Statistics and Probability Letters 136 (2018) 68–72

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the role of statistics in the era of big data: A computer science perspective

Stefano Ceri

DEIB, Politecnico di Milano, Milano, Italy

ARTICLE INFO

Article history:
Available online 22 February 2018

Keywords:
Big data
Paradigm shift in statistics and computer science
Fourth paradigm
Two cultures
Teaching data science

ABSTRACT

Statistics and computer science are facing remarkably similar discussions on the role of big data. In this article, I advocate that the computer science community has taken advantage of big data since about five decades, thereby building the main commercial companies of today's computer industry, and specifically I describe the new emphasis on data as the emergence of the so-called Fourth Paradigm. Then, I draw a parallel between the debates on big data occurring within the statistics and computer science community; and finally I advocate for a joint, new and pervasive approach to data science, in which both communities can capitalize on each other's skills.

1. Big data in the last five decades

"Big data" is a successful buzzword, used (and abused) to denote a paradigm shift of the scientific approach. As I am a computer scientist, and more specifically a member of database community, I do not consider big data as a revolution. Rather, big data have been the main ingredient of more than five decades of data management research; what happened in recent years is just the natural evolution of what happened in the past.

The database community is aware of the need of managing big data since its infancy. Turing award Edward T. Codd invented relational databases in the early seventies (Codd, 1970). The Very Large Data Bases (VLDB) Conference, which is the most prestigious in the field, started in 1975. Since the mid eighties, companies such as Oracle, IBM and Microsoft built a substantial portion of their commercial success thanks to data management software; data management initially applied to financial world (banks, trading) and then extended to all commercial activities. More recently, search methods over big data are key to the success of Google and Yahoo, while data analysis methods such as association rules are key to the success of Amazon. In the new century, social networks such as Facebook, LinkedIn, Twitter and Instagram are large data collections associated to a business model which emphasizes data sharing; similarly, the so-called "sharing economy" (AirBNB, Huber) is based on social use of big data. The growth of big data can be witnessed by looking at the Go-Globe web site,[1] which describes the information generated on Internet in one minute — including 44 million messages, 2,3 million Google queries, 3 million likes and 3 million shares on Facebook, 2,7 million downloads from YouTube.

The exponential growth of data (both in quality and size) has influenced many other sectors of computer science; among them, artificial intelligence is providing renewed emphasis on machine learning and deep learning as data-driven model of knowledge creation. Requirements of big data management of companies such as Google and Yahoo are leading to the development of cloud computing, which in turn has reshaped the way in which we use computer technologies nowadays.

More in general, the dualism between data and algorithms has shaped the entire evolution of computer science (recall 1975 Wirth's book *Algorithms* + *Data Structures* = *Programs*), and emphasis has periodically been given to either one aspect or the other; recent years have featured an increased interest in data.

E-mail address: stefano.ceri@polimi.it.

[1] https://www.go-globe.com/blog/60-seconds/ consulted on 7/25/2017.

https://doi.org/10.1016/j.spl.2018.02.019
0167-7152/© 2018 Elsevier B.V. All rights reserved.

2. Fourth paradigm

The emphasis on big data in computer science is described in the *Fourth Paradigm* book (Fourth Paradigm, 2009), dedicated to the memory of Turing award Jim Gray. In the preface, the history of science is historically separated into four phases: the first based upon empirical science and observations, the second upon theoretical science and mathematically-driven insights, the third upon computational science and simulation-driven insights, the fourth upon data-driven insights of modern scientific research. Accordingly, we have entered the fourth phase, featuring the data-driven approach.

It is worthwhile to recall the words of Jim Gray about what makes big data amenable to effective processing. He claims the importance that all data being used, no matter how assembled, should be self-describing and should have a schema. In this way, it is possible to address data semantics, e.g. to talk about stars and galaxies and their properties. Once a schema is well-defined, data can be indexed, aggregated and searched, and it is easier to build both ad-hoc queries and generic visualization tools. If instead big data are just stored as files, it is not possible to use the concepts of star or galaxy; the data scientist has to understand the data content in each file, in a bottom-up and unstructured fashion. Essentially, these words are calling for a layer expressing data semantics and organization which should be separate from data content, in contrast to using data without understanding their structure.[2]

The legacy of Jim Gray is huge. I still remember when, about fifteen years ago, he was explaining to me his joint work with astronomers while at the same time I was looking into classical computer science problems, and I could not understand his enthusiasm. A posteriori, I was still trapped into a disciplinary silo, he had already moved towards interdisciplinary data science.

