

Introduction – The Contributions of Big Data

Philippe Tassi*

Abstract – The revolution, which is quite recent, brought about by digital convergence and connected objects, has enabled a homogenisation of data types which would historically have been considered as different, for example: digital data, texts, sound, still images, and moving images. This has encouraged the Big Data phenomenon, the volume of which includes two related parameters: quantity and frequency of acquisition; quantity can extend as far as exhaustivity and frequency can be up to and including real time. This Special Issue features a series of articles that examine its uses and implications, as well as the challenges faced by statistical production in general, and especially that of official statistics. Just like any innovation, Big Data offer advantages and raise questions. The obvious benefits include “added” knowledge – a better statistical description of the economy and the society. They are also a driver for development in computer science in the broadest sense, and in applied mathematics. However, we cannot do without some degree of vigilance, since data and how they are used can affect individuals, their freedoms and the preservation of their privacy.

JEL Classification: C1, C8

Keywords: digital data, Big Data, statistics, official statistics

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

* *Médiamétrie* (ptassi@mediametrie.fr)

Received on 21 March 2019

Translated from the original version: “Introduction – Les apports des Big Data”

To cite this article : Tassi, T. (2018). Introduction – The Contributions of Big Data. *Economie et Statistique / Economics and Statistics*, 505-506, 5–16.
<https://doi.org/10.24187/ecostat.2018.505d.1963>

Some History – and Histories

Although the term “data” may seem modern, especially when preceded by “big”, we ought to remember that data is no other than the plural form of the supine of the Latin verb “do”, “das”, “dare”, “dedi”, “datum”. Moving beyond the Latin origins of the word, mass or even exhaustive data collection is not a phenomenon from the present digital era; it is an activity that began as soon as writing emerged, since the latter was a necessary pre-requisite for it. Most historians and archaeologists believe that writing first appeared in Lower Mesopotamia (present-day Iraq) approximately 5,000 years ago, at a time when nomadism was in decline and the first settlements were being established, which led to the birth of the cities of Sumer. Since memory alone would no longer suffice to understand, manage, and govern these cities, written marks had to be used. The site at Uruk (Erech in the Bible) has yielded many clay tablets dating from the fourth millennium BCE, tablets covered with signs traced using a reed stylus - the origin of cuneiform script, a structured writing system involving several hundred signs.

Data collection could then begin, starting with two main areas of interest: astronomy and the counting of populations. As Jean-Jacques Droysbeke wrote: “[...] the Mesopotamians were very early adopters, [...] and in ancient Egypt also, from the end of the third millennium BCE [...] to know how many men were available to participate in the construction of temples, palaces, pyramids [...] or even [...] for tax purposes.” Data collection was not limited to these city-states. China and India had systems covering large territories in the last millennium BCE. China had its “directors of the masses”. In India, the Maurya Empire covered a vast territory similar to present-day India, and its first emperor, Chandragupta, established a census in the 4th century BC.

Turning to data processing, and given that the expression “artificial intelligence” (AI) is now in common usage, let’s give it a definition and an historical perspective. Yann LeCun, who was the 2016 Chair of “Informatics and Computational Sciences” at the College de France as well as the first Director of the Facebook Artificial Intelligence Center in New York and later in Paris, and also a leading figure in AI and Deep Learning has defined AI as follows: “making machines complete tasks that would normally be assigned to people and animals”. For the historical perspective, we might look back to Babylon or the Chinese Empire, since even at that early stage, it seemed natural to try to model human brain behaviour and depict man as a machine as a precursor to the design of learning machines.

One forerunner of Artificial Intelligence was the Catalan philosopher and theologian, Ramon Llull (1232 - 1315), who invented “logic machines”. Predicates, subjects and theories were organised into geometric figures that were deemed perfect (circles, squares, triangles). With the aid of dials, cranks and levers to turn a wheel, the propositions and theses were moved into position to reveal their truth or falsehood. Llull was a major influence on his contemporaries and even beyond, since four centuries later, Gottfried Leibniz would find inspiration in his work.

