cost.
3.	Pentaho www.pentaho.com	Pentaho's big data integration and analytics solutions turn information into insights to help your organization gain a competitive advantage. Pentaho Data Integration prepares and blends data to create a complete picture of the business that drives actionable insights. The platform delivers accurate, analytics-ready data to end users from any source. With visual tools to eliminate coding and complexity, Pentaho puts big data and all data sources at the fingertips of business and IT users.
4.	Trifacta www.trifacta.com	Trifacta's mission is to create radical productivity for people who analyze data. Data Wrangling by Trifacta allows you to discover, wrangle, and visualize complex data quickly.
5.	Tamr www.tamr.com	Tamr's patented software fuses the power of machine learning with knowledge of data to automate the rapid unification of data silos.
6.	Paxata www.paxata.com	Paxata provides an Adaptive Information Platform that enables business analysts with an enterprise-grade, self-service data preparation system to support the on-demand and ad-hoc business data needs for analytics, operations and regulatory requirements.
7.	StreamSets www.streamsets.com	StreamSets software delivers performance management for dataflows that feed the next generation of big data applications. It brings operational excellence to the management of data in motion, so that data continually arrives on-time and with quality, empowering business-critical analysis and decision-making.
8.	Alation www.alation.com	Alation's enterprise collaborative data platform empowers employees inside of data-driven enterprises to find, understand, and use the right data for better, faster business decisions. Alation combines the power of machine learning with human insight to automatically capture information about what the data describes, where the data comes from, who's using it and how it's used.
9.	Informatica www.informatica.com	Informatica can integrate, secure, and govern next-generation big data with repeatable, reliable, and maintainable processes to add value to your organization. Informatica Cloud can manage and integrate your data across Salesforce, Microsoft Azure, Amazon Web Services (AWS), Amazon Redshift, Workday, and hundreds of other data sources, regardless of their location.
10.	MuleSoft www.mulesoft.com	MuleSoft provides the most widely used integration platform (Mule ESB & CloudHub) for connecting SaaS & enterprise applications in the cloud and on-premise. Delivered as a unified integration experience, CloudHub™ and Mule ESB™ (Enterprise Service Bus) are built on proven open source technology for fast and reliable on-premise and cloud integration without vendor lock-in.
11.	SnapLogic www.snaplogic.com	SnapLogic is a unified data and application integration platform as a service (iPaaS). Their hybrid cloud architecture is powered by 300+ Snaps, which are pre-built integration components that simplify and automate complex enterprise integration patterns and processes.
12.	Bedrock Data www.bedrockdata.com	Bedrock Data is the software who connects, cleans and synchronizes all your cloud systems in real-time. It integrates SaaS systems.
13.	Xplenty www.xplenty.com	Xplenty's platform allows organizations to integrate, process, and prepare data for analytics on the cloud. By providing a coding and jargon-free environment, Xplenty's scalable platform ensures businesses can quickly and easily benefit from the opportunities offered by big data without having to invest in hardware, software, or related personnel. With Xplenty, every company can have immediate connectivity to a variety of data stores and a rich set of out-of-the-box data transformation components.

S. No.	Company Name	Notes
14.	Tealium www.tealium.com	Tealium customer data platform, comprised of an enterprise tag management solution, omnichannel customer segmentation and action engine, and suite of rich data services, creates a vendor-neutral data foundation that spans web, mobile, offline and IoT.
15.	Enigma www.enigma.com	The Enigma ecosystem of Data Infrastructure is built to organize the world's public data and designed for flexibility, Enigma's infrastructure connects disparate datasets across organizational silos. Their Concourse and Assembly platforms enable you to design resilient data pipelines, improve metadata, and search and build on top of data assets.
16.	Podium Data www.podiumdata.com	Podium is a turnkey, end-to-end big data management platform that revolutionizes how enterprises manage, prepare, and deliver data. It transforms your traditional data supply chain into a self-service, on-demand marketplace.
17.	Zaloni www.zaloni.com	Zaloni provides enterprise data lake management, governance and self-service data solutions to ensure a clean, actionable data lake. Their Mica and Bedrock platform enables customers to gain their own competitive advantage through organized, actionable data lakes.
18.	StitchData www.stitchdata.com	Stitch is a simple, powerful ETL built for developers with all the flexibility and control software developers need to provision data, without the maintenance headaches. Their clean, flexible UI allows developers to configure their data pipeline in a way that balances data freshness with cost and production database load.
19.	Alooma www.alooma.com	Alooma's powerful data integration platform performs any integration on any data warehouse on real-time ETL at scale. Alooma enables data teams to have visibility and control. It brings data from your various data silos together to your data warehouse, all in real-time.
20.	Segment www.segment.com	Segment is an analytics API and Customer Data Platform. It collects customer data from client's web, mobile, server and cloud apps, integrates the data and loads it to the client data warehouse.
21.	Vertica www.vertica.com	The HPE Vertica Analytics Platform is purpose built from the very first line of code for Big Data analytics. It is designed for use in data warehouses and other big data workloads where speed, scalability, simplicity, and openness are crucial to the success of analytics. Vertica relies on a tested, reliable distributed architecture and columnar compression to deliver fast speed.
22.	Kognitio www.kognitio.com	Kognitio provides massive query acceleration and high concurrency by pulling the data of interest into high speed memory so that highly efficient massively parallel processing algorithms can be applied to each and every operation. The Kognitio Platform has very rich and mature SQL support as well as an extremely flexible and powerful NOSQL capability. It can be implemented on both Hadoop and existing data warehouses.
23.	Exasol www.exasol.com	EXASOL helps companies to run their businesses smarter and drive profit by analyzing data and information at unprecedented speeds. They have developed database for analytics and data warehousing, and offer first-class know-how and expertise in data insight and analytics.

3) Hadoop and Spark framework

S. No.	Company Name	Notes
1.	Cloudera www.cloudera.com	Cloudera is a Big Data company with analytics abilities. It allows companies to huge amounts of data on a low cost hardware. They provide solutions in data processing, machine learning, stream processing and exploratory data science.
2.	Hortonworks www.hortonworks.com	The Hortonworks data management platform gives solutions for big data analysis in the open-source architecture for all types of data.
3.	Pivotal www.pivotal.io	Pivotal provides development services on an open source platform. Services include mobile banking, trading and transactions, compliance and process automation, personalized sales and marketing & Risk, fraud and security management.
4.	MapR www.mapr.com	The MapR Converged Data Platform integrates Hadoop, Spark, and Apache Drill with real-time database capabilities, global event streaming, and scalable enterprise storage to power big data applications. The MapR Platform delivers security, reliability, and real-time performance, while lowering both hardware and operational cost.
5.	IBM InfoSphere www.ibm.com/software/in/data/infosphere	The InfoSphere Platform provides data integration, data warehousing, master data management, big data and information governance. The platform provides foundation for information-intensive projects, providing the performance, scalability, reliability and acceleration needed to simplify difficult challenges.
6.	BlueData www.bluedata.com	BlueData deploys Big Data by making it easier, faster, and more cost-effective for organizations of all sizes to use Hadoop and Spark infrastructure on-premises (or in the public cloud). You can 1) Spin up Hadoop or Spark clusters for test or production environments. 2) Deliver the agility and efficiency benefits of virtualization, with the performance of bare-metal. 3) Work with any Big Data analytical application, any Hadoop or Spark distribution, and any infrastructure. 4) Provide the enterprise-grade governance and security required, in a multi-tenant environment.
7.	Jethro www.jethro.io	Jethro makes Real-Time Business Intelligence work on Hadoop, integrates with Tableau, Qlik & SaaS analytics dashboards. Jethro transparently supports all SQL queries, for thousands of concurrent users analyzing tens of billions of rows. All that with interactive response times measured in seconds.
8.	INRIX www.inrix.com	INRIX is a global SaaS and DaaS company which provides a variety of Internet services and mobile applications pertaining to road traffic and driver services.
9.	Microsoft Azure www.azure.microsoft.com	Microsoft Azure is an open, flexible, enterprise-grade cloud computing platform. It is a growing collection of integrated cloud services that developers and IT professionals use to build, deploy, and manage applications through our global network of datacenters. With Azure, you can build and deploy wherever you want, using the tools, applications, and frameworks of your choice.
10.	Amazon Web Services www.aws.amazon.com	Amazon Web Services offers reliable, scalable, and inexpensive cloud computing services. Free to join, pay only for what you use. It is a secure cloud services platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow.
11.	Google Cloud Platform www.cloud.google.com	Google Cloud Platform lets you focus on what's next for your business. Google Cloud Platform frees you from the overhead of managing infrastructure, provisioning servers and configuring networks. They provide Future proof infrastructure, powerful data and analytics. Also, it lets you build and host applications and websites, store data, and analyze data on Google's scalable infrastructure.
12.	CAZENA www.cazena.com	Cazena handles a wide range of enterprise data and analytics. Its services include 1) Data Science Sandbox to run a wide range of analytics in the cloud, without assembling or maintaining the underlying technology. 2) Cloudera Data Lake to collect, stage and pre-process raw data like log files or streaming data or cost-efficiently archive historical data. 3) Data Mart to improve access, share data or augment data warehouses.

S. No.	Company Name	Notes
13.	IBM BigInsights www.ibm.com/us-en/marketplace/biginsights	IBM BigInsights combines 100% open source Hadoop and Spark for the open enterprise to cost-effectively analyze and manage big data. You can get more insights by combining a variety of data sources in creative and innovative ways using Hadoop. IBM provides a complete solution, including Spark, SQL, Text Analytics and more to scale analytics quickly and easily.
14.	Treasure Data www.treasuredata.com	Treasure Data makes over 100+ integrations to unify your data using data connectors, machine learning and more. It connects, unifies, analyzes, scales and automates data infrastructure.
15.	Altiscale www.altiscale.com	Altiscale provides Big Data as a Service using Hadoop, Spark, and other data science tools in the cloud. Use cases for Financial Services include: 1) Aggregating structured and unstructured data from multiple customer interactions to obtain a holistic view of customers. 2) Analyzing a combination of historical and real-time market data to make investment decisions. 3) Accurately modeling the credit and investment risk the firm is exposed to at any point in time. 4) Using advanced analytics to set insurance rates, particularly from usage-based insurance programs. 5) Identifying instances of fraud, such as unauthorized transactions, false insurance claims, and money laundering.
16.	Qubole www.qubole.com	Qubole Data Service (QDS) turns big data analytics into actionable insights across multiple big data use cases like Sentiment Analysis, 360-Degree Customer View, Ad-hoc Analysis, Real-Time Analytics, Multi-Channel Marketing, Customer Micro-Segmentation, Ad Fraud Detection and Clickstream Analysis.
17.	Databricks www.databricks.com	Databricks provides a Virtual Analytics Platform on top of Apache Spark that empowers anyone to easily build and deploy advanced analytics solutions.
18.	CenturyLink www.centurylink.com/business/data.html	CenturyLink provides Cloud and Hosting Services. They provide tools to economically start, grow, and support applications using on-demand server resources in a virtualized environment that is easily accessible over a public network. It also helps to create secure and reliable data center colocation services, helping businesses increase efficiency, cut costs, increase growth opportunities and improve service levels.
19.	GridGain www.gridgain.com	GridGain Systems offers in-memory computing software and services solutions. As an in-memory solution, query times are 1,000 to 1,000,000 times faster than disk-based systems. You can modernize your existing data-intensive architecture by inserting GridGain between your existing application and database layers. It features a unified API which supports SQL, C++, .NET, PHP, MapReduce, JAVA/Scala/Groovy, and Node.js access for the application layer. GridGain and applications and databases can run on premise, in a hybrid environment, or on a cloud platform such as AWS, Microsoft Azure or Google Cloud Platform.
20.	Alluxio www.alluxio.org	Alluxio, formerly Tachyon, is an Open Source Memory Speed Virtual Distributed Storage. It enables any application to interact with any data from any storage system at memory speed. It is a memory-centric distributed storage system enabling reliable data sharing at memory-speed across cluster frameworks, such as Spark and MapReduce.
21.	Hadoop HDFS www.hortonworks.com/apache/hdfs	The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a wide variety of concurrent data access applications, coordinated by YARN
22.	Hadoop MapReduce www.hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html	Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.
23.	Yarn www.yarnpkg.com	Yarn is a fast, reliable, and secure dependency management system. It is a package manager for your code. It allows you to use and share code with other developers from around the world. Yarn allows you to use other developers' solutions to different problems, making it easier for you to develop your software.
24.	Spark	Apache Spark is a fast and general engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. It runs programs up to 100x faster

S. No.	Company Name	Notes
	www.spark.apache.org	than Hadoop MapReduce in memory, or 10x faster on disk. Spark runs on Hadoop, Mesos, standalone, or in the cloud.
25.	Mesos www.mesos.apache.org	Apache Mesos abstracts CPU, memory, storage, and other compute resources away from machines (physical or virtual), enabling fault-tolerant and elastic distributed systems to easily be built and run effectively.
26.	Tez www.tez.apache.org	The Apache Tez™ project is aimed at building an application framework which allows for a complex directed-acyclic-graph of tasks for processing data. It is currently built atop Apache Hadoop YARN.
27.	Flink www.flink.apache.org	Apache Flink® is an open-source stream processing framework for distributed, high-performing, always-available, and accurate data streaming applications.
28.	CDAP www.cask.co/products/cdap	CDAP is 100% open source platform that provides both data integration and app development capabilities on Apache Hadoop and Spark. The platform helps you future proof your big data investments, provides rapid time to value, and empowers your business users with a self-service user experience.
29.	Apache Kylin www.kylin.apache.org	Apache Kylin™ is an open source Distributed Analytics Engine designed to provide SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets
30.	Trafodion www.trafodion.apache.org	Apache Trafodion (incubating) is a webscale SQL-on-Hadoop solution enabling transactional or operational workloads on Apache Hadoop. Trafodion builds on the scalability, elasticity, and flexibility of Hadoop. Trafodion extends Hadoop to provide transactional integrity, enabling new kinds of big data applications to run on Hadoop.

4) Data Analysis Infrastructure (Cluster and Clouds)

S. No.	Company Name	Notes
1.	Amazon Web Services www.aws.amazon.com	Amazon Web Services offers reliable, scalable, and inexpensive cloud computing services. Free to join, pay only for what you use. It is a secure cloud services platform, offering computing power, database storage, content delivery and other functionality to help businesses scale and grow.
2.	Docker www.docker.com	Docker is an open platform for developers and sysadmins to build, ship, and run distributed applications, whether on laptops, data center VMs, or the cloud.
3.	Pepperdata www.pepperdata.com	Pepperdata is the big data performance company that provides solutions to complex performance problems, helps to run more jobs on your cluster, and guarantee on-time executions of critical jobs.
4.	CoreOS www.coreos.com	CoreOS is a lightweight Linux operating system designed for clustered deployments providing automation, security, and scalability for your most critical applications.
5.	StackIQ www.stackiq.com	StackIQ gives you the ability to deploy and manage large-scale Big Data infrastructure quickly and efficiently. It helps customers build, run, and manage large distributed systems and private cloud infrastructure with a complete automation platform.
6.	Mesosphere www.mesosphere.com	Mesosphere is democratizing the modern infrastructure we use at Twitter, AirBnB, and other webscale companies to quickly deliver data-driven services on any datacenter or cloud. Apache Mesos is the open-source distributed systems kernel at the heart of the Mesosphere DC/OS. It abstracts the entire datacenter into a single pool of computing resources, simplifying running distributed systems at scale.
7.	HPCC Systems www.hpccsystems.com	HPCC Systems helps process and analyze big data more efficiently and reliably than with any other comparable solution available. The platform delivers superior performance, agility, and scalability.
8.	Kubernetes www.kubernetes.io	Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications
9.	Microsoft Azure www.azure.microsoft.com	Microsoft Azure is an open, flexible, enterprise-grade cloud computing platform. It is a growing collection of integrated cloud services that developers and IT professionals use to build, deploy, and manage applications through our global network of datacenters. With Azure, you can build and deploy wherever you want, using the tools, applications, and frameworks of your choice.
10.	Pivotal www.pivotal.io	Pivotal provides development services on an open source platform. Services include mobile banking, trading and transactions; compliance and process automation; personalized sales and marketing; and risk, fraud and security management.
11.	Google Cloud Platform www.cloud.google.com	Google Cloud Platform lets you focus on what's next for your business. Google Cloud Platform frees you from the overhead of managing infrastructure, provisioning servers and configuring networks. They provide Future proof infrastructure, powerful data and analytics. Also, it lets you build and host applications and websites, store data, and analyze data on Google's scalable infrastructure.
12.	Snowflake www.snowflake.net	Snowflake provides a data warehouse built for the cloud, delivering a solution capable of solving problems for which legacy, on-premises and cloud data platforms were not designed.
13.	Waterline Data www.waterlinedata.com	Waterline is an enterprise data catalog company. They help data professionals discover, organize and curate data and also expose the newly organized data for business professionals to use.
14.	Info Works www.infoworks.io	InfoWorks provides a complete functionality in a single platform from data ingestion, data synchronization and the building of data models and cubes. It scales your data warehousing and analytics on Hadoop by using advanced machine intelligence.
15.	Panasas	Panasas is the premier provider of hybrid scale-out NAS storage for technical computing and

S. No.	Company Name	Notes
	www.panasas.com	enterprise workloads. All Panasas ActiveStor storage appliances leverage the Panasas PanFS parallel file system for superior performance, data availability and reliability, linear scalability, and easy management.
16.	Nimble Storage www.nimblestorage.com	Nimble storage provides All Flash and Hybrid Flash solutions with its predictive flash platform to give quick and reliable access to data.
17.	COHO Data www.cohodata.com	Coho Data offers an integrated Software Defined Infrastructure (SDI) solution that brings public cloud simplicity, true scale-out and low cost economics into the data center. The Coho DataStream platform simplifies data center infrastructure operations and management, by abstracting storage, network and compute, thereby allowing for the right mix of resources for varied application demands.
18.	Qumulo www.qumulo.com	Qumulo Core is the leader in data-aware scale-out NAS. Its real-time analytics help storage administrators obtain instant answers about their data footprint by explaining usage patterns and which users or workloads are impacting their performance and capacity.
19.	Veeva www.veeva.com	Veeva Systems Inc. is cloud-based software for the global life sciences industry committed to innovation, product excellence, and customer success.
20.	Tencent Cloud www.qcloud.com	Tencent Cloud is a secure, reliable and high-performance cloud compute service provided by Tencent.
21.	Alibaba Cloud www.intl.aliyun.com	As the cloud computing arm and business unit of Alibaba Group (NYSE: BABA), Alibaba Cloud provides a comprehensive suite of global cloud computing services to power both our international customers' online businesses and Alibaba Group's own e-commerce ecosystem. Alibaba Cloud's international operations are registered and headquartered in Singapore, and the company has international teams stationed in Dubai, Frankfurt, Hong Kong, London, New York, Paris, San Mateo, Seoul, Singapore, Sydney and Tokyo.
22.	Huawei Cloud www.hwclouds.com	Since the implementation of Huawei's "cloud- pipe-device" strategy, Huawei Enterprise Cloud have focused on the IaaS layer, enable the PaaS layer, and aggregate the SaaS layer. Huawei Enterprise Cloud services are committed to providing secure and reliable cloud services for enterprises, governments, and innovative/startup groups. Using ICT services is as easy as 1-2-3.
23.	Inspur www.en.inspur.com	Inspur is one of the largest global data center and cloud computing solutions provider in the world. Our technological capabilities enable us to deliver full Cloud service stack from IaaS and PaaS to SaaS. We are widely recognized as the leaders in computing and server hardware design and production, and work with more than 1,000 enterprises around the world in end to end HPC, cloud computing, and private and public cloud management.

5) Data Management and Security

S. No.	Company Name	Notes
1.	New Relic www.newrelic.com	New Relic provides deep performance analytics for every part of software environment. One can easily view and analyze massive amounts of data, and gain actionable insights in real-time for apps, users and business.
2.	AppDynamics www.appdynamics.com	AppDynamics delivers real-time access to every aspect of your business and operational performance, so you can anticipate problems, resolve them automatically, and make smarter, more certain business decisions. AppDynamics now a part of Cisco
3.	Amazon Web Services www.aws.amazon.com	Amazon Web Services offers reliable, scalable, and inexpensive cloud computing services. Free to join, pay only for what you use. It is a secure cloud services platform, offering computing power, database storage, content delivery and other functionality to help businesses scale and grow.
4.	Actifio www.actifio.com	Actifio's data virtualization software lets enterprises deliver data-as-a-service, instantly, anywhere across a hybrid cloud environment. It results in faster testing for DevOps and better data protection SLAs, all while reducing cost and complexity.
5.	Numerify www.numerify.com	Numerify is a cloud-based analytics solution that delivers intelligence for IT service, operations & development.
6.	Splunk www.splunk.com	Splunk Inc. provides the leading platform for Operational Intelligence. Customers use Splunk to search, monitor, analyze and visualize machine data.
7.	Anodot www.anodot.com	Anodot is a real time analytics & automated anomaly detection solution that discovers & turns outliers in time series data into valuable business insights.
8.	Datadog www.datadoghq.com	The Datadog team is on a mission to bring sanity to IT Management. With turn-key integrations, Datadog seamlessly aggregates metrics and events across the full devops stack.
9.	Rocana www.rocana.com	Rocana's mission is to help technology leaders be the catalyst for digital transformation within their companies. Rocana's objective is to arm technologists with total visibility and the power to outperform the competition. Their innovative operations software eliminates the boundaries of legacy IT monitoring tools, challenging all assumptions about the scale and economics of total visibility.
10.	Tanium www.tanium.com	Tanium invented a fundamentally new approach to endpoint security and systems management, where security and IT operations teams can achieve 15-second Visibility and Control.
11.	iSocialSmart www.isocialsmart.com	Social Smart Software, LLC (Social Smart) is a leading vendor of enterprise social media marketing and compliance solutions for highly regulated industries. Mission is to provide quality social networking software products that allow everyone to confidently leverage social networks in a riskfree compliant manner.
12.	Cylance www.cylance.com	Cylance® is revolutionizing cybersecurity with products and services that proactively prevent, rather than reactively detect the execution of advanced persistent threats and malware. Protect your endpoints against advanced malware with the antivirus built on Artificial Intelligence and machine learning.
13.	StackPath www.stackpath.com	StackPath is the intelligent web services platform for security, speed and scale. It unifies enterprise security solutions by leveraging collaborative intelligence that makes each service smarter and more secure with every threat detected, in addition to vastly improving the customer experience.
14.	Darktrace www.darktrace.com	Darktrace is an Enterprise Immune System technology company for cyber security. Darktrace is the cyber defense technology that is capable of detecting anomalous behaviors, without any prior

S. No.	Company Name	Notes
		knowledge of what it is looking for.
15.	ThreatMetrix www.threatmetrix.com	ThreatMetrix, The Digital Identity Company, helps businesses prevent online fraud, while enhancing the customer experience and boosting revenue.
16.	Guardian Analytics www.guardiananalytics.com	Guardian Analytics provide fraud prevention solution using behavior-based anomaly detection. They automatically and immediately protect 100 percent of users and account holders against all types of threats and schemes, and dynamically adapts to new attacks and changes in legitimate user behavior.
17.	Anomali www.anomali.com	Anomali is a Threat Intelligence Platform that enables businesses to integrate security products and leverage threat data to defend against cyber threats.
18.	Sift Science www.siftscience.com	Sift Science is powered by data, technology, and people. Apps worldwide send data in real-time to Sift Science using their modern REST APIs & JavaScript snippet. Their large-scale machine learning technology analyzes data instantly connecting clues left behind by fraudsters.
19.	SecurityScorecard www.securityscorecard.com	SecurityScorecard provides the most accurate security ratings & continuous risk monitoring for vendor and third party risk management.
20.	Recorded Future www.recordedfuture.com	Recorded Future is the leading dedicated threat intelligence provider, powered by patented machine learning, and driven by world-class threat researchers.
21.	Feedzai www.feedzai.com	Feedzai prevents fraud and keep commerce safe. It uses a machine learning platform for fighting Fraud and Risk.
22.	Illumio www.illumio.com	Illumio helps you stop cyber threats, improve understanding of risk, and simplify security operations for applications in and across you data center and cloud.
23.	Code42 www.code42.com	Code42 protects more than 39,000 organizations worldwide. Code42 enables IT and security teams to centrally manage and protect critical data for some of the most recognized brands in business and education. From monitoring endpoint data movement and use, to meeting data privacy regulations, to simply and rapidly recover from data incidents.
24.	Data Gravity www.datagravity.com	DataGravity enables organizations to identify sensitive data across their virtual environments and protect it from theft, misuse, and abuse. DataGravity understands data and helps companies of all sizes secure and protect their sensitive information through actionable insights and timely intelligence about their data.
25.	Cipher Cloud www.ciphercloud.com	CipherCloud provides full-scale enterprise cloud security solutions for cloud monitoring, encryption, key management, malware detection and compliance.
26.	Vectra www.vectranetworks.com	Vectra provides real-time attack visibility and non-stop automated threat hunting powered by Artificial Intelligence. The result is blind-spot-free threat detection coverage across the entire network infrastructure and all devices, including IoT and BYOD. Vectra also lets you respond quickly and decisively to attacks by putting the most relevant threat information and context at your fingertips.
27.	Sqrrl www.sqrrl.com	Sqrrl is the threat hunting company that enables organizations to target, hunt, and disrupt advanced cyber threats. Sqrrl Enterprise enables the ingestion and analysis of disparate datasets to facilitate proactive threat detection, which is also known as cyber threat hunting.
28.	Blue Talon www.bluetalon.com	BlueTalon is a leading provider of data-centric security for next-gen data platforms. BlueTalon keeps enterprises in control of their data by allowing them to give users access only to the data they need.

S. No.	Company Name	Notes
29.	Reltio www.reltio.com	Reltio help companies turn their data into information and knowledge assets in the efficient way. Reltio manages all data types including multi-domain master data, transaction and interaction data, third party, public and social data.
30.	Clarity www.clarityinsights.com	Clarity's solutions include pricing sophistication and accuracy for property and casualty insurance, database marketing and campaign management, and customer data integration. The company also provides advisory services, such as strategy, capability maturity assessment, data analytics, master data management/data governance, and program/project management; architecture services, including solution blueprints, information architecture, data virtualization, search, mobile business intelligence, and technology selection; build services, including data stores, data integration, information delivery and advanced analytics.
31.	BrightPlanet www.brightplanet.com	BrightPlanet is a deep web data collection & software-as-a-service company that specializes in harvesting large amounts of unstructured web data and preparing it for analysis.

