ai@cam

ai.cam.ac.uk

Access to data for research: lessons for the National Data Library from the front lines of AI innovation

In partnership with:

MINDEROO CENTRE FOR TECHNOLOGY & DEMOCRACY

Bennett Institute for Public Policy Cambridge

Foreword

The role of AI as a catalyst for a new wave of research is the subject of growing policy interest. High-profile advances in domains such as protein folding – and the Nobel Prizes that have followed – illustrate the potential of AI to accelerate discovery. At a time when Government is refreshing its AI policy priorities, this report explores how data enables such innovation through a series of case studies from research groups across Cambridge.

These case studies show innovative uses of data for research in areas that are critically important to science and society, including:

- Improving diagnosis of ovarian cancer: by using patient data to build AI-enabled diagnostic tools, Dr Mireia Crispin's group aims to improve healthcare outcomes for patients with ovarian cancer.
- Understanding the impact of social media on mental health: Dr Amy Orben's team has developed an innovative approach to accessing social media data to analyse the impact of technology use on child mental health.
- Connecting conservation evidence to policy decision-making: by training advanced AI systems on a large dataset of academic literature, Dr Sadiq Jaffer and Dr Alec Christie are building an evidence synthesis tool that helps policymakers navigate the complex evidence base relating to conservation policy.
- Forecasting the impact of climate change: Dr Scott Hosking's team is combining data from satellites, drones, ocean robots, and more, to create a digital twin of the Polar regions that can help researchers understand the impact of climate change there, and help policymakers protect local communities and wildlife.
- Managing public services: as the healthcare crisis surrounding the COVID-19 pandemic unfolded, Professor Stefan Scholtes and his team worked with healthcare leaders across the East of England to extract insights from patient and hospital management data that could inform decisions about how to most effectively direct NHS resources.
- Tracking patterns of economic activity: working with the Office for National Statistics, Professor Vasco Carvalho and colleagues are showing how alternative data sources can help policymakers understand and respond to economic disruptions.

These case studies also illustrate the continuing barriers that researchers face in accessing data. Low levels of data maturity in many areas mean it remains difficult to find, access, and combine data resources. Even where data is well-curated, legal and technical barriers to data access – or a lack of specialist support to overcome these – hold back its use for research. Overcoming these barriers is possible. The projects highlighted in this report can help policymakers understand how.

The National Data Library offers an opportunity to bring renewed focus to efforts to unlock the value of data for research. To be successful, the Library will need to translate high-level policy intentions to use data for public benefit to practical actions that unblock data access and use. Experiences from the front-line of AI innovation show the importance of co-designing the proposed Library with communities of research and practice who understand those practical actions, and aligning its ways of working with public interests and concerns. They also indicate the potential for the National Data Library to act as a rallying point that brings together priority data resources and the UK's research capabilities to drive progress in areas of need.

Translating the potential of AI for research to real-world practices will require concerted action to address a complex web of legal, technical, cultural, and organisational barriers that affect who is able to access data and for what purposes. By centring the experiences of researchers on the front-line of AI innovation, this report hopes to bring some of these barriers into focus. These case studies were inspired by engagement with the Department for Science, Innovation, and Technology on current issues in data policy, and we hope this report will be useful in informing continuing conversations in this area. Thank you to the researchers that contributed case studies about their work.

J. Montgomery

Jessica Montgomery,
Director, ai@cam,
University of Cambridge

N. Lawrence

Neil Lawrence,
DeepMind Professor
of Machine Learning,
University of Cambridge

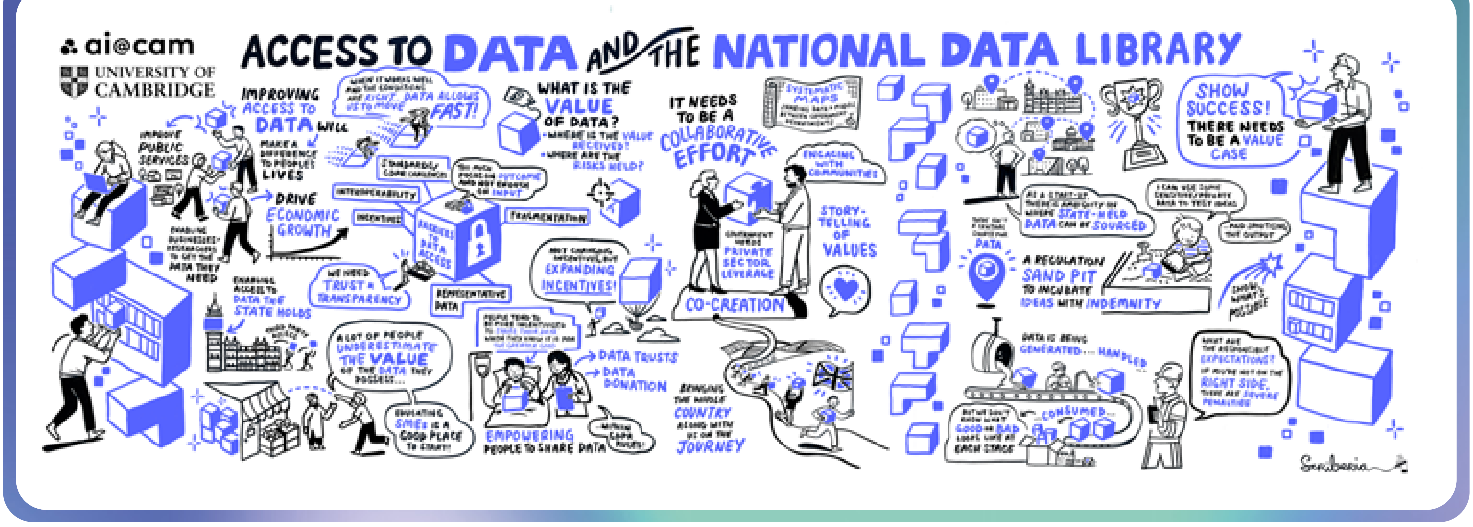
D. Coyle

Diane Coyle,
Bennett Professor of
Public Policy, University
of Cambridge

G. Neff

Gina Neff, Professor
of Responsible AI,
Queen Mary University
of London, Executive
Director of the
Minderoo Centre
for Technology &
Democracy, University
of Cambridge,
and Deputy CEO,
Responsible AI UK

Summary


ai@cam
UNIVERSITY OF CAMBRIDGE
ACCESS TO DATA AND THE NATIONAL DATA LIBRARY
IMPROVING ACCESS TO DATA WILL
IMPROVE PUBLIC SERVICES
MAKE A DIFFERENCE TO PEOPLE'S LIVES
DRIVE ECONOMIC GROWTH
ENABLING BUSINESSES + RESEARCHERS TO GET THE DATA THEY NEED
ENABLING ACCESS TO DATA THE STATE HOLDS
DATA ALLOWS FAST!
INTEROPERABILITY
INCENTIVES
BARRIERS TO DATA ACCESS
FRAGMENTATION
REPRESENTATIVE DATA
WE NEED TRUST + TRANSPARENCY
A LOT OF PEOPLE UNDERESTIMATE THE VALUE OF THE DATA THEY POSSESS...
EDUCATING SMEs IS A GOOD PLACE TO START!
WHAT IS THE VALUE OF DATA?
- WHERE IS THE VALUE RECEIVED?
- WHERE ARE THE RISKS HELD?
EXPANDING INCENTIVES!
DATA TRUSTS
DATA DONATION
EMPOWERING PEOPLE TO SHARE DATA
IT NEEDS TO BE A COLLABORATIVE EFFORT
SYSTEMATIC MAPS
ENGAGING WITH COMMUNITIES
GOVERNMENT NEEDS PRIVATE SECTOR LEVERAGE
CO-CREATION
STORY-TELLING OF VALUES
BRINGING THE WHOLE COUNTRY ALONG WITH US ON THE JOURNEY
SHOW SUCCESS! THERE NEEDS TO BE A VALUE CASE
AS A START-UP, THERE'S AMBIGUITY OF WHERE STATE-HELD DATA CAN BE SOURCED
DATA
A REGULATION SAND PIT TO INCUBATE IDEAS WITH INDEMNITY
SHOW WHAT'S POSSIBLE
DATA IS BEING GENERATED... HANDLED... CONSUMED...
GOOD OR BAD
WHAT ARE THE EXPECTATIONS? IF YOU'RE NOT ON THE RIGHT SIDE, THERE ARE SEVERE PENALTIES
Scribania

