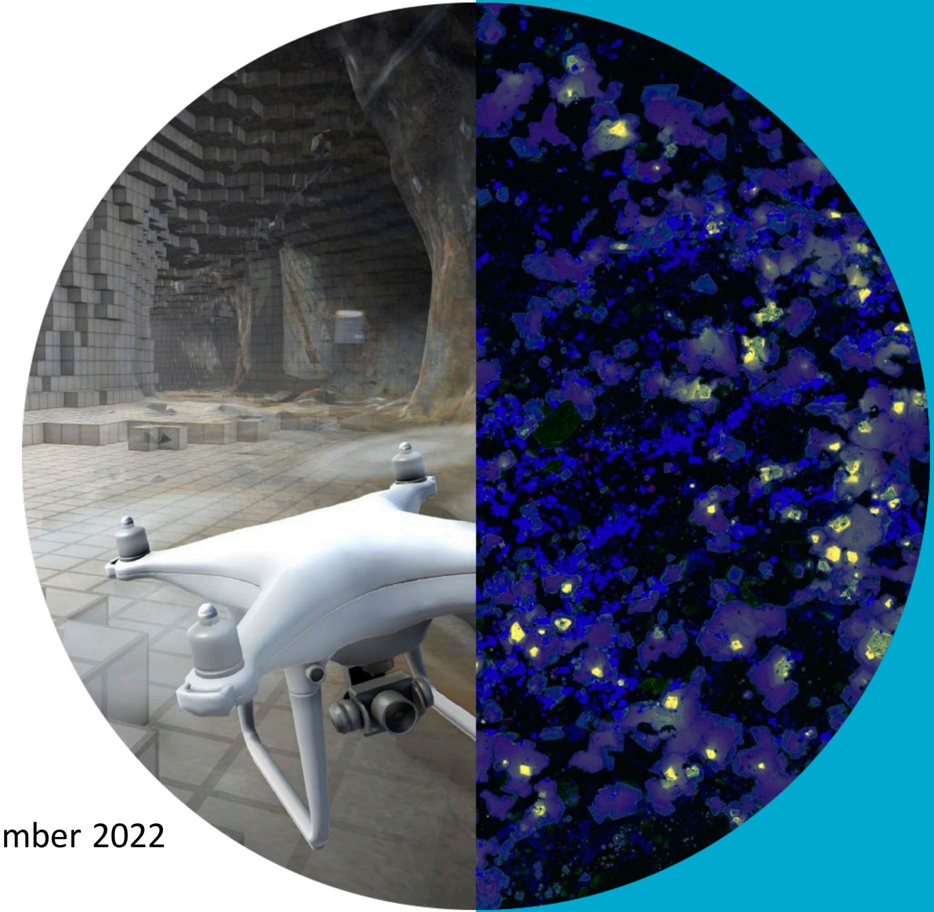
How AI is changing discovery



Ong Cheng Soon | 21 September 2022
Roche AG – Kuala Lumpur

Australia's National Science Agency

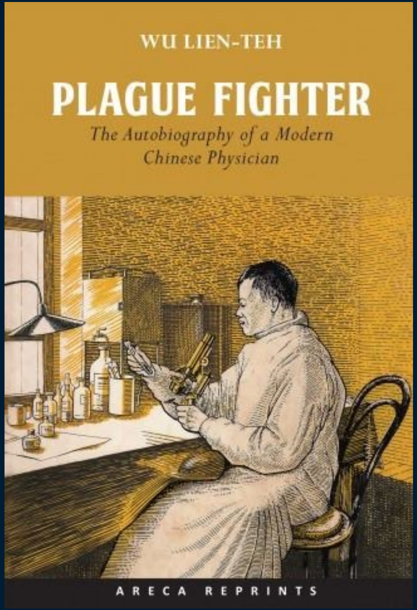
Do you know this famous person?

10 March 2021 google doodle

Goh Lean Tuck (Wu Lien The, 伍連德)

- Invented face mask
- Prevented a plague
- In 1910!

10 March 2021 google doodle



- Father of modern medicine
- Malaysians are innovative!

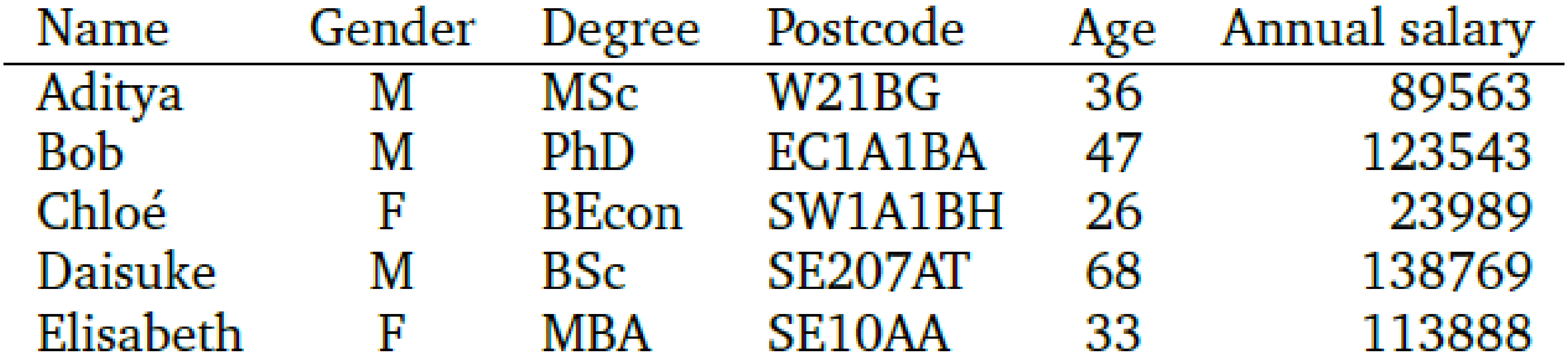
http://wulientehsociety.org

How AI is changing discovery

- How to deal with text and image data?
  - Case study in medical imaging
- Where does data come from?
  - Case study in genome biology

- What is data?