From Sampling to Big Data: Complementary Paradigms

We could say that the world has lived under the near-total reign of exhaustive data collection, however, a few rare approaches at sampling did exist in the mid-17th

century: John Graunt and William Petty's school of political arithmetic in England and Vauban's advances in France. The 20th century featured a slow decline in census and exhaustive data collection, and an ever more assertive rise of the sampling paradigm. The founding act was the speech by Anders N. Kiaer, Director of the Central Bureau of Statistics in Norway, during the International Statistical Institute's Berne Congress in August 1895: the first recognition for the *pars pro toto*.

In 1925, the ISI validated Kiaer's approach, and developments followed swiftly: the benchmark paper on surveys would appear in 1934 (Neyman, 1934). Operational applications rapidly ensued: in the economic field, following J. M. Keynes's articles in the early 1930s, the first consumer and distributor panels emerged in 1935, run by companies such as Nielsen in the United States, GfK in Germany. In 1935, George Gallup launched his company in the United States: the American Institute for Public Opinion (AIPO). He went on to achieve fame among the general public after he used a sample of voters to predict Franklin D. Roosevelt's victory over Arthur Landon in the 1936 presidential election. In 1937, Jean Stoetzel created the French equivalent organisation – l'Institut Français d'Opinion Publique (IFOP), the first opinion research company in France.

In the post-war years, sampling became the reference due to its operational speed and reduction of costs. With the advancement of probability, statistics and information technology, we also witnessed a general expansion into new fields such as economics, official statistics, health, marketing, sociology, media audiences, political science, etc. For most of the 20th century, therefore, the sampling paradigm prevailed and exhaustive censuses went into decline; we should note that in the 1960s, official statistical censuses of the population, agriculture and industry were still in existence.

Since the end of the 20th century and the start of the 21st century, digital convergence has favoured the automated collection of data for ever larger populations, generating databases that hold a growing mass of information, and heralding a return to exhaustive collection. Additionally, under the digital transition, it has become possible to harmonise information that was previously considered distinct and heterogeneous such as: quantitative data files, text files, (audio) sound files, photos and moving images (video). The two main parameters that help to define the volume of Big Data are: quantity and frequency of acquisition; quantity can extend to exhaustivity, and frequency can be up to and including real time.

Issues Raised by Big Data

Among the various issues raised by Big Data, some are old and some are new. They concern processing methods, storage, protection and security, property rights, etc. What statistical processing or algorithms should be applied to the data? What status does the data have and what is the status of the data author/owner? What is the regulatory or legislative framework like?

An enduring phenomenon

Big Data is clearly more than just a fad. We are only beginning to exploit it. Every day, new examples of Big Data appear across ever-expanding areas: medicine,

epidemiology, health, insurance, sport, marketing, culture, and human resources, not to mention official statistics.

Digitalisation has lent weight to methodologies, modelling and technologies, as well as to their related professions. Innovations in algorithmics and machine learning when applied to Big Data represent a rapidly growing field, from the genius of Alan Turing to Arthur Samuel, Tom Mitchell, Vladimir Vapnik and Alexey Chernovenkis (Vapnik, 1995, 1998). The digital world is everywhere, investments are not fleeting, and the policy orientation of states is clear. In France, clear guidance was given in the thirty-four proposals for an industrial renaissance in France (François Hollande, September 2013). The report by the Innovation 2030 Commission chaired by Anne Lauvergeon placed particular emphasis on the excellent reputation of French training in mathematics and statistics. This was further demonstrated by the strong showing of “French Tech” at the Consumer Electronics Show (CES) in Las Vegas. As part of its strategic reflection “Insee 2025”, Insee addressed access to private data and its use in official statistics. Connected objects and the “Internet of Things” are strengthening this phenomenon (Nemri, 2015).

Trust

In general, data and statistics produced by governments or companies are based on personal information, which raises questions about how to protect these sources, i.e. their privacy. Given the constant advances in science and data processing procedures, how can we establish and maintain the confidence of the general public, who are the leading stakeholder, whilst respecting the balance between the promise of confidentiality and the use of the collected data? The answer lies in two complementary approaches: one regulatory, since States have long been aware of the need to establish legal safeguards; and one technological, by erecting technical barriers to prevent the unwarranted dissemination of data.

