A Future for Data Analytics in Ireland

A Summary Report on a one-day event held at the Helix, Dublin City University, hosted by the Insight Centre for Data Analytics

Friday, 16 September 2016

Edited by

Dr. Brian Mac Namee  
Insight Centre for Data Analytics,  
University College Dublin

Prof. Noel O’Connor  
Insight Centre for Data Analytics,  
Dublin City University

Prof. Alan Smeaton  
Insight Centre for Data Analytics,  
Dublin City University (Chair)

Contact:  
Prof. Alan F. Smeaton  
Insight Centre for Data Analytics  
Dublin City University  
Alan.Smeaton@DCU.ie
Background and Context

- Data analytics underpins many aspects of our society and it is therefore essential that researchers, policy makers and industry work together to help shape the direction of its evolution to realise the potential benefits to Ireland and globally.
- The document is a report of the main conclusions drawn from a horizon scanning exercise that took place at the “Future of Data Analytics Forum” held at the Helix, Dublin City University on 16 September 2016.
- Content and conclusions are based on contributions from almost 50 attendees representing research, industry and government agencies.

Key points for policy makers

- Data-driven decision making techniques can now be used to generate reliable, objective insights that can have a transformative effect on society.
- Ultimately, data analytics is about empowering humans to make more informed decisions by uncovering previously inaccessible actionable information.
- Automated decision making can only take place in a very small number of non-critical applications where there is no threat to human safety or no potential negative societal impact.
- It is important to ensure that a human understandable explanation is associated with every decision suggested by a machine - this is not only desirable but will be necessary with the enforcement of the EU legislation on personal data in 2018 (European Commission Regulation (EU) 2016/679).
- Industry values a strong and close relationship with academic research to access talent and to ensure bi-directional flow of knowledge, yet it is difficult for publicly-funded research to keep up with industry as large companies are investing significantly more funding in data analytics than most funding agencies.
- Data privacy and ownership is particularly relevant in Ireland given the companies based here whose business is about personal data and it is thus important that Ireland shows international leadership in this area.

Key points for industry

- Ireland has the potential to be world leader in a number of sectors, although some sectors such as those associated with the life sciences, have yet to experience the full potential of data analytics.
- While data analytics techniques may seem generic, when they are used they are specialised for specific applications and thus can vary significantly across different sectors.
- The hot topic in machine learning is called “deep learning” and with the right input data and the right problem to tackle, this approach can operate at human levels of performance.
- The growth of the Internet of Things (IoT) where IoT devices will generate or capture huge amounts of data, presents significant commercial opportunity. However this will need real time analytics and edge computing, where processing is carried out on the device as opposed to the cloud or sometimes via a combination of the two. This provides a potential scalable solution, but needs close cooperation between industry and academia to make this work in practice.
• Standardisation of data collection from devices so that it can be usable for all is still a long way off and should continue to be prioritised by industry.
• Challenging aspects such as privacy can present opportunities for new business models if companies “think outside the box” (e.g. payment models for the privilege of data privacy) and could cause significant disruption in the market.

**Key points for the research community**

• Deep learning is not a silver bullet and features some important weaknesses and has actually moved us further away from explanation-based analytics which becomes an important legal requirement in the near future.
• New approaches to learning from noisy or sparsely labeled data are required to address the challenges faced in any field involving subjective inputs from humans such as behavioural science, life sciences, etc.
• Bias is a key challenge for future research. Further work is needed to target new kind of algorithmic decision-making that can identify the sweet spot between security and accountability.
• We are still a long way short of delivering technological solutions for privacy management as will be legally required, and even when available we will face the problem of widespread deployment and adoption of such techniques.
• Other key areas for future research include:
  ○ Investigating the inherent vulnerability of deep learning under adversarial conditions
  ○ Investigating the integration of elements of control theory, particularly for data analytics targeting complex, real time systems
  ○ Developing approaches that can perform continuous learning upon deployment so that the learned models evolve over time.
Executive Summary

We live in a society which is characterised by the constant roll-out of new technologies for work, for our health, for entertainment, for leisure, for sports, in fact for all aspects of our lives. Many of these technologies are digital and they generate data trails from their use. Almost by stealth, we now find there is significant exploitation of this data for purposes which go well beyond the initial reasons for such data capture. Many of these uses are for good, though some have downsides of which we must be cognizant.

Data analytics is the term used to describe the process of categorising and analysing data in order to identify patterns and to visualise and take action as a result of those patterns. It underpins the exploitation of the massive amounts of data now available and is carried out in many organisations in Ireland, from multi-national corporations to SMEs and from government agencies to educational institutions. Many of the services we now avail of are either “born digital” and driven by data analytics or have shifted to be data-driven. These range from advertising (e.g. Google AdWords) to entertainment (e.g. the storylines on Netflix TV series) and from medicine (e.g. automated classification of x-ray images) to finance (e.g. algorithmic trading).

Ireland is investing in data analytics. The Insight Centre for Data Analytics is the largest single investment in a research program in the history of the State and has more than 200 PhD students at various stages of completion of their research. There are taught Masters programmes in data analytics in most of Ireland’s Higher Education Institutions and undergraduate degree programmes in Data Science and Data Analytics are under development or just launched.

In such an environment which is fast-changing, global and talent-dependent, keeping abreast of technical developments in the field is not enough, Ireland needs to be ahead of that developmental curve. Yet, no single individual or organization has the capacity, the breadth of expertise, or the multiple viewpoints to see where that developmental curve is taking us.

