# Data Science at the University of Oxford

**Draft version 3.0**

**20th October 2014**

# Contents

**Overview** 2

1. **Underpinning Tools, Methods and Technology** 3
2. **Society and Public Policy** 8
3. **Finance** 15
4. **Business** 17
5. **Biomedical and Life Science** 18
6. **Climate and the Environment** 22
7. **Astronomy** 26
8. **Physical Sciences** 27
9. **Arts and Humanities** 28
10. **Education** 31
11. **Facilities and Infrastructure** 32
12. **Innovations in IP Management** 33

**Directory of Researchers** 35

## Overview

In almost every sector of commercial and public endeavour there has been or will be a data deluge. Unparalleled opportunities for innovation and exploitation are driven by the availability of data from emerging and converging digital platforms and their explosive uptake by the public. These new and growing forms of data include:

- data from expanded digital access and empowerment of individuals world-wide
- ever more digital traffic of all kinds, including social interactions and commercial transactions
- the production of novel kinds of scientific data, from earth observation, control systems, networks, and high throughput biomedical science
- the growth of open data initiatives
- corporate imperatives to create greater value from existing customers and to explore disruptive business models.

Data Science is changing the way we think about important problems such as intelligence, complexity, and our universe. The problems associated with managing data and deriving value from them have created an explosion of new ideas in the fields of statistics, computer science, mathematics, engineering, and across the social sciences. Crucially, new branches of knowledge, likely to have a profound impact on humanity, are emerging that combine and draw deeply from these distinct disciplines.

The ability of companies and institutions to analyse and exploit their data to the mutual benefit of customers and stakeholders will provide a powerful lever for future innovation, competitiveness, and growth within the global digital economy. The necessary skills available to achieve this will be at a premium. As the world becomes ever better connected digitally, the level of international competition and innovation will intensify. The United Kingdom must therefore build on its strengths of acknowledged excellence across science, engineering, medicine, and the social sciences to provide thought leadership and game-changing innovation in data sciences and technology. The creation of methods to capture, analyse and exploit data intelligently through the efficient distillation of actionable insights urgently requires a concerted effort at many interacting levels. It needs to deliver an imaginative growth in the development of concepts, algorithms and technologies, and successfully translate them into operational environments.

The breadth and strength of Oxford University's data research makes it extremely well-placed to address these challenges. All four academic divisions – Humanities, Mathematical, Physical and Life Sciences, Medical Sciences and Social Sciences – have a keen interest and involvement in data science: from building fundamental data handling tools and applying the techniques of machine learning and high-dimensional statistics to extract meaning from large, heterogeneous sources, to examining the social and ethical issues created by the growth of big data. Much of our research is interdisciplinary and collaborative, and research agendas are often determined in partnership with key stakeholders in business, government and social enterprise, ensuring that the research helps to answer the pressing questions of society, industry and the wider world. Oxford's engagement with all aspects of data science means that it is uniquely positioned to play a leading role in responding to opportunities in this area.

# 1. Underpinning Tools, Methods and Technology

Critical to successful Data Science are the core issues of efficient and intelligent processing, reliable storage, and management of data. Data alone are meaningless - it is these mathematical, statistical, computer science and information engineering tools which enable data to be organised and transformed into knowledge, and which allow the integration and coordination of overall systems. As the volume of available data continues to grow, and as increasingly varied types of data need to be collected and analysed, new technologies need to be developed to deal with them.

## a. Intelligent processing: Machine learning

Machine Learning (ML), the construction and study of systems that can learn from and act upon data rather than merely follow programmed instructions, is one of the fastest growing areas of science and has profoundly impacted and changed our technological world. It is largely responsible for the rise of giant data companies such as Google, and it has been central to the development of lucrative products and technologies that are now becoming mainstream and commonplace. Spam detection, advertising engines, image and video search, credit card fraud detection and news personalisation are all made possible through ML. The potential future applications of ML are vast and include the use of automatically-driven cars, more efficient energy management systems, and improved systems of health care management.

A machine learning class of algorithms called Ensemble Classifiers are the driving technology behind Microsoft's Kinect sensor, and also drive recommendation and personalisation in online magazines, apps and websites such as Amazon. Face detectors, taken for granted by users with cameras and mobile phones, are built with ensemble classifiers, and the same detectors are used to identify abnormalities in medical imaging. Another aspect of ML, Deep Learning Methods, promises even deeper impact. All state-of-the-art speech recognition systems, including those of Microsoft and Google, rely on deep learning; improvements are so great that many researchers now believe the speech recognition problem has been solved. Deep learning also drives semantic image and video search at all the big search engines. The economic impact in terms of copyright infringement, advertising and marketing is vast.

The rise of big data makes machine learning systems essential. Our ability to acquire and store data has surpassed our ability to understand them; individuals and organisations are swamped with a deluge of data, which creates enormous research and business opportunities, but also new threats. It impacts medicine and healthcare, our understanding of mind and intelligence, supply-chain management, privacy, public policy, and our capacity to improve energy usage and manage our natural resources. With this flood of data, the need for statistical machine learning has never been greater.

Machine learning and statistical data analysis research at Oxford brings together groups of researchers from across Engineering, Computer Science and Statistics, many of whom have long-standing collaborations, to exchange ideas, network, and provide a forum for the coordination of research developments across the University in this fast-moving domain. Machine Learning in Oxford is particularly strong in: Bayesian optimisation, the use of intelligent algorithms to perform

complicated design tasks (de Freitas, Osborne). The aim is to nourish both industry and other academic disciplines which are drowning in data, but starved of knowledge. In the medium to longer term the intention is to establish a Data Analytics hub which will form the epicentre of expertise in the UK and beyond in data-driven statistical modelling, big data and data analysis applications.

### b. Efficient processing: High performance data analysis

Improved hardware is also critical to the progress of data science. Over the last decade computing technologies have undergone dramatic change, from single core monolithic Central Processing Units (CPUs) to multi-core CPUs, which are now assisted by many-core Graphics Processing Units (GPUs) or co-processors such as Intel's Xeon Phi. Heading into the future, the landscape of High Performance Computing (HPC) solutions widens: ultra-low power devices to enable the gathering of large volumes of data at very low energy costs, easily-accessible Field Programmable Gate Arrays (FPGAs) to allow for low power supercomputing in inhospitable environments, and closely coupled CPU and GPU units which will bring the power of supercomputing a decade ago into the palm of your hand today.

These different types of HPC hardware address the processing challenges of one or more of the four V's of big data processing (Volume, Velocity, Variety and Veracity) and lead to the possibility of increasingly rapid analysis of large and complex data sets. A vital technique is to use these cutting-edge HPC technologies to enable real-time data processing and reduction of big data in situations where the volume of data produced is too large to be stored, allowing for interesting data to be retained and data containing no meaningful information to be discarded. Often the key to efficient computation on this scale relies on the ability to factor the problem into subsets which may be computed in parallel. Many-core devices such as GPUs, Xeon Phi and FPGAs have a highly parallel structure that makes them especially effective for such tasks.

Recent work by researchers in Oxford (Giles) has begun to show how speedups of many orders of magnitude may be achieved by judicious use of GPUs, and furthermore how very large arrays of GPUs may be able to rapidly perform analysis on terabytes of data. This opens the way for ultra large scale data analysis of complex, real-world problem sets without compromise. The applications are wide ranging, extending to any discipline that advances through the generation and analysis of vast datasets. Other work (Armour) focuses on using many-core technologies such as GPUs, FPGAs and Xeon Phi to enable real-time data processing of big data. The upcoming international Square Kilometre Array radio telescope project has been used as the problem domain for testing the feasibility of such technologies to perform this task. This work has begun to show how many-core technologies are extremely well suited to big data processing in inhospitable environments due to their ability to achieve high GFLOPs/Watt ratios when used correctly.

### c. Data management in the age of big data

Data is everywhere, generated by increasing numbers of applications, devices and users. The growth in the number and diversity of data sources is compounded by an increase in the scale of the data. However, for many user communities and applications, the idea that the data relevant to a task resides in a well maintained database or data warehouse is no longer tenable. Data typically comes from constantly evolving, heterogeneous, and unreliable sources.

The challenges that face data management in the current era of increasingly rapid and widespread data production and publication are referred to as the four V's of big data, namely Volume - the scale of the data, Velocity - speed of change, Variety - different forms of data, and Veracity - uncertainty of data. These helpfully indicate that although size matters, it isn't everything. Although the untapped value of big data is considered to be huge, and the growth in companies in the associated marketplace is rapid, a substantial fraction (sometimes cited as up to 80%) of the time of data scientists is devoted to data wrangling - the process of data identification, reorganisation and cleaning that allows meaningful analysis to begin.

Current research in the Oxford Database Group (Olteanu) seeks to enable systematic, efficient and effective use of data in domains that manifest several or all of the four V's. A few examples of active research projects follow.

- The DIADEM system provides intelligent domain-specific extraction of high-value structured data from a large set of web pages of semi-structured text. Further effort is on extracting heterogeneous data behind web forms.
- MayBMS/SPROUT/ENFrame are scalable systems that allow complex processing (querying and data mining) on uncertain, incomplete, or inconsistent data. The datalog meta-engine in the LogicBlox commercial database system tackles the aspect of velocity and investigates how to efficiently and incrementally maintain the result of large declarative code in the face of continuous and fast-pacing changes to the code itself and to the input data.
- The FDB project puts forward relational data compression that can be used in very large-scale commercial distributed database systems like Google F1 to reduce network communication cost and improve response time for complex queries and Google Ad analytics.

### d. Numerical analysis

Numerical analysis concerns the development of algorithms for solving all kinds of mathematical problems. It is a wide-ranging discipline having close connections with computer science, mathematics, engineering, and the other sciences. Big data analytics has changed the nature of the algorithms needed to be able to extract meaning from large distributed data sets. Within numerical analysis there are four research strands which are especially relevant to data science and analytics:

- numerical linear algebra, used for example in online recommending systems where there is a dependency on incomplete information
- numerical optimisation (mathematical programming), which deals with problems of minimising or maximising functions of finitely or infinitely many variables, the central task in nearly all operations research questions
- single value decomposition of distributed data: modern computational approaches, such as those of Amazon, deal specifically with the distributed nature of the available computational resources
- information and data compression, focusing on the design, analysis, and application of numerical algorithms for information-inspired applications in signal and image processing.

Oxford's Numerical Analysis Group has long been a leader within the UK, and its expertise in these areas enables the solution of many of the problems inherent in modern-day data analysis.

### e. Visualisation and visual analytics

Our ability to observe data effectively and efficiently is a primary tool which enables us to understand data, to extract information from it, and to transform it into knowledge. Regardless of how other advanced technologies may help us in dealing with large volumes of data, without initial and regular observation of data samples such technologies cannot be adaptively deployed in a context, objectively validated, dynamically evaluated, or continuously improved. Visualisation techniques play an indispensable role in helping us to observe data.

Visualisation (or more precisely, computer-supported data visualisation) is the study of transformation from data to visual representations in order to facilitate effective and efficient cognitive processes in performing tasks involving data. Data visualisation began with the invention of the line graph in the 10$^{th}$ or 11$^{th}$ century and progressed through the establishment of statistical graphics in the 18$^{th}$ and 19$^{th}$ centuries to become a ubiquitous technique in science and engineering. Today, however, computer-supported visualisation techniques have advanced much further than classical statistical graphics, and can handle a very wide variety of data – textual, tabular, network, geometric, image-based, geographical, and temporal. For example, parallel coordinate visualisations often handle multivariate data with tens and sometimes hundreds of variables. Field-based techniques are routinely used in medicine, science and engineering to depict complex scalar, vector and tensor fields.

