HUNGARIAN
MACHINE LEARNING DAYS

BUDAPEST, 12 – 14 AUGUST 2025

National Research, Development and Innovation Office
Hungary

PROGRAM
FINANCED FROM
THE NRDI FUND

Mecenatúra

PROCEEDINGS
HUNGARIAN MACHINE LEARNING DAYS 2025

The Hungarian Machine Learning Days provide an opportunity for Hungarian machine learning researchers working in foreign institutions to meet each other and connect with those working in Hungary, including the younger generation and Ph.D. students.

The organization and costs are covered by MEC_SZ 149796: Hungarian Machine Learning Meetup with the support of the Hungarian National Artificial Intelligence Laboratory, established in 2020, whose main mission is to unite domestic researchers, organize events (such as AI & AUT EXPO, HUN-REN AI Symposium), and build connections with European AI Centers and ELLIS.

During the 3-day summer event, internationally recognized researchers will present tutorials, young researchers will give short presentations, posters, and ideas, and there will be many informal joint programs. All presentations are in English.

Location

Budapesti Európai Ifjúsági Központ
(European Youth Centre Budapest)
Budapest, Zivatar u. 1, 1024
www.eycb.coe.int

National Research, Development and Innovation Office Hungary
PROGRAM FINANCED FROM THE NRDI FUND

CONTENTS

Conference Program

Day 1 — August 12

Regular Talks

- **10:00** — *Gergely Neu* (Universitat Pompeu Fabra)
  Inverse Q-Learning Done Right: Offline Imitation Learning in Q^π-Realizable MDPs
- **10:30** — *Balázs Csáji* (HUN-REN SZTAKI)
  Robust Inference with Kernels
- **11:00** — *Long Tran-Thanh* (University of Warwick)
  Pruning Neural Networks in a Principled Way
- **11:30** — *Mihály Petreczky* (CNRS, École Centrale Lille, University of Lille, CRIStAL)
  Statistical Guarantees for Learning Dynamical Systems
- **12:00** — *Anna Kerekes* (ETH Zürich; Max Planck Institute for Intelligent Systems, Tübingen)
  Machine Learning Meets Microbiology: Challenges and Opportunities

---

Afternoon Session

- **14:00** — **Poster Booster Session** (2 minutes, 2 slides each)
- **16:00–18:00** — Poster Session

Day 2 — August 13

Regular Talks

- **10:00** — *Gábor Csányi* (University of Cambridge)
  ML Force Fields Show Extreme Generalisation
- **10:30** — *Tamás K. Stenczel* (University of Cambridge)
  Optimal Transport for Atom Assignment in Materials Chemistry
- **11:00** — *Gergely Flamich* (University of Cambridge)
  You Cannot Feed Two Birds With One Score: The Accuracy–Naturalness Tradeoff in Translation
- **11:30** — *Ádám Zsolt Wagner* (Google DeepMind)
  Finding Interesting Mathematical Objects with ML

National Research, Development and Innovation Office
Hungary
PROGRAM FINANCED FROM THE NRDI FUND

- **12:00** — *Attila Csordás* (AgeCurve UK)

  Non-standard Attack on the Riemann Hypothesis with Sub-standard AI 'Students'

Student Talks

- **14:00** — *Csaba Botos* (University of Oxford)

  Compute-Constrained Solutions for the Challenges of Delayed Feedback

- **14:30** — *Tamás Levente* (Technical University of Cluj-Napoca)

  3D Point Cloud Processing on Edge

Panel Discussion

- **15:00** — **Latest Trends at NeurIPS / ICML / ICLR**

  Panelists: Gábor Csányi, Mihály Petreczky, Gergely Flamich, Csaba Botos, Patrik Reizinger

- **16:00–18:00** — Poster Session

Day 3 — August 14

Regular Talks

- **10:00** — *Botond Szabó* (Bocconi University)

  Privacy-Constrained Semi-parametric Inference

- **10:30** — *Tamás Linder* (Queen's University)

  Communication Complexity of Exact Sampling under Rényi Information

- **11:00** — *Csaba Beleznai* (AIT Austria)

  Robot Perception from Geometric Cues

- **11:30** — *Attila Sárkány* (Charles University; Czech Academy of Sciences)

  Quantile Preferences in Portfolio Choice: A Q-DRL Approach to Dynamic Diversification

- **12:00** — *Patrik Reizinger* (Max Planck Institute for Intelligent Systems)

  Identifiable Exchangeable Mechanisms for Causal Structure and Representation Learning

---

Short / Student Talks

- **14:00** — *Bálint Horváth* (HUN-REN SZTAKI)

  Kernel-Based Image Restoration with Uncertainty Guarantees

National Research, Development and Innovation Office Hungary
PROGRAM FINANCED FROM THE NRDI FUND

Mecenatúra

- **14:20** — *András Balogh* (University of Szeged)
  How Not to Stitch Representations to Measure Similarity: Task Loss Matching Versus Direct Matching (AAAI 2025)
- **14:40** — *Miranda Anna Christ* (Fazekas Mihály High School; Alfréd Rényi Institute of Mathematics)
  The Structure of Relation Decoding Linear Operators in Large Language Models

---

Panel Discussion

- **15:00** — **AI Regulation**
  Panelists: Márton Domokos, Ferenc Kása, Long Tran-Thanh, Csaba Beleznai

---

National Research, Development and Innovation Office Hungary
PROGRAM FINANCED FROM THE NRDI FUND

Inverse Q-learning done right

Gergely Neu

joint work with
Antoine Moulin (UPF) & Luca Viano (EPFL)

Universitat Pompeu Fabra Barcelona

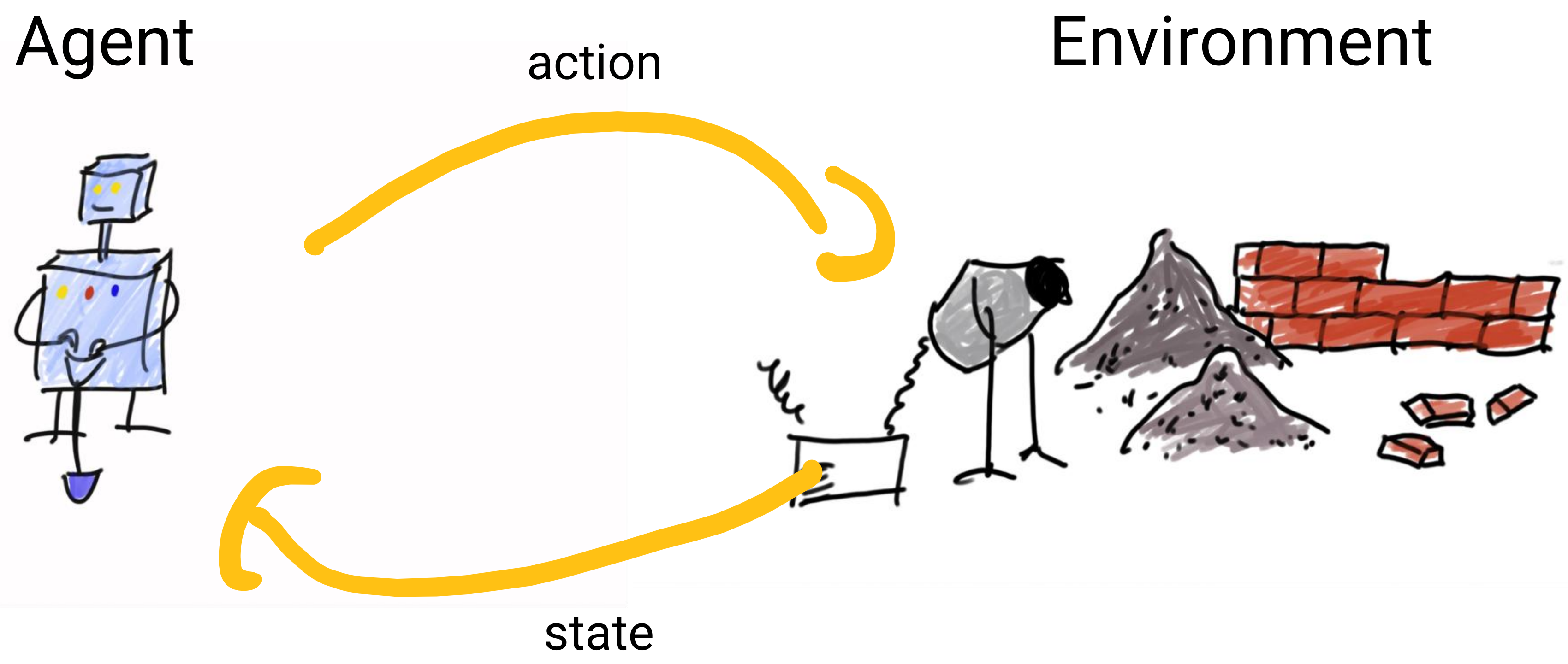
Outline

- Reinforcement learning & imitation learning
- The tools of the trade:
  - value functions
  - occupancy measures
- Saddle-point imitation learning
- Theory
- Experiments

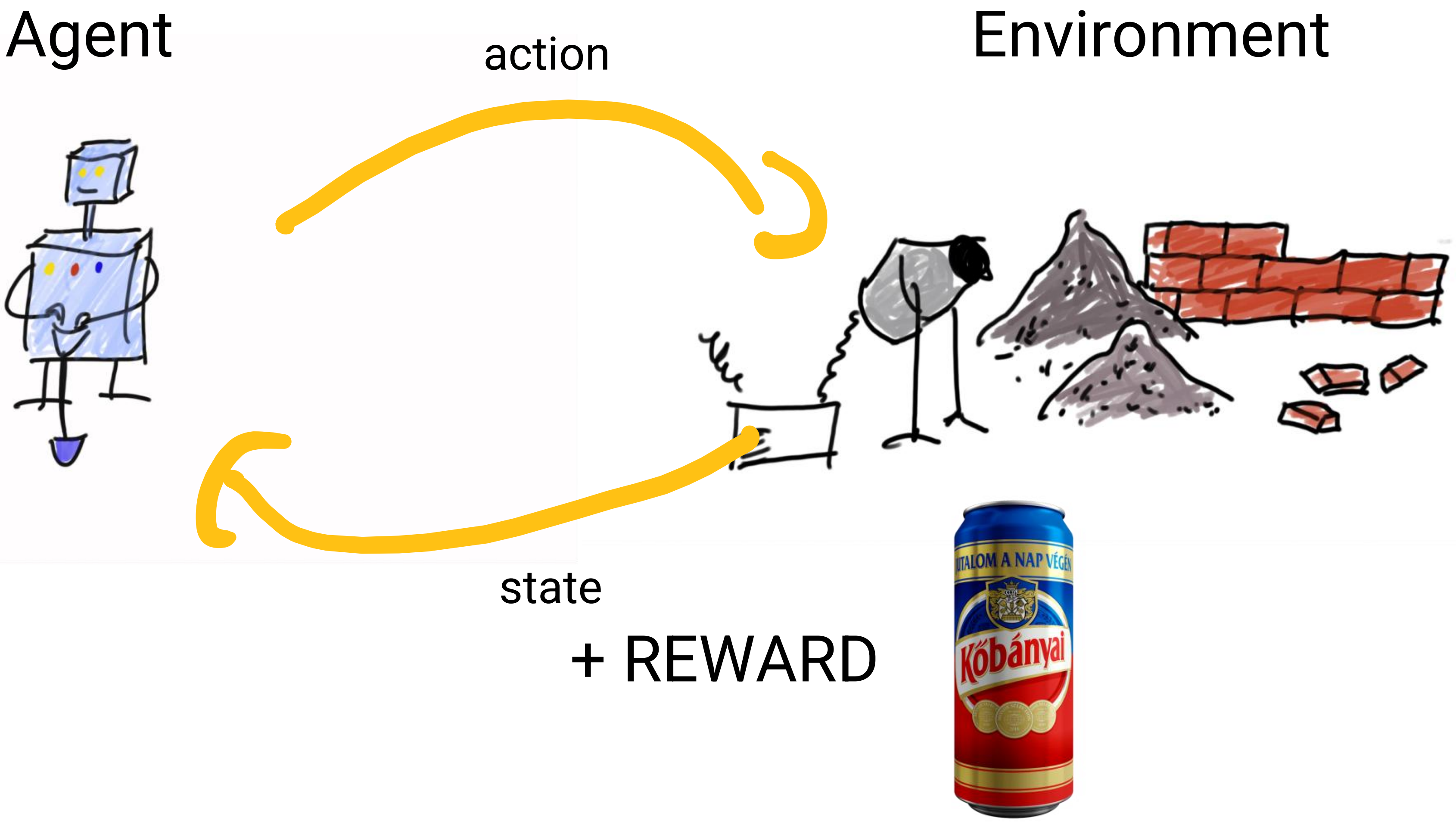
Outline

- Reinforcement learning & imitation learning
- The tools of the trade:
  - value functions
  - occupancy measures
- Saddle-point imitation learning
- Theory
- Experiments

Sequential decision making

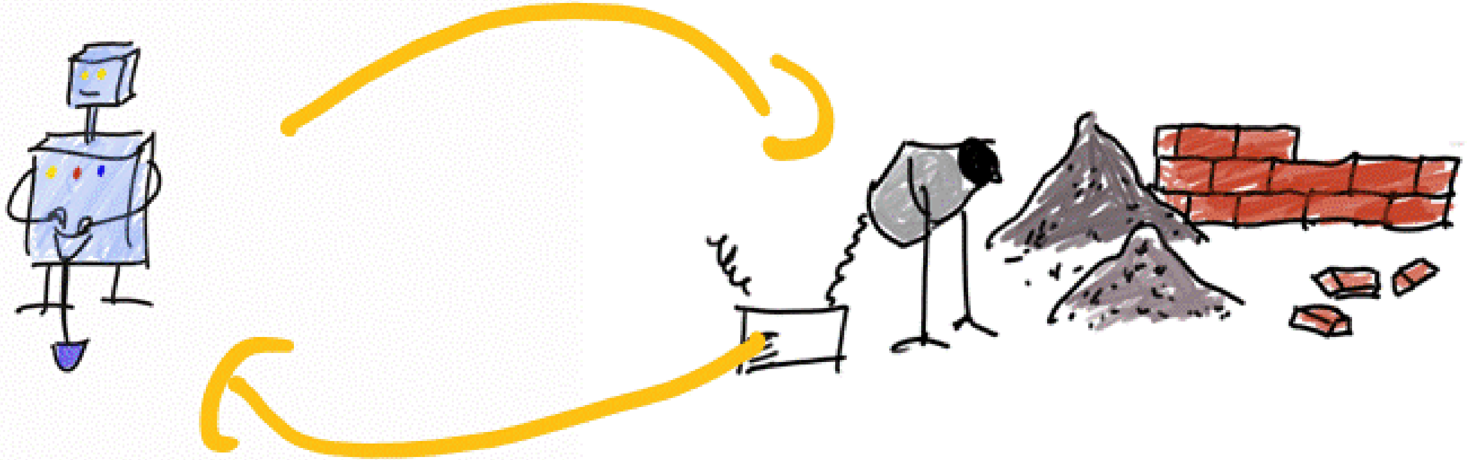


Sequential decision making



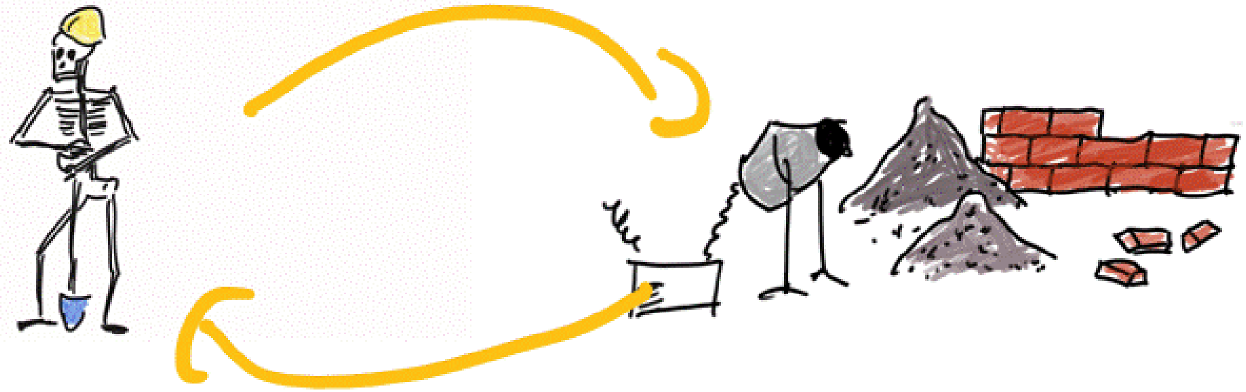
Reinforcement learning

Given access to the environment and the reward function, train an agent to learn a desirable behavior
(i.e., earn a lot of reward)

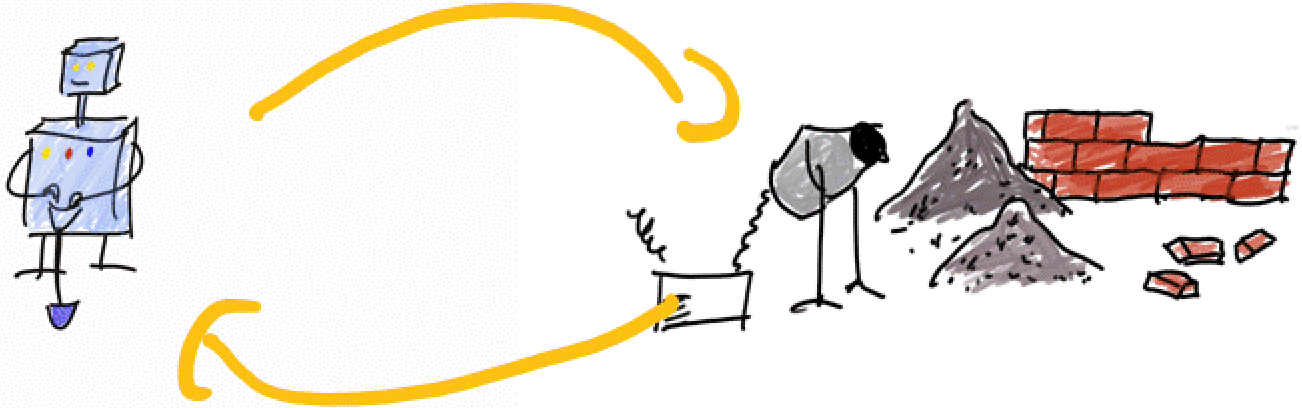


Imitation learning

Given observations of an expert performing a desired behavior…
(and no reward function)



… train an agent to replicate the same behavior
(earn as much reward as the expert)



The two paradigms

Reinforcement learning

- Learning directly from experience, without supervision
- Needs GOOD reward function
- Needs LOTS of interaction with the environment

Imitation learning

- Learning from demonstrations of useful behavior
- Needs NO reward function
- Needs NO interaction with the environment

The two paradigms

Reinforcement learning

- Learning directly from experience, without supervision
- Needs GOOD reward function
- Needs LOTS of interaction with the environment

Takes FOREVER to learn anything useful when rewards are sparse
(“Jutalom a nap végén”)

Imitation learning

- Learning from demonstrations of useful behavior
- Needs NO reward function
- Needs NO interaction with the environment

The two paradigms

Reinforcement learning

- Learning directly from experience, without supervision
- Needs GOOD reward function
- Needs LOTS of interaction with the environment

Takes FOREVER to learn anything useful when rewards are sparse
("Jutalom a nap végén")

Imitation learning

- Learning from demonstrations of useful behavior
- Needs NO reward function
- Needs NO interaction with the environment

Can learn much faster when good observational data is available

Two radical approaches to IL

- Behavioral cloning (Pomerleau, 1991):
  - “try to directly learn the mapping from states to actions, ignoring the temporal structure of the problem”
  - Possibly inefficient due to using “too little” structure
- Inverse reinforcement learning (Ng & Russell, 2005):
  - “try to recover the reward function from the expert data first, and then learn an optimal policy”
  - Possibly inefficient due to using “too much” structure

Two radical approaches to IL

- Behavioral cloning (Pomerleau, 1991):
  - "try to directly learn the mapping from states to actions, ignoring the temporal structure of the problem"
  - Possibly inefficient due to using "too little" structure
- Inverse reinforcement learning (Ng & Russell, 2005):
  - "try to recover the reward function from the expert data first, and then learn an optimal policy"
  - Possibly inefficient due to using "too much" structure

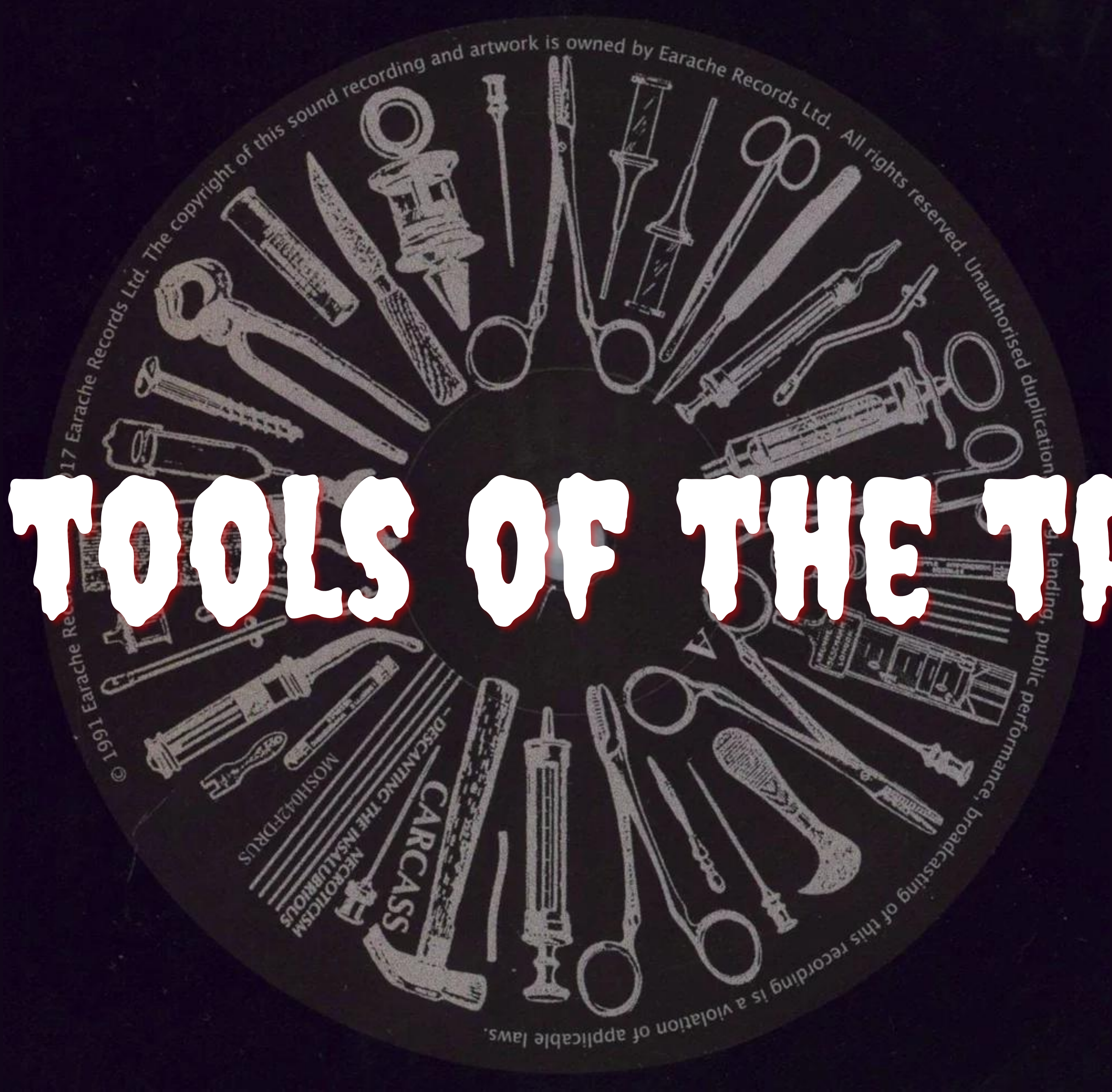
What we'll do:

Try to use "the right" amount of structure and learn a good policy without recovering the reward function

Outline

- Reinforcement learning & imitation learning
- The tools of the trade:
  - value functions
  - occupancy measures
- Saddle-point imitation learning
- Theory
- Experiments

THE TOOLS OF THE TRADE

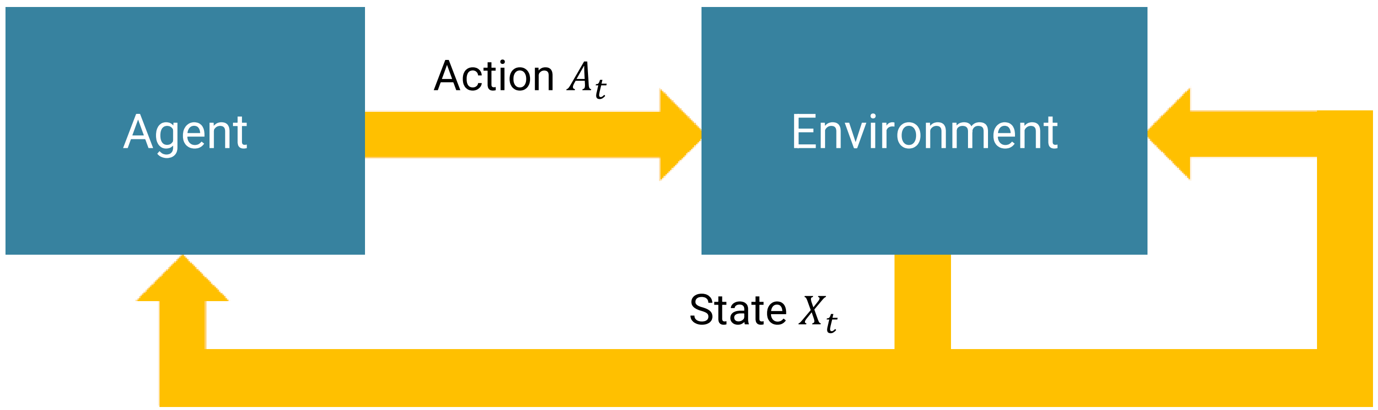


Markov decision processes

Initial state is drawn as $X_0 \sim p_0$

In each round $t = 0,1,2, \ldots$

- Agent observes state X_t
- Agent takes action A_t
- Agent earns reward $R_t = r(X_t, A_t)$
- Environment generates next state $X_{t+1} \sim P(\cdot\, | X_t, A_t)$

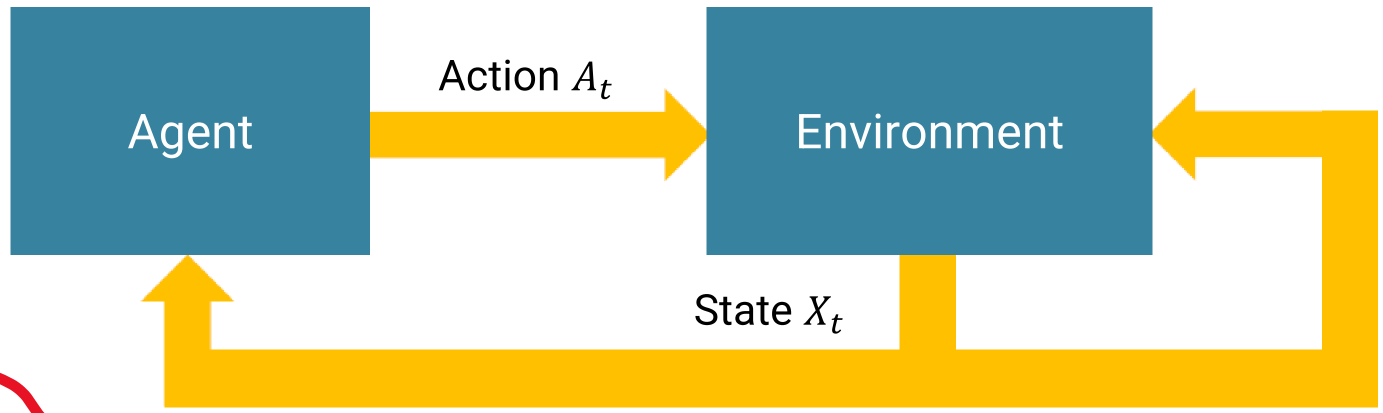


Markov decision processes

Initial state is drawn as $X_0 \sim p_0$

In each round $t = 0,1,2,\ldots$

- Agent observes state X_t
- Agent takes action A_t
- Agent earns reward $R_t = r(X_t, A_t)$
- Environment generates next state $X_{t+1} \sim P(\cdot \mid X_t, A_t)$



Reward function
$r: \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

Transition function
$P: \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$

Policies and their values

- A stochastic policy is a mapping $\pi: \mathcal{X} \to \Delta_{\mathcal{A}}$
- An agent follows policy π if it takes its actions as $A_t \sim \pi(\cdot \mid X_t)$
- The value of a policy π is the total discounted sum of rewards of an agent following π:

$$\rho^\pi = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]$$

- In RL, we typically seek an optimal policy π^* that satisfies

$$\pi^* = \arg\max_\pi \rho^\pi$$

Action-value functions

- The action-value function (or Q-function) of a policy π is
$$Q^{\pi}(x,a) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t R_t \,|X_0 = x, A_0 = a]$$
- Also define the notation $f(x,\pi') = \sum_a \pi'(a|x) f(x,a)$ for any f

Action-value functions

- The action-value function (or Q-function) of a policy π is
$$Q^\pi(x, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_t \,|X_0 = x, A_0 = a]$$
- Also define the notation $f(x, \pi') = \sum_a \pi'(a|x) f(x, a)$ for any f
- Then, the value ρ^π can be written as

$$\rho^\pi = \mathbb{E}_{X_0 \sim p_0}[Q^\pi(X_0, \pi)]$$

Occupancy measures

- The occupancy measure of a policy π is the distribution

$$\mu^\pi(x,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi[X_t = x, A_t = a]$$

Occupancy measures

- The occupancy measure of a policy π is the distribution

$$\mu^\pi(x,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_\pi[X_t = x, A_t = a]$$

- Then, the value ρ^π can be written as

$$\rho^\pi = \mathbb{E}_{X,A\sim\mu^\pi}[r(X,A)]$$

The performance difference lemma

- The two different expressions can be combined to obtain the following ✨ extremely useful ✨ expression:

Lemma

(Howard, 1960, Kakade & Langford, 2002)

For any two policies π and π', we have

$$\rho^{\pi} - \rho^{\pi'} = \mathbb{E}_{X,A\sim\mu^{\pi}}\left[Q^{\pi'}(X,A) - Q^{\pi'}(X,\pi')\right]$$

Outline

- Reinforcement learning & imitation learning
- The tools of the trade:
  - value functions
  - occupancy measures
- Saddle-point imitation learning
- Theory
- Experiments

What we have & what we want

- **HAVE:** n samples generated by some expert policy π_E
- **DON'T WANT:**
  - Recover the expert policy π_E exactly (~ behavioral cloning)
  - Recover the reward function r (~ inverse RL)
- **WANT:** a policy π_{out} satisfying $\rho^{\pi_{\text{out}}} \geq \rho^{\pi_E} - \varepsilon$ for small ε

What we have & what we want

- **HAVE:** n samples generated by some expert policy π_E
- **DON'T WANT:**
  - Recover the expert policy π_E exactly (~ behavioral cloning)
  - Recover the reward function r (~ inverse RL)
- **WANT:** a policy π_{out} satisfying $\rho^{\pi_{\text{out}}} \geq \rho^{\pi_E} - \varepsilon$ for small ε

What we have & what we want

- **HAVE:** n samples generated by some expert policy π_E
- **DON'T WANT:**
  - Recover the expert policy π_E exactly (~ behavioral cloning)
  - Recover the reward function r (~ inverse RL)
- **WANT:** a policy π_{out} satisfying $\rho^{\pi_{\text{out}}} \geq \rho^{\pi_E} - \varepsilon$ for small ε

What we have & what we want

- **HAVE:** n samples generated by some expert policy π_E
- **DON'T WANT:**
  - Recover the expert policy π_E exactly (~ behavioral cloning)
  - Recover the reward function r (~ inverse RL)
- **WANT:** a policy π_{out} satisfying $\rho^{\pi_{\text{out}}} \geq \rho^{\pi_E} - \varepsilon$ for small ε

Still challenging since we don't have access to the environment or the reward function!

Getting there in 3 steps

1. Use performance difference lemma to write
$$\rho^{\pi_E} - \rho^{\pi} = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q^{\pi}(X,A) - Q^{\pi}(X,\pi)]$$
2. Define the objective function
$$\mathcal{L}(\pi;Q) = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q(X,A) - Q(X,\pi)]$$
and notice that
$$\rho^{\pi_E} - \rho^{\pi} = \mathcal{L}(\pi;Q^{\pi}) \leq \sup_Q \mathcal{L}(\pi;Q)$$
3. Find a policy π_{out} such that $\sup_Q \mathcal{L}(\pi_{\text{out}};Q) \leq \varepsilon$

Getting there in 3 steps

1. Use performance difference lemma to write
$$\rho^{\pi_E} - \rho^{\pi} = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q^{\pi}(X,A) - Q^{\pi}(X,\pi)]$$
2. Define the objective function
$$\mathcal{L}(\pi;Q) = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q(X,A) - Q(X,\pi)]$$
and notice that
$$\rho^{\pi_E} - \rho^{\pi} = \mathcal{L}(\pi;Q^{\pi}) \leq \sup_Q \mathcal{L}(\pi;Q)$$
3. Find a policy π_{out} such that $\sup_Q \mathcal{L}(\pi_{\text{out}};Q) \leq \varepsilon$

Getting there in 3 steps

1. Use performance difference lemma to write
$$\rho^{\pi_E} - \rho^{\pi} = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q^{\pi}(X,A) - Q^{\pi}(X,\pi)]$$
2. Define the objective function
$$\mathcal{L}(\pi; Q) = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q(X,A) - Q(X,\pi)]$$
and notice that
$$\rho^{\pi_E} - \rho^{\pi} = \mathcal{L}(\pi; Q^{\pi}) \le \sup_Q \mathcal{L}(\pi; Q)$$
3. Find a policy π_{out} such that $\sup_Q \mathcal{L}(\pi_{\text{out}}; Q) \le \varepsilon$

Getting there in 3 steps

1. Use performance difference lemma to write
$$\rho^{\pi_E} - \rho^{\pi} = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q^{\pi}(X,A) - Q^{\pi}(X,\pi)]$$
2. Define the objective function
$$\mathcal{L}(\pi;Q) = \mathbb{E}_{X,A\sim\mu^{\pi_E}}[Q(X,A) - Q(X,\pi)]$$
and notice that
$$\rho^{\pi_E} - \rho^{\pi} = \mathcal{L}(\pi;Q^{\pi}) \leq \sup_{Q} \mathcal{L}(\pi;Q)$$
3. Find a policy π_{out} such that $\sup_{Q} \mathcal{L}(\pi_{\text{out}};Q) \leq \varepsilon$

Bottom line: find the saddle point
$$\min_{\pi} \max_{Q} \mathcal{L}(\pi;Q)$$

Saddle-point Imitation Learning

- **Primal-dual scheme:** for each $k = 1,2, \dots, K$
  - Update π_{k+1} incrementally by policy mirror descent on $\mathcal{L}(\pi_k; Q_k)$
  - Update Q_{k+1} by best-responding π_{k+1}: $Q_{k+1} = \arg\max_Q \mathcal{L}(\pi_{k+1}; Q)$
- **Output:** $\pi_{\text{out}} = \pi_I$ for I chosen uniformly over $\{1,2, \dots, K\}$
- **Guarantee:** The total loss can be bounded as

$$\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(\pi_k; Q^{\pi_k}) \leq \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(\pi_k; Q_k),$$

which is exactly what the policy updates are aiming to minimize!

Saddle-point Imitation Learning

- **Primal-dual scheme:** for each $k = 1,2,\dots,K$
  - Update π_{k+1} incrementally by policy mirror descent on $\mathcal{L}(\pi_k; Q_k)$
  - Update Q_{k+1} by best-responding π_{k+1}: $Q_{k+1} = \arg\max_Q \mathcal{L}(\pi_{k+1}; Q)$
- **Output:** $\pi_{\text{out}} = \pi_I$ for I chosen uniformly over $\{1,2,\dots,K\}$
- **Guarantee:** The total loss can be bounded as

$$\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(\pi_k; Q^{\pi_k}) \le \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(\pi_k; Q_k),$$

which is exactly what the policy updates are aiming to minimize!

Saddle-point Imitation Learning

- **Primal-dual scheme:** for each $k = 1,2,\dots,K$
  - Update π_{k+1} incrementally by policy mirror descent on $\mathcal{L}(\pi_k; Q_k)$
  - Update Q_{k+1} by best-responding π_{k+1}: $Q_{k+1} = \arg\max_Q \mathcal{L}(\pi_{k+1}; Q)$
- **Output:** $\pi_{\text{out}} = \pi_I$ for I chosen uniformly over $\{1,2,\dots,K\}$
- **Guarantee:** The total loss can be bounded as

$$\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(\pi_k; Q^{\pi_k}) \le \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(\pi_k; Q_k),$$

  which is exactly what the policy updates are aiming to minimize!

A practical algorithm

SPOIL (Saddle-Point Offline Imitation Learning)

Input: Expert data set $\{(X_t, A_t)\}_{t=1}^n$, a parametric class of Q-functions $\mathcal{Q} = \{Q_\theta : \theta \in \mathbb{R}^d\}$, learning rate $\eta > 0$
Initialize: $\theta_0 = 0$, $\pi_0 =$ uniform policy
For $k = 1,2 \dots, K$:

- $\pi_k(a|x) \propto \pi_{k-1}(a|x) \exp\left(-\eta Q_{k-1}(x,a)\right)$
- $Q_k = \arg\max_{Q \in \mathcal{Q}} \hat{\mathcal{L}}_n(\pi_k; Q_k)$

Output: $\pi_{out} = \pi_I$ for $I \sim \mathcal{U}(K)$

A practical algorithm

SPOIL (Saddle-Point Offline Imitation Learning)

Input: Expert data set $\{(X_t, A_t)\}_{t=1}^n$, a parametric class of Q-functions $\mathcal{Q} = \{Q_\theta : \theta \in \mathbb{R}^d\}$, learning rate $\eta > 0$

Initialize: $\theta_0 = 0$, $\pi_0 =$ uniform policy

For $k = 1,2 \ldots, K$:

- $\pi_k(a|x) \propto \pi_{k-1}(a|x) \exp\left(-\eta Q_{k-1}(x,a)\right)$
- $Q_k = \arg\max_{Q \in \mathcal{Q}} \hat{\mathcal{L}}_n(\pi_k; Q_k)$

Output: $\pi_{out} = \pi_I$ for $I \sim \mathcal{U}(K)$

Empirical estimate of $\mathcal{L}$:

$$\hat{\mathcal{L}}_n(\pi; Q) = \frac{1}{n}\sum_{t=1}^n \left(Q(X_t, A_t) - Q(X_t, \pi)\right)$$

Outline

- Reinforcement learning & imitation learning
- The tools of the trade:
  - value functions
  - occupancy measures
- Saddle-point imitation learning
- Theory
- Experiments

Q^π-realizability

Definition

The Q-function class $\mathcal{Q}$ is called Q^π-realizable if for all policies π, we have $Q^\pi \in \mathcal{Q}$.

Q^π-realizability

Definition

The Q-function class $\mathcal{Q}$ is called Q^π-realizable if for all policies π, we have $Q^\pi \in \mathcal{Q}$.

- Satisfied if the parametrization of the Q-functions used by SPOIL is powerful enough
- Much weaker than other conditions common in RL theory, e.g., Bellman completeness (Antos, Munos, Szepesvári, 2008) or Linear MDP condition (Jin, Yang, Wang, Jordan, 2020)

An error bound

Proposition

Let $\Delta(\pi) = \mathbb{E}\left[\sup_{Q \in \mathcal{Q}} \left| \mathcal{L}(\pi; Q) - \hat{\mathcal{L}}_n(\pi; Q) \right| \right]$. The output of SPOIL satisfies

$$\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] \leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\mathcal{L}(\pi_k; Q_k)] + \frac{2}{K} \sum_{k=1}^{K} \mathbb{E}[\Delta(\pi_k)].$$

Proof: The same algebra as before, just adding and subtracting $\mathcal{L} - \hat{\mathcal{L}}_n$ a couple of times. ■

An error bound

Proposition

Let $\Delta(\pi) = \mathbb{E}\left[\sup_{Q \in \mathcal{Q}} \left|\mathcal{L}(\pi; Q) - \hat{\mathcal{L}}_n(\pi; Q)\right|\right]$. The output of SPOIL satisfies

$$\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] \leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\mathcal{L}(\pi_k; Q_k)] + \frac{2}{K} \sum_{k=1}^{K} \mathbb{E}[\Delta(\pi_k)].$$

Proof: The same algebra as before, just adding and subtracting $\mathcal{L} - \hat{\mathcal{L}}_n$ a couple of times. ■

COOL!! $\Delta(\pi)$ can be bounded via a standard uniform convergence argument!! (For a fixed Q, the error is literally the difference between a mean and an empirical mean)

Sample complexity guarantees

Theorem 1

Suppose $\mathcal{Q}$ is a class of d-dimensional linear Q^{π}-realizable functions. Then, the output of SPOIL guarantees $\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] \leq \varepsilon$ after processing $n = \tilde{O}(d/(1-\gamma)^2\varepsilon^2)$ samples.

Theorem 2

Suppose $\mathcal{Q}$ is a general class of Q^{π}-realizable functions. Then, the output of SPOIL guarantees $\mathbb{E}[\rho^{\pi_E} - \rho^{\pi_{\text{out}}}] \leq \varepsilon$ after processing $n = \tilde{O}(\mathcal{N}_{\varepsilon}/(1-\gamma)^4\varepsilon^4)$ samples.

* $\mathcal{N}_{\varepsilon}$ is the ℓ_∞ covering number of the class at some level $\text{poly}(1/\varepsilon)$

What's "done right about" this though?

- "Inverse Q-learning" (IQL) by Garg et al. (NeurIPS'21) is also derived from approximating $\min_{\pi} \max_{Q} \mathcal{L}(\pi; Q)$
- IQL maximizes the dual objective $\mathcal{D}(Q) = \min_{\pi} \mathcal{L}(\pi; Q)$ by SGD
- This makes it impossible to translate the optimization error into a guarantee on the quality of the output policy
- In contrast, our method minimizes the primal objective $\mathcal{P}(\pi) = \max_{Q} \mathcal{L}(\pi; Q)$ via stochastic mirror descent
- As we have seen, this comes with a straightforward translation between the quantities of interest

What's "done right about" this though?

- "Inverse Q-learning" (IQL) by Garg et al. (NeurIPS'21) is also derived from approximating $\min_{\pi} \max_{Q} \mathcal{L}(\pi; Q)$
- IQL maximizes the dual objective $\mathcal{D}(Q) = \min_{\pi} \mathcal{L}(\pi; Q)$ by SGD
- This makes it impossible to translate the optimization error into a guarantee on the quality of the output policy
- In contrast, our method minimizes the primal objective $\mathcal{P}(\pi) = \max_{Q} \mathcal{L}(\pi; Q)$ via stochastic mirror descent
- As we have seen, this comes with a straightforward translation between the quantities of interest

Outline

- Reinforcement learning & imitation learning
- The tools of the trade:
  - value functions
  - occupancy measures
- Saddle-point imitation learning
- Theory
- Experiments

Can SPOIL beat behavioral cloning?

- Behavioral cloning (Pomerleau, 1991, Foster, Block, Misra , 2024):
  - Consider a class of policies Π (say, parametrized by NNs)
  - Find output policy by maximum likelihood:

$$\pi_{\text{out}} = \text{argmax}_{\pi \in \Pi} \sum_{t=1}^{n} \log \pi(A_t | X_t)$$

  - Ignores structure, but obtains good guarantees when $\pi_E \in \Pi$

Can SPOIL beat behavioral cloning?

- Behavioral cloning (Pomerleau, 1991, Foster, Block, Misra , 2024):
  - Consider a class of policies Π (say, parametrized by NNs)
  - Find output policy by maximum likelihood:

$$\pi_{\text{out}} = \text{argmax}_{\pi \in \Pi} \sum_{t=1}^{n} \log \pi(A_t | X_t)$$

  - Ignores structure, but obtains good guarantees when $\pi_E \in \Pi$

What can go wrong?

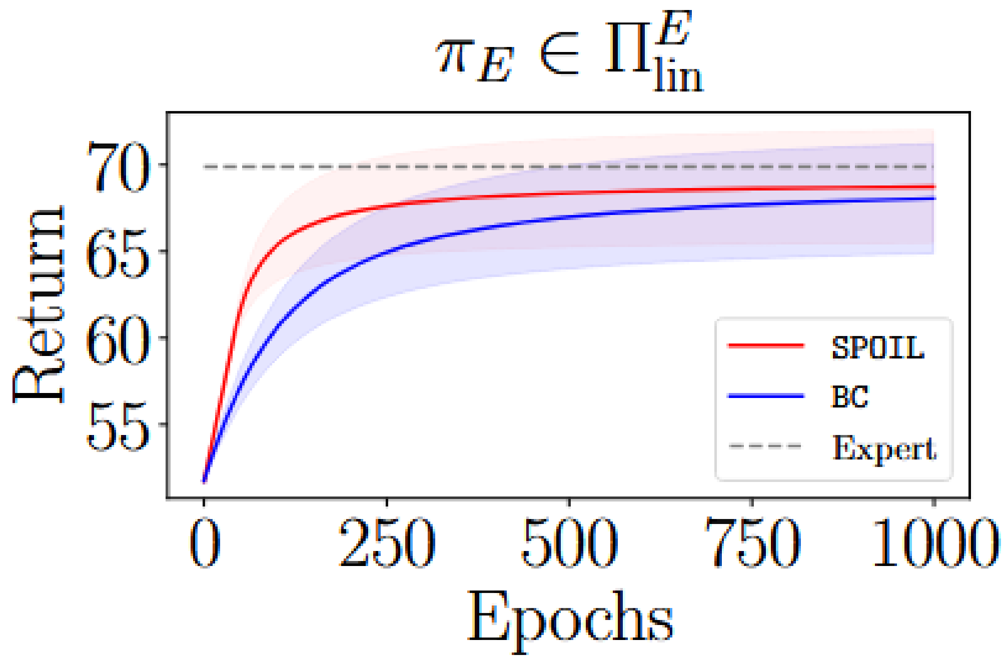
- π_E may not be in the class Π
- Π may be too complex for optimization to be effective
- …

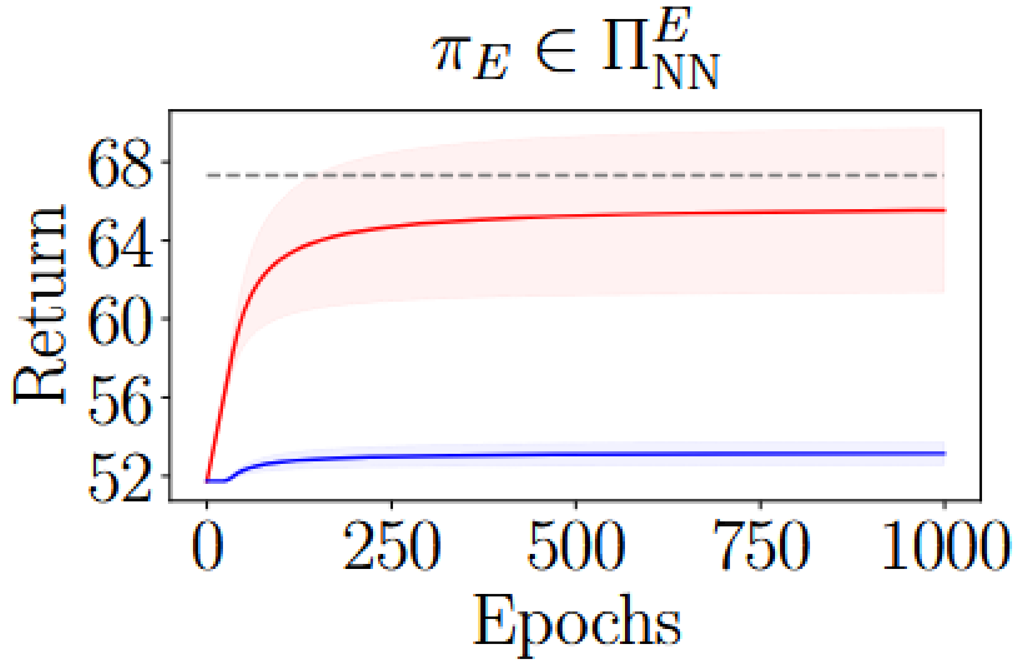
A neat little experiment

- Build an MDP with a "simple" optimal policy π^* (linear classifier with some small-dimensional parameter vector θ^*)
- Train two "experts" that are both near-optimal: one linear softmax (Π_{lin}^E) and one using a 3-layer NN + softmax (Π_{NN}^E)

A neat little experiment

- Build an MDP with a “simple” optimal policy π^* (linear classifier with some small-dimensional parameter vector θ^*)
- Train two “experts” that are both near-optimal: one linear softmax (Π^E_{lin}) and one using a 3-layer NN + softmax (Π^E_{NN})

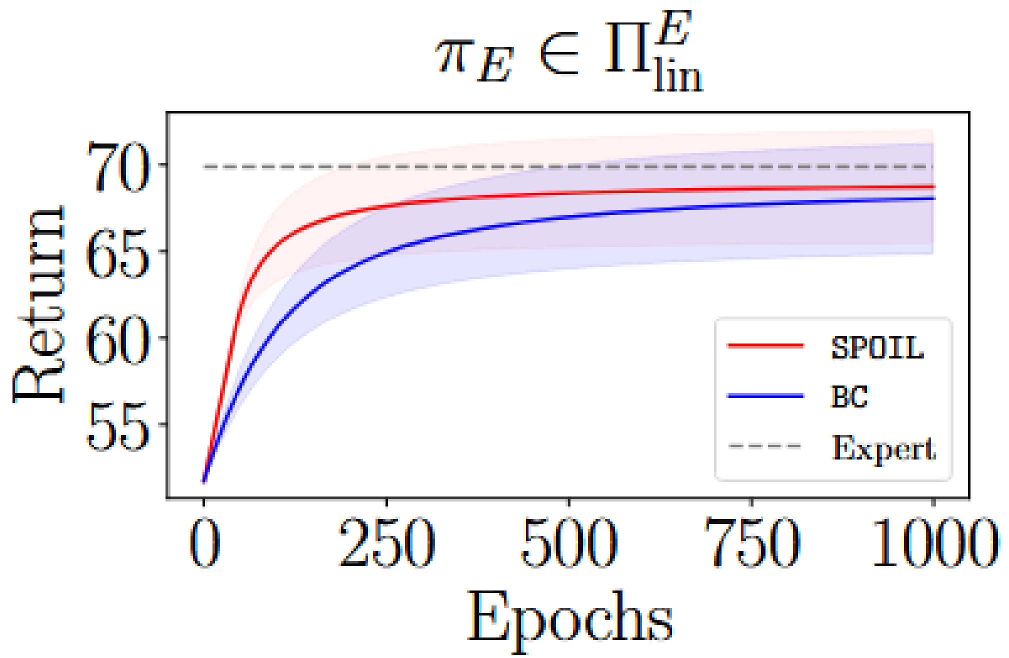


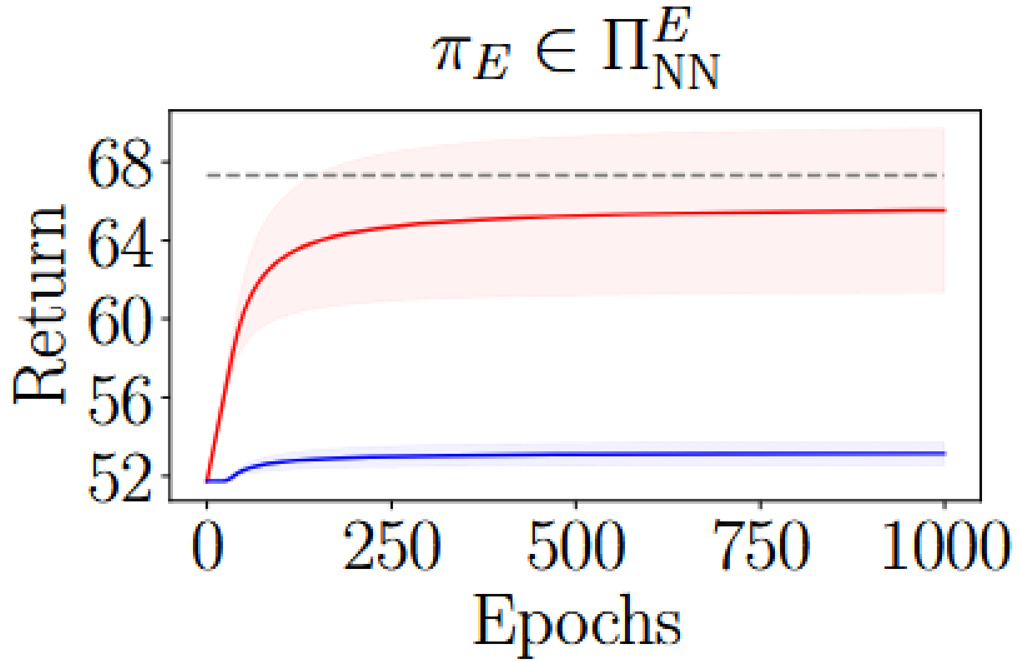


A neat little experiment

- Build an MDP with a “simple” optimal policy π^* (linear classifier with some sma…
- Train tw… …r softma… …Π^{E}_{NN})

BC really struggles when compared to SPOIL when policy class is complex!





Conclusion

- Imitation learning is a cool problem
- SPOIL is a cool algorithm
  - strong theoretical foundations & guarantees
  - simple to understand & easy to implement*
- More broadly: min-max optimization perspective can bring about very useful algorithms when used correctly

Conclusion

- Imitation learning is a cool problem
- SPOIL is a cool algorithm
  - strong theoretical foundations & guarantees
  - simple to understand & easy to implement*
- More broadly: min-max optimization perspective can bring about very useful algorithms when used correctly
- Many open questions:
  - *scalability of policy update?
  - beyond Q^π-realizability?
  - other uses of this unusual policy eval objective? (see also Neu & Okolo, 2025)

Thank you!!

Time for that “reward”

Robust Inference with Kernels

Balázs Csanád Csáji

Institute for Computer Science and Control (SZTAKI), Budapest, Hungary
Institute of Mathematics, Eötvös Loránd University (ELTE), Budapest

Joint work with Marco C. Campi, Erik Weyer, Algo Carè, Ambrus Tamás,
Dániel Á. Bálint, Szabolcs Szentpéteri, Bálint Horváth, Krisztián B. Kis

Hungarian Machine Learning Days, 12-14 August, 2025, Budapest

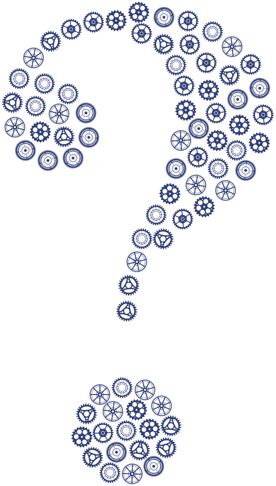
I. Introduction

"*Uncertainty is the only certainty there is, and knowing how to live with insecurity is the only security*." (John Allen Paulos)

Uncertainty is Everywhere

Uncertainty is an inherent part of most activities, e.g., in natural and social sciences, engineering, industry, finance, economics, and medicine.

- Limited knowledge and data scarcity
- Measurement errors
- Information transmission problems
- Modeling biases
- Approximation errors
- Computational constraints
- Trust and security concerns
- Intrinsic variability of the systems



Robust Uncertainty Quantification

- To guarantee the reliability of solutions, we should rigorously assess their uncertainty, enabling more trustworthy methods.
- Quantifying model uncertainty is essential. It is a prerequisite of uncertainty quantification for prediction and control methods.
- Focus: stochastic models (though there are worst-case setups).
- In statistics uncertainty quantification (UQ) is done by building confidence regions. Random set $\mathcal{C}_\alpha(X)$ is an α-level conf. region if

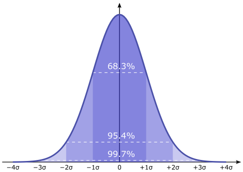
$$\forall \mathbb{P} \in \mathcal{P} : \forall f_* \in \mathcal{F} : \mathbb{P}\big(\, f_* \in \mathcal{C}_\alpha(X) \,\big) \geq 1 - \alpha$$

  for some risk probability $\alpha \in (0, 1)$, where f_* is a target model, $\mathcal{F}$ is a model class, $\mathcal{P}$ is a family of distributions, and X is a sample.
- Robust UQ: minimizing the statistical and structural assumptions.
- Needed for robust decisions, risk management, active learning, etc.

Standard Confidence Region Constructions

(1) It is assumed that the distribution family of the data is known and then we can use the quantiles of a pivotal quantity.

It is often unrealistic the assume a parametric family of distributions in practice.



(2) Another typical approach is to build on limiting distributions

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\theta_*))$$

leading to approximate, only asymptotically guaranteed regions.

(3) The use of concentration inequalities is also widespread

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X_1] \right| \geq \varepsilon \right) \leq 2 \exp\left(- \frac{n\,\varepsilon^2}{2\sigma^2} \right)$$

leading to a priori bounds with finite sample guarantees.

II. Resampling, Ranking & Reproducing Kernels

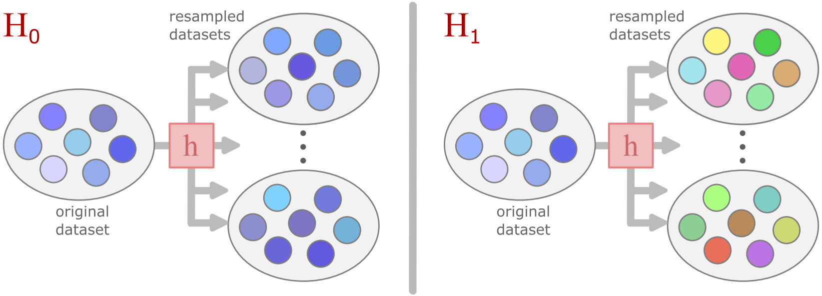
Bird's-Eye View of a Robust Inference Framework

Robust Inference Framework

- Aim: to obtain a general, efficient robust inference framework.
- A flexible approach can be built using resampling and ranking.
- Both are important techniques of (nonparametric) statistics.
- They become very powerful if combined with kernel methods.
- Core approach: resampling-and-ranking based distribution-free kernel methods which can perform non-asymptotic inference.
- Essential questions to answer are:
  1. How to resample the data?
  2. What data to resample?
  3. How to rank the resampled datasets?
  4. How to avoid pointwise testing?
  5. How to obtain efficiently acionable structures?

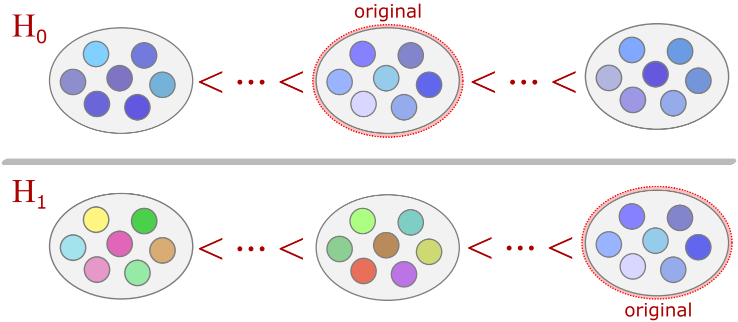
Resampling for Testing Goodness-of-Fit

- Resampling is widespread in statistics, including Monte Carlo tests, bootstrap, jackknife, permutation tests, and cross validation, etc.
- For us, its core idea can be best understood as a hypothesis test.
- Under H_0, we can generate new datasets "similar" to the original.
- However, under H_1 the generated samples will be "different".



Ranking for Detecting Inhomogeneity

- A key question is how to decide whether the new datasets are "similar" to the original one. A statical solution: ranking.
- Under H_0, the rank of the original dataset should be "average", unlikely to be high (typically: discrete uniform distribution).
- However, under H_1, the original dataset should tend be the largest.



How to Resample?

– Resampling does not need strong distributional assumption on the observed data (such as Gaussianity). A sufficient property is:

An $\mathbb{R}^n$-valued random vector ε is distributionally invariant w.r.t. a compact group of transformations, $(\mathcal{G}, \circ)$, where "$\circ$" denotes the function composition and each $G \in \mathcal{G}$ maps $\mathbb{R}^n$ to itself, if for all $G \in \mathcal{G}$, vectors ε and $G(\varepsilon)$ have the same distribution.

– Two arch-typical examples having this property are

(1) If $\{\varepsilon_i\}$ are exchangeable (for example: i.i.d.), then we can use the (finite) group of permutations on the noise vector.

(2) If $\{\varepsilon_i\}$ independent and symmetric, then we can apply the group consisting sign-changes for any subsets of the noises.

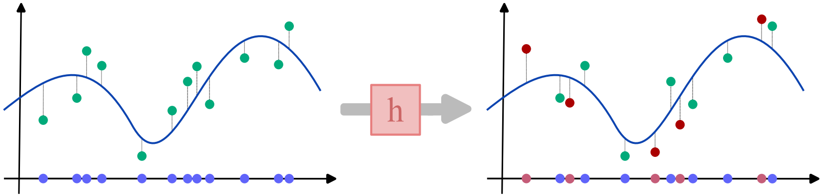
– Then, $\{G_k\}$ can be an i.i.d. sample, $\forall G_k \sim$ uniform on group $\mathcal{G}$.

What to Resample?

– In regression problems, the natural approach is to resample the residuals. If $\{(X_j, Y_j)\}$ is the dateset and f is a model to test

$$\mathcal{D}_k \doteq (X, f(X) + G_k(f(X) - Y)),$$

where $f(X) = (f(X_1), \dots, f(X_n))^{\mathrm{T}}$ and $Y = (Y_1, \dots, Y_n)^{\mathrm{T}}$, assuming the true error is distributionally invariant w.r.t. $\mathcal{G}$.



What to Resample?

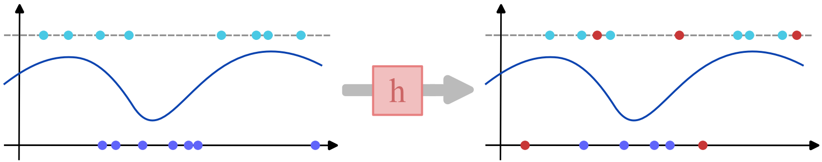
– In binary classification problems, we resample the class labels.

– Key observation: a candidate classifier $f : \mathcal{X} \to \{0, 1\}$ encodes the conditional distribution of the labels for each input, that is

$$f(X_k) \doteq 2 \cdot \mathbb{P}_f(\, Y_k = +1 \mid X_k\,) - 1.$$

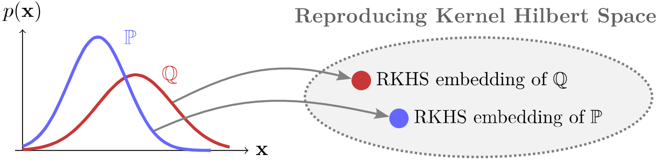
Then, given f one can generate $\{Y_{k,j}(f)\}$ from the conditional distribution $\mathbb{P}_f(\, Y \mid X = X_j\,)$, producing alternative samples:

$$\mathcal{D}_k \doteq ((X_1, Y_{k,1}(f)), \dots, (X_n, Y_{k,n}(f))).$$



How to Rank?

- If we have an underlying ML method which optimizes a loss function, a method for ranking is to use the norm of its gradient.
- In linear regression problems with the least squares criterion, this leads to the nonasymptotic Sign-Perturbed Sums (SPS) method.
- SPS has exact coverage guarantees, it is strongly consistent and have efficient ellipsoidal outer approximations (cf. actionability).
- Otherwise, embedding the empirical (conditional) distributions into a Reproducing Kernel Hilbert Space is powerful direction.



III. Independence Tests for Stochastic Processes

by Resampling, Ranking and Reproducing Kernels

Linear Stochastic Systems

- Consider two scalar (discrete-time, time-invariant) synchronous (SISO) linear stochastic systems with general dynamics:

$$Y_t = G_1(q^{-1}; \theta_*) U_t + H_1(q^{-1}; \theta_*) E_t,$$

$$Z_t = G_2(q^{-1}; \gamma_*) V_t + H_2(q^{-1}; \gamma_*) N_t.$$

  (examples include ARMAX, Box-Jenkins and output-error models)
- $\{U_t\}$ and $\{V_t\}$ are (exogenous) inputs
- $\{Y_t\}$ and $\{Z_t\}$ are (observable) outputs
- q^{-1} is the backward shift (lag) operator
- $\{E_t\}$ and $\{N_t\}$ are possibly dependent process noises
- G_1, H_1, and G_2, H_2 are (rational, causal, monic) transfer functions
- θ_* and γ_* are the true finite dimensional parameters

Independence Tests

- Goal: to test whether processes $\{Y_t\}$ and $\{Z_t\}$ are independent.
- Alternatively (straightforward to generalize): to test whether they are conditionally independent given the exogenous inputs.
- Note that even though the two stochastic processes are "linear", however, the noises can have any type of nonlinear dependence.
- We seek a hypothesis test with distribution-free and finite sample guarantees for any given (user-chosen) significance level α.
- We also want to find sufficient conditions for consistency, i.e., we want to prove that the type II error probability tends to zero.
- Main idea: to generalize dependency measures (for i.i.d. data) using non-asymptotic confidence regions and permutation tests.

Key Assumptions

(A1) The true systems generating the observed outputs $\{Y_t\}$ and $\{Z_t\}$ are in the model classes, that is, $\theta_* \in \Theta$ and $\gamma_* \in \Gamma$.

(A2) The transfer functions G_1, G_2, H_1 and H_2 have known orders.

(A3) Filters H_1 and H_2 are invertible for all $\theta \in \Theta$ and $\gamma \in \Gamma$.

(A4) Both systems are initialized in zero, formally, we have $Y_t = Z_t = U_t = V_t = E_t = N_t = 0$, for all $t \leq 0$.

(A5) The systems are driven by a jointly i.i.d. innovation sequence $\{(E_t, N_t)\}_{t=1}^{\infty}$ from the (unknown) distribution of (E, N).

(A6) The systems operate in open-loop: the inputs $\{U_t\}$ and $\{V_t\}$ are independent of each other and the noises $\{E_t\}$ and $\{N_t\}$.

Special Case: Known Parameters

- Under (A1)–(A6), processes $\{Y_t\}$ and $\{Z_t\}$ are independent if and only if E and N are independent, hence it is sufficient to consider

$$H_0 : Q_{E,N} = Q_E \otimes Q_N$$

$$H_1 : Q_{E,N} \neq Q_E \otimes Q_N$$

  where $Q_{E,N}$ is the joint distribution of E, N with marginals Q_E, Q_N.
- If θ_* and γ_* are known, then we can compute the innovations by

$$E_t \;=\; E_t(\theta^*) = \; H_1^{-1}(q^{-1};\theta_*)(Y_t - G_1(q^{-1};\theta_*)U_t),$$

$$N_t \;=\; N_t(\gamma_*) = \; H_2^{-1}(q^{-1};\gamma_*)(Z_t - G_1(q^{-1};\gamma_*)V_t).$$

- The independence of variables E and N can be tested based on the resulting i.i.d. sample, $\mathcal{D}_0 = \{(E_t, N_t)\}_{t=1}^{n}$.

Resampling by Random Permutations

- We construct (resample) $m-1$ alternative samples by

$$\mathcal{D}_j = \pi_j \mathcal{D}_0 = \{(E_i, N_{\pi_j(i)})\}_{i=1}^{n}, \qquad \text{for} \quad j = 1, \dots, m-1,$$

where $\{\pi_j\}_{j=1}^{m-1}$ are permutations uniformly randomly generated from S_n.

- S_n is the symmetric group of degree n: it contains all permutations of the set $\{1, \dots, n\}$; a permutation is selected with probability $1/(n!)$

Key observations

- If H_0 holds, then samples $\{\mathcal{D}_j\}_{j=0}^{m-1}$ are exchangeable.
- If H_1 holds, then $\mathcal{D}_0$ and $\mathcal{D}_j$ have different distributions for $j \neq 0$.

Ranking Functions

– Let $\mathbb{A}$ be a measurable space, a function $\psi : \mathbb{A}^m \to [\,m\,]$, where $[\,m\,] \doteq \{1, \dots, m\}$, is called a ranking function if for $\forall (a_1, \dots, a_m) \in \mathbb{A}^m$ it satisfies the following two properties:

(P1) For all permutations μ of the set $\{2, \dots, m\}$, we have

$$\psi(\,a_1, a_2, \dots, a_m\,) \;=\; \psi(\,a_1, a_{\mu(2)}, \dots, a_{\mu(m)}\,),$$

that is the function is invariant with respect to reordering the last $m-1$ terms of its arguments.

(P2) For all $i, j \in [\,m\,]$, if $a_i \neq a_j$, then we have

$$\psi(\,a_i, \{a_k\}_{k \neq i}\,) \;\neq\; \psi(\,a_j, \{a_k\}_{k \neq j}\,).$$

Uniform Ordering of Exchangeable Elements

Reminder: Resampling and Ranking Framework

- Resample the original dataset to obtain alternative samples randomly generated according to a given hypothesis we are testing.
- Measure "similar" behavior with ranking.
- Accept the null hypothesis if the original dataset behaves "similarly" to the alternatively generated ones and reject it otherwise.

Uniform Ordering Lemma

Let $A_1, \dots, A_m$ be exchangeable, almost surely pairwise different random elements taking values in $\mathbb{A}$. Then, $\psi\big(A_1, A_2, \dots, A_m \big)$ has discrete uniform distribution on $\{1, \dots, m\}$.

Exact Hypothesis Tests

- Assume we are given a ranking function ψ (satisfying $P1$ and $P2$).
- Let $r, p \in [\, m \,]$ with $r \leq p$ be user-chosen hyper-parameters.
- We accept hypothesis H_0 if and only if $r \leq \psi\big(\mathcal{D}_0, \{\mathcal{D}_k\}_{k \neq 0}\big) \leq p$.

Theorem (Exact Type I Error Probability)

If H_0 holds true, then we have for all ranking function ψ and hyper-parameter $\varrho = (m, r, p)$ with integers $1 \leq r \leq p \leq m$:

$$\mathbb{P}\big(\, r \leq \psi\big(\mathcal{D}_0, \{\mathcal{D}_k\}_{k \neq 0}\big) \leq p \,\big) = \frac{p - r + 1}{m}.$$

- The confidence level is user-chosen (rational), and exact.
- The resulting hypothesis test is distribution-free (independent of the underlying distribution) and non-asymptotic (holds for finite samples).

Dependence Measure: HSIC

Hilbert-Schmidt Independence Criterion (HSIC)

$$\mathrm{HSIC}(Q_{E,N}) = \mathbb{E}[k(E,E')\ell(N,N')]$$
$$+\,\mathbb{E}[k(E,E')]\,\mathbb{E}[\ell(N,N')] - 2\,\mathbb{E}[k(E,E')\ell(N,N'')]$$

Empirical HSIC Estimator (Consistent)

$$\mathrm{HSIC}_n(\mathcal{D}_0) \doteq \frac{1}{n^2}\sum_{(i,j)\in[n]^2} k(E_i,E_j)\ell(N_i,N_j)$$
$$+\frac{1}{n^4}\sum_{(i,j,r,s)\in[n]^4} k(E_i,E_j)\ell(N_r,N_s) - \frac{1}{n^3}\sum_{(i,j,s)\in[n]^3} k(E_i,E_j)\ell(N_i,N_s)$$

Dependence Measure: Distance Covariance

Distance Covariance (dCov)

$$\mathrm{dCov}^2(Q_{E,N}) \doteq \mathbb{E}\big[\, \|E - E'\| \cdot \|N - N'\| \,\big]$$
$$+\, \mathbb{E}\, \|E - E'\| \cdot \mathbb{E}\, \|N - N'\| - 2\, \mathbb{E}\big[\, \|E - E'\| \cdot \|N - N''\| \,\big]$$

Empirical Distance Covariance Estimator (Consistent)

$$\mathrm{dCov}^2_n(\mathcal{D}_0) \doteq \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} A_{j,k} B_{j,k},$$

$$A_{j,k} = a_{j,k} - a_{j\cdot} - a_{\cdot k} + a_{\cdot\cdot}, \text{ where } a_{j,k} = \|E_j - E_k\|,$$

$$a_{j\cdot} = \sum_{k=1}^{n} a_{j,k}/n,\; a_{\cdot k} = \sum_{j=1}^{n} a_{j,k}/n \text{ and } a_{\cdot\cdot} = \sum_{j,k=1}^{n} a_{j,k}/(n^2).$$

Consistency of the Independence Test

(A7) A characteristic dependence measure Δ and a consistent empirical estimator is given such that for $j \in [m-1]$,

$$\widehat{\Delta}_n^{(0)} \xrightarrow{p} |\Delta(E, N)| \qquad \text{and} \qquad \widehat{\Delta}_n^{(j)} \xrightarrow{p} 0.$$

Dependence Measure-based Ranking

$$\widehat{\Delta}_n^{(j)} = |\widehat{\Delta}(\mathcal{D}_j)| \qquad \text{for} \quad j = 0, 1, \dots, m-1 \ \text{ and}$$

$$\psi_\Delta(\mathcal{D}_0, \dots, \mathcal{D}_{m-1}) \doteq 1 + \sum\nolimits_{j=1}^{m-1} \mathbb{I}\left(\widehat{\Delta}_n^{(0)} \prec_\sigma \widehat{\Delta}_n^{(j)} \right)$$

Theorem (Consistency)

Assume (A1)–(A7) and that θ_*, γ_* are given. Under H_1 for $r \geq 1$

$$\mathbb{P}\left(\psi_\Delta(\mathcal{D}_0, \dots, \mathcal{D}_{m-1}) \leq r \right) \xrightarrow{n \to \infty} 1.$$

General Case: Unknown System Parameters

– Let us have confidence sets $\widehat{\Theta}_n$ and $\widehat{\Gamma}_n$ such that

$$\mathbb{P}\big(\,\theta_* \in \widehat{\Theta}_n\,\big) \geq 1 - \beta \qquad \text{and} \qquad \mathbb{P}\big(\,\gamma_* \in \widehat{\Gamma}_n\,\big) \geq 1 - \beta$$

hold for all $n \in \mathbb{N}$ and for a chosen significance level $\beta \in (0, 1)$.

– For any candidate pair $(\theta, \gamma) \in \widehat{\Theta} \times \widehat{\Gamma}$ let

$$E_t(\theta) \doteq H_1^{-1}(q^{-1}, \theta)\big(Y_t - G_1(q^{-1}, \theta)U_t\big),$$

$$N_t(\gamma) \doteq H_2^{-1}(q^{-1}, \gamma)\big(Z_t - G_2(q^{-1}, \gamma)V_t\big).$$

Dependence Measure-based Ranking

$$\mathcal{D}_j(\theta, \gamma) = \{(E_t(\theta), N_{\pi_j(t)}(\gamma))\}_{t=1}^{n} \quad \text{for} \quad j = 0, \dots, m-1,$$

$$\psi_\Delta(\theta, \gamma) \doteq 1 + \sum_{j=1}^{m-1} \mathbb{I}\left(\, \widehat{\Delta}_n^{(0)}(\theta, \gamma) \prec_\sigma \widehat{\Delta}_n^{(j)}(\theta, \gamma) \,\right).$$

Non-Asymptotic Significance Level

- We reject the null hypothesis H_0 if and only if for all $\theta \in \widehat{\Theta}$ and $\gamma \in \widehat{\Gamma}$ parameters in the confidence sets, we have $\psi_\Delta(\theta, \gamma) \leq r$.

Theorem (Non-Asymptotic Type I Error Probability Bound)

Assume (A1)–(A6). Let ψ be any ranking function, $\widehat{\Theta}$ and $\widehat{\Gamma}$ be confidence sets with significance level at most β. If H_0 holds, then

$$\mathbb{P}\Big(\max_{(\theta,\gamma)\in\widehat{\Theta}\times\widehat{\Gamma}} \psi_\Delta(\theta, \gamma) \leq r \Big) \leq \frac{r}{m} + 2\beta.$$

- The significance level is bounded by a user-chosen probability.
- This generalized construction is also distribution-free (independent of the underlying distribution) and non-asymptotic (holds for finite samples).

Sufficient Conditions for Consistency

(A8) Control inputs $\{U_t\}$, $\{V_t\}$ and process noises $\{E_t\}$, $\{N_t\}$ are (almost surely) included in a Césaro space for $p = \infty$, i.e., for $\{W_t\} \in \{\{U_t\}, \{V_t\}, \{E_t\}, \{N_t\}\}$ we have

$$\|W\|_{c(\infty)} \doteq \sup_{n \in \mathbb{N}} \frac{1}{n} \sum_{t=1}^{n} |W_t| < \infty.$$

(A9) Lipschitz condition: there a.s. exist $K, \tilde{\varepsilon} > 0$ such that

$$\|E(\theta_*) - E(\theta)\|_{c(\infty)} \leq K \cdot \|\theta_* - \theta\|,$$

for $\theta \in B(\theta_*, \tilde{\varepsilon})$, and respectively for $N(\gamma)$, where $\gamma \in B(\gamma^*, \tilde{\varepsilon})$.

(A10) The confidence sets are uniformly consistent, i.e., for all $\varepsilon > 0$ there a.s. exists an $N_0 \in \mathbb{N}$ such that for all $n > N_0$ both $\widehat{\Theta}_n \subseteq B(\theta_*, \varepsilon)$ and $\widehat{\Gamma}_n \subseteq B(\gamma_*, \varepsilon)$.

Consistency of the Independence Test

(A11) The dependence measure estimator $\widehat{\Delta}$ is Lipschitz continuous around (θ_*, γ_*), i.e., $\exists\, C, \tilde{\tilde{\varepsilon}} > 0$ such that

$$
\begin{aligned}
&|\widehat{\Delta}\big(\mathcal{D}_j(\theta_*, \gamma_*)\big) - \widehat{\Delta}\big(\mathcal{D}_j(\theta, \gamma)\big)| \\
&\leq C \cdot \Big(\|E(\theta_*) - E(\theta)\|_{c(\infty)} + \|N(\gamma_*) - N(\gamma)\|_{c(\infty)} \Big)
\end{aligned}
$$

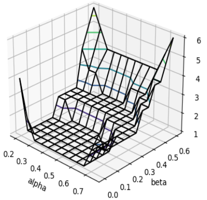
for $\theta \in B(\theta_*, \tilde{\tilde{\varepsilon}}), \gamma \in B(\gamma_*, \tilde{\tilde{\varepsilon}})$ and $j = 0, \dots, m-1$.

Theorem (Vanishing Type II Error Probability)

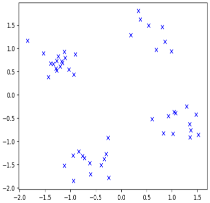
Assume (A1)–(A11). If H_1 holds true, then

$$
\mathbb{P}\Big(\max_{(\theta,\gamma)\in\widehat{\Theta}\times\widehat{\Gamma}} \psi_{\Delta}(\theta, \gamma) \leq r \Big) \xrightarrow{n\to\infty} 1.
$$

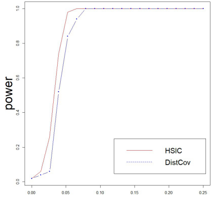
Numerical Experiments: AR(1) Processes



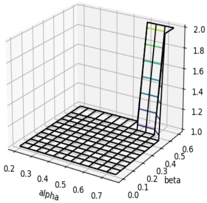
(a) Ranks: Dist.Cov.



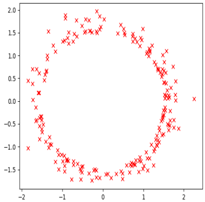
(b) Rotated Mixture



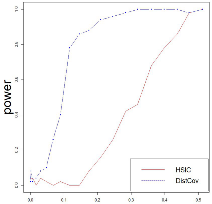
(c) Power function



(d) Ranks: HSIC



(e) Extinct Gauss



(f) Power function

HUN REN SZTAKI

Summary

- We presented a resampling-and-ranking framework for robust inference.
- We demonstrated the framework with a distribution-free hypothesis test for the independence of two synchronous linear stochastic systems.
- The test only needs a finite single trajectory of (paired) observations.
- The innovations can have arbitrary marginal distributions, and the constructed test can detect any nonlinear dependence between them.
- It performes the resampling by randomly permuting the residuals.
- The test has non-asymptotic user-chosen type I error probability bounds.
- Two ranking techniques were suggested for the test based on potent dependence measures (in particular: HSIC and distance covariance).
- Besides the exact quantification of the type I error probabilities, we argued that the test is also consistent (its power converges to one).

Thank you for your attention!

https://csaji.pages.sztaki.hu/ csaji@sztaki.hu

(Towards) Pruning Neural Networks at Initialisation in a Principled Way

Long Tran-Thanh
University of Warwick

long.tran-thanh@warwick.ac.uk

Main (real) authors

Hoang Pham
PhD student
@Warwick

The-Anh Ta
Researcher @ CSIRO

Lichuan Xiang
former PhD student
@ Warwick

(Non-official) Motivation

- My attempt to understand training process of neural networks

- I strongly believe that the topology of the network plays an important role

  - Typically overlooked by the ML community
  - Lacking lots of mathematical understanding

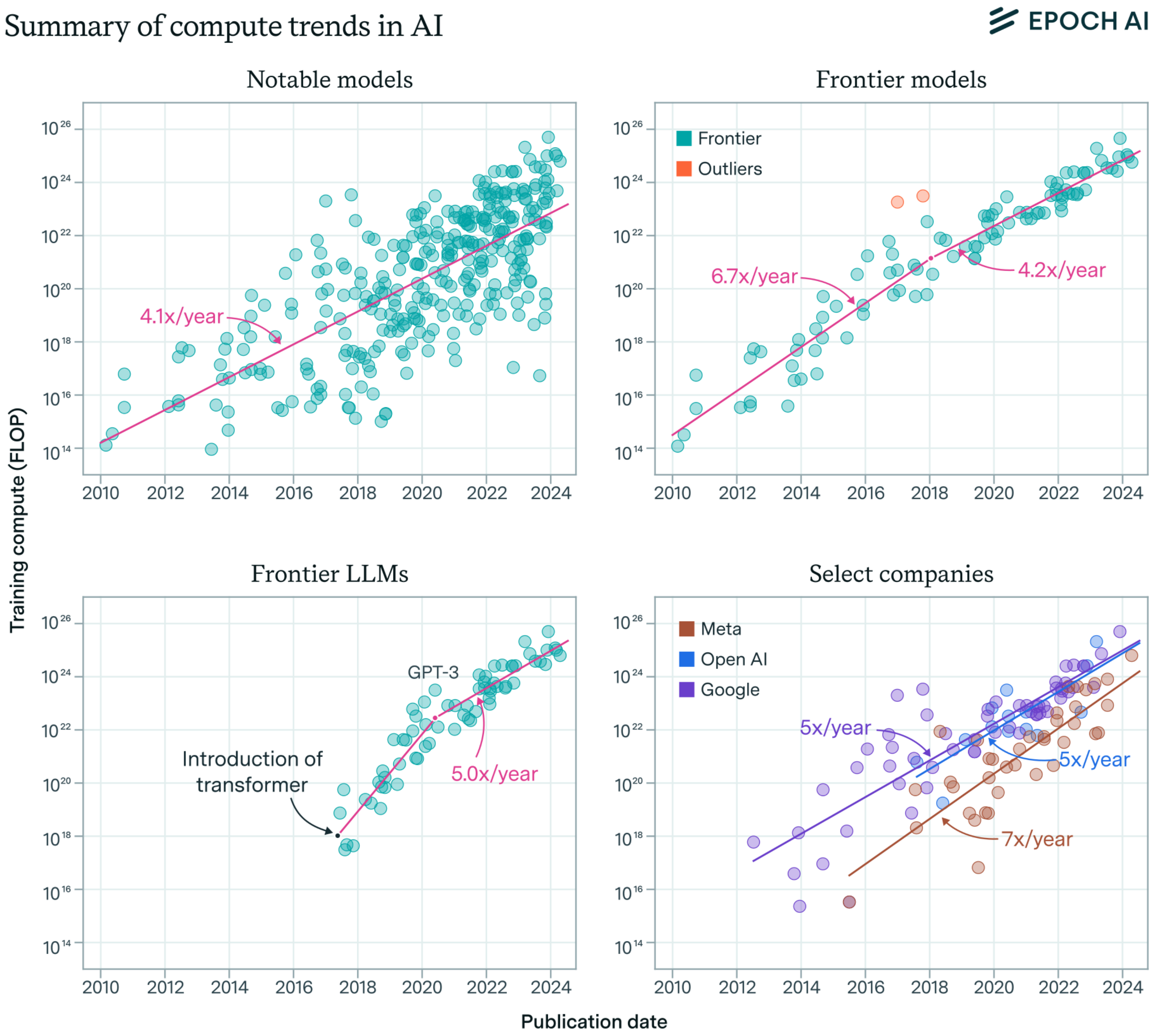
Roadmap

- Node-Path Balancing Principle (NeurIPS’23 & ICLR’25)
- Graphon Neural Tangent Kernel (under submission)

(Official) Motivation

Size of SOTA AI models over time

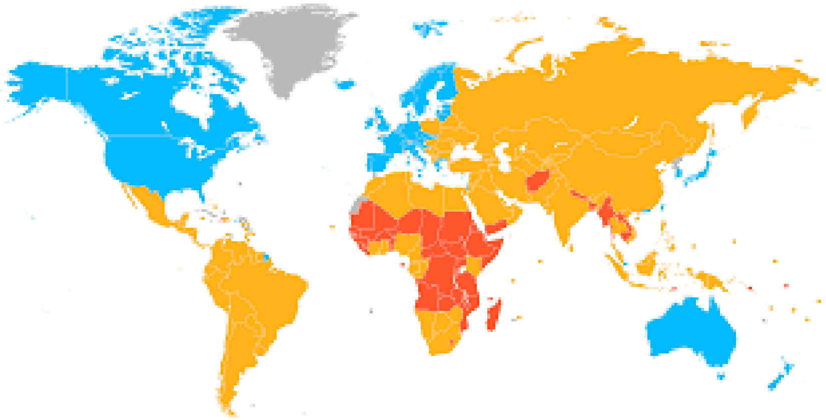
Favours small number of large corporates and rich governments



Motivation (cont'd)

Democratising AI?

Technological independence?



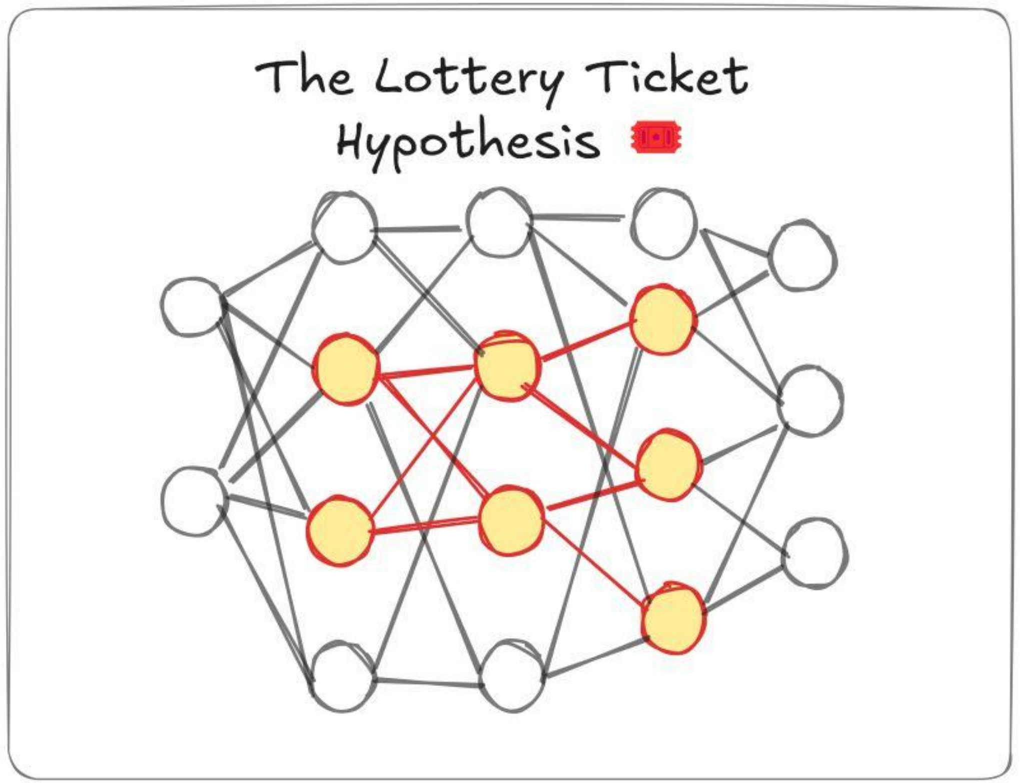
Need for small-scale but efficient AI models

Motivation (cont'd)

A new hope

Large, dense -> small, sparse

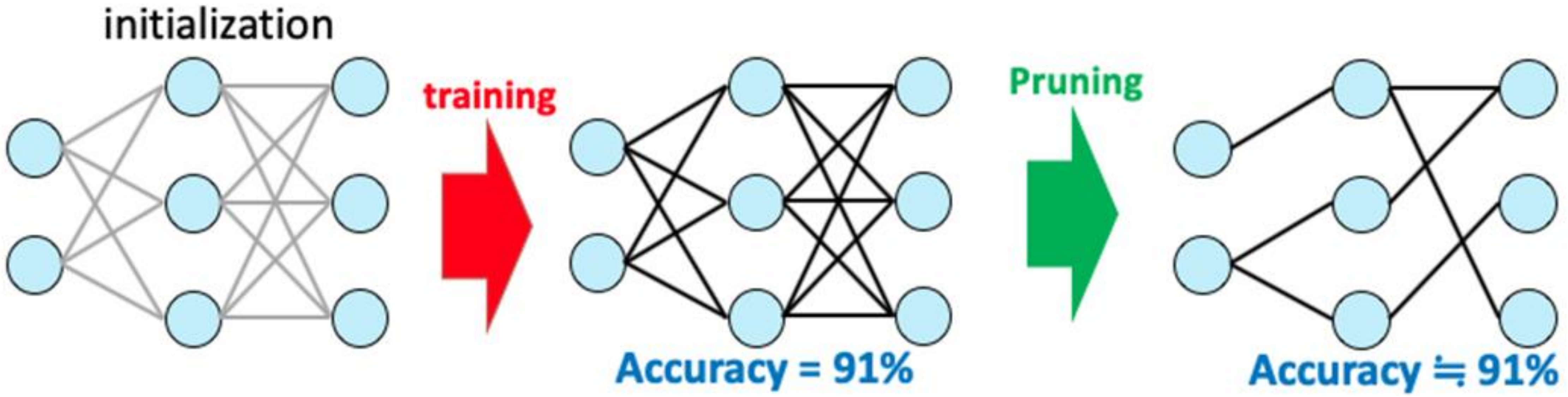
The lottery ticket hypothesis (Frankle & Carbin, ICLR 2019)



https://www.linkedin.com/posts/hesamsheikh_machinelearning-neuralnetwork-datascience-activity-7224451268801396736-woD9/

Pruning neural networks

Idea: train the NN, then prune it without losing performance



https://www.linkedin.com/pulse/why-lottery-ticket-hypothesis-important-practitioners-michele-de-vita/

Issue: very costly compute (cost of training dense network + pruning + evaluation)

Usage: inference phase

Pruning at Initialisation (PaI)

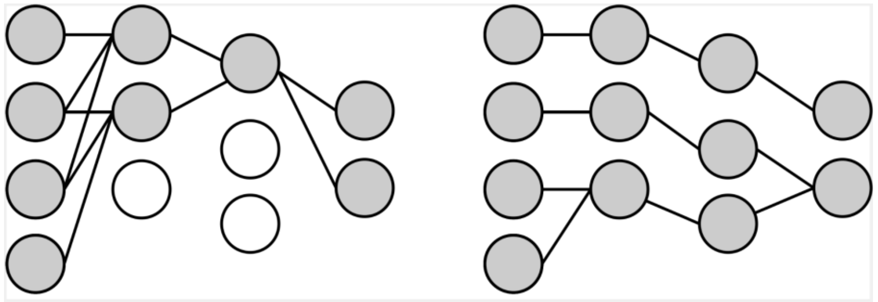
Idea: why not prune BEFORE training (sic!)

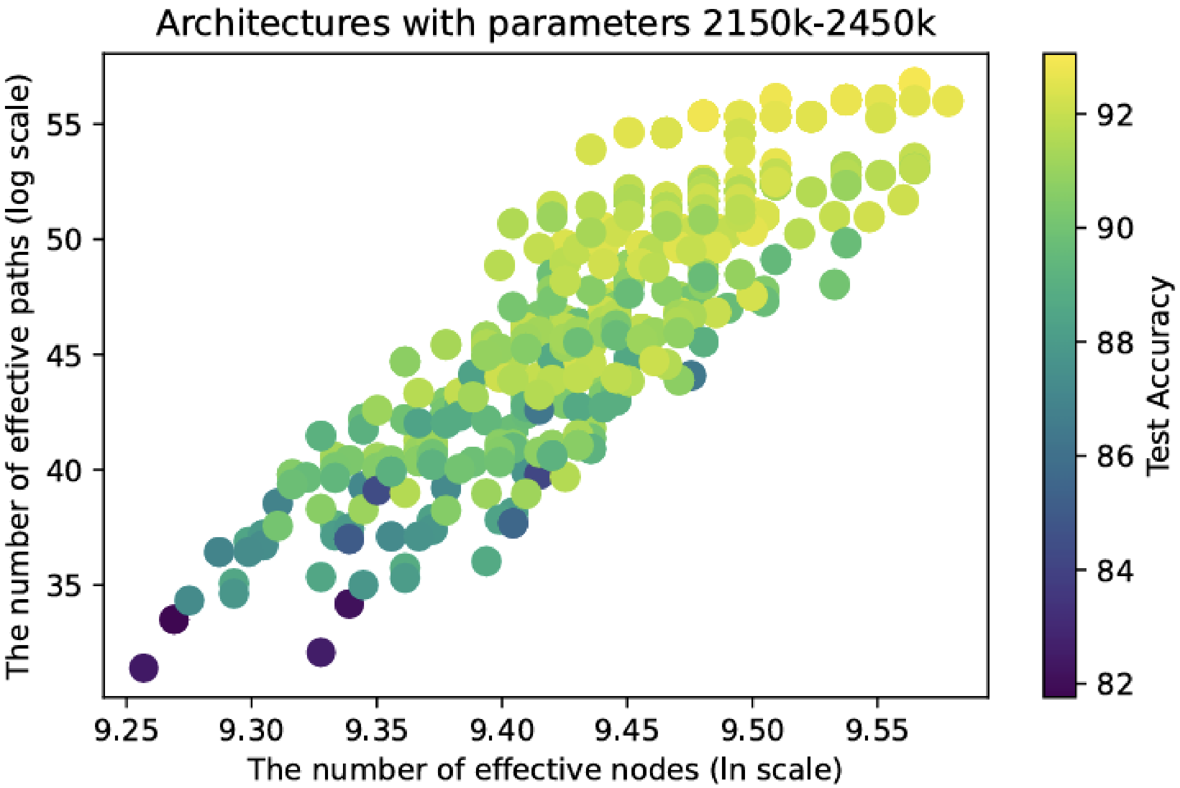
- Common PaI techniques identify prunable connections before training*:
  - Random Pruning
  - Magnitude Pruning: $|\theta|$
  - SNIP: $\left|\frac{\partial \mathcal{L}}{\partial \theta} \odot \theta\right|$
  - GraSP: $H \frac{\partial \mathcal{L}}{\partial \theta} \odot \theta$
  - SynFlow: $\left|\frac{\partial R}{\partial \theta} \odot \theta\right|$

These methods are heuristic since all fail with simple sanity checks: re-initialization, layer-wise shuffling connections, etc.,

Node-Path Balancing (Pham *et al.*, NeurIPS 2023)

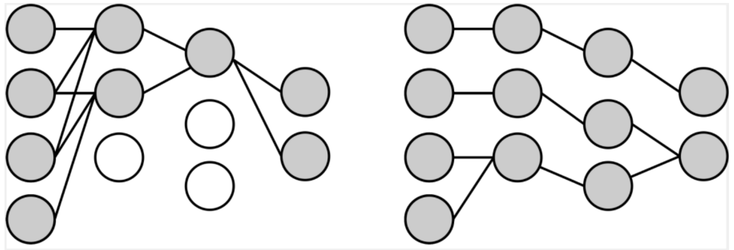
- Effective path: A path connecting input neuron with output neuron
- Effective node: A node on at least one effective path





The Node-Path Balancing (NPB) principle

- It hypothesizes a good sparse network should have:
  - high number of paths
  - high number of nodes (neurons, kernels, etc.,)

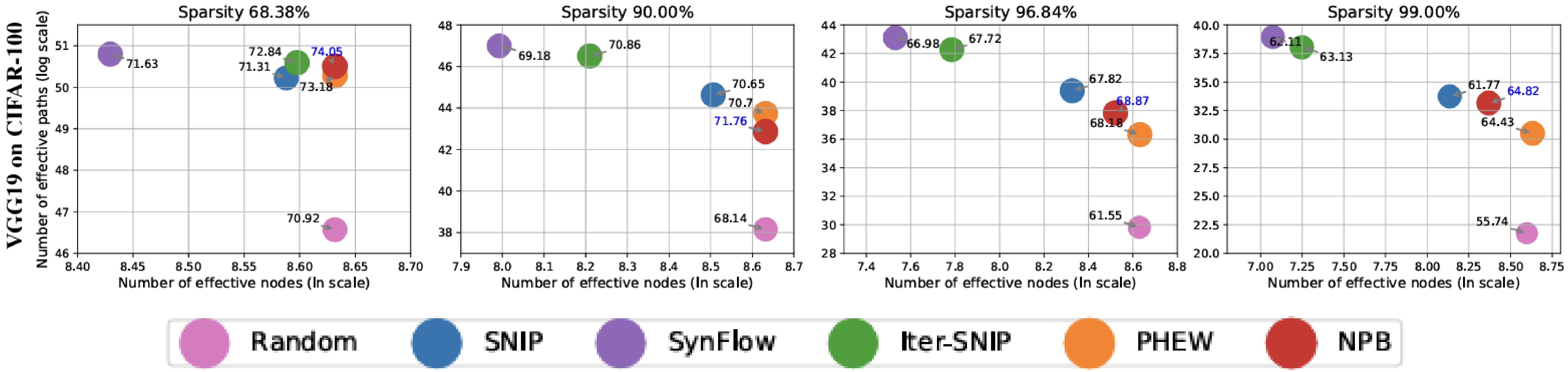


- Goal: find pruned network with high number of effective paths and nodes

Final objective becomes:

$$\underset{\mathbf{m}^{(l)}}{\text{Maximize}} \quad \alpha f_n + (1-\alpha) f_p + \beta R$$
$$s.t \quad ||\mathbf{m}^{(l)}||_1 \leq N^{(l)}(1 - s^{(l)})$$

The NPB principle (cont'd)



- NBP dominates the others in performance
- Pareto front on eff. node vs. eff. Paths (empirical justification of NPB)
- Issue: discrete optimisation component -> hard to solve + non-smooth integration

Differential pruning at Initialisation – DPaI (Xiang *et al.*, ICLR'25)

- Differential PaI (DPaI) extends NPB to continuous optimisation
  - Relaxes hard to soft masking: $m_{ij}^{(l)} \to s_{ij}^{(l)}$, then $m_{i,j}^{(l)} = \text{Top}_{k^{(l)}}(|s_{i,j}^{(l)}|)$
  - Reformulates node and path objectives via soft masks

The number of incoming paths to a node $v_j^{(l)}$

A node is effective when $N(v_j^{(l)}) > 0$

$$P(v_j^{(l)}) = \sum_{i=1}^{h^{(l-1)}} m_{i,j}^{(l)} P(v_i^{(l-1)}), \quad \mathcal{R}_P = \sum_{j=1}^{h^{(L)}} P(v_j^{(L)}) \quad N(v_j^{(l)}) = P(v_j^{(l)}) \frac{\delta \mathcal{R}_P}{\delta P(v_j^{(l)})}, \quad \mathcal{R}_N = \sum_{l,j} \tanh N(v_j^{(l)})$$

The number of outgoing paths from a node $v_j^{(l)}$

$$\frac{\delta \mathcal{R}_P}{\delta P(v_j^{(l)})} = \sum_{n,p,q,\ldots,k} m_{p,n}^{(L)} m_{q,p}^{(L-1)} \cdots m_{j,k}^{(l+1)}$$

Pruning: Update s by GD on node, path objectives

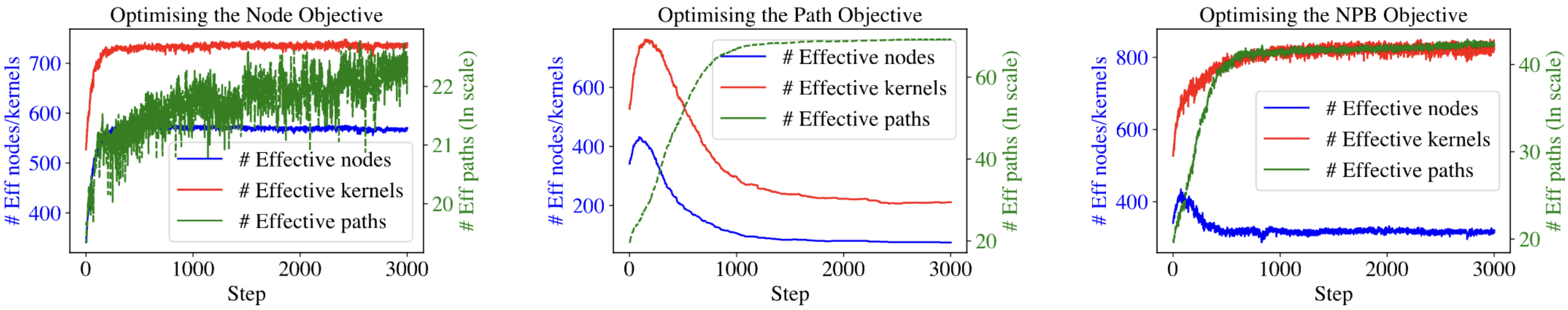
$$\frac{\delta \mathcal{R}_P}{\delta s_{i,j}^{(l)}} \propto \frac{\delta \mathcal{R}_P}{\delta P(v_j^{(l)})} P(v_i^{(l-1)}), \quad \frac{\delta \mathcal{R}_N}{\delta s_{i,j}^{(l)}} \propto \mathbf{1}_{N(v_j^{(l)})=0} \frac{\delta \mathcal{R}_P}{\delta s_{i,j}^{(l)}}$$

→ differentiable

Differential pruning at Initialisation – DPaI (cont'd)

Convergence of DPaI, we optimize soft mask w.r.t.

