

“Doing statistics is like doing crosswords
except that one cannot know for sure
whether one has found the solution.”
– *John Tukey*

The Annals of Statistics
2002, Vol. 30, No. 6, 1535–1575

**JOHN W. TUKEY: HIS LIFE AND PROFESSIONAL
CONTRIBUTIONS¹**

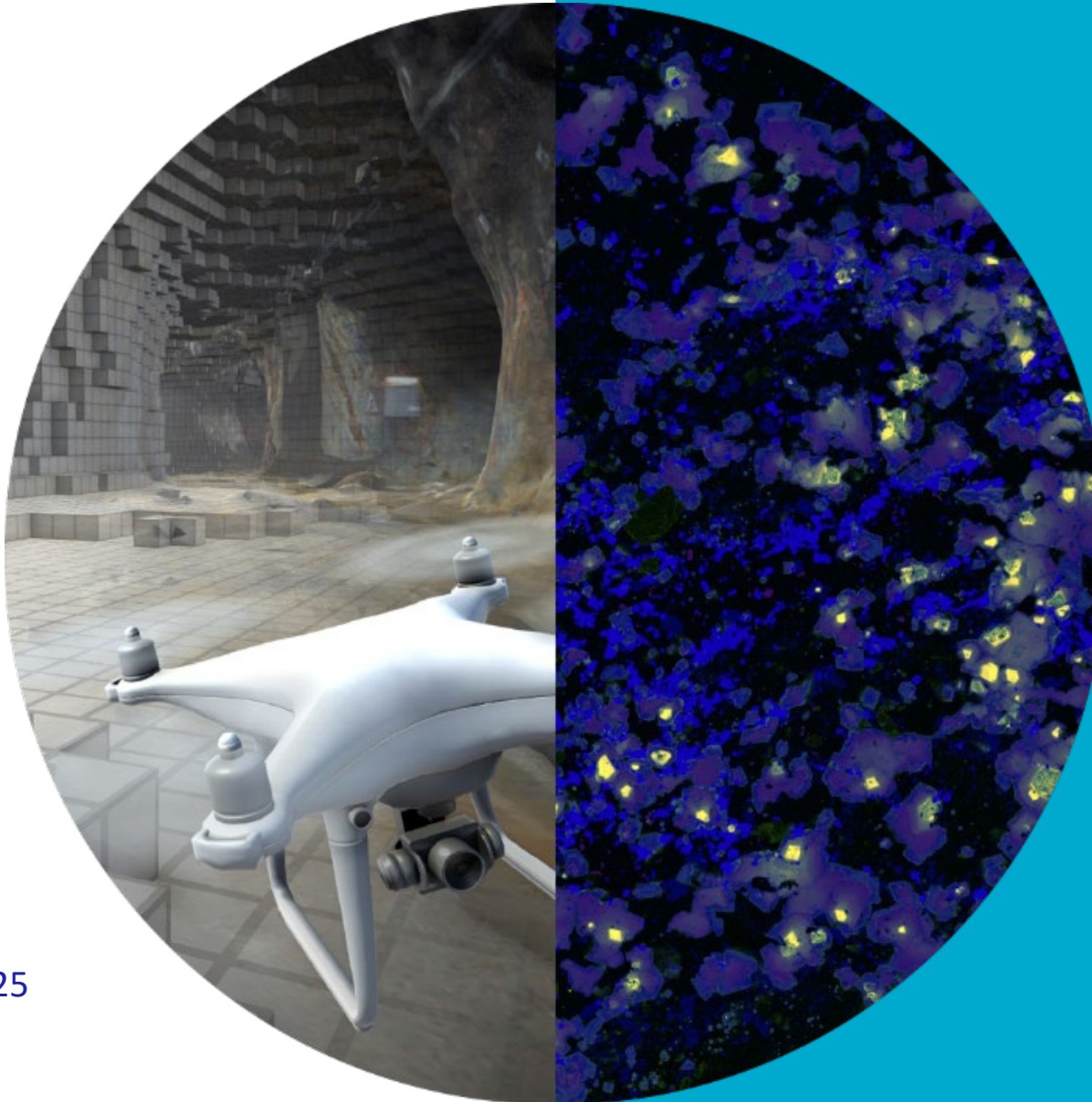
BY DAVID R. BRILLINGER



On Finding Good Experiments (in life science)

Cheng Soon Ong,
Data61 and ANU
25 November 2025

Biometrics in the Bush Capital 2025
International Biometric Society

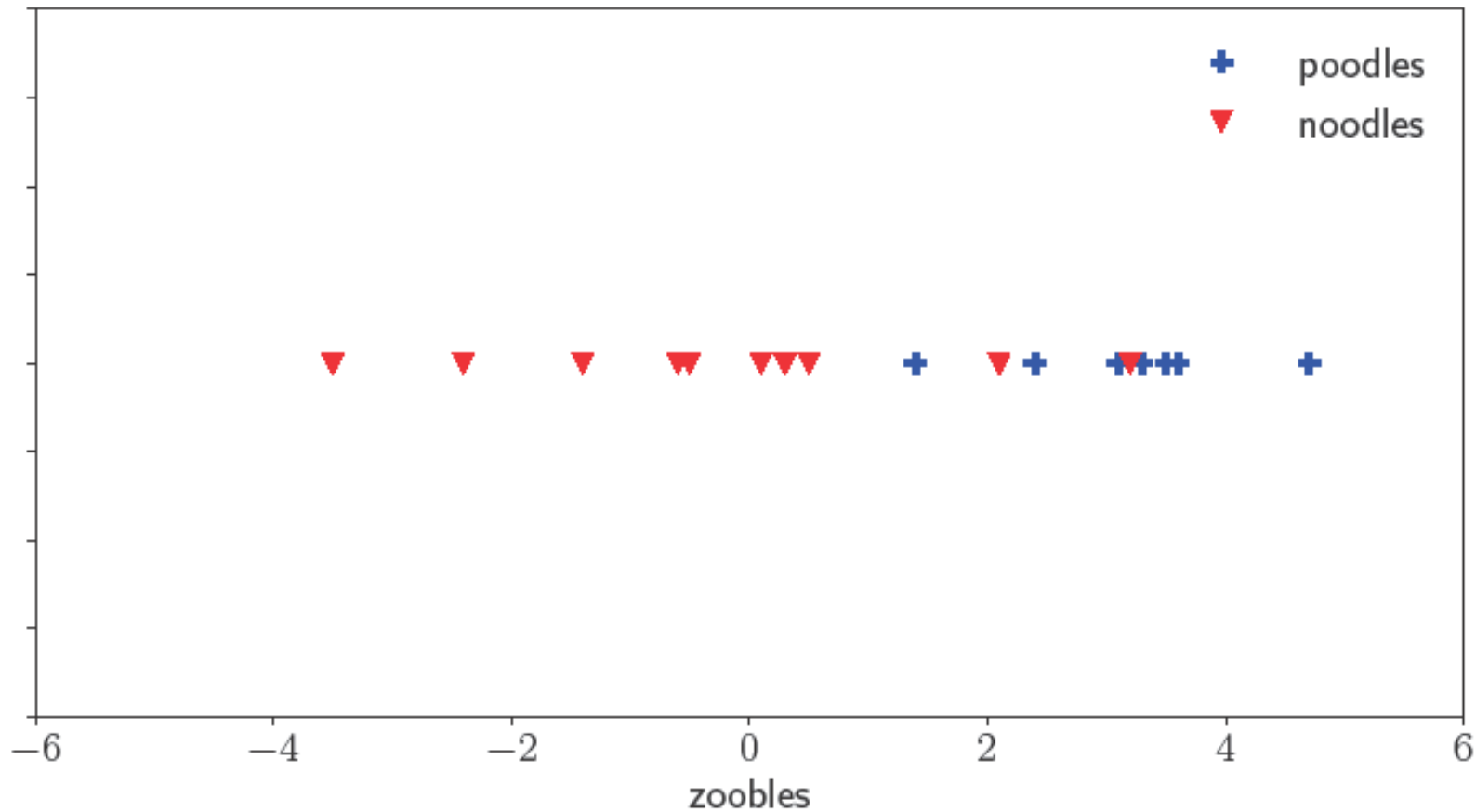


I would like to acknowledge the Ngunnawal people, the traditional custodians whose ancestral lands we're meeting on today, and pay my respect to their Elders past and present.



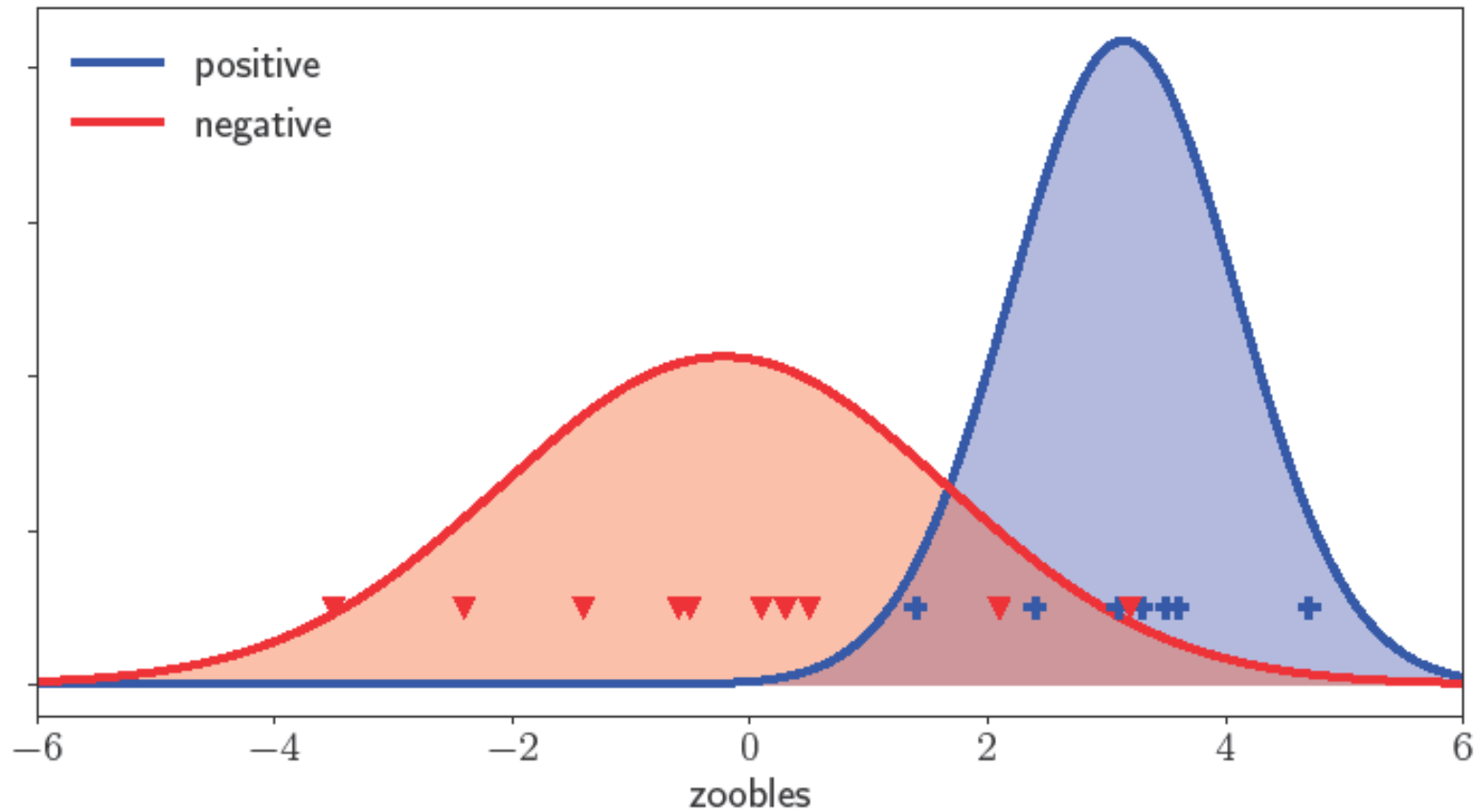
Given some data

- Classify **blue plus** vs **red triangles**, based on features



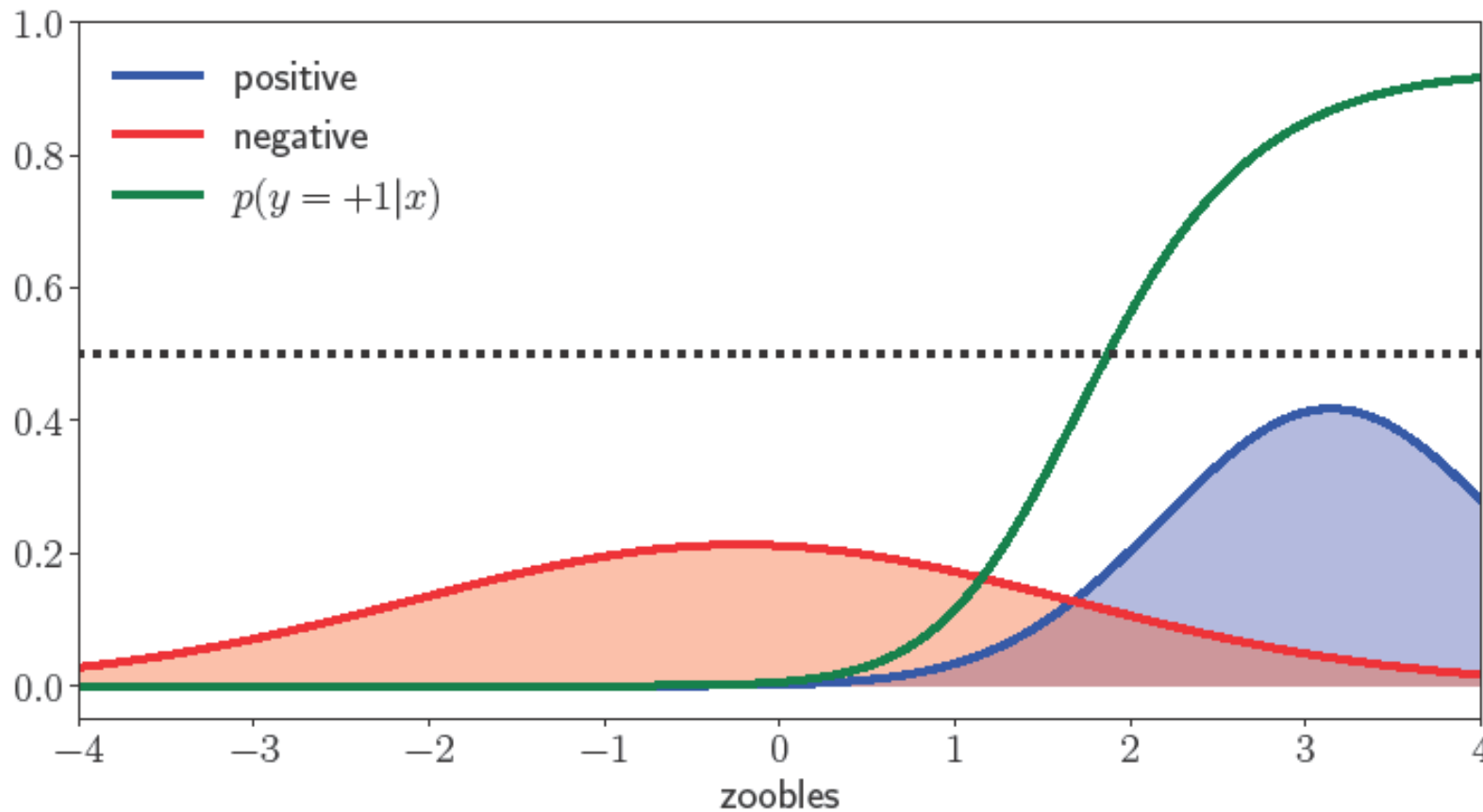
Fit a model to data

- Estimate a Gaussian for each class conditional



Build a classifier

- Compute the posterior probability of **blue plus**



What is machine learning?