Visual analytics is a new subfield of visualisation. It rests on four assertions: (i) Statistical methods alone cannot convey an adequate amount of information for people to make informed decisions. (ii) Algorithms alone cannot encode an adequate amount of human knowledge about relevant concepts, facts, and contexts. (iii) Visualisation alone cannot effectively manage levels of details about the data or prioritise different information in the data. (iv) Direct interaction with data alone is not scalable to the amount of data available. Hence we need equip ourselves with statistics, algorithms, visualisation and interaction in an integrated manner in order to extract meaningful information from large and complex data sets.

Oxford is at the forefront in several areas of visualisation and visual analytics research (Chen). It is an international leader in developing theories of visualisation, which bring mathematics and cognitive sciences together to explain why and how visualisation works. It is also internationally-leading in video visualisation, a collection of innovative techniques for summarising videos into images for rapid processing. It creates new concepts (e.g. four levels of visualisation), defines new methodologies (e.g. systematising glyph designs), sets new records (e.g. encoding 20 dimensions visually), and offers new visualisation solutions to different disciplines (e.g. analysing cell videos, phonetic dynamics in poems, work flows in biological experiments, threats in cybersecurity, energy consumption by software, pathological signatures in medical imaging, or events in sports).

### f. Developing search engines with data-extraction capabilities

The proliferation of data-intensive websites means that users are faced with countless sources of data whose interpretation is clearly beyond human processing capabilities. These data are mostly hidden behind web interfaces and appear to the users in semi-structured (e.g. HTML) or

unstructured (e.g. PDF) formats. Researchers at Oxford (Horrocks) aim to develop a semantic and context-aware search and annotation engine with data-extraction capabilities, which uses domain-knowledge and context models to analyse and annotate web documents and extract the relevant data – an issue of paramount importance to decision makers.

### g. Information security and privacy

The growth of connected services and their collection of data has given rise to a growing number of adversaries whose objectives include invasions of privacy, the prevention of legitimate processing, theft, fraud, and coercion, as well as high-grade threats to society and national security. Techniques to combat these threats draw upon longstanding research in communications security - and also upon careful design so that systems and data collection are security-positive or privacy-positive from their inception. Insight into these technical measures needs to be matched with understanding of human behaviour, risk appetite, and political, social, and economic factors: such an inter-disciplinary approach lies at the heart of Oxford's Cyber Security Centre.

Recent work in Oxford has explored both how to control adequately access to large datasets - so that, for example, privacy can be preserved in the presence of successive queries, via selective disclosure and evolving access control. Early applications of these approaches were seen in the management of large collections of medical data. Other work is successfully exploring how new commodity hardware technologies can be used to enhance the trustworthiness of data processing in cloud and other contexts - so that, for example, the data owner can pass the processing to a third party without risk of its accidental disclosure, or so that two mutually-distrusting parties can arrange for a third safely to process data on their joint behalves.

Other work explores security inferences which can be drawn from examining large datasets - for example in the analysis of all data crossing a network, in order to build a visualisation of its present state, or to discover rogue actors or 'insider threats'. Tracking normal and anomalous behaviour - in a network or in the mass of interactions an individual makes with a system (type, touch, gaze) - is gaining prominence as a means of ensuring 'continuous authentication' (in contrast to one-off login authentication of users), and Oxford researchers are making leading contributions in this, through applying machine learning techniques to such interaction data.

### h. Theoretical foundation of data science

While classical mathematical knowledge in disciplines such as probability, statistics, information theory and game theory will continue to underpin data science, there is an urgent need to enhance the theoretical foundation of data science, so that it can explain and model various data-intensive phenomena. Oxford is perhaps best placed to lead such a development on the international stage. The collective capacity of Oxford is reflected in its world leading research in probability and statistics (Mathematics), information theory in communication (Engineering), information theory in visualisation (Oxford e-Research Centre), algorithms and quantum information theory (Computer Science), and philosophy of information (Oxford Internet Institute). Combining all this expertise makes Oxford a highly productive theoretical think tank for data science.

## 2. Society and Public Policy

As people conduct more and more of their lives in digital settings, they leave a growing volume of digital trails which can be harvested to generate big data of a kind previously unavailable to social research. Data science is increasingly important to understand individual behaviour, social relationships and societal trends, with far-reaching implications across the social sciences in terms of the type of research questions that can be asked and the methodologies for answering them. There is a huge potential to apply natural science models and concepts to social behaviours, leading to wide-ranging benefits for civil society and public good – but also posing challenges to social science research, in terms of the multi-disciplinary expertise required to study social behaviour. In addition the internet and social media are themselves important subjects for social research, since they can tell us a great deal about how society is changing in response to technological development. The insights that data science methods can generate feed directly into policy-making, provide an extremely valuable source of information for businesses and governments, and informing planning and policy decisions across sectors.

Oxford's expertise in the area of data in the social sciences is reflected in the Oxford Internet Institute, a multidisciplinary social science department wholly dedicated to the study of life online. It has a concentration of expertise in working with real-world large-scale data and new algorithms from the fields of computer science and physics to tackle social science questions, and works in close collaboration not just with all other social science departments at Oxford but with researchers from other disciplines such as mathematics, and engineering. The wide range of departments and university institutes involved indicate that this is a flourishing, dynamic area of research in Oxford, encompassing a wide range of internal collaborations as well as creative partnerships with external organisations and companies.

### a. Understanding individual and social behaviour

Widespread use of internet- and phone-based technologies has changed the way we live, work and even think over the past two decades. The data we generate as we go about our daily lives are now key to understanding peer-to-peer interactions and social and political behaviour. As well as the digital trails left by online activities, the explosion of peer-to-peer digital communication via social media, email, phone and texts means that every event at a personal, local, national or international level now receives an unprecedented amount of reaction and comment from individual people. Reaction is often virtually instantaneous and can lead to a rapid cascade in communications which may significantly affect trends and attitudes on a national or even worldwide scale.

Oxford social scientists are developing new techniques for investigating individual behaviour in a range of digital environments, and investigating the influence that different platform designs have on that behaviour. The scope of this research is very wide and includes work on:

- Micro-labour markets, such as internet-based crowdsourcing marketplaces (Lehdonvirta)
- Understanding social interactions in virtual currencies such as Bitcoin by triangulating transactional data with other data from the social web (Lehdonvirta, Lyons)
- Understanding behaviour in online social networking sites, including the relationship between anonymity and uncivil, rude and aggressive behaviour on social media (Hogan)

- Interactions in crowd-like online learning environments (such as MOOCs, online distance learning courses aimed at unlimited participation and open access) and their effect on learning interactions and outcomes (Eynon)
- The effect of controls and platform design on aggressive behaviour in gaming environments (Przybylski)

In the area of politics and government, an Oxford research team (Margetts) examines political behaviour and government-citizen interactions, generating large-scale transactional data from social media and government sites. Online experiments are used to test the effect of different information environments on civic engagement and political participation, feeding back into the design of government online services and public policy. The project has accumulated a 'big' dataset of all signatures to all petitions launched on the UK and US government petitions platform from 2010, developed models of the growth rates and dissemination networks of petitions on social media, and analysed natural experiments around changes in website design. This is feeding directly into the design of petitions platforms in the UK (through collaboration with the Cabinet Office and House of Commons), Canada and the US.

Another key research project (Reed-Tsochas) explores the impact that big data generated about human behaviour (from social media, mobile phone usage, and other transactional data) has on social science, at a time of increasing computational power and novel computational techniques often imported from other disciplines. This provides the opportunity for a more scientific exploration of individual and collective human behaviour, with implications for a broad range of business and other activities. Research themes include investigations into how dependent we have become on modern technologies such as mobile phones, and whether social networking sites are significantly changing the nature of our relationships with other people. The project brings together an international team of experts to examine the impact of online social networking at individual, group and societal levels, with the aim of discovering whether our digital social life enhances, replaces or threatens face-to-face relationships.

Analysis of digital peer-to-peer traffic can provide a wealth of information that helps us to understand how trends are formed, who influences them and how they impact on the 'real world'. High volumes of communication during times of national emergencies such as flooding can be especially difficult for organisations to deal with. Oxford researchers (Grindrod) have a long-term collaboration with Strathclyde University and many digital media marketing companies, designing and deploying algorithms that can identify separate thematic conversations between transient groupings of individuals and locate key influencers within such conversations, in order to provide high-resolution analytics during communication avalanches. This is particularly effective as the public reacts to both planned and spontaneous events, and it can also help to channel response resources immediately to where they are needed. The research has led to a number of high impact applications and opportunities in partnership with UK based advertising/media agencies. For example, Bloom Agency in Leeds has developed the 'Whisper' service that deploys these methods over data such as the Twitter feed, enabling Bloom to grow their analytics team and provide novel services for large corporate clients such as ITV, Sky, Virgin Atlantic, SSE plc, and UKPN plc. Oxford researchers also advise other UK agencies on analytics, models and forecasting consumers' behaviour and decisions on digital platforms. Recently this research has led to novel services for the

energy sector using real time analytics of Twitter data to track power cuts and to keep the public better informed, and analytics tracking rumours regarding public companies that may indicate real or perceived issues that will drive share price movements.

Behavioural insights provided by big data can be so complex that visualisation and mapping are an essential part of the narrative. Oxford researchers have developed a range of tools to visualise and map big data in a wide variety of settings, including:

- Research into the geography of the internet, mapping the production and consumption of the world's knowledge, including the European Research Council funded project GeoNet: Internet Geographies (Graham).
- Taking a critical look at data visualisation in the context of migration, exploring how far data visualisation can provide impartial, evidence-based analysis of data on migration and migrants in the UK in order to inform media, public and policy debates (W Allen).

The commercial interest in having novel analytics research done over their own proprietary data, most common in customer facing industries, has led Grindrod and co-researchers at Strathclyde to develop an engagement model that allows for confidential research access to data, whilst the commercial partners also produce publishable versions of data once new methods have been shown to be successful. The commercial partners have rights to exploit the methods and academic staff put some effort into translational activities, in return for some downstream transparency. To date this has attracted support from BT, dunnhumby/Tesco, Capita, PA Consulting, Lloyds Banking Group, SingTel, Lockheed Martin, Unilever, and FirstGroup.

## b. Societal change

The data generated by domestication of digital technologies in everyday life is bringing new ways to analyse and understand social change, enabling researchers to address many complex questions which have previously been difficult to study.

In the field of **health sociology**, a recently awarded ERC-funded project at Oxford (Mills) will be the first ever to engage in a comprehensive study of the role of genes and gene environment interaction on reproductive behaviour. Until now, social science research has primarily focused on sociological and social science explanations to understand fertility outcomes, largely ignoring the role of biology and genetics. Due to unprecedented advances in molecular genetics over the last decades, for the first time it is possible to examine whether there is a genetic component to reproductive outcomes such as age at births, number of children and infertility. The project will focus on examining fertility outcomes in relation to classic social science determinants, but also genetic and lifestyle factors (e.g., smoking, stress, BMI). This transdisciplinary project draws upon research within sociology, demography, molecular genetics and medical research to examine gene and environment interaction, new types of causality and aims to produce fundamentally new results. In a related field, researchers are using big data and modelling to develop, evaluate and refine methodologies that can assess the effects of social, economic and health policies on the pattern and magnitude of health inequalities among socioeconomic groups (Stuckler).

A key area of social change that can be understood through data science is **migration**. The mobility of people is now firmly recognised as a key dimension shaping society today, but the relationship between migration and societal change is only partly understood. Researchers at Oxford conduct research covering a spectrum of global migration processes and phenomena, and in particular have developed expertise in relation to migration and the labour market, and migration and urban change. The Migration Observatory, which involves experts from a wide range of disciplines and departments at the University of Oxford, provides impartial, independent, authoritative, evidence-based analysis of data on migration and migrants in the UK, to inform media, public and policy debates, and to generate high quality research on international migration and public policy issues. There are a number of extant projects that utilise big data techniques to research and better understand aspects of migration and its impacts upon society. For example, a long-term project (W Allen) investigates how UK newspapers portray immigrants and refugees, and this has necessitated the design, building and management of a large textual dataset of migration-related news coverage, totalling over 43 million words and 58,000 items.