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

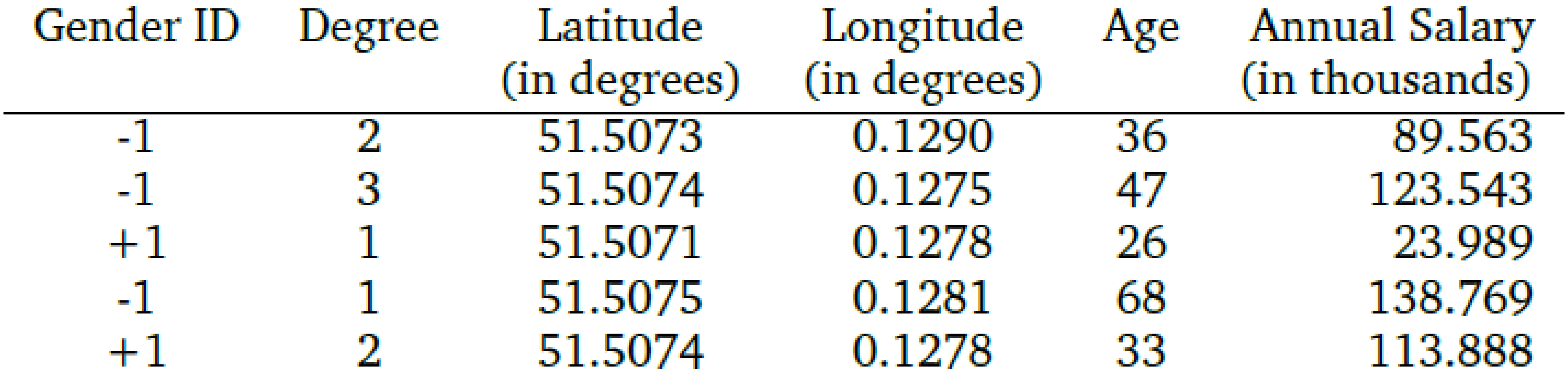
Gender ID	Degree	Latitude (in degrees)	Longitude (in degrees)	Age	Annual Salary (in thousands)
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888

binary

ordered category

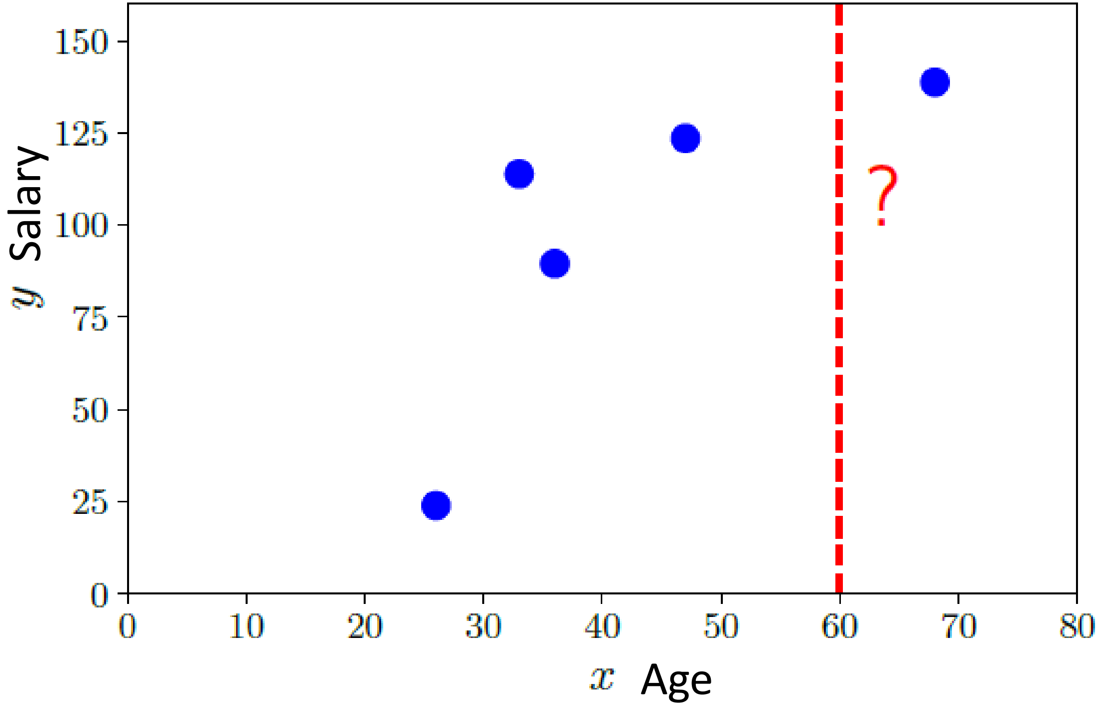
postcode

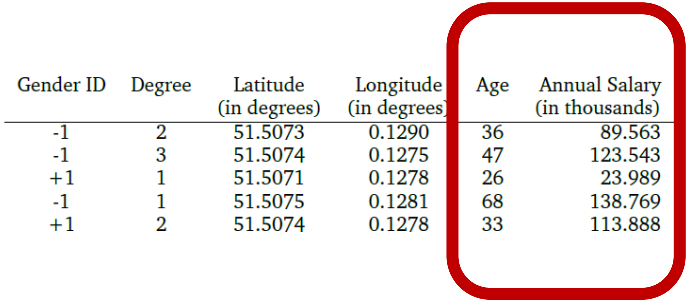
Data in numerical format



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

Predict salary given age





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

What is Machine Learning?

