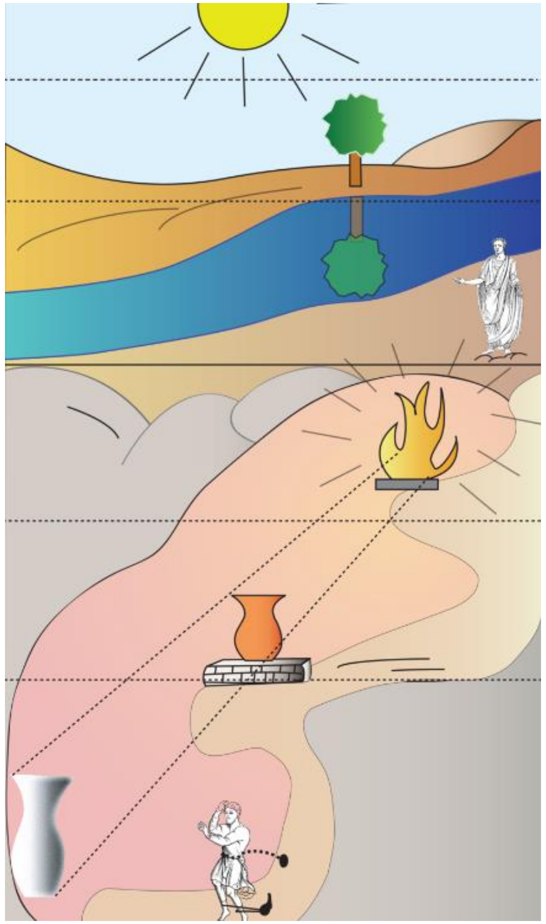
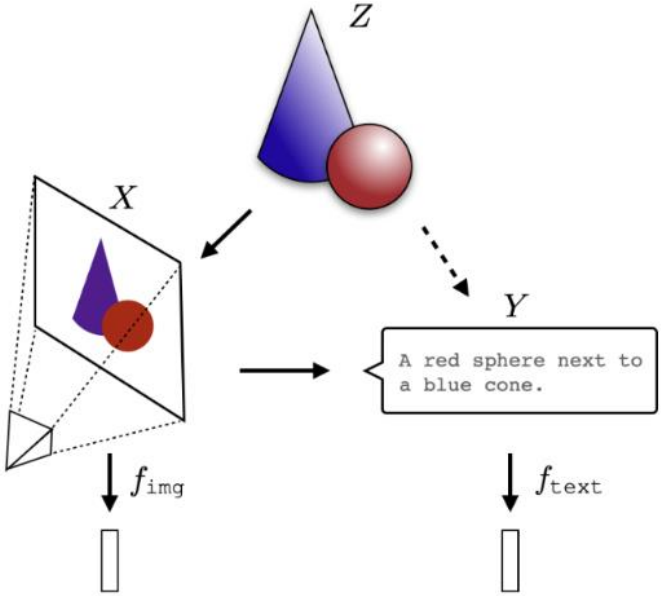
The Platonic Representation Hypothesis

Minyoung Huh*, Brian Cheung*, Tongzhou Wang*, Phillip Isola*. ICML2024-Oral





UBC MLRG
Yi (Joshua) Ren
25-July-2024

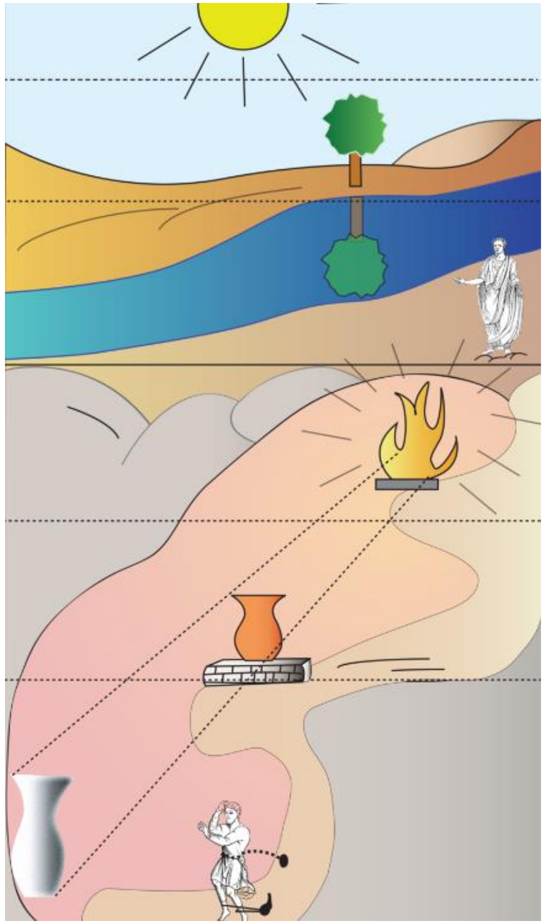
OUTLINES

1. What is the Platonic Representation Hypothesis
2. How they find that & Experimental supports
3. Why and how converge to GT
4. Limitations and Implications

OUTLINES

1. What is the Platonic Representation Hypothesis
2. How they find that & Experimental supports
3. Why and how converge to GT
4. Limitations and Implications

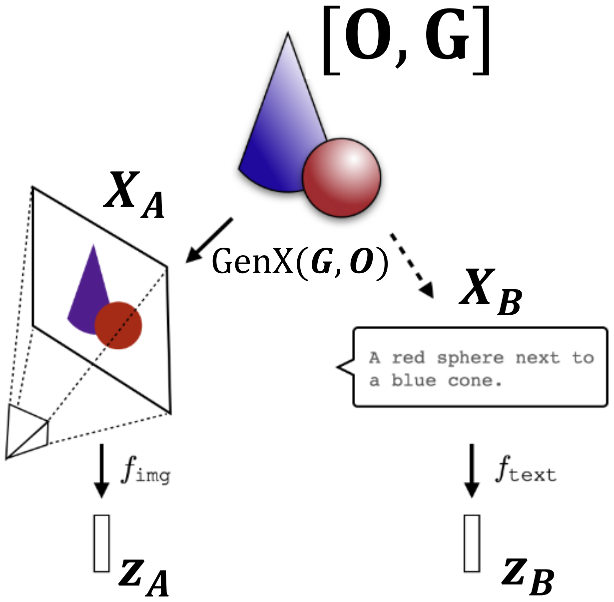
How we understand the world?