6) Machine Learning Tools

S. No.	Company Name	Notes
1.	Domino www.dominodatalab.com	Domino provides data science teams with best practice knowledge management and reproducibility, and rapid development and deployment of models. It is a front-end to the cloud, automating elastic compute designed for data science workloads, while letting IT control resource usage.
2.	SparkBeyond www.sparkbeyond.com	SparkBeyond has built an AI-powered research engine, capable of asking questions and discovering complex patterns in data by understanding their meaning. They combine state-of-the-art Artificial Intelligence technology with large-scale computing to accelerate breakthroughs that are hidden in the data.
3.	Rapidminer www.rapidminer.com	RapidMiner makes data science teams more productive through a unified platform for data prep, machine learning, and model deployment. Its platform accelerates the building of complete analytical workflows – from data prep to modeling to business deployment – in a single environment, dramatically improving efficiency and shortening the time to value for data science projects.
4.	DataRobot www.datarobot.com	DataRobot automates the entire modeling lifecycle, enabling users to quickly and easily build highly accurate predictive models.
5.	Yhat www.yhat.com	Yhat (pronounced Y-hat) provides an end-to-end data science platform for developing, deploying, and managing real-time decision APIs. With Yhat, data scientists can transform static insights into production-ready decision making APIs that integrate seamlessly with any customer- or employee-facing app. Yhat also created Rodeo, an open source integrated development environment (IDE) for Python.
6.	Ayasdi www.ayasdi.com	Ayasdi is an advanced analytics company that offers a machine intelligence platform and intelligent applications. Ayasdi is used to solve big data and complex data analytics challenges and to automate formerly manual processes using unique data. Ayasdi's machine intelligence platform combines scalable computing and big data infrastructure with the latest machine learning, statistical and geometric algorithms and Topological Data Analysis to enable data scientists, domain experts and business people to be exponentially more productive.
7.	Dataiku www.dataiku.com	Dataiku DSS is the collaborative data science software platform for teams of data scientists, data analysts, and engineers to explore, prototype, build, and deliver their own data products more efficiently.
8.	Seldon www.seldon.io	Seldon is the open-source machine learning pipeline for real-time recommendations and enterprise-grade predictive analytics. It includes proven algorithms, industry models and a microservices API to productionize your own. Seldon is platform-agnostic with no lock-in. You keep full ownership and control of sensitive data on-premise or in the cloud.
9.	Yseop www.yseop.com	Yseop is an Artificial Intelligence enterprise software company whose natural language generation products automate reasoning, dialog, and writing in multiple languages. It offers Yseop Compose, the self-service enterprise-level language generation technology. It also builds bespoke solutions helping businesses leverage data, automate business processes, and aid in their digital transformation.
10.	BigML www.bigml.com	BigML is a consumable, programmable, and scalable Machine Learning platform that makes it easy to solve and automate Classification, Regression, Cluster Analysis, Anomaly Detection, Association Discovery, and Topic Modeling tasks. BigML is helping thousands of analysts, software developers, and scientists around the world to solve Machine Learning tasks "end-to-end", seamlessly transforming data into actionable models that are used as remote services or, locally, embedded into applications to make predictions.
11.	CognitiveScale www.cognitivescale.com	CognitiveScale's Augmented Intelligence platform and products pair humans and machines so they can achieve something new and exponentially valuable together: Engage users intelligently at the edge and Amplify process intelligence at the core

S. No.	Company Name	Notes
		through self-learning, self-assuring business processes.
12.	GoogleML www.cloud.google.com/ml-engine	Google Cloud Machine Learning Engine is a managed service that enables you to easily build machine learning models, which work on any type of data, of any size.
13.	Context Relevant www.contextrelevant.com	The Context Relevant platform uses machine learning to enable the rapid creation of intelligent applications that deploy at enterprise scale, reacting automatically to trends and changes that occur in the underlying environment, updating dynamically, and improving over time.
14.	Cycorp www.cyc.com	The Cyc software combines common sense knowledge base with powerful inference engines and natural language interfaces to deliver human-like understanding and transparent explanations of its output and reasoning. Cyc applications can stand alone or work in concert with pattern matching AI tools, such as Machine Learning, to deliver truly differentiated value.
15.	HyperScience www.hyperscience.com	HyperScience is an Artificial Intelligence company specializing in the automation of office work. Their machine learning software takes over menial work that people are doing today and frees employees to focus on more complex tasks.
16.	Nara Logics www.naralogics.com	Nara Logics builds a synaptic network of explicit and inferred connections to create an intelligence layer on top of chaotic, siloed enterprise data for real-time, context relevant recommendations and give the reasons behind them.
17.	Clarabridge www.clarabridge.com	Utilize sentiment and text analytics to automatically collect, categorize and report on structured and unstructured data. It's more than text analytics, more than speech analytics, more than customer experience intelligence: Clarabridge enables a real-time, continuous feedback loop for constant, consistent performance across your business.
18.	H2O.ai www.h2o.ai	H2O.ai is the maker behind H2O, the open source Deep Learning platform for smarter applications and data products. H2O operationalizes data science by developing and deploying algorithms and models for R, Python and the Sparkling Water API for Spark.
19.	Scaled Inference www.scaledinference.com	Scaled Inference is the platform for intelligent computing. Their platform interfaces with both machines (software) and people (developers) to enable a new generation of goal-driven, context-adaptive, self-improving software applications.
20.	SparkCognition www.sparkcognition.com	SparkCognition is a Cognitive Security and Analytics company. They use Machine Learning & AI techniques for Cloud Security and IoT.
21.	Deepsense.io www.deepsense.io	Deepsense.io provides Deep Learning and machine learning solutions. It has two products: 1) Neptune is designed to give data science teams the freedom to explore, while remaining organized and collaborative throughout the creative process. 2) Seahorse is designed to help build Spark applications in a visual and interactive way. Users who want to create data workflows, from ETL to predictive modeling, can do it quickly and easily: no programming skills are required.
22.	Skymind www.skymind.ai	The Skymind Intelligence Layer (SKIL) is an open-source enterprise distribution. It contains all of the necessary open-source components and proprietary vendor integrations to build production-grade Deep Learning solutions. Skymind is the company behind Deeplearning4j, the commercial-grade, open-source, distributed deep-learning library written for Java and Scala. Integrated with Hadoop and Spark, DL4J is specifically designed to run in business environments on distributed GPUs and CPUs.
23.	Bonsai www.bonsai.ai	Bonsai abstracts away the complexity of machine learning libraries and algorithms, making the programming and management of AI models more accessible to developers.
24.	Agolo	Agolo uses Artificial Intelligence (machine learning and natural language processing) to create summaries from the world's information in real-time.

S. No.	Company Name	Notes
	www.agolo.com	
25.	AYLIEN www.aylien.com	AYLIEN provides a text analysis and Sentiment Analysis solutions to unlock the hidden value of your text. AYLIEN Text Analysis API is a package of Natural Language Processing, Information Retrieval and Machine Learning tools for extracting meaning and insight from textual and visual content with ease.
26.	Lexalytics www.lexalytics.com	Lexalytics process billions of unstructured documents every day, globally. Translate text into profitable decisions; make state-of-the-art cloud and on-premise text and sentiment analysis technologies that transform customers' thoughts and conversations into actionable insights. The on premise Saliency® and SaaS Semantix® platforms are implemented in a variety of industries for social media monitoring, reputation management and voice of the customer programs.
27.	Clarifai www.clarifai.com	Clarifai is an Artificial Intelligence company that uses visual recognition, solving real-world problems for businesses and developers alike. Clarifai applies machine learning to image & video recognition, helping customers understand and manage their media.
28.	Deep Vision www.deepvisionai.com	DeepVision provides agile, innovative and profitable state of the art Deep Learning based computer vision solutions applied to several industries by leveraging cutting edge technologies and trends.
29.	Cortica www.cortica.io	Cortica.io has wrapped its Retina Engine into an easy-to-use, powerful platform for fast semantic search, semantic classification and semantic filtering. With the Retina Platform, you can process any kind of text, independently of language and length.
30.	Algocian www.algocian.com	Algocian provides state-of-the-art video analytics for security and automotive applications. Algocian's detection system reduces bandwidth costs for providers, increasing profitability.
31.	Captricity www.captricity.com	Captricity's handwriting recognition technology helps organizations in insurance, government and healthcare unlock access to the customer data they need to optimize business processes, improve decision making and enhance the customer experience.
32.	Netra www.netra.io	Netra makes sense of imagery, understands how consumers are engaging with the brand or category on social media & improve targeting – using Netra's Visual Intelligence Solution. Netra discovers influential images of your brand on social in real-time.
33.	Deepomatic www.deepomatic.com	Deepomatic API integrates the power of Deep Learning and computer vision into projects. They provide ready-to-use specialized detection APIs: fashion, weapon detection, furniture detection, urban scene detection (cars, pedestrian, traffic signs), as well as a visual search engine. They also customize APIs tailored to the use-case.
34.	Gridspace www.gridspace.com	Gridspace is a platform for Conversational Intelligence. The company makes software that tells businesses about their mission-critical voice communications. The company's speech processing system, which combines new techniques in the fields of speech recognition, natural language processing, and Artificial Intelligence, turns conversational interactions into structured business data.
35.	TalkIQ www.talkiq.com	TalkIQ offers voice-to-text transcription and AI-driven insights for clients across a wide variety of industries. Using deep-learning algorithms, proprietary speech recognition and natural language processing, they give our clients visibility into all of their phone conversations, extracting insights for uses ranging from sales optimization, to customer service, compliance, and more. Their recommendations enable teams to save time and maximize productivity, through quick access to action-oriented data & analysis.
36.	Nexidia www.nexidia.com	Nexidia has applied years of research and development to deliver a comprehensive range of customer interaction analytics solutions. Nexidia holds numerous patents for phonetic indexing and searching capabilities which lie at the heart of their analytic products. Their proprietary technology and services support seamless integration with existing infrastructure, workflow processes, and innovative applications.

S. No.	Company Name	Notes
37.	Twilio www.twilio.com	Twilio is a cloud communications platform for building messaging, voice & video in the web and mobile applications.
38.	Capio www.capio.ai	Capio is focused on developing the engine that drives human-to-machine experiences. Powered by the latest advancements in Machine Learning and High Performance Computing, their contextually-aware ASR solutions power applications across multiple industries, including Automotive, Robotics, Call Center, Voice Biometrics, Home Automation and Transcription.
39.	MindMeld (formerly Expect Labs) www.mindmeld.com	MindMeld is a deep-Domain Conversational AI to power the next generation of Voice and Chat Assistants.
40.	Mobvoi www.mobvoi.com	Mobvoi is a Chinese Artificial Intelligence company providing voice recognition, semantic analysis and search technologies.
41.	Qurious www.qurious.io	Qurious leverages the power of Artificial Intelligence to give sales reps real-time answers to real-time questions and objections from customers.
42.	Pop Up Archive www.popuparchive.com	Pop Up Archive uses cutting edge speech-to-text technology to make sound searchable. For any audio file they tag, index & transcribe it automatically.
43.	Predix www.predix.io	Predix helps develop, deploy, and operate industrial apps at the edge and in the cloud. One can securely connect machines, data, and analytics to improve operational efficiency.
44.	Maana www.maana.io	Maana's Knowledge Platform™ enables companies to increase profitability. Maana's knowledge graph and algorithms dramatically accelerate knowledge discovery to provide a holistic view of the assets or processes enterprises want to optimize.
45.	Sentenai www.sentenai.com	Sentenai automates data engineering for data science. Sentenai's sensor data cloud database uses machine learning to automatically store and index data based on structure and patterns in the data.
46.	Planet OS www.planetos.com	Planet OS provides big data infrastructure for the energy industry to help them transform the way data-driven decisions are made. With specialized applications to easily integrate, exchange, and visualize their proprietary as well as third party data, renewable energy can become more competitive.
47.	Uptake www.uptake.com	Uptake's predictive platform enables major industries to improve uptime, streamline operations and spot growth opportunities. Uptake leverages its strategic partners' expertise to develop products that solve industry pain points and enable new data-driven business models and revenue sources.
48.	Imubit www.imubit.com	Imubit provides a platform for security, predictive analytics and big analog data of Internet of Things (IoT).
49.	Preferred Networks www.preferred-networks.jp	The focus of Preferred Networks is to apply real-time machine learning technologies to new applications in the emerging field of the Internet of Things (IoT).
50.	Thingworx www.thingworx.com	The ThingWorx Technology Platform was built from the ground up for the Internet of Things. It makes it easy to develop and deliver powerful IoT solutions that deliver transformative business value.
51.	Konux	KONUX is an IoT company that integrates smart sensor systems and AI-based analytics to continuously monitor asset and infrastructure condition, and enable predictive maintenance. With them, companies know their assets' health in real time and turn their

S. No.	Company Name	Notes
	www.konux.com	data into quality increasing and cost-saving actions.
52.	Alluvium www.alluvium.io	Alluvium delivers real-time collective intelligence to expert-driven industrial operations.
53.	GloT www.giotnetwork.com	GloT, Green IoT, is LPWAN (Low Power Wide Area Network) IoT total solution provider. They design and manufacture GloT devices and AP/Gateway, optimize and operate base station, manage gateway and data forwarding, develop application cloud system.
54.	Narrative Science www.narrativescience.com	Narrative Science is humanizing data like never before, with technology that interprets your data, then transforms it into Intelligent Narratives at unprecedented speed and scale. With Narrative Science, data becomes actionable—a powerful asset one can use to make better decisions, improve interactions with customers and empower employees.
55.	Loopai www.loop.ai	Loop Q is a software and hardware platform that consists of two core components: the Loop Learning Appliance and the Loop Reasoning Appliance. These components use proprietary unsupervised Deep Learning algorithms designed to iteratively learn language and concepts directly from source data -- without being explicitly told what to look for, or where.
56.	spaCy www.spacy.io	spaCy is a free open-source library featuring state-of-the-art speed and accuracy and a powerful Python API. spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython.
57.	Luminoso www.luminoso.com	Luminoso Technologies, Inc. is an AI-based deep analytics company that enables clients to rapidly discover value in their data. Analytics and Compass, their flagship products, reveal an accurate, unbiased, real-time understanding of what consumers are saying, uncovering actionable insights and the “unknown unknowns.” These insights are used to increase business performance and build better customer experiences.
58.	MonkeyLearn www.monkeylearn.com	MonkeyLearn builds highly scalable Machine Learning API to automate text classification.
59.	SigOpt www.sigopt.com	SigOpt is the optimization platform that amplifies research. SigOpt takes any research pipeline and tunes it. Their cloud-based ensemble of optimization algorithms is proven and seamless to deploy, and is used within the insurance, credit card, algorithmic trading and consumer packaged goods industries.
60.	Fuzzyio www.fuzzy.io	Fuzzy.io takes your knowledge and intuition about your business and converts that into an algorithm that improves automatically.
61.	Kite www.kite.com	Kite augments a coding environment with all the web's programming knowledge - intelligently sorted completions, all the documentation, and thousands of great examples – into the editor, helping to write the code faster.
62.	Rainforest www.rainforestqa.com	Rainforest QA unleashes the full potential of fast-moving development teams. Code more and ship faster with the only AI-powered Crowdttest Platform built for agile testing and development. Rapidly execute high-quality regression, functional and exploratory testing for web and mobile apps.
63.	Anodot www.anodot.com	Anodot is a real time analytics & automated anomaly detection solution that discovers & turns outliers in time series data into valuable business insights.
64.	Layer6ai www.layer6.ai	With Layer6ai's personalization engine, one can access Deep Learning to deliver personalized recommendations, search results and insight with unprecedented accuracy — in real time.
65.	1Qbit	1QBit mission is to apply breakthroughs in computation to machine intelligence and

S. No.	Company Name	Notes
	www.1qbit.com	optimization science through a widely accessible, quantum-ready software platform.
66.	AIDYIA www.aidyia.com	Aidyia deploy cutting edge artificial general intelligence (AGI) technology to identify patterns and predict price movements in global financial markets. Aidyia Limited is licensed by the Securities and Futures Commission of Hong Kong as an asset manager.
67.	Alation www.alation.com	Alation's enterprise collaborative data platform empowers employees inside of data-driven enterprises to find, understand, and use the right data for better, faster business decisions. Alation combines the power of machine learning with human insight to automatically capture information about what the data describes, where the data comes from, who's using it and how it's used. Customers include eBay, MarketShare, Square, and some of the world's largest finance and retail firms
68.	AlgoDynamix www.algodynamix.com	AlgoDynamix is a pioneering portfolio risk analytics company focusing on financially disruptive events. The deep data algorithms underpinning the AlgoDynamix analytics engine use primary data sources (the world's global financial exchanges) and proprietary unsupervised machine learning technology.
69.	Amplitude www.amplitude.com	Amplitude has analytics that helps companies of any size gain valuable insights from user behavior. Make better product decisions based on user insights and get the web and mobile analytics your team needs.
70.	Apcera www.apcera.com	Apcera is an enterprise-grade container management platform with workflow, orchestration, scheduling, storage, networking, plus a container engine, delivering a complete, enterprise-ready solution for businesses.
71.	BeyondCore www.beyondcore.com	Salesforce Has Completed Its Acquisition of BeyondCore. BeyondCore analyzes millions of data combinations in minutes, for unbiased answers, explanations and recommendations—to improve business metrics that matter, today.
72.	Big xyt www.big-xyt.com	Big xyt is a service provider for smart, flexible, efficient and smart data solutions dedicated to interactive analytics of large data sets. Big xyt engineers and operates solutions for storing and analyzing large amounts of data, enabling customers to transform data into information and decisions instantaneously.
73.	Bitfusion www.bitfusion.io	Bitfusion provides software that makes managing and using Deep Learning and AI infrastructure easy, elastic, and efficient. Bitfusion is completely changing that dynamic, enabling all organizations, data scientists, and developers to leverage Deep Learning software and high-performance hardware like GPUs quickly, productively, and cost-effectively.
74.	Blue Pool www.bluepool.tech	BluePool enables transformation of business models in the financial industry through sophisticated AI and Machine Learning based decision algorithms.
75.	BlueData www.bluedata.com	BlueData simplifies and accelerates your Big Data deployment, with lower costs and with fewer resources. You have a highly flexible, agile, and secure Big-Data-as-a-Service environment to enable faster time-to-insights and faster time-to-value – now available either on-premises or on AWS.
76.	BMLL Technologies www.bmltech.com	BMLL Technologies Ltd hosts full depth limit order book data on the AWS cloud. Machine learning API allows generalized pattern recognition. Also provides unlimited processing power through the AWS cloud with up to 80% reduction over usual cloud costs.
77.	Cambridge Semantics www.cambridgesemantics.com	Cambridge Semantics Inc. is an enterprise analytics and data management software company. Our software, allows IT departments and their business users to semantically link, analyze and manage diverse data whether internal or external, structured or unstructured, with speed, at big data scale and at the fraction of the implementation costs of using traditional approaches.

S. No.	Company Name	Notes
78.	Cask www.cask.co	Cask focus on Apps and Insights, not on Infrastructure and Integration. Cask makes building and running big data solutions easy, provides the first unified integration platform for big data that cuts down the time to production for data applications and data lakes by 80%.
79.	CircleCI www.circleci.com	CircleCI provides every developer state-of-the-art automated testing and continuous integration tools. Thousands of leading companies including Facebook, Kickstarter, Shyp, and Spotify rely on CircleCI to accelerate delivery of their code and enable developers to focus on creating business value fast.
80.	Citrine Informatics www.citrine.io	Citrine Informatics is the operating system for advanced materials and chemicals. Our data driven platform brings together the world's materials data and creates a corporate asset from your internal data. The result is that our Global 1000 customers hit their development, product design, and manufacturing targets 2-5 times faster, creating billions of dollars in market advantage.
81.	Declara www.declara.com	With Declara, you can manage your learning in one unified, intelligent knowledge engine that helps you get smarter faster. Declara provides a better way to discover the content that matters to you, your teams, and communities. Their proprietary CognitiveGraph™ engine uses semantic analysis and predictive analytics to generate your Learning Profile and power your personalized learning feed. The more it learns about what you find valuable, the more valuable it becomes in aiding your discovery of new content and insights.
82.	Deltix www.deltixlab.com	Deltix is a provider of software and services for quantitative research, algorithmic and automated systematic trading. Deltix software enables a complete straight through processing environment for the development and deployment of closely-integrated alpha generation and/or execution strategies.
83.	Elastic Search www.elastic.co/products/elasticsearch	Elasticsearch is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases. As the heart of the Elastic Stack, it centrally stores your data so you can discover the expected and uncover the unexpected.
84.	Extrahop www.extrahop.com	ExtraHop makes data-driven IT a reality with real-time analytics and cloud-based machine learning. ExtraHop analyze every single digital interaction occurring in the IT environment and turn that data into the most accurate and timely source of intelligence for everyone from IT Operations to Security to the CIO.
85.	GraphQL www.graphsql.com	GraphQL empowers fast real-time analysis to solve the toughest Big Data problems. Its main products includes, Anti-Fraud/Money Laundering, Customer Intelligence, Supply chain intelligence and energy efficient analytics
86.	IBM Watson www.ibm.com/watson	Watson products and APIs can understand data to reveal business-critical insights, and bring the power of cognitive computing to your organization.
87.	Illumio www.illumio.com/home	Illumio Adaptive Security Platform (ASP) helps you stop cyber threats, improve understanding of risk, and simplify security operations for applications in and across data center and cloud environments.
88.	KeenIO www.keen.io	Create intelligent apps APIs for capturing, analyzing, and embedding event data in everything you build. Keen was designed to give developers the flexibility and extensibility of a custom analytics stack, without the hassle and risk of managing big data infrastructure.
89.	Loggly www.loggly.com	Loggly is a cloud-based, enterprise-class log management service, serving more than 10000 customers.
90.	MapD www.mapd.com	MapD's mission is to not just make queries faster, but to create a fluid and immersive data exploration experience that removes the disconnect between an analyst and their data. The MapD analytics platform leverages the power of highly parallel GPUs.

S. No.	Company Name	Notes
		Traditional CPUs consists of a relatively small number of compute cores optimized for sequential processing.
91.	Mesosphere www.mesosphere.com	Mesosphere is democratizing the modern infrastructure we used at Twitter, AirBnB, and other webscale companies to quickly deliver data-driven services on any datacenter or cloud. Apache Mesos is the open-source distributed systems kernel at the heart of the Mesosphere DC/OS. It abstracts the entire datacenter into a single pool of computing resources, simplifying running distributed systems at scale.
92.	MetaMarkets www.metamarkets.com	Metamarkets is the provider of interactive analytics for programmatic marketing. Customers such as Twitter, AOL and LinkedIn use the Metamarkets platform to drive their business performance through intuitive access to real-time information. As an independent analytics software provider, Metamarkets gives its users the ability to see what's happening in the media marketplaces where they operate and provides the high-speed processing power needed to gain a competitive edge.
93.	Opera Solutions www.operasolutions.com	Opera Solutions is a global provider of advanced analytics software solutions that address the persistent problem of scaling Big Data analytics. Signal Hub™, the company's flagship technology platform, extracts and applies insights from the most complex data environments to address multiple business problems and opportunities, accelerating targeted business actions and radically reducing time-to-value.
94.	Pachyderm www.pachyderm.io	Pachyderm lets you store and analyze your data using containers. Pachyderm is a data lake that offers complete version control for data and leverages the container ecosystem to provide reproducible data processing.
95.	Priceonomics www.priceonomics.com	Priceonomics is helping companies turn their data into content marketing that performs. They come up with interesting topics and turn the insights into great content marketing that spreads. Their approach uses a key piece of software called Content Tracker, a content measurement dashboard.
96.	Saama www.saama.com	Saama's Fluid Analytics solutions maximize the customer's existing infrastructure, allowing them to focus on the white space between existing capabilities and the critical business questions that need to be answered.. Saama has broad experience in projects including visualization, MDM, Hadoop, cloud and other advanced analytics solutions, in industries such as life sciences, healthcare, insurance, financial services, CPG, high-tech and media.
97.	Scry Analytics www.scryanalytics.com	Scry Analytics provides its platforms, solutions and services for creating decision support systems for profit and not-for-profit organizations, the human society at large, and for giving back to the open source community. It combines subject matter expertise with automated computation and Artificial Intelligence to enhance data driven decision making under uncertainty.
98.	Semantic Web Companies www.semantic-web.at	Semantic Web Company GmbH (SWC) is the leading provider of graph-based metadata, search, and analytic solutions. A team of Linked Data experts provides consulting and integration services for semantic data and knowledge portals.
99.	Sentient www.sentient.ai	Sentient Investment Management is developing and applying proprietary quantitative trading and investment strategies built using the Sentient Technologies distributed Artificial Intelligence system, the most powerful system of its kind. Sentient Investment Management utilizes Sentient's Artificial Intelligence platform to continually evolve and optimize its investment strategies.
100.	TempoIQ www.tempoi.com	TempoIQ is platform for storing, analyzing, visualizing and monitoring the time series data that powers connected applications. Also help IoT companies produce insights in milliseconds instead of months.
101.	Thinknum www.thinknum.com	Analysts from around the world are building the largest repository of financial models on ThinkNum. ThinkNum monitors companies' websites and can be used to access the web's financial knowledge.

S. No.	Company Name	Notes
102.	ThoughtSpot www.thoughtspot.com	ThoughtSpot combines a new relational search engine for data analytics with a custom-built, in-memory relational data cache to provide sub-second response times to search queries run over billions of rows of data. ThoughtSpot scales to billions of records and thousands of users, and returns results in milliseconds.
103.	Tidemark www.tidemark.com	Collaborate and create accurate plans, timely budgets, and predict risk-weighted forecasts to drive informed decisions with a financial planning and analytics platform built for the digital era.
104.	Zuora www.zuora.com	Zuora is an end-to-end subscription management platform with the capabilities and insights you need to grow a healthy subscription business and create customers for life. Zuora's applications are designed to automate billing, commerce, and finance operations.
105.	Interana www.interana.com	Interana is defining and leading a new category of behavioral analytics for the digital economy which enables people to obtain insights from the actions people, products, or machines make over time
106.	LiftIgniter www.liftigniter.com	LiftIgniter uses machine learning to enhance personalization & recommendation. LiftIgniter optimizes for the highest click-through-rate, engagement, reduced bounced, sharing, and conversion. Create brand affinity and Personalization which is revolutionizing the Internet.
107.	Quantenstein www.quantenstein.org	Quantenstein is an integrated software platform for automated long-term value investing that builds on the latest developments in Deep Learning technology. For a given investment universe and set of constraints, Quantenstein optimizes client-specific financial performance metrics based on large quantities of fundamental accounting data to assemble tailored investment portfolios.
108.	Pit.AI www.pit.ai	Pit.AI has developed a core machine learning technology that is based on a nonconventional quantitative finance approach and novel machine learning techniques. Aim is to develop Artificial Intelligence agents that learn how to invest by themselves and that can develop more granular market insights than human experts using massive data sets
109.	ASI Datascience www.asidatascience.com	ASI Datascience empowers organisations to become more data-driven by providing first class software, skills, and advisory solutions. To unlock the power of data in gaining competitive advantage, we help organisations to make sense of the data, big and small. We believe firmly in providing innovative, simple and easy to implement solutions that generate business value.
110.	Thought Machine www.thoughtmachine.net	Thought Machine is changing the technology underlying our banking systems. With a world class team expert in cloud computing, machine learning, finance, design and app building, they are creating the banks of the future.
111.	Baidu Big Data Lab www.bdl.baidu.com	Baidu Big Data Lab (BDL) was founded in July 2014, which is one of the three interrelated teams of Baidu Research. BDL focuses on large-scale machine learning algorithms, Core Search Technologies and big data applications in areas such as predictive analytics, vertical business solutions, large data structure algorithms, and intelligent systems research. Currently, BDL has established world-leading large-scale machine learning platform, Deep Learning based personal secretary platform, intelligent spatial-temporal data analytics platform, and intelligent business big data solution platform. Meanwhile, BDL also obtained prolific research achievements and industrial experience in a variety of big data application domains, such as Core Search Technologies, health-care, retail, finance, tourism, and real estate.