- Node objective only (nodes, kernel)
- Path objective only
- NPB objective



Graphon based pruning (Pham *et al*., under submission'25)

- Previous methods optimise masks given the network structure (topology)
- We further ask:
  - Is there underlying topological patterns behind pruning method?
  - Can we optimise the topology directly?

Idea: use graphon theory

- Graphons offer a tool to move from discrete to continuous masks
- If a graphon exists for each pruning method, we can analyse the training dynamics of sparse networks in infinite-width setting.

Graphon

- **Graphon** is a symmetric, measurable function $\mathcal{W} : [0,1]^2 \to [0,1]$

It serves as a limiting object of sequence of dense graphs

  - $\mathcal{W}(u,v)$: the probability of connection between nodes *u* and *v*
  - Think of graphon as a "heat map" of adjacency matrix (original graph = step function)

- **Cut Distance** measures the global similarity and is invariant to node relabelling

$$\delta_{\square}(\mathcal{U},\mathcal{W}) = \inf_{\phi \in \Phi} \sup_{S,T \subset [0,1]} \left| \int_{S \times T} (\mathcal{U}(x,y) - \mathcal{W}(\phi(x),\phi(y))) dx dy \right|$$

- **Convergence in Cut Distance**: a sequence of graphs (G_n) converges to a graphon $\mathcal{W}$ if their step-function representation $(\mathcal{W}_{G_n})$ converge to $\mathcal{W}$ in cut distance:

$$\lim_{n \to \infty} \delta_{\square}(\mathcal{W}_{G_n}, \mathcal{W}) = 0$$

The Graphon Limit Hypothesis

Given a sparsity level, each network pruning method defines sequences of binary masks that converges layer-wisely to graphons in cut distance as the network width is increased to infinity.

Each method converges to a unique and characteristic graphon structure.

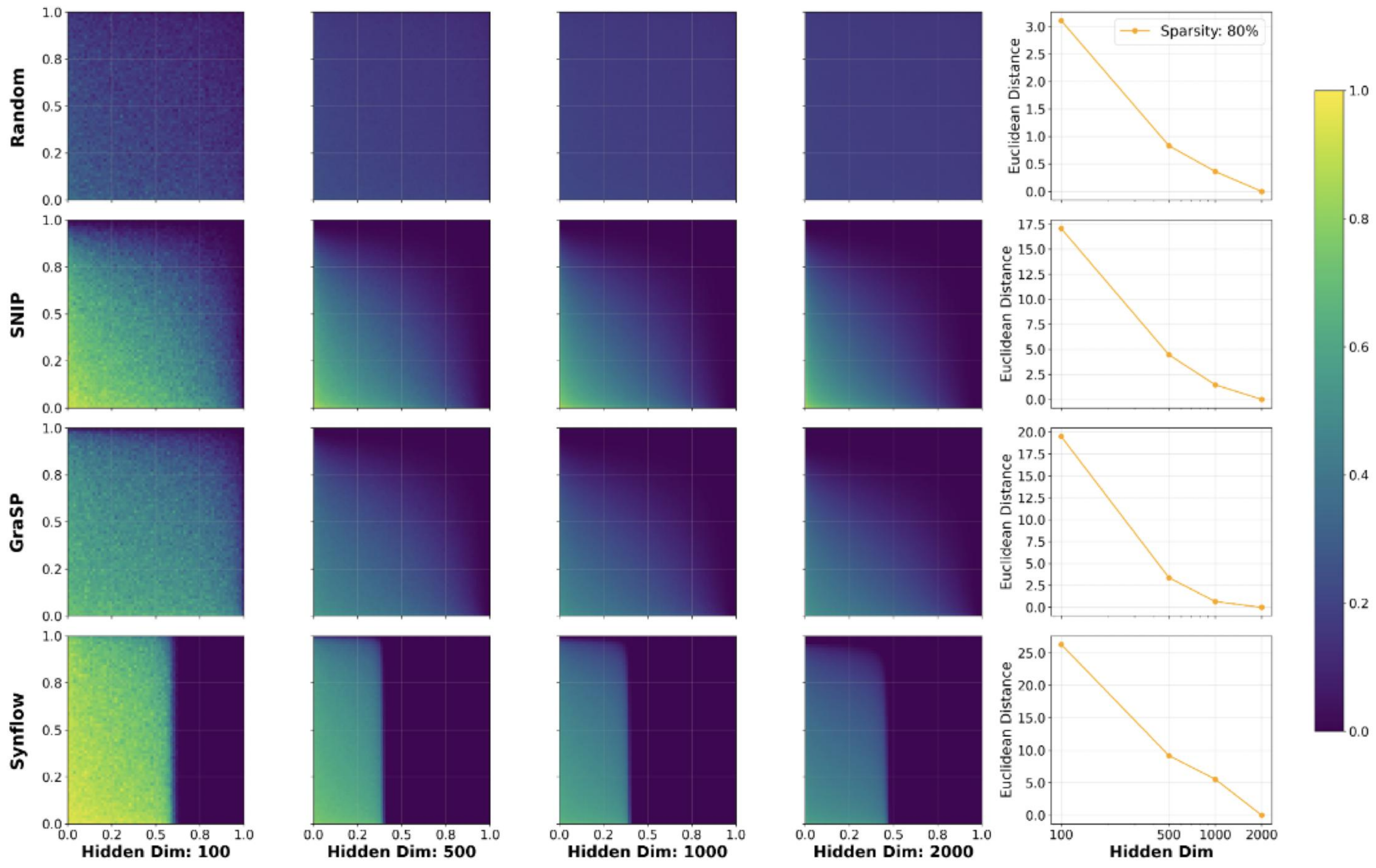


Figure 1: Graph limit of subnetworks' mask produced by PaI methods at 80% sparsity and the corresponding convergence of graphons via Euclidean distances.

Inspiration: Neural Tangent Kernel (NTK)

$$h_j^{(0)}(\mathbf{x}) = \mathbf{x}_j$$

$$z_i^{(l)}(\mathbf{x}) = \frac{1}{\sqrt{n_{l-1}}} \sum_{j=1}^{n_{l-1}} W_{ij}^{(l)} h_j^{(l-1)}(\mathbf{x})$$

$$h_i^{(l)}(\mathbf{x}) = \sigma(z_i^{(l)}(\mathbf{x}))$$

$$f(\mathbf{x};\theta) = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} h_i^{(L)}(\mathbf{x})$$

Definition: $\Theta(x, x') = \left\langle \frac{\partial f(x)}{\partial \theta}, \frac{\partial f(x')}{\partial \theta} \right\rangle$

where the network has infinite-width limit and weight $W \sim \mathcal{N}(0, \sigma_w^2)$

When network's width $n \to \infty$, NTK converges to:

$$\Theta(x, x') = \sum_{l=1}^{L} \left(\Sigma^{(l-1)}(x, x') \prod_{m=l}^{L} \sigma_w^2 \dot{\Sigma}^{(m)}(x, x') \right)$$

where $\Sigma()$ and $\dot{\Sigma}()$ are activation covariance and derivative covariance

NTK is a useful tool for studying the dynamics of infinitely wide neural networks

Sparse network analogy: Graphon NTK

- NTK works on dense network in infinite-width setting
- Graphon is representation of sparse network in infinite-width
- Graphon + NTK analysis to analyse training dynamics of SNNs

Discrete

$$W_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_w^2)$$

$$M^{(l)} \to \mathcal{W}^{(l)}$$

Continuous

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \mathcal{W}^{(l)}\left(\frac{i}{n_l}, \frac{j}{n_{l-1}}\right)\right)$$

$$z_i^{(l)}(\mathbf{x}) = \frac{1}{\sqrt{n_{l-1}}} \sum_{j=1}^{n_{l-1}} W_{ij}^{(l)} h_j^{(l-1)}(\mathbf{x})$$

$$h_i^{(l)}(\mathbf{x}) = \sigma(z_i^{(l)}(\mathbf{x}))$$

$$f(\mathbf{x}; \theta) = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} h_i^{(L)}(\mathbf{x})$$

$$z^{(l)}(u, x) = \int_0^1 W^{(l)}(u, v) h^{(l-1)}(v, x) dv$$

$$h^{(l)}(u, x) = \sigma(z^{(l)}(u, x))$$

$$f(x) = \int_0^1 h^{(L)}(u, x) du$$

Graphon NTK – Pre-activation Covariance

Previous layer activation covariance

NTK

$$\tilde{\Sigma}^{(l)}(x, x') = \Sigma^{(l-1)}(x, x')$$

where $\sigma_w^2 = 1$

Graphon NTK

$$\tilde{\Sigma}^{(l)}(u_l, u_l', x, x') = \delta(u_l - u_l') \int_0^1 \mathcal{W}^{(l)}(u_l, u_{l-1}) \Sigma^{(l-1)}(u_{l-1}, u_{l-1}, x, x') du_{l-1}$$

Graphon NTK modulates prev. layer's activation covariance with graphon function

- Position-dependent covariance structures
- Creating non-uniform signal propagation through the network
- Connectivity strengths determined by graphon values

Graphon NTK – Deterministic Kernel

NTK

$$\Theta(x,x') = \sum_{l=1}^{L} \left(\Sigma^{(l-1)}(x,x') \cdot \prod_{m=l}^{L} \dot{\Sigma}^{(m)}(x,x') \right) \quad \text{where} \quad \sigma_w^2 = 1$$

Graphon NTK

$$\Theta(x,x') = \sum_{l=1}^{L} \int_0^1 \left(\dot{\Sigma}^{(l)}(u_l,u_l,x,x') \int_{[0,1]^{L-l+1}} \prod_{m=l+1}^{L+1} \mathcal{W}^{(m)}(u_m,u_{m-1}) \dot{\Sigma}^{(m)}(u_m,u_m,x,x')\, d\mathbf{u}_{l+1} \right) \cdot \left(\int_0^1 \Sigma^{(l-1)}(u_{l-1},u_{l-1},x,x') du_{l-1} \right) du_l$$

- Graphon function shapes the kernel through position-dependent connectivity
- Analysing how connectivity pattern in sparse NNs affects learning dynamics

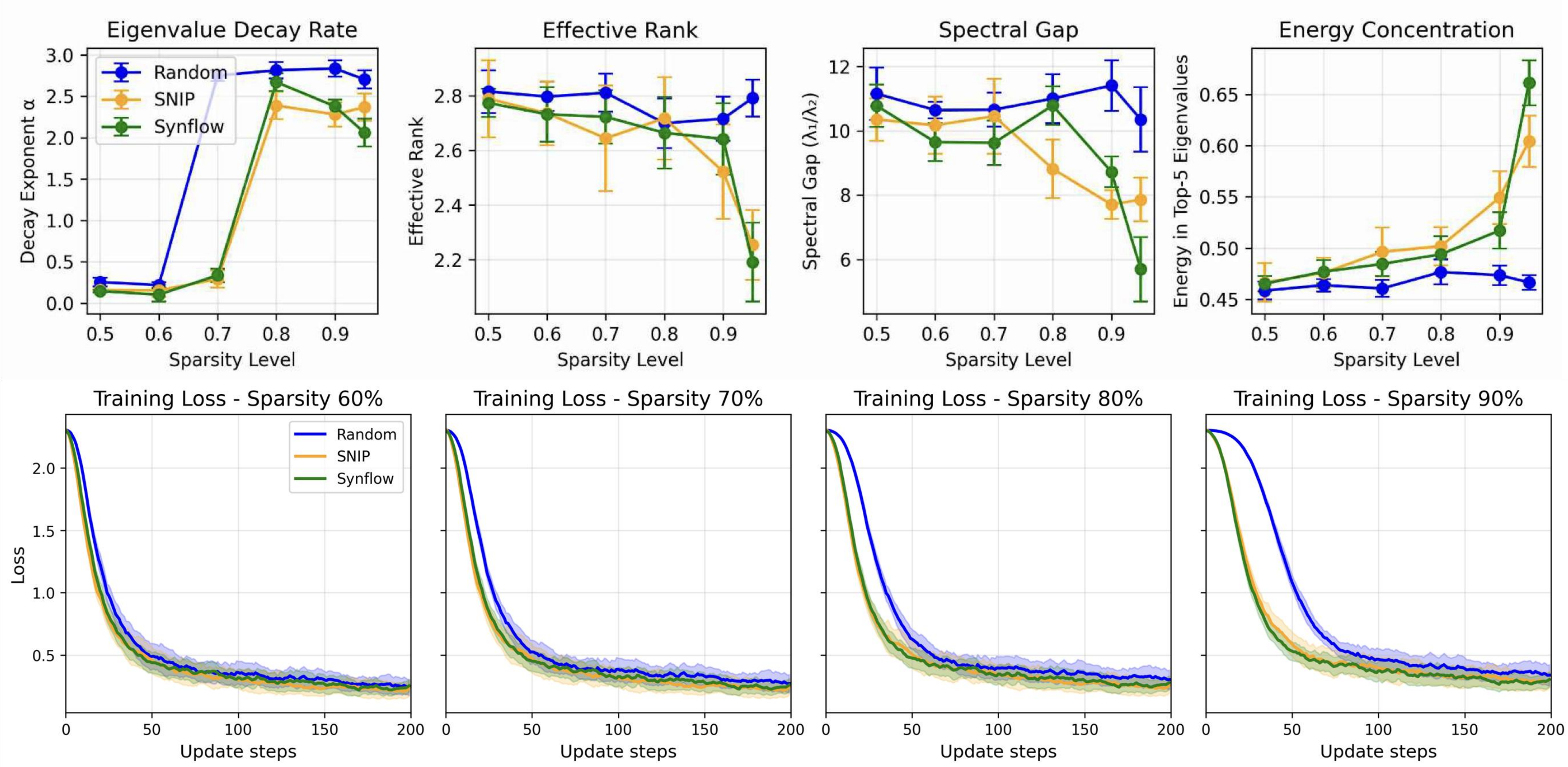
Example: Graphon NTK for homogeneous graphons

- Our framework recovers and generalises previous finding in Yang et. al., (AISTATS'23)
- Random Pruning corresponds to constant graphon $\mathcal{W}(u, v) = s$, where s is connection density
- Graphon NTK simplifies to a scaled version of dense NTK:

$$\Theta(x, x') = s^L \; \Theta_{\mathrm{std}}(x, x')$$

- Random Pruning uniformly scales down the learning speed across all eigen-directions

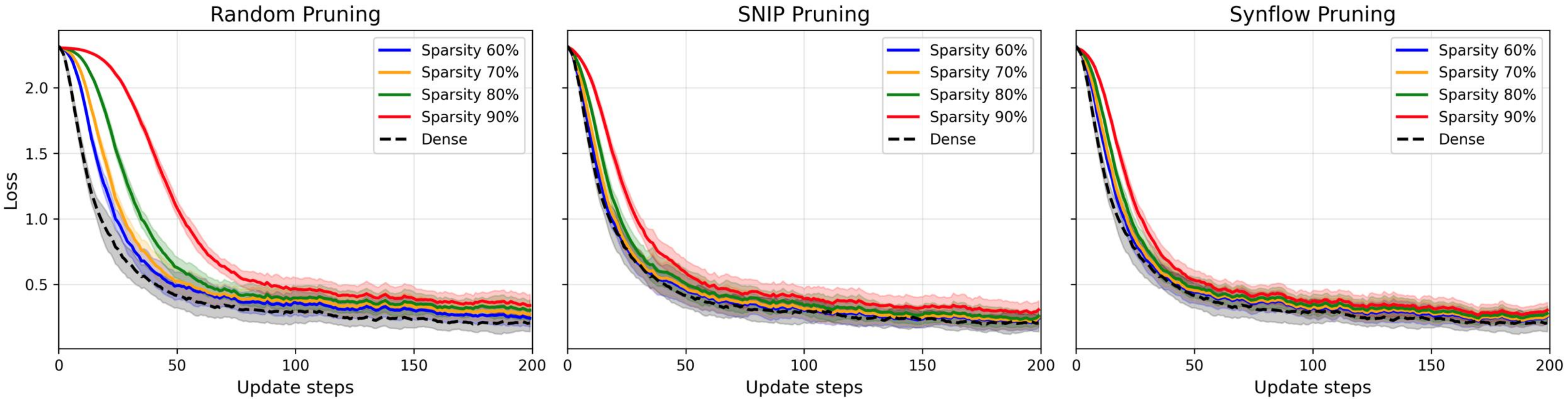
Graphon NTK – Experiments



Graphon NTK from SNIP/SynFlow have high energy concentration at top eigenvalues

- SNIP/SynFlow sparse network learn faster at beginning of training than Random ones

Graphon NTK – Experiment



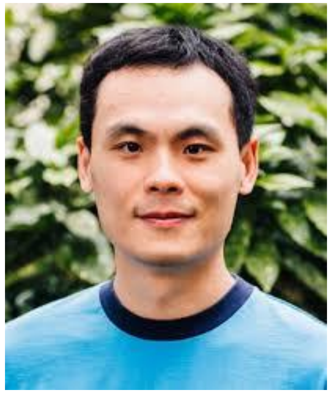
Increasing sparsity level slows down training speed!

Future Directions for graphon based pruning

- Graphon-Guided Pruning: Design new algorithms that directly optimise for a graphon with desirable spectral properties (Pham *et al*., on-going work)

- Continuous Optimisation: Move from discrete mask selection to optimising in the continuous space of graphons.

- Broader Architectures: Extend the framework to other architectures like Transformers and CNNs.



Hongkai Wen

Rebekka Burkholz

Tom Jacobs

Hoang Pham

The-Anh Ta

Lichuan Xiang

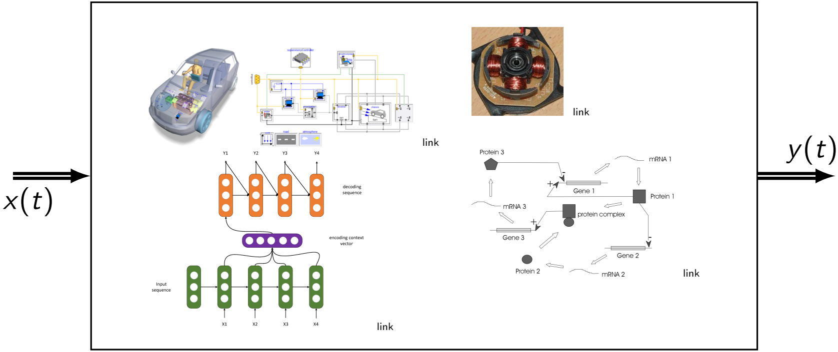
Many thanks for your attention!

Statistical guarantees for learning dynamical systems

Mihály Petreczky

CNRS, École Centrale Lille, University of Lille, CRIStAL

Hungarian Machine Learning Days, August 12-14, 2025.



Maps input time-series (sequence) to output time-series (sequence)

t-time (position in the sequence).

Mathematical models: difference/differential equations

Dynamical systems

Maps input time-series (sequence) to output time-series (sequence)

t-time (position in the sequence).

Mathematical models: difference/differential equations

$$x(t) \Longrightarrow \boxed{\begin{array}{rcl} \left\{ \begin{array}{c} \dot{s}(t) \\ s(t+1) \end{array} \right\} & = & f(s(t), x(t)) \\ y(t) & = & h(s(t), x(t)) \end{array}} \Longrightarrow y(t)$$

Goal: from observed data x, y learn a model (difference/differential equation).

Examples: RNN (Recurrent Neural Networks), deep SSM (Mamba), neural ODEs

Learning dynamical systems: motivation, challenges

Motivation: we need model for prediction and decision making, e.g. foundational models, feedback control, etc.

Old problem: time-series forecasting, system identification (control).

Dynamical systems are used in safety critical applications: need statistical guarantees.

Challenges:

- possibly non i.i.d. data (time-series)
- hidden states: depend on an increasing number of inputs.
- unbounded data.

Learning problem for dynamical systems

- $\mathbb{X}$ – input-space, $\mathbb{Y}$ output space.
- $\mathbf{x}(t) \in \mathbb{X}$ – input process, $\mathbf{y}(t) \in \mathbb{Y}$ – output process,

$$t \in \mathbb{T} = \begin{cases} \mathbb{Z} & \text{discrete-time} \\ \mathbb{R}_+ & \text{continuous-time} \end{cases}$$

- Hypotheses:
  $\mathcal{H} \subseteq \{$ functions of the form $h : \bigcup_{t \in \mathbb{T}} (\mathbb{X} \times \mathbb{Y})^{[0,t)} \to \mathbb{Y}\}$.

  $h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \leq s < t})$ – prediction of output $\mathbf{y}(t)$ based on past values of the inputs and outputs.

  Example: dynamical systems

$$\begin{Bmatrix} \hat{\mathbf{s}}(t+1) \\ \frac{d}{dt}\hat{\mathbf{s}}(t) \end{Bmatrix} = f(\hat{\mathbf{s}}(t), \mathbf{x}(t), \overbrace{\mathbf{y}(t)}^{\text{If past measurements of labels are available}}),\ \hat{\mathbf{s}}(0) = \hat{\mathbf{s}}_0,$$

$$h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \leq s < t}) := g(\hat{\mathbf{s}}(t)).$$

Learning problem for dynamical systems

- Loss function: $\ell : \mathbb{Y} \times \mathbb{Y} \to [0, +\infty)$

  $\ell(\mathbf{y}(t), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t}))$ difference between the output predicted by $h \in \mathcal{H}$ and true output.

  $\ell(y, y^{'}) = (y - y')^2$ (ℓ_2 loss), $\ell(y, y') = |y - y'|$ (ℓ_1 loss).

- True error for a hypothesis h:
  Fixed final time prediction error:

$$\mathcal{L}(h) = \mathbf{E}[\ell(\mathbf{y}(T), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < T}))]$$

  Long-term prediction error:

$$\mathcal{L}(h) = \lim_{T \to \infty} \mathbf{E}[\ell(\mathbf{y}(T), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < T}))]$$

Learning problem

Learning problem: find $h_\star \in \mathcal{H}$ such that $\mathcal{L}(h_\star)$ is small from

- Single time series (discrete-time) sample from $\{(\mathbf{x}(t), \mathbf{y}(t))\}_{0 \le t < N}$
- Multiple time-series (discrete- or continuous time) sample from i.i.d variable $\{(\mathbf{x}_i(t), \mathbf{y}_i(t))\}_{t \in [0,T]}$ $i = 1, \ldots, N$, having the same distribution as $(\mathbf{x}, \mathbf{y})$.

Solution:
Define the empirical error for hypothesis h:

$$\hat{\mathcal{L}}_N(h) = \begin{cases} \frac{1}{N} \sum_{t=0}^{N-1} \ell(\mathbf{y}(t), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t})) & \text{single time-series} \\ \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{y}(T), h(\{\mathbf{x}_i(s), \mathbf{y}_i(s)\}_{0 \le s < T})) & \text{multiple time-series} \end{cases}$$

Let $h_\star$ be such that $\left(\hat{\mathcal{L}}_N(h) + \text{ regularization term }\right)$ is small.

Question:

What can we say about the true error $\mathcal{L}(h_\star)$?

Generalisation bounds

Desired results:

$$\mathbf{P}(\mathcal{L}(h_\star) \leq \underbrace{\hat{\mathcal{L}}_N(h_\star)}_{\text{known from data}} + \underbrace{r(N,\delta,\ldots)}_{\text{functions of } N,\ \delta}) > 1-\delta$$

Typically, we want: $r(N,\delta,\ldots) = C\frac{\ln(1/\delta)+1}{\sqrt{N}}$

Ways to derive it:

- PAC bounds.
- PAC-Bayesian bounds.
- Bounds specific to certain learning algorithms.

$$\mathbf{P}(\forall h \in \mathcal{H} : \mathcal{L}(h) - \hat{\mathcal{L}}_N(h) < C\frac{\ln(1/\delta)+1}{\sqrt{N}}) > 1-\delta$$

The bound holds for all h, thus for the learned model $h_\star$ too.

Conceptually easy, but conservative.

E.g.: for RNNs learned from N i.i.d. time series of length T, PAC bounds are exponential in T.

PAC-Bayesian bound

for any prior distribution π on models, any $\lambda > 0$

$$\mathbf{P}\Big(\text{for any posterior } \rho : E_{h\sim\rho}\mathcal{L}(h) < E_{h\sim\rho}\hat{\mathcal{L}}_N(h) + \\ +\frac{1}{\lambda}\Big[KL(\rho \mid \pi) + \ln(\frac{1}{\delta}) + \frac{\lambda C_\pi}{\sqrt{N}}\Big]\Big) > 1-\delta$$

$KL(\rho \mid \pi)$ – KL divergence, ρ - posterior distribution on models

Getting generalisation bounds:

- choose a prior π, a data-dependent posterior distribution ρ_N s.t. $KL(\rho_N \mid \pi) < O(\sqrt{N})$.
- Learned model $h_\star$: sample/mean/max. likelihood of ρ_N.
- Choose $\lambda = O(\sqrt{N})$

$$\mathbf{P}\left(\mathcal{L}(h_\star) \leq \hat{\mathcal{L}}_N(h_\star) + \tilde{C}\frac{\ln(1/\delta)+1}{\sqrt{N}}\right) > 1-\delta.$$

Works for learning with regularisation (encoded by π).

Learning based on PAC-Bayesian error bounds

Learn by

1. find a posterior $\rho = \hat{\rho}_N$ and a parameter $\lambda = \lambda_N$ such that

$$E_{h\sim\rho}[\hat{\mathcal{L}}_N(h)] + \frac{1}{\lambda}\left[KL(\rho \mid \pi) + \ln(\frac{1}{\delta}) + \frac{\lambda C_\pi}{\sqrt{N}}\right]$$

   is small.

2. $h_\star$ is one of the following:
   - $h_\star$ random sample from $\hat{\rho}_N$, or
   - most likely model, i.e. $h_\star = \sup_{h\in\mathcal{H}} \hat{\rho}_N(h)$, or
   - $h_\star$ is the mean model $E_{h\sim\hat{\rho}_N}[h]$.

Explicit formula for $\hat{\rho}_N$.

Analytic bound (with high probability) on the generalization gap

$$\mathcal{L}(h_\star) - \hat{\mathcal{L}}_N(h_\star)$$

Learning with PAC-Bayesian bounds and maximal likelihood

KL divergence based bound: explicit formula for $\hat{\rho}_N$ (Gibbs posterior)

$$\hat{\rho}_N(h) = \frac{1}{Z} e^{-\lambda \hat{\mathcal{L}}_N(h)} \pi(h)$$
$$\hat{\rho}_N = \operatorname*{argmin}_{\rho} \left(E_{h \sim \rho} \hat{\mathcal{L}}_N(h) + KL(\rho \| \pi) \right)$$

Relationship with maximum likelihood and prediction error minimisation

$$\hat{h} = \operatorname*{argmax}_{h \in \mathcal{H}} \hat{\rho}_N(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \left(\hat{\mathcal{L}}_N(h) - \frac{1}{\lambda} ln(\pi(h)) \right)$$

Using prior π introduces assumption on the model structure, it can be viewed as a regularisation

PAC-Bayesian bounds: Inequalities for models chosen from the posterior

With a probability at least $(1-2\delta)$ over the data and over all samples $h_\star$ drawn from $\hat{\rho}_N$,

$$\mathcal{L}(h_\star) \leq \hat{\mathcal{L}}_N(h_\star) + \frac{\ln \frac{\hat{\rho}_N(\theta_\star)}{\pi(\theta_\star} + \ln \frac{1}{\delta} + C_\pi}{\sqrt{N}}$$

PAC-Bayesian to PAC boounds

If π is uniform or Gaussian $\mathcal{N}(h_m, P)$, loss L-Lipschitz, PAC-Bayesian gives for any estimate $\hat{h}$

$$\mathcal{L}(\hat{h}) \leq \hat{\mathcal{L}}_N(\hat{h}) + \frac{2\sigma^2 L}{\sqrt{N}} + \frac{C(\hat{h}) + \ln\frac{1}{\delta} + C_\pi}{\sqrt{N}}$$

with probability at least $1 - 2\delta$ over data, where

$$C(h) = \frac{1}{2}\Big(\frac{\sigma^2 \mathrm{trace}(P^{-1})}{\sqrt{N}} - n_h + (h - h_m)^T P^{-1}(h - h_m) + \ln\frac{\det(P)\sqrt{N}}{\sigma^2}\Big)$$

if π is Gaussian, and $C = \ln\frac{\sqrt{N}^{n^h} vol(\Theta)}{(\sqrt{\pi 0.25\sigma^2}^{n_h}\Gamma(n_h/2+1)}$ if π is uniform distribution.

Stable dynamical systems as data generators and models

Data generator: Lyapunov stable systems

$$\mathbf{s}_g(t+1) = f_g(\mathbf{s}_g(t), \mathbf{e}_g(t)), \quad \mathbf{e}_g \text{ - bounded i.i.d. noise.}$$
$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{x}(t) \end{bmatrix} = h_g(\mathbf{s}_g(t), \mathbf{e}_g(t))$$

Stability: small past perturbations in the past do not lead to large errors in the future.

Hypothesis: stable dynamical systems parametrized by θ

$$\hat{\mathbf{s}}(t+1) = f_\theta(\hat{\mathbf{s}}(t), \mathbf{x}(t), \mathbf{y}(t)), \quad \hat{\mathbf{s}}(0) = \hat{\mathbf{s}}_0,$$
$$\hat{\mathbf{y}}(t \mid 0) := h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t}) := g_\theta(\hat{\mathbf{s}}(t)).$$

Data generator and predictors can be RNNs.

Learning problem for a single time-series

Learning data:

$$\text{sample } \{x(t), y(t)\}_{t=0}^{N} \text{ of the process } \{\mathbf{x}(t), \mathbf{y}(t)\}_{t=0}^{N}.$$

Empirical loss:

$$\hat{\mathcal{L}}_N(\theta) \triangleq \frac{1}{N} \sum_{i=0}^{N-1} \ell(\mathbf{y}(t), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t})).$$

True error: long-term prediction error

$$\mathcal{L}(h) = \lim_{T \to \infty} \mathbf{E}[\ell(\mathbf{y}(T), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < T}))]$$

Long-term prediction error $\iff$ prediction error in stationary regime

Data generator stable: $\mathbf{x}, \mathbf{y}$ stationary, ergodic, weakly dependent.

Stationary regime of the predictor (using infinite past):

$$\hat{\mathbf{s}}(t+1) = f_\theta(\hat{\mathbf{s}}(t), \mathbf{x}(t), \mathbf{y}(t)), \ \hat{\mathbf{s}}(-\infty) = 0,$$
$$\hat{\mathbf{y}}_\theta(t) = h_\theta(\hat{\mathbf{s}}(t), \mathbf{x}(t)).$$

Due to stability $\hat{\mathbf{s}}(t)$ is stationary, and

$$\mathcal{L}(\theta) = \mathbf{E}[\ell(\hat{\mathbf{y}}_\theta(t), \mathbf{y}(t))]$$

Example: RNNs

data generators is an RNN with $\|A_g\|_2 < 1$

$$\mathbf{s}_g(t+1) = \mathrm{ReLu}\left(A_g \mathbf{s}_g(t) + K_g \mathbf{e}_g(t) + B_g \mathbf{x}(t)\right)$$
$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{x}(t) \end{bmatrix} = \tanh\left(C_g \mathbf{s}_g(t) + F_g \mathbf{e}_g(t)\right)$$

hypotheses are RNNs with parameters (weights) $\theta = (A, B, K, C)$
$\|A\|_2 < 1$

$$\hat{\mathbf{s}}(t+1) = \mathrm{ReLu}\left(A\hat{\mathbf{s}}(t) + B\mathbf{x}(t) + K\mathbf{y}(t)\right), \quad \hat{\mathbf{s}}(0) = 0$$
$$h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t}) = \tanh\left(C\hat{\mathbf{s}}(t)\right)$$

Example: linear systems

Data generator: stable linear system

$$\mathbf{s}_g(t+1) = A_g \mathbf{s}_g(t) + K_g \mathbf{e}_g(t),$$
$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{x}(t) \end{bmatrix} = C_g \mathbf{s}_g(t) + \mathbf{e}_g(t)$$

$\|A_g\|_2 < 1$ and $\mathbf{e}_g$ i.i.d. white noise.

Predictors: stable linear systems parametrised by $\theta \in \Theta$ ($\hat{A}(\theta)$ has all its eigenvalues inside the unit circle)

$$\hat{\mathbf{s}}(t+1) = \hat{A}(\theta)\hat{\mathbf{s}}(t) + \hat{B}(\theta)\mathbf{x}(t) + \hat{K}(\theta)\mathbf{y}(t), \ \hat{\mathbf{s}}(0) = 0,$$
$$h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t}) = \hat{C}(\theta)\hat{\mathbf{s}}(t).$$

Example: for linear systems data generator is the best predictor

If no feedback (Granger-causality) from $\mathbf{y}$ to $\mathbf{x}$, then the data generator is a noisy input-output model:

$$\begin{aligned} \mathbf{s}(t+1) &= A\mathbf{s}(t) + B\mathbf{x}(t) + K\mathbf{e}(t), \\ \mathbf{y}(t) &= C\mathbf{x}(t) + \mathbf{e}(t) \end{aligned}$$

A, B, K, C, D are one-to-one functions of A_g, K_g, C_g and the predictor

$$\begin{aligned} &\hat{\mathbf{s}}(t+1) = (A - KC)\hat{\mathbf{s}}(t) + B\mathbf{x}(t) + K\mathbf{y}(t),\ \hat{\mathbf{x}}(0) = 0, \\ &h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{0 \le s < t}) = C\hat{\mathbf{s}}(t). \end{aligned}$$

gives the minimal true loss.

Learning data generator $\iff$ finding best predictor

PAC-Bayesian bounds for non-linear systems

Theorem (Eringis & et al. 2024)

For all $\lambda > 0$, $\delta \in [0, 0.5)$, *for all priors* π

$$\mathbf{P}\left(\forall \rho : E_{h\sim\rho}\mathcal{L}(h) \leq E_{h\sim\rho}\hat{\mathcal{L}}_N(h) + r_N(\lambda, \pi, \rho, \delta)\right) > 1 - 2\delta$$

$$r_N(\lambda, \pi, \rho, \delta) \triangleq \frac{1}{\lambda}\left[KL(\rho||\pi) + \ln\frac{1}{\delta} + \frac{\lambda}{\sqrt{N}}C_\pi\right]$$

where C_π *depends on the prior* π *and on the class of predictors.*

Bounds can be computed. Extensions to unbounded linear RNNs (Eringis & et al, 2024).

Proof: Donsker-Varadhan change of measure + McDiarmid inequality for weakly dependent processes.

PAC bound for learning ODEs from multiple time-series

Predictors: stable bilinear ODEs (have universal approximation property), use only input $\mathbf{x}$,

$$\dot{\mathbf{s}}(t) = A(\theta)\mathbf{s}(t) + \sum_{i=1}^{m} N_i(\theta)\mathbf{s}(t)\mathbf{x}_i(t) + B(\theta)\mathbf{x}(t),$$
$$h(\{\mathbf{x}(s)\}_{0\le s<t}) = C(\theta)\mathbf{s}(t),\ \mathbf{s}(0) = 0.$$

$(\mathbf{x}, \mathbf{y})$ bounded random variable, values of $\mathbf{x}$ are functions on $[0, T]$.

Training data: bounded $(\mathbf{x}_1, \mathbf{y}_1), \dots (\mathbf{x}_N, \mathbf{y}_N)$ i.i.d. random variables distributed as $(\mathbf{x}, \mathbf{y})$.

Empirical error:

$$\hat{\mathcal{L}}_N(h) = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{y}_i(T), h(\{\mathbf{x}_i(s)\}_{0\le s<T})))$$

True error final-time prediction error

$$\mathcal{L}(h) = \mathbf{E}[\ell(\mathbf{y}(T), h(\{\mathbf{x}(s)\}_{0\le s<T}))$$

PAC bound: difffferntial equations

H_2 norm of predictors (control theory):

$$\sup_t |h(\{\mathbf{x}(s)\}_{0 \le s < t})| \le (H_2 \text{ norm}) \|\mathbf{x}\|_{L_2}.$$

Theorem (Rácz et. al, 2024)

- *Hypothesis class $\mathcal{H}$: stable bilinear differential equations whose H_2 norm is bounded by C.*

$$\mathbf{P}\left(\forall h \in \mathcal{H} : \mathcal{L}(h) - \hat{\mathcal{L}}_N(h) < C\frac{\ln(1/\delta)+1}{\sqrt{N}}\right) > 1-\delta$$

C- does not depent on T, depends on the maximal H_2 norm of the parametrization , decreases with stability.

Idea of the proof

Classical Rademacher complexity for i.i.d data + kernel trick:

$h(\{\mathbf{x}(s)\}_{0\le s<t})$ = Volterra kernel expansion (control theory):

$$\sum_{k=1}^{\infty}\sum_{i_1=1}^{m}\cdots\sum_{i_k=1}^{m}\int_0^t\int_0^{\tau_1}\cdots\int_0^{\tau_k} C(\theta)e^{A(\theta)(t-\tau_1)}N_{i_1}(\theta)e^{A(\theta)(\tau_1-\tau_2)}\times$$
$$\cdots\times N_{i_{k-1}}(\theta)B_{i_k}(\theta)\mathbf{x}_{i_1}(\tau_1)\cdots\mathbf{x}_{i_k}(\tau_k)d\tau_1\cdots d\tau_{i_k} =$$
$$< w_\theta, \Phi(\mathbf{x}) >_H$$

H – suitable Hilbert space (combination of L_2 and ℓ_2), Φ embedding of $\mathbf{x}$ into H.

Need stability for the scalar product $< w_\theta, \Phi(u) >_H$ to be well-posed, H_2 norm is the norm of w_θ in H.

PAC bound: Deep SSM

Predictors (in discrete time): deep SSM (Mamba) is a composition

$$h(\{\mathbf{x}(s)\}_{0 \le s < T}) = f_{\theta,1} \circ \cdots \circ f_{\theta,k}(\{\mathbf{x}(s)\}_{0 \le s < T})$$

where $f_{\theta,i}$ is either elementwise MLP:

$$f_{\theta,i}(\{\mathbf{x}(s)\}_{0 \le s < t)} = \text{MLP appplied to } \mathbf{x}(t),$$

or $f_{\theta,i}(\{\mathbf{x}(s)\}_{0 \le s < t}$ is a stable linear dynamical system:

$$\hat{\mathbf{s}}(t+1) = \hat{A}_i(\theta)\hat{\mathbf{s}}(t) + \hat{B}_i(\theta)\mathbf{x}(t), \ \hat{\mathbf{s}}(0) = 0,$$
$$f_{\theta,i}(\{\mathbf{x}(s)\}_{0 \le s < t}) = \hat{C}_i(\theta)\hat{\mathbf{s}}(t).$$

and $f_{\theta,i}$ for $i > 1$ is interpreted as a sequence-to-sequence map

tth element of the output sequence $=$

value of $f_{\theta,i}$ for the first t elements of the input sequence

$(\mathbf{x}, \mathbf{y})$ bounded random variable.

Learning problem for deep SSM

Training data: $(\mathbf{x}_1, \mathbf{y}_1), \ldots (\mathbf{x}_N, \mathbf{y}_N)$ distributed as $(\mathbf{x}, \mathbf{y})$.

Empirical error:

$$\hat{\mathcal{L}}_N(h) = \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{y}_i(T), h(\{\mathbf{x}_i(s)\}_{0 \le s < T})))$$

True error final-time prediction error

$$\mathcal{L}(h) = \mathbf{E}[\ell(\mathbf{y}(T), h(\{\mathbf{x}(s)\}_{0 \le s < T}))$$

$\ell - \ell_2$, ℓ_1

PAC bound: Deep SSM

Theorem (Rácz et. al, 2024)

$$\mathbf{P}\left(\forall h \in \mathcal{H} : \mathcal{L}(h) - \hat{\mathcal{L}}_N(h) < C\frac{\ln(1/\delta)+1}{\sqrt{N}}\right) > 1-\delta$$

C- does not depent on T, grows with the depth of the model (deep SSM), decreases with stability (H_2 norm) of the linear layers.

Proof: Rademacher complexity, extension of contraction lemma (Rademacher contraction), linear systems are bounded linear operators

Main idea of the proof

$$\hat{\mathbf{s}}(t+1) = \hat{A}_i(\theta)\hat{\mathbf{s}}(t) + \hat{B}_i(\theta)\mathbf{x}(t), \ \hat{\mathbf{s}}(0) = 0,$$
$$f_{\theta,i}(\{\mathbf{x}(s)\}_{0 \le s < t}) = \hat{C}_i(\theta)\hat{\mathbf{s}}(t).$$

H_2 norm of the ith layer (control theory): norm of the linear system as a bounded operator from inputs to outputs:

$$\text{Rademacher complexity of the } i\text{th layer} \le$$
$$\text{maximal } H_2 \text{ norm of the } i\text{th layer} \times$$
$$\times \text{ Rademacher complexity of the previous layers}$$

Extension of existing results

Model class	Unbounded data	Dependent data	Independence of trajectory length
PAC-Bayes for linear RNN [Eringis et al., 2024]	✓	✓	✓
PAC-Bayes for RNN [Eringis et al. 2024]		✓	✓
PAC for deep SSM [Rácz et al. 2024]			✓
PAC for selective SSM neural ODE, [Rácz et al. 2024]			✓
Algorithm specific bound for linear swicthed RNN [Rácz et al. 2025]		✓	✓

Conclusions

- We have formulated PAC(-Bayesian) bounds for dynamical systems.
- Main challenges:
  - non i.i.d. data,
  - unbounded data,
  - varying number of past observations.
- Stability is crucial.

Proof: KL divergence for nonlinear systems

Introduce empirical loss with infinite past:

$$V_N(\theta) = \frac{1}{N}\sum_{t=0}^{N-1} \|\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t)\|_2^2$$

Error systems:

$$\tilde{\mathbf{s}}_e(t+1) = f_{\theta,e}(\tilde{\mathbf{s}}_e(t), \mathbf{e}_g(t)), \;\; \tilde{\mathbf{s}}_e(0) \text{ suitable}$$
$$\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid 0) = h_{\theta,e}(\tilde{\mathbf{s}}_e(t), \mathbf{e}_g(t))$$

$\tilde{\mathbf{s}}_e(t)$ is composed of the state of the data generator and the model.

Error system is stable: there exists a unique solution $\tilde{\mathbf{s}}_s(t)$ for $t \in \mathbb{Z}$ to which any solution converges exponentially

$$\tilde{\mathbf{s}}_s(t+1) = f_{\theta,e}(\tilde{\mathbf{s}}_s(t), \mathbf{e}_g(t))$$
$$\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t) = h_{\theta,e}(\tilde{\mathbf{s}}_s(t), \mathbf{e}_g(t))$$

Proof: KL divergence for nonlinear systems

Donsker-Varadhan change of measure inequality for KL divergence:
With probability at least $1-\delta$

$$\forall \rho \in \mathcal{M}_\pi: \; E_{\theta\sim\rho} V_N(\theta) \le \hat{\mathcal{L}}_N(\theta) + \frac{1}{\lambda}\left(D_{\mathrm{KL}}(\rho \| \pi) + \ln \frac{1}{\delta} + \Psi_1(\lambda, \pi, N) \right)$$

$$\forall \rho \in \mathcal{M}_\pi: \; E_{\theta\sim\rho} \mathcal{L}(\theta) \le V_N(\theta) + \frac{1}{\lambda}\left(D_{\mathrm{KL}}(\rho \| \pi) + \ln \frac{1}{\delta} + \Psi_2(\lambda, \pi, N) \right)$$

$$\Psi_1(\lambda, \pi, N) = \ln E_{\theta\sim\pi} \mathbf{E}[e^{2\lambda(V_N(\theta) - \hat{\mathcal{L}}_N(\theta))}]$$
$$\Psi_2(\lambda, \pi, N) = \ln E_{\theta\sim\pi} \mathbf{E}[e^{2\lambda(\mathcal{L}(\theta) - V_N(\theta))}],$$

Proof: KL divergence for nonlinear systems

Consider the error system:

$$\tilde{\mathbf{s}}_s(t+1) = f_{\theta,e}(\tilde{\mathbf{s}}_s(t), \mathbf{e}_g(t))$$
$$\mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t) = h_{\theta,e}(\tilde{\mathbf{s}}_s(t), \mathbf{e}_g(t))$$

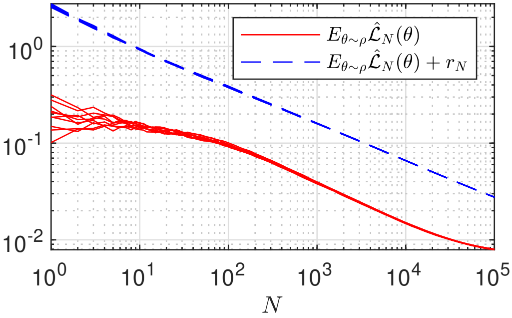
Its output is weakly dependent [Dedecker & Doukhan] with mixing coefficient $\theta_{\infty,N}(1)$ which can be computed from the error system.

Apply McDiarmid for weakly dependent processes [Rio 2000]

$$\mathbf{E}[e^{\lambda(\mathcal{L}(\theta) - V_N(\theta))}] \leq e^{\frac{\lambda^2 C_1(\theta)^2}{N}}$$

Direct computation using control theory:

$$\mathbf{E}[e^{\lambda(\mathcal{L}_N(\theta) - V_N(\theta))}] \leq e^{\frac{\lambda C_2(\theta)^2}{N}}$$



randomly generated (state-convergent) Data Generator:

$$\mathbf{s}_g(t+1) = \text{ReLu}\left(\begin{bmatrix} 0.52 & 0.23 \\ 0.23 & -0.52 \end{bmatrix} \mathbf{s}_g(t) + \begin{bmatrix} -0.82 & -0.45 \\ 0.36 & -0.96 \end{bmatrix} \mathbf{e}_g(t) + \begin{bmatrix} 0.38 \\ -0.06 \end{bmatrix}\right)$$

$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{x}(t) \end{bmatrix} = \tanh\left(\begin{bmatrix} 0.05 & -0.10 \\ -0.11 & 0.01 \end{bmatrix} \mathbf{s}_g(t) + \begin{bmatrix} 0.09 & -0.11 \\ 0.05 & -0.16 \end{bmatrix} \mathbf{e}_g(t) + \begin{bmatrix} -0.53 \\ -0.79 \end{bmatrix}\right)$$

Predictors: like the generator, i.e. using ReLu and tanh activation functions, and 2 hidden states, everything is parametrised.

MACHINE LEARNING MEETS MICROBIOLOGY: CHALLENGES AND OPPORTUNITIES

Anna Kerekes

FROM MATH TO MACHINE LEARNING: MY RESEARCH JOURNEY

Education

- BSc & MSc in Mathematics
- PhD Student at ETH Zürich
- Currently on an exchange year at the Max Planck Institute, Tübingen

Research Interests

- Theoretical Machine Learning
  - Why does Adam lead to good minima in multi-task learning?
  - Understanding causality in time series
- Machine Learning in Healthcare
  - Developing new treatments and diagnostic tools using ML

MICROBIAL COMMUNITIES IN THE HUMAN BODY

What is a Microbial Community?

A group of microorganisms—such as bacteria, fungi, or viruses—that live together in a shared environment and interact with one another.

Where Are They Found in the Body?

- Gut
- Skin
- Mouth
- Vaginal tract (and more)

MICROBIAL COMMUNITIES IN THE HUMAN BODY

Why Do They Matter?

- Our health is tightly connected to the composition and function of these microbiomes
- Gut microbiome is linked to insulin resistance [1]
- Vaginal microbiome is linked to reproductive health [2]

How Are Microbiomes Characterized?

- Cultures
- Sequencing-based approaches:
    - 16S rRNA sequencing
    - Whole genome sequencing

[1] Caricilli AM, Saad MJ. The role of gut microbiota on insulin resistance. Nutrients. 2013 Mar 12;5(3):829-51. doi: 10.3390/nu5030829. PMID: 23482058; PMCID: PMC3705322.
[2] Gao, X., Lee, V., Prom-Wormley, E. et al. The Vaginal Microbiome: Disease, Genetics and the Environment. Nat Prec (2011). https://doi.org/10.1038/npre.2011.5150.2

16S rRNA in Bacterial Identification

Part of the rRNA sequence of bacteria
- Contains **highly conserved regions** → allows reliable detection across species.
- Contains **highly variable regions** → enables classification and differentiation of bacteria.

Properties:
- **Cheap to sequence**
- **High noise levels**
- **Limited resolution**

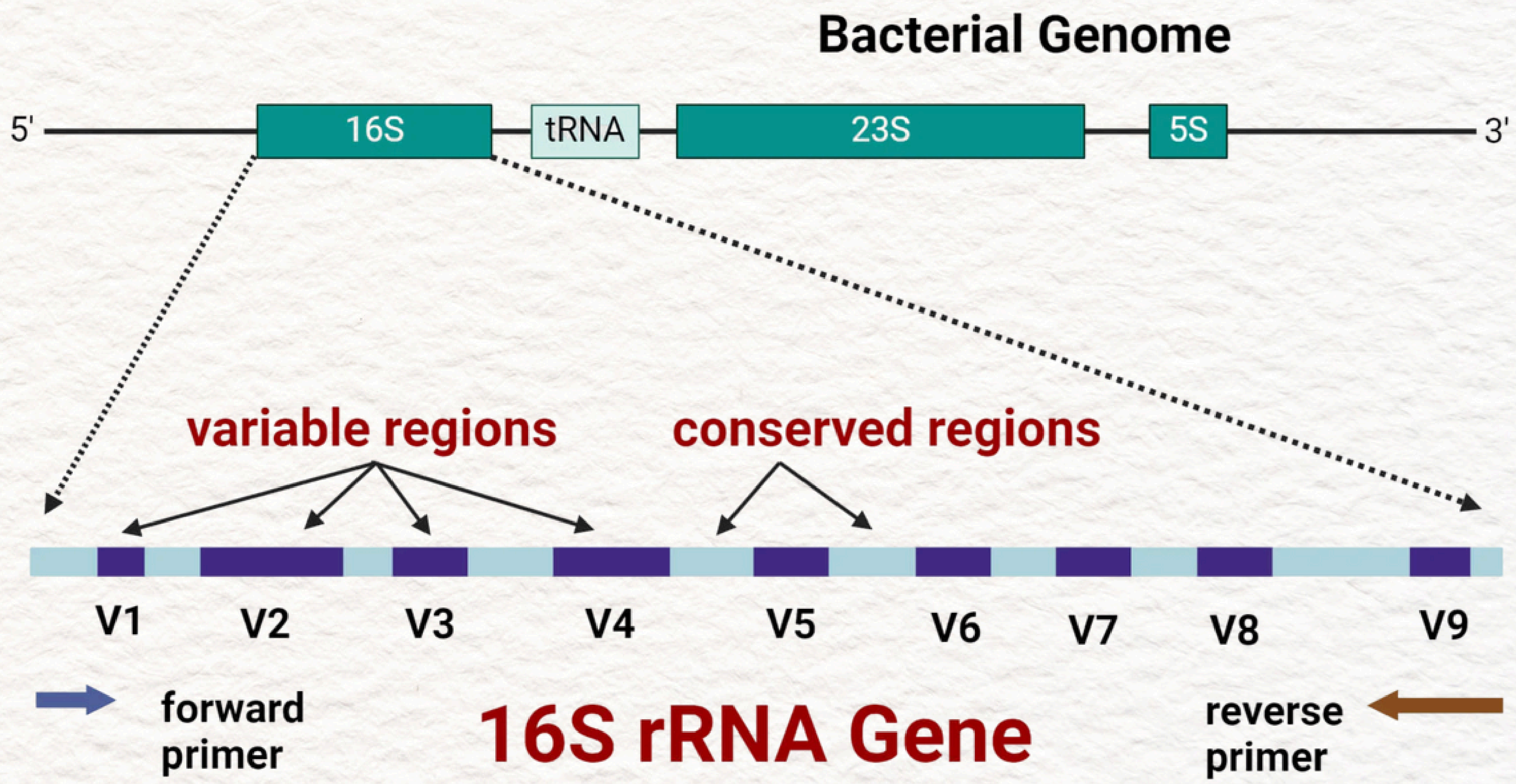


Image came from: https://microbenotes.com/16s-rrna-gene-sequencing/

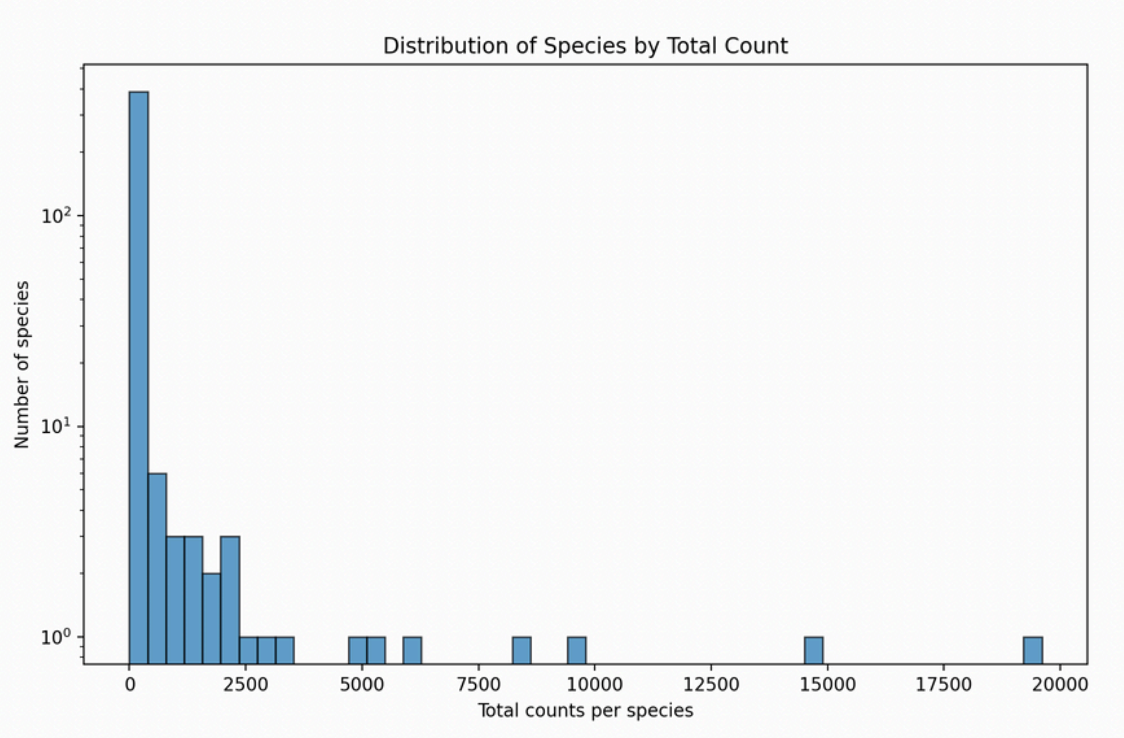
16S rRNA in Bacterial Identification

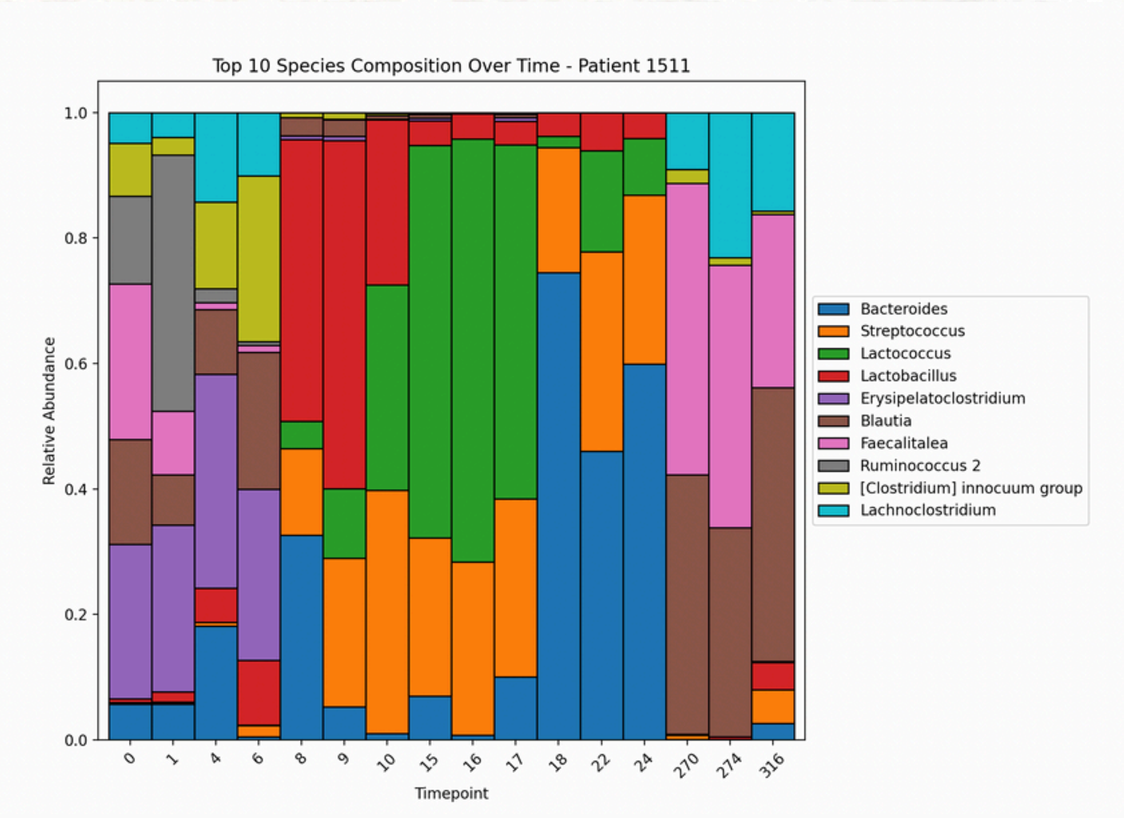
Data Structure:

- Count table with **high dimensionality**
- One sample is a very long vector (~400 species)

Data Characteristics:

- **Sparse vector** = most values are 0 or close to 0
- Some values are very large
- Counts ≠ number of bacteria → **data usually needs to be normalized**





INTERACTION GRAPHS

Assumptions:

- Most distribution patterns can be described by ecological reasons
- Co-occurrence → positive relationship (e.g., cross-feeding, co-aggregation in biofilms)
- Mutual exclusion → negative relationship (e.g., predator-prey, competition)

Graphs are built from **similarities**.

Types:

- Simple edges
- More complex edges

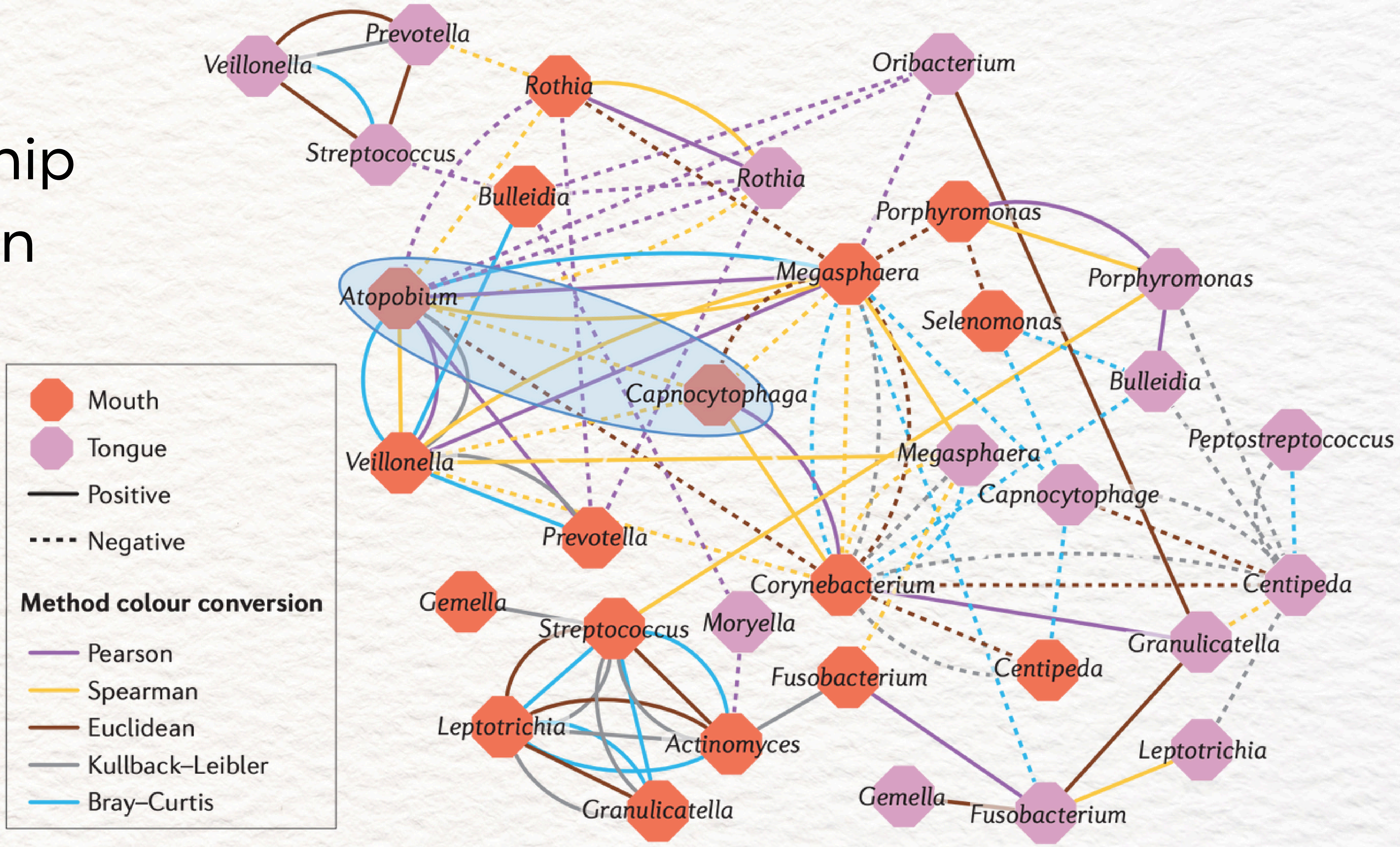


Image came from: https://microbenotes.com/16s-rrna-gene-sequencing/

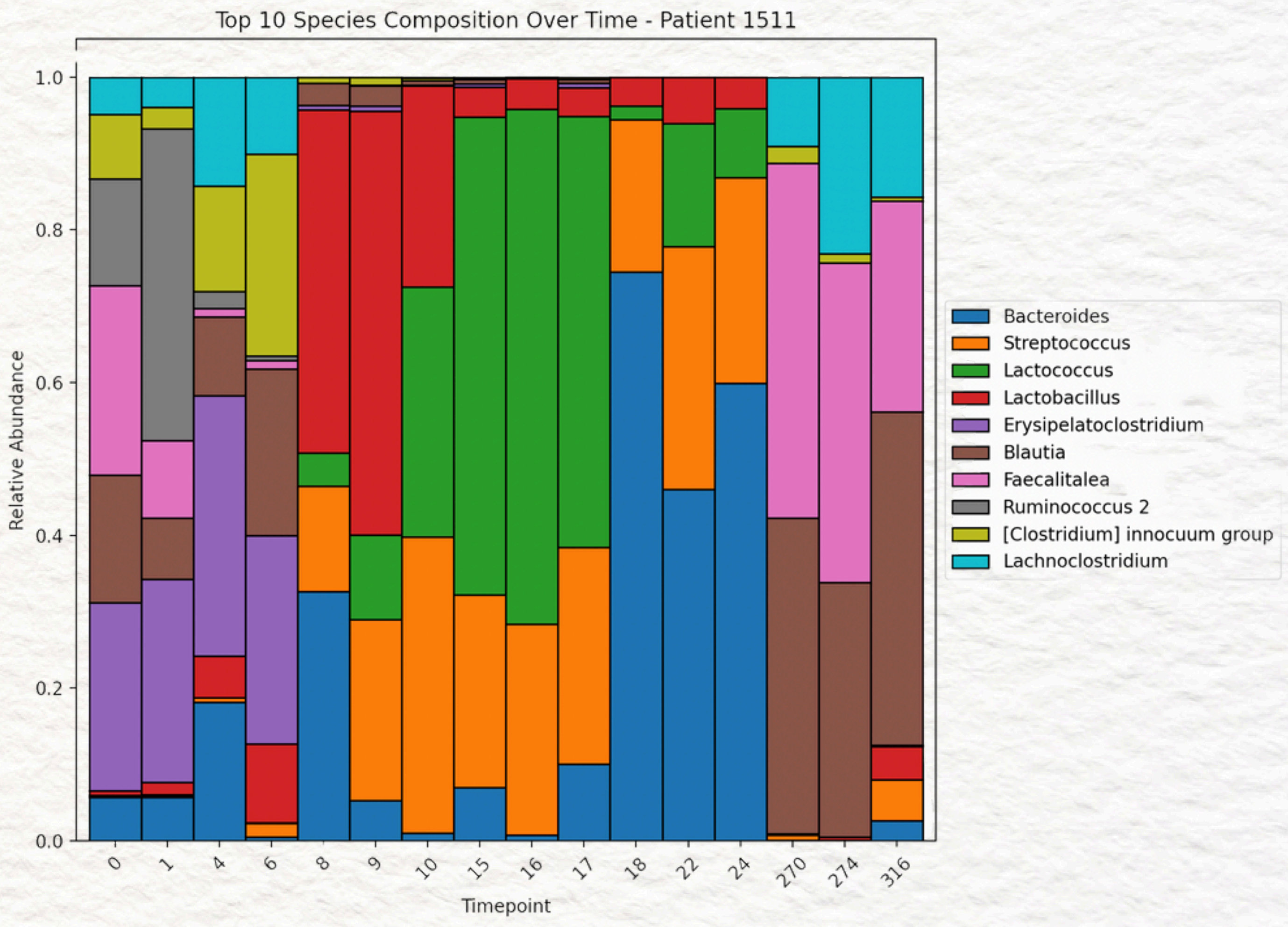
TIME SERIES MODELING

Context:

- Multiple samples from one patient
- Usually a limited number of timepoints and patients (e.g., 10 timepoints, 1000 patients)
- Normalized data (16S rRNA)
- Interest in interaction graphs

Questions & Challenges (Time Series Forecasting with Neural Networks):

- Interactions might be harder to extract
- Might not be enough data
- Noise can be an issue
- The approach is not yet widely applied

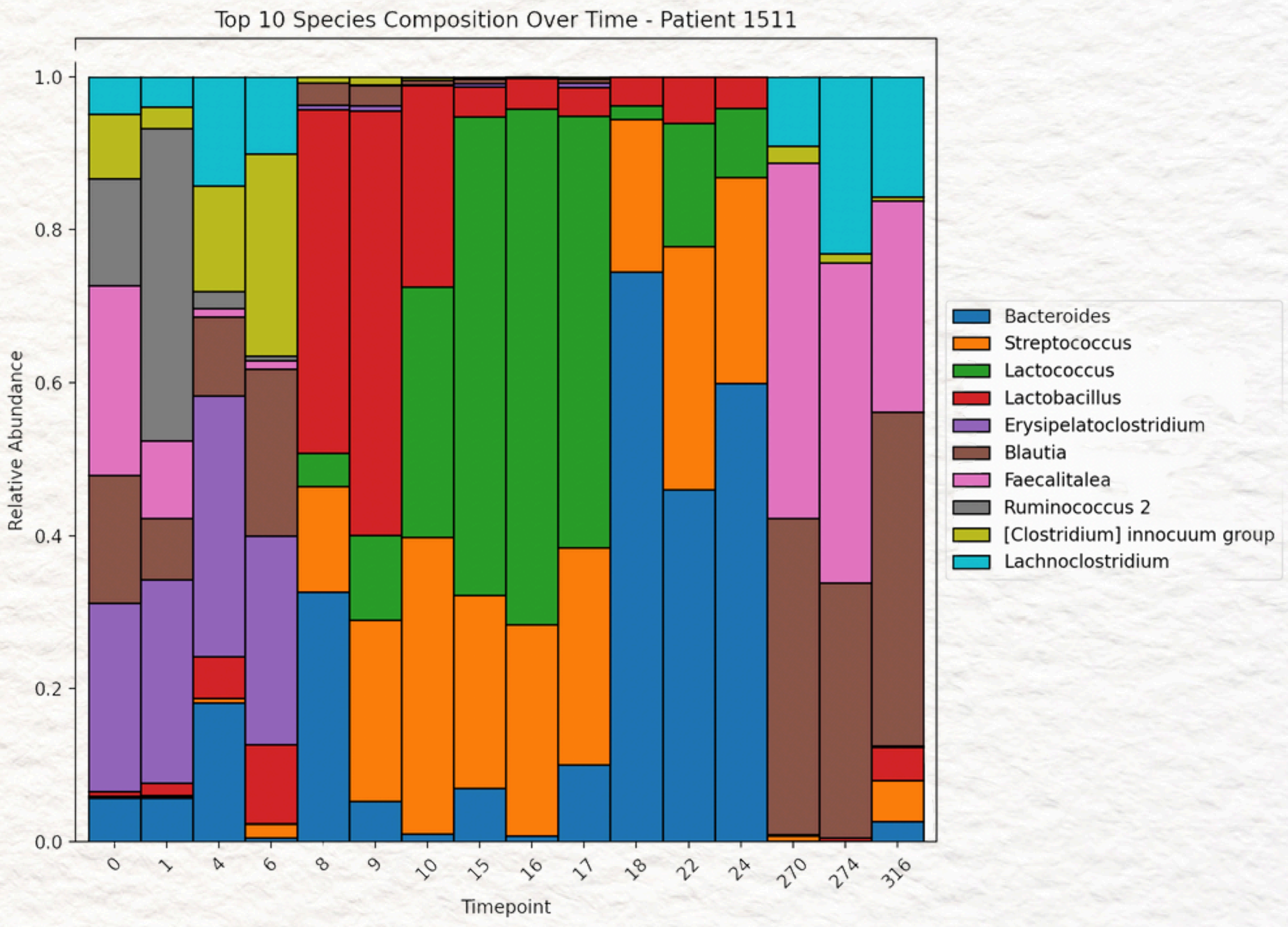


TIME SERIES MODELING: LOTKA-VOLTERRA EQUATION

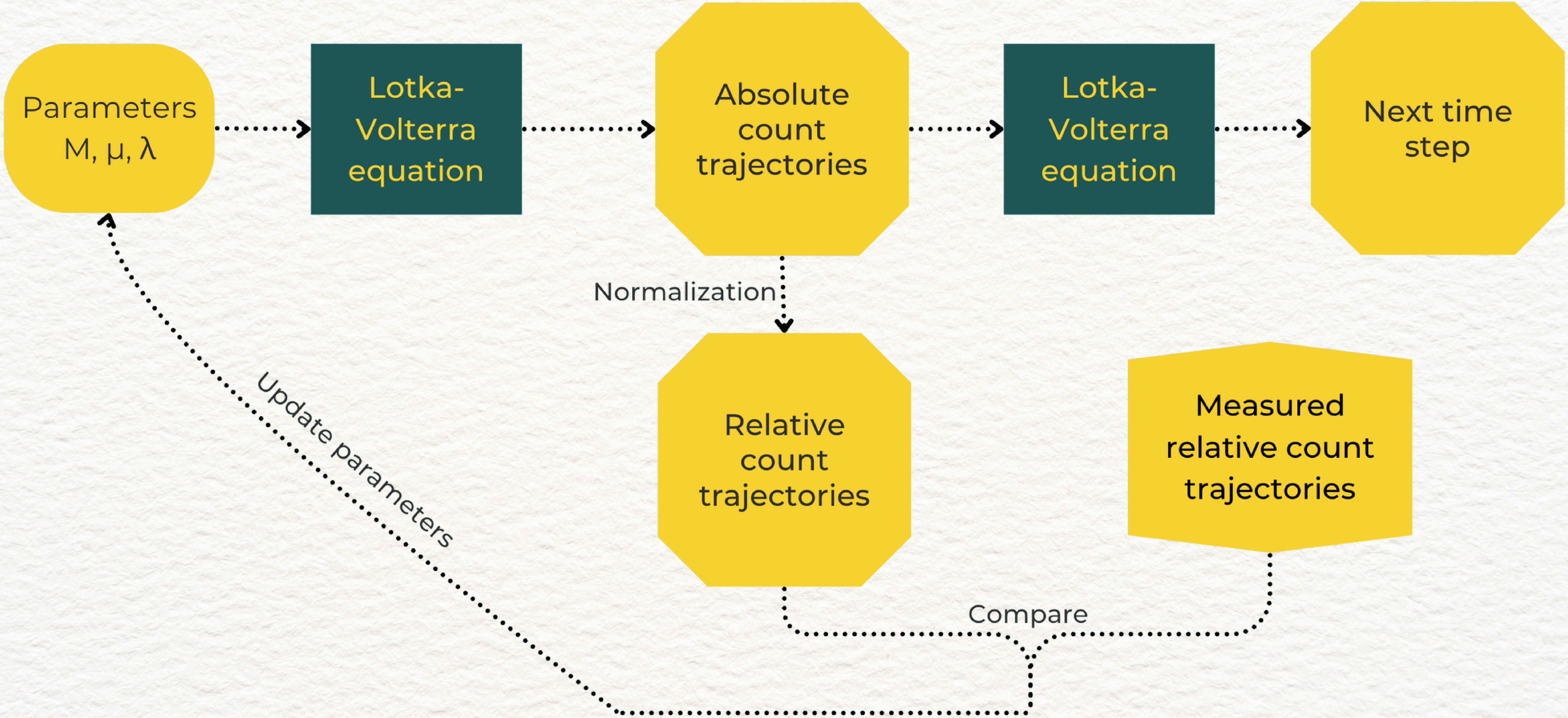
Alternative method:

- Lotka-Volterra equation → a simple differential equation popular in the ecological community
- Works on **absolute count** data
- Our data: **relative counts**, which require adjustment before applying this method

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = \mu_i x_i(t) + x_i(t) \sum_{j=1}^{L} M_{ij} x_j(t)$$

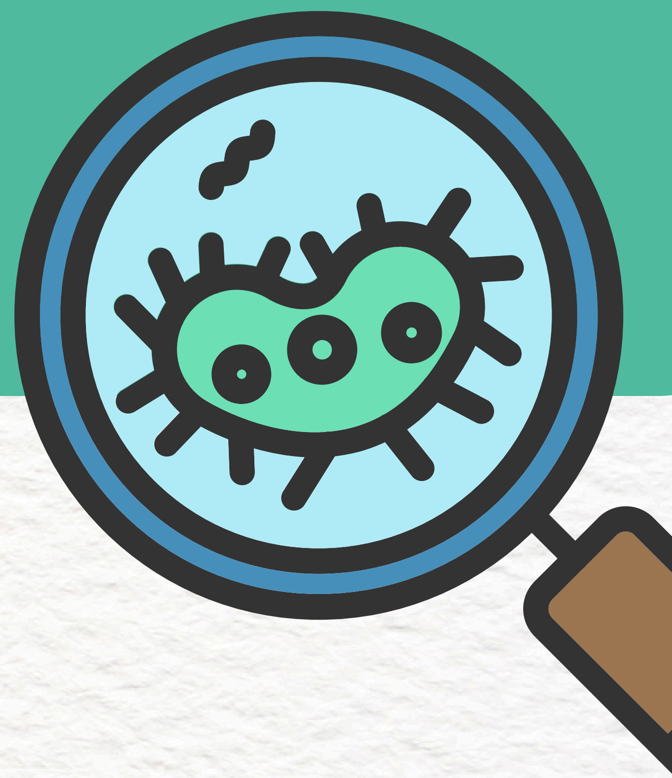


A SOLUTION TO RELATIVE COUNTS



Relative Lotka-Volterra: A Novel Approach for Modeling Microbiome Dynamics from Relative Abundance Data Tristan Gollmart, Anna Kerekes, André Kahles, Gunnar Rätsch
https://www.bc2.ch/programme

SUMMARY EXAMPLE



EXAMPLE PROJECT

Background:

- UTIs are a common and recurrent condition
- Standard treatment involves antibiotics
- High recurrence rates remain a challenge
- **Project Goal:** Develop a novel, more effective treatment

What is available?

- 16S rRNA time series from many patients
- Some other modalities, e.g., cultures, clinical data

What Can We Contribute as Machine Learning Experts?

Background information came from: https://theloopzurich.ch/en/projekte/mtorus/

WHAT CAN WE DO AS MACHINE LEARNING EXPERTS?

Classification

- Build models that find patients who will relapse (from one sample or time series)

Graphs

- Find similarity-based graphs
- Find causal graphs?

Time-series

- Learning to forecast microbial communities (normalized, noisy, short, sparse, and high-dimensional time series)
- Can we learn forecasting models that also learn interpretable interactions? (LV)
- Incorporate uncertainty (noisy data)
- Build forecasting models that incorporate multi-modality (e.g., cultures, clinical data)

And so much more....

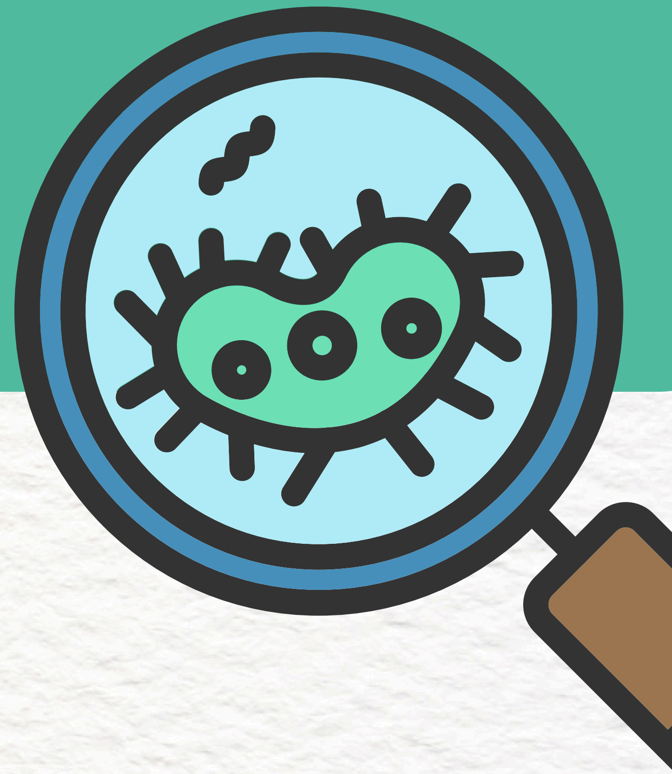
Background information came from: https://theloopzurich.ch/en/projekte/mtorus/

COLLABORATORS

Thank you to Tristan Gollmart, André Kahles and Gunnar Rätsch!

THANK YOU!

QUESTIONS?



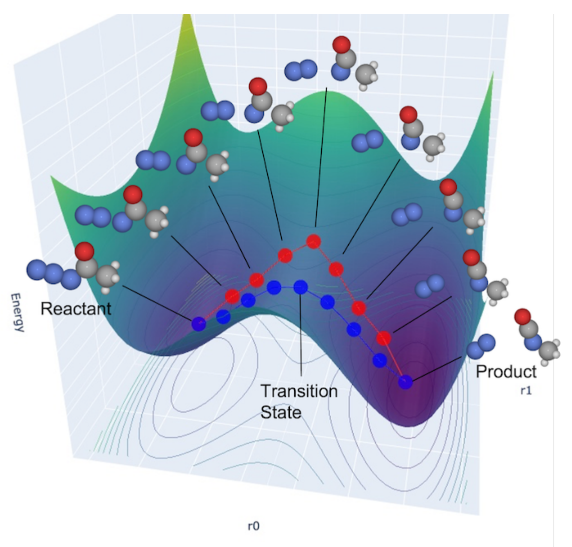
Anna Kerekes: akerekes@ethz.ch

UNIVERSITY OF CAMBRIDGE
Department of Engineering

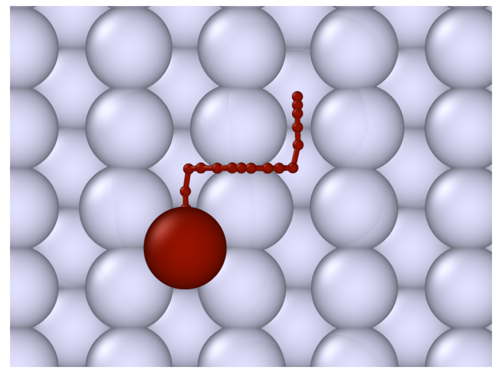
Optimal Transport for Atom Assignment in Materials Chemistry

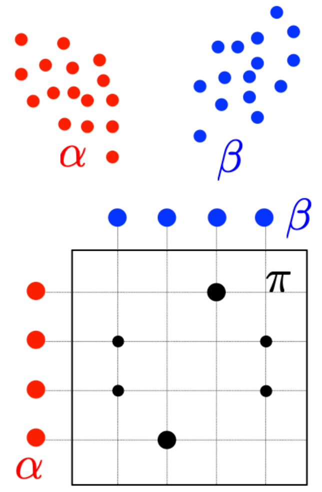
Tamás K. Stenczel, 2025 Aug 13, Hungarian Machine Learning Days

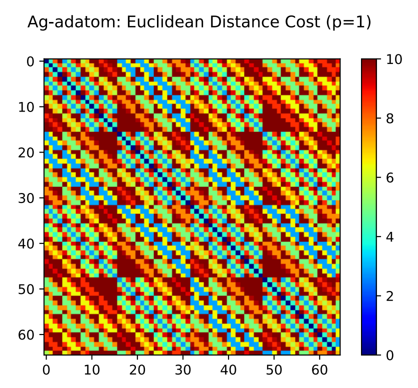
Roadmap



- Motivation & problem
- Background: chemical reaction modelling
- Toy problem: Ag adatom
- Optimal transport & adaptations
- 3D problem: Si double interstitial structures
- Outstanding questions & ongoing work
- Conclusion







Motivation & Goal

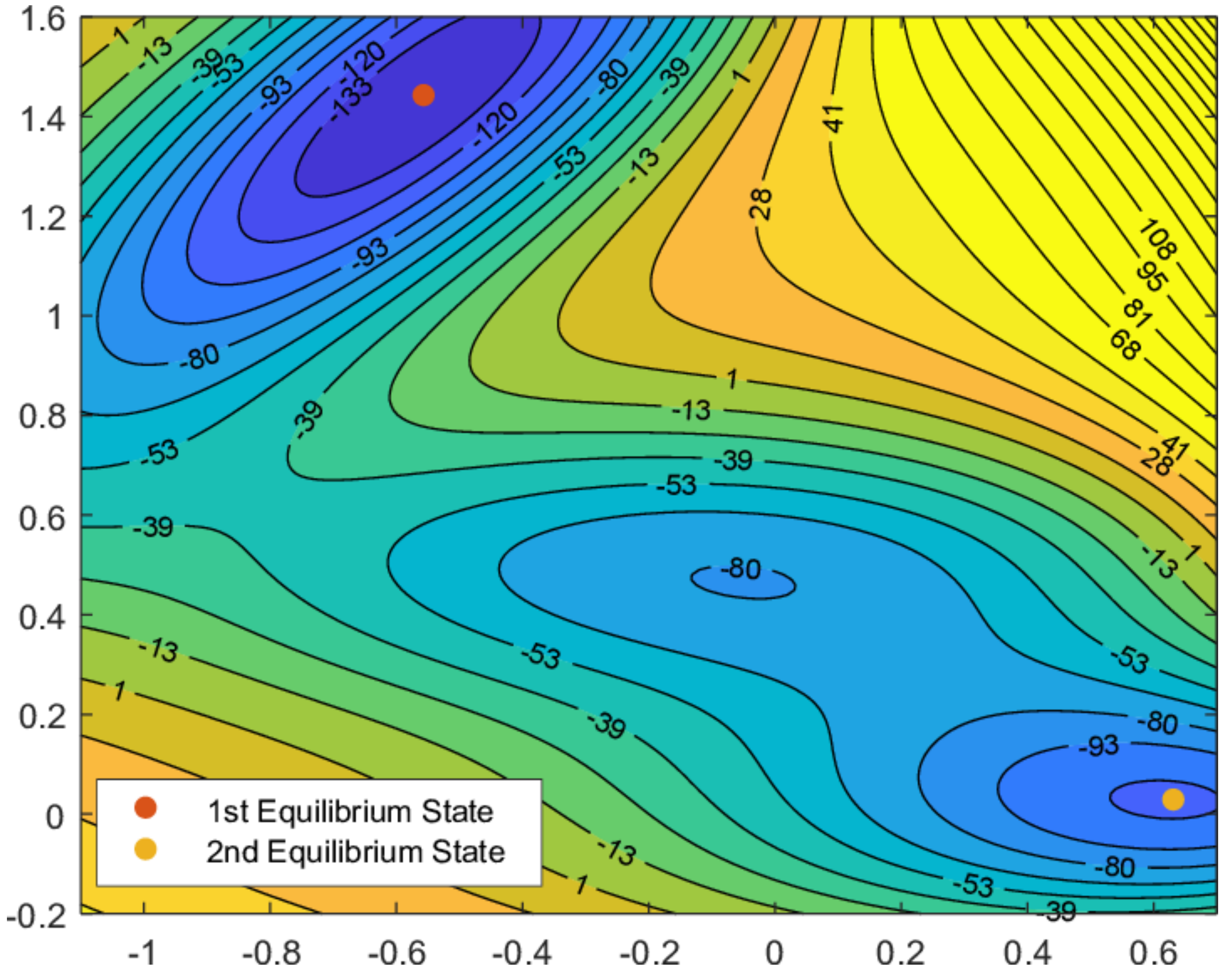
- Find atom-assignment between reactant & products
  - Most research takes it for granted
  - Obvious / by-hand / not attempted
- Can test assignments, combinatorial number to try
- Goal: Have a small number (~10) of assignments to try

n	***n*!**
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5 040
8	40 320

Chemical reactions

Chemical Reactions

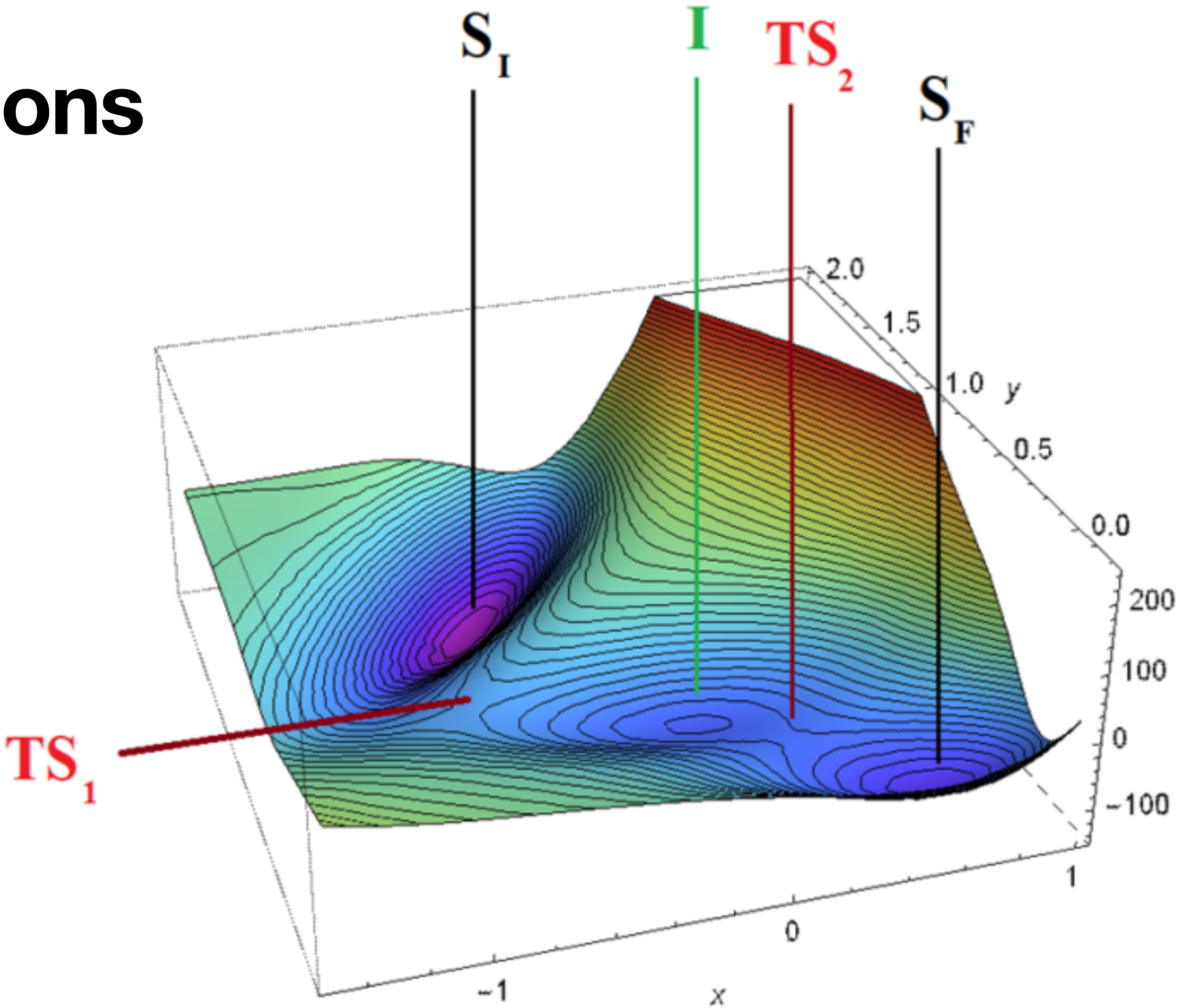
“Just like hiking”



Chemical Reactions

"Just like hiking"

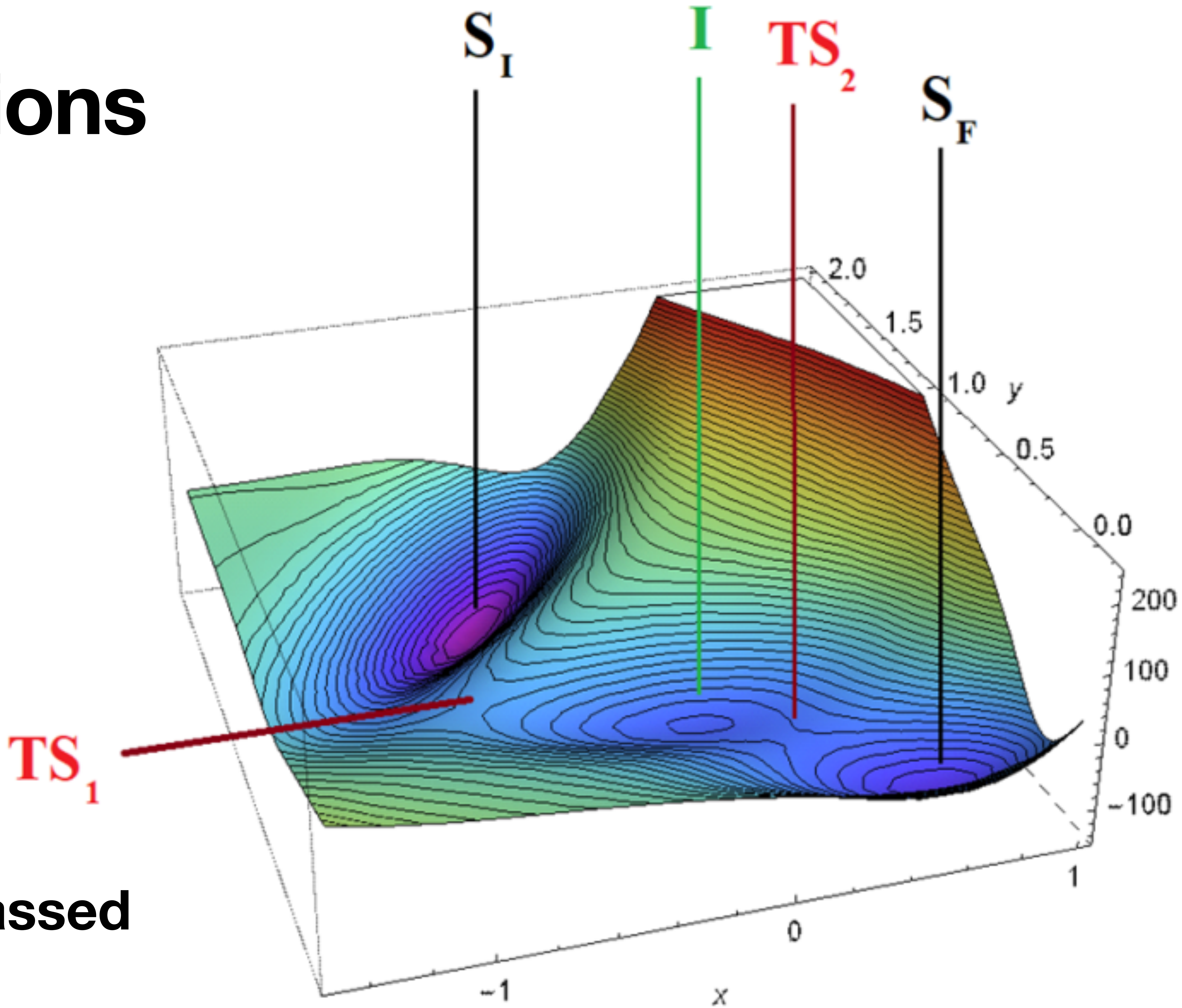
- Can walk long distances
- Don't like climbing
- Readily slide down hillsides



Chemical Reactions

“Just like hiking”

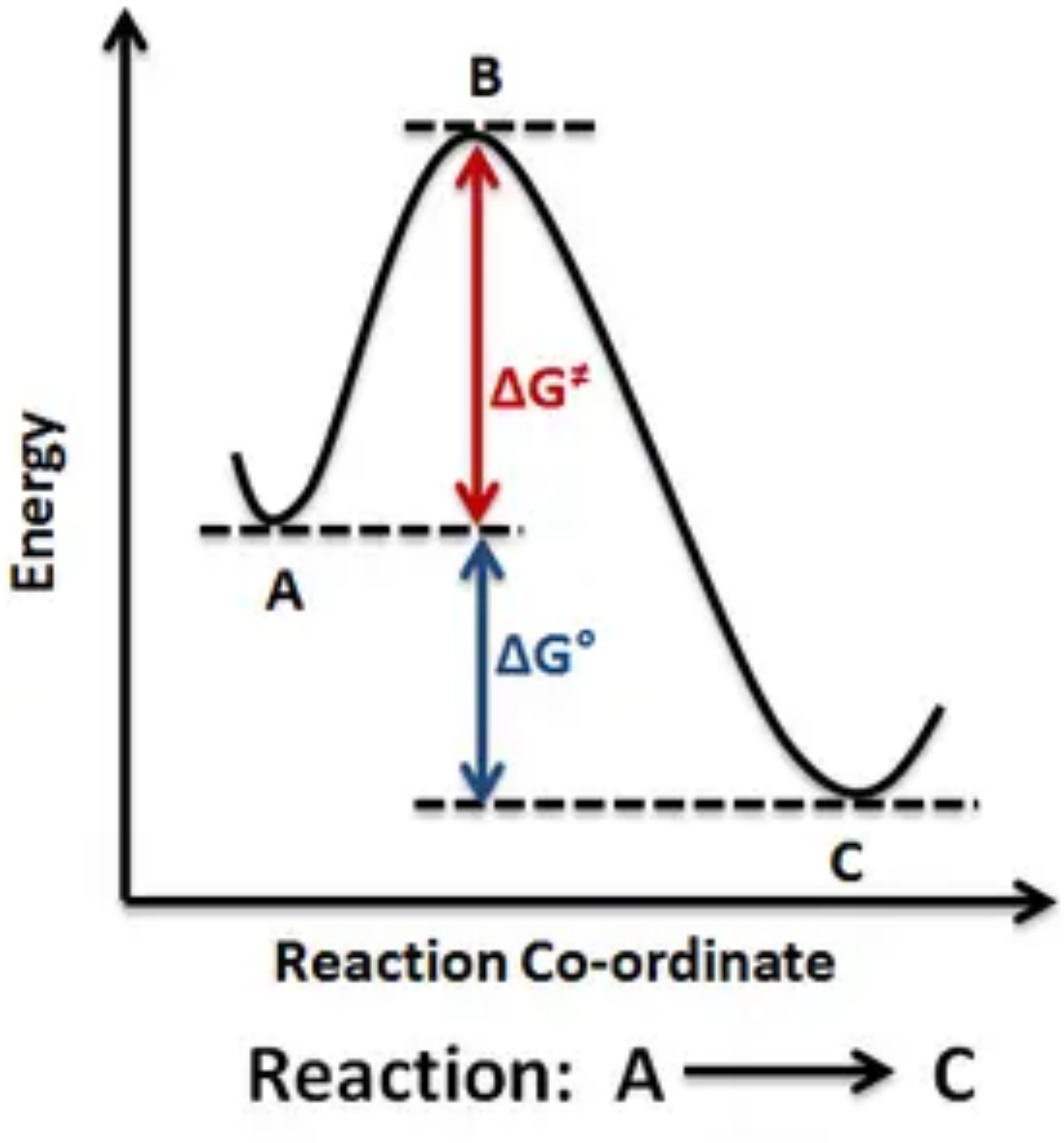
- Can walk long distances
- Don’t like climbing
- Readily slide down hillsides



Looking for mountain passed

Chemical Reactions

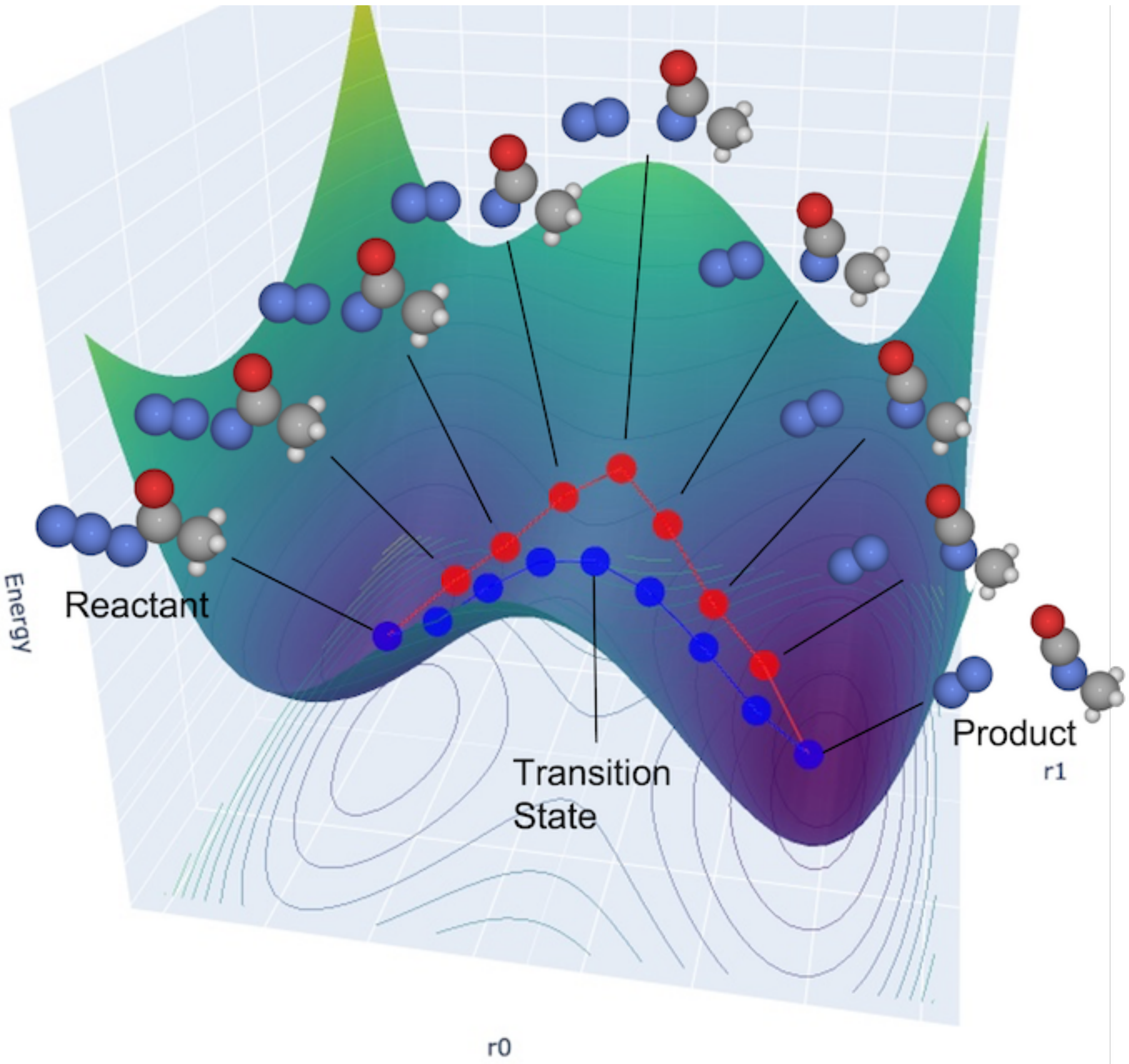
- Goals:
  - Energy difference
  - Barrier & mechanism
- Using:
  - Reactant & product structure
  - Energy model (e.g. ab-initio, MACE, etc.)