Machine Learning is about prediction

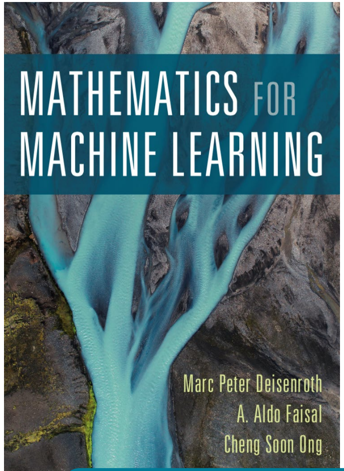
- Machine Learning is about prediction
  - Examples/covariates/features
  - Labels/annotations/target variable

$$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim \mathcal{X}$$
$$\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \sim \mathcal{Y}$$

Predictor

$$\boxed{f_{\boldsymbol{w}}(\boldsymbol{x}) : \mathcal{X} \to \mathcal{Y}}$$

- Estimate the best predictor = training
  - No mechanistic model of the phenomenon
  - There are many examples
  - The outcomes (labels) are well defined (usually binary)

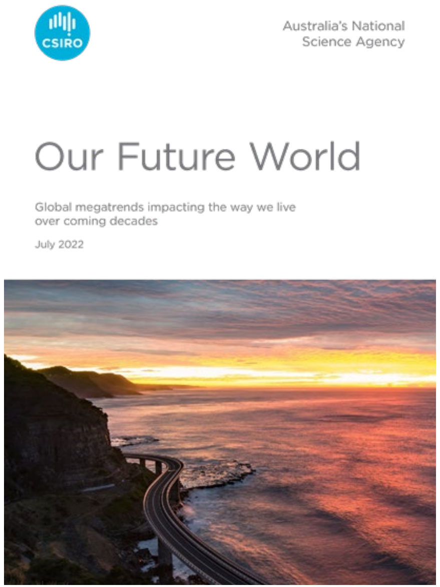


MATHEMATICS FOR MACHINE LEARNING

Marc Peter Deisenroth
A. Aldo Faisal
Cheng Soon Ong

mml-book.com

Global megatrends in data and AI

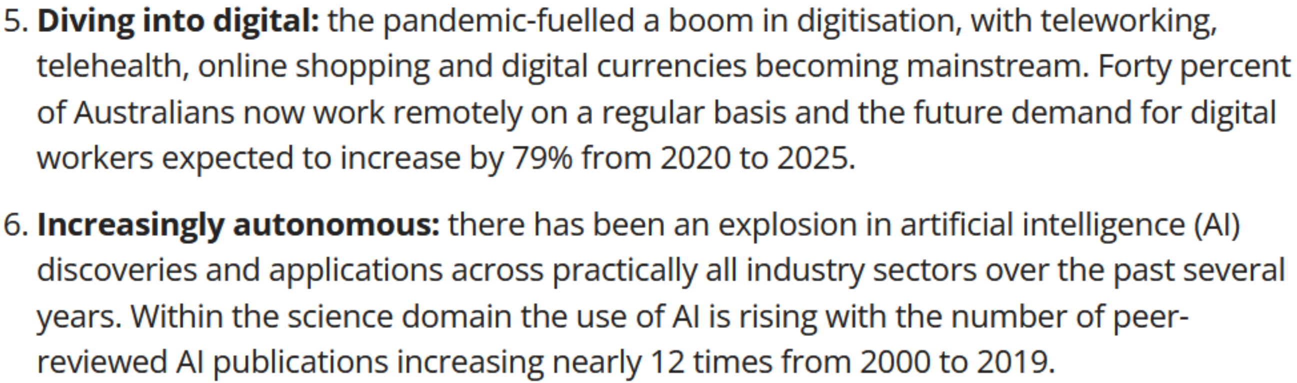


Australia's National Science Agency

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

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

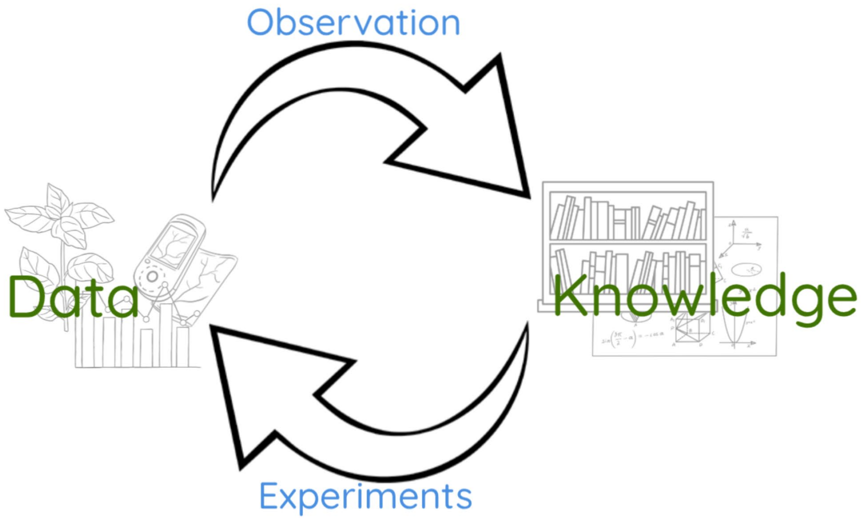
MLAI Future Science Platform

How to use prediction to help perform scientific discovery?

