Abstract Book

Click on the presenter's name to view the abstract

Thursday 16 September

10:00 - 10:25	**Mihi, Conference Welcome, Housekeeping & Navigation**
10:25 - 10:30	**Break**
10:30 - 11:30	**Presentation session A**
	Theme: Workflows
10:30	*Rose Pearson* Flood resilience Aotearoa: Engaging Aotearoa's wealth of public geospatial data
10:50	*Alena Malyarenko* Coupled Climate – Ice – Ocean Modelling: Setting up a framework
11:10	*Richard Lupat* Janis: A Python framework for workflow translation
11:30 - 11:40	**Break**
11:40 - 12:07	**Presentation session A continued**
	Theme: Workflows
11:40	*Adam Hyde* Building efficient workflows for the rapid review of time sensitive research
12:00	*Jian Liu* Automated workflow aids biophysical model development – Couple R targets and APSIMX
12:07 - 12:30	**Break**
12:30 - 1:05	**Presentation session B**
	Theme: Community and engagement
12:30	*Paula Andrea Martinez* Translating complex topics: Software containers
12:50	*Matt Plummer* Static Tactics: Using static website workshops to develop capability and collaboration
12:57	*Arindam Basu* Communicating epidemiological insights using explorable explanations
1:05 - 2:00	**Lunch Break**
2:00 - 2:50	**Presentation session C**
	Theme: HPC and Code optimisation part 1
2:00	*Manodeep Sinha* Corrfunc: Blazing fast correlation functions on the CPU
2:20	*Wolfgang Hayek* How to draw an owl – Transitioning from tutorials to the real world with Dask
2:50 - 3:00	**Break**
3:00 - 4:00	**Keynote Speaker**
3:00	*Karaitiana Taiuru* Māori Data Sovereignty: An introduction and implementation in the research sector.
4:00 - 4:15	**Break**
4:15 - 5:00	**Birds of a Feather**
4:15	*Nooriyah Lohani* Metrics for measuring the contributions of Research Software Engineers
5:00 - 5:30	**Networking happy hour**
5:30	**End of Day**

Click on the presenter's name to view the abstract

Friday 17 September

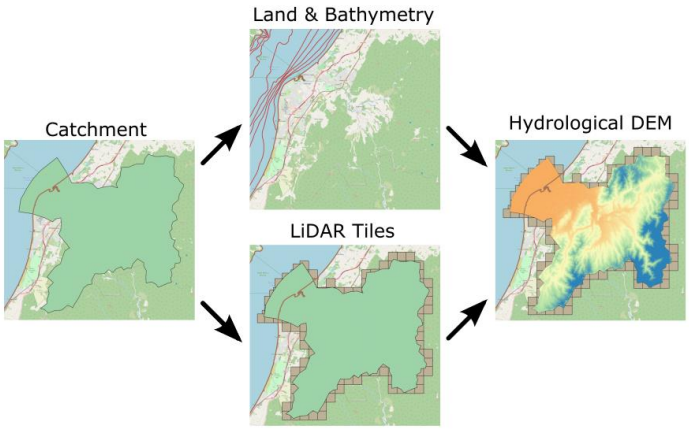
9:00 - 9:15	**Welcome, Housekeeping & Navigation**
9:15 - 10:00	**Birds of a Feather**
9:15	*Linley Jesson , Alan Tan, Richard Dean* Opportunities and challenges for ML-Ops for NZ Research Institutes
10:00 -10:05	**Break**
10:05 - 11:05	**Presentation session D** **Theme: Tools showcase**
10:05	*Simon Anastasiadis* Accelerating data wrangling: A researcher-built dataset assembly tool
10:25	*Hamish Campbell* Data version control & collaboration for reproducible science with Kart
10:45	*Mike Laverick* Automating instrument data workflows: Integrating Globus into MyTardis
11:05 - 11:15	**Break**
11:15 - 12:03	**Presentation session E** **Theme: HPC and Code optimisation part 2**
11:15	*Matija Cufar* Using Julia in a high performance computing environment
11:35	*Alexander Pletzer* Accelerating a physics code with OpenACC
11:55	*Zhihan Wang* Distributed computing strategies for training deep networks on a high performance computing server
12:03 - 1:00	**Lunch Break**
1:00 - 2:00	**Keynote Speaker**
1:00	*Jannat Maqbool* AI and the opportunity for NZ to lead rather than follow
2:00 - 2:10	**Break**
2:10 - 3:00	**Presentation session F** **Theme: Research Software**
2:10	*Tom Honeyman* A national agenda for research software in Australia
2:30	*Paula Andrea Martinez, Georgina Rae* FAIR 4 RS - What, why and what next?
2:50	*Matthias Liffers* Co-opting academic recognition systems for research software
3:00 - 3:15	**Break**
3:15 - 3:55	**Presentation session G** **Theme: ML/ Mlops**
3:15	*Deepak Karunakaran* Building an MLOps pipeline for Lumi Drug Scan – A real-time drug screening system
3:35	*Jan Schindler* Experiences developing an operational workflow for large-scale instance and semantic segmentation of remote sensing imagery using CNNs
3:55 - 4:15	**Closing remarks and conference wrap-up**
4:15 - 4:45	**Networking happy hour**
4:45	**End of Conference**

Flood resilience Aotearoa: Engaging Aotearoa's wealth of public geospatial data

Rose Pearson, Emily Lane, Matthew Wilson, Pooja Khosla, Cyprien Bosserelle, Yu-Ching Lee
NIWA, LINZ
rose.pearson@niwa.co.nz

ABSTRACT / INTRODUCTION

The Endeavour-funded programme Mā te Haumaru ō te Wai aims to produce nationally-consistent flood inundation maps for Aotearoa, while adhering to principles of open science and access. Key inputs include hydrologically conditioned digital elevation models representing the topography excluding land- cover, while ensuring river connectivity; and roughness maps indicating the flow resistance of land-covers.
We are developing GeoFabrics, a GitHub-hosted open-source Python package, to provide an integrated reproducible workflow for generating these inputs. GeoFabrics takes LiDAR point clouds (e.g. LINZ datasets hosted on OpenTopography) and combines them with other geographical datasets (e.g. bathymetry depth contours hosted on the LINZ Data Service) to produce these maps. Due to the national scale and long-term focus of the programme, it is critical that this workflow can be run over any catchment as new LiDAR surveys become available. GitHub Actions trigger automated testing with each push to monitor the computational reproductivity of the workflow during its development to ensure the mature workflow will produce acceptably similar results independent of the operating environment.
We are also developing geoapis, a GitHub-hosted open-source Python library for enabling scientists to access publicly-available web-accessible geospatial data by encapsulating API calls and URL-wrangling within a clean modular interface.



