PhUSE 2016

Adventures in Data Science

Kieran Martin, Roche, Welwyn Garden City, UK

ABSTRACT

For the last few years, big data has been big news; promising a revolution in how we think and work. In this paper I will describe my own experiences as a data scientist. I will try to explore what exactly a data scientist is, what big data is, and why we should (or shouldn't) be getting excited about the new opportunities available to us. I will outline the skillset and knowledge required of a data scientist, and talk about some of the tools they use to solve the problems they face. I'll discuss some of the benefits that can come from a data-orientated approach to analysis, as well as the pitfalls that await the unprepared. Finally, I will look at how we, as statistical programmers, can be more like data scientists, and what we can do to be ready for the challenges of this new data focused world.

INTRODUCTION

Before joining the pharmaceutical industry I worked as a Data Scientist at Zurich Insurance. I recently joined Roche as a statistical programmer but believe that statistical programmers can learn much from data scientists, and in so doing make sure they remain relevant for a new era where data is king. In this paper I will aim to argue for why data science is an exciting new field and why we should all be paying attention to it.

I will give a brief history of data science, then try to explain what a data scientist is, and how they might differ from more traditionally defined roles such as statisticians. I will look at some of the pitfalls of the data orientated approach taken by data scientists, but discuss the gains that can be made as well. I will help illustrate this with examples from my career, and then discuss what programmers might learn from data scientists.

A BRIEF HISTORY OF DATA SCIENCE

Data science is a very recent concept, although it draws on older ideas of managing and analysing data to obtain the greatest value from it. John Tukey (Tukey, 1962) spoke about having a central interest in data analysis rather than being a statistician. Peter Naur then used the term "data science" as far back as 1974 (Naur, 1974), but it's fair to say that the field really got going in the last couple of decades, as larger data sets and the tools to obtain value from these data sets began to become available.

In 1996 the International Federation of Classification Societies met and included data science in the title of their conference. The Data Science Journal was launched in 2002, followed by the Journal of Data Science the year after.

The number of people calling themselves "Data Scientists" has grown tremendously even within the last 5 years: on LinkedIn it has grown from 4,540 in 2010 to at least 11,430 in 2015 (stitchdata.com). This reflects the growth in both the collection of data by companies such as Google and Amazon, but also the tools to analyse them such as R and Python® becoming more mature and widely used. IBM shared statistics in 2013 indicating that 90% of the world's data had been created in the preceding 2 years

While data scientists are often employed by technology focused companies (Microsoft, Facebook and IBM are the top 3 employers of data scientists listed on LinkedIn), Pharmaceutical companies are hiring self-described data scientists in greater numbers: in 2015 at least 90 were employed at GlaxoSmithKline and 32 at Roche according to Stitch Data's analysis. Of course, LinkedIn is a mere subset of existing roles, and many people in data science roles do not necessarily have the title "data scientist" so this is likely to be an underestimate of the true figures.

It seems clear that data science is an idea whose time has come, while it has existed as a term for around 60 years, computing power has now become sufficient that it is a reality. But what is data science?

DEFINING DATA SCIENCE

In this section I will expand on what a data scientist can be considered to be, and the roles they take on. The perspective I take throughout this paper is personal, but is hopefully reflective of the industry at large. Different data

scientists can be wildly different, coming from different backgrounds and having different approaches when it comes to handling data.

A data scientist can be considered to have three broad roles: acquiring and manipulating large amounts of data, analysing said data and finally visualising the results. Each of these roles represent a great deal of work, and one of the great problems with defining what it is a data scientist does hinges on the amount of effort allocated to each role.

I will discuss these roles in some detail, and the challenges associated with them, but perhaps the major difference between a data scientist and a more traditional analyst might be a data orientated mind-set, which I will discuss in more detail later in this section.

ACQUIRING DATA

Acquiring data is of course vital for any analysis. In a traditional scientific approach, something which those within the pharmaceutical industry are perhaps more accustomed to, we collect data via experiment. We establish a hypothesis then collect data to prove or disprove said hypothesis.

