We know how to do binary classification…

Muffin vs chihuahua

Cinnamon rolls vs tails

Some ways to use a binary classifier

Cheng Soon Ong,
Data61 and ANU
10 December 2025

Mathematics in Data Science Workshop,
University of Sydney

CSIRO

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

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 degrade plastic (e.g. PET).

Challenge – search space is **VAST**

Atoms in universe: 10^{80}
Seconds since Big Bang: 10^{17}

ppEST: **Aryl esterase**

MGTLLVVGDSISAAFGLDSRQGWVALLEKRLSEEGFEHSVVNASISGDTSAGGAARLSALLAEHKPELVIIELGGNDGL
RGQPPAQLQQNLASMVEQSQQAGAKVLLLGMKLPPNYGVRYTTAFAQVFTDLAEQKQVSLVPFFLEGVGGVPGM
MQADGIHPAEAAQEILLDNVWPTLKPML

20^{200} possible (substitution) variants!

Dan Steinberg

Asiri Wijesinghe

Allen Zhu

Lu Zhang

Asher Bender

UNOFFICIAL

CSIRO's BioFoundry

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

Hafna Ahmed

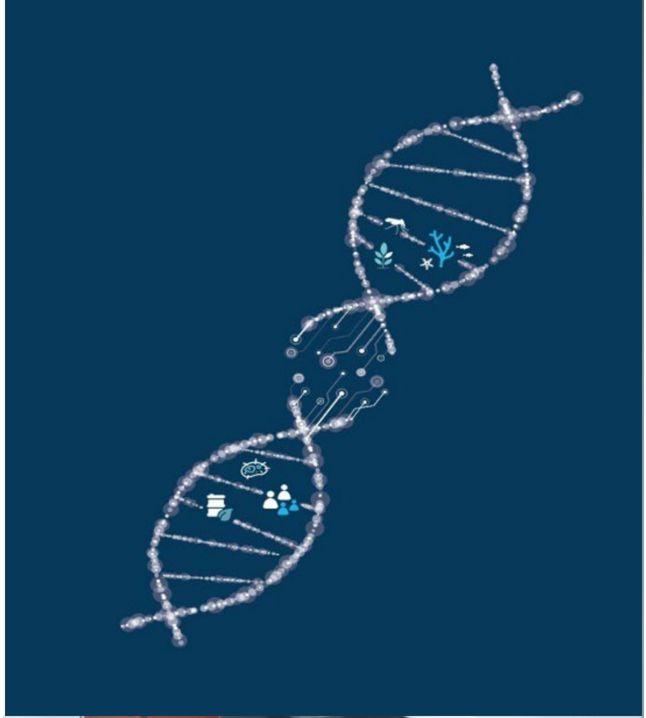
Chie Ishitate

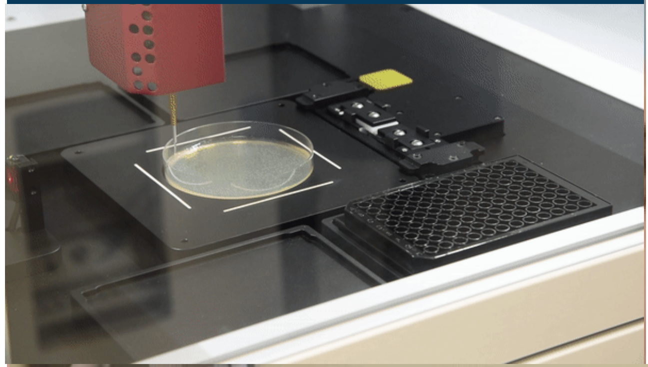
Candice Jones
MPI Marburg

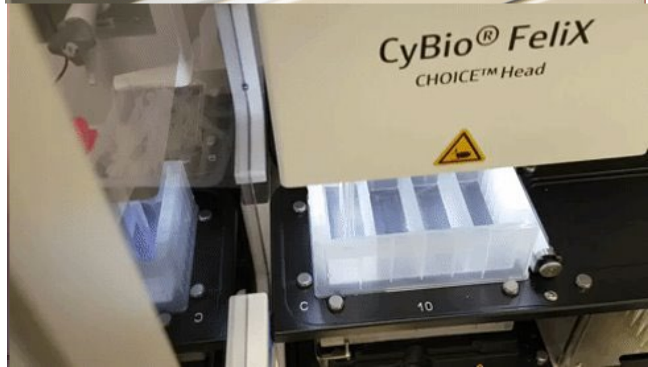
Adrian Marsh

Robert Speight

https://research.csiro.au/aeb

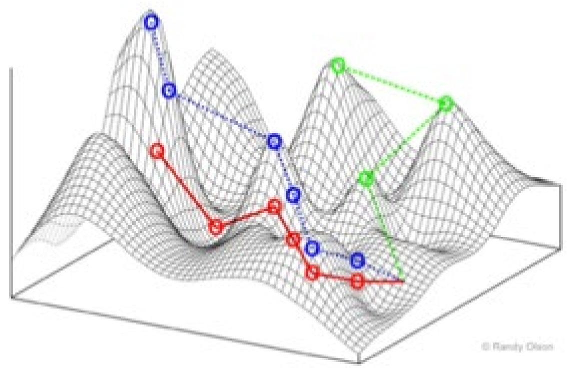




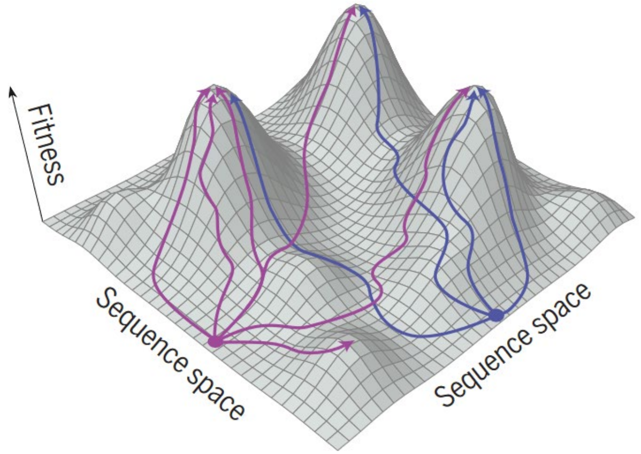


Fitness landscape

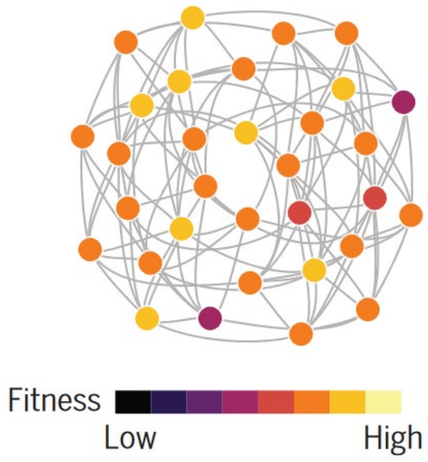
- A metaphor in evolutionary biology
