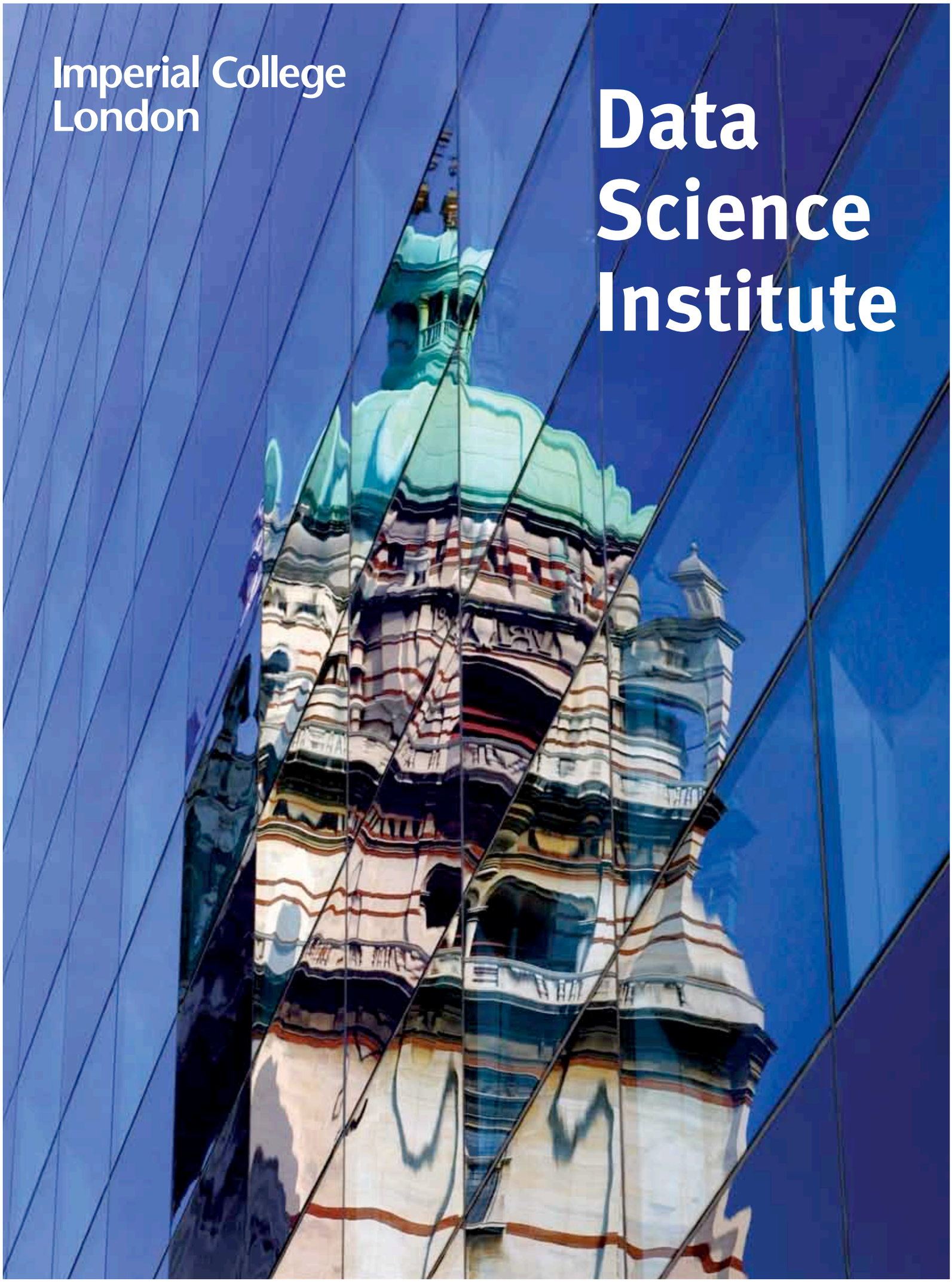
Imperial College London

Data Science Institute



Foreword

We live in a world where billions of gigabytes of data are generated every day about all aspects of our lives. New techniques and technologies continue to emerge which enable the generation of new insight, faster and with greater accuracy. Data is a tremendous asset in our search for solutions to the grand challenges in science, engineering, medicine and business but the existence of large and increasingly accessible data sets has created ethical debates around privacy and security.

The scale of the challenges and opportunities ahead in data science are clear, and this sets the context for the relevant and timely work of the Data Science Institute at Imperial College London.

The Institute is distinctive because of its multi-faceted and collaborative approach, which encompasses work across a broad spectrum of disciplines from the application of statistical methods to large and complex data sets, to data visualisation, and the economic aspects of data. The Institute continues to assemble a diverse and vibrant community of collaborators from across the globe. This approach will enable the Institute to deliver leading research and education across the whole data cycle, from data collection to communicating and acting upon insight generated.

I hope you share my enthusiasm for the plans set out in this booklet, and I would encourage you to engage with the Institute.

Professor Alice Gast
President of Imperial College London

Contents

Welcome to the Data Science Institute

Data science is the discipline that deals with collecting, preparing, managing, analysing, interpreting and visualising large and complex datasets. The discipline has its roots in the integration of statistics and computer science, where it is driving scientific and technological advancement in diverse areas such as, astrophysics, particle physics, biology, meteorology, medicine, finance, healthcare, and social sciences.

Modern science typically involves big data, taking advantage of high-throughput data capture and high-performance computing capabilities. Data science is therefore an essential element of all modern interdisciplinary scientific activities. It acts as the glue to facilitating collaborative scientific discovery and involving the whole life cycle of data, from acquisition and exploration to analysis and communication of the results. Data science is not only concerned with the tools and methods to obtain, manage and analyse data, it is also about extracting value from data and translating it from asset to insight.

The Data Science Institute has been established to conduct research on the foundations of data science and to foster the development of advanced theory, technology and systems that contribute to the state-of-the-art in data science and big data.

At the Data Science Institute, Imperial College London, our objectives are:

- To develop data management and analysis technologies and services for supporting data-driven research at Imperial College.
- To act as a focal point for coordinating data science research at Imperial College by facilitating access to funding, engaging with global partners, and stimulating cross-disciplinary collaboration.
- To promote the training and education of the new generation of data scientists by developing and coordinating new degree courses, and conducting public outreach programmes on data science and to advise Imperial College on data strategy and policy by providing world-class data science expertise.
- To enable the translation of data science innovation by close collaboration with industry and supporting commercialisation.
- To promote data science and its applications to Imperial education and general public.

In this booklet we highlight some of the research that we have begun to conduct at the Data Science Institute, and how data science is being applied across faculties at Imperial College, in various application areas.

I am sure you will find our work both informative and inspiring, and look forward to collaborating with you in the near future.

Professor Yike Guo
Founding Director of the Data Science Institute

Introduction to the Data Science Institute

The Data Science Institute (DSI) at Imperial College London opened in April 2014 and provides a hub for data-driven research and education across the College.

Its mission is to provide a focal point for Imperial College's capabilities in multidisciplinary data-driven research by coordinating advanced data science research for College scientists and partners, alongside educating the next generation of data scientists.