The Allegory of the cave

- There exist an unique ground truth (GT)
  - GT include the factors (the existance of specific objects)
  - GT include the mechanisms (the function projecting obj. to shadow)
- Our observations are the “shadow” of GT
  - Different objects have different shadow (generally)
  - Different light source create different shadow
- We use the observations to understand the world (GT)
  - We see, we learn, and we verify our knowledge
  - We have our own bias when learning
- We use our understanding to predict the future
  - The knowledge that have more accurate predictions are closer to GT
  - Our knowledge is also constraint by the observations

How the model understand the world?



ViT on IN1K GPT on language

A general machine learning system

✓ Assume the existence of stable $[\mathbf{O}, \mathbf{G}]$ and $\mathrm{GenX}(\mathbf{G}, \mathbf{O})$

- There exist an unique ground truth (GT)

✓ Observations $\boldsymbol{X_A}, \boldsymbol{X_B}$ of different modalities use different GenX

- Our observations are the "shadow" of GT

✓ We learn models $f: \mathcal{X} \rightarrow \mathcal{Z}$ to "guess" GT, ($\boldsymbol{Z}$ is our understanding)

- We use the observations to understand the world (GT)

✓ We use the learned function f to make predictions (with task head)

- We use our understanding to predict the future

Overview of this Platonic Representation Hypothesis

Assumptions

Different f tries their best to achieve good generalization ability

Observations on different modalities are different projections of the same GT

↓

f_A and f_B tend to become more similar

↓

The similarity lies in how they get representations, not the model

↓

The representations converge to GT

- Knowing GT means perfect generalization → better generalization means closer GT
- GT is unique, but have many **equivalent** Z
- Similarity is defined on function space, should consider the **equivalence** above
- Converge is a **trend**, not a **guarentee**

Q: How to define the similarity between representation spaces (IN1K vs CIFAR10)?

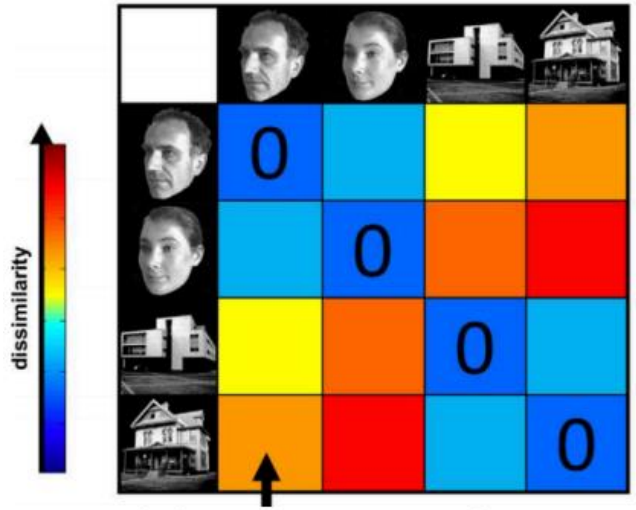
- Step 1: feature extractor to get dense representation $f: \mathcal{X} \to \mathbb{R}^n$

- Step 2: define the kernel measuring the similarity between the representations given two inputs, $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

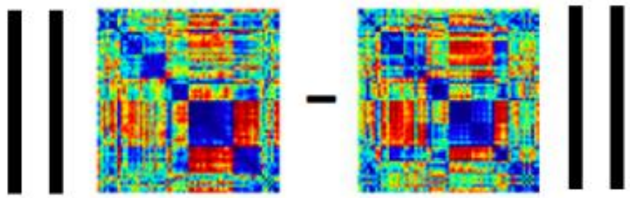
  E.g., the L2 distance, we have $K(x_i, x_j) = \|f(x_i) - f(x_j)\|_2^2$

- Step 3: for two different representation spaces (e.g., vision and language) measure their similarity using a kernel-alignment metric