- Mostly about prediction
 - Examples/covariates/features
 - Labels/annotations/target variables

$$x_1, \dots, x_n \sim \mathcal{X}$$
$$y_1, \dots, y_n \sim \mathcal{Y}$$

- Predictor

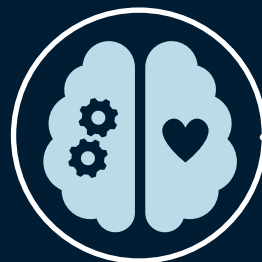
$$f_{\theta}(x) : \mathcal{X} \rightarrow \mathcal{Y}$$

- Estimate the best predictor = training = $\min_{\theta} \ell(f_{\theta})$
(given loss)

Finding good experiments



What is an **experiment**?



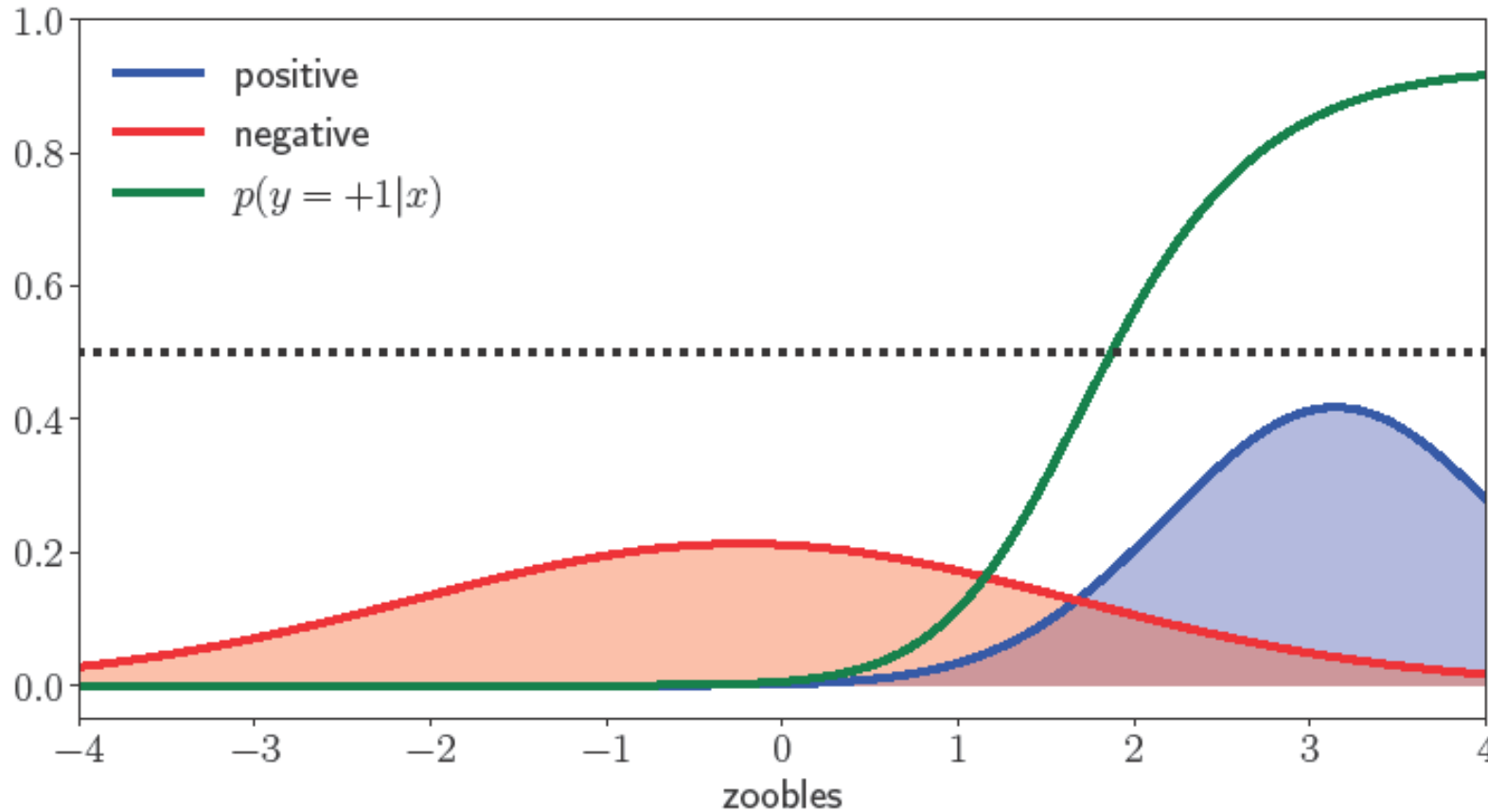
How do we **find** good experiments?



What do we mean by “**good**”?

