Mathematical and statistical foundations and challenges of (big) data sciences

AUTHOR:
Loyiso G. Nongxa[1]

AFFILIATION:
[1]Centre for Mathematical and Computational Sciences, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Loyiso Nongxa

EMAIL:
Loyiso.nongxa@wits.ac.za

KEYWORDS:
high dimensionality; heterogeneity; data analytics; modelling; data sets

HOW TO CITE:
Nongxa LG. Mathematical and statistical foundations and challenges of (big) data sciences. S Afr J Sci. 2017;113(3/4), Art. #a0200, 4 pages. http://dx.doi.org/10.17159/sajs.2017/a0200

The hype around data sciences in general and big data in particular and the focus either on the potential commercial value of data analytics or on promoting its adoption as a new paradigm in conducting research, may crowd out important discussions that need to take place about the theoretical foundations of this 'emerging' discipline. In South Africa, discussions around (or the mere mention of) big data, especially within the National System of Innovation, often go hand in glove either with the Square Kilometre Array project and astrophysics, or eResearch or cyberinfrastructure. In his excellent essay '50 Years of data science', David Donoho of Stanford University remarks:

> *The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning which adds some technology for 'scaling' up to 'big data'. This chosen superset is motivated by commercial rather than intellectual developments. Choosing this way is likely to miss out on the really important intellectual development of the next fifty years.*[1]

It has now been recognised in some academic circles that important advances in the rapidly evolving 'discipline' of data sciences will depend significantly on contributions from communities in mathematics, statistics and computer science. This is reflected, amongst other things, by the number of conferences, workshops and thematic programmes that have been organised under themes like 'mathematical challenges of big data', or 'thematic programme on statistical inference, learning and models for big data'. Conversely, questions and open problems arising out of big data (will) spur new research activities and directions for the mathematical, statistical and computational sciences, and could lead to the opening of new frontiers in mathematics, statistics or theoretical computer science. The main objective in this commentary is to argue for the positioning of computational, mathematical and statistical sciences in South Africa at the centre of the heralded big data revolution. These disciplines are strategically important to provide a solid intellectual and academic foundation upon which to build a vibrant and successful project in big data analysis, especially in South Africa. This could lead to the renewal and rejuvenation of mathematical and statistical sciences in South Africa and facilitate collaboration and bringing closer researchers to work on problems at the boundaries or intersections of the respective disciplines. It also provides an opportunity to produce graduates with the breadth and depth of knowledge in all three major disciplines who are versatile enough to either be capable of pursuing fundamental research in these three broad disciplines or work in areas (public or private sectors) focusing on applications of (big) data sciences.

Background and terminology

The views expressed in this commentary are based on a desktop analysis of four different types of documents. The first are reports of workshops and thematic programmes funded and hosted by, for example, some National Science Foundation funded institutes like ICERM, SAMSI and IMA in the USA; the Fields and Banff Institutes in Canada; IMA in the United Kingdom and ACEMS in Australia. The second are abstracts of research papers presented at some of these workshops and survey articles in online research journals. The third type of documents considered are reports produced by Committees of the National Research Council of The National Academies – two worth highlighting are 'The Mathematical Sciences in 2025' and 'Frontiers in Massive Data Analysis'. And fourthly, syllabi and course outlines of mainly postgraduate courses that have been implemented at universities that have introduced new master's programmes in data sciences. In September 2016, the Academy of Science of South Africa convened a 2-day workshop under the theme 'Finding Synergies in the Mathematical Sciences'. This commentary is an overview of a background document prepared for a plenary presentation at the workshop.

There are terms that frequently crop up in discussions about (big) data analysis that often implicitly are used interchangeably as if they were synonymous. For the purposes of this commentary, we record the meaning that will be attached to these terms, although there is no consensus on the formal definition of these terms and how they are used. 'Data sciences' is an emerging 'science' concerned with extracting knowledge and information and gaining insights from data sets 'arising from experimental, observational, and/or simulated processes in the natural and social sciences and other areas'[2] which may be structured or unstructured and collected under diverse circumstances and environments. The 'theoretical foundations of data sciences' is a new area of academic interest at the intersection of statistical sciences and computer science, founded on a strong base of the mathematical sciences (pure and applied) and ranges from developing the theories, algorithms and methodologies, to paying attention to and understanding the applications of these within various domains. 'Data analytics' is the process of examining raw data with the purpose of knowledge discovery, gaining insights of information and hidden value from the data using software tools, techniques, processes and algorithms that have been specifically developed for such purposes. It is often argued that data scientists should have some specific domain knowledge, be it, for example in business management or health sciences. Data analytics in different domains gives rise to terms that signify the domain of the application: 'business analytics' is the application of data sciences in the business environment and has disproportionately received more attention, possibly more than applications in any other domain. One could (erroneously?) ascribe this phenomenon to the popularity of the book *Competing on Analytics: The New Science of Winning* by Jeanne G. Harris and Thomas H. Davenport and the highly cited and popular McKinsey Global Institute report, 'Big data: The next frontier for innovation, competition and productivity'. Other applications that have received significant attention would be health analytics, text analytics, fraud and risk analytics and financial analytics, to name just a few. On the other side of the spectrum, the influential book *The Fourth Paradigm: Data-intensive Scientific Discovery*, edited by Tony Hey, Stewart Tansley and Kirstin Tolle, makes

the argument that scientific advances are becoming more and more data driven and researchers will more and more have to think of themselves as consumers of data. This has possibly influenced eResearch, defined by Whitmore[3] as:

> *the use of information technology to support existing and new forms of scholarly research in all academic disciplines…encompass*[ing] *computational and eScience, cyberinfrastructure and data curation…usually data-intensive but the concept also includes research performed digitally at any scale.*

Big data challenges:

There is an emerging consensus about the meaning of 'big data'[4-6], using the so-called four V's: volume, velocity, variety and veracity. Simply put, big data refers to massive data sets that are so voluminous and/or move so fast and/or vary in quality and structure, or may be 'questionable', which may make it difficult (to impossible) to store, manage and process using traditional methods. It is also generally accepted that one has to be mindful of the entire 'big data analysis pipeline' (sometimes called the big data analytics life cycle). This pipeline or life cycle involves two major phases: data management and data analytics. Each of these consists of clearly identifiable steps, each of which may present major challenges. Data management involves (1) acquisition and recording; (2) extraction, cleaning and annotation; and (3) integration, aggregation and representation. The analytics phase involves (1) data modelling and analysis; (2) interpretation; and (3) decision-making.

Some of the common challenges that underlie many, and sometimes all, of the different phases and processes above, require responses that depend significantly on ideas from the mathematical and statistical sciences.

Big data have unique features that are often not shared by or found in small or traditional data sets: (1) high dimensionality; (2) heterogeneity and incompleteness; (3) scale; (4) timeliness; and (5) security and privacy.

High dimensionality

A data point represents an object with, say p features, where p could be a very large positive integer. Geometrically the data point lives in a high-dimensional vector space. The geometry of high-dimensional vector spaces exhibits peculiarities which can be counter-intuitive when one attempts to extrapolate from lower-dimensional vector spaces and this is a major aspect of what is known as the 'curse of dimensionality'. For example, the volume of an n-ball tends to zero as n tends to infinity. If we have n data points, then in traditional statistical analysis, p may be fixed and much smaller than n (the data set can be represented by 'tall-and-skinny' matrices). But if p is much larger than n (data set represented by 'short and fat' matrices), or p increases as n increases, then the traditional methods of analysis could break down. It has been observed that high dimensionality could lead to spurious correlation, where variables that are (theoretically) independent may have high sample correlations. Spurious correlation may cause false scientific discoveries and wrong statistical inferences. The mathematical and statistical properties of high-dimensional data spaces are often poorly understood and inadequately considered.

Heterogeneity and incompleteness

One of the steps identified above in the big data analysis pipeline is the integration, aggregation and representation of data sets. These will be data sets corresponding to different populations that could have been collected from different sites and different environments, using different platforms, methodologies and technologies. Algorithms deployed on computational devices 'expect' homogeneous data and the systems have been designed to analyse data more efficiently if the data are structured and have identical format and size. A data set could have points with missing records and 'educated' estimates of these missing points could introduce biases and skew conclusions. Also omitting points with missing records could lead to different conclusions and recommendations.

Scale

The growth in computational and storage power has made it possible to work with massive data sets rather than small samples one normally deals with in traditional statistical analysis. It is common to discard or ignore outliers when analysing data sets of small size (which may be the case in traditional statistical analysis). However, these could in fact be representatives of what could turn out to be important subpopulations. Big data analysis allows for analyses that may reveal hidden structures of each subpopulation of the data. It may also reveal important common features across many subpopulations even when there are large individual variations. There are technical challenges that have been identified when it comes to working with data at peta-exascale. The 'fundamental shift underway now [is]: data volume is scaling faster than compute resources, and CPU speeds are static'[7].

Timeliness

The increase in the size of data sets to be analysed using the traditional methods and techniques means that it often takes longer to analyse. There could be situations in which there are time constraints to analysis, interpretation and decision-making and the result of the analysis is required immediately. This would be the case when a high-quality answer that is obtained slowly can be less useful (and more costly) than a medium-quality answer that is obtained quickly. An example is credit card fraud detection: it would bring down costs and prevent financial loss if this is flagged before a transaction is completed, potentially preventing the transaction from taking place. It is obvious that a full analysis of the background of the credit card owner's spending patterns may not be feasible in real-time and rather a decision would be based on a partial analysis.