**Risk** is another area that is affected by use of digital technologies, posing new challenges for evaluation, which in turn may be tackled through data science techniques. Big data allows near real-time prospective (rather than retrospective) risk analysis; Oxford researchers (McSharry) use both transactional data and data harnessed from social media to investigate the relationship between the social and natural environment and individual outcomes, and to monitor and forecast real-world events. The research takes a multi-model adaptive approach using sophisticated machine learning algorithms to generate multiple plausible scenarios for decision-making, which can be used to generate predictive models that aid our understanding of risk, and develop early warning systems for extreme events, from natural disasters to extreme social events such as riots or healthcare crises caused by high demand.

As with technologies such as stem cell research or cloning, developments in data and computing often become available before there is any opportunity to properly assess their impact on society. One area of concern, for example, is that inexpensive and ubiquitous sensors could enable an 'Internet of things' in which all kinds of ordinary physical objects will be connected to the Internet and constantly in communication with one another, raising many issues related to privacy, vulnerability to hacking, opportunities for new types of scientific data gathering and collaborative information filtering. As machine learning advances and artificial intelligence develops, there is a small but nevertheless real possibility of machines evolving beyond the ability of humans to control them. In addition there is a need to question the assumptions underlying 'big data' itself: Is it always a 'force for good'? Does it elevate some forms of knowledge and information above others? Does it champion some versions of society and undermine others? Who has access to it, and who is excluded? A critique of the way we define and make use of data is essential in any thinking society.

Oxford's 'Future Impacts of Technology' programme aims to investigate basic questions of the predictability of technological change and the knowability of latent systemic risks that might accumulate as a result of gradual and seemingly innocuous change. The project also investigates the challenges of wise policy formulation under conditions of unpredictable, rapid, and potentially disruptive technological change. This overall subject area is fast becoming a resource for research in a variety of disciplines with direct application to policy-making, particularly as industry and

government invest greater amounts in new data science tools and methodologies. For example, a collaboration between Engineering and the Oxford Martin School (Frey, Osborne) led to research suggesting that half of the US workforce could be at risk of computerisation, with huge potential implications for society. The dangers inherent in artificial intelligence are also considered under this programme. Additionally, key Oxford researchers collaborate on cybersecurity research through the Cybersecurity Centre (Brown, Creese) in order to keep up-to-date with this rapidly developing field.

Work by Oxford economics researchers (Hendry) recognizes that big data offer many potential benefits but also a range of problems, including an excess of false positives, mistaking correlations for causes, ignoring sampling biases, and analysing by inappropriate methods. Two Oxford projects ('Big Talk about Big Data' and 'Big Data, Big Visions') aim to interrogate the values underlying the use of 'big data' as a concept, and show how and to what extent British civil society organisations harness data to achieve their missions. They take a critical look at how civil society organisations increasingly demand research that is 'data-driven' or 'evidence-based', with the intention of documenting the extent to which perceived advantages of data 'bigness' (volume, variety, and velocity) influence these demands. The research will examine the particular meanings that are attached to concepts like 'Big Data', 'data-driven research', and 'evidence' in the context of British politics and policy, and determine what kinds of civil societies are being promoted or demoted as a result.

### c. Informing public policy

As data science generates insights into individual behaviour and social change, it can feed directly into the design of public policy and government services, aligning them more closely with citizens' needs and actual behaviour and societal trends. There is widespread interest across government in what data science can do to make public policy better and to guide decisions. There is also recognition of a real need for capacity building, in terms of research methods, tools and algorithms - but also data science education and training. Oxford researchers are working with the nascent data science group in the Cabinet Office and Government Digital Service to work out how to build that capacity. The project 'Urban Data 2 Decide: Integrated Data Visualisation and Decision Making Solutions to Forecast and Manage Complex Urban Challenges' (Bright) is one example of using innovative data science methods to solve policy problems; a collaboration between research institutions in Austria, Sweden and Denmark, along with city councils in Oxford, Manchester and Copenhagen, the project explores ways of utilising data from social media and open data libraries to aid urban decision-making. This work has involved a range of stakeholder groups such as strategists for digital government, policymakers in data governance, and NGOs using big data in developing world and disaster relief contexts.

Work underway at Oxford in specific policy sectors includes:

**Education:** the Government Department for Education in England has developed the National Pupil Database (NPD), one of the largest and most detailed national datasets based on student-level data. For over 15 years the NPD has been collating the records of over six million students aged 4-16. The data cumulates every term with a Unique Pupil Identifier so that students can be followed over their school careers. There are thousands of data items including demographic data on students' age,

gender, ethnicity, special educational needs and socio-economic circumstances, together with the results of multiple educational assessments, and data on attendance or exclusion from school. Matched to this is a database on the characteristics of 30,000 England schools and a growing database on teachers and other members of the schools workforce. Links to Post-16 Learner records and higher education datasets are also being created. Researchers at Oxford (Strand, Sammons, Melhuish) have been at the forefront in applying quantitative techniques for analysing large and complex datasets (such as multi-level modelling) to the NPD to address important social science questions such as the relationship between student background and educational progress, the size and stability of school effects on student outcomes, and the impact of different kinds of pre-school experience on long term outcomes.

**Energy:** Realising potential efficiency gains in energy requires a better understanding of how energy users (individual and corporate) behave, which factors affect their energy use and how energy suppliers can manage supply in the most efficient way. As technology evolves, electricity distribution network operators (DNOs) must shift from being merely network infrastructure providers to become the 'traffic controller and enabler' of many different types of future activity on the networks. Smart meters have a crucial role to play here, not only by offering the consumer a way of monitoring and controlling their own energy usage, but also by collecting data that can give electricity providers vital information about customer usage and demand, thus informing their planning. Researchers at Oxford (Grindrod) are working together with Scottish and Southern Energy Power Distribution (the local DNO) and a range of academic, local government and business partners to develop smart meter analytics that will deliver improved performance for the electricity sector. Key aspects of the research include: developing a behavioural segmentation of domestic and light commercial consumers, based on smart meter data, for planning and targeting purposes; developing adaptive, real-time, rolling forecast methodologies that can drive smart control of generation and storage solutions; quantitative full-scale simulation of network performance under new technology scenarios; and providing next-generation demand planning solutions. This work has led to a number of new opportunities for Oxford to collaborate with other UK DNOs, as well as customer-facing smart meter analytics projects abroad. It is also a leading part of a wider pan-Oxford effort across the sciences and social sciences under the Energy theme.

Other work in this area through projects supported by EPSRC, the European Commission and the Energy Technologies Institute (Wallom) covers the development and deployment of services to utilise fundamental data mining algorithms and tools such that user communities are able to take raw data and generate actionable information. These include:

- Developing and leading national and international collaborations with academia and industry on how to transform the energy networks of today into the smart networks of tomorrow with minimal or no disruption.
- Understanding the impact of transformative energy pricing policies on consumers within the private and domestic sectors.
- Development and deployment of e-infrastructure services to support computational and data analysis as a service to move these from back office function to go-to services for operators.

**Transport:** Oxford researchers (Rauch, Willems) have been granted access to a data set that contains all individual travel movements on the London public transport system for several weeks before and after the most recent Tube strike in 2014. They are using the occurrence of the strike as to investigate how that disruption affected travellers' experimentation to find new routes: did people switch back to their original routes after the strike was over, or did some of them stick to their new route? The answer to this question can elucidate the rationality of individuals, their inclination to experiment during 'normal times', and their ability to find optimal paths in networks.

**National Security and Counter Terrorism:** Internet-based activities and the data they generate have profound implications for policy initiatives tackling crime and counter-terrorism. Effective national security relies on tracking and monitoring of data of all kinds, and there is thus an obvious and pressing need for the UK's security services, particularly GCHQ, to be at the forefront of research into big data. In fulfilling this need, as with other of their core activities, it is natural that the security services should retain the services of leading academic researchers with expertise in the relevant areas. The extent to which Oxford researchers are engaged in such work is a highly sensitive issue, but it is a matter of public record that some of the impact statements submitted by the department to REF2014 cover classified research. In addition, Grindrod has expertise in models for counter-terrorism and real time recognition of anomalies within vast communications data sets. Ker works in information hiding, particularly the hiding of messages in digital media such as images, and develops methods to analyse millions of images in order uncover covert communication channels.

## d. An ethical data science

Data about human behaviour poses ethical challenges for research which are distinct from the challenges involved in analysing data about galaxies or particles, for example. The risk is that of a 'double bottleneck': ethical mistakes or misunderstandings may lead to distorted legislation, which may cripple the usability of big data for the public good. There is a widely acknowledged need for better understanding of, and the elaboration of a national framework for, the ethical use of big data. Key issues to be tackled include: transparency (particularly for the secondary uses of personal data); consent; privacy; ensuring that ethical guidelines cover all big data eventualities; balance (e.g. between individual rights to privacy and public health threats or national security); and management, in terms of who has the right to access, use, audit and control the release of data.

Oxford philosopher Luciano Floridi, founder of the field of the Ethics of Information, is working towards the development of such a framework, establishing a team to work on national guidelines for working with big data. This includes a major two year University-funded project entitled 'The Ethics of Biomedical Big Data' with colleagues across Medical Sciences and the Mathematical, Physical and Life Sciences. He works with policy-makers including the European Commission (as Chairman of their 'Onlife Initiative') and Google (as a member of their Advisory Board on the 'Right to be Forgotten'). The ultimate aim is to mainstream ethics in decision-making relating to big data. Important ethical issues have also been explored in the award winning bestselling book 'Big Data' by Oxford's Viktor Mayer-Schönberger, and in a Sloan foundation project on big data.

# 3. Finance

The turmoil witnessed in financial markets in recent years has illustrated important links between seemingly disparate markets, and a high level of connectivity of the global financial system. These interdependencies between financial institutions or assets are often poorly understood and can have large and unforeseen consequences. Analysis of financial data and markets can lead to a greatly improved understanding of changing risk states, and help financial institutions make sense of new global market conditions. In addition, banks and other financial institutions are faced with an avalanche of data from customers and other sources, which needs proper analysis if it is to help them develop good financial and investment strategies.

## a. Quantitative finance

Oxford has a wide range of researchers working on all aspects of quantitative finance. Several groups from the Mathematical Institute are active in applying analytics to the financial sector, including the Numerical Analysis Group, and Mathematical and Computational Finance. The latter is one of the largest and most dynamic research environments in mathematical finance in the world, with internationally recognised experts in core mathematical fields. This is reflected in its permanent links with the Nomura Centre for Mathematical Finance (based within the Mathematical Institute) and the Oxford-Man Institute, an externally-funded financial research centre. These collaborations generate lively interactions between researchers coming from different backgrounds, and a truly impressive seminar programme.

The Oxford-Man Institute (OMI) focuses on interdisciplinary quantitative finance (mathematics as applied to the financial markets), and has impact on businesses; it is actively involved in four of Oxford's CDTs. The strong social context of the data from OMI's business partners across the financial, insurance and pensions sectors has the potential to impact on public policy. As well as working with the Mathematical Institute, the OMI operates in collaboration with the Saïd Business School, Oxford sociologists, engineers, computer scientists, and data-driven research in law. For example, collaborative research (Lehdonvirta, Lyons) is combining the Oxford Internet Institute's leading expertise in large-scale Internet data collection methods with OMI's expertise in novel financial data analysis methods to facilitate a better understanding of cryptocurrency markets (virtual currencies that use cryptography for security). In addition, research in Engineering Science on Bayesian inference (Roberts, Osborne) has found a powerful application in extracting and combining sparse information in huge arrays of financial data, including natural-language Twitter data.

