SCAsia 2019
Supercomputing
Gathering the **Best of HPC** in Asia

Australia's precision medicine
powered by machine learning and cloud-computing

Hon A/Prof Denis Bauer | PhD
11 March 2019, SCA

Building a healthier world together

Precision medicine

… in Australia

Find Disease Genes

… using Apache Spark

VariantSpark
Machine Learning for
Genomic Variants

Share Disease Insight

… using serverless architecture

GT-Scan2
Computationally Guiding
Genome Engineering

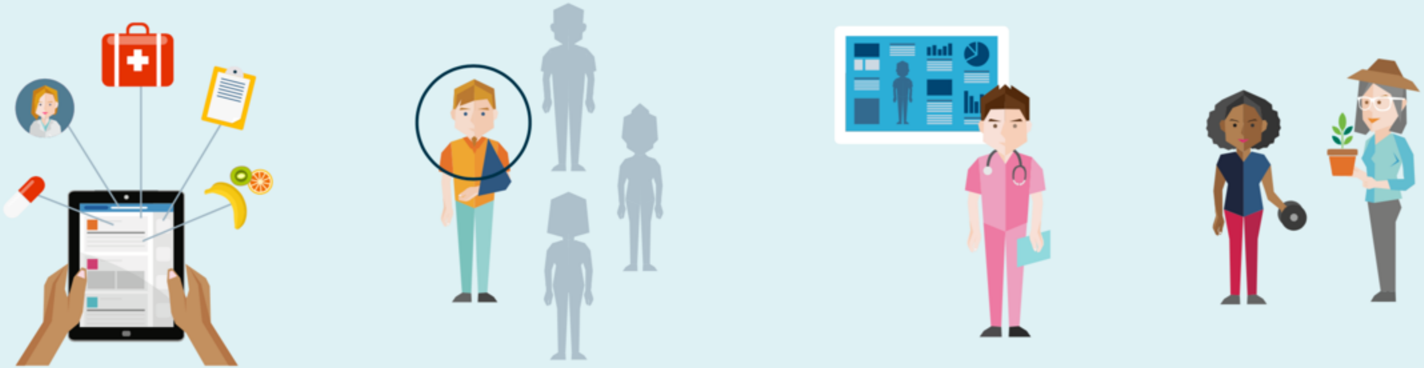
CSIRO: Top 1% of global research agencies

- Invented **WiFi**, used in five billion devices globally.
- Developed the vaccine for the **Hendra Virus**.
- Developed the **Total Wellbeing** & Low-Carb Diets.

Credit https://toolstotal.com/

CSIRO's vision for the #FutureOfHealth

The health system will shift...

