UNIVERSITY OF LEEDS






LEEDS *Institute for Data Analytics*

ANNUAL REPORT

2018 / 19


LIDA: AN INTRODUCTION

LIDA was established in 2014 as a hub for multidisciplinary data science, combining world-class research with new initiatives in training and education, in partnership with multiple stakeholders.

There is a growing movement around the world to ensure the effective use of emerging, vast data collections to drive research, policy development and improve lives. The Leeds Institute for Data Analytics (LIDA) brings together research groups and data scientists from a range of disciplines to open up new opportunities to understand health and human behaviour. In doing so, helping to better understand what is required to tackle a wide range of social and environmental problems.

The LIDA remit is three-fold:

- **Engage in world-class research.** Our approach is one of collaboration, allowing for interdisciplinary methodologies to address complex sociological and environmental challenges, facilitating the transformation of information into knowledge.
- **Develop and strengthen data analytics expertise.** At the core of our activity is a desire to address the skills gap in data analytics. We do this through a diverse portfolio of education, training and events. We provide opportunities for researchers, students and partners across disciplines and organisational boundaries, to acquire new skills and knowledge.
- **Work with partners for the benefit of society.** The cornerstone of LIDA is working in close collaboration with those organisations who are themselves the generators and custodians of data, to address real-world challenges and build capacity.

WELCOME FROM PROFESSOR LISA ROBERTS, DEPUTY VICE-CHANCELLOR: RESEARCH AND INNOVATION

Five years ago Leeds laid the foundations in data analytics, and since then research in this area has grown significantly. This has been enhanced through investment in people and through our partnership with the Alan Turing Institute. This partnership, led by LIDA but with contributions from researchers and data scientists across the University, has been established to make great advances in data science research to change the world for the better.

Being a partner of the Turing builds on Leeds' strengths in making a real and telling difference to the world around us, by working across traditional boundaries to find innovative solutions to some of the greatest challenges facing us today.

The importance of data analytics and artificial intelligence (AI) to national prosperity and well-being has become more prominent than ever over the last year. Earlier in 2019, the UK Government announced a one billion pound sector deal for artificial intelligence. The sector deal includes the promise of continued research funding, the appointment of outstanding academic leaders as AI Fellows within UK universities, and support for a new industrial Masters programme to develop new educational opportunities that create stronger links with business and government to power the economy. This increased focus on AI plays to the strengths established within LIDA and beyond across the University.

LIDA has a strong portfolio of education and training opportunities that includes a range of multidisciplinary MScs and Doctoral Training Centres, and we are very excited about the new CDT in AI for Medical Diagnosis and Care funded by UKRI.

We have always considered our relationship with external partners from business, government, health and social care, and transport, to be a major strength of LIDA and key to ensuring that our research has real-world impact. As well as benefitting from our academic programmes, our partners share data, expertise and insight, collaborate through research events, workshops, and student supervision, and in many cases are generous in their financial support for our research. The formal launch of our new Nexus Innovation Hub in May 2019 and our new Business Engagement Framework are clear signals of the University's intent to create a step-change in the way we support, develop and celebrate applied research with impact.

I am extremely proud of LIDA's success to date and I am very excited about how our research in data analytics and AI continues to shape the future.

INTRODUCTION FROM PROFESSOR MARK BIRKIN, LIDA CO-DIRECTOR

LIDA was established in 2014 with high hopes and bold ambitions from the union of an ESRC Consumer Data Research Centre and MRC Centre for Medical Bioinformatics.

In the intervening period we have been the beneficiaries of generous support from the University of Leeds, Research Councils and other funders. The pages of this report show convincing evidence of the productivity of these investments. National and regional programmes have built on our core themes with respect to consumption, for example in the Centre for Doctoral Training in Data Analytics and Society (pp 56-57), the BBSRC PigSustain project (pp 36-37) and in five years continuation funding for the CDRC. Continued activity in Medical Bioinformatics embraces Pathology (pp 34-35), Cardiology (pp 30-31), and our established national centre for Bowel Cancer Intelligence. Most exciting of all, we have begun to connect between projects (e.g. using loyalty cards for public health (pp 32-33), and in the SIPHER and ActEarly projects (p20). We are joining up with foundational expertise in mathematics, computation and AI e.g. in advancing data assimilation for next generation urban analytics (pp 40-41), pattern mining and visualisation of data (pp 42-43), and AI for Medical Diagnosis and Care (pp 54-55). Leeds partnership in the Alan Turing Institute (pp 14-15) has provided further impetus, partly in helping to connect 24 Turing Fellows across multiple faculties and career stages, but also to place Leeds at the heart of a major national initiative with new mechanisms to collaborate with our peers at the forefront of the revolution in data science and AI.

Academic innovation in LIDA is not only promoted through grants from Research Councils. Doctoral training, taught postgraduate programmes and a unique data science intern scheme are vital elements of the strategic mix. All of these have prospered richly in the last year with the award of a new CDT, the introduction of new MSc programmes and our third and largest cohort of eleven interns. Importantly, these activities are also all strongly focused to external partners, providing data, advice, problem orientation and often direct financial support for LIDA's activities. LIDA's impact transcends peer-reviewed research publications to the provision of data infrastructure and the delivery of value to our partners in time, money, healthy lives, effectiveness and efficiency. We are hugely grateful to all the partners who have shared with us in this genuinely collaborative process. If there are readers who have to date preferred to observe than to engage, we can only encourage that a warm and constructive welcome awaits should you be motivated to move closer in the next year.

Perhaps most exciting of all, LIDA sits on a wave of innovation in data science and AI which is only just beginning to swell. Again, the current mix demonstrates how widespread these interests are becoming across disciplines, and through business and government. In order to maximise its potential, LIDA will need to continue to adapt rapidly. I am delighted that this last year has seen the appointment of a new co-Director and four deputy Directors to strengthen our management group. I'm sure that I speak for us all in expressing an optimism at what can be achieved with the foundations that have now been laid, and how much we look forward to reporting on the progress to come in future years.

SCALING UP OUR ACTIVITIES

NEW APPOINTMENT IN HEALTH INFORMATICS IN LIDA

Professor Chris Gale has been appointed as Interim Co-Director of Leeds Institute for Data Analytics.

Since LIDA's inception five years ago, the technology and data landscape has evolved very rapidly. Professor Gale will lead the development of a new vision for biomedical informatics and the supporting infrastructure. Chris will initially take up the post for a minimum of six months while a full time Chair of Health Informatics is recruited.

Professor Gale said: *"I am delighted to take on this important role at this critical time in health data research – Leeds has real strength in data analytics and this is an excellent opportunity to build on the great work to date and create a world-leading Institute. Given the increasing volume and availability of population-wide structured and unstructured health information, the reality of data driven discovery for patient benefit is upon us."*

Professor Gale is Chair of Cardiovascular Medicine and until his LIDA appointment was Head of Clinical and Population Sciences at the Leeds Institute of Cardiovascular and Metabolic Medicine (LICAMM). Professor Gale's research incorporates the efficient use of observational and randomised data to deliver population-based studies of cardiovascular quality of care and clinical outcomes. This includes the use of multi-source electronic health records for trial design, outcomes capture and taxonomy of cardiovascular survivorship. He holds major research awards predominantly from the National Institute for Health Research and the British Heart Foundation, and has published over 140 research manuscripts in peer reviewed journals, including Journal of the American Medical Association, Lancet, British Medical Journal and European Heart Journal. He is honorary Consultant Cardiologist at Leeds General Infirmary where he practices clinical cardiology with particular interests in general cardiology, post myocardial infarction survivorship, and chronic heart failure.

NEW DEPUTY DIRECTORS EMBEDDED WITHIN LIDA

In autumn 2018 LIDA introduced four new Deputy Director roles to support the development of the Institute's Research & Innovation, Education & Training and Research Technology portfolios.

LIDA Deputy Director of Research & Innovation
Professor Mark Gilthorpe,
Professor of Statistical Epidemiology, School of Medicine

Mark is a Fellow of the Alan Turing Institute for Data Science and Artificial Intelligence. Trained as a mathematical physicist, Mark's driving interest centres on improving our understanding of the observable world through modelling. After his PhD, he spent time as a consultant data analyst before being recruited into academia. Mark has since fashioned a programme of interdisciplinary research that spans the gap between theoretical and applied data analytics, focussing particularly on modelling complexity and highlighting and solving common data analytic problems. More recently, Mark's research and teaching interests have converged around the insights and utility of causal inference methodology, addressing the often under-recognised third pillar of Data Science, which comprises 'description', 'prediction' and 'causal inference'. Mark is particularly keen to explore how causal inference might be integrated with machine learning and artificial intelligence.

LIDA Deputy Director of Education & Training
Dr Luke Burns,
Lecturer in Quantitative Human Geography, School of Geography

Having worked in both industry and academia, Luke has developed expertise in several areas of quantitative spatial analysis including the advanced application of geographical information systems (GIS) and development of geodemographic classifications and composite indicators. Luke leads an innovative online Masters programme in GIS and teaches on a broad selection of analytical courses comprising undergraduate, taught postgraduate, open distance learning and continuing professional development. He also holds a visiting lectureship at the University of Strathclyde. Luke is a member of the Leeds Institute for Teaching Excellence and a Fellow of the Higher Education Academy and has undertaken considerable work on the importance of embedding quantitative and data skills into university curricula. Luke has been recognised by the Association for Geographic Information and Times Higher Education for contributions to student education.

LIDA Deputy Director of Research Technology

LIDA Deputy Director of Education & Training
Dr George Ellison,
Associate Professor of Epidemiology, School of Medicine

George's research spans the biological and social sciences, with a specific interest in interdisciplinary work exploring bias and error within scientific method and practice. Throughout his career in South Africa and the UK, he has been involved in the design, delivery and quality assurance of award-bearing courses and skills-based training; focussing in particular on continuing professional development for primary, community and public health practitioners, and those delivering urgent and emergency care. More recently, George played a central role in implementing the flagship Research, Evaluation and Special Studies strand of the School of Medicine's MBChB curriculum at Leeds, and was the first lecturer in epidemiology and biostatistics to be nominated twice for the Medical Students' Representative Council (MSRC) Teacher of the Year. His current work includes working with LIDA's external partners to develop innovative postgraduate pathways in data science.

LIDA Deputy Director of Research Technology
Professor Roy Ruddle,
Professor of Computing,
School of Computing

Roy has worked in both academia and industry; interested in visualization, visual analytics and human-computer interaction in spaces that range from virtual reality and the real world to high-dimensional data. Currently, he is Principal Investigator on the EPSRC QuantiCode 'making sense of data' project and Co-Investigator on the NIHR QualDash project which is developing dashboards for National Clinical Audit data. He helped to develop the Leeds Virtual Microscope for the visualization of tera-pixel image collections, leading to its use for pathology training in NHS hospitals and commercialisation by Roche. His petrophysics research led to the spin-off company Petriva and his current collaborators include NHS Digital, Leeds City Council, Leeds Teaching Hospitals NHS Trust, and J Sainsbury PLC.

OUR ALAN TURING PARTNERSHIP

Turing Fellows
In Autumn 2018 we were pleased to announce that 24 researchers from the University of Leeds begun Fellowships at The Alan Turing Institute. These prestigious Fellowships marked the second phase in our partnership that was announced earlier in the year between the University of Leeds and the UK's national centre for data science and artificial intelligence.

The 24 Fellows, who cover a range of disciplines and represent multiple faculties across the University, will join the Turing's existing community of researchers who are advancing data science and artificial intelligence to address a number of ambitious challenges facing science, society and the economy.

