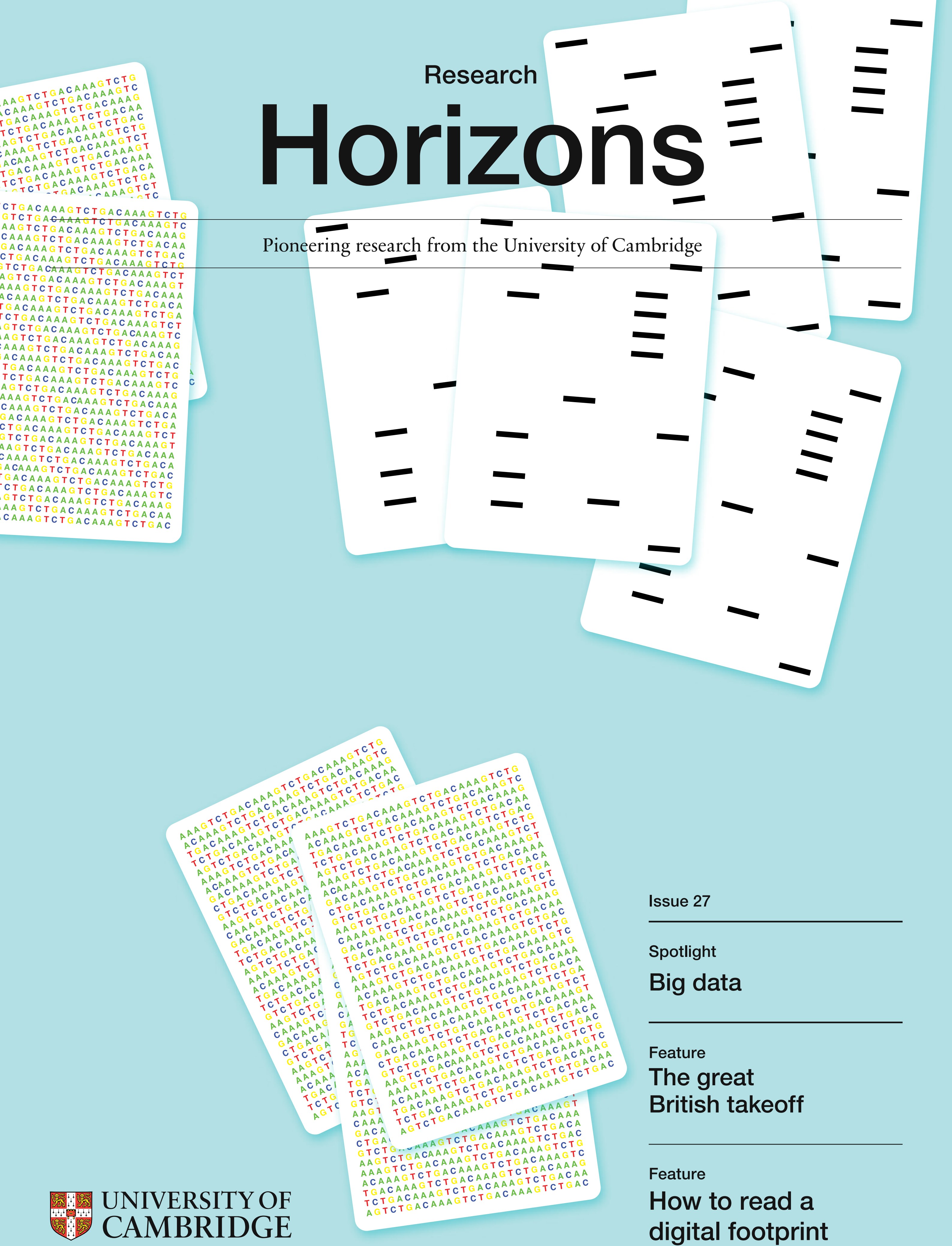
Research Horizons

Pioneering research from the University of Cambridge



Issue 27

Spotlight

Big data

Feature

The great British takeoff

Feature

How to read a digital footprint

UNIVERSITY OF CAMBRIDGE
www.cam.ac.uk/research

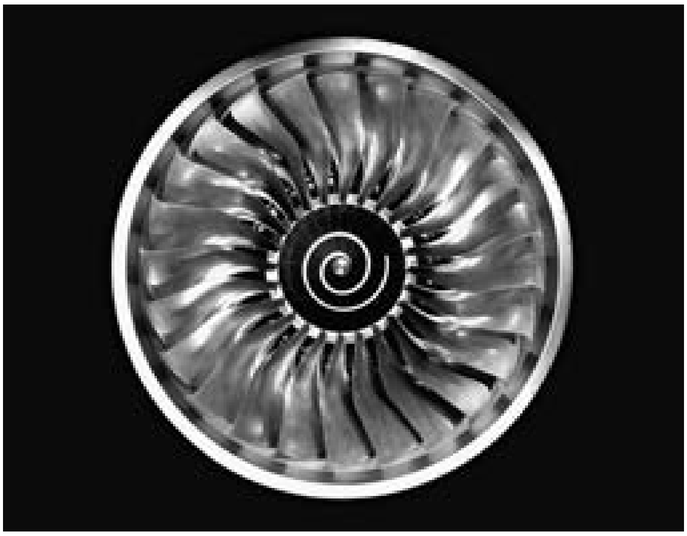
Contents

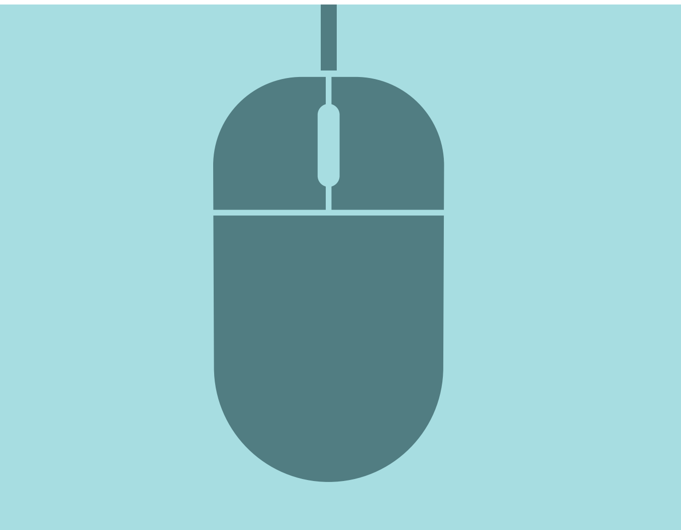
A News

B Features

C Things

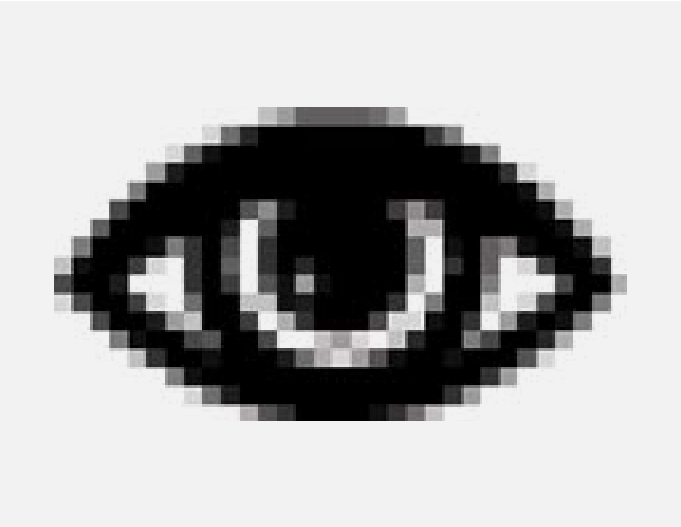
D Spotlight: Big data





D

E Inside out



Welcome

At the heart of almost all research lies data and its interpretation. But today's datasets are the largest, most diverse and fastest accumulating ever experienced – so much so, they have acquired a moniker all of their own, 'Big Data', the focus of this issue.

The Square Kilometre Array of telescopes, for instance, will produce more data than the entire traffic on the global internet at any given moment.

Global comparisons of DNA databases, meanwhile, are helping researchers find patients with some of the rarest of diseases, so clinicians don't have to start from scratch each time they encounter a new case.

Some data is of a type unthinkable a few years ago. Social media – Twitter and Facebook, for example – is providing information that could revolutionise psychological profiling, employment and commerce.

Getting the most out of big data requires new methods to handle large volumes of information and the clever use of statistical algorithms to distil meaningful knowledge out of disorder. In 2013, we launched a University-wide initiative, Cambridge Big Data, to help researchers respond to challenges like these. Cambridge is also one of five universities that will carry out research in organising, storing and interrogating big data as part of the Alan Turing Institute.

Of course, not all data is 'big', but can still be incredibly difficult to gather – such as understanding the impact, decades later, of how much a child plays. Quantifiable evidence in this area is needed for educational practice and urban development, as described in this issue, and a new research centre in the Faculty of Education will help to provide it.

We also cover research on the most extreme environment known to engineering – the jet engine – as well as air quality monitoring, how sheep are helping us to understand a devastating infant brain disease, a new theory of the industrial revolution and the fascinating story of a 'haunted' medieval book.

Professor Lynn Gladden
Pro-Vice-Chancellor for Research

Editor
Dr Louise Walsh

Design
The District

Printers
Micropress

Contributors
Craig Brierley, Sarah Collins, Jennifer Hayward, Tom Kirk, Fred Lewsey, Stuart Roberts, Louise Walsh

T +44 (0)1223 765 443
E research.horizons@admin.cam.ac.uk
W cam.ac.uk/research

News

Words of Waterloo

Documents forming the “first draft of history” in the aftermath of the Battle of Waterloo go on display.

“The field of Battle exhibits this morning a most shocking spectacle too dreadful to describe….” This letter written from the body-strewn battlefield at Waterloo, together with an invasion map of the UK and a book from Napoleon’s personal library in exile, is among the exhibits that have gone on display in Cambridge University Library during one of the first major Waterloo exhibitions of the bicentenary commemorations.

Looking at how Waterloo was written about in the immediate aftermath of the battle fought on 18 June 1815, the exhibition draws on the rich and varied collections at the Library and includes political propaganda, broadsheets, military drillbooks, coloured engravings and early historical accounts of the bloodshed.

“The exhibition really shows us the first draft of history as it was being written in the days, months and years after the battle,” explains historian Dr Mark Nicholls, who co-curated the exhibition.

The exhibition also features artefacts and mementoes from the battlefield itself, including a musket ball and a charred fragment of Hougoumont, the farmhouse which occupied a vital position in the Duke of Wellington’s line. The relics were collected by a teenage girl visiting the field 10 years after the battle.

Credit: University Library

Image
Boy’s Book of British Battles

“Waterloo is the most famous battle in modern European history, and from the very first moment soldiers and civilians alike wanted to put their experiences and emotions into words,” adds co-curator John Wells. “We examine how the battle’s impact was expressed through the written word, and how the documentary records of the time continue to have resonance for us today. Waterloo is still news, 200 years later.”

‘A Damned Serious Business: Waterloo 1815, the Battle and its Books’ runs until 16 September 2015

‘Big data’ social

A new Centre is teaching undergraduate social scientists the quantitative skills they will need to tackle ‘big data’.

The UK lags behind other countries in preparing social scientists for the world of ‘big data’ says Dr Brendan Burchell, Director of a Centre set up to teach the advanced quantitative skills they will need to work with large datasets.

The Cambridge Undergraduate Quantitative Methods Centre (CUQM), rooted in the Department of Sociology, aims to ensure that at least 25% of social scientist graduates leaving Cambridge will have some statistical expertise.

“The UK is already way ahead of many other countries in the availability of large datasets that can be used to inform both policy and social science research,” he adds. “Over the next few decades – the career span of current undergraduates – we are likely to see huge advances in the use of quantitative data including ‘messier’ datasets that can only be analysed with recent advances in big data techniques.

“These skills will become increasingly vital for careers in social science research, but they will also make students much more employable in most other sectors as well.”

CUQM is extending the exposure to statistics in the social science undergraduate courses at Cambridge, as well as providing vacation courses and work placements. It is part of a wider initiative to train social scientists in research methods. Cambridge’s Social Science Research Methods Centre, for instance, complements the work of CUQM by teaching quantitative methods to graduate students.

www.cuqm.cshss.cam.ac.uk

News in brief

More information at www.cam.ac.uk/research

01.04.15

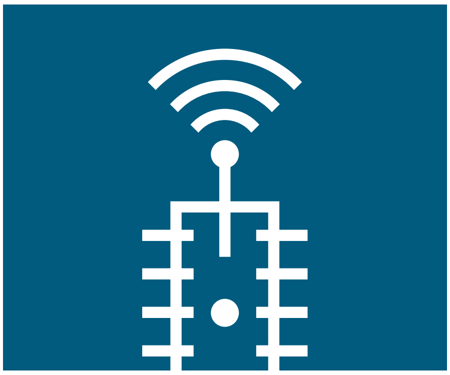
Archaeologists unearth one of Britain’s largest medieval hospital cemeteries, containing over 1,000 human remains.

25.03.15

The Chemistry of Health programme is awarded £17 million to research Alzheimer’s and Parkinson’s diseases.

Antennas and chips

Discovery of the 'last frontier' of semiconductor design could be a massive leap forward for wireless communications.



Researchers from the Department of Engineering have unravelled one of the mysteries of electromagnetism, which could enable the design of antennas small enough to be integrated into an electronic chip.