7) Technology Consulting Firms

S. No.	Company Name	Notes
1.	Ufora www.ufora.com	Ufora provides expertise to design and engineer distributed systems infrastructure that allows data science to scale. The platform acts as a dynamic code optimizer, figuring out how to parallelize operations as they happen. They have made their kit open source and it is really great for everyone, as technology will likely find its way into all sorts of places.
2.	System2 www.sstm2.com	System2 is a quantitative extension of internal investment teams. They help analysts leverage big data to improve model estimates and answer seemingly impossible questions.
3.	Psych Signal www.psychsignal.com	Psych Signal provides real time 'Trader Mood', data, analytics and indices for financial institutions & investment professionals seeking an edge. They created a natural language processing engine which interprets social media text in the context of stock prices. Their technology parses millions of online conversations every day in order to quantify the public's mood about specific stocks and other securities.
4.	Clearstory Data www.clearstorydata.com	ClearStory Data speeds data blending from multiple disparate sources, intelligently combining them into holistic insights, and delivering interactive business insights.
5.	Equity Data Science www.equitydatascience.com	Equity Data Science (EDS) is a quantitative platform for professional investors, which includes investment analytics, portfolio management, and risk modeling. EDS saves time, is easy-to-use, affordable, and helps investors generate better performance.
6.	Capital Cube www.online.capitalcube.com	CapitalCube provides financial research and content for investors, information providers, finance portals and media. Their automatically generated narratives and predictive analytics empower investment decisions based on in-depth analysis of a company's peer performance, earnings quality, and dividend strength.
7.	Lexalytics www.lexalytics.com	Lexalytics process billions of unstructured documents every day, globally. Translate text into profitable decisions; make state-of-the-art cloud and on-premise text and sentiment analysis technologies that transform customers' thoughts and conversations into actionable insights. The on premise Salience® and SaaS Semantria® platforms are implemented in a variety of industries for social media monitoring, reputation management and voice of the customer programs.
8.	Ripple www.ripple.com	Ripple provides global financial settlement solutions to enable the world to exchange value like it already exchanges information – giving rise to an Internet of Value (IoV). Ripple solutions lower the total cost of settlement by enabling banks to transact directly, instantly and with certainty of settlement. Banks are partnering with Ripple to improve their cross-border payment offerings, and to join the growing, global network of financial institutions and market makers laying the foundation for the Internet of Value.
9.	minds.ai www.minds.ai	minds.ai is a full service Artificial Intelligence design consultancy. They design and build custom neural networks for businesses.
10.	RailsBank www.railsbank.com	Railsbank is a banking and compliance platform that connects together a global network of partner banks with companies who want API access to global banking. Railsbank simplifies on-boarding companies to our banking partners, then gives access via the Railsbank API to banking services such as creating digital ledgers, connecting digital ledgers to real bank accounts, issuing IBANs for ledgers, receiving money, sending money, converting money (FX), collecting money (direct debit), issuing cards, and managing credit.
11.	M Science www.msscience.com	M Science is a data-driven research and analytics firm, uncovering new insights for leading financial institutions and corporations. M Science is revolutionizing research, discovering new data sets and pioneering methodologies to provide actionable intelligence. And combine the best of finance, data and technology to create a truly unique value proposition for both financial services firms and top corporations.

APPENDIX

Techniques for Data Collection from Websites

Collection of website data is enabled through many bespoke and generic packages. We list commonly used packages in R, Python and Java below.

R- Packages

Source	R Package	Package/Data URL
Yahoo, FRED, Oanda, Google	Quantmod	CRAN/ http://www.quantmod.com/
Quandl	Quandl	CRAN/ http://www.quandl.com/help/packages/r
TrueFX	TFX	CRAN/ http://rpubs.com/gsee/TFX
Bloomberg	Rblpapi	CRAN.rstudio.com/web/packages/Rblpapi/index.html
Interactive Broker	IBrokers	CRAN/ https://www.interactivebrokers.com/en/main.php
Datastream	rdatastream	https://github.com/fcocoemas/rdatastream
Penn World Table	pwt	CRAN/ https://pwt.sas.upenn.edu/
Yahoo, FRED, Oanda	fImport	CRAN/ http://www.rmetrics.org/
ThinkNum	Thinknum	CRAN/ http://thinknum.com/
Webscraping	rvest	CRAN
	xml	CRAN
Twitter	twitterR	CRAN
LinkedIn	Rlinkedin	CRAN

Python – Packages

Source	Python Package	Package/Data URL
Quandl	Quandl	https://www.quandl.com/tools/python
Bloomberg	BLPAPI	https://www.bloomberglabs.com/api/libraries/
Webscraping	Beautiful Soup	PyPi
	Selenium	PyPi
Twitter	twitter	PyPi
LinkedIn	python-linkedin	PyPi

Java-Packages

Source	Java Package	Package/Data URL
Webscraping	Jaunt	http://jaunt-api.com/
	jsoup	https://jsoup.org/

We further illustrate common web scraping techniques through the following ten use-cases.

Example One: Getting Financial Data from Yahoo/Google using Pandas DataReader

Pandas DataReader helps you download data from a number of data sources via a simple code. Following is an example to pull stock price levels from Yahoo Finance using the Datareader. The urllib2 library helps you make all requests through your corporation's proxy.

```
import datetime
import pandas as pd
import pandas_datareader.data as web
import urllib2

def readDataFromWeb(assets, startDate, endDate, source = 'yahoo'):

    PROXIES = {'https' : "https://proxy.companynamename.net:8080"}
    Proxy = urllib2.ProxyHandler(PROXIES)
    Opener = urllib2.build_opener(proxy)

    urllib2.install_opener(opener)
    prices = {}
    volumes = {}
    for asset in assets :
        try:
            df = web.DataReader(asset,source,start=startDate,end=endDate)
            prices[asset] = df['Adj Close']
            volumes[asset] = df['Volume']
        except:
            print "Error: skipping",asset
    prices = pd.DataFrame(prices)
    volumes = pd.DataFrame(volumes)
    return pd.Panel({'Price': prices, 'Return' : prices.pct_change(), 'Volume': volumes})

def main():
    start = datetime.date(2016, 12,20)

    end = datetime.date.today()-datetime.timedelta(1)

    AssetList = ['YHOO','AAPL','IBM','F']

    Data = readDataFromWeb(AssetList,start,end)

    Data.Price
```

Output:

```
>>> Data.Price
```

	AAPL	F	IBM	YHOO
Date				
2017-02-15	135.509995	12.63	181.679993	45.650002
2017-02-16	135.350006	12.54	181.429993	45.160000
2017-02-17	135.720001	12.58	180.669998	45.099998
2017-02-21	136.699997	12.69	180.259995	45.500000

Example Two: Scraping Population Data from Wikipedia using Rvest

The “Rvest” library provides a simple solution to scrape data through any website. Following is an example of getting population data through Wikipedia’s page on U.S. states and territories. The library “magrittr” helps to use pipes that make the code easier to interpret.

```
library(magrittr)      # To use pipes that would make coding easier and elegant
library(rvest)         # Version 'rvest_0.3.2'
setInternet2()        # Use internet.dll so that R url request appears as an
Internet Explorer request
Sys.setenv(https_proxy="xxxx.xxxx:8080")      # Setting proxy in R

url <- "http://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population"
population <- url %>%
  html() %>%
  html_nodes(xpath='//*[@id="mw-content-text"]/table[1]') %>%
  html_table()

population <- population[[1]]
Data<-population[,c('State or territory','Population estimate, July 1, 2016')]
head(Data)
```

Output:

```
> head(Data)
  State or territory Population estimate, July 1, 2016
1      California      39,250,017
2         Texas      27,862,596
3       Florida      20,612,439
4       New York      19,745,289
5  Pennsylvania      12,802,503
6       Illinois      12,801,539
```

Example Three: Get all S&P 500 companies from Wikipedia

Another illustration of using Rvest and Magritte is provided by the example below.

```
library(magrittr)      # To use pipes that would make coding easier and elegant
library(rvest)         # Version 'rvest_0.3.2'
setInternet2()        # Use internet.dll so that R url request appears as an Internet
Explorer request
Sys.setenv(https_proxy=" https://proxy.companynamename.net:8080")  # Setting proxy in R

url <- "https://en.wikipedia.org/wiki/List_of_S%26P_500_companies"
AllCompanies <- url %>% read_html() %>%
  html_nodes(xpath='/html/body/div[3]/div[3]/div[4]/table[1]') %>%
  html_table(fill=TRUE)
AllCompanies <- AllCompanies[[1]]
head(AllCompanies)
```

Output:

Ticker symbol	Security	SEC filings	GICS Sector	GICS Sub Industry
MMM	3M Company	reports	Industrials	Industrial Conglomerates
ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment
ABBV	AbbVie	reports	Health Care	Pharmaceuticals
ACN	Accenture plc	reports	Information Technology	IT Consulting & Other Services
ATVI	Activision Blizzard	reports	Information Technology	Home Entertainment Software
AYI	Acuity Brands Inc	reports	Industrials	Electrical Components & Equipment
ADBE	Adobe Systems Inc	reports	Information Technology	Application Software
AAP	Advance Auto Parts	reports	Consumer Discretionary	Automotive Retail
AES	AES Corp	reports	Utilities	Independent Power Producers & Energy Traders
AET	Aetna Inc	reports	Health Care	Managed Health Care

Example Four: Getting all the references from a Wikipedia page

```
library(magrittr)      # To use pipes that would make coding easier and elegant
library(rvest)         # Version 'rvest_0.3.2'
setInternet2()         # Use internet.dll so that R url request appears as an Internet Explorer request
Sys.setenv(https_proxy=" https://proxy.companyname.net:8080") # Setting proxy in R

page <- read_html("https://en.wikipedia.org/wiki/Julienning")
sources <- page %>%
html_nodes(".references li") %>%
html_text()
head(sources)
```

Output:

```
"^ \"The S&P 500 is becoming the S&P 502.\". USA Today. 2014-07-30. Retrieved 2014-12-02. "
"^ \"S&P U.S. Indices Methodology Update\" (PDF). www.spice-indices.com. 2015-01-12. Retrieved 2015-03-13. "
"^ \"Our Company\". AvalonBay. 1994-12-31. Retrieved 2012-02-10. "
"^ http://www.bostonscientific.com/en-US/about-us/new-marlborough-headquarters.html"
"^ \"Helmerich & Payne, Inc\". Hpinc.com. Retrieved 2012-02-10. "
"^ \"Offices | Southwestern Energy\". Swn.com. Retrieved 2012-02-10. "
```

Example Five: Scraping Population Data from Wiki using R/XML

XML is an old library which can also be used to pull out data by nodes. Rvest is preferable to using XML library, since it can also use CSS selectors to find elements in a page.

```
library(magrittr)      # To use pipes that would make coding easier and elegant
library(XML)           # Version 'XML_3.98-1.5'
library(RCurl)         # Use R's Curl library to call URLs
setInternet2()         # Use internet.dll so that R url request appears as an Internet
                        # Explorer request
Sys.setenv(https_proxy=" https://proxy.companyname.net:8080") # Setting proxy in R

XML_doc<-
htmlParse(getURL('https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_pop
ulation'),asText=TRUE)
XML_table_node <- getNodeSet(XML_doc,'/html/body/div[3]/div[3]/div[4]/table[1]')[[1]]
population <- readHTMLTable(XML_table_node, stringsAsFactors=FALSE)
Data<-population[,c('State or territory','Population estimate, July 1, 2016')]
head(Data)
```

Output:

```
> head(Data)
  State or territory Population estimate, July 1, 2016
1      California      39,250,017
2         Texas      27,862,596
3       Florida      20,612,439
4       New York      19,745,289
5   Pennsylvania      12,802,503
6       Illinois      12,801,539
. |
```

Example Six: Using Glassdoor API to pull out company reviews

You need to have created an account in Glassdoor to be able to access its API. Please note that the API for a trial account gives limited output. The output below gives the Featured Reviews of the companies. The API as of now only supports output in JSON structure. Glassdoor API has REST-ful architecture within the MVC (Model-View-Controller) paradigm. Standard packages exist in Python (e.g. package json) and Java/Spring framework to bind JSON to objects.

```
library(RCurl)         # Version 'RCurl_1.95-4.8'
library(tidyjson)      # Version 'tidyjson_0.2.1'
library(magrittr)      # Version 'magrittr_1.5'
setInternet2()         # Use internet.dll so that R url request appears as an Internet
                        # Explorer request

Sys.setenv(https_proxy=" https://proxy.companyname.net:8080") # Set company's proxy
settings

# Go through the Glassdoor documentation to create the required URL
```



```
url
="https://api.glassdoor.com/api/api.htm?t.p=126685&t.k=g896Yh2jrEe&userip=170.148.132.135&
useragent=Mozilla&format=json&v=1&action=employers&q="

data<-getURL(url)                # Get output in JSON

# Get data from different nodes
Output1<- data %>% enter_object("response") %>% enter_object("employers") %>% gather_array
%>%
enter_object("featuredReview")%>% gather_keys %>% append_values_string(column.name='Data')
Output2<- data %>% enter_object("response") %>% enter_object("employers") %>% gather_array
%>%
gather_keys %>% append_values_string(column.name='Data')

# Data Formatting
RawOutput<-Output1[Output1$key %in% c('pros','cons'),c('key','Data')]
pros<-RawOutput[RawOutput$key %in% c('pros'),]
cons<-RawOutput[RawOutput$key %in% c('cons'),]
Name<-Output2[Output2$key %in% c('name'),c('key','Data')]
FinalOutput<-as.data.frame(cbind(Name$Data,pros$Data,cons$Data))
names(FinalOutput)<-c("Name","Pros","Cons")
write.csv(FinalOutput,"CompanyReviews.csv") # Write to csv
print(FinalOutput,right=F)                # Print the output to console
```

Output:

Name	Pros	Cons
IBM	Disclaimer: A lot of what I'm writing below of course	1. Unfortunately, IBM still uses the "normal distribution"
Walmart	Advancement opportunities, great at developing skills, gr	Understaffing issues negatively affects all parts of their bu
Accenture	There are a lot of pros working for Accenutre. They have	This is not an opportunity for those that do not want to wc
Target	Target is really making some great improvements right n	With all of the transformation, there is a little uncertainty ;
Tata Consultancy Services	one of the best company in india	long working hours for people
Infosys	Best Environment to work at	Work more than 12 hours at times
Deloitte	These folks know exactly what they are doing. They set	Work-life balance can become poor, especially during tigl
Microsoft	Collaboration, Innovation, a great place to work	Can't think of any cons
Hewlett-Packard	Innovative Company, lots of muscle to flex in the IT world	If you are a remote employee, forget advancement. Force

Example Seven: Using Quandl library to pull out financial data

Quandl collects financial and economic datasets from hundreds of publishers on one convenient platform. The R library 'Quandl' helps you get this data in the simplest way possible. A similar library 'Quandl' exists to fetch data from quandl using Python.

```
install.packages("Quandl")
library(Quandl)                # Version 'Quandl_2.8.0'
```

```
#This call gets US GDP from FRED and puts it in a data frame (Note that you need to know the "Quandl code" of each dataset you download. In the above example, it is FRED/GDP):
```

```
mydata = Quandl("FRED/GDP")      # Get all the data
```

```
mydata = Quandl(c("FRED/GDP.1", "WIKI/AAPL.4"),start_date="2011-12-31", end_date="2017-02-28",collapse="annual")
```

```
# Use the start/end date settings to define for which dates data is to be pulled out.  
'Collapse' settings help you to determine the frequency of the data
```

Output:

	DATE	FRED.GDP - VALUE	WIKI.AAPL - Close
1	2012-12-31	16297.3	532.1729
2	2013-12-31	16999.9	561.0200
3	2014-12-31	17692.2	110.3800
4	2015-12-31	18222.8	105.2600
5	2016-12-31	18855.5	115.8500
6	2017-12-31	NA	136.9900

Example Eight: Using Selenium to scrape data from NASDAQ website

Selenium automates browsers. Primarily, it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can also be automated as well.

```
import pandas as pd
from numpy import nan
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC

## create a pandas dataframe to store the scraped data
df = pd.DataFrame(index=range(40),
                  columns=['company', 'quarter', 'quarter_ending',
                           'total_revenue', 'gross_profit', 'net_income',
                           'total_assets', 'total_liabilities', 'total_equity',
                           'net_cash_flow'])

## launch the Chrome browser
path = "C:/geckodriver.exe"
browser = webdriver.Firefox(executable_path=my_path)
browser.maximize_window()
url_form = "http://www.nasdaq.com/symbol/{}/financials?query={}&data=quarterly"
financials_xpath = "//tbody/tr/th[text()='{}']/../td[contains(text(), '$')]"

## company ticker symbols
symbols = ["amzn", "aapl", "fb", "ibm", "msft"]

def get_elements(xpath):
    ## find the elements
    elements = browser.find_elements_by_xpath(xpath)
    ## if any are missing, return all nan values
    if len(elements) != 4:
        return [nan] * 4
    ## otherwise, return just the text of the element
    else:
        text = []
        for e in elements:
            text.append(e.text)
        return text

for i, symbol in enumerate(symbols):
    url = url_form.format(symbol, "income-statement")
    browser.get(url)

    company_xpath = "//h1[contains(text(), 'Company Financials')]"
    company = WebDriverWait(browser, 10).until(EC.presence_of_element_located((By.XPATH,
    company_xpath))).text

    quarters_xpath = "//thead/tr[th[1][text()='Quarter:']/th[position()>=3]"
    quarters = get_elements(quarters_xpath)

    quarter_endings_xpath = "//thead/tr[th[1][text()='Quarter Ending:']/th[position()>=3]"
    quarter_endings = get_elements(quarter_endings_xpath)

    total_revenue = get_elements(financials_xpath.format("Total Revenue"))
    gross_profit = get_elements(financials_xpath.format("Gross Profit"))
    net_income = get_elements(financials_xpath.format("Net Income"))
    ## navigate to balance sheet quarterly page
```

```
url = url_form.format(symbol, "balance-sheet")
browser.get(url)
total_assets = get_elements(financials_xpath.format("Total Assets"))
total_liabilities = get_elements(financials_xpath.format("Total Liabilities"))
total_equity = get_elements(financials_xpath.format("Total Equity"))
## navigate to cash flow quarterly page
url = url_form.format(symbol, "cash-flow")
browser.get(url)

net_cash_flow = get_elements(financials_xpath.format("Net Cash Flow"))

for j in range(4):
    row = i + j
    df.loc[row, 'company'] = company
    df.loc[row, 'quarter'] = quarters[j]
    df.loc[row, 'quarter_ending'] = quarter_endings[j]
    df.loc[row, 'total_revenue'] = total_revenue[j]
    df.loc[row, 'gross_profit'] = gross_profit[j]
    df.loc[row, 'net_income'] = net_income[j]
    df.loc[row, 'total_assets'] = total_assets[j]
    df.loc[row, 'total_liabilities'] = total_liabilities[j]
    df.loc[row, 'total_equity'] = total_equity[j]
    df.loc[row, 'net_cash_flow'] = net_cash_flow[j]

print df
```

Output:

	company	quarter	quarter_ending	total_revenue	gross_profit
AMZN	Company Financials	4th	12/31/2016	\$43,741,000	\$14,782,000
AAPL	Company Financials	1st	12/31/2016	\$78,351,000	\$30,176,000
FB	Company Financials	4th	12/31/2016	\$8,809,000	\$7,761,000
IBM	Company Financials	4th	12/31/2016	\$21,770,000	\$10,893,000
MSFT	Company Financials	2nd	12/31/2016	\$24,090,000	\$14,189,000
MSFT	Company Financials	1st	9/30/2016	\$20,453,000	\$12,609,000
MSFT	Company Financials	4th	6/30/2016	\$20,614,000	\$12,635,000
MSFT	Company Financials	3rd	3/31/2016	\$20,531,000	\$12,809,000

Example Nine: Using 'twitterR' to get data from Twitter through R

TwitterR requires you to create your OAuth credentials to be able to use twitter API. An example of getting data using the 'twitterR' package is provided below.

```
library(dplyr)
library(purrr)
library(twitterR) # Version 'twitterR_1.1.9'
key = "xxxxxxxxxxxxxxxxxxxxx"
secret = "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
```

```
token = "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
token_secret = "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
setup_twitter_oauth(key, secret, token, token_secret)

trump_tweets <- userTimeline("realDonaldTrump", n = 3200)
trump_tweets_df <- tbl_df(map_df(trump_tweets, as.data.frame()))
library(tidyr)

tweets <- trump_tweets_df %>%
  select(id, statusSource, text, created) %>%
  extract(statusSource, "source", "Twitter for (.*)<") %>%
  filter(source %in% c("iPhone", "Android"))
```

Example Ten: Using 'BeautifulSoup' to get ETF expense ratios from Wisdom Tree through Python

BeautifulSoup is a popular Python library for web-scraping. It can parse XML/HTML pages and includes common tree-traversal algorithms.

```
import urllib2
from bs4 import BeautifulSoup
import re
import os
import pandas as pd
from pandas import ExcelWriter
import sys

proxy_support = urllib2.ProxyHandler({'https' : "https://proxy.companynamename.net:8080"})
opener = urllib2.build_opener(proxy_support)
urllib2.install_opener(opener)

def soup_from_url(url):
    res_html = urllib2.urlopen(url).read()
    soup = BeautifulSoup(res_html.decode('utf-8','ignore'), 'html.parser')
    return soup

types = ["currency","equity","fixed-income","alternative"]
list_of_etfs = ["DGR","DGR","DQD","DQI","DXC","DXG","DXO","DXP","DXU"]
outfile = open("wisdomtree_etfs_ER_latam.csv","w")
for etf in list_of_etfs:
    for type_of_etf in types:
        url = r"https://www.wisdomtree.com/etfs/"+ type_of_etf+ "/" + etf.lower()
        try:
            soup = soup_from_url(url)
        except Exception, e:
            continue
    rows = soup.findAll('td')
    for i in range(0,len(rows)):
        txt = rows[i].text.strip()
        if (txt == "Expense Ratio"):
            er = rows[i+1].text.strip()
            if len(er)>0:
```

```
                break;
            if (txt == "Net Expense Ratio, amount charged to shareholder1" ):
                er = rows[i+1].text.strip()
                if len(er)>0:
                    break;
            to_write = etf+", "+str(er)+"\n"
            print "writing Expense Ratio for etf", etf, str(er)
            outfile.writelines(to_write)
```

Output:

```
'extract etf expense ratio from wisdomtree website'
writing Expense Ratio for etf DGR 0.45%
writing Expense Ratio for etf DGR 0.45%
writing Expense Ratio for etf DQD 0.45%
writing Expense Ratio for etf DQI 0.45%
writing Expense Ratio for etf DXC 0.45%
writing Expense Ratio for etf DXG 0.45%
writing Expense Ratio for etf DXO 0.45%
writing Expense Ratio for etf DXP 0.45%
writing Expense Ratio for etf DXU 0.45%
```

Packages and Codes for Machine Learning

In much of applied data science, practitioners do not implement Machine Learning directly. Implementations of common techniques are available in various programming languages. We list popular examples below in C++, Java, Python and R. For a comprehensive list of algorithm implementations, see the websites of [Awesome-Machine-Learning](#) and [MLoss](#).

C++	
Package	Description
OpenCV	Real-time computer vision (Python, Java interface also available)
Caffe	Clean, readable and fast Deep Learning framework
CNTK	Deep Learning toolkit by Microsoft
DSSTNE	Deep neural networks using GPUs with emphasis on speed and scale
LightGBM	High performance gradient boosting
CRF++, CRFSuite	Segmenting/labeling sequential data & other Natural Language Processing tasks

JAVA	
Package	Description
MALLET	Natural language processing, document classification, clustering etc.
H2O	Distributed learning on Hadoop, Spark; APIs available in R, Python, Scala, REST/JSON
Mahout	Distributed Machine Learning
MLlib in Apache Spark	Distributed Machine Learning library in Spark
Weka	Collection of Machine Learning algorithms
Deeplearning4j	Scalable Deep Learning for industry with parallel GPUs

PYTHON	
Package	Description
NLTK	Platform to work with human language data
XGBoost	Extreme Gradient Boosting (Tree) Library
scikit-learn	Machine Learning built on top of SciPy
keras	Modular neural network library based on Theano/Tensorflow
Lasagne	Lightweight library to build and train neural networks in Theano
Theano /Tensorflow	Efficient multi-dimensional arrays operations
MXNet	Lightweight, Portable, Flexible Distributed/Mobile Deep Learning with Dynamic, Mutation-aware Dataflow Dep Scheduler; for Python, R, Julia, Go, Javascript and more
gym	Reinforcement learning from OpenAI
NetworkX	High-productivity software for complex networks
PyMC3	Markov Chain Monte Carlo sampling toolkit
statsmodels	Statistical modeling and econometrics

R	
Package	Description
glmnet	Penalized regression
class::knn	K-nearest neighbor
FKF	Kalman filtering
XgBoost	Boosting
gam	Generalized additive model
stats::loess	Local Polynomial Regression Fitting
MASS::lda	Linear and quadratic discriminant analysis
e1071::svm	Support Vector Machine
depmixS4	Hidden Markov Model

stats::kmeans	Clustering
stats::prcomp, fastICA	Factor Analysis
rstan	Markov Chain Monte Carlo sampling toolkit
MXnet	Neural Network

Python Codes for Popular ML Algorithms

Below we provide sample Python codes, demonstrating use popular Machine Learning algorithms.