3. Two cultures in two contexts

A remarkably similar discussion about *two cultures* occurs in the scientific communities of computer science and statistics. Piercesare Secchi recalls the 2016 meeting of the Italian Statistical Society, and specifically a discussion of how statistical modeling reaches conclusions from data; citing Leo Breiman, there are two cases: "one assumes that data are generated by a given stochastic model"; the other "uses algorithmic models and treats the data mechanism as unknown". According to Breiman (2001), in the past "the statistical community has committed to the almost exclusive use of stochastic models". However, algorithmic modeling "can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets".

A similar debate occurred at the European Computer Science Summit (EECS 2015 in Vienna); EECS is a series of meetings where the computer science community (with over 115 member institutions across 29 countries) annually debates about research and education in Information and Computer Sciences in Europe. At EECS 2015, I defended *data science* (Ceri, 2017) and Moshe Vardi defended *formal science* (Vardi, 2015) during a two-headed session; we actually preferred to discuss the dualism of *data-driven* and *model-driven* approaches to science.[3]

According to the *model-driven* approach, models are built in the mind of scientists, then they are formally expressed, and then can be tested; experiments confirm or falsify them. In statistical terms, data are generated by a stochastic model, which must be understood; according to Breiman (2001), in the past "the statistical community has committed to the almost exclusive use of (such) data models".

With the advent of big data, however, the classic approach to science hypothesize, model, test has been put in question, and a *data-driven* approach is emerging, which consists of being driven by data rather than by a-priori assumptions. An extreme point of view (citing Chris Anderson) is that "We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot".

In our debate, no extreme position prevailed; we essentially agreed with Moshe Vardi's abstract that "the data-driven approach does not replace the formal-model approach"; he cited many recent research experiences where "the data-driven approach stands on the shoulders of the formal-model approach". Breiman came to similar conclusions in predicting the future of statistics. "If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools".[4]

An example of shift from model-driven to data-driven approach is offered by the pharmaceutical industry. While the main discoveries in medicine of the last century are due to enlightening intuitions of scientists, the typical initial step of drug discovery, e.g. described in Bayer (2015), uses massive high-throughput screening. Proteins play a significant role in the course of a disease, as drugs can either switch these proteins off or enhance their function. Once a target has been successfully identified, systematic test procedure is used to look for substances which could be a suitable starting point for a new active ingredient, by applying an in-house compound library (e.g. Bayer's currently has over three million chemical substances). Robots fill thousands of microtiter plates on which thousands of tests are performed simultaneously. Thus, the initial phase of drug discovery is mostly the outcome of massive screening, and only partially intuition-driven; a data-driven approach is prevalent over a model-driven approach. As in computing, a brute-force approach is made possible by a huge technological progress.

[2] Data integration requires the availability of such layer across multiple data sources; it is perhaps the most challenging aspect of big data management.

[3] http://www.informatics-europe.org/ecss/about/past-summits/ecss-2015.

[4] Note the use of the term *data model* as representative of model-driven design.

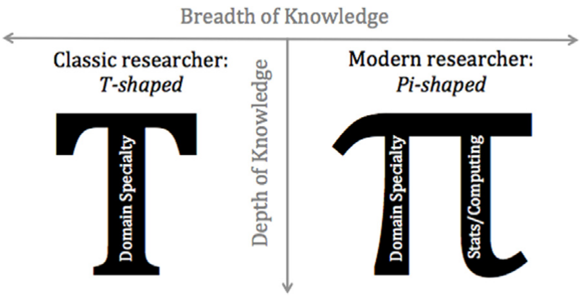
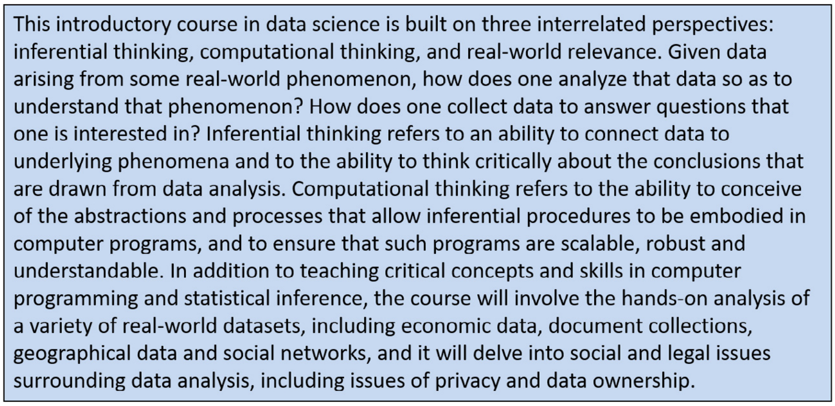


Fig. 1. Representations of T vs. Pi-shaped education.