ABOUT THE AUTHOR(S)

Rose Pearson is currently a postdoctoral fellow at NIWA and a visiting researcher at the Geospatial research institute. Her postdoctoral research focuses on combining geospatial data, primarily LiDAR point clouds, to produce hydrologically conditioned DEMs and roughness maps for use in river flood modelling. Her research interests centre on surface generation and attribute mapping from a wide array of spatial and geospatial dataset

RETURN TO PROGRAMME

Coupled Climate – Ice – Ocean Modelling: Setting up a framework

Alena Malyarenko[1,2], Alexandra Gossart[2]
[1] NIWA, Wellington
[2] National Modelling Hub, Antarctic Research Center, VUW
alena.malyarenko@niwa.co.nz

ABSTRACT / INTRODUCTION

Coupled Earth System Modelling is key to predicting climate change and future sea level rise. The Antarctic Science Platform is aiming at producing future projections with a fully coupled atmosphere – ocean – sea ice – ice sheet model. We have set up a configuration for the Ross Sea Region using three components: MITgcm (ocean – sea ice – ice sheet), WRF (atmosphere) and ESMF/NUOPC (coupler) on NESI and are conducting test runs. Our challenges included coordinating compilers, fortran versions, file management and precision of calculations. Our present work is focused on benchmarking different model configurations and obtaining estimates of resources needed to make future projections.

ABOUT THE AUTHOR(S)

Alena Malyarenko and Alexandra Gossart are Modelling Hub Postdoctoral fellows within the Antarctic Science Platform.

RETURN TO PROGRAMME

Janis: A Python Framework for Workflow Translation

Richard Lupat[1], Michael Franklin[1,2], Evan Thomas[3], Juny Kesumadewi[1,2], Jiaan Yu[1], Grace Hall[2], Mohammad Bhuyan[3], Tony Papenfuss[1,3], Andrew Lonie[4], Daniel Park[2], Bernard Pope[2], Jason Li[1]

1. Peter MacCallum Cancer Centre
2. The University of Melbourne
3. Walter and Eliza Hall Institute
4. Australian BioCommons Email:

richard.lupat@petermac.org

ABSTRACT / INTRODUCTION

There are many standards for building bioinformatics workflows, including Common Workflow Language (CWL), Workflow Description Language (WDL), Nextflow, and more. Each standard brings a community and a set of resources, including its execution engine, platform and other tools. The incompatibility of these standards poses challenges for portability, where changing between systems requires re-engineering efforts and is an inhibitor to sharing workflows. Janis is an open-source Python framework that addresses this interoperability problem by abstracting both the workflow and execution model to generate CWL, WDL, or Nextflow workflows. Janis simplifies many aspects of building workflows and can mask idiosyncrasies of the target specifications while still allowing for rich workflow logic to be represented. The ability to target multiple workflow specifications unlocks tools from their respective communities and mitigates the risks and effects of pipeline frameworks becoming unsupported. Janis also partially supports the ingestion of existing CWL and WDL workflows, and this feature will help in migrating popular workflows across different communities.

In this demo, we will discuss how Janis can provide a key piece of technology for workflow interoperability. We will demonstrate how abstraction can provide valuable benefits to workflow authors and the community.

ABOUT THE AUTHOR:

Richard Lupat is a senior bioinformatics software engineer at Peter MacCallum Cancer Centre (Melbourne, Australia). He is primarily working on developing bioinformatics pipelines and automating the analysis for next-generation sequencing data.

RETURN TO PROGRAMME

Building efficient workflows for the rapid review of time sensitive research

Adam Hyde
Coko (https://coko.foundation)
adam@coko.foundation

ABSTRACT / INTRODUCTION

This talk will focus on Adams real world experience designing and building (open source) collaboration systems for the rapid sharing and review of research (preprints and micropublications). The discussion will draw from system design of Kotahi (https://kotahi.community) and its' implementation for the Novel Coronavirus Research Compendium, eLife, and the Organisation for Human Brain Mapping.

This talk is suitable for both strategic and technical staff. Adam will focus on high level design architecture, benefits of Single Source Publishing Systems, the value of workflow concurrency in sharing research rapidly, and how the landscape of academic publishing is changing.

ABOUT THE AUTHOR

Adam is a NZ born serial social entrepreneur. Founder of Book Sprints, Coko Foundation, Kotahi, Editoria, Pagedjs and other open source technologies. For the last 17 years he has been building platforms and communities to accelerate the sharing of scholarly research. Most recently Adam has drawn satisfaction from designing and building systems that accelerate the review of novel coronavirus research (preprints). Adam is a Shuttleworth Fellow and has been awarded the NZ Open Source Award for Community Building.

RETURN TO PROGRAMME

Automated workflow aids biophysical model development – couple R targets and APSIMX

Jian Liu
Plant & Food Research
jian.liu@plantandfood.co.nz

ABSTRACT / INTRODUCTION

Parameterising Agricultural production simulators next generation (APSIMX) models are challenging. Conventional parameterisation often involved multiple ad-hoc processes to estimate parameter values. Thinking logic could be lost or lacking reproducibility in these manual exercises. The R package “target” is a pipeline toolkit that can orchestrate codes, files, various data sources, and more importantly, document the thinking logic. A case study was done in parameterising the APSIMX-Lucerne model via targets package in R. The case study utilised three main features of the target package including function-orient programming, caching and parallel computing. The function-orient programming provided an apparent pathway for package development. The caching feature allows users to do quality checking and only update objects that have modified dependencies. Lastly, the parallel computation feature reduced the computing time considerably. Moreover, the script-based workflow contributes to efficient version control and increase reproducibility. The implementation of the automated workflow in R yielded three outcomes. First, a customised R package was developed to derive soil water parameters for APSIMX models. Secondly, one could quantify time expense on parameterisation for APSIMX models. Thirdly, the case study demonstrated the possibilities of objective and reproducible parameterisation for APSIMX models.