Python

Lasso

```
>>> from sklearn.linear_model import Lasso
>>> model = Lasso(alpha=0.1)
>>> model.fit([[-1,-1],[0,0],[1,1]],[-1,0,1])
Lasso(alpha=0.1, copy_X=True, fit_intercept=True, max_iter=1000,
      normalize=False, positive=False, precompute=False, random_state=None,
      selection='cyclic', tol=0.0001, warm_start=False)
>>> print(model.coef_)
[ 0.85  0. ]
>>> print(model.intercept_)
0.0
>>> model.predict([[3,-3]])
array([ 2.55])
```

Ridge

```
>>> from sklearn.linear_model import Ridge
>>> model = Ridge(alpha=0.1)
>>> model.fit([[-1,-1],[0,0],[1,1]],[-1,0,1])
Ridge(alpha=0.1, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=None, solver='auto', tol=0.001)
>>> print(model.coef_)
[ 0.48780488  0.48780488]
>>> print(model.intercept_)
0.0
>>> model.predict([[3,-3]])
array([ -4.21884749e-15])
```

ElasticNet

```
>>> from sklearn.linear_model import ElasticNet
>>> model = ElasticNet(alpha=0.1)
>>> model.fit([[-1,-1],[0,0],[1,1]],[-1,0,1])
ElasticNet(alpha=0.1, copy_X=True, fit_intercept=True, l1_ratio=0.5,
      max_iter=1000, normalize=False, positive=False, precompute=False,
      random_state=None, selection='cyclic', tol=0.0001, warm_start=False)
>>> print(model.coef_)
[ 0.44604258  0.44554178]
>>> print(model.intercept_)
0.0
>>> model.predict([[3,-3]])
array([ 0.00150239])
```

K-Nearest Neighbors (Python)

```
>>> from sklearn.neighbors import NearestNeighbors
>>> import numpy as np
>>> X = np.array([[ -1, -2], [-2, -2], [-3, -5], [1, 1], [2, 2], [4, 4]])
>>> model = NearestNeighbors(n_neighbors=2, algorithm='ball_tree').fit(X)
>>> distances, indices = model.kneighbors([[0,0]])
>>> distances
array([[ 1.41421356,  2.23606798]])
>>> indices
array([[3, 0]], dtype=int64)
```

Logistic Regression

```
>>> from sklearn.linear_model import LogisticRegression
>>> model = LogisticRegression(penalty="l2")
>>> model.fit([[-2,-3],[1,0],[1,1]],[1,0,1])
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                    intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                    penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                    verbose=0, warm_start=False)
>>> print(model.coef_)
[[-0.36284928 -0.09783526]]
>>> print(model.intercept_)
[ 0.28002799]
>>> model.predict([[3,3]])
array([0])
```

SVM

```
>>> from sklearn.svm import SVC
>>> import numpy as np
>>> X = np.array([[-3, -2], [-4, -5], [3, 4], [4, 5]])
>>> y = np.array([1, 1, 2, 2])
>>> model = SVC()
>>> model.fit(X,y)
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape=None, degree=3, gamma='auto', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
>>> print(model.predict([[0,0]]))
[1]
```

Random Forest Classifier

```
>>> from sklearn.ensemble import RandomForestClassifier
>>> import numpy as np
>>> X = np.array([[-3, -2], [-4, -5], [3, 4], [4, 5]])
>>> y = np.array([1, 1, 2, 2])
>>> model = RandomForestClassifier(n_estimators=3)
>>> model.fit(X,y)
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                        max_depth=None, max_features='auto', max_leaf_nodes=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, n_estimators=3, n_jobs=1,
                        oob_score=False, random_state=None, verbose=0,
                        warm_start=False)
>>> print(model.predict([[0,0]]))
[1]
```

K-Means

```
>>> from sklearn.cluster import KMeans
>>> import numpy as np
>>> model = KMeans(n_clusters=2, random_state=0).fit(X)
model.labels_
array([0, 0, 1, 1])
>>> model.predict([[1, 2], [-1, -1]])
array([1, 0])
>>> model.cluster_centers_
array([[ -3.5, -3.5],
       [ 3.5,  4.5]])
```

PCA

```
>>> from sklearn.decomposition import PCA
>>> import numpy as np
>>> X = np.array([[-3, -2], [-4, -5], [3, 4], [4, 5]])
>>> model = PCA(n_components=2)
>>> model.fit(X)
PCA(copy=True, n_components=2, whiten=False)
>>> print(model.explained_variance_ratio_)
[ 0.99388963  0.00611035]
```

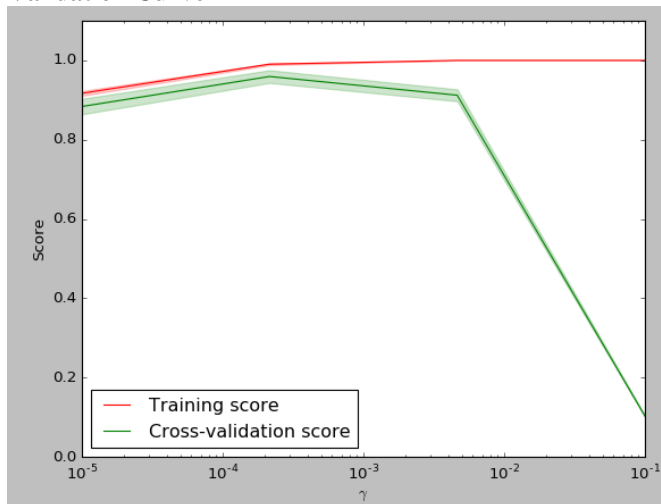
Mathematical Appendices

Model Validation Theory

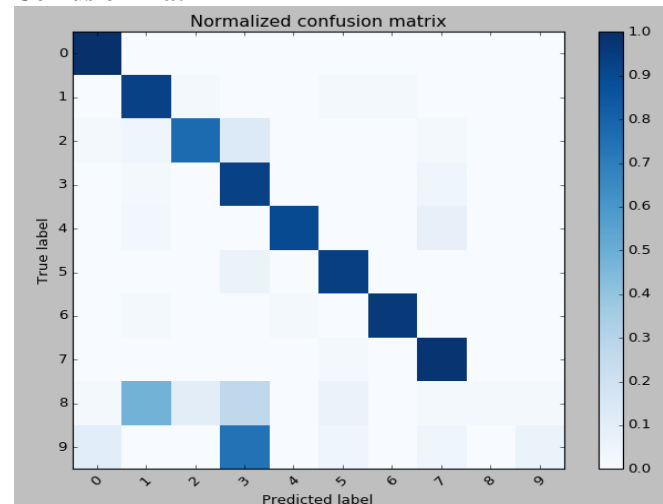
Validation Curve: Optimal value for hyperparameters can be visually inspected through a graph called “Validation Curve”. Here, the input hyper-parameter is varied along a range of values, and an accuracy score is computed both over the entire training set and through cross-validation. Graph below shows the validation curve for support vector machine classifier as the parameter gamma is varied. For high values of gamma, SVM overfits yielding a low cross-validation accuracy score and a deceptively high training accuracy score.

Confusion Matrix: Another way to visualize the output of a classifier is to evaluate its normalized confusion matrix. On a database of hand-written digits, we employed a linear SVM model. Alongside i -th row of the confusion matrix (denoting a true label of $i+1$), the j -th element represents the probability that predicted digit is equal to $j+1$.

Validation Curve



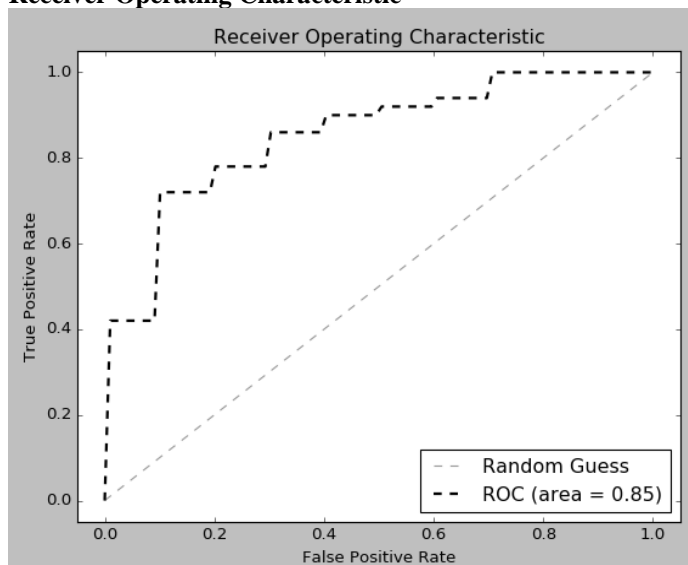
Confusion Matrix



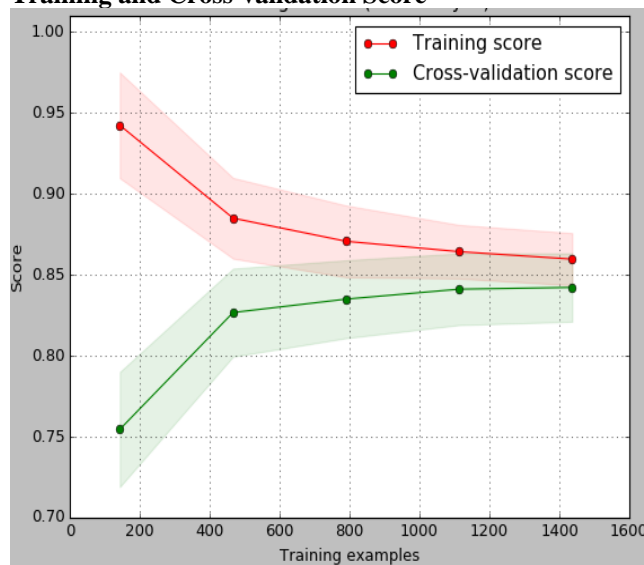
Receiver Operating Characteristic: Another common tool to measure the quality of a classifier is to use the Receiver Operating Characteristic. We use a binary-valued dataset and used a linear SVM to fit the data. We used 5-fold cross-validation. To compare classifier via the ROC curve, choose the curve with higher area under the curve (i.e. the curve increases sharply and steeply from origin).

Training and Cross-validation Score: In many complex datasets, we find that increasing the number of training examples increases score through cross-validation. The training score does not have a fixed behavior as the number of training examples increases.

Receiver Operating Characteristic

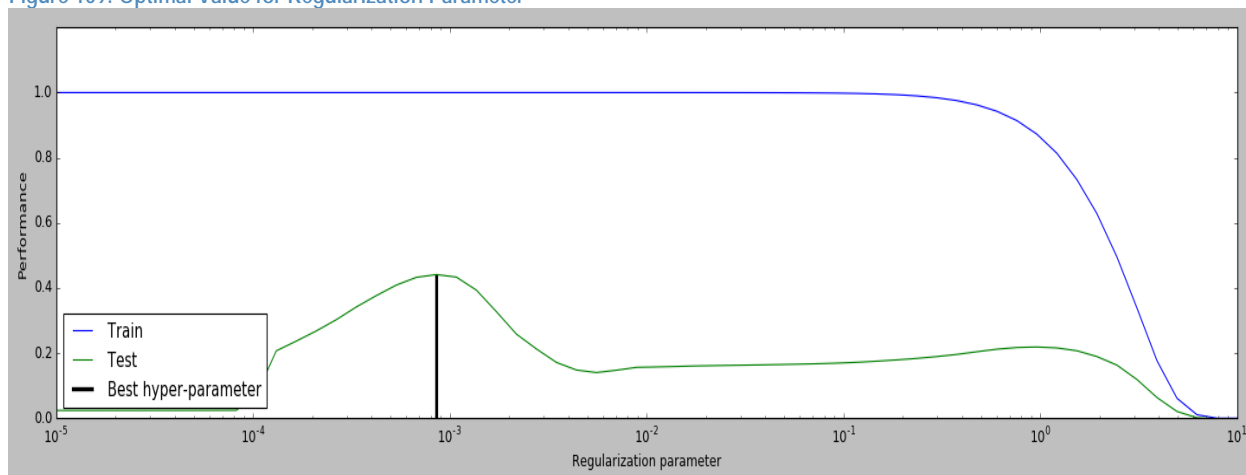


Training and Cross-validation Score



Optimal Value for Regularization Parameter: Another tool to choose a model is to note the value of regularization parameter where performance on test set is the best.

Figure 109: Optimal Value for Regularization Parameter



Source: J.P. Morgan Quantitative and Derivatives Strategy

Model Validation Theory : Vapnik-Chervonenkis Dimension

We can address both the questions through the notion of Vapnik-Chervonenkis dimension⁵⁸.

Even without invoking learning theory, we can use the Chernoff Bound (or Hoeffding inequality) to relate the training error to the test error in the case where samples are drawn i.i.d. from the same underlying distribution for both the training and test error. If $\{z_i\}_{i=1}^m$ were m samples drawn from Bernoulli(ϕ) distribution, then one would estimate ϕ as

$$\hat{\phi} = \frac{1}{m} \sum_{i=1}^m z_i$$

following the usual maximum likelihood rule. For any $\gamma > 0$, it can be shown that

$$P(|\phi - \hat{\phi}| > \gamma) < 2e^{-2\gamma^2 m}.$$

This tells us that as sample size increases, the ML estimator is efficient and the discrepancy between training and test error is likely to diminish.

Consider the case of binary classification, where we have m samples $S = \{(\underline{x}^{(i)}, y^{(i)})_{i=1}^m\}$, with $y^{(i)} \in \{0,1\}$. Further, assume that these samples are drawn i.i.d. from a distribution D . Such an assumption, proposed by Valiant in 1984, is called the PAC or Probably Approximately Correct assumption. We can define the training error as

$$\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^m 1\{h(\underline{x}^{(i)}) \neq y^{(i)}\}$$

and the test/generalization error as

$$\epsilon(h) = P_{(x,y) \sim D}(h(x) \neq y).$$

Consider further a hypothesis class H of binary classifiers. Under empirical risk minimization, one seeks to minimize the training error to pick the optimal classifier or hypothesis as

$$\hat{h} = \arg \min_{h \in H} \hat{\epsilon}(h).$$

If $|H| = k$, then it can be shown for any fixed m, δ that

$$\hat{\epsilon}(h) \leq \left(\min_{h \in H} \epsilon(h) \right) + \left(2 \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}} \right)$$

with probability exceeding $1 - \delta$.

The first term in RHS above is the bias term that decreases as k increases. The second term in RHS above represents the variance that increases as k increases. This again indicates the Variance-Bias tradeoff we alluded to before. More importantly, we can reorganize the terms in the inequality above to show that as long as

$$m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} = O\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right),$$

⁵⁸ VC dimension is covered in Vapnik (1996). The PAC (Probably Approximately Correct) framework was developed in Valiant (1984) and Kearns and Vazirani (1994). AIC and BIC were proposed in Akaike (1973) and Schwarz (1978), respectively. For further discussion on cross-validation and Bayesian model selection, see Madigan and Raftery (1994), Wahba (1990), Hastie and Tibshirani (1990).

we can always bound the generalization error of the optimal classifier by

$$\hat{\epsilon}(h) \leq \left(\min_{h \in H} \epsilon(h) \right) + 2\gamma.$$

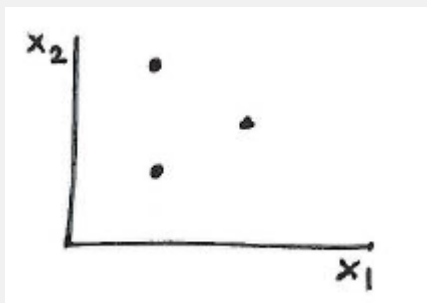
This result shows that the number of training samples must increase logarithmically with number of classifiers in the class H . If $|H| = k$, then we need $\log(k)$ parameters to describe it, which in turn implies that the number of input examples to grow only linearly with the number of parameters in the model.

The above analysis holds true for simple sets of classifiers. If we wish to choose the optimal linear classifier

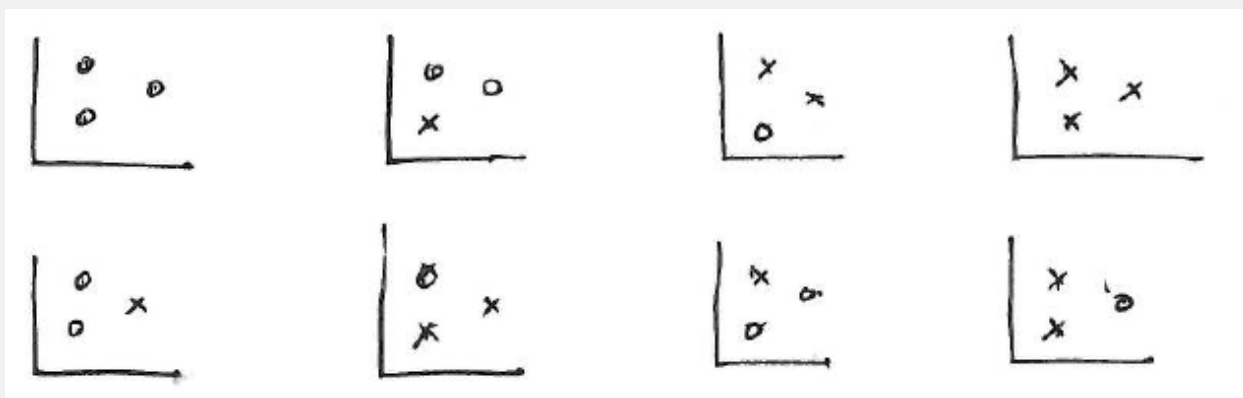
$$H = \{h_{\underline{\theta}} : h_{\underline{\theta}}(\underline{x}) = 1(\underline{\theta}^t \underline{x} \geq 0); \underline{\theta} \in \mathbb{R}^n\},$$

then $|H| = \infty$ and the above simplistic analysis does not hold. To address this practical case, we need the notion of Vapnik-Chervonenkis dimension.

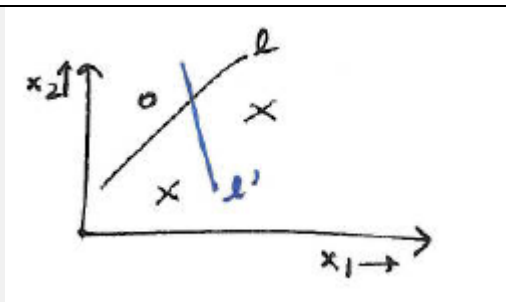
Consider three points as shown.



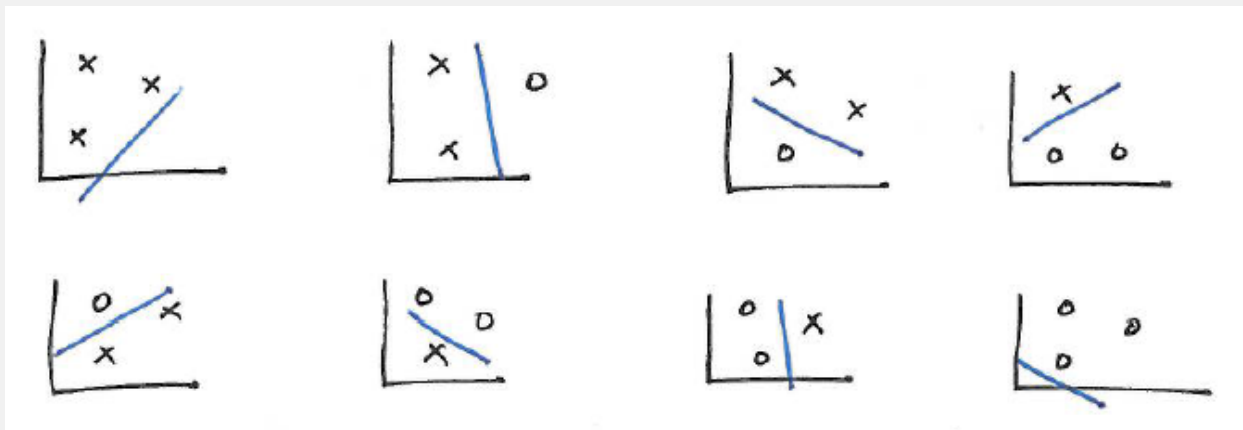
A labeling refers to marking on those points as either 0 or 1. Marking zero by O and one by X, we get eight labeling as follows →



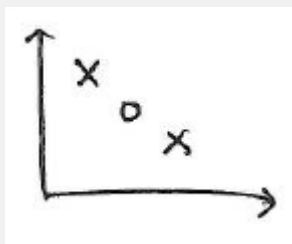
We say that a classifier – say, a linear hyperplane denoted by l – can realize a labeling if it can separate the zeros and ones into two separate blocks and achieve a zero training error. For example, the line l in figure is said to realize the labeling below, while the line l' fails to do so.



We can extend the notion of *realizing a labeling* to a set of classifiers through the notion of *shattering*. Given a set of points $S = \{\underline{x}^{(i)}\}_{i=1}^d$, we say that H shatters S if H can realize any labeling on S . In other words, for any set of labels $\{y^{(i)}\}_{i=1}^d$, there exists a hypothesis $h \in H$ such that for all $i \in \{1, \dots, d\}$, we have $h(\underline{x}^{(i)}) = y^{(i)}$. For example, the set of linear classifiers can shatter S shown in the figure above, since we can always fit a straight line separating the O and X marks. This is illustrated in the figure below.



Note that linear classifiers cannot shatter S' below.



Further, the reader can try and check that linear classifiers cannot shatter any set S with four or more elements. So the maximum size of a set that, under some configuration, can be shattered by the set of linear classifiers (with two parameters) is 3. We say formally that the Vapnik-Chervonenkis dimension of H is 3 or $VC(H) = 3$. The Vapnik-Chervonenkis dimension $VC(H)$ for a hypothesis class H is defined as the size of the largest set that is shattered by H .

With the above definitions, we can state the foundational result of learning theory. For H with $VC(H) = d$, we can define the optimal classifier as

$$h^* = \arg \min_{h \in H} \epsilon(h),$$

and the classifier obtained by minimizing the training error over m samples as $\hat{h} = \arg \min_{h \in H} \hat{\epsilon}(h)$.

Then with probability exceeding $1 - \delta$, we have

$$\epsilon(\hat{h}) \leq \epsilon(h^*) + O\left(\sqrt{\frac{d}{m} \log \frac{m}{\delta} + \frac{1}{m} \log \frac{1}{\delta}}\right).$$

This implies that, for

$$\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$$

to hold with probability exceeding $1 - \delta$, it suffices that $m = O(d)$.

This reveals to us that the number of training samples must grow linearly with the VC dimension (which tends to be equal to the number of parameters) of the model.

Particle Filtering

Signal modelling and state inference given noisy observations naturally leads us to stochastic filtering and state-space modelling. Wiener provided a solution for a stationary underlying distribution. Kalman provided a solution for non-stationary underlying distribution: the optimal linear filter (first truly adaptive filter) based on assumptions on linearity and Gaussianity. Extensions try to overcome limitations of linear and Gaussian assumptions but do not provide closed-form solutions to the distribution approximations required. Bayesian inference aims to elucidate sufficient variables which accurately describe the dynamics of the process being modeled. Stochastic filtering underlies Bayesian filtering and is an inverse statistical problem: you want to find inputs as you are given outputs (Chen 2003). The principle foundation of stochastic filtering lies in recursive Bayesian estimation where we are essentially trying to compute the joint posterior. More formally, recovering the state variable \mathbf{x}_t given F_t with data up to and including time t , to essentially remove observation errors and compute the posterior distribution over the most recent state: $P(\mathbf{X}_t | \mathbf{Y}_{0:t})$.

There are two key assumptions underlying the recursive Bayesian filter: (i) that the state process follows a first-order Markov process:

$$p(\mathbf{x}_n | \mathbf{x}_{0:n-1}, \mathbf{y}_{0:n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1})$$

and (ii) that the observations and states are independent:

$$p(\mathbf{y}_n | \mathbf{x}_{0:n-1}, \mathbf{y}_{0:n-1}) = p(\mathbf{y}_n | \mathbf{x}_n)$$

From Bayes rule given \mathbf{Y}_n as the set of observations $\mathbf{y}_{0:n} := \{\mathbf{y}_0, \dots, \mathbf{y}_n\}$ the conditional posterior density function (pdf) of \mathbf{x}_t is defined as:

$$p(\mathbf{x}_n | \mathbf{Y}_n) = \frac{p(\mathbf{y}_n | \mathbf{x}_n) p(\mathbf{x}_n | \mathbf{Y}_{n-1})}{p(\mathbf{y}_n | \mathbf{Y}_{n-1})}$$

In turn, the posterior density function $p(\mathbf{x}_n | \mathbf{Y}_n)$ is defined by three key terms:

Prior: the knowledge of the model is described by the prior $p(\mathbf{x}_n | \mathbf{Y}_{n-1})$

$$p(\mathbf{x}_n | \mathbf{Y}_{n-1}) = \int p(\mathbf{x}_n | \mathbf{x}_{n-1}) p(\mathbf{x}_{n-1} | \mathbf{Y}_{n-1}) d\mathbf{x}_{n-1}$$

Likelihood: $p(\mathbf{y}_n | \mathbf{x}_n)$ essentially determines the observation noise

Evidence: the denominator of the pdf involves an integral of the form

$$p(\mathbf{y}_n | \mathbf{Y}_{n-1}) = \int p(\mathbf{y}_n | \mathbf{x}_n) p(\mathbf{x}_n | \mathbf{Y}_{n-1}) d\mathbf{x}_n$$

The calculation and or approximation of these three terms is the base of Bayesian filtering and inference.

Particle filtering is a recursive stochastic filtering technique which provides a flexible approach to determine the posterior distribution of the latent variables given the observations. Simply put, particle filters provide online adaptive inference where the underlying dynamics are non-linear and non-Gaussian. The main advantage of sequential Monte Carlo methods⁵⁹

⁵⁹ For more information on Bayesian sampling, see Gentle (2003), Robert and Casella (2004), O'Hagan and Forster (2004), Rasmussen and Ghahramani (2003), Rue, Martino and Chopin (2009), Liu (2001), Skare, Bolviken and Holden (2003), Ionides (2008), Gelman and Hill (2007), Cook, Gelman and Rubin (2006), Gelman (2006, 2007). Techniques to improve Bayesian posterior simulations are covered in van Dyk and Meng (2001), Liu (2003), Roberts and Rosenthal (2001) and Brooks, Giudici and Roberts (2003). For adaptive MCMC, see Andrieu and Robert (2001) and Andrieu and Thoms (2008), Peltola, Marttinen and Vehtari (2012); for reversible jump MCMC, see Green (1995); for trans-dimensional MCMC, see Richardson and Green (1997) and Brooks, Giudici and Roberts (2003); for perfect-simulation

is that they do not rely on any local linearization or abstract functional approximation. This is at the cost of increased computational expense though given breakthroughs in computing technology and the related decline in processing costs, this is not considered a barrier except in extreme circumstances.

