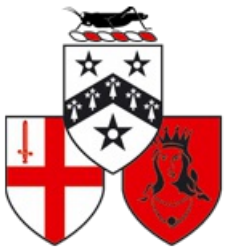

GRESHAM COLLEGE

Founded 1597

The Challenge of Big Data Transcript

Date: Tuesday, 15 November 2016 - 1:00PM

Location: Museum of London

15 November 2016

The Challenge of Big Data

Professor Chris Budd OBE

Introduction

The subjects of **Big Data** and **Data Analytics** are much in the news at the moment. According to the press it is all around us, will make a huge difference to our lives, and has massive ethical issues which should worry us all. Anyone who has a Smart Phone, a LapTop or who uses Google is already interacting with Big Data. But it is much more pervasive than that. Indeed, we are meeting and generating Big Data (all of which can be analysed) whenever we drive, turn on a light switch, see the doctor or even go to the shops. On a very positive note, the use of Big Data in medicine has been advertised as a means of curing many diseases including cancer [1]. Mathematical algorithms linked to Big Data lie at the heart of software such as Google which now seems quite indispensible to our lives. Other mathematical algorithms are used extensively by many organisations, such as retailers, as a vital part of our modern lives, including (according to the new film *Bridget Jones' Baby)* helping us to decide on our life partner. However, Big Data also has a darker side linked to personal privacy. For example, the rumours are that by using Big Data a supermarket can tell whether you are pregnant by the foods that you are buying, before even you or your doctor know about it yourselves. For example if a woman searches for or buys items such as prenatal vitamins and unscented lotions, then she is much more likely to be pregnant. It is claimed that the algorithms used by the retailers to do this are so precise that they can even predict when the baby is due. See the New York Times article [2].

So, what is all the fuss about, what really is Big Data, and is the press really right to be worried about it? To a certain extent I think that there is cause for concern. Big Data really will change our lives, for both good and bad, and we must certainly be alert to the ethical issues concerning it. However there is still a lot of confusion about exactly what Big Data is and what it can do. I will attempt to explore these issues in this talk, explain how mathematics can make sense of big data and (for example) advance medicine in the process, as well as looking at the very big ethical issues involved.

Eight Great Technologies

In 2012 HM Government identified a list of Eight Great Technologies, which it saw as the future technologies in which the UK will be a world leader. These were launched in a speech by the former minister for science, The Rt. Hon David Willetts MP. This speech has led to an HM Government Industrial Strategy report and a flurry of activity on many websites. More information on the eight great technologies is given in the government report [3]. In the original speech in 2012 the Eight Great Technologies were identified as being

- Big Data
- Satellites and space
- Robotics and autonomous systems
- Synthetic biology
- Regenerative medicine
- Agricultural science
- Advanced materials
- Energy and its storage.

More recently, quantum based technology has been added to this list, and it is likely to grow further. Similar lists, which overlap considerably in content with those from the UK have been compiled in other countries' government publications as well as in the popular media (such as the MIT Technology Review or the Scientific American). It is worth noting that the HM Government list was identified by the *Policy Exchange Think Tank* and the *Technology Strategy Board,* in collaboration with research scientists and members of the research funding bodies. A technology made it on the list if:

- It represented an important area of scientific advance
- There was already some existing capacity for it in the UK
- It was likely that new commercial technologies would arise from it
- There was 'some' popular support for it

Later on in this course of lectures I will look in some details at the mathematics associated with Agricultural Science, Advanced Materials, and Energy, but in this lecture we will look at the first, and arguably most important of these technologies, namely Big Data.

What is Big Data and where does it come from?

We live in the information age, and most of what we do is hugely influenced by our access to massive amounts

of data, whether this is through the Internet, on our computers, or on our mobile phones. About 100 years ago when we were transmitting information by Morse code, the transmission rate was 2 bytes per second. This improved with the use of the tele-printer to 10 bytes per second, and then with the modem to 1 kilobyte per second. In contrast, with modern data we are looking at transmission rates of over **1 Gigabytes per second**. To put this into context, a byte is one letter, so in one Giga bytes we have one billion letters, which is about 150 million words, or about 1000 whole books, arriving for us to read **every second**. Personally I'm doing well if I read one book a week! Similarly early computers (such as the one I used to do my PhD back in 1983) had about 1 kilobyte of random access memory (RAM)

(with more data having to be stored, unreliably, on magnetic or even paper tape). Whereas a

modern lap top has several gigabytes of RAM and up to 1 terabyte of memory (that is one trillion bytes or about 1 million books, far bigger than the average library). Access to such a large amount of data leads in turn to large technological and ethical problems. Mathematics can help us with the former, and I will argue strongly in this lecture that we should all be aware of the latter. So, what does the 'Challenge of Big Data' mean?