The purpose of any antenna, whether in a communications tower or a mobile phone, is to launch energy into free space in the form of electromagnetic or radio waves, and to collect energy from free space to feed into the device.

One of the biggest problems, however, is that antennas are still quite big and are incompatible with electronic circuits – which are ultra-small and getting smaller all the time.

"An aerial's size is determined by the wavelength associated with the transmission frequency of the application, and in most cases it's a matter of finding a compromise between aerial size and the characteristics required for that application," explains Professor Gehan Amaratunga, who led the recently published research.

Working with researchers from the National Physical Laboratory and Cambridge-based dielectric antenna company Antenova Ltd, the team used thin films of piezoelectric materials, a type of insulator which is deformed or vibrated when voltage is applied. At a certain frequency, these materials become not only efficient resonators, but efficient radiators as well, meaning that they can be used as aerials.

Future applications of the discovery include implementation of the 'Internet of Things', where almost everything in our homes and offices, from toasters to thermostats, is connected to the internet.

'Serial killers' caught on film

A dramatic video has captured the behaviour of white blood cells as they destroy cancer cells.

Inside all of us lurks an army of serial killers whose primary function is to kill again and again. Cytotoxic T cells, a type of white blood cell, 'hunt down' and destroy cancer cells and virally infected cells before moving on to their next target.

Now, the moment of killing has been captured on film in 3D by a collaboration of researchers from the UK and the USA. The research was led by Professor Gillian Griffiths at the Cambridge Institute for Medical Research with funding from the Wellcome Trust.

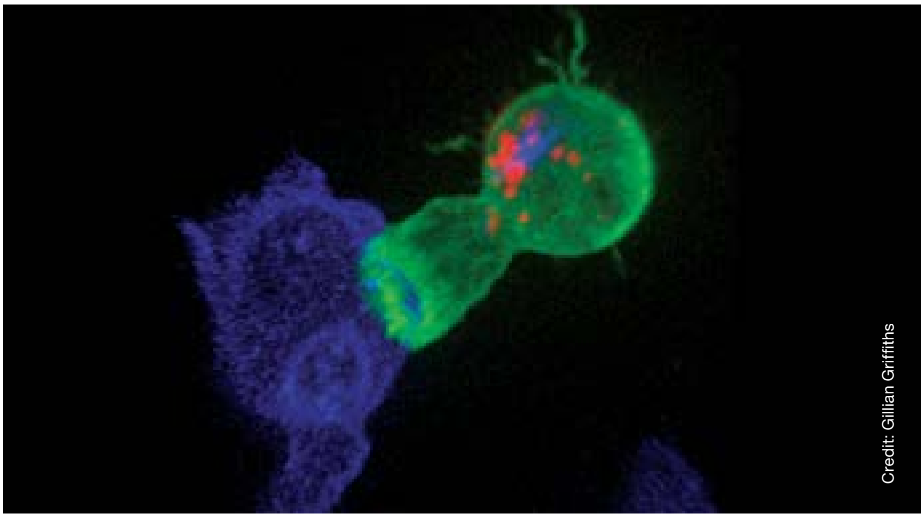
The researchers used high-resolution time-lapse multi-colour imaging techniques that capture slices through an object and then 'stitch' them together. As a result, they have managed to elucidate the order of the events that lead to delivery of the lethal 'hit' from these serial killers.

There are billions of T cells within our blood, each of which is engaged in the ferocious and unrelenting battle to keep us healthy. The T cell extends membrane protrusions that explore the surface of the cell, checking for tell-tale signs that it is an uninvited guest. It then injects poisonous proteins known as cytotoxins between the T cell and the cancer cell, before puncturing the surface and delivering its deadly cargo.

"Once the cytotoxins are injected into the cancer cell, its fate is sealed and we can watch as it withers and dies," explains Griffiths. "The T cell then moves on, hungry to find another victim."

Image
A T cell (green) delivers the lethal hit

Film available
bit.ly/1GMgBJU



Credit: Gillian Griffiths

24.03.15

A National Research Facility for Infrastructure Sensing will recieve £18 million in funding to support the application of sensor technologies.

12.03.15

Research on new family forms, such as same-sex parents, looks at what these forms mean for the parents and children involved.

16.02.15

Cambridge is one of three flagship Drug Discovery Institutes that will fast-track development of new treatments for dementia.

A real piece of work

Image
Flames and smoke billow from the open coke hearths of the Bedlam furnaces in Coalbrookdale, Shropshire, as imagined by Philippe Jacques de Loutherbourg in 1801

Credit: Science Museum

In 2003, researchers embarked on a project to piece together a picture of changes in British working life over the course of 600 years. The emerging results seem to demand a rewrite of the most important chapter in our social and economic history.

There comes a point when talking with Dr Leigh Shaw-Taylor at which it seems necessary to go over the facts again, if only to establish that he really does mean what he appears to have just said.

While many historians will spend their careers chipping away at the past with gentle care, 12 years into his research project, The Occupational Structure of Britain, 1379–1911, Shaw-Taylor seems to be calling for a wholesale rewrite. If his emerging results are correct, then they have the potential to transform not only the most important chapter in our social and economic history – the industrial revolution (so-called) – but with it the wellspring of much of our local and national identity.

So isn't this a little drastic? “We're talking about a fundamental change in what we understand about the past,” he says. “That is a fairly widespread view of our work. I've always felt that you can do more with historical research than people think, but I never thought that we could do this much. And it's nothing compared with what we could achieve if we can keep the project going.”

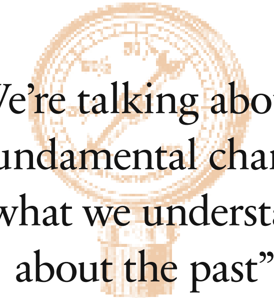
The project, as its name suggests, is a hugely ambitious, wide-scale attempt to reconstruct the picture of how working life changed and developed in Britain from the late Middle Ages through to the early 20th century. Co-directed by Shaw-Taylor and his Cambridge colleague Professor Sir Tony Wrigley, the research team has spent years assembling information about matters such as population size, transport infrastructure and sector-by-sector employment, at different points in time.

It's a complex job and, before this, nobody had really tried it. Much of what we know about social and economic history is based on records such as wills and parish registers, which are patchy, inconsistent or highly selective. As well as collating information, the team therefore had to develop a method of controlling for this lack of coherence, to avoid distorting the resulting picture of the past. “We had to develop a system of weighting the importance of the data when analysing it,” Shaw-Taylor explains. “We still can't be sure that it's right, but it puts a limit on the extent to which we can be wrong.”

Textbook orthodoxy says that, before the industrial revolution, most people in Britain worked in primary sector employment, overwhelmingly in agriculture. During the ‘revolutionary’ 80-year period starting in about 1760, this landscape was transformed as secondary industries – like processing and manufacturing – took off. Only in the 1950s did Britain supposedly begin to evolve into the tertiary, service-based economy that we have today.

On such things are national and local myths founded – tales of a green and pleasant land that rapidly became black with the smog of industry, for example, or of a country that used to make things, but doesn't any more.

When Shaw-Taylor and colleagues looked at the data that they had assembled, however, they found that it didn't fit the existing picture. Nationally, for example, secondary sector employment seems to have grown more between 1500 and 1750 than between 1750 and 1850. “We've always presumed that the major structural shift in employment from the primary to the secondary sector took place between 1750 and 1850,” he says. “Well, according to what we've found, that change took place about 100 years earlier than we thought.”



“We're talking about a fundamental change in what we understand about the past”

Similarly, the data transforms our picture of the evolution of tertiary, service-based industries in Britain. Rather than taking off in the mid-20th century, these seem to have been growing all the way through the 18th and 19th. By 1911, one man in 10 was, for example, working in transport – others were shopkeepers, merchants, clerks or professionals.

If this is true, it means an adjustment to our ‘island story’ that has some radical implications for the history of places far beyond these shores as well. For instance, it is often argued that Britain's industrialisation was made possible thanks to the raw materials gathered by the slaves of Empire. If industrialisation began before the Empire existed, however, as these findings suggest, the story changes. “Moreover, for a small island off the coast of north-west Europe to start projecting its power around the world, something unusual must have happened internally before that, not after,” Shaw-Taylor points out.

Equally, if the shift to secondary sector employment happened before the dark, Satanic mills that populate the nation's consciousness as temples of the industrial revolution even existed, then we need to modify our picture of what people were actually doing. If not farming, then what?

It seems likely that more early-modern Brits than we thought were carpenters, shoemakers, bakers, butchers, tailors and masons. This, in turn, raises puzzles about when and why agricultural and primary labour ceased to be dominant. The likelihood is that the evolution of more productive, less labour-intensive farming led to a decline in the relative importance of primary work. Over time, the children and grandchildren of agriculturalists would have been drawn to new opportunities in the secondary sector, or even tertiary, service industries.

Much remains to be done and there are still significant gaps in the research, most notably around the role of women in British employment history. Many historians associate the industrial revolution with new opportunities for female employment; others believe, just as fervently, that female employment collapsed. Only with more work and more funding will the team be able to establish exactly how women's lives, and the family, changed during this period, and the consequences that this had for women's social status.

What exists at the moment is, nevertheless, a compelling case for a data-led approach to writing the story of the past. “Methodologically, explaining why things happened in history is very difficult because it only happens once and you can't run it under controlled conditions,” Shaw-Taylor observes. “Yet the processes historians are trying to describe are often vastly more complex than those described by science. Our approach has been to eschew questions of why until we have the data at our disposal. Until you have those patterns, you're just trying to explain things that may or may not have happened, and that's a waste of time.”

Funded by the British Academy, the Economic and Social Research Council, The Leverhulme Trust and the Isaac Newton Trust

I
Dr Leigh Shaw-Taylor
lmws2@cam.ac.uk
Faculty of History

Counting on sheep

Sheep are smarter than we might think, with brains surprisingly similar to ours. These similarities are helping researchers to study a devastating and incurable infant brain disease.

"Shall we take one of the sheep for a walk?" asks Professor Jenny Morton before we head down to the farmyard.

This seems a strange question at first: we're all familiar with sheep behaving with a flock mentality, unable to think for themselves. So much so, in fact, that 'follow like a sheep' is a commonly used, derogatory phrase in the English language.

Yet, on meeting the sheep, it is immediately clear that these are not just dumb animals. The individual characters portrayed in the animated film Shaun the Sheep might be closer to the truth. "These animals are really smart," explains Morton, who leads a team in the Department of Physiology, Development and Neuroscience. "They all have their own personalities."

Morton's colleague Dr Nicholas Perentos lets Isabella, one of his sheep, out of her pen. She is excited to be out, but doesn't bound off; rather, she follows Perentos closely at heel, like a Labrador following its master. Once outside, she runs up and down the farmyard, stopping 'to say hello' to other sheep before returning expectantly to her handler. "She's definitely Nic's sheep," says Morton. "She knows who I am, but I'm not wearing my usual farm clothes today, so she's a little wary of me."

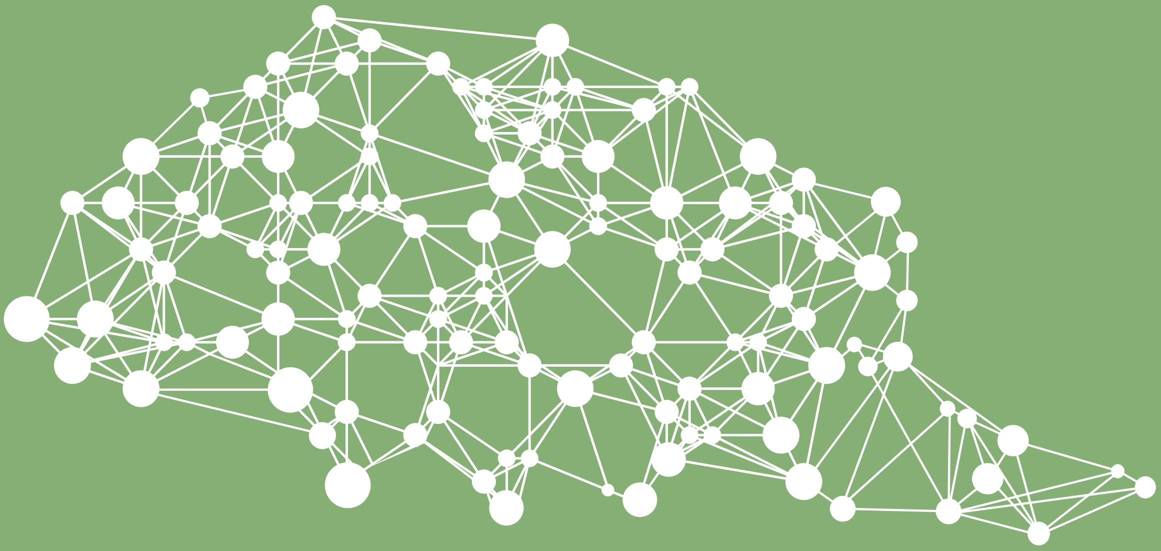
Morton and colleagues are studying the cognitive skills and behaviour of these sheep, using experiments adapted from those carried out with humans. A standard task they use is to give the sheep two options and measure their behaviour: choose option A and they receive pellets, choose B and they receive nothing.