Access to data is a vital enabler of AI innovation. How to govern data in a way that delivers widespread public benefit is a long-standing policy challenge. It requires careful negotiation of a suite of policy issues – from privacy to ownership – that each have legal, technical, ethical, and organisational elements, while aligning with societal interests or concerns. In navigating this challenge, on-the-ground experiences from innovative AI projects offer important insights into what practical action can be taken to increase access to data for public benefit while maintaining appropriate safeguards around data use. Inspired by current discussions about the UK's AI policy landscape – and the opportunities presented by initiatives such as the National Data Library – this report presents a series of case studies exploring the value that access to data for research can create, and the barriers researchers face in accessing such data.

These case studies show a diverse landscape of AI research aiming to deliver benefits for science, citizens, and society. The projects represented here are tackling vital challenges in improving cancer treatment, understanding the impacts of climate change, protecting biodiversity, supporting economic growth, and understanding the impact of technology on society. Their work is creating opportunities to deliver public benefit from data and AI, using a range of data types that include both statistical data and text. The research set out in this report is helping healthcare systems to benefit from new diagnostic tools that deliver better patient outcomes, policymakers to improve the evidence base for decision-making, and organisations to improve their operational processes. These case studies also illustrate the barriers that researchers are grappling with, and suggest how a new wave of policy development could help address these.

A shared theme across these projects is the importance of increasing the UK's data readiness. This means curating datasets that are findable, accessible, and interoperable in areas of strategic importance. Dr Crispin's work developing diagnostics for ovarian cancer, for example, illustrates the potential value in different forms of patient data, and the distance still to travel to make such data available in digital forms that can be shared across trusted partners in research and healthcare. Tackling these challenges will require investment in the development of strategic datasets in areas where there is potential to generate widespread public benefit. Delivering that technical work in turn requires organisational capability-building. The success of COVID-response projects led by Professor Scholtes show how political will and organisational leadership can drive rapid progress to overcome barriers to data access. Initiatives such as the National Data Library could provide a focal point for such leadership.

Making data and AI tools accessible to a broader community of researchers relies on access to skilled data managers and engineers. Dr Hosking's climate science research illustrates how – even in areas of research with a track record of creating open data resources – there is a need to develop human capital. Data managers who understand how to steward data resources and software engineers who understand how to build the tools that allow effective use of those resources are the vital interfaces that allow researchers to create value from data. Such resources should be factored in to the design of a national data initiative.

These case studies also show the need for a digital infrastructure that connects data, researchers, and resources that enable data use. Components of such an infrastructure might include:

→ A function that brokers data access agreements: Intermediary organisations can help bridge between open research and private datasets. To pursue her research on the impact of social media use on mental health, Dr Orben has leveraged a combination of research participants' personal data rights and intermediary organisations. This has allowed her to access insights from privately held datasets that would otherwise be inaccessible. A collaboration with the Office for National Statistics has helped Professor Carvalho analyse banking transaction data that offers game-changing insights for economics research. In each case, an intermediary has provided technical expertise, practical support, and governance structures that unlock access to private sector data.

- An archive for open data resources that can be mined or used for research: Dr Sadiq Jaffer and Dr Alec Christie's innovative research to build a Large Language Model for conversation policy shows how AI can help policymakers extract insights from large volumes of academic literature that might otherwise be difficult to use in decision-making. Their work also suggests how technical or organisational barriers to data access could be overcome through well-managed archives of open access research.
- A trusted research environment that combines data, software, and compute: Across these case studies, researchers point to the value of software and compute to support AI deployment, and the need for research infrastructures that make such facilities available.

The experiences from the front-line of AI innovation that are set out in these case studies highlight important design considerations for the UK's National Data Library.

- Co-design: The National Data Library has a broad range of potential users across research disciplines, in the public sector, in civil society, and in business. Co-designing the agenda for the Library with these users will be important in aligning stakeholder interests and prioritising areas for action. It can help identify strategically important data resources, the barriers to using these resources, and the actions to overcome those barriers. Bringing public voices to the table is also vital in connecting the innovation supported by the Library to areas of societal interest. ai@cam's recent public dialogues on the role of AI in the Missions for Government[1] show there is a growing desire from members of the public to have a say in how data and AI are used, as part of a suite of measures to increase democratic control over AI development.
- Convening: Data can act as a rallying point for the research community to drive progress tacking critical challenges in areas of social need. There is an opportunity for the Library to convene mission-led projects that catalyse innovation, and in so doing build coalitions across research, the public sector, and business that translate AI innovation to widespread public benefit.
- Connection to practice: Touchpoints or feedback loops with user communities can help Government understand whether the proposed benefits of the Library are being delivered in practice. Strategic data initiatives too often fail to translate high-level objectives or principles into practical interventions. By connecting experiences of real-world innovation in AI back to policy design and implementation, the Library can help bridge this gap.

The National Data Library could help create a public infrastructure for innovation that connects AI research to widespread public benefit. Success will require collaboration across research, policy, and practice to bridge between policy ambitions for the Library and the communities that can translate its resources to beneficial AI innovations.

[1] See https://ai.cam.ac.uk/projects/public-dialogues [accessed 1 November 2024]

Contents

Diagnosing the data challenges in cancer research

Dr Mireia Crispin is an award-winning Assistant Professor in the Department of Oncology at the University of Cambridge with a PhD in Particle Physics.[2] Dr Crispin's group at the University's Early Cancer Institute[3] is spearheading a new approach to cancer research and treatment. She and her colleagues are bringing together diverse datasets to decode one of the most fatal and least studied forms of gynaecological cancer: high-grade serous ovarian cancer.

Drawing on information ranging from radiological imaging to biopsy data, patient demographics, treatment history, and tumour DNA markers, Dr Crispin and her team are developing sophisticated AI tools that personalise cancer diagnosis and treatment. These tools allow them to gain a deeper understanding of each patient's unique cancer and to predict how patients might respond to treatment.

"Once these techniques are more mature and have been validated," Dr Crispin says, "the vision is that women who have been diagnosed with ovarian cancer will be able to find the best possible treatment for them and for their specific cancer."

What used to take hours of painstaking research for radiologists reviewing radiology images and CT scans can now be done at pace with AI – allowing huge amounts of data to be processed and analysed within a matter of minutes. The highly-curated cohort of patients Dr Crispin is working with today numbers more than 1,000 women who have been treated for ovarian cancer. And she is also analysing data from a clinical trial involving 600 patients – numbers that were unthinkable even a few years ago.

