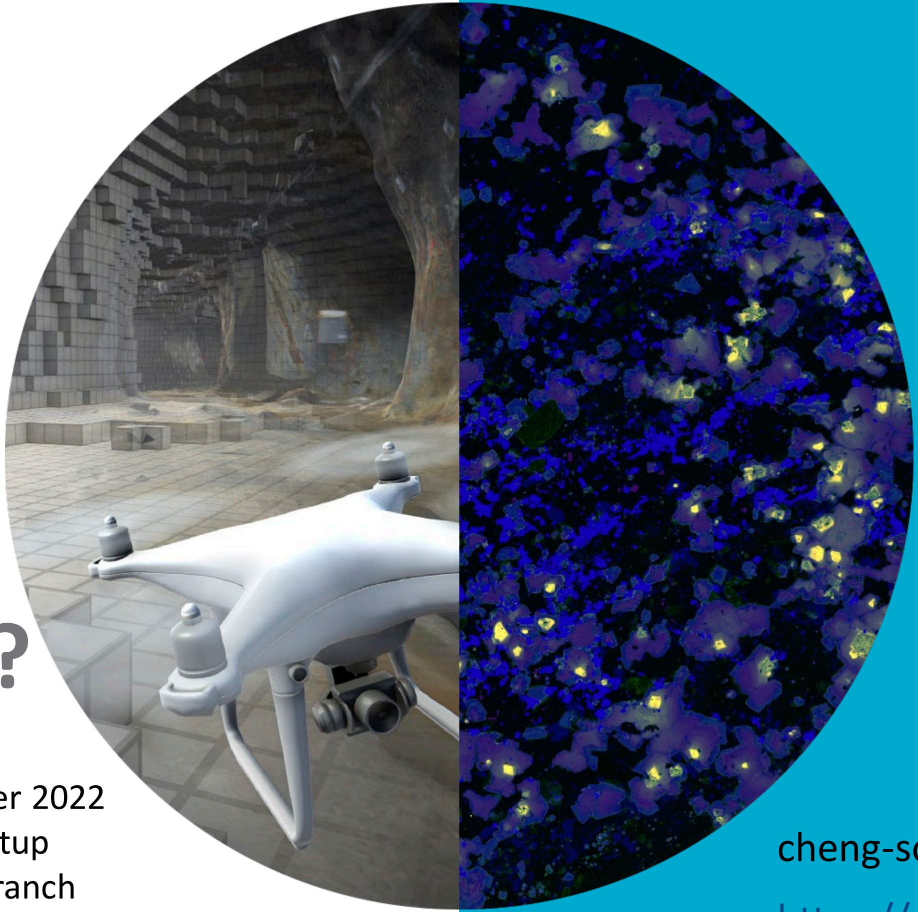
What is a scientific instrument?



Cheng Soon Ong | 11 October 2022
Canberra Data Scientists Meetup
Statistical Society Canberra Branch

Australia's National Science Agency

cheng-soon.ong@data61.csiro.au

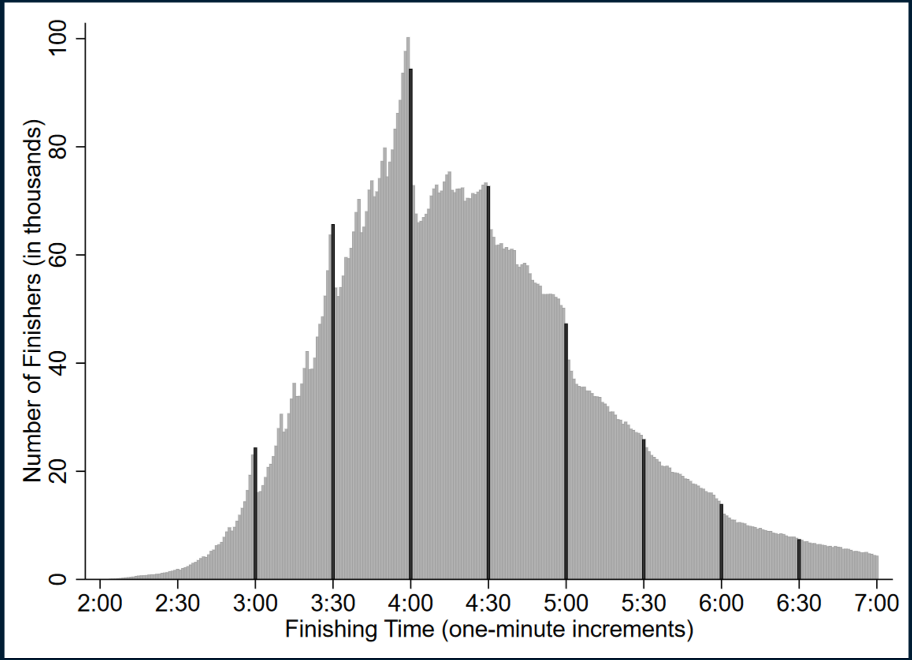
https://research.csiro.au/mlai-fsp/

I would like to acknowledge the Ngunnawal - Ngambri people as the Traditional Owners of the land that we’re meeting on today, and pay my respect to their Elders past and present.

What is the distribution of running speeds?

a group of academics running a marathon while holding a microscope, anime style (DALL-E)

Distribution of marathon finishing times (n=9.5m)



How you measure affects what you observe

Allen, Dechow, Pope, Wu (2016)
Reference-Dependent Preferences: Evidence from Marathon Runners

What is a scientific instrument?

- What is data?

- How to deal with non-tabular data?
  - Case studies in environmental observation (microscope, satellite, computer)
- Where does data come from?
  - Case study in genome biology (organism)
- Opportunities and challenges for data science

A fake HR database

Name	Gender	Degree	Postcode	Age	Annual salary
Aditya	M	MSc	W21BG	36	89563
Bob	M	PhD	EC1A1BA	47	123543
Chloé	F	BEcon	SW1A1BH	26	23989
Daisuke	M	BSc	SE207AT	68	138769
Elisabeth	F	MBA	SE10AA	33	113888

Data in numerical format

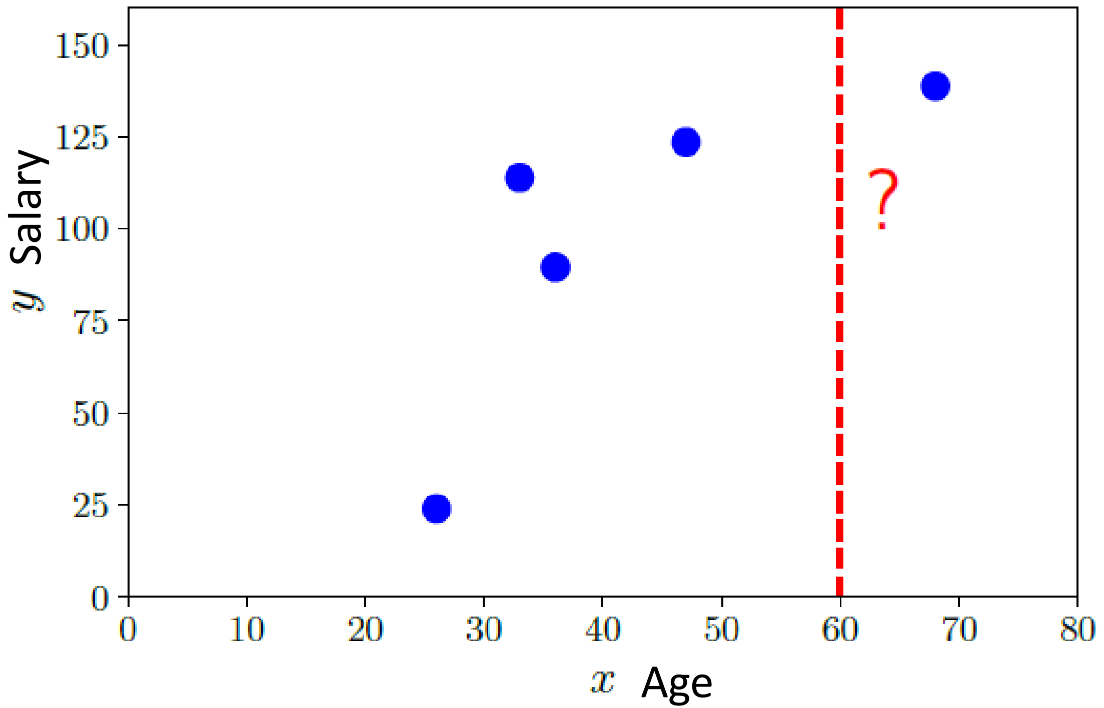
Gender ID	Degree	Latitude (in degrees)	Longitude (in degrees)	Age	Annual Salary (in thousands)
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888

binary

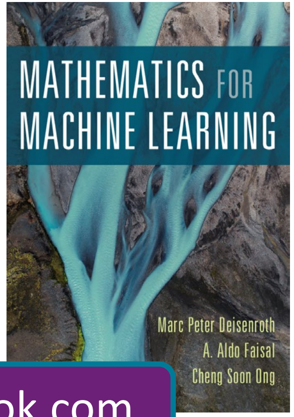
ordered category

postcode

Predict salary given age (ML is about prediction)



Gender ID	Degree	Latitude (in degrees)	Longitude (in degrees)	Age	Annual Salary (in thousands)
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888



mml-book.com

Who we are

Australia's national science agency

One of the world's largest multidisciplinary science and technology organisations

5,200+ dedicated people working across 58 sites globally

State-of-the-art national research infrastructure

We delivered $7.6 billion of benefit to the nation in FY21

Global megatrends in data and AI

Australia's National Science Agency

Our Future World

Global megatrends impacting the way we live over coming decades

July 2022

5. **Diving into digital:** the pandemic-fuelled a boom in digitisation, with teleworking, telehealth, online shopping and digital currencies becoming mainstream. Forty percent of Australians now work remotely on a regular basis and the future demand for digital workers expected to increase by 79% from 2020 to 2025.

