THE BEHAVIOURAL INSIGHTS TEAM

Using Data Science in Policy

A report by the Behavioural Insights Team

Contents

Foreword

John Manzoni
Chief Executive of the Civil Service

The huge growth in data, and in tools for analysing it and putting it to use, has changed the world and continues to do so. Whether targeting adverts to us online, or using search engine data to predict where flu outbreaks will happen and ensuring drugs are available, all of us are being touched by these changes.

There is great potential for government to improve the performance and productivity of services through the smarter use of data. This data includes outcomes, use patterns, costs, and citizen experiences. With this wealth of data, we have an obligation to make government services the best they can be. This means learning from where services are working well, and improving where they are not. It means personalising and targeting public services around the needs and wishes of the individual and business. And it means using experimentation, and an understanding of within service variation, to quickly establish how services and systems can be improved.

The full realisation of this potential will not happen overnight. Like any new endeavour there will be challenges along the way. We need to be prepared to innovate, expanding approaches that work, and learning from and adapting things that don’t. This report reflects that ethos, and highlights where approaches should be refined further. But it also makes clear why this perseverance is justified – the work in this report is just the start of what we can do, and yet the results are already demonstrating the great impact that will be achieved in core government services like schools, health, and social care.

Across government we are already transforming the way in which citizens and the state engage, through the expansion of digital technology. The next step is to ensure that the data gained is constantly driving improvement. For that, the application of data science will be key.

John Manzoni
Chief Executive of the Civil Service

Foreword

David Halpern

Chief Executive of the Behavioural Insights Team

Since the creation of the Behavioural Insights Team in 2010, we have conducted over 400 randomised controlled trials and become known for our commitment to empirical methods and robust evaluation. Trials are a powerful tool to establish 'what works' and are integral to increasing the innovation and performance of government. They provide a yardstick against which both 'marginal' and 'disruptive' innovations can be measured.

If trials are a key engine of innovation, then the fuel that powers them is data.

Yet powerful though trials are, they are not the only 'tool in the box' of empirical and innovative governments. Predictive analytics and machine learning are opening new pathways to improve public services through the systematic study of subtle variations, and complex relationships, in public service experiences and outcomes.

As the results in this report show, there are many ways in which data science techniques can be used to improve government policy and practice. Regulators can use these techniques to greatly improve the detection and targeting of underperforming services, as we have shown that a machine learning model was able to direct a 20 per cent inspection rate to identify 95 per cent of inadequate GP practices. Practitioners can be helped to make more informed decisions, as we have shown that machine learning can be used to help identify children at high risk of re-entering the social care system even though their cases have been closed. Predictive analysis also offers the prospect of developing much more targeted interventions, in essence helping to identify not just 'what works' on average, but which intervention is likely to work best for whom, such as in our study of student support at King's College London.

It is a decade since Bloomberg demonstrated that smarter use of data could improve New York's public services, from making better judgements about where to send inspectors to targeting crime. In the years since, the volume and availability of data has grown across the world, as has our understanding of decision-making and behavioural science. In this report, we seek to show the potential that this has begun to unlock.

David Halpern
Chief Executive of the Behavioural Insights Team

Executive summary

The range of techniques that make up data science – new tools for analysing data, new datasets, and novel forms of data – have great potential to be used in public policy. However, to date, these tools have principally been the domain of academics, and, where they have been put to use, the private sector has led the way.

At the same time, many of the uses of machine learning have been of fairly abstract interest to government. For example, identifying trends on Twitter is helpful but not inherently valuable. Projects showcasing the power of new data and new tools, such as using machine learning algorithms to beat human experts at the game Go, or to identify the prevalence of cat videos supporting one political candidate or another, have been some distance from application to government ends. Even when they have been applicable, often they have not been adequately tested in the field and the tools built from them have not been based on an understanding of the needs of end users.

Hence, along with many others, over the past year we have been working to conduct rapid exemplar projects in the use of data science, in a way that produces actionable intelligence or insight that can be used not simply as a tool for understanding the world, or for monitoring performance, but also to suggest practical interventions that can be put into place by governments.

We have conducted eight such exemplars, focused on four areas: targeting inspections, improving the quality of randomised controlled trials (RCTs), helping professionals to make better decisions, and predicting which traffic collisions are most likely to lead to someone being killed or seriously injured. This report covers six of these eight exemplars.

Targeting inspections

- We found that **65 per cent of ‘requires improvement’ and ‘inadequate’ schools were within the 10 per cent of schools identified as highest risk by our model**. Increasing this to the riskiest 20 per cent, our model captured 87 per cent of these schools.
- Using publicly available data published by the Care Quality Commission and other sources, **95 per cent of inadequate GP practices can be identified by inspecting only one in five practices.**
- By only using the public part of the CQC’s Intelligent Monitoring system, which is based on several clinical indicators, a similar model would only pick up 30 per cent of inadequate practices for the same inspection effort.[1]
- We have also built a model to predict the inspection results of care homes, but this model is much less successful, suggesting either that more data are needed or that machine learning techniques could be of limited use here.

Improving randomised controlled trials

- Previously, we have used RCT data to study how the effectiveness of interventions varies for specific sub-groups, enabling interventions to be better targeted.
- These sub-groups tended to be broadly defined by one or two pre-determined characteristics and combinations of characteristics were largely ignored.
- By applying causal machine learning algorithms to data from RCTs, we are able to identify differential impacts of an intervention across all observable characteristics, ensuring that people get the best intervention for them, and helping to prevent backfires.
- We replicated an experiment conducted in 2016 with King's College London, in which students were encouraged to attend a Welcome Fair by being sent text messages emphasising either employability or social belonging, with the belonging condition performing best.
- In our replication study, participants were randomly assigned to receive either one of the messaging arms allocated randomly or the message that the machine learning algorithm predicted would give them the best outcome based on their observable characteristics.
- In our first study using these techniques, we found **a small positive but not statistically significant effect from allocating messages by algorithm**, which we believe is due to poorly regulated model complexity. We are improving the design of our targeting by using a consensus of models rather than one.

Helping professionals to make better decisions

- Social workers need to make a large number of decisions, very often with little time and incomplete information.
- Our previous work in this area has shown that high caseloads for assessment social workers can influence the decisions that are taken.
- Working with one local authority, we used natural language processing to predict which cases that were flagged for no further action would return within three months and result either in a child protection plan or a child being taken into care.
- Analysis using both text and structured data allowed us to predict, **8.3 times better than chance**, which cases were likely to be referred back into the system.
- Using just analysis of the text, **we can detect 45.6 per cent of cases that will return from just under 6 per cent of all cases**, allowing interventions to be precisely targeted to support the families most in need.
- We are working with social workers to build a digital tool that can be used to help inform their decisions.

Predicting serious traffic collisions

- Traffic collisions in East Sussex have bucked a national trend for fewer incidents of killed and seriously injured (KSI) casualties.
- We are able to predict which accidents will result in someone becoming KSI, **with drivers' behavioural factors, and not road conditions**, contributing the most to the explanation.
- We have been able to bust some myths – for example, about older drivers, and goods vehicles.
- Motorcyclists, the young, and people in early middle age are disproportionately more likely to be involved in KSI incidents in East Sussex.

If you would like to be kept informed of our latest work, findings and publications, subscribe: http://www.behaviouralinsights.co.uk/subscribe.

Stay in touch: email us at info@bi.team.

Introduction

The Behavioural Insights Team (BIT) has always been a part of the "what works" movement, supporting evidence based policymaking. To date, this has largely meant running randomised controlled trials (RCTs) to try to provide gold-standard evidence of whether or not policy interventions – drawn from both behavioural science and other areas – are effective. In the past seven years, we have run more than 450 of these RCTs across the policy spectrum, ranging from tax to education to international development to healthcare, in countries from the UK to Australia and from the USA to Syria.

There are, however, times when an RCT isn't appropriate, and there is a great deal that can be learned about how to make policy better without running one. Using data intelligently, and turning it to face serious problems, can help to improve the effectiveness and efficiency of existing programmes. It was for this reason that we set up our Data Science team a little under a year ago. In the first year we had a clear mission: to get out and work with government departments, with local governments, and with other organisations, to demonstrate quickly that data science can have a positive impact on public services across a range of dimensions.

What do we mean by 'data science'?

