1

From Big Data to Deep Learning

The first chapter of the book begins with a presentation of big data, which is both a growing field of study and application (with algorithms and tools) and its raw material (data sets). The term data is in the plural (of datum) as it should be in English, and we can talk about the plurality of data sources, but we can also think of the discipline in the singular, with a set of concepts and methods, a mixture of machine learning[1] and computer technology. In the first sense of the term, big data can be understood as large collections of data, too large for traditional data-processing systems, but this translation only emphasizes the voluminous aspect of these data, whereas their variety is often also essential, as the following pages will show. Of the machine learning methods, we will focus on deep learning methods, which are the best for dealing with big data such as text, voice, image, and video. We will then go on to examine some key points in the processing of data, especially big data, and we will conclude with questions of personal data protection and some elements on open data.

1.1 Introduction

Big Data covers all the issues associated with the collection and exploitation of very large sets of data, of very varied natures and formats (texts, photographs, clicks, signals from sensors, connected objects, etc.), and in very rapid, even continuous evolution. They are invading many fields of activity: health, industry, transport, finance, banking and insurance, retail, public policy, security, scientific research, etc.

The economic stakes are high: McKinsey[2] estimates that big data could save healthcare policies in the United States $300 billion per year and generate $600 billion in consumption by using consumer location data.

We will see that the impacts of big data are very important for both people and business, and that the technological challenges are formidable.

Before going into detail about big data, let's take a quick historical look at the developments that led to them. For a several decades, the rise of computing power has accompanied the explosion of data production. This rise has developed over several eras:

- before 1950, the beginnings of statistics with a few hundred individuals and a few variables, collected in a laboratory according to a strict protocol of experimental design for a scientific study;

Deep Learning: From Big Data to Artificial Intelligence with R, First Edition. Stéphane Tufféry.

Companion website: www.wiley.com/go/Tuffery/DeepLearning

- in the 1960s–1980s, data analysis with a few tens of thousands of individuals and a few dozen variables, rigorously collected for a specific survey;
- in the 1990s and 2000s, data mining with several million individuals and several hundred variables, collected in the information system of companies for decision-making;
- from the 2010s, big data with several hundred million individuals and several thousand variables, of all types, collected in companies, systems, the Internet, for decision-making and new services.

From the end of the nineteenth century until the 1950s was the era of classical statistics. It precedes the invention of computers, so the means of calculation are manual and very limited, and it is characterized by:

- small volumes studied;
- the strong hypotheses on the statistical distributions followed (triad: linearity, normality, homoscedasticity);
- models derived from theory and confronted with data;
- the probabilistic and statistical nature of the methods;
- laboratory use.

The predominance of the hypothetico-deductive method and the importance of tests and inferential statistics can be noted. Data are collected and analyzed within a strict, often scientific, framework in order to verify a theory, which can be refuted by the result of a test. The foundations of mathematical statistics, but also important predictive methods, such as linear discriminant analysis and logistic regression, date from this period.

In the 1960s and 1980s, the emergence of computers revolutionized the discipline, allowing for much more complex and rapid calculations than before, on thousands of analyzed individuals and dozens of variables, with the construction of "Individuals x Variables" tables. This was a time of great theoretical creativity, during which many fundamental methods were invented that are still widely used today. This is the golden age of data analysis and especially of factorial analysis, and visual representation begins to take on great importance.

The 1990s saw the advent of the concept of *data mining,* which is not only characterized by an explosion of computing resources and the quantity of data to be processed, but also by a profound change in the role of quantitative analysis. Even if it uses the tools of statistics, data mining differs from statistics in many qualitative and quantitative aspects:

- millions or tens of millions of individuals, and hundreds or thousands of variables are analyzed:
- many variables are non-numerical, sometimes textual;
- weak assumptions are made about the statistical distributions followed;
- data are collected prior to the study, and often for other purposes;
- the populations studied are constantly changing;
- outliers (out of the norm, at least in terms of the distributions studied) are present;
- the data are imperfect, with data entry errors, coding errors, missing values;
- fast calculations are needed, sometimes in real time;
- we are not always looking for the mathematical optimum, but sometimes for the model that is easiest to understand for non-statisticians;

- the models are derived from the data and sometimes attempts are made to derive theoretical elements from them;
- some methods are starting to come from information theory or *machine learning;*
- used in companies, and not only in labs and universities.

Two major changes have occurred. On the one hand, the data analyzed often were not collected for the needs of the study and were collected for management needs, with all that this implies in terms of redundancy, imprecision, and even errors. Data mining involves a long and important work of "cleaning" the data prior to modeling. On the other hand, the aim is not essentially scientific, or even sometimes explanatory, but rather to assist in decision-making. It is not a question of formulating theoretical hypotheses and then verifying them with the help of observations, but of finding models that fit the observed data as well as possible, and can, in some very specific cases, suggest theoretical clues.

For example, in the commercial field, we do not try to elaborate a theory of customer behavior and its hidden motives, and we simply try to characterize the profiles of those customers most likely to buy a given product, and possibly know when and at what price.

A new era began in the 2010s when big data appeared, with hundreds of millions of individuals and thousands of variables, of all types, sometimes very noisy, collected in companies, objects, the Internet, to provide decision support, and also, more than with data mining, new services. The concept of big data was born out of the explosion in the production of all kinds of data: traces left on the Web (sites visited, videos seen, clicks, keywords searched for, etc.), opinions expressed in social networks (about a person, a company, a brand, a product, a service, a movie, a restaurant), posts in social media and content shared on websites (blogs, photographs, videos, music tracks, etc.), geolocation by GPS or IP address, information gathered by industrial, road and climate sensors, RFID chips, NFC devices and connected objects (smartphones, connected watches and wristbands, intelligent personal assistants,[3] cameras, electricity meters, household appliances, scales, clothes, medical devices, cars, glasses with augmented reality ...) that form what is called the *Internet of Things* (IoT).

These big data are characterized by what Doug Laney has called[4] the "three Vs": volume, velocity and variety.[5]

The volume of data involved has given Big Data its name, with an order of magnitude that can reach the petabyte (10^{15} bytes). The increase in data volume comes from the increase in:

- the number of individuals observed (more numerous or at a finer level);
- the frequency of observation and data recording (from monthly to daily or hourly);
- the number of observed features.

This increase also comes from the observation of new data, especially from connected objects, geolocation, and the Web.

Thus, in one minute on the Internet:[6]

- 120 new accounts are created on LinkedIn;
- 571 websites are created;
- 600 Wikipedia pages are created or updated;
- 1,500 blogs are posted
- 46,740 photographs are posted on Instagram;

- 154,200 calls are made on Skype;
- 752,000 dollars are spent online;
- $258,750 in sales are made by Amazon;
- 342,000 smartphone applications are downloaded;
- 456,000 tweets are sent;
- 900,000 people log on to Facebook;
- 4.1 million videos are viewed on YouTube;
- 3.6 million searches are made on Google;
- 16 million SMS are sent;
- 156 million emails are sent (including over 100 million spam emails).

This aspect of volume is perhaps the most visible and spectacular feature of big data, but it is not entirely new, since the retail, banking, and telecom industries have long been handling huge volumes of data, with annual numbers of transactions routinely exceeding one billion. However, if the number of objects or individuals processed (the rows of the databases) was very large, the number of their observed characteristics (the columns of the databases) was not so large, and this is rather where the novelty of big data lies, and their theoretical and practical difficulties. The main theoretical difficulty is the so-called "curse of dimensionality," to which we return in Section 1.5.1. From a practical point of view, it is sometimes said that big data begin when we can no longer load data into memory and, more generally, when we can no longer process it by "conventional" means (in a rather vague sense, which can, for example, refer to non-distributed computer architectures). What sounds like a joke is at the same time pragmatic, and refers to the technical and software tools that can manipulate data that cannot reside in memory.

The variety of big data is due to the fact that these data are of very diverse natures and forms: numerical, graphs, web logs, texts in various forms (documents, emails, SMS, etc.), sounds, images, videos, functional data ... This variety makes it difficult to use the usual databases and requires a variety of methods: graph analysis, deep neural networks, text mining, web mining ... Heterogeneous data can be crossed: for example, sales data matched with social network data, or pollution sensor data matched with weather data and traffic sensors.

The velocity of big data comes from the fact that these data are updated rapidly, sometimes in real time and in a continuous flow, or at least at high frequency, and must often be processed just as quickly. By data streams, we mean the continuous flow of data from industrial, meteorological, and astronomical sensors. An autonomous vehicle must be able to continuously and immediately process an enormous quantity of data, estimated at 30 terabytes per day. On merchant sites, the customer's decision on the Web is made quickly because it only takes one click to change site, so it is necessary to instantly make them the best commercial offer. Another example: credit card fraud must of course be detected in real time. In some cases, the limitation of the calculation time can induce an error to be estimated and controlled.

Note that in some cases, it is not only the application of the statistical model but its update that is done in real time or at least very frequently. Velocity thus has three components:

- data velocity;
- the speed of the treatments to be implemented;
- the speed of updating the models.

The concept of data science has recently developed to address the theoretical and technical challenges raised by big data. The modeling methods applied to big data are far from being new but they have seen strong motivation and important work related to big data. We can mention in particular some advanced techniques of sampling, optimization, machine learning, deep learning, penalized estimators of the Lasso or Danzig type, incremental learning, spatial statistics, the use of graphs, and of course the analysis and generation of natural language, both written and spoken.

Sampling issues are important, as they can help to reduce the volume of data to make it more easily manipulated, and to infer general conclusions from partial observations. But the representativeness of samples is difficult to establish, with multiple data sources, which do not cover the same populations and have a significant number of missing values. This raises problems of sampling techniques and sample adjustment. For example, it is necessary to assign appropriate weights to the observed units in order to obtain a representative sample of the population studied. There is also the issue of matching individual data from different sources and using auxiliary information. These methods are already used by the national statistical institutes (such as French INSEE), which uses numerous data sources (surveys, population census, and administrative files), as well as by media audience measurement institutes.

Incremental learning refers to methods that allow us to build a model, not on the basis of a complete sample, but on the basis of data that arrive in smaller chunks and from which we must update the model without having to take into account past observations. These methods make it possible to analyze data arriving in streams or which are too large to be processed at once. They are found in certain decision trees, the Very Fast Decision Trees, which rely on a theoretical limit, the Hoeffding bound, to determine a sufficient number of observations to obtain a split of each node of a tree close enough to the split that would have been obtained with all the observations. By this sampling, these trees can treat data so massive that the calculations would be too long or even impossible. Another widely used incremental algorithm is Alan Miller's "Memory Bounded Algorithm" AS 274 which is used for linear regression and is implemented in the `biglm` package of R. Incremental algorithms also exist in deep learning.

Machine learning methods, such as model aggregation (ensemble methods), support vector machines, and neural networks discussed in this book, are used for their high predictive power, in situations where model readability is not required and their "black box" characteristic is not an insurmountable obstacle, especially since some methods of explainability (Section 1.5.7) of their predictions are beginning to develop.

Deep learning is used for problems as complex as image, video, text and speech recognition, it uses sophisticated machine learning techniques such as convolutional neural networks, and it relies on massively parallel computer architectures, especially for computations with graphics processors, which we discuss later. Along with big data, deep learning is the main topic of this book because it is at the heart of modern data science. Most big data could not be analyzed without deep learning.

Whether the images are from social networks or medical imaging, advanced algorithms are needed to process them, not just to retouch or enhance them, but to recognize faces, places, or tumor cells. Unfortunately, deep learning can also create fake images or videos, manufacturing *deepfakes*. "Multimodal artificial intelligence" methods, combining image,

voice and text recognition, are being implemented to try to flush out these deepfakes. Deep learning is also becoming indispensable for processing data as numerous and complex as those produced by genetic analysis, astronomy, or particle physics. Deep learning has also made it possible to achieve unequaled performance in text and voice processing, whether it is answering human questions in writing or orally, or helping users in their searches on the Web: today, huge transformer neural networks (Section 9.7) analyze the requests of Internet users to find the most relevant information.

1.2 Examples of the Use of Big Data and Deep Learning

Examples of the use of big data and deep learning abound in a wide variety of sectors. This section gives a quick overview before we come back to discuss several of them.

In the field of transport, it is about the improvement of road traffic by geolocation, the search for free parking spaces, the billing of parking in paying areas thanks to the reading and the optical recognition of characters on the license plates, the dynamic fixing of the price of airline tickets.

This last application is part of dynamic pricing, also called yield management or revenue management. It concerns activities with fixed available capacities, high fixed costs, perishable products (to be sold before a certain date) and that can be sold in advance at differentiated prices. Dynamic pricing determines in real time the appropriate quantities to put on sale, at the appropriate price, at the appropriate time, in order to maximize the profit generated by the sale. It is about maximizing margin without reducing demand. It was born in the 1980s in the airline industry with American Airlines, but has since spread to many other areas of transport, hotels, advertising space, tourism, entertainment. For example, it is applied by Uber to regulate the number of drivers behind the wheel at a given time: lowering fares when supply exceeds demand should encourage Uber drivers not to drive at that time. Dynamic pricing is also used in fashion, where it is more difficult to implement for new products with less predictable sales.

To take the example of air transport, here is what Stéphane Ormand, head of the "revenue management" department at Air France-KLM Group, says about dynamic pricing in an interview in *La Croix* on June 27, 2016:

> Nearly a year in advance, they [pricers] virtually carve up an aircraft for each flight into a multitude of fares, taking into account a multitude of factors – everything from the economic crisis in Brazil to falling exchange rates in Venezuela or Nigeria, or the Euro soccer tournament in France ... Analysts must keep a constant watch on the competition, local or global events that can disrupt demand for a given flight and adapt the sales strategy.

A single economy class cabin can be divided into fifteen or so fares, with fare differentials of 1 to 10, and the number of seats between these fares can be adjusted according to demand.

In a sense, dynamic pricing is a return to the age-old practice of prices that were not fixed and posted, but were the result of negotiation. Here, we cannot speak of negotiation

because the price is set without the buyer's knowledge, but it is the overall behavior of all buyers that influences the price. Dynamic pricing must be legally regulated so as not to discriminate against certain customers on the basis of their individual profile or the place where they live.

In marketing, geolocation allows you to send a promotion or a coupon on your smartphone when you pass near a business. This geolocation (or *geofencing*) uses the GPS of smartphones, Wi-Fi hotspots, beacons.

Another marketing application of big data is the analysis of preferences, recommendations, possibly linked to sales data, to target consumers more efficiently (Section 3.6).

A classic in the retail industry is the analysis of receipts and its cross-referencing with loyalty program data. Consider that these analyses are done in real time, and that it is at the moment of checkout that the customer is warned, based on the contents of their shopping cart, that they may have forgotten to buy an item. There is a lot of talk about Internet big data, but 90% of sales in the world are still made in physical stores, and Wal-Mart remains the leading retail group in the world, even if Amazon comes in third place in Deloitte's Top 250 in 2020.[7] It was in sixth place in 2016 and in 186th place in 2000. Moreover, it should be noted that Amazon is starting to open physical stores and to make agreements with mass retail chains. Under the terms of an agreement between Google and Carrefour, since 2020, Google's voice assistant can be used to order purchases at Carrefour, through a smartphone or a connected speaker. Such an agreement was made in August 2017 in the United States between Google and Wal-Mart. With big data, physical stores can try to exploit their advantages over online businesses.