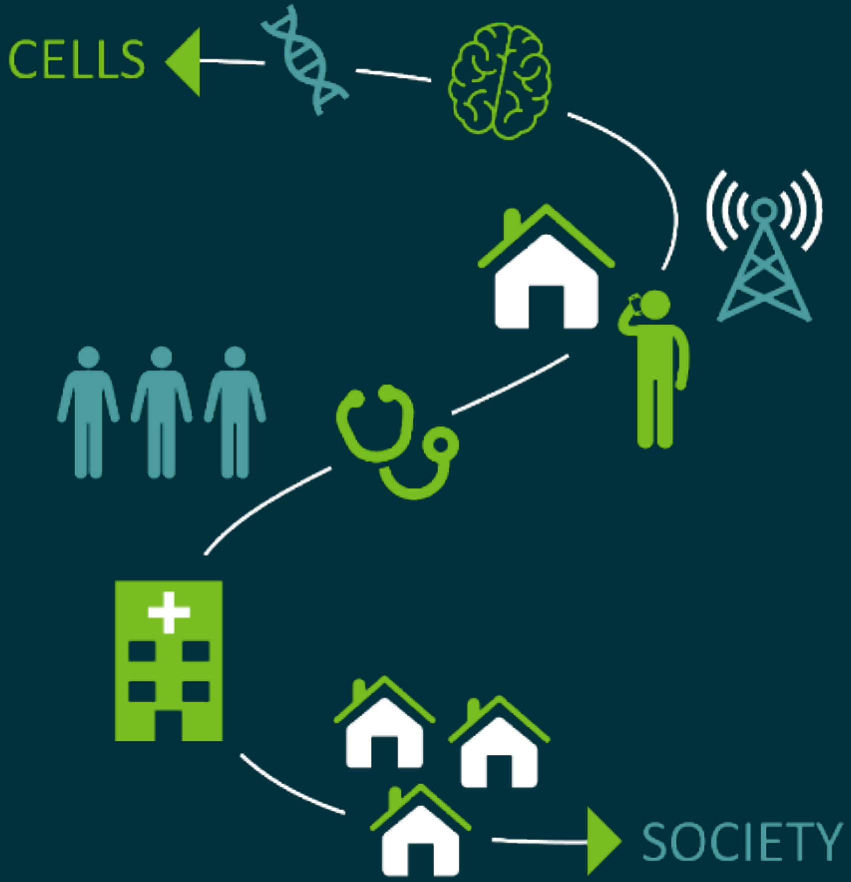


...from treating patient illness to managing consumer health and wellbeing

...from accepting one-size-fits-all to precision health solutions

...from a reactive system to a holistic and predictive approach

...from extending life to improving quality of life over a lifetime



THE AUSTRALIAN E-HEALTH RESEARCH CENTRE

CSIRO

AU's leading national e-Health centre

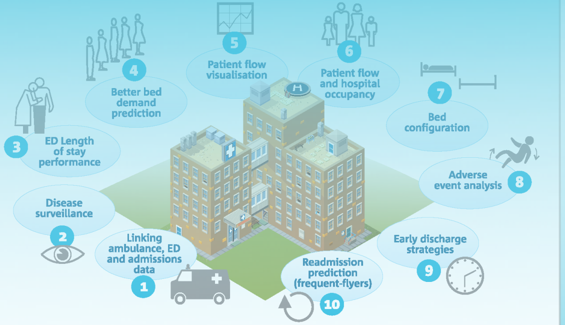
World-first clinical substantiated mobile health app

CARDIHAB

THE AUSTRALIAN
E-HEALTH
RESEARCH CENTRE

Precision medicine is at the heart of our research

Hospital forecasts


1
Linking ambulance, ED and admissions data
2
Disease surveillance
3
ED Length of stay performance
4
Better bed demand prediction
5
Patient flow visualisation
6
Patient flow and hospital occupancy
7
Bed configuration
8
Adverse event analysis
9
Early discharge strategies
10
Readmission prediction (frequent-flyers)
H

QUEENSLAND EMERGENCY MEDICINE RESEARCH FOUNDATION

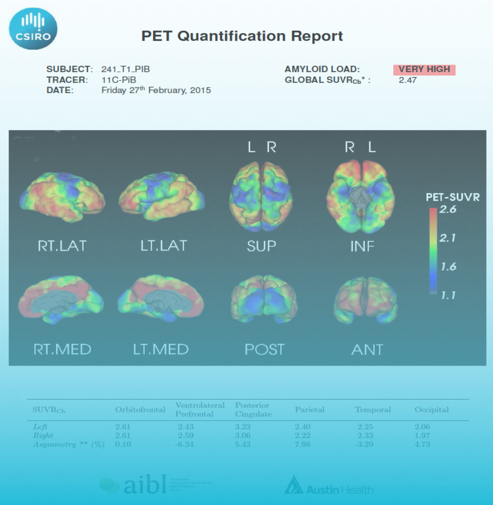
The 2009 iAWARDS Queensland State Award Winner

CeBIT AUSTRALIA

donwalker AWARDS hisa effectiveness efficiency access

The 2013 iAWARDS

Quantitative Imaging



CSIRO

PET Quantification Report

SUBJECT: 241_T1_PIB
TRACER: 11C-PiB
DATE: Friday 27th February, 2015

AMYLOID LOAD: VERY HIGH
GLOBAL SUVR$_{Cb}$* : 2.47

SUVR$_{Cb}$	Orbitofrontal	Ventrolateral Prefrontal	Posterior Cingulate	Parietal	Temporal	Occipital
Left	2.61	2.43	3.23	2.40	2.25	2.06
Right	2.61	2.59	3.06	2.22	2.33	1.97
*Asymmetry ** (%)*	0.10	-6.34	5.43	7.98	-3.29	4.73

