The Alan Turing Institute Scoping Programme

THE ALAN TURING INSTITUTE

To inform the development of its research programmes, in the summer of 2015 the Alan Turing Institute launched a scoping process, inviting researchers and other stakeholders to submit proposals for possible research topics. Some 100 proposals were peer-reviewed to determine the 30 or so most promising topics, which were then discussed at specially convened workshops.

The result of this process is the following scoping report. It is unusual in its breadth and technical span, encompassing disciplines as varied as the digital humanities, social data science, computer science, mathematics and statistical science. Areas in which data science could potentially be applied are similarly diverse, ranging from sports analytics to health and the geosciences.

This report focuses explicitly on the UK. It illustrates not only the UK's scientific strengths and expertise but also our strong culture of collaboration. The university affiliations of workshop organizers ranged from Imperial College London in the south to Strathclyde University in the north and, of course, included the Institute's founding partner universities.

The next step is to take this map of opportunities based on existing expertise and build from it the field of data science in the UK. The Institute looks forward to encouraging many collaborations across this map in the coming years and to being able to report on the innovations that will follow. This is an exciting task given the potential rewards and speed of intellectual progression of data science and its potential application for the benefit of UK citizens and society more broadly. To quote a visionary remark of Ada Lovelace, the mother of computer science: in data science

"A new, a vast, and a powerful language is developed for the future use of analysis, in which to wield its truths so that these may become of more speedy and accurate practical application for the purposes of mankind than the means hitherto in our possession have rendered possible."

We hope we can build on the opportunities offered to us together and create a world full of data opportunities for the next generation.

Andrew Blake
Director

Howard Covington
Chair

Foreword to the scoping process

The scoping process of the Alan Turing Institute was initiated in 2015, with a general call for scoping workshops issued in May 2015. Some 100 proposals were submitted over three calls, and peer-reviewed in a process organized by the Institute's Interim Programme Committee (see below). About 30 workshops were selected, and ran from the start of September 2015 to the end of February 2016. Organizers and participants were drawn from across the UK and overseas.

As reflected in the membership of the Interim Programme Committee, proposals covered a wide range of disciplines. During their evaluation, opinions from disciplines with different traditions and cultures had to be balanced and assessed. It is my pleasure to thank the Interim Programme Committee members who ran the call, the researchers who contributed to the peer review of proposals and all those who submitted proposals in response to this call.

Sofia Olhede
Chair of the Alan Turing Institute Interim Programme Committee

Alan Turing Institute Interim Programme Committee

Simon Arridge	Computer Science, UCL
Graham Cormode	Computer Science, Warwick
Jon Crowcroft	Computer Laboratory, Cambridge
David Firth (Gareth Roberts)	Statistics, Warwick
Zoubin Ghahramani	Engineering, Cambridge
Mark Girolami	Statistics, Warwick
Ben Leimkuhler	Mathematics, Edinburgh
Terry Lyons	Mathematical Institute, Oxford
Helen Margetts	Oxford Internet Institute, Oxford
Steve Renals	Informatics, Edinburgh
Steve Roberts	Engineering, Oxford
Carola-Bibiane Schönlieb	Applied Mathematics and Theoretical Physics, Cambridge
John Shawe-Taylor	Computer Science, UCL
Robin Williams	Social and Political Science, Edinburgh
Patrick Wolfe	Statistical Science and Computer Science, UCL

Report on the 2015/6 scoping programme

Andrew Blake and Sofia Olhede

The Alan Turing Institute's scoping programme

The Alan Turing Institute is the UK's new national institute for data science. It has been established as an independent charity whose founding partners are the Universities of Edinburgh, Oxford, Cambridge, Warwick and UCL and the Engineering and Physical Sciences Research Council (EPSRC). To shape the direction of its research programme, the Institute ran a series of more than 30 scientific scoping workshops in the final quarter of 2015 and early 2016 (Appendix 1). These brought together academics and other stakeholders to discuss the scientific roadmap for the Institute and to map out the theoretical and applied areas that hold the most promise and align well with the Institute's context, mission and expertise. The events also provided an opportunity for participants to register an interest to be involved in the Institute's programmes.

In addition to the scoping programme, the Institute's partner universities held 12 summits focused on applications of data science, which provided valuable insights from industry and commerce. The Institute also ran an internal set of symposia, to build its awareness of responsible research and other crosscutting issues. These programmes are listed in Appendix 1.

National and international context

Several landmark publications, in the UK and internationally, have highlighted strategic opportunities and associated challenges in data science. These include "Science as an Open Enterprise" (Royal Society)[1], "Frontiers in Massive Data Analysis" (US National Academies)[2], "Digital Agenda 2014–2017" (Federal Government of Germany)[3], "Seizing the Data Opportunity" (UK Department of Business, Innovation and Skills)[4], "The Big Data Dilemma" (House of Commons Science and Technology Committee)[5], and the White House report "Big Data: Seizing Opportunities, Preserving Values"[6]. The UK is not alone in making investments in data science. Appendix 2 provides an overview of the global community of data science research centres.

Workshop themes

A range of coherent areas of inquiry emerged from discussions at the scoping workshops, and provide a possible foundation for the Institute's future work.

