Accepted Talks and Keynotes

1. Richard Dean, ESR - *The New Zealand Data Science Accelerator*
   I’ll explain what data science accelerators are and why I see them as a really great way to develop a rich and collaborative data science community within NZ. Richard’s pilot accelerator within ESR is due to wrap up in September so it’ll be a really exciting time to hear how things are going right at the coalface.

2. Thomas Etherington, Manaaki Whenua Landcare Research - *A research institution framework for publishing open code to enable reproducible science*
   Reproducible science is greatly aided by open publishing of scientific computer code. Recently, a group of staff at Manaaki Whenua - Landcare Research wanted to establish an institutional GitHub account so that we would have a place to publish the code we were generating. However, while there are many institutional benefits for encouraging the publication of scientific code, there are also institutional considerations around things such as intellectual property and risk. Therefore, taking the perspective of a research institution, we tried to understand what were the benefits and obstacles to open code publishing by asking ourselves: who will be involved, how should code be licensed, where should code be published, how to get credit, what standards, and what costs? In reviewing advice and evidence relevant to these questions we propose a research institution framework for publishing open scientific code to enable reproducible science.

3. KEYNOTE: Celine Cattoen-Gilbert, NIWA - *National scale flood forecasting in the world of data, models, HPC and AI – shaping a more resilient tomorrow.*
Flooding is the most frequent natural disaster in New Zealand. Two trends have recently emerged in the development of new flood forecasting systems: i) a shift from deterministic to ensemble (probabilistic) streamflow predictions, and ii) a move towards national/continental scale systems that attempt to forecast flows for all streams over a given domain. We will present a selection of NIWA's work on forecasting flood and hazards, including key aspects of the development of a national short-term flow forecasting system prototype for New Zealand rivers (0-2 days). The forecast system, running on HPC, is updated every 6 hours and produces river flow forecasts for nearly 60,000 rivers across the hydrologically diverse New Zealand environment. National scale forecasting systems require complex automated workflows, with data assimilating models, and uncertainty quantification techniques. We will highlight some of the key science and technical challenges of this work, and the role of data, models, HPC and AI, in shaping a more resilient tomorrow.

4. Chris Scott, NeSI - Improving NeSI researchers' productivity with a consultancy service

The primary goals of NeSI's consultancy service are to enable increased scale and complexity of research and to help researchers use NeSI's computing capabilities to increase research outputs and impact. Here we will present some insights and examples of the types of work we do in the NeSI consultancy service. Recent trends in the types of work we have been undertaking have shown that researchers are increasingly interested in software sustainability, which covers topics such as build systems, version control and testing. We often get called in to boost the performance of code written in scripting languages, such as Python and R, where writing compiled language extensions that utilise shared memory parallelism can often make a large difference to performance. Increasingly we are also being asked to help with the visualisation of data, including using ParaView to visualise 3d data and implementing in-situ visualisation in simulation codes using ParaView Catalyst.

5. Jonny Williams, NIWA - Extreme earth system modelling on Māui
With the availability of NeSI’s new platforms, specifically Māui in this case, I will present results from our new earth system model. The model is complex to say the least, having a stratosphere-resolving atmosphere, global ocean, dynamic sea ice, atmospheric chemistry and ocean biology. This model is closely related to the one developed in the UK - the UKESM - but with the addition of a nested high resolution ocean model surrounding the New Zealand region. We plan to use this model to investigate the effect of ocean model resolution on marine heatwaves in our region. Getting this model running was a significant technical challenge and I will give an overview of how we got the model running. This is a highly collaborative exercise involving the UK climate modelling community, the Globus high speed data transfer network, a NeSI consultancy project and from the Deep South National Science Challenge.

6. Marie Armstrong, Malaghan Institute of Medical Research - Data management: implementing change at the Malaghan Institute

Facing rapid changes in research technology and the corresponding data explosion, the Malaghan Institute is embarking on a project to better manage our research data. Typically, researchers are encouraged to create data management plans on a project-by-project basis, but we believe that this is a core service to be implemented organisation-wide by the IT department. We believe that for any change to stick, all staff need to be included on the journey – one system to rule them all, rather than heading in different directions.