aibl The Australian Imaging, Biomarkers & Lifestyle Flagship Study of Ageing

Austin Health

Disease risk prediction

aws

Microsoft Azure

FHIR

SAMSUNG

Genomics england

databricks

2017 iAWARDS Australian Digital Technology Award Winner

CSIRO

Precision medicine is enabled by genomics

project MinE
Make it yours

$70 Million

Global Alliance for Genomics & Health

Queensland Genomics Health Alliance

$25 Million

NSW GOVERNMENT | Genomics Strategy Health

$25 Million

$25 Million

Australian Genomics Health Alliance

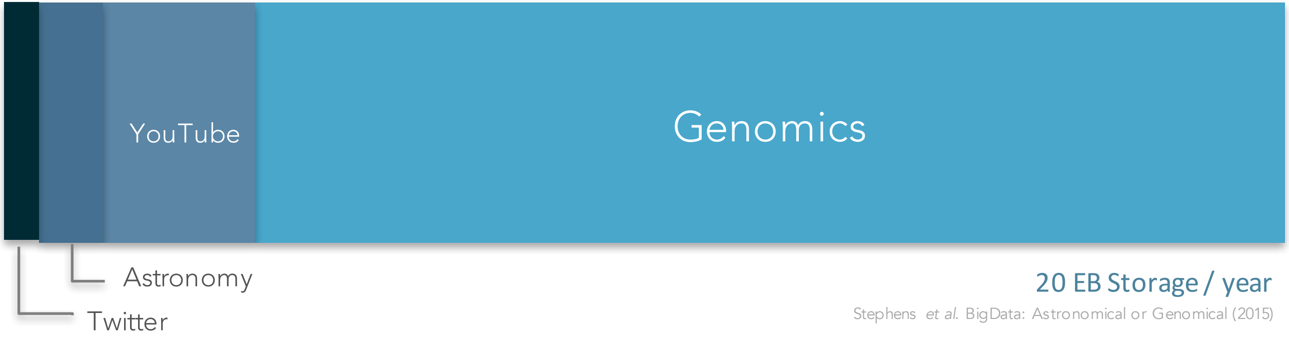
Melbourne Genomics Health Alliance

$25 Million

By **2025** it is estimated that **50%** of the world population will have been sequenced.

Frost&Sullivan

Data acquisition of BigData disciplines in 2025



20 EB Storage / year

Stephens *et al.* BigData: Astronomical or Genomical (2015)

