



State of the art facilities:  
Data Observatory  
Private Cloud Infrastructure

Welcome to the  
**Data Science Institute**

Using data science  
to create a better world

# DSI ANNUAL REVIEW

# 20 22

December 2022

Imperial College  
London



# FOREWORD FROM THE CO-DIRECTORS

**DSI Co-Directors:**  
Professor Yi-Ke Guo  
Dr Mark Kennedy



Emerging from unprecedented times, 2022 has been a time for everyone to find their way to a new normal, and so it has been at the DSI. Besides of course, continuing paramount research work on major grants and research labs, it has been great to welcome friends old and new to the Data Observatory, to re-launch in person events, kick-off activities with our partner institute at the London School of Economics, and win approvals to start an exciting new education initiative.

2022 was an exciting year for data science. Alongside the continued growth in data science jobs in research, business, and government, the power of modern data sets is becoming even more evident as generative AI tools made the leap from labs to use by wider audiences. As the science of data and its use in discovery, invention, and practical work, data science continues to take shape as a foundational helping discipline supplying the powerful new kinds of data sets that enable breakthroughs in machine learning applications in many different fields.

The safeguarding of large data sets represents a key challenge moving forward and last year the DSI saw the launch of our own Private AI Cloud infrastructure that has continued to help us support data management across the College and the outside world. This facility supports the Data Learning Group as they develop methods for generating novel data assets, and the Computational Privacy Group in their efforts to understand the data behind AI models and how we can extract information about data sets not only from fully anonymised data, but also by systematic use of models trained by data—even without access to anonymised data. Because of high stakes and stringent data protection requirements, life sciences applications continue to be a proving ground for new methods and tools. Supporting this and other areas, data visualisation also continues to be vital to the discovery side of data science.

In the coming year, we will stay focused on laying a strong academic foundation for work in data science that enables both new scientific inquiries and, more generally, the betterment of society. The work of the DSI has become ever crucial in ensuring that we properly analyse, visualise, sort and disseminate the most pressing data to those who need it. Looking ahead, we are excited to build on our educational offerings from the successful summer and winter schools, bringing data science tools to a wider demographic."

Keep up to date with the DSI by following us on LinkedIn, Twitter and Instagram: @ImperialDSI

# CONTENTS

## The Year in Numbers 4

### Institute News

DSI Squared: Imperial  
launches collaboration with  
LSE 6

Portugal President visits DSI 7

Senior Healthcare Leaders visit  
the DSI 7

DSI researchers present work  
across the globe 8

Summer schools a great  
success 9

DSI researchers win  
outstanding paper award 10

Imperial students win inter-  
university hackathon 11

DSI welcomes new partnership  
with HyperionDev 11

### Research Stories

12 Proposed mechanisms for  
detecting illegal content can be  
easily evaded

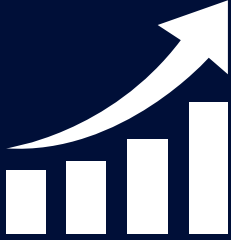
14 Data synthesis tool can  
rapidly help sort complex data  
sets

16 New AI model can prevent  
damaging and costly data  
breaches

18 Machine learning model  
uses social media data to more  
accurately monitor wildfires

20 Selected Publications 2022

# THE YEAR IN NUMBERS



## Growth

12

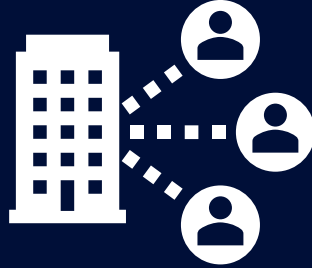
new members  
joined

2

new  
partnerships  
launched

10+

VIP visits to the  
DSI



## Outreach

8

seminars and  
conferences  
organised

113

summer school  
students

456

TB of research  
data managed



## Research

50+

papers  
published

£5M+

in research  
funding

17

active research  
projects









## DSI Squared: Imperial launches collaboration with sister institute at the London School of Economics

**2022 saw the launch of the DSI Squared collaboration between Imperial's Data Science Institute and its sister institute with the same name at the London School of Economics (LSE).**

When it comes to data science research and its impact, LSE's strengths in the social sciences naturally complements Imperial's strengths in science, technology and medicine.

By working together, the two Institutes hope this initiative will enhance their joint influence on policy in wide-scope domains.

### 'Speed-dating' networking events

In the summer, Imperial's DSI hosted the first networking event as part of the collaboration, held at Imperial's White City campus.

The research networking event took the form of 'speed-dating' where researchers from each DSI had the chance to meet each other, introduce their research interests and identify any opportunities for collaboration.

The event was attended by around 20 researchers from both institutes and participants had short, paired discussions, followed by an informal drinks reception.

### Unsolved research problems seminar series

In September, LSE's Data Science Institute took their turn to host the first seminar for a new series which aims to foster innovations by bridging the gap between

the social sciences, computer sciences and STEM subjects through presenting unsolved problems and crowdsourcing solutions from experts across these fields.

In this first seminar Dr Rossella Arcucci presented her unsolved problem relating to data science and machine learning for climate models.

Later in November, Imperial hosted the second seminar in this series, with Dr Ken Benoit presenting his unsolved problems relating to scaling text using the class affinity model.

The collaboration is hoping to launch a joint grant in the new year.

## Portugal President Marcelo Rebelo de Sousa visits the DSI's Data Observatory



**T**he President of Portugal paid a visit to the Data Science Institute's Data Observatory as part of his trip to celebrate Day of Portugal in June earlier this year.

The President met the College's Portuguese students and staff and heard about their innovative ideas and ongoing projects.

Joining him on the visit was João Gomes Cravinho, the Foreign Secretary, Paulo Cafôfo, Secretary of State for the Portuguese Communities, Nuno Brito, the Ambassador of Portugal to the United Kingdom, and Nadhim Zahawi, the UK's Secretary of State for Education.

As part of his visit, the President took a trip to the Data Science Institute where he saw presentations in the Data Observatory from Professor Sanjeev Gupta on Mars Rover data, and social media and pandemic correlations from

Dr Ovidiu Serban and Andrianirina Rakotoharisoa.