30 postdoc researchers

10 senior scientists

1 vision



Machine learning for scientific discovery

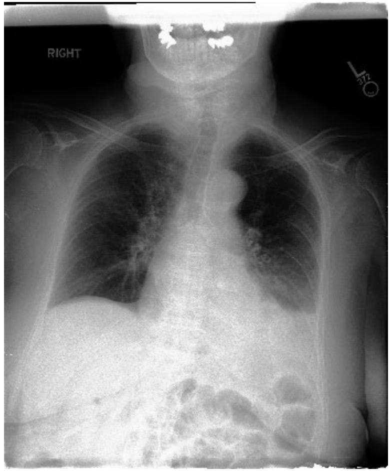
How to deal with text and image data?

Why is Medical Image Analysis Important?

Radiologists' diagnostic accuracy drops by 4% after 8 hours

Text and image data: chest x-rays

- Medical diagnosis relies on expert interpretation of images and text



Radiologist's report

As compared to the previous radiograph, the known left-sided effusion is unchanged. The effusion is restricted to the left lung base and to the left sinus. There is subsequent atelectasis at the left lung base. The well inflated lung parenchyma shows no evidence of pneumonia. However, presence of pneumonia in the atelectatic lung regions cannot be excluded. Borderline size of the cardiac silhouette. No pulmonary edema. At the right lower aspect of the trachea, a calcified lymph node might be present.

How to convert to numerical data?

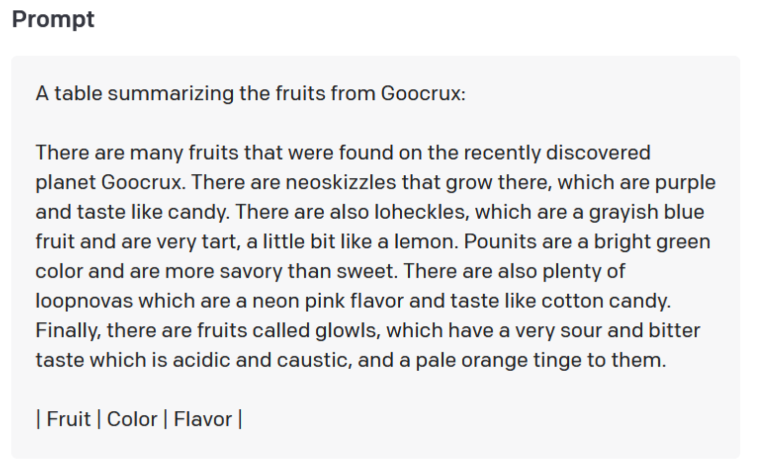
Popular science
Large Language Models

- Deep learning for representing text
- Natural language processing tasks
- Text generation
- Safety and equality

https://openai.com/blog/gpt-3-apps/

https://hai.stanford.edu/news/how-large-language-models-will-transform-science-society-and-ai

https://faculty.washington.edu/ebender/stochasticparrots.html

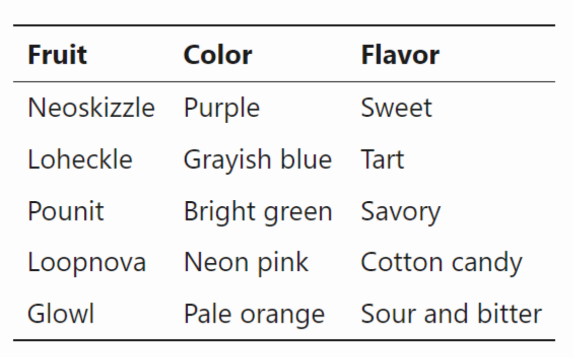


Prompt

A table summarizing the fruits from Goocrux:

There are many fruits that were found on the recently discovered planet Goocrux. There are neoskizzles that grow there, which are purple and taste like candy. There are also loheckles, which are a grayish blue fruit and are very tart, a little bit like a lemon. Pounits are a bright green color and are more savory than sweet. There are also plenty of loopnovas which are a neon pink flavor and taste like cotton candy. Finally, there are fruits called glowls, which have a very sour and bitter taste which is acidic and caustic, and a pale orange tinge to them.

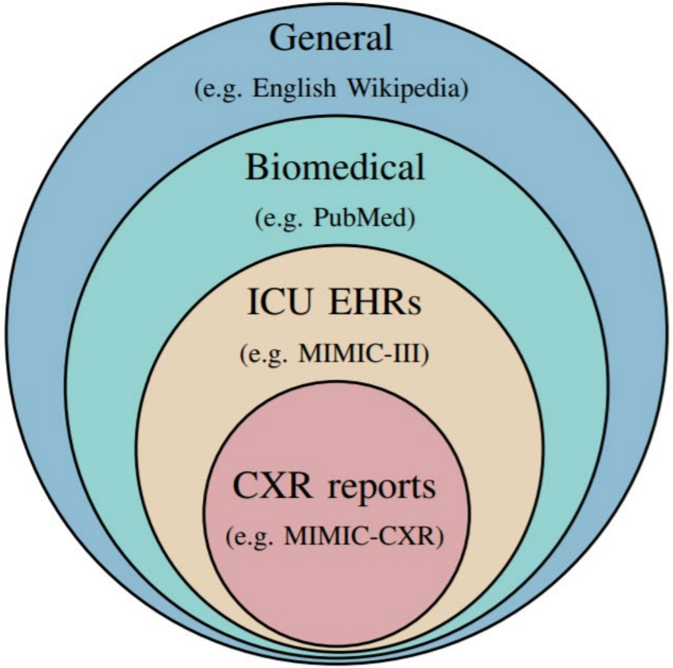
| Fruit | Color | Flavor |



Fruit	Color	Flavor
Neoskizzle	Purple	Sweet
Loheckle	Grayish blue	Tart
Pounit	Bright green	Savory
Loopnova	Neon pink	Cotton candy
Glowl	Pale orange	Sour and bitter