'Data science' has become something of a buzzword for 'anything innovative that uses data'. This might include visualising data, summarising it in various ways, or using it to predict events or outcomes. Our focus is largely on the last of these three, which is generally called 'predictive modelling'. This usually consists of gathering a large set of historical data, using computer algorithms to find patterns in the data that it would be impractical or impossible for a human to find, and then using those patterns either to understand the process in question or to predict where and when specific events are likely to happen, in order to plan for and respond to those events.

In a proud BIT tradition, the Data Science team was set up with a sunset clause: we had 12 months from 16 January 2017 to meet certain objectives, or the team would be disbanded. These objectives were:

1. to produce at least three exemplars, across at least two policy areas;
2. to make use of publicly available data, web scraped data, and textual data, to produce better predictive models to help government;
3. to test the implications of these models using RCTs;
4. to begin developing tools that would allow us to put the implications of our data into the hands of policymakers and practitioners.

Between January and November 2017, we completed **eight** exemplar projects, in health, education, social care, road safety. We worked with local and national governments and with inspectorates.

We worked to determine how inspections of schools, GP practices and care homes for the elderly could be made more effective, allowing regulators and inspectorates to use publically available datasets to predict which institutions are most likely to fail and thereby target their inspections accordingly. We showed that this data, married to machine learning techniques such as gradient boosted decision trees, can significantly outperform both random and systematic inspection targeting. However, in the case of care homes, the impact is not large, which helped us to articulate the limits of machine learning, or at least of the current data. We are excited to be working with Ofsted to put the insights from this work into action, and to improve on the work that has already been done.

Working with King's College London (KCL) as part of our KCLXBIT project in the 2016–2017 academic year, we've used causal machine learning techniques to identify various groups within our sample that benefitted particularly highly, or particularly little, from the messages we sent students to get them to sign up to societies or to use online learning environments. This let us work out the best intervention to give any individual, based on their characteristics and how they responded during the trial.

However, a prediction is only as good as the decisions that can be made based on it. By again partnering with KCL, we were able to test the predictions of our model in the real world. We found that, while the algorithm did improve the targeting of messages by a small amount, the effect was not significant. We are iterating the process that we use to target interventions and will continue to test and adapt our algorithms as we have with interventions since our inception.

As well as developing targeted, one-off interventions, it is important that data science be made useful to professionals and practitioners every day. In our project working with one local authority to predict future escalations of cases referred to children's social care, we conducted qualitative work alongside the machine learning to help uncover the best way of presenting findings for an audience of (often rightly sceptical) social workers, so that it could be useful to them. By bringing the human element of practice together with machine learning, we are now building a digital tool to help social workers make use of these insights in real time.

One important overall lesson we learned is that projects tend to succeed best when there is a clear predictive problem; high-quality, large-scale and preferably individual-level data; buy-in to implement findings in practice; and clear ethical and legal clearance. We recommend scoping projects using the University of Chicago's Center for Data Science and Public Policy's Data Maturity Framework.[2]

The use of data science in policy is new, exciting and full of potential that has yet to be realised. There is much work left to be done in developing actionable insights from this data and then, vitally, putting them to work. We do believe, however, that the projects we present here represent a step in the right direction.

Identifying underperformers

Care Quality Commission inspections of GP practices

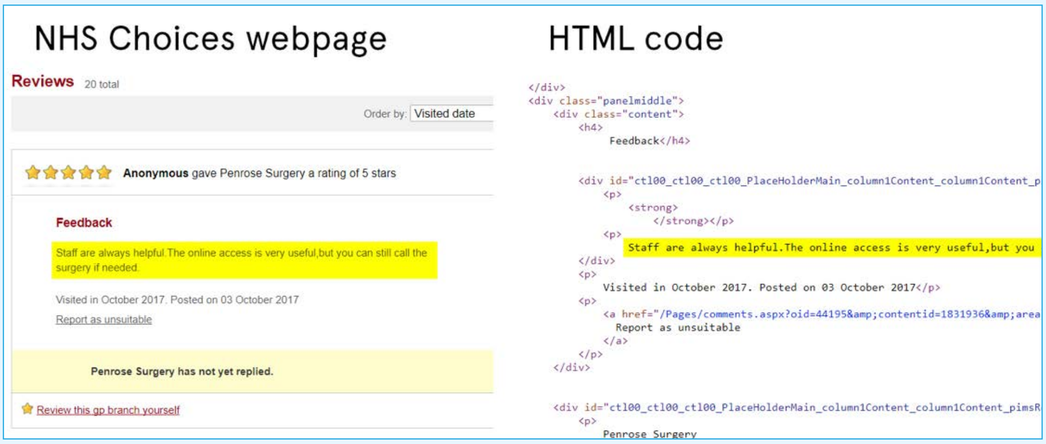
General practitioners (GPs) have a profound impact on our health – they are the first point of contact for many non-urgent medical attendances and a vital channel for public health campaigns. However, due to the high volumes and wide variety of conditions GPs see, as well as varying levels of professional skill, some practices in the UK do not meet the required standard.

The Care Quality Commission (CQC) has a duty to inspect GP practices but has finite resources in terms of personnel and time to do so and inspections also place a burden on GP practices. The CQC rates GP practices on a four-point scale from outstanding to inadequate. These inspections focus not only on clinical outcomes (i.e. patient health) but also on five broad domains, specifically whether practices are safe, effective (clinically), caring, responsive and well-led.

We investigated how we could improve existing targeting systems by trying to predict past inspection ratings using data generated from before those inspections were conducted. In particular, we looked at whether practices received a positive outcome (an 'outstanding' or 'good' rating) or a less desirable one ('requires improvement' or 'inadequate'). We found that 11 per cent of practices received low ratings, including 2 per cent that received inadequate ratings, which trigger automatic re-inspection within six months and should lead to remedial action by GPs.

We used a wide variety of publicly available data for this project. As well as clinical indicators published by the CQC, we used data from the Office for National Statistics and data on the type and number of drugs GPs prescribe. We also scraped the text from reviews left on the NHS Choices website by patients (see box, below).

'Scraping' information from websites



NHS Choices webpage

Reviews 20 total

★★★★★ **Anonymous** gave Penrose Surgery a rating of 5 stars

Feedback

Staff are always helpful.The online access is very useful,but you can still call the surgery if needed.

Visited in October 2017. Posted on 03 October 2017

Penrose Surgery has not yet replied.

HTML code

```
</div>
<div class="panelmiddle">
    <div class="content">
        <h4>
             Feedback</h4>


        <div id="ctl00_ctl00_ctl00_PlaceHolderMain_column1Content_column1Content_p
            <p>
                <strong>
                    </strong></p>
            <p>
               Staff are always helpful.The online access is very useful,but you
        </div>
        <p>
            Visited in October 2017. Posted on 03 October 2017</p>
        <p>
            <a href="/Pages/comments.aspx?oid=44195&amp;contentid=1831936&amp;area
              Report as unsuitable
            </a>
        </p>
    </div>


    <div id="ctl00_ctl00_ctl00_PlaceHolderMain_column1Content_column1Content_pimsR
        <p>
            Penrose Surgery
```

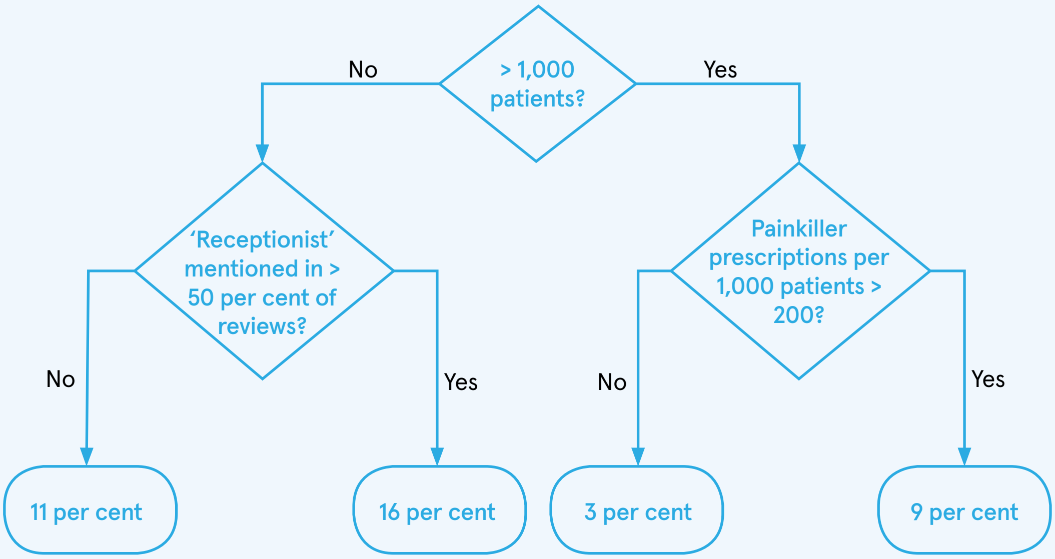
Scraping is the process of extracting and converting data from websites so that it can be used for analysis. In our work on GP practices, we used scraping to obtain the star ratings and reviews for all practices that had ratings on the NHS Choices website. This allowed us to include some aspects of the public's opinion of each practice in our analysis, which we hypothesised could help us identify underperformers. Patient (or user) reviews are often valuable publically available sources of information, but are typically published only on websites and not in a dataset ready for analysis. Scraping allows us to access the value locked away in these proprietary websites.