Significant Regulatory Framework

In the field of statistics, many countries have a legislative framework, among them France which played a pioneering role with its Data Protection Act (*Loi Informatique et Libertés*) from 1978. An even earlier French law of 7th June 1951 related to obligation, coordination and secrecy in the statistical domain. It defined statistical secrecy as the “impossibility of identification” in the context of official statistics (censuses and surveys). Accordingly, any communication of personal, family or private data was prohibited for 75 years. The 23rd October 1984 Post and Electronic Telecommunications Code (*Code des Postes et Télécommunications électroniques*) and its various amendments addresses the processing of personal data in the context of electronic communication services, in particular via networks that support devices that collect and identify data. France’s Conseil d’État also published a work entitled “Fundamental Rights in the Digital Age”, containing fifty proposals to ensure that digital technology supports the rights of individuals and the public interest and including a chapter on “predictive algorithms” (Rouvroy, 2014). We should also mention professional codes of ethics, such as that of the European Society for Opinion and Market Research (ESOMAR), established in 1948, and regularly updated to clarify best practice when conducting market and opinion research.

The most famous law known to the general public in France is probably the law of 6th January 1978 on data processing, data files and individual liberties, known in

French as “*Loi Informatique et Libertés*” (Data Protection Act). It specifies the rules that can apply to personal data. The first article of the 1978 law clarifies: “personal data is taken to mean any information relating to a natural person who is or can be directly or indirectly identified by reference to an identification number or to one or more factors that are specific to them.” These personal data may either be kept raw or may be processed and then stored. The law stipulates that processing means: “any operation or set of operations performed on such data using any mechanism whatsoever, and in particular the collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or interconnection, as well as locking, deletion or destruction.” The significance of this lies in the fact that Big Data is “massive” in both of the aforementioned senses: quantity and variety (the 6 Vs). The extent of the analysis that can be deduced from data calculated by inference is also important.

One special category of personal data is sensitive data, the collection and processing of which are prohibited as a matter of principle. Sensitive information is taken to mean information that directly or indirectly reveals a person’s racial or ethnic origins, political, philosophical or religious opinions, trade union affiliations, or information that concerns their health or sexual life (Article 8). Finally, the GDPR (General Data Protection Regulation) enacted in 2016 and in force across EU Member States since May 2018, has been the focus of all attention; all the more so since it will be followed by the ePrivacy regulation, a *lex specialis* to the GDPR.

Technology and Data Confidentiality

The relationship between information technology, privacy, personal data and databases is a fairly long-established research area, having been formally addressed since the 1970s. Respect for privacy is also a principle on which the whole world seems to agree *a priori*. Is it possible to guarantee this respect in the technology arena?

Cybersecurity and methods of encryption have evolved a lot since their beginnings over three thousand years ago. These methods can be used to render a document (here used in the broadest sense of the term) unreadable – i.e. incomprehensible to anyone who does not possess the encryption key. Julius Caesar encrypted the messages he sent to his generals and the Rossignol des Roches family (Antoine Rossignol, his son Bonaventure and grand-son Antoine-Bonaventure) ran Louis XIV’s “*Cabinet Noir*” (Black Chamber) and operated the world famous 17th century “Great Cipher”. Lastly, we all know of Claude Chappe’s telegraph coding from the late 18th century, as well as Samuel Morse’s electric telegraph which arrived a few years later.

The vision of Tore Dalenius

In the context of the databases in existence prior to 1980, the Swedish statistician Tore Dalenius laid down some principles on ethics and respect for privacy. In his article (Dalenius, 1977), he set out the following principle: “Accessing a database should not allow you to learn more about an individual than could be learned without access to that database.” He added: $X(i)$ being the value of the variable X for an individual i , if the publication of a statistical aggregate T helps to determine $X(i)$ specifically, without access to T , then a breach of confidentiality has occurred. This principle seems acceptable. Unfortunately, we can demonstrate that it cannot be

generalised: any third party who would like to collect personal data about individual i can do so by taking advantage of auxiliary information accessible from outside the database.

Anonymisation

One *a priori* intuitive technique of data protection consists of rendering the personal data anonymous. This is tantamount to removing all of the variables in the database that could identify a particular person. It is a reiteration of the personal data concept as expressed in French data protection law. Of course, a natural person is identifiable by his or her name, but also through other characteristic variables such as a registration code, an address (postal or IP), phone numbers, a PIN code (Personal Identification Number), photographs, or biometric components such as a fingerprint or DNA. More generally, identification is possible through any variable that can be used in crossing or cross-checking to find an individual in a group (e.g. their place of birth, date of birth or local polling station). This represents a less perfect and less immediate identification than using their name, however, it remains highly likely that the person would be identified, which is a far cry from complete ignorance.