To make some inroads into this, almost 50 of Ireland’s leading researchers and industry practitioners in data analytics gathered at the Insight Centre for Data Analytics for a one-day Forum meeting to discuss some of the important topics in data analytics, with particular emphasis on Ireland’s position. The topics covered do not represent comprehensive coverage of the field, rather they result from the experience and interests of the research and industry leadership gathered. Nonetheless many valuable insights were revealed during the event. The many topics discussed during the meeting are grouped into 4 main areas, machine learning algorithms, platforms for data analytics namely edge vs. cloud hosting, data privacy, ownership and legal issues and finally, the important application areas for data analytics in Ireland.

This report presents a summary of the discussions, edited by 3 of the attendees and has been seen and approved by all attendees and some other contributors listed at the end of the document.
1. Introduction and Setting the Context

Data analytics is the term used to describe the process of categorising and analysing data in order to identify patterns and to visualise and take action as a result of those patterns. It underpins the exploitation of the massive amounts of data now available and is carried out in many organisations in Ireland, from multi-national corporations to SMEs and from government agencies to educational institutions.

Because of the importance of data analytics and the fact that it has only recently, and very rapidly, emerged as a technology underpinning so many aspects of our society, it is essential that we are not just familiar with it and how to use it but that we understand it deeply and that we help shape the direction of its evolution.

The Insight Centre for Data Analytics is Ireland’s largest SFI-funded research centre with more than 400 researchers across four co-lead institutes, Dublin City University, National University of Ireland Galway, University College Cork and University College Dublin. In September 2016 Insight hosted a series of tightly coupled events at Dublin City University including:

- The annual Insight student conference where more than 100 Insight students presented their work either as plenary presentations, posters or demos;
- A meeting of the Insight Governance Board;
- A joint meeting between Insight’s Scientific Advisory Board and Insight’s Industry Advisory Board.

The latter of these events brought together experts from across the world, from Ireland, from the West and East Coasts of the US, from Singapore and from throughout Europe, to Ireland, to help Insight focus on its future. With so many worldwide experts in data analytics in Dublin anyway, we seized the opportunity to build a one-day Forum event around their availability, specifically focusing on emerging trends in data analytics, from a technical viewpoint, and with a perspective on Ireland. Participation was composed of almost 20 senior industry representatives, almost 20 members of Insight’s research leadership, all active in data analytics work, as well as some from government agencies including Enterprise Ireland, IDA Ireland and SFI, and they joined some of the members of our advisory groups for the day-long event. The meeting was formed as an open discussion forum with no structured or formal presentations where all participants could, and did, contribute to a free flowing discussion, moderated by Prof. Alan Smeaton, Director of Insight at Dublin City University.

Capturing the wide range of discussion topics in a single short report is extremely challenging, but the discussion topics can be categorised into 4 areas: Data-driven analytics (the technical aspects of), platforms for data analytics (cloud, server or edge based), data privacy, ownership and legal issues, and finally emerging areas for data analytics. The Forum meeting was not intended to conclude on a set of action items to be followed but instead it highlighted topic areas, mostly technical, where research and further development is needed. As such, the Forum can be regarded as contributing to a research roadmap, a form of horizon-scanning which, for the participants at least and perhaps for the readers of this report, will help us all to start thinking about what is next in this broad suite of techniques.
which comprise the area of data analytics. By being aware of and by thinking about this near-term future, then that will surely help to maximise its use.

2. Data-Driven Decision Making

2.1 What is Data-Driven Decision Making?

In data-driven decision making analytical techniques are used to generate insights by recognising complex patterns hidden within data. This is done by transforming complex data and multi-modal data streams into actionable events for humans and machines alike. These insights enable people to make more intelligent decisions by removing much of the subjectivity and emotion that is usually used in decision-making. There are many analytics tools used in data-driven decision making usually summarised using the terms data analytics, data science or, more recently, big data. The specific technologies used include data mining, pattern recognition, visualisation, machine learning, classification and others and the discipline includes inputs from probability theory, statistics, logic, optimisation, computer science, search, and control theory.

The Forum spent a lot of time talking specifically about machine learning because, even though data analytics is a much broader topic, it is the recent advances in machine learning and their application to data-driven decision making that has accelerated the use of all these techniques in combination. Machine learning is particularly interesting to researchers from all backgrounds because of the way it changes the nature of how computers are used, as shown in the diagram below. In “traditional” computer programming, programmers write programs which are executed on data, using a computer, in order to generate an output. In machine learning, data is also fed into the computer but so also is the known outputs from this data, and the computer’s role is to learn the patterns which determine the outputs, to generate a model for the outputs, and this constitutes what we would traditionally have called a “program”.

The applications for data analytics and machine learning in particular are becoming ubiquitous. It started really in computer vision, trying to recognise content in images and video, and then moved to natural language processing for applications like determining
sentiment and translation to other languages, and is now applied to forecasting, games, data mining, robotics, navigation, and so many more.

There is lots of media speculation recently about machine learning and neural networks replacing human decision making leading to loss of jobs. Machine learning, especially deep learning, requires massive amounts of correctly labelled data to be used as training data and while some domains are data-rich, like advertising or internet searching and advertising, others like healthcare will never have these kinds of massive labelled datasets and so the Forum was firmly of the belief that this scaremongering is unfounded and that we need to embrace, learn about, and use data analytics techniques fully.

2.2 What can we do with Machine Learning?

The field of machine learning can be broadly categorised into three sub-fields: supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning an algorithm is used to learn a prediction model from a dataset containing a large number of examples and their associated predictions. This is achieved by inferring from the dataset the patterns that relate to different predictions that are possible in the dataset. For example, from a large set of past emails (examples) labelled as spam or non-spam (associated predictions) a supervised machine learning algorithm could be used to learn a prediction model that would be able to automatically categorise future emails as spam or non-spam. Supervised machine learning has been applied in areas ranging from recognising the contents of an image, predicting the future price of a stock, predicting the likelihood that a customer will buy a particular product, or making diagnoses from medical images.