This introductory course in data science is built on three interrelated perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon? How does one collect data to answer questions that one is interested in? Inferential thinking refers to an ability to connect data to underlying phenomena and to the ability to think critically about the conclusions that are drawn from data analysis. Computational thinking refers to the ability to conceive of the abstractions and processes that allow inferential procedures to be embodied in computer programs, and to ensure that such programs are scalable, robust and understandable. In addition to teaching critical concepts and skills in computer programming and statistical inference, the course will involve the hands-on analysis of a variety of real-world datasets, including economic data, document collections, geographical data and social networks, and it will delve into social and legal issues surrounding data analysis, including issues of privacy and data ownership.

Fig. 2. Syllabus of the undergraduate course Foundations of Data Science, Berkeley University (2015 edition).

4. A new paradigm: data science

Professional life requires the ability to draw information from large datasets, using a broader set of skills. Modern education is advocating the evolution from T-shaped to Pi-shaped models of knowledge.

When student's education supports a *T-shaped approach*, the student masters both domain specialization (on the vertical axis) and cross-disciplinary competences (on the horizontal axis). Such model is discussed in a book (Banerjee, 2015) which I edited with Benny Banerjee, from the Stanford Design School. The so-called horizontal (soft) skills include: how to speak in public, how to actively participate to teams and excite a new empowered form of leadership, how to make decisions, how to approach and organize projects with appropriate methods, enhance creativity. Politecnico di Milano and Torino, the two largest school of engineering in Italy, are engaged since 2004 in *Alta Scuola Politecnica* fostering the T-shaped approach.

A new emerging model, denoted as *Pi-shaped*, adds another vertical competence, relative to the statistical and computational abilities which are required in order to deal with the analysis of (big) data (see Fig. 1). Such competence is useful in all fields, including humanities: think to the relevance of a correct data science approach in schools of journalism, whose students should be trained to understand not only the factual truth of reported news, but also their statistical relevance.

Besides specific curricula in data science, it is remarkable to note that data science is being offered at undergraduate level and traversal to all majors. An example is offered by Berkeley University: *Foundations of Data Science*, a course originally offered by the departments of statistics and of computer science, is now offered by the Data Science Division, created in December 2016, with a focus on undergraduate education. The course was collaboratively taught by two professors from the Statistics and EECS departments in parallel/alternative to basic courses in statistics or computer science, and attended in 2016 by approximately 500 students from all disciplines; an interesting evaluation of this course edition is available (Berkeley, 2016). A similar trend is now occurring at Harvard, where *Data Science 1: Introduction to data science* was opened last year to all the disciplines, and was attended by about 300 students.

Another important aspect of the current shift in data science has to do with pragmatics: data science is applied to solving concrete problems, as clearly described by the syllabus of the Berkeley course *Foundations of Data Science* (see Fig. 2).

Many graduate data science programs were started recently. Among them, the Harvard Faculty of Arts and Sciences launched a new Master of Science (SM) degree in Data Science, under the joint academic leadership of the Computer Science and Statistics faculties. The program is administered through the Institute for Applied Computational Science (IACS) of the School of Engineering and Applied Sciences (SEAS). The program is described as follows: "Data Science lies at the intersection of statistical methodology, computational science, and a wide range of application domains. The program will offer strong preparation in statistical modeling, machine learning, optimization, management and analysis of massive data sets, and data acquisition. The program will also focus on topics such as reproducible data analysis, collaborative problem solving, visualization and communication, and security and ethical issues that arise in data science". Politecnico di Milano participates to this program through Data-Shack, a highly experimental program where students from Politecnico and Harvard participate to the Harvard Capstone Project Course through interdisciplinary project groups of students from both universities. [5]

From a pedagogical perspective, it is interesting to note that at Berkeley, in June 2017 "more than 30 faculty and instructors across a range of disciplines from English to Sociology to Neuroscience to Physics took a week out of their summer to explore how to incorporate tools developed for data science into their own teaching".[6]

5. On the role of statistics and computer science

It is clear from the previous section that joining statistics and computer science would provide high opportunities and responsibility towards education of students from all disciplines, starting at the undergraduate level, including humanities.

Computer science contributes the data science infrastructure, as it deals with the aspects of: data acquisition, representation, cleaning and integration; information/knowledge modeling and semantic and context/based enrichment; information retrieval, data mining, knowledge discovery; artificial intelligence methods and machine learning; data visualization and exploration. With respect to all these aspects, computer science is not a mere technology provider, as it brings about suitable abstractions and foundations. Today's attention on big data is primarily the outcome of progress in computer science.