Monte Carlo approximation using particle methods calculates the expectation of the posterior density function by importance sampling (IS). The state-space is partitioned into which particles are filled with respect to some probability measure. The higher this measure the denser the particle concentration. Specifically, from earlier:

$$p(x_t | y_{0:t}) = \frac{p(y_t | x_t) p(x_t | y_{0:t-1})}{p(y_t | y_{0:t-1})}$$

We approximate the state posterior by $f(x_t)$ with i samples of $x_t^{(i)}$. To find the mean $\mathbb{E}[f(x_t)]$ of the state posterior $p(x_t | y_{0:t})$ at t , we generate state samples $x_t^{(i)} \sim p(x_t | y_{0:t})$. Though theoretically plausible, empirically we are unable to observe and sample directly from the state posterior. We replace the state posterior by a proposal state distribution (importance distribution) π which is proportional to the true posterior at every point: $\pi(x_t | y_{0:t}) \propto p(x_t | y_{0:t})$. We are thus able to sample sequentially independently and identically distributed draws from $\pi(x_t | y_{0:t})$ giving us:

$$\begin{aligned} \mathbb{E}[f(x_t)] &= \int f(x_t) \frac{p(x_t | y_{0:t})}{\pi(x_t | y_{0:t})} \pi(x_t | y_{0:t}) dx_t \\ &\approx \frac{\sum_{i=1}^N f(x_t^{(i)}) w_t^{(i)}}{\sum_{i=1}^N w_t^{(i)}} \end{aligned}$$

When increasing the number of draws N this average converges asymptotically (as $N \rightarrow \infty$) to the expectation of the true posterior according to the central limit theorem (Geweke 1989). This convergence is the primary advantage of sequential Monte Carlo methods as they provide asymptotically consistent estimates of the true distribution $p(x_t | y_{0:t})$ (Doucet & Johansen 2008).

IS allows us to sample from complex high-dimensional distributions though exhibits linear increases in complexity upon each subsequent draw. To admit fixed computational complexity we use sequential importance sampling (SIS). There are a number of critical issues with SIS primarily the variance of estimates increases exponentially with n and leads to fewer and fewer non-zero importance weights. This problem is known as weight degeneracy. To alleviate this issue, states are resampled to retain the most pertinent contributors, essentially removing particles with low weights with a high degree of certainty (Gordon et al. 1993). It addresses degeneracy by replacing particles with high weight with many particles with high inter-particle correlation (Chen 2003). The sequential importance resampling (SIR) algorithm is provided in the mathematical box below:

Mathematical Box [Sequential Importance Resampling]

1. Initialization: for $i = 1, \dots, N_p$, sample

$$\mathbf{x}_0^{(i)} \sim p(\mathbf{x}_0)$$

with weights $W_0^{(i)} = \frac{1}{N_p}$.

For $t \geq 1$

2. Importance sampling: for $i = 1, \dots, N_p$, draw samples

$$\hat{\mathbf{x}}_t^{(i)} \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(i)})$$

MCMC, see Propp and Wilson (1996) and Fill (1998). For Hamiltonian Monte Carlo (HMC), see Neal (1994, 2011). The popular NUTS (No U-Turn Sampler) was introduced by Gelman (2014). For other extensions, see Girolami and Calderhead (2011), Betancourt and Stein (2011), Betancourt (2013a, 2013b), Romeel (2011), Leimkuhler and Reich (2004).

- set $\hat{\mathbf{x}}_{0:t}^{(i)} = \{\mathbf{x}_{0:t-1}^{(i)}, \hat{\mathbf{x}}_t^{(i)}\}$.
- Weight update: calculate importance weights
$$W_t^{(i)} = p(\mathbf{y}_t | \hat{\mathbf{x}}_t^{(i)})$$
 - Normalize weights
$$\tilde{W}_t^{(i)} = \frac{W_t^{(i)}}{\sum_{j=1}^{N_p} W_t^{(j)}}$$
 - Resampling: generate N_p new particles $\mathbf{x}_t^{(i)}$ from the set $\{\hat{\mathbf{x}}_t^{(i)}\}$ according to the importance weights $\tilde{W}_t^{(i)}$.
 - Repeat from importance sampling step 2.

Resampling retains the most pertinent particles however destroys information by discounting the potential future descriptive ability of particles – it does not really prevent sample impoverishment it simply excludes poor samples from calculations, providing future stability through short-term increases in variance.

Our Adaptive Path Particle Filter⁶⁰ (APPF) leverages the descriptive ability of naively discarded particles in an adaptive evolutionary environment with a well-defined fitness function leading to increased accuracy for recursive Bayesian estimation of non-linear non-Gaussian dynamical systems. We embed a generation based adaptive particle switching step into the particle filter weight update using the transition prior as our proposal distribution. This enables us to make use of previously discarded particles ψ if their discriminatory power is higher than the current particle set. [More details on the theoretical underpinnings and formal justification of the APPF can be found in Hanif (2013) and Hanif & Smith (2012).]

$$W_t^{(i)} = \max[p(\mathbf{y}_t | \hat{\mathbf{x}}_t^{(i)}), p(\mathbf{y}_t | \check{\mathbf{x}}_t^{(i)})] \text{ where } \check{\mathbf{x}}_t^{(i)} \sim p(\mathbf{x}_t | \psi_{t-1}^{(i)}) \text{ and } \check{\mathbf{x}}_{0:t}^{(i)} = \{\mathbf{x}_{0:t-1}^{(i)}, \check{\mathbf{x}}_t^{(i)}\}$$

Mathematical Box [Adaptive Path Particle Filter]

- Initialization: for $i = 1, \dots, N_p$, sample
$$\mathbf{x}_0^{(i)} \sim p(\mathbf{x}_0)$$

$$\psi_0^{(i)} \sim p(\mathbf{x}_0)$$

with weights $W_0^{(i)} = \frac{1}{N_p}$

For $t \geq 1$
- Importance sampling: for $i = 1, \dots, N_p$, draw samples
$$\hat{\mathbf{x}}_t^{(i)} \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(i)}) \text{ set}$$

$$\hat{\mathbf{x}}_{0:t}^{(i)} = \{\mathbf{x}_{0:t-1}^{(i)}, \hat{\mathbf{x}}_t^{(i)}\} \text{ and draw}$$

$$\check{\mathbf{x}}_t^{(i)} \sim p(\mathbf{x}_t | \psi_{t-1}^{(i)}) \text{ set}$$

$$\check{\mathbf{x}}_{0:t}^{(i)} = \{\mathbf{x}_{0:t-1}^{(i)}, \check{\mathbf{x}}_t^{(i)}\}$$
- Weight update: calculate importance weights
$$W_t^{(i)} = \max[p(\mathbf{y}_t | \hat{\mathbf{x}}_t^{(i)}), p(\mathbf{y}_t | \check{\mathbf{x}}_t^{(i)})]$$

Evaluate:

if $p(\mathbf{y}_t | \hat{\mathbf{x}}_t^{(i)}) > p(\mathbf{y}_t | \check{\mathbf{x}}_t^{(i)})$ then

$$\hat{\mathbf{x}}_t^{(i)} = \psi_t^{(i)}$$

end if
- Normalize weights

⁶⁰ More details on the theoretical underpinnings and formal justification of the APPF can be found in Hanif (2013) and Hanif and Smith (2012).

$$\tilde{W}_t^{(i)} = \frac{W_t^{(i)}}{\sum_{j=1}^{N_p} W_t^{(j)}}$$

5. Commit pre-resample set of particles to memory:

$$\{\psi_t^{(i)}\} = \{\hat{x}_t^{(i)}\}$$

6. Resampling: generate N_p new particles $\mathbf{x}_t^{(i)}$ from the set $\{\hat{x}_t^{(i)}\}$ according to the importance weights $\tilde{W}_t^{(i)}$.
7. Repeat from importance sampling step 2.

Financial Example: Stochastic Volatility Estimation

Traditional measures of volatility are either market views or estimated from the past. Under such measures the correct value for pricing derivatives cannot be known until the derivative has expired. As the volatility measure is not constant, not predictable and not directly observable it is best modeled as a random variable (Wilmott 2007). Understanding the dynamics of the volatility process in tandem with the dynamics of the underlying asset in the same timescale enable us to measure the stochastic volatility process. However, modelling volatility as a stochastic process needs an observable volatility measure: this is the stochastic volatility estimation problem.

The Heston stochastic volatility model is among the most popular stochastic volatility models and is defined by the coupled two-dimensional stochastic differential equation:

$$\begin{aligned} dX(t)/X(t) &= \sqrt{V(t)}dW_X(t) \\ dV(t) &= \kappa(\theta - V(t))dt + \varepsilon\sqrt{V(t)}dW_V(t) \end{aligned}$$

where $\kappa, \theta, \varepsilon$ are strictly positive constants, and W_X and W_V are scalar Brownian motions in some probability measure; we assume that $dW_X(t) \cdot dW_V(t) = \rho dt$, where the correlation measure ρ is some constant in $[-1, 1]$. $X(t)$ represents an asset price process and is assumed to be a martingale in the chosen probability measure. $V(t)$ represents the instantaneous variance of relative changes to $X(t)$ – the stochastic volatility⁶¹. The Euler discretization with full truncation⁶² of the model takes the form:

$$\begin{aligned} \ln \hat{X}(t + \Delta) &= \ln \hat{X}(t) - \frac{1}{2}\hat{V}(t)^+\Delta + \sqrt{\hat{V}(t)^+}Z_X\sqrt{\Delta} \\ \hat{V}(t + \Delta) &= V(t) + \kappa(\theta - \hat{V}(t)^+)\Delta + \varepsilon\sqrt{\hat{V}(t)^+}Z_V\sqrt{\Delta} \end{aligned}$$

where \hat{X} the observed price process and \hat{V} the stochastic volatility process are discrete-time approximations to X and V , respectively, and where Z_X and Z_V are Gaussian random variables with correlation ρ . The operator $x^+ = \max(x, 0)$ enables the process for V to go below zero thereafter becoming deterministic with an upward drift $\kappa\theta$. To run the particle filters we need to calibrate the parameters $\kappa, \theta, \varepsilon$.

Experimental Results – S&P 500 Stochastic Volatility

To calibrate the stochastic volatility process for the S&P 500 Index we ran a 10,000 iteration Markov-chain Monte Carlo calibration to build an understanding of the price process (observation equation) and volatility process (state equation). We

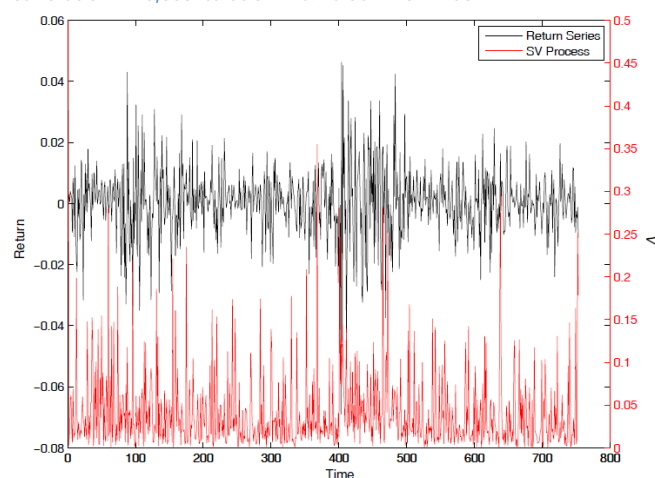
⁶¹ SV is modeled as a mean-reverting square-root diffusion, with Ornstein-Uhlenbeck dynamics (a continuous-time analogue of the discrete-time first-order autoregressive process).

⁶² A critical problem with naive Euler discretization enables the discrete process for V to become negative with non-zero probability, which makes the computation of $\sqrt{\hat{V}}$ impossible.

took the joint MAP (maximum a posteriori) estimate⁶³ of κ and θ from our MCMC calibration as per Chib, et al. (2002). The Heston model stochastic volatility calibration for SPX can be seen in the first figure below, where we can see the full truncation scheme forcing the SV process to be positive, and the associated parameter evolution can be seen in the second figure (Hanif & Smith 2013).

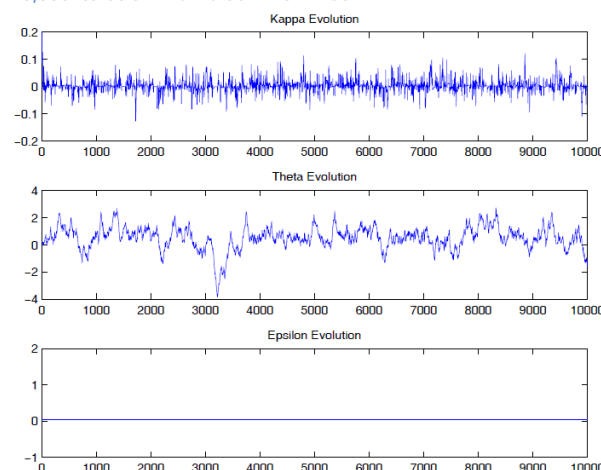
Of note, we can see ε is a small constant throughout. This is attributable to the fact ε represents the volatility of volatility. If it were large we would not observe the coupling (trend/momentum) between and amongst securities in markets as we do.

Figure 110: Heston model SPX daily closing Stochastic Volatility calibration – 10,000 iteration MCMC Jan '10 – Dec '12



Source: Hanif (2013), J.P. Morgan QDS.

Figure 111: Heston model SPX Parameter Estimates and Evolution – 10,000 iteration MCMC Jan '10 – Dec '12



Source: Hanif (2013), J.P. Morgan QDS.

Given the price process we estimate the latent stochastic volatility process using the SIR, MCMC-PF⁶⁴, PLA⁶⁵ and APPF particle filters run with $N = 1,000$ particles and systematic resampling⁶⁶. Results can be seen in the table and figure below. We can clearly see the APPF providing more accurate estimates of the underlying stochastic volatility process compared to the other particle filters: the APPF provides statistically significant improvements in estimation accuracy compared to the other filters.

Figure 112: Heston model experimental results: RMSE mean and execution time in seconds

Particle Filter	RMSE	Exec. (s)
PF (SIR)	0.05282	3.79
MCMC-PF	0.05393	59.37
PLA	0.05317	21.30
APPF	0.04961	39.33

Source: Hanif (2013), J.P.Morgan QDS

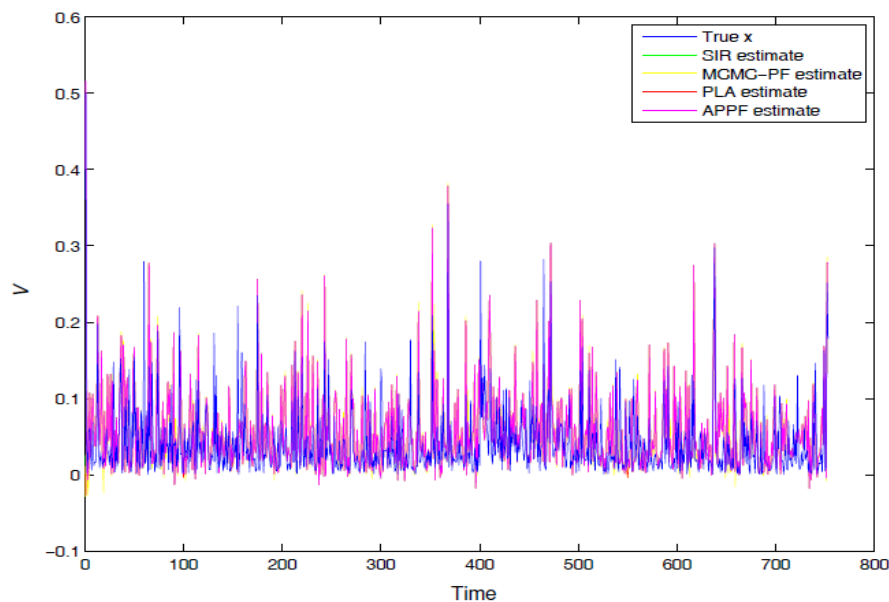
⁶³ The MAP estimate is a Bayesian parameter estimation technique which takes the mode of the posterior distribution. It is unlike maximum likelihood based point estimates which disregard the descriptive power of the MCMC process and associated pdfs.

⁶⁴ The Markov-chain Monte Carlo particle filter (MCMC-PF) attempts to reduce degeneracy by jittering particle locations, using Metropolis-Hastings to accept moves.

⁶⁵ The particle learning particle filter (PLA) performs an MCMC after every 50 iterations.

⁶⁶ There are a number of resampling schemes that can be adopted. The three most common schemes are systematic, residual and multinomial. Of these multinomial is the most computationally efficient though systematic resampling is the most commonly used and performs better in most, but not all, scenarios compared to other sampling schemes (Douc & Cappé 2005).

Figure 113: Heston model estimates for SPX – filter estimates (posterior means) vs. true state



Source: Hanif (2013), J.P.Morgan QDS

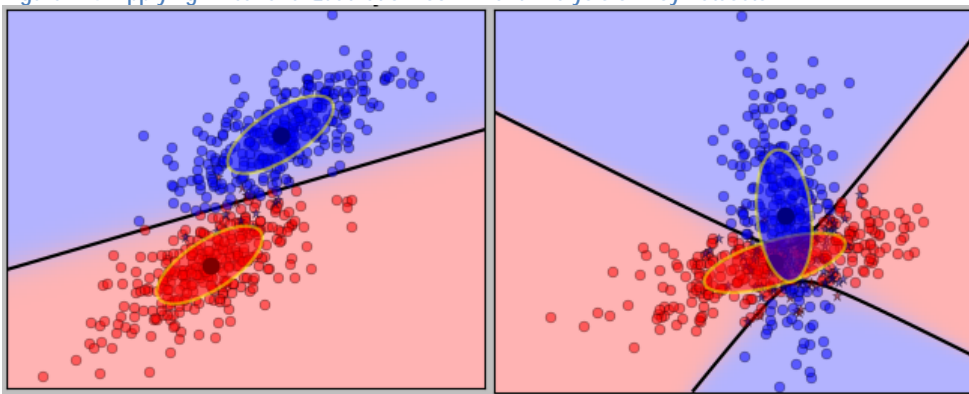
These results go some way in showing that selective pressure from our generation-gap and distribution-recombination method does not lead to premature convergence. We have implicitly included a number of approaches to handling premature convergence in dynamic optimization problems with evolutionary computation (Jin & Branke, 2005). Firstly, we generate diversity after a change by resampling. We maintain diversity throughout the run through the importance sampling diffusion of the current and past generation particle set. This generation based approach enables the learning algorithm to maintain a memory, which in turn is the base of Bayesian inference. And finally, our multi-population approach enables us to explore previously, possibly unexplored regions of the search space.

Linear and Quadratic Discriminant Analysis

Learning algorithms can be classified as either discriminative or generative algorithms⁶⁷. In Discriminative Learning algorithms, one seeks to learn the input-to-output mapping directly. Examples of this approach include Rosenblatt's Perceptron and Logistic Regression. In such discriminative learning algorithms, one models $p(y|\underline{x})$ directly. An alternative approach would be to learn $p(y)$ and $p(\underline{x}|y)$ from the data, and use Bayes theorem to recover $p(y|\underline{x})$. Learning algorithms adopting this approach of modeling both $p(y)$ and $p(\underline{x}|y)$ are called Generative Learning algorithms, as they equivalently learn the joint distribution $p(\underline{x}, y)$ of the input and output processes.

Fitting Linear Discriminant Analysis on data with same covariance matrix and then Quadratic Discriminant Analysis on data with different covariance matrices yields the two graphs below.

Figure 114: Applying Linear and Quadratic Discriminant Analysis on Toy Datasets.



Source: J.P.Morgan Macro QDS

⁶⁷ For discriminant analysis (linear, quadratic, flexible, penalized and mixture), see Hastie et al (1994), Hastie et al (1995), Tibshirani (1996b), Hastie et al (1998) and Ripley (1996). Laplace's method for integration is described in Wong and Li (1992). Finite Mixture Models are covered by Bishop (2006), Stephens (2000a, 2000b), Jasra, Holmes and Stephens (2005), Papaspiliopoulos and Roberts (2008), Ishwaran and Zarepour (2002), Fraley and Raftery (2002), Dunson (2010a), Dunson and Bhattacharya (2010).

Mathematical Model for Generative Models like LDA and QDA

In Linear Discriminant Analysis or LDA (also, called Gaussian Discriminant Analysis or GDA), we model

$y \sim \text{Bernoulli}(\phi)$, $\underline{x}|y = 0 \sim N(\underline{\mu}_0, \Sigma)$, and $\underline{x}|y = 1 \sim N(\underline{\mu}_1, \Sigma)$. Note that the means are different, but the covariance matrix is same for $y=0$ and $y=1$ case. The joint log-likelihood is given by

$$l(\phi, \underline{\mu}_0, \underline{\mu}_1, \Sigma) = \log \prod_{i=1}^m p(\underline{x}^{(i)}, y^{(i)}; \phi, \underline{\mu}_0, \underline{\mu}_1, \Sigma).$$

Standard optimization yields the maximum likelihood answer as

$$\phi = \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\}, \underline{\mu}_0 = \frac{\sum_{i=1}^m \mathbf{1}\{y^{(i)}=0\} \underline{x}^{(i)}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)}=0\}}, \underline{\mu}_1 = \frac{\sum_{i=1}^m \mathbf{1}\{y^{(i)}=1\} \underline{x}^{(i)}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)}=1\}} \text{ and } \Sigma = \frac{1}{m} \sum_{i=1}^m (\underline{x}^{(i)} - \underline{\mu}_{y^{(i)}})(\underline{x}^{(i)} - \underline{\mu}_{y^{(i)}})^T.$$

The above procedure fits a linear hyperplane to separate regions marked by classes $y = 0$ and $y = 1$.

Other points to note are:

- If we assume $\underline{x}|y = 0 \sim N(\underline{\mu}_0, \Sigma_0)$ and $\underline{x}|y = 1 \sim N(\underline{\mu}_1, \Sigma_1)$, viz. we assume different covariance for the two distributions, then we obtain a quadratic boundary and the consequent learning algorithm is called Quadratic Discriminant Analysis.
- If the data were indeed Gaussian, then it can be shown that as the sample size increases, LDA asymptotically performs better than any other algorithm.
- It can be shown that Logistic Regression is more general than LDA/QDA; hence logistic regression will outperform LDA/QDA when the data is non-Gaussian (say, Poisson distributed).
- LDA with the covariance matrix restricted to a diagonal leads to the Gaussian Naïve Bayes model.
- LDA coupled with the Ledoit-Wolf shrinkage idea from portfolio management yields better results than plain LDA.

A related algorithm is Naïve Bayes with Laplace correction. We describe it briefly below.

Naïve Bayes is a simple algorithm for text classification, which works surprisingly well in practice in spite of its simplicity. We create a vector \underline{x} of length $|\mathcal{V}|$, where $|\mathcal{V}|$ is the size of the dictionary. We set $x_i = 1$ in the vector if the i^{th} word of the dictionary is present in the text; else, we set it to zero. The naïve part of the Naïve Bayes title refers to the modeling assumption that the different x_i 's are independent given $y \in \{0,1\}$. The model parameters are

- $y \sim \text{Bernoulli}(\phi_y) \leftrightarrow \phi_y = P(y = 1)$,
- $x_i|y = 0 \sim \text{Bernoulli}(\phi_{i|y=0}) \leftrightarrow \phi_{i|y=0} = P(x_i|y = 0)$, and
- $x_i|y = 1 \sim \text{Bernoulli}(\phi_{i|y=1}) \leftrightarrow \phi_{i|y=1} = P(x_i|y = 1)$.

To calibrate the model, we maximize the logarithm of the joint likelihood of training set of size m $l(\phi_y, \phi_{i|y=0}, \phi_{i|y=1}) = \prod_{i=1}^m p(\underline{x}^{(i)}, y^{(i)})$. This yields the maximum likelihood answer as

$$\begin{aligned} \phi_{j|y=1} &= \frac{\sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\}} \\ \phi_{j|y=0} &= \frac{\sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 0\}} \\ \phi_y &= \frac{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\}}{m} \end{aligned}$$

Naïve Bayes as derived above is susceptible to 0/0 errors. To avoid those, an approximation known as Laplace smoothing is applied to restate the formulae as

$$\begin{aligned} \phi_{j|y=1} &= \frac{\sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\} + 2} \\ \phi_{j|y=0} &= \frac{\sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 0\} + 2} \\ \phi_y &= \frac{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\} + 1}{m + 2} \end{aligned}$$

Other points to note are:

- Naïve Bayes is easily generalizable to the multivariate case; the model there is also called the multivariate Bernoulli event model.
- It is common to discretize continuous valued variables and apply Naïve Bayes instead of LDA and QDA.

For the specific case of text classification, a multinomial event model can also be used. A text of length n is represented by a vector $\underline{x} = (x_1, \dots, x_n)$, where $x_i = j$ if i^{th} word in the text is the j^{th} word in the dictionary V . Consequently, $x_i \in \{1, \dots, |V|\}$. The probability model is

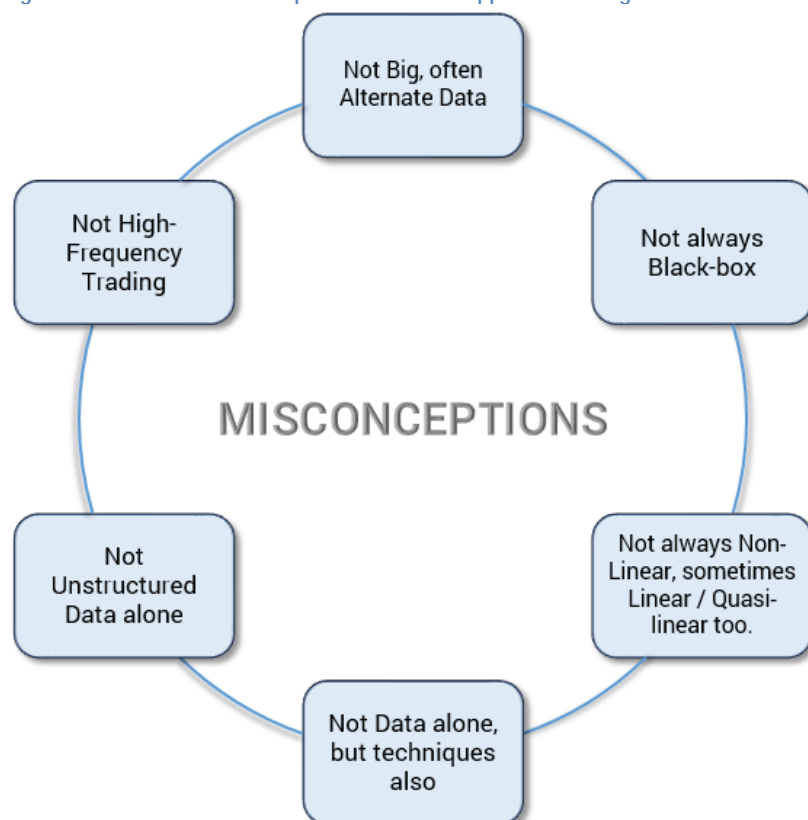
$$\begin{aligned} y &\sim \text{Bernoulli}(\phi_y) \\ \phi_{i|y=0} &= P(x_i|y=0) \\ \phi_{i|y=1} &= P(x_i|y=1) \end{aligned}$$

Further, denote each text $\underline{x}^{(i)}$ in the training sample as a vector of n_i words or $\underline{x}^{(i)} = (x_1^{(i)}, \dots, x_{n_i}^{(i)})$. Optimizing and including the Laplace smoothing term yields the answer as

$$\begin{aligned} \phi_y &= \frac{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = 1\} + 1}{m + 2} \\ \phi_{k|y=1} &= \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{1}\{x_j^{(i)} = k \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^m n_i \mathbf{1}\{y^{(i)} = 1\} + |V|} \\ \phi_{k|y=0} &= \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{1}\{x_j^{(i)} = k \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^m n_i \mathbf{1}\{y^{(i)} = 0\} + |V|} \end{aligned}$$

Common Misconceptions around Big Data in Trading

Figure 115: Common misconceptions around the application of Big Data and Machine Learning to trading



Source: J.P.Morgan Macro QDS

1. **Not Just Big, But Also Alternative:** Data sources used are often new or less known rather than just being 'Big' – size of many commercial data sets is in Gigabytes rather than Petabytes. Keeping this in mind, we designate data sources in this report as Big/Alternative instead of just Big.
2. **Not High Frequency Trading:** Machine Learning is not related to High Frequency Trading. Sophisticated techniques can be and are used on intraday data; however, as execution speed increases, our ability to use computationally heavy algorithms actually *decreases* significantly due to time constraints. On the other hand, Machine Learning can be and is profitably used on many daily data sources.
3. **Not Unstructured Alone:** Big Data is not a synonym for unstructured data. There is a substantial amount of data that is structured in tables with numeric or categorical entries. The unstructured portion is larger; but a caveat to keep in mind is that even the latest AI schemes do not pass tests corresponding to Winograd's schema. This reduces the chance that processing large text boxes (as opposed to just tweets, social messages and small/self-contained blog posts) can lead to clear market insight.
4. **Not new data alone:** While the principal advantage does arise from access to newer data sources, substantial progress has been made in computational techniques as well. This progress ranges from simple improvements like the adoption of the Bayesian paradigm to the more advanced like the re-discovery of artificial neural networks and subsequent incorporation as Deep Learning.