- Each of the horizontal axes represents some notion of sequence variation
- The vertical axis captures some property of interest (so-called fitness)
- Open question how to represent



Wikipedia: Sewall Wright



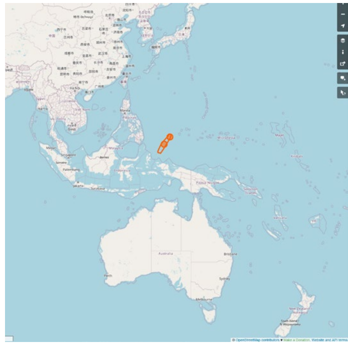
Papkou et al., Science 382, 901 (2023)



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

$$p(x|y > \tau)$$

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

Three messages

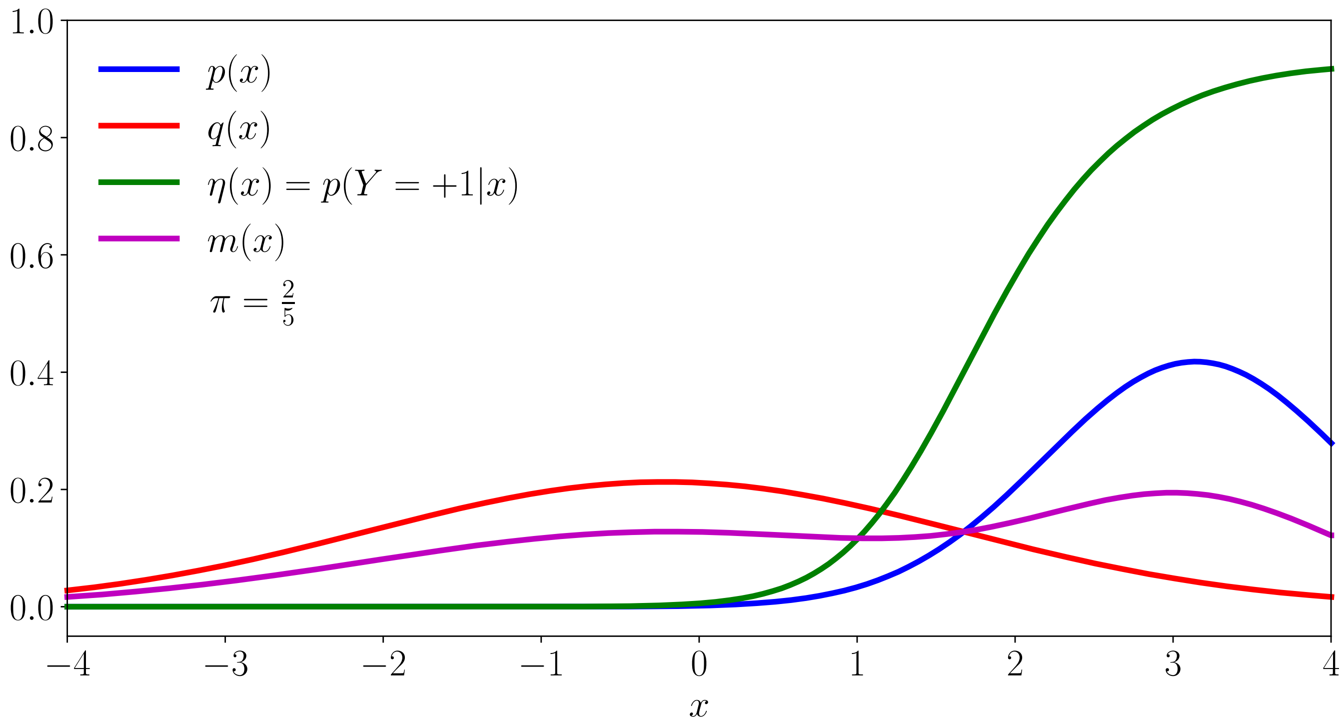
Class probability estimation to density ratios

Density ratio to Bayesian optimisation

Searching the space of densities

Two ways to represent binary classification

- Two densities P and Q, and a class prior $\mathcal{D} = (P, Q, \pi)$
- A marginal density M, and a conditional probability $\mathcal{D} = (M, \eta)$



Two ways to represent binary problems

- Positive density $p(x|z = 1)$
- Negative density $p(x|z = -1)$
- Class prior $\pi := p(z = 1)$

- Marginal density $p(x)$
- Class probability = conditional probability of positive

$$\eta(x) := p(z = 1|x)$$

Density ratios

Aditya Menon
Google Deepmind

- Going from class probability estimation to density ratio estimation
- From Bayes' Theorem

$$\forall x \in \mathcal{X} \quad \frac{\eta(x)}{1-\eta(x)} = \frac{p(x|z=1)}{p(x|z=-1)} \cdot \frac{\pi}{1-\pi}$$