ABOUT THE AUTHOR

Jian Liu works in Plant & Food Research as a data scientist. He is also doing his master thesis at Lincoln University as a part-time student in the area of crop modelling. Jian picked up data science skills through the working experience and grow his passion in the applications of data science on investigating the interaction between plants and environment

RETURN TO PROGRAMME

Translating Complex Topics: Software Containers

Paula Andrea Martinez[1], Sam Muirhead[2]
[1]Australian Research Data Commons, [2]CameraLibre
[1]paula.martinez@ardc.edu.au, [2]sam@cameralibre.cc

ABSTRACT / INTRODUCTION

A short video has the power to enable accessibility.

A complex topic can be translated for a learning audience using language and use cases that reflect them. The Australian Research Data Commons and Camera Libre worked together to develop an explainer video highlighting the benefits of using software containers in research.

Technical explanations can get bogged down in jargon and implementation details. By focusing on clear language and the needs of researchers, we were able to introduce the basics of software containers quickly and effectively.

We opted for an animation video because it captures the attention of viewers, and can fluidly link abstract concepts, human stories and tangible use cases. Using a consistent visual language allowed us to explore different aspects of software containers while continuously reinforcing the core concept.

Following the Mozilla Open Leaders directive “openness as the norm in innovation and research”, we worked in the open, and made the process and output publicly available via a CCby 4.0 Attribution licence.

We’re happy to share the lessons learned in developing an engaging, insightful and creative video! Watch it here: https://www.youtube.com/watch?v=HelrQnm3v4g and cite the DOI: https://doi.org/10.5281/zenodo.5091259.

ABOUT THE AUTHOR(S)

Paula Andrea Martinez is currently the Software Project Coordinator at ARDC. Her work supports the proposed Software Agenda in Australia leading to research software recognition as a first-class scholarly output of research. She is a former ARDC consultant whose work focused on development of National Training Materials in various selected topics from which the Software containers video is one output. She is also an advocate of Open Science, better research, diversity and community projects like the FAIR4RS, The Carpentries, ROpenSci and R-Ladies. She is also working with the Research Software Alliance as part time Community Manager.

Sam Muirhead is an animator, activist and technologist who makes complex topics accessible, promotes open collaboration and supports social and environmental justice activism. His creative work and digital tools are openly licensed — built upon the work of others, and then shared for others to take further. Sam was a 2018-19 Mozilla Fellow, working with Creative Commons.

RETURN TO PROGRAMME

Static Tactics

Using static website workshops to develop capability and collaboration

Matt Plummer
Victoria University of Wellington
matt.plummer@vuw.ac.nz

ABSTRACT / INTRODUCTION

There are a range of introductory workshops designed to teach digital research skills, notably those provided by the Carpentries. But although frequently pitched at novices, for some, particularly those from humanities and social science backgrounds, concepts taught in these workshops – such as version control with git – can be tricky to grasp. As a complement or precursor to Carpentries-style workshops, static website workshops in which attendees design and publish a git-based, academic portfolio website, provide an immediate, accessible and enjoyable experience. Through an incremental, scaffolded lesson structure, attendees learn about version control, git and GitHub/Lab, text editors, web protocols and styling (markdown, CSS and HTML), and configuration files (yml). This lightening talk will explain how these workshops develop capability for a wider range of academics to utilise tools such as git, and thus prime them for interdiscplinary collaboration and better utilisation of research computing resources.

ABOUT THE AUTHOR

Matt's background spans the arts and technology. In his current role as a Digital Research Consultant, he acts as a 'digital interpreter', working with researchers from different disciplines to utilise technology in innovative and transformative ways. He has assisted with the development of open source projects and research tools, coordinated numerous community-building and training events, and enjoys the opportunity to introduce researchers to new approaches and collaborators.

RETURN TO PROGRAMME

Communicating Epidemiological Insights using Explorable Explanations

Arindam Basu
University of Canterbury School of Health Sciences
arindam.basu@canterbury.ac.nz

ABSTRACT / INTRODUCTION

COVID19 pandemic has resulted in increased volume of interactive visualisations and data analyses. Epidemiologists and Public Health Data scientists need to present epidemiological data analyses in ways that enable individuals to interact with the data and work and simulate scenarios meaningful to them. This is challenging for Epidemiologists who are unfamiliar with programming skills to produce such visualisations. This learning curve can be flattened with explorable explanations that are web-based interactive data presentations but explorable explanations are not common among health scientists either as tools are not easy to identify.
The goal of this presentation is to demonstrate how Jupyter notebooks can be integrated in a web-based notebook to create explorable explanations, and embed these explanations in publicly available websites. The presenter will demonstrate this using real-time analyses of COVID19 statistics enabling users interact with data and visualisation. This can be extended to other tpes of epidemiological data analyses.

ABOUT THE AUTHOR

Arindam Basu is an Associate Professor of Epidemiology and Public Health Sciences at the School of Health Sciences, University of Canterbury.

RETURN TO PROGRAMME

Corrfunc: Blazing Fast Correlation Functions on the CPU

Dr. Manodeep Sinha, Dr. Lehman Garrison
Swinburne University of Technology,
ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D)
Center for Computational Astrophysics, Flatiron Institute
msinha@swin.edu.au
lgarrison@flatironinstitute.org

ABSTRACT / INTRODUCTION

How galaxies are distributed in space is determined by a combination of universal cosmological parameters, gravity, and the physics of galaxy formation. Quantifying galaxy clustering requires computing pair-wise separations -- an inherently quadratic process. Consequently, comparing the observed clustering of galaxies to that theoretically predicted is both useful to advance our understanding of physics and technically challenging. Here I present Corrfunc – an open-source suite of OpenMP-parallelized clustering codes that target current CPU micro-architecture with custom Advanced Vector Extensions (AVX512F, AVX) and Streaming SIMD Extensions (SSE) intrinsics. By design, Corrfunc is highly optimized and is at least a factor of few faster than all existing public galaxy clustering correlation function routines. While Corrfunc was developed primarily with astrophysical applications in mind, the basic algorithm within Corrfunc can be easily extended for applications that require looping over neighbours up to a certain maximum spatial extent. For instance, molecular dynamics simulations, game development for modelling flocking behaviour, any cross-matching between multiple datasets based on spatial separation, can potentially benefit from the Corrfunc algorithms. Corrfunc is covered by a suite of tests, extensive documentation and is publicly available at https://github.com/manodeep/Corrfunc.