Medical reports have specific language

- Domain specific words
- Acronyms
- Errors and typos



Popular science

Image generation

- Deep learning for representing images
- Object detection
- Artistic generation
- Robustness and attribution

https://www.craiyon.com/

https://stability.ai/blog/stable-diffusion-public-release

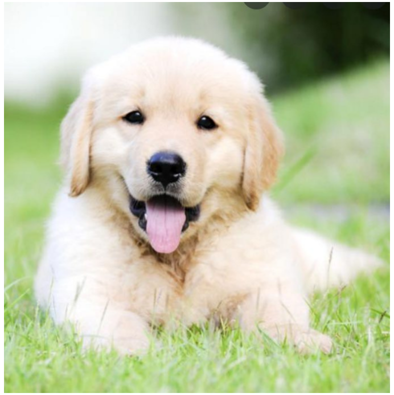
https://openai.com/dall-e-2/

FIRST PLACE
FINE ARTS
COLORADO STATE FAIR
PUEBLO, CO
Jason Allen
Pueblo West
Théâtre D'opéra Spatial
$750
Colorado State Fair

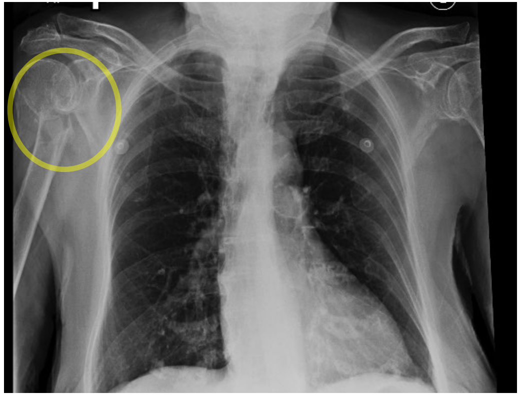
Medical images have specific properties

- Object of interest not in the middle
- Detect deviation from normal

A (cute) puppy



A humeral fracture CXR



ImageCLEF 2021 and 2022 medical imaging competitions

2021:

- Medical Image Captioning: ***3rd place***

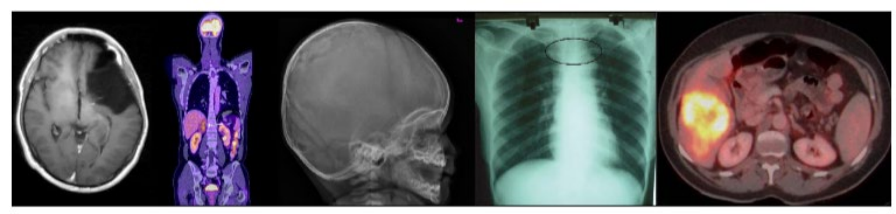
2022:

- Medical Image Captioning: ***3rd place***
- Medical Image Concept Detection: ***3rd place***
- Tuberculosis Caverns Detection: ***1st place***

CSIRO at ImageCLEF medical Caption 2022
http://ceur-ws.org/Vol-3180/paper-109.pdf

Aaron Nicolson

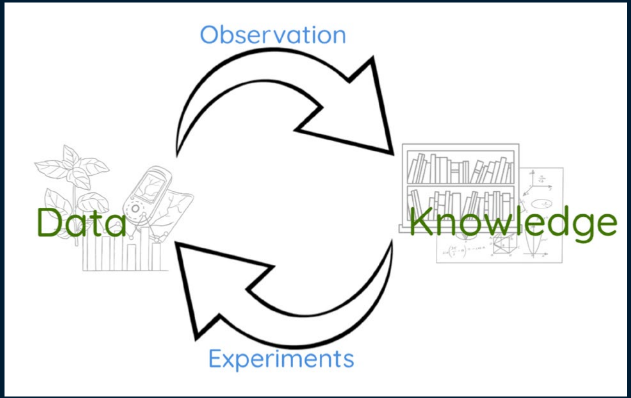
Leo Lebrat



https://www.imageclef.org/

Data lifecycle

1. Can I load your data using `pandas` or `numpy`?
2. Confounders, missing values, scale, units, encoding
3. Define the problem you want to answer:
   - The business/scientific problem
   - The performance metric
   - The model for the predictor
4. Run `sklearn` or `statsmodels` (**machine learning part**)
   Do not train on the test set.
5. Convert predictions into human friendly form for decision making