## b. Financial big data

With the fast development of information technology, financial markets are generating huge volumes of data at a rapid pace. Faced with massive amounts of internal information and an ever-growing pool of unstructured data, financial institutions have yet to take advantage of all that that data has to offer. Research on financial big data has until now been less prominent, but has the potential to make a major contribution in shaping investment strategies and creating new businesses. The externally-funded Oxford-Nie Financial Big Data Lab is the first research laboratory

in this area at a major university, and will be at the front line of data science research applied to finance. It will provide a platform for research collaboration between academics, practitioners and regulators, and generate research that informs how financial firms can manage big data and benefit from it.

### c. Financial networks

The collapse of Lehman Brothers in 2008 led to a dramatic change in the way global markets operate. Prior to 2008, markets were more stable and predictable and there was no marked correlation between assets. However, as the financial crisis unfolded investors began to panic and move to assets that were perceived to be less risky, generating a very high level of correlation. This prompted Oxford mathematics researchers, in collaboration with the investment bank HSBC, to consider how correlation between assets evolved over time and how this was related to events at a macroeconomic and geopolitical level.

Using HSBC data, researchers investigate the structure and dynamics of financial networks using tools which cluster the data into densely connected groups, thereby revealing underlying structure in the network and detecting functionalities or relationships between the nodes (Porter). In collaboration with HSCB, Oxford was instrumental in creating the 'Risk on, risk off' (RORO) paradigm, a new index for characterising the behaviour of markets in the wake of the 2008 crisis. By analysing large datasets of asset prices over 12 years, the researchers were able to examine how correlation between assets evolved over time and how this was related to events at a macroeconomic and geopolitical level, revealing a structure in the data that was not present in the models used in standard financial theory. Having begun as a specialised research tool within HSBC's foreign exchange team, the RORO methodology was quickly disseminated and has had profound consequences for all other market participants, since it provides a replacement framework which enables investors to construct new strategies for the allocation of assets in a more unstable and unpredictable global market.

### d. Limit order books

The Mathematical Institute also conducts research on models of limit order trading in foreign exchange spot markets, where trades are conducted via bilateral trade agreements (rather than through a central counterparty). Researchers have introduced a new method of measuring prices in such markets, and have identified several 'stylized facts' of price formation which enable better understanding of the process of price formation. Models like these can help an understanding of the latent structure of the network of bilateral trade agreements, and help to explain how market participants make decisions when assessing the market.

# 4. Business

Timely and appropriate data analysis is critical to the success of businesses across the board, from retail to insurance. In the UK the retail sector contains the largest non-public employers and represents a particular area of growth and national excellence within data science and analytics. Understanding consumer behaviour is important for a wide range of applications, from developing more successful marketing strategies to economic policy design. Modern retailers now have access to data streams about their customers from both traditional shop-based outlets (loyalty cards or EPOS transactions), and e-commerce platforms. The immediate and obvious application of this is in designing targeted communications to individual customers (for example, personalised vouchers), but such data can also be deployed for strategic applications that benefit the business.

## a. Retail analytics and consumer behaviour

Research at Oxford (Grindrod) aims to extract and understand the patterns that are present in shopping data, using large data sets composed of customers' transaction histories: when and how much they bought of a particular product, together with additional information on demographics and the products themselves. The research focuses on models of behavioural dynamics including customer lifetime value modelling, identification of common shopping patterns (often called shopper missions), and the incorporation of ideas from behaviour economics (heuristics used in decision making by consumers of bounded rationality) into future retail analytics applications. Partners in the research include Tesco and its subsidiaries as well as large suppliers such as Unilever e-commerce and Unilever Research. The know-how generated extends to relevant problems owned by retail banking (Lloyds, Customer Decisions and Analytics), mobile telephone operators (Vodafone), and internet service providers (Telefonica). A separate research strand (Smith) analyses very large sets of purchasing data in order to understand supermarket pricing incentives, and additionally studies millions of transactions in the construction industry to understand the determination of prices in business-to-business markets where the buyers are large firms.

## b. Big data in the business sector

In a more general business context, the Saïd Business School (SBS) researches the potential impacts of big data on business sectors such as insurance, healthcare, logistics and retail, as well as analysing large-scale financial datasets through collaborative research across Oxford departments. As an example, a recent collaboration between SBS and IBM explored how organisations are dealing with the enhanced volume of data about their clients and customers. One key finding is that the impediment to greater adoption of big data is the lack of a clear business rationale and business owner for the activity. The team is continuing to investigate how SMEs and the public sector deal with this issue, and will also look at the societal considerations that arise around security and consent. SBS's Complexity Economics Programme aims to deepen the understanding of important economic phenomena, such as financial crises, economic growth, inequality, technological innovation, and the management of systemic risk, by applying an interdisciplinary perspective grounded in the study of complex systems, using tools such as complex network analysis and agent-based modelling. SBS also engages in collaborative financial data research with other Oxford departments and institutes such as the Oxford-Man Institute for Quantitative Finance.

# 5. Biomedical and Life Science

Health is perhaps the area in which the greatest advances have been made in the application of data science. Huge data sources such as UK Biobank, the UK's 100,000 genome project and new electronic patient records provide enormous scope for big data analysis which look set to transform all aspects of healthcare: the way public health is delivered, the development of personalised medicine, and the identification of new drugs and therapies.

Oxford has been at the forefront of biomedical research for two decades and is extremely well-placed to capitalise on these new developments in healthcare. The Old Road campus, established in 1992, now houses eight distinct institutes based around either platform technologies (genomics, epidemiology, bioengineering) or therapeutic areas such as cancer and inflammatory joint disease, and has more than 2000 scientists exploiting the unique collaboration opportunities offered by the Campus and the adjacent hospitals. The institutes offer ample opportunity for partnerships between academics from health and non-health disciplines (e.g. Engineering and Statistics). At the heart of the campus is the newly-established Li Ka Shing Centre for Health Information and Discovery, which will focus specifically on the potential of 'big data' to revolutionise health research and offer safer and more personalised treatments for patients. The centre brings together leading University researchers from across genetics, epidemiology and public health, clinical medicine, computer science and IT, statistics and bioinformatics, working alongside drug companies.

## a. Identifying new drug targets

High throughput biology is capable of generating very large amounts of data and provides the opportunity to analyse biological pathways systematically in order to understand, at a fundamental level, how they could be manipulated to treat disease. One of the major challenges facing pharmaceutical companies is identifying and validating potential drug targets before launching hugely costly commercial drug discovery programmes.

Oxford's Target Discovery Institute (TDI) is establishing high throughput biological approaches, including genomics, proteomics, small molecule screening, structural genomics and computational biology, and will work in partnership with the pharmaceutical industry to define and characterise better drug targets. This is exemplified by the Structural Genomics Consortium, where work in conjunction with Engineering has demonstrated the value of Machine Learning techniques to spot protein crystals in images from high-throughput experiments. The TDI represents a new approach for academia interacting with industry and is already engaged with 10 major collaborations with pharmaceutical companies in the area of target discovery and validation.

## b. Computational statistics and genomics

Computational statistics applied to high-throughput genetic data has been a major driver behind recent advances in biomedical science and in our understanding of human population history and evolution. These fields typically involve the analysis of genetic data to look forward in time, to predict those individuals likely to be at risk of a particular disease or those responding to a particular treatment, as well as look backward in time to chart our common human ancestry, right back to the

dawn of life on earth. For example, by adapting hidden Markov models, a tool originally developed in signal processing, using approximate probability structures, researchers at Oxford analysed hundreds of complete human genome sequences from around the world to reconstruct an ancestral map of human migration and historical (genetic) mixture events. Comparing human genomes to those of other species is an area of growing research, and Oxford researchers (Holland, Ponting) have used such analyses to reveal major genetic changes in evolution. In genetic epidemiology, new technology allowing measurement at the individual single-cell level coupled with new experimental methods which allow for genetic editing has opened up the prospect for causal analysis on the functional consequences of naturally occurring genetic variation. This has demanded tailored computational statistical methods to facilitate the information gathering from such high-dimensional, noisy but rich, experimental data.

Related techniques have also been applied to genetic data from human viruses, including HIV, influenza, hepatitis and dengue viruses. Data sets from these organisms already comprise tens or hundreds of thousands of genetic sequences. Zoology researchers (Pybus) are developing computational tools that mine the rich epidemiological information contained in viral genomes. These tools reveal the rate and nature of viral transmission and are used to inform public health decisions during epidemics.

### c. Data analytics and healthcare

Healthcare systems world-wide are entering a new, exciting phase: ever-increasing quantities of complex, massively multivariate data concerning all aspects of patient care are starting to be routinely acquired and stored, throughout the life of a patient. This exponential growth in data far outpaces the capability of clinical experts to cope, resulting in a 'data deluge' in which the data are largely unexploited. There is huge potential for using advances in 'big data' machine learning methodologies to exploit the contents of these complex datasets. Robust, scalable, automated inference can improve healthcare outcomes significantly by using patient-specific probabilistic models, a field in which there is little existing research, and which promises to develop into a new industry supporting the next generation of healthcare technology. Data integration across spatial scales, from molecular to population level, and across temporal scales, from fixed genomic data to a beat-by-beat electrocardiogram, will be one of the key challenges for exploiting these massive, disparate datasets.

Current national interest focuses on linking disparate databases within the hospital electronic patient record (EPR), such as patient demographics, blood chemistry, and physiology, using conventional statistical methods. However, conventional methods do not scale to such incomplete, noisy, terabyte-scale, massively multivariate datasets. The research agenda of the Biomedical Signal Processing group (Tarassenko) is to build complex multi-scale models and extract clinically-useful information from very large healthcare datasets in order to optimise patient treatment. Models are built using either a bottom-up approach (for example, using Bayesian Gaussian processes for modelling multivariate time-series data) or a top-down approach based on massively-multivariate probabilistic graphical models. In both cases, learning takes place within a Bayesian framework which provides the optimal approach for quantifying the uncertainty associated with the noisy and missing data typical of large healthcare datasets.

Existing projects include the fusion of genomic data with time-series EPR data to (i) understand and track resistance to antibiotics in infectious diseases; (ii) identify new strains of known pathogens, such as tuberculosis, E. coli, and C. diff; (iii) better understand the causal factors in inflammatory disease, and identify how existing treatments may be tailored in a patient-specific manner to improve patient outcomes.

Another area of research of considerable import is that of health informatics and translational medicine. Here, scientific progress will rely increasingly upon the availability and re-use of detailed, comparable data on large numbers of people. Electronic patient records are a valuable resource, but are often lacking in detail, quality, and consistency. By augmenting legacy records systems with new, lightweight technologies for data capture and integration, Oxford researchers (Davies) are providing support for translational medicine within the NHS.

A key feature of the Oxford approach is the emphasis upon linked data: every piece of data can be linked automatically to structured metadata describing the intended interpretation and usage, the context of collection, and also any subsequent processing. Software systems can read the metadata, and use it to determine how the data is indexed, transformed, integrated, secured, presented, or managed. This allows for a high degree of automation and, simultaneously, a high degree of customisation. This is precisely what is required to achieve detail, quality, and consistency at scale, and this approach will be at the heart of data management for future health informatics and research.