$$\mathrm{m}: \mathcal{K} \times \mathcal{K} \to \mathbb{R}$$

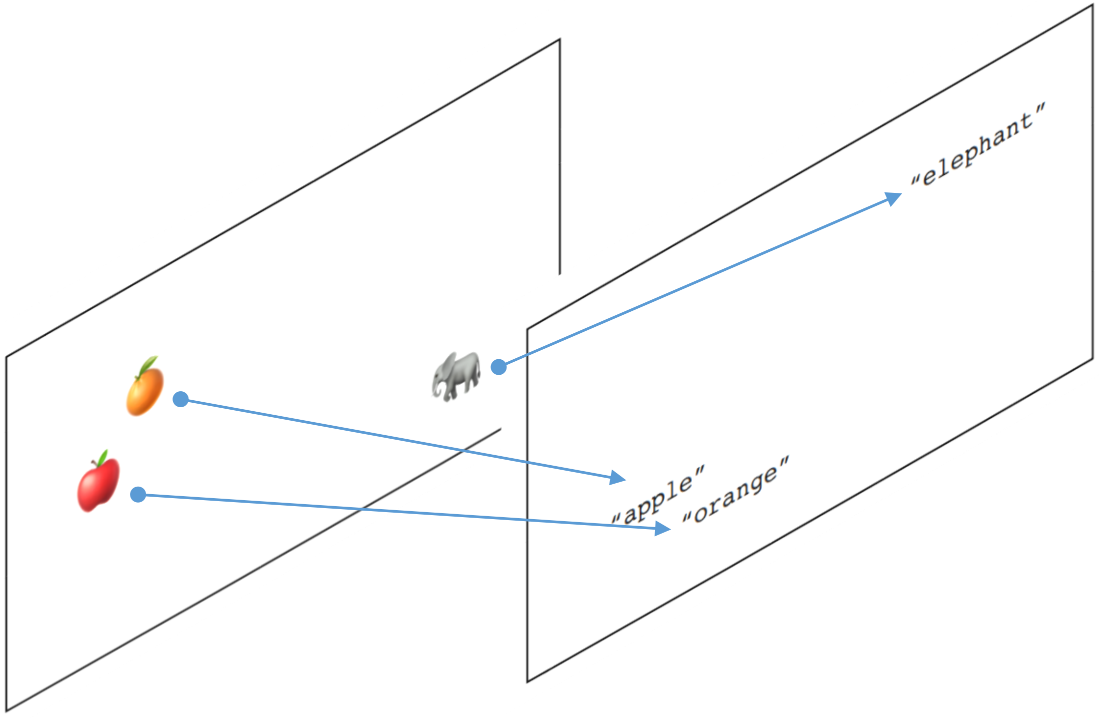


[fig from: Kriegeskorte, Mur, Ruff, et al. 2008]



1. What is the Platonic Representation Hypothesis

For example, topological similarity: $\mathbf{m}(\mathbf{K_A}, \mathbf{K_B}) \triangleq \mathbf{Corr}\left(\mathbf{K}\left(f_{text}\left(\mathbf{x}_\mathbf{A}^{(\mathbf{i})}\right), f_{img}\left(\mathbf{x}_\mathbf{A}^{(\mathbf{j})}\right)\right), \mathbf{K}\left(f_{text}\left(\mathbf{x}_\mathbf{B}^{(\mathbf{i})}\right), f_{img}\left(\mathbf{x}_\mathbf{B}^{(\mathbf{j})}\right)\right)\right)$



- Step 1: find feature extractors f_{text}, f_{img}
- Step 2: define their kernel as L2 distance on $\mathcal{Z}$ space

$f_{text}(\text{apple}) - f_{text}(\text{orange}) = 1$ $\qquad$ f_{img}(🍎) $- f_{img}$(🍊) $= 2$

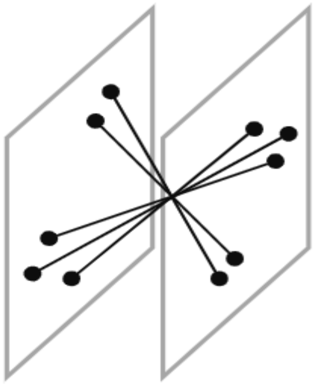
$f_{text}(\text{apple}) - f_{text}(\text{elephant}) = 10$ $\qquad$ f_{img}(🍎) $- f_{img}$(🐘) $= 15$

$f_{text}(\text{orange}) - f_{text}(\text{elephant}) = 8$ $\qquad$ f_{img}(🍊) $- f_{img}$(🐘) $= 12$

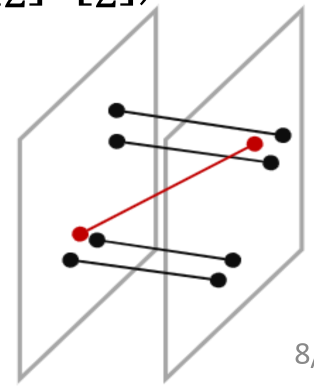
- Step 3: calculate their ranking correlation

$$\mathbf{m}(\mathbf{K_A}, \mathbf{K_B}) = \text{Spearman}\left(\begin{bmatrix}1\\3\\2\end{bmatrix}, \begin{bmatrix}1\\3\\2\end{bmatrix}\right) = 1$$

High $\mathbf{m}$:

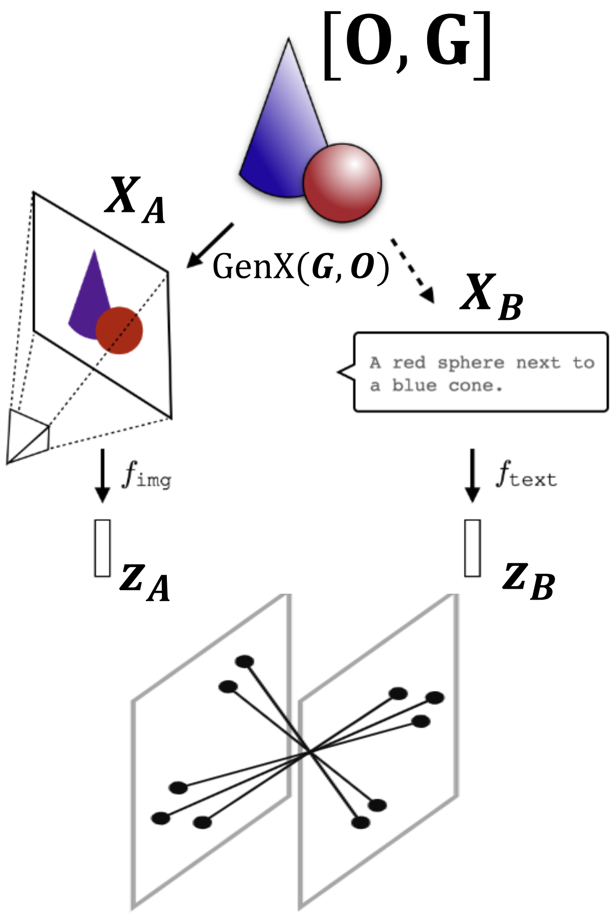


Low $\mathbf{m}$:



Summary:

- Althought trained **sepearately**, **independently**, with different datasets, target, models, etc.
- Still converge to similar representation space
- The converged structure is determined by GT



A, B are well-trained models → Big $\mathbf{m}(\mathbf{K_A}, \mathbf{K_G})$; $\mathbf{m}(\mathbf{K_B}, \mathbf{K_G})$ → Only one GT → Big $\mathbf{m}(\mathbf{K_A}, \mathbf{K_B})$

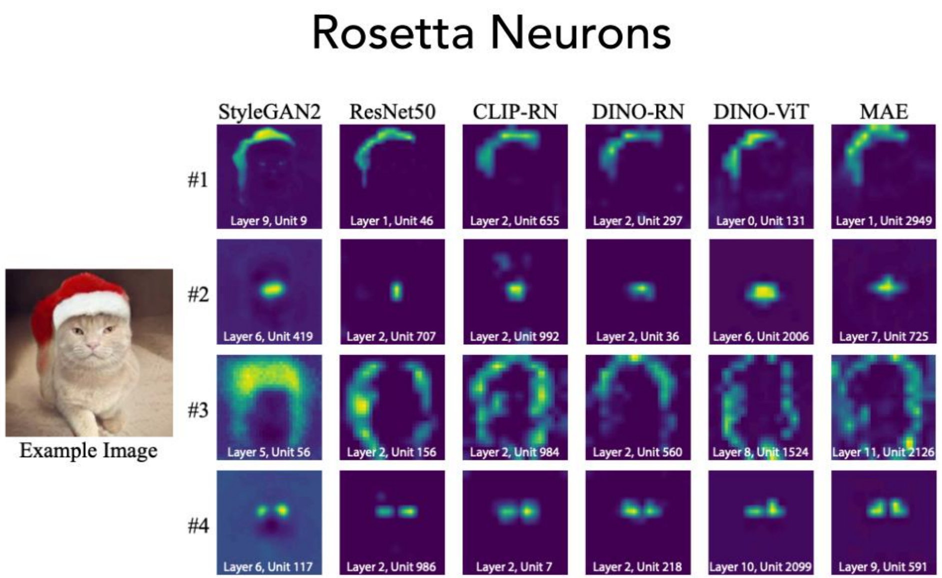
OUTLINES

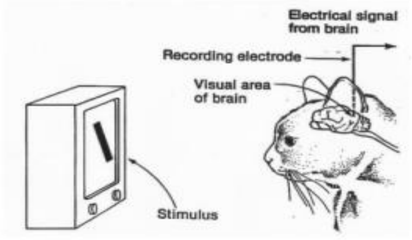
1. What is the Platonic Representation Hypothesis
2. How they find that & Experimental supports
3. Why and how converge to GT
4. Limitations and Implications

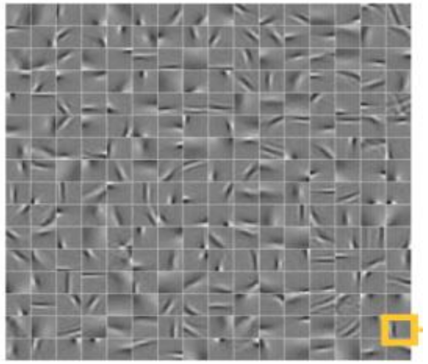
On a ***single modality***, good representations are shared among different tasks.

- Fact 1: common features across different **tasks**: pretrain + finetune style
- Fact 2: common features across different **models**
- Fact 3: common features across different **species**

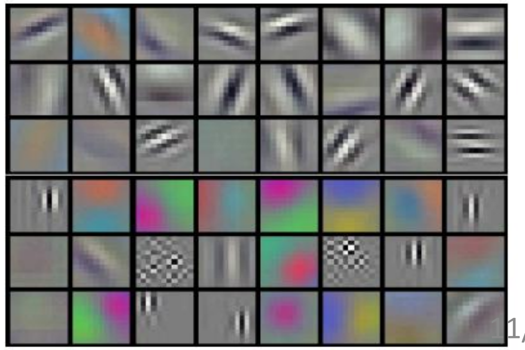
✓ Possible explanation: these good features are similar to GT