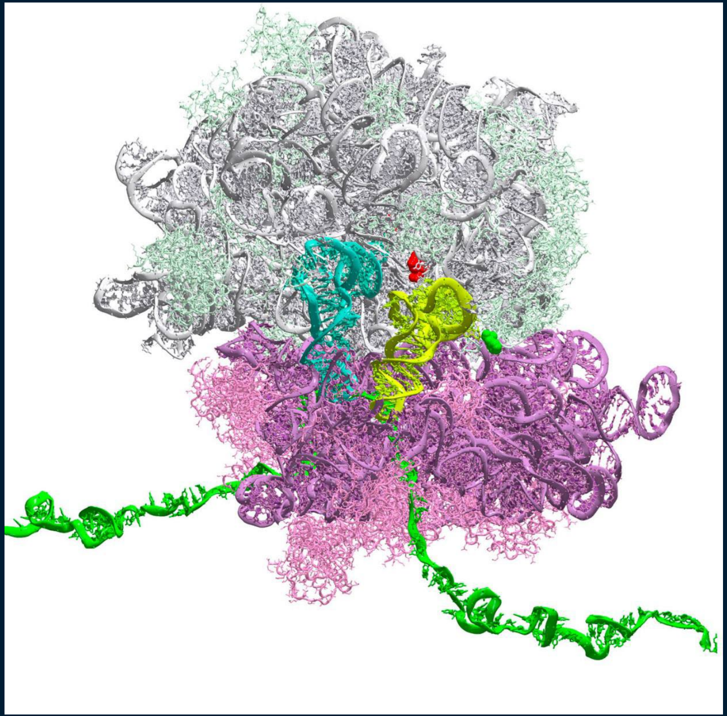


Where does data come from?

Adaptive design

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

Which genome should we grow?



What’s the objective?

Design Ribosome Binding Site (RBS) sequences → Optimize the protein expression level.

RBS sequence	**Normalized[*] Protein Expression Level**
TTTAAGAGTTATATATACAT	1.58
TTTAAGAATATGCTATACAT	1.42
TTTAAGACTCGGATATACAT	0.14
TTTAAGAGTTTTTTATACAT	2.88

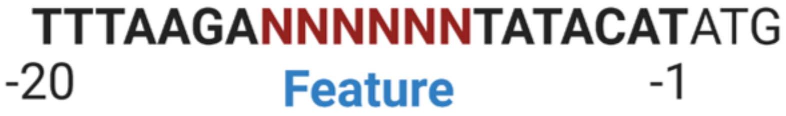
Core part (design space): 4^6 = 4096 possibilities in total

[*] zero mean and unit variance normalization $z = \frac{x-\mu}{\sigma}$

Practical challenge

How do we perform lots of TIR measurements (within budget)?

- 6 positions, 4 bases per position = 4096
- Likely that many RBS sequences give low translation rates



- **Want to only try a small fraction**
- Sequences very similar to consensus likely to perform well, but worse than consensus
- Interesting to consider drastically different core RBS sequence

High throughput experiments needed

MLAI augmented SynBio

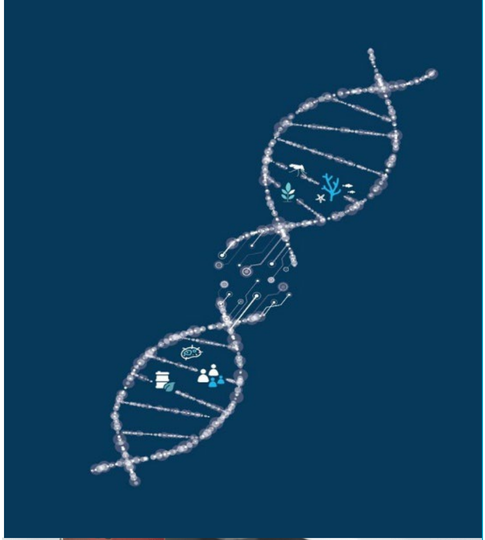
- **Working definition of ‘synthetic biology’:** *The design and construction of DNA-encoded parts, devices, machines, and organisms; and their application for useful purposes.*
- Experimental science domains
  - Integrative Biological Modelling
  - Engineering Novel Biological Components
  - Assembling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

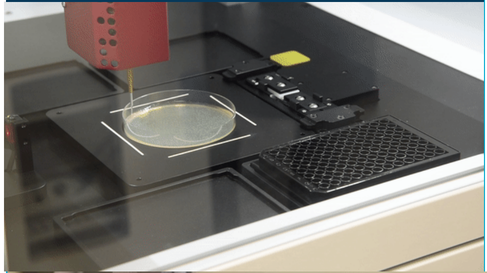
https://research.csiro.au/synthetic-biology-fsp

Claudia Vickers

Janet Reid

Alison Rice

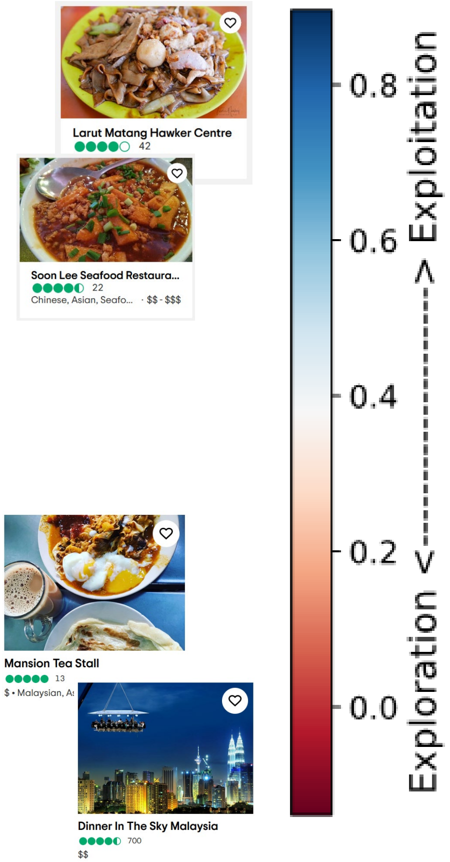




Still too many options to try!

