Centre for Doctoral Training for Innovation in Data Intensive Science

Contents

Welcome

Welcome to the Liverpool Centre for Doctoral Training for Innovation in Data Intensive Science (LIV.INNO).

The amount of digital data that exists in the world is growing at a rapid rate. Recent years have witnessed a dramatic increase of data in many fields of science and engineering, due to the advancement of sensors, mobile devices, biotechnology, digital communication and internet applications. Data is generated continuously from multiple sources by companies, users and devices in a huge velocity, volume and variety.

The large volume of data inundates businesses on a day-to-day basis. However, it is not only the amount of data that matters; it is how to analyse big data for insights that lead to better decisions, business strategies and innovations.

Based on the success of Liverpool's first Centre for Doctoral Training (CDT) in Data Intensive Science, LIV.DAT, the new CDT LIV.INNO has quickly established itself as a hub for training students in managing, analysing and interpreting large, complex datasets and high rates of data flow. The CDT features a unique training approach, addressing some of the biggest challenges in data intensive science to tackle a growing skills gap in this important area. The training centre is supported by the Science and Technology Facilities Council (STFC) and hosted by the University of Liverpool and Liverpool John Moores University / Astrophysics Research Institute.

We offer our students comprehensive training in data intensive science through cutting edge research projects and a targeted academic training programme, complemented by placements in industry.

In this publication, you can find out more about the many collaborations and partnerships that we have, as well as discover some of the ways in which our students are working with industry during their placements.

We are also pleased to share with you the stories of many of our students about why they chose the CDT at Liverpool, the wide range of research areas they are involved in and how they helped the organisations they worked with.

We hope you enjoy reading about our centre and look forward to collaborating with you on your data-related research challenges.

Carsten P Welsch
LIV.INNO Director

Our expertise

Prof Carsten P Welsch
Director

Professor Carsten P Welsch has initiated and led the two Liverpool STFC Centres for Doctoral Training, LIV.DAT and LIV.INNO, as well as six pan-European research networks that have trained more than 100 postgraduate researchers. His R&D covers the design and optimisation of accelerators and light sources and underpinning technologies, in particular advanced beam diagnostics and high gradient acceleration, as well as the application of accelerators with a focus on healthcare instrumentation and data science.

As past Chair of STFC's Education Training and Careers Committee, member of the UKRI Skills Advisory Group, member of STFC Council and scientific advisor, he helps improve researcher training and drives the advancement of physics.

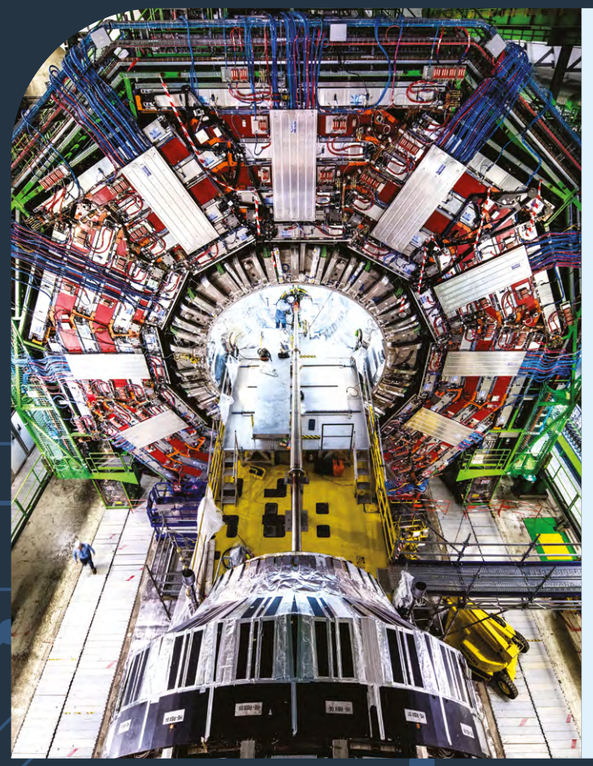
Dr David Hutchcroft
Deputy Director

Dr David Hutchcroft is an experimental high-energy physicist working as part of the LHCb collaboration. He writes software to simulate, trigger, reconstruct and analyse data from the LHCb experiment at CERN to look for new physics processes seen in the decays of the b quarks. He is also working on building the new silicon pixel vertex detector for the upgrade to the experiment in Liverpool. Dr Hutchcroft runs the GridPP site at Liverpool, part of the international computer network that is used to provide the computing resources for the LHC and other science experiments.