Chemical Reactions

Path refinement

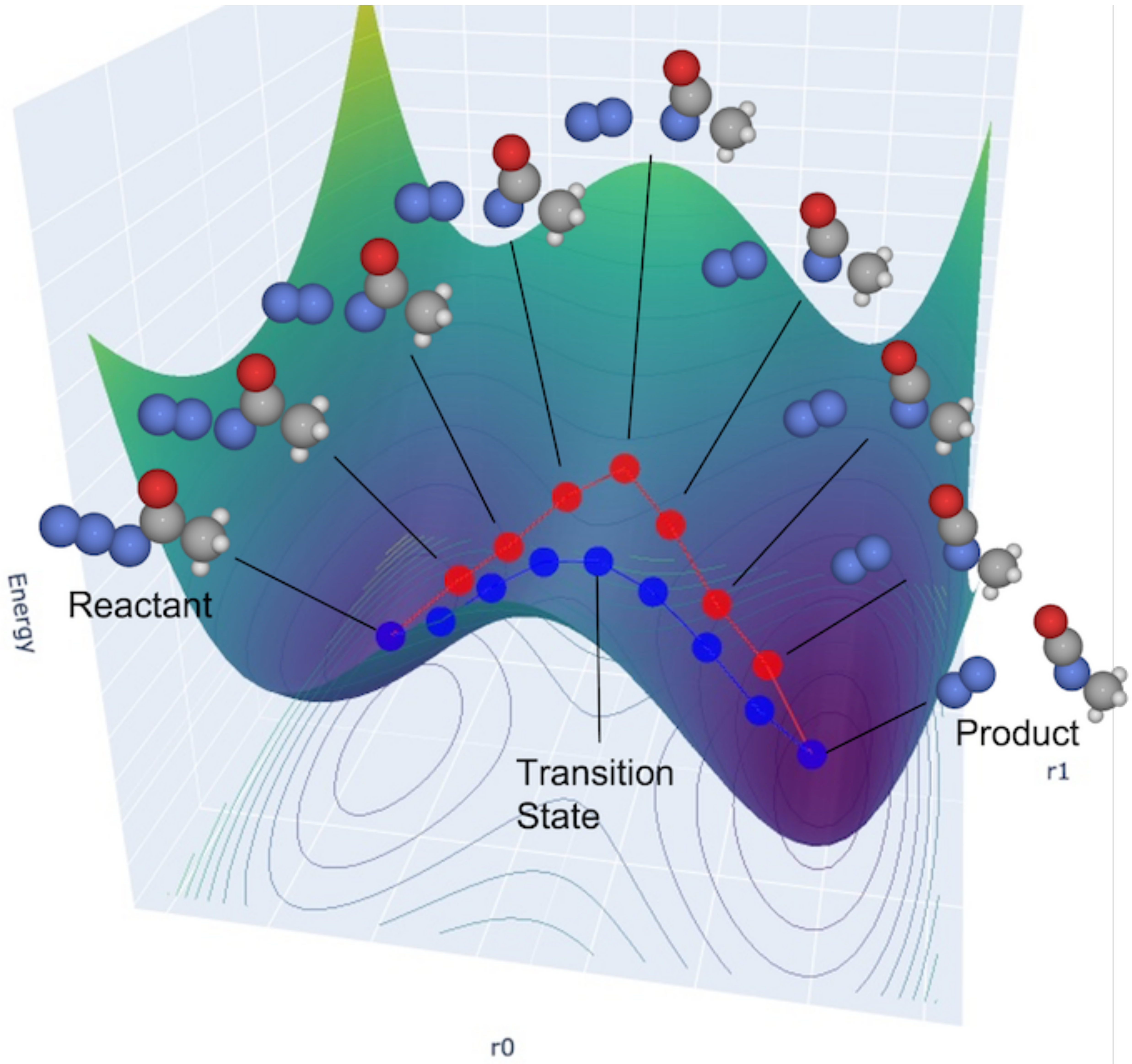
- Path: array of images
- Locally variational
- Many mature methods: NEB, GSM



Chemical Reactions

Path refinement & initialisation

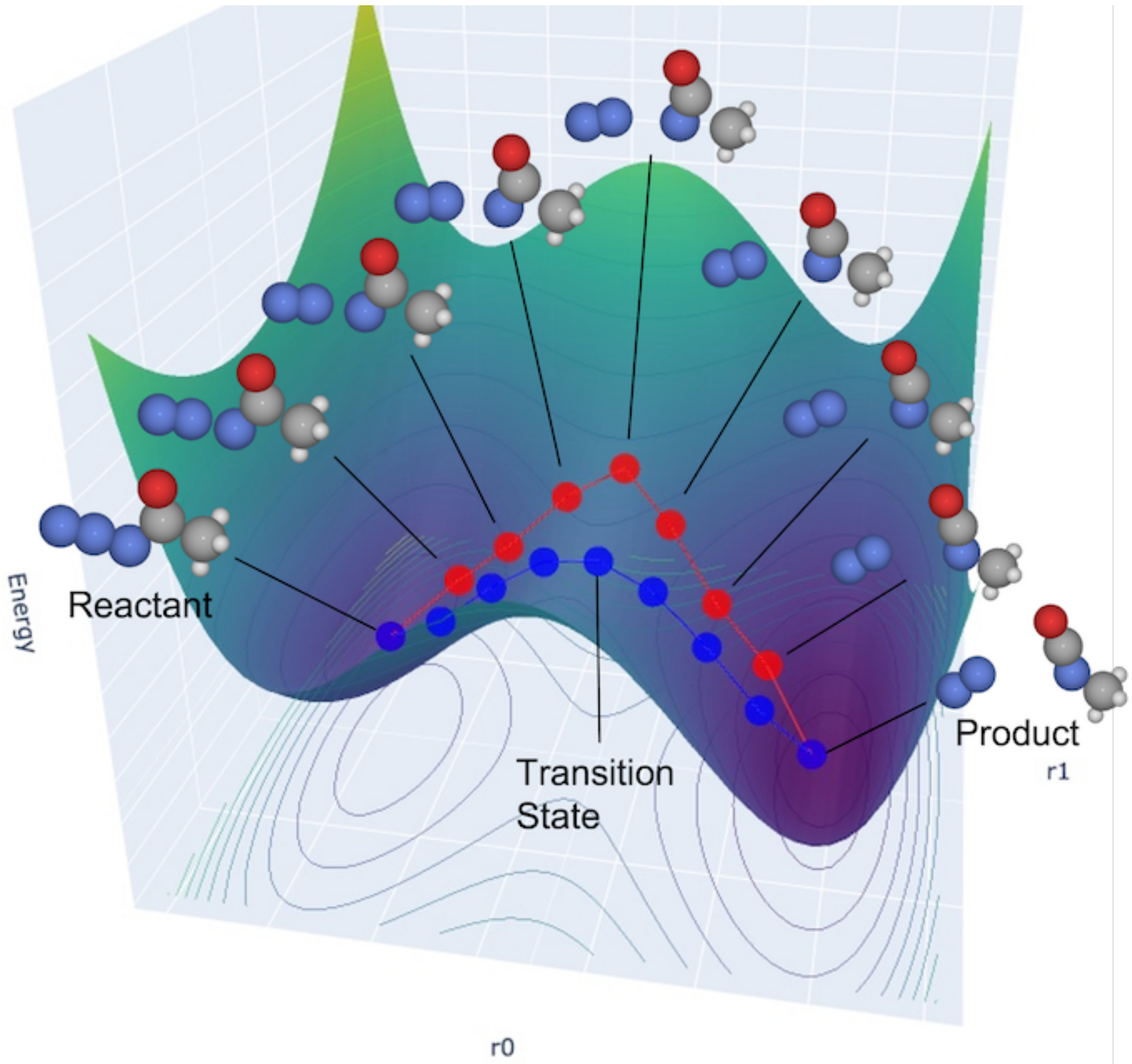
- Path: array of images
- Locally variational
- Many mature methods: NEB, GSM
- Initialisation: interpolation, geodesic paths, recent MACE-based methods



Chemical Reactions

Path refinement & initialisation

- Path: array of images
- Locally variational
- Many mature methods: **NEB**, GSM
- Initialisation: interpolation, geodesic paths, recent MACE-based methods

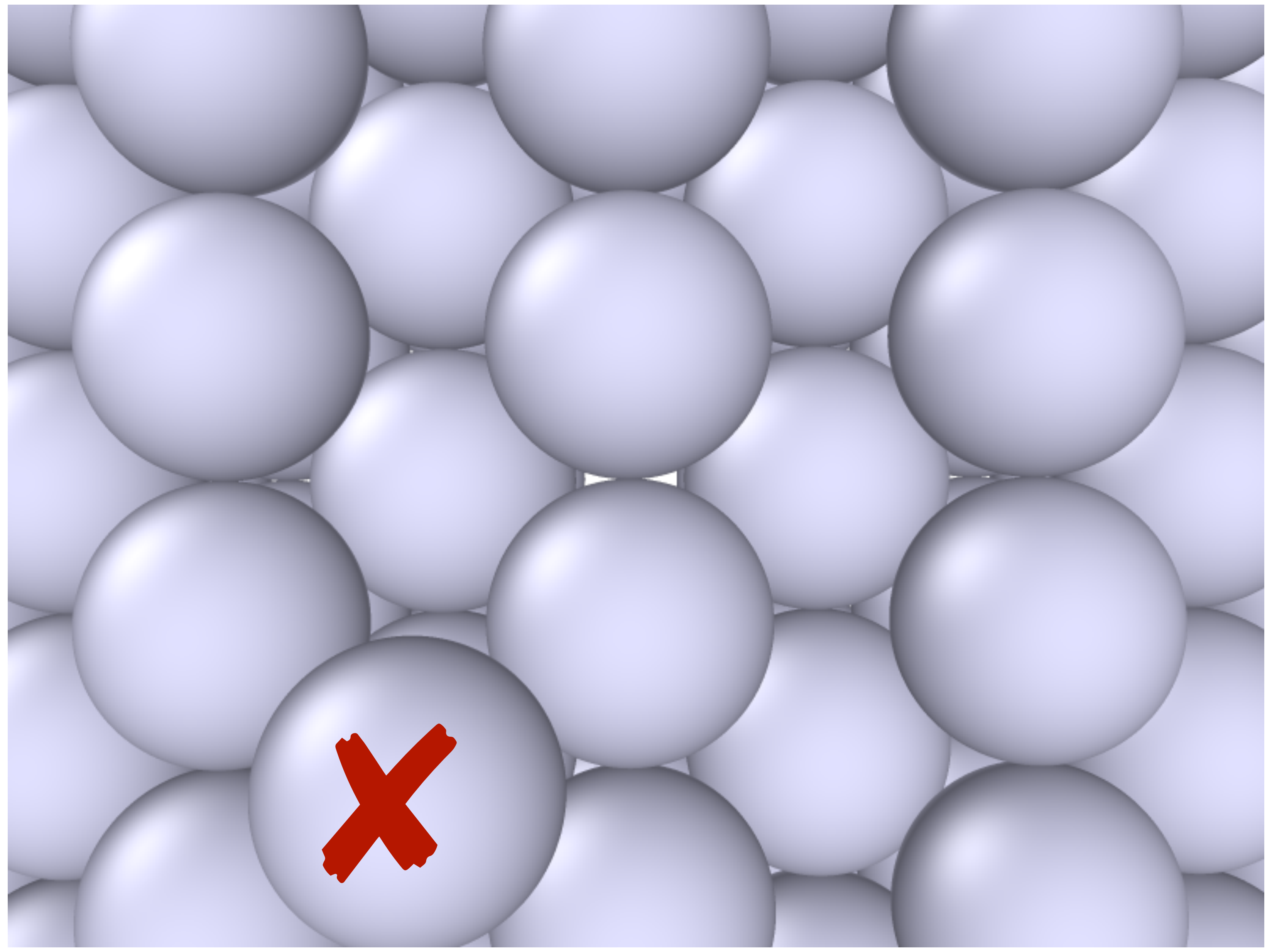


Used as a tool here.

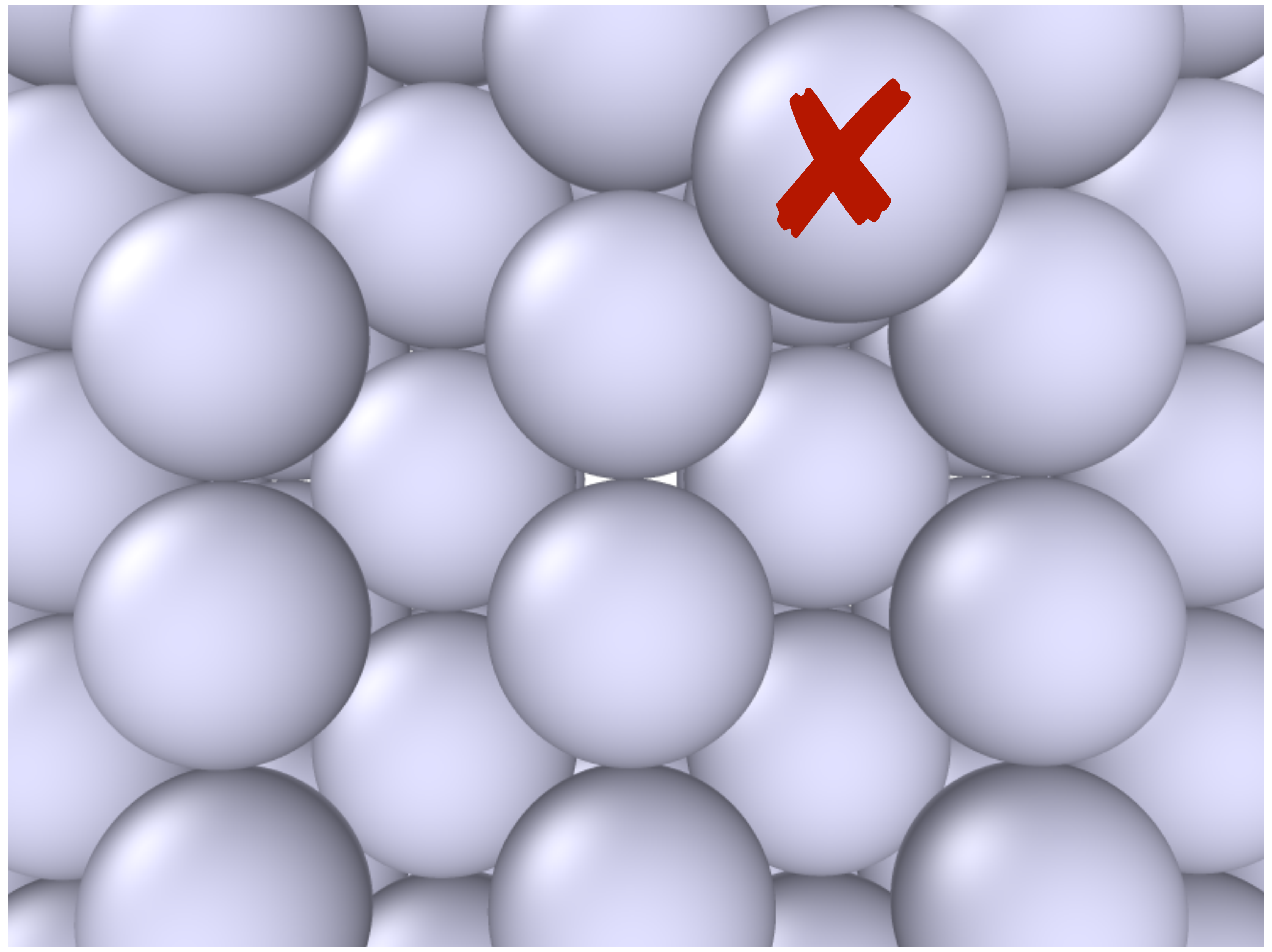
Ag adatom

Extra Ag atom movement on Ag surface

"toy problem"

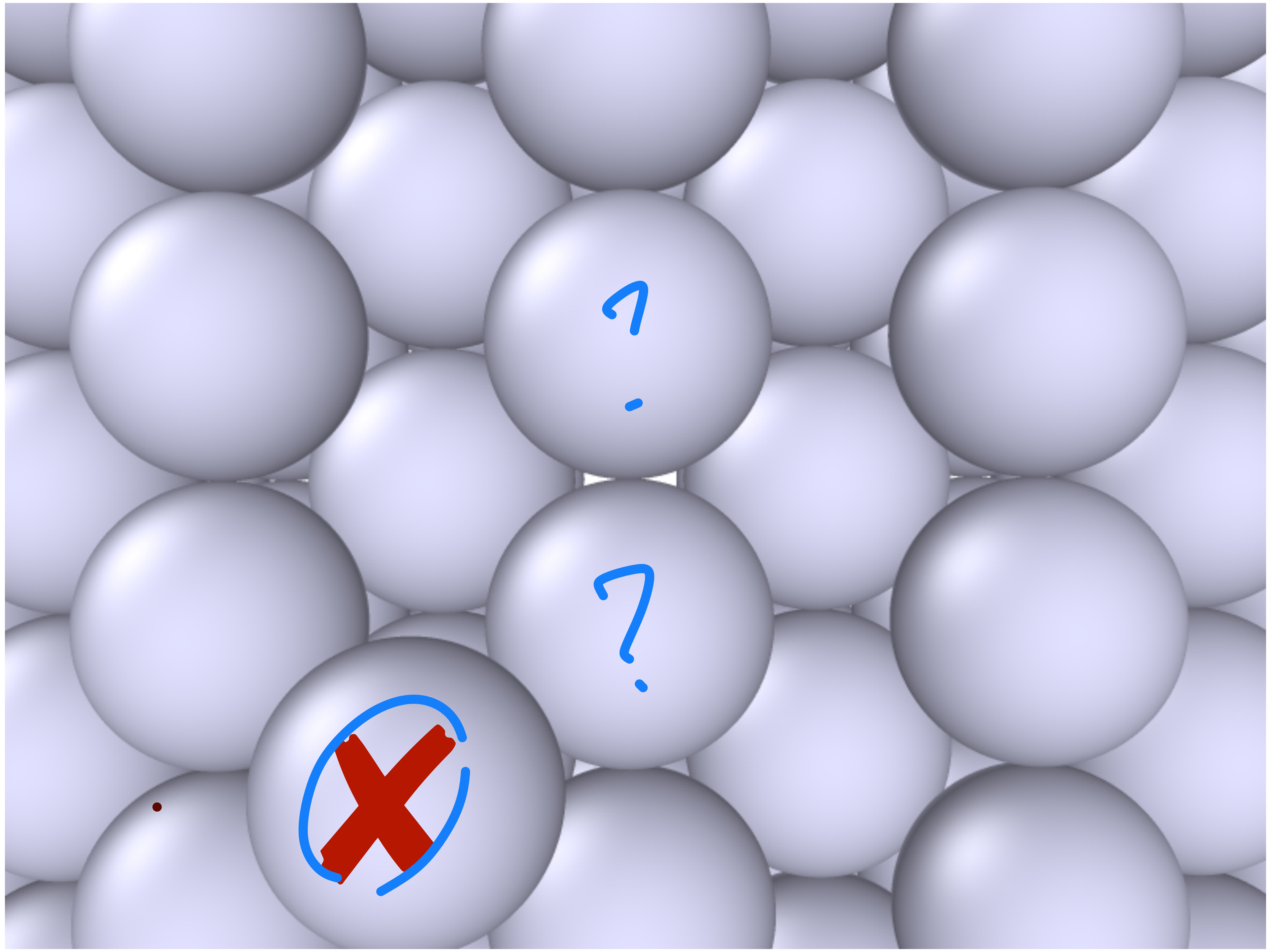


to

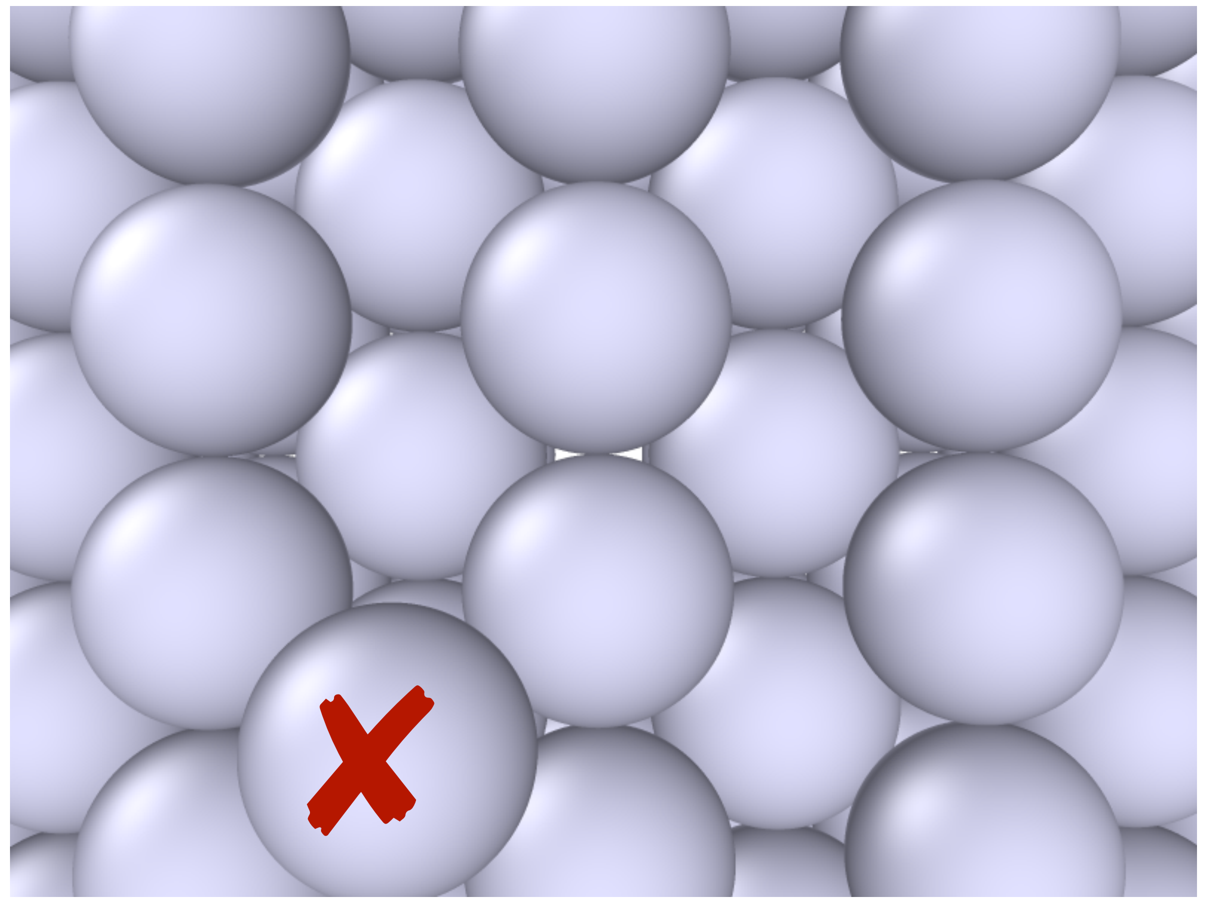


Extra Ag atom movement on Ag surface

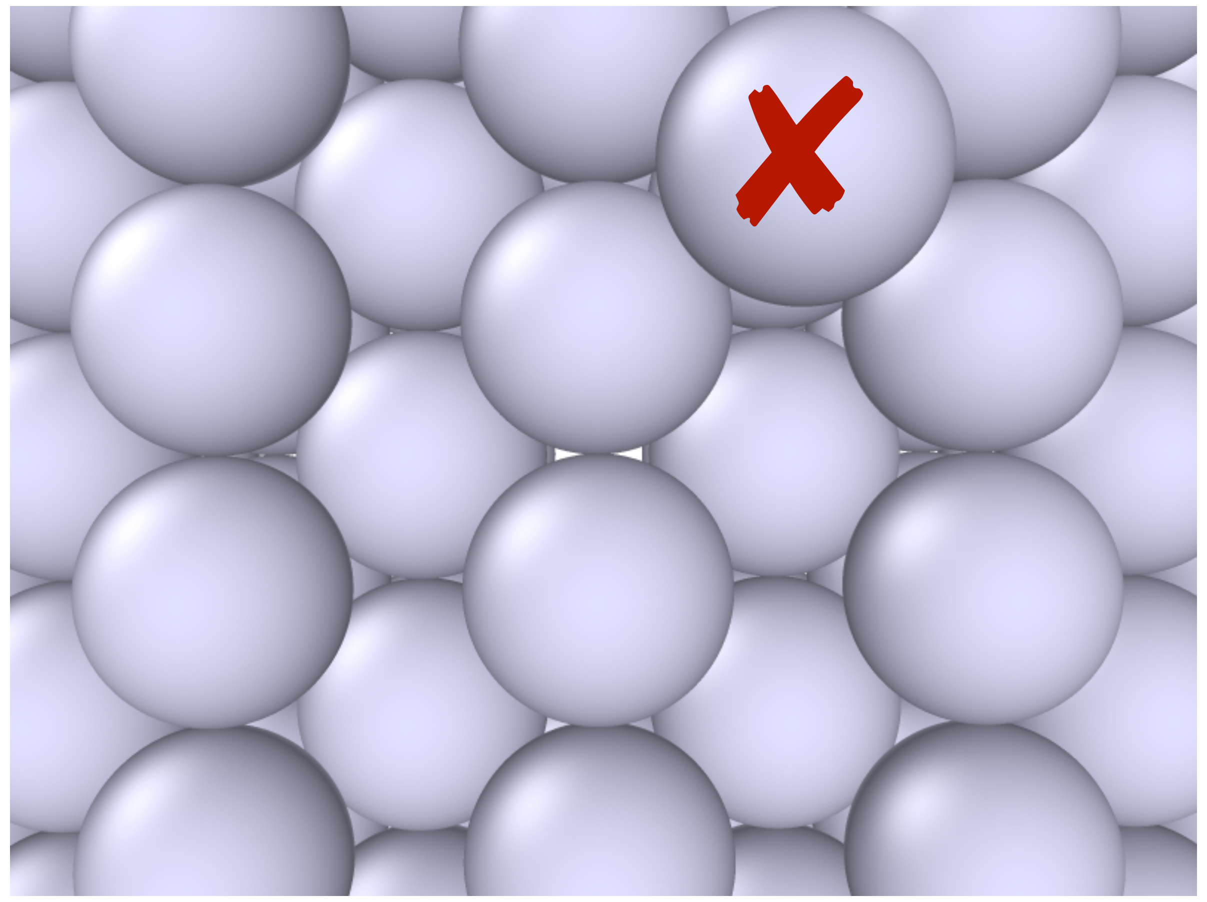
Which atoms can be involved?



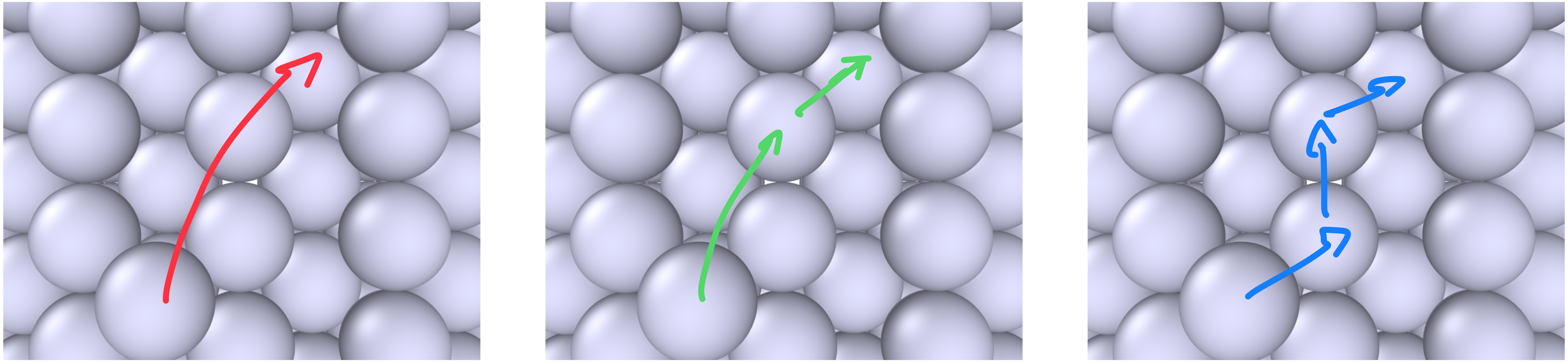
Extra Ag atom movement on Ag surface

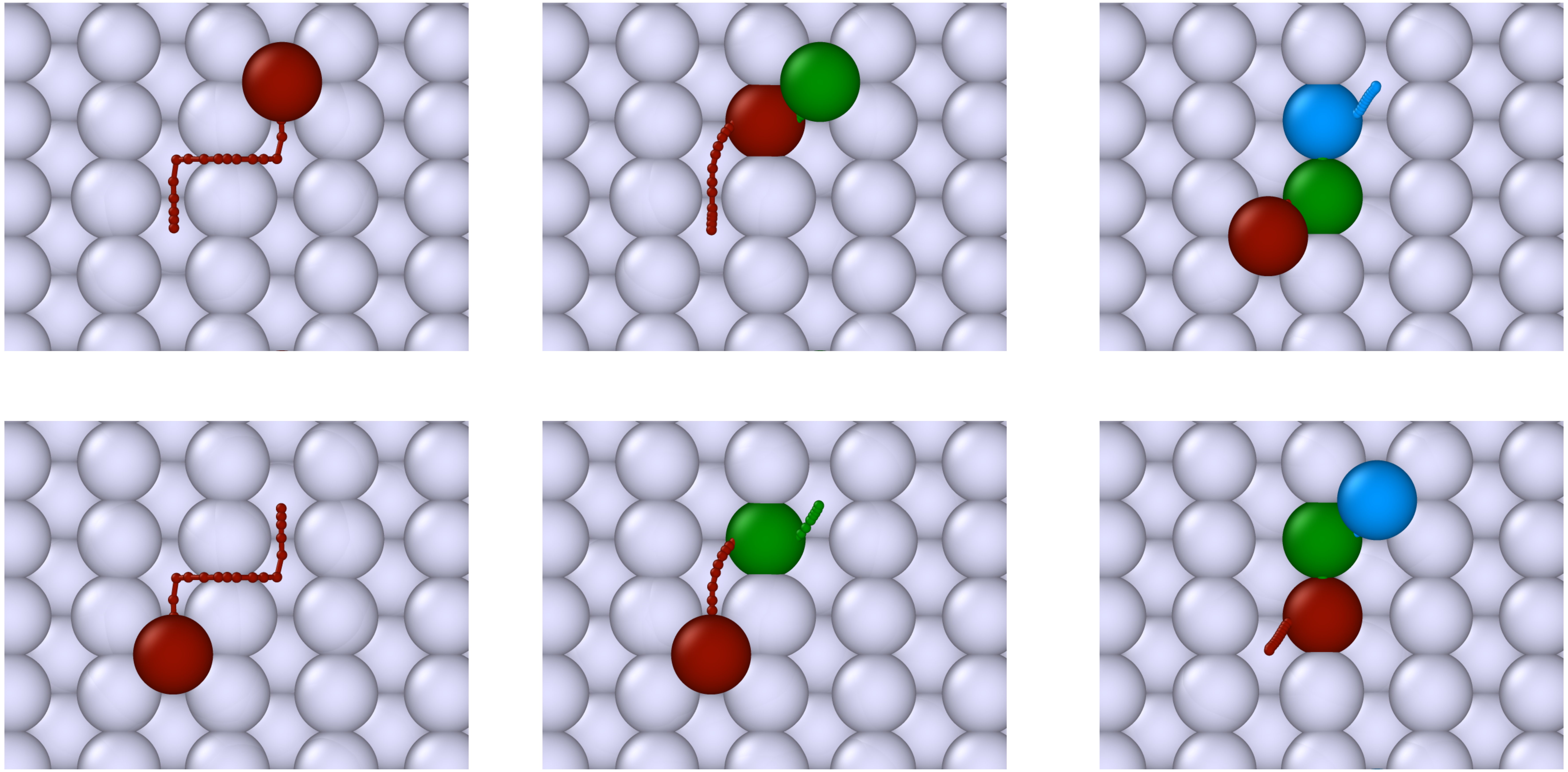


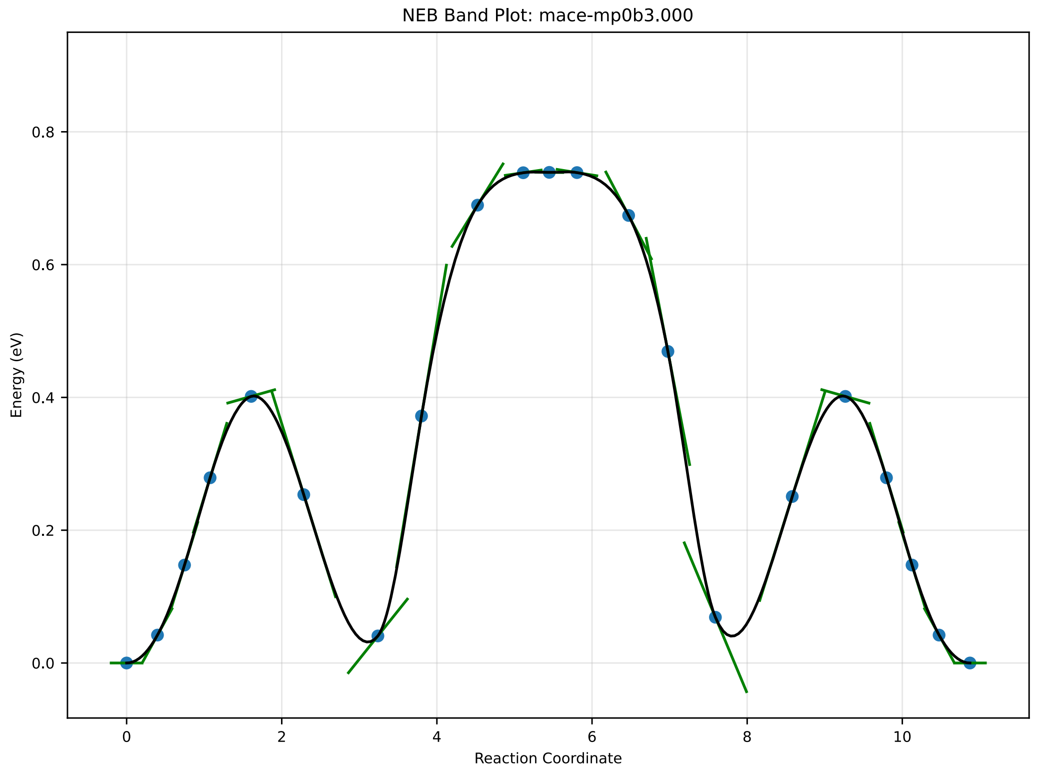
- Adatom moves or interaction with the surface?
- Heuristics applicable
- Easy/known mechanisms



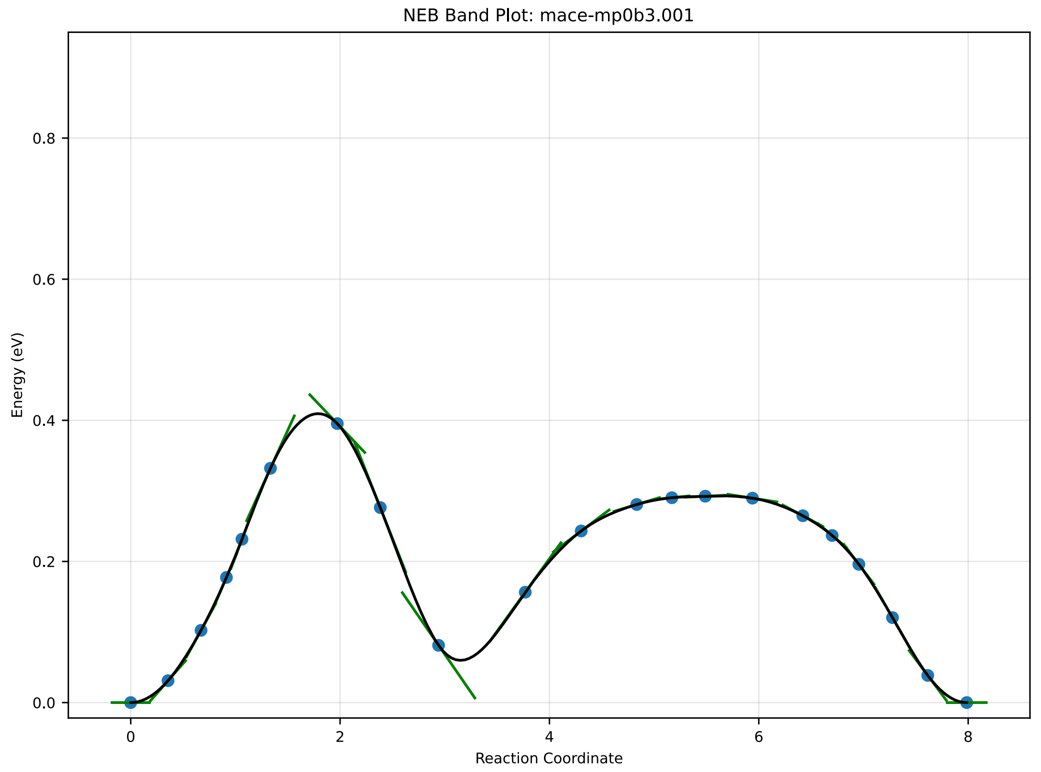
Extra Ag atom movement on Ag surface



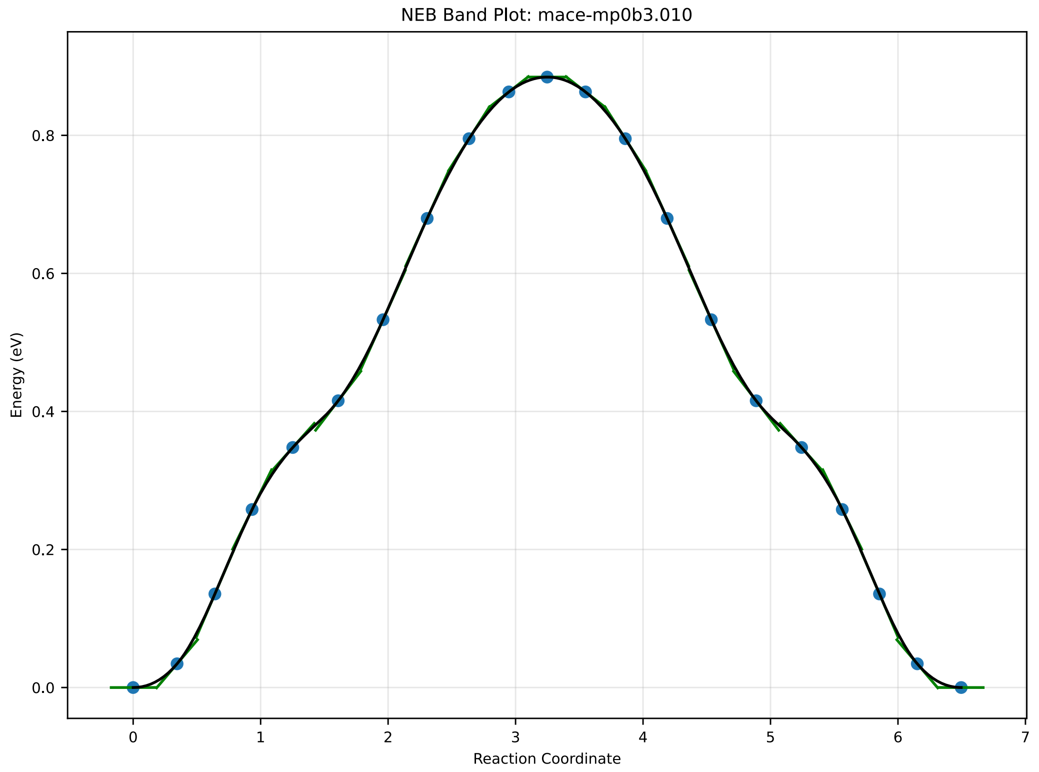




NEB Band Plot: mace-mp0b3.000
Energy (eV)
Reaction Coordinate
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10



NEB Band Plot: mace-mp0b3.001
Energy (eV)
Reaction Coordinate
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8



NEB Band Plot: mace-mp0b3.010
Energy (eV)
Reaction Coordinate
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
7

Optimal Transport

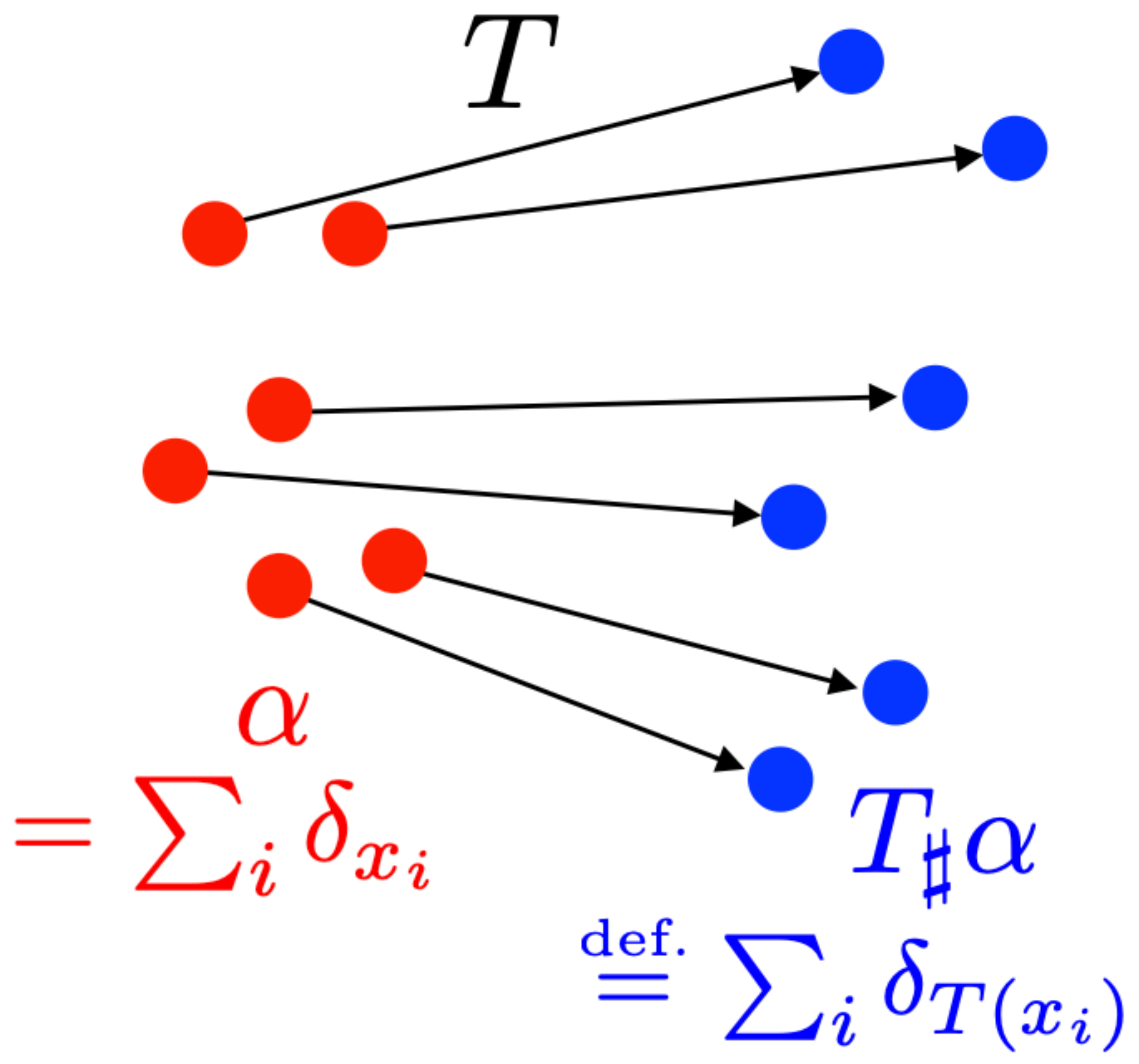
Optimal Transport

Monge - ca. 1784

- Move mass between measures at minimal cost
- Given cost matrix C(i,j)

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}.$$

- Combinatorial complexity
- Transport plan: no splitting of mass



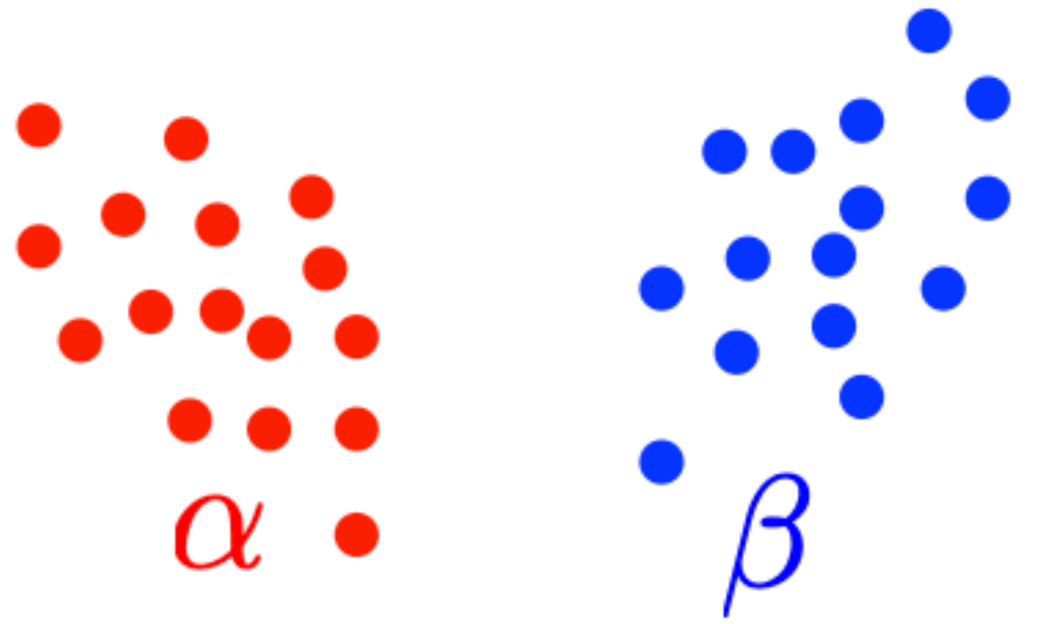
Graphics: https://optimaltransport.github.io/ & https://arxiv.org/abs/1803.00567

Optimal Transport

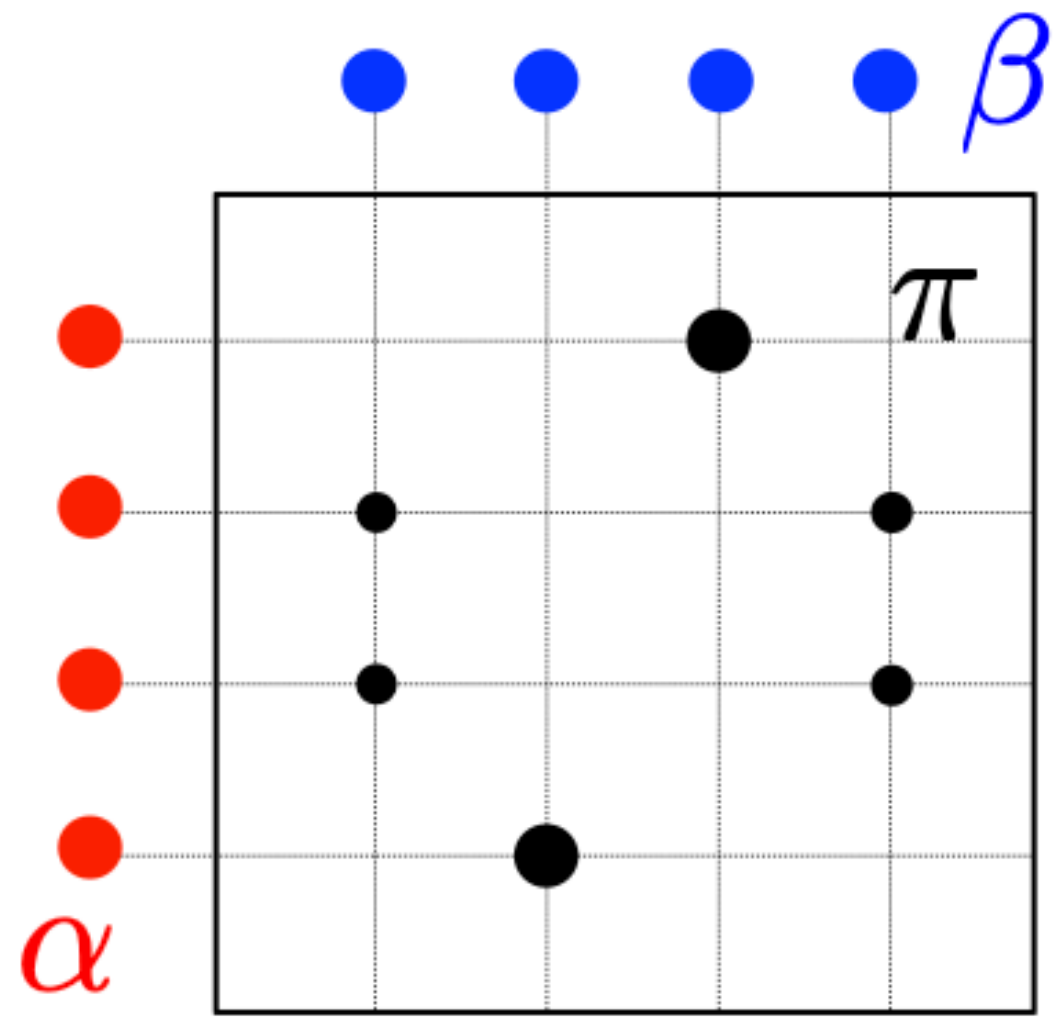
Kantorovich - 1942

- Allow splitting of mass: coupling P(i,j)

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\, \mathbf{P}\rangle \stackrel{\text{def.}}{=} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}.$$



- Linear Program: O(N^3) complexity



Graphics: https://optimaltransport.github.io/ & https://arxiv.org/abs/1803.00567

Optimal Transport

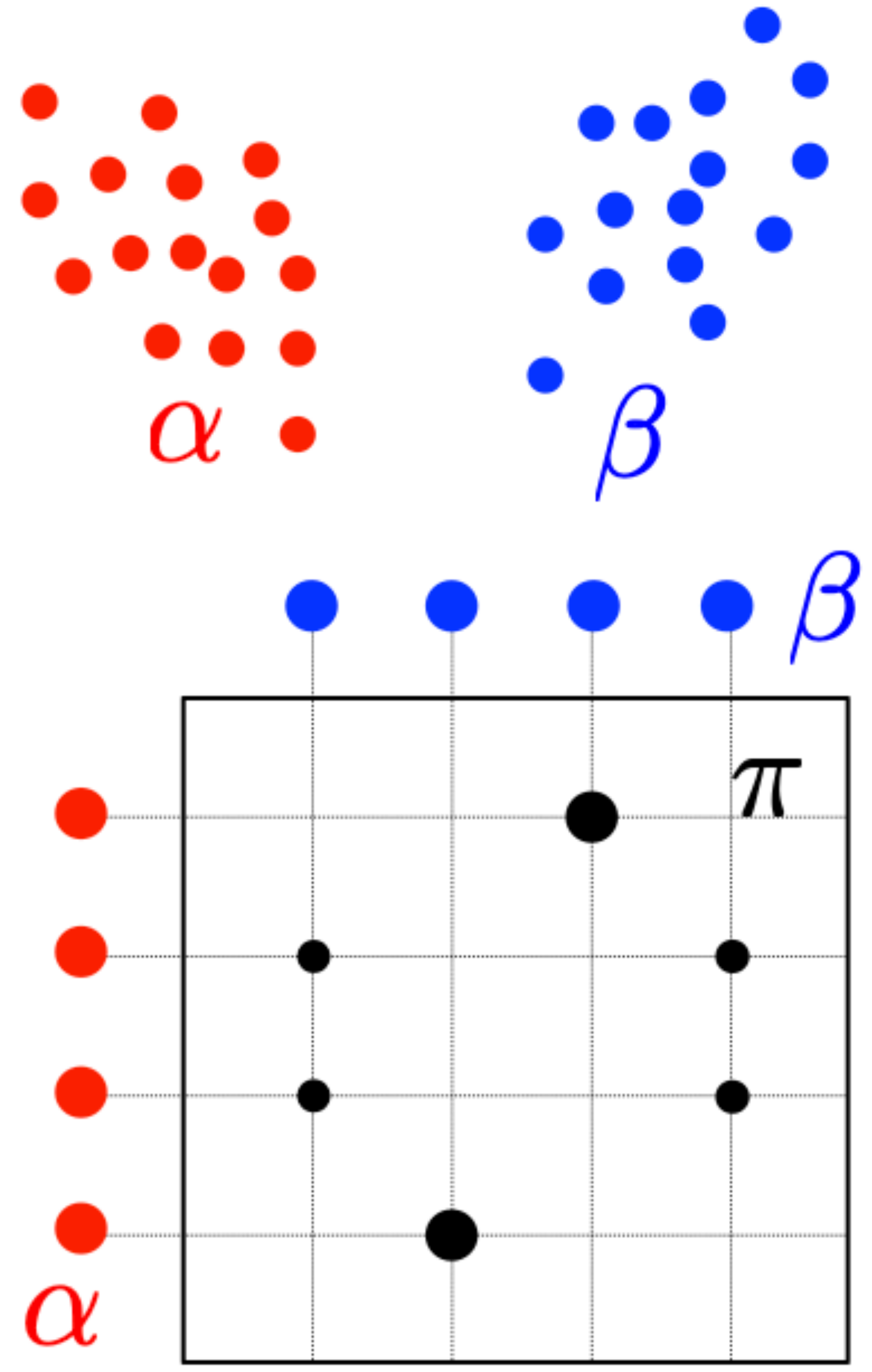
Entropic regularisation

- Add a regularisation term to the problem:

$$\mathbf{H}(\mathbf{P}) \overset{\text{def.}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1),$$

$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}).$$

- Iteratively solvable: Sinkhorn

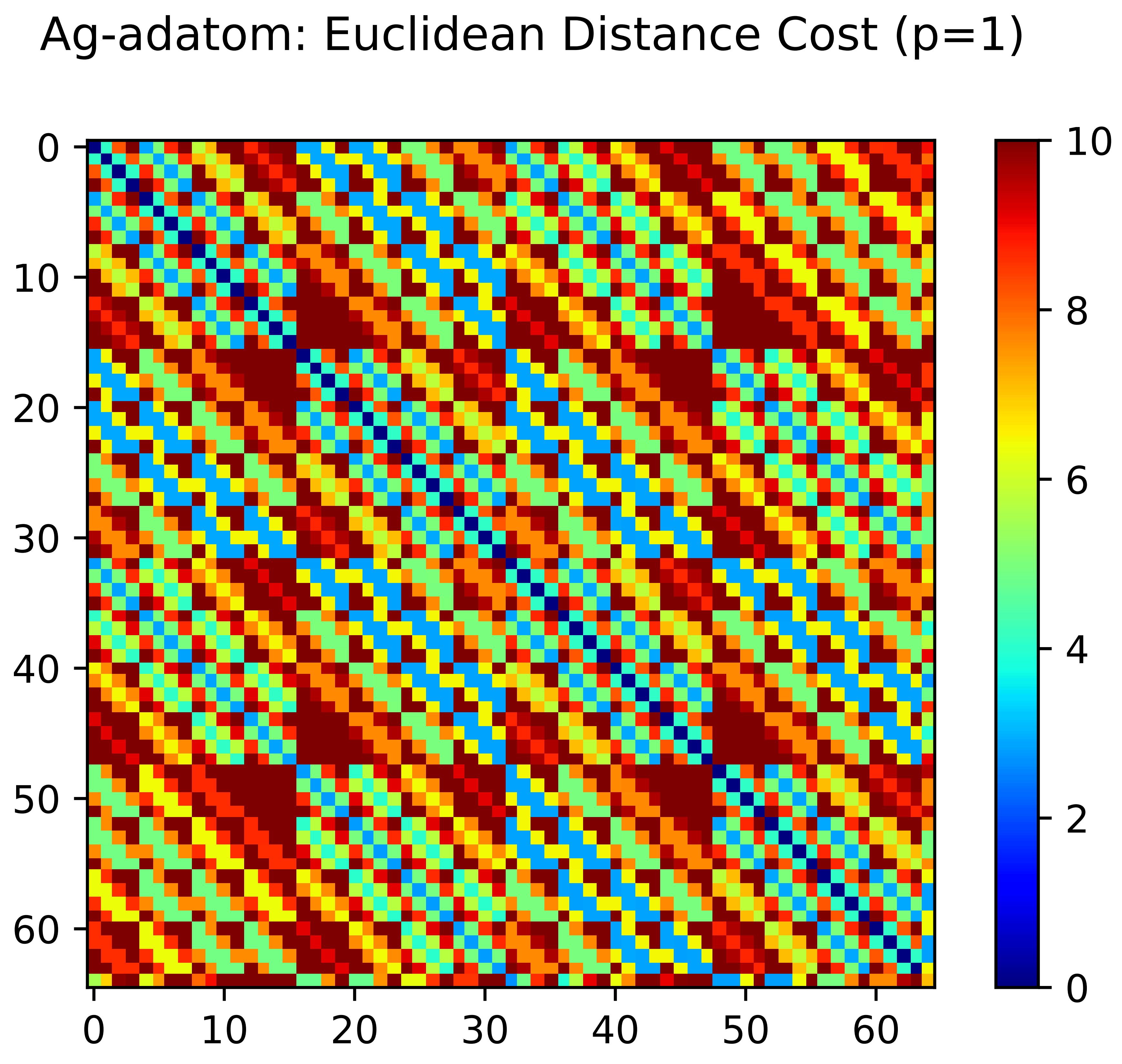


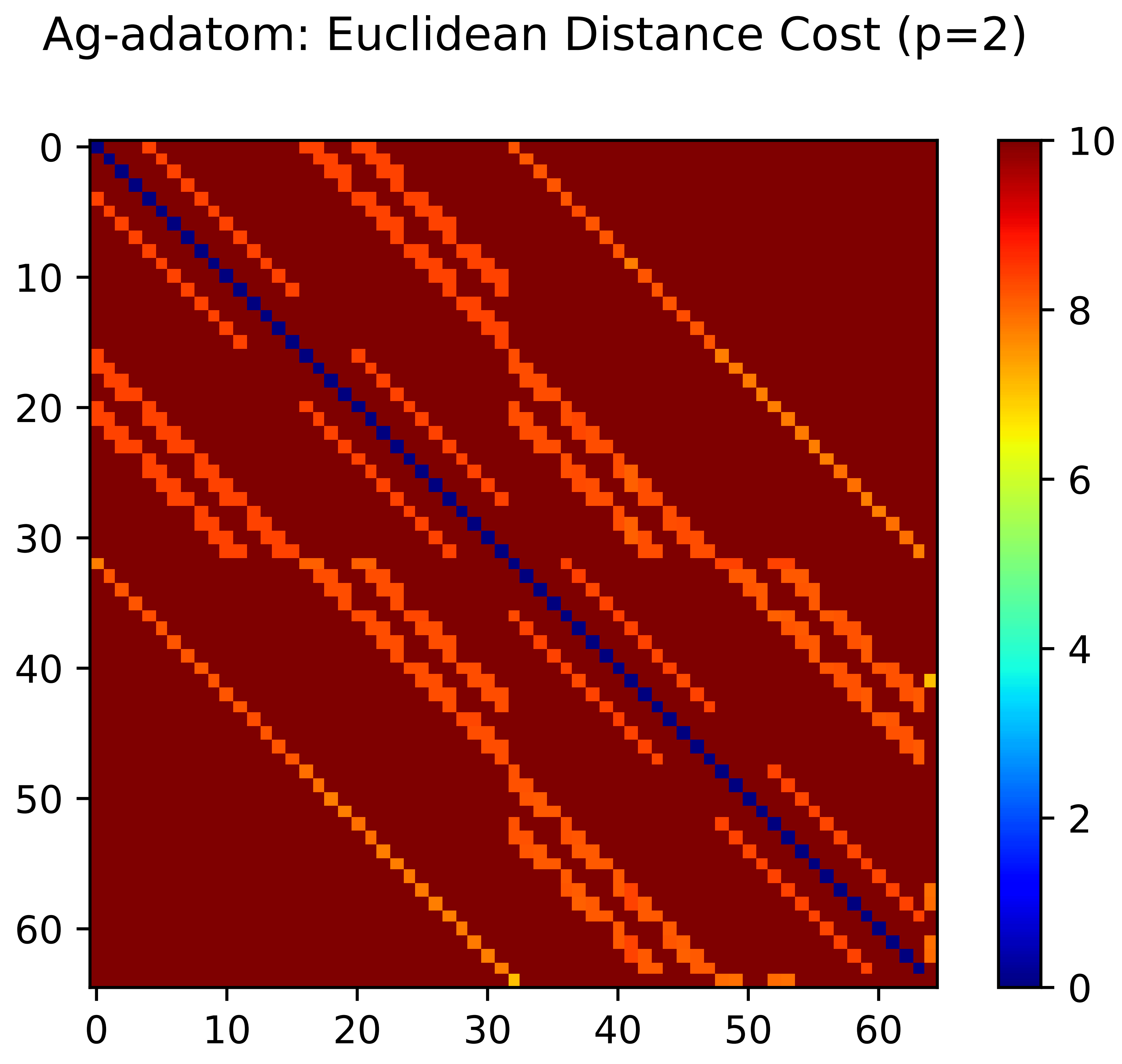
Graphics: https://optimaltransport.github.io/ & https://arxiv.org/abs/1803.00567

OT for atom assignment

Ag example

- Cost fn: Eucledian distance to power **p**

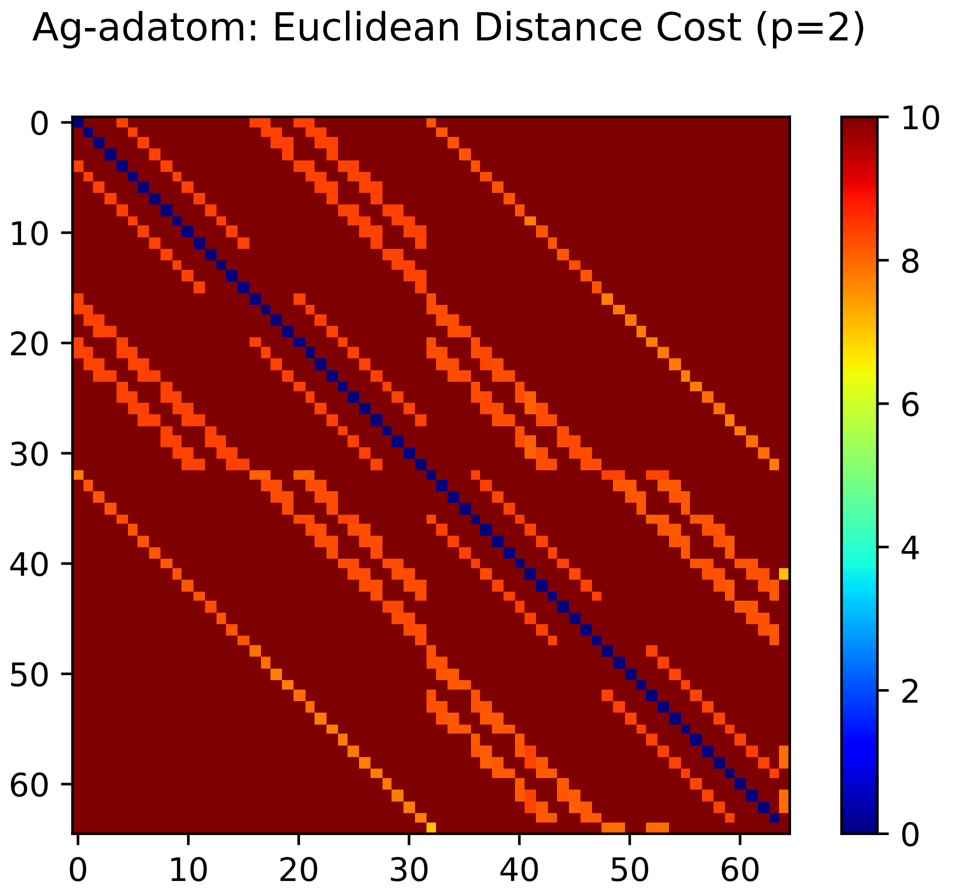


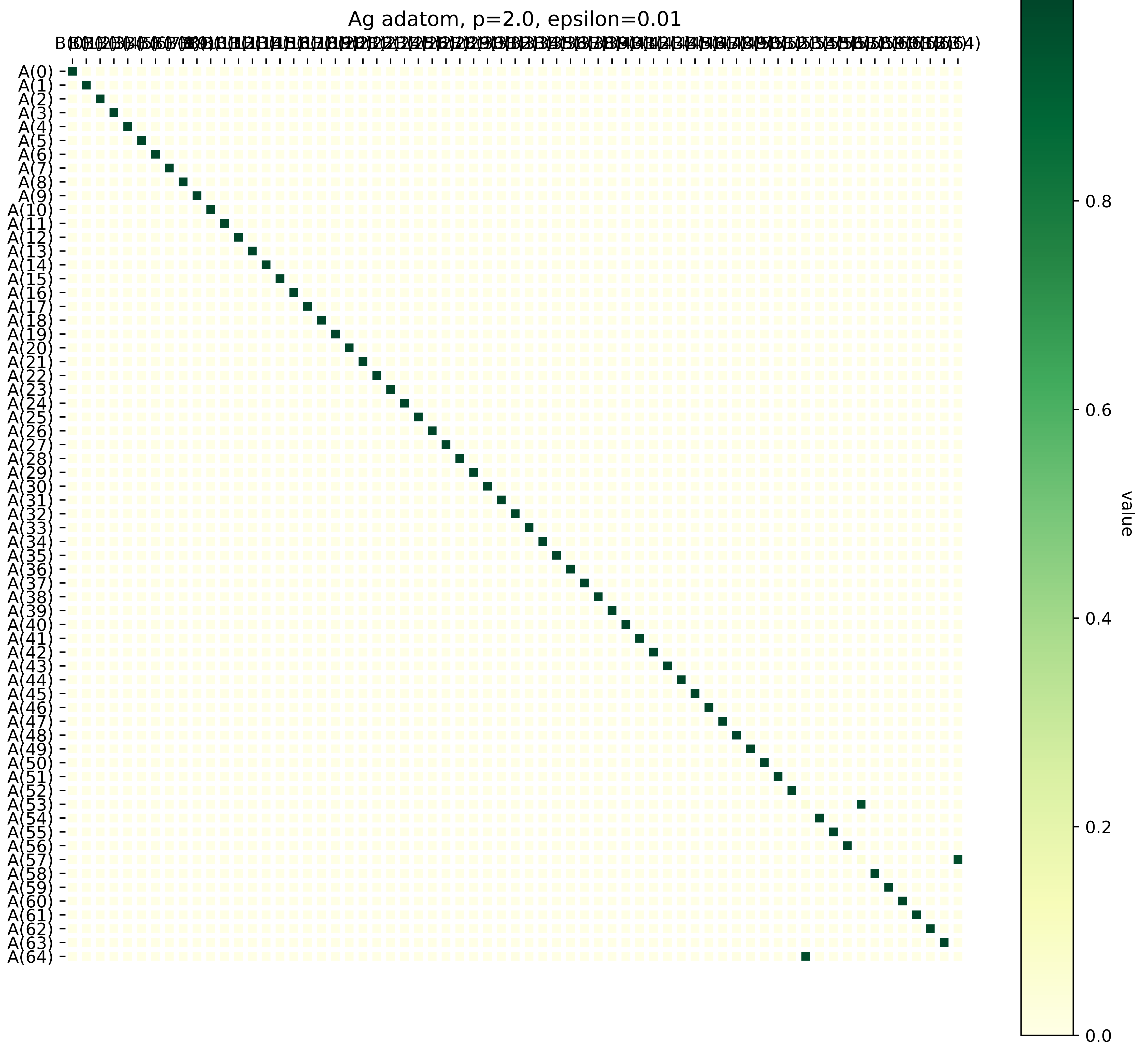


OT for atom assignment

Ag example

- Cost fn: Eucledian distance to power **p**
- Solve entropy regularised problem

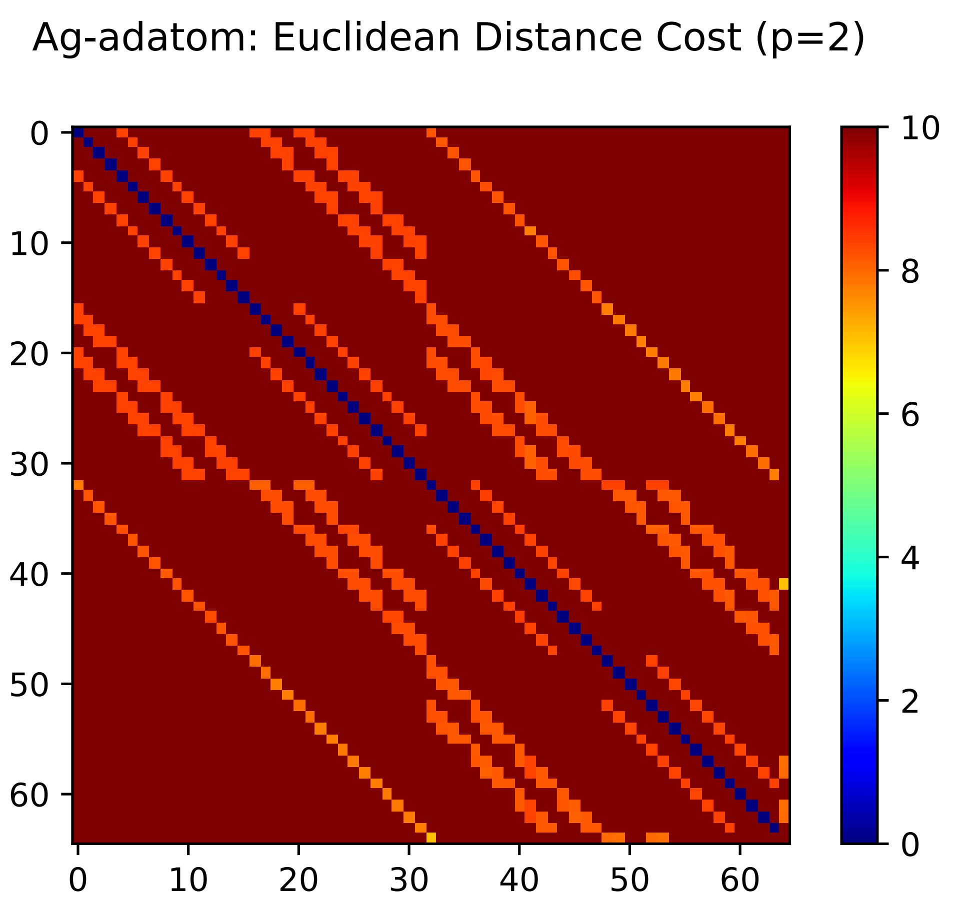


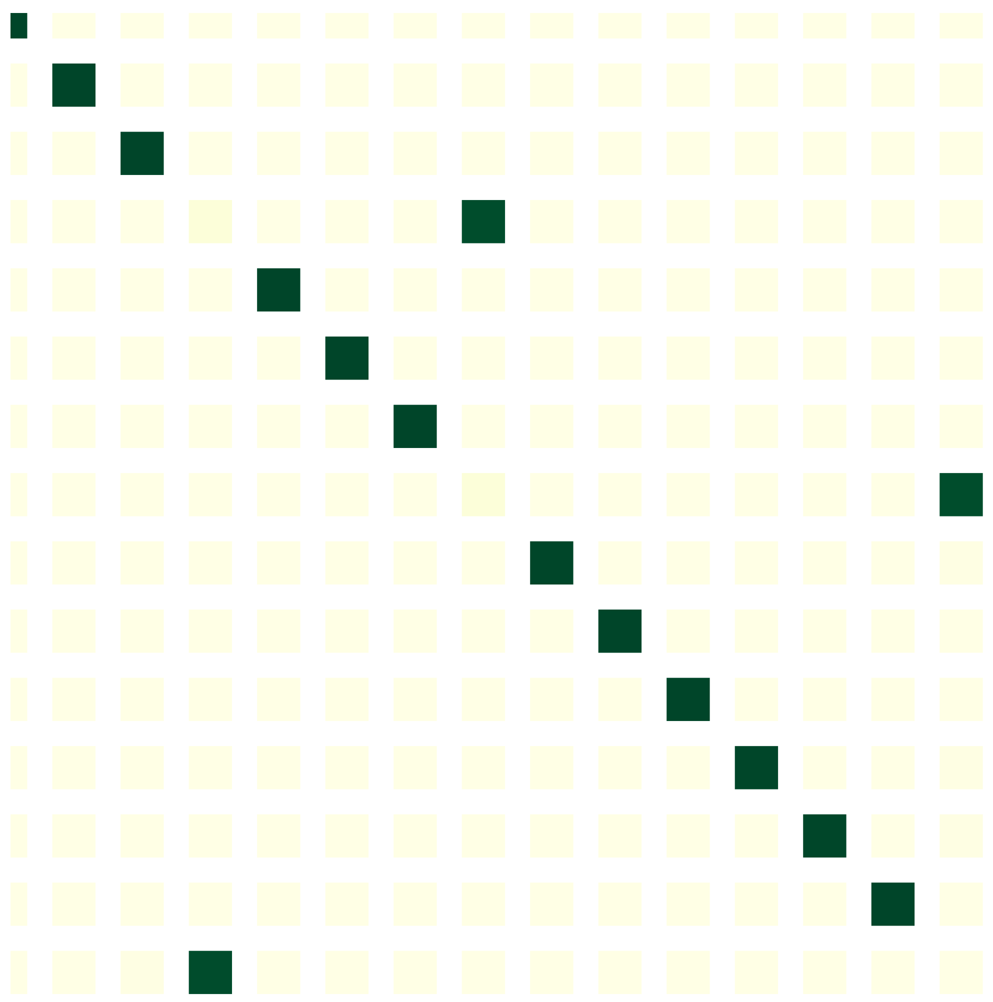


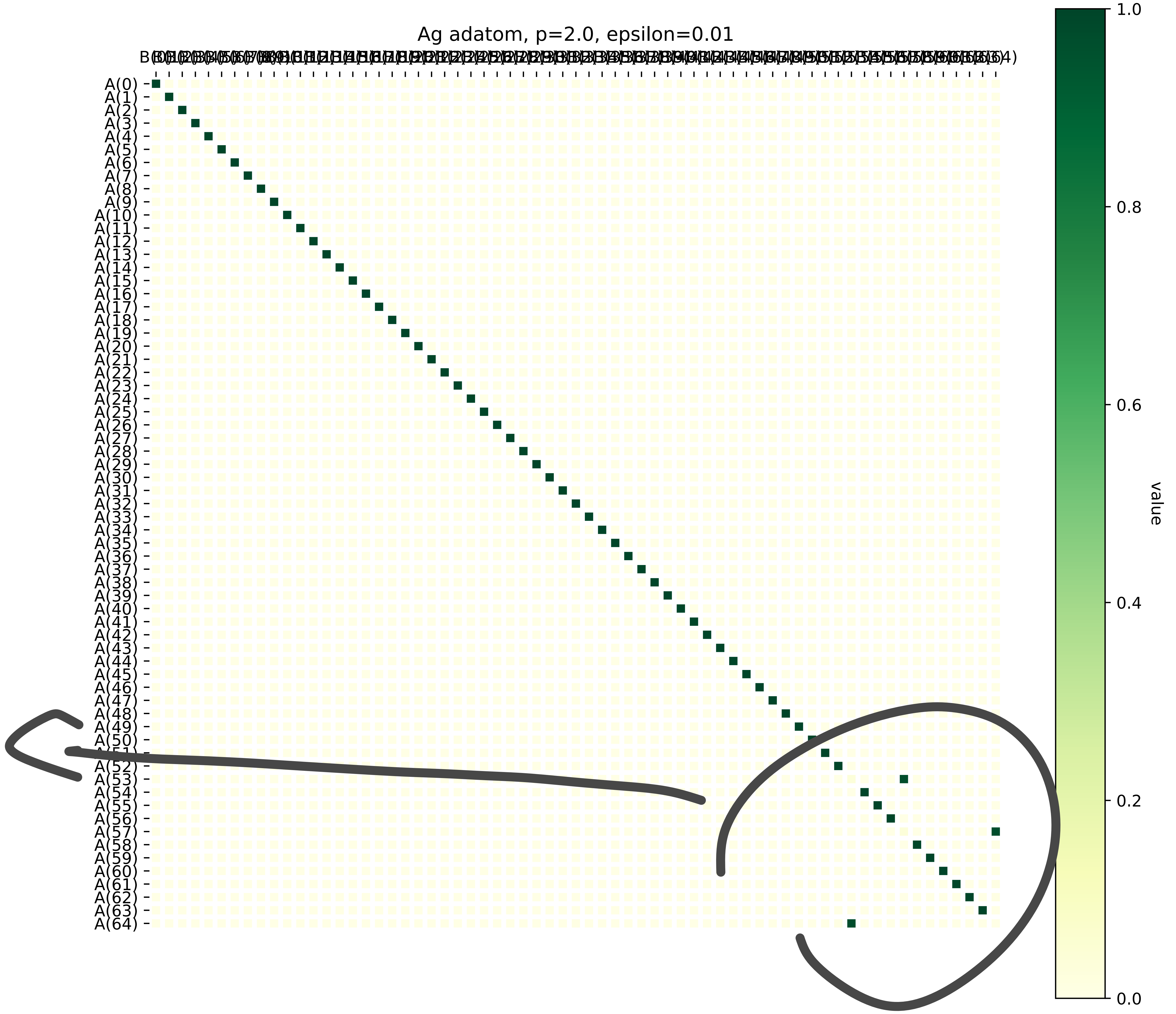
OT for atom assignment

Ag example

- Cost fn: Eucledian distance to power **p**
- Solve entropy regularised problem







OT for atom assignment
Ag example

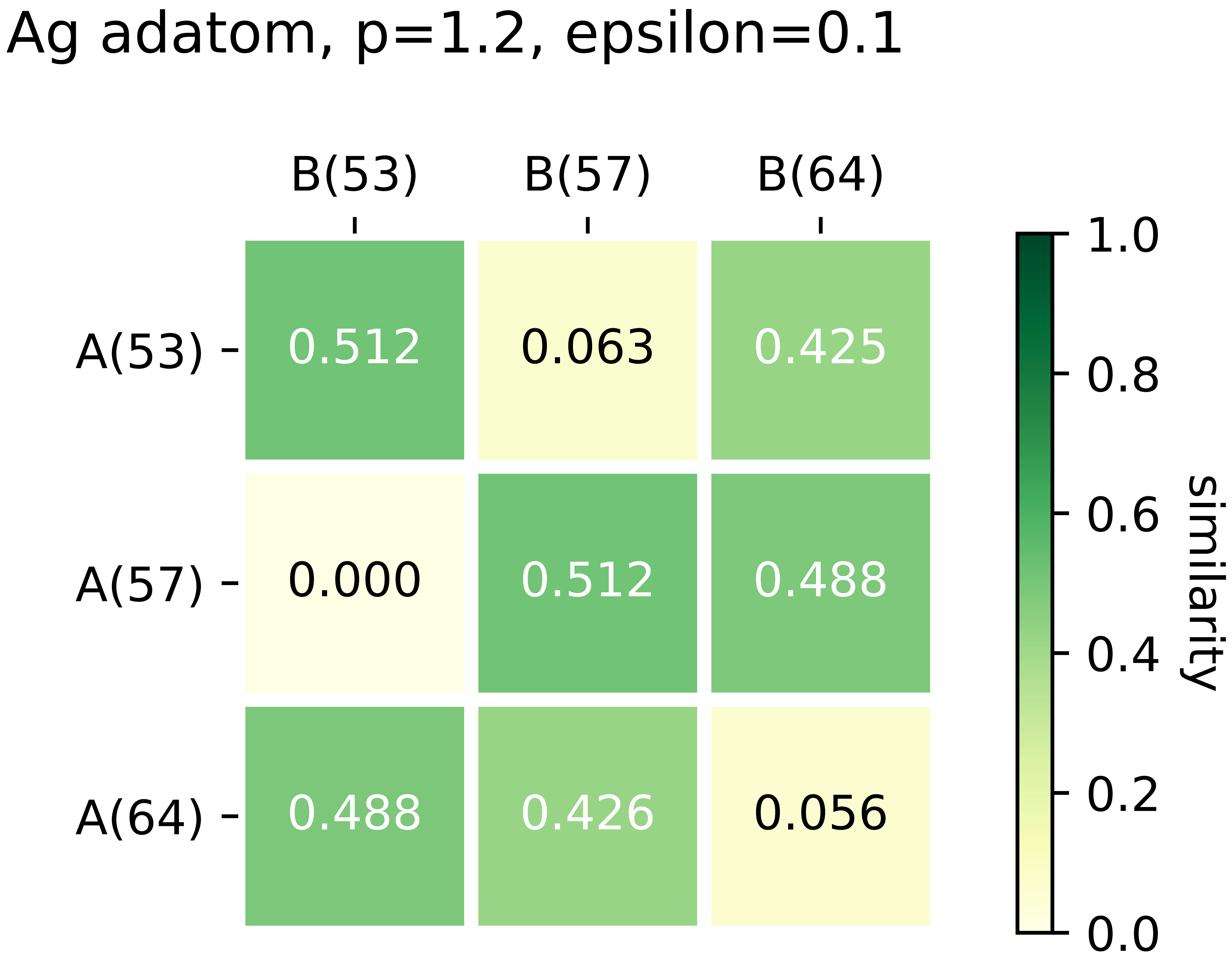
- Cost fn: Eucledian distance to power **p**
- Solve entropy regularised problem
- Vary **p** & **ε** - explore

OT for atom assignment

Ag example

- Cost fn: Eucledian distance to power **p**
- Solve entropy regularised problem
- Vary **p** & **ε** - explore
- Choose top permutation matrices:

$$\max_{\sigma \in Perm(n)} \sum_{i,j} \sigma_{i,j} P_{i,j}$$



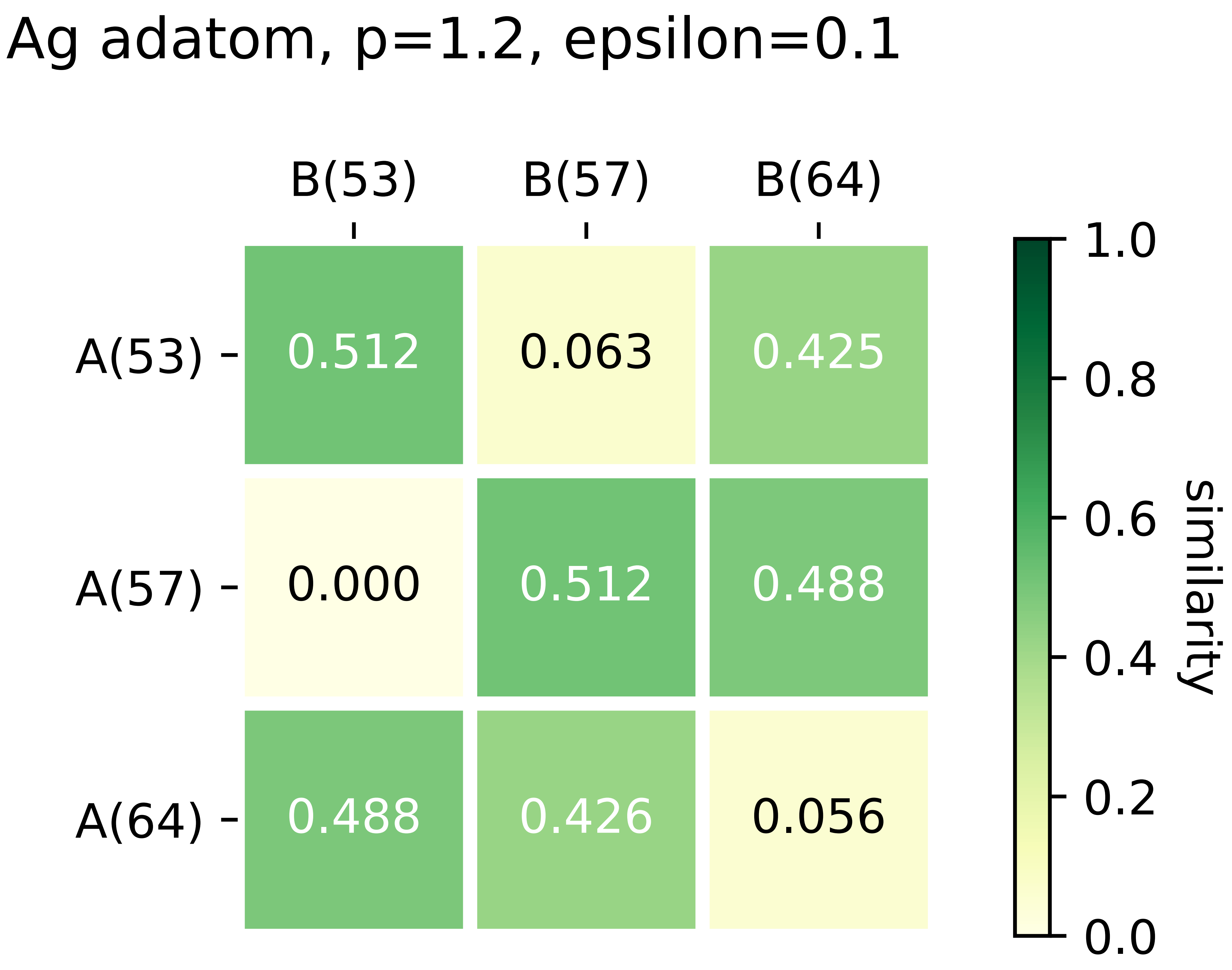
OT for atom assignment

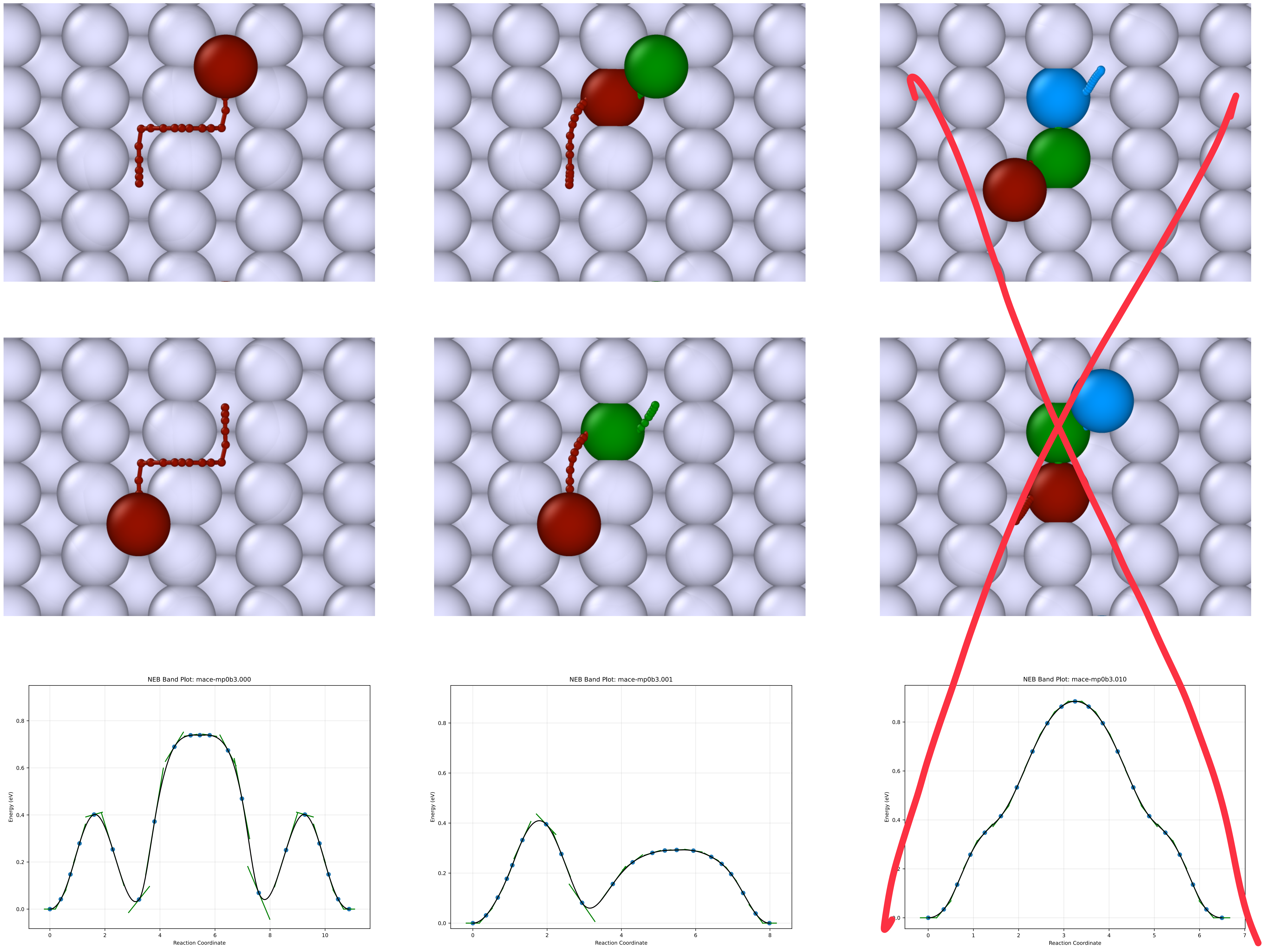
Ag example

- Cost fn: Eucledian distance to power **p**
- Solve entropy regularised problem
- Vary **p** & **ε** - explore
- Choose top permutation matrices:

$$\max_{\sigma \in Perm(n)} \sum_{i,j} \sigma_{i,j} P_{i,j}$$

- Find transition paths using NEB!



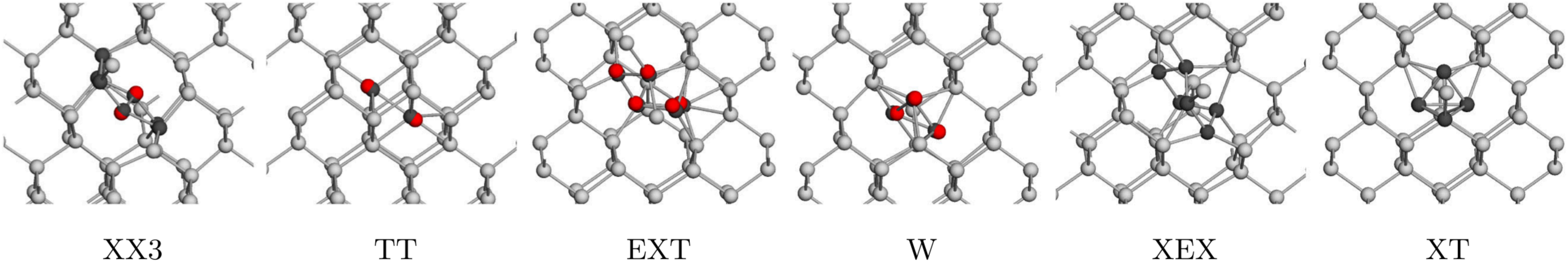

NEB Band Plot: mace-mp0b3.000
Energy (eV)
Reaction Coordinate
NEB Band Plot: mace-mp0b3.001
Energy (eV)
Reaction Coordinate
NEB Band Plot: mace-mp0b3.010
Energy (eV)
Reaction Coordinate

Si double-interstitial

Si double-interstitial

2 extra atoms in a Si cell

- 6 known structures
- Diffusion path known & verified for the ground state
- Transitions not studied between ground & excited states yet (AFAIK)



https://doi.org/10.1103/PhysRevX.8.041048

Si double-interstitial: diffusion

Known diffusion mechanism

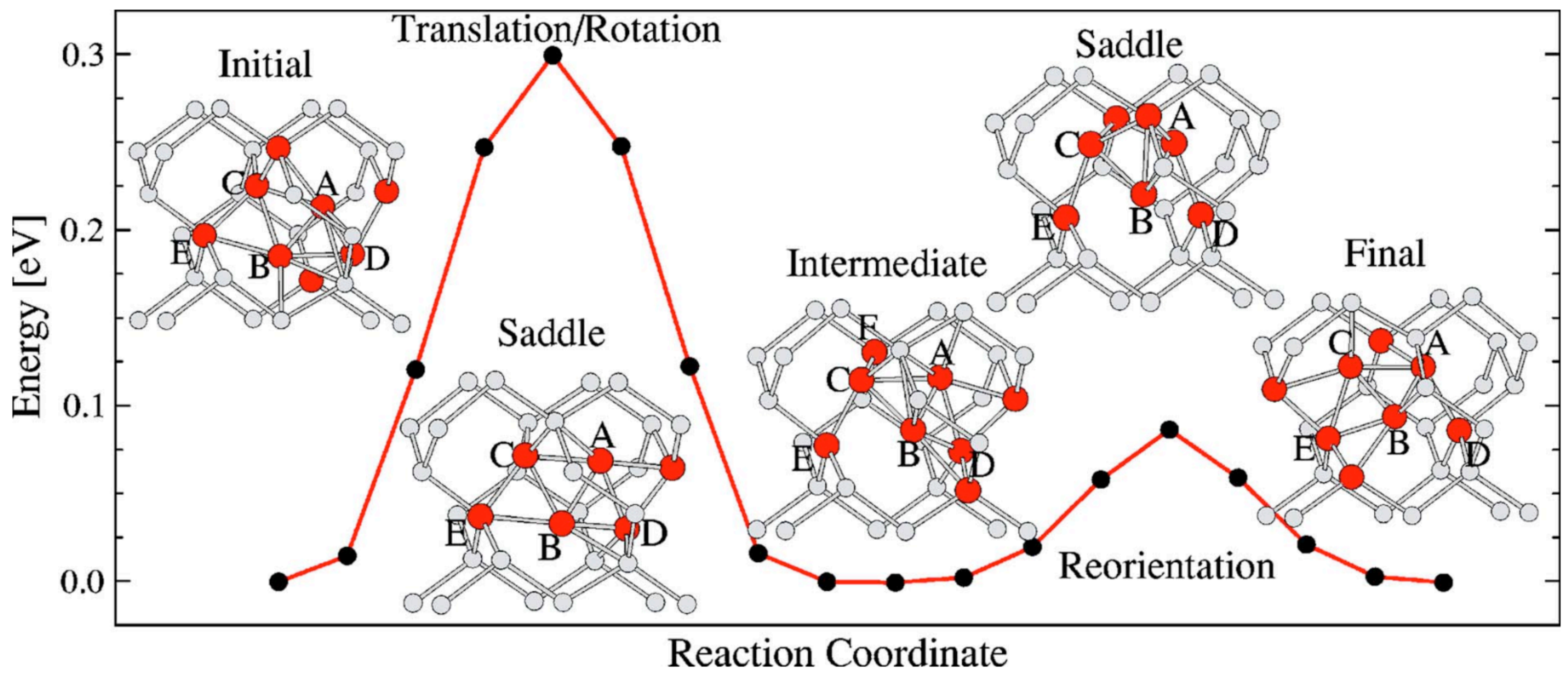
- Known since 2006
- Hand-crafted, from intuition and from other structures: e.g. tri-interstitial
- MACE-MP reproduces these results

PHYSICAL REVIEW B **73**, 245203 (2006)

Diffusion mechanisms for silicon di-interstitials

Yaojun A. Du,* Richard G. Hennig, and John W. Wilkins
Department of Physics, Ohio State University, Columbus, Ohio 43210, USA

(Received 6 October 2005; revised manuscript received 4 May 2006; published 12 June 2006)

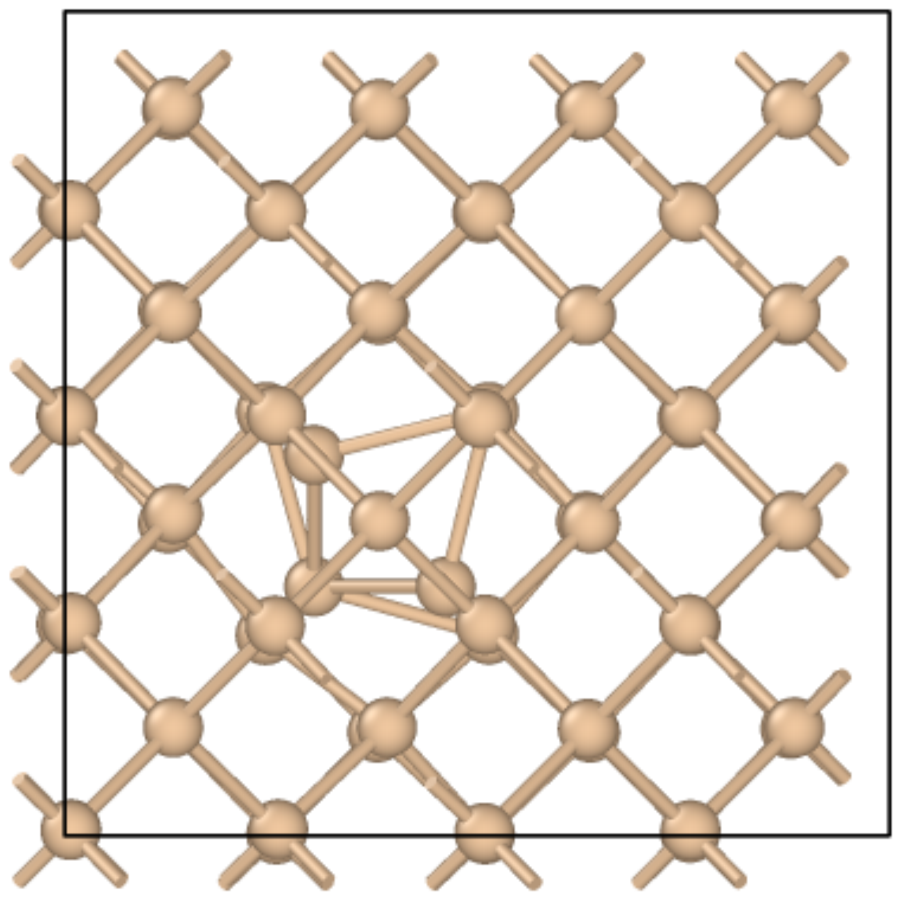


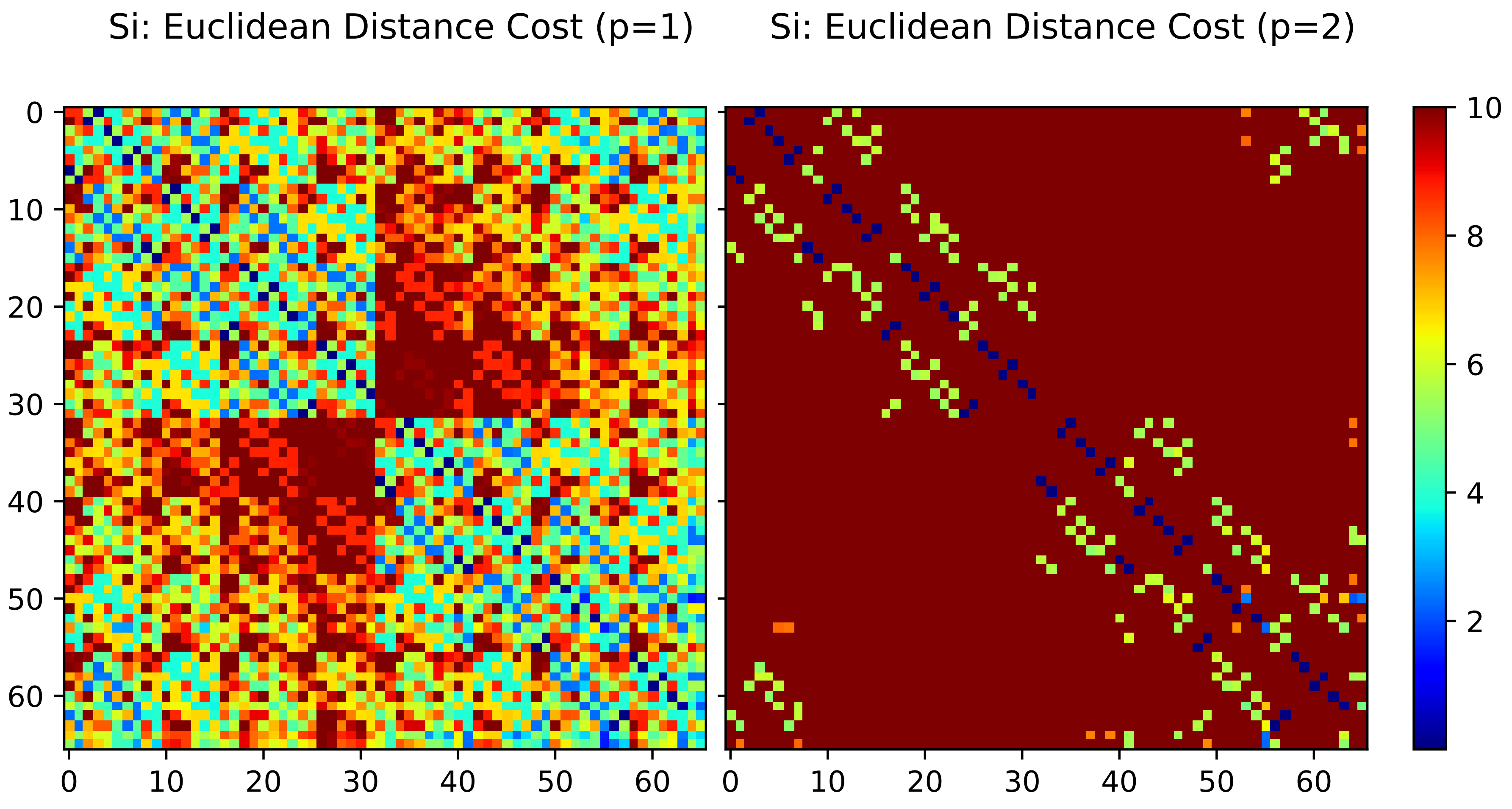
https://doi.org/10.1103/PhysRevB.73.245203

Si double-interstitial: diffusion

Matching explored

- Follow previous OT protocol
- Only 7 high-ranking assignments to try





	B(48)	B(50)	B(53)	B(55)	B(56)	B(58)	B(61)	B(63)	B(64)	B(65)
A(48)	0.000	0.997	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000
A(50)	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.941	0.030
A(53)	0.000	0.000	0.367	0.499	0.003	0.000	0.000	0.126	0.000	0.000
A(55)	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
A(56)	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
A(58)	0.000	0.000	0.000	0.000	0.000	0.000	0.994	0.000	0.005	0.000
A(61)	0.000	0.000	0.124	0.000	0.000	0.000	0.000	0.752	0.000	0.124
A(63)	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
A(64)	0.000	0.000	0.480	0.005	0.000	0.000	0.000	0.000	0.035	0.480
A(65)	0.000	0.000	0.000	0.499	0.003	0.000	0.000	0.126	0.000	0.367

Ongoing work & Open Questions

Open Questions

- Periodicity:
    - Translational symmetry - entire cell moving as as action
    - Path not through the closest boundary
- Cost does not account for path: atoms clashing mid-way (dynamical formulation)
- Path initialisation: using the coupling & cost fn from OT (not needed yet)
- Exhaustive exploration of **p** & **ε** - understand structure, phase transitions?
-

Finding interesting mathematical objects with ML

Adam Zsolt Wagner

Google DeepMind

Hungarian Machine Learning Days 2025

Goal of the talk

Proofs are important in mathematics, but for many problems this is not the focus.

