Ubiquity Symposium

Big Data

Big Data: Business, Technology, Education, and Science

By Jeffrey Johnson, Luca Tesei, Marco Piangerelli, Emanuela Merelli, Riccardo Paci, Nenad Stojanovic, Paulo Leitão, José Barbosa, Marco Amador

Editor's Introduction

Transforming the latent value of big data into real value requires the great human intelligence and application of human-data scientists. Data scientists are expected to have a wide range of technical skills alongside being passionate self-directed people who are able to work easily with others and deliver high quality outputs under pressure. There are hundreds of university, commercial, and online courses in data science and related topics. Apart from people with breadth and depth of knowledge and experience in data science, we identify a new educational path to train "bridge persons" who combine knowledge of an organization's business with sufficient knowledge and understanding of data science to "bridge" between non-technical people in the business with highly skilled data scientists who add value to the business. The increasing proliferation of big data and the great advances made in data science do not herald in an era where all problems can be solved by deep learning and artificial intelligence. Although data science opens up many commercial and social opportunities, data science must complement other science in the search for new theory and methods to understand and manage our complex world.

Ubiquity Symposium

Big Data

Big Data: Business, Technology, Education, and Science

By Jeffrey Johnson, Luca Tesei, Marco Piangerelli, Emanuela Merelli, Riccardo Paci, Nenad Stojanovic, Paulo Leitão, José Barbosa, Marco Amador

The emergence of big data underlies one of the major global and local socio-political-economic revolutions of the twenty-first century. However, the important lesson is not that the data are big, but that all data produced by humans and nature are becoming digital [1]. Today all human activity leaves a digital trace that, by accident or design, can contain extremely valuable latent information.

A huge new technology-enabled business called data science is sweeping the globe to address practical problems and exploit the immense latent commercial and social value of digital data. Data science combines information technology and computational techniques with individual and collective human intelligence.

This article considers data science from its applications in the private and public sectors, where a strong demand for qualified practitioners has stimulated the creation of new educational programs by universities and commercial organizations. Job advertisements for data scientist demand not only technical skills and knowledge, but also highlight soft skills including personal drive, the ability to learn quickly in a self-directed way and, above all, interpersonal skills and teamwork. Data science lies at the intersection of business applications, technology and computational problem solving techniques, and individual and collective human problem-solving intelligence.

There are many practical problems that data science can solve today, but some require new science to lay the foundations for future solutions. The fast moving commercial world of data science and engineering will have an increasing impact on the social world in the short term of

a decade or so, but there are many areas where the business of data science will stall without new scientific knowledge and understanding.

Data Science in Business and Administration

The information implicit in data has to be made explicit by synthetic processes that require human intelligence and ingenuity. The value in the data lies in it being able to answer questions such as:

- Will this person buy our product?
- Is this drug effective?
- Is this person engaged in criminal activity?
- Will these shares increase or decrease in value?
- Will this machine fail?
- Should we invest in this proposition?
- Will this disease spread to our country?
- How many votes will result from this political campaign?
- Do we have enough hospital beds for this emergency?

In some of these examples the monetary or social value of the answers to the questions is enormous and this generates huge incentives to answer them.

Data science is the combination of scientific knowledge, scientific methods, and computational processes that can answer the kind of questions above and, more generally, in combination with individual and collective human intelligence extract the latent value from the ever-increasing flux of digital data for a wide variety of applications. The new and relatively highly paid profession of data scientist reflects the demand by companies and organizations to increase their revenue and reduce their costs by extracting relevant information from data, be

it big or small, distributed or local, homogeneous or heterogeneous, and owned in-house or paid for externally.

