Mathieu Rosenbaum

Big data: A Game Changer for Quantitative Finance?

Prisme N°35
March 2017

The Cournot Centre and Foundation

The Cournot Centre and Foundation

The Cournot Centre is an independent, non-profit organization. It has created a forum where specialists meet to advance theories and increase understanding of the economic and social sciences and their epistemology. Like the Centre, the Foundation pursues catalysing work in the tradition of Augustin Cournot, accelerating theoretical formulations. Under the aegis of the Fondation de France, the Foundation puts into perspective the probabilistic paradigm shift originating in mathematics and spreading across disciplines.

Big Data: A Game Changer for Quantitative Finance?[1]

Mathieu Rosenbaum

Prisme N°35

March 2017

[1] Transcript by Jean-Gabriel Brin of Mathieu Rosenbaum's talk at the Cournot Centre workshop on big data. English translation by Lynne Forest. The video is available on the Cournot Centre's website: www.centre-cournot.org. The Centre thanks Serge Chanchole for his careful proofreading.

© Cournot Centre, March 2017

Summary

The increasingly widespread availability of massive volumes of digital data is challenging the mathematical sciences. While the inflow of big data modifies our ways of accessing and managing data, quantitative methods remain, on the other hand, largely unchanged, even if they must hitherto be applied on a very large scale over very short time periods.

The aim of this *Prisme* is to evaluate the reformulations currently imposed by big data in the various fields of mathematics and the developments they are making possible in finance.

Any attempt to describe the full range of financial activities in just a few pages would be doomed to failure. It is, however, possible to take an objective look at some of their developments. The aim of this *Prisme* is to analyse the development considered to be the most significant in recent years: that of big data. They have made their appearance in numerous fields, and massive investments have been made in the techniques for collecting and processing them, from the necessary equipment to advertising. Many companies have launched out onto markets that rely on the idea of an ongoing revolution. At first glance, the term "big data" would appear to designate a commercial offer. Although the general mathematical framework used to study them has not changed, the approaches and methods applied in finance have been affected by the arrival of big data. How can we describe this transformation of procedures, and how can they be placed within a frame of reference? This text aims to trace the recent history of methods applied to financial markets in order to define and characterize the changes that have taken place in financial practices.

A brief look at the history of the development of derivative products will allow us to trace the evolution of the treatment of uncertainty on financial markets, today partially attributed to big data. The appearance of derivative products in their current form dates back to the 1970s. Indeed, the year 1971 marked the end of US dollar convertibility to gold. The first options contract market was created two years later. The *Chicago Board Options Exchange* was set up to deal with the uncertainties surrounding the end of the fixed exchange rates system established in Bretton Woods in 1944. It was also in 1973 that Fischer Black, Robert Merton and Myron Scholes finalized their options pricing formula. The approach of these researchers, based on continuous-time stochastic processes, made the pricing of financial derivatives possible. With such products, companies could deal with the fresh uncertainties they were facing. In France, the idea of a financial derivatives market was put forward in 1978, but the first French futures market, MATIF,[2] did not open until 1986. Black, Scholes and Merton's idea spread, and their probabilistic methods were quickly adopted. Their findings made it possible to create hedge and investment products,

[2] *Marché à terme international de France*

whose risks they were able to control perfectly thanks to dynamic management.[3]

Let us take a concrete example to better understand the logic behind this idea. We will take the case of a buyer who wishes to supply himself with a ton of kerosene in a year's time, at today's price, of, let's say, 1 euro a kilo. The client and the financial institution sign a contract, known as a call option, which gives the client the right, if he wishes, to demand the ton of kerosene at the price stipulated in the contract (1 euro a kilo) after one year. If the ton of kerosene costs less than 1 euro after the year has passed, the client does not use his call option; if it is more expensive, he exercises it. Let us imagine that the price of kerosene after one year is equal to s; if $s > 1$, the option is exercised, otherwise not. At the end of the year, the client receives a profit of $1000*(s - 1)$ if $s - 1 > 0$, otherwise nothing. Obviously, the financial institution receives payment on the day of the signing of the contract: this is the price of the call option. How can the price of the option, c, be fixed?