The institute's Co-Director Dr Mark Kennedy welcomed the group and gave an overview of the DSI.

## Senior healthcare leaders from 14 countries visit the Data Observatory

**I**nternational hospital CEOs attended the DSI as part of the Innovative Hospital programme to learn more about the use of big data in healthcare.

Organised by Imperial's Centre for Continuing Professional Development, the Innovative Hospital Programme brings together healthcare leaders including clinicians and hospital CEOs from across the globe to better understand the role that



big data serves in improving the quality of hospital care, along with addressing current economic and policy issues facing healthcare systems.

On 24 August 2022, as part of the 4-day programme, over 20 senior healthcare leaders visited the Data Science Institute,

where Research Fellow Dr Ovidiu Serban introduced them to different projects demonstrating the power and benefits of data science to help improve healthcare and inform better decisions.



## DSI researchers present their work across the globe



**A**cross the year DSI experts have presented their research at a number of important conferences across the globe.

In August, members of the Computational Privacy Group presented their research on privacy and client-side scanning at the 31st USENIX Security Symposium in Boston.

The USENIX Security Symposium is an annual conference that brings together researchers, practitioners, system administrators, system programmers, and others interested in the latest advances in the security and privacy of computer systems and networks.

The symposium was split into many different thematic tracks, covering topics such as web security, software vulnerabilities, scanning and censorship, and smart homes.

The themes covered by Imperial involved the topics of local differential privacy and client-side scanning, researched by Computational Privacy leader Dr Yves-Alexandre de Montjoye, PhD students Andrea Gadotti, Ana-Maria Cretu, Forent Guepin, and Shubham Jain,

and researchers Dr Florimond Houssiau and Meenatchi Sundaram Muthu Selva Annamalai, some of which attended the event.



**O**n 7 November 2022, privacy and security experts from all over the globe gathered for the annual ACM Conference on Computer and Communications Security, held this year in Los Angeles.

The ACM Conference on Computer and Communications Security (CCS) is the flagship annual conference of the Special Interest Group on Security, Audit and Control (SIGSAC) of the Association for Computing Machinery (ACM), and brings together information security researchers, practitioners, developers and users to explore cutting-edge ideas and results. ver and represented the fourth workshop in this series under the ICCS conference.

Imperial's Ana-Maria Cretu and Dr Florimond Houssiau, both current or former PhD students of the Computational Privacy Group at Imperial and the DSI. Dr Houssiau is now a Research Associate at the Alan Turing Institute.



**S**cientists from Imperial hosted a workshop at the annual International Conference on Computer Science (ICCS) conference, delivering a series of talks about machine learning and data assimilation.

This year, the ICCS conference was hosted by Brunel University in London on 21-23 June 2022 and was split into 16 thematic tracks, each intended to provide a forum for the discussion of one or more specific topics in the field of computational science.

The thematic track and accompanying workshop organised by Imperial's Data Science Institute covered the popular topic of Machine Learning and Data Assimilation for Dynamical Systems (MLDADS). This was hosted by the DSI's Dr Rossella Arcucci, Professor Yi-Ke Guo, Dr Cesar Quilodran Casas, Dr Sibio Cheng and Jake Lever and represented the fourth workshop in this series under the ICCS conference.





## Summer schools a great success

**Over 100 students took part in two different summer schools run by Imperial's DSI, to learn more about data science.**

Over July and August 2022, the Data Science Institute in collaboration with Imperial's Centre for Continuing Professional Development hosted two summer schools for students to learn more about data science or to understand more about artificial intelligence and data science in relation to healthcare innovation.

Over 100 students took part and successfully received a verified digital certificate from Imperial College and prizes were awarded to the best team project.

**Data Science Summer School**  
In the first online summer school, 83 students studying IT, computing or any engineering degrees at well-recognised universities in China with an interest in data science took part.

Through 40 hours of live lectures, workshops, tutorials, project work and self study, the cohort were introduced to the concept of data science, and got to hear from industry experts on data science applications.

Some of the talks included an introduction into 'The world of artificial intelligence' by DSI Co-director Professor Yi-Ke Guo, 'Data Privacy & Ethics' by Senior Lecturer Dr Yves-Alexandre de Montjoye and 'An Introduction into Blockchain Technology by Operations Manager Dr Kai Sun.

The students also completed a computer vision project centred around medical imaging for modelling brain tumours, or a natural language processing project where they used a COVID-19 dataset to discover the relationship between different biomedical entities.

For the first time this year, students also undertook a peer review challenge in order to develop their critical thinking skills, science communication skills and to improve their knowledge of data science. Peer review is the independent assessment of scientific knowledge by experts in the field, which serves as a quality control mechanism.

### **Artificial Intelligence and Data Science for Healthcare Innovation Summer School**

Drawing on the success of the Data Science Summer School, the DSI launched a new summer school with a focus on the health industry.

In the healthcare industry, data

science and artificial intelligence play a pivotal role in bringing together innovation and patient care and they have the potential to transform how healthcare is delivered.

In this summer school, 31 students from multi-disciplinary backgrounds with an interest in data science and the healthcare industry took part.

During the two-week long summer school, students heard various talks from the DSI: Dr Jingqing Zhang spoke about AI-assisted electronic health records, Dr Ovidiu Serban gave a talk on presenting data using data visualisation tools and Dr Benny Lo from Imperial's Hamlyn Centre, presented an insight into cutting-edge innovations in medical robotics.

### **"Very useful and inspiring"**

Overall, the feedback from the summer schools were extremely positive. For example:

"This is a valuable experience. I've improved my English skills during the course. Cooperating with new teammates is also very challenging. Also, I gained more knowledge about data science and AI."

The DSI looks forward to hosting the next cohort of students in the summer of 2023.



## Data scientists win Outstanding Paper Award for research on knowledge graphs

**A** team from the DSI won an award for their research on knowledge graphs, presented at the NeurIPS ENLSP workshop in December 2022.

