Finance and new technologies at the heart of information acquisition and processing; how has informational efficiency been modified and what is the impact on fund managers' performance?

Adélaïde Maitre

GRENOBLE iae Graduate school of Management

Master Thesis

Finance and new technologies at the heart of information acquisition and processing; how has informational efficiency been modified and what is the impact on fund managers' performance?

Presented by: MAITRE Adélaïde

University advisor: JIMENEZ-GARCES Sonia

Master 2
Program Advanced in Finance and Accounting
2019 - 2020

Summary

Abstract

The aim of this research is to understand the impact of new technologies on financial markets. The question is to determine whether the profusion of information increases informational efficiency. And how this changes the way prices are formed and, above all, how it affects the performance of fund managers.

Introduction

The financial markets have always been the first to adopt new technologies in order to have more and more information and profits, and this at an ever-increasing speed. The telegraph was the first transmission technology, followed by the telephone in the 1970s and finally by increasingly powerful computers. These technologies have reduced latency time, managing data from collection to analysis to the conclusion of the transaction. The emergence of the web, then social networks, has fostered the profusion of information that is now the playground of data scientists. All this abundance of digital information has changed the approach to gathering information for financial decision-making but has also changed the dynamics of the markets. (Datafloq)
Excerpt from the famous Fama 1970 Hypothesis of the Informational Efficiency of Financial Markets:
"An information-efficient market is one whose price instantaneously and fully reflects all available information". Thus, past information is directly incorporated, but so are expectations about future events. Therefore, the price of a share reflects its intrinsic value. There are three forms of market efficiency, the weak form; only past information is incorporated into prices, the semi-strong form; past and public information are incorporated into prices, and the strong form; prices incorporate past, public and private information.

The market evolves in a semi-strong form, but what about the evolution of this form with the evolution of information technology? Is it possible for technological progress to lead to a strong form of the market? Has the arrival of new technologies changed and will it continue to change the way information is incorporated and the dynamics of financial markets? Does not the quality and efficiency of markets depend on the quantity and quality of information and its dissemination process? How can Big Data technologies improve information, or at least increase the quantity of information? Is there any benefit? And how is it possible to take advantage of it? How do fund managers benefit from these technological improvements? Will their performance decrease while market efficiency increases? Or is it the reverse?

Finally, do Big Data technologies fundamentally increase these parameters, so we want to see if Big Data improves market efficiency and to what extent this can impact the performance of investment funds?

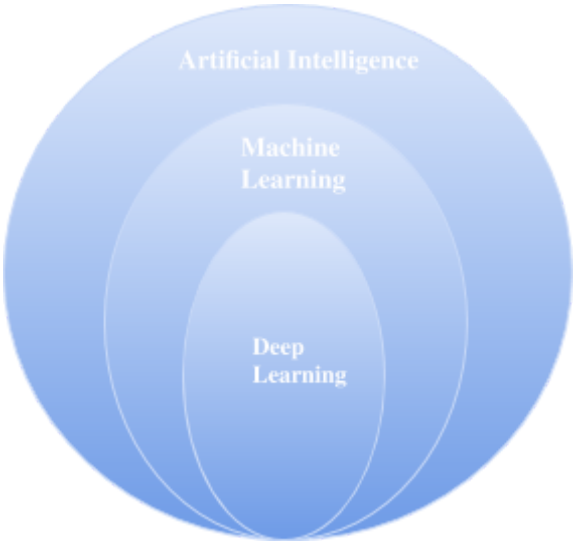
The objective of this research is to analyse if there is an improvement in market efficiency thanks to new technologies. We would like to see whether technological innovations and, more specifically, whether the increase in the volume of data tends to improve market efficiency through the use of data technologies. The question is to determine if the profusion of data and the emergence of increasingly sophisticated technologies, such as artificial AI, are leading to a move towards a strong informational form of markets. Or does this profusion tend to make more and more noise and to what extent can it have an impact on

the performance of portfolio managers who base their strategies on information acquisition and analysis.

Most studies focus on whether or not a market is efficient. However, few studies examine if the efficiency of market changes over time. In addition, most studies focus on a particular event, but very few analyse a set of events, in this case, the increasing arrival of technological innovations. This is one of the first limitations; it is difficult to understand the impact of the arrival of Big Data technologies on the markets as a whole.
The information has an impact on overall risk: we know that risk can come from a lack of information, so the more information there is, the less information risk there is. This shows that market volatility tends to decrease over time. Total risk is the combination of the systematic risk and the specific risk. The informational risk is a component of the specific risk. For security, having more information should decrease the informational risk.
The amount of data available is increasing rapidly, as are the sources and the power of the technologies that can process it. We will see that some asset management companies are using Big Data. It seems that the use of this data combined with artificial intelligence offers higher returns than standard management. This could be proof that Big Data technologies make possible to capture more information and therefore to make gains.
An analysis of abnormal returns over a very wide window could allow us to highlight whether the market tends to be more informationally efficient. Normally, if information tends to be fully and instantaneously incorporated, then abnormal returns should decrease because the potential for gains decreases as informational efficiency increases.

Two main objectives will be analysing in this research. First, we want to know if tools of Big Data increase the informational efficiency and then in which manner tools of Big Data affect the performance of portfolio manager.

In order to process all this data more efficiently, an additional technological layer has been added to the Big Data. This layer is called Artificial Intelligence (AI). Advances in deep learning and the increase in machine power have enabled AI to become an essential technology in data processing.
According to F. Iafrate (2018), big data is what feeds artificial intelligence, which is a response to the massive exploitation of data through autonomous and self-learning solutions. Machine learning and deep learning are forms of artificial intelligence. The basis of its technologies is the execution of tasks through programming. It has been followed by the execution of tasks through learning. It should be noted that the learning of machines is enabled by data. The more data, the more the machine will be able to learn.
Artificial intelligence is all techniques that allow computers to mimic human intelligence. One finds in this framework the machine learning.
Machine learning is a subset of AI that includes techniques that allow machines to improve in tasks with experience. A great deal of training and exercise is required for artificial intelligence to learn. It includes deep learning that is a higher degree in the execution of tasks. Deep learning is a subset of machine learning based on neural networks that allow a machine to train itself to perform a task. (Microsoft)
A diagram will make it possible to understand the link between these technologies.



We will mainly discuss the technologies of Big Data as a whole, which includes Artificial Intelligence, machine learning and deep learning.
We will first look at what Big Data is, its history and technologies. Then we will see through the analysis of the literature on how and to what extent data from the Internet and connected objects are an invaluable source of information. Finally, we will see some elements that can allow us to affirm that the market is evolving towards a new form of efficiency. Through research, we are going to see a duality between informational efficiency and fund managers' performance.

According to Grossman and Stiglitz (1980), “prices reflect the information of informed individuals but only partially, so that those who expend resources to obtain information do receive compensation”.
Therefore, the Big Data tools should allow incorporating more information; more information should then decrease the informational risk and so increase efficiency.
On the other hand, more information can increase market noise. This noise will increase informational risk and ultimately decrease efficiency.
Regarding the performance of fund managers; performance is only possible if there is information asymmetry since active mutual funds managers job resides in acquiring and analysing information, thanks to their financial expertise, in order to generate private information and invest based on this private information (that other marketers may not have). G. Cullen et al. show that funds which use the private information that contradicts the public information exhibit superior average performance. Then, the performance of mutual funds that trade using private information is higher than those that trade not using private information.
It follows from the first part that performance is allowed if informational efficiency is not high. Therefore, if Big Data increases informational efficiency, fund managers will have lower performance, but if, on the contrary, too much information creates a lot of noise, it will be less difficult for fund managers to make more profits.

I. Big Data

I.1. Definition

From the book of S. Yu and S. Guo, 2016. *Big Data Concepts, Theories, And Applications*, we are able to define Data and more precisely Big Data.
Data is a set of numbers, figures, facts, words, observations, measurements or descriptions of something that can be analysed. Note that there are three main types of data. Structured data which is data represented or stored in a predefined format, i.e. in databases and computer languages, unstructured data which is data in any format that is understandable by humans at first glances, such as text, and semi-structured data which is a form of structured data that does not follow the formal structure of data models.
Finally, data is an element that gives information, one looks for the sense of the data. In other words, putting data in context creates added value to constitute information. Thus, the value of data lies in the ability to know how to exploit and link them together. The result is a simple scheme: sets of data are captured and analysed, and the information is then generated and used.
Big Data is about data entry, the first phase of information gathering, but on a completely different scale. We talk about mega data or massive data. The volume of data is such that it is beyond human intuition and analytical capacity. The quantitative explosion of digital data in the age of the Internet has forced researchers to find new ways of seeing, capturing, understanding and analysing the world. New orders of scale in capturing, searching, sharing, storing, analysing and presenting data are now being talked about, given the massive amount of mega data. This is how the Big Data developed, but above all their technologies, those that make it possible to store an indescribable amount of information, but also those that make it possible to process them.

It can, therefore, be said that the rise of computers, storage devices and the explosion in the volume of information has given rise to Big Data and its use. Big Data has revolutionized the way in which every decision-making process is handled, as well as the approach to many business and research issues.

I.2. History

The book of E. Lazard (2016) helped us to summarize a brief history of the technologies.
It was in the United States, in Palo Alto, that the digital revolution began in 1939 thanks to W. Hewlett and D. Packard. HP is focused on measuring instruments but gradually, HP realizes that data acquisition is essential when you have many measuring instruments. Due to the technology available at the time, all instruments were developed using discrete components and the first specialized integrated circuits were designed. HP develops its first minicomputers and personal computers for measurement acquisition and data processing.
The ENIAC follows in 1946, which is the first computer in the world: it weighs 30 tons and occupies 167 m2. It has the capacity to perform 5,000 additions per second.
It was in 1950, in the article "Computing Machinery and Intelligence", that Alan Turing highlighted artificial intelligence.
The invention of the transistor paved the way for the miniaturization of components and in 1958, J. Kilby's invention of the integrated circuit allowed the entry for nanotechnology. In the 1960s, the process of miniaturizing components continued to reduce costs.

At the same time, the development of information technology in large groups gave rise to the need to organise the data collected. This is when the first databases were born, relational databases, i.e. an organization of data in two-dimensional tables. It was Edgard F. Codd, a computer scientist at IBM, who provided this solution. These relational databases can process queries but in a specific language, SQL, this language only understands structured data.
In 1969, the ARPANET project was born, the first packet-switched network, i.e. a connection is established between different locations using packet switching to transfer data.
It was really in 1971 that the digital revolution began with the combination of the first microprocessor and the networking of some twenty geographically distant computers. This marked the genesis of the Internet. Files began to be digitized.
In 1983, the TCP/IP protocol was adopted and the name Internet appeared.
In the 1980s, in order to overcome certain limitations such as slow data processing and visualization difficulties, R. Kimball, B. Inmon proposed Business Intelligence (BI). It is a solution that allows data to be collected, extracted and transformed in order to analyse it according to several criteria and present it synthetically to decision-makers. It is a Big Data tool. The data, in multiple sources and formats, go through an ETL (extract-transform-load), which is responsible for centralizing the information before storing it in a data warehouse. The advantage of BI is the ability to classify, historise and analyse all kinds of data.
In 1990, ARPANET disappeared and the World Wide Web (WWW), which is the public hypertext system, appeared. It allows us to consult pages on sites.

Technological advances were not long in coming, and the increase in storage capacity, memory, and processors, as well as the internal architecture of computers and networks, made it possible to keep pace with the massive production of data.
In the year 2000, 368 million computers were connected worldwide (Supinfo). We will talk about Web 2.0. Data production has accelerated sharply in recent years with the development of the digital economy and the emergence of web giants such as Google, Facebook, Amazon, and Twitter. Cloud computing will make it possible to store all this mass of data.
From 2010, the data lake will appear and this is a big change, unlike the Data warehouse, which does not centralize data in all its forms, the data lake will make it possible to keep all the potential of the original data in any form, i.e. structured or unstructured, to refine the analysis of the data by scientists.

I.3. 4Vs Value

The big data allows us to group a family of tools that answers the 4Vs problem. Zikopoulos and his collaborators (2012), in an IBM publication, describe Big Data as a set of three "V-words": volume, velocity, and variety. In a later IBM publication, Zikopoulos et al (2013) introduced the additional concepts of veracity and value into Big Data.
Thus, it is interesting to define Big Data in multidimensional terms. They can then be defined as follows; volume i.e. the large scale of the data, velocity i.e. the real-time data flow, variety i.e. the different data formats, veracity i.e. the certainty of the data.

I.3.1. Volume

The volume corresponds to the mass of information produced every second, that is to say, the amount of data generated at each instant. Yu, S. and Guo, S., (2016).
Big data are gigantic in size: at the end of 2018, the global volume of digital data reached 33 zettabytes, compared to only 2 zettabytes in 2010 (lebigdata). One zettabyte is equivalent to 1

billion terabytes, or 10^{21} bytes. The forecast for 2035 (Statista) is 2,142 zettabytes, which represents an astronomical amount of information. On average, we generate 2.5 exabytes (1-exabyte = 10^{18} bytes) of data every day.
All types of data have greatly increased over time. As far as structured data is concerned, in finance, algorithmic trading has greatly contributed to this expansion. High-frequency trading in 2009 represented 2% of transactions but 60-70% of the volume traded for US equity. In 20 years, the total number of transactions has increased 50-fold and up to 120-fold during the financial crisis (lebigdata).
Unstructured data has also increased significantly, and this data is very widely used by financial institutions both in investment strategies and in the choice of financial or banking products. The number of unstructured data doubles every 24 months.

Source: Blackrock (2018)

I.3.2. Velocity

Velocity is the speed of development and deployment of new data.
One decade ago, the stocks OHLC (Open-High-Low-Close chart) prices were reported the following day. In the current financial market, stock can experience about 500 quote changes and about 150 trades in 1 ms, 1800 quotes and 600 trades in 100 ms, 5500 quotes and 700 trades in 1 s according to M. Lewis (2014).
The analysis and ability to effectively leverage large data technology, advanced statistical modelling, and predictive analytics to support real-time decision-making across all channels and business operations will create value for companies that master these tools and for the financial markets as well.
One of the first technologies that have improved the velocity and the volume of the data in Finance is the High-Frequency Trading (HFT). Computers are programmed to send orders to capture revenue through millions of automated micro-transactions. Traders with the fastest execution speeds are more profitable than traders with slower execution speeds. The order-passing capacity of high-frequency software is 21 million orders in less than 2 minutes. The transmission speed of an order is 0.0065 seconds for 1,200 kilometres. A. Baddou (2017).

Everything is moving faster these times, so the amount of information per second is increasing, the speed of execution is increasing the volume of data. The frequency and speed of everyday transactions are also increasing. This opens up new perspectives for financial institutions, but also many challenges. How can the execution speed of computers be increased even more? How can we continue to increase the storage capacity of information? How can we innovate and further improve the ability to analyse and predict this data? Can this perpetual quest, speed, and innovation lead to an informational efficiency?

I.3.3. Variety

Only 20% of the data is structured data that is stored in relational database tables. The remaining 80% is unstructured data stored in flexible databases such as Data Lake. This can be images, videos, text, voice, conversations or messages on social networks, photos on different sites and much more. Big Data technology enables the analysis, comparison, recognition, and classification of these different types of data. These are the different elements that make up the variety offered by Big Data (Datascience).
The same is true for financial data, whether structured or unstructured. Structured financial data is in the form of time series. These types of data depend on the type of instrument, be it equities, futures, options, ETFs or OTCs. It should also be noted that for the same information, the data can take different forms, for example, if we are not in the same market, although the market tends to become more homogeneous. Yu, S. and Guo, S., (2016).
The combination of these different data formats, therefore, allows us to have indications and a better understanding of the real value of a share, the needs of clients, the types of transactions and the market sentiment.

