ResBaz 2020 - the lollie-scramble.
What can we learn from the experience?

Laura Armstrong, Yvette Wharton, Matt Plummer

Centre for eResearch, University of Auckland l.armstrong@auckland.ac.nz
Centre for eResearch, University of Auckland y.wharton@auckland.ac.nz
Centre for Academic Development, Victoria University of Wellington matt.plummer@vuw.ac.nz

ABSTRACT / INTRODUCTION

The *Research Bazaar* is a worldwide festival promoting the digital literacy emerging at the centre of modern research. We and our research community LOVE the usual 3-day cohort event full of digital skills workshops, social activities and amazing food. However, we had to adapt the event in 2020 (and forgo the nachos)!

In responding to the uncertainty of lockdowns, the push to deliver/attend online, increasingly financially challenging environments, and our community needs, we joined forces to deliver ResBaz 2020: Pick n Mix (https://resbaz.auckland.ac.nz). Initially coming together over shared dates, the event went on to offer 35 online sessions with hosts and speakers from the University of Auckland, Victoria University of Wellington, University of Canterbury, NeSI, ESR, National Library of New Zealand, University of Leipzig, Academic Consulting and The New Media Studio.

Laura, Yvette and Matt will share how *ResBaz 2020: Pick n Mix* came to be, what worked and what didn't, and what we can learn from the experience. We'll explore the following questions:

- *Can we deliver ResBaz online as the new normal?*
- *What is lost from the lack of a cohort experience, in-person attendance, serendipitous interactions over yummy food?*
- *What is gained from the wider NZ researcher community coming together?*
- *What can be shared for others to adapt/adopt?*
- *Is their space for different types of ResBaz experiences - in-person and online, local and national, generalist and discipline/community specific?*

Join us for a short presentation and discussion on these and your own questions.

AUDIENCE

This will be of interest to those designing and delivering researcher skills and development.

ABOUT THE AUTHOR(S)

Laura Armstrong is a Senior eResearch Engagement Specialist at the Centre for eResearch, University of Auckland working to engage researchers in eresearch, and deliver research data management services and researcher enablement projects. https://orcid.org/0000-0003-2370-3924

Yvette Wharton is the eResearch Solutions Lead at the Centre for eResearch, University of Auckland, working on research data management services and researcher enablement projects. She has extensive experience in University teaching, research and IT environments and is passionate about using her broad knowledge to facilitate people to achieve their aspirations.http://orcid.org/0000-0002-6689-8840

Matt Plummer is a Digital Research Consultant in Victoria University of Wellington’s Centre of Academic Development. In this role he utilises a background that spans the arts and technology to act as a ‘digital interpreter’, working with researchers from different disciplines to utilise technology in innovative and transformative ways. https://orcid.org/0000-0002-2737-2707

Towards an Institution-Wide Research Data Management Framework

Laura Armstrong, Brian Flaherty, Mark Gahegan

Centre for eResearch, University of Auckland l.armstrong@auckland.ac.nz
New Zealand eScience Infrastructure (NeSI) b.flaherty@auckland.ac.nz
Centre for eResearch, University of Auckland m.gahegan@auckland.ac.nz

ABSTRACT / INTRODUCTION

Research Data Management (RDM) is increasingly recognised as a critical knowledge gap for researchers as international and domestic funders, publishers, and ethics committees introduce more stringent requirements regarding Data Management Plans and the collection, storage and sharing of research data.

The University of Auckland is responding to the evolving research landscape with a strategic initiative to develop an integrated Research Data Management framework that is consistent with international standards, including FAIR data principles to improve data sharing whilst also adding in the principles of Māori Data Sovereignty. The project is sponsored by our Deputy Vice-Chancellor (Research) and includes an Advisory Panel and a Māori Data Sovereignty Kāhui.

Members of the project team will provide details of the work to:

- engage researchers, including socialising FAIR, CARE and Māori Data Sovereignty principles, through a roadshow, online survey, and interviews,
- identify challenges and opportunities presented by research data across the various different faculties of the university,
- develop and use a model to establish current and desired RDM maturity, and
- align with funders, code of conduct policies and ethics requirements.

AUDIENCE

This will be of interest to those interested in developing services and increasing maturity in the management of research data.

ABOUT THE AUTHOR(S)

Laura Armstrong is a Senior eResearch Engagement Specialist at the Centre for eResearch, University of Auckland working to engage researchers in eresearch, and deliver research data management services and researcher enablement projects. https://orcid.org/0000-0003-2370-3924

Brian Flaherty is Product Manager for Data Services at NeSI, including repository development, data transfer and storage. He has a background in digital libraries.
https://orcid.org/0000-0002-9278-146X

Professor Mark Gahegan is the director of the Centre for eResearch at the University of Auckland, and a professor in Computer Science.

FAIR for Research Software

Michelle Barker1, Georgina Rae2, Maxime Rio2,3 & Nooriyah Lohani2 1

Research Software Alliance (ReSA), 2New Zealand eScience Infrastructure, 3NIWA

michelle@researchsoft.org , georgina.rae@nesi.org.nz , maxime.rio@nesi.org.nz , nooriyah.lohani@nesi.org.nz

ABSTRACT / INTRODUCTION

We are seeing a significant movement to ensure research software is recognised as a fundamental and vital component of research.

This BoF aims to utilise audience expertise to identify examples of best practice in developing FAIR/sustainable software for research, to advance community knowledge and networks.

The FAIR For Research Software Working Group (FAIR4RS WG) is leading the research software community in the crucial step of agreeing how to apply the FAIR principles to research software by mid-2021. This BoF will engage the audience by identifying local examples of best practice in creating FAIR software and utilising the Lamprecht et al paper to provide guidance on what the FAIR principles might include. The best practice examples can be used to promote the FAIR principles for research software when finalised.

This BoF brings together 3 organisations:

- RSE-AUNZ in sharing the best research software practices across the community
- Research Software Alliance (ReSA)'s co-convening of the FAIR4RS WG with Research Data Alliance (RDA) and FORCE11 to create international community-agreed standards and guidelines
- NeSI in working towards research software as a first class research output

Why attend?

- Engagement with work developing the FAIR principles for research software
- Sharing of best practice to enable knowledge transfer across projects
- Development of networks across the RSE community
- Contribution to a set of impact stories from the community demonstrating best practice for sustainable sesearch software.
- Increased awareness of the activities of RSE-AUNZ, ReSA and NeSI

Format:

- 3 x short talks from each organisation
- A topic identification activity to guide breaking into smaller groups
- Breakout into groups

A report back from each group

ABOUT THE AUTHOR(S)

Michelle Barker

Dr Michelle Barker is the Director of the Research Software Alliance (ReSA). She has extensive expertise in open science, research software, digital workforce capability and digital research infrastructure. As a sociologist, Michelle is passionate about building collaborative partnerships to achieve system change. She recently chaired the OECD Global Science Forum expert group on digital skills for the research sector, and is a former Director of the Australian Research Data Commons, where she led the national research software infrastructure investment program.

Georgina Rae

Georgina 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.

Maxime Rio Maxime

Rio is a data science engineer at NeSI and a data scientist at NIWA. Over the last few years, he has helped scientists by developing probabilistic models, adapting machine learning tools for their needs, scaling imaging processing pipelines for large datasets and providing training. He easily gets excited by scientists' research and wants to help them to get the most out of their data.

Nooriyah Lohani

Nooriyah is a Bioinformatician by training and after working for a few years in a commercial and academic realm, is now a research communities advisor at NeSI passionate about understanding research needs in the eScience sector. She is also Co-chair of the RSE Australia New Zealand steering committee.

eResearch@QUT: Reflections on the implementation of an institution-wide strategy

Matthew Bellgard
Queensland University of Technology
matthew.bellgard@qut.edu.au

ABSTRACT

Background/Context

eResearch is the term given to the role of digital transformation in disciplinary, multi-, inter- and trans- disciplinary research where every research question can be framed in the context of eResearch touch points at different stages: digital platforms; data management, open data, curation, governance, contracts, access, and privacy; Internet of Things; artificial intelligence; decision support; bioinformatics; security; visualisation; algorithms, analytics; network latency challenges; high performance computing and cloud computing. Critical to the quantifiable success of eResearch is ongoing close engagement with end-users to deliver solutions, rapidly leveraging expertise, and repurpose solutions, continual improvements in work practices based upon feedback, as well as obtaining efficiencies through automation.

A review of the literature reveals that, depending on the context, the term eResearch can mean different things to a diverse range of stakeholders and end-users. For instance, eResearch is couched in the context of: i) the research data life cycle (1); ii) being pivotal to research collaboration (2); iii) the focus for electronic collaboration tools (2); iv) critical tools for a specific activity or computational analysis (3); v) contained in conference titles and themes as well as national research funding initiatives[1]; vi) being identified as a core function/Centre within a research institution or organisation[2]; vii) essential training for researchers to improve research digital literacy skills and design of eResearch tools and strategies to enable research data management, analytics, advanced computing and research infrastructure (4); and viii) design and deployment of advanced digital platforms to support and drive research (5).

Results and Discussion

No matter the context, the field of eResearch is critical to all research endeavours and its role can be defined as follows: to work closely with researchers and end users to understand the research question, the breadth of technical and socio-technical challenges in order to devise innovative strategies to ultimately ensure technological solutions are fit-for-purpose and best-of-breed.

This presentation showcases a confluence of successful eResearch interrelated exemplars highlighting the fast evolving best practices in: i) enabling and driving research across multiple disciplines; ii) optimised advanced computing within a heterogenous computing environment; iii) delivery of scalable and reusable eResearch solutions to support diverse institutional research infrastructure including imaging, HASS and biotechnology; iv) deployment of advanced data and analytics platforms for real-world challenges for external partners; v) a revised research data management policy and associated functional guidance material to support internal institutional processes; and vi) developing eResearch standard operating procedures for continual improvement across the eResearch team for rapid response and promote feedback for end-users be they, researchers, students, faculties, schools, Centres and Divisional key stakeholders (4).

[1] Australian Research Data Commons https://ardc.edu.au/

[2] Australasian eResearch Organisations (https://aero.edu.au/)

Measures of eResearch success are defined and presented by which it becomes possible to share and exchange best of breed eResearch approaches across institutions and organisations.

Acknowledgments

The author wishes to acknowledge the contributions of the staff within the Office of eResearch at Queensland University of Technology.

References

1. Gupta S, Müller-Birn C. A study of e-Research and its relation with research data life cycle: a literature perspective. Benchmarking: An International Journal. 2018;25(6):1656-80.
2. Anandarajan M. e-Research Collaboration: Theory, Techniques and Challenges. Anandarajan M, editor: Springer-Verlag Berlin Heidelberg; 2010.
3. Joyce EL, DeAlmeida DR, Fuhrman DY, Priyanka P, Kellum JA. eResearch in acute kidney injury: a primer for electronic health record research. Nephrol Dial Transplant. 2019;34(3):401-7.
4. Bellgard MI. ERDMAS: An exemplar-driven institutional research data management and analysis strategy. International Journal of Information Management. 2020;50:337-40.
5. Bellgard MI, Snelling T, McGree JM. RD-RAP: beyond rare disease patient registries, devising a comprehensive data and analytic framework. Orphanet journal of rare diseases. 2019;14(1):176.

ABOUT THE AUTHOR(S)

Name:

Professor Matt Bellgard

Bio

Professor Matt Bellgard is the inaugural eResearch Director at Queensland University of Technology. He has personally attracted over $45m in research funding, is co-inventor of 5 full/20 provisional patents, co-designed and commissioned a world's top 100 supercomputer, co-authored over 152 peer reviewed articles in areas including human/animal/plant genomics, bioinformatics, health informatics, AI, biosecurity, eResearch, HASS, remote sensing and radio astronomy. He leads the design and development of digital health solutions for government, industry and academia and is Chair of the APEC Life Science Innovations Forum Rare Disease Network.

Staying connected in an evolving eResearch ecosystem

Robin Bensley
Business Operations Manager, New Zealand eScience Infrastructure (NeSI)
robin.bensley@nesi.org.nz

ABSTRACT / INTRODUCTION

HPC has always operated at the frontier of computing and data technologies, pushing the boundaries of computing power and bandwidth. As globally we become ever more connected, the computing and data systems we depend on every day need greater power and performance.

Cloud computing is now evolving to incorporate high performance offerings, though these offerings are still niche in scale and expensive to consume in volume. However the direction is clear. Cloud-native technologies are now highly relevant to those operating HPC and data platforms, causing a rapid evolution in the platforms underpinning modern computational science.

NeSI's advantage is that it has always been more than just a provider of computational power. Its value and impacts extend beyond its specialised infrastructure because NeSI's team of domain experts, research software engineers, and HPC specialists are passionate about supporting research in New Zealand. NeSI's people power delivers tailored support experiences, collaborative partnerships with communities, and dedicated engagement across the research ecosystem.

In this session, we'll share some of the milestones NeSI celebrated in 2020 and how it remains a connected, responsive and essential component of New Zealand's eResearch ecosystem.

ABOUT THE AUTHOR(S)

Robin Bensley, *Business Operations Manager, New Zealand eScience Infrastructure (NeSI)*
Robin joined NeSI as its Business Operations Manager in 2016. Prior to NeSI, Robin worked at the University of Auckland running research finance and contracting teams. He has also worked in commercial roles in New Zealand and Europe, focused on the use of data to drive business strategy, running and setting up teams consulting in business analysis, solution development and IT operations.

Harnessing the disruptive nature of cheap, portable technology for community empowerment

Authors name(s): Miles Benton and Matt Storey
Organisation: Institute of Environmental Science and Research (ESR), Wellington, NZ
Authors Email(s): miles.benton@esr.cri.nz, matt.storey@esr.cri.nz

ABSTRACT

The current 'climate' is full of buzz words, such as: AI (artificial intelligence); deep learning; cloud computing, and the 'Internet of Things'. As consumers, and even research specialists, this can all be overwhelming. At ESR we are endeavouring to provide our staff, clients, and hopefully the wider community, with some insight into the technologies behind this jargon. In this talk I will discuss our experiences with the Nvidia Jetson family of small embedded computing platforms. What started as an idea to address a very personal need has developed into an international, collaborative, cross-programme study to develop and deploy an innovative, disruptive, portable and affordable sequencing technology into the hands of the community to empower their health and well-being. Additionally, the affordable, easy to source components provide exciting opportunities for such endeavours as community outreach and education. If you want a sneak peak visit this GitHub repository:
https://github.com/sirselim/jetson_nanopore_sequencing

ABOUT THE AUTHOR(S)

- Dr Miles Benton
- Bio: - Dr Benton is a Senior Bioinformatics Scientist within the Human Genomics group at ESR. He is interested in the development of methods to deal with ever expanding genomic data sets and their access and interpretation back to the people that matter (i.e. clients, clinicians, researchers, public, etc). Part of his role at ESR has been implementing bioinformatics workflows in both research and clinical settings. He is also developing machine learning/AI technology on portable Nvidia modules for field deployment in various areas. Dr Benton is a member of the Genomics Aotearoa Bioinformatics Leadership Team, where he is responsible for overseeing bioinformatics support for human health projects. He is also heavily involved in the Data Carpentries as an instructor and facilitator, as well as a mentor on ESR's data science accelerator programme. He is deeply committed to making data science and it's tools accessible, with the belief that everyone should be able to 'play' with and interpret their data.

Building a(n) (almost) sustainable institutional training culture

Mik Black
University of Otago / Genomics Aotearoa
mik.black@otago.ac.nz

ABSTRACT / INTRODUCTION

The delivery of digital skills training to the New Zealand research community has become a major shared undertaking for two of our nation's major science infrastructure providers, the NZ eScience Infrastructure (NeSI) and Genomics Aotearoa (GA). Although neither organisation has "training" as a primary objective, both NeSI and GA have embraced the idea that increased digital literacy within a research community invariably leads to better utilisation of science capability, and ultimately to the acceleration and expansion of the research being undertaken.

How these lofty goals are accomplished, however, is not necessarily straightforward. Over the past six years, NeSI, and more recently Genomics Aotearoa, have adopted the Open Source, volunteer-driven teaching framework popularised by the international Carpentries training movement. While both of these national organisations have invested considerable resources into both the delivery of training, and the support of NZ's research communities, there is also a need for strong commitment of resources by researchers' own institutions, in order to ensure that the benefits of digital literacy training are fully realised.

In this talk I will describe the work that the University of Otago (in partnership with GA and NeSI) has undertaken in this space over the past six years, as we have (slowly) moved towards our goal of developing a sustainable local training infrastructure that supports our researchers. This will include lessons that we have learned along the way, advice to others who would tread the same path, and also some speculation on where things may be heading in the future.

ABOUT THE AUTHOR

Associate Professor Mik Black
Mik received a BSc(Hons) in statistics from the University of Canterbury, and a MSc (mathematical statistics) and PhD (statistics) from Purdue University. After completing his PhD in 2002, Mik returned to New Zealand to work as a lecturer in the Department of Statistics at the University of Auckland. An ongoing involvement in a number of Dunedin-based collaborative genomics projects resulted in a move to the University of Otago in 2006, where he now leads a research group focused on the development and application of statistical methods for the analysis of data from genomics experiments, with a particular emphasis on human disease. Mik has also been heavily involved in major initiatives designed to put in place sustainable national research

infrastructure for NZ: Genomics Aotearoa and NZ Genomics Limited for genomics, digital literacy training via The Carpentries, and NeSI (New Zealand eScience Infrastructure) for high performance computing and eResearch.

Embracing cloud-native architectures: What's been done & what's next

BoF

ABSTRACT / INTRODUCTION

Powering today's data-centric and data-intensive research requires computing capabilities that are responsive and high-performing. More and more we're seeing scientific cloud deployments popping up as solutions to these needs, particularly as the international Open Infrastructure (nee OpenStack) community has matured and stabilised in recent years.

This BoF aims to bring together research infrastructure specialists from across the country to discuss what strategies, models, and technical architectures could help New Zealand level up its research cloud capability.

ABOUT THE AUTHOR(S)

Blair Bethwaite, *Solutions Manager, New Zealand eScience Infrastructure (NeSI)*
Blair has worked in distributed computing for over a decade; both in research and for research; for institutional and national projects; from applications, through grid & cloud middleware, to full HPC & cloud systems design, implementation, and operations. Previously over the ditch at Monash University, Blair most recently led Monash's use of OpenStack to underpin research computing. Originally from Christchurch, in mid-2018 Blair returned to take up the opportunity of becoming NeSI's Solutions Manager, focusing back up the technology stack closer to the user. Blair's role within NeSI covers Application Support and Collaboration & Integration.

Otago Bioinformatics Spring School

Murray Cadzow[1], Ludovic Dutoit[1], Dinindu Senanayake[2], Matt Bixley[1], and Ngoni Faya[3]
[1] University of Otago
[2] New Zealand eScience Infrastructure (NeSI)
[3]Genomics Aotearoa
murray.cadzow@otago.ac.nz, ludovic.dutoit@otago.ac.nz, dinindu.senanayake@nesi.org.nz,
matt.bixley@otago.ac.nz, and ngoni.faya@otago.ac.nz

ABSTRACT / INTRODUCTION

Leveraging on the success of two-day workshops modelled on The Carpentries, we decided to “level up” and create a week-long training event hosted at the University of Otago with support from Genomics Aotearoa and NeSI. The Otago Bioinformatics Spring School was an opportunity to symbiotically combine bioinformatic themed workshops, Genomic Data Carpentry, Genotype-by-Sequencing, RNA sequence analysis, and environmental DNA analysis. Separately, each workshop has a degree of required introductory content e.g. *Introduction to the Unix shell*. By grouping this content we could include topics to give a fuller example of reproducible research techniques to attendees, such as incorporation of version control into their workflows.

This talk will highlight successes and challenges from the inaugural Otago Bioinformatics Spring School.

ABOUT THE AUTHOR(S)

Dr Murray Cadzow

- Murray is a Teaching Fellow and Scientific Officer at the University of Otago. He is both a Carpentries instructor and instructor trainer. His teaching focus is on delivering digital literacy training to researchers, and the development and support of the local Carpentries community at Otago. His research involves the use of large datasets to investigate the genetic basis of Gout in Māori and Polynesian populations.

Dinindu Senanayake

- An Applications Support Specialist,HPC at NeSI with a particular interest in Bioinformatics and Computational Biology. Joined NeSI following half a decade of research experience gained in the field of Cancer Genetics, Chemical Genetics and Bioinformatics

Dr Ludovic Dutoit

- Ludovic is a Research Fellow at the University of Otago. His research encompasses evolutionary and ecological genomics, from conservation genetics to environmental DNA. He is dedicated to growing the “omics” capability in New Zealand by delivering training through workshops and individual support.

Ngoni Faya

- Ngoni is Genomics Aotearoa’s Training Coordinator, tasked with supporting and building capacity and capability in bioinformatics for New Zealand.