Modern scientific research is largely data driven. The DSI conducts research on core data science to develop advanced theory, technology, and systems that will contribute to the state-of-the-art in data science and support world-class research at Imperial and beyond.

The DSI acts as a focal point for expertise in data-driven research at Imperial, to help tackle grand challenges by encouraging the sharing of data and technologies for analysis and management. Data science aims to deliver tangible value from data assets. The DSI will empower Imperial and its industrial partners to collaborate in the pursuit of data-driven innovation.

The DSI is the hub for data science and engineering research across Imperial and organises a series of research networks in the form of virtual or physical research laboratories.

Our focus is to support College-wide cross-faculty collaborative research programmes addressing data-driven scientific grand challenges.

We are developing College-wide computational infrastructure for managing and processing scientific research data. This will enable world-leading data science research at Imperial and will be the world leader in driving data science platform development.

The Institute is participating in the establishment of College policies and strategy for building and strengthening its research data assets. We offer technology support for data stewardship, software platform development, training and project-specific collaborations. We provide a focal point for building a global alliance of academic and industrial partners to address major data science challenges and applications.

The Institute aims to generate significant intellectual property and, through strategic partnerships, to translate this into social and economic impacts as well as offering an advanced education programme to train a new generation of data scientists.

Research at the Data Science Institute

Progress depends on understanding. Understanding is built on evidence. And evidence comes from data. To be useful, data must be combined with knowledge extraction technologies to yield insights, understanding, and predictive power.

This is the aim of the DSI: to facilitate the extraction of knowledge and understanding from data. The DSI encourages the sharing of data and collaboration in research methods across Imperial College and with the wider research community as well as with industry.

Professor David Hand, OBE
Chairman of the DSI Research Board

Foundations of Data Science at the DSI

Advanced Data Analytics

Analytics is at the core of data science. Analytics research focuses on understanding data in terms of statistical models. It is concerned with the collection, analysis and interpretation of data, as well as the effective organisation, presentation and communication of results relying on data. As an essential part of data science, analytics research makes important advances in many scientific areas such as biology, medicine, economics, and finance. Machine learning is a subfield of computing and statistics that concerns the development of computational systems that can learn from data. In recent years, machine learning has become a central field in scientific and engineering research where mathematical models can be learned from big data. Modern machine learning algorithms, such as pattern discovery based on cognitive neural networks and Bayesian statistics, have made significant contributions to the big data revolution.

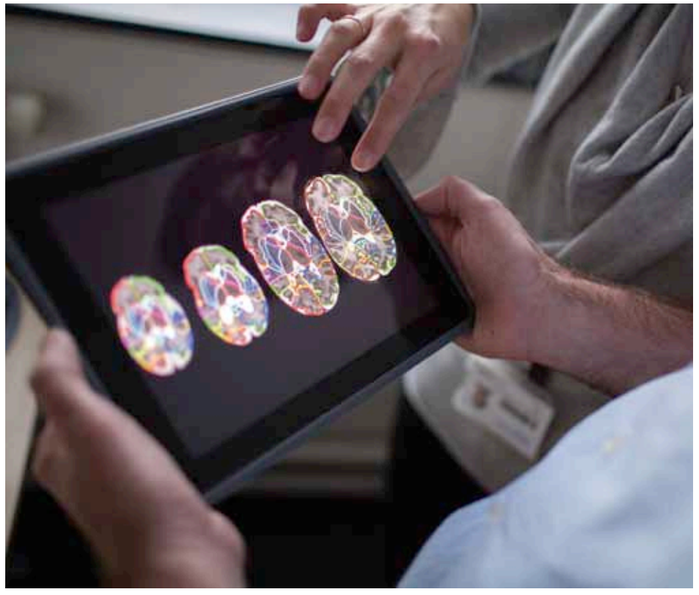
At the DSI, advanced analytics and its applications are the main research focuses. The DSI is building a College-wide data science research network across the disciplinary spectrum, led by pioneering researchers in statistics and computer science, to catalyse the broad research activities in advanced analytics at Imperial College, to enable the cross fertilisation of methods in different research, to develop new algorithms and build analytical applications that raise profound scientific and engineering challenges. Our work includes applications to: translational medicine, neuroscience, sensor informatics, urban informatics, human behavioural analysis, and social network analysis. In all these research areas we gather large amounts of data from heterogeneous sources such as, DNA sequencing, brain images, physiological sensors, environmental sensors, and social media. Analytical algorithms are developed to discover hidden patterns in the data and build predictive systems that generalise knowledge extracted from these datasets to previously unseen cases. We are particularly focusing on the real applications where learning processes take place in real-time and need to be adaptive to changing conditions. Examples include deep learning for neurological diseases (epilepsy, multiple sclerosis) and compressive-sensing for functional MRI (fMRI) medical imaging analysis.

Big Data Management

Underlying all data science methods, especially in the age of big data, is how to store, manage, and process data at the petabyte level and beyond.

Data management poses significant challenges with large, heterogeneous datasets that are commonly constantly changing. Data is generated on the fly through interaction with users of computing systems, from machines through their normal operation, and also through gathering data from various sensors and devices. In the past, data was typically harvested and stored in carefully designed structures however one of the characteristics of big data is the use of unstructured ‘dirty’ data. Traditional databases and data warehouses are not adequate for dealing with such data. To address the challenges of big data, new kinds of data management technologies and methods are required. Some examples of these are classed as ‘NoSQL’ database solutions that include in-memory databases, schema-less databases (e.g. key-value store, document-oriented), and graph databases. To deal with very large datasets, these database technologies support distributed data processing to spread data management over distributed computing resources.

At the DSI, our researchers are investigating how to best utilise new data management technologies, where we are experimenting with NoSQL solutions for biomedical data, such as large-scale molecular profiling and medical imaging data, as well as disparate and unstructured data taken from sensor networks. Our experimental software is developed on our own petabyte-scale test bed, enabling us to be at the forefront of data management research and application. Among our efforts is the tranSMART data warehouse, which is an international effort led by the DSI to build a standardised big data management and analytics platform for translational medicine research, that has been adopted by medical research organisations and pharmaceutical companies worldwide.



Visualising Data

Visualisation is a critical part of the data science for two reasons. Firstly, the human visual system excels at pattern recognition and we are best able to make sense of big data sources through the visualisation of data and subsequent analytics. Secondly, once insight has been gained from data visualisation, visual representations of data as well as analysis results are typically the most effective means of communicating results.

The unprecedented scale of modern big data will require matching large-scale visualisation environments. Such environments must be collaborative to enable interactive discussion of data and further exploration of the implications. In history, the collection of large scale data by radar and observation during the Battle of Britain led to the creation of huge aircraft plotting rooms. Similarly the modern increase in the scale of data must lead to large-scale visualisation capabilities that will enable us to understand as much of the 'big picture' of big data as possible.

Towards this goal the DSI is building a Global Data Observatory in partnership with KPMG. This will consist of a visualisation studio allowing teams of researchers and analysts to collaborate in the exploration of data and the derivation of actionable insight in real time with a cutting edge visualisation environment.