Black–Scholes[4] and Merton's[5] theories allow us to calculate c in such a way that by investing c on the markets in a dynamic manner, the financial institution is certain to be able to remunerate the client at the end of the year, in other words, to transform the value of c today into $1000*(s - 1)$, if $(s - 1)$ has a positive value tomorrow, otherwise 0. Therefore, there is no longer risk involved if the hypotheses of this probabilistic model are respected. This way of eliminating the risk factor has proved to be a scientific revolution on the markets.

The time interval remains fundamental in all cases. Indeed, in the world of derivative products, as well as for structural reasons, the natural time unit is a day. Until the end of the 1990s, a daily value was registered, every day at 5 pm, for example. The trader had to use a pricing model to fix a price observed once during the day, at a fixed time, then at the same time the next day, and so on. The collected data were reviewed and analysed at the close of the day, and the process was

---

[3] El Karoui, Nicole (2009), *A Moment of the Probabilistic Experience: The Theory of Stochastic Processes and their Role in the Financial Markets, Prisme* No. 17, Cournot Centre. This article describes and analyses the history of the development of derivative products.

[4] Black, Fischer, and Myron Scholes (1973), "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*. 81(3): 637–54.

[5] Merton, Robert (1973), "Theory of Rational Option Pricing", *Bell Journal of Economics and Management Science,* The RAND Corporation, 4 (1): 141–83.

renewed every day. When the price was recorded every day at 5 pm for two years, this is what we obtained:

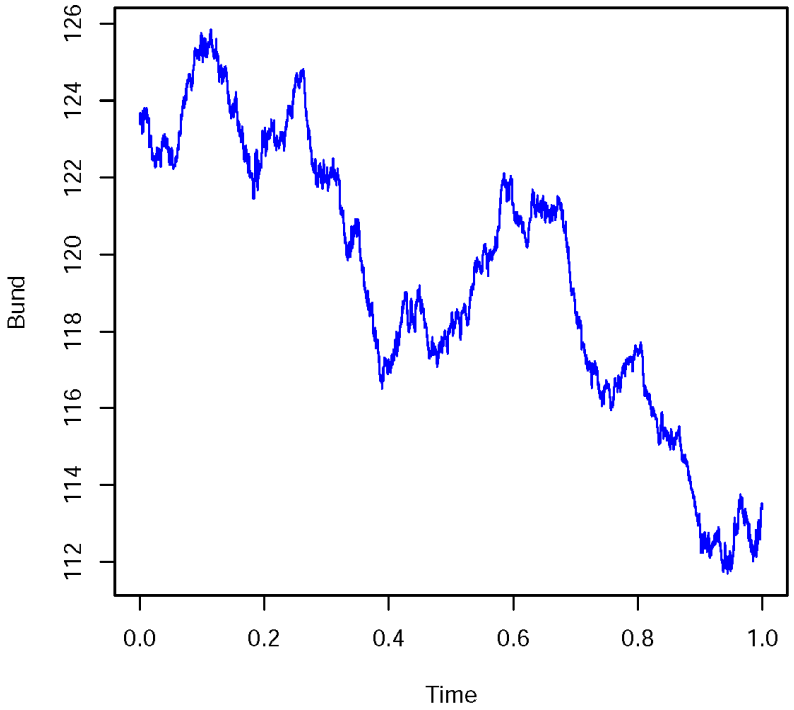


Figure 1: Example of a standard graph for the price of a share over a two-year period after daily recording.