"Being a university partner of The Alan Turing Institute provides opportunities for the University's researchers to work closely with the Institute's academic, industry and policy partners and undertake the most ambitious, impactful research possible.

I am confident that The Alan Turing Institute will benefit from LIDA's interdisciplinary approach and the fantastic engagement we have achieved with national partners across the retail, energy and healthcare sectors."

Professor Lisa Roberts,
Deputy Vice-Chancellor:
Research and Innovation

Programme Lead for Urban Analytics
Towards the end of 2018 the Alan Turing Institute has announced a new urban analytics research programme, led by Professor Mark Birkin, Co-Director of LIDA.

The Alan Turing Institute is the UK's national institute for data science and artificial intelligence, urban analytics as the latest of nine strategic research programmes in areas such as artificial intelligence, health and public policy.

"The Institute has announced its urban analytics programme at an opportune time. New data from 'smart cities' is providing transformative insights all over the world.

Devices ranging from wearable tech to smart tickets will permit deeper understanding of behaviour and lifestyles, economic prosperity, mobility and health – with positive impacts for business, planners and policymakers as well as the scientific community. The Turing is ideally placed to exploit these opportunities through the advancement of methods ranging from visual simulation to artificial intelligence."

"I am looking forward to working with the Turing's unique network of universities and external partners to create an ambitious programme of academic investigation with real world impact."

Professor Mark Birkin, Co-Director of LIDA

21 NOVEMBER 2017
University of Leeds announced as a Partner of the Alan Turing Institute

12 OCTOBER 2018
Turing Fellows announced

28 NOVEMBER 2018
Programme lead for Urban Analytics announced

CASE STUDY: UNDERSTANDING THE CITY THROUGH THE INDIVIDUAL

Only two joint fellowships between the Economic and Social Research Council (ESRC) and The Alan Turing Institute were awarded in their first year, and LIDA's Professor Alison Heppenstall holds one of them.

The fellowships were established to bridge the gap between the social sciences and big data and ensure that data science can be harnessed to address societal challenges. With a research background across both disciplines, Professor Heppenstall was an obvious candidate for this prestigious award.

The fellowships focus on 'smart cities'. The aim is to draw on the huge datasets generated within an urban environment to better understand how cities work and find ways to improve the lives of those who live, work and play within them.

Urban datasets can come from public transport, footfall cameras, retail transactions, weather stations or air quality monitors, for example. Professor Heppenstall is looking at using information from social media and mobile phones. Although it's not possible to identify individuals through these datasets, to protect privacy, they can show time, location and, through keywords in tweets, identify characteristics such as age or gender.

Professor Heppenstall explains: *"Census data provides information on where people live and mostly spend their nights; the journey to work data provides information on who is moving through the city at rush hour. But we know far less about who is using our cities during the day, such as students, non-working parents, the unemployed and retired. Merging footfall and social media data could help to fill that gap."*

Professor Heppenstall's research involves data-wrangling – bringing multiple data sets together in order to extract more meaningful conclusions. In addition, machine learning methods that can spot the patterns and gaps in the data are under development.

These patterns can be used to inform the design of what's known as an agent-based model to interrogate chosen datasets. These models are given their name because they assign behaviour and characteristics to individuals and then see the impact of that behaviour, thereby allowing the model to more accurately represent the reality of a city, where thousands of individuals are interacting all the time.

The applications are numerous, from determining what encourages different ethnic groups to live in certain areas, identifying who is most exposed to air pollution, or how big events impact on traffic flow.

RESEARCH AT LIDA

INTRODUCTION FROM PROFESSOR MARK GILTHORPE, LIDA DEPUTY DIRECTOR FOR RESEARCH AND INNOVATION

When LIDA was established five years ago as a creative interdisciplinary hub for data analytics expertise, it was funded through two new multi-million-pound centres: the ESRC Consumer Data Research Centre (CDRC) and the MRC Medical Bioinformatics Centre (MBC).

We are immensely proud of the tremendous work both Centres have achieved, both individually and by working together, and this has been rewarded by a steady stream of additional funding in excess of £50 million. This includes substantial follow-on funding for the CDRC until 2024. As well as new funding streams for a host of applied and methodological projects and centres, notably the Centre for Immersive Technologies. This has also led to the addition of a fourth theme – visualisation and immersive technologies – which works across a range of disciplines to help companies and organisations optimise the use of virtual reality and augmented reality in their products and services. With an annual global market predicted to be worth over £100 billion, immersive technologies are poised to influence and enhance every area of our lives, transforming how we live, work and play.

LIDA's success, and its commitment to data science skills development, are also reflected in the recent award of Centres for Doctoral Training in: Artificial Intelligence for Medical Diagnosis and Care (funded by the UKRI); and Fluid Dynamics (funded by the EPSRC); and by substantial investment from HDR UK in the MSc in Precision Medicine and new MRes in Data Science and Analytics for Health, each of which were developed in collaboration with industry and NHS partners. This partnership approach also lies behind the new FinTech MSc, developed in consultation with leading industry bodies, including the Chartered Banker Institute.

Clearly, there has never been a more important time to work in data science. Digital technologies are now widespread and pervasive, and the data these gather are recognised as instrumental in our economic and social futures. Here at LIDA we are focused on driving forward our applied and methodological research capabilities, working with external partners to generate robust and novel insights from data to deliver genuine impacts for the public and private sectors – insights designed to benefit the lives of individuals, communities and the environments that sustain us.

In this Annual Report, as well as reflecting upon our achievements to date, we look ahead to LIDA's key strategic initiatives for the year ahead and beyond – initiatives planned to strengthen LIDA's reach and impact, and the insights that its research provides to external partners and policy makers.

FUNDING AWARDED TO LEADING DISEASE PREVENTION PROJECTS

For its first ever funding round the UK Prevention Research Partnership (UKPRP) awarded £25 million into understanding and influencing the social, economic and environmental factors that affect our health.

The research projects awarded funding will each look at the 'bigger-picture' factors behind the prevention of non-communicable diseases – which are estimated to account for 89 percent of all deaths. All of the projects aim to deliver real-world changes to reduce the burden of these diseases on our health and social care systems and enable people to live longer, healthier lives.

In total, 8 projects received the UKPRP funding with 2 directly linked with LIDA:

- **ActEarly:** a city collaboration approach to the early promotion of good health and wellbeing. Led by the Bradford Institute of Health Research received £6.6 million. ActEarly will conduct research into improving the life chances of children in two deprived areas in the UK, Bradford (Yorkshire) and Tower Hamlets (London), focused on healthy places, healthy learning and healthy livelihoods. Professor Mark Mon-Williams from the University of Leeds will lead the healthy learning theme.
- **The SIPHER Consortium** – Systems science In Public Health Economic Research, led by the University of Sheffield was awarded £4.9 million. SIPHER will research systems-based economic evaluation methods to provide a common basis on which to appraise the effectiveness, costs and benefits of policy measures across social and health outcomes in different policy areas, including inclusive economic growth; housing; mitigating the effects of adverse childhood experiences and the promotion of mental wellbeing.

"By investing in these interdisciplinary teams and drawing on a wide range of knowledge and expertise, UKPRP is supporting work that will have real life benefits for both policy makers and the wider public alike.

Non-communicable diseases place a huge burden on us all and we hope that this investment will help to provide practical and tangible solutions that will positively impact people's lives and health."

Professor Dame Sally Macintyre,
Chair of the UKPRP Scientific Advisory Board and Expert Review Group Panel

UNIVERSITY OF LEEDS ACADEMIC NAMED AS ROYAL ACADEMY OF ENGINEERING CHAIR

Investigations into an alternative and complementary method of discovering, developing, and testing new medical devices has received funding for ten years as part of the Academy's Chair in Emerging Technologies scheme.

Professor Alejandro Frangi, Diamond Jubilee Chair in Computational Medicine and associate member of LIDA has been named Chair and been awarded almost £2.7 million for his research.

He is using techniques developed in the emerging discipline of computational medicine where imaging, sensing, modelling and simulation converge. The ten-year support provided to the Chairs will enable them to progress their pioneering ideas from basic science through to full deployment and commercialisation.

The Academy has made awards totalling over £20 million in research funding through its Chairs in Emerging Technologies programme, providing long-term support to nine world-leading engineers across the UK to advance emerging technologies.

The new technology areas developed by the Chairs in Emerging Technologies have the potential to considerably benefit society and the UK economy, and enable the nation to remain at the global forefront of engineering innovation. The areas of research funded reflect the UK's wider technological priorities, with many of the projects directly aligned to the government's Industrial Strategy and designed to tackle some of the biggest industrial and societal challenges of our time.

Professor Frangi said: *"Computational medicine can bring about a complete shift in the way devices are conceived, developed, and ultimately tested for the market."*

"We are developing methods and systems to realise the vision of 'in-silico' trials where computer analysis is used to engineer medical devices from their conception. Computer models will be looking at ways, and under what circumstances, a device could fail, cause harm or be ineffective for some groups – and all of this will be happening long before testing with real patients.

"These in-silico trials are based on populations of virtual patients representing the natural variation, for instance, in people's anatomical, physiological and biological make-up found in real-life or target populations."

UNIVERSITY ACADEMIC FELLOWS

Great Minds, the University's biggest ever recruitment scheme which was launched in 2014, aimed to attract a wealth of new talent to Leeds.

The scheme will recruit up to 250 University Academic Fellows (UAFs), each aligned to existing disciplinary strengths or future strategic aspirations; and each hosted within innovative interdisciplinary teams, including those developed within LIDA. These five-year, tenure-track posts offer extraordinary opportunities for professional and career development, culminating in the appointment to the rank of Associate Professorship. To date, over 150 outstanding UAFs have been recruited, with more than one in five affiliated with LIDA, as demonstrated in the following table:

Matthew Allsop	Leeds Institute of Health Sciences
Julie Aspden	Biological Sciences
Dan Birks	Quantitative Policing and Crime Data Analytics
Ron Chen	Biological Sciences
Elizabeth Duncan	Biological Sciences
Niamh Forde	Leeds Institute of Cardiovascular and Metabolic Medicine
Heather Ford	Performance, visual arts & communications
Jiaqi Ge	Geography
Marlous Hall	Leeds Institute of Cardiovascular and Metabolic Medicine
Eva Heinen	Transport Studies
Antreas Kalli	Molecular and Cellular Biology
Nik Lomax	Geography
Robin Lovelace	Transport Studies
Laura Matthews	Cancer and pathology
Richard Mann	Statistics
Jessica Meyer	Legacies of War
Michelle Morris	Health Informatics
Vincent Muller	Theory and ethics of disruptive technology
Faisal Mushtaq	Psychology
Mary O'Connell	Computational and Molecular Evolutionary Biology
Carlo Perrotta	Digital Learning
Mar Pujades-Rodriguez	Biomedical and Clinical Sciences
Viktoria Spaiser	Political Science Informatics
Peter Tennant	Health Data Analytics
Mark Trigg	Water related risk
Paul Townend	Institute of Computational & System Sciences
Lucy Ziegler	Palliative Care

This year we spotlight the activities of four of these UAFs to illustrate how their diverse strengths are helping to expand the scope, reach and capability of Data Science here at Leeds.

Jiaqi Ge
Since starting my LIDA-affiliated UAF post in May, I have been struck by the extraordinary research support available here at the University of Leeds. This includes a designated team established specifically to support early career researchers (several of whom are UAFs), which has helped many of us to secure research funding and establish interdisciplinary research teams. Indeed, at one of the very first of the team's events I met a wide range of career mentors and was able to discuss ideas with other early career researchers from a diverse array of backgrounds – discussions that have since developed into concrete ideas for many of the projects we are now progressing (more on these below). I have also benefitted from the regular practice-based workshops and drop-in sessions the team provides, many of which are specifically designed to support early career researchers generate successful grant proposals; support I have found to be invaluable.