6. **Increasingly autonomous:** there has been an explosion in artificial intelligence (AI) discoveries and applications across practically all industry sectors over the past several years. Within the science domain the use of AI is rising with the number of peer-reviewed AI publications increasing nearly 12 times from 2000 to 2019.

https://www.csiro.au/en/research/technology-space/data/our-future-world

MLAI Future Science Platform

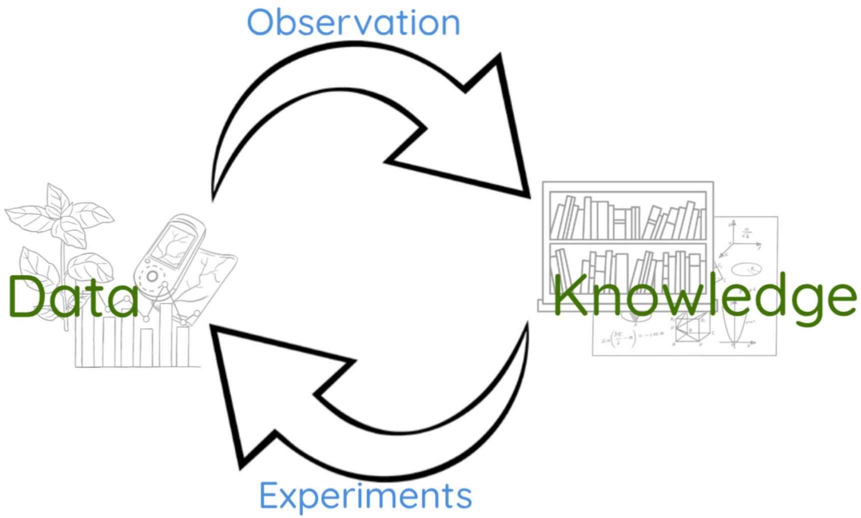
How to use prediction to help perform scientific discovery?

30 postdoc researchers

10 senior scientists

1 vision

Machine learning for scientific discovery



Gender ID	Degree	Latitude (in degrees)	Longitude (in degrees)	Age	Annual Salary (in thousands)
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888

How to deal with non-tabular data?

Towards automated detection of harmful algae and toxic blooms

- Microscopy on water samples
- Computer vision
  - Create bounding boxes
  - Classify algae species

https://blog.csiro.au/using-artificial-intelligence-to-detect-harmful-algae/


Chris Jackett

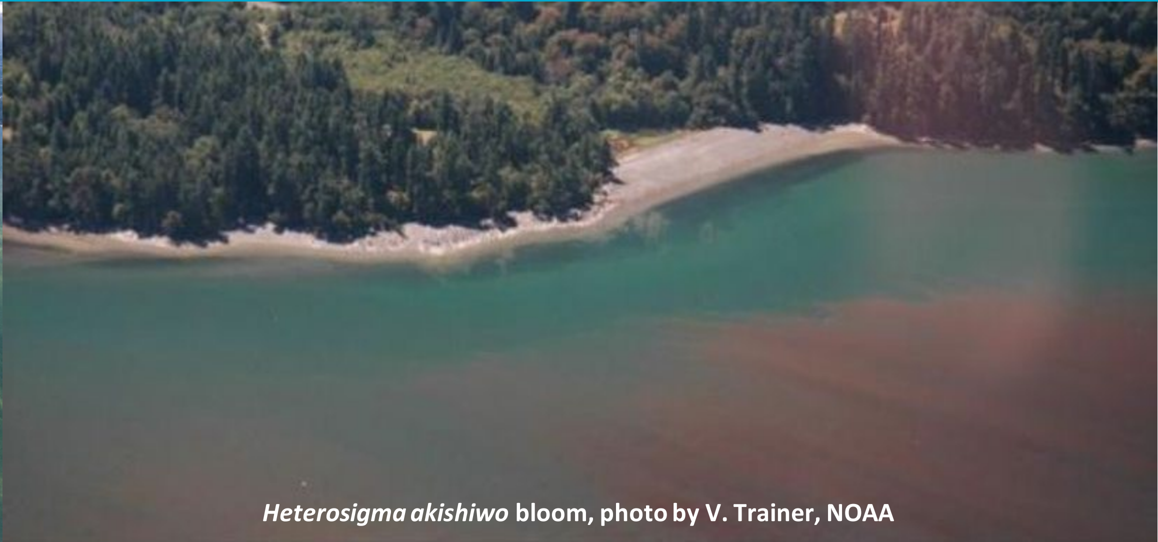

Viv Rolland


Pete Thrall

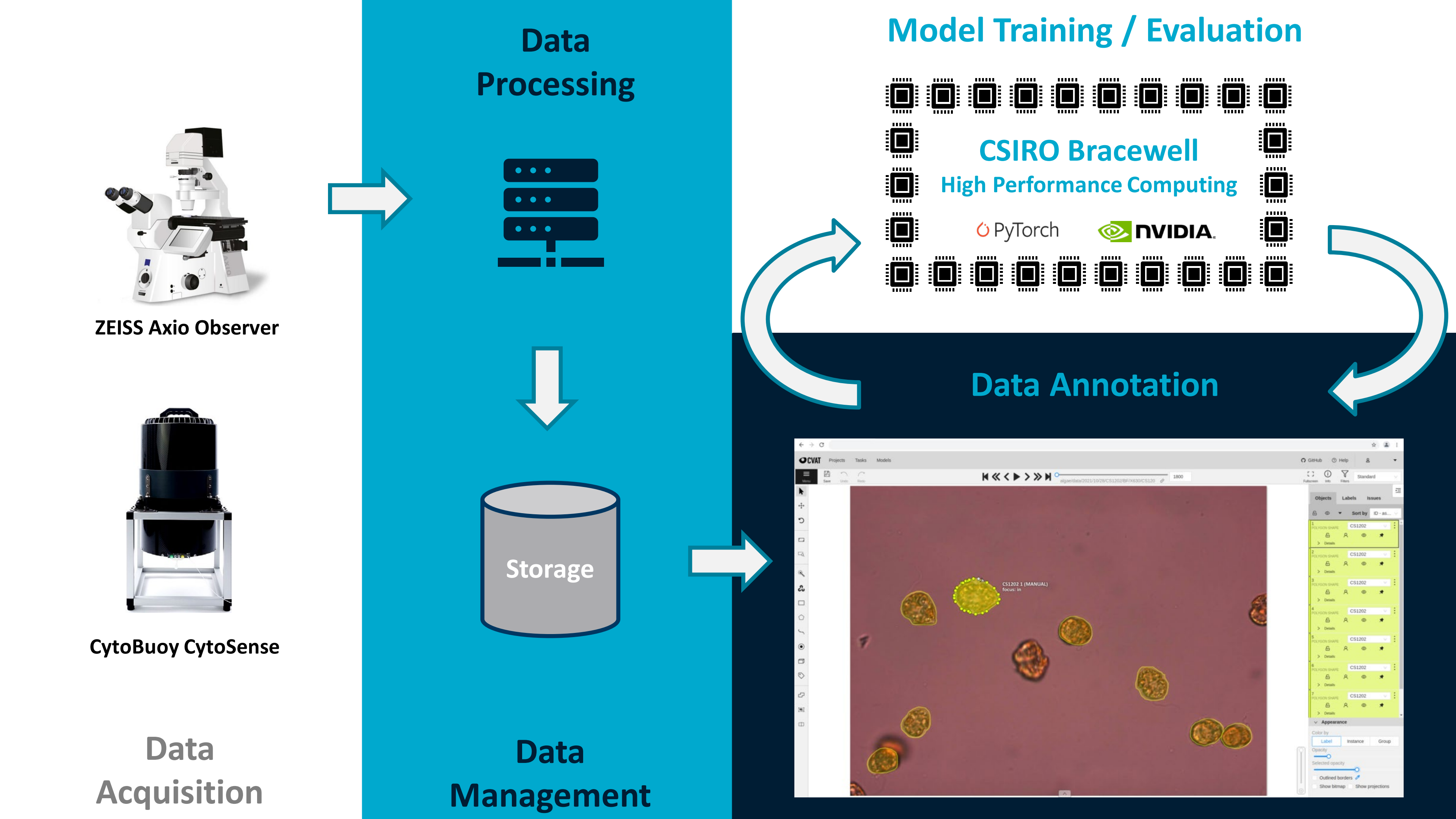
Harmful Algal Blooms (HABs)

- Damaging impact on the environment and aquatic organisms
- **2012** – Tasmanian east coast - dinoflagellate *Alexandrium catenella* closed the seafood industry
- **2016** - Murray Darling River – toxic blue-green algae impacted drinking water, agriculture or recreation
- **2018/19** - Murray Darling River - toxic blue-green algae resulted in high fish mortalities