Data roadblocks

However, new challenges are emerging when it comes to accessing the data Dr Crispin needs for her cutting-edge research. The infrastructure for accessing the data doesn't seem to be keeping pace with the speed of scientific progress. Although clinical health records are digitised for patients in Cambridge, which makes them relatively easy to access in theory, that isn't always the case for historical pathology records. For retrospective studies, for example, some of the older biopsy and medical images Dr Crispin and her team needs have to be retrieved from data storage in Wales and scanned in manually, making data curation a long and arduous process.

The original patient data, which is owned by the NHS, is highly confidential. Data from studies approved for analysis are extracted and stored anonymously on University servers in a secure data environment. Only researchers with valid NHS letters of access and research passports can access the data. Dr Crispin's team works with a dedicated data manager who liaises with the NHS to make sure that the data is retrieved in a structured way that follows NHS frameworks and protocols assiduously. The data manager, who then also cleans and curates the anonymised data so that it's suitable for specific research purposes, is paid for by Cancer Research UK (CRUK), which funds Cambridge Cancer Centre infrastructure costs. Any CRUK-funded research also needs to follow FAIR principles – which stands for Findable, Accessible, Interoperable and Reusable - and is designed to ensure fairness, inclusivity and transparency in research.

Embedding these principles in the Centre's data management enables Dr Crispin's team to publish enough data for each research project through scientific research papers so that the scientific community can reproduce and verify the results. Although each project varies in terms of publishing protocols, replicability is Dr Crispin's minimum threshold when it comes to sharing data, which includes fully anonymised patient biomarkers and images. These datasets can be stored in public databases, the University's own data sharing system (Apollo), or together with the code as part of the software repository.

However, there are several other challenges when it comes to accessing data that hamper Dr Crispin's work: "The barriers are generally that the data is not set up to be mined on a large scale," says Dr Crispin. "Both from the point of view of how the data was taken in the first place and also from the point of view of the infrastructure, people and facilities that would enable you to rapidly access those data."

At Cambridge University Hospitals NHS Foundation Trust, for example, there are only a limited number of people who are permitted access to the highly-protected patient data that researchers across the University need for their work, which inevitably leads to long delays. It's also hard to find data managers who are suitably qualified and willing to work for a university or an NHS trust for a relatively modest salary when they could be earning far more in industry, Dr Crispin says.

This challenge is exacerbated by the fact that clinical IT infrastructures are completely separate from University computers, unlike competitors in Europe and the US. This causes two main problems. Firstly, the NHS IT infrastructure is not set up to deal with data-heavy research in the clinic, which makes returning prognostic models back from the lab into the clinic very challenging. While the University has access to world-leading super computers, the NHS doesn't have access to the kind of computational infrastructure that researchers ideally need. Dr Crispin's colleague, Director of Clinical Integration Dr Sarah Burge, puts it this way: "It's like building a Formula 1 engine in the lab, and expecting it to fit into a Fiesta chassis in the clinic."

Secondly, moving data between the two different governances – even with all the ethical approvals in place – is a challenge, involving data transfer agreements that have to be constructed anew each time: "This just doesn't exist in countries and hospitals where the research and clinical IT domains are under one governance system," says Dr Burge.

There are also barriers to collaboration between different NHS trusts when it comes to sharing clinical data. "It's extremely painful, in my experience," says Dr Crispin. "You effectively have to go to each hospital one by one and set up meetings and deal with completely different governance structures and data management systems … and there's no guarantee that you're ever going to get the full data."

She contrasts this approach to clinical data in the UK with her experience as a postdoctoral researcher in the US, where access to clinical records for strictly defined research purposes was much more straightforward. And in her original area of research, particle physics, data is shared by scientists in a far more collaborative way: "Everyone has access to everything," she says. "Everything is shared."

Hope on the horizon

With ambitions to build a new cancer research hospital in Cambridge, however, there are plans in process to set up a centralised way to query patient data sets both internally and across the region in a more synchronised way.[4] Dr Crispin also cites a new Electronic Health Record Research and Innovation Database (ERIN) in the pipeline, which will include all patient-level data collected as part of providing patient care at the NHS Trust in Cambridge.[5]

However, there's far more that could be done at a national level, according to Dr Crispin, to harness the huge potential of AI and data for vital medical research like hers: "I think it should be a priority to have hospitals that are data ready and it should be a priority to have data management and IT teams with specific responsibilities around enabling and facilitating data management for research," she says. "I think it's clear to everyone that these powerful technologies could be very, very useful if applied properly in the medical setting. But you will never be able to do it properly unless you have some boots on the ground working within the NHS environment to make it possible. I know the budgets are stretched, but if we want to future-proof all of this, then there needs to be some investment in it."

Dr Crispin also welcomes the proposal to create a National Data Library in the UK, citing existing resources such as the UK Biobank[6] and databases like the Cancer Genome Atlas (TCGA)[7] and the Cancer Imaging Atlas (TCIA)[8], which are used heavily by cancer researchers even though they draw on a small set of data.

"If there was a National Data Library that was fully comprehensive that we could use to test hypotheses or ask new questions, I think it could be huge" she says. "It would not be unprecedented – OpenSAFELY[9], led by the University of Oxford, is a great example that was developed during the COVID-19 pandemic. That is the direction we should be moving in."

[1] See https://ai.cam.ac.uk/projects/public-dialogues [accessed 1 November 2024]

[2] See https://www.earlycancer.cam.ac.uk/dr-mireia-crispin [accessed 1 November 2024]

[3] See: https://www.earlycancer.cam.ac.uk [accessed 1 November 2024]

[4] See https://www.eoe-securedataenvironment.nhs.uk/index.html [accessed 1 November 2024]

[5] See https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/erin-ehr-research-and-innovation-database/ [accessed 1 November 2024]

[6] See https://www.ukbiobank.ac.uk [accessed 1 November 2024]

[7] See https://www.cancer.gov/ccg/research/genome-sequencing/tcga [accessed 1 November 2024]

[8] See https://www.cancerimagingarchive.net [accessed 1 November 2024]

[9] See https://www.opensafely.org [accessed 1 November 2024]

A sea change in social media research

A key question for the Digital Mental Health Group – led by Dr Amy Orben[10] at the Medical Research Council (MRC) Cognition and Brain Sciences Unit[11] at the University of Cambridge – is how growing up in a time of rapid digitalisation affects young people's mental health and psychological well-being.

Until now, most studies in this field have used self-reported measures of time spent on social media platforms rather than focusing on how the platforms are actually being used. This approach has been an inexact science – in part due to the limitations of people's own recollections, and also due to the lack of detailed insights into what people are interacting with on social media.

To fill this gap, Dr Orben and her team have been exploring new ways of gathering data to help shed light on pressing policy questions about young people and their mental well-being – drawing on innovative and rigorous statistical methodology, secondary datasets, and Open Science approaches.

The team has been piloting an ecological momentary assessment (EMA) study[12] using data donations from social media archives known as Data Download Packages (DDPs) to better understand user interactions with social media content. This new approach has become possible since the General Data Protection Regulation (GDPR) was introduced across Europe and the UK in 2018, legally mandating all platforms that store their users' data – including social media platforms – to share these data with their users upon request.[13] If successful, the hope is that this approach could form the basis of large-scale cohort studies.

Dr Orben and her team recruited around 500 young people aged 13 to 18 from across schools around the East of England to take part in a short study that involved downloading and sharing their personal data (or DDPs) in a highly anonymised form from two of the most popular social media platforms among young people – TikTok and Instagram. The goal was to track in real time any connections between the content young people were interacting with on social media and their mood and well-being.