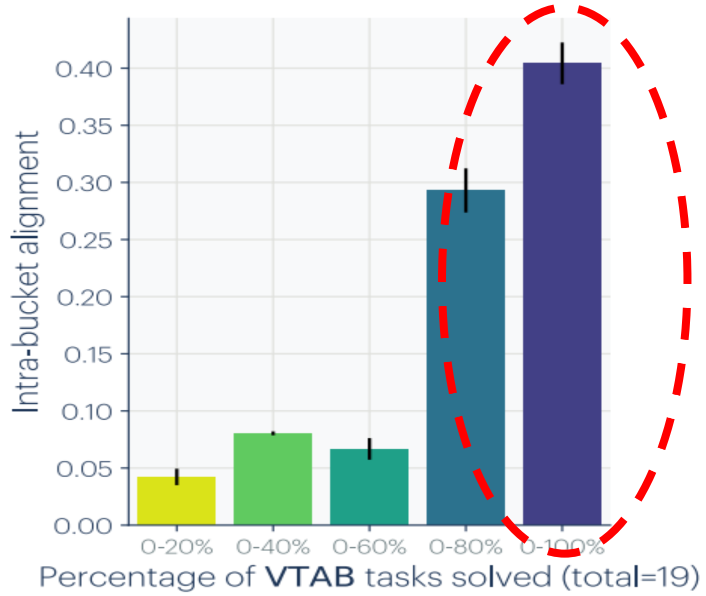
Filters in AlexNet

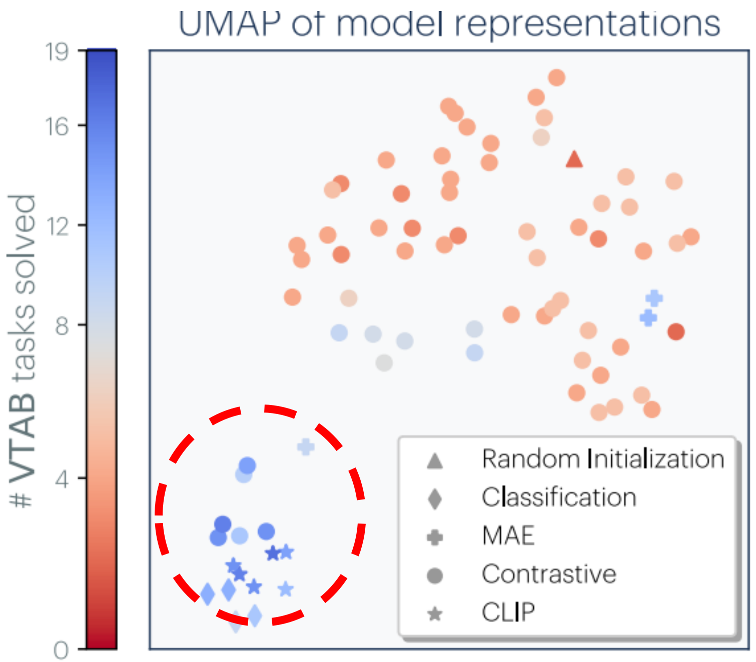


Experiments in the paper (single modality)

- Step 1: collect 78 vision models with **different** architectures (MLP, CNN, Transformer), objectives (classification, segmentation, SSL), training data distributions (CIFAR, IN21K, etc.)
- Step 2: fix $f(x)$ and train their linear head on 19 different VTAB tasks
- Step 3: calculate the representation alignment score $\mathbf{m}(\mathbf{K_A}, \mathbf{K_B})$ for all models
- Step 4: group their performance on VTAB

All strong representations are alike, each weak representation is weak in its own way.

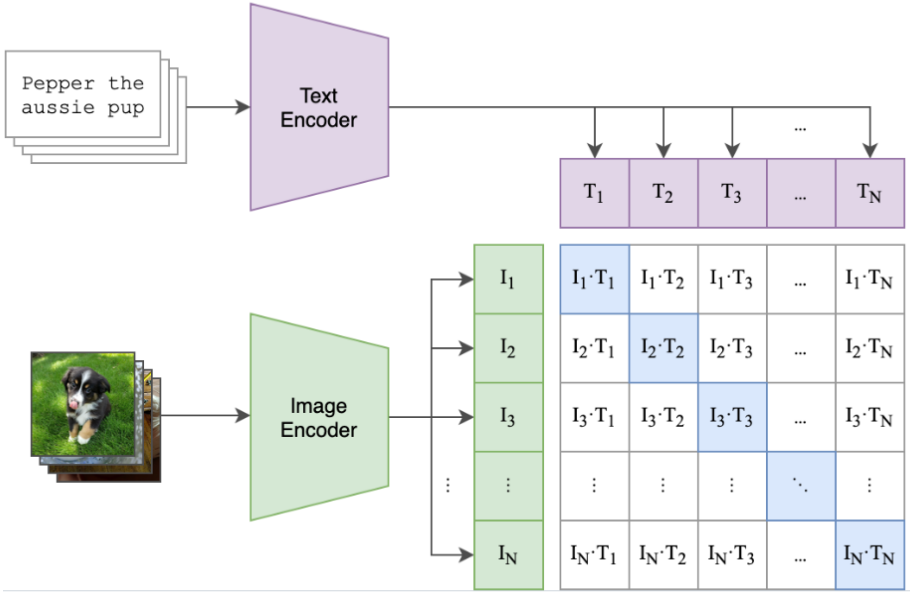




On ***multi-modality***, training together can bring benefits

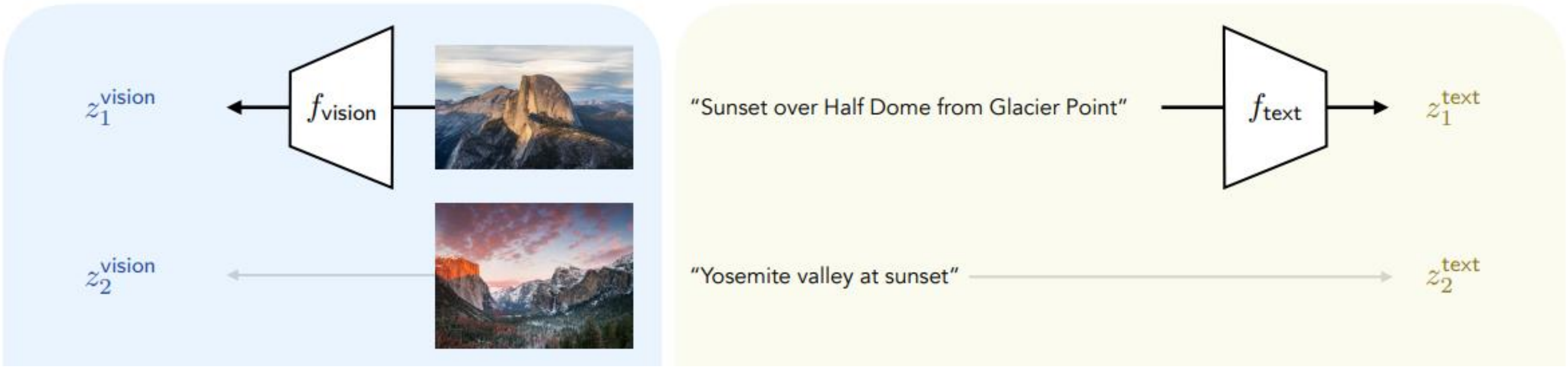
- Fact 1: CLIP is trained using paired image and language captions
- Fact 2: GPT4o and other SOTA LLM, VLM, claims they use multi-modal data
- Fact 3: carefully designed experiments in the paper