Security and privacy

New capabilities to gather, analyse, disseminate, and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected.[8] There is justifiable public fear regarding the inappropriate uses of personal data. There is a conflict between the use of personal information for a 'greater' (?) public good on the one hand and disadvantaging individuals as a result of 'insights' gained from big data analysis on the other hand. This is not new and has occurred throughout our history and continues today. Today's concern about big data reflects both the substantial increase in the amount of data being collected and the fact that in many instances ordinary citizens are not even aware of the data that are being collected about them, nor aware of who has the data and how the data may be used against them. The sharing of data about individuals, without their consent, by those who hold such data and might be interested in aggregating the data to gain new insights is a big concern. The same data and analytics that provide benefits to individuals and society if used appropriately can also create potential harm. GPS tracking of individuals might lead to a better understanding of traffic problems which could lead to innovative solutions; but can also be used inappropriately in tracking the movement of individuals and their whereabouts. The recently published book by a former Wall Street quant, Cathy O'Neil, titled *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, raises some legitimate concerns about some of these issues.

Big data techniques

It was noted earlier that a data point can be viewed as an n-dimensional vector, where n represents the number of features that are being observed. If we consider a set of m such data points, then these can be represented as an m by n matrix. The data points can be thought of as being randomly selected from a probability distribution of this n-dimensional vector space. In certain instances, tensors, which are multidimensional generalisations of matrices, have been found to be useful in representing multidimensional data, for example in neurosciences. Data points can sometimes be viewed as 'nodes' of a graph, where the 'edges' represent, for example, 'relationships' between the nodes. The number of nodes and/or edges may increase or decrease with time and the edges may carry 'weights'. Topological methods can sometimes be deployed to determine the structural features of the data sets and geometry employed

to determine the 'distance' between the data points. In some cases the data set can be too large to store or difficult to capture because it is arriving at high frequency with respect to the analysis resources available and sophisticated sampling and other statistical analysis techniques are required. The analysis of the data has to be as efficient and cost effective (in time, resources, etc.) as possible. This provides some indication as to why optimisation is ubiquitous in (big) data analytics.

Many of the big data techniques that often are deployed to efficiently process large volumes of data are founded on some of the mathematical and statistical concepts implicit in the previous paragraph and others. These include, but are not restricted to: (1) dimension reduction; (2) clustering; (3) data visualisation; (4) optimisation methods; (5) statistical methods (computing, inference, modelling and learning); (6) data mining; (7) machine learning; and (8) social network analysis.[9]

Mathematical and statistical challenges

The first issue to address is that of workforce development in an area that has been described as an emerging discipline and covers a broad range of learning areas that currently are taught in silos. The three communities – mathematics, statistics and computer science – in South Africa need to rise to the challenge of training the next generation of graduates who will have the necessary breadth and depth of knowledge in these three disciplines which are necessary to keep abreast of developments in the theoretical foundations of (big) data sciences and contribute to future developments of this new discipline. The graduates will also have to be comfortable working in multidisciplinary teams both in the public and private sectors. The core and indispensable knowledge areas include linear algebra, multivariable calculus, elementary probability and statistics, as well as experience in writing code in one of the main programming languages. These would normally be taught in the first 2 years at the undergraduate level. At the senior undergraduate level up to honours, matrix analysis, optimisation, graph theory and algebraic topology (emphasis on homology) would be foundational knowledge areas and other important topics would include harmonic analysis, time series analysis, approximation theory and a firm grasp of functional analysis. Stand-alone modules for each of these would lead to a 'bloated' curriculum and the challenge is for the experts to rethink the core knowledge and reorganise it into new packages that aggregate and integrate topics that are usually taught in silos. The more advanced topics are either not prominent or absent on the national research landscape in South Africa. In the mathematical sciences, these topics would include: (1) randomised numerical linear algebra; (2) topological data analysis; (3) matrix and tensor decompositions; (4) random graphs; (5) random matrices; and (6) complex networks.

Over the last 4 years, there has been a big focus on the foundations of (big) data sciences both in North America and the United Kingdom. There have been extended long-term thematic programmes hosted mainly (but not exclusively) by institutes in the USA funded by the National Science Foundation and privately, and government- and private-funded institutes in Canada. In the United Kingdom, the Engineering and Physical Sciences Research Council (EPSRC) has established (and is funding) Centres for Doctoral Training (CDTs) with a focus on various aspects (mainly statistical and computational) of data sciences. These offer a 4-year PhD with a compulsory large volume of coursework in the first year. Examples of thematic programmes are 'Theoretical Foundations of Big Data Analysis' hosted by the Simons Institute for the Theory of Computing (22/08/2013 – 20/12/2013); 'Statistical and Computational Methodology for Massive Data Sets', hosted by the Statistical and Mathematical Sciences Institute (2012/2013) and the 'Thematic Programme on Statistical Inference, Learning, and Models for Big Data' hosted by the Fields Institute for Research in Mathematical Sciences (06/2015 – 12/2015). The EPSRC funds at least three CDTs and their themes are 'Data Sciences' hosted by the University of Edinburgh, 'Cloud Computing for Big Data' hosted by the University of Newcastle and 'Next-generational Statistical Science' co-hosted by Oxford University and Warwick University. The reports of the thematic programmes and the coursework component of the CDT PhDs are a rich source of information both for topics for coursework in statistical sciences as well as open problems and research directions. The broad and common topics for coursework include (1) statistical modelling; (2) statistical and machine learning; (3) statistical inference; (4) computational statistics; (5) probability and approximation; (6) Bayesian methods for big data analysis; and (7) probabilistic methods for big data.