Using electroencephalography (EEG), the researchers can measure patterns of electrical activity across the brain to see what is happening when the sheep make decisions. Recently, they have begun making measurements from deep inside the brain. "We can now record from individual neurons as they fire," says Perentos. "This might be in response to a particular task or a decision they're making, or it might be cells that 'fire' depending on where they are standing or which way they are turning." The discovery of these location-specific cells in mice – so-called 'place cells' – last year won Professor John O'Keefe from University College London a Nobel Prize.

Once the animal knows the task, the researchers will reverse the choices: now option B gives the pellets, but nudging the lever for option A offers no reward. Rats, monkeys, sheep and humans all learn to switch; but, compared with rodents, sheep react very differently, explains Morton. "When they don't get their reward they'll turn around and walk up to Nic, baa-ing, as though they're saying 'The apparatus isn't working, go and sort it out'."

The sheep's intelligence is one reason why Morton believes they are a useful animal to help us understand how the brain works. There are some practical reasons – their docile nature makes them easy to manage and their large body size means they can easily carry equipment such as GPS trackers in a harness on their backs, allowing researchers to measure their natural behaviour – but it is the size and structure of their brains that is key.

Sheep's brains are much larger than those of rodents, similar in size to the brain of a rhesus macaque, and with the complex folds that are seen in primate brains. Crucially, their brains also have



“We can now record from individual neurons as they fire”

basal ganglia similar to ours – this is the area deep in the brain that, along with the cerebral cortex, is responsible for important functions such as the control of movement and ‘executive functions’ such as decision-making, learning and habit formation. It’s this latter facet that makes sheep a useful model for studying brain diseases such as Huntington’s disease and Batten disease that affect the basal ganglia and cerebral cortex.

You may never have heard of Batten disease: it’s extremely rare, and only a handful of infants or children are diagnosed each year in the UK. It is a genetic disease caused when a child carries two copies of an aberrant gene – one copy from each parent. But it is also extremely serious – symptoms include progressive blindness, severe seizures and the loss of language, swallowing and motor skills. Death at a young age is inevitable and there is no cure.

Although Batten disease affects humans, it has never been seen in other primates. It does, however, occur naturally in sheep, though it’s unclear how common it is, as most farmed sheep are killed as lambs for human consumption. The disease was identified in sheep in New Zealand, and it is from these sheep that Morton’s animals were bred. Some of her sheep are imported, others are studied in New Zealand.

Batten disease is very similar in sheep and humans. At first, it is difficult to spot a Batten sheep, but after about a year, they begin to lose their eyesight and show unusual behaviour. After 18 months to two years, they show signs of dementia, often standing motionless in space, and can become agitated if handled by someone other than their usual handler.

Recording brain activity, particularly in areas such as the hippocampus, which is crucial for memory and learning, will give Morton and her team insights into what goes wrong in the disease in sheep. This is one step along the long path towards treating – even curing – the disease in humans.

With collaborators in Australia, Morton is also studying Huntington’s disease, a more common but equally devastating disease. Unlike those with Batten disease, people – and sheep – with Huntington’s do not begin showing symptoms until adulthood. “We have good mouse models for studying Huntington’s disease, but mice are short-lived animals, whereas sheep can live to at least 12 years. This is another huge benefit of studying the disease in sheep.”

There is no question that research using animals remains controversial. There are some who believe that animal research can never be justified. Morton has herself encountered extreme examples of such people in the past and has faced death threats because of her work. But she knows that her work is extremely important for the families of children with Batten disease.

“There’s only one thing worse than being a parent with a child who is blind, losing their motor skills and developing dementia,” she says, “and that’s being a parent with a child who is blind, losing their motor skills and developing dementia, and thinking that no one is asking why. That’s why we have a duty to do our research.”

Professor Jenny Morton
ajm41@cam.ac.uk
Dr Nicholas Perentos
Department of Physiology,
Development and Neuroscience

The Great British Takeoff

“Increasing one ingredient might produce one sought-after property, but at the sake of another... we need to find the perfect chemical recipe”

Credit: Rolls-Royce plc

The Periodic Table may not sound like a list of ingredients but, for a group of materials scientists, it's the starting point for designing the perfect chemical make-up of tomorrow's jet engines.

Inside a jet engine is one of the most extreme environments known to engineering.

In less than a second, a tonne of air is sucked into the engine, squeezed to a fiftieth of its normal volume and then passed across hundreds of blades rotating at speeds of up to 10,000 rpm; reaching the combustor, the air is mixed with kerosene and ignited; the resulting gases are about a third as hot as the sun's surface and hurtle at speeds of almost 1,500 km per hour towards a wall of turbines, where each blade generates power equivalent to the thrust of a Formula One racing car.

Turbine blades made from 'super' materials with outstanding properties are needed to withstand these unimaginably challenging conditions – where the temperatures soar to above the melting point of the turbine components and the centrifugal forces are equivalent to hanging a double-decker bus from each blade.

Even with these qualities, the blades require a ceramic layer and an air cooling system to prevent them from melting when the engine reaches its top temperatures. But with ever-increasing demands for greater performance and reduced emissions, the aerospace industry needs engines to run even hotter and faster, and this means expecting more and more from the materials they are made from.

This, says Dr Cathie Rae, is *the* materials grand challenge. "Turbine blades are made using nickel-based superalloys, which are capable of withstanding the phenomenal stresses and temperatures they need to operate under within the jet engine. But we are running close to their critical limits."

An alloy is a mixture of metals, such as you might find in steel or brass. A superalloy, however, is a mixture that imparts superior mechanical strength and resistance to heat-induced deformation and corrosion.

Rae is one of a team of scientists in the Rolls-Royce University Technology Centre (UTC) at the Department of Materials Science and Metallurgy. The team's research efforts are focused on extracting the greatest possible performance from nickel-based superalloys, and on designing superalloys of the future.

Current jet engines predominantly use alloys containing nickel and aluminium, which form a strong cuboidal lattice. Within and around this brick-like structure are up to eight other components that form a 'mortar'. Together, the components give the material its superior qualities.

"Even tiny adjustments in the amount of each component can have a huge effect on the microscopic structure, and this can cause radical changes in the superalloy's properties," explains Dr Howard Stone. "It's rather like adjusting the ingredients in a cake – increasing one ingredient might produce one sought-after property, but at the sake of another. We need to find the perfect chemical recipe."

Stone is the Principal Investigator overseeing a £50 million Strategic Partnership on structural metallic systems for advanced gas turbine applications funded jointly by Rolls-Royce and the Engineering and Physical Sciences Research Council (EPSRC), and involving the Universities of Birmingham, Swansea, Manchester, Oxford and Sheffield, and Imperial College London.

The researchers melt together precise amounts of each of the different elements to obtain a 5cm bar, then exhaustively test the bar's mechanical properties and analyse its microscopic structure. Their past experience in atomic engineering is vital for homing in on where the incremental improvements might be found – without this, they would need to make many millions of bars to test each reasonable mixture of components.

Now, they are looking beyond the usual components to exotic elements, although always with an eye on keeping costs as low as possible, which means not using extremely rare materials. "The Periodic Table is our playground… we're picking and mixing elements, guided by our computer models and experimental experience, to find the next generation of superalloys," he adds.

The team now have 12 patents with Rolls-Royce. One of the most recent has been in collaboration with Imperial College London, and involves the discovery that the extremely strong matrix structure of nickel-based aluminium superalloys can also be achieved using a mixture of nickel, aluminium, cobalt and tungsten.

"Instead of the cake being flavoured with two main ingredients, we can make it with four," Stone explains. "This gives the structure even better properties, many of which we are only just discovering."

"We've also been looking at new intermetallic reinforced superalloys using chromium, tantalum and silicon – no nickel at all. We haven't quite got the final balance to achieve what we want, but we're working towards it."

Stone highlights the importance of collaboration between industry and academia: "New alloys typically take 10 years and many millions of pounds to develop for operational components. We simply couldn't do this work without Rolls-Royce. For the best part of two decades we've had a collaboration that links fundamental materials research through to industrial application and commercial exploitation."

It's a sentiment echoed by Dr Justin Burrows, Project Manager at Rolls-Royce: "Our academic partners understand the materials and design challenges we face in the development of gas turbine technology. Improvements like the novel nickel and steel alloys developed in Cambridge are key to helping us meet these challenges and to maintaining our competitive advantage."

The Cambridge UTC, which was founded by its Director Professor Sir Colin Humphreys in 1994, is one of a global network of over 30 UTCs. These form part of Rolls-Royce's £1 billion annual investment in research and development, which also includes the Department of Engineering's University Gas Turbine Partnership. Rolls-Royce and EPSRC also fund Doctoral Training Centres in Cambridge that help to ensure a continuing supply of highly trained scientists and engineers ready to move into industry.

The UK aerospace industry is the largest in Europe, with a turnover in 2011 of £24.2 billion; worldwide, it's second only to that of the USA. Meanwhile, increasing global air traffic is estimated to require 35,000 new passenger aircraft by 2030, worth about $4.8 trillion.

For the researchers, it's fascinating to see global engineering challenges being solved from the atom up, as Rae explains: "The commercial success of a new engine can be dependent on very small differences in fuel efficiency, which can only be achieved by innovations in materials and design. There's something really exciting about working at the atomic scale and seeing this translate into innovation with big powerful machines."

Film available online

Dr Cathie Rae
cr18@cam.ac.uk
Dr Howard Stone
hjs1002@cam.ac.uk
Department of Materials Science
and Metallurgy

Credit: All images, National Library of Wales

Haunting of the *Black Book*

The 16th-century owner of one of Wales' oldest manuscripts probably thought they were 'tidying up' when they assiduously erased ancient doodles and verses scribbled in its margins. Now, Cambridge researchers have brought them back to life.

Professor Paul Russell and PhD student Myriah Williams had been peering at the ancient manuscript for several hours, methodically turning page after page and adjusting the ultraviolet (UV) lamp in the hope of casting new light and understanding on a 750-year-old masterpiece.

Other readers and researchers had come and gone from the Reading Room at the National Library of Wales as the pair ploughed on. But, despite their efforts, the vellum pages had revealed only the medieval Welsh poetry they knew so well, plus a few tiny fragments of text in the margins, none of which were particularly remarkable or noteworthy.

Then, as the UV light fell on folio 39v of the manuscript, Russell turned in astonishment to his colleague and asked: "Are you seeing what I'm seeing?"

There, invisible to the naked eye but appearing under the glare of UV, were a pair of ethereal faces and a line of accompanying text. With image-enhancement techniques, they were to find an entire page of erased verse that was (and remains) unknown in the canon of Welsh poetry.

The manuscript containing the ghostly images was *The Black Book of Carmarthen* – the earliest surviving medieval manuscript written solely in Welsh. Containing some of the earliest references to the legendary tales of King Arthur and Merlin, the Black Book (so-called because of the colour of its binding) is a collection of 9th- to 12th-century religious and secular poetry, and draws on the traditions of the Welsh folk-heroes and legends from the early medieval period.

However, despite its importance and decades of scholarly research, it is the work of the two Cambridge researchers that is illuminating new glimpses of verse from this ancient book.

"We knew that there had been significant erasure in the margins of the manuscript but we never expected to find two faces staring out at us," says Williams. "We thought we might recover some text but not images. You never find images."

Williams and Russell, from the Department of Anglo-Saxon, Norse and Celtic, have worked together on the Black Book for the past three years. Russell has studied the language – the nuts and bolts of spelling, punctuation, grammar, and so on – whereas for Williams the book as a whole is the subject of her PhD.

The 54-page book, which dates from 1250 and is only just larger than a hand's length, is thought to have been written by a single scribe who was probably collecting and recording poetry during a long period of his life. Then, as the manuscript changed hands over the centuries that followed, its various owners made their own additions.

Until, that is, it was 'tidied up'. They believe that a 16th-century owner of the book, possibly a man named Jaspar Gryffyth, summarily erased the marginalia.

Among the erased material is some previously unrecorded Welsh verse. Although the text is fragmentary and in need of more analysis, it seems to be the continuation of a poem on the preceding page together with a new poem at the foot of the page.

"It's easy to think we know all we can know about a manuscript like the Black Book," adds Williams. "But to see these ghosts from the past brought back to life in front of our eyes has been incredibly exciting. The drawings and verse that we're in the process of recovering demonstrate the value of giving these books another look.

"The margins of manuscripts often contain medieval and early modern reactions to the text, and these can cast light on what our ancestors thought about what they were reading. The Black Book was particularly heavily annotated before the end of the 16th century. For instance, Welsh scholar Dr John Davies of Mallwyd wrote in the margin 'I don't understand the Black Book'. This is wonderful! This was a man who wrote one of the first Welsh grammars and dictionaries! This type of reaction brings the pages to life."

The pair also recovered a drawing of a fish underneath a poem about the drowning of Cardigan Bay. The bay was flooded, so legend says, through the wrath of God, and both Russell and Williams believe the fish was drawn in connection with the poem's subject matter. Ironically, the page containing the poem about flooding shows some evidence of water damage.

Contents of the Black Book range from religious verse to praise poetry to narrative poetry. An example of the latter is the earliest poem concerning the adventures of Arthur, which sees the famed hero seeking entrance to an unidentified court and expounding the virtues of his men in order to gain admittance.

Other heroes are praised and lamented in a lengthy text known as *Englynion y Beddau*, the Stanzas of the Graves, in which a narrator presents geographic lore by claiming to know the burial places of upwards of 80 warriors. Arthur makes an appearance here as well, but only in-so-far as to say that he cannot be found: *anoeth bid bet y arthur*, 'the grave of Arthur is a wonder'.