### d. Infrastructure, standards and training for life science and biomedicine

Life science is becoming increasingly collaborative and complex: diverse technologies are used to understand organisms and diseases, and data come from multiple sources in a heterogeneous manner and grows at an unprecedented scale. In translational medicine, there is a rapidly expanding interest in the use of integrative analysis approaches, i.e. systems biology, that require Knowledge Management platforms that not only store data but also facilitate the comparative analysis of different data types and the use of advanced analytical and modelling tools. There is also expanding enthusiasm from the private sector to harness the potential of open data paradigms for better exploitation of shared datasets. A major hurdle is linked to lack of interoperable infrastructure and common data standards, and often to the use of commercially available platforms; the legacy of data generated within publicly funded projects becomes challenging.

To meet these challenges, two Europe-wide initiatives have been set up: ELIXIR, providing a coherent infrastructure to enable researchers to seamlessly navigate the ecosystem of life-science data services; and the Innovative Medicines Initiative (IMI), Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients. Oxford researchers (Sansone, Ponting) make a significant contribution to these initiatives, driving the curation and standards training sector in the ELIXIR-UK node, and working on IMI's eTRIKS project, which aims to develop solutions to data-sharing problems and create and run an open, sustainable research informatics and analytics platform. Oxford also contributes to the newly funded Centre of Excellence for Annotation and Data Retrieval (CEDAR), one of the 11 centres funded in the NIH Big Data to Knowledge Initiative, and works with international partners to pilot new methods and solutions to

maximize data interoperability, reproducibility and reuse, and to set a common data standards agenda.

### e. Human brain analytics

Researchers in the mathematics department (Grindrod) are working with fMRI and MRI data on a variety of problems inferring high resolution structure of the most active pathways for information processing in the brain, in collaborations with applied researchers (including clinical language sciences, psychology and engineering/cybernetics). This has led to a number of innovations in understanding how human brains do what they do based on relatively small scale irreducible sub networks of cells; the relationships between learning and the evolving dynamic architecture; the effects of various drugs on schizophrenia patients or the recovery of damaged brains; the prediction of the onset of Alzheimer's disease versus mild cognitive impairments; and in informing new paradigms of non-binary, low energy, computing (working with IBM and others). Dominantly based on the theory of high-resolution evolving network representations for the brain, such ideas may be relevant to the development of novel clinical devices (and currently participation in the international CAD Dementia Grand Challenge).

# 6. Climate and the Environment

A stream of observations from sources such as satellites and weather sensors across the globe means that climate and environment researchers have access to an unprecedented volume of climate and environment data, and increasingly sophisticated simulations and models are built on such data. In combination these allow scientists to understand and predict events such as weather, climate change, earthquakes and volcanic eruptions with greater precision, to evaluate environmental quality and disease risk factors, and to support wildlife tracking and conservation.

## a. Weather and climate forecasting

Meteorology and climate science involve the manipulation and analysis of enormous amounts of data. The vast majority of observations of Earth's climate system come from satellites which measure a range of geophysical variables, not only at Earth's surface, but at many different levels in the atmosphere. Additionally, computer models of the climate system produce forecasts on a range of timescales, from a few hours ahead to many centuries ahead. Predictions of weather and climate are inherently uncertain: initial conditions can never be known perfectly, and weather and climate simulators are necessarily imperfect representations of the underlying laws of physics.

Oxford physicists (Palmer) are at the forefront of research that aims to reliably quantify the uncertainty in weather and climate predictions, and identify what is needed to reduce current levels of uncertainty. Palmer has pioneered the development of ensemble forecasting, where models are run multiple times varying uncertain initial conditions or model parameters in order to mitigate the effects of chaos. Very short range weather forecasts are made with extremely high-resolution models which can resolve individual cloud systems. Longer climate-change forecasts are made with much lower-resolution models, but these produce estimates of future weather each day over the coming century. All meteorological forecasts are made in ensemble mode, and this alone increases the generation of meteorological data by roughly to two orders of magnitude.

Such probabilistic prediction systems have the potential to be highly reliable. However, the climate is an extremely complex physical system, which means that forecast models are always approximations of reality. In order that users of weather and climate models can rely on the predictions made, researchers perform detailed analyses of the output of these models, comparing them with the observations and using the results to determine ways to improve the models. Then, when improved representations have been developed, further analysis can be conducted to verify that forecast errors have been reduced. In order to undertake this work effectively, efficient ways of manipulating vast amounts of data have to be developed - pushing forward the data science agenda. Innovative approaches to this at Oxford include an investigation of the use of stochastic processors, which allow hardware-induced faults in calculations in exchange for improvements in performance or a large reduction in power consumption – a potentially important trade-off in the context of the huge energy requirements of supercomputers.

Oxford researchers are involved in the analysis of observational and model data (as well as in the development of the satellite instruments and models themselves), collaborating with the Met Office and with the European Centre for Medium-Range Weather Forecasts at Reading, as well as with other centres in the UK and around the world.

## b. Climate modelling

Oxford established and continues to operate the world's largest climate modelling experiment, Climateprediction.net, a simulation that runs on hundreds of thousands of private computers (M Allen, Wallom). The data generated allow researchers to answer important and difficult questions about how climate change is affecting the world now and how it will affect the world in the future.

A recent development of Climateprediction.net is the regional climate modelling experiment weather@home. By embedding a regional model in a global climate model the impact of climate change on local weather can be assessed. weather@home continues to utilise the large network of private computers which allows for many hundreds of thousands of simulations of daily weather to be realised. weather@home allows the project scientists to explore future scenarios of the changing climate and also to pose the question of what the weather would be like if climate change had not occurred. One recent Oxford study quantified the role in which man-made climate change increased the risk of occurrence of the UK floods in 2013 and 2014.

weather@home also brings a new problem of how to effectively analyse a large amount of high-spatial and high-temporal resolution data. Through a move from global to higher resolution regional modelling, researchers are building on their previous experience to explore new experiments and answer new questions. A project carried out in conjunction with the reinsurance industry to investigate the probability of occurrence of European winter wind storms found an effective way of reducing this data. Using data-mining, objective feature recognition and tracking and data packing techniques a catalogue of 800,000 potential European wind storm footprints was produced in a dataset of ~80GB. This was a large reduction from the initial data output by the regional climate model, which was about 12TB in size. The dataset is now in use by five reinsurance companies to inform their underwriting decisions. Through support from groups such as Climate Central, this information can be made widely available and used in ways that will be visible to the general public.

In addition, engineering researchers in Oxford (Roberts, Osborne) have introduced methods that leverage sensor networks to provide real-time estimates of dynamic environmental phenomena such as air temperature. The use of flexible non-parametric algorithms allows effective inference even with minimal domain knowledge, and a probabilistic approach provides uncertainty estimates to support the decision-making of human operators. The methods have been applied in cases where the data are delayed, intermittently missing, censored or correlated, and validated using data collected from multiple real networks of weather sensors, including Bramblemet, MIDAS land surface stations and the Wannengrat Alpine Observatory. Algorithms have been developed that parsimoniously select only the most valuable observations from sensor networks. Real observations are usually associated with some cost, such as the battery energy required to power a sensor or transmit a reading: it is desirable to reduce the number of such observations. Oxford researchers have demonstrated such active data selection in real sensor networks, allowing the algorithms to intelligently concentrate sparse observations in the most informative spatio-temporal locations, even in the presence of dynamic, missing and faulty data.

### c. Natural catastrophes

There is a breadth and depth of research excellence at Oxford in various areas of big data that relate to environmental activities, specifically the risks associated with changes to the environment. Research into the quantification and financing of risk associated with natural catastrophes includes research funded by the Risk Prediction Initiative to understand the probability of extreme wind storms in Europe for the reinsurance industry, and the development of techniques to understand better interconnected risks using time series modelling, machine learning, signal processing, and systems analysis.

Activities with the Saïd Business School emphasise the global context, challenges and opportunities in which businesses are embedded; its Global Opportunities and Threats Oxford (GOTO) programme is inherently interdisciplinary, providing expert perspectives on world-scale challenges such as Big Data, demographic change, and water scarcity. The work of Oxford's Global Cyber Security Capacity Centre has many applications to environmental risks including risk propagation and communication and visual analytics.

Researchers in Earth Sciences are key players in the Centre for the Observation and Modelling of Earthquakes, Volcanoes and Tectonics, which uses Earth Observation data to study earthquakes and volcanoes. They are also active in collaborative projects such as Strengthening Resilience in Volcanic Areas (STREVA – Mather, Pyle) and Earthquakes without Frontiers (EwF - Parsons), both of which make use of large volumes of seismic and volcanic data to help communities in hazard areas to improve their preparedness for eruptions or earthquakes.

### d. Wildlife conservation, biodiversity management, and collective and emergent behaviour

A wide diversity of fields in ecology and animal behaviour have been transformed by the data-rich opportunities afforded by biologging technology. Traditionally, information about animals and their habitat has been acquired using labour-intensive direct observation and manual tracking of VHF tags. Recent advances in sensor technology have resulted in miniature tracking tags that combine a number of sensing modalities, such as GPS and accelerometers with wireless uplinks. Together with static sensors (such as camera traps and bioacoustic loggers) these provide a wealth of data on wildlife and farm animal behaviour, and human-wildlife interaction and pressure. Animals can be tracked interacting in real time, resulting in millions of observations for thousands of individuals, and distributions and behaviour can be inferred from machine-learning algorithms analysing remotely-sensed or video images.

These approaches have opened new fields of inquiry into the causes of emergent properties of groups, from the spread of disease and information, to the development of an 'early-warning' system for animal welfare. Oxford's expertise in this area ranges from sensor device development through to statistical analysis and inference of animal behaviour, and is heavily cross-disciplinary, covering Zoology, Computer Science and Information Engineering (Macdonald, Sheldon, Coulson, Guilford, Dawkins, Markham, Roberts). The broad challenges revolve around handling noisy and potentially incomplete sensor data, fusing data from multiple sensor sources and modalities and optimizing collection of data, given the limited size and weight of animal-borne devices.

### e. Ecological assessment and human risk factors

The vast data sets generated by remote sensing and satellite imaging have applications in environmental science, with applications relevant to conservation science and human health. Oxford researchers are at the forefront of both applications. For example, researchers in Zoology are developing tools that integrate vast spatial data sets to evaluate the ecological and environmental quality of all points on the Earth's land surface, to generate tools used by industry seeking to minimize environmental damage when harvesting natural geological resources (Willis). Similar mapping tools are also used to plot current distribution, and predict the future spread, of infectious diseases and their insect vectors, used in financial decision-making by governments and international agencies (Hay).

# 7. **Astronomy**

Data science has a vital role to play at the very largest scale of physics research. Modern telescopes are capable of capturing enormous quantities of data whose analysis can help to reveal hidden truths about the solar system, the Milky Way and the distant universe. Because of their sheer scale and cost, major advances in the collection of astronomical data usually involve international collaborations of researchers.

Oxford physicists and engineers are involved with the development of many of the next generation of telescopes that are currently being installed across the globe. The management, processing, storage and assessment of the vast amounts of data that will result typify the challenges facing data science. Teams of researchers from across Oxford are collaborating to develop appropriate solutions in these areas (Dunkley, Verma). The University is strongly represented on the Square Kilometre Array (SKA) project, and in various facilities that are testing the technology and are in themselves revolutionary telescopes: eMERLIN in the UK, and LOFAR and APERTIF in the Netherlands. Oxford is also making a key contribution to the development of the two telescopes that will eventually turn into different parts of the SKA, the South African MeerKAT and the Australian SKA Pathfinder telescopes.

Oxford astronomers were the originators of Galaxy Zoo, the first Zooniverse citizen science project (Lintott). Unlike many of the previous crowd sourcing projects which merely relied on the idle computing power of participants' computers, Galaxy Zoo required the active participation of human volunteers, and thus generated large amounts of data which needed to be fused to provide reliable answers. The needs of this and subsequent Zooniverse projects have driven the development of innovative mathematical and statistical solutions which are able to extract accurate information from the dynamic and uncertain data generated by hundreds of thousands of individual decision makers. Roberts' group in Engineering Science has created elegant and efficient solutions using Bayesian probability theory; using these methods, researchers have been able to extract up to 98% accuracy from data which showed only 75% accuracy when a simple averaging of scores was applied.

