Data science: an emerging discipline

Ixent Galpin

PhD MSc BA, Profesor Titular, Universidad de Bogotá-Jorge Tadeo Lozano.

E-mail: ixent@utadeo.edu.co

ABSTRACT

The role of data scientist has been described as the "sexiest job of the 21st Century". While possibly there is a degree of hype associated with such a claim, there are factors at play such as the unprecedented growth in the amount of data being generated. This paper characterises the already established disciplines which underpin data science, viz., data engineering, statistics, and data mining. Following a characterisation of the previous fields, data science is found to be most closely related to data mining. However, in contrast to data mining, data science promises to operate over datasets that exhibit significant challenges in terms of the four Vs: Volume, Variety, Velocity and Veracity. This paper notes that the current emphasis, both in industry and academia, is on the first three Vs, which pose mainly scientific or technological challenges, rather than Veracity, which is a truly scientific (and arguably a more complex) challenge. Data Science can be seen to have a more ambitious objective than what traditionally data mining has: as a *science*, data science aims to lead to the creation of new theories and knowledge. This paper notes that, ironically, the veracity dimension, which is arguably the closest one relating to this objective, is being neglected. Despite the current media frenzy about data science, the paper concludes that more time is needed to see whether it will emerge as discipline in its own right.

Keywords: Data science, data mining, data engineering, statistics, Big Data.

Introduction

According to the magazine *Harvard Business Review,* the term Data Scientist describes a new role set to become the "sexiest job of the 21st Century" (Davenport & Patil, 2012). While possibly there is a degree of hype associated with such a claim, there are factors at play such as the unprecedented growth in the amount of data being generated. The International Data Corporation (IDC) expert committee has predicted that the digital universe will grow by a factor of 44 over the time period of 2009 to 2020, to over a trillion gigabytes (Gantz & Reinsel, 2013). However, various experts have noted that there is no consensus about what exactly the new discipline, *data science*, entails. According to a *Forbes* magazine article, it lies at the intersection of traditional disciplines such as statistics and computer science (Press, 2013). Indeed, some have questioned whether data science merits the creation of a new discipline as such, given the important commonalities shared with existing disciplines such as data mining and statistics. This paper presents a brief characterisation of the disciplines related to the data science field, and aims to identify the main traits which distinguish them. Finally, it aims to draw some conclusions about what sets data science apart from these already established disciplines.

Data Engineering

The Bulletin for the Technical Committee in Data Engineering of the Institute of Electrical and Electronics Engineers (IEEE) defines the scope of data engineering as "the design, implementation, modeling, theory and application of database systems and their technology" (Friedman, 1998). A similar perspective is given by the webpage of the *International Conference on Data Engineering* (ICDE), which states that the relevant research issues addressed by the conference are "designing, building, managing, and evaluating advanced data systems and applications" (ICDE, 2016). The topics mentioned in the most recent call for papers of this conference are diverse and include "Cloud computing and Database-as-a-Service, Big Data and Data Warehousing [...] Metadata Management [...] In-Memory Database Architecture and Systems, [...] Crowdsourcing [...] Streams and Sensor Networks" and "emerging database issues such as post-disk database architectures, Internet-of-Things support, and vertical applications" (ICDE, 2016).

George Fletcher, of the Technological University of Eindhoven in Holland, describes data engineering as being "an established research area concerned with data intensive systems and algorithms" and "a core discipline of data science" (SIKS, 2015). Another definition is proposed by the University of Aalborg in Denmark who, in its description of its Masters programme in data engineering,

states that this field is concerned with the technologies for data management, conceptual modelling of data and the design of databases, data models, query languages, query processing and optimization, and data indexing (University of Aalborg, 2015). Furthermore, the University of Dundee, who also offers a Masters in this field, explains that the role of data scientist and data engineer are complementary, given that the role of the data engineer is to develop infrastructure to store, manage and analyse the wave of data which is currently being generated (University of Dundee, 2015). This point of view coincides with that expressed by Provost and Fawcett (2013), who also note that data engineering and data processing are critical to support the activities associated with data science, even if they are more general and useful for many other endeavours.