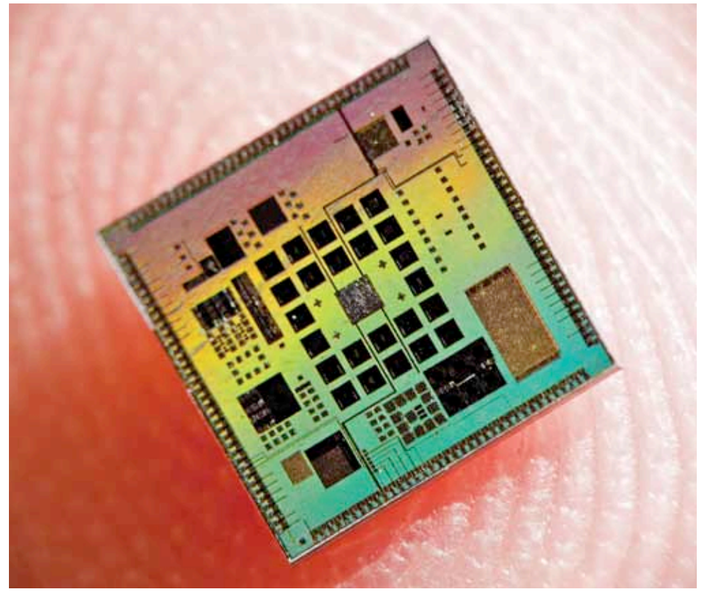
The Data Economy

The essence of the big data revolution comes from the realisation that data is a new form of natural resource. Markets are emerging in which datasets are sold and traded as commodities. Large public datasets are available through government open-source websites such as the US-based www.data.gov or UK-based www.data.gov.uk. In both the private and public sectors, data science is the catalyst for advances in the physical sciences including: astrophysics, particle physics, biology, climate science and meteorology; the human sciences including medicine and healthcare; and the social sciences such as economics, business, and finance, as well as sociology, and political science. The generation of data in each of these fields offers businesses, governments, and non-profit organisations unparalleled opportunities to divine new insights into human behaviour and social dynamics.

Often traditional economic and legal frameworks that are used to harness data ignore data's own intrinsic qualities. For example, data may be classified as an economic good from at least two perspectives using an economic theory of value. Data is a club good and, even more specifically, an information good which is to say that it is non-rival yet excludable. The standard definition of private goods as rival and excludable, or public goods as non-rival and non-excludable, is being challenged in the information age. Among other considerations, digital data by its own nature is not excludable. Rather, datasets are less scarce than other economic goods and so excludability constraints placed on access to data tend to be externally imposed.

This reflection gestures to key questions: How should data be treated as an economic good? How should an agent assign value to data? Not only commercial data, but even more importantly, personal data? And how might individual citizen-consumers learn to exercise their rights as owners of personal data that is of incredible potential to external parties (businesses and governments) when aggregated with population-level data?

At the Data Science Institute, we are carrying out research into the economic, legal and policy mechanisms required for the emerging Data Economy in the UK and worldwide. Working with our colleagues in economics and social science, we are pioneering in the key areas of data economy such as open data business model, digital money and digital service exchanges.



Security & Ethics of Data

Setting aside these practical issues in the data economy, consumer advocates understandably worry about treating data as an asset. Much of the anxiety associated with data protection, concerns not only how datasets are collected, maintained, disseminated, and destroyed (i.e. the security issues), but also how personal data is anonymised when used in the interest of public welfare, such as medical research. Technologies, such as anonymisation and pseudonymisation have been developed to address the ethics challenge. We understand that the extent to which these technologies are enforced is subject to the demands of the legal environment in which data exists—reflecting local, national, and international protection priorities. Data privacy, therefore, has both criminal and civil law liability, and it straddles both the public and private legal domains. If both expected and hitherto unimagined benefits of large-scale data analytics are to be enjoyed by society, laws and policy frameworks then governing data must continually consider how to assign value to data while respecting individual privacy.

At the DSI, we have built a strong research agenda in data security and data ethics research in collaboration with technical research groups (e.g. security research in Department of Computing), policy expertise (e.g. Institute for Security Science and Technology) and the legal community. This research becomes even more crucial in our biomedical applications.

Multidisciplinary Applications

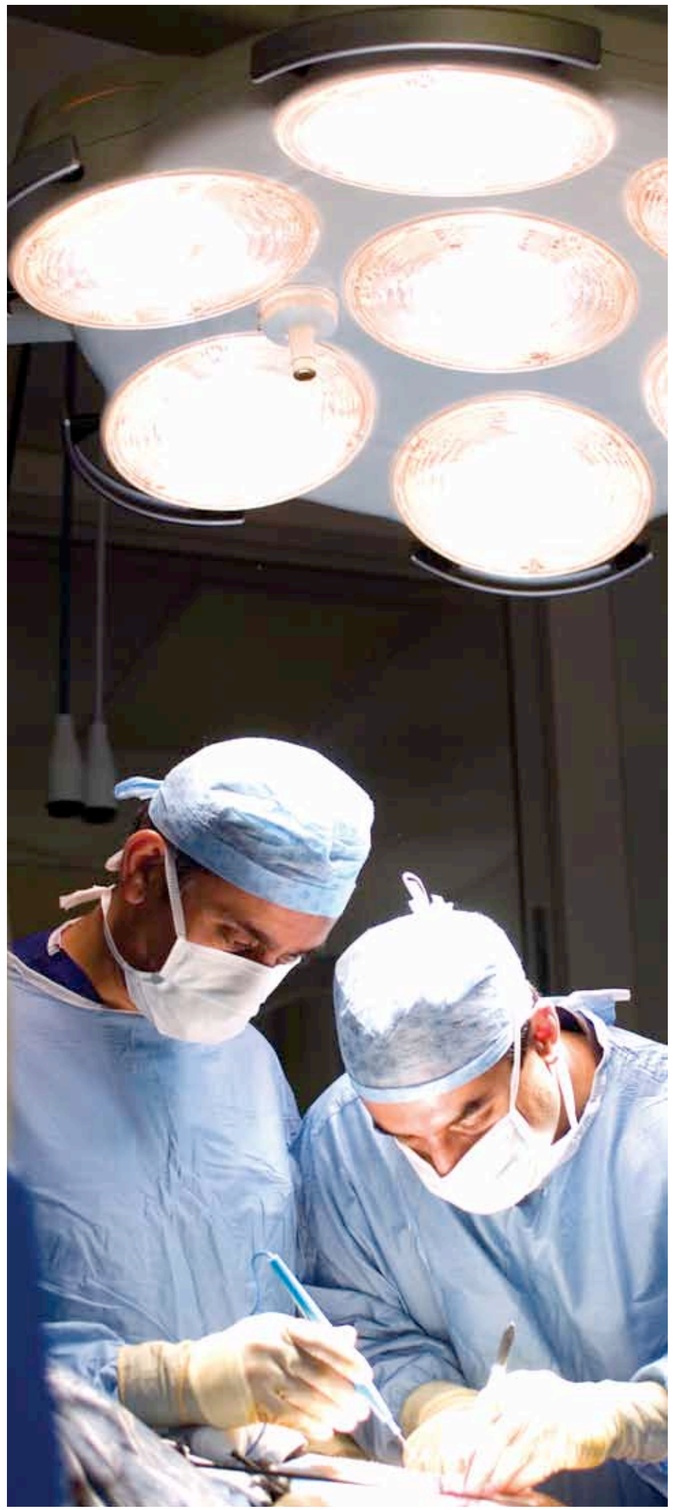
Imperial College's strengths lie in multidisciplinary research and innovation. The DSI will be all-encompassing, not only developing foundational research in data science methods, technologies, and policy, but also in applying data science in all areas of science, engineering, business and medicine.