Combining the two types of businesses, Amazon Go is a physical store that relies on sensors to know which items have been picked up by customers (and not put back on the shelf). Payment is made automatically when leaving the store, if one has an Amazon account and a smartphone with the Amazon Go app. Like online visits and purchases, those made in this store can be tracked by Amazon, which can use them to derive relevant recommendations.

Perhaps less mediatized than the previous applications of big data, the scientific applications are important and varied in fields that are big users of massive data, such as meteorology, seismology, oceanography, ecology, genomics, epidemiology, medical imaging, astronomy, and nuclear physics.

For example, the Large Synoptic Survey Telescope (LSST) records 30 terabytes of images each day, with real-time alerts for changes in the position or brightness of celestial objects. The Large Hadron Collider (LHC) used to discover the Higgs boson records 60 terabytes of data every day, at each of its 100 million collisions per second. The Discovery supercomputer at the NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate data and climate model simulations.

In literature, the millions of digitized texts cover the entire history of literature and can be analyzed for comparative literature or to study the diffusion of cultural movements in a society, according to the historical or social context. They are also available and downloadable on sites such as Gutenberg[8] for texts in the public domain. In archeology, computer vision can be applied to the identification of patterns on ancient objects, their comparison with other known objects, their dating, and the study of their technical and stylistic evolution.[9] These applications of deep learning are part of the disciplines called *digital humanities*.

Artificial intelligence comes to the aid of paleography, that is, the study of ancient writings, to distinguish minute differences in the style of two texts. Thus, in April 2021, scientists from the University of Groningen in the Netherlands published an article in *PLOS One* revealing that two scribes, and not just one, had probably written one of the Dead Sea Scrolls, the Great Isaiah Scroll.

In journalism, data journalism relies on the analysis of digital data to verify, produce, and distribute information. These data are increasingly numerous, notably thanks to Open Data (Section 1.7) and social networks (Section 5.1). This is different from infographics, which only aims to represent data. It can also rely on natural language processing tools, machine translation, and automatic text generation to comment on data. Among the applications of data journalism, we can mention the identification of individuals or events in photographs, the detection of fake photographs or videos, the detection of fake news and fact-checking, the exploration and analysis of archives and large databases, the automatic monitoring of social media to detect emerging topics, the "breaking news" to be announced before the others, the analysis of political speeches, or the detection of propaganda, of influences. We can also mention the recommendation systems (Section 3.6) that suggest to readers new articles likely to interest them, based on those they have already read.

In the field of human resources, job boards rely on machine learning methods, and companies use résumé analysis enriched by searching CV libraries, detecting links made by the candidate on social networks, events in which he or she participates, his or her career path, etc. This sourcing allows start-ups to detect candidates who best match the position offered and the company's culture more quickly than headhunters, and at a lower cost. Interview reports with employees can also be analyzed to automatically detect positive or negative tones, situations of discomfort, the expression of difficulties. More recently, AI has been used to automate telephone or video-conference interviews with candidates, but with results that are sometimes unreliable in evaluating the aptitudes of candidates for a position. Recruitment is a field where particular attention is paid to the absence of bias and discrimination.

Applications of big data to education are the development of applied technologies to education (EdTech) and the analysis of social networks to learn about the popularity of lessons and student satisfaction, and to adapt teaching to the progress of each student.

In computer science, big data tools are used to monitor machines and networks, and to detect failures or security incidents. The Swiss company DeepCode uses artificial intelligence trained on open source lines of code to help programmers, detecting bugs in their lines of code and suggesting fixes based on other programmers' fixes. It does this for JavaScript, TypeScript, Java, C/C++, and Python. In 2021, the French startup SourceAI uses GPT-3 (Section 9.7) to translate natural language queries (in English, French, German, or Spanish) into computer code in over 40 programming languages. Some envision the end of the programmer's job in a few decades, as programming can be done entirely by artificial intelligence.

In June 2021, Google engineers announced that they had successfully used reinforcement learning (Section 7.1) to design the layout of their next TPU chips (Section 2.6.2), i.e., to organize the arrangement of the billions of transistors that make them up, so as to minimize the surface area, the length of the wires, and the power consumption, and to maximize the performance.[10] Six hours of calculations produce a result at least as good as several months of work by specialized engineers.

In their hunt for fraud, tax authorities are increasingly relying on information found on the Web and social networks, which can reveal a taxpayer's lifestyle, address, commercial and financial transactions, more effectively than reading the celebrity pages of magazines.

Security, along with video surveillance and intelligence, is also increasingly using big data and artificial intelligence tools, with the possible drifts mentioned in Section 1.4.3. Some intelligence agencies admit to using artificial intelligence to prevent cyberattacks, to fight against disinformation by detecting deepfakes, to combat security threats and crime, while relying on ethical principles. But security can also be endangered by big data, whether it is the myriad of connected objects that can fall under the control of malicious people (to direct them in server attacks, for example), or even connected autonomous vehicles (Section 1.4.2) that could also be remotely controlled and used in terrorist attacks. Since 2016, the Mirai malware has been regularly used to infect connected objects such as video surveillance cameras to form networks, botnets, capable of massive attacks on servers, in the form of message streams saturating those servers (known as a *denial of service attack*). The advent of 5G and the growth in the number of connected objects have led to fears that the frequency of these attacks will increase. Video surveillance facial recognition is also of interest to banks, which are beginning to see it as a way to identify dangerous individuals and enhance the security of their employees. Facial recognition algorithms were first defeated by the wearing of health masks during the Covid-19 crisis, before being refined to work under these new conditions, and they can also be used to detect the presence of people without masks.

Big data are present even in sports, in soccer, basketball, baseball or tennis, in the collection and analysis of data on the position, the movements of the ball and the players, their heart rate, the history of the matches played, these big data being used to develop new game tactics, to predict the results of competitions, to improve performance, to study the trajectories on the field, to measure the impact of fatigue, to limit the risks of injury, or to adapt training. They were used by the German national team during the 2014 World Cup, and we can also mention the use of data from the Hawk-Eye system on the trajectory of balls and players in sports such as cricket, tennis, badminton or volleyball. A lot of data has also been collected for a long time in motorsports. The collection and communication of physiological data from sportsmen and women to their coaches and team managers, and even to sports journalists, raise obvious confidentiality problems.

1.3 Big Data and Deep Learning for Companies and Organizations

The contribution of big data to business is fourfold:

- increase in audience and market share, loyalty, with, for example, the increase in sales induced by recommendations;
- decision support, with the example of forecasting stock market and macroeconomic trends using the analysis of web searches and social media posts;
- improving operational processes, for example, by forecasting the company's needs and the price of raw materials (according to various meteorological and economic criteria) over time, allowing the company to purchase the right quantities at the right time, or

forecasting store traffic according to weather conditions and external events, or even predictive maintenance;
- the creation of new innovative products and services, in which the customer himself benefits from a decision-making system, whether for downloading films or music (recommendation systems) or for driving, as we will see in Section 1.4.2.

With big data, we are witnessing a shift from analytics to decision-making. Data are no longer used only to help the company make decisions, but are used to offer new products and services to the customer, as personalized as possible, and to set prices with precision thanks to yield management.

The second industrial revolution with the invention of the assembly line by Henry Ford in 1913 led to mass production and a standardization of the offer, thus to mass consumption, mass media, and mass culture. In the opposite direction, big data allow companies to adapt their offer more specifically to each customer and allow a tailor-made production. Moreover, it is no longer just a matter of performing complex analyses on large volumes of data, but of creating new functions at the heart of real-time operational applications. However, there is a caveat to this optimistic observation: recommendation systems tend to lock people into profiles determined by algorithms and only offer them products, contracts, and information that are supposed to correspond to their profile. This is what is called algorithmic containment (Section 10.5).

1.3.1 Big Data in Finance

Big data can be used in the financial industry to analyze stock market risk, financial risk and fraud risk.

On the subject of stock market risk, a study published in *Scientific Reports*[11] in 2013 shows a correlation between certain keywords entered on Google and the evolution of stock market prices: before a fall in stock market indices, investors are concerned and search the Web for information helping them decide whether to hold or sell their shares. We will come back to this example later.

In the analysis of financial risk, social media posts and press articles can provide information on the sentiment of economic operators. For example, what people say about a company, its image among its partners, financial analysts or the general public, its difficulties, its reputation, its image in terms of quality, innovation, social and environmental respect, all these elements can provide information on its financial health and be integrated into the analysis. The management of certain funds relies on this type of analysis, also provided by economic information specialists such as Ellisphere (formerly Coface Services).

For credit card fraud risk detection, geolocation data of smartphone holders can be compared with payment terminal information to ensure consistency.

1.3.1.1 Google Trends

Google Trends[12] allows us to see how often a term has been searched on Google, with a visualization by region of the associated searches and certain headlines in the news. The frequencies associated with several terms can be overlaid as in Figure 1.1. The data can be downloaded in CSV (comma-separated values) format. Google Trends is used in various fields such as health (Section 1.4.1) and financial markets.

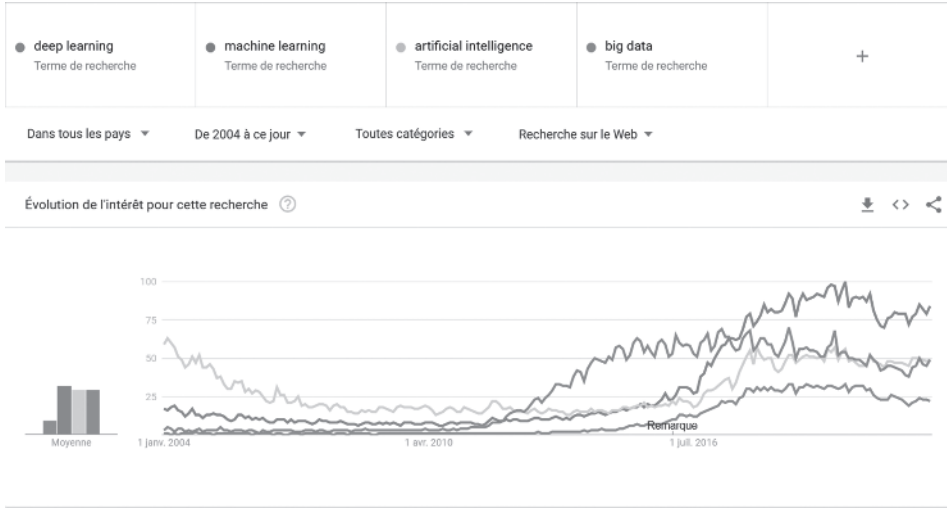

Figure 1.1 Google Trends.

1.3.1.2 Google Trends and Stock Prices

The use of Web searches for forecasting stock market prices is based on the principle that decision-making is preceded by a search for information, which is increasingly done on the Web.

Here is how the study published in *Scientific Reports* (2013) was conducted. From January 2004 to February 2011, 98 terms were analyzed on Google Trends: debt, stocks, portfolio, inflation, Dow Jones, etc.,. and for each term, the proportion $n(t)$ of the total searches of week t was recorded. During the same period, the closing prices $p(t)$ of the Dow Jones Industrial Average (DJIA, often shortened to Dow Jones, the oldest index of the New York stock exchanges) were taken at the beginning of each week t.

The hypothesis tested is that $p(t+1)$ is correlated with $n(t)$, specifically with:

$$\Delta n(t, \Delta t) = n(t) - \text{average of } n(t) \text{ between } t-1 \text{ and } t-\Delta t,$$

with Δt varying between 1 and 6 weeks in the study.

The test was translated into a "Google Trends strategy" consisting of:

- sell the DJIA at the price $p(t+1)$ on the 1st day of week $t+1$ if $\Delta n(t, \Delta t) > 0$, then buy the DJIA at the price $p(t+2)$ at the end of the 1st day of week $t+2$;
- buy the DJIA at price $p(t+1)$ on the 1st day of week $t+1$ if $\Delta n(t, \Delta t) < 0$, then sell the DJIA at price $p(t+2)$ at the end of the 1st day of week $t+2$.

The highest gain, 326% between 2004 and 2011, was obtained with the term "debt" and $\Delta t = 3$, while a "Dow Jones strategy" of replacing $\Delta n(t, \Delta t)$ with $\Delta p(t, \Delta t)$ ('buy at the sound of the gun and sell at the sound of the bugle') yields a gain of only 33% over the same period.

1.3.1.3 The `quantmod` Package for Financial Analysis

Stock market and currency prices and other financial information can be found in sources such as Yahoo Finance. Quantitative data are represented by symbols, such as ^FCHI for

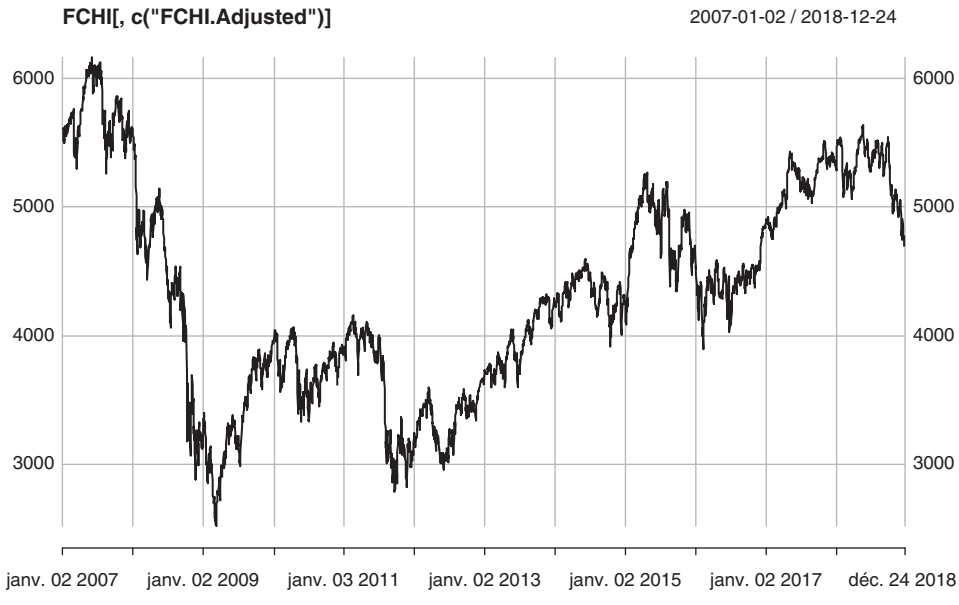

Figure 1.2 Stock market price.