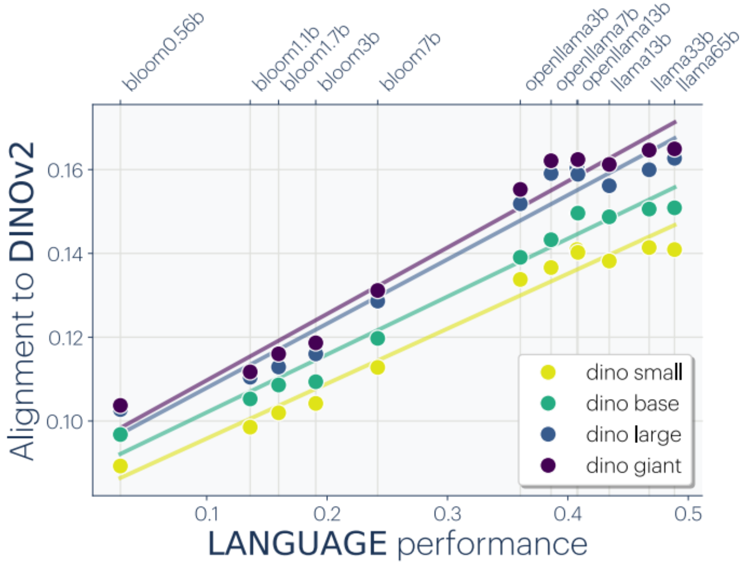
✓ Possible explanation: the features are similar to GT even for different modalities



Experiments in the paper (multiple modality)

- Step 1:, select **5** ViT models f_{img} (DINO, MAE, CLIP, etc.) and **11** LLMs f_{text}
- Step 2: on **wikipeadia image text dataset** generate the corresponding $z_{img}[i]$ and $z_{text}[i]$
- Step 3: measure the language performance using **log likelihood**
- Step 4: measure $\mathbf{m}(\mathbf{K}_{\mathbf{img}}, \mathbf{K}_{\mathbf{text}})$ of all 11 LLMs to each vision models to each ViT





Larger LLM ⇔ Better performance ⇔ Better Alignment with ViT
(Similar trend for MAE, CLIP, CLIP-ft, Supervised ViT)

Summary:

- On both single and multiple modalities, **representations** of good models **converge**
- The converged space is very likely to **represent GT** (since all models generalize well)
- Some advanced systems already applies multi-modal training

OUTLINES

1. What is the Platonic Representation Hypothesis
2. How they find that & Experimental supports
3. Why & how converge to GT
4. Limitations and Implications

Remember all models considered here are trained independently on their own modality, so they start with the following general loss.

$$\underbrace{f^*}_{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}} [\underbrace{\mathcal{L}}_{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

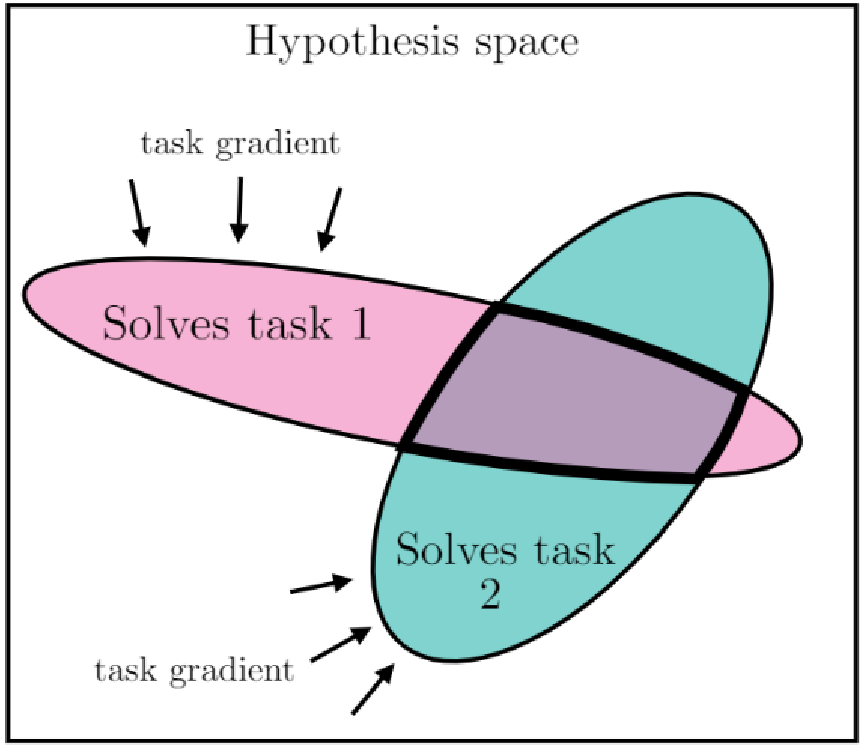
Based on that, they propose three possible explanations for the convergence:

A. Task Generality (GT can generalize to arbitrary tasks)
B. Model Capacity (GT require the model complex enough to encode GT)
C. Simplicity bias (GT is the simplest representation that explains all training examples)

3. Why & how converge to GT

A. Task Generality (GT can generalize to arbitrary tasks)

$$\overbrace{f^*}^{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}}[\overbrace{\mathcal{L}}^{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$