Further famous figures also appear throughout, including Myrddin, more familiarly known by the English as 'Merlin'.

> *"We thought we might recover some text but not images. You never find images."*

There are two prophetic poems attributed to him during his 'wild man' phase located in the middle of the manuscript, but additionally the very first poem of the book is presented as a dialogue between him and the celebrated Welsh poet Taliesin. Ever since Geoffrey of Monmouth composed *Historia Regum Britanniae* in the 12th century there has been a connection between Carmarthen and Merlin, and it may be no accident that the Black Book opens with this text.

Russell believes that the new discoveries may only be the tip of the iceberg in terms of what can be recovered as imaging techniques are enhanced: "These drawings and other marginalia help us to go beyond the text to show what people thought about it, sometimes seriously, sometimes in a playful way. The manuscript is extremely valuable and incredibly important – yet there may still be so much we don't know about it."

Images
Images of *The Black Book of Carmarthen*, including the erased faces (left)

Myriah Williams
mjw202@cam.ac.uk
Professor Paul Russell
pr270@cam.ac.uk
Department of Anglo-Saxon, Norse and Celtic

Play’s the thing

Children’s play is under threat from increased urbanisation, perceptions of risk and educational pressures. The first research centre of its kind aims to understand the role played by play in how a child develops.

Brick by brick, six-year-old Alice is building a magical kingdom. Imagining fairy-tale turrets and fire-breathing dragons, wicked sorcerers and gallant heroes, she’s creating an enchanting world. Although she isn’t aware of it, this fantasy will have important repercussions in her adult life: it is helping her take her first steps towards her capacity for abstract thought and creativity.

Minutes later, Alice has abandoned the kingdom in favour of wrestling with her brother – or, according to educational psychologists, developing her capacity for strong emotional attachments. When she bosses him around as ‘his teacher’, she’s practising how to regulate her emotions through pretence. When they settle down with a board game, she’s learning about rules and turn-taking.

“Play in all its rich variety is one of the highest achievements of the human species,” says Dr David Whitebread from Cambridge’s Faculty of Education. “It underpins how we develop as intellectual, problem-solving, emotional adults and is crucial to our success as a highly adaptable species.”

Recognising the importance of play is not new: over two millennia ago, Plato extolled its virtues as a means of developing skills for adult life, and ideas about play-based learning have been developing since the 19th century.

But we live in changing times, and Whitebread is mindful of a worldwide decline in play. “Over half the world’s population live in cities. Play is curtailed by perceptions of risk to do with traffic, crime, abduction and germs, and by the emphasis on ‘earlier is better’ in academic learning and competitive testing in schools.

“The opportunities for free play, which I experienced almost every day of my childhood, are becoming increasingly scarce. Today, play is often a scheduled and supervised activity.”

International bodies like the United Nations and the European Union have begun to develop policies concerned with children’s right to play, and to consider implications for leisure facilities and educational programmes. But what they often lack is the evidence to base policies on, as Whitebread explains: “Those of us who are involved in early childhood education know that children learn best through play and that this has long-lasting consequences for achievement and well being. But the kind of hard quantifiable evidence that is understood by policy makers is difficult to obtain. Researching play is inherently tricky."

"The type of play we are interested in is child-initiated, spontaneous and unpredictable – but, as soon as you ask a five-year-old 'to play', then you as the researcher have intervened," explains Dr Sara Baker. "And we want to know what the impact of play is years, even decades, later. It's a real challenge."

Dr Jenny Gibson agrees: "Although some of the steps in the puzzle of how and why play is important have been looked at, there is very little, high-quality evidence that takes you from the amount and type of play a child experiences through to its impact on the rest of its life."

Now, thanks to the new Centre for Research on Play in Education, Development and Learning (PEDaL), Whitebread, Baker, Gibson and a team of researchers hope to provide evidence on the role played by play in how a child develops.

"A strong possibility is that play supports the early development of children's self-control," explains Baker.

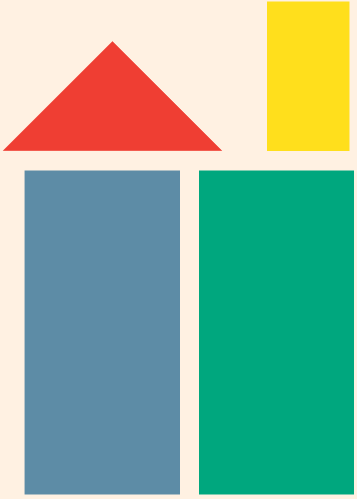
"These are our abilities to develop awareness of our own thinking processes – it influences how effectively we go about undertaking challenging activities."

In a study carried out by Baker with toddlers and young pre-schoolers, she found that children with greater self-control solved problems quicker when exploring an unfamiliar set-up requiring scientific reasoning, regardless of their IQ. "This sort of evidence makes us think that giving children the chance to play will make them more successful and creative problem-solvers in the long run."

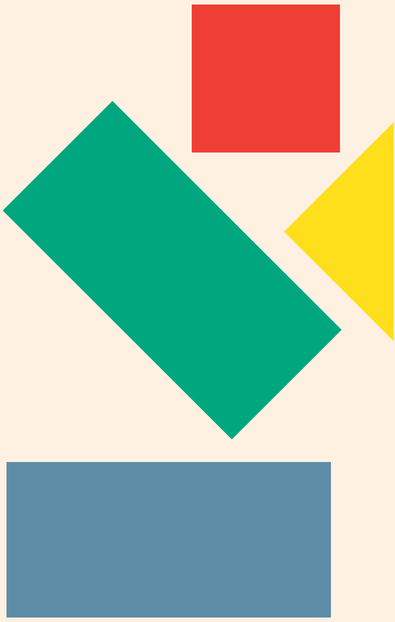
If playful experiences do facilitate this aspect of development, say the researchers, it could be extremely significant for educational practices because the ability to self-regulate has been shown to be a key predictor of academic performance.

Gibson adds: "Playful behaviour is also an important indicator of healthy social and emotional development. In my previous research, I investigated how observing children at play can give us important clues about their well being and can even be useful in the diagnosis of neurodevelopmental disorders like autism."

Whitebread's recent research has involved developing a playful approach to supporting children's writing. "Many primary school children find writing difficult, but we showed in a previous study that a playful stimulus was far more effective than an instructional one." Children wrote longer and better structured stories when they first played with dolls representing characters in the story. In the latest study, children first built their story with LEGO, with similar results. "Many teachers commented that they had always previously had children saying they didn't know what to write about. With the LEGO building, however, not a single child said this through the whole year of the project."



"The opportunities for free play, which I experienced almost every day of my childhood, are becoming increasingly scarce"



The strand of research he leads in the Centre will focus on the results of large-scale longitudinal studies, such as the University of London's Millennium Cohort Study, which is charting the social, economic and health conditions of individual children. Whitebread hopes to determine how much a child plays, the quality of their playfulness, and with what end result.

Even when this evidence is known, it is often difficult to develop practices that best support children's play. The two research strands led by Gibson and Baker will aid this: Gibson will be developing an understanding of the cognitive processes involved in play and measures of playfulness, and Baker will be constructing and evaluating play-based educational interventions.

Whitebread, who directs PEDaL, trained as a primary school teacher in the early 1970s, when, as he describes, "the teaching of young children was largely a quiet backwater, untroubled by any serious intellectual debate or controversy." Now, the landscape is very different, with hotly debated topics such as school starting age and the introduction of baseline assessment to those starting school in September 2015.

"Somehow the importance of play has been lost in recent decades. It's regarded as something trivial, or even as something negative that contrasts with 'work'. Let's not lose sight of its benefits, and the fundamental contributions it makes to human achievements in the arts, sciences and technology. Let's make sure children have a rich diet of play experiences."

Left to right
Dr David Whitebread
dgw1004@cam.ac.uk
Dr Sara Baker
stb32@cam.ac.uk
Dr Jenny Gibson
jlg53@cam.ac.uk
Faculty of Education

Bad air days

Pollution causes 30,000 people a year in the UK to die early yet most of us are unaware of the degree to which we are exposed to it. Low-cost pollution detectors could provide the answer.

Rush hour can be maddening. Roads congested with traffic, public transport overcrowded, pavements heaving with people. But as well as the frustration, there's a sinister side to the commute to work: every breath you take could be adding to your risk of dying prematurely.

Air pollution is the world's largest single environmental health risk, causing one in every eight deaths according to figures released last year by the World Health Organization. In the UK, 30,000 people die prematurely every year as a result of poor air quality, and it costs the NHS and wider economy many billions each year.

Traffic is the main culprit; however, industry, domestic heating, power generation and burning are all contributors to pollution. And although the effects of pollution might be noticeable on a particularly smoggy day in a large city, decades of exposure to only slightly higher levels – a level we wouldn't even notice – can increase the risk of heart and lung diseases, stroke and cancer.

"To work out the factors we should be worried about, and how we can intervene, we need to rethink how we measure what's going on," explains atmospheric scientist Professor Rod Jones.

In the UK, the Automatic Urban and Rural Network provides valuable hour-by-hour assessments of air quality. But with only 171 monitoring stations at fixed sites nationwide, large areas of the country remain uncovered. Cost is the main limitation to developing a higher density network.

With this in mind, Jones' team, together with industrial partners and other universities, has been developing low-cost pollution detectors that are small enough to fit in your pocket, stable enough to be installed as long-term static detectors around a city, and sensitive enough to detect small changes in air quality on a street-by-street basis. Their findings are now informing research projects aimed at improving air quality in major cities across Europe and North America.

The detectors are based on electrochemical sensors developed by project partner Alphasense for industrial safety, where detection of toxic gases is needed at the parts-per-million level. Monitoring air quality, however, requires parts-per-billion sensitivity. "Rod and I had the confidence to believe that we could push our sensors to lower concentration levels, and yet keep sensor costs low," says Dr John Saffell, Technical Director at Alphasense.

The electrochemical devices the team developed can measure a wide range of pollutants, including carbon monoxide, nitrogen dioxide and ozone, and they contain laser technology (developed by the University of Hertfordshire) to detect particulates from cars and lorries. The addition of a GPS aerial allows air quality data and location to be mapped simultaneously.

A series of proof-of-concept studies followed. Personal devices were strapped to bicycles, carried in cars and on buses, and static devices were attached to lampposts and stationed at roadsides and at critical pollutant sites. Fifty static devices were also deployed around London Heathrow Airport to record 22 months in the life of one of the busiest airports in the world.

"This was the first time technology like this had been tested in real-world situations as a high-density network," says Jones, whose research at Heathrow

> *"Air pollution is the world's largest single environmental health risk, causing one in every eight deaths"*

was funded by the Natural Environment Research Council. "We could see huge variability in the exposure to pollution that people encounter as they move around the urban environment, including 'hotspots'.

At Heathrow, we could see the airport turning on and off during the day, individual aircraft taxing and taking off, and the effects of wind direction and the perimeter and M25 motorway road traffic."

They also discovered that sensor performance can create new opportunities. Jones and colleagues had to develop new smart software methods capable of separating local pollution events from background signals (pollution transported from long range) and then to calibrate sensors across networks. Plus, they needed to move from being able to process the data after it has been collected to doing so in real-time.

The team has been working with Cambridge Environmental Research Consultants – developers of world-leading air quality modelling software – combining the unprecedented level of data created by the pollution-monitoring studies with model output to enhance the understanding of pollution dispersion.

For instance, sensors can be used to ask whether pollution along bus routes is improved by upgrading the exhaust processing on a bus fleet; whether people living at the top of high-rise buildings experience more or less pollution than people at street level; and to what extent changing a route to work, even from one side of the road to another, can affect an individual's exposure.

Last year, the first commercial product (AQMesh) was released by UK manufacturer Geotech, which specialises in environmental monitoring equipment. AQMesh uses Alphasense sensors to sample every 10 seconds, and data processing is carried out in real-time using cloud computing software similar to that developed by the Cambridge team.

"When the project started in 2006 there were lone voices calling for a different approach to air quality monitoring," explains Geotech's Commercial Manager Amanda Randle. "The Cambridge team and Alphasense helped us to understand the sensor's full potential, and now we have a product that can be placed exactly where it's needed and provides valuable information."

And now the approach pioneered in Cambridge is helping to inform two of the largest air quality research studies of their kind.

The AirSensa project, run by the non-profit organisation Change London, aims to deploy large numbers of air quality sensors across the whole of Greater London. Alphasense is providing the sensors and supporting the engineering; and Cambridge is helping with data interpretation in a project whose ethos is "you can't manage what you can't measure."

Meanwhile, the methodologies the researchers developed in the pilot study at Heathrow are contributing to CITI-SENSE, an EU-funded €12.7 million project providing wireless networks to eight cities across Europe. CITI-SENSE involves 27 partner institutions from academia, the healthcare sector and industry (including Alphasense and Geotech), as well as the general public. Citizens across Europe will be involved in data collection through personal monitors and in community decision-making to choose monitoring solutions for spaces such as schools and urban public spaces.

"Even though the effects of poor air quality on health are well known, irrefutable evidence of the scale of the air quality issue and the benefits of ameliorating strategies is urgently needed," adds Jones. "CITI-SENSE provides a test-bed for both rolling out the new technologies that are coming online and for drawing on the 'power of the Citizen' to guide how society responds."