Algorithms and architecture co-design

In five years' time, data centres will be exploiting new networking hardware, better energy efficiency and new switch hardware. Building new data centres will create opportunities for innovation, as

[1] Royal Society. *Science as an Open Enterprise.* 2012. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

[2] National Academies. *Frontiers in Massive Data Analysis.* 2013. http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis

[3] Federal Ministry for Economic Affairs and Industry, Germany. *Digital Agenda 2014–2107.* 2014. https://www.digitale-agenda.de/Webs/DA/DE/Home/home_node.html

[4] Department of Business, Innovation and Skills. *Seizing the Data Opportunity: A strategy for UK data capability.* 2013. https://www.gov.uk/government/publications/uk-data-capability-strategy

[5] House of Commons Science and Technology Committee. *The Big Data Dilemma.* 2016. http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf

[6] White House. *Big Data: Seizing Opportunities, Preserving Values.* 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

current centres were built from existing technology (e.g. commodity hardware, TCP/IP networking). In particular, starting afresh will remove the need to redevelop code multiple times for different uses (e.g. for testing, large batch processing, and high-speed streaming). Next-generation systems will allow code to be written only once and then run, with the compiler stack adjusting. Data-processing systems have also become increasingly complex, calling for research into simplification. Storage and database design must be considered in this context. The challenge of building in security and privacy for storage and processing also needs to be addressed; this could include, for example, homomorphic encryption and enclave computing, while issues related to the scaling up of security mechanisms will also need to be considered. Systems should also be matched to the full lifecycle of the analysis and treatment of large data volumes.

There is scope to exploit developments in hardware capabilities, such as field programmable gate arrays (FPGAs), network interface controllers (NICs) and graphics processing units (GPUs), together with rack-scale computer architecture. Key questions include how these developments could support stream and graph processing beyond simple programming models such as MapReduce.

Developing capability in hardware and system architecture has potential to drive algorithm development for analytics. Opportunities exist to couple algorithm development with an architecture programme through co-design, in which architecture and algorithm development each influence the other. This would be likely to require input from multiple aspects of a data science programme, including algorithm developments, systems design, statistics and mathematics.

One candidate for co-design is work on data analytics. Ideally, such tools will be informed by the latest developments in computer architectures, and will allow such architectures to be optimised for data analytics workloads, both on the client side and in the cloud. Conversely, the engineering and implementation of algorithms in software should aim to exploit fully architecture and hardware, including central processing units (CPUs), GPUs and other coprocessors. The aim could be to run software on hardware and system architectures that have been co-designed, providing a showcase of the possibilities offered by integrated hardware/software co-design.

Another opportunity is provided by deep learning, where innovative new approaches are being applied to artificial intelligence. The coprocessors currently used are predominantly GPUs designed for graphics rather than learning; although GPUs customised for learning are emerging, there are possibilities for further customisation. Probabilistic programming, the design of new languages and runtimes with built-in inference and learning algorithms will also require well-matched architectures in order to generate inference and learning tools that can be widely used by the software industry.

Distributed machine learning addresses issues related to distributed storage and processing through use of distributed algorithms. Very little is known about the properties of, for example, federated optimization, which attempts to protect user privacy while dealing with distributed computation. Federated optimization violates most assumptions normally made in communication-efficient optimization, an issue that is likely to generate significant opportunities for research. Other popular algorithms, such as principal component analysis and approximate inference, need to be understood in the distributed setting, offering additional potential topics of inquiry.

Workshops:

Improving the data analytics process	Computing systems research for big data	Distributed machine learning and optimization.	Deep learning	Probabilistic programming

Data analytics platforms

Research in platform and tool building could provide an essential pipeline of talent from foundational work in mathematics, statistics and computation, through computer systems research to applied work and engineering. Such developments could power the translation of ideas into practice and achieve societal impact.

Platform building should acknowledge the whole analytics process. Going from data to knowledge starts with wrangling, and simplifying data by removing anomalies, interpolating or otherwise, treating missing observations, and extracting features or summaries. Platform building would ideally automatically handle such early data cleaning, and make it less analyst dependent.

Platforms also could benefit by building in methods based on formal logic. There is, for example, interest in combining logic and learning in relational databases. Platforms could also automate other aspects of analysis; for example, recent research in probabilistic programming takes a well-established line of research in automated inference and combines it with programming language research, creating opportunities to broaden considerably the adoption of learning and modelling technologies in mainstream software development.

Many forms of analytics are underpinned by mathematical optimization that enables model parameters to be determined automatically from data, and estimates to be computed from models. This is critical when the models and/or the datasets are sizeable and computation needs to scale up. An opportunity exists to apply this fundamental research in tools for popular platforms such as R and Python, or even a new platform, exploiting the power of modern cloud computing. This could enable analytics users to take advantage of state-of-the-art sophisticated inference techniques, considerably enhancing the impact of analytics methods.

Workshops:

Improving the data analytics process	Logical foundations of data science	Distributed machine learning and optimization.	Probabilistic programming

Predictive modelling

This strand of research is in contrast with model-based inference, in which a constructive model is built to explain a particular set of data and from which the data could in principle be synthesised. Predictive modelling encompasses black box inference methods, such as neural networks and deep learning systems, and the technologies needed to support them. These methods are central to the processing of large volumes of data needed to train systems for automated analysis of text, video and audio streams. It is particularly interesting to teach systems to handle multiple modes of data simultaneously – for example, text annotation of audio or video, or aligning audio with video. There are opportunities for collaboration with the British Library, which holds important data archives, including the history of the UK view of the worldwide web, which the British Library stores as a national library of deposit, and an extensive audio archive currently being digitised.