ABOUT THE AUTHOR(S)

Dr. Manodeep Sinha is a computational astrophysicist based at the Centre for Astrophysics & Supercomputing at Swinburne University, Melbourne. Dr. Sinha completed his PhD in Astronomy from The Pennsylvania State University, and is currently a Senior Research Software Scientist, working with the ARC Centre of Excellence All-Sky Astrophysics in 3D (ASTRO 3D). Dr. Sinha works at the intersection of astrophysics, statistics, high-performance computing and software engineering. Dr. Sinha is a passionate advocate for sustainable research infrastructure and a champion for open and inclusive communities. Dr. Sinha is the founder and co-chair of the Research Software Engineers (Australia & New Zealand) community.

Dr. Lehman Garrison is a Flatiron Research Fellow at the Flatiron Institute's Center for Comptational Astrophysics in New York City. He specializes in high-performance simulations of the clustering of dark matter and galaxies, with emphasis on parallel programming and GPU computing. As co-chair of the Cosmological Simulations Working Group of the Dark Energy Spectroscopic Instrument (DESI) collaboration, Dr. Garrison combines simulations with galaxy survey observations to infer physical properties of the universe.

RETURN TO PROGRAMME

How to draw an owl – transitioning from tutorials to the real world with Dask

Maxime Rio[1,2], Wolfgang Hayek[1], Glen Reeve[1], Yalu Wen[3], Wendy Li[3]
[1] NIWA, [2] NeSI, [3] University of Auckland
maxime.rio@nesi.org.nz, wolfgang.hayek@niwa.co.nz

ABSTRACT / INTRODUCTION

The Dask library for parallel processing is a powerful tool for scaling up data-driven workflows and scientific computation in Python. Among its core strengths is its frontend-backend separation, allowing users to run their workflows on laptops, workstations, and high-performance computers (almost) unchanged. The frontend features multiple interfaces with varying complexity, supporting a wide range of applications. Dask has also been integrated into several well-known packages, such as scikit-learn and xarray.
In this talk, we will showcase practical applications where we used Dask to scale up data analyses with parallel processing. We will discuss some of the difficulties and pitfalls that users face when trying to apply examples from tutorials to the real world and highlight the many useful features that make Dask a powerful asset in a data science engineer's and research software engineer's toolbox.

ABOUT THE AUTHOR

Maxime Rio is a data science engineer and data scientist at NeSI and NIWA. He enjoys helping researchers to analyse their data, from visualisation to probabilistic modelling.
Wolfgang Hayek is a research software engineer at NIWA, and group manager of NIWA's scientific programming group, with many years of experience in scientific computing and HP

RETURN TO PROGRAMME

Keynote Speaker - September 16, 2021, 3:00 – 4:00

Māori Data Sovereignty: An introduction and implementation in the research sector

Dr. Karaitiana Tairuru
Christchurch Heart Institute, University of Otago
karaitiana@taiuru.maroi.nz

ABSTRACT / INTRODUCTION

Using legal and moral obligations to recognise and enact Māori Data Sovereignty rights and principles using legal instruments such as Te Tiriti/Treaty of Waitangi and United Nations Declaration of the Rights of Indigenous Peoples (UNDRIP), I discuss how Crown Research Institutes, universities, and other public sector organisations can decolonise their perspectives of data and ownership and understand and implement Māori Data Sovereignty principles and Māori Data Ethics as a platform.

Māori Data is a Taonga Māori will be explained using Te Ao Māori perspectives and examples and introducing six new Māori Data Sovereignty licences to protect Māori Data.

To conclude, a real-world example of how an Otago University institute is implementing Māori Data Sovereignty principles with both digital and bio data samples.

ABOUT THE AUTHOR

My career spans the past 28 years, originally in the ICT industry and with Māori property rights. I have been involved as an advocate and proponent for digital Māori rights, cultural appropriation, data sovereignty/digital colonialism, te reo Māori revitalisation with technology, Māori representation and Intellectual Property Rights for the past 24 years. In more recent times raising tikanga Māori and mātauranga Māori awareness issues and academic research focusing on Māori cultural rights with gene research.

I am well versed in Māori culture, tikanga, te ao Māori and how those rights and beliefs are applied to the digital and biological sciences.

Iwi affiliations: Ngāi Tahu (Koukourarata, Puketeraki, Rāpaki, Taumutu, Tūāhuriri, Waewae, Waihao, Waihopai, Wairewa); Ngāti Rārua; Ngāti Kahungunu (Ngāti Pāhauwera); Ngāti Hikairo (Ngāti Taiuru); Tūwharetoa (Tamakopiri); Ngāti Hauiti (Ngāti Haukaha); Ngāti Whitikaupeka; Pākehā.

RETURN TO PROGRAMME

Metrics for Measuring the Contributions of Research Software Engineers

Ms Nooriyah Lohani[1], Mr Justin Baker[2], Dr Manodeep Sinha[3,4],Mr Nick May, Dr Rebecca Lange[5], Ms Heidi Perrett[6]

1New Zealand eScience Infrastructure, New Zealand, 2CSIRO, Melbourne, Australia, 3Centre for Astrophysics & Supercomputing, Swinburne University, Melbourne, Australia, 4ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), 5Curtin University, Bentley, Australia, 6CERES, Brisbane, Australia

nooriyah.lohani@nesi.org.nz, justin.baker@csiro.au, msinha@swin.edu.au, nicholasmay2@gmail.com, rebecca.lange@curtin.edu.au, heidi.perrett@cerestag.com

ABSTRACT / INTRODUCTION

"A Research Software Engineer (RSE) combines professional software engineering expertise with an intimate understanding of research." (https://society-rse.org/about/)

There is an increasing awareness of the value RSEs provide to the research community. RSE communities themselves are growing rapidly both within the New Zealand research sector and internationally. Many research-based organisations are working towards having the RSE role recognised and defined as a career path in its own right. However, the multifaceted nature of the role makes it a complicated task to capture the various kinds of RSE contributions.

To be able to recognise the contributions of RSE's, this BoF will raise questions and encourage discussions such as: Does it make sense to attach metrics to a complex role like the RSE, and if so, what should such metrics encompass? Consider the following list, which highlights a subset of the diverse range of skills, services, and capabilities common to many RSEs:

- Science domain/computational knowledge
- Training and outreach
- Best practice software development and methodologies
- Mentoring
- Adherence to software and data standards
- Knowledge about domain-specific as well as general-purpose software tools
- Cybersecurity
- Data analytics and visualisation
- Publication support for supplementary software/data

In this BoF, moderated by members of the RSE-AUNZ Steering Committee, we would like to better understand and quantify the contributions of RSEs. We encourage all members of the RSE community in New Zealand, Australia and other interested stakeholders to attend and help identify the various facets of RSE contributions and associated suitable metrics.

