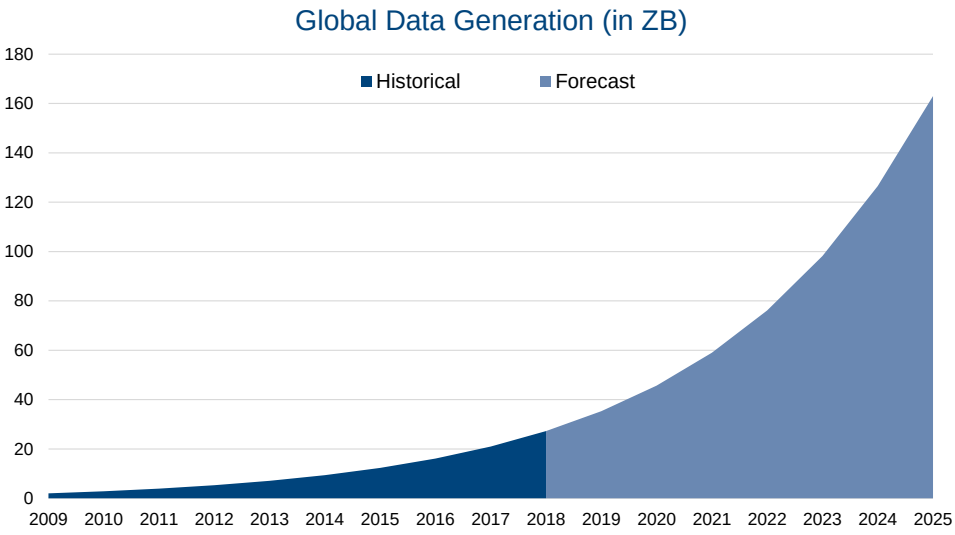
Primer: Big Data and Machine Learning

The financial markets have become increasingly efficient, making it harder to beat them using conventional methods. While we are still in the early days, some professionals are already using the wealth of available data to apply sophisticated quantitative techniques such as machine learning. These techniques will never solve the market; but, they may help money managers to generate alpha in an increasingly competitive industry.

Shane Obata, Chris Kerlow, Craig Basinger, Derek Benedet – May 2018

Big Data and the Future of Finance

We are in the midst of an exponential rise in data. IBM estimates that 90% of the world's data was created in the past two years[1]. The flood of data is coming from three primary sources: individuals, companies and sensors. Individuals are contributing with every Instagram post & Uber hail, companies are generating more transaction data than ever before and people are putting sensors into everything from washing machines to wind turbines[2]. Global data generation amounted to ~21 ZettaBytes (ZB) in 2017 and is expected to rise to more than 160 ZB by 2025, according to Seagate.



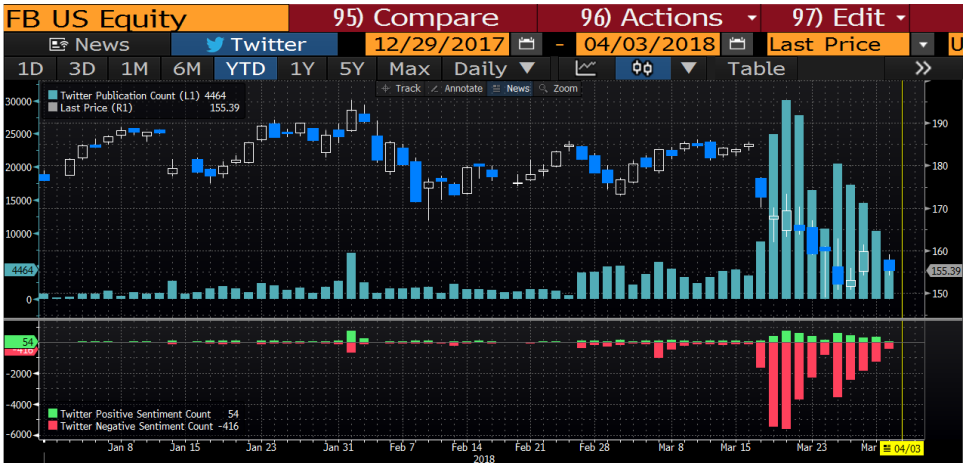
Note: 1 zettabyte (ZB) = 1 trillion gigabytes (GB) *Source: Seagate*

The financial world is also awash in data. Financial markets, economies and news provide us with a near limitless supply. Most investors have access to the same data at around the same time. This is making it harder to beat the market using conventional methods. In response, some money managers are exploring alternative data sources as they try to find an edge. For example, some investors are examining satellite imagery to assess oil rig & shipping activity and credit card transactions to evaluate sales trends. Others are analyzing social media to gauge sentiment.