5. **Not always non-linear:** Many techniques are linear or quasi-linear in the parameters being estimated; later in this report, we illustrate examples of these including logistic regression (linear) and Kernelized support vector machines (quasi-linear). Many others stem from easy extensions of linear models into the non-linear domain. It is erroneous to assume that Machine Learning deals exclusively with non-linear models; though non-linear models certainly dominate much of the recent literature on the topic.
6. **Not always black box:** Some Machine Learning techniques are packaged as black-box algorithms, i.e. they use data to not only calibrate model parameters, but also to deduce the generic parametric form of the model as well to choose the input features. However, we note that Machine Learning subsumes a wide variety of models that range from the interpretable (like binary trees) to semi-interpretable (like support vector machines) to more black box (like neural nets).

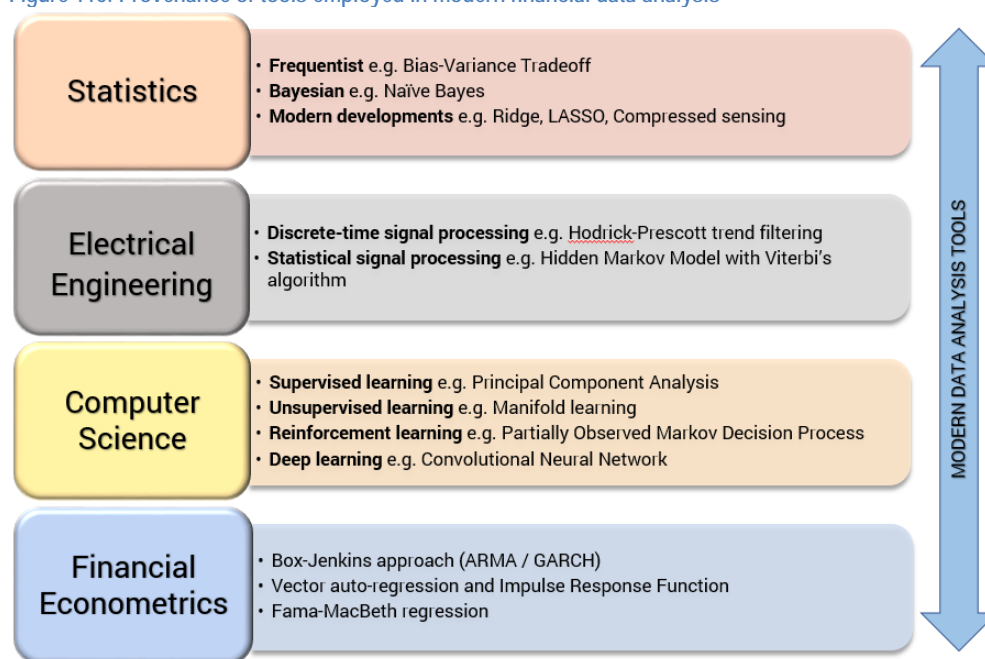
Provenance of Data Analysis Techniques

To understand Big Data analysis techniques as used in investment processes, we find it useful to track their origin and place them in one of the four following categories:

- 'Statistical Learning' from Statistics;
- 'Machine/Deep Learning' and 'Artificial Intelligence' from Computer Science;
- 'Time Series Analysis' from Econometrics; and
- 'Signal Processing' from Electrical Engineering.

This classification is useful in many data science applications, where we often have to put together tools and algorithms drawn from these diverse disciplines. We have covered Machine Learning in detail in this report. In this section, we briefly describe the other three segments.

Figure 116: Provenance of tools employed in modern financial data analysis



Source: J.P.Morgan Macro QDS

Statistical Learning from Statistics

Classical Statistics arose from need to collect representative samples from large populations. Research in statistics led to the development of rigorous analysis techniques that concentrated initially on small data sets drawn from either agriculture or industry. As data size increased, statisticians focused on the data-driven approach and computational aspects. Such numerical modeling of ever-larger data sets with the aim of detecting patterns and trends is called 'Statistical Learning'. Both the theory and toolkit of statistical learning find heavy application in modern data science applications. For example, one can use Principal Component Analysis (PCA) to uncover uncorrelated factors of variation behind any yield curve. Such analysis typically reveals that much of the movement in yield curves can be explained through just three factors: a parallel shift, a change in slope and a change in convexity. Attributing yield curve changes to PCA factors enables an analyst to isolate sectors within the yield curve that have cheapened or richened beyond what was expected from traditional weighting on the factors. This knowledge is used in both the initiation and closing of relative value opportunities.

Techniques drawn from statistics include techniques from frequentist domain, Bayesian analysis, statistical learning and compressed sensing. The simplest tools still used in practice like OLS/ANOVA and polynomial fit were borrowed from frequentists, even if posed in a Bayesian framework nowadays. Other frequentist tools used include null hypothesis testing, bootstrap estimation, distribution fitting, goodness-of-fit tests, tests for independence and homogeneity, Q-Q plot and the Kolmogorov-Smirnov test. As discussed elsewhere in the report, much analysis has moved to the Bayesian paradigm. The choice of prior family (conjugate, Zellner G, Jeffreys), estimation of hyperparameters and associated MCMC simulations draw from this literature. Even simple Bayesian techniques like Naïve Bayes with Laplace correction continue to find use in practical applications. The statistical learning literature has substantial intersection with Machine Learning research. A simple example arises from Bayesian regularization of ordinary linear regression leading to Ridge and Lasso regression models. Another example lies in the use of ensemble learning methods of bagging/boosting that enable weak learners to be combined into strong ones. Compressed sensing arose from research on sparse matrix reconstruction with applications initially on reconstruction of sub-sampled images. Viewing compressed sensing as L_1 -norm minimization leads to robust portfolio construction.

Time Series Analysis from Econometrics

Time-series Analysis refers to the analytical toolkit used by econometricians for the specific analysis of financial data. When the future evolution of an asset return depended on its past own values in a linear fashion, the return time-series was said to follow an auto-regressive (AR) process. Certain other variables could be represented as a smoothed average of noise-like terms and were called moving average (MA) processes. The Box-Jenkins approach developed in the 1970s used correlations and other statistical tests to classify and study such auto-regressive moving average (ARMA) processes. To model the observation that volatility in financial markets often occurred in bursts, new processes to model processes with time-varying volatility were introduced under the rubric of GARCH (Generalized Auto-Regressive Conditional Heteroskedastic) models. In financial economics, the technique of Impulse Response Function (IRF) is often used to discern the impact of changing one macro-economic variable (say, Fed funds rate) on other macro-economic variables (like inflation or GDP growth). In this primer, we make occasional use of these techniques in pre-processing steps before employing Machine Learning or statistical learning algorithms. However, we do not describe details of any time-series technique as they are not specific to Big Data Analysis and further, many are already well-known to traditional quantitative researchers.

Signal Processing from Electrical Engineering

Signal processing arose from attempts by electrical engineers to efficiently encode and decode speech transmissions. Signal processing techniques focused on recovering signals submersed in noise, and have been employed in quantitative investment strategies since the 1980s. By letting the beta coefficient in linear regression to evolve across time, we get the popular Kalman filter which was used widely in pairs trading strategies. The Hidden Markov Model (HMM) posited the existence of latent states evolving as a Markov chain (i.e. future evolution of the system depended only on the current state, not past states) that underlay the observed price and return behavior. Such HMMs find use in regime change models as also in high-frequency trend following strategies. Signal processing engineers analyze the frequency content of their signals and try to isolate specific frequencies through the use of frequency-selective filters. Such filters – for e.g. a low-pass filter discarding higher frequency noise components – are used as a pre-processing step before feeding the data through a Machine Learning model. In this primer, we describe only a small subset of signal processing techniques that find widespread use in the context of Big Data analysis.

One can further classify signal processing tools as arising from either discrete-time signal processing or statistical signal processing. Discrete-time signal processing dealt with design of frequency selective finite/infinite impulse response or FIR/IIR filter banks using Discrete Fourier Transform (DFT) or Z-transform techniques. Use of FFT (an efficient algorithm for DFT computation) analysis to design an appropriate Chebyshev/Butterworth filter is common. The trend-fitting Hodrick-Prescott filter tends to find more space in financial analysis than signal processing papers. Techniques for speech signal processing like Hidden Markov Model alongside the eponymous Viterbi's algorithm is used to model a latent process as a Markov chain. From Statistical signal processing, we get a variety of tools for estimation and detection. Sometimes studied under the rubric of decision theory, these include Maximum Likelihood/Maximum A-Posteriori/Maximum Mean-Square Error (ML/MAP/MMSE) estimators. Non-Bayesian estimators include von Neumann or minimax estimators. Besides the Karhunen-Loeve expansion (with an expert use illustration in digital communication literature), quants borrow

practical tools like ROC (Receiver Operating Characteristic). Theoretical results like Cramer-Rao Lower Bound provide justification for use of practical techniques through asymptotic consistency/convergence proofs. Machine Learning borrows Expectation Maximization from this literature and makes extensive use of the same to find ML parameters for complicated statistical models. Statistical signal processing is also the source for Kalman (extended/unscented) and Particle filters used in quantitative trading.

A Brief History of Big Data Analysis

While the focus on Big Data is new, the search for new and quicker information has been a permanent feature of investing. We can track this evolution through four historical anecdotes.

- a. The need for reducing latency of receiving information provided the first thrust. The story of Nathaniel Rothschild using carrier pigeons in June 1815 to learn about the outcome of the Battle of Waterloo to go long the London bourse is often cited in this aspect.
- b. The second thrust came from systematically collecting and analyzing “big” data. In the first half of the 20th century, Benjamin Graham and other investors collected accounting ratios of firms on a systematic basis, and developed the ideas of Value Investing from them.
- c. The third thrust came from locating new data that was either hard or costly to collect. Sam Walton – the founder of Walmart – used to fly in his helicopter over parking lots to evaluate his real estate investments in the early 50’s.
- d. The fourth thrust came from using technological tools to accomplish the above objectives of quickly securing hard-to-track data. In the 1980s, Marc Rich – the founder of Glencore – used binoculars to locate oil ships/tankers and relayed the gleaned insight using satellite phones.

Understanding the historical evolution as above helps explain the alternative data available today to the investment professional. Carrier pigeons have long given way to computerized networks. Data screened from accounting statements have become standardized inputs to investments; aggregators such as Bloomberg and FactSet disseminate these widely removing the need to manually collect them as was done by early value investors. Instead of flying over parking lots with a helicopter, we can procure the same data from companies like Orbital Insight that use neural networks to process imagery from low-earth orbit satellites. And finally instead of binoculars and satellite phones, we have firms like CargoMetrics that locates oil ships along maritime pathways through satellites and use such information to trade commodities and currencies.

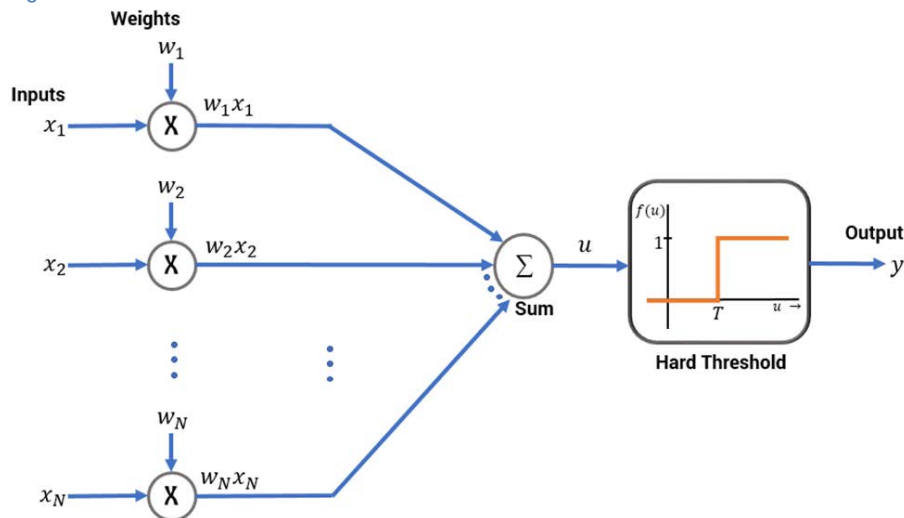
In this primer, we refer to our data sets as big/alternative data. Here, Big Data refers to large data sets, which can include financial time-series such as tick-level order book information, often marked by the three Vs of volume, velocity and variety. Alternative data refers to data – typically, but not-necessarily, non-financial – that has received lesser attention from market participants and yet has potential utility in predicting future returns for some financial assets. Alternative data stands differentiated from traditional data, by which we refer to standard financial data like daily market prices, company filings and management reports

The notion of Big Data and the conceptual toolkit of data-driven models are not new to financial economics. As early as 1920, Wesley Mitchell established the National Bureau of Economic Research to collect data on a large scale about the US economy. Using data sets collected, researchers attempted to *statistically* uncover the patterns inherent in data rather than *formulaically* deriving the theory and then fitting the data to it. This statistical, a-theoretical approach using novel data sets serves as a clear precursor to modern Machine Learning research on Big/Alternative data sets. In 1930, such statistical analysis led to the claim of wave pattern in macroeconomic data by Simon Kuznets, who was awarded the Nobel Memorial Prize in Economic Sciences (hereafter, ‘Economics Nobel’) in 1971. Similar claims of economic waves through statistical analysis were made later by Kitchin, Juglar and Kondratiev. The same era also saw the dismissal of both a-theoretical/statistical and theoretical/mathematical model by John Maynard Keynes (a claim seconded by Hayek later), who saw social phenomena as being incompatible with strict formulation via either mathematical theorization or statistical formulation. Yet, ironically, it was Keynesian models that led to the next round of large-scale data collection (growing up to hundreds of thousands of prices and quantities across time) and analysis (up to hundreds of thousands of equations). The first Economics Nobel was awarded precisely for the application of Big Data to Jan Tinbergen (shared with fellow econometrician Ragnar Frisch) for his comprehensive national model for Netherlands, United Kingdom and the United States. Lawrence Klein (Economics Nobel, 1980) formulated the first global large-scale macroeconomic model; the LINK project spun off from his work at Wharton continues to be used till date for forecasting purposes. The most influential critique of such models – based on past correlations, rather than formal theory – was made by Robert Lucas (Economics Nobel, 1995), who argued for reestablishment of theory to account for evolution in empirical correlations triggered through policy changes. Even the Bayesian paradigm, through which a researcher can systematically update his/her prior beliefs based on streaming evidence, was formulated in an influential article by Chris Sims (Economics Nobel, 2011) [Sims(1980)].

Apart from employment of new, large data sets, econometricians have also advanced the modern data analysis toolkit in a significant manner. Recognizing the need to account for auto-correlations in predictor as well as predicted variables, the Box-Jenkins approach was pioneered in the 1970s. Further, statistical properties of financial time-series tend to evolve with time. To account for such time-varying variance (termed ‘heteroskedasticity’) and fat tails in asset returns, new models such as ARCH (invented in Engle (1982), winning Robert Engle the Economics Nobel in 2003) and GARCH were developed; and these continue to be widely used by investment practitioners.

A similar historical line of ups and downs can be traced in the computer science community for the development of modern Deep Learning; academic historical overviews are present in Bengio (2009), LeCun et al. (2015) and Schmidhuber (2015). An early paper in 1949 by the Canadian neuro-psychologist Donald Hebb – see Hebb (1949) - related learning within the human brain to the formation of synapses (think, linking mechanism) between neurons (think, basic computing unit). A simple calculating model for a neuron was suggested in 1945 by McCulloch and Pitts – see McCulloch-Pitts (1945) – which could compute a weighted average of the input, and then returned one, if the average was above a threshold and zero, otherwise.

Figure 117: The standard McCulloch-Pitts model of neuron



Source: J.P.Morgan Macro QDS

In 1958, the psychologist Franklin Rosenblatt built the first modern neural network model called the Perceptron and showed that the weights in the McCulloch-Pitts model could be calibrated using the available data set; in essence, he had invented what we now call a *learning algorithm*. The perceptron model was designed for image recognition purposes and implemented in hardware, thus serving as a precursor to modern GPU units used in image signal processing. The learning rule was further refined through the work in Widrow-Hoff (1960), which calibrated the parameters by minimizing the difference between the actual pre-known output and the reconstructed one. Even today, Rosenblatt’s perceptron and the Widrow-Hoff rule continue to find place in the Machine Learning curriculum. These results spurred the first wave of excitement about Artificial Intelligence that ended abruptly in 1969, when the influential MIT theorist Marvin Minsky wrote a scathing critique in his book titled “Perceptrons” [Minsky-Papert (1960)]. Minsky pointed that perceptrons as defined by Rosenblatt can never replicate a simple structure like a XOR function, that is defined as $1 \oplus 1 = 0 \oplus 0 = 0$ and $1 \oplus 0 = 0 \oplus 1 = 1$. This critique ushered in, what is now called, the first *AI Winter*.

The first breakthroughs happened in the 1970s [Werbos (1974), an aptly titled PhD thesis of “Beyond regression: New tools for prediction and analysis...”], though they gained popularity only in the 1980s [Rumelhart et al (1986)]. The older neural models had a simple weighted average followed by a piece-wise linear thresholding function. Newer models began to have multiple layers of neurons interconnected to each other, and further replaced the simple threshold function (which returned

one if more than threshold and zero otherwise) with a non-linear, smooth function (now, called an *activation function*). The intermediate layers of neurons hidden between the input and the output layer of neurons served to uncover new features from data. These models, which could theoretically implement any function including the XOR⁶⁸, used regular high-school calculus to calibrate the parameters (viz., weights on links between neurons); the technique itself is now called backpropagation. Readers familiar with numerical analysis can think of backpropagation (using, ‘gradient descent’) as an extension to the simple Newton’s algorithm for iteratively solving equations. Variants of gradient descent remain a workhorse till today for training neural networks. The first practical application of neural networks to massive data sets arose in 1989, when researchers at AT&T Bell Labs used data from the US Postal Service to decipher hand-written zip code information; see LeCun et al (1989).

The second AI winter arose more gradually in the early 1990s. Calibrating weights of interconnections in a multi-layer neural network was not only time-consuming, it was found to be error-prone as the number of hidden layers increased [Schmidhuber (2015)]. Meanwhile, competing techniques from outside the neural network community started to make their impression (as reported in LeCun (1995)); in this report, we shall later survey two of the most prominent of those, namely Support Vector Machines and Random Forests. These techniques quickly eclipsed neural networks, and as funding declined rapidly, active research continued only in select groups in Canada and United States.

The second AI winter ended in 2006 when Geoffrey Hinton’s research group at the University of Toronto demonstrated that a multi-layer neural network could be efficiently trained using a strategy greedy, layer-wise pre-training [Hinton et al (2006)]. While Hinton’s original analysis focused on a specific type of neural network called the Deep Belief Network, other researchers could quickly extend it to many other types of multi-layer neural networks. This launched a new renaissance in Machine Learning that continues till date and is profiled in detail in this primer.

⁶⁸ For the universality claim, see Hornik et al (1989).

References

- Abayomi, K., Gelman, A., and Levy, M. (2008), “Diagnostics for multivariate imputations”, *Applied Statistics* 57, 273–291.
- Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I.(1995), “Fast discovery of association rules”, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Cambridge, MA.
- Agresti, A. (2002), “Categorical Data Analysis”, second edition, New York: Wiley.
- Akaike, H. (1973), “Information theory and an extension of the maximum likelihood principle”, *Second International Symposium on Information Theory*, 267–281.
- Amit, Y. and Geman, D. (1997), “Shape quantization and recognition with randomized trees”, *Neural Computation* 9: 1545–1588.
- Anderson, T. (2003), “An Introduction to Multivariate Statistical Analysis”, 3rd ed., Wiley, New York.
- Andrieu, C., and Robert, C. (2001), “Controlled MCMC for optimal sampling”, Technical report, Department of Mathematics, University of Bristol.
- Andrieu, C., and Thoms, J. (2008), “A tutorial on adaptive MCMC”, *Statistics and Computing* 18, 343–373.
- Ba, J., Mnih, V., & Kavukcuoglu, K. (2014), “Multiple object recognition with visual attention”, arXiv preprint arXiv:1412.7755.
- Babb, Tim, “How a Kalman filter works, in pictures”, Available at [link](#).
- Banerjee, A., Dunson, D. B., and Tokdar, S. (2011), “Efficient Gaussian process regression for large data sets”, Available at [link](#).
- Barbieri, M. M., and Berger, J. O. (2004), “Optimal predictive model selection”, *Annals of Statistics* 32, 870–897.
- Barnard, J., McCulloch, R. E., and Meng, X. L. (2000), “Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage”. *Statistica Sinica* 10, 1281–1311.
- Bartlett, P. and Traskin, M. (2007), “Adaboost is consistent, in B. Schölkopf”, J. Platt and T. Hoffman (eds), *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA, 105–112.
- Bell, A. and Sejnowski, T. (1995), “An information-maximization approach to blind separation and blind deconvolution”, *Neural Computation* 7: 1129–1159.
- Bengio, Y (2009), “Learning deep architectures for AI”, *Foundations and Trends in Machine Learning*, Vol 2:1.
- Bengio, Y., Courville, A., & Vincent, P. (2013), “Representation learning: A review and new perspectives”, *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828
- Bengio, Y., Goodfellow, I. J., & Courville, A. (2015), “Deep Learning”. *Nature*, 521, 436–444.
- Berry, S., M., Carlin, B. P., Lee, J. J., and Muller, P. (2010), “Bayesian Adaptive Methods for Clinical Trials”, London: Chapman & Hall.
- Betancourt, M. J. (2013), “Generalizing the no-U-turn sampler to Riemannian manifolds”, Available at [link](#).

- Betancourt, M. J., and Stein, L. C. (2011), “The geometry of Hamiltonian Monte Carlo”, Available at [link](#).
- Bigelow, J. L., and Dunson, D. B. (2009), “Bayesian semiparametric joint models for functional predictors”, *Journal of the American Statistical Association* 104, 26–36.
- Biller, C. (2000), “Adaptive Bayesian regression splines in semiparametric generalized linear models”, *Journal of Computational and Graphical Statistics* 9, 122–140.
- Bilmes, Jeff (1998), “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, Available at [link](#).
- Bishop, C. (1995), “Neural Networks for Pattern Recognition”, Clarendon Press, Oxford.
- Bishop, C. (2006), “Pattern Recognition and Machine Learning”, Springer, New York.
- Blei, D., Ng, A., and Jordan, M. (2003), “Latent Dirichlet allocation”, *Journal of Machine Learning Research* 3, 993–1022.
- Bollerslev, T (1986), “Generalized autoregressive conditional heteroskedasticity”, *Journal of econometrics*, Vol 31 (3), 307–327.
- Bradlow, E. T., and Fader, P. S. (2001), “A Bayesian lifetime model for the “Hot 100” Billboard songs”, *Journal of the American Statistical Association* 96, 368–381.
- Breiman, L. (1992), “The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error”, *Journal of the American Statistical Association* 87: 738–754.
- Breiman, L. (1996a), “Bagging predictors”, *Machine Learning* 26: 123–140.
- Breiman, L. (1996b), “Stacked regressions”, *Machine Learning* 24: 51–64.
- Breiman, L. (1998), “Arcing classifiers (with discussion)”, *Annals of Statistics* 26: 801–849.
- Breiman, L. (1999), “Prediction games and arcing algorithms”, *Neural Computation* 11(7): 1493–1517.
- Breiman, L. (2001), “Random Forests”, *Journal of Machine Learning*, Vol 45(1), 5–32. Available at [link](#).
- Breiman, L. and Spector, P. (1992), “Submodel selection and evaluation in regression: the X-random case”, *International Statistical Review* 60: 291–319.
- Brooks, S. P., Giudici, P., and Roberts, G. O. (2003), “Efficient construction of reversible jump MCMC proposal distributions (with discussion)”, *Journal of the Royal Statistical Society B* 65, 3–55.
- Bruce, A. and Gao, H. (1996), “Applied Wavelet Analysis with S-PLUS”, Springer, New York.
- Bühlmann, P. and Hothorn, T. (2007), “Boosting algorithms: regularization, prediction and model fitting (with discussion)”, *Statistical Science* 22(4): 477–505.
- Burges, C. (1998), “A tutorial on support vector machines for pattern recognition”, *Knowledge Discovery and Data Mining* 2(2): 121–167.

