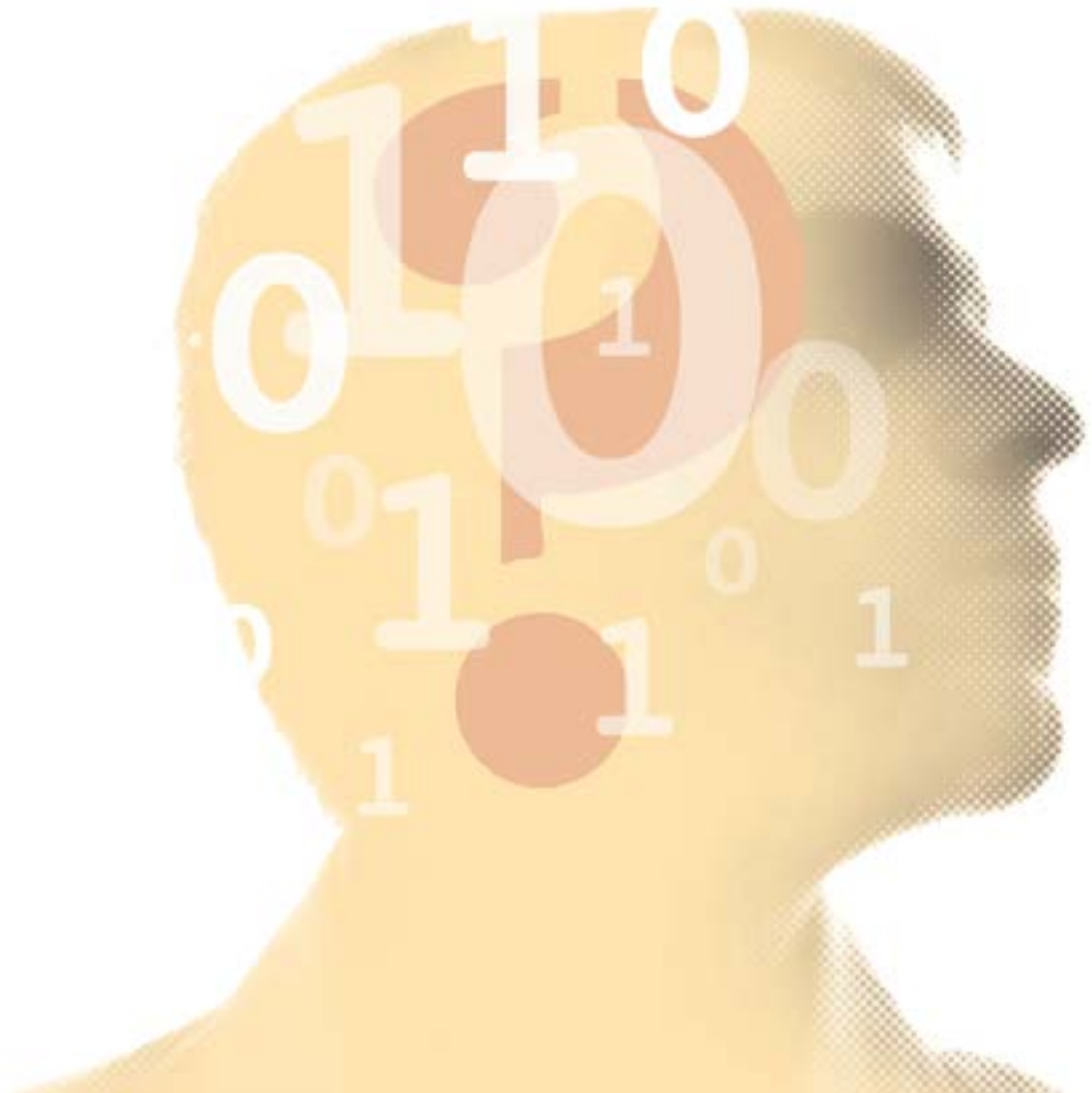
Big data and analytics – what's new?