For a data scientist, however, the data is often the source of the hypothesis and as such the data usually already exists. It might be customer data collected during sales, or impressions on a webpage, or additional data from external companies: credit checks, address registers. For life sciences, it might be anonymised hospital data or prescription information. The only surety is that there will be plenty of it!

The nature of the data required will depend on the particular questions being asked. Often, data scientists may find themselves in a consultancy role working on multiple projects, so will find themselves relying on the owners of the data, which may involve multiple individuals, both those who gather the data and those who understand the underlying logic behind it. In the pharmaceutical industry, this would be Clinical Science plus data management, and possibly statistics as well. A good data scientist needs to be able to communicate with all parties to ensure that they get the right data to solve the questions they are attempting to answer.

Gathering the data will require manipulation using some kind of data skills; data sources will likely come from multiple sources and they will need to be linked in some way. Truly big data, of the order of millions or more rows, will require specialised software such as Hadoop ®, but in all cases a knowledge of data manipulation tools of some kind will be required. In many cases the data may need cleaning and manipulation before it is in a fit state for analysis. Knowing the best way to approach this in an efficient manner is key; many different approaches can be taken here, of varying efficiency.

ANALYSING DATA

Analysing data is often considered the key part of a data scientist's work, and it is certainly essential, but in many ways it is subordinate to collecting the correct data set in the first place. There are multiple tools available for analysing data, and the approach an individual analyst might take will depend somewhat on the problem at hand but also their particular skillset. Data Scientists from a computer science background might go for a machine learning approach, deploying tools such as neural networks and classification trees, while data scientists from a more traditional statistical background might go for multiple regressions or even a Bayesian approach.

Every analytical approach taken will have drawbacks associated with them which a good data scientist should be aware of. In many cases the power of the tools will depend on the level of analysis required. In many industries, there is a great deal of data which simply isn't being analysed at all, and even the most basic of analyses might provide value; in industries where data has already been carefully analysed, more in depth analytical strategies will be required.

Value can often be generated from data in a fashion similar to that seen in Figure 1