Overview

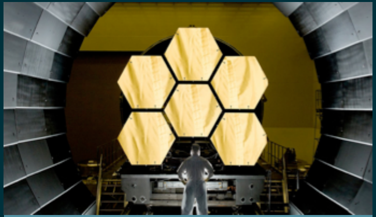
Precision medicine

… in Australia

Find Disease Genes

… using Apache Spark

VariantSpark
Machine Learning for Genomic Variants

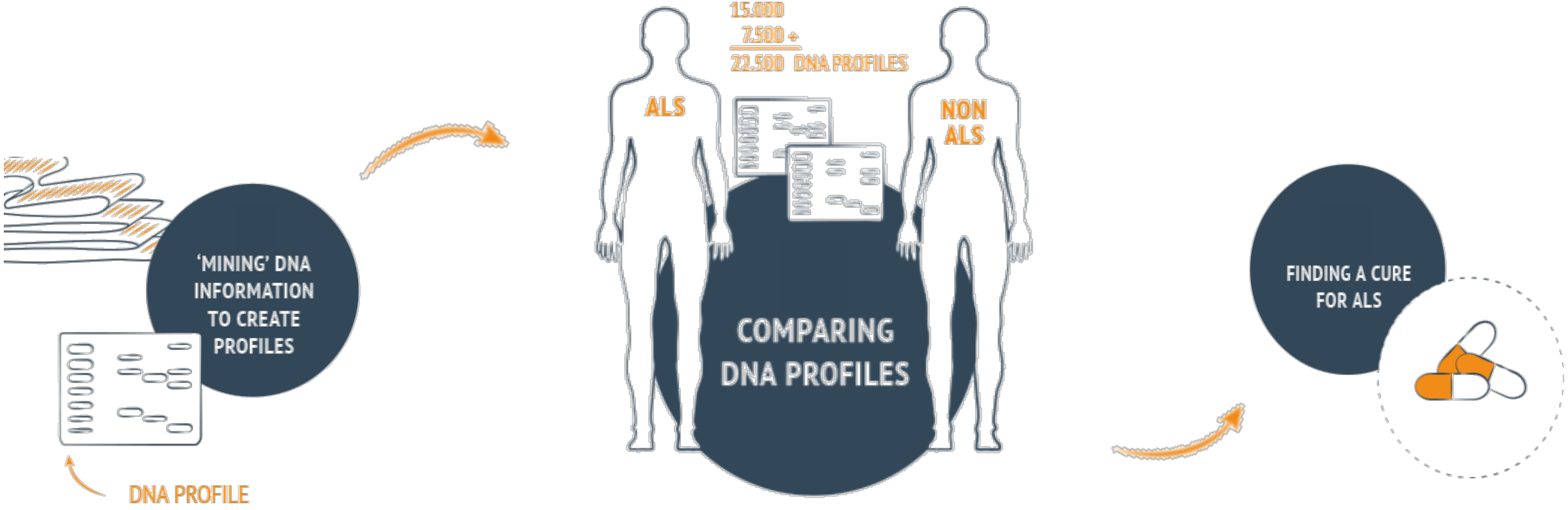


Share Disease Insight

… using serverless architecture

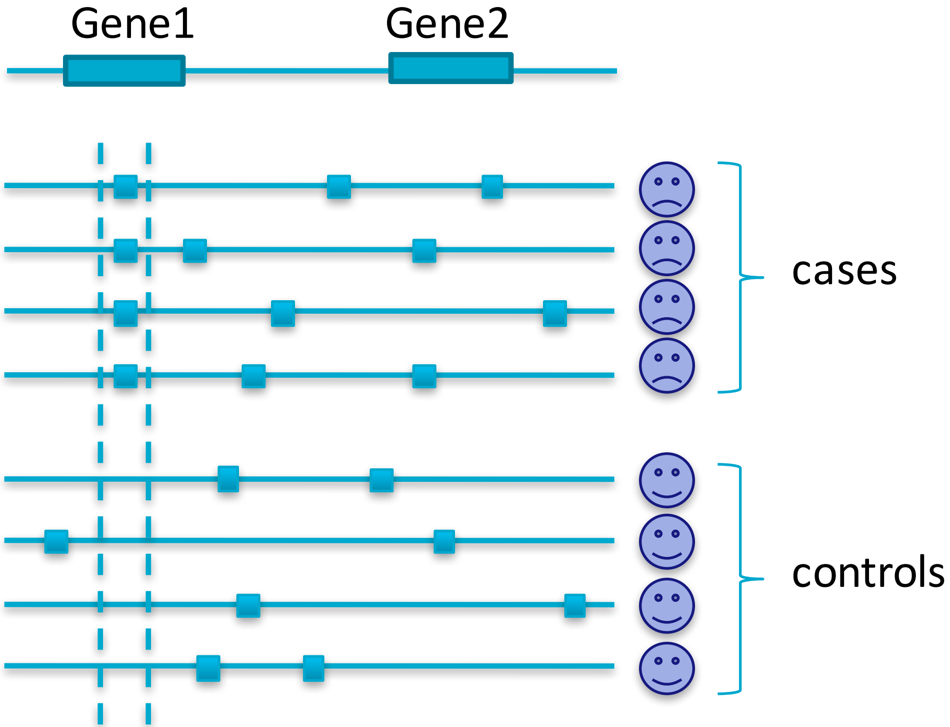
GT-Scan2
Computationally Guiding Genome Engineering

Finding the cure for ALS

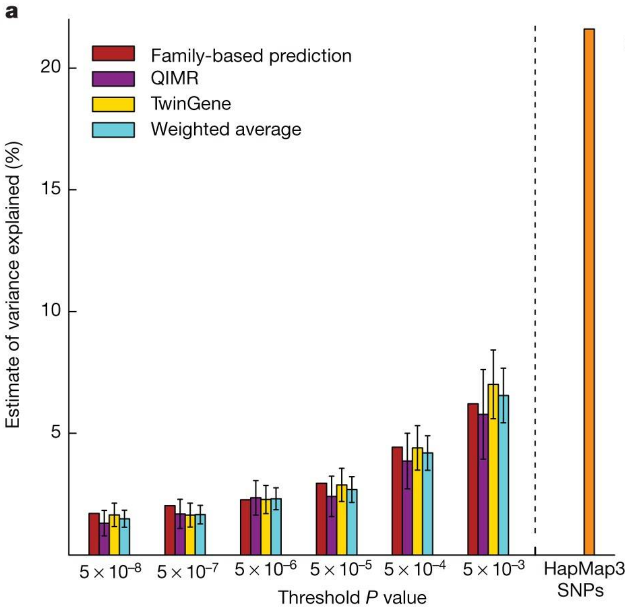


https://www.projectmine.com/about/

Finding the disease gene(s)



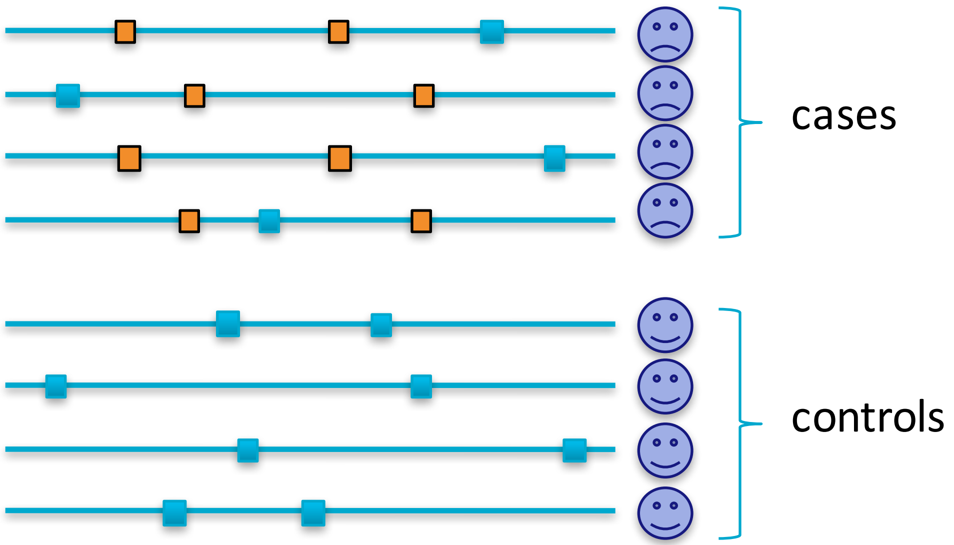
Complex diseases have polygenic risk factors