The image above on the left shows the way the webpage normally appears when someone is browsing https://www.nhs.uk for GP reviews. On the right is the underlying HTML code that generates the view on the left. We wrote programmes that read this HTML code and extracted the parts of it that we were interested in. For example, for each review, the programmes would grab the free-form written portion, highlighted in yellow above. They would also read how many stars each user had given the GP practice. HTML code, and websites in general, are not usually written to make information easy to extract, so custom programmes needed to be written to interpret each individual element on each page of the NHS Choices website that we were interested in.

The use of the large dataset of prescription data along with the text of 99,644 NHS Choices online reviews allowed us to understand the nuances of doctors' behaviour that would otherwise be undetectable. To capture this contingent behaviour, we analysed the data using gradient boosted decision trees (see box, below).

Using decision trees as statistical models

A decision tree is a flexible object that can describe many real-world processes and allows different data to be used as predictors depending on the context. In this example, the decision tree predicts the chance of a poor inspection result:[3]

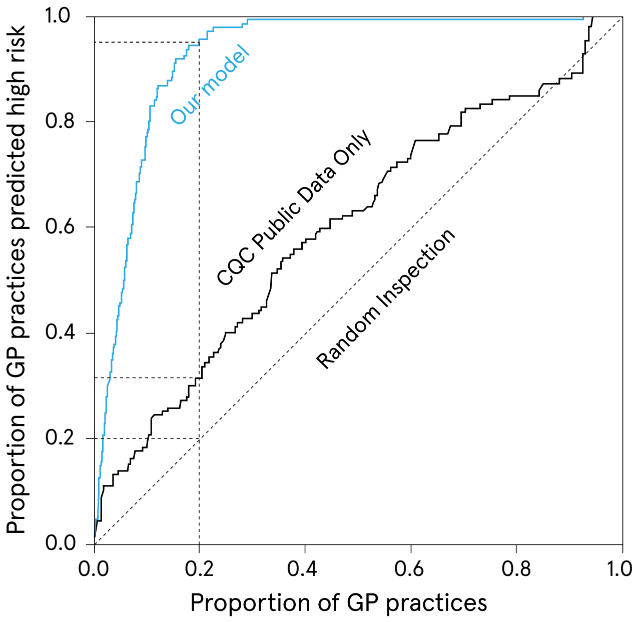

> 1,000 patients?
No
Yes
'Receptionist' mentioned in > 50 per cent of reviews?
Painkiller prescriptions per 1,000 patients > 200?
No
Yes
No
Yes
11 per cent
16 per cent
3 per cent
9 per cent

Notice how the text data is used for smaller practices and the prescriptions data for larger practices. This context-awareness is not found in many other techniques and allows us to achieve high predictive power in large and complex problems, as here.

Any statistical model can be improved by a process of iteratively refocusing it on the cases where its predictions were least accurate. As an analogy, imagine a teacher giving a child homework: if the teacher is setting follow-up work, it is better for this to cover the material the student got wrong than the topics the student didn't struggle with. This process is called 'boosting', and 'gradient boosted decision trees' are the result of applying this process to a decision tree model. The result of this is a large collection of different decision trees, which are then averaged to produce predictions.

The results from the model were surprising (see Figure 1). We could identify nearly all (95 per cent) of inadequate clinics (those that were given the lowest inspection rating) by only inspecting the 20 per cent that our model identified as most risky. In contrast, we were only able to identify 30 per cent of inadequate practices if we restricted our input data to sources published by the CQC, demonstrating the massive predictive power of these extra data sources.

Figure 1: The gain curve for the CQC GP practices model.



It was harder to detect GP practices rated as 'requires improvement': we could detect around 55 per cent of the 'inadequate' and 'requires improvement' practices by inspecting 30 per cent of them. Using just public CQC data, we could achieve only 40 per cent.

Within our model, more decisions are based on the text data than any other data source. The model appears able to pull out particular words and phrases that occur in reviews and indicate a good or bad practice.

Example of text associated with a good GP practice

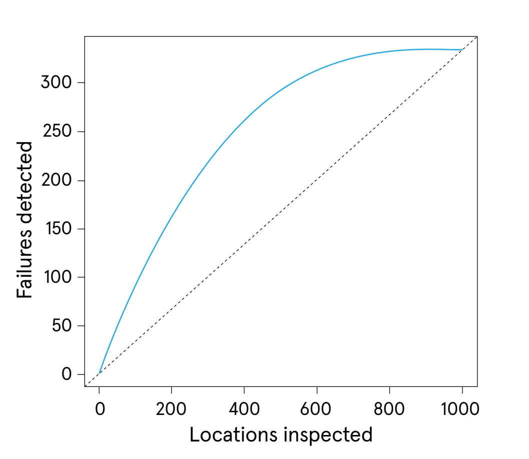
'As a result of my attendance and the excellent nursing care the condition has improved markedly for which I am very grateful. When I have needed to see a General Practitioner I have also received excellent and courteous care.'

How do we measure the performance of a predictive model?

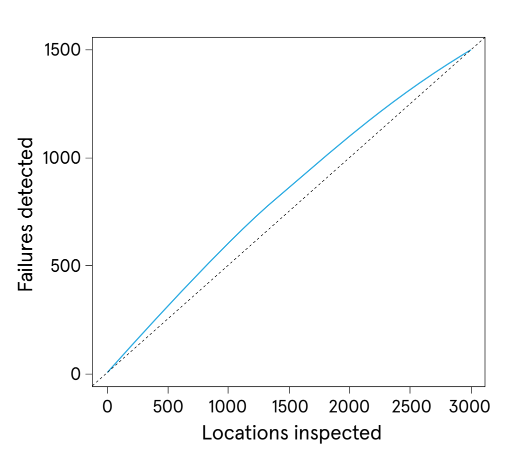
We can use historical data to compare the modelled outcome with the actual outcome and produce rates of false positives (where the model predicts an outcome which does not happen) and false negatives (where the model predicts no outcome when one does happen). These unfortunately depend on where we set the sensitivity of the model, and can be traded off against each other: a model with high sensitivity will often predict an outcome and will have a high false positive rate and a low false negative rate. The converse is true for a model with low sensitivity.

To assess the performance of a predictive model independently of this choice of where to set the sensitivity, we look at the number of correctly identified positive cases over all the possible values of this threshold, and plot that against the number of cases where the model predicts a positive outcome (regardless of the truth). For a strongly predictive model, this curve will arc up strongly, as in the left-hand example below. For a weakly predictive model, it will only be slightly above the diagonal line representing the average behaviour of selecting cases at random. This curve is usually called a 'gain curve'.

Strong



Weak



Clinical indicators were less predictive than we might have expected, in part because they miss something that the text and more detailed demographic indicators pick up: the standard of care.

Sometimes, the associations were surprising – for instance, when GPs prescribed fewer doses of certain drugs, improvement was likely to be required. This shows how a rigorous data-informed process can be used to avoid motivated thinking, and highlights the potential in inspection targeting for reducing cognitive load by automatically sifting through many different factors.

What can we learn from this? We found that it is feasible to learn from prior inspection results to target future inspections more efficiently. It is also possible to find more GP practices that are inadequate or that require improvement with the same number of inspections.

Care homes

We also looked at predicting ratings of care homes for the elderly, using a similar approach to the one we used for GP practices. Unfortunately, there is much less public data on care homes – the CQC's indicators are not public, and workforce data is optional to collect and can only be obtained on a regional not a care home level from Skills for Care[4].

Another difficulty with care homes is that reviews are overwhelmingly positive, with over 99 per cent having three stars or higher average ratings. This is perhaps because serious negative complaints with care homes tend to be reported to regulators; in contrast, with GP practices, a lot of negative reviews focused on lower-level issues.