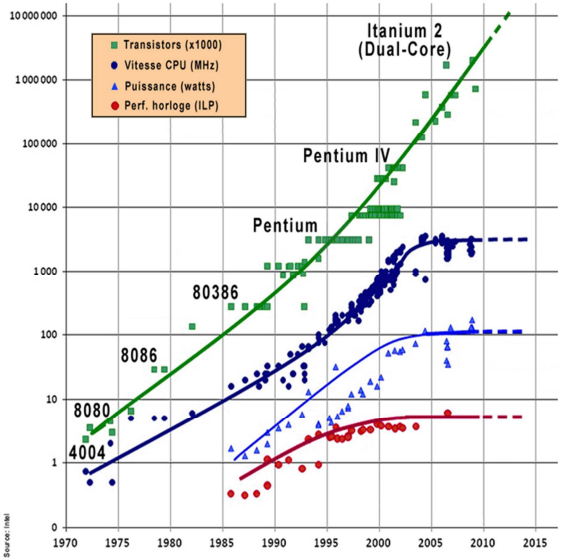
This graph resembles that of another particular random movement: Brownian motion. Before the arrival of big data, the world of financial mathematics conserved this basic modelling. On a statistical level, Brownian motion is an appropriate process, just as it is suitable from the point of view of visual perception. Furthermore, some of its theoretical properties make it particularly adapted to this time scale.

The modernization of financial mathematics that followed Black–Scholes and Merton's work coincided with the arrival of increasingly powerful computers and software. The gradual launch of faster machines and the arrival of personal

computers in the mid-1980s led to the disappearance of the IBM mainframe computers used to centralize calculations. This replacement was accompanied by a proliferation of the data available to traders. New software was designed and new professions created: computer specialists and engineers flocked to the trading rooms in the 1990s. The arrival of big data became palpable at the beginning of the 2000s. Gradually, the professional lexicon was affected too. "Statistics", for example, disappeared and gave way to "Data Science". In the French *Grandes Ecoles*, there were no longer any specializations in statistics or statistics course modules, which were replaced by the various branches of data science. "Classification" disappeared too, as "clustering" became the new buzzword. The outdated exploration of data – "data mining" – gave way to automatic learning, better known as "machine learning". There is clearly a fashion component in the advent of big data, and we must question what this trend conceals. Where do the tangible transformations lie?

To attempt to answer that question, we must first define "big data" in the context of financial mathematics and develop the right hypotheses to help us to understand the consequences of their use. Most importantly, mass data are derived from the systematic recording of market activity. These data are conserved and can be easily consulted. In practical terms, all the prices, transactions and order books are systematically memorized. The recording processes allow for the conservation and time-stamping of all versions, with the details of each access. This makes the data easier to use by traders and provides complete traceability. The reason for these recordings is that information potentially lies in the recorded data, with ***potentially*** being the operative word here. The data are only of interest if they can be easily accessed and analysed in a pertinent manner by the systems. The use of big data relies heavily on the appropriate storing of data. Consequently, the accumulation of data has proven to be of great use in high-frequency trading.

High-frequency transactions involve quasi-permanent operations on the market, where a very high number of exchanges takes place in a very short time. They are a logical development of the uninterrupted progress in processing capacity. This progression is illustrated by the two graphs below.



Source: Intel

Figure 2: Growth in data processing capacity 1970–2015

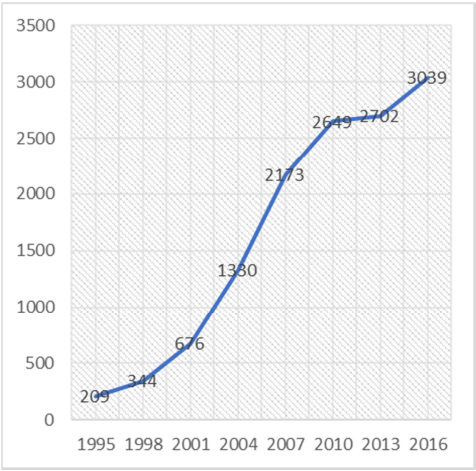


Figure 3: Turnover of the over-the-counter market for derivative products according to the *BIS Statistics Explorer* (April 2016, daily averages in billions of U.S. dollars)

In most cases, the quasi-permanent interventions on the market are not the result, as we can sometimes hear, of ill-intentioned hackers attempting to steal investors' money; today, they are present in all practices of the financial markets. Moreover, thanks to powerful computers, anyone can have access to high-, or more-or-less high-frequency computing power, and the financial markets are adapting. The scale is in milliseconds or even nanoseconds. At such speed, all the transactions in an order book can be recorded, and it is these data that traders focus on. The market is thus, at time *t*, a virtual place where buyers and sellers exchange a financial asset. Let us look now at how this works in concrete terms.