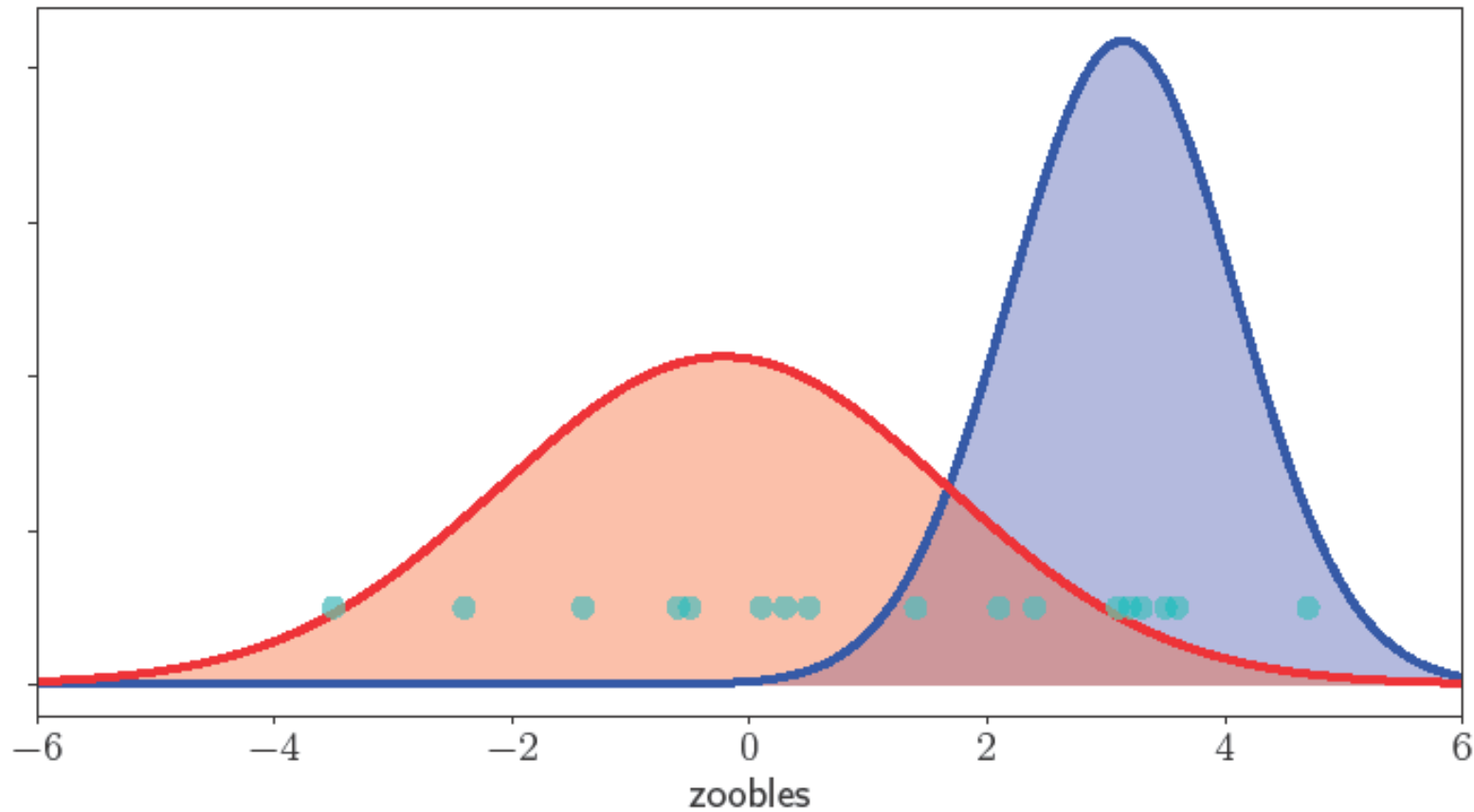
Recall what is a predictor

- Compute the posterior probability of blue plus



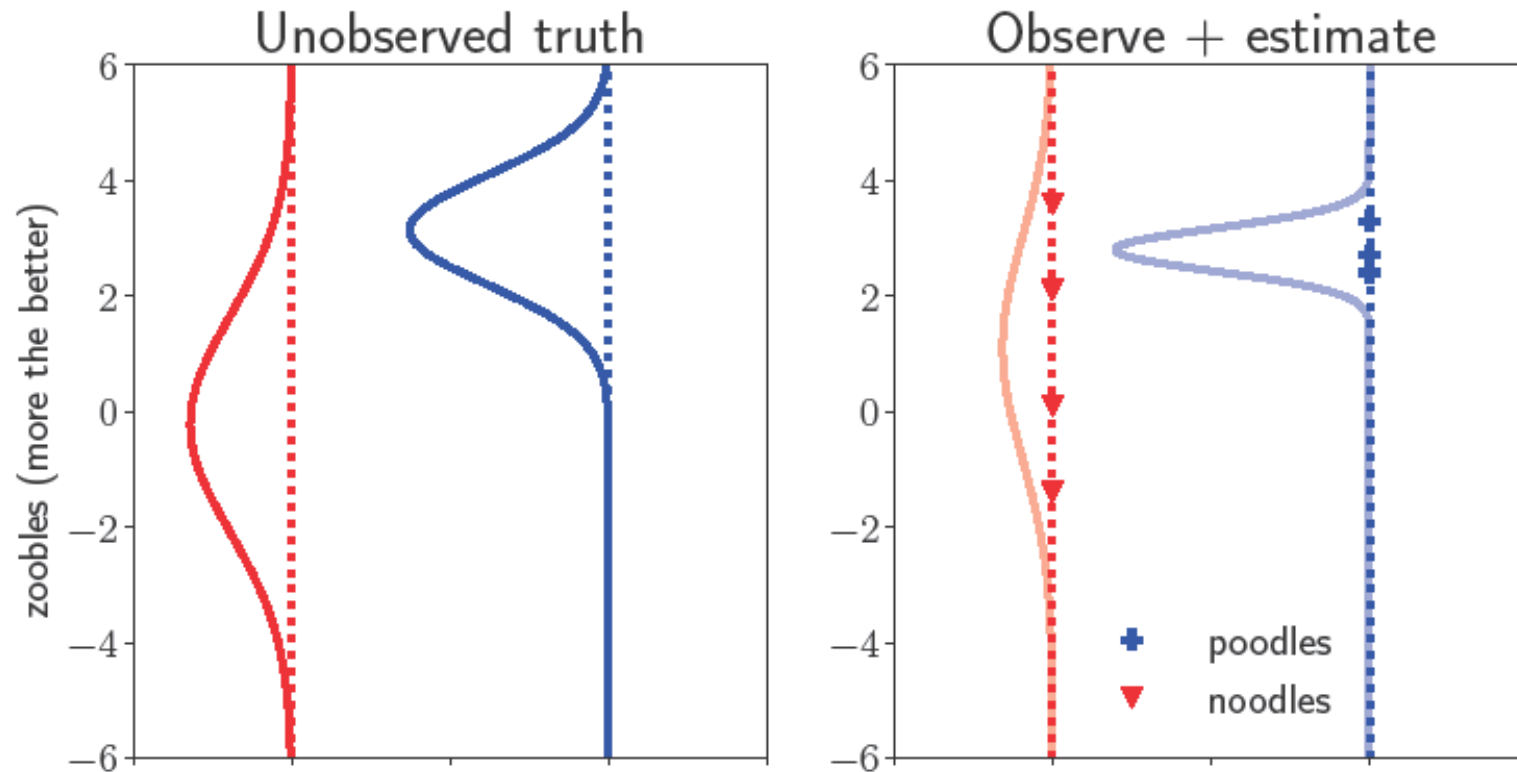
Active Learning

- Want to build a predictor without paying for a lot of labels

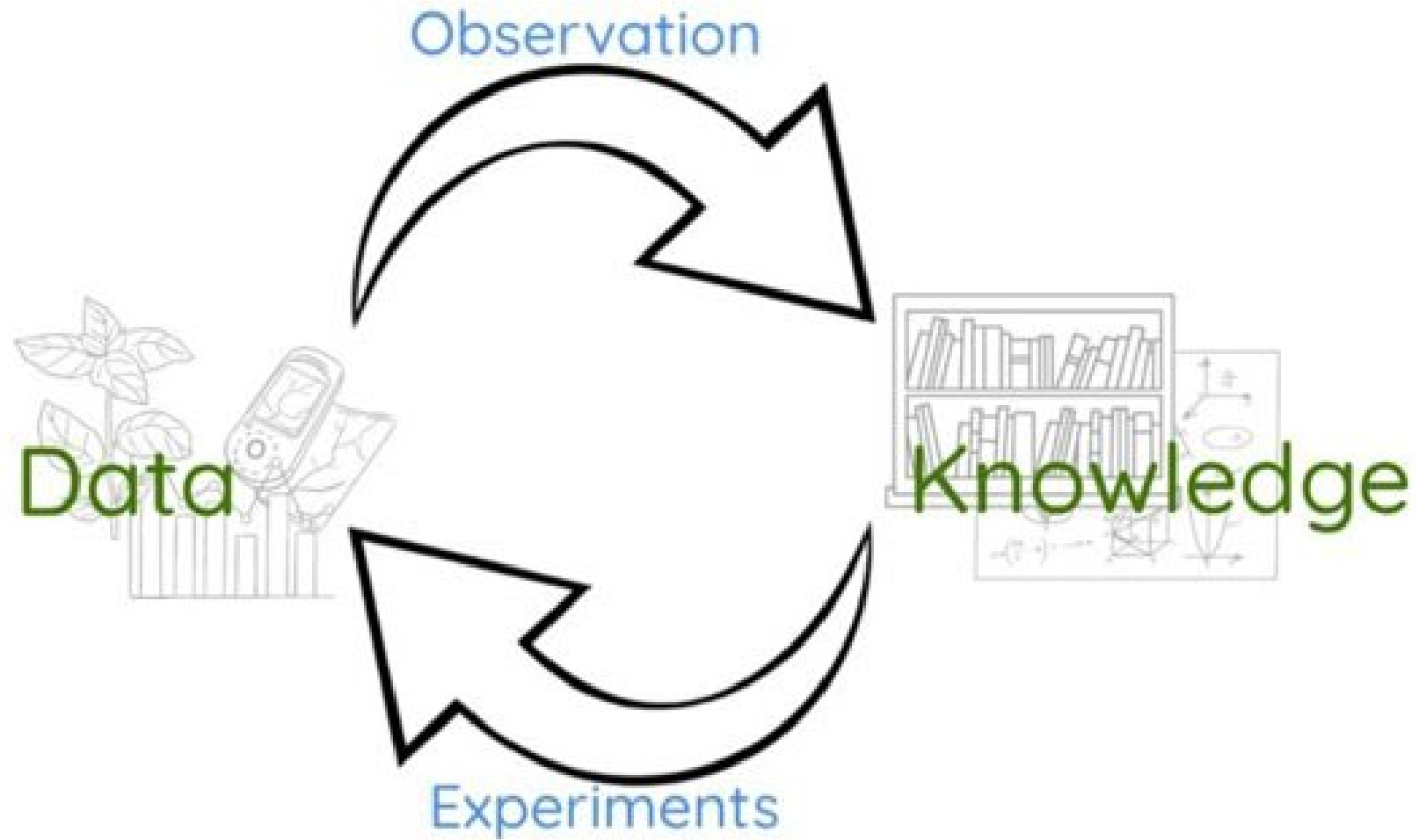


Bandits / Bayesian optimisation

- Want to maximise the outcome of different choices



The scientific method



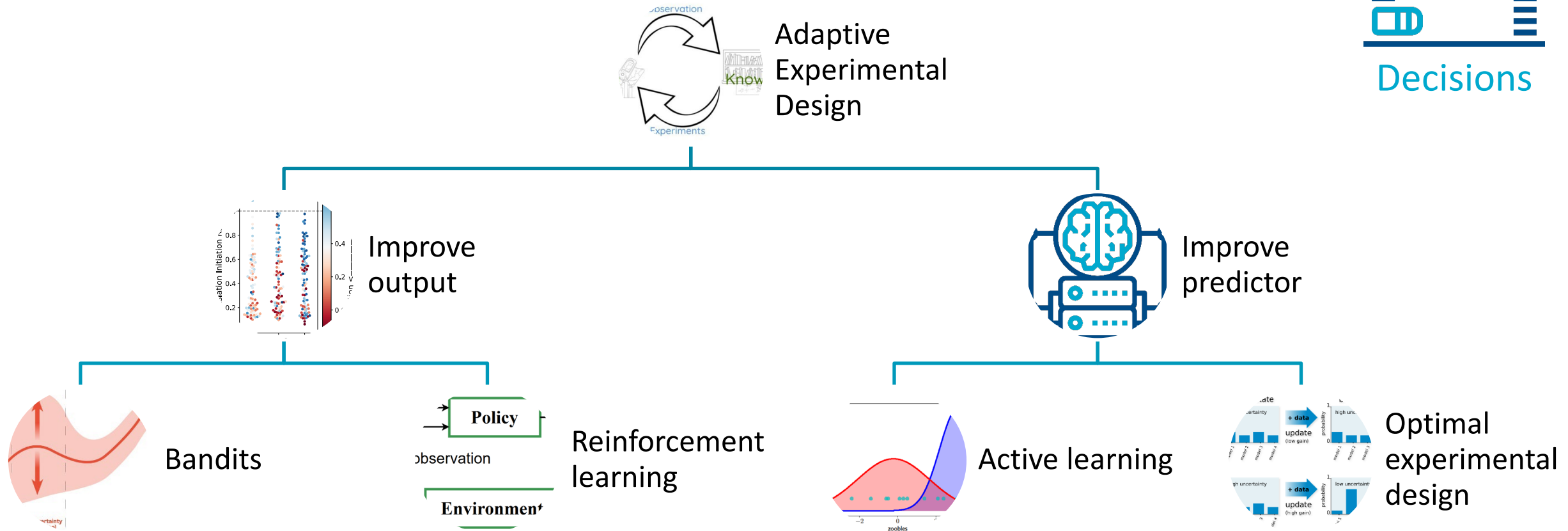
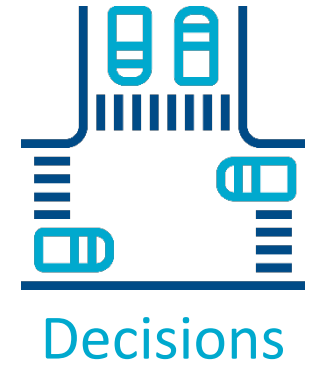
Using a predictor to adaptively choose data

- Given a predictor

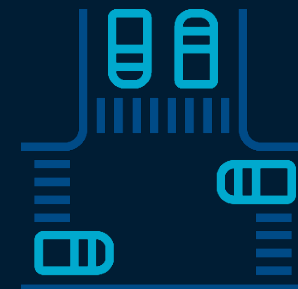
$$f_{\theta}(x) : \mathcal{X} \rightarrow \mathcal{Y}$$

- We could collect data (perform an experiment) to:
 - Bandits
Improve the output of the predictor, that is better $f(x)$
 - Active Learning
Improve the predictor/model itself, that is better θ

Decision making



- <https://researchoutreach.org/articles/adaptive-experiments-machine-learning-help-scientific-discovery/>
- Chades, Blau, Ong, Machine Learning for Biological Design, Chap 19, Synthetic Biology (2nd ed), 2024

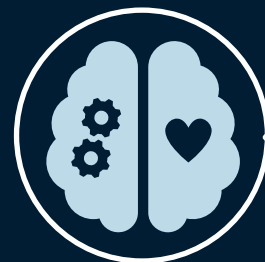


Decisions

Three messages



What is an experiment?



How do we find good experiments?

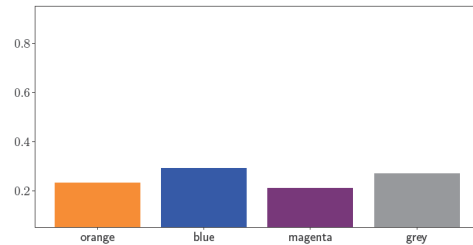


What do we mean by “good”?

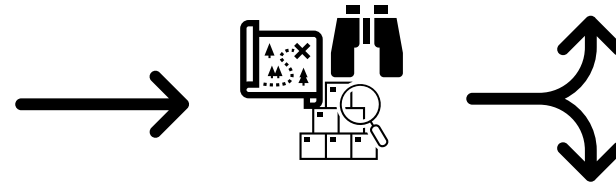
Information gain – ecological modelling

- Find experiments to improve the model's parameters

Imagine parameters have only 4 states

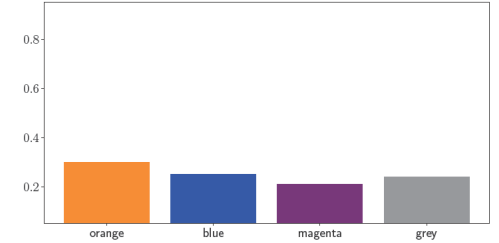


Before – prior
 $p(\theta)$

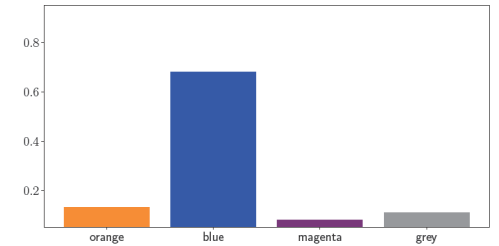


Take a sample $\mathbf{D}$

$p(\theta|\mathbf{D})$



Uninformative experiment



Informative experiment

$$\mathbb{E}_{\mathbf{D} \sim \mathcal{X}, \mathcal{Y}} \mathcal{D} \left[p(\theta|\mathbf{D}) \parallel p(\theta) \right]$$

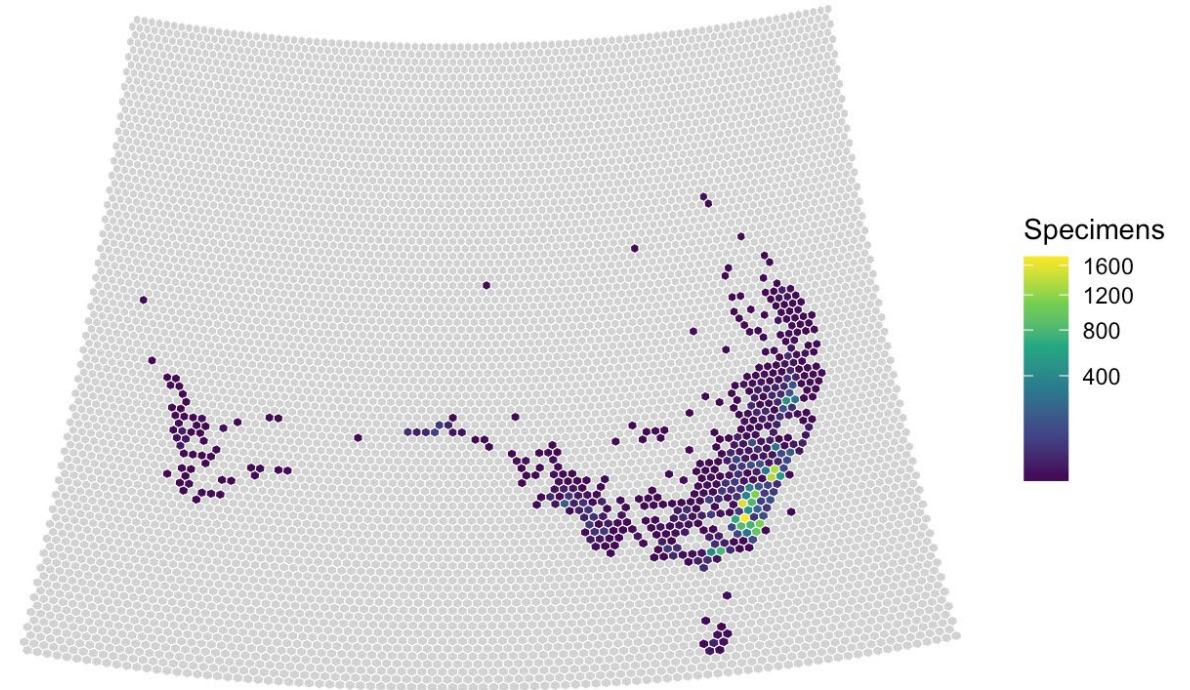
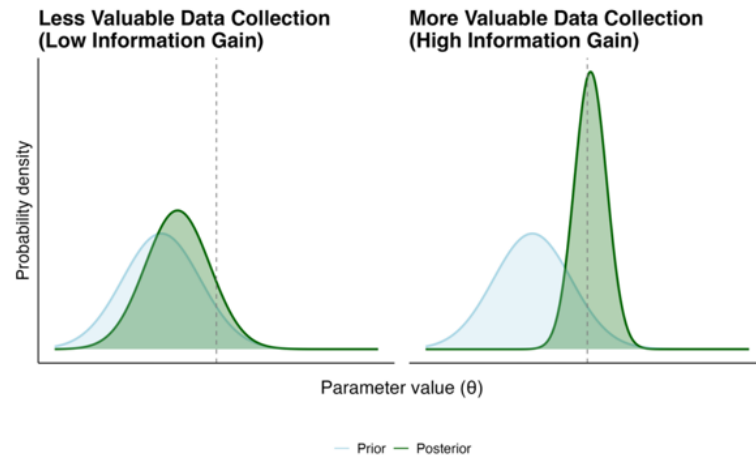
Don't know which data we're going to get