Film available online

Professor Rod Jones
rlj1001@cam.ac.uk
Department of Chemistry

Things
A is for Albatross

Images
Edward Wilson's sketches of albatrosses

Feature and film available
bit.ly/1EmWZdL

A fascinating 26-part series – the Cambridge Animal Alphabet – has launched online. It celebrates Cambridge's connections with animals through literature, art, science and society.

Horses frolic on the plaster cast of the Parthenon frieze in the Museum of Classical Archaeology; research into chickens is helping to understand a major source of food poisoning; a new book explores the making of the modern dog as a pampered urban pet; and, every day, millions of fruit flies are fed by the fly kitchen in the Department of Genetics.

These are just some of the stories we discovered when we set out to create a Cambridge Animal Alphabet: an A to Z of animals with a Cambridge connection. The series, which is available on our website, launches with *A is for Albatross…*

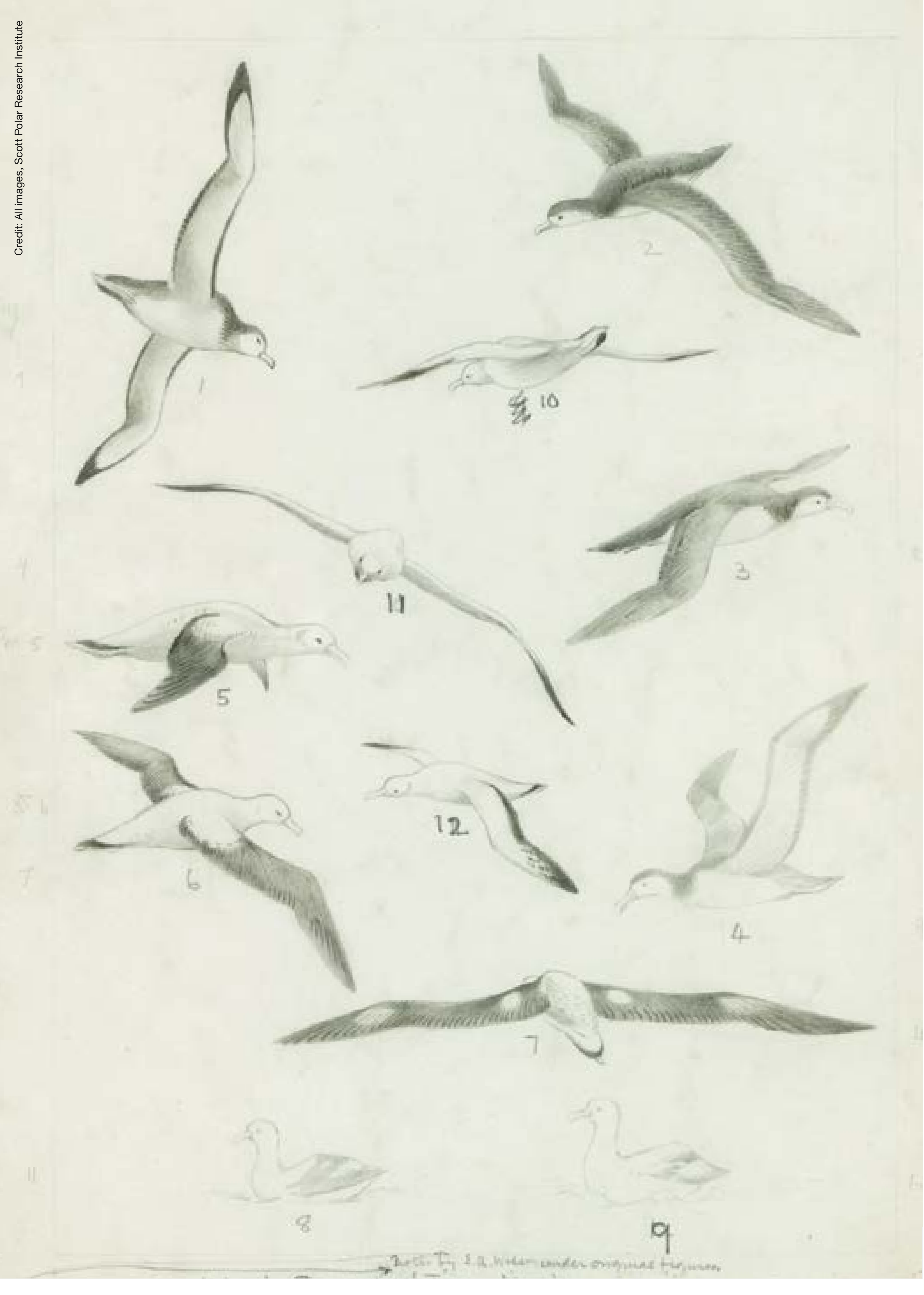
In June 1910, Dr Edward Wilson set sail to Antarctica on board the *Terra Nova* on the British Antarctic Expedition led by Captain Scott. A supremely talented artist, Wilson sketched what he saw – including the majestic albatross.

The expedition ended in tragedy. The members of the British expedition perished on their return from the pole having discovered that the Norwegians had got there first. Wilson's sketchbook was retrieved from the tent where he and his companions spent their last days.

Today, around 1,900 of Wilson's drawings and sketches are held by Cambridge's Scott Polar Research Institute (SPRI), which houses a unique collection of materials illustrating polar exploration, history and science.

"Wilson is one of the greatest artists of the heroic age of polar exploration," says Heather Lane, former Keeper of the Polar Museum at SPRI. "He captured with stunning accuracy both the anatomical structure and the fragile beauty of living things."

Read more about Wilson's sketches and our research on albatrosses at www.cam.ac.uk/research, and watch out for *B is for Bear*, *C is for Chicken*, *D is for Dragon* as we work our way through the Cambridge Animal Alphabet.



Credit: All images, Scott Polar Research Institute

With more information than ever at our fingertips, statisticians are vital to innumerable fields and industries. Welcome to the world of the datarati, where humans and machines team up to crunch the numbers.

Let’s get statted

Researchers are now refining the system to cope with the messy, incomplete nature of real-world data

"I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding," Hal Varian, Chief Economist at Google famously observed in 2009. It seems a difficult assertion to take seriously, but six years on, there is little question that their skills are at a premium.

Indeed, we may need statisticians now more than at any time in our history. Even compared with a decade ago, we can now gather, produce and consume unimaginably large quantities of information. As Varian predicted, statisticians who can crunch these numbers are all the rage. A new discipline, 'Data Science', which fuses statistics and computational work, has emerged.

"People are awash in data," reflects Zoubin Ghahramani, Professor of Information Engineering at Cambridge. "This is occurring across industry, it's changing society as we become more digitally connected, and it's true of the sciences as well, where fields like biology and astronomy generate vast amounts of data."

Over the past few years, Richard Samworth, Professor of Statistics, has watched the datarati step out from the shadows. "It's probably fair to say that statistics didn't have the world's best PR for quite a long time," he says. "Since this explosion in the amount of data that we can collect and store, opportunities have arisen to answer questions we previously had no hope of being able to address. These demand an awful lot of new statistical techniques."

'Big data' is most obviously relevant to the sciences, where large volumes of information are gathered to answer questions in fields such as genetics, astronomy and particle physics, but it also has more familiar applications. Transport authorities gather data from electronic ticketing systems like Oyster cards to understand more about passenger movements; supermarkets closely monitor customer transactions to react to shoppers' predilections. As users of social media, many of us disclose data about ourselves that is as valuable to marketing as it is relevant to psychoanalytics. Increasingly, we are also 'lifeloggers', monitoring our own behaviour, health, diet and fitness, through smart technology.

This information, as Ghahramani points out, is no use on its own: "It fills hard drives, but to extract value from it, we need methods that learn patterns in the data and allow us to make predictions and intelligent decisions." This is what statisticians, computer scientists and machine learning specialists bring to the party – they build algorithms, which are coded as computer software, to see patterns. At root, the datarati are interpreters.

Despite their 'sexy' new image, however, not enough data scientists exist to meet this rocketing demand. Could some aspects of the interpretation be automated using artificial intelligence instead, Ghahramani wondered? And so, in 2014 and with funding from Google, the first incarnation of The Automatic Statistician was launched online. Despite minimal publicity, 3,000 users uploaded datasets to it within a few months.

Once fed a dataset, the Automatic Statistician assesses it against various statistical models, interprets the data and – uniquely – translates this interpretation into a short report of readable English. It does this without human intervention, drawing on an open-ended 'grammar' of statistical models. It is also deliberately conservative, only basing its assessments on sound statistical methodology, and even critiquing its own approach.

Ghahramani and his team are now refining the system to cope with the messy, incomplete nature of real-world data, and also plan to develop its base of knowledge and to offer interactive reports. In the longer term, they hope that the Automatic Statistician will learn from its own work: "The idea is that it will look at a new dataset and say, 'Ah, I've seen this kind of thing before, so maybe I should check the model I used last time'," he explains.

While automated systems rely on existing models, new algorithms are needed to extract useful information from evolving and expanding datasets. Here, the role of human statisticians is vital.

To characterise the problem, Samworth presents a then-and-now comparison. During the past century, a typical statistical problem might, for instance, have been to understand the relationship between the initial speed and stopping distance of cars based on a sample size of 50.

These days, however, we can record information on a huge number of variables at once – the weather, road surface, make of car, wind direction, and so on. Although the extra information has the potential to yield better models and reduce uncertainty, in many areas, the number of features measured is so high it may even exceed the number of observations. Identifying appropriate models in this context is a serious challenge, which requires the development of new algorithms.

To resolve this, statisticians rely on a principle called 'sparsity'; the idea that only a few bits of the dataset are really important. The statistician identifies these needles in the haystack. Various algorithms have been developed to select the important variables, so that the initial sprawl of information starts to become manageable and patterns can be extracted.

"It fills hard drives, but to extract value from it, we need methods that learn patterns in the data"

Together with his colleague Dr Rajen Shah in the Department of Pure Mathematics and Mathematical Statistics, Samworth has developed a method for refining any such variable selection technique called 'Complementary Pairs Stability Selection'. This applies the original method to random subsamples of the data instead of the whole, and does this over and over again. Eventually, the variables that appear on a high proportion of the subsamples emerge as those meriting further attention.

Scanning Google Scholar for citations of the paper in which this was proposed, Samworth finds that his algorithm has been used in numerous research projects. One looks at how to improve fundraising for disaster zones, another examines potential biomarkers for breast cancer survival, and a third identifies risk factors connected with childhood malnutrition.

How does he feel when he sees his work being applied so far and wide? "It's funny," he says. "My training is in mathematics and I still get a kick from proving a theorem, but it's also rewarding to see people using your work. It's often said that the good thing about being a statistician is that you get to play in everyone's back yard. I suppose this demonstrates why that's true."

I **Left to right**
Professor Zoubin Ghahramani
zg201@cam.ac.uk
Department of Engineering
Professor Richard Samworth
rjs57@cam.ac.uk
Department of Pure Mathematics and Mathematical Statistics

The 'world's largest IT project' — a system with the power of one hundred million home computers — may help to unravel many of the mysteries of our universe: how it began, how it developed and whether humanity is alone in the cosmos.

Imagine having to design a completely automated system that could take all of the live video from all of the hundreds of thousands of cameras monitoring London, and automatically dispatch an ambulance any time any person falls and hurts themselves, anywhere in the city, without any human intervention whatsoever. That is the scale of the problem facing the team designing the software and computing behind the world's largest radio telescope.

When it becomes operational in 2023, the Square Kilometre Array (SKA) will probe the origins, evolution and expansion of our universe; test one of the world's most famous scientific theories; and perhaps even answer the greatest mystery of all — are we alone?

Construction on the massive international project, which involves and is funded by 11 different countries and 100 organisations, will start in 2018. When complete, it will be able to map the sky in unprecedented detail — 10,000 times faster and 50 times more sensitively than any existing radio telescope — and detect extremely weak extraterrestrial signals, greatly expanding our ability to search for planets capable of supporting life.

The SKA will be co-located in South Africa and Australia, where radio interference is least and views of our galaxy are best. The instrument itself will be made up of thousands of dishes that can operate as one gigantic telescope or multiple smaller telescopes — a phenomenon known as astronomical interferometery, which was developed in Cambridge by Sir Martin Ryle almost 70 years ago.

"The SKA is one of the major big data challenges in science," explains Professor Paul Alexander, who leads the Science Data Processor (SDP) consortium, which is responsible for designing all of the software and computing for the telescope. In 2013, the University's High Performance Computing Service unveiled 'Wilkes' — one of the world's greenest supercomputers with the computing power of 4,000 desktop machines running at once, and a key test-bed for the development of the SKA computing platform.

During its projected 50-year lifespan, the SKA will carry out several experiments to study the nature of the universe. Cambridge researchers will focus on two of these, the first of which will follow hydrogen through billions of years of cosmic time.

"Hydrogen is the raw material from which everything in the universe developed," says Alexander. "Everything we can see in the universe and everything that we're made from started out in the form of hydrogen and a small amount of helium. What we want to do is to figure out how that happened."

The second of the two experiments will look at pulsars — spinning neutron stars that emit short, quick pulses of radiation. Since the radiation is emitted at regular intervals, pulsars also turn out to be extremely accurate natural clocks, and can be used to test our understanding of space, time and gravity, as proposed by Einstein in his general theory of relativity.

By tracking a pulsar as it orbits a black hole, the telescope will be able to examine general relativity to its absolute limits. As the pulsar moves around the black hole, the SKA will follow how the clock behaves in the very strong gravitational field.