As a result of the University's strong commitment to early career researchers, I have already been involved in several exciting projects and associated grant proposals. Given my own research focus on developing simulation models – and in particular, agent-based models – to study complex social and urban phenomenon, the large number of well-supported UAFs and ECRs across the University of Leeds means that I have been able to develop collaborations across a swathe of applied areas (including: housing bubbles; urban decline; gentrification; and work-related commuting patterns). For example: I am currently working with colleagues in the School of Computing on opportunities for linking artificial intelligence and machine learning with agent-based modelling; I have been invited to participate in a European H2020 project with colleagues from Geography's Ecology and Global Change Group; I am developing a grant proposal on smart rural transport with colleagues at the Institute for Transport Studies; and I have also developed a collaboration with colleagues from University College London, Oxford University, and Edinburgh Napier University, together with international colleagues in China, Canada, USA and Germany.

Clearly, the collaborative culture and diverse networks that LIDA and the University of Leeds offer, are offering wonderful opportunities for the interdisciplinary approaches that my work so often features. I am looking forward to developing these, and to delivering impact of tangible societal value, in the years ahead.

Peter Tennant
Since joining LIDA, the focus of my work has increasingly converged on understanding, applying, adapting and translating novel methods for 'causal inference'. LIDA has provided me with a powerful platform for international collaboration on related applied projects, including an international consortium aiming to develop guidelines for the application of directed acyclic graphs (DAGs) to applied research. It has also facilitated the development of my teaching expertise which – somewhat unsurprisingly – has focussed on causal inference training. As a result, I have taken up the posts of: Deputy Programme Lead for the MSc in Health Data Analytics, on which I deliver Introduction to Health Data Science and Modelling Strategies for Causal Inference; and Co-Lead of the oversubscribed LIDA/Turing Summer/September School in Causal Inference with Observational Data (see p.68).

My research – and the focus of my Turing Fellowship – centres on using DAGs to understand, and thereby resolve, longstanding challenges in the analysis of observational data. Recent work, for example, has helped to unravel several persistent problems with: the analysis of change scores; compositional data; and longitudinal data. This approach has helped improve our collective understanding of findings based on biased (naïve) analyses of composite variables (such as Body Mass Index/BMI and weight gain/loss during pregnancy), and I will shortly begin supervising a Turing Doctoral Scholar who will be examining similar issues in the analysis of composite anthropometric variables commonly used in cardio-metabolic research.

I also lead the Turing Institute's Interest Group on 'Causal Inference', serve on the Cancer Research UK Epidemiology Expert Review Panel, and I am Honorary Secretary of the Society for Social Medicine and Population Health, as well as the University of Leeds lead for the UK Reproducibility Network.

Faisal Mushtaq
I recently completed my University Academic Fellowship in Health Engineering and am delighted to have been appointed to a Fellowship of the prestigious Alan Turing Institute, not least because my background is in experimental psychology while the research questions I address (which are centred around human decision-making) sit at the interdisciplinary interface between neuroscience, engineering and computing – all of which benefit from the data-driven methodologies that the Turing Institute aim to support, and which LIDA has helped to foster at the University of Leeds.

Indeed, over the course of my LIDA-affiliated UAF I implemented a range of different techniques to generate the data necessary for understanding the processes underlying the selection and execution of 'actions'. These methods ranged from capturing millisecond-precise indices of neural activity through to extracting measures of real-world performance from large-scale datasets – datasets that, over the last few years, have been radically impacted upon by the increasing availability, capability and utility of virtual – and augmented-reality (VR and AR) technologies.

My research group – the Immersive Cognition Laboratory (ICON) – has been quick to take advantage of the scientific power afforded by these technological advances. Our strong conviction has been that the portability, experimental control, measurement precision, and visualisation capabilities provided by VR and AR will transform the study of human behaviour; and I am very excited that this potential has been recognised by LIDA and the University of Leeds in their support for the establishment of a dedicated Centre for Immersive Technologies (CfIT) in June this year. The Centre constitutes a major new research initiative which aims to address scientific challenges in VR and AR that – like my own research – specifically require cross-disciplinary working. As such, I am delighted to have been appointed the Centre's first Associate Director, and that the Centre will be housed within LIDA. There is a natural symbiosis between immersive technologies, AI and data analytics, and I am excited about the novel opportunities that will be created for our new Centre by working within the vibrant interdisciplinary research environment that LIDA has succeeded in developing.

Robin Lovelace
My LIDA-affiliated UAF post has offered extraordinary opportunities for collaborative research, and the last three years have been very (very) busy … though extremely rewarding in terms of the activities and outputs undertaken and delivered. These include the Propensity to Cycle Tool (https://www.pct.bike) – a publicly available planning support system, the development of which I led in 2017 – which has been used by tens of thousands of people to better understand active transport, and by policy makers and planners to inform strategical cycle networks nationwide. The Tool has influenced more than £500 million of investment in Manchester, Liverpool, Birmingham and cities further afield; and has led to several high-impact research outputs and publications, as well as consultations with international organisations (such as the World Health Organisation) which, in turn, have led to follow-on funding. These applied successes aside, LIDA has also provided an ideal methods-focussed interdisciplinary context for the development of a suite of new analytical packages in R, and an associated book entitled Geocomputation with R, which was published earlier this year and has already received more than 50,000 unique visits (and strong sales of the hard copy as well: Lovelace R, Nowosad J, Muenchow J. Geocomputation with R. CRC Press; 2019). LIDA's role in supporting UAFs such as myself to develop, network, collaborate and achieve such impacts continues to be invaluable.

CASE STUDY: USING DATA SCIENCE TO FIGHT CRIME

We expect our police forces to deal with an increasingly diverse number of problems – from tackling burglary and violent crime to safeguarding vulnerable communities and responding to critical incidents. While the nature of crime is changing, the tools and datasets we have for understanding and combatting it are also evolving.

Police agencies are data-rich organisations and data analytics is one tool that is becoming invaluable in supporting police to make the most of their information. The goal is to effectively and ethically deliver the resources they have in ways that maximise safety and minimise harm in our communities.

Researchers at LIDA are at the forefront of this new science, working with a number of different partners developing ways to use data analytics to better understand, predict and prevent crime.

LIDA is working with Safer Leeds, investigating ways to analyse free text data from crime reports – a process which requires natural language processing and machine learning techniques to look for patterns and similarities within the reports. These tools will help crime analysts better understand patterns in types of offending and aim to offer an early warning system for new emerging criminal behaviours.

"Vast amounts of rich unstructured text data are collected by police and their partners on a day-to-day basis," says David Jackson, Partnership Intelligence Lead at Safer Leeds.*"These large datasets present significant analytical challenges, but also offer huge opportunities. The work we're doing with LIDA will help us harness this resource to better understand and ultimately, we hope, reduce crime."*

Dr Daniel Birks, Academic Fellow in Quantitative Policing and Crime Data Analytics at the School of Law and Fellow of the Alan Turing Institute, says:*"A typical example of these new types of offending are recent spikes in thefts and muggings committed by offenders riding mopeds. Police intelligence has always had an operational understanding of these offences, but using data analytics we can proactively identify these new criminal behaviours more clearly and rapidly, and start to understand the types of conditions under which they're most likely to take place."*

Working with the N8 Policing Research Partnership, LIDA has established an alliance across the N8 group of universities and 11 UK police services to address the challenges of modern policing.

Through the partnership, LIDA has teamed up with West Yorkshire Police to use machine learning to tackle vehicle crime. Researchers are developing ways of recognising vehicles with falsified number plates, by comparing the genuine make, model, year and colour of a vehicle with the records accessed via automatic number plate recognition systems. This will allow them to flag vehicles that are likely to have been cloned and might go on to be involved in criminal activity.

Taking a broader look at criminal behaviour, Dr Birks is also developing advanced computer simulations of crime using a technique called agent-based modelling. These models, which he describes as 'synthetic societies', can be used to better understand the link between the individual offender and victim behaviour and widespread trends in crime across society.

"Using these models, we can carry out experiments that would otherwise be impossible in the real world. We can use them to better understand how, for example, different street network configurations or neighbourhood facilities might increase or decrease people's risk of victimisation," says Dr Birks.

The team has recently received funding from The Alan Turing Institute to explore how these simulation techniques might also be used in practice. Working with partners at the University College London, the Metropolitan Police and the National Police Co-ordination Centre, the goal is to see if tools can be developed to help police better understand the complex challenges of resourcing and demand.

CASE STUDY: THE BIGGER PICTURE BEHIND HEART ATTACKS

The number of people surviving heart attacks has improved enormously in recent decades – but little is known about what happens to patients afterwards. Helping healthcare professionals, patients and researchers understand what other diseases heart attack survivors may have, or might go on to develop, could lead to new treatments and better health outcomes.

Dr Marlous Hall, a senior epidemiologist at LIDA, is leading a Wellcome Trust-funded project to investigate using electronic health record data gathered by NHS Digital – accessing around 145 million records of hospitalisations in England.

Through previous work funded by the British Heart Foundation (BHF), the team identified that nearly 60 per cent of patients hospitalised with a heart attack will also have at least one other chronic condition, likely to reduce their life expectancy by around three years. Conditions such as heart failure, hypertension and peripheral vascular disease, a circulation disorder that affects blood vessels outside the heart.

"We want to find out why people might be susceptible to these other types of disease – is it just because they are living longer, or are there other reasons connected to their heart condition, that we need to be considering?" explains Dr Hall.

Identifying these patterns and links could help inform patients, doctors and the wider research community about what combinations of diseases might occur together and which treatment combinations may need to be developed in future.

"Current treatment guidelines will typically look at the individual diseases," says Dr Hall. *"But if we can understand how, for example, a heart attack might progress towards heart failure, we should be able to devise better long-term treatment plans that could manage these diseases in combination – reducing the treatment burden on patients – or prevent their onset altogether."*

The work has already been used by the British Heart Foundation in a campaign to improve awareness of heart failure and associated illnesses.

"As the population gets older, more and more people who experience a heart attack are already suffering from a number of other illnesses," says British Heart Foundation Associate Medical Director, Jeremy Pearson.

"We need to make sure that we're providing the best possible care for people with these conditions, to both reduce their chance of having a heart attack and to give them the best possible chance of recovering from a heart attack should the worst happen."

In future, the team plans to broaden out the approach, looking at different diseases to understand the risks patients with one disease have of developing other health conditions.

LIDA's IT infrastructure is fundamental to the success of research programmes like these since the ability to store sensitive information securely is key to the partnership with NHS Digital.

Dr Hall says: *"There are great resources available, including hospital and GP records, but it's not easy to access or process such vast quantities of data and use it to provide meaningful insights whilst also sticking to extremely rigorous ethical standards. Through the collaborative environment in LIDA, we're able to do that, so all that data becomes extremely powerful for both researchers and for the overall benefit of patients."*

CASE STUDY: COULD LOYALTY CARDS IMPROVE OUR PUBLIC HEALTH?

The use of supermarket loyalty cards and mobile phone apps tracking diet or exercise are increasingly pervasive. Not only do they provide a useful service for the user but they also gather information about the habits of individuals. What if all that data could also be used to feed into public health research that could improve the overall health of the community?