Dr Andreea Font
Deputy Director

Dr Andreea Font is a Reader at the Astrophysics Research Institute (ARI) at Liverpool John Moores University and a member of the Computational Cosmology group at ARI. Her research interests are in the area of formation and evolution of galaxies and on the nature of dark matter. A lot of her research focuses on modelling the properties of our own Galaxy, the Milky Way and of other 'Milky Way analogues'. To this end, she has recently constructed a suite of Milky Way-type simulations called ARTEMIS, which enabled her to compare theoretical models with observations of Milky Way-type galaxies.

Driving innovation through data science



Data-driven physics

Data intensive science in physics is generating extremely large and complex data sets that require to be analysed computationally, whether it is the Astrophysics Research Institute simulating the evolution of the Universe from just after the Big Bang all the way to present day or a research team trying to improve cancer diagnosis. Physicists working in astronomy, particle, nuclear physics, accelerator, mathematical and computer science develop novel methodological approaches ranging from machine learning to data analysis techniques to inference and modelling with applications in science and industry.

The impact of data science on our lives

The last 10 years have witnessed big changes in data science technologies like Artificial Intelligence (AI), quantum computing and 5G. These developments transformed society by enhancing connectivity, providing access to online education and developing amongst others new AI-enabled health technologies that diagnose diseases and extend life expectancy.

The integration of data science technologies into all areas of a business has resulted in fundamental changes to how businesses operate and how they deliver value to customers. Big Data has been the underpinning driving force behind these global technological developments in society and businesses.

LIV.INNO CDT: Preparing the next generation of data scientists

The LIV.INNO training programme is designed to address a wide range of employment skills, including research skills and techniques, project management, networking, communication and presentation skills, with the aim to provide all students with the skills set required for a future career in both, academia and industry.

This involves schools, seminars, data science forums, secondments to industry of several months duration, as well as high impact outreach events.

Areas of research in data intensive science

Recent years have witnessed a dramatic increase of data in many fields of science and engineering, due to the advancement of sensors, mobile devices, biotechnology, digital communication, and internet applications. Managing, analysing and interpreting large, complex datasets and high rates of data flow is a growing challenge for many areas of science and industry.

Our students at LIV.INNO receive a comprehensive training programme in data intensive science to address data challenges presented by research in astronomy, nuclear, particle and accelerator physics. This training takes place via modules at both universities as well as via the students' individual research projects. The following pages show a few selected case studies that highlight the enormous breath of research activity that is being carried out in the centre:

1

'Monte Carlo and High Performance Computing (HPC)' is led by LIV.INNO Deputy Director Dr David Hutchcroft, Senior Lecturer at the Department of Physics at the University of Liverpool, where he researches High Energy Physics.

2

'Artificial Intelligence (AI) and Machine Learning (ML)' is led by Dr Anh Nguyen, Lecturer in Artificial Intelligence in the Department of Computer Science at the University of Liverpool.

3

'Data Analysis' is led by LIV.INNO Deputy Director Dr Andreea Font, Reader at the Astrophysics Research Institute at Liverpool John Moores University where she researches Astrophysics.

1 Monte Carlo and High Performance Computing (HPC)

LIV.INNO students receive a thorough grounding in Monte Carlo theory. Through R&D projects in astrophysics, nuclear, accelerator and particle physics, they learn about their application in these scientific domains, all of which are linked with Big Data challenges. Monte Carlo methods and HPC are powerful tools for everything from modelling the birth and evolution of the universe to discreet-event systems in manufacturing and stress /strength stochastic modelling in engineering design. The technique is used by professionals in such widely disparate fields as finance, healthcare, energy, manufacturing, engineering, research and development, insurance, oil & gas, transportation and the environment.

Lauryn Eley

Project: Optimisation of low-dose, low cost mobile 3D x-ray imaging

Supervisor: Prof Carsten P Welsch (UoL)

Lauryn's research looks at Digital Tomosynthesis (DT), an imaging technique that can give a stack of slices through a volume, giving many of the 3D benefits of CT scanners but at dramatically lower cost and dose. She simulates and optimises the system using Monte Carlo (MC) models and then carries out scans to identify optimum system parameters.