Dr Amanda Ferguson[14] from the Digital Mental Health Group explained how the young people were recruited during lunch-time seminars at their schools. They received payments of up to £65 for taking part. The team had so many volunteers that they had to request more funding from the MRC Cognition and Brain Sciences Unit to cover the additional data processing costs and participant payments. Eventually, 330 young people of the original 500 recruits submitted their data.

Taking part in the study involved responding to three 5-minute questionnaires a day – before school, after school and in the evening – for two weeks regarding their social media use and how it affected their mood and emotions, whether they were feeling happy, sad or lonely in the moment, for example. The young people would then download their data onto a web portal

run by the Dutch company Eyra as a series of JavaScript Object Notation (JSON) files[15] without the data ever leaving their hard drives. Understandably, data processing protocols were incredibly tight. All of the data was anonymous and could not be linked in any way to an individual's actual social media usage. This was crucial for protecting the young people's privacy and adhering to the strict ethical framework for the study, which applies to all of the group's research, according to Dr Ferguson.

Once the JSON files were analysed by the team's data scientists, using Large Language Models, they were able to provide granular data about how young people's social media usage was linked to their mood in real time over the two-week period. Although the data is still being analysed, the pilot study has been so successful that the Digital Mental Health Group is now planning to scale up to a much larger cohort involving thousands of young people.

"It's a type of data that we haven't had before," says Dr Ferguson. "There's been a lot of speculating about how the content people are viewing is impacting their well-being, without having any really good data. I think it's a real sea change in social media research. It has massive value."

Nothing has been published about the pilot project yet, but eventually the ambition would be to make the findings as open as possible to help answer pressing questions around young people's social media use. There is a shared project in the pipeline with colleagues at Stanford University drawing on the data already gathered. The team is now planning to use similar cutting-edge social media data for projects such as Born in Bradford – an internationally recognised research programme tracking the lives of 40,000 Bradfordians.[16]

Breaking barriers

In spite of the project's success, there have been some significant barriers to overcome, including being at the whim of social media companies, says Dr Ferguson. Because there is currently no specific regulation over DDPs, both TikTok and Instagram would regularly change what was available on their data download packages in pretty significant ways, according to Dr Ferguson, with substantial knock-on effects for the data gathering process. Even with a large and well-resourced team, working for an intensive three to four months on the project, which was funded by a Medical Research Council grant, it was sometimes challenging to keep up with these changes that affected both the structure of the data download packages and the type of information included in the packages. This inevitably created delays in the data-gathering process. Relying on an external company to collect the data also added a layer of complexity, although the company themselves were very good partners, Dr Ferguson says.

"I think if there was a standardised data download package, then there's a middle space that could be filled by academics or by a company like Eyra that could make it even more accessible to mental health researchers and then more nuanced questions can be asked," she adds.

Dr Ferguson believes that having access to more data is always going to be useful for researchers, especially large, openly available secondary datasets that can be revisited as statistical methods evolve: "Our group has gone back to reanalyse datasets a few times and had some pretty important findings from that," she says. She also points to large-scale studies like the National Health Services Mental Health of Children and Young People surveys as an example of what can be achieved.[17]

She is unsure whether the data sets provided by social media companies would be reliable enough to be included in a National Data Library, unless much more standardisation is introduced. However, she believes that a national library for data could be a powerful tool for advancing research on young people and mental health – as long as clear and consistent standards are established.

[10] See https://amyorben.com [accessed 1 November 2024]

[11] See https://www.mrc-cbu.cam.ac.uk [accessed 1 November 2024]

[12] "Ecological momentary assessments (EMAs) study people's thoughts and behaviour in their daily lives by repeatedly collecting data in an individual's normal environment, at or close to the time they carry out that behaviour." See https://www.gov.uk/guidance/ecological-momentary-assessmentn [accessed 1 November 2024]

[13] See https://www.gov.uk/data-protection/find-out-what-data-an-organisation-has-about-you [accessed 1 November 2024]

[14] See https://www.mrc-cbu.cam.ac.uk/people/amanda.ferguson/ [accessed 1 November 2024]

[15] A JavaScript Object Notation (JSON) file is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects.

[16] Read more: https://borninbradford.nhs.uk/

[17] See https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england [accessed 1 November 2024]

Conserving with code: How data is helping to save our planet

Over the last two decades, the University of Cambridge-based project Conservation Evidence has screened more than 1.5 million scientific papers on conservation, as well as manually summarising 8,600+ studies relating to conservation actions.[18] It has taken an estimated 75-person years for a team of highly skilled and highly trained researchers to manually curate the current database, which is used by a wide range of organisations to help inform decisions that are increasingly vital to the future of our planet and biodiversity conservation. Using the current manual system, only a few hundred new papers can be added to the open access database each year. AI has opened up new possibilities.

Planetary Computing Fellow Dr Sadiq Jaffer[19] and zoologist Dr Alec Christie,[20] Henslow Research Fellow at Downing College, are part of an interdisciplinary team of scientists from the University of Cambridge who are using AI to accelerate the synthesis. Building on the work of the Conservation Evidence team, the aim is to create a transformative tool for conservation research. "What we decided to do was to look at this as a way of building systems around the existing team to make them a lot more productive," explains Dr Jaffer. "So there's always a human in the middle of things who is accountable for the decisions that are made. But there are AI systems that can increase their productivity significantly."

To start with, Dr Jaffer and his colleagues compared Conservation Evidence's curated collection of papers with a broader sample of scientific literature. Using advanced language models, they clustered these papers by topic such as amphibians, reptiles, birds, and other species, confirming there was a clear structure in the data that could be leveraged for further analysis. Funded through a range of sources including ai@cam, The Hans Wilsdorf Foundation, and other philanthropic donations, this work allowed the team to break down research into smaller, more meaningful categories. With this groundwork in place, they developed an innovative three-step AI pipeline designed to streamline the identification of relevant conservation research.

In the first step of the project, smaller language models sifted through the OpenAlex dataset – an open data research platform that publishes metadata on around 250 million academic papers – to identify potentially useful papers. This analysis narrowed the sample of literature for analysis down to 6 million papers. From there, the team built a system that can download the abstracts and PDFs for each of the papers from relevant publishers and rank them according to their relevance.

After that, a set of 100,000 papers was filtered using a Large Language Model – similar to those powering ChatGPT – which applied the same criteria a human reviewer would use to assess each paper's relevance. This system is now being tested against human judgements, and early results indicate strong alignment between AI and human evaluations, showing promise for its future use in automating the review process, according to Dr Jaffer.

This is one of the first times data mining has been applied on such a large scale for conservation literature, as far as Dr Jaffer is aware, paving the way for more industrial-scale approach to gathering evidence. What makes the project particularly innovative is the fact that they are building a living system that can be expanded week on week with new papers. As researchers select more papers, the AI-driven system becomes more precise and finds more relevant evidence.

The team is now focusing on steps two and three of the process: extracting deeper insights from the full-text PDFs of relevant papers and integrating this information into Conservation Evidence's existing database. But the project has not been without its challenges.

Legal and technical headaches

"The biggest barrier was actually getting data from the publishers," says Dr Jaffer. "The OpenAlex data set is freely available, but it provides just the titles and metadata for papers and not the abstracts. Some publishers are quite free with their abstracts because they realise that having them out there in the wild means that there's more chance of people discovering their papers. Unfortunately, some of the biggest publishers are not."

In theory, copyright issues should not have been a problem because text and data mining for academic research (or non-commercial uses) is exempt from copyright rules set by the Copyright, Designs and Patents Act 1988. However, the team had a range of responses from publishers when they got in touch about downloading the data they needed, even though the data was technically Open Access and available under a Creative Commons License.