I.3.4. Veracity

Veracity concerns the reliability and credibility of the information collected. The veracity of information is essential to create value from the data. Valid and authentic data is essential. Besides, a large amount of data collected can lead to statistical errors and misinterpretations. Yu, S. and Guo, S., (2016).
Since Big Data allows for the collection of unlimited numbers and forms of data, it is difficult to justify the authenticity of its content.
In financial terms, the veracity of the information is important, but we can temper this because sometimes the value of a company in the markets is not equal to its intrinsic value but rather to what we believe to be the real value of the company. Thus, a significant amount of misinformation or false beliefs can change the value of a company. This false news can change the value of companies both up and down. In the article "Fake News, Investor Attention, and Market Reaction", Jonathan Clarke and Hailiang Chen show that machine learning algorithms are capable of identifying false news and, as a result, the stock market seems to value false news correctly. The abnormal trading volume increases with the publication of false news, but it is less than that of legitimate news. Thus, while the data collected may not necessarily be true, algorithms may have the ability to separate the true from the false thanks to artificial intelligence.

Finally, it follows that the notion of value corresponds to the profit that can be derived from the use of Big Data as the result of the combination of these 4 dimensions but also and above all to the use of technologies related to Big Data.

This technology represents a privileged commercial stake given its capacity to have a profound impact on the entire integrated world economy.
However, what interests us above all is that Big Data plays an essential role in the transformation of processes, "Machine-to-Machine" type exchanges and thus allows the development of a better "informational ecosystem". It also enables faster and more credible decisions to be made, taking into account information both internal and external to the organisation, i.e. in a utopian way of all existing global information in the near future. These Big Data technologies prove how easy it is today to access an infinite source of information. But they also show us how easy and, mainly, how fast it is to translate this profusion of data into usable information. This is why it may be relevant to ask whether we are heading towards a high informational efficiency form. Indeed, do the dimensions of Big Data make it possible to overcome the practical and theoretical limitations that have made the market until now semi-strong in terms of information?

I.4. Technologies at the age of Big Data

The technology around Big Data has played a very important role in digital transformation. Firstly, data traffic has been redirected to the Cloud, thanks to a shared pool of connected storage devices; the parallel computing method has made it possible to compute a larger amount of data, achieving greater accuracy for the same costs; and finally, unstructured data has been taken into account, thanks to the implementation of new developments in the data architecture space.

Therefore, the technological innovations that have enabled the development of Big Data are classified into three key points: F. Iafrate (2018), S. Yu and S. Guo (2016).

- ★ Storage technologies, which have enabled the implementation of storage systems, considered being more efficient than traditional SQL for the analysis of mass data. An example is the use of NoSQL databases such as MongoDB, Cassandra or Redis. NoSQL does not use a structured query language and allows it to take into account data that would not correspond to a standard table (SQL). These technologies allow us to improve timing and predictive intelligence. The degree of innovation of these technologies and techniques is really important, especially when it comes to real-time simulation and high-volume forecasting, which are the key elements of an effective market assumption. Markets are more complex and interconnected, and information flows faster than ten years ago, as a result of interconnections. It is no longer possible to obtain a complete picture of a portfolio from a single data source. Companies must store and disseminate various types and huge amounts of data and efficiently link disparate data to each other to obtain actionable information. High-volume data management technologies offer solutions for effective data management.
- ★ Massively parallel processing with server infrastructures to spread processing across hundreds or sometimes thousands of nodes. An example is a Hadoop framework, specifically developed for new databases for unstructured data. It combines the HDFS distributed file system, the HBase NoSQL database, and the MapReduce algorithm, which enables the development of high-performance computing modes.
- ★ The storage of data in memory (memtables) speeds up query processing times.

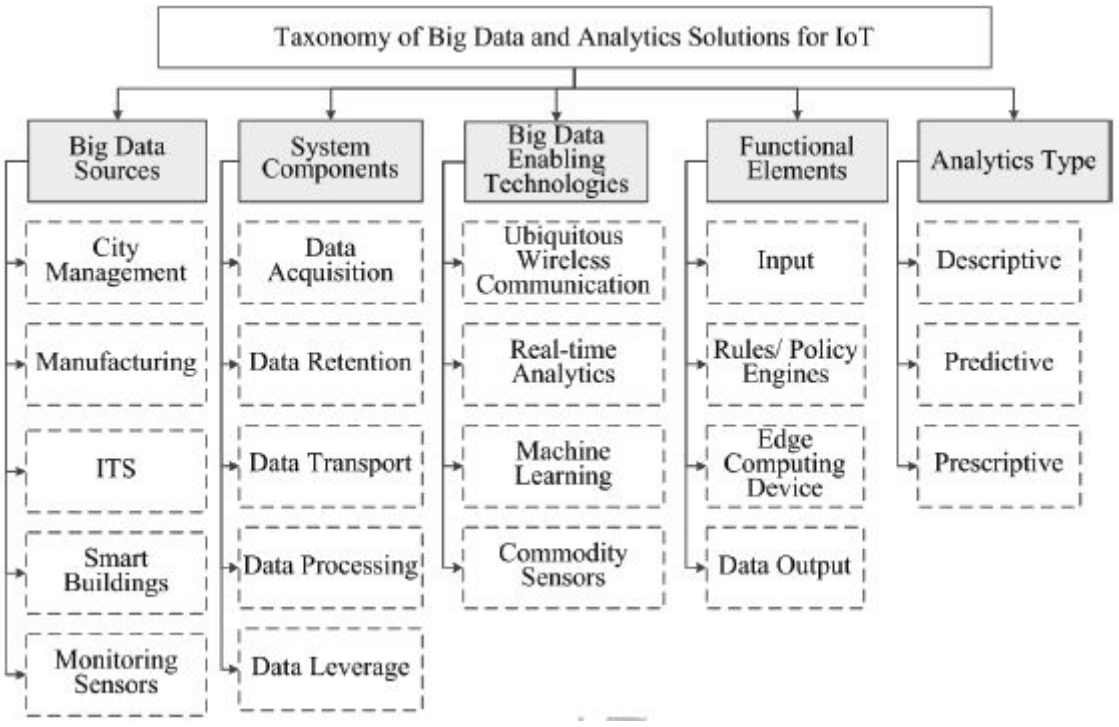
Subsequently, the technological evolution has shown that it is possible to move from the storage of structured information (SQL) in relational databases to very large volumes of data, not necessarily standardized and non-relational (NoSQL). All this is complemented by

very fast analysis and clear visualization (BI). Enabling the creation of predictive models and further analysis through Artificial Intelligence (AI).

These technologies are essential for data collection, storage and analysis. Possession of these technologies is a sign of the use of Big Data in the decision-making process. Thus, technologies are a prerequisite for the use of Big Data. However, the importance of these technologies can vary depending on the capacity to collect and store the data collected and their use. The use can range from simple statistical analysis to more advanced learning methods to the prediction of human behaviour.

I.4.1. Internet of things

Internet of things (IoT) is the basics of the data collection, it is in virtue of the Internet of Things that such a mass of data can be collected on all subjects: climate, consumption trends, physical activity, lifestyle... All this data is collected by the different connected objects and sent to the Big Data technologies to be used. O'Leary, D., (2013).
The Internet of things is what links human data to Big Data technologies. Chui et al (2010) define the Internet of Things as "... sensors and actuators embedded in physical objects - from roadways to pacemakers - are linked through wired and wireless networks, often using the same Internet Protocol (IP) that connects the Internet". Most of the objects are linked to the Internet and therefore all of these objects are connected to each other. The aim is to transform this very large volume of data into valuable insights. The Internet of things is the intermediary, it is by reason of this that Big Data is collected but it is not the finality, the data has to be stored and analysed in order to be used in the decision-making process. This will bring information and enable its analysis, it will also enable enhanced situational awareness and sensor-driven decision analytics. All of this will significantly improve decision-making. Subsequently, the data collected will allow these objects to react and sometimes even make decisions without human intervention, thanks to automation and control. This will allow optimization processes but also the optimization of resources. All this is made possible by complex object autonomy systems. "The IoT in Banking and Financial Services market size is expected to grow from USD 167.3 Million in 2017 to USD 2030.1 Million by 2023, at a Compound Annual Growth Rate (CAGR) of 52.1% during the forecast period". (MarketsandMarkets).



Source: E. Ahmed et al. (2017)

I.4.2. Hadoop

Apache's Hadoop is open-source software originally designed for an online search engine to capture and process information from all over the Internet. It consists of two well-known components, MapReduce and the Hadoop Distributed File System (HDFS). It was designed to distribute distributed data packets to different storage systems and perform data processing in parallel, the name of the first step being "Map"; it then consolidates the processed output on a single server to perform the second step called "Reduce". Hadoop is an open-source project that provides a platform for Big Data. The problem in analysing the data was the difficulty for the computers to record it because it is mainly unstructured or semi-structured. Hadoop addresses this problem; it is designed to solve the problems caused by a huge amount of data that mixes different types of structures. It solves the most common problem associated with Big Data: storing and accessing large amounts of data in an efficient fashion way. (Hadoop)
The data is stored not on memory or disk but in a cloud. Moreover, the software is able to transform data into usable data and keeps in memory where the data was taken and keeps copies, this is resilience: if a server fails it is possible to find the data elsewhere. This process is what makes Hadoop so good at dealing with large amounts of data: Hadoop spreads out the data and can handle complex computational questions by harnessing all of the available cluster processors to work in parallel.

I.4.3. Complex event processing

As we have seen, traditional databases have their limitations. Complex event processing (CEP) is a family of technologies that allows the analysis of a stream of data

flowing from live sources. This allows us to identify the data path but also the most relevant business indicators. This procedure compensates for the previously mentioned limitation: CEP allows companies to analyse and act quickly in the face of changes and this, in real-time. This allows them to act before the opportunity to gain is lost, but it also allows them to move from a batching process, i.e. an automatic sequence of orders made by a computer to real-time analysis and decision-making. A. Adi et al (2006).
The advantages offered by these CEPs are immense; the diversity of the data is such that the information is very rich. These CEPs can provide information that contains data about present and past events as well as trends about the likely future. All this in record time: less than a few milliseconds with high throughput: a hundred or a few thousand events processed per second for extremely complex events.

In finance, the CEP has been used very early on to help traders or automatic trading systems in their buy or sell decisions.

I.4.4. Data Lake

The Data Lake is what replaces Data Warehouses for semi and unstructured data. It is, therefore, a storage place that contains the Big Data. Data Lake is capable of storing all types of data. In a Data Warehouse, most of the data preparation usually takes place before processing; in a data lake, data preparation takes place later (only when the data is used). Therefore, the speed of data processing is faster for data lakes because only useful data will be processed. Data lakes have the characteristic of being very flexible. F. Iafrate (2018)

I.5. Limits

The main limitation of Big Data and its tools is the lack of privacy among individuals. Westin (1970) discusses the four basic rights and states of privacy: the right to solitude, the right to intimacy, the right to anonymity in crowds, and the right to reserve.
To be efficient, Big Data needs data from anywhere, anytime and from anyone. Some of this data can be taken from individuals with or without an explicit and clear agreement. The use of this Big Data, therefore, goes against these rights. Along these lines, if Big Data is a promising tool, it does not necessarily respect the basic rights inherent to human life. The PRISM scandal is a case in point. This surveillance program established by the NSA and revealed by Edward Snowden, which consists of collecting all personal data (Facebook, e-mail, the information contained on phones...) and storing them in a single place for information and tracking purposes. The European Union has since decided to protect its citizens through a law, the GDPR (General Data Protection Regulation). The GDPR obliges, from 28 May 2018, companies to strengthen the protection of personal data and allows a right to forget. Note that, it is only true within the European Union.

Another important limitation in Big Data technologies is that some very sensitive data are stored on clouds, specifically public cloud providers. This paves the way for cybercrime.

Moreover, although big data technologies are very powerful, sometimes there is a lack of compatibility between the financial needs and the possibilities of big data technologies. Hadoop implementation via MapReduce as we have seen above has been a success, however for the financial markets, its implementation is more difficult because it relies on offline batch processing whereas the important thing on the financial markets is the reactivity to

real-time information. X. Tian et al. (2015). However, this limitation should be qualified because innovations and technological developments are moving very quickly.

In addition, Big Data technologies require versatile individuals, whose skills must be highly specialized, including programming, mathematics, statistics, and finance. Individuals must have broad knowledge, master all these tools, and have the ability to understand IT, obtain results, analyse them and communicate them. It is a question of having a wide range of skills and edge knowledge in a very technical field.

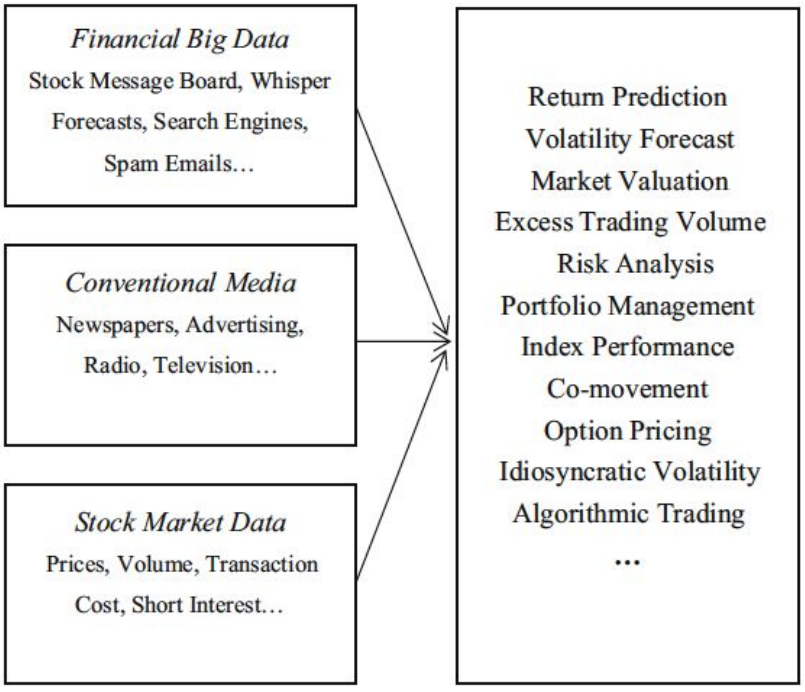
Finally, there is not enough research on the subject applied in finance, in practice, the advances are very strong but at the academic level, there is a gap. Big data and these tools are very complex and as said before require a wide range of skills both technical and theoretical. Moreover, most of the research and advances in this field come from private laboratories.