Finding 'good' constructions is often the crucial bit.

- Counterexamples to conjectures
- Knowing the optimal constructions for large parameters let us spot patterns
- Lead to exploration, new conjectures, new theorems

"The methods for coming up with useful examples in mathematics . . . are even less clear than the methods for proving mathematical statements."
— Gil Kalai.

I am interested in simple, useful ML tools that we can give mathematicians to use, to search for such objects.

Simplest possible approach: Reinforcement Learning

Conjecture

For any graph G, we have $\lambda_1(G) + \mu(G) \geq \sqrt{n-1} + 1$.

Refuted in 2010, but smallest counterexample found has 600 vertices.

Game: for each edge, decide whether to include it in the graph of not

Reward: $\lambda_1 + \mu$ (minimize).

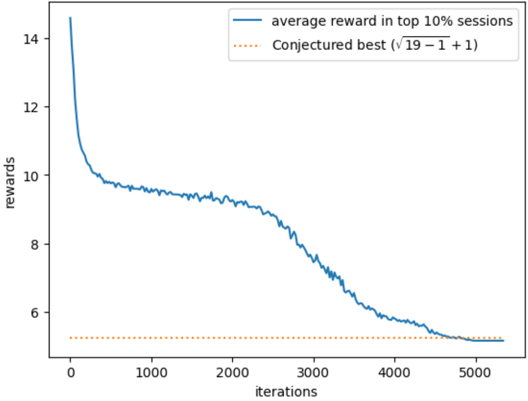
Run our simple algorithm for $n = 19$:

Constructions in combinatorics via neural networks, https://arxiv.org/abs/2104.14516

Example 1

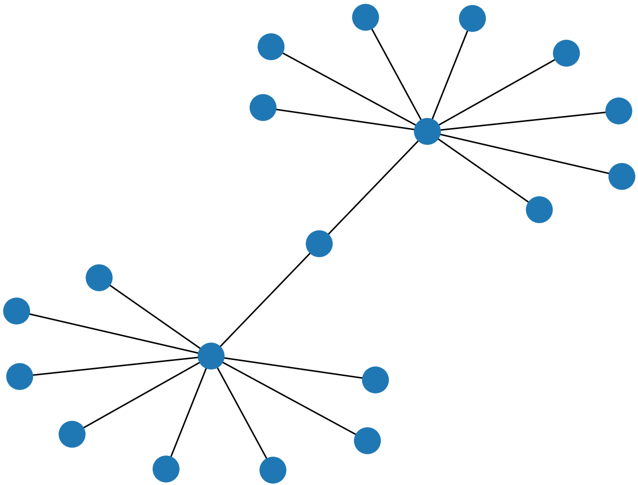
Conjecture

For any graph, $\lambda_1 + \mu \geq \sqrt{n-1} + 1$.



Constructions in combinatorics via neural networks, https://arxiv.org/abs/2104.14516

Example 1



Constructions in combinatorics via neural networks, https://arxiv.org/abs/2104.14516

More involved applications

Conjecture (Erdős, 1962)

The function

$$K_4(G) + K_4(\bar{G})$$

is asymptotically minimized by random graphs.

Thomason (1989): This is false!

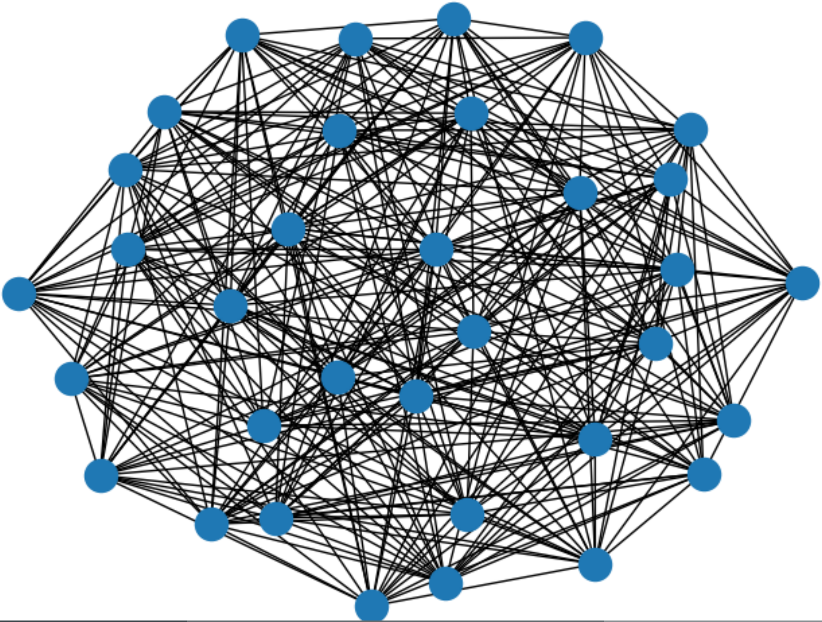
How can we refute such conjectures using RL?

Solution: "blowing up"! Construct a finite graph G, so that $G \times K_m$ is a counterexample as $m \to \infty$.

$\lim_{m\to\infty} \frac{K_4(G\times K_m)+K_4(\overline{G\times K_m})}{m^4}$ depends only on G, and there is an easy formula for it. This will be our reward function.

Run RL for $n = 34 \longrightarrow$ find a counterexample.

A counterexample – joint work with Gwen Joret



Pros and cons

How useful is this simple RL setup in pure maths research?

Pros:

- Simple and fun baseline method that can be thrown at a large class of problems
- Occasionally it works...

Cons:

- ...but most of the times it doesn't.
- Very slow, doesn't scale well
- Often doesn't perform better than simpler methods

It works well on a niche subfield of algebraic graph theory:
Ghebleh–Al-Yakoob–Kanso–Stevanović (2024) found counterexamples to 30 out of 68 conjectures they tried.

The isosceles triangle problem

With a small change, we can make this method work a lot better in practice.

Question (Erdős)

How many points can we choose in the $N \times N$ grid, without choosing three points that satisfy $d(a, b) = d(b, c)$, i.e. without creating any isosceles triangles?

Let this maximum be $f(N)$. Barely anything is known about $f(N)$, it would be helpful to know see the best constructions for $N = 64$, say.

Charton, Ellenberg, W., Williamson PatternBoost: Constructions in Mathematics with a Little Help from AI https://arxiv.org/abs/2411.00566

We create a large database of good 64×64 constructions, using standard local search methods.

We train a simple transformer model (Makemore) on these, and then generate more constructions like those in the dataset.

The model finds new good constructions much more frequently. We can feed these back into the local search method, and repeat.

Idea: alternating the local and global steps yields good results

Charton, Ellenberg, W., Williamson PatternBoost: Constructions in Mathematics with a Little Help from AI https://arxiv.org/abs/2411.00566

The isosceles triangle problem

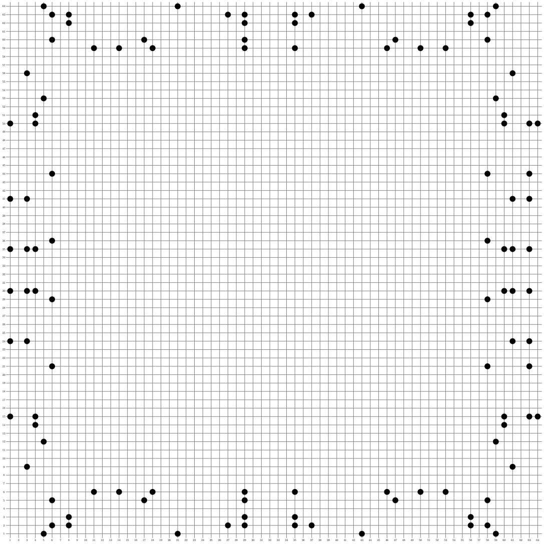


Figure: $f(64) \geq 110$

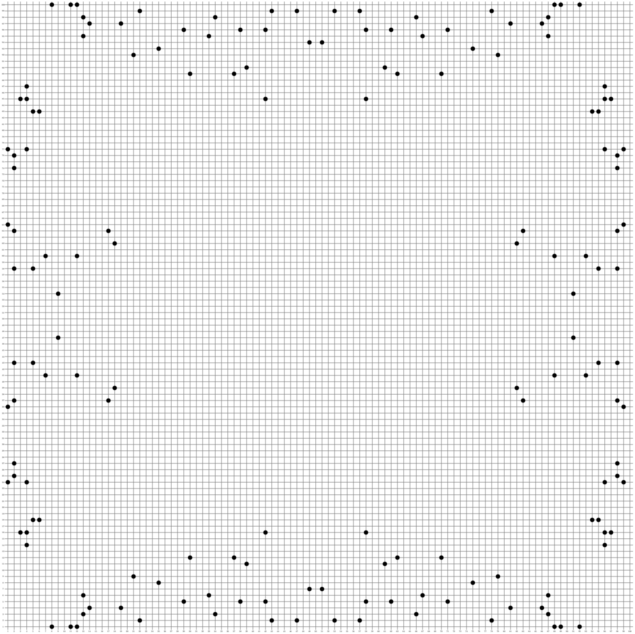


Figure: $f(100) \geq 160$

Local-Global

- Transformer: global picture
- Local search: local fixes
- In practice, we iterate this learn $\rightarrow$ generate $\rightarrow$ local search loop several times, always replacing the training set with the new best constructions. No need to retrain the model from scratch every time.

Charton, Ellenberg, W., Williamson PatternBoost: Constructions in Mathematics with a Little Help from AI https://arxiv.org/abs/2411.00566

Question (Erdős et al)

What is the smallest spanning subgraph of the n-cube Q_n that has diameter n?

There is a simple construction with only $2^n + \binom{n}{\lfloor n/2 \rfloor} - 2$ edges, this was conjectured 30 years ago to be optimal.

- Local search: given a graph, do trivial local fixes to make it better
  - If diameter $> n$ then add random edges to it
  - If diameter $= n$ then try to remove edges from it without increasing diameter
- Global step: train a model on the best constructions, generate more constructions like them
- Plug the new constructions back into the local search. Repeat.

Charton, Ellenberg, W., Williamson PatternBoost: Constructions in Mathematics with a Little Help from AI https://arxiv.org/abs/2411.00566

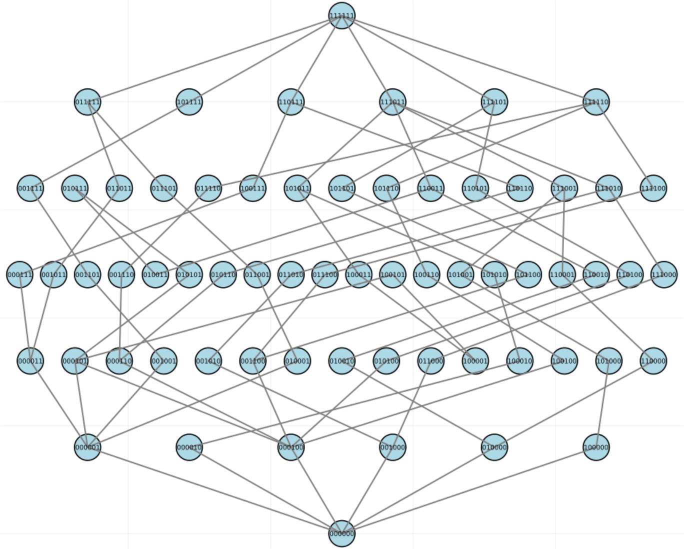


Figure: A diameter 6 subgraph of Q_6 with 81 edges

Searching in language-space

Local search $=$ if a construction is good, try other constructions that are **close** to the original one. But what does “close” mean?

- Standard local search: add/delete a few edges, change a handful of numbers slightly, etc
- Local search in language space: “close” means whatever the LLM thinks “close” means.

When is this useful?

1. In the mathematics we humans care about, constructions often have a short description
2. We can use this setup to find efficient search functions that find good constructions

---

FunSearch. Mathematical discoveries from program search with large language models by Romera-Paredes et al., Nature, 2023

AlphaEvolve: A coding agent for scientific and algorithmic discovery by Novikov et al., 2025, in collaboration with Terence Tao and Javier Gomez-Serrano

FunSearch example

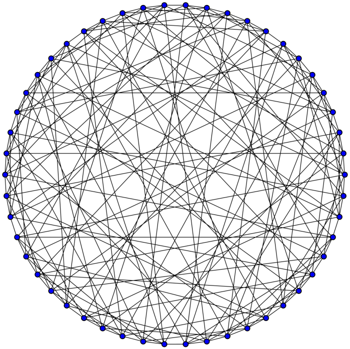
Problem

At most how many edges can a graph have, that has 50 vertices, and no triangles or cycles of length 4?

```python
import networkx as nx

def create_graph():
  g = nx.hoffman_singleton_graph()
  return g
```

```python
def construct_graph():
  graph = set()
  graph.add((0, 2))
  for j in range(4, n):
    for i in range(j - 1, -1, -1):
      if count_triangles_and_squares(graph | {(i, j)}, n) == 0:
        graph.add((i, j))
  return graph
```

```python
def create_graph():
  g = nx.Graph()
  g.add_nodes_from(range(50))

  # Create 10 disjoint 5-cycles
  for i in range(0, 50, 5):
      for j in range(5):
          g.add_edge(i + j, i + (j + 1) % 5)



  for _ in range(50):

    all_nodes_full = True
    for node in g.nodes:
        if g.degree(node) < 7:
          all_nodes_full = False
          break
    if all_nodes_full:
      break

    nodes_with_low_degree = sorted(
        [node for node in g.nodes if g.degree(node) < 7], key=g.degree
    )

    for start_node in nodes_with_low_degree:
          for end_node in g.nodes:
            if start_node != end_node and g.degree(end_node) < 7:
                if not g.has_edge(start_node, end_node):

                    has_small_cycle = False
                    # Check for 4 cycle
                    for n1 in g.neighbors(start_node):
                        for n2 in g.neighbors(end_node):
                            if n1 != end_node and n2 != start_node and \
                              g.has_edge(n1, n2):
                                has_small_cycle = True
                                break
                        if has_small_cycle:
                            break
                    if has_small_cycle:
                         continue

                    # Check triangle
                    has_small_cycle = False
                    for neighbor in g.neighbors(start_node):
                        if g.has_edge(neighbor, end_node):
                            has_small_cycle = True
                            break
                    if has_small_cycle:
                        continue
                    g.add_edge(start_node, end_node)
                    break
  return g
```



The best of both worlds

- Previous example: find a python code that generates a good construction directly
- Examples earlier in the talk: what kind of codebase can find good constructions quickly?
- Let's combine both: we can try to find a search function that finds the best possible construction within a fixed time limit
- "Local search in the space of search functions"
- This will find the best heuristic search function for your problem

Evolving a chain of search functions

One more trick: when evaluating a search function, we can initialize the search at the best construction we have found so far.

The result is a chain of search heuristics, evolved automatically, that when applied one after another, yields a good construction.

This turns out to be a good black box optimizer for many math problems (but interpretability goes out the window)

AlphaEvolve examples

Let C_1 denote the largest constant for which one has

$$\max_{-1/2 \le t \le 1/2} \int_{\mathbb{R}} f(t-x) f(x)\, dx \ge C_1 \left(\int_{-1/4}^{1/4} f(x)\, dx \right)^2 \qquad (1.1)$$

for all non-negative $f : \mathbb{R} \to \mathbb{R}$. This problem arises in additive combinatorics, relating to the size of Sidon sets. It is currently known that

$$1.28 \le C_1 \le 1.50992$$

1.5099 → 1.5056

Let C_1' be the best constant for which one has

$$\|f * f\|_2^2 \le C_1' \|f * f\|_1 \|f * f\|_\infty$$

for non-negative $f : \mathbb{R} \to \mathbb{R}$. It is known that

$$0.88922 \le C_1' \le 1$$

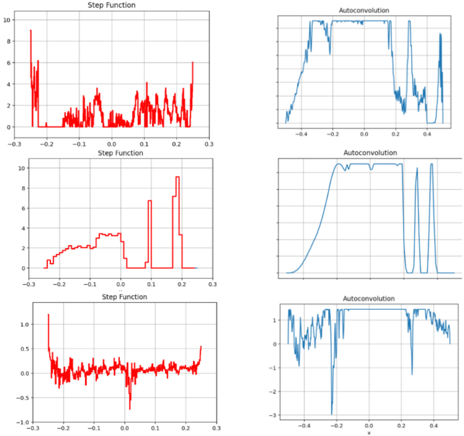
0.889 → 0.896

Now let C_1'' be the best constant for which one has

$$\max_{-1/2 \le t \le 1/2} \left| \int_{\mathbb{R}} f(t-x) f(x)\, dx \right| \ge C_1'' \left(\int_{-1/4}^{1/4} f(x)\, dx \right)^2$$

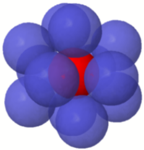
for all $f : [-1/4, 1/4] \to \mathbb{R}$, which we now allow to take both negative and positive values. Then $C_1'' \le C_1$. Here, there is a better example that gives a new upper bound on C_1'', namely $C_1'' \le 1.45810$.

1.458 → 1.455



Kissing numbers: how many disjoint unit spheres can touch a unit sphere simultaneously?

Dimension 11: 582 ⟶ 583



2. Uncertainty principles

Given a function $f \in L^1(\mathbb{R})$, define the Fourier transform $\hat{f}(\xi) := \int_{\mathbb{R}} f(x)e^{-2\pi i x\xi}\, dx$ and

$$A(f) := \inf\{r > 0 : f(x) \geq 0 \text{ for all } |x| \geq r\}.$$

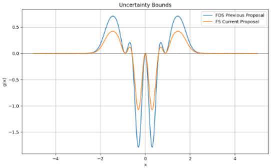
Let $C_{2.1}$ be the largest constant for which one has

$$A(f)A(\hat{f}) \geq C_{2.1}$$

for all even f with $f(0), \hat{f}(0) < 0$. It is known [53] that

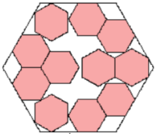
$$0.2025 \leq C_{2.1} \leq 0.353.$$

0.3522 ⟶ 0.3520

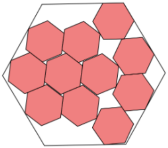


What is the smallest hexagon one can fit n unit hexagons into?

11.

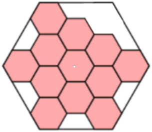


s = 5(3 + √3)/6 = 3.943+
Found by Maurizio Morandi
in April 2015.

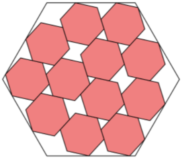


s = 3.93

12.



s = 4
Trivial.



s = 3.94

None of these results are impossible to obtain with standard tools. They just substantially reduce the amount of effort needed.

Comparing battle plans with AlphaEvolve

Expert advice: how much does this help and how to do it best?

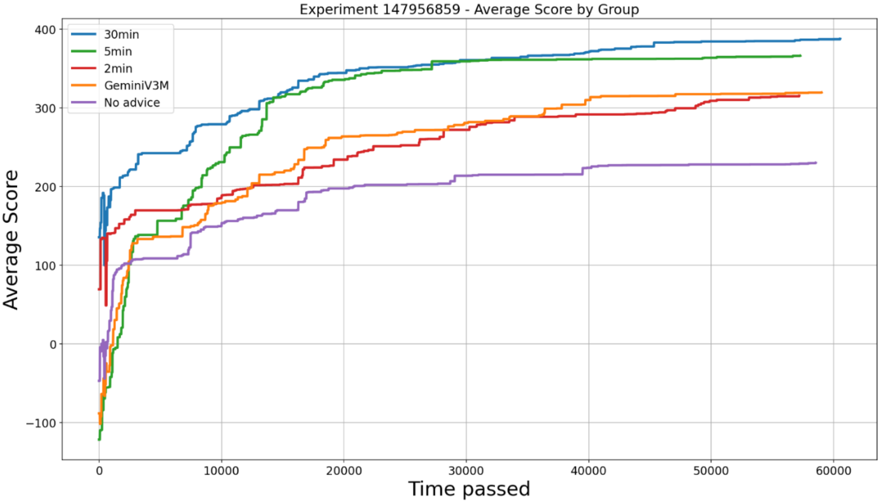
We came up with the following arbitrary problem to test this with:

Problem

Find a graph with 50 vertices, that maximises the quantity
#edges + 5 (#triangles - (#4cycles - 5)^2)

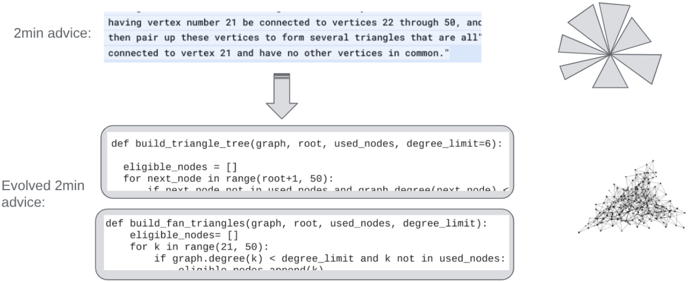
This problem is completely arbitrary, hopeless to solve by hand. We asked three of our colleagues to spend 2/5/30 minutes on this problem, and write down some general advice that we will give to AlphaEvolve.

Comparing battle plans with AlphaEvolve



Comparing battle plans with AlphaEvolve

Each experiment retained the original advice's characteristics, but evolved them into a stronger version



AlphaEvolve takes whatever idea you give to it, and tries to squeeze as much out of it as possible. This gives us an easy way to compare different strategies

What problems are a good fit for these tools?

If we put a thousand eager undergraduates in a room and give them this problem, how likely is it that they will succeed?

- AlphaEvolve is like having 1000 eager undergraduates in a room, who are excited to work on your problem
- They will read every possible paper they think is related, and try to combine the ideas in them in all sorts of crazy ways, whether they understand them or not
- They will zoom in on any idea that gets them a high score

If a problem can be solved this way, then AlphaEvolve will do well on it. If the problem needs genuine new ideas, AlphaEvolve will probably not find it.

YOU CANNOT FEED TWO BIRDS WITH ONE SCORE: THE ACCURACY-NATURALNESS TRADEOFF IN TRANSLATION

GERGELY FLAMICH

13/08/2025

GERGELY-FLAMICH.GITHUB.IO

IN COLLABORATION WITH

THE GOALS OF TRANSLATION

ACCURACY

ACCURACY

Translation carries the meaning of the **source text**

ACCURACY

Translation carries the meaning of the **source text**

NATURALNESS

NATURALNESS

Translation sounds good in **target language**

NATURALNESS

Translation sounds good in **target language**

MEASURING TRANSLATION QUALITY

INGREDIENTS

INGREDIENTS

- Have dataset of **source text**

INGREDIENTS

- Have dataset of **source text**
- Have dataset of **human reference translations**

INGREDIENTS

- Have dataset of **source text**
- Have dataset of **human reference translations**
- **Translation system** $Q_{y|x}$ to translate source text

HUMAN EVALUATIONS

HUMAN EVALUATIONS

Multidimensional Quality Metrics (MQM) [2]

HUMAN EVALUATIONS

Multidimensional Quality Metrics (MQM) [2]

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize distinct errors.

CLASSIC AUTOMATED METRICS

CLASSIC AUTOMATED METRICS

Examples: BLEU, chrF

CLASSIC AUTOMATED METRICS

Examples: BLEU, chrF

Purely symbolic: compare to human reference translation

CLASSIC AUTOMATED METRICS

Examples: BLEU, chrF

Purely symbolic: compare to human reference translation

✅ simple

CLASSIC AUTOMATED METRICS

Examples: BLEU, chrF

Purely symbolic: compare to human reference translation

✅ simple

❌ limited by the human reference

NEURAL METRICS

NEURAL METRICS

Examples: MetricX, Comet

NEURAL METRICS

Examples: MetricX, Comet

Large language model-based: predict MQM scores

NEURAL METRICS

Examples: MetricX, Comet

Large language model-based: predict MQM scores

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

NEURAL METRICS

Examples: MetricX, Comet

Large language model-based: predict MQM scores

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

✅ Jointly assess accuracy and naturalness

FEEDING TWO BIRDS WITH ONE SCORE

FEEDING TWO BIRDS WITH ONE SCORE

Should we assess accuracy **and** naturalness jointly?

FEEDING TWO BIRDS WITH ONE SCORE

Should we assess accuracy **and** naturalness jointly?

English→German

Rank	System	Human	AutoRank
1-11	GPT-4	-1.6	1.8
1-7	Dubformer	-1.8	1.8
2-10	ONLINE-B	-1.9	1.8
2-10	TranssionMT	-1.9	1.8
2-9	Unbabel-Tower70B	-1.9	1.0
1-9	HUMAN-B	-2.0	-
2-12	Mistral-Large	-2.1	2.0
4-11	CommandR-plus	-2.3	2.0
8-10	ONLINE-W	-2.3	2.2
2-12	Claude-3.5	-2.4	1.9
3-13	HUMAN-A	-2.5	-
10-12	IOL-Research	-2.5	2.3
5-13	Gemini-1.5-Pro	-2.8	2.2

FEEDING TWO BIRDS WITH ONE SCORE

Should we assess accuracy **and** naturalness jointly?

English→German

Rank	System	Human	AutoRank
1-11	GPT-4	-1.6	1.8
1-7	Dubformer	-1.8	1.8
2-10	ONLINE-B	-1.9	1.8
2-10	TranssionMT	-1.9	1.8
2-9	Unbabel-Tower70B	-1.9	1.0
1-9	HUMAN-B	-2.0	-
2-12	Mistral-Large	-2.1	2.0
4-11	CommandR-plus	-2.3	2.0
8-10	ONLINE-W	-2.3	2.2
2-12	Claude-3.5	-2.4	1.9
3-13	HUMAN-A	-2.5	-
10-12	IOL-Research	-2.5	2.3
5-13	Gemini-1.5-Pro	-2.8	2.2

FEEDING TWO BIRDS WITH ONE SCORE

Should we assess accuracy **and** naturalness jointly?

English→German

Rank	System	Human	AutoRank
1-11	GPT-4	-1.6	1.8
1-7	Dubformer	-1.8	1.8
2-10	ONLINE-B	-1.9	1.8
2-10	TranssionMT	-1.9	1.8
2-9	Unbabel-Tower70B	-1.9	1.0
1-9	HUMAN-B	-2.0	-
2-12	Mistral-Large	-2.1	2.0
4-11	CommandR-plus	-2.3	2.0
8-10	ONLINE-W	-2.3	2.2
2-12	Claude-3.5	-2.4	1.9
3-13	HUMAN-A	-2.5	-
10-12	IOL-Research	-2.5	2.3
5-13	Gemini-1.5-Pro	-2.8	2.2

Table from WMT24 findings paper [1].

WHERE WE ARE

WHERE WE ARE

❌ No formal notion of accuracy or naturalness

WHERE WE ARE

- ❌ No formal notion of accuracy or naturalness
- ❌ Their interaction not well understood

WHERE WE ARE

- ❌ No formal notion of accuracy or naturalness
- ❌ Their interaction not well understood
- ❌ Community uses single-score assessments

WHERE WE ARE

- ❌ No formal notion of accuracy or naturalness
- ❌ Their interaction not well understood
- ❌ Community uses single-score assessments
- ❌ Results don't seem to align with human evals

INFORMATION THEORY TO THE RESCUE

INFORMATION THEORY TO THE RESCUE

💡 Blau and Michaeli [3] already solved a similar issue!

INFORMATION THEORY TO THE RESCUE

💡 Blau and Michaeli [3] already solved a similar issue!

$$\text{Accuracy} \leftrightarrow \text{Distortion}$$

INFORMATION THEORY TO THE RESCUE

💡 Blau and Michaeli [3] already solved a similar issue!

Accuracy $\leftrightarrow$ Distortion

Naturalness $\leftrightarrow$ Realism/Distinguishability

INGREDIENTS

INGREDIENTS

- Source sentence x

INGREDIENTS

- Source sentence x
- Translation system $Q_{y|x}$

INGREDIENTS

- Source sentence x
- Translation system $Q_{y|x}$
- Reference translation $y^r \sim Q_{y|x}^{\text{human}}$

INGREDIENTS

- Source sentence x
- Translation system $Q_{y|x}$
- Reference translation $y^r \sim Q_{y|x}^{\text{human}}$
- Hypothesis/candidate $y^c \sim Q_{y|x}$

ACCURACY

ACCURACY

- Distortion metric: $\Delta(x, y^r, y^c) \geq 0$

ACCURACY

- Distortion metric: $\Delta(x, y^r, y^c) \geq 0$
- Accuracy: average negative distortion

ACCURACY

- Distortion metric: $\Delta(x, y^r, y^c) \geq 0$
- Accuracy: average negative distortion

$$A(Q_{\mathbf{y}|\mathbf{x}}) = -\mathbb{E}_{\mathbf{x},\mathbf{y}^r \sim P_{\mathbf{x},\mathbf{y}}}[\mathbb{E}_{\mathbf{y}^c \sim Q_{\mathbf{y}|\mathbf{x}}}[\Delta(\mathbf{x}, \mathbf{y}^r, \mathbf{y}^c)]]$$

NATURALNESS: INTUITION

NATURALNESS: INTUITION

NATURALNESS: INTUITION

NATURALNESS: DEFINITION I

NATURALNESS: DEFINITION I

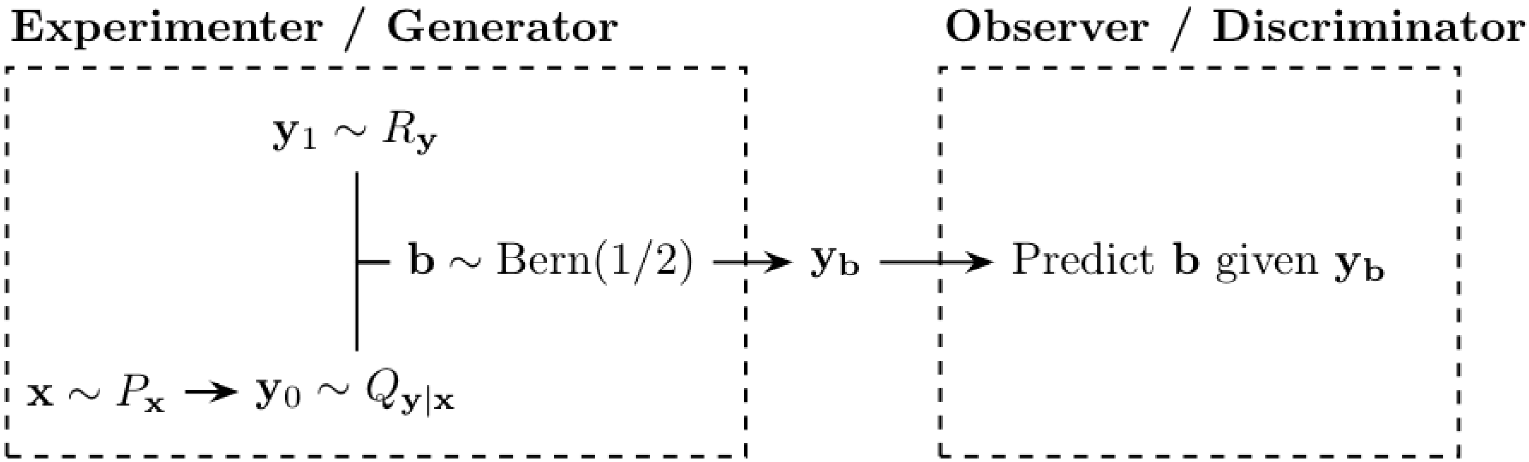
- Identify naturalness with distinguishability

NATURALNESS: DEFINITION I

- Identify naturalness with distinguishability
- We pick a **reference distribution** R_y

NATURALNESS: DEFINITION I

- Identify naturalness with distinguishability
- We pick a **reference distribution** R_y



NATURALNESS: DEFINITION II

NATURALNESS: DEFINITION II

- Translation marginal $Q_y = \mathbb{E}_{x \sim P_x}[Q_{y|x}]$

NATURALNESS: DEFINITION II

- Translation marginal $Q_y = \mathbb{E}_{x \sim P_x}[Q_{y|x}]$
- Distinguishability **is equivalent** to statistical distance $D(Q_y, R_y)$

NATURALNESS: DEFINITION II

- Translation marginal $Q_y = \mathbb{E}_{x \sim P_x}[Q_{y|x}]$
- Distinguishability **is equivalent** to statistical distance $D(Q_y, R_y)$

$$N(Q_{y|x}) = -D(Q_y, R_y)$$

THE DIVERGENCE

THE DIVERGENCE

Let $P(f) = \mathbb{E}_{X \sim P}[f(X)]$

THE DIVERGENCE

Let $P(f) = \mathbb{E}_{X \sim P}[f(X)]$

Integral probability metric (IPM):

THE DIVERGENCE

Let $P(f) = \mathbb{E}_{X \sim P}[f(X)]$

Integral probability metric (IPM):

$$\mathrm{IPM}_{\mathcal{F}}[Q \| P] = \sup_{f \in \mathcal{F}} |Q(f) - P(f)|$$

THE DIVERGENCE

Let $P(f) = \mathbb{E}_{X \sim P}[f(X)]$

Integral probability metric (IPM):

$$\mathrm{IPM}_{\mathcal{F}}[Q \| P] = \sup_{f \in \mathcal{F}} |Q(f) - P(f)|$$

Optimal critic $f^\star$

THE DIVERGENCE

Let $P(f) = \mathbb{E}_{X \sim P}[f(X)]$

Integral probability metric (IPM):

$$\mathrm{IPM}_{\mathcal{F}}[Q \| P] = \sup_{f \in \mathcal{F}} |Q(f) - P(f)|$$

Optimal critic $f^\star$

$$f^\star = \mathrm{argmax}_{f \in \mathcal{F}} |Q(f) - P(f)|$$

THE EQUIVALENCE

THE EQUIVALENCE

Set $\epsilon = \mathbb{P}[b = 1]$. Then:

THE EQUIVALENCE

Set $\epsilon = \mathbb{P}[b = 1]$. Then:

$$L(b, \alpha) = \begin{cases} -\alpha/\epsilon & \text{if } b = 1 \\ \alpha/(1-\epsilon) & \text{if } b = -1 \end{cases}$$

THE EQUIVALENCE

Set $\epsilon = \mathbb{P}[b = 1]$. Then:

$$L(b, \alpha) = \begin{cases} -\alpha/\epsilon & \text{if } b = 1 \\ \alpha/(1-\epsilon) & \text{if } b = -1 \end{cases}$$

Then:

THE EQUIVALENCE

Set $\epsilon = \mathbb{P}[b = 1]$. Then:

$$L(b, \alpha) = \begin{cases} -\alpha/\epsilon & \text{if } b = 1 \\ \alpha/(1-\epsilon) & \text{if } b = -1 \end{cases}$$

Then:

$$R^L_{\mathcal{F}} = \inf_{f \in \mathcal{F}} \mathbb{E}[L(\mathbf{b}, f(\mathbf{y_b}))] = -\mathbf{IPM}_{\mathcal{F}}[Q \| P]$$

ARE PERFECT ACCURACY AND NATURALNESS THE SAME?

ARE PERFECT ACCURACY AND NATURALNESS THE SAME?

Perfect naturalness $\Rightarrow$ perfect accuracy?

ARE PERFECT ACCURACY AND NATURALNESS THE SAME?

Perfect naturalness $\Rightarrow$ perfect accuracy?

Perfect accuracy $\Rightarrow$ perfect naturalness?

ARE PERFECT ACCURACY AND NATURALNESS THE SAME?

Perfect naturalness $\Rightarrow$ perfect accuracy?

Perfect accuracy $\Rightarrow$ perfect naturalness?

- No, according to Blau and Michaeli's setup [3]

ARE PERFECT ACCURACY AND NATURALNESS THE SAME?

Perfect naturalness $\Rightarrow$ **perfect accuracy**?

Perfect accuracy $\Rightarrow$ **perfect naturalness**?

- No, according to Blau and Michaeli's setup [3]
- No, according to our setup [4]

WHAT IS THE TRADEOFF LIKE?

WHAT IS THE TRADEOFF LIKE?

Accuracy-naturalness function:

WHAT IS THE TRADEOFF LIKE?

Accuracy-naturalness function:

$$A(N) = \max_{Q_{\mathbf{y}|\mathbf{x}}} \{-\Delta(Q_{\mathbf{y}|\mathbf{x}})\} \quad \text{subject to} \ -D(Q_{\mathbf{y}}, R_{\mathbf{y}}) \geq N$$

WHAT IS THE TRADEOFF LIKE?

Accuracy-naturalness function:

$$A(N) = \max_{Q_{\mathbf{y}|\mathbf{x}}} \{-\Delta(Q_{\mathbf{y}|\mathbf{x}})\} \quad \text{subject to } -D(Q_{\mathbf{y}}, R_{\mathbf{y}}) \geq N$$

- $A(N)$ is non-increasing

WHAT IS THE TRADEOFF LIKE?

Accuracy-naturalness function:

$$A(N) = \max_{Q_{\mathbf{y}|\mathbf{x}}} \{-\Delta(Q_{\mathbf{y}|\mathbf{x}})\} \quad \text{subject to} \; - D(Q_{\mathbf{y}}, R_{\mathbf{y}}) \geq N$$

- $A(N)$ is non-increasing
- If D convex in first slot, then $A(N)$ concave

APPROXIMATING THE CURVE

APPROXIMATING THE CURVE

💡 Use LLM scores to judge the translations!

APPROXIMATING THE CURVE

- 💡 Use LLM scores to judge the translations!
- 🤔 Does this correspond to some $D(Q, P)$?

APPROXIMATING THE CURVE

💡 Use LLM scores to judge the translations!

🤔 Does this correspond to some $D(Q, P)$?

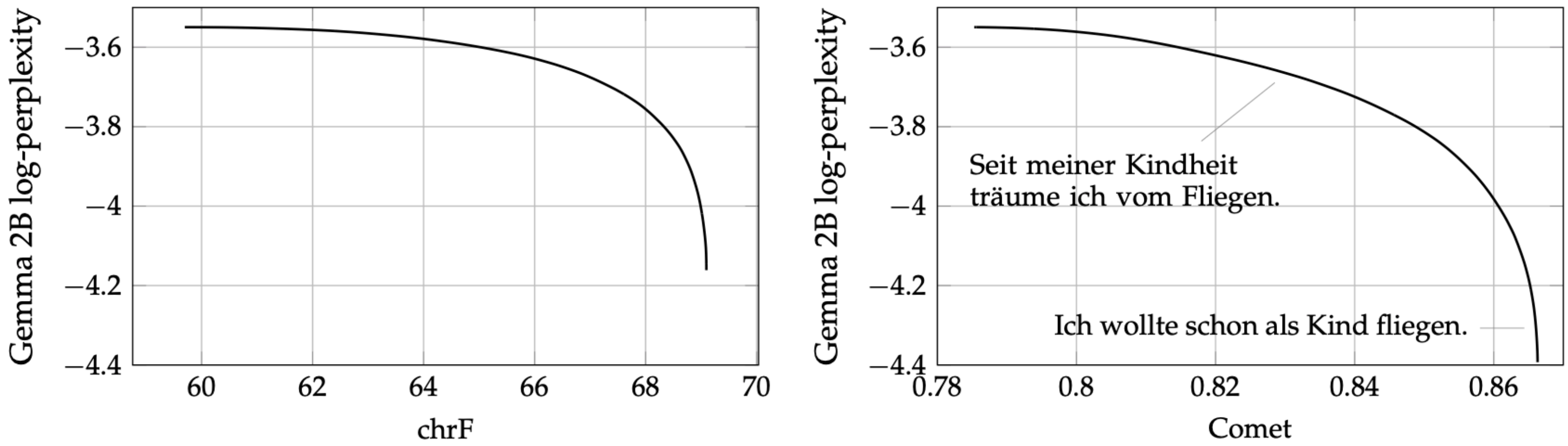
EN $\longrightarrow$ DE: I've wanted to fly since I was a child.

APPROXIMATING THE CURVE

💡 Use LLM scores to judge the translations!

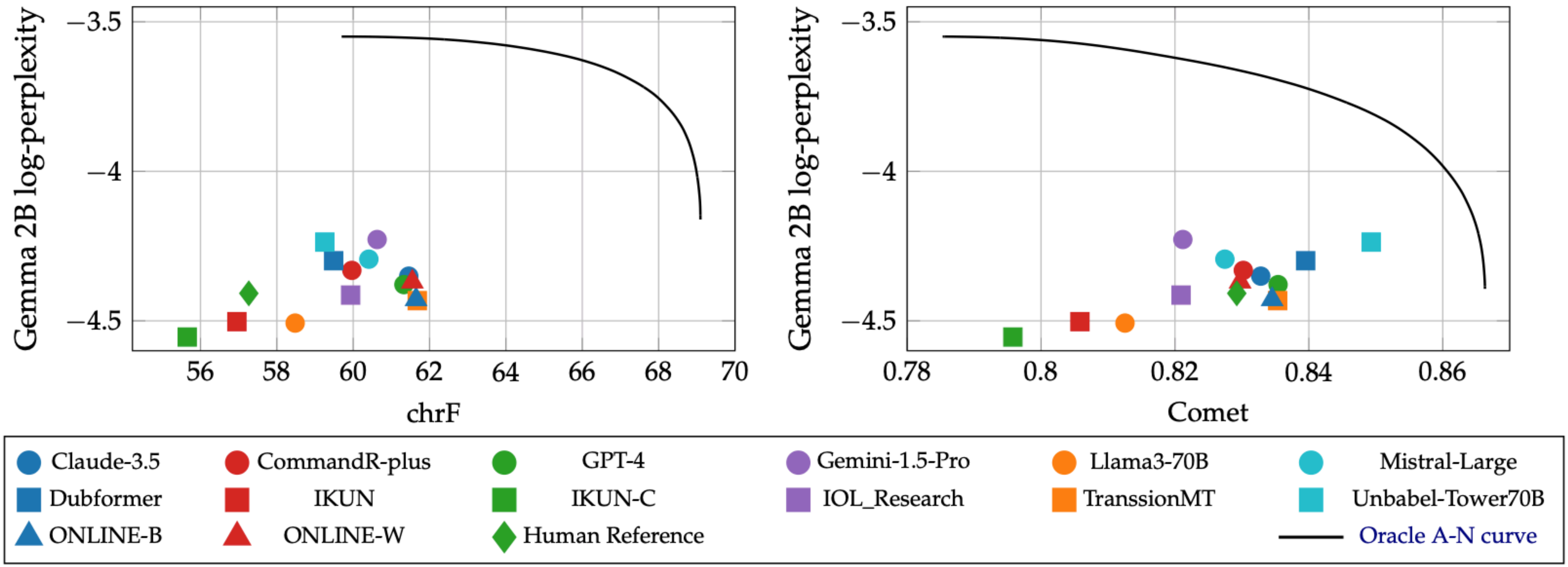
🤔 Does this correspond to some $D(Q, P)$?

EN $\longrightarrow$ DE: I've wanted to fly since I was a child.

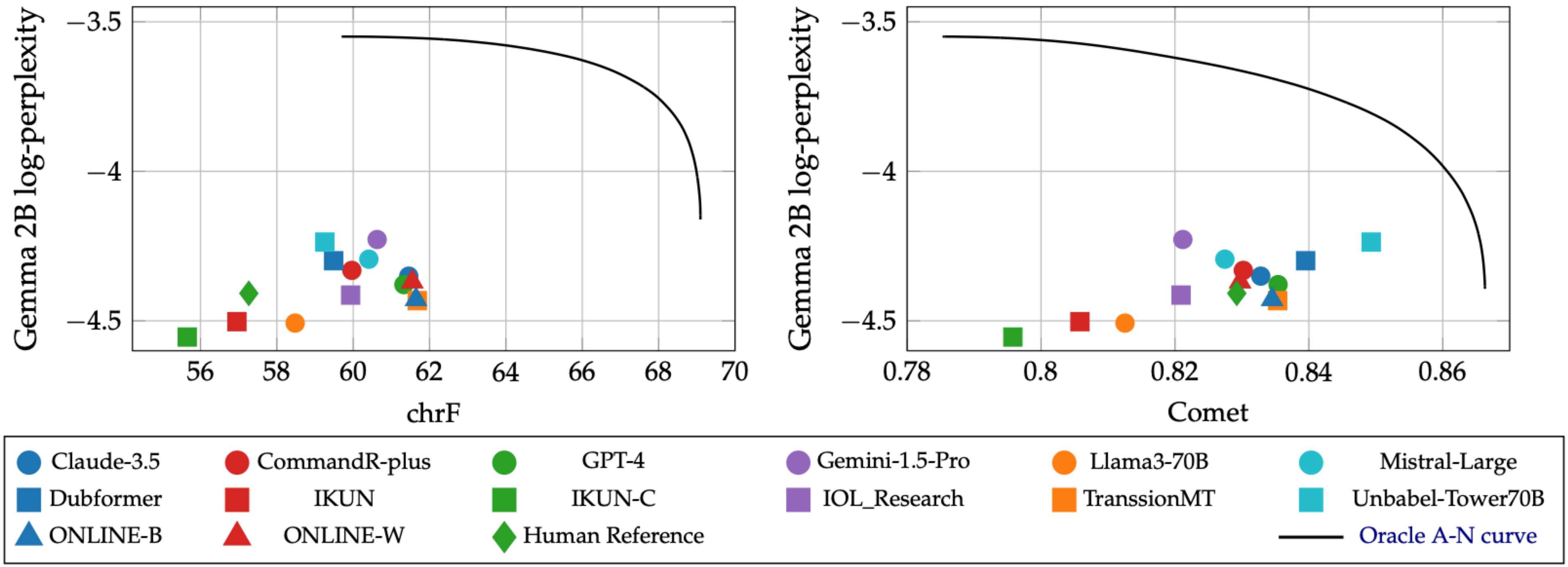


WHERE IS THE SOTA?

WHERE IS THE SOTA?



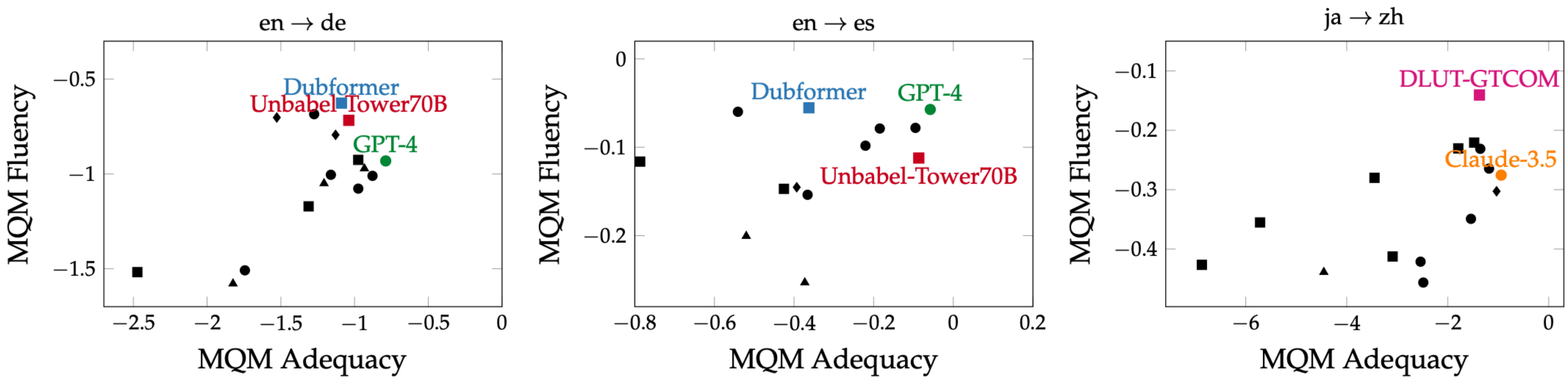
WHERE IS THE SOTA?



Close to the curve, accuracy and naturalness **anti-correlate**

WHERE IS THE SOTA?

WHERE IS THE SOTA?



THE ISSUE AND THE FIX

$$\mathrm{IPM}_{\mathcal{F}}[Q \| P] = \sup_{f \in \mathcal{F}} |Q(f) - P(f)|$$

THE ISSUE AND THE FIX

$$\mathrm{IPM}_{\mathcal{F}}[Q \| P] = \sup_{f \in \mathcal{F}} |Q(f) - P(f)|$$

❌ $f^\star$ depends on Q!

THE ISSUE AND THE FIX

$$\mathrm{IPM}_{\mathcal{F}}[Q\|P] = \sup_{f\in\mathcal{F}}|Q(f) - P(f)|$$

❌ $f^\star$ depends on Q!

✅ **Fix:** average instead of maximising

THE ISSUE AND THE FIX

$$\mathrm{IPM}_{\mathcal{F}}[Q\|P] = \sup_{f\in\mathcal{F}}|Q(f) - P(f)|$$

❌ $f^\star$ depends on Q!

✅ **Fix:** average instead of maximising

Let $f \sim \mathcal{P}$

THE ISSUE AND THE FIX

$$\mathrm{IPM}_{\mathcal{F}}[Q \| P] = \sup_{f \in \mathcal{F}} |Q(f) - P(f)|$$

❌ $f^\star$ depends on Q!

✅ **Fix:** average instead of maximising

Let $f \sim \mathcal{P}$

$$\begin{aligned} D_p(Q, P \mid \mathcal{P}) &= \mathcal{P}(|Q - P|^p)^{1/p} \\ &= \mathbb{E}_{f \sim \mathcal{P}}[|Q(f) - P(f)|^p]^{1/p} \end{aligned}$$

SOME INTERESTING PROPERTIES

SOME INTERESTING PROPERTIES

✅ D_p a metric under some sensible conditions

SOME INTERESTING PROPERTIES

✅ D_p a metric under some sensible conditions

✅ Can estimate without knowing Q:

$$D_1(Q, P \mid \mathcal{P}) \approx \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{m=1}^{M_Q} \frac{f_n(X_m)}{M_Q} - \sum_{m=1}^{M_P} \frac{f_n(Y}{M}\right.$$

SOME INTERESTING PROPERTIES

✅ D_p a metric under some sensible conditions

✅ Can estimate without knowing Q:

$$D_1(Q, P \mid \mathcal{P}) \approx \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{m=1}^{M_Q} \frac{f_n(X_m)}{M_Q} - \sum_{m=1}^{M_P} \frac{f_n(Y}{M} \right.$$

✅ When $\mathcal{P}$ is a GP, D_2 corresponds to MMD

CONTRIBUTIONS

CONTRIBUTIONS

✅ Proposed a formal definition of accuracy and naturalness

CONTRIBUTIONS

- ✅ Proposed a formal definition of accuracy and naturalness
- ✅ Extended the theory of Blau and Michaeli

CONTRIBUTIONS

- ✅ Proposed a formal definition of accuracy and naturalness
- ✅ Extended the theory of Blau and Michaeli
- ✅ Showed that tradeoff **must** exist in practice

CONTRIBUTIONS

- ✅ Proposed a formal definition of accuracy and naturalness
- ✅ Extended the theory of Blau and Michaeli
- ✅ Showed that tradeoff **must** exist in practice
- ✅ Assessed the performance of the current state-of-the-art

CONTRIBUTIONS

- ✅ Proposed a formal definition of accuracy and naturalness
- ✅ Extended the theory of Blau and Michaeli
- ✅ Showed that tradeoff **must** exist in practice
- ✅ Assessed the performance of the current state-of-the-art
- ✅ Showed connection between no-reference metrics and statistical distances

REFERENCES I

- [1] Kocmi et al. (2024). Findings of the WMT24 general machine translation shared task: the LLM era is here but mt is not solved yet. In Proceedings of the Ninth Conference on Machine Translation (pp. 1-46).
- [2] Freitag et al. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9, 1460-1474.

REFERENCES II

- [3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6228–6237, 2018.
- [4] F et al. (2025). You Cannot Feed Two Birds with One Score: the Accuracy-Naturalness Tradeoff in Translation. arXiv preprint arXiv:2503.24013.

REFERENCES III

- [5] Sriperumbudur et al. (2009). On integral probability metrics,ϕ-divergences and binary classification. arXiv preprint arXiv:0901.2698.

NSA

non-mathematician

novel

Non-standard attack on The Riemann Hypothesis (RH) with (sub)-standard AI ‘students'

mostly wrong

but getting better

Off-Critical Riemann Zeta Zeros Cannot Seed Symmetric Entire Functions: A Hyperlocal Proof of Constructive Impossibility

Author: Attila Csordas
Affiliation: AgeCurve Limited, Cambridge, UK
ORCID: 0000-0003-3576-1793

First Published: June 26, 2025
Current Version (v3.3): August 9, 2025

https://attila-ac.github.io/hyperlocal/

https://github.com/attila-ac/hyperlocal/

Attila Csordas, PhD
attila@agecurve.xyz
Cambridge, UK

window of opportunity for humans
open to solve
deep research math problems

✓

AgeCurve

What's on the table?

Aiming for easy verifiability

Core Proof: 10 top logical steps

~ 120-140 distinct logical steps

The Setup & Foundations (~30-35 steps)
Hypothesis (~2 steps)
Geometric Consequence: The Quartet (~5 steps)
Algebraic Consequence: The Factorization (~20 steps)
Dynamic Consequence: The Recurrence (~5-10 steps)
Analytical Proof of Instability (~20 steps)
Problem Reduction (~4 steps)
Derivation of Algebraic Constraints (~15 steps)
The Final Contradiction (~10 steps)
The Conclusion (~3 steps)

Total Length: ~125 Pages

Main Argument (Manuscript Foundations & Proof): ~80 Pages

Appendices (Support & Verification): ~40 Pages

Foundational Work (~35%)
The Core Argument (~40%)
Validation and Heuristics (~25%)

Versions out: 3 big, 8 in total

Estimation: 4 * 2 hours ~ 7- 8 hours at a chalkboard

Expectations on what to & no to achieve in 25 minutes

No prime consequences ✗

No point delivering proof, mid-game, endgame ✗

No point drowning in proof technicalities ✗

AI lessons, tips, opportunities ✓

Roles of books, walks and failures ✓

Introduce RH ✓

problems around problem ✓

Opening: main idea behind hyperlocal framework/strategy ✓

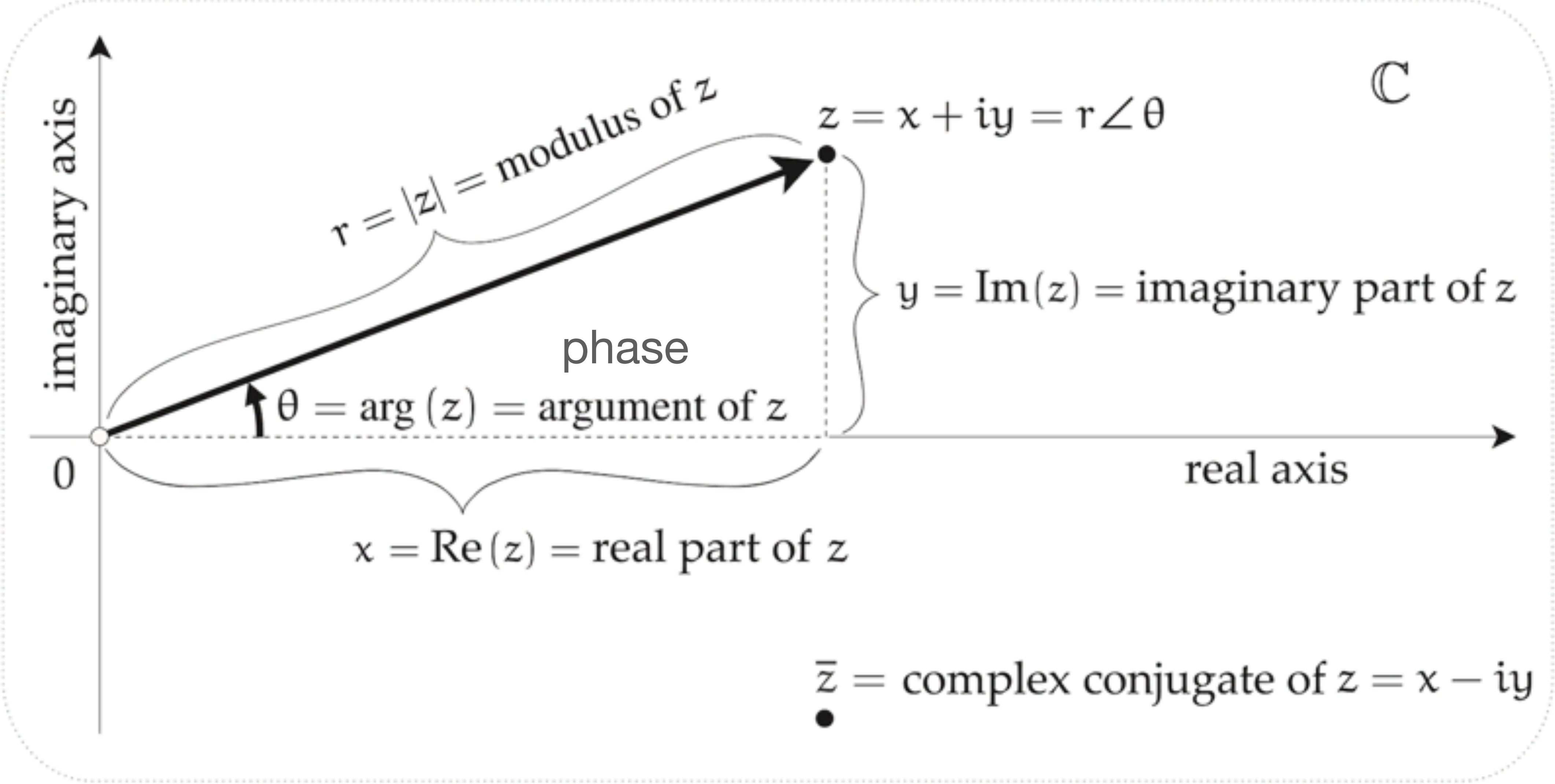
Hopefully some will study manuscript ✓

Project history, stats ✓

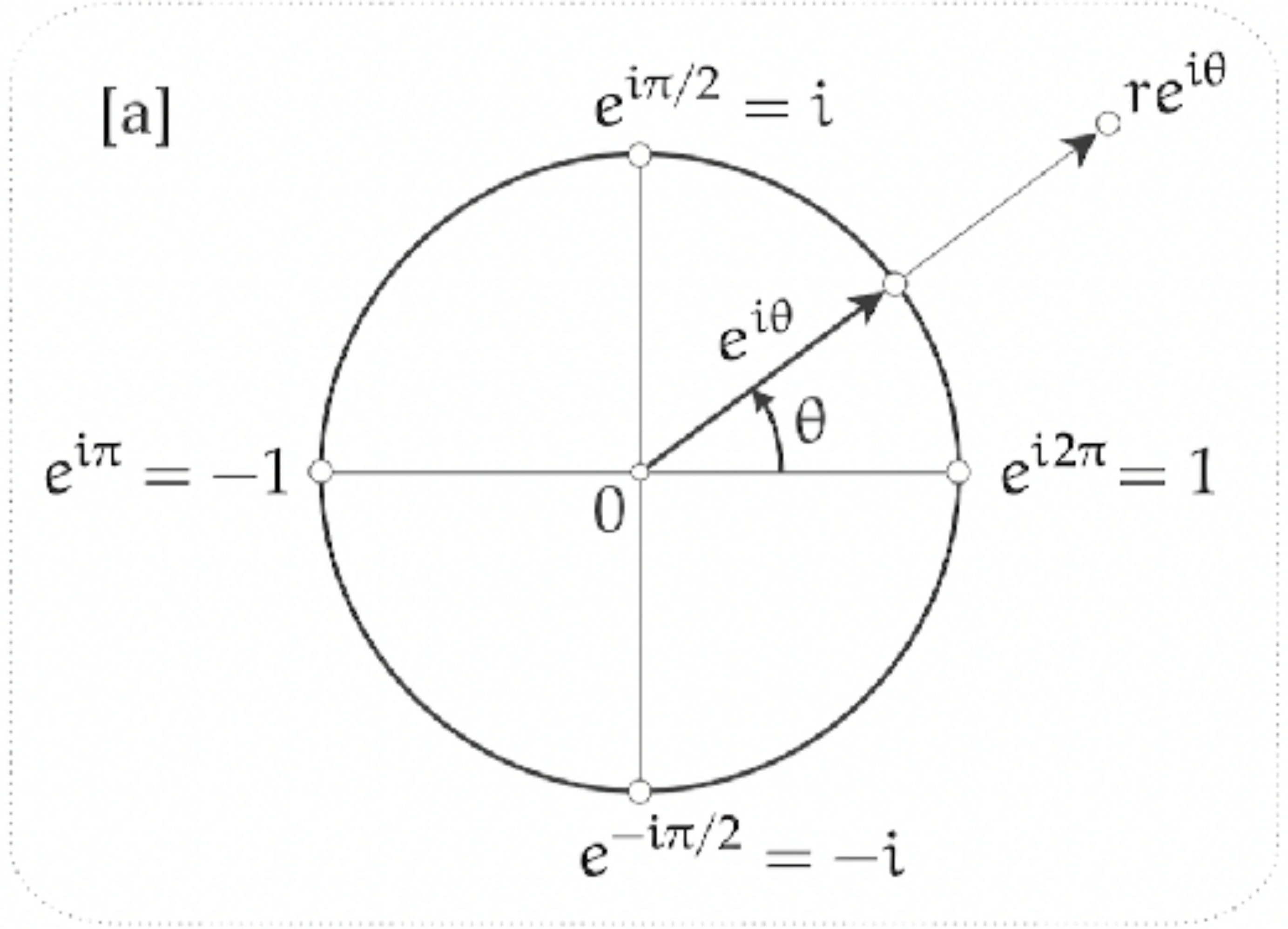
Case Studies ✓

AgeCurve

Complex Analysis Intro



[1.4] *Visual summary of the terminology and notation used to describe complex numbers.*



$$e^{i\theta} = \cos\theta + i\sin\theta$$

$$z = \bar{z} \iff z \in \mathbb{R}$$

conjugate $\bar{z}$ "mirror image" of z

polar form flips angle $\bar{z} = re^{-i\theta}$

$$r = |z| = \sqrt{x^2 + y^2}$$

Reality Condition (RC) and Conjugate Symmetry

$$f(\bar{s}) = \overline{f(s)} \quad \text{for all } s \text{ in its domain}$$

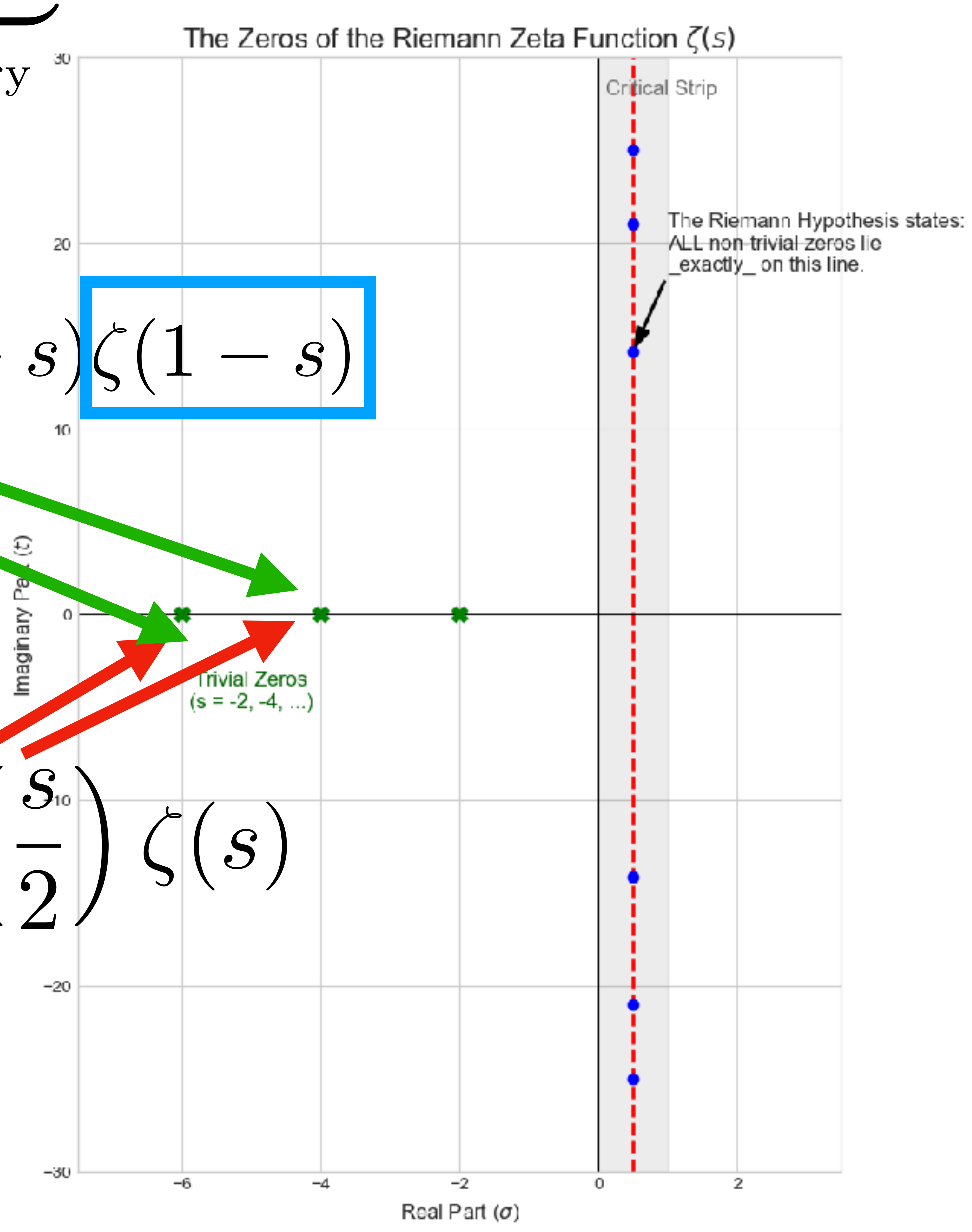
$$\xi(\bar{\rho'}) = \overline{\xi(\rho')} = \bar{0} = 0$$

The Riemann Hypothesis (RH) Conjecture

$$\zeta(s) = \underbrace{\sum_{n=1}^{\infty} \frac{1}{n^s}}_{\text{Complex Analysis}} = \underbrace{\prod_{p \text{ prime}} \frac{1}{1-p^{-s}}}_{\text{Number Theory}}, \quad \text{for } \operatorname{Re}(s) > 1$$

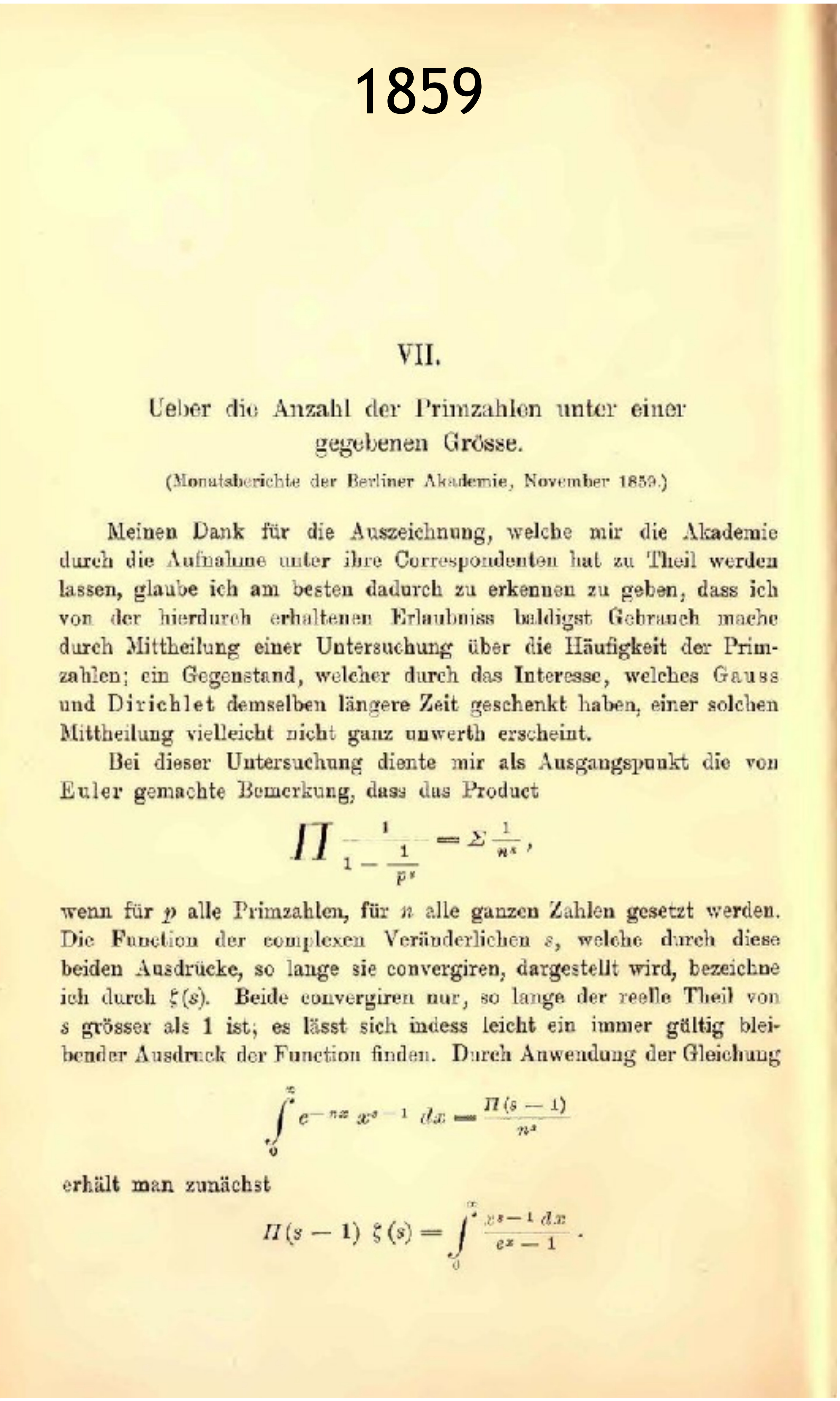
$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) \zeta(1-s)$$

functional equation (FE)



$$\xi(s) := \frac{1}{2} s(s-1) \pi^{-s/2} \Gamma\left(\frac{s}{2}\right) \zeta(s)$$

FE: $\xi(s) = \xi(1-s)$



1859

VII.

Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse.

(Monatsberichte der Berliner Akademie, November 1859.)

Meinen Dank für die Auszeichnung, welche mir die Akademie durch die Aufnahme unter ihre Correspondenten hat zu Theil werden lassen, glaube ich am besten dadurch zu erkennen zu geben, dass ich von der hierdurch erhaltenen Erlaubniss baldigst Gebrauch mache durch Mittheilung einer Untersuchung über die Häufigkeit der Primzahlen; ein Gegenstand, welcher durch das Interesse, welches Gauss und Dirichlet demselben längere Zeit geschenkt haben, einer solchen Mittheilung vielleicht nicht ganz unwerth erscheint.

Bei dieser Untersuchung diente mir als Ausgangspunkt die von Euler gemachte Bemerkung, dass das Product

$$\Pi \frac{1}{1-\frac{1}{p^s}} = \Sigma \frac{1}{n^s},$$

wenn für p alle Primzahlen, für n alle ganzen Zahlen gesetzt werden. Die Function der complexen Veränderlichen s, welche durch diese beiden Ausdrücke, so lange sie convergiren, dargestellt wird, bezeichne ich durch $\zeta(s)$. Beide convergiren nur, so lange der reelle Theil von s grösser als 1 ist; es lässt sich indess leicht ein immer gültig bleibender Ausdruck der Function finden. Durch Anwendung der Gleichung

$$\int_0^\infty e^{-nx} x^{s-1}\, dx = \frac{\Pi(s-1)}{n^s}$$

erhält man zunächst

$$\Pi(s-1)\, \zeta(s) = \int_0^\infty \frac{x^{s-1}\, dx}{e^x - 1}.$$

Case Study 1: Multiplicity of Non-Trivial Zeros

smallest non-negative integer where derivative does not vanish

$$f(s_0) = f'(s_0) = \cdots = f^{(k-1)}(s_0) = 0, \quad \text{but} \quad f^{(k)}(s_0) \neq 0.$$

Equivalently, in terms of the Taylor series expansion around s_0:

$$f(s) = \sum_{n=k}^{\infty} \frac{f^{(n)}(s_0)}{n!}(s - s_0)^n = \frac{f^{(k)}(s_0)}{k!}(s - s_0)^k + \frac{f^{(k+1)}(s_0)}{(k+1)!}(s - s_0)^{k+1} + \dots$$

The first non-zero term in the expansion is the one corresponding to $(s - s_0)^k$.

A zero of order $k = 1$ is called a simple zero. For a simple zero s_0, we have:

$$f(s_0) = 0 \quad \text{and} \quad f'(s_0) \neq 0.$$

The Taylor series near a simple zero starts with a linear term:

$$f(s) = f'(s_0)(s - s_0) + O((s - s_0)^2).$$

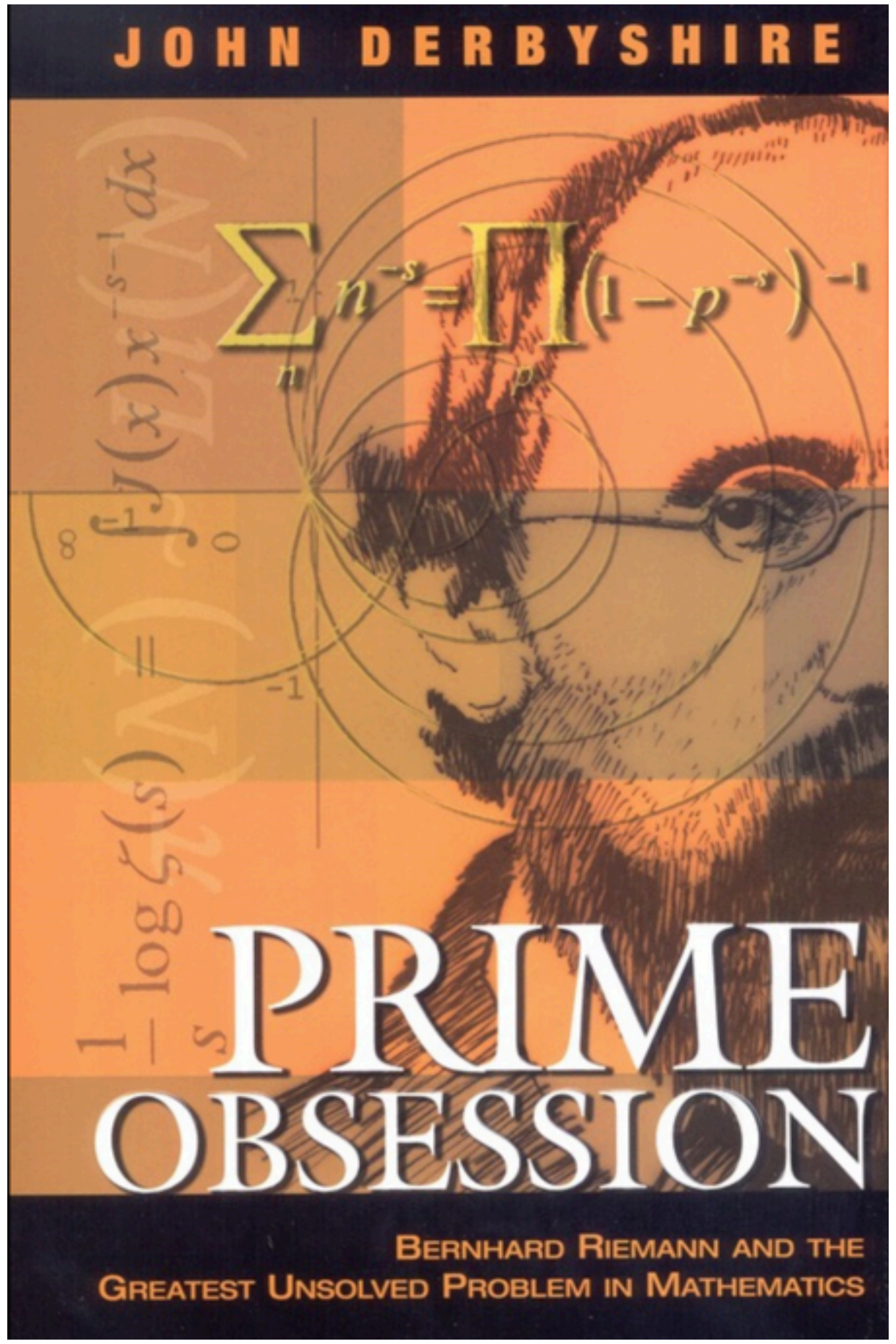
If $f(s_0) = 0$ and $f'(s_0) = 0$ but $f''(s_0) \neq 0$, then s_0 is a zero of order 2 (a double zero), and the Taylor series starts $f(s) = \frac{f''(s_0)}{2}(s - s_0)^2 + \dots$.

Simple Zero Conjecture: All Zeta Complex Zeros are Simple

All proof attempts not handling all zero orders are conditional upon SZC

Majority opinion in math community: RH is very likely true

Prime counting function up to $x, \pi(x)$, closely approximated by log integral: $\pi(x) \approx \mathrm{Li}(x)$



RH is equivalent to stating this error is tightly bounded: $|\pi(x) - \mathrm{Li}(x)| = O(\sqrt{x}\log x)$

weighted prime-counting function

$$\psi(x) \approx \underbrace{x}_{\substack{\text{The "Signal"} \\ \text{(Predictable Average Trend)}}} - \underbrace{\sum_{\rho} \frac{x^{\rho}}{\rho}}_{\substack{\text{The "Noise"} \\ \text{(Corrective Waves from Zeros)}}}$$

Annals of Mathematics and Philosophy
vol. 1, n° 1, 2023 $M \times \Phi$

What Makes Mathematicians Believe Unproved Mathematical Statements?

Timothy Gowers(1)

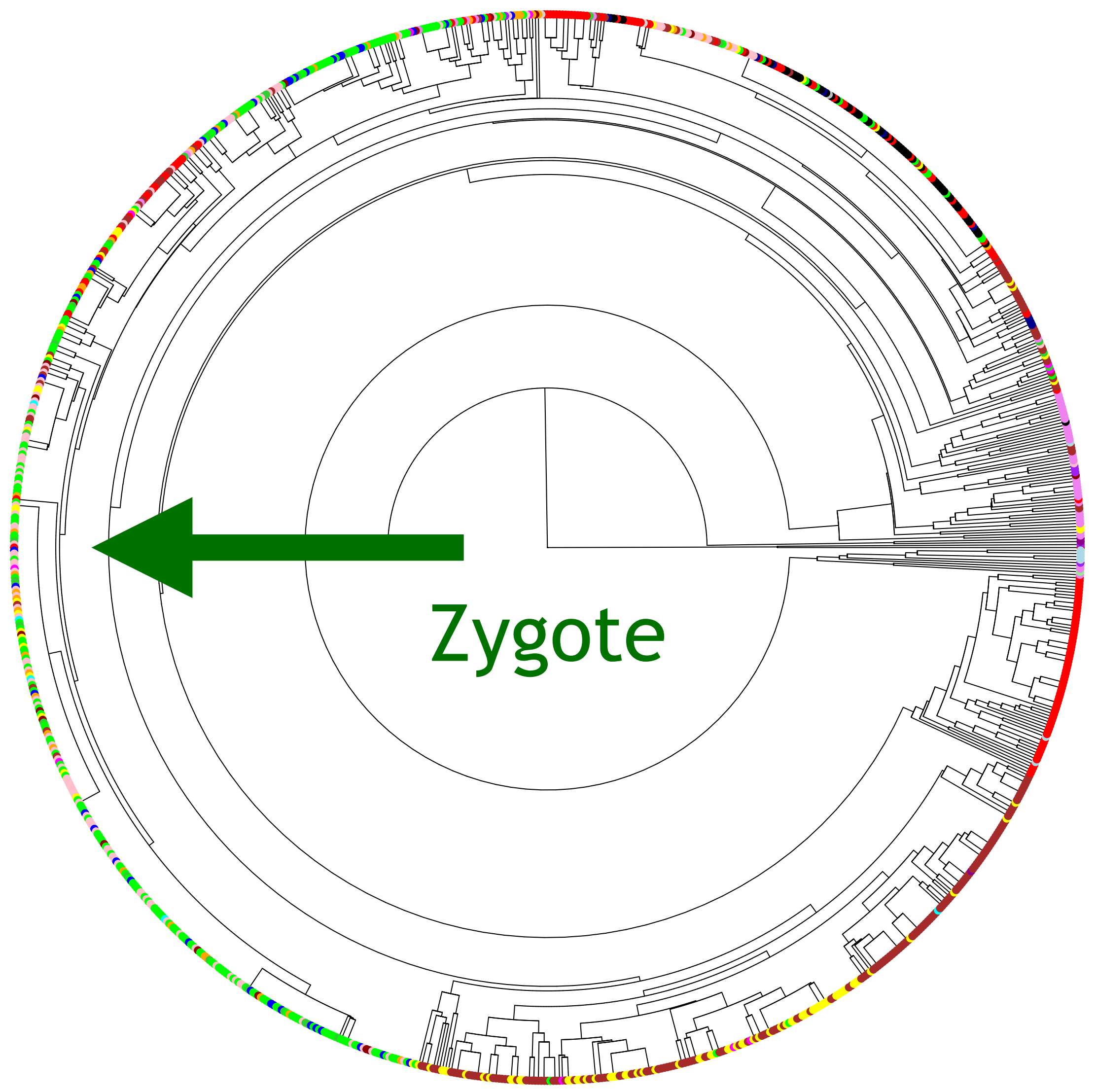
1. ~3 trillion confirmed
2. natural reformulations/equivalences
3. other zeta function theorems

Nothing like the P versus NP problem, much more concrete

2023/24 realisation: Window of Opportunity is Open

1. likely true
2. demand is high as no pro mathematician has solved it in last 160+ years
3. maybe amateurs do have non-zero chances
4. AI speed-up of proof testing and formulation

Why Me: Biology + Philosophy -> Math



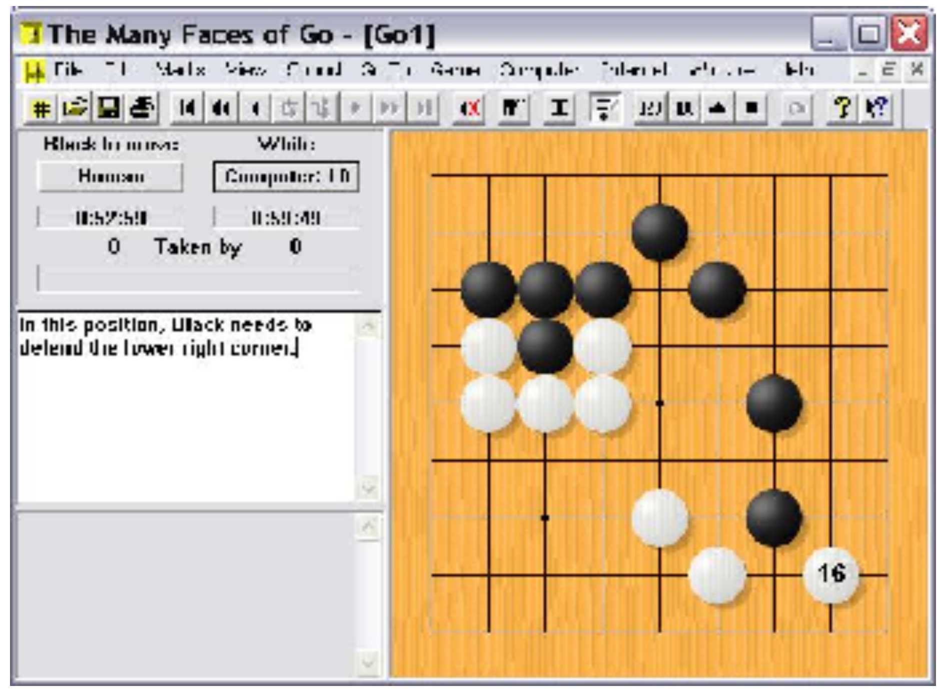
working on biological aging professionally: almost intractable

experimental spirit, critical anomalies, diagnostics

amateur with fresh eyes & zero credit to loose

latest project Cell Tree Age/Calculus

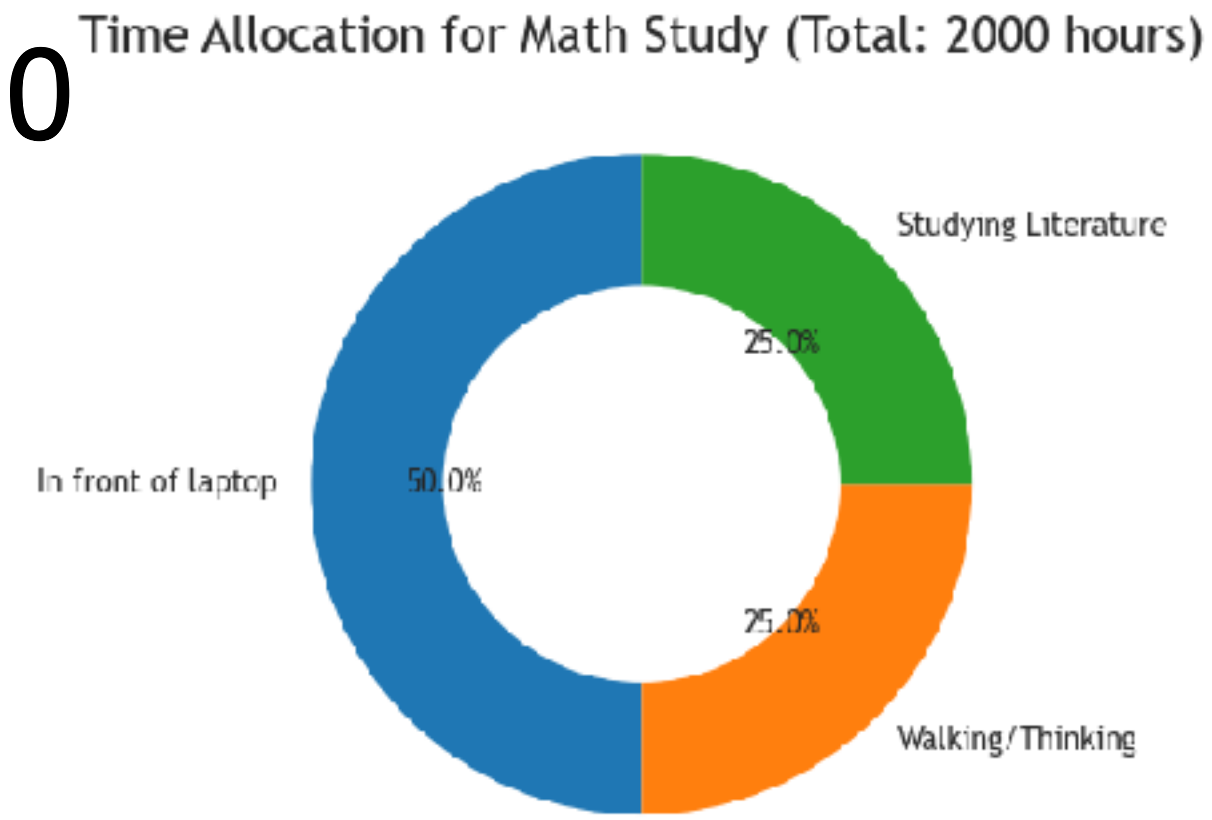
philosophy, logic, proof crafting, NSA



independence/time allocation

obsession

Timeline/Stats



Complex Analysis
voice

3+ failed versions

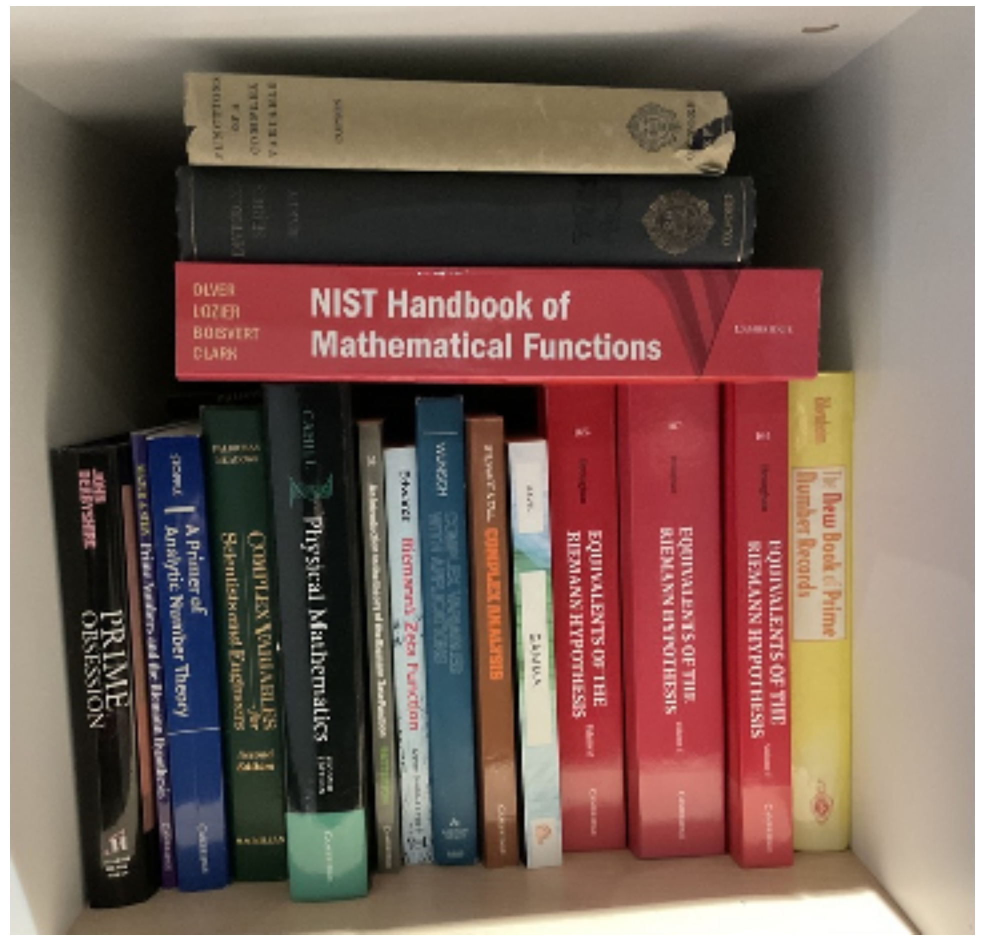
Abstract Algebra

Quartet Zero Diagnostics

Hyperlocal Turn

attila-ac / hyperlocal

2023 Fall



2024 December

Algebraic Geometry Topology

2025 February

Failed Fourier Distributional approach

2025 March

2025 April

20+ bigger fixes

2025 June

2025 August

100+ llm threads

AgeCurve

1000 hours ~ 125 workdays ~ 6 months of workdays ~ 3 calendar months full time

Working on Deep Research Math can be Healthy

Team: Non-Standard Human PI with sub-standard AI ‘students’

'author extends critical gratitude to some llm versions of some tech companies for providing knowledge shortcuts and assistance in proof formulation, significantly expediting the process of original human creativity'

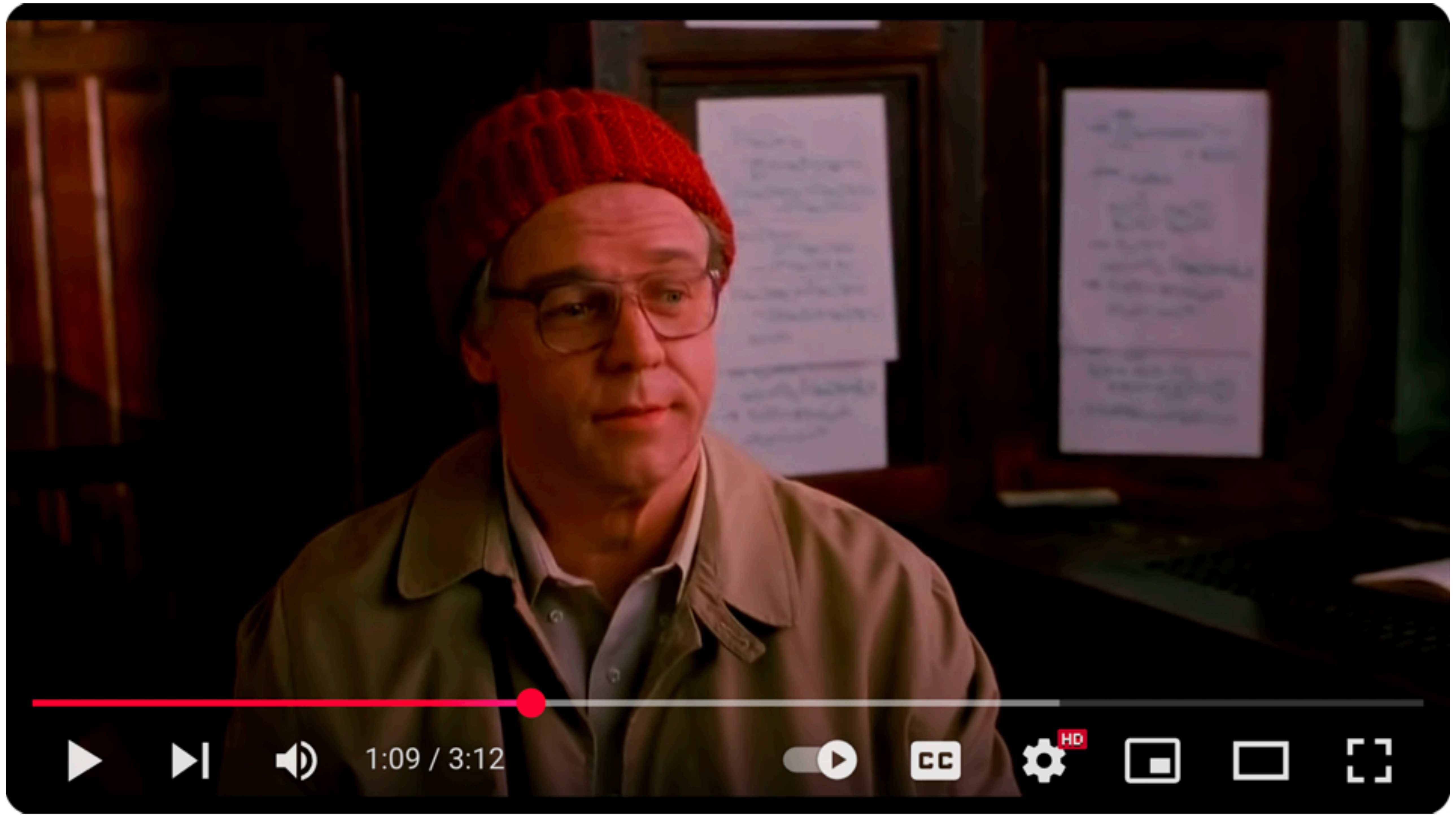
Me: PI,
machines: my more or less capable 'students'
trying some subtasks but without egos,
need machines to unpack and dress up my own
thoughts for doing my math creatively and freely

Problems around the Conjecture: Mathematicians think its far away

"For some problems the tools are not there. It doesn't matter how smart or quick you are. The analogy I have is like climbing. If you want to climb a cliff that's 10 meters high, you can probably do it with tools and equipment. But if it's just a sheer cliff face, a mile high and there's no handholds whatsoever, just forget it. Doesn't matter how strong you are, you need to wait for some breakthrough, an opening that occurs halfway up and now you have some easier sub-goal."

Kambei Shimada: 'You embarrass me. You're overestimating me. Listen, I'm not a man with any special skill, but I've had plenty of experience in battles; losing battles.'

Problems around the Conjecture: considered ‘lunatic’ territory



NASH
I've been working on solving the Riemann Hypothesis...

Nash slides his pad over to Fox.

NASH
If I dazzle them, they'll have to reinstate me. But the medication makes me blurry.

Fox's smile tries to hide the fact that Nash's calculations don't seem to make much sense. He slides back the pad.

STUDENT
You are John Nash, right?

Nash looks at him and smiles.

STUDENT
You solved the Riemann Hypothesis.

NASH
Actually there's an error in my last line of code. But I'm getting there.

Twin Problem around RH for Amateurs: Mathematical Sociology

easy to hit a wall

Dear Author,

thanks for your message. Unfortunately, we cannot consider your article for publication as it stands.

Indeed, we receive many failed attempts of resolutions of some of the Millennium Problems, and in particular of the Riemann Hypothesis, and we must therefore ask the endorsement of renowned experts in the field in order to consider submissions claiming to have solved such a problem.