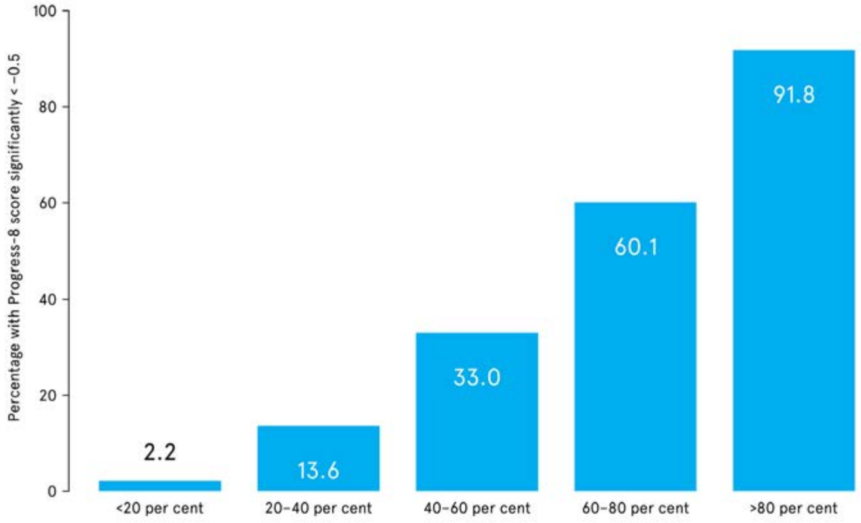
In spite of this, we can still detect about 45 per cent of the 'requires improvement' or 'inadequate' care homes if we inspect 30 per cent of the care homes, showing that we can achieve reasonable predictive power even with less available or relevant data. The equivalent detection rate for GP practices was 55 per cent.

Progress-8

Progress-8 is a headline indicator of school performance which measures the progress a pupil makes between the end of primary school and the end of Key Stage 4 (GCSEs) compared to pupils with similar primary school performance. It is used as a floor standard,[5] meaning that it identifies schools where students are doing worse than they 'should do' relative to their earlier life performance.

Progress-8 is of interest to Ofsted and serves as a useful comparison point to the inspections themselves. We found that Progress-8 is highly predictable (with an almost perfect relationship) with just one variable: the percentage of Key Stage 4 pupils eligible for pupil premium (see Figure 2). The same relationship is not true for Ofsted inspection results.

Figure 2: Percentage of schools with Progress-8 scores failing the floor standard by percentage of Key Stage 4 pupils eligible for pupil premium.



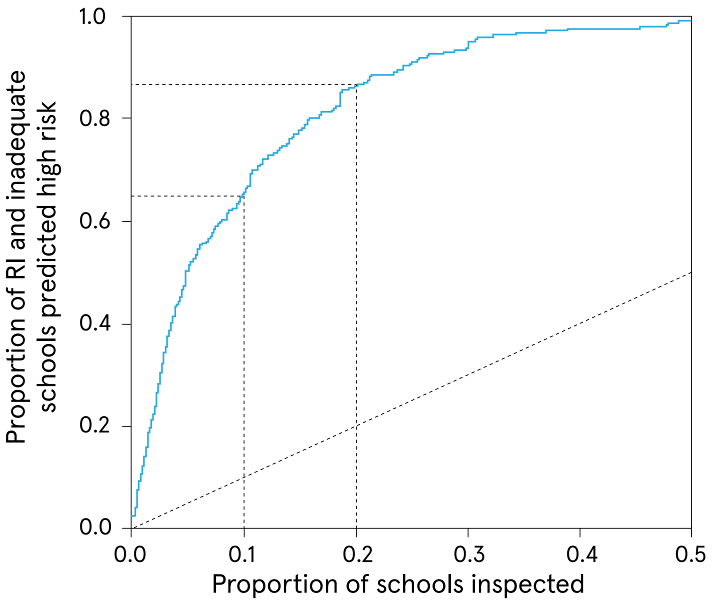
Ofsted school inspections

School inspections are a vital part of ensuring that standards are maintained across the state school system. Further, we know from previous research that interventions in schools with low ratings improve outcomes.[6] Ofsted is already a leader in setting, validating and using rating scales and in the use of statistical models to target inspections. We attempted to predict the results of Section 5 reinspections of schools previously rated 'good' using machine learning, because Ofsted has a degree of control over the timing of these reinspections.[7]

We used data which is publicly available from the year before an inspection happened, including workforce data, UK census and deprivation data from the local area, school type, financial data (sources of finance and spending), performance data (Key Stages 2 for primary schools and Key Stages 4 and 5 for secondary schools) and Ofsted Parent View answers to survey questions. We found that 65 per cent of 'requires improvement' and 'inadequate' schools were within the 10 per cent of schools identified as highest risk by our model. Increasing this to the riskiest 20 per cent, our model captured 87 per cent of these schools (see figure 3).

Note that our model, presented here, is based on pre-2015 data. It is not the model currently in use by Ofsted.

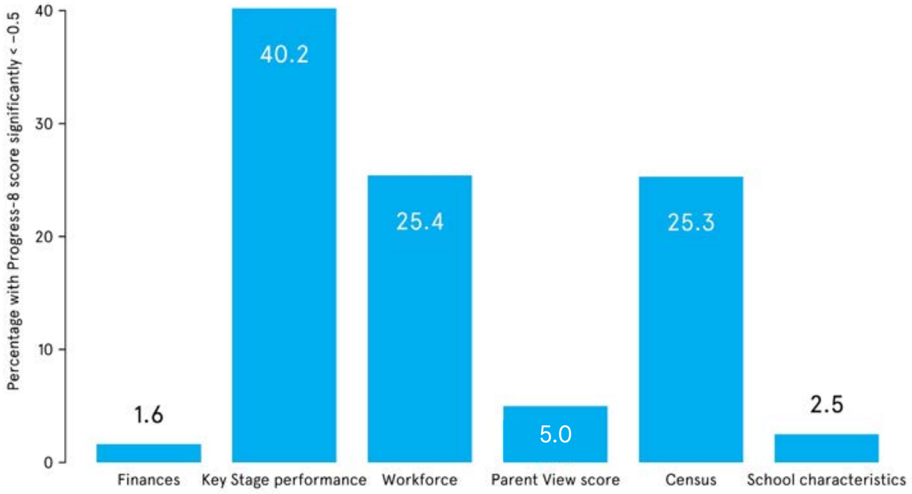
Figure 3: The gain curve for our model.



The overall differences between the two groups of schools are quite subtle – the predictive power of the model is derived from the interactions between the variables rather than the power of any one variable by itself. For example, having good reviews and low staff turnover might be much more predictive than having either indicator alone. This nuanced picture, compared to the results we saw from predicting Progress-8 scores, suggests that Ofsted inspections are picking up something deeper about schools than just educational performance. Figure 4 shows the most important categories for predicting Ofsted performance.

Some of the most important factors are around the school's finances, which are published through the school workforce census. One interesting predictor is missing data. If a school did not provide some types of information to, for example, the Department for Education, it was more likely to fail its inspections than schools that did provide this information. This requires further investigation, but it could point to a general paucity of administration at these schools.

Figure 4: Proportions of predictive power contributed to the model by the various data sources.



Decision aids

Helping children's social workers to accurately escalate cases

Social workers have some of the hardest jobs in the public sector. Individual social workers who conduct assessments could be handling as many as 50 cases at a time. They are responsible for quickly assessing whether a child is at risk of harm and in need of protection, and ultimately, in conjunction with the courts, whether a child needs to be taken into care.

They must do this with scarce resources, under fierce time pressure and often in the face of hostile opposition. Unlike in other fields working with children, such as teaching, there is no clear outcome measure (such as grades) by which social workers are assessed, and individual failures of social worker decision-making attract extreme scrutiny.

Our work on children's social care

We have been working on children's social care for several years, focusing on the 'front door' of children's social care: the decision that is made when the council is first contacted about a potential incident. Our main finding, from a behavioural science perspective, was that social workers in assessment teams make hundreds or thousands of decisions over the course of a career but receive relatively little feedback on what happens next; their cases either leave the social care system or are referred on to another team. This lack of feedback makes it difficult for social workers to learn efficiently from their previous decisions.

We have also investigated factors associated with a child's pathway through care: Did they receive multiple assessments? Was the child's case closed? Did they subsequently return to the social care system? A key finding was that, across all three local authorities we worked with, when the initial contact was made about the child on a weekend, it was less likely to progress through the social care system than when it was made on a weekday.

