Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI
Discover. Innovate. Grow.™

agresearch

SCION
forests · products · innovation

Bringing New Zealand's eResearch in parity with that established by North America's Human Genomics Community

Helge Dzierzon, PhD

Team Leader Bioinformatics, *Plant and Food*

Nauman Maqbool, PhD, MBA

Group Leader Knowledge & Analytics, *AgResearch*

Tatiana Lomasko, PhD, MBA

Research Leader, Computer Science and Bioinformatics, *Scion*

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI
Discover. Innovate. Grow.™

agresearch

SCION
forests · products · innovation

What kind of Questions are our Industries facing nowadays?

==

What are our Researchers working on these days?

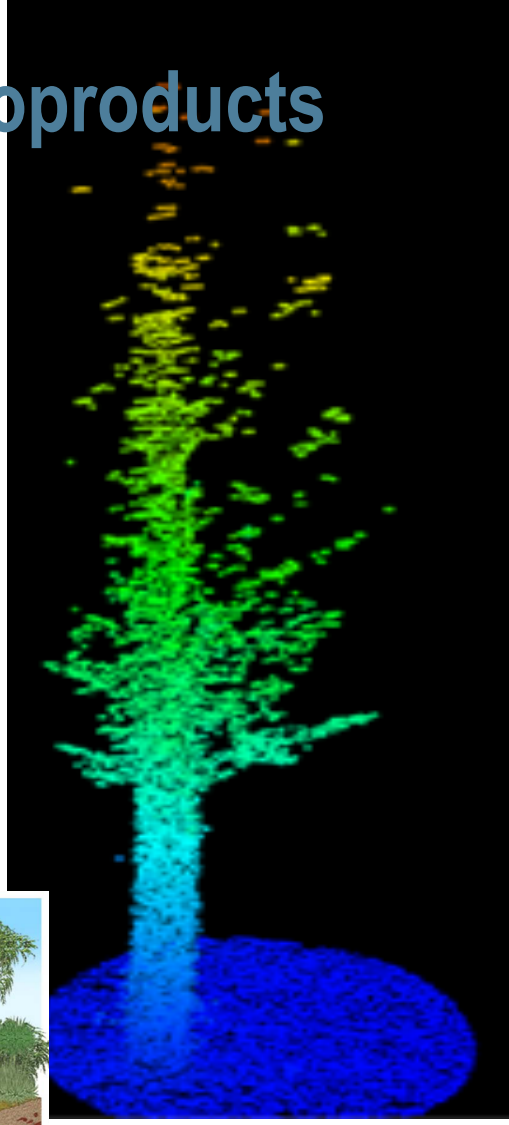
Scion

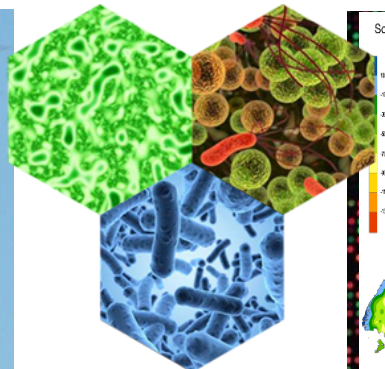
Prosperity from trees

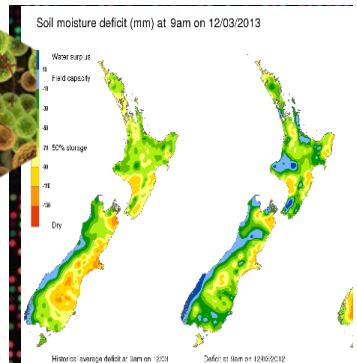
Mai i te ngahere oranga

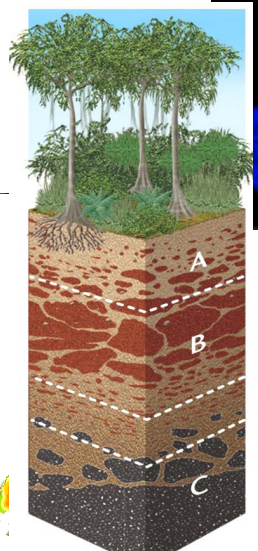
Forest Science + Manufacturing and Bioproducts