the French CAC 40. The R `quantmod` package allows one to access, analyze, and represent these financial data. Here is how we can retrieve the CAC 40 prices between 2007 and 2018, and display them (Figure 1.2).

```
> library(quantmod)
> getSymbols("^FCHI", src="yahoo")
> head(FCHI)
           FCHI.Open FCHI.High FCHI.Low FCHI.Close FCHI.Volume FCHI.Adjusted
2007-01-02   5575.76   5621.65  5575.63    5617.71    85910000       5617.71
2007-01-03   5621.00   5623.67  5596.82    5610.92   118580700       5610.92
2007-01-04   5573.73   5585.54  5547.17    5574.56   130465700       5574.56
2007-01-05   5552.64   5566.24  5517.35    5517.35   126420500       5517.35
2007-01-08   5532.57   5555.67  5509.06    5518.59   115053800       5518.59
2007-01-09   5549.01   5563.48  5532.62    5533.03   151688200       5533.03

> plot(FCHI[,c("FCHI.Adjusted")], type="l")
```

For the rest of our analysis, we transform the `quantmod` data into a data frame, and remove any rows with missing data. This is the case here for May 1, 2014. We put the date, which is the name of the rows, in a specific variable, and we extract the month.

```
> cac40_value <- as.data.frame(FCHI)
> cac40_value <- cac40_value[which(!is.na(cac40_value$FCHI.Volume)),]
> cac40_value <- cbind("date" = rownames(cac40_value), cac40_value)
> rownames(cac40_value) <- NULL
> cac40_value$mois <- substr(cac40_value$date,1,7)
> head(cac40_value)
         date FCHI.Open FCHI.High FCHI.Low FCHI.Close FCHI.Volume FCHI.Adjusted    mois
1  2007-01-02   5575.76   5621.65  5575.63    5617.71    85910000       5617.71 2007-01
2  2007-01-03   5621.00   5623.67  5596.82    5610.92   118580700       5610.92 2007-01
3  2007-01-04   5573.73   5585.54  5547.17    5574.56   130465700       5574.56 2007-01
```

```
4  2007-01-05   5552.64   5566.24  5517.35    5517.35   126420500       5517.35 2007-01
5  2007-01-08   5532.57   5555.67  5509.06    5518.59   115053800       5518.59 2007-01
6  2007-01-09   5549.01   5563.48  5532.62    5533.03   151688200       5533.03 2007-01
```

We then calculate the monthly average of the adjusted closing price, and transform the month-year into a 01-month-year date of the usual POSIXct class. This data frame can be merged with Google Trends data.

```
> cac40_monthly <- aggregate(cac40_value[,"FCHI.Adjusted"],
   list(date=cac40_value$mois), mean)
> cac40_monthly$date <- as.POSIXct(paste(cac40_monthly$date,
   "01", sep="-"), format='%Y-%m-%d', tz= "UTC")
> colnames(cac40_monthly) <- c("date", "value")
> head(cac40_monthly)
        date    value
1 2007-01-01 5586.986
2 2007-02-01 5684.217
3 2007-03-01 5495.406
4 2007-04-01 5831.803
5 2007-05-01 6053.903
6 2007-06-01 6014.926
```

1.3.1.4 Google Trends in R

The `gtrendsR` package allows access to the Google Trends statistics, i.e. the frequency of queries normalized on a scale from 0 to 100,[13] by entering one or several search keywords, and possibly a region and dates. With the first versions of the package on CRAN, up to version 1.3.5, you had to have a Gmail address and log in with your address and password. Around mid-2017, the Google Trends API[14] changed and the `gtrendsR` package was updated accordingly. It no longer requires authentication, but it has the disadvantage of only providing monthly, not weekly, series if you go back more than five years. A Python package exists, `pytrends`, with the same parameters as the R package.

You can specify a geographical area and a resolution:

- with a resolution of "1h," "4h", "1d," "7d," Google Trends returns data for the last one hour, four hours, one day, and seven days;
- a resolution not shown is weekly for any period between three months and five years, and daily for any period between one and three months.

A parameter (`gprop`) of the `gtrends()` function can specify the Google product on which the search should be performed: Web (default), news, images, YouTube... Here, we are interested in web searches, and we extract the search statistics in France between January 1, 2007 and December 31, 2015.

```
> library(gtrendsR)
> cac40_trend <- gtrends("CAC 40", geo="FR", time="2007-01-01
   2015-12-31")
> head(cac40_trend$interest_over_time)
```

```
        date hits keyword geo gprop category
1 2007-01-01   10  CAC 40  FR   web        0
2 2007-02-01   10  CAC 40  FR   web        0
3 2007-03-01   12  CAC 40  FR   web        0
4 2007-04-01   11  CAC 40  FR   web        0
5 2007-05-01   10  CAC 40  FR   web        0
6 2007-06-01    9  CAC 40  FR   web        0
> plot(cac40_trend)
```

The `plot()` function displays the evolution of the search frequencies of an expression on Google over time (Figure 1.3).

1.3.1.5 Matching Data from `quantmod` and Google Trends

We now merge the adjusted closing prices and the queries frequencies on Google Trends in order to obtain a double set of monthly values.

```
> cac <- merge(cac40_monthly, cac40_trend$interest_over_time[,1:2])
> head(cac)
        date    value hits
1 2007-01-01 5586.986   10
2 2007-02-01 5684.217   10
3 2007-03-01 5495.406   13
4 2007-04-01 5831.803   10
5 2007-05-01 6053.903   10
6 2007-06-01 6014.926    9
```

As explained above, there is a negative correlation between the frequency of queries on Google Trends and the CAC 40 prices, slightly stronger for the correlation of ranks.

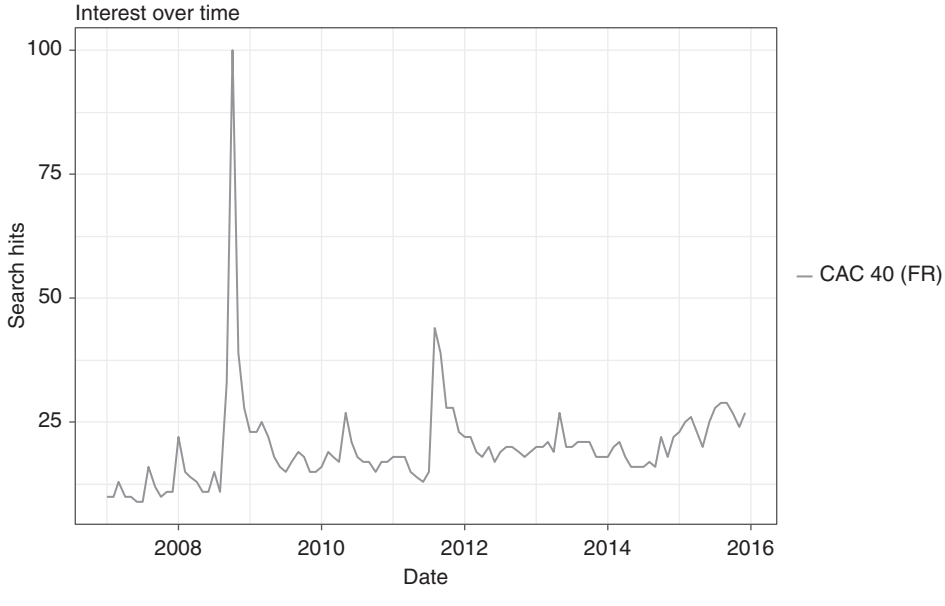

Figure 1.3 Google search queries.

```
> cor(cac$value, cac$hits)
[1] -0.3378188
> cor(cac$value, cac$hits, method="spearman")
[1] -0.3956836
```

By transforming the previous data frame into a multiple time series using the `ts()` function, we can search for the lag of the Google search series that maximizes its linear cross-correlation with the CAC 40 price series. We indicate that the series starts in January 2007 and that it is monthly. The graphs of the two time series can be plotted on the same time axis (Figure 1.4) or the two time series can be overlaid on the same graph (Figure 1.5) by first standardizing (centering and scaling) the data with the `scale()` function in order to put the two series on the same scale.

```
> cac.ts <- ts(data = cac, start = c(2007,1), frequency = 12)
> cac.ts
               date    value hits
Jan 2007 1167606000 5586.986   10
Feb 2007 1170284400 5684.217   10
Mar 2007 1172703600 5495.406   13
Apr 2007 1175378400 5831.803   10
May 2007 1177970400 6053.903   10
Jun 2007 1180648800 6014.926    9
…
> plot(cac.ts[,-1])
> plot(scale(cac.ts[,-1]), plot.type = "single", lty = 1:2)
```

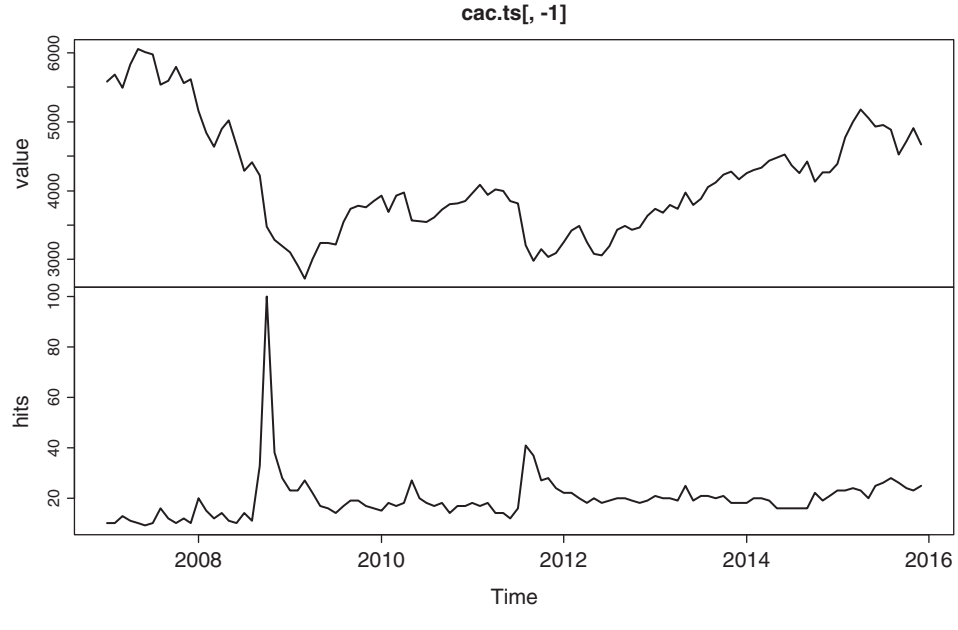

Figure 1.4 Stock market prices and search queries time series.

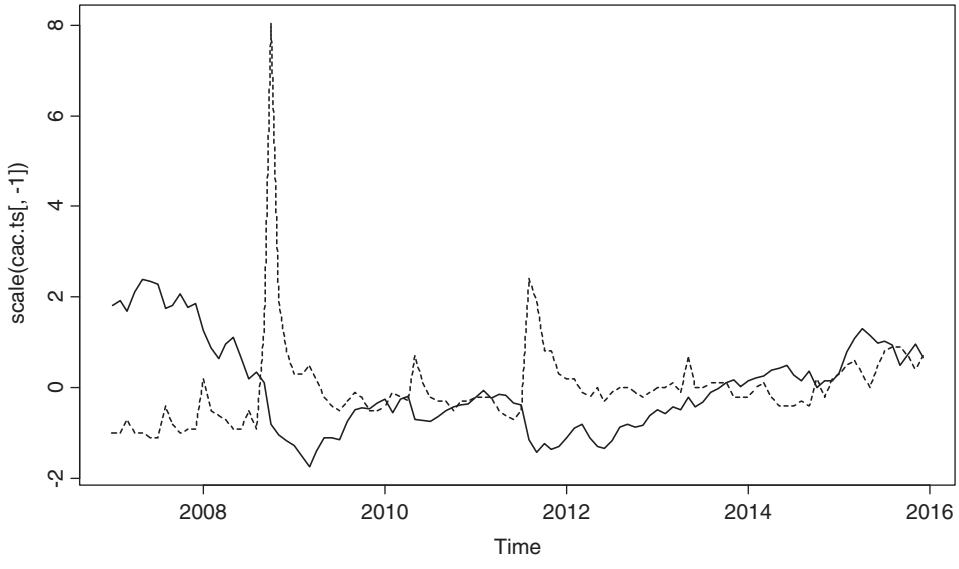

Figure 1.5 Stock market prices and search queries overlay.

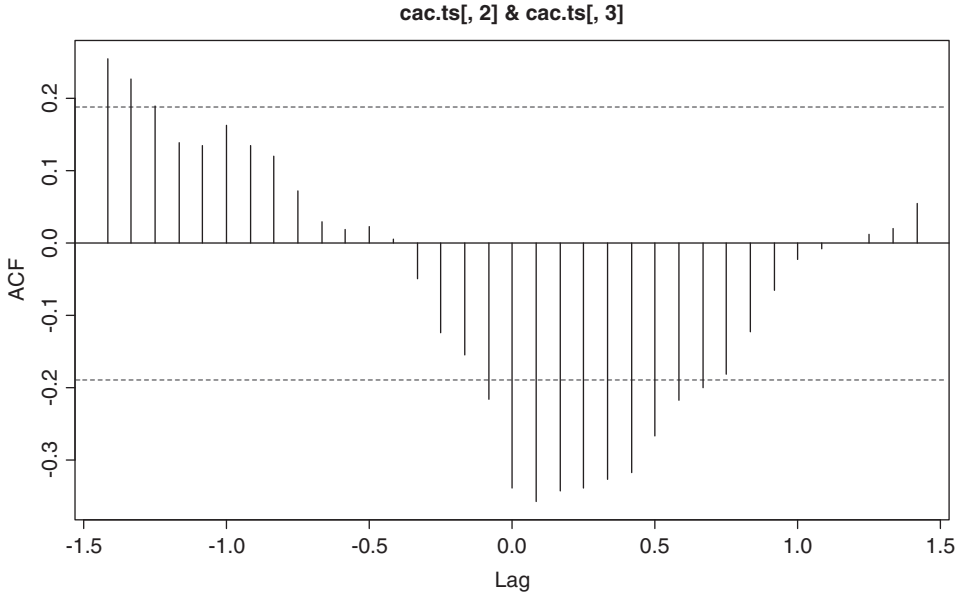

Figure 1.6 Correlogram.

We plot the correlogram (Figure 1.6) and we see that the linear correlation is maximal (in absolute value) when the trend series precedes the CAC 40 price series by one month. This correlation is then −0.357.

```
> cross <- ccf(cac.ts[,2],cac.ts[,3], type = c("correlation"),
   plot = TRUE)
> cross$lag[which.max(abs(cross$acf))]*12
[1] 1
> cross$acf[which.max(abs(cross$acf))]
[1] -0.3569695
```

We can also calculate the Spearman correlogram of correlations and see that the correlation is maximal (in absolute value) when the trend series precedes by one month that of the CAC 40 prices.

```
> cross <- ccf(rank(cac.ts[,2]),rank(cac.ts[,3]),
   type = c("correlation"), plot = TRUE)
> cross$acf[which.max(abs(cross$acf))]
 [1] -0.3963044
```

This rank correlation equal to -0.396 is not surprisingly stronger than the linear correlation.

