AWALEE NOTES

Breaking down Data Science:
Concepts, Methods, and Business Applications in Finance

Study carried out by the Data Science Practice
Special thanks to Lionel Lecesne

awalee
FINANCIAL CONSULTING

NOVEMBER 2018

Table of contents

For questions or suggestions, the reader may contact the authors at ALisbonne@awaleeconsulting.com and LLecesne@awaleeconsulting.com.

ABSTRACT

The present note defines Data Science and clarifies how it is related to other concepts among which Big Data and Machine Learning. Basic principles of Machine Learning are exposed and the different types of learning explained. The document ends with a presentation of some Data Science business applications in the finance sector.

INTRODUCTION

The fast *expansion* of innovations relying on digital technologies from the beginning of the 2000's is at the origin of the dissemination of large amounts of data (Lynch, 2008). In addition, improvements of hardware regarding computing power and storage now make it possible to run very demanding algorithms utilizing large data sets. The conjunction of these two behaviors and the easy availability of data offers new opportunities for businesses and research to analyze human and natural behaviors. Still, these huge amounts of data pose a number of technical challenges among which data collection, storage, and processing. This entails that data scientists are required to have a wide range of skills to cope with data mining and machine learning problems (Lemberger *et al.,* 2016).

Data science is a broad notion whose bounds and components are quite elusive. Further, a number of other very actual notions like machine learning and big data seem to be thinly related to data science. It is hence a challenging task to formulate a comprehensive and well-delimited definition. The present note aims at defining data science and claryfying how it is interrelated to other notions like Big Data and Machine learning. In addition, it provides illustrations of how data science can be used for businesses in a variety of financial segments.

The remainder of this paper is as follows. First, the concept of Big Data is introduced and it is then explained why new methods need be used to deal with it. Next, Data science is defined and it is discussed whether it is a new discipline or not. Different classes of machine learning algorithms are then introduced. Finally, concrete instances of data science applications are provided to the reader.

I. WHAT IS BIG DATA?

The concept of *Big Data* started becoming popular at the beginning of the 2010's. Big Data may broadly be defined as data sets whose size and complexity imply that traditional data analysis methods, hardwares, and softwares are inappropriate to efficiently extract knowledge from these data.

However, even with this broad definition, it remains tough to distinguish Big Data from common large data sets. Three elements need to be specified in order to precisely characterize a Big Data set, these elements being usually referred to as the three V's: *Volume*, *Variety,* and *Velocity.*

Volume: For each variable composing a data set, the number of observations is significantly large and poses the technical challenge of storing the data.

Variety: The data set should comprise a large number of gross variables, which may significantly differ in their quality and format.

Velocity: Data sets are generated, collected, and updated at relatively high frequency. As an illustration, stock market data are updated, collected, and processed at millisecond granularity.

Strictly speaking, it follows that a data set that does not fulfill the three V's cannot be considered as being a Big Data set. The mining of Big Data requires new techniques belonging to Data Science, this latter area we define in the next section.

II. DEFINING DATA SCIENCE

Having previously introduced the notion of Big Data, the purpose of this section is to define Data Science and explain how it is related to Big Data.

Data Science is commonly defined as the *activity of drawing out knowledge from data*. The very general nature of this definition results from the fact that Data Science actually comprises a variety of methods to explore a great diversity of data types. As exhibited by figure 1, two main disciplines may actually be recognized as being part of Data Science, namely *Data Mining* and *Statistics*.

Statistics is actually not a recent discipline as it has been existing for more than one century. Statistics may be defined as the science of collecting, preparing, processing, and interpreting data.