- **Big Data - Omics**
  - Genomics, Proteomics, Metabolomics
- **Big Data - Phenotyping**
  - LiDAR, Spectral
- **Big Data - GeoSpatial**
  - Metagenomes, Terrain, Meteorological



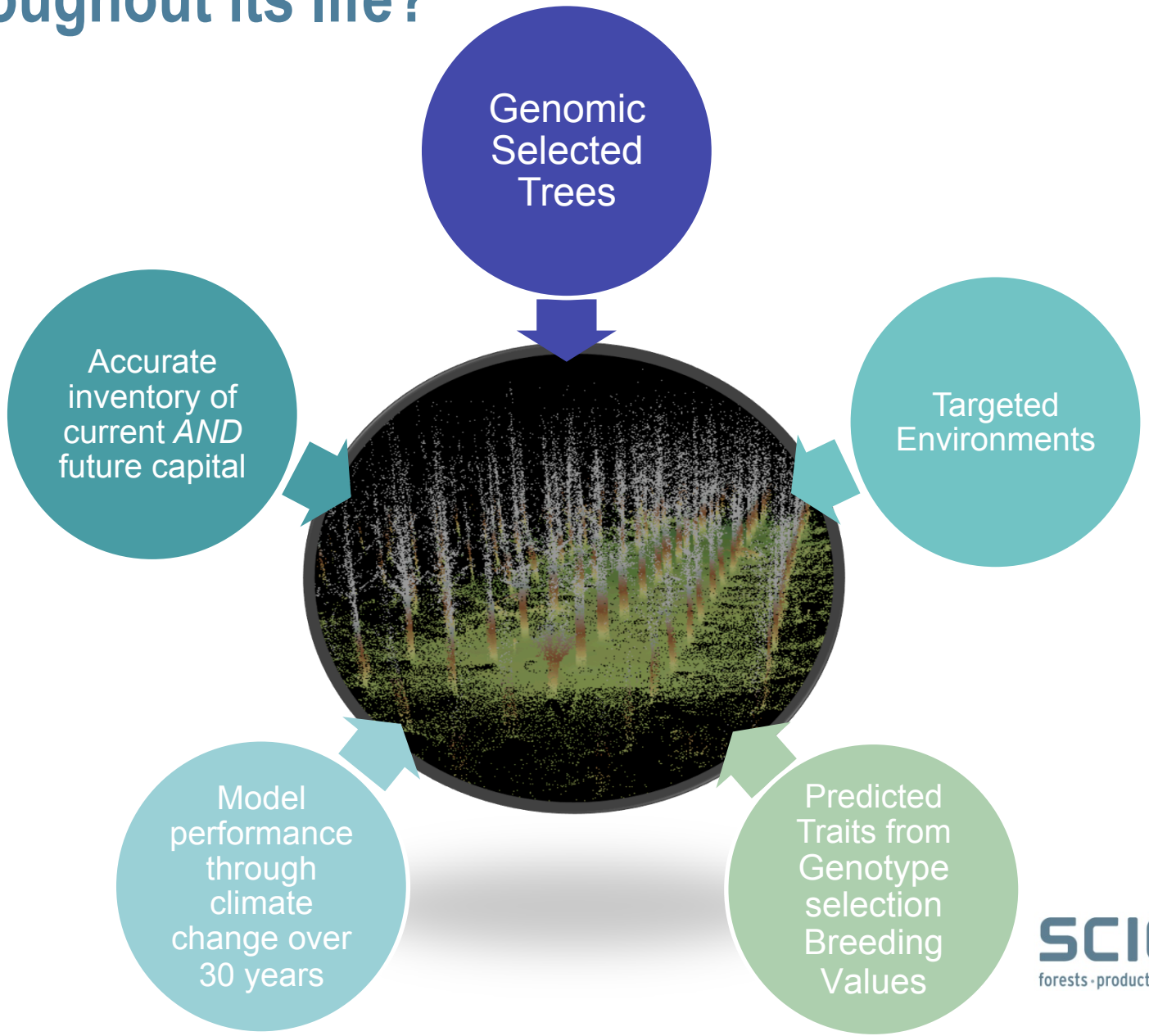






SCION
forests · products · innovation

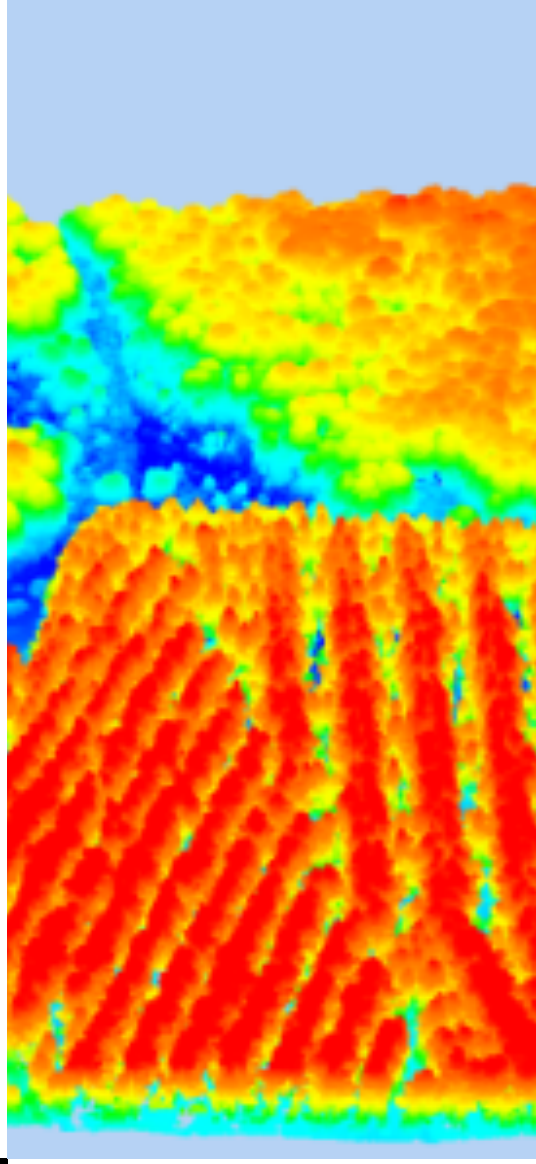
BigData Challenge: Can we model a whole forest throughout its life?

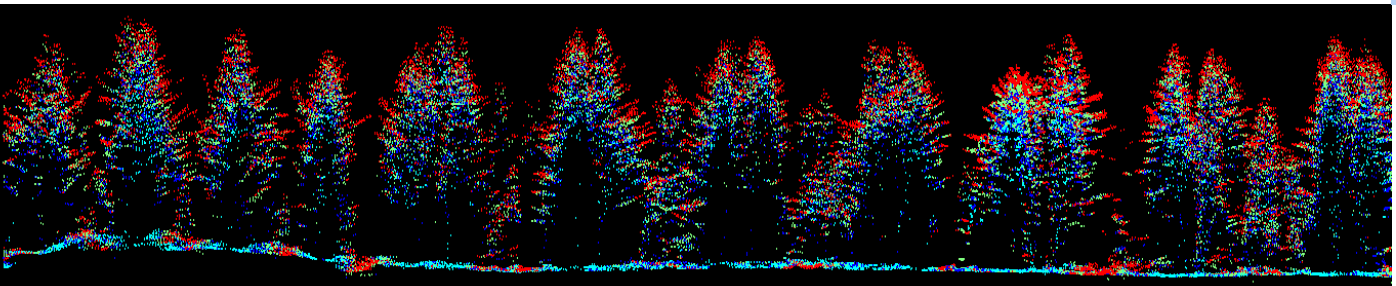


SCION
forests · products · innovation

Current Data Challenges

- Large, Un-linked, Datasets
  - How do we even start combining them (genomics/ LiDAR)?
- Data Analytics
  - How do we make sense of BigData?
- Internal knowledge share
  - Geospatial: pattern recognition, machine learning applied to Genomic predictions
- How much information is enough?
  - Is precision forestry cost effective? ROI?