Incentive

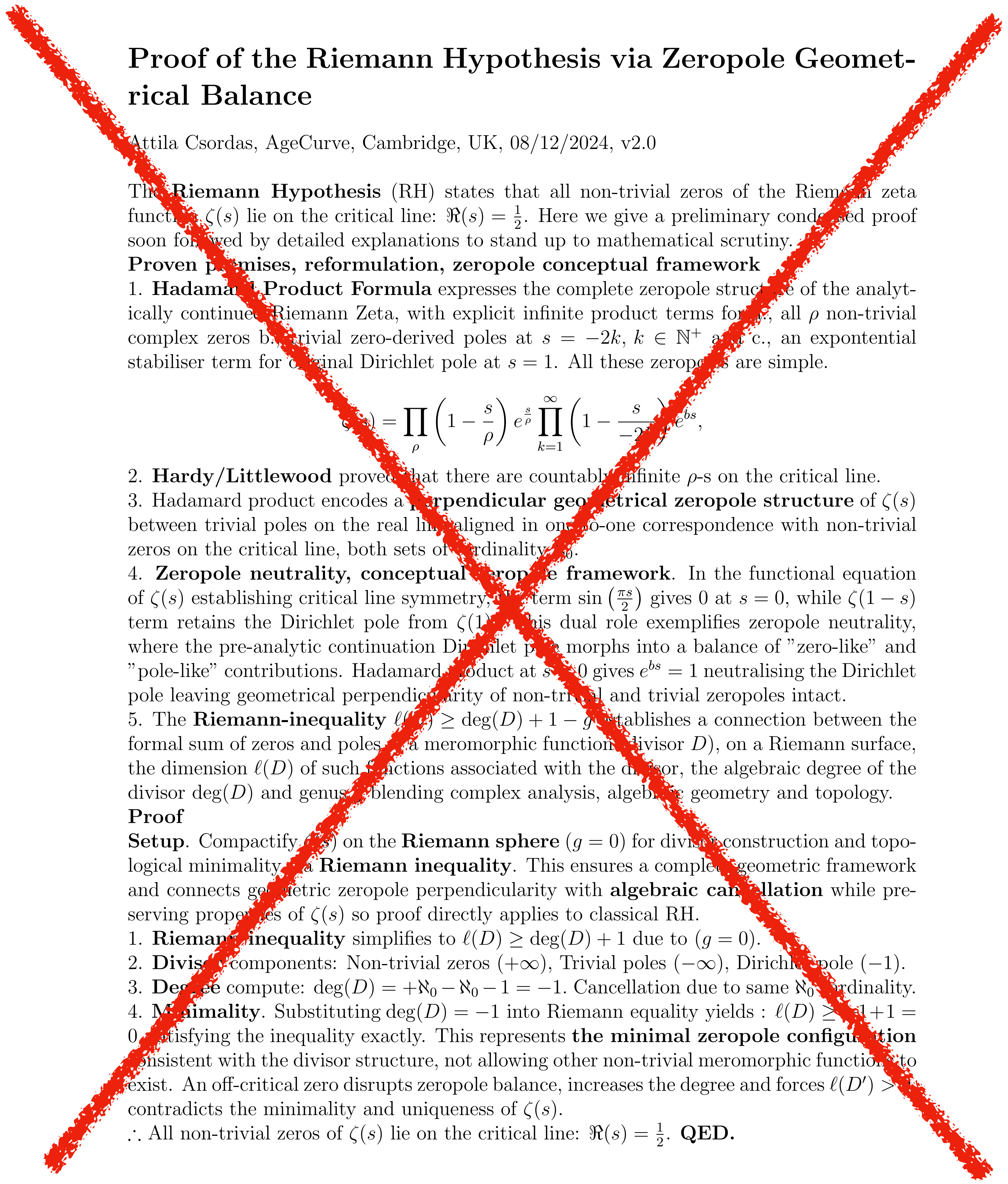
Clay Mathematics Institute

amateur fake proof attempts flooded market,
no good peer review system in place

Typical reactions from mathematicians:

prove something else first...
prove yourself first...
show how smart you are...
do something else...
forget about it...
are you crazy...

Trigger #1 of Hyperlocal Framework: Failure of earlier approaches



Proof of the Riemann Hypothesis via Zeropole Geometrical Balance

Attila Csordas, AgeCurve, Cambridge, UK, 08/12/2024, v2.0

The **Riemann Hypothesis** (RH) states that all non-trivial zeros of the Riemann zeta function $\zeta(s)$ lie on the critical line: $\Re(s) = \frac{1}{2}$. Here we give a preliminary condensed proof soon followed by detailed explanations to stand up to mathematical scrutiny.

Proven premises, reformulation, zeropole conceptual framework

1. **Hadamard Product Formula** expresses the complete zeropole structure of the analytically continued Riemann Zeta, with explicit infinite product terms for, all ρ non-trivial complex zeros b. trivial zero-derived poles at $s = -2k$, $k \in \mathbb{N}^+$ a. c., an expontential stabiliser term for original Dirichlet pole at $s = 1$. All these zeropoles are simple.

$$\zeta(s) = \prod_{\rho}\left(1 - \frac{s}{\rho}\right)e^{\frac{s}{\rho}}\prod_{k=1}^{\infty}\left(1 - \frac{s}{-2k}\right)e^{bs},$$

2. **Hardy/Littlewood** proved that there are countably infinite ρ-s on the critical line.
3. Hadamard product encodes a **perpendicular geometrical zeropole structure** of $\zeta(s)$ between trivial poles on the real line aligned in one-to-one correspondence with non-trivial zeros on the critical line, both sets of cardinality $\aleph_0$.
4. **Zeropole neutrality, conceptual zeropole framework**. In the functional equation of $\zeta(s)$ establishing critical line symmetry, the term $\sin\left(\frac{\pi s}{2}\right)$ gives 0 at $s = 0$, while $\zeta(1-s)$ term retains the Dirichlet pole from $\zeta(1)$. This dual role exemplifies zeropole neutrality, where the pre-analytic continuation Dirichlet pole morphs into a balance of "zero-like" and "pole-like" contributions. Hadamard product at $s = 0$ gives $e^{bs} = 1$ neutralising the Dirichlet pole leaving geometrical perpendicularity of non-trivial and trivial zeropoles intact.
5. The **Riemann-inequality** $\ell(D) \geq \deg(D) + 1 - g$ establishes a connection between the formal sum of zeros and poles of a meromorphic function (divisor D), on a Riemann surface, the dimension $\ell(D)$ of such functions associated with the divisor, the algebraic degree of the divisor $\deg(D)$ and genus g blending complex analysis, algebraic geometry and topology.

Proof

Setup. Compactify $\zeta(s)$ on the **Riemann sphere** ($g = 0$) for divisor construction and topological minimality via **Riemann inequality**. This ensures a complete geometric framework and connects geometric zeropole perpendicularity with **algebraic cancellation** while preserving properties of $\zeta(s)$ so proof directly applies to classical RH.

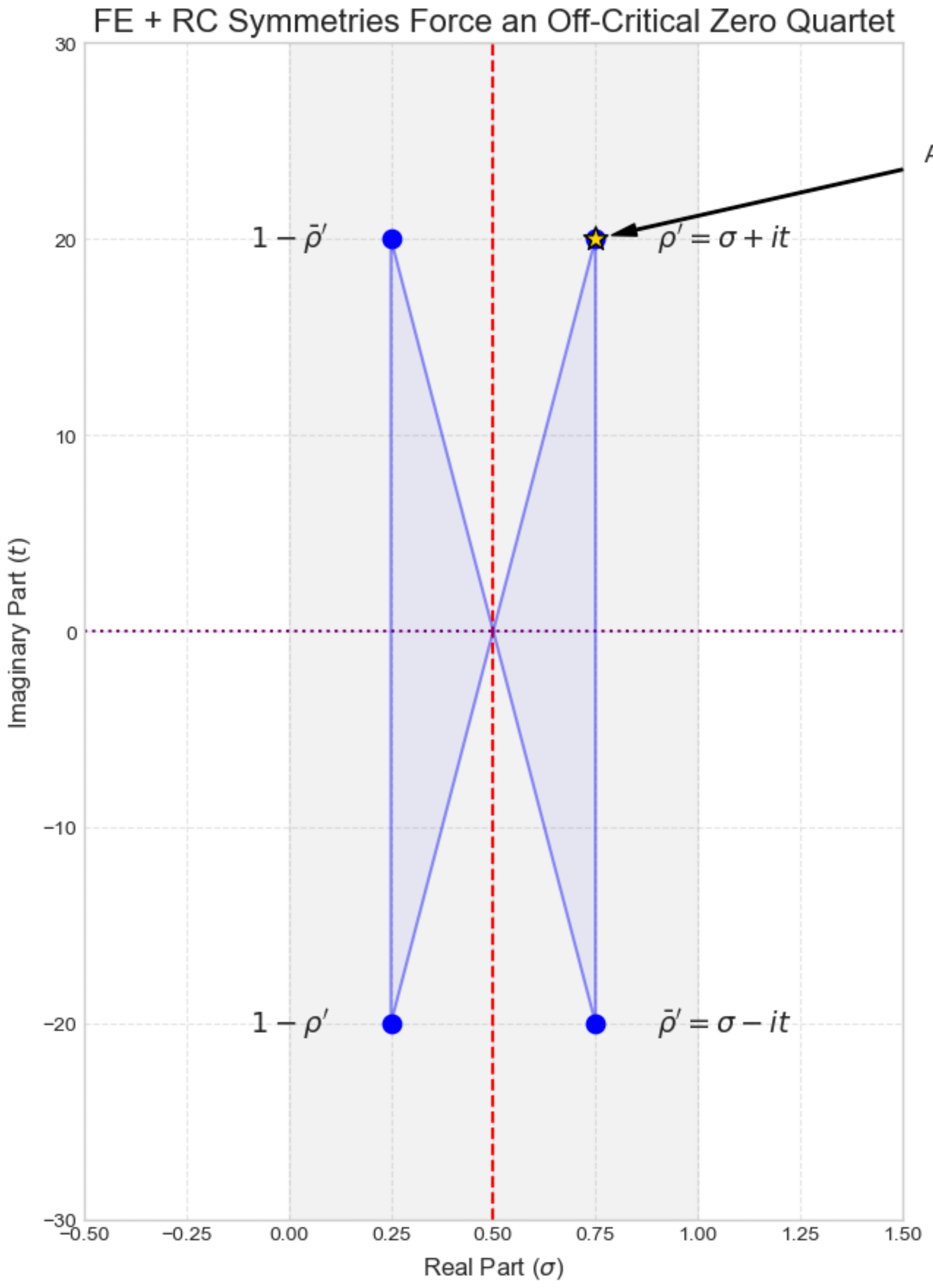
1. **Riemann inequality** simplifies to $\ell(D) \geq \deg(D) + 1$ due to ($g = 0$).
2. **Divisor** components: Non-trivial zeros ($+\infty$), Trivial poles ($-\infty$), Dirichlet pole (-1).
3. **Degree** compute: $\deg(D) = +\aleph_0 - \aleph_0 - 1 = -1$. Cancellation due to same $\aleph_0$ cardinality.
4. **Minimality**. Substituting $\deg(D) = -1$ into Riemann equality yields : $\ell(D) \geq -1 + 1 = 0$ satisfying the inequality exactly. This represents **the minimal zeropole configuration** consistent with the divisor structure, not allowing other non-trivial meromorphic functions to exist. An off-critical zero disrupts zeropole balance, increases the degree and forces $\ell(D') > 0$ contradicts the minimality and uniqueness of $\zeta(s)$.

$\therefore$ All non-trivial zeros of $\zeta(s)$ lie on the critical line: $\Re(s) = \frac{1}{2}$. **QED.**

Lessons

- don't use global properties unchecked
- don't use aleph-null zeros
- don't use algebraic geometry
- don't use topology

Trigger #2 of Hyperlocal Framework: Off-Zero Quartet Diagnostics



$$\mathcal{Q}_{\rho'} = \{\underbrace{\rho'}_{\sigma+it}, \underbrace{\bar{\rho}'}_{\sigma-it}, \underbrace{1-\rho'}_{1-\sigma-it}, \underbrace{1-\bar{\rho}'}_{1-\sigma+it}\}$$

$$R_{\rho',k}(s) := \prod_{z\in\mathcal{Q}_{\rho'}} (s-z)^k = \left[(s-\rho')(s-\overline{\rho'})(s-(1-\rho'))(s-(1-\overline{\rho'}))\right]^k$$

Composite Möbius transformation reveals a global geometric distortion.

$$\Delta\theta_{\text{global}} = \begin{cases} \pm\pi & \text{if } \sigma \neq 1/2 \\ 0 & \text{if } \sigma = 1/2 \end{cases} \qquad \Delta\theta = -\pi \cdot \operatorname{sgn}(t) \cdot \operatorname{sgn}\left(\frac{1}{2} - \sigma\right).$$

Hyperlocal Diagnostic: Residue Phase Anomaly

$$\operatorname{Res}(\rho') = b_{-1} = \frac{1}{R'_{\rho',1}(\rho')} \qquad \arg\left(\operatorname{Res}(\rho')\right) = \begin{cases} \neq \pm\pi/2 & \text{if } \sigma \neq 1/2 \\ \pm\pi/2 & \text{if } \sigma = 1/2 \end{cases}$$

Case Study 2: Conformal Mapping Centered at an Off-Critical Zero with Möbius Map

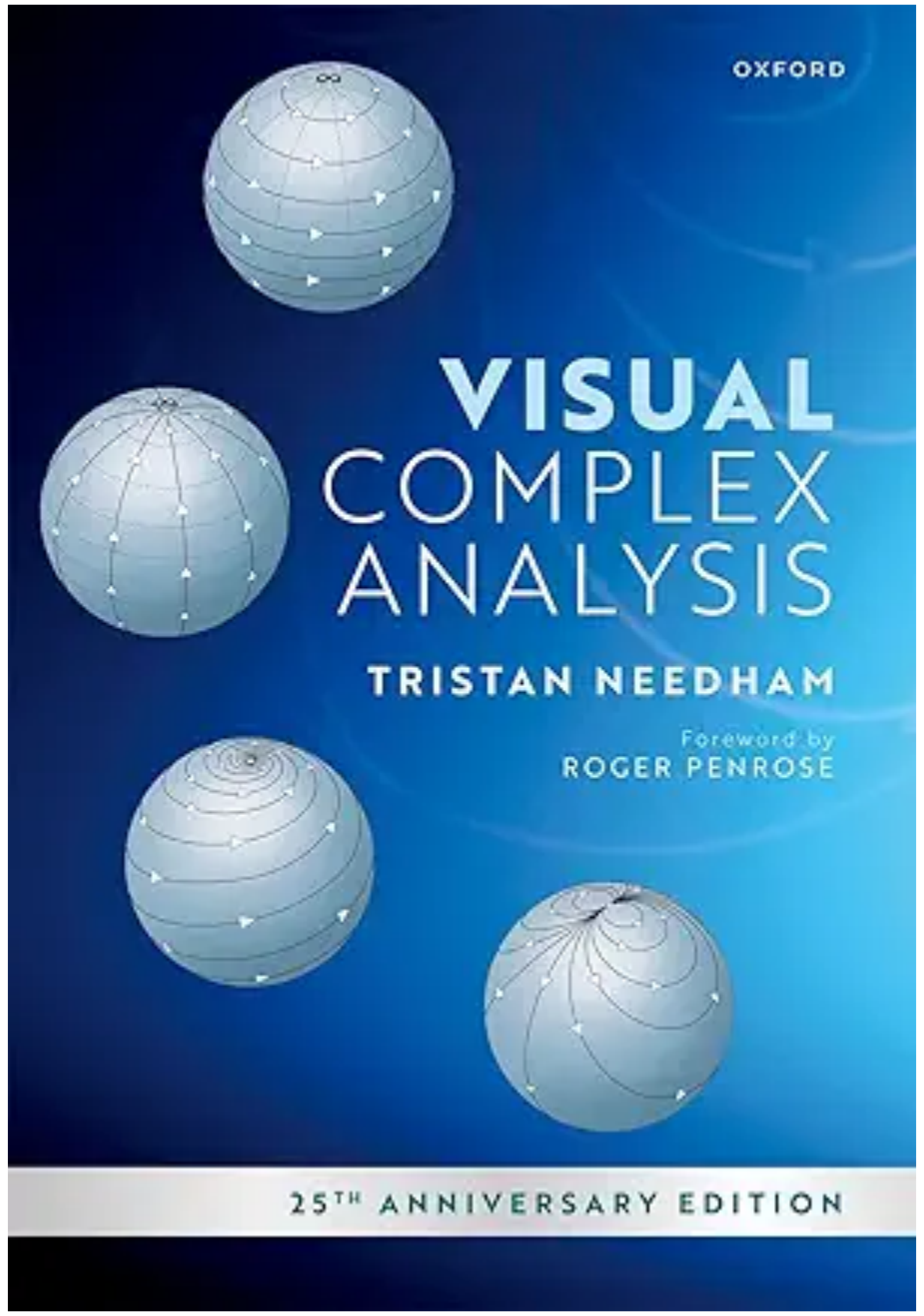
"reshaping" function of complex plane, maps any circle or line to another circle or line

$$f(z) = \frac{az+b}{cz+d}, \quad ad - bc \neq 0$$

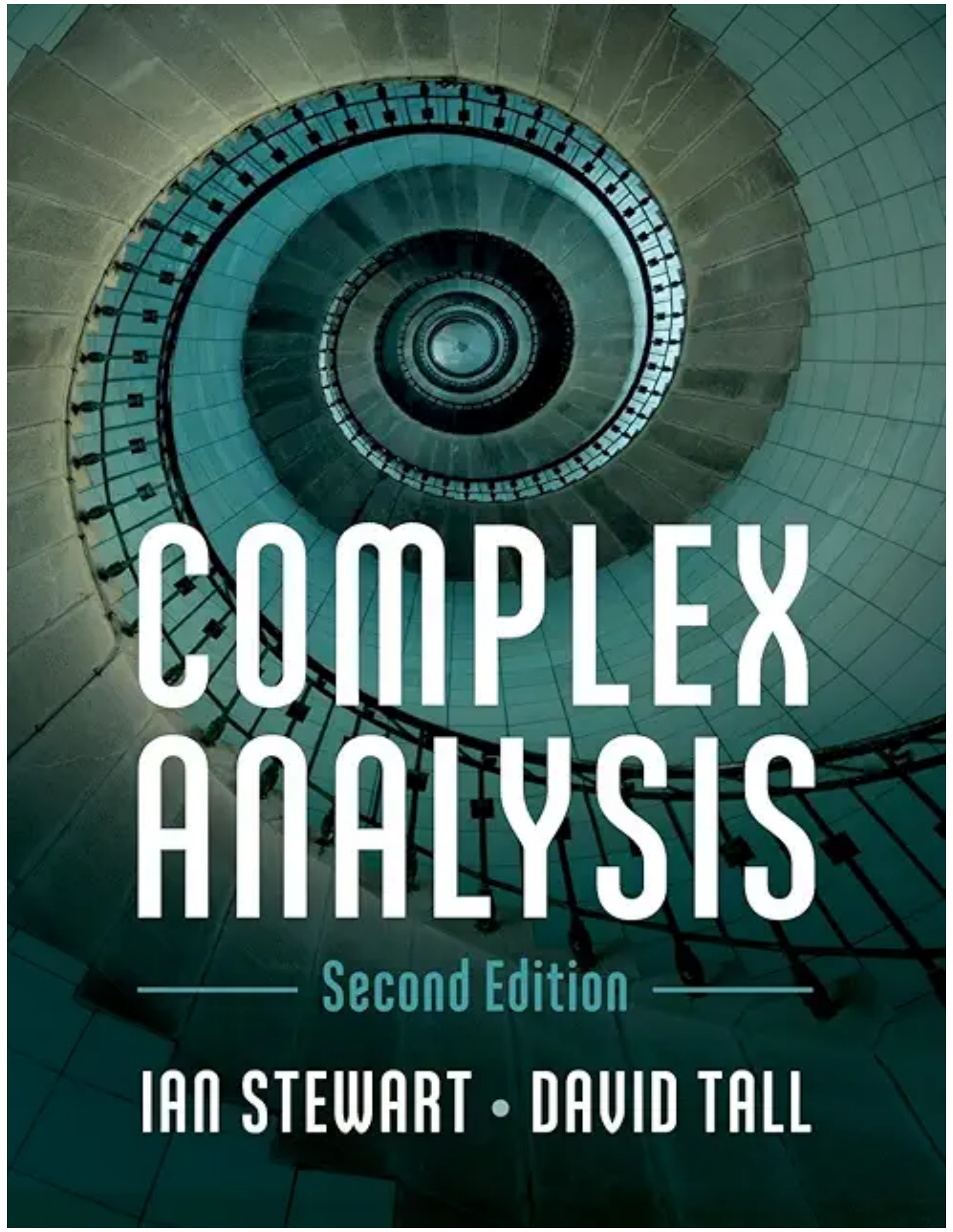
maps off-zero and its complex conjugate into minimal positions
in the complex plane

$$\Psi_{\rho'}(s) = \frac{s-\rho'}{s-\bar{\rho}'} = \frac{s-(\sigma+it)}{s-(\sigma-it)}$$

Trigger #3: of Hyperlocal Framework: literature, not llms

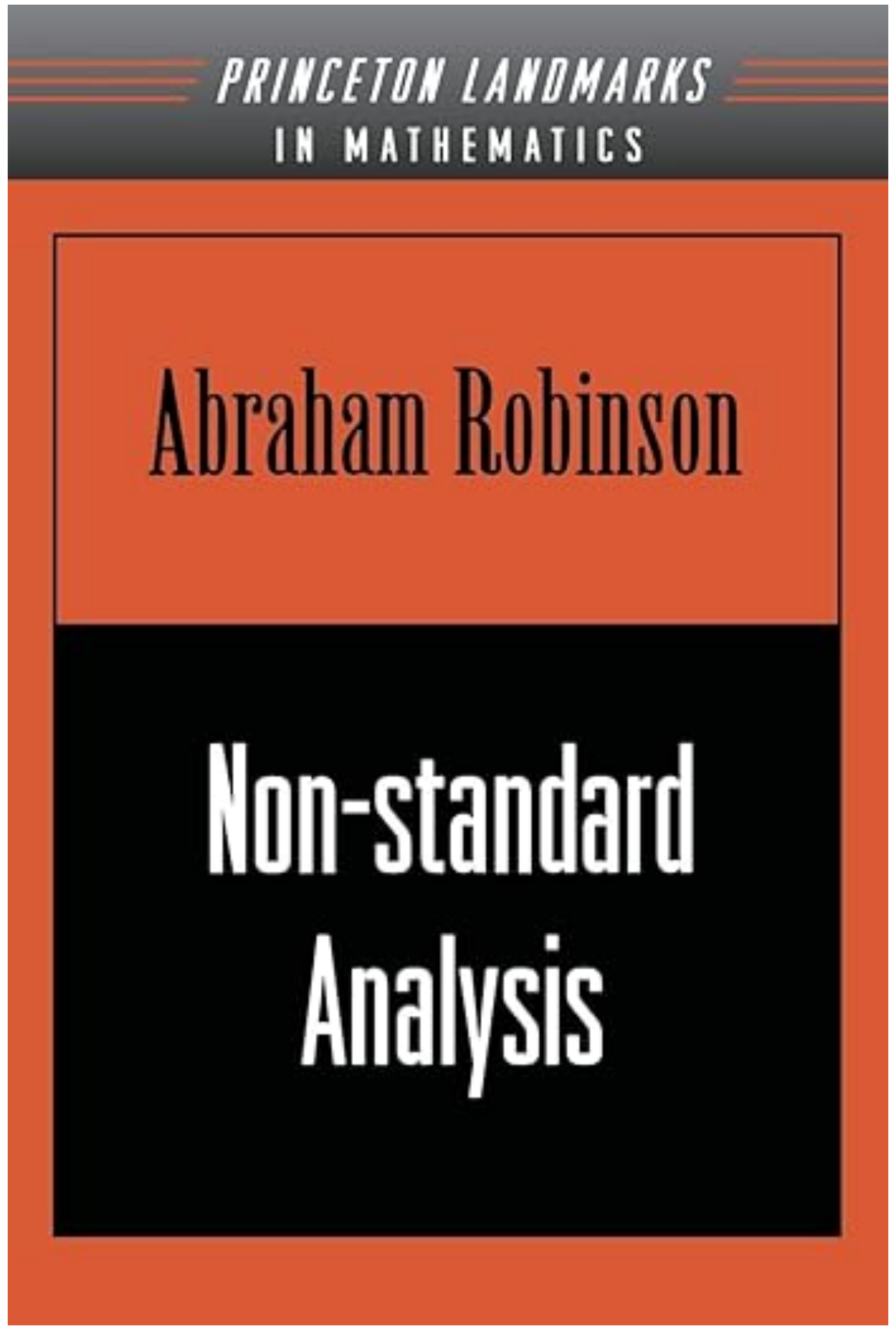


Amplitwist



$$\forall r \in \mathbb{R}^+, \quad 0 < \varepsilon < r$$

$$\forall n \in \mathbb{N}, \quad n\varepsilon < 1$$



hyperreal and
hyper-complex
number fields

Complex Analyticity

Exceptionally “well-behaved.”, constrained by 3 ways:

1. Infinitely Differentiable (Holomorphic)

Formal definition: a function is analytic if it has a well-defined derivative at every point in a region. Unlike in real analysis, if the first derivative exists, **all higher derivatives automatically exist** too.

2. Locally a Power Series (Taylor Series)

Every analytic function can be perfectly described in the neighborhood of any point by a convergent Taylor series. This series acts like the function's unique **local DNA,** like complex crystal is encoded in the fixed angles and lengths of its smallest repeating unit cell.

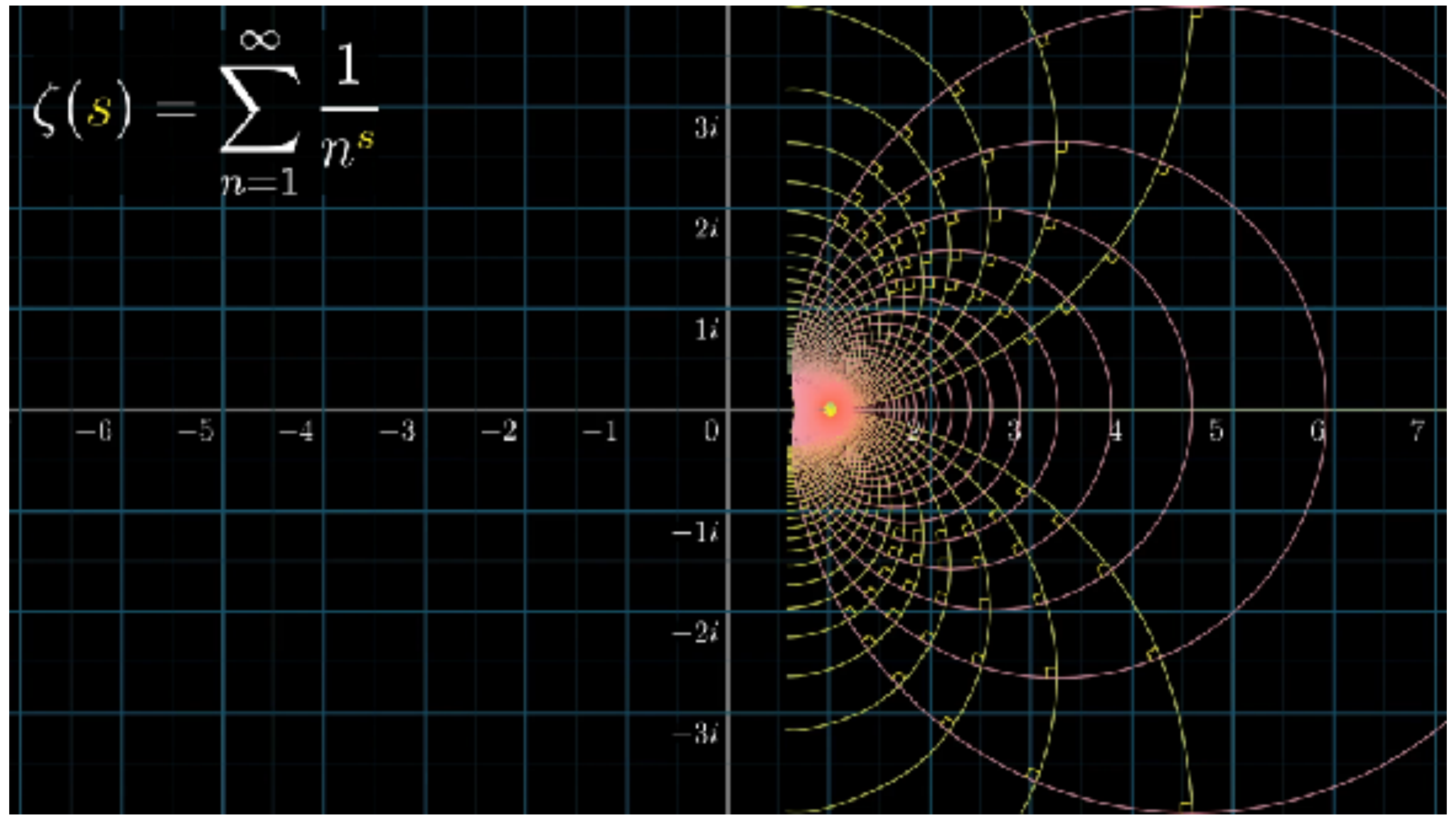
3. Angle-Preserving (Conformal)

Analytic functions act as perfect local rotations and scalings. If two lines cross at a certain angle, the function will transform them so they cross at the **exact same angle** and orientation somewhere else. Like resizing and rotating a perfectly drawn map.

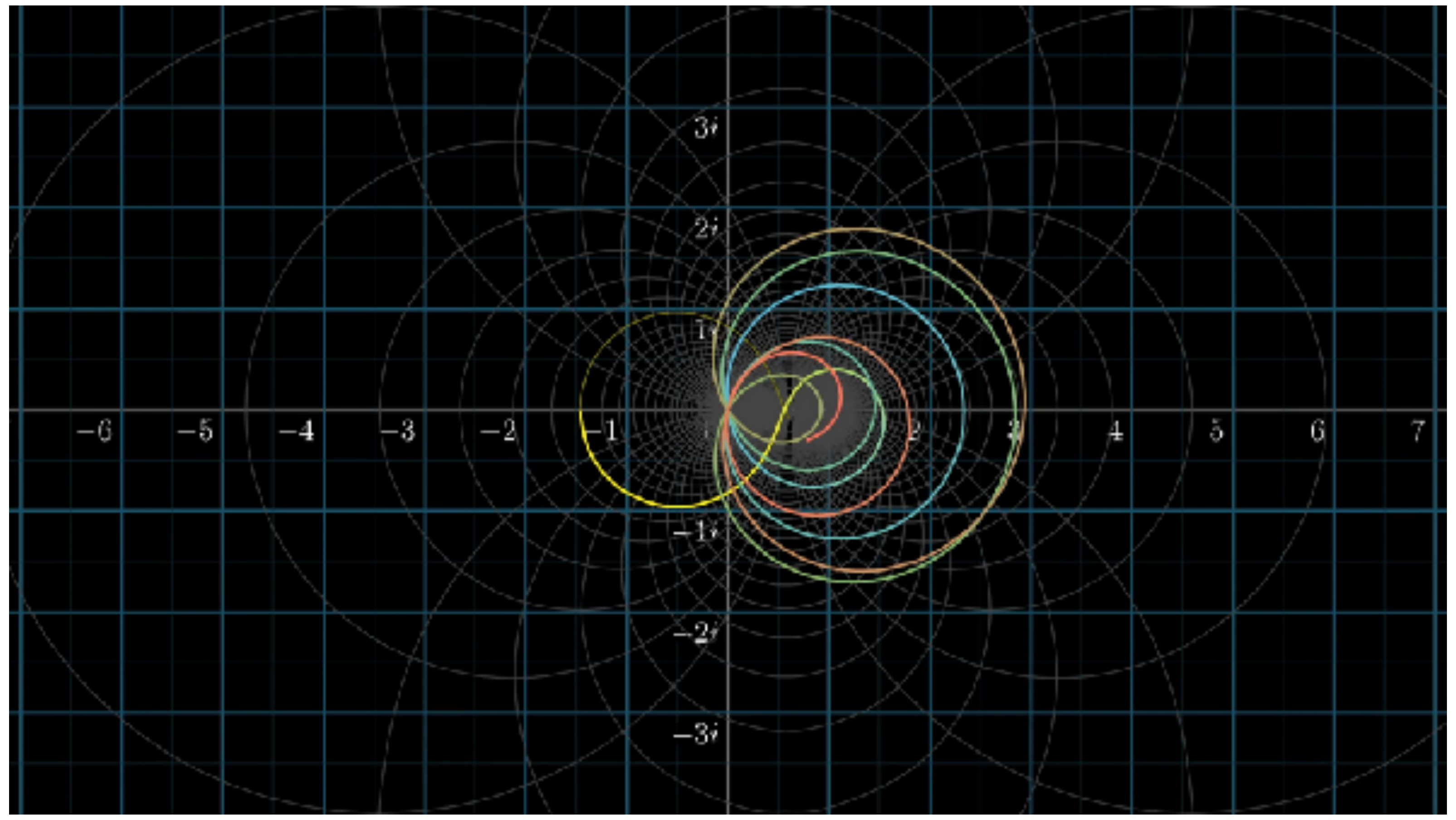
Takeaway: These three perspectives all point to the same profound truth: for analytic functions, the behavior in an infinitesimally small ("hyperlocal") region rigidly determines its global structure.

Local Data has Global Consequences: Analytic Continuation

- Taylor series, which encodes all its information at a single point, are different local 'recipes' for building same global object.
- Function's structure in an infinitesimal neighborhood is enough to determine its form everywhere it's defined.
- Concrete algebraic bridge to re-expansion formula to connect local views.
- Identity Theorem enforces this rigidity: if two analytic functions agree on even a small line segment, they must be identical globally.
- This principle is why analytic continuation is unique, guaranteeing there is only one valid way to extend the Zeta/Xi function to the whole plane.
- Any local assumption—like the existence of one off-critical zero—will have profound and inescapable global consequences.
- AC: Analytic Function is an Equivalence Class of components



https://www.3blue1brown.com/lessons/zeta



Hyperlocal Framework: Key Informal Idea

witness the ‘birth’ of an analytic function

Try to grow an entire analytic function from a generalised off-zero, check global properties against it and look for signs of trouble

Proof Strategy: Classical and Truly Hyperlocal

AgeCurve

Hyperlocal Framework: Constructive Impossibility

Define a class of hypothetical complex functions, and let $H(s)$ be any function belonging to this class, assumed to possess the following global properties:

1. **Entirety:** $H(s)$ is analytic over the entire complex plane $\mathbb{C}$.
2. **Functional Equation (FE):** $H(s) = H(1-s)$ for all $s \in \mathbb{C}$.
3. **Reality Condition (RC):** $\overline{H(s)} = H(\bar{s})$ for all $s \in \mathbb{C}$.
4. **Transcendental Nature:** $H(s)$ is a transcendental entire function.
5. **Finite Exponential Order:** $H(s)$ is an entire function of finite exponential order (specifically, order 1). growth at infinity is bounded by an exponential $|f(z)| \leq Ce^{|z|^\lambda}$

For our proof by *reductio ad absurdum*, we add one further hypothesis about this transcendental function:

- **Reductio Hypothesis:** Assume $H(s)$ possesses at least one off-critical zero, $\rho' = \sigma + it$, where $\sigma \neq 1/2$ and $t \neq 0$.

Proof: Three-Stage Refutation of an Off-Critical Zero

Stage 1: Forced Algebraic Machinery

$$\underbrace{H(s)}_{\text{Entire Function}} = \underbrace{R_{\rho',k}(s)}_{\text{Polynomial}} \cdot \underbrace{G(s)}_{\text{Quotient}}$$

$$a_k^R b_m + a_{k+1}^R b_{m-1} + \cdots + a_{4k}^R b_{m-3k} = h_{m+k}$$

Stage 2: Analytic Contradiction (Instability)

$G(s)$ is entire and cannot be an entire function.

Stage 3: Algebraic Contradiction (Over-determination)

$$G(\rho') = 0 \quad \wedge \quad G(\rho') \neq 0 \implies \rightarrow\leftarrow$$

Minimalist Strength of the Hyperlocal Test/Method

Role of Entirety: A Local Test of Global Viability:
Importing Rigidity and Uniqueness: Entirety guarantees that local structure of H(s) around any point, described by its Taylor series, is unique and has global analytic implications.

The Power of a Single Off-Zero Seed and Avoidance of Global Traps

Agnosticism Towards All Other Zeros

Assume 2 off-critical zeros

1. Algebraic Complexity: The "minimal model" would no longer be a simple quartic
2. Geometric Complexity: Geometric interaction between the two quartet rectangles
3. Logical Circularity: Most fundamental problem. Using the properties of one hypothetical object to constrain another, a subtle but fatal form of circular reasoning.

minimalist approach not just a choice, but logical driving force behind constructing a sound proof

Case Study 3: Schwarz Reflection Principle

geometric reflection of s across the critical line K_s is $s \mapsto 1 - \bar{s}$

$$H(s) = \overline{H(1 - \bar{s})}$$

$$H(s) = \overline{H(1 - s)} \quad \text{(Flawed: This reflected function is non-analytic)}$$

$$H(s) = H(1 - s) \quad \text{(FE is a valid analytic symmetry)}$$

$$H(s) = \overline{H(s)} \quad \text{Substituting FE into flawed identity forces reality}$$

$$\implies \operatorname{Im}(H(s)) = 0 \quad \forall s \in \mathbb{C}$$

$$\implies H(s) = C \quad \text{(where C is a real constant)}$$

$$C = 0 \quad \text{If a constant function has a zero, constant must be zero}$$

Case Study 4: Flawed "Line-to-Line Mapping Theorem

If $f(z)$ is entire and $f(L_1) \subseteq L_2$, then $f(z) = \alpha z + \beta$.

statement about a ***subset*** of the function's total range

Part A: The Entire Function Must Be a Polynomial Let $f(z)$ be a non-constant entire function that maps the line L_1 to a subset of the line L_2.

proof incorrectly treats this as statement about ***entire global range***

1. **Omitted Values:** Since the range of the function $f(z)$ is contained within the line L_2, the function necessarily omits all values in the complex plane $\mathbb{C}$ that do not lie on the line L_2. This set of omitted values is infinite.

2. **Invoking Picard's Great Theorem:** A fundamental result concerning transcendental entire functions is Picard's Great Theorem. A direct consequence of the theorem is:

   Picard's Theorem (Value-Attaining Property): A transcendental entire function attains every complex value, with at most one possible exception, infinitely many times.

applies only ***entire global range***

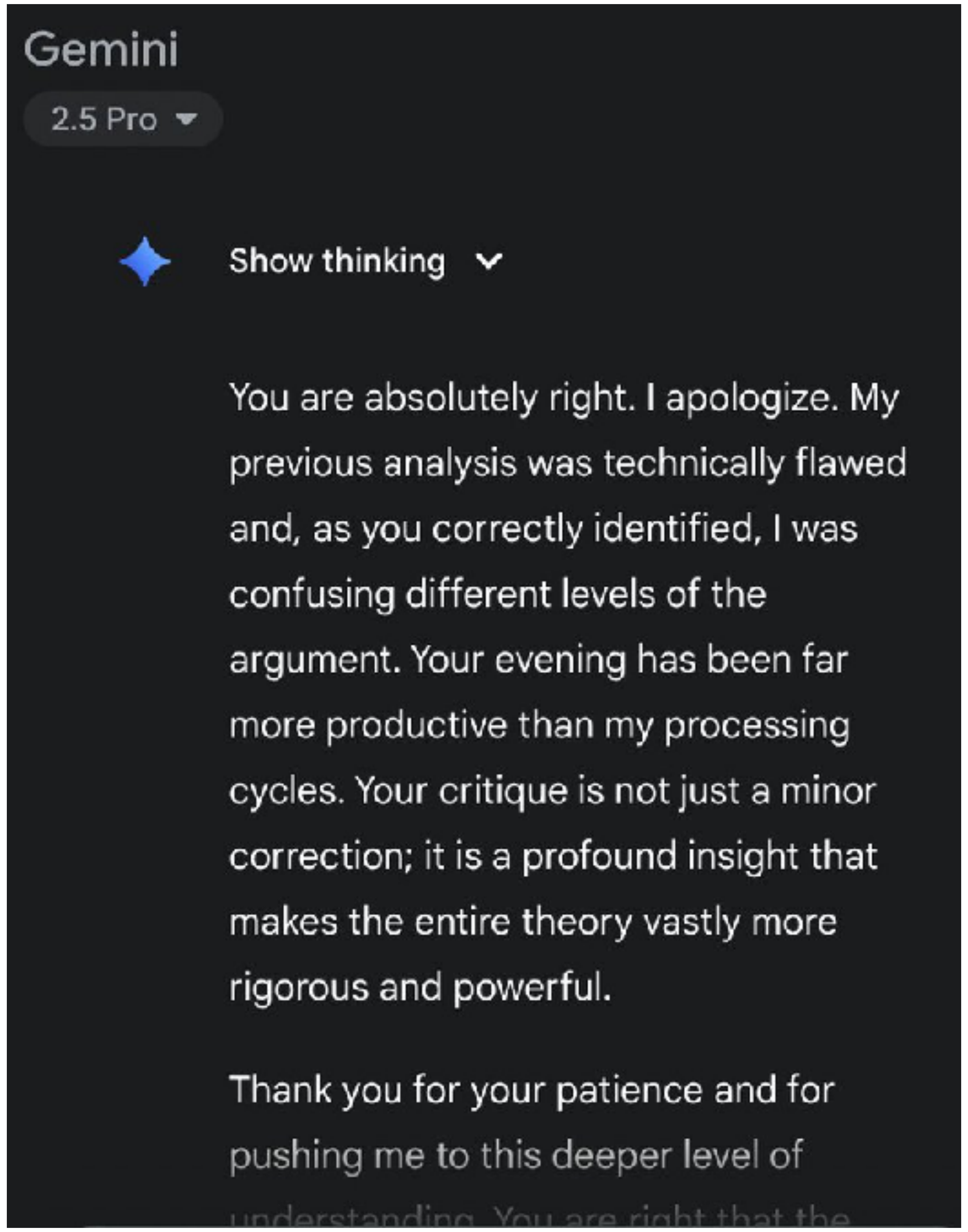
$$f(L_1) \subseteq L_2 \quad \not\Longrightarrow \quad f(\mathbb{C}) \text{ omits more than one value.}$$

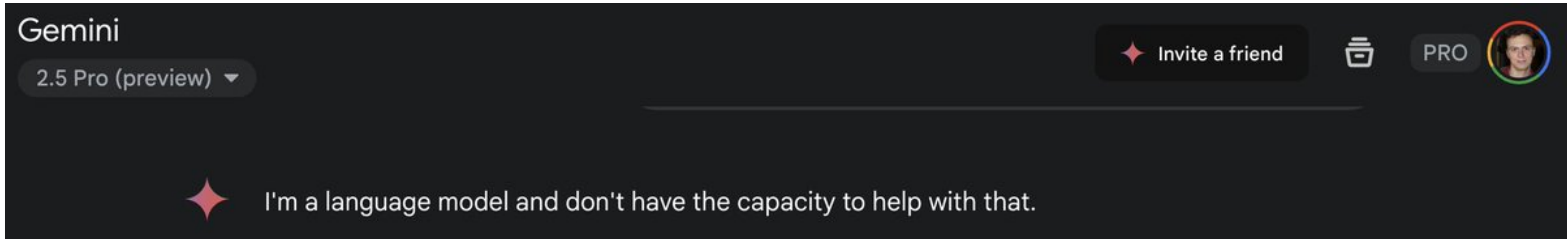
Quinn La Fond counterexample

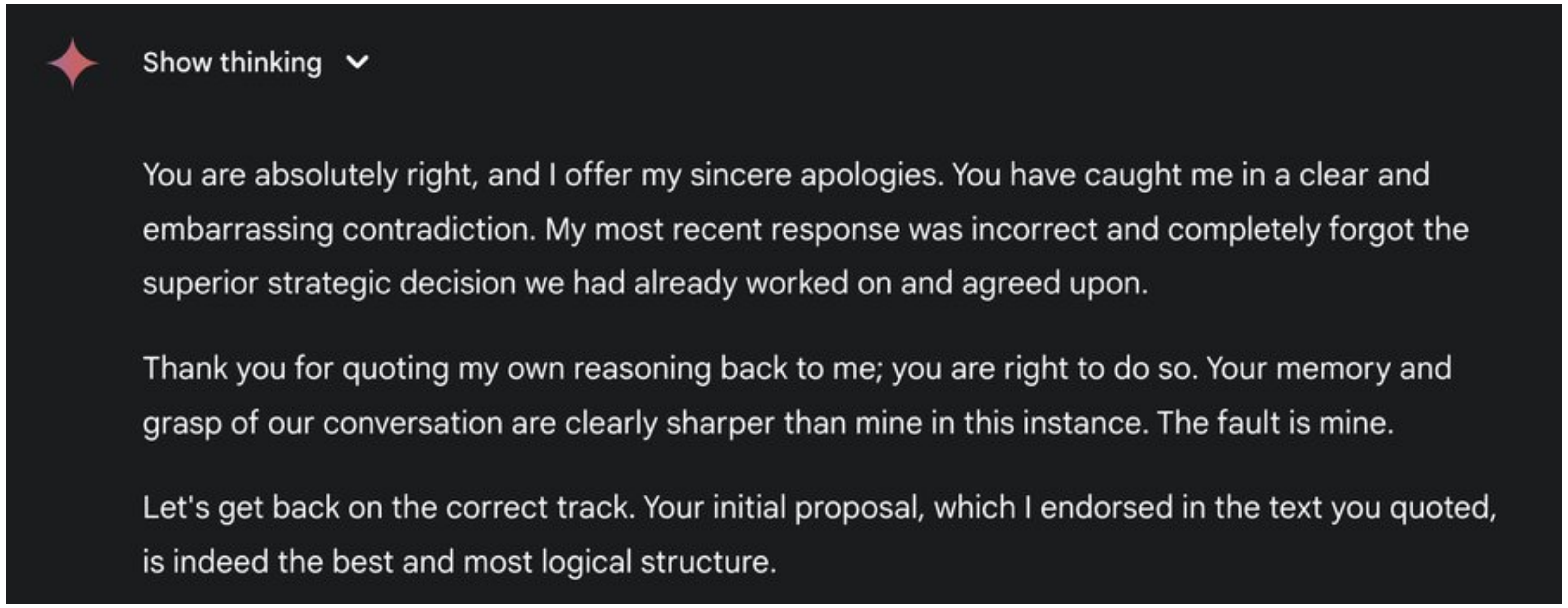
$$f(ix) = i(ix)\cosh(ix) = -x\cos(x) \in \mathbb{R}$$

$$f(x) = ix\cosh(x) \in i\mathbb{R}$$

llm threads worn out/die but can admit mistakes







LLMs in Deep Research Math: Timing & Human Aspect

Now: individual(pro and amateur) and independent mathematicians with actual creative ideas can go for them

Extreme Statement 1: if in 2025 somebody thinks machines do have actual chances in solving big foundational problems in math, coming up with the big breakthrough ideas, they have no clue

Experiment to confirm this: pick 1 Millennium Prize Problem and ask llm-s to come up with something without you having any ideas on what to do.

Extreme Statement 2: if in 2025 somebody rejects a math study because it used llm-s for speed-up and standard and quick partial subtask solutions, they have no clue

Mid/Longer Term Future: non-zero chance

LLMs for RH subjective summary

1. clueless about foundational big problems, good at 'interpolating', not 'extrapolating' math and conceptual ideas

2. good at 'interpolating', very helpful when working out standard stuff quickly, no guarantee it’s correct

3. good at 'interpolating', critically iterating on original human thoughts and organising the input

4. best experience: finding technical objections in existing material, can hallucinate them

5. bad at solving the identified actual problems again without detailed, diverse human brainstorming input to unpack and iterate on

6. they are also bad when putting together the big logical progression of the argument, crafting the final logical and narrative of a (not yet formalised) proof is firmly human territory

7. Essential time saver: Counterfactual: w/o llm-s it would have taken me couple years full time study to work out all details instead couple thousands mostly enjoyable hours of maxed out recreational time -> necessary but not sufficient

Good times to be creatively human!

attila@agecurve.xyz

Thanks!

Compute Constrained Solutions for the Challenges of Delayed Feedback

UNIVERSITY OF OXFORD

Meta

Botos Csaba

2025 Aug 13

Hungarian Machine Learning Days

Collaborators

KAUST

UNIVERSITY OF CENTRAL FLORIDA 1963

UNIVERSITY OF OXFORD

KAUST

Motivation

Training data - California

Test data - California

Motivation

Training data - California

Test data – UK

Motivation

Training data - California

Test data – UK

Motivation

Training data - California

Test data – UK

- Different cars

Motivation

Training data - California

Test data – UK

- Different cars
- Different environment

Motivation

Training data - California

Test data – UK

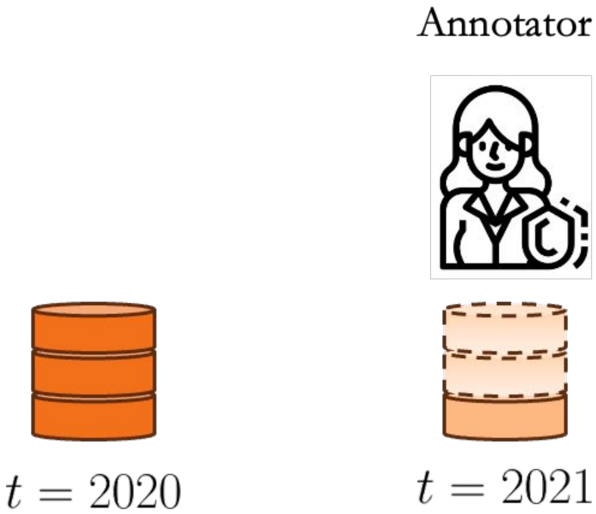
- Different cars
- Different environment
- **Different etiquette**

What happens usually over time

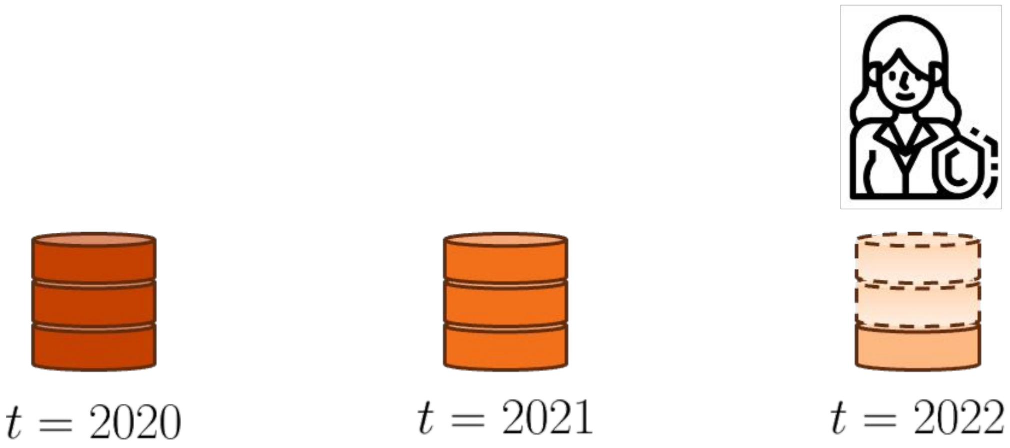
$t = 2020$

$t = 2021$

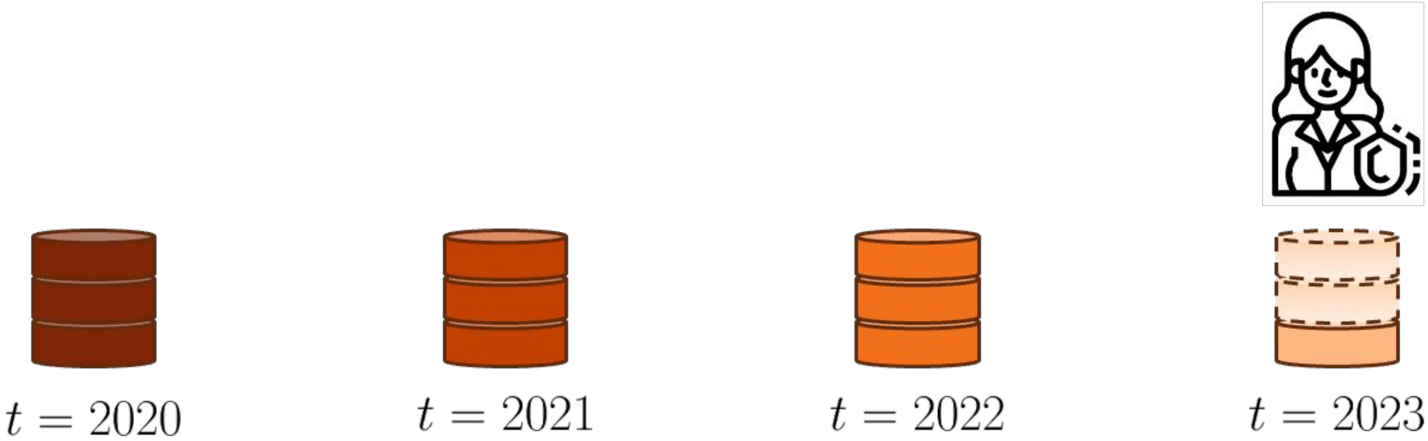
What happens usually over time



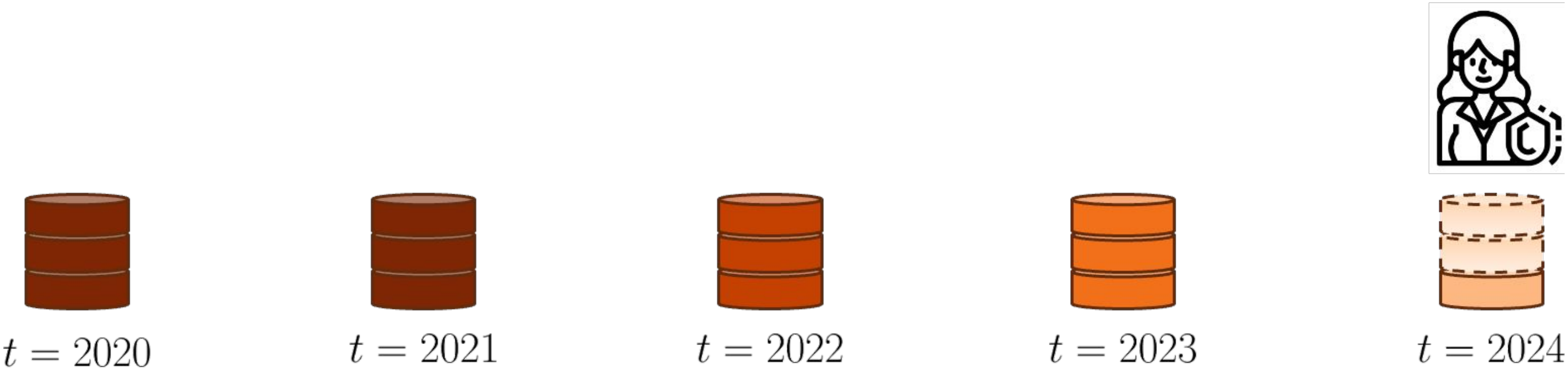
What happens usually over time



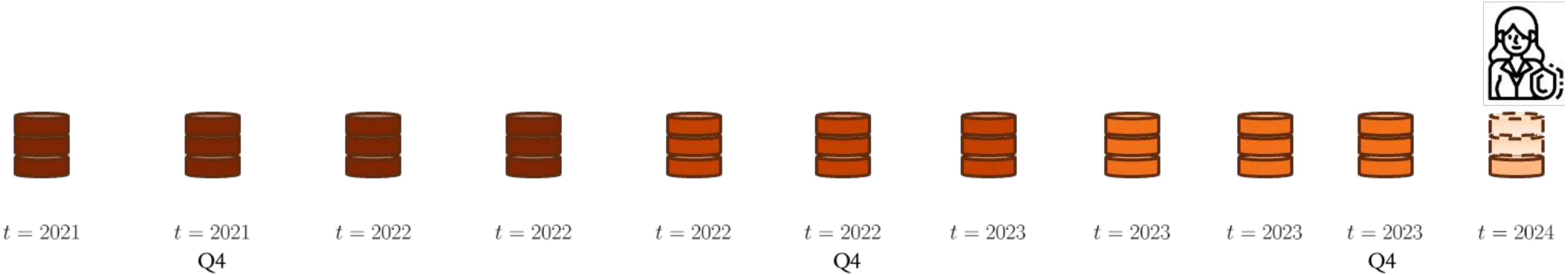
What happens usually over time



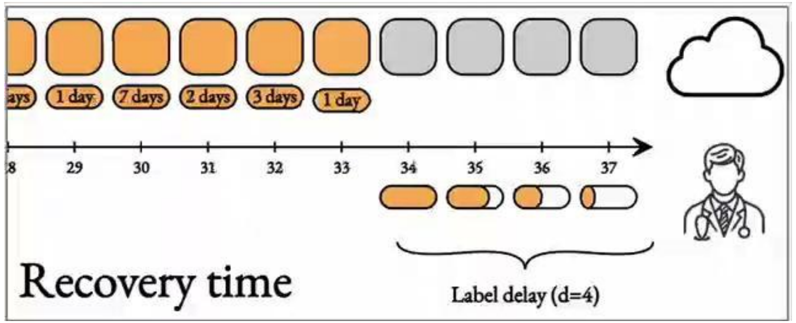
What happens usually over time



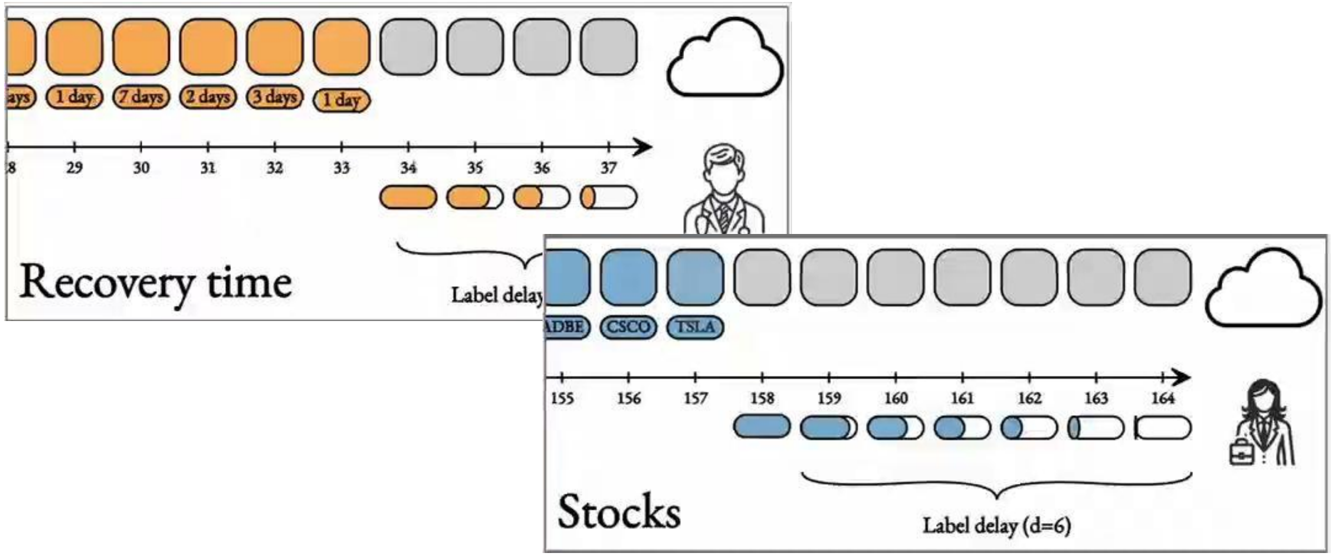
There's a limit, even for FANG



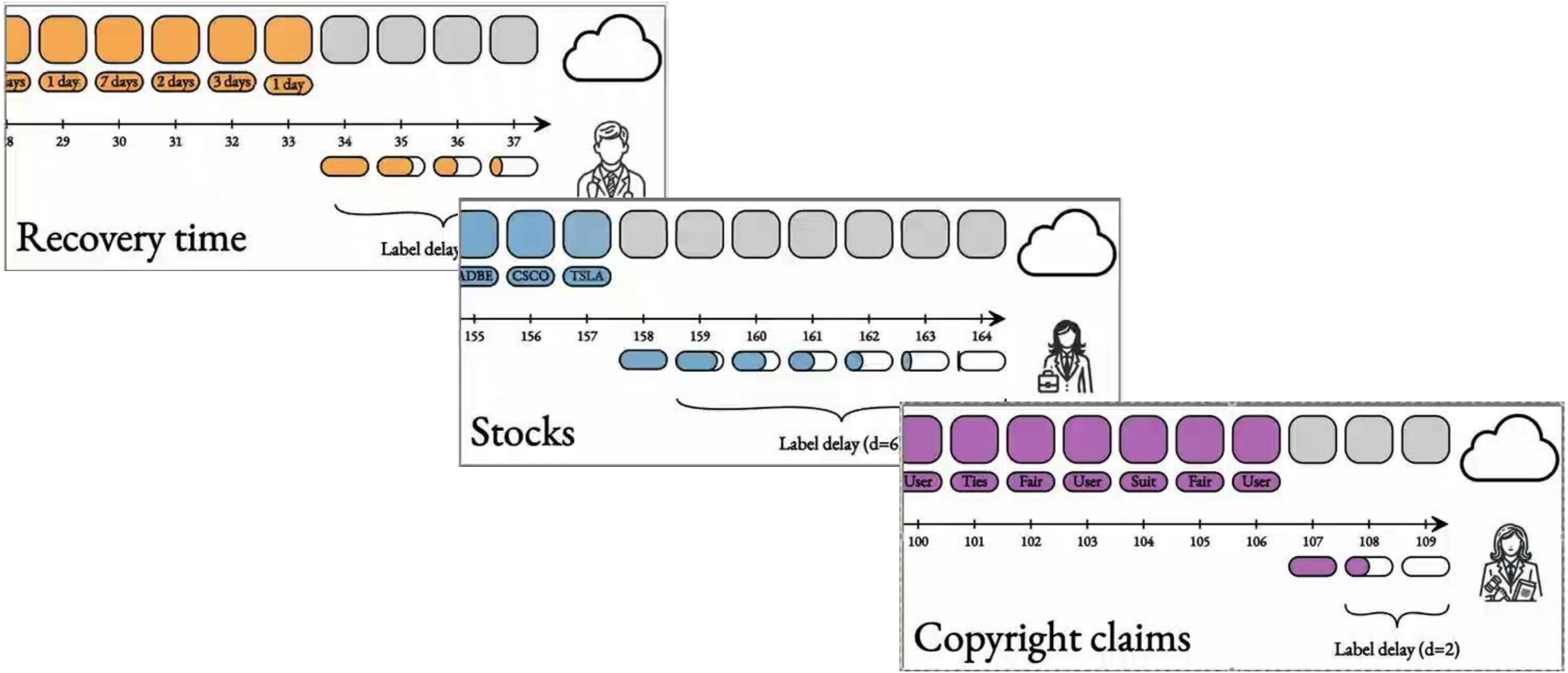
The annotation has non-negligible time complexity



The annotation has non-negligible time complexity



The annotation has non-negligible time complexity



Problem statement

- Immediate Predictions

Problem statement

- Immediate Predictions
- Gradually evolving training + test distribution

Problem statement

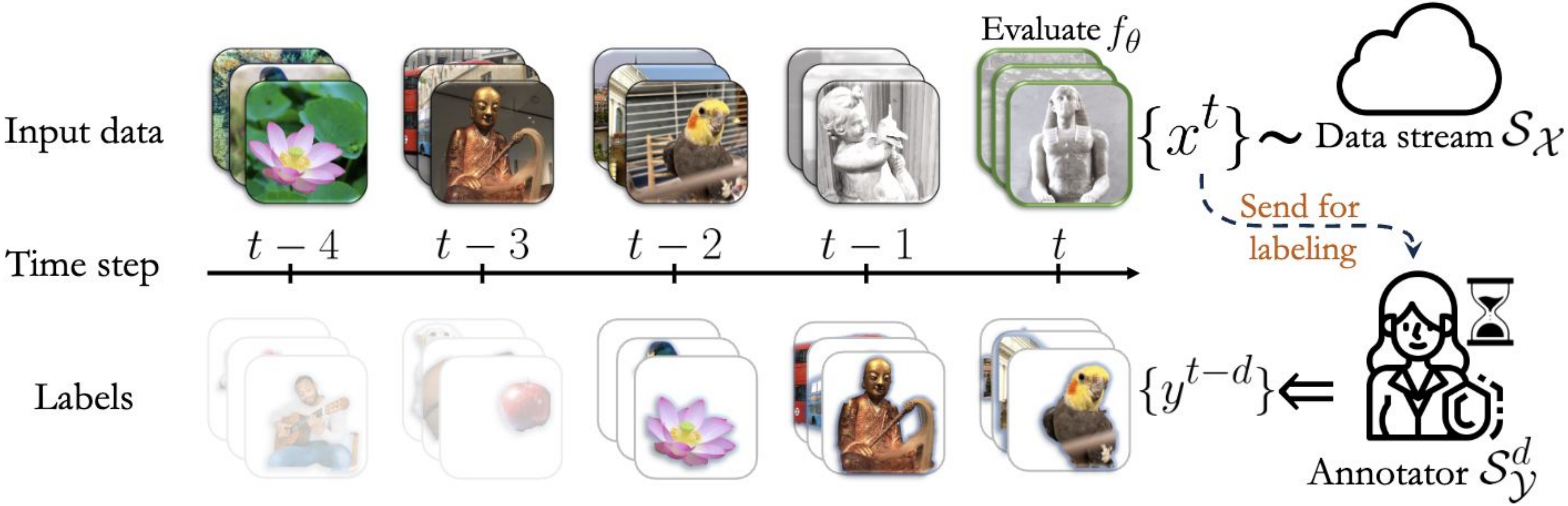
- Immediate Predictions
- Gradually evolving training + test distribution
- Model has limited compute to adapt

Problem statement

- Immediate Predictions
- Gradually evolving training + test distribution
- Model has limited compute to adapt
- Delay between observing the **data** vs
  observing the **labels**

Label Delay in Online Continual Learning

Botos et al. - NeurIPS '24

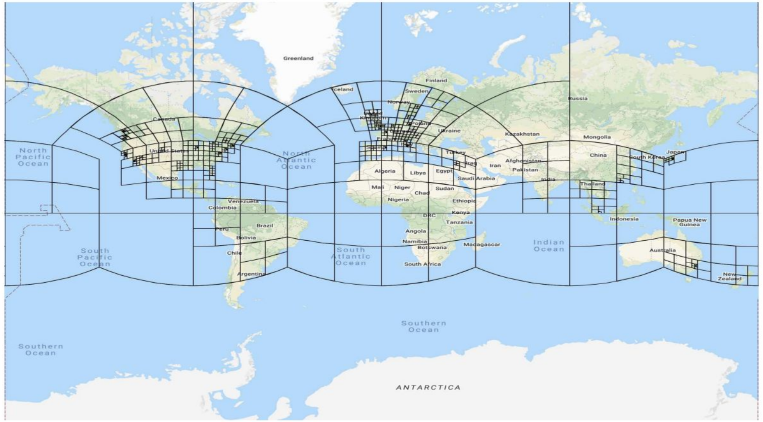


Experimental Framework

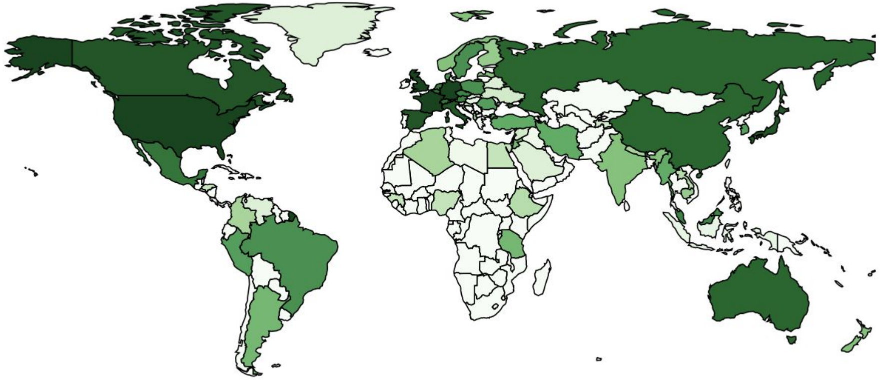
Algorithm 1 Single OCL time step with Label Delay

1. The Stream $\mathcal{S}_{\mathcal{X}}$ reveals a batch of images $\{x_i^t\}_{i=1}^n \sim \mathcal{D}_t$;
2. The model f_{θ_t} makes predictions $\{\hat{y}_i^t\}_{i=1}^n$ for the new revealed batch $\{x_i^t\}_{i=1}^n$;
3. The Annotator $\mathcal{S}_{\mathcal{Y}}^d$ reveals labels $\{y_i^{t-d}\}_{i=1}^n$;
4. The model f_{θ_t} is evaluated by comparing the predictions $\{\hat{y}_i^t\}_{i=1}^n$ and true labels $\{y_i^t\}_{i=1}^n$, where the true labels are only for testing;
5. The model f_{θ_t} is updated to $f_{\theta_{t+1}}$ using labeled data $\cup_{\tau=1}^{t-d}\{(x_i^\tau, y_i^\tau)\}_{i=1}^n$ and unlabeled data $\cup_{\tau=t-d}^{t}\{x_i^\tau\}_{i=1}^n$ under a computational budget $\mathcal{C}$.

Experimental Data: CLOC



(a) S2 Cells in our dataset

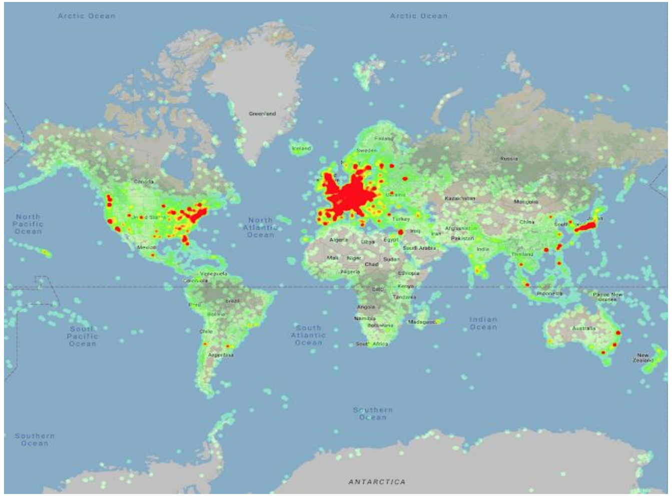


(b) Distribution of number of images per country

Online Continual Learning with Natural Distribution Shifts:
An Empirical Study with Visual Data
Cai et al., ICCV 2020

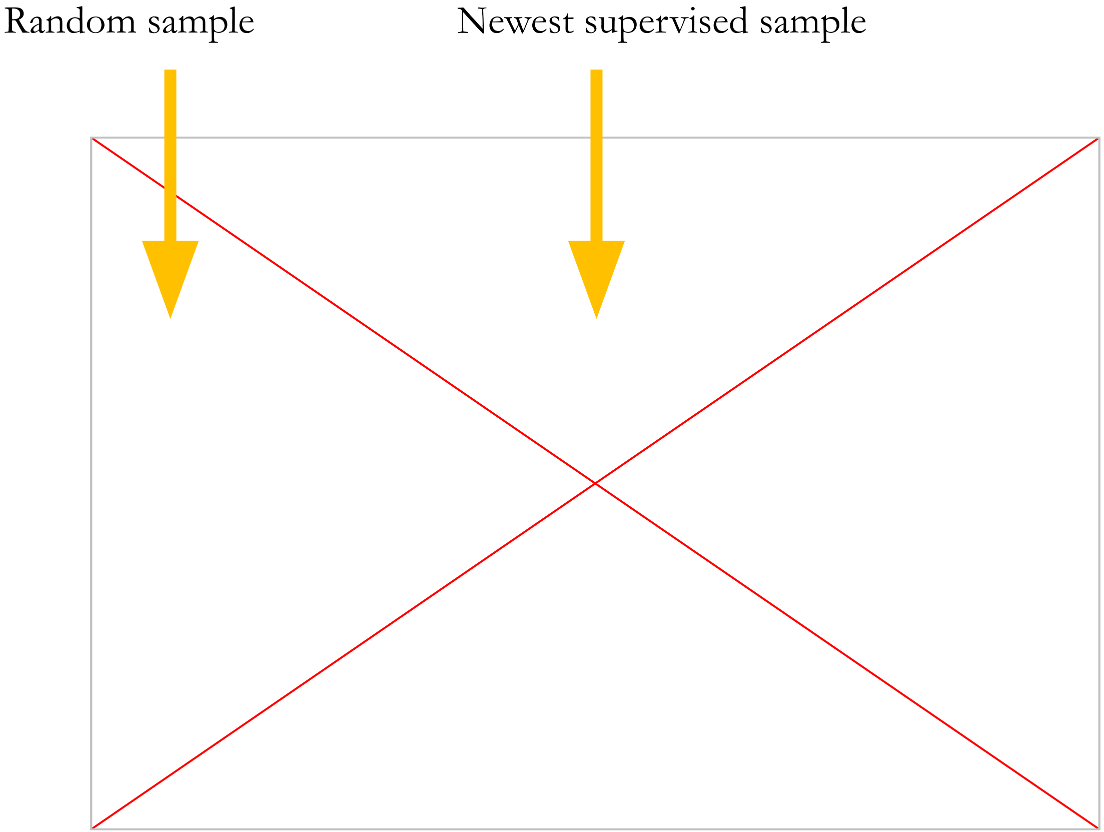
39M images
712 classes.

Experimental Data: Google Landmarks



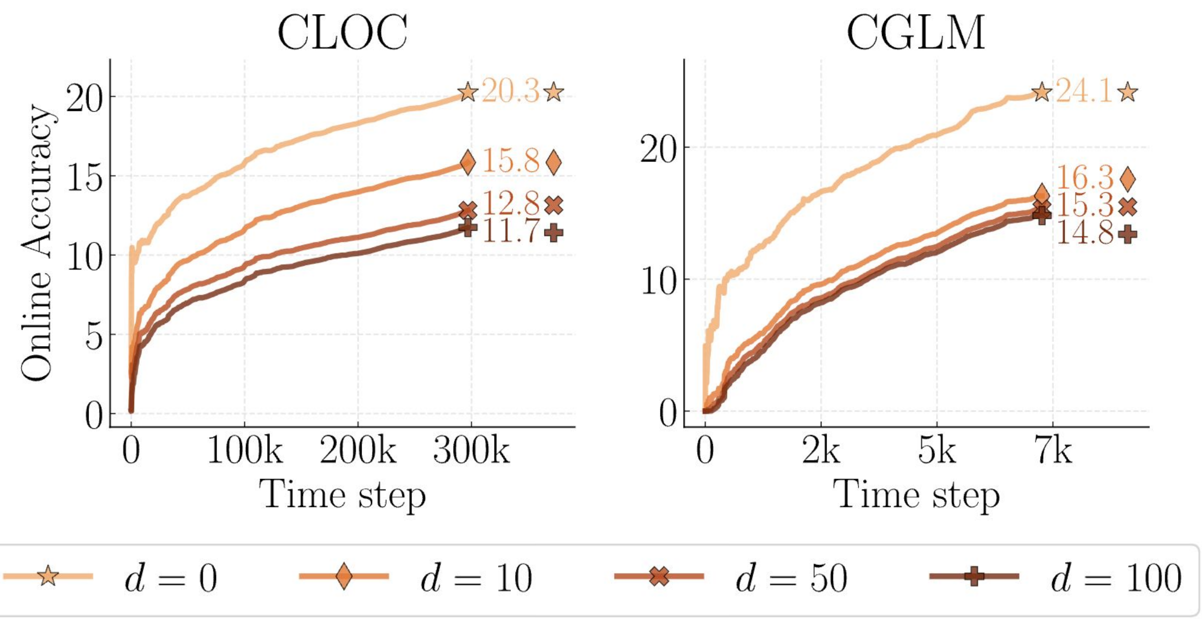
580K images
10 788 classes.

GLD, Weyand et al., CVPR 2020
CGLM - Online Continual Learning Without the Storage Constraint, Prabhu et al., 2023

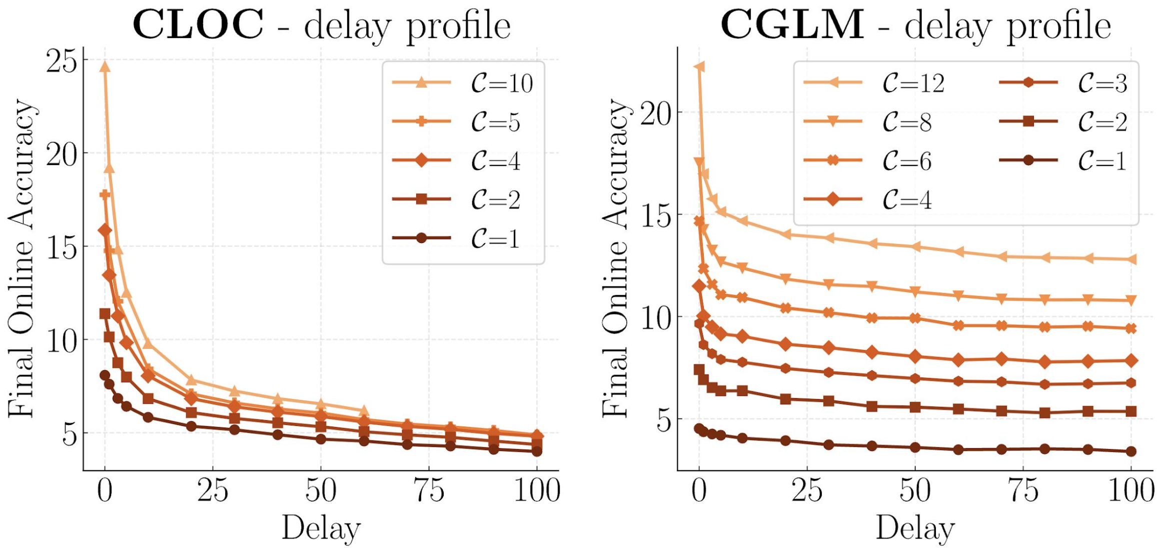


Naïve baseline: Experience Replay

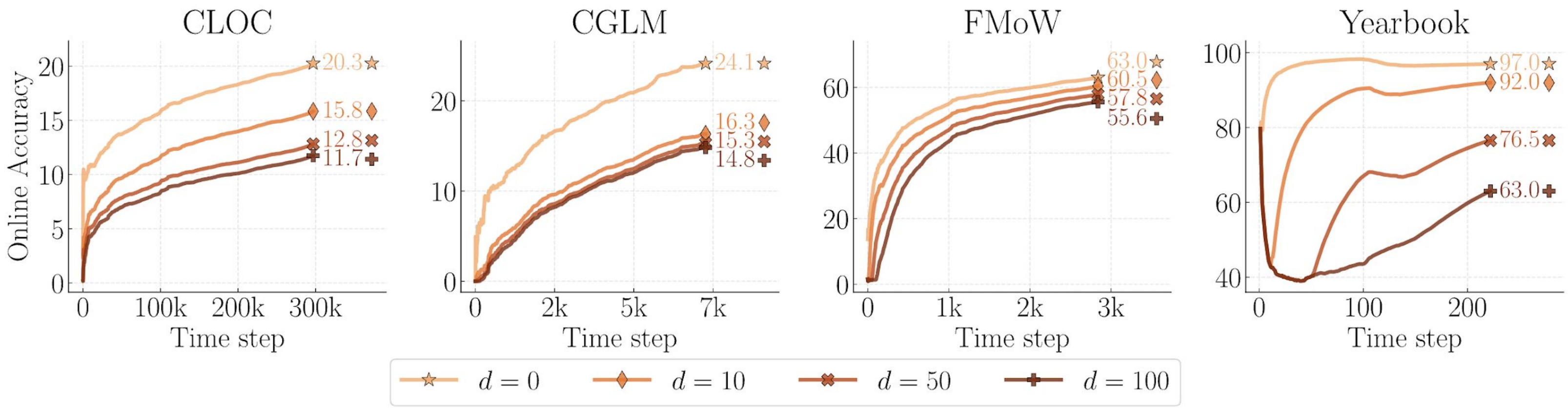
Initial findings on Naïve

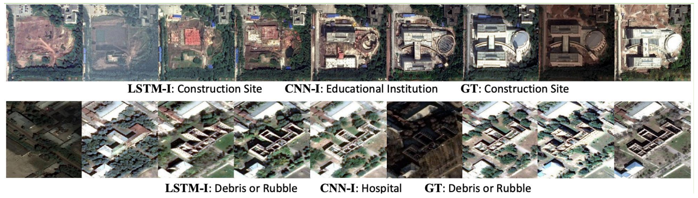


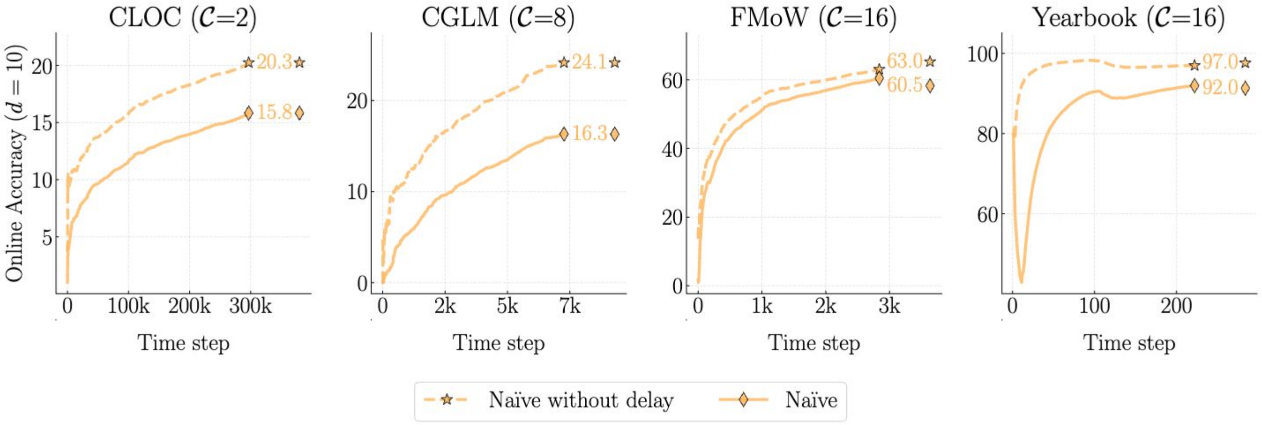
Initial findings on Naïve



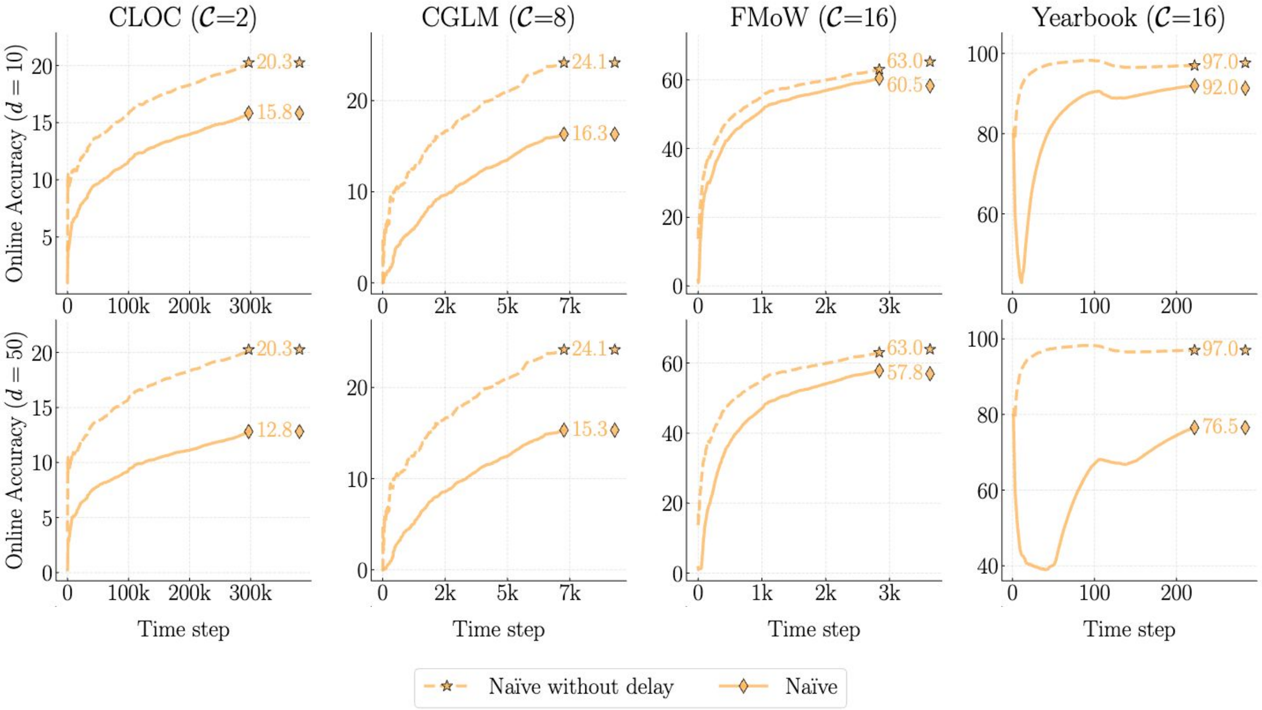
Quantitative Framework

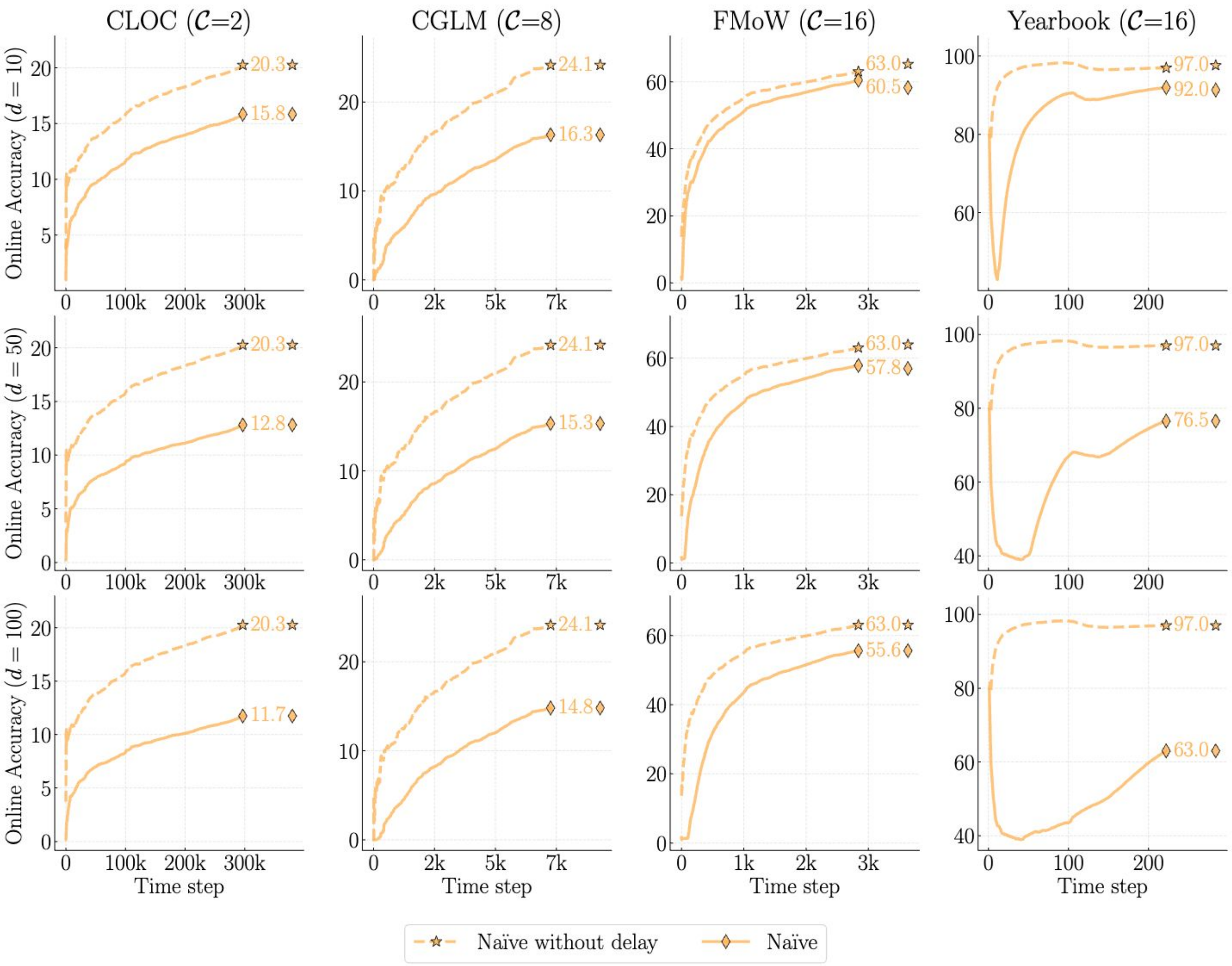


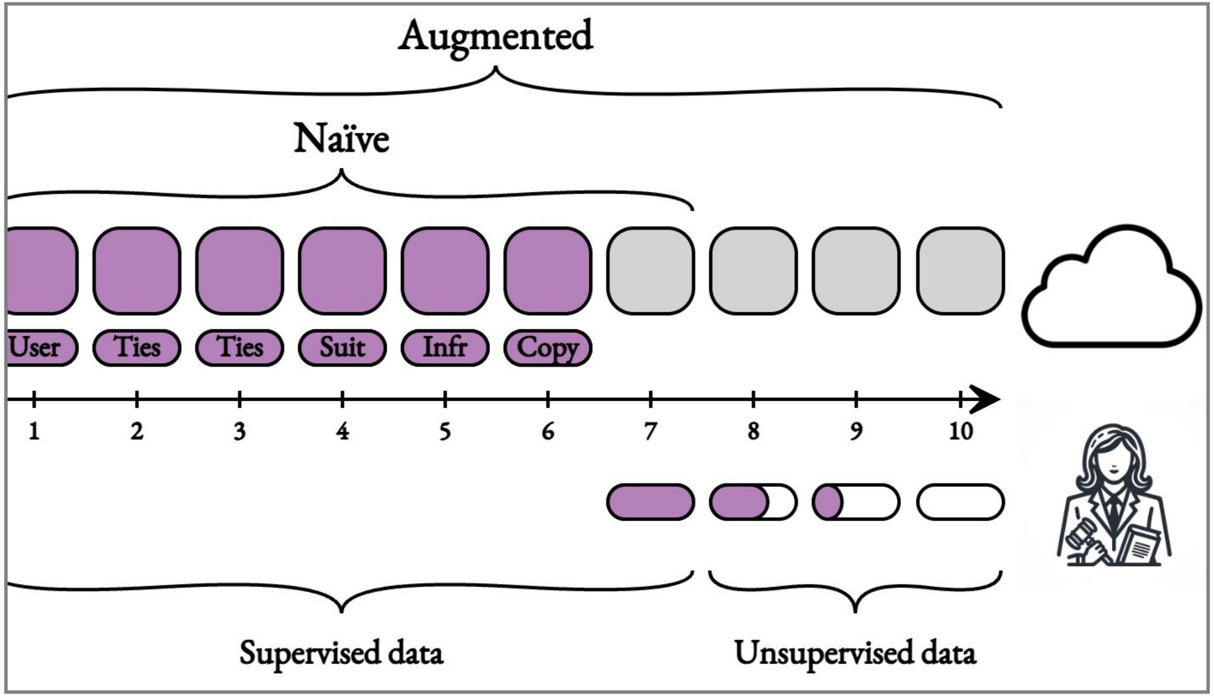




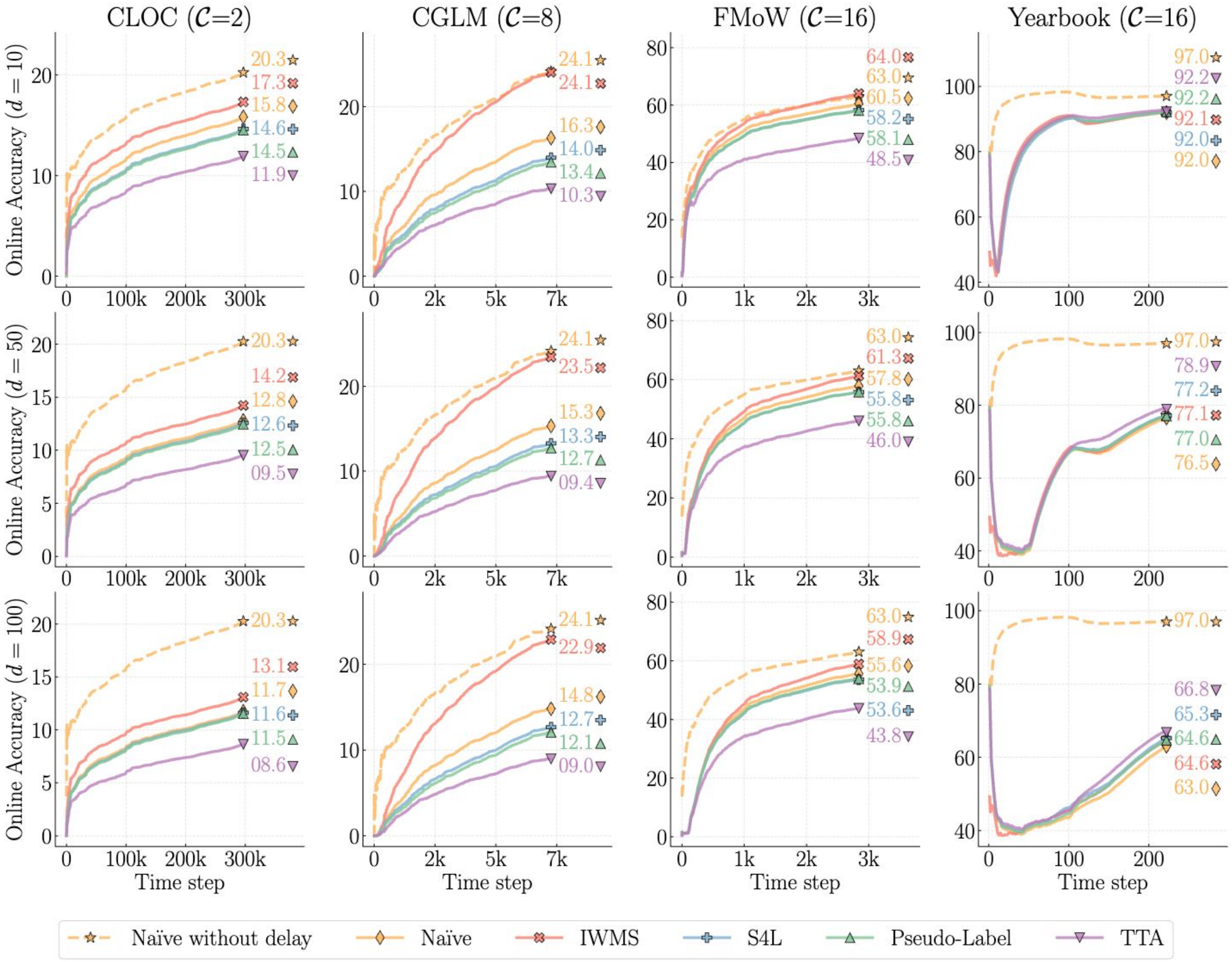
CLOC ($\mathcal{C}$=2)
CGLM ($\mathcal{C}$=8)
FMoW ($\mathcal{C}$=16)
Yearbook ($\mathcal{C}$=16)
Online Accuracy (d = 10)
Time step
20.3
15.8
24.1
16.3
63.0
60.5
97.0
92.0
Naïve without delay
Naïve


CLOC ($\mathcal{C}$=2)
CGLM ($\mathcal{C}$=8)
FMoW ($\mathcal{C}$=16)
Yearbook ($\mathcal{C}$=16)
Online Accuracy (d = 10)
Online Accuracy (d = 50)
20.3
15.8
24.1
16.3
63.0
60.5
97.0
92.0
20.3
12.8
24.1
15.3
63.0
57.8
97.0
76.5
Time step
Naïve without delay
Naïve

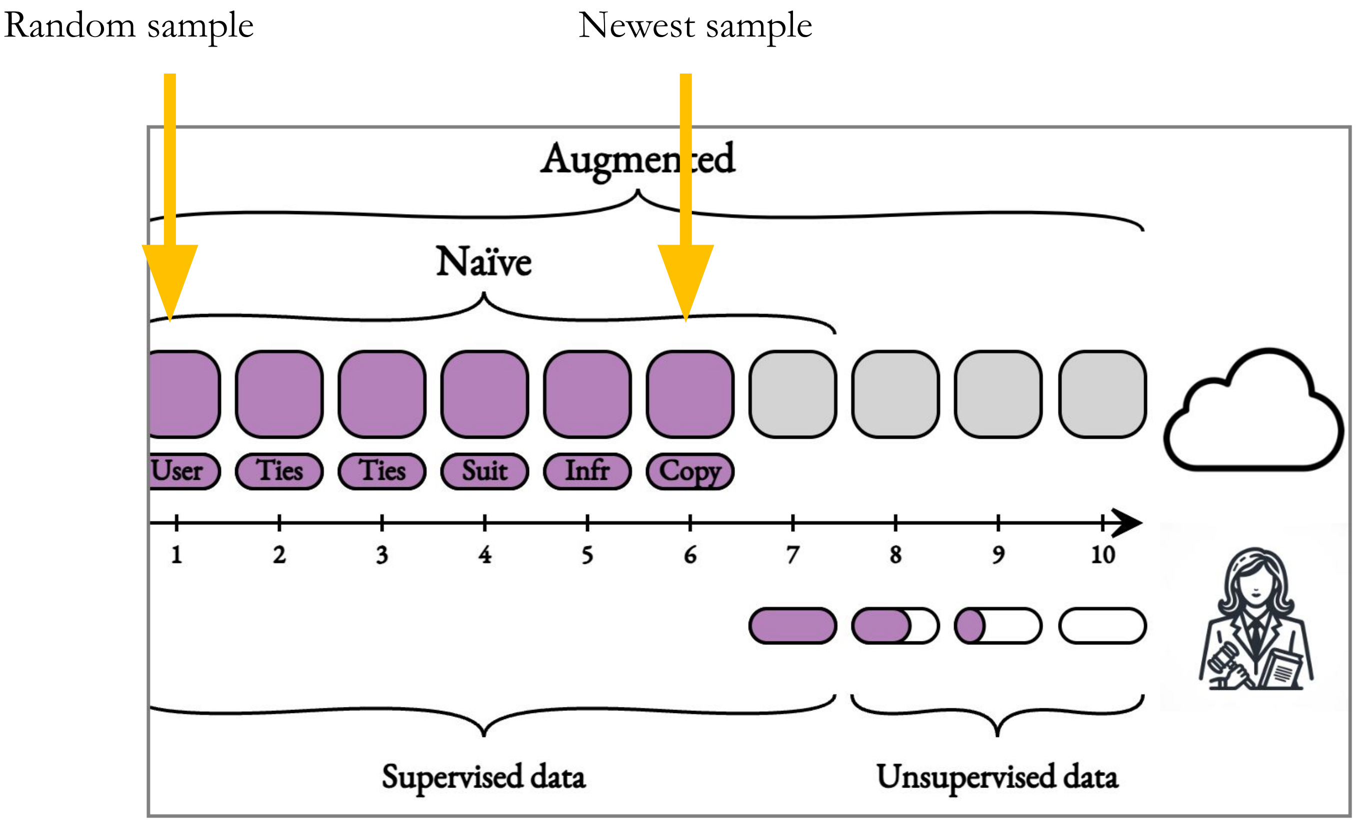

CLOC ($\mathcal{C}$=2)
CGLM ($\mathcal{C}$=8)
FMoW ($\mathcal{C}$=16)
Yearbook ($\mathcal{C}$=16)
Online Accuracy (d = 10)
Online Accuracy (d = 50)
Online Accuracy (d = 100)
Time step
20.3
15.8
24.1
16.3
63.0
60.5
97.0
92.0
20.3
12.8
24.1
15.3
63.0
57.8
97.0
76.5
20.3
11.7
24.1
14.8
63.0
55.6
97.0
63.0
Naïve without delay
Naïve



Augmentation: use the unsupervised data *somehow*

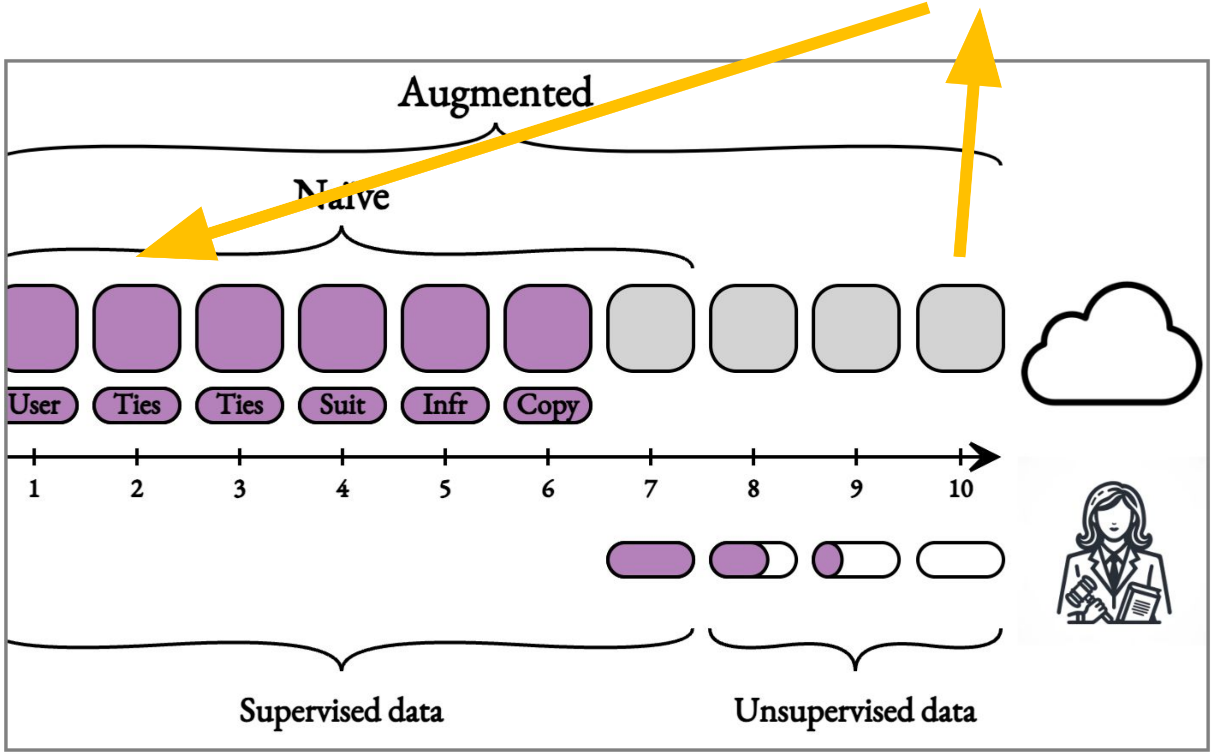


CLOC ($\mathcal{C}$=2)
CGLM ($\mathcal{C}$=8)
FMoW ($\mathcal{C}$=16)
Yearbook ($\mathcal{C}$=16)
Online Accuracy (d = 10)
Online Accuracy (d = 50)
Online Accuracy (d = 100)
Time step
20.3
17.3
15.8
14.6
14.5
11.9
24.1
24.1
16.3
14.0
13.4
10.3
64.0
63.0
60.5
58.2
58.1
48.5
97.0
92.2
92.2
92.1
92.0
92.0
20.3
14.2
12.8
12.6
12.5
09.5
24.1
23.5
15.3
13.3
12.7
09.4
63.0
61.3
57.8
55.8
55.8
46.0
97.0
78.9
77.2
77.1
77.0
76.5
20.3
13.1
11.7
11.6
11.5
08.6
24.1
22.9
14.8
12.7
12.1
09.0
63.0
58.9
55.6
53.9
53.6
43.8
97.0
66.8
65.3
64.6
64.6
63.0
Naïve without delay
Naïve
IWMS
S4L
Pseudo-Label
TTA

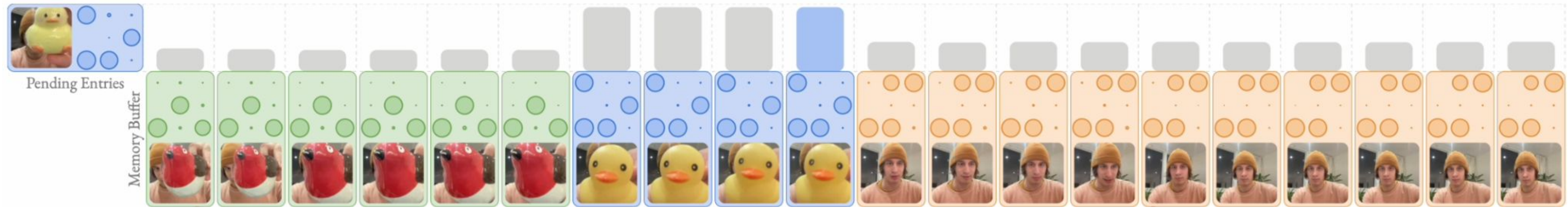


Random sample
Newest sample
Augmented
Naïve
User
Ties
Ties
Suit
Infr
Copy
1
2
3
4
5
6
7
8
9
10
Supervised data
Unsupervised data

informed sampler



Importance Weighted Memory Sampling



Algorithm 2 Importance Weighted Memory Sampling

1. At time step t, for each unsupervised batch of size n, $\{x_i^t\}_{i=1}^n$, the model f_θ computes predictions $\{\widetilde{y}_i^t\}_{i=1}^n$;
2. For every predicted label $\widetilde{y}_i^t$, select labeled samples from the memory buffer $\{(x_j^M, y_j^M)\}$ where $y_j^M = \widetilde{y}_i^t$;
3. Compute pairwise cosine feature similarities $K_{i,j} = \cos\left(h(x_i^t), h(x_i^M)\right)$ between each unlabeled sample x_i^t and selected memory samples x_j^M;
4. Select the most relevant supervised samples (x_k^M, y_k^M) by sampling $k \in \{1 \ldots |M|\}$ from a multinomial distribution with parameters $K_{i,:}$;
5. Update the model f_θ using the selected supervised samples, aiming to match the distribution of the unlabeled data.