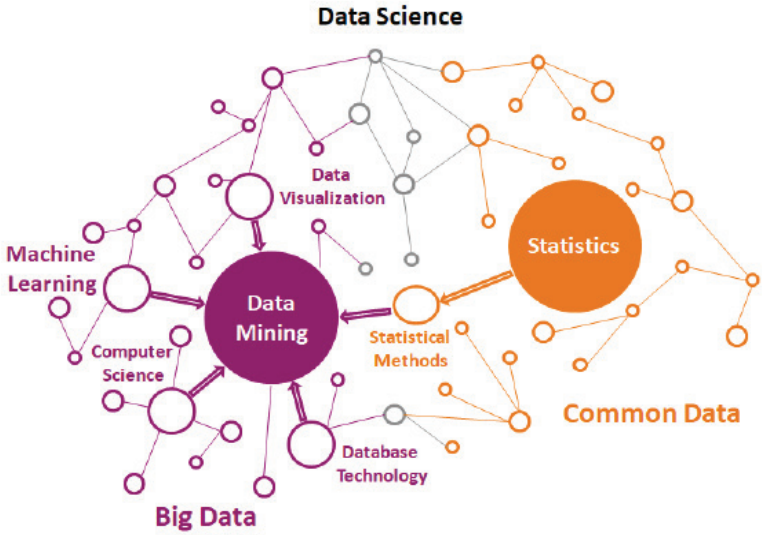

Figure 1 : **Data Science overview**

One main aspect of statistics is that it heavily relies on probability theory to draw conclusions about an entire population based on samples of this population. Hence dealing with very large data sets like Big Data does intrinsically not match the definition of statistics. Further, Statistics has a strong interest in finding causality patterns between variables and thus heavily relies on models and hypotheses testing.

Although there are some bridges between Data Mining and Statistics (Friedman, working paper), these two fields differ from each other in many respects:

• A first noticeable difference obviously regards the type of data used by these two areas of Data Science. Statistics rather relies on traditional data sets that cannot be considered as Big Data, even when large sets are used. In opposition, Data Mining utilizes Big Data even if traditional data sets may also be used.

• Next, Data Mining and Statistics serve different goals. Statistics essentially aims at detecting causality patterns between variables and inferring results for an entire population basing on samples. In contrast, Data Mining is mainly focused on the prediction of variables' future value, objects classification, clustering, and patterns recognition.

• An other point relates to methods employed to conduct analyzes. Statistics is essentially based on probability theory and hypotheses testing. Contrarily, Data Mining uses a lot of methods from other fields, among which Machine Learning, Database Technologies, and Data Vizualization. Notice that Data Mining also makes use of some statistical methods which emphasizes that it is not completely independent from Statistics.

• Finally, Statistics and Data Mining differ in the techniques used to assess accuracy of analyzes. That is, Statistics utilizes specific tests which depend on underlying model assumptions. These tests are usually applied ex-post, that is after estimates have been conducted. A completely alternative approach is applied in Data Mining. For some types of methods, the model is trained on training data before being applied to test data.

Machine Learning has become very popular for the extraction of knowledge from data and nowadays represents a large portion of data mining techniques. Next section deals with providing the reader an overview of main Machine Learning algorithms.

III. MACHINE LEARNING: THE BIG PICTURE

Machine learning is a subset of algorithms belonging to artificial intelligence (The reader may refer to Hastie *et al.,* 2009 for a comprehensive presentation of Machine Learning). Arthur Samuel, one of the pioneers in artificial intelligence, defined machine learning as the *ability to learn without being explicitly programmed*. That is, in Machine Learning a system is given the capability to learn by itself from new data and experiments. Four main types of problems may be addressed by machine learning algorithms, namely *data classification, data clustering*, *regression*, and *dimensionality reduction*. The choice of a machine learning algorithm then depends on the type of problem one wishes to solve and on the nature of the data.

As depicted by figure 2, three major classes of machine learning algorithms may nowadays be acknowledged: Supervised learning, unsupervised learning, and reinforcement learning. Each class contains a number of specific algorithms, which often rely on popular mathematical or statistical methods. Machine learning has known a fast expansion over the recent period thanks to the increase of computers' speed and data availability. However, some mathematical and statistical techniques that are used were already established in the first half of the twentieth century. Thus, techniques like *linear regression* are relatively simple to understand and implement. We present in the following paragraphs the main machine learning categories and the types of problems they allow to solve.

