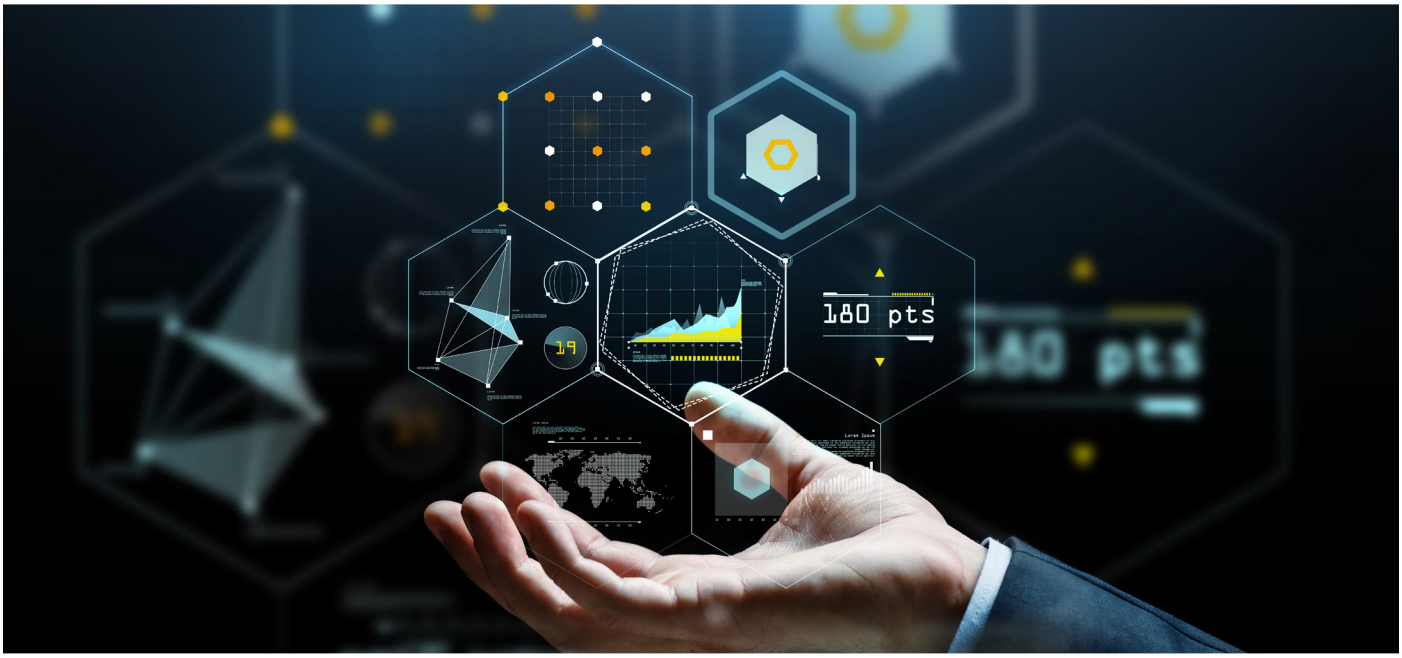
BIG DATA
CONFERENCE & WORKSHOP

25 - 26 May 2017
Plant & Food Research
Lincoln
New Zealand



BIG DATA CONFERENCE & WORKSHOP
DATA SCIENCE: WHEN BIG DATA MATTER

ABSTRACT AND PROGRAMME BOOK

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI

DRAGONFLY Data Science

The Elshire Group Ltd.

LANDCARE RESEARCH
MANAAKI WHENUA

Thank you to our conference sponsors

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI

DRAGONFLY
Data Science

The Elshire Group Ltd.

LANDCARE RESEARCH
MANAAKI WHENUA

BIG DATA
CONFERENCE & WORKSHOP

Welcome

On behalf of Plant & Food Research I would like to welcome you to our inaugural Data Science Workshop. Never before has such a workshop been more timely, with everyone grappling with the management and analysis of ever increasing amounts of data. I hope you enjoy the presentations, events and networking opportunities that the workshop has been designed to deliver.

Richard Newcomb
Plant and Food Research

Conference Venue

Plant & Food Research – Lincoln

Canterbury Agriculture & Science Centre
Gerald Street
Lincoln 7608
New Zealand

Contents

Social Events

Thursday 25 May

Conference Dinner | 6 - 9.30pm

White House Black
Coach transport will be provided from the conference venue.
This will be a 2 course meal with drinks.