Data Science as Technique and Technology

Consider the general case of an organization that requires information that is not readily available, for example an answer to one of the questions above—where the relevant data are distributed over its intranet and the internet, are heterogeneous, are partly or wholly owned by others, whose existence is not known *a priori*, and where considerable modeling, processing, and analysis may be required to extract the information required. This will involve the following iterative data lifecycle processes:

a) searching to find relevant data
b) data preparation
c) data analysis
d) data visualization

Searching to find relevant data is a fundamental task in data science. The digital nature of big data enables it to be searched by semi-automatic or automatic means. Most of us initiate semi-automatic searches every day using search engines, trying to find information relevant to the business at hand and often experiencing the unpredictability of the search process: You enter a search term, the search engine provides a selection of links, you choose to follow one or more using your human intelligence to synthesize what you observe, possibly following new links or entering new search terms, and so on until you find what you think you want or give up. More automated searches can be performed by web crawlers and the programs that control them. Data science has many search tools but, ultimately, the quality and value of the data depends on human supervision and engagement in the search process.

Data preparation is a necessary and significant task since data discovered on the internet can be messy and need cleaning. The first problem to be addressed is that data is heterogeneous. This is the case for free-formatted text, and can be true of numerical and tabulated data, which may be formatted differently or based on different classification schemes. Also it is subject to a

variety of errors, including it being incomplete and inconsistent, e.g. tables may have missing or contradictory values, companies may publish the minimum about their activities, while deliberate falsehoods propagate such as the malicious lies of trolls and obfuscation of "fake news." Data science has many tools for cleaning data but, ultimately, the quality of the data and the information synthesized from it will depend on human supervision of and engagement in the cleaning process, transforming it into (more) homogeneous forms suitable for analysis.

Data analysis is the process of combining and processing data to transform it into information. It requires the analyst has some form of model of the system; this may be as simple as a mathematical equation giving some future value of a number or as complicated as a land-use transportation model used in city planning. The many kinds of model that could be used are as follows:

- *Network and systems models* investigate future behaviors through interactions. Network analysis enables statistics to be abstracted from large data sets including degree distributions, the analysis of paths between parts, the emergence of giant components, and many structural relationships.
- *Agent based models* represent individuals as agents with their own, possibly changing, attributes and behaviors that govern their interactions with other agents. Agents may be individual people or inanimate objects such as houses, vehicles, and organizations. Examples include the spread of diseases as agents interact locally through everyday contact and globally through transportation networks.
- *Statistical modeling* involves analyzing populations and stabling relationships between data, e.g. through regression analysis and the establishment of correlations.
- *Time series models* have one or more values analyzed through time, and future values in time are forecast on the basis of previous values in time. For example, many human activities correspond to the calendar enabling forecasts to be made of the number of people booking holidays, the number of people requiring hospital treatment, or the amount of beer that will be consumed at any given time in the future.

- *Classification models* assign individuals to classes or clusters to forecast behaviors, e.g. personalized advertisements appearing on web sites, or the diagnosis of a medical condition and forecasts of the most appropriate treatment.
- *Neural networks models* are based on a biological analogy with networks of neurons learning from data without being explicitly programmed, typically performing classifications after being trained on data—e.g. diagnosis based on clinical measurements—and instantaneous based on a live stream of market data.
- *Artificial intelligence* covers a wide range of algorithms designed to emulate human reasoning, including programs that find logic-based patterns in data or use reasoning to make forecasts.
- *Topological models* are becoming better known for their ability to find significant topological structure in large data sets [3], for example persistent entropy [4].
- *Deep learning models* combine all techniques, neural networks, and artificial intelligence to find structure in (often massive) data sets. In this scenario, there may be no top-down model of a particular phenomenon and analyzing data attempts bottom-up modeling based on the only thing available: data. Most of the time, data are generated from complex social, economic, technical, and environmental systems and data analysis provides a way to discover hidden paths and emergent behaviors.
- *Predictive modeling* combines many of the approaches above to forecast possible future behaviors and events.

This list is not exhaustive and there are intersections between many of its items. Also the data search-preparation-analysis-visualization life cycle sketched above is not sequential but the stages interact with feedback and iterations. This discussion illustrates the fact that there is no single procedure or algorithm that can convert data into useful information or knowledge, and abstracting information from data is far from an algorithmic or automatic process—it requires the expertise and professional knowledge of data scientists.

Data analysis often requires bespoke programming to combine data sets, to combine software from the many libraries implementing the modeling approaches described above, and where

necessary to write original code to implement new or system-specific models for a given application.