2 Artificial Intelligence (AI) and Machine Learning (ML)

LIV.INNO students are trained in AI/ML techniques for solving complex problems in scientific research and discovery. AI/ML training for science is challenging because it has to deal with massive quantities of multi-dimensional data. HPC is increasingly being employed for meeting the exponentially growing demand for AI/MI. Scientists run ML techniques on HPC to extract valuable information from complex data sets. For example, physicists analyse complex data that is produced by particle accelerators, analyse medical images and predict protein structure properties with the help of ML models. Deep learning enabled AI technology is widely used in industry to recognise images, process natural languages, play games better than humans and drive cars.

Andrea Sante

Project: Reconstructing the assembly history of our Galaxy using neural networks

Supervisor: Dr Andreea Font (LJMU)

Andrea's research looks at the cosmological model of how galaxies were formed by using the evidence left behind today in the form of debris from the destruction of smaller dwarf galaxies. From the number and shapes of these stellar streams it is possible to reconstruct the accretion history of galaxies and constrain the nature of dark matter. ML techniques are critical in this project and Andrea uses a combination of convolutional and recurrent neural networks to extract the contextual information related to the tidal streams and to model their time evolution.

Robert McNulty

Project: AI: from high energy physics to medical applications

Supervisor: Dr Nikos Rompotis (UoL)

Robert's research looks at using AI techniques in two very different fields. Firstly, he works with members of the ATLAS team at Liverpool to commission and improve ML algorithms used for efficient identification of objects including tau-lepton and heavy-flavour jets. These techniques are then applied in the future Higgs-pair analysis to improve its sensitivity. Secondly, Robert works alongside experts at ARO to evaluate the performance of convolutional neural networks to Magnetic Resonance (MR) images data. He then utilises the acquired AI knowledge to further improve the sensitivity of the Higgs-pair production analysis.

3 Data Analysis

LIV.INNO's approach to data analysis is driven by the development of efficient numerical Monte Carlo and Deep Learning techniques and their application to solve complex data problems. As organisations continue to generate enormous amounts of data, they recognise the importance of data analytics to make key business decisions. The development of efficient Monte Carlo techniques provides optimal design, scheduling and control of industrial systems, the development and analysis of new materials and structures and the risk analysis of large portfolios of financial products. This is complemented by AI technologies such as Deep Learning methods which seek greater insight among the ever-increasing amount of data in several key industries and powered by technological advancements as in, for example, computer vision, natural language processing, Internet of Things (IoT) and computer hardware.

Alexander Jury

Project: Longitudinal Density Monitor for the Large Hadron Collider (LHC)

Supervisor: Prof Carsten P Welsch (UoL)

Alexander's research uses the Longitudinal Density Monitor (LDM) on the LHC. The LDM is a key diagnostic in the LHC for particle physics experiments and measures the luminosity profile with very good resolution. This monitor produces a vast amount of data used for luminosity calibration and machine control. Alexander's PhD project focuses on the development of novel approaches to analyse this data and develop a novel tool for precise real-time luminosity measurements.

Katherine Ferraby

Project: Data Analysis with Deep Learning technique and real-time event reconstruction in MUonE

Supervisor: Prof Thomas Teubner (UoL)

Katherine's research is working on MUonE, a proposed project at CERN, to measure the shape of the differential cross section of μ-e elastic scattering as a function of the space-like squared momentum transfer. This project produces large amounts of data with around 10^{12} events to be analysed. Katherine focuses on performing quality tracking in the real time trigger reconstruction. Algorithms of track finding techniques based on ML are developed and applied to the reconstruction process at the online and offline stages.

How are students **making a difference?**

Training in LIV.INNO is based on the model developed as part of the successful delivery of the first CDT LIV.DAT and a number of pan-European training networks that were all coordinated by Professor Welsch.

The LIV.INNO PhD schemes specialise in their industry links and the industry placement is an excellent opportunity for the students to apply their bespoke LIV.INNO training to the real world, while also gaining knowledge and experience of working outside of academia. The R&D of a placement project is in an area outside of the student's core PhD research project to give the student a new experience.

Our track record in big data science and innovation stretches back to the LIV.DAT CDT, the predecessor of the LIV.INNO CDT. Here we showcase a selection of the placement projects completed by LIV.DAT PhD students in data intensive science projects.

Mental health and wellbeing

Selina Dhinsey

Selina completed her six month placement at Chanua Ltd, an organisation that creates mental health and wellbeing tools using digital innovation. The project she worked on, NeuroLove, was born out of COVID-19 and the lockdown during the pandemic. It is an online platform, aimed at young people aged 8-25, supporting them to stay mentally and physically well during these difficult and overwhelming times, especially with so many out of school. It has a whole library of resources to keep them busy and active with a team of social therapists available to talk to. There are daily sessions ranging from baking and augmented reality workshops to dance classes, yoga and low mood workshops.