ICAEW Making information systems work initiative

The IT Faculty is the focal point for ICAEW's activities on technology and the future of the accountancy profession. The faculty's work focuses on four transformational technology trends for accountants – Artificial intelligence, Blockchain, Cyber security and Data (ABCD). We take a research-based approach to understanding these technologies, how they apply across different areas of the accountancy profession and what impact they may have in the future. Our work brings together leading thinkers and practitioners from business, research, technology and public policy through panel discussions, reports and lectures. Reports include *Providing leadership in a digital world*, *Artificial intelligence and the future of accountancy*, *Blockchain and the future of accountancy*, *Audit insights: cyber security*, *Big data and analytics: what's new?*, and *Big data in Chinese businesses: international perspectives*

The IT Faculty's thought leadership programme, Making information systems work, looks at how technology is transforming the way that we do business and interact with each other. It draws on three themes which are essential to the success of IT – value, trust and standards. Previous reports include *Measuring IT returns*¸ and *Building trust in a digital age: rethinking privacy, property and security.*

For more information on the IT Faculty, please visit icaew.com/joinitf

For more information on the IT Faculty and how to get involved, please visit icaew.com/itfac or contact Richard Anning at richard.anning@icaew.com, or on +44 (0)20 7920 8635.

ISBN: 978-1-78363-377-7

1. Introduction

OVERVIEW

- The trend of big data is being propelled by enormous growth in computing power, new sources of data and the infrastructure to enable innovative knowledge creation.
- Applying analytics to big data creates many opportunities for businesses to gain greater insight, predict future outcomes and automate non-routine tasks.
- We need to ensure the use of big data and analytics is appropriate and subject to robust challenge, especially where predictive models are being relied upon.
- This requires greater skills in statistics, both in specialist disciplines and across a wide range of business functions that are using big data.
- We also need new thinking about the ethics, governance and regulatory framework around big data to ensure sufficient transparency and confidence in its use.
- But all businesses can get started with big data by asking good questions about their operations, strategy and stakeholders, and understanding how different data can help to answer them.

There has been a lot of talk about big data and analytics in recent years, but what is really new about it? After all, businesses have always used data and information to support decision making and manage operations. Indeed, some argue that information flows define what organisations do, how they are structured, and how they generate value.

When analytical techniques are applied effectively to big data, businesses can potentially achieve many incremental improvements – personalised services, optimised operations and better risk management, for example. Furthermore, the high pace of technology development and the sheer volume and breadth of opportunities to use data will increasingly influence the business environment and provide opportunities for disruptive new business models.

As a result, it's important to build our understanding around what big data means, what analytical techniques can do and how businesses can exploit these trends. This short report considers what is new and not so new about big data and analytics. It aims to inform decision makers in business and government about the opportunities and risks in this area.

Our framework for analysis is shown in the image below and focuses on three questions.

- What's creating big data?
- What are the opportunities and risks?
- How do we exploit big data?

FIGURE 1

2. What's creating big data?

There is no doubt that the world is producing enormous amounts of data. But 'big data' isn't just about volume. After all, scientists and industries such as banking have been coping with very large amounts of data for many years. Big data is also about complexity and speed, and is often characterised by the '3 Vs' – large volumes of data, high-velocity data flows, and a wide variety of data, especially unstructured and semi-structured data such as text and images.

The trend of big data is being propelled by three factors:

- growth in computing power;
- new sources of data; and
- infrastructure for knowledge creation.

These three elements result in significant advances in machine learning, a well-established field in artificial intelligence which is increasingly applied in a wide range of business contexts.

COMPUTING POWER

The core enabler of big data is the enormous growth in computing power and storage in recent years, which is making possible the capture and processing of entire data sets, regardless of their size and complexity. This is often described in terms of exponential growth in computing power.

MAKING SENSE OF EXPONENTIAL GROWTH

The impact of exponential growth is commonly illustrated through an ancient Chinese story about a chessboard. As a reward from the emperor for a particular task, a man puts one grain of rice on the first square of a chessboard and asks for the emperor to double it for each square – going from one to two to four and so on. While this seems an innocuous request, which is quickly granted, the rice quickly mounts up. By the last square on the chessboard, the man is owed 18,446,744,073,709,551,616 grains of rice. This story illustrates that while growth by doubling starts off with small changes, it soon leads to very large increases which can become difficult to comprehend.

In their book *The Second Machine Age*, Erik Brynjolfsson and Andrew McAfee argue that this is exactly what we are now seeing with computing technology. Rules of thumb such as Moore's Law show that all kinds of computing capabilities have been doubling every year or two since the early days of business computers in the 1960s. Brynjolfsson and McAfee argue that we are now effectively in the second half of the chessboard, which is resulting in very large improvements in short periods of times.

The cloud computing model further supports the widespread use of big data. Cloud computing is based on a model of pooling and sharing computing resources across a business (private cloud) or between a number of different customers (public cloud). By using a cloud, a business doesn't need to buy all the computing resources it might use; it simply accesses them when needed. Therefore, the cloud model potentially provides a business with access to vast computing resources on an efficient and flexible basis.