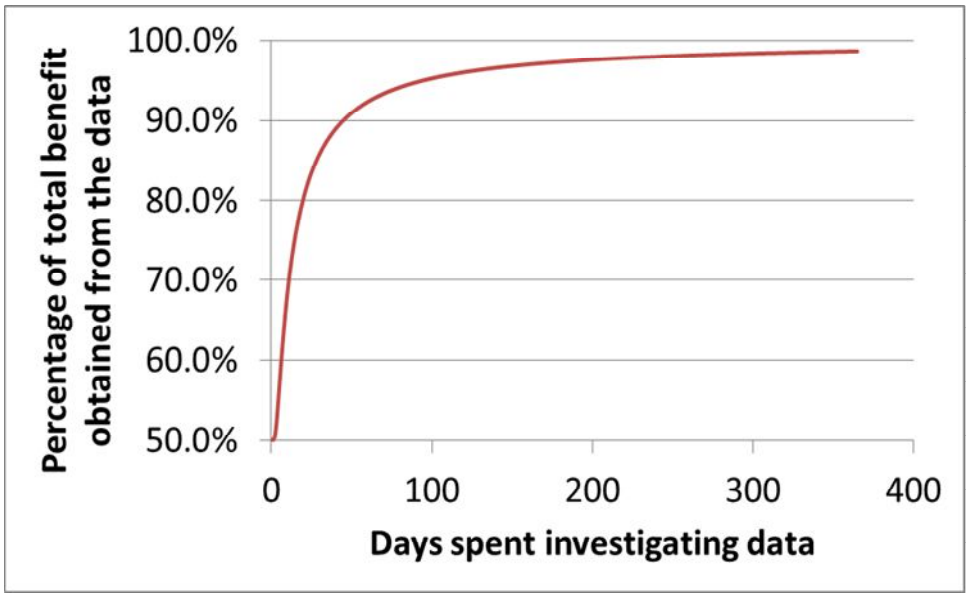

Figure 1: Value gained from data versus time spent investigating it

A lot of the value contained in data can be obtained very rapidly. Simply plotting data can tell many stories about it without even having to do further analysis. Further to that, simple analysis can produce basic insights into the data which get much of its value, while as more time is spent the gains become proportionally less. For an idea of this, see Figure 2: a linear then quadratic fit is applied to the same data. The quadratic fit is clearly slightly better, but the respective R^2 values are 0.95 and 0.98. In practical terms, the linear fit may often be sufficient for prediction, and might have been easier to find given the data at hand. It's also worth mentioning that simpler models can be less prone to the dangers of over fitting. Of course in this example the quadratic fit wasn't difficult to find, but in practice these more complex models can be more difficult to produce and more time consuming.

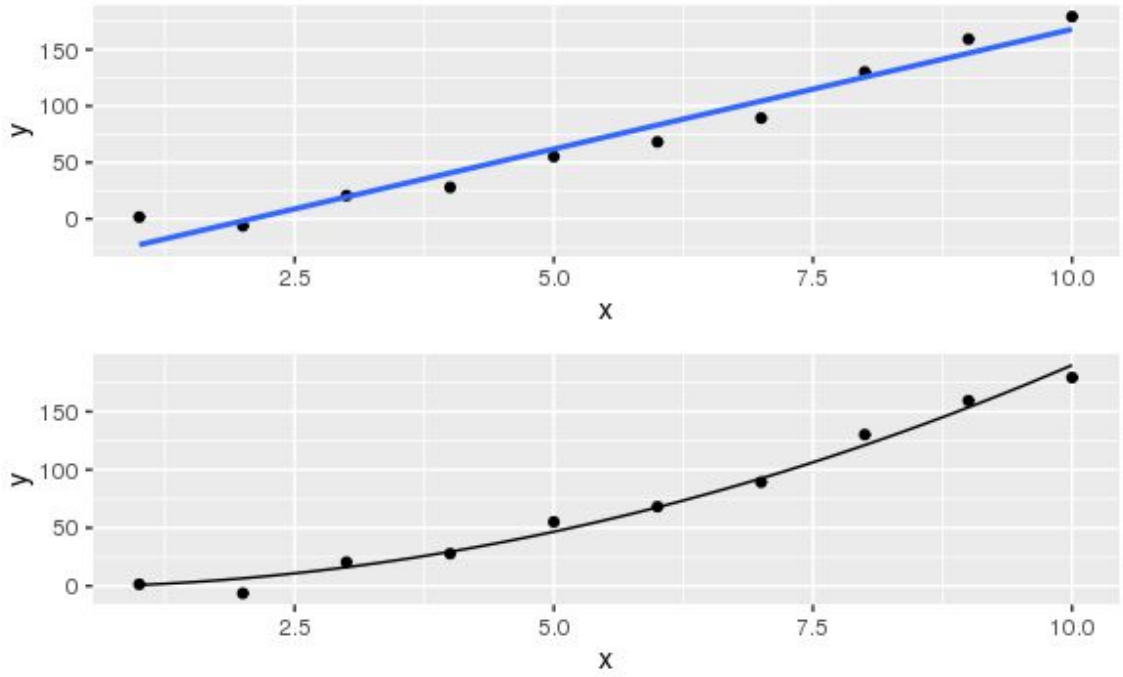

Figure 2: The same data fitted with a linear then quadratic fit

In some instances, getting the last 5% of insight from data may be extremely valuable, especially if competing with other analysts using the same or similar data. In other instances, more value might be obtained from doing fast

analyses and spending one's time profitably on investigating different data. A good data scientist will need to be able to distinguish between these scenarios, and spend the appropriate amount of time on each project.

VISUALISING DATA

Good visualisation of data was something that was often lacking even twenty years ago in the general field of data analysis, but recently the tools to produce excellent visualisations, as well as an increasing awareness of their value, means that the ability to produce excellent visualisation of your data is becoming more and more key. An analysis is only useful if it can be understood and utilised. Providing those who access you results with the tools to understand the data and the key results associated with the analysis is an important skill that all data scientists need.