Selina worked on developing the chat function and training social therapists on how to use it. She was also the data analytics lead, exploring how the site is used and by whom. Her work allowed access to the site to those who needed it most.

> "During my placement, I've worked on a range of projects mainly around using the data collected to see how improvements and adjustments can be made to better serve people. I've had to get to grips with new systems and platforms from the very beginning, which has been challenging but so rewarding to see my own growth and progression. I've spent a lot of time becoming familiar with the Amazon Web Services architecture and seeing how they can aid data analytics. Some work I've done has even surrounded the social media and YouTube realm, something that I really didn't expect but have thoroughly enjoyed.
>
> I've also been able to participate in Neuro Champions workshops, which aim to educate the young through the cross-section of neuroscience, mental health and technology."

Developing a railway worker safety tool

Phillip Marshall

Phillip undertook a placement with OnTrack – a software company that specialises in the area of track worker safety – to produce a system that enables contractors to plan work safely on live railway lines. For the past decade OnTrack have been collecting data and recently they decided to utilise it. Phillip was asked to take a deep dive into their data to see what he could find. Phillip worked with colleagues to extract, analyse and visualise the data and prototype a product that could help contractors see if they were meeting targets for productivity and most importantly safety - without having to collect and compile data manually from their employees. During the project, it became evident how disconnected parties holding incompatible data creates difficulties – for example, last minute notices to track workers only existed in pdf format file documents.

> “I was astonished at the possibilities to improve efficiency and safety in the industry if more data was collected and if the data already collected was used to its potential. It gave me an insight into how many industries must be adapting to the new world of big data and how great it is to work on projects involving big data. I also learnt the challenges to building a big data future.
>
> Currently data is often held by many different parties and in many different formats. Managing this challenge to accumulate large, structured and useful datasets is the first important step needed to really see the benefits of big data. For me, moving into industry has always been the goal. Before I started, I had very little experience of big data or computer science and my PhD has taught me almost everything I know.
>
> The chance to do an industry placement has been incredibly valuable and has opened my eyes to the big data world available to me.”

Networking platform for technology companies

Alberto Acuto

Alberto Acuto completed his placement at IQBlade (Liverpool Science Park) where a team of data scientists and tech channel experts work together to help companies that want to find new partners, clients and network growth from a data-driven perspective in a self-contained platform. The strength of this database is the size and quality of the data available on UK-based companies and this is now being expanded to include data on European-based companies. The rate at which this platform is growing, as well as the continued need for improving search results for clients, has highlighted a number of technical challenges in the big data realm.

Alberto’s project at IQBlade was to classify in a smarter and quicker fashion, over four hundred thousand ‘not-yet-classified’ listed companies. The idea behind the project was to use data available on the database to create models to make predictions on the unlabelled data. During his placement, Alberto worked on building a classification algorithm based on the descriptive features obtained from text to classify similar types of companies.

> “The task was challenging for several reasons. First of all, coming from a different background and the lack of basic knowledge was something that I had to deal with at the beginning. Luckily the team helped me out a lot in getting to the right pace to progress with the work. That was my first hands-on experience in text mining (Natural Language Processing) and database queries and handling, so I was really curious to experiment with those new tools.
>
> It was really interesting to see and apply what I have learnt in the last few years and to finally make the most out of that knowledge. These six months have been interesting, challenging and helped me develop a lot in many different aspects.”

Developing a Natural Language Processing Product

Tom Williams Harrison

In early 2019 Tom Williams Harrison did his placement at Exgence Ltd, a start-up that is aiming to provide solutions to software companies that spend a lot of resources processing "Invitation to Tender" (ITT) documents, which are created by a prospective customer of the software vendor. The ITT document contains requirements set by the customer which is then sent to many software vendors which in turn complete them with details about how their software meets the requirements. This process is often not trivial, particularly for larger companies with many software products and whose sales teams are separated from the software engineers.

Tom helped develop a solution that works by analysing a set of existing ITT documents and performing Natural Language Processing (NLP) to extract the semantic information from previously completed answers. He also created suitable pre-processing steps and developed a number of NLP models to process the structured data. This stage was a valuable learning experience in machine learning methods on language-like data. His project then focussed on the software engineering and development stage to turn the product into a completed package that would integrate with the editor of the ITT document (such as MS Excel).