Software advances have complemented these developments in processing and storage capability. For example, new types of software support large and unstructured data sets better than traditional database management systems. Software such as Apache Hadoop supports the management of very large data sets by splitting the processing between many computers. Capabilities in handling unstructured data, such as video and text, have improved greatly. Visualisation tools have also progressed, enabling better analysis and presentation of data.

DATA SOURCES

This increase in computing power is making it economically viable to collect and process data from many new sources, such as the following.

- The internet provides a variety of clickstream data, such as searches, sites visited and goods viewed as well as actual transactions.
- Social media has created new types of data, including status updates, comments and likes, photos, videos and networks of contacts.
- Mobile technology is providing more opportunity to create social media and internet data, and generates new data about the location of individuals.
- Open data refers to the release of large amounts of primarily public sector data, such as geo-spatial data, transport data, government financial data and public service data.
- The 'internet of things' is the embedding of computer chips and sensors in physical assets (including machines, buildings, domestic appliances and clothes), all of which generate data.

As businesses increasingly use digital technology in areas such as sales and marketing, customer management, supply chain and internal communications, they are also generating more data internally which they can use. Furthermore, the improvements in the management of semi-structured and unstructured data enable businesses to make better use of a variety of existing and new data sources such as email and text, CCTV, pictures and voice.

Consequently, we are seeing a massive 'datafication' of activities, for example, goods that we look at but don't buy, our daily travel routes or photos we take. While these activities have always happened, it has never been technically possible or economically viable to capture and analyse data about them on a systematic basis.

INFRASTRUCTURE

The digital infrastructure has enabled new types of collaboration and knowledge creation, as evidenced through trends such as crowdsourcing and open source software. This sharing of knowledge has brought together new communities and led to insights in data from unexpected places. Sometimes, insights have come from data specialists who know nothing of the topic but can spot patterns in data. Other times, insights have come from domain specialists who really understand the field and have used fairly basic data techniques to solve problems. But the flexible nature of the digital environment enables all kinds of new knowledge sharing and creation.

The emphasis on applied research, especially in conjunction with commercial use and research challenges, has also enabled significant progress in specific fields. Advances in automated language translators, for example, have flowed from the insight that word-for-word translation is not very effective – in many cases, a single word translates into a number of words. As a result, researchers focused instead on phrase-to-phrase translation, and this has been a far more successful approach. Therefore, advances have come from new understanding of specific problems rather than breakthroughs in general theory.

IMPROVEMENTS IN MACHINE LEARNING

Machine learning is a technique in artificial intelligence whereby a computer teaches itself the answer to a question. This approach contrasts with most computer programming, which predefines rules and lets the computer calculate the best answer based on those rules.

Although rules-based programming has worked well in many cases, it has not always been possible to define rules clearly. Recognising language or images, for example, is usually complex and nuanced, with as many exceptions as rules. Similarly, where tasks are deeply grounded in tacit knowledge, rather than repeatable routines, they have not easily been described and therefore computerised.

Machine learning finds patterns in data sets and matches new pieces of data to them on the basis of probability. By doing this process over and over again, the computer can validate and improve the accuracy of the model and identify the most likely answer to the question. Therefore, it is not trying to answer the question on the basis of logic, understanding or following the rules. It tries to learn the answer through matching data to patterns and working out what the rules might be.

This approach relies on advanced models built on sophisticated algorithms. But it also benefits from having very large data sets – the more data points there are, the more times the model can run, learn and test the accuracy of its results. For example, another aspect of improved automated language has been a massive increase in the data available to run through the models. And as errors are corrected, this generates even more data and leads to greater accuracy. Consequently, while machine learning techniques have been around for a long time, they have improved greatly in recent years.

These models rely particularly on Bayesian thinking about statistics. The Bayesian approach aims to avoid false positive correlations by building on prior beliefs and probabilities and refining the output based on new evidence. Therefore, it takes a more contextual and dynamic approach instead of simply relying on the presence of strong correlations. While Bayesian thinking has been criticised for many years as being subjective, it is at the centre of most machine learning and is thought to produce more accurate results.

3. What are the opportunities and risks?

These trends combine to change the economics around data, and enable businesses to use data in ways that were not previously possible or viable. We separate out three broad ways in which big data is being used:

- to gain insights;
- to predict the future; and
- to automate non-routine decision making.

In each of these areas there are opportunities to create value for a business. However, there are also risks in using data in these ways and care needs to be taken to avoid unjustified conclusions, ensure appropriate reliance on predictive models and manage the impact of growing automation.

INSIGHT

Businesses can use big data to enhance their understanding of their operations and stakeholders, including the following.