Programme

Thursday 25 May 2017 | PRESENTATIONS

Registration Opens	**8.30 am**
The O-Team – Introduction	**8.50 - 9.00 am**
SESSION 1: EXPLORATION	
Shannon Schlueter *The University of North Carolina at Charlotte, USA; Leader of the UNCC Data Science Initiative* Universal Data Science Challenges: Perspective from Disparate Data Domains	**9.00 - 9.30 am**
Mingshu Cao *AgResearch* Transforming Big Data into Big Insights – a Journey of Endless Paradigm Shifting?	**9.30 - 9.50 am**
John McCallum *Plant & Food Research* Better not Bigger: Thoughts on Foundation Data Science in Biology	**9.50 - 10.10 am**
Coffee Break	**10.10 - 10.30 am**
SESSION 2: APPLICATIONS	
Charles Unsworth *Auckland University* Machine Learning for Neural Application	**10.30 - 11.00 am**
Sara Jaeger *Plant & Food Research* Analysing Twitter and Unstructured Text Data	**11.00 - 11.20 am**
Peter McAtee *Plant & Food Research* Lessons Learnt from Implementing KiwiScan, an Analysis Platform for Assessing Fruit Images	**11.20 - 11.40 am**
Ben Pyluc Jolly, David Medyckyj-Scott, Raphael Spiekermann and Anne-Gaelle Ausseil *Landcare Research* A Tool for the Repeatable Generation and Automated Documentation of Land-use Classification Maps	**11.40 am - 12.00 pm**

Programme *continued*

Lunch break	**12.00 - 1.00 pm**
SESSION 3: APPLICATIONS	
Richard Newcomb *Chief Scientist, Plant & Food Research* Data to do with Smell	**1.00 – 1.30 pm**
Farhana Pinu *Plant & Food Research* First Draft Metabolic Reconstruction of a Wine Yeast	**1.30 - 1.50 pm**
Gadi Bental, Richard Oliver, Robert Valkenburg and Andrew Wilson *Plant & Food Research, Compac sorting equipment, Soul Machines* TaDFu – 3D modelling of Fruit	**1.50 - 2.10 pm**
SESSION 4: BIG DATA IN THE SCIENCE INDUSTRY	
Tatiana Lomasko *Center of Space Science Technology (CSST)* Big Space Data: What you Do with It?	**2.10 - 2.40 pm**
Finlay Thompson *DragonFly* Data Science Practices in the Real World	**2.40 - 3.00 pm**
Bruno Lago *Catalyst* Cloud Computing for Data Scientists	**3.00 - 3.20 pm**
Eric Burgueño *Plant & Food Research* Making Science Better: A Deep Dive into powerPlant	**3.20 - 3.40 pm**
Coffee break	**3.40 - 4.00 pm**
SESSION 5: DEMONSTRATIONS	
Demo Session 1 *Dragonfly* A Live Demonstration of Modelling Incidental Common Dolphin Captures in New Zealand's Commercial Trawl Fisheries	**4.00 - 4.30 pm**
Demo Session 2 *Bruno Lago, Catalyst* Demonstrating the Catalyst Cloud Computing Platform	**4.30 - 5.00 pm**
Drinks & Nibbles	**5.00 - 6.00 pm**
Conference Dinner	**6.00 - 9.30 pm**

Programme *continued*

Friday 26 May 2017 | WORKSHOPS

Linley Jesson, Marcus Davy and Mark Wohlers *Plant & Food Research* Data Visualisation	**9.00 am - 12.00 pm**
Morning tea	**10.15 - 10.30 am**
Shannon Schlueter *The University of North Carolina at Charlotte, USA; Leader of the UNCC Data Science Initiative* Synergy from Heterogeneous Data	**9.00 am - 12.00 pm**
Morning tea	**10.15 - 10.30 am**
Lunch Break	**12.00 - 1.00 pm**
Shannon Schlueter *The University of North Carolina at Charlotte, USA; Leader of the UNCC Data Science Initiative* Machine Learning for the Non-Computer Scientist: Where's the Easy Button?	**1.00 - 4.00 pm**
Afternoon tea	**2.15 - 2.30 pm**
NeSI Managing Large Scale Data in a High-Performance Computing Environment Using the NeSI Computational Infrastructure	**1.00 - 4.00 pm**
Afternoon tea	**2.15 - 2.30 pm**