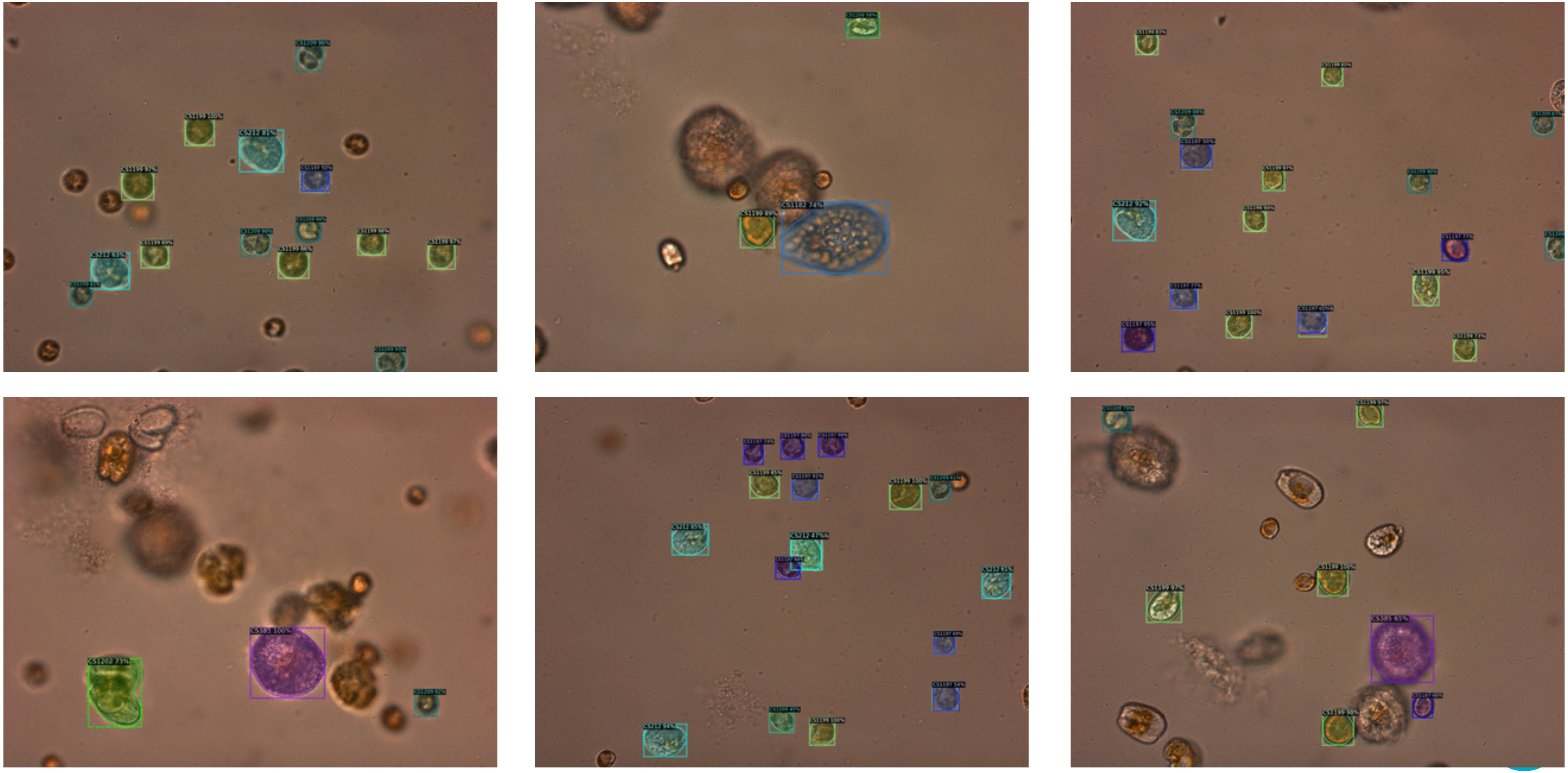
Aerial Associates Photography, Inc. photo by Zachary Haslick



Heterosigma akishiwo **bloom, photo by V. Trainer, NOAA**


Data Acquisition
ZEISS Axio Observer
CytoBuoy CytoSense
Data Processing
Storage
Data Management
Model Training / Evaluation
CSIRO Bracewell
High Performance Computing
PyTorch
NVIDIA
Data Annotation
CVAT

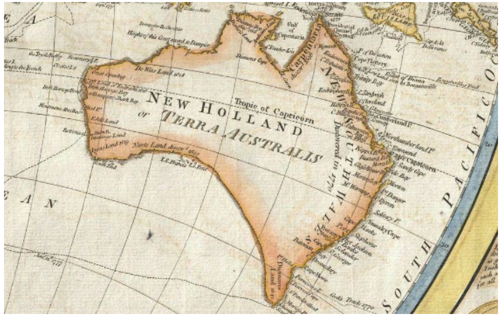
Mask R-CNN Results – Mixed Species Assemblage



What is a map?

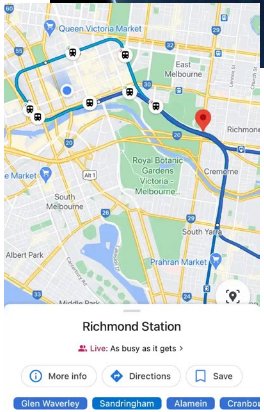
Modern maps augment our understanding of events

Paper



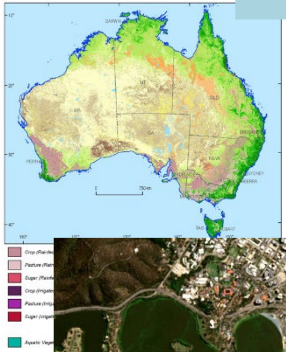
Samuel Dunn (1794)

Electronic



The Guardian (25/2/21)

Multi modal prediction



Geoscience Australia

Canberra Times (11/3/21)

Plant biodiversity mapping in Australia with satellite imagery

- Diversity of plants is key in maintaining stability and productivity of ecosystems
- Spaceborne remote sensing

Yiqing Guo

Cindy Ong

Shaun Levick

Peyman Moghadam

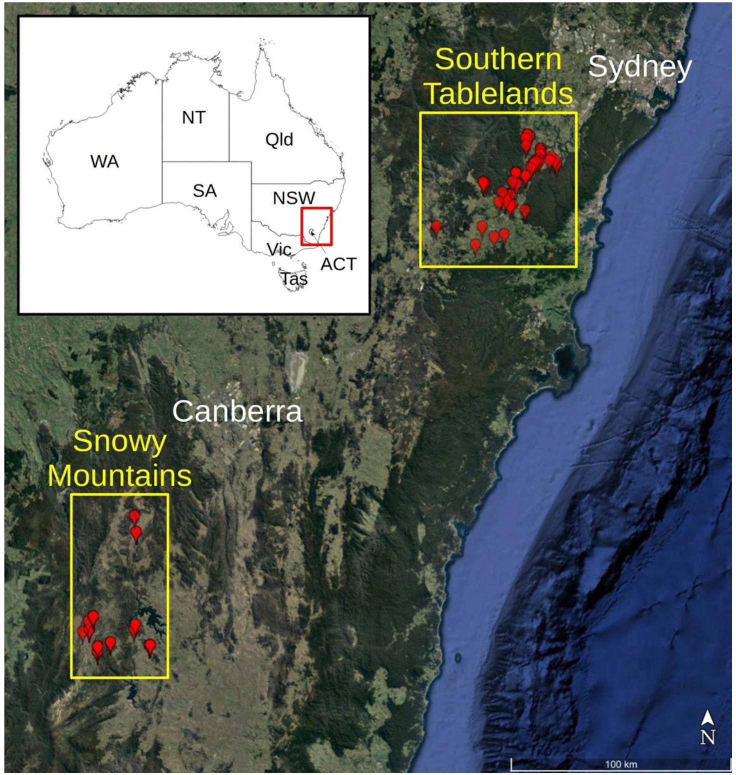
Y. Guo, K. Mokany, C. Ong, P. Moghadam, S. Ferrier, and S. R. Levick (2022). Quantitative assessment of DESIS hyperspectral data for plant biodiversity estimation in Australia. IGARSS 2022.

