

---

# What data science means for the future of the actuarial profession

---

## Abstract of the London Discussion

[Institute and Faculty of Actuaries, Sessional Research Event, London, 3 April 2017]

**The Chairman (Mr M. H. Tripp, F.I.A.):** A very warm welcome. First, just to introduce myself, I am Michael Tripp. I chair the modelling analytics and insights and data (MAID) working party. I come from a general insurance background but I stress that the working party is cross-practice.

There is not one phrase that really sums up what we are talking about, although big data, data visualisation and predictive analytics are used on occasion. We have tended to use the words “data science universe” as the best catch-all.

One of the questions that I keep asking, and I think the profession keeps asking, is “what makes an actuary”? I think there are probably two key elements: mathematics, particularly statistics and financial modelling; plus the main elements of expertise in general insurance, life insurance, pensions, health insurance and so on.

But what has been going on over the last 15 years or so in the academic world is that computer science has been advancing in terms of analytics and what it does, and gradually has become data science, and data science is just starting to learn what mathematicians have been doing for a long time. Mathematicians and operations research people have been coming the other way, perhaps being a bit slow, to understand what computers can do now. Thus, there is a very rapid convergence of computer science, data science and maths. That is what presents us with the challenge of the moment.

We had an open discussion with the council of the Institute and Faculty of Actuaries (IFoA) at the end of February. It concluded that actuaries, in order to do the best job possible for their employers or their clients, must know what a data scientist would do in dealing with the problems that the actuaries are tackling.

Other actuarial professions around the world have taken steps to address these issues. I think that Henry (Mungalsingh) will touch on that later. The IFoA is now at the point where the Council have asked us to produce a strategy paper to help the profession determine what it should be doing.

There are two parts to what we are trying to achieve tonight. First, we want to share some things that we have been doing for the past year related to machine learning and data science. In the second part of the evening we would like to obtain some feedback about what you would like to see the IFoA do. Whether it is exams, continuous professional development (CPD), qualifications; whether it is the quality of our work, the risk, the regulation; whether it is the ethics; whether it is public relations (PR) and image; whether it is recruitment and the skills we need, collaboration with other people, research. Whatever it is, we should like to hear your views in order to help inform the strategy paper that we are putting together.

MAID started about 12 months ago and has four work streams. Work stream one is about research. It has done a really good job and obtained feedback from about 400 people. This suggests that, not surprisingly, actuaries know about data science. There is a general awareness and an appreciation of its importance although many actuaries may not have specific knowledge about the technicalities and the methodologies. Older actuaries tend to have less awareness of the approach and there tends to be a higher awareness amongst actuaries working in general insurance than in other areas and, perhaps, lower awareness amongst pensions or life actuaries.

Clearly, there is a demand for learning and support for members of the profession for their own career development. People particularly feel that we need to have experts to whom they can go if they have problems that are beyond traditional mathematical techniques. There is also a recognition that collaboration is really important.

If you speak to chief executives of an insurance company and ask them about data scientists, they will probably say: “Yes, we have used data scientists. They have been helpful but not necessarily as helpful as an actuary, because they do not always understand how insurance works.” Nevertheless, data science is a buzzword and we need to be aware of it and be able to deal with it. Clearly we are a profession, we have the public interest at our hearts and people are aware of ethics and legal considerations.

Figure 1 shows the actions resulting from the research. Several work streams are shown on the slide.

Work stream two is well represented here and we will discuss it in a moment. There are four associated case studies that have been worked on using new techniques to deal with traditional actuarial problems.

Work stream three is not represented here tonight. It is concerned with whether combining data science with actuarial science will create opportunities in the wider sense for actuarial professionals.

Work stream four is about the implications for the profession and is represented tonight by Henry (Mungalsingh).

To start the ball rolling, we are going to ask Alan Chalk to speak. Alan has done jobs for many large insurance companies on pricing and recently did a Masters degree in machine learning at University College. He is going to give some personal thoughts on machine learning, what it means and how fascinating he found it.

1. Promotion of data science methods and opportunities for actuaries
2. Pre- and post-qualification learning to equip current and future actuaries with required tools and techniques
3. Support advancement of the subject matter through research and thought leadership
4. Collaborate with other professionals and disciplines to share knowledge and advance techniques
5. Ensure regulatory and ethical implications arising from data science universe are understood

**Figure 1.** Possible actions identified from member survey

**Mr A. Chalk, F.I.A.:** I am going to discuss what machine learning means to me, and my involvement in this particular working party.

I finished the exams 25 or 30 years ago. So I no longer know anything about life assurance because I spend all my time working in general insurance. But this working party covers pretty much everything done by actuaries. I work together with Valerie (du Preez), who introduced me to a life assurance company.

The people working there said that they do experience analysis. In summary, if you are a life assurance company you need to know what proportion of people are going to die given a certain age. This is important for product pricing as the benefits depend on whether people die or not. This is a difficult problem for life assurance companies because they do not insure sufficient people at every age for them to obtain an accurate view of what proportion of people die.

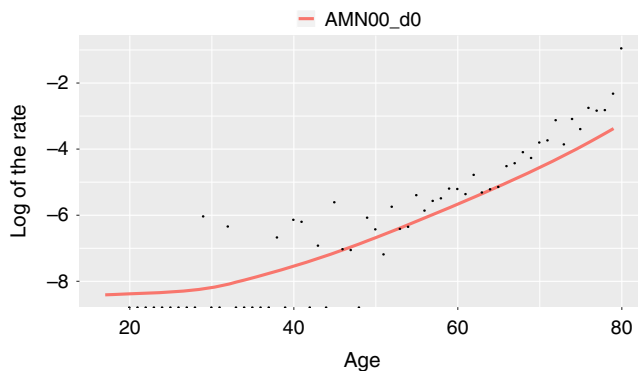
In Figure 2 the black dots represent the actual experience at each individual age. You can see that they are all over the place. If a life assurance company were to use its actual experience, then you would have weird outcomes, such as, for example, nobody dying between age 20 and 30 and then a massive mortality rate at age 31.