LIDA's LifeInfo survey, led by Dr Michelle Morris, surveyed 10,000 people about the types of loyalty card or apps they use, and their thoughts on allowing health researchers to access this information and link it with their health records, in the future.

If people are open to sharing their data, the project could be the first step in a much longer-term programme looking at how the records might be used in public health research.

"In this first phase, we really want to find out what concerns people might have about how their personal information is used," says Dr Morris.

"Researchers need to have insight into what's available, but also how they need to work with people to alleviate concerns and build trust around how the information will be used."

The ultimate aim is to build a research tool that can cross-reference thousands of lifestyle and health records to uncover links and patterns. This could help researchers and healthcare professionals to develop interventions and treatments for conditions such as diabetes, obesity or cancer. Data would be anonymised and stored in a secure research environment so that there is no risk of identification of individuals.

"We think there's huge potential here to uncover some really useful insights," explains Dr Morris. *"Supermarkets gather this data for marketing purposes, to increase sales – why aren't we also using it to help improve people's health too?*

"If we can look in detail at the differences in lifestyle between people who have a serious health condition such as diabetes, and those who don't, we might be able to uncover new patterns that could help us treat or prevent these conditions."

Existing large surveys, such as the National Diet and Nutrition Survey, already collect some valuable information, with around 9,000 responses gathered over the past nine years. But gathering this information is labour intensive for participants and there will be errors with people not remembering, or choosing to give inaccurate answers.

"The type of information we might get from people's shopping trips, or from their apps will also have limitations," says Dr Morris. *"But the information has already been collected and is not subject to someone remembering what they ate or how far they walked – it's potentially a hugely valuable resource if people are open to donating the information."*

The team is working in partnership with Leeds Teaching Hospitals. Adam Glaser, Professor of Paediatric Oncology and Late Effects at the University of Leeds, and Consultant Paediatric Oncologist and Late Effects physician at the Leeds Teaching Hospitals NHS Trust, is a co-lead researcher on the project. He says: *"Gaining these precise and detailed insights would give us a much clearer picture of how lifestyle affects health.*

"One clear example of where this might benefit paediatric medicine is in premature birth where there are many health risks. We don't always know why some babies are born early, but there may be clues in the diet and physical activity patterns of the mother. Defining these patterns is the first step in identifying guidance or interventions that could improve the health of expectant mothers and babies."

CASE STUDY: BRINGING PATHOLOGY PRACTICE UP TO DATE

Digital pathology is by no means a new practice; pathologists have had the ability to scan and digitise an entire pathology glass slide for over twenty years. The proffered advantages of digital pathology are many; allowing for a more efficient and collaborative diagnosis and yet hospitals haven't made the switch from using microscopes and glass slides.

The Northern Pathology Imaging Co-operative (NPIC), led by the University of Leeds and Leeds Teaching Hospitals NHS Trust received an investment of £17.1 million from UK Research and Innovation and involved industry partners to roll out a programme of digital pathology and artificial intelligence across the north of England.

Step one of the transformative programme is already underway; from September 2018 Leeds Teaching Hospital NHS Trust began to digitise all pathology, and over the course of the project all of the hospitals in the West Yorkshire Association of Acute Trusts will follow.

The benefits are clear; speeding up and improving collaboration between pathologists and researchers. Instead of sending pathology glass slides from one lab to another, a link to an image can be shared amongst teams, wherever they are based, with markers plotting areas to study like a longitude and latitude on online maps.

Changing these ways of working is no mean feat, without considering the practical changes for the workforce, the sheer quantity of data this will produce is huge. To put this in context, if you were to print out a full resolution digital pathology image it would be the size of a tennis court.

NPIC will generate an average of 760,000 images per year, about 1.2 petabytes of data. If one byte is the size of a grain of rice, one petabyte is the Island of Manhattan covered in rice.

Geoff Hall, Professor of Digital Health and Cancer Medicine and Chief Clinical Information Officer for Leeds Teaching Hospitals is leading the area of work focused on how to securely store and process data at that scale.

Another key part of the project is to consider the ethics of data sharing to ensure NPIC partners abide by the highest professional standards when images are used for research purposes.

The work will also look at applying artificial intelligence and machine learning to digital pathology, transforming how cancer and other diseases are diagnosed. In the near future AI can be trained to recognise patterns in pathology images, on a massive scale and potentially predict patient outcomes.

Professor Hall is also considering how to link the pathology images with patient clinical information to gain further insight into the disease trajectory.

Dr Darren Treanor, a Pathologist at the University of Leeds and Leeds Teaching Hospitals NHS Trust, is leading the project, said: *"Digital pathology is a technology with a huge potential to improve healthcare.*

"NPIC will allow us to use digital pathology to help patients across the region, and provide a platform on which we will develop artificial intelligence tools for pathology diagnosis to be used around the world."

The consortium includes a network of nine NHS hospitals, seven universities and 10 industry-leading medical technology companies.

"This is a huge opportunity for Yorkshire to lead in this new area and further enhance our position as a hub for medical technology. We can explore how to use digital pathology as part of precision medicine to ensure patients receive treatments tailored to their disease."

Professor Geoff Hall

CASE STUDY: DATA HELPS THE PIG INDUSTRY PREPARE FOR THE FUTURE

The PigSustain project received £2 million to create a model to support the UK pig industry to respond to future challenges posed by the intensification of production, fluctuations in consumer demand, climate change, global production levels and international trade.

PigSustain is part of the UK's cross-government programme of food security research and involves biologists, economists, spatial scientists, statisticians, computer scientists, vets and industry representatives.

The team from LIDA includes Professor Mark Birkin, Dr Nik Lomax and Dr Will James, who are developing the part of the model that will forecast consumer trends and market stability. The research draws on the Office for National Statistics data from 2008-2016 as well as attitude surveys by government and polling companies.

"Essentially we're looking at how much people spend on pork products, which products they buy and how that has changed over time," explains Dr James. *"We also want to understand the spending habits of different groups within the population and the reasons those habits change."*

To drill down to this level of detail, the team is taking data from the annual Living Cost and Food Survey, which provides information on everything bought by 12,000 people over a single fortnight. Because the sample used for the survey is representative of the population as a whole, this data can be mapped against the UK population using the National Census, based on variables such as age, gender, ethnicity and household income. Enabling the team to calculate what pork products are bought by whom, and where. The data covers all food products, so they can see what items replace the pork products if consumption levels fall.

"This will allow us to quantify the impact of certain consumer trends," says Dr James. *"For example, with the older generation preferring to buy joints of pork, how important to the market are new products, such as pulled pork, which are of more interest to younger consumers?"*

The analysis has also highlighted huge variation across the UK, with certain areas of London seeing weekly expenditure on ham and bacon of just 20p per person, rising to 60p in some rural areas.

"Based on the attitude surveys, which include some demographic information, we can see how much pork consumption is affected by different factors, including religious beliefs, disposable income and environmental beliefs," says Dr James.

The next step is to use this historical data to map future trends of pork consumption and expenditure across the UK, drawing on the expertise of colleagues in LIDA who develop population projections. This will then feed into the larger PigSustain model to help the UK pig industry prepare for the future.

PigSustain is funded by the Biotechnology and Biological Sciences Research Council, Economic and Social Research Council, Natural Environment Research Council and Scottish Government, through the Global Food Security's 'Resilience of the UK Food System Programme'

"Being a university partner of the Alan Turing Institute provides opportunities for the University's researchers to work closely with the Institute's academic, industry and policy partners and undertake the most ambitious, impactful research possible."

Professor Lisa Roberts, the University's Deputy Vice-Chancellor: Research and Innovation

PARTNERSHIPS AND COLLABORATIONS

LIDA is matching the world class capabilities of university research with the needs and opportunities of external partners in business, government and the third sector. Ensuring our research priorities are driven by real-world challenges in partnership with the organisations that are facing them.

IMPROBABLE WORLDS LIMITED

Founded in 2012, Improbable World Limited is a British multinational technology company headquartered in London that makes distributed simulation software for video games and corporate use. It believes in a future where new, virtual worlds will augment human experience and become as meaningful as the physical world.

LIDA has developed a strong partnership with Improbable through a number of collaborations, with the aim of developing a computational and mathematical framework for data assimilation using agent-based models. The projects integrate data emerging from smart cities (e.g. traffic counters, social media activity, environment sensors, etc.) into large-scale urban simulations in real time.

LEEDS CITY COUNCIL

Since the foundation of LIDA in 2014, Leeds City Council (LCC) has been one of LIDA's key partners. The partnership currently comprises a number of collaborative projects, including two PhD projects, one of which is aiming to quantify the ambient population of Leeds city centre. Through undertaking internships with LCC, both PhD students have had the opportunity to analyse footfall trends at Kirkgate Market (one of the largest covered markets in Europe) and within Leeds city centre. By investigating these trends at various, times and locations throughout the Market and the city centre, they have created visualisations to explore the optimal times and days of the week to hold events, with the aim of boosting visitor numbers.

Investing in projects such as this can also be of potential benefit to planning domains as varied as emergency management, city planning and retail sector development. For example, the second PhD project is developing methods to perform real-time simulations of pedestrian travel systems (directions of flow, routes, crossings, etc.), with a particular emphasis on integrating real-time data into agent-based models of travel flows. In an age where technology is increasingly driving planning and infrastructure innovation, the value of these techniques and their applications are still being realised, and such techniques have the potential to inform not only emergency planning and management, including real-time response, but also data-driven civic and health service provision.

Health and care providers and commissioners in Leeds have identified the need for a single agreed population count for Leeds as the basis of coherent service delivery plans. Currently estimates for the Leeds resident population from ONS differ from GP registers by tens of thousands. To tackle this disparity, LIDA and LCC are also working together as part of the LIDA Data Scientist Internship Programme, on the question: What is the definitive population count for Leeds residents? This internship project is using these ONS and GP register data to attempt to answer this.

NEXUS
Successful innovation is nothing new to the University. Leeds has a proven track record of commercialisation, creating over 110 companies in the last 20 years, six of which are AIM market listed with a combined value in excess of £500m.

As a key part of the University's approach to business engagement, Nexus is a community of innovators located on Campus and based in a new state-of-the-art building, and provides easy access to the University's research and expertise.

Business partners can engage in a variety of ways, ranging from being a member of the Nexus community, to co-working or locating your business here, in the highest quality workspace and labs.

In May 2019, Nexus was launched and presented a new report in partnership with CBI focused on advanced data analytics and the creation of a new frontier for business research and development.

Professor Lisa Roberts, Deputy Vice-Chancellor: Research & Innovation at the University of Leeds, said: *"The CBI/Nexus report shows the extraordinary opportunities for UK businesses to harness data for R&D, as well as the vital role universities have to play in supporting innovation and productivity.*

"Unlocking this potential will supercharge R&D and that's why we're making a step-change at Leeds in how we work with business.

"This is what Nexus, our innovation hub, is all about – giving businesses seamless access to this support. There has never been a better time for universities and businesses to collaborate."

IMPACT STORY

LIDA Data Scientist Internship Project, partnered with Improbable Worlds Limited – Luke Archer, LIDA Data Scientist Intern

Probabilistic programming and data assimilation for next generation city simulation

The field of social simulation is dominated by Agent-Based Models (ABMs), typically ABMs are restricted to a one-shot approach to calibration, using historical data. Whilst these models might act very similarly to the real-world phenomena they mimic, they diverge from the true state (the actual position and movement of all real-world agents) almost immediately.

A solution to this would be to update the model with real-time data, following a process well studied in numerical weather prediction called Dynamic Data Assimilation (DDA). Data assimilation techniques combine the output of a predictive model with noisy real-world observations to produce a more accurate model state.