Matt Bixley

- Matt is a Teaching Fellow at the University of Otago helping to deliver the Carpentries Programs. He has a background in Quantitative Genetics and Genomics and is about to start with NeSI as a Support Specialist.

funcX: Managed and Federated FaaS for Research

Kyle Chard and Ian Foster

University of Chicago and Argonne National Laboratory

chard@uchicago.edu, foster@anl.gov

ABSTRACT / INTRODUCTION

The Function as a Service (FaaS) model offers a new paradigm for remote computing in which applications, implemented as a set of programming functions, are executed without regard for the resources on which they execute. Users register a function by specifying the function signature and function body in a high-level programming language as well as any system or language dependencies. Later, they may then call that function one or more times by specifying input arguments and the id of the registered function. Traditionally, cloud providers operate the hardware, virtual machines, and containers on which the function executes. Users are then billed, in fine-grain increments, for the time that the function executes.

While the FaaS model provides a simple abstraction for scientific computing, unfortunately existing implementations are not designed to support research computing infrastructure which is often heterogenous and distributed. Conversely, research computing infrastructure is not designed to support execution of lightweight programming functions. To address this impedance mismatch, we are developing funcX—a federated FaaS system that provides managed execution of functions across a distributed set of function serving endpoints. funcX implements a Globus-like model, in which a cloud-hosted service enables users to register and manage functions and endpoints, as well as enabling them to securely execute functions on remote endpoints. The funcX Python SDK provides simple abstractions that enable users to write, register, and invoke functions from any Python programming environment (e.g., in a Jupyter notebook). The funcX endpoint software is delivered as a Python package that can be installed by users (in user space) on various computing systems—from laptops and clouds to clusters and supercomputers. The endpoint software manages the secure and scalable execution of functions by provisioning resources, staging function code and data, optionally managing execution in containers, and returning results to users. funcX leverages Globus Auth to enable seamless authentication and group-based authorization for registering functions and endpoints, executing functions, and retrieving results.

In this talk we will describe the funcX system and outline how it is being used in various scientific use cases. For example, we will describe how funcX is used to enable high throughput screening of small molecules in the search for effective COVID-19 therapeutics; automated and event-based computing for real-time analysis of data obtained from X-ray diffraction crystallography; and as an execution runtime in high energy physics analysis tools for transparent remote.

ABOUT THE AUTHOR(S)

Kyle Chard is a Research Assistant Professor in the Department of Computer Science at the University of Chicago. He also holds a joint appointment at Argonne National Laboratory. His research focuses on a broad range of problems in data-intensive computing and research data management. He leads various projects related to distributed and parallel computing, scientific reproducibility, research automation, and cost-aware use of cloud infrastructure.

Ian Foster is the Director of Argonne's Data Science and Learning Division, Argonne Senior Scientist and Distinguished Fellow, and the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago. Foster's research contributions span highperformance computing, distributed systems, and data-driven discovery. He has published hundreds of scientific papers and eight books on these and other topics. Methods and software developed under his leadership underpin many large national and international cyberinfrastructures.

Gladier: A Programmable Data Capture, Storage, and Analysis Architecture for Experimental Facilities

Kyle Chard and Ian Foster

University of Chicago and Argonne National Laboratory

chard@uchicago.edu, foster@anl.gov

ABSTRACT / INTRODUCTION

The extraordinary volume and velocity of data produced by scientific instruments presents new challenges to efficiently organize, process, and share data without overburdening researchers. To address these needs we are developing Gladier (Globus Architecture for Data-Intensive Experimental Research), a data architecture that enables the rapid development of customized data capture, storage, and analysis solutions for experimental facilities. We have deployed a Gladier at Argonne's Advanced Photon Source (APS) and Leadership Computing Facility (ALCF) to enable various solutions, including: delivery of data produced during tomographic experiments to remote collaborators; capture, analysis, and cataloging of data from X-ray Photon Correlation Spectroscopy (XPCS) experiments; and feedback based on analysis of data from serial synchrotron crystallography (SSX) experiments to guide data acquisition.

The Gladier architecture leverages a data/computing substrate based on data and compute agents deployed across computer and storage systems at APS, ALCF, and elsewhere, all managed by cloudhosted Globus services. All components are supported by the Globus Auth identity and access management platform to enable single sign on and secure interactions between components. This substrate makes it easy for programmers to route data and compute requests to different storage systems and computers. Other services support the definition and management of flows that coordinate data transfer, analysis, cataloging, and other activities associated with experimental activities. Each service can be accessed via REST APIs, and/or from Python via a simple client library (which calls the REST APIs). Scientists can then develop experiment-specific data solutions by coding to these APIs or library—or reuse or adapt solutions developed by others. Importantly, both the overall architecture and specific solutions can easily be replicated at other institutions and extended to provide additional capabilities.

We describe three examples to illustrate how Gladier can be used to implement powerful data collection, analysis, and cataloging capabilities.

1. DMagic: Automated data delivery to experimentalists. The DMagic system uses a combination of Globus APIs and APS administrative APIs to 1) automatically create and configure shared storage space on the ALCF Petrel data service before an experiment begins; and 2) automatically copy over experimental data from the beamline to Petrel storage as they are produced during the experiment.

2. XPCS data collection, analysis, and cataloguing. This example uses Globus Automate to automatically collect data at an XPCS experiment, transfer the data to an HPC computer for processing, and then load processed data into a catalog, from where it can be searched and retrieved by authorized individuals 3. Rapid feedback for SSX experiments. This example guides SSX experiments by generating statistics and images of the sample being processed and providing them to the scientists in near real-time. These results can then be used to determine whether enough data have been collected for a sample, whether a second sample is needed to produce suitable statistics, or whether the sample is producing enough data to warrant continued processing.

In this talk we will present the Gladier architecture, highlight the major components used in the architecture, discuss three example data solutions deployed at APS and ALCF, and describe how the Gladier architecture can be replicated in other environments.

ABOUT THE AUTHOR(S)

Kyle Chard is a Research Assistant Professor in the Department of Computer Science at the University of Chicago. He also holds a joint appointment at Argonne National Laboratory. His research focuses on a broad range of problems in data-intensive computing and research data management. He leads various projects related to distributed and parallel computing, scientific reproducibility, research automation, and cost-aware use of cloud infrastructure.

Ian Foster is the Director of Argonne’s Data Science and Learning Division, Argonne Senior Scientist and Distinguished Fellow, and the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago. Foster’s research contributions span highperformance computing, distributed systems, and data-driven discovery. He has published hundreds of scientific papers and eight books on these and other topics. Methods and software developed under his leadership underpin many large national and international cyberinfrastructures.

Patient-specific computational modelling for optimal coronary artery bypass grafting design in cardiac surgery patients

Authors name(s): Krish Chaudhuri [a,b], Alex Pletzer [c,d], Callum Walley [a,d], Chris Scott [a,d], Nic Smith [a]

Organisation(s): [a] The University of Auckland; [b] GreenLane Cardiothoracic Surgical Unit at Auckland City Hospital; [c] The National Institute of Water and Atmospheric Research (NIWA); [d] New Zealand eScience Infrastructure (NeSI)

Authors Email(s): kcha636@aucklanduni.ac.nz

ABSTRACT / INTRODUCTION

Background/Context

Ischemic heart disease is the major cause of death globally[1] and occurs due to blockages in the coronary arteries that supply blood to the heart. To prevent mortality and morbidity from this significant disease, cardiac surgeons perform coronary artery bypass grafting surgeries. This operation is the most common procedure performed by cardiac surgeons. Since the first such operation in 1964[2] it has been routinely performed, yet surgeons themselves differ in their decisions on how to arrange the bypass grafts for a particular patient. The biophysics of the coronary circulation and the consequences of multiple bypass grafts on the flow dynamics down the blocked (stenotic) vessels is complex. To address this issue, in this study we present a numerical simulation based framework for guiding surgeons in selecting the best grafting strategy before undertaking operation.

Method

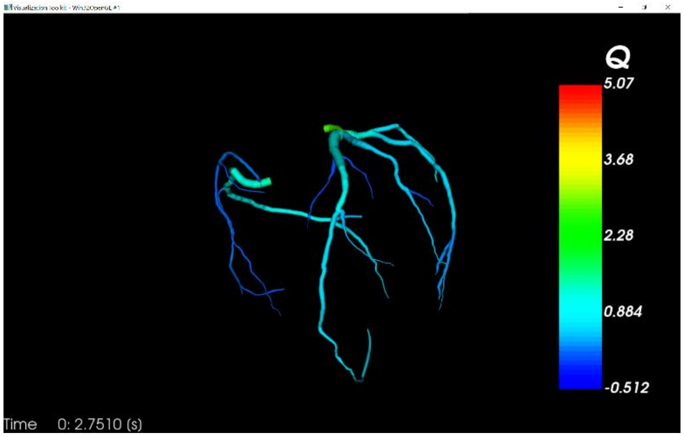
Patients with coronary artery disease had their coronary artery circulations segmented from CT coronary angiograms. A standalone Python code package has been developed to solve 1D Navier-Stokes equations for blood flows and pressures in elastic coronary arteries with stenoses. A lumped parameter model was applied at the end point of large vessels to represent terminal circulation.

Typical grafting strategies are then added to the network circulation models. The models were interrogated to determine the optimal bypass graft strategy to provide the maximal flows in the native coronary arteries and bypass grafts. This was investigated by varying the graft type and arrangements of the inlet and outlet vessels for each graft. Each patient specific model then had the degrees of the stenoses varied to determine the impact of the grafting strategies in these cases.

A group of expert cardiac surgeons were surveyed to note their intended grafting strategy to provide optimal flows for each patient-specific diseased circulation. They were asked whether they would change their grafting design strategy for variations in the degrees of stenosis for a particular patient. The accuracy of their predictions were then compared as both individuals and a group of experts with the computational model. Factors involved in influencing their decision-making were then explored.

Results

Initially a Python code was developed by the researcher, a Cardiothoracic Surgeon, to solve the networks of blood vessels, using an existing Python package VaMpy for guidance[3]. Although a steady state solution was reached, a network of 150 vessel segments would have taken over 4 months to solve. With the help of New Zealand eScience Infrastructure's (NeSI's) consultancy service, the execution time significantly improved to 4 and a half hours. Additional code improvement from NeSI involved allowing complex vessel topologies to be specified using a simple string format. The application of high-performance computing with embarrassing parallelisation, using Mahuika, allowed multiple such networks to be solved within the same timeframe. Improvements in the code were facilitated by the input of the data scientists along with educating the student researcher. Further refinements are currently being made to facilitate the computer model to further support the study involving the cardiac surgeons.



Conclusion

The use of the Python language for programming, facilitated the accessibility and understandability of the coding process for surgeons. This project demonstrates the benefit of medical professionals working alongside with computer scientists to produce software that can assist surgeons in decision making. Whilst the application of this project is in cardiac surgery, other types of surgeries or cognitive psychology could potentially benefit from a similar approach.

Acknowledgments

Royal Australasian College of Surgeons for Surgeon Scientist Scholarship
The Centre for eResearch at The University of Auckland for guidance
New Zealand eScience Infrastructure (NeSI) for consultancy over two phases and high-performance computing

References

1. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *Plos med*. 2006;3(11):e442.
2. Head SJ, Kieser TM, Falk V, Huysmans HA, Kappetein AP. Coronary artery bypass grafting: Part 1--the evolution over the first 50 years. *Eur Heart J*. 2013;34(37):2862-2872.

3. Diem AK, Bressloff NW. VaMpy: A python package to solve 1D blood flow problems. *Journal of Open Research Software*. 2017;5(1).

ABOUT THE AUTHOR(S)

Krish Chaudhuri

Krish is a Cardiothoracic Surgeon at the GreenLane Cardiothoracic Surgical Unit at Auckland City Hospital and student at The University of Auckland

Alex Pletzer
Alex is a Research Software Engineer at NIWA and NeSI

Callum Walley
Callum is an Application Support Analyst at The University of Auckland and NeSI

Chris Scott
Chris is a Research Software Engineer at The University of Auckland and NeSI

Nic Smith
Nic is an Honorary Academic and Research Fellow at The Auckland Bioengineering Institute at The University of Auckland

Climate, Weather and Avocados

Jelena Cosic, Louis Ranjard

PlantTech Research Institute
jelena@pri.co.nz

louis@pri.co.nz

ABSTRACT / INTRODUCTION

Agriculture is highly dependent on climate and, as such, crop yield variability is affected by year-to-year climatic variability, with regards to both extreme events and changes in historical patterns of regional climate. Understanding the effect of weather variables on crop production is key to forecast and manage production. Currently, multiple systems, including remote and proximal sensing, collect data at high temporal frequency. However, it is challenging to identify the specific variables that can specifically have an impact on production.

The project we are currently working on uses NOAA GFS historical longitudinal data of the more than last ten years. We identify and extract multivariate weather factors that are known to have an impact on the avocado yield. New methodologies, based on machine learning and artificial neural networks, are investigated to identify the most important set of variables that have an effect on the avocado production. It is known that interactions between weather variables can have more impact than single variables effect, therefore it is important to build models that can directly consider these interactions. These weather factors are then used in the model along with the avocado yield that was recorded for the last ten growing seasons.

ABOUT THE AUTHOR(S)

Jelena Cosic

Jelena is a PhD Candidate at the University of Auckland. During her PhD, Jelena was working on developing a Bayesian network (BN) as a method of representing vineyard ecosystem. BN was used to model vineyard ecosystem incorporating chemical profiles, meteorological information, and other data at different time points in the life cycle of vineyards in order to discover the differences vineyard management techniques make with respect to resilience and profit.
Last January Jelena joined PlantTech Research team as research scientist. Jelena's research focus on the development and application of machine learning methods to provide answers to some of the interesting horticultural problems.

Louis Ranjard

Louis completed a PhD (Biology) from the University of Auckland in 2010. After his PhD, Louis worked as a postdoc at the Department of Statistics at the University of Auckland and also as bioinformatician for New Zealand Genomics Ltd. After a postdoc experience at the Australian National University, Louis returned to New Zealand to work as a machine learning engineer at Biomatters Ltd. As principal scientist for PlantTech Research Institute Ltd, Louis's research focus on the development and application of machine learning methods to the analysis of biological datasets, with a particular emphasis on the integration of multiscale and heterogenous data sources, from genomics to sensors data.

Data Science Accelerator Level 2 - Power-up or reboot?

Richard Dean

Institute of Environmental Science and Research

richard.dean@esr.cri.nz

ABSTRACT / INTRODUCTION

If 2020 were a computer game, it would be a pretty difficult level. There were more twists and turns than Super Mario Kart, and there was the ever-present threat from an end of level baddie like no other. Despite the gameplay being set to 'hard', during 2020, ESR's data science initiative battled on – we trained more mentors, taught data skills remotely and worked through more data science accelerator projects than ever before.

This presentation reports on how the second year of ESR's data science accelerator programme has gone – was it game over or power up?

The presentation follows on from last year's eResearch 2020 keynote where Richard outlined ESR's data science initiative within the context of the evolution of Public Health Data Science - from Bills of Mortality in 1592, through the cholera years of the 19th century to genomics and critical risks of the current day such as climate emergency, inequalities, lifestyle factors and emergent pandemics.

ESR's data science accelerator programme is critical for driving organisational change and a new focus on emergent skills in areas such as engineering, robotics, innovation, coding and automation – Erica for short.

ABOUT THE AUTHOR(S)

Richard Dean

Richard is a data nerd and ESR's data science accelerator lead - which means he gets to work with lots of very clever scientists who keep kiwis safe from harm and protect our environment. His favourite computer game when a kid was SimCity 2000, which in many ways was simply a precursor to a role at ESR where he still gets a little bit excited when looking at a map of the sewers.

Data Science at Crown Research Institutes

Richard Dean (1), Linley Jesson (2), Alan Tan (3)

1 - Institute of Environmental Science Research (ESR), 2 – Plant and Food Research, 3 – Scion

richard.dean@esr.cri.nz, linley.jesson@plantandfood.co.nz, alan.tan@scionresearch.com

ABSTRACT / INTRODUCTION

New Zealand's Crown Research Institutes present many opportunities for data scientists across a diverse science system which covers everything from predicting weather and climate change to earthquakes, image processing and disease surveillance.

In this birds of a feather session, data scientists from across CRIs have an opportunity to discuss types of problems tackled using data science in their respective domains, common interests and look at opportunities for collaboration and consolidation of data science tools, methods and systems that have applications across many domains.

ABOUT THE AUTHOR(S)

Richard, Linley and Alan are data scientists who believe that New Zealand will achieve it's best through collaboration and sharing our expertise across our respective institutions. In November 2020, they set up a cross-CRI data science meetup group with the aim of improving collaboration and developing career pathways for data scientists who work at research organisations within New Zealand.

Richard Dean leads the Data Science Accelerator initiative at the Institute of Environmental Science and Research (ESR), a scheme he unashamedly stole from his experience while a senior data scientist with Public Health England's Public Health Data Science team. Richard's research interests focus on the translation of data in to knowledge and in assessing the potential of emergent technology to transform data capture and reporting processes in new and novel ways.

Linley Jesson has a PhD in Evolutionary Genetics from the University of Toronto and then a lecturer at Victoria University of Wellington and University of New Brunswick in Canada. Following a 2-year stint in Marburg Germany on a Humboldt Fellowship she returned to New Zealand and joined Plant and Food Research as a statistical scientist. She has been the group leader of Data Science at Plant and Food since 2017. Her research interests include communication of statistical concepts, evolutionary genetics and the development of new predictive models. She is currently leading a project on developing best practices for taonga data in Plant and Food Research.

Dr. Alan Tan is a senior data scientist in Scion, where he contributes strategically and technically to Scion’s data science research initiatives. Dr. Tan’s research interest is in the applications of Deep Learning in 3D remote sensing data, data visualisation, distributed systems and high-performance computing. He obtained his Ph.D. in Computer Science from the University of Waikato.

REANNZ Connect – BoF session

Hannah Edwards
REANNZ
engagement@reannz.co.nz

ABSTRACT / INTRODUCTION

As a part of eResearch NZ 2021 REANNZ would like to invite members to join us in Wellington for the next instalment of REANNZ Connect. The session is open to all REANNZ members who would like to know more about who we are, what we do and they ways that REANNZ supports and functions within the eResearch community.

Last years' REANNZ Connect brought together network engineers, systems engineers and architects who use or operate networks across the research and education sector. This community-building event emphasised technical collaboration, skills development and the open exchange of ideas.

This year we would like to broaden the forum to include technologists, researchers and end users and members of the wider research and education sector to connect and network at a people level, not just at a high-speed network level.

REANNZ will prepare updates from the wider team, but we encourage open discussion and our members to bring along any questions or discussion points. Members of our Network Operations team and Engagement team will be attending in order to support members to utilise the network and community to its full capacity.

ABOUT THE AUTHOR(S)

REANNZ

Moving data, mobilising knowledge - REANNZ, the Research and Education Advanced Network New Zealand, is New Zealand's designated National Research and Education Network (NREN). We provide a specialist network, tools and services to enable NZ's scientists, researchers and educators to connect, collaborate and contribute.

Hannah Edwards

Hannah is the Communications and Marketing Manager at REANNZ and develops their external marketing and communications. She supports and coordinates engagement initiatives and activities with members and partners, spreading the word of how REANNZ members utilise the network and services to support the crucial work that they conduct.

A Survey of iRODS Rules to Enforce Site Policies and Enable Automated Workflows

David Fellinger

iRODS Consortium

davef@renci.org

ABSTRACT / INTRODUCTION

Data management and curation in research sites has become increasingly more complex in the last decade. A large portion of this complexity is caused by the ever-increasing number of instruments and sensors producing huge amounts of data. At the same time, researchers and data curators have joined to enforce FAIR (Findability, Accessibility, Interoperability, and Reuse) Principles of data products. This is a worldwide initiative which will increase the potential of collaboration through data sharing across diverse sites.

Managing large amounts of data requires rules-based automation to ensure that site policies are respected, and, at the same time, workflows must be enabled from data ingestion through the distribution of data products.

The evolution of iRODS technology

Initially, the basis for iRODS (The Integrated Rule-Oriented Data System) was a university project funded by a government agency. The concept was to build a searchable data base with entries that were linked to specific data represented in a plurality of POSIX compliant file systems. These file systems would appear as a single namespace so that a user did not have to be concerned with data locality but rather, just the descriptive metadata which was provided by a researcher when the file was generated. This original implementation of open source cataloging was successful to the extent that it became the basis of a commercial product that was sold to various government organizations. The evolution of iRODS as it exists today began with the founding of the iRODS Consortium about 7 years ago. Today iRODS can be described both as a technology platform and, through the consortium, a vibrant and diverse community with many mutual goals.

Currently, iRODS is built to support 8 essential capabilities. These include automated data ingestion, data integrity checking, storage tiering, and auditing enabling indexing, provenance tracking, and compliance checking. Finally, an interface to publication completes the feature set.

Use cases requiring policy enforcement were described at the most recent iRODS Consortium User Group Meeting [1].

First, let’s look at the Victoria Department of Agriculture in Australia. Their overall goal is nothing less than increasing farm efficiency in the entire state. Their policies include automated data gathering and migration to a processing site so that data products can be analyzed. The goal will be to federate iRODS locations located on individual farms. Sensor data will be automatically ingested on the farm then the aggregate data will be ingested to a central site also utilizing iRODS rules to enforce the collection policy [2]. This managed use of eResearch and iRODS will directly affect the GDP of the state.