- Hence the density ratio

$$\forall x \in \mathcal{X} \quad \frac{p(x|z=1)}{p(x|z=-1)} = \frac{\eta(x)}{1-\eta(x)} \cdot \frac{1-\pi}{\pi}$$

Menon, Ong, Linking losses for density ratio and class-probability estimation, ICML 2016

Linking losses

- Observe that the density ratio is a transformation of the class probability estimate, where the analytic relation is

$$\Psi_{\mathrm{DR}}(a) := \frac{1-\pi}{\pi} \cdot \frac{a}{1-a}$$

- There is a direct link between the losses of density ratio estimation and class probability estimation

$$r(x) = \frac{\eta(x)}{1-\eta(x)} \cdot \frac{1-\pi}{\pi} = \Psi_{\mathrm{DR}}(\eta(x))$$

Linking losses

- There is a direct link between the losses of density ratio estimation and class probability estimation
- We analyse the properties of imperfect estimates
- Via a novel Bregman identity of a perspective transform
- We provide a way to design new losses

Loss	$\ell_{-1}(v)$	$\ell_1(v)$	$\Psi^{-1}(v)$
KLIEP	v	$-\log v$	$\frac{v}{1+v}$
LSIF	$\frac{1}{2}v^2$	$-v$	$\frac{v}{1+v}$
Power$_\alpha$	$\frac{v^{1+\alpha}-1}{1+\alpha}$	$\frac{1-v^\alpha}{\alpha}$	$\frac{v}{1+v}$
Square	$(1+v)^2$	$(1-v)^2$	$2v-1$
Logistic	$\log(1+e^v)$	$\log(1+e^{-v})$	$\frac{1}{1+e^{-v}}$
Exponential	e^v	e^{-v}	$\frac{1}{1+e^{-2v}}$

Menon, Ong, Linking losses for density ratio and class-probability estimation, ICML 2016
Nock, Menon, Ong, A scaled Bregman theorem with applications, NeurIPS 2016

Three messages

- Class probability estimation to density ratios
- Density ratio to Bayesian optimisation
- Searching the space of densities

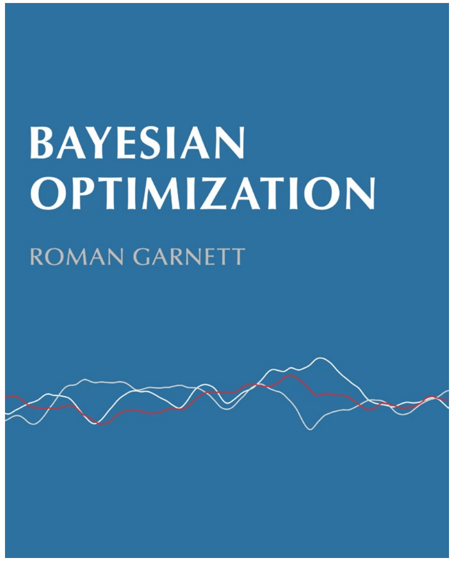
Probability of improvement

- Bayesian optimisation is an adaptive experimental design approach
- Following Bayesian decision theory, optimise expected utility

$$\mathbb{E}_{y \sim p(y|x,\mathbf{D})}\left[u(y;\tau)\right]$$

- One utility is probability of improvement

$$u_{\mathrm{PI}}(y;\tau) := \mathbb{I}(y > \tau)$$



Probability of improvement = density ratio

- Expectation of PI utility

$$\mathbb{E}_{y\sim p(y|x,\mathbf{D})}\left[u_{\mathrm{PI}}(y;\tau)\right]$$

- By the definition of expectation

$$\begin{aligned} &= \int_{-\infty}^{\infty} \mathbb{I}(y>\tau)p(y|x,\mathbf{D})\mathrm{d}y \\ &= \int_{y=\tau}^{\infty} p(y|x,\mathbf{D})\mathrm{d}y \\ &= p(y>\tau|x,\mathbf{D}) \end{aligned}$$

Probability of improvement = density ratio

- The same as density ratio (by Bayes' theorem)

$$p(y > \tau | x) p(x) = p(x | y > \tau) p(y > \tau)$$

- Observe that $p(y > \tau)$ is free of x
- Therefore we have a density ratio

$$\mathbb{E}_{y \sim p(y|x, \mathbf{D})} \left[u_{\mathrm{PI}}(y; \tau) \right] \propto \frac{p(x | y > \tau)}{p(x)}$$

Song, Yu, Neiswanger, Ermon, A General Recipe for Likelihood-free Bayesian Optimization, ICML 2022
Tiao, Klein, Seeger, Bonilla, Archambeau, Ramos, BORE: Bayesian Optimization by Density-Ratio Estimation, ICML 2021

Probability of improvement = binary classifier

- Score

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

Class probability

$$\eta(x) : \mathcal{X} \to [0, 1]$$

- Thresholding the score

$$z := \mathbb{I}(y > \tau) \in \{0, 1\}$$
$$\eta(x) \approx p(z = 1|x) = p(y > \tau|x)$$