In evaluating Project Crewe under the Department for Education's Social Work Innovation Fund,[8] we conducted our first RCT in social care, which was published in 2017. Although this trial was small in scale, we were able to extract insights through a qualitative analysis of social worker case notes, which were manually coded based on protective and harmful factors in the text. This provided a risk score based on the number of factors present and allowed us to understand how risk changed over the course of the evaluation.

Can we predict escalation of closed cases?

In 2017, we embarked on a project to enhance our previous work in children's social care using cutting-edge data analysis techniques from the fields of machine learning and natural language processing. Building on our earlier analysis, we investigated the social worker decision-making process to close a case and recommend 'no further action'. Our core aim was to explore how far these techniques and tools could provide insights that could be practical for social workers on the ground. We therefore consulted social workers on the data science findings, conducting six semi-structured interviews with current social workers to understand their interpretation of the data.

The core predictive problem was this: given the text of the initial referral and assessment, and structured data relating to the case, could we predict whether the case would be re-referred and escalated if it were closed?

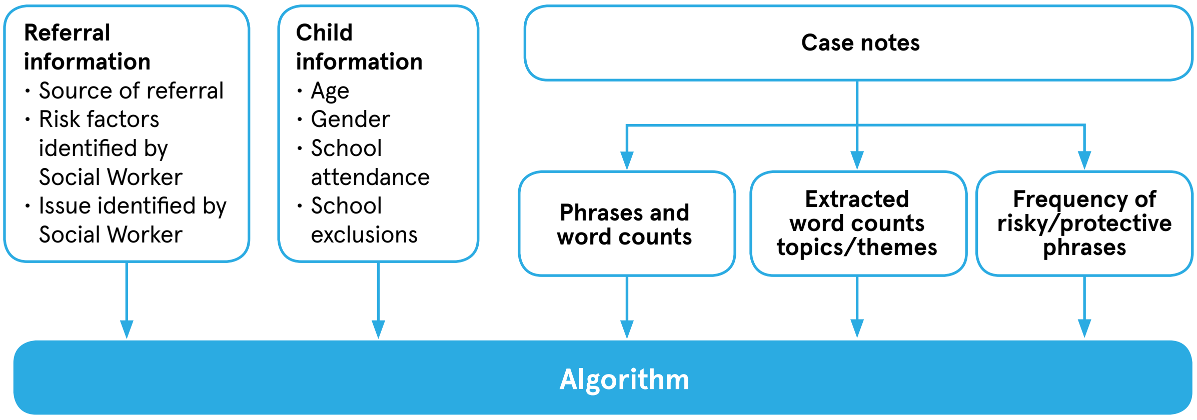
Of the 11,000 children's cases that were referred into the system in the two years we analysed, 5,117 were immediately closed, with no further action. We sought to determine which of these cases were likely to come back into the social care system later (1,693 children) and, of those, which would return with serious case activity (583), such as a Child Protection Enquiry (child identified as having a risk to their life, or being at risk of immediate harm).

What we did

While some information on the initial referral and the child was available for analysis, the most substantial data was the social workers' case notes, which are free-form notes made by social workers under headings such as 'Assessment Analysis', to document their findings and the details of each case. Case notes are largely unstructured – there are many idiosyncrasies in how information is recorded and how issues are described. This makes them challenging to analyse using traditional text analysis techniques, such as counting how often certain words or phrases appear in the text.

We therefore analysed the text using topic models (see box below) to extract themes from the text. These topics, together with more traditional structured data relevant to the case, were fed into a machine learning algorithm (a gradient boosting machine, see box on page 11). This machine learning algorithm was used to identify cases which had a high risk of returning into the social care system after being closed. The various inputs are illustrated in figure 5.

Figure 5: The inputs of the machine learning algorithm used to detect escalated closed cases.



Topic modelling

Topic modelling is the automated process of finding groups of words that commonly occur together (in topics). They typically generate easier to comprehend findings, for policymakers and other professionals, than other text analysis methods, because topics tend to be more coherent than lists of words or phrases considered independently.

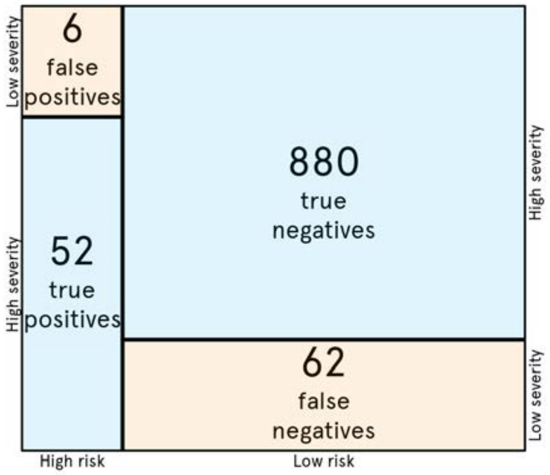
We rely on a technique called ‘structural topic modelling’[9], which improves the coherence of these groups based on the fact that topic prevalence varies across speakers and what is being spoken about. For instance, the original paper found that the propensity to discuss the Iraq War varied by the political ideology of blogs. In our social work context, topics were allowed to vary by child characteristics, social worker experience, identified risks and who referred the case to social services.

To help us interpret this data and situate findings from the algorithm in real-world experiences, we sought feedback and conducted a series of semi-structured interviews with four social workers and two team managers from within the local authority. This was important for this project because social workers would need to understand the reasons behind the algorithm’s suggestions for any particular case in order to could combine these insights with their own expertise.

What we found

Using this combination of structured and unstructured data, our algorithm was able to identify a small (6 per cent) set of cases that were closed as ‘high risk’. This high-risk set contained nearly half of the cases that would later return and escalate, with very few (0.6 per cent) false positives (here, a false positive is a high-risk case which does not in fact return and escalate). Figure 6 illustrates the model’s expected performance on 1,000 closed cases. A case is ‘high severity’ if it is re-referred and escalates; otherwise it is ‘low severity’.

Figure 6: Expected distribution of true and false positives and negatives for 1,000 previously unclassified cases.



‘Risk’ refers to the model’s prediction and ‘severity’ to the actual outcome. For example, a ‘false negative’ is a case that escalated but that the model categorised as ‘low risk’.

There is always a trade-off between the false-positive rate and the false-negative rate; however in this case it is difficult to decrease the false negatives without there being a large increase in false positives, which would create a lot of unnecessary work. There are a lot of reasons why this group of outcomes is hard to predict. Some families, fearful of social workers, hide issues or pretend to be more compliant than they are to avoid being associated with the stigma of children's services' involvement. In addition, interviews with the social workers revealed that finding actual evidence of the problem can also be challenging in the short time window they are given.

The topic extracted from the text that was most useful in identifying cases that were likely to return is exemplified by the following anonymised extract from case notes:

> "I, [Name of social worker], am recommending no further action from CSC [children's social care]. I feel the children have not shown any worries or changes of behaviour from the verbal argument. I feel even though CSC have had involvement with the family over the years, [Name] is capable of being a person of safety for [child, half-sibling and sibling]."

Looking at multiple examples, this topic seems to correspond to case notes where the social worker feels that it is necessary to spend time justifying why they are closing a case, often due to insufficient evidence or consent from the family. This reflects some of the challenges social workers face, where there may not be enough evidence to substantiate issues that the social worker suspects may be present, or the evidence is not clear cut, or the families may be hiding the extent of the issues.