BIG DATA
CONFERENCE & WORKSHOP

ABSTRACTS

PRESENTATIONS

Universal Data Science Challenges: Perspective from Disparate Data Domains

Shannon Schlueter (University of North Carolina, USA)

Data science is largely a domain agnostic discipline. Best practices and novel solutions to once intractable challenges are routinely gleaned from seemingly disparate data domains. Methodologies once rarely identified outside of the realm of bioinformatics and computational biology are now commonly used to address challenges in fields such as image recognition and social sentiment analysis. But is this synergy mutually beneficial? Can the life sciences and bioinformatics claim benefit from practices found in other data rich domains?

Transforming Big Data into Big Insights – a Journey of Endless Paradigm Shifting?

Mingshu Cao (AgResearch)

Driven by modern instrumentation, big data in Omics has brought many opportunities and the fundamental changes. The changes include the approaches we store, analyse and share data; conclusions we can effectively reach and validate; and even in the way we communicate complex information to stakeholders. Based on my experience working with large data in Metabolomics, Transcriptomics and Genomics I present case studies to demonstrate the power of big data in generating exciting leads and the challenges we need to face. Making better use of big data in Omics demands enhanced computing environment, capability of dealing with heterogeneous information, and novel algorithms to reveal new connections and to model system effects of complex traits.

Better not Bigger: Thoughts on Foundation Data Science in Biology

John McCallum (Plant & Food Research)

Few biologists handle ‘big data’ (> 1Tb) but now frequently encounter ‘large data’ that stretches their computational toolset and skills. There are often financial, social, technological or business barriers to data aspirants wishing to scale their practice. Realising value from data-rich technologies for our lives, business and communities will require that we work together to raise standards of digital scholarship, support diversity, and adopt lifelong learning practices. This presentation will provide an opinionated review of experience gained with large data and highlight the foundation data science skills that are helping us. In particular we will demonstrate examples of how the Python ecosystem is enabling choice, collaboration and scalability.

Notes

Machine Learning for Neural Applications

Charles Unsworth (Auckland University)

In the first part of the seminar, I will discuss how we have applied ANNs for the detection of volatile organic compounds (VOCs) from the in vivo firings of insect olfactory neurons. In the second part of the seminar, I will discuss how we have applied Machine Learning approaches to identify Sharp wave transients in the EEG in order to aid the early detection of Hypoxic Ischemia in Newborn Infants.

Analysis of Unstructured Consumer Insights

Sara Jaeger (Plant & Food Research)

Analysis of social media data for consumer insights had been gaining popularity. In this talk, a study that analyses Tweets about eating situations is presented, alongside a discussion of the challenges encountered in the analysis process. Emoji are commonly used by users of social media, and are said to play the role of non-verbal cues in computer mediated communications. Hence, emoji can alter the meaning of posts/ messages and it is important to understand their meanings. Top-line results are presented from a study that investigates emoji sentiment. The presentation concludes with a brief perspective on the value of social media data in food-related consumer research.

Lessons Learnt from Implementing KiwiScan, an Analysis Platform for Assessing Fruit Images

Peter McAtee (Plant & Food Research)

Selective breeding in fruit trees, arable crops and finfish currently involves raising populations of tens of thousands of individuals obtained from multiple controlled crosses of elite parents. Such large populations mean that during the early stages of selection (e.g. stage 1 in PFR's pip- and kiwifruit breeding programmes) only a subset of individuals are evaluated based on multiple traits. Phenomic data is used for calculating the individuals' breeding potential and for the breeders to make decisions about which individuals to move to further stages or cull. Unfortunately, the number of individuals that can be phenotypically assessed is limited and a large proportion of raised individuals often get deemed non-suitable for evaluation in the field. As a consequence, the phenomic information from these individuals is not captured, weakening the knowledge gained from our programs. Phenotypic assessment of all individuals is a time-consuming process, subject to human error, is currently not reproducible, and uses progressive thresholds.