ABOUT THE AUTHOR(S)

Nooriyah Lohani has a background in Genetics and Computer Science with 8 years experience working in commercial R&D and academic research in New Zealand. Nooriyah is now part of the engagement and communications team at NeSI alongside pursuing a PhD in Data Science. Nooriyah is passionate about recognising the contributions of researchers/research software engineers bringing

RETURN TO PROGRAMME

together research and software skills, which she actively works towards as co-chair of the Research Software Engineers - Australia and New Zealand community.

Dr. Manodeep Sinha is a computational astrophysicist based at the Centre for Astrophysics & Supercomputing at Swinburne University, Melbourne. Dr. Sinha completed his PhD in Astronomy from The Pennsylvania State University, and is currently a Senior Research Software Scientist, working with the ARC Centre of Excellence All-Sky Astrophysics in 3D (ASTRO 3D). Dr. Sinha works at the intersection of astrophysics, statistics, high-performance computing and software engineering. Dr. Sinha is a passionate advocate for sustainable research infrastructure and a champion for open and inclusive communities. Dr. Sinha is the founder and co-chair of the Research Software Engineers (Australia & New Zealand) community.

RETURN TO PROGRAMME

Opportunities and challenges for ML-Ops for NZ Research Institutes

Linley Jesson1, Alan Tan2, Richard Dean3
1Plant and Food Research, 2NIWA, 3ESR

ABSTRACT / INTRODUCTION

Productionising Data Science solutions can result in important impact for CRIs and universities. Key challenges faced commonly are how researchers can push models and research into production easily and maintaining the ability to update models deployed in production environment. The means by which this production can happen is varied, and a huge range of solutions exist. As a result, it can be time consuming and challenging to identify a suitable solution as there is no one-sized fits all solution. There has been a recent suggestion that researchers can contribute to writing Software Carpentry or other training resources to help new researchers understand the key steps involved and to speed up-skilling.

In this birds-of-a-feather session we will share people's experiences with learning how to build data science pipelines, identify key needs for training, and other ways that New Zealand research institutes can work together to provide greater impact for science in Aotearoa/New Zealand.

RETURN TO PROGRAMME

Accelerating data wrangling: A researcher-built dataset assembly tool

Simon Anastasiadis
Social Wellbeing Agency
Simon.Anastasiadis@swa.govt.nz

ABSTRACT / INTRODUCTION

Good analytic and research projects combine information in ways that lead to new insights and actions. The preparation for such projects often involves drawing data together into a single dataset ready for analysis. However, as the number of data sources increases so does the complexity of preparation. Without a consistent method for assembling analysis-ready datasets, this process can become time-consuming, expensive, and error prone.

In response to this challenge, the Social Wellbeing Agency has developed the Dataset Assembly Tool. By standardising and automating the data preparation and dataset assembly stages of analytic projects, the tool helps staff deliver higher quality work faster. We have already found the use of this tool significant, more than halving the time spent in preparation and wrangling.

The Dataset Assembly Tool is now available for other researchers and analysts to use. While the tool can be used within a single project, greater impact is anticipated from its reuse across projects and organisations. As more researchers use the tool and its patterns, the increased consistency will strengthen the research community by making it easier to collaborate, sharing knowledge and code.

ABOUT THE AUTHOR

Simon Anastasiadis is a data scientist, computational problem solver, and Integrated Data Infrastructure (IDI) researcher. He seeks to enable faster delivery by delivering research in a way that maximise the reuse of project components.

RETURN TO PROGRAMME

Data Version Control & Collaboration for Reproducible Science with Kart

Hamish Campbell
Koordinates
hamish.campbell@koordinates.com

ABSTRACT / INTRODUCTION

Kart (kartproject.org) is an open source tool that allows you to quickly and easily manage history, branches, data schemas, and synchronisation for large & small datasets – especially geospatial data – between different working copy formats and operating systems.
Modern open source software has given us excellent code version control and reproducible build environments. However data remains difficult to handle, changes are badly tracked if at all, and data state is difficult to verify.
In this demonstration Koordinates Product Manager Hamish Campbell will show how Kart provides this critical missing link in the reproducible-science toolchain. He will cover how Kart can import and manage large datasets, provide detailed & verifiable change control, handle merge conflicts and more.

ABOUT THE AUTHOR

Hamish Campbell is a software product manager with a background in civil engineering and geospatial applications. He cares about open data and building great software that is good for society and the plan

RETURN TO PROGRAMME

Automating instrument data workflows: Integrating Globus into MyTardis

Mike Laverick
Centre for eResearch, University of Auckland
mike.laverick@auckland.ac.nz

ABSTRACT / INTRODUCTION

Research instruments are producing larger and larger volumes of data that often require substantial processing before being science-ready. This means it is becoming ever more important to automate data transfer and processing workflows, while simultaneously documenting metadata pertaining to instrument/experiment configurations and pre-/post- processing steps.

This talk discusses our recent efforts in integrating Globus into the University of Auckland's deployment of MyTardis, facilitating instrument data transfer between our centralised instrument data store and personal/managed end-points such as New Zealand eScience Infrastructure (NeSI).

We explore some of the technical challenges involved in coupling these two technology solutions, and discuss our future development plans for MyTardis, Globus, and FuncX to help reduce the ever-growing research burden of transferring and processing instrument data.