CyVerse, based in Tucson is a perfect example of a multi-national application of iRODS. The site hosts a diverse range of research data starting several years ago with plant genomics [3]. The data is mirrored to an iRODS based site at TACC (Texas Advanced Computing Center) in Austin [4]. Data is ingested from partner sites worldwide including Melbourne, Sydney, Brisbane, Canberra, Adelaide, Perth, and Hobart in Australia [5]. CyVerse also offers compute services using the “data to compute” model.

The KTH Royal Institute of Technology is utilizing iRODS to transition data between GPFS file systems allowing file system upgrades while maintaining availability [6]. The iRODS based checksumming was employed to assure that data integrity was maintained through the entire process. Storage at KTH is utilized to host data from the entire country of Sweden.

At Utrecht University data site policies stress data discovery to enable research. To that end, a custom interface called Yoda was written based upon iRODS. “Yoda deploys iRODS as its core component, customized with more than 10,000 lines of iRODS rules”. This site is successfully hosting a very large research data archive [7].

At KU Leuven iRODS is being utilized to allow researchers to have active data utilization enabling project work before publication. FAIR principles are stressed in their data policies with iRODS tools for compliance [8].

At Bristol Myers Squibb, iRODS is being utilize to manage AWS cloud-based data sets to enable worldwide project progress. The iRODS technology is used to manage data flows and to maintain a catalog of the available data in real time interfacing with AWS Lambda functions [9].

At the NIEHS (National Institute of Environmental Health Sciences) discoverability of diverse data sets with auditable data governance is stressed. Metadata templates have been written to guarantee standardization in the description of the hosted data [10].

There are many research sites both academic and commercial that use iRODS to enforce policies. It’s important to note, however, that a site policy does not have to fully evolve immediately. The iRODS open source technology with its plug-in framework allows policies to grow and evolve over time and effectively “future-proof” eResearch archives worldwide.

About the author

Dave Fellinger is a Data Management Technologist and Storage Scientist with the iRODS Consortium. He has over three decades of engineering experience including film systems, video processing devices, ASIC design and development, GaAs semiconductor manufacture, RAID and storage systems, and file systems. He attended Carnegie-Mellon University and holds patents in diverse areas of technology.

References

1. The agenda for the iRODS User Group Meeting 2020 is available from; https://irods.org/images/irods_ugm2020_agenda.pdf, accessed 11 November 2020 2. A presentation from the Victoria Department of Agriculture is available from; https://irods.org/uploads/2020/Murphy-AgVic-SmartFarm_Data_Managementslides.pdf, accessed 11 November 2020 3. A presentation from CyVerse is available from; https://irods.org/uploads/2020/Roberts-CyVerse-Discovery_Environment-slides.pdf, accessed 11 November 2020 4. A presentation from TACC is available from; https://irods.org/uploads/2020/Jordan-TACCThe_Past_Present_and_Future_of_iRODS_at_TACC-slides.pdf, accessed 11 November 2020 5. A description of EMBL in Australia is available from; https://www.embl-abr.org.au/wp-content/uploads/2016/05/EMBL-nodes-hubs-flyeronline-final.pdf, accessed 11 November 2020 6. A presentation from KTH is available from; https://irods.org/uploads/2020/Korhonen-KTHMigration_Between_GPFS_Filesystems-slides.pdf, accessed 11 November 2020 7. A presentation from Utrecht University is available from; https://irods.org/uploads/2020/Westerhof-Smeele-UtrechtUniYoda_and_iRODS_Python_rule_engine_plugin-slides.pdf, accessed 11 November 2020 8. A presentation from KU Leuven is available from; https://irods.org/uploads/2020/Barcena-KULeuven-VSCiRODS_Data_Management_Platform-slides.pdf, accessed 11 November 2020 9. A presentation from Bristol Myers Squibb is available from; https://irods.org/uploads/2020/Shaikh-BMSiRODS_for_scientific_applications_in_AWS_Cloud-slides.pdf, accessed 11 November 2020 10. A presentation from NIEHS is available from; https://irods.org/uploads/2020/Conway-NIEHS-Applications_of_iRODS-slides.pdf, accessed 11 November 2020

Coaching great practices of describing a problem

José Filipe Gonçalves Higino
NIWA / NeSI
jose.higino@nesi.org.nz

ABSTRACT / INTRODUCTION

From asking a question to someone, to opening a request case at a support channel, we often witness the arduous path of how to better describe our problem or how to ask for the right information in case you helping solve one, in order to avoid that inefficient "ping-pong" communication.

Working through support services, resolving problems which require a complex troubleshooting skills and technical knowledge are areas where usually reducing the scope of the investigation is the best course of action. Although, finding the right questions to ask or providing the right information is often seen as a difficult task to master, especially when we know nothing about it.

Using a methodology based on "**Kepner**-**Tregoe** Problem Solving and Decision Making (PSDM)" we can train the mind to ask questions that narrow the scope of the problem, eventually improving the way we describe it, which leads to a potential reduction of time to find a solution for it.

I will be presenting a short version of an adaptation of the methodology for real practical use cases that help both requester and helper to better describe a problem.

References:

- The New Rational Manager, by Charles H. Kepner and Benjamin B. Tregoe

ABOUT THE AUTHOR(S)

José Higino

José moved from Portugal to New Zealand in 2014 to work with NIWA's Supercomputer, Fitzroy, as a High Performance Computing (HPC) Systems Engineer in Wellington. His has a background in electrotechnics and computers engineering and worked for 7 years at IBM Portugal as a software and services IT Specialist. Passionate about troubleshooting, logic, distributed and parallel systems, he followed the roots of HPC and distributed storage. Currently, he is part of the NeSI's Platforms Team supporting the Maui (Cray XC50) and Mahuika (Cray CS500/400) supercomputers.

Moving data: getting up to speed with Globus and Science DMZ

Brian Flaherty, Richard Tumalian, Megan Guidry, Hannah Edwards
NeSI, REANNZ
brian.flaherty@nesi.org.nz, megan.guidry@nesi.org.nz, richard.tumalian@reannz.co.nz,
hannah.edwards@reannz.co.nz

ABSTRACT / INTRODUCTION

Supporting a national scientific Data Transfer platform is a collaborative effort, NeSI, New Zealand eScience Infrastructure, and REANNZ, the Research and Education Advanced Network New Zealand, work in partnership to provide the digital tools that support collaboration and knowledge-sharing within New Zealand's research sector and beyond, with international connectivity via the high-speed network and the worldwide network of Data Transfer Nodes powered by Globus.

In this session Brian Flaherty, Data Services Product Manager, and Richard Tumaliuan, Senior Network Engineer, will each provide an overview of the tools that support and enable end users to make fast, secure, and reliable transfers with Globus and Science DMZ, as well as how these tools fit together to improve performance across the whole data transfer journey.

- How can researchers move data on a global scale?
- What is a Science DMZ?
- What is Gloubs?

There will also be an opportunity for discussion groups to delve deeper into the user experience of Globus and the technical aspects of Science DMZ:

Globus breakout group - Globus users and potential Globus users can see how they can get value from the service and the steps for getting started. The group will discuss the difference between managed vs. personal endpoints and how users can engage with more advance transfer options.

Science DMZ breakout group – will consider the technical aspects of the Science DMZ for example transfer node reference design and the REANNZ managed network edge, with examples of how the Science DMZ has been deployed and technical requirements.

At the end of the session groups will come back together for an opportunity to ask any questions, hear the key points to take away from each group's discussion.

Audience –

This session is open to NeSI's contributors and REANNZ members, end users and researchers interested in using Globus or learning more about what it can enable. The session is also open to technologists that work within the R&E sector (network engineers, software engineers, systems administrators within REANNZ membership) .

ABOUT THE AUTHOR(S)

Brian Flaherty

Brian is Product Manager, Data at NeSI. He has a background in digital libraries digital scholarship, research infrastructure & support and discovery services.

Richard Tumalian

Richard is a Senior Network Engineer at REANNZ, working alongside the wider Network Operations team to support the core services and connectivity solutions.

Megan Guidry

Megan Guidry is the Regional Coordinator for the Carpentries in New Zealand and also works as the training coordinator for the New Zealand eScience Infrastructure (NeSI). Her main priority is raising the eResearch capability in New Zealand through training delivery and community building.

Hannah Edwards

Hannah is the Communications and Marketing Manager at REANNZ and develops their external marketing and communications. She supports and coordinates engagement initiatives and activities with members and partners, spreading the word of how REANNZ members utilise the network and services to support the crucial work that they conduct.

Taonga: building a data repository for genomics research in New Zealand

Brian Flaherty and Jun Huh from New Zealand eScience Infrastructure

Genomics Aotearoa's data repository working group are authors as well. They are:

* Mik Black from UoO

* Ben Te Aika from UoO

* Ben Curran from UoA

* Miles Benton from ESR

* Libby Liggins from Massey

* Rudi Brauning from AgResearch

ABSTRACT / INTRODUCTION

Genomics Aotearoa and NeSI formed a partnership in 2018 to support the computational needs of genomics researchers, with NeSI providing high performance compute and storage capabilities. Over the past two years the partnership has broadened to include the management of research outputs, with indigenous data sovereignty being a key driver for onshore hosting of genomic data sets.

In 2019, NeSI undertook a repository discovery project, building a locally managed Globus platform, which provided high speed transfer and a secure, role based access control for a small number of genomics files. In 2020, with more datasets on the horizon, Genomics Aotearoa and NeSI initiated a capability development project to design a solution that would enable researchers to securely store, archive and share genomic data from taonga species through a NZ-based portal with customised search, browse, and access functionality. The data repository prototype was built using Gen3, an open source genomics repository solution developed by the University of Chicago in partnership with the US National Cancer Institute.

The discovery and development process included end user workshops to identify researchers' needs; in particular, functionality and metadata requirements that could support FAIR and CARE data principles, and

the fundamental requirement to support managed access and governance of taonga species data in line with the principles of Maori Data Sovereignty.

The aim of this BoF is to introduce the work that has been done so far, and to open up the discussion for future challenges and opportunities around managing research data in New Zealand. Some of the known challenges include supporting FAIR and CARE principles, handling of security and privacy, especially with sensitive data. centralised v distributed repository infrastructure & funding, discovery of data hosted offshore, and provision of analysis and visualisation tools close to the data.

SciDataMover: Reliable, secure data movement for Australian researchers

Ryan Fraser[1], Chris Myers[1]
[1]AARNet
ryan.fraser@aarnet.edu.au, chris.myers@aarnet.edu.au

ABSTRACT

Despite extremely high speed and low latency network, Australian researchers still resort to using portable storage devices to move data around. Generally, the read/write operations of the endpoints are not up to the speed to which data volumes transfer on the network and hence data or packets are lost, resulting incomplete data for researchers to analyse.

AARNet is provisioning a service in collaboration with Globus called SciDataMover that will aid Australian Researchers move large data between endpoints with ease in a reliable and secure manner. The goal is to deploy Globus across the research sector in Australia, particularly at "large data facilities" to support researchers.

We will present the early results and learnings of the utilisation of the service, along with future-plans for rollout across Australia to support growing data needs of institutional researchers. Further, we will highlight the opportunities this service present for other developers of services and Institutions.

ABOUT THE AUTHOR(S)

Ryan Fraser is Manager of Research and Commercial at AARNet. He is working with researchers at AARNet member institutes and affiliates to develop and provision infrastructure required to support research activities. Formerly, Ryan was at CSIRO and had an extensive career in developing and leading e-Infrastructure and data science programs, which delivered to clients in government and industry.

Digital collecting vs. digital research – Are they compatible?

Andrea Goethals
National Library of New Zealand
Andrea.goethals@dia.govt.nz

ABSTRACT / INTRODUCTION

This presentation will explore with eResearch attendees the following question: How can researchers inform the National Library of New Zealand's digital collecting, processing, description and delivery so that these collections are more useful to researchers?

The first half of the presentation will give eResearch attendees a high-level overview of the Library's digital collections, highlighting the Web and social media collections; the Library's activities to make its digital collections open for data analysis by researchers; and the legal, social, ethical and methodological challenges encountered.

In the second half of the presentation, we will lead eResearch attendees in a discussion of how digital collecting institutions like the Library, can work more closely with digital researchers to provide data that is usable for research purposes. The Library's collecting processes include many decisions that ultimately shape what gets collected and what can be done with it. What are the types of decisions made by collectors that shape the digital collections? Which collecting decisions do researchers want documented? Where in the collecting, processing, description, packaging, delivery pipeline do researchers want to be involved? How can we create a feedback loop between the Library's collecting and users of the collections, so that we are enabling and not inhibiting research in NZ?

ABOUT THE AUTHOR

Andrea Goethals manages the digital preservation team at the National Library of New Zealand. She has primary responsibility for the overall day-to-day operations of the National Digital Heritage Archive and contributes to the strategic direction of the Library's digital preservation programme. She champions digital preservation issues and collaborates closely with others at the Library and around the world to advance digital preservation standards and practices. She co-chairs the Library's Digital Research Working Group, runs the NZ DOI Consortium, and is a co-organiser of Australasia Preserves.

Digital outreach – communications in a post-Covid world

Lucy Guest, Chris Wilkinson, Aidan Muirhead, Adam Huttner-Koros
NCI Australia Communications & Outreach Team
lucy.guest@anu.edu.au

Jana Makar, Megan Guidry
NeSI Engagement Team
jana.makar@nesi.org.nz

ABSTRACT / INTRODUCTION

A note to reviewers
We would like to submit this paper as a BoF session, but understand that you require presenters to attend in person. Given both the nature of the presentation, a live virtual tour of our facility, and the unfortunate fact that we are unable to travel to NZ due to covid and budget restrictions, we are hoping you may consider this as an option. Many thanks!

TIME REQUIRED: One hour.

ABSTRACT / INTRODUCTION

The abrupt halt to in-person outreach activities in 2020 inspired organisations to trial innovative online experiences for education and outreach activities. NCI Australia launched their new supercomputer, Gadi, which debuted at #24 in the June 2020 Top500 list, during Australia's Covid19 lockdown period. With researchers, school groups and conference attendees no longer allowed to visit the site, the Communications and Outreach Team began running live virtual tours. The first tour, held during National Science Week, ran on Twitter's Periscope platform, and subsequent tours have used both WebEx and Zoom.

Join this presentation and the NCI team will take you on a virtual tour of Australia's #1 supercomputer followed by the New Zealand eScience Infrastructure (NeSI) team making their tour debut with a virtual walk through their High Performance Computing Facility in Wellington. Join communications and outreach professionals and share your experiences of running digital outreach in a post-covid landscape.

ABOUT THE AUTHOR(S)

Lucy Guest – Communications and Outreach Manager, NCI Australia
Lucy's passion for STEM began on a sheep farm in Northern NSW where her childhood was spent exploring, experimenting and investigating. The National Youth Science Forum cemented 'science' as a career path, and a Bachelor Science/Law undertaken at UNE. It was the NYSF that brought her to Canberra, where she worked as the Marketing and Communications Officer, relishing the opportunity to introduce the joy of STEM to next generations. Lucy joined NCI as their Communications Manager in 2012 and is committed to championing women in HPC.

Chris Wilkinson, NCI Communications Officer has a background in journalism and media production, and has been with NCI Australia as a Communications and Outreach Officer for over four years. First and foremost a storyteller, Chris translates the incredible scientific achievements made possible by high-performance computing and data services in Australia and shares these amazing stories with the world through moving images, research highlights and other media.

Aidan Muirhead, NCI Communications Officer grew up in two Australian territories – the ACT and the NT – as well as Singapore and Serbia. She loved that maths gave her a universal language and has always wanted to know more about how things work. Passion for STEM and sharing stories led her to complete a Bachelor and Graduate Diploma in Science Communication at ANU. After 8 years at Questacon developing and delivering STEM programs across Australia, Aidan moved to NCI in 2019. Aidan is proud to support diversity in HPC, HPD, and eResearch.

Adam Huttner-Koros, NCI Communications Officer, is a science communicator from Canberra, currently working at the National Computational Infrastructure. He came to do comms in the supercomputing world after some time dabbling with freelance writing. With interests spanning across science and in particular to linguistics, his broad experience in languages and science writing often comes in handy for cross-cultural and cross-disciplinary communication. Currently attempting to add German to a language portfolio including English, French, Hungarian and Japanese, Adam finds the science of language use, sharing and development fascinating

Jana Makar, Communications Manager, New Zealand eScience Infrastructure (NeSI)
Based at the University of Auckland, Jana coordinates a variety of engagement initiatives and external communications to raise the profile of NeSI's activities, impacts, and collaborations. Prior to joining NeSI, Jana spent more than a decade in communications roles with various organisations in Canada's digital research infrastructure sector, from provincial research and education networks to regional and national high performance computing platforms. She has a degree in Communications from the University of Calgary and spent the early part of her career working as a newspaper journalist.

Megan Guidry, Research Communities Advisor, New Zealand eScience Infrastructure (NeSI)
Megan Guidry is the Regional Coordinator for the Carpentries in New Zealand and also coordinates the training activities of New Zealand eScience Infrastructure (NeSI). Her main priority is raising eResearch capability in New Zealand through training delivery and community building.

BoF

Challenge Accepted: Responding to community feedback for supporting diversity in HPC & eResearch

Lucy Guest, Aidan Muirhead

NCI Australia Communications & Outreach Team

lucy.guest@anu.edu.au

Jana Makar, Megan Guidry

NeSI Engagement Team

jana.makar@nesi.org.nz

Aditi Subramanya Pawsey

Communications

aditi.subramanya@pawsey.org.au

Kerri Wait Monash

eResearch Centre, Monash University

kerri.wait@monash.edu

Loretta Davis Executive Officer,

Australasian eResearch Organisations (AeRO)

loretta@aero.edu.au

Dr Jenni Harrison

Director of Strategic Projects and Engagement, Pawsey and Chair of the AeRO Executive Committee

jenni.harrison@pawsey.org.au

ABSTRACT / INTRODUCTION

In November 2020, an Australasian Chapter of the global organisation Women in High Performance Computing (WHPC) was launched to better support diversity across the Australian and New Zealand HPC and eResearch sectors. As one of the first steps to connect with the community, the Chapter's founding organisations — NeSI, Australasian eResearch Organisations (AeRO), Monash University, NCI Australia, and Pawsey Supercomputing Centre — polled community members on what activities and initiatives they'd like the Chapter focus on in 2021. In this session, we will review the results of that community consultation, discuss how the top-ranked activities can be actioned, as well as dive deeper into what can be learned from past and other initiatives related to mentorship, recruiting & retention, professional development, and community-building for women in HPC and eResearch.

more valuable experience for participants.

ABOUT THE AUTHOR(S)

Jana Makar, *Communications Manager, New Zealand eScience Infrastructure (NeSI)*

Jana coordinates a variety of engagement initiatives and external communications to raise the profile of NeSI's activities, impacts, and collaborations. Prior to joining NeSI, Jana spent more than a decade in communications roles with various organisations in Canada's digital research infrastructure sector, from provincial research and education networks to regional and national high performance computing platforms.

Megan Guidry, *Research Communities Advisor, New Zealand eScience Infrastructure (NeSI)*

Megan Guidry is the Regional Coordinator for the Carpentries in New Zealand and also coordinates the training activities of New Zealand eScience Infrastructure (NeSI). Her main priority is raising eResearch capability in New Zealand through training delivery and community building. Lucy Guest, NCI Communications & Outreach Manager Lucy's passion for STEM began on a sheep farm in Northern NSW where her childhood was spent exploring, experimenting and investigating. The National Youth Science Forum cemented 'science' as a career path, and a Bachelor Science/Law undertaken at UNE. It was the NYSF that brought her to Canberra, where she worked as the Marketing and Communications Officer, relishing the opportunity to introduce the joy of STEM to next generations. Lucy joined NCI as their Communications Manager in 2012 and is committed to championing women in HPC.

Aidan Muirhead, *NCI Communications Officer*

Aidan grew up in two Australian territories – the ACT and the NT – as well as Singapore and Serbia. She loved that maths gave her a universal language and has always wanted to know more about how things work. Passion for STEM and sharing stories led her to complete a Bachelor and Graduate Diploma in Science

Communication at ANU. After 8 years at Questacon developing and delivering STEM programs across Australia, Aidan moved to NCI in 2019. Aidan is proud to support diversity in HPC, HPD, and eResearch.

Kerri Wait, *HPC Consultant, Monash eResearch Centre, Monash University*

Kerri's HPC journey began as an electronic engineering student simulating semiconductor devices during an industrial experience placement in Germany. Kerri has worked at a number of HPC and research computing facilities in Australia, collaborating with researchers to deliver scientific research that is faster, less painful, more robust, and repeatable. Kerri attended IBM's EXITE program as a high school student, returning to speak as an early career professional, and is particularly interested in supporting women from low socioeconomic backgrounds to explore careers in STEM.

Aditi Subramanya, *Marketing & Events Officer, Pawsey Supercomputing Centre*

Aditi is a creative marketing and communications professional with more than 10 years' experience in her chosen profession. She holds a Bachelor of Commerce specialising in Public Relations and Tourism and Event Management. She has been instrumental in providing global visibility in order to showcase Pawsey's capabilities and services via key exhibitions at conferences worldwide, and plays a pivotal role in increasing market presence and overall brand awareness.