Statistics

Statistics, a branch of mathematics, is the most established of the disciplines being considered. Tukey, a mathematician from the United States of America best known for having invented the Fast Fourier Transform algorithm, describes statistics as a process in which inferences are made from the particular to the general, viz., from a sample to the general population (Tukey, 1962).

Friedman, a statistician from the University of Stanford, notes that there are various fields who had seminal origins in statistics, such as pattern recognition, database management systems, neural nets, machine learning, and graphical models such a Bayesian networks, genetic programming, chemometrics, and data visualization (Friedman, 1998). However, he considers that these fields were subsequently ignored by the statistics community, due mainly to the fact that many statisticians considered that the field of statistics should be limited in scope to mathematical-based probabilistic inference. Friedman acknowledges that this stance may have the advantage that it requires few changes to practices and current statistics academic programmes. However, he notes that such a stance will imply that statisticians will need to resign themselves to the fact that the role of the field of statistics as a protagonist in the information revolution will gradually diminish over time.

As early as 1962, John Tukey already had a wider vision of what statistics as a field should encompass (Tukey, 1962). Tukey considered that statistics should be related to data analysis, and that the field should not be defined as a set of tools but rather, in terms of the problems which is set to solve. Friedman notes that this would lead to significant changes in academic programmes and the everyday practice of statisticians. Importantly, it would involve adopting computation, and topics such as numerical linear algebra, numerical and combinatorial optimization, data structures, algorithm design, hardware architecture and programming methodologies, among others. Friedman argues that if, since the outset, statisticians had adopted computational methodologies as a fundamental statistical tool, other fields related to data would not exist as such, as they would have been subsumed by the field of statistics.

In his article published in 1998 Friedman calls out to his statistician colleagues, pointing out that there exists a crisis in the statistics world, and that possibly it is necessary to change the culture within the statistics community, especially in light of the widespread rejection of computational methods. From one side, he attributes the problem to being one of “marketing”, given that experts of other fields are unable to understand the importance of statistics. However, Friedman acknowledges that this posture would imply that he assumes the problem to be merely cosmetic, and that fundamental changes to paradigms related to statistics are unnecessary (Friedman, 1998).

Data Mining

Alvaro Fernandes from the University of Manchester defines data mining as “A new generation of tools and techniques for automated, intelligent data analysis [whereby] knowledge, rather than data, is what is mined. The goal is to generate models for deployment. DM [Data mining] builds on machine learning and statistics but with more emphasis on large volumes and on

practical deployment of models. The motivation is similar to data warehousing and OLAP [Online Analytic Processing], but [with] different goals, users and processing mode. [It is] Not just for description, [but] prediction too." (Fernandes, 2004).

However, other definitions have been proposed. Fayyad defines *Knowledge Discovery in Databases* (KDD), a term generally accepted to be synonymous with data mining, as being "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, Piatetsky-Shapiro, & Smyth 1996). According to Friedman (1998), data mining "is used to discover patterns and relationships in data, with an emphasis on large observational data bases [sic]. It sits at the common frontiers of several fields including Data Base [sic] Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization".

The field of data mining started to flourish in the 80s. Dhar notes that several books about data mining that date from the 90s describe how machine learning methods can be applied to resolve diverse business problems. In that decade, the was an explosion of tools that leveraged transactional and behavioural data to explain and predict phenomena. An important lesson learnt in the 90s is that machine learning is effective at detecting patterns with relative ease without having to make strong assumptions about the nature of the data and, in particular, the relationships between variables (as is the case, for example, in statistics). However, Dhar claims that a disadvantage of these methods is that they are often unable to distinguish between signal and noise (Dhar, 2013).

Dhar considers that those methods which do not force us to make prior assumptions about the relationship between different variables have great potential. He makes the point that traditionally, we have been trained to believe that new theories have their origin in the human mind, based on previous theory, and subsequently data collection takes place to prove or disprove the validity of the theory. Dhar argues that machine learning inverts the process: one starts with a large dataset, and instead of posing specific queries, expects the computer to identify unexpected hidden patterns in the data (Dhar, 2013).