Let us take the shares of Company X as an example. At time *t*, there are buyers ready to purchase Company X's shares at 110.53€, some wanting to buy at 110.52€, others at 110.51€, and so on. The table below shows the prices, with the quantities in the column opposite.

Table: Simplified order book

Bid price	Bid size (number of shares)	Ask price	Ask size (number of shares)
110.53€	95	110.54€	17
110.52€	90	110.55€	50
110.51€	5	110.56€	117
110.50€	2	110.57€	18
110.49€	10	110.58€	100

According to this order book, there is at least one trader who is ready to buy "95 Company X shares at a unit price of 110.53€". There may be two traders: one who is

willing to buy 40 shares, the other 55. Most of the time the breakdown is not explicit, but, in any case, we have 95 shares that will potentially be purchased at 110.53€. On the sell side, it is the same principle: there are sellers ready to sell at 110.54€, at 110.55€, at 110.60€, and so on.

A transaction occurs when someone says, for example, "I am prepared to buy at 110.53€". If you are prepared to sell at 110.53€, you have found a buyer. Someone else may say, "I am prepared to buy at 110.54€," and a seller is found. These are the operations that can be observed in trading rooms, where traders scrutinize the order books on their screens. Basically, when a datum changes, an indicator lights up and flashes on the trading screen. A bank thus records all the orders for the thousands of assets available. Terabytes of data are consequently conserved. The data from the four columns are recorded for each asset. What needs to be recorded is of the order of $5+5+5+5=20$, and they change roughly every millisecond. The data are thus truly massive and arrive at very high frequency.

In this environment, it has become reasonable, for example, to ask a bank for the content of the order book for a given share three days previously, at 16 hundred hours, 43 minutes, 55 seconds, and 96 milliseconds, since the bank is technically capable of giving an answer. This major innovation in storage techniques has been accompanied by a more precise grasp of the market, as its substance has become more easily accessible. How can models be improved and tested so that they can be adapted to these new hypotheses? If more data are available to stimulate researchers, there are also a lot more data to calibrate their new models and put them to the test, under far more difficult conditions.

The exploitation of data recorded every millisecond, or even every nanosecond, is not possible within the framework of the Brownian motion described above. Indeed, it is not pertinent for the short timescales of many trading problems. A client may call a broker and give the instructions to place an order such as: "you have two hours to sell 10 million shares for me". The sale of 10 million shares can neither be performed haphazardly, nor in one go. How then can the client's profit be maximized? If the trader strictly follows the order book in the table, he will sell 95 lots at 110.53€, 90 lots at 110.52€, and so on, so that the average sale price per share is now worth, let us say 100€. The client, who is scrutinizing the market, observes that the share price is roughly 110.53€, and therefore does not accept that

the average price works out to 100€. The trader must thus organize his or her sales strategy over the two-hour interval.

The new modelling problems that are emerging are thus a direct consequence of the comprehension and use of a very large volume of data over a very short time period. Even if only considering the first line of an order book (110.53€ in the table, for example), during the two hours allowed for the transaction, the price will vary according to a trajectory that bears no resemblance to a Brownian motion trajectory as shown in Figure 4 below.