Advancing Personal Medicine

Personalised medicine is about tailoring medical treatment to the individual characteristics of each patient, by classifying individuals into groups that differ in their susceptibility to a particular disease or their response to a specific treatment. With the amount of data being mined and analysed, it will be easier to identify genetic correlations, identify patterns in patient and population data, identify patient specific patterns and predict physiological conditions, discovering biomarkers that present signs of normal or abnormal processes and provide better patient self-management for enhanced clinical outcomes. The proliferation of data, generated from high-throughput molecular profiling to physiological sensing, offers great opportunities for personalised medicine by offering more precise diagnoses and more effective treatments, as researchers are able to drill down to see what is happening and create more targeted therapies, specifically at the molecular and tissue levels.

At the DSI we are working with Imperial College's biomedical researchers in the medical school and the worldwide medical research community to build big data technologies to enable personal medicine. This research includes building the European Translational Knowledge Management and Service (eTRIKS) platform as the gold standard personalised medicine big data research platform for the European medical research community. Besides building the capacity to manage clinical research data of millions patients, the platform also supports advanced analytics and visualisation that can combine clinical and genomic information with unstructured, non-clinical, or even analogous data to provide rich, actionable insights for healthcare decision makers. Such an example would be mapping lifestyle trends and genetic patterns within a specified age group, and their correlation to incidence of diabetes in a given population. The platform also supports the integration of multiple types of sources used in medical studies, such as genomics, radiology, biometrics, patient data from lab systems, HIEs (Health Information Exchanges), RIS (Radiology Information Systems), imaging from PACs (picture archiving and communication systems) and patient portals.

eTRIKS has been now used in a rich portfolio of medical research projects at Imperial College and across Europe and covers diverse disease areas, spanning inflammatory and autoimmune disorders to oncology and cardiovascular diseases.

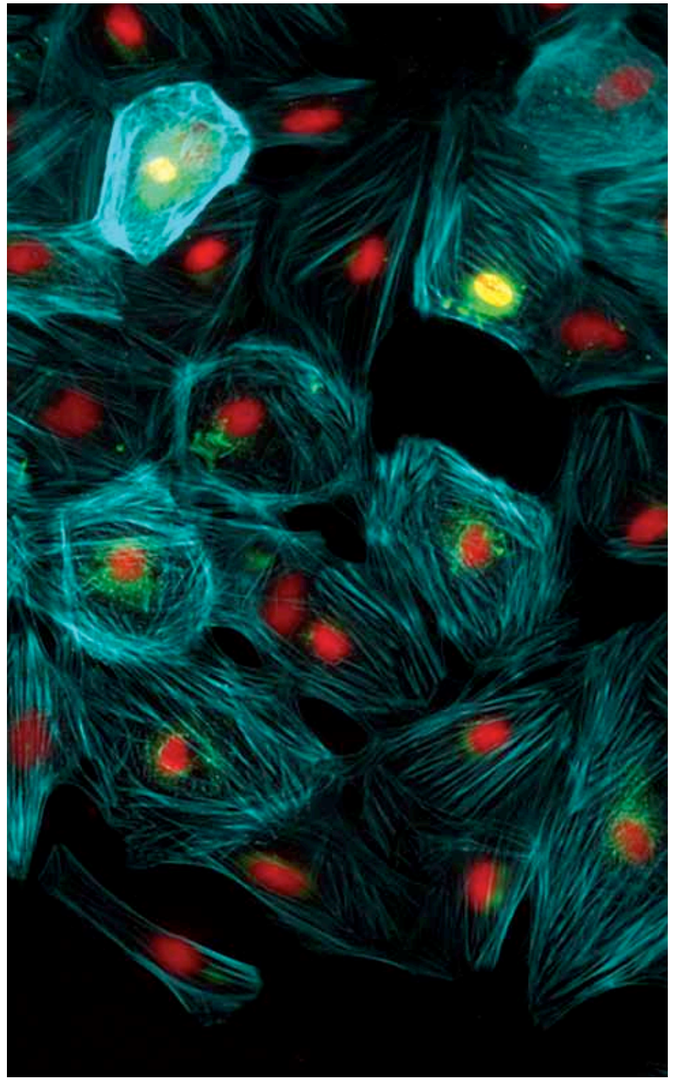


Case study:
Tackling Severe Asthma with Big Data

U-BIOPRED (Unbiased BIOmarkers in PREDiction of respiratory disease outcomes) is a €27M research project founded by the Innovative Medicines Initiative (IMI). IMI is Europe's largest public/private initiative aiming to hasten the development of better and safer medicines for patients; and which aims to speed up the development of better treatments for patients with severe asthma. The U-BIOPRED consortium includes representatives of all stakeholder groups by involving 20 academic institutions, 10 bio-pharmaceutical industry partners (EFPIA; the European Federation of Pharmaceutical Industries and Associations), six patient organisations, three small to medium enterprises and one multinational industry, in 12 European countries.

The project consists of longitudinal clinical studies on adult and paediatric subject cohorts, as well as associated translational studies using animal, in vitro and in silico models of asthma. To add to the study complexity, samples collected from the aforementioned subjects and models are profiled using high-throughput "omics" technologies, producing vast amounts of data, which may be able to produce novel molecular classifiers that better describe the disease heterogeneity.

The ultimate goal of the project is the identification of novel biomarkers of severe asthma acquired by both invasive and non-invasive sample collection techniques, as well as to enhance the mechanistic understanding of asthma aetiopathogenesis leading to the discovery of novel therapeutic targets.