For the last ten years or so, information and communication technologies have been generating lots of data that is useful for the previously mentioned analysis, for example, calls from a mobile device or connections to the internet. All of these “computer traces”, or “logs”, can easily be exploited thanks to advances in software and search engines. On the face of it, anonymisation is a simple concept to understand and to implement, however, it can become complex and also risks the deletion of useful or relevant variables from the database. Furthermore, as science progresses, the number of privacy breaches is increasing and the probability of identifying an individual from a database containing personal data is higher too, even after anonymisation.

Destruction or Aggregation of Data

Another method is to delete the data once a certain period of operational use has elapsed. However, deleted data can be valuable long after its “working life”, e.g. for historians and researchers. If we reuse the principle of France’s 1951 law on statistical secrecy in companies, it would then be possible to aggregate the individual data and only divulge these aggregated results once a certain amount of time has passed.

Data Masking

Obscuring data (also called data obfuscation or masking) involves maintaining the confidentiality of data by deliberately “altering” it. This can be done indirectly by burying the data in a bigger environment (along the same lines as diluting meaningful data) or else directly by transforming the data to make it insignificant. For the first of these methods, we might, for example, create additional variables that increase the data vector’s size, thereby creating a “fog” in which to hide our data. The second group of methods is characterised by techniques that are non-disruptive: masking the value of some cells in a table of results; removing variables concerning certain individuals; dividing a sample of data extracted from the database; or combining certain categories for variables with modalities, etc.

Particular methods also exist that directly intervene on the data to create noise, in the broadest sense of the term, and modify certain variables by rounding or blocking them *via* truncation at maximum or minimum thresholds. The transformation of variables can also be achieved by applying a homomorphism, swapping the value of the same variable between two individuals, or by means of data perturbation through random noise injection. Applied to the original data, some transformations (e.g. swapping, rotation) will leave the linear statistics invariant, whereas others will not. Arising out of work into missing data (Little, 1993; Rubin, 1993, 2003), this approach is especially relevant to synthetic data.

A New Approach: Differential Privacy

Since the mid-2000s, there has been another perspective on privacy protection (Dwork, 2004, 2006), and its philosophy owes a lot to Dalenius: “The probability of a negative consequence of any kind for individual i (for example: being refused credit or insurance) cannot significantly increase because of the representation of i in a database.”

We should nuance the adverb “significantly” here because it is very hard to predict what information – or what combination of information – might have negative consequences for the individual in question, were this information to be made public. All the more so since this information cannot be observed but rather is estimated using a calculation, and also because some consequences deemed negative by one person may on the contrary seem positive to someone else! This approach, which we could name “privacy” or “differential privacy” is based on probabilistic and statistical suppositions. Could this approach be expanded? The idea is to quantify the risk of a possible privacy breach, whilst at the same time measuring the impact of effective data protection on privacy in statistical terms. This opens up a new field of research that will analyse data post-obfuscation, alteration or modification of the original in order to maintain confidentiality.

Mathematical Statistics, Econometrics and Big Data: An Inevitable Convergence

Statisticians and econometricians have been slow to familiarise themselves with the volumetry and techniques derived from machine learning, which did not provide direct answers to classic problems such as the accuracy of estimates or causality. Change is under way with the creation of bridges to machine learning and artificial intelligence. In terms of data, it is pointless to pit *sampling data* against *Big Data*. Far better to try to bring them closer, hybridising these two information sources to obtain a third, richer source.

Similarly, in methods and tools, it serves no purpose to set econometrics against machine learning. These approaches have been developed in response to different yet complementary questions, and there is real convergence between these disciplines: econometrics borrows some machine learning methods and vice versa; the causality concepts that econometricians hold dear are among those themes that have been identified as ways to advance machine learning research. Data scientists now have a wider range of tools at their disposal: convolutional neural networks (deep learning), support vector machine approaches (SVM); random forests and boosting,

not to mention adapted software and libraries. Nevertheless, we need to remain aware of the possible limitations of Big Data and new tools, and know that predictive machine learning technology could possibly predict what is observed in the data. This convergence has become all the more inevitable with the emergence of quantum computing.