The following calculation shows that the correlation is less strong in 2012, 2013, and 2015. In fact, we even see that the direction of the correlation reverses slightly in 2012 and cancels in 2013. These differences in correlation across years can be seen in Figure 1.5.

```
> library(lubridate)
> cac$an <- year(cac$date)
> sapply(2007:2015, function(x) {cor(cac[which(cac$an==x),]
   $value, cac[which(cac$an==x),]$hits, method="spearman")})
[1] -0.8297550 -0.6208354 -0.6926184 -0.5920817 -0.6678821
      0.1043598  0.0000000 -0.6250023 -0.3697275
```

The same calculation can also be done using annual windows in the time series.

```
> sapply(2007:2015, function(x) {scac <- window(cac.ts,
   start=c(x,1), end=c(x,12)) ; cor(scac[,2],scac[,3],
   method="spearman")})
[1] -0.8297550 -0.6208354 -0.6926184 -0.5920817 -0.6678821
     0.1043598  0.0000000 -0.6250023 -0.3697275
```

Finally, we can check that the CAC 40 prices are correlated to other Google searches, but the `gtrends()` function only allows at most five terms simultaneously to be queried.

```
> cac40_trend_all <- gtrends(c("crise", "dette", "chômage",
   "inflation", "croissance"), geo="FR", time="2007-01-01
   2015-12-31")
> CAC <- merge(cac40_monthly,
   cac40_trend_all$interest_over_time[,1:3])
> sapply(c("crise","dette","chômage","inflation","croissance"),
   function(x) {cor(CAC[which(CAC$keyword==x),]$value,
   CAC[which(CAC$keyword==x),]$hits, method="spearman")})
      crise       dette     chômage   inflation  croissance
-0.50732425 -0.19880401  0.20049101 -0.08707338  0.19189605
```

The negative correlation with "crisis" is expected, but the positive correlation with "unemployment" is more surprising.

1.3.2 Big Data and Deep Learning in Insurance

Big data are present in both branches of insurance: property and casualty (P&C) insurance, and health insurance.

In P&C insurance, auto insurance, with the possibility of geolocating insured drivers, is very interested in the contribution of big data.

Some insurers have developed a smartphone application that analyzes the driving style of drivers in order to offer them appropriate rates. This application analyzes the number of kilometers driven, the time, the type of road, and takes measurements (acceleration, braking, speed in curves, and speed in relation to traffic) allowing the calculation of a monthly score influencing the pricing. GPS data can also be used. You have to be careful about a radical change in behavior, which could lead to the suspicion of fraud.

Some apps allow you to analyze your driving in order to adjust your insurance premium and offers automatic assistance calls in case of an accident.

Sensors on the car could even signal the risk of breakdown, telling the driver what to do and where to find the nearest breakdown mechanic. These solutions are in the interest of the insurer and the insured (and society): reduction of the risk of breakdown and accident.

Another application of deep learning algorithms for computer vision is the analysis of photos of crashed vehicles to estimate their damage and repair costs. This automatic analysis is not always efficient enough but it is very fast and can help experts in their work.

In health insurance, "pay as you live" contracts are developing in the same way, the aim of which is to suggest lifestyles deemed better for the insured and the insurer. In order to know if you have a really healthy lifestyle, the insurer could be tempted to look at your activities as they appear on social networks, at least as soon as the progress of artificial intelligence will make it possible to use these data in a sufficiently fast and reliable way.

1.3.3 Big Data and Deep Learning in Industry

The numerous sensors (temperature, pressure, vibration, wear) placed on the components of a machine, an elevator, an engineering structure (bridge, tunnel, building) or a transport vehicle permit data to be retrieved in real time and remotely, which, once analyzed and modeled, can provide a probability of failure, of breakage of a part, and allow an arbitration between:

- unnecessarily heavy and frequent maintenance operations, leading to unnecessary expenses and unavailability of the productive apparatus;
- insufficient maintenance operations, leading to costly and even dangerous failures.

The interest in these predictive maintenance devices is the fact that the measured stresses are the real ones and not the ones foreseen in the design. Predictive maintenance allows us to intervene earlier, at the first signs of weakness, and therefore at a lower cost and for less time than if we wait for the actual degradation of the device. Some estimate that predictive

maintenance and the industrial internet of things could save several hundred billion euros per year worldwide.

In addition to productive devices, examples of applications are of course aeronautics (the sensors on an airplane generate 1 gigabyte of data during a Paris-New York flight), automotive, bridge monitoring, with their tension, corrosion, temperature and wind speed sensors, and smart meters.

Smart meters allow real-time prediction of consumption of electric energy, but also of malfunctions, and faster and more economical billing. These smart meters can measure and transmit in near real time (at short intervals throughout the day) the energy consumption of a building.

The uses of smart meters are many: to monitor networks in real time, to balance supply and demand, to regulate consumption, to read consumption remotely and to produce accurate bills, to identify breakdowns and defaults, and to refine forecasting models. It is a component of smart grids (intelligent electrical networks).

The data transmitted continuously (readings every ten minutes) are big data in their scope, and they will provide fairly accurate information on the lifestyle of the building's occupants, especially since it is planned that various connected objects will communicate their data to the smart meter: the boiler, thermostat, heat pump, radiators, etc. It will be possible to identify devices in operation from their load curve.

Examples of big data use abound in the power generation and distribution industry:

- sensors in the plants;
- measures on the networks (regulation, optimization);
- results of simulations (production planning);
- connected devices' load curves ;
- letters from customers ;
- Web site log files ;
- social networks, forums ;
- telephone records.

Voluminous data, coming from various sources and often in a continuous flow (electrical consumption, sensors), thus big data, intervene in the following:

- smart grids: smart meters, electric vehicles;
- smart home: energy management, Internet of Things, connected speakers;
- smart cities: traffic management, lighting, energy management.

Smart meters also exist to measure water or natural gas consumption, and they can detect abnormal consumption and especially leaks.

Other contributions of big data are the optimization of the supply chain and the improvement of the design of future aircraft.

Deep learning is used in computer vision, useful in particular to learn how to distinguish a defective part from a compliant part. This is a binary classification problem on images, as we will see in Section 8.3, except that here it is not a question of distinguishing animals or objects, but defective parts or not. The learning of a deep neural network is done on a sample of images of already classified parts. These models are used in real time during the welding

of parts monitored by a camera. The camera is connected to an algorithm that triggers an alert if a bad weld is detected. The stakes are high when we know that a defective part can cause a breakdown or an accident, and cost the manufacturer a lot of money, and that at the same time there are not enough human inspectors to ensure this level of surveillance. This system is used in particular by automotive and agricultural equipment manufacturers. Other manufacturing processes can benefit from the same type of algorithm.

1.3.4 Big Data and Deep Learning in Scientific Research and Education

The applications of big data in science are not the most publicized but are not the least important.

1.3.4.1 Big Data in Physics and Astrophysics

Particle physics and astrophysics are interested in objects that are so numerous that the contribution of big data is decisive. The DEUS (Dark Energy Universe Simulation) project seeks to model the evolution of the entire observable universe, from the Big Bang to the present day. It applies the standard cosmological model to simulate the behavior of 550 billion particles, in order to gain a better understanding of the nature of dark energy and its influence on the structure of the universe, as well as the origin of the distribution between dark matter and galaxies. The calculations were performed on the CURIE supercomputer of the GENCI consortium (Grand Équipement National de Calcul Intensif), which includes the French government (49%), the CEA (20%), the CNRS (20%), universities (10%) and INRIA (1%). Housed at the CEA, CURIE has 92,000 processors and a capacity of 2 petaflops: it is one of the five most powerful computers in the world. The calculations generate 150 petabytes of data but only 1 petabyte of data needs to be stored. The first simulations were carried out in April 2012 and won the Big Data prize from the American magazine *HPCwire* in 2013.[15] The *Large Hadron Collider* (LHC) is also a huge source of data: 60 terabytes per day.

Modern telescopes provide a gigantic quantity of images and therefore of data: for example, the *Very Large Telescope* (VLT), in fact is composed of four optical telescopes in the Atacama desert in the north of Chile, which each collect 15 terabytes per night. We can also mention the future Vera C. Rubin Observatory (or LSST: *Large Synoptic Survey Telescope*) built in Chile, which will record 30 terabytes of data per night, with a camera of 3.2 billion pixels, the largest existing digital camera. From 2023, it will provide the precise position of more than 10 billion galaxies and new information on dark matter and dark energy, but to do this will require computing power adapted to big data.

As for Euclid, the space mission of the European Space Agency (ESA), whose objective is to understand the origin of the acceleration of the expansion of the universe and the nature of its source: dark energy, Euclid will be based on a telescope to be launched in 2022 and will be able to observe ten billion astronomical sources in visible light and in the infrared, going back 10 billion years to the period when dark energy played a significant role in the acceleration of the expansion of the universe. This telescope will produce in ten years about 10 petabytes of images that will be analyzed in nine computing centers working for more than one hundred European laboratories, and will provide a wealth of data to astronomers for decades to come.

1.3.4.2 Big Data in Climatology and Earth Sciences

The study of the climate, the forecasting of its evolution and in the shorter term the weather forecasts naturally involve huge amounts of data. Examples include the Discovery supercomputer at the NASA Center for Climate Simulation (NCCS), which stores 32 petabytes of climate data and climate model simulations. In 2017, the UN launched the *Data for Climate Action* challenge to encourage researchers to use big data to find solutions for climate change mitigation and adaptation.[16]

In addition, work is being done to estimate the probability of an earthquake occurring at a given location.

1.3.4.3 Big Data in Education

Data analysis is increasingly being used to monitor students, particularly in distance learning which has recently become widespread. Faced with the difficulties of this type of teaching, where the teacher does not see the student, data analysis is used to adapt teaching to the student's progress and to identify the risks of dropping out and send alerts to the teaching team. The objective is to increase the success rate by individualizing the teaching.[17] This analysis can be based on the time spent, the way of following the program, of watching videos, of searching on the Web ... and even on the analysis of clicks and data entry. It is important to beware of crossing the line between these analyses and those of student behavior or automatic evaluation, which is tempting to reduce costs but carries the risk of incorrect evaluations.

These methods can be applied in particular to MOOCs ("massive online open courses"), which aim to bring together as many learners as possible online. OpenClassrooms[18] was created in 1999 (under the name "Le site du Zéro") and offers free online courses accessible to all, on computer science, science, and business. France Université Numérique[19] (FUN) is a MOOC that was launched in October 2013 and has offered about 100 courses since January 2014. One of the best-known MOOCs, Coursera,[20] offers several hundred courses.

1.4 Big Data and Deep Learning for Individuals

Following the previous section, which looked at big data applications from the perspective of companies and organizations, this section takes the perspective of individuals, in whose lives big data are becoming increasingly important.

1.4.1 Big Data and Deep Learning in Healthcare

Just as medicine has long made extensive use of statistics, known as biostatistics, so it was one of the first users of deep learning, particularly in medical imaging, and of big data generated by medical devices and measuring instruments, the processing of which is of course of capital interest for public health. The use of big data is part of the *4 P's of medicine*: personalized, preventive, predictive, and participative.

1.4.1.1 Connected Health and Telemedicine

Big data are first encountered in remote medical monitoring, for example, for the detection of heart attack risks, or Diatelic for the remote monitoring of dialysis patients. DeepMind

has developed the Streams application, which analyzes the medical data of patients suffering from kidney disease, in real time, in order to alert their doctors of a risk of acute renal failure.

Remote monitoring tools, or even simple applications for smartphones, can analyze the data transmitted by sensors (heart rate, blood pressure, oxygen saturation, blood sugar, temperature, etc.). They can reduce the costs of patient care while improving their quality of life, especially for diabetics. In 2018, the first connected watch with an electrocardiogram sensor appeared. Artificial intelligence is increasingly present, through numerous start-ups, to analyze medical images, electrocardiograms or genetic data, optimize radiotherapy treatments, or reduce the duration of MRI examinations.

A distinction must be made between connected health, or eHealth, and telemedicine. Connected health is reserved for minor ailments and the monitoring of patients by themselves, using algorithms. It can also simply help to adopt a healthier lifestyle or increase the autonomy of the elderly. Telemedicine is aimed at serious illnesses, focuses on clinical aspects and uses much greater resources, with secure networks permanently connected to medical teams. The same device can be used for both connected health and telemedicine, for example, a blood pressure monitor can be connected to a doctor to monitor a high risk of high blood pressure (following certain medical treatments) or be connected to a simple algorithm if the risk of blood pressure is limited. A connected scale can be used to track a baby's weight after a difficult birth or to track the weight of a person on a diet. We can guess that the commercial stakes are not far from connected health. Neither are the economic and ethical stakes: should we stop reimbursing a patient whose remote monitoring has shown that he or she is not respecting his or her medical treatment and is not taking his or her medication? We know that the rate of non-adherence reaches 50% in certain pathologies. The fear of penalties could also slow down the acceptance of connected health, as well as the fear of not respecting medical confidentiality.

Other questions also arise about connected health: its usefulness, its effectiveness and of course its reliability. The latter is crucial when a connected object is an insulin pump that will automatically inject a diabetic patient (artificial pancreas) when required.

1.4.1.2 Geolocation and Health

In Kenya, the geolocation of cell phone users (15 million users and 12,000 antennas) can analyze the movements of the population and discover that some movements had little impact on the spread of malaria (near Nairobi) or, on the contrary, had a lot (near Lake Victoria).

With geolocation data from more than 150,000 cell phones, in 2014, the Swedish NGO Flowminder analyzed the movements of populations in West Africa and tried to predict the location of Ebola fever outbreaks, which allowed for better location of treatment centers.[21] The difficulty with these analyses is that cell phone equipment is uneven across countries, and also that people's behavior changes during the epidemic, which complicates their modeling.

After the Haiti earthquake in 2010, geolocation helped facilitate aid distribution by analyzing crowd movements based on cell phone data.[22] During the subsequent cholera outbreak, phone data were used to identify areas at risk.

During the 2015 earthquakes in Nepal, Flowminder again distinguished itself by using cell phone data to locate population movements, as a large proportion of Nepalese were equipped with them.

In March 2020, data from European mobile operators were used to track the spread of the Covid-19 epidemic and conduct spatial epidemiology analyses.

1.4.1.3 The Google Flu Trends

By analyzing the keywords entered on its search engine, Google was able to establish a correlation between certain requests and the appearance of an epidemic of influenza (Figure 1.7). This correlation was corroborated by health monitoring organizations and was published in the journal *Nature*.[23]

This example illustrates the V of big data velocity (Section 1.1), with daily data updates and not weekly updates as in traditional health monitoring: yet, the sooner an epidemic is detected, the better it is fought.

Google began this project in 2008, following other work on searches or clicks on certain sites. Google analyzed billions of queries on its search engine in the United States between September 2003 and March 2007, and extracted the 50 million most frequent queries (keyword or combination of keywords, for example "indications of flu" or "cold/flu remedy"). Note that keywords with no apparent connection to influenza were not excluded *a priori*.

These queries were aggregated on a weekly basis, and normalized across regions of the United States by dividing the *RC* number of queries containing a certain keyword combination by the total *RT* number of queries in the region.

Specifically, nine regions were considered: those of the Centers for Disease Control and Prevention (CDC), which provide epidemiological data, including the weekly percentage of physician visits with influenza-like illness (ILI): temperature greater than 100°F (37.8°C) with a cough or sore throat.