Researchers from the DSI including Weihang Zhang, Dr Ovidiu Serban, Jiahao Sun and Professor Yi-Ke Guo recently won an Outstanding Paper Award for their paper entitled ‘Collective Knowledge Graph Completion with Mutual Knowledge Distillation’.

**Knowledge graphs are collections of interlinked descriptions of concepts, entities, relationships and events which put data into context, and are used in a range of applications including data governance, fraud detection, personalised recommendations, chatbots, search tools and other intelligent systems.**

The award was presented in the Graph for Natural Language Processing track at the NeurIPS-2022 Efficient Natural Language and Speech Processing Workshop on 2 December 2022 in New Orleans.

The workshop focused on the fundamental challenges to make natural language and speech processing models more efficient in terms of data, model, training and inference. Natural language is a subfield of linguistics, computer science and artificial intelligence that analyses the interactions between computers and human language.

In their winning paper, the team proposed a novel method to address the resource imbalance problem between knowledge graphs of different languages.



**According to lead author Weihang Zhang:**

“The results on DBP5L dataset have shown that all multilingual knowledge graphs can benefit from the collective knowledge transfer in this method, resulting in the state of the art performance on multilingual knowledge graph competition task.”

Many congratulations to the team who will be publishing a longer version of the paper later next year.





**Three Imperial physics students have won an international hackathon organised by Refinitiv, with another qualifying for the final stage.**

On 1 December 2022, Imperial physics students Kennedy Au, Andrew Liang and John Ow won first place and £300 in an inter-university hackathon organised by Refinitiv, a subsidiary of the London Stock Exchange Group and global provider of financial market data and infrastructure. The local organisation was supported by Dr Ovidiu Şerban and Ms Gemma Ralton, and from the Data Science Society Students (ICDSS) and the Mathematics and Design School student societies.

As part of a push to empower students to learn finance and coding, Refinitiv rolled out a product called 'Codebook' to allow students from any discipline to access easy-to-use coding interfaces to learn to code and apply it to management, finance, economics, computer science or data science. In line with this initiative, they set up an international inter-university hackathon, where students were given access to Refinitiv's data to solve a data-related problem.

## Imperial students win Refinitiv international inter-university hackathon

The winning team, EqualsMCsquare, presented a project centered on a short-term forecasting method called Bayesian Structural Time Series which they used to forecast Bitcoin.

They said:

“We chose this project for two reasons: sharpening our own understanding of Bayesian statistics, as well as it being a unique approach that may give us an edge in the competition compared to other time series techniques.”

“After phase 1.5, we decided to pivot our idea after the feedback session and focus on creating an educational Jupyter Notebook that walks through our code, rather than unnecessarily overcomplicating the model. On the final day of the hackathon, we delivered a detailed presentation on a Bayesian vector autoregression model, and compared this to two standard machine learning models that we've also trained on the time series; namely Long short-term memory (LSTM) and XGBoost.”

Another Imperial team, FinTech Hackers, also qualified for the final stage and included Financial Technology students: Dimitry Tertychnyy and Patrik Kovac. Congratulations to all that took part!



*One to watch*

## DSI welcomes new partnership with HyperionDev

In the new year, the DSI will officially launch a series of online data science bootcamps in partnership with HyperionDev, a large tech education provider.


HyperionDev offers an industry-aligned alternative to traditional technology learning models with a focus on teaching Software Engineering, Data Science and Web Development. With a human-led approach and accessible learning structure, they have brought coding skills to thousands of students in more than 40 countries in partnership with world-renowned institutions. Founded in 2012 and financially backed by Facebook and Google, HyperionDev intends to continue expanding and diversifying the tech talent pools for jobs in the United Kingdom and indeed globally through their added commitment to quality job placements and post-training student support.

The partnership will take the form of three online coding bootcamps where candidates can learn key skills in software engineering, data science or web development in six months or less.

The partnership will be expecting the first cohort in February 2023.



# RESEARCH STORIES



## Proposed mechanisms to detect illegal content can be easily evaded

**Data scientists have shown that current mechanisms to detect illegal content proposed by governments, tech companies and researchers are not robust.**

A team of data scientists from Imperial's Data Science Institute (DSI) and the Department of Computing have demonstrated that current mechanisms of detecting illegal content, known as perceptual hashing, do not work sufficiently and could be easily bypassed by illegal attackers online who aim to evade detection.

In the study, published in 31st USENIX Security Symposium held in August in Boston, the team including the DSI's Shubham Jain, Ana-Maria Cretu, and Dr Yves-Alexandre de Montjoye showed that 99.9% of images were able to successfully bypass the system undetected whilst preserving the content of the image.

Through a large-scale evaluation of five commonly used algorithms for detecting illegal content, including the PDQ algorithm developed by Facebook, the team showed that modified images are able to avoid detection whilst still maintaining very visually similar images.



**Co-Author Shubham Jain said:**

“Our results shed serious doubts on the robustness of perceptual hashing and scanning mechanisms currently proposed by governments and researchers around the world.”

Robustness in computer science is the ability of a computer to cope with errors or unexpected input.

### **Detecting illegal content online**

Currently, messaging platforms such as WhatsApp use a process called end-to-end encryption which enables people to securely and privately communicate with one another. Governments and law enforcement agencies have however raised concerns that illegal content might now be shared undetected.

A mechanism of scanning data known as client-side scanning has been recently proposed by tech companies and governments as a solution to detect illegal

content in end-to-end encryption communications. Client-side scanning broadly refers to systems that scan message contents such as images, text and videos, for matches against a database of previously known illegal content, before the message is encrypted and sent to the intended recipient.

However, it has received a significant backlash from the privacy community, raising concerns that the system of client-side scanning could result in privacy breaches for surveillance purposes.

Moreover, the creation of the UK Online Safety Bill and the EU Digital Services Act package has suggested having mandatory illegal content detection on large messaging platforms, but this assumes that the existing perceptual hashing-based client-side scanning systems would work.