This project investigated the use of Keanu - a Probabalistic Programming LIbrary (PPL) developed by Improbable Worlds Ltd. - as a framework for data assimilation on a spatial AMB called StationSim (see figure 1). StationSim utilises the MASON framework to produce a simplified representation of passengers exiting a train and moving towards a specified exit.

Explaining the science

In DDA, statistical techniques are used to combine a predictive model with observations of the model state, accounting for the level of uncertainty in both. It is at heart a Bayesian process; we combine a prior (model prediction) with a likelihood (observations) to derive a more accurate posterior (assimilated model).

Whereas there are a number of well-defined DDA techniques, we saw an opportunity to investigate a new method. Probabilistic programming is the marriage of programming languages and statistical theory, that allows for complex statistical models to be defined and evaluated in hours instead of days or weeks by hand. The key task for a PPL is statistical inference, which in turn usually requires a data structure that can accurately represent a probability distribution.

Keanu relies on what is called a computational graph to carry out inference, which it calls a Bayesian Network. This network represents the conditional dependency relationships between all vertices, probabilistic and non-probabilistic. In this graph, Keanu can 'observe' (set) the value of certain vertices using observations. Keanu can then calculate the most likely value of the parent and child vertices, causing a cascade of calculations until it has calculated the most likely state of the entire network. The Bayesian Network here is our prior prediction, and after applying observations, our posterior is the updated model state.

Figure 1. StationSim. This image shows the AMB we used to develop the DA algorithm. Small circular agents (blue) start an entrance on the left (in green), moving towards one of the two exits on the right hand side (red).

Findings

Both Keanu and MASON are complex libraries for very different reasons; MASON attempts to handle the complicated parts of ABM behind the scenes hidden from the user, whereas Keanu is in a pre-Alpha release stage with limited documentation.

Using a digital twin, the first twin producing data and the second twin attempting to assimilate it, we have managed to produce a functional data assimilation algorithm with Keanu. However, the algorithm is not yet finished and requires some improvement to be useful. The algorithm shows a clear transformation of the state between assimilation windows when observing data but needs to be more accurate to apply to a real-world system.

Applications

There are many tangible applications to this research; an assimilated spatial ABM would prove useful wherever real-time knowledge of people's position and movement are valued. For example, ABMs have been used to improve evacuation and disaster planning. Knowing the exact position of people in a coastal town during a tsunami warning would mean better traffic management and more efficient use of shelters.

CASE STUDY: FINDING AND VISUALISING THE PATTERNS AND THE GAPS IN BIG DATA

Large organisations routinely collect vast amounts of data from the public, but finding the right tools to organise and interrogate that data is a huge challenge.

Each time you go shopping, visit your GP or interact with your local council, information about those interactions is collected and stored. The scale and complexity of all that data require ever more powerful and sophisticated tools to deliver meaningful insights.

At LIDA, researchers have been developing new techniques to deliver this information in formats that are both highly detailed – at the level of an individual – and highly useable.

The QuantiCode project started in 2016 to identify and tackle some 'real-world' data challenges, working with large organisations, including NHS Digital, Leeds City Council, and a major retailer.

Led by Roy Ruddle, Professor of Computing at The University of Leeds and Fellow of the Alan Turing Institute, with a cross-disciplinary team from computing, maths, ethics, geography, medicine and health, QuantiCode project has taken two primary approaches.

In the first strand, researchers have developed new temporal pattern mining techniques, which can search for patterns in data to characterise individual behaviours over a particular time period. What is more, the tool is robust to inaccuracies in the recorded timestamps (a common feature in many data sets).

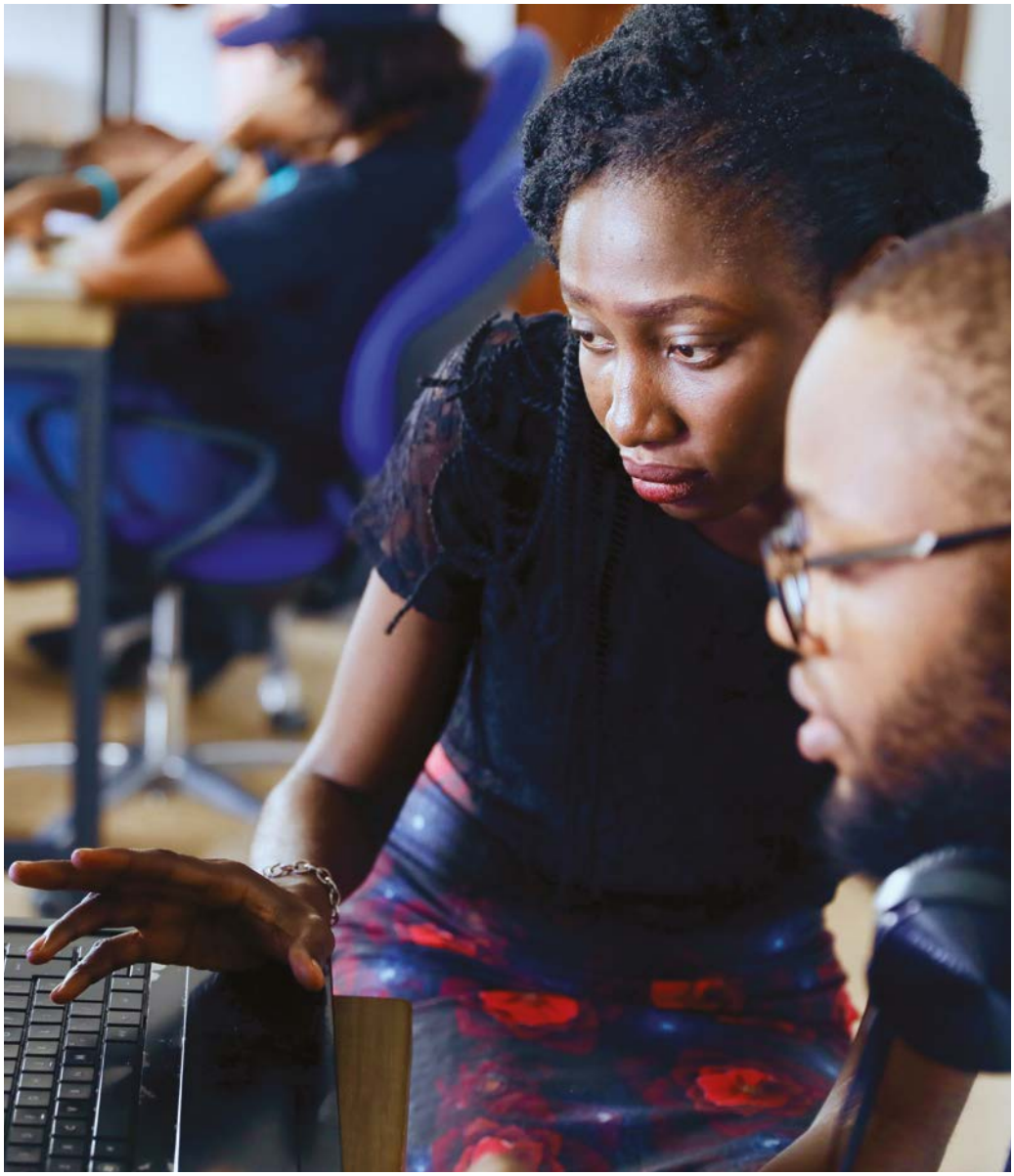
The team are using the technique to build a tool to help Leeds City Council (LCC). "*QuantiCode has enabled us to use data we collect routinely on the types of social care that people receive to better understand whom we need to support and how we can best meet their needs. We have collaborated on the project to develop new, innovative techniques to support the people receiving social care,"* says Will Ridge, a Business Intelligence Manager with LCC.

QuantiCode's second strand focuses on visualisation and developing tools to investigate data quality. Working with NHS Digital, the team has produced a data visualisation tool to investigate missing data in hospital statistics.

The tool – called ACE – is capable of rapidly interrogating datasets with millions of records. On one occasion, in just 30 minutes with ACE, users were able to identify patterns of missing data that were completely unexpected and localised to a particular unit in one hospital.

"These gaps could be due to human error or a software glitch and they occur in less than one per cent of the records," says Professor Ruddle. *"Being able to identify the source means the quality of future data will be improved. That's a win not only for the NHS but also for all of the scientists who subsequently use that data in their research."*

Both of these techniques came together in a third strand to provide a leading retailer with more sophisticated customer intelligence. QuantiCode developed analytic techniques capable of analysing hundreds of thousands of shopping combinations to uncover insights about the purpose of any individual shopping trip. This type of intelligence is invaluable in improving customer experience, driving sales and making strategic investments.



"Supermarkets are pretty good at using information from loyalty cards to understand what people are buying and when," says Professor Ruddle. *"However, they're not accurate at classifying the purpose of that shopping trip – whether I'm buying my weekly groceries, whether I've just popped out to pick up a couple of things, or whether I'm planning a party."*

Taken together, the QuantiCode team expect the techniques to offer a step change in the ever-expanding field of data analysis and continue to refine the tools and explore new ways to deliver richer insights from complex event sequence data.

CDRC INNOVATION FUND

In July 2017, ESRCs Consumer Data Research Centre launched an initiative aimed at supporting research that would capitalise on our core consumer data sets, extend our network of partners and drive substantive and innovative research across a broad range of disciplines. Through an ESRC Innovation Fund valued at £500,000, the Centre commissioned ten projects that aligned not only with the Centre's existing research themes in health and urban mobility, but extended into the ESRC's strategic priority areas such as housing and productivity.

"Working with partners at the University of Stirling and University of Sheffield has been a great way to unleash the power of our data and share the findings with the world. Without the ESRC's funding this would not have been possible, so it's a win all-round."

Alex Parsons, mySociety

INNOVATION FUND IMPACT STORY:

FixMyStreet: Micro-geographies of civic engagement and neighbourhood environmental quality

Professor Alasdair Rae, University of Sheffield
Providing local environmental services, such as repairing streets and collecting rubbish are the most basic and necessary tasks of local government. Such services are commonly addressed with through citizen-initiated reporting.

Research highlights the links between a good-quality local environment and good health and wellbeing of residents, with increased social capital and empowerment felt by residents.

Evidence suggests that residents in more affluent areas are more likely to report a problem. By relying on citizens to report problems more prosperous neighbourhoods could receive a higher level of service from their local council.

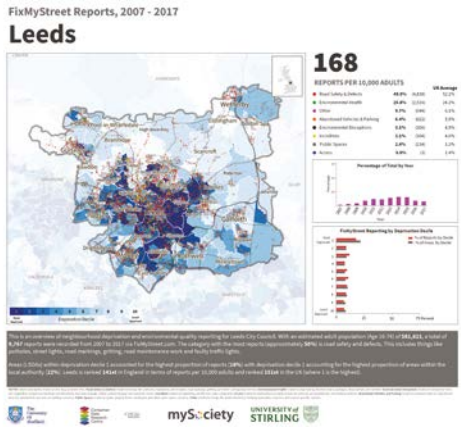
The project assesses where and what kind of requests for local environmental services come from.

Project aims
Using the data provided by FixMyStreet (c.1m records) to explore local neighbourhood conditions across the UK in relation to income, deprivation, household moves, transport, health and internet penetration.

Explaining the science
This project makes use of a large *street quality* reporting dataset (from FixMyStreet) over an eleven year period in combination with existing CDRC data assets, to explore micro-geographies of civic engagement and neighbourhood environmental quality.