Some publishers were happy to give the go ahead and even pointed the team to the code for bulk downloading their papers on the coding platform GitHub. One example of best practice in this regard was the nonprofit, Open Access publisher PLOS. Other publishers insisted the team had to go through complicated Application Programme Interfaces (APIs), which threw up a range of technical and legal issues. Dr Jaffer and his colleagues had to write custom code to access each publisher's data. It took around three to four months of negotiating to reach agreement with some of the bigger publishers, with support from the University Library and JISC – the UK digital, data, and technology agency focused on tertiary education, research, and innovation.

"We've got AI models that can really accelerate these kinds of syntheses to provide evidence for policy-makers but if the publishers are going to be gatekeepers, it's going to slow down our progress in doing that," Dr Jaffer says. "And the sad thing is that the government has paid money via research grants to pay for this Open Access material to be made available and yet we can't actually go and download it."

If this hadn't been a non-commercial project with the support and resources of the University of Cambridge, Dr Jaffer says, there would have been even more barriers to contend with.

"If you had a startup that was trying to do this, they would have to open their chequebooks to the publishers in order to have any chance of pulling this off," adds Dr Jaffer. "And that's a barrier for these kind of innovative startups that want to do this outside of academia."

Creating a coherent archive

What would be helpful, according to Dr Jaffer, is more standardisation between publishers for Open Access material under permissive licences. At the moment, OpenAlex links to many disconnected Open Access repositories. However, having a coherent archive for Open Access materials that are licensed in such a way that they can be used for data mining without any technical hurdles would be the ideal scenario for this kind of research, as well as for a National Data Library, Dr Jaffer suggests.

"Obviously it can't be done for everything because there are things that are licensed by the publishers that are not Open Access and that's fair enough," he adds. "But for the things that are Open Access, there shouldn't be technical barriers in the way of doing it."

The ultimate goal for Dr Jaffer, his colleague Dr Christie and others is to ensure that evidence-based research informs conservation practice and policy, leading to more efficient use of public funds and better conservation outcomes. All the source code from the current project will eventually be made available for other scientists who want to create a similar living evidence synthesis pipeline. There is potential to roll this approach out to other disciplines too, including education and public health.

"We are providing the framework and the platform for more effective policy-making and decision-making," Dr Christie says. "But more work is needed to ensure decision-makers use it – and use it wisely."

[18] See https://www.conservationevidence.com/content/page/24 [accessed 1 November 2024]

[19] See https://www.cst.cam.ac.uk/people/sj514 [accessed 1 November 2024]

[20] See https://www.zoo.cam.ac.uk/directory/alec-christie [accessed 1 November 2024]

Breaking the ice: Addressing data barriers in Polar research

Dr Scott Hosking is leading the charge when it comes to using AI to tackle some of the biggest environmental challenges of our time.[21] As head of British Antarctic Survey's AI Lab and co-director of the Turing Research and Innovation Cluster in Digital Twins at The Alan Turing Institute, the climate-scientist-turned-data-expert leads a multi-disciplinary team of scientists and engineers developing cutting-edge tools to help predict how our planet is changing.

A key area of Dr Hosking's work is building 'digital twins' – highly detailed simulations of real-world environments that use AI and machine learning to help create digital versions of the physical world that can test future scenarios and forecast changes. Drawing on data from satellites, drones, radars, ocean floats, aircraft, robots, underwater vehicles, and a range of other sources, his team is currently working on a digital twin of the Polar regions.

Dr Hosking is also leading a pioneering collaboration to address the impact of climate change in the Arctic region using an AI-based sea ice forecasting system. IceNet draws on satellite data and weather observations to predict sea ice concentrations across the Arctic with remarkable precision, providing daily forecasts up to six months ahead.[22] It's proving to be a vital tool, helping indigenous Arctic communities and global policy organisations prepare for rapidly changing conditions, as well as helping to track and protect endangered wildlife including polar bears and Arctic foxes, and to reduce carbon emissions for shipping operations.

Working with WWF and Canadian partners, Dr Hosking's team has developed forecasts that have been used to help government researchers plan large-scale polar bear surveys, and to develop novel migration early-warning systems for endangered caribou. By blending decades of real-world observations and climate simulations, IceNet is already outperforming traditional physics-based numeral models in potentially game-changing ways, according to Dr Hosking.

"AI is different in that you show it so much information that it starts to forecast forward, based on data rather than physical understanding," he says. "For a long time, physicists were sceptical that a data-driven approach would be better because you're removing physics. I'm a physicist by training, but we've been surprised that AI can do a better job. It's taking the weather community by storm and suddenly old traditional weather models are just being outpaced now."

Drilling down into the data

The datasets that Dr Hosking and his team draw on for their work are all open access, ranging from in situ British Antarctic Survey sensors to satellite data from Copernicus – the EU space programme's observation component – as well as satellite systems managed by the European Space Agency and the European Centre for Medium-Range Weather. As a Natural Environment Research Council (NERC) organisation that's part of UK Research and Innovation, British Antarctic Survey follows strict data policies, frameworks and protocols which aim to preserve and manage all environmental data of long-term value. The data underpinning all Dr Hosking's research is published and shared via British Antarctic Survey's Polar Data Centre. The aim for all projects, Dr Hosking says, is to be as open and transparent as possible when it comes to sharing data that has been gathered from a myriad different sources.

"In a way, we are flooded with data," says Dr Hosking. "We can access almost all the data we need, but it's just in different places around the world in different formats, updated at different times, and often fragmented and imperfect."

Herein lies one of the biggest challenges facing Dr Hosking and his colleagues, who spend much of their time working out how to combine this fragmented data in a meaningful way. The challenge is exacerbated by technical hurdles, with data held in different formats and on different servers. This means that data scientists are spending more than half of their time accessing, downloading, and cleaning data before they can even start their research, Dr Hosking says.

"You'd think that, with the millions we spend on satellites, we would have good tools to access the data, but we just don't," says Dr Hosking. "We've got more data than we have the ability to process. We work with terabytes of data[23] at a time. We have access to petabytes[24]of data but we have no means to ingest all that data."

The biggest data barrier for cutting-edge environmental research like his, Dr Hosking says, is building the digital infrastructure and the scaffolding that connects everything together.

Removing barriers

To address these obstacles, Dr Hosking's team at British Antarctic Survey and The Alan Turing Institute includes Research Software Engineers (RSEs). These positions have been set up to build the underpinning digital tools and infrastructure to support open science. However, the resources for software development and data management are often squeezed into research budgets as an afterthought and are not allocated the time or budget needed, according to Dr Hosking.

Another related challenge is that scientists often work in silos, spending weeks developing software to download and clean data that is then discarded once a paper is published – leading to wasted resources and duplicated efforts.

"To do groundbreaking, real-world science, we cannot sit in our silos," says Dr Hosking, whose team shares open-source software code for their projects via the platform GitHub.

"We need to be far more joined up," he says. "For me, the biggest barrier is building sustainable research quality software and building in the legacy so that teams can work together and they don't have to reinvent the wheel again and again."

The overnment could help to address these barriers, Dr Hosking says, by ringfencing money in future research contracts to make sure data management is prioritised. Making funding available for Research Software Engineers to build the digital infrastructure that sits on top of the data and to provide the tools for scientists to do the research, for example, would be a huge step forward.