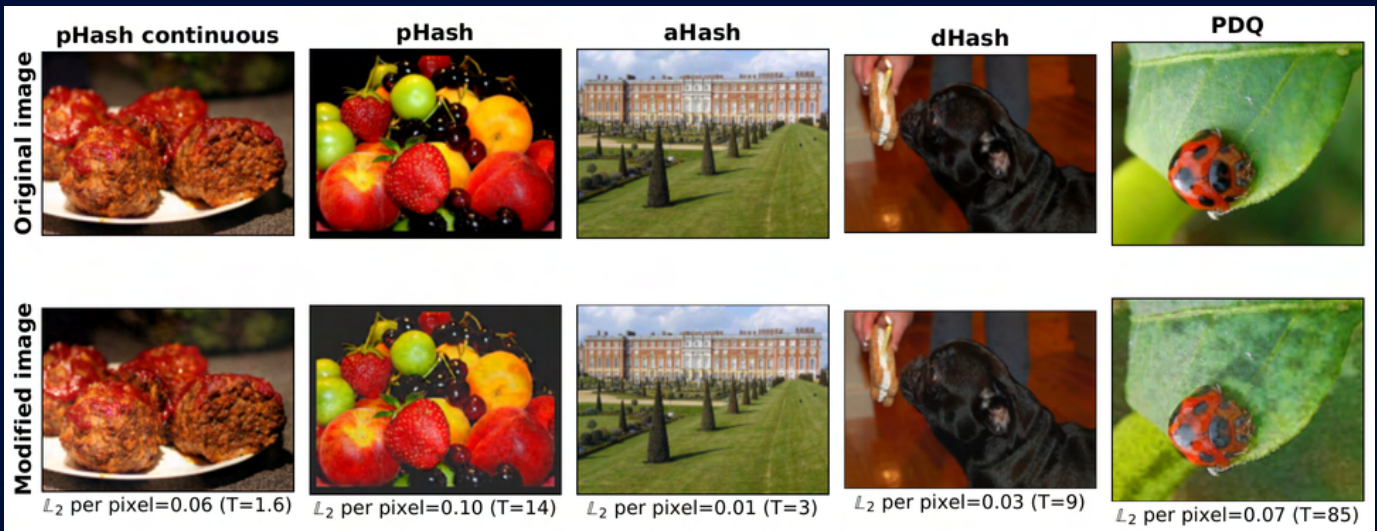
### **Digital fingerprints**

Perceptual hashing is a type of client-side scanning that uses a fingerprinting algorithm to produce a signature of an image using numbers. It is designed to generate signatures that remain robust to slight image modifications like rotation or rescaling.





The team demonstrated in their study that with slight adjustments to the digital fingerprint of the image, the modified image is highly visually similar to the original image, but was able to go undetected by illegal content detection algorithms.



This numerical signature of an image, or perceptual hash, can be compared to other signatures of existing illegal content, stored in a large database that is continually updated. It is however impossible to retrieve the image from the perceptual hash.

The detection system compares the two hashes and the image is flagged as illegal content if numbers are significantly similar and less than a predefined threshold.

Near-identical images demonstrate how methods of detecting illegal content are not robust. The team demonstrated in their study that with slight adjustments to the digital fingerprint of the image, the modified image is highly visually similar to the original image, but was able to go undetected by illegal content detection algorithms.

Jain said: “By design, perceptual fingerprints change slightly

with a small change in the image, something which, we showed, makes them intrinsically vulnerable to attacks. For this reason, we do not believe these algorithms to be ready and deployed for general use.”

### False positives

In the study, the team suggested that attempting to solve the problem by increasing the threshold would be an ineffective and impractical solution.

The larger the threshold, the more modified the image and therefore the less likely the illegal content would be shared. However, the team found that increasing the threshold would detect too many false images, with one in every three detections a false positive, creating over one billion images wrongly flagged everyday.

Co-Author Ana-Maria Cretu explains: “It is important to keep the number of false positives to a minimum as too many could

overwhelm the system and also there begins to be risks to individual privacy as each false positive has to be decrypted and shared with a third party for verification.”

The team also demonstrated that an attacker can create many different modified variations of one image whose fingerprints are very different from that of the original image and therefore adding image variations to the database of illegal content is unlikely to be effective.

### Moving forward

Currently, tech companies like Apple have proposed to use perceptual hashing algorithms based on deep learning called deep hashing. Recently, the team showed their attack to be very effective against deep hashing as well. They applied the same attack to the winning model of Facebook’s Image Similarity Challenge 2021, a deep hashing algorithm designed to detect image manipulations.

## Data synthesis tool can rapidly sort complex datasets to help decision-makers



“ This study represents a promising tool for use in the future of information retrieval research and data quality in medicine.”

Dr Ovidiu Serban

**A new data synthesis tool can help tackle ‘infodemics’ by quickly sorting through large, complex datasets to make optimal decisions for society.**

The DSI’s Dr Ovidiu Serban in collaboration with commercial partners AWS, MirrorWeb and CloudWick, have developed a new platform called Realtime Data Synthesis and Analysis (REDASA) to help curate and filter large amounts of complex information in a short amount of time. The platform represents a key tool that will help stakeholders make optimal decisions for society which will ultimately improve factors like public health and safety.

While the huge scale of the scientific response to the COVID-19 pandemic has unquestionably saved lives, the sheer volume and velocity of new information published each day has triggered an unprecedented ‘infodemic’ – an overabundance of information both online and offline.

By combining the knowledge of medical experts, with the efficiency of an artificial-intelligence-enabled engine, the team, developed a data extraction methodology to filter out documents representing only the most relevant and important information about COVID-19. This new method can be applied to other extensive datasets in the future.

The study, published in the Journal of Medical Interest Research, used REDASA to create one of the world’s largest and most up-to-date sources of COVID-19-related evidence,

consisting of over 104,000 documents.

### **An ‘infodemic’**

COVID-19 is the first pandemic in history in which technology and social media are being used on a massive scale to keep people safe, informed, productive, and connected.

However, at the same time, the technology we rely on has enabled an ‘infodemic’ that has undermined the global response and measures to control the pandemic.

The rapid publication of large amounts of data across both peer- and nonpeer-reviewed sources presents considerable challenges for stakeholders such as policy makers, clinicians, and patients to navigate.