In unsupervised machine learning an algorithm is used to automatically infer groups based on patterns in a large dataset, without the nature of these groups being known in advance. For example, a company's customer base could be automatically arranged into groups displaying similar behaviour. Example applications of unsupervised machine learning include customer segmentation, fraud detection, anomalous event detection in sensor networks and finding communities in social networks.

In reinforcement learning a decision making policy is learned and improved over time based on the successes and failures of decisions made so far. Example applications of reinforcement learning include controlling robot behaviours, controlling autonomous vehicles including cars and helicopters, automated players in games, and vehicle routing systems.

Most of the discussion of machine learning for data-driven decision making at the Forum focused on insights generated by models learned using supervised learning algorithms. Supervised machine learning algorithms are used in data-driven decision making in applications ranging from personalising advertising for visitors to websites, determining the likelihood that a customer will repay a loan for credit decisions, predicting the propensity of customers to renew a service or leave for another provider, automatically categorising documents and images, translating text between languages, and recognising the content of
images. The real advantage of using machine learning for these tasks is that models can be used to generate reliable, objective insights (and even make automated decisions) quickly and efficiently and in massive numbers.

The Forum discussed how the capability of machine learning systems has improved significantly in the last decade based on a combination of algorithmic improvements, the availability of massive datasets, improvements in processing power, and significant investment by industry in machine learning solutions to specific problems. This is particularly evident in the performance of models trained for tasks such as image recognition and voice transcription.

The forum also recognised, however, that there is still plenty of room for further improvements in machine learning - there is still plenty that we can’t do. While massive improvements have been seen in certain narrow domains (e.g. image recognition) the same is not the case across all domains. It is still often the case that it is not possible to achieve human-levels of prediction accuracy for machine learning models built for specific tasks. Similarly, many of the models we train are still only able to do relatively low level tasks. For example, the much heralded improvement in performance of speech recognition models have allowed systems to be built that can reliably transcribe spoken text, but is still not possible to train models to reliably understand the meaning or intent of that text.

Moreover, large questions in areas including bias, accountability, and regulation surround the use of machine learning models in data-driven decision making. The next section will discuss these in detail.

2.3 Challenges with Machine Learning

2.3.1. Accountability

As we see more and more decision-making processes becoming evidence-based and using data as inputs then no matter what the domain, we need to know what is the correctness of the decision that the analytical solution has determined. The use of data-driven algorithms in science and engineering such as computer vision or robotics is well established and can be measured in logical ways. In humanities and behavioural science and anything involving subjective inputs from humans, there is poor validation of the input data sets and so poorer quality of the output decisions.

The motivation behind this move, apart from efficiency, is that it is part of human nature that we have inherent established biases to various degrees. The biases are usually unconscious biases such as women having to tackle bias in gender gaps in technology workplaces, but by isolating the human emotion and bias from decision-making we reduce the impact of these biases.

The idea of using formulae to perform decisions is not new. For example, credit scores have been in place since the 1960s based on statistical models and banks are using algorithms to decide loans, insurance etc., but what has become topical these days is making algorithms accountable. Systems with algorithm-driven decision-making should be able to explain and justify their decisions and made accountable for those decisions.
Machine learning from past criminal records is being used in some parts of the United States to decide parole for re-offending criminals, and this is being delivered as a traffic light system (green, orange and red) but with no visibility of the probabilities that underlie those recommendations. This is a good example to illustrate insulating the decision output from any bias that a judge may have - gender, race, religion, geography, whatever. In theory the decisions can be more consistent as there is no emotional input and that is both its strength and its glaring weakness. There needs to be an explanation associated with every decision, a justification and a rationale.

Perhaps the technology needed behind the decision-making is a spectrum of models that can be used to give non-biased outputs, the thinking being that if you use many different algorithms to solve the same problem then you reduce bias. This is not the way that machine learning or classification is done today and certainly there is not much use of algorithms that continuously evolve themselves and their models. When used without evolution, algorithmic-based decisions become self-perpetuating, with no variance, which is not a good thing. In cases like judicial decisions on criminal parole, there is little support for audits being put in place to balance the algorithmic decision/recommendation with the personal experience and wisdom of the decision-maker.

So while the Forum agreed that a newer kind of algorithmic decision-making is very desirable, this comes into sharp focus when we consider the enforcement of the upcoming EU legislation on personal data in 2018, one that is able to account for its decisions. We also recognise the threat of “gaming” the explanations in order to figure out how the algorithm works, a form of adversarial reverse-engineering. A black-box decision maker that gives no insights into its operation is a far more secure system than one with opens its lid and allows people to see how and why it made certain decisions, so a newer kind of algorithmic decision-maker needs to find the sweet spot between security and accountability.

2.3.2 Deep Learning Limitations

The hot topic in machine learning is called "deep learning" or deep neural networks, convolutional neural networks to give it its technical title. Deep learning is at the more complex end of the machine learning spectrum and operates by using huge amounts of correctly annotated data in order to build recognisers or classifiers that require huge amounts of computational resources to train. The computational requirements are so large that local onsite implementations of the training phase require more than desktop computing resources and in many cases use general purpose graphics processing units (GPUs) in order to deliver the required performance.

What makes deep learning so attractive is that with the right input data and the right problem to tackle and in a narrow enough domain, they can operate at human levels of performance, or better. In computer vision problems, for example, deep neural networks are used by companies like Google and Facebook to automatically tag images and videos. Likewise for applications in language processing, in autonomous vehicles, and others.