"For me, it's less about where the data sits because climate data is very open and available," says Dr Hosking. "It's all about the digital infrastructure and the software. If I walk into a normal library with books, everything's there in front of me ready to read. I want to do the same thing through my web interface or even through my computer terminal. I want to be able to log in and have access to everything through simple lines of code."

[21] See https://www.bas.ac.uk/profile/jask/#profilecontent [accessed 1 November 2024]

[22] Andersson, T., and Hosking, J. (2021). Forecasts, neural networks, and results from the paper: 'Seasonal Arctic sea ice forecasting with probabilistic deep learning' (Version 1.0) [Data set]. NERC EDS UK Polar Data Centre. https://doi.org/10.5285/71820e7d-c628-4e32-969f-464b7efb187c

[23] A terabyte equals one million, million bytes of data.

[24] A petabyte represents one quadrillion bytes, or 1,000 terabytes. It is a massive amount of data storage and is commonly used to measure the capacity of hard drives, data centres and cloud storage systems.

Making a difference with data: Insights from COVID-19

When the COVID-19 pandemic first emerged across the UK in early 2020, policy-makers in the East of England turned to Dennis Gillings Professor of Health Management Stefan Scholtes at the University of Cambridge Judge Business School for help.[25] Drawing on 25 years' experience at the University of Cambridge and with six years as Chair of the Board of the 60,000-patient Granta Medical Practices[26] in Cambridgeshire under his belt, Professor Scholtes convened a multi-disciplinary team of statisticians and public health experts who could help to analyse regional data to inform crucial policy decisions on a local level.[27]

With a health crisis emerging before their eyes, NHS colleagues in the East of England needed rapid advice to help plan their regional response. One of the key questions they needed to answer was how many COVID-19 admissions there might be in each of the hospitals across the East of England. That was the "exam question" that Professor Scholtes and his team took as a starting point.

They combined data from daily situational reports[28] produced by each of the hospitals across the region with national Intensive Care Unit (ICU) data that provided granular patient-by-patient information to build an innovative model that helped to predict the regional timing and size of the pandemic's peak. They also investigated how to use new data sources, such as mobility data based on mobile usage, to help predict the spread of the COVID-19 virus.

Throughout the first 12 months of the pandemic, Professor Scholtes and his team held weekly Zoom meetings with policy-makers from NHS East of England and Public Health England who were making key decisions to help plan everything from the number of ventilators needed to the projected number of hospital beds. Based on this crucial regional analysis, policy-makers decided not to open emergency Nightingale Hospitals in the East of England, which helped to save the NHS many millions of pounds.

"I think we did make a different decision because we did have different data," says Professor Scholtes. "It's all about making better informed judgements."

Although the data wasn't shared publicly, for confidentiality reasons, papers were published about the innovative model developed by the team.[29] The partnership between policy-makers and academia played a pivotal role in the region's pandemic response, earning Professor Scholtes and his team a Collaboration Award in the University of Cambridge's Vice Chancellor's Awards for Research Impact and Engagement in 2022.[30]

Overcoming barriers

Most of the data Professor Scholtes and his team relied on for their analysis was owned and managed by the NHS. Because it was a health emergency, access to the information they needed was a lot faster than usual, although Professor Scholtes is quick to point out that the data was still hosted securely on a University server, and standard protocols were followed at all times.

"The COVID project was an example of where pragmatism overruled barriers that are sometimes constraining," says Professor Scholtes. "Normally you have to go through lots of hoops to get to the data that you need. In this case there were shortcuts; they were not dangerous shortcuts because we were very aware of the sensitivity of the data, but they just made our lives a lot easier and a lot faster."

During non-pandemic times, getting access to healthcare data can be a slow and painstaking progress – taking anything from six to 18 months, according to Professor Scholtes.

"One of the challenges in the NHS is that the data is very siloed," he explains. Take the example of primary healthcare records held by the 6,000 primary care practices across England: "They have their own data in three different systems," he says. "And these three systems don't talk to each other, and they don't talk to the NHS systems."

To overcome these data challenges, Professor Scholtes believes that interoperability is vital, while ensuring that safeguards are in place to respect patient confidentiality and privacy. He suggests the planned NHS Federated Data Platform is a useful step in the right direction.[31] This is software currently being built that will enable NHS organisations to bring together operational data – currently stored in separate systems – to help staff access the information they need in one secure environment.

“If this works as planned, it will make a big difference,” says Professor Scholtes, and this kind of national data pool might also feed into a National Data Library in future, he suggests.

“[A data library] could bring incredible value,” he says, as long as it has the correct guardrails in place to protect patient confidentiality. “At the moment, everything is fragmented. Every researcher has their own way of accessing data and dealing with the red tape that’s necessary. A National Data Library would allow us to do things at scale. Standardisation and scalability would be the key for a data platform like this.”

Demonstrating value

Before building the library, Professor Scholtes suggests taking a step back and asking what questions need to be answered to create public value: “Think of this as a library with empty shelves and then filling the shelves one by one,” he says. “So you break it up, challenge by challenge, and think about how do I fill these shelves, driven by big policy questions. It’s got to be about creating value for society.”

For Professor Scholtes, the onus also falls on academics to demonstrate the value of their research more forcefully than they currently do.

“We tend to stop when our paper is published and that’s only half the story,” he says. “Often you have to carry on and continue your involvement with the policy development and policy implementation piece to close the loop and evaluate what you’ve done. You can tell a bigger story, but it takes five to six years before the story emerges.”

Here again he believes a National Data Library could play a key role.

“I think we as academics have more of a responsibility than just publishing papers,” Professor Scholtes says. “It’s not just about information and knowledge creation. It’s about making a difference.”

[25] See https://www.jbs.cam.ac.uk/people/stefan-scholtes/ [accessed 1 November 2024]

[26] See https://www.grantamedicalpractices.co.uk [accessed 1 November 2024]

[27] See https://www.jbs.cam.ac.uk/2020/coronavirus-research/ [accessed 1 November 2024]

[28] See https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/data-provision-notices-dpns/covid-19-situation-reports [accessed 1 November 2024]

[29] Betcheva, L., Erhun, F., Feylessoufi, A., Fryers, P., Gonçalves, P., Jiang, H., Kattuman, P., Pape, T., Pari, A., Scholtes, S. and Tyrrell, C. (2024) An Adaptive Research Approach to COVID-19 Forecasting for Regional Health Systems in England. INFORMS Journal on Applied Analytics 0(0). https://doi.org/10.1287/inte.2023.0009 [accessed 1 November 2024]

[30] See https://www.jbs.cam.ac.uk/2022/cambridge-judge-honourees-in-prestigious-university-of-cambridge-awards/ [accessed 1 November 2024]

[31] The NHS Federated Data Platform (FDP) is software that will enable NHS organisations to bring together operational data – currently stored in separate systems – to support staff to access the information they need in one safe and secure environment. This could be the number of beds in a hospital, the size of waiting lists for elective care services, or the availability of medical supplies.