- Probability of improvement

$$\mathbb{E}_{p(y|x)}\left[u_{\mathrm{PI}}(y; \tau)\right] = p(y > \tau|x)$$

Three messages

Class probability estimation to density ratios

Density ratio to Bayesian optimisation

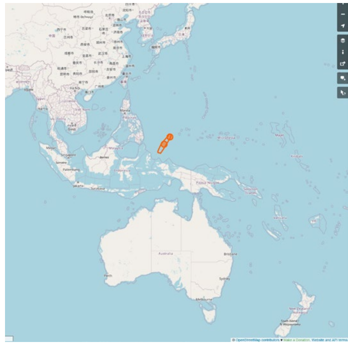
Searching the space of densities

UNOFFICIAL

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



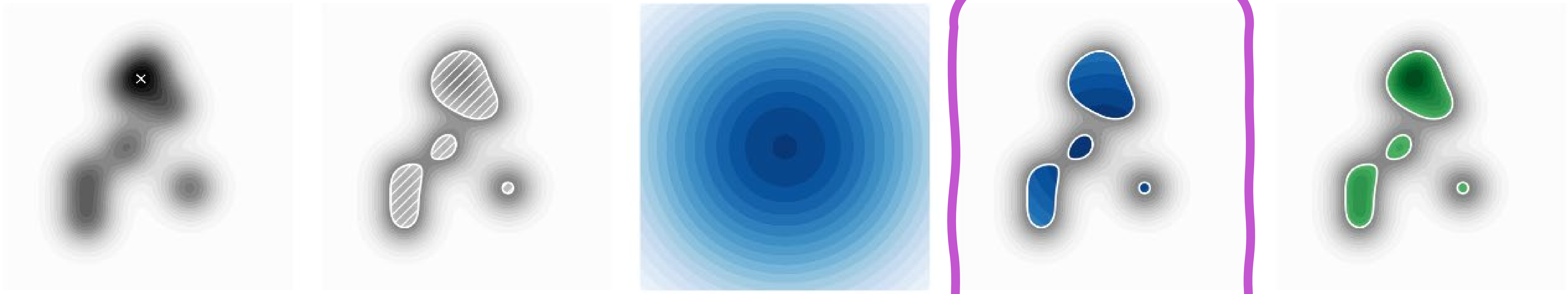
Active Generation

- Goal: We want to generate from a (conditional) probability density

$$p(x|y > \tau)$$

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) $\mathrm{argmax}_{\mathbf{x}}\, y(\mathbf{x})$ | (b) $\mathcal{S}$ | (c) $p(\mathbf{x})$ | (d) $p(\mathbf{x}|y > \tau)$ | (e) $\mathcal{F}$

Find the fittest

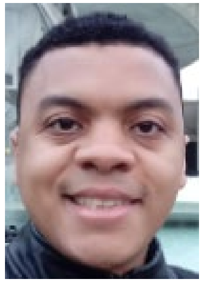
Set of viable sequences

Prior

Posterior

Ground truth

Dan Steinberg



Rafael Oliveira

Edwin Bonilla

https://arxiv.org/abs/2409.06142

Flexible surrogate models

- Build on advances in generative modelling
- Minimise reverse KL divergence of model to unknown density

$$\mathrm{argmin}_{\phi} \mathbb{D}_{\mathrm{KL}} \left[\, q(x|\phi) \; || \; p(x|y > \tau) \, \right]$$

- Surrogate $q(x|\phi)$
- Islands of fitness $p(x|y > \tau)$

Expected log probability of improvement

- Expanding KL divergence and applying Bayes' theorem

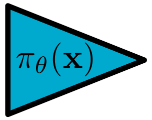
$$\mathrm{argmin}_{\phi} \mathbb{E}_{q(x|\phi)} \left[\log \frac{q(x|\phi)}{p(x)} - \log p(y > \tau | x) \right]$$

- Recall probability of improvement

$$\mathrm{argmax}_{\phi} \mathbb{E}_{q(x|\phi)} \left[\log u_{\mathrm{PI}}(y; \tau) \right] - \mathbb{D}_{\mathrm{KL}} \left[\ q(x|\phi) \ || \ p(x) \ \right]$$

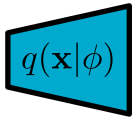
Solving active generation

- 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

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

- Since we use CPE, direct generalisation to multi-objective optimisation and finding Pareto-sets

UNOFFICIAL

Theoretical analysis

- The learned distribution approaches the true distribution

Theorem 2. *Let assumptions 1 to 5 hold. Then VSD equipped with GP-PI approaches the level-set distribution at the following rate:*

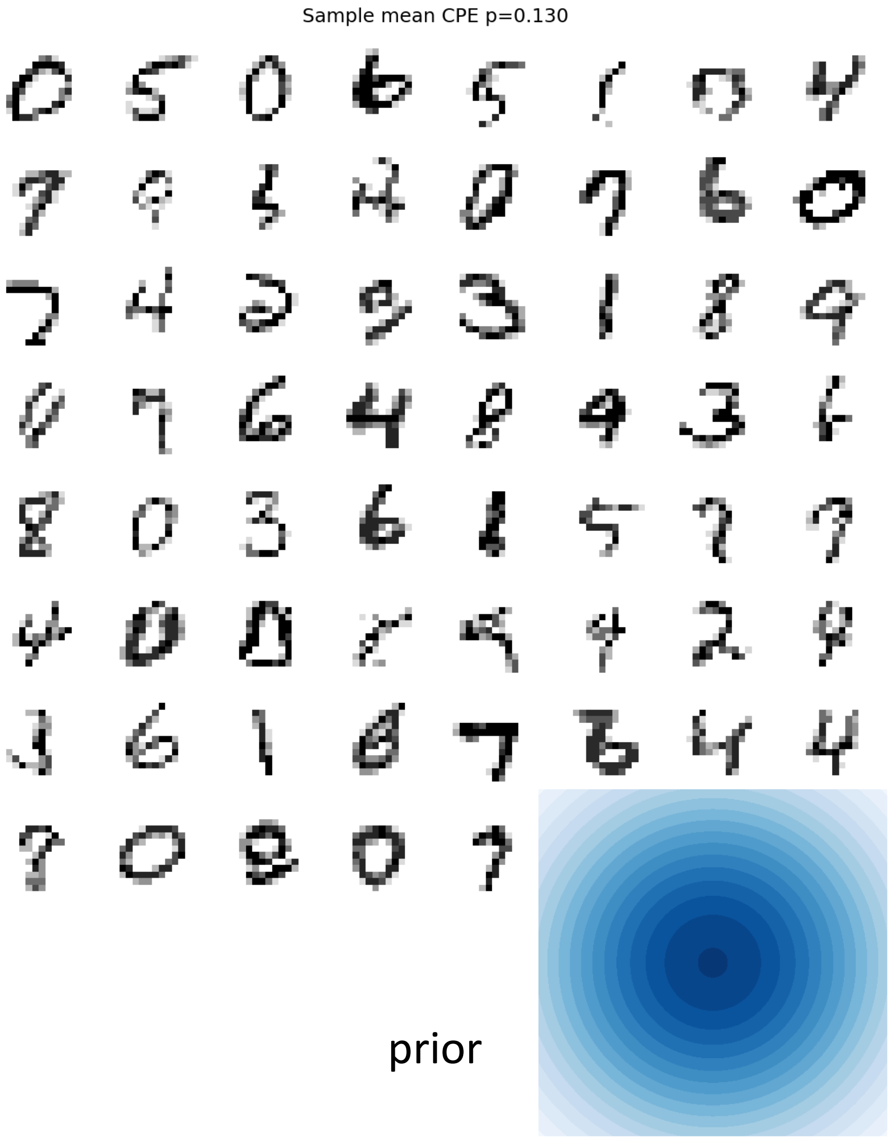
$$\mathbb{D}[p(\mathbf{x}|y > \tau_t, \mathcal{D}_t) \| p(\mathbf{x}|y > \tau_t, f_\bullet)] \in \mathcal{O}_\mathbb{P}(t^{-1/2}) \,. \tag{69}$$