The actual results delivered by Galaxy Zoo, Planet Hunters and other Zooniverse astronomy projects have exceeded all expectations; for example, it resulted in the recent discovery by committed Planet Hunter volunteers of a double binary star system with a transiting planet, previously unknown to science. The volume of data analysed on the various projects would have been impossible for any individual researcher to achieve in a lifetime, and has given scientists access to valuable information that drives forward their research. The expertise developed at Oxford has blossomed into dozens of projects which now extend into zoology, history and other disciplines, creating a revolution in research endeavour which looks set to continue and expand still further.

# 8. Physical Sciences

## a. Particle physics

Oxford's particle physicists are engaged in experiments across the globe, probing the properties of the newly discovered Higgs boson, teasing out the behaviour of elusive neutrinos, dark matter, and dark energy, and searching out as-yet undiscovered features of physics and spacetime. Many of these experiments are themselves global enterprises, involving hundreds or thousands of physicists worldwide, analysing large datasets for tiny correlations; for instance, even after extensive online data filtering, the experiments of the Large Hadron Collider (LHC) in Switzerland still churn out approximately 30 petabytes of data each year. In this sense, 'big data' is a natural part of experimental particle physics, and while the field has pursued some large-scale computing directions, such as 'grid' computing, largely focused on the community's data processing and data sharing needs, the Oxford group has pioneered more generic approaches (Tseng). The group was one of the first to apply virtualisation to scientific analysis code, and later to demonstrate x86 emulation inside the Java Virtual Machine. The group also developed technology which dramatically streamlined secure 'citizen science' deployment. These projects found use in particle physics simulations as well as in astrophysics data analysis.

At the same time, the Oxford group plays a key role in managing the LHC's ATLAS experiment's 'metadata', i.e., the data about data which enables scientists to organize, find, and normalise the petabytes per year of experimental data streams so that they can be combined for maximal statistical power. This metadata is itself taking on 'big Data' dimensions, and is prompting new ideas and models, as well as updates of older ones, to meet the analysis challenges of fundamental physics.

## b. Chemistry

The neutron and X-ray facilities (ISIS and Diamond Light Source respectively) housed at the nearby Rutherford Appleton Laboratory are sources of significant data sets. For example, recent developments in rapid data acquisition techniques allow direct experimental observation of key fundamental physio-chemical processes (in particular crystallisation and vitrification). However, rapid acquisition has implications for both data quality and potential resolution of the data in terms of constituent chemical identities. Oxford chemistry theorists (Wilson) are at the forefront of developing (atomistic) simulation models which are crucial for interpretation and eventually will drive these experiments.

Oxford theorists are themselves the authors of large data sets through extensive computational work using local and national computational resources, covering subjects as diverse as bioinformatics, the modelling of DNA self-assembly, and the formation of low dimensional nanotubular structures.

# 9. Arts and Humanities

Digital Humanities research takes place at the intersection of digital technologies, humanities, and social sciences. It creates new data, methods, pedagogies and applications across the humanities disciplines, in computer science and its allied technologies and in a range of other human-centric disciplines. It also studies the impact of these shifts on cultural heritage, memory institutions, libraries, archives and digital culture.

Digital Humanities (DH) is by its very nature necessarily highly collaborative, marrying the novel use of technology with fundamental humanities research, and engaging with library and museum collections.

The University is an internationally-leading centre for DH research and teaching. Oxford has the largest community of digital humanities researchers anywhere in the world and hosts more digital projects than anywhere else in the UK: more than 200 are identified on the Digital Humanities at Oxford website (DH@Ox), many more than any other UK institution. The University also hosts the enormously successful Digital Humanities at Oxford Summer School annually. Over £14.5M of external grant income has been awarded since 2008-09 to DH projects within Humanities faculties, and a further £1M has been invested by the University in pump-priming DH research to develop new databases and DH resources. Oxford has had funding in all three rounds of the international 'Digging into Data' programme (Round 1 - Digging into the Enlightenment, Mining a Year of Speech, and Structural Analysis of Large Amounts of Music; Round 2 - Imagery Lenses for Visualizing Text Corpora; Round 3 - Commonplace Cultures, and Resurrecting Early Christian Lives).

Unlike many other institutions Oxford has not formed a Digital Humanities centre, reflecting its dispersed ecology. Instead a Digital Humanities working group has developed in parallel to the DH@Ox website to provide overarching coordination and communication. Among others, it involves IT Services, the Bodleian Libraries, Oxford e-Research Centre (the only national e-Science centre that remains of the 10 originally funded by RCUK in 2001), Oxford Internet Institute, Oxford's libraries, museums and collections, and a wide cross section of humanities researchers. Oxford's digital humanities is organised as a network and a partnership, which pervades the University's museums, libraries, and faculties and departments. The Oxford Research Centre in the Humanities (TORCH) is leading the development and implementation of the university's strategy in Digital Humanities, and provides support for this highly interdisciplinary activity. As the publisher and IP owner of a substantial proportion of Britain's literary, linguistic and historical cultural assets, the University also has a vital role to play in creating resources, standards, infrastructure and services to support digital data in this wide-ranging area.

## a. Arts and humanities data sets

DH collaboration in Oxford has demonstrated that its competitive advantage lies in its ability to aggregate academic excellence, technical expertise and a critical mass of exceptional collections. A major achievement is the creation and publication of a portfolio of very large arts and humanities data sets, the size and scale of which stand comparison with that of datasets in the sciences. Oxford's digital assets include the Oxford Text Archive; the British National Corpus, 100 million

words of contemporary English text and 1,500 hours of spoken audio (part of the Year of Speech corpus, which requires twice the storage of one year's data from the Hubble space telescope); the CLAROS Web of Art, images and metadata relating to over 20 million artefacts, which is over 1½ times the size of the Sloan digital sky survey; Electronic Enlightenment, 64,000 historical letters written by 8,008 historical figures from the early 17th to the mid-19th century; scans of 25,000 Early English Books Online; and hundreds of thousands of out-of-copyright books from the Bodleian library that were scanned and published in partnership with Google Books. These and other such resources are openly available, contributing the highest-quality educational and cultural resources to the public, enriching the quality of life of school students, teachers, and other citizens.

### b. The spoken word

The collection and analysis of very large samples of speech is necessary because there are so many sources of variation in language: variation between people from different places, between males and females, young and old. In order to tackle the problem presented by human and social variation, it is statistically necessary to engage with increasingly large samples of data. Yet finding segments of interest in masses of audio is far more challenging than text-based search; it is for this reason that advanced signal processing techniques need to be developed and scaled up to deal with processing Petabyte-scale audio.

Oxford leads humanities research in the development and application of speech and natural language processing technologies to the analysis of digital audio resources (Coleman). The Phonetics Laboratory leads the world in the application of Speech Recognition technology to the automatic indexing (time-aligned mark-up) of huge corpora of spoken audio recordings, such as the on-line Audio edition of the British National Corpus, which provides the means for researchers to search for and retrieve 2-3 orders of magnitude more of spoken language data than previous collections, at the level of single words or even individual vowels and consonants. As a consequence, Oxford researchers are redrawing the lines between humanities and science concerning the analysis of language – especially UK English in all its rich variety. Just as in the past century Oxford came to be seen as the trusted recorder of written English, the University now has its sights set on the collection, analysis and documentation of recorded spoken English, in all styles and accents.

Digital audio recordings represent a massive and rapidly-growing sector of internet traffic; this is partly because audio consumes several orders of magnitude of storage and bandwidth than text. Beside the University's language resources, Oxford anthropologists and musicologists have published extensive collections of audio material from other cultures, notably the Pitt Rivers Museum's Reel to Real archive, which brings hugely important cultural assets out of the back rooms of the museum for public enjoyment and scholarly appreciation.

### c. Music

The analysis of digital audio recordings of musical performance is an important new opportunity for musicology. The international Digging into Data project 'Structural Analysis of Large Amounts of Music Information' conducted web-scale music analysis, and collaborative research work at Oxford is now exploring how emerging technologies for working with music as sound and score can

transform musicology, both as an academic discipline and as a practice outside the university. This computational musicology expertise also underpins the project 'Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption', providing an important intersection with the creative industries sector.

### d. Citizen science

Digital Humanities in Oxford exemplifies new data analysis methodologies through citizen science and crowdsourcing. The Zooniverse platform began with Galaxy Zoo, an astronomy project, but has since expanded into dozens of projects requiring the active participation of human volunteers, with well over a million users. Zooniverse supports multiple humanities projects: the data gathered by Ancient Lives helps scholars study the Oxyrhynchus collection, Operation War Diary explores soldiers' diaries from the First World War through annotation and tagging, and Constructing Scientific Communities is investigating Citizen Science in the 19th and 21st Centuries.

# 10. **Education**

Oxford has a strong tradition of graduate research and training in all areas of data science. Seven of its EPSRC- and ESRC-funded Centres for Doctoral Training offer courses which are highly relevant, enabling graduates to gain the skills necessary to tackle the data science problems of the future:

- The Autonomous Intelligent Machines and Systems CDT provides the opportunity to develop in-depth knowledge, understanding and expertise in autonomous intelligent systems, mixing both practical and theoretical aspects of intelligent machines and systems.
- The Cyber Security CDT has core research themes including cyber-physical security, real-time security, assurance and big-data security.
- The Industrially Focused Mathematical Modelling CDT trains students in a broad range of techniques spanning mathematical modelling, analysis and computation relevant to addressing challenges that face modern companies. You will be actively engaged with the CDT's company partners through courses, mini projects and research projects.
- The Statistical Science CDT runs a programme in the theory, methods and applications of next-generation statistical science for 21st century data-intensive environments and large-scale models.
- The Systems Approaches to Biomedical Science CDT is an innovative open collaboration between the University of Oxford and 15 partner industrial organisations, working together to develop novel computational, mathematical and physical techniques to solve biomedical research problems.
- The Theory and Modelling in Chemical Sciences CDT trains students in theoretical and computational chemistry, covering the development of new theory, implementation of methods as reliable software, and application of such methods to a host of challenges in chemical and related sciences.
- The ESRC Doctoral Training Centre provides a Social Science of the Internet pathway which develops an in-depth understanding of the concepts, theories and methods (including Internet-specific methods) required to undertake and assess rigorous empirical research and policy analysis of Internet-related issues.

In addition, many other graduate courses contain a strong element of data science research, or are entirely focused on teaching the data-related skills needed by business and industry. The Mathematical & Computational Finance MSc, for example, is a 10 month full-time intensive programme which prepares students for a career in quantitative finance in the financial industry.

## 11. **Facilities and Infrastructure**

There will be an ongoing requirement for infrastructure to support the development of data science across Oxford. This will include both storage facilities and the interconnecting networks to provide secure and efficient channels for data transfer. The magnitudes of data that will be created will demand ever more innovative approaches to storing and transferring information efficiently, especially where external regional or national computational facilities are utilised in any overall workflow.

The IT Services and Oxford e-Research Centre provisioned ARC facility (Advanced Research Computing) already provides access to, and advice on using, a range of high-performance computing infrastructure both within the University and via external, regional, collaborations in the Science Engineering South e-Infrastructure (SES). The ARC facility acts as one component in the processing of large data volumes for numerous projects, and can act as a foundation on which to build larger data repositories for active processing. Through other components of the IT Services portfolio, combined with the expertise of the e-Research Centre and Bodleian services, ARC is well placed to interface with other storage facilities for the long term retention and curation of data (Trefethen, Richards).