Loretta Davis, *Executive Officer, Australasian eResearch Organisations (AeRO)*

Loretta is a seasoned IT professional with 25+ years experience in the eResearch, commercial and government sectors in Australia, Africa and the USA. When not working part time for AeRO, Loretta consults as a Solutions Specialist to a number of private clients.

Dr Jenni Harrison, *Director of Strategic Projects and Engagement, Pawsey Supercomputing Centre and Chair of the AeRO Executive Committee*

Jenni is a passionate leader in technology and a positive role model. Jenni is an inclusive, strategic thinker who leads on national STEM initiatives, whilst mentoring others (presently a mentor for IMNIS and AIM WA). On 30th October 2020, Jenni was recognised by Women in Technology WA as a Tech [+] 20 Award Winner for 2020. Jenni is passionate about women in STEM and inclusion, is a Member of STEM Women, Women in STEMM, UN Women, WiTWA and is a Women in Data Science Ambassador for 2020. Jenni has presented on inclusion in STEM at several international conferences and events. An AICD graduate, with substantial governance experience, Jenni uses her skills to promote inclusion. In this regard Jenni is Chair of SHINE, a remarkable Not for Profit organisation based in the Geraldton region that collaborates with business and schools to actively engage with young female students who are at risk of disengaging from the conventional education system. Jenni is a lifelong learner and published author.

Sowing the seeds of capability: Experience what Carpentries instructor training is all about

Megan Guidry, New Zealand eScience Infrastructure, megan.guidry@nesi.org.nz

Murray Cadzow, The University of Otago, murray.cadzow@otago.ac.nz

Arindam Basu, The University of Canterbury, arindam.basu@canterbury.ac.nz

ABSTRACT / INTRODUCTION

The Carpentries is a global community that is dedicated to raising the capability of researchers. It does this by enabling communities to run or request workshops for coding and data management skills - whether it is R for social scientists, python for ecologists, version control with git, or one of the many other collaboratively constructed lessons maintained by the community.

The Carpentries is dependent on local champions instructing and coordinating events for their region/ institution. So the best way to get more events happening locally is to prepare individuals in the community to teach!

Those that have the desire to share their coding or data management skills can request to attend a 2-day Carpentries instructor training workshop for free, which teaches them pedagogy and practical tips for preparing for and running a workshop - things like how to create a positive environment for learners, motivation and demotivation, and how to get feedback from learners. The Carpentries instructor training workshops give researchers in the community the tools and guidance they need to become confident and capable instructors.

What is this session about?

In this session, we will explain why NeSI and the other NZ Caprentries member institutions buy into the Carpentries instructor training model. We then will teach you one of our favourite instructor training lessons. Come along if you are curious about the instructor training model, want to know more about the Carpentries, or if you want to experience (or re-experience) a lesson from the much-loved instructor training curriculum.

ABOUT THE AUTHOR(S)

Megan Guidry is the Regional Coordinator for the Carpentries in New Zealand and works as Research Communities Advisor for New Zealand eScience Infrastructure (NeSI). Her main priority is raising eResearch capability in New Zealand through training delivery and community building.

Murray Cadzow is a Teaching Fellow and Scientific Officer at the University of Otago. He is both a Carpentries instructor and instructor trainer. His teaching focus is on delivering digital literacy training to researchers, and the development and support of the local Carpentries community at Otago. His research involves the use of large datasets to investigate the genetic basis of Gout in Māori and Polynesian populations.

Arin Basu is a medical doctor, lecturer, and an epidemiologist. He works at the University of Canterbury at Christchurch as a senior lecturer in the Health Sciences Centre and serves as a senior researcher at the Health Services Assessment Collaboration, Health Sciences Centre. Arin is currently leading two research projects at the Health Sciences Centre - the first one is funded by the Tertiary Education Commission on immersive learning through virtual reality applications for the training of physicians and nurses and the second project, funded by Ako Aotearoa, is on the technology used for the professional training of nurses, physicians and other health care workers who are involved in the provision of Telehealth services. Arin's research interests include the use of virtual worlds in medical education, training, and provision of medical care.

Data-driven horticulture

Istvan Hajdu and Ian Yule
PlantTech Research Institute
istvan@pri.co.nz, ian@pri.co.nz

ABSTRACT / INTRODUCTION

Smart data collection is already happening on a real time basis and growers have access to several data gathering devices. While the sensors deliver vital information about weather, plant and soil parameters as well as management activities, drawing conclusions from such a large amount of spatiotemporal data requires an agro-expert with cooperative knowledge.

However, the endless strings of temporally dense tabular data can be overwhelming and cumbersome to translate to meaningful and actionable information. Therefore, it is of high importance to develop new ways of expressing analyses and integrate findings into a highly visual form to help growers and most users of these large, potentially complex data systems.

To guide management practices, the use of cognitive computing and artificial intelligence (AI) has been of great interest in recent years. Although, when aggregating remotely sensed (spaceborne and airborne) and ground-based datasets, due to the temporal and spatial data density, the time and spatial scale of the region of interest, geospatial techniques need to be introduced. The blend of certain AI tools and geospatial methods led to the birth and evolution of GeoAI, an interdisciplinary concept. By leveraging AI and applied spatial technologies, the benefits of smart orchard monitoring can be taken to a significantly higher level.

AI is capable of recognising trends, identifying patterns, and relationships in complex multi-seasonal datasets and the produced information can be ingested into geographic information systems (GIS). Thus, the otherwise hidden interrelations between a high number of variables can be revealed. During the last few decades, cartographic maps dominated as standard GIS outputs. However, the commonly dynamic nature of environmental modelling has driven the adoption of web maps, real-time viewing platforms such as dashboards with mapping elements to most effectively present insights.

The concept of data driven horticulture aims to propose an ambitious attempt to utilise these modern tools to revolutionise New Zealand’s horticultural sector. To boost productivity and to improve environmental outcomes the on-orchard management and value-chain intelligence can be reformed by sensing more, optimising decision making and tailoring actions.

ABOUT THE AUTHOR(S)

Dr. Istvan Hajdu

Istvan Hajdu has built up a career applying his particular skill set to help public bodies and commercial organisations translate complex location-based information and find solutions to particular business challenges. Currently working as a Research Scientist at PlantTech Research Institute, he arrived with a wealth of experience in handling geospatial data and Geographical Information Systems (GIS), particularly the integration of GIS with artificial intelligence.

Istvan's own scientific career started in Hungary by completing a master's in geography and then he achieved a second master's in geological engineering. Istvan has a strong background in geospatial data sciences due to his various roles as GIS analyst, data and mapping officer in the United Kingdom. This knowledge was further extended by completing a Ph.D. within the Primary Growth Partnership programme, which explored soil water modelling in New Zealand's hill country.

After finishing his PhD thesis, as a spatial analyst at Massey University, he gained experience in implementing research and geospatial analytics in industry-driven applications by using data from airborne hyperspectral imaging.

Prof. Ian Yule

Ian Yule is the Research Director for PlantTech Research, he is an experienced leading researcher with a strong track record of working in higher education conducting industry relevant research and commercialisation. Skilled in Precision Agriculture and Agritech, Ian has spent the bulk of his recent career working on contract research in the areas of precision agriculture, agri-technology and remote sensing; with a particular focus on hyperspectral imaging and image analysis.

Ian has a strong commercial history. He is co-founder of Hyperceptions, a data processing company analysing aerial hyperspectral imaging data to deliver agricultural insights. Ian is also the founder of Stoneleigh Consulting, which operates both within New Zealand and internationally. They offer technical assistance and consultancy in the areas of Agritech, precision agriculture, agricultural development and technology.

Strong education professional with a BSc(Hons), MSc, PhD focused in Agricultural Engineering from University of Newcastle-upon-Tyne. Chartered Engineer, Fellow of the IAgrE, President the International Society of Precision Agriculture

NeSI Consultancies - Evolving a Scientific Programming Service

Wolfgang Hayek, Chris Scott, Alexander Pletzer, Maxime Rio
NeSI
wolfgang.hayek@nesi.org.nz, chris.scott@nesi.org.nz, alexander.pletzer@nesi.org.nz, maxime.rio@nesi.org.nz

ABSTRACT / INTRODUCTION

High Performance Computing (HPC) is an essential tool for modern research, underpinning a growing diversity of disciplines with simulations, data processing, and data analysis. Despite ongoing efforts to provide powerful and easy-to-use software packages and user interfaces, HPC and its efficient utilisation remains a fundamental challenge, often requiring specialist knowledge and expert assistance to enable ambitious research projects to achieve their goals. From experience, even small programming tweaks in research codes have resulted in performance improvements of many factors.

The NeSI consultancy service was created to address these issues and has delivered more than 6000 hours of scientific programming support to over 70 research projects during the last 3 years alone. It has been continuously refined to make it as easy and accessible as possible for the users, and it needs to keep evolving to accommodate the latest developments in technology, emerging user needs, and reach fields of science that have not used HPC before.

This presentation provides an overview of the types of projects and scientific disciplines that have used the service so far, highlights its recent expansion into data-driven work, and explores challenges that may transform the service in the future, to ensure that it continues to provide useful support and reaches as many researchers across New Zealand as possible.

ABOUT THE AUTHOR(S)

Wolfgang Hayek is a research software engineer at NeSI and NIWA, and group manager of NIWA's scientific programming group. Wolfgang has expertise in radiative transfer modelling, visualisation, data analysis, and high performance computing.

Chris Scott is a research software engineer for NeSI at University of Auckland. Currently lead of the computational science team, Chris has a background in molecular dynamics, Monte Carlo methods, finite element analysis, visualisation and parallel computing.

Alex Pletzer is a research software engineer for NeSI at NIWA. Originally a physicist, Alex drifted towards high performance computing during a career that spans research in plasma physics, working for a private company in Colorado, and supporting users at university in Pennsylvania.

Maxime Rio is a data scientist at NeSI and NIWA. He enjoys helping researchers to analyse their data, from visualisation to probabilistic modelling.

Staying ahead of the data deluge

David Honey
HPE
david.honey@hpe.com

ABSTRACT / INTRODUCTION

Because data is the new Intellectual Property, its care and preservation are more important than ever.

David's talk is an introduction to the latest data management software created by HPE for managing unstructured data at extreme scale; Data Management Framework version 7 (DMF7).

DMF7 integrates with popular parallel filesystems commonly used in HPC and AI environments. It is a tool that helps administrators implement site policies consistently across one or many filesystems. It can act upon a system event, like a filesystem dropping below a threshold for free space. It can operate on a schedule, like taking a snapshot of the namespace reflection at prescribed intervals. It can automate site policies, like staging datasets or migrating files that have been untouched for 30 days. Taken together, the tools for managing filesystems in DMF7 help customers lower the total cost of ownership (TCO) for high performance storage systems, increase administrator productivity, and lower the cost of storing data.

ABOUT THE AUTHOR(S)

The HPC Sales business unit is the HPC consulting and integration arm of HPE. HPE is successfully delivering customer consulting and projects with particular expertise in High Performance Computing, Visualisation and Complex Data Management. HPE is providing solutions to the Government, Manufacturing, Sciences, Communications, Entertainment and Resources sectors.

David is a Data Management Expert within the HPC team.

David's experience covers all aspects of the system life cycle from requirements analysis, infrastructure design, capacity planning, continuity planning, integration planning and benchmarking for new systems through to configuration management, change management and problem management for mature systems.

He has provided consultancy to clients on technology and architecture options, designing solutions in conjunction with multiple vendors and managing complex implementations.

David's knowledge covers TV broadcast and film post production, HPC, storage performance management with expert knowledge of data sharing architectures, hierarchical storage and data protection.

Prior to joining HPE, David managed ICT Infrastructure for Telecom NZ Ltd, worked as a business analyst for the Post Office and worked as a technical consultant for Mobil Oil. David has a Bachelor of Science in Physics from Victoria University of Wellington and is a certified PMI Project Management Professional.

Future of eResearch
(Oral Presentation)

Nick Jones
Director, New Zealand eScience Infrastructure (NeSI)
nick.jones@nesi.org.nz

ABSTRACT / INTRODUCTION

Over the last decade New Zealand has seen some major investments into eresearch infrastructures and capabilities.

This talk explores what bets we've placed, how has our thinking evolved, and what do we see elsewhere in the international landscape of national research e-infrastructure investments.

This session will open discussions to lead into a followup BoF / panel session that will explore where we are, and more importantly, where we might need to go next.

ABOUT THE AUTHOR(S)

Nick Jones, *Director, New Zealand eScience Infrastructure (NeSI)*
Nick Jones is NeSI's founding Director, having established and led NeSI alongside a team of colleagues and peers since inception in mid-2011. Nick is responsible for NeSI's strategic directions and performance overall, bringing together a talented and diverse array of people, and their institutions and interests. Nick has over 20 years' experience in innovating in advanced information/computing technology in sectors including education, science and research. Nick established the eResearch NZ conference series in 2010 to support the sector coming together in the spirit of community to share experiences and explore directions in an area so critical to our future prosperity as a nation.

Future of eResearch
(BoF / panel session)

Nick Jones, New Zealand eScience Infrastructure (NeSI)
Dianna Taylor, REANNZ
Jo Lane, University of Waikato

ABSTRACT / INTRODUCTION

Over the last decade New Zealand has seen some major investments into eResearch infrastructures and capabilities.

This panel will explore what we've learned over the initial decade of eResearch infrastructure investments in NZ and elsewhere internationally. The discussion will be used to prompt debate on where we are, and more importantly, where we might need to go next.

ABOUT THE AUTHOR(S)

Nick Jones, *New Zealand eScience Infrastructure (NeSI)*
Nick Jones is NeSI's founding Director, having established and led NeSI alongside a team of colleagues and peers since inception in mid-2011. Nick is responsible for NeSI's strategic directions and performance overall, bringing together a talented and diverse array of people, and their institutions and interests.

Dianna Taylor, *REANNZ*
Dianna joined REANNZ as Chief Executive Officer in August 2019. She was previously General Manager Technology/CIO at the New Zealand Racing Board. She brings over 20 years' experience in the banking and financial services industry, including in CIO and General Manager roles at Kiwibank.

Jo Lane, *University of Waikato*
Associate Professor Jo Lane is a computational chemist at the University of Waikato and is currently Deputy Dean for the School of Science. Jo obtained his BSc(Hons) and PhD from the University of Otago.

The path to growing New Zealand's clinical genomics capacity

Kemp, L., deLigt, J., Macartney-Coxson, D. and Benton, M.
Institute of Environmental Science and Research (ESR)
leah.kemp@esr.cri.nz

ABSTRACT / INTRODUCTION

Genetics plays an important role in many diseases afflicting New Zealanders. There is a shift internationally towards more high-throughput genetic testing. However currently in New Zealand, genetic testing routinely involves gene panels (testing multiple genes involved in a disorder the patient is suspected to have) or testing for variation in chromosomal structure and number. However, there is much more information available within a person's genome that has powerful diagnostic and treatment implications. This is where clinical genetics becomes clinical genomics!

In this talk, I will discuss how the Human Genomics group at ESR has been working in collaboration with the Capital & Coast District Health Board (CCDHB) and Genomics Aotearoa (GA) to grow New Zealand's clinical genomics capacity by adopting established practices and infrastructure used overseas. We built open-source analysis pipelines based on best practices in bioinformatics. These utilise workflow languages and portable software allowing the analysis of genomic data to be standardised New Zealand wide. Parts of these analysis pipelines run on cutting edge NVIDA GPU's in HPC environment that massively speed up the analysis. When demand for clinical genomics expands, we can scale up with more GPU's, the software is ready to adapt.

We envision a future where clinical genome sequencing can be done in New Zealand. With this 'last piece of the puzzle', we can go from sample collection to clinical interpretation in days instead of months or years, a crucial speedup for people living with undiagnosed disease. We have already had a potential diagnosis of an individual who was unable to be diagnosed through years of traditional testing. Stories such as this continue to spur us on to make clinical genomics and precision medicine accessible to all New Zealanders.

ABOUT THE AUTHOR(S)

- Leah Kemp
- I'm a bioinformatician at ESR working predominantly on analysing human genomic data. Although my background is in fisheries population genetics, New Zealand Trevally to be exact!

The Research Data Alliance – A Global Collaboration Forum to Tackle Data Challenges

Stefanie Kethers, Andrew Treloar)
Australian Research Data Commons
Stefanie.Kethers@ardc.edu.au, Andrew.Treloar@ardc.edu.au

ABSTRACT

The Research Data Alliance (RDA) [1], a community-driven initiative with the goal of building the social and technical infrastructure to enable open sharing and re-use of data, currently has over 11,000 members - including data scientists, librarians, researchers, funders, policy makers - from more than 140 countries [2]. RDA members come together through focused Working Groups and Interest Groups, formed by international experts from academia, the private sector and government, to collaborate on tackling global data challenges. To date, RDA Working and Interest Groups have produced about 50 Outputs, which have been adopted in over 100 settings [2].

RDA holds biannual Plenaries, highly interactive work meetings where members meet and network, advance the work of their groups, and propose or establish new groups. Due mainly (but not exclusively) to the COVID-19 pandemic, RDA Plenaries have moved from mainly physical meetings with some remote access to mainly virtual meetings with no, or very small, face-to-face components. This provides an opportunity, in particular for people for whom travel to Plenaries is difficult, to engage with RDA and participate in the RDA Plenaries, and in RDA more generally.

In this poster, we will present an overview of the outcomes, activities, and opportunities for engaging with and benefitting from RDA, including RDA outputs, adoption stories, and current data challenges being tackled by RDA groups. We will also outline different pathways to engaging with RDA, and will also look forward to the RDA Plenary 17, to be held from 20-22 April 2021, and a series of regional events associated with the Plenary, which should be beneficial to the New Zealand eResearch community.

References

[1] Berman, F., & Crosas, M. (2020). The Research Data Alliance: Benefits and Challenges of Building a Community Organization. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.5e126552

[2] Research Data Alliance (2020). RDA in a Nutshell (October 2020). https://www.rd-alliance.org/sites/default/files/attachment/RDA-in-a-nutshell-October-2020.pptx (retrieved 3 Dec 2020).

ABOUT THE AUTHOR(S)

- Dr Stefanie Kethers

Dr Stefanie Kethers is the Director of Operations of the Research Data Alliance and has been a member of the RDA Secretariat since before the RDA's launch in 2013. She has a strong interest in supporting cooperation and collaboration in the workplace, and has previously worked as a researcher on a variety of related projects, including investigating archival services for Koorie communities, researchers' data management practices and needs, and improving handover processes in hospitals.

- Dr Andrew Treloar
  Dr Andrew Treloar is the Director Platforms and Software for the Australian Research Data Commons. He was co-chair of the Research Data Alliance (http://rd-alliance.org/) Technical Advisory Board from 2013-2020. His twitter bio describes him as "Data-tragic, urban-greenie, home-gardener, cycling-commuter, BodyPump-addict, lapsed-linguist", which probably isn't a bad summary. He never seems to be able to make enough time for practising his 'cello or reading, but does try to prioritise talking to his chickens and working in his vegetable garden and orchard. Further details at http://andrew.treloar.net/ or follow him on Twitter as @atreloar.

Jupyter Notebooks for Absolute Beginners

Sara King

AARNet

sara.king@aarnet.edu.au

ABSTRACT / INTRODUCTION

This 3-hour online workshop will introduce you to Jupyter Notebooks, a digital tool that has exploded in popularity in recent years for those working with data.

You will learn what they are, what they do and why you might like to use them. It is an introductory set of lessons for those who are brand new, and have little or no knowledge of coding and computational methods in research. By the end of the workshop, you will have a good understanding of what Jupyter Notebooks can do, how to open one up, perform some basic tasks and save it for later. If you are really into it, you will also be able to continue to experiment after the workshop by using other people's notebooks as springboards for your own adventures!

This online workshop is targeted at those who are absolute beginners or 'tech-curious'. It includes a hands-on component, using basic programming commands, but requires no previous knowledge of programming.

Please check that you have access to www.cloudstor.aarnet.edu.au . If you do not have CloudStor access you can still attend, just advise the trainer beforehand so arrangements can be made.

ABOUT THE AUTHOR

- Dr Sara King
- Dr Sara King is the Training and Engagement Lead for AARNet. She is focused on outreach within the research sector, developing communities of interest around training, outreach and skills development in eResearch. She is currently working on creating reusable guidance information for Jupyter Notebooks and other AARNet services to be adapted for Carpentry training workshops. She is passionate about helping others develop the infrastructure and digital literacies required for working in a data-driven world, translating technology so it is accessible to everyone.