- Using new sources of data to gain deeper understanding, for example using more granular data about customers to understand their preferences, activities and location.
- Exploiting the real-time nature of big data to improve services and operations, for example through personalising responses and offers.
- Applying analytics to gain new insights and interrogate entire data sets eg,
  – recognising new associations and patterns;
  – linking data from disparate sources; and
  – identifying exceptions, unexpected behaviour and outliers.

This provides many opportunities to improve customer management, supply chain management and risk management. It also provides opportunities for regulators to focus their resources more effectively and identify what needs further investigation.

REGULATORY EFFICIENCY
The UK Government's tax authority, HMRC, has been at the forefront of using big data to combat tax fraud. Its Connect system has won industry awards and is credited with generating £1.4bn additional tax revenues in the course of its first year. The investment required was just £45m.

It works by drawing on data from 28 different sources and enabling sophisticated analysis to identify outliers and anomalies. As well as data from tax returns, it uses data from the internet, social media sites, land registry records, international tax authorities and banks. Inspectors have even used images from Google Street View to validate information about properties. The ability to connect together disparate data about individuals is seen as the key to success. While this was always possible, the time involved in making such connections manually was prohibitive.

These new capabilities can be extremely powerful, but there is nothing radically new about the concept of using data to give us new insights. There are also well-established risks of using data in this way: for example, management information systems have always been hampered by data which is inaccurate, inconsistent, duplicated or out of date. These problems can be significantly amplified by big data, as many of the new sources of data can be unreliable or become outdated very quickly, such as social media.

Traditional responses to poor quality data emphasise cleansing data, or disregarding it entirely where the quality is very poor. But big data commentators argue that the sheer volume of data makes granular quality far less important. Analysis will still show the general trend, even if individual data items are of

variable quality. In order to make the most of opportunities with the data, it is argued, we instead need to work some degree of ambiguity around the accuracy of some data.

However, there is a trade-off between data volume, speed and granular quality, and there will be different conclusions depending on the specific context. Where data is being relied upon to make important decisions about specific individuals or organisational resources, ensuring appropriate levels of quality is likely to remain vital. By contrast, where analysis aims to identify trends, or respond quickly to customer demands, some data inaccuracy is more likely to be acceptable. Decision makers need to understand the standard of quality required in different contexts and ensure that the data used meets that standard.

Another risk relates to the selection of data and parameters of the analysis. One advantage of big data techniques is that they enable the analysis of entire data sets, rather than samples of data. This may allow users to spot new patterns or outliers. It also overcomes the risks of using smaller data samples. While the analysis of samples is valid, the samples need to be selected entirely randomly and have no bias. Otherwise the results are skewed and should not be extrapolated to a wider population. By using entire data sets, big data can avoid these sampling risks.

But all data sets, to some extent, represent a specific selection of data and therefore care needs to be taken to ensure that broad conclusions are justified. There may be other relevant data that is not captured in the data set, for example data from an earlier period or data about individuals who do not participate in particular activities. Social media data is often cited in this context – researchers may be able to interrogate entire data sets, but they are not representative of the entire population.

Care also needs to be taken to understand what the data really tells us. Big data may highlight new associations and patterns but they may be spurious. The urban myth of a correlation between buyers of nappies and beer at particular times of the day is well known, but it's not clear what insight that really gives. There are also many statistical traps to be avoided, such as relying on averages which hide large variations.

PREDICTION

Big data analytics particularly focus on the prediction of future outcomes and embedding predictive models in business operations. While businesses have always been trying to predict the future, big data opens up new possibilities. New sources of data mean that it is possible to make predictions in new areas and machine learning techniques increase the accuracy of prediction models. As a result, greater reliance can be placed on models.

Search engines, for example, predict the information that will be most useful to the individual. That prediction will be based on other links to the material, the proximity to the search term and the extent to which others have clicked on the link. It may be personalised, based on location, gender, age and previous search history, where these are known. Increasingly, search engines predict what you're searching for, as you type.

Many customer service functions personalise services based on predictions about individual customers. Recommendation engines use information about previous transactions, similar customers and the qualities of products and services to predict products you are likely to enjoy. They may be able to predict when customers are likely to shift loyalties elsewhere or how they will respond to a particular offer.

Other businesses are making use of predictions about the wear and tear of physical assets. This enables efficient scheduling of repair and maintenance to minimise faults and prolong the life of assets. Sports teams and commentators are using data to predict the future performance of individuals and teams.