For each region of the United States, and for each of the 128 weeks considered, the percentage *P* of influenza-like visits, the RC/RT ratio, and a linear regression of their logit was sought:

$$logit(P) = \beta_0 + \beta_1 . logit\left(^{RC}/_{RT}\right) + \varepsilon.$$

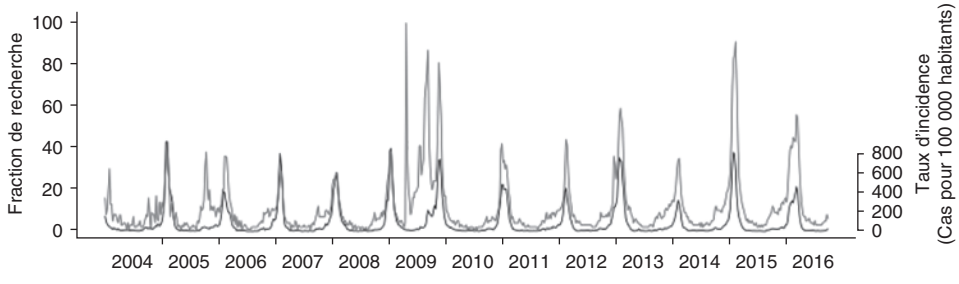

Figure 1.7 Google Flu Trends. Source: Sentinel Network; http://www.sentiweb.fr/france/fr/?page=google.

For each of the 50 million queries:

- nine linear regressions were each fitted on 128 points (i.e. 450 million regressions, distributed over several hundred machines);
- the correlation coefficients of the predicted and observed values of P were calculated for each of the nine regions;
- the average correlation of the nine linear regressions was calculated.

The queries with the highest average correlation were kept as they best corresponded to the evolution of influenza symptoms (simultaneously in all nine regions, to reinforce the significance of the result), and then they were combined together in new linear models. The same combinations were then tested in all nine regions. The combination of the 45 best queries gave the best overall result. In particular, when a combination of 81 queries was reached (the last being "Oscar nominations"), the correlation of the overall model dropped (Figure 1.8).

In these top 45 queries, 91% are directly related to the flu. In the next 55 queries, only 55% are directly related to the flu. For example, there is "high school basketball" which is in season in winter.

A linear model was finally fitted on these 45 best queries, over all nine regions, i.e. $1152 = 128 \times 9$ observations. It was validated on 42 weeks more recent than the training data sample (March 2007 to May 2008) and on all nine regions, showing a correlation coefficient = 0.97 (average between regions). This model can be applied continuously (including outside the periods monitored by the CDC) and provides predictions for the next day, whereas the CDC data are available in one to two weeks. This time saving is precious to take preventive measures: alerting populations, vaccination, detection of a new strain. But traditional networks are complementary, providing more accurate data and also data such as age and address that are absent from web queries.

As can be seen on its website,[24] Google Flu Trends has been distributed in 29 countries. It was also followed in 2011 by a Google Dengue Trends.

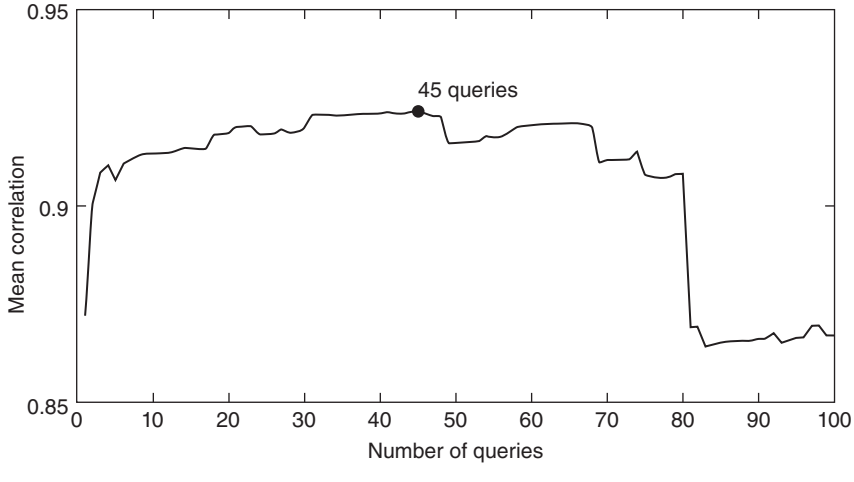

Figure 1.8 Number of queries in Google Flu Trends. Source: *Nature*, 457 (2009), 1012–1014.

Thus, the Google Flu Trends (GFT) model forecasts were initially 97% correlated with CDC forecasts. The quality of these forecasts declined from 2013 onwards, with an overestimation (sometimes double) of influenza cases.[25] Several causes have been identified:

- internal to Google: the Google engine has evolved since 2008, in particular with the auto-completion of additional search terms, i.e. semi-automatic entry based on the previous searches of Internet users, which artificially increases the prevalence of certain terms (searches no longer depend solely on the state of health of the Internet user, but also on the operation of the search engine);
- external to Google: the number of searches can also increase if the news talks more about the disease (as in the case of H1N1) or a drug related to the disease, or even when a media outlet, such as *The New York Times* in 2008, publishes an article on the GFT giving an example of a query analyzed by Google, a query immediately entered by many Internet users (this is the reason given by Google for not publishing the list of GFT queries).

Other causes are not new: Google users are not a representative sample of the whole population, symptoms may be poorly described, and moreover some words with the same search pattern as the flu have no relation to the disease.

Google revised its model in 2009 (after the H1N1 flu epidemic), in 2013 (after the H3N2 flu epidemic) and in 2014, trying to take into account the impact of media coverage of the flu on Google searches. It also refined its regression model by substituting an elastic net regression[26] (Section 3.3). But Google eventually closed its Google Flu Trends website to the public in August 2015, though the data remained available to researchers.

Despite the use of elastic net, there was a probable overfitting of the GFT model, especially on winter epidemics, resulting in poor generalization to other periods, as the GFT model detects cold and winter and not only influenza.

Google's work has been criticized for its forecasting errors and lack of information provided to the scientific community,[27] but the GFT can be useful in countries without a health surveillance network as efficient as the Centers for Disease Control and Prevention in the United States, and in conjunction with health surveillance data.

It was shown[28] that the Google Flu Trends model improved the use of only historical CDC data in an ARIMA model, decreasing the mean absolute error (MAE) by 16% with a 32-week learning window. Other work[29] showed that GFT data could be used in an auto-regressive model incorporating a hidden Markov chain, which is more robust, more accurate, and incorporates seasonality: the ARGO (AutoRegression with GOogle search data) model.

The use of health surveillance data is important to ensure the reliability of local-level predictions, which are both the least accurate in the GFT model and the most important for prevention, vaccine delivery, and distribution actions.

Other projects include Flu Near You (US) and GrippeNet.fr (France). Some projects are based on analysis of Twitter (MappyHealth, Crowdbreaks, Sickweather), Yahoo! and Wikipedia, but the signal-to-noise ratio is low, and the young age of users makes the sample unrepresentative.

Another example is provided by the Covid-19 outbreak in 2020. Recall that it appeared at the end of November 2019 in the Chinese region of Wuhan and that the Chinese authorities recognized the existence of this new pneumopathy at the end of December 2019. The World Health Organization (WHO) announced on January 9, 2020 that this epidemic was

due to a new RNA virus, SARS-CoV-2, of the coronavirus family like SARS-CoV (responsible for the SARS epidemic in 2003) and MERS-CoV (responsible for the MERS epidemic in 2012). This new virus is the causative agent of this new respiratory infectious disease called Covid-19 (for **C**orona**VI**rus **D**isease). The human-to-human transmissibility and the number of actual cases in Wuhan were guessed in mid-January 2020, notably from studies by Imperial College London based on cases reported far from Wuhan, in Thailand and Japan. The late announcement of the WHO and its initial underestimation of the pandemic were surprising, and have since been attributed to the influence of China and its desire to exonerate its responsibility in the dissemination of the virus.

The Canadian company BlueDot had detected the epidemic and its spread in Asia a few days before the WHO alert, based on an algorithm that takes into account not only data from Google searches for symptoms, but also from online forums, news feeds in 65 languages, health bulletins and official statements, as well as air travel. The algorithm's analyses are verified by epidemiologists who validate the information before sending it to health authorities, airlines, and public hospitals in several countries.

1.4.1.4 Research in Health and Medicine

The collection of patient data makes it possible to detect epidemics or to evaluate the side effects of a treatment. It is possible to match medical data with environmental data to study, for example, the effect of exposure to pesticides or endocrine disruptors. Genomic data can also be matched with clinical, medical imaging or socio-economic data to explain the onset of a disease or the response to a treatment.

In addition to these public projects, there are private projects such as Google's *Baseline* project, which wants to map human health by collecting medical data from volunteers equipped with connected sensors.

In population genomics, we seek to identify the regions of the genome affected by selection processes, from domestication to the recent improvement of breeds used in production, and understanding artificial selection can help to improve natural selection. Genome-wide association studies (GWAS) are the analysis of genetic variants in individuals, and the study of their correlations with phenotypic traits such as diseases (rheumatoid arthritis, Crohn's disease, breast cancer, Alzheimer's disease). These studies require massive data, which can involve hundreds of thousands of individuals and millions of variables: the nucleotide polymorphisms (SNPs: single-nucleotide polymorphisms) which are the genetic variants studied. These SNPs represent about 90% of human genetic variants and correspond to the substitution of one nucleotide for another at a specific location in the genome. For example, at a specific position in the genome, the C (cytosine) nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by a T (thymine) nucleotide. This means that there is a SNP at this position in the genome. On average, a SNP is encountered every 100 to 1000 nucleotides and there are about 5 million SNPs in a human genome (i.e., compared to a reference genome), distributed irregularly across the chromosomes. Fortunately, not all of these genetic variants are associated with pathologies, but some are associated with pathologies and responses of the body to treatments, hence the interest of genome-wide association studies. One difficulty with these studies is that genome-wide databases are scarce and lack representativeness for certain populations.

Neuroimaging benefits greatly from deep learning techniques. Brain imaging, for example, by MRI, allows us to map the areas activated during the performance of certain tasks and thus associated with certain mental processes. It is also possible to link the observed shape of the brain to certain characteristics, such as age, and to learn to distinguish by imaging the normal aging of a brain from a neurodegenerative disease. For a long time, brain imaging was limited to small datasets because of acquisition and processing costs, and the conclusions were unreliable and sometimes contradictory, due to the importance of experimental variations. It can now handle larger datasets, thanks to image compression methods and the use of distributed computing, and can rely on computer vision algorithms using deep neural networks. These algorithms can detect hidden patterns indicative of a pathology such as Alzheimer's disease, allowing an earlier diagnosis than the appearance of clinical signs and thus likely to increase the chances of successful therapies.

Deep learning, with methods based, for example, on generative models (Section 7.19), is beginning to be used to design new molecules for therapeutic purposes. In December 2018, DeepMind announced some very interesting results on protein folding, a challenging and important question in understanding the mechanisms of life and certain diseases. Deep neural network models were trained to predict the 3D shape of proteins based on their constituent amino acids. Proteins are chains of amino acids and folding can occur between each amino acid in the chain, leading to an extremely large number (about 20^{300} for 300 amino acids in a protein) of different combinations and structures. With the same amino acids, two proteins can have different shapes and therefore different functions. So the three-dimensional shape of a protein is very important, and a wrong folding of a protein can prevent it from fulfilling its function properly and be the cause of diseases such as diabetes or Alzheimer's. In order to understand this protein folding mechanism, and the diseases caused by its defects, that researchers are organizing competitions to predict the shape of proteins. The challenge is to predict the structures of recently discovered, but not yet made public, proteins based on their amino acid sequence. The team that comes closest to reality wins. DeepMind's AlphaFold algorithm won a competition (CASP13) by finding the most accurate prediction for 24 of the 43 proposed structures,[30] while the runner-up managed to do so for only 14 structures. It had been trained on thousands of known proteins, taking into account the distance between the amino acids and the angles formed by the folds, to find the shape that requires the least energy, which is the one taken by the protein. It was also able to predict in March 2020 the 3D structure of several proteins associated with the SARS-CoV-2 coronavirus, and an improved version trained on 170,000 proteins, AlphaFold 2, won the CASP14 competition in November 2020. The results of this competition are measured by a global distance test (GDT) between the predicted structure and the actual shape of a protein observed in laboratory. This distance is normalized on a scale of 0 to 100 and AlphaFold 2 obtained a GDT score above 90 for two-thirds of the proteins. Its GDT score for the most difficult protein to predict was even 25 points higher than the next team, while in 2018 the lead was only 6 points. Now, a distance as small as that reflected by a GDT score above 90 falls within the measurement uncertainty interval and means that AlphaFold 2 predictions are as good as the imaging data. AlphaFold 2 has been described in a paper published in *Nature*,[31] along with an open source software and a database of the structures of hundreds of thousands of proteins from several species.

It is envisaged to use this type of algorithm to create proteins fulfilling a particular function, useful for certain health or environmental problems, and start-ups are being created in this field. A new Alphabet company called Isomorphic Labs has thus been launched in November 2021 to apply these methods to the discovery process of new drugs.[32]

BERT models have also been applied to this question of the 3D structure of proteins,[33] their attention mechanisms connecting amino acids that are distant in sequence but close in 3D structure.

We will see in Section 7.20.3 that very promising projects of brain-machine interfaces allow a paralyzed person to pilot an exoskeleton by thought, as a kind of artificial spinal cord, or a mute person to express himself or herself using a vocal device decoding their brain signals.

Of course, its successes cannot hide for the moment the limits of artificial intelligence in the medical field, as we will see in Section 10.2.

1.4.2 Big Data and Deep Learning for Drivers

Big data and deep learning are present in automotive driving assistance, using:

- intelligent guidance system (integrating vehicle location, traffic status and history, presence of hazards, weather conditions and driver's time constraints);
- economic driving assistance (by integrating the consumption of the vehicle in its various phases of operation: stop, acceleration, cruising, braking);
- driving assistance systems based on information communicated by other vehicles, such as the presence of a broken-down vehicle, a hill or a descent allowing to anticipate an increase or a decrease of the engine speed (this is under study in some road transport companies);
- detection of driver fatigue by analyzing the driver's behavior.

Big data can be used for tracking systems in case of theft, similar to cell phones.

Big data also enables personalized maintenance based on information collected by multiple sensors: brake wear, engine speed, fluid pressure, etc. Analyzing these data and comparing them with data collected on other vehicles helps determine the optimal time to perform a service. These data are sent to a cloud to feed statistical models and send alerts to the driver if an anomaly is detected.

Agreements were announced in 2018 between automakers and GAFAMs: a Google/Renault-Nissan-Mitsubishi agreement for the use of Google Assistant artificial intelligence, and a Microsoft/Volkswagen agreement for the Volkswagen Automotive Cloud to use the Microsoft Azure platform.