However, there is a major downside, apart from the computational requirements and the large dataset requirements, and that is that deep neural networks are effectively a black box
in how they operate. Instead of storing what they have learned in some form of queryable format, by their nature they diffuse what they have learned as a series of weights across a very large connected network, mimicking the way the human brain organises information as synapses or links between neurons or nodes. This means that there is no way for them to explain or justify any decisions they make and their use has actually moved us further away from explanation-based analytics, even though they can provide better accuracy and faster processing speeds. Given that there is an increasing requirement for algorithmic-based decisions to be able to explain and justify their decisions as we will see later in this report, this makes deep learning a two-edged sword … very accurate but we don’t know how it works! This is called the black-box problem, and by “baking in” the knowledge learned from training data into the network as weights on links which correspond to features then we actually need to do a form of neuroscience in order to understand how the network of nodes and links which has been learned, actually makes its decisions.

A second downside to deep neural networks is that they are not a silver bullet for data-driven decision-making and the limitations of deep learning are now starting to emerge, which should be worrying for those who believe in this technology so faithfully. In computer vision it is easy to fool concept detectors built using deep learning in cases where the training data is not exhaustive and the network has over-fitted to the data. For example, if we have used a deep learning approach to detect motor cars using a collection of images which does not include a sofa or a snooker table as not being an example of a motor car, for then when we apply this to pictures which include a sofa or motor car they may be incorrectly classified as a motor car. This is a trivial example and arises from the unsupervised nature of the learning of features, but it becomes far more serious when deep learning is applied in applications using CCTV footage, or medical images or scientific applications. The vulnerability becomes particularly acute under adversarial conditions when those with malevolent intentions try to exploit the mis-classifications and we currently have no defense against this.

These problems with deep learning can only be resolved with a much deeper understanding of how and why deep learning implementations work so well, when they work well and when they do not, a revision of its mathematical underpinnings, and an extension of deep learning theory.

2.3.3 Human in the Loop

Using data-driven approaches is a scientific approach to decision making which is the best choice, but only sometimes. There is a spectrum from algorithm-only decisions to human interpretation of results of algorithms (i.e. decision support tools) through to human-only. These are important distinctions. There was a very strong feeling at the Forum that the human is the key decision maker, informed by the machine, and that while data-driven approaches can inform decision making for a complex question, the individual should be the decision maker and this goes back to earlier points made about accountability.

Dealing with audits of automated decisions or evaluations of decision-support algorithms and built-in biases that they may have is an issue. The Forum was very mindful of questions like who is responsible for a decision, be it made by a machine, a human or a combination? The consensus was that it must be a human because you can’t hold a machine accountable.
In all of this “human in the loop”, data is key in that the quality of the data will inform the accuracy of the machine algorithm, and its biases. This is why systems like Facebook and Google are in a great position to exploit machine learning because they have the best datasets where “best” means largest, and cleanest, so their sheer data sizes smooth over any errors in their algorithms and also because their applications are not life-critical, there is no major downside to getting the wrong kind of advert on a webpage or posting to your timeline.

2.4 The Challenge of Industrial Research in Data Analytics

There are still many more technical developments required in the data science / data analytics area with both basic research work and integrative or applied work that needs to be done. However for the most part, the data analytics community is operating in silos and there is a need for linking the different areas and applications for data analytics so that we better understand them, allowing us to better use them and ultimately get better societal benefit.

There is a two-tier system when it comes to data analytics development as the world of privately funded research where techniques like machine learning is being used in real products it is forging ahead very speedily while in public research institutions it is much slower. In data analytics, industry driven research exerts pressure in terms of fast developments whereas publicly funded research has to achieve the same and even more advanced results with more limited resources. The industry-led view can sometimes suggest that academic research may not be able to perform at the level as industry funded research can. In turn, these differing views sometimes create the illusion that things develop so fast that is quite difficult to know what it is really happening in the field. It is really difficult for the (publicly-funded) research community to keep up with what is happening in industry because we do not know what industry is doing all the time.

Companies have the data, and the problems while government-funded and public service-funded research lags behind. The availability of open public data has been a step in the right direction but moves too slowly compared to progress in the private sector where companies like Google and Facebook have budgets for data analytics which are far larger than most funding agencies.

So the challenge here is finding the sweet spot between public and private research funding in an area which is so lucratively attractive for industry and where researchers can see their work deployed to masses of people and making real impact, something that is not always the case under publicly-funded research. Can academics really keep up with the field when so much is carried out behind company firewalls? A parallel can be drawn to the example of the chemical industry in Germany where industry commercialised academic developments but where the two research areas now support each other symbiotically.

Industry does not want to put publicly funded research out of business and industry does benefit from innovative research closely related or far away from their product lines. It is not in industry’s interest to alienate publicly funded research and it is important to keep open the relationship between industry and academia. This co-dependency between academic research and industry research is a necessary handshake to keep the flow of skills, talent,
knowledge and information going in one direction, and the flow of access to real data, industry training and funding support going back in the other direction.

Meanwhile the open source community is providing powerful analytics tools and disseminating learning outside of the major web companies and also the fact that researchers will move around companies will cross-fertilise ideas between companies and back into academia.

2.5 Shared and Communal Benefits

There is a basic phenomenon in science that when we measure something, we change it. This comes from quantum theory and is the basis for Heisenberg’s uncertainty principle, and a variation of it applies to applications of machine learning and recommender systems. When we use machine learning to make a model for some decision-making that affects people and we apply that model we affect and change people’s behaviour, thus invalidating the model we have generated. Most uses of data analytics simply apply a model, the same model, to multiple cases without considering the overall global effect. For example, consider routing cars to vacant spaces in a busy car park. Every time a recommendation is issued for an incoming car to park at a specific vacant space, the model of space availability changes, and that makes sense. For simple systems like this, the overall global constraints, car park space availability is the overriding concern. Waiting list systems in hospitals can also fall into this category, where the model changes dynamically as the recommendation is taken up or a place on the list is filled.