This is inappropriate so you somehow have to smooth these mortality rates. Life assurance companies do this by taking a mortality curve which has been fitted over population data and therefore is very reasonable.

You can see that the red curve is nice and smooth. That represents the actual experience of the population. But the company cannot use the actual experience of the population because that might not represent the people they assure. So they have to try and do something to make the red curve remain smooth but to be closer to the black dots. You can see that the black dots are higher than the red curve at the older ages. It is not quite clear what is going on at the younger ages.

The life assurance actuaries deal with this problem by grouping some ages together. For example, if ages 75–80 are grouped together we will see that at age 75–80 the experience is 20% above the red line.

If ages 20–30 are grouped together the experience is about 10% below the red line. Given the results, the company will start adjusting the red line downwards at some ages and upwards at some other ages to try to make it more closely reflect their own experience in a sensible way.



**Figure 2.** Experience analysis

```
gam( death ~ s(age, bs = "cr", k = 20),
      quasipoisson(link = "log"),
      data = dt_all[fold %in% 1:5],
      offset = m1_pred_rate_exp_log)

user      system  elapsed
5.169     1.302    6.495
```

**Figure 3.** Experience analysis ... quicker

There is one problem with this approach. It is not clear how to group the black dots. For example, should you group 75–80 together? Or maybe you should group 70–80 together because that will be a bit smoother? Maybe you should put all of them together and just have one adjustment overall?

In practice, life actuaries try a few different groups and see which look best. That is a case of learning from the data. Humans learn from the data by looking at it and seeing what looks reasonable. This particular exercise is a case of human learning.

Machine learning involves a machine learning from the data. One possible way this can happen is shown in Figure 3. In this case, one line of code will solve the problem in a very optimal manner.

Let us examine the line of code: “gam” stands for generalised additive model. This is a statistical model and is not especially related to machine learning. Then in brackets it says “death” which refers to the mortality rate.

Death depends on some kind of function of age. “s” stands for smooth. In effect you are telling the computer you want it to be smooth. The parameter “ $k=20$ ” tells the computer how “wobbly” the mortality rates are allowed to be. The computer does not necessarily use “ $k=20$ ”. That is just the maximum wobble that is allowed. The computer looks very carefully at the data and sees what is predictive and what is not predictive. In fact, in this particular case the computer chose the final parameter of  $k$  to be 1.98.

The question is whether the human built model or the computer built model performs better. I do not know because I did not have access to the insurance company data. But the computer approach is a pretty optimal way of doing it.

The machine learning is frighteningly fast. Actuaries are doing things in old-fashioned ways which are potentially slow and potentially perform less well in practice, where performance actually matters.

I felt strongly about this. I am very interested in other people getting involved in the area.

If you want to become involved one approach is to listen to probably the best machine learning lecturer in the world. Google “Yaser S. Abu-Mostafa” and listen to his first 4 hours of lectures. You will have covered the fundamental high-level concepts of machine learning.

A second suggestion is to use Kaggle as a really easy way to become involved. Kaggle is a forum that puts up competitions. They put up starter scripts and give you data. All you have to do is download the data, download a script, and try to compete and do what teams of data scientists around the world are doing.

One competition which is up at the moment is called “Two Sigma Connect: Rental Listing Inquiries”. What you have to do, if you compete, is to try and predict which listings on some house rental sites are going to attract low, medium or high interest from the people going through that website.

The only reason why you might not want to do this is because you are afraid of being bottom of the scoreboard. But within the next 2 or 3 days I am going to put a really poor submission there under my own name. You will see Alan Chalk at the bottom of the leader board across the world so no one needs to be afraid of looking bad in public.

The final way to become involved is through this working party. We are going to listen to talks on case studies from Alex (Panlilio), Zhixin (Lim) and Valerie (du Preez), who have used machine learning on a problem which is relevant to what they are doing. This work still an ongoing part of this working party and I think there are an opportunities to become involved. Have a listen to these case studies and be enthused.

**The Chairman:** We are going to move on now to the three people who are going to present the four case studies. Unfortunately Ben Canagaretna, who has previously chaired this section, is not available. In his stead is Alex Panlilio who is assistant catastrophe developer in exposure management at Barbican Insurance, which is a general insurer. Alex did engineering mathematics at Bristol with first class honours. Alex will talk about exposure management and machine hull underwriting.

After that Valerie du Preez, a member of the senior management at Grant Thornton in their actuarial risk team, will talk about mortality. She has done a lot of work in the insurance industry, in particular, leveraging tools and using various disciplines to help to do things differently and in a better way. During a recent project on financial processes she saw that there were certain processes that were suitable for automation. So she is well used to applying technology and will discuss a practical use of machine learning.

Then Zhixin Lim will talk about asset liability management. He is very keen to explore the use of artificial intelligence and machine learning.

**Mr A. Panlilio:** I will discuss work stream two. We are looking to try to find new approaches to current actuarial work. We started off by looking at a list of problems across the pricing areas and tried to find which one is the most applicable to us. From this we came up with four case studies, the first one of which is the exposure management case study.

The case study aims to use machine learning to fill in and complete missing data. Normally in catastrophe management we receive property data from Risk Management Solutions and AIR. Coming from the broker, this varies in quality. Some data are missing and some are quite good. We tried to train on the data we had to make a prediction. We did this by looking at how tall a building was in the property data and how many storeys a building had in terms of continuous variables.

What we found was, with the geographical information, latitude and longitude were the best predictors for the age of a building, whereas, property value was the best predictor for the height of a building in storeys.

With this process we defined the problem, uploaded the data, tested and refined the model using quite iterative steps.

What we would like to do as the next stage for this project is to try to apply this method to multiclass fields. Occupancy codes, residential, commercial and construction codes, and the material from which a building is constructed are influential on the pricing. Unfortunately, the software we were using does not support this work although we would like continue.

Moving onto the next case study which is on marine hulls. I am presenting on behalf of one of my colleagues who conducted this project. In this project we were trying to look for a more sophisticated way of rating risks. We used machine learning and applied it to ship data and claims data and looked specifically at the frequency and severity of claims separately and then in aggregate.