In cases where datasets have a large volume and are multi-dimensional, Dhar emphasises the difficulty of identifying "good" queries, viz., knowing what to ask in order to obtain a potentially interesting and actionable revelation. For a model to be useful, both from a practical and from a scientific point of view, the patterns found need to have predictive power. Dhar emphasises the importance of applying Occam's razor on the models generated, given that simpler models tend to be preferred to complex ones (Dhar, 2013).

Data Science vs. Data Engineering

Dhar argues that traditional database methods are inadequate for discovering knowledge, given that they are optimized for access and retrieval of data, based on user queries. For example, database queries aim to identify data which satisfy a pattern which has been previously identified by a user. On the other hand, to carry out knowledge discovery, the reverse process takes place. In other words, the aim is to identify the patterns which satisfy the data. Dhar proposes two properties which the patterns must have: they must be interesting and robust. By interesting is meant that the discovered pattern must be unexpected, as if it is obvious the effort invested in discovering the pattern would be in vain. By having an interesting pattern is meant, specifically, that concrete actions can be made based on the knowledge acquired. The other important property, robustness, refers to the fact that the pattern discovered, in order to be useful, must continue to exist in the future. If this is the case, the pattern will don the knowledge discoverer with predictive power.

Data Science vs. Statistics

Dhar identifies various factors that distinguish data science, a field that has existed for less than a decade, and statistics, a discipline that has existed for centuries. Firstly, Dhar notes that the raw material in question (i.e., data) tends to have a volume which is several orders of magnitude

greater that what was previously imagined to be possible. Traditional statistical methods tend to operate over data volumes that can be handled by an everyday calculator. In fact, at present, the rate at which many sources generate data have led many to believe that traditional data models, such as the classical relational model (Codd, 1970) are no longer appropriate, and led new movement to propose data models and new approaches, such as NoSQL (Cattell, 2011) and NewSQL (Stonebraker and Weisberg, 2013).

Large volumes of data have implications with respect to the use of traditional statistical methods, as pointed out by Martin Theus, who makes reference to a limitation of statistics known as the *upscaling problem* (Unwin, Theus and & Hofmann, 2006). An illustrative example is that, given a sufficiently large sample, a $\chi 2$ test always yields a significant result. Put another way, given a sufficiently large dataset, there is no pairwise interaction between two categorical variables which is not significant. Theus claims that the upscaling problem is present in many statistical tests, which leads to the relevance of a large part of statistical theory developed over the last century being questioned in cases of problems which do not have a small dataset size. Theus concludes that the problem has become more urgent recently, due to both advances in database technologies and increases in data volumes.

Friedman considers that statisticians and data mining practitioners could unite to take on the future challenges of data analysis. As well as being critical towards the statistics community, he notes that some paradigms in the data mining community may require adjustments. Friedman cites, for example, the obsession with large volumes of data, which has led to the belief that large quantities of data are necessary in order to carry out analyses which are worthwhile. Friedman points out that sampling methodologies, which have a long tradition in the field of statistics, may be used to improve the accuracy and at the same time reduce computational requirements of those analyses. He further argues that a computationally intensive procedure over a sample of the data may yield a more accurate result than a less sophisticated technique using the full dataset.

According to Dhar, another factor which distinguishes statistics and data science is heterogeneity with the types of data handled. Statistics traditionally handles numeric data. However, nowadays most generated data is non-structured and may comprise text, images, and video. This leads to important challenges in the data engineering field, as it becomes necessary to develop (semi-automated) integration techniques for data which are highly diverse: an open problem in the database research community. Finally, Dhar considers that it is important to highlight the fact that statistics tends to be focussed on describing data and explaining past phenomena. In contrast, data science tends to have a strong emphasis on the prediction of future phenomena.