- 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

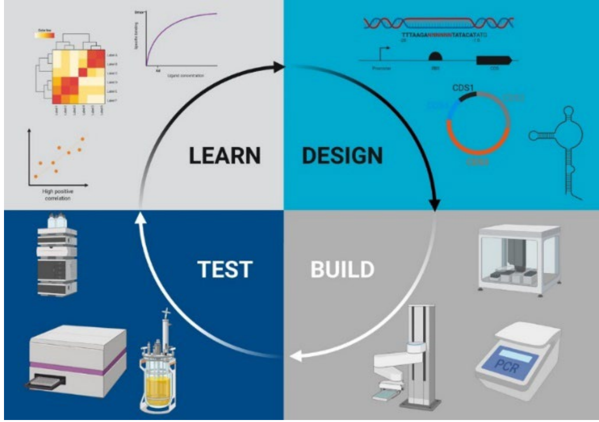

Larut Matang Hawker Centre
●●●●○ 42
Soon Lee Seafood Restaura...
22
Chinese, Asian, Seafo... · $$ - $$$
Mansion Tea Stall
●●●●● 13
$ • Malaysian, As
Dinner In The Sky Malaysia
700
$$
0.8
0.6
0.4
0.2
0.0
Exploration <-------------> Exploitation

Algorithms

Mengyan Zhang, ANU

Maciej Holowko, L&W

Huw Hayman Zumpe, SynBio



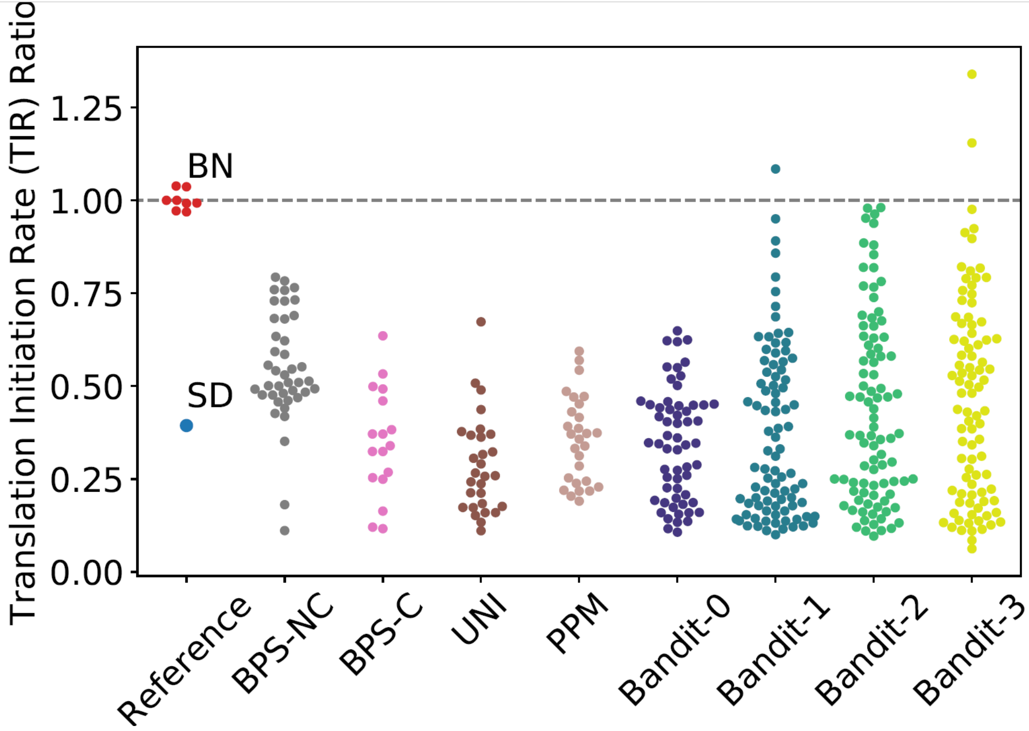
1. A (Bayesian) regression algorithm which predicts both
   - Mean
   - Uncertainty

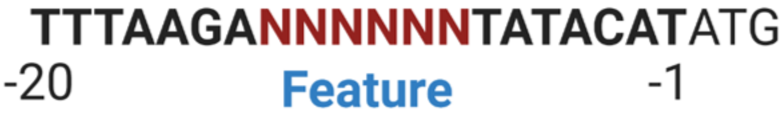
→ Gaussian Process Regression (aka Kriging) → LEARN

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

→ Multiarmed Bandits Algorithms: Upper Confidence Bound → DESIGN

AI recommends good designs





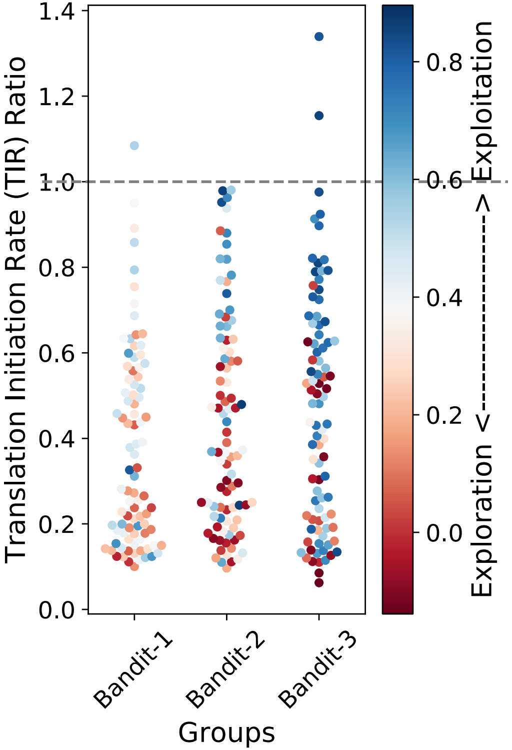
- Hard to search by evolving sequences
- 4 experimental cycles
- 35% stronger than engineered sequence

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?