These applications shift from individual to collective interests. The collection of information provided in real time by drivers' smartphones also provides information to the community about road traffic conditions, smoothing the traffic and even avoiding accidents. This information can also send alerts on road conditions, saving detection costs. Weather sensors on in-vehicle smartphones can provide more detailed information than the national weather service.

The biggest innovation expected in the next decades in the transport sector is the advent of the autonomous vehicles: train, bus, truck, car and maybe even planes. The self-driving

car should free up time, reduce traffic jams, reduce the number of parking spaces needed, facilitate vehicle sharing, and of course reduce the number of accidents (the majority of which are due to human error).[34] More generally, the financial and human cost of transport will decrease. However, some people will suffer: taxi drivers and truck drivers, perhaps garage owners.

The self-driving vehicle uses sensors and cameras to reconstruct the road situation in 3D and the position of other vehicles, and artificial intelligence algorithms to analyze these data in real time and act accordingly on the vehicle's controls to accelerate, slow down, brake, turn, depending on the road configuration, traffic conditions, external disturbances and of course obstacles. It must ensure both the safety and comfort of passengers.

Several levels of autonomy have been defined and must be reached over time by the vehicles.[35]

- At level 0, the vehicle is not automated at all, and the driver has total control of the main functions of the vehicle (engine, gas pedal, steering, brakes) at all times.
- In Level 1, the vehicle has some driver assistance features, such as Anti-lock Braking System (ABS) or Electronic Stability Program (ESP) that assist the driver but still leave the vehicle under the driver's control.
- At Level 2, partial automation of the vehicle's main functions allows the driver to be replaced in some situations, such as parking the vehicle, but leaves the driver fully responsible for driving in all other situations.
- Level 3 is limited autonomous driving, where in certain situations, for example, on the highway, the driver can hand over full control of the vehicle to the automated system, which will then be responsible for the main functions and will be able to accelerate, brake, and change lanes. The driver must nevertheless remain attentive and be ready to recover control of the vehicle at any time.
- Level 4 is autonomous driving throughout a journey, where the driver does not need to watch the road or hold the steering wheel, but only on appropriate routes (good enough road) and provided that the weather conditions are suitable (no fog or snow, for example).
- At Level 5, driving is fully autonomous in all circumstances and the driver can even leave the driver's seat or even be absent.

Developing these autonomous vehicles employs an enormous number of possible driving scenes, including vehicle malfunctions, and requires massive simulations of millions of kilometers, on any type of road, in any type of weather conditions and in any type of circumstance, including the failure of a computer or sensor. A particular difficulty for artificial intelligence is predicting human behavior: while a driver will guess that a pedestrian running after a bus is likely to cross the road suddenly, this will be much more difficult for an artificial intelligence.

1.4.3 Big Data and Deep Learning for Citizens

Local authorities can analyze social networks, discussion forums, blogs, and the online press to help them build dashboards and opinion barometers to help them define public policies. The analysis of these documents allows a city to detect the expectations and concerns of its inhabitants, in order to better target its investments, its renovation works, its new services, or its communication.

We must not hide the risk of drifting toward surveillance of citizens' opinions and personal networks, or of manipulation as shown by the cases that have involved Facebook since 2016, particularly around the American presidential elections.

Particularly worrying is the example of China, where a mass surveillance system is being set up using video surveillance cameras and facial recognition algorithms to track everyone in the country. The regime's goal is to cover all of China's public places with 450 million cameras by 2020, allowing it to find any person in a city within minutes (as tested by a BBC journalist who was found in seven minutes in the city of Guiyang from a photograph). Even more worrying, it is planned that this system will be coupled with a "social credit system" evaluating the behavior of each citizen in order to grant or withdraw rights. This algorithm is kept secret, but it is known that it is based on financial data, purchasing habits, and social contacts, among other things. This evaluation could lead to denying a person access to certain means of transport, hotels, purchase of an apartment, etc. Blacklists are disseminated on the Web and people calling them are informed by a message that their correspondent is on the blacklist. In 2019, China tested the extension of this social credit system to foreign companies, with a system that constantly monitors compliance with three hundred standards and downgrades the rating of a company that does not comply with them, which could make this system a political weapon.

The abuses of facial recognition are not limited to states, with applications allowing anyone to find on the Web all the photographs of themselves or of any person from whom they have obtained a photograph, even without their knowledge, even providing for alerts to be sent as soon as a new photograph is detected.

Social networks can also be scrutinized to anticipate election results, but polls remained more effective in predicting the results of the first round of the French presidential election in 2017. We know the importance of big data in the 2012 American presidential elections. Both in the campaign of the candidates to target potential voters by the right means (email, phone call, home visit) and focus on voters supposedly more undecided, adapt the communication of a candidate in real time based on messages on Twitter. And both in the prediction of results by statistician Nate Silver in his blog FiveThirtyEight.com, correct prediction in each of the 50 states (his predictions were correct in 49 of the 50 states in 2008, and in 31 of the 33 states where Senate elections were held in 2012). His model is known to rely heavily on polling but has not been disclosed. Social networks can also be used to estimate election results before the close of polling stations.

In everyday life, deep learning algorithms are used in many applications: voice recognition and synthesis to communicate with a chatbot, automatic reading of checks, early detection of fires from a network of ground cameras, infrastructure monitoring, etc.

1.4.4 Big Data and Deep Learning in the Police

The use of big data and machine learning in predictive policing has developed a lot in the United States, where there is a logic of "proactive policing" whose objective is to prevent crimes and not only to react and arrest the perpetrators. This policy is not quite the same in Europe, and in particular in France, where the arrest of a person can only be carried out in the event of the "beginning of execution" and where the simple control of a person can only be done on the basis of objective elements. The prediction provided by a software is

not always considered to be such an element (whereas the nose of a dog will be for drug detection), but France and Europe are nevertheless beginning to take an interest in such software.

In France, the PredVol software was tested in 2016–2017 to predict vehicle crime. It provided a cartographic representation of the tested department and pointed out the risk areas according to a weekly forecast, from a five-year history. It specified the offence: car theft, motorcycle or bicycle theft or theft from a vehicle. Its evaluation showed that it tended to predict that risky areas would remain risky. As it did not provide any really new information, it was abandoned. The national police force then acquired another software system, Paved, which predicts a higher risk of vehicle theft or burglary in a given area, perhaps a little more accurately than PredVol. It is not based on personal data but on the history of delinquency in that area.

PredPol (now Geolitica)[36] is software developed by researchers at the University of California Los Angeles (UCLA) by analyzing 13 million crimes, and operated since 2011 by the Los Angeles Police Department and other US police departments. This provides them with a daily updated mapping of 200 x 200 meter areas where there is a high probability of a crime or offense occurring, which allows them to define patrol zones. The details of PredPol's algorithm are not published, but it is known that it is based on a seismology model by French researcher David Marsan that predicts aftershocks from an earthquake. PredPol explains future events by past events: people break in more often where they have already broken in. Like the previous software mentioned, PredPol makes predictions about areas and not about people. This police prediction system has been quite successful and has spread to other cities. Compared to the classic pins planted on a map, it has an undeniable practicality with its predictions sent to police officers' smartphones.

It has since been criticized[37] in two opposite directions: the duster effect and the streetlight effect. The duster effect means that the predictions of the algorithm may be accurate but that they do not reduce crime, which would only shift to a new location. The streetlight effect implies that the algorithm's predictions may not even be accurate, but they give the appearance of being accurate because police forces concentrate according to the predictions and thus record crime more often in predicted areas than in others. The crimes might not be more where the algorithm says they are, but they would simply be recorded better. In any case, these are the places where crime is usually concentrated, which could be predicted to be at risk without the help of software.

It should be considered that the purpose of this software is actually less to predict crime itself than to optimize police patrols according to risk, like its competitor Hunchlab from the American company Azavea. Nevertheless, it continues to be regularly accused of perpetuating economic and racial bias.[38]

We can also mention the Key Crime system in Italy and Precobs (Pre Crime Observation System, but this is also a nod to the *precogs* of *Minority Report*) in Germany.

Unlike PredPol and its emulators, the American company Palantir's model (named after the "vision stones" in *The Lord of the Rings*) relies on the personal data of previously convicted individuals, including of course their backgrounds, and establishes a network between them like those studied in Chapter 5, making it possible to highlight central individuals and increase their risk score. This solution interested the French Direction Générale de la Sécurité Intérieure (DGSI), before being extended to other activities and

other French and European clients, despite the reluctance fueled by Palantir's relationship with American intelligence and fears of data leakage to an American agency.

1.5 Risks in Data Processing

The challenges of big data and deep learning lie not only in the computing power to store and process data, but also in its security and quality. There are several pitfalls to avoid:

- too little training data;
- poor data quality;
- missing values;
- non-representative samples;
- spurious correlations;
- overfitting;
- lack of explainability of models.

1.5.1 Insufficient Quantity of Training Data

An essential difficulty of Big Data comes from what is called the "curse of dimensionality". This curse comes from the fact that the individuals studied are points in a space whose number of dimensions is equal to the number of characteristics observed: each of these characteristics is a coordinate of the individual. With big data, the individuals are therefore represented in high-dimensional spaces, and of course the individuals must be sufficiently close, the space must be sufficiently dense, so that the learning of rules and models can be carried out on a representative basis. However, it is very difficult to fill a high dimensional space. For example, in the subspace $[0,1]^d$, 10^d points are needed to be separated by a distance equal to 0.1. Thus, in dimension 1, ten points in the interval [0,1] make it possible to cover a phenomenon with a step equal to 0.1, but it takes one hundred points to obtain the same coverage in $[0,1]^2$ and it takes ten billion in $[0,1]^{10}$, which is much greater than the number of individuals commonly present in a dataset. This explains why the predictive power of a model increases with the number of explanatory variables used up to a certain point before decreasing (the Hughes effect).

Now, if we consider images, we realize that a training sample can only contain a tiny proportion of all possible pixel configurations: even with extremely simple images of 28 × 28 pixels, each taking only 2 values (white/black), the number of configurations, the dimension of the solution space, is $2^{28\times28}$, i.e. 10^{236}, a huge number!

1.5.2 Poor Data Quality

Data quality is always critical, and even more so with big data because:

- many data come from external and heterogeneous sources;
- their rapid evolution makes it more difficult to control them;
- they may be used in automatic algorithms;
- some data, typically those collected on the Web and social networks, contain a large amount of noise, missing values or even errors.

Let's recall the main quality requirements when processing data:

- the definition of the data must be known and adequate;
- the data must not be erroneous;
- non-missing values must be present in the attributes that require them;
- the matching of data between different data sources must not reveal any inconsistencies;
- the data values must be up to date;
- data should not be unduly duplicated;
- the history, processing and location of the data in question must be easily traceable.

It is also important to avoid errors that could occur when reconciling data sources between which no universal keys exist.

1.5.3 Non-Representative Samples

A basic rule of modeling is to train the model on a representative sample of the population to which the model will be applied. A lack of representativeness is the most common cause of systematic errors in model prediction, commonly referred to as "algorithmic bias."

Representative sampling is not necessarily simple random sampling, and it is not uncommon to stratify a sample to ensure that each type of individual, even the rarest ones, are adequately represented. This is a common practice in public opinion polling, where one wishes to control for gender and age representation in order to improve the accuracy of the results for the entire population. But there are rules for determining the optimal size of strata, and in all cases, one should not exclude from the sample a stratum on which one will then want to make predictions or obtain results.

It is also known that a linear regression model is only valid for the range of observed values. We know that medical results cannot be extrapolated from one type of patient to another. We know that an image recognition algorithm will sometimes fail miserably to recognize an image absent from its training sample, as recalled below (Section 1.5.6).

1.5.4 Missing Values in the Data

A particular cause of unrepresentativeness of a sample is the presence of missing values for some variables, or even missing individuals altogether, which is the case of non-respondents in surveys or credit applicants who have been denied credit. In this case, since these individuals did not obtain credit, we cannot assume that they would have repaid it or not, and therefore they do not have a good or bad payer label and are necessarily absent from the training dataset of a scoring model. Non-response is then total, whereas it is only partial when some of the responses were provided or some of the variables were observed. If the values or responses are not missing at random, as is often the case, they render the sample unrepresentative and lead to the existence of a bias in the results.

This is what we try to remedy with imputation methods, i.e. the replacement of missing values. There are deterministic imputation methods, in which the imputed value is always the same for each individual, and random methods, in which the imputed value varies with each application of the method. Among the deterministic methods, we find the imputation of Y by its mean or median over the respondents, or imputing Y by a regression $Y = f(X_i)$ on other variables X_i without missing values.

These methods are useful for obtaining values for each individual, but sometimes one is not specifically interested in individual predictions but rather in results over the whole population, whether it is to calculate correlation coefficients, regression coefficients or other statistical parameters. The objective of imputation is not to obtain the best possible predictions of the missing values, but to replace them with plausible values in order to make the best use of the available information to make inferences about the parameters of the population. However, deterministic methods obviously do not respect the distributions of the imputed variables or their relationships with the other variables. In particular, they underestimate the variance of the imputed variables. This is why we often turn to random imputation methods.

The first method consists of adding a random residue to the previous regression, which is then written $Y = f(X_i) + \varepsilon$. This residual can be drawn from a parametric distribution fixed *a priori*, or drawn randomly from the prediction errors of the deterministic imputation method observed on the respondents. Another method is the "hot deck" imputation[39] in which the missing value of an observation is replaced by the non-missing value of another observation drawn at random from the same sample. A third method is Donald B. Rubin's multiple imputation.[40] In a single imputation, each missing value is replaced by an assumed value. In multiple imputation, each missing value is imputed by several (often of the order of five) values. This results in several completed datasets without missing values, which allows the desired statistical analyses to be performed on each of these datasets, and then the results obtained are combined into a set of estimated parameters with their variance. This allows us to add the variance between the datasets to the variance within each dataset.

The advantage of these random imputation methods is that they preserve the distributions of the imputed variables, but this comes at the cost of additional variance, due to the random selection of residuals or individuals. One can try to limit this variance by a good implementation of these methods, but other approaches have been proposed to limit the imputation variance, such as the balanced imputation of Chauvet *et al.*[41] This has several advantages: it is as simple as simple imputation, it can be applied to qualitative or quantitative variables, and it preserves the distribution of the imputed variables

In computer vision, each image is defined by a set of pixels and missing parts of the image correspond to pixels with missing values. We can use the previous imputation methods but a more natural solution is to represent an image as a graph, where each pixel is a vertex, and two neighboring pixels are connected by an edge. When pixels are missing in an image, the corresponding graph has fewer vertices but this does not change the way it is processed. In particular, we can apply to it the methods of Spatial Graph Convolutional Networks (SGCN) and perform on these graphs the classical convolution operations on images. Compared to simple graph networks that only take into account the neighborhood of the pixels, spatial graph networks take into account their spatial coordinates, which is more precise information. This approach has achieved good results.[42]

1.5.5 Spurious Correlations

It is known that a strong correlation between margarine consumption and the divorce rate in Maine, or between crude oil imports from Norway and driver deaths in collision with railway train, does not reflect any causality. These are two of many examples presented on

a well-known website.[43] It would never occur to anyone to use such spurious correlations in a model, even if their R^2 is greater than 0.95. And yet, there is a great risk that tools designed, and even praised, to look for any correlation in big data, detect such absurd correlations and implement them in an "AutoML" approach, i.e. automatic machine learning.