For the analysis to be useful, and make sense at a local level, the project produced a high resolution map and data poster for every local authority in the UK. These show the location of reports, what category they fall into, how they compared to other areas. Comparisons were done with the number of reports in an area, to their deprivation profile.

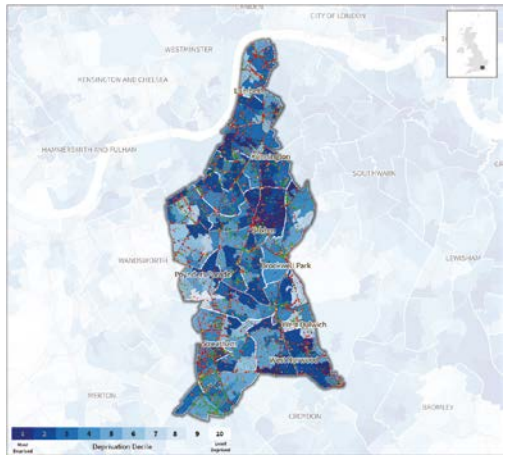
391 graphics were produced – one for each UK local authority (except county councils as these overlap with other local authorities). Each graphic covers the full period from 2007 to 2017, shows the number of reports per 10,000 adults and a breakdown of reports by type.



Findings
What the analysis of FixMyStreet data actually shows:

- Detailed visualisations of where more than 1 million neighbourhood environmental quality reports have been submitted in the UK over more than a decade.
- Improved understanding of what local residents find important in relation to neighbourhood quality. By looking at the break-down of reports by type across different areas we get an insight into what matters to people. This predominantly seems to be things like potholes, fly-tipping and dog fouling.
- How the uptake of the technology itself (i.e. the FixMyStreet website and app) has developed in different parts of the country.
- How reporting rates vary between different types of neighbourhood. This study used neighbourhood deprivation levels as a proxy for neighbourhood type, but there are other ways. Nonetheless, some important differences in reporting levels in relation to deprivation, both at the national and local level were observed.

Applications
The results will be interesting and useful for local people, councils and anyone with an interest in neighbourhood environmental quality. Interested parties can also download the data which mySociety have made available as part of this project, and conduct further analyses.



These data could be used by councils to understand whether service provision is equitable and neighbourhoods are assessed on the need rather than where the residents are most engaged. From what is known about neighbourhood environmental quality in the nation's most deprived areas, this analysis of FixMyStreet data would suggest that there is significant under-reporting of issues in those neighbourhoods which fall into the most deprived decile.

"LIDA has put the University of Leeds at the forefront of data science. It has become a national asset as a leading research and development institute, with the capacity to drive many aspects of the UK Industrial Strategy and to change lives through harnessing the power of data."

Professor Lisa Roberts, Deputy Vice-Chancellor: Research and Innovation

INFORMING PUBLIC POLICY

Connected Places Catapult

Last year the Connected Places Catapult (then the Transport Systems Catapult) introduced its first group of Business Fellows, who will build closer ties between the UK's leading transport related academic departments and the transport industry. Paul Evans, based in the Consumer Data Research Centre within LIDA was announced as one of the Business Fellows. One of the aims of the Business Fellows network is to help understand how universities engage with industry, government and the wider transport innovation ecosystem and establish their current regional, national and international impact.

IMPACT STORY:

LIDA DATA SCIENTIST INTERNSHIP PROJECT, PARTNERED WITH THE ALAN TURING INSTITUTE AND CONNECTED PLACES CATAPULT

Stelios Theophanous, LIDA Data Scientist Intern, Professor Richard Romano, Professor Nick Malleson and Dr Nik Lomax

Synergy PRIME: Multi-level modelling, simulation and visualisation

It has been estimated that more than two thirds of the world's population will live in urban areas by 2050. The ongoing and future rapid urbanisation is likely to exacerbate existing problems such as air pollution, noise pollution and traffic congestion. Current research in the field of smart mobility aims to integrate big data, innovative ideas and technologies to design a more sustainable mobility system that can provide potential solutions to these problems. A step towards achieving this is through the development of novel traffic modelling and simulation techniques that are able to generate accurate predictions of future travel demand.

Synergy PRIME is a proof of concept for integrating population growth projections, agent-based modelling and interactive traffic simulation to assist the development of future intelligent transport systems.

Project aims

Explore the possibility of integrating population projections (Mistral), agent-based modelling (Surf) and interactive simulations (Aimsun) with existing transport design methodologies. This could ultimately strengthen the potential of Aimsun as a tool for developing and testing future projected scenarios, therefore supporting the design of future intelligent transport systems.

Explaining the science

This project focuses on integrating data sources and software from three components: Mistral, Surf and Aimsun. ITRC Mistral is a large project involving the generation of high resolution population growth projections using data from the UK census and microsimulation techniques.

Simulating Urban Flows (Surf) is a project that uses 'big' data and implements agent-based modelling with the aim of improving our understanding of movement patterns in urban areas. It incorporates real footfall data to model the typical 9-5 workday travel patterns of commuters in Otley, West Yorkshire, UK.

The third component is Aimsun, a traffic simulation tool that facilitates the analysis of numerous traffic phenomena and supports the prediction of future traffic scenarios based on existing traffic data.

An integration workflow has been designed where Mistral produces the population for the Surf model, which in turn generates the trips to be simulated and visualised in Aimsun. Lastly, Aimsun generates the route that each agent follows to reach its destination.

Figure 1. A small section of the simulated road network developed in Aimsun (north part of Otley).

Findings

The Mistral/Surf integration involved generating population projections (2018 to 2041) at the Local Authority District (LAD) level and then converting them to the Output Area (OA) level, to be used as input for Surf. Upon running the Surf model using the synthetic population from Mistral, trip data was collected and converted to an XML file of traffic arrivals that is compatible with Aimsun. A digital twin of Otley and its surrounding areas was developed in Aimsun (see Figure 1), where the trips were simulated.

The case study comparing travel demand between 2019 and 2039 revealed that the integration was feasible at a basic level, however, numerous aspects require further improvement before beneficial insights regarding future travel behaviour can be yielded.

Applications

This research could assist government agencies, local authorities and consultancies design future intelligent transport systems within the smart mobility framework. The outcomes could also be beneficial to vehicle manufacturers and service providers who need to understand the future environments within which their products and services will need to operate.

Patrizia Franco, Senior Technologist in Transport Modelling, Connected Places Catapult said: *The workshops organised by the Connected Places Catapult allowed the project team to meet with all the interested parties and to define what questions the model should be able to answer in order to support Local and Transport Authorities in their priorities and needs around mobility and future trends."*

"The integration between the different tools, developed in Synergy PRIME project, allows the use of a systems of systems approach to explore future scenarios in emerging mobility trends and technologies. This is an exciting step change in the evolution of Agent Based Models as a comparable alternative to traditional strategic transport modelling approaches."

DATA ANALYTICS EXPERTISE

EDUCATION AND TRAINING

Introduction from Dr George Ellison and Dr Luke Burns, LIDA Deputy Directors for Education and Training

Our appointments as Deputy Directors for Education and Training has brought a renewed drive to LIDA's ambitions in this area. We joined the team at a time when LIDA had already firmly established itself as a unique interdisciplinary space through which the University's rich polytechnic skills base and strong network of external partners had generated a diverse and innovative portfolio of education and training opportunities.

Through traditional education routes, LIDA academics deliver no fewer than twelve research-intensive Masters programmes, two of which are embedded within its specialist Centres for Doctoral Training in: 'Data Analytics and Society' (funded by the UK's Economic and Social Research Council) and 'Artificial Intelligence for Medical Diagnosis and Care' (funded by UK Research and Innovation).

LIDA is also home to the successful Data Science Internship Programme (which welcomes its fourth cohort in 2019/20) and ever-popular Summer School (which in the coming year will run twice in July and September). At the same time, LIDA's monthly Seminar Series and regular Training & Capacity Building Workshops continue to thrive, whilst the Leeds Data Science Society is once again the Student Union's most popular, and the Leeds Critical Data Studies Group offers incisive interdisciplinary insights into the proliferation of 'big data' and data science techniques.

Our focus over the past 6-9 months has been to explore how LIDA might broaden and strengthen the delivery of training for individuals and organisations so that we remain agile and can keep pace with the rapidly evolving environment for data science. Partnerships have been developed with key local and national stakeholders across a range of public and commercial sectors, each focussing on the design and development of flexible routes through higher education. These include a new Masters of Research (MRes) scheme with separate pathways (including those in health, media, retail and finance); and a flexible and adaptable cross-programme 'credit-accumulation' scheme which will offer bespoke professional development in data science and analytics tailored to the specific needs of individual students, employers and agencies. Such approaches aim not only to widen the appeal of research-led, skills-based training but also to drive the co-production of such training with external partners to ensure that this meets the specific needs and aspirations of each workplace and each workforce. For example, working closely with the Royal Geographical Society, the Geospatial Commission and a range of graduate employers has led to the development of dedicated frameworks to enable degree apprenticeship programmes; and our partnership with NHS Digital will deliver workplace-based training in data science to 22 of its frontline data analysts over the next four years.

While these initiatives reflect LIDA's ongoing engagement in flexible approaches to postgraduate education, we are also exploring a range of ways to widen participation in data science amongst undergraduate students at the University and members of the public further afield. This has led to the ongoing development of a Year 2 'Discovery Module' in data science and analytics (which will be available from 2020), intended to support the continuing success of the Leeds Data Science Society and the appetite for data skills amongst the University's student body. Further afield, LIDA is exploring the provision of online resources developed in partnership with colleagues across the Worldwide Universities Network, which boasts members in all four corners of the globe.

The future is bright, and LIDA remains at the forefront of innovation in data science education and training.

EDUCATION AND TRAINING

One of our core priorities at LIDA is training the next generation of data scientists; in 2019 we are pleased to report that our existing CDT in Data Analytics and Society will be joined by a new centre.

UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care
September 2019 will see the launch of a new Centre for Doctoral Training hosted in LIDA, focussing on the early detection, diagnosis, treatment and care of cancer.

The Centre will support 50 fully funded PhD researcher places to unlock the potential of artificial intelligence in medical diagnosis and care. Researchers will work with key national partners including Leeds Teaching Hospitals Trust, the NHS and industry. Working with academic staff, clinicians and patients will ensure they are well-placed, highly skilled employees, ready to become the next generation of AI healthcare pioneers.

"We are recruiting talented students from a range of backgrounds, and from across science, engineering and health. While we will be focusing on cancer – where Leeds has an outstanding research and clinical reputation – our PhD graduates will be equipped with broad skills in AI to innovate in diagnosis and care in cancer and beyond." **Professor David Hogg, School of Computing at the University of Leeds.**

"We believe that AI will transform the way that medicine is practiced and how patients are managed over the next 20 years, and we want to be at the forefront of that revolution." **Dr Christopher Herbert, Director of Operations: Research and Innovation at LTHT.**

Research in the centre will span three themes where AI can be applied to advance cancer care and associated morbidities:

- **Screening and Early Detection:** exploring the use of AI in epidemiology, risk stratification and digital phenotyping, to improve screening and prevention at scale; and developing AI algorithms to process multi-faceted patient data for early detection of cancer before symptoms present.
- **Diagnosis:** exploring the application of AI to process data from pathology, radiology, wearable sensors, patient records, and genomics, leading to faster, more precise and efficient diagnosis.
- **Therapy and Care:** exploring the role of AI in the development of precision medications and novel therapies that meet the complex needs of individual patients; and improving the quality of life for patients living with and beyond cancer through the development of automated decision support tools informed by self-reported patient outcomes and audio-visual recordings.