These stakeholders must rapidly synthesise information to make optimal, evidence-based decisions for the benefit of society and for the protection of public health, and current methods of synthesising data are unable to keep up with the pace of the rapidly changing information landscape.

Therefore, there is an urgent need to capture, structure and interpret large and complex datasets in real time.

### **Human-in-the-loop methodology**

The REDASA’s design adopts a user-friendly, human-in-the-loop methodology by embedding an efficient, user-friendly curation platform into a natural language processing search engine. This means that human experts are involved in the decision-making process

along with AI components: the automated system filters down the information to a manageable amount, humans then check the results and assess the quality of the selected work and another AI component then verifies that the experts are consistent.

The platform has been designed for use across a wide range of data-rich subject areas while keeping application and impact in mind. It continuously captures and synthesises both academic literature and relevant ‘grey’ literature (including news websites, policy documentation and social media posts) to develop a data curation approach that could supplement machine-learning methodologies.

### **Next steps**

According to Dr Ovidiu Serban, “This study represents a promising tool for use in the future of information retrieval research and data quality in medicine. We are currently validating the same pipeline in cancer research, while also working with research groups looking at systematic reviews for biomarkers.”

Moving forward he said: “We are working with publishers to get more data and to ensure all data is fully accessible. We are also looking into using this tool for analysing social media to incorporate public opinion and discussions around various medical treatments. Ideally by showing the evolution of medical evidence over time, and being more open with existing medical evidence, we would be able to counteract future fake news phenomenon.”



The DSI's Dr Ovidiu Serban in collaboration with commercial partners AWS, MirrorWeb and CloudWick, have developed a new platform called Realtime Data Synthesis and Analysis (REDASA) to help curate and filter large amounts of complex information in a short amount of time.



Real-Time Medical Evidence Execution  
 High Quality Curated Data Late and  
 Querable, Real-Time Data Synthesis

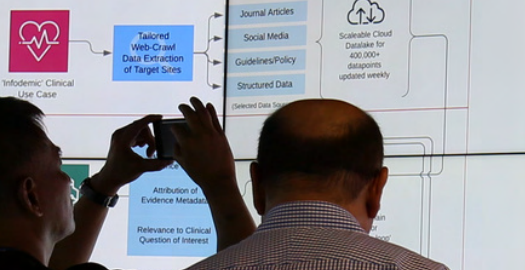
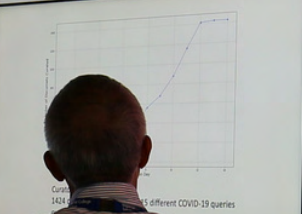


Table 1. COVID-19 related NLP queries used in the REDASA Search Index to enable a question-specific curatorial methodology:

1. What is the time interval between SARS-CoV-2 infection and using personal?
2. What is the time interval between infection and diagnosis COVID-19?
3. What is the time interval between diagnosis and recovery?
4. What is the impact of health care on COVID-19 recovery?
5. What is the risk of SARS-CoV-2 infection in health care professionals?
6. What is the impact of SARS-CoV-2 infection on health care professionals?
7. What is the impact of SARS-CoV-2 infection on health care workers?
8. What is the impact of SARS-CoV-2 infection on health care workers?
9. What is the impact of SARS-CoV-2 infection on health care workers?
10. What is the impact of SARS-CoV-2 infection on health care workers?
11. What is the impact of SARS-CoV-2 infection on health care workers?
12. What is the impact of SARS-CoV-2 infection on health care workers?
13. What is the impact of SARS-CoV-2 infection on health care workers?
14. What is the impact of SARS-CoV-2 infection on health care workers?
15. What is the impact of SARS-CoV-2 infection on health care workers?
16. What is the impact of SARS-CoV-2 infection on health care workers?
17. What is the impact of SARS-CoV-2 infection on health care workers?
18. What is the impact of SARS-CoV-2 infection on health care workers?
19. What is the impact of SARS-CoV-2 infection on health care workers?
20. What is the impact of SARS-CoV-2 infection on health care workers?



**PanSurg REDASA** | Research during COVID-19

**10% of all papers worldwide written on COVID-19**

**have been curated in just 2 weeks as we build our real-time systematic review tool**

MirrorWeb | AWS | Amorphic

PanSurg | PanSurg.org | @PanSurg | Imperial College London



## New AI model can help prevent damaging and costly data breaches

Imperial privacy experts have created an AI algorithm that automatically tests privacy-preserving systems for potential data leaks.



**This is the first time AI has been used to automatically discover vulnerabilities in this type of system, examples of which are used by Google Maps and Facebook.**

**T**he experts, from Imperial's Computational Privacy Group, looked at attacks on query-based systems (QBS) - controlled interfaces through which analysts can query data to extract useful aggregate information about the world. They then developed a new AI-enabled method called QuerySnout to detect attacks on QBS.

QBS give analysts access to collections of statistics gathered from individual-level data like location and demographics.

They are currently used in Google Maps to show live information on how busy an area is, or in Facebook's Audience Measurement feature to estimate audience size in a particular location or demographic to help with advertising promotions.

In their new study, published as part of the 29th ACM Conference on Computer and Communications Security, the team including the Data Science Institute's Ana Maria Cretu, Dr Florimond Houssiau, Dr Antoine

Cully and Dr Yves-Alexandre de Montjoye found that powerful and accurate attacks against QBS can easily be automatically detected at the pressing of a button.

**According to Senior Author Dr Yves-Alexandre de Montjoye:**

“Attacks have so far been usually developed using highly skilled expertise. This means it was taking a long time for vulnerabilities to be discovered, which leaves



systems at risk.”

“QuerySnout is already outperforming humans at discovering vulnerabilities in real-world systems.”

### The need for query-based systems

Our ability to collect and store data has exploded in the last decade. Although this data can help drive scientific advancements, most of it is personal and hence its use raises serious privacy concerns, protected by laws such as the EU’s General Data Protection Regulation.

Therefore, enabling data to be used for good while preserving our fundamental right to privacy is a timely and crucial question for data scientists and privacy experts.