Our results show that ships that were older than 1980 and younger than 2010 had fewer claims than ships constructed between 1980 and 2010. We think this is linked to mileage: how far a ship has been travelling. The older ships would be docked and the younger ships would not be used as much. We would like to explore this further as well as adding further claims data because we had several thousand ships and only a few hundred claims. Obviously adding more ships into the model would help refine the model and improve it.

I am going to hand you over now to my colleague Valerie (du Preez), who will speak on mortality.

**Mrs V. du Preez, F.I.A.:** Thank you, Alex (Panlilio). Over the past few months, I have been fortunate enough to work with Alan (Chalk), Alex (Panlilio) and Zhixin (Lim) to understand what data science would mean for an insurance actuary. I should warn you, if you talk to Alan after this session he is very good at persuading you to open the black box of programming. I am thankful that he persuaded me to open it and start exploring.

Our case study on mortality involved the step-by-step analysis of applying supervised machine learning. I will now take you through the steps involved.

It is fascinating to see how much data are publicly available. Through Kaggle, we identified that there were death statistics available on a 2014 data set of US deaths. When we explored further, we realised that the Centre for Disease Control and Prevention in the United States annually publishes specific death data on each of the recorded deaths at quite a granular level of detail.

Data on each individual's death was recorded including the age of death, the gender, educational information, racial information, as well as the month and the day of the week of death.

Before we applied any predictive analytics to that data set, we applied an interesting approach and looked at visualising the data in a different way. Through the use of R, we looked at word cloud analysing to find out which types of death stand out, perhaps heart disease or cancer. We also looked at population data within the United States and considered seasonal variances.

We wanted to see what machine learning could do with this data set. We quickly realised that we had to define our problem very specifically in terms of what we wanted to predict. We brainstormed a couple of case studies.

The one that came up is quite an interesting one to look into: whether there was any link between the number of suicides and external factors within the US economy. We looked at external data and we

enriched our data set, which now consisted of data from 1980 to 2014, by looking at three additional data sets. First, as a proxy for market performance, we looked at the Dow Jones. In addition to that, we looked at the Consumer Confidence Index in the United States and patterns of savings and spending. Lastly, we considered a new innovation: the Mood Index. Based on social media posts, such as Twitter, there are a couple of indices being developed that show the optimism of a specific area in a country. There are various companies that derive the data and give you quite a unique way of looking at it.

We wanted to look at these three additional data sets and apply a machine learning model to our data. It involved numerous iterations. We came up with a regression type model that fitted the data quite nicely. We concluded that there was a slight correlation between the increase in suicides for men and a decrease in the Consumer Confidence Index. The other two indices did not prove statistically significant.

For me that was one part of the conclusion. A more important part is that we have been able to take these steps in order to analyse the data in a different way than has been done before.

Of course, there are limitations to our project and we have identified the next steps that we would want to take to improve it. The 2015 data sets have just been released. We have also started thinking about what additional views would be interesting to see. Different scenarios to investigate, and what additional external data we could add to this data set.

We have identified certain areas, related to reserving and pricing, that we want to investigate, in order to understand how they would be applied for a life insurance company.

On that note, over to Zhixin (Lim).

**Mr Z. Lim, F.I.A.:** In this case study I am going to show you an example of how machine learning can save you time and effort. To set the scene, this started as a project to forecast interest rates using machine learning techniques. As I was training the model, it became quite apparent that central banks' monetary policy, especially quantitative easing (QE), plays a huge part in the fitted model.

This is not surprising, given how influential central banks are on the level of interest rates. Central banks control the supply of money through monetary policies and with QE they literally print money. The communications published by the banks are an important source of information. These communications include speeches, press releases, inflation reports and other forms of guidance. The tone or sentiment of these communications, from dovish to hawkish, set expectations in the market. If you are in the business of predicting interest rates, you have to read these communications.

I love reading but central banks communications do not make exciting reading material. Wouldn't it be great if a model can read and perform sentiment analysis on central banks' communications? One way you could achieve this is by tediously coding prescribed rules, or you can use machine learning techniques to save you time and effort.

How do we go about doing this?

The first step is to build up a lexicon, a vocabulary of words from which the model can learn. In this case, the monetary policy committee (MPC) meeting minutes form a convenient source of text data because they come classified.

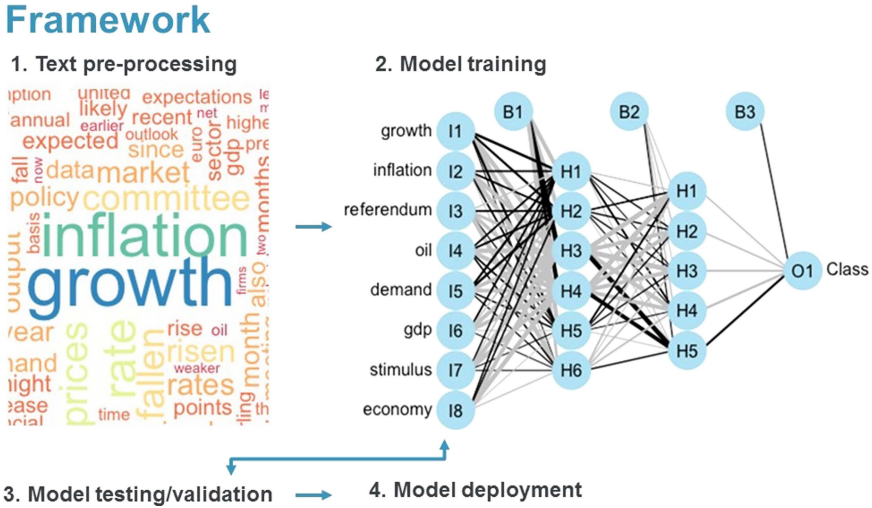


Figure 4. Framework

For example, if the MPC decides to reduce interest rates or increase QE, the minutes should reflect a dovish sentiment. These minutes are in the form of PDF files and they have to be pre-processed for it to be used by a supervised machine learning technique.