For complex, real time systems however, there are elements of control theory, which are not integrated into models of systems learned by machine learning approaches. This issue arises in, for example, the finance industry where everybody gets the same information about stocks and share prices at exactly the same time and all the prediction algorithms react in the same way, which can lead to thrashing.

Another example where this arises is in vehicle route planning where, in real time, our phone apps or satnav systems tell us the shortest or fastest route to take to get from A to B. Some of these even factor in live data about average speeds on roads along the route to give finer-grained predictions of travel time. Yet every commuter gets the same information and there is no application of control theory to optimise overall throughput for all commuters. Perhaps this is too large a computational problem to merit the investment in this case, but as our systems get larger and more complex, the knock-on effects and interactions among individual predictions without overall control theory will come into play. Some will be obvious, like car commuting while others will be more subtle and longer-term.

When a vendor recommends products to an individual customer based on past purchases, and that customer makes those purchases then the customer has changed, or perhaps reinforced the model the system has of that person. Models need to evolve and change constantly every time they are deployed, models learned from data need to learn from their own use and need to consider individual cases or decisions as communities or ensembles of decisions for the benefit of the whole system rather than just as sets or sequences with benefits of short-term goals. This requires re-thinking and extending data analytics techniques across all sectors, not just those listed above.
3. Platforms for Data Analytics

Increasingly in big data and data analytics applications, data usually arises as data streams, coming from activities happening in real time as they are being logged or recorded. Algorithms for data mining and pattern recognition, which form part of what makes up the field of data analytics, used to be executed on desktops or servers or on the cloud because the demand for real time responses was less than it is now. These days we see an increasing trend and demand for real time responses. Smart cities, smart grids, applications in transport, financial trading, these are all examples where real time responses are built into the application. Yet data from sensors in our environment, from sensors monitoring our health and physiology and many others all require analytics that may at best not be executed on big servers or cloud computing. Do we see the implementation of analytics processing moving from cloud servers down to new small devices? Is there a case for wanting easy access to some data directly on the network instead of having to shuffle a lot of low-level data to big servers for processing and analysis?

This idea is referred to as “edge computing” and examples of this would be the on-device data processing done by smart devices like phones, or even on less smart devices like wearable personal sensors which do the computation to count steps or distances walked or calories burned, on the device itself. The technology evolution is now at the stage where we are seeing general-purpose programmable computing and even hardware support for machine learning algorithms, directly on devices. Intel’s Knights Mill chip, whose formal announcement is expected shortly, will be able to run memory-hungry convolutional neural network implementations directly in silico, pushing data analytics capabilities down onto the device. Samsung have indicated similar plans for their smartphones. Low cost Raspberry Pi devices, credit card computers with low energy requirements, are also helping to promote edge computing as a viable architecture for many applications. In this scenario the cloud is used to do the training and learning of models, these models are then pushed out to the devices for edge computing and decision-making leading to faster reaction time, lower latency, and reduced networking demands with less raw data being transmitted.

The reason this is an important issue, and even more so for Ireland, is because of the growth of the internet of things (IoT) where IoT devices will generate or capture huge amounts of data which will need real time analytics for which edge computing is a very viable architecture. In IoT implementations when we put data analytics “on the edge” then one of the threats is that the security element will become an issue where the device will not be able to process the analytics models and manage any required security encryptions. An example is in smart energy grids where local edge processing could be done on data for the home user while cloud-based data analytics like data mining and pattern recognition could be applied in other ways for other business applications and business intelligence.

The discussion on this topic at the Forum meeting went back and forth, reflecting that the benefits of edge computing are not at all clear-cut as yet and not everybody agrees that this is actually a good idea. Some of the positives of edge computing are indicated above, and some of the negatives are:
If we divest decision-making onto a device with edge computing then what does that mean for being able to justify or explain decisions, important in the context of the EU regulation on the demands for explanations set to become law in mid-2018? With edge computing, all intermediate data, all intermediate decisions, the things that are needed in order to provide explanations, is lost once the calculations are completed.

Even with implementations directly in silicon there are limits on the processing capabilities for edge computing as with Machine Learning implementations now directly available in silicon, these just run the models, not necessarily train them. Even though there is intelligence on the device, which then transmits part-processed data back to the cloud to support further decision-making, these implementations do not evolve or dynamically learn those models.

Standardisation of data collection from devices so that it can be usable for all is still a long way off. The amount of work needed for completion of standardisation during data collection is a priority because it requires a lot of work right now where standardisation improvements will be required or else it will prove difficult to get the data in the right format to carry out analytics.

As of now we take the possible benefits of edge analytics on a case by case basis, but whatever advantages of edge computing that may exist, its use is not a black and white answer. Edge vs. cloud decisions are decided based on the solution need at the point of use, trying to find the sweet spot between conflicting demands. Like many other issues in data analytics, this question of edge vs. cloud does not sit in isolation – we need shared challenges to drive both the industry and academic discussion and the development of specific platforms and algorithms.
4. Data Privacy and Ownership

4.1 The Importance of Data Privacy

There has been a polarising debate over the privacy and ownership rights associated with personal data. On the one hand, when we generate data, or data is generated about us, then the issue is who owns that data, who has permission to use that data, and for what purposes, who can see that data, has a person the right to know answers to some or all of these questions and if data about a person is erroneous then has a person the right to know this and to change the data.