The UK has made important contributions to recent innovations in this area, including deep learning methods that use large volumes of data to train neural nets. This approach has also been successfully combined with reinforcement learning to great effect. Outstanding problems include

determining provable properties of such algorithms and coordinated approaches to data repositories. There is clearly a need to train more deep learning researchers, given their current market demand. Techniques from distributed machine learning and optimization can also be utilized in this setting.

Workshops:

Understanding multimodal data at scale	Distributed machine learning and optimization.	Deep learning

Events and anomalies

Recognizing unusual or anomalous behaviour is key to the analysis of large collections of data. Developments in this area have applications in privacy and security for recognizing intrusions. Another important application is monitoring of the extensive telemetry data generated by large industrial machines used in areas such as mining and by turbines and transportation systems. There are also connections to healthcare monitoring and sports analytics, cyber-security and financial technologies. Development of new approaches will depend on the availability of heterogeneous and well-curated collections of data, and will require close interactions between those working on analysis and algorithm development.

This topic encompasses sketching and streaming algorithms, as well as mathematical models and characterisations of change or anomaly detection. It represents an opportunity to connect development of algorithms with an understanding of their mathematical performance. Good data wrangling and pre-processing are also important.

A range of expertise could contribute to this area, which lies at the interface between computational and mathematical sciences, from sketching and streaming to change-point detection. There is also potential to exploit functional and harmonic analysis in mathematics, to recognize disconnected or abrupt changes in behaviour.

Workshops:

Improving the data analytics process	Anomaly & change detection in streaming big data	Data protection and security at scale

Geometry and topology of data

Mathematical innovations have the potential to transform data science. Observed structure in data sets can be viewed in terms of the individual mechanisms that generated those data, but an analysis of the mathematical properties of such mechanisms can also reveal underlying similarities. This perspective is important to data science, generating a better understanding of which data structures are similar and can therefore be viewed as having closely related properties and underlying characteristics. It draws on underpinning disciplines such as combinatorics, geometry and topology to identify common features, as well as linear algebra, statistics and machine learning to extract conclusions from unified representations. This approach also provides a potential theoretical underpinning for analytics.

There are many open problems in characterising models in terms of their geometry and topology. There is a significant gap between our understanding of the mathematical description of structure and how best to use this understanding to draw conclusions from data. The UK is well placed to

address this issue, with its expertise in algebraic topology, rough path theory, mildly non-Euclidean data, combinatorics for networks and functional data analysis.

Work in this area could find practical application in areas such as finance, neuroscience and computational anatomy. There is an opportunity to bring together theoretical data scientists with applied data scientists and their collaborators, to develop new practically motivated analysis techniques and to generate theoretical advances in the mathematical sciences.

Workshops:

Networks & big data	Topological data analysis (TDA) – theory, computation and application	Developing the foundations of learning for non-Euclidean objects.

Indexing, labelling and retrieval

This is a core area for machine learning technologies, with significant opportunities for innovative research and practical application. One area is automated analysis of text, video and audio streams. A particular challenge in this area is that much of the richest data are owned by companies and regarded as a valuable asset. It is difficult for academic researchers to gain access to these data sets – Twitter being a notable exception. In this respect, opportunities for collaboration with the British Library are particularly appealing, given its important data archives.

Data in this setting are often multimodal, spanning text, images, video recordings, sensor recordings and logs of human interactions. Analysis of such varied forms of data requires interdisciplinary groups, applying techniques from natural language processing, machine learning, computer vision, speech processing, digital humanities and other data analytics such as statistics. Areas of potential impact include the media, urban analytics, libraries and archives, and interactive systems.

These technologies could also have important applications in healthcare, particularly in medical imaging. Multimodality is inherent across different imaging technologies, such as computed tomography, magnetic resonance imaging (several different imaging modalities) and ultrasound, with joint image registration and interpretation core problems. There is also a broader technical context, with a wide range of medical data stored in electronic patient records. That broader view of data, including imaging data, has relevance to clinical decision-making, and draws in other aspects of data science such as modelling, inference, databases and privacy, offering opportunities for interdisciplinary collaborations. There is significant UK expertise in organizations that specialize in imaging, as well as many initiatives inside UK universities.

Workshops:

Understanding multimodal data at scale	Big data and big problems in image-based healthcare technology.

Internet of things

The internet of things is a broad theme of potentially major importance to society[7]. It draws together instrumentation in diverse forms – in cities measuring pollution, traffic and weather, on vehicles, in smart devices and wearables – and requires interpretations to be made of diverse sets of

[7] Government Office for Science. *Internet of Things: Making the most of the second digital revolution*. A report by the UK Government Chief Scientific Adviser. 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/409774/14-1230-internet-of-things-review.pdf

signals and patterns. It is an integrative endeavour of interest to numerous agencies and programmes, including the Future Cities Catapult and EPSRC programmes. It is also of considerable interest to government for public services and to industry for next-generation machines. To be effective in this area, researchers would need to work with the various agencies to create bridges between methodology and practical problems and data sources.

The internet of things highlights a diversity of technical themes. For example, anomaly and change detection in temporal data is challenging because of the one-sided, unbalanced nature of the available training data, which are almost entirely devoted to detection of negative events. Because sensors are highly heterogeneous, in the type or physical dimensioning of their signals and in their spatiotemporal resolutions and sampling patterns, expertise in multimodal data analysis is needed. Internet of things systems may operate at an individual level, as in personal fitness devices, or operate at much larger scales, such as city-wide deployment in congestion and route-planning control systems and in large-scale engineering systems. Lastly, collation of people's personal data raises important issues related to the ethics of data distribution and the security of data transmission in a large and heterogeneous environment, involving both client devices and cloud computing.