According to the HM report on the Eight Great Technologies it is

The collection, handling, assurance, curation, analysis and use of:

- *Large amounts of existing data using existing methods and technology*
- *Existing data using new methods and technology*
- *New data using new methods and technology*

The challenge of dealing with such data is always to derive value from large signals, *where the useful data may be buried in an avalanche of noise.* One of the big problems associated with this is that one person's noise may well be another person's signal. For example, if I am talking on the phone then the sound of a car starting up outside is just noise. Until I realise that the car starting up is my own and it is being stolen. At that point the phone conversation becomes the noise! This example also illustrates the subjective nature of Big Data. The real problem of dealing with modern data is not so much its volume (although this is still an issue) but rather in dealing with its highly imprecise and subjective nature. It is, for example, generally easier to talk about the weather, than it is to assess someone's reasons for wanting to buy chocolate in a supermarket.

Perhaps the leading source of current Big Data comes from *the Internet.* According to a recent estimate, about bytes (a Zettabyte) of information are added to the Internet every year, much of which is graphical in content. The 'internet penetration' in the UK, Canada and Korea in particular is over 80% of the total population, and in all but a few countries is over 20 %. This wealth of data leads in turn to the huge mathematical problems of how we identify, search and organize this information. A major source of this data comes from the ever growing, and very fuzzily defined, content on Social Media websites. For example*, Facebook* was launched in 2004. It now has 2 Billion registered users (about one quarter of the world's population!) of which 1.5 Billion are active. Around 2.5 Billion pieces of content (around 500 Terabytes of information) are added every day to Facebook sites, with most of this data stored as pictures. The search engine Google is estimated to somewhere around 1-15 Exabytes (bytes) of data, which it searches by using an algorithm based in part on finding eigenvectors of very large matrices. Another source of Big Data comes from mobile and smart phones. There are now more mobile phones than people in the world, with the potential for

simultaneous conversations. The forthcoming plans for a 5G network will operate on millimetre wavelengths at frequencies as around 70GHz (and are already being piloted in my home town of Bristol, UK). The 5G networks will offer data rates at 1 Gigabyte per second offered simultaneously to tens of workers on the same office floor and with several hundreds of thousands of simultaneous connections supporting massive deployments of sensors. Such sensors can provide constant monitoring of, say, our state of health, energy and life-style, with

significant ethical implications. Indeed the future is rapidly approaching (such as in the *Internet of Things*) in which our domestic devices simply communicate with each other (for example the cooker talks to the dishwasher and also to the supermarket every time a meal is prepared) with little or no human interference.

As well as the devices above, significant amounts of data, of significant interest to the social sciences, comes from the way that we use devices, which in turn generate lots of data, and the information that it gives about our lifestyles. Again there are significant ethical issues here. Every time that we make a purchase with Amazon, shop on line, use our bank on-line, switch on an electrical device, or simply use a mobile phone or write an email, we are creating data which contains information which can in principle be analysed. For example our shopping habits can be determined (and they can find out if we are pregnant), or our location tracked and recorded. Mathematics can be used at all stages of this, but we must never lose sight of the moral dimension in so doing. I will return to this topic at the end of this lecture.

The nature of Big Data

In one sense, Big Data has been the subject of mathematical investigation for at least 100 years. Any mathematical model described by a partial differential equation with an infinite number of degrees of freedom, naturally leads to a source of a large amount of data. A classical example of this is meteorology, in which the current meteorological models (typically based on extensions of the Navier-Stokes equations of fluid mechanics) are solved on super computers with discretisations with over a Billion degrees of freedom informed (in a typical 6 hour forecast window) by over a Million observations of the state of the atmosphere and oceans. Similar large data sets arise in climate models, geophysics and (especially) astronomy. However, the data in these problems, whilst very large, is also well structured and well understood (with known levels of uncertainty), as befits its origins in the physical sciences for which we have good and well understood mathematical models.

The real challenges of understanding and dealing with Big Data do not come from these

data sets, however large they may be. In contrast the real difficulties arise from data, which has its origins (as described above), in the biological sciences, the social sciences and in particular in people based activity.

Such data is **Challenging** in that it is: garbled, partial, unreliable, complex, soft, fast arriving, and (of course) big.

It is also **Novel** and very different from much of the data arising from physical models in that it is: heterogeneous, qualitative, relational, and partial.