II. Big Data: Application in Finance

II.1. Backdrop

Big Data technologies in finance can be used in countless situations, from risk control with machine learning tools to financial market analysis with Business intelligence (BI). But also passing by visualization tools to creating new financial and market indicators such as indexed sentiment by exploiting public sentiment.
In financial market, information is the key element of every transaction. Many financial institutions and financial companies have adopted Big Data technologies to get more information and speed it out of the market. The purpose of these tools is to create value by standardizing data collection processes to enable comparability of information, reduce the latency time between data acquisition, data transformation into information and investment decisions. To achieve this, algorithms and software have been improved through new architectures as we have seen. D. Shen, S. Chen, (2018)
Market finance is a field where data is vital information for decision-making in all its aspects. The major data tools in finance are present everywhere: from the front office with algorithms to the back office with data analysis. Big Data tools aim at discovering market opportunities that are not possible through simple data analysis. The ability to process large and complex events at ultra-fast speed has removed virtually all obstacles to the rapid capture of market trends and risk management in record time.
The only limit in Big Data technologies is the imagination and creativity of experts and scientists.
The four natures of Big Data: volume, speed, variety and veracity are used as a prism through which to understand the pitfalls and opportunities of emerged and emerging technologies in financial services. Today, the profusion of data has radically changed the financial sector, not only in portfolio analysis and risk management but also in retail banking and credit rating.
As we have said, the efficiency of financial markets is largely attributed to the amount of information available to agents, but also to the process of disseminating that information. The main financial data, therefore, depends on the specificities of the company, but also on market news, information appearing on social networks, but also on all other media and websites. All this information can have an impact on the formation and updating of investor expectations and can, consequently, influence investment decisions and thus influence share prices. Big financial data increases the volume of information, the speed of processing, the variety, and veracity of this information. This can lead to an increase in the quality of

financial markets since, by definition, the quality of a market depends on its ability to fully and instantaneously integrate information into prices however in the other hand it can also lead to a decrease of the overall quality of the market if this amount of information leads to higher noises. Hence, it would be interesting to know whether and to what extent the overall quality of financial markets has changed as a result of increased technological innovation in data and data processing. As can be seen in the figure, data for the major financial markets complement the traditional media and stock market data already available. It is therefore now possible to go even further in understanding the proponents of financial markets; from predictability to dynamics, for the sake of an even more comprehensive source of information.
What is important in the formation of prices on financial markets is how information is generated, disseminated, used and integrated. Thus, with an additional colossal source of information, as well as increasingly efficient technologies, the question arises, if, in the near future, market dynamics might not tend towards a greater form of markets efficiency. In addition to obtaining numerical information, Big Data tools can infer emotions through data analysis. Thus, these tools can couple the fields of quantitative and behavioural finance.



Source: D. Shen, S. Chen, (2018) Big Data Finance and Financial Markets

II.2. Literature review

There are three main determinants in the valuation of a share price: firstly, the price is determined by the financial information available on the companies, with fundamental analysis (analysis of the company, quarterly, annual publication, etc.) these data are available on the companies' website, the analysis is done by specialists. In addition to this, we add the

analysis of past performance. It is a set of financial data that gives indications on the value of a share. The Big Data tools have access to these financial data as well as institutional investors. We are not going to focus the literature review on the latter since the Big Data tools will "only" allow us to gather all this data, analyse it and draw conclusions almost instantaneously once the financial information is published or when data are stored in management software.
Two other more or less related elements are involved in pricing: the economic situation, everything related to macroeconomic and investor sentiment. Investor sentiment can be derived from the economic situation and the events that influence it (strong growth can lead to euphoria and the formation of a bubble, just as fear can lead to a crisis with a crash) or from what investors subjectively think of a company. This is how the less rational part in price formation appears, which is more subject to expectation and supposition under the influence of emotions. Consequently, if the price of a share is no longer based on numerical analysis: ratios, coefficients, etc., then it becomes more difficult to predict its intrinsic value and its evolution.
It can be said that the understanding of price formation on financial markets is the result of both financial analysis with the analysis of companies and behavioural finance with the analysis of investors' irrational behaviour on the markets and their reactions to the environment. The sum of these elements then determines the price that results from the quantity of supply and demand influenced by financial analysis and cognitive biases.

Accordingly to the literature review, we will see whether investor sentiment influences the price of stock returns, to what extent and how in order to be able to master its components as precisely as possible. The literature review will focus on the impact of this unstructured and structured data from the Internet on stock prices. We will examine whether and to what extent information from the Internet is integrated into prices.
Therefore, we will see how the Internet, search engines and social networks can be an unavoidable source of information nowadays. This can lead us to ask ourselves if this sentiment information can help portfolio manager to create value, in what manner? We can also wonder if this sentiment analysis can help in the discovering of the fundamental price.

After demonstrating from the literature review that prices are influenced by data from the Internet and that investor sentiment plays an important role in the formation of stock prices, we will try to show whether this source of information improves the efficiency of the financial information and also if it can create value for a portfolio manager. For this purpose, we will use a literature review to see how market efficiency has been measured so far.
We will draw inspiration from this to propose a method for analysing market efficiency in order to see whether the increase in technological innovations allows for greater integration and an increase in the quality and quantity of information. In other words, whether technological innovations increase informational efficiency.

Before seeing what has been done in the analysis and prediction of data from the Internet, it seems relevant to define what sentiment analysis is to understand investor market sentiment. From data, it is possible through textual and language analysis to know the quality of the subjectivity of information. Thus, the technical tools of Big Data make possible to say whether data is rather objective or subjective. If it is subjective, then it will be possible to deduce its orientation: positive or negative and therefore to deduce the feeling behind the data, i.e. the state of mind of the investor at the time when he generated the information.

Data mining can be done in various ways, more or less efficient. So far, articles use statistics to analyse and draw conclusions from their analysis. It is also possible to use artificial intelligence with deep learning (or machine learning) for example.
"One of the greatest changes brought by big data concerns financial markets. For example, many works have been implemented using sentiment analysis, i.e., whether a piece of the text reflects a positive/negative sentiment". Bollen et al., 2011; Corea and Cervellati, 2015.

II.2.1. What research has been conducted in the area of big data and risks analysis?

Tools of Big data had allowed studying the mechanism of outbreak and contagion of systemic risk in the financial area. The European central bank defines systemic risk as "the risk of experiencing a strong systemic event. Such an event adversely affects a number of systemically important intermediaries or markets". The trigger of the event could be an exogenous shock (idiosyncratic, i.e. limited in scope, or systematic, i.e. widespread), which means from outside the financial system. Alternatively, the event could emerge endogenously from within the financial system or from within the economy at large. Many authors use Big Data analysis in order to detect systemic risk. Contrary to those who use homogeneous data in aims to have a trend, here the interest is the heterogeneous financial information.
As Shiller (2000) rightly said, "The history of speculative bubbles begins roughly with the advent of newspapers. (...) Although the news media present themselves as detached observers of market events, they are themselves an integral part of these events. Significant market events generally occur only if there is similar thinking among large groups of people, and the news media is an essential vehicle for the spread of ideas". Through this, we can see the importance and above all the influence of the media on public opinion. Overwhelming the Internet and newspapers with headlines in the same direction (positive or negative) will inevitably orient public opinion and thus also influence the market.
Based on this, some authors have chosen to study Big Data and mainly those of the media in anticipation of systemic shock and more particularly in the influence on systemic risk. R. Nyman et al (2018) have implemented an algorithm that analyses a large number of financial market text-based data. The goal is to see how narratives and sentiment play a role in driving developments in the financial system. The authors determined that the formation of high levels of sentiment like anxiety is coming just before the global financial crisis. In this article, the authors relate the link between narratives, sentiment and systemic risk.
P. Cerchiello and P. Giudici (2016) also attempt to create a model that quantifies systemic risk and predicts shocks using Big Data. The authors focus on how bank default risks are transmitted among banks. They presented a systemic risk model based on big data and showed that such a model can shed in light the interrelationships between financial institutions. The data used are those from the financial market and financial tweets. This analysis shows that big data whose tweets are predictive in the development of a model that measures systemic risk. Thus, it is possible to assess the risk of contagion between banks and financial institutions through Big Data.
P. Sarlin (2016a) focuses on the context and theory of visualization and visual analysis of information as a potential means of risk communication in macro-prudential supervision. The author concludes by saying that two essential tools needed to assess systemic risks are the analytical visualisation of big data and interactive interfaces.
P. Salin (2016b) tells us that there are 4 main areas that give us an idea of financial risks and more specifically systemic risks. These are machine learning, networks analytics, investment management, simulation and fuzzy systems. The author focuses on machine learning, one of

the tools of the Big Data, the goal is to combine self-organizing maps and graph-based approach to project financial correlation networks.

We can, therefore, say that one of a hot topic in terms of Big Data and risks concerns analysing and mapping systemic risk.

II.2.2. Is information from the Internet included in the price?

Daniel E. O'Leary's article (2012) helps us understand how and why Big Data generates the Internet of Signs. An example to understand the Internet of Signs and Big Data, in a car park, the use of sensors allows us to know or not if there is any space left. The data is the number of available spaces, i.e. the number of occupied spaces. Another piece of data is the location of the car park, let's imagine in the city centre. The sign behind this data is that if the car park is full, an event will occur. Historically, semiotics has focused on the analysis of information generated by man, that is, the signs that man receives through his behaviour. Today, thanks to Big Data, signs based on this information are integrated throughout the Internet and this data leads to signs. They can be present in different Internet media, such as blogs, wikis, comments, Twitter messages, YouTube, etc. Another example, a summary of activity on social networks gives signals of what is happening or will happen by showing which events are considered by humans to be interesting. Therefore, the signs are interpreted not as the result of the direct behaviour of the human being, but as the result of the behaviour of the human being through the data generated by his or her behaviour.

Consequently, it is no longer necessary to observe the human being to understand and predict his behaviour, but simply to observe a flow of data.

N. Jegadeesh and S. Titman (1993) have shown that past performance is predictive of future returns. The analogy can be made with micro-blogging and Twitter posts - the messages that have been posted can predict future returns. Thus, many authors have sought to highlight the link between twitter posts and the financial market. One of the first examples of the use of unstructured data in stock price prediction dates back to 1998 with the analysis of P. Wysocki (1999). He studied the relationship between the volumes of messages posted on Yahoo! Finance and the stock market performance of more than 3,000 listed stocks. The author finally highlighted the fact that the volume of messages posted the day before predicts the change in the value of stock returns the next day. G. Ranco et al (2015) were able to determine that there was a link between Twitter sentiment and abnormal returns during the peaks of Twitter volume. This is true both for expected Twitter volume peaks such as quarterly announcements, but it is also true for less significant events. In this study, the authors adapted the "event study" to the analysis of Twitter posts. Then, the authors showed a link between stock price returns and Twitter sentiment in tweets about the companies.
J. Bollen et al (2011) go further in the analysis of market sentiment; the authors analyse the impact of mood on the market. Here, the mood is inferred from the Tweet posted. It emerges that the accuracy of the Dow Jones Industrial Average prediction can be significantly improved by taking mood into account. The lexical field of the words has its importance so, the use of pejorative or meliorative words influences the future stock returns as shown by Loughran and McDonald (2011), the more negative words are reported in the posts for a given stock the more this stock is associated with low returns in the future.

Furthermore, H. Chen et al (2014) highlight that peers' opinions are more listened to than professional opinions. The paper investigates the extent to which investor opinions transmitted via social networks impact future stock prices but also earnings surprises. Thus, views and comments predict future stock returns and earnings surprises. According to Cogent research 2008, more than one in four adults in the U.S says they consider investment advice published on social networks. Along these lines, we can still conclude that the Internet has an increasingly important place as a source of information but especially in the integration of information by investors.
Moreover, W. Sanger, T. Warin (2016) have highlighted results about the quality of the information. One of their results is that tweeting by professionals using tickers provides more useful information than tweeting by ordinary people who do not use tickers. On another note, financial tweets have more impact on predictive models, regarding the fundamental value of a stock, but tweeting can highlight the sentiment of investors.

From these two studies, we can see that the opinion of peers is more listened to but that their information is of lower quality. Therefore, we see here that information from tweets generates noise.

Finally, thanks to the authors F. Corea and E.M. Cervellati (2015), among others, it has been proven that advanced technologies allow for more accurate predictive accuracy. This article is more up to date than the previous ones because technology has evolved faster, so new techniques and new data will allow us to complete the research already done on the link between investor sentiment and stock price variation. The study shows that it is possible to infer some important insights on how to incorporate new information into trading and predictive models. M. Skuza's paper, A. Romanowski (2015) continues on technological advances and the authors also evaluate the predictive power of tweets on stock prices but using machine learning to rank sentiment and best predict stock prices. Calculations were made with MapReduce. This allowed to further improving the predictive power of the model.

Although Twitter is very popular and offers almost instantaneous communication, there are other sources of information such as stock message boards (Antweiler and Frank, 2004), blogs (De Choudhury et al. 2008), online newspaper articles (Schumaker and Chen, 2009; Lavrenko et al. 2000), security analyst recommendations (Barber et al. 2001), web search queries (Bordino et al. 2012). This last source has been very widely studied, as can be seen from the various searches that follow: Da et al (2011), in this article, the authors propose an indicator for direct measurement of investor attention. Attention is a key element in the asset-pricing model, but it turns out that attention is a scarce cognitive resource (Kahneman (1973)) because investors have limited attention, which can affect price dynamics. Understanding how investor attention is formed on the Internet can help improve predictive models. Thus, the authors to measure attention use the volume of search in Google. Google is the most widely used search engine, hence the study of it. So, the searches made on Google are representative of the searches made in general i.e. the general state of mind. Searching the Internet is direct evidence that attention is paid to a given subject. For the authors, the volume of ticker searches will induce an increase in the demand for action for the corresponding company. Authors, Barber and Odean (2008) explain this by the fact that on average only buyers are looking for information because they have to choose among several alternatives. By comparability, they will glean information from the Internet to help them make their choices among many alternatives. And the more they look for a particular ticker

the more interested they are in buying that stock. On the other hand, when an investor wants to sell, he already owns the stock so he will not research the stock he wants to sell.

Consequently, an increase in research should on average lead to an increase in stock prices. Finally, unlike indirect proxies such as turnover, extreme returns, news or advertising spending. The authors propose a direct measure of investor attention (especially individual attention) and show that the volume of research is directly related to attention.
In this way, adding analysis from other sources of information can contribute to more accurate forecasts.

K. Joseph, M. Babajide Wintoki, Z. Zhang (2011) expand on other research on the link between the volume of research on the Internet and the forecasting of future prices. The authors examine the predictive power of online ticker searches to predict abnormal stock returns and trading volumes. This provides an indication of what search intensity can mean. The results also confirm what was found by Da and Al (2011) but add a dimension that confirms the work of Baker and Wurgler (2007). For instance, the authors have shown that the sensitivity to stock returns differs from that of volatility returns. In other words, research intensity is greater for stocks with higher volatility because they are more difficult to arbitrage. Therefore, the sensitivity of returns to the search intensity is lowest for easy-to-arbitrage because volatility is low.
According to K.Joseph, all of these three papers together enable us to state "the intensity of searches for ticker symbols serves as a valid proxy for investor sentiment, which in turn is useful for forecasting stock returns and volumes".
M. Bank et al (2011) also validate the previous results, it comes out of this study that an increase in Google searches is associated with an increase in trading activity and also stock liquidity. For the authors, the increase in liquidity comes from the fact that the costs associated with information asymmetry decrease. They finally conclude that measuring the volume of searches allows us to measure the attention of uninformed investors. This article complements the previous articles in the sense that this time investors are characterized as uninformed. Thus, the results obtained are consistent with those obtained by Da et al (2011), who determine that the volume of research allows highlighting the attention of investors for a specific type of share or company.