Workshops:

Understanding multimodal data at scale	The ethics of data science	Deep learning	Anomaly & change detection in streaming big data	Data protection and security at scale

Large-scale optimization

Optimization is key to many modern data science problems, such as likelihood inference, penalized likelihood inference, inverse problems and other problems formulated in terms of a maximum and/or minimum to gain information from observed data. Optimization problems are difficult to disentangle from underlying data architectures. A challenging issue is how best to solve an optimization, given the sometimes fixed data architecture.

This field faces many challenges. For example, optimization over large numbers of variables remains challenging; initial developments led to the formation of the field of high-dimensional data analysis, where underdetermined systems are solved using penalization. There is recognition of the need to combine algorithmic development of optimization with selection of an appropriate optimization criterion. In particular, choosing tuning parameters remains an important and difficult challenge.

Distributed optimization also remains a pressing area for future development. Distributed storage, and hence distributed algorithms, are now increasingly widely used and need to be studied further. Very little is known about their theoretical properties. Work in this area is likely to result in novel paradigms for distributed computing, and the scaling up of mathematical and statistical developments via new computer systems and algorithms is one of the most dynamic areas of data science globally.

In inverse problems also the balance between statistical properties and computational constraints is important. Current research topics include combining statistical and numerical approaches to solve inverse problems, coping with model uncertainty and using sparse approximation techniques.

Workshops:

Theoretical and computational approaches to large scale inverse problems	Statistical and computational challenges in large-scale data analysis	Distributed machine learning and optimization.

Logic and machine learning for databases

There is an opportunity to combine logic and machine learning in database development. Commercial users normally access data via relational database systems that are based on logical models of data in the form of schemas. Schemas capture the structure of data in a form that is human readable. They can be used to convert unstructured data, for example free text, into the structured form that database systems use. Schemas can also be used to detect illegal structures in data and possibly to correct them. Schemas have to be constructed by hand, however, which is laborious. In machine learning, it is uncommon to use relational databases. Learning systems are most often based on pattern recognition, probabilistic modelling or neural networks but learning in a logical context is currently unusual, though not entirely unknown.

The excitement in this area of research is twofold. First, the need to build schemas by hand significantly slows down development, and machine learning approaches might enable schemas to be constructed automatically. Secondly, bringing commercial practice in relational databases into the realm of machine learning could increase the practical impact of machine learning. These interactions could also address the issue that much of the practical effort in extracting value from data is not in the analysis itself but in the data wrangling that takes disorganised or otherwise inappropriately structured data and converts it into a usable form. Machine learning that is normally directed at analysis could thus be applied in other areas of data processing. Work in this area also has the potential to be translated into practice via the building of tools and platforms.

Workshops:

Improving the data analytics process	Logical foundations of data science

Models, inference and learning

This strand of work addresses learning and inference from observed data points. The field is based on the combined expertise of statistics and machine learning, strengthened by other fields such as inverse problems, which use regularization to make stable estimates for model selection and parameter estimation.

Machine learning is traditionally strong in the UK. UK-based specialists have made important contributions to the field, especially related to kernel methods. In addition, the UK has strengths in computational inference, based on a legacy of computational algorithms for Monte Carlo methods, and in likelihood inference. For large data sets, likelihood inference and computational methods based on evaluating likelihood become less feasible and alternative methods are needed, so that computational constraints can be rationally balanced with statistical performance.

There are many exciting topics to explore. The topic of probabilistic numerics, for example, addresses computation using the mechanisms of inference. Outstanding challenges include the development of algorithms that achieve optimal rates of numerical approximation and low variance, as well as partial differential equation and ordinary differential equation solvers for engineering problems. There is an opportunity for collaborations across the mathematical sciences and machine learning, to create methods founded in both fields.

Evaluating likelihood is key to most rational approaches to quantifying uncertainty. Likelihood-based analysis becomes difficult when sets of observations become too massive for likelihood to be evaluated, the model used is too complicated, the data are of poor quality, or the model is misspecified. Avenues of exploration include approximate methods for intractable likelihood, including approximate Bayesian computation, variational methods, as well as new directions for sequential Monte Carlo and particle Markov chain Monte Carlo (MCMC) methods. Other adaptions of MCMC are promising, including gradient-based, non-reversible, pseudo-marginal and adaptive approaches.

There is significant scope to develop new theory and methods in areas that lie between the fields of statistics and computer science, requiring new theoretical tools as well as new algorithms to manage modern volumes of data. High-dimensional inference is one such area, where the number of parameters or the complexity of the generating mechanism grows with the observed sample size. Heterogeneity and personalized inference is another topic requiring significant innovation to draw repeatable conclusions from sets of observations. Often such observations can be thought of as large arrays or tensors with missing observations. Another interesting area relates to general inference methods for functional and object data. While there are approaches for fixed modalities, combining information across many modalities is an important and challenging problem.

Information theory treats the understanding of information from an axiomatic perspective, which is especially relevant in sensing problems. Information theory provides methods to understand trade-offs between different desirable conditions, such as trading privacy versus utility, or good representation for a single data set versus any possible underlying population. Tools from information theory can be used to obtain a better understanding of fundamental properties of different algorithms in terms of large sample performance.