On the other hand, statistics provides the pillars for data analysis and interpretation; we need to solve the new challenges listed in Secchi (2018) by Piercesare Secchi (among them: parametric and non-parametric statistics, dimensional reduction, distributed inference, divide et impera approaches, data integration) in order to meet the new requirements of big data and data science. It is interesting to note that data integration is seen mostly as a problem of semantic interpretation by the computer science community and as inferring a joint distribution of random elements extracted from distinct databases by the statistics community — clearly the two problems overlap and both need to be solved in approaching many global data science problems.

I agree with Piercesare's conclusions: big data do not speak for themselves, correlation is not enough, hence (Anderson, 2008) should be taken just as a provocative statement.[7] A rigorous approach to data science requires a deep knowledge of theory and methods from statistics.

Thus, I advocate that computer science and statistics should embrace the data science revolution together. We live an era in which computers can be considered a commodity, although necessary for any human activity; computer programming should be mastered by all students of scientific faculties — although coding is now regarded as a particular form of reasoning, to be learn at schools. Data science brings about a traversal education, applicable to all sciences: to business management when we consider the economic value of information; to sociology when we consider its social value; to biology and medicine when we consider life sciences, and so on. The primary outcome of data science is value creation, and the beneficiary is the whole mankind.

6. Conclusions

I like to conclude with a personal note. Ten years of work in the Alta Scuola Politecnica have convinced me that multidisciplinarity and interdisciplinarity are key to success in modern science. The work embraced with designers has given me more evidence; and my current Advanced ERC Grant on *Data-Driven Genomic Computing* is another significant step in this direction. In these days, I am deeply interested in life sciences, studying aspects such as the tridimensional organization of DNA, or how gene expression and mutations may cause tumors, or how new drugs can be obtained by joining multiple molecules in polypharmacological networks; at the same time, I use my computer science background to build next-generation infrastructures for genomic computing. Milano will soon host the Human Technopole (https://www.htechnopole.it/en/), a research infrastructure for life science established by the Italian Government which will "achieve his mission using genomics, the analysis of increasingly large data sets, and new diagnostics techniques". This is the ideal battlefield for taking advantage of data science.

Acknowledgments

Most of the thoughts, issues and perspectives reported in the previous pages are the result of fruitful discussions with my fellows of the Data Science group at DEIB (http://datascience.deib.polimi.it/); specifically, I am grateful to Marco Brambilla, Emanuele della Valle and Florian Daniel for their suggestions on a draft of this paper. This research is supported by the Advanced ERC Grant No. 693174, Data-Driven Genomic Computing (GeCo).

[5] https://iacs.seas.harvard.edu/co-curricular-programming and http://datashack.deib.polimi.it/.

[6] http://data.berkeley.edu/.

[7] Although the statistics community is not without sin, recall that "if you torture the data long enough, it will confess to anything" (Huff, 1954).

References

Anderson, C., 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired, https://www.wired.com/2008/06/pb-theory/ (access: 5.17.17).

Banerjee, B., 2015. In: Banerjee, B., Ceri, S. (Eds.), Building Innovation Leaders: A Global Perspective. In: Springer series on Understanding Innovation, Springer-Verlag.

Bayer, 2015. Bayer, From Molecules to Medicine, http://pharma.bayer.com/en/research-and-development/technologies/small-and-large-molecules/index.php (access: 1.5.15).

Berkeley, 2016. Berkeley's Undergraduate Data Science Curriculum: Year 1 Pedagogical Overview, http://data.berkeley.edu/sites/default/files/dsepsummaryreportyear1.pdf (access: 5.17.17).

Breiman, L., 2001. Statistical modeling: the two cultures. Statist. Sci. 6 (3), 199–231.

Ceri, S., 2017. On the big impact of big computer science. In: Werthner, F., van Harmelen, H. (Eds.), Informatics in the Future. Springer Verlag.

Codd, E.T., 1970. A relational model of data for large shared data banks. Commun. ACM 13, 6.

Fourth Paradigm, 2009. In: Hey, T., Tansley, S., Tolle, K. (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

Huff, D., 1954. How to Lie with Statistics. W. W. Norton & Company.

Secchi, P., 2018. On the role of statistics in the era of big data: a call for a debate. Statist. Probab. Lett. 136, 10–14. Special Issue on "The role of Statistics in the era of Big Data".

Vardi, M., 2015. From Model-Driven Computer Science to Data-Driven Computer Science and Back, slides presented at EECS 2015. http://www.informatics-europe.org/ecss/about/past-summits/ecss-2015/ (access:5.17.17).

Wirth, N., 1975. Algorithms + Data Structures = Programs. In: Prentice-Hall Series in Automatic Computation.