Building on the participation in the European Grid Infrastructure (EGI), researchers from within the Oxford e-Research Centre have led since its inception the creation of the world's largest federation of public and private cloud computing resources (Wallom). The EGI-federated cloud system provides underpinning resources to research communities across the EU, allowing the storage of research data in a persistent manner, but also (unlike traditional methods) allowing the movement of the necessary computing processes close to the data, thus reducing unnecessary congestion on the continent's research and education networks and speeding paths to discovery. Current users of the system include the various European Strategic Frameworks for Research Infrastructure alongside national, local and private sector researchers in fields as diverse as Musicology and Ecology. Alongside this European effort Oxford researchers are also instrumental in the design and implementation of cloud computing resources at a national scale, working with institutions such as the European Bioinformatics Institute and STFC. One example is the Environmental Omics System Cloud to construct community cloud computing resources which can support the use of transformative research methodologies.

# 12. Innovations in IP Management

Data Science has the strong potential to generate economic growth through next generation analytics. It is proposed that the guiding principle should be that the IP process should not delay or inhibit the exploitation of the outcomes.

Current experiences across the University in open IP will be adapted to create the most appropriate approach for data analytics. Three related approaches are provided below by way of example of the type of arrangements that could be adopted; two are from EPSRC funded Centres for Doctoral Training (CDT) and one from a consortium of researchers and pharma companies.

## a. Open Innovation

The Systems Approach to Biomedical Science CDT currently works with a consortium of thirteen companies and two not-for-profit research institutes, all of whom have signed a single governing agreement, the basic principles of which are 'open innovation' with an emphasis on the students' ability to share and discuss their work within the programme and also to publish their results. The results of students' projects are available to all the partners involved, and there is a short lead-time on clearing research for publication (21 days). The agreement with companies helps protect all parties and gives the following rights in respect of IP: (i) The University will own all results and arising IP from the projects and have the on-going right to use it in research and teaching. (ii) Each company will have a free, non-exclusive right (which would generally be royalty-free) to use any of the results and IP from any of the projects for R&D and business purposes. (iii) However, if a company includes our IP in a product or process which is then commercialised, the University will receive a royalty (on terms to be agreed) in proportion to inventive contributions. (iv) The parties that take forward patenting will share the patenting costs.

## b. Open IP

The Structural Genomics Consortium has an open IP approach: it publishes everything quickly and disseminates research findings freely. The advantages of this approach include: (i) The ability to work with multiple private organisations; (ii) The ability to work with expert academics quickly; (iii) Reduction of duplication and wastage; (iv) Allowing the pooling of infrastructures and expertise; (v) Acceleration of science and drug discovery.

## c. Open Access

The CDT in Synthesis for Biology and Medicine (SBM) has an Open Access Programme based on cooperative academic and industrial training. Building on the department's existing CDT experiences and doctoral training networks, a genuinely integrated public-private partnership for doctoral training is being implemented.

To exploit fully the opportunities offered by its industry and government partners, the SBM CDT has adopted a model where all parties agree to publish research outputs in the public domain rather than protecting in patents to allow completely transparent exchange of information, know-how and

specific expertise not only between students and supervisors on different projects but between different companies.

The nature of the programme allows all stakeholders barrier-free, real-time access to all the information that the programme generates. This will also facilitate dissemination of scientific information through seminars, posters and on-line methods, as the SBM CDT will not be constrained by IP agreements.

## Directory of Researchers

**Researchers are listed alphabetically under each theme heading. Those who are active in several areas of research may be listed more than once.**

### 1. Underpinning Tools, Methods and Technology

**Dr Phil Blunsom**, University Lecturer in Computer Science, Department of Computer Science
*Intersection of machine learning and computational linguistics*
http://www.cs.ox.ac.uk/people/phil.blunsom/

**Professor Min Chen**, Professor of Scientific Visualisation, Oxford e-Research Centre (OeRC)
*Visualisation and Visual Analytics*
https://sites.google.com/site/drminchen/home

**Professor Nando de Freitas**, Professor of Computer Science, Department of Computer Science
*Machine learning and big data*
http://www.cs.ubc.ca/~nando/

**Professor Arnaud Doucet**, Professor of Statistics, Department of Statistics
*Bayesian statistics and novel Monte Carlo methods*
http://www.stats.ox.ac.uk/~doucet/

**Professor Mike Giles**, Professor of Scientific Computing, Mathematical Institute
*Parallel Computing, Monte Carlo methods*
http://www.oxford-man.ox.ac.uk/people/mike-giles

**Professor Chris Holmes**, Professor of Biostatistics, Department of Statistics
*Bayesian statistics and statistical modelling*
http://www.stats.ox.ac.uk/~cholmes/

**Professor Ian Horrocks,** Professor of Computer Science, Department of Computer Science
*Knowledge representation and ontologies*
http://www.cs.ox.ac.uk/people/ian.horrocks/

**Professor Thomas Lukasiewicz**, Professor of Computer Science, Department of Computer Science
*(Personalised) semantic search and query answering on the Web*
http://www.cs.ox.ac.uk/thomas.lukasiewicz/

**Professor Andrew Martin,** Associate Professor, Department of Computer Science
*Security of systems; trustworthy data processing*
http://www.cs.ox.ac.uk/andrew.martin/

**Dr Ivan Martinovic,** Associate Professor, Department of Computer Science
*Machine learning techniques applied to security analysis, continuous authentication*
http://www.cs.ox.ac.uk/people/ivan.martinovic/

**Professor Dan Olteanu,** Associate Professor, Department of Computer Science
*Databases and data management*
http://www.cs.ox.ac.uk/dan.olteanu/

**Professor Michael A. Osborne**, Associate Professor in Machine Learning, Department of Engineering
*Engineering: machine learning*
http://www.robots.ox.ac.uk/~mosb/

**Dr Kasper Rasmussen,** Departmental Lecturer, Department of Computer Science
*Machine learning techniques applied to security analysis, continuous authentication*
http://www.cs.ox.ac.uk/people/kasper.rasmussen/

**Professor Steve Roberts**, Professor of Machine Learning, Department of Engineering Science
*Machine learning approaches to data analysis, Bayesian statistics*
http://www.robots.ox.ac.uk/~sjrob/

**Dr Andrew Simpson,** University Lecturer, Department of Computer Science
*Evolving access controls, data protection and privacy*
http://www.cs.ox.ac.uk/andrew.simpson/

**Professor Lionel Tarassenko**, Director of the Institute for Biomedical Engineering, Department of Engineering Science
*Signal processing, biomedical engineering, patient monitoring*
http://www.ibme.ox.ac.uk/research/biomedical-signal-processing-instrumentation/prof-l-tarassenko

**Professor Yee Whye Teh**, Professor of Statistical Machine Learning, Department of Statistics
*Statistical machine learning*
http://www.stats.ox.ac.uk/people/academic_staff/yee_whye_teh

## 2. Society and Public Policy

**Dr William Allen**, Research Officer, The Migration Observatory
*Migrants and migration policies*
http://www.migrationobservatory.ox.ac.uk/about-us/william-allen

**Dr Jonathan Bright**, Research Fellow, Oxford Internet Institute
*Computational and 'big data' approaches to the social sciences*
http://www.oii.ox.ac.uk/people/?id=323

**Professor Ian Brown,** Professor of Information Security and Privacy and Associate Director, Oxford Internet Institute
*Surveillance, privacy-enhancing technologies, and Internet regulation*
http://www.oii.ox.ac.uk/people/brown/

**Professor Sadie Creese**, Professor of Cybersecurity, Department of Computer Science
*Cybersecurity*
http://www.cs.ox.ac.uk/people/sadie.creese/

**Professor Rebecca Eynon**, Senior Research Fellow and Associate Professor, Oxford Internet Institute
*Sociology of education, technology enhanced learning, everyday life and learning, digital and social exclusion*
http://www.oii.ox.ac.uk/people/?id=21

**Professor Luciano Floridi**, Director of Research and Professor of Philosophy and Ethics of Information, Oxford Internet Institute
*Information and computer ethics, philosophy of information, philosophy of technology*
http://www.philosophyofinformation.net/

**Dr Carl Benedikt Frey**, James Martin Fellow, Oxford Martin School
*Transition of industrial nations to digital economies, and subsequent challenges for economic growth and employment*
http://www.oxfordmartin.ox.ac.uk/people/453

**Professor Mark Graham,** Senior Research Fellow and Associate Professor, Oxford Internet Institute
*Geography of the internet*
http://www.oii.ox.ac.uk/people/?id=165

**Professor Peter Grindrod**, Mathematical Institute
*Peer to peer digital communication, energy networks, counter-terrorism*
http://www.maths.ox.ac.uk/people/peter.grindrod

**Professor Sir David Hendry**, Director, Economic Modelling, The Institute for New Economic Thinking, Oxford Martin School
*Econometric methodology, time-series econometrics, applied macroeconometrics*
http://www.oxfordmartin.ox.ac.uk/people/79

**Dr Bernie Hogan,** Research Fellow, Oxford Internet Institute
*Intersection of social networks and media convergence*
http://www.oii.ox.ac.uk/people/?id=140

**Dr Andrew Ker,** University Lecturer in Computer Security, Department of Computer Science
*Information hiding and covert communication*
http://www.cs.ox.ac.uk/andrew.ker/home.html

**Dr Vili Lehdonvirta**, Research Fellow, Oxford Internet Institute
*Virtual goods, virtual currencies and online labour markets*
http://vili.lehdonvirta.com/

**Professor Helen Margetts**, Director and Professor of Society & the Internet, Oxford Internet Institute
*Government and digital era governance and politics*
http://www.oii.ox.ac.uk/people/?id=2

**Professor Viktor Mayer-Schönberger**, Professor of Internet Governance and Regulation, Oxford Internet Institute
*Information in a networked economy*
http://www.oii.ox.ac.uk/people/?id=174

**Dr Patrick E McSharry,** Head of Catastrophe Risk Financing, Smith School of Enterprise and the Environment
*Risk, forecasting, decision-making, machine learning, big data*
http://www.oii.ox.ac.uk/people/?id=316

**Professor Edward Melhuish,** Professor of Education, Department of Education
*Educational research involving large and complex datasets*
http://www.education.ox.ac.uk/about-us/directory/professor-edward-melhuish/

**Professor Eric T. Meyer**, Associate Professor, Oxford Internet Institute
*Social informatics, big data, computational research, digital humanities*
http://www.oii.ox.ac.uk/people/?id=120

**Professor Melinda Mills,** Nuffield Professor of Sociology
*Combining social science and genetic approaches*
http://www.sociology.ox.ac.uk/academic-staff/melinda-mills.html

**Dr Andrew Przybylski**, Research Fellow, Oxford Internet Institute
*Psychology, human motivation, video games, virtual environments*
http://www.oii.ox.ac.uk/people/?id=328

**Dr Ferdinand Rauch**, Associate Professor of Economics, Department of Economics
*Applied international economics, economic geography*
http://www.economics.ox.ac.uk/Academic/ferdinand-rauch

**Dr Felix Reed-Tsochas**, Director, Oxford Martin Programme on Complexity, Saïd Business School
*Complex networks*
http://www.sbs.ox.ac.uk/community/people/felix-reed-tsochas

**Professor Pam Sammons,** Professor of Education, Department of Education
*Educational research involving large and complex datasets*
http://www.education.ox.ac.uk/about-us/directory/professor-pam-sammons/

**Professor Ralph Schroeder**, Senior Research Fellow, Oxford Internet Institute
*Virtual environments, social aspects of e-Science, sociology of science and technology*
http://www.oii.ox.ac.uk/people/?id=26

**Professor Steve Strand,** Professor of Education, Department of Education
*Educational research involving large and complex datasets, equity gaps in educational outcomes*
http://www.education.ox.ac.uk/about-us/directory/professor-steve-strand/