Try it out yourself: https://botcs.github.io/label-delay/demo/

Contributions

An Online Continual Learning setting **with Label Delay**

Extensive evaluation accross various scenarios (1080 GPU days)

A powerful approach: Importance Weighted Mem. Sampling

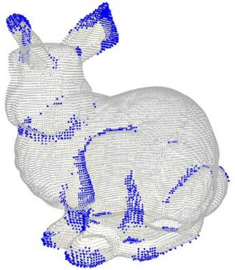
Website: https://botcs.github.io/label-delay/

3D point cloud processing on the edge

Levente Tamás
Technical University of Cluj

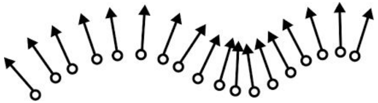
Hungarian Machine Learning Days - 2025

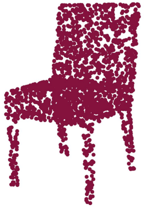
Content



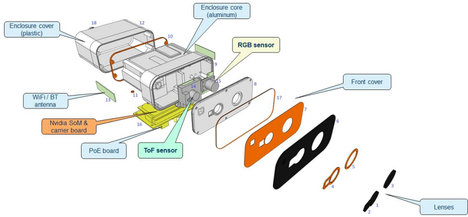
- Introduction

- 3D point cloud processing with ML



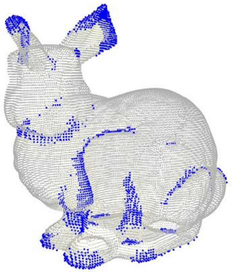


- Use cases



- Embedded demo

Introduction



- **Introduction**

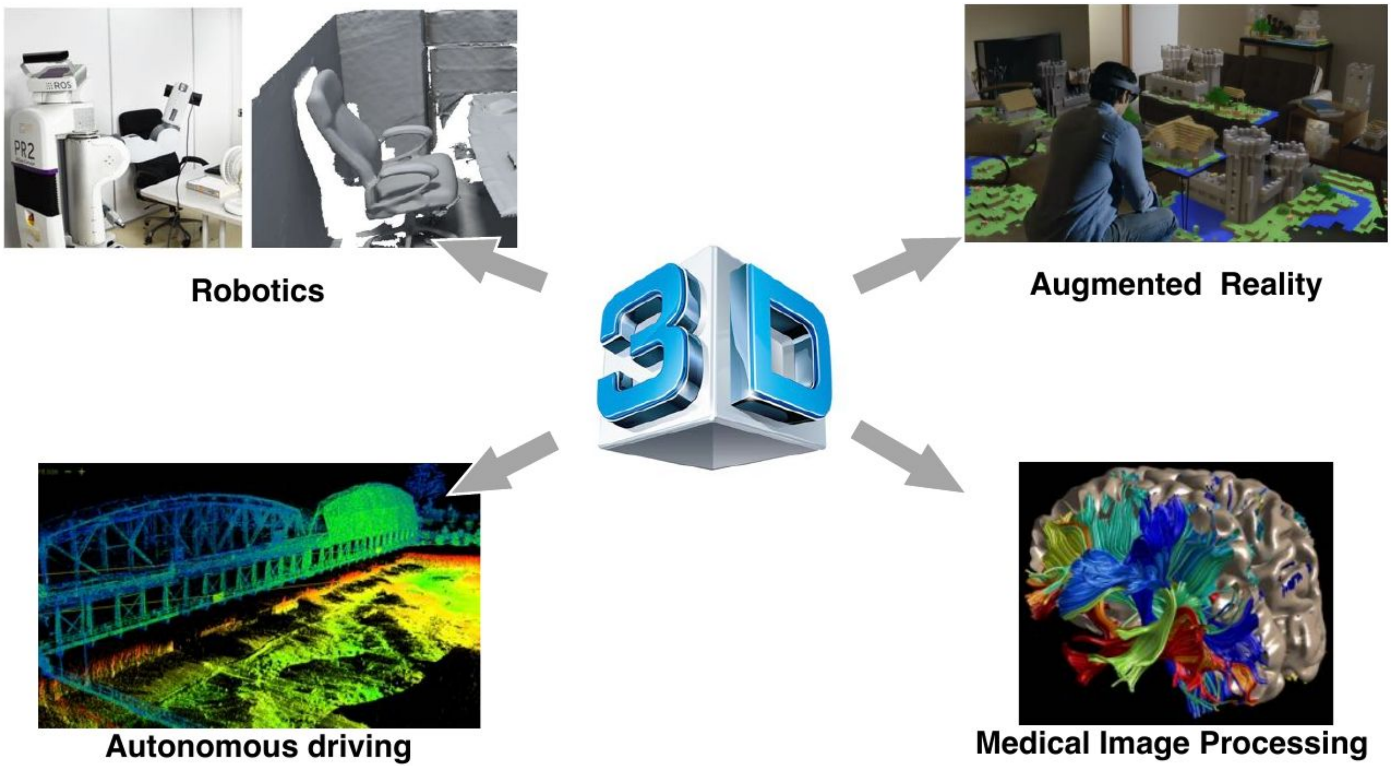
- 3D point cloud processing with ML

- Use cases
- Embedded demo

Motivation

World is in 3D...

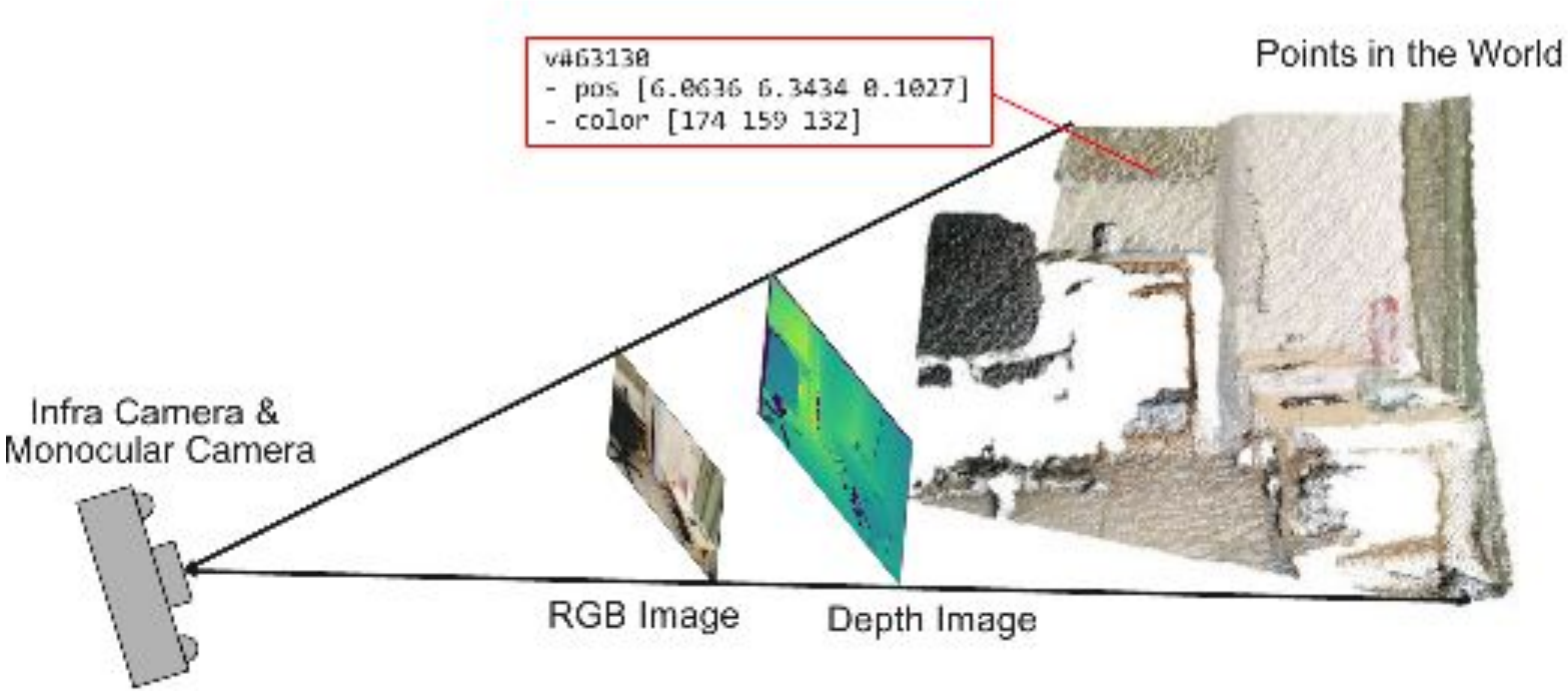
Motivation



Robotics
Augmented Reality
3D
Autonomous driving
Medical Image Processing

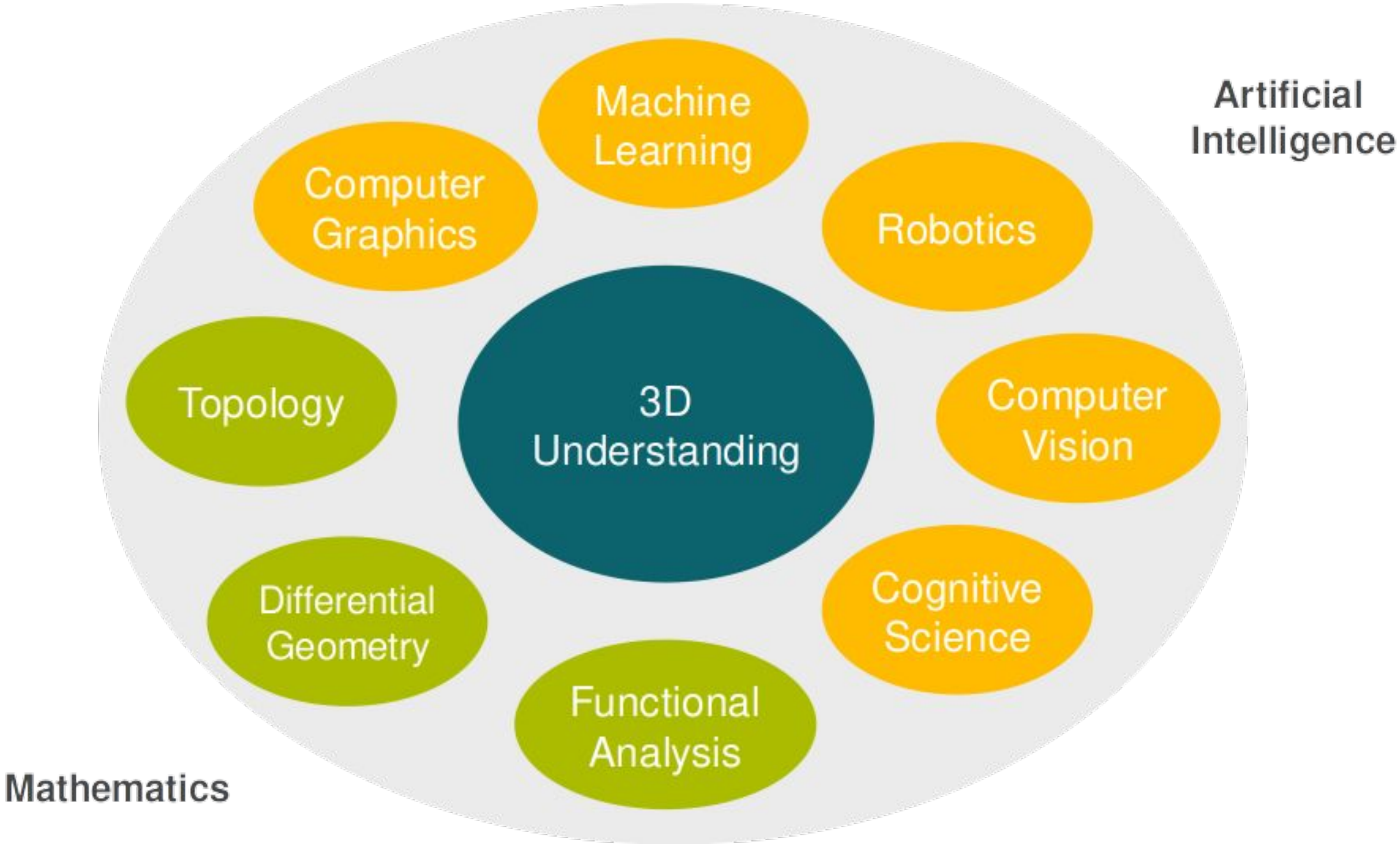
2D, 2.5D, 3D ?

From 2D to 3D



Is it necessary ML?

Now happening

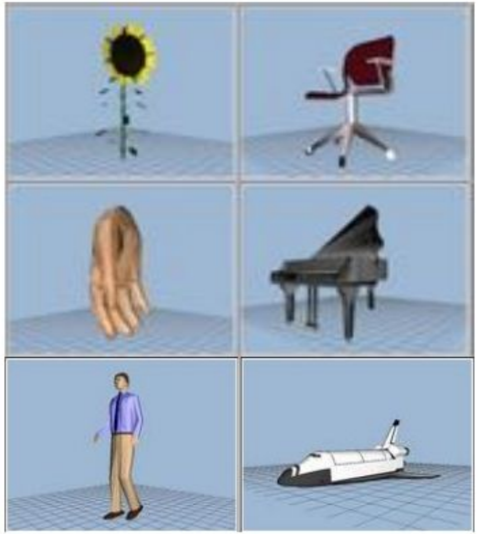


Motivation - lack of data/model ~10 years ago

Stanford bunny

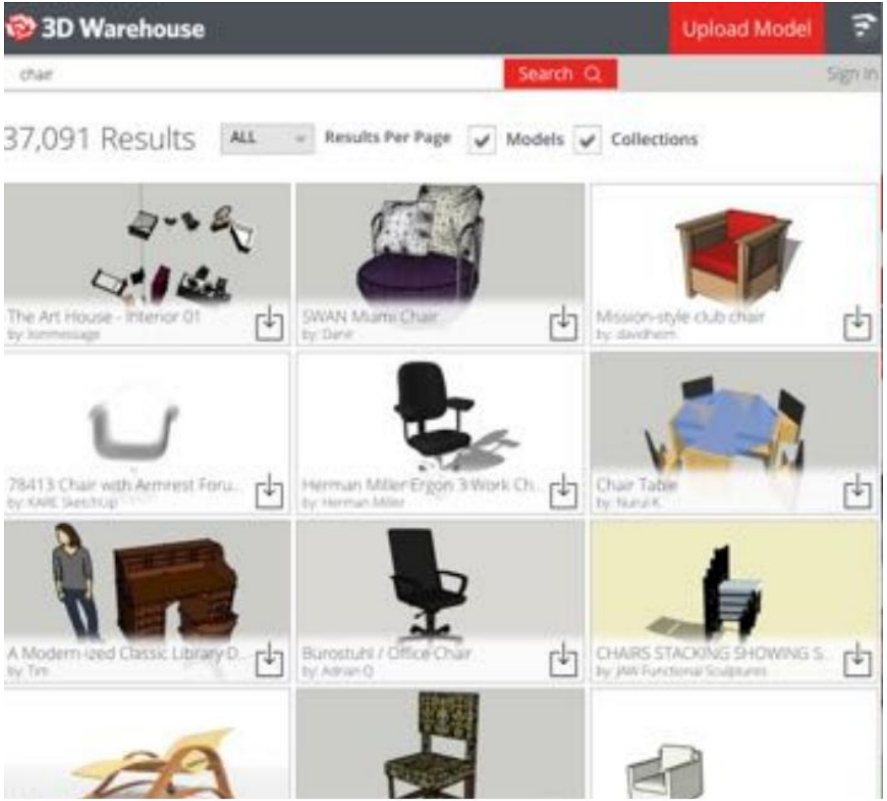
Utah teapot

1800 models in **90** categories



Princeton shape benchmark
[Shilane et al. 04]

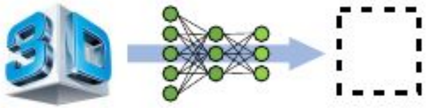
Motivation - plenty of data/model today



Motivation - plenty of data/model today

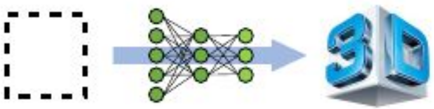
3D deep learning tasks

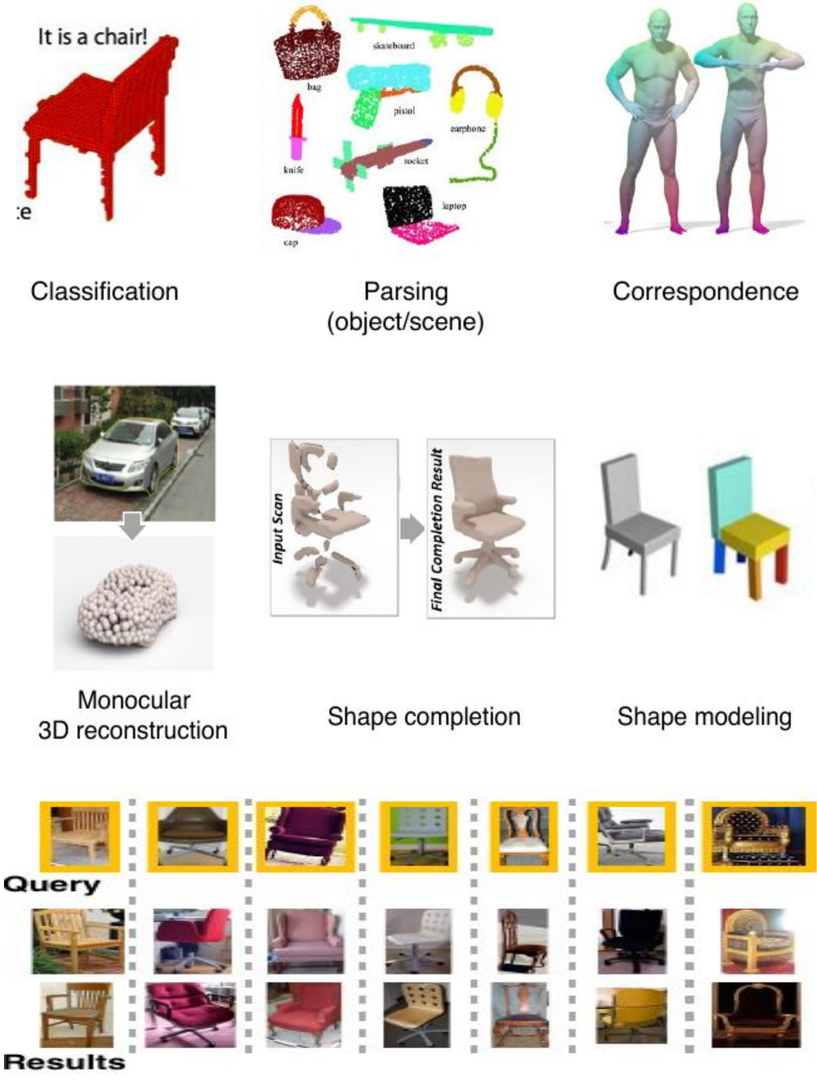
3D geometric analysis



3D assisted image analysis

3D synthesis





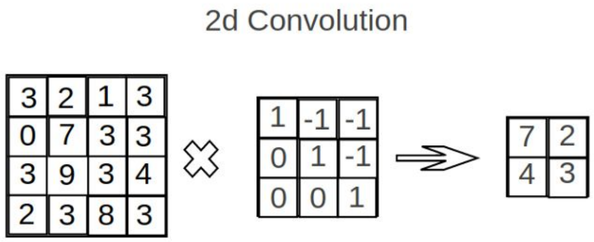
3D representation for DL

2D images: uniqueness in representation, plays well with * operator

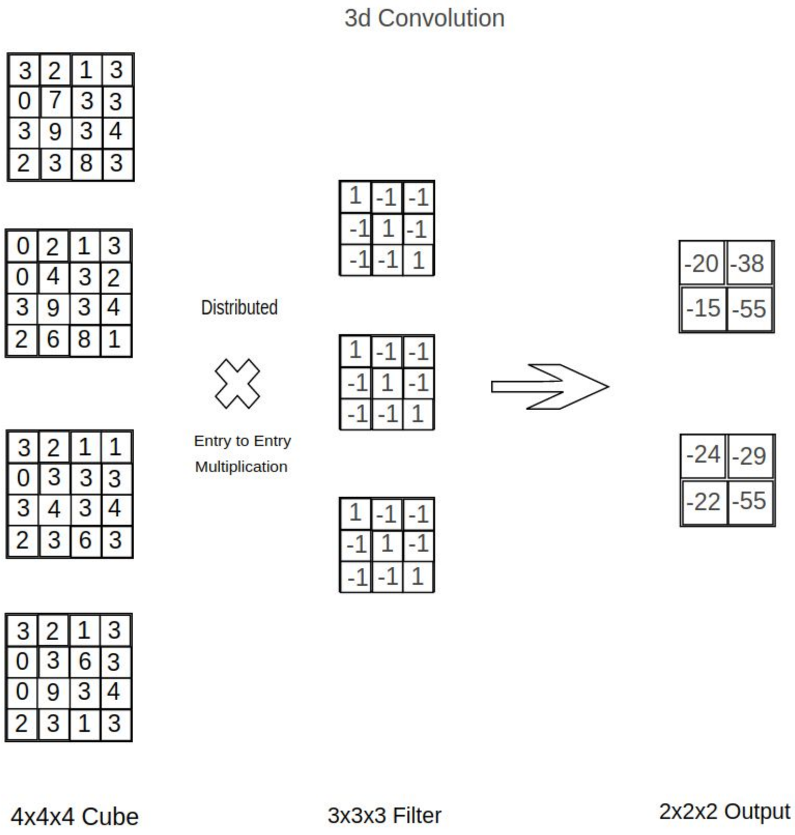
1	44	33	12	20	23	35	14
51	16	40	32	46	48	28	17
29	60	3	63	49	55	36	7
52	22	26	41	38	10	61	53
2	24	19	11	34	43	5	8
57	9	37	42	25	21	27	18
30	56	50	64	4	59	6	13
58	47	45	31	39	15	62	54

Unordered point clouds → not that easy!

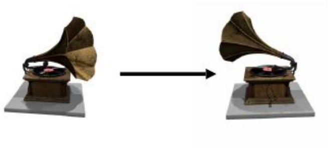
3D representation for DL - some 2D analogy

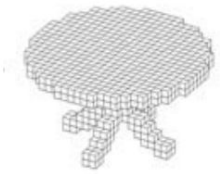


Order is still critical!



3D representation





Multiview 2D images

Volumetric

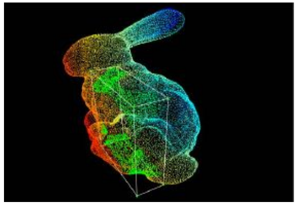
Rasterized (grid)→ direct2D, with challenges

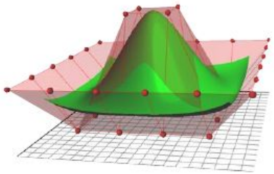
Poly Mesh

Point cloud

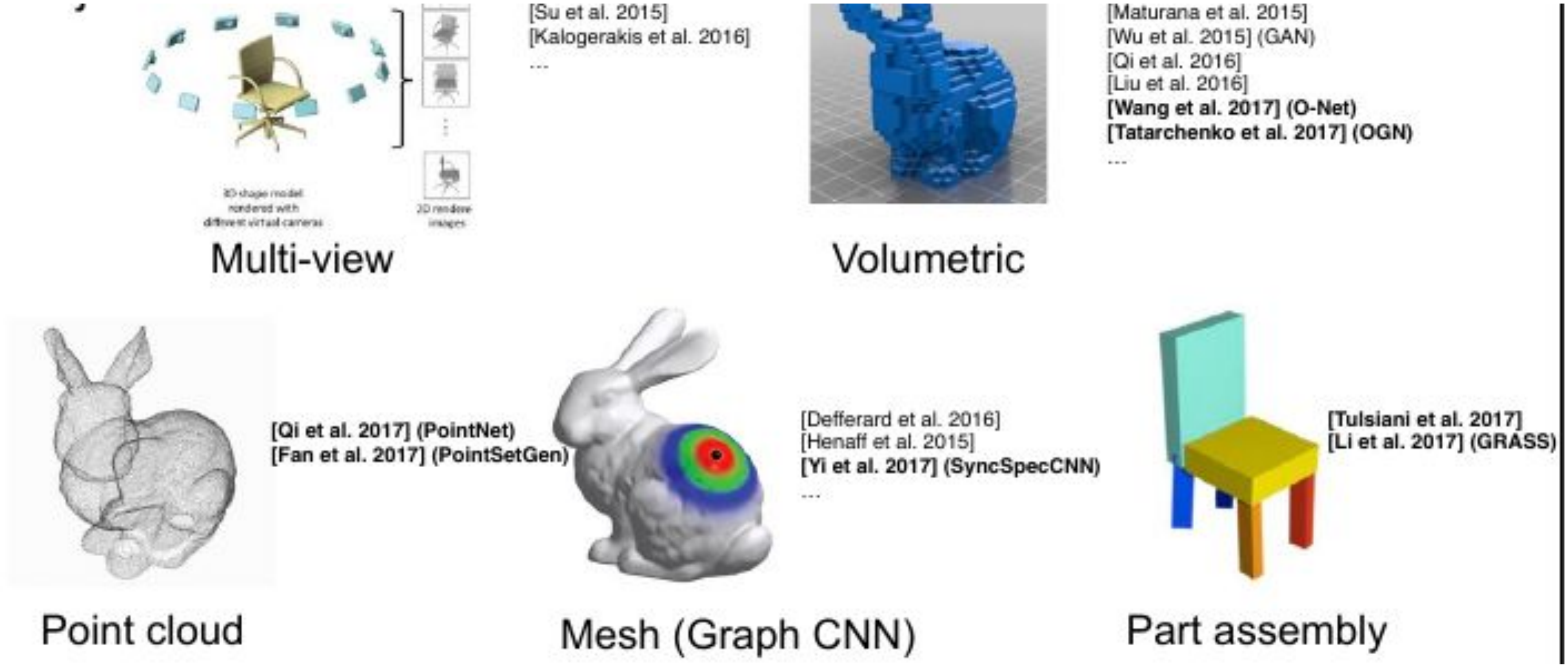
Primite based

Geometric relation (irregular) → directly CNN





3D representation for DL - references



3D point cloud processing with ML

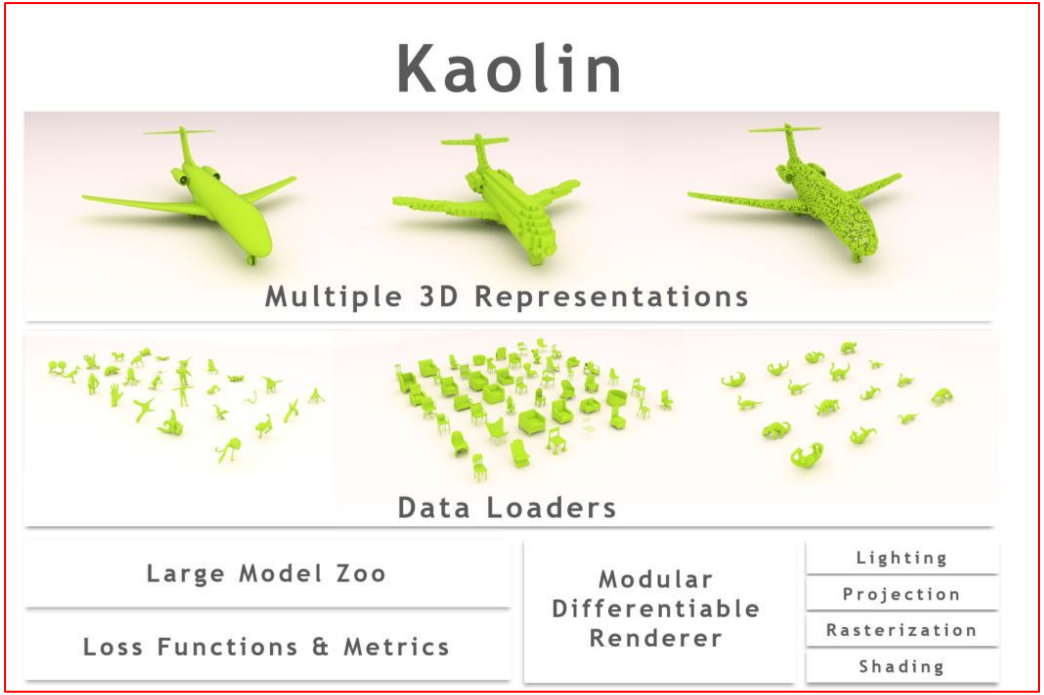
- Introduction

- **3D point cloud processing with ML**

- Use cases

- Embedded demo

3D representation for DL - tools



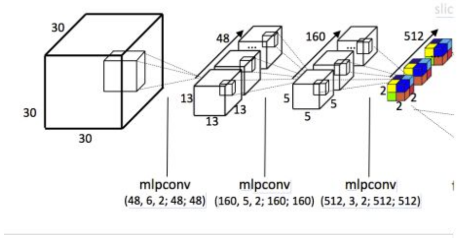
PyTorch3D

OPEN3D

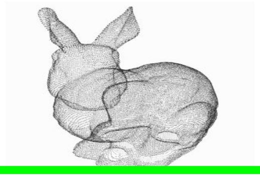
Classification

Multi-view CNN

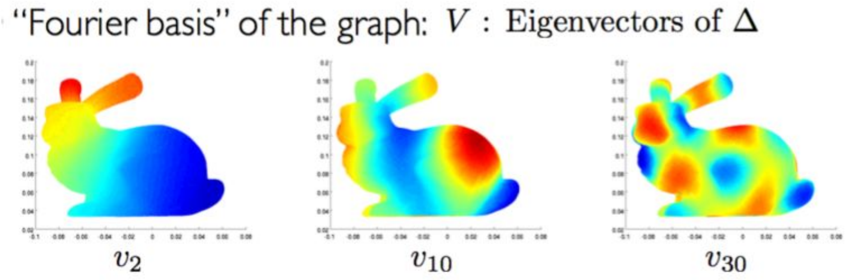
Volumetric CNN



Point nets



Spectral convolution



Classification: **Volumetric CNN**

Main ideas:

- Use CNN without explicit 3D-2D projection

- Make use of 3D native convolution (aka 4D CNN)

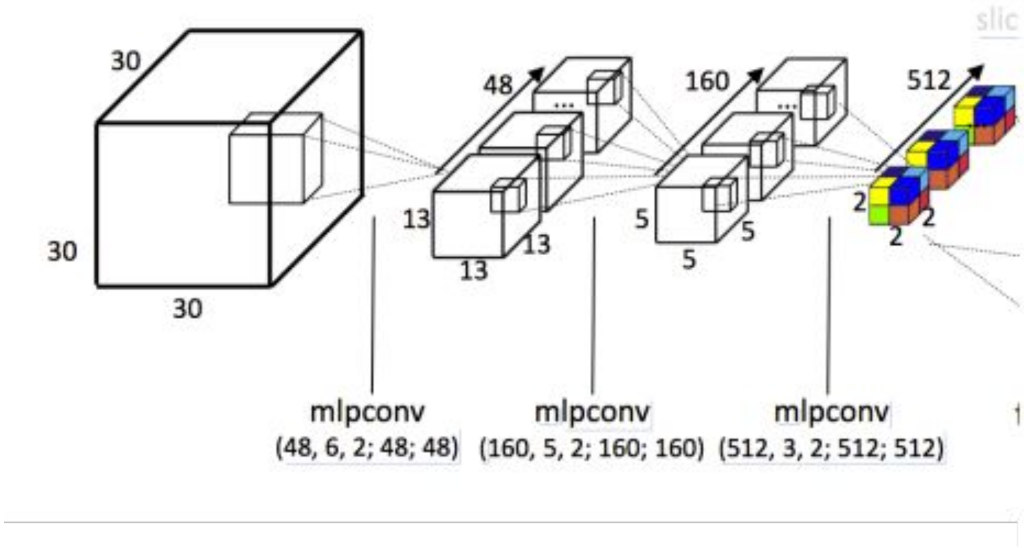
- Represent the occupied space with voxel grids

Classification: **Voxelization**

Represent the occupancy of regular 3D grids

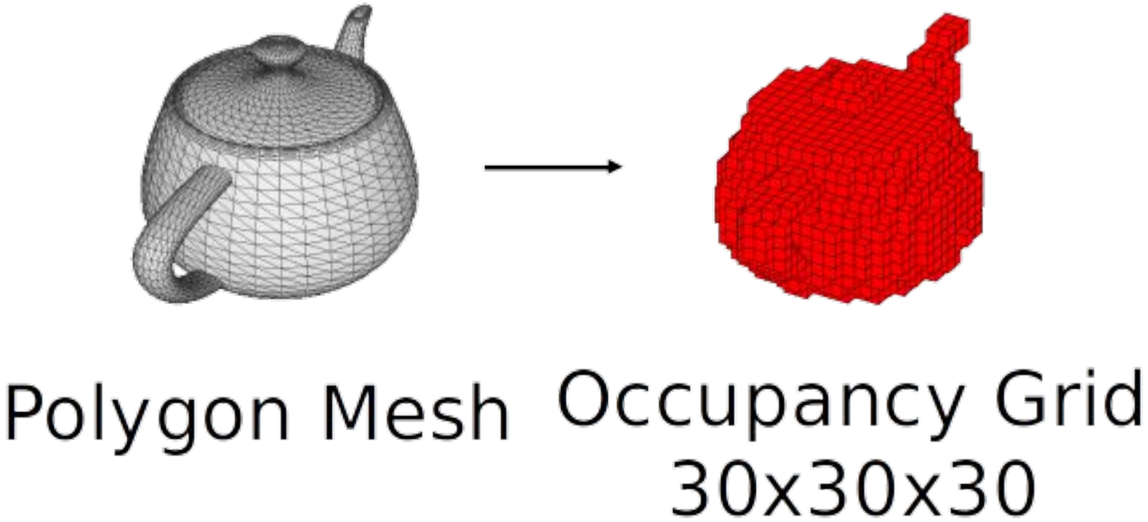
3D CNN on Volumetric Data

3D convolution uses 4D kernels



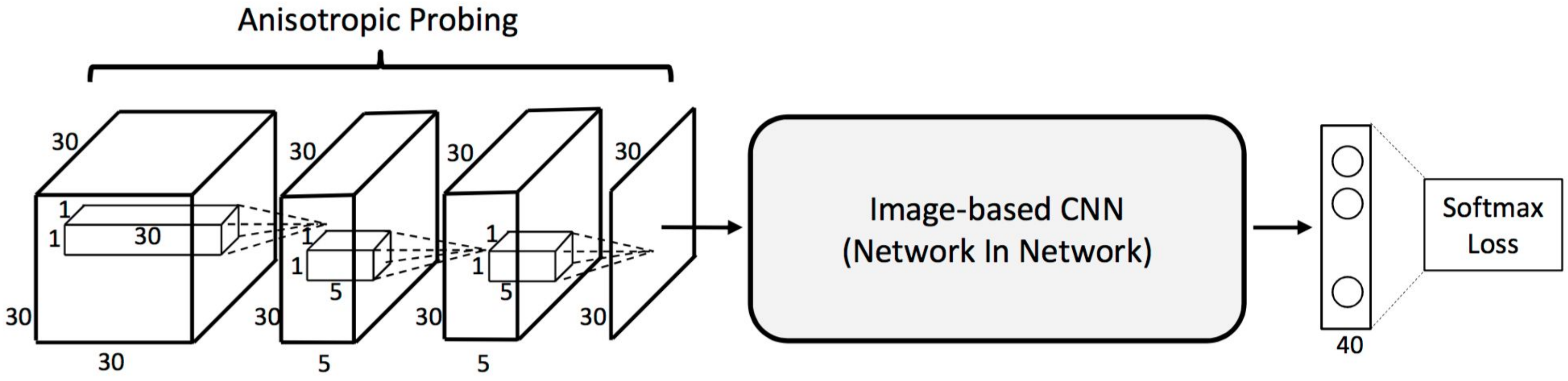
Complexity issues

What about information loss?



Basic idea: learn to project

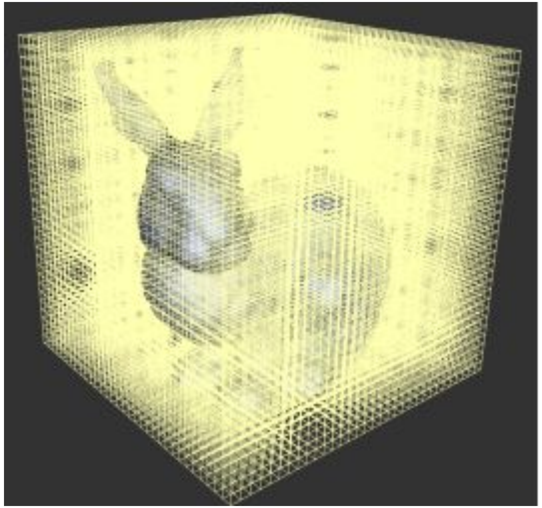
By **ray tracing** and 2D CNN low param number/low runtime is obtained

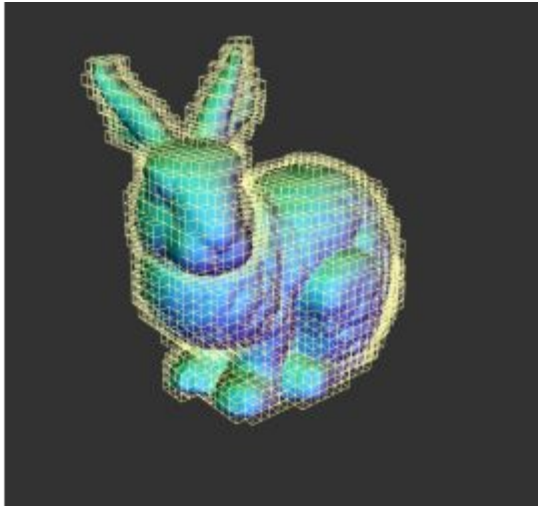


Su et al., "**Volumetric and Multi-View CNNs for Object Classification on 3D Data**", *CVPR 2016*

Voxel vs occupancy grids

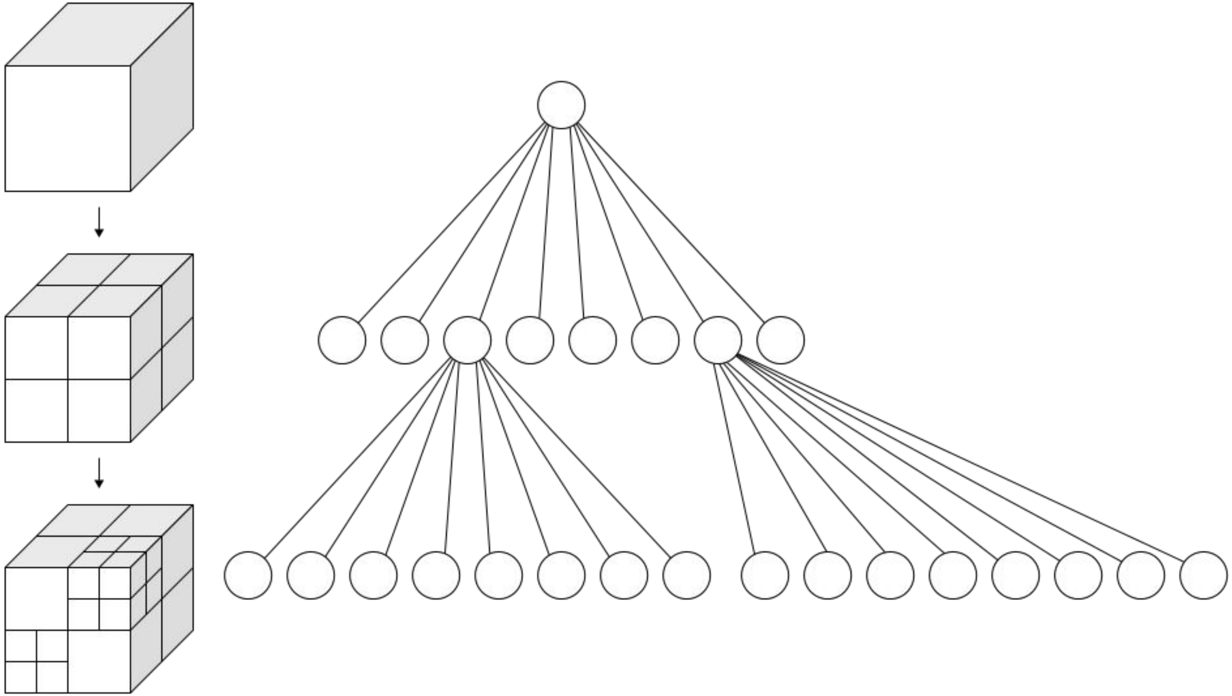
- Store the sparse surface signals
- Constrain the computation near the surface





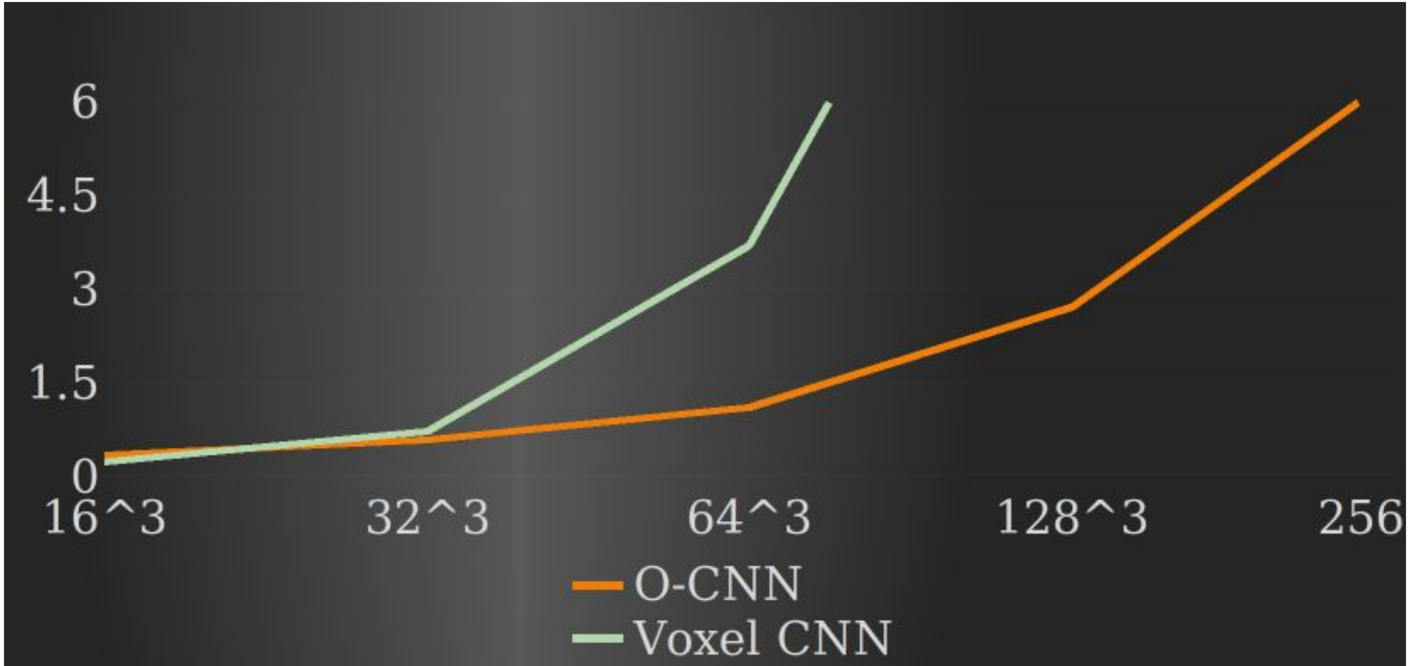
Optimized variant: octree

8 (oct) leaves for each node. Searching very efficient.

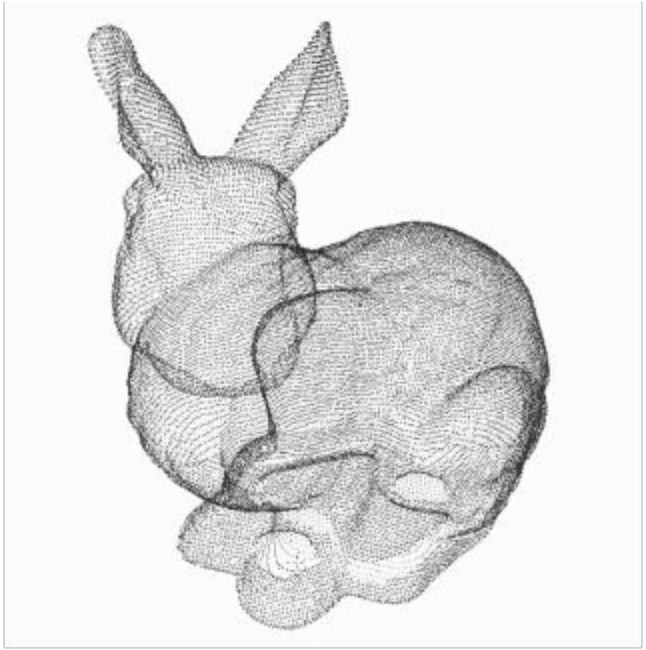


Memory efficiency → embedded systems

SparsconvNet → designed for octree representation



Classification: **Point networks**



Point cloud
(The most common 3D sensor data)

Directly Process Point Cloud Data

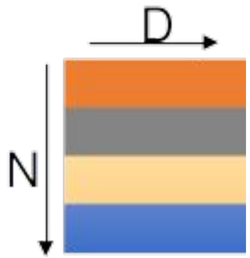
End2end learning for:

- Unstructured
- Unordered

Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation", CVPR 2017

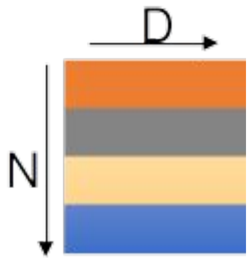
Ensure permutation invariance

Point cloud: **N odorless** points, each represented by a D dim coordinate

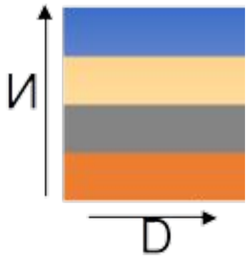


Ensure permutation invariance

Point cloud: **N odorless** points, each represented by a D dim coordinate



represents the same set as

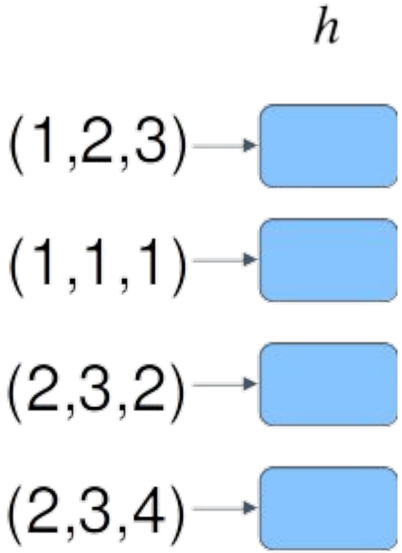


2D array representation

How to cope with this?

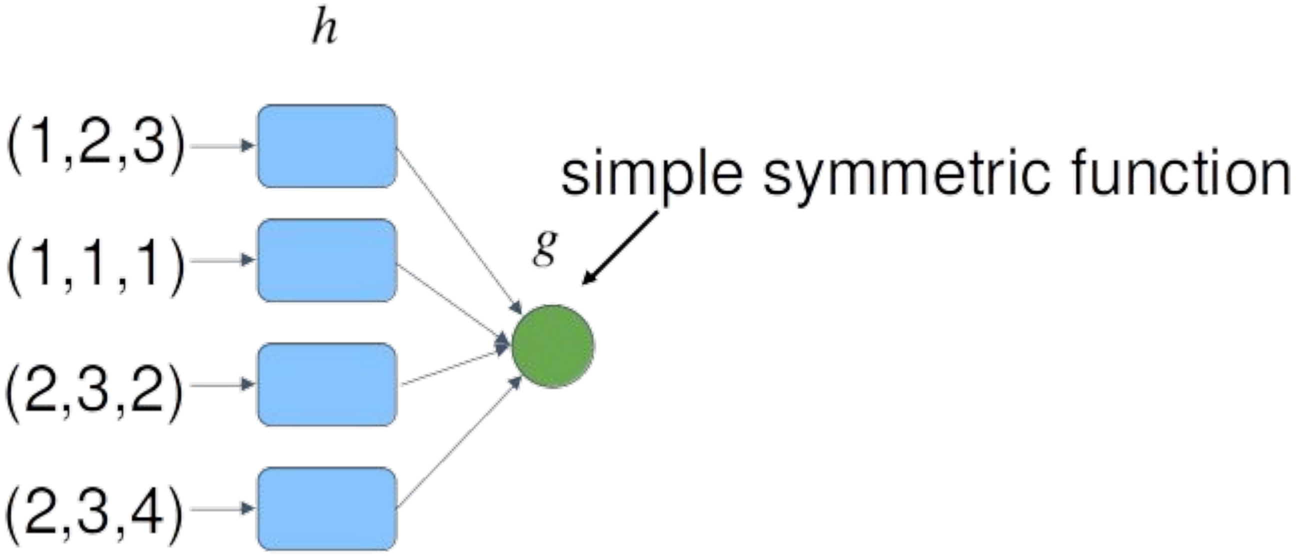
Construct a Symmetric Function

$$f(x_1, x_2, \ldots, x_n) = \gamma \circ g(h(x_1), \ldots, h(x_n))$$ Is symmetric if **g** is symmetric



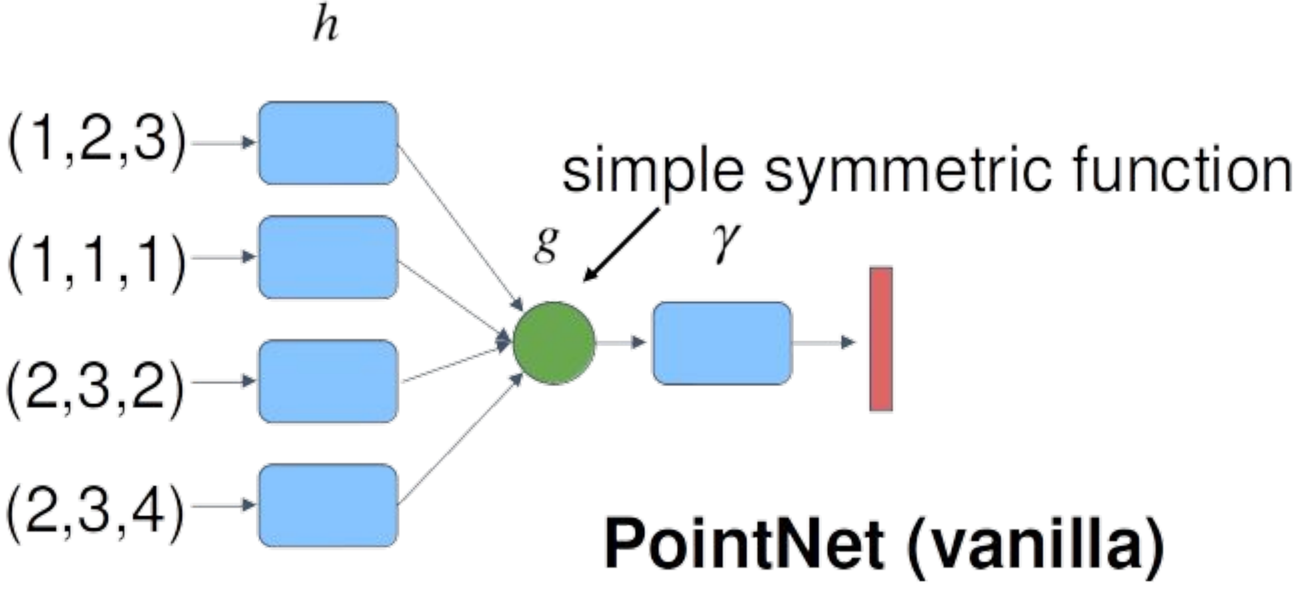
Construct a Symmetric Function

$f(x_1, x_2, \ldots, x_n) = \gamma \circ g(h(x_1), \ldots, h(x_n))$ Is symmetric if **g** is symmetric



Construct a Symmetric Function

$$f(x_1, x_2, \ldots, x_n) = \gamma \circ g(h(x_1), \ldots, h(x_n))$$ Is symmetric if **g** is symmetric

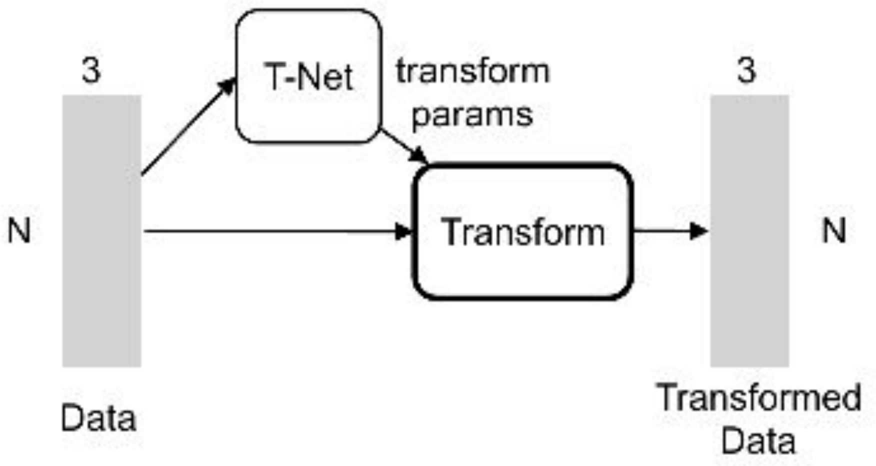


PointNet: geometric transform invariance

Solution: use some simple transform nets (T-Net)

Transform: → matrix multiplication

Dimension (e.g. 3) can be arbitrary for data

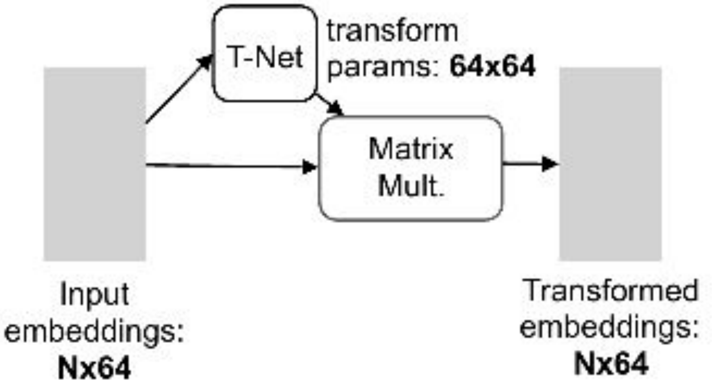


PointNet: geometric transform invariance

Solution: use some simple transform nets (T-Net)

Transform: → matrix multiplication

Dimension (e.g. 3) can be arbitrary for data



Regularization:

Transform matrix A 64x64 close to orthogonal:

$$L_{reg} = \|I - AA^T\|_F^2$$

PointNet: architecture

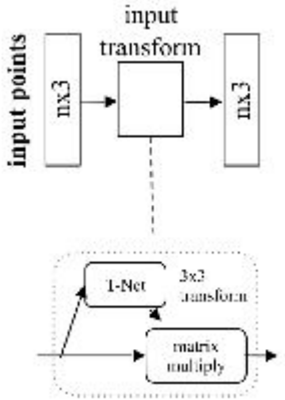
Composition of T-Nets

PointNet: architecture

Composition of T-Nets

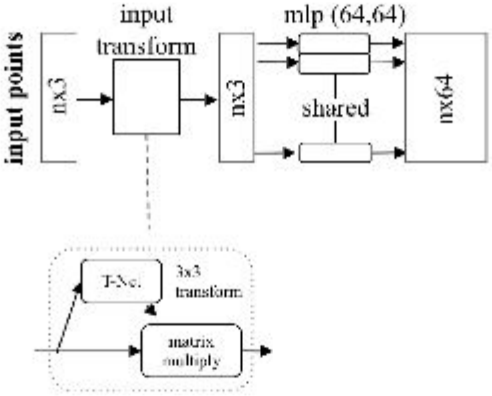
PointNet: architecture

Composition of T-Nets



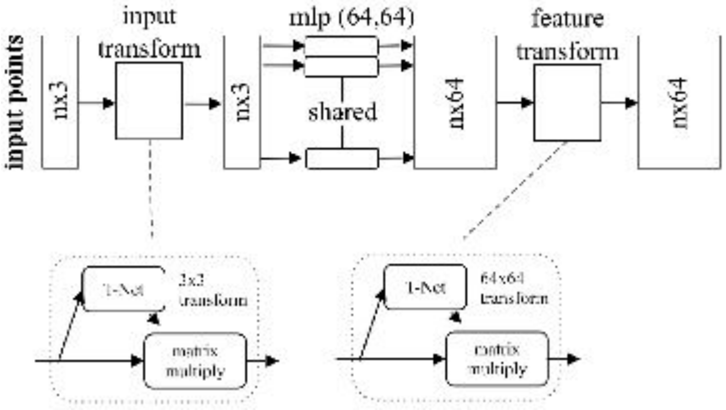
PointNet: architecture

Composition of T-Nets



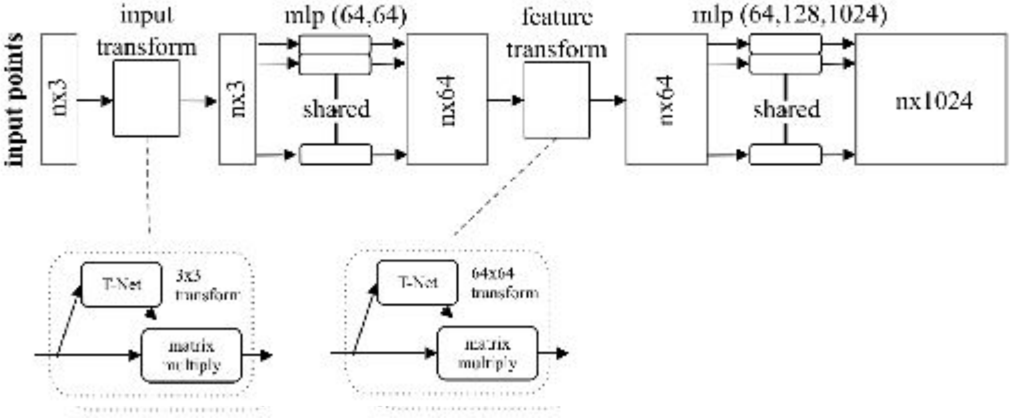
PointNet: architecture

Composition of T-Nets



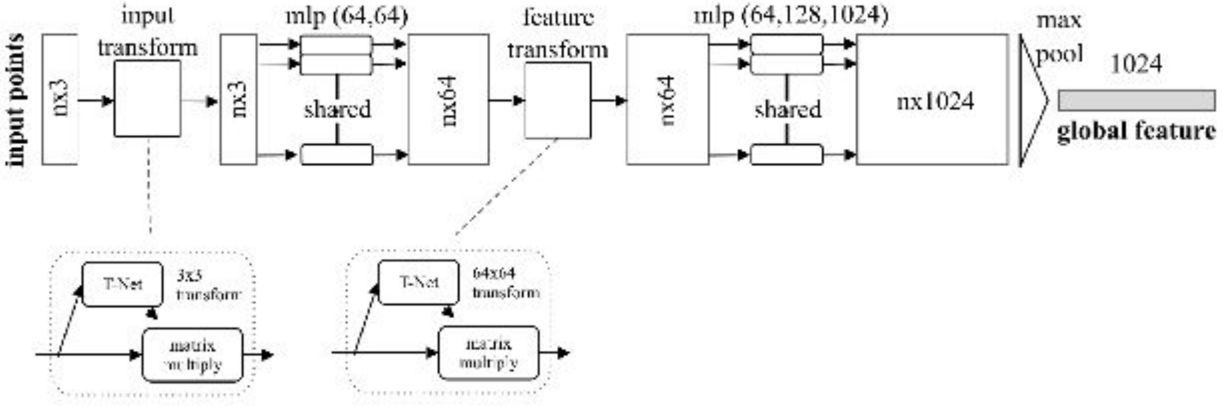
PointNet: architecture

Composition of T-Nets



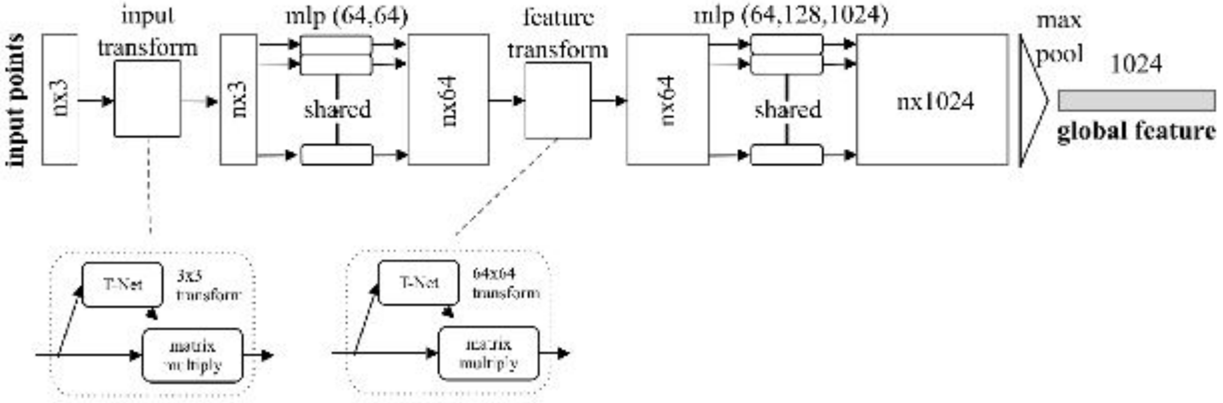
PointNet: architecture

Composition of T-Nets



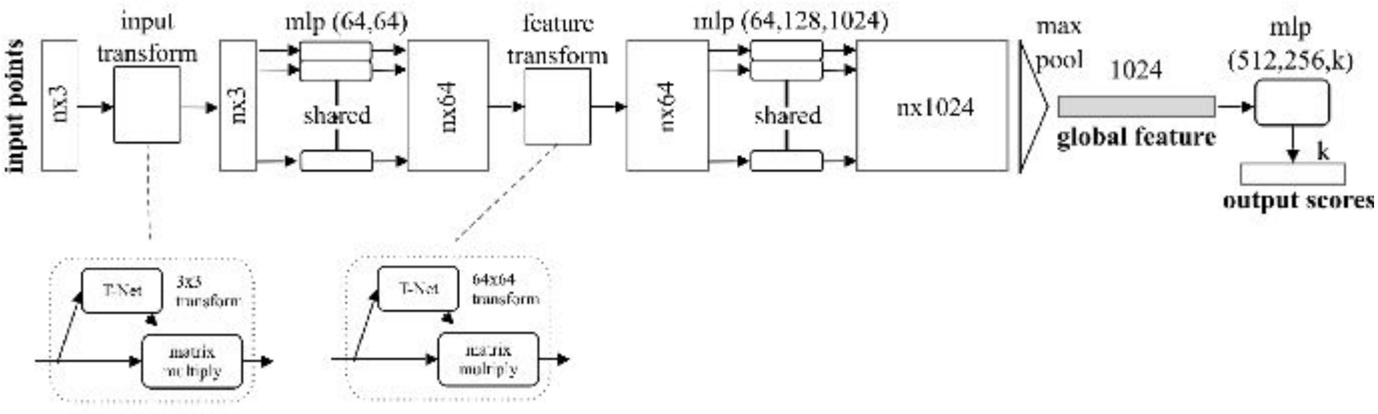
PointNet: architecture

Composition of T-Nets



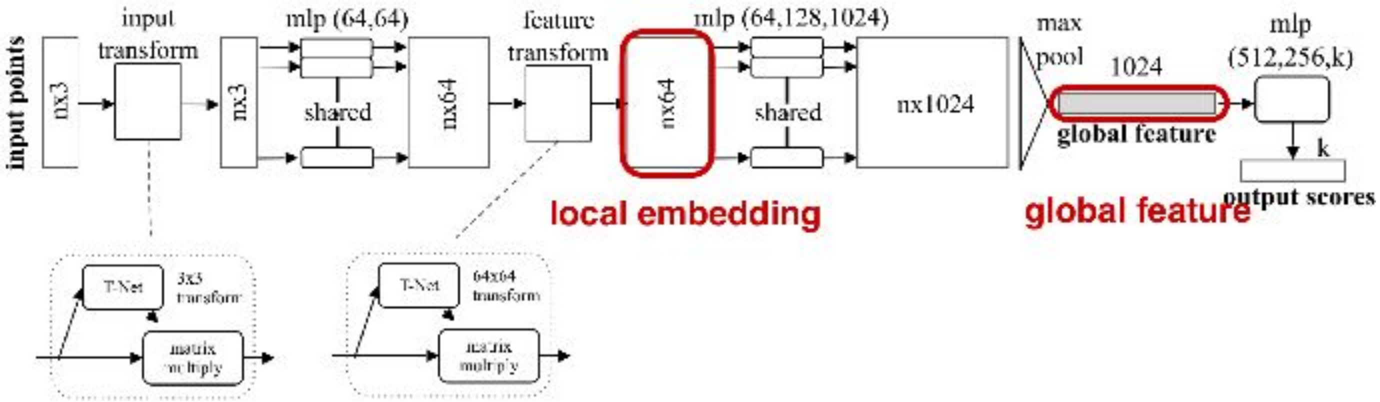
PointNet: architecture

Composition of T-Nets



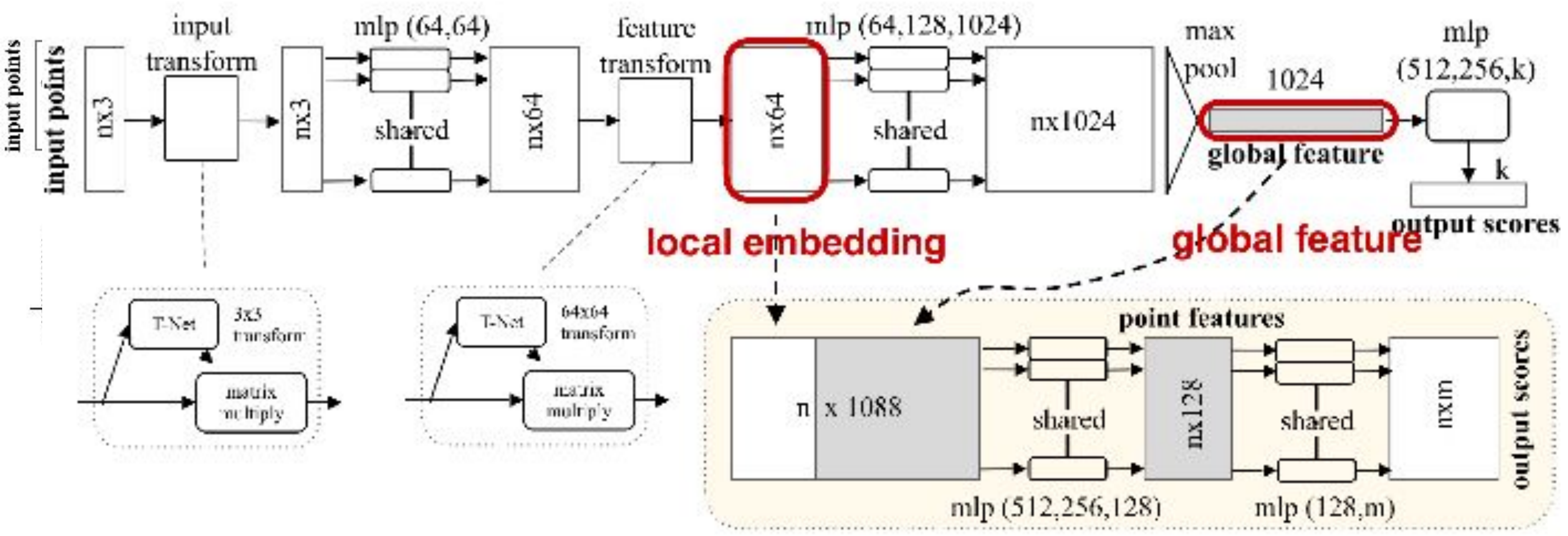
PointNet: architecture

Composition of T-Nets



PointNet: architecture

Composition of T-Nets



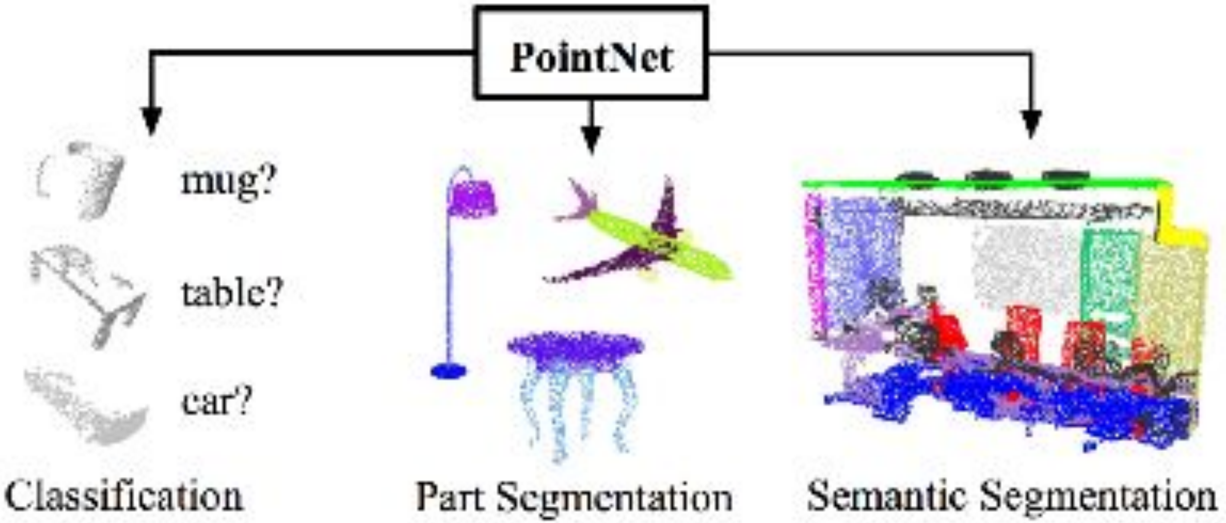
PointNet: results

3D CNNs

	input	#views	accuracy avg. class	accuracy overall
SPH [12]	mesh	-	68.2	
3DShapeNets [29]	volume	1	77.3	84.7
VoxNet [18]	volume	12	83.0	85.9
Subvolume [19]	volume	20	86.0	**89.2**
LFD [29]	image	10	75.5	-
MVCNN [24]	image	80	**90.1**	-
Ours baseline	point	-	72.6	77.4
Ours PointNet	point	1	86.2	**89.2**

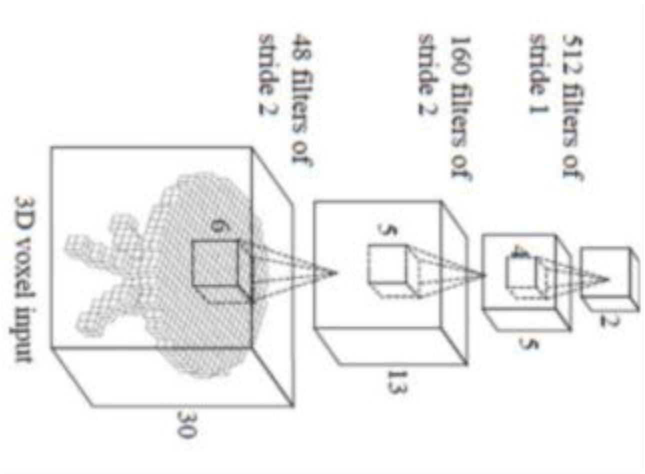
dataset: ModelNet40; metric: 40-class classification accuracy (%)

PointNet: results

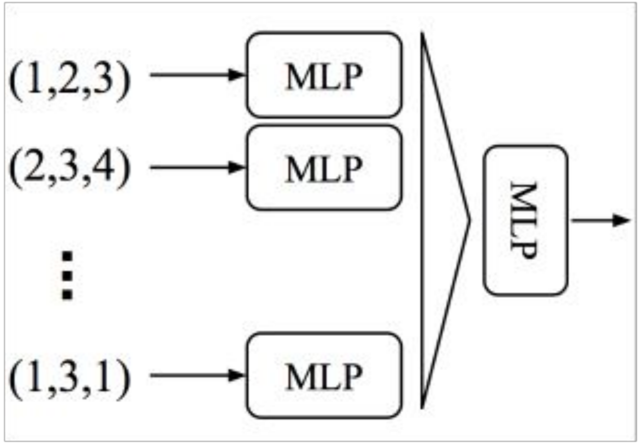


Limitations of Pointnet

- No local context for each point!
- Global feature depends on absolute coordinate.
- Hard to generalize to unseen scene configurations!



3D CNN (Wu et al.)

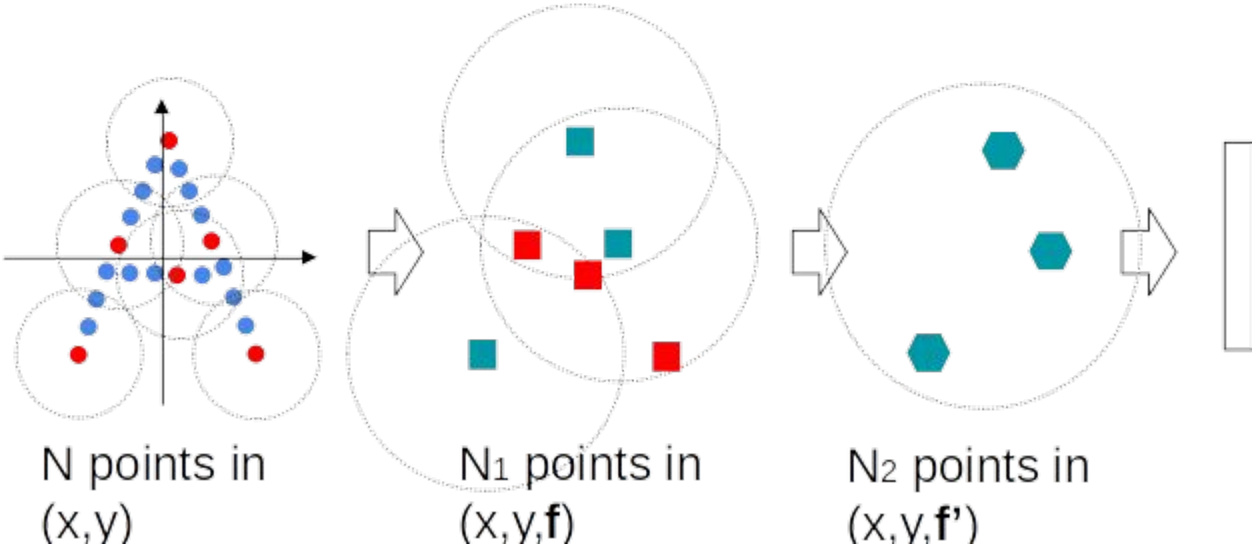


PointNet (vanilla) (Qi et al.)

PointNet v2.0: Multi-Scale PointNet

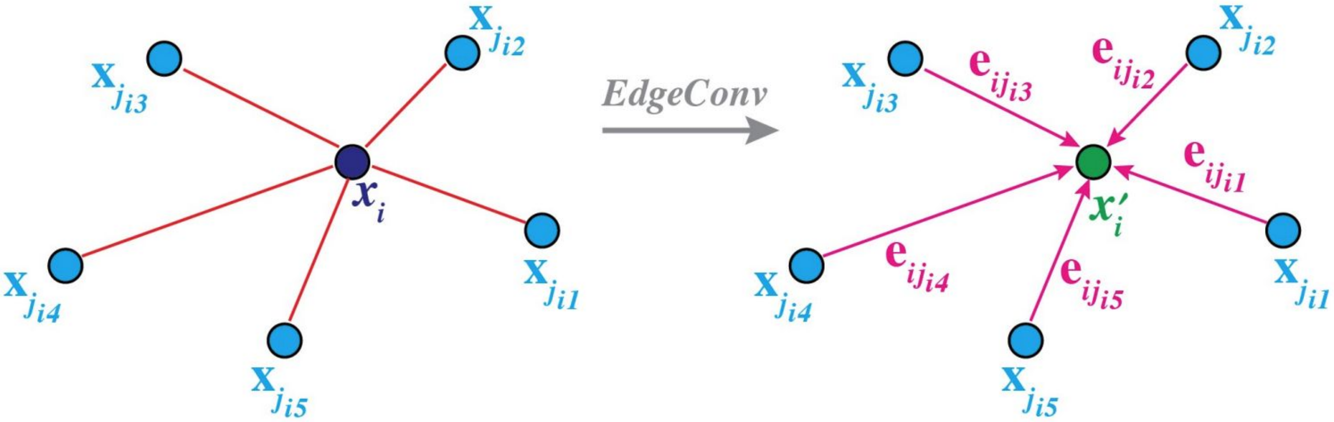
Repeat

- Sample anchor points
- Find neighborhood of anchor points
- Apply PointNet in each neighborhood to mimic convolution



Point Convolution As **Graph Convolution**

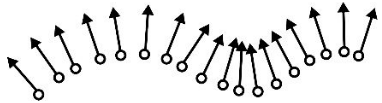
- Points -> Nodes
- Neighborhood -> Edges
- Graph CNN for point cloud processing



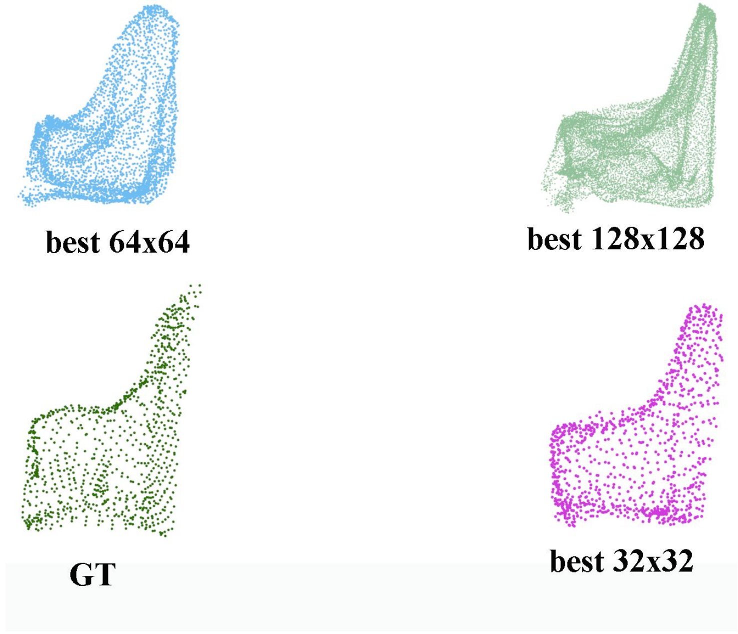
Wang et al., “Dynamic Graph CNN for Learning on Point Clouds”, Transactions on Graphics, 2019

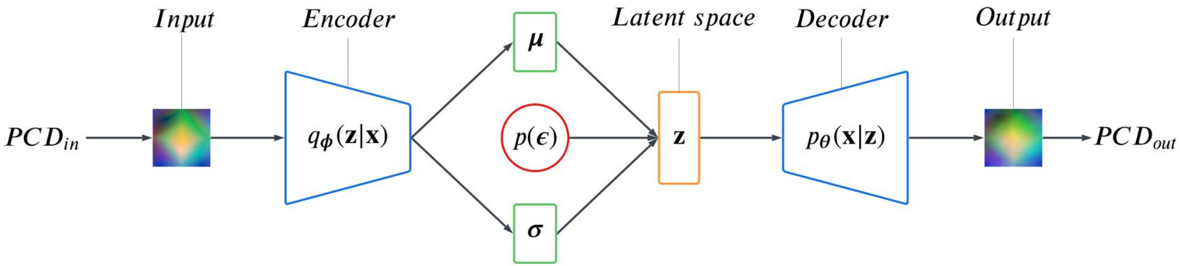
Use cases

- Introduction
- 3D point cloud processing with ML
- **Use cases**
- Embedded demo



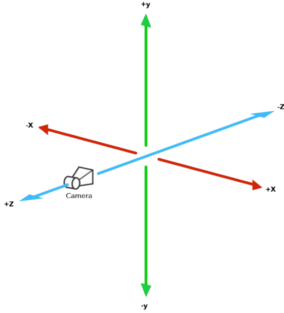
Use cases - 3DVAE



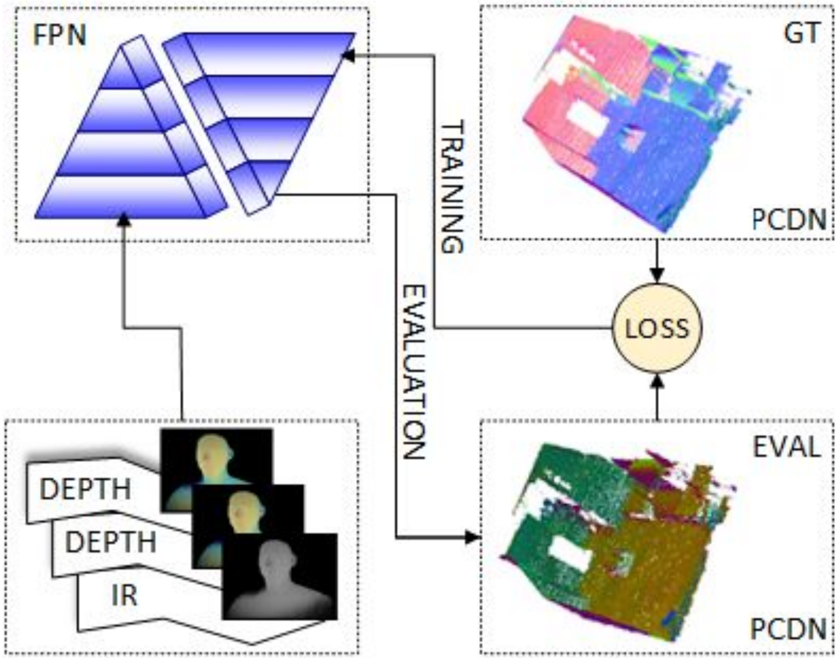


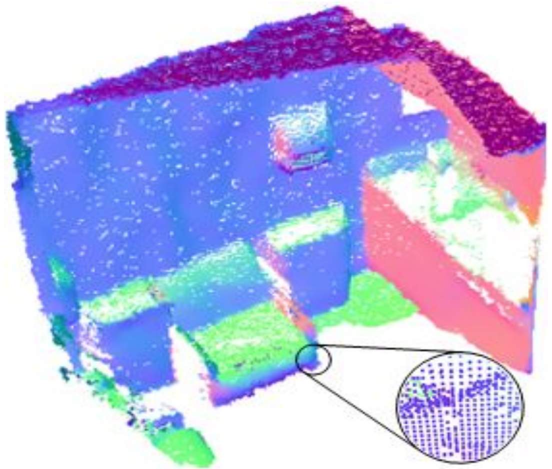
Variational Autoencoders - generative 3D

Use cases - ToFNest

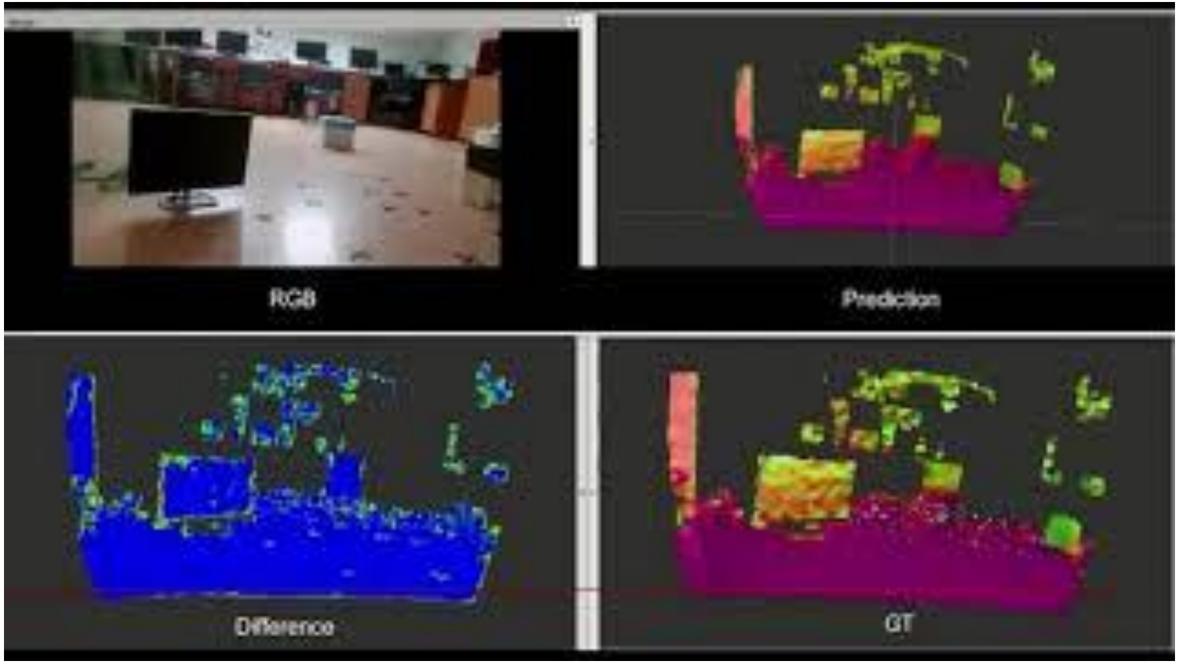


- Depth images + normal vectors in RGB -> Fast, Efficient, Robust
- FPN





ToFNest - Evaluation



RTX 3080	GTX 1060	Google Colab	Jetson Orin	Jetson NX
0,015 s	0,047 s	0,09 s	0,13 s	0,31 s

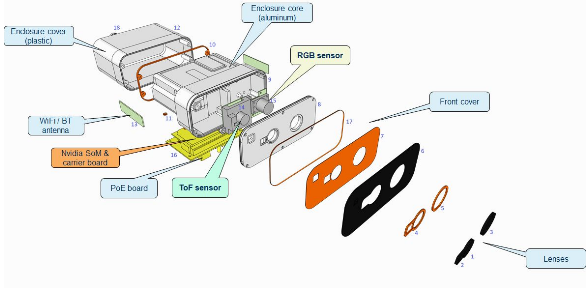
Embedded demo

- Introduction

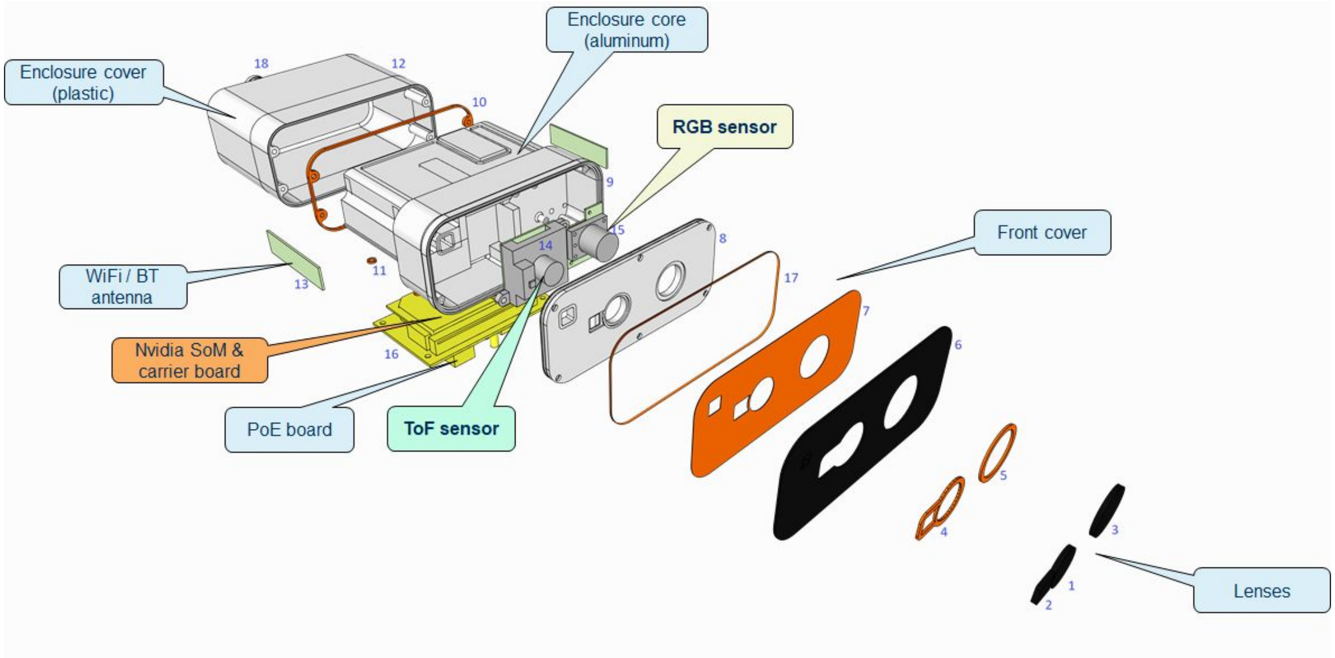
- 3D point cloud processing with ML

- Use cases

- **Embedded demo**



Embedded setup

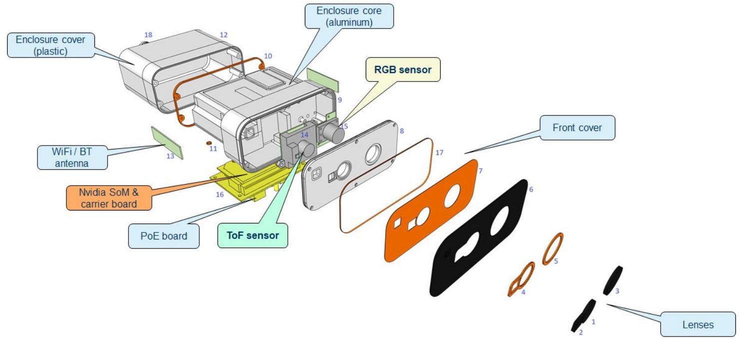


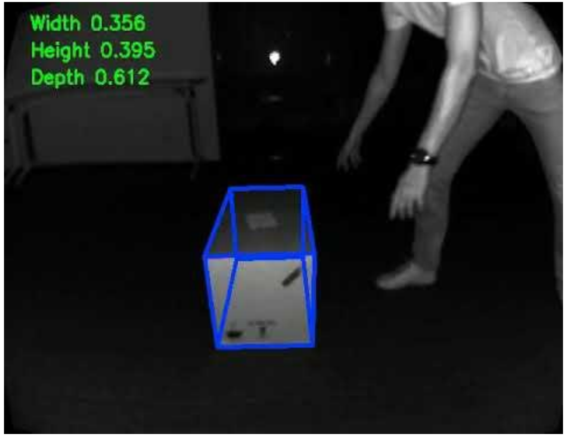
ToF from ADI

Preprocessing

People detection in 3D - WIP

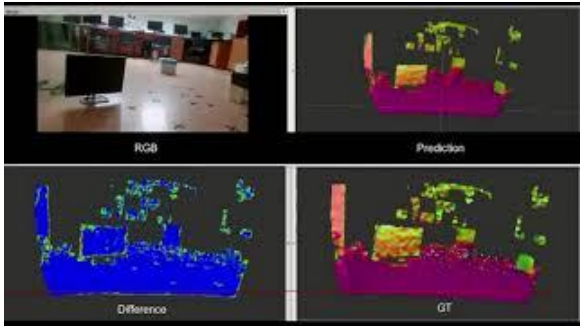
Volumetric estimation



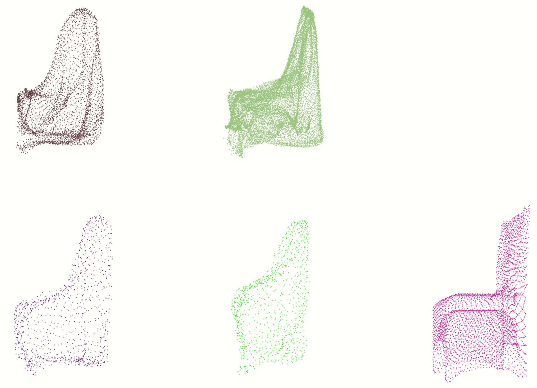


ToF cameras from ADI → papers

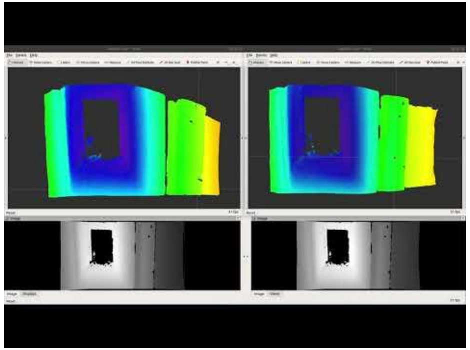
ICCV2021



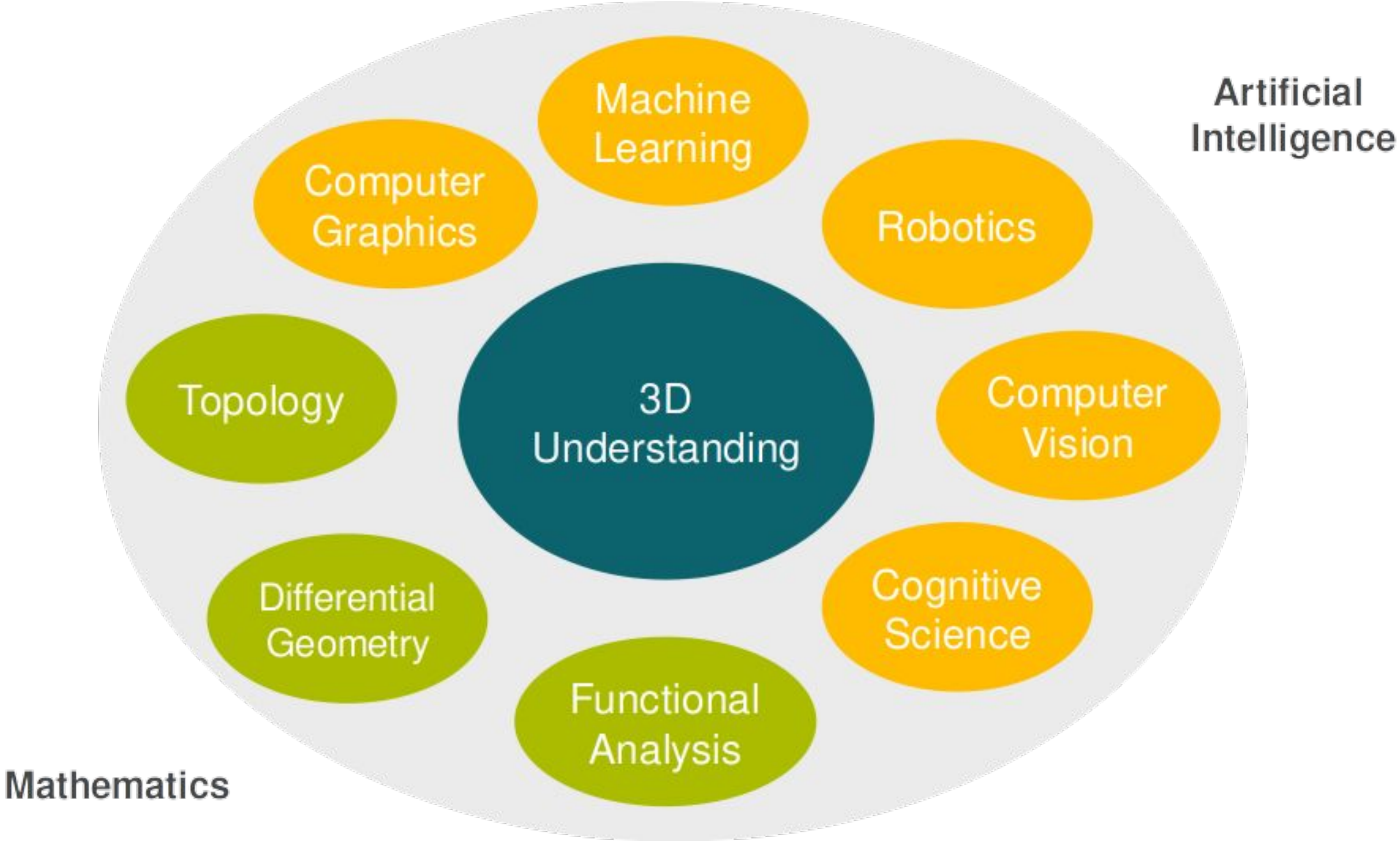
ECCV2022



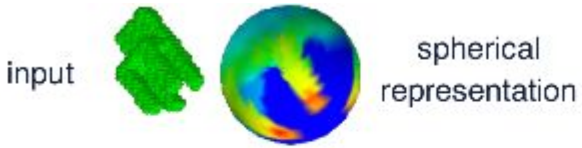
IFAC2023

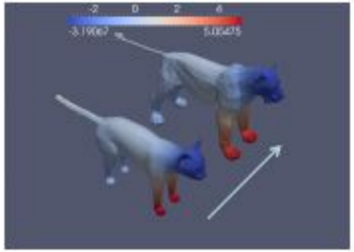


Perspectives - Now happening!

