Data Science, AI and the Turing Institute – workshop report

On 29th June 2018, around 80 academics attended the workshop to find out more about Data Science and Artificial Intelligence at the University of Birmingham, and for an update on our partnership with the Alan Turing Institute.

Following a number of presentations (for which, see slides provided), we had breakout sessions, which aimed at identifying relevant areas of research, finding commonalities of interest, and exploring ways in which future collaborative research can be encouraged.

A list of research areas identified both at the workshop and through previous scoping exercises is provided at the end of this report.

The main issues arising from the breakout sessions were:

Qualitative vs quantitative research. Much research that uses large data sets is numerical by nature and quantitative (often statistical) methods are used. However, there is also often a considerable amount of qualitative information, and domain knowledge, that is essential in provide the appropriate context for understanding this data. Finding an approach to joint research that includes both aspects would be of great benefit.

Match-making. While there is a lot of common interest across the university, finding exactly the right person to collaborate with is still difficult. Suggestions for "speed-dating" events might help. We could also collect a database on the web of interests to enable experts to be found more easily.

Training. Non-specialists need to be able to make sure they are using state-of-the-art techniques. However, it is difficult to keep up to date with these. Providing training pitched to a non-specialist audience could be very useful.

The need for computational expertise. Related to the above, many researchers would like a computational or mathematical expert to work alongside them in a project. Academics from those areas often do not have time or motivation for this, as there is often limited research opportunity for them. Having a pool of "research engineers" whose job is to help write and deliver projects could be invaluable.

Ideas for future, more focussed, activities:

- Workshops focussed around a particular research theme (including speed-dating, and "open mike" sessions).
- Establish mailing list and web page for sharing news, events, opportunities.
- Advertise in-house training more widely where appropriate.
- Advertise events at Turing Institute more widely. Establish travel fund to enable people to attend.
- Develop follow-on project via Institute for Advanced Studies.
- Continue discussions to establish a University Institute for Data Science and AI

Research related to Data Science and Artificial Intelligence

- **Data-centric Engineering**

  Manufacturing: use of machine learning and data mining in improving manufacturing processes; robotics and artificial intelligence techniques, e.g. for robotic disassembly and remanufacturing of complex products.

  Formulation engineering: Uncertainty-based models and physics-based models in combination to provide predictive capabilities in multiphase flow systems in industry (nuclear, oil and gas, fmcg).

  Energy storage: the Birmingham Centre for Energy Storage has a pilot scale liquid air energy storage facility which provides large experimental datasets.

  Materials characterisation: using imaging, simulation and computational modelling techniques to characterise different functional properties of materials, ranging from batteries to alloys to biomaterials.

- **Defence and Security**

  Causes of conflict: data analytics applied to quantitative social data in predicting ethnic war; monitoring the rise and influence of political corruption.

  Critical infrastructure security: security of rail network as it becomes digitised, security of the power grid, security of future autonomous vehicles.

  Blockchain and distributed ledger technologies: support key revocation in web-systems and end-to-end encrypted messaging systems; accountability of decryption.

  Cyber-security**:** research on human-related risks in cybersecurity for organisations.

- **Urban Analytics**

  Digital geographies the spatiality of the digital economy, the effects it can generate on space and the position of cities within spatial, complex networks.

  Temporary urbanism: research on planned or unplanned actions designed with the ambition of activating a space in need of transformation and impacts on the surrounding socio-economic environment.

  Environmental indicators: link genome dynamics to ecosystem functions – to understand how the information encoded in genomes gives rise to complex and responsive ecological systems.

- **Financial services**

  Simulated economies: use of big data sets in driving realistic simulations; artificial intelligence techniques used in understanding economic markets and auctions

  Economic forecasting: big data and AI-machine learning methods for forecasting purposes with applications to energy, real estate and the environmental economy.

  Understanding flows of money: use of open procurement data from the NHS and local government to understand flows of public monies to private companies and not-for-profits; capturing and classification of charity accounts data

- **Health**

  Health data: Development and evaluation of analytical approaches to address integration of multimodal, multi-dimensional, patient-level healthcare data, generating a trusted research platform that acquires, represents, and processes large, multi-dimensional and distributed research data to generate novel means by which to improve diagnosis, refine prognosis and personalise treatment.

  Imaging: Spanning the molecular through to medical and large-scale brain communication networks, pushing forward the technological capacity to generate, integrate and analyse the outputs of new and existing platforms to understand human and non-human function and action.

  Omics: Understanding the role of novel technologies – including genomics, metabolomics and proteomics – and the use of data generated by these to monitor, understand and predict outcomes.

- **Government**

  Government and third-sector organisations and funding: using machine learning to better understand the influence of charities' missions on their fundraising success;

  Tax administration: work with the NAO to explore the role of machine learning in transforming and enhancing the efficiency of tax administration and compliance agencies (e.g. HMRC).

  Crime: looking at a very big data base of crimes committed in two police areas and trying to determine factors that affect the solvability of crime in order to develop an algorithmic approach to allocation of resources to crime investigations.

  Political polling: using social media feeds to predict political polling outcomes, and various socio-demographic gradients therein.

- **System architecture for data science**

  Security in the cloud: Building on our work on confidentiality from cloud providers and on trusted computing platforms, we are investigating cloud-side hardware technologies to ensure security properties for cloud users.

  Programming language infrastructure for big data: compilation techniques for unconventional architectures (especially reconfigurable and distributed); domain-specific languages for machine learning ("abductive functional programming").

  Software Architecture: ultra-large scale autonomous data-driven software architectures.