Workshops:

Probabilistic numerics	Intractable likelihood	Statistical and computational challenges in large-scale data analysis
High-dimensional statistical models & big data: methodology & applications	Probabilistic programming	Information-theoretic foundations and algorithms for large-scale data inference

Networks

The mathematical foundations of graphs and networks are key to problems as varied as understanding large volumes of relational data, designing complex computer architectures, and representing phenomena in computational social science.

The agenda for this area stretches from problems in combinatorics, defining the rich mathematical frameworks that allow large networks and their features to be compared and modelled, to inferring given models and interpreting data summaries for applications. A key challenge is understanding sparse and heterogeneous networks, typical of most real-world networks. There is also a need for statistical models of network structure, inspired by complicated real data sets, with structured behaviour, which will require substantial mathematical innovation.

Algorithms for processing graph-structured data are being developed, particularly to match data to local network structure. Networks are also suitable structures for software systems that scale. Network data structures could also be used in the design of new big data processing paradigms. In

social science, a key topic is explaining patterns of influence in social networks. Here the aim is to use tools to construct an analytic framework, describing the mechanisms of social interactions.

The UK is well represented in this field, with strengths in algorithms and mathematics for networks, in graph processing, and in related social science disciplines. The area is highly interdisciplinary, lying at the nexus of several disciplines.

Workshops:

Social data science	Networks & big data	Computing systems research for big data

Physical sciences

The physical sciences have produced massive volumes of data. CERN alone processes one petabyte of data every day and has developed distributed approaches to its complex and intensive computational requirements. Other fields, such as astrophysics, materials science, chemistry and geoscience, need to develop mechanisms to implement large-scale data acquisition, with appropriate experimental design, data architectures and analytics techniques. There is also a need for numerical simulation methods, requiring a combination of algorithm development, numerical analysis and statistical computing.

Geophysics has a particular need for inversion and imaging techniques, able to cope with very large volumes of data, thus intersecting with inverse problems and statistical methods for penalization. Sensing and sensing technologies may have similarities with the technologies developed for the internet of things. A number of methods used are generic, rather than area-specific, and there is considerable scope to apply mathematical approaches to solve data-acquisition problems in the earth sciences.

Other models in the physical sciences are based on the use of partial differential equations, such as data assimilation for weather forecasting and continuum mechanics. Incorporating data within such models is challenging and an open problem. Additional numerical problems requiring considerable computational resources include data-driven chemistry and materials science.

The UK has centres of excellence in extreme scale computation and numerical simulation, including Science and Technology Facilities Council sites such as the Hartree Centre, Distributed Research utilising Advanced Computing (DiRAC) and the GridPP collaboration.

Workshops:

Partial differential equations for modelling, analysing and simulating data rich phenomena	Big data in geoscience	The challenges of data intensive & extreme scale simulation in physics, materials & chemistry

Privacy and security

There are opportunities in the area of privacy and security both to pursue research and to contribute to national debate and formulation of policy and practice[8].

[8] House of Commons Science and Technology Committee. *The Big Data Dilemma.* 2016. www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf

Outstanding research challenges include the need to ensure or quantify privacy, especially understanding the potential for re-identification of individuals, and to avoid hidden discrimination. Furthermore, the increasing complexity of algorithms means new methods are needed to ensure responsibility and accountability, and to ensure that algorithms are designed to operate ethically. Algorithms may need to be ethically audited to ensure that are not inherently discriminatory.

Key outstanding research challenges in security include managing security, especially in the context of potential day-to-day attacks. Detecting intrusion and abuse rapidly and at scale is also vital. The use of analytics developed especially for security is an area of current interest and has overlaps with anomaly detection.

This area also has relevance to the digital economy, especially the design of secure systems for peer-to-peer transactions. Secure and anonymous management of personal data becomes key in such settings. Cryptography and transaction protocols are vital for secure peer-to-peer transactions and for decentralised currencies such as bitcoin, as well as for decentralised ledgers for recording various kinds of authenticated information.

The agenda for this area stretches from social science and particularly ethics through to areas of computing such as privacy in large databases, including differential privacy (related to personal information that is not explicitly published but is inferable indirectly).

Workshops:

The ethics of data science	Data science for the digital economy: Digital currencies and peer-to-peer economics	Data protection and security at scale

Social data science

Social data science provides an opportunity to forge links between the technical world of mathematics, statistics and computing, and the understanding of human behaviour[9]. The study of human behaviour in a connected world is a large part of social data science. As citizens are now in an environment of nearly ubiquitous sensing, where almost all actions and interactions are recorded, the potential for understanding human behaviour is considerable. Social data science will need to build on algorithms, understanding of human behaviour and application-specific knowledge.

Work in this area needs to consider responsible design from the outset, embracing data science ethics to identify what users will find acceptable[10]. Other research themes at the Institute will also need to consider these ethical and experimental design issues, highlighting the influence of the social sciences on mathematical and computational science. It is also increasingly being recognised that algorithms and use of analytics influence society and the way we understand how algorithms and decisions based on data intersect. This in turn emphasises the need to incorporate feedback into the design of algorithms.