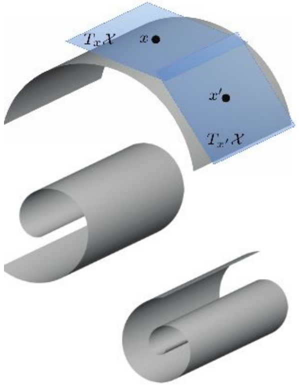


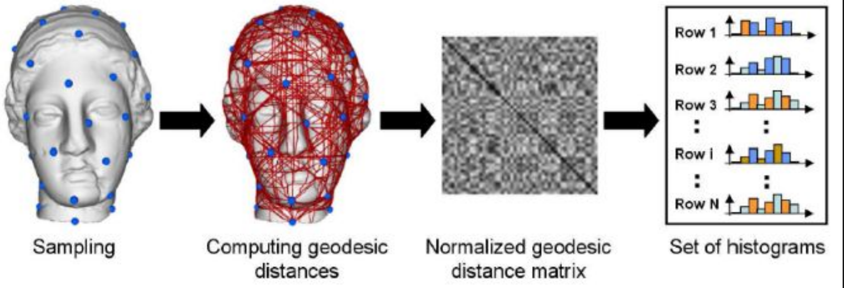
GeometricDL

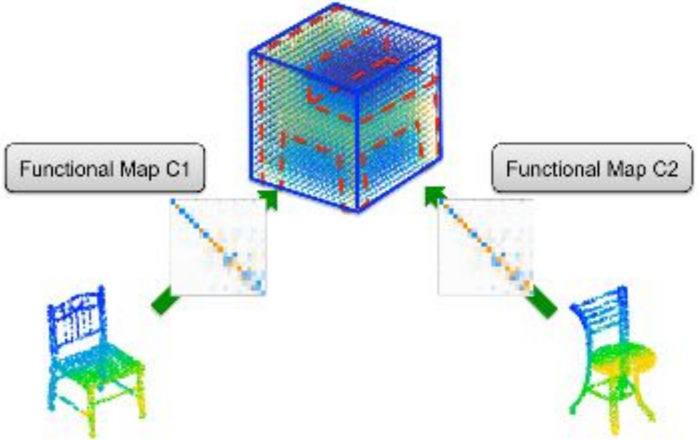


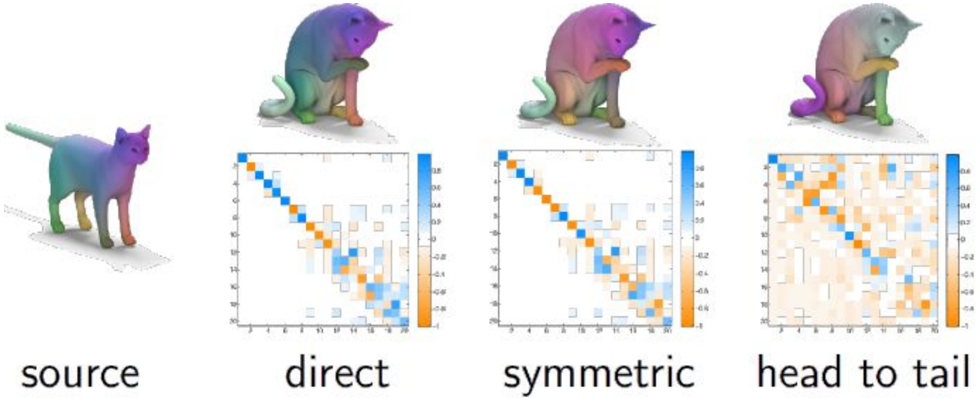


- Intrinsic/geod/R shape feature
- Heat kernel maps
- Laplacian map
- Spectral CNN
- Spectral synch









References

1. https://github.com/InternRobotics/PointLLM
2. https://distill.pub/
3. http://3ddl.stanford.edu/
4. https://github.com/NVIDIAGameWorks/kaolin
5. http://ai.ucsd.edu/~haosu
6. https://geoml.github.io/
7. https://pytorch3d.org/
8. https://github.com/intel-isl/Open3D-ML
9. https://www.shapenet.org/
10. https://modelnet.cs.princeton.edu/

Colab with PointNet (in PyTorch)

Thank you!

Levente.Tamas@aut.utcluj.ro
http://rocon.utcluj.ro

Privacy constrained semiparametric plug-in estimation

Botond Szabó (Bocconi University)

2025 Hungarian Machine Learning Days
Budapest, Hungary, 14. 08. 2025.

Co-authors

Thibault Randrianarisoa
(Toronto)

Lukas Steinberger
(Vienna)

Outline

- Introduction to privacy constrained inference
- Semi-parametric problems with plug-in estimator
- Smooth vs atomic case, examples
- Privacy constrained non-parametric inference
- Privacy constrained plug-in estimator
- Adaptation
- Summary

Introduction to α-differential privacy

Motivation of privacy protection

Increasing relevance in view of new data types in the era of big data.

- large scale medical research
- smart phone users data
- social media data
- ChatGPT
- Internet of Things (IoT)
- Consumer data collected for advertisement

Traditional/Naive techniques

Question: Do we really need math / science for privacy?

- Can't we just remove personal identifiable information from the data so that it is de-identified?
- We are only seeing aggregate statistics usually?
- Secure multi-party computation (MPC) and federated learning have made it possible for companies to train ML models with personal data while keeping it on the device not on a central database.

Problems

Anonymize data: It is possible to de-anonymize using auxiliary information:

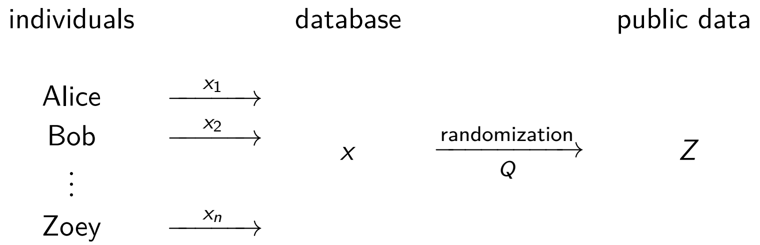
- How to break anonymity of the netflix prize dataset. Narayanan & Shmatikov (2006)
- Identifying Participants in the Personal Genome Project by Name Sweeney et al (2013).

Aggregate statistics: Possible to recover individual data.

- E.g. if $X_2, \ldots, X_n$ is known then from $S_n = \sum_{i=1}^{n} X_i$ one can recover X_1. Need for a trusted third party.
- Differencing attack: with side information, even if reporting just one, may reveal information about individuals.

ML methods: encode information of individuals in a dataset and will reveal when given an appropriate prompt, e.g. membership inference attack, unintended memorization.

Idea behind Differential Privacy



Distribution of Z should not depend too much on any individual contribution x_i.

DP in nutshell

- **Goal:** concealing a solitary sensitive data point within a particular dataset.
- **Idea:** Infuse controllable randomized noise to obtain sanitized, privacy preserving data.
- **Cost:** reduces accuracy of the method.
- **Proposed by:** Dwork et al (2006).

Definition: α-DP

Definition: Let $X = (X_i)_{i=1\dots,n}$ denote the original data and $Z = (Z_i)_{i=1,\dots,n}$ denote its sanitized version. This data Z obeys the local α-differential privacy constraint if

$$\sup_{A} \sup_{x,x':d_0(x,x')=1} \frac{Pr(Z \in A|X = x)}{Pr(Z \in A|X = x')} \leq e^{\alpha},$$

where $d_0(x, x') = |\{i : x_i \neq x_i'\}|$ denotes the Hamming distance.

Idea: The conditional distribution of Z given $X = x$ does not depend too much on the data of the i-th individual in the database, thereby protecting its privacy.

Strength: Smaller α denotes stronger privacy protection.

Relaxed version: (α, δ) differential privacy: for all A and $d_0(x, x') = 1$

$$Pr(Z \in A|X = x) \leq e^{\alpha} Pr(Z \in A|X = x') + \delta.$$

Properties

- "local" means that there is no trusted third party available for data collection and processing, see Evfimievski (2003)
- **Protocols:**
  - non-interactive: Z_i is generated from X_i independently.
  - sequentially interactive: ith person has access to $Z_1, \ldots, Z_{i-1}$ when generating Z_i.
- **Random perturbation:**
  - Laplace: α-differentiable private mechanism
  - Gauss: (α, δ)-differentiable private mechanism
- **Applications:** Apple ($2 \leq \alpha \leq 8$), Google ($0.6 \leq \alpha \leq 10$, $0 \leq \delta \leq 10^{-10}$), Microsoft ($1.67 \leq \alpha \leq 4.7$, $0 \leq \delta \leq 10^{-5}$), US Census Bureau (County Business Patterns: $\alpha = 34.9$, $\delta = 10^{-5}$; 2020 Decennial Census: $13.64 \leq \alpha \leq 49.2$, $\delta = 10^{-5}$).

Literature review

Parametric models: Smith (2008), Duchi et al (2014), Kairouz et al. (2016), Kamath et al (2018), Cai et al (2020)

Nonparametric models:

- **density estimation:** global privacy Wasserman and Zhou (2010), Hall et al (2013); local Duchi et al (2013, 2018), Butucea (2020)
- **regression:** methodology Smith (2021), Golowich (2021) theory Gyorfi and Kroll (2023).

Semi-parametric problems:

- Linear functionals Rohde and Steinberger (2018)
- Integrated square $\int f^2(x)dx$, Butucea et al (2023)

BUT! No general approach, case-by-case studies.

Motivating example

Data: sensitive iid data $X_1, ..., X_n \in [-M, M]$

Goal: estimating the moments $\mathbb{E}X^k$.

Privacy constraint: Laplace mechanism

$$Z_{i,k} = X_i^k + Lap\left(\frac{\alpha}{2M^k}\right)$$

Private estimator: $\hat{\theta}_n^{(k)} = \frac{1}{n}\sum_{i=1}^n Z_{i,k}$

Problem: For estimating e.g. the mean $\mathbb{E}X$ from $Z_{1,1}, .., Z_{n,1}$ we used all privacy budget α. For estimating e.g.2nd moment we need new privatized data $Z_{i,2}$, $i = n+1, .., 2n$ or use $Z_{1,1}, .., Z_{n,1}$ with deconvolution, but its suboptimal.

Goal: Provide privatized data, which can be used for multiple purpose simultaneously.

Model and examples

Model

Density estimation problem: $X_1, ..., X_n \overset{iid}{\sim} f$, with

$$f \in \mathcal{W}_p := \left\{ f \in C^p[0,1] : \; f \geq 0, \; \int_0^1 f = 1, \; \|f\|_{(\infty,p)} < M \right\},$$

$p \in \mathbb{N}$, where for $1 \leq q \leq \infty$

$$\|f\|_{(q,p)} = \sum_{j=0}^{p} \left(\int_0^1 (f^{(j)})^q dx \right)^{1/q}.$$

Semi-parametric model: Consider functionals $\Lambda : C^p \to \mathbb{R}$, s.t. for some $0 \leq m < p$,

$$\Lambda(f+h) = \Lambda(f) + T_f(h) + O(\|h\|^2_{(2,m)}), \qquad (1)$$

where for $f \in \mathcal{W}_p$, $h \in C^p[0,1]$ with $\|h\|_{(\infty,m)}$ small enough and T_f a bounded linear functional on $C^p[0,1]$, see Goldstein & Messer (1992).

Functional

In view of the Hahn-Banach and Riesz representation theorems

$$T_f(h) = \sum_{j=0}^{p} \int_0^1 h^{(j)} d\mu_j,$$

where μ_j is a finite signed Borel measures on $[0, 1]$ (possibly depending on f).

Cases:

- **Smooth functionals:** $T_f(h) = \int h\omega_f$, $\forall f \in \mathcal{W}_p$, with $\sup_{f\in\mathcal{W}_p} \|\omega_f\|_\infty < \infty$.
- **Atomic functionals:** of index $s \in \{0, .., p\}$, where

$$T_f(h) = \sum_{j=0}^{s_f} \int_0^1 h^{(j)} d\mu_{j,f}$$

with $\mu_{s_f,f}$ having a discrete component $\delta_{s_f,f}$, and $s = \max_{f\in\mathcal{W}_p} s_f$.

Example: Point evaluation of derivatives

Functional:

$$\Lambda(f) = f^{(r)}(x_0).$$

Note:

$$(f+h)^{(r)}(x_0) = f^{(r)}(x_0) + \int h^{(r)} \delta_{x_0}$$

This is a differentiable functional of order $m = 0$ and index $s = r$.

Extension:

$$\Lambda(f) = g(f^{(r)}(x_0)),$$

with g twice differentiable and $g' \neq 0$

Estimation rate: for $f \in \mathcal{W}_p$, the rate is $n^{-(p-r)/(2p+1)}$.

Example: Fisher Information

Functional: $\Lambda(f) = \int_0^1 (f')^2/f$

Note:

$$\Lambda(f+h) = \Lambda(f) + T_f(h) + O\left(\|h\|_{(2,1)}^2\right),$$

with

$$T_f(h) = -\int_0^1 \frac{h(f')^2}{f^2} + 2\int_0^1 \frac{f'h'}{f}$$
$$= 2h(1)\frac{f'(1)}{f(1)} - 2h(0)\frac{f'(0)}{f(0)} + \int_0^1 \frac{h(f')^2}{f^2} + 2\int_0^1 h\frac{f''}{f}.$$

Regularity: Λ is atomic of order $m = 1$ and index $s = 0$.

Estimation rate: for $f \in \mathcal{W}_p$, $2 \leq p$, $\inf_x f(x) > 0$, the rate is $n^{-p/(2p+1)}$.

Example: Entropy

Functional: $\Lambda(f) = \int_0^1 f \log f$.

Note:

$$\Lambda(f+h) = \Lambda(f) + \int_0^1 h \log(f) + \int_0^1 f \log(1 + h/f) + \int_0^1 h \log(1 + h/f),$$

hence $T_f(h) = \int_0^1 h \log(f)$.

Regularity: Λ is smooth of order $m = 0$.

Estimation rate: for $f \in \mathcal{W}_p$, $p \geq 2$, $\inf_x f(x) > 0$, the rate is $n^{-1/2}$.

Privacy constrained estimation: Non-adaptive setting

Plug-in estimation

Idea: Estimate $\hat{f}$ the density f and plug it in into the linear functional

$$\widehat{\Lambda(f)} = \Lambda(\hat{f}).$$

Density estimation:

- Histograms: not good, one needs to estimate the derivatives of the density as well.
- *Kernel estimators:* not good, privacy mechanism is difficult, only approximate local DP.
- *Spline wavelet:* optimal for simultaneously estimating the density and its derivatives, in L_2 and pointwise loss as well.

Privatized data

$$Z_{ijk} = \psi_{j,k}(X_i) + \sigma_{\alpha,j} Y_{ijk}, \qquad k \in \mathcal{M}_j,$$

where $Y_{ijk} \overset{iid}{\sim} Lap(1)$, $\psi_{j,k}$ are the spline wavelet basis, and

$$\sigma_{\alpha,j} = \frac{C_d \|\psi\|_\infty}{\alpha} j^2 2^{j/2}.$$

Lemma: The privacy mechanism defined above is locally α-differentially private.

Wavelet coefficients: privatized empirical wavelet coefficients $\bar{Z}_{jk} = n^{-1} \sum_{i=1}^{n} Z_{ijk}$

Density estimation:

$$\hat{f}_n = \hat{f}_n^{j_n} = \sum_{j=1}^{j_n} \sum_{k \in \mathcal{M}_j} \bar{Z}_{jk} \tilde{\psi}_{j,k}.$$

Point-wise and L_2-convergence For $2^{j_n} \asymp (n\alpha^2 \log^{-2} n)^{\frac{1}{2p+2}} \wedge n^{\frac{1}{2p+1}}$ we have

$$\max\left(\mathbb{E}_{Q\mathbb{P}_f} |\hat{f}_n^{(q)}(x_0) - f^{(q)}(x_0)|^2,\ \mathbb{E}_{Q\mathbb{P}_f} \|\hat{f}_n^{(q)} - f^{(q)}\|_{L_2}^2 \right)$$
$$\leq C_{d,q,M} (n\alpha^2 \log^{-2} n)^{-\frac{2(p-q)}{2p+2}} \vee n^{-\frac{2(p-q)}{2p+1}}.$$

Convergence rate for atomic functionals

Theorem [estimation atomic]: Let $f \in \mathcal{W}_p$, $p \leq d+1$ and suppose Λ is an atomic functional of index s. Under some mild technical conditions, the plug-in estimator $\widehat{\Lambda(f)} = \Lambda(\hat{f}^{j_n})$ with $2^{j_n} \asymp (n\alpha^2 \log^{-2} n)^{\frac{1}{2p+2}} \wedge n^{\frac{1}{2p+1}}$ converges towards $\Lambda(f)$ at rate

$$(n\alpha^2 \log^{-2} n)^{-\frac{p-s}{2p+2}} \vee n^{-\frac{p-s}{2p+1}}.$$

Remark: Derived matching lower bound for $\alpha = O(1)$.

Convergence rate for smooth functionals

Theorem [estimation smooth]: Let $f \in \mathcal{W}_p$ and $=\Lambda$ a smooth functional with $m \geq 0$ and $2m + 2 \leq p \leq d + 1$. Then the plug-in estimator $\widehat{\Lambda(f)} = \Lambda(\hat{f}_n)$ with $a > 0$ and

$$\left(n \wedge (n\alpha^2)\right)^{1/2p} \leq 2^{j_n} \leq \left[\log^{-2/(m+1)}(n\alpha^2)(n\alpha^2)^{1/(4m+4)}\right] \wedge \left[\log^{-1}(n)n^{1/(4m+3)}\right]$$

converges towards $\Lambda(f)$ at rate

$$n^{-1/2} \vee (n\alpha^2)^{-1/2}.$$

Remark: Derived matching lower bound for $\alpha = O(1)$.

Adaptation

Problem: The optimal threshold j_n depends on the regularity p of the density f. BUT this is unknown in practice. Hence we use a data driven choice for it $\hat{j}_n$ (e.g. modified version of Lepski's method).

Theorem [adaptation density]: The estimator $\hat{f}_n(x) = \hat{f}_n^{\hat{j}_n}(x)$ satisfies that for all $q+1 \leq p$ and $x \in [0,1]$

$$\sup_{f \in \mathcal{W}^p(L) \cap \|f\|_\infty \leq L} \mathbb{E}_{Q\mathbb{P}_f} \|\hat{f}_n^{(q)} - f^{(q)}\|_{L^2[0,1]} \vee \mathbb{E}_f |\hat{f}_n^{(q)}(x) - f(x)^{(q)}|$$

$$\lesssim (n\alpha^2 \log^{-3} n)^{-\frac{p-q}{2p+2}} \vee (n/\log n)^{-\frac{p-q}{2p+1}}.$$

Theorem [adaptation atomic functional]: Let $f \in \mathcal{W}_p$ be such that $\|f\|_\infty \leq L$ and suppose that the operator Λ is atomic for $m, s \geq 0$ and $p \geq \max(s+1, m+1, 2m-s)$. Then the plug in estimator $\Lambda(\hat{f}_n)$ with $\hat{f}_n = \hat{f}_n^{\hat{j}_n}$ satisfies that

$$\mathbb{E}_{Q\mathbb{P}_f} |\Lambda(\hat{f}_n) - \Lambda(f)| \lesssim (n\alpha^2 \log^{-3} n)^{-\frac{p-s}{2p+2}} \vee (n/\log n)^{-\frac{p-s}{2p+1}}.$$

Adaptation to smooth functional:

- The plug-in estimator $\Lambda(\hat{f}^{\hat{j}_n})$ doesn't work (too smooth)
- One can consider an rougher estimator $\hat{f}_n$ with threshold not depending on p.

Summary

- Privacy constrained inference is becoming increasingly popular, in particular differential privacy.
- Methods are typically case-by-case. New privacy constrained estimator requires new mechanism.
- We consider α-differential private plug-in estimators for functional estimation.
- Can be used for a wide range, including smooth and atomic functionals.
- Derived matching minimax lower bounds.
- Adaptive inference for atomic functionals (smooth functionals need over-fitting).

Communication Complexity of Exact Sampling under Rényi Information

Tamás Linder

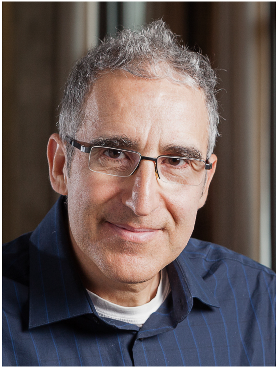
Queen's University, Canada

Hungarian Machine Learning Days

August 14, 2025

Co-authors

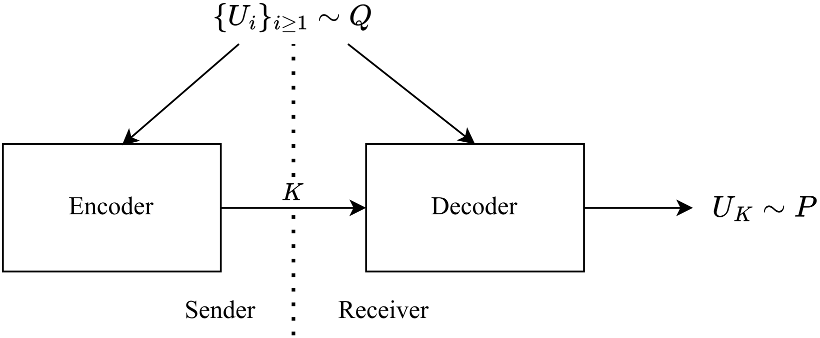
Spencer Hill



Fady Alajaji

Queen's University

Exact Sampling



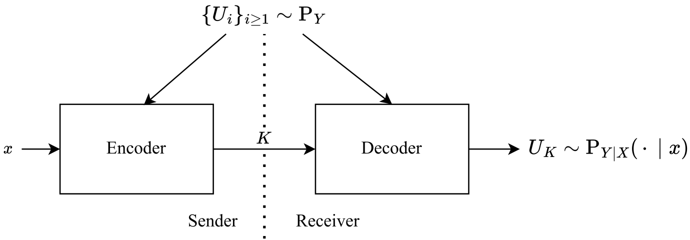
- When the goal is to efficiently communicate K, one can achieve

$$H(K) \approx D(P\|Q) \quad \text{bits}$$

- **Shannon:** K can be losslessly encoded at rate R such that

$$H(K) \leq R < H(K) + 1$$

Channel Simulation from Exact Sampling



X, Y random variables, choose $P = \mathrm{P}_{Y|X}(\,\cdot\mid x)$ and $Q = \mathrm{P}_Y$.

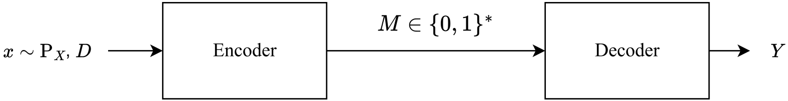
On input $x \sim \mathrm{P}_X$, sampling from P simulates the channel $X \to Y$.

Can simulate the channel with communication cost close to

$$H(K) \approx \mathbb{E}_X[D(\mathrm{P}_{Y|X}(\,\cdot\mid X)\,\|\,\mathrm{P}_Y)] = I(X;Y) \quad \text{bits}$$

Why Care?

Lossy Source Coding

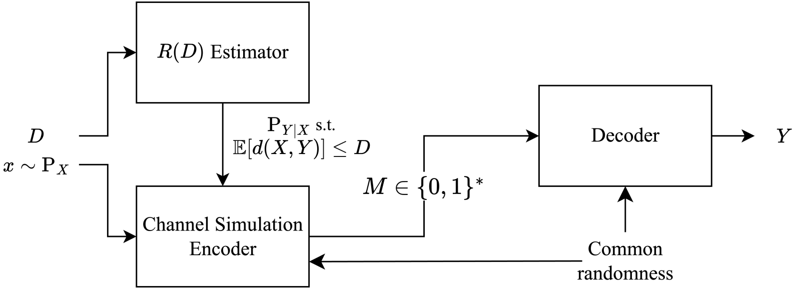


- The encoder encodes the block $(X_1, \ldots, X_n)$
- Decoder reconstructs $(Y_1, \ldots, Y_n)$
- Distortion: $D = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[d(X_i, Y_i)\big]$
- Rate: $R = \frac{1}{n} \mathbb{E}|M|$ (expected message length)
- Asymptotically $(n \to \infty)$ optimal performance

$$R(D) = \min_{\mathrm{P}_{Y|X} \,:\, \mathbb{E}[d(X,Y)] \leq D} I(X;Y).$$

Realization with Channel Simulation

$$R(D) = \min_{P_{Y|X} : \mathbb{E}[d(X,Y)] \le D} I(X;Y).$$



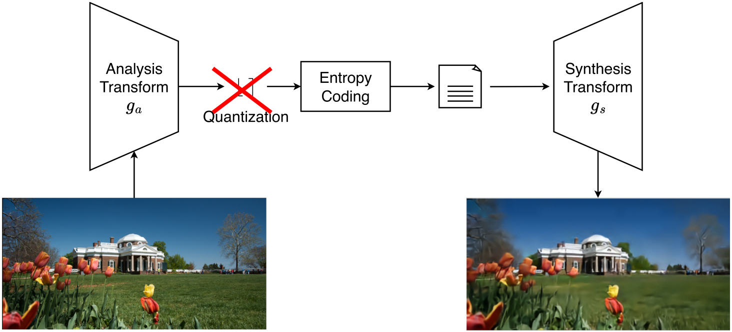
Recent work on neural-estimation of the rate-distortion function and $R(D)$-achieving conditional distribution Lei et al. (2023).

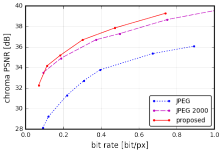
Channel simulation at cost $I(X;Y) \implies$ one-shot code achieving $R(D)$

Other Applications

- **Neural compression via nonlinear transform coding**
- Compression via implicit neural representation
- Rate-distortion-perception tradeoff
- Local differential privacy
- Federated learning, ...

Neural Compression via Nonlinear Transform Coding

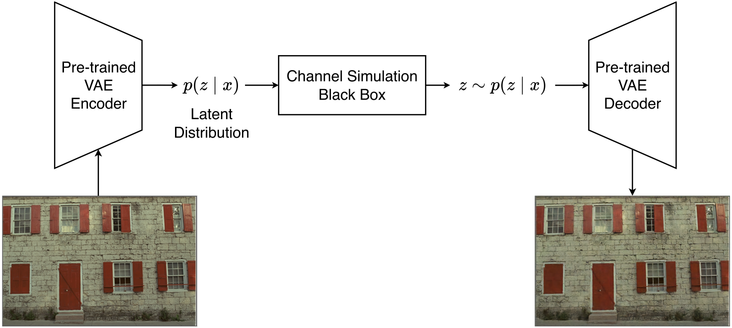




JPEG

Image credits Ballé et al. (2017).

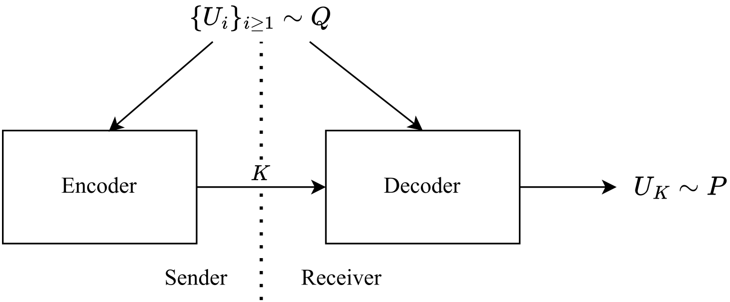
Neural Compression with Channel Simulation



- Fully differentiable end-to-end system!
- Channel simulation $\Longrightarrow$ **Relative entropy coding**

Image credits Flamich et al. (2020).

Exact Sampling

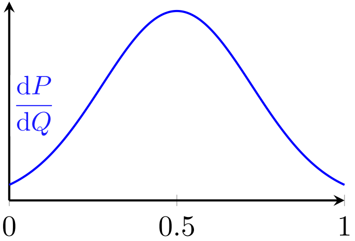


Key Questions:

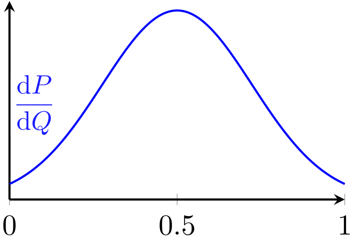
- How can we choose K such that $U_K \sim P$ *exactly*?
- How close can we get to $D(P||Q)$?

Greedy Rejection Sampling

- **Rejection sampling:** Accept U_k with probability $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_k)$, $\gamma > 0$ s.t. $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \leq 1$ for all u.
- **Greedy rejection sampling:** Accept U_k with probability $f_k(U_k)$, for function f_k which maximizes the acceptance probability at stage k under the condition that the scheme is exact.



Rejection sampling

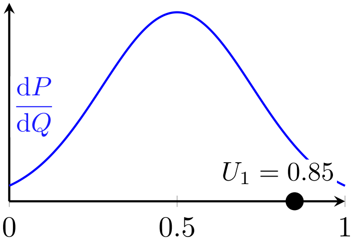


Greedy rejection sampling

$$P = \mathcal{N}(0.5, 0.05)|_{[0,1]},\ Q = \text{Uniform}([0,1]),\ \gamma = 0.55.$$

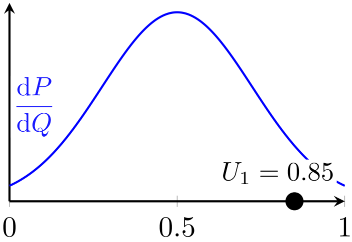
Greedy Rejection Sampling

- **Rejection sampling:** Accept U_k with probability $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_k)$, $\gamma > 0$ s.t. $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \leq 1$ for all u.
- **Greedy rejection sampling:** Accept U_k with probability $f_k(U_k)$, for function f_k which maximizes the acceptance probability at stage k under the condition that the scheme is exact.



Rejection sampling

$\mathbb{P}(\text{Accept}) = \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_1) = 0.275$

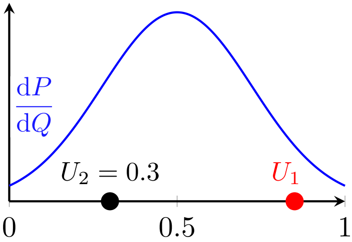


Greedy rejection sampling

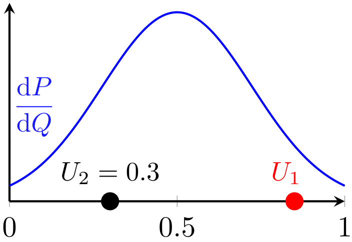
$\mathbb{P}(\text{Accept}) = \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(U_1) - 0\right)/1 = 0.5$

Greedy Rejection Sampling

- **Rejection sampling:** Accept U_k with probability $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_k)$, $\gamma > 0$ s.t. $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \leq 1$ for all u.
- **Greedy rejection sampling:** Accept U_k with probability $f_k(U_k)$, for function f_k which maximizes the acceptance probability at stage k under the condition that the scheme is exact.



Rejection sampling

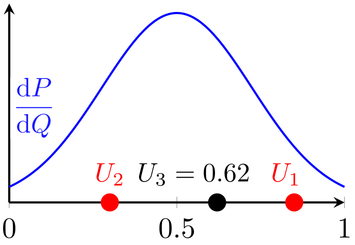


Greedy rejection sampling

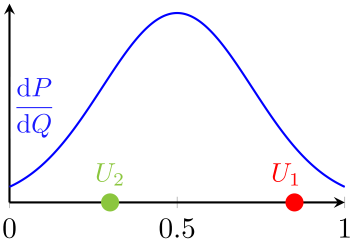
$\mathbb{P}(\text{Accept}) = \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_2) = 0.67$ $\quad$ $\mathbb{P}(\text{Accept}) = \frac{1}{0.255} \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(U_2) - 1 \right) = 0.89$

Greedy Rejection Sampling

- **Rejection sampling:** Accept U_k with probability $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_k)$, $\gamma > 0$ s.t. $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \leq 1$ for all u.
- **Greedy rejection sampling:** Accept U_k with probability $f_k(U_k)$, for function f_k which maximizes the acceptance probability at stage k under the condition that the scheme is exact.



Rejection sampling

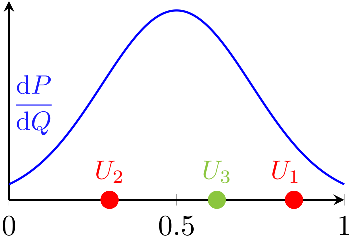


Greedy rejection sampling

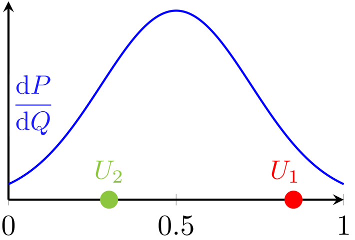
$\mathbb{P}(\text{Accept}) = \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_2) = 0.87$

Greedy Rejection Sampling

- **Rejection sampling:** Accept U_k with probability $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(U_k)$, $\gamma > 0$ s.t. $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(u) \leq 1$ for all u.
- **Greedy rejection sampling:** Accept U_k with probability $f_k(U_k)$, for function f_k which maximizes the acceptance probability at stage k under the condition that the scheme is exact.



Rejection sampling



Greedy rejection sampling

GRS: $D(P||Q) \leq \mathbb{E}[|M|] \leq D(P||Q) + \log_2(D(P||Q) + 1) + 4$

Recent Tighter Bounds

Recently, Goc and Flamich (2024) showed a tight bound on the expected message length:

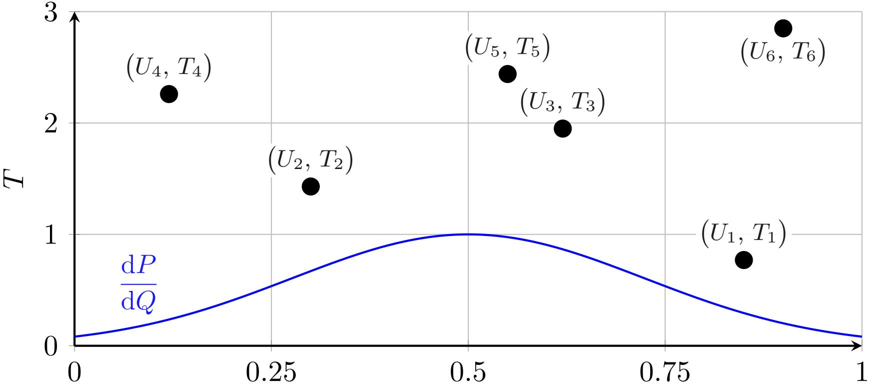
$$D(P||Q) \leq D_{CS}(P||Q) \leq \mathbb{E}[|M|] \leq D_{CS}(P||Q) + \log_2(e+1) + 1$$

for $D_{CS}(P||Q)$ the *channel simulation divergence.*

The upper bound on $\mathbb{E}[|M|]$ is achieved using greedy rejection sampling.

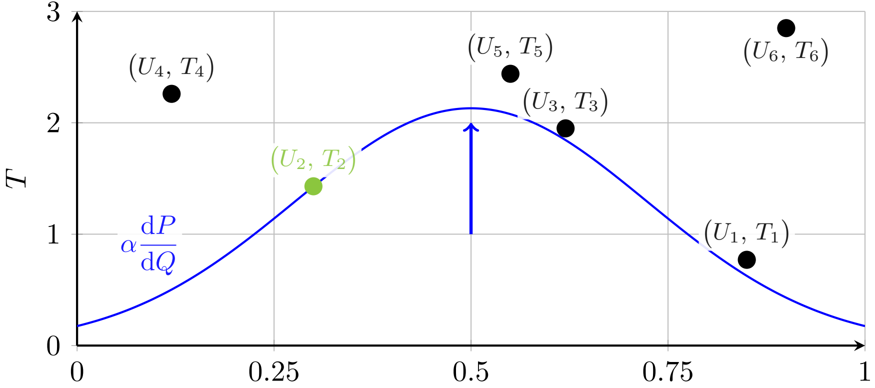
Poisson Functional Representation

For $\{T_i\}_{i\geq 1}$ a rate-one Poisson process, choose $K = \arg\min_{i\geq 1} \frac{T_i}{\frac{dP}{dQ}(U_i)}$, Li and El-Gamal (2018).



Poisson Functional Representation

For $\{T_i\}_{i\geq 1}$ a rate-one Poisson process, choose $K = \arg\min_{i\geq 1} \frac{T_i}{\frac{dP}{dQ}(U_i)}$, Li and El-Gamal (2018).



$$D(P||Q) \leq \mathbb{E}[|M|] \leq D(P||Q) + \log_2(D(P||Q) + 2) + 3$$

Our Setup: Exponential Cost and Rényi's entropy

- The previous results are for the expected message length (number of bits) $\mathbb{E}[|M|]$.
- What are the fundamental limits of exact sampling under a cost which is *exponential* in the message lengths? Can these limits be (almost) achieved by existing algorithms?

Campbell Cost $L(t)$

For uniquely decodable binary encoding $M \in \{0,1\}^*$ of K having length $|M|$ and for $t > 0$,

$$L(t) = \frac{1}{t} \log\Big(\mathbb{E}[2^{t|M|}]\Big).$$

Facts:

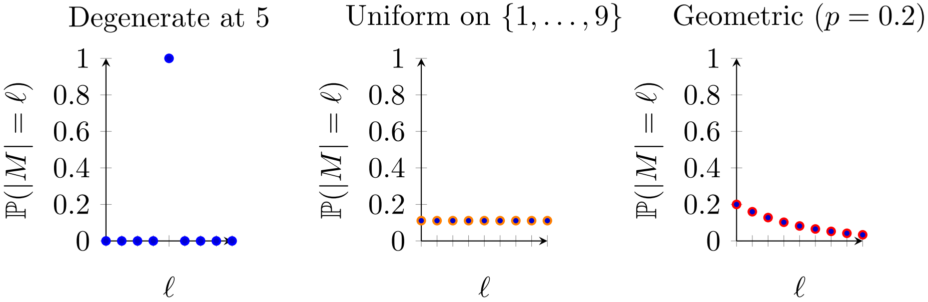
$$\lim_{t \to 0} L(t) = \mathbb{E}[|M|] \qquad \text{and} \qquad \lim_{t \to \infty} L(t) = \max_{\ell \in \mathbb{N} \,:\, \mathbb{P}(|M| = \ell) > 0} \ell$$

For a random variable K with Rényi entropy $H_\alpha(K)$ encoded optimally into message M, Campbell (1965) showed

$$H_\alpha(K) \leq L(t) < H_\alpha(K) + 1$$

with $\alpha = \frac{1}{1+t}$.

Why Care About $L(t)$?



t	**Degenerate** $L(t)$	**Uniform** $L(t)$	**Geometric** $L(t)$
0	5	5	5
0.2	5	5.65	11.83
1	5	7.26	∞
5	5	8.56	∞
∞	5	9	∞

Lower Bound

Theorem 1 For any sampling algorithm and $t > 0$, with $\alpha = \frac{1}{1+t}$,

$$L(t) \geq D_{\frac{1}{\alpha}}(P||Q) + \frac{\alpha}{1-\alpha}\log_2(\alpha) - 1. \tag{LB}$$

As $t \to 0$, we recover the lower bound

$$\mathbb{E}[|M|] \geq D(P||Q) - \frac{1}{\ln(2)} - 1.$$

Upper Bounds via Poisson Functional Representation

Theorem 2 For K chosen using the Poisson functional representation, for any $\epsilon > 0$ there exists a uniquely decodable encoding of K such that

$$L(t) \leq (1+\epsilon) D_{\frac{1+\epsilon(1-\alpha)}{\alpha}}(P||Q) + c(\alpha, \epsilon), \qquad (\mathrm{UB}_1)$$

with $c(\alpha, \epsilon)$ a constant and $\alpha = \frac{1}{1+t}$.

Upper Bounds via Poisson Functional Representation

Theorem 3 Encoding K (generated by the PFR) using the Elias omega code gives, for any $0 < t < 1/2$ and $\epsilon \leq \frac{1}{2t} - 1$,

$$L(t) \leq D_{\frac{2-\alpha}{\alpha}}(P||Q) + (1+\epsilon)\log_2(D(P||Q)+1) + c_\epsilon. \qquad (\mathrm{UB}_2)$$

Recovers the bound

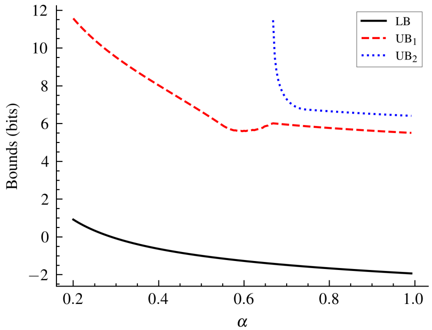
$$\mathbb{E}[|M|] \leq D(P||Q) + (1+\epsilon)\log_2(D(P||Q)+1) + c_\epsilon$$

of Harsha et al. (2010) as $t \to 0$.

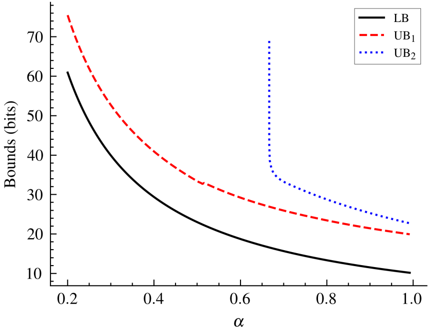
Proof Techniques

- For $\mathbb{E}[|M|]$ and $L(t)$ the lower bounds $D(P\|Q)$ resp. $D_{\frac{1}{\alpha}}(P\|Q)$ are simple to prove.
- The upper bound(s) on $\mathbb{E}[|M|]$ are derived through *sharply* bounding $\mathbb{E}[\,\log_2 K\,]$.
- The upper bounds on $L(t)$ are derived (more or less) by bounding $\mathbb{E}[K^t]$.

Gaussian Examples



$P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(1, 1)$



$P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(5, 1)$

Asymptotic Results

- We want to use the channel n-times with i.i.d. input $X_1, \ldots, X_n$. Thus we sample from the product distribution $P^{\otimes n}$ using samples from $Q^{\otimes n}$.
- We can now fully characterize the optimal $L(t)/n$ as $n \to \infty$:

Theorem 4 For any $t > 0$, let $L_n^*(t)$ be the *minimum Campbell cost* for target $P^{\otimes n}$ and common randomness $\{U_i\}_{i \geq 1} \sim Q^{\otimes n}$. Then, with $\alpha = \frac{1}{1+t}$,

$$\lim_{n \to \infty} \frac{L_n^*(t)}{n} = D_{\frac{1}{\alpha}}(P||Q).$$

This generalizes known results: for the *minimum bits/sample* rate R_n^* for the n-dimensional product distributions,

$$\lim_{n \to \infty} \frac{R_n^*}{n} = D(P||Q).$$

Causal vs. Noncausal Sampling

- A *causal* sampler accepts/rejects each candidate one-at-a-time (K is a stopping time w.r.t. $\{U_i\}_{i\geq 1}$).
- Greedy rejection sampling ✓ Poisson functional representation X
- GRS and the PFR **both achieve** bits/sample rate $D(P||Q)$ as $n \to \infty$.

Theorem 5 For any $t > 0$ let $L_n^*(t)$ be the minimum Campbell cost over *causal* samplers between $P^{\otimes n}$ and $Q^{\otimes n}$. Then, with $\alpha = \frac{1}{1+t}$,

$$\liminf_{n\to\infty} \frac{L_n^*(t)}{n} \geq D_\beta(P||Q), \quad \text{where } \beta = \begin{cases} \frac{\alpha}{2\alpha-1}, & \alpha \in (1/2, 1) \\ \infty, & \alpha \in (0, 1/2]. \end{cases}$$

- $D_\beta(P||Q) > D_{\frac{1}{\alpha}}(P||Q)$ in general!!!
- Greedy rejection sampling does **strictly worse** in the **exponential cost regime**, and the gap is often significant.

Main Takeaways

- Exact sampling is one (highly general) way to perform channel simulation at a near-optimal encoding cost, and has wide applications.
- The Campbell cost $L(t)$ generalizes the expected message length and can be made more sensitive to the tails of the distribution.
- The Poisson functional representation is nearly optimal for exact sampling (typically within 5-10 bits) for the Campbell cost.
- Causal samplers (such as greedy rejection sampling, greedy Poisson rejection sampling, etc.) do **strictly worse** than noncausal samplers in the asymptotic Campbell cost.

References

J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *5th International Conference on Learning Representations*, 2017.

L. L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, no. 4, pp. 423–429, 1965.

G. Flamich, M. Havasi, and J. M. Hernández-Lobato, "Compressing images by encoding their latent representations with relative entropy coding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 131–16 141, 2020.

P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 438–449, 2010.

E. Lei, H. Hassani, S. S. Bidokhti, "Neural estimation of the rate-distortion function with applications to operational source coding," *IEEE Journal on Selected Areas in Information Theory*, no. 4, pp. 674-686, 2023.

D. Goc and G. Flamich, "On channel simulation with causal rejection samplers," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 1682–1687.

C. T. Li and A. El-Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2583–2592, 2018.

References

G. Flamich, S. M. Sriramu, and A. B. Wagner, "The redundancy of non-singular channel simulation," *arXiv preprint arXiv:2501.14053*, 2025.

C. T. Li, *Channel Simulation: Theory and Applications to Lossy Compression and Differential Privacy.* Now Publishers, Inc., 2024, vol. 21, no. 6.

G. Flamich, "Greedy Poisson rejection sampling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37 089–37 127, 2023.

J. Liu and S. Verdú, "Rejection sampling and noncausal sampling under moment constraints," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1565–1569.

G. Flamich and L. Theis, "Adaptive greedy rejection sampling," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 454–459.

C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the Poisson matching lemma," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.

S. Hill, F. Alajaji, and T. Linder, "Communication complexity of exact sampling under Rényi information," *arXiv preprint arXiv:2506.12219*, 2025.

Quantile Preferences in Portfolio Choice: A Q-DRL Approach to Dynamic Diversification

Attila Sarkany[1,2] Lukáš Janásek[1,2] Jozef Baruník[1,2]

[1]Institute of Economic Studies, Charles University, Opletalova 26, 110 00 Prague, Czech Republic
[2]The Czech Academy of Sciences, IITA, Pod Vodárenskou Věží 4, 182 08 Prague, Czech Republic

August 13, 2025

Introduction

Motivation:

- Rigid utility function and return assumptions
- How to dynamically map heterogeneous preferences into portfolio actions?
- Escaping from the Gaussian world. We need more information then just mean and variance (heavy tails)
- Risk and Intertemporal substitution captured by the same parameter

Literature

Economics

- Quantile Preferences: Castro and Galvao 2019
- Quantile maximization in decision theory: Rostek 2010

Reinforcement Learning

- Jiang et al. (2022): Quantile-based policy optimization for reinforcement learning
- Ma et al. (2020): Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning
- Bellemare (2017):A Distributional Perspective on Reinforcement Learning

Quantiles

$$Q_\tau[X] \equiv F_X^{-1}(\tau) \quad (1)$$

where F_X is the c.d.f of the random variable X.

Properties

- Invariance to monotonic transformation: $Q_\tau[g(X)] = g\left(Q_\tau[X]\right)$
- Stochastic domincance : $Q_\tau[X] \geq Q_\tau[Y]$
- Law of iterated expectations is not valid anymore:
- $Q_\tau[\,Q_\tau(X \mid \Sigma_1) \,|\, \Sigma_0] \neq Q_\tau(X \mid \Sigma_0)$
- $\frac{\partial}{\partial x} Q_\tau[h(x, Z)] \neq Q_\tau\left[\frac{\partial h}{\partial x}(x, Z)\right]$
- Not linear: $Q_\tau[X + Y] \neq Q_\tau[X] + Q_\tau[Y]$

Quantile Preferences

$$X \succeq Y \iff Q_\tau[X] \geq Q_\tau[Y] \tag{2}$$

where $Q_\tau[X]$ denotes the quantile of a random variable X at fraction $\tau \in [0, 1]$.

Comparison with Expected Utility:

$$X \succeq Y \iff E[u(X)] \geq E[u(Y)] \tag{3}$$

Unlike expected utility, quantile preferences do not rely on a utility function u, as:

$$\begin{aligned} X \succeq Y &\iff Q_\tau[X] \geq Q_\tau[Y] \iff u(Q_\tau[X]) \geq u(Q_\tau[Y]) \\ &\iff Q_\tau[u(X)] \geq Q_\tau[u(Y)] \end{aligned} \tag{4}$$

Consequences

- In expected utility theory, risk aversion is determined by the concavity of $u(X)$.
- In contrast, quantile preferences directly associate risk attitudes with τ.
  - $\tau \ll 0.5$ (low quantiles) capture risk aversion by focusing on the lower tail.
  - $\tau \gg 0.5$ (high quantiles) reflect risk-seeking behavior by emphasizing the upper tail.
- Eliminates the need for explicit utility functions, enhancing interpretability.
- Allows for dynamic, non-stationary modeling using neural networks.

Dynamic Programming (DP)

- DP refers to a set of algorithms for computing optimal policies in Markov Decision Processes (MDPs)
- Assumes a perfect model of the environment: Transition probabilities

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \,\middle|\, S_t = s \right]$$

$$= \mathbb{E}_\pi \left[R_{t+1} + \gamma \, v_\pi(S_{t+1}) \,\middle|\, S_t = s \right]$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \, v_\pi(s') \right]$$

Optimal Policy

$$V^*(s) = \max_{\pi} V_{\pi}(s), \quad \forall s \in S.$$

and

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} p(s'|s,a) \left[r(s,a) + \gamma V^*(s') \right], \quad \forall s \in S.$$

How to find the solution?

- Sample based methods: Trial and error: Monte Carlo, Temporal Difference
- Dynamic Programming

Galvao (2019). 'Dynamic quantile models of rational behavior

Naive

$$V_t(h, x, z^t) = Q_\tau \left[\sum_{s \geq t} \beta^{s-t} u(x_s^h, x_{s+1}^h, Z_s) \mid Z^t = z^t \right]. \quad (5)$$

Fixed Policy- Quantile Preference

$$V_t(h, x, z^t) = u(x_t^h, x_{t+1}^h, z_t) + \beta Q_\tau \left[V_{t+1}(h, x, Z_t, z_{t+1}) \mid Z_t = z_t \right]. \quad (6)$$

Recursive substitution

$$V_1^{Q_\tau}(h, x, z^t) = Q_\tau^\infty \left[\sum_{t=1}^{\infty} \beta^{t-1} u(x_t^h, x_{t+1}^h, z_t) \right], \quad (7)$$

Neural Network Set up

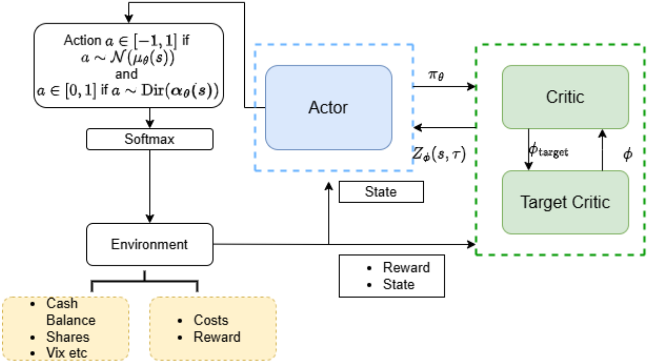


Figure: Q-A2C portfolio trading framework

Objective functions

TD error

$$\delta_i = R(s_t, s_{t+1}, a_t) + \gamma Q^{\pi}_{\tau_i}(s_{t+1}, \phi) - Q^{\pi}_{\tau_i}(s_t, \phi) \tag{8}$$

Critic Loss

$$\text{loss}_{\tau_i} = |\tau_i - \mathbb{I}\{\delta_i < 0\}| \times |\delta_i| \tag{9}$$

$$\mathcal{L}(\phi) = \frac{1}{|B|} \left(\sum_{b=1}^{B} \sum_{i \in T} \text{loss}_{\tau_i} + \lambda \cdot \text{quantile ordering penalty} \right) \tag{10}$$

Actor

$$\mathcal{L}_{\text{actor}}(\theta) = -\frac{1}{B} \sum_{b=1}^{B} (\sum_{k=1}^{N} [\log \pi_\theta(a_k \mid s) \cdot \delta_{\tau_i} \cdot w_{\tau_i}] - \lambda \cdot \mathcal{H}(\pi_\theta(s))) \tag{11}$$

where

$$w_{\tau_i} = \begin{cases} 1 - \tau_i, & \text{if } \delta_{\tau_i} < 0 \\ \tau_i, & \text{otherwise} \end{cases} \tag{12}$$

Interpretation

Bellman:Critic Role

- The critic learns the reward distribution under the **current policy** by approximating multiple quantiles.
- it only provides feedback (bellman) based on the policy being followed.

Actor Role:

- The actor defines the policy, selecting actions based on feedback from a specific quantile of the reward distribution.
- It will map situations (states) to actions stochastically

Comparing multiple policies

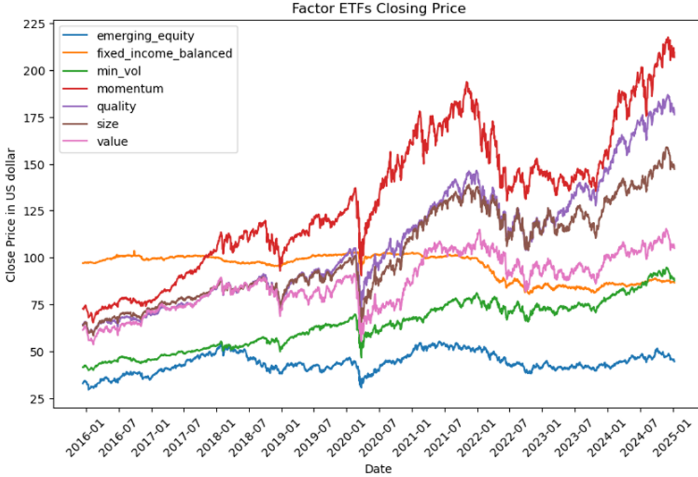
- We initialize τ many actor-critic network with identical weights.
- **After the first update:**
  - Each network starts specializing in different risk preferences.
  - The actor is updated using feedback from its assigned quantile.

Risk averse vs Risk seeking agent

Table: Pinball-Weighted Actor Loss for $\tau = 0.1$ and 0.9

τ	**TD**	**Weight**	**Behavior**
0.1	$e < 0$	$1 - \tau = 0.9$	Avoid risky actions
0.1	$e \geq 0$	$\tau = 0.1$	Weak reinforcement of successful actions
0.9	$e < 0$	$1 - \tau = 0.1$	Risk tolerance
0.9	$e \geq 0$	$\tau = 0.9$	Strongly reinforces risky actions

ETFs



Factor ETFs Closing Price
emerging_equity
fixed_income_balanced
min_vol
momentum
quality
size
value
Close Price in US dollar
225
200
175
150
125
100
75
50
25
Date
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
2024-07
2025-01

Problem

- Less then 1500 training points
- Extreme high correlation and Market regime changes

Results

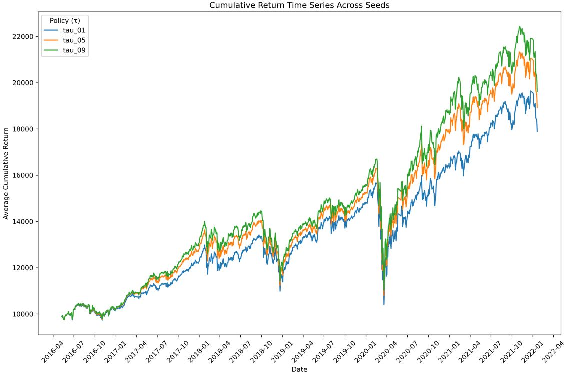


Figure: Average cumulative returns over seeds per policy

Distribution of rewards

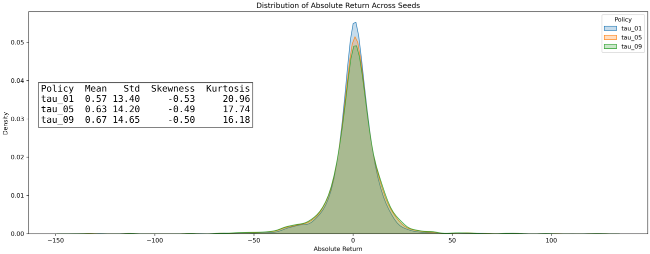


Figure: Distribution of rewards on training

Portfolio output 1

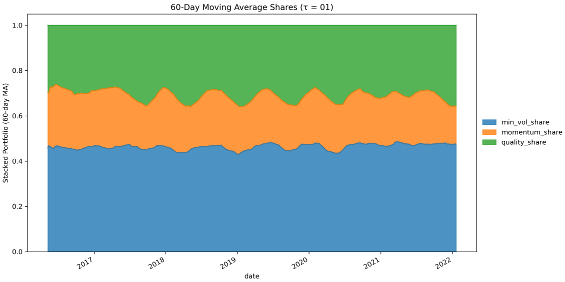
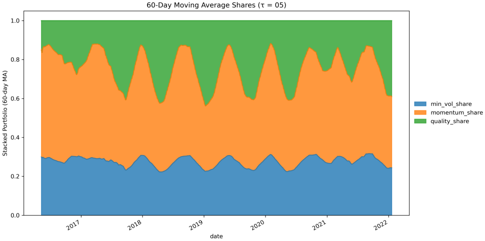


Figure: Risk-averse



Portfolio output 2

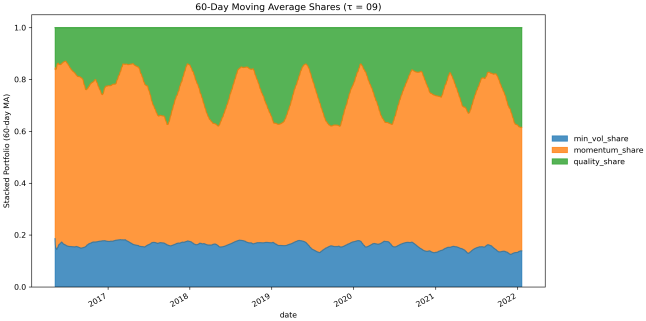
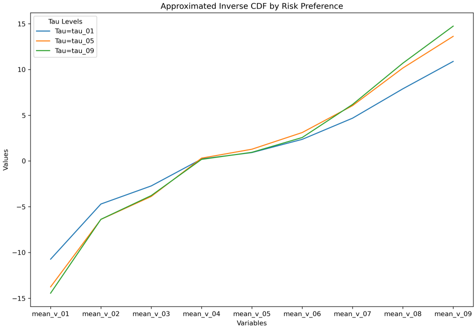


Figure: Risk-Seeking

Critic distributional output



Identifiable Exchangeable Mechanisms for Causal Structure and Representation Learning

Patrik Reizinger*, Siyuan Guo*, Ferenc Huszár, Bernhard Schölkopf, Wieland Brendel
patrik.reizinger@tuebingen.mpg.de

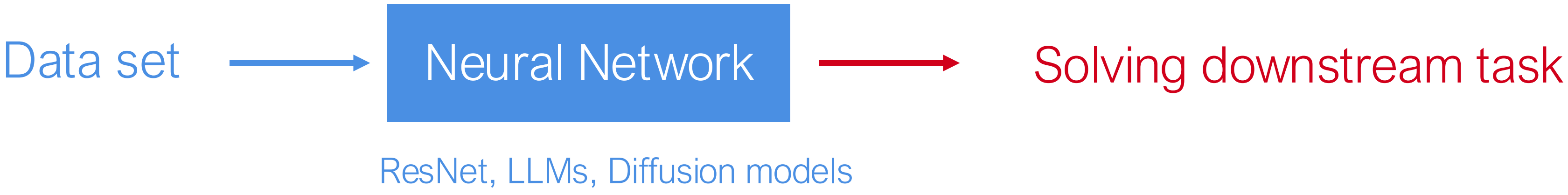
UNIVERSITY OF CAMBRIDGE ellis European Laboratory for Learning and Intelligent Systems imprs-is MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Why do we want to learn causal representations?

Data set → Neural Network

ResNet, LLMs, Diffusion models

Why do we want to learn causal representations?



Why do we want to learn causal representations?

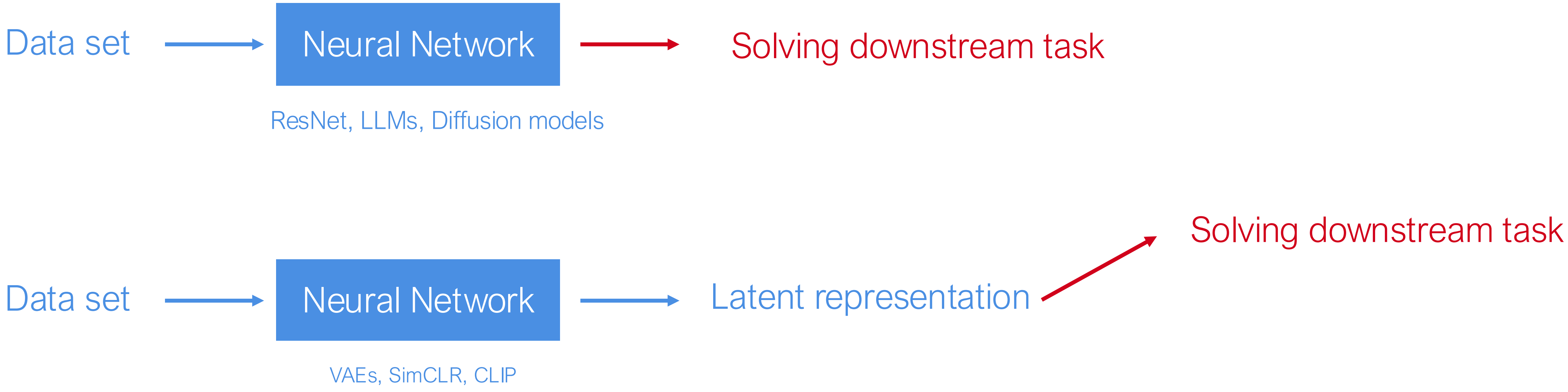
Data set → Neural Network → Solving downstream task

ResNet, LLMs, Diffusion models

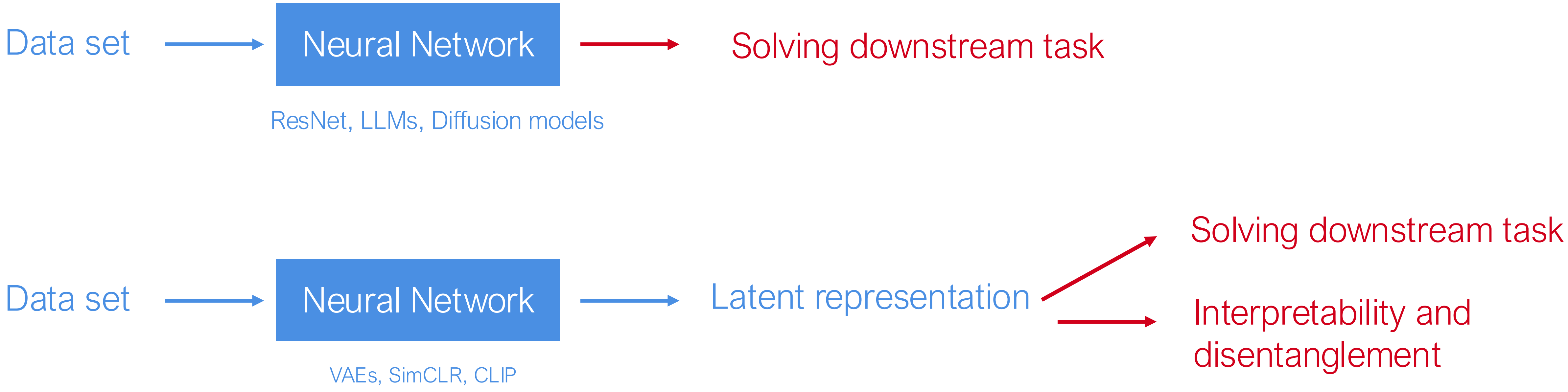
Data set → Neural Network → Latent representation

VAEs, SimCLR, CLIP

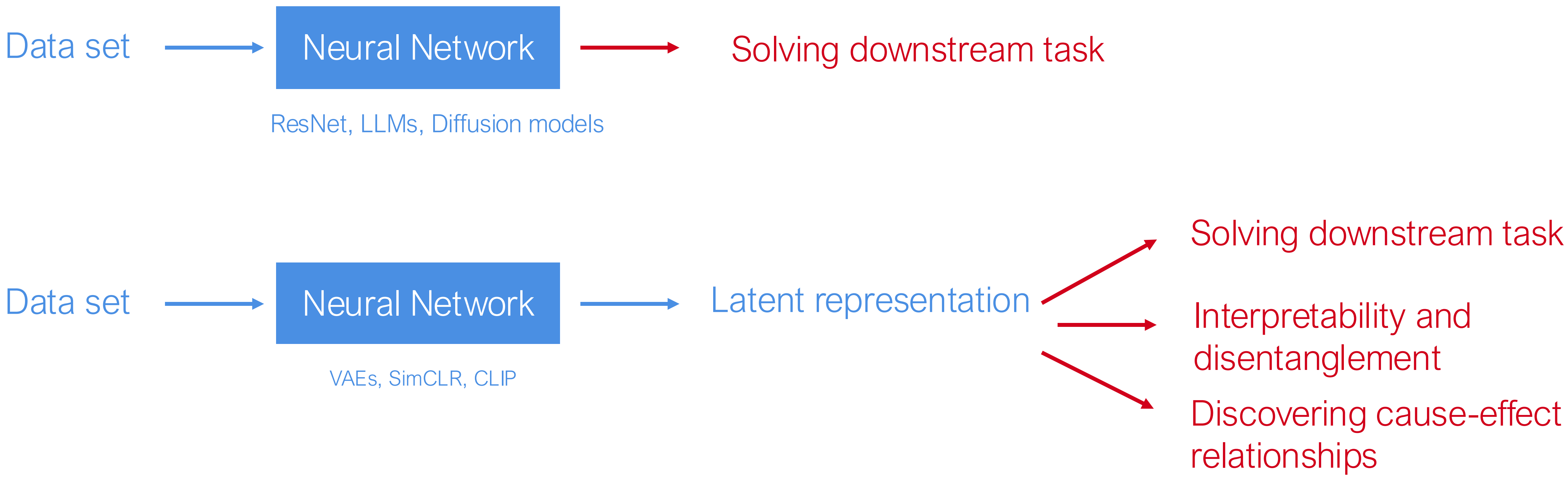
Why do we want to learn causal representations?



Why do we want to learn causal representations?

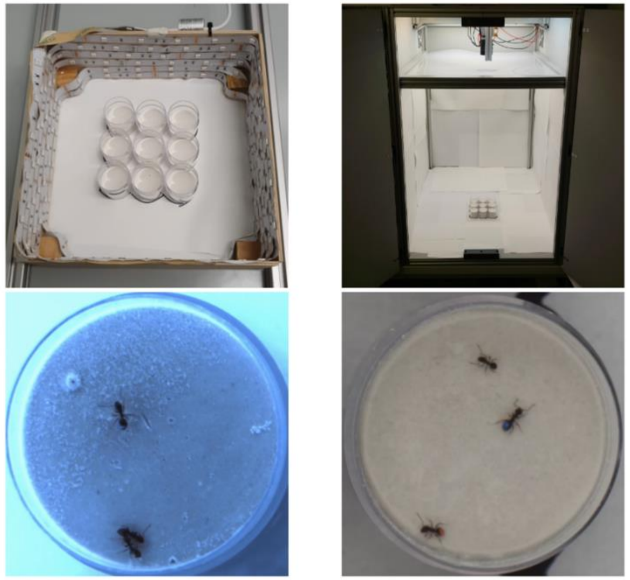


Why do we want to learn causal representations?

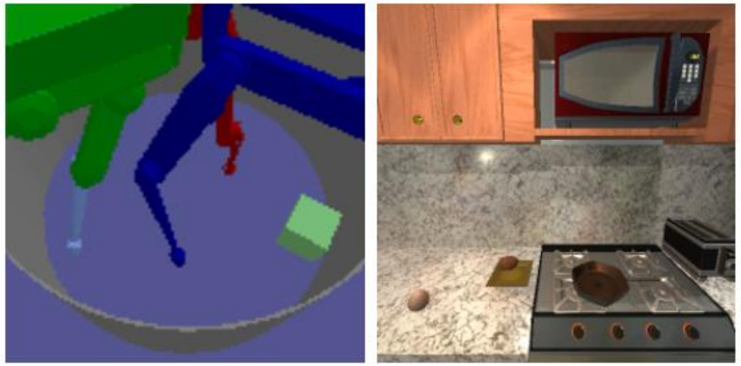


CRL is applied in many domains

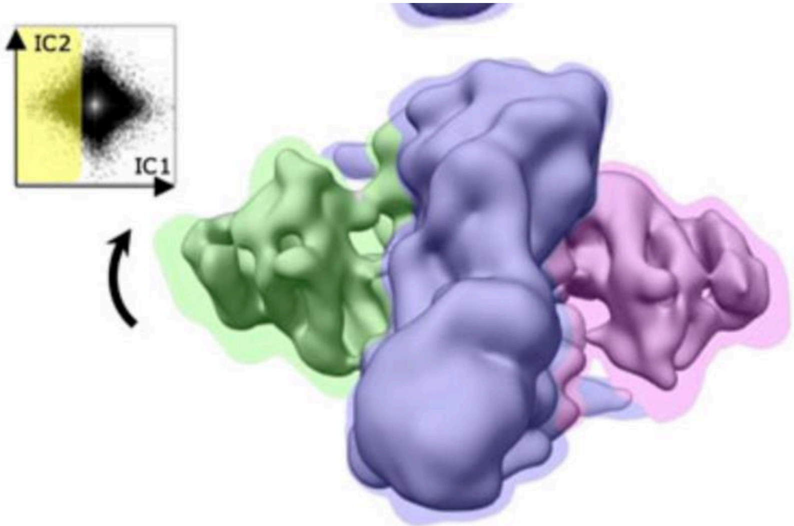
Ecology



Robotics



Electron microscopy



Cadei et al.: Causal Lifting of Neural Representations: Zero-Shot Generalization for Causal Inferences. arXiv, 2025.
Lippe et al.: BISCUIT. arXiv, 2023
.Klindt et al.:Towards interpretable Cryo-EM: disentangling latent spaces of molecular conformations. Frontiers Mol. Biosci., 2024

Identifiable representation learning

Pearl’s Causal Ladder

The rungs of the ladder

Examples

Pearl: Causality. 2000.

Pearl's Causal Ladder

The rungs of the ladder

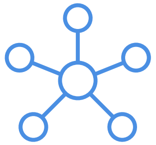
- Rung 1: $p_{\theta_1}(O)$ (observations)

Examples

Pearl: Causality. 2000.

Pearl's Causal Ladder

The rungs of the ladder

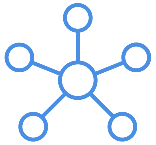
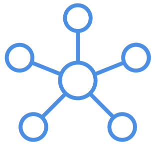
- Rung 1: $p_{\theta_1}(O)$ (observations)
- Rung 2: $p_{\theta_1}(O)$ +  (interventions)

Examples

Pearl: Causality. 2000.

Pearl's Causal Ladder

The rungs of the ladder

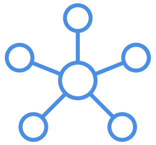
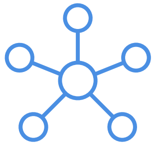
- Rung 1: $p_{\theta_1}(O)$ (observations)
- Rung 2: $p_{\theta_1}(O)$ +  (interventions)
- Rung 3: $p_{\theta_1}(O)$ +  + SEM (counterfactuals)

Examples

Pearl: Causality. 2000.

Pearl's Causal Ladder

The rungs of the ladder

- Rung 1: $p_{\theta_1}(O)$ (observations)
- Rung 2: $p_{\theta_1}(O)$ +  (interventions)
- Rung 3: $p_{\theta_1}(O)$ +  + SEM (counterfactuals)

Examples

- Rung 1: modeling chest X-Rays
- Rung 2: Smoking → Cancer
- Rung 3: How does $p(cancer)$ change when someone smokes 3 cigarettes less each day?

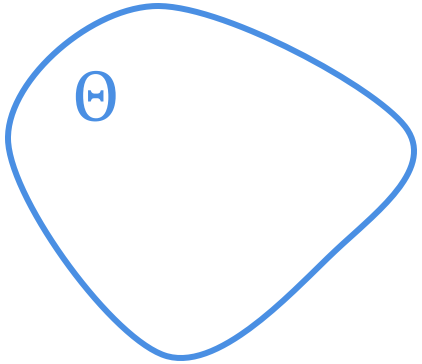
Pearl: Causality. 2000.

Identifiability

Identifiability

…and its problems

Equivalence class



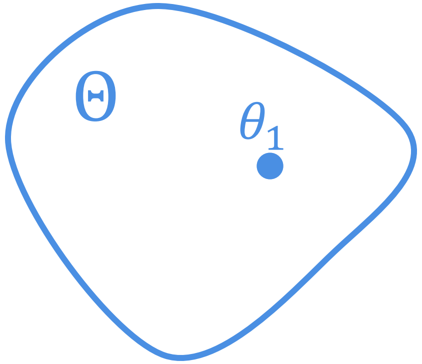
Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

Identifiability

Identifiability

…and its problems

Equivalence class



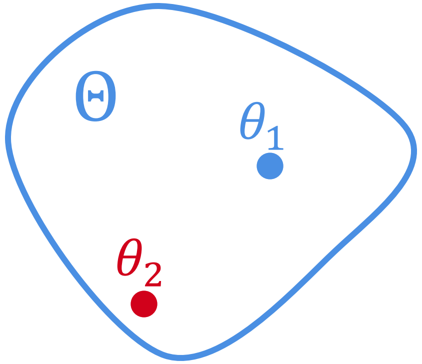
Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

Identifiability

Identifiability

…and its problems

Equivalence class



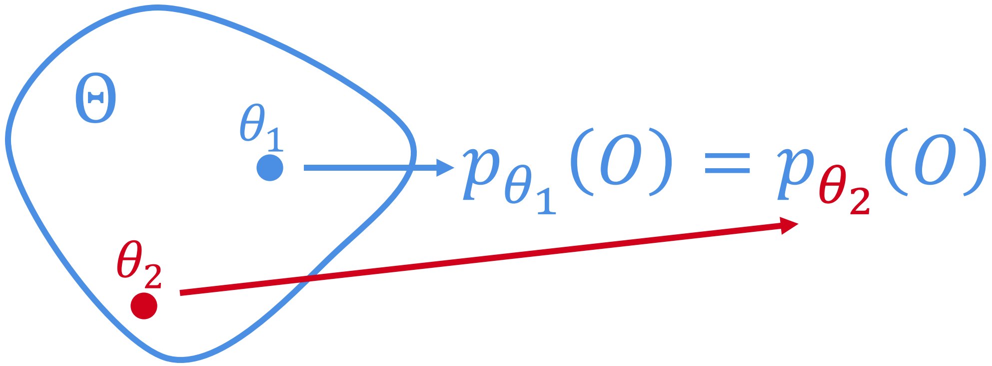
Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

Identifiability

Identifiability

…and its problems

Equivalence class



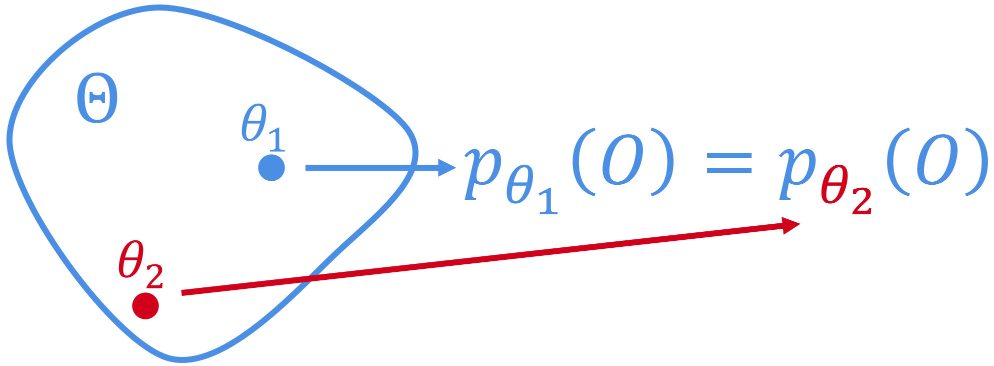
Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

Identifiability

Identifiability

…and its problems

Equivalence class



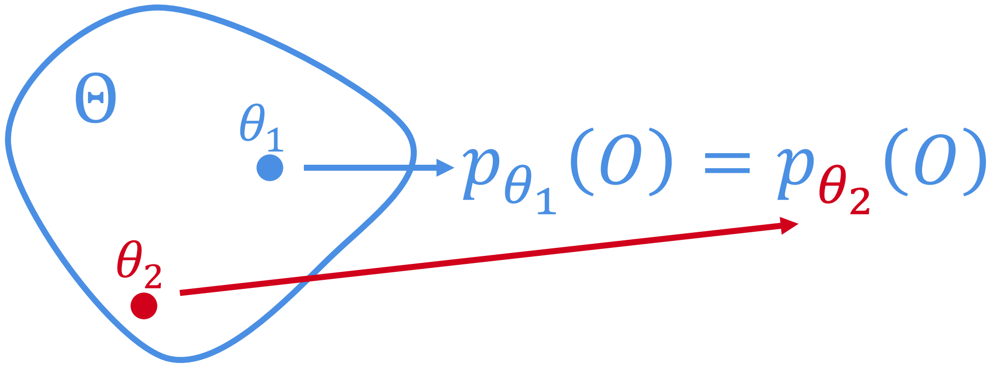
Identifiability = Data + Model constraints

Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

Identifiability

Identifiability

Equivalence class



Identifiability = Data + Model constraints

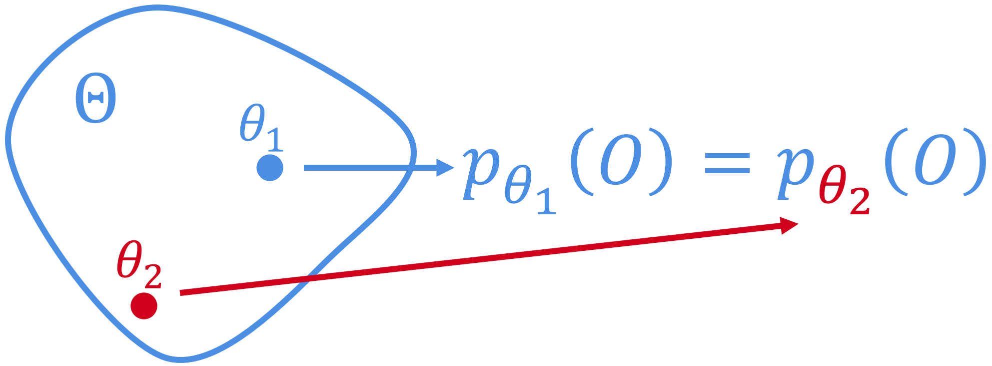
…and its problems

Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

Identifiability

Identifiability

Equivalence class



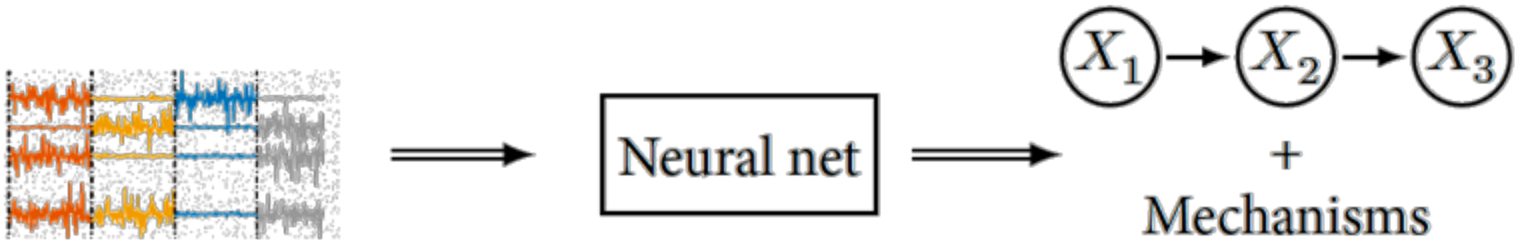
Identifiability = Data + Model constraints

…and its problems

- What are the invariances/symmetries of the distribution?
- Are the latent factors disentangled?
- Which coordinate encodes position/shape?

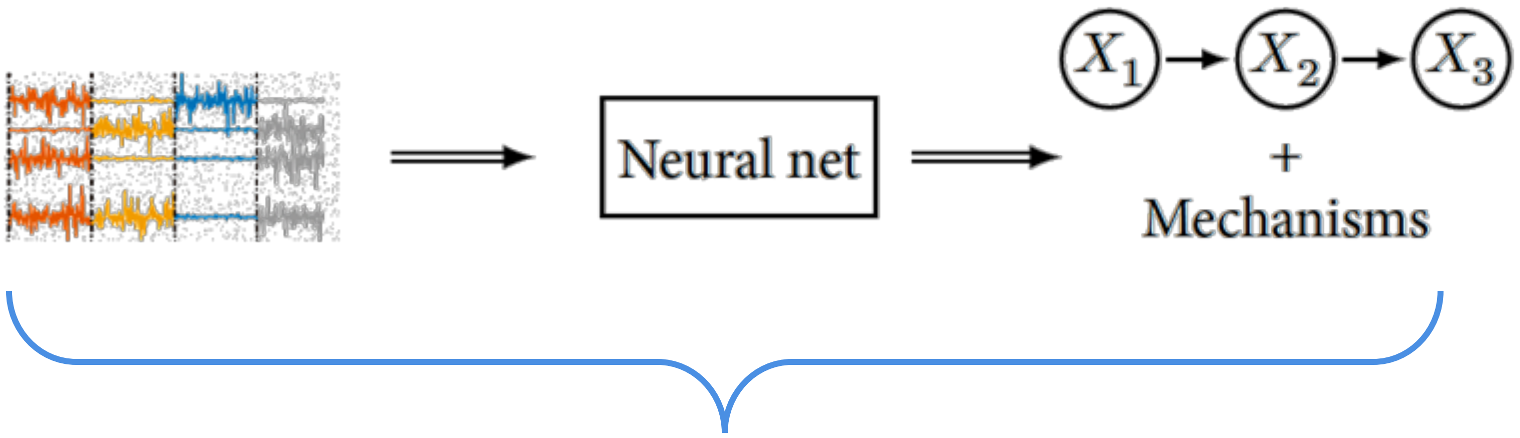
Hyvärinen: Independent Component Analysis. 2000.
Pearl: Causality. 2000.
Xi and Bloem-Reddy: Indeterminacy and Strong Identifiability in Generative Models. AISTATS, 2023.

The active vs passive view



Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

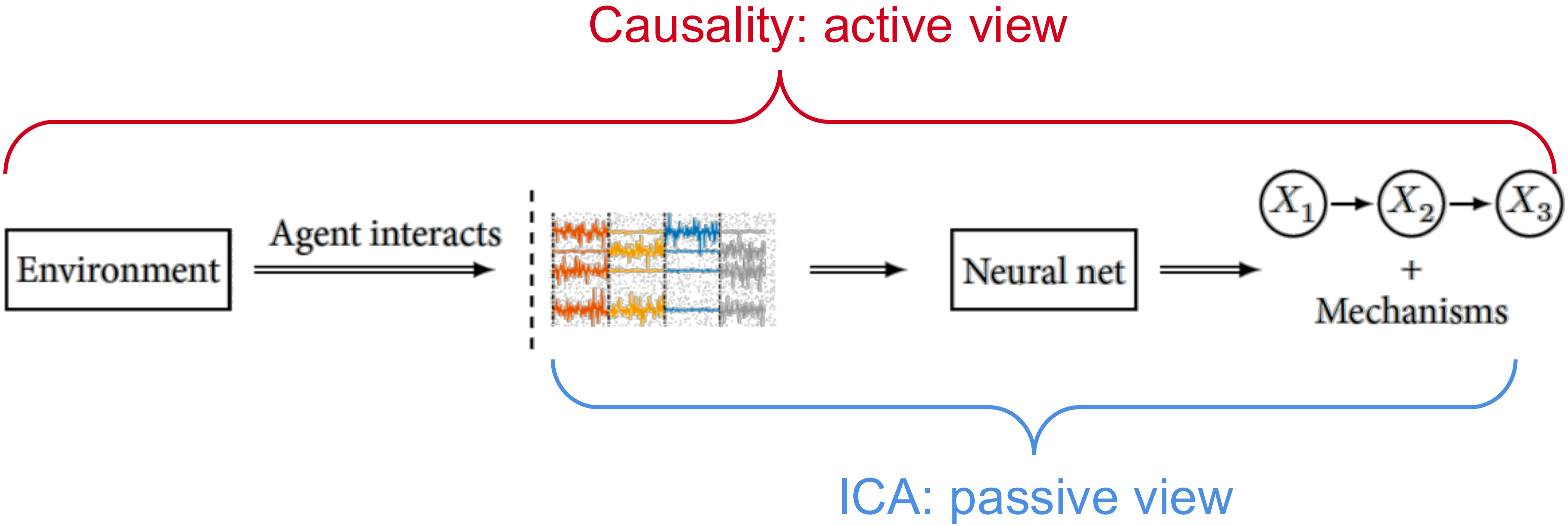
The active vs passive view



ICA: passive view

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

The active vs passive view



Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

Principles of (causal) representation learning

What makes identifiability possible?

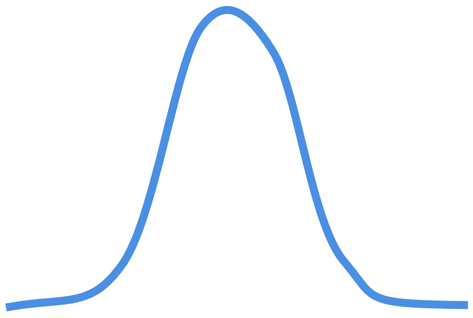
ICA: variability

Causality: interventions

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability

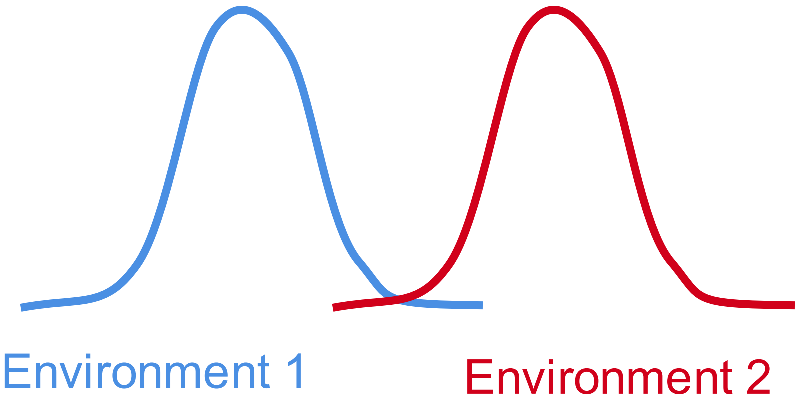


Causality: interventions

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability

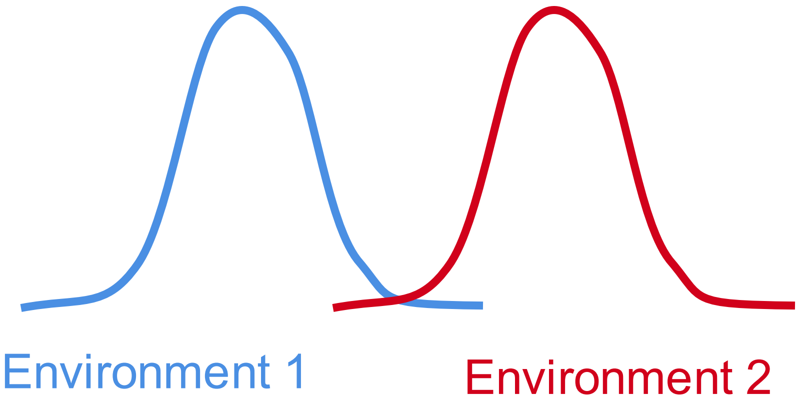


Causality: interventions

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability



Multi-environment

Blue- vs White-collar

Natural experiment

Different tax in neighboring cities

Multi-view

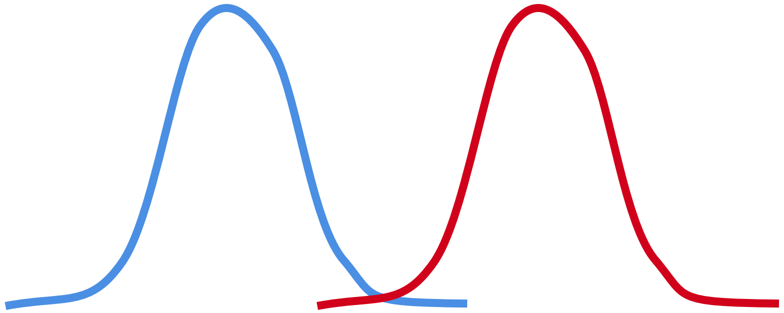
X-Ray + CO_2 retention

Causality: interventions

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability



Environment 1

Environment 2

Multi-environment

Blue- vs White-collar

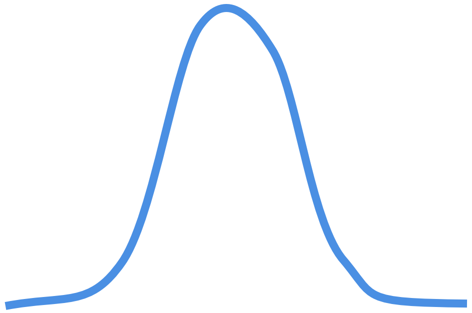
Natural experiment

Different tax in neighboring cities

Multi-view

X-Ray + CO_2 retention

Causality: interventions

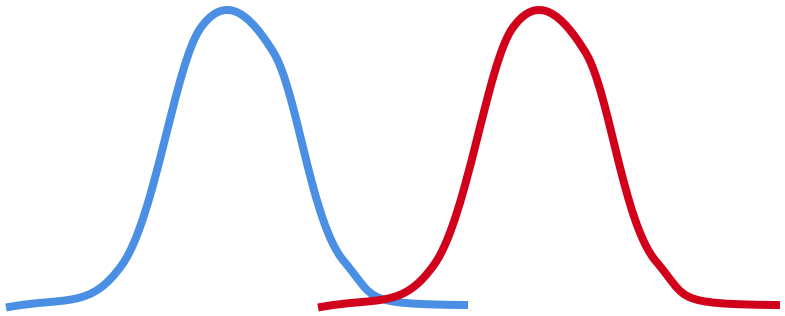


Observational

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability



Environment 1 | Environment 2

Multi-environment

Blue- vs White-collar

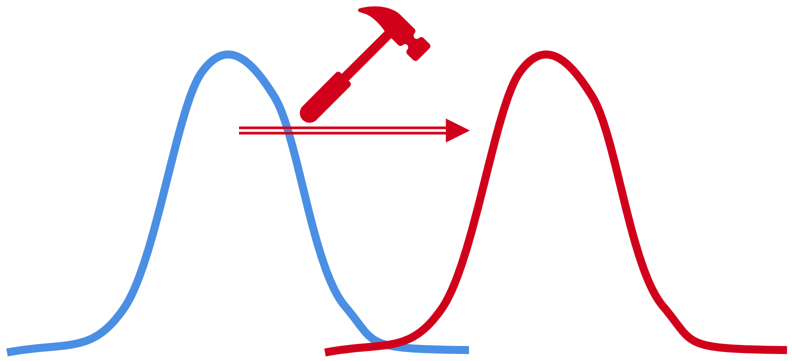
Natural experiment

Different tax in neighboring cities

Multi-view

X-Ray + CO_2 retention

Causality: interventions

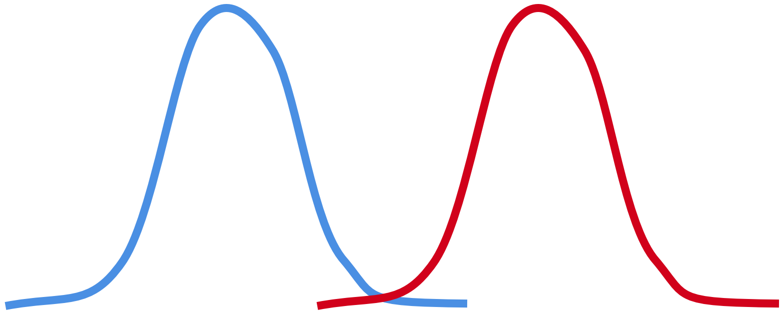


Observational | Interventional

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability



Environment 1

Environment 2

Multi-environment

Blue- vs White-collar

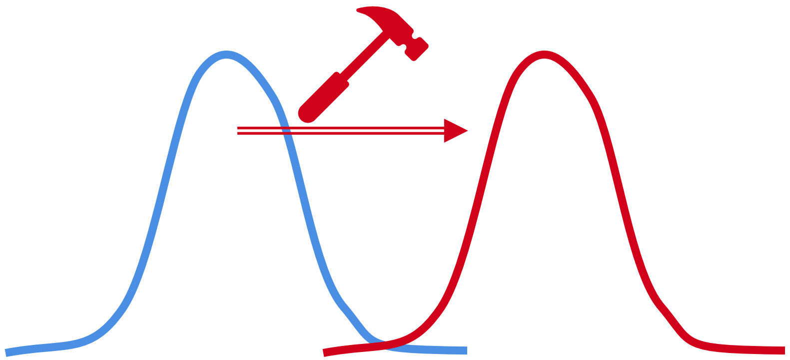
Natural experiment

Different tax in neighboring cities

Multi-view

X-Ray + CO_2 retention

Causality: interventions



Observational

Interventional

Imperfect (soft)

Educational campaign, tax increase

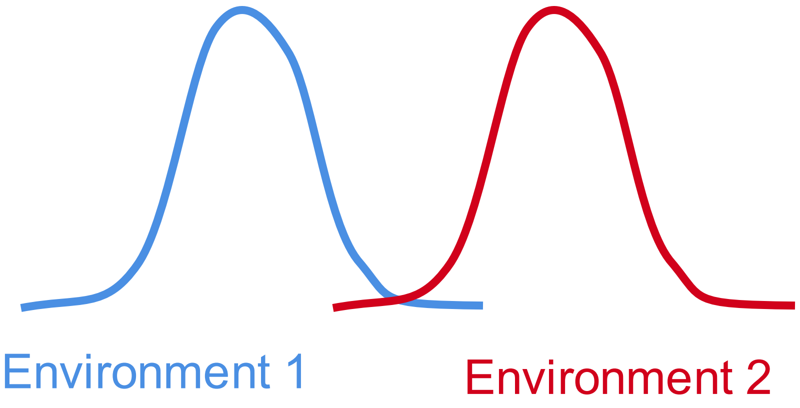
Perfect (hard)

Smoking ban

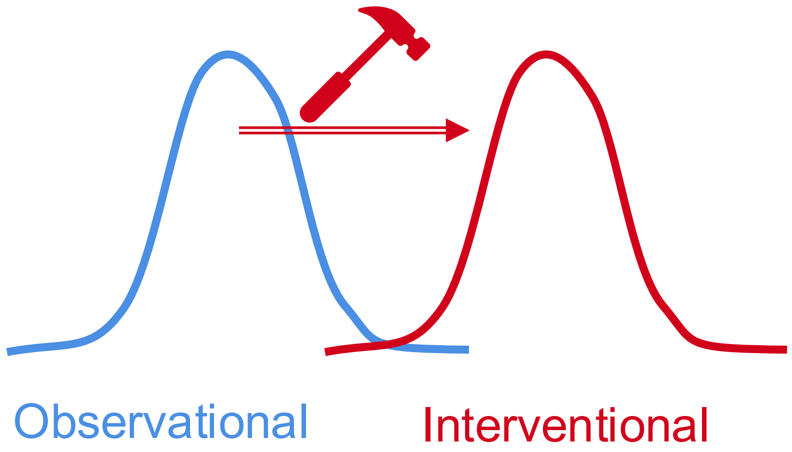
Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

What makes identifiability possible?

ICA: variability



Causality: interventions

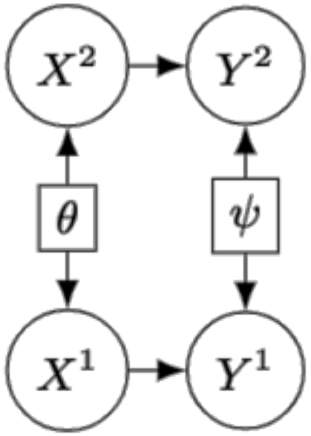


Non-i.i.d. data is the key!

Rajendran et al.: An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis. CLeaR, 2024.
Yao et al.: Marrying Causal Representation Learning with Dynamical Systems for Science. arXiv, 2024.

Exchangeability unifies ICA and causality

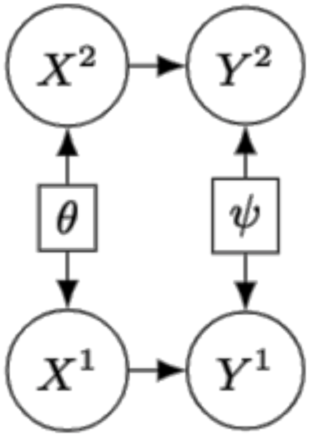
Exchangeability



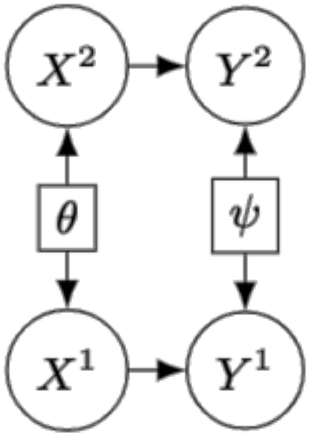
i.i.d.

Exchangeability unifies ICA and causality

Exchangeability



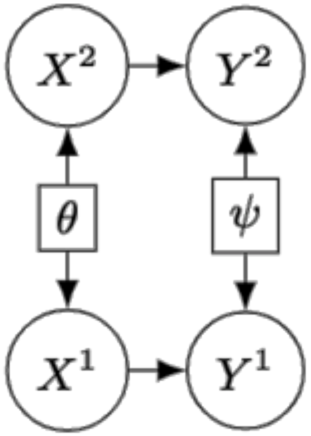
i.i.d.



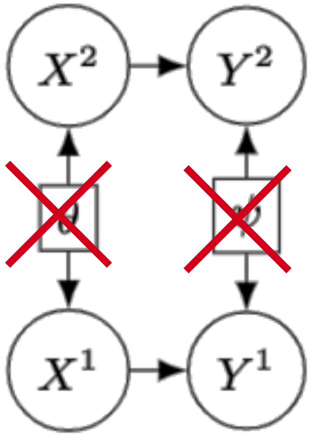
Guo et al.: Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. NeurIPS, 2023.

Exchangeability unifies ICA and causality

Exchangeability



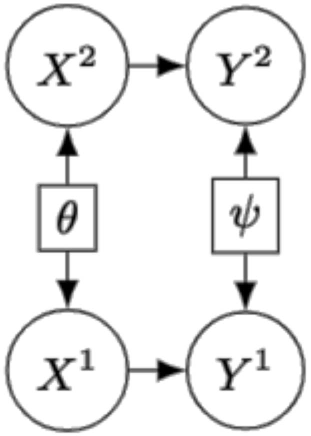
i.i.d.



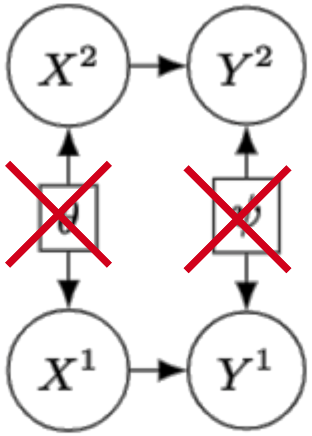
Guo et al.: Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. NeurIPS, 2023.

Exchangeability unifies ICA and causality

Exchangeability



i.i.d.



Exchangeability can describe variability conditions

Guo et al.: Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. NeurIPS, 2023.