ABOUT THE AUTHOR

Mike Laverick is a former atomic astrophysicist turned eResearch Solutions Specialist at the University of Auckland. Hailing from the UK, he completed his PhD at KU Leuven, Belgium, before moving to New Zealand and joining the Centre for eResearch. Mike is currently working on the development of the university's new instrumentation data platform, and the Space Payload Operation Centre. Mike is also a keen advocate of all things Python-related, contributing to digital research skills training and community events.
https://orcid.org/0000-0002-9220-2982

RETURN TO PROGRAMME

Using Julia in a High Performance Computing Environment

Matija Čufar[1,2], Mingrui Yang[1,2,3], Elke Pahl[3,4], Joachim Brand[1,2]

[1]New Zealand Instituite for Advanced Study and Centre for Theoretical Chemistry and Physics, Massey University

[2]Dodd-Walls Centre for Photonic and Quantum Technologies

[3]MacDiarmid Institute for Advanced Materials and Nanotechnology

[4]Department of Physics, University of Auckland

matijacufar@gmail.com

ABSTRACT

High-performance computing (HPC) software is almost exclusively written in low-level programming languages such as Fortran, C, or C++. Julia is a relatively new language that promises the flexibility and usability of high-level languages, without sacrificing any performance.
When starting a quantum Monte-Carlo project, we were presented with a choice – extend an existing Fortran code, or start from scratch. We picked the latter and decided to do it in Julia.
In this presentation, we will present our software Rimu.jl[1] and share our experiences in using a high-level programming language in an HPC environment. We will lay out the pros and cons of our approach and perhaps help you decide whether Julia is a good fit for your next project.

ABOUT THE AUTHORS

Matija Čufar is a research technician at Massey University. He is working on Rimu.jl, a quantum Monte Carlo package. Prior to this, he developed the persistent homology package Ripserer.jl[2].

Mingrui Yang is a PhD student from Massey University with a background in computational physics and chemistry.

Dr Elke Pahl is a senior lecturer from the University of Auckland. Her research spans topics in computational physics and chemistry, and solid-state physics with an emphasis on computational modelling.

Prof. Joachim Brand is a theoretical physicist from Massey University. His research expertise is in ultracold atomic gasses.

1 https://github.com/joachimbrand/Rimu.jl

2 https://github.com/mtsch/Ripserer.jl

RETURN TO PROGRAMME

Accelerating a physics code with OpenACC

Alexander Pletzer[1,2], Chris Scott[1,3] and Gilles Bellon[3]
[1]New Zealand eScience Infrastructure (NeSI), [2]NIWA, [3]University of Auckland
Alexander.Pletzer@nesi.org.nz

ABSTRACT / INTRODUCTION

Graphical Processing Units (GPUs) offer the prospect of significantly accelerating scientific software. How much human labour is required and what kind of speedup can be realistically achieved is a question which we will try to address. Our testbed is a quasi-equilibrium tropical circulation model developed by researcher Gilles Bellon, which is written in Fortran. Here we reveal the code changes that were required to run on a GPU. As part of a NeSI consultancy, we applied a mix of OpenACC directives and calls to CUDA libraries, which instruct the compiler to offload data and computation to the GPU. In order to achieve good performance on the GPU, we had to pay particular attention to data locality, minimising the amount of data copies to and from the GPU. At the end of the consultancy, a 60X performance improvement was achieved compared to a single threaded execution of the code on CPU.

ABOUT THE AUTHOR(S)

Alexander and Chris a research software engineers for NeSI, working with researchers to run better and faster on NeSI computing platforms. Gilles Bellon is senior lecturer in the faculty of science at the University of Auckland. His interests are in better understanding the interaction between tropical clouds and tropical atmospheric circulation and their impact on the observed climate variability.

RETURN TO PROGRAMME

Distributed Computing Strategies for Training Deep Networks on High Performance Computing Server

Vincent Wang
Massey University
zwan076@gmail.com

ABSTRACT / INTRODUCTION

The professional project of my master degree in Information Sciences (Computer Science) is regarding speech dataset augmentation to improve the performance of speech recognition deep learning networks. We have proposed a new speech data augmentation approach in speech data. For a fair comparison and suitable for publication, we used the same setup and configurations for training deep networks with and without our new data augmentation approach. This requires large amount of GPU computational resource to implement distributed computing strategies to speed up experiments, as the deep networks size are quite large for training. Meanwhile, we have been offered chances for trial using High Performance Computing (HPC) servers with Lambda Stack (NeSI's HPC) and DGX Stack (Waikato University's HPC) to implement experiments of our project. In the conference, I would like to make 5 minutes lightening talk to present the practical experience we have gained for the distributed computing optimisation strategies from our research project, which includes discussion of low level HPC server architecture, user interfaces and runtime containers for different platforms, code implementations with Pytorch framework for achieving distributed training of deep networks, and communication overheads among GPUs.

ABOUT THE AUTHOR

Vincent Wang is a Master Student in Information Sciences (Computer Science) , Massey University

RETURN TO PROGRAMME

Keynote Speaker - September 17, 2021, 1:00 – 2:00

AI and the opportunity for NZ to lead rather than follow

Jannat Maqbool
University of Waikato's Artificial Intelligence Institute

ABSTRACT / INTRODUCTION

New Zealand is internationally recognised in the area of machine learning and in fact for more than a decade and even now people are developing skills in machine learning using software developed here. The timing is right for us to now leverage this existing reputation and strength together with our uniqueness as a nation to lead rather than following the rest of the world in leveraging AI to benefit our people, the environment and economy. In this talk we will hear about the work underway to connect AI researchers, provide a voice for the research ecosystem and share knowledge across the AI ecosystem and wider to encourage greater awareness, the building of talent and capability, and to drive collaboration in positioning New Zealand and New Zealanders as frontrunners when it comes to the potential of AI.

ABOUT THE AUTHOR

Jannat was previously the Waikato tech sector lead at Te Waka and Smart Cities Advisor at Hamilton City Council, and is currently the Associate Director at the University of Waikato's Artificial Intelligence Institute. Jannat is also a Trustee at Web Access Waikato Trust, on the Executive Council at NZ IoT Alliance and TechWomen NZ, a board member at NZ Tech, and Director – NZ at Smart Cities Council ANZ.

RETURN TO PROGRAMME

A National Agenda for Research Software in Australia

Tom Honeyman, Paula Andrea Martinez
Australian Research Data Commons, Australian Research Data Commons
tom.honeyman@ardc.edu.au, paula.martinez@ardc.edu.au

ABSTRACT / INTRODUCTION

The Australian Research Data Commons sees research software as a critical but underappreciated piece of national infrastructure. Building on this, we have recently produced an agenda mapping out a pathway to recognition of research software as a first-class output of research.

The “National Agenda for Research Software” (https://bit.ly/rs-agenda) presents a set of 12 actions to achieve recognition. These 12 actions are grouped under three high level aims to “See, Shape and Sustain Research Software” which consider software availability and visibility, the suitable application of software engineering best practice and the maintenance of critical research software infrastructure respectively. For each of these three aims there are corresponding actions considering the necessary infrastructure, guidance, communities and advocacy needed to achieve it.