KiwiScan is a platform that was developed to address some of these issues. It was developed in-house to collect, storage, and computationally assess cross-sectional images of kiwifruit. This platform utilises open source image analysis software (ImageJ) to estimate tissue boundaries and then quantify the amount of colour present within each tissue selection. The use of this quantification platform in conjunction with a previously genotyped kiwifruit block has validated the selection of key red QTL locations and highlighted some potential candidate QTL's for other important breeding traits.

Notes

A Tool for the Repeatable Generation and Automated Documentation of Land-use Classification Maps

Ben Pyluc Jolly, David Medyckyj-Scott, Raphael Spiekermann, and Anne-Gaelle Ausseil (Landcare Research)

Land-use classifications (LUCs) provide spatial information on different land uses over a given area, from local to national scale. They are typically constructed by applying complex rules over multiple spatial data sets, some of which require normalisation or pre-classification. While generally of good quality, documentation of this process is still variable and repeatability is a concern given the complexity involved. Most steps happen inside geographic information system (GIS) tools that do not strictly enforce documentation and data sets can be passed around researchers so tracking provenance can be problematic. Processing time for national scale LUCs is typically in the order of days. We present and discuss a tool created specifically to address these issues while optionally moving the creation of LUCs into the high-performance computing space. Called 'pyluc', it uses a single simplified Python script to define inputs and rules in a self-documenting way. Using this script, pyluc acquires input data from the original sources, applies the defined classification rules, and produces the resulting classification map along with supporting human-readable documentation and machine-readable W3C PROV data provenance information. Processing is raster-based (automatic rasterisation or reprojection of input data is provided) and tileable, with optional parallelisation that significantly reduces processing time. The only requirement to replicate a given LUC map with supporting documentation is the single pyluc definition script and the pyluc tool itself.

Data to do with Smell

Richard Newcomb (Plant & Food Research)

Smell is one of the most complex of our senses. Thousands of odours, hundreds of receptor genes, a few sensory perceptions and all with just one nose. In this talk I outline how we have used analysis and visualisation of data from genomes and odour space to unravel how odours are perceived in insects and humans.

Notes

First Draft Metabolic Reconstruction of a Wine Yeast

Farhana Pinu (Plant & Food Research)

In this project, we are developing a metabolic model that will ultimately assist winemakers in identifying the effects of variable juice compositions and winemaking practices on the chemical composition of wines. In collaboration with the University of Queensland, Australia, we first developed the first CYC (online) database of *Saccharomyces cerevisae* EC1118 based on the publicly available genome annotations. This database was then used to generate the reconstruction of a mathematical model. This platform helped us to generate knowledge of wine yeast metabolism and fermentation behaviour. We curated and refined the draft reconstruction based on existing literature. Deletion, addition and modification of reactions, genes and metabolic pathways were carried out where appropriate to improve the draft model. The next step was to perform in silico juice manipulation experiments using an existing dataset that contains information about over 400 Sauvignon blanc grape juices and wines collected over four different harvesting seasons from different wine growing regions of New Zealand. We also used data obtained from the literature. Based on this information, the model is now able to generate over 1000 hypotheses regarding primary metabolism of *Saccharomyces cerevisae* EC1118. These hypotheses can be used to design and perform winemaking experiments to answer specific scientific questions. The results will be included in the metabolic model to validate and improve the prediction capability.

TaDFu – 3D Modelling of Fruit

Gadi Bental, Richard Oliver (Plant & Food Research), Robert Valkenburg (Compac sorting equipment), Andrew Wilson (Soul Machines)

At packing time fruit are being graded by a sorting machine (or human graders) before they are packed and stored. Compac Sorting Equipment make grading machines based on computer vision. Fruit are graded for colour, shape, size and visual blemishes. The fruit are moving and rotating in front of several cameras at a rate of 10 fruits per second. About 200 images are taken of each fruit and the information from all these images need to be merged into a coherent picture of the fruit. We present here an algorithm to build an accurate 3D model of the fruit using these images. The algorithm was used to build models of over 1000 fruit using a small sorter at Ruakura.

Notes

Big Space Data: What you do with it?