So, there are many uses for prediction models and they are becoming increasingly widespread across all kinds of businesses. But there is a big step from merely gaining insight from data to predicting future activities. Can such models really be relied upon?

4. How do we exploit big data?

At the heart of this debate is the risk of relying on correlations rather than understanding causation. Big data predictive models use associations between elements (correlations) and patterns to predict future outcomes. They are not necessarily based on any deep understanding of why particular outcomes occur.

While correlation models can be extremely accurate, there are dangers of relying entirely on correlations with no reference to underlying theories of causation. The links found may be spurious and result in false positives or negatives. Systems based on correlations can be gamed to get a particular answer. If conditions change but the underlying model assumptions remain the same, the model can suddenly produce highly inaccurate predictions.

The debate over the relative importance of understanding correlation and causation is not new. However, Chris Anderson took it to a new level when he declared that big data means 'the end of theory'– we no longer need to understand why something happens when we can predict whether it will happen again, often with high degrees of accuracy. Few would agree with such bold statements though, and the case of Google Flu Trends demonstrates both the opportunities and dangers of predictive models.

GOOGLE FLU TRENDS
Google created a stir when it built a model to predict outbreaks of flu, based on correlations in its search data. Searches relating to flu symptoms were a good early indication of a flu outbreak and the Google model was extremely accurate in its predictions. It was also available earlier than the official data, which had to be gathered from various sources and consolidated. Google's predictions, using real-time search data, were produced a week earlier. Consequently, this was hailed as a great example of the power of big data.

However, there have also been problems with the model. During a flu epidemic in 2013, for example, the model overestimated the incidence of flu by 100%. This was due to many more people checking flu symptoms even though they were not actually suffering from flu. Therefore, the correlation – between looking up symptoms and having flu – no longer worked. You needed to understand why people were looking up the symptoms in order to make an accurate prediction.

While the model was subsequently fixed, the accuracy has again reduced as search habits and Google's algorithm have evolved. It is thought that, for example, Google's autocomplete feature for typing search terms encourages people to search for flu symptoms when they don't have flu.

Google may well be able to refine the model in the future and increase the accuracy to regain previous accuracy levels. However, it illustrates the risks of relying on correlations where conditions or behaviour change.

In order to understand when predictive models are most valuable, we need to consider when having more data leads to more accurate predictions.

Areas which have seen substantial improvements in prediction include short-term weather forecasts, sporting performance and political polling. The path of Hurricane Sandy, for example, was accurately predicted, which enabled better mitigation of the effects of the storm. The US data journalist Nate Silver accurately predicted the outcome of every state in the 2012 US Presidential election through data models, even though he had no prior experience of politics or polling – his background was sports analysis.

However, there are areas where prediction has not improved, such as with earthquakes or economic growth. In the cases where our understanding of causal factors is poor, the environment is particularly complex, or the pace of change is high, prediction is likely to remain difficult, regardless of the amount of data available. In these cases, having more data and identifying new patterns can help us to improve our understanding, but may not increase our ability to predict accurately.

AUTOMATION

The third broad use of big data builds on these predictive capabilities to automate non-routine decisions and tasks.

Driverless cars have received a lot of publicity as the technology has improved rapidly in recent years. They rely on a wide variety of data from sensors, mapping applications, satellites and others to navigate their way, and they are incredibly accurate in their decision making – very few accidents have been reported. Indeed, they are increasingly seen as a way to improve road safety, as they remove the risks related to human fatigue, carelessness and poor-quality driving.

We also see increasing automation of professions such as law and medicine, as computers take on more tasks which have previously been difficult to computerise on the basis of predefined rules.

For example, healthcare companies are starting to exploit machine learning techniques to automate medical diagnosis. Computers can hold far more information than humans and can quickly and accurately work through the possible scenarios based on the presenting symptoms to identify the most likely cause. Disciplines like radiography are already heavily automated as computers can spot patterns or exceptions more accurately. In the legal profession, models can scan through vast amounts of potential evidence much more quickly and accurately than humans. As a result, the discovery phase of cases is becoming more automated.

There are clear risks where automation goes wrong. How do you know something has gone wrong? Who is responsible? How do you correct errors? How do you manage systemic risks?

But there are also deeper issues around how far automation can go – when are computers better decision makers and when does human knowledge remain vital?

While there has been a lot of publicity around big data, it has been exploited mainly by big companies, often those at the leading edge of data and technology, such as internet companies or major retailers. Many other businesses, especially SMEs, are a long way from utilising big data.

This section outlines some of the challenges and priorities in exploiting big data for three distinct groups:

- businesses (split into businesses getting started with data and those which have more experience with data and are looking to maximise its impact);
- accountants; and
- policymakers.