SCION
forests · products · innovation

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI
Discover. Innovate. Grow.™

Plant and Food

- Plant Breeding
- Human Health
- Food
- Bioprotection
- …

PFR Research Complexity

- Sustainable Production, Food Innovation, Seafood Techn., Bioprotection, Breeding and Genomics
- Genomics, Metagenomics, Proteomics, Metabolomics, Phenomics
- Species: Apple, Pear, Kiwifruit, Snapper, Fungi, Bacteria, …

Challenges

- Managing the 4(5) Vs of Big Data
  - Volume, Velocity, Veracity, Variety, Value (IBM)
- Statistical variability high particular in humans
- We are dealing with highly complex systems with a limited amount of information about them
  - Collaboration of essence
- High competitive environment (1.4 Billion against 4 Million)
- Thus: We are working on automation for a higher productivity

AgResearch

Our Research focuses on:

- Pasture-based animal production systems
- New pasture plant varieties
- Agriculture-derived greenhouse gas mitigation and pastoral climate change adaptation
- Agri-food and bio-based products and agri-technologies

Challenges…

- Data getting bigger
  - Data Storage & Analytics
  - Bioinformatics
    - Is internal HPC the way to go?
  - Precision Agriculture
  - Data transfer speed
- Research Data Management

agresearch

Example Project

1000 Sheep Genomes

- Utah State University, Baylor College of Medicine (US), AgResearch (NZ), CSIRO & the DPI Victoria (Aus)
- Whole genome sequencing for SNP identification
  - gene mutation discovery
  - GWAS and genomic prediction
  - genome evolution, domestication & the consequences of selection etc.
- Analytical challenges
  - 30TB of raw data (of 1000 genomes)
  - Data transfer speed Internationally & Nationally
  - CPU requirements – multiple analyses/month
    - 7000 CPU hours/month on average, overall 50,000
  - Storage

agresearch

Plant & Food RESEARCH RANGAHAU AHUMĀRA KAI Discover. Innovate. Grow.™ agresearch SCION forests · products · innovation

BUT, where are we today?

+

First steps to move forward

- Update our eResearch Strategies to be aligned with international eResearch best practice
  - Establish eResearch mechanisms and practices that would maximize efficiency and usability while minimizing costs
- What kind of Data Analytics do we need?
  - Machine/deep learning, Statistics, R, Hadoop, Spark
  - What do we already have? How to bridge the gap(s)?
- What can our CRI's learn from each other?

How has North America's Human Genomics Community tackled it? *What can we learn from them?*

Ontario Institute for Cancer Research

OICR is an innovative translational research institute dedicated to research on the prevention, early detection, diagnosis and treatment of cancer.

International Projects:

- ICGC: International Cancer Genome Consortium
- Genomic Data Commons
- Cancer Genome Collaboratory
- Pancancer Analysis of Whole Genomes

