Chapter 2
On the Big Impact of “Big Computer Science”

Stefano Ceri

Abstract  Big science is bringing unprecedented progress in many fundamental fields, such as biology and medicine. While progress cannot be questioned, when looking at the foundations and models of big science one wonders if such new approach is in contrast with critical thinking and model-driven scientific methods—which has shaped for decades higher education in science, including computer science. In this paper, after a discussion on how big science is shaping drug discovery and modern biology, I trace the start of this new interest on data science as outcome of the “fourth paradigm” and I discuss how CS education is changing due to the impact of big science, and question where/how it will be hosted within universities and if Academia is a good fit for data scientists.

2.1 Introduction

“Big science” is a popular, perhaps abused term. It indicates not only that massive amounts of data are nowadays available in an unprecedented way, but also that the approach to science is shifting from being model-driven (where modeling and abstractions are foundational and data are just supportive of given initial hypotheses) to data-driven (where models can be directly extracted from data). A new figure of “data scientists” is emerging (and very much requested by the labor market), with strong computational and statistical background, that is particularly good at extracting domain-specific knowledge from big datasets.

The strength of a data driven approach is evident if one looks at the main players of the Computer Science Industry, such as Google and Facebook, whose business model relies on making profits out of data, the so-called “socially produced content”. Google is a “data” company, as (a) its computational approach is data driven and heavily based in statistics and math (see the original competitive advantage upon other search companies based upon the “PageRank” algorithm, invented by the co-founders); and (b) it owns no data itself, but it converts user data

S. Ceri
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio, 34/5, 20133 Milano, Italy
e-mail: stefano.ceri@polimi.it

© The Author(s) 2017
H. Werthner, F. van Harmelen (eds.), Informatics in the Future,
DOI 10.1007/978-3-319-55735-9_2
into information that is then provided essentially to the same users (in a sense, Google managed to convert steel into gold, as in many legends of ancient times).

In this paper, I start with a discussion of prototypical examples of big science in the medical and biological domains, which serves the purpose of introducing the data-driven approach in contrast to the model-driven approach; then I trace the start of the emphasis on “big data” to the so called “fourth paradigm” and recall some illuminating principles by Jim Gray, who first inspired the fourth paradigm revolution. Then I turn to discussing how big science should be taught, and whether big science experts can find a good fit in the current academic value system. I conclude with some considerations about the general relevance of multi-disciplinary and a problem-driven approach in computer science curricula, which go beyond the specific discussion on “big data”, although they fit well with the main theme of the paper.

2.2 Where “Big Science” Is Really Big: Pharma Industry and Genomics

The pharmaceutical industry has been impacted from a “big data approach” since a long time. We all remember the big discoveries in medicine in the last century due to scientists who had enlightening intuitions, driven by their acute observations; but now, the process of drug discovery is a highly standardized one, that has been described, e.g., by Bayer (2015), as a fixed sequence of 10 steps. In particular, the first two steps, called “DNA testing for target discovery” and “high-throughput screening”, consist of discovering the proteins that might be playing 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—known as lead candidates—which could be a suitable starting point for a new active ingredient. In this step, high throughput screening (HTS) applies an in-house compound library (currently containing over three million chemical substances) for suitable lead candidates. Robots fill thousands of microtiter plates on which up to 1536 tests can be performed simultaneously (see Fig. 2.1). In essence, drug discovery is now in most cases the outcome of massive screening, rather than being intuition-driven.

The drug discovery method described above seems to question the classic scientific method, built around testable hypotheses, or “models”. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. Scientists are trained to recognize that correlation is not causation, that no conclusion should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence), that “data without a model is just noise.” However, the brute-force approach seems to require no a-priori model, or at least requires just a generic knowledge about the generic processes leading to the production of proteins and the mechanisms for compound screening. Generally speaking, faced with massive data, the classic approach to science—hypothesize, model, test—is
becoming obsolete. Data availability in the range of petabytes allows Chris Anderson, editor in chief of Wired, to say that: “Correlation is enough: 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 summary: is correlation without causation good enough when data sizes are big enough? This dilemma is current among computer scientists.