Other authors have shown that Google searches can predict investor sentiment. In the article of Da et al (2015), unlike the others previously seen, the authors use the volume of internet research to create an index as a measure of investor sentiment, this index measures household fears about the economic environment, the keywords used are related to the lexical field of fears in economics "recession", "bankruptcy", "unemployment" and "gold" for example. This index has enabled them to predict short-term return reversals: an increase in the FEARS index (financial and economic attitudes revealed by search) corresponds to low market-level returns for the day but high returns for the following days. This index also predicts the increase or decrease in volatility and also predicts mutual fund flows out of equity funds and bond funds. Thus, one can measure economic uncertainty by seeing how often certain keywords are searched for, those who normally reflect an uncertain state of mind and lead to an increase in market volatility. The volume of enquiries is thus positively related to uncertainty and negatively related to investor confidence for all types of investors. In another hand, it also shows the importance of the reversal effect, resulting in increased volatility in the market. Thus, investors react to uncertainty by selling risky assets and asking for premium risk. Another index that shows investor confidence in the markets is the fear and

greed index of CNN money. Investors are driven by two main types of emotions: fear and greed. This index is based on the fact that excessive fear will lead to stocks trading well below their intrinsic values and excessive greed will lead to overvaluation of the asset. To do so the index evaluates the stock price momentum, the stock price strength, the stock price breadth, the put, and the call options, the junk bond demand, the market volatility and the safe-haven demand. These indicators are very useful to know if the stock could be over or undervalued.

Dzielinski (2012) showed also, that in the face of uncertainty, investors intensify their research on the subject of uncertainty. We can, therefore, see from Google trend that the stock return will vary and volatility will increase. In this last paper, the author shows that the reaction is the same for individual and institutional investors. About the paper: "Can Internet search queries help to predict stock market volatility?" T. Dimpfl, S. Jank, as in previous works, also highlight the positive link between the volume of Internet search queries and the increase in market volatility.
In this way, the volume of research allows us to capture the attention of investors. W. Zhang, D. Shen, Y. Zhang, X. Xiong (2013) shows us that this is also true for Asian search engines. The authors have shown that the frequency of searching for stock names in the Baidu engines can be a new proxy for measuring investor attention. The empirical results show that quantified investor attention is an explanatory variable for abnormal performance even when trading volume is taken into account. Thus, they have shown that investor attention has a strong relationship with abnormal returns. Finally, they postulate that open-source information can improve the speed of information dissemination and besides make the market efficient. The authors study the impact of the amount of research on a type of stock and its relationship to the change in the price of that stock on the Chinese stock market. This overlaps with the Internet of signs seen above. The attention of investors on a stock is an indicator of the variation of the price of this stock. This paper is related to the paper by Da et al 2011 in the meaning that both study the impact of investor attention on the variation of stock prices. Here the proxy is the frequency of search for a stock name in the engine Baidu the second on Google. According to the authors "Baidu Index, based on Baidu Web searches by tens of millions of Internet users, provides more scientific, authentic and objective data, which comes from users' authentic retrieving record without any simulation". Note that in China the most used search engine is Baidu, that's why its analysis seems more relevant to evaluate the stock price in China.

Thanks to the literature review we could see that the Internet and more precisely most used search engine and social network are a huge source of data and that the information derived from this data is predictive of future stock returns. We also saw that the data from the Internet allows us to have an overview of investor sentiment, so that the Big Data tools can, thanks to the analysis of the mood transmitted by the data, know if the market tends to under- or overvalued assets but can also predict the future trend.
It is up to the analysts to get the best out of it with the help of artificial intelligence and more precisely using machine learning and deep learning tools.
Finally, now that we know that information from the Internet is integrated into prices, that this source of information is tremendous and that it allows access to data in real-time, we may wonder if this does not increase the efficiency of the market. In order to answer this, we must first look at what work has been done to measure market efficiency and performance.
As we have seen from the literature review, there is a lot of research that shows the influence of information from the Internet on prices. Numerous articles have proven that the

information that is posted on the networks and on the Internet influences the future prices of securities. Some research focuses on the link between Big Data and risk. That is to say, to what extent do Big Data tools allow a better understanding of risk?

Very little, if any, research focuses on the variation in informational risk arising from Big Data and how this can be a lever for fund managers. In other words, if informational efficiency is low, due to increased noise; because of big data, to what extent can fund managers' benefit of that. Or conversely, if informational efficiency is high due to big data, what are the consequences for the future performance of portfolio managers.
Here, the gap to be resolved is whether big data and its tools increase informational efficiency. And what are the consequences of portfolio managers' performance?
We will, therefore, see what has been done in the literature regarding the measurement of efficiency in order to be able to calculate whether it has varied with the arrival of Big Data tools. Then we will see what has been done in the literature with regard to measuring fund performance. This will give us indications on how to proceed in order to demonstrate a variation in informational efficiency with the emergence of Big Data tools and whether or not their use increases the returns of fund managers.

II.2.3. Measure of efficiency

The purpose of this literature review is to see how market efficiency has been measured: what are the tools, methods, and limitations. The aim is to compare the evolution of abnormal returns over time and also to measure information asymmetry, to validate or not the hypothesis of an improvement in market efficiency. Market efficiency is a central theme in financial theory, but it is still debated. E.F. Fama (1998) points out that so far, the literature has failed to prove market efficiency through the study of long-term return anomalies. The result of his research is that market anomaly arises from overreactions and under reactions of stock prices to information. Furthermore, E.F. Fama tells us that depending on the method used to measure long-term returns anomalies, they can tend to disappear. It is, therefore, necessary to adopt the right measurement method in order to avoid biased results.
Usually, market efficiency has been measured after and just before an event, there are very many event studies to see if and how the information is integrated. However, there is little new research on the evolution of information integration by the market over time; also in different markets. Until now, studies have mainly focused on a particular event and a specific market and not on a set of events with an overview of the different markets.
An approach is to see if gains in the market are possible. Assuming that an efficient market leaves no opportunity for gains, then no-arbitrage can be found. To prove this, no profitable trading algorithm could have been found. Kang et al (2002) proved that profit opportunities were possible thanks to momentum and contrarian strategies on the China stock market. They later showed that this was mainly due to information overreaction.
Irrational behaviour can be explained by behavioural finance. Barberis and Thaler (2003) speak of market inefficiency caused by biases and reasoning errors on the part of agents. Different tests have been made to measure market efficiency.

The main question remains how quickly and to what extent the stock price reacts to information. That is to say, how quickly the information is integrated and the price reaches the expected price. For this, we consider the assumptions of the accurate model price. These are the classical tests of information-efficiency, which present a joint hypothesis problem (the model imposes what the target price should be).

According to K.R.L Godfrey (2017), one way to avoid this problem i.e. joint hypothesis problem is to see if there are characteristics that proxy for the desired phenomenon. And the most widely used proxy for assessing market efficiency is random walk theory; an efficient market is one that follows a random walk in which returns cannot be predicted (Samuelson's (1965)). Thus, if we find a presence of serial correlation or return predictability then we can say that the market is inefficient. With this approach, it seems more complicated to quantify the degree of efficiency of a market. It would be more in the order of knowing whether a market is efficient or not because of its randomness or not. Nevertheless, it is tricky to say whether it is more or less random and therefore more or less efficient. Random walk and simulated trading can be done without the basis of the informational model. Thus, these different methods mainly provide evidence that a market is not efficient but not that a market tends to be efficient.
K.R.L Godfrey seeks for a new measure of market efficiency by addressing the problems outlined above. Thus, he favoured a numerical scale model rather than a binary model. Moreover, it can be calculated from available information and not from predicted values. Therefore, the author eliminates all the barriers shed in the light before.
Information is used to obtain the value of a security. Fundamental pressures, financial behaviour plus non-fundamental pressures (noise trading) create the price. Fundamental pressures create efficiency, bias creates inefficiency and non-fundamental pressures can be both depending on the trading strategy chosen.
Consequently, the method used here by K.R.L Godfrey is to compare the performance of passive and active trading. If the market is efficient, active traders should not outperform passive traders. The results are as follows: when transaction costs are high, there are fewer opportunities to make profits and market efficiency is high because active and passive trading are close to each other.
Furthermore, market efficiency increases with diversification. The less volatile the market is, the more efficient it is. The more volatile the market is, the less we tend to have a strong form of efficiency. There is an inverse relationship between volatility and efficiency.
It helps to know how efficient a market is and not if the market is efficient or not. We can gauge the efficiency of a market.

According to the literature review, we have seen that three main categories can send back a signal i.e. price information. The first is fundamental analysis, with the analysis of relative valuation, financial quality and non-financial quality through quarterly earnings calls, company presentations, and product releases, for example. All this data can be available online and then read by an IA and analysed. Then comes sentiment, which can be analysed through the markets, with the trend but also the timing, to do this we can analyse data from social networks, research activities on the net but also from blog posts. Finally, there is the macroeconomic signal with the analysis of the country, the region, the industry but also the style. The information comes from prices, volumes, and flows but also from the press and business updates.
We have seen, for instance, that there is limited research on how market efficiency is achieved, particularly over time, and also if market efficiency tends to evolve favourably over time with the arrival of technology and the constant innovation. Therefore, the next step is to try to find out and to find ways to solve this gap.

Let’s see first the definition of market efficiency according to Grossman and Stiglitz (1980):”(…) prices reflect the information of informed individuals (arbitrageurs) but only partially so that those who expend resources to obtain information do receive compensation.

How informative the price system depends on the number of individuals who are informed". Therefore from this definition, we can deduce that efficiency is the amount of information revealed by price, the more there are arbitrageurs the more the price is informative. In extent, if we have no information asymmetry, then all information is revealed by price.
Prizes are partly revealing: they make it possible to highlight public information. In order to take advantage of this partial balance, it is necessary to acquire private information. Thus, in order to take advantage, fund managers will spend resources on information. This private information will allow them to increase performance. If the performance generated by the use of this information is higher than the cost of acquiring it, then the funds will create profits. And the market is not efficient. On the other hand, if the performance is equal to the acquisition cost, then the market is efficient.

How has information asymmetry been studied over time and how has it evolved with Big Data tools?

S. Clarke (2001) lists 4 major groups of proxies for measuring information asymmetry: analysts' forecast measures, investment opportunity set measures, stock return measures and market microstructure measures.
S. Krishnaswami and V. Subramaniam (1999) analysed information asymmetry by studying the forecast error in earnings measured before the announcement of the spin-off. Thus, firms for which the information held by the manager (cash flow and value) is greater than that held by the outside market, therefore the forecast errors will be high. The result is that following the spin-off, the asymmetry of information decreases. Elton et al (1984) first demonstrated the measurement of asymmetry by measuring errors in analysts' forecast of earnings. According to S. Clarke (2001), analysts' forecast measures are not the most appropriate way of measuring information asymmetry, since the riskiest firms have, by definition, greater variations in their cash flow and value, so that firms have more forecast errors.
K. Chung et al. (1995) highlight information asymmetry on the assumption that the firm with the highest asymmetry has a higher earning potential for those who have private information so that this type of firm will attract more analysts. The more financial analysts there are around a firm, the stronger the information asymmetry. Moreover, the authors also point out that the size of the spread set up by market makers is an indication of the informational asymmetry. The larger the bid-ask spread, the stronger the asymmetry. Market makers take more risks when asymmetry information is high.

The investment opportunity set measures are more criticized. Usually, authors such as R. McLaughlin et al. (1998) use the ratio of market value to book value of equity to see the asymmetry of information. Here, the authors analyse the relationship between information asymmetry and long-run changes in firm operating performance. They find that firms with more information asymmetry have larger post-issue performance declines. However, market-to-book measures are not directly related to information asymmetry. It is also a measure of corporate performance (Tobin's Q), growth potential (J. Gaver and K. Gaver 1993), risk (S. Penman 1996) and market power.

The stock return is another proxy for asymmetry and more precisely the residual volatility. S. Bhagat et al (1985) and DW. Blackwell et al. (1990) show that informational asymmetry can be detected with the analysis of the residual volatility in the daily stock returns. The residual volatility is a sign of uncertainty, so the less information we have about

a firm, the higher the residual volatility will be. The results of A. Kyle (1985) also support this view.

Finally to measure informational asymmetry one method is to focus on microstructure measures. As S. Clarke (2001) says, the bid-ask spread consists of three components: the order processing, the inventory and the adverse selection component. Part of the bid-ask spread is used to compensate the market maker for the risk it takes to trade with more informed traders.
Later the bid-ask spread was divided into two components: the part attributable to information asymmetry and the cost of inventory (plus monopoly rent and risk aversion).
D. Easly et al (1996) have shown that less active stocks are riskier because private information is more important for such stocks. Thus, as shown by Y. Amihud and H. Mendleson's (1986) the risk increases with the bid-ask spread. Moreover, it appears that small market values presentment more asymmetry because there is more informed trade.

From these different studies, it appears that the most relevant measure is the bid-ask spread. Another frequently used measure is the firm-specific return variation (FSRV). It is a measure of idiosyncratic risk in the way that it helps to know the informativeness of a stock price and so to gauge the informational risk of a firm.
A. Durnev et al. (2004) propose a tool to gauge how the information about a firm is quickly and accurately reflected in share prices. This is the firm-specific stock return variation (FSRV). FSRV is a measure of the integration of information into prices; it measures the degree of asymmetry among investors. In other words, it is a measure of the informativeness of stock prices. This tool is based on the work of R. Morck et al. (2000). The authors find out that "Among developed economy stock markets, higher firm-specific returns variation is associated with stronger public investor property rights".
Better protection of property rights could, therefore, make firm-specific risk-arbitrage more attractive, so that strong property rights promote informed arbitrage.
From Grossman and Stiglitz (1980) we know that a low cost of private information leads to a higher intensity of trades. And finally, the price will be more informative. Therefore, high firm-specific return variation is linked to higher informative prices: high FSRV leads to low asymmetry.
From A. Durnev et al. (2004) the firm-specific return variation is measured as follow:

$$r_{i,j,t} = \beta_{j,0} + \beta_{j,m}\, r_{m,t} + \beta_{j,i}\, r_{i,t} + \varepsilon_{i,j,t},$$

$r_{i,j,t}$ is the return of the stock j over the period t,
$r_{m,t}$ is the return of the market portfolio,
$r_{i,t}$ is the return of a value-weighted portfolio for securities of the same industry,
$\varepsilon_{i,j,t}$ is the noise term.

This latter, the noise term is the firm-specific part. According to R. Burlacu, P. Fontaine and S. Jimenez-Garcès (2005), the specific risk is the variance of the firm-specific part relative to the total stock variance. In mathematical term, it is one minus the R2 obtained from the regression see above.
Therefore the measure of FSRV is :

$$FSRV_i = ln\left(\frac{1 - R^2 i}{R^2 i}\right)$$

The movements in stock market returns can be explained by three factors. One coming from the market in general and related to systematic risk is called macro factors. The second, which is the micro factor, comes from the risk inherent in the type of industry but which is also a systematic risk and the last is the specific risk resulting from a particular company. We talk about specific factors. The price movement, therefore, comes from public information as regards the market and the industry and from private information as regards the company. Thus it is possible to know how much information is revealed by the price thanks to the ratio between the specific return variation and the total return variation. R. Burlacu, P. Fontaine and S. Jimenez-Garcès 2005.

This measure, therefore, gives us a measure of the extent of information asymmetries. In order to obtain price informativity, the size factor must be taken into account. Therefore, the higher the FSRV, the higher the share of returns not explained by systematic and specific factors. In other words, the higher the FSRV the stronger the information asymmetry. Demonstrating informational efficiency is equivalent to demonstrating a low FSRV.

II.2.4. Measure of performance

There are many performance measures for investments. Modern portfolio theory has highlighted the link between the risk of a portfolio and its return. The first performance indicator was developed using the Capital Asset Pricing Model (CAPM) developed by Sharpe (1964) and then the risk-adjusted ratios: the Sharpe ratio (1966). Treynor's ratio (1965) and the Value at Risk (VaR) based measure are also absolute risk-adjusted performance measures. Then measures relative to the benchmark were developed to overcome the limitations of the first models, which are absolute measures. This is the set of alpha.
With the evolution of the CAPM, the risk measures have also evolved in order to have measures that are more and more accurate in practice and not only in theory.
It is important that the choice of performance measures is simple, fair and effective in order to understand the resulting information.