BUSINESSES GETTING STARTED WITH DATA

Many businesses struggle to know where to start with data, especially big data. For smaller businesses in particular, the concept of big data may appear to be irrelevant, as they may not have very much data. Alternatively, they may have lots of data but struggle to identify what is useful. This is made harder by the way that the real value of data often comes from reuse or in combination with other data.

Rather than focusing on data, businesses can start by focusing on good questions that will help them be more successful. While there can be value from experimenting with data and seeing what turns up, the broad advice from experts is to frame useful questions and focus analysis on answering those questions.

Once the questions are clear, management can then identify all the possible sources of data that could help it answer those questions. Some of those may be existing sources of internal or external data. Some of the data may not yet exist but could easily be collected by tweaking systems or processes. Some data may need longer-term planning to collect and therefore require some cost-benefit analysis.

HUMAN DECISION MAKING

We place high value on human knowledge, but we can also be very poor decision makers. In his book *Thinking, Slow and Fast*, Nobel award-winning psychologist Daniel Kahneman outlines many biases that result in poor decision making. For example, confirmation bias leads us to look for information which supports our existing opinion and ignore inconsistent information. We often see patterns where they don't exist. We usually fail to appreciate risks properly and overestimate the likelihood of rare but catastrophic events, while underestimating more mundane risks.

As a result, we often make very fast and very poor decisions which can be much better done through a computer which does not exhibit such biases.

But models cannot replace human knowledge and judgement entirely. We still have the ability to make very good decisions based on reason and complex pattern recognition. Decisions involving emotional intelligence and observation of behaviour will remain human. People will be needed to interpret data and relate it to real-life problems. People will also continue to be the key source of creativity and innovation. The real challenge is to identify the higher-order problems where human knowledge and capacities are important and can work with computer technologies to deliver the best overall solutions.

Businesses also need to access all the necessary skills. These are spread across three broad areas:

- statistical skills to build the algorithms and understand the robustness of models;
- data and technology skills to extract and manipulate the data; and
- domain knowledge to ask the right questions and gain insight from the analysis.

Some of these skills can be bought from third parties and smaller businesses are likely to benefit from data and analytics service providers, rather than attempting to build high levels of technical skill internally. The cloud model also provides a way for smaller businesses to access the technical resources required without investing in substantial hardware themselves. However, exploiting big data also requires significant internal resources to ask the right questions and interpret the results based on the specific context of the individual business strategy and operations.

As a result, businesses are likely to make most progress by starting with small, targeted projects which have tangible benefits and then build on successes to exploit further sources of data.

BUSINESSES MAXIMISING THE IMPACT OF DATA

Where businesses have more experience in exploiting data, the challenges shift to embedding a data-conscious culture and building the right structures to maximise the opportunities and manage the risks.

There is, for example, a shift from traditional decision-making cultures, sometimes termed 'highest-paid person's opinion', to cultures which are more reliant on data. Indeed, academic research suggests that companies which are data-driven are likely to have higher output and higher productivity than companies which are not making use of new technologies.

However, there are tensions in becoming more data driven. For example, data can be overwhelming and result in paralysis. It can also stifle innovation and risk-taking.

GOOGLE AND DATA-DRIVEN DECISION MAKING

Google is the poster child for data-driven decision making. Indeed some of the non-engineering senior managers have left because they found it too difficult to work in a culture which was entirely focused on data. In one well-publicised example, their visual design leader, Douglas Bowman, resigned because of its 'design philosophy that lives or dies strictly by the sword of data'. As an example, the company insisted on running an experiment to test which shade of blue (out of a possible 41 shades) got better responses from users. Design decisions needed to be backed up by data; human experience and judgement was not enough.

In his resignation blog, Bowman said, 'When a company is filled with engineers, it turns to engineering to solve problems. Reduce each decision to a simple logic problem. Remove all subjectivity and just look at the data ... And that data eventually becomes a crutch for every decision, paralysing the company and preventing it from making any daring design decisions.'

Another challenge relates to the organisational structure around big data, given the need to access many different skills. Building interdisciplinary teams is a key element of success and a variety of approaches have been taken to achieve this. Some businesses have set up centres of excellence that lead initiatives and share experience and best practice, working with other business areas as needed. In other cases, functional areas, such as marketing or operations, which are making real use of big data, may lead organisational capabilities. IT functions may also take the lead in some cases.

But many businesses struggle to build cross-organisational ownership and sponsorship of data projects and investments. Business units are typically focused on their own projects and data, which means that building common tools and data definitions and sharing skills and knowledge can be challenging.