[1] "Big Data and AI Strategies" – J.P. Morgan Quantitative and Derivatives Strategy – May 2017

[2] http://www.wired.co.uk/article/internet-of-things-what-is-explained-iot

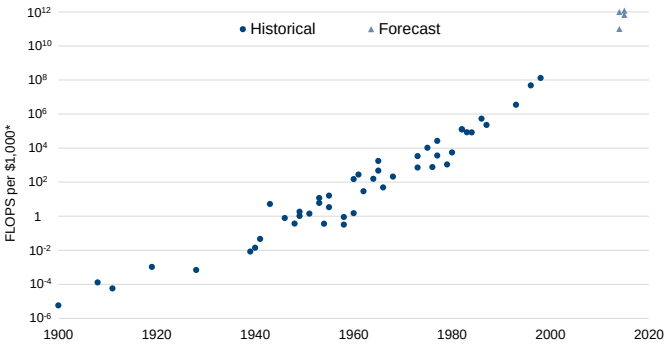
The following is a Bloomberg Twitter Activity Chart of Facebook (FB) from December 29th, 2017 to April 3rd, 2018. The top panel shows the Twitter publication count and the price while the bottom panel shows the sentiment breakdown. As we can see, there was a dramatic increase in the publication count, and more specifically negative sentiment, following the Facebook-Cambridge Analytica data scandal, which involved the collection of personal data from up to 87 million FB users.



source: Bloomberg

The big data revolution has been driven by advancements in computing power & storage and by falling costs. PCs are faster than ever before and our iPhones store as much as old computers did. Access to and use of data has also improved dramatically. Public, private and hybrid clouds are allowing multiple users to access data simultaneously, with limited overlap of storage.

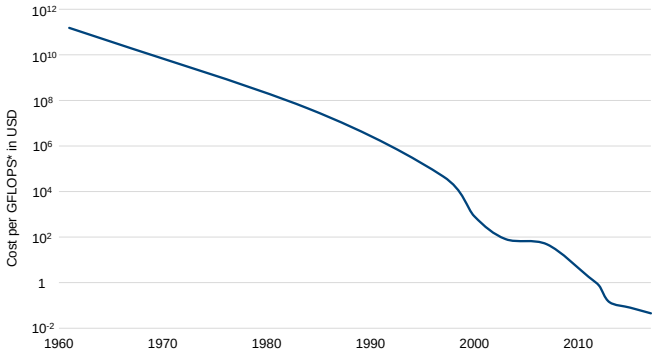
Exponential Growth in Computing Power



Notes: *FLOPS per $1,000: Floating Point Operations Per Second (FLOPS) = The amount of computing that a machine does for each $1,000 of costs. The graph has been logarithmized to a base of 10.

Sources: Singularity, AI Impacts

Cost of Computing - Falling Since 1960



Notes: *GFLOPS = GigaFLOPS - 1 GFLOPS = 10^9 FLOPS. The graph has been logarithmized to a base of 10.

Source: Wikipedia

More usable data has brought about novel applications. Autonomous vehicles, facial recognition and the native ads in your Facebook feed are just some examples. Many industries are using data science to advance their businesses. Finance is one of them; however, it has been a relative laggard, particularly in Canada. The main reason is that there are not many professionals who possess both programming and finance skills.

Applications of Machine Learning

The rise in availability and use of data has given way to advancements in data analysis, specifically in machine learning. This technique is used to analyze data, to "learn" from that data and improve performance. It is also used to uncover "hidden insights" from the data[3].

Machine learning generally falls into three broad categories:

1) **Supervised learning** involves giving the machine not only the outcomes but also the data necessary to arrive at those outcomes. If the machine does a good job of learning then we may use it for predicting future outcomes. The healthcare sector provides us with some of the easiest-to-understand examples of real-world machine learning applications. For instance, supervised learning has been used to "analyze data from electronic health records to predict heart failures" with a high level of accuracy[4].
2) **Unsupervised learning** is different in that it does not provide the machine with an outcome. Instead, the model is given unlabeled data which it then describes as it sees fit. Researchers at Google used unsupervised learning to detect high-level features in random YouTube images[5]. Their models were able to identify concepts such as "faces, human bodies and cats", despite the fact that the corresponding images were not labeled to begin with.
3) The last category of machine learning is **reinforcement learning**. In this kind of learning, the machine is provided with a set of rules. It then learns which actions it ought to take so as to maximize a given "award" using a trial-and-error process. Scientists at Alphabet-owned DeepMind used reinforcement learning to conquer the ancient Chinese game of Go[6]. They did this by developing a program that uses deep learning, a subset of machine learning that is loosely related to a biological nervous system[7], and reinforcement learning. The program has been highly successful. The early iterations, which trained on thousands of human games, quickly surpassed human levels and beat the best players in the world. The latest version, AlphaGo, surpassed all previous versions in just 40 days. Remarkably, it did this by playing games against itself, free from the constraints of the limits of human knowledge.