A Special Issue on Big Data and Statistics

Economie et Statistique / Economics and Statistics is devoting two volumes of a special issue to Big Data. This first volume is wider in scope, featuring eight articles with a mix of areas for reflection, applications and methodology. The second volume (forthcoming) will address the theme of price indexes.

The first article by **Clément Bortoli, Stéphanie Combes and Thomas Renault** deals with France's quarterly GDP growth forecast adjusted for seasonal variations and working days. The authors compared the use of a simple autoregressive model (AR) with an AR model featuring a "business climate" variable and a "media sentiment" variable. Elaborating a media sentiment indicator allows us to gauge the overall tone of a media base, and a press title in particular. The integration of this indicator in the model provided some promising results, whether we are looking at the advance GDP forecast (forecasting) or the immediate forecast (nowcasting).

François Robin's article examines the modelling of e-commerce turnover based on data sourced from FEVAD. The Banque de France traditionally uses a SARIMA (12) model, and the author's approach is to complement this model with data from the Monthly Business Survey and data from Google Trends. Illustrating a major advantage of Big Data, Google Trends data is available almost in real time. It analyses the mass of search queries on Google's search engine to construct monthly indices for the terms employed. As independent sources, they are available before the FEVAD results and make a nowcasting approach possible. The technique used stems from machine learning (adaptive lasso method).

Pete Richardson's paper focuses on short-term macroeconomic forecasting and immediate forecasting, (also called nowcasting) that was conducted using Big Data, internet search queries, social media, and financial transactions – i.e. a wider set of databases than the ones traditionally used by national statistical institutes. In a broad-spectrum piece, he analyses a variety of applied research: labour market, consumption, housing market, travel and tourism, and financial markets. The author explains the limitations on what data from web searches can contribute, and he seems to have a preference for data sourced from social media. In his conclusion, he focuses on four particular areas in which to improve these new models and new data: quality and accessibility, information extraction methods, comparison of measurement methods, and improvements to testing and modelling.

The next two articles analyse the contributions of a special type of big data – data sourced from mobile phone operators, which is all the more interesting in light of the penetration rate of these phones among the population. **Guillaume Cousin and Fabrice Hillaireau** tackle the tourism sector and, more particularly, attempt to estimate foreign tourist numbers by counting the number of foreign visitors and their overnight stays. Currently, the survey of visitors from abroad is based on traffic data according to their mode of transport. Using this as a starting point for their estimates, the authors also used counting and surveys. For the time being, this trial conducted

since the summer of 2015 has come to the conclusion that mobile phone data is relevant as a complement to the current mechanism rather than a replacement for it. The experiment also identified the limitations of and areas for improvement in this new information source.

The use of mobile phone data to estimate the resident population is studied and analysed by **Benjamin Sakarovitch, Marie-Pierre de Bellefon, Pauline Givord and Maarten Vanhoof**. This exploratory article manages to construct a detailed overview of the current limitations and issues raised regarding this kind of data, as well as its significance and potential. One example of the difficulties they encountered was the uneven territorial coverage caused by variable antenna density which led them to resort to using a Voronoï tessellation (division of the area into polygons of varying sizes). A second problem was how to adjust the data to move away from the subscriber population to the total population. This first exploration has demonstrated that, at the current stage of development, it is still complex and premature to align exact counting statistics such as those currently produced by official statistics. Nevertheless, this mobile phone input source is of potential relevance to some approaches, such as the study of social and spatial segregations.

In the context of media audience measurement, **Lorie Dudoignon, Fabienne Le Sager and Aurélie Vanheuverzwyn** approached a concrete example of the complementarity of panel data and Big Data from a methodological perspective, thereby offering an illustration of the hybridisation of these two database types. Although traditionally based on data from individual sampling, nowadays, the mechanisms for measuring media performance, at least as far as the internet and potentially some television services are concerned, have integrated big data found in real time via equipment such as broadband routers. Once the Big Data found in the objects has been cleaned up (“big” does not necessarily mean “perfect”), the methodological basis for the hybridisation of the two data types is provided by a hidden Markov model, which is used to arrive at an equivalent level of granularity for the two sources, i.e. at the level of individuals, the state of an object such as a router that is supplying no information about the number of viewers or their socio-demographic variables.