> *'The gut feeling of me and social workers are at times, this will come back because we haven't effectively changed the dynamics of the situation ... But we know at the moment that it isn't enough, for example, to go to court.'*
>
> *(An interviewed social worker)*

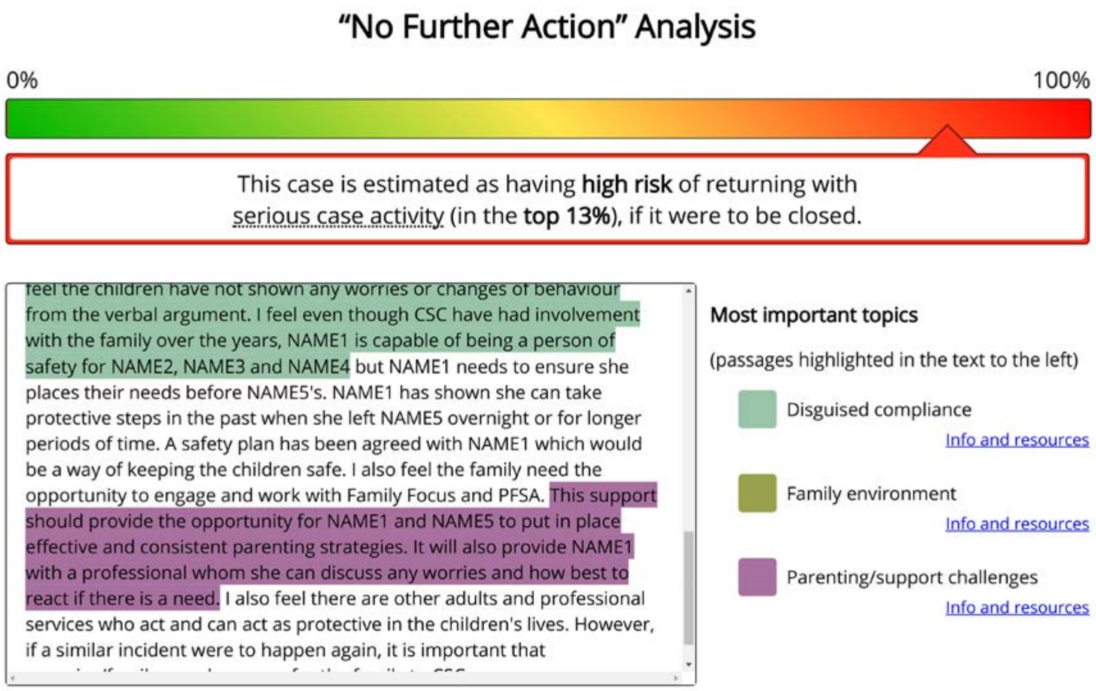
In other cases, this language is used by social workers when they believe that a family might benefit from their help but the family refuses to cooperate or accept help from social care. Social workers interviewed were clear that they had to make a decision based on the current available evidence but would appreciate a tool which helped them overcome their internal biases or see whether cases had the hallmarks of being at a high risk of re-referral.

What's next? A digital decision aid

We are currently building a digital tool that will allow social workers to see the algorithm's estimated risk for a particular case. This will allow us to make our findings accessible and useful to social workers in their daily practice. Following feedback from social workers and managers, the best use for the algorithm may actually be in providing an evidence base to justify spending more time on potentially risky cases, when the decision is not clear cut but the social worker typically wouldn't have sufficient grounds for keeping a case open.

The first version of this tool, which we hope will be tested with the participating local authority in early 2018, will allow a case worker or manager to copy text from a person's case notes and paste it into the tool, which will then analyse the text and provide a risk estimate. As well as the risk rating, which will be red, amber or green, the tool will identify which sentence fragments and topics in the text were most indicative of the risk. The basic design for the tool can be seen in Figure 7.

Figure 7: Design of the prototype children's social care risk assessment tool.



"No Further Action" Analysis

This case is estimated as having **high risk** of returning with serious case activity (in the **top 13%**), if it were to be closed.

feel the children have not shown any worries or changes of behaviour from the verbal argument. I feel even though CSC have had involvement with the family over the years, NAME1 is capable of being a person of safety for NAME2, NAME3 and NAME4 but NAME1 needs to ensure she places their needs before NAME5's. NAME1 has shown she can take protective steps in the past when she left NAME5 overnight or for longer periods of time. A safety plan has been agreed with NAME1 which would be a way of keeping the children safe. I also feel the family need the opportunity to engage and work with Family Focus and PFSA. This support should provide the opportunity for NAME1 and NAME5 to put in place effective and consistent parenting strategies. It will also provide NAME1 with a professional whom she can discuss any worries and how best to react if there is a need. I also feel there are other adults and professional services who act and can act as protective in the children's lives. However, if a similar incident were to happen again, it is important that

Most important topics

(passages highlighted in the text to the left)

🟩 Disguised compliance
Info and resources

🟩 Family environment
Info and resources

🟪 Parenting/support challenges
Info and resources

The design of this tool is informed by our previous interviews with social workers, and the tool will be subject to additional stages of feedback and adaptation. It is key that this tool offers practical insight and is extremely easy for social workers to incorporate into their day-to-day work, so we are taking particular care to include social workers and their managers in the design process.

Designing targeted interventions

In the spirit of 'what works', we can use data science to understand what makes someone likely to perform an undesirable behaviour or not. This allows us to design interventions which are targeted at those individuals who are most at risk, as well as maximising the efficiency with which we deliver our interventions. Conversely, we can avoid targeting people who are unlikely to do the behaviour we want to stop.

Reducing traffic collisions in East Sussex

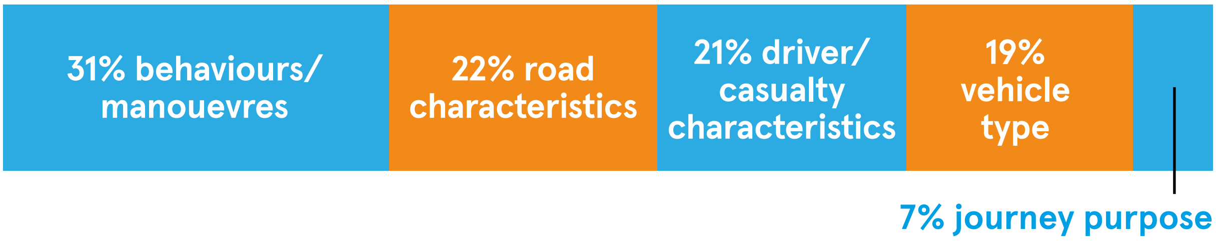
While the UK's roads are very safe by international standards, there was a 4 per cent recorded increase in fatalities from 2015 to 2016[10] and transport accidents remain a leading cause of death among individuals aged 20–34.

There is a wealth of data available to design interventions with: in addition to the drivers' and any casualties' details, an attending officer at a collision forms and records an opinion of what the 'contributory factors' were that led to the collision. We can use these to form a picture of which behaviours cause the most serious of crashes, where someone is killed or seriously injured (KSI). Common contributory factors include 'careless, reckless or in a hurry', 'impaired by alcohol' and 'exceeding the speed limit'.

We used all of these data sources as the basis for a predictive model for the severity of a collision. The aim was not to produce predictions themselves – there is no use in predicting the severity of a collision once it has happened. Instead, this model is able to tell us the *importance* of each of its inputs into predicting collision severity, so that we can target those behaviours which have a real impact.

In particular, we can now quantify the extent to which collision severity is influenced by the behaviour that caused the collision. This is shown in Figure 8.

Figure 8: Relative predictive importance of categories of data.



The bars add up to 100 per cent and each one indicates the total amount of importance (similar to correlation, but more general) for all variables in that category. The colours are for cosmetic purposes only.

Driver behaviour is the most predictive element in collision severity, more so than the characteristics of the road (including the speed limit and prevailing weather conditions), the age and gender of the driver, and how far they were from home.

Another important insight is that the purpose of the journey (commuting, work or personal) does not strongly predict collision severity.[11] This means that targeting only occupational drivers is unlikely to be as effective as having all journey types in scope.

Armed with the knowledge of what the most dangerous behaviours are, we can target people who are doing them the most often. In East Sussex, these fall into three categories.

The first category is *young drivers*. Drivers under the age of 25 are known to be a high-risk group nationally, and in East Sussex we found that young male drivers were significantly overrepresented in collision data compared with the number of driving licences held: 16.8 per cent of KSI collisions were caused by them but they held only 5.6 per cent of licences. Young drivers disproportionately caused severe crashes when they were within three miles of home, and being under the influence of alcohol was much more common for this age group than for others.