What you see in Figure 4 is the word cloud of the processed text data. One element of the pre-processing that I have done is what is called tokenisation, meaning breaking down the text into individual words.

Data or text pre-processing is the difficult part. Once that is done, machine learning takes over building the model, learning from the words and the associated classification.

In the figure is a visualisation of an artificial neural network, the poster child of machine learning techniques.

This illustrates the learning process which the model goes through. The tokenised words are the inputs. They are given weight and are passed through layers of non-linear functions or neurons, arriving at an output function which classifies text into one of three categories: dovish, neutral or hawkish.

This learning process will be repeated until the training sample classification error is minimised.

Once you are satisfied with the first iteration of learning, you can validate the model by making it classify text it has not seen before. If the out-of-sample performance is not satisfactory, you repeat the learning process. This learning and validation process can best be described using an analogy with which we are all familiar.

Say you are studying for your Actuarial Fellowship exams, you probably start with the core reading as the input. You learn the core reading over a few months arriving at an output function which is your understanding of the exam material. You then validate the output function on exam day by doing the exam. If you fail, you go back to learning. This time, perhaps, including previous exam questions as one of your inputs, or you can learn the exam material more deeply.



This process continues until you qualify as an actuary. Similarly, once it is trained and validated, the model can be deployed to do the specialist job it is trained to do.

**The Chairman:** I know that Alex (Panlilio), Zhixin (Lim) and Valerie (du Preez) are working on writing up these case studies as part of a paper.

Henry Mungalsingh is now going to talk about work stream four. Henry works for Capita Employee Benefits. His original background was at the Cass Business School.

**Mr H. Mungalsingh, F.I.A.:** I am going to cover some of the implications for professional affairs. We are trying to decide on the appropriate narrative. To give two extreme narratives, actuarial science could fit into the data science set or data science could fit into our actuarial profession.

Once we have established our professional identity we can start tackling harder issues like the implications for the exams, CPD and external regulation.

Figure 5 covers what we have done so far. In the third quarter of 2016 the original group had five members and they drafted four opinion-based essays covering the five areas of education, internal/external regulation, thought leadership and membership needs.

Based on that work we produced an update which went to the IFoA management board.

In the last quarter of 2016 the membership survey was carried out and we provided an update to the General Insurance Research Organising Committee. We recruited four new members, of whom I was one.

By the first quarter of 2017 the plan was quite clear. It is broken up into short-term wins and longer-term objectives. In the short term we have managed to obtain approval to produce a quarterly data science bulletin. It is pretty straightforward, will be one to two pages in length and the intention is to provide information about some of the latest things that are happening. The purpose is not so much to educate the membership but to excite and engage them. The first issue is ready to go out and is going to be put on the IFoA website. It is going to complement the IFoA general newsletter. There are longer-term plans to develop a data science hub that is going to be located on the IFoA website in

May 16	Agree on scope and plan for short paper ✓		
June 16	Drafting of four short opinion based essays covering the five areas of interest.		
July 16	Discuss essays and write paper for IFoA management board meeting ✓		
Aug 16			
Sep 16	Member Survey ✓	MAID update to GIRO ✓	Workstream expanded to include four new members (nine in total) ✓
Oct 16			
Nov 16			
Dec 16	Short term – quick wins: production of a quarterly data science bulletin for IFoA members	Construction of a data science organisation relationship map	Longer term - hypotheses for testing
Jan 17			
Feb 17			
Mar 17			
2017+	Strategic Paper to Management Board + Council		

**Figure 5.** Implications for the profession – update; MAID, modelling analytics and insights and data; GIRO, General Insurance Research Organising Committee; IFoA, Institute and Faculty of Actuaries

order to improve member engagement. We have also managed to construct a data science organisation map which I will discuss shortly.

Finally, in respect of the longer-term objectives, one of our colleagues, Alexander Hanks, has put together a number of hypotheses. We currently have 13 but we are hoping to reduce that number and collect data, evidence and feedback as an iterative process.

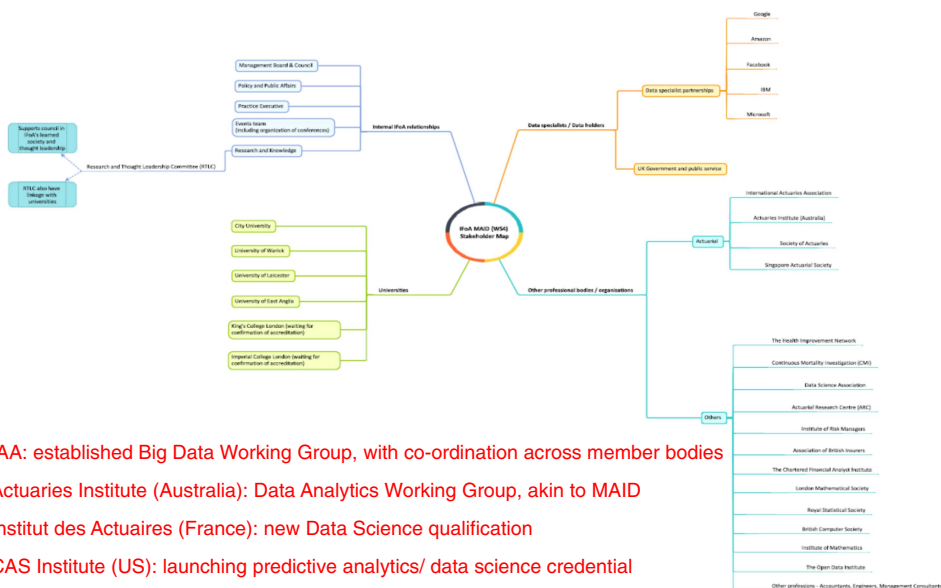
The last thing for 2017 is that there are plans for our findings to go into a strategic paper with some recommendations.

There is an organisational relationship map in Figure 6. It is as comprehensive as we can make it, but naturally we expect it to be narrowed down over time.