Source: DeepMind

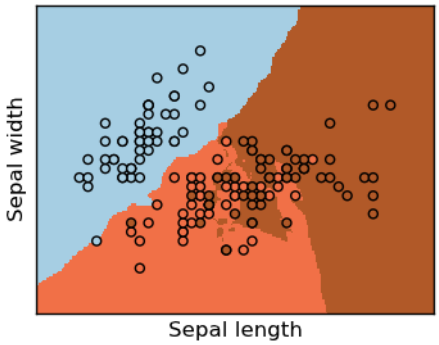
[3] https://en.wikipedia.org/wiki/Machine_learning
[4] https://harvardsciencereview.com/2017/05/16/machine-learning-the-future-of-healthcare/
[5] https://research.google.com/pubs/pub38115.html
[6] https://deepmind.com/blog/alphago-zero-learning-scratch/
[7] https://en.wikipedia.org/wiki/Deep_learning

The world of finance is ripe for machine learning because it is overflowing with data. Even so, we believe that finance is still in the early stages of adoption, especially relative to other industries. The application of machine learning in finance requires a theoretical knowledge of the techniques involved, skills in programming and experience developing quantitative strategies. Few individuals possess all of the required skills.

That said, things are moving quickly. RBC, BlackRock and J.P. Morgan are just some of the companies that have recently announced new ventures dedicated to data science. Machine learning will never solve the market. Still, we believe that it will become a crucial part of the analysis toolkit.

Inside the Machines

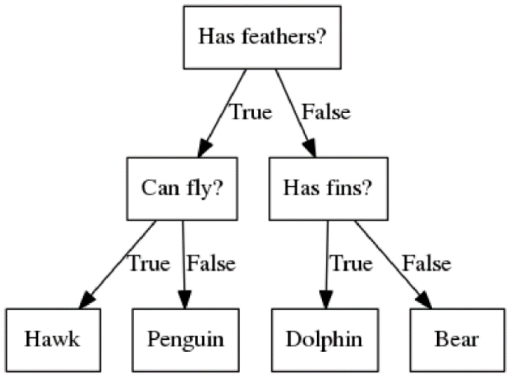
In this section, we will briefly introduce some of the most common machine learning algorithms used today. Each category of machine learning involves its own set of models. In supervised learning, there are k-Nearest Neighbors (KNN), linear models, decision trees and support vector machines. KNN is the simplest algorithm and is used for classification and regression. The basic idea is that it will make a prediction on a data point by comparing it to similar instances, or "nearest neighbors". For example, we might use KNN to classify different species of flowers based on sepal (the part of a flower that encloses the petals) widths and lengths.



Source: "KNN classification example" – scikit-learn

Linear models (OLS, ridge, lasso, etc.) have been used extensively for a long time. Generally, they are used to predict a value (y) using a linear function of the input features ($w_o + w_1x_1 + \ldots + w_px_p$) or to make a classification using similar logic. In regression, we could use linear models to predict Boston house prices[8] based on characteristics such as percentage lower status of the population, average number of rooms per dwelling and weighted distances to five Boston employment centers.

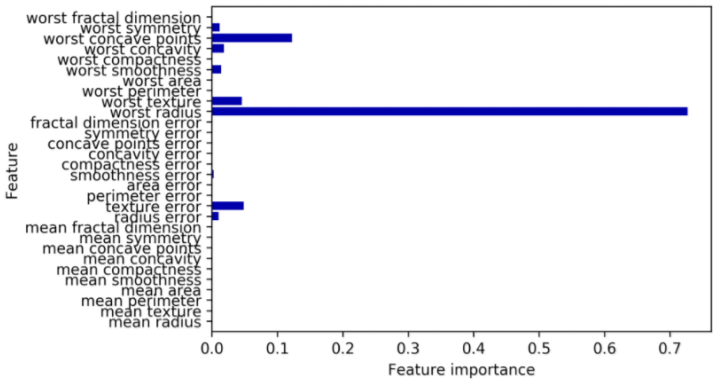
Decision tree models, and their extensions (random forests, gradient boosting machines, etc.), are also widely used for classification and regression. The logic behind these models is quite intuitive. Essentially, the models arrive at an answer by iterating through a series of if/else questions. In a simple task, we might use a decision tree to distinguish among several animals.