Tatiana Lomasko (Centre of Space Science Technology)

The focus of this seminar is presenting the opportunities rasising from the establshment of the NZ Centre of Space Science Technology (CSST). New Zealand needs to capitalise on the rapidly developing global opportunity provided by satellites, both for Earth observation and for data telemetry. This opportunity is relevant for many sectors of New Zealand's economy and society including primary production, energy and water management, national security, urban planning and hazard management. I will present the CSST's plans to profit from extracting the full value available from existing space-based measurements that are often made freely available by international space agencies as well as from developing a national capability to own and operate our own satellites to enhance our country's wealth and well-being.

There is exponential growth in the volume of satellite data, primarily imagery, being made available. This Earth Observation market was estimated at NZ$2.7 billion in 2015 (by North-ern Sky Research) and is predicted to double by 2024. The market segment comprises data, data processing, and information products, with value-added information products undergoing the most rapid growth and having the greatest future potential.

Data Science Practices in the Real World

Finlay Thompson (DragonFly)

While doing research in a reproducible way is recognised as desirable goal, this ambition can sometimes be neglected when confronted with the heat of tight deadlines and busy research programmes. At Dragonfly we have developed/discovered a set of practices that ensures our work is always reproducible. I will talk about what those principles are, and how committing to reproducible research has increased the reliability of our work, and the productivity of our team.

Notes

Cloud Computing for Data Scientists

Bruno Lago (Catalyst)

Cloud Computing is a disruptive technology, but only when done right. In this presentation Bruno will explain how data scientists and researchers can benefit from cloud computing, describe options to store and process massive amounts of data, and what software testing methods can be used to ensure the integrity of the code, process and results.

Making Science Better: A Deep Dive into powerPlant

Eric Burgueño (Plant & Food Research)

Have you ever heard of powerPlant and wondered exactly what it is? Are you using powerPlant already but want to know more? Then come along! In this presentation I will explain the reasons behind why powerPlant exists, why (and when) it is useful, and how you can use it to make your Science better. Also, something about a cat.

Notes

WORKSHOPS

Data Visualisation

Linley Jesson, Marcus Davy, Mark Wohlers

This course wll comprise the following:

- Principles of data visualisation. How do we compare sizes, shapes, areas, colours and how to use them to create visually informative graphs. Extending this to look at multi-dimensional and time series data using the NIWA climate database.
- Examples of vizualising Twitter data in a 2-d scatter plot using the T-SNE data reduction algorithm, then run through a more practical t-sne example based on a few versions found online using the MNIST dataset.
- How to run javascript examples using D3/D3plus to show networks like this http://d3plus.org/examples/basic/9034389/, and a vega example where a 2-d scatter plot where you can follow points over time https://vega.github.io/vega/examples/global-development/
- How to dynamically visualise data using R, and evolving technologies plotly, rbokeh and shiny applications

Synergy from Heterogeneous Data

Shannon Schlueter

Data integration is a demanding necessity of bioinformatics. With an ever growing count of new data sources and thoughts on the analysis of such data, many technologies are offered as a panacea for the resolution of this ‘big data’ challenge. Which of these choices add value to your data? Which provide insight and which are fraught with bias and misinterpretation? This workshop will explore a selection of data integration techniques through hands-on data science. Participants will analyze disparate data sets through a series of tutorials all the while focused on the question “Does the inclusion of this data provide more/better insight?”

Notes

Machine Learning for the Non-Computer Scientist: Where's the Easy Button?

Shannon Schlueter

Machine learning has become a tool atht s ised more often in the research community to analyse Big Data. This course will introduce you to that technique on a beginner level. It will give you a basic understanding of the ML concept.

Managing Large Scale Data in a High-Performance Computing Environment Using the NeSI Computational Infrastructure

NeSI

This session will cover a range of aspects regarding research data, in particular large scale data. We will cover best practices around storage and management of big datasets (including Data Management Plans) and advantages of storing research data within the compute infrastructure such as the one provided by the New Zealand eScience Infrastructure (NeSI). We will discuss main caveats regarding handling large-scale datasets for research. NeSI will also present its planned data services and solutions.

Notes

Notes

Thank you to our conference sponsors

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI

DRAGONFLY
Data Science

The Elshire Group Ltd.

LANDCARE RESEARCH
MANAAKI WHENUA

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI

DRAGONFLY
Data Science

The Elshire Group Ltd.

LANDCARE RESEARCH
MANAAKI WHENUA

SCIENCEEVENTS.CO.NZ/DATASCIENCE