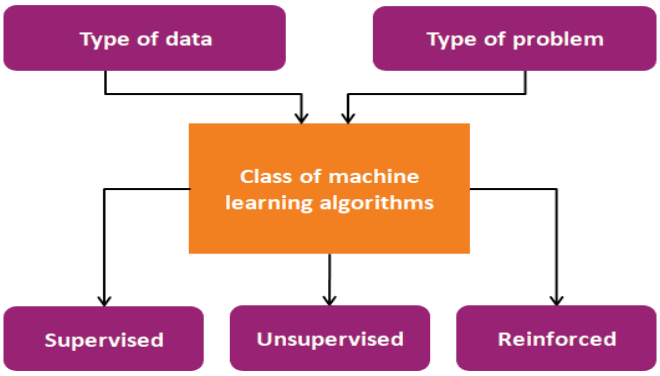

Figure 2: **Deciding factors for ML type**

Supervised learning

The basic idea of supervised learning is that the selected machine learning model (or equivalently algorithm) is trained on data to achieve a given task. Two main types of problems may be addressed with supervised learning. A first type of problem is variable prediction, that is inferring future value of a target variable relying on past observations of both this target variable and a set of predictive variables correlated with the target variable. An other pattern of supervised learning is classification (Kotsiantis, 2007).

Classification consists in assigning labels to objects in the data. In other terms, in a classification problem one has a number of predefined categories and the goal of the algorithm is to distribute objects among them. Even if a number of statistical methods exist to perform prediction and classification, models dealing with these patterns frequently rely on regression techniques (linear regression, logistic regression,...).

To illustrate the functioning of supervised learning, let us consider the case of a prediction problem. The basic principle is that a number of observable variables, referred to as *predictive variables*, are used to predict the future value of an other variable called *target variable*. To make these predictions, supervised learning relies on past realizations of both the target variable and the set of predictive variables. As a simple example, a car company may wish to predict the number of car sells in the next year using GDP growth of the current year. Relying on past observations of GDP growth at year $\mathcal{N}$ and car sells at year $\mathcal{N}+1$, a relation between both variables may be build which enables predicting next year's car sells.

Formally, let $\mathcal{Y}$ denote the target variable that is to be predicted. The goal of machine learning in a prediction model is to predict $\mathcal{Y}$ using m predictive variables $\mathcal{X} = (\mathcal{X}1, ..., \mathcal{X}m)$. The observed value of $\mathcal{Y}$ is given as

$$\mathcal{Y} = \varphi(\mathcal{X}) + \varepsilon$$

where $\varphi(.)$ is an unobservable function that mainly explains $\mathcal{Y}$ from $\mathcal{X}$ and where ε is a random variable describing the difference between predicted and realized values of the target variable. The true function $\varphi(.)$ is unobservable, hence the goal of machine learning is to obtain an approximation of φ that we denote φ^*.

The prediction function φ^* is inferred in two steps using historical observations of both the target variable and the set of predictive variables. Available data are segmented in two samples, namely the training data and the test data. First, the prediction function is trained using the training data until some level of performance is reached. Performance is assessed by a metric that compares the realization of the target variable with its predicted value. Next, the prediction function is tested using the test data. Test data are used to evaluate how well the predictive function performs at predicting the target variable on a sample on which it has not been trained. This step is critical as some supervised learning algorithms suffer from overfitting, that is they very well predict the target variable within the training data but have poor performance at predicting the target variable outside the training data set.

Notice that the handy feature of supervised learning is that output data are labelled, which entails that it is possible to test the algorithm by comparing predicted with realized values of the target variable. Supervised learning gathers a number of popular mathematical and statistical techniques among which *decision trees*, *ordinary least squares regression, logistic regression*, and *support vector machines*.