Conversely, there is influence in the reverse direction, with new methods in statistics and computation revolutionizing the way that social science experiments are carried out and hypotheses are generated, based on the modern tools of robust inference at scale, crowd sourcing, and large-scale modelling of human behaviour. The availability of ever-increasing volumes of human-generated data[11], coupled with epidemiological databases, has the potential to generate new

[9] Lazer D *et al.* Computational social science. *Science*. 2009;323(5915):721–3.
[10] House of Commons Science and Technology Committee. *The Big Data Dilemma.* 2016. http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf
[11] UK Data Forum. *UK Strategy for Data Resources for Social and Economic Research 2013–2018*. 2013. http://www.esrc.ac.uk/files/research/uk-strategy-for-data-resources-for-social-and-economic-research/

understanding and ultimately deliver social benefits. Synergies can be found with other large-scale UK initiatives such as the Economic and Social Research Council (ESRC) Administrative Data Research Centres.

Workshops:

Social data science	The ethics of data science	Networks & big data	Algorithm society	Data protection and security at scale

Monitoring and analytics

Urban analytics is also a very promising area, with a strong intersection with monitoring. Human activities generate vast swathes of information, both rich and complex. For example, in city life, data can be acquired through online social media, telecommunications, geolocation, crime, health, transport, air quality, energy, utilities, weather, CCTV, wi-fi usage, retail footfall and satellite imaging. One challenge is how to adapt existing mathematical technologies to integrate and analyse these data. Another is to develop specially adapted tools to collect data in urban environments. Adaption opportunities exist in sectors focused on customer interactions, and possible collaboration partners could range from local councils to commercial entities. Exploring the sensitivity to scale is one specific question that could be addressed, and how the behaviour of cities depends on their size. Exploring the power of crowd sourcing, and the inherent biases in such sampling, would lead to both practical innovation and deep theoretical challenges.

Monitoring is also important in environmental applications. Such applications present considerable technical challenges, intersecting with anomaly and change detection in streaming big data, understanding multimodal data at scale, and statistical and computational challenges in large-scale data analysis. Underpinning methodologies include statistical models and methods for time series, point processes and random fields, as well as the more novel area of functional data analysis. Analysis is hampered by distributed data storage and the need to compress data before storage.

Monitoring and analytics also overlap with sensing for data-rich sports. Challenges are diverse in the latter area, ranging from decision-making online from real-time data streams to optimal design for training regimes based on monitoring. There are clear intersections with understanding multimodal data at scale as well as with anomaly and change detection. Overlaps also exist with privacy and security, as openness in elite sport can lead to a loss of competitive advantage. Research relationships could be established with some of the major web-based training-analytics platforms, such as Strava and TrainingPeaks, and with open-source platforms such as Golden Cheetah. These relationships could potentially provide access to mass data from non-elite athletes, opening up opportunities for research on a larger scale. The area is also an opportunity for public communication of data science, raising awareness of the field.

Workshops:

Data science for data-rich sports	Urban analytics	Anomaly & change detection in streaming big data

Public sector

The public sector presents enormous opportunities for the application of data science. For example, many cultural endeavours generate abundant multimodal data such as music, speech and video. The importance of methods for the conservation and preservation of the world's digital heritage is outlined in the UNESCO charter on the protection of digital heritage[12]. Typically data sets are noisy, large scale and very heterogeneous. Their analysis will require new models of data heterogeneity and novel methods of analysis. In addition, when inferences are made about global collections, the problem of preferential sampling must be addressed, as many collections have not been digitised, and appropriate meta-analysis methods will be required. Automated methods to incorporate scholarly protocols for interpretation are also a potential area of investigation. The British Library has both expertise and extensive sources of data in this area, offering opportunities for collaboration.

Energy drives the modern economy. Understanding the energy system means modelling and predicting both energy production and energy consumption. This understanding is of importance to system operators, policy-makers, regulators and the private sector. The data science challenges span data gathering, data processing and data use for operations and investment, as well as data valuation and uncertainty quantification. Drawing from the UK Government's National Policy Statements for Energy Infrastructure, the need for advances in this area is evident, as is also made clear by the recent Paris agreement on climate change[13].

Work in this area would have close links with social data science, as well as privacy and security.

Workshops:

The foundations of social data science	Cultural and heritage Informatics	Whole-energy systems

Financial services

There are many opportunities at the interface of data science and financial services. Key research topics include blockchain technologies, privacy and cybercrime, and the Financial Conduct Authority's Project Innovate initiative. A major challenge for public research is access to data, where secure gateways have the potential to play an important role. Important public documents in this arena include the Blackett reviews from the Government Office for Science[14]. Financial modelling and forecasting are linked to multivariate and time series methods, both areas undergoing significant innovation in econometrics.

The peer-to-peer digital economy is growing rapidly. An Accenture report[15] estimates that US$75 million was invested in blockchain in 2015, up from US$30 million in 2014, and investment is likely to hit US$400 million in 2019. Applications such as cryptocurrencies, the sharing economy, peer-to-peer lending, distributed ledgers, crowdfunding, resource-sharing platforms and digital marketplaces are developing rapidly. Risks include new forms of fraud, information asymmetry and increasing

[12] UNESCO. *Charter on the Preservation of Digital Heritage.* 2003. http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html
[13] United Nations. Framework Convention on Climate Change. http://unfccc.int/2860.php
[14] Government Office for Science. *FinTech Futures: The UK as a world leader in financial technologies.* A report by the UK Government Chief Scientific Adviser. 2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/413095/gs-15-3-fintech-futures.pdf
[15] Accenture Consulting. Blockchain-Enabled Distributed Ledgers: Are investment banks ready? 2016. https://www.accenture.com/gb-en/insight-blockchain-enabled-distributed-ledgers-investment-banks

systemic risk. As peer-to-peer systems are highly complex and therefore likely to lead to unanticipated and altered behaviours, applying data science methods in this setting is challenging, and will require bespoke adaptation. As individuals are increasingly able to engage directly in transactions without mediation by banks or other organizations, the societal importance of this area is clear.