Visual analytics is the science of presenting information in ways that people unfamiliar with data science can understand. Visualization includes presenting data in the usual form of graphs and charts in two and three dimensions, but also includes animations and interactive graphics that allow users to interact with the data and information in ways that they find natural an intuitive. Visual analytics can go beyond this, helping people to find new patterns in data that had not been observed before.

Data Science Education

The interest in data analysis reflects an increasing awareness by companies and organizations that recognize to maintain or improve their sustainability they must extract new information and knowledge from existing and new forms of digital data. In this, the education of data scientists plays a crucial role. The monetary or social value of data analytics can be enormous and this generates huge incentives for the creation of data-science capabilities, and an increasing demand for people with data-science skills.

A recent study of data scientist recruitment and education showed in the U.K. there are hundreds of jobs being advertised across a wide range of applications [2]:

- Manufacturing: modeling and forecasting, modeling production, and supply chains
- Retail: marketing, modeling and forecasting sales
- FinTech: banking and financial services
- Supermarkets: marketing, modeling and forecasting sales
- Airlines: marketing, modeling and forecasting sales
- Engineering: control, system monitoring, repair and lifecycle management
- Government, e.g. the library of the House of Commons has data science capabilities
- Marketing: designing campaigns, analyzing data on sales, footfall, web clicks etc.

- Education: analyzing data on web clicks, study times, marks gained, study paths
- Scientific: analyzing large quantities of multidimensional numerical data
- Medicine: classification for diagnosis and treatments, statistical analysis
- City planning: modelling and mapping to forecast land use, transport, housing, and services

The jobs advertised required a variety of technical knowledge, skills, and experience, including:

- technical issues, e.g. setting up virtual machines in the cloud with generic tools
- programming: use of computer languages, e.g. Python, JavaScript, C, C++. C#
- data structures and algorithms, distributed, data lakes, trees, graphs, search
- databases and query languages: SQL, noSQL, distributed databases
- modeling: types of model, e.g. network models, systems models, AI
- statistics: statistical theory and packages, e.g. SPSS, R
- web design: user interface design, HTML, CSS, front- and back-end programming
- visualization: using visualization tools, graphics, maps GIS

The jobs also specified a range of soft skills and attributes including:

- team working
- passion/proactive/self-starter
- analytic and problem solver
- work with business/customers
- curious/hacker/open/independent

- work in a fast-paced environment
- attention to details/quality work
- developing one's skills, leadership and mentoring
- creative/entrepreneurial

These underline the importance of the human element in data science. Data scientists are expected to have a wide range of technical skills alongside being passionate self-directed people who are able to work easily with others and deliver high-quality outputs under pressure.

In the U.K. there are more than 100 university masters courses (EQF[1] Level 7 [5])) and about a dozen bachelors' courses in data science (EQF Level 6). The position for doctoral research (Ph.D., EQF Level 8) in the U.K. is more complicated since at this level the title may or may not indicate data science. It is estimated that each year in the U.K. about 500 to 1,000 people are trained to doctoral level in data science and related areas.

Apart from formal university education, there are many boot camps in the U.K. offering intense data-science courses over periods of one to three months. Furthermore there are many MOOCs on various aspects of data science, e.g. Python, Jupyter Notebooks and R. The various platforms for e-learning include hand-crafted websites; Moodle-based specialist sites, such as that used by the Open University; and commercial or not-for-profit MOOC platforms, such as U.S.-based EdX and Coursera and FutureLearn, which is based in the U.K. Typically these courses are free or have low fees and last four to six weeks with about six study hours per week.

The European Erasmus Da.Re project has identified a new educational path to train bridge persons, i.e. persons who combine knowledge of an organization's business with sufficient knowledge and understanding of data science to "bridge" between non-technical people in the business with highly skilled data scientists able to add value to the business [2]. For example, an advertising company with its in-house data science department designed its data analytics after a campaign was designed. A bridge person working within the marketing team could have used

[1] European Qualifications Framework.

their knowledge of data analytics to advise how the campaign could be adapted to produce the most useful data before launching the campaign. In other cases a bridge person in a company could inform senior managers of the potential gains of data analytics and interface the company to external consultants able to offer the required services [2].