- We discover (in expectation) the true fit sequences (hits H)

Corollary 1. *Under the settings in Theorem 2, we also have that:*

$$\mathbb{E}[|H_T - H_T^*|] \in \mathcal{O}(\sqrt{T}) \,.$$

Generating unrolled images



prior

posterior({3,5})

Active generation - desiderata

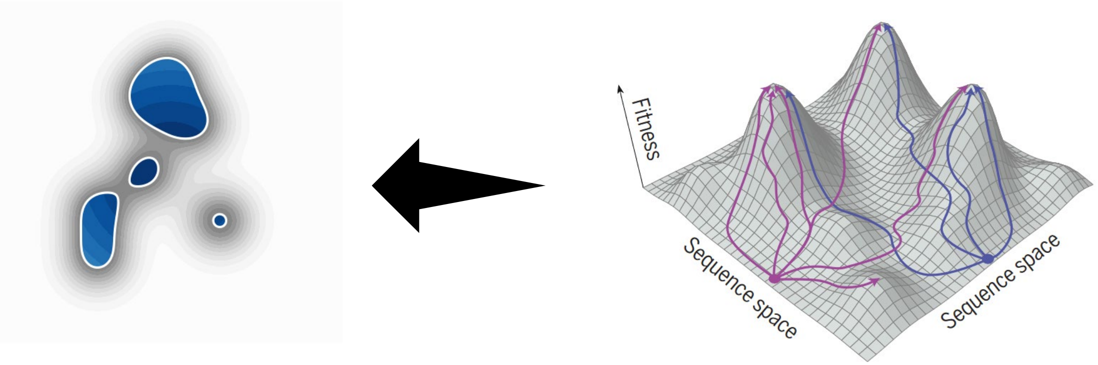
Requirements

- R1 – Rare feasible designs
- R2 – Sequential non-myopic candidate generation
- R3 – Discrete search
- R4 – Batch generation of diverse candidates
- R5 – Generative models

Desiderata

- D1 – Guaranteed convergence
- D2 – Gradient based optimisation
- D3 – Scalable predictive models

Contributions



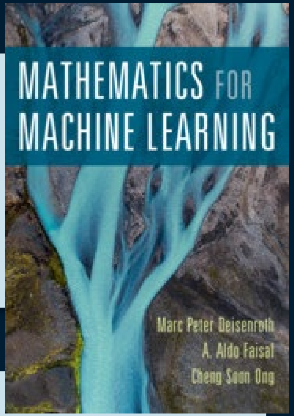
1. Requirements and desiderata for finding islands of fitness
2. Batch active generation objective over a (practically) innumerable discrete design space as an instance of variational inference
3. A modular algorithm, VSD, which solves this objective
4. Theoretical bounds, generalisation to multi-objective setting
5. VSD works in practice ☺

Steinberg, Oliveira, Ong, Bonilla, Variational Search Distributions, ICLR 2025
https://arxiv.org/abs/2409.06142
Steinberg, Wijesinghe, Oliveira, Koniusz, Ong, Bonilla, Amortized Active Generation of Pareto Sets, NeurIPS 2025
https://arxiv.org/abs/2510.21052

https://ong-home.my

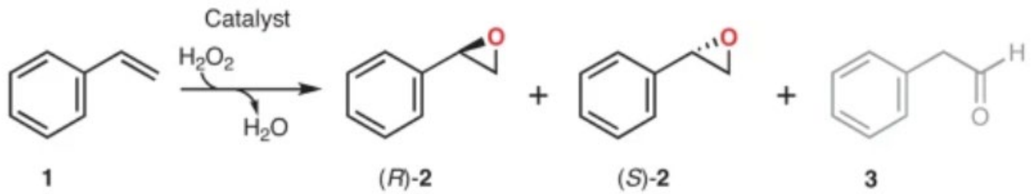
@cheng@masto.ai
@ml4x.bsky.social
chengsoon.ong@anu.edu.au