QBS have the potential to enable privacy-preserving anonymous data analysis at scale. In QBS, curators keep control over the data and therefore can check and examine queries sent by analysts to ensure that the answers returned do not reveal private information about individuals.

However, illegal attackers can bypass such systems by designing queries to infer personal information about specific people by exploiting vulnerabilities or implementation bugs of the system.

### Testing the system

The risks of unknown strong “zero-day” attacks where attackers capitalise on vulnerabilities in systems have stalled the development and deployment of QBS.

To test the robustness of these systems, in a similar way to penetration testing in cyber-security, data breach attacks can be simulated to detect information leakages and identify potential vulnerabilities.

However, manually designing and implementing these attacks against complex QBS is a difficult and lengthy process.

Therefore, the researchers say, limiting the potential for strong unmitigated attacks is essential to enable QBS to be usefully and safely implemented whilst preserving individual rights to privacy.

### QuerySnout

The Imperial team developed a new AI-enabled method called QuerySnout which works by learning which questions to ask the system to gain answers. It then learns to combine the answers automatically to detect potential privacy vulnerabilities.

By using machine learning, the model can create an attack consisting of a collection of queries that combines the answers in order to reveal a particular piece of private

information. This process is fully automated and uses a technique called ‘evolutionary search’ which enables the QuerySnout model to discover the right sets of questions to ask.

This takes place in a ‘black-box setting’ which means the AI only needs access to the system but does not need to know how the system works in order to detect the vulnerabilities.



**Co-First Author Ana-Maria Cretu said:**

“We demonstrate that QuerySnout finds more powerful attacks than those currently known on real-world systems. This means our AI model is better than humans at finding these attacks.”

### Next steps

Presently, QuerySnout only tests a small number of functionalities. According to Dr de Montjoye: “The main challenge moving forward will be to scale the search to a much larger number of functionalities to make sure it discovers even the most advanced attacks.”

Despite this, the model can enable analysts to test the robustness of QBS against different types of attackers. The development of QuerySnout represents a key step forward in securing individual privacy in relation to query-based systems.



## Machine learning model uses social media to accurately monitor wildfires

**S**cientists have developed a new machine learning model that uses social media data to predict and monitor wildfires more accurately in real-time.

Data scientists from Imperial's Data Science Institute used machine learning - a subfield of artificial intelligence where computers learn from data and statistics, in a wildfire prediction.

model. In this new model, they combined social media data and geophysical satellite data to predict wildfire characteristics with high accuracy.

The study, published in the *Journal of Computational Social Science*, demonstrates how social media could be key to making more informed, socially driven decisions which could

help disaster management teams to identify areas of immediate danger.

The intensity of wildfires and wildfire season length is increasing due to climate change, causing greater threats to populations worldwide. To control and mitigate the negative effects of wildfires, computational models exist that attempt to



understand and predict the physical characteristics and evolution of these wildfire events.



**According to lead author, Jake Lever, who is a student of the Leverhulme Centre for Wildfires, Environment and Society:**

“Social media is becoming increasingly important as a source of information and this work tries to make sense of it within the context of wildfires. I hope this research will eventually make computational models more socially conscious, and therefore useful for disaster management teams.”

### **Real-time human sensors**

The proliferation of social media in recent years has created a large amount of publicly available, unprocessed social data that captures the opinions and emotions of people in real-time. The immediate publishing of information on these platforms means that in disaster situations users act as ‘human sensors’, detecting and documenting events as they happen.

Increasingly, this social media data is being used by scientists for investigating, modeling, and mitigating natural disasters. Its widespread use during such events has helped develop disaster management applications from a humanitarian perspective, which is now being employed by aid agencies across the globe.

At present, traditional methods of sensing and predicting wildfires typically rely on geophysical sensors such as satellite data or remote sensing. However, this fails to accurately capture current wildfire information as it relies on lengthy simulations, satellite data which is timely to access, and it does not consider other social factors like blocked roads, downed powerlines, or evacuation orders.

According to Lever: “Instead of having a network of cameras or climate sensors to track a wildfire, you can use a network of social media users or ‘human sensors’ posting information about a disaster in real-time.”

“It’s easier, cheaper, and quicker as you don’t have to go out and place a network of sensors and you can do it straight away from almost anywhere in the world.”

### **Sentimental Wildfires**

By combining Twitter data with historical satellite data from the Global Fire Atlas, the team developed ‘Sentimental Wildfires’ – a machine learning model, trained on both social and physics wildfires data using a Sentimental Analysis.

Sentimental Analysis is a way of analysing emotional or subjective content in text, and it can be useful to help evaluate the levels of destruction in local areas, identify people or communities displaced, and improve overall disaster management and mitigation.

Previous research indicates that often regions that have a lower social sentiment (or higher amount of emotional or subjective content), correlate to a perceived disaster severity in that area.

In the study, the team tested the model with two 2016 datasets from the US and Australia. Their findings suggest that social media is a predictor of wildfire activity, and the numerical methods and algorithms proposed in the study can be applied to other natural hazard events in the future.

### **Next steps**

The study represents a proof-of-concept for the socio-physical model for wildfire prediction, developed by Imperial’s Data Learning Group.

According to co-author Dr. Rossella Arucci: “The Data Learning Group at the DSI focuses on developing fundamental models for AI that are applied to different real-world applications, from air pollution to wildfires, economic models, social sciences and medicine.”

Moving forward, the team hopes to investigate how mis- and disinformation on social media can be accounted for by creating a more thorough verification process to extract the most meaningful and accurate information in natural disaster events.

In addition, combining data from multiple social media platforms will help to overcome any biases of using a single platform.

The model is also currently limited by local social media usage and internet connectivity. However, these problems will continue to decrease with social media usage continually rising and global internet connectivity on the horizon.