"General relativity tells us that massive objects like black holes warp the space–time around them, and what we call gravity is the effect of that warp," says Alexander. "This experiment will enable us to test our theory of gravity with much greater precision than ever before, and perhaps even show that our current theories need to be changed."

Although the SKA experiments will tell us much more than we currently know about the nature of the universe, they also present a massive computing challenge. At any one time, the amount of data gathered from the telescope will be equivalent to five times the global internet traffic, and the SKA's software must process that vast stream of data quickly enough to keep up with what the telescope is doing.

Masters of the universe

how it began, how it developed
and whether humanity is alone in the cosmos

Credit: SKA Organisation

Moreover, the software also needs to grow and adapt along with the project. The first phase of the SKA will be just 10% of the telescope's total area. Each time the number of dishes on the ground doubles, the computing load will be increased by more than the square of that, meaning that the computing power required for the completed telescope will be more than 100 times what is required for phase one.

"You can always solve a problem by throwing more and more money and computing power at it," says Alexander. "We have to make it work sensibly as a single system that is completely automated and capable of learning over time what the best way of getting rid of bad data is. At the moment, scientists tend to look at data but we can't do that with the SKA, because the volumes are just too large."

The challenges faced by the SKA team echo those faced in many different fields, and so Alexander's group is working closely with industrial partners such as Intel and NVIDIA, as well as with academic and funding partners including the Universities of Manchester and Oxford, and the Science and Technology Facilities Council. The big data solutions developed by the SKA partners to solve the challenges faced by a massive radio telescope can then be applied across a range of industries.

One of these challenges is how to process data efficiently and affordably, and convert it into images of the sky. The target for the first phase of the project is a 300 'petaflop' computer that uses no more than eight megawatts of power: more than 10 times the performance of the world's current fastest supercomputer, for the same amount of energy. 'Flops' (floating point operations per second) are a standard measure of computing performance, and one petaflop is equivalent to a million billion calculations per second.

"The investment in the software behind the SKA is as much as €50 million," adds Alexander. "And if our system isn't able to grow and adapt, we'd be throwing that investment away, which is the same problem as anyone in this area faces. We want the solutions we're developing for understanding the most massive objects in the universe to be applied to any number of the big data challenges that society will face in the years to come."

Image
Artist's impression of the SKA, which will be made up of thousands of dishes that operate as one gigantic telescope

Professor Paul Alexander
pa@mrao.cam.ac.uk
Department of Physics

The Big Dating Game

When is a rare disease not a rare disease? The answer: when big data gets involved. An ambitious new research project aims to show patients that they are not alone.

At some point in their career, every doctor will encounter a patient whose condition perplexes them, requiring detailed investigation and discussion with colleagues before diagnosis is possible. After all, not every disease is as common as cancer, which affects around one in three of us, or depression, which affects one in 10.

Dr Lucy Raymond from the Department of Medical Genetics specialises in rare diseases. Technically, this means diseases that affect fewer than one in 2,000 people, but in fact, Raymond sees children with learning disabilities so rare that they may be the only person in the UK to be affected.

These conditions are usually caused by one of two scenarios: a spontaneous change to their DNA, not inherited, or a ‘recessive disorder’ where two copies of the same, rare variant are necessary for the disease and each parent unwittingly passes on a copy. Comparing the child’s and their parents’ genomes enables the researchers to pinpoint the gene responsible. In extremely rare cases – where the patient appears to be truly unique – the researchers need to study whether the same variant in mice or zebrafish creates a similar condition.

“Or,” Raymond explains, “we might essentially generate a ‘dating agency’

to try to match our patient with a similar case somewhere else in the world.” With these diseases as rare as they are, the only way for this to be viable would be to have access to tens, possibly hundreds, of thousands of potential matches: something the era of ‘big data’ makes possible.

But this presents a potential problem: how to share information about the patient without breaking their confidentiality. Unlike in the USA, where projects such as the Broad Institute’s Exome Aggregation Consortium (ExAC) place genome data in the public domain, data in the UK is deposited in a ‘managed-access’ database: bona fide researchers with a clear research proposal are allowed access, and only then after signing a commitment saying they will not attempt to identify individual patients.

“We have to remember that big data is great, but it isn’t our data: it’s people’s data and we need to be respectful of this. People in the UK are often altruistic; we have free blood donation, we have a tremendous tradition of patients giving to help others. We must not jeopardise this relationship.

“Parents know that even if finding the gene abnormality that is responsible will not immediately help their child, it may help ensure that others don’t have to wait 20 years before their child receives a diagnosis. They’re happy to share the data on that basis, but are less keen on the idea that they’ll lose control of the information.”

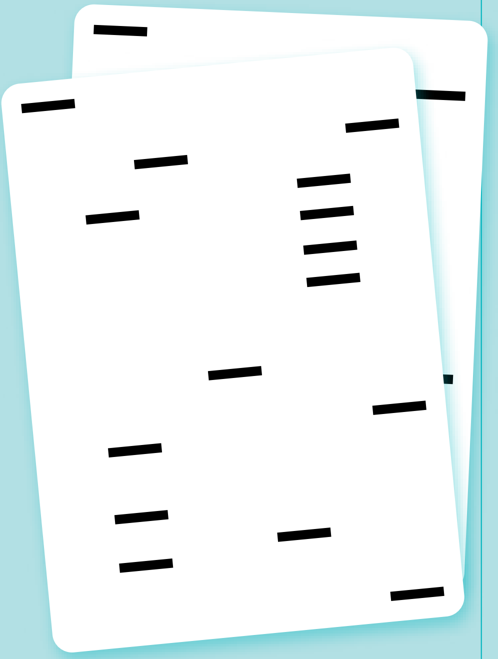
For several years, Raymond, Professor Willem Ouwehand and Dr John Bradley have been leading the National Institute for Health Research BioResource for Rare Diseases in Cambridge, which has recruited some 5,800 patients. They are now part of a major initiative launched by Prime Minister David Cameron: the 100,000 Genomes Project. Cambridge University Hospitals NHS Foundation Trust will lead the East of England Genomic Medicine Centre, one of 11 centres across the UK aimed at realising this project and sequencing the genomes of patients affected by cancer or rare diseases.

“The 100,000 Genomes Project is about going forward to having a truly national health service, not a provincial, regional health service,” explains Raymond. “The data will be central, will be national, will be available to researchers and healthcare professionals across the country.”

The sheer number of people recruited will create a powerful dataset and ensure that clinicians and researchers don’t have to start from scratch each time they encounter a new case. In fact, the value of a patient’s genome extends beyond just helping identify the cause of their disease: it’s also important as a ‘control’ to compare against and help find the cause of another patient’s disease. “It’s a form of ‘enforced altruism’. Having all the data stored in a central place means that everybody’s data acts as a control for everybody else’s. It has a multiplying effect.”

Big data also reveals an otherwise glaringly obvious fact that the name ‘rare diseases’ obscures: one in 2,000, even in a population of 64 million, is not an insignificant number of people. “Ten years ago people used to ask ‘Why study rare diseases when they’re so rare?’ It’s only recently that people are coming round to see that, with big data, rare is common.

“Rare diseases are becoming increasingly tractable, too, so now there’s a huge interest in them, which is good: it’s not your fault if your disease is rare. Solving these problems is the next big challenge,” says Raymond with a glint in her eye. “If it was all easy, we wouldn’t be doing it – in typical Cambridge style.”



I

Left to right
Dr Lucy Raymond
flr24@cam.ac.uk
Department of Medical Genetics
Dr Lydia Drumright
lnd23@cam.ac.uk
Department of Medicine

Trust me, I’m an e-doctor

Big data ‘dating agencies’ are not just for people with rare conditions. A similar concept could help patients with far more common conditions receive the best possible hospital treatment.

Addenbrooke’s Hospital in Cambridge is one of the first ‘eHospitals’ in England, explains Dr Lydia Drumright from the Department of Medicine. Everything that happens to you within the hospital – every test result, every diagnosis, every drug prescribed – is captured in an electronic record. Drumright and her colleague Dr Afzal Chaudhry believe that the wealth of information in these records can be used to better inform the treatments of individuals.

“Around 10–20% of our patients may have diabetes or acute kidney injury, but that’s not necessarily why they’re here,” explains Drumright. “They might have had a heart attack, so they’re being cared for by the cardiology team, but the drugs they’re prescribed might have an impact on their other conditions. Added to that, they’re now more susceptible to infection.

“It’s the junior doctors that have to look after the patients and do the basic prescribing. They’re still learning, but need to know which drugs work best and the hospital’s policy for prescribing antibiotics.”

Could a patient ‘dating agency’ not dissimilar to that suggested by Raymond, based on everyone’s medical records, help these junior doctors? “The doctor can search for other patients that look like their own. They can go back historically and see what drugs were prescribed and what their outcomes looked like.”

Drumright is mindful of setting up a system that tells doctors what to prescribe; the literature about how we interface with technology suggests that people can too easily surrender their responsibility. Instead, it’s about building on collective knowledge, “What we’re trying to do is enhance the doctor’s experience so that it’s not ‘my experience as me’, it’s the experience of every prescriber in the hospital.”

MINING FO ₵¤RRU₱

Researchers have developed a new technique that trawls the enormous amounts of public procurement data now available across the EU to highlight unscrupulous uses of public funds: from national and regional levels to individual contracts, companies and politicians.

The American economist Alan Greenspan once described corruption as “the way human nature functions”, it’s just that successful economies manage to keep it to a minimum. The question, of course, is how.

In the digital age, with its ‘freedom of information’, corrupt uses of public finance for political and corporate cronyism should have fewer dark corners to hide in.

Since the late 2000s, virtually all developed countries digitised and made available public procurement data. However, this data deluge can create the illusion of transparency, with a fog of information so vast as to seem impenetrable.

Previously, exposing corruption often relied on the diligence of journalists and campaigners to sift through data and make connections. Such investigations require time and luck, and can be biased.

But now a team of data-driven sociologists have created a new measurement system for detecting exploitation of public finance, designed to take advantage of the new data avalanche. It’s a system that is likely to rattle those profiting corruptly at the public’s expense (and give activists good cause to salivate).

The team defined key ‘red flags’: contractual situations that suggest high risks of corrupt behaviour. By unleashing ‘creeper’ algorithms and sophisticated text-mining programs on public procurement data to sniff these flags out, the team can map levels of corruption risk at regional and national scale, track corrupt behaviour in tendering

“Corruption is probably the number one complaint about people in power”

> **Using our methodology, institutionalised corruption can be measured right down to the level of individual contracts and tenders in about 50 countries around the globe**

organisations, and pinpoint suppliers and even individual contracts that look fishy.

The Corruption Risk Index (CRI) mines available information about expenditure of public finances for political collusion, competition rigging and crony capitalism, all with unrivalled speed and accuracy. Developed by Dr Mihály Fazekas and Professor Lawrence King from the Department of Sociology, it forms the basis of the Digital Whistleblower, or ‘DigiWhist’, led by Cambridge with a consortium of European institutes, and which has just secured €3 million of European Union (EU) Horizon 2020 funding.

“Corruption is probably the number one complaint about people in power, but there were no really objective ways to measure corruption,” explains King.

“Using our methodology, institutionalised corruption can be measured right down to the level of individual contracts and tenders in about 50 countries around the globe since 2008 to 2009 – opening up a whole universe of scientific and policy applications. We aim to make CRI available to citizens, civil society groups and journalists, to hold politicians and political parties accountable for corrupt behaviour.”

The project began when Fazekas had a brainwave while working on his PhD with King. In many developed nations since 2007, whenever the government purchased something over around €20,000 (or equivalent), the contract and tender data were made digitally available. In many countries, this is around 7% of the GDP – a big chunk of the economy.

Fazekas spoke to experts on public procurement to uncover the box of tricks often employed to fleece the public purse. Cannily, he also talked to companies who had fallen out of favour since their country’s government changed, “so they were happy to tell me how it was back in the day”. This work eventually led to the CRI’s 13 ‘red flags’ of corruption.

For example: very short tender periods (“if a tender is issued on a Friday and awarded on a Monday – red flag”); very specific or suspiciously complex tenders compared with the field (“like writing a job description for a role you want your friend to get”); tender modifications leading to bigger contracts; inaccessible tender documents; very few bidders in highly competitive markets. Different scales and combinations of flags allow researchers to create the risk rankings of the CRI.

Using an initial EU grant, the team conducted a proof of principle with data from Hungary, Slovakia and the Czech Republic. They found that firms with a higher CRI score made more money: the final contract value frequently came in much higher than the original estimate. These companies are also more likely to have politicians involved – either managing or owning them – and be registered in tax havens.

Over the next three years, the team aims to do this for procurement data across 34 European countries and the EU institutions, creating a corruption ranking that ranges from national to contract level. “Previous corruption indicators tended to be very blunt instruments. We can analyse regions and sectors but also individual organisations and loan officers. It’s an enormously powerful and fine-grained tool,” adds King.

The DigiWhist project will encompass four different data labs across Europe to collect and ‘clean’ data, and build databases. While their current mechanism has manual elements, the next version – developed by Dr Eiko Yoneki’s team in Cambridge’s Computer Laboratory – will have self-learning algorithms that recognise errors and link to existing solutions from the database. “After an initial teaching phase, it will kind of run on its own,” says Fazekas.

All their findings will be made publicly available, with downloadable databases that can be interrogated by academics, journalists and, indeed, anyone with an interest in what happens to public money and in holding businesses and political parties accountable for corrupt behaviour.