It is broken down into four areas: those that hold data or specialise in data, for example, Google; other professional bodies both actuarial and non-actuarial; the universities; and finally internal elements within the IFoA. The idea is that we will need to determine the meaningful relationships and how we can maximise value out of them.

Regarding what the other professions are doing, it is pretty clear that most of the other actuarial professions have initiated a response. For example, the International Actuarial Association has established a big data working group that is looking to collaborate with various other professions.

The Actuaries Institute in Australia have established a working group very similar to MAID. They have a slightly different focus, although there are some overlaps. They are focusing on vision, their own community and employer engagement.



- IAA: established Big Data Working Group, with co-ordination across member bodies
- Actuaries Institute (Australia): Data Analytics Working Group, akin to MAID
- Institut des Actuares (France): new Data Science qualification
- CAS Institute (US): launching predictive analytics/ data science credential
- SoA (US): developing predictive analytics certificate

Figure 6. What others are doing; MAID, modelling analytics and insights and data; IFoA, Institute and Faculty of Actuaries

In both France and the United States they are developing their own data science qualification. For example, the Institute of Actuaries in France are developing something called the Data Science Award and that is going to cover data science techniques.

The institute in the United States is producing something similar. There will be a general data science qualification applying to both actuaries and non-actuaries. They have plans to produce some more specialist qualifications going forward.

So, it is pretty clear that the other actuarial bodies have initiated their responses. I think that is progressive.

The last two figures (Figures 7 and 8) are tied together and relate to our identity as a profession. We know that our work is being impacted by data science but it is hard to see very far into the future in this area. We are, however, making sensible steps forward. As we continue to gather more evidence and data we need to address some key issues.

For example, we want to make sure that all our members have skills in data science techniques. We still need to determine what skills and what depth of expertise is required. This is, naturally, going to have an impact on the exams syllabus and CPD.

The work of actuaries is being directly impacted by advances in data science and this trend is accelerating – need to be clear as a profession on our role in this new world.

**Exam system, CPD and life long learning:** Need to ensure new and current members have the right skills in data science techniques.

**Data security and ethics:** The key issue for everyone, including actuaries, to address.

**Regulation and reserved roles:** Need to consider how to balance this with working in evolving areas of data science.

**Collaboration and coordination of efforts:** changes in data science are impacting everyone – need to identify the best way to collaborate and coordinate efforts.

**Public relations message on data science:** How do we define and present ourselves as a profession.

**Research programme:** Need to focus research and make advances in actuarial areas of data science.

Figure 7. Key thoughts



Figure 8. Key things to think about; IFoA, Institute and Faculty of Actuaries

We need to think about things like data security, the risks associated with handling data and the related ethical issues.

We want carry out our own research and advance actuarial science, data science techniques, or what we consider advancing data science, bringing in the skills that actuaries have.

We also want to consider what the most meaningful collaborations are. There are potentially quite a lot, as shown in Figure 6.

We also want to consider the public relations message. I will leave you with these philosophical questions: what is our professional identity? Who are we as actuaries? Going back to the two narratives. One view is that we are a long-standing profession that operates in a number of specialised areas and that we want to continue to go down that road. Given this, do we want to equip ourselves with data science techniques that provide us with meaningful insights in our specialist areas? Another view is to expand our remit and move out of the specialised disciplines towards being “Jack of all trades”.

These are the questions that we have to try to answer. We are confident that we can do so as we continue to gather more evidence and test the various hypotheses.

Based on our investigations we should be able to draw meaningful insights as to the impact on internal and external regulation, our PR profile, exams/CPD and how much research to undertake.

I conclude by saying that we are all working together and we hope to work with the entire membership. That is why the next 20 minutes should generate some meaningful discussion. We are really keen to find out from you what you think.

**The Chairman:** I will be very happy to take any feedback on what you have seen. But, more particularly, we would be really interested to hear your perspective on the questions floated and what should be the strategic approach for the IFoA.

**Prof A. D. Wilkie, F.F.A.:** We have heard about a lot of interesting applications. Most of the applications were described in such a way that I can imagine that one could apply classical or Bayesian statistics to them instead, or as well.

The only methodology you really talked about was when Zhixin (Lim) showed a picture of a neural network. Is that the methodology that you are using throughout and calling machine learning or is machine learning a much wider thing than simply using neural networks?

The advantage of classical methods is you have a precise thing which, generally, you are optimising, for example, maximum likelihood. You can obtain standard errors, parameter estimates, and lots of other relevant information. You can tell which of the various inputs are most useful and, within the model, which of the parameters and terms you are choosing to input are significant or which ones you can do without.

For example, within the neural network, shown in Figure 4, in the first column of inputs how many of those were useful and how many would you want to miss out? You then have two columns of neurons, but was that the best structure? Would you want three columns of neurons? Would you

want more neurons or fewer neurons? Is this all well documented so that you can tell? At the end of it, do the results differ significantly in their accuracy from classical statistical methods?

What Alan (Chalk) described earlier on was not quite what actuaries at the Continuous Mortality Investigation (CMI) Bureau have been doing on mortality. They have been using classical statistics using maximum likelihood estimation. In Figure 2 it would have been useful to see not only the dots showing the estimates but arrows showing the upper and lower confidence intervals of where the mortality rates might be.

The black dots obviously did not go in the same line as the position of the red line which represents assured lives, male non-smokers, centred on the year 2000. You did not go beyond age 80. There is quite a bit of data beyond age 80 in CMI for assurances and certainly for annuities. Not knowing the actual numbers, however, it was not so easy to see the significance of not examining the oldest ages.

With the traditional CMI methods and modern computers, the graduation only takes a few seconds. It competes on an equal basis with machine learning so far as speed is concerned.

I have made a lot of statements. My questions are: have you compared the neural network methodology with classical statistical methodology or Bayesian or Markov chain Monte Carlo approaches? I do not mind which statistical methods you want to use. Can you get the same answers out of the neural network system as you can from classical methods?