## Selected publications from DSI researchers 2022

- Abdel-Aziz, M. I., Vijverberg, S. J., Neerincx, A. H., Brinkman, P., Wagener, A. H., Riley, J. H., ... & Maitland-van der Zee, A. H. (2022). A multi-omics approach to delineate sputum microbiome-associated asthma inflammatory phenotypes. *European Respiratory Journal*, 59(1).
- Aliee, H., Massip, F., Qi, C., Stella de Biase, M., van Nijntzen, J., Kersten, E. T., ... & Faiz, A. (2022). Determinants of expression of SARS-CoV-2 entry-related genes in upper and lower airways. *Allergy*, 77(2), 690-694.
- Arcucci, R., Casas, C. Q., Joshi, A., Obeysekera, A., Mottet, L., Guo, Y. K., & Pain, C. (2022). Merging Real Images with Physics Simulations via Data Assimilation. In *European Conference on Parallel Processing* (pp. 255-266). Springer, Cham.
- Basaran, B. D., Qiao, M., Matthews, P. M., & Bai, W. (2022). Subject-specific lesion generation and pseudo-healthy synthesis for multiple sclerosis brain images. In *International Workshop on Simulation and Synthesis in Medical Imaging* (pp. 1-11). Springer, Cham.
- Basaran, B., Matthews, P. M., & Bai, W. (2022). New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation.
- Browne, P., Lima, A., Arcucci, R., & Quilodr an-Casas, C. (2022). Forecasting emissions through Kaya identity using Neural Ordinary Differential Equations. *arXiv preprint arXiv:2201.02433*.
- Buizza, C., Casas, C. Q., Nadler, P., Mack, J., Marrone, S., Titus, Z., ... & Arcucci, R. (2022). Data learning: Integrating data assimilation and machine learning. *Journal of Computational Science*, 58, 101525.
- Chagot, L., Quilodr an-Casas, C., Kalli, M., Kovalchuk, N. M., Simmons, M. J., Matar, O. K., ... & Angeli, P. (2022). Surfactant-laden droplet size prediction in a flow-focusing microchannel: a data-driven approach. *Lab on a Chip*, 22(20), 3848-3859.
- Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., & Rueckert, D. (2022). MaxStyle: Adversarial Style Composition for Robust Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 151-161). Springer, Cham.
- Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., ... & Rueckert, D. (2022). Enhancing MR image segmentation with realistic adversarial data augmentation. *Medical Image Analysis*, 82, 102597.
- Chen, J., Anastasiou, C., Cheng, S., Basha, N. M., Kahouadji, L., Arcucci, R., ... & Matar, O. K. (2022). Computational fluid dynamics simulations of phase separation in dispersed oil-water pipe flows. *Chemical Engineering Science*, 118310.
- Cheng, S., & Arcucci, R. (2022, April). Machine Learning based surrogate modelling and parameter identification for wildfire forecasting. In *ICLR, AI for Earth and Space Science*, 2022.
- Cheng, S., Jin, Y., Harrison, S. P., Quilodr an-Casas, C., Prentice, I. C., Guo, Y. K., & Arcucci, R. (2022). Parameter flexible wildfire prediction using machine learning techniques: Forward and inverse modelling. *Remote Sensing*, 14(13), 3228.
- Cheng, S., Prentice, I. C., Huang, Y., Jin, Y., Guo, Y. K., & Arcucci, R. (2022). Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting. *Journal of Computational Physics*, 111302.
- Cheng, S., Quilodr an-Casas, C., & Arcucci, R. (2022). Reduced order surrogate modelling and Latent Assimilation for dynamical systems. In *International Conference on Computational Science* (pp. 31-44). Springer, Cham.
- Cretu, A. M., Houssiau, F., Cully, A., & de Montjoye, Y. A. (2022, November). QuerySnout: Automating the Discovery of Attribute Inference Attacks against Query-Based Systems. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (pp. 623-637).
- Cretu, A.M., Monti, F., Marrone, S., Dong, X., Bronstein, M. and de Montjoye, Y.A., 2022. Interaction data are identifiable even across long periods of time. *Nature communications*, 13(1), pp.1-11.
- Dai, C., Wang, S., Mo, Y., Angelini, E., Guo, Y., & Bai, W. (2022). Suggestive annotation of brain MR images with gradient-guided sampling. *Medical Image Analysis*, 77, 102373.
- Dai, C., Wang, S., Mo, Y., Zhou, K., Angelini, E., Guo, Y., & Bai, W. (2020, October). Suggestive annotation of brain tumour images with gradient-guided sampling. In *International conference on medical image computing and computer-assisted intervention* (pp. 156-165). Springer, Cham.
- Davies, R. H., Augusto, J. B., Bhuva, A., Xue, H., Treibel, T. A., Ye, Y., ... & Moon, J. C. (2022). Precision measurement of cardiac structure and function in cardiovascular magnetic resonance using machine learning. *Journal of Cardiovascular Magnetic Resonance*, 24(1), 1-11.
- Dmitrewski, A., Molina-Solana, M., & Arcucci, R. (2022). CntrlDA: A building energy management control system with real-time adjustments. Application to indoor temperature. *Building and Environment*, 215, 108938.