Difference between posterior and prior

Species distribution model

- African lovegrass
- Beta-Bernoulli presence model

Eragrostis curvula (African Lovegrass) Distribution - 50km Hexagons
23249 specimens in 474 hexagons



Owen Forbes



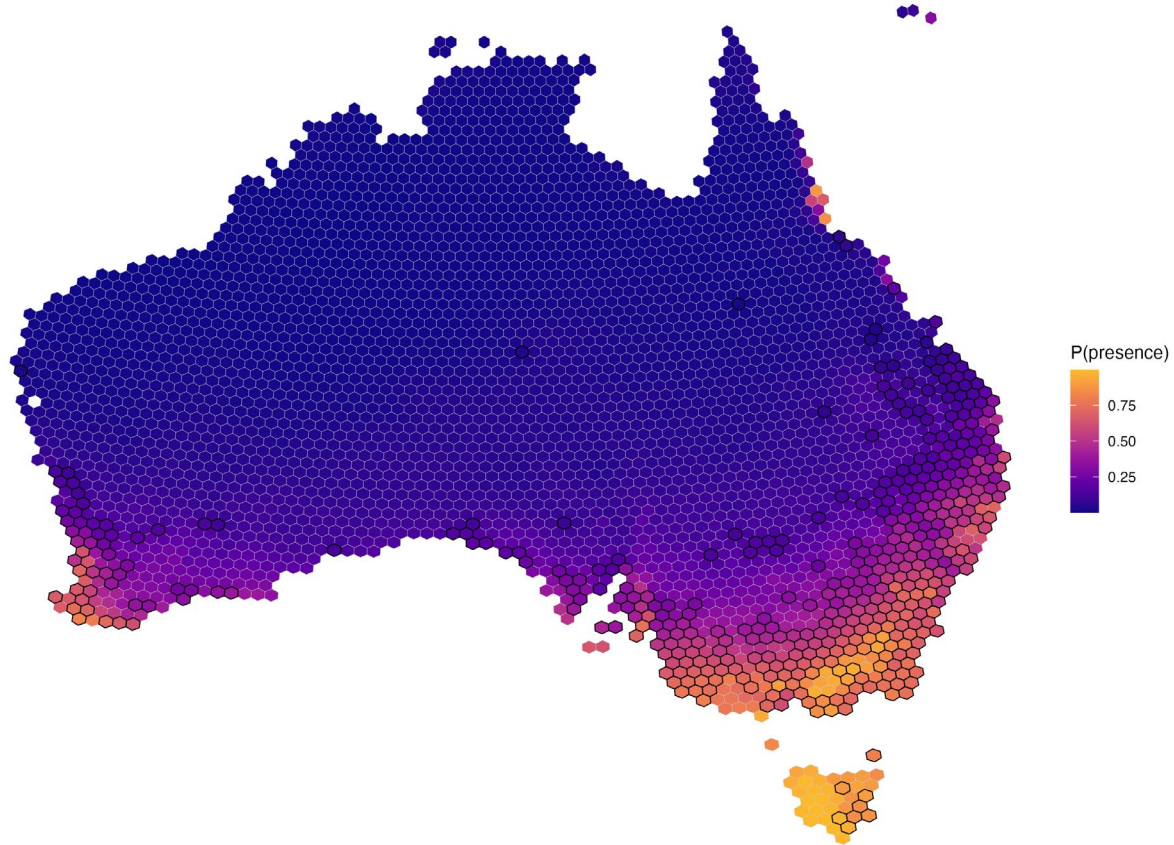
Pete Thrall



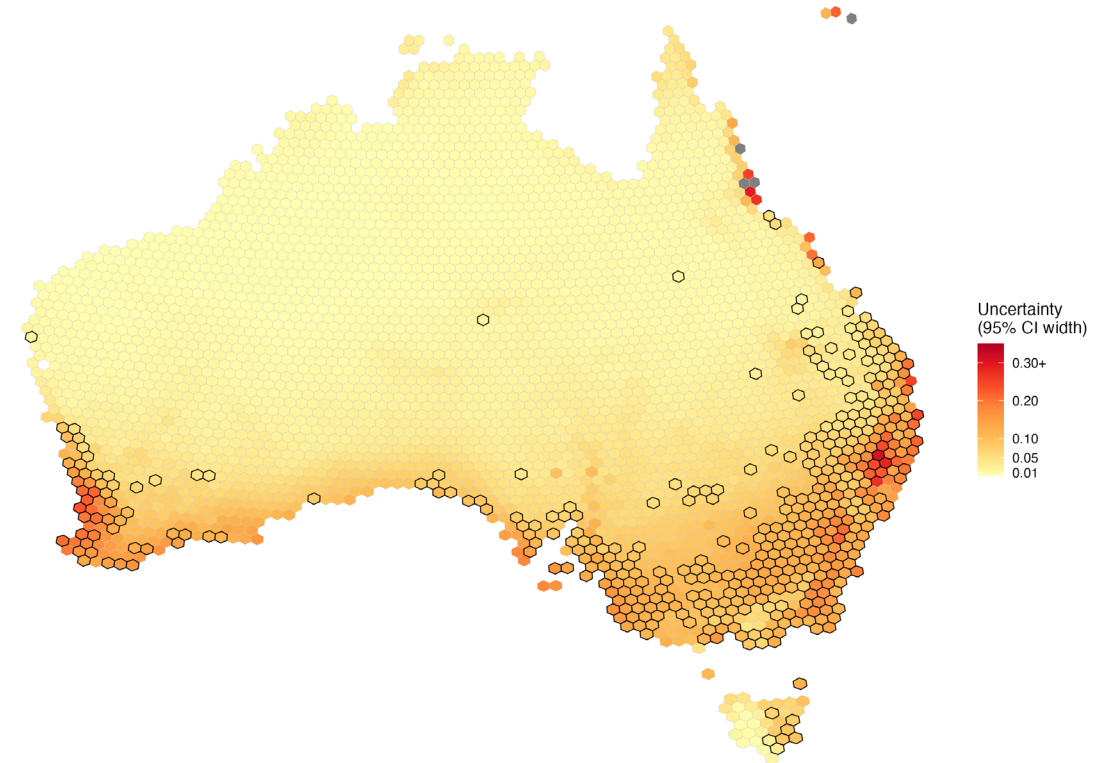
Andrew Young

Prediction and uncertainty

Occupancy Probability



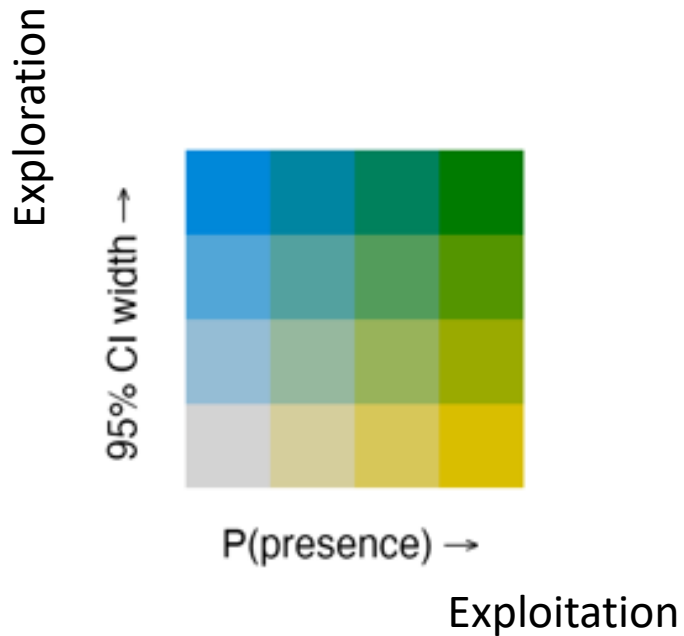
Prediction Uncertainty



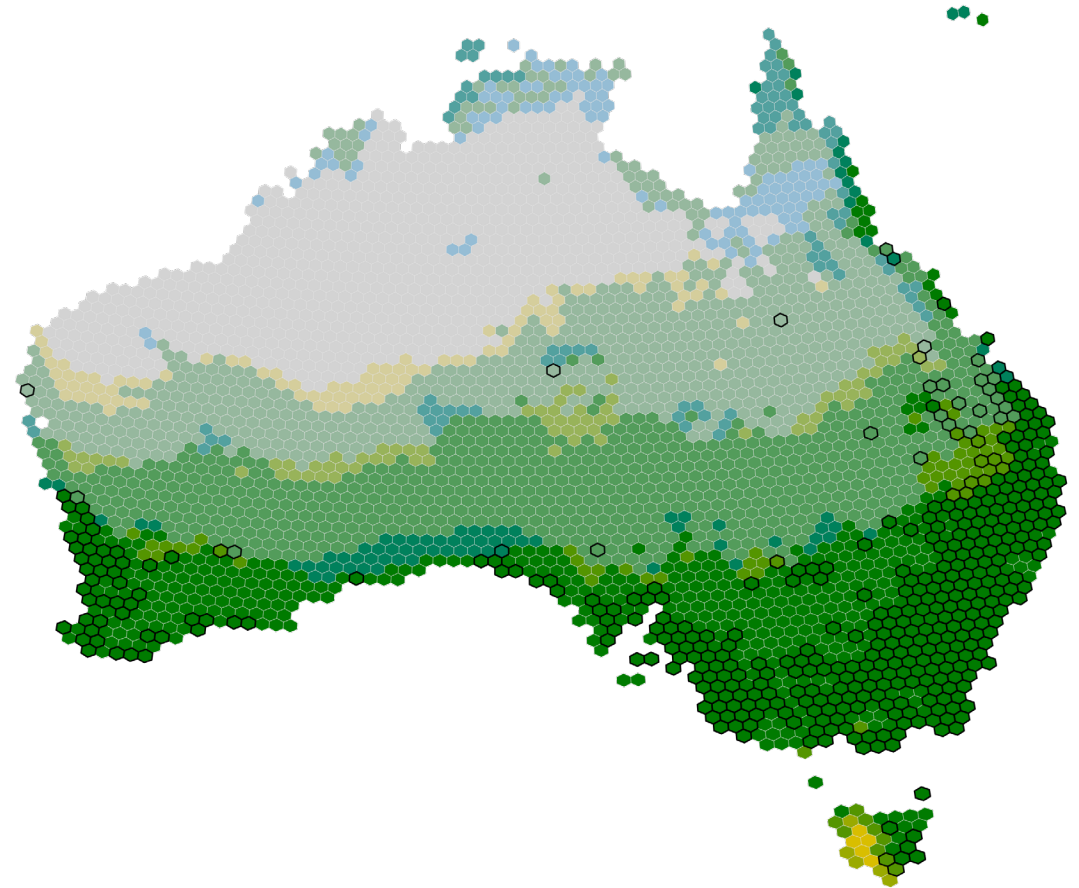
Hexagons with black borders have confirmed observations

Prediction and uncertainty

- Unify into one plot



Combined: Estimate + Uncertainty

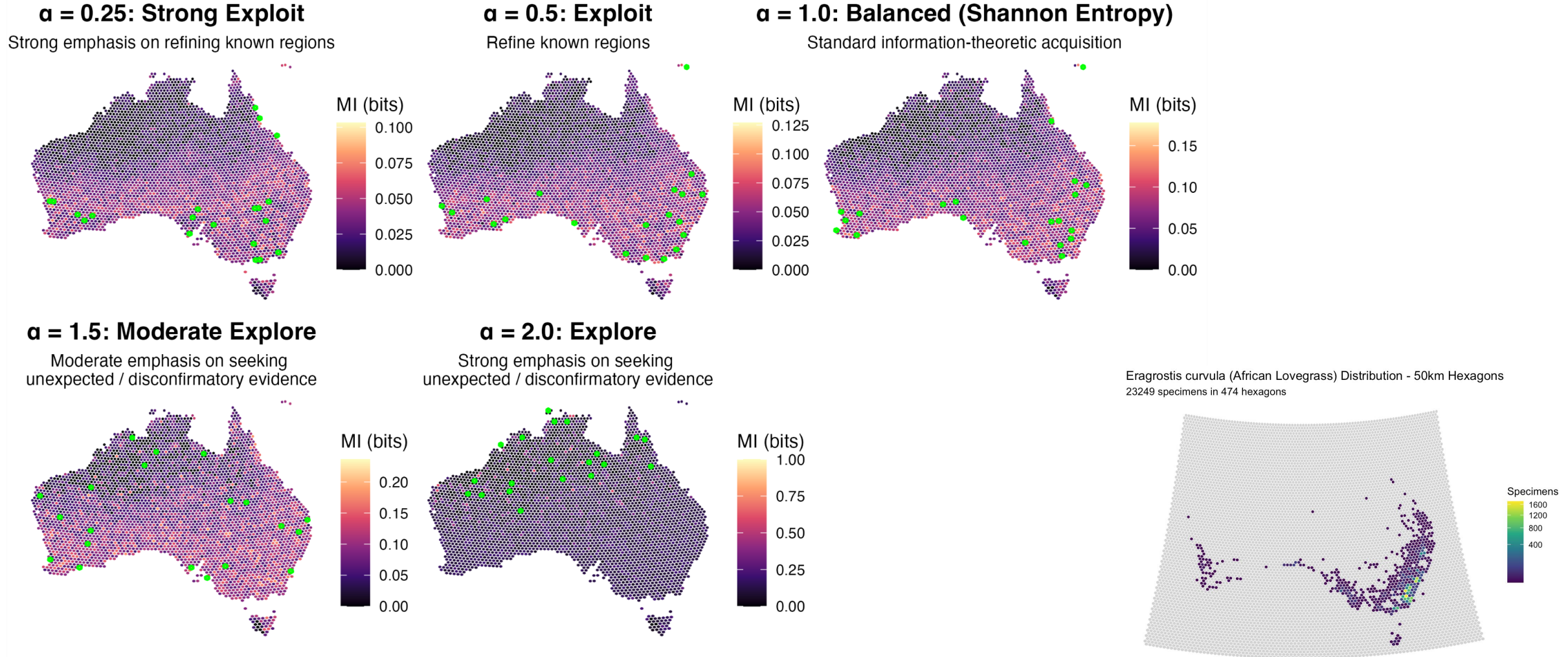


Vizumap: an R package for visualising uncertainty in spatial data
Journal of Open Source Software, 6(59), 2409, 2021