Financial modelling and forecasting are based on traditional methods in time series analysis, as well as stochastic processes, but adapted to large numbers of time series observations, making both models and observations high-dimensional. Recent developments include use of Twitter data to capture and exploit market sentiment, providing links to analysis of multimodal data.

Customer-facing sectors have already been adapting advanced data science methods for problems such as recommender systems, optimizing the pricing of services, determining sector areas for promotions, and spatial analysis to guide localization of new initiatives. These areas all raise novel mathematical and computer system challenges. There is also potential for the analysis of open data for these types of applications. Use of customer data creates opportunities for personalized initiatives, but also raises privacy and security issues.

Workshops:

Cultural and heritage informatics	Data science for the digital economy	Statistical and computational challenges in large-scale data analysis	Networks & big data

Health and biomedicine

Health and biomedicine are sources of large and complex data sets with complicated patterns and structure. High-throughput biology and precision medicine are being driven by technological advances such as massive parallel sequencing, automated microscopy and wearable medical devices. Analysing such data introduces a number of technical challenges, such as modelling of correlated and high-dimensional data, data integration and data heterogeneity. Explanatory as well as predictive power are important assessment criteria for posited generative mechanisms in this field. Outstanding challenges include the development of effective software for analysis and the need for multi-stakeholder collaboration across industry, the charitable sector, research councils and the academic sector. We again note complementary expertise in the Francis Crick Institute, the Farr Institute of Health Informatics Research, the Wellcome Trust Sanger Institute and the EMBL European Bioinformatics Institute, as well as many bodies inside UK universities.

Healthcare[16,17] is an important source of data. Better use of data has the potential to improve healthcare provision and disease prevention. Use of healthcare data raises important questions related to data protection and security at scale, as well as the ethics of data science. Areas to develop potentially include the construction of new methodologies for the analysis of high-dimensional heterogeneous datasets (mix of continuous, numeric, ordinal, categorical and free text data, all of which have an irregular time component) and new approaches to deal with complex non-random missing data. This strand of development could focus on mental health, cancer, metabolic diseases, ageing, and neurodegenerative and chronic diseases, and their impact on quality of life and the wider economy.

[16] Association of the British Pharmaceutical Industry. *Big Data Road Map*. 2013. http://www.abpi.org.uk/our-work/library/industry/Pages/big-data-road-map.aspx

[17] Center for US Health System Reform Business Technology Office. *The 'Big Data' Revolution in Healthcare: Accelerating value and innovation*. 2013. http://www.mckinsey.com/~/media/mckinsey/industries/healthcare%20systems%20and%20services/our%20insights/the%20big%20data%20revolution%20in%20us%20health%20care/the_big_data_revolution_in_healthcare.ashx

Imaging is an area of UK strength. It is also an underpinning technology for healthcare. Key challenges in imaging that could be addressed through data science approaches include diagnosis and prognosis, patient stratification for drug development and use, basic understanding of disease, and image-guided intervention. Large amounts of imaging data are available from routine use, the value of which has not yet been fully exploited. The opportunities for development are manifold and include bespoke analytics tools for medical imaging problems using machine learning and computational statistics, as well as automated quality control and harmonization of data acquired from different sources. Novel analytics methodologies have the potential to be applied in multiple areas, including non-invasive cancer imaging, imaging for neurological disease, orthopaedics and ophthalmology.

Workshops:

High-throughput biology and precision medicine	Data science for data-rich sports	Data-intensive healthcare	Image-based healthcare technologies

Appendix 1: Institute scoping programme workshops and associated events

For more details, see the Institute's events pages[18].

Ref No	Workshop title
4	Data science challenges in high-throughput biology and precision medicine
5	Probabilistic numerics
6	Data science for data-rich sports
7	Intractable likelihood
10	Understanding multimodal data at scale
12	Opportunities and challenges for data-intensive healthcare.
13	Theoretical and computational approaches to large-scale inverse problems
14	The foundations of social data science
15	The ethics of data science: the landscape for the Alan Turing Institute
16	Partial differential equations for modelling, analysing and simulating data-rich phenomena
19	Networks and big data
20	Statistical and computational challenges in large-scale data analysis
26	Improving the data analytics process
29	Advancing cultural and heritage informatics
30	Big data and big problems in image-based healthcare technology
37	Data science methods and tools for urban analytics
38	Computing systems research for big data
39	Algorithm society
48	Logical foundations of data science
50	High-dimensional statistical models and big data: methodology and applications
51	Topological data analysis (TDA) – theory, computation and application
52	Data science for the digital economy: digital currencies and peer-to-peer economics
54	Distributed machine learning and optimization
56	Deep learning scoping workshop
59	Anomaly and change detection in streaming big data
62	Developing the mathematical foundations of learning for non-Euclidean objects
72	Probabilistic programming
73	Data science applications in whole-energy systems
74	Data protection and security at scale
80	Big data in geoscience
82	The challenges of data-intensive and extreme-scale numerical simulation in physics,

[18] https://turing.ac.uk/past-events/

	materials and chemistry
87	Information-theoretic foundations and algorithms for large-scale data inference

Summits organised by the Institute's partner universities focusing on applications of data science

Summit title
Summit for the citizen/customer-facing sectors
Data science for health
Data science for media
Financial summit
Data analytics for credit risk
Big data for small and medium-sized enterprises
Big data in the physical sciences
Future cities
High-value manufacturing
Data science for government and policy
Energy summit
Privacy summit

Symposia run by the Institute to build awareness of responsible research and other cross-cutting issues.