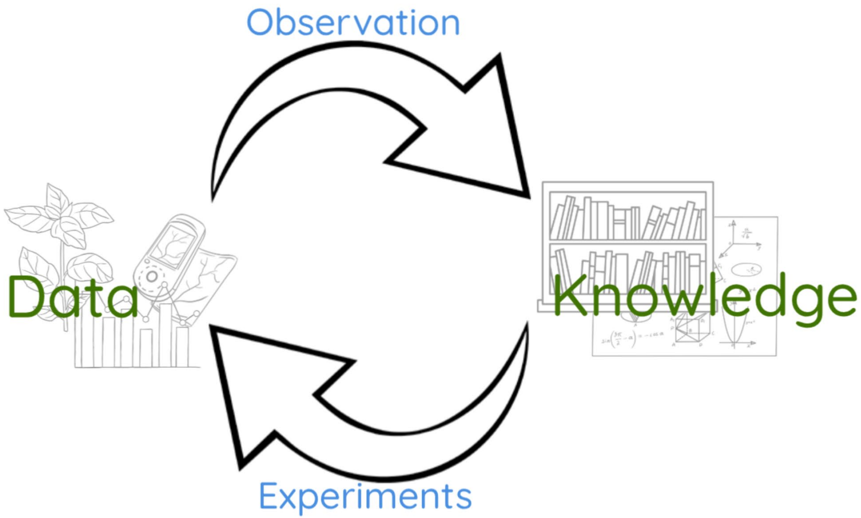
AI for Scientific Discovery

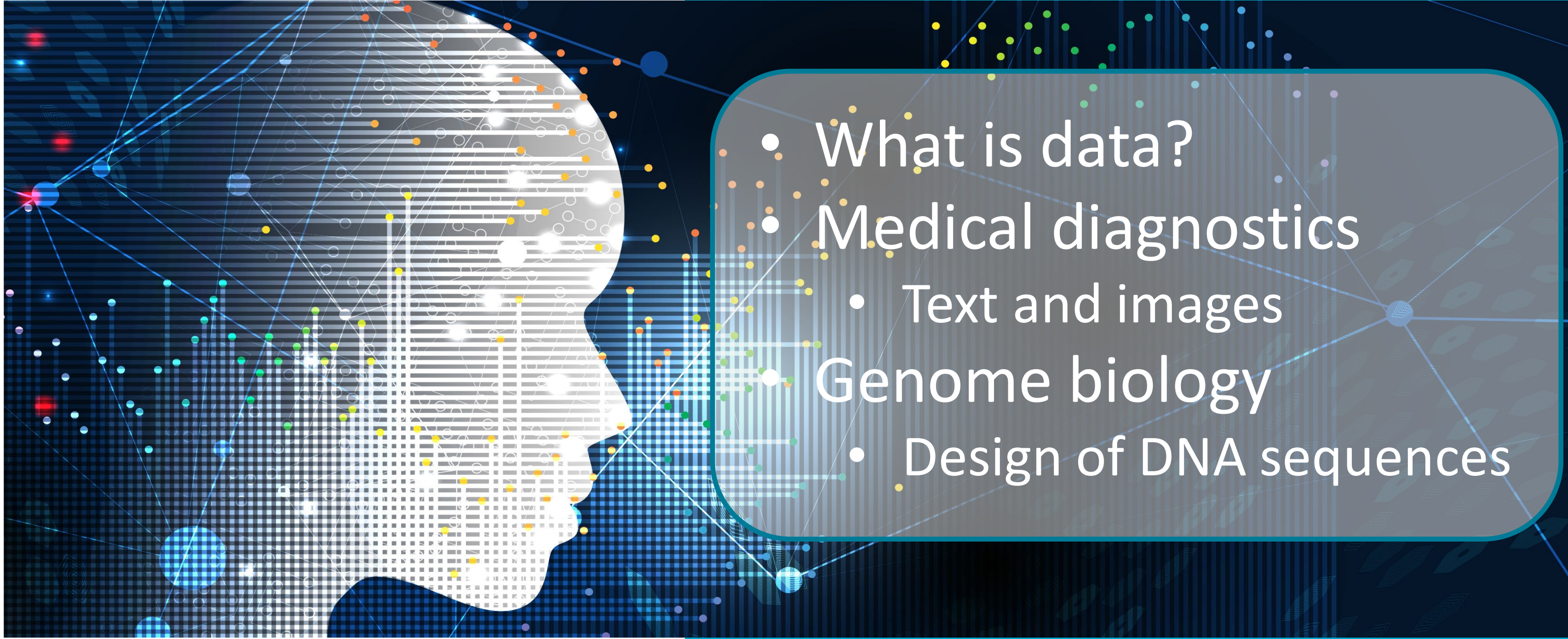
How to use prediction to help perform scientific discovery?

- Scientific discovery has two phases
  - Observation
  - Experimentation

- Observation:
  Converts data to knowledge

- Experiments:
  Use knowledge to measure better data





- What is data?
- Medical diagnostics
    - Text and images
- Genome biology
    - Design of DNA sequences

How AI is changing discovery

September 2022

Ong Cheng Soon
cheng-soon.ong@data61.csiro.au

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

Machine Learning & Artificial Intelligence

Australia's National Science Agency