Exploration - exploitation

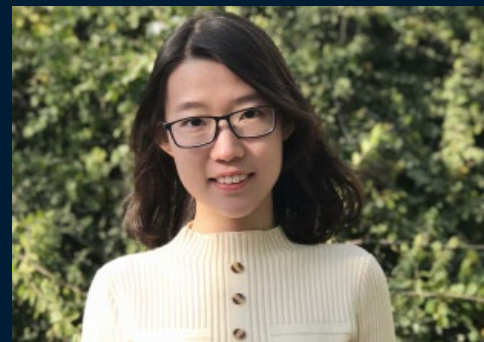
Fig 2: Acquisition Function Recommendations by Alpha for Renyi Entropy Search

Green outlines show top 20 recommended sampling locations for each policy

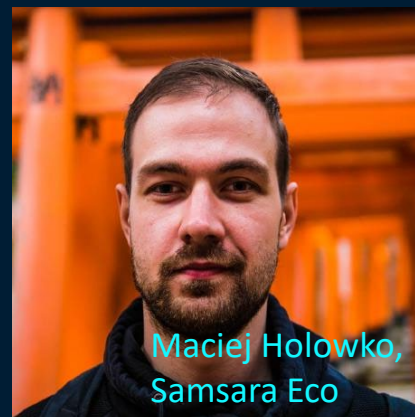


Higher gene expression

- Ribosomal binding site
- Given 6 letters of a genome, predict gene expression
- Sequence kernel
- Use Bayesian optimisation (GP-BUCB, kriging believer)
- 4 experimental cycles
- 35% stronger than engineered sequence



Mengyan Zhang, Oxford



Maciej Holowko,
Samsara Eco



Huw Hayman Zumpe,
Nourish



ACS Synthetic Biology
Volume 11, Issue 7
July 15, 2022
Pages 2221-2526

ARTICLE

Machine Learning Guided Batched Design of a Bacterial Ribosome Binding Site

[View article page](#)

Mengyan Zhang, Maciej Bartosz Holowko, Huw Hayman Zumpe and Cheng Soon Ong

CITE

© 2022 The Authors. Published by American Chemical Society
<https://doi.org/10.1021/acssynbio.2c00015>



CSIRO's BioFoundry

UNOFFICIAL

- **Engineering biology:**
*is the set of methods for **designing, building, and testing** engineered biological systems*



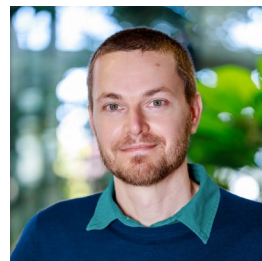
Hafna Ahmed



Chie Ishitate



Candice Jones
MPI Marburg



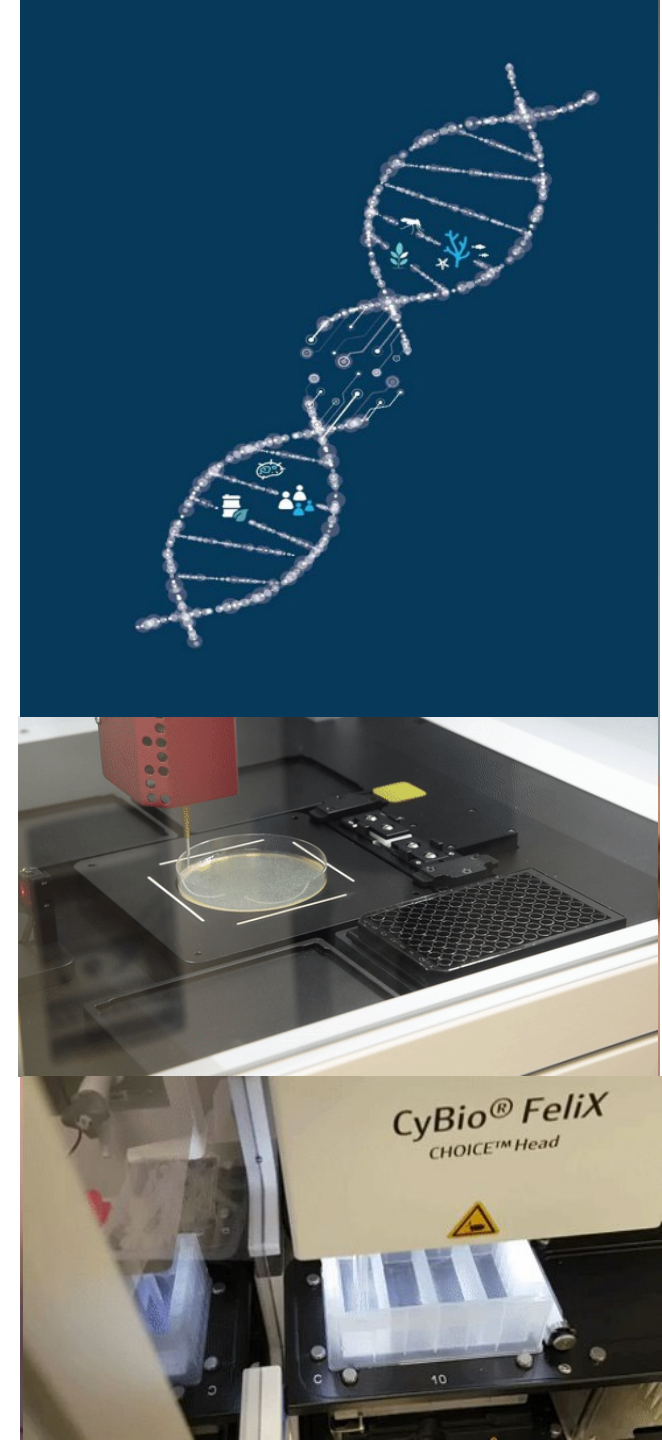
Adrian Marsh



Robert Speight

<https://research.csiro.au/aeb>

UNOFFICIAL



What's the objective?

Design Ribosome Binding Site (RBS) sequences



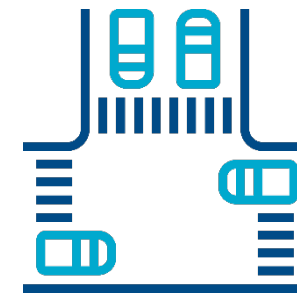
Optimize the protein expression level.

RBS sequence	Normalized* Protein Expression Level
TTTAAGAGTTATATATACAT	1.58
TTTAAGAAATATGCTATACAT	1.42
TTTAAGACTCGGATATACAT	0.14
TTTAAGAGTTTTTATACAT	2.88