The issue has been particularly important in Ireland because of the number of large companies based in Ireland whose business is about personal data - gathering it, mining it, learning from it, extracting usable knowledge from it and then using that knowledge. This it is important that Ireland takes seriously the issues of consideration of data ownership and needs to show leadership internationally in this area. We have started well with the appointment of Minister Dara Murphy as the first Minister of data protection in the EU but we need to think beyond legal frameworks and how to use the blunt instruments of legislation, some of which is unenforceable.

In broad terms, we need a model that can handle both data ownership and privacy, in other words the model should take into account elements of privacy to be included in subsequent analysis and use of personal data as it may subsequently reveal hidden information.

To highlight the importance of this issue, the Forum was told of Insight’s launch of a Magna Carta for Data in Brussels in 2015, to start an EU level conversation about data ethics. The aim was to lift the conversation above the level of data privacy and protection, and to encourage policy makers to look at the rights of citizens in balance with the very real benefits that responsible data research can return for those same citizens.

Subsequently, Insight participated in a Day of Action on Data for Health and Science, once again in Brussels, that focused on the specific ethics challenges faced by patients, researchers and healthcare providers when capturing, retaining and using patient data. Since then, Insight has identified the need to broaden the conversation beyond the business, health science and computing community, and to draw in perspectives from social science and the humanities. This notion of a Magna Carta for data is gathering momentum and focusing attention on the issue.

4.2 The EU Directive on the Protection of Personal Data

In January 2012, the European Commission proposed a comprehensive reform of data protection rules for citizens within the EU. The objective of this new set of rules is to give citizens back control over their personal data, and to simplify the regulatory environment for business. On 4 May 2016, the official texts of the Regulation and the Directive were published. The Directive entered into force on 5 May 2016 and EU Member States, including Ireland, have to transpose it into their national law by 6 May 2018.
Formally, this is defined as Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

At the core of the legislation is the need to think about the right of the citizens and treating the current 28 member states with the same rules in relation to data privacy. Different member states have very different cultural opinions across the EU. Even though the legislation is the product of 4 years of negotiation the member states could not agree on some parts of the regulation such as the age for consent for children on social media. Initially this age was proposed as 16 years, and then it was brought back to 13 years but will not be consistent across the EU. There is a harmonisation process to try to address such discrepancies and find common agreement but not everything can be resolved.

Another important part of the legislation is to do with a citizen’s power to request to have the outcome of any automated decision made about them, explained. Obvious cases where this is important are in credit rating, for example, and this creates a major headache to companies who use, for example, machine learning to classify people not just for credit rating but for targeted advertising or for product recommendation. As we saw earlier, in deep learning, whose deployment is increasingly popular, by the nature of the technique we don't actually know what the internals of the algorithm are doing, we're not supposed to, so how will citizen explanations work there. What about self driving cars which are already on roads (in the US) and are making driving decisions that affect us? Where does consent and explanation fit in there?

Nationally, we are trying to understand the EU regulation because once the regulation becomes a law it will have huge impact. Ireland will want to interpret it with the most pragmatic view as the Irish Commissioner for data regulation has to enforce it for the companies which are based here. Consent is the tricky aspect of the data privacy regulation, not just with the original consent given to use data but also to be able to withdraw consent subsequently, and a key challenge is making this national & European data protection law, functional, i.e. making it work. The legislation also needs to be flexible to work in the nuanced deployment of the law to everything from Facebook, to data from agri-food production.

Clearly the legislation needs the support of technologies because continued deployment of deep learning, which does not embed re-tracing and explanations into its *modus operandi*, is not a way forward. What support can technology offer the legislation and what is acceptable, for example the issue of digital trust? Anonymisation methods are currently a tool used in data protection but cross-referencing methods can get around anonymisation techniques and therefore in general anonymisation does not work. An alternative solution to anonymisation is by adding noise to data but still preserving the statistical and distributional properties of the data. In order to apply this method the sampling strategy is very important.

So while the incoming legislation is based on a model of one-size fits all which may bring disadvantages for specific population subgroups, there is now an imperative to researchers and technology developers to focus on techniques to support this legislation.
4.3 Data Privacy as a Business Model

An issue which had a lot of attention during the Forum discussions was the idea of privacy as a business model – paying for the privilege of data privacy. This would require a technological solution based on having a more complex representation of a data point (in contrast to a simple data point with a single value) by attributing additional parameters such as protection, permissions, access, perhaps even accuracy and reliability, provenance, or cost for usage. If this was already in place today we could imagine a world with a completely different model of managing our data, and where data-rich systems like Facebook and Google which use our data and sell it on to others, would operate in a completely different way. For example, Facebook computes 92 data points for each of their users so that they can predict characteristics like gender, politics, technology ability, and this model of each user is what is being sold to Facebook advertisers. Who owns this model, should the user, and should the user benefit in some financial way from this model?

Economists regularly say that there needs to be a model for consumer data which offers reward rather than restriction on use of data and that if this was in place then issues like ownership, privacy, sharing, access and monetisation would all disappear. However re-engineering current personal data management into such an utopia may not be possible. Instead the Forum felt that doing this in a narrow domain where data analytics, mining, exploitation and usage has not yet taken off, might be feasible and data management in the agri-food production sector was suggested as one possibility and a second possibility was suggested as the data streams coming from automobiles. In the case of agri-food production the attraction is that there is very little work done to date in this sector and the attraction of using automobile data is that it less personal than personal data but has a lot of potential value.

4.4 Technologies for Privacy Management

Since there is such a gaping need for technology solutions to the problems of privacy management then what technologies are currently available and used?

At all times, the traceability of data is as important, as how we store data. By that we mean that privacy is not about encryption and security of access, in the context of data analytics and big data applications it is about reverse engineering analytics results and outputs and re-tracing data back to its origins, identifying individual people and their attributes, which should not have been identifiable and should have been kept secret.