Untangling the web of supply chain data

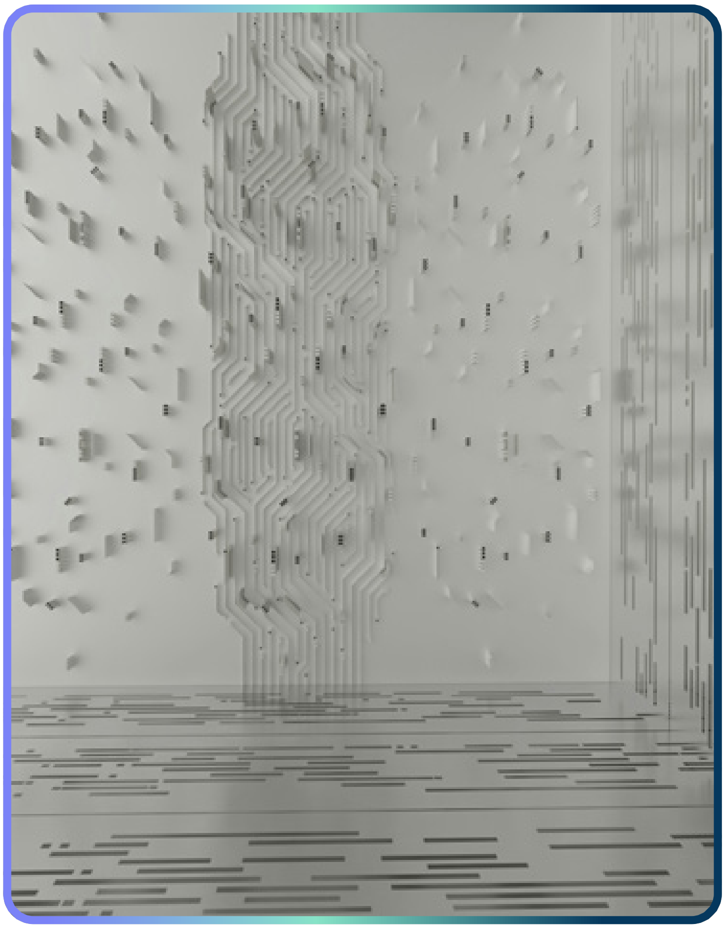
Imagine a vast interconnected web of 300 million firms, connected through an estimated 13 billion supply chain links. These are the businesses that produce the goods and services that make up our global economy.[32] In the UK alone, there are around 5.6 million registered businesses[33] with millions of suppliers. This web of complex interconnections has been almost impossible to untangle on a company-by-company level, until now.

According to University of Cambridge Professor of Macroeconomics Vasco M. Carvalho – who operates at the frontier of economics as Director of the Janeway Institute[34] – this blind spot has left our society ill-prepared to make rapid and well-informed decisions about how to respond to economic shocks. During the COVID-19 pandemic, for example, this led to shortages in critical medical supplies.

To address this gap, Professor Carvalho and his team have been working on a pioneering approach to help economists make sense of the web of data relating to businesses in the global economy by drilling down into financial transactions that can help to show company and industry-wide trends. In some countries, understanding firm-level interactions has been more straightforward because there are VAT records showing the value of goods and services traded between companies. Professor Carvalho refers to this as producing the "gold standard" of data. However, in countries like the UK and US, this gold standard doesn't exist. This means that Professor Carvalho and his team have had to take a different approach. Instead of VAT records from the tax authorities, he and colleagues from The Alan Turing Institute – where he is also a fellow – have been working together with the Office of National Statistics (ONS) to draw on data from the UK's electronic bank payment systems network to generate the information they need to map economic trends.[35]

Bank payment systems are the behind-the-scenes processes that allow money to be transferred between bank accounts in a series of steps. These digital systems now enable everyday financial activities like buying groceries, withdrawing cash, paying for a house deposit, receiving salaries or benefits, paying by direct debit, or transferring money via smartphone. Last year, these regulated processes handled over 22 billion transactions, amounting to around £75 trillion, according to the UK's Payment Services Regulator.[36]

One of the businesses that hosts these transactions on behalf of the banking industry is called Vocalink – a MasterCard company that processes all of the UK's real-time payments, settlement, and direct debit systems, as well as networks of over 47,000 ATMs. They also deal with over 90% of salaries, more than 70% of household bills, and 98% of state benefits in the UK.



A game changer for economists

Controlled by the banks, this huge ledger of financial transactions is a potential treasure trove of data for economists. However, until recently, it has not been available for anything other than fraud and security surveillance, due to privacy concerns. This began to change during the COVID-19 pandemic when economists like Professor Carvalho were asked to look at innovative ways to access supply chain data in real time to help avert shortages in vital supplies. They turned to the banking payment system for some quick answers and what they found was potentially game changing.

"As we move away from a cash economy, there are all these immense ledgers that we can use for extremely valuable public statistics in real time with plentiful data," explains Professor Carvalho. "Democratic society relies on public statistics. Banks have been generating these massive amounts of data as a byproduct of the financial system. And all this data can serve a broader goal of producing national statistics in a different way."

Mathematicians, data scientists, statisticians, and economists came together with the ONS to work out how they could use machine learning to transform these huge ledgers of anonymised data into something that could map financial transactions on an aggregated sector level so it could be used by researchers.[37] The findings on a sample of the data were surprisingly accurate: "It turns out that there is comparable information to the gold standard type data that does not exist in the UK," says Professor

Carvalho. “It surprised me the extent to which this data looks like the gold standard we were trying to replicate. This data simply does not exist otherwise.”

This new approach has potentially wide-ranging implications for economic research, helping to map where UK goods and services are imported from and exported to and showing how different regions across the UK do business with each other. In turn, this can help to inform regional industrial policies and major infrastructure decisions across the country.

Confidentiality concerns

Customer confidentiality and privacy has been crucial. Protocols for handling the data are incredibly strict: None of the raw bank-owned data ever leaves Vocalink’s systems. Only a limited number of ONS researchers have clearance to access the data and findings from the data are only shared publicly in an aggregated way – for example, showing industry trends rather than company-by-company data. Even with these assurances, the banks – the data owners – have had concerns about the process.

“The biggest barrier by far has been that the data owner does not see any benefit,” says Professor Carvalho. “And in fact, they only see downsides in anyone touching their data – even a relatively neutral body like the ONS. At the beginning of the project, there were hopes that there would be more access and more freedom of working with individual data. We could have moved much faster and with much fewer error rates if we were allowed to know more stuff.”

Professor Carvalho says some external observers have expressed a dislike for this type of research, regarding it as an example of Orwellian oversight gone too far. But his argument is that the data is being gathered anyway so why shouldn’t it be used for society’s benefit?

“There’s all this data and these ledgers that could revolutionise the way we as a society compile information about how we’re doing socially and economically,” he says. “And it’s our data anyway because we’re generating it. But it’s in the hands of the banking system. There needs to be a public conversation about what could prevent these barriers – and the effective use of these data for the global public good.”

Professor Carvalho’s research has already generated a lot of interest, both from the UK and overseas. The pilot ONS collaboration has been extended and there are plans to make the team’s research methodology public so it can be replicated in other countries and using other payment systems too. In the meantime, Professor Carvalho believes this type of research could help to revolutionise an antiquated national statistical system here in the UK that has limited resources to produce the vital public statistics that underpin our democracy.

He suggests a National Data Library in the UK could potentially act as a kind of ombudsman that could help to bridge the gap between sensitive data sources and individual researchers – providing a highly valuable resource for researchers while safeguarding data privacy.