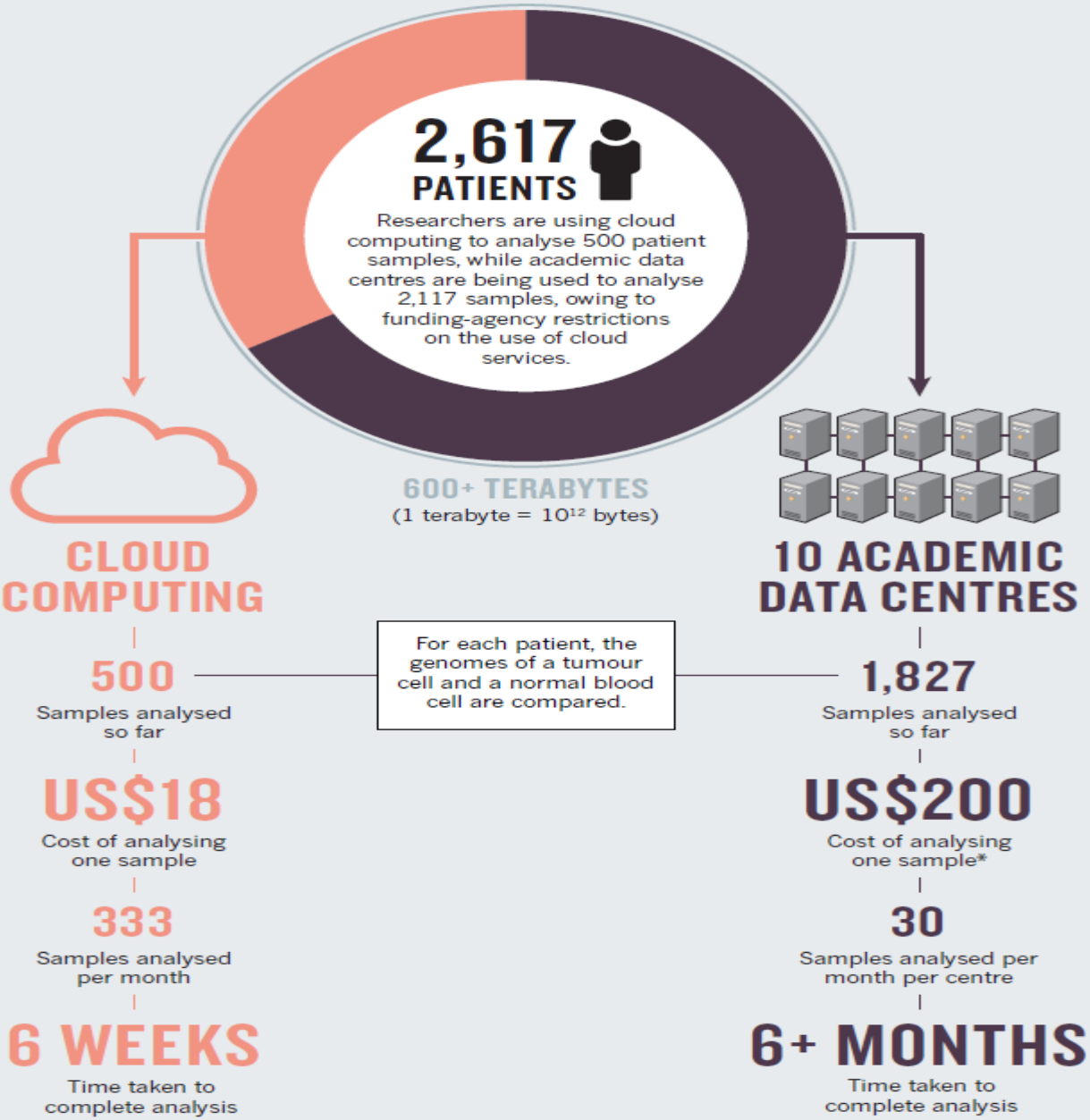
- 10-15PB, ~1000 cores
- ~20-30 working groups, over 1000 researches
- Research, IT, technical and legal

The Pan-Cancer Analysis of Whole Genomes

- Joint effort: coordinated, uniformed “Blueprint” submission
  - ~5000 whole genomes from ~2500 donors to investigate the non-coding parts of the genome in cancer
- Uniformed data processing and analysis pipeline(s)
  - Deploy and monitor workflows at multiple compute sites
  - Protocols for authorizing access to sensitive data in the cloud