Unsupervised learning

While supervised learning is essentially used for prediction and data classification, the purpose of unsupervised learning is mainly data clustering and dimensionality reduction. Unsupervised learning describes the structure of unlabelled data according to some features hidden in the data. It hence strongly differs from supervised learning in a number of respects. One major feature of unsupervised learning is that the model cannot be trained as with supervised learning since data are unlabelled. This is a key distinction from supervised learning algorithms. An other consequence of dealing with unlabelled data is that accuracy of the model cannot be evaluated with test data. This type of machine learning is particularly convenient for identifying patterns in the data and for anomaly detection. A variety of algorithms may be used for unsupervised learning, among which the most popular are *k-means* (Chawla *et al*., working paper), *neural networks*, and *principal component analysis*. We briefly introduce the problems of data clustering and dimensionality reduction (Hinton and Salakhutdinov, 2006) in the following paragraphs.

Data clustering (Flynn *et al.*, 1999) may be defined as the organization of a collection of objects based on their similarities. Hence, an object belonging to a given cluster is more similar to other objects of this same cluster than to any other object of an other cluster. Figure 3 provides a pictorial representation of data clustering. Objects of interest are depicted by points in an $\mathcal{X}$-$\mathcal{Y}$ space. Each object is completely defined by its features which are values taken by $\mathcal{X}$ and $\mathcal{Y}$. In the present case one may note that objects are organized in five clusters. However, a different number of clusters might have been chosen. Actually, the number of clusters may be chosen arbitrarily but there exist specific methodologies to determine an optimal number of clusters.

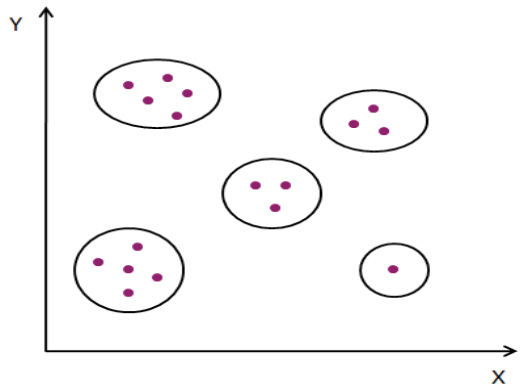

Figure 3: **Illustration of data clustering**

An other type of data processing that may be addressed using unsupervised learning algorithms is dimensionality reduction. This is a methodology that enables reducing the number of variables to explain a certain behavior. Machine learning indeed often relies on an important amount of variables, some of them being usually highly correlated with each other. Hence, it may occur that many of these variables do not bring much information for the problem to solve and only increase complexity and computation time. Dimensionality reduction then only retains the most significant variables, referred to as the principal variables. One very popular dimensionality reduction technique is the principal component analysis, which already existed in statistics before being implemented within machine learning algorithms.

Reinforcement learning

Reinforcement learning is a type of machine learning in which an algorithm learns how to act to achieve a certain goal in a given environment. It is hence classified as a behavioral learning model. As there are no training data in this type of machine learning, the algorithm learns from experience. Specifically, the algorithm learns from making trials and errors to achieve a task. At each trial the outcome is analyzed and gives a feedback to the algorithm. Training experiments are collected with the goal of maximizing a long-term reward.

A trial leading to an improvement of the reward reinforces the algorithm which progressively converges to the best strategy. A well-known application of reinforcement learning is self-driving cars.

To help the reader grasp the functioning of reinforcement learning, this class of algorithms may be illustrated by an analogy to some basic human behavior. One may imagine the case of a young child going to sleep and wanting his mother to read him a story. Getting the story read is the child's target. The child will try different behaviors (= trials) to get the story read. A first attempt might be to ask his mother kindly. If the mother is tired, she will decline. Integrating the fact that being kind does not succeed (= feedback to the algorithm), the child will try an other way by being insolent. The mother will then reply by being severe and the child will integrate this information (= new feedback). As a last trial, the child will start crying loudly to get the story read. This immediately leads the mother to capitulate and read the story to her child. This basic human behavior well illustrates the way reinforcement learning works.

In addition, notice that an other type of learning, referred to as Deep Learning can be implemented in any of three aforementioned learning types. Deep learning is exposed in the appendix.

IV. DATA SCIENCE APPLICATIONS IN FINANCE

The huge amount of data available in the financial sector has made it possible to develop applications relying on machine learning. Four significant fields of applications in the banking and asset management industries are exposed in the following paragraphs.