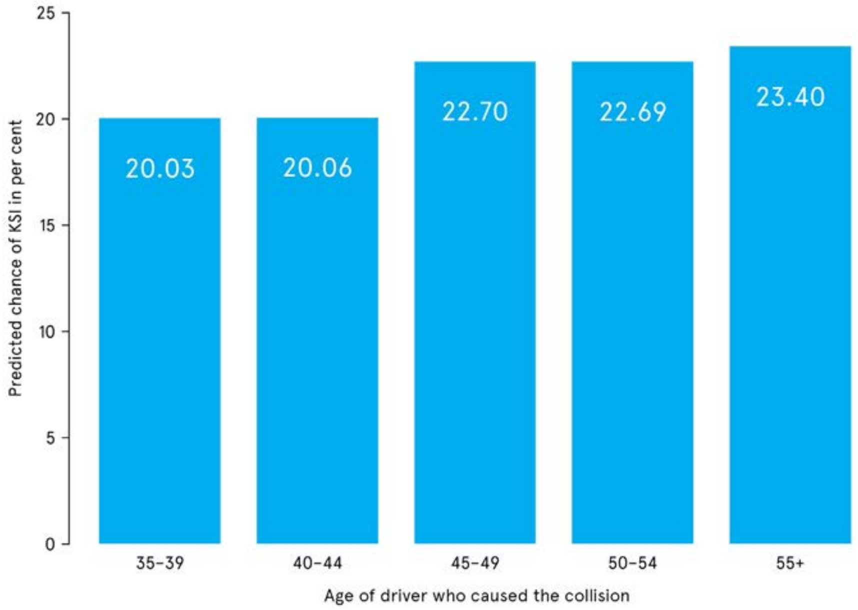
If a young driver causes a collision, the speed limit and the severity of the collision are almost statistically independent. This is not true for any other category of driver, and strongly suggests that altering speed limits is not an effective tactic for reducing collision severity in young drivers.

The second category is *motorcyclists*. Nationally, a mile ridden on a motorcycle is twelve times as likely to result in death or serious injury than a mile driven in a car.[12] In East Sussex, that figure is *26 times*.[13] When a collision occurs, the expected severity is unsurprisingly high, but in high-speed-limit zones this confluence of risks leads to a very rare phenomenon: if a motorcyclist in East Sussex is involved in a collision in a 60 mph speed limit zone or higher, it is *more likely than not* there will be a serious injury or fatality.

We know from interviews with road safety professionals that motorcyclists will often organise themselves to avoid speed enforcement measures, and that exceeding the speed limit is a common contributory factor to collisions for motorcyclists. This does not mean that reducing speed limits will not be effective, as they may act as an 'anchor'[14] or a reference point that affects our perception of high and low speeds. However, we must recognise that motorcyclists are engaging in more wilfully dangerous behaviours as opposed to lapses in concentration and that this will require different techniques to combat.

The third target category is more nuanced: *drivers aged 45–65 interacting with vulnerable road users*. If a collision involves, but is not caused by, a vulnerable road user, then there is approximately a one-in-five chance of a death or serious injury if the other party is aged 35–44 (see Figure 9). Increase that age to 45 and over and there is a sudden increase in this chance to almost one in four. These collisions are often close to the home of the driver and it is very common for them to be caused by inattention or being 'careless, reckless or in a hurry'.

Figure 9: Sudden increase in serious collisions/fatality with vulnerable road users for drivers aged 45+



Chance of serious or fatal collision (where person is killed or seriously injured: KSI) where that collision is caused by a car driver and involves another, vulnerable road user. The age is of the driver.

The behaviours exhibited by these three groups are very different in nature and we cannot expect to tackle them effectively using the same techniques. As part of our commitment to understanding 'what works', we will be trialling variations of two letters for car drivers at risk of dangerous driving as well as testing the effects of targeted communications on the anniversaries of driving offences or minor collisions for all three groups.

The value of our data science approach is that we have avoided designing interventions that are doomed to fail because the target audience is not doing the behaviour the interventions are designed to inhibit, or because the interventions target a group of people who are not behaving dangerously.

Making RCTs better:
collaborating with KCL×BIT

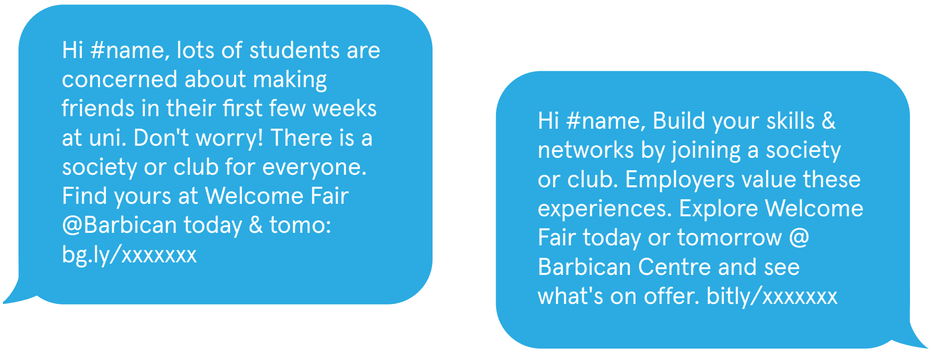
We have run over ten RCTs with KCL hoping to learn what works in increasing participation in various university activities and in the overall sense of belongingness, using institutional datasets to measure efficacy. KCL×BIT is a two-year collaborative project between the Widening Participation Department, Policy Institute at KCL and BIT.[15] Recognising that getting into university is not the final step for students, particularly those from underrepresented groups, this project focused on student success as well as student access.

Trial: getting students to sign up to societies

In 2016, we ran a trial which aimed to increase the likelihood that young people would attend the Welcome Fair run at the start of term by the KCL Students' Union. At these fairs, students can learn about and then sign up to clubs and societies, and also meet fellow students.

We ran an RCT to test the effect of different messages on whether students attended the fair and whether they joined societies. One third of eligible students received no texts prior to the fair. A third of students were in the 'belonging' group and received three texts focused on reducing perceived barriers associated with belonging, while the final third (in the 'employability' group) received three messages emphasising the employment benefits of societies. Figure 10 shows these two texts.

Figure 10: Texts sent for the KCL Welcome Fair trial (belonging on left and employability on right).

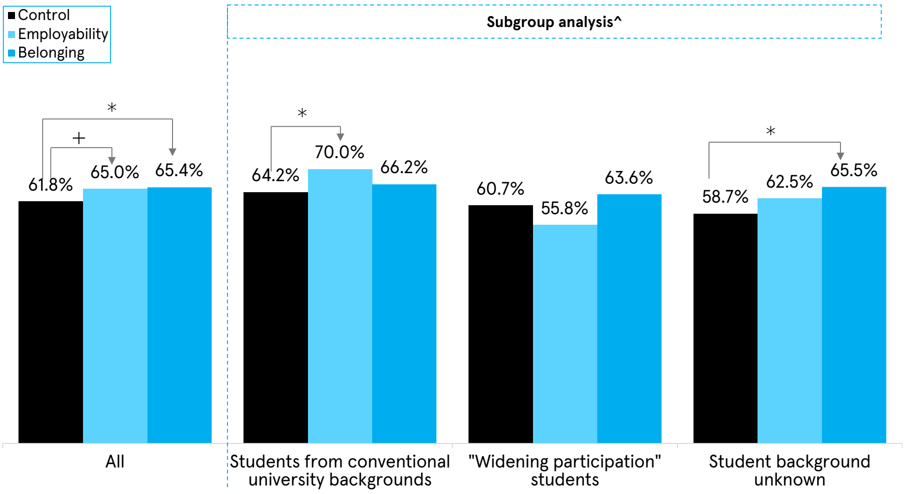


Hi #name, lots of students are concerned about making friends in their first few weeks at uni. Don't worry! There is a society or club for everyone. Find yours at Welcome Fair @Barbican today & tomo: bg.ly/xxxxxxx

Hi #name, Build your skills & networks by joining a society or club. Employers value these experiences. Explore Welcome Fair today or tomorrow @ Barbican Centre and see what's on offer. bitly/xxxxxxx

We found that the messages made students significantly more likely to attend the Welcome Fair, although which message was most effective appeared to depend on student characteristics. Across all students, the messages centred on belonging had the greatest impact on attendance, resulting in a 6 per cent increase. Employability messages, by contrast, increased attendance by just over 5 per cent, and the increase was not significant at conventional levels (see Figure 11). The employability messages increased sign-ups to societies but the belonging messages did not.

Figure 11: Sub-group analysis of the effect of behavioural messages on attendance at the Welcome Fair in 2016 by Widening Participation Status.



**: p < 0.05 +: p < 0.1 ^: Students were grouped by their Acorn categorisation, a measure of social background. Those from conventional backgrounds were in Acorn categories 1-3 (which reflects higher socioeconomic backgrounds) whilst "widening participation" students were in Acorn categories 4-5. Some students did not have an Acorn categorisation, likely because they were international students, and have been grouped under "unknown".*

Follow-up trial: using machine learning to tailor messages

Because the results of the Welcome Fair trial proved promising – and because the effects of the two interventions seemed to vary for particular sub-groups – we decided to run a version of the trial for the 2017 fair using a causal machine learning approach. We used the results of the 2016 trial to predict which condition new students should be assigned to based on their characteristics.