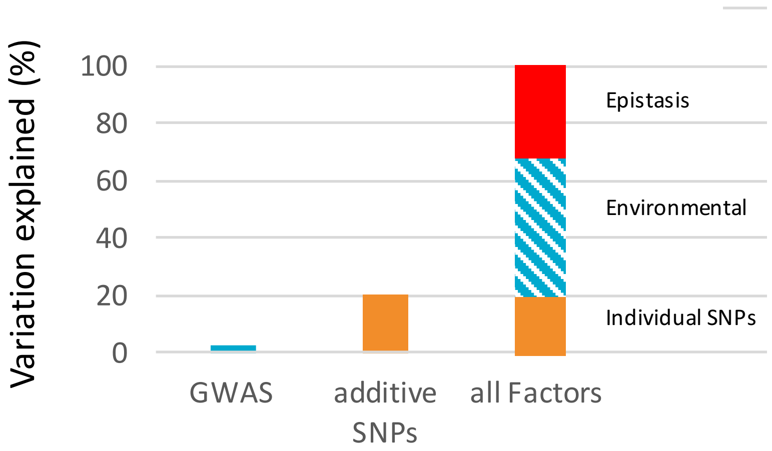


Polygene analyses suggest that SNPs with *P*-values **well below GWS add significantly** to [obesity] variance explained.

Locke *et al.* Nature 2015

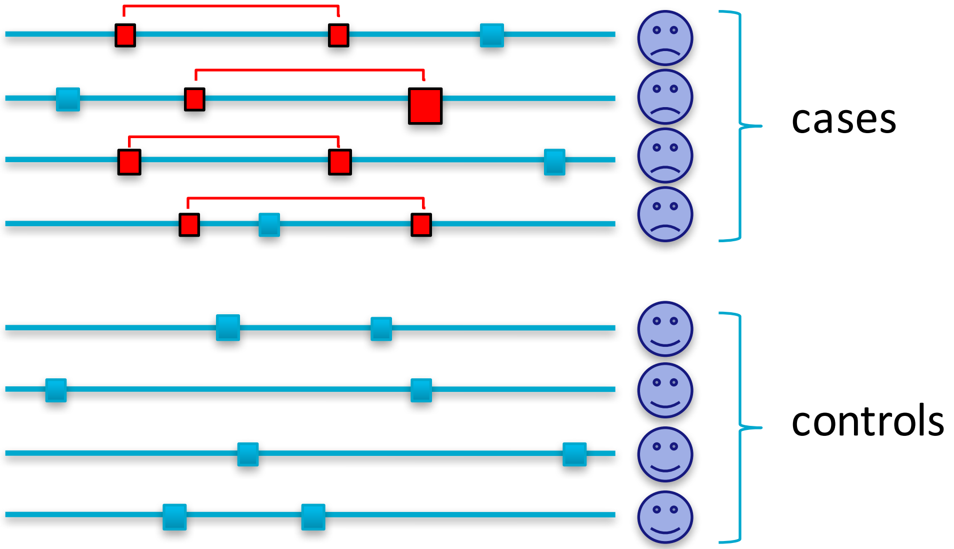
Complex diseases are driven by multiple genes





