ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Royal Statistical Society response to the European Commission on Mathematics and Digital Science

The Royal Statistical Society is a learned society for statistics, a professional body for statisticians, and a charity that promotes statistics for the public good, with more than 6000 members around the world. We support that there should be more collaboration across mathematics and digital science to advance and apply research. Many of the methods that should be highlighted for this are their heart statistical, and in our answers below we address what statistics in particular can contribute to discussions convened by the European Commission.

1. The role of statisticians in big data. Could new mathematical methods (from e.g. topology, stochastic, probability theory etc.) help Europe fully profit from available data and solve main problems in data sciences?

a) What methods can statistics contribute?

"Big data" is the latest phrase for what has been previously or otherwise called "data analytics" or "data mining". Statistics is fundamental to and at the very heart of new cutting edge developments in the field, but it equally contributes well-established methods and theories that are just as important and of ongoing usefulness. These include the following:

1) Big data needs to be made smaller. Many of the findings that emerge from big data are based on only a small fraction of the data collected, and are in fact reached through sampling methods. Statistics offers models and tools to screen and split big data in a 'divide and conquer' approach, to decide how much of the data is useful and how much to discard.

2) Statistics helps us to be less wrong. Analysis using statistics is needed to ensure that results produced by models and tools are not over-interpreted as important or decisive, or interpreted wrongly.

For example, at the Large Hadron Collider at CERN, the massive quantities of data that were collected had to be reduced first by a factor of 10,000 and then by a further factor of 100, to reach a level of 'big' data that could be analysed.[1] The number of potential false findings grew exponentially as the data got bigger. Statistical inference methods and careful practice were crucial to take account of the false discovery rate, and to judge with any confidence whether a discovery had really been made.[2]

Statistics is also at the forefront of more sophisticated dealings with uncertainty and unknowns. As mentioned in the EC's work programme on Future and Emerging Technologies, there is a need to develop algorithms that quantify uncertainty and noise for 'exascale', multi-scale data.[3] Big data collection is subject to biases and is always some form of abstraction from what is really and currently taking place. Data can be empirically analysed by digital processes, but mathematics and statistics are crucial for substantive interpretation. Probabilistic modelling is needed to interpret

stochastic (non-deterministic) events, and to take account of changing interventions over time. Pattern recognition becomes highly sophisticated as networks of relationships between data-points are mapped.

b) How can mathematicians, politicians, businesses and society best work together in making use of data to tackle societal challenges?

Collaborative consortia are needed across research disciplines including mathematics and statistics. These should reach out to support industrial applications in businesses, in government and in non-governmental organisations. The development of data for society should also include widening the benefits from collecting and curating data, for example by seeking to open up data where this would be of benefit, and working to address privacy concerns with regard to the use of personal and administrative data. It is also necessary to strengthen official statistical systems and for research consortia to liaise with these. We see a positive relationship between the extent and robustness of official data and the scope for development that is of benefit for society to take place.[4]

2. The role of mathematics in high performance computing (HPC), in particular 'exascale' computing. In the light of the changes imposed on computing due to the data deluge, how can mathematicians help Europe advance towards data-centered HPC?

The European Commission defines high performance computing as the development of multiple processors for processing more data more quickly, with an expected speed of 10^{18} operations per second by 2020. This agenda excludes distributed data, such as cloud computing, and is all about processing power.[5]

An understanding of the statistical applications of HPC is needed to inform the design of HPC services. We welcome that the European Commission envisages "ambitious applications and close cooperation with the scientific disciplines and stakeholders concerned" so that different features can be optimised for different scientific or industrial challenges. Providing a data-centred model of particular functions of the human brain is one challenge for example, whereas using HPC to inform the design of batteries for electric cars would be another. Statistical methods help to analyse data for the purpose of designing models. Statistics are also needed to assess the uncertainty in new high-performance simulations, and the extent to which they are useful.

3. The role of e-infrastructures in maths. Could e-infrastructures help resolve the biggest challenges in maths? How e-infrastructures could help mathematicians manage the existing level of complexity of mathematical problems in ways that are not feasible today and may result in significant scientific breakthroughs? What are the needs in terms of specific e-infrastructure services?