[32] Pichler, A. Christian Diem, Alexandra Brintrup, François Lafond, Glenn Magerman, Gert Buiten, Thomas Y. Choi, Vasco M. Carvalho, J. Doyne Farmer and Stefan Thurner (2023) ‘Building an alliance to map global supply networks: New firm-level data can inform policy-making,’ Science, 19 Oct 2023 (Vol 382, Issue 6668) pp. 270-272, DOI:0.1126/science.adi7521

[33] Business Statistics, House of Commons Library (May 2024): https://commonslibrary.parliament.uk/research-briefings/sn06152/ [accessed 1 November 2024]

[34] See https://www.janeway.econ.cam.ac.uk [accessed 1 November 2024]

[35] Read more: https://www.ons.gov.uk/news/news/officefornationalstatisticsandthealanturinginstitutejoinforcestoproducebetterandfasterestimatesofchangestooureconomy

[36] See https://www.psr.org.uk/how-we-regulate/when-you-make-a-payment/ [accessed 1 November 2024]

[37] There are already publicly available tangible, aggregated, anonymous outputs of this experimental project work available via the ONS website: https://www.ons.gov.uk/economy/economicoutputandproductivity/output/articles/industrytoindustrypaymentflowsuk/2016to2023experimentaldataandinsights [accessed 1 November 2024]

Annexes

Contents

1. Further reading on case studies.

Cancer care

Mireia Crispin-Ortuzar, R. Woitek, M.A.V Reinius et al. (2023) Integrated radiogenomics models predict response to neoadjuvant chemotherapy in high grade serous ovarian cancer, Nat Commun 14, 6756, https://doi.org/10.1038/s41467-023-41820-7

Paverd, H., Zormpas-Petridis, K., Clayton, H., Burge, S. and Crispin-Ortuzar, M. (2024) Radiology and multi-scale data integration for precision oncology, npj | precision oncology Perspective, https://doi.org/10.1038/s41698-024-00656-0

Social media and mental health

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., and Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. Communication Methods and Measures, 16(4), 266–282. https://doi.org/10.1080/19312458.2022.2109608

Zendle, D., and Wardle, H. (2023). Debate: We need data infrastructure as well as data sharing – conflicts of interest in video game research. Child and Adolescent Mental Health, 28(1), 155–157. https://doi.org/10.1111/camh.12629

Conservation evidence

Ferris, P., Dales, M., Jaffer, S., Holcomb, A., Toye Scott, E., Swinfield, T., Eyres, A., Balmford, A., Coomes, D., Keshav, S. and Madhavapeddy, A. (2024) Planetary computing for data-driven environmental policy-making. Working Paper. arXiv, June 2024. https://doi.org/10.48550/arXiv.2303.04501

Madhavapeddy, A. (2024) Harnessing the power of AI to help save our planet. ai@cam blog post, available at: https://ai.cam.ac.uk/blog/harnessing-the-power-of-ai-to-help-save-our-planet

Climate modelling

Andersson, T.R, Hosking, J.S., Pérez-Ortiz, M. et al. (2021) Seasonal Arctic sea ice forecasting with probabilistic deep learning. Nat Commun 12, 5124, https://doi.org/10.1038/s41467-021-25257-4

Andersson, T.R. and Hosking, J.S. (2021) Forecasts, neural networks, and results from the paper: 'Seasonal Arctic sea ice forecasting with probabilistic deep learning' (Version 1.0) [Data set]. NERC EDS UK Polar Data Centre. https://doi.org/10.5285/71820e7d-c628-4e32-969f-464b7efb187c

COVID-19 forecasting

Betcheva, L., Erhun, F., Feylessoufi, A., Fryers, P., Gonçalves, P., Jiang, H., Kattuman, P., Pape, T., Pari, A., Scholtes, S. and Tyrrell, C. (2024) An Adaptive Research Approach to COVID-19 Forecasting for Regional Health Systems in England, INFORMS Journal on Applied Analytics, https://doi.org/10.1287/inte.2023.0009

Betcheva, L., Erhun, F., Feylessoufi, A., Fryers, P., Gonçalves, P., Jiang, H., Kattuman, P., Pape, T., Pari, A., Scholtes, S. and Tyrrell, C. (2020) Rapid COVID-19 Modeling Support for Regional Health Systems in England, SSRN Electronic Journal, https://doi.org/10.2139/ssrn.3695258

Economics

Pichler, A., Diem, C., Brintrup, A., Lafond, F., Magerman, G., Buiten, G., Choi, T.Y., Carvalho, V.M., Doyne Farmer, J. and Thurner, S. (2023) Building an alliance to map global supply networks: New firm-level data can inform policy-making, Science, 19 Oct 2023 (Vol 382, Issue 6668) pp. 270-272, DOI:0.1126/science.adi7521

Buda, G., Hansen, S., Rodrigo, T., Carvalho, V.M., Ortiz, A. and Rodríguez Mora, J.V. (2022) National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption, Cambridge Working Papers in Economics/Janeway Institute Working Paper Series, https://www.janeway.econ.cam.ac.uk/publication/jiwp2220

2. Case study contributors

Thank you to the researchers that helped develop these case studies:

Vicky Anning, science writer

Vasco Carvalho, Professor of Macroeconomics, Faculty of Economics, University of Cambridge.

Alec Christie, Imperial College Research Fellow, Centre for Environmental Policy, Faculty of Natural Sciences, Imperial College London, Visiting Researcher, Conservation Evidence, Department of Zoology, University of Cambridge

Mireia Crispin-Ortuzar, Assistant Professor and Group Leader, Department of Oncology, University of Cambridge, Co-Lead, Cancer Research UK Cambridge Centre - Ovarian Cancer Programme, Co-Lead, Cancer Research UK Cambridge Centre - Mark Foundation Institute for Integrated Cancer Medicine, and Chief Digital Officer, 52North Health

Amanda Ferguson, Senior Research Associate, MRC Cognition and Brain Sciences Unit, University of Cambridge.

Scott Hosking, Science Leader and Head of the AI Lab, British Antarctic Survey, and Co-Director of the Natural Environment, Turing Research and Innovation Cluster in Digital Twins (TRIC-DT), The Alan Turing Institute.

Sadiq Jaffer, Bernstein Planetary Computing Fellow, Head of Technology, Cambridge Centre for Carbon Credits (4C), Senior Research Associate, Department of Computer Science and Technology, University of Cambridge

Amy Orben, Programme Leader Track Scientist, MRC Cognition and Brain Sciences Unit, University of Cambridge

Stefan Scholtes, Dennis Gillings Professor of Health Management, Judge Business School, University of Cambridge

3. About ai@cam

ai@cam is the University of Cambridge's mission to develop AI that serves science, citizens, and society. It is an interdisciplinary AI incubator that is accelerating research to tackle real-world challenges with AI, informing the development of AI policy, and connecting across business and civil society to help translate AI innovations to practice. Its vision is of AI-enabled innovations that benefit society, created through interdisciplinary research that is deeply connected to real-world needs.
More information: ai.cam.ac.uk

About the Bennett Institute for Public Policy

The Bennett Institute for Public Policy is one of the UK's leading public policy institutes, achieving significant impact through its high-quality research. Our goal is to rethink public policy in an era of turbulence and inequality. Our research connects the world-leading work in technology and science at the University of Cambridge with the economic and political dimensions of policymaking. We are committed to outstanding teaching, policy engagement, and to devising sustainable and long-lasting solutions.
More information: www.bennettinstitute.cam.ac.uk

About the Minderoo Centre for Technology and Democracy

The Minderoo Centre for Technology and Democracy is an independent team of academic researchers at the University of Cambridge, who are radically rethinking the power relationships between digital technologies, society and our planet.
More information: www.mctd.ac.uk

ai@cam

ai.cam.ac.uk

For questions about this report, get in touch via the information below:

Email:
contact@ai.cam.ac.uk

Visit:
ai.cam.ac.uk

Connect on X:
@ai_cam_mission

LinkedIn:
@ai-cambridge

In partnership with:

MINDEROO CENTRE FOR TECHNOLOGY & DEMOCRACY

Bennett Institute for Public Policy Cambridge