**Professor David Stuckler,** Professor of Political Economy and Sociology, Department of Sociology
*Integration of political economy and public health*
http://www.sociology.ox.ac.uk/academic-staff/david-stuckler.html

**Dr David Wallom**, Associate Director, Oxford e-Research Centre
*Energy and ICT, cloud computing, volunteer computing*
http://www.oerc.ox.ac.uk/people/david-wallom

**Dr Tim Willems**, Postdoctoral Research Fellow, Department of Economics
*Macroeconomics, political economics, learning, optimal experimentation, imperfect information, and applied econometrics*
https://sites.google.com/site/twillems85/

## 3. Finance

**Professor Ian Goldin**, Professor of Globalisation and Development, Director of the Oxford Martin School
http://www.oxfordmartin.ox.ac.uk/about/director/

**Professor Georg Gottlob**, Professor of Informatics, Department of Computer Science
*Information Systems, Algorithms*
http://www.cs.ox.ac.uk/people/georg.gottlob/userweb/

**Dr Vili Lehdonvirta**, Research Fellow, Oxford Internet Institute
*Virtual goods, virtual currencies and online labour markets*
http://vili.lehdonvirta.com/

**Professor Terry Lyons**, Director, Oxford-Man Institute
*Rough Paths, Stochastic Analysis, and applications*
http://people.maths.ox.ac.uk/tlyons/

**Dr Marek Musiela,** Deputy Director, Oxford-Man Institute
*Commercial exploitation of term structure models*
http://www.oxford-man.ox.ac.uk/people/marek-musiela

**Professor Michael A. Osborne**, Associate Professor in Machine Learning, Department of Engineering
*Engineering: machine learning*
http://www.robots.ox.ac.uk/~mosb/

**Professor Mason Porter**, Professor of Nonlinear and Complex Systems, Mathematical Institute
*Complex networks*
http://people.maths.ox.ac.uk/porterm/

**Professor Steve Roberts**, Professor of Machine Learning, Department of Engineering Science
*Machine learning approaches to financial data analysis, Bayesian statistics*
http://www.robots.ox.ac.uk/~sjrob/

## 4. Business

**Professor Peter Grindrod**, Mathematical Institute
*Mathematics, Social Networks, Behavioural Analytics*
http://www.maths.ox.ac.uk/people/peter.grindrod

**Professor Howard Smith**, Associate Professor in Economics, Department of Economics
*Industrial Organization, Empirical Industrial Organization, Applied Microeconometrics, Differentiated Products*
http://www.economics.ox.ac.uk/Academic/howard-smith

**Professor Peter Tufano**, Peter Moores Dean and Professor of Finance, Saïd Business School
*Consumer finance; risk management and corporate financial engineering; and mutual funds*
http://www.sbs.ox.ac.uk/community/people/peter-tufano

## 5. Biomedical and Life Science

**Dr Gavin Band,** Statistician, Nuffield Department of Medicine
*Statistician on large datasets*
http://www.well.ox.ac.uk/gavin-band-2

**Professor Sir John Bell**, Regius Professor of Medicine, University of Oxford
*Genetics and genomics*
http://www.sbs.ox.ac.uk/community/people/sir-john-bell

**Professor Rory Collins**, Professor of Medicine and Epidemiology, Nuffield Department of Clinical Medicine
*Bioinformatics and Statistics, Cardiovascular Science, Clinical Epidemiology, Genetics and Genomics*
http://www.cardioscience.ox.ac.uk/bhf-centre-of-research-excellence/researcher-profiles/rory-collins

**Professor Jim Davies,** Professor of Software Engineering, Department of Computer Science
*Semantics-driven technology for medical research*
http://www.cs.ox.ac.uk/Jim.Davies/

**Professor Peter Donnelly,** Professor of Statistical Science, Department of Statistics
*Applications of probability and statistics in genetics, Gene mapping, population genetics*
http://www.stats.ox.ac.uk/people/academic_staff/peter_donnelly

**Professor Chris Holmes**, Professor of Biostatistics, Department of Statistics
*Statistical genomics and genetic epidemiology*
http://www.stats.ox.ac.uk/~cholmes/

**Professor Peter Holland,** Department of Zoology
*Genomics and evolution*
http://www.zoo.ox.ac.uk/people/view/holland_pwh.htm

**Professor Jonathan Marchini**, Professor in Statistical Genomics, Department of Statistics
*Statistical genetics, genome-wide association studies, Bayesian statistics, image analysis*
http://scholar.google.co.uk/citations?user=YdS6szoAAAAJ&hl=en

**Professor Gil McVean**, Professor of Statistical Genetics, Department of Statistics
*Computational statistics and genomics*
http://www.stats.ox.ac.uk/people/academic_staff/gilean_mcvean

**Dr Simon Myers**, University Lecturer in Bioinformatics, Department of Statistics
*Statistical genetics*
http://www.stats.ox.ac.uk/~myers/

**Professor Chris Ponting**, Professor of Genomics, Department of Physiology, Anatomy and Genetics
*Intersection between disease genomics, computational biology, molecular mechanism determination*
http://www.dpag.ox.ac.uk/team/group-leaders/chris-ponting

**Professor Oliver Pybus,** Department of Zoology
*Evolution and infectious disease*
http://www.zoo.ox.ac.uk/people/view/pybus_og.htm

**Professor Sir Peter J Ratcliffe,** Nuffield Professor of Clinical Medicine, Head of the Nuffield Department of Clinical Medicine
http://www.ndm.ox.ac.uk/principal-investigators/researcher/peter-ratcliffe

**Dr Susanna-Assunta Sansone,** Associate Director of Life, Natural and BioMedical Sciences, Oxford e-Research Centre
*Data management, biocuration, standards, ontology*
http://www.oerc.ox.ac.uk/people/Susanna-Assunta-Sansone

**Professor Lionel Tarassenko**, Professorial Fellow in Electrical and Electronic Engineering, Department of Engineering Science
*Signal processing, biomedical engineering, patient monitoring*
http://scholar.google.co.uk/citations?user=Zs5pBlMAAAAJ&hl=en

## 6. Climate and the Environment

**Professor Myles Allen**, Professor of Geosystem Science, School of Geography and the Environment
*Climate prediction modelling*
http://www.geog.ox.ac.uk/staff/mallen.html

**Professor Tim Coulson,** Professor of Zoology, Department of Zoology
*Emergent behaviour: population dynamics, mathematical modelling*
http://www.zoo.ox.ac.uk/people/view/coulson_t.htm

**Professor Marian Dawkins,** Professor of Animal Behaviour, Department of Zoology
*Emergent behaviour: automated assessment of farm animal welfare*
http://users.ox.ac.uk/~snikwad/

**Professor Tim Guildford,** Professor of Animal Behaviour, Department of Zoology
*Emergent behaviour: sensing and tracking*
http://www.zoo.ox.ac.uk/people/view/guilford_tc.htm

**Professor Simon Hay**, Professor of Epidemiology, Department of Zoology
*Malaria Atlas project, infectious disease mapping*
http://www.zoo.ox.ac.uk/people/view/hay_si.htm

**Professor David Macdonald,** Director of the Wildlife Conservation Research Unit (WildCRU)
*Behavioural ecology*
http://www.wildcru.org/members/professor-david-macdonald-cbe-dsc-frse/

**Professor Andrew Markham,** Associate Professor in Software Engineering, Department of Computer Science
*Sensing and communication in extreme and challenging application*
http://www.cs.ox.ac.uk/people/andrew.markham/

**Dr Neil Massey,** Researcher, Department of Physics
*Climate prediction modelling*
http://www2.physics.ox.ac.uk/contacts/people/massey

**Dr Tamsin A. Mather,** Academic Fellow in Physics and Chemistry of the Earth and Environment, Department of Earth Sciences
*Earth Sciences: volcanoes and STREVA*
http://www.earth.ox.ac.uk/~tamsinm/

**Professor Tim Palmer**, Royal Society Research Professor in Climate Physics
*Weather and climate, climate physics*
https://www2.physics.ox.ac.uk/contacts/people/palmer

**Professor Barry Parsons**, Professor of Geodesy and Geophysics, Department of Earth Sciences
*Earth Sciences: Earthquakes without Frontiers*
http://www.earth.ox.ac.uk/people/profiles/academic/barry

**Professor David Pyle**, Professor of Earth Sciences, Department of Earth Sciences
*Tectonics, volcanoes and hazards*
http://www.earth.ox.ac.uk/people/profiles/academic/davidp

**Professor Steve Roberts**, Professor of Machine Learning, Department of Engineering Science
*Machine learning approaches to financial data analysis, Bayesian statistics*
http://www.robots.ox.ac.uk/~sjrob/

**Professor Ben Sheldon,** Luc Hoffman Professor of Field Ornithology, Department of Zoology
*Emergent behaviour: population dynamics*
http://www.zoo.ox.ac.uk/people/view/sheldon_bc.htm

**Dr David Wallom**, Associate Director, Innovation, Oxford e-Research Centre (OeRC)
*Energy and ICT, cloud computing, volunteer computing*
http://www.oerc.ox.ac.uk/people/david-wallom

**Professor Kathy Willis**, Professor of Biodiversity, Department of Zoology
*Ecological and environmental datasets*
http://www.zoo.ox.ac.uk/people/view/willis_k.htm

## 7. Astronomy

**Professor Joanna Dunkley**, Professor of Astrophysics, Department of Physics
*Cosmology, studying the origins and evolution of the universe*
http://www-astro.physics.ox.ac.uk/~Dunkley/Home.html

**Professor Chris Lintott**, Professor of Astrophysics & Citizen Science Lead, Department of Physics
*Galaxy formation, citizen science*
https://www2.physics.ox.ac.uk/contacts/people/lintott

**Dr Aprajita Verma**, Research Fellow: E-ELT, Department of Physics
*Telescopes, galaxy surveys*
https://www2.physics.ox.ac.uk/contacts/people/verma

## 8. Physical Sciences

**Professor Alexander A. Schekochihin,** Professor of Theoretical Physics, Department of Physics
*Theoretical astrophysics and plasma physics*
http://www-thphys.physics.ox.ac.uk/people/AlexanderSchekochihin/

**Dr Jeff Tseng**, University Lecturer in Physics, Department of Physics
*Large Hadron Collider and ATLAS Project*
https://www2.physics.ox.ac.uk/contacts/people/tseng

**Professor Mark Wilson,** Professor of Chemistry, Department of Chemistry
*Physical and Theoretical Chemistry, computational chemistry*
http://research.chem.ox.ac.uk/mark-wilson.aspx

## 9. Arts and Humanities

**Professor John Coleman**, Professor of Phonetics, Faculty of Linguistics, Philology and Phonetics
*Phonetics, speech technology, laboratory phonology and computational linguistics*
http://www.jcoleman.co.uk/

**Dr Kathryn Eccles**, Research Fellow, Oxford Internet Institute
*Digital humanities*
http://www.oii.ox.ac.uk/people/?id=138

**Professor David De Roure**, Professor of e-Research, Oxford e-Research Centre
*Interface with libraries and humanities*
http://www.oerc.ox.ac.uk/people/David%20De%20Roure

## 11. Facilities and Infrastructure

**Dr Andrew Richards**, Associate Director - Operations and Service, Oxford e-Research Centre
http://www.oerc.ox.ac.uk/people/richards-andy

**Professor Anne Trefethen**, Professor of Scientific Computing and Director, Oxford e-Research Centre
http://www.oii.ox.ac.uk/people/?id=93

**Dr David Wallom**, Associate Director-Innovation, Oxford e-Research Centre
*Cloud utilisation, research data management, green IT, ICT security)*
http://www.oerc.ox.ac.uk/people/david-wallom