As an example, suppose that a Tsunami is approaching a coastline. As it approaches we can gather data on it from satellites and ocean buoys. This data is well structured and we can combine it with good models of tsunami motion to get a good estimate both of the magnitude of the tsunami and when it will hit land. Once it hits land and people, in large numbers, start to escape from it, there is likely to be a mass of social media data (for example mobile phone conversations from those caught up in the evacuation) related to it. This data will be very confused and imprecise. However it still gives vital information about the passage of the tsunami and of its effects. The challenge is how to couple this mass of confused data with the more precise data gathered earlier.

What questions do we want to ask of Big Data?

The novel aspects of Big Data lead in turn to many challenges in how we deal with it, how we

visualise it, make speculations from it, model it, understand it, experiment on those systems which generate it, and ultimately how we might control those systems. The mathematical and scientific challenges behind these questions are as varied as they are important.

As examples of such questions we can include the following:

- Ranking information from vast networks in web browsers such as Google.
- Identifying retail consumer preferences, loyalty or even sentiment and making personalised recommendations.
- Modelling uncertainties and patterns in health trends for many individual patients.
- Monitoring health in real time, especially in the environment that 5G will lead to.
- Using smart data gathered from energy usage to optimise the way that the energy is then supplied to consumers.

Essentially we are trying by doing this to detect patterns in this huge mass of information. Humans have been doing this forever, which is how we make sense of, and survive in, the complex environment which we all live in. Long may this continue! However, the very scale of big data makes automation necessary and this, in turn, necessarily relies on sophisticated mathematical algorithms. This neatly brings us onto the next section.

The mathematics of Big Data, and how some of its algorithms work.

It is fair, I think, to say, that many of the future advances in modern mathematics (together

with theoretical computer science and related topics such as Machine Learning) will either be stimulated by the applications of Big Data or will driven by the need to understand Big Data. Of course many existing mathematical and statistical techniques (some of which until recently were considered as 'pure mathematics') are now finding significant applications in our understanding of Big Data. According to Andreas Weigend, formerly the chief scientist at Amazon.com, “It’s like an arms race to hire statisticians nowadays.” He said further that “Mathematicians are suddenly sexy.” This is a most exciting thing to hear said if you (like me) are a mathematician! Indeed retailers such as the North American company *Target*, are now one of the major employers of statisticians and other types of ‘data analysts’.

A loose description of the use of statistics in the context of Big Data is the extraction of useful information by combining dodgy models with even dodgier data (and then using this information to predict into the future). Early models of processes were linear, in that they assumed a linear relationship of the form between an input

x and an output y. Usually the measured data z was then given by the output y plus a degree of noise. Provided that the noise was not too great, it was possible to make reliable estimates of the coefficients m and c in the linear model above. The beauty of this is that it is then possible to use these values of m and c to *predict* the output from a completely new value of an input x. Linear models such as this have a long track record of giving useful predictions. They are used, for example, by utility industries to help predict future demand for supplies. In the form of Kalman Filters they are used to continuously process data as it arrives. This technology is vital in, for example, mobile phones and GPS systems. More modern algorithms are based on fitting more sophisticated *nonlinear models* to data. An important example of such are the *neural networks* used in many machine learning systems to find patterns in data. In a manner very similar to the linear models described above, neural networks are firstly *trained* on a large data set, and then tested on another, before being used to aid in future decision making. One application of them might be to aid a doctor making a diagnosis of a patient's health given certain medical data. Another application is their use in face recognition by (for example) social media websites. However such algorithms have a potentially sinister side. They could, for example, be used to make decisions on who to hire for a job based on a set of answers to a questionnaire. Our understanding of the behaviour of nonlinear systems is, however, very incomplete. Thus, big decisions affecting people's lives, could potentially be made by algorithms, which are ill understood and for which much research is needed before they can be completely trusted.

Another example of a mathematical theory of huge importance to Big Data is the mathematics of **network theory**. This describes objects, described by nodes, and the connections between them, described by edges. Network theory explains the connections between the objects (often formulated through an adjacency matrix) and allows us to search the network for connections between the data by finding structures in the adjacency matrix. Such algorithms are used by search engines, such as Google. Indeed, at the heart of Google is a very fast algorithm for finding the leading eigenvector of the adjacency matrix describing the connexions in the World-Wide-Web. Network theory and can also describe (via differential equations) the movement of information around the network itself. As an example the nodes could be computers or website on the computer, and the edges, connections between the computers or links between the websites. The nodes could also be people and the connections to their 'friends' on Facebook or Twitter, or they could be mobile hand sets and the link a conversation or simply a close proximity which might lead to interference. This latter issue is particularly important, as with 7 Billion people in the world, there are a potential of conversations over a mobile phone network, each of which must not interfere with any other. Indeed, managing the mobile phone network (which is of course also used to download huge amounts of data) is a significant and growing application of the theory of graph colouring which until recently was regarded as firmly in the domain of pure mathematics. Other examples of networks, which lead to Big Data include:

Social networks: Friendship, sexual partners, Facebook and other social media,

Organisational networks: Management, crime syndicates, Eurovision,

Technological networks: World-wide-web, Internet, the power grid, electronic circuits,

Information networks: DNA, Protein-Protein interactions, citations, word-of-mouth, myths and rumours,

Transport networks: Airlines, food logistics, underground and overground rail systems,

Ecological networks: Food chains, diseases and infection mechanisms.

For many more examples of networks see the review article by Newman [4].

Network theory can be used to address more of the many questions related to Big Data as described above. Specifically network theory based algorithms can be used to segment data and find clusterings in data. Such information is vital in data mining and pattern recognition, and is especially important to the retail industry, segmenting graphs (which can include images) into meaningful communities, finding friendship groupings, investigating the organisation of the brain, and even finding **Eurovision voting patterns**. These voting patterns look at links described by the process of *who voted for who.* A careful analysis of the resulting network shows that the rumours of voting blocks really are true! [5]

Such analysis can also help with the very significant problem encountered in many applications of linking databases with different levels of granularity in space and time

Equally important is the question of how connected the network is, and what is the shortest length of a path through the network. This is essential for efficient routing in the Internet, interpretation of logistic data, speed of word of mouth communications and marketing. Network theory is also essential in searching for influential nodes in huge networks (of huge importance to search engines), and in finding the resilience of a network, which can then be used to break a terrorist organisation, or even to stop an epidemic.

Of course, network theory, whilst important, is just one of a variety of the many mathematical techniques used to study Big Data. As much of Big Data (particularly that found on social media websites) takes the form of **images**, mathematical algorithms, which classify, interpret, analyse and compress images are extremely important in all Big Data studies. They are already used by social media websites, which make extensive use of

neural networks and other machine learning algorithms, to implement them. Linear signal processing, and related statistical methods have long been used to analyse and interpret images. But there has recently been a significant growth in novel mathematical algorithms, drawing again on ideas in 'pure mathematics'. Some of these algorithms, particularly those for image segmentation or denoising, are based on the analysis of nonlinear partial differential equations, leading to some powerful and unexpected applications of such obscure areas of analysis as the p-Laplacian. Algebraic topology plays a very useful role in classifying images, and in particular the field of *persistent homology* [6] can be used to find ways of classifying objects in an image which do not depend upon the orientation, or even the scale of the object. Cohomology and tropical geometry, in particular combinatorial skeletal allow for a different form of object classification. Finally, techniques from category theory can be used to 'parse' an image to see how the various components fit together, and also (in the context of machine learning) to allow for machines to 'perceive' what the objects are in an image and to make 'reasoned' decisions about it. One application of this is the technology already used (by such organisations as social media networks and the Home Office) to recognise people's faces. Somewhat more worrying is related technology used to detect people's emotions (with one application the means of detecting a potential terrorist in a large crowd).

A recent and exciting development in the mathematical analysis of Big Data, due to Emmanuel Candes, Justin Romberg, Terence Tao and David Donoho [7] is the area of *compressed sensing*. Traditional signal processing has used Fourier or wavelet based methods to represent data, and compression is the achieved by a suitable truncation of this representation. In contrast, compressed sensing aims to exploit sparsity in the data and to achieve compression by direct sampling. (One mechanism for doing this is to use more 'blocky' representations of figures using piecewise constant representations. Compressed sensing is finding very important applications including the representation of large data sets arising in medical applications as well as the more domestic application of restoring old photographs.

Big Data is of course also a significant driver for advances in computer science and machine learning, and the development of novel computing algorithms. These include machine learning, encrypted computation, (which relies heavily on results in number theory), quantum annealing and quantum algorithmics. This is only a short list. Other areas of mathematics and computer science which have found applications in the study of Big Data include: segmentation clustering, optimal and dynamic sampling, uncertainty modelling and generalised error bounds, trend tracking and novelty detection, context awareness, integration of multi-scale models, real-time forecasting, data integrity and provenance methods, visualization methods, data compression and visualistion, dimension reduction, logic and reasoning, and optimisation and decision.