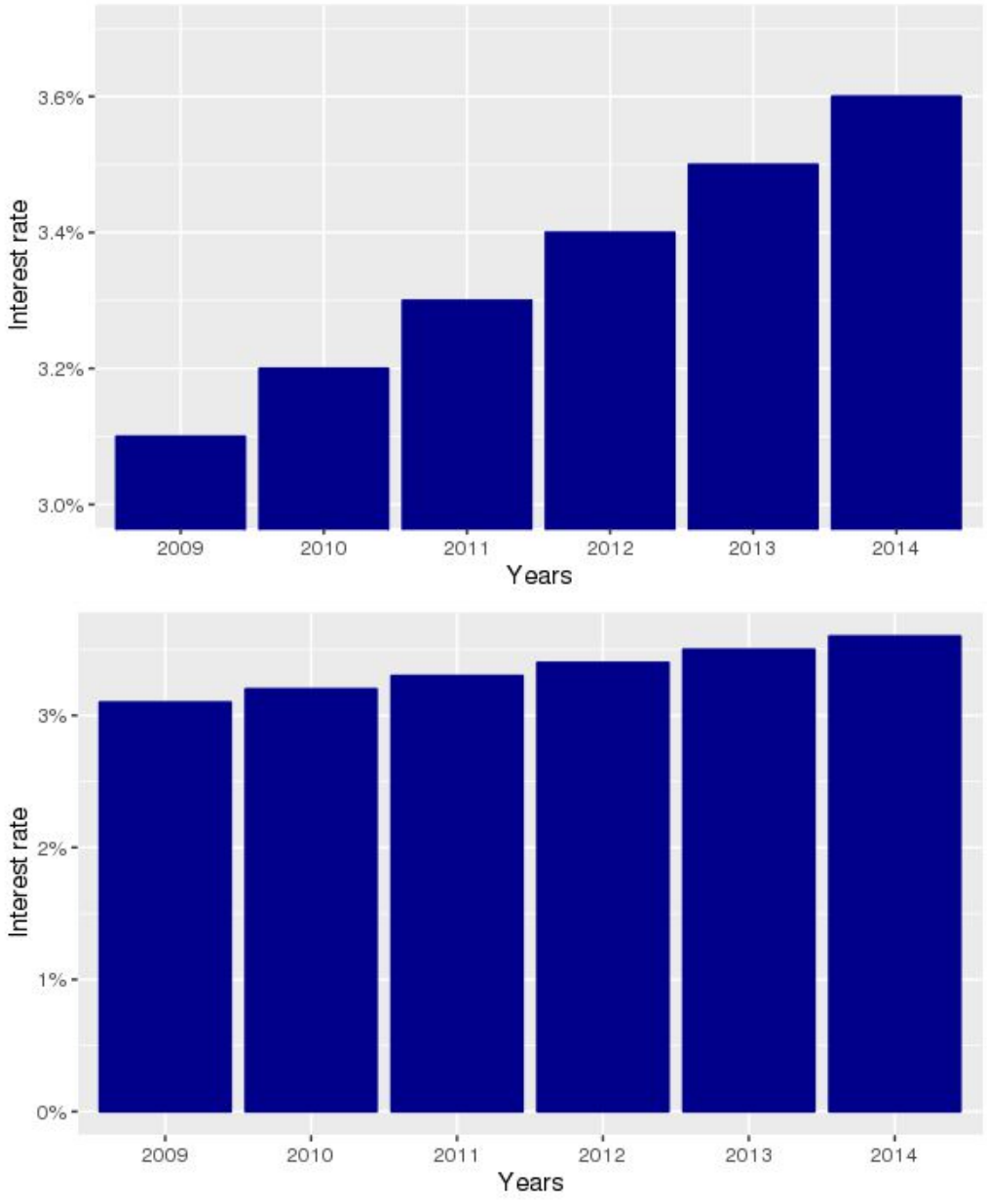

Figure 3: Two plots of the same data. One implies a gradual rise in interest rate; one implies a very moderate one. Note that more helpful than either of these charts would be an extended axes showing the historical trend, which could indicate if the recent change is unusually large or small

Fortunately the tools for producing excellent visualisations are increasing in ease of use and availability. Tools such as Spotfire® and R Shiny present exciting opportunities to involve your audience/customer in exploring the data, but in all cases it is vital for the analyst to be involved; a bad visualisation can mislead just as much as a good visualisation can inform.