Study Area

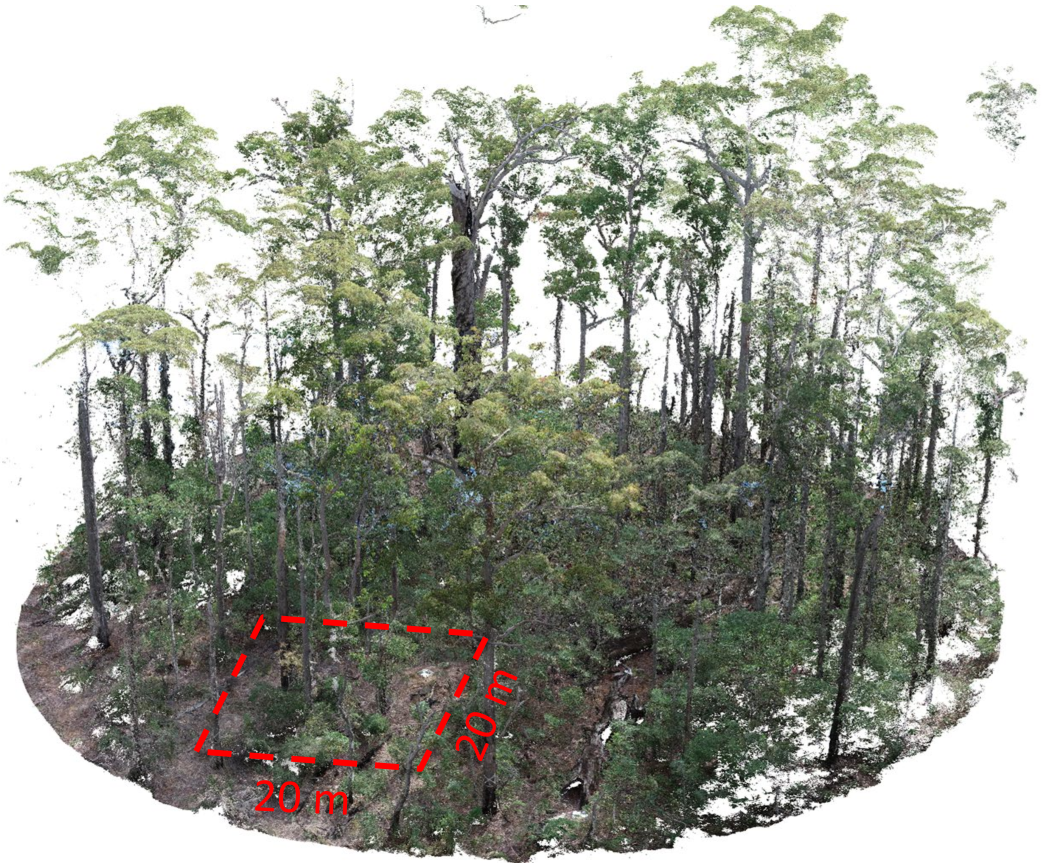
We focused on two regions in southeast Australia

- **Southern Tablelands**
- **Snowy Mountains**

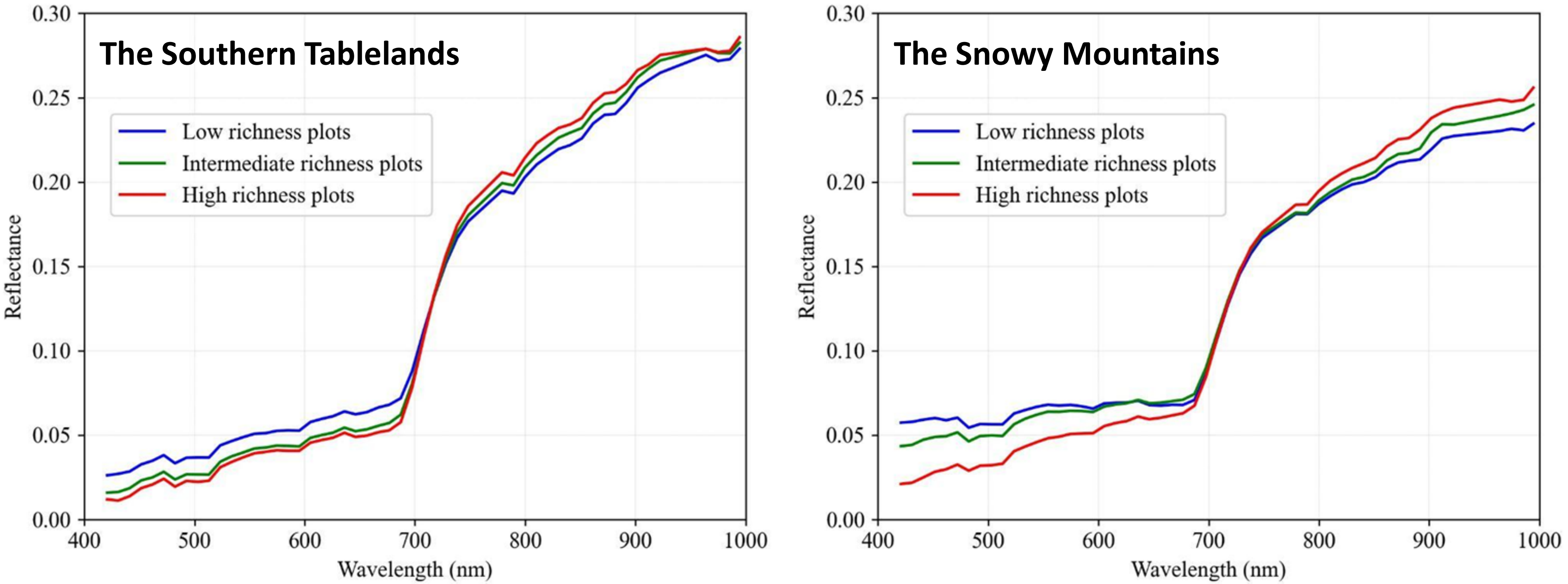
	Southern Tablelands	Snowy Mountains
Number of Samples	44	29
Geo-extent	34°12’26”–34°39’07”S 150°05’57”–150°40’51”E	35°43’58”–36°16’30”S 148°23’16”–148°39’02”E
Sampling Time	Feb 19, 2017~Dec 07, 2017	Feb 24, 2016~Dec 13, 2017
Plot Size	400 sq m	400 sq m



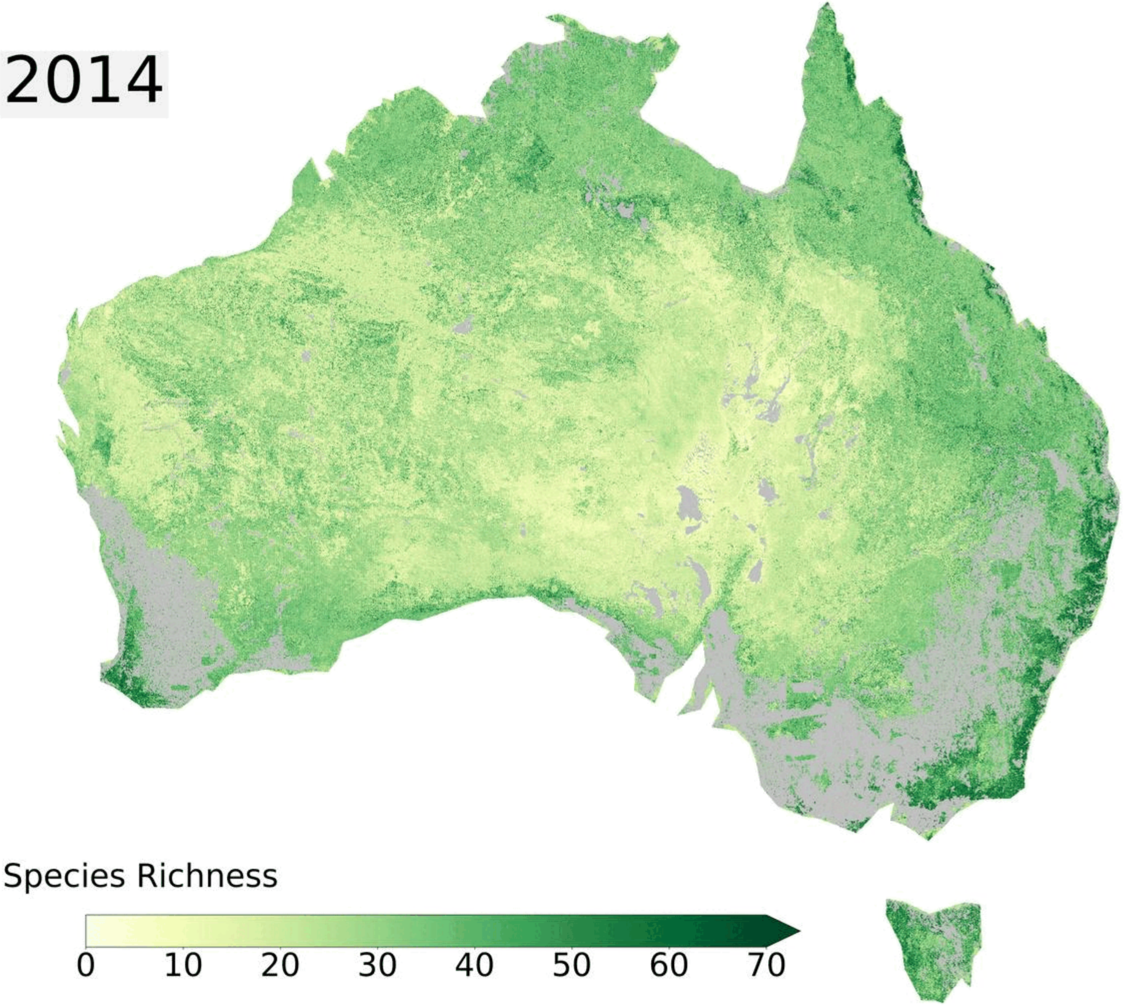
Plant Species Richness (Alpha Diversity)



Spectral Reflectance for Low, Intermediate, and High Species Richness Plots



Mapping Result

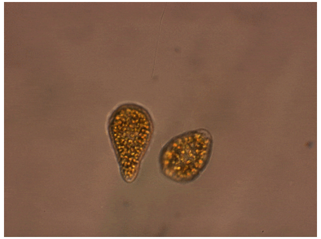


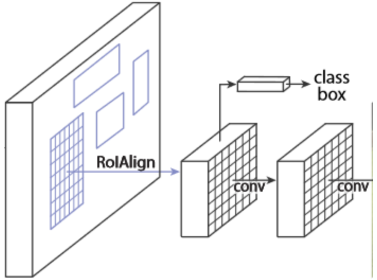
Key Info:

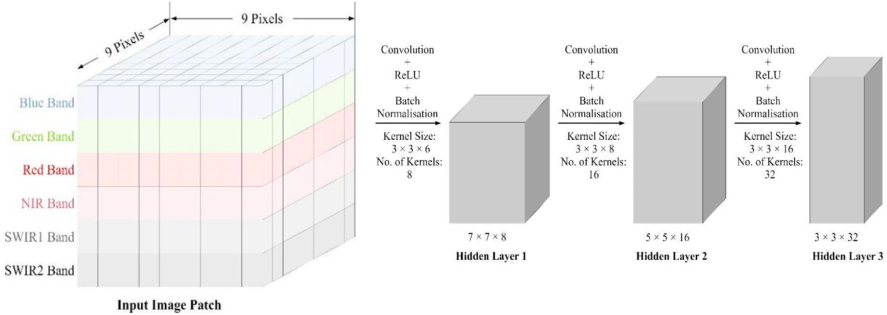
(1) > 70k ground truth samples in total over Australia

(2) Time of survey ranges from 1986 to date

Deep learning provides tools to create embeddings







Bounding box,
Category
Real value

"Predictor"

Labels

Vector data

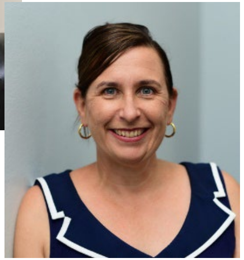
Representation learning

Raw data

A Spatio-Temporal Neural Network Forecasting Approach for Emulation of Firefront Models

- Use domain knowledge
- Evolution of spatial information
  - More efficient
  - Predictive uncertainty
  - Enable scenario planning

Carolyn Huston



Petra Kuhnert

Spark

Wildfire simulation toolkit

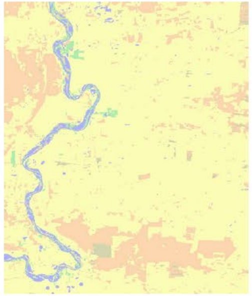
https://research.csiro.au/spark/

Weather timeseries

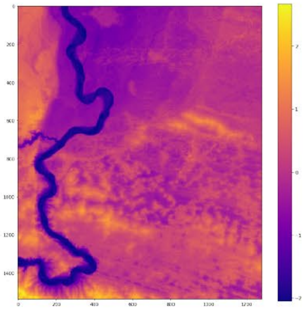
$$\frac{\partial \psi}{\partial t} = \vec{v}|\nabla \psi|$$

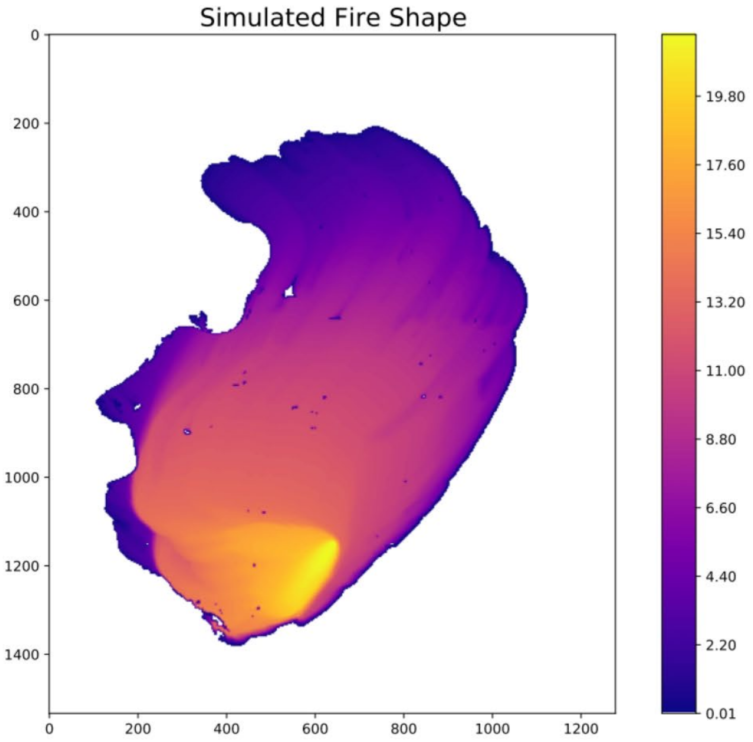
Level set

Fuel Class

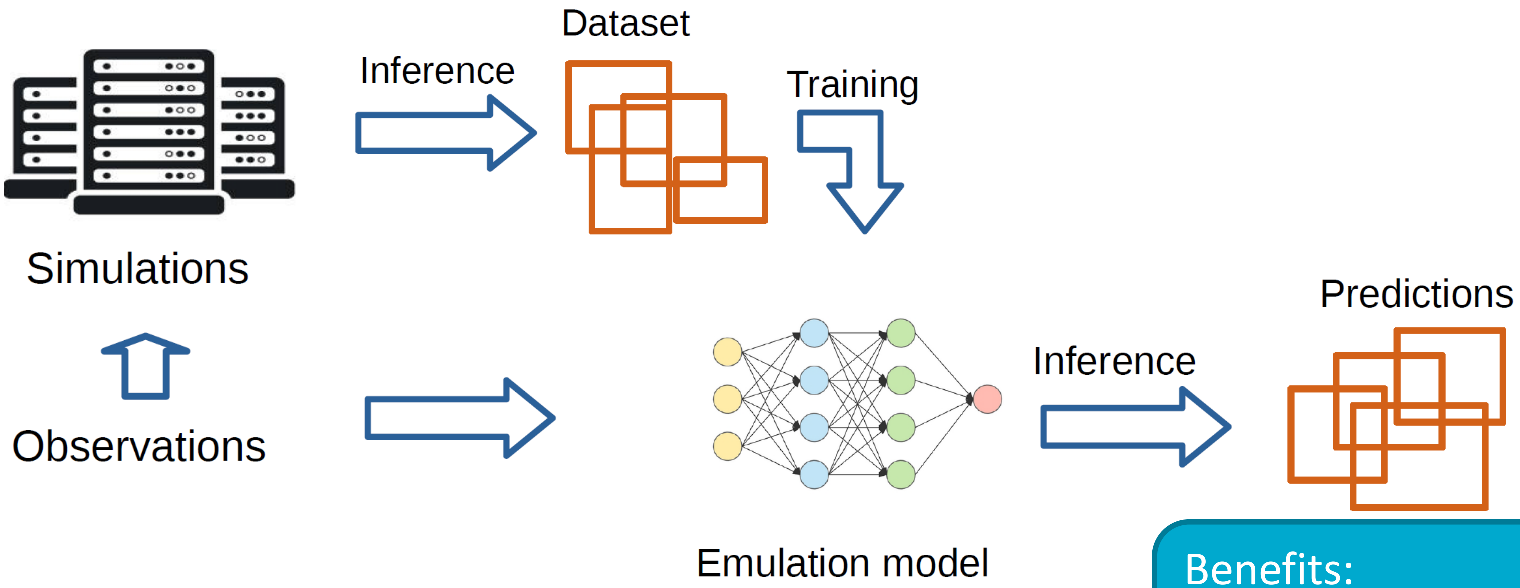


Height-map

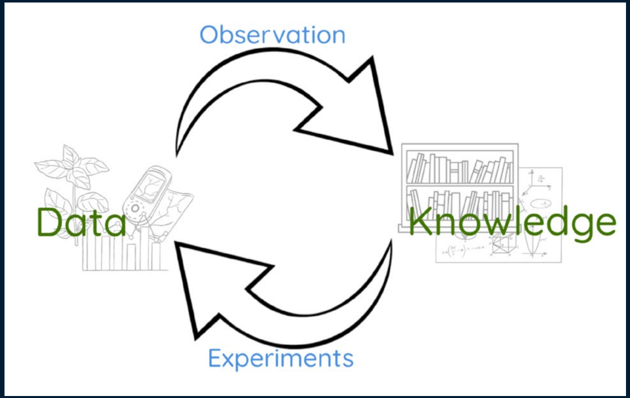




Model emulation



Benefits:
- More efficient
- Predictive uncertainty

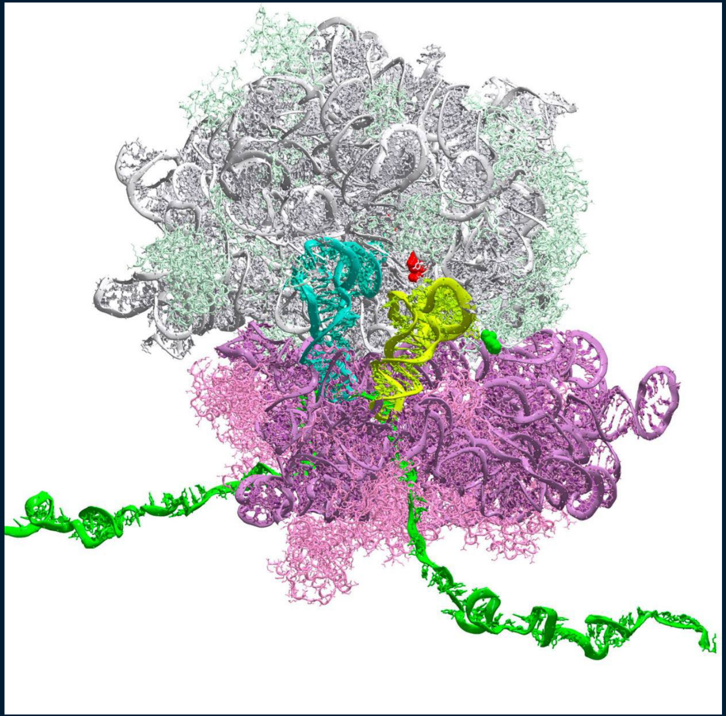


Where does data come from?