E-infrastructure can decisively change the extent and availability of data to which advanced statistics and mathematics can readily be applied. An example given in EC documents is the Research Data Alliance, which seeks to ensure data interoperability and exchange worldwide.[6] Work under this agenda could help to share information across borders and improve international statistics. The way in which e-infrastructure operates however differs across and even within European states, so it is right that the EC has invited contributions of new sustainable models for data sharing rather than seeking to take a top down approach. Discussion of e-infrastructure for research also needs to be joined up to discussion of legislation such as EU data protection law, as this has implications for how research data are managed.

4. The impact of applied and industrial mathematics on innovation. How can we maximise it?

With regard to statistical methods and theory in particular, the American Statistical Association's working group on big data research and development has found that the most productive innovation is likely to occur at the interface between statistics and computer science, in multi-disciplinary teams.[7] Their recent paper on 'discovery with data' offers numerous ideas of areas for development. They look in turn at biological sciences/bioinformatics, healthcare and public health, civic infrastructure and governance, business analytics, the social sciences, and physical and geosciences. Across these fields they find a need for researchers who are computationally savvy and statistically literate. This is a valuable combined skill set for innovation, and skills in this area should be supported through research funding opportunities and training.

5. **New mathematics-related and statistical science topics to be discussed online or in an upcoming workshop.**

An issue that has often arisen is the lack of sufficient communication between different specialisms in mathematics and statistics. In many application areas, there are potential gains from bringing together insights from applied mathematical modelling (deterministic or stochastic) and statistical modelling approaches. This has been done successfully in, for instance, very large scale climate and environmental modelling, where, despite advances in HPC, there are still (and always will be) limitations on the spatial and temporal scale and detail of the models that can be run. Statistical emulation of the modelling, as an addition to the overall modelling process, can add invaluable insights on the uncertainty in predictions from the model. Online discussion or workshop activity on other areas where varying statistical and mathematical approaches are currently used could lead to similar advances of understanding and capability. Indeed for any areas where prediction (even short-term) is involved, some sort of modelling and consideration of cause will be fruitful. Current data, however abundant and comprehensive, do not speak for themselves about the future.

Causal inference may be a further area for discussion as it is developing as a major topic in statistics and in artificial intelligence/machine learning, as well as in econometrics. The roots of causal inference lie in the need to deal with data that do not arise from carefully controlled experiments, so it is relevant to the observational data that arise in many social and business big-

data contexts. There is potential for further work to apply causal inference in larger and more diffuse data contexts.

Response submitted 29 September 2014 by the Royal Statistical Society's Policy and Research Manager.

---

[1] Hand, D.J. (2013) Data not dogma: big data, open data, and the opportunities ahead. In *Advances in Intelligent Data Analysis, XII, 12th International Symposium, LNCS 8207*. Berlin, Springer. Pp. 1-12.
[2] Spiegelhalter, D.J. (2014) The future lies in uncertainty. *Science 345*(6194). Washington, AAAS. Pp. 264-265.
[3] European Commission (2014). *Horizon 2020 Work Programme 2014-2015: Future and Emerging Technologies, Revised 22 July 2014* [PDF]. Available from: ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-fet_en.pdf (Accessed: September 2014)
[4] Shah, H. (2014) Data revolution: why we mustn't blow it [online article]. *Post2015.org*, 10 April 2014. Available at: http://post2015.org/2014/04/10/data-revolution-why-we-mustnt-blow-it/ (Accessed: September 2014)
[5] European Commission (2012) *High Performance Computing: Europe's place in a global race* [PDF]. Available from: http://ec.europa.eu/digital-agenda/futurium/sites/futurium/files/futurium/library/CommunicationHigh-PerformanceComputingEuropesplaceinaGlobalRace-COM.pdf (Accessed: September 2014)
[6] European Commission (2014) *HORIZON 2020 WORK PROGRAMME 2014–2015: 4. European research infrastructures (including e-Infrastructures), Revised 22 July 2014* [PDF]. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-infrastructures_en.pdf (Accessed: September 2014)
[7] American Statistical Association (2014) *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society* [PDF]. Available from: http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf (Accessed: September 2014)