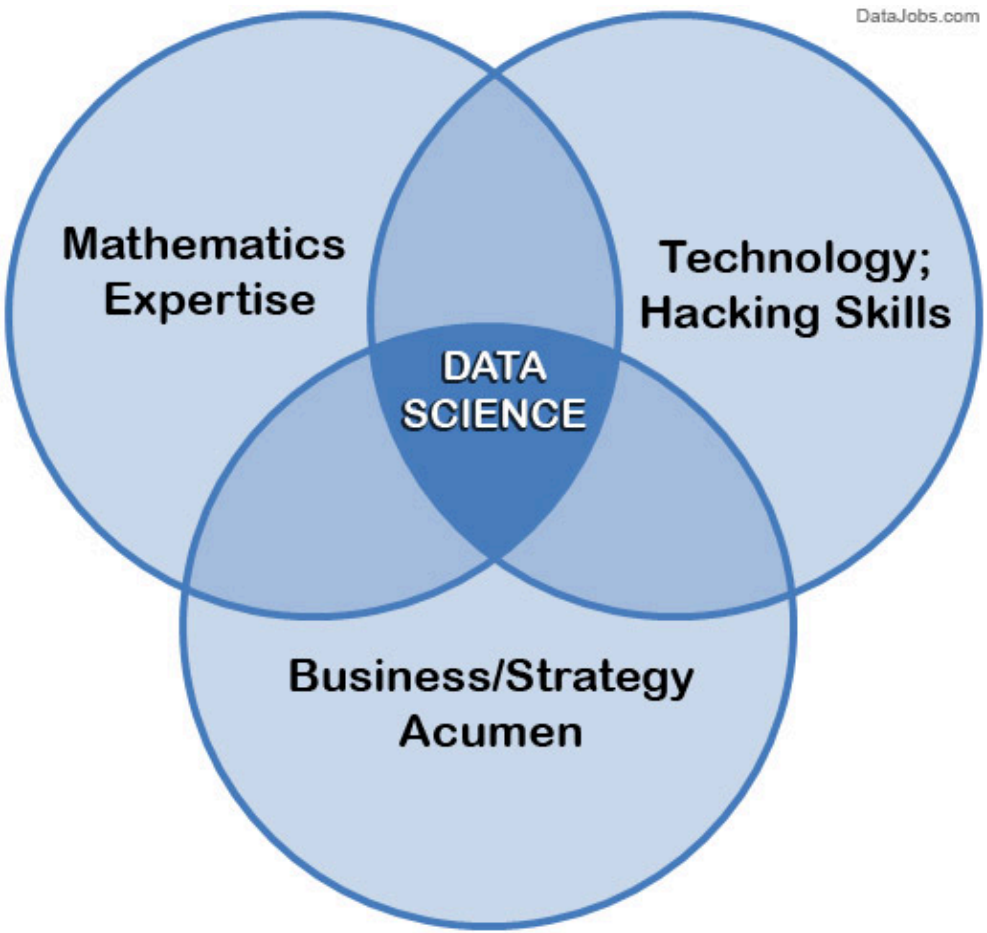
EXPRESS LANE (Lincoln Stein, et.al., July 2015. Create a cloud commons. *Nature*, Volume 523, 149-151)

The Pan Cancer Analysis of Whole Genomes project (in which L.D.S., P.C., G.G. and J.O.K. are involved), an effort to investigate the role of non-coding parts of the genome in cancer, demonstrates how much faster and cheaper it is to use cloud computing than to use conventional academic data centres when analysing vast biological data sets.



Data Science

- R, Statistics
- Apache Hive
- Impala
- Hadoop
- Text Analytics
- Predictive Analytics
- Machine/Deep Learning
  - Spark mllib
  - Amazon ML
  - Azure



Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI
Discover. Innovate. Grow.™

agresearch

SCION
forests · products · innovation

Our Strategic Plan

==

Users Oriented Approach

+

Distributive Learning/Hybrid Model/ Outsourcing

+

Knowledge Sharing/Collaboration

+

Working with REANZZ, NeSI, NZGL on matching demand & supply

Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI
Discover. Innovate. Grow.™

agresearch

SCION
forests · products · innovation

Acknowledgements

Scion:

Warren Parker
Emily Telfer
FII Team
IT Team

Plant and Food:

Plant and Food
BI Team
Biom. Team
Tatiana

AgResearch:

Rudi Brauning
Alan McCulloch
Russell Smithies

Thank You!

Q&A