**The Chairman:** Thank you, David. Excellent points, but with all due respect, I think we will wait until the end to respond. We ought to start by surfacing the wide variety of issues. But you make a very good point. I know that there are some very strong views about making sure that people understand classical methods. As I said right at the start, computer scientists do not necessarily understand maths in the way trained mathematicians do, while some do, it's not 100%, and we need to be open-minded on their contribution.

**Dr L. M. Pryor, F.I.A.:** I am an actuary. I am also a computer scientist. My PhD some years ago was in artificial intelligence and I have spent the past 25 years moving between the two areas. I think that the work that the MAID working party are doing is absolutely excellent. I am very pleased to see this coming together of my two disciplines.

I have some points which I should like to make which I think might help us all think about the relationship between actuaries and big data. The first thing is that we were pioneers of big data. Back in the days when the actuarial profession started, we essentially looked at mortality statistics, and in those days the data that we actuaries were handling were absolutely huge for the technologies that we had available. We developed some very effective ways of handling it using what was available at the time.

What we are seeing now is much more data being available but also new technologies coming on stream. If you look at the power of computers nowadays there is a complete sea change in what is available.

In the past, actuaries have adapted to new technologies and developed new ways of doing things. For example, we do not use commutation functions any longer. Let us adapt to these new technologies.

There are nowadays an awful lot of machine learning big data toolkits out there. You can download a package in practically any programming language you want to use, throw your data at it and the people who write these toolkits say that you will get useful stuff out.

To my mind, that is simply not true. One of the really important things about handling data is recognising (a) what you want to learn from it and (b) which bits of it are likely to be useful.

For example, supposing that you are looking at burglary statistics and you want to find out something to do with burglary, such as when burglaries are likely to be most prevalent. Further suppose, you have all sorts of statistics including latitude and longitude. Perhaps you can see some varying effects in that different localities have different burglary frequencies at different times, and there does not seem to be any rhyme or reason.

In fact, if you pre-processed that data so that instead of putting in latitude and longitude, you put in day length or hours of darkness, you might very well see some useful information. You can do that because you have essentially postulated the form of a relationship in the data and that can be very powerful. It means that the machine learning is not having to come up with that relationship, you have done it straight away. That can be both very powerful and a huge limitation in that you have limited what is possible for your algorithm to produce.

However, experience shows that in many more complex domains that is absolutely the right approach. When you are doing speech recognition, you have a vast number of data points based on soundwaves. Speech recognition experts use machine learning to do all sorts of things such as determining if two samples represent the same person speaking. They use regular signal processing techniques. They do the Fourier transforms themselves and feed the results of the Fourier transforms in as data.

They also feed in grammars about what you can expect to happen. Different languages have different combinations of phonemes which are possible in various orders. You can get fantastic results like that but you do not get them just by throwing the raw soundwave at a machine learning algorithm.

This I think is where actuaries can really come into their own in our traditional fields of life insurance and pensions. We know a lot about what things are likely to be significant and likely to be helpful.

I have just been involved in a project where we are looking at pensions reform in an emerging Asian nation and we have a certain amount of data, about 16 years of data, of about 13 million to 15 million people. We wanted to think about how earnings progression worked over peoples' careers.

This was not machine learning in that we split the data up into test data and training data, trained on the training data and then compared the results against the test data. What we were looking at was trying to find ways of grouping, patterns that one could see in the data.

We used *K*-means, a simple clustering technique. In fact, we found various sets of patterns. We came up with about five or six different sets of patterns.

We used those sets of patterns in our projections to see how much difference it made whether we just used 512 different patterns or 2048 different patterns, or whatever.

So, big data is not just about machine learning. It is about analytics. It is about different ways to gain insights into the data. This is an area where I think actuaries have a lot to say because we know what sort of information we might find useful for other techniques.

I do not think that big data analytics should be seen as competing with traditional statistical techniques. It is a question of adding to our toolset so that we have a whole armoury of tools at our disposal that we can use to bludgeon the data into submission.

**Mr A. M. Slater, F.I.A.:** I am glad that I am speaking after Louise (Pryor), because she has said a lot of what I wanted to say.

If anyone has been unfortunate enough to go to a FinTech Conference, you will see that there are loads of 20 year olds in black roll neck jumpers getting excited that they can solve the problems of the world by just getting data and throwing it at a machine learning algorithm. It is quite clear that they have not got a clue about what they are getting excited about.

The fact that they do not understand the data is self-evident. I totally support what Louise (Pryor) has been saying: understanding the data is absolutely critical. Let us be blunt. The unique selling point of actuaries has been gradually undermined by spreadsheets. There are better IT platforms available, of course. But fundamentally, spreadsheets have undermined what was our unique selling point of being able to cope with long-term pension and insurance liabilities. And it is about time that we woke up to that.

More fundamentally, what we are observing is a view of mathematics which is moving from looking for close formed solutions to looking for algorithmic solutions. I totally support algorithmic solutions. They can do far more powerful things than close formed solutions. They are both alternative variations of the same mathematics. Do not think one is superior to the other.

It is overdue to have a discussion in this Hall questioning our identity as a profession. My view is that societal challenges other than data science could equally be debated. I look forward to hearing views from others.

**The Chairman:** What other topics would you like us to challenge ourselves with?

**Mr Slater:** Clustered regularly interspaced short palindromic repeats. This is the gene editing technique which has caused great excitement in the biological sciences. Fundamentally the development in science from the biological field is very important and will influence us in a major way at some stage.

**Mr J. M. Affolter, F.I.A.:** I work for Charles Taylor. I have been a GI Actuary for about 10 Years. I would just like to confirm that I agree with some of the other points made. I think actuaries are the original data scientists. We should have confidence going into this area.

One of the benefits we can bring is helping with the choice of methods to apply. Sometimes a simple method is appropriate, sometimes a complex one is appropriate. Actuaries are well equipped to make those type of choices. It is important that actuaries learn the practical computer programming techniques.

I think if the profession wants to do more than just become Solvency II experts, this is an important crossroads. Maybe learning these techniques should be compulsory as part of CPD.

**The Chairman:** I am sure, having sat down with Council 6 weeks ago, that there is a lot of enthusiasm for being more than just regulatory Solvency II experts.