This can be taken back to basic principles of why we aggregate and process sets of data. For example, if we are doing a comparison of people’s salaries across different job types and I earn less than you, then I don’t need to know how much more you earn than I do, just that you do. Likewise we might only need to know pairwise salary comparisons across a subset of the people in our survey. Knowing just this minimal amount of information would be sufficient to carry out the comparison across job types but we don’t step back and ask what are the real questions we need to have answers to in order to carry out whatever analysis we are doing. Instead we take a blunt and lazy approach and we gather everybody’s salary and then do our calculations. In other words, we gather and store more data than we need.
Once this over-abundance of data has been gathered we then layer a blanket of privacy over it by anonymising it, or trying to do so. However, when the data relates to real people and given there is now so much public and open information available about people like census information, voting registers, property registers, school rolls, etc., it is possible to use this information to cross reference what should be private data and thus undermine the anonymity of this data. There are many examples of this which garner media coverage and create negative publicity and sentiment towards applications for big data.

As a next step beyond anonymisation we can add controlled noise or errors to data sets and with such data sets super-imposed with simulated noise, we can preserve anonymity. This has to be done carefully, adding noise to data sets while preserving the properties of the data set overall. In such cases we have data correlated centrally and packaged incompletely in order to protect anonymity but even this is not a fool-proof solution as there are ways around it.

Yet another possible technology for privacy management is differential privacy which is learning the characteristics about a group of entities while protecting the privacy of each individual by using statistics to learn probability distributions for characteristics while not recording anything about any individual entity. This is equivalent to generating language models in text retrieval and is well studied in that domain but its downside is that it requires large samples of the data from which it can generalise, a situation that isn’t always true in real life applications.

In summary the Forum was of the view that while there is much work and lots of interest in technology solutions for privacy management, we are a long way from delivering the kind of techniques needed, and even then we have the problem of widespread deployment and adoption of such techniques.
5. Sectors for Data Analytics

Data Analytics and Big Data applications range right across all sectors including healthcare both public and personal, education, transport, energy, entertainment, advertising and marketing, science and research, manufacturing, security and law enforcement, finance and trading, and more. Some sectors, however, offer particular attraction and a few of these were highlighted during the Forum meeting.

5.1 Data Analytics in the Agriculture and Food Production Sector

The Irish agri-industry has a value of €20 Billion per annum and counts for 40% of our exports, it is bigger than the pharmaceutical sector for example. Yet, across the world, food production needs to increase massively in order to feed the world’s growing population. For example, annual cereal production will need to rise to about 3 billion tonnes from 2.1 billion today.

Farming and food production has huge potential for the use of IoT, drones, in-situ sensing and remote sensing of the environment from satellite data, with data tracking throughout the food production process. The integration and analytics applied to such data allows traceability of produce from farming to fork.

Ireland can be a world leader in data analytics for agri-tech, the process of synergising this data, mining it for production information and then visualising and acting on it in order to improve the final product as it moves from the farm to the market. Examples of this include crop monitoring, counting livestock, using remote satellite measurements to test water body quality or soil quality, and more. Environmental sensors linked to agriculture are a big opportunity as the resolution of satellites are improving and we now have massive data sets available. The issues include development of more effective strategies for monitoring crop/food/water/soil condition in real time, for example, by combining information from satellite imaging, with drone based data and in-situ sensing. The overall challenge is how to structure this heterogeneous data in a way that it can be shared and made accessible as part of a hybrid solution that combines different tools for the agri- and food production sector.

However, Ireland needs to move fast in the AgriTech space in order to capitalise on data analytics opportunities as part of a future for Ireland. Already some companies like John Deere are producing high tech solutions for agritech and as the commercial players are moving into the agritech sector we are seeing opportunities for growth. Much of this work would fall under the umbrella of computational sustainability and the development of business models for such an ecosystem, a topic which should attract funding opportunities.

5.2 Data Analytics in Healthcare

The potential benefits for data analytics in the broad life sciences and healthcare sector is enormous in terms of personal healthcare, population health, genomics, and the economics of healthcare. Patient monitoring is one area with particularly strong potential for healthcare cost reductions, personalised medicine and more.
However, any initiative in using analytics in healthcare raises the kinds of questions around data protection and ownership discussed earlier and this is a massive limiting factor, especially in the space where people are widely using the cloud and open architectures for data processing.

Much of the debate surrounds the issue of the patient electronic record, the single integrated view of a patient’s entire record, the development of which has been ongoing in Ireland for years with very slow progress. This is seen as a limiting factor and stifles really innovative work such as the ability to integrate large (patient record) datasets. In other areas where data has fewer usage constraints such as advertising and transport, we have seen the application of data analytics reveal real business insights and intelligence and create commercial opportunities. If the same levels of data integration and the "single customer view" were in place for healthcare data then similar insights would be revealed, ultimately to be used for improvements in healthcare.

The Forum also discussed that similar situations also exist in any other scenarios besides healthcare where a single integrated customer / patient / client record does not exist, and an example of this other scenario is our higher education institutions.

5.3 Data Analytics Across Application Areas

Having discussed a range of application areas for data analytics, one thing that emerged during the Forum discussion was the realisation that the technologies used in data analytics across different areas are not the same and are very different when you move from one sector to another. While machine learning techniques, for example, may seem generic, in fact when they are used they are specialised for specific applications.