Ways to use a binary classifier

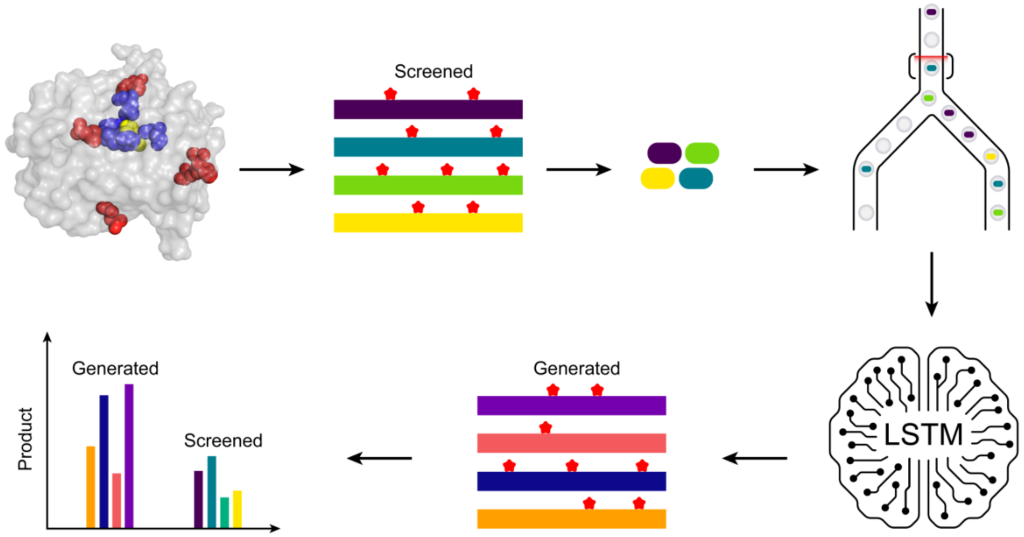


- Class probability estimation to density ratios
- Density ratio to Bayesian optimisation
- Searching the space of densities

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

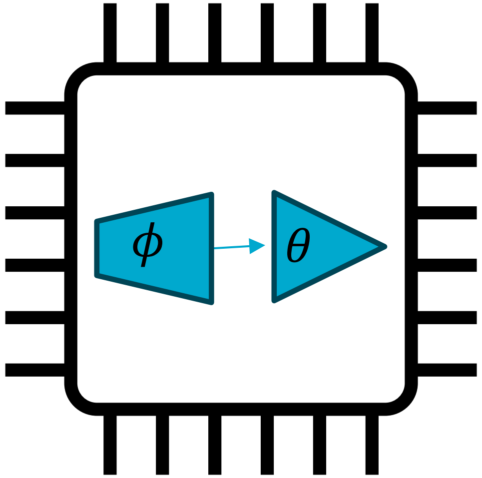
allozymes accelerating the evolution

CSIRO

DATA 61

Active Generation

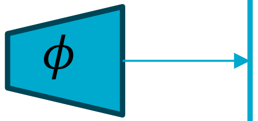
Active generation: find best ϕ^* for generating "good" (or best) **x**



Predictor + generator learning

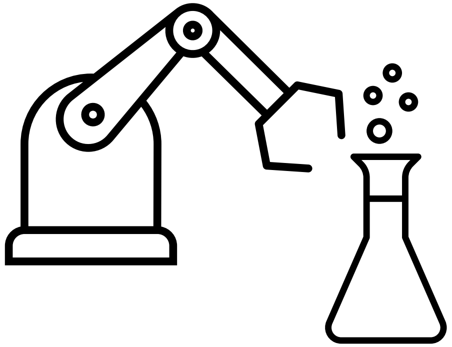
ϕ

Generate good candidates instead of selecting from a list



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

Unlabelled data *generation*, **design is generation**



Build + Test

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

Labelled Data

*Can "select" from a **vast number** of enzymes (e.g. 20^{289}), since they are **generated***

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

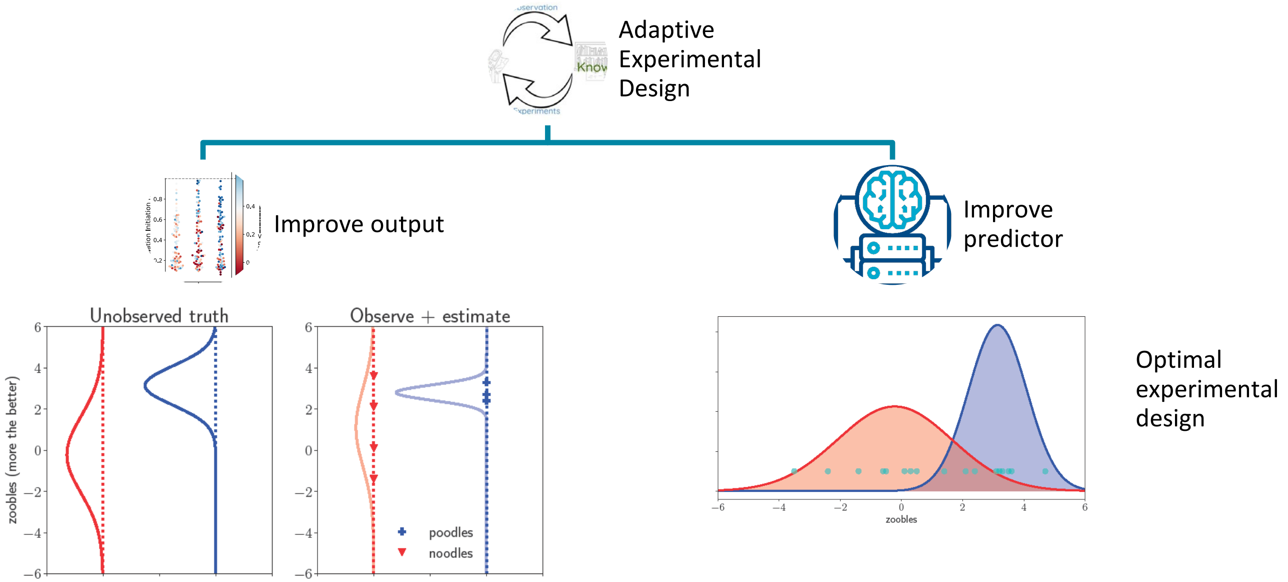
On finding good experiments

What is an experiment?