There were two conditions in this trial. The first group was randomly allocated to receive either the belonging or the employability message. In the second group, students received the message that the algorithm predicted would be more effective at prompting them to turn up. We removed students who did not want to receive messages or who had numbers that bounced.

Random assignment (*N* = 2,085)	Algorithmic assignment (*N* = 2,085)
50 per cent received belonging messages; 50 per cent received employability messages.	Allocation to belonging or employability conditions was based on the algorithm.

This algorithm was trained by predicting which students would be best sent the employability message, rather than the more effective (on average) belonging message as the default.

In the random-assignment condition, 60.1 per cent of students attended the Welcome Fair – 59.1 per cent for the employability condition and 61 per cent for the belonging condition. By contrast, 60.5 per cent of students attended the Welcome Fair in the algorithmic-assignment condition, which is not significantly different from the random-assignment condition.

It might be that, for some of the participants who were assigned (by default) into the 'belonging' group by the algorithm, the control condition would actually have performed better. We are iterating our choice of algorithms in this space, but an important takeaway is that the exact design of targeting is important, so that we cause an overall improvement. As this is to our knowledge the first RCT to test this algorithm, we have also contributed to understanding how machine learning algorithms for sub-group analysis might be tested.

What makes a good data science project?

We have found that the following four ingredients are key in a data science project:

1. a predictive problem or the need to understand unstructured data;
2. high-quality, large-scale, appropriate data;
3. the ability, and departmental buy-in, to implement findings in practice;
4. ethical and legal clearance.

What makes a good predictive problem? There needs to be a measured behaviour or rating that we can predict (ideally at an individual level – unfortunately a drawback of official statistics is they are often at an aggregate level), and it must be directly relevant to our partner's priorities. For instance, with school inspections, the predictive problem was predicting which schools would receive a requires improvement or inadequate inspection rating, which is clear and measurable from historical data.

Data quality is the next issue, as there needs to be a good fit between the problem and the data we can use to solve it. While each project is different, two common themes are:

- Data that is at the same level as the behaviour we are interested in is more useful than data that covers a wider area or group.
- If text or image data is required, it is worth being particularly careful to ensure that the data is high-quality, reasonably complete and readily extractable.

Buy-in is crucial so that decision-making does not eventually fall back into traditional ways of thinking.

It is, however, also important to think early on about the ethics of the project:

1. Data science has particularly strong considerations around privacy.
2. Legislation requires that any decision taken by automated means may be requested to be reconsidered just on that basis (with some exemptions).
3. Data which is being used to train algorithms may itself be subject to bias (for example, if only African-Americans have been historically disproportionately charged with certain offences due to police bias then the algorithm will exacerbate the bias by predicting that African-Americans are at a higher risk of offending).[16]
4. Algorithms which use protected characteristics, such as race or gender, must be very carefully designed so they are not less accurate in relation to vulnerable groups even if the data is not biased.

Future directions for data science in policy

Forecasting

Traditionally, forecasting and its companion task, resource allocation, have been the domains of meteorologists and military personnel. However, due to the availability of big data, they could be usefully applied across the NHS and other large government systems.

We are currently working with Transforming Systems (a health data analytics company specialising in whole system data) and Medway Hospital on a project to predict waiting times for accident and emergency patients using machine learning. Hospitals need an accurate way of knowing waiting times both for short-term planning and medium-term rostering, which both have real impacts on severe and acute health outcomes. In this project, we are working with Transforming Systems' analysts using a methodology similar to Google's methods for predicting advertising success based on thousands of search queries over time. We expect to have results within six months.

Cost-benefit analysis and what works for whom

We view it as important to go beyond 'what works' to look at 'what works for whom'. We will be working with government departments and agencies to examine how the different components of national programmes, particularly those focused on vulnerable groups, worked for different groups of people. This will allow interventions to be better targeted, and will allow governments to get more information out of expensive evidence-gathering exercises.

Cost-benefit analysis is routine in government. However, the availability of new tools and increased computing power have meant that we can now answer more interesting questions than simply 'on average, was the benefit provided more than the cost?'

We are interested in examining when we should stop collecting new data on or doing new evaluations of interventions and instead move towards scaling them up. This depends in part on whether we have information on enough parts of the country or sub-groups of people.

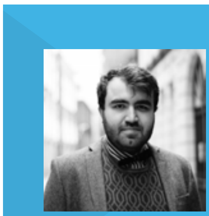
About the Authors

Michael Sanders
Chief Scientist and Head of Research and Evaluation

Dr. Michael Sanders is Chief Scientist and Head of Research and Evaluation at the Behavioural Insights Team (BIT). Alongside his work at BIT, Michael works as an associate fellow at the Blavatnik School of Government and as a senior research associate at University College London, where he co-directs the Behavioural Science and Policy PhD programme. Michael completed his PhD in Economics at the University of Bristol, and postdoctoral studies at Harvard's John F. Kennedy School of Government.

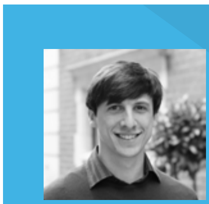
James Lawrence
Head of data science

James Lawrence is BIT's Head of Data Science. Before coming to BIT and leading the work described in this report, James worked in research and development for a major UK insurer. James has a mathematics and statistics background, with an MMath from the University of Cambridge.



Daniel Gibbons
Research Advisor

Daniel Gibbons is a Research Advisor for BIT who works on data science and evaluation projects, with particular interests in text analysis and predictive modelling. They have an MPhil in Economic Research from the University of Cambridge and an MSc. in Mathematics and Statistics from the University of Queensland.



Paul Calcraft
Technical Lead and Data Scientist

Dr. Paul Calcraft is the Technical Lead for BI Ventures and a Data Scientist in the Research and Evaluation Team. Before joining BIT, Paul was a programmer in the web development industry, latterly moving into machine learning and research. He holds a PhD in Computer Science from the University of Sussex.

Endnotes

1. This is based on our own modelling, using the same techniques but with a limited set of data. The CQC does use more data than this in its own targeting and does not solely target based on a model.
2. http://dsapp.uchicago.edu/resources/datamaturity
3. Note that the decision tree is a dummy tree and is **not** reflective of the actual statistical model.
4. Skills for Care (http://www.skillsforcare.org.uk/Home.aspx) is the strategic body for workforce development and training for adult social care in England and is the home of the National Skills Academy for Social Care.
5. This is true if the score is less than −0.5 and the confidence interval is entirely below zero.
6. Allen, R., & Burgess, S. (2012). *How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England. Working Paper No. 12/287*. Centre for Market and Public Organisation. Available at http://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/wp287.pdf
7. We should emphasise that Ofsted does not currently take this approach to targeting its inspections and that no Ofsted inspections have been instigated by this work at time of publication.
8. Department for Education. (2017). *Children in need: Project Crewe*. Available at https://www.gov.uk/government/publications/children-in-need-project-crewe
9. Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(151), 988–1003.
10. Department for Transport. (2017). *Reported road casualties in Great Britain: 2016 annual report*. Statistical Release. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/648081/rrcgb2016-01.pdf
11. This does not necessarily mean that, for example, occupational journeys result in fewer collisions than others. We are making statements about severity, namely that these collisions are neither unexpectedly severe nor unexpectedly slight.
12. Department for Transport. (2016). *Road traffic estimates in Great Britain: 2016*. Available at https://www.gov.uk/government/statistics/road-traffic-estimates-in-great-britain-2016
13. Department for Transport. (2016). *Traffic counts: East Sussex traffic profile for 2000 to 2016*. Available at https://www.dft.gov.uk/traffic-counts/area.php?region=South+East&la=East+Sussex
14. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New Series)*, 185(4157), 1124–1131.
15. For more information on this wider collaboration, see the project blog: https://blogs.kcl.ac.uk/behaviouralinsights
16. For a series of examples, some of which do not relate to protected characteristics (see point 4 above), see Sample, I. (2017). AI watchdog needed to regulate automated decision-making, say experts. *The Guardian*, 27 January. Available at https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions

THE
BEHAVIOURAL
INSIGHTS TEAM

Data Science Team:

Michael Sanders, James Lawrence, Dan Gibbons, Paul Calcraft.

Contributors:

Doireann O'Brien, Clare Delargy, Edward Flahavan, Jessica Heal, Min-Taec Kim, Lucy Makinson, David Nolan, Sean Sheehan, Handan Wieshmann.

Behavioural Insights Team
4 Matthew Parker Street
London
SW1H 9NP