The Multitask Scaling Hypothesis

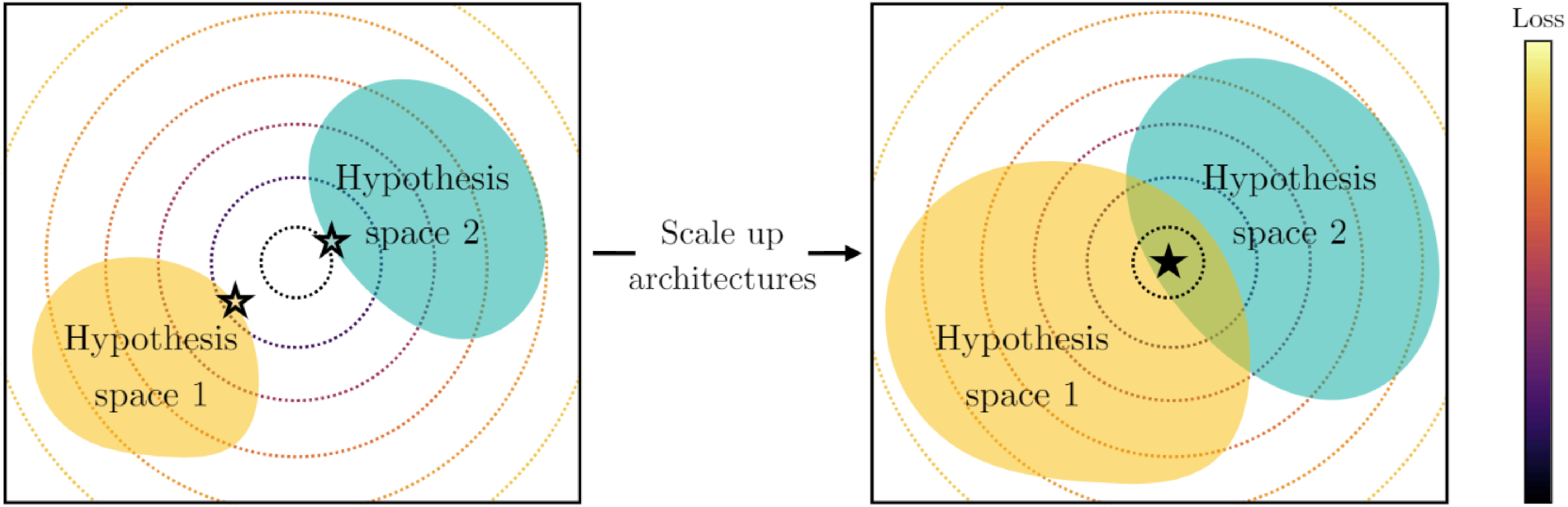
There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

GT must among these solutions, because we require the model generalize well on these N tasks.

3. Why & how converge to GT

$$\underbrace{f^*}_{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}}[\underbrace{\mathcal{L}}_{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

B. Model Capacity (GT require the model to be complex enough)



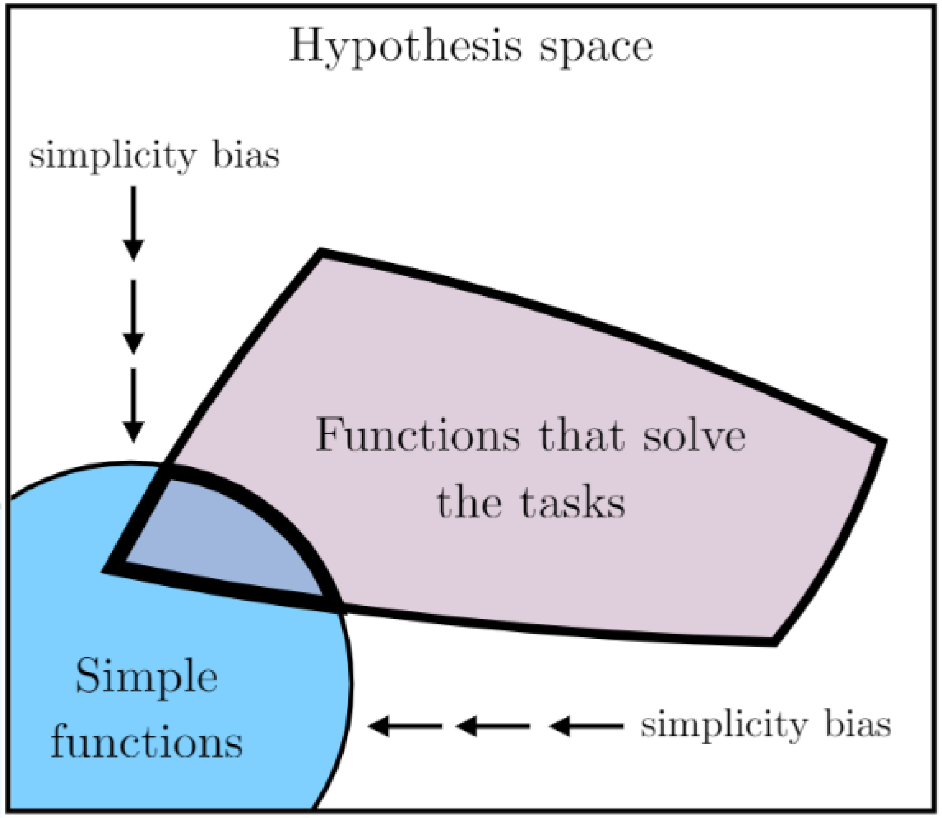
Bigger models are more likely to converge to a shared representation than smaller ones

3. Why & how converge to GT

$$\overbrace{f^*}^{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x\sim \text{dataset}}[\overbrace{\mathcal{L}}^{\text{training objective}}(f,x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

C. Simplicity bias (GT is the simplest representation that explains all training examples)

Spontaneously



> **The Simplicity Bias Hypothesis**
>
> Deep networks are biased toward finding simple fits to the data, and the bigger the model, the stronger the bias. Therefore, as models get bigger, we should expect convergence to a smaller solution space.