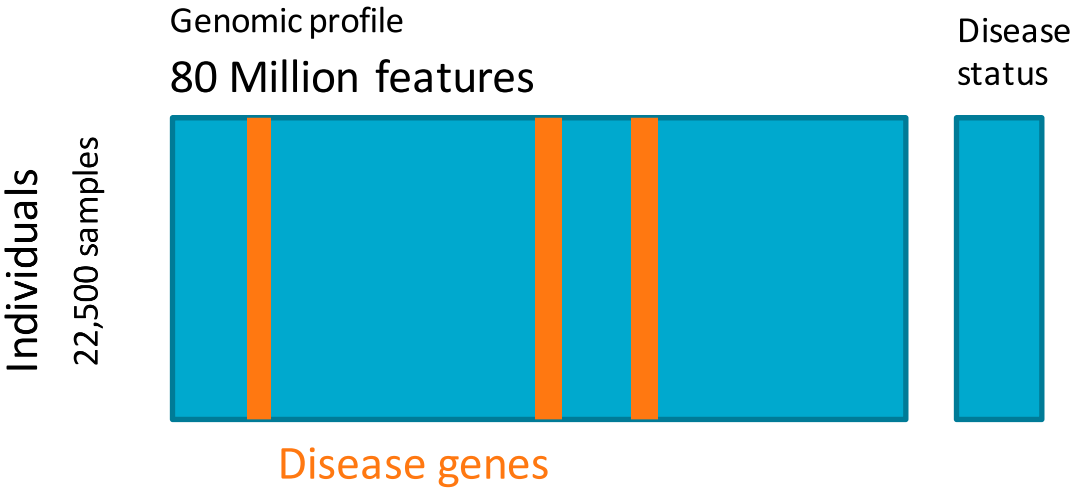
Complex diseases are driven by

multiple interacting genes with variable contribution



Need a more sophisticated ML approach, such as **Random Forest**

Machine learning on 1.7 Trillion datapoints



Genomic profile

80 Million features

Disease status

Individuals

22,500 samples

Disease genes

project MinE Make it yours

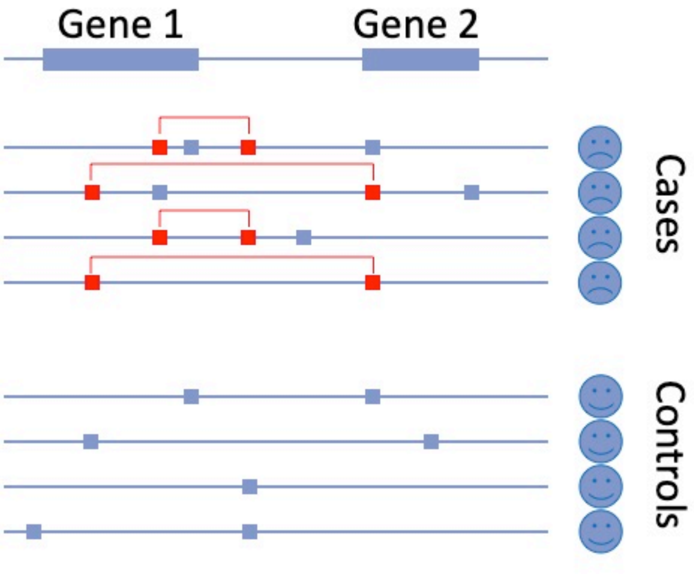
Population-scale genomic data analysis requires BigData solutions

	Desktop compute	High-performance compute cluster	Hadoop/Spark compute cluster
Focus	small data	Compute-intensive	Data-intensive
Node-bound	Yes	Yes	No
Parallelization	10 CPU	100+ CPU	1000+ CPU
Parallelization procedure	bespoke	bespoke	standardized

CSIRO solution

VariantSpark
Machine Learning for
Genomic Variants

VariantSpark: Machine Learning to find markers for complex diseases



Faster

“Analyzes 3000 individuals with 80M features in 30 minutes“

Smarter

“Requires 80% fewer samples to detect statistical significant signal“

Used by

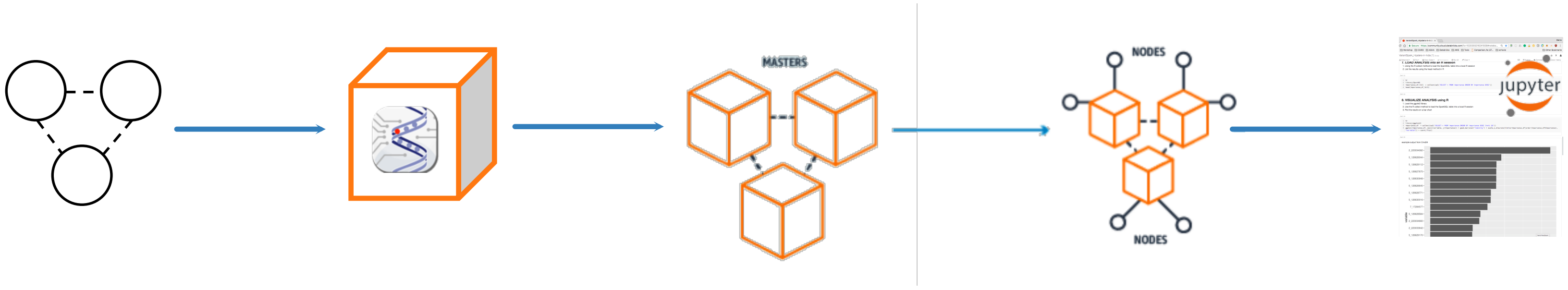
project MinE Make it yours | MACQUARIE University SYDNEY·AUSTRALIA | Microsoft Azure | aws | databricks | PATHWAY GENOMICS | SAMSUNG SAMSUNG SDS

Compute sovereignty becomes of growing importance

- Retain domestic HPC capability but enable **cloud-like flexibility** (e.g. multi-tenancy)
- **Connect international** datasets without lag
- **-> SuperCloud**