- Carvalho, C. M., Lopes, H. F., Polson, N. G., and Taddy, M. A. (2010), “Particle learning for general mixtures”, *Bayesian Analysis* 5, 709–740.
- Chen, S. S., Donoho, D. and Saunders, M. (1998), “Atomic decomposition by basis pursuit”, *SIAM Journal on Scientific Computing* 20(1): 33–61.
- Chen, Z (2003), “Bayesian filtering: From Kalman filters to particle filters”, Tech. rep., and beyond. Technical report, Adaptive Systems Lab, McMaster University.
- Cherkassky, V. and Mulier, F. (2007), “Learning from Data (2nd Edition)”, Wiley, New York.
- Chib, S et al. (2002), “Markov chain Monte Carlo methods for stochastic volatility models”, *Journal of Econometrics* 108(2):281–316.
- Chipman, H., George, E. I., and McCulloch, R. E. (1998), “Bayesian CART model search (with discussion)”, *Journal of the American Statistical Association* 93, 935–960.
- Chui, C. (1992), “An Introduction to Wavelets”, Academic Press, London.
- Clemen, R. T. (1996), “Making Hard Decisions”, second edition. Belmont, Calif.: Duxbury Press.
- Clyde, M., DeSimone, H., and Parmigiani, G. (1996), “Prediction via orthogonalized model mixing”, *Journal of the American Statistical Association* 91, 1197–1208.
- Comon, P. (1994), “Independent component analysis—a new concept?”, *Signal Processing* 36: 287–314.
- Cook, S., Gelman, A., and Rubin, D. B. (2006), “Validation of software for Bayesian models using posterior quantiles”, *Journal of Computational and Graphical Statistics* 15, 675–692.
- Cox, D. and Wermuth, N. (1996), “Multivariate Dependencies: Models, Analysis and Interpretation”, Chapman and Hall, London.
- Cseke, B., and Heskes, T. (2011), “Approximate marginals in latent Gaussian models”, *Journal of Machine Learning Research* 12, 417–454.
- Daniels, M. J., and Kass, R. E. (1999), “Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models”, *Journal of the American Statistical Association* 94, 1254–1263.
- Daniels, M. J., and Kass, R. E. (2001), “Shrinkage estimators for covariance matrices”, *Biometrics* 57, 1173–1184.
- Dasarathy, B. (1991), “Nearest Neighbor Pattern Classification Techniques”, IEEE Computer Society Press, Los Alamitos, CA.
- Daubechies, I. (1992), “Ten Lectures in Wavelets”, Society for Industrial and Applied Mathematics, Philadelphia, PA.
- Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), “Bayesian Methods for Nonlinear Classification and Regression”, New York: Wiley.
- Dietterich, T. (2000a), “Ensemble methods in machine learning”, *Lecture Notes in Computer Science* 1857: 1–15.
- Dietterich, T. (2000b), “An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization”, *Machine Learning* 40(2): 139–157.

- DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001), “Bayesian curve-fitting with free-knot splines”, *Biometrika* 88, 1055–1071.
- Dobra, A., Tebaldi, C., and West, M. (2003), “Bayesian inference for incomplete multi-way tables”, Technical report, Institute of Statistics and Decision Sciences, Duke University.
- Donoho, D. and Johnstone, I. (1994), “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika* 81: 425–455.
- Douc, R & Cappé, O (2005), “Comparison of resampling schemes for particle filtering”, In *Image and Signal Processing and Analysis*, 2005. ISPA 2005.
- Doucet, A & Johansen, A (2008), “A tutorial on particle filtering and smoothing: Fifteen years later”.
- Duda, R., Hart, P. and Stork, D. (2000), “Pattern Classification” (2nd Edition), Wiley, New York.
- Dunson, D. B. (2005), “Bayesian semiparametric isotonic regression for count data”, *Journal of the American Statistical Association* 100, 618–627.
- Dunson, D. B. (2009), “Bayesian nonparametric hierarchical modeling”, *Biometrical Journal* 51, 273–284.
- Dunson, D. B. (2010a), “Flexible Bayes regression of epidemiologic data”, In *Oxford Handbook of Applied Bayesian Analysis*, ed. A. O’Hagan and M. West. Oxford University Press.
- Dunson, D. B. (2010b), “Nonparametric Bayes applications to biostatistics”, In *Bayesian Non-parametrics*, ed. N. L. Hjort, C. Holmes, P. Muller, and S. G. Walker. Cambridge University Press.
- Dunson, D. B., and Bhattacharya, A. (2010), “Nonparametric Bayes regression and classification through mixtures of product kernels”, In *Bayesian Statistics 9*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 145–164. Oxford University Press.
- Dunson, D. B., and Taylor, J. A. (2005), “Approximate Bayesian inference for quantiles”, *Journal of Nonparametric Statistics* 17, 385–400.
- Edwards, D. (2000), “Introduction to Graphical Modelling”, 2nd Edition, Springer, New York.
- Efron, B. and Tibshirani, R. (1993), “An Introduction to the Bootstrap”, Chapman and Hall, London.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), “Least angle regression (with discussion)”, *Annals of Statistics* 32(2): 407–499.
- Ekster, G (2014), “Finding and using unique datasets by hedge funds”, Hedge Week Article published on 3/11/2014.
- Ekster, G (2015), “Driving investment process with alternative data”, White Paper by Integrity Research.
- Elliott, R.J. and Van Der Hoek, J. and Malcolm, W.P. (2005) “Pairs trading”, *Quantitative Finance*, 5(3), 271-276. Available at [link](#).
- Engle, R (1982), “Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation”, *Econometrica*, Vol 50 (4), 987-1008.
- Evgeniou, T., Pontil, M. and Poggio, T. (2000), “Regularization networks and support vector machines”, *Advances in Computational Mathematics* 13(1): 1–50.
- Fan, J. and Gijbels, I. (1996), “Local Polynomial Modelling and Its Applications”, Chapman and Hall, London.

- Faragher, R (2012), “Understanding the Basis of the Kalman Filter via a Simple and Intuitive Derivation”.
- Fill, J. A. (1998), “An interruptible algorithm for perfect sampling”. *Annals of Applied Probability* 8, 131–162.
- Flury, B. (1990), “Principal points”, *Biometrika* 77: 33–41.
- Fraley, C., and Raftery, A. E. (2002), “Model-based clustering, discriminant analysis, and density estimation”, *Journal of the American Statistical Association* 97, 611–631.
- Frank, I. and Friedman, J. (1993), “A statistical view of some chemometrics regression tools (with discussion)”, *Technometrics* 35(2): 109–148.
- Freund, Y. (1995), “Boosting a weak learning algorithm by majority”, *Information and Computation* 121(2): 256–285.
- Freund, Y. and Schapire, R. (1996b), “Game theory, on-line prediction and boosting”, *Proceedings of the Ninth Annual Conference on Computational Learning Theory, Desenzano del Garda, Italy*, 325–332.
- Friedman, J. (1994b), “An overview of predictive learning and function approximation”, in V. Cherkassky, J. Friedman and H. Wechsler (eds), *From Statistics to Neural Networks*, Vol. 136 of NATO ISI Series F, Springer, New York.
- Friedman, J. (1999), “Stochastic gradient boosting”, Technical report, Stanford University.
- Friedman, J. (2001), “Greedy function approximation: A gradient boosting machine”, *Annals of Statistics* 29(5): 1189–1232.
- Friedman, J. and Hall, P. (2007), “On bagging and nonlinear estimation”, *Journal of Statistical Planning and Inference* 137: 669–683.
- Friedman, J. and Popescu, B. (2008), “Predictive learning via rule ensembles”, *Annals of Applied Statistics*, to appear.
- Friedman, J., Hastie, T. and Tibshirani, R. (2000), “Additive logistic regression: a statistical view of boosting (with discussion)”, *Annals of Statistics* 28: 337–307.
- Gelfand, A. and Smith, A. (1990), “Sampling based approaches to calculating marginal densities”, *Journal of the American Statistical Association* 85: 398–409.
- Gelman, A. (2005), “Analysis of variance: why it is more important than ever (with discussion)”, *Annals of Statistics* 33, 1–53.
- Gelman, A. (2006b), “The boxer, the wrestler, and the coin flip: a paradox of robust Bayesian inference and belief functions”, *American Statistician* 60, 146–150.
- Gelman, A. (2007a), “Struggles with survey weighting and regression modeling (with discussion)”, *Statistical Science* 22, 153–188.
- Gelman, A. (2007b), “Discussion of ‘Bayesian checking of the second levels of hierarchical models’”, by M. J. Bayarri and M. E. Castellanos. *Statistical Science* 22, 349–352.
- Gelman, A., and Hill, J. (2007), “Data Analysis Using Regression and Multilevel/Hierarchical Models”, Cambridge University Press.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995), “Bayesian Data Analysis”, CRC Press, Boca Raton, FL.

- Gelman, A., Chew, G. L., and Shnaidman, M. (2004), “Bayesian analysis of serial dilution assays”, *Biometrics* 60, 407–417.
- Gelman, A; Carlin, J. B; Stern, H. S; Dunson, D. B; Vehtari, A and Rubin, D. B., “Bayesian Data Analysis”, CRC Press.
- Gentle, J. E. (2003), “Random Number Generation and Monte Carlo Methods”, second edition. New York: Springer.
- George, E. I., and McCulloch, R. E. (1993), “Variable selection via Gibbs sampling”, *Journal of the American Statistical Association* 88, 881–889.
- Gershman, S. J., Hoffman, M. D., and Blei, D. M. (2012), “Nonparametric variational inference”, In *proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.
- Gersho, A. and Gray, R. (1992), “Vector Quantization and Signal Compression”, Kluwer Academic Publishers, Boston, MA.
- Geweke, J (1989), “Bayesian inference in econometric models using Monte Carlo integration”, *Econometrica: Journal of the Econometric Society*, 1317–1339.
- Gilovich, T., Griffin, D., and Kahneman, D. (2002), “Heuristics and Biases: The Psychology of Intuitive Judgment”, Cambridge University Press.
- Girolami, M., and Calderhead, B. (2011), “Riemann manifold Langevin and Hamiltonian Monte Carlo methods (with discussion)”, *Journal of the Royal Statistical Society B* 73, 123–214.
- Girosi, F., Jones, M. and Poggio, T. (1995), “Regularization theory and neural network architectures”, *Neural Computation* 7: 219–269.
- Gneiting, T. (2011), “Making and evaluating point forecasts”, *Journal of the American Statistical Association* 106, 746–762.
- Gordon, A. (1999), “Classification (2nd edition)”, Chapman and Hall/CRC Press, London.
- Gordon, N et al. (1993), “Novel approach to nonlinear/non-Gaussian Bayesian state estimation”, In *Radar and Signal Processing*, IEE Proceedings F, vol. 140, 107–113. IET.
- Graves, A. (2013), “Generating sequences with recurrent neural networks”, arXiv preprint arXiv:1308.0850.
- Graves, A., & Jaitly, N. (2014), “Towards End-To-End Speech Recognition with Recurrent Neural Networks”, In *ICML* (Vol. 14, 1764-1772).
- Green, P. and Silverman, B. (1994), “Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach”, Chapman and Hall, London.
- Green, P. J. (1995), “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”, *Biometrika* 82, 711–732.
- Greenland, S. (2005), “Multiple-bias modelling for analysis of observational data”, *Journal of the Royal Statistical Society A* 168, 267–306.
- Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015), “DRAW: A recurrent neural network for image generation”, arXiv preprint arXiv:1502.04623.
- Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., eds. (2002), “Survey Nonresponse”, New York: Wiley.
- Hall, P. (1992), “The Bootstrap and Edgeworth Expansion”, Springer, New York.

- Hanif, A & Smith, R (2012), “Generation Based Path-Switching in Sequential Monte Carlo Methods”, IEEE Congress on Evolutionary Computation (CEC), 2012 , pages 1–7. IEEE.
- Hanif, A & Smith, R (2013), “Stochastic Volatility Modeling with Computational Intelligence Particle Filters”, Genetic and Evolutionary Computation Conference (GECCO), ACM.
- Hanif, A (2013), “Computational Intelligence Sequential Monte Carlos for Recursive Bayesian Estimation”, PhD Thesis, Intelligent Systems Group, UCL.
- Hannah, L., and Dunson, D. B. (2011), “Bayesian nonparametric multivariate convex regression”, Available at [link](#).
- Hastie, T. (1984), “Principal Curves and Surfaces”, PhD thesis, Stanford University.
- Hastie, T. and Stuetzle, W. (1989), “Principal curves”, Journal of the American Statistical Association 84(406): 502–516.
- Hastie, T. and Tibshirani, R. (1990), “Generalized Additive Models”, Chapman and Hall, London.
- Hastie, T. and Tibshirani, R. (1996a), “Discriminant adaptive nearest neighbor classification”, IEEE Pattern Recognition and Machine Intelligence 18: 607–616.
- Hastie, T. and Tibshirani, R. (1996b), “Discriminant analysis by Gaussian mixtures”, Journal of the Royal Statistical Society Series B. 58: 155–176.
- Hastie, T. and Tibshirani, R. (1998), “Classification by pairwise coupling”, Annals of Statistics 26(2): 451–471.
- Hastie, T., Buja, A. and Tibshirani, R. (1995), “Penalized discriminant analysis”, Annals of Statistics 23: 73–102.
- Hastie, T., Taylor, J., Tibshirani, R. and Walther, G. (2007), “Forward stagewise regression and the monotone lasso”, Electronic Journal of Statistics 1: 1–29.
- Hastie, T., Tibshirani, R. and Buja, A. (1994), “Flexible discriminant analysis by optimal scoring”, Journal of the American Statistical Association 89: 1255–1270.
- Hastie, T; Tibshirani, R and Friedman, J (2013), “The elements of statistical learning”, 2nd edition, Springer. Available at [link](#).
- Hazelton, M. L., and Turlach, B. A. (2011), “Semiparametric regression with shape-constrained penalized splines”, Computational Statistics and Data Analysis 55, 2871–2879.
- Hebb, D. O (1949), “The organization of behavior: a neuropsychological theory”, Wiley and sons, New York.
- Heskes, T., Opper, M., Wiegerinck, W., Winther, O., and Zoeter, O. (2005), “Approximate inference techniques with expectation constraints”, Journal of Statistical Mechanics: Theory and Experiment, P11015.
- Hinton, GE and Salakhutdinov, RR (2006), “Reducing the dimensionality of data with neural networks”, Science 313 (5786), 504-507.
- Hinton, GE; Osindero, S and Teh, Y-W (2006), “A fast learning algorithm for deep belief nets”, Neural Computation.
- Ho, T. K. (1995), “Random decision forests”, in M. Kavavaugh and P. Storms (eds), Proc. Third International Conference on Document Analysis and Recognition, Vol. 1, IEEE Computer Society Press, New York, 278–282.

- Hodges, J. S., and Sargent, D. J. (2001), "Counting degrees of freedom in hierarchical and other richly parameterized models", *Biometrika* 88, 367–379.
- Hoerl, A. E. and Kennard, R. (1970), "Ridge regression: biased estimation for nonorthogonal problems", *Technometrics* 12: 55–67.
- Hoff, P. D. (2007), "Extending the rank likelihood for semiparametric copula estimation", *Annals of Applied Statistics* 1, 265–283.
- Hornik, K; Stinchcombe, M; White, H, "Multilayer feedforward networks are universal approximators", *Neural Networks*, Vol 2 (5), 359-366.
- Hubert, L and Arabie, P (1985), "Comparing partitions", *Journal of Classification*.
- Hyvärinen, A. and Oja, E. (2000), "Independent component analysis: algorithms and applications", *Neural Networks* 13: 411–430.
- Imai, K., and van Dyk, D. A. (2005), "A Bayesian analysis of the multinomial probit model using marginal data augmentation", *Journal of Econometrics*. 124, 311–334.
- Ionides, E. L. (2008), "Truncated importance sampling", *Journal of Computational and Graphical Statistics*, 17(2), 295-311.
- Ishwaran, H., and Zarepour, M. (2002), "Dirichlet prior sieves in finite normal mixtures", *Statistica Sinica* 12, 941–963.
- Jaakkola, T. S., and Jordan, M. I. (2000), "Bayesian parameter estimation via variational methods", *Statistics and Computing* 10, 25–37.
- Jackman, S. (2001), "Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference and model checking", *Political Analysis* 9, 227–241.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), "An Introduction to Statistical Learning", *Springer Texts in Statistics*.
- Jasra, A., Holmes, C. C., and Stephens, D. A. (2005), "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling", *Statistical Science* 20, 50–67.
- Jiang, W. (2004), "Process consistency for Adaboost", *Annals of Statistics* 32(1): 13–29.
- Jin, Y & Branke, J (2005), "Evolutionary optimization in uncertain environments-a survey", *Evolutionary Computation*, *IEEE Transactions on* 9(3):303–317.
- Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999), "Introduction to variational methods for graphical models", *Machine Learning* 37, 183–233.
- Kadanoff, L. P (1966), "Scaling laws for Ising models near T_c ", *Physics* 2, 263.
- Kalman, R.E. (1960), "A New Approach to Linear Filtering and Prediction Problems", *J. Basic Eng* 82(1), 35-45.
- Karpathy, A. (2015), "The unreasonable effectiveness of recurrent neural networks", Andrej Karpathy [blog](#).
- Kaufman, L. and Rousseeuw, P. (1990), "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley, New York.

- Kearns, M. and Vazirani, U. (1994), "An Introduction to Computational Learning Theory", MIT Press, Cambridge, MA.
- Kitchin, Rob (2015), "Big Data and Official Statistics: Opportunities, Challenges and Risks", Statistical Journal of IAOS 31, 471-481.
- Kittler, J., Hatef, M., Duin, R. and Matas, J. (1998), "On combining classifiers", IEEE Transaction on Pattern Analysis and Machine Intelligence 20(3): 226-239.
- Kleinberg, E. M. (1996), "An overtraining-resistant stochastic modeling method for pattern recognition", Annals of Statistics 24: 2319-2349.
- Kleinberg, E.M. (1990), "Stochastic discrimination", Annals of Mathematical Artificial Intelligence 1: 207-239.
- Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection", International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, 1137-1143.
- Kohonen, T. (1989), "Self-Organization and Associative Memory (3rd edition)", Springer, Berlin.
- Kohonen, T. (1990), "The self-organizing map", Proceedings of the IEEE 78: 1464-1479.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, A. and Saarela, A. (2000), "Self-organization of a massive document collection", IEEE Transactions on Neural Networks 11(3): 574-585. Special Issue on Neural Networks for Data Mining and Knowledge Discovery.
- Koller, D. and Friedman, N. (2007), "Structured Probabilistic Models", Stanford Bookstore Custom Publishing. (Unpublished Draft).
- Krishnamachari, R. T (2015), "MIMO Systems under Limited Feedback: A Signal Processing Perspective ", LAP Publishing.
- Krishnamachari, R. T and Varanasi, M. K. (2014), "MIMO Systems with quantized covariance feedback", IEEE Transactions on Signal Processing, 62(2), Pg 485-495.
- Krishnamachari, R. T and Varanasi, M. K. (2013a), "Interference alignment under limited feedback for MIMO interference channels", IEEE Transactions on Signal Processing, 61(15), Pg. 3908-3917.
- Krishnamachari, R. T and Varanasi, M. K. (2013b), "On the geometry and quantization of manifolds of positive semi-definite matrices", IEEE Transactions on Signal Processing, 61 (18), Pg 4587-4599.
- Krishnamachari, R. T and Varanasi, M. K. (2009), "Distortion-rate tradeoff of a source uniformly distributed over the composite $P_F(N)$ and the composite Stiefel manifolds", IEEE International Symposium on Information Theory.
- Krishnamachari, R. T and Varanasi, M. K. (2008a), "Distortion-rate tradeoff of a source uniformly distributed over positive semi-definite matrices", Asilomar Conference on Signals, Systems and Computers.
- Krishnamachari, R. T and Varanasi, M. K. (2008b), "Volume of geodesic balls in the complex Stiefel manifold", Allerton Conference on Communications, Control and Computing.
- Krishnamachari, R. T and Varanasi, M. K. (2008c), "Volume of geodesic balls in the real Stiefel manifold", Conference on Information Science and Systems.
- Kuhn, M. (2008), "Building Predictive Models in R Using the caret Package", Journal of Statistical Software, Vol 28(5), 1-26. Available at [link](#).

- Kurenkov, A (2015), “A ‘brief’ history of neural nets and Deep Learning”, Parts 1-4 available at [link](#).
- Laney, D (2001), “3D data management: Controlling data volume, velocity and variety”, META Group (then Gartner), File 949.
- Lauritzen, S. (1996), “Graphical Models”, Oxford University Press.
- Leblanc, M. and Tibshirani, R. (1996), “Combining estimates in regression and classification”, Journal of the American Statistical Association 91: 1641–1650.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015), “Deep Learning. Nature”, 521(7553), 436-444.
- LeCun, Y; Boser, B; Denker, J; Henderson, D; Howard, R; Hubbard, W; Jackel, L (1989), “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation , Vol 1(4), 541-551.
- LeCun, Y; Jackel, L.D; Bottou, L; Brunot, A; Cortes, C; Denker, J.S.; Drucker, H; Guyon, I; Muller, U.A; Sackinger, E; Simard, P and Vapnik, V (1995), “Comparison of learning algorithms for handwritten digit recognition”, in Fogelman, F. and Gallinari, P. (Eds), International Conference on Artificial Neural Networks, 53-60, EC2 & Cie, Paris.
- Leimkuhler, B., and Reich, S. (2004), “Simulating Hamiltonian Dynamics”,. Cambridge University Press.
- Leonard, T., and Hsu, J. S. (1992), “Bayesian inference for a covariance matrix”, Annals of Statistics 20, 1669–1696.
- Levesque, HJ; Davis, E and Morgenstern, L (2011), “The Winograd schema challenge”, The Thirteenth International Conference on Principles of Knowledge Representation and Learning.
- Little, R. J. A., and Rubin, D. B. (2002), “Statistical Analysis with Missing Data”, second edition. New York: Wiley.
- Liu, C. (2003), “Alternating subspace-spanning resampling to accelerate Markov chain Monte Carlo simulation”, Journal of the American Statistical Association 98, 110–117.
- Liu, C. (2004), “Robit regression: A simple robust alternative to logistic and probit regression. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives”, ed. A. Gelman and X. L. Meng, 227–238. New York: Wiley.
- Liu, C., and Rubin, D. B. (1995), “ML estimation of the t distribution using EM and its extensions”, ECM and ECME. Statistica Sinica 5, 19–39.
- Liu, C., Rubin, D. B., and Wu, Y. N. (1998), “Parameter expansion to accelerate EM: The PX-EM algorithm”, Biometrika 85, 755–770.
- Liu, J. (2001), “Monte Carlo Strategies in Scientific Computing”, New York: Springer
- Liu, J., and Wu, Y. N. (1999), “Parameter expansion for data augmentation”, Journal of the American Statistical Association 94, 1264–1274.
- Loader, C. (1999), “Local Regression and Likelihood”, Springer, New York.
- Lugosi, G. and Vayatis, N. (2004), “On the bayes-risk consistency of regularized boosting methods”, Annals of Statistics 32(1): 30–55.
- MacQueen, J. (1967), “Some methods for classification and analysis of multivariate observations”, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds. L.M. LeCam and J. Neyman, University of California Press, 281–297.

- Madigan, D. and Raftery, A. (1994), “Model selection and accounting for model uncertainty using Occam’s window”, *Journal of the American Statistical Association* 89: 1535–46.
- Manning, C. D (2015), “[Computational linguistics and Deep Learning](#)”, *Computational Linguistics*, Vol 41(4), 701-707, MIT Press.
- Mardia, K., Kent, J. and Bibby, J. (1979), “*Multivariate Analysis*”, Academic Press.
- Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012), “Approximate Bayesian computational methods”, *Statistics and Computing* 22, 1167–1180.
- Martin, A. D., and Quinn, K. M. (2002), “Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999”, *Political Analysis* 10, 134–153.
- Mason, L., Baxter, J., Bartlett, P. and Frean, M. (2000), “Boosting algorithms as gradient descent”, 12: 512–518.
- McCulloch, W.S and Pitts, W. H (1945), “A logical calculus of the ideas immanent in nervous activity”, *Bulletin of Mathematical Biophysics*, Vol 5, 115-133.
- Mease, D. and Wyner, A. (2008), “Evidence contrary to the statistical view of boosting (with discussion)”, *Journal of Machine Learning Research* 9: 131–156.
- Mehta, P and Schwab, D. J. (2014), “An exact mapping between the variational renormalization group and Deep Learning”, Manuscript posted on Arxiv at [link](#).
- Meir, R. and R’atsch, G. (2003), “An introduction to boosting and leveraging”, in S. Mendelson and A. Smola (eds), *Lecture notes in Computer Science, Advanced Lectures in Machine Learning*, Springer, New York.
- Meir, R. and R’atsch, G. (2003), “An introduction to boosting and leveraging”, in S. Mendelson and A. Smola (eds), *Lecture notes in Computer Science, Advanced Lectures in Machine Learning*, Springer, New York.
- Meng, X. L. (1994a), “On the rate of convergence of the ECM algorithm”, *Annals of Statistics* 22,326–339.
- Meng, X. L., and Pedlow, S. (1992), “EM: A bibliographic review with missing articles”, In *Proceedings of the American Statistical Association, Section on Statistical Computing*, 24–27.
- Meng, X. L., and Rubin, D. B. (1991), “Using EM to obtain asymptotic variance-covariance matrices:The SEM algorithm”, *Journal of the American Statistical Association* 86, 899–909.
- Meng, X. L., and Rubin, D. B. (1993), “Maximum likelihood estimation via the ECM algorithm:A general framework”, *Biometrika* 80, 267–278.
- Meng, X. L., and van Dyk, D. A. (1997), “The EM algorithm—an old folk-song sung to a fast new tune (with discussion)”, *Journal of the Royal Statistical Society B* 59, 511–567.
- Minka, T. (2001), “Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*”, ed. J. Breese and D. Koller, 362–369.
- Minsky, M and Papert, S. A (1960), “*Perceptrons*”, MIT Press (latest edition, published in 1987).
- Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013), “Bayesian Gaussian copula factor models for mixed data”, *Journal of the American Statistical Association*.
- Neal, R. (1996), “*Bayesian Learning for Neural Networks*”, Springer, New York

- Neal, R. and Hinton, G. (1998), “A view of the EM algorithm that justifies incremental, sparse, and other variants”; in *Learning in Graphical Models*, M. Jordan (ed.), Dordrecht: Kluwer Academic Publishers, Boston, MA, 355–368.
- Neal, R. M. (1994), “An improved acceptance procedure for the hybrid Monte Carlo algorithm”, *Journal of Computational Physics* 111, 194–203.
- Neal, R. M. (2011), “MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*”, ed. S. Brooks, A. Gelman, G. L. Jones, and X. L. Meng, 113–162. New York: Chapman & Hall.
- Neelon, B., and Dunson, D. B. (2004), “Bayesian isotonic regression and trend analysis”, *Biometrics* 60, 398–406.
- Nelder, J. A. (1994), “The statistics of linear models: back to basics. *Statistics and Computing*“, 4,221–234.
- O’Connell, Jared and Højsgaard, Søren (2011), “Hidden Semi Markov Models for Multiple Observation Sequences: The mhsmm Package for R”, *Journal of Statistical Software*, 39(4). Available at [link](#).
- O’Hagan, A., and Forster, J. (2004), “Bayesian Inference”, second edition. London: Arnold.
- Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007), “Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons”, *Statistics in Medicine* 26, 2088–2112.
- Ormerod, J. T., and Wand, M. P. (2012), “Gaussian variational approximate inference for generalized linear mixed models”, *Journal of Computational and Graphical Statistics* 21, 2–17.
- Osborne, M., Presnell, B. and Turlach, B. (2000a), “A new approach to variable selection in least squares problems”, *IMA Journal of Numerical Analysis* 20: 389–404.
- Osborne, M., Presnell, B. and Turlach, B. (2000b), “On the lasso and its dual, *Journal of Computational and Graphical Statistics* 9”: 319–337. Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions, *Statistics and Probability Letters* 33: 291–297.
- Papaspiliopoulos, O., and Roberts, G. O. (2008), “Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models”, *Biometrika* 95, 169–186.
- Park, M. Y. and Hastie, T. (2007), “ l_1 -regularization path algorithm for generalized linear models”, *Journal of the Royal Statistical Society Series B* 69: 659–677.
- Park, T., and Casella, G. (2008), “The Bayesian lasso”, *Journal of the American Statistical Association* 103, 681–686.
- Pati, D., and Dunson, D. B. (2011), “Bayesian closed surface fitting through tensor products”, Technical report, Department of Statistics, Duke University.
- Pearl, J. (2000), “Causality Models, Reasoning and Inference”, Cambridge University Press.
- Peltola T, Marttinen P, Vehtari A (2012), “Finite Adaptation and Multistep Moves in the Metropolis-Hastings Algorithm for Variable Selection in Genome-Wide Association Analysis”. *PLoS One* 7(11): e49445
- Petris, Giovanni, Petrone, Sonia and Campagnoli, Patrizia (2009), “Dynamic Linear Models with R”, Springer.
- Propp, J. G., and Wilson, D. B. (1996), “Exact sampling with coupled Markov chains and applications to statistical mechanics”, *Random Structures Algorithms* 9, 223–252.