Furthermore, to what extent do businesses have the controls and governance in place to ensure that data models are used appropriately and in a way that will create sustainable value?

There are many governance issues to consider here. When computer programs are based on rules, it is possible to unpick what has been done, identify any errors and exercise control over the outcome. However, when outcomes derive from algorithms and machine learning, it is more complex. There may be many different correlations that are involved and the algorithm will evolve as it refines its models and hones the probabilities involved. Who will understand exactly how the algorithm works and the assumptions that have been made in the process?

As data increasingly impacts on people's lives, greater transparency and assurance may be needed over what algorithms are doing. Where, for example, mistakes are made these need to be identified and corrected. As data and algorithms increasingly contribute to corporate value, investors may want assurance around their long-term sustainability. There is also growing reliance on businesses 'doing the right thing' when using data or models, and consideration needs to be given to the ethical framework.

ACCOUNTANTS

Information and data are at the heart of what accountants do and therefore the accounting profession can engage with big data and analytics in many different ways.

At a high level, these trends may enable different fields to adopt the structured approach to data developed by the accounting profession. National accounts, for example, have historically been developed from the top down and not rooted in the real transactions of the economy, due to difficulties in extracting and processing more granular data. Big data techniques potentially enable bottom-up approaches which could be more accurate and grounded in real-time analysis, enabling better macro-economic analysis.

There are many opportunities for accountants to make use of big data and analytics techniques to enhance their contributions to businesses, such as:

- using predictive models and other sources of data to improve budgeting and forecasting;
- using more sophisticated outlier and exception analysis to improve internal control and risk management;
- improving the efficiency and quality of audit activities through analysis of whole data sets.

Accountants are also experienced in issues around quality of data and analysis. They are trained to have natural prudence and scepticism and therefore are well-placed to ensure the use and analysis of big data is robust.

However, this would require greater knowledge in the theory and practice of statistics than many accountants currently have. While the degree of knowledge would vary, depending on the specific role, accountants would need at least enough statistical knowledge to be an 'intelligent buyer' and ask good questions of suppliers or other parts of the business.

Growing automation also presents long-term challenges to the profession. While computerisation of mundane tasks may free up accountants to focus on more value-adding activities, it raises questions about the training of the profession. With fewer opportunities for hands-on experience and learning, the traditional training path may need to evolve to reflect this.

POLICYMAKERS

There are a number of different elements for policymakers to consider. Big data provides many opportunities to improve policy making, public services and government. Governments are some of the biggest users of IT systems and generate enormous amounts of data about citizens and services. This will increase as more government services move online, providing many opportunities to personalise and improve services, target resources and interventions and inform citizens. Big data can support evidence-based decision making, by enabling deeper analysis on the impact of policies. It can enhance democracy and transparency through the release and use of open data.

Policymakers should also be concerned about building the right skills across the economy to enable widespread use of big data. 'Data scientist' describes an individual who combines some of the skills highlighted in this paper to analyse big data. There are many concerns about a shortage in these specialist skills in the coming years. For example, a 2011 study by the McKinsey Global Institute predicted a skills gap of 140,000 to 190,000 people by 2018 in the US alone. Similar fears are expressed about the UK, which will severely limit the ability of businesses to exploit the technology.

However, it's not just specialist skills. Many workers will need to gain better statistical skills in order to make appropriate use of big data. This report has outlined a variety of dangers around big data, such as data quality, selection of data sets, and reliance on correlations. Business use of big data needs to be grounded in a good understanding of what the models are doing, the assumptions that have been made in building the models and their limits. Without better statistical skills across many business functions, there are significant risks of inappropriate reliance on data and models. This needs to be matched with a natural scepticism about what data is telling us and an ability to challenge data specialists on the results.

It also requires the ability to ask and frame good questions in a particular area. Creativity and imagination are therefore important characteristics. Fostering these capabilities and encouraging experimentation will be vital.

Furthermore, the regulatory framework needs to be carefully considered. In a world of big data, decisions about individuals will increasingly be made on the basis of patterns and profiling. Therefore, big data has deep social implications about when we want to prejudge people based on data about past behaviour, personal characteristics and similarities to others.

This is strongly linked to debates about privacy. Data profiling, especially where large amounts of personal data are aggregated together, provides very deep insights into individuals. The benefit of these activities are cheap (or free) personalised services, and to date, many consumers have been content with this trade-off. However, greater concern may be shown as analysis goes deeper into our activities and personal lives.