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

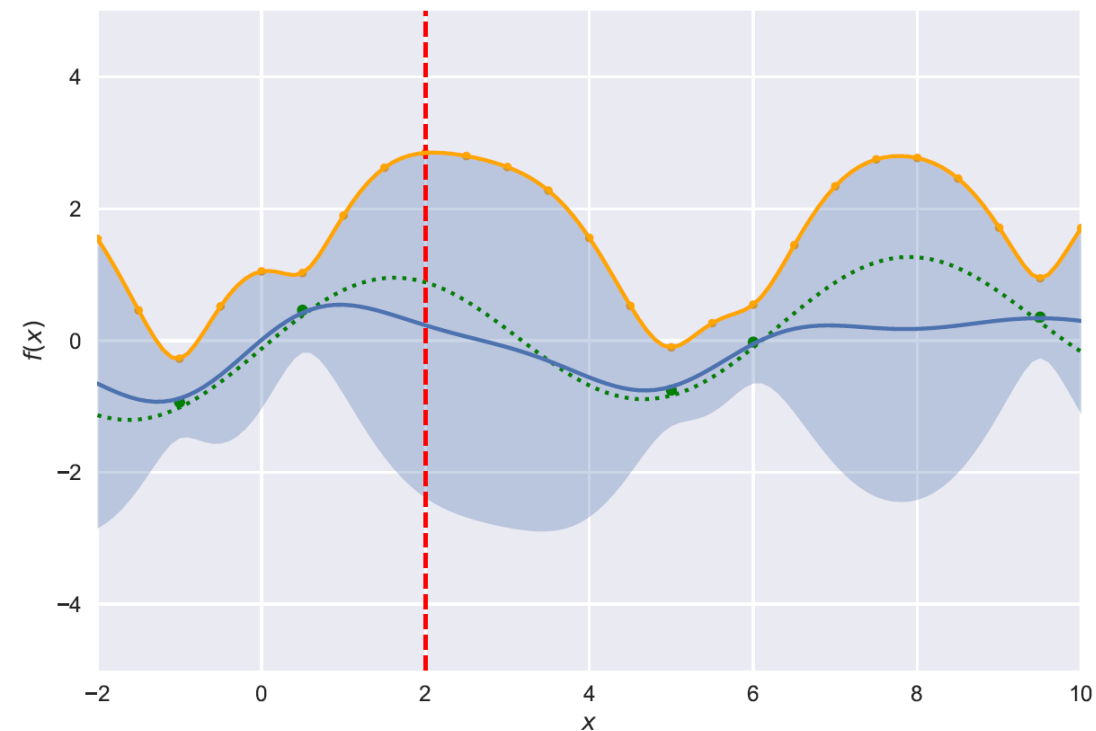
Upper confidence bound

A particular bandit algorithm

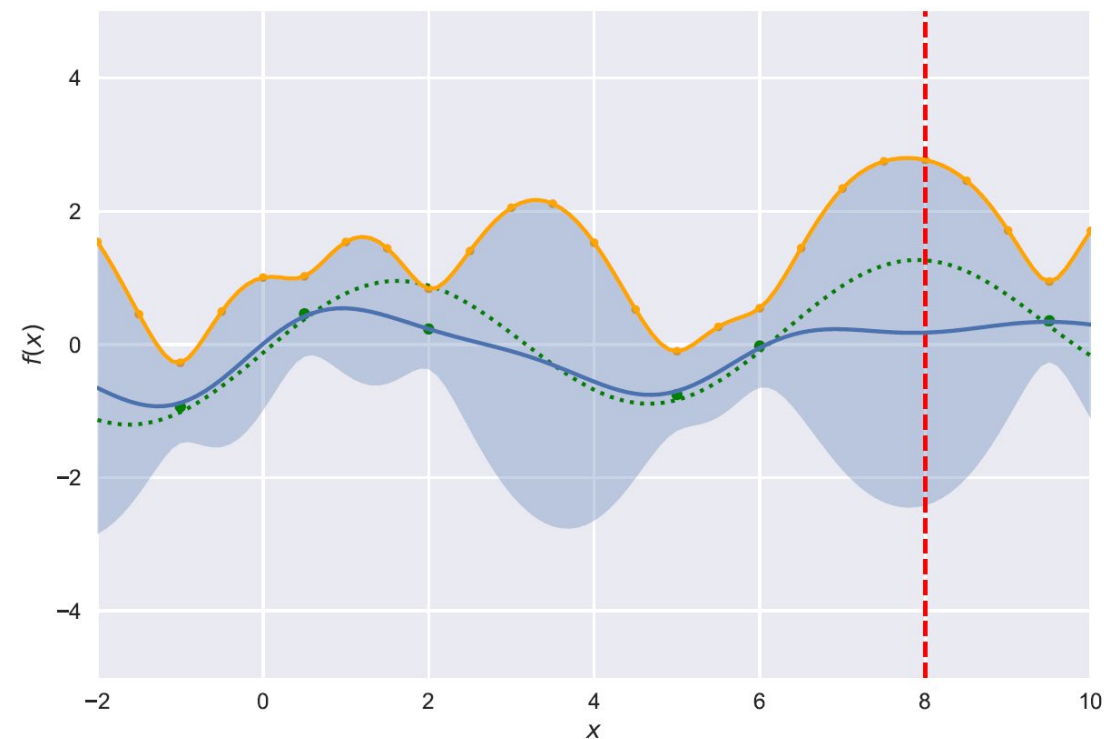


Decisions

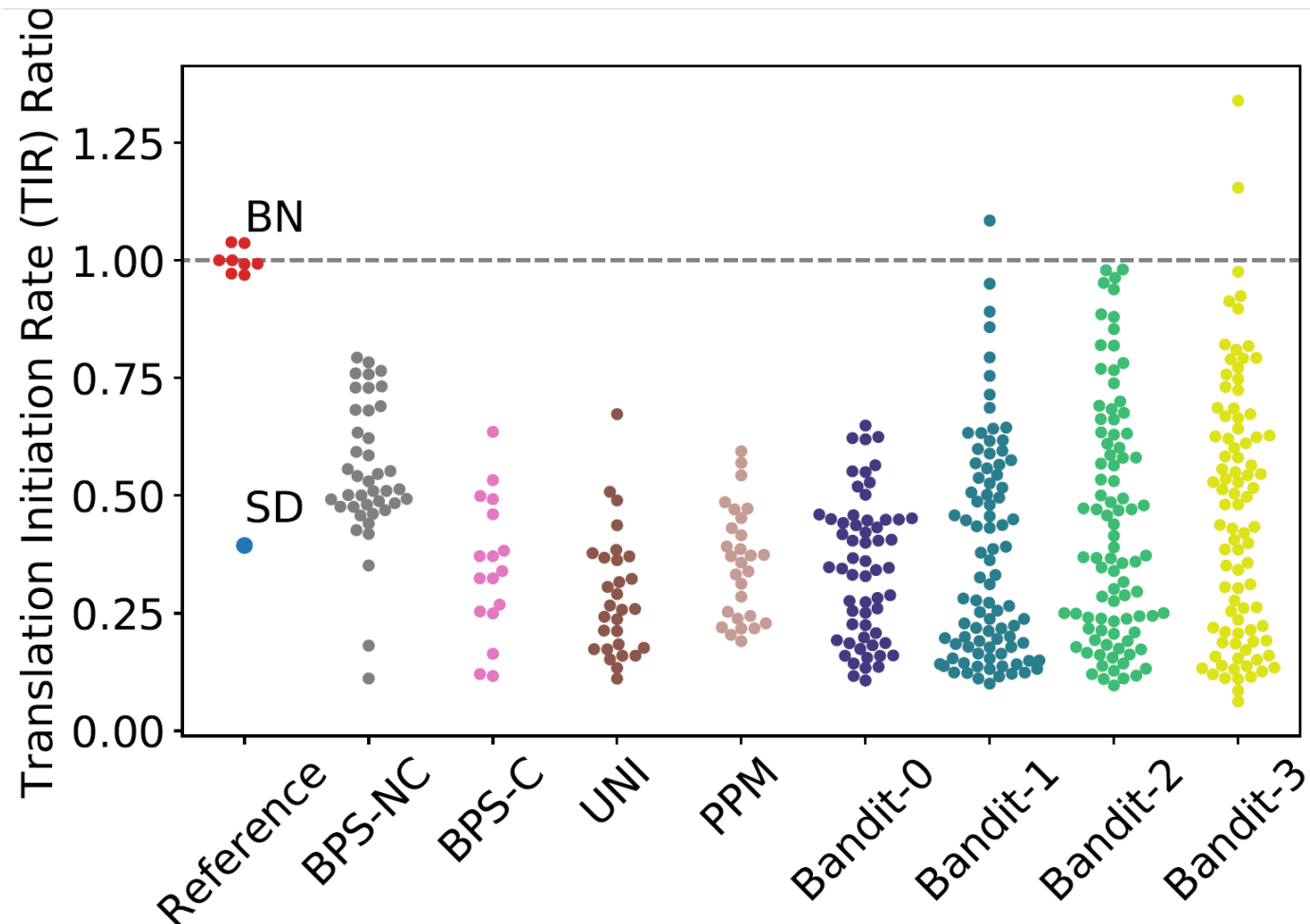
GP-BUCB: Recommend 1st Data Point



GP-BUCB: Recommend 2nd Data Point



ML recommends good designs



TTTAAGANNNNNTATACATATG
-20 Feature -1

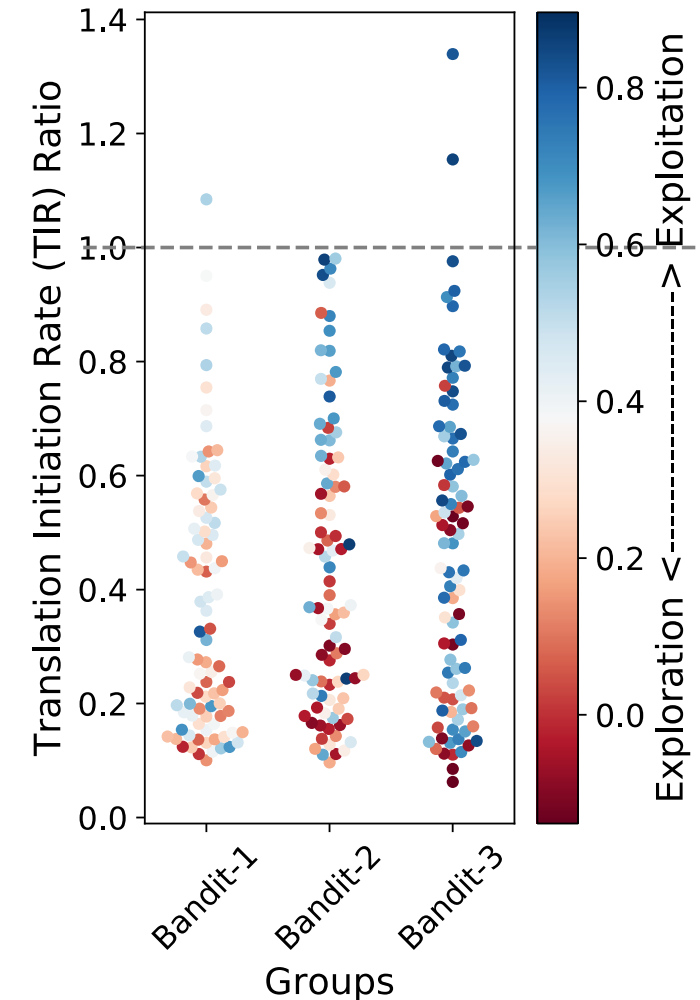
- Hard to search by evolving sequences
- 4 experimental cycles
- 35% stronger than engineered sequence

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

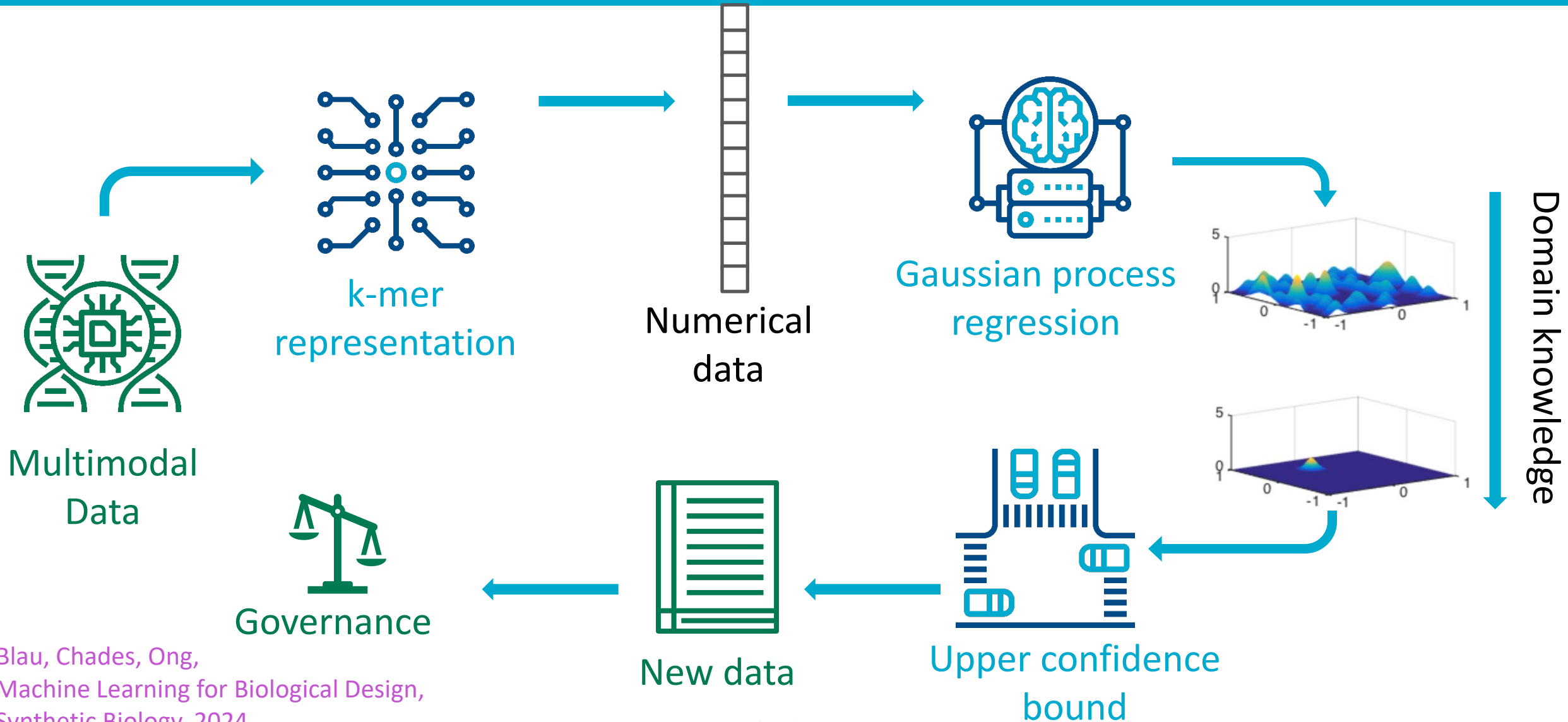
Exploration-Exploitation Trade-off

- 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?



Machine learning for genomics





UNOFFICIAL

Work in progress: find plastic degrading enzymes

Enzyme – a short protein sequence, acts as a catalyst for a chemical reaction. Represented as a string of amino acid letters but has a 3D form.

Aim: we want to be able to **find useful** enzymes that can digest plastic

Challenge – search space is **VAST**

Atoms in universe: 10^{80}

Seconds since Big Bang: 10^{17}

ppEST: Aryl esterase

```
MGTLVVGDSISAAFGLDLSRQGWVALLKRLSEEGFEHSVVNASISGDTSGAGGAARLSALLAEHKPELVIIELGGNDGL  
RGQPPAQLQQNLASMVEQSQQAGAKVLLGMKLPPNYGVRYTTAFAQVFTDLAEQKQVSLVPFFLEGVGGVPGM  
MQADGIHPAEAAQEILLDNVWPTLKPML
```



20^{200} possible (substitution) variants!



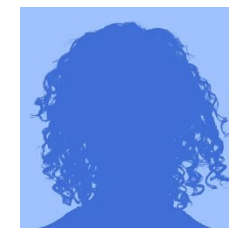
Asiri Wijesinghe



Allen Zhu



Lu Zhang

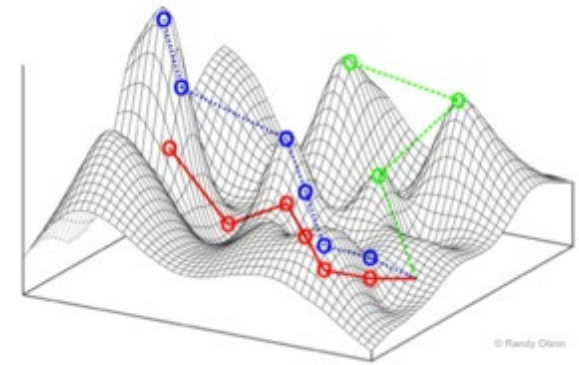


Asher Bender

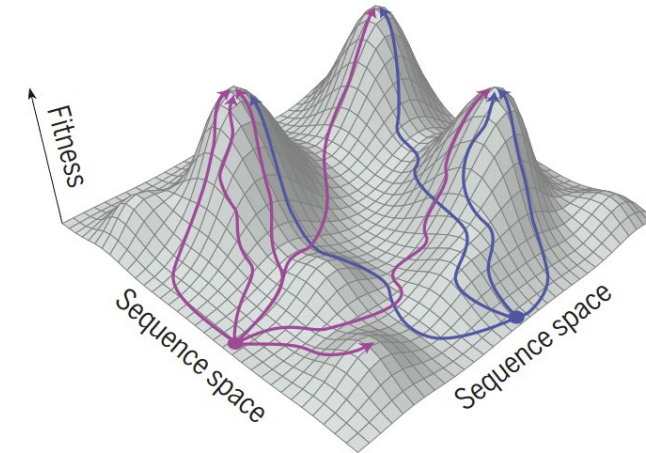
UNOFFICIAL

Fitness landscape

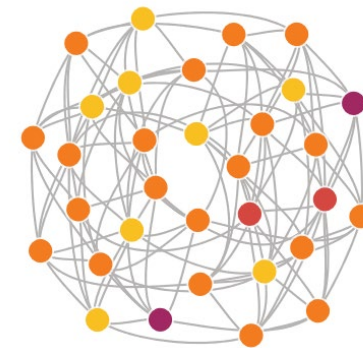
- A metaphor in evolutionary biology
- Each of the horizontal axes represents some notion of sequence variation
- The vertical axis captures some property of interest (so-called fitness)
- Open question how to represent



Wikipedia: Sewall Wright



Papkou et al., Science 382, 901 (2023)



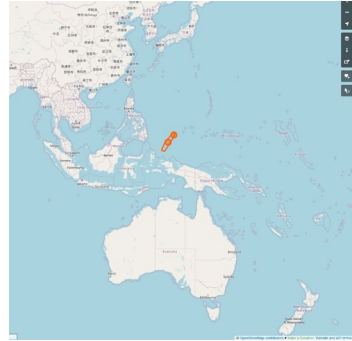
Fitness 
Low High

Our idea: Islands of fitness

- Desiderata: want to find only “fit” sequences
- Intuition: Many sequences are not viable,
and we cannot measure their fitness



Aerial View of Seventy Islands, Micronesia, Palau
by Reinhard Dirscherl

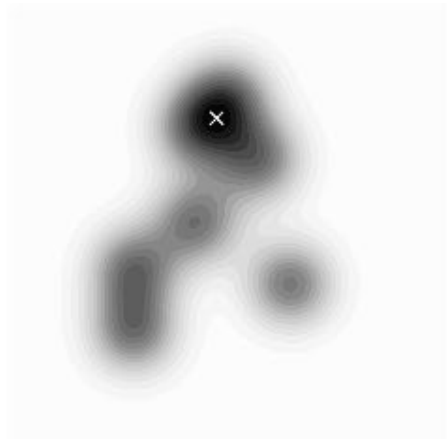


Active Generation

- Goal: We want to generate from a (conditional) probability density
$$p(x \mid y > \tau)$$

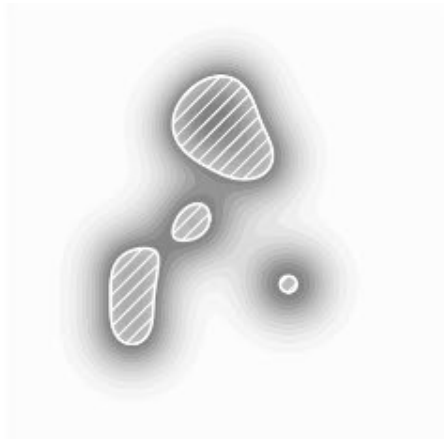
Where x is the space of sequences and y is the fitness value
 τ is a parameter that identifies “fit” sequences