A Data-driven Cloud Classification Framework Based on a Rotationally Invariant Autoencoder

Takuya Kurihana, Elisabeth Moyer, Rebecca Willett, and Ian Foster
The University of Chicago
foster@uchicago.edu

ABSTRACT / INTRODUCTION

Advanced satellite-born remote sensing instruments produce high-resolution multi-spectral data for much of the globe at a close to daily cadence. These datasets open up the possibility of improved understanding of cloud dynamics and feedbacks, which remain the biggest source of uncertainty in global climate model projections. As a step towards answering these questions, we describe an automated *rotation-invariant cloud classification* (RICC) method that leverages deep learning autoencoder technology to identify cloud types within large datasets in an unsupervised fashion, free from assumptions about predefined classes. We describe both the design and implementation of this method and its evaluation, which uses a sequence of testing protocols to determine whether the classes that it identifies: (1) are separable (i.e., are cohesive in latent space and separated from each other), (2) are stable, i.e., produce similar or identical classes when different subsets of the data are used; (3) capture information on spatial distributions, such as textures; (4) are rotationally invariant, i.e., insensitive to the orientation of an image; and (5) are physically reasonable, i.e., embody scientifically relevant distinctions. Results obtained when these protocols are applied to RICC outputs suggest that the resultant novel cloud classes are appropriately spatially coherent and invariant to orientations of input images, and that they capture meaningful aspects of cloud physics. The next steps in the work will involve applying the method to several decades of data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Terra and Acqua satellites.

This use of unsupervised learning for cloud classification represents a departure from most previous work, which has applied supervised learning methods to label clouds according to predefined classes, and thus by definition cannot discover new cloud types.

This work is in progress, with excellent results obtained recently from the evaluation process described above; scaling up to larger datasets is the next step. Foster will be attending eResearch NZ 2021 and hopes to engage with other participants with interests in deep learning for image processing and/or climate science.

ABOUT THE AUTHOR(S)

The authors are in the departments of Computer Science (Foster, Kurihana, Willett) and Geophysical Sciences (Moyer) at the University of Chicago. Kurihana is a PhD student working with Foster as his advisor, in consultation with Moyer and Willett.

Interactive HPC Computation with Open OnDemand and FastX

Lev Lafayette, Sean Crosby
University of Melbourne
lev.lafayette@unimelb.edu.au, scrosby@unimelb.edu.au

ABSTRACT / INTRODUCTION

As dataset size and complexity requirements grows increasingly researchers need to find additional computational power for processing. A preferred choice is high performance computing (HPC) which, due to its physical architecture, operating system, and optimised application installations, is best suited for such processing. However HPC systems have historically been less effective at visual display, and least of all in an interactive manner, leading into a general truism of "compute on the HPC, visualise locally". This is primarily due to the tyranny of distance, but also with additional latency introduced by contemporary graphics when remote display instructions are sent to a local X-server. With a demand for both HPC computational power and interactive graphics, the University of Melbourne has implemented two technologies, FastX and Open OnDemand, on their general purpose HPC system "Spartan". This allows users to run graphical applications on Spartan, by submitting a job to the batch system which executes an XFCE graphical environment. In the Spartan environment FastX has been coupled with Open OnDemand which provides to web-enabled applications (e.g., RStudio, Jupyter Notebooks). In illustrating how this environment operates at the Spartan HPC system, the presentation will also illustrate recent research case studies from the University of Melbourne that have utilised this technology.

ABOUT THE AUTHOR(S)

- Lev Lafayette
    - Lev Lafayette is an HPC systems administrator and educator at the University of Melbourne, where has been for the past five years. Prior to that, he held a similar role at the University of Melbourne for eight years. He has also worked for the Ministry of Foreign Affairs (Timor-Leste) and the Parliament of Victoria, was active in Linux community development for over fifteen years. He collects post-graduate degrees for fun and profit and is currently studying at the University of Otago (his sixth degree) and the University of London, London School of Economics (his seventh).

  Sean Crosby

    - Sean is the HPC leader for the HPC team at the University of Melbourne.

ARCOS: Findings from a roadmap to establish a national capability for containers

Authors name(s): Dr Steven Manos
Organisation(s): University of Melbourne eResearch Centre, Melbourne, Australia
Authors Email(s): steven@biocommons.org.au

ABSTRACT / INTRODUCTION

Containers have been described as a computer within a computer, which allows researchers to run software without installation, saving time and effort. Kubernetes is the emerging open source project that has become the most commonly used approach for multicloud application deployment using containers.

The Australian Research Container Orchestration Service (ARCOS) seeks to emulate the transformational impact of the ARDC (OpenStack) Core Services competency by establishing a national collaborative Kubernetes Core Service to support the use of containers and container orchestration in the research sector. The ARDC's Storage and Compute theme are working with the research community to develop the requirements for ARCOS and a plan for implementing the service.

The primary objectives ARCOS are:

- To engage with Australian research infrastructure providers, developers and researchers to collate and understand their use of Kubernetes
- Collect information on how Kubernetes is being used internationally as an input into ARCOS
- Coordinate efforts and foster collaboration across research communities
- Bring together service providers from the academic and research community to establish and exchange best practice on the application of Kubernetes within the Australian research sector
- Make recommendations on national services and community activities
- Provide input into the implementation of a national Kubernetes Service for researchers.

During this virtual presentation we will outline the vision of ARCOS, provide an update of the current progress of the initiative, solicit feedback from the audience and encourage those interested in this ground-breaking initiative to participate in its development.

ABOUT THE AUTHOR(S)

- Dr Steven Manos
- Steven has 15 years of experience working at the intersection of research practice and digital technologies. He brings a mix of skills in facilitation, strategy and tech, and has a big focus on

partnerships and community building. His ambition is to deliver a more united national workforce of research support specialists providing valuable expertise and new services to the research community.

Starting an eResearch revolution with Deep Learning

Brent Martin and Aleksandra Pawlik
Manaaki Whenua Landcare Research
martinb@landcareresearch.co.nz, pawlika@landcareresearch.co.nz

ABSTRACT / INTRODUCTION

Background

Manaaki Whenua Landcare Research (MWLR), like most research institutes, both consumes and generates an ever-increasing amount of data. In particular, spatial data (images, hyperspectral data, spatial samples) are central to much of what MWLR does. MWLR has a strong track record of producing spatial data for consumption by research, including cleaned satellite imagery and GIS layers such as the Land Cover Database (LCDB)[1]. MWLR also holds many nationally and internationally significant physical collections of flora and fauna. As well as the physical samples themselves, the metadata associated with these specimens is invaluable for further analysis, such as species distribution[2].

In recent years, machine learning (ML) has increasingly matured from a branch of computer science to a respected tool in the researchers' toolbox. Most recently, deep learning has revolutionised computer vision, unlocking new opportunities to extract knowledge from images and other spatial data. For example, whereas ten years ago it was considered reasonable to be able to identify pollen grain species from images with 65-70% accuracy by *both humans and computers*, it is now straightforward to achieve accuracies exceeding 95% using deep learning[3].

Organically growing deep learning

At MWLR, we have begun a journey to dramatically increase the impact of our research by consuming/reconsuming data using machine learning, with a particular focus on deep learning. This is being achieved through a collaboration between the Informatics research team and the wider scientific cohort at MWLR. This process began with a small number of projects where the potential benefit was clear. Early results of these projects have been disseminated internally through webinars, leading to further projects being identified. In this initial stage, the emphasis is on rapidly achieving results and developing broad knowledge of tools and techniques. To date we have focussed on image classification through feature extraction[4], segmentation (U-net[5]) and object detection (Mask R-CNN[6]).

Through this initial exploratory phase, we have identified two fundamental barriers to uptake. First, researchers distrust "black box" models that do not add to our understanding of "why", and that cannot explain how they reach a conclusion. We are addressing this first concern by exploring how classification decisions can be visualised to show what contributed to the outcome. For example, we have shown that a deep learning model can classify beech pollen species from images to over 80% accuracy, a task considered too difficult for even specialist humans. Because the researchers have been sceptical of this outcome, we

have used occlusion sensitivity visualisations to demonstrate that for correct classifications the deep learning network is focussing on expected areas of the image, such as the pollen grains' edge or texture, unlike for incorrect classifications. We are now investigating whether similar techniques exist that are suitable for image segmentation and object detection tasks. For segmentation problems, we can also manually investigate differences between the training data and predictions; in some cases the error may be inaccuracies in the training data, highlighting the potential for deep learning models to augment manual processes as a further benefit.

The second barrier to uptake is the quantity and quality of data needed. For image classification tasks, we have developed a novel method of utilising deep learning models for feature extraction that dramatically reduces both the number of training examples required as well as the processing requirements[7]; we have successfully built species identification models with good accuracy from as little as a few hundred images for domains such as fungal spores, coprosmas, moths and beech pollen. For segmentation tasks, we are experimenting with methods for bootstrapping imperfect training data through an iterative process of training weak models and using them to refine the training data with some additional manual correction where required[8]. We are experimenting with this technique for identifying tree species from UAV orthomosaics where the class polygons are weakly inferred from tree stem positions obtained through ground-based surveys, and then subsequently refined based on the segmentation suggested by the model. It is hoped that such techniques will dramatically lower the effort required to build training sets for such tasks, increasing the value obtained from localised ground surveys by using the data to make inferences at regional or national scale. Finally, we are also exploring the impact of resolution on accuracy, to quantify the limits of scaling up small-scale surveys to be repeatable at the national level from more easily available spatial data such as hyperspectral satellite imagery.

We have so far identified 12 projects, half of which are being actively pursued. We have also organised an internal "mini-symposium" which will present two case studies, as well as discussing machine learning and deep learning techniques. A "panel" session will then discuss further potential project ideas submitted by the audience. This approach has been successful in engaging further researchers; to date six further projects have been identified ranging from counting manuka flowers in images to extracting text from historical documents, and it is anticipated that the panel discussion will generate significant further interest.

Conclusion

As the amount of data available rapidly increases, we need increasingly sophisticated tools to take advantage of the opportunities it presents. Deep learning has dramatically increased the potential to make use of imagery and other spatial data. By automating previously time-consuming tasks, from segmenting satellite images by land cover to counting specific pollen grains on microscope slides, there is an opportunity to greatly increase the impact of the science we perform. While demonstrating this potential through exploratory projects at Manaaki Whenua, we have observed that the scientific community is ready and willing to enter the data-driven world of eResearch. The next steps are to grow our support services and tools to lower the effort required for researchers new to the field to be able to quickly experiment with machine learning/deep learning techniques so they can develop exciting new areas of data-driven research.

References

- https://lris.scinfo.org.nz/layer/104400-lcdb-v50-land-cover-database-version-50-mainland-new-zealand/

1. I. Bartomeus, J. R. Stavert, D. Ward and O. Aguado (2018): Historical collections as a tool for assessing the global pollination crisis, Philosophical Transactions of the Royal Society B: Biological SciencesVolume 374, Issue 1763

2. Sevillano V, Holt K, Aznarte JL (2020): Precise automatic classification of 46 different pollen types with convolutional neural networks. PLoS ONE 15(6): e0229751. https://doi.org/10.1371/journal.pone.0229751

3. Liang, H., Sun, X., Sun, Y. et al (2017). Text feature extraction based on deep learning: a review. J Wireless Com Network 2017, 211. https://doi.org/10.1186/s13638-017-0993-1

4. Ronneberger, Olaf; Fischer, Philipp; Brox, Thomas (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597

5. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969

6. V. Vetrova, S. Coup, E. Frank and M. J. Cree (2018). Hidden Features: Experiments with Feature Transfer for Fine-Grained Multi-Class and One-Class Image Categorization. 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), Auckland, New Zealand, pp. 1-6, doi: 10.1109/IVCNZ.2018.8634790.

7. Weinstein, B.G.; Marconi, S.; Bohlman, S.; Zare, A.; White, E. (2019). Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks. *Remote Sens.* 2019, *11*, 1309.

ABOUT THE AUTHOR(S)

Brent Martin

Brent is a 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. Brent holds a PhD in Computer Science from the University of Canterbury, New Zealand focussing on artificial intelligence in education.

Aleksandra Pawlik

Aleksandra is an eResearch capability specialist at Manaaki Whenua Landcare Research, where she is assisting with the development of strategy, procedures and tools that promote data-driven science and research data management. She also organises and instructs workshops assisting researchers to develop their research data skills. In her career Aleksandra has been active in the UK’s Software Sustainability institute, where she led the institute’s training activities. Outside of academia, Aleksandra has worked as a Research Community Manager for the New Zealand eScience Infrastructure (NeSI), as a researcher for NHS Lothian projects and as a freelance IT consultant in the commercial sector. She is also an instructor for the Software Carpentry Foundation. Aleksandra holds a PhD in Computing from the Open University focussing on documentation in scientific software.

AUNZ AI4LAM – Research Collaboration

Ingrid Mason, Alexis Tindall & Adam Moriarty
Independent, University of Adelaide, Auckland War Memorial Museum
ingrid.b.mason@gmail.com, alexis.tindall@adelaide.edu.au, amoriarty@aucklandmuseum.com

ABSTRACT / INTRODUCTION

This presentation announces (1) the emergence of the Australian and New Zealand regional chapter of the AI4LAM as a Trans-Tasman community of practice; (2) highlights an opportunity for data and computer science research and eResearch; and (3) it is a call out to computer and data scientists in New Zealand and Australia.

Building working bridges between data custodians and research communities across sectors and jurisdictions requires a design, plan, and people that sense they have a shared interests and needs in common. In 2019 an initiative kicked off to set up a regional chapter AUNZ (Australia and New Zealand) of the AI4LAM (Artificial Intelligence, for Libraries, Archives and Museums) community commenced. The grassroots initiative begin the process of community engagement, putting out the welcome mat, and inviting people to pitch in, help coordinate, share their experiences, knowledge and expertise. The AUNZ chapter of AI4LAM has a charter guiding community work on collaboration, ethics and raising knowledge and skills is targeted at GLAMR (Gallery, Library, Archive, Museum and Recordkeeping) professionals and humanities researchers.

Opportunities are emerging to explore computation and analytics techniques and AI in digital heritage and digital humanities research practices in partnership with computer and data scientists. In computer and data science literature there have been extraordinary inroads in the development of new computation and analytics techniques and the application of this knowledge and skill is already driving change in research and industry. This expertise is now aiding and informing the transformation of cultural heritage and humanities research practices. The people represented by the regional chapter are passionate about cultural heritage and the humanities, and AI4LAM is working to increase awareness and accessibility of these new approaches. We anticipate that bringing computer scientists, humanities researchers, data scientists, and GLAMR professionals together on equal footing will lead to fantastic research collaboration, in artificial intelligence and machine learning, robotics, information systems, systems engineering, analytics and visualisation.

This is a call from the community that have formed around the AUNZ chapter of AI4LAM to the computer and data scientists in New Zealand and Australia – to walk and work with us.

ABOUT THE AUTHOR(S)

- Ingrid Mason, eResearch and Heritage Consultant. Ingrid drives change in the digital transformation of humanities research and cultural heritage through the development of new technologies and national infrastructure.
- Alexis Tindall, Manager, Digital Innovation, University of Adelaide Library. Alexis works to support data-enabled humanities and arts research in project management and training roles. She is passionate about digitisation, open scholarship and digital access to GLAM collections and research data.
- Adam Moriarty, Head of Collection Information and Access. Adam works to open the multidisciplinary collections of the Museum online in a way that their stories can be freely accessed and shared. He and his team have been experimenting with AI to understand how to make GLAM collections more discoverable and accessible.

Deploying secure and scalable VPN connections with eduVPN

Vladimir Mencl
Research and Education Advanced Network New Zealand
vladimir.mencl@reannz.co.nz

ABSTRACT / INTRODUCTION

Since early 2020, working from home has become the new social norm. Employees need to be able to work from home at ideally the same level of productivity as when working from the office – and for that, will need access to resources which are often accessible only from the office network. The access is typically facilitated with a Virtual Private Network (VPN), providing a secure connection from the user's computer to the office network.

Most organisations have VPN solutions in place, but often, these were not designed for the situation where most or almost all staff are working from home, and did not cope well with the sudden increase in demand.

This is partly because hardware VPN appliances have longer replacement cycles, and often lag behind the refreshes (and throughput increases) of other network equipment.

The eduVPN software package (also known as Let's Connect) is an open-source VPN solution, covering both the server side and the client side, designed by the R&E community for the R&E community. An eduVPN server can be easily deployed on a Linux server.

During the lockdown in early 2020, we have been approached by a member institution as their VPN solution was not coping with the demand, especially for large data sets transfers.

We have deployed a managed eduVPN server for this institution – and even though it was deployed on an ordinary VM within REANNZ infrastructure, it significantly outperformed the existing VPN solution and allowed the staff of this organisation to work from home, including research data set transfers that were not possible with the existing VPN solution.

This talk will present both the client and server side of eduVPN, demonstrate their ease of use and deployment, describe the key use cases that can be addressed by eduVPN, and outline what deployment options are available.

ABOUT THE AUTHOR(S)

Dr. Vladimir Mencl has been part of the New Zealand R&E community since 2006 and has been involved in identity and access management projects since the early days of the BeSTGRID project. When the Tuakiri project moved to REANNZ, Vlad joined REANNZ where he is part of the Systems team as a Senior Software Engineer.

ORCID Data Dashboard: A Snapshot of NZ Research

Brian Minihan
ORCID
b.minihan@orcid.org

ABSTRACT / INTRODUCTION

ORCID has provided a service to individual researchers by providing a disambiguative identifier and to grant funders, publishers and research organisations by supporting an authoritative registry of these researchers works as well as trusted connections to that registry via API. However, obtaining useful data from the ORCID registry was previously limited to complex API queries, which was often inaccessible and cumbersome to many parties. In 2020, ORCID endeavoured to level UP by creating a Data Dashboard with useful statistics to ORCID consortia lead organisations, such as Royal Society Te Apārangi. This oral presentation will introduce the Dashboard, examine lessons learned about ORCID metadata and illustrate aspects about the current state of research in New Zealand through connections to ORCID.

Acknowledgements:

Tom Demeranville, Product Director, ORCID

ABOUT THE AUTHOR(S)

- Brian Minihan
  Based in Hong Kong, Brian is a point of contact for expanding ORCID member organisations in South Asia and Hong Kong, supporting existing ORCID Consortia in Oceania and conducting scholarly communications outreach and events throughout the Asia Pacific. Before joining ORCID, Brian was a scholarly communications librarian at Hong Kong Baptist University and a Research Associate at University of San Francisco's Ricci Institute of Chinese-Western Cultural History.

Learning and growing from our mistakes and charting a new course for a better future

Jymal Morgan, Donia Macartney-Coxson, et al TBC

Organisation(s) Institute of Environmental Science and Research, NZ, and TBC

Authors Email(s) Jymal.Morgan@esr.cri.nz; Donia.Macartney-Coxson@esr.cri.nz and TBC

ABSTRACT / INTRODUCTION

ESR [a Crown Research Institute] and a Community group will traverse their shared experiences associated with a lived experience of when research goes wrong and how they are collectively working towards building the bridge towards a better future. When information/data 'leaves' the community – it is still theirs, does the data really leave? – learnings and experiences in this space.

How we envisage the presentation – if the abstract is accepted
An informal live discussion between an ESR scientist, and a community group member with a facilitator – all sat on a couch. While this could be a stand alone talk we think it would work really well coupled to a Data Sovereignty talk i.e. data sovereignty talk first (suggested speaker Ben Te Aika) followed by this discussion piece.

Please note: if this abstract/concept is accepted for oral presentation at eResearch 2021 we will provide a more detailed abstract and bios for the authors. Please do not publish/make publically available as is.

The Australian BioCommons Community Engagement Strategy: Understanding Community-scale Challenges to Inform Solution Delivery

Tiff Nelson[1], Andrew Lonie[1], Johan Gustafsson[1,2], and Jeff Christiansen[1]

[1]Australian BioCommons, [2]Bioplatforms Australia

tiff@biocommons.org.au

andrew@biocommons.org.au

johan@biocommons.org.au

jeff@biocommons.org.au

ABSTRACT / INTRODUCTION

Background

The Australian BioCommons develops digital capacity, training, and bioinformatics infrastructure to support Australia's life scientists. So how can we identify the greatest needs of many thousands of geographically dispersed researchers, and also deliver useful infrastructure? Strong user engagement is paramount to understand community needs and direct the deployment and resourcing of appropriate infrastructure to ensure maximum impact.

Method

We have developed a five-step process of engagement that maximises community interaction, from initiation to deployment.

Results

1/ Identify meaningful communities of manageable scope around focus areas with infrastructure challenges, such as genome assembly and annotation of non-model organisms, and microbiome analysis;

2/ Research the community topic area to understand broad needs and challenges to engage members;

3/ Communicate with the broad community, inclusive of everyone from any expertise level or any institution, to identify issues, roadblocks, and solutions/suggestions through electronic surveys, shared discussion boards, and virtual meetings;

4/ Document the challenges and, in discussion with infrastructure specialists, detail conceptual solutions with an endorsement from a subset of practitioners from the community, formalised in an Infrastructure Roadmap, which forms a blueprint for solutions that can be deployed to address the community challenges; and,

5/ Deploy and implement solutions with our infrastructure partners, with testing and feedback from the community.

Conclusion

Through this engagement process, the Australian BioCommons has identified and then coordinated work to deploy essential infrastructure that was previously lacking to support critical communities (e.g. those undertaking genome annotation). Successful outcomes are measured by positive responses from the community (e.g. turning up in large numbers, actively joining the discussion), and active use by early adopters, and uptake of deployed services that have been identified by various communities

The consultation method is now being applied to engage a diverse range of communities.

ABOUT THE AUTHOR(S)

Dr Tiff Nelson is the researcher community engagement officer with the Australian BioCommons. Coming from a background in microbiome analyses, Tiff uses her domain-specific knowledge to engage researchers to identify and document challenges for community-backed infrastructure solutions.