Table 1
Jigsaw puzzle of basic risk-adjusted performance measures

interpretation \ risk	total risk	market risk
Ratio dividing the excess return of the fund by its risk	Sharpe Ratio (SR)	Treynor Ratio (TR)
Differential return between fund and risk-adjusted market index	Total Risk Alpha (TRA)	Jensen Alpha (JA)
Return of the risk-adjusted fund	Risk-adjusted Performance (RAP)	Market Risk-adjusted Performance (MRAP)
Differential return between risk-adjusted fund and market index	Differential Return based on RAP (DR^{RAP})	Differential Return based on MRAP (DR^{MRAP})

H. Scholz and M. Wilkens (2005)

See **Appendices** for more information about measures of performance.

III. Research Project

We are attempting to determine whether the information tends to be perfectly and instantaneously integrated into prices; in other words, whether we are moving towards an efficient market. In addition, if and how performances are impacted.

An efficient market is an effective market. There are two main components in the market efficiency, first, an efficient allocation of resources with an optimal distribution among agents and second, an informational efficiency, i.e. how much information is revealed by prices. Thus, in an efficient market, the price must reveal all available information and must not allow for gains, i.e. profitable trading opportunities.
However, agents have bounded rationality, tools were not efficient enough to allow comparability and some agents had private information. Markowitz established the theoretical framework for the financial asset-pricing model and strong assumptions were made. An investor will act rationally to maximize returns while minimizing risk. A few years later, Fama assumed that the information efficiency of the market is characterized by the fact that there should be no profit opportunities in the financial markets, which are defined by random market models, and that all information is known at all times because prices instantaneously reflect new events. These two assumptions have shown their limitations: Firstly, because there is information asymmetry, private information: in theory, there is no way to beat the market unless one investor has more information than another. Thus, thanks to Big Data technologies, it is questionable whether it is possible to extract additional information from an almost unlimited amount of data. The latter will provide additional

information that can help beat the market, i.e. succeed thanks to the technologies in detecting private information, information that only certain investors were aware of.
On the other hand, if, with these technologies, we too come to possess information that has been concealed until now, then everyone will be able to have this information and the information that was initially private will no longer be private; no one will be able to beat the market.
Secondly, because an agent is not rational, or at least has limited rationality. The emotional involvement of agents is a response to limited rationality. Shiller (1981) found that the volatility of stock prices was far too high to be attributed solely to the entry of new information. Therefore, this excessive volatility calls into question the informational efficiency hypothesis. This excessive volatility of stock prices is attributed to a reaction of investors to information that is not related to fundamentals (Shiller, 1989; Barberis and Thaler, 2003). High price volatility is due to the presence of irrational investors. These latter make asset valuation errors that are difficult for rational investors to correct.
Behavioural finance interprets financial markets to reflect social or "animal spirits", we are not rational, and most of our decisions are influenced by our "animal spirits" J.M Keynes, 1936. Emotions, in addition to information, play an important role in human decision-making, therefore, understanding human reactions to uncertainty, the main source of irrational behaviour, can then allow us to anticipate investors' reactions. Advances in behavioural finance research will improve predictive models of financial markets. Including the results in Big Data technologies could then make it possible to have an even more accurate predictive model.
Here, we would not arrive at a perfectly efficient market in the Markowitz sense, but rather thanks to the fact that we know that agent has limited rationality and knowing this, we can anticipate his reactions, then, at that moment, we will manage to find the equilibrium price.

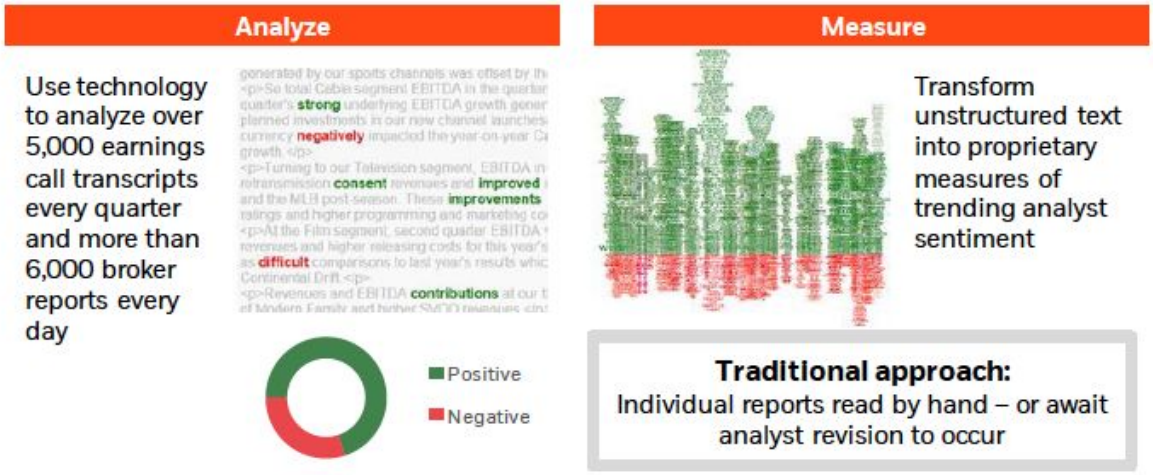
Until a few years ago, data on economic activity were not readily available. But the situation has changed radically in record time. One of the reasons for this is the growth of the Internet and its technological tools. Virtually everything on the Internet is recorded. When you search on Google or another search engine, all queries and clicks are recorded. When shopping on Amazon, eBay, Cdiscount or others, every purchase, every click, and every item viewed and compared is captured and recorded. Reading an online diary, watching videos on YouTube, tracking personal finances, each behaviour is recorded. However, recording individual behaviour is not limited to the Internet, SMS, mobile phones and geolocation are analysed around the clock, just as scanned data, employment records, and electronic medical records are all part of the data set that every individual using the technology leaves behind. As a result, the sphere between private and public no longer exists. The world tends to automate its tasks more and more, the amount of data available will continue to grow and the information about companies will continue to increase at the same time.
For example, studies tend to show the value of making accounting and auditing tasks automatic, through software, which will reduce the time of action but also reduce fraud. All these financial data recorded will then be available in real-time and with the help of the BI software, it will be possible to see the health of a company and calculate its stock market value at the same time: instantaneously. The automation and digitization of tasks will allow access to all data and therefore to extract the most optimal information thanks to a constant improvement of the dimensions of the Big Data, i.e. volume, velocity, variety, and veracity. Consequently, one may wonder if these technological advances and this access to an infinite field of information would not make a greater informational efficiency? Are markets tending towards a strong informational form as information technologies evolve?

To epitomize, first, we showed that information from the Internet was integrated into prices, that investor sentiment allowed us to anticipate future prices. We also know that for professional investors, the main source of information comes from specialized platforms (Bloomberg, Reuters).
Furthermore, we know that Big Data technologies can gather all types of information, in gigantic quantities, anywhere, anytime. And that it is even possible to anticipate previously unpredictable macroeconomic events thanks to early indicators extracted from social media and the Internet (blogs, Twitter feeds, etc.) that can help predict the evolution of various economic and business indicators.
Therefore, all these elements may lead us to believe that all information, whether public or private, will be contained in the Big Data (in one way or another, more or less legally or by respecting rights).

As we have said, if a market is not efficient it is possible to make profits because there are profit opportunities due to information asymmetry.
Thus, to see if the market tends to be more efficient we could study abnormal returns over a large window of time. If these returns decrease over time, we can say that we are trending towards a more efficient market. And if it increases we can say that Big Data makes more noises and enhance trade and arbitrage opportunities.
Moreover, as we have seen, excessive volatility is a sign of lack or low informational efficiency.
Hence, seeing whether this volatility decreases could point to an increase in informational efficiency or in the opposite confusions between noises and signals.
Moreover, it would be wise to separate the periods of crisis in the analysis, because, during a crisis, agents exhibit even more irrational behaviour.
Abnormal return is the difference between the actual return and the expected return of a security. With the technologies of Big Data, in theory, both should be equal. Big Data strengthen the capacity of prediction of the expected return. And if the technologies dominate the market and their decisions are their own (deep learning, machine learning) then logically their expectations should be equal to the intrinsic value of the asset.

III.1. Example with Blackrock

Blackrock is the world's largest asset manager. What interests us here is that Blackrock has implemented a system of Big Data and artificial intelligence tools called SAE (systematic active equities), it provides distinctive access to global investment opportunities. SAE is based on the observation that the data generated is exploding, that more and more information is needed to make decisions and that the sources of information are constantly increasing. SAE technology takes into account all sources of information in order to establish a price and choose an investment strategy. It takes into account the combination of fundamental value, sentiment, and macro themes. All this is achieved through the analysis of structured and unstructured data.
This technology is a concrete example of the theories seen upstream. This technology has only recently been implemented because it requires considerable investment, know-how, techniques, and advanced technology.
We can see above, how the use of AI-related Big Data allows us to see the feeling that results from the earnings call. How unstructured data are transformed into global sentiment.

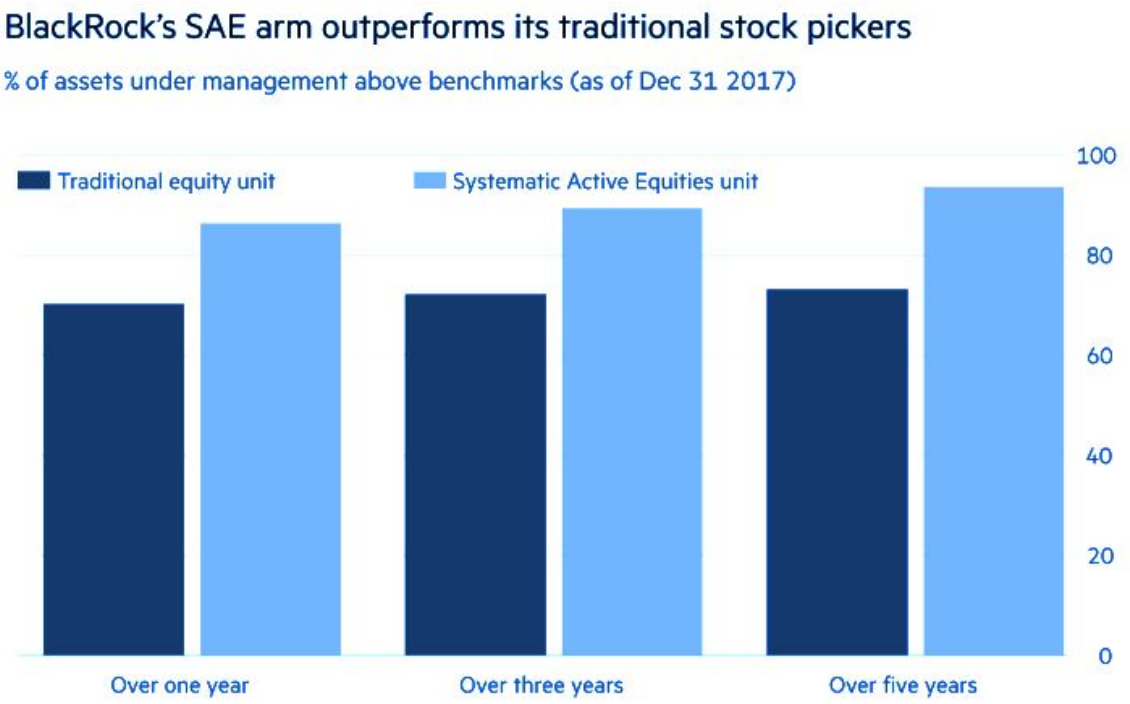


Source: Blackrock

As another example: analysts have to go to a company to investigate the activity or read all the broker reports or earnings announcements to have an idea of the value of a company. Machines can read at a faster pace and in any language and use satellite data to analyse the activity of a company. The large data technologies can be applied in any field and therefore it allows to understand, recognize and master all topics, all companies and therefore to invest where alpha is the highest.

Blackrock in its SAE solution uses consumer sentiment and creates consumption models with the help of Internet searches from websites, reports, comments, and social networks. They get millions of pieces of information captured per second. Blackrock has created its own research laboratory dedicated to artificial intelligence and more specifically to research the optimal use of Big Data technologies. As we can see, these technologies create returns, Blackrock's objective is to generate alpha no longer with traditional techniques but by taking advantage of the abundance of information and technological innovations.

The objective of SAE is to analyse Big Data using technology, artificial intelligence and more specifically machine learning. Blackrock's objective is to constantly improve investment results. In addition to its SAE technology, Blackrock combines its software called Aladdin to minimize the risks of its portfolios. The data allow anticipating and what better way than anticipating risks avoiding them. So we can optimize returns and minimize risk by combining these two tools.



Source: Blackrock

As can be seen, Blackrock's SAE unit outperformed traditional Blackrock share units in both the short and long term. 93% of SAE assets have outperformed their benchmarks over the past five years.
This means that large data technologies allow for superior returns. These superior returns are possible due to better information holding if one follows the assumption of market efficiency and if the information is not instantly and fully integrated into prices.
Furthermore, we can see that volumes have to be increasingly large to generate equal or even lower returns (compared to the past), which means that we tend to have a more perfect market because the profit opportunities diminish over time: we have to increase the volume of transactions to expect at least equal returns.
In addition, because the SAE unit outperforms the traditional unit of shares, we have a second proof that Big Data technology allows us to have superior information and beat the market.
Although in the future, if all investment companies and investors use the same information-based approach to the market, the market will tend to be perfect: as soon as the set-up costs, ease of access and skills are adapted, management companies will all use the same tools. In the end, the information can be the same for everyone.
This is an ideal, however, as technological advances continue to evolve and the competitive advantage of some will continue to make the difference in terms of performance and so in term of market efficiency.

III.2. Research questions and hypothesis

Firstly, we will propose methods to evaluate the quality of the market: we will calculate abnormal returns for the longest possible period of time. It may also be interesting to do this in different markets in order to find out if there are differences in technology or any other explanation.

Then, we will see whether, according to the data obtained, abnormal returns decrease over time.
Finally, we will conclude on the increase in informational efficiency thanks to the growth in the use of Big Data tools.

In a second experiment, we will see whether volatility has decreased over time.
If information increases, the over/under reaction should normally decrease, especially since it is possible to anticipate irrational reactions thanks to behavioural finance. Thus, alongside the study of abnormal returns, we can also study volatility over time. We know that risk in the markets is associated with high volatility and that low risk means low volatility. We also know that the less volatile the market is, the more efficient it is. The more volatile the market is, the less efficient we tend to be. There is an inverse relationship between volatility and efficiency. So it helps to know how efficient a market is or is not. We can gauge the efficiency of the information.
So information helps to reduce uncertainty, and in turn, this reduces volatility. So little information means a lot of risks and a lot of information means little risk and therefore low volatility. Hence, if markets are efficient in the informational sense then volatility should be relatively low.
So we will see if over time the volatility of the markets tends to decrease. If from the analysis we see that overall volatility tends to decrease over time then we can potentially conclude that Big Data technologies can increase the quality of the market through a decrease in uncertainty due to an increase in the amount of information.
The goal is, therefore, to see whether market efficiency has changed over time. But also to compare this evolution with the different markets. Information reduces uncertainty since information reduces volatility.
To do so, we will take the different volatility indices proposed by the major world markets and see the trend that emerges, i.e. see if the trend curve's steering coefficient is positive, negative or zero. We will also have to see if this is significant.
We will then be able to conclude or not on an increase in the efficiency of the markets with a decrease in volatility. Therefore, if the volatility decreases it means information tend to lead to a more efficient market but at the opposite we can say that more information tend to lead to more noises.