Basic concepts such as remembering to set the scale at 0, or being clear when it is not (see Figure 3 for an example of a misleading chart), and keeping the chart clear and simple, and not full of extraneous information are very important for every data scientist to understand if they want their audience to engage with their data.

WHAT IS A DATA SCIENTIST?

Given that a data scientist will often be called upon to do all three of the above in their careers, in some senses a data scientist can be considered a combination of data manager, statistician and statistical programmer: certainly anyone calling themselves data scientist will probably have a selection of skills from these three fields. Each Data Scientist's specialisation will necessarily depend on where they work and the gaps they are filling.

Perhaps the most important mind set a data scientist has is a flexible one. They need to be ready to understand different kinds and types of data, and decide on the best method to analyse said data. They also need to understand who will be using their results, and the best way to communicate those results to that audience.

A DIFFERENCE IN OUTLOOK?

One thing which might highlight the differences between data science and more traditional analysis is that data science has often been more data focused. In a classically scientific approach, we should ideally approach the data with hypotheses which we confirm or deny with the data. In a data science approach, we are using data to generate the hypothesis in the first place. This follows from the idea that much data we collect simply isn't used, or isn't used to its full capacity, and we could generate value from that data by looking for the stories it is telling us. We can use both classical and newer approaches to analysis

Of course, this is a practice we should only engage in with caution, as with sufficient quantities of data we can tell ourselves almost any story we want too: for amusing examples see http://www.tylervigen.com/spurious-correlations where numerous absurd correlations are collected. For example between 1999 and 2009:

- The numbers of films Nicholas Cage appeared in in a given year has a 67% correlation with the number of people who drowned by falling into a pool (see Figure 4)
- The age of the Miss America of each year has an 87% correlation with murders by steam, hot vapours and hot objects
- The number of letters in the winning "Word of Scripps" National Spelling Bee has an 81% correlation with the number of people killed by venomous spiders.

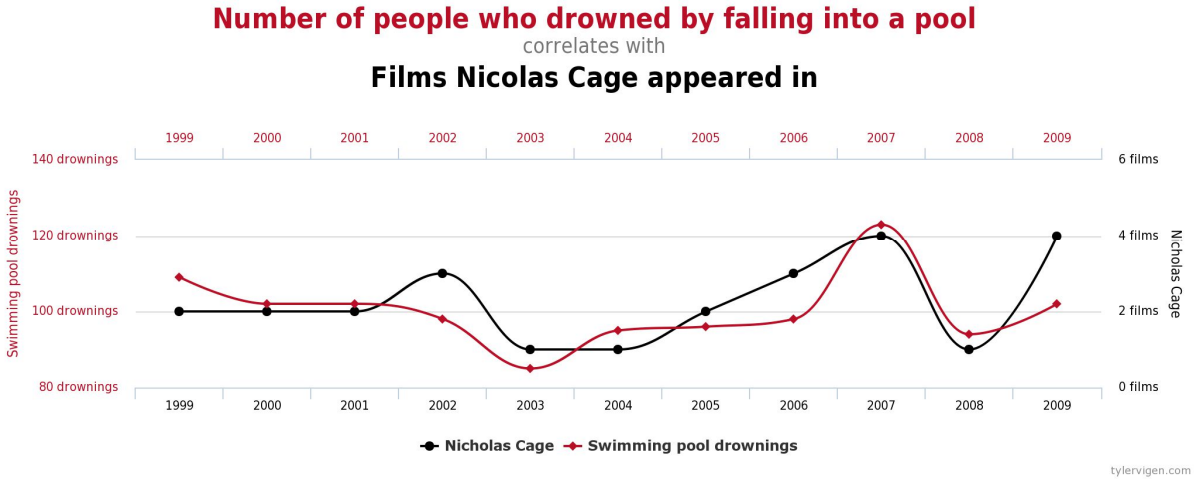

Figure 4: Nicholas Cage films against swimming pool drownings

A more public example of a failure of data science was the Google Flu tool. This was initially very successful in predicting the amount of flu that was being experienced across the US using search terms… until it wasn't (Figure 5). The reason for this sudden failure wasn't completely transparent, and was probably related to the lack of science behind the model being used. The correlation was based on a model with perhaps a lack of understanding as to why search terms and actual levels of flu were so well correlated, so when circumstances changed for whatever reason, the model could not cope.