**Mr P. J. Lee, F.I.A.:** I am on Council, as indeed Louise (Pryor) and Andrew (Slater) are. I think that this meeting is very timely. There is a threat and an opportunity. The threat is accelerating. I was at a Microsoft webinar last week, a full day conference that they had in Redmond. They are claiming that in the last year they have made progress that they thought would take another 5 years thanks to deep learning.

The theme of the conference was geeks versus hippos. Hippos are highly paid professionals. Think perhaps of us versus geeks who are the people that Andrew and Louise were talking about, who understand modelling but not necessarily what the data means. They were saying that, across the board in their experience, geeks are beating hippos.

If that is the message coming from Microsoft, then we need to consider the other side of that argument. It is not all bad news. Accenture did a study of 1000 companies that have been using machine learning for quite a while. They found that some new jobs are going to be created from deep learning. They fall into three categories. There are trainers, explainers and sustainers. There is an article which you can search for on that topic.

In brief, the explainers are likely to be very important because with the EU's new general data protection regulation coming in from next year, consumers will have a right to explanation, which means that they can question and fight any decision that affects them made purely on an algorithmic basis. There will be roles for people to explain what the models have been doing and why decisions have been taken. That may bring us into the picture. There will also be new ethical roles.

I think that the threat has been accelerating. From the point of view of the future of our members we also need to accelerate this conversation.

**Mr T. G. L Jowett, F.I.A.:** I should just like to ask that the IFoA really push forward with developing education for data science. I look back and remember when I was asking people if we could set up something for risk management. It took several years and I think we somewhat missed the boat on risk management.

I would like to see something developed very quickly on data science. I agree with everything that has been said, that many of the people that download software packages and analyse data have not got a clue about the data or the maths but they do know how to press buttons. I think we should be in there adding our intelligence to what is going on.

As the power of computers and software increases with Moore's Law, I wonder how soon it is going to be before some of these analyses are being done so automatically that we do not quite know what is behind them. I think there could be ethical issues. For example, I am sure that with distinct data sets that individually contain anonymised information, some machine learning techniques will find ways of connecting them in such a way that it is no longer anonymised. If the people and the companies that are using these databases do not really know the potential implications, how long is it before we lose public confidence and there is some kind of scandal? I think the ethical area is



something that is going to be critical going forward. Maybe that is an area where actuaries can really add value.

**Mr M. Walters:** I have a friend who is a data scientist and it is an ongoing debate between the two of us whether the actuarial profession is catching up with the progress that data scientists have managed to accomplish over the past few years.

The narrative, as you have said, is regarding transparency. Data scientists create a black box whereas actuaries will be able to explain how models work and the results that models they produce. I think transparency and regulation in the end will bring actuaries to the forefront.

There must be protection for the public. Various forms of automation and machine learning are able to provide efficiencies but at what expense to the public? The narrative must be that actuaries are entering the data science world not entirely to compete and bring efficiencies but also to bring the element of the actuarial profession where we build integrity and transparency into any work that we do, and how documentation of that is possible. It is great that the profession is doing all this work in terms of engaging both industries.

I hope at some point it will be the norm that data scientists and actuaries are working together as part of a business model. I commend this initiative.

**Mrs N. King, F.I.A.:** My background is as a pensions actuary. For the last 4 years I have been working in workforce analytics. It would be great to hear what the views of the working party are about non-traditional professions or people working in non-traditional fields of actuarial work and how data science impacts them.

I have experience in lots of different subject areas. We work with data scientists. I echo the idea that actuaries can play a role in understanding a business problem, shaping the analysis to be carried out and, more importantly, interpreting the results.

An important issue is whether you can make an intervention that makes sense in the business context. I also echo the ethical, legal and data privacy concerns that have been expressed. We work with many companies to help them decide on their policies. They need to consider how they use data on their employees and whether it is truly anonymous and if it can be disaggregated.

We have come across some interesting issues when looking at things like recruitment and promotions. How do you ensure that you are starting from clean, raw data and the results coming out of your model are true and do not reflect inherent bias, for example, around diversity and inclusion?

**The Chairman:** Mr Trevor Llanwarne, a representative of work stream 4 has asked us to obtain some feedback from the audience. If you feel excited by what is going on in the data science movement put your hands up. If you feel frightened by it put your other hand up.

The audience finds the developments generally positive.

**Mr Llanwarne:** It would be useful to have some thoughts on what people think is the top priority for work stream 4. Is it education? Is it professionalism and regulation? Or members' needs or thought leadership, which Henry (Mungalsingh) mentioned earlier.

**The Chairman:** Other possibilities might be: regulation and risk management; ethics; PR and image; or collaboration.

I think we should find a way of giving people a chance to give more considered views after this, ideally in the next 2 or 3 weeks, because that is the timeframe for putting together our strategy paper for Council.

**Mr T. J. Birse, F.I.A.:** To follow-up a few points about education in the 2019 curriculum, which is currently being developed, there are modules which require students to deal with statistical questions, and also to do much more work modelling in Excel.

This is a move towards data science. The interesting thing is that in developing the material for these subjects, the volunteers that we have had so far have all been from academia. It would be really nice if there were practitioners who felt that they were expert in these areas who would volunteer to help. They would contact the education team in Oxford who would happily take any volunteers to help with the 2019 curriculum development. This is an important area and it would be good to receive broader inputs than just the views of academia.

**The Chairman:** I am now going to give Alan (Chalk) the challenge of responding to the comments of Professor Wilkie.

**Mr Chalk:** Professor Wilkie made about 10 or 15 different, complicated points in his brief comments. I only recorded five of them. But I think they are fascinating and those points do address a wide variety of issues. Each of those five could take about an hour to talk about.

I will just repeat them back. David said first that Bayesian techniques could be used to solve much of what we saw today. He then said that those techniques were in some ways superior to what people call the machine learning techniques because you can obtain the standard deviation of the parameter estimates and the standard deviation of your predictions very clearly.

He then went on to say that machine learning techniques are also inferior because you cannot tell formally which of your input variables are most important and thus you cannot communicate this information.