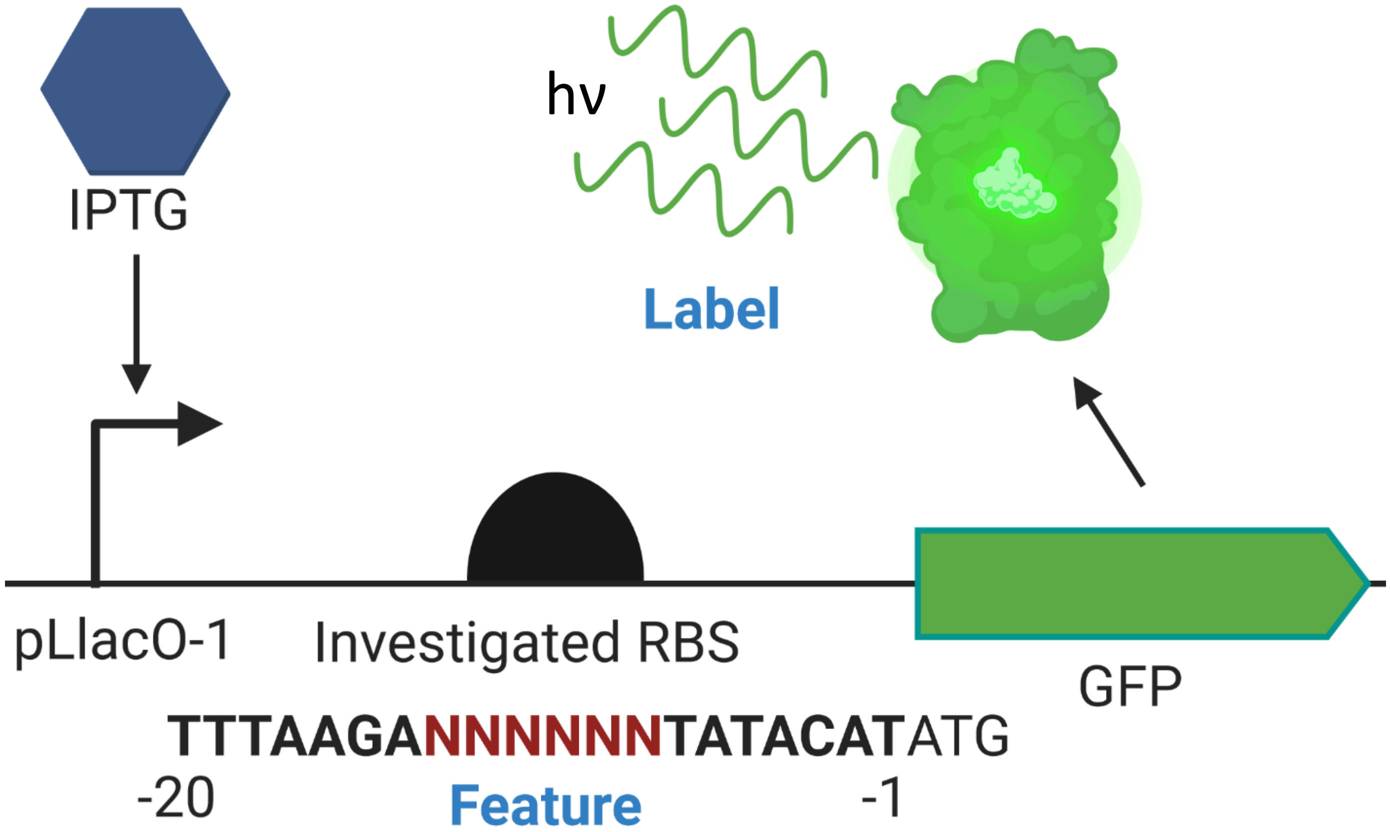
Adaptive design

- Genomic sequencing revolution
  - Fast and cheap
  - Portable
- Biological factories
  - Drug design
  - Alternative foods

Which genome should we grow?



Genetic Device Design



MLAI augmented SynBio

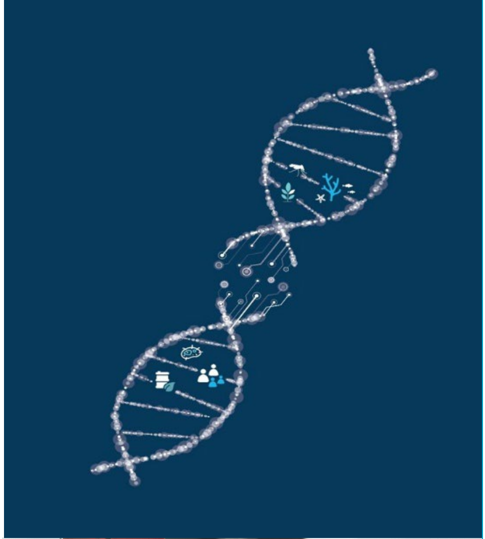
- **Working definition of 'synthetic biology':**
  The design and construction of DNA-encoded parts, devices, machines, and organisms; and their application for useful purposes.
- Experimental science domains
  - Integrative Biological Modelling
  - Engineering Novel Biological Components
  - Assembling Novel Biosystems
- Application areas
  - Mosquito borne diseases
  - Bacterial biofilms
  - Chemical synthesis using yeast

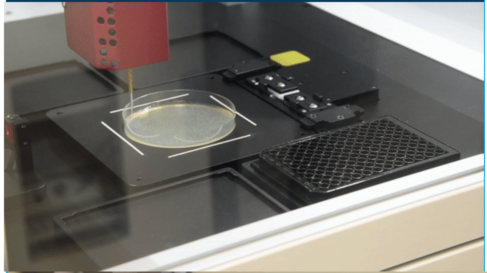
Claudia Vickers

Janet Reid

Alison Rice

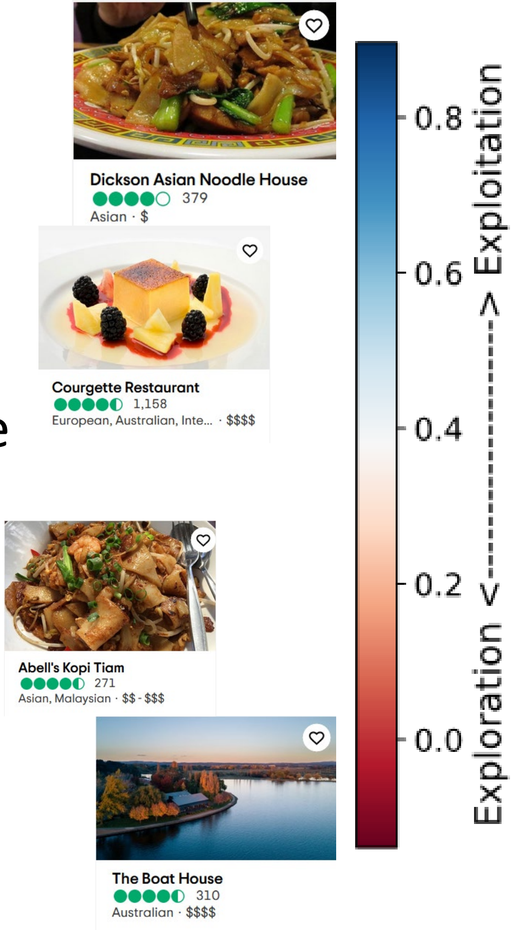
https://research.csiro.au/synthetic-biology-fsp





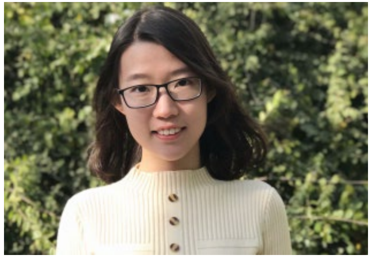
Still too many options to try!