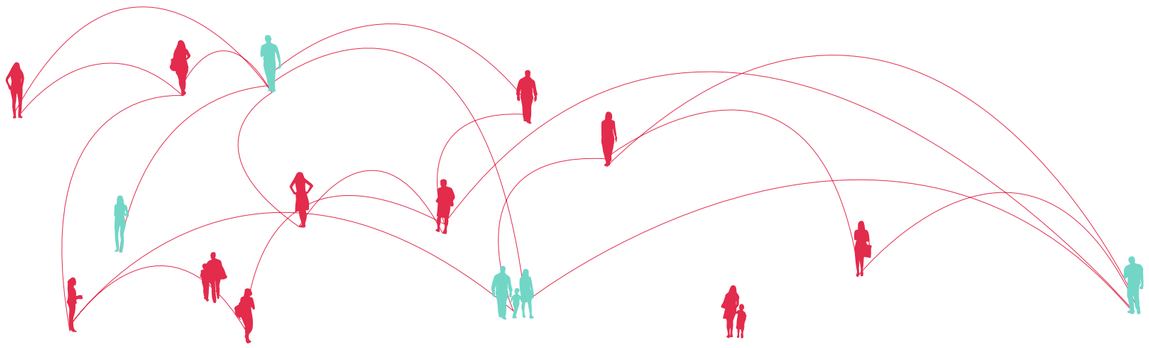


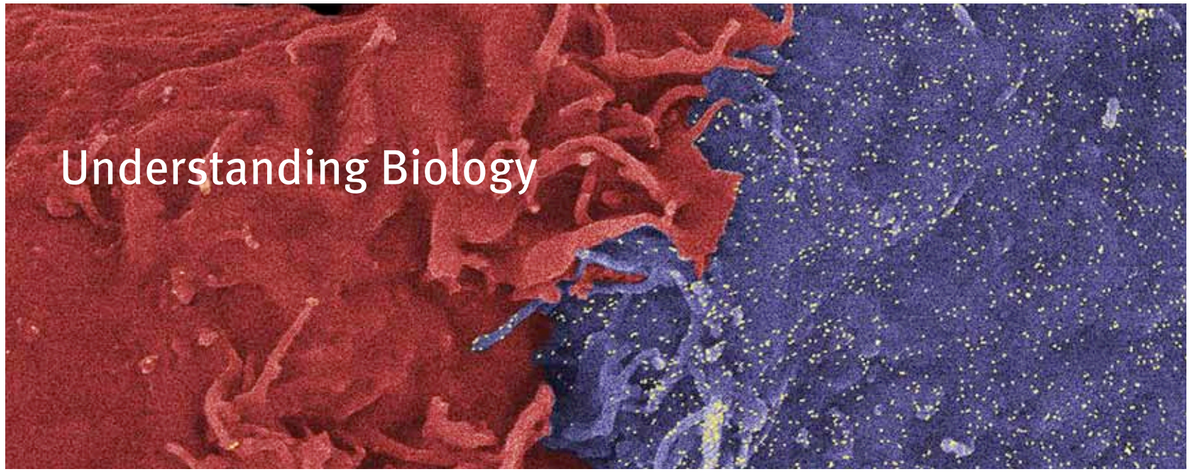
Combining the open source tranSMART software platform (to which the DSI is a major contributor) with a wiki-based collaboration portal, developed by the DSI, led to expedition of data harmonisation, access and analysis for this large scale international research collaboration in biomarker discovery. In addition to the data integration and analysis support, the extended platform enabled the management of scientific hypothesis generation, analytic workflow and publication process, whilst ensuring full transparency to a multi-stakeholder project. This is the first real collaborative knowledge management platform in EU translational research. Its success inspired major investment to the knowledge management technology and service support in the IMI program through the funding of eTRIKS and EMIF projects. The DSI leads the eTRIKS project with total funding of €24M over five years.

In scientific terms the return on investment has been the following:

- Ten statistical analysis workflows linked to source data reviewed and commented.
- Avoided redundancy and inconsistency in results.
- Enabled a large consortium to work effectively in collaboration.
- Ten papers in publication pipeline with significant scientific achievements (biomarkers and methods).
- Data added to the UK Asthma Registry by a simple tranSMART download.

U-BIOPRED is now fully support by eTRIKS. For this work the U-BIOPRED project was awarded the 2014 Bio-IT World Best Practices Award for Research & Drug Discovery.





Understanding Biology

With the advances in high-throughput experimental technologies, biologists are starting to grapple with massive data sets. Multi-omics data, such as genomics, transcriptomics, proteomics, and metabolomics data, is growing astronomically. This data creates a potential goldmine of insights as they provide system-level measurements for all types of cellular components in a model organism, yielding unprecedented views of the cellular inner workings. Meanwhile, this data also raises many new research challenges, not only because of the size of the data but also its increasing complexity. It becomes more and more difficult to handle this data with traditional software and hardware. For example, functional genomics, using sequencing technology to investigate the expression levels of thousands of genes simultaneously at the whole genome level, produces data of very high dimensionality, with variability from experiment to experiment. Such data requires tools such as advanced sequencers, laboratory information management systems (LIMS) for data capture, automated data processing for normalisation and analysis, and data annotation and curation tools for interpreting and integrating the data and analysis results in a biological context. 'Omics'-scale data presents significant opportunities in uncovering hidden patterns of molecular behaviour and in building predictive models for practical biomedical applications, and requires efficient solutions for its capture, curation, search, analysis, storage, transfer, sharing and visualisation.

As a biology-based interdisciplinary field of study, systems biology focuses on the computational and mathematical modelling of complex interactions within biological systems. To deal with large volumes of biological data, scientists are creating data management platforms that allow flexible handling of the data generated from high-throughput experimental technologies such as next generation sequencing (NGS) and processing of the data into information with biological semantics. Meanwhile, novel analytical algorithms and cutting-edge computational technologies have also been developed to accelerate the speed of data processing and analysis. Integrative approaches are being explored for combining different types of molecular information, such as genetic, genomics, proteomics, cellular signalling, and phenotypical information. These new tools and techniques for handling big data promise to turn masses of information into a better understanding of the basic biological mechanisms. Such understanding is then applied to various applications, for example, drug discovery and diagnosis.

At the DSI, we are working with biologists and statisticians, mathematicians, engineers and computer scientists, to enable next generation bioinformatics. To make sense of 'omics'-scale data and gain insights from this data that will advance our understanding of the underlying biological mechanisms, statistics, machine learning and high performance computing are brought together at the DSI for systems biology research to explore fundamental questions in biology. Furthermore, it is expected to lead to practical applications in medicine, drug discovery, and bioengineering.

Our Environment

Where understanding the human body and biological processes is a complex and data-intensive challenge, consider scaling up our understanding of the complexities of nature to the environment around us, from the climate to our planet.