Our proposed system utilizes concepts borrowed from the film industry and is aligned to the FAIR data principles. The challenge is both technical and people-related, with the “old way” a comfortable and simple system. However, improved ability for centralised data analysis, and realising the value of historic data currently locked away in archives is a tantalising promise for our research staff. As well as implementing the technical components, the IT department is preparing a prescriptive plan to guide staff through that most languishing prospect: change.
7. Alexander Pletzer, NeSI - *Scripting at the speed of compiled code*
Scripting languages such as Python, R and Matlab have a rich set of data structures and offer extensive ecosystems of freely available packages and libraries. These languages are therefore ideally suited for solving hard and complex problems - as such they are an excellent point of entry into the world of high performance computing. A drawback, however, is that programs written in a scripting language may run an order of magnitude slower than compiled Fortran or C/C++ code. Here we'll compare various strategies for making your scripts run on par with compiled code.

8. Jade Arnold, NIWA - *MatLab vs Julia: trials and tribulations*
Post-processing results from the open-source dispersal model Disp_GPU has traditionally been completed using MatLab. The development of a standard post-processing toolbox using free software is very appealing. Julia language, with good performance and similar syntax to MatLab, is a great candidate. This short presentation covers the trials and tribulations of re-coding a short script from MatLab into Julia. Using the example of a script that loops through millions of particles, and finding them in or out of a polygon, over 481 timesteps. What are the benefits? Can you gain insights from our experience?

9. KEYNOTE: Tammy Steeves, University of Canterbury and Roger Moraga, Tea Break Bioinformatics - TBT

10. Robbie Price, Manaaki Whenua Landcare Research - *SMAP Workflows on and off the cluster*
We present an approach to automating workflows for reproducible science than can be run on or off the NeSI cluster with minimal alteration. This mechanism allows us to us the same workflow methodology for jobs that can be run locally as those utilising the NeSI facilities.
We outline the core methodology, and I discuss number of workflow approaches we are implementing to standardise our workflows, code engineering, and metadata production.

11. Wolfgang Hayek, NeSI - Optimising TensorFlow performance on multicore CPUs

TensorFlow is a popular Python library for developing machine learning models, with a wide range of applications. Machine learning, and in particular deep learning, can be computationally very demanding. TensorFlow is therefore typically used with GPUs or specialised hardware. However, almost every modern computer comes with multiple CPU cores with considerable computational power. Running TensorFlow on multicore CPUs can be an attractive option, e.g., where a workflow is dominated by IO and faster computational hardware has less impact on runtime, or simply where no GPUs are available.

This talk will discuss which TensorFlow package to choose, and how to optimise performance on multicore CPUs. We will also compare runtimes of training and inference tasks of a deep learning model between different CPU and GPU configurations as an example of a real-world application.


Free and open source software (FOSS) is increasingly popular in both research and production environments. Traditional closed-source software is often costly or even unavailable on the market, and in a research context the ability to inspect the source code is essential for reliable and reproducible eResearch. But, proverbially, it takes a village. FOSS projects with niche user bases and small core development teams often become moribund due to loss of key personnel and funding support, which is a business risk to consumers of the FOSS project.
Here, we present the NeSI Consultancy Service as a solution to such a software continuity problem, affecting the ecological modelling software Zonation used by the Department of Conservation. We will share journey through diagnosis to solution, and will present possible strategies to make eResearch pipelines that currently incorporate FOSS software more resilient.


The availability of medical domains publicly has made medical information searching a popular type of search on the World Wide Web. However, medical information searchers face many issues and challenges when performing a medical search. The aim of this study is to understand how typical medical searchers (non-medical professionals, medical professionals and medical students) perform medical information searching. Specifically, this research study focuses on how i) task difficulty and ii) topic familiarity affect medical information searching behaviour from a quantitative and qualitative aspect. Results of this research study indicates that classic methods of categorising a searcher based on domain expertise and knowledge are not suitable on an expert type domain instead narrower and more streamlined methods are necessary.

14. Rand Husso, GNS Science - AI for forecasting landslides that result from earthquakes

This presentation will be on the progress we've made trying to forecast landslides caused by earthquakes and precipitation. Good quality input datasets are used to train a Tensorflow / Keras Sequential neural network whose parameters are determined by using Bayesian hyperparameter optimization (BayesianOptimization), and whose training details are established by using KFold (sklearn) iterations with the EarlyStopping monitor. The resulting trained neural network is then applied to all the available data, and some graphs are created to show the results.
15. Alan Tan, Scion - *Building a Visual Analytics System for Spatio-temporal Analysis*

Ability to extract knowledge from complex spatio-temporal datasets is key for understanding events happening in our surrounding environment, such as forest growth, disease spread, climate change and socio-economics. However, there is a lack of efficient tools for analysing these complex datasets from a holistic perspective due to challenges such as data size, quality and multi-dimensional complexities. In this talk, we first briefly introduce our work on a visual recommender system for exploring large complex spatio-temporal datasets. We then touch on the tools and libraries, such as D3 and Vega, used for constructing an interactive and responsive front-end interface for visual analytics.