  Ontologies: novel approaches for ontology-based multimodal multidimensional data integration.

- **Security and robustness in data science**

  Fairness, transparency, privacy: making data usages transparent to the subjects they apply to.

  Transparency for public key infrastructure: ensuring that certificate transparency is robust even in the face of very powerful attackers, harnessing data about web browsing habits to ensure that gossip protocols are effective in disseminating snapshots of the certificate transparency ledger.

Privacy-supporting data sharing for health care: looking at more powerful methods of privacy-supporting data sharing. Current data anonymisation techniques are not effective in supporting privacy, and also limit the usefulness of data since they do not allow cross-references to be made.

- **Machine learning and artificial intelligence**

  Machine learning: High dimensional data mining & machine learning, probabilistic modelling; dimensionality reduction; applications to health, neuroscience, biosciences, chemistry, physics.

  Natural computation: large-scale black box optimization, theoretical development, applications in engineering design, software engineering, medical imaging, materials characterization.

  Computer vision: hierarchical decomposition of scenes, object tracking, learning labels.

  Natural language processing: sentiment analysis, scientific document analysis, applications to health informatics.

  Planning and robotics: planning for autonomous vehicles, planning manipulation, applications in warehouse planning and in extreme environments (nuclear decommissioning).

- **Complex structure in data**

  Harmonic analysis: multiscale analysis (e.g. frequency/wavepacket decompositions). Harmonic analysis underpins much of image/signal processing, for example..

  Graph theory: combinatorial property testing, an area whose aim is to gain information about large networks and related structures by taking only a small number of local random samples.

  Complex networks: stochastic models for (large) complex networks. Such models can be used to infer and predict relationships between key parameters of large networks and (infection) processes on such networks. Complex networks with applications in neuroscience, gene regulation and ecology.

  Topological Data Analysis: integration of diverse data (Ethnicity, Culture, Health, Geography, Socio-economics) from distinct cohort studies

- **Understanding humans in a connected world**

  Understanding human decision making: understanding understand how people and organisations make decisions and how they can optimize their behaviour to achieve higher profit, better social outcomes, as well as flourish and bolster their well-being.

  Human-computer interaction: quantitative approaches to studying human interaction with technology; socio-technical systems.

  Understanding crime: predicting risk of criminality from adverse childhood experiences; comparison of model-based algorithms with human judgement in solving crimes.

  Understanding conflict*:* quantitative and qualitative approaches to understanding conditions under which conflict (such as civil war) begins; including geographical, social, and governance aspects.

- **Ethics and data science**

  Consent and confidentiality: exploring issues raised by the creation of data sets and use of big data.

  Ownership issues: for example, with regard to the data and profits derived from data (e.g. data-mining).

  Use of big data and analytics: particularly with regard to respecting minority groups and interests (which may be drowned out in the results of big data), and in recognising valuable but non-consequential outcomes.

- **Corpus Linguistics**

Understanding language variation and change: how do words, meanings and linguistic patterns change over time and change according to groups of speakers reflecting changes in culture and society.

Studying discourses: how does language create and shape realities, e.g. how are social groups talked about, how are political and economical problems presented through discourses evidenced in the form of Big Data.

Literary texts as data: how do corpus linguistic methods, stylometry and distant reading identify patterns in literature.

- **Data Intensive Physics**

Gravitational waves: data science and data intensive applications of a broad range of Bayesian and stochastic sampling techniques, Markov Chain Monte Carlo and Nested Sampling Methods, to infer the properties of compact objects in the Universe through gravitational wave observations.

Optical astronomy: the Large Synoptic Survey Telescope (LSST) is the flagship next generation optical survey telescope, which from ~2022 will survey the entire sky every four nights. LSST data processing presents a substantial "big data" challenges.

Exoplanet discovery: extracting empirical evidence about the physical processes that lead to the formation and evolution of planets; investigating planetary systems that are different to our own, either by the type of planets that compose those systems, by their architectures, or because of the type of star(s) they orbit.

Particle physics: central involvement in the ATLAS experiment at the Large Hadron Collider, studying the Higgs boson, and matter-antimatter asymmetry in the Universe using beauty quarks at the LHCb experiment. We also are part of the Worldwide LHC Computing Grid - distributed computing for LHC data analysis and Monte Carlo simulation production, where we host a medium-sized site by UK university standards.

Local Cluster Substructure Survey (LoCuSS): an international collaboration whose aims include to calibrate galaxy clusters as a probe of the mysterious “dark energy” responsible for the accelerating expansion of the universe.

- **Robotics for nuclear decommissioning**

National Centre for Nuclear Robotics: A consortium of eight universities, led by University of Birmingham, has secured £42 million of new investment to fund the National Centre for Nuclear Robotics (NCNR).

State-of-the-art robotics, sensing and AI technologies: to address the major societal challenges posed by nuclear environments and materials. Much of this work must be done by robots, because the materials are too hazardous for humans, however, many of the necessary robotic solutions have not yet been developed

- **Detecting, tracking and understanding single molecules**

Deep learning for single-molecule science: using deep-learning methods to make predictions based on single-molecule data.

Vector-based classification of single-molecule charge transport data: multivariate pattern analysis applied to simulated and experimental single-molecule charge transport data

Understanding, imaging and modifying DNA: using enzymes for DNA marking; detecting the epigenome using DNA sequencing.

Protein visualization: novel methods for visualising single membrane proteins, COMPARE: a unique multi-million pound collaboration between the universities of Birmingham and Nottingham.