Associate Professor Andrew Lonie is the Director of the Australian BioCommons and A/Prof in the Faculty of Medicine, Dentistry and Health Sciences, and the School of Computing & Informatics Systems at the University of Melbourne.

Dr Johan Gustafsson is the Bioinformatics Engagement Officer at the Australian BioCommons. Johan works closely with members of the Bioplatforms Australia Framework Initiatives, as well as the BioCommons communities of practice: to identify bioinformatics requirements and subsequent opportunities for developing fit-for-purpose solutions.

Dr Jeff Christiansen is the Associate Director of Engagements and Operations with the Australian BioCommons and brings expert biological knowledge from a molecular biology research background and over 20 years of hands-on experience distilling researcher's challenges and then building fit-for-purpose and easy to use data-centric solutions for various communities of life science researchers.

GigaByte: A New Workflow for Rapid Dissemination of Datasets & Tools- Bringing Papers to Life

Nicole A Nogoy, PhD
Executive Editor,
GigaScience Press
nicole@gigasciencejournal.com

ABSTRACT

In the era of data-intensive research and the COVID-19 pandemic, the importance of open, and rapid dissemination of research has drawn greater attention. With this comes challenges, such as balancing speed with the quality of peer review, and the requirements of archiving a traditionally static “version of record” being difficult to balance with the advantages of web-based interactivity. Aiming to tackle these challenges and to speed up the publishing process using data infrastructure from a non-profit research organisation in Hong Kong, GigaScience Press has been helping develop a new custom-built platform and workflow using end-to-end publishing technology that enables accepted manuscripts to be converted to an online and PDF-ready article within a day.

GigaByte is a new journal that speeds up the publishing process by using new custom-built, end-to-end publishing technology. A streamlined editorial effort enhances speed by focusing on publishing short-format data and software-centric articles, which greatly reduces researcher writing time. And in addition, a questionnaire-style peer review process ensures rapid review with a focus on reuse potential over narrative value. The publication process integrates with the open GigaScience Database (GigaDB) that serves as a broad-spectrum repository to display data and tools associated with these publications with the added benefit of curators being on hand to help the submitter rapidly curate metadata and host data files. Papers are brought to life by embedding numerous "widgets" in the manuscript and dataset pages. This custom-built infrastructure and workflow has enabled *GigaScience* Press to move beyond the traditional static, descriptive journal article by including embedded content.

ABOUT THE AUTHOR

Nicole Nogoy is the Executive Editor at GigaScience Press currently based in Wellington, NZ. A research scientist by training with a degree in Biochemistry and Physiology from Victoria University of Wellington (NZ), and a PhD in Natural Sciences from the University of Goettingen (Germany), Nicole moved into publishing and was the launch and Managing Editor at *Genome Medicine* (BioMed Central). Having seen the light of open access and open science, Nicole is an open data and open science advocate with over 10 years of open access and STEM Publishing experience.

REANNZ Update – Moving data, mobilising knowledge

Carl Olsen
REANNZ
carl.olsen@reannz.co.nz

ABSTRACT / INTRODUCTION

REANNZ's mission is to design, build and operate New Zealand's high-performance network and provide member's with network services that enable them to connect and collaborate with each other and the global NREN ecosystem.

At a high level, what does this mean for the research and education ecosystem in New Zealand?

During this talk Carl Olsen, Member Engagement Manager at REANNZ, will cover the fundamental aspects of what makes up the REANNZ network. He will introduce the people in the team that work to support and enable members with connectivity solutions and collaboration opportunities within the wider research and education community. The session will also look at the core network that connects members nationally and the subsea cable system that connects New Zealand to the international network of global NRENs (National Research and Education Network). REANNZ acts as a part of the science and research system in New Zealand, connecting people, knowledge and capability to support developing ideas and innovations. The talk will look at use cases and examples of how the team enables research outcomes via the high speed network and demonstrate how members use the network.

ABOUT THE AUTHOR(S)

Carl Olsen

Carl leads the engagement team at REANNZ joining from Kiwibank where he focused on building strong, high performing teams to deliver better outcomes for customers and the organisation. His focus is to build deeper relationships with members to understand the unique needs of science and data intensive research, to ensure that members have access to the networking capabilities, people and resources they need.

Automating the grunt work out of epigenetics research

Brook O’Reilly
University of Canterbury
brook.oreilly@pg.canterbury.ac.nz

Dr Amy Osborne
University of Canterbury
amy.osborne@canterbury.ac.nz

Dr Miles Benton
Institute of Environmental Science and Research (ESR)
miles.benton@esr.cri.nz

ABSTRACT / INTRODUCTION

One of the perennial problems encountered in large research projects is making sense of a number of different data sources. An ongoing project aims to make use of data from previous cohort studies and combine this with insights derived from online databases to find associations in gene expression that haven’t been explicitly identified in past studies. This is achieved via a software pipeline which can take data from numerous sources and process it into a format through which hypotheses can be tested and insights derived. Part of this involves an automated internet search tool which can look for information in institutional databases or via search engines and use this to either direct pipeline activities or provide supplementary information to researchers. With a plethora of information available online, automated approaches present an attractive opportunity to increase the speed and efficiency of scientific research.

The focus of this project so far has been the study of DNA methylation, which is one of a number of epigenetic alterations, which are known to alter gene expression without changing the underlying genomic sequence. We can assess DNA methylation using a number of different tools, including whole-genome methods which can essentially see all 28 million possible DNA methylation sites in entirety, or the more-common methylation array technology which only looks at a subset (typically ~450,000 or ~850,000) of the possible locations where methylation can occur. The use of array technology, such as that employed by Illumina in their EPIC platform, is becoming widespread owing to its relatively low cost and fast throughput. Epigenome-wide association studies (EWAS) tend to focus on the relationship between epigenetic marks and specific traits, but researchers will often make the recorded epigenetic data available online in either processed or raw format. One of the key motivators for this project was the fact that uploaded data can be used to support findings unrelated to the study for which the data was originally produced.

Combining data from different sources poses a significant difficulty, especially in the context of DNA methylation – a number of different normalisation and processing methods are available and the calculated methylation intensities can vary substantially based on which of these methods were used. The research component of this project aims to counteract this by testing hypotheses with comparative methods, rather than directly assessing methylation intensity.

Additionally, an automated internet search and analysis application was built to scour online databases for information and insights – these are either directly relevant to the projects findings or are used to direct further pipeline activities. Though this automated approach to research isn’t perfect, it can yield a good idea of the big picture and identify potential avenues of interest significantly faster than what would be achievable by manually chasing up every lead. This can provide a boon to smaller groups that can’t justify expending resources on ‘rabbit holes’ that might not be immediately useful, and also may provide benefits to larger research programmes looking to understand the broader scope of findings as they come in.

ABOUT THE AUTHOR(S)

Brook O’Reilly is a MSc student, finishing up his thesis through the University of Canterbury. Formerly a systems and software engineer in the medical device and healthcare informatics industries, he undertook a masters in biological sciences with the intention of applying an engineering approach to the problems faced by researchers today. His scientific work is primarily based on analysis of patterns of DNA methylation in the human genome, and when not working on research or freelancing, he has been known to enjoy teaching applied data science and machine learning at UC.

Dr Amy Osborne is a Senior Lecturer in Genomics at the University of Canterbury.

Dr Miles Benton is a Senior Scientist (Computational Genomics) at ESR.

The role of temperature control in the kiwifruit cool chain revealed by Artificial Intelligence

Alvaro Orsi
PlantTech Research Institute Limited
alvaro@pri.co.nz

ABSTRACT / INTRODUCTION

Zespri, the world largest marketer of kiwifruit, exports to more than 50 countries, aiming to ensure that the product meets the different offshore market quality standards. After harvesting, many different stages in the cold chain, including temperature control during shipping, can affect the fruit quality in different ways. Despite the huge variety of factors there is limited sampling of fruit in pallets making it challenging for Zespri to identify the key stages that affect fruit quality and thus address them appropriately. This work is motivated by the need to develop a data-driven solution for this business challenge, which will allow Zespri to improve their decision process to improve the kiwifruit quality that reaches all offshore markets.

Previous attempts at understanding how the different stages in the cool chain can impact fruit quality had limited success. A full picture based on quantitative, data-driven analysis is yet to emerge. Here I present a novel approach to address this problem. Making use of data from the 2018, 2019 and 2020 seasons I incorporate state-of-the-art machine learning models, time series analysis, Bayesian Networks and a mechanistic fruit softening model to reveal how the different stages in the cool chain impact fruit quality.

I demonstrate that it is possible to link the fruit quality outcomes in offshore markets in terms of cold chain properties in its different stages, and also identify the key cold chain properties that are most responsible for affecting fruit quality metrics.

The outcomes of this research allow Zespri to plan to allocate resources efficiently in the cold chain stages that are most critical for the market/fruit variety/period of interest to address fruit quality outcomes. More importantly, it constitutes the backbone knowledge over which Zespri can develop a new decision process for its Vessel Assessment Committee, using a robust data-driven quantitative foundation.

ABOUT THE AUTHOR(S)

- Alvaro Orsi. Principal Research Scientist at PlantTech.
  More than 12 years of experience in Scientific computing, AI and Machine learning applied to address both scientific and data-driven industry challenges. The former by pursuing an academic career as a Computational Cosmologist in research centres in the UK, Chile and Spain. The latter

since 2019 as a Principal Scientist at PlantTech in New Zealand, where I find solutions for the horticulture industry through artificial intelligence technology.

I have published over 50 articles in top scientific astrophysical journals, and mentored several postgraduate students (Masters and PhDs). I have also led scientific groups in large international collaborations and organised international scientific meetings.

Who needs research software engineers?

Alexander Pletzer[1], Nooriyah Lohani[2], Chris Scott[2] and Georgina Rae[2]
[1]NIWA/NeSI, [2]University of Auckland/NeSI
alexander.pletzer@nesi.org.nz

ABSTRACT / INTRODUCTION

Research Software Engineers (RSEs) combine the knowledge of good software engineering practice with a degree of subject matter expertise (e.g. from physics, genomics, chemistry, applied mathematics or biology). The role of a RSE complements that of the scientist in ways that allows the latter to concentrate on their research while reaping the benefit of recent advances in digital technologies and best practices. A key difference between scientists and RSEs is that the former tend to work in a single research area for many years while the latter often move from one research area to another on a monthly or yearly pace, enabling cross-pollination of tools, methods and ideas. This, and other factors, will determine whether the interaction of RSEs with scientists will be fruitful. Through examples taken from NeSI's consulting activities we will attempt to sketch the ideal RSE profile and outline a set of conditions which will increase the likelihood of a successful collaboration between RSEs and researchers.

ABOUT THE AUTHOR(S)

- Alex Pletzer helps researchers run better and faster on NeSI platforms when he's not playing ping-pong with colleagues or windsurfing around Wellington

Just Add Context: How you can level up your analysis with Dimensions on Google BigQuery

Danu Poyner and Jared Watts
Digital Science
d.poyner@digital-science.com, j.watts@digital-science.com

ABSTRACT / INTRODUCTION

Levelling up means doing things you haven't been able to do before. To reach New Zealand's strategic goals for research impact, we're going to have to unlock new skills and capabilities – in ourselves, and our organizations. This means empowering a diverse workforce, and finding unique ways to contribute within the research sector. These types of goals require insight and evidence to translate into plans of action – but the contextual information required isn't always easily accessible. In this talk we will share two case studies, that demonstrate how you can use Dimensions, the world's largest and most diverse database of research insight, to help you provide the necessary strategic insight across your organisation. The first case study demonstrates how joining Dimensions data with your internal institutional data can support efforts to enable a highly skilled, inclusive and diverse research workforce. The second case study demonstrates how global contextual analysis can level up understanding of your institution's true research strengths and expertise, so you can identify unique opportunities to contribute to the research environment. By enabling direct access to the underlying data via Google BigQuery, you can now leverage Dimensions' strengths in contextualising research activity while connecting to and levelling up your existing reporting and analysis infrastructure to meet your organization's requirements.

ABOUT THE AUTHOR(S)

- Danu Poyner | Dimensions Product Specialist, Asia-Pacific

  Danu has worked in research management roles in Australian and New Zealand universities, focussing on research systems, performance and policy. As a Dimensions Product Specialist based in Auckland, he supports academic institutions throughout the region with data-driven approaches to analysing and improving research performance and impact.

- Jared Watts | Lead Developer, Dimensions Data Platform

  Jared Watts is the Lead Engineer Dimensions Data Platform and BigQuery project. Jared came to Digital Science from the University of Auckland, where he had worked for the past 10 years in library, research information and central information technology roles. Jared has experience managing software development teams, translating end-user requirements into quality software and architecting reliable and dependable solutions. A true polyglot of programming languages, Jared is always ready to learn a new language or technology.

Galaxy Australia - an analysis, upskilling and collaborative platform for the life sciences

Gareth Price[1], Simon Gladman[2] and Andrew Lonie[3]

1 - QCIF Facility for Advanced Bioinformatics, Institute of Molecular Biology, University of Queensland, Australia
2 - Melbourne Bioinformatics, Faculty of Medicine Dentistry and Health Sciences, University of Melbourne, Australia
3 - Australian BioCommons, Faculty of Medicine Dentistry and Health Sciences and School of Computing and Information Systems, University of Melbourne, Australia

g.price@qcif.edu.au; simon.glagman@unimelb.edu.au; andrew@biocommons.org.au

ABSTRACT / INTRODUCTION

The global data analysis platform - Galaxy - has evolved to meet the changing needs of life scientists, over its 15 year existence. The platform allows researchers to analyse data with clear and persistent meta-data captured for all their processing steps, allows for easy collaboration via controlled sharing and makes an ideal platform for the continual need to train researchers on the latest analytical workflows available on any one of the global publicly accessible usegalaxy.* services. Australia runs one of the three primary global usegalaxy services - Galaxy Australia (https://usegalaxy.org.au) and in the 3 years of national operation has shown steady growth in total and active users (total users - Dec 2020 - 10,600) as well as number of tool submissions, in the order of 20-30% growth per year. The success of the platform has come with challenges to resourcing the operational team, the underpinning infrastructure, governance and support, all of which have only demonstrated the national criticality of the platform to Australian researchers and their overseas collaborators. Through careful planning and community engagement Galaxy Australia plans reach to 2025 and beyond in a clear program of work to continue to support accessible, reproducible, and transparent computational biological research. This presentation will show the journey taken and planned for the platform and how researchers can plan to use it for many years to come.

ABOUT THE AUTHOR(S)

Dr Gareth Price is Head of Computational Biology at QCIF Facility for Advanced Bioinformatics, with 20 years' experience as a Bioinformatician and Genomics Scientist. In this role Gareth manages the diverse spectrum of researcher lead questions involving genomic data, provides training in genomic data analysis, as well as leading Galaxy Australia as Service Manager.

Simon Gladman is a senior bioinformatician at Melbourne Bioinformatics at the University of Melbourne. He has over 12 years experience in bioinformatics - mostly in microbial genomics, infrastructure development and bioinformatics training. Simon is currently the lead Engineer of Galaxy Australia, a free, publicly available, web-based platform for life sciences analysis.

Associate Professor Andrew Lonie is Director of the Australian BioCommons and Associate Professor in the Faculty of Medicine, Dentistry and Health Sciences, and School of Computing and Information Systems, University of Melbourne. The Australian BioCommons is a national infrastructure initiative to build the computational systems, expertise and training that Australia's life science researchers need to be globally competitive in the age of digital biology.

Galaxy Australia Workshop - showcasing a Leveling Up data analysis and training platform

Gareth Price[1], Simon Gladman[2]

[1]*QCIF Facility for Advanced Bioinformatics, Institute of Molecular Biology, University of Queensland, Australia*
[2]*Melbournre Bioinformatics, Faculty of Medicine, University of Melbourne, Australia*
g.price@qcif.edu.au; simon.glagman@unimelb.edu.au

ABSTRACT / INTRODUCTION

The Galaxy platform enables accessible, reproducible, and transparent computational biological research. The usegalaxy.* global community of Galaxy servers has enabled researchers globally to perform complex analyses and also importantly to upskill in best practice methodologies through their local usegalaxy.* platform and the Galaxy Training Network (https://training.galaxyproject.org/). Galaxy Australia has recently deployed a Training Infrastructure as a Service (TIaaS) to power larger and more ambitious workshop events. This workshop will take advantage of the Australian TIaaS to showcase how using the Galaxy platform allows researchers to rapidly upskill in over 15 different life science communities of practice, covering 100 plus topics. The workshop will cover the infrastructure supporting Galaxy Australia and how this scales to support analytical demands from single core to high-memory cluster jobs, streamlined data ingest functionality (both files and metadata) and how the TIaaS enables efficient face-to-face and virtual training workshops. Finally, the workshop will highlight the broadening fields of research support enabled through Galaxy such as proteomics, metabolomics, ecology and climate modelling.

Availability and Requirements:

* Project home page: https://usegalaxy.org.au

* Operating system(s): Linux, Windows, Mac OS X – requirement: internet browser only

ABOUT THE AUTHOR(S)

Dr Gareth Price is Head of Computational Biology at QCIF Facility for Advanced Bioinformatics, with 20 years' experience as a Bioinformatician and Genomics Scientist. In this role Gareth manages the diverse spectrum of researcher lead questions involving genomic data, provides training in genomic data analysis, as well as leading Galaxy Australia as Service Manager.

Simon Gladman is a senior bioinformatician at Melbourne Bioinformatics at the University of Melbourne. He has over 12 years experience in bioinformatics - mostly in microbial genomics, infrastructure development and bioinformatics training. Simon is currently the lead Engineer of Galaxy Australia, a free, publicly available, web based platform for life sciences analysis.

Building Partnerships for eResearch

Georgina Rae

New Zealand eScience Infrastructure

georgina.rae@nesi.org.nz

ABSTRACT / INTRODUCTION

NeSI's purpose is to build the High Performance Compute (HPC) capability of New Zealand researchers - to build skills, tools, ambitions and communities. NeSI has chosen to deliver on this goal through a network of collaborative relationships, or 'partnerships'. Partnerships could be with researchers, research groups and organisations.

In this session I will outline why we value partnerships and describe the richness of NeSI's partnership landscape using case studies of some current partnerships. I will then delve into some of the challenges and lessons learned of this relationship-based approach, including Collective Impact, choosing where to focus our energy and share some of the tactics we use to keep track.

ABOUT THE AUTHOR(S)

Georgina 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.

Regional Downscaling of Climate Data using Deep Learning and Applications for Drought / Rainfall Forecasting.

Neelesh Rampal1, Abha Sood1, Stephen Stuart1, Maxime Rio1,2 and Alexander Pletzer1,2

1National Institute of Water and Atmospheric Research, New Zealand

2 New Zealand eScience Infrastructure (NeSI)

neelesh.rampal@niwa.co.nz

ABSTRACT / INTRODUCTION

Dynamical downscaling of global climate model (GCM) data from ~150 km to 12 km resolution or smaller requires running a computationally expensive regional climate model (RCM) using GCM forcing data at the lateral and surface boundaries. In addition, RCM output data (temperature, precipitation, etc.) are biased with respect to in-situ observations due to various physical processes that are not adequately represented in the model, sometimes due to sub-grid scale effects. Methods in machine learning have the potential to extract more abstract relationships between in-situ observations and GCMs in addition to using RCM simulation reference data, and thus could improve the representation and accuracy of downscaled variables. In particular, precipitation is notoriously difficult to predict due to complex sub-grid scale processes and local features such as orography. In this study, we explore and test different machine learning approaches with the aim of improving the accuracy of regionally downscaled GCM output.

To test the effectiveness of methods in machine learning, we downscaled a variety of regional circulation indices to monthly rainfall anomalies (mm/day) for a single location (Whenuapai, Auckland). The circulation indices used were the M1 and Z1 Trenberth indices - which describe both zonal and meridional flow across New Zealand, and the Southern Oscillation Index (SOI) - which describes the atmospheric phase of the El Niñ˜o Southern Oscillation. These results were tested against a baseline multivariate linear regression.

With the help of NeSI's consultancy service, we developed a scalable pipeline to automatically run a variety of experiments including varying the number of lagged circulation indices and training a large selection of models. For all our linear models (e.g., OLS, Gradient Boosting) the minimum root mean square error (rms) was achieved using approximately 96 lagged months of the circulation indices, which in turn explained approximately 10 -15 % of the variance in the rainfall anomalies. However, through using a deep neural network, we can explain approximately 50% of the variance in rainfall. The significant improvement in accuracy is a strong indication that deep neutral networks can extract more abstract relationships from the lagged history of circulation indices.

Since our initial results are promising, we have applied the trained model to data from past and future climate model projections and compared the estimates of climate change signal at the study site from machine learning with dynamical regional climate models output. Future work will include downscaling circulation and synoptic flow patterns, that is two-dimensional synoptic fields to both daily and monthly gridded rainfall anomalies.