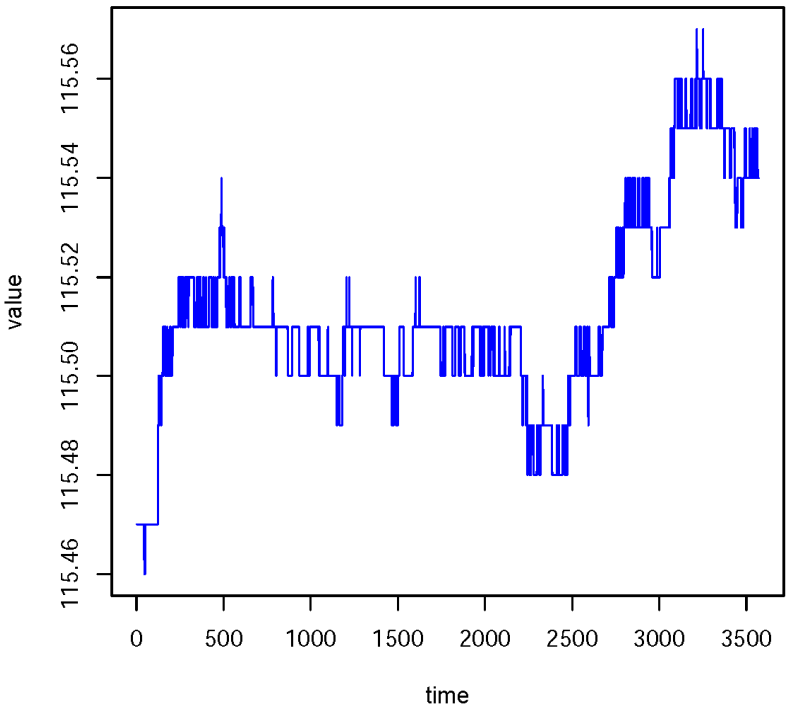


Figure 4: Standard evolution of an asset over a one-hour period

The challenges posed by the explosion of big data in finance are first and foremost those of the field of mathematics, which must provide answers to the questions asked by practitioners. Among such questions, that of the volatility of the market is primordial. We are able to interpret market volatility as represented in Figure 1; it can be measured according to the framework of Brownian motion. We

are, however, less well-equipped to interpret the graph in Figure 4, where the notion of volatility is essential in the calculation of risk in high-frequency trading. Another difficulty concerns the correlation between two different assets. Ten years ago, it was possible to establish empirical correlations of price increases on a daily basis in order to obtain an approximate estimation of the price differences between assets. This is a far more delicate operation over a two-hour timespan. Even if the price variations are comparable, the fact that the transaction times differ between two companies leads to problems of time discontinuity, which are extremely difficult to solve for statisticians. Thus, we may observe two charts showing sharp fluctuations, but at different times (in itself quite an insignificant observation). Non-synchronicity alone is an obstacle that makes it extremely difficult to calculate such correlations. The need for such calculations remains, nevertheless, very strong in high-frequency trading.

Big data in finance have brought up fresh research questions and led to new mathematical developments. Progress has notably been made in stochastic control. Observations of bank practices invite us to revisit, through the prism of big data, some of the techniques and problems identified since the development of derivative products. This phenomenon is reflected in the creation of new academic programs in universities. When I was a professor at University Paris VI, we were asked to set up a big data degree programme. Once we had reviewed all the mathematics courses already on offer, we realized that approximately 30 of them fit under the big-data label. What was lacking was a common umbrella under which to place them, and especially a better understanding of the interdependency between statistics and other digital disciplines, particularly computer science. This experience brought us to the realization that the computer-science component must be designed and taught in synergy with the other components.

The confusion arising from the misconception among the users of mega data should also be dispelled. One of the dangers facing finance is a belief in the all-powerfulness of big data and the idea that they represent a solution to every problem. One of the most marked effects of this belief is the recognition of new qualifications, particularly concerning the capacity to treat these digital data. A sharp interest for applicants able to demonstrate these capacities can be observed in job descriptions. Competency in statistics and informatics are obviously necessary for jobs in finance. Nevertheless, there is a tendency today to believe that financial models are useless, and that any

gifted computer scientist who knows about machine learning is competent in finance. That is false. We must not forget that finance is an institution with agents interacting with markets, which are also institutions, whose presence is not always easily discernible! If we forget this, we end up paying for it sooner or later. Moreover, big data alone are of no use in the financial industry; they are a means and not an end. It is interesting, for example, to look at the fate of structures that have not acknowledged this: today the average lifespan of new high-frequency trading companies relying purely on IT techniques is six months.

Ultimately, the challenge presented to us by this abundance of data lies in elaborating suitable high-frequency models. How can we achieve this? A good high-frequency model must accurately reproduce the empirical high-frequency characteristics of the markets, but must also be a useful model for resolving difficult trading problems (such as the one posed by our client wanting to sell 10 million shares in two hours) or for improving our understanding of the markets. Combining these two requirements is no easy task. We have seen of what little use are those discarnate models that merely reproduce abstract phenomena.