- Rabiner, L.R. and Juang B.H. (1986), “An Introduction to Hidden Markov Models”, IEEE ASSp Magazine, Vol 3, Issue 1, P.4-16. Available at [link](#).
- Ramsay, J., and Silverman, B. W. (2005), “Functional Data Analysis”, second edition. New York: Springer.
- Rand, W.M (1971), “Objective criteria for the evaluation of clustering methods”, Journal of the American Statistical Association, Vol 66 (336), Pg 846-850.
- Rasmussen, C. E., and Ghahramani, Z. (2003), “Bayesian Monte Carlo”, In Advances in Neural Information Processing Systems 15, ed. S. Becker, S. Thrun, and K. Obermayer, 489–496. Cambridge, Mass.: MIT Press.
- Rasmussen, C. E., and Nickish, H. (2010), “Gaussian processes for machine learning (GPML) toolbox”, Journal of Machine Learning Research 11, 3011–3015.
- Rasmussen, C. E., and Williams, C. K. I. (2006), “Gaussian Processes for Machine Learning”, Cambridge, Mass.: MIT Press.
- Rasmussen, C. E., and Williams, C. K. I. (2006), “Gaussian Processes for Machine Learning”, Cambridge, Mass.: MIT Press.
- Ray, S., and Mallick, B. (2006), “Functional clustering by Bayesian wavelet methods”, Journal of the Royal Statistical Society B 68, 305–332.
- Regalado, A (2013), “The data made me do it”, MIT Technology Review, May Issue.
- Reilly, C., and Zeringue, A. (2004), “Improved predictions of lynx trappings using a biological model”, In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, ed. A. Gelman and X. L. Meng, 297–308. New York: Wiley.
- Richardson, S., and Green, P. J. (1997), “On Bayesian analysis of mixtures with an unknown number of components”, Journal of the Royal Statistical Society B 59, 731–792.
- Ripley, B. D. (1996), “Pattern Recognition and Neural Networks”, Cambridge University Press.
- Robert, C. P., and Casella, G. (2004), “Monte Carlo Statistical Methods”, second edition. New York: Springer.
- Roberts, G. O., and Rosenthal, J. S. (2001), “Optimal scaling for various Metropolis-Hastings algorithms”, Statistical Science 16, 351–367.
- Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2009), “Bayesian nonparametric functional data analysis through density estimation”, Biometrika 96, 149–162.
- Romeel, D. (2011), “Leapfrog integration”, Available at [link](#).
- Rosenbaum, P. R. (2010), “Observational Studies”, second edition. New York: Springer.
- Rubin, D. B. (2000), “Discussion of Dawid (2000)”, Journal of the American Statistical Association 95, 435–438.
- Rue, H. (2013), “The R-INLA project”, Available at [link](#).
- Rue, H., Martino, S., and Chopin, N. (2009), “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)”, Journal of the Royal Statistical Society B 71, 319–382.

Rumelhart, D. E.; Hinton, G. E., and Williams, R. J. (1986), “Learning representations by back-propagating errors”, *Nature*, 323, 533–536.

Schapire, R. (1990), “The strength of weak learnability”, *Machine Learning* 5(2): 197–227.

Schapire, R. (2002), “The boosting approach to machine learning: an overview”, in D. Denison, M. Hansen, C. Holmes, B. Mallick and B. Yu (eds), *MSRI workshop on Nonlinear Estimation and Classification*, Springer, New York.

Schapire, R. and Singer, Y. (1999), “Improved boosting algorithms using confidence-rated predictions”, *Machine Learning* 37(3): 297–336.

Schapire, R., Freund, Y., Bartlett, P. and Lee, W. (1998), “Boosting the margin: a new explanation for the effectiveness of voting methods”, *Annals of Statistics* 26(5): 1651–1686.

Schmidhuber, J (2015), “Deep Learning in neural networks: an overview”, *Neural Networks*, Vol 61, Pg 85-117.

Schutt, R. (2009), “Topics in model-based population inference”, Ph.D. thesis, Department of Statistics, Columbia University.

Schwarz, G. (1978), “Estimating the dimension of a model”, *Annals of Statistics* 6(2): 461–464.

Scott, D. (1992), “Multivariate Density Estimation: Theory, Practice, and Visualization”, Wiley, New York.

Seber, G. (1984), “Multivariate Observations”, Wiley, New York.

Seeger, M. W. (2008), “Bayesian inference and optimal design for the sparse linear model”, *Journal of Machine Learning Research* 9, 759–813.

Senn, S. (2013), “Seven myths of randomisation in clinical trials”, *Statistics in Medicine* 32, 1439–1450.

Shao, J. (1996), “Bootstrap model selection”, *Journal of the American Statistical Association* 91: 655–665.

Shen, W., and Ghosal, S. (2011), “Adaptive Bayesian multivariate density estimation with Dirichlet mixtures”, Available at [link](#).

Siegelmann, H. T. (1997), “Computation beyond the Turing limit”, *Neural Networks and Analog Computation*, 153-164.

Simard, P., Cun, Y. L. and Denker, J. (1993), “Efficient pattern recognition using a new transformation distance”, *Advances in Neural Information Processing Systems*, Morgan Kaufman, San Mateo, CA, 50–58.

Sims, C. A (1980), “Macroeconomics and reality”, *Econometrica*, Vol 48 (1), Pg 1-48.

Skare, O., Bolviken, E., and Holden, L. (2003), “Improved sampling-importance resampling and reduced bias importance sampling”, *Scandinavian Journal of Statistics* 30, 719–737.

Spiegelhalter, D., Best, N., Gilks, W. and Inskip, H. (1996), “Hepatitis B: a case study in MCMC methods”, in W. Gilks, S. Richardson and D. Spiegelhalter (eds), “Markov Chain Monte Carlo in Practice”, *Inter disciplinary Statistics*, Chapman and Hall, London, 21–43.

Spielman, D. A. and Teng, S.-H. (1996), “Spectral partitioning works: Planar graphs and finite element meshes”, *IEEE Symposium on Foundations of Computer Science*, 96–105.

Stephens, M. (2000a), “Bayesian analysis of mixture models with an unknown number of components: An alternative to reversible jump methods”, *Annals of Statistics* 28, 40–74.

- Stephens, M. (2000b), “Dealing with label switching in mixture models”, *Journal of the Royal Statistical Society B* 62, 795–809.
- Su, Y. S., Gelman, A., Hill, J., and Yajima, M. (2011), “Multiple imputation with diagnostics (mi) in R: Opening windows into the black box”, *Journal of Statistical Software* 45 (2).
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014), “Sequence to sequence learning with neural networks”, In *Advances in neural information processing systems* (3104-3112).
- Sutton, R. S., & Barto, A. G. (1998), “Reinforcement learning: An introduction (Vol. 1, No. 1)”, Cambridge: MIT press.
- Tarpey, T. and Flury, B. (1996), “Self-consistency: A fundamental concept in statistics”, *Statistical Science* 11: 229–243.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*”, Series B 58: 267–288.
- Tibshirani, R. and Knight, K. (1999), “Model search and inference by bootstrap bumping”, *Journal of Computational and Graphical Statistics* 8: 671–686.
- Tokdar, S. T. (2007), “Towards a faster implementation of density estimation with logistic Gaussian process priors”, *Journal of Computational and Graphical Statistics* 16, 633–655.
- Tokdar, S. T. (2011), “Adaptive convergence rates of a Dirichlet process mixture of multivariate normal”, Available at [link](#).
- United Nations (2015), “Revision and Further Development of the Classification of Big Data”, *Global Conference on Big Data for Official Statistics at Abu Dhabi*. See links [one](#) and [two](#).
- Valiant, L. G. (1984), “A theory of the learnable”, *Communications of the ACM* 27: 1134–1142.
- Van Buuren, S. (2012), “Flexible Imputation of Missing Data”, London: Chapman & Hall.
- van Dyk, D. A., and Meng, X. L. (2001), “The art of data augmentation (with discussion)”, *Journal of Computational and Graphical Statistics* 10, 1–111.
- van Dyk, D. A., Meng, X. L., and Rubin, D. B. (1995), “Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance”, *Statistica Sinica* 5, 55–75.
- Vanhatalo, J., Jylänki, P., and Vehtari, A. (2009), “Gaussian process regression with Student-t likelihood”, *advances in Neural Information Processing Systems* 22, ed. Y. Bengio et al, 1910–1918.
- Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013b), “GPstuff: Bayesian modeling with Gaussian processes”, *Journal of Machine Learning Research* 14, 1005–1009. Available at [link](#).
- Vapnik, V. (1996), “The Nature of Statistical Learning Theory”, Springer, New York.
- Vehtari, A., and Ojanen, J. (2012), “A survey of Bayesian predictive methods for model assessment”, *selection and comparison. Statistics Surveys* 6, 142–228.
- Vidakovic, B. (1999), “Statistical Modeling by Wavelets”, Wiley, New York.
- von Luxburg, U. (2007), “A tutorial on spectral clustering”, *Statistics and Computing* 17(4): 395–416.
- Wahba, G. (1990), “Spline Models for Observational Data”, SIAM, Philadelphia.

- Wahba, G., Lin, Y. and Zhang, H. (2000), “GACV for support vector machines”, in A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans(eds), *Advances in Large Margin Classifiers*, MIT Press, Cambridge,MA., 297–311.
- Wang, L., and Dunson, D. B. (2011a), “Fast Bayesian inference in Dirichlet process mixture models”, *Journal of Computational and Graphical Statistics* 20, 196–216.
- Wasserman, L. (2004), “All of Statistics: a Concise Course in Statistical Inference”, Springer, New York.
- Weisberg, S. (1980), “Applied Linear Regression”, Wiley, New York.
- Werbos, P (1974), “Beyond regression: New tools for prediction and analysis in the behavioral sciences”, PhD Thesis, Harvard University, Cambridge, MA.
- West, M. (2003), “Bayesian factor regression models in the “large p, small n” paradigm”, In *Bayesian Statistics 7*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A.F. M. Smith, and M. West, 733–742. Oxford University Press.
- Whittaker, J. (1990), “Graphical Models in Applied Multivariate Statistics”, Wiley, Chichester
- Wickerhauser, M. (1994), “Adapted Wavelet Analysis from Theory to Software”, A.K. Peters Ltd, Natick, MA.
- Wilmott, P (2007), “Paul Wilmott on Quantitative Finance”, 3 Volume Set. Wiley.
- Wolpert, D. (1992), “Stacked generalization”, *Neural Networks* 5: 241–259.
- Wong, F., Carter, C., and Kohn, R. (2002), “Efficient estimation of covariance selection models”, Technical report, Australian Graduate School of Management.
- Wong, W. H., and Li, B. (1992), “Laplace expansion for posterior densities of nonlinear functions of parameters”, *Biometrika* 79, 393–398.
- Yang, R., and Berger, J. O. (1994), “Estimation of a covariance matrix using reference prior”, *Annals of Statistics* 22, 1195–1211.
- Zeiler, M. D. and Fergus, R (2014), “Visualizing and understanding convolutional networks”, *Lecture Notes in Computer Science* 8689, Pg 818-833.
- Zhang, J. (2002), “Causal inference with principal stratification: Some theory and application”, Ph.D. thesis, Department of Statistics, Harvard University.
- Zhang, P. (1993), “Model selection via multifold cross-validation”, *Annals of Statistics* 21: 299–311.
- Zhang, T. and Yu, B. (2005), “Boosting with early stopping: convergence and consistency”, *Annals of Statistics* 33: 1538–1579.
- Zhao, L. H. (2000), “Bayesian aspects of some nonparametric problems”, *Annals of Statistics* 28, 532–552.

Glossary

Accuracy/Error Rate

The deviation between the accepted value and the model output expressed as a percentage of the accepted value. It is usually averaged over all the outputs.

Active Learning

This is a subset of semi-supervised Machine Learning where the learning algorithm is able to interactively get more information on the fly. It is usually used when getting the 'labels' to the data is computationally expensive so the algorithm can be more frugal by asking only for the labelling that it needs.

Alternative Data

Data not typically used by practitioners and model builders and can be sourced typically from Individuals, Business Processes and Sensors. These data sets typically have minimal aggregation and processing making them more difficult to access and use.

Anomaly Detection

This is a special form of Machine Learning where the algorithm is specifically look for outliers. These are observations that do not conform to the expected outcome.

Artificial Intelligence

This term is colloquially used to denote the 'intelligence' exhibited by machines. This 'intelligence' will take inputs to a problem; and through a series of linkages and rules, the AI will present a solution that aims to maximise its chance of successfully solving the problem. It encompasses the techniques of Big Data and Machine Learning.

Attribute

See Feature. It is also referred to as a field, or variable.

Auto-regression

This is a regression model where past values have an effect on current values. If there is only *correlation* (not causation) between past and current values it is called auto-correlation.

Back Propagation

This is a common method used to train neural networks, in combination with optimisation or gradient descent techniques. A two phase training cycle is used; 1) an input vector is run through the NN to the output, 2) a loss function is used to traverse back through the NN and apply an error value to each neuron, representing its contribution to the original output. These losses or gradients represent the weights of the neurons, which attempt to minimise the total loss function.

Bayesian Statistics

This is a branch of statistics that uses probabilities to express 'degree of belief' about the true state of world objects. It is named after Thomas Bayes (1701-1761).

Bias (Model)

A systematic difference between the model estimate and the true value of the population output (also known as Systematic Error). It arises due to erroneous assumptions in the learning algorithm (e.g. assuming the forecast model is linear when it is not). This is related to Variance.

Big Data

A term that has become mostly associated a large Volume of data, both Structured and Unstructured. However it is also commonly used to imply a high Velocity and Variety of data. The challenge is how to make sense of and create effective business strategies from this data.

Boosting

A technique in Machine Learning that aggregates an ensemble of weak classifiers into a singular strong classifier. It is often used to improve the overall accuracy of the model.

Classifier

A function or algorithm that is used to identify which category a set of observations belongs to. It is built using labelled training data containing observations where the category is known. The task at hand is known as 'classification'.

Cloud Computing

Storing and processing data using a network of remote servers (instead of using a local computer). This computing paradigm often includes technology to manage redundancy, distributed access, and parallel processing.

Clustering

This is a form of unsupervised learning in which the learning algorithm will summarize the key explanatory features of the data using iterative Knowledge Discovery. The data is unlabelled and the features are found using a process of trial and error.

Complexity (Model)

This term typically refers to the number of parameters in the model. A model is perhaps excessively complex if it has many more parameters relative to the number of observations in the training sample.

Confusion Matrix

See Error Matrix. It is called such because it makes it easy to see if the algorithm is 'confusing' two classes (or mislabelling).

Convolutional Neural Network

This is a type of Neural Network (feed-forward) which 'convolves' a sub-sampling layer over the input matrix – popular with machine vision problems.

Cost Function

This is one of the key inputs to most Machine Learning approaches and is used to calculate the cost of 'making a mistake'. The difference between the actual value and the model estimate is the 'mistake', and the cost function for example could be the square of this error term (like it is in ordinary least squares regression). This cost function is then what needs to be minimised by adjusting the model parameters.

Cross-Validation Set

A subsample of the data put aside for training validation, hyper parameters, and classifier selection. It is used *after* the training set and *before* the testing set. This is also called the 'hold-out method'

Curse of Dimensionality

This refers to the problems that arise when moving into higher dimensions that do not occur in low-dimensional settings. It can be easily seen how forecast complexity increases when moving from 2D (plane) to 3D and this continues to be the case as we move into even higher dimensions.

Decision Trees

These are a tool for supporting decisions that can be arranged in a tree-like fashion. They are typically very fragile and sensitive to the training set but have the advantage of being very transparent.

Deep Learning

This is a Machine Learning method that analyzes data in multiple layers of learning (hence 'deep'). It may start doing so by learning about simpler concepts, and combining these simpler concepts to learn about more complex concepts and abstract notions. See Neural Networks

Dependent Variable

This is the variable being forecasted and responds to the set of independent variables. In the case of simple linear regression it is the resultant Y to the input variable X.

Dummy Variable

Typically used when a Boolean input is required to the model, and will take a value of 1 or 0 to represent true or false respectively.

Error Matrix (Confusion Matrix)

This is a specific table that contains the performance results of a supervised learning algorithm. Columns represent the predicted classes while rows are the instances of the actual class.

Error Matrix		
Actual vs. Predicted	Negative	Positive
Negative	A	B
Positive	C	D

The above Error (or Confusion) matrix depicts a simple L2 case with two labels.
Accuracy: $(A+D)/(A+B+C+D)$, fraction of correctly labelled points
True Positive: $D/(C+D)$, Recall or Sensitivity rate for positive values over all actually positive points
True Negative: $A/(A+B)$, Specificity rate for negative values over all actually negative points
False Positive: $B/(A+B)$, incorrect positive labels over all negative points
False Negative: $C/(C+D)$, incorrect negative labels over all positive points

Error Surface

Used in Gradient Descent, the error surface represents the gradient at each point.

Expert System

A set of heuristics that try to capture expert knowledge (usually in the form of if-then-else statements) used to help make advice or decisions (popular in the field of medicine).

Feature

This is a measurable input property of the relationship being observed. In the context of supervised learning, a feature is an input, while a label is an output.

Feature Reduction

The process of reducing the number of input variables under consideration. One popular approach is using Principal Component Analysis to remove correlated input variables and isolate the pure and orthogonal features for inputs.

Feature Selection

The input variables selected for processing with the aim being that this subset of features should most efficiently capture the information used to define or represent what is important for analysis or classification. This can also be automated or done manually to create a sub-set of features for processing.

Forecast Error

See Accuracy/Error rate.

Gradient Decent

This is an optimisation technique that tries to find the inputs to a function that produce the minimum result (usually error) and is often used in NNs applied to the error surface. Significant trade-offs between speed and accuracy is made by altering the step-size.

Heteroscedasticity

Heteroscedasticity occurs when the variability of a variable is unequal across the range of values of a second variable, such as time in a time-series data set.

Hidden Markov Models (HMM) and Markov chain

A Markov Chain is a statistical model that can be estimated from its current state just as accurately as if one knew its full history, i.e. the current and future states are independent of past states, and the current state is visible. In a HMM the state is not visible, while the output and parameters are visible.

Independent Variable

Most often labelled the X variable, the variation in an independent variable does not depend on the changes in another variable (often labelled Y).

In-Sample Error

Can be used to test between models, the In-Sample Error measures the accuracy of a model 'in-sample' (and is usually optimistic compared to the error of the model out-of-sample).

Knowledge Discovery / Extraction

The ultimate aim of Machine Learning is to extract knowledge from data and represent it in a format that facilitates inferencing.

Linear Regression

Aims to find a simple, linear, relationship between the dependant variable (Y) and the independent variable (X), usually of the form; $Y = aX + b$. This relatively simple technique can be extended to multi-dimensional analysis.

Model Instability

Arises when small changes in the data sample (sub) set cause large changes in the model parameters. This can be caused by wrong model form, omitted variables or heteroskedastic data

Multivariate Analysis

Is concerned with the estimation of multiple variables influence over each other simultaneously and should not be confused with *multivariable* regression (which is only concerned with predictions of one dependant variable given multiple independent variables).

Natural Language Processing

NLP systems attempt to allow computers to understand human speech in either written or oral form. Initial models were rule or grammar based but couldn't cope well with unobserved words or errors (typo's). Many current methods are based on statistical models such as hidden Markov models or various Neural Nets

Neural Network

A computer modelling technique loosely based on organic neuron cells. Inputs (variables) are mapped to neurons which pass via synapses to various hidden layers before combining to the output layer. Training a neural network causes the weights of the links between neurons to change, typically over thousands of iterations. The weighted functions are typically not linear.

Logistic Regression

A modified linear regression which is commonly used as a classification technique where the dependent variable is binary (True/False) and can be extended to multiple classifications using the 'One vs Rest' scheme (A/Not A, B/Not B, etc) where predictions are probability weighted.

Loss Function

See Cost Function.

Machine Learning

This is a field of computer science with the aim of modelling data so that a computer can learn without the need for explicit programming. ML benefits from large data sets and fast processing with the aim of the system to generalise beyond the initial training data. Subsequent exposure to earlier data should ideally result in different, more accurate output.

Multi-layer Perceptron

A form of Neural Network where the inputs are often transformed by a sigmoid function and the model utilises at least 1 hidden layer (or many more for Deep Learning), where the hidden layers are structured in a fully connected directed graph.

Null Hypothesis

The null hypothesis is generally set such that there is no relationship between variables or association amongst groups. H_0 must then be disproved by appropriate statistical techniques before an alternative model is accepted.

Over fitting

This occurs when an excessively complex model is used to describe an underlying process, where the excess parameters closely map the data in-sample but reduces performance out-of-sample.

Orthogonality

A term used to describe perpendicular vectors (or planes) in multi-dimensional data. By extension it can also be used to describe non-overlapping, uncorrelated or otherwise independent data.

Perceptron

A simple Neural Network modelling a single neuron with multiple binary inputs that ‘fires’ when the weighted sum of these is greater than or equal to zero (above a fixed threshold).

Precision

True positive values divided by all predicted positive values in a confusion matrix or result set.

Principal Component Analysis (PCA)

This is a statistical technique to reduce the dimensionality of multivariate data to its principal, uncorrelated or orthogonal components. The dimensions are ordered such that the first component has the highest variance (data variability) as possible. The transformed axes are called eigenvectors and the data is represented with eigenvalues.

P-Value

The Probability-Value of a statistical ‘Null Hypothesis’ test. For example we may hypothesis there is *no* relationship between X and Y, and this model is rejected if the p-value of a linear model is $< 5\%$. Smaller p-values suggest a stronger result *against* the null hypothesis.

Random Error (Systematic Error, Measurement Error)

This is a component of Measurement Error, the other being Systematic Error (or bias). Random error is reduced by increasing sample sizes and operations such as averaging while systematic error is not.

Random Forest

This supervised learning technique uses multiple decision trees to vote on the category of a sample.

Regression

Fitting a random variable Y using explanatory variables X.

Reinforcement Learning

This Machine Learning technique is based on behavioural psychology. Software agents actions (adjustment to model coefficients) within an environment (dataset) are designed to maximise a cumulative notional reward. The key difference with other supervised learning techniques is that with reinforcement learning the correct input/output pairs are not presented.

Response Variable (Dependent Variable)

The variable that depends on other variables. It is also called the dependent variable.

Semi-Supervised Learning

A type of learning algorithm that lies between unsupervised learning (where all data are unlabelled) and supervised learning (where all data are labelled with outcome /response Y).

Symbolic AI

A branch of Artificial Intelligence (AI) research that are based on an approach that formulate the problem in a more symbolic and human-readable format

Supervised Learning

This is a category of Machine Learning in which the Training Set includes known outcomes and classifications associated with the feature inputs. The model is told a-priori of the features to use and then is concerned with only the parameterisation.

Support Vector Machine

Support vector machine is a statistical technique that looks for a hyperplane to separate different classes of data points as far as possible. It can also perform non-linear classification using the kernel trick to map inputs into a higher dimensional feature space.

Test Set

A test set is a set of data points used to assess the predictions of a statistical model

Time Series

A collection of data points that are ordered by time.

Time-Series Analysis: Long Short-Term Memory

A type of Recurrent Neural Network architecture that is suited to classification, time-series and language tasks such as those in smart-phones.

Training Set

A training set is a set of data points used to estimate the parameters of a statistical model

True/False Positive/Negative

See Error Matrix

Univariate Analysis

A type of statistical analysis that looks at the relationship between the dependent variable (or response variable) and a single predictor

Unstructured Data

Unstructured data refers to data that is not well organized with a pre-defined format. Usually they include text and multimedia contents. Due to the ambiguities, unstructured data tends to be more difficult to analyse.

Unsupervised Learning

A category of Machine Learning where the training set has no known outcome or structure. The technique is attempting to learn both the significant features as well as the parameters of the model.

Utility function

A utility function measures the preference as a function of choices. For example, the choices can be the weights allocated to different assets, and the preference can be the expected returns of the portfolio minus expected risk of the portfolio.

Validation Set

A validation set is a set of data points used to tune and select the parameters of the model. We can use the validation set to choose a final model, and test its performance using a separate test set.

Variance (Model)

This is the error a model has from small changes in the input training data. It is the main problem behind over fitting. Related to Bias.

Variance-Bias tradeoff

This is the tradeoff that applies to all supervised Machine Learning – both the Bias and the Variance need to be minimised and less of one will usually mean more of the other. If the bias or variance is too high it will confound a learning algorithm from generalizing beyond its training set.

Web Scraping

Web scraping refers to the procedure to extract data from the web. It involves fetching and downloading data from webpages, as well as parsing the contents and reformatting the data.