High-Throughput Phenotyping for the Orchard

Author: Louis Ranjard

ABSTRACT / INTRODUCTION

Abstract: Modern machine vision systems allow for high-resolution imagery to be collected in the field, creating an opportunity to characterise multiple plant traits (phenotype) on large scale at a significantly smaller cost than manual visual assessment. In horticulture, high-throughput phenotyping of the fruit and canopy is performed by collecting images at very high temporal and spatial resolution. Such data informs of the spatial and temporal dynamic of the fruits, e.g. growth and maturity stages. PlantTech Research Institute is currently developing multiple computational solutions to accurately phenotype orchards and extract useful information from images in combination with other data, such as remote, proximal sensing, weather and microbial community profiles. However, challenges remain to efficiently gather, transmit, process and analyse the resulting datasets. For example, current machine vision systems generate hundreds of megabytes of data per second, which is saved to SSDs, which are then transported from the orchard to the processing facility by courier, making the whole process highly impractical. Moreover, the different sources generate data at different time resolutions, and high level of noise is often found. We will present some of the solutions currently being developed to mitigate these issues.

Author's bio: Louis completed a PhD (Biology) from the University of Auckland in 2010. After his PhD, Louis worked as a postdoc at the Department of Statistics at the University of Auckland and also as bioinformatician for New Zealand Genomics Ltd. After a postdoc experience at the Australian National University, Louis returned to New Zealand to work as a machine learning engineer at Biomatters Ltd. As principal scientist for PlantTech Research Institute Ltd, Louis's research focus on the development and application of machine learning methods to the analysis of biological datasets, with a particular emphasis on the integration of multiscale and heterogenous data sources, from genomics to sensors data.

3D Virtual Atlas of the Uterus

Jonathan Reshef, Hanna Allerkamp, Jo James, Alys Clark
Auckland Bioengineering Institute
jres129@aucklanduni.ac.nz, h.allerkamp@auckland.ac.nz, j.james@auckland.ac.nz,
alys.clark@auckland.ac.nz

ABSTRACT / INTRODUCTION

Introduction: Creating a 3D atlas of an organ's structure is a highly desirable resource to have as this helps to improve in vivo anatomical understanding. One of the most poorly understood organ systems in the human body is the utero-placental circulation. A major limitation to understanding the utero-placental circulation is the inability to investigate the blood vessels in the pregnant uterus, which change rapidly and are inaccessible to direct measurement during pregnancy. Because of these limitations, the extent of their remodelling/adaptation to pregnancy has largely been described only qualitatively. A historic collection of gravid uterine samples our research group has accessed presents a unique opportunity to directly explore these structures for the first time.

Methods: The Boyd and Dixon Collections (Cambridge, UK), contain rare historical collections of preserved and sectioned gravid uterine tissue, treated with a range of histology stains to visualise different parts of the tissue. A slide scanner was used to digitally image specimens containing up to 510 serial sections from samples of 6 – 20 weeks of gestation. These samples were prepared in the 1950s, well before standardised sample preparation and digitisation was available, so are prepared inconveniently and often in poor quality to perform digital analysis. To reconstruct these images into a 3D stack and perform segmentation, an open source custom-written programme was designed specifically to address the novel issues associated with this data. This programme performs significant pre-processing to isolate the samples in their frames. Linear and non-linear registration methods have been developed to ensure the continuity of the structures. Interpolation between missing samples is performed to compensate for tissue damage. Machine Learning methods are used to learn both hand-picked and automatically collected features from the 3D model to segment the entire volume.

Results: This programme has created 3D reconstructions of multiple specimens of gravid human uteri at different stages of gestation. The continuity of the 3D model reveals structures and vessel connectivity that were not previously evident in the 2D sections. Segmentation of specific features of interest is ongoing.

Conclusion: The 3D reconstruction of gravid uterine tissue significantly improves the ability to observe anatomical features. Segmenting individual vessels and features will further enable improve parameterisation of computational models of this circulation, to determine the impact of the dynamic changes in structure on blood flow haemodynamic over the first half of pregnancy.

ABOUT THE AUTHOR(S)

- Jonathan Reshef

  Having studied Biomedical Engineering as his undergraduate at the UoA, Jonathan has continued his studies as a masters student with a particular interest in image processing. This work forms the body of his thesis.

- Hanna Allerkamp

  A post-doc research fellow with the Auckland Bioengineering Institute, Hanna continues to investigate the utero-placental circulation using animal models and understanding how these models relate to human development.

- Jo James

  Jo is a Senior Research Fellow in the Department of Obstetrics and Gynaecology at the Faculty of Medical and Health Sciences. She co-leads of the Placenta Modelling group which develops in vitro and in vivo silico tools to understand how a healthy placenta forms in early pregnancy, and also leads a research group focussed on the role of placental stem cells in fetal growth restriction.

- Alys Clark

  Senior Research Fellow in the Auckland Bioengineering Institute, Alys is one of the co-leads of the Placenta Modelling group which develops in vitro and in vivo silico tools to understand how a healthy placenta forms in early pregnancy. She also develops computational models of the lungs and ovaries with the goal of developing efficient and reliable methods for modelling physical processes that occur simultaneously in complex networks of tissue and blood vessels.

Machine Learning on NeSI 101

Maxime Rio[1,3], Alexander Pletzer[1,3], Nooriyah Lohani[1,2], Megan Guidry[1,2]
[1]NeSI, [2]University of Auckland, [3]NIWA
maxime.rio@nesi.org.nz, alexander.pletzer@nesi.org.nz, nooriyah.lohani@nesi.org.nz,
megan.guidry@nesi.org.nz

ABSTRACT / INTRODUCTION

In this three hour, hands-on workshop, the NeSI team will introduce you to the wonderful world of machine learning via the user-friendly Jupyter on NeSI platform. Come along to acquaint yourself with amazing algorithms and conquer your fear of obscure machine learning jargon.

You will be taken on a whirlwind tour of the Scikit-learn Python library during which you will encounter the usual machine learning suspects (aka things that are easy to google). You will then test-drive your new skills on a real machine learning problem.

Learners will practice model fitting, evaluation and model selection. At the end of this session they should know:

- the definitions of key machine learning terms,
- the main phases of a machine learning project,
- the types of ML tasks and associated classes of models.

For the best experience, we require attendees to come ready to play by pre-registering for the workshop.

Who is this event for

Researchers or individuals that know a bit of coding but not much about machine learning. If you want to brush up on your coding skills beforehand please consider reviewing this programming with python material.

Registration requirements

You need to pre-register for this event to secure a seat. Please email training@nesi.org.nz if you have questions about this workshop

ABOUT THE AUTHOR(S)

Maxime Rio is a data science engineer at NeSI and a data scientist at NIWA. Over the last few years, he has helped scientists by developing probabilistic models, adapting machine learning tools for their needs, scaling imaging processing pipelines for large datasets and providing training. He easily gets excited by scientists' research and wants to help them to get the most out of their data.

Alex Pletzer is a research software engineer for NeSI at NIWA. Originally a physicist, Alex drifted towards high performance during a career that spans research in plasma physics, working for a private company in Colorado and supporting users at university in Pennsylvania.

Nooriyah is a Bioinformatician by training and after working for a few years in a commercial and academic realm, is now a research communities advisor at NeSI passionate about understanding research needs in the eScience sector. She is also Co-chair of the RSE Australia New Zealand steering committee.

Megan Guidry is the Regional Coordinator for the Carpentries in New Zealand and also works as the training coordinator for New Zealand eScience Infrastructure (NeSI). Her main priority is raising eResearch capability in New Zealand through training delivery and community building.

Data science consultancies at NeSI:

a whirlwind tour of case studies

Maxime Rio and Alexander Pletzer
NeSI / NIWA
maxime.rio@nesi.org.nz,alexander.pletzer@nesi.org.nz

ABSTRACT / INTRODUCTION

The volume of scientific data has exploded in recent years as well as the complexity of numerical tools to exploit them. Making use and sense of a deep learning model, scaling and automating a machine learning workflow on HPC, deploying a remote visualization for large datasets are but some of the difficult but rewarding data science skills to master to succeed in modern science.

To help researchers facing the new challenges of this data area, the NeSI consultancy service is increasing its capacity to support researchers for data science related aspects of their research. This presentation will give you a tour of recent projects addressed by the service and highlight domains in which it can assist scientists. Among other things, you'll discover the joy of porting a Tensorflow model, the struggle of making sense of a car crashes dataset, the excitement of automating a (small) weather forecast model, the enthusiasm of engaging with the community in a friendly challenge.

ABOUT THE AUTHOR(S)

Maxime Rio is a data science engineer at NeSI and a data scientist at NIWA. Over the last few years, he has helped scientists by developing probabilistic models, adapting machine learning tools for their needs, scaling imaging processing pipelines for large datasets and providing training. He easily gets excited by scientists' research and wants to help them to get the most out of their data.

Alex Pletzer is a research software engineer for NeSI at NIWA. Originally a physicist, Alex drifted towards high performance during a career that spans research in plasma physics, working for a private company in Colorado and supporting users at university in Pennsylvania.

Deep learning in a clinical context: the big picture from “big data”

Nathan Russell (PhD Student) Supervised by Assoc. Prof Mik Black, Assoc. Prof Peter Larsen & Dr Miles Benton
University of Otago and Institute of Environmental Science and Research

Nathan.russell@esr.cri.nz

ABSTRACT / INTRODUCTION

Clinical data is a source of large volumes of data: it is not uncommon to have orders of magnitude more variables than observations. The amount of collected information is continuing to grow as technology constantly pushes the boundaries of information that can be collected. With such overwhelming volumes of data it is not a surprise that all of this data isn’t fully utilised throughout the clinical decision making process. For example, a critical patient admitted into the intensive care unit undergoes a barrage of tests during their stay generating a large amount of data points, but only single pieces of these are actually acted upon[i]. An alternative, and potentially much more powerful, approach would be to develop a tool that allowed clinicians to visualise this big data in an integrated platform and see the “big picture”, helping to guide clinical decision making, diagnosis and even indicate prognosis or future medical events.

Such tools may come in the form of artificial intelligence/machine learning[ii], something of a trending topic. The development of (un)supervised learning tools would offer the possibility of improving diagnosis, prognosis and best course of treatment, and would enable physicians to tailor their response in real-time. For years now machine learning has been successfully implemented in complex tasks by stock traders and programmers for data analysis and statistical models. However, successful applications for this technology in health and clinical medicine are still fairly limited[iii]

Work conducted during the course of my PhD will explore how machine learning tools can be applied to improve the clinical decision making process and potentially offer deeper insights to aid diagnosis, treatment and prognosis.

This work is currently ongoing and is a key part of my thesis looking at implementation of machine learning tools for “big data” integration, analysis and visualisation to improve clinical decision making.

ABOUT THE AUTHOR(S)

- Nathan Russell
- A 1st year PhD Student at the University of Otago investigating how the clinical decision making process can be improved through implementation of machine learning tools for “big data” integration, analysis and visualisation. With a background in biology and clinical immunology my interests lie in what information can potentially be extracted from medical data as efficiently as possible which can be useful both at the bedside but also in biomedical research.

GPUs on NeSI

Chris Scott, Wolfgang Hayek, Alex Pletzer, Maxime Rio, Georgina Rae
NeSI
chris.scott@nesi.org.nz

ABSTRACT / INTRODUCTION

GPUs (Graphics Processing Units) have been used for many years to successfully accelerate scientific applications. The amount of work required to take advantage of GPU acceleration can vary significantly depending on the code you are using, the problem you are trying to solve, etc. In some cases it may be as simple as recompiling your code to enable GPU support; other cases may require modifying the code, for example using an API such as OpenACC to offload loops to the GPU, or even porting code to CUDA to take full advantage of the parallelisation offered by the GPU.

NeSI's consultancy service (https://www.nesi.org.nz/services/consultancy) can help researchers take advantage of the GPUs on our HPC systems. Here we will present an overview of NeSI's GPU capability, including our recent investment into the latest NVIDIA A100 cards, and then highlight some case studies where we have helped researchers take advantage of GPU acceleration, to give the audience an idea of the amount of work involved and potential speedups. Some examples of recent work we have done in this area include: enabling GPU support in the molecular dynamics application NAMD to accelerate protein modelling simulations; using OpenACC to accelerate a code for computing the log-determinant of a matrix, a fundamental kernel of some data science applications; linking against cuBLAS to accelerate a tropical circulation model; and writing CUDA code to accelerate N-body simulations of the solar system.

ABOUT THE AUTHOR(S)

- Chris Scott is a Research Software Engineer for NeSI
- Wolfgang Hayek is a Research Software Engineer for NeSI and NIWA
- Alex Pletzer is a Research Software for NeSI
- Maxime Rio is a Data Science Engineer for NeSI and NIWA
- Georgina Rae is NeSI's Science Engagement Manager

Paving the way for Bioinformatics excellence in New Zealand

Dinindu Senanayake, New Zealand eScience Infrastructure, dinindu.senanayake@nesi.org.nz

Ngonidzashe Faya, Genomics Aotearoa, ngoni.faya@otago.ac.nz

ABSTRACT / INTRODUCTION

Genomics Aotearoa and NeSI have had their sights set on raising Bioinformatics capability in New Zealand since delivering their first joint workshop in June 2019. 1.5 years later, 491 people have attended workshops run by NeSI and Genomics Aotearoa (with help from other institutions). As the duo approaches the 500 learner mark, they consider what lessons have been learned along the way and what the future of genomics training in New Zealand looks like.

In this talk, Dinindu Senanayake and Ngoni Faya will talk about the successes and challenges of training in 2020 and beyond including:

- Maintaining momentum during a sudden shift to online delivery
- The pros and cons of running week-long training workshops
- How training has improved over months of trial-and-error
- What training initiatives NeSI and Genomics Aotearoa will be focused on in 2021

ABOUT THE AUTHOR(S)

Dinindu Senanayake

- Dini is an Applications Support Specialist (High Performance Computing) at NeSI with a particular interest in Bioinformatics and Computational Biology. He joined NeSI following more than a half a decade of research experience gained in the field of Cancer Genetics, Chemical Genetics and Bioinformatics/Computational Biology.

Ngoni Faya

- Ngoni is Genomics Aotearoa's Training Coordinator, tasked with supporting and building capacity and capability in bioinformatics for New Zealand. He is mainly involved in the development and delivering of Genomics workshops throughout the country in partnership with NeSI and Otago Carpentries. His PhD in Genetics was obtained in 2019 at Massey University where he was looking at the reproduction and venom system of a parasitoid wasp *Nasonia vitripennis*. In the Masters in

Bioinformatics thesis, Ngoni did a comparative analysis of the humans heat shock protein 90s against the human parasite’s proteins. His interests are in gene regulation and function studies using both bioinformatics and molecular biology techniques. Currently, he is also involved in collaboration work where he is analysing transcriptomic and metagenomics datasets.

Telling Stories with Data

Dr Sydney J Shep
Wai-te-ata Press, Te Herenga Waka Victoria University of Wellington
sydney.shep@vuw.ac.nz

ABSTRACT / INTRODUCTION

The technological infrastructure and intellectual capital required to process big data often obscures the need for narrative. Yet, as data journalists emphasise, "Data does not just provide neutral and straightforward representations of the world, but is rather entangled with politics and culture, money and power." Critical narratives help unpick the assumptions we make and the techno-socio-cultural domains with which we interact. Join me on a journey of rich data collaboration and its discontents through algorithmic storytelling and scrollytelling.

ABOUT THE AUTHOR(S)

- Dr Sydney J Shep

  - Sydney is a Reader in Book History and The Printer, Wai-te-ata Press. She focuses on the interdisciplinary study of transnational and cross-cultural book history and print culture in the contexts of the history of empire, history of technology, and the history of reading. Technological convergence is an additional platform for research and practice, bringing both historic and contemporary media into creative conversation though explorations into the digital handmade, generative computer art, and typographically-situated augmented reality experiences. Her current research focuses on big cultural data and collaborative kaupapa Māori approaches and is grounded in the theories, methods, and practices of digital humanities, spatial history, and cultural informatics. In 2014, she was awarded a Marsden Fund grant (her third) to study William Colenso and the Victorian Republic of Letters, with a focus on personal geographies and global networks. Sydney is also a practising letterpress printer, exhibiting book artist, and designer bookbinder who undertakes creative research commissions at Wai-te-ata Press.

If you liked it then you shoulda put on a PID on it BoF Session

Natasha Simons, Anton Angelo, Shiobhan Smith, Laura Armstrong, Yvette Wharton, Siobhann McCafferty

Australian Research Data Commons, Brisbane, Australia, natasha.simons@ardc.edu.au

University of Canterbury, anton.angelo@canterbury.ac.nz

University of Otago, shiobhan.smith@otago.ac.nz

Centre for eResearch, University of Auckland, l.armstrong@auckland.ac.nz

Centre for eResearch, University of Auckland, y.wharton@auckland.ac.nz

Australian Research Data Commons, Brisbane, Australia, siobhann.mccafferty@ardc.edu.au

ABSTRACT / INTRODUCTION

"If you liked it then you shoulda put a PID on it" is the riff off theme song for self-professed persistent identifiers (PIDs) nerds around the world. Adoption of PIDs is key for research outputs that are FAIR and "not just for Christmas". Persistent Identifiers such as ORCIDs and DOIs are critical to enabling FAIR research and lay the foundation for improved citation and tracking of research impact. The PID landscape nationally and internationally is varied, dynamic and evolving which can make it both exciting and challenging for researchers and research institutions to navigate.

The goals of this raPID fire BoF session are to enable you to dip into this dynamic topic and:

- Broaden your understanding of the value proposition of PIDs
- Hear about a range of PID types and initiatives in use in research
- Create a space where you can raise challenges you may be having in adopting, integrating or using PIDs in your research and/or institutional research systems
- Provide an opportunity for you to hear about new developments in the PID landscape from the perspective of the organisation's who enable and support PID adoption in Australia

In this session, you will hear a number of short raPID fire presentations from "PIDs nerds" involved in a range of PID initiatives including but not limited to ORCIDs for people, DOIs for publications and data, PIDs for research instruments, RAID for projects and more. We'll

have plenty of time for discussion, bad puns, references to apt song titles (should you want to sing along) and we're hoping there will be wine and cheese involved!

AUDIENCE

This BoF will be of interest to those designing, implementing, maintaining and supporting PID services including eResearch professionals, repository managers, developers and librarians. Participants should come along prepared to exchange knowledge, share experiences and contribute to discussions about optimising the 'power of PIDs'.

The session will kick off with brief lightning talks presented by those working at the cutting edge of global developments in PID services and infrastructure. Following these, participants will be encouraged to contribute to an open discussion to share experiences, explore ideas and ask questions.

OUTCOMES

Participants will leave the BoF with a fresh perspective on the opportunities PIDs can offer researchers and research organisations. We envisage that many participants will be prompted to explore in greater depth, ideas raised during the session as they might apply to their organisation. The BoF will also offer participants the opportunity to establish or strengthen connections with the broader PID community in New Zealand, Australia and internationally.

ABOUT THE AUTHOR(S)

<u>Natasha Simons</u>

Natasha Simons has her head in the clouds - literally, technically and figuratively. She loves research data and making good stuff happen. As Associate Director, Data & Services at the Australian Research Data Commons she is incredibly serious and responsible, running programs that support the development of data infrastructure and of course have persistent identifiers (PIDs) at the heart. As a "PIDs nerd" she contributes to the development of many PID initiatives in Australia and globally including ORCID, DOI, IGSN, RAiD and more. She is passionate about enabling FAIR data and a corresponding change in scholarly communication culture.

https://orcid.org/0000-0003-0635-1998

Anton Angelo

Anton Angelo is a data librarian working at the University of Canterbury. He managed Canterbury's effort to be among the first NZ Universities in the NZ DOI consortium, and adopting the NZ Orcid Hub, verifying over 80% of Canterbury's scholars' affiliations. He also manages the UC Research Repository, the Canterbury Institutional Repository, and has been very active in supporting Open Access. He has two cats and three chickens. https://orcid.org/0000-0002-2265-1299

Shiobhan Smith

Shiobhan has over 10 years' experience working in Libraries and Museums. Prior to being appointed as the University of Otago Library's Research Support Unit Manager, Shiobhan was Subject Librarian to a number of Humanities departments including Sociology, Anthropology, Geography, and Theology. As Subject Librarian to the Centre for Sustainability, Shiobhan was involved in the development of the Otago Data Management Planning tool and has an interest in Research Data Management. Shiobhan also has knowledge and skills in Digital Humanities, Bibliometrics, and Information Literacy. https://orcid.org/0000-0003-1738-9836

Laura Armstrong

Laura Armstrong is a Senior eResearch Engagement Specialist at the Centre for eResearch, University of Auckland working to engage researchers in eresearch, and deliver research data management services and researcher enablement projects. http://orcid.org/0000-0003-2370-3924

Yvette Wharton

Yvette Wharton is the eResearch Solutions Lead at the Centre for eResearch, University of Auckland, working on research data management services and researcher enablement projects. She has extensive experience in University teaching, research and IT environments and is passionate about using her broad knowledge to facilitate people to achieve their aspirations. http://orcid.org/0000-0002-6689-8840

Siobhann McCafferty

Siobhann is a Project Manager for the ARDC and a fan of all things PID. Amongst other things she manages the ARDC RAiD service and coordinates the ARDC Instruments for Identifiers in Australasia Community of Practice (i4iOZ). http://orcid.org/0000-0002-2491-0995

An Exploration of Social Media Analytics Solutions for Diverse Application Domains

Prof. Richard O. Sinnott

University of Melbourne

rsinnott@unimelb.edu.au

ABSTRACT / INTRODUCTION

Over 1000 Masters-level students at the University of Melbourne have been taught big data analytics on the NeCTAR Research Cloud since 2013 as part of the Cluster and Cloud Computing course taught by the presenter. This course covers HPC programming including MPI as well as the hands-on experiences in dynamic deployment and scaling of applications on the Cloud. Students are exposed to technologies such as noSQL systems such as CouchDB, Hadoop/HDFS and Spark, as well as how to write scalable Cloud solutions using scripting approaches such as Boto and Ansible, as well as latest technologies such as Docker, Docker SWARM and Kubernetes.