Another example of “big” approach is the construction of massive repositories of genomic data. Next Generation Sequencing is a technology for reading the DNA which is producing huge amounts of DNA sequences—at an exponentially decreasing cost and exponentially increasing processing speed, at a much faster pace than the Moore law for processors. Several worldwide consortia have been created in order to accumulate sequence data and make them available to the research community, through the combined efforts of hundreds of laboratories in the world. Among them, 1000 Genomes: Deep Catalog of Human Genetic Variation (1000 Genomes Project Consortium et al. 2010), with the goal of finding most of the genetic variants that have frequencies of at least 1% in the populations; the Cancer Genome Atlas (Weinstein et al. 2013), which presents a comprehensive genomic characterization and analysis of several cancer types; the 100,000 Genomes Project, a UK project which will sequence 100,000 genomes from around 70,000 people, chosen amongst NHS patients with a rare disease or with cancer plus their families; and ENCODE: Encyclopedia of DNA Element (ENCODE Project Consortium 2012), with the goal to

---

1http://www.genomicsengland.co.uk/
build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

In recent years, I have become strongly interested in genomic data management. Figure 2.2, extracted from Paradigm 4 (2015), indicates the current state of genomic computing, where a lot of progress has been done in the so called “primary analysis” (essentially the reading of DNA) and “secondary analysis” (essentially, the alignment of DNA reads to a reference genome for each species, and the discovery of DNA features, such as mutations and peaks of expressions); my interest is on “tertiary analysis”, i.e. making sense of datasets resulting from multiple, heterogeneous experiments. With my research team, I am currently working on a genomic data model and query language supporting querying of heterogeneous genomic datasets on the cloud, with a tight integration between genomic data (relative to the regions of the whole genome) and metadata (relative to the experiment preparation and sampling, including possibly the patient phenotype). In the near future, we expect genome sequencing to be a key to understanding many pathologies, and we forecast bridging genomics to personalized healthcare; this could be among the most interesting “big data” problems of mankind.

2.3 Where It All Started: “Fourth Paradigm”

The emphasis on “big data” in computer science can be traced back to the fundamental contribution of the “Fourth Paradigm” book (Hey et al. 2009), and to the legacy of Jim Gray (Fig. 2.3). The book’s preface presents a historical view of computer science as 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 properly address content within a collection of information, e.g., by saying: “I want all the genes that have this property” or “I want all of the stars that have this property” or “I want all of the galaxies that have this property.” Once a schema is well-defined, data can be indexed, aggregated, searched in parallel, and it is easier to build both ad-hoc queries and generic visualization tools. If instead big data are just a “bunch of files”, it is not even possible to see the concept of gene, or star, or galaxy; and 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 organization which should be separate from data content, in contrast to the current trend of just using data without any concern upon understanding their structure and quality.

The legacy of Jim Gray is huge, in this as in many other fields. I still remember when, about ten years ago, he was explaining to me his joint work with astronomers while at the same time I was looking into classic computer science problems, and I couldn’t quite understand his enthusiasm. A posteriori, I was still trapped into a disciplinary silo, he had already moved towards interdisciplinary data science.

Scientific research more and more dependent on the careful analysis of large datasets, requiring a broad skill-set: scientists must be experts not only in their own domain, but in statistics, computing, algorithm building, and software design.
In order to understand the skills that we expect from a “next-generation data scientists”, we refer to two popular diagrams, illustrated in Fig. 2.4. The first one shows the evolution from T-shaped to Pi-shaped models of knowledge: with the former model, a researcher had to provide both support domain specialization (on the vertical axis) with horizontal knowledge [i.e. general and cross-disciplinary competences: how to speak in public, participate to teams and become team leader, make decisions, approach and organize projects with appropriate methods, enhance creativity, see Banerjee and Ceri (2015)]. The new emerging model, denoted as Pi-shaped, adds another vertical competence, relative to the mathematical, statistical, and computational abilities which are required in order to deal with data science.