Variational Search Distributions



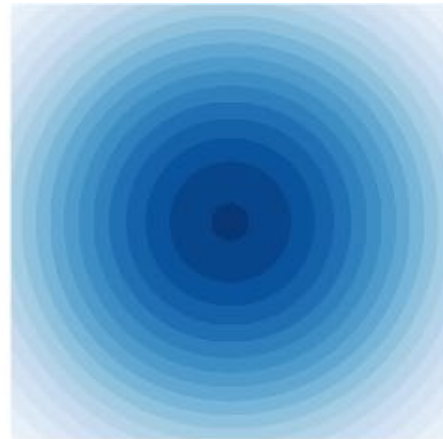
(a) $\operatorname{argmax}_{\mathbf{x}} y(\mathbf{x})$

Find the fittest



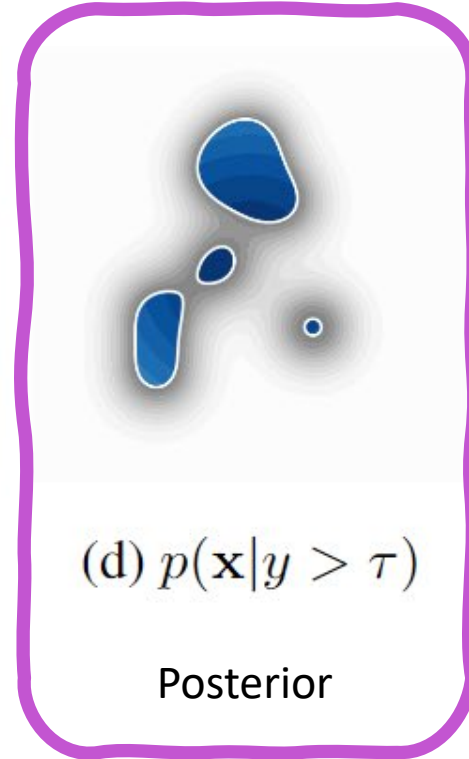
(b) $\mathcal{S}$

Set of viable
sequences



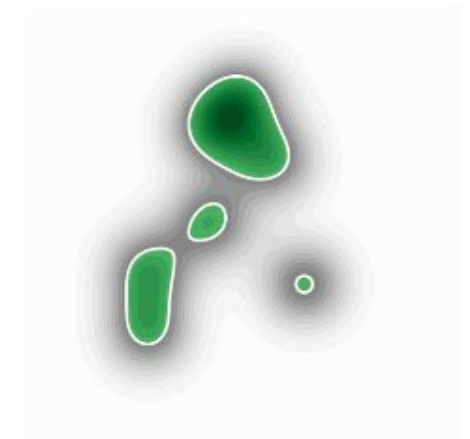
(c) $p(\mathbf{x})$

Prior



(d) $p(\mathbf{x}|y > \tau)$

Posterior



(e) $\mathcal{F}$

Ground
truth

Steinberg, Oliveira, Ong, Bonilla,
Variational Search Distributions, ICLR 2025
<https://arxiv.org/abs/2409.06142>



Dan Steinberg



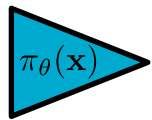
Rafael Oliveira



Edwin Bonilla

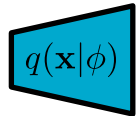
Solving active generation

- Frame online black box optimization as sequential learning of conditional generation
- In each round of the sequence, there are two steps



Fit a binary classifier (CPE), $z := \mathbf{1}[y > \tau]$ indicates “good”

$$\pi_{\theta}(\mathbf{x}) \approx p(z = 1|\mathbf{x})$$



Update the generative model

$$\phi_t^* \leftarrow \underset{\phi}{\operatorname{argmax}} \mathcal{L}_{\text{ELBO}}(\phi, \theta_t^*)$$

- We choose to maximise the evidence lower bound (variational inference)

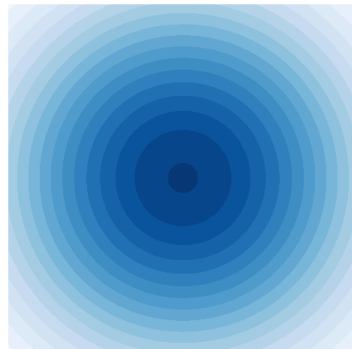
$$\mathcal{L}_{\text{ELBO}}(\phi, \theta) = \mathbb{E}_{q_{\phi}(\mathbf{x})}[\log \pi_{\theta}(\mathbf{x})] - \mathbb{D}_{\text{KL}}[q_{\phi}(\mathbf{x}) \| p(\mathbf{x}|\mathcal{D}_0)]$$

Generating unrolled images

Sample mean CPE $p=0.130$



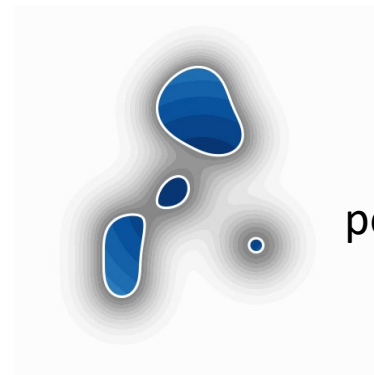
prior



Sample mean CPE $p=0.997$

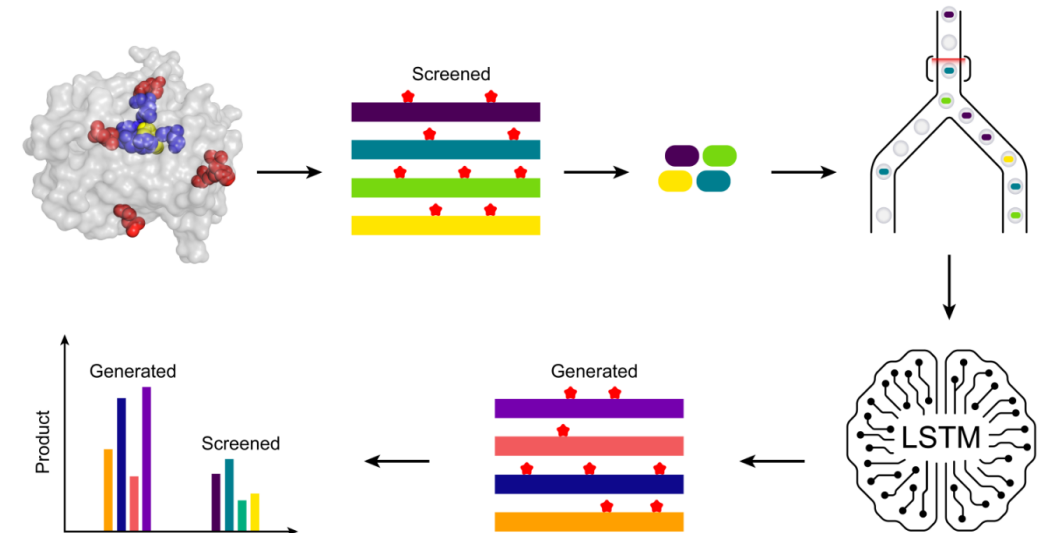
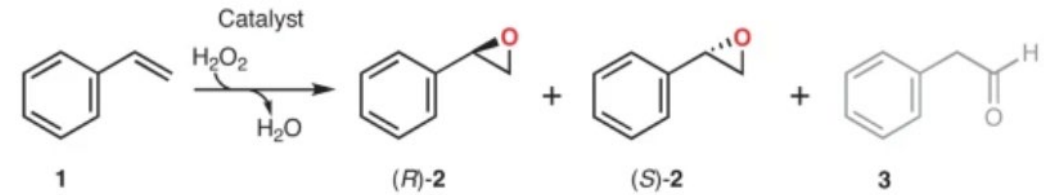


posterior($\{3,5\}$)



Early results - peroxygenases

- Peroxygenase is a catalyst that inserts an oxygen atom
- Want to **engineer specificity** into unspecific peroxygenase
- Enzyme library is screened using microfluidic sorting
- Active generation consistently outperformed direct selection from the same screening data



Ultrahigh throughput screening to train generative protein models for engineering specificity into unspecific peroxygenases
 Nair, Steinberg, et. al.
<https://www.biorxiv.org/content/10.1101/2025.11.02.685536v1>

Active Generation

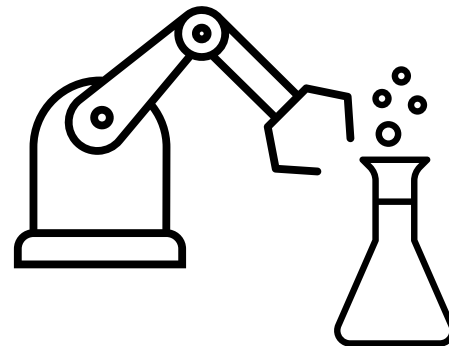
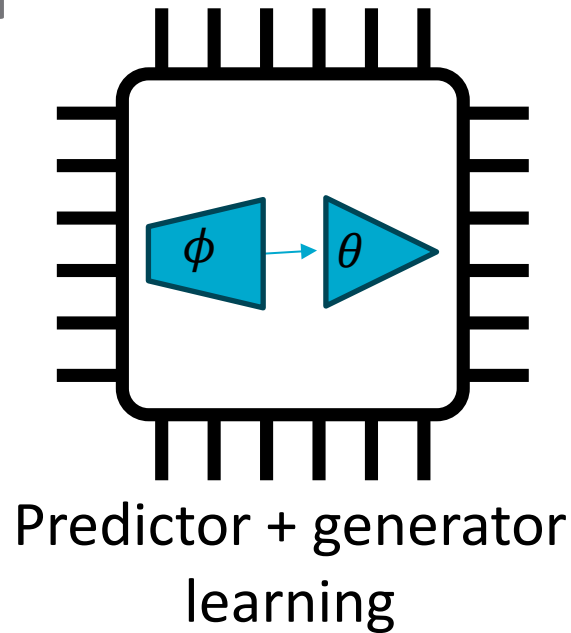
Active generation: find best ϕ^* for generating “good” (or best) $\mathbf{x}$

MKTTTL...LFLVGALTQ	1.2
MKTTTL...LFLVGTLTQ	3.6
...	...
MKTTTL...LFLVGALTT	0.3

Labelled
Data

$$\pi_{\theta}(\mathbf{x}) \approx p(z = 1|\mathbf{x})$$

$$\phi_t^* \leftarrow \underset{\phi}{\operatorname{argmax}} \mathcal{L}_{\text{ELBO}}(\phi, \theta_t^*)$$



Build + Test

Can “select” from a **vast number** of enzymes (e.g. 20^{200}), since they are **generated**

** we are **not** doing latent space optimisation!

ϕ

Generate good candidates instead of selecting from a list

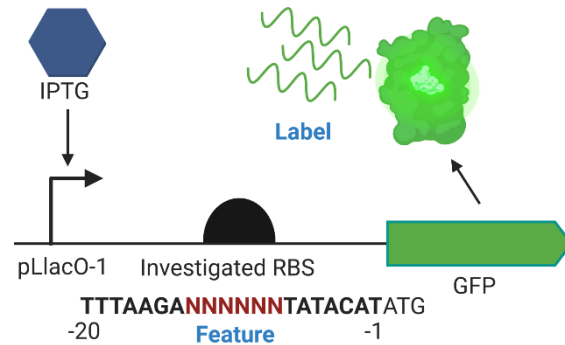
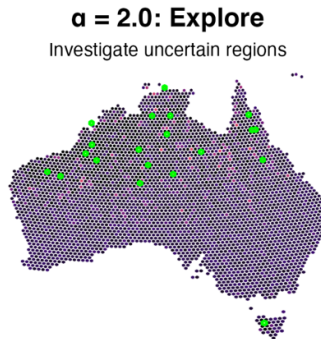


MKTTTL...LFLVGALTQ
MKTTTL...LFLVGTLTQ
MKTTIL...LFLVGTLTQ
MKTSTL...LFLVGTLTQ
MKTTTL...LFLVGTLTQ
...
MKTTTL...LFLVGALTT

Unlabelled data
generation,
design is generation

Data is measured for a reason

- Consider the set of all possible things to measure

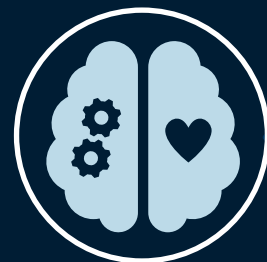


- Think of the predictor output as producing features
 - Each calculated feature demonstrates the “importance” of an experiment
 - Can get multiple features by using different predictors
 - Estimate predictive uncertainty to inform decisions
- Adaptively choose the next thing to measure by maximising an objective
(machine learning is about defining good objective functions)

On finding good experiments



What is an experiment?

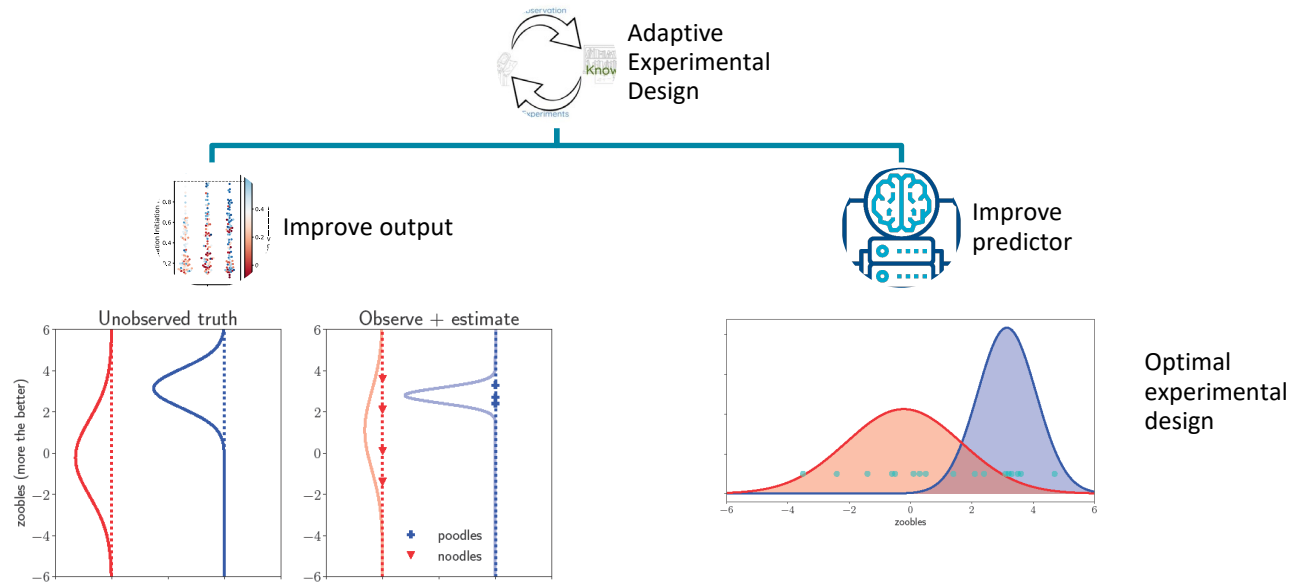


How do we find good experiments?