16. Maxime Bombrun, Scion - *Gradient Boosting comparison: The importance of categorical features*

Characterising trees is challenging due to the size, the growing cycle and the structure complexity, therefore phenotyping at forest-scale in such variable environments requires outstanding technologies. Advances in remote sensing applied in a well-characterised forest have resulted in a deluge of data which provide new insight of the complex interactions in biological systems.

Our principal aim was to implement a model capable of handling a forest-scale (2.7 million observations) dataset with complex and noisy features with mixed data types (categorical/continuous) to predict forest productivity. To successfully achieve this objective, we investigated three recent implementations of gradient boosting machines, XGBoost, LightGBM, and CatBoost. We have shown the importance of including categorical predictors in the predictive model, the impact on the important drivers of...
productivity, and the methodology to select the best genetic material to optimise the deployment of genetic material across the forestry landscape.

The overall aim of the phenotyping platform is to increase forest productivity and sustainability by matching trees with improved genetics to specific sites for deployment. This precision-based approach is the foundation of future forest planting that will ensure the right tree goes in the right place, with the right purpose and management to realise the best outcomes.

17. Yuda Munarko, Auckland Bioengineering Institute - Automatic Annotation of Query in Physiome Model Repository (PMR)
The Physiome Model Repository (PMR) is a collection of physiological and anatomical models written in XML based form. The PMR goal is to provide a robust platform for scientists in the biology-related domain such as bioengineering and biomedical so they can reuse, reproduce, collaborate, and exchange simulation experiments consistently and unambiguously. The stored models consist of elements of mathematical equations along with all variables, and description. Currently, a large number of elements has been annotated using ontology URIs and has been stored as RDF triples for easy management and retrieval. By using SPARQL, scientists are able to find relevant elements and use these for their works. However, the use of SPARQL in PMR needs sufficient knowledge about ontology URIs and elements needed which may cause difficulties. Therefore, we have developed an automatic annotation to map the user’s text query to ontology URIs. We have utilised textual information inside the PMR and ontology URIs' label, definition, and synonym from BioPortal to extract text-based features. We have also used the NLP parser to divide the query into candidate phrases. Utilising these features, we are able to annotate the candidate phrases and then select the final phrases with relatively high accuracy.
18. Ben Jolly, Manaaki Whenua Landcare Research - Remote sensing on NeSI

The remote sensing team at Manaaki Whenua has an interesting history of working on the NeSI HPCs and during the last 8 years has developed a number of systems and practices that make day-to-day life on a shared HPC environment easier and more productive. Remote sensing analysis is typically data intensive and has traditionally been more embarrassingly parallel than true HPC, with users often coming from physics and GIS backgrounds instead of pure computer science. These aspects have introduced a number of challenges for us to overcome in the wider NeSI system, with some important lessons learned along the way. This presentation covers a brief history of our experiences on the old and new systems before providing a broad overview of our current setup for Manaaki Whenua’s remote sensing users on Mahuika, including how we structure our satellite dataset. Among other topics it covers our use of a virtual lab, our own repository of modules, how we use Miniconda and more recently Singularity alongside the lmod environment, and a number of other custom and open source tools.

19. Cyprien Bosserelle, NIWA - In-situ visualization and VTK capability for hydrodynamics model, a NeSI visualisation consultancy

Hydrodynamics models predicting inundation play an important role in hazard assessments and forecast. The BG model uses a block uniform quadtree layout to define the computational grid and allow some adaptive mesh refinement (AMR) at the block level to increase model resolution where needed. ARM grids can be challenging to render and quickly visualize model results for large domain can be challenging with traditional software making it difficult to quickly assess the quality and validity of large-scale models. In order to allow rapid rendering of model variables, a NeSI consultancy was setup to create in-situ visualization routines that allow the BG model to directly output. The presentation will demonstrate the visualisation techniques now available for BG and discuss the NeSI visualisation consultancy.