The article by **Arthur Charpentier, Emmanuel Flachaire and Antoine Ly** illustrates the necessary convergence between econometric techniques and learning models. The proximity and differences between learning and econometrics are demonstrated. The authors introduce neural networks, the SVM approach, classification trees, bagging, and random forests, as well as sketching out the impact of big data on models and techniques in several application fields. They conclude that, although the two cultures – econometrics and learning – developed in parallel, an ever-increasing number of bridges exists between the two.

Finally, the article by **Evelyn Ruppert, Francisca Grommé, Funda Ustek-Spilda and Babi Cakici** examines the significant issue of trust in official statistics, in the current big data context. The authors return to the importance of respect of privacy, data protection, and especially of the need to rethink the relationship with the public, who supply the raw material used to produce statistical indicators, in particular through national institutes. Big Data, which are not from public sources, influence the notion of trust; the co-production of “citizen data”, defined as the participation of citizens in all stages of production, is a fundamental principle. □

BIBLIOGRAPHY

- Dalenius, T. (1977).** Towards a methodology for statistical disclosure control. *StatistikTidskrift*, 15, 429–444.
- Desrosières, A. (1993).** *La politique des grands nombres. Histoire de la raison statistique*. Paris : La Découverte.
- Droesbeke, J.-J., Saporta, G. (2010).** Les modèles et leur histoire. In : Droesbeke, J.-J. & Saporta, G. (Eds), *Analyse statistique des données longitudinales*, pp. 1–14. Paris : Technip.
- Droesbeke, J.-J., Tassi, P. (1990).** *Histoire de la Statistique*. Paris : PUF.
- Dwork, C. (2006).** Differential Privacy. *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, 1–12.
https://link.springer.com/chapter/10.1007/11787006_1
- Executive Office of the President (2014).** *Big Data: Seizing Opportunities, Preserving Value*.
<https://obamawhitehouse.archives.gov>
- Fisher, R. A. (1922).** *On the Mathematical Foundations of Theoretical Statistics*. *Philosophical Transactions of the Royal Society*, 222(594-604), 309–368.
<https://doi.org/10.1098/rsta.1922.0009>
- France Stratégie & CNNum (2017).** Anticiper les impacts économiques et sociaux de l’Intelligence Artificielle. Rapport du groupe de travail 3.2.
<https://www.strategie.gouv.fr/publications/anticiper-impacts-economiques-sociaux-de-lintelligence-artificielle>
- Hamel, M.-P., Marguerit D. (2013).** Analyse des big data. Quels usages, quels défis ? France Stratégie, *Note d’analyse* N 08.
<https://strategie.gouv.fr/publications/analyse-big-data-usages-defis>
- Jensen, A. (1925).** Report on the Representative Method in Statistics. *Bulletin de l’Institut International de Statistique*, 22(1), 359–380.
- Kiaer, A. N. (1895).** Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l’Institut International de Statistique*, 9(2), 176–183.
- Little, R. (1993).** Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2), 407–426.
- Nemri, M. (2015).** Demain l’internet des objets. France Stratégie, *Note d’analyse* N° 22.
<https://strategie.gouv.fr/publications/demain-linternet-objets>
- Neyman, J. (1934).** On the Two Different Aspects of Representative Method Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–625.
<https://doi.org/10.2307/2342192>
- OPECST (2017).** Pour une intelligence artificielle maîtrisée, utile et démystifiée. Rapport N°464.
<https://www.senat.fr/notice-rapport/2016/r16-464-1-notice.html>
- PCAST (2014).** Big Data and Privacy: A Technological Perspective. Report to the President.
https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy

Rouvroy, A. (2014). Des données sans personne : le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data. In : *Étude annuelle du Conseil d'État : le numérique et les droits fondamentaux*, pp. 407–422. La Documentation Française

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*, 9(2), 461–468.
<https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf>

Rubin, D. B. (2003). Discussion on Multiple Imputation. *International Statistical Review*, 71(3), 619–625.
<https://www.jstor.org/stable/1403833>

Singh, S. (2000). *The Code Book*. London: Fourth Estate Ltd.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Villani, C. (2018). Donner un sens à l'Intelligence Artificielle. Rapport public.
<https://www.ladocumentationfrancaise.fr/rapports-publics/184000159/index.shtml>.