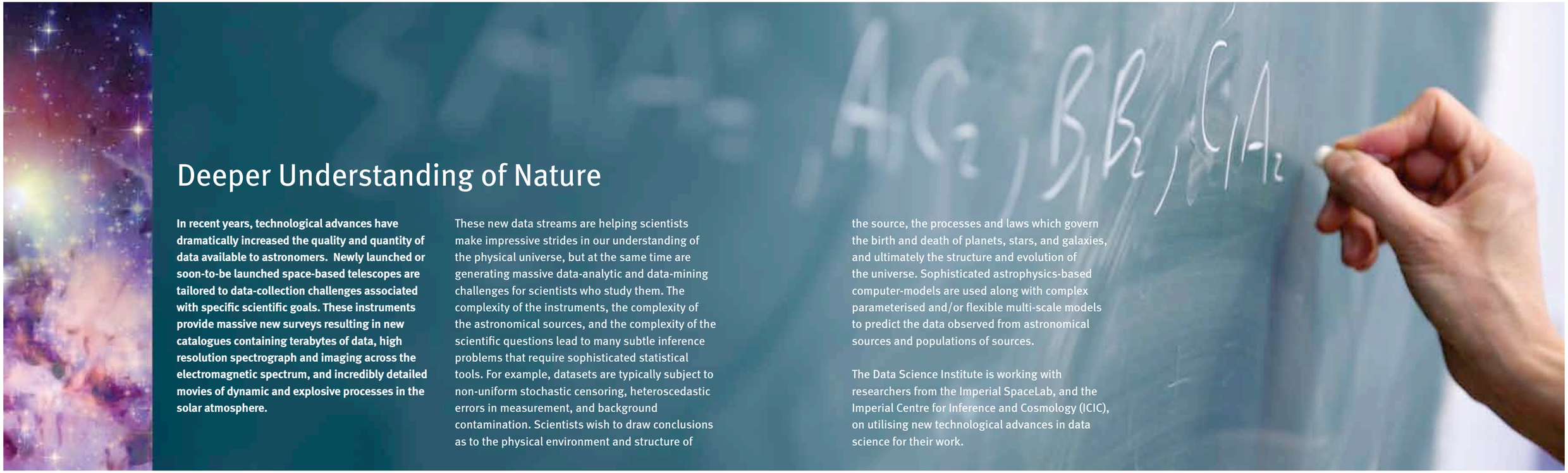
Simulation, modelling and knowledge discovery needs to operate at every biological scale extending up to a global scale that not only includes biological scales (from microscopic upwards), but also behavioural complexities of interacting species, geophysical and meteorological systems, and biodiversity and even the spread of disease. For example, climate scientists have long worked on understanding global processes and create simulations based on computational models to predict how such processes, for example weather, might progress. A new, and welcomed, challenge is that the barriers to gathering, storing and processing large amounts of environmental data are decreasing, but the technical barriers for domain scientists have increased. At the DSI, we are working with environmental researchers to understand how to deal with their data, apply new statistical techniques, and lower the barriers to harnessing complex data in their research by building new tools and training scientists in data science methods. We are collaborating with world-leading research groups and centres, such as Imperial's own Space and Atmospheric Physics Group, and the Grantham Institute for Climate Change, to help them harness the power of big data in climate prediction, risk mitigation, and natural disaster planning. The DSI is now leading the development of a doctoral training programme in the area of big data for risk management with our College partners and other London universities.

Data-Driven Engineering

Engineering and understanding complex systems has always been data intensive, where the main branches of engineering: chemical engineering, materials and earth science and engineering, civil, electrical and mechanical engineering, all utilise simulation and modelling to make advances in technology. With the complexity of computer-based models ever increasing, the simulations on such models are increasingly data-intensive and require a range of numerical methods such as the finite and analytical element methods, ODEs (ordinary differential equations) and PDEs (partial differential equations), as well as agent-based methods and machine learning to make models and simulations more adaptive and accurate. Also, comparing the simulation results with real world observations provides insights to improve models. Thus, the developing new methods and technologies of such Bayesian style data-driven modelling is a common trend in engineering research and development.

DSI works closely with our colleagues to use data in a broad range of engineering areas including in multiphase flow research, sensor networks, robotics, high-performance computing hardware design such as in FPGAs, aeronautics, and bioengineering. The DSI aims to support engineering research by assisting in the application of big data across engineering disciplines.

Beyond progresses driven by data, engineering also enables the invention of new hardware devices to harvest data and harness the power of data. Computing devices equipped with sensors today are pervasive, where the Internet of Things is a driving force for generating data and in-turn pushing forward the forefront of innovations in big data technologies. The DSI stays at the forefront of data-driven engineering research - for example, we are actively engaged in developing pervasive sensing technology for healthcare by combining data streams taken from body sensors (EEG, actimetry, gait), environmental (temperature, light, pollution), and behavioural sensing (emotional state, geo-location, consumer behaviour). This research brings together engineering sub-disciplines: bioengineering, civil engineering, computing, electrical and electronic engineering, with the Data Science Institute acting as the focal point to make innovations in well-being research and social dynamics. The WikiHealth system developed by DSI is now becoming a platform for generic data management and the analysis of physiology and lifestyle data.



Deeper Understanding of Nature

In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. Newly launched or soon-to-be launched space-based telescopes are tailored to data-collection challenges associated with specific scientific goals. These instruments provide massive new surveys resulting in new catalogues containing terabytes of data, high resolution spectrograph and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere.

These new data streams are helping scientists make impressive strides in our understanding of the physical universe, but at the same time are generating massive data-analytic and data-mining challenges for scientists who study them. The complexity of the instruments, the complexity of the astronomical sources, and the complexity of the scientific questions lead to many subtle inference problems that require sophisticated statistical tools. For example, datasets are typically subject to non-uniform stochastic censoring, heteroscedastic errors in measurement, and background contamination. Scientists wish to draw conclusions as to the physical environment and structure of the source, the processes and laws which govern the birth and death of planets, stars, and galaxies, and ultimately the structure and evolution of the universe. Sophisticated astrophysics-based computer-models are used along with complex parameterised and/or flexible multi-scale models to predict the data observed from astronomical sources and populations of sources.

The Data Science Institute is working with researchers from the Imperial SpaceLab, and the Imperial Centre for Inference and Cosmology (ICIC), on utilising new technological advances in data science for their work.

Imperial West (pictured) is a new 25 acre innovation district in west London where much of College's smart cities research will take place.

Sensing Smart Cities

Smart Cities encompass approaches to increasing efficiency in an urban infrastructure, where the efficiency gains are sought through the intelligent management, use of computing technology, and citizen participation, much of which is based around utilising urban data. The drive toward Smart Cities alongside the rising adoption of pervasive sensors is leading to big sensor data, which is so large and complex that it becomes difficult to use traditional methods to utilise it. For a modern city, a torrent of data is collected from sensors in various domains everyday. Sensors are now the dominant source of worldwide-generated data, with 1,250 billion gigabytes (1.25M petabytes) in 2010. And it's growing exponentially. Sensor data has also high-velocity (collected and processed in real-time) and high-variety (collected by networks of diverse kinds of sensors). As these sensor datasets are interconnected with each other, the volume of the integrated data becomes even larger. Sensor data platforms should be able to support such a high volume of data, as well as the large scale of its application. While systems already exist for storing and managing large-scale sensor data, the real value of such data is the insight that can be generated from it along with digital services can be devised as data products. However, there is currently no platform that enables sensor data to be taken from collection through its use in computational models to produce useful data products. At the DSI, we explore these key challenges and provide a response through a sensor data software platform called Concinnity. Concinnity takes sensor data from collection to final product via a cloud-based data repository and an easy-to-use workflow designer. It supports rapid development of applications built on sensor data using data fusion and the integration of models to form novel workflows. These features enable value to be quickly derived from sensor data to drive the development of digital services in a smart city. Concinnity is key to the infrastructure used in the UK Digital Economy project, Digital City Exchange.

Economics, Finance & Value

Data in financial markets is constantly generated and quickly changing, where traditional statistical models quickly become out-of-date, and increasingly high performance real time machine learning algorithms that can adapt and change based on the new data are giving a new edge to trading. This is an active area of research where the DSI and Imperial College's quantitative financial researchers are working together to make exciting progress. Beyond using data science for exploring competitive advantages in economies, new models for the economy itself are emerging. For example, money itself is increasingly exchanged digitally, which brings new research questions such as: How to realise a new commercial structure with digital money? Does digital money adoption make a difference? What are the big data implications of digital money? Is it possible to quantify the benefits to governments, corporations and individuals? What are the factors that affect the outcome of a digital money initiative? Going beyond this, data itself could be considered a currency. At the DSI we are beginning to investigate new models for the digital economy as well as supporting data-driven research in economics and finance.