Future directions

As previously remarked, the reports of the thematic programmes hosted by national institutes provide a reservoir of information about, amongst other things, current research as well as research questions that would make a significant impact on and contribute to the theoretical foundations of data sciences. To provide a taste of some of these, below is an extract from a Call for Proposals by the US National Science Foundation under its 'Critical Techniques and Technologies for Advancing Foundations and Applications of Big Data Sciences and Engineering' programme:

- Sophisticated computational/statistical modelling for simulation, prediction, and assessment in computation-intensive and data-intensive scientific problems.
- State-of-the-art tools and theory in statistical inference and statistical learning for knowledge discovery from massive, complex, and dynamic data sets.
- General theory and algorithms for advancing large-scale modelling of problems that present particular computational difficulties, such as strong heterogeneities and anisotropies, multiphysics coupling, multiscale behaviour, stochastic forcing, uncertain parameters or dynamic data, and long-time behaviour.
- Study of mathematical, statistical, and stochastic properties of networks.
- Mathematical and statistical challenges of uncertainty quantification.
- Development of numerical, symbolic and statistical theory and tools to uncover and study analytical, topological, algebraic, geometric, and number-theoretic structures relevant for large-scale data acquisition, data security and cybersecurity.

Conclusion

The main objective of this short commentary was to propose to the mathematics, statistics and computer science communities at South African universities as well as private and public sectors to take an interest in the theoretical foundations of data sciences. This interest has potential to foster dialogue and collaboration amongst members of these communities. Such collaboration could spur rejuvenation and renewal in these three disciplines through incubating new areas of study on the South African higher education landscape and graduating the next generation of scholars with breadth and depth of knowledge in mathematics, statistics and computer science. Real and meaningful progress in big data does not only require 'whiz kids' in Hadoop, MapReduce, Spark, Python and R but also graduates who understand the algorithmic, mathematical and statistical underpinnings of these programs. The call we are making has potential for new collaborations between the university sector and users of mathematics, statistics and computer science in the private and public sectors and non-governmental organisations.

Acknowledgements

This commentary is based on a presentation I gave at a 2-day workshop organised by the Mathematical Sciences Committee of the Academy of Science of South Africa. I would like to thank Professor Fritz Hahne for bringing to my attention the report of the National Academy of Sciences titled 'The Mathematical Sciences in 2025' when I was on a fellowship at the Stellenbosch Institute for Advanced Studies (STIAS). I would also like to express my gratitude to STIAS for their generous financial support.

References

1. Donoho D. 50 Years of data science [document on the Internet]. c2015 [cited 2016 Jun 07]. Available from: http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
2. Workshop on "Theoretical Foundations of Data Science (TFoDS)" [document on the Internet]. c2016 [cited 2017 Mar 09]. Available from: http://www.cs.rpi.edu/TFoDS/TFoDS_v5.pdf
3. Whitmore AL. Thoughts on eResearch: A scientist's perspective. J eScience Librariansh. 2013;2(2), e1045, 5 pages. http://dx.doi.org/10.7191/jeslib.2013.1045
4. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293–314. https://doi.org/10.1093/nsr/nwt032
5. Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Shahabi C. Big data and its technical challenges. Commun ACM. 2014;57(7):86–94. https://doi.org/10.1145/2611567
6. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, et al. Challenges and opportunities with big data: A white paper prepared for the Computing Community Consortium Committee of the Computing Research Association [document on the Internet]. c2012 [cited 2017 Mar 09]. Available from: http://cra.org/ccc/resources/ccc-led-whitepapers/
7. Lodha R, Jan H, Kurup L. Big data challenges: Data analysis perspective. Int J Current Eng Technol. 2014;4(5):3286–3289.
8. Moura J, Serrao C. Security and privacy issues of big data [article on the Internet]. c2016 [cited 2017 Jan 21]. Available from: https://arxiv.org/abs/1601.06206
9. Chen CLP, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. Inform Sci. 2014;275:314–347. http://dx.doi.org/10.1016/j.ins.2014.01.015