Technical details: March 13th 1630 – 1700
SuperCloud – The evolution of HPC to Software Defined Computing
Mr. Jacob Anders, CSIRO and
Mr. Garry Swan, CSIRO

Deploying cloud-like workflows on HPC



Access to bare metal instances

Create Docker for VariantSpark

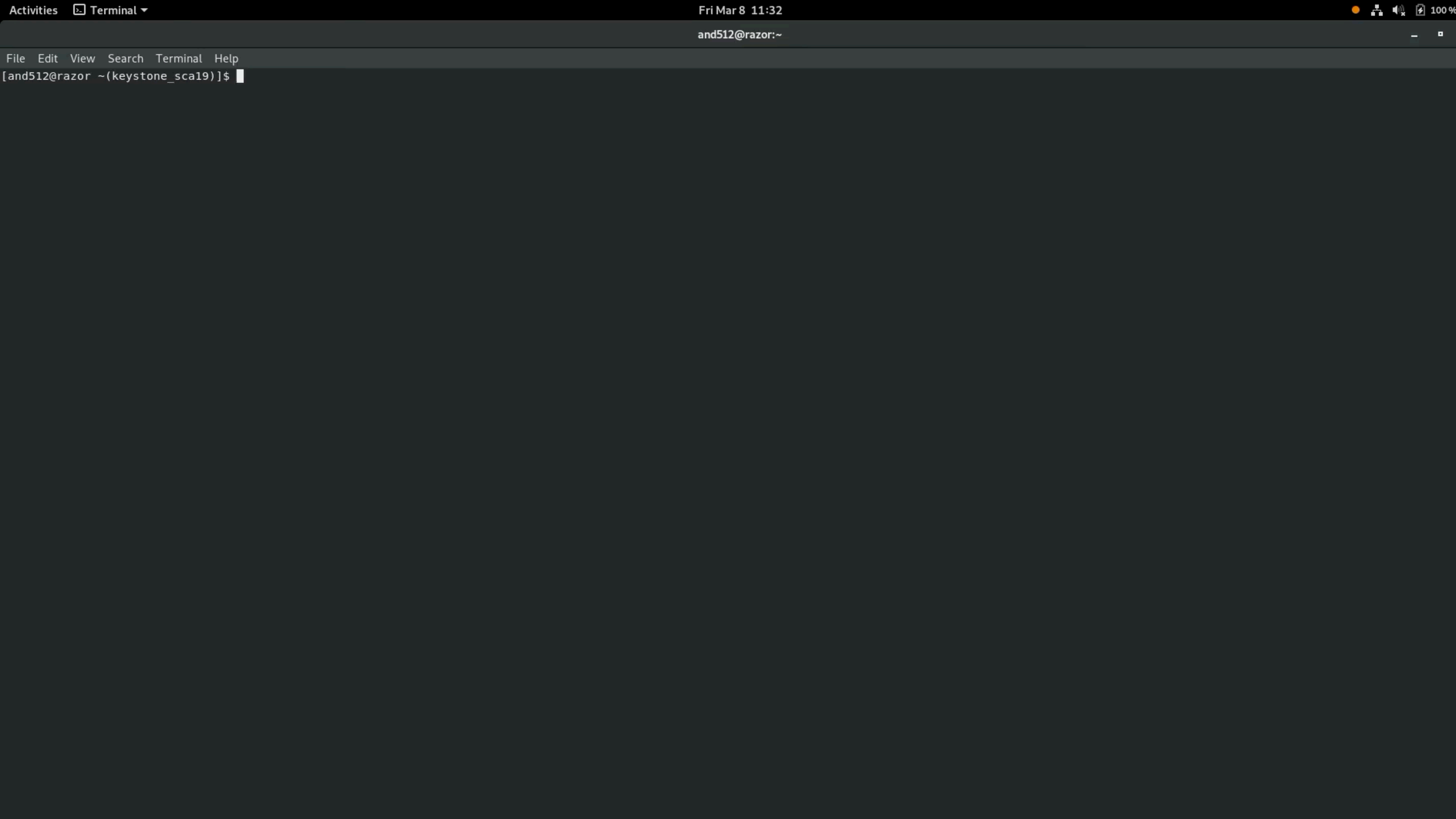
Enable instances to run docker

Provision Spark-cluster

Connect Jupyter to trigger runs and capture output

Demo

Genetic
Hipster-Index
Powered by
VariantSpark
Machine Learning for
Genomic Variants



```
[and512@razor ~(keystone_sca19)]$ 
```

Overview

Precision medicine

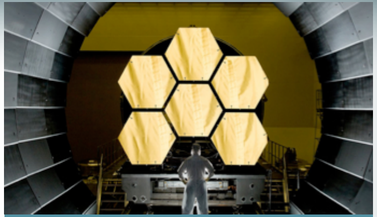
… in Australia

CSIRO

Find Disease Genes

… using Apache Spark

VariantSpark
Machine Learning for Genomic Variants



Share Disease Insight

… using serverless architecture

GT-Scan2
Computationally Guiding Genome Engineering

Beacon:
sharing genomic data

- Today, **70 Beacon are lit** to share information about rare genetic diseases.
- **Reduce cost** to enable more sharing.
- Using **serverless (FaaS)** technology.