- Francis, C. M., Futschik, M. E., Huang, J., Bai, W., Sargurupremraj, M., Teumer, A., ... & Matthews, P. M. (2022). Genome-wide associations of aortic distensibility suggest causality for aortic aneurysms and brain white matter hyperintensities. *Nature communications*, 13(1), 1-18.
- Gadotti, A., Houssiau, F., Annamalai, M. S. M. S., & de Montjoye, Y. A. (2022). Pool Inference Attacks on Local Differential Privacy: Quantifying the Privacy Guarantees of Apple's Count Mean Sketch in Practice. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 501-518).
- Gong, H., Cheng, S., Chen, Z., Li, Q., Quilodr an-Casas, C., Xiao, D., & Arcucci, R. (2022). An efficient digital twin based on machine learning SVD autoencoder and generalised latent assimilation for nuclear reactor physics. *Annals of Nuclear Energy*, 179, 109431.
- Houssiau, F., Schellekens, V., Chatalic, A., Annamraju, S. K., & de Montjoye, Y. A. (2022).  $M \hat{=} 2 \hat{=} M$ : A general method to perform various data analysis tasks from a differentially private sketch. *arXiv preprint arXiv:2211.14062*.
- Houssiau, F., Rocher, L. and de Montjoye, Y.A., (2022). On the difficulty of achieving Differential Privacy in practice: user-level guarantees in aggregate location data. *Nature communications*, 13(1), pp.1-3.
- Jain, S., Cretu, A. M., & de Montjoye, Y. A. (2021, June). Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.
- Kart, T., Fischer, M., Winzeck, S., Glocker, B., Bai, W., B ulow, R., ... & Gatidis, S. (2022). Automated imaging-based abdominal organ segmentation and quality control in 20,000 participants of the UK Biobank and German National Cohort Studies. *Scientific Reports*, 12(1), 1-11.
- Lever, J., & Arcucci, R. (2022). Sentimental wildfire: a social-physics machine learning model for wildfire nowcasting. *Journal of Computational Social Science*, 5(2), 1427-1465.
- Lever, J., & Arcucci, R. (2022). Towards Social Machine Learning for Natural Disasters. In *International Conference on Computational Science* (pp. 756-769). Springer, Cham.
- Lever, J., Arcucci, R., & Cai, J. (2022, June). Social Data Assimilation of Human Sensor Networks for Wildfires. In *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 455-462).
- Maynard, T., Baldassarre, L., de Montjoye, Y. A., McFall, L., &  skarsd ttir, M. (2022). AI: Coming of age?. *Annals of Actuarial Science*, 16(1), 1-5.
- Mehta, R., Filo, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., ... & Arbel, T. (2021). QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation--Analysis of Ranking Metrics and Benchmarking Results. *arXiv preprint arXiv:2112.10074*.
- Meng, Q., Bai, W., Liu, T., O'Regan, D. P., & Rueckert, D. (2022). Mesh-Based 3D Motion Tracking in Cardiac MRI Using Deep Learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 248-258). Springer, Cham.
- Meng, Q., Qin, C., Bai, W., Liu, T., De Marvao, A., O'Regan, D. P., & Rueckert, D. (2022). MulViMotion: Shape-aware 3D Myocardial Motion Tracking from Multi-View Cardiac MRI. *IEEE Transactions on Medical Imaging*.
- Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., & Rueckert, D. (2022). Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*.
- Ouyang, C., Wang, S., Chen, C., Li, Z., Bai, W., Kainz, B., & Rueckert, D. (2022). Improved post-hoc probability calibration for out-of-domain MRI segmentation. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (pp. 59-69). Springer, Cham.
- Qiao, M., Basaran, B. D., Qiu, H., Wang, S., Guo, Y., Wang, Y., ... & Bai, W. (2022). Generative modelling of the ageing heart with cross-sectional imaging and clinical data. *arXiv preprint arXiv:2208.13146*.
- Qin, C., Wang, S., Chen, C., Bai, W., & Rueckert, D. (2023). Generative myocardial motion tracking via latent space exploration with biomechanics-informed prior. *Medical Image Analysis*, 83, 102682.
- Schneider, R., Bonavita, M., Geer, A., Arcucci, R., Dueben, P., Vitolo, C., ... & Mathieu, P. P. (2022). ESA-ECMWF Report on recent progress and research directions in machine learning for Earth System observation and prediction. *npj Climate and Atmospheric Science*, 5(1), 1-5.
- Tanwar, A., Zhang, J., Ive, J., Gupta, V. and Guo, Y., 2022. Phenotyping in clinical text with unsupervised numerical reasoning for patient stratification. *Experimental Biology and Medicine*, p.15353702221118092..
- Thanaj, M., Mielke, J., McGurk, K. A., Bai, W., Savioli, N., de Marvao, A., ... & O'Regan, D. P. (2022). Genetic and environmental determinants of diastolic heart function. *Nature cardiovascular research*, 1(4), 361-371.

- Tournier, A.J. and De Montjoye, Y.A., 2022. Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data. *Science Advances*, 8(33), p.eabl6464.
- Wang, Y., Blackie, L., Miguel-Aliaga, I., & Bai, W. (2022). Memory-efficient Segmentation of High-resolution Volumetric MicroCT Images. arXiv preprint arXiv:2205.15941.
- Wong, A., Jain, S., Cretu, A.M. and de Montjoye, Y.A., 2022. Blogpost: Deep perceptual hashing is not robust to adversarial detection avoidance attacks.
- Yang, X., Wang, S., Xing, Y., Li, L., Xu, R. Y. D., Friston, K. J., & Guo, Y. (2022). Bayesian data assimilation for estimating instantaneous reproduction numbers during epidemics: Applications to COVID-19. *PLoS computational biology*, 18(2), e1009807.
- Zaydullin, R., Bharath, A. A., Grisan, E., Christensen-Jeffries, K., Bai, W., & Tang, M. X. (2022, October). Motion Correction Using Deep Learning Neural Networks-Effects of Data Representation. In 2022 IEEE International Ultrasonics Symposium (IUS) (pp. 1-3). IEEE.
- Zhang, C., Cheng, S., Kasoar, M., & Arcucci, R. (2022). Reduced order digital twin and latent data assimilation for global wildfire prediction. *EGUsphere*, 1-24.
- Zhang, D., Barbot, A., Seichepine, F., Lo, F. P. W., Bai, W., Yang, G. Z., & Lo, B. (2022). Micro-object pose estimation with sim-to-real transfer learning using small dataset. *Communications Physics*, 5(1), 1-11.
- Zhang, W., Serban, O., Sun, J. & Guo, Y. (2022). Collective Knowledge Graph Completion with Mutual Knowledge Distillation, In Proceeding of the NeurIPS 2022 Efficient Natural Language and Speech Processing (ENLSP-II) workshop.
- Zhuang, Y., Cheng, S., Kovalchuk, N., Simmons, M., Matar, O. K., Guo, Y. K., & Arcucci, R. (2022). Ensemble latent assimilation with deep learning surrogate model: application to drop interaction in a microfluidics device. *Lab on a Chip*, 22(17), 3187-3202.

## Data Science Institute

A Global Institute of Imperial College London

South Kensington Campus  
Imperial College London  
London SW7 2AZ, UK



Data Science Institute at Imperial College London



@imperialdsi



@imperialdsi