The reason this was raised as an important point was that there is a perception that data analytics is agnostic in terms of areas or types of data to be analysed. In fact, as demonstrated in practice, there is not such a concept or technology leading into the reusability of models for data analytics but there is reusability of the analyst that move from one sector to another and is able to create and modify algorithms for specific purpose. This leads to limitations with regards to the long-term sustainability of tools or established solutions in data analytics, but demonstrates the value of the data analyst who brings his or her know-how and also existing models to different domains, and thus demonstrates the transferability from sector to sector. It is thus an important research topic, and improving the transferability of techniques, probably by automating in some way, needs to be addressed.
<table>
<thead>
<tr>
<th>Attendees and Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prof. Michael L Brodie</strong> Research Scientist</td>
</tr>
<tr>
<td><strong>Prof. Ken Brown</strong> Principal Investigator</td>
</tr>
<tr>
<td><strong>Paul Buitelaar</strong> Principal Investigator</td>
</tr>
<tr>
<td><strong>Margaret Burgraff</strong> Vice President Business Development and Global Geographies Developer Relations</td>
</tr>
<tr>
<td><strong>Joe Butler</strong> Senior Director of Engineering</td>
</tr>
<tr>
<td><strong>Prof. Tat-Seng Chua</strong> KITHCIT Chair and Professor of Computer Science</td>
</tr>
<tr>
<td><strong>Mike Conroy</strong> Industrial Partnering and Alliances</td>
</tr>
<tr>
<td><strong>Medb Corcoran</strong> Intelligent Enterprise Analytics Lead</td>
</tr>
<tr>
<td><strong>Ollie Daniels</strong> CEO</td>
</tr>
<tr>
<td><strong>Prof. Dermot Diamond</strong> Principal Investigator</td>
</tr>
<tr>
<td><strong>Ernesto Diaz-Avilies</strong> Chief Data Scientist</td>
</tr>
<tr>
<td><strong>Dr. John Doyle</strong> Lead Analyst Developer Research &amp; Development COO Chief Data Office</td>
</tr>
<tr>
<td><strong>Prof. Gene Freuder</strong> Funded Investigator</td>
</tr>
<tr>
<td><strong>Prof. Nial Friel</strong> Principal Investigator</td>
</tr>
<tr>
<td><strong>Ronan Gavin</strong> Solution Lead - Data Platform &amp; Analytics</td>
</tr>
<tr>
<td><strong>Dr. Carol Gibbons</strong> Director of ICT Programmes &amp; Sector Manager for Electronic Clients</td>
</tr>
<tr>
<td><strong>Gareth Hanley</strong> Managing Director</td>
</tr>
<tr>
<td><strong>Adrienne Harrington</strong> Head of Data Protection Unit</td>
</tr>
<tr>
<td><strong>Dr. Ali Intizar</strong> Funded Investigator</td>
</tr>
<tr>
<td><strong>Prof. Mark Keane</strong> Principal Investigator</td>
</tr>
<tr>
<td><strong>Prof. Tahar Kechadi</strong> Principal Investigator</td>
</tr>
<tr>
<td><strong>Dr. Philip Kelly</strong> Senior Data Scientist</td>
</tr>
<tr>
<td><strong>Spyros Kotoulas</strong> Manager, Health and Person-Centred Systems</td>
</tr>
<tr>
<td><strong>Sean Lennon</strong> TSS R&amp;D Analytics Architect</td>
</tr>
<tr>
<td><strong>Flavio Livide</strong> Team Lead, Interactive Analytics Platform Research</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Dr. Brian MacNamee</td>
</tr>
<tr>
<td>Dr. Michael Madden</td>
</tr>
<tr>
<td>Dr. Elena Martines</td>
</tr>
<tr>
<td>Dr. Andrew McCarren</td>
</tr>
<tr>
<td>Pat McCarthy</td>
</tr>
<tr>
<td>Joe McGill</td>
</tr>
<tr>
<td>Mark McKeon</td>
</tr>
<tr>
<td>Cronan McNamara</td>
</tr>
<tr>
<td>Gearoid Mooney</td>
</tr>
<tr>
<td>Prof. Noel O'Connor</td>
</tr>
<tr>
<td>Eamon O'Doherty</td>
</tr>
<tr>
<td>Dr. Simon O'Regan</td>
</tr>
<tr>
<td>Prof. Barry O'Sullivan</td>
</tr>
<tr>
<td>Dr. Paul Palleth</td>
</tr>
<tr>
<td>Prof. Gregory Provan</td>
</tr>
<tr>
<td>Prof. Dietrich Rebholz-Schumann</td>
</tr>
<tr>
<td>Dr. Alessandra Sala</td>
</tr>
<tr>
<td>Prof. Robert Shorten</td>
</tr>
<tr>
<td>Prof. Alan Smeaton</td>
</tr>
<tr>
<td>Prof. Padhraic Smyth</td>
</tr>
<tr>
<td>Paddy White</td>
</tr>
</tbody>
</table>
Some of the attendees at the Data Analytics Future Forum event, September 2016, DCU

Other contributors

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Karen Dawson</td>
<td>Research Integration Coordinator</td>
<td>Insight Centre for Data Analytics</td>
</tr>
<tr>
<td>Dr. Ernesto Diaz-Avilies</td>
<td>Chief Data Scientist</td>
<td>Citi Innovation Lab – Dublin</td>
</tr>
<tr>
<td>Dr. Antonella Ferrecchia</td>
<td>EU Programme Manager</td>
<td>Insight Centre for Data Analytics</td>
</tr>
<tr>
<td>Keith Fingleton</td>
<td>Chief Technology Advisor - Financial Services</td>
<td>IDA Ireland</td>
</tr>
<tr>
<td>Ken Finnegan</td>
<td>Senior Technology and Research Advisor</td>
<td>IDA Ireland</td>
</tr>
<tr>
<td>Keith Griffin</td>
<td>Principal Engineer</td>
<td>Cisco Systems, Inc.</td>
</tr>
</tbody>
</table>