Source: "Introduction to Machine Learning with Python" – Andreas C. Muller & Sarah Guido

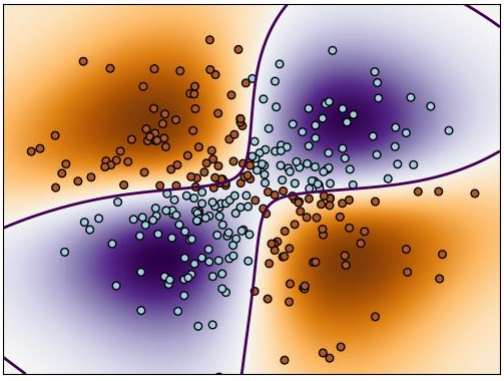
[8] http://www.scipy-lectures.org/packages/scikit-learn/auto_examples/plot_boston_prediction.html

In a more complex task, we could use a decision tree to help us to determine whether or not a patient has breast cancer by asking multiple statistical questions about that patient's tumor. Decision tree algorithms are especially useful because they allow us to look inside the models and to determine which features are the most important. In regards to sklearn's breast cancer Wisconsin dataset, the models show that "worst radius" is the most important determinant for separating malignant and benign tumors.



Source: "Introduction to Machine Learning with Python" – Andreas C. Muller & Sarah Guido

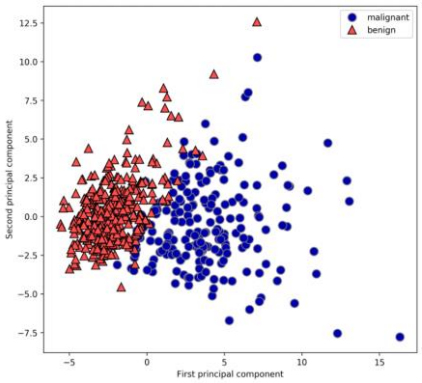
The last set of models is support vector machines. These algorithms are used in classification and regression. With respect to the former task, SVMs are used to separate data into classes using "decision boundaries." The right figure shows how a non-linear SVM classifies blue versus brown points with a high degree of accuracy based on such boundaries.



Source: "Non-linear SVM" – scikit-learn

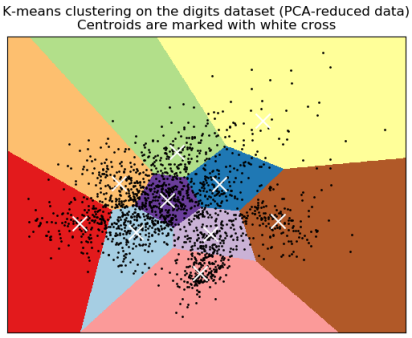
Unsupervised learning is more complex, since there is no known output. These algorithms are simply given data and then are asked to explain it in some way. Generally, unsupervised learning involves either transformation or clustering tasks.

In transformation, the goal is to create a new representation of some data that will, hopefully, provide better insights than the raw data does. Principal Component Analysis (PCA) is one of the most common transformational algorithms. It is used to "rotate" data into features that still explain the dataset but that are uncorrelated with each other. Typically, a subset of those features is then taken to best explain the data using fewer components. We could demonstrate this logic on the same cancer dataset as above. For example, PCA does a good job of separating the two classes based on just two principal components.



Source: "Introduction to Machine Learning with Python" – Andreas C. Muller & Sarah Guido

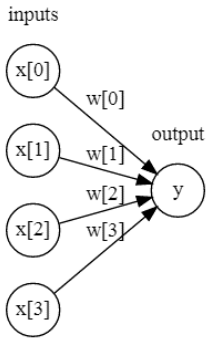
In clustering, the aim is to separate data into groups, called clusters. The simplest clustering algorithm is k-Means. It works similarly to the one described above, but with a few differences. The k-Means clustering algorithm alternates between two steps. First, it identifies "cluster centers". Second, it classifies data points that are nearest to those clusters. This process continues until, eventually, all of the data points are classified. As an example, we could use k-Means clustering to accurately classify different handwritten digits (1, 2, 3…9). In the next figure, each black dot represents a specific person's handwriting. The k-Means algorithm works by first identifying the cluster centers, marked with white crosses, and then by classifying the nearest neighbors accordingly. The end result is 10 different clusters, represented by 10 different colors.