The second image of Fig. 2.4 illustrates the data science skills as currently available on the professionals market, where a first required ability is concerned with “hacking skills” (according to Drew Conway, “being able to manipulate text files at the command-line, understanding vector operations, thinking algorithmically”); then, appropriate math and statistics methods are needed to extract information from data (e.g., knowing what an ordinary least squares regression is and how to interpret it.) Combining hacking skills to math and statistics “only gets machine learning”, whereas a substantive expertise in the application domain is needed as third dimension “which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods.” Interestingly, Drew Conway denotes those with hacking skills plus substantive expertise as a danger zone, as people in this area may be perfectly capable of extracting and structuring data, but they lack any understanding of what computation means. Fortunately, “it requires near willful ignorance to acquire hacking skills and substantive expertise without also learning some math and statistics along the way. As such, the danger zone is sparsely populated, however, it does not take many to produce a lot of damage.”

Fig. 2.4 Representations for data science research expertise: (a) “T” vs. “Pi-shaped” education and (b) multidisciplinary contributions to data science (from: DrewConway.com, retrieved 25/6/2015)
2.4 Data Science and Academia: Issues and Opportunities

The push towards the creation of a new focus on data sciences in academic programs is very strong. Perhaps the most radical change is occurring at Berkeley University, where a new course Foundations of Data Science, jointly offered by the departments of statistics and of computer science, is currently being experimented with the goal of being offered to all the freshmen students in the near future (Data Sciences @ Berkeley 2015). The course will present key elements of introductory computational and inferential thinking in an integrated fashion, cementing conceptual understanding through direct experience with data (see the syllabus in Fig. 2.5).

Many universities are starting 1-year, intensive programs to create data scientists. Among them, Harvard University offers a 1-year Master of Science in Computational Science and Engineering (CSE) (Fig. 2.6) targeted towards the construction of data scientists' skills which seem to be inspired both by T and Pi-shaped education principles, reported in Fig. 2.4; note the emphasis on communicating across disciplines and collaborate with teams, which are typically regarded as horizontal skills, but note also the emphasis on real-life problems. Other one-year master programs are spreading worldwide; among them, in Italy, three new programs offered by Cefriel (jointly with Politecnico di Milano), by the University of Pisa and of Bologna.

At Politecnico di Milano, we created a “Big Data” track in our regular Master Degree, which includes suitable CS courses (such as Advanced Data Management, Data Mining, Machine Learning), some application-oriented courses (e.g. in Computational Biology) and interdisciplinary contributions from the schools of Mathematics (Applied Statistics) and of Management (Business Intelligence). Small innovative experiments are ongoing, e.g. the Data-Shack program\(^2\) jointly organized

---

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.

---

\(^2\)datashack.deib.polimi.it

---

Fig. 2.5 Syllabus of the undergraduate course Foundations of Data Science, Berkeley University, from: http://databears.berkeley.edu/content/stat-94-cs-94-foundations-data-science
by Harvard’s Institute for Applied Computational Science and by our Master schools of Computer Science and of Design, where students are exposed to hands-on data science problems.

The growing interest in academic courses and programs in data science seems not to be matched by offering of academic positions to the data science; according to Jake Vanderplas,3 time spent developing high-quality reusable software for solving concrete problems translates to less time writing and publishing, which under the current system translates to little hope for academic career advancement. In particular, it is argued that industry may be a better fit for data scientist, due to a number of factors: salary, stability, opportunity for advancement, respect of peers, freedom from the burden of publishing and teaching, given that also in industry there is the opportunity to work on interesting projects and to contribute to open source software. In his blog, Jake Vanderplas calls for an increase of the pay of post-doctoral scientific research positions—so as to become more competitive with industry—and for a change of the Academic value system to defend the data scientists career, by pushing for a new standard for tenure-track evaluation criteria which emphasizes the importance of producing reproducible software and rewards the development of open, cross-disciplinary scientific software tools.