This risk can only be avoided by careful analysis of the data and knowledge of the business behind it. This knowledge is necessary because not all spurious correlations, which do not correspond to any causality, present themselves in such a burlesque light. For example, in the field of risk prediction, certain apparent risk factors are not due to an inherent risk but to the company's management rules in the face of this risk, rules that are valid at a given moment but not immanently. A model built on these rules will appear reasonable and will provide correct predictions at first, but will quickly show its limits and a decrease in its predictive power.

In an attempt to discern true causality among correlations, epidemiologist Bradford Hill set out nine criteria for causality,[44] including temporality, correlation strength, plausibility, and reproducibility. Temporality (cause precedes effect) is obvious, and the intensity of the correlation can be measured by the data scientist, but plausibility requires business expertise and therefore exchanges between data scientists and business experts.

Beware, however: not all unexpected correlations are spurious.

1.5.6 Overfitting

We speak of overfitting of a model when too much complexity leads it to follow all the fluctuations of the training sample, to learn the noise in addition to the signal, to detect false links with the risk of applying them wrongly to other samples.

When we talk about a model being too complex, it is normally in comparison with the size of its training sample, and a larger sample can allow the training of a more complex model, in the larger number of cases the model allows a larger number of parameters to be fitted. However, the quantitative size of the training sample is not enough, and its representativeness is also important: its cases must represent the largest possible number of cases. This seems obvious but is not always easy to obtain. For example, in computer vision, a deep neural network will be able to learn to identify fish, but as these fish will have been represented in the vast majority of cases on a seabed, the network will have learned to recognize a fish on a seabed and may not be able to recognize it in an unusual context. Examples are quite frequent of animals or objects not recognized when they are not in their usual environment. Examples include husky dogs mistaken for wolves when photographed in the snow.[45] A famous paper shows how a photograph of a panda can be imperceptibly modified to be mistaken for a photograph of a gibbon by the GoogLeNet neural network.[46] This kind of "attack" that could mislead deep learning models is taken seriously and studied, as it could lead to accidents if it involves autonomous vehicles which are made to take one road sign for another.

1.5.7 Lack of Explainability of Models

By moving from classical regression models to deep neural network models or ensemble models aggregating large numbers of decision trees, we have moved from "white boxes"

to "black boxes". The direct apprehension that one has of the coefficients of a regression gives the impression that one understands the model well, and it is true that they allow one to justify one's predictions perfectly. But we refer to the previous discussion on correlations that are not necessarily causalities, to warn against the illusion of believing that the mathematical justification of predicted values provides a real explanation of the model. "Predicting is not explaining," to quote the title of a famous book by the mathematician René Thom.[47] And most methods of "model explainability" consist solely of determining the influence (intensity and meaning) of the variables in the predictions of the model. However, the importance of the variables measured in this way does not mean the existence of causality, and the change in one variable may imply changes in other variables: one cannot often measure the influence of a variable "all other things being equal"/ (*ceteris paribus*). We have thus obtained a form of explanation of the model, in the sense that we can justify our predictions with arguments that are understandable to a human being, but this justification may not be satisfactory if it does not indicate the underlying causes at play. It will answer the question "how was such a high score calculated for this individual?" rather than "why is this score so high for this individual?". This is why, alongside purely predictive models, causal models are being developed, such as Bayesian networks.

Recognizing their limitations, model explainability methods can nonetheless be useful, particularly in a regulatory setting, for responding to auditors and ensuring that no criterion has a disproportionate weight in a prediction or is likely to cause bias. The simplest methods estimate the global weight of a variable in a model (for example, by measuring the decrease of the predictive power of the model after randomly swapping the values of the variable), but more recent methods, such as LIME[48] and Shapley's[49] method, can be applied in machine learning and even deep learning models to estimate the contribution of each characteristic in the prediction made for a specific individual (Section 8.3.4). They are implemented in R (`shapr, fastshap, lime, iml`) and Python (`shap, lime, skater`) packages and more general tools such as *Google AI Explanations* available on Google Cloud Platform, and *AI Fairness 360* from IBM available in R and Python.

1.6 Protection of Personal Data

All that is technically possible is not legally and even less ethically possible, and the overuse of personal data would be counterproductive for any company that engaged in it. Negligence is not acceptable either: data security is crucial when large amounts of personal data are stored. This is both for financial reasons (these data are valuable) and for the protection of personal data, and therefore for reasons of trust. The data must be safe from theft, loss or damage, and protection against hacking and cyberattacks must be reinforced.

1.6.1 The Need for Data Protection

Legislation on personal data protection will probably have to evolve to take into account the existence of the immense amounts of personal data available on the Internet, of which the holder is probably not even aware. The massive use of data collected by social networks has, however, been the subject of awareness on the occasion of several recent cases,

including the scandal revealed in March 2018 of the processing of personal information of more than 85 million people, collected through Facebook by Cambridge Analytica. This company was able to collect a lot of data from Facebook users thanks to a personality questionnaire available on Facebook, which was to be used by Cambridge researchers to study the psychological profile of a person thanks to their activity on Facebook. One of the researchers sold these data, which belonged not only to the people who had installed the application but also to their friends on Facebook, to Cambridge Analytica, which allowed political figures to use it to influence the vote for Donald Trump during the 2016 US presidential election, by sending elements that could orient their vote to the voters' Facebook feed. Since then, it has also been discovered that Facebook allowed advertisers to target their marketing messages by using the phone number provided by Facebook users to secure their account, and also by using contacts from the user's phone address book (which the social network has access to after the user has authorized it to scan their contacts for "friends").

In addition, numerous cases of data theft have highlighted security breaches:

- the theft of 100 million LinkedIn user email addresses and passwords revealed in May 2016;
- the theft of 68 million Dropbox user email addresses and passwords revealed in August 2016;
- the hacking of at least 500 million Yahoo! accounts in 2014, revealed only in September 2016;
- the theft of at least 143 million consumers' personal data from Equifax databases revealed in September 2017;
- the theft of 57 million Uber customers' personal data in October 2016, revealed only in November 2017;
- the hacking of at least 50 million Facebook accounts in September 2018;
- the theft of personal data (login, password, and in some cases name, phone number, place of residence, birthday) of 533 million Facebook accounts of users from 106 countries, theft carried out in 2019 but revealed only in April 2021 by its posting on a hackers' forum;
- the theft of card and bank account data, and other personal information of 100 million U.S. customers of Capital One Bank in March 2019, during the migration of its IT system to the cloud, by a former AWS employee.

The users concerned were not always notified by the company, or beyond the legal deadlines.

Particular attention should be paid to objects connected to the Internet, video surveillance cameras, routers, thermostats ... whose knowledge of the IP address can allow the takeover by an unwanted third party and be used in denial of service attacks, which consist in saturating the target's servers by concentrating toward it a huge flow of requests from the hacked objects (which then constitute a *botnet*). An even more direct and serious danger could come from taking control of a connected autonomous vehicle (Section 1.4.2), as it was already partially the case in 2016 when Chinese researchers from the Keen Security Lab took remote control of some functions of a Tesla Model S car, including the dashboard lock and the braking! Terrorists could even take control of an entire fleet of vehicles and turn them into deadly weapons. We are beginning to advocate the presence of a button on these vehicles that would allow them to be disconnected from the Internet.

The strengthening of legislation on the protection of personal data (Section 1.6.3) and its adaptation to the Internet era will have to accompany and provide a framework for the development of a digital economy based on the exploitation of enormous quantities of data provided by Internet users, social networks, and even all citizens, most often without their knowledge, because of their activity, their expression, their purchases, their movements, but also quite simply because of their lives in a world of increasingly connected sensors and devices. The acceptance of the digital economy model will require better control of its protagonists.

Thus, geolocation by smartphones can track individuals, and know their habits, the places they frequent to the nearest address. This geolocation is based on the recovery of the addresses of the relay antennas to which the smartphones are connected. The development of connected sensors also generates large amounts of individual data, especially in the field of health, where data are sensitive. The use of the cloud raises the issue of data transfer to countries that do not have legislation protecting personal data, and under conditions that do not always ensure total data security.

Cryptography is essential for encrypting data to ensure that it cannot be misused if it falls into the possession of the wrong people. Some data are only stored and exchanged in encrypted form. For example, payment data transmitted over the Internet is encrypted. It is even possible to perform certain operations on encrypted data without having to decrypt it. Unfortunately, cryptographic methods can also be used by hackers who manage to encrypt all the data on a computer, or even an entire company network, and demand a ransom in exchange for the decryption key: this is known as ransomware.

1.6.2 Data Anonymization

Not to mention criminal cases, anonymization may be desired in order to study sensitive data, such as medical data, without making it possible to identify the individuals concerned. In other fields, such as commercial, banking, there is a "right to be forgotten" provided for by the legislation so that the data of a former customer is no longer accessible by name a certain time after the customer's departure. Nevertheless, the company possessing these data may wish to be able to use it for statistical studies that do not contain names, or to meet regulatory obligations for risk analysis in the case of a bank.

Anonymization methods do exist, but they do not always offer total protection. On the one hand, each Internet user, even if not identified, can be associated with a very precise profile of behavior or even personality. On the other hand, it has been demonstrated[50] that the combination of certain information, known as quasi-identifiers, often corresponds to a unique person, who can then be identified if we have a database in which this person appears with an identifier and their quasi-identifiers. It is enough to match the two databases containing these quasi-identifiers, one containing the identifier, the other not, to be able to identify the individual in the second database. The quasi-identifiers can be the sex, the date of birth and the postal code together.

A famous example of de-anonymization of a database is that of Netflix, which in October 2006 launched its prize for the best recommendation model (see Section 3.6) by providing a database of 100 million ratings of movies seen by 480,000 users whose name had been replaced by an anonymous identifier. A few weeks later, two researchers from the

University of Texas, Narayanan and Shmatikov, announced that a de-anonymization algorithm allowed them to find a user in the Netflix database:[51]

- in 84% of the cases by knowing 6 ratings of the user,
- in 99% of the cases by knowing 6 ratings with their date within 14 days,
- in 68% of the cases by knowing 2 ratings with their date to within 3 days.

They applied their result to the public IMDb (Internet Movie Database) and the ratings it contains, assuming that the ratings of the same user of Netflix and IMDb are highly correlated. They showed that their de-anonymization is effective even if the intersection of the two databases is small.

The problem with this breach in anonymity is that a Netflix user does not want their ratings to become public, revealing their personality traits.

The same misfortune happened the same year to AOL, which had given access to 20 million web searches of 650,000 users, some of whose identities were found by two journalists of *The New York Times*.

The basic anonymization known as pseudonymization is therefore often not sufficient. This method consists of replacing an individual's identifier with a pseudonym or a random key (which must, however, allow for file matching).

More sophisticated methods of anonymization do exist,[52] such as data noise (e.g. by adding Gaussian noise to continuous variables) and especially *k*-anonymization[53] which consists of a kind of "blurring" of the data. In this method, we start by determining the quasi-identifiers, whose combination of values can be unique, and by making the values of the quasi-identifiers of each individual equal to those of at least (*k-1*) other individuals: we replace these values by intervals of values by grouping adjacent quasi-identifiers. For example, we group the individuals by age groups and we replace the age of each individual by its age group. The value of *k* must be well chosen: small enough not to degrade the precision of the analyses that will be performed on the data, but large enough not to reveal too precisely the values of the sensitive data.

The *k*-anonymization may present a technical difficulty for the replacement of some data, for which a good "blurring", whether it is about the constitution of bins for a quantitative variable or the grouping of categories for a categorical variable, may require a non-trivial and somewhat long human or machine processing.

But the main limitation of this method arises when a sensitive data item varies little in each grouping of quasi-identifiers. In this case, it would be enough to know the value of this sensitive data item for an individual that we have managed to identify, to have a fairly accurate idea of the value of this data item for the (*k-1*) other individuals of their group. If we want to avoid this, we can increase the level of aggregation of the individuals and constitute larger groups.

Another possibility is to group individuals according to another quasi-identifier, for example, the postal code, which will better mix the values of the sensitive data to be protected. However, this mixing can be detrimental to the analyses: it will not be possible to verify the existence of a link between age and a phenomenon if we only know the occurrence of the phenomenon for heterogeneous age groupings.

In both cases, the objective is that of the *l*-diversity method: to add a constraint so that not only each group of quasi-identifiers contains at least *k* individuals, but that the sensitive variable takes at least *l* distinct values.

A third method is *t*-proximity since *l*-diversity does not remove all useful information, it is further disguised by constructing the groups of quasi-identifiers so that the distribution of values of the sensitive variable is approximately the same in each group of quasi-identifier. The disadvantage of *t*-proximity is that it limits the possible analyses since it reduces the possibility of discovering useful correlations between the quasi-identifiers and the sensitive variable.

A fourth method is differential privacy.[54] It is appreciated by specialists because it provides theoretical bounds on the risk of identifying an individual. However, it is difficult to implement, and it creates fictitious data that are not always plausible (but which can be for geolocation data).

A fifth method consists of generating a dataset that recreates all or part of the variables in the dataset but preserves their distributions and relationships. This makes it possible to fit a model whose parameters are very close to those that would have been calculated on the initial data. At the same time, since the data are synthetic, they can no longer be linked to a person, which completely preserves anonymity. This method can be implemented using the R package `synthpop`.

Here's how it's done: the set of variables is divided into two subsets:

- those which are not synthesized (possibly empty set);
- those that are synthesized.

The non-synthesized variables are not modified (be careful not to leave any quasi-identifier in them, unless you delete them).

For synthesized variables:

- the first variable is sampled with replacement from the initial dataset;
- the second variable is sampled conditionally to the first variable according to a method:
  - non-parametric: we build a CART decision tree on this variable whose predictor is the first variable;
  - parametric: a model is built on this variable whose type determines the model (linear regression for a continuous variable, logistic regression for a binary variable, multinomial or ordinal logistic regression);
  - sampling with replacement from the initial dataset (risk of not respecting the links between variables);
  - from a user-defined model (e.g. a neural network).
- the *pth* variable is sampled in the same way conditionally to the *p-1* previous variables.

Obviously the synthesis models (conditional distributions) are calculated on the observed data and not on the synthesized data.

Note that in all methods, it is better to start by removing the extreme values that are easier to identify and more difficult to anonymize.

Another avenue for privacy is *federated learning*. It was born in 2016 from a Google paper[55] proposing to replace the learning of deep learning models on large centralized databases with collaborative learning among users relying on a decentralized architecture.

In federated learning, a global model trained on all individuals is replaced by the aggregation of several local models each trained on a subset of individuals. The model training is first initialized (randomly or by transfer learning) on a server, then deployed

on each client terminal (e.g. smartphones) where the training continues based on the client data only, then the model updates are encrypted and sent to the server, which then aggregates all the updates into an optimized global model that is deployed on each client terminal. It is assumed that reverse engineering does not allow the central actor to guess the data of a local terminal from its updates.