How do we find good experiments?

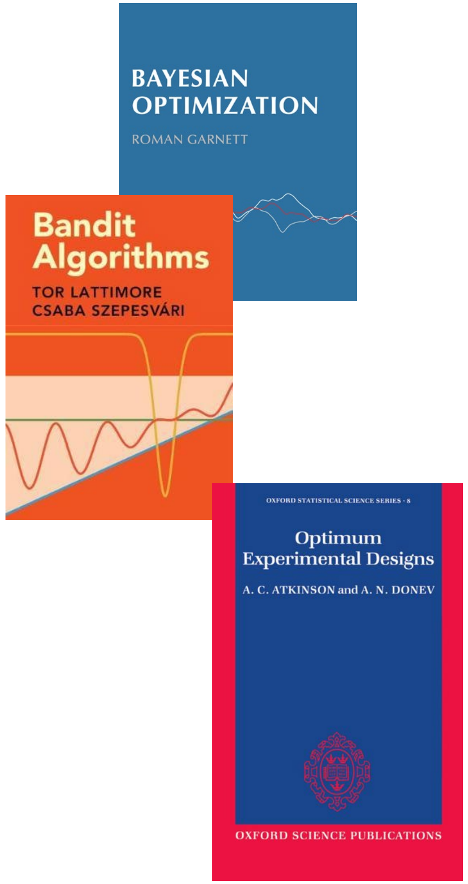
What do we mean by good?

Better measured values or better models

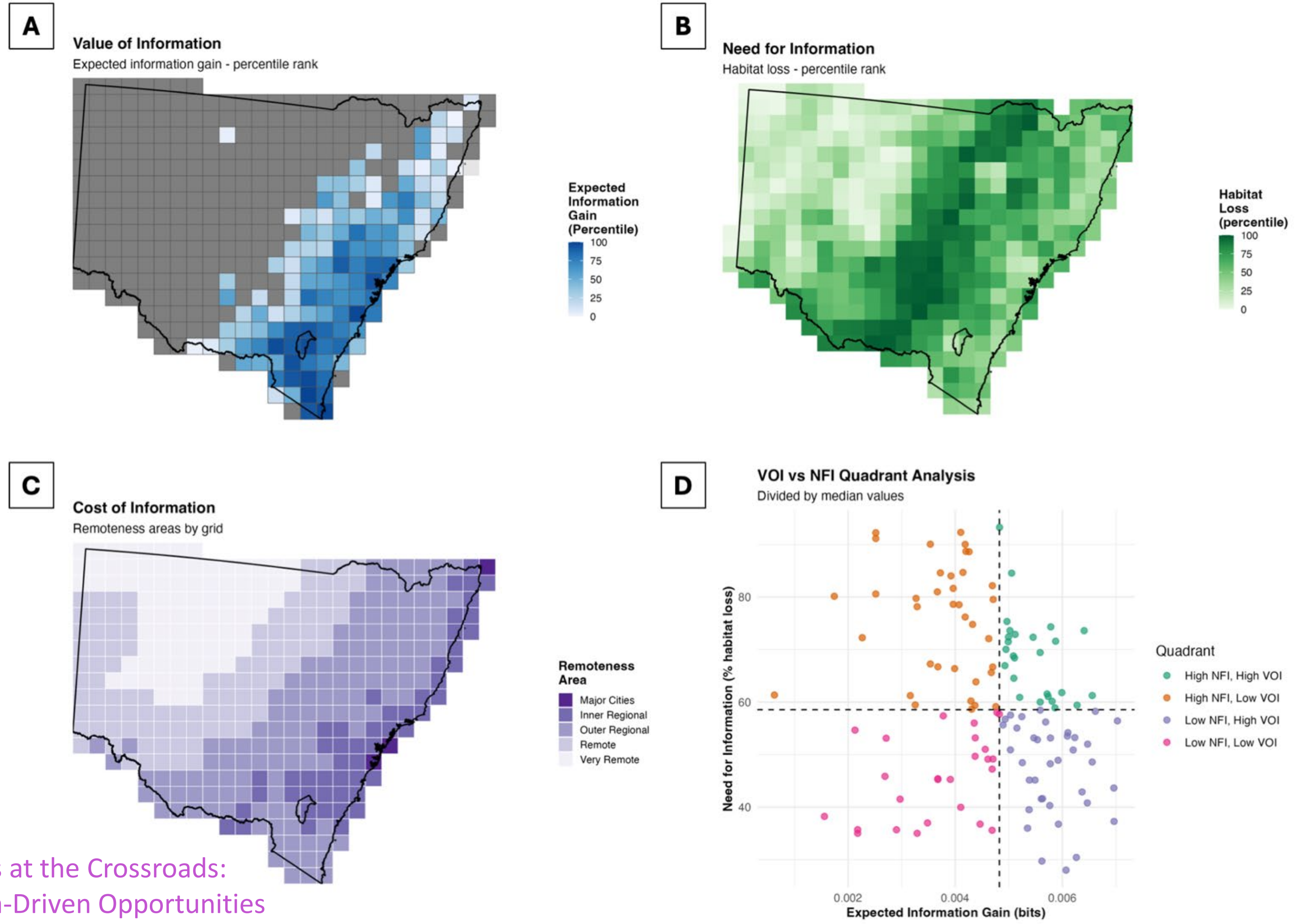


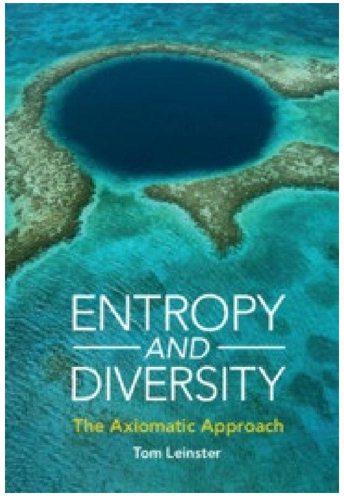
Recall that we can use a machine learning predictor in two ways:

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



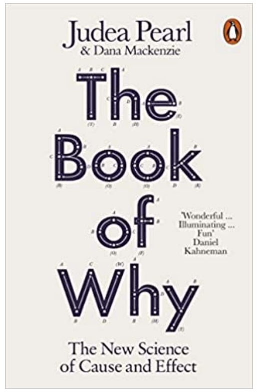
Value, need, cost



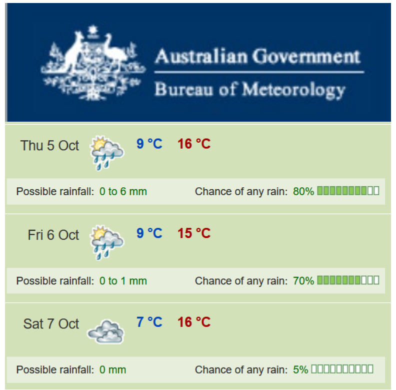


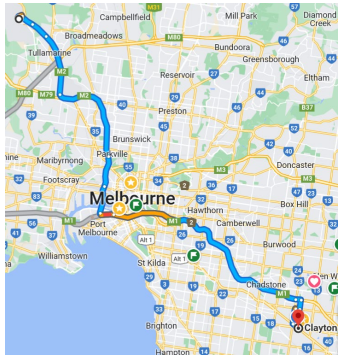
Natural History Collections at the Crossroads:
Shifting Priorities and Data-Driven Opportunities
Forbes, Thrall, Young, Ong, Ecology Letters, vol 28, no 8, 2025

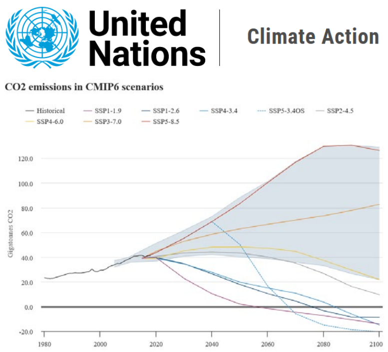
ML is **not only** about predictions



Predictions vs Decisions vs Actions







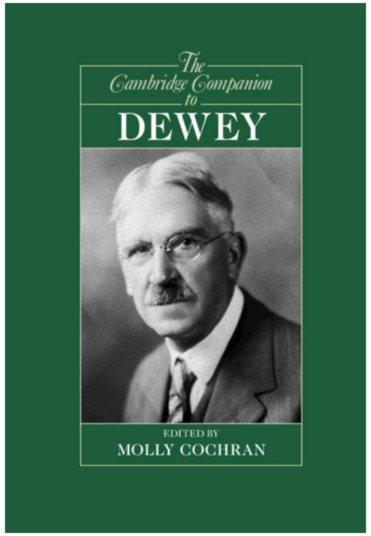
	Weather	Traffic	Climate
Predictions	Will it rain tomorrow?	Jam on M1?	Risk in 2050?
Decisions	Take umbrella?	Train or taxi?	Plan for net zero
Actions	Does not affect weather	Affects traffic!	Want wrong predictions!

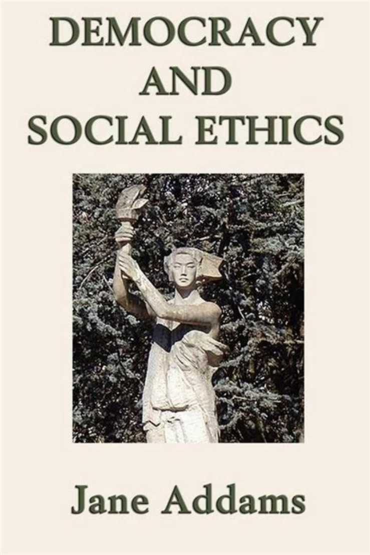
When accurate prediction models yield harmful self-fulfilling prophecies, Patterns, 2025
https://doi.org/10.1016/j.patter.2025.101229

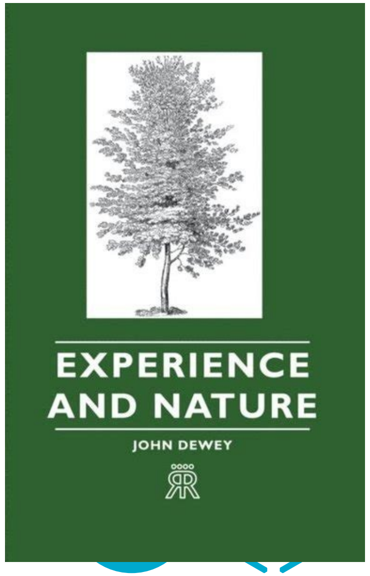
Pragmatism in International Relations

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 Chades, Melanie McGrath, Erin Bohensky, Lucy Carter, Rebecca Coates, Ben Harwood, Md Zahidul Islam, Sevvandi Kandanaarachchi, Cheng Soon Ong, Andrew Reeson, Samantha Stone-Jovicich, Cécile Paris, Mitchell Scovell, Kirsty Wissing, and David M. Douglas

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

Sarah Hartman [1] Cheng Soon Ong [2 3] Julia Powles [4 5] Petra Kuhnert [1]

Opportunities and Challenges in Designing Genomic Sequences

Mengyan Zhang [1 2] Cheng Soon Ong [2 1]