I have not, therefore, identified any form of scientific discontinuity with the arrival of big data, but rather a discipline that is moving forward: problems are better formulated, research questions more clearly defined, interaction models more accurate and digital optimization more efficient. Overall, many aspects are progressing rapidly. We cannot speak in terms of a scientific discovery, but rather of research that is advancing and delving deeper. Additionally, the industries and services affected by developments in financial mathematics have become much more aware of the changes needed. Employees must have the necessary qualifications to be able to deal with the large quantities of data that need to be more efficiently recorded, sorted and interpreted. The financial industry very quickly requested the creation of general academic courses dealing with all aspects of big data, and principally the statistical, analytical and computing elements. Adapting course content to techniques that have already demonstrated their effectiveness presents a structural challenge due to the synergies created between the various disciplines. Despite all this, no paradigm shift has occurred in the process.

We should remember that big data did not emerge from a powerful new theorem, knowledge of which has suddenly become a *sine qua non* for working in the

financial industry. It is for the above reasons that, in my opinion, there is a real difference between the birth of derivative products and the recent availability of big data. Fundamentally, big data have allowed us to understand that the notion of time is not a determining factor in itself; it is, however, the essential element in data that enables us to solve practical problems, and this problem-solving dimension depends on the timescale. The model put forward by Black–Scholes and Merton is a continuous-time model for low-frequency, low-volume trading, whereas today trading is definitely high frequency. The modelling of the founding fathers is no longer adapted, and it has become necessary to account for higher dimensions.

To conclude, we should bear in mind that the increasingly widespread availability of huge volumes of digital data initially brought fresh challenges to finance: data access and management, the relevance of new mathematical models and their connection to theories, the reorganization of academic disciplines, and the transformation of operational methods. In the banking industry, the arrival of big data has made us aware of the need for more efficient data access in order to organize and process them better and more accurately. The availability of big data is also an invitation to take a closer look at the history of mathematical developments. We must follow the development of probabilistic methods to understand the transformations that big data have imposed. Their arrival has led researchers, traders and the banks themselves to reformulate questions. They have been forced to rethink a number of seemingly well-established practices dating back to the 1990s. Nevertheless, the emergence of big data is by no means comparable to the revolution brought about by the application of Black–Scholes and Merton's ideas for derivative products.

The *Prisme* Series

The *Prisme* Series is a collection of original texts that focus on contemporary theoretical issues. The authors are contributors to the Cournot series of conferences, panels and seminars.

Latest releases:

34. Saturation and Growth: When Demand for Minerals Peaks
Raimund Bleischwitz and Victor Nechifor

33. A Time-Frequency Analysis of Oil Price Data
Josselin Garnier and Knut Sølna

32. Why Speech Technology (almost) Works
Mark Liberman

31. How to Flee Along a Straight Line: Tracking Self-Repelling Random Walks
Laure Dumaz

30. The Evolving Connection between Probability and Statistics
Noureddine El Karoui

A complete list of publications can be found at
www.centre-cournot.org

Mathieu Rosenbaum is a Professor at Ecole Polytechnique, where he is in charge of the Master 2 programme in probability and finance. The programme is co-organized with University Pierre and Marie Curie (Paris VI) where he was previously a professor.

Co-founder and editor-in-chief of the journal *Market Microstructure and Liquidity*, he is also managing editor of *Quantitative Finance* and associate editor of *Electronic Journal of Statistics*, *Journal of Applied Probability*, *Mathematical Finance, Mathematics and Financial Economics, Statistical Inference for Stochastic Processes, SIAM Journal on Financial Mathematics* and *Statistics & Risk Modeling*. He received the Europlace Award for best young researcher in finance in 2014 and a European Research Council grant in 2015.

Rosenbaum's research mainly focuses on statistical finance problems, such as market microstructure modelling or designing statistical procedures for high-frequency data. He is also interested in regulatory issues, particularly in the context of high-frequency trading.

Series Editor: Jean-Philippe Touffut
Illustration Artist: Gerald Wassen / Cover Designer: Sophie Otrage