The agenda also characterises the necessary stakeholders in this change and their relative interest in the various actions of the agenda. In this talk, the agenda framework, the results of a validation of the actions, measured relative interest of stakeholders in these actions, and sets of priorities and gaps will be presented. All this combined informs a path forward for Australia to recognise research software as a first class output of research.

ABOUT THE AUTHOR(S)

Tom Honeyman is the manager of the Software Program at the Australian Research Data Commons. In this role he oversees a set of activities to work towards recognition of research software as a first class output of research. His research background is in descriptive linguistics and he has previously worked as a research software engineer, mostly in linguistics and anthropology, and has also worked in various research data roles.

Paula Andrea Martinez is currently the Software Project Coordinator at ARDC. Her work supports the proposed Software Agenda in Australia leading to research software recognition as a first-class scholarly output of research. She is also an advocate of Open Science, better research, diversity and community projects like the FAIR4RS, The Carpentries, ROpenSci and R-Ladies. She is also working with the Research Software Alliance as part time Community Manager.

RETURN TO PROGRAMME

Supporting FAIR4RS Adoption Guidelines

Paula Andrea Martinez[1,2], Daniel S. Katz[3], Michelle Barker[2], Carlos Martinez[4], Tom Honeyman[1], Georgina Rae[5]

[1]Australian Research Data Commons, [2]Research Software Alliance, [3]NCSA & CS & ECE & iSchool, University of Illinois, [4]The Netherlands eScience Center, [5] New Zealand eScience Infrastructure

paula.martinez@ardc.edu.au, dskatz@illinois.edu, barkermd@outlook.com, c.martinez@esciencecenter.nl, tom.honeyman@ardc.edu.au, georgina.rae@nesi.org.nz

ABSTRACT / INTRODUCTION

Research software is a significant and vital component of research. The FAIR for Research Software Working Group (FAIR4RS WG) is a global community that advocates for and raises awareness of FAIR (Findable, Accessible, Interoperable, Reusable) research software through regional and international collaboration.

The FAIR4RS Principles recently developed by the FAIR4RS WG are relevant to the larger ecosystem that supports research software. As RSEs and advocates for RSEs, we encourage global involvement in developing a clear and measurable path to progressively improve the FAIRness of research software, leading to better research practices, such as replicability, reproducibility, and increased recognition and transparency.

This session will outline how the RSE AUNZ community can participate in co-producing guidelines for implementation of the FAIR4RS Principles, to help encourage and streamline adoption. The emergence of community implementation strategies will assist in demonstrating the value placed upon adoption of the FAIR4RS Principles and encourage policy makers to also embrace implementation.

This work is valuable to RSEs, researchers, the scientific community, organisations that create, modify, manage, share, protect, fund and preserve research software, and others with an interest in the FAIR4RS Principles.

ABOUT THE AUTHOR(S)

Dr. Paula Andrea Martinez is currently the Software Project Coordinator at ARDC. Her work supports the proposed Software Agenda in Australia leading to research software recognition as a first-class scholarly output of research. She is a Co-Chair of the FAIR4RS WG, also an advocate of Open Science, better research, diversity and community projects like the FAIR4RS, The Carpentries, ROpenSci and R-Ladies. She is also working with the Research Software Alliance (ReSA) as part-time Community Manager.

Dr. Daniel S. Katz is Chief Scientist at the National Center for Supercomputing Applications (NCSA), Research Associate Professor in Computer Science, Electrical and Computer Engineering (ECE), and the School of Information Sciences (iSchool) at the University of Illinois Urbana-Champaign. He is also the Steering Committee Chair of the Research Software Alliance (ReSA), co-chair of the FAIR4RS WG and the FORCE11 Software Citation Implementation Working Group, and an Associate-Editor-in-Chief of the Journal of Open Source Software (JOSS).

RETURN TO PROGRAMME

Dr. Michelle Barker has extensive expertise in open science, digital research infrastructure and digital workforce capability. As a sociologist, Michelle is passionate about building collaborative partnerships to achieve system change. She works as both an Open Science Consultant, and Director of the Research Software Alliance (ReSA), which brings together software communities together to collaborate to increase recognition of research software.

Dr. Carlos Martinez-Ortiz is Scientific Community Manager at the Netherlands eScience center. Previously he worked on several projects as Research Software Engineer. His current focus is in Software Sustainability and is a co-chair of the FAIR4RS WG.

Dr. Tom Honeyman is the manager of the Software Program at the ARDC, where he is working on a program of activities to seek recognition of research software as a first-class output of research. As part of that work the ARDC has recently released a National Agenda for Research Software, which Tom was a co-author on. The Agenda defines a set of 12 actions which combined represent a call to “See, Shape and Sustain Research Software”.

Georgina Rae is the Science Engagement Manager at NeSI where she ensures that NeSI is building strong relationships with the research sector. Prior to NeSI she has worked in molecular biology and intellectual property. She is passionate about enabling research and is interested in the fundamental shifts required to level up scientific research.

RETURN TO PROGRAMME

Co-opting academic recognition systems for research software

Matthias Liffers, Tom Honeyman, Paula Andrea Martinez,

Australian Research Data Commons

matthias.liffers@ardc.edu.au, tom.honeyman@ardc.edu.au, paula.martinez@ardc.edu.au[

ABSTRACT / INTRODUCTION

The Australian Research Data Commons (ARDC) is a transformational initiative that enables Australian research community and industry access to nationally significant, leading edge data intensive eInfrastructure, platforms, skills and collections of high-quality data.

It is not common for researchers to be recognised for the time they put into developing research software. By subverting the most common method of academic recognition - citations - research software developers can hook into citation metrics infrastructure.

The ARDC has developed a range of openly-licensed resources that can be reused and remixed by researchers and research support professionals to raise broad awareness of software citation and normalise the practice.

This presentation will introduce the resources and the motivation behind developing them.

ABOUT THE AUTHOR(S)

Paula Andrea Martinez is currently the Software Project Coordinator at ARDC. Her work supports the proposed Software Agenda in Australia leading to research software recognition as a first-class scholarly output of research. She is a former ARDC consultant whose work focused on development of National Training Materials in various selected topics. She is also an advocate of Open Science, better research, diversity and community projects like the FAIR4RS, The Carpentries, ROpenSci and R-Ladies. She is also working with the Research Software Alliance as part time Community Manager.