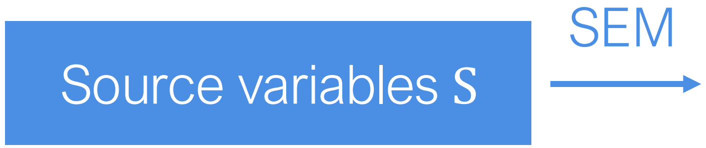
Identifiable Exchangeable Mechanisms

The Independent Exchangeable Mechanisms (IEM) framework

Source variables S

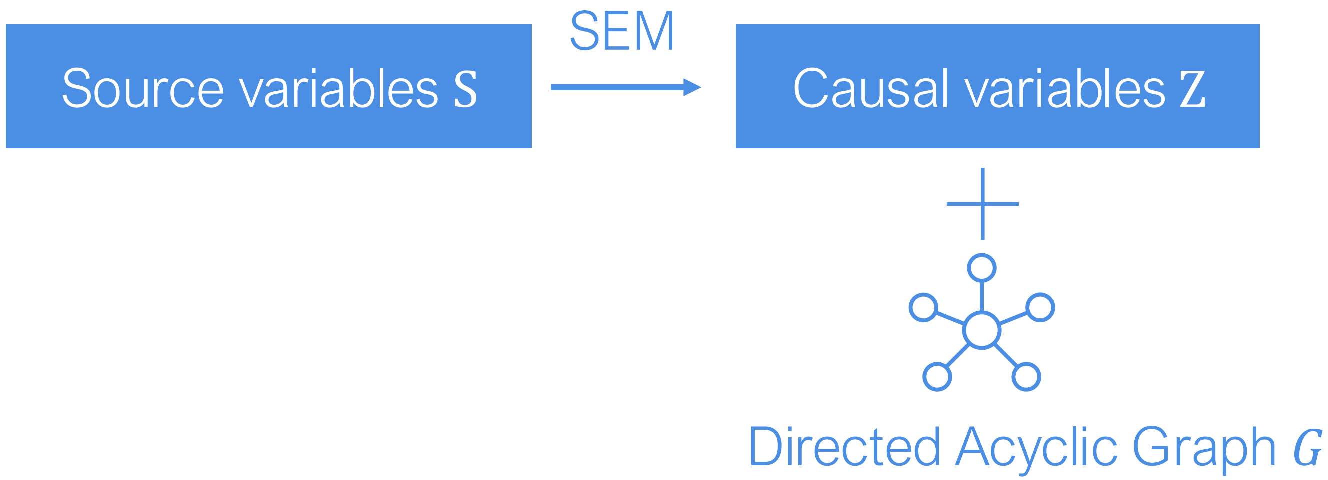
S – exogenous source variables
Z – causal variables
O – observations

The Independent Exchangeable Mechanisms (IEM) framework



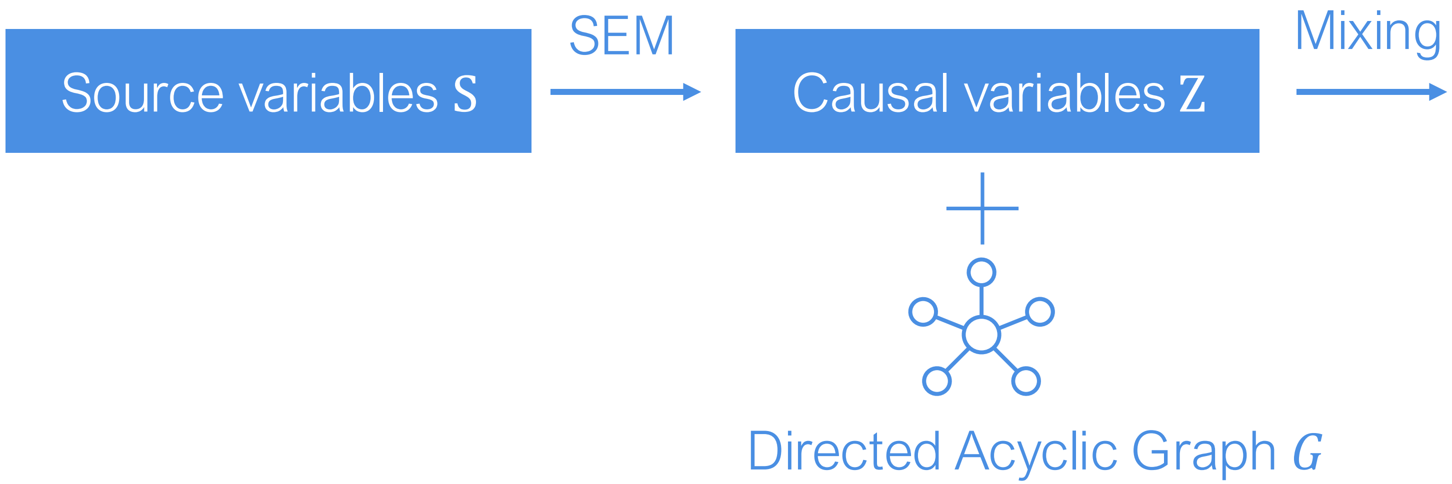
S – exogenous source variables
Z – causal variables
O – observations

The Independent Exchangeable Mechanisms (IEM) framework



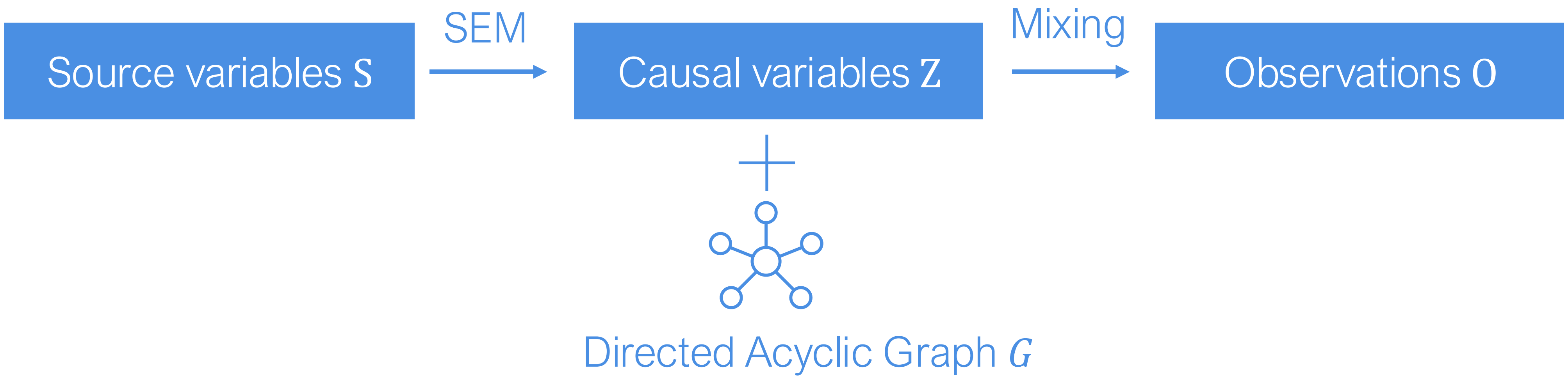
S – exogenous source variables
Z – causal variables
O – observations

The Independent Exchangeable Mechanisms (IEM) framework



S – exogenous source variables
Z – causal variables
O – observations

The Independent Exchangeable Mechanisms (IEM) framework

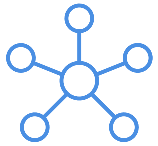


S – exogenous source variables
Z – causal variables
O – observations

The Independent Exchangeable Mechanisms (IEM) framework

Source variables S → (SEM) → Causal variables Z → (Mixing) → Observations O

+



Directed Acyclic Graph G

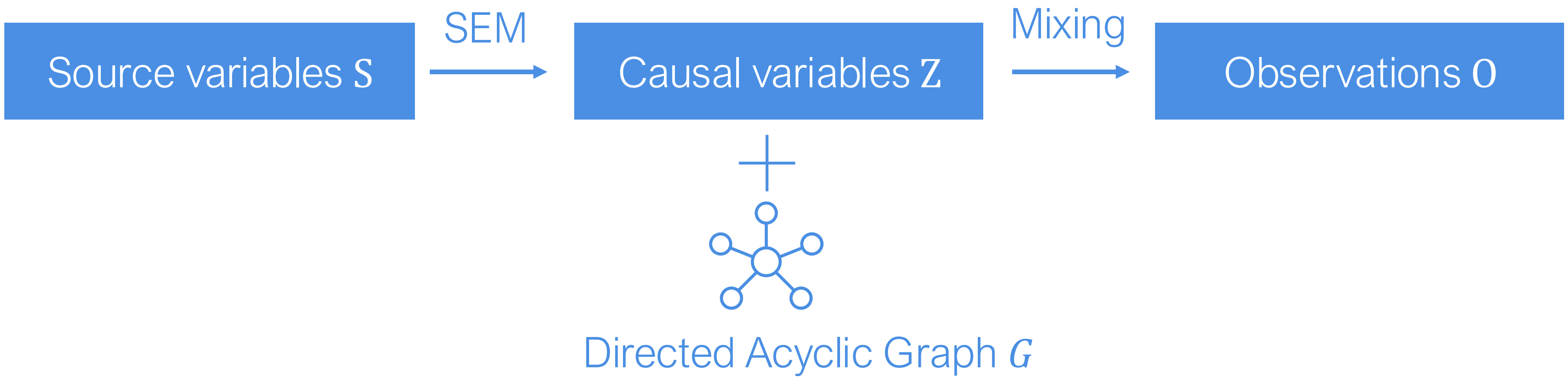
Example

Socioeconomic status → (SEM) → Smoking, Cancer → (Mixing) → Chest X-Ray

Smoking → Cancer

S – exogenous source variables
Z – causal variables
O – observations

The Independent Exchangeable Mechanisms (IEM) framework

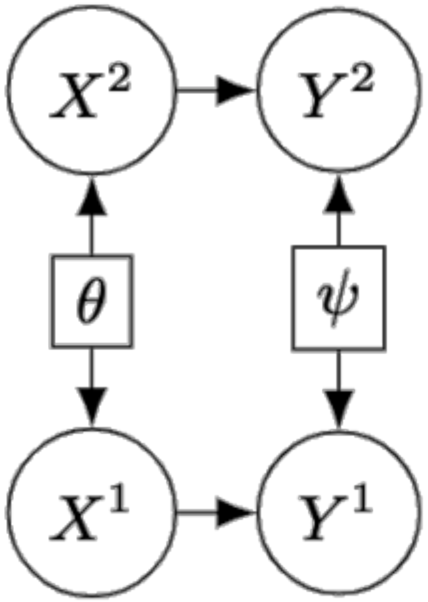


The difference between causal discovery, ICA, and causal representation learning is which components are modeled/observed

S – exogenous source variables
Z – causal variables
O – observations

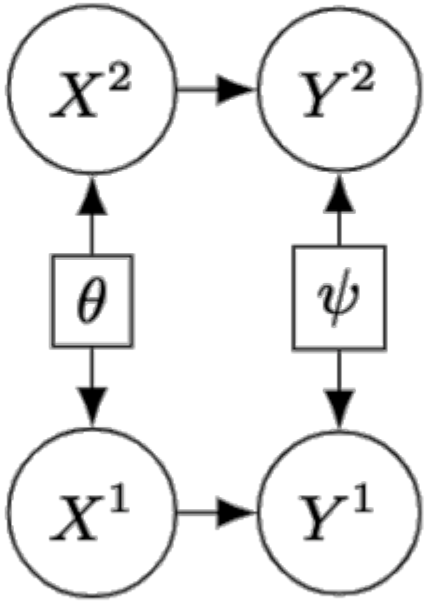
Extending Causal de Finetti

CdF

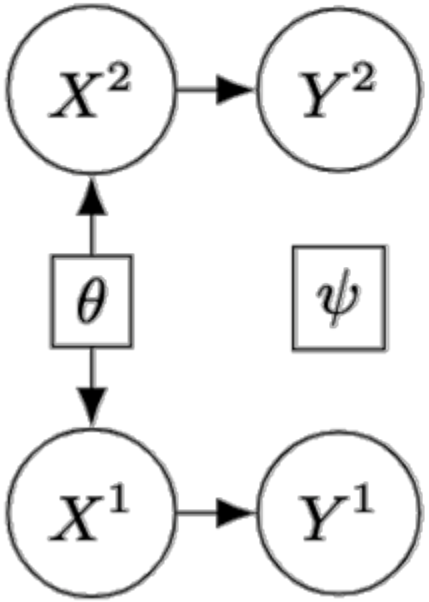


Extending Causal de Finetti

CdF

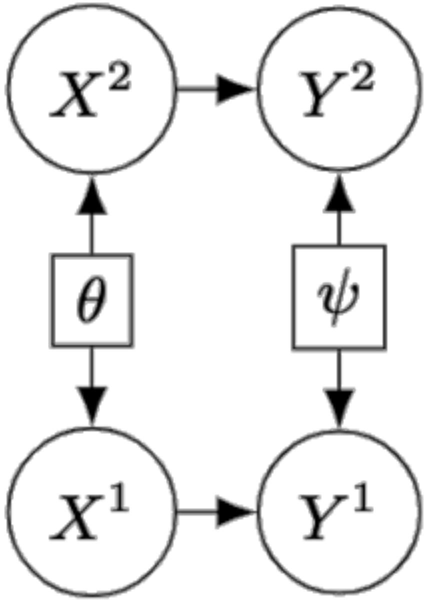


Cause variability

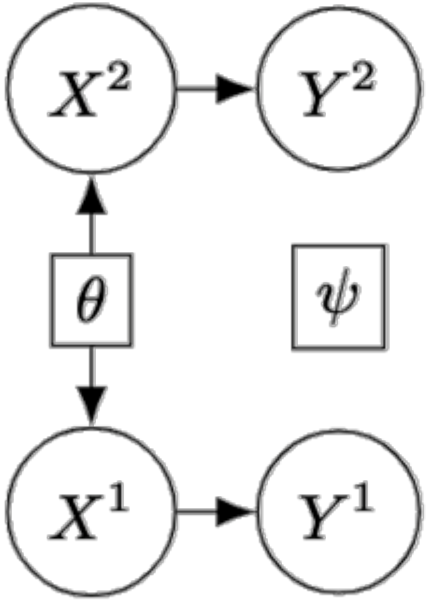


Extending Causal de Finetti

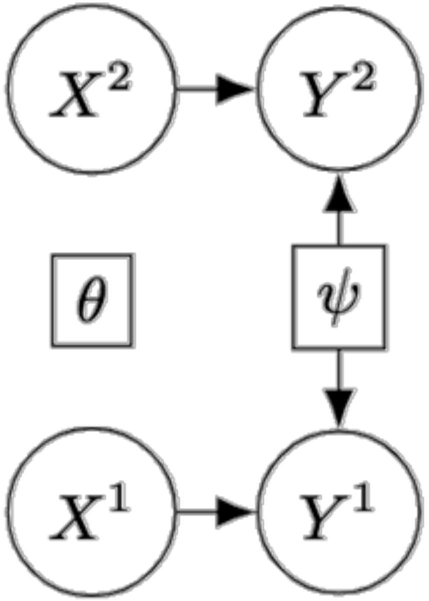
CdF



Cause variability

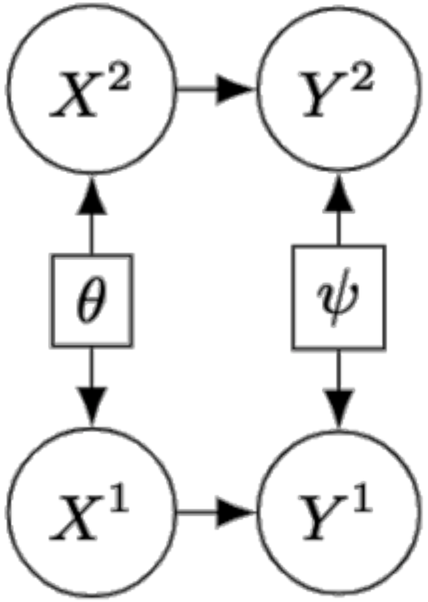


Mechanism variability

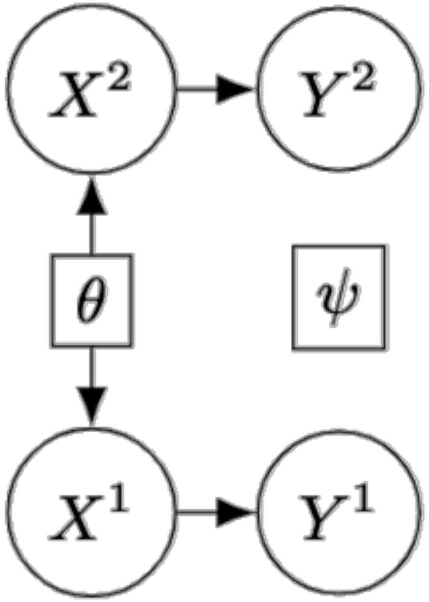


Extending Causal de Finetti

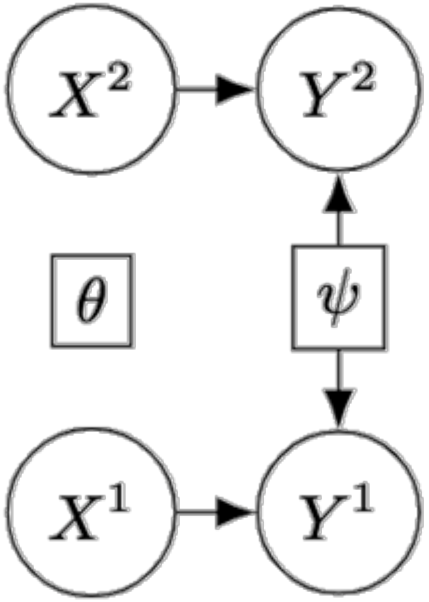
CdF



Cause variability



Mechanism variability

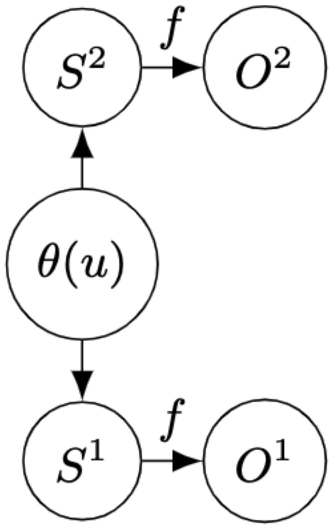


Causal discovery is possible iff any of these two conditions hold

Cause/mechanism duality for Time-Contrastive Learning

Original

- Non-stationary sources
- Deterministic mixing function



S – exogenous source variables
Z – causal variables
O – observed variables
Observed, don't care, target quantity

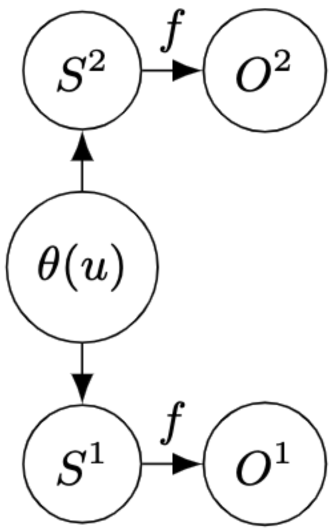
Dual

Hyvärinen et al.: Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. NeurIPS, 2016.

Cause/mechanism duality for Time-Contrastive Learning

Original

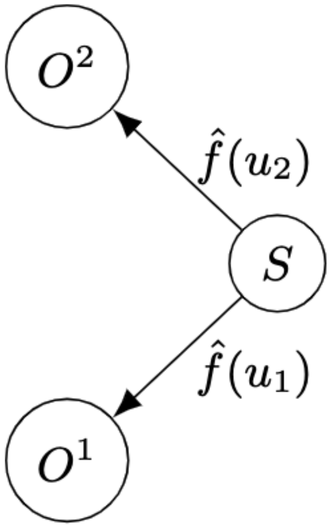
- Non-stationary sources
- Deterministic mixing function



S – exogenous source variables
Z – causal variables
O – observed variables
Observed, don't care, target quantity

Dual

- Stationary sources
- Stochastic mixing function



Hyvärinen et al.: Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. NeurIPS, 2016.

Conclusion for IEM

- We categorize the different types of non-i.i.d. data enabling identifiability
- We apply an exchangeable lens to unify
  - Causal discovery,
  - Identifiable representation learning (Independent Component Analysis)
  - Causal representation learning
- Our framework, Independent Exchangeable Mechanisms (IEM), is not universal, but focuses on the multi-environment case
  - For a more comprehensive unifying framework of CRL, see, Yao et al., 2024
- So what?
  - IEM leads to relaxed conditions of the Causal de Finetti theorem
  - IEM also suggests new identifiability results for both source and causal variables

Yao et al.: Unifying Causal Representation Learning with the Invariance Principle. arXiv, 2024.

Identifiable Exchangeable Mechanisms for Causal Structure and Representation Learning

Patrik Reizinger*, Siyuan Guo*, Ferenc Huszár, Bernhard Schölkopf, Wieland Brendel
patrik.reizinger@tuebingen.mpg.de

Kernel-Based Image Restoration with Uncertainty Guarantees

Bálint Horváth

– Institute for Computer Science and Control (SZTAKI),
Hungarian Research Network (HUN-REN)
– Budapest University of Technology and Economics (BME)

Joint work with: Balázs Csanád Csáji

Hungarian Machine Learning Days, Budapest, 2025

I. Problem setting

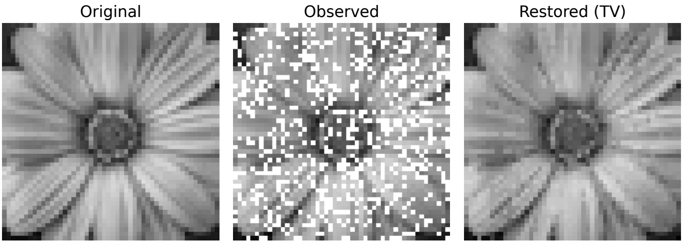
Main Assumptions and Objectives

Problem Setting

- **Images:** functions with $d = 2$ dimensional input vectors.
- **Pixels:** outputs of these functions at some observed inputs (typically distributed on a grid).
- **Observations:** the inputs are from $\mathcal{D} \doteq [0,1] \times [0,1]$, the outputs are centered and scaled, from $[-1,1]$ (grayscale images).
- **Generalization:** can be extended to multi-dimensional cases, (RGB or CYMK codes), for example, by handling each coordinate separately, then combining the results to a hyperrectangle.

Single Image Inpainting: Problem Setting

- **Main task:** estimate the value of missing pixels based on one image (without any additional information).
- **Well-known methods:** total variation based techniques, methods using biharmonic equations, several deep learning approaches.



Single Image Super-Resolution: Problem Setting

- **Aim:** map a low resolution image to a higher resolution one.
- **Recent solutions:** based on deep convolutional networks (e.g., EDSR: Enhanced Deep Residual Networks for Super-Resolution).

Main Assumption

Assumption (A1)

We have a finite sample $(x_1, y_1), \dots, (x_n, y_n)$ of inputs and outputs, where $x_k \in \mathcal{D} \doteq [0,1]^d$, $y_k \in [-1,1]$, and, $y_k = f_*(x_k)$, for $k \in [n] \doteq \{1, \dots, n\}$. Inputs $\{x_k\}$ are distinct and f_*, the "true" data generating function, belongs to a known RKHS $\mathcal{H}$ with a continuous and universal reproducing kernel k.

Def. Let $\mathbb{X}$ be a metric space, let $\mathcal{Z} \subseteq \mathbb{X}$ be a compact subset and let $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a kernel, i.e., symmetric and positive definite. Let $\mathcal{K}(\mathcal{Z}) \doteq \overline{span}\{\, k(z,.) : z \in \mathcal{Z}\,\}$, i.e., the closure w.r.t. the supremum norm of the linear span of all $k(z,.)$ functions, where $z \in \mathcal{Z}$. Let $\mathcal{C}(\mathcal{Z})$ be the set of all continuous $f : \mathcal{Z} \to \mathbb{R}$ type function. Kernel k is called universal if and only if for all compact $\mathcal{Z} \subseteq \mathbb{X}$, we have $\mathcal{K}(\mathcal{Z}) = \mathcal{C}(\mathcal{Z})$.

Paley-Wiener Spaces

Definition

A Paley-Wiener space is a subspace of $\mathcal{L}^2(\mathbb{R}^d)$, where for each $\varphi \in \mathcal{H}$ the support of the Fourier transform of φ is included in a given interval $[-\eta, \eta]^d$, where $\eta > 0$ is a hyper-parameter.

- These space are Reproducing Kernel Hilbert Spaces (RKHSs).
- The reproducing kernel of a Paley-Wiener space takes the form

$$k(u, v) = \pi^{-d} \prod_{j=1}^{d} \frac{\sin(\eta(u_j - v_j))}{u_j - v_j},$$

  where (for convenience) $\sin(\eta \cdot 0)/0$ is defined to be η.
- Motivation: we limit the frequencies of the target function ("band limited"), e.g., to ensure that it cannot change arbitrarily fast.

Objectives

- **Main goals:** estimate the missing pixel value of any given query input point and provide uncertainty bounds for the estimation.
- **Formally:** construct a point estimate $\bar{f}$ of f_* together with confidence band, i.e., a (data-dependent) function $I : \mathcal{D} \to \mathbb{R}^2$, where $I(x) = (\, I_1(x), I_2(x)\,)$ specifies the endpoints of a (closed) interval estimate for $f_*(x)$ and contains $\bar{f}(x)$ for any $x \in \mathcal{D}$.
- Therefore, the required properties are

$$\forall x \in \mathcal{D},\ \bar{f}(x) \in [I_1(x), I_2(x)],$$

$$\mathbb{P}\Big(\forall x \in \mathcal{D} : I_1(x) \leq f_*(x) \leq I_2(x)\Big) \,\geq\, 1 - \gamma,$$

where $\gamma \in (0, 1)$ is a (user-chosen) risk probability, and $\nu(I)$ is the reliability of the confidence band.

II. Known Norm Bounds

Point Estimation and Uncertainty Quantification

Point Estimation

– **'Key" function:** The element from $\mathcal{H}$ with minimum norm, which interpolates each output y_k at the corresponding input x_k:

$$\bar{f}(x) \,=\, \sum_{k=1}^{n} \hat{\alpha}_k k(x, x_k),$$

where the weights are $\hat{\alpha} = K^{-1}y$ with $y \doteq (y_1, \dots, y_n)^{\mathrm{T}}$ and $\hat{\alpha} \doteq (\hat{\alpha}_1, \dots, \hat{\alpha}_n)$, and $K_{i,j} = k(x_i, x_j)$ is the Gram matrix.

– By the reproducing property and the Cauchy-Schwartz inequality:

$$|\, f(x) - f(x')\,| \,\leq\, \|f\|_{\mathcal{H}}\, \|k_x - k_{x'}\|_{\mathcal{H}} \,\leq\, \|f\|_{\mathcal{H}}\, d(x, x')$$

for all $f \in \mathcal{H}$, where $k_x \doteq k(\cdot, x)$ and the distance is

$$d(x, x') \,=\, \sqrt{k(x,x) + k(x', x') - 2k(x, x')}$$

Uncertainty Quantification: Idea

Assumption (B1)

For the given risk probability $\gamma \in (0,1)$, we known a constant $\kappa \geq \| \bar{f} \|_{\mathcal{H}}^2$, where $\bar{f}$ is the minimum norm interpolant, with

$$\mathbb{P}\left(\|f_*\|_{\mathcal{H}}^2 \leq \kappa \right) \geq 1 - \gamma.$$

- **Simplified problem:** (B1) provides a priori known bounds for the kernel norm of the data-generating function (later: relaxation).
- For a candidate (x_0, y_0) pair, we calculate the norm square of the minimum norm interpolation of $\{(x_k, y_k)\}_{k=1}^n \cup \{(x_0, y_0)\}$.
- If this norm (square) is less than or equal to the upper bound (κ), we include (x_0, y_0) in our confidence band.
- **Required:** an efficient method to decide the endpoints of the confidence interval for every input $x_0 \in \mathcal{D}$.

Uncertainty Quantification: Construction

- **Optimization problems:** for a candidate query input point x_0, we are looking for the minimal/maximal y_0 value, where the norm of the interpolation function is less than or equal than κ.

- These lead to the following two (convex) optimization problems:

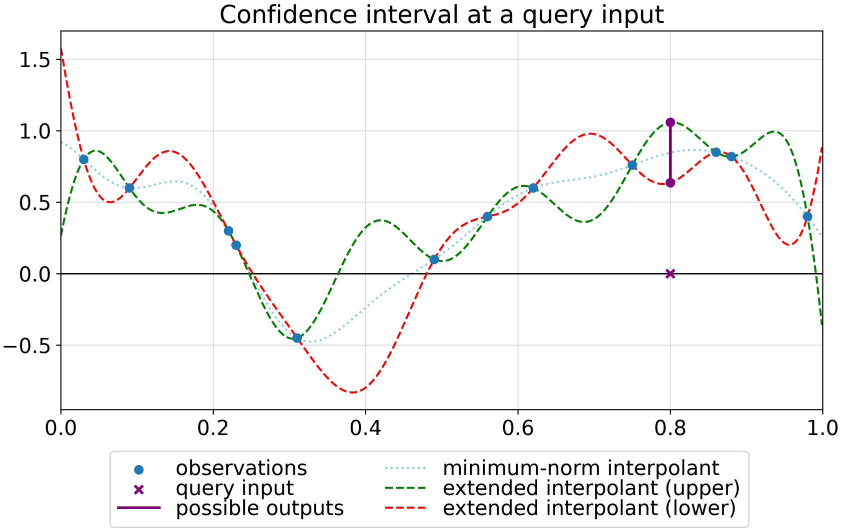
$$\begin{aligned} \min / \max \quad & y_0 \\ \text{subject to} \quad & (y_0, y^{\mathrm{T}}) K_0^{-1} (y_0, y^{\mathrm{T}})^{\mathrm{T}} \leq \kappa, \end{aligned}$$

where $K_0(i+1, j+1) \doteq k(x_i, x_j)$ for $i, j \in \{0, 1, \dots, n\}$.

- **Special structure:** they can be solved analytically.

- The optimal values, $y_{\min}$ and $y_{\max}$, determine the endpoints of the confidence interval at x_0: $I_1(x_0) \doteq y_{\min}$ and $I_2(x_0) \doteq y_{\max}$.

Uncertainty Quantification: Illustration



Theoretical Guarantees

Lemma 1: Point Estimation Inclusion

Assuming A1 and B1, it holds true that $\forall x \in \mathcal{D}$,

$$\bar{f}(x) \in [(I_1(x), I_2(x))].$$

Theorem 1: Guaranteed Coverage

Assume A1 and B1. Then, for any $\gamma \in (0, 1)$ be a risk probability and any finite sample size n, the construction of SGKI guarantees

$$\nu(\mathcal{I}_n) \geq 1 - \gamma.$$

III. Unknown Kernel Norm Bounds

Assumptions, Bounding the Norm

Unknown Kernel Norm Bounds: Assumptions

In order to relax (B1), we introduce the following assumptions:

(A2) Function f_* is from a Paley-Wiener space $\mathcal{H}$; and f_* is almost time-limited to $\mathcal{D}$, that is

$$\int_{\mathbb{R}} f_*^2(x)\, \mathbb{I}(x \notin \mathcal{D})\, \lambda(\mathrm{d}x) \,\leq\, \delta_0,$$

where $\mathbb{I}(\cdot)$ is an indicator and $\delta_0 > 0$ is a universal constant.

(A3) Sample $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{D} \times \mathbb{R}$ is i.i.d.

(A4) The inputs, $\{x_k\}$, are distributed uniformly on $\mathcal{D}$.

Bounding the Kernel Norm

Lemma 2: Upper Bound for the Kernel Norm

Assuming A2, A3 and A4, for any user-chosen risk probability $\gamma \in (0,1)$, we have (for any finite sample size)

$$\mathbb{P}(\|f_*\|_{\mathcal{H}}^2 \leq \kappa) \geq 1 - \gamma,$$

with the following choice on the upper bound κ:

$$\kappa \doteq \frac{1}{n}\sum_{k=1}^{n} y_k^2 + \sqrt{\frac{\ln(1/\gamma)}{2n}} + \delta_0.$$

IV. Reducing the Computational Complexity

Schur Complements

Complexity Reduction: Schur Complements (1)

- **Issue:** for every query input $x_0 \in \mathcal{D}$, we need to calculate the inverse of the extended Gram matrix, namely K_0^{-1}.
- **Alternative solution:** for a given $x_0 \in \mathcal{D}$, partition K_0 as

$$\begin{bmatrix} r_0 & k_0 \\ k_0^{\mathrm{T}} & K \end{bmatrix},$$

  where $K \in \mathbb{R}^{n\times n}$, $k_0 \in \mathbb{R}^n$, $r_0 \in \mathbb{R}$, with $r_0 = k(x_0, x_0)$ and $k_{0,i} = k(x_0, x_i)$, for $i \in [n]$.
- **Schur complement:** we can write

$$K_0^{-1} = \begin{bmatrix} g_0^{-1} & -K^{-1}k_0 g_0^{-1} \\ -g_0^{-1}k_0^{\mathrm{T}}K^{-1} & K^{-1} + K^{-1}k_0 g_0^{-1} k_0^{\mathrm{T}} K^{-1} \end{bmatrix},$$

  where $g_0 \doteq (K_0/r_0) \doteq r_0 - k_0^{\mathrm{T}} K^{-1} k_0$ is the Schur complement.

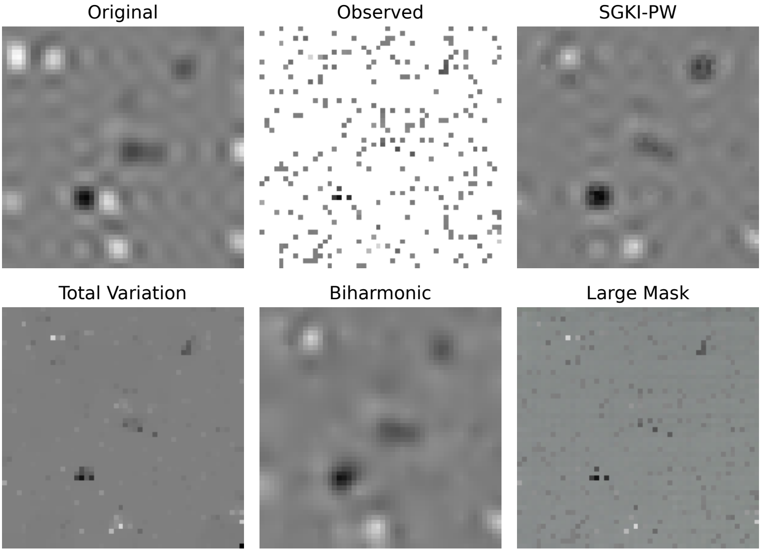
Complexity Reduction: Schur Complements (2)

- **Complexity:** the new approach to compute the confidence interval needs $\mathcal{O}(n^2)$ floating point operations (flops), instead of $\mathcal{O}(n^3)$ flops, assuming the matrix K^{-1} is available.
- K^{-1} only needs to be computed once, then K^{-1} can be used for each possible query input x_0 to construct the inverse of K_0.
- **Example:** If $n = 100$, then the speedup of computing K_0^{-1}, given K^{-1}, could be even $100\times$, depending on the implementation.
- The (parameterized) complexity of the original method is $\mathcal{O}((hw - n)n^3)$ flops, while the Schur complement based version requires $\mathcal{O}(n^3 + (hw - n)n^2))$ flops (the image has $h \times w$ pixels).

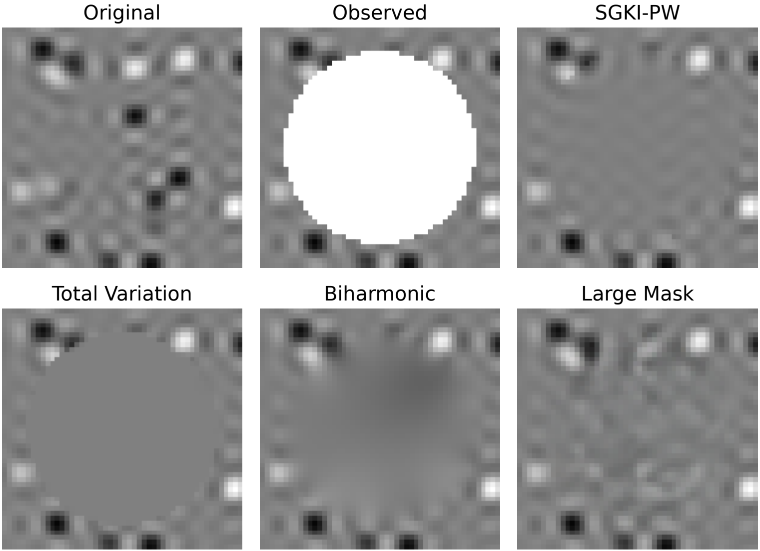
V. Empirical Validation

Single Image Inpainting and Super-Resolution

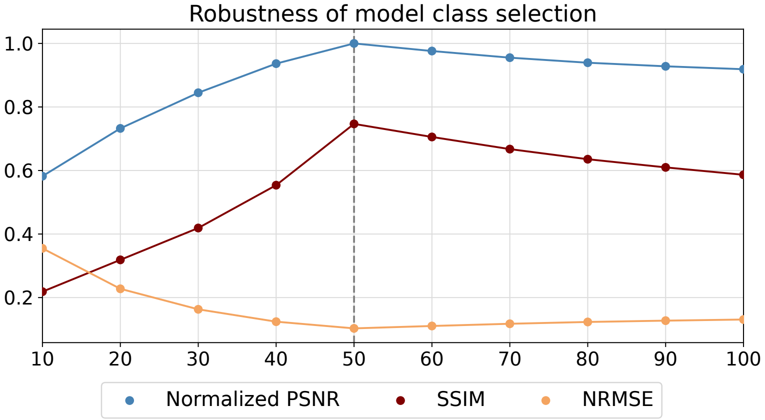
Single Image Inpainting: Random Pixels



Single Image Inpainting: Deterministic Cut-out

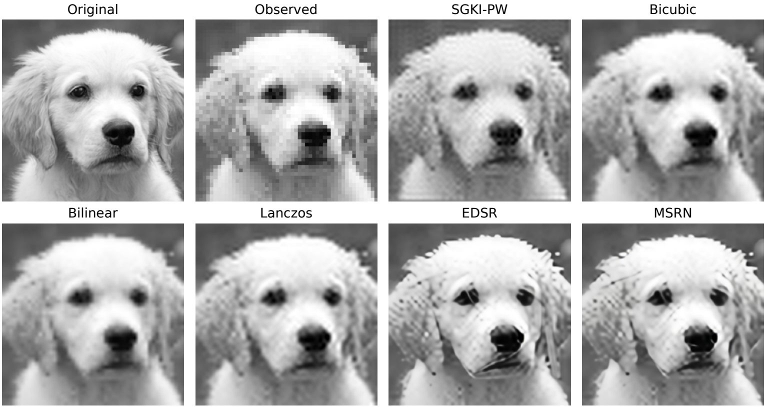


Misspecification Robustness



Uncertainty Quantification: Relative Diameters

Single Image Super-Resolution



Quantitative Results (1)

Table: Synthetic single image inpainting experiments using 100 randomly generated "band-limited" images with 10 % observations.

Inpainting	PSNR (avg)	SSIM (avg)	NRMSE (avg)
SGKI-PW	**26.0489**	**0.7499**	0.1082
Total Variation	23.3874	0.5354	0.1356
Biharmonic	25.4460	0.6872	**0.1074**
Large Mask	22.6320	0.4711	0.1476

Quantitative Results (2)

Table: Synthetic single image super-resolution experiments using 20 "band-limited" images with the aim to double their resolution.

Super-resolution	PSNR (avg)	SSIM (avg)	NRMSE (avg)
SGKI-PW	**38.1143**	**0.9831**	**0.0258**
EDSR	36.1177	0.9741	0.0311
MSRN	36.1017	0.9740	0.0312
Bicubic	35.9471	0.9734	0.0318
Bilinear	35.8641	0.9725	0.0321
Lanczos	36.0334	0.9745	0.0315
Nearest Neighbor	33.2433	0.9447	0.0433

Quantitative Results (3)

Table: Single image inpainting experiments, with 10 % pixels observed, on real-world images from the grayscale Set12 dataset. SGKI was used with the Gaussian kernel, which is defined as $k(z, s) \doteq \exp(-||z - s||^2/(2\sigma^2))$.

Inpainting	PSNR (avg)	SSIM (avg)	NRMSE (avg)
SGKI-G	16.7601	0.3792	0.2765
Total Variation	17.0084	0.4201	0.2693
Biharmonic	**17.3096**	**0.5260**	**0.2605**
Large Mask	14.7450	0.2237	0.3477

Quantitative Results (4)

Table: Single image super-resolution experiments on real-world images from the grayscale Set12 dataset; the target scale was $\times 4$.

Super-resolution	PSNR (avg)	SSIM (avg)	NRMSE (avg)
SGKI-PW	**21.6117**	**0.6263**	**0.1597**
EDSR	16.7301	0.5049	0.2773
MSRN	16.8671	0.5078	0.2729
Bicubic	19.9679	0.6119	0.1911
Bilinear	20.4913	0.6211	0.1801
Lanczos	19.8083	0.5967	0.1946
Nearest Neighbor	18.8881	0.5624	0.2166

Summary

- The SGKI method can construct simultaneous, non-asymptotic non-parametric confidence bands if the data-generating function is from an RKHS having a continuous and universal kernel.
- The constructed point estimation, provided by the minimum norm interpolant, is always included in the confidence band.
- The uncertainty of the missing pixels can be quantified.
- By using Schur complements, the computational complexity of the method can be significantly improved.
- Several numerical experiments and comparisons were presented to support the viability of the approach.

Related Publications

(J1) Csáji, B. Cs.; Horváth, B.: Nonparametric, Nonasymptotic Confidence Bands with Paley-Wiener Kernels for Band-Limited Functions, IEEE Control Systems Letters, IEEE Press, Vol. 6, 2022, pp. 3355–3360.

(J2) Horváth, B., Csáji, B. Cs.: Single Image Inpainting and Super-Resolution with Simultaneous Uncertainty Guarantees by Universal Reproducing Kernels, Machine Learning, Springer Nature, 114, 179, 2025.

(J3) Csáji, B. Cs.; Horváth, B.: Derandomizing Simultaneous Confidence Regions for Band-Limited Functions by Improved Norm Bounds and Majority-Voting Schemes, IEEE Control Systems Letters, IEEE Press, Vol. 9, 2025, pp. 1381-1386.

Thank you for your attention!

How Not to Stitch Representations to Measure Similarity: Task Loss Matching Versus Direct Matching

András Balogh[1], Márk Jelasity[1, 2]

[1]University of Szeged [2]HUN-REN-SZTE Research Group on AI

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

SZTE SZEGEDI TUDOMÁNYEGYETEM

RGAI

Introduction

- Methods for studying the similarity of internal representations are important tools for understanding neural networks
- Model stitching allows us to study the functional similarity of representations
- Recent observations[1] question the reliability of model stitching as a measure of similarity, and its interpretability and inner working is under-investigated

[1]: Hernandez et al. Model stitching: Looking for functional similarity between representations. *In SVRHM Workshop @ NeurIPS*, 2022.

Introduction

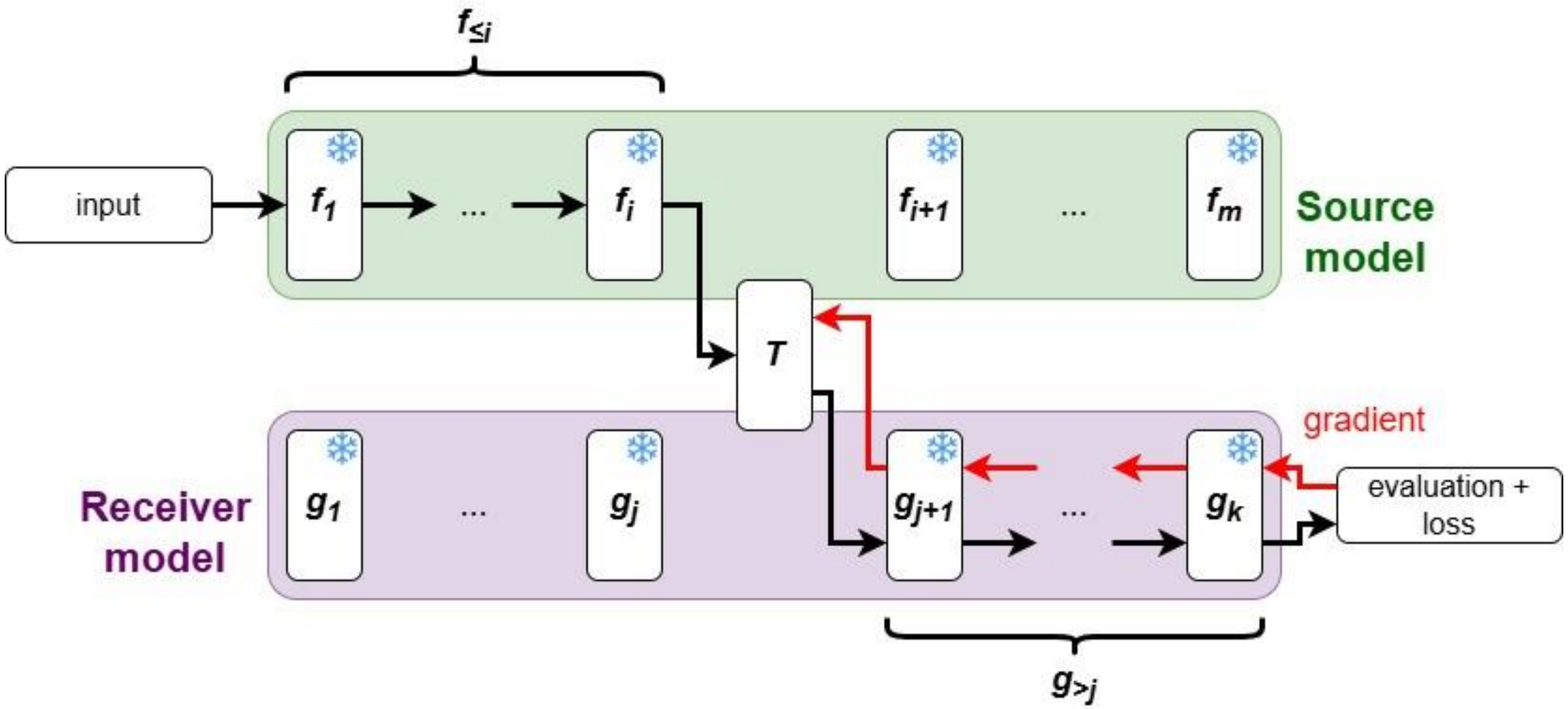
- Methods for studying the similarity of internal representations are important tools for understanding neural networks
- Model stitching allows us to study the functional similarity of representations
- Recent observations question the reliability of model stitching as a measure of similarity, and its interpretability and inner working is under-investigated

Our main question:

Is model stitching a reliable way to measure similarity of neural network representations?

Spoiler: yes and no – depending on **how** we stitch models.

Model Stitching



How Can We Stitch Models?

- **Task Loss Matching (TLM)**: $\arg_\theta \min \mathbb{E}_{p(x,y)}\left[\mathcal{L}\left(\left[g_{>j} \circ T_\theta \circ f_{\leq i}(x)\right], y\right)\right]$
  - We only care about task performance
  - No requirements on structural alignment
  - Only functional similarity
- **Direct Matching (DM)**: $\arg_\theta \min \mathbb{E}_{p(x)}\left[\left\| T_\theta \circ f_{\leq i}(x) - g_{\leq j}(x)\right\|_F\right]$
  - We only care about structural alignment
  - No requirements on task performance
  - Functional and structural similarity

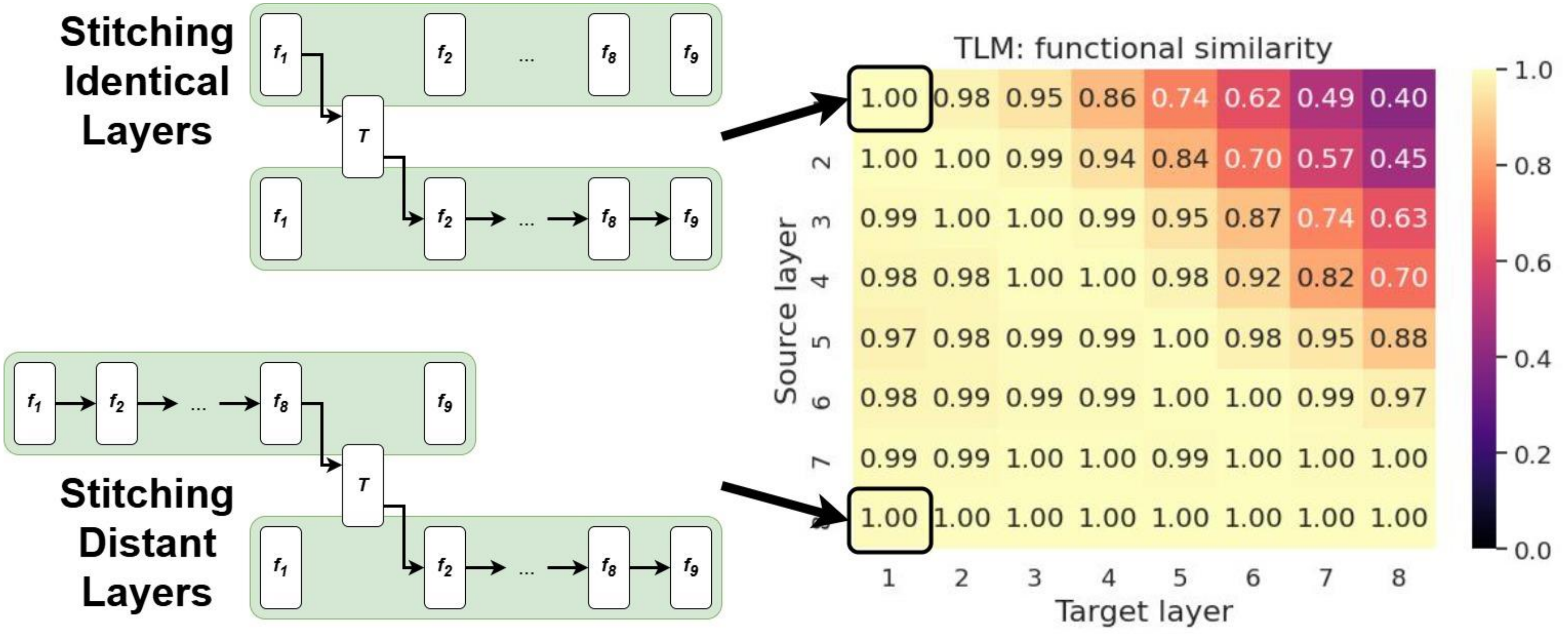
Task Loss Matching is Unreliable

- **Intra-network similarity**: Within a network, each layer should be most similar to itself.
- **Inter-network similarity**: Between two architecturally identical networks trained from different initializations, each layer in one network should be most similar to the corresponding layer in the other network. (Kornblith et al. 2019)

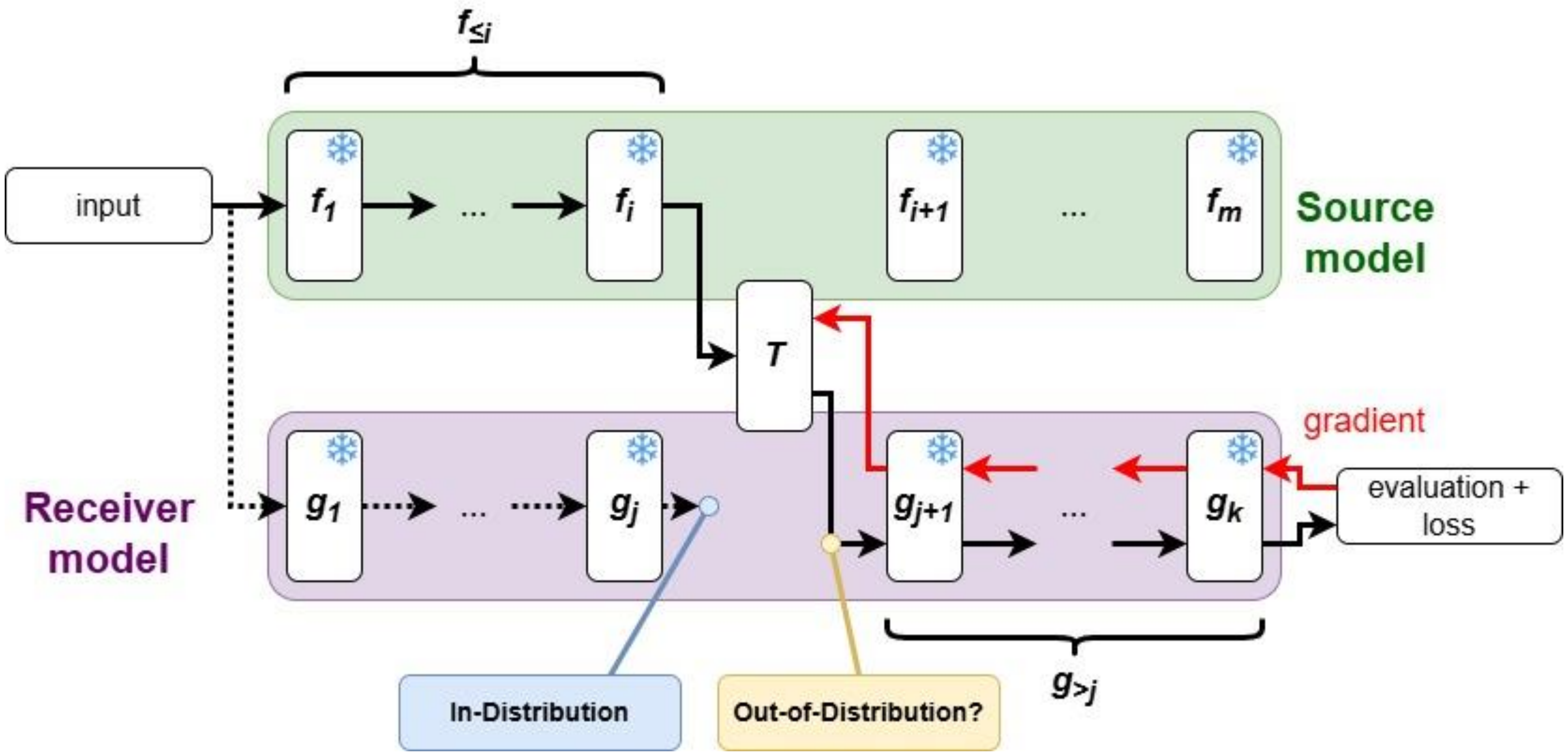
Similarity index	Intra-network		Inter-network	
	RN-18	ViT-Ti	RN-18	ViT-Ti
PWCCA	100%	100%	12.50%	8.33%
OPD	100%	100%	19.17%	18.33%
LCKA	100%	100%	96.11%	34.81%
TLM	63.75%	24.17%	35.28%	10.00%
DM (struct.)	100%	100%	92.50%	25.19%
DM (func.)	100%	100%	66.39%	11.85%

PWCCA: Morcos et al. Insights on representational similarity in neural networks with canonical correlation. *In NeurIPS*, 2018.
LCKA: Kornblith et al. Similarity of neural network representations revisited. *In ICML*, 2019.
OPD: Ding et al. Grounding representation similarity through statistical testing. *In NeurIPS*, 2021

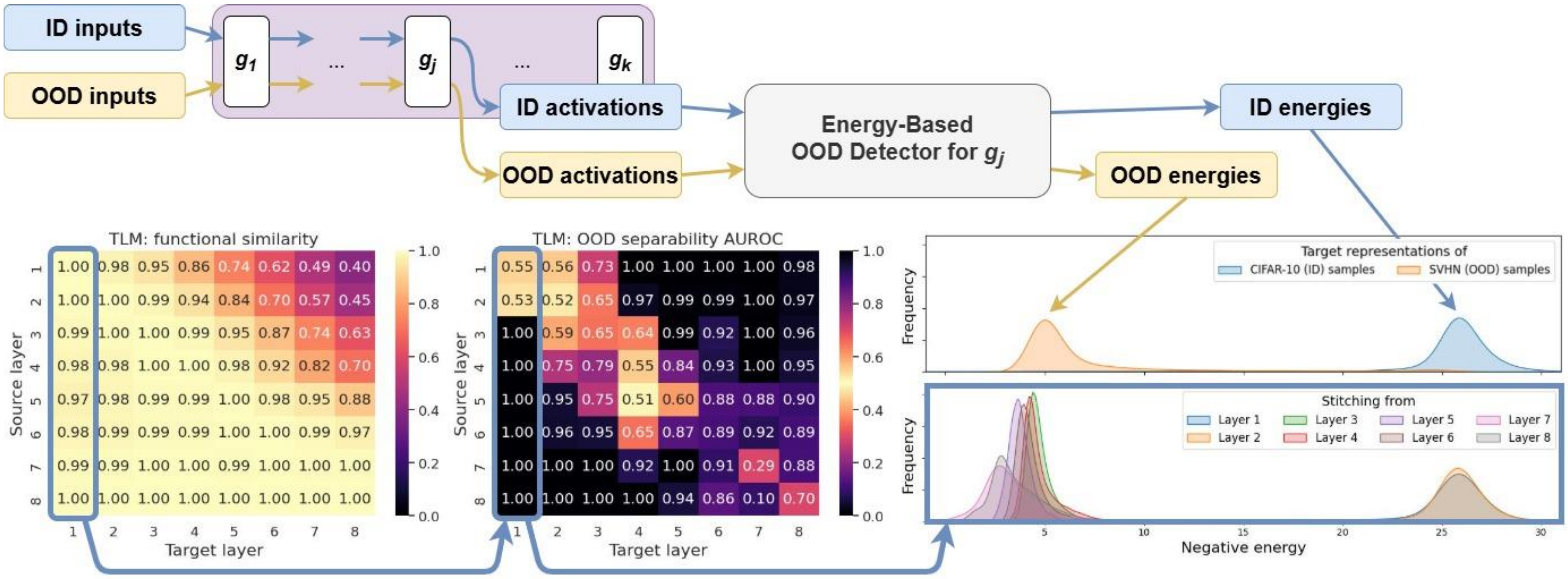
Intra-Network Similarities



Out-of-Distribution Representations

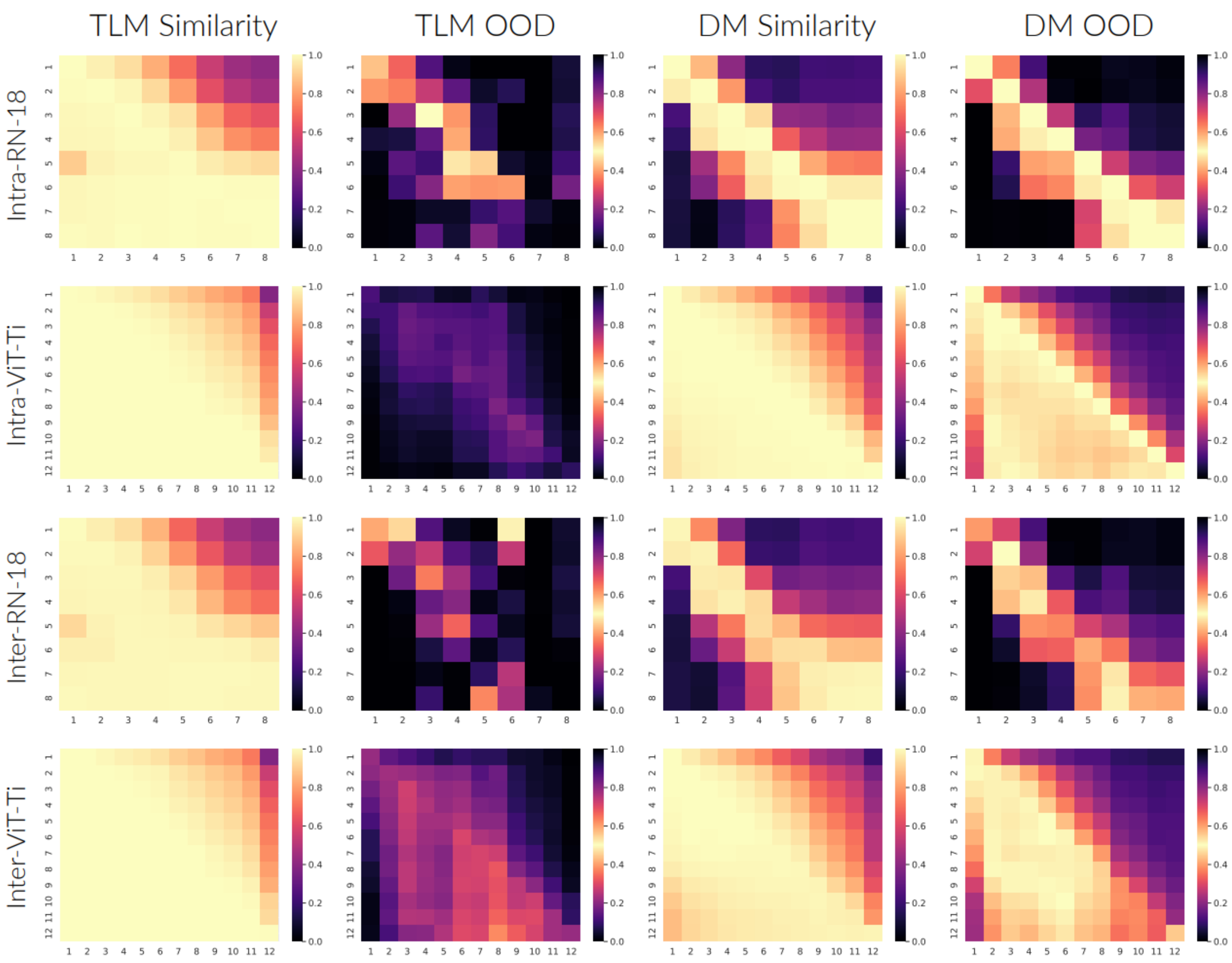


Out-of-Distribution Representations



Liu et al. Energy-based out-of-distribution detection. *In NeurIPS*, 2020

OOD Representations: TLM vs. DM



Statistical Tests for Functional Similarity

- We compare direct matching to structural similarity indices using the framework of (Ding et al. 2021), to see how well correlated they are to functional similarity.
  - Similarity indices should be **sensitive** to changes that affect functionality
  - Similarity indices should be **specific** against changes that don't affect functionality

Task	Model	LCKA		PWCCA		OPD		DM (func.)		DM (struct.)	
		τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
CIFAR-10	ResNet	0.569	0.722	0.628	0.777	0.548	0.718	**0.790**	**0.917**	0.658	0.807
	ViT	0.485	0.632	**0.826**	**0.945**	0.732	0.862	0.745	0.889	0.786	0.906
SVHN	ResNet	0.692	0.833	0.691	0.828	0.574	0.734	**0.762**	**0.893**	0.705	0.836
	ViT	0.704	0.838	0.618	0.782	0.697	0.846	0.672	0.864	**0.719**	**0.879**
ImageNet	ResNet	0.776	0.890	**0.832**	**0.924**	0.670	0.846	0.810	0.914	0.6631	0.751
	ViT	0.595	0.752	**0.836**	**0.946**	0.642	0.815	0.776	0.921	0.774	0.903

Sensitivity test

Task	Model	LCKA		PWCCA		OPD		DM (func.)		DM (struct.)	
		τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
CIFAR-10	ResNet	0.859	0.9700	0.729	0.896	0.768	0.920	0.857	0.970	**0.876**	**0.9750**
	ViT	0.719	0.870	0.650	0.785	0.531	0.682	**0.782**	**0.921**	0.664	0.802
SVHN	ResNet	**0.929**	0.992	0.761	0.914	0.927	**0.992**	0.890	0.982	0.901	0.985
	ViT	0.715	0.876	0.302	0.425	≈ 0	≈ 0	**0.742**	**0.893**	0.290	0.410
ImageNet	ResNet	0.844	0.952	0.659	0.821	0.597	0.678	**0.856**	**0.964**	≈ 0	≈ 0
	ViT	0.714	0.865	0.451	0.597	0.499	0.636	**0.791**	**0.918**	0.491	0.644

Specificity test

Ding et al. Grounding representation similarity through statistical testing. *In NeurIPS*, 2021

Conclusions

- Purely functional measures of similarity are unreliable
  - Strong inconsistency with basic sanity checks
  - Unintuitively high similarities between functionally different representations
- The reason behind this is out-of-distribution (OOD) alignment
  - As a result of optimizing for task performance, even with simple linear stitching
  - When similarities seem intuitive, purely functional alignment might still be OOD
- Direct Matching effectively combines structural and functional perspectives
  - In-distribution alignment
  - High correlation with functional similarity

Thank you for your attention!

This work was supported by the University Research Grant Program (in Hungarian: EKÖP) of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund, the European Union project RRF-2.3.1-21-2022- 00004 within the framework of the Artificial Intelligence National Laboratory, and by the project TKP2021-NVA-09, implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

Questions?

The Structure of Relation Decoding Linear Operators in Large Language Models

Miranda Anna Christ[*1,2], Adrián Csiszárik[*2,3], Gergő Becsó[3], Dániel Varga[2]*

[1]*Fazekas Mihály High School;* [2]*Alfréd Rényi Institute of Mathematics;* [3]*Eötvös Loránd University*

**Equal contribution*

Relations in LLMs

"Michael Jordan plays the basketball"

subject relation object

(Michael Jordan; person plays sport; basketball)

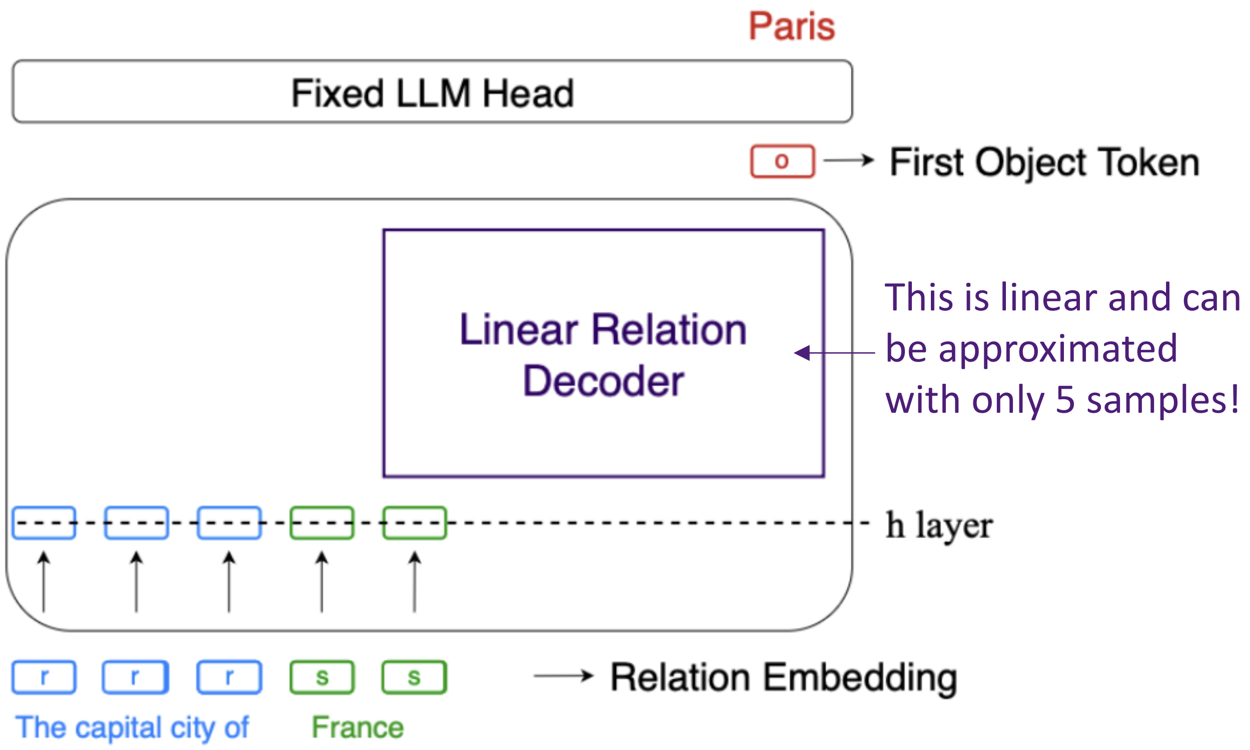
The samples of a given relation $\mathcal{R}$ are all $\{(s_0, o_0); (s_1, o_1); \ldots\}$ pairs.

Linear Relation Decoders [1]

One linear relation decoder is:

- ⊳ Relation specific
- ⊳ More than 16M parameters

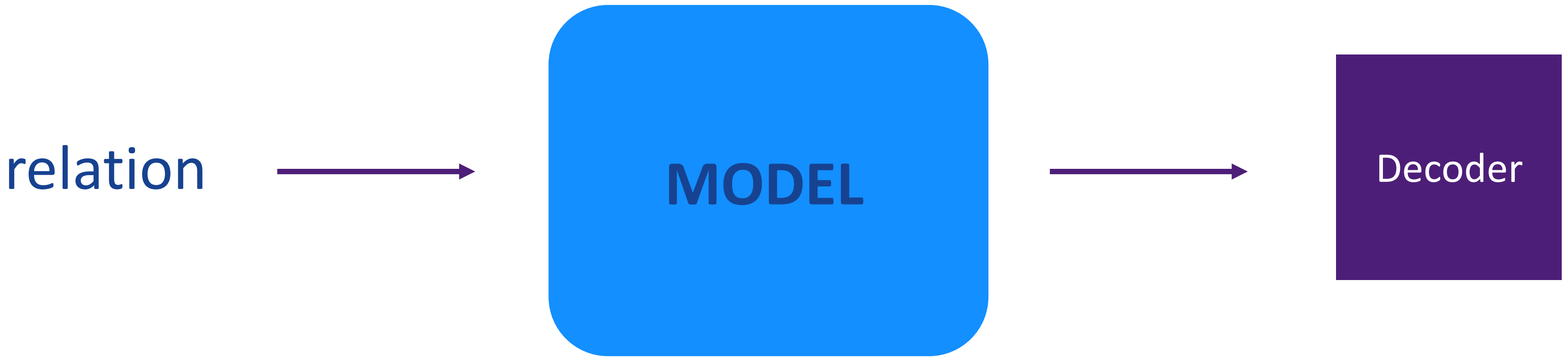
► All of the 47 relation decoders are $47 * 16M = 790M$ parameters



[1] Hernandez *et al.* 2024. *Linearity of Relation Decoding in Transformer Language Models*. ICLR 2024 (Spotlight)

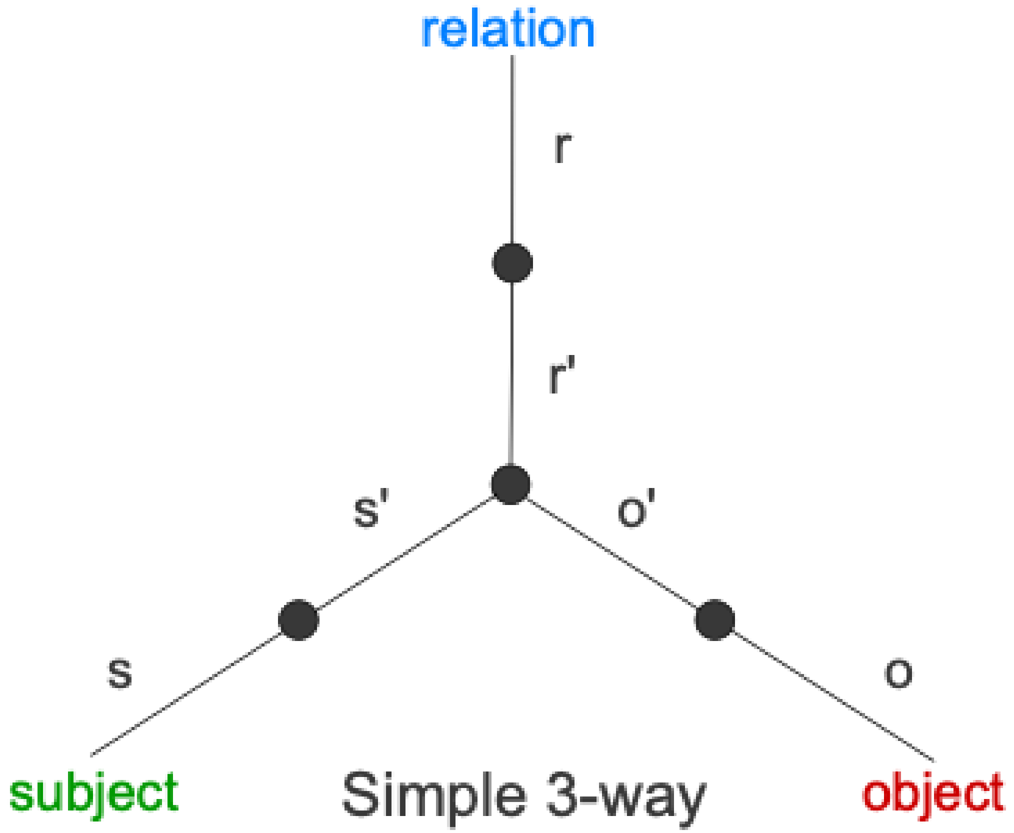
The Structure of Relation Decoding

- *How do the collection of relation decoders relate?*
- *Can we compress them? Or even model them?*



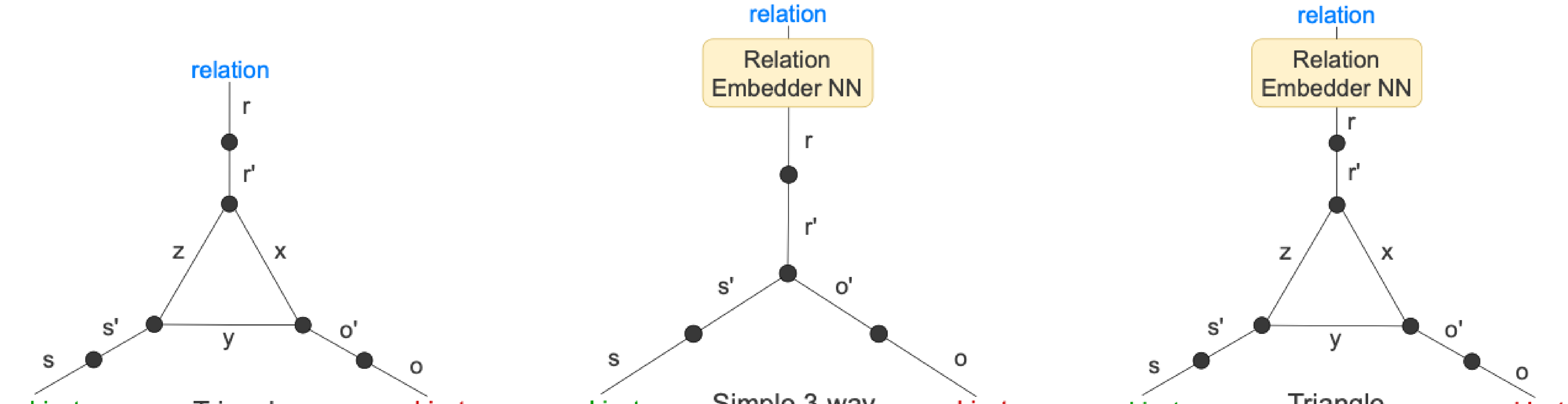
Can we handle relations collectively by compressing the decoders effectively?

Order-3 tensor networks



Simple 3-way
tensor network

Order-3 tensor networks



Triangle tensor network

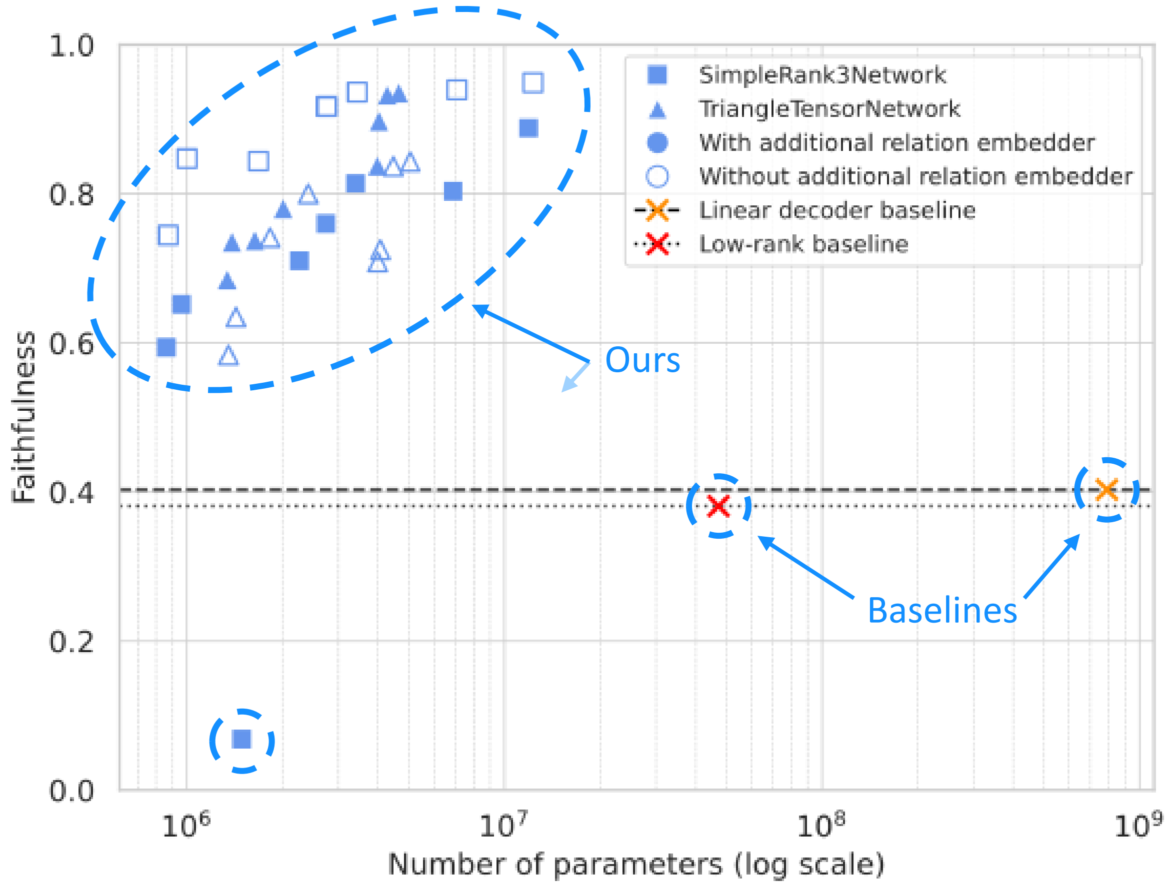
Simple 3-way tensor network with relation embedder

Triangle tensor network with relation embedder

Linear Relation Decoders are Substantially Compressible!

The measure of performance: Faithfulness

⊳ Accuracy of object prediction

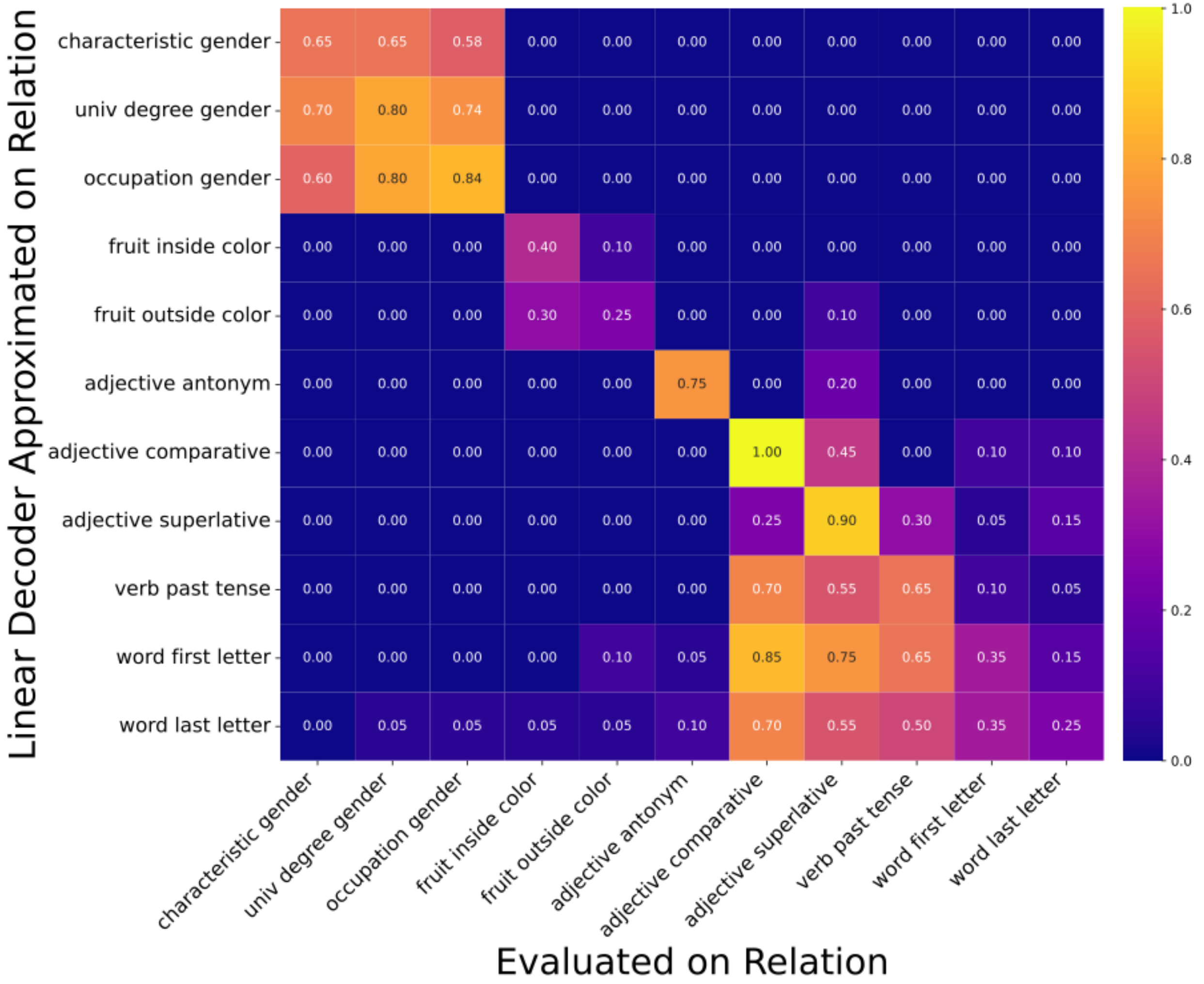


Understanding Compressibility: Semantic Structure

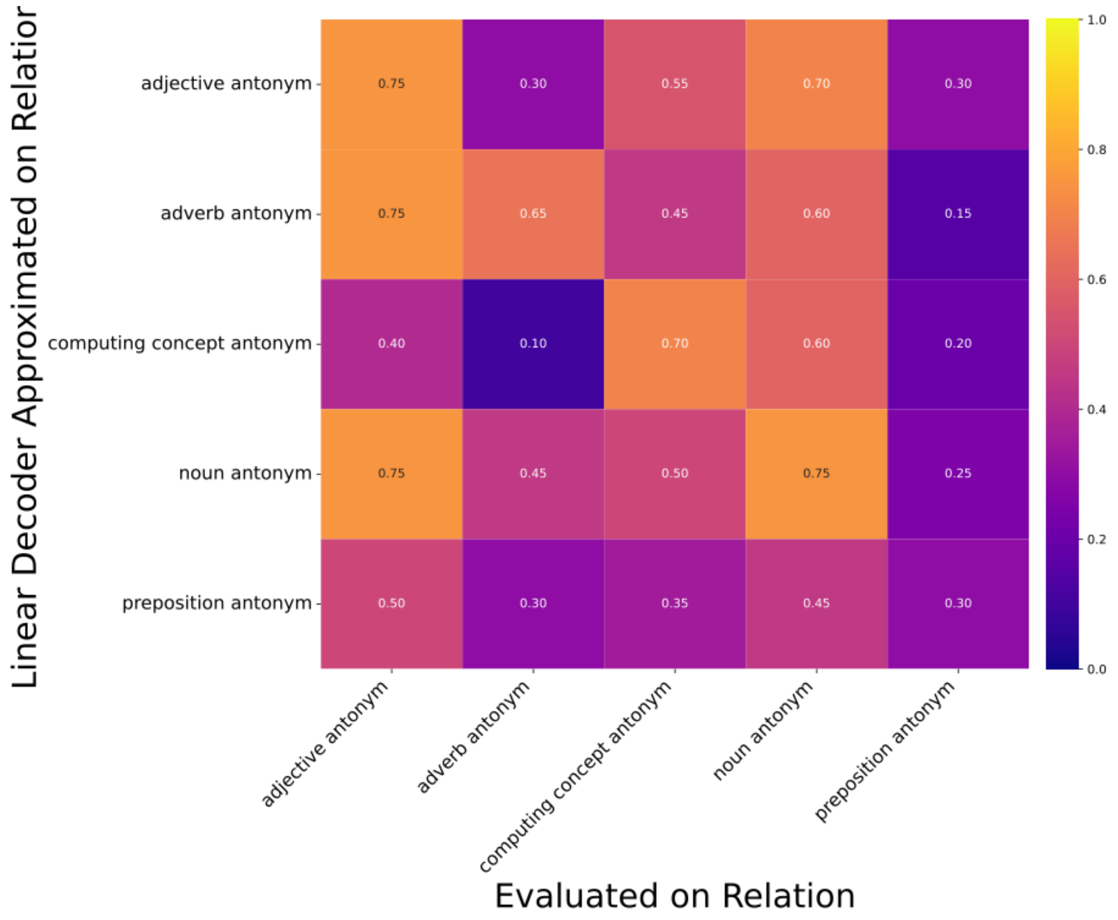
Cross-evaluation protocol:

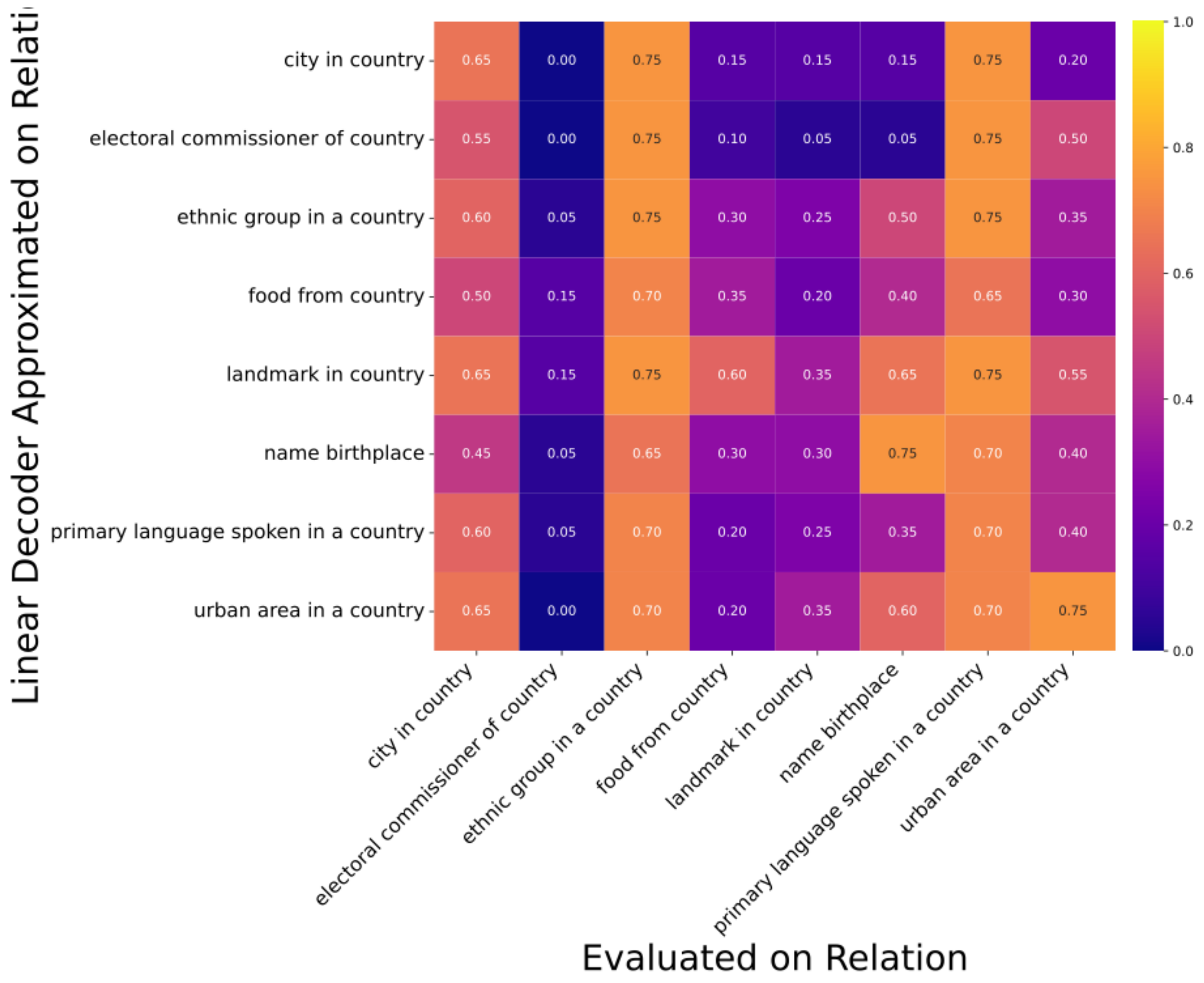
We evaluate

- ⊳ every $\mathcal{R}_i \in \mathcal{R}$ relation on
- ⊳ every $\mathcal{M}_j \in \mathcal{M}$ decoder and
- ⊳ compute faithfulness $f(\mathcal{R}_i, \mathcal{M}_j)$



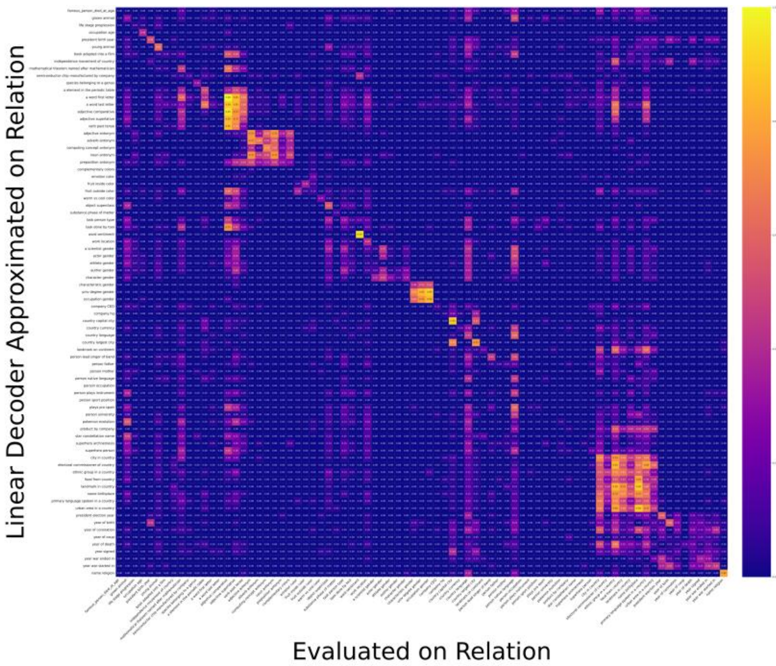
Extending the Dataset



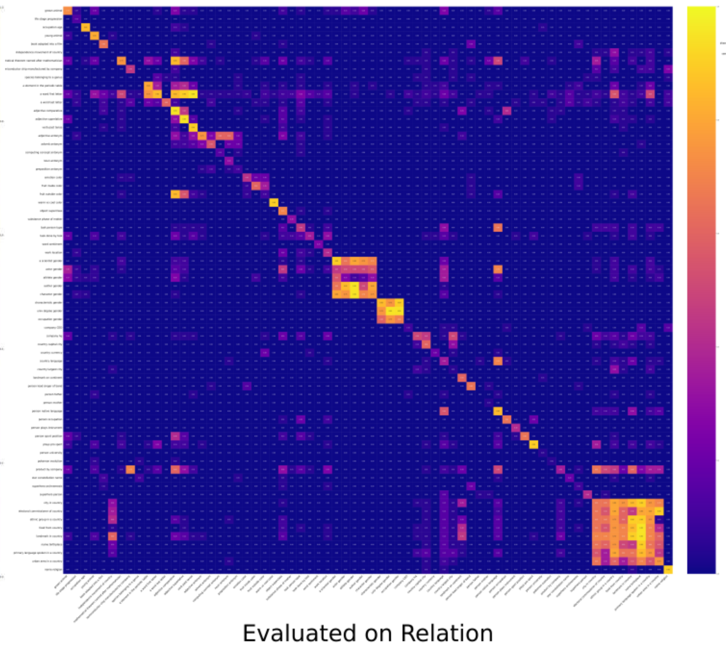


Linear relation decoders are primarily property-based, rather than relation-specific

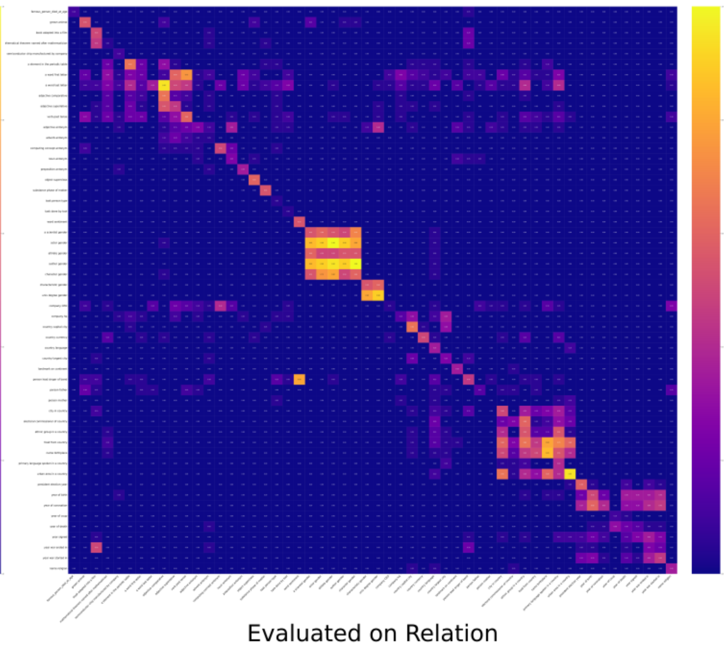
Different LLM Architectures



GPT-J 6B



Llama 3.1 8B



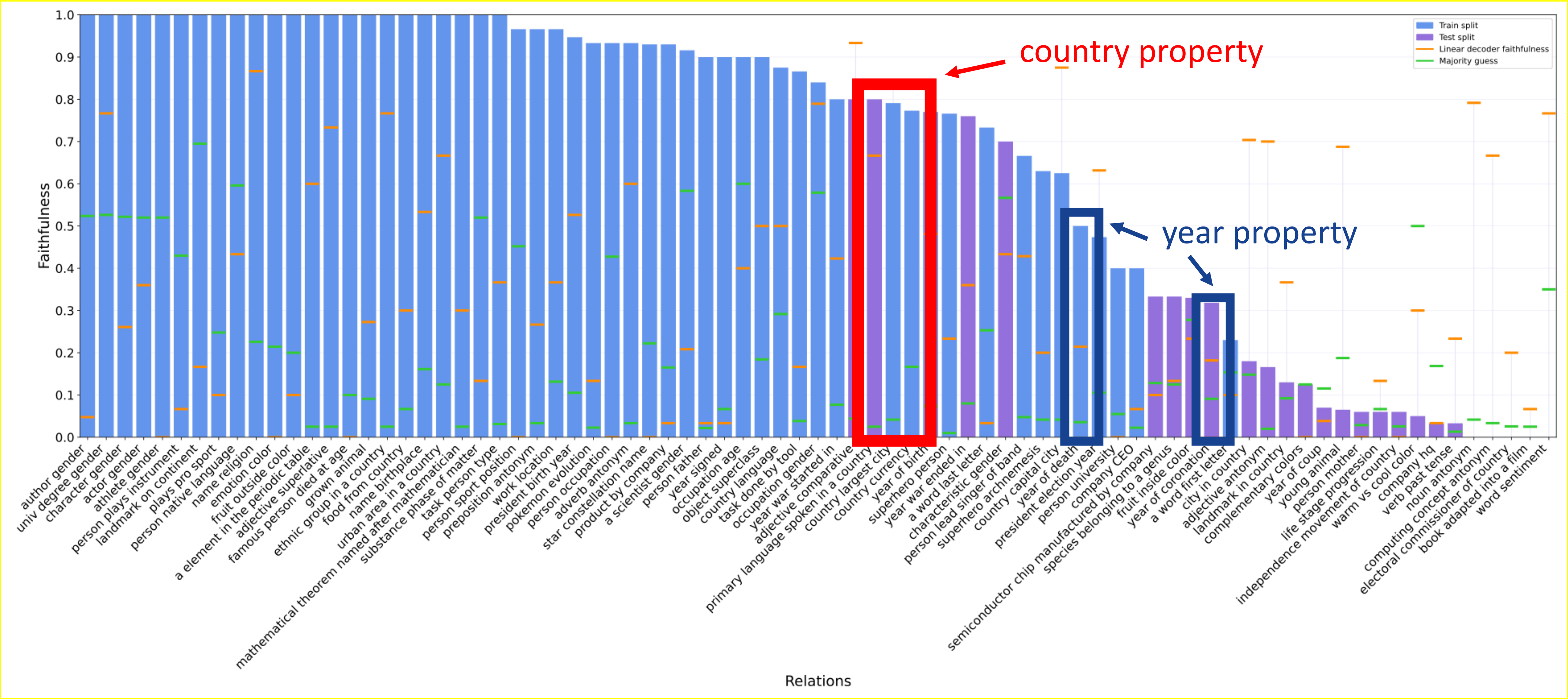
GPT NeoX 20B

Train set
Test set

Three Levels of Generalization

Held-out relations	Held-out relations corresponding to one property	Sample-wise held-out (subject, object)
City in country	*Number plus 6*	*1 plus 6 is 7*
Word sentiment	*Number minus 19*	*2 plus 6 is 8*
Adjective antony	*Number minus 3*	*3 plus 6 is 9*
Fruit inside color	*Number plus 21*	*4 plus 6 is 10*
Fruit outside color	*Number plus 9*	*5 plus 6 is 11*
Person mother	*Number minus 5*	*6 plus 6 is 12*

Generalization on Held-Out Relations



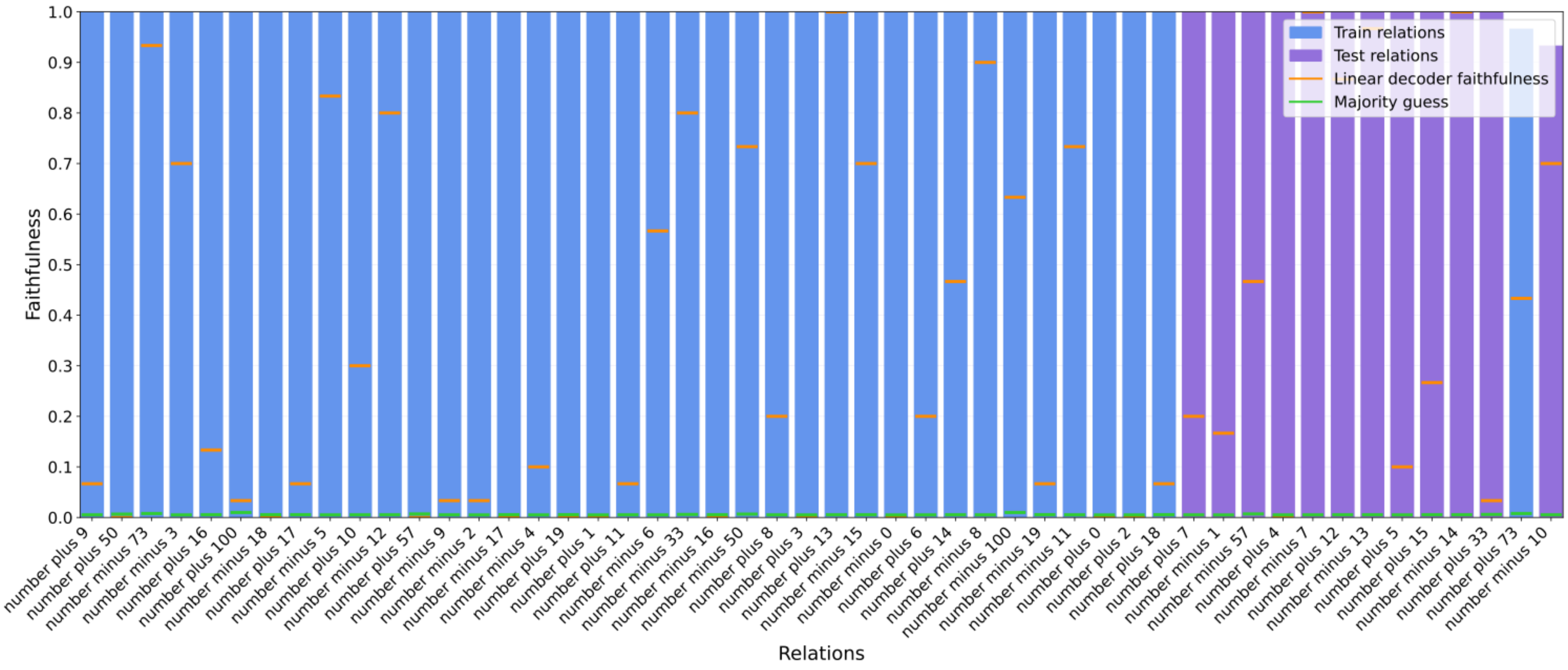
Generalization on Held-Out Relations Corresponding to One Property

Tightly controlled mathematical dataset

Number minus 5
Number plus 9
Number plus 21
Number minus 3
Number plus 6
Number minus 19

Generalization on Held-Out Relations Corresponding to One Property

99.1% test faithfulness!



Thank you!




Miranda Anna Christ
chrm@berkeley.edu

Adrián Csiszárik
csadrian@renyi.hu

Gergő Becsó
becsogergely@student.elte.hu

Dániel Varga
daniel@renyi.hu

CONTENTS

Mecenatúra

National Research, Development and Innovation Office
Hungary

PROGRAM
FINANCED FROM
THE NRDI FUND

HUNGARIAN MACHINE LEARNING DAYS

BUDAPEST, 12 – 14 AUGUST 2025

Mecenatúra

National Research, Development and Innovation Office
Hungary

PROGRAM
FINANCED FROM
THE NRDI FUND