In a third experiment our aim will be to measure informational asymmetry by measuring the information risk. We want to know if the publication of information on the Internet influences information efficiency.
In other words, if the profusion of information decreases asymmetry and increases efficiency or on the contrary, the profusion of information creates noise and decreases efficiency.

III.2.1. Securities analysis

Following what has been seen previously, we propose several hypotheses that can be verified through the analysis of the securities. Our hypothesis are based on the assumption that new technologies allow to tend to a more efficient market. We can therefore say that:
H1: New technologies reduce abnormal long-term returns,
H2: New technologies reduce volatility,
H3: New technologies reduce information risk.

III.2.2. Portfolio analysis

Based on our previous findings, we can now test whether new technologies lower the performance of active mutual funds.
Assuming that market efficiency increases, the earning opportunities for fund managers are diminishing.
H4: Informational efficiency reduces the performance of fund managers.

III.3. Methodology and Expected results

III.3.1. Temporality

III.3.1.1. Abnormal returns

Data descriptions:

The sample is made up of all CAC 40 companies and the CAC 40 index for the period from March 1, 1987, to December 31, 2019.

Methodology:

Methodology derived from equity strategy research. Value Relevance of Analysts' Earnings Forecasts September 1, 2003.
Expected returns should be equal to the returns. On a long return window.

Calculate daily abnormal returns (AR) for each firm for each year. We have to be careful because, during the whole period, the firms are not the same.

- Calculate the daily stock return such that
  $R_{i,t} = Ln(Pd/Pd-1)$
- We calculate the daily-expected stock return (we can use different models, the Fama and French 3 factors models would be the best because it is the most accurate).
  $E(R_{i,t}) = Rfi,t + ßi,1(E(Rm,t) - Rfi,t) + ßi,2(SMBt) + ßi,3(HMLt)$

The data will have to be found for each firm in the CAC 40. For simplicity it is possible to do this with the CAPM, $E(R_{i,t}) = Rfi,t + ßi,t[(E(Rm,t) - Rfi,t)]$ but here too the Rf for each period will have to be determined, taking the 10-year OAT, and the ß corresponding to each of the firms for each of the periods.

- Calculate the abnormal returns such that
  $AR_{i,t} = R_{i,t} - E(R_{i,t})$
- Calculate the average abnormal return (AAR) for each year all over the period for the N stocks. It helps to eliminate idiosyncrasies in measurement due to particular stocks.
  $AAR_t = 1/ NN\sum_{i=1} AR_{i,t}$
- It may also be interesting to calculate the cumulative average abnormal return (CAAR) for each year. CAAR helps to get a sense of the aggregate effect of the abnormal returns.

It will then be necessary to do the same for other markets in order to allow comparability around the world.

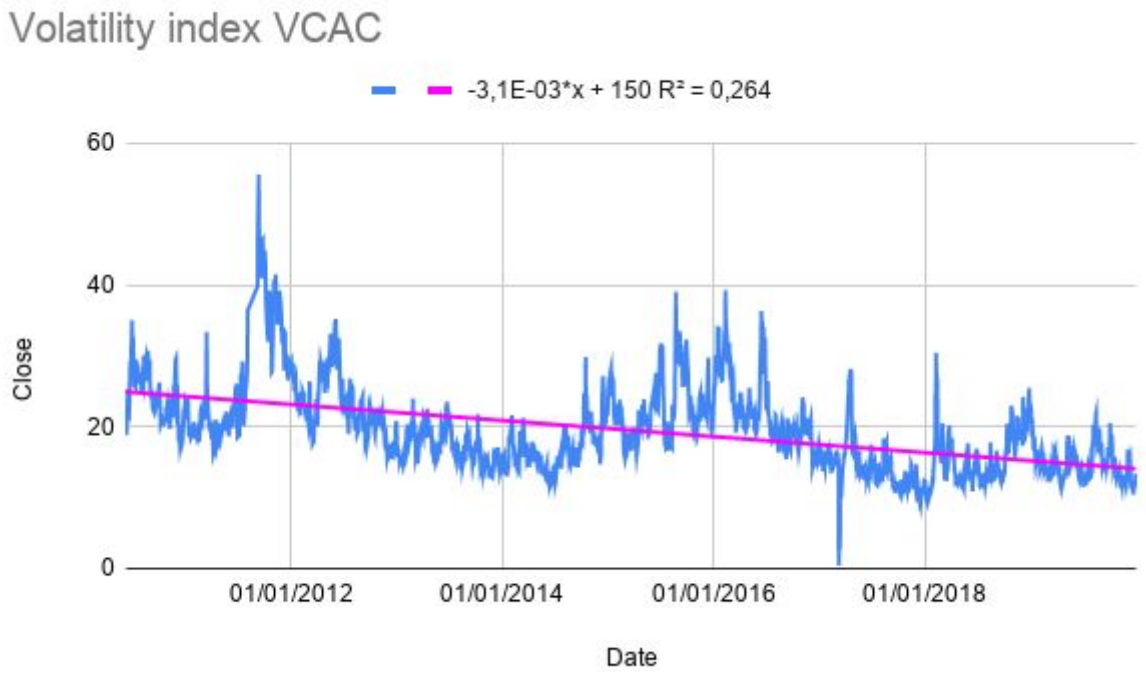
III.3.1.2. Volatility

Data descriptions:

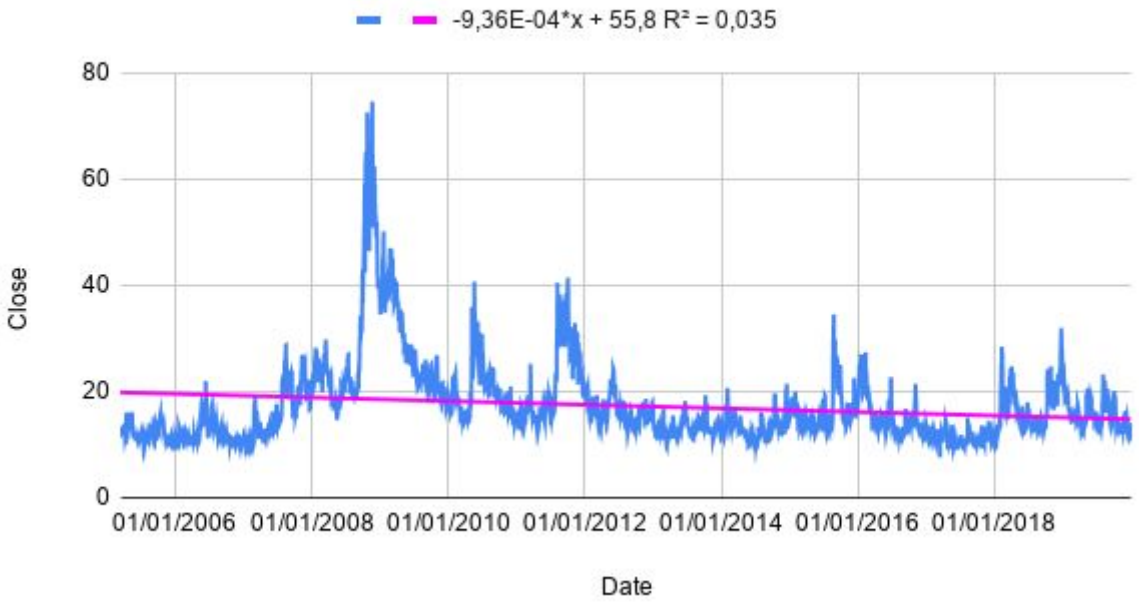
Using the volatility index of the world's major stock exchanges over the entire period available. Determine the trend curve and study its direction and significance. Due to the lack of available data, we were only able to experiment with a few indices.

Methodology:

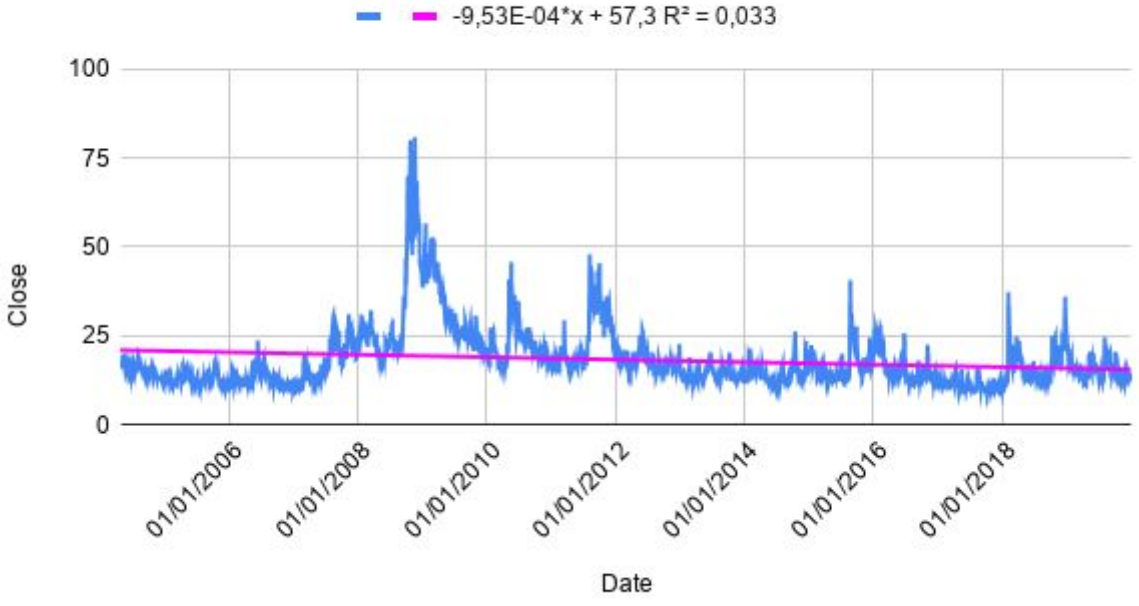
We collected volatility index data for the main markets: VCAC for CAC 40, VXD for DJI, VIX for S&P 500, VXFXI for China ETF. Then we plot the regression line. We study direction and significance. We conclude on the influence or not of Big Data technologies on market volatility.



Volatility index VXD

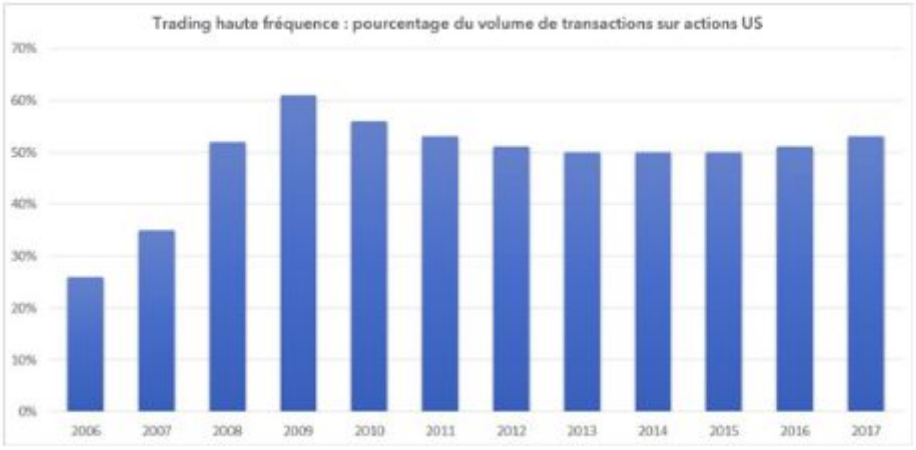


Volatility undex VIX



From the analysis of the different volatility indices, it can be seen that volatility tends to decrease. Nevertheless, it should be noted that the R2 is low, which means that our results may be questioned.
For some, like the CAV, the period begins during the crisis and therefore the volatility is much higher at the beginning.
Crises and so irrational reactions have a strong impact on the volatility analysis. It might be relevant to neutralize periods of crises.
Moreover, the study period is not long enough to be able to draw conclusions. The analysis begins when Big Data technologies were already on the market. It would have had to start at an earlier period. The impact of technologies is not yet sufficiently clear.
The experiment is inconclusive.

We can also link the results and draw a parallel with high-frequency trading. As can be seen in the graph, the percentage of US equity trading volume with HFT increased significantly and then stabilized.



(Source: TABB Group, Deutsche Bank, ResearchGate)

In the second graph, we can see that revenues from this high-frequency trading on the same market in the same years have decreased considerably.

(Source: TABB Group, Deutsche Bank)

The reasons for this are reduced volatility, costs but also greater competition. The latter means that the speed of market players tends to be the same because the use of this technology tends to become more homogenous.
Transaction volume increases but revenues decrease as the number of technologies increases. The objective of high-frequency trading is for computers to spot new trends and anomalies in the markets and act automatically before the rest of the market becomes aware of them. However, if all the players have the same technologies they will act at the same time and the information will then be shared by all instantaneously.

III.3.2. Information risk

Does Big Data increase informational efficiency? In other words, does information asymmetry decrease thanks to Big Data? Or is it the inverse relationship?

To see this relationship we will start from the hypothesis that the most tweeting financial securities give more information advantage, therefore for this latter, the information asymmetry is lower and the informational efficiency is higher.
In order to test this hypothesis, we will compare two companies. A first company whose tweet volume is very high, to do so we will take the most quoted company (high number of hashtags). Secondly, we will take a company of size, turnover and similar activity but whose activity on the Internet and especially tweeting is low.
In order to compare the informational efficiency, we will compare the degree of information asymmetry. Then, we will compare the FSRV for the very tweeter and the very low tweeter companies.

We conclude on the relationship between the amount of information available on the Internet and information asymmetry. In other words, to what extent does the publication of information on the Internet allow influencing information efficiency?
On the other hand, does the profusion of information decrease asymmetry and increase efficiency or on the contrary, does the profusion of information create noise and decrease efficiency?
It will also allow us to shed in light if tweeting allows for the release of private information or just creating noises.

Step 1: Select the companies and compute returns

Find very tweeted company and compute returns of the market portfolio one year then, the return of a value-weighted portfolio for securities of the same industry over the same period.

Step 2: Times series regression

We then do a regression.
We need to do the same with a low tweeted company. We will get two regressions.

$$r_{i,j,t} = \beta_{j,0} + \beta_{j,m}\, r_{m,t} + \beta_{j,i}\, r_{i,t} + \varepsilon_{i,j,t},$$

$r_{i,j,t}$ is the return of the stock j over the period t,
$r_{m,t}$ is the return of the market portfolio,
$r_{i,t}$ is the return of a value-weighted portfolio for securities of a same industry,
$\varepsilon_{i,j,t}$ is the noise term.

Step 3: Compute the firm-specific return variation

With these results, we will be able to compute the degree of asymmetry thanks to the FSRV.

$$FSRV_i = ln\left(\frac{1 - R^2_i}{R^2_i}\right)$$

We will have also two results of FSRV and by comparing we will be able to say if tweeted companies are more informationally efficient than low tweeted companies.
Following our reasoning, we should normally find that the most tweeted company presents less information asymmetry. New technologies reduce information risk.

III.3.3. Size

Hypothesis: there is more information asymmetry for small companies because there are more informed traders.

Step 1: Select companies and compute returns

We select all companies from CAC 40 in 2014. And all companies from the CAC PME in 2014. (2014 because this is the first year where CAC PME exists).
We proceed exactly in the same way but for the year 2019.

Step 2: Times series regression

We need to regress our results. In total, we will get two regressions for CAC 40: one for 2014 and one for 2019. Two regressions for CAC PME: one for 2014 and one for 2019.