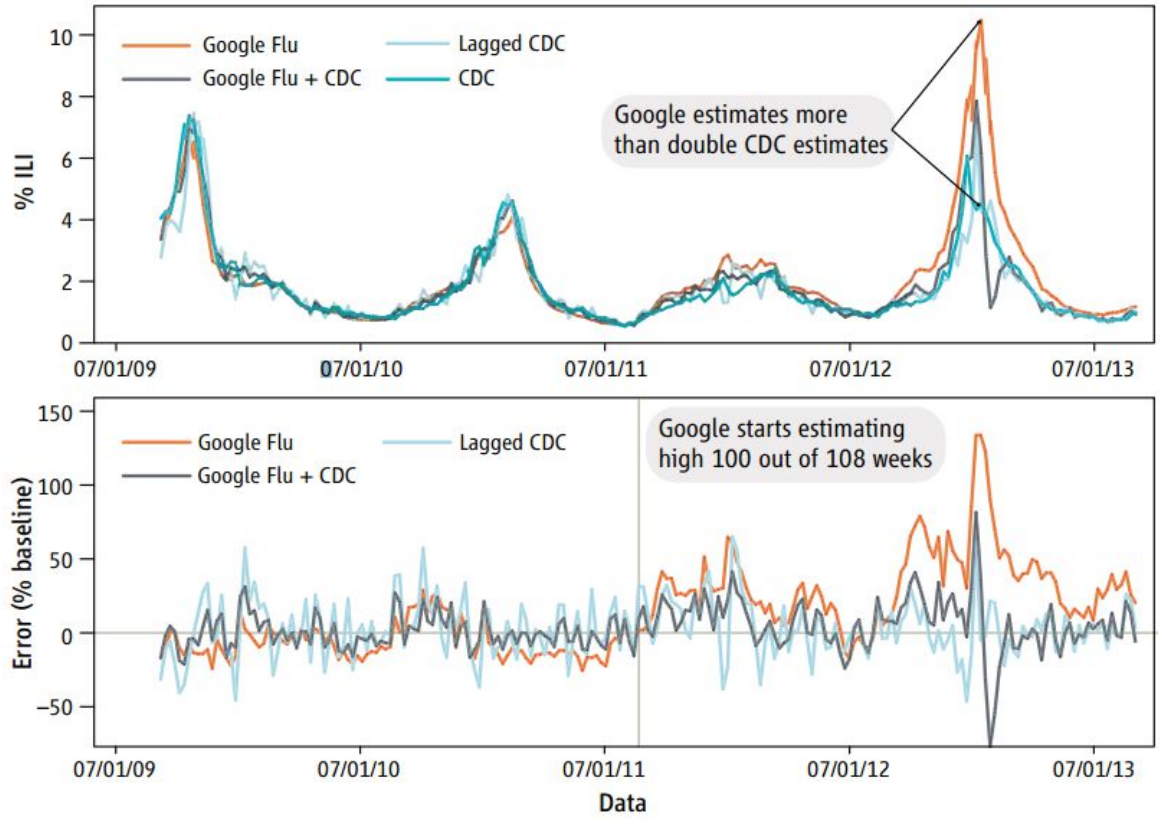

Figure 5: Google Flu against the Centre for Disease Prevention's estimates of incidence of flu. This was extremely close until 2012, when it started going wrong (Image sourced from Lazer et al 2014)

This example, and others, points to a danger in data science. Just because there is a wealth of data available, it doesn't mean we can abandon the basic scientific and statistical principles that underline good analysis, and if we do we will see large failures like this one.