Recruiting instantaneous appropriately powered compute

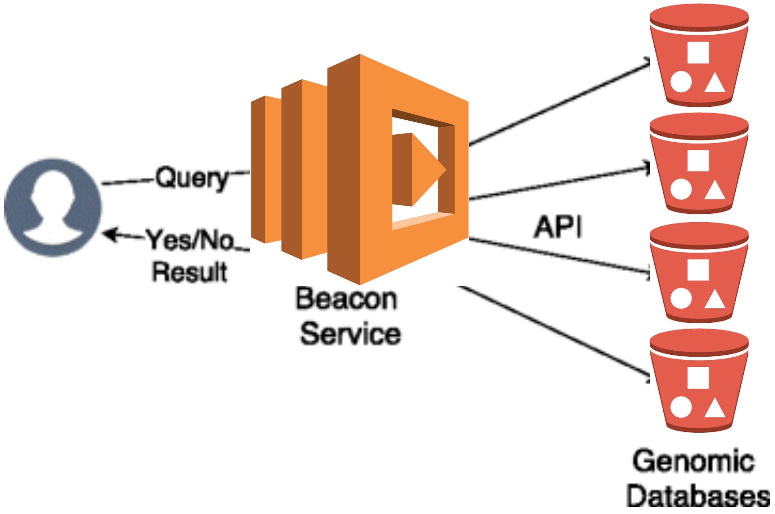
	Desktop compute	High-performance compute	Hadoop/Spark	Serverless
Focus	small data	Compute-intensive	Data-intensive	Agility
Node-bound	Yes	Yes	No	No
Parallelization	10 CPU	100+ CPU	1000+ CPU	1-1000+ CPU
Parallelization procedure	bespoke	bespoke	standardized	standardized
Overhead in the cloud	NA	spin-up lag	spin-up lag	instantaneously

CSIRO solution

VariantSpark
Machine Learning for
Genomic Variants

Beacon
Serverless lookup of geno-
types and frequencey

Serverless-Beacon to scale up discovery across continents



Powerful

“Scaling up to large volumes of distributed variant data.“

Cheaper

“Only pay for the resources consumed – zero downtime cost.“

Used by

Global Alliance for Genomics & Health

aws

Australian Genomics Health Alliance

Melbourne Genomics Health Alliance

Beacon
Serverless lookup of geno-types and frequencey

Three things to remember

- Complex multigenic diseases should be studied using **‘wide’ ML** (VariantSpark).
- **Serverless architecture** makes even data-intensive web-apps affordable (Serverless Beacon).
- SuperCloud offers an exciting and potentially cheaper supplement to public cloud providers: **let’s build a healthier future together!**

Let's build a healthier world together

Team

Denis Bauer, PhD

Oscar Luo, PhD

Laurence Wilson, PhD

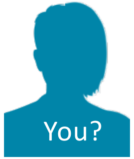
Aidan O'Brien

Natalie Twine, PhD

Arash Bayat

Brendan Hosking

We are hiring...



...email Denis

DATA 61 CSIRO

Rob Dunne, PhD

Piotr Szul

Collaborators

MACQUARIE University SYDNEY·AUSTRALIA

Lynn Langit

Microsoft Azure

databricks

ANU

aws

Alibaba Cloud

Epsagon

cloudera

DIUS

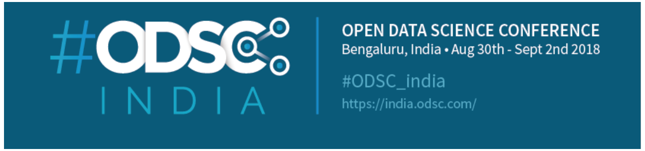
Software

- GT-Scan2 – Computationally Guiding Genome Engineering
- VariantSpark – Machine Learning for Genomic Variants
- Tribes – Detecting distantly related individuals
- NGSANE – Production Informatics for High Throughput Data
- GenPhen-Insight – Genome-Phenome Discovery Framework
- Beacon – Serverless lookup of genotypes and frequency

News

Top 10 Australian IT stories of 2017

ComputerWeekly.com Top 10 IT stories of 2017 CW+ Content


#ODSC INDIA

OPEN DATA SCIENCE CONFERENCE
Bengaluru, India • Aug 30th - Sept 2nd 2018

#ODSC_india
https://india.odsc.com/

KEYNOTE

aws PUBLIC SECTOR SUMMIT 2018