Case study:
Data as a Currency

Personal data collection has become pervasive with the widespread and ever growing use of smartphones equipped with sensors like GPS. These datasets have great value not only for the data owners and creators, but also for companies and governments after fusing and analysing them for example, the crowdsourcing-based intelligent traffic monitoring application Waze. However, the current mobile application marketplace model has caused a critical issue that the application can declare a minimal set of data permissions, whilst in time, collect vast quantities of valuable and sensitive user data. On the other hand, the user typically cannot get reward from their contributed data. We propose to use a novel model to protect these sensitive user datasets, and fluidify the data to maximise their value.

An important problem of the current consumption of user-generated data is that it's not explicitly indicated before user purchasing applications. And although in some marketplaces like Google Play where the access of user data by the application is indicated, the granularity of data access authentication is not enough. For example, even though an application indicates it will access the user's location data, it doesn't state how much and how frequently it would collect such data. It is not reasonable for a location-based application that a user might use once a day to run in the background and continuously collect a user's location data. Last year, Facebook, Apple, Twitter, Yelp and 14 other companies were served a lawsuit accusing them of distributing privacy-invading mobile applications. In February 2012, Path, a mobile social networking service, was widely criticised for uploading contacts from iPhone users without permission.

To address this problem, we designed a pay-by-data model in which the usage of user-generated data is explicitly indicated, and the access to user's data is controlled by new authentication service that supports more detailed granularity. The model has five different components: data collection service, data pricing agreement, authentication service, mobile platform interface, and application marketplace. The data collection component provides an infrastructure (based on our WikiSensing data integration platform) where a user can store and manage their data in a secure manner. The data pricing enables the agreement between the application and user on how user-generated data will be consumed by the application. Once the agreement is made, the access to the data is controlled by the authentication service; only serving data within the scope and amount in the data pricing agreement. This interface between application and the application marketplace provides a way for current applications and marketplaces to integrate with pay-by-data model. They cut off the channels for developers to get the data directly from mobile phone, and put all the data access in a controlled and authenticated context.

Based on this pay-by-data model, we propose to further develop the data concurrency model. In this model, data collected by the mobile application is viewed as the payment from the user to purchase the utility or service from the application. So the data pricing agreement sets the value of data, or the price of the service. And such an agreement can be extended for "data-data", as the agreement between data exchange, and thus enables the data fluidity between applications. So a data owner can exchange its data with some other's data at the rate they agree. Only through this can the data collected be exchanged and their value maximised, while the benefit of the users are safeguarded.

We have created a prototype system of pay-by-data platform, based on a revised version of OAuth2 protocol and service for data usage authentication, and a customised version of the Android platform to redefine the data collection process. Based on this work, the future research would build the comprehensive platform, and then investigate the mechanism to enable data currency.

Facilities

State-of-the-art facilities at the DSI

The DSI is housed in purpose built facilities in the heart of Imperial College London's campus in the Borough of South Kensington. Such a central location provides excellent access to collaborators across the College and across London.

Our purpose built facilities are designed to enable collaboration across Imperial College and with our industrial partners, the academic community and the wider public. Our dedicated space within the William Penney Laboratory has facilities for over 50 researchers and collaborators. The DSI also has a dedicated data centre, largely donated by Huawei Technologies, to facilitate the institute's research activities. Uniquely, visualisation was a key focus in the design of our facilities, knowing that data can only be comprehended to the degree it can be visualised we have incorporated a number of visual communication areas into our building.

Our central location on the main thoroughfare of campus means that we will be able to incorporate visualisation technology in the form of a video wall onto the facade of our building to enable communication with students and academic staff. To complement this we will create a touch sensitive display that will be accessible to passers-by allowing two-way communication. Internally we have an innovation display space to showcase the research within the Data Science Institute.

The centrepiece of our facility is the KPMG Global Data Observatory. This is a data visualisation and decision-making studio at the heart of the Data Science Institute enabling academic and industrial teams to interactively visualise, analyse and gain insight from datasets. The heart of the studio is a large 100 mega-pixel immersive display environment. Consisting of 48 monitors, this space will follow the decision theatre model allowing participants to be surrounded by data, models, and analytics which are all visualised and modifiable to show them the impact of their decision-making process. The studio will allow for 'hackathon' style events in small groups each working with high-end display equipment to gain insight into data. This unique space will enable great innovation in data visualisation and analysis.

Education

Educating the next generation of data scientists

One of the core objectives of the Data Science Institute at Imperial College is to promote the training and education of the new generation of data scientist by developing and coordinating new degree courses, and building international academic joints labs.