ANONYMISATION OF PERSONAL DATA

A common response to privacy concerns around big data is to emphasise anonymity. Where data is personally identifiable, it is subject to regulation such as the Data Protection Act, so it is stripped of any personal elements which can be identified, such as name. However, real anonymisation is difficult to achieve, especially where data from multiple sources is being combined.

Netflix provides a good example of failed anonymisation, when it ran a competition to improve its recommendation algorithm. It released a lot of anonymised data about its users and how they had rated different films. However, a couple of researchers then matched this data against publicly-available data on another website and were quickly able to re-identify specific individuals.

Therefore, while anonymisation can be an important part of a privacy strategy, it needs to be carefully applied and the risks understood.

5. Summary

The trend of big data is being propelled by an enormous growth in computing capability, our ability to capture, store and apply sophisticated analytics to data from many new sources, and innovative ways to share and develop our knowledge. This is creating many opportunities for businesses to derive greater insight, predict future outcomes and automate non-routine tasks.

But the use of big data and analytics needs to be appropriate and subject to robust challenge. There are many dangers around the quality of data, selection of data sets and construction of models that need to be properly understood when using big data. This requires greater skills in statistics, both in specialist disciplines and across business functions which are using big data.

We also need new thinking about the ethical and regulatory framework around big data, as it will increasingly impact on the lives of individuals and underpin customer service, innovation, quality and business operations. Businesses will need to have appropriate governance to manage the risks and ensure data is used in acceptable ways. Policymakers also need to consider the regulatory framework carefully, and encourage the range of skills needed to exploit big data.

But all businesses can get started with big data by asking good questions about their operations, strategy, and stakeholders and understanding how different data can help to answer them.

Further reading

Anderson, C. (2008), The End of Theory: *Will the Data Deluge Makes the Scientific Method Obsolete?* Wired Magazine, 23 June 2008.

Bowman, D. (2009), Goodbye Google. Available at http://stopdesign.com/archive/2009/03/20/goodbye-google.html [Accessed 13 October 2014].

Brynjolfsson, E. and McAfee, A. (2014), *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York: W. W. Norton & Company.

Brynjolfsson, E., Hitt, L. and Kim, H.H. (2011), *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* Available at http://ssrn.com/abstract=1819486 [Accessed 13 October 2014].

Cukier, K. and Mayer-Schonberger, V. (2013), *Big Data: A Revolution That Will Transform How We Live*, Work and Think, London: John Murray.

Gillon, K., Aral, S., Lin, C., Mithas, S. and Zozulia, M. (2014), *Business Analytics: Radical Shift or Incremental Change?* Communications of the Association for Information Systems, Vol. 34, Article 13.

Haghighi, A. (2012), *What it takes to build great machine learning products.* Available at http://radar.oreilly.com/2012/04/great-machine-learning-products.html [Accessed 13 October 2014].

Kahneman, D. (2012), *Thinking, Fast and Slow*, London: Penguin.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. and Kruschwitz, N. (2011), *Big Data, Analytics and the Path from Insights to Value*, MIT Sloan Management Review, Vol 52(2), pp. 21–31.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014), *The Parable of Google Flu: Traps in Big Data Analysis.* Science 343 (6176), pp.1203–1205.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh C. and Beyers, A. (2011), *Big Data: The Next Frontier for Innovation, Competition and Productivity,* McKinsey Global Institute.

Silver, N. (2012), *The Signal and the Noise: The Art and Science of Prediction*, London: Allen Lane.

ICAEW IT FACULTY

ICAEW's IT Faculty is a leading authority on technology and the finance profession. It provides its members with information that allows them to make the best possible use of IT and keep up to date with IT issues and developments. Membership is open to finance professionals with an interest in technology, to join visit icaew.com/joinitf

---

ICAEW connects over 147,000 chartered accountants worldwide, providing this community of professionals with the power to build and sustain strong economies.

Training, developing and supporting accountants throughout their career, we ensure that they have the expertise and values to meet the needs of tomorrow's businesses.

Our profession is right at the heart of the decisions that will define the future, and we contribute by sharing our knowledge, insight and capabilities with others. That way, we can be sure that we are building robust, accountable and fair economies across the globe.

ICAEW is a member of Chartered Accountants Worldwide (CAW), which brings together 11 chartered accountancy bodies, representing over 1.6m members and students globally.

ICAEW
Chartered Accountants' Hall
Moorgate Place
London
EC2R 6EA
UK

T +44 (0)20 7920 8100
E generalenquiries@icaew.com
icaew.com

Chartered Accountants Worldwide
Member

GAA
Global Accounting Alliance