Of course, a high profile failure doesn't invalidate the whole field of data science, and it would be very easy to pick out examples of models built using more traditional approaches failing as unexpected events invalidated the assumptions behind them. Models, after all, are just that, and a large shift in circumstance can always make models stop working. Still, this indicates that those who think of big data and data science as a panacea should be cautious. The problems which all scientists have faced throughout time still exist.

There is no doubt that there have been large successes in data science being applied in the last few years. One notable example was IBM's Watson, which successfully won the quiz show Jeopardy against some of the best former contestants, and the recent victory of Google's Go AI against Lee Sedol. Both of these used big data plus neural networks to beat the best humans at highly complex tasks. It is likely that these successes will only continue over the next few decades, and we should expect to see novel solutions to complex problems in unexpected areas. It behoves all of us to be prepared for these changes.

MY ADVENTURES IN DATA SCIENCE

To help illustrate what a data scientist might do in their day to day work, I will look at my career, exploring my background, and then my role as a data scientist and some of the challenges I faced.

MY BACKGROUND

I come from a fairly traditional statistical background. I studied a masters in statistics with applications in medicine at the University of Southampton before studying a PhD in Experimental Design there as well. During that time I got a background in traditional statistical approaches to modelling, as well as learning to program in R, a skill which was invaluable in my future career. In addition, I had to frequently present my work to a variety of audiences, and visualise it in a clear and transparent way. Being able to communicate with audiences of all levels is a vital skill for any data scientist.

Escaping academia, I joined the Office for National Statistics (ONS) as an Assistant Methodologist in Small Area Estimation. While my job title was not data scientist, much of what I did could qualify as data science. Applying my analysis skills to different and varied data, understanding the context and applying methods appropriately; all of these things prepared me for work as a data scientist.

I left the ONS and headed off to Zurich Insurance to begin work as a data scientist. More recently I joined Roche as a statistical programmer.

MY ROLE AS A DATA SCIENTIST

I joined the analytical development team at Zurich, a new team devoted to getting value out of the data that Zurich collected. Like many companies, Zurich collects a great deal of data, both from their customers and from external partners; our role was to obtain value from that.

Within every industry the challenges facing a data scientist are likely to be different, and I found that the challenges even varied from project to project. I worked in a number of different areas, each related to quite different parts of the business: fraud, pricing and predicting the likelihood of customers to purchase products. I needed to understand the area involved, and ask the right questions to make sure the correct analytical solution was provided. In each case, a different analytical solution was required and collecting and merging the data presented varied challenges

The data available hadn't always been collected with analysis in mind, so things like unique identifiers were missing: on one occasion, I needed to match different companies together only by their name. This was not always recorded consistently, and as it was entered manually, typos and different capitalisation meant that matching different data sets together was challenging.

This was one example where the philosophy behind Figure 1 needed to be applied. The challenge of matching dirty data is something which one can spend a considerable amount of time on, trying to optimise between false positives and negatives. I settled on a holistic approach, using several different methods, including the function *compged* in SAS ® which scores two strings on their similarities. I tested this method, but not extensively, with an understanding that time spent on this problem could be distributed elsewhere. This is always true for any analyst trying to solve a problem, perfection is an ideal which should be aimed for but should not stand in the way of delivering timely work.

To illustrate why spending too much time on this approach would be a mistake, we can look at what occurred when the analytical solution was delivered: we immediately received plenty of feedback and suggestions from the users of the tool. While we had elicited feedback beforehand, in practice end users really can't judge how good a solution is until it is provided to them. Much of the comments and constructive criticism would have been difficult to anticipate in advance, and the issues highlighted would probably have been present even if we had spent another six months honing the solution. By not spending too much time initially we were able to go back and make changes without having wasted unnecessary time on solutions which might have proved irrelevant when feedback was provided.

One example of data scientists easily adding value by looking at data in new ways was looking at how we tracked fraud. By analysing the customers we were pursuing for potential fraud using classification techniques I was able to find groups of customers where we were rarely able to detect fraud and thus were currently wasting resource. The technique I used here was a data science one, but theoretically any analytical technique could have been used. The principle being that sometimes when one investigates data there are simple wins that can be obtained.