- Each option has a measurable outcome
  - Efficacy of drug
  - Amount of protein
- Study conditions limit the precision we can measure



- Multi armed bandits
  - Maximise outcomes
  - Trade of exploration and exploitation

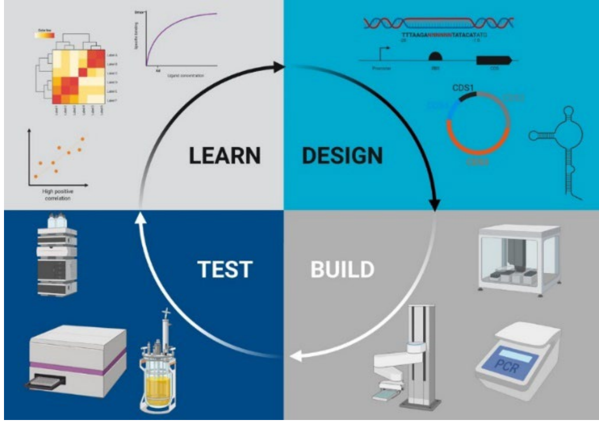
Algorithms



Mengyan Zhang, ANU

Maciej Holowko
Nourish Labs

Huw Hayman
Zumpe, SynBio



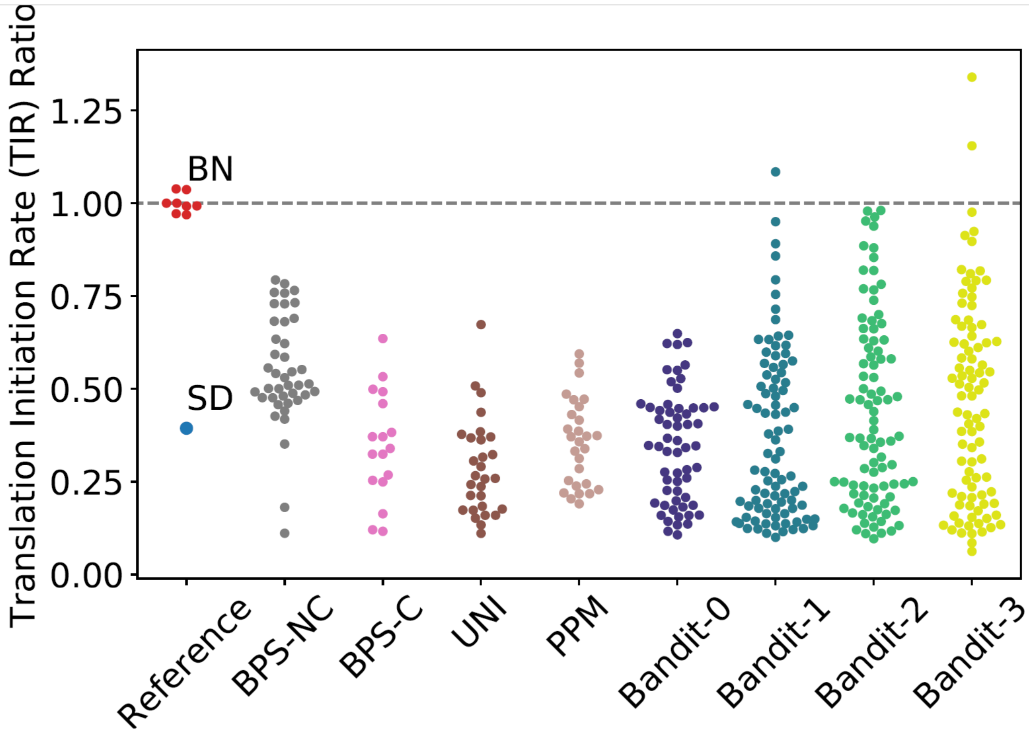
1. A (Bayesian) regression algorithm which predicts both
   - Mean
   - Uncertainty

→ Gaussian Process Regression (aka Kriging) → LEARN

2. An online/batch algorithm which recommends sequences to design

→ Multiarmed Bandits Algorithms: Upper Confidence Bound → DESIGN

AI recommends good designs



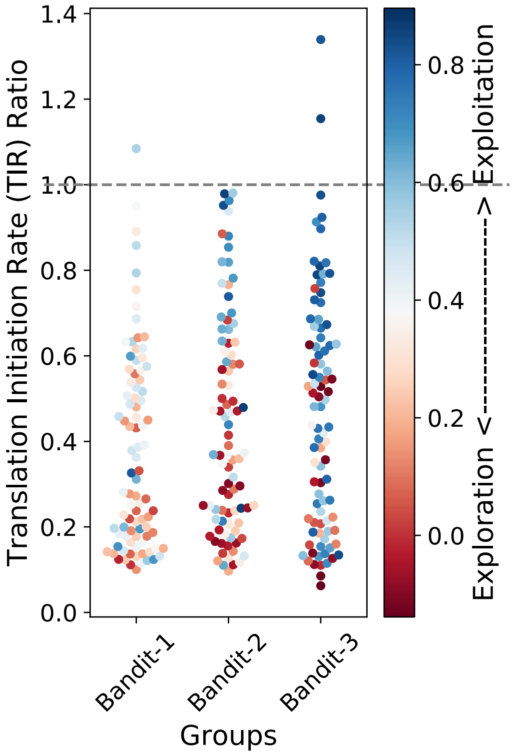
- Hard to search by evolving sequences
- 4 experimental cycles
- 35% stronger than engineered sequence

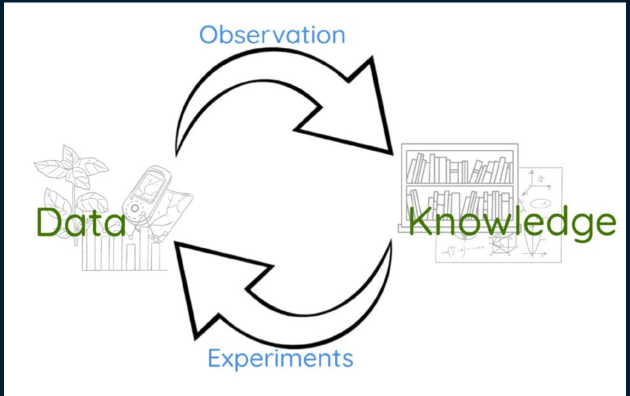
Zhang, Holowko, Hayman Zumpe, and Ong,
Machine learning guided design for ribosome binding site.
ACS Synthetic Biology, 2022

Exploration-Exploitation Trade-off

- Exploration: unknown (untested) RBS design space with potentially high label
- Exploitation: querying areas that are predicted to give relatively high labels.

Which genome should we grow?

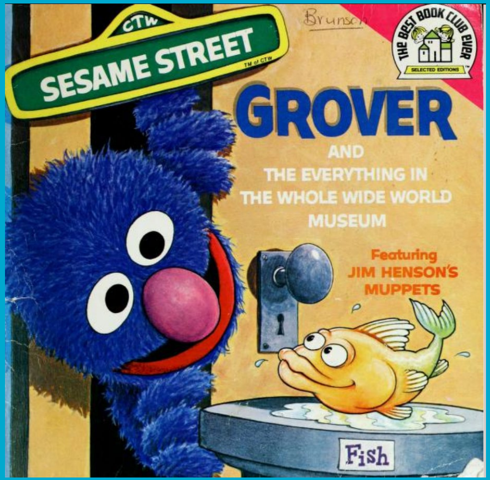




How can a statistician help?

What is deep learning?

- Current progress is driven by benchmarks
- Categories are slippery
- Define your tasks carefully



AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender
Department of Linguistics
University of Washington

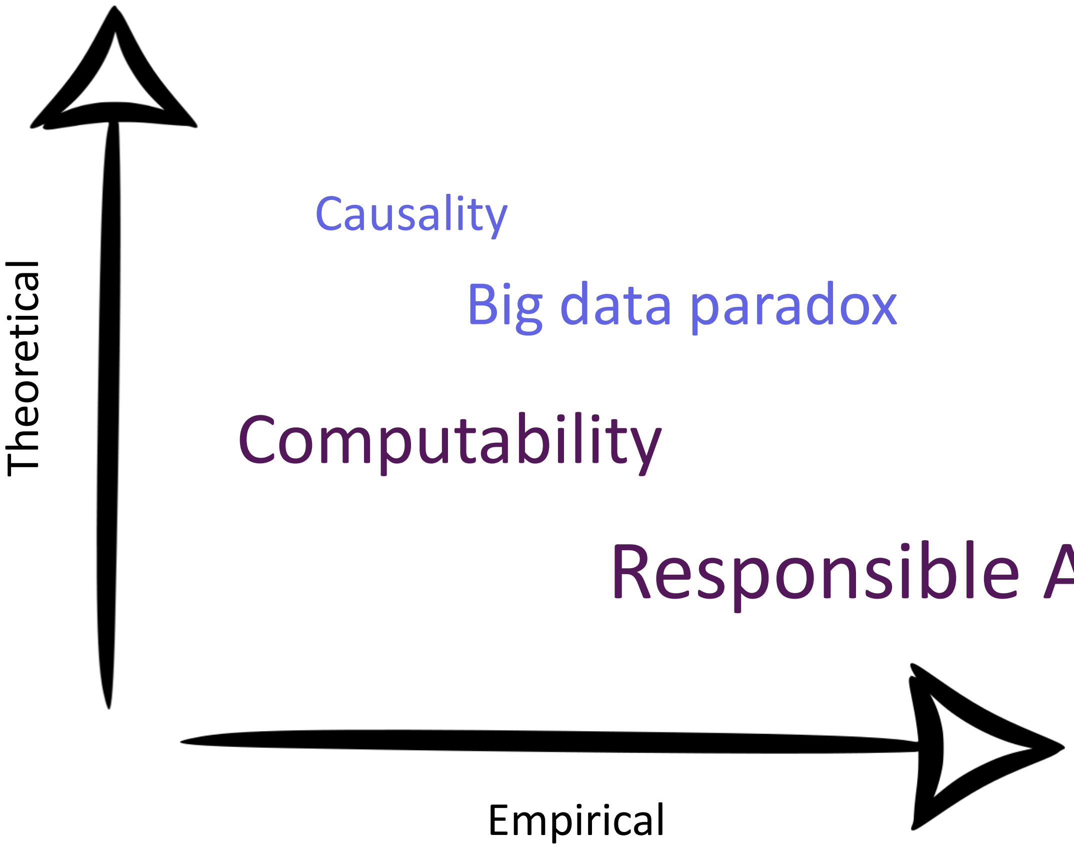
Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