Algorithmic Trading

Machine learning has now been around for long time in the banking industry with automated trading. With first implementations starting in the 1990's, banking is one of the first business sectors where machine learning has been deployed. This was enabled thanks to several coincident factors, namely the arrival of computers for trading activity, the increase of market data availability, and the contest for trading at higher speed. Machine learning for trading is referred to as algorithmic trading and relies on prediction algorithms belonging to supervised learning (labelled data). It consists in using artificial intelligence to predict intraday prices (usually at minutes or seconds granularity) and make highspeed automated trading decisions according to these predictions. Algorithmic trading is constantly getting improved within the banking industry.

The actual tendency is to move towards deep learning which relies on neural networks with multiple layers. This allows dealing with massive amounts of data and should yield better and faster mining of these data.

Portfolio Management

Machine learning is progressively getting implemented in asset management. This technology first attracted hedge funds for algorithmic trading but also slowly draws attention of investment funds. In asset management, machine learning is essentially used to predict future asset prices[2], detect arbitrage opportunities, and decide of optimal portfolio allocations. The benefit of machine learning in this field is that it can process large amounts of available data to make enlightened investment choices. Data may be of many different types, incorporating for instance market data, financial statements, and macroeconomic variables.
Particularly, a growing trend in portfolio management activties is to use machine learning technology to build customized portfolios. Investors have to respond to questionnaires in order to establish their goals and risk-aversion profile. Machine learning algorithms then process investors' profiles together with market and economic data to suggest an optimal allocation. This technology which is sometimes available online is referred to as robo-advisor.

Corporate and Investment Banking

Data Science and Machine Learning also have useful applications in the area of corporate and investment banking. An overview of these applications is exposed in the following lines.

At front office, thanks to machine learning there is an enrichment of the corpus of methodologies designed for the pricing of derivatives. In particular, Neural Networks are very useful for front office applications. Accurate pricing of derivatives is notably a key issue for traders for market making and to take efficient trading decisions.

An other useful application of Machine Learning is data quality for risk management. Data quality is a concern that cuts accross several banking and asset management activities. The issue consists in detecting outliers in historical data sets, these latter being utilized for risk metrics computations. Having outliers-free historical data sets is then a critical issue for reliability of risk metrics.

Machine learning algorithms may also be implemented for the prediction of credit risk. A number of both supervised and unsupervised methods offer the opportunity to estimate credit risk metrics.

Retail Banking and Insurance

Retail banking and insurance are segments increasing reliance on Data Science to extract useful informations about their customers and insureds. A central issue for banks and insurance companies is to accurately grasp the risk profile of their clients to accurately price contracts they propose. Retail banks are particularly concerned with assessing the default risk of their clients to fairly decide of an interest rate for loan attribution. On their side, insurance companies face the same type of problem as they need to evaluate the lifestyle of their insureds (risk taking, life hygiene,...) to fairly price insurance premiums.

Statistical classification methods like score assignment have long been used in these areas to meet business needs. In greater detail, classification algorithms are used to organize clients between predefined categories. Nowadays the amount of clients data is so large that machine learning algorithms are more and more used as they enable considerably enhancing scoring models. As an illustration some insurance companies now use machine learning prediction algorithms to predict lifetime of their insureds.

Fraud Detection

An other promising field of machine learning implementation is fraud detection. It has already been applied with success in non-financial sectors but its development in the financial and insurance segments is currently growing.
Outliers detection algorithms are utilized to detect unusual data and hence identify risk of fraudulent activity, scammers, and scam techniques. Digital payment companies are particularly exposed to financial scam and hence particularly attracted by machine learning technologies.

CONCLUSION

It is nowadays clear that Data Science offers many technological opportunities to extract knowledge from Big Data. A revolution is on the move and is not limited to research and purely industrial segments. Many businesses have already implemented artificial intelligence for applications in marketing, logistics, or fraud detection. However, this move is relatively in its infancy in the financial industry. A number of applications are yet to come, especially due to the huge amount of data available. Two main types of uses are notably promising, namely prediction and data quality management. Further, these applications cross many financial departments, among which front office, risk management, and asset management. Finally, it is likely that Data Science will disseminate to back and middle offices.