Matthias Liffers (he/him) is the Research Software Skills Specialist at the Australian Research Data Commons. With a background in computer science and informatics, he believes that software should be considered a first-class scholarly output along journal papers and datasets.

Tom Honeyman is the manager of the Software Program at the Australian Research Data Commons. In this role he oversees a set of activities to work towards recognition of research software as a first class output of research. His research background is in descriptive linguistics and he has previously worked as a research software engineer, mostly in linguistics and anthropology, and has also worked in various research data roles.

RETURN TO PROGRAMME

Building an MLOps Pipeline for Lumi Drug Scan – A Real-time Drug Screening System

Deepak Karunakaran, Janet Stacey
Institute of Environmental Science and Research
deepak.karunakaran@esr.cri.nz, janet.stacey@esr.cri.nz

ABSTRACT / INTRODUCTION

There has been an exponential growth in the application of data science in industry and commerce, bringing in the need to adopt software development standards into machine learning (ML) projects. Unlike a typical software development project which is code-based and has a well understood lifecycle of coding, testing and deployment, ML projects are data-centric and therefore inherently uncertain. To deal with the challenges in productionizing machine learning, machine learning operations (MLOps) is emerging as a discipline for managing ML project lifecycle including data management, model training and evaluation, model deployment and its maintenance. Every ML project comes with its unique characteristics. Being an evolving discipline, MLOps presents multiple perspectives and ways to manage them. Lumi drug scan is a drug screening system developed together by ESR and NZ Police which use ML models to classify drugs in real-time using a portable device on-field. In this talk, we present our experiences in building an MLOps pipeline and developing a unique process maturity model while achieving the goal of moving Lumi from the proof-of-concept to production. We discuss some of the best practices, choice of algorithms, challenges faced and the learnings from this work, which is currently under progress.

ABOUT THE AUTHOR(S)

Deepak Karunakaran is currently working as a post-doctoral researcher at the Institute of Environmental Science and Research in the Forensics R&D group. He completed his PhD in computer science at Victoria University of Wellington. Before migrating to Aotearoa, he worked as a scientist at ABB India Research, a control systems and robotics company, in Bangalore, India. He is passionate about machine learning and artificial intelligence and his current focus is on tackling challenging problems in forensics. While keen on developing new machine learning algorithms, he is also invested in the emerging discipline of Machine Learning Operations (MLOps).

Janet Stacey is a Digital Sciences Engineer and trained bioinformatician at the Institute of Environmental Science and Research (ESR). Janet holds a Master of Science in Forensic Science and a Master of Science in Bioinformatics from the University of Auckland. She has 14 years' experience working in forensic casework and now works to move the industry forward using data science and next generation sequencing. Janet is highly involved in creating automated workflows, machine learning, training of staff and students through initiatives such as the data accelerator program and promoting collaboration efforts within the NZ data science community.

RETURN TO PROGRAMME

Experiences developing an operational workflow for large-scale instance and semantic segmentation of remote sensing imagery using CNNs

Jan Schindler[1], Brent Martin[2], Alexander Amies[2], Ben Jolly[3] and David Pairman[2]

1 Manaaki Whenua – Landcare Research, Informatics, Wellington 6011, New Zealand
2 Manaaki Whenua – Landcare Research, Informatics, Lincoln 7608, New Zealand
3 Manaaki Whenua – Landcare Research, Informatics, Palmerston North 4472, New Zealand
schindlerj@landcareresearch.co.nz

ABSTRACT / INTRODUCTION

Convolutional Neural Network (CNN) architectures offer unprecedented opportunities for improved mapping of land cover, change and individual features in satellite and aerial imagery compared to traditional machine learning algorithms (e.g., Random Forests). We have utilized state-of-the art CNN architectures for instance segmentation of objects using MaskRCNN and semantic segmentation of land cover using ResUNet and have developed an operational workflow for mapping exercises at the local-, regional- and national scale.

While a plethora of open-source Deep Learning frameworks exist, they usually target non-spatial imagery, small case studies or reference datasets and cannot be used on specific data archives and HPC infrastructures without major rework.

We have developed a fit-for-purpose reusable software pipeline that both runs on the NeSI HPC and on local PCs with GPU support. It allows for flexible training from geospatial layers and large volumes of remote sensing imagery stored in the efficient KEA-file format, keeps inode usage at a minimum, and includes cross-validation visualization and tile-based predictions routines.

The pipeline evolved while researching these techniques for a range of environmental applications, including mapping of urban trees and forests, forest destocking, landslides and historic urban green spaces. We discuss our experience with taking this agile approach.

ABOUT THE AUTHOR(S)

Jan Schindler is a remote sensing & data scientist at Manaaki Whenua – Landcare Research. He investigates natural features, processes, and human impacts on the Earth's surface and in the atmosphere by working across disciplines. He completed his doctoral studies at the Max Planck Institute for Chemistry on the retrieval of air pollutants (NO_2, HCHO) using Differential Optical Absorption Spectroscopy and has expertise in optical, Radar and LiDAR remote sensing from ground- and air- and space-based sensors. He employs data science techniques including ML/DL for spatial pattern analysis and environmental modelling and solves scale and integration challenges of disparate environmental datasets. His work supports data-led land cover mapping for New Zealand, novel methods for 2- & 3D tree detection and characterization of indigenous forests and scaling these on HPC facilities.

RETURN TO PROGRAMME

Brent Martin is a senior data scientist and machine learning specialist at Manaaki Whenua Landcare Research. His career has spanned both academic research as a senior lecturer at Canterbury University, as well as software engineering and R&D roles in various commercial companies. Brent's research in AI and machine learning includes developing new ML classification algorithms; applying ML to real-world problems such as electricity demand forecasting; research and development in Intelligent Tutoring Systems; developing social network analysis techniques for criminal investigation, and applying deep learning to environmental problems. Brent holds a PhD in Computer Science from the University of Canterbury, New Zealand focussing on artificial intelligence in education.

Alexander Amier is a data scientist at Manaaki Whenua – Landcare Research, specialising in machine learning applications to remote sensing problems. He has worked on numerous land cover identification and classification problems, including winter forage mapping, bare ground identification, and pasture productivity modelling. He has also employed temporal deep learning techniques for national-scale detection of exotic forest felling. He holds a Ph.D. from the University of Canterbury in Mechanical Engineering, for which he developed a novel radar-based structural health monitoring device for determining building deformation during earthquakes.

RETURN TO PROGRAMME