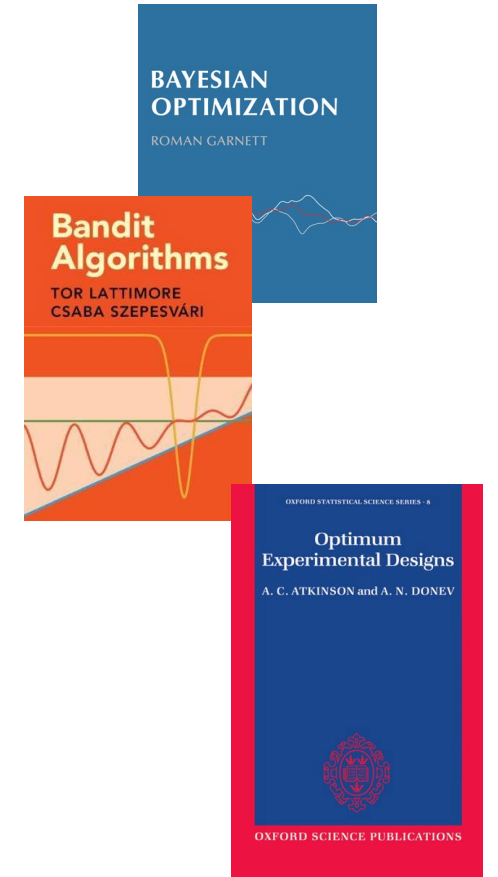
What do we mean by good?

Better measured values or better models

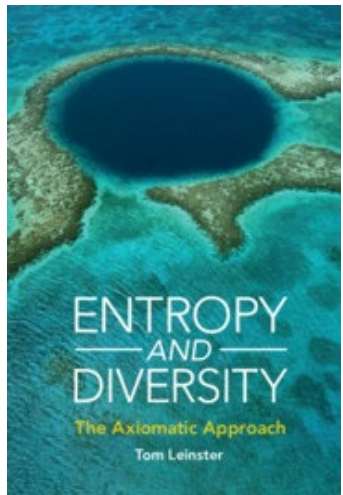
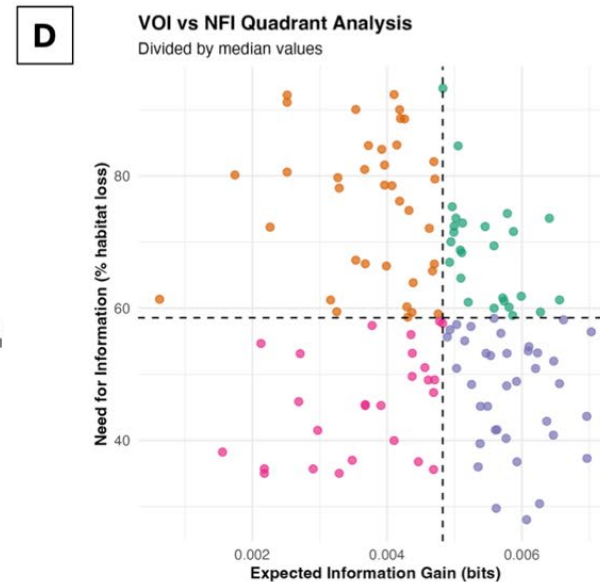
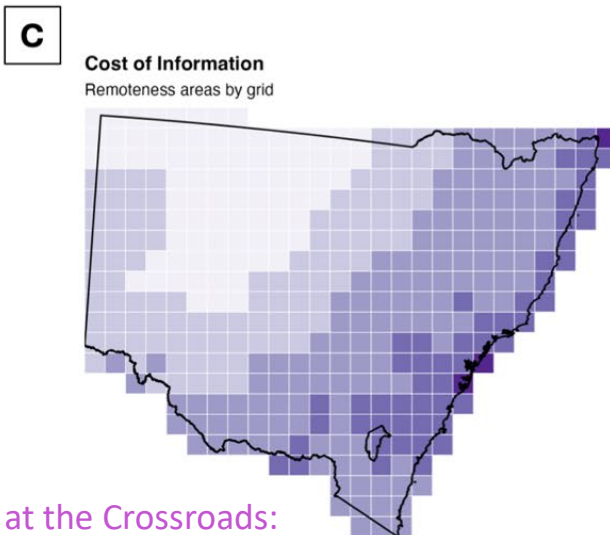
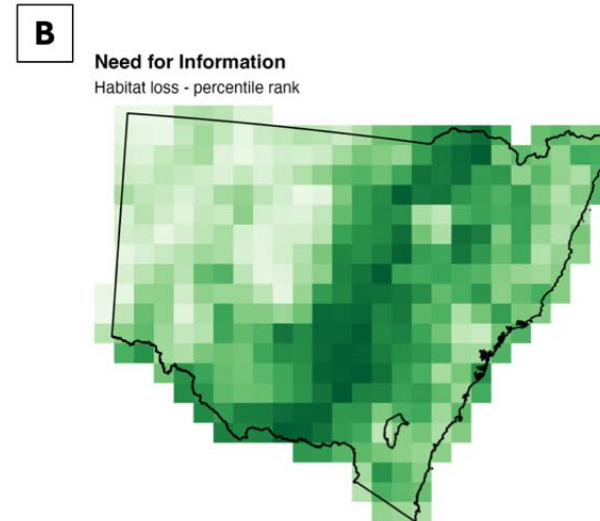
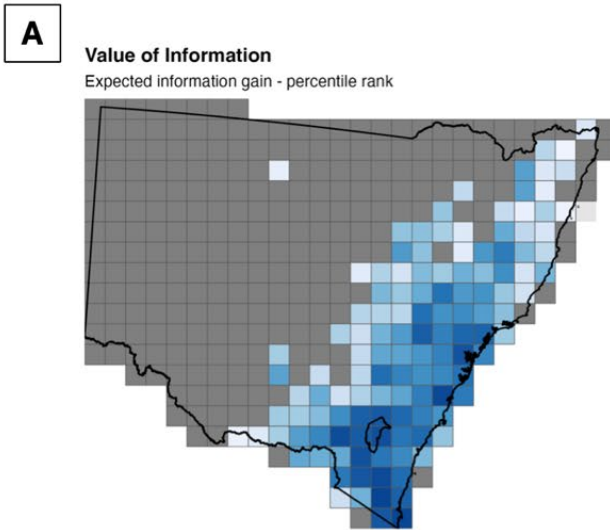


Recall that we can use a machine learning predictor in two ways:

1. The parameters of the model
2. The output values on a test set



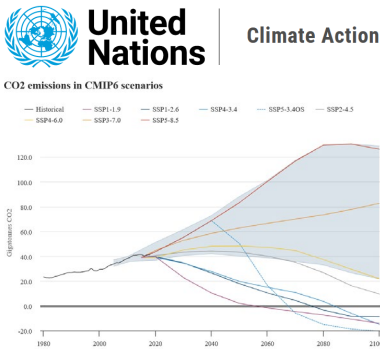
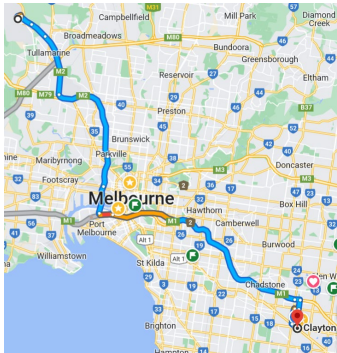
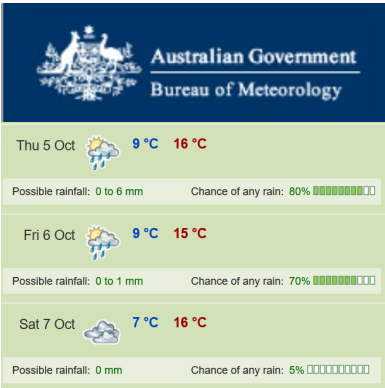
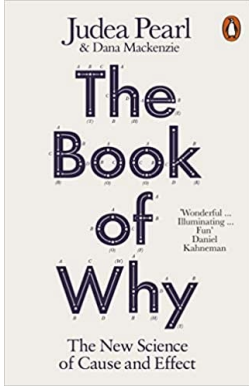
Value, need, cost



Natural History Collections at the Crossroads:
Shifting Priorities and Data-Driven Opportunities
Forbes, Thrall, Young, Ong, Ecology Letters, vol 28, no 8, 2025

ML is **not only** about predictions

Predictions vs Decisions vs Actions



	Weather	Traffic	Climate
Predictions	Will it rain tomorrow?	Jam on M1?	Risk in 2050?
Decisions	Take umbrella?	Train or taxi?	Plan for net zero
Actions	Does not affect weather	Affects traffic!	Want wrong predictions!

When accurate prediction models yield harmful self-fulfilling prophecies, Patterns, 2025
<https://doi.org/10.1016/j.patter.2025.101229>



Pragmatism in International Relations



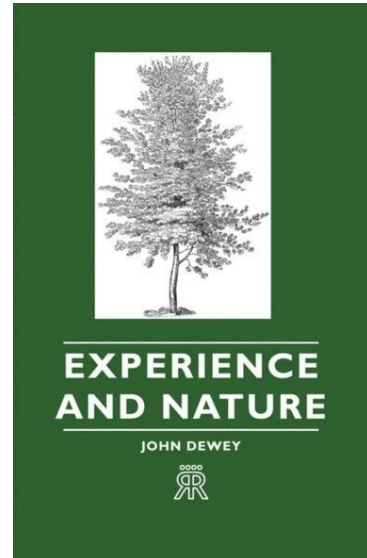
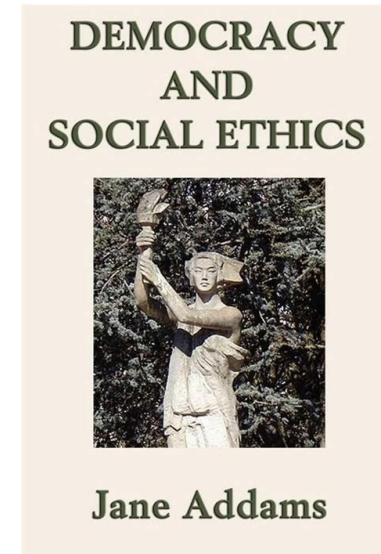
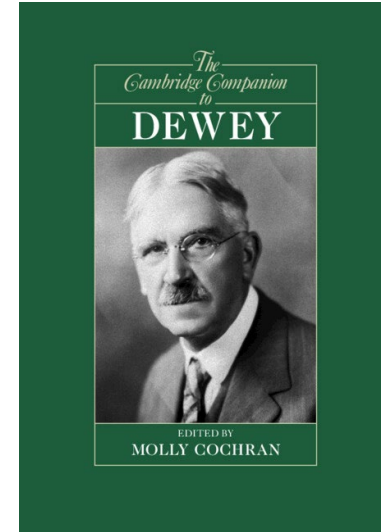
Toni Erskine, ANU



Xueyin Zha, ANU

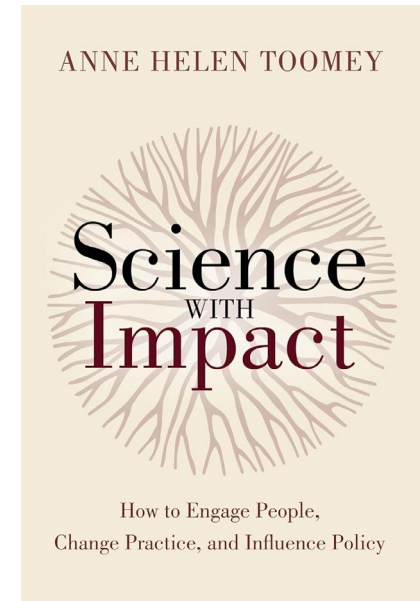
- Need a way to say “good” and “true”

PhD thesis: Normative Truth-Seeking from the Ground Up:
Experiential Pathway to Global AI Governance



We should learn from each other

- Need more than data science
- How to foster cross disciplinary projects?
- π shaped research teams



IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY

Four Compelling Reasons to Urgently Integrate AI Development With Humanities, Social and Economics Sciences

Iadine Chades¹, Melanie McGrath², Erin Bohensky, Lucy Carter³, Rebecca Coates⁴, Ben Harwood, Md Zahidul Islam, Sevvandi Kandanaarachchi⁵, Cheng Soon Ong⁶, Andrew Reeson⁷, Samantha Stone-Jovicich⁸, Cécile Paris⁹, Mitchell Scovell¹⁰, Kirsty Wissing¹¹, and David M. Douglas¹²

Position: We need responsible, application-driven (RAD) AI research

Opportunities and Challenges in Designing Genomic Sequences

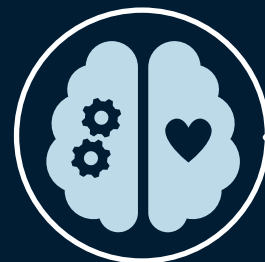
Sarah Hartman¹ Cheng Soon Ong^{2,3} Julia Powles^{4,5} Petra Kuhnert¹

Mengyan Zhang^{1,2} Cheng Soon Ong^{2,1}

On Finding Good Experiments



What is an experiment?



How do we find good experiments?



What do we mean by good?