$r_{i,j,t} = \beta_{j,0} + \beta_{j,m}\, r_{m,t} + \varepsilon_{i,j,t}$,

$r_{m,t}$ is the return of the market portfolio,
$\varepsilon_{i,j,t}$ is the noise term.

Step 3: Compute the firm-specific return variation

We compute the FSRV for the fourth regressions.

$FSRV_i = ln(\frac{1-R^2 i}{R^2 i})$

With the results, we will be able to say if asymmetry decreases with time but also if size matter in the asymmetry of information.
Normally, FSRV for CAC PME should be higher than FSRV for CAC because information asymmetry for small companies is higher.
Also, FSRV for the year 2014 should be higher than FSRV for the year 2019 because new technologies reduce informational asymmetry and new technologies increase over time.

III.3.4. Technology

Hypothesis: New technologies of Big Data generate more profits.

Here we will see whether the use of the profusion of information leads to higher profits.
If it leads to higher profits, then market efficiency is low.
Do the results obtain match those obtained in the previous experiment? Whether the previous experiment concluded that there was an increase in efficiency and the new experiment concluded that there was an increase in profits is a question that needs to be asked.

We select management societies that use big data tools such as Hadoop and some Data Lake for example.

We select the ones that do not use very much.
An analysis of the performance of the funds is carried out using the TRA.
We will obtain a TRA per fund. We compare. We conclude on the performance of funds using Big Data tools.

If $\gamma > 0$, then the Big Data tools can generate performance.
We project the result of the alpha at a given moment and we regress at each date.

Step 1: Select the funds

The funds will be classified according to their use of Big Data tools. To do so, we will define a scale from not at all to a lot. To calculate we can compare for example the storage surface of the funds. (Useful for Data Lake for example).

Step 2: We calculate the Total Risk Alpha for each fund.

$\text{TRAi} = \mu i - rf - \frac{\mu m - rf}{\sigma m} \sigma i$
$\text{TRAi} = \sigma i\,(\text{SRi} - \text{SRm})$

Each TRA is calculated for each fund for each year/month.

Step 3: Time series regression

We regress the data we get; we'll have as much regression as funds.

$$R_{1,t} = \alpha_1 + \beta_{1,F1} F_{1,t} + \varepsilon_{1,t}$$
$$R_{2,t} = \alpha_2 + \beta_{2,F1} F_{1,t} + \varepsilon_{2,t}$$
$$\vdots$$
$$R_{n,t} = \alpha_n + \beta_{n,F1} F_{1,t} + \varepsilon_{n,t}$$

F is the factor Use of Big data Technologie

Step 4: Cross-sectional regression

We regress the previous regression in cross-sectional.

$$R_{i,1} = \gamma_{1,0} + \gamma_{1,1} \hat{\beta}_{i,F1} + \varepsilon_{i,1}$$
$$R_{i,2} = \gamma_{2,0} + \gamma_{2,1} \hat{\beta}_{i,F1} + \varepsilon_{i,2}$$
$$\vdots$$
$$R_{i,T} = \gamma_{T,0} + \gamma_{n,1} \hat{\beta}_{i,F1} + \varepsilon_{i,T}$$

γ is the regression coefficient it quantifies the returns explained by the use of Big Data tools.

Step 5: Average coefficient

The regression coefficient is averaged for each fund. We will finally have a coefficient for each fund that will be the benefit or not of using the Big Data tools. The higher the coefficient is, the more the returns of the funds are explained by their use of the Bi Data tools.

$$E(R_i) = \gamma_0 + \gamma_1 \hat{\beta}_{i,F1} + \varepsilon_i$$

We can then conclude on the impact of Big Data tools in the ability to generate more performance or not.

This experiment will allow us to see if Big Data tools can generate more profitability. The goal is to see if funds that use Big Data tools outperform funds that do not use them.
We also want to see how much of the performance is attributable to the use of Big Data if they generate higher profits.

Conclusion

Based on the analysis of the theories as well as various experiences, I think it is still too early to talk about an efficient information market. The major information technologies are not yet fully present everywhere and at all times in the major markets. On the other hand, I think we can observe advancement in market efficiency. Although most studies focus on the efficiency of a market at one moment, we need to examine whether a market tends to be more or less efficient than in the past, and then see whether the degree of divergence tends to decrease in the markets.
It should be noted that non-fundamental pressures from noise traders are having an increasingly strong impact on markets. Volumes traded are increasing, but noise traders do not contribute to market efficiency. Noise traders are not involved in price determination. And it can be very dangerous if Big Data tools mislead signals and noises. Therefore, their presence will divert the market from its efficient form. However, without these noise traders, there will be no trade in the markets because there is no opportunity for profit, and it is, therefore, questionable whether they do not ultimately allow trade. We are faced with a dilemma: if the markets are efficient, the objective of arbitrage will no longer exist.

As large data evolves, artificial intelligence will be able to do its own fundamental analysis, but also to anticipate the behaviour and biases of individuals because these tend to be repeated in their anomalies. Limited rationality is due to the impossibility of comparing "all available baskets". Therefore, unlike humans, major data technologies have access to all data anywhere in the world. By analogy, Big Data technologies would be more inclined to anticipate major crises because a machine cannot have limited rationality in its decision-making processes. To be effective, Big Data technologies must take this component into account. Thus, thanks to machine learning, artificial intelligence is increasingly able to integrate human behaviour and therefore the biases it can show in its models.

Indeed, if Big Data technologies come to dominate financial markets, and knowing that forecasts cannot take into account their own impact, then, whether this is true or not, they

influence the market. The market can then follow a totally wrong trajectory if wrong indicators also lead the way. How can we be sure that the data collected is reliable?
I think that the two components that go into price formation help to answer this question. So the reliable data will be the fundamental value of the company, which is obtained by the calculation, by the financial analysis of the company. That is something true, something real. The other element that comes into play is the value of a company in the eyes of investors. You have to consider that the fundamental value will be a reliable figure, but it will not necessarily be the value of the company in the markets. As J.M Keynes (1936) showed with the beauty contest, the value does not depend on intrinsic value but on the perception that market players have of what the value of the company is. Here, value is no longer based on financial data, on real bases, but on anticipations and representations of the world, of a value that agents have.
It would then be more relevant to ask at what point one can know whether a representation of reality, of value, is reliable data.
In addition, technologies can anticipate unpredictable human behaviour in different ways. One of them is that thanks to the data that humans transmit, these signals give warning. The second is made possible by the understanding of humans and their psychological mechanisms in general and in markets.
I believe that if artificial intelligence technologies are to be increasingly effective, the indispensable contribution of researchers is needed upstream. Without research, it would not be possible to program intelligence and teach it behaviour. Data creates information, this information enables us to have the knowledge, and knowledge can be transmitted. For example, it is possible to map human intention using digital traces of human behaviour and to ask the machine to help us with this task.

As we have seen with high-frequency trading, transaction volumes increase significantly but not revenues because the information discovery process tends to become more homogenous.
Furthermore, it also tends to reduce the revenues of the players on the market. They are therefore constantly looking for solutions to overcome this. And to do so, they need to find new technologies that are ever more innovative and ever faster.
Consequently, as long as mankind is capable of innovation, the market will tend towards a strong form but will never reach it.

An important point to note is that Big Data technologies free humans from complex and repetitive tasks to allow the improvement of high value-added tasks. Artificial intelligence makes man smarter; man allows this artificial intelligence to become more and more efficient over time. It can, therefore, be said that the only limit to technological innovation lies in the limits of human imagination.

Therefore, large data technologies will allow for more accurate information, but also for anticipating risks, in theory.
Although, despite that, in reality, more information could mean more risks: “the less, the better”.

And more, against all intuition, the data will reduce the field of possibilities, as the future is already predicted by the data. Further, the data will all point in the same direction, which can be a big problem if the results are wrong, for example, a major crisis due to the formation of a bubble. Or worst an unpredictable event: randomly cannot be predicted.

Analysts will end up analysing the same data with the same tools and forecasting the same patterns.
It is very interesting to see that in the past, one of the limitations of decision-making was that forecasts were generally based on past data, both private and public. There is no guarantee that historical data; conditions in the past would be repeated in the future. A typical example is that of exogenous shocks. Until now, it has been impossible to incorporate unexpected events.
As D. Hume (1739) said: “No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion.”
The black swan was highlighted by Nassim Nicholas Taleb (2007), in his essay he defines the black swan as an unpredictable event with major consequences and whose event is a posteriori simplified. According to the author, "This a posteriori rationalisation comes from the fact that the information that would have allowed the event to be predicted was already present but not taken into account by risk mitigation programmes. The same is true for the perception of individuals".

Last but not least, I think that the increased use of technology and especially of programs based on the same models can greatly weaken the market. This inevitably leads to market fragility and contagion effects. We can see this with the increase in mini flash crashes. This is partly explained by the fact that the market is globalized and if there is a failure in one country it will be reflected in the others. Technologies and programs are not shock resistant because they are not foreseen in their programs (rare events that are part of the distribution queue). However, the number of technologies obeying the market and the models tend to increase and these rare events too. One can then wonder to what extent the increase in the amount of information and the inevitable noise weakens the market. But also the lack of reliable data about the financial system, and even more important the ability to understand available data. There are key factors in the inability to prevent the ongoing crisis.
Moreover, as technology becomes more widespread in financial markets, profits tend to decline. On the other hand, volumes are increasing to offset these "loss of earnings". This is based on the assumption that in order to compensate for this "loss" it is also possible to take more risks. Thus, one may ask whether the risk of crashes or failures does not increase considerably?

We may wonder if we are not moving towards a uniform market, using the same technology and the same strategies, increasing volumes and taking more and more risks? What will be the consequences if a rare and unforeseen event hits this increasingly interdependent and fragile market with full force?

References

Adi, A., Botzer, D., Nechushtai, G. and Sharon, G., 2006. Complex Event Processing for Financial Services. *2006 IEEE Services Computing Workshops*.

Ahmed, E., Yaqoob, I., Hashem, I., Khan, I., Ahmed, A., Imran, M. and Vasilakos, A., 2017. The role of big data analytics in Internet of Things. *Computer Networks*, 129, pp.459-471.

Amihud, Y., and H. Mendleson, 1986, Asset pricing and the bid-ask spread, Journal of Financial Economics 17, 223-249.

Antweiler, W. and Frank, M., 2001. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *SSRN Electronic Journal*.

Baddou,A., 2017. *Les Nouveaux Loups de Wall Street*. 2017. [video].

BAKER, M. and WURGLER, J., 2006. Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, 61(4), pp.1645-1680.

Baker, M. and Wurgler, J., 2007. Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, 21(2), pp.129-151.

Bank, M., Larch, M. and Peter, G., 2011. Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management*, 25(3).

Barber, B., Lehavy, R., McNichols, M. and Trueman, B., 2003. Reassessing the Returns to Analysts' Stock Recommendations. *Financial Analysts Journal*, 59(2), pp.88-96.

Barberis, N., Thaler, R.H., 2003. A survey of behavioural finance. In: Constantinides, G.M., Harris, M., Stulz, R. (Eds.), *Handbook of the Economics of Finance. Elsevier Science BV*, pp. 1052–1121 (Ch 18).

Beckhart, B. and Keynes, J., 1936. The General Theory of Employment, Interest and Money. *Political Science Quarterly*, 51(4), p.600.

Bhagat, S., W. Marr, and M. Spivey, 1985, "The rule 415 experiment: equity markets." *Journal of Finance* 40, 1385-1401.

Blackwell, D., W. Marr, and M. Spivey, 1990," Shelf registration and the reduced due diligence argument: Implications of the underwriter certification and the implicit insurance hypotheses." *Journal of Financial and Quantitative Analysis* 25, 245-259.

Bollen, J. and Mao, H., 2011. Twitter Mood as a Stock Market Predictor. *Computer*, 44(10), pp.91-94.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. and Weber, I., 2012. Web Search Queries Can Predict Stock Market Volumes. *PLoS ONE*, 7(7), p.e40014.

Burlacu, R., Fontaine, P. and Jimenez-Garcès, S., 2005. The "firm-specific return variation": a measure of price informativeness or information asymmetry?. *Annals of Financial Economics*, 01(01), p.0550004.

Cerchiello, P. and Giudici, P., 2016. Big data analysis for financial risk management. *Journal of Big Data*, 3(1).

Chen, H., De, P., Hu, Y. and Hwang, B., 2014. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, 27(5), pp.1367-1403.

Chui, M., Loffler, M., Roberts, R., 2010. The internet of things. McKinsey quarterly.

Chung, K., McInish, T., Wood, R. and Wyhowski, D., 1995. Production of information, information asymmetry, and the bid-ask spread: Empirical evidence from analysts' forecasts. *Journal of Banking & Finance*, 19(6), pp.1025-1046.

Clarke, J., Chen, H., Du, D. and Hu, Y., 2018. Fake News, Investor Attention, and Market Reaction. *SSRN Electronic Journal*.

Corea, F. and Cervellati, E., 2015. The Power of Micro-Blogging: How to Use Twitter for Predicting the Stock Market. *Eurasian Journal of Economics and Finance*, 3(4), pp.1-7.

Corea, F., 2016. Big Data and Risk Management in Financial Markets: A Survey. *Canadian Derivatives institute*.

Cullen, G., Gasbarro, D., Monroe, G.S., 2009. Mutual fund trades and the value of contradictory private information, Journal of Banking & Finance

DA, Z., ENGELBERG, J. and GAO, P., 2011. In Search of Attention. *The Journal of Finance*, 66(5), pp.1461-1499.

DA, Z., ENGELBERG, J. and GAO, P., 2011. In Search of Attention. *The Journal of Finance*, 66(5), pp.1461-1499.

De Choudhury, M., Sundaram, H., John, A. and Seligmann, D., 2008. Can blog communication dynamics be correlated with stock market activity?. *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia - HT '08*.

Dzielinski, M. (2012). Measuring economic uncertainty and its impact on the stock market. *Finance Research Letters*, 9(3), 167–175.

Easley, D., N. Kiefer, M. O'Hara, and J. Paperman, 1996, "Liquidity, information, and infrequently traded stocks," *Journal of Finance* 51, 1405-1436.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1–24.

Elton, E., Gruber, M., Gultekin, M., 1984. Professional expectations: accuracy and diagnosis of errors. Journal of Financial and Quantitative Analysis 19, 351}363.

Elton E. J. and Gruber M. J., Modern Portfolio Theory and Investment Analysis, 5th ed., Wiley, 1995

Fama, E., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), p.383.

Fama E. F. and French K. R., "Multifactor Explanations of Asset Pricing Anomalies", Journal of Finance, vol. 51, n°1, March 1996, pp. 55-81.

Gaver, J. and K. Gaver, 1993, "Additional evidence on the association between the investment opportunity set and corporate financing, dividend, and compensation policies," *Journal of Accounting and Economics* 16, 125-160.

Godfrey, K., 2017. Toward a model-free measure of market efficiency. *Pacific-Basin Finance Journal*, 44, pp.97-112.

Gressis N., Philippatos G. C. and Vlahos G., "Net Selectivity as a Component Measure of Investment Performance", Financial Review, vol. 21, n°1, 1986.

Grossman, S., & Stiglitz, J. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review, 70*(3), 393-408. Retrieved May 5, 2020, from www.jstor.org/stable/1805228

Iafrate, F., 2018. *INTELLIGENCE ARTIFICIELLE ET BIG DATA*. ISTE.

JEGADEESH, N. and TITMAN, S., 1993. Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, 48(1), pp.65-91.