Data Science

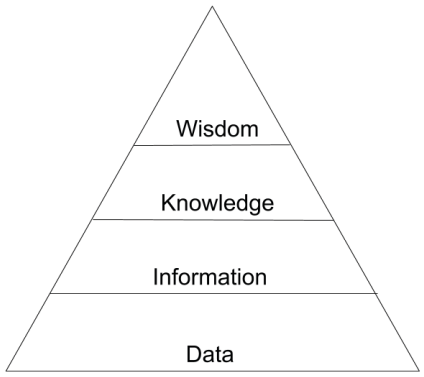
The challenges for data science are considered by many to be a consequence of the *Big Data* phenomenon, a buzzword frequently used to describe datasets which exhibit challenges associated with the four Vs (Volume, Variety, Velocity and Veracity). It is frequently argued that the first V, which relates to the unprecedented *volume* of generated data, implies the need to develop new techniques for the massively parallel processing of data. For example, this view is supported by Friedman (1998), who makes reference to Chuck Dickens, the ex-director of the SLAC (Stanford Linear Accelerator Centre), who stated that every time that computational power increases by a power of ten, it becomes necessary to completely rethink how and what we should compute. Friedman proposes a corollary, stating that the same could be said about data, viz., that every time it increases by an order of magnitude, we should go back to the drawing board and rethink the design of data processing systems (Friedman, 1998). This would imply that the recent increases in the amount of data available would imply the need to completely redesign data processing systems, although not necessarily that a new discipline is required.

The second V, *Variety*, implies the development of integration techniques for very heterogeneous data, from completely non-structured data to video and audio, using techniques such as natural language processing and ontologies.

Velocity implies that it is necessary to analyse data, which often arrives in unpredictable bursts, and also implies the development of non-trivial techniques to manage data streams. However, these first three Vs imply engineering or technological challenges, rather than truly scientific ones. In contrast, *Veracity*, which relates to the creation of knowledge, (i.e., inductively justifiable generalizations over data), poses a genuine scientific challenge. Ironically, despite recent research activity in the area of automated knowledge base construction (e.g., see (Weikum & Theobald, 2010) for a survey), most of the current focus seems to currently be on the first three Vs.

The use of the term *science* is important, given that, as pointed out by Dhar, it implies knowledge acquired through the systematic study of something (Dhar, 2013). Dhar cites a definition proposed by Heilbron, in which he defines science as a systematic tasks in which knowledge is constructed and organized through provable explanations and predictions (Heilbron, 2003). As such, the importance of veracity is apparent in the pyramid of knowledge, also known as the DIKW hierarchy (Rowley, 2007), which represents the layers *Data* (at the base of the pyramid), *Information, Knowledge and Wisdom,* shown in Figure 1. Broadly speaking, data engineering could be viewed as corresponding to the layer at the base of the pyramid, as in this context it refers to techniques that are employed to prepare data for subsequent analysis. Statistics is mainly descriptive, given that it is used to describe and understand past phenomena, and could be seen to correspond to the information layer. Data mining may, based on past phenomena, be used to carry out predictions about future events, which would correspond to the knowledge layer. Data science would aspire to the knowledge/wisdom layers of the pyramid, given that it aims for new theories and knowledge to be created.

***Figure 1**: The knowledge pyramid, also known as the DIKW hierarchy.*



Conclusions

Data science has as foundation the well-established field of statistics, and can be viewed as having two fundamental pillars which complement each other: data engineering and data mining. This paper observes that these three fields have significant intersections and are strongly inter-connected with each other. However, they have some important differences which set them apart. In summary, data engineering is concerned with the design of systems to store, index, query, integrate and transport data, and forms the backbone of any data processing system. It is concerned with the infrastructure which may enable statistical calculations or data mining to take place. Statistics is concerned with processing samples of the full dataset, tends to be descriptive, and to be used for numeric data. It requires prior assumptions about the nature of the data to be made. On the other hand, data mining involves automating the process of finding patterns in data, and tends to operate over larger datasets, generally without the need for prior assumptions about the data to be made. The data tends to be numeric as well. From the characterisations presented, it can be observed that data mining is the established field which most closely resembles data science.