Another discussion (from 1) concerns where data science should be housed by Academia; five solutions are considered.

1. Data science is simply a label for a new skill-set, and shouldn’t be separated from the departments in which it is useful. Departments across the university should simply incorporate relevant data science techniques into their normal curriculum.

---

3https://jakevdp.github.io/blog/2014/08/22/hacking-academia/
2. Data science might be organized as a consulting service, in a similar way to the IT infrastructures. We can’t expect every scientist to be fluent in the statistical and computational methods required to work with large datasets, so these tasks would be outsourced to data science experts.

3. Data science could be seen as an applied branch of computer science or of statistics which should give rise to a separate department, similarly to existing departments of “applied math” and “applied physics”, which distinguish themselves from the non-applied version by employing the techniques of the field within practical rather than theoretical contexts.

4. Data science could be similar to an evolution of library science: as digitization is changing the role of libraries on university campuses, focus of a modern Library and Information Science departments could be moved from hosting printed books to data curation.

5. A middle ground to the above approaches may be to organize data science within an interdisciplinary institute, along a cross-departmental organization which is becoming common in Academia.

2.5 Assessment and Conclusions

Computer science is still a growing discipline, that has successfully overcome a decline (e.g. in enrollment) observed about 10 years ago. In these days, most of computer science schools around the words have lots of undergraduate and graduate applicants, typically of very high quality, and the job market offers good jobs to a potentially large mass of students exiting from our education systems, both at undergraduate and graduate level. Computer science schools as they stand could be self-referential, and stand on their models, methods and technologies; but computer science could be even more successful by opening to other fields. The creation of data science curricula within computer science is not only timely (it captures the “taste of time”) but also strategic, as a means for outreaching to other scientific communities, given that data science is intrinsically interdisciplinary.

Compared with other foundational disciplines, such as mathematics, biology, physics and chemistry, computer science has a clear advantage in being immediately applicable to solve concrete problems. Therefore, computer science can be very much problem-driven, both in the case of a broad offer to undergraduate students and in the development of graduate curricula; the push towards data science is making this aspect even more evident. In the past, many openings of computer science to the multidisciplinary dimension from inside computer science have been quite difficult: such disciplines were considering computer scientists not much as “peer scientists”, but rather as “service providers” who could offer their skills in order to solve disciplinary problems of the hosting discipline. A push towards data science will put our community in the position of driving the change towards a true multi-disciplinary approach, where data scientists with a strong
computational background will be recognized as a key success factor for solving data-driven problems.

In conclusion, an educational model of big science is emerging, combining computational and inferential thinking in an integrated fashion, cementing conceptual understanding through direct experience with data (Data Sciences @ Berkeley 2015). This approach stems from a new emphasis on “big data”, which can be traced back to the Fourth Paradigm book (Hey et al. 2009) and which is becoming more and more relevant with the growth and worldwide organization of big data repositories; I offered some examples of them in the biology sector. One big question which is left somehow opened is whether computing should be driven by models rather than by data; according to some opinion-makers and colleagues, traditional computer science models should be used when/if needed but no longer be the key foundational aspect of problem solving.

References

1000 Genomes Project Consortium et al (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073
Banerjee B, Ceri S (eds) (2015) Building innovation leaders: a global perspective. Understanding innovation. Springer, Cham
Bayer (2015) From molecules to medicine. http://pharma.bayer.com/en/research-and-develop ment/technologies/small-and-large-molecules/index.php. Retrieved 15 July 2015
Data Sciences @ Berkeley (2015) The undergraduate experience. http://ls.berkeley.edu/files/file/ Data%20Sciences%20Education%20Sketch%201.2_0.pdf. Retrieved 11 Nov 2015
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74
Hey T, Tansley S, Tolle K (eds) (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond
Paradigm 4 (2015) Accelerating bioinformatics research with new software for big data knowledge. White paper. http://www.paradigm4.com/, April 2015
Weinstein JN et al (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45(10):1113–1120

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.