Symposium title
Reproducibility, sustainability and preservation
Theoretical foundations of visual analytics
Responsible, ethics-aware research and innovation in data science

Future related events.

Follow from	**Forthcoming summer workshops**
	Innovative finance for social good in the data economy
	Semantic web and data integration
7, 50	MCMC and diffusion techniques
37	Urban analytics
26	Improving the data analytics process
38	Systems research agenda

Forthcoming summer symposium
Communicating machine learning

Appendix 2: International centres working in data science

In the USA, a number of third-sector initiatives have fuelled the growth of data science. The Gordon and Betty Moore Foundation with the Alfred P. Sloan Foundation have supported the University of Washington, the University of California, Berkeley, and New York University for work in this area. The University of Washington initiative seeks to advance data-driven discovery by building on an existing eScience Institute. The University of California, Berkeley, was funded to establish the Berkeley Institute for Data Science to bring together scientists from natural and social sciences with computer scientists and mathematicians. Finally, New York University established a new collaboration for data management and data analytic techniques, focused on methodological advances. Initiatives focusing on the theoretical foundations of data science include the Simons Institute for the Theory of Computing and the Center for Discrete Mathematics and Theoretical Computer Science at Rutgers University. Several other data science centres have been established in the USA, listed in the Table below.

In Europe, multiple data science centres have been established. These include the Insight Centre for Data Analytics in the Republic of Ireland, the Paris Saclay Centre for Data Science in France, Instituto Mixto de Investigación UC3M-Santander on Financial Big Data in Spain. Germany has several well-established centres, including the Max Planck Institute for Intelligent Systems, the practically oriented Fraunhofer Institutes, and the more science-oriented Helmholtz Institutes. Australia has expertise in big data in the information communications technology (ICT) sector at National ICT Australia (NICTA, currently subject to reorganization).

In Asia, important centres include the Fudan School of Data Science, the Beijing Institute of Big Data Research led by Beijing University, the Institute of Statistics and Big Data at Renmin University, Beijing, the Center for Statistical Science at Tsinghua University, a Big Data and Statistics Research Center set up by China's National Bureau of Statistics and the Shanghai University of Finance and Economics, and the Kawarabayashi Large Graph Project at the National Institute of Informatics, Japan.

International centres working in data science

Country	Name	Website
USA	Information Initiative at Duke University	http://bigdata.duke.edu/
USA	Michigan Institute for Data Science	http://midas.umich.edu
USA	Simons Center for Data Analysis	https://www.simonsfoundation.org/simons-center-for-data-analysis/
USA	Bloomberg campus of Cornell Tech	http://tech.cornell.edu
USA	Massachusetts Institute of Technology Institute for Data, Systems and Society	https://idss.mit.edu
USA	Network Science Institute at Northeastern University	http://www.networkscienceinstitute.org
USA	Center for Statistics and Machine Learning at Princeton University	http://sml.princeton.edu/
USA	Data Science Institute at Columbia University	http://datascience.columbia.edu
USA	Simons Institute for the Theory of Computing	https://simons.berkeley.edu
USA	Center for Discrete Mathematics and Theoretical Computer Science at Rutgers University	http://dimacs.rutgers.edu
USA	eScience Institute, University of Washington	http://escience.washington.edu/
USA	Berkeley Institute for Data Science	https://bids.berkeley.edu
USA	The NYU Center for Data Science	http://datascience.nyu.edu
Republic of	Insight Centre for	https://www.insight-centre.org

Ireland	Data Analytics	
France	Paris Saclay Centre for Data Science	https://www.universite-paris-saclay.fr/en/research/project/lidex-cds
Spain	Instituto Mixto de Investigación UC3M-Santander on Financial Big Data	http://www.uc3m.es/ss/Satellite/UC3MInstitucional/en/Detalle/Organismo_C/1371213288068/1371206581851/UC3M-BS_Institute_of_Financial_Big_Data_(IFiBiD)
Germany	Berlin Big Data Center	http://www.bbdc.berlin/start/
Germany	Max Planck Institute for Intelligent Systems	http://www.is.mpg.de/de
Germany	Fraunhofer Institutes	http://www.fraunhofer.de/en.html
Germany	Helmholtz Institutes	http://www.helmholtz.de/en
Australia	National ICT Australia	https://www.nicta.com.au
China	Fudan School of Data Science	http://www.sds.fudan.edu.cn/wp/
China	Beijing Institute of Big Data Research	www.bibdr.org
China	Institute of Statistics and Big Data at Renmin University	http://isbd.ruc.edu.cn/more.php?cid=16
China	Center for Statistical Science at Tsinghua University	http://www.stat.tsinghua.edu.cn/outlineDetail.jsp?initmenuid=37
China	Big Data and Statistics Research Center	http://ssm.shufe.edu.cn/Home/Index/content?id=88&cid=205
Japan	Kawarabayashi Large Graph Project at the National Institute of Informatics	http://bigdata.nii.ac.jp/wp/english/