However, data science bandwagon promises to offer more than what traditionally data mining does. It aims to deal with significantly more

complex data, particularly in terms of the volume, variety, velocity dimensions. Less emphasis is made on the veracity dimension. Considering that it is being cast as a science, it arguably should have greater emphasis on the veracity dimension, so that it can aspire to reach the Wisdom layer of the DIKW hierarchy.

As a final thought, it is worth mentioning that about 10-15 years ago, there was a lot of hype about *e-science* (Hey & Trefethen, 2002), a research area with similar purported goals to data science. The grid was also a hot research topic, and is seldom mentioned nowadays, as the focus is on cloud computing. A sceptic may argue that buzzwords are invented every few days, by people in industry hoping to market new products, and academics who yearn for research funding. At present, these are early days, and it seems imprudent to conclude that data science will emerge as a new discipline in its own right, or that it will be subsumed by data mining. Perhaps, the outcome depends on how effectively the Veracity dimension is handled.

Acknowledgements

The author is grateful to Alvaro A. A. Fernandes for providing feedback and suggestions during the writing of this paper.

REFERENCIAS

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record, 39*(4), 12-27. Retrieved from http://www.sigmod.org/publications/sigmod-record/1012/pdfs/04.surveys.cattell.pdf

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387. Retrieved from https://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf

Davenport, T. H., & Patil, D. J. (2012, October). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* (70). Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64-73. Retrieved from http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/abstract

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37-54. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Fernandes, Alvaro A. A. (November 2004). *Advanced Database Technologies: Data Mining. Slides.* Manchester: School of Computer Science, University of Manchester.

Friedman, J. H. (1998). Data Mining and Statistics: What's the connection? *Computing Science and Statistics, 29*(1), 3-9. Retrieved from http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf

Gantz, J., & Reinsel, D. (2013). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East. *IDC iView: IDC Analyze the Future, 2007,* 1-16. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf

Heilbron, J. L. (Ed.). (2003). *The Oxford companion to the history of modern science.* Oxford: Oxford University Press.

Hey, T., & Trefethen, A. E. (2002). The UK e-science core programme and the grid. *Future Generation Computer Systems, 18*(8), 1017-1031.

ICDE (2016). *32nd IEEE International Conference on Data Engineering.* Retrieved from http://icde2016.fi/

Press, G. (2013). *A very short history of data science.* Forbes Technology. Retrieved from http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51-59. Retrieved from http://online.liebertpub.com/doi/pdf/10.1089/big.2013.1508

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science, 33*(2), 163-180. Retrieved from http://jis.sagepub.com/content/33/2/163

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE. Retrieved from https://xa.yimg.com/kq/groups/72986399/1585974627/name/06567202.pdf

SIKS (2015). *Data engineering for data science: challenges and opportunities.* http://wwwis.win.tue.nl/~gfletcher/luo/

Stonebraker, M., & Weisberg, A. (2013). The VoltDB Main Memory DBMS. *IEEE Data Eng. Bull., 36*(2), 21-27. Retrieved from http://sites.computer.org/debull/a13june/voltdb1.pdf

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. Retrieved from https://www.jstor.org/tc/accept?origin=/stable/pdf/2237638.pdf

University of Aalborg (2015). *Masters in Data Engineering.* http://www.en.aau.dk/education/master/data-engineering

University of Dundee (2015). *Postgraduate Courses - Data Engineering MSc.* http://www.dundee.ac.uk/study/pg/data-engineering/

Unwin, A., Theus, M., & Hofmann, H. (2006). *Graphics of large datasets: visualizing a million.* New Jersey: Springer Science & Business Media.

Weikum, G., & Theobald, M. (2010, June). From information to knowledge: harvesting entities and relationships from web sources. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 65-76). ACM. Doi: 10.1145/1807085.1807097.

Revista de divulgación científica
CITAS
CIENCIA INNOVACIÓN TECNOLOGÍA AMBIENTE SOCIEDAD