Fazekas believes their results could be married with public crowdsourcing to build a more complete picture of the consequences of siphoning public funds. “Imagine a mobile app containing local CRI data, and a street that’s in bad need of repair. You can find out when public funds were allocated, who to, how the contract was awarded, how the company ranks for corruption. Then you can take a photo of the damaged street and add it to the database, tagging contracts and companies,” says Fazekas, who is already working with DigiWhist advisors on prototypes.

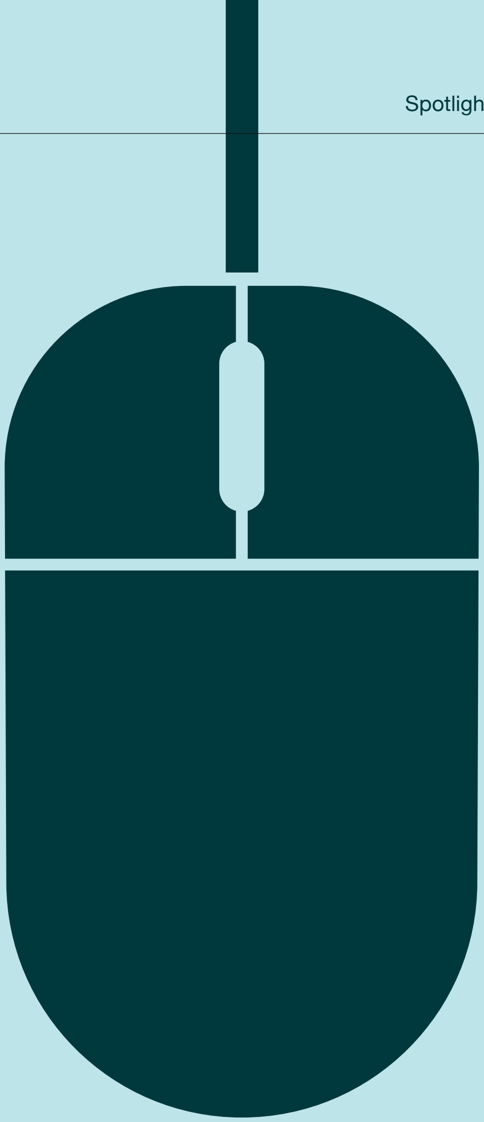
“The idea that the public are going to be able to interrogate this data on a very localised basis and contribute to it themselves through things like smartphone apps is a compelling one!” Fazekas adds.

For King, health will be a big focus. “One of the big debates is around deregulation and privatisation of health, and whether it increases efficiency. But does it increase corruption?

“There’s been a lot of talk of big data for a while now but not much has come out of it… By having researchers like Mihály, who straddle both tech and social science, I think we’ll start to see the potential for big data to turn into important findings that really do make the world better,” says King.

Left to right
Professor Lawrence King
lk285@cam.ac.uk
Dr Mihály Fazekas
mf436@cam.ac.uk
Department of Sociology

Computer Tutor



“Machines are good at dealing with routine things and large amounts of data… these tools can free up the teacher’s time to focus on actual teaching”

Millions of English language tests are taken each year by non-native English speakers. Researchers at Cambridge’s ALTA Institute are building ‘computer tutors’ to help learners prepare for the exam that could change their lives.

“We arrived to our destination and we looked each other.”

To a native English speaker, the mistakes in this sentence are clear. But someone learning English would need a teacher to point them out, explain the correct use of prepositions and check later that they have improved. All of which takes time.

Now imagine the learner was able to submit a few paragraphs of text online and, in a matter of seconds, receive an accurate grade, sentence-by-sentence feedback on its linguistic quality and useful suggestions for improvement.

This is Cambridge English Write & Improve – an online learning system, or ‘computer tutor’, to help English language learners – and it’s built on information from almost 65 million words gathered over a 20-year period from tests taken by real exam candidates speaking 148 different languages living in 217 different countries or territories.

Built by Professor Ted Briscoe’s team in Cambridge’s Computer Laboratory, it’s an example of a new kind of tool that uses natural language processing and machine learning to assess and give guidance on text it has never seen before, and to do this indistinguishably from a human examiner.

“About a billion people worldwide are studying English as a further language, with a projected peak in 2050 of about two billion,” says Briscoe. “There are 300 million people actively preparing for English exams at any one time. All of them will need multiple tests during this learning process.”

Language testing affects the lives of millions of people every year; a successful test result could open the door to jobs, further education and even countries.

But marking tests and giving individual feedback is one of the most time-consuming tasks that a teacher can face. Automating the process makes sense, says Dr Nick Saville, Director of Research and Validation at Cambridge Assessment.

“Humans are good teachers because they show understanding of people’s problems, but machines are good at dealing with routine things and large amounts of data, seeing patterns, and giving feedback that the teacher or the learner can use. These tools can free up the teacher's time to focus on actual teaching."

Cambridge Assessment, a not-for-profit part of the University, produces and marks English language tests taken by over five million people each year. Two years ago, they teamed up with Briscoe's team and Professor Mark Gales in the Department of Engineering and Dr Paula Buttery in the Department of Theoretical and Applied Linguistics to launch the Automated Language Teaching and Assessment (ALTA) Institute, directed by Briscoe. Their aim is to create tools to support learners of both written and spoken English.

Underpinning Write & Improve is information gleaned from a vast dataset of quality-scored text – the Cambridge Learner Corpus. Built by Cambridge University Press and Cambridge Assessment, this is the world's largest collection of exam papers taken by English language learners around the world.

Each test has been transcribed and information gathered about the learner's age, language and grade achieved. Crucially, all errors (grammar, spelling, misuse, word sequences, and so on) have been annotated so that a computer can process the natural language used by the learner.

Write & Improve works by supervised machine learning – having learnt from the Corpus of errors, it can make inferences about new unannotated data. Since its launch as a beta version in March 2014, the program has attracted over 20,000 repeat users. And each new piece of text it receives continues this process of learning and improving its accuracy, which is already running at almost equal to the most experienced human markers.

Briscoe believes that this sort of technology has the potential to change the landscape of teaching and assessment practices: "Textbooks are rapidly morphing into courseware where people can test their understanding as they go along. This fits with pedagogical frameworks in which the emphasis is on individual profiling of students and giving them tailored advice on what they can most usefully move onto next."

He regards the set-up of ALTA as the "best type" of technology transfer: "We do applied research and have a pipeline for transferring this to products. But that pipeline also produces data that feeds back into research."

The complex algorithms that underpin Write & Improve are being further developed and customised by iLexIR, a company Briscoe and others set up to convert university research into practical applications; and a new company, English Language iTutoring, has been created to deliver Write & Improve and similar web-based products via the cloud and to capture the data that will feed back into the R&D effort to improve the tutoring products.

Now, the researchers are looking beyond text to speech. Assessing spoken English brings a set of very different challenges to assessing written English. The technology needs to be able to cope with the complexities of the human voice: the rhythm, stress and intonation of speech, the uhms and ahhs, the pauses.

"The fact that you can get speech recognition on your phone tends to imply in some people's minds that speech recognition is solved," says Gales, Professor of Information Engineering. "But the technology still struggles with second language speech. We need to be able to assess the richness in people's spoken responses, including whether it's the correct expression of emotion or the development of an argument." Gales is developing new forms of machine learning, again using databases of examples of spoken English.

"The data-driven approach is the only way to create tools like these," adds Briscoe. "Building automated tests that use multiple choice is easy. The stuff we are doing is messy, and it's ever-changing. We've shown that if you train a system to this year's exam on data from 10 years ago the system is less accurate than if you train it on data from last year."

This is why, says Briscoe, it's unimaginable to reach a point where the machines have learned enough to understand and predict almost all of the typical mistakes learners make: "Language is a moving target. English is constantly being globalised; vocabulary changes; grammar evolves; and methods of assessment change as progress in pedagogy happen. I don't think there will ever be a point when we can say 'we are done now'."

I **Professor Ted Briscoe**
ejb@cl.cam.ac.uk
Computer Laboratory

Built on information from almost

65 million

words gathered over a 20-year period from tests taken by real exam candidates

speaking

148

different languages

living in

217

different countries or territories

Researchers are using social media data to build a picture of the personalities of millions, changing core ideas of how psychological profiling works. They say it could revolutionise employment and commerce, but the work must be done transparently.

In 2007, Dr David Stillwell built an application for an online networking site that was starting to explode: Facebook. His app, myPersonality, allowed users to complete a range of psychometric tests, get feedback on their scores and share it with friends. It went viral.

By 2012, more than six million people had completed the test, with many users allowing researchers access to their profile data. This huge database of psychological scores and social media information, including status updates, friendship networks and 'Likes', is the largest of its kind in existence. It contains the moods, musings and characteristics of millions – a holy grail of psychological data unthinkable until a few years ago.

Stillwell and colleagues at Cambridge's Psychometrics Centre provided open access to the database for other academics. Academic researchers from over 100 institutions globally now use it, producing 39 journal articles since 2011.

Meanwhile, the Cambridge Psychometrics team devised their own complex algorithms to read patterns in the data. Resulting publications caused media scrums, with a paper published in early 2015 generating nervous headlines around the world about computers knowing your personality better than your parents.

But how surprising is this really, given the amount we casually share about ourselves online every day? And not just through social media, but also through web browsing, internet purchases, and so on. Every interaction creates a trace, which all add up to a 'digital footprint' of who we are, what we do and how we feel.

We know that, behind closed doors, corporations and governments use this data to 'target' us – our online actions mark us out as future customers, or even possible terrorists – and, for many, this reduction in privacy is a disturbing fact of 21st-century life.

The Cambridge researchers believe that the new era of psychological 'big data' can be used to improve commercial and government services as well as furthering scientific research, but openness is essential.

"If you ask a company to make their data available for research, usually it will go to some corporate responsibility office which deems it too risky – there's nothing in it for them. Whereas if you tell them you can improve their business, but as part of that they make some data available to the research community, you find a lot more

How to read a digit footpr

It contains the moods, musings and characteristics of millions – a holy grail of psychological data unthinkable until a few years ago

open doors," says Stillwell, who co-directs the Centre.

Around half of the Centre's current work involves commercial companies, who come to them for "statistical expertise combined with psychological understanding" – often in an attempt to improve online marketing, an area still in its infancy.

The team has recently launched an interface called Apply Magic Sauce, based on the myPersonality results, which can be used as a marketing and research tool that turns digital 'footprints' into psycho-demographic profiles.

"If you use the internet you will be targeted by advertisers, but at the moment that targeting happens in the shadows and isn't particularly accurate," says Vesselin Popov, the Centre's development strategist.

"We all have to suffer advertising, so perhaps it's better to be recommended products that we might actually want? Using opt-in anonymous personality profiling based on digital records such as Facebook Likes or Last.fm scores could vastly improve targeted advertising and allow users to set the level of data-sharing they are comfortable with," says Popov. "This data could then, with the permission of users, be used to enrich scientific research databases."

Measuring psychological traits has long been difficult for researchers and boring for participants, usually involving laborious questionnaires. This will sound familiar to anyone who has used an employment agency or job centre. The team are now building on their previous work with algorithms to take psychometric testing even further into uncharted territory – video games. Job centres might be the first to benefit.

"A job centre gets about seven minutes with each job seeker every two weeks, so providing personalised support in that time is challenging," explains Stillwell. "We are working with a company to build a game that measures a person's strengths in a 'gamified' way that's engaging but still accurate."

In 'JobCity', currently an iPad proof of concept, users explore job opportunities in a simulated city. The game measures psychological strengths and weaknesses along the way, offering career suggestions at the end, and providing the job centre with feedback to help them guide the applicant. The team has tested the game with a group of under-25s and the results are promising.

For the Centre's Director Professor John Rust, the team's background in psychology means they don't lose sight of the people within the oceans of data: "We're dealing with organisations that are using 'big data' to make actuarial decisions about who gets lent money, who gets a job – you don't want this left solely to computer engineers who just see statistics."

"We want machines that can recognise you as a person. Much of the information for doing that already exists in the servers of Google, Facebook, Amazon, and so on. Your searches and statuses are all reflections of questions, experiences and emotions you have: all psychometric data. It's the basis for a future where computers can truly interact with human beings."

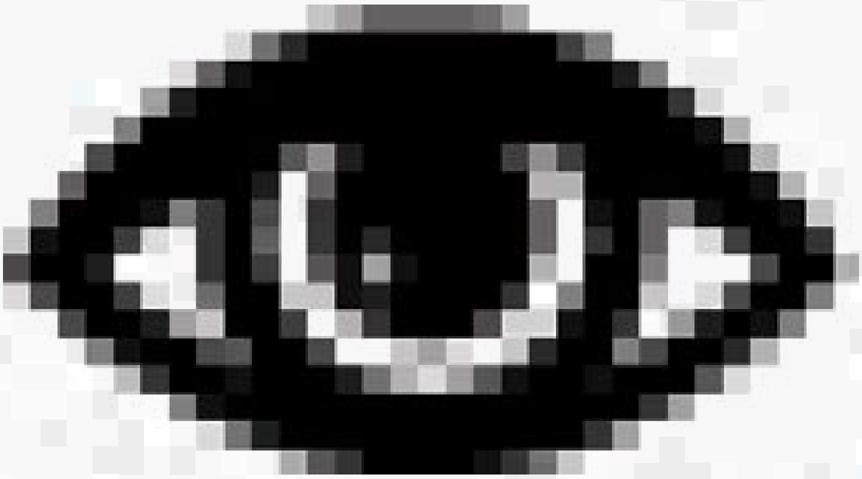
Cyberspace has, for Rust, opened a 'Pandora's box' that's taken psychological testing to a new level. But, he says, the current explosion in big data bears comparison to a previous shift that happened a century ago – the advent of IQ tests shortly before the First World War. Millions of servicemen were tested to determine role allocation within the military. Suddenly, says Rust, overexcited scientists had massive psychological datasets. IQ tests influenced societies long after the war, leading he says to some of the most shameful episodes of the 20th century including scientific racism and sterilisation of the 'feebleminded'.