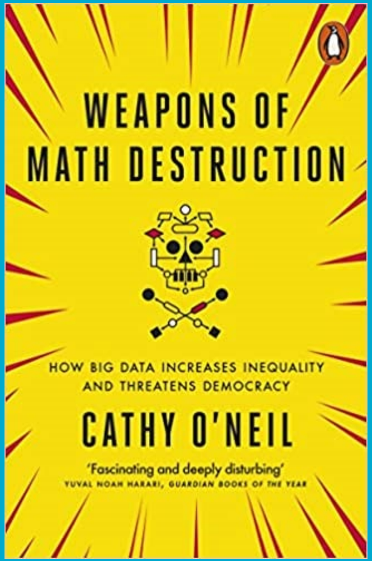
NEURAL INFORMATION PROCESSING SYSTEMS

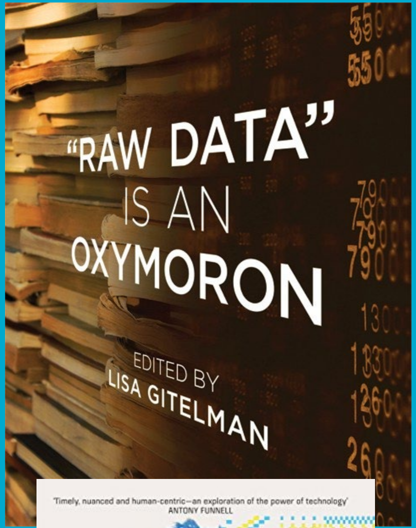
Opportunities and Challenges

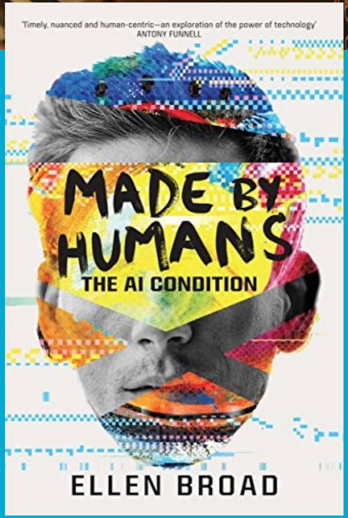


Responsible AI

- Predict based on observations
- Observations may not be suitable for the task
- Task may be poorly specified

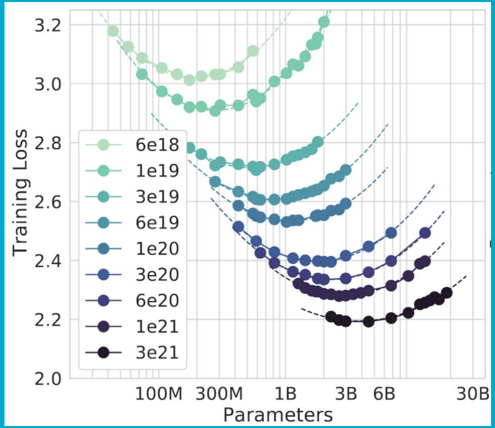






Computability

- Scaling laws still growing
- Some exponential time problems can be solved efficiently
- Compositions and backpropagation



In 2022:
(6 Apr) Dall-E 2
(23 May) Imagen
(22 Aug) Stable Diffusion
(29 Sep) Make-A-Video
(6 Oct) Imagen-video

There and Back Again

A Tale of Slopes and Expectations

Cheng Soon Ong
Data61, CSIRO
chengsoon.ong@anu.edu.au
@ChengSoonOng

Marc Peter Deisenroth
University College London
m.deisenroth@ucl.ac.uk
@mpd37

NeurIPS Tutorial, December 2020

Machine learning for combinatorial optimization: A methodological tour d'horizon

Yoshua Bengio[c,b], Andrea Lodi[a,b,*], Antoine Prouvost[a,b]

European Journal of Operational Research

ARTIFICIAL INTELLIGENCE IN KNOT THEORY

AUTHOR : C. Adams :

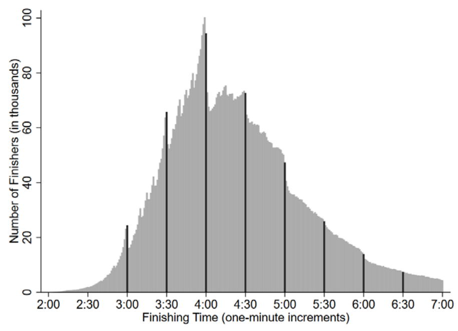
EDITOR/ART : R. Ghrist :

AMR THE ASSOCIATION FOR MATHEMATICAL RESEARCH

Big data paradox

- Observation affects data
- Law of large populations
  - US elections 2016
  - Sample size : n=2.3m → 400

- Adaptive experiments result in non-independent data



Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election

Xiao-Li Meng

THE ANNALS of APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

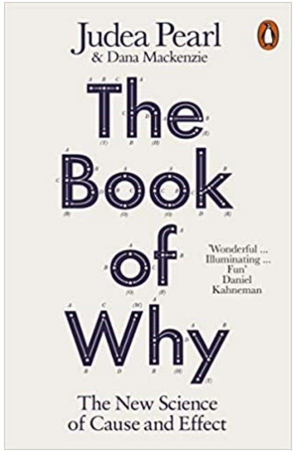
Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry

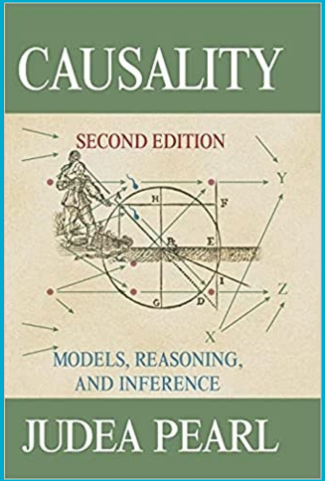
DECEMBER 2019

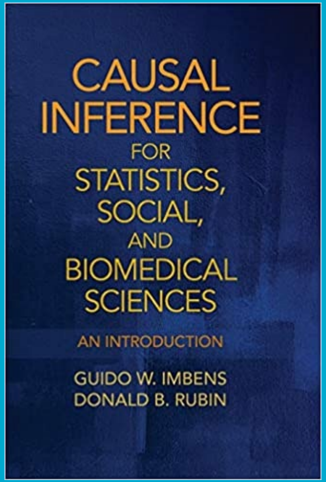
FDA U.S. FOOD & DRUG ADMINISTRATION

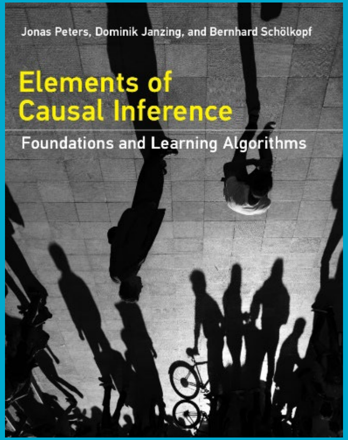
Causality

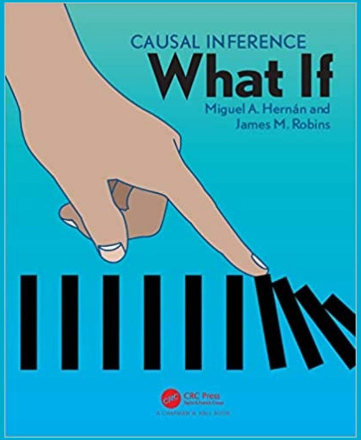
- Reinvent the language of statistical inference







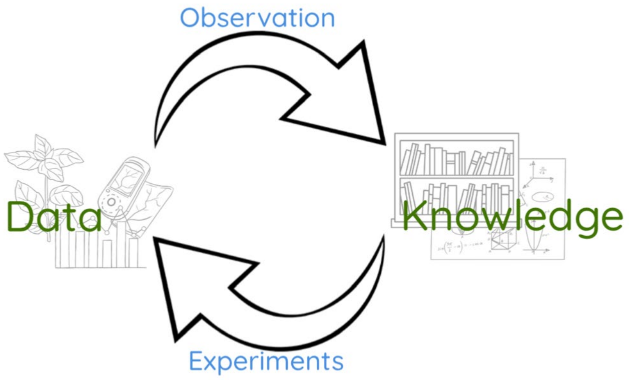


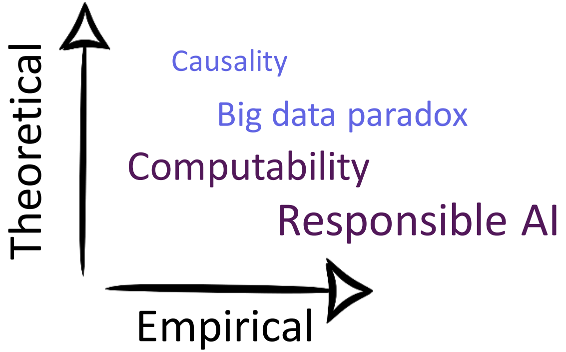


What is a scientific instrument?

How to use prediction to help perform scientific discovery?

- Scientific discovery has two phases
  - Observation
  - Experimentation
- Non-tabular data
  - Deep learning to find good embeddings
  - Model complex labels
  - Include domain knowledge
- Exploration-exploitation tradeoff
  - Use knowledge to measure better data





cheng-soon.ong@data61.csiro.au
https://research.csiro.au/mlai-fsp/