EXAMPLE PHD PROJECT TITLES

Screening and Early Detection:
Using the CORECT-R UK-wide colorectal cancer data repository curated at Leeds to identify high-risk digital phenotypes for targeted screening.

Training recognition of skin cancer from 5 million images held at LTHT, to create a smartphone app that can be used by patients for rapid self-screening.

Diagnosis:
Using machine learning to identify which digital phenotypes could be fed into automatic image analysis of tissue scans to better highlight and categorise suspected tumours.

Analysing images automatically within digital pathology, digital radiology and digital photography to accelerate and improve accurate diagnosis.

Therapy and Care:
Risk-stratifying patients into those who would benefit from major surgery or those where minor interventions would be a better option, using machine learning from pathology images linked to genetic and clinical outcome data;

Using AI to develop individualized risk-stratified models for cancer survivor surveillance after treatment, based on supported self-management and patient-initiated follow-up.

Find out more about more about the UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care:

https://ai-medical.leeds.ac.uk/

CENTRE FOR DOCTORAL TRAINING IN DATA ANALYTICS AND SOCIETY

The ESRC-funded CDT for Data Analytics and Society provides postgraduate research training across the Universities of Leeds, Liverpool, Manchester and Sheffield.

"The CDT programme had so much more to offer than other PhD programmes. The integrated MSc develops so many technical skills: learning programming languages (Python and R); learning how to use appropriate software (Anaconda, R studio, and SAS enterprise); increasing future employability." **Jenny Grey – PhD candidate in the CDT for Data Analytics and Society**

"The CDT cohort draws together students from a wide range of academic disciplines and life experiences, allowing problems to be approached from a variety of viewpoints, thus seeing skills and approaches shared across disciplines." **Keiran Suchak – PhD candidate in the CDT for Data Analytics and Society**

Find out more about the ESRC Centre for Doctoral Training in Data Analytics and Society:

https://datacdt.org/

CASE STUDY:

Identifying the activity and habits of individuals in a large geospatial dataset – Franscesca Pontin, PhD candidate in the CDT for Data Analytics and Society

As of 2018 a predicted 21 per cent of the UK population were using fitness apps on their phone, using the inbuilt accelerometer to track their activity. Moreover, 6.3 million people in the UK own a fitness wearable; a wrist or waist worn device that tracks their activity. To date this remains a vastly underutilised source of data in physical activity research, despite the usefulness of its size and temporal coverage which exceeds the capacities of predesigned studies.

This study looks into the value of using smartphone app data as a tool to measure physical activity at a population level. Using the Bounts app data to determine firstly, how representative users are of the general population and secondly, what insights can be made from such a large dataset.

Time series analysis will be conducted to better understand the user's physical activity behaviour patterns, and use the GPS data from the users to identify how aspects of the built environment are modifying their activity behaviours.

MASTERS COURSES

Masters Studentship Awards

Two of LIDA's newest Masters programmes (MSc Precision Medicine; and MRes Data Science & Analytics for Health) – both developed in collaboration with external partners (including: GlaxoSmithKline; Viapath; Leeds Teaching Hospitals NHS Trust; and NHS Digital) – were recently awarded substantial funding from Health Data Research UK (HDR UK) to provide full-time funded studentships to three cohorts of students from 2020 onwards.

"The distinctive focus of the MSc Precision Medicine on 'omics' data will provide students with a critical understanding of the 'omics' technologies, their interpretation and application in key areas of healthcare such as cancer, rare inherited diseases and infectious diseases, as well as research." **Maha Younes – Clinical Scientist in Molecular Genetics Viapath Genetics Laboratory**

"By working in partnership with the largest local provider of training in health and health care data science skills, the translational workplace-based projects undertaken by MRes students under the supervision of interdisciplinary teams of specialists in computing, health and biostatistics, following advanced data science skills training, will reap a wealth of benefits for both partners – not least those NHS Digital data analysts and data scientists who will be supported to participate as students in the programme." **Tom Denwood (Executive Director) and Daniel Ray (Director of Data) – NHS Digital**

Announcing the studentships available for LIDA's programmes, and those delivered at five other successful applicants across the UK, **Professor Peter Diggle (Director of Training at HDR UK)** wrote:

"The programmes will genuinely integrate statistics, informatics and health science and… bring us one step closer to building a community to lead the health data science revolution."

Health:
For more information or to be notified of the call for applications, please contact: **lida@leeds.ac.uk**

Interdisciplinary Masters in Data Science and Analytics
Elsewhere, LIDA's growing portfolio of research-driven Masters courses attracts students from around the world to develop their skills and understanding in the application of advanced data science and analytics techniques, including:

- MSc Advanced Computer Science (Data Analytics)
- MSc Business Analytics and Decision Sciences
- MSc Consumer Data Analytics
- MSc Data Science and Analytics
- MSc Geographical Information Science – NEW ONLINE FOR 2019
- MSc Financial Technology – NEW for 2019
- MSc Health Informatics
- MSc Health Data Analytics
- MSc Precision Medicine: Genomics and Analytics – NEW FOR 2019
- MRes Data Science and Analytics for Health – NEW FOR 2019

NEW MASTERS COURSES FOR 2019/20

MSc in Financial Technology
Financial Technology – or fintech – combines technological innovations with financial systems including banking, capital markets and global payment infrastructures. It is rapidly reshaping the financial services industry: from the reinvention of retail banking to providing services via smartphones; to statisticians collaborating with artificial intelligence experts to design and build new systems.

Delivered through the Leeds University Business School this part-time Masters degree is designed to ensure working professionals are equipped to navigate, and adapt to, the shifting financial landscape, covering areas such as cryptofinance, blockchain, big data, artificial intelligence, computer programming, cybercrime and machine ethics.

MSc in Geographical Information Science (online)
The MSc in Geographical Information Science (GIS) offers a part-time, online Masters programme providing in-depth specialist knowledge of GIS techniques. Delivered through the School of Geography, students will benefit from a flexible approach to learning designed to fit the needs of working professionals anywhere in the world. Teaching will use a mix of approaches including videos, handbooks, practical guides and discussion forums; with all assessment completed via coursework and submitted online.

Core modules will introduce students to key GIS packages and a range of data sources/techniques, of direct use to those undertaking social and environmental research and those keen to understand the application of geotechnology as practitioners. Students will study the theories and concepts underpinning GIS and will be able to select from a range of contemporary optional modules including those on advanced visualisation, digital image processing, and geodemographics.

Find out more about more about the online MSc in Geographical Information Science
https://courses.leeds.ac.uk/d985/geographical-information-science-msc

MSc in Precision Medicine: Genomics and Analytics
Biomedical research generates large volumes of complex data, encompassing genomics, proteomics, metabolomics, phenotypic data, epidemiology and clinical trial investigations, to inform our understanding of disease mechanisms.

This programme was developed in collaboration with partners in industry (including GlaxoSmithKline) and the NHS. It aims to ensure that scientists have the skills to utilise these data sets in the development of precision medicine. These data impact on the creation of new tailored therapies, aiding earlier diagnosis and the selection of optimal treatment regimens for patients who may be suffering from disorders ranging from common disorders of complex aetiology (such as cancer or cardiovascular disease) to the more rare Mendelian disease (such as muscular dystrophy).

Delivered through the Faculty of Biological Sciences, this programme has been designed to meet research needs in industry and academia where there is a demand for scientists with both biological knowledge and the computational, statistical and analytical skills to drive genomic precision medicine.

Find out more about more about the MSc in Precision Medicine: Genomics and Analytics
https://courses.leeds.ac.uk/i661/precision-medicine-genomics-analytics-msc

MRes in Data Science & Analytics for Health
The MRes in Data Science & Analytics for Health is the first programme in LIDA's innovative MRes scheme, which has been developed in partnership with the Data, Insights and Analytics Directorate at NHS Digital and will be delivered through the School of Computing.

The programme will focus on providing workplace-based training in data science and analytics for health with full – and part-time students (including data analysts and scientists from NHS Digital) learning core competencies in: data science; prediction and causal inference; machine learning; and artificial intelligence techniques.

Students will then be free to choose up to two additional, optional modules (including those in research software engineering and professional data science practice, and those developed for LIDA's Masters in Precision Medicine and Health), before embarking on workplace-based applied research projects in health improvement for patient benefit with NHS Digital and their health sector partners.

Find out more about more about the MRes in Data Science & Analytics for Health
https://courses.leeds.ac.uk/i661/data-science-analytics-health-mres

DATA SCIENTIST INTERNSHIP PROGRAMME

As part of our ongoing commitment to developing data science capability, LIDA's Data Scientist Internship programme offers 11 paid interns first-hand work experience with a diverse range of multi-disciplinary applied research teams. The aim of the programme is to grow data science excellence, foster innovation and collaboration, and provide interns with opportunities to better understand the challenges and potential benefits of applying complex analytical techniques to real-world data in real-world contexts.

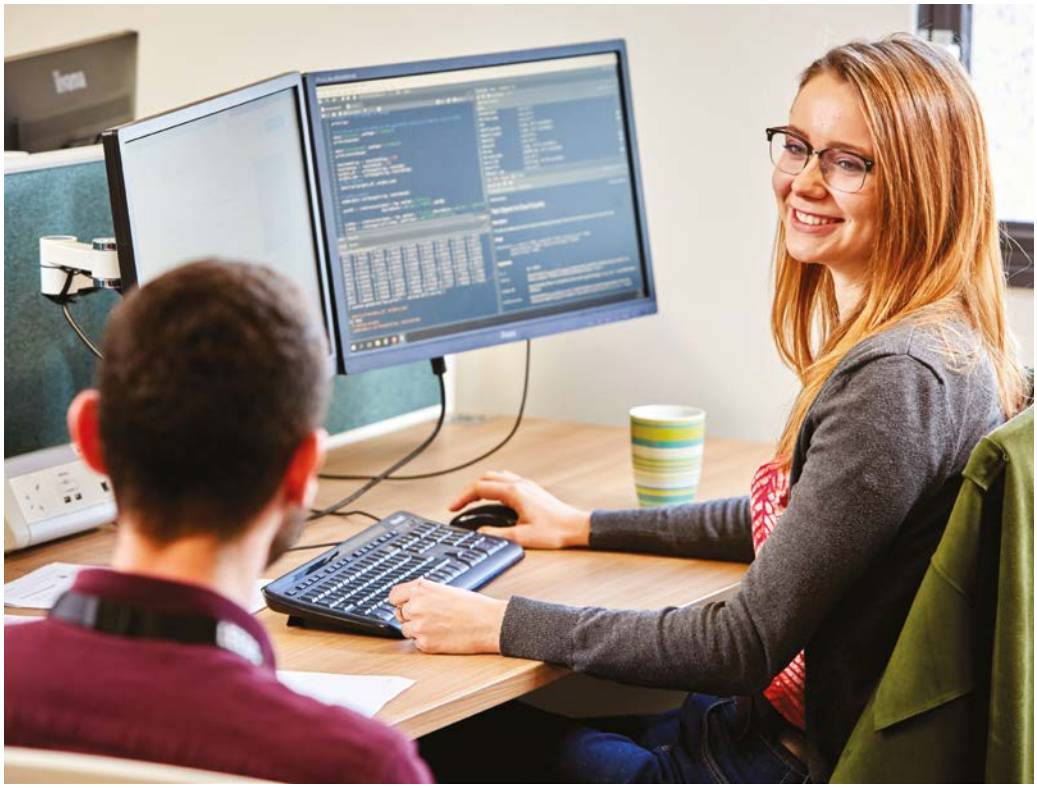
During their year-long placement, LIDA's Data Science Interns work on two research projects, joining existing research teams and external partners from the public, private and voluntary sector, under the supervision of experienced academics from the University of Leeds. To date, Interns have worked on projects co-supervised by medical technology research teams, high street retailers, international technology companies and local authorities – exploring, specifying and harnessing data science solutions to pressing social and commercial issues.