"Today you have another psychological big data situation being used to challenge a perceived global threat: terrorism. Government data scientists hunting would-be terrorists are enthusiastically adopting big data, but there will be social consequences again. In many ways, we already have Big Brother – whatever that now means," Rust says.

"The new psychological data revolution needs serious research, and ethical debates about it need to be happening in the public arena – and they're not. We have a responsibility to say to people working on this in secret in companies and institutions: 'You've got to come and discuss this in an open place'. It's what universities are for."

I

Left to right
Dr David Stillwell
ds617@cam.ac.uk
Professor John Rust
jnr24@cam.ac.uk
Vesselin Popov
vp288@cam.ac.uk
The Psychometrics Centre
Department of Psychology



I always feel like somebody’s watching me…

“What we’re trying to do is develop processing frameworks that would allow this data to be useful and to be used, without the somewhat creepy feeling that you’re constantly being watched”

What power can individuals have over their data when their every move online is being tracked? Researchers at the Cambridge Computer Laboratory are building new systems that shift the power back to individual users, and could make personal data faster to access and at much lower cost.

It’s a fact of modern life – with every click, every tweet, every Facebook Like, we hand over information about ourselves to organisations who are desperate to know all of our secrets, in the hope that those secrets can be used to sell us something.

Companies have been collecting every possible scrap of information from their customers since long before the internet age, but with more powerful computers, cheaper storage and ubiquitous online use, the methods organisations use to gather information about people have become ever-more sophisticated. And sometimes those organisations know us better than our own families or friends.

For example, several years ago, data analysis tools used by the US retailer Target had become so precise that they were able to determine, with astonishing accuracy, whether a woman was pregnant and how far along she was, based on her purchase of certain products. And in one particularly embarrassing incident, Target knew that a teenage girl was pregnant before her father did, much to her father’s displeasure.

“What Target learned from that incident is that marketing too accurately can really make people squeamish,” says Professor Jon Crowcroft of the University’s Computer Laboratory. “But if they made their marketing a little less accurate by increasing the amount of privacy they give their customers, they found they can still retain or increase their customer base without making people feel as if they’re being spied on.”

Crowcroft's research is in the area of 'privacy by design' – systems that allow us to live in the digital world and protect our privacy at the same time. As the concept of the Internet of Things – internet-connected washing machines, toasters and televisions – becomes reality, Crowcroft insists that privacy by design is needed to address the massive power imbalance that occurs when our personal data is shared with, and sold by, corporations, governments and other organisations.

But privacy by design doesn't mean disconnecting from the online world and putting on a tinfoil hat – far from it. "There's already a lot of data stored about each and every one of us – the things we buy, the food we eat, the health issues we have – and for each of these market segments, there are perfectly legitimate uses for that data," adds Crowcroft. "Collecting healthcare data is fantastically useful for tracking pandemics, preventative care, more-efficient treatment, public health – those are all perfectly reasonable and positive uses for big data. At the same time, most sites gather information in order to target ads more accurately, and most people are actually okay with that. So the question then becomes, what is privacy by design?"

"What we're trying to do is develop processing frameworks that would allow this data to be useful and to be used, without the somewhat creepy feeling that you're constantly being watched," says Crowcroft's colleague Dr Richard Mortier.

The type of system that Crowcroft and Mortier envision is one in which the user has the scope to allow access to their data on a case-by-case basis, rather than it be harvested whether they like it or not: computations are performed where the data is gathered, and the results are pushed back to the organisation that wants the data.

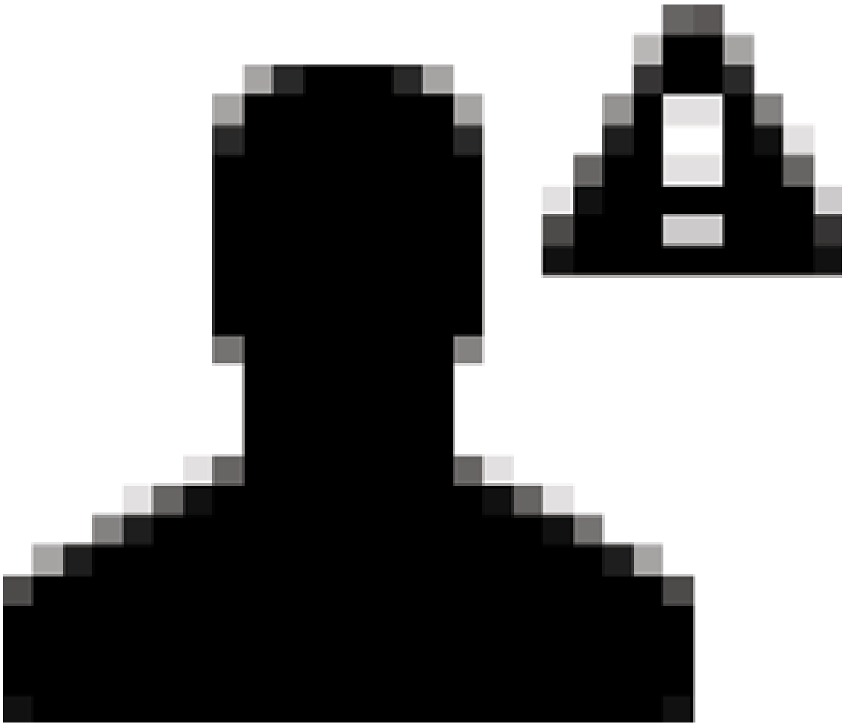
"We can change the big data problem completely by moving where the data is processed," explains Mortier. "Rather than having systems where all of the data is gathered in some huge central location and processed, if you reconstruct the system so that the data is processed in the same place it's gathered, individuals would be able to take some of the control of their information back from corporations and surveillance organisations. Instead of one huge central processing node, we want to see billions of smaller nodes, which would make information quicker to access, and could potentially be stored at lower overall cost."

Crowcroft and Mortier have designed and partially built systems where a person's data stays local to them, and they can have the option to decide what is shared and with whom. For example, a patient can share their healthcare data with their GP, but the GP would have to get authorisation from the patient before sharing that data with a pharmaceutical company.

"People realise they're being marketed to, but I don't think they realise the scale of it – it really is a hidden menace," says Crowcroft. "The point is that we could build systems that could stop that completely, and re-enable it on the basis of a level playing field. We want to see systems where people have agency over their data, giving them the ability to allow or prevent certain types of access."

Contrary to what some people may assume about the nature of digital life, adds Crowcroft, the vast majority of people highly value their own privacy. He points to the launch and then recall of Google Glass, a wearable computer worn like eyeglasses. "People started wearing these things into restaurants and other diners wouldn't put up with it, because they didn't want to be recorded while eating their lunch – it really creeped people out," he says. "And that's in a public space: imagine the same sort of thing happening in a private space. It's about the asymmetry and the idea that this is being done to you and you have no comeback. The problem with digital infrastructures is you don't see them, and to a certain extent companies depend on people not understanding them – we can build systems where there are mechanisms through which they can be understood."

Crowcroft and Mortier recognise that they'll never convince everyone to ditch cloud computing and switch to a decentralised system. But that isn't their goal. "It takes a while to show that new ways of doing things can really work," says Crowcroft. "If these sorts of systems become a reasonably widely used alternative, it will go a long way towards keeping companies and cloud storage providers honest. The very small number of providers leads to the exploitation of the network effect, where they have a strong monopolistic position over a certain type of data. And monopolies are not good for economies. If a decentralised system is more ethical, enough people using it may incentivise the big providers to be more ethical too."



Left to right
Professor Jon Crowcroft
Jon.Crowcroft@cl.cam.ac.uk
Dr Richard Mortier
richard.mortier@cl.cam.ac.uk
Computer Laboratory

Inside out

Tragedy in Nepal

Following the recent devastation in Nepal, Evan Miles reflects on Himalayan glaciology and the natural hazards faced by those living in an area to which he was about to return for fieldwork.

April 25, 2015. I awoke in Cambridge keen to finish preparations and packing for my fifth season of fieldwork in the Langtang Valley of Nepal. I was due to fly out the following day. I checked my email for last-minute updates, and was jerked out of any sense of routine by the title 'Big earthquake just hit Nepal'.

A barrage of Twitter, Facebook, news and media searches, Skype calls and inquiries lasted for the rest of the day as my colleagues and I began to realise the scale of the event. At first we were wondering if we could get to our field site but, in a Skype conference that evening, it became clear that our scientific work was out of the question. Our priorities shifted from scientific to humanitarian – trying to contact our friends and colleagues, and gather and share the scant information from beyond Kathmandu.

Geoscientists had highlighted the likelihood of an 8.0-magnitude earthquake for decades, and reports had specifically assessed the human and financial consequences of such a tremor. But Nepal is a very poor country – it struggles to pave roads connecting its widespread mountainous terrain and has no real means for emergency preparedness in the modern sense of the word. Earlier this year, a Turkish Airlines plane crash-landed in Kathmandu, and it was several days before the country's single international runway could be reopened to commercial flights.

As scientists, we have to acknowledge the risks of any study site, and had not overlooked the potential for an earthquake, or the severe avalanches and landslides that occur frequently in the mountains.

However, the Nepalese Himalaya also present a knowledge gap within glaciology: when the International Panel on Climate Change published its 4th Assessment Report in 2007, the community of glacier scientists had little understanding of Himalayan glaciers owing to remoteness, weather and altitude. It was therefore a priority for further research and the motivation for my PhD, working with Drs Ian Willis and Neil Arnold at the Scott Polar Research Institute, and collaborating with researchers in Switzerland, the Netherlands and Nepal.

> Geoscientists had highlighted the likelihood of an 8.0-magnitude earthquake for decades

The focus of our studies has been the Lirung Glacier in the upper Langtang Valley north of Kathmandu. It's a good example of the debris-covered glaciers that are common in High Mountain Asia, where avalanches and rockfalls deposit large amounts of rubble onto the ice.

The rubble accumulates at the surface as the ice melts, forming a blanket comprising sand, gravel, cobbles and boulders, which substantially alters how the glacier interacts with the atmosphere. This type of glacier is much less understood than its clean-surface counterparts.

Our studies on Lirung Glacier are aimed at resolving a conundrum about debris-covered glaciers that can help us understand how these glaciers will respond to climate change.

The thick debris should reduce the melt of glacier ice, but many debris-covered glaciers seem to be melting much faster than expected – nearly on par with clean-ice glaciers. We think this is due to the presence of exceptional surface features – bare ice-cliffs and small lakes on top of the glacier – that, although covering only a fraction of the glacier's surface, appear to melt much faster than the surface under the debris layer.

After years of detailed observations, our team is developing numerical models of ice-cliffs and ponds, which do look to be partly responsible for the high rates of ice loss from High Mountain Asia's glaciers. This is an important step to understanding the region's response to climate change, as they are not yet accounted for in projections of glacier melt in a future climate.

Now, though, science has to take a back seat. It's unclear how or when our observations will continue, but the earthquake is not simply another obstacle for our research. During four field seasons, we've built relationships with Nepali villagers along the Langtang Valley and scientists in Kathmandu. The destruction in Kathmandu is terrible – large numbers of casualties, World Heritage Sites destroyed – but reports suggest outlying villages have fared even worse, with few buildings having withstood the tremors, and devastating avalanches and landslides widespread. The Langtang Valley is no exception, as several villages appear to have been wiped out entirely by landslides burying hundreds of villagers and a long Tibetan heritage.

We wonder about our instruments, but we are much more concerned about the villagers we've got to know. For the present, we are trying to map landslides, to prioritise for immediate rescue operations and then eventual rebuilding.

Evan is funded by a Gates Cambridge PhD Studentship, the University's Fieldwork Fund, Trinity College's Rouse Ball and Eddington Fund, the Department of Geography's Philip Lake and William Vaughan Lewis Fund, and the Scott Polar Research Institute B.B. Roberts Fund

I Evan Miles
esm40@cam.ac.uk
Scott Polar Research Institute
Department of Geography

Data Mining and Analysis
Analysis of Multivariate and High-Dimensional Data
Mining of Massive Datasets
Privacy, Big Data, and the Public Good

An Introduction to Galaxies and Cosmology
SECOND EDITION
An Introduction to the Sun and Stars
SECOND EDITION
interstellarum Deep Sky Atlas
DESK EDITION
Atlas of Great Comets

T +44 (0)1223 765 443
E research.horizons@admin.cam.ac.uk
W cam.ac.uk/research
f facebook.com/cambridge.university
twitter.com/cambridge_uni
youtube.com/cambridgeuniversity
instagram.com/cambridgeuniversity

Contact
Research Horizons
Office of External Affairs
and Communications
The Pitt Building, Trumpington Street
Cambridge, CB2 1RP

Cover
Big data ‘dating agencies’ are being used to match patients who have rare diseases worldwide to help clinicians diagnose and treat them; find out more on p. 24 this issue.