Essentially, watch this space! I am confident that we will see great advances in pure, applied and computational maths arising from these challenges. However we must not forget other areas of science also involved in understanding Big Data. A big reason why retailers such as Target can analyse our shopping habits is that, over the past two decades, the science of habit formation and related human behaviour has become a major field of research in neurology and psychology departments at hundreds of well financed institutions.

So, what can and cannot be achieved?

So, is this all hype, or can the analysis of Big Data really deliver the advances in understanding that it promises? In one sense it is 'easy'. There is no doubt that there are patterns in data and that by finding those patterns we can learn more. Hence the success of algorithms in Big Data in working out our shopping habits. However, I would argue that there is always the case of 'garbage in garbage out'. The algorithms are only likely to be as good as the models on which they are built, and not everything can be modelled easily. For example I was once taken to task by a computer scientist who asked me why I was even bothering developing careful physical models for the weather, when all that was needed was a 'simple' application of machine learning in order to forecast the weather with perfect accuracy. However, even apparently sophisticated algorithms such as neural nets have at their heart some fairly basic models, and there is no guarantee that nature will really behave that way. I personally do not in any way believe that Big Data will ever show us the perfect way to true love, or even indeed the weather in a weeks time.

The ethical dimension

So far I have concentrated on the mathematical, computational, and technological challenges of Big Data. This is after all my profession. However, the ethical challenges are arguably greater, and here I have much less to offer by way of a professional opinion. The key problem that I see with the use of Big Data is a huge loss of privacy. There is now data your age, whether you are married, whether you have children, where you live, your social class, your estimated salary, what credit cards you use and what Web sites you visit. Data is also available on your ethnicity, job history, financial status, health, education, social life, food intake, political leanings, reading habits giving, religion, health and the number of cars you own. All this can be used in 'behavioural research' in an attempt to work out how and what we are thinking. Not only is this an invasion of privacy, we can, and should, argue strongly about the ethical dimension of companies using this information to make decisions about us, informed by algorithms few of the users of them understand. On the contrary side, the creators of these algorithms (and I confess that I am one of them), do not necessarily have the ethical and philosophical training to

deal with the challenges that they raise. There is, I think, a pressing need for mathematicians, lawyers, and policy makers to work closely together to tackle the huge ethical issues, which the use of Big Data brings.

The UK response

The UK government has responded positively to the importance of funding mathematically focused research into Big Data. To this end the Engineering and Physical Sciences Research Council (EPSRC) (which is the rough equivalent of NSERC or NSF) has committed around £40 Million to found the *Alan Turing Institute* (ATI). This is a collaboration between the founding partner universities of Oxford, Cambridge, Edinburgh, Warwick and University College London (UCL) together with non-academic partners including GCHQ (the UK equivalent of the NSA) with Andrew Blake from Microsoft Research as the first director.

The site of the ATI will be the British Library close to St. Pancras station in North London. See the website https://turing.ac.uk/ for more details.

According to this website

The work of the Alan Turing Institute will enable knowledge and predictions to be extracted from large-scale and diverse digital data. It will bring together the best people, organisations and technologies in data science for the development of foundational theory, methodologies and algorithms. These will inform scientific and technological discoveries, create new business opportunities, accelerate solutions to global challenges, inform policy-making, and improve the environment, health and infrastructure of the world in an 'Age of Algorithms'.

I expect to see similar developments in many other countries in the near future.

Some References

[1] T. Fielden, *Maths becomes biology's magic number,* (2016), BBC website, http://www.bbc.co.uk/news/science-environment-37630414

[2] C. Duhigg,, *How Companies learn your secrets*, (2012), New York Times, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=3&hp&

[3] D. Willetts, *Eight Great Technologies*, (2013), Policy Exchange.

http://www.policyexchange.org.uk/images/publications/eight\%20great\%20technologies.pdf

[4] M. Newman*, The structure and function of complex networks,* (2003), SIAM Review

45, 167-256.

[5] D. Fenn, O. Suleman, J. Efstathiou, N.F. Johnson, *How does Europe Make Its Mind Up? Connections, cliques, and compatibility between countries in the Eurovision Song Contest*, (2006), Physica A: Statistical Mechanics and its Applications, **360**, 576--598.

[6] H. Edelsbrunner, *Persistent homology in image processing,* (2013),

in: Graph-based representations in pattern recognition, Proceedings of the 9th IAPR-TC-15 International Workshop, GbRPR 2013, Vienna, Austria, May 15-17, 2013., Springer.

[7] E.J. Candes, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements}, (2006), Communications on Pure and Applied Mathematics, **59**,

1207--1223.