"I have thoroughly enjoyed my time as a LIDA Data Science Intern. This programme is exactly what I needed to combine my PhD research skills with industry projects. I am confident I will finish this year perfectly positioned for my next career move." **Dr Kevin Minors – LIDA Data Science Intern**

"The LIDA Data Science Internship programme has provided me with the unique opportunity to develop valuable skills through being able to deliver my own data science project using real-world data and having the opportunity to attend data science workshops. I have thoroughly enjoyed being able to build upon my data analysis skills through attending LIDA workshops and being able to gain experience on how data science is used in differing fields by work on two separate projects which were health – and geography-based. One of the highlights of taking part in this internship is working within a multi-disciplinary team and having the ability to share ideas and work through problems faced as part of a team." **Rizwana Uddin – LIDA Data Science Intern**

Intern Natacha Chevenoy received an award from the International Association of Law Enforcement Intelligence Analysts (IALEIA) for her work with the Lancashire Constabulary identifying online hate crime. The IALEIA Award for Excellence recognises individuals for "outstanding contributions as intelligence analysts, investigators, or prosecutors utilising intelligence products leading to the achievement of the organisation's objectives."

The majority of those who completed the programme have gone on to further research or employment in data science roles in a variety of sectors including: government agencies; engineering, software and consultancy firms; and in the banking and financial sector.



SAMPLE PROJECTS

Extracting actionable insights from free text police data
This project is developing new methods for analysing free text police data. It explores state-of-the-art natural language processing techniques to develop machine-learning methods that can be used by the Safer Leeds initiative and its policing partners to leverage valuable information recorded in free text police data that might otherwise be lost.

SPENSER – A synthetic population estimation and scenario projection model
SPENSER is a synthetic population estimation and projection model generated using dynamic microsimulation. It provides the framework for estimating characteristics of a population that are both dynamic and at high resolution (i.e. at household – and individual-level); together with a comprehensive set of tools for user-customisable scenario projections. The interactive interface allows users to set assumptions about the future (e.g. around economic, policy, health changes) which are then translated to underlying demographic constraints (e.g. mortality, fertility, migration).

Predicting and warning extreme wind response of bridges using advanced data analytics
Using unique wind response information for two major bridges – the historic Clifton Suspension Bridge (UK) and the airport-linking Ting Kau Bridge (HK) – the principal aim of this project is to understand and model extreme wind-induced bridge behaviours that can be implemented within rationalised multi-measurement early warning systems that exploit existing monitoring systems – systems that might then be generalised to other bridges globally, contributing to better informed decision-making for critical building infrastructure. Of particular merit to the approach adopted by this project is the understanding it will provide of complex topographical effects and unanticipated anomalies such as instances where high wind speeds do not substantially exit these structures.

IMPACT STORY:

LIDA DATA SCIENCE INTERNSHIP PROJECT, UNDERTAKEN AS PART OF THE QUANTICODE RESEARCH PROJECT

Ivana Kocanova, LIDA Data Scientist Intern, Dr Muhammad Adnan, Dr Georgios Aivaliotis, Professor Roy Ruddle

A visual analytics workflow for investigating customers' transactions in convenience stores

"What products do our customers buy together?" is a common question that retailers want to answer because it provides them with insights about customers' shopping behaviour that underpin strategic investments. However, extracting such information is often a challenging task due to the sheer complexity of the data.

This project investigates how novel data mining and visualization techniques might speed up the analysis of shopping behaviour, and proposes a visual analytics workflow for investigating customers' transactions.

Project aims

The work was conducted with a major retailer. Our contributions are:

(a) describing a visual analytics workflow that combines multiple event mining methods together with visualization to analyse large-scale event data; and

(b) evaluating the workflow with real-world data.

Explaining the science

Efficient algorithms for discovering relationships among the products present in separate transactions are crucial for understanding shopping behaviour. The dataset used contained 140,986 unique product combinations from 365,756 separate transactions. The high number of unique transactions is the reason why insight discovery is such a complex task in retail settings. Consequently, we aim to simplify the transactions by dividing them into smaller, more manageable building blocks.

High Utility Itemset Mining (HUIM) searches for recurring product combinations that customers place in their shopping baskets. Since HUIM could potentially produce as many as 2m-1 unique combinations of products (or 'itemsets') for dataset with m unique products (thus 512 combinations for just 10 products, or 524,288 for only 20 products, and so on…), some sort of user-specified utility threshold is usually required. The threshold represents a trade-off between the number of itemsets and the percentage of the dataset that is covered by those itemsets.

Results
By using itemsets as the building blocks of transactions we have shown that the complexity of the data can be significantly reduced. Setting the utility threshold to 1,068 enabled us to cover 95 percent of transactions with only 1,191 itemsets. This presents a 100-fold reduction in the number of unique patterns.

Having obtained these itemsets, a user may then visually investigate the composition of transactions and create various transaction 'sketches'. Such sketches analyse the transactions from the product perspective (e.g. what products are often bought together), but also from a time-perspective (e.g. what products are bought at a particular time of day).

Applications
Understanding interrelations amongst products could be leveraged by retailers in numerous ways, particularly in smaller convenience stores. Here the space available for products is limited and it is therefore important to offer a range of products which best satisfies the majority of customers' needs; while a related benefit concerns the layout of products within a store which is heavily influenced by, and can influence, the product interactions observed.

COURSES AND CAPABILITY

LIDA invests significant resources into developing data science capability at all levels, across both academia and industry, and over the past 12 months we have hosted courses attended by over 300 data scientists. These range from introductory courses for postgraduate students from the University of Leeds and further afield, through to advanced training for data scientists working in local, regional and national public, private and voluntary sector organisations, as outlined beow:

Introductory courses
These one day courses are suitable for academic and non-academic researchers, and provide an introduction to popular tools for spatial and data analytics, including:

- R
- Python
- ARC GIS (academic only)
- QGIS
- GIS for Retail Data Analytics
- Spatial methods for public health researchers
- Tableau (academic only)

Advanced training for data scientists
These intermediate and advanced courses are designed for academic and non-academic researchers who have completed our related introductory course or have first-hand practical experience.

- Intermediate R
- Geocomputation with R
- R for Transport
- Causal Inference with Observational Data: the challenges and pitfalls (in collaboration with The Alan Turing Institute)

"NHS Digital has identified Python as a key data handling and analysis platform. It aims to use it as a key analysis tool in the future and therefore this course gave me a good initial insight into how to use it." **Python Introductory Course attendee**

"Incredibly well taught course in Python. Would definitely return for more from the same tutor and assistants. Really insightful and helpful. The fact it was over two days gave time for the information to sink in. I really enjoyed it and feel I got a lot out of it." **Python Introductory Course Attendee**

"Well organised course in R. Structured with little didactic teaching but encouraged straight away to try practical examples with a very clear prompt sheet. Really emphasised the 'learn by doing' method which worked very well in this instance. I work in the clinical aspects of hospital medicine where we are routinely expected to do data analysis but do not have access to statistical packages. Learning R is a great way to access open source tools for free and the data visualisation is definitely much better than many other programs." **Introduction to R Course Attendee**

CAUSAL INFERENCE SUMMER AND SEPTEMBER SCHOOLS

This five-day summer school has grown in popularity since it was established four years ago. Such is the oversubscription that an additional September School has been introduced. The school is delivered in partnership with the Alan Turing Institute and has recently been accredited with a Continuous Professional Development (CPD) certification.

The school is run by Professor Mark Gilthorpe and Dr Peter Tennant (both Fellows of the Alan Turing Institute) through a mix of lectures, discussions, and interactive workshops – blending theory with real-world examples, providing an essential introduction to the analysis of 'big data'.

By exploring the philosophy and utility of directed acyclic graphs (DAGs), participants learn to recognise and avoid a range of common pitfalls in the analysis of complex causal relationships, including the longitudinal analyses of change, mediation, nonlinearity and statistical interaction.

"The course is an absolute must for anyone who is serious about improving their own knowledge of data analysis." **Causal Inference Summer School Attendee**

"The course was a paradigm shift for me in terms of thinking around causal inference and gave me the tools to think about some important pitfalls in analysis that I would otherwise have missed." **Causal Inference Summer School Attendee**

LIDA Seminar Series
The LIDA seminar series provides a regular opportunity for colleagues to come together and hear about the latest developments in and applications of data analytics across a wide range of disciplines.

In the last year we have welcomed over 40 speakers, 38 percent of which have visited from other higher education institutes, research or business organisations, with just over 500 attendees.

Range of seminar talks from the last year:

- Dealing with variable misclassification in the results section, not the discussion section: An introduction to quantitative bias analysis – **Professor Matthew Fox, Boston University**
- Data quality and causal inference in Learning Health System: some challenges for developing reliable algorithms in an imperfect world – **Professor Jeremy Wyatt, University of Southampton**
- Building a Digital Twin: Testing the effectiveness of telecommunications policies in a virtual world – **Dr Edward J Oughton, University of Oxford**
- The continental divide? Economic exposure to Brexit in regions and countries on both sides of the Channel – **Professor Raquel Ortega-Argilés, University of Birmingham**

LEEDS DATA SCIENCE SOCIETY

Formed in 2015, the student-led Data Science Society now has more than 500 members across the University – Leeds Student Union's most popular student society. Members include developers and data scientists as well as those who are simply curious about data science and are keen to learn more about the systems and processes involved in the coming information age.

The society hosts regular training courses, networking and careers events for members; and over the past year introduced the Data Science Engagement and Employability Conference (Data SEEC), held on 8 May in partnership with LIDA. The Data SEEC was designed to facilitate successful job placements and encourage stronger connections between the University, the Data Science Society and the vibrant local data science industry; and was attended by over 60 students and businesses.

This event was specifically targeted at the data science and analytics sector, providing an exclusive opportunity for students to showcase their data science skills and work to representatives from prestigious employers, who in turn discussed relevant employment opportunities.

https://www.leedsdatascience.com/

LEEDS CRITICAL DATA STUDY GROUP

The Leeds Critical Data Studies Group is a multidisciplinary collective of postgraduate and early career researchers from across the University of Leeds who are working on the critical interrogation of data systems, practices and logics.

The core activities for the group include organising events and workshops to discuss current research, and to learn new research techniques. Events and workshops include: regular seminars hosted within LIDA's weekly Seminar Series where issues relating to critical data studies in each of our parent disciplines are debated and discussed; and specialist workshops focussing on the extraction, analysis and interpretation of data from online platforms and other websites. Since its inception in 2016, the academic diaspora of the Group's original members has provided an international network of critical data scholars offering unique insights and opportunities for cross-cultural interdisciplinary collaboration in this rapidly evolving and critically relevant field.

http://datastudies.leeds.ac.uk/

UNIVERSITY OF LEEDS

University of Leeds
Leeds, United Kingdom
LS2 9JT
+44 (0) 113 243 1751
www.leeds.ac.uk

Leeds Institute for Data Analytics
Level 11, Worsley Building
Clarendon Way
Leeds
LS2 9NL
+44 (0) 113 343 9680
www.lida.leeds.ac.uk
lida@leeds.ac.uk
@LIDA_UK