Jensen M. C., "The Performance of Mutual Funds in the Period 1945-1964", Journal of Finance, vol. 23, May 1968, pp. 389-419.

Joseph, K., Babajide Wintoki, M. and Zhang, Z., 2011. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, 27(4), pp.1116-1127.

Kang, J., Liu, M.H., Ni, S.X., 2002. Contrarian and momentum strategies in the China stock market: 1993–2000. Pac. *Basin Financ*. J. 10, 243–265.

Krishnaswami, S. and Subramaniam, V., 1999. Information asymmetry, valuation, and the corporate spin-off decision. *Journal of Financial Economics*, 53(1), pp.73-112.

Kyle, A., 1985, "Continuous auctions and insider trading," *Econometrica* 53, 1315-1336.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J., 2000. Language models for financial news recommendation. *Proceedings of the ninth international conference on Information and knowledge management - CIKM '00*,.

Lazard, E., Mounier-Kuhn, P. and Berry, G., 2016. *Histoire Illustrée De L'informatique*. [Les Ullis]: EDP Sciences.

Le Sourd, V., 2007. Performance measurement for traditional investment. Edhec Risk and asset management research centre.

Lewis, M. and Baker, D., 2014. *Flash Boys*. New York: Simon & Schuster Audio.

LOUGHRAN, T. and MCDONALD, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35-65.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P. and Smith, R., 2018. News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment. *SSRN Electronic Journal*.

McLaughlin, R., A. Safieddine, and G. Vasudevan, 1998, "The information content of corporate offerings of seasoned securities: An empirical analysis," *Financial Management* 27, 31-45.

Morck, Randall, Bernard Yeung, and Wayne Yu, 2000, The information content of stock markets: Why do emerging markets have synchronous stock price movements?, Journal of Financial Economics 58, 215–238.

O'Leary, D., 2013. 'BIG DATA', THE 'INTERNET OF THINGS' AND THE 'INTERNET OF SIGNS'. *Intelligent Systems in Accounting, Finance and Management*, 20(1), pp.53-65.

Ohlhorst, F., 2013. *Big Data Analytics*. Hoboken, N.J.: Wiley.

Penman, S., 1996, "The articulation of price-earnings ratios and market-to-book ratios and the evaluation of growth," *Journal of Accounting Research* 34, 235-258.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. and Mozetič, I., 2015. The Effects of Twitter Sentiment on Stock Price Returns. *PLOS ONE*, 10(9), p.e0138441.

Samuelson, P., 1965. Proof that properly anticipated prices fluctuate randomly. *Ind. Manag. Rev*. 6 (2).

Sanger, W., Warin, T., 2016. Frequency and unstructured data in Finance: an exploratory study of Twitter". *Journal of global research in computer science.*

Sarlin, P., 2016a. Macroprudential oversight, risk communication and visualization. *Journal of Financial Stability*, 27, pp.160-179.

Sarlin, P., 2016b. Computational Tools for Systemic Risk Identification and Assessment. *Intelligent Systems in Accounting, Finance and Management*, 23(1-2), pp.3-5.

Scholtz H. and Wilkens M., "A Jigsaw Puzzle of Basic Risk-adjusted Performance Measures", Journal of Performance Measurement, spring 2005.

Schumaker, R. and Chen, H., 2009. Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, 27(2), pp.1-19.

Sharpe W. F., "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk", Journal of Finance, vol. 19, September 1964, pp.425-442.

Shen, D. and Chen, S., 2018. Big Data Finance and Financial Markets. *Computational Social Sciences*, pp.235-248.

Shiller, R.1981. Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends? *The American Economic Review*, Vol. 71, No. 3, PP. 421-436.

Shiller, R. 1990. Market Volatility and Investor Behavior. *The American Economic Review*, 80(2), 58-62.

Shiller, R., 2001. *Irrational Exuberance*. Princeton, N.J.: Princeton University Press.

Skuza, M. and Romanowski, A., 2015. Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction. *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*.

Subrahmanyam, A., 2019. Big data in finance: Evidence and challenges. *Borsa Istanbul Review*, 19(4), pp.283-287.

Sun, Y., Shi, Y. and Zhang, Z., 2019. Finance Big Data: Management, Analysis, and Applications. *International Journal of Electronic Commerce*, 23(1), pp.9-11.

Taleb, N.N., 2007. The black swan: the impact of the highly improbable. *Choice Reviews Online* 45(03), pp.45-1430-45-1430.

Tian, X., Han, R., Wang, L., Lu, G. and Zhan, J., 2015. Latency critical big data computing in finance. *The Journal of Finance and Data Science*, 1(1), pp.33-41.

Westin, A., 1970. *Privacy And Freedom*. London: Bodley Head.

Wysocki, P. (1998). Cheap talk on the web: The determinats of postings on stock message boards. *University of Michigan Business School Working Pape*r (98025).

Wysocki, P., 1999. Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. *SSRN Electronic Journal*.

Yu, S. and Guo, S., 2016. *Big Data Concepts, Theories, And Applications*. SPRINGER.

Zhang, W., Shen, D., Zhang, Y. and Xiong, X., 2013. Open source information, investor attention, and asset pricing. *Economic Modelling*, 33, pp.6 3-619.

Zikopoulos, P. and Melnyk, R., 2013. *Harness The Power Of Big Data*. New York: McGraw-Hill.

Websites:

BlackRock. 2020. [online] Available at: <https://www.blackrock.com/hk/en/investment-ideas/systematic-active-equity>

Blackrock.com. 2020. [online] Available at: <https://www.blackrock.com/corporate/literature/whitepaper/viewpoint-artificial-intelligence-machine-learning-asset-management-october-2019.pdf>

Datafloq.com. 2020. A Short History Of Big Data. [online] Available at: <https://datafloq.com/read/big-data-history/239>

Docs.microsoft.com. 2020. Apprentissage Profond Et Apprentissage Automatique - Azure. [online] Available at: <https://docs.microsoft.com/fr-fr/azure/machine-learning/concept-deep-learning-vs-machine-learning>

Ecb.europa.eu. 2009. [online] Available at: <https://www.ecb.europa.eu/pub/pdf/fsr/art/ecb.fsrart200912_02.en.pdf>

Ft.com. 2020. Blackrock Bets On Algorithms To Beat The Fund Managers. [online] Available at: <https://www.ft.com/content/e689a67e-2911-11e8-b27e-cc62a39d57a0>

Hadoop.apache.org. 2020. Apache Hadoop. [online] Available at: <http://hadoop.apache.org>

marketsandmarkets.com. 2020. IoT in Banking and Financial Services Market Solution (Security, a., 2020. Sample Request - Iot In Banking And Financial Services Market Size, Share And Global Market Forecast To 2023 | Marketsandmarkets. [online] Marketsandmarkets.com. Available at: <https://www.marketsandmarkets.com/requestsampleNew.asp?id=172304505>

Marketsandmarkets.com. n.d. Humanoid Robot Market Worth 3,962.5 Million USD By 2023 Growing With A CAGR Of 52.1%. [online] Available at: <https://www.marketsandmarkets.com/PressReleases/humanoid-robot.asp>

Medium. 2020. Everything A Data Scientist Should Know About Data Management*. [online] Available at: <https://towardsdatascience.com/everything-a-data-scientist-should-know-about-data-management-6877788c6a42>

Statista Infographies. 2020. Infographies. [online] Available at: <https://fr.statista.com/graphique-du-jour/>

SUPINFO - Ecole Informatique - Formation en Informatique - Paris, .., 2020. L'évolution D'internet | SUPINFO, École Supérieure D'informatique. [online] Supinfo.com. Available at: <https://www.supinfo.com/articles/single/5256-evolution-internet>

2025, B. and L, +., 2020. Big Data : Le Volume De Données Mondial Multiplié Par 5 D'Ici 2025 - Lebigdata.Fr. [online] LeBigData.fr. Available at: <https://www.lebigdata.fr/big-data-2025-idc>

Appendices

Measures of performance

Notations:

rf = risk-free interest rate
μi = average return of fund i
μm = average return of the market index
σ i = standard deviation of the returns of fund i i
σ m = standard deviation of the returns of the market index
ßi = market risk of fund i
μbpi = average return of the benchmark portfolio

Absolute risk-adjusted performance
Total risk

Sharpe Ratio 1966

It is a measure that evaluates funds' risk-adjusted returns with no reference to the benchmark. It allows gauging the return of investment compared to its risk. The Sharpe ratio measures the return of a portfolio in excess of the risk-free rate, compared to the total risk of the portfolio, measured by its standard deviation. (V. Le Sourd, 2007).

Sharpe Ratio (SRi) = $\frac{(\mu i - rf)}{\sigma i}$

The Sharpe ratio gives a clearer idea of the profit associated with the risks taken. All other things being equal, higher Sharpe Ratios translate into higher performance.

Relative-risk adjusted performance
Systematic risk

Alpha Jensen

Jensen's alpha was proposed in 1968 by M.C. Jensen and this is a relative risk-adjusted performance measure. This method is based on the CAPM, which is "the difference between the portfolio return above the risk-free rate and the return explained by the market model". (V. Le Sourd, 2007).
Jensen's alpha is a risk-adjusted performance measure that represents the average return on a portfolio or investment. It shows whether the portfolio's return is above or below that predicted by the CAPM.

Alpha = μi – [rf + ß (μm – rf)]

Note that contrary to the Sharpe and Treynor measures this is a measure relative to the benchmark however only the systematic risk is taking into account because the specific risk is eliminated by diversification. This is one limit of this model, as we have seen diversification

does not necessarily eliminate specific risk. Moreover, the Jensen measure is based on the CAPM which is also limited because based on strong assumptions.
The Jensen measure has the advantage to be easy but we face some limitations. Therefore the Elton and Gruber (1995) measure can fill the gap. They proposed a performance measure in the continuity of Jensen's alpha but they take into account the total risk as well as a portfolio that is not on the Security market line (taking into account the possibility of lending and borrowing).

Total risk

Total risk alpha (TRA) Fama 1972

Total risk alpha measures the performance of a fund by comparing its returns to those of the portfolio benchmark by taking into account the total risk of the funds. This total risk is obtained by combining the market portfolio and the risk-free asset, the portfolio is therefore situated on the Capital Market Line.

$TRAi = \mu i - \mu bpi$

$\mu bpi = rf + \frac{\mu m - rf}{\sigma m} \sigma i$

A portfolio can have a higher return for a fixed level of risk if the manager chooses to borrow money at the risk- free rate and reinvest it. As explained by H. Scholz and M. Wilkens (2005): « fund B has a higher total risk alpha than fund A. However, the TRA is not a suitable device for ranking performances. An investor in fund A could have created a portfolio which exactly matches the return and risk of fund B simply by borrowing at the risk-free rate and investing these proceeds in fund A. This effect is called the leverage bias ». Therefore the Jensen alpha a by a lever.

We can therefore write :

$\mu i - \mu bpi = \mu i - rf - \frac{\mu m - rf}{\sigma m} \sigma i$

$TRAi = \mu i - rf - \frac{\mu m - rf}{\sigma m} \sigma i$

Total risk alpha (TRA) Gressis, Philippatos and Vlahos (1986)

Gressis, Philippatos and Vlahos (1986) from the total risk alpha formula propose an alternative formula to derive the total risk alpha in order to understand the link between TRA and Sharpe ratio.

$TRAi = \mu i - rf - \frac{\mu m - rf}{\sigma m} \sigma i$

Knowing $SRi = \frac{(\mu i - rf)}{\sigma i}$

$TRAi = \sigma i \, (SRi - SRm)$

These performance analysis methods will help us in our experimentation. It should be noted that total risk alpha will be favoured because the informational risk we are studying is part of the specific risk. Moreover, the total risk alpha remains simple to use. Moreover, it is based on the CAPM hypothesis but takes into account the reality of the portfolio as it is not necessarily on the Capital Market Line.

Fama Mcbeth regression

From Fama-Macbeth Two-Step Regression by IHS EViews 2014.

The two-step Fama-MacBeth regression is a way to test how factors affect asset returns. In theory, returns are explained by the risk associated with certain factors. Therefore, the objective of the Fama MacBeth regression is to highlight the premium i.e. the coefficient of exposure to these factors. It can help assess the risk of the factor relative to the level of the premium, i.e., the coefficient. The objective is to find the premium associated with exposure to the selected risk factors.
Two steps in the Fama MacBeth regression. The first step regresses the performance of selected portfolios against one (or more) time series of selected factors.
The second step the cross-section of portfolio returns is regressed against the factor exposures. This is done for each time step in order to give a time series of the coefficient for each factor.
Then we must average the coefficients obtained for each time series, for each factor. This average of each of the coefficients for each of the factors will be respectively the expected premium corresponding to the risks of each of the factors.
In other words, the average of the coefficient will be the expected premium for exposure to a risk unit.

Thus, for n portfolio or asset returns and m factors, the exposure of the ß factors is obtained by calculating n regressions, each on these m factors

First step: time-series regression for each stock i

$$R_{1,t} = \alpha_1 + \beta_{1,F1}F_{1,t} + \beta_{1,F2}F_{2,t} + ... + \beta_{1,Fm}F_{m,t} + \varepsilon_{1,t}$$
$$R_{2,t} = \alpha_2 + \beta_{2,F1}F_{1,t} + \beta_{2,F2}F_{2,t} + ... + \beta_{2,Fm}F_{m,t} + \varepsilon_{2,t}$$
$$\vdots$$
$$R_{n,t} = \alpha_n + \beta_{n,F1}F_{1,t} + \beta_{n,F2}F_{2,t} + ... + \beta_{n,Fm}F_{m,t} + \varepsilon_{n,t}$$

Ri,t = return of portfolio or asset i (n total) at time t,
Fj,t = factor j (m total) at time t,
ßi,Fm = factor exposures, (that describe how returns are exposed to the factors,
t = time goes from 1 through T.

The next step consists in calculating T cross-sectional regressions of the returns of the equations of step 1.

Second step : cross-sectional regression

$$R_{i,1} = \gamma_{1,0} + \gamma_{1,1}\hat{ß}_{i,F1} + \gamma_{1,2}\hat{ß}_{i,F2} + ... + \gamma_{1,m}\hat{ß}_{i,Fm} + \varepsilon_{i,1}$$
$$R_{i,2} = \gamma_{2,0} + \gamma_{2,1}\hat{ß}_{i,F1} + \gamma_{2,2}\hat{ß}_{i,F2} + ... + \gamma_{2,m}\hat{ß}_{i,Fm} + \varepsilon_{i,2}$$

.
.
.

$$R_{i,T} = \gamma_{T,0} + \gamma_{n,1}\hat{ß}_{i,F1} + \gamma_{n,2}\hat{ß}_{i,F2} + ... + \gamma_{n,m}\hat{ß}_{i,Fm} + \varepsilon_{i,T}$$

ß^= estimated factor exposures
γ = coefficient of regression

γ are the regression coefficients that will be used to calculate the average in order to achieve the expected premium for exposure to the factor.

To calculate this premium, the average of all the coefficients obtained for each factor must be calculated. We will have 1 premium per factor but as much premium as a factor.
It is possible to skip the second step and to go directly to the third. The second step with T regressions can be replaced by a single regression of n portfolio returns, averaged over time, against m factor exposures with lengths n.

Third step: average return over time

$$E(R_i) = \gamma_0 + \gamma_1\hat{ß}_{i,F1} + \gamma_2\hat{ß}_{i,F2} + ... + \gamma_m\hat{ß}_{i,Fm} + \varepsilon_i$$

E(Ri) = average return over time

(Go back to **Measure of performance**)

Table of content