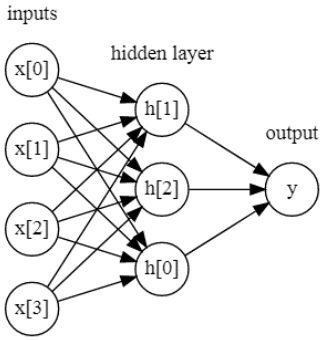
Source: "A demo of K-means clustering" – scikit-learn

The last category of machine learning algorithms, reinforcement learning, depends on what are known as neural networks. With linear regressors, the weighted sum of the inputs results in the output. With neural networks, there is an intermediate step in which weighted sums are used to calculate "hidden units". These units, which comprise the hidden layer, are then combined to form the final output layer.

Linear Regressor



Neural Network



Source: "Introduction to Machine Learning with Python" – Andreas C. Muller & Sarah Guido

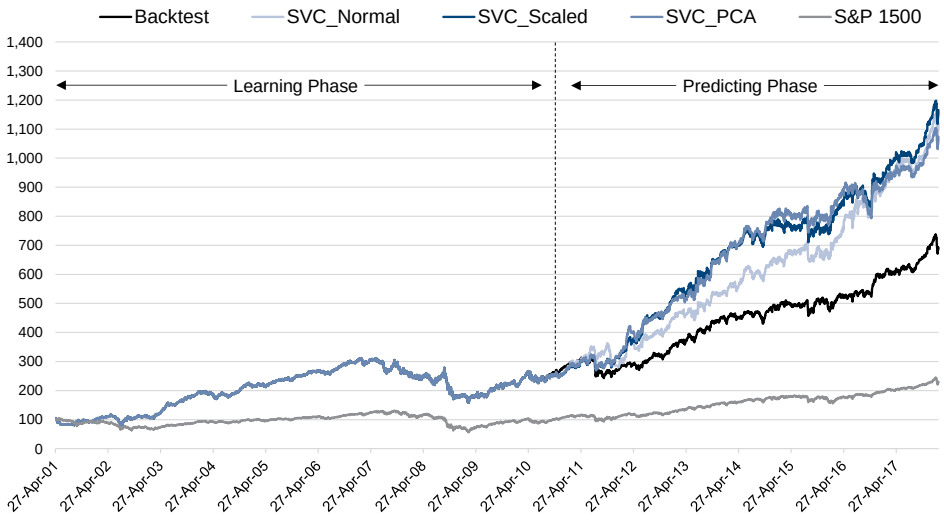
The above right example shows a neural network with one hidden layer. Neural networks with more than one hidden layer are known as deep learning models. As mentioned previously, these networks form the basis of highly complex models such as the ones used to develop AlphaGo and to recognize images.

There are many algorithms in the machine learning toolkit. Each one has its strengths and weaknesses. That said, it is usually impractical to apply most if not all of them at once. In our work, we first applied various models but eventually settled on the set that consistently worked the best for our data.

Applying Machine Learning in Practice

We used machine learning in practice to refine one of our research strategies. The strategy was already good to begin with and outperformed the S&P 1500 benchmark in our backtest from April 27, 2010 to February 14, 2018. Still, we believed that applying machine learning would help to improve trade selection.

After trying several models, we settled on SVMs. The results were very promising. Three of our four SVM models outperformed the base strategy, while only taking on about 90% of the trades.



Backtested, not actual results – April 27, 2001 to February 14, 2018

The above chart demonstrates the potential for applying machine learning in finance. Even so, we realize that our explanation is light on details. *If you are interested in learning more then please read our new white paper, "Unloved to Less Unloved".*

Conclusion

The rise in availability and use of data are already having an impact in finance. It is still early; however, some professionals are already using the wealth of available data to apply sophisticated techniques. Machine learning will never solve the market. Even so, it may help practitioners to find an edge in an increasingly competitive industry. We are optimistic that data science will continue to add value in our investment process. In our view, the more data the better…

Backtesting / Hypothetical Performance – The performance presented is not the actual performance. Hypothetical performance is provided for illustrative purposes only. Hypothetical/backtested performance is used to illustrate historical performance had the portfolio been available over the indicated time period. Hypothetical performance has inherent limitations, such as the use of fixed assumptions, that could have materially impacted performance and is created with the benefit of hindsight. As this data is provided for information purposes only and actual figures may differ, it should not be relied upon as investment advice. There is no assurance that the portfolio will achieve or exceed its investment objectives. Management fees and other expenses were not deducted from hypothetical portfolio performance.