- GT has the smallest Kolmogorov complexity
- Model trained using GD natually favor that
- Learning dynamics and implicit regularization of SGD might be a possible explanation (happy to discuss later)
- More discussions in appendix AB of [1] and [2]

[1] Ren, Yi, et al. "Improving compositional generalization using iterated learning and simplicial embeddings." *NeurIPS 2023*
[2] Goldblum, Micah, et al. "The no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning." arXiv 2023

Summary:

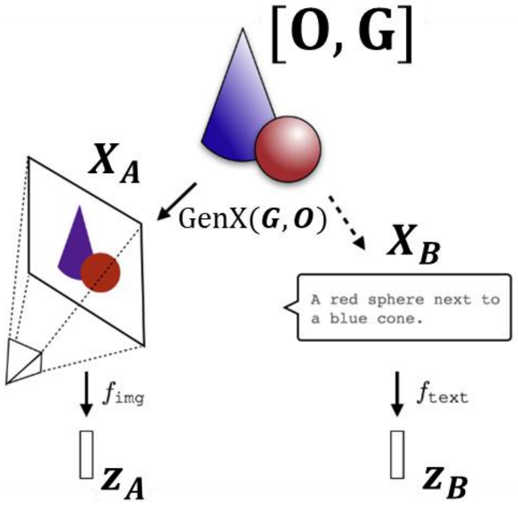
$$\overbrace{f^*}^{\text{trained model}} = \operatorname*{arg\,min}_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}} [\overbrace{\mathcal{L}}^{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

Based on that, they propose three possible explanations for the convergence:

A. Task Generality (GT can generalize to arbitrary tasks)
B. Model Capacity (GT require the model complex enough to encode GT)
C. Simplicity bias (GT is the simplest representation that explains all training examples)

OUTLINES

1. What is the Platonic Representation Hypothesis
2. How they find that & Experimental supports
3. Why & how converge to GT
4. Limitations and Implications



Limitations:

- GenX on two modalities needs more considerations
  - Degenerated dimensions across different modalities
    (e.g., GenX_A ignore the color while GenX_B ignore the shape)
  - They might describing fundamentally different information
    (Vision: a bird flying in the sky; Language: praise the freedom)

- Current models' alignment level is relatively low
  - In the paper, best alignment with DINOv2 is 0.16
    (but perfect alignment should be 1!)
  - Alignment need the data has more semantic overlap
    (but mainstream dataset cannot achieve that)

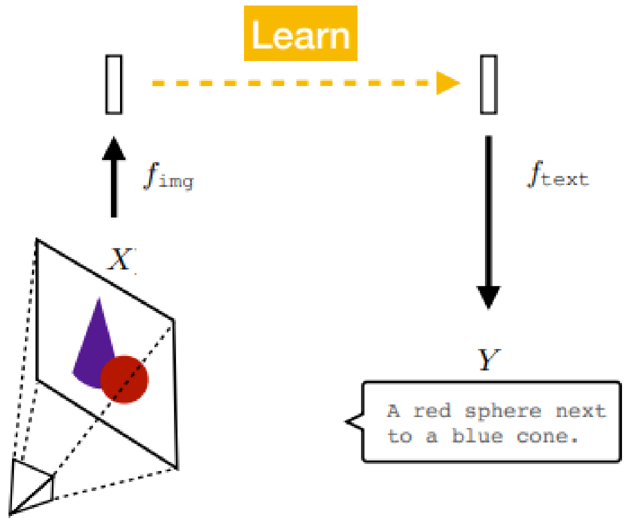
- No theory links all these pieces yet
  - Why simplicity bias exist? Relationship to K-Complexity?
  - How to formally describe this process, even on toyish setting?

Implications:

- All data modalities should help all model modalities
  - A word should be worth n pixels for training a vision model. A pixel should be worth m words for training an LLM.
  - Many multimodal works already show these benefits (e.g., LlaVA, GPT-4v, etc)
- Ease of cross-modal learning
  - A common representation can serve as a bridge for translation
  - Abundant paired data may be unnecessary for grounding [1]
- Good representation → Knowning GT → Uncover causality
  - Help us understand why model behave like this
  - Help us uncover more rules of the nature
  - Compression for AGI [2]

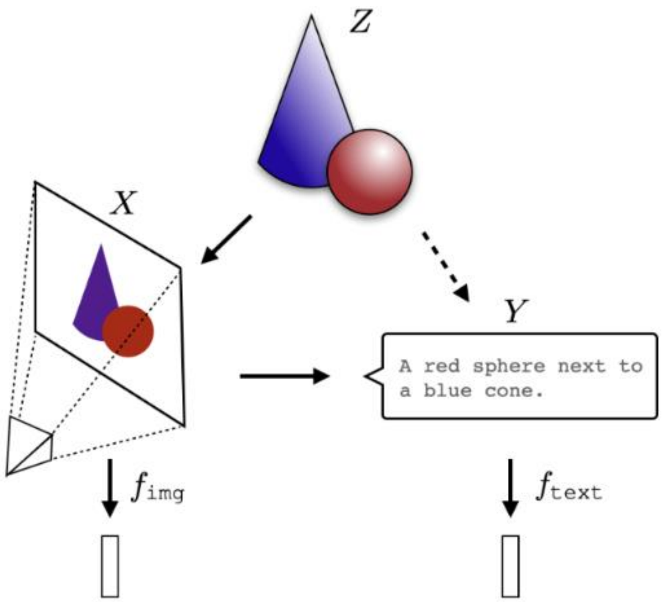
Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th

[https://openai.com/index/gpt-4-research/]



[1] Sorscher, Ben, Surya Ganguli, and Haim Sompolinsky. "Neural representational geometry underlies few-shot concept learning." PNAS 2022
[2] Jack Rae, Compression for AGI (https://www.youtube.com/watch?v=dO4TPJkeaaU)

Thanks for your attention



The slides borrow figures from:

- Huh, Minyoung, et al. "The platonic representation hypothesis." ICML – Oral 2024
- Their project page (https://phillipi.github.io/prh/)
- Their slides and talk at UCB (https://www.youtube.com/watch?v=1_xH2mUFpZw)