Exgence Ltd brought their first product to the market in July 2019.

> This was an interesting challenge, since there is no set standard for these documents, so they are completely arbitrary in structure and formatting and can even be different filetypes.
>
> Due to the very small size of the start-up and the relatively early time in the start-up cycle when I joined them, I was able to enjoy a lot of freedom in creating the framework, as well as very fast communication and decision times. I was also able to develop most of the product backend essentially from scratch and being familiar with the whole codebase meant that experimenting with new ideas was much easier."

Data simulation modelling

Alex Hill

Alex Hill completed his placement at IBM Research (Hartree Centre at Daresbury), working on surrogate modelling. He started with an introduction to this subject including researching the fundamentals of surrogate methods: predominantly Polynomial Chaos Expansion and Gaussian Process Emulation. The particular motivation of Alex's research was rare event modelling - where the simulation modelled takes input values sampled from a Pareto distribution. The challenge was to produce a surrogate which not only reproduced the expected output statistics of the simulation - i.e. closely approximate the simulator at low values - but also accurately reproduce the simulator outputs along the tail. These two considerations were often in tension with each other and posed a significant challenge.

Alex's key finding was the decomposition of the Pareto input distribution into 'peak' and 'tail' components, followed by the creation of two separate surrogates which were used in tandem to approximate the full model. Significant improvements in the approximation of the simulator output across the support range were found, as well as in the approximation of output statistics. Alex also worked on the 'Curse of Dimensionality' which refers to the unfortunate fact that as the number of input variables of a simulation increases, the computational cost of approximating it increases more dramatically still. Here he focused on optimising the selection of training inputs in higher dimensions, as well as to visualise and quantify uncertainty in the simulation and surrogate outputs. Good progress was made in applying this to more complicated simulations for the client, as well as to epidemiological models for Covid-19, when the internship drew to a close.

> I fully enjoyed my time at IBM, and only wish that I could have experienced the full person-to-person programme envisaged. Despite the challenges, it has changed the way I view approaches to coding, the potential for cross-subject collaboration and the possibilities in working outside academia."

How to **collaborate** with us?

Our partnerships and collaborations with external organisations are a vital part of the CDT's commitment to train doctoral students. We work closely with industry, the public sector and other organisations on a wide range of activities including skills development, sponsorship, and research sandpits.

Here are some of the ways you can work with us:

Joint studentships

Individual PhD projects can be proposed and jointly funded by partner organisations. These students will be supervised jointly by a university academic supervisor and a member of the partner organisation.

Training

There is an opportunity for partners to contribute to a wide range of technical and advanced research skills training. The LIV.INNO students have access to the state-of-the-art multi-purpose training area LIV.HUB as an industry hack space.

Industry placements

As part of their PhD, students complete a 3-6 month industry placement at an external organisation in a field outside of their core PhD project. Placements encourage mutually beneficial research collaboration between the CDT academic researchers, students and partner organisations.

Mentoring

Partners can act as independent mentors to our students for career advice and contribute to student supervisions on research projects of mutual interest.

Research sandpits

Partners can discuss industry relevant problems with students and interdisciplinary groups of academic researchers to spark ideas on industry related problems or challenge themes and form new collaborative research projects.

Guest lectures

We invite a wide range of guest speakers to offer specialist keynote sessions, seminars or workshops featuring a topic relevant to data intensive science and Big Data.

Pilot projects

Organisations may wish to run small scale pilot projects related to their core business. We can identify students with relevant experience and interests to carry out research or implement activities.

Sponsorship opportunities

We welcome support in the form of sponsored events and operational activities such as sponsored prizes or co-funding a one-day event, workshop, mini-conference or research visit.

If you would like to discuss any of the above opportunities, project ideas or explore particular areas of research in 'Data Intensive Science', contact us and we can put you in touch with the right people to assist you.

LIV.INNO

Contact and further details

Prof Dr Carsten P Welsch
LIV.INNO Director

University of Liverpool
Department of Physics
Oliver Lodge Building
L69 7ZE Liverpool, UK

and

Cockcroft Institute
Sci-Tech Daresbury
QUASAR Group
4 Keckwick Lane
WA4 4AD Warrington, UK

www.livinno.org

UNIVERSITY OF LIVERPOOL

LIVERPOOL JOHN MOORES UNIVERSITY

UKRI Science and Technology Facilities Council

LIV.INNO is supported by STFC under grant ST/W006766/1.