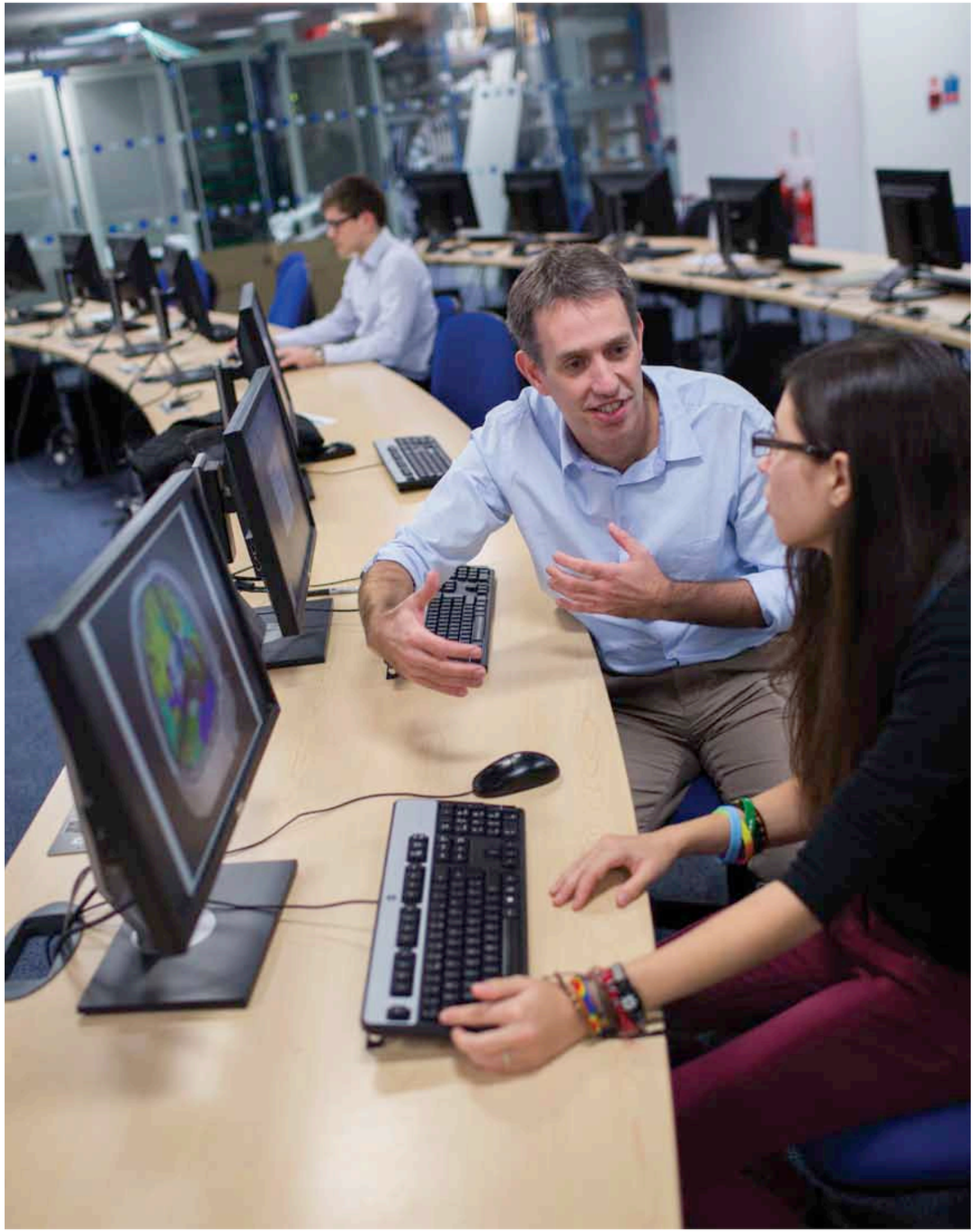
Education and training programmes

Setting up new data science education and training programmes at Imperial College provides a unique opportunity to innovate in the way that courses are designed and delivered. We are developing three different models for delivery, which will set Imperial apart from the rapidly growing number of data science courses offered at university level across the United Kingdom and worldwide. These models will build on Imperial's multidisciplinary data-driven research. We are planning to introduce data science courses by:

- Introducing data science modules, with a balance derived from computing and mathematics, as additional electives alongside existing MSc programmes across Imperial's four faculties, to teach core topics in data science within domain-specific specialisms during the advanced stages of an MSc course. We aim to launch these modules for the academic year 2015/16.
- Developing a flexible part-time MSc in Data Science directed at professional students who would continue to be active in industry while studying, delivered in partnership with the Centre for Continuing Professional Development. We aim to launch this degree programme for the academic year 2016/17.
- Developing a unique collaborative international degree programme in data science, partnering Imperial College with universities in the United States and China. We aim to launch this degree programme for the academic year 2017/18.

Joint Academic Labs

International academic partnerships are vitally important as the global community confronts challenges in energy, the environment, health, data, and security that are well beyond the ability of individual universities or nations to solve. The DSI is therefore looking to create links with other academic institutions worldwide to more effectively tackle these grand challenges. For example, the recent establishment of the Imperial College London-Zhejiang University Joint Lab for Applied Data Science creates a critical mass of expertise that will drive data innovation and create new solutions and technologies. The lab will facilitate student and academic researcher exchanges between Imperial and Zhejiang University to enable joint research projects funded by the Chinese and UK governments as well as hosting Zhejiang University-funded scholars to study towards their doctorates at the DSI.



Industrial Collaboration and Translation

Innovation is one of Imperial's key strengths, taking basic research and applying it to new domains thereby producing new products and services. DSI's industrial partnerships form the basis for innovation by enabling our researchers to tackle real-world problems, putting theory into practice.

Larry Hirst, CBE
Chairman of the DSI Advisory Board

Strong industrial collaborations are at the centre of the DSI's development. We focus on translating our foundational and technological research into new products, services and businesses, leading to high visibility and impact of the Institute. At the DSI we devote resource and effort to work closely with our industrial partners to pursue world-class research and encourage innovation with them to fast track our foundational data science research into tangible value.

Joint Industry Labs

With DSI acting as a focal point for data-driven research and application, we have partnered with several key industry organisations in order to facilitate new research driven by industry needs as well as giving us a route to commercialisation with established market leaders.

Imperial College-Huawei Data Science Innovation Lab

The Huawei Data Science Innovation Lab is a joint partnership between Imperial College and Huawei Technologies hosted in, and coordinated by, the Data Science Institute. Its mission is to foster cooperation between researchers at Imperial and Huawei Research by funding interdisciplinary research projects in the area of big data. In particular, the lab aims to demonstrate interdisciplinary innovation by taking data analytics methods developed in one application domain (e.g. healthcare) and reapplying them successfully in another (e.g. telecoms).

For 2014-2016 five research themes have been identified:

- **High-performance deep learning.**
- **Scalable models of information propagation in networks.**
- **Scalable algorithms for mining high-dimensional / high-frequency data.**
- **Body sensor networks and informatics.**
- **Big data for big science.**

Former President of Imperial Sir Keith O'Nions and Huawei's William Xu.

KPMG Global Data Observatory

Imperial's Provost Professor James Stirling and KPMG's Jim Marsh.

The Global Data Observatory, housed in the Data Science Institute, is a data visualisation studio and decision-making space funded as part of a £20M investment by KPMG to Imperial College London. The Global Data Observatory will facilitate decision-making through immersive, multi-dimensional presentation of data and analysis that allows decision makers to participate in exploring data, weighing the findings of analytical exercises, translating analytical findings into insights, and deriving the implications and actions suggested by insights from the analytics. As well as providing state-of-the-art visualisation facilities for the Imperial's academic and industrial partners, it will act a test bed for novel data visualisation research and development.

Outreach

DSI in the Community

Events

Key to DSI's engagement activities, both within Imperial College as well as with the wider research community, is a rich series of events that include our own 'Distinguished Speaker' lectures that have so far included author of Big Data: A Revolution That Will Transform How We Live, Work and Think Professor Viktor Mayer-Schönberger of the Oxford Internet Institute, as well as renowned data mining expert Dr Usama Fayyad, the Chief Data Officer of Barclays Bank. The DSI also runs a seminar series with invited speakers from the academic community, industry, and third-sector organisations that has so far featured speakers from the Allen Institute for Brain Science, École Polytechnique Fédérale de Lausanne (EPFL), SAS Institute, eBay, Thomson Reuters, dunnhumby, Teradata and DataKind. The DSI also helps prepare students for careers in data science by facilitating student-centred events to engage the student community in data science.

Student Competitions

As part of our student engagement activities, the DSI also runs a number of competitions.

The DSI established the 'Summer Data Challenge' – an annual contest to find new insights in a selected collection of datasets by combining them with any open data of choice, and to propose how these new insights can be translated into social or economic benefit, for example the creation of new start-ups or business units in existing companies, or the development of new products. We worked with our industry partners Starcount, Purple Seven, and Transport for London, who together provide the competition access to a collection of exclusively prepared datasets. A cash prize is provided through funding from Research Councils UK, a mentoring session with an industry leader in data science from our partner, Starcount, and start-up support from Imperial's own business pre-accelerator, Imperial CreateLab, are all offered as prizes for the three top-ranked entries.

The DSI is also creating a Best Thesis in Data Science award that opens to all doctoral theses that utilise a data-driven approach or contribute directly to the data science literature. Theses will be judged by a panel of Imperial College experts with a cash prize and an opportunity to rewrite the winning thesis into a book to be published as part of a new book series – New Frontiers in Data Science.

Imperial College London

Data Science Institute
William Penney Laboratory
Imperial College London
South Kensington Campus
London SW7 2AZ
United Kingdom
www.imperial.ac.uk/data-science

Designed by jcreative.co.uk