One thing to remember about data science and all analysis is that not every analysis will be a success. Just as some trials will not have the results expected, so will some data simply be of too low quality to answer questions of interest. Again, the correct approach is to dedicate the appropriate amount of resource to each problem so not too much effort is wasted on ultimately unsuccessful investigations.

I enjoyed my work as a data scientist, in particular the flexibility in the work I approached. My hope is that much of what I learned during that time can be applied as a statistical programmer.

STATISTICAL PROGRAMMING VERSUS DATA SCIENCE

In this final section I will look at the differences between statistical programmers and data scientists and try to explore what we as programmers might gain from emulating data scientists.

Statistical programmers are, generally, part of an organisation with clearly codified roles and responsibilities. In terms of similarities between statistical programming versus data science, a focus on correctly preparing data for analysis, and visualising it, is very much shared. Analysis is something programmers will engage in, but usually with the support of statisticians.

So what marks the difference between data science and a statistical programmer? Currently, I believe that this is in agility. The environment we work in is changing, and data science is a reflection of that. As a profession, data scientists exist to take advantage of new data and new tools. There is a focus on the new, on looking at unexplored avenues of research and making use of data in new and innovative ways. Programmers are often focused on making sure that the required outputs are produced for filing. This requires precision, care, and dedication to making sure that the correct procedures are followed. This work is vital, but sometimes gets in the way of more agile work.

As Statistical programmers, can we learn from data scientists? I believe so, although the lessons we can learn depend on what we believe our role to be. Are we just there to produce the tables, listings and figures as per specification, or can we challenge what's required, and suggest new approaches? Can we get involved in exploring our rich data sets to provide value to our companies and, ultimately, patients? I think that as an industry we must move more and more to partnership with scientists and statisticians, not just meeting the basic requirements, but finding new ways to explore and visualise data.

Data Scientists need to have an understanding of their data, and so do we. We need to be having conversations with statisticians and scientists to understand the problems before they arise, and be able to find new opportunities where they are available. We don't necessarily need to become data scientists, but we do need to be ready for changes that are coming, in terms of the amount of data and the tools we use to process it. With increased standardization of our work coming from industry collaborations (CDISC, Transcelerate), then statistical programmers need to broaden our horizons so as to keep adding continued value to our business. With better visualisation techniques becoming available, are we content to produce static listings when we could produce interactive ones? Are the graphs we produce as clear and as easy to read as possible? Perhaps we could create interactive graphs? The tools are all there, we just need the will to use them.

CONCLUSION

The field of data science is certain to continue to expand over the next few decades. More and more data is being gathered, and the tools to analyse it are getting better. We all need to be ready to face this new world, or find traditional approaches outpaced by the new. There are advantages in the old, and discarding them would be a mistake, but we can and should adapt to make use of new techniques where it makes sense to do so.

In a way the tool we choose to use matters much less than the mind-set we as analysts or programmers need. There will always be a new exciting way to visualise and analyse data, and new data and ways to look at it, and what matters less is a mastery of the new tool than a willingness to approach the new, and to keep on learning

about new methods as they become available. It can be intimidating to embrace the new, but it's essential if we do not want to be left behind.

I believe that statistical programmers are in a good position for facing the future, provided we seek the opportunities that are out there. We are at a tipping point where large shifts in the industry are likely to happen. We must ensure that we are not left behind.

REFERENCES

The Future of Data Analysis, (1962), Vol33, Number 1, 1-67, Ann. Math. Statist.

Concise Survey of Computer Methods, (1974), Peter Nuar, Studentlitteratur

The State of Data Science. Retrieved from https://www.stitchdata.com/resources/reports/the-state-of-data-science

The Parable of Google Flu: Traps in Big Data Analysis, (2014), David Lazer, Ryan Kennedy, Gary King and Alessandro Vespiganni, Vol 343, Issue 6176, pp 1203-1205 Science

What is big data? Retrieved from https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html