The cornerstone of this course is teaching students how to develop and scale big data solutions. Social media has been used for this course throughout as the basis of a live and challenging big data resource. This talk will illustrate examples of student work that focuses on social media data analytics including Twitter, Instagram, Flickr, Foursquare and Reddit. A range of scenarios and solutions will be presented including use of such data to better under the way in which individuals move around the city; commuting patterns; the adult (sex) industry; identification of the gender of social media users; the dietary habits of individuals; linking users across different platforms, through to historic data mining to identify the social media use of suicide completers and the challenges that arise in the reliable reidentification of accounts. Wherever possible the social media data scenarios are compared and validated with official data sources from the AURIN platform (www.aurin.org.au) also developed, supported and maintained by Prof Sinnott's team.

The talk will also cover a recent grant funded by ARDC to capture all social media data across Australia through the Digital Data Observatory. The architecture of this platform is shown in Figure 1. The images are examples of student works that will form part of the talk.

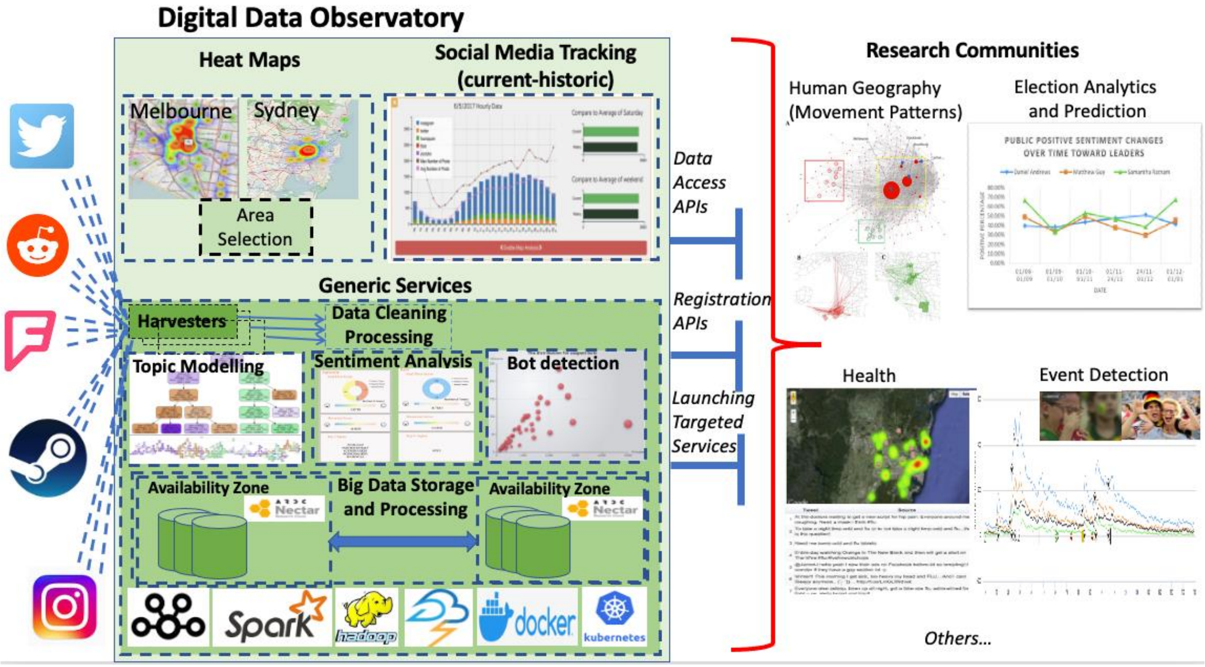


Figure 1: Architecture of the ARDC funded Digital Data Observatory to commence in 2021

ABOUT THE AUTHOR(S)

Professor Richard O. Sinnott is Professor of Applied Computing Systems and Director of the Melbourne eResearch Group at the University of Melbourne. He has been lead software engineer/architect on an extensive portfolio of national and international projects, with specific focus on those research domains requiring finer-grained access control (security) and those dealing with big data challenges. He has over 400 peer reviewed publications across a range of applied computing research areas.

SorTR (Automated Reference DNA Analysis and Rework selection)

Janet Stacey , Anna Lemalu and Maria van der Salm
ESR
janet.stacey@esr.cri.nz

ABSTRACT / INTRODUCTION

ESR interprets DNA profiles from graphs of different coloured peaks (called Electropherograms (EPGs)) that indicate the presence of pieces of DNA of different sizes. The process of DNA profiling causes artefacts so resulting EPGs consist of a baseline signal with a number of peaks which may be artefacts or true DNA fragments. Analysis is required to remove all artefacts based on a set of rules leaving only allelic peaks. However, the interpretation of EPGs can be difficult, consuming resources and time. In addition, there can be a number of reasons why suitable profiling results are not achieved, and therefore an analyst needs to make the decision on the appropriate rework for each failed sample.

SorTR is a prototype workflow using two machine learning models that automatically interpret reference DNA profiles to assign the alleles present and, where required, recommends appropriate rework options. Utilising this prototype system would reduce licence costs and create a minimum time saving of approximately 27 work days in a year.

ABOUT THE AUTHOR(S)

Janet Stacey (presenting author) – Digital Sciences Engineer (ESR)
Anna Lemalu - Senior Scientist – Forensic Biology (ESR)
Maria van der Salm – Scientist – Forensic Biology (ESR)

AARNet’s Sensitive Data Service Journey

Dr Frankie Stevens[1], Robert Pocklington[1], Michael D’Silva[1], Gavin Kennedy[1], **Mike Baker[3], Dr Adele Haythornthwaite[2], Dr Ilka Kolodziej[2]**

[1]AARNet, , Australia

[2]The University of Sydney, Sydney, Australia

[3]Children’s Medical Research Institute, Sydney, Australia

ABSTRACT / INTRODUCTION

As a critical element of Australia’s research infrastructure landscape, Australia’s Academic and Research Network (AARNet) is intricately involved in providing research relevant infrastructure solutions to Australia’s research community. Such services include CloudStor, the national storage, analysis and management platform in use across Australia by over 90K researchers. AARNet has a strategic priority to support and invest in health and medical research infrastructure, and has been approached by the research community to provide sensitive data services, given the perceived gap in services that provide both the security and privacy required for sensitive data research, and the need to collaborate around this. This presentation will describe the journey AARNet has undertaken with respect to developing a sensitive data service, not only for the health and medical community, but also for other communities around Australia who deal with sensitive data, such as ecologists, cultural and human genomics researchers. The sector analysis and subsequent Proof of Concept (POC) project will be described in detail during the presentation, with demonstrations of the platform, including authorisation workflows, multi-factor authentication, audit trails and more. AARNet’s plans for the subsequent pilot and production systems will also be described. This presentation will inform current AARNet service users of the upcoming new features, and give institutions a preview of functionality that they have been requesting to host and manage their institution’s sensitive research data assets.

ABOUT THE AUTHOR(S)

Dr Frankie Stevens is AARNet’s Associate Director, eResearch, and leads the AARNet’s Health and Medical Strategic Priority. Frankie has 20 years’ experience in the Higher Education Sector, having worked in both the Australian and overseas university sectors. She sits on the Technical Advisory Board for the Global Research Data Alliance, and the Australian eResearch Organisations (Aero) Executive Committee.

Dr Adele Haythornthwaite leads a team of research data consultants at Sydney Informatics Hub (University of Sydney), and formulates research data policy and strategy. Having a background in ecology and IT, Adele has a particular interest in helping researchers work with sensitive data.

Ilka Kolodziej is the Clinical Data Systems Manager at the NHMRC Clinical Trials Centre at The University of Sydney. She leads the team that develop clinical data systems for a variety of clinical trials and health-related research projects. These projects often involve international collaboration and require careful consideration towards the collection and management of participant data.

Mike Baker is Head of IT for Children's Medical Research Institute. Mike has a long history of working with transformational national research infrastructure via the eScience program and the University of Edinburgh in the UK and the University of Sydney and AARNet in Australia.

Robert Pocklington is a full-stack software developer at AARNet working on the Sensitive Data Project.

ResBaz: joining the dots for sustainable community effort

Liz Stokes, ARDC, liz.stokes@ardc.edu.au; Laura Armstrong, University of Auckland, l.armstrong@auckland.ac.nz; Matthias Liffers, ARDC, matthias.liffers@ardc.edu.au; Kathryn Napier, Curtin University, kathryn.napier@curtin.edu.au; Matt Plummer, Victoria University of Wellington, matt.plummer@vuw.ac.nz, Yvette Wharton, University of Auckland, y.wharton@auckland.ac.nz

ABSTRACT / INTRODUCTION

Please include any of the following points, but not limited to:

- *Background/Context*

Many ResBaz events were cancelled in the wake of the 2020 COVID-19 pandemic. The collaboratively organised ResBaz digital skills festival model has worked well as a common framework because it is flexible enough for local organising teams to tailor their efforts to their immediate, or city-based communities. But the question remains: are we happy with this distributed model? What kind of regional coordination would help LEVEL UP the impact of these collaborative skills training events?

- *Objective*

The objective of this BoF is to host an information exchange and share experiences between ResBaz event organisers across NZ and Australia, Participants will collaborate on specific challenges in break out room sessions, and prioritise key pieces of inter-organisational coordination which would help ensure sustainability of these skills training communities.

- *Outline*

Th 60 minute BoF agenda is as follows

- Introduction (5)
- Calendar of Awareness – who is planning to do what when? (10)
- Break out room themes (30)
  - Innovating the ResBaz social experience online: what just might work?
  - Meeting learner needs online: what does the long pre-set up tail look like?
  - Supporting instructors: how do you prepare yours?
  - Negotiating with partners: ROI and WIIFM
- What coordination and support would sustain your efforts? (10)
- Wrap up and next steps (5)

- *Outcomes*
  - Calendar of Awareness shares respective timetables for regional ResBazzes
  - Breakout room discussion unpacks problems and suggests creative solutions
  - Collective polling on support and coordination priorities.
- *Acknowledgments*

This BoF continues the momentum for ResBaz coordination from the ARDC eResearch and Data Skills Summit in October 2020.

ABOUT THE AUTHOR(S)

Liz Stokes is a Senior Research Data Skills Specialist at the Australian Research Data Commons. She runs training for the research support professionals in research data management tools and techniques for local and international audiences. Liz has been involved in running the Sydney ResBaz since 2017.
https://orcid.org/0000-0002-2973-5647

Laura Armstrong is a Senior eResearch Engagement Specialist at the Centre for eResearch, University of Auckland working to engage researchers in eresearch, and deliver research data management services and researcher enablement projects.

Matthias Liffers is a Research Software Skills Specialist at the Australian Research Data Commons. He thinks that researchers should be recognised for the software they write. He has been involved in running Perth ResBaz since 2016.

Kathryn Napier is a Senior Data Scientist at the Curtin Institute for Computation, with a research background in ecological sciences and bioinformatics.

Matt Plummer is a Digital Research Consultant based in the Centre for Academic Development at Victoria University of Wellington.

Yvette is the team lead of the Centre for eResearch's Solutions team. With over 20 years' experience in University teaching, research and IT environments she is passionate about using her broad knowledge to facilitate people to achieve their aspirations.

Looking at Forestry from a Digital Lens: Challenges and Lessons Drawn

Alan Tan
Scion
alan.tan@scionresearch.com

ABSTRACT / INTRODUCTION

Forestry is one of the key industries in New Zealand, delivering an export worth of approximately $6.4 billion in 2018 [1]. However, management of forest assets and forestry operations in New Zealand have stuck to the "boots-on-the-ground" approach; relying on sending ground-crews into the field to assess and report on the situation on the ground. Challenges such as scale of forestry plots, size of trees and complexity of work environment in forestry already pose issues for innovation in forestry. With COVID-19 exacerbating existing challenges such as labour and freedom-to-operate, there is now an even stronger motivation to innovate within forestry operations.

In line with the theme for eResearch 2021, we will present current work within our remote sensing and data science research focusing on innovating how New Zealand forestry operates. We will also present some of the challenges and lessons we drew from our experience and journey, thus far, in transforming New Zealand forestry into a high-productive, efficient and technological advance industry sector.

ABOUT THE AUTHOR(S)

Dr. Alan Tan is a senior data scientist in Scion, where he contributes strategically and technically to Scion's data science research initiatives. Dr. Tan's research interest is in the applications of Deep Learning in 3D remote sensing data, data visualisation, distributed systems and high-performance computing. He obtained his Ph.D. in Computer Science from the University of Waikato.

[1] Ministry for Primary Industries. (2020). *Data on forestry imports and exports and indicative log prices*. Available: mpi.govt.nz/forestry/new-zealand-forests-forest-industry/forestry/wood-product-markets/

Virtual Desktops for HPC

Callum Walley
New Zealand eScience Infrastructure
callum.walley@nesi.org.nz

ABSTRACT

Getting started in the world of High Performance Computing can be a daunting task, especially for those unfamiliar with a command line environment.

Many researchers may also want to perform some degree of visualisation without the hassle of downloading large files, or the high latency of using a GUI through X11.

In this session a Virtual Desktop environment solution on NeSI is presented. An overview & demo will be given, and a discussion of what could be coming next will follow.

ABOUT THE AUTHOR

- Callum Walley works for NeSI in the Applications support team. They are also working towards a Master of Engineering at University of Auckland.

Software on NeSI

Callum Walley
New Zealand eScience Infrastructure
callum.walley@nesi.org.nz

Albert Savary
New Zealand eScience Infrastructure
albert.savary@nesi.org.nz

ABSTRACT / INTRODUCTION

NeSI's approach to HPC software will be discussed.

- *Current NeSI software stack.*
- *Maintenance, updating*
- *Building and maintaining your own software.*
- *EasyBuild basics.*
- *Containerisation.*

ABOUT THE AUTHOR(S)

- Callum Walley works for NeSI in the Applications support team. They are also working towards a Master of Engineering at University of Auckland.
- Albert Savary works for NeSI in the Applications support team.

ARDC Data Retention Project: Building the Foundation of Impact in eResearch Infrastructure

Dr J Max Wilkinson
Australian Research Data Commons
max.wilkinson@ardc.edu.au

ABSTRACT / INTRODUCTION

The Australian Research Data Commons (ARDC) believes that to maximise the impact of research data output of meritorious research, researchers must have timely access to high quality data collections stored on stable and persistent infrastructure. A primary framework to deliver this outcome are the FAIR data principles as applied to content, specifically valuable national data collections. Through partnerships with eligible organisations, the Data Retention Project focuses on the F, A and R aspects of FAIR and lays the foundations that embed incentives in both business and data management practices.

Partnerships in the Data Retention project will embed contemporary research data management processes that enrich data collections with controlled and consistent metadata into common infrastructure business models.

This presentation will discuss the challenges of bridging the gap between information management and service provision in the context of the ARDC strategic vision and explain the approaches we have taken to realise the benefits to the Australian research sector.

ABOUT THE AUTHOR

Dr Wilkinson has worked for 20 years in the research data management, research data governance and research infrastructure domains. For the last 5 years he has provided services to numerous eResearch organisations in Australiasia including the National eScience Infrastructure (NeSI), Council of New Zealand Research Librarians (CONZUL), AgResearch, eResearch 2020, MBIE, the Australian Research Data Commons (ARDC) and Microscopy Australia. Prior to this he lived in the UK, most recently as Head Of Research Data and Network Services at University College London, the Datasets Programme Manager at the British Library and Informatics coordinator at the NCRI. He received his PhD in Molecular Nephrology from UCL in 2003.

Creating a living archive of Aotearoa New Zealand's climate model data

Dr Jonny Williams[1], Dr Alexander Pletzer[1,2], Dr Hilary Oliver [1]

1 - NIWA, Wellington, 2 – NeSI,
email - jonny.williams@niwa.co.nz

ABSTRACT / INTRODUCTION

Over the last year, approximately 0.5 petabytes of climate model data has been generated using the New Zealand Earth System Model, or 'NZESM' [Behrens et al.]. Needless to say, this amount of data cannot be analysed easily without postprocessing. It also requires archiving and for this we have the new NeSI 'nearline' archiving system.

Climate model data are the product of a concerted effort between dozens of centres around the world and the data is extremely valuable to determine the impact of climate change on society. Due to the amount of data, postprocessing is numerically expensive, which justifies the need for archiving results. This is equally true for other climate modelling work done at NIWA, for example as part of the next Intergovernmental Panel on Climate Change (IPCC) report; part of the larger 'coupled model intercomparison project', CMIP6.

The NZESM itself is made up of dynamic ocean, sea-ice and atmospheric 'general circulation models' and each of these requires their own methods or postprocessing, each of which will be briefly presented and demonstrated. I will briefly discuss the model itself before illustrating how the main results can be processed and presented in point-and-click format before the data is archived. I will also briefly illustrate how the nearline system is used. Together these represent the 'living archive' referred to in the presentation title; one that can be examined easily using already-generated output.

The individual simulation themselves can take several months to run and so it is essential that we are able to monitor the progress of the model in real time. I will also present results from a successful NeSI consultancy project where the run speed of this monitoring software ('afterburner') was sped up by a factor of almost 40 [Pletzer, 2020] using the Slurm workload manager and Cylc [Oliver et al., Pletzer] on NeSI's Māui platform.

<u>References</u>

- Erik Behrens et al., Local Grid Refinement in New Zealand's Earth System Model: Tasman Sea Ocean Circulation Improvements and Super-Gyre Circulation Implications, Journal of Advances in Modeling Earth Systems, https://doi.org/10.1029/2019MS001996, 2020.

1. H. Oliver et al., "Workflow Automation for Cycling Systems: The Cylc Workflow Engine", Computing in Science & Engineering Vol 21, Issue 4, https://doi.org/10.1109/MCSE.2019.2906593, July/Aug 2019.

2. Pletzer, ‘rosesnip’ software, https://github.com/pletzer/rosesnip, 2020.

Acknowledgements

1. The UK Met Office-led Unified Model Consortium.
2. NeSI scientific and technical staff.
3. The Deep South National Science Challenge for funding.

ABOUT THE AUTHORS

- Jonny has a PhD in computational physics from the University of Bath and has worked in UK academia and government as a climate scientist as well as in private industry as an environmental consultant. He plays classical viola and is a very slow runner.
- Alexander Pletzer is High Performance Computing Research Software Engineer for NeSI at NIWA. Alex plays ping-pong and windsurfs whenever he’s not running jobs on NeSI.
- Hilary Oliver holds a PhD in Astrophysics (Computational Plasma Physics) from Princeton University. He works on software infrastructure for environmental forecasting systems, leads the open source Cylc Workflow Engine project, and co-chairs the Unified Model Consortium’s Technical Advisory Group. In his spare time Hilary rides a motorcycle and wishes he was in a band.

Taking Advantage of Technology Innovations in the Next Generation of NeSI HPC Infrastructure

Jeff Zais

NeSI / NIWA
jeff.zais@niwa.co.nz

ABSTRACT / INTRODUCTION

NeSI, the New Zealand eScience Infrastructure, provides HPC access for users throughout New Zealand. The primary systems are Mahuika and Maui, Cray clusters based on Intel "Broadwell" and "Skylake" processors, respectively. NeSI regularly updates capabilities through the addition of both specialized nodes and expansion of clusters. Usage on both Maui and Mahuika has steadily grown since they were introduced in 2018, and added capacity is required to handle future user demand. This talk will review that usage, plus the options evaluated for an expected 2021 addition to Mahuika. In particular, HPC technology advances in processors, system interconnect, and memory technology will be surveyed. By combining these innovations, an expansion of Mahuika will be summarized, which should serve to enable research for the New Zealand scientific community in the next years.

ABOUT THE AUTHOR(S)

- Jeff Zais

-

- Jeff Zais serves both NeSI and NIWA as the Senior High Performance Computing Architect and Science Advisor. His academic background includes a B.S. degree from the University of Wisconsin, and M.S. and Ph.D. degrees from Stanford University in Aerospace Engineering. Professional experience includes technical and management roles at Ford Aerospace, Cray Research, IBM, and Lenovo, focused on high performance computing application performance and system architecture.

[i] Sittig, D. F., Wright, A., Osheroff, J. A., Middleton, B., Teich, J. M., Ash, J. S., . . . Bates, D. W. (2008). Grand challenges in clinical decision support. *Journal of biomedical informatics, 41*(2), 387-392.

[ii] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869-8879.

iii Ng, A. (2016). What artificial intelligence can and can’t do right now. *Harvard Business Review, 9*.