He then commented that if we are using neural networks the best structure is difficult to determine. For example, how many neurons do we need and how deep does the network need to be?

Finally, he pointed out that the generalised additive model I used to discuss mortality modelling was relatively simplistic compared to what the CMI board are currently doing.

That is just a very broad summary. Each of those points needs some discussion. But to say one or two things, the American Casualty Actuarial Society is introducing an exam. They are not calling it data science. They are calling it modern actuarial statistics. In that exam there is a vast element related to Bayesian work. It is about giving actuaries the exposure and the toolkit to use Markov chain and Monte Carlo approaches in very easy ways, given the software which is current available so that actuaries feel comfortable applying those approaches.

I think we have to be careful when we talk about the fact that we need to know data science. In fact, what is the difference between data science and statistics? A cynical view is that the only difference

between statistics and data science is how big the grant is. If you say you are doing statistics, you get a £50,000 grant. If you say you are doing data science, you get half a million grant! However, there are a great many non-linear techniques, which are not traditional statistics, which we need to have in our toolkit so we can approach problems in the best possible way.

Standard deviation is a difficult problem in data science. It is not easy to derive them, it has to be done in quite a convoluted way. Although, you can actually take some of these black boxes and figure out what is going on inside of them, which was not the case a few years ago.

How do we optimise the number of neurons, and so on when using neural nets? Optimisation is a very big and complicated area, but it can be done and we actuaries have to know that it can be done, and roughly how it should be done so that if you have inexperienced data scientists in your team you can give them appropriate guidance.

Overall, we are not necessarily saying that machine learning is the only way, or even defining machine learning. Just that actuaries need to be aware of the modern techniques that are being used.

In terms of the power of these techniques, I fitted the best possible generalised linear models that I could do using machine learning techniques in this case. I then fitted some non-linear techniques. Many people have done this. The additional power in a general insurance, personal lines arena, is usually in the region of 7%–10%, depending on exactly how you measure performance. In this particular arena that was significant in terms of the overall profitability of the client so that it was worthwhile to consider whether we should move forward on that basis.

The techniques that are available are more powerful than the generalised linear model techniques that are easily available. I do, however, acknowledge that the CMI do things in a far better and more complicated way than I fully understand.

**The Chairman:** It did occur to me, David (Professor Wilkie), when you were speaking, that we could do with further detailed discussion. Part of our problem is that we do not necessarily explain the classical methods and their power clearly enough to our users. There are some challenges there in addition to those directly related to machine learning and the new tools and techniques. Maybe it is a conversation to be continued.

**Prof Wilkie:** I did not say that the classical techniques were better. I said they can do some of the same things, and it would be useful to compare them with other methods using some criteria (perhaps different ones) of which method could be said to be better.

**The Chairman:** Yes, and if we eventually produce an intellectual paper, then that is one of the issues that I think will be really important.

**Mr Lim:** I would add to that. David will be pleased to know that the Bayesian method actually performs better in the classification model on a standalone basis. The sentiment analysis model is based on an ensemble of models including neural network and random forest models.

**Mr Mungalsingh:** I have one comment. It draws on one of my experiences a couple of months ago when I met with the head of Capita's big data team. We were trying to determine whether any kind

of collaboration could work. He had about six or seven data scientists who worked with him. He said: “Henry, one of the problems we are encountering right now is my guys love maths but they do not know how to monetise business. They do not know how to take these complex statistics and solve business problems with them”.

I think that it is unfair to say that that is a microcosm of all data scientists, but as actuaries I like to think we are known to solve business problems. Also, I do hope personally, however, we evolve, that we will place great emphasis on communication.

**Mrs du Preez:** I think David (Professor Wilkie)’s comments reinforce the point that we should be using the experience that we have as a profession. David said these tools have been there for years. We should be using the experience within the room and embracing the technology in order to give more efficient analysis.

**The Chairman:** Just a couple of thoughts about how the work will progress. First, we will be producing a paper for Council and we will try to obtain and build in further feedback. Tonight has been really helpful in that process. Second, I am sure that the MAID working party will be producing a paper, or a series of papers, to keep the discussion going.

I have a couple of other things to say. First, I have been away on holiday in India. I visited, among other things, an astronomy observatory, built probably 100 years ago. One could reflect that Galileo, Newton and so on, produced an awful lot of thinking about astronomy which is still highly relevant. It is just that we have become better and we are doing things building on that logic. I have a relatively new grandson and he has the challenge of adapting to things that can be done now in a way that I never would have. The way that we use computers, tools and techniques, will be so different. But as he goes through life learning, I hope that he will learn from the classical methods such as those of Newton and Galileo.

Our challenge, I think, as human beings is how to keep the process of learning but yet not always be reinventing the wheel. That is a challenge.

In conclusion, tools and techniques are very important. We need to make sure that we are always up-to-date with the most modern and the most relevant ones. We need to understand the assumptions of the tools and techniques used. We need to be able to explain those assumptions to others. As professionals, making judgements, we need to put all the different available methodologies into play so that we are using the best tools and techniques. Lifelong learning is going to be highly relevant. I am sure we would all want to help the profession to achieve these objectives.

Another issue is judgement, ethics and professional insight. I know that there are some people who say that ethical standards and so on are not a differentiator. Nevertheless, I think as actuaries we are, on the whole, really well-intentioned. We do want to do the right thing for the society in which we live, for the people we work for and with whom we work.

That sense of integrity is very important and will determine how we use data, how we judge that we have the right data, how we make judgements and how we explain to people why we are doing what we are doing. Some of the values we stand for will, I am sure, endure regardless of the tools, techniques and technologies.

It does seem highly timely that council have the opportunity to consider what these developments mean for the following matters: education; lifelong learning; the way that we are regulated; the risks; the quality image that we have; our PR reputation; recruitment; the skills that we need; collaborations; thought leadership; research; and membership needs.

Thank you ever so much to all of the presenters tonight. Thank you, too, to those who took part. We look forward to the ongoing development of this subject for the greater good of society and for us as professionals. Thank you very much.