2 Contrarily to predictions for algorithmic trading, predictions in asset management applications are conducted for longer time periods ranging from days to years.

APPENDIX

Deep Learning is a subfield of Machine Learning in which algorithms imitate the functioning of human brain. Precisely, Deep Learning relies on deep neural networks, that is artificial neural networks with multiple neuron layers. As exhibited by figure 4, several (hidden) layers may be found between the input layer (which receives gross data) and the output layer (which ends the data processing). Deep learning uses a sequence of multiple layers to extract some features in the data. Each subsequent layer uses the output from the previous layer as input.

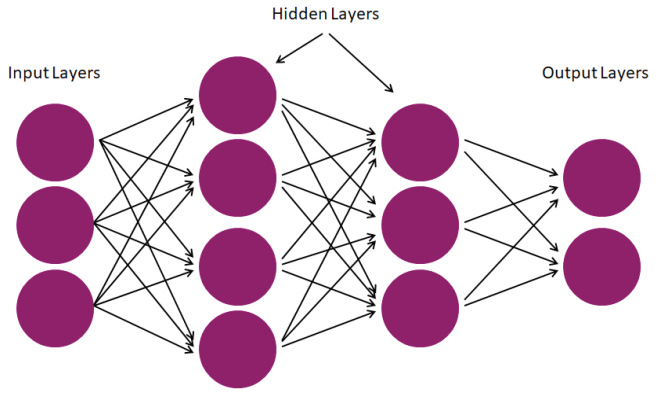

Figure 4 : **Neural Network with multiple layers**

The attraction for Deep Learning stems from the shortcomings of Machine Learning. It appears that performance of Machine Learning attains a plateau as more data are given as input. In opposition, Deep Learning keeps improving as algorithms are fed with additional data which justifies increasing interest for this technology. Further, it is found that Deep Learning performs very well with supervised learning algorithms (labelled data).

REFERENCES

[1] Chawla, S., and Gionis, A.. k-means–: A unified approach to clustering and outlier detection. Working paper.

[2] Flynn, P.J., Jain, A.K., and Murty, M.N., 1999. Data clustering: A review. Working paper.

[3] Friedman, J. H.. Data Mining and Statistics: What's the Connection? Stanford University working paper.

[4] Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Second Edition, Springer.

[5] Hinton, G. E., Salakhutdinov, R. R., 2006. Reducing the Dimensionality of Data with Neural Networks. Science 313, 504.

[6] Kotsiantis, S. B., 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31, 249-268.

[7] Lemberger, P., Batty, M., Morel, M., and Raffaëlli, J.-L., 2016. Big Data et Machine Learning, Manuel du data scientist, Dunod.

[8] Lynch, C., 2008. Big Data: How do your data grow. Nature 455, 28-29.

ABOUT US

Awalee est un cabinet de conseil indépendant spécialiste du secteur de la Finance, créé en 2009 et qui compte plus de 80 collaborateurs.

Nous sommes en mesure à la fois d'adresser des sujets relatifs à l'expertise des métiers de la Finance (Consulting) et de conduire des projets d'organisation et de transformation (Advisory). Et nous le faisons grâce à la synergie agile de ces deux savoir-faire.

Nos expertises s'exercent dans la conformité réglementaire, la finance quantitative, la fonction finance, la gouvernance des outils & systèmes, le management des risques et les marchés financiers. Au-delà de ce que nous faisons, il y a comment nous le faisons : viser l'excellence et repousser nos limites tout en cultivant la convivialité et en favorisant l'esprit d'équipe.

Nous sommes Awalee : nous sommes AWARE & AWESOME.

Awalee consulting
59 avenue Marceau
75016 Paris

www.awaleeconsulting.com

twitter.com/awaleeconsulting

linkedin.com/awaleeconsulting