The Limits of Science and the Limitations of Data Science

It is tempting to believe the increasing proliferation of big data and the great advances made in data science herald in an era where all problems can be solved. As discussed above, data science is a combination of technical systems and human ingenuity that can give astonishing insights into human behavior. Although these insights open up many opportunistic commercial and social opportunities, commercial data science cannot answer scientific questions such as those proposed in 2013 by Hayley Birch, Mun Keat Looi, and Colin Stuart [6]:

1. What is the universe made of?
2. How did life begin?
3. Are we alone in the universe?
4. What makes us human?
5. What is consciousness?
6. Why do we dream?
7. Why is there stuff?
8. Are there other universes?
9. Where do we put all the carbon?
10. How do we get more energy from the sun?
11. What's so weird about prime numbers?
12. How do we beat bacteria?

13. Can computers keep getting faster?
14. Will we ever cure cancer?
15. When can I have a robot butler?
16. What's at the bottom of the ocean?
17. What's at the bottom of a black hole?
18. Can we live forever?
19. How do we solve the population problem?
20. Is time travel possible?

Seen this way, data science is at best just a small part of science. This list, which could contain many other open questions, shows our limited knowledge of the cosmos, of technology, of biology, and of ourselves.

Conclusion

Although there are many examples of big data being the source of great commercial and social value, transforming that latent value into real value requires great human intelligence and application of human-data scientists. Alongside this a science of data is emerging, but this brings to bear existing scientific knowledge and analytic techniques rather than establishing a new science. Data scientists are expected to have a wide range of technical and soft skills, and there are many courses in data science to provide training. Complementing people with breadth and depth of knowledge and experience in data science, a new educational path is required to train bridge persons able to connect organizations with data science services. Although they mark a milestone in the information revolution of the 20th and 21st centuries, big data and data science cannot answer all questions and they have their limitations.

References

[1] Johnson, J. H., Denning, P., Delic, K., and Sousa-Rodrigues, D. Prologue: Big data, digitization and social change. *Ubiquity*, December 2017.

[2] Cristalli, C., Gatto, M., Isidori, D., Paci, R., Merelli, E., Piangerelli, M., Tesei, L., Johnson, J. H., Barbosa, J., Leitão, P., Piras, F., Kavšek, B., Romero, C. J., Amador, M., Borlinić, J., Horvat, B., and Stojanovic, N. New Big Data Initiatives - Towards a data driven mindset. Da.Re Intellectual Output 1. August 2017.

[3] Merelli, E., Rucco, M., Sloot, P., and Tesei, L., Topological characterization of complex systems: Using persistent entropy. *Entropy* 17, 10 (2015), 6872-6892; doi: 10.3390/e17106872.

[4] Rasetti, M., and E. Merelli. The topological field theory of data: A program towards a novel strategy for data mining through data language. *Journal of Physics: Conference Series* 626, 1 (2015).

[5] The European Commission. Descriptors defining levels in the European Qualifications Framework (EQF). 2017.

[6] Birch, H., Loo, M. K., and Stuart, C, *The Big Questions in Science: The Quest to Solve the Great Unknowns*. Andre Deutsch, London, 2014.

Biographies

Jeffrey Johnson, Ph.D. is Professor of Complexity and Design at the Open University in the UK and Deputy President of the UNESCO UniTwin Complex Systems Digital Campus.

Luca Tesei, Ph.D. Is Professor of Computing at the University of Camerino in Italy.

Marco Piangerelli, Ph.D. is a postdoctoral researcher at the University of Camerino in Italy.

Emanuela Merelli, Ph.D. is Professor of Computing at the University of Camerino in Italy.

Riccardo Paci is Innovation Fund Manager at Loccioni Group in Italy.

Nenad Stojanovic, Ph.D. is CEO of Nissatech in Serbia.

Paulo Leitão, Ph.D. is Professor in the Department of Electrical Engineering of the Polytechnic Institute of Bragança in Portugal.

José Barbosa, Ph.D. is a researcher in the Department of Electrical Engineering of the

Marco Amador, Ph.D. is Senior Software Engineer, CTO and Managing Partner of Maisis Information Systems in Portugal.

DOI: 10.1145/3158350