The interest is in preserving the privacy of users, whose data do not leave their device. This is interesting for models trained on the Internet of Things, on medical data that will not leave the processing center, or on data from a consortium of companies that would like to work together without sharing their data. Moreover, the cost of data transfers is reduced.

A variant is where the variables, not the individuals, are divided into subsets. In all cases, the idea is that no one sees the whole data set.

The limitation of federated learning is the quality of the data. On the one hand, the theoretical distributions presupposed for the modeling are less frequently met locally than globally. On the other hand, by not gathering all the variables on all the individuals, we do not permit cross-tabulations which often allow us to detect inconsistencies and data quality problems. Finally, by only sending back to the central server the results of the local models and not the data that were used to train them, the quality of the backtesting *of* the models is severely limited: we can possibly follow their performance but we can no longer link this performance to the data of the model and note that a performance drop is linked to such or such data that should be corrected or replaced.

There is a Python library for federated learning: `PySyft`.

1.6.3 The General Data Protection Regulation

Harmonizing and strengthening the various existing European regulations, the General Data Protection Regulation (GDPR) which came into force on May 25, 2018, establishes new rules for the protection of personal data of European citizens, including:

- the right to erasure ("right to be forgotten"), i.e. the possibility for an individual to obtain the erasure of all personal data concerning him or her;
- the right of every individual not to be subject to a decision based exclusively on automated processing that produces legal effects concerning him or her or significantly affects him or her in a similar way;
- a new right to data portability: the individual must be able to recover, in a structured and commonly used computer format, his or her data communicated to an operator in order to keep them or transmit them to another operator;
- a duty to fully inform the person whose data are being collected in order to obtain their explicit and informed consent;
- taking into account the protection of personal data (*privacy by design*), from the very first stages of the design of the product or service that will process these data, in a way that is inherent to the company's/organization's information system (for example, by using pseudonymization and encryption);
- the application by default of the highest level of personal data protection (e.g. only necessary data should be processed, short retention period and limited and controlled accessibility) without any intervention from the user (whose default profile should preserve the confidentiality of his privacy) (*privacy by default*);

- the replacement of *a priori* declarations of processing by the obligation (*accountability)* for any data controller to be able to demonstrate compliance with data protection rules at any time, and therefore to have an internal control system;
- the appointment of a data protection officer in companies and organizations carrying out large-scale personal processing;
- the obligation for companies and organizations to notify as soon as possible to the national protection authority the cases of serious personal data breaches, and this authority may require the company to notify individually each person concerned;
- in the event of non-compliance with the GDPR, financial penalties of up to 4% of the company's annual worldwide turnover or €20 million, whichever is higher.

The GDPR applies to all companies processing personal data of EU citizens, even if they are established outside the EU or if their processing is carried out outside the EU.

The challenge of this regulation is to protect data without preventing its legitimate use, which requires, among other things, the ability to match data. If medical data cannot be matched with patients' socio-demographic and geographic data, it is obvious that their analysis loses some of its acuity.

The aim is to make those responsible for processing personal data accountable, in order to restore the confidence of those who entrust their data and, for those who use them, confidence in the quality and lawfulness of the collection of these data.

On April 21, 2021, the European Commission published its proposals for the regulation of artificial intelligence.[56] It wishes to both promote the development of artificial intelligence while addressing the risks it poses to the security and fundamental rights of Europeans. It classifies artificial intelligence systems into four categories according to their potential risk, which is judged:

- unacceptable, if AI aims to do the following:
  - manipulate human behavior to deprive users of their free will;
  - allow for social scoring by states;
- high, for AI use in:
  - biometric identification of persons;
  - critical infrastructures that could endanger the life and health of citizens;
  - education or vocational training;
  - product safety components (e.g. in robot-assisted surgery);
  - employment, workforce management and access to self-employment (e.g., résumé screening);
  - essential private and public services (e.g., credit risk assessment);
  - law enforcement;
  - the management of migration, asylum and border controls;
  - administration of justice and democratic processes;
- limited, if AI contributes:
  - for example, to conversational agents such as chatbots;
- minimal, if AI enters a majority of applications, including:
  - anti-spam filters;
  - video games.

The latter category is not regulated. To the limited risk category, specific transparency obligations apply: when talking to a conversational agent, a user must know that he or she is interacting with a machine in order to make an informed decision about whether to proceed.

High-risk artificial intelligence systems will have to meet strict conditions in order to be put on the market, including:

- adequate risk assessment and mitigation systems (risk management system);
- high quality of the data fed into the system;
- registration of activities in a public European database to guarantee the traceability of results;
- detailed documentation providing all the necessary information on the system and its purpose to enable the authorities to assess its compliance;
- clear and adequate information for the user;
- appropriate human control;
- a high level of robustness, security, and accuracy.

These rules still need to be adopted by the Member States and the European Parliament, but we can already see that the algorithms are in America and the standards are in Europe. This can lead to embarrassing situations, such as when regulatory authorities have difficulty understanding that in epidemiology a medical cohort must last a long time, and when in the midst of the Covid-19 pandemic they slow down the creation of a new cohort intended to study the locations of contamination. Moreover, the Cloud Act adopted in 2018 in the United States allows the American administration to seize all data relating to the electronic communications of a company or an individual, stored on the servers of an American company or operating in the United States, whether these servers are located in the United States or in another country, and without that individual being informed. This concerns, among others, all clients of GAFAM and the Cloud Act obviously conflicts with the GDPR since the data of a European citizen can be seized by the American authorities.

A new legal approach to data protection began to emerge from 2018: the patrimonialization of data that, belonging to the individual, could only be sold on a data market with the owner's consent (which is already the consequence of the legislation) and possibly with to be paid for, justified by the benefits derived from it by their users, starting with GAFAM.

1.7 Open Data

Open data are free public data, numerous and reputed to be reliable. However, they require a lot of work to widely share data collected for a specific use. They are mainly used in the fields of transport, health, libraries. They satisfy the growing demand of transparency of citizens and their wish to participate in public life.

Open data are the basis of a new business model, which has seen the creation of start-ups to exploit and value data better than the large companies or organizations that originate the data, and to provide services to individuals or other companies. For example, there are many smartphone applications on the topic of transport, using geolocation and data on infrastructure and schedules.

Crowdsourcing is the use of contributions from a large number of people to perform a task. Contributors may bring varying degrees of expertise, machine time from individual computers, or low-skilled but numerous volunteer labor. The motivations are diverse: personal development, distraction, play, self-promotion or altruism.

The general public can participate in the advancement of certain research, including:

- Wikipedia projects;
- papyrus transcription (https://www.ancientlives.org/#/);
- the Old Weather project (https://www.oldweather.org/) to capture nineteenth-century ship's logs containing interesting Arctic weather information;
- the Galaxy Zoo project (https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/) to classify more than one million galaxies;
- more generally, the Zooniverse portal (www.zooniverse.org) and its dozens of collaborative projects in medicine, biology, climatology, science, history, literature, art.

We can also say that Google Flu Trends (Section 1.4.1.3) is a passive form of crowdsourcing. The same is true for certain authentication systems on the Internet (CAPTCHA) that ask the user to recognize a car, a bridge, an animal in a series of photographs that are presented. This system is used not only to confirm that the user is a human being and not a robot, but also to make the user label images, which will then be used to train computer vision algorithms.

Crowdsourcing is most often voluntary and in several of the examples cited, it is qualified. Another form of activity, described as digital labor, is (poorly) paid and unskilled. It is the one performed by people who connect to platforms (the best known being Amazon Mechanical Turk), sometimes as a main activity but most often in addition to another salaried activity, to perform repetitive tasks that are necessary to train supervised algorithms: identifying and classifying images and videos, writing reviews, searching for information in documents, translating texts, transcribing recordings of intelligent voice assistants, etc. They are millions of "Turkers" (contraction of worker and Turk) and digital workers on other microwork platforms, who supplement their income in conditions that may seem undignified (low pay, lack of employment contract, lack of welfare benefit) but which are flexible and useful to them. It is not without humor that Amazon has given its platform a name that refers to the "Mechanical Turk," a so-called automaton from the end of the eighteenth century that could play chess and beat many human players. In reality, a real human player was hiding inside the cabinet and it can be seen as a metaphor for human intelligence, which is still indispensable in tasks that artificial intelligence alone cannot always perform well enough, whether it is analyzing images or words.

Notes

1 We include here in the same term statistics, statistical learning, and machine learning, for the sake of brevity and because the boundaries are shifting, despite our attempt to draw them in the Introduction.

2 McKinsey, Big Data, the Next Frontier for Innovation, Competition and Productivity, 2011. www.mckinsey.com/business-functions/mckinsey.

3 The best-known being Google Home, Amazon Echo, and Apple HomePod, allowing voice control of devices in a home. Amazon claims in early 2019 to have connected its Alexa smart voice assistant to more than 100 million devices worldwide, from its Echo connected speaker, of course, to home appliances and soon Volkswagen cars.

4 Laney, D., 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note. 6 (70) (2001).

5 Two more "V's" have since been added: veracity and value, but veracity has always been at the heart of statisticians' concerns, and the importance of value was already affirmed in data mining.

6 See, for example, https://www.domo.com/learn/data-never-sleeps-5, http://www.visualcapitalist.com/happens-internet-minute-2017/ and also http://www.internetlivestats.com.

7 https://www2.deloitte.com/content/dam/Deloitte/at/Documents/consumer-business/at-global-powers-retailing-2020.pdf.

8 http://www.gutenberg.org.

9 Pawlowicz, L.M. and Downum, C.E., Applications of Deep Learning to Decorated Ceramic Typology and Classification: A Case Study Using Tusayan White Ware from Northeast Arizona, *Journal of Archaeological Science*, 130 (2021), 105375.

10 Mirhoseini, A., Goldie, A., Yazgan, M. *et al.* (2021). A Graph Placement Methodology for Fast Chip Design. *Nature*, 594 (2021), 207–212.

11 Preis, T., Moat, H.S, and Stanley, H.E., Quantifying Trading Behavior in Financial Markets Using Google Trends, *Scientific Reports* 3 (2013), 1684, http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html.

12 https://trends.google.com/.

13 https://support.google.com/trends#topic=6248052.

14 An API (Application Programming Interface) is a programming interface that allows software to provide services to other software. We will discuss several of them in this book.

15 http://www2.cnrs.fr/en/2013.htm.

16 https://www.unglobalpulse.org/data-for-climate-action.

17 http://www.scientificamerican.com/article.cfm?id=how-big-data-taking-teachers-out-lecturing-business&page=2

18 https://openclassrooms.com/.

19 https://www.fun-mooc.fr/.

20 https://www.coursera.org/.

21 http://www.flowminder.org/publications/containing-the-ebola-outbreak-the-potential-and-challenge-of-mobile-data.

22 http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1001083

23 Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L., Detecting Influenza Epidemics Using Search Engine Query Data, *Nature* 457 (2009), 1012–1014, http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html. See also http://www.google.org/flutrends/intl/en_us/about/how.html and http://websenti.u707.jussieu.fr/sentiweb/?page=google.

24 https://www.google.org/flutrends/.

25 *Nature*, 494 (2013), 155–156, 14 February 2013 (http://www.nature.com/news/when-google-got-flu-wrong-1.12413).

26 https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41763.pdf.

27 Justified by Google, which argues that the disclosure of the search terms used in its model immediately leads to an artificial increase in the input of these terms.

28 Preis, T. and Moat, H.S., Adaptive Nowcasting of Influenza Outbreaks Using Google Searches. *Royal Society Open Science*, 1 (2014), 140095, http://rsos.royalsocietypublishing.org/content/1/2/140095.

29 Yanga, S., Santillanab, M., and Koua, S., Accurate Estimation of Influenza Epidemics Using Google Search Data Via ARGO, *PNAS*, 112(47) (2015), 4473–4478.

30 Senior, A.W. *et al*., Improved Protein Structure Prediction Using Potentials from Deep Learning, *Nature*, 577 (2020), 706–710.

31 Jumper, J. *et al*., Highly Accurate Protein Structure Prediction with AlphaFold, *Nature*, 596 (7873) (2021), 583–589.

32 https://www.isomorphiclabs.com/.

33 Vig, J., Madani, A., Varshney, L.R., Xiong, C., Socher, R., and Rajani, N.F., BERTology Meets Biology: Interpreting Attention in Protein Language Models, arXiv:2006.15222. (2020).

34 In this regard, we should not forget the first fatal accident in May 2016 when a Tesla car hit a truck that it had not seen turning. Tesla said at the time that this was the first fatal accident in more than 200 million miles already driven by its vehicles. In March 2018, another accident involved an autonomous vehicle and a pedestrian who was hit by the vehicle and killed by the collision.

35 SAE International (J3016) Automation Levels.

36 https://geolitica.com/.

37 https://cortecs.org/wp-content/uploads/2014/10/rapport_stage_Ismael_Benslimane.pdf.

38 https://themarkup.org/show-your-work/2021/12/02/how-we-determined-crime-prediction-software-disproportionately-targeted-low-income-black-and-latino-neighborhoods.

39 Andridge, R. and Little, R., A Review of Hot Deck Imputation for Survey Nonresponse, *International Statistical Review*, 78 (2010), 40–64.

40 Rubin, D.B., Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse, in *Proceedings of the Survey Research Methods Section* (American Statistical Association, 1978), pp. 20–34.

41 Chauvet, G., Deville, J.C., and Haziza, D., On Balanced Random Imputation In Surveys, *Biometrika*, 98 (2011), 459–471.

42 Danel, T. *et al*., Processing of Incomplete Images by (Graph) Convolutional Neural Networks (2020). arXiv:2010.13914v1

43 http://www.tylervigen.com/spurious-correlations.

44 Hill, A. B., The Environment and Disease: Association or Causation?, *Proceedings of the Royal Society of Medicine* 58 (5) (1965), 295–300.

45 Ribeiro, M. T., Singh, S., and Guestrin, C., "Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016). arXiv:1602.04938.

46 Goodfellow, I.J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples (2014). arXiv:1412.6572.

47 Thom, R., *Prédire n'est pas expliquer* (Éditions ESHEL, 1991).

48 Ribeiro, *et al*., op. cit.

49 Štrumbelj., E. and Kononenko, I., Explaining prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems* 41(3) (2014), 647–665.
50 Allard, T., Nguyen, B., and Pucheral, P. L'art de préserver l'anonymat, *Pour la science*, 98 (2018), 98–103.
51 Narayanan, A., and Shmatikov, V., How to Break the Anonymity of the Netflix Prize Dataset (2007). arXiv:cs/0610105.
52 Nguyen, B., Techniques d'anonymisation. *Statistique et Société*, Société française de statistique, 2(4) (2014), 53–60.
53 Samarati, P., and Sweeney, L. Generalizing Data to Provide Anonymity When Disclosing Information (Abstract), in *PODS '98: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (1998).
54 Dwork, C., Differential Privacy. Paper presented at International Colloquium on Automata, Languages and Programming (ICALP) (2006).
55 McMahan, H.B., Moore, E., Ramage, D., and Aguera y Arcas, B. Federated Learning of Deep Networks Using Model Averaging (2016). arXiv:1602.05629v1.
56 https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682.