What Representational Similarity Measures Imply about Decodable Information

Sarah Harvey[1,2] David Lipshutz[1] Alex Williams[1,2]

FLATIRON INSTITUTE

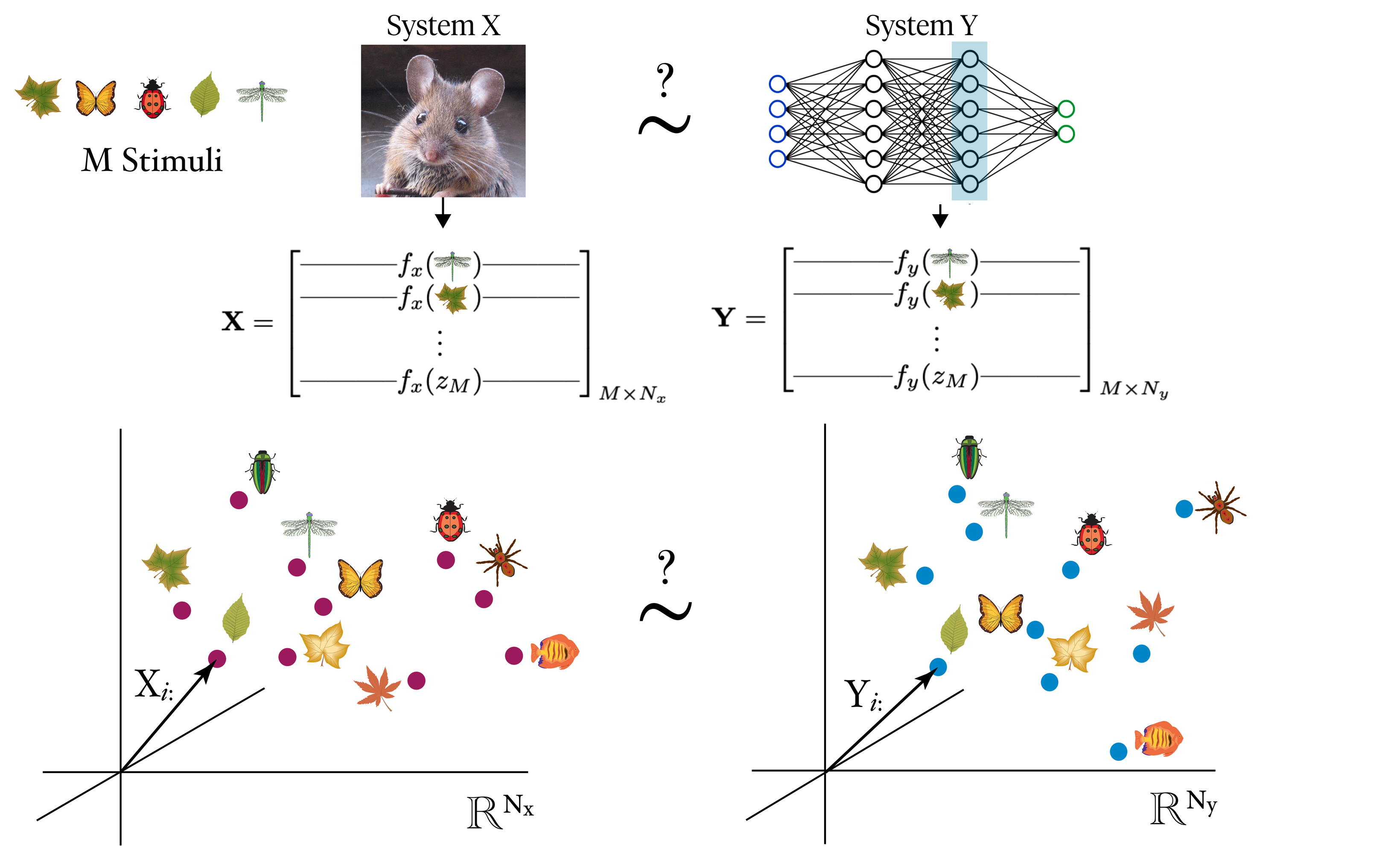
NYU NEUROSCIENCE

[1] Flatiron Institute, Center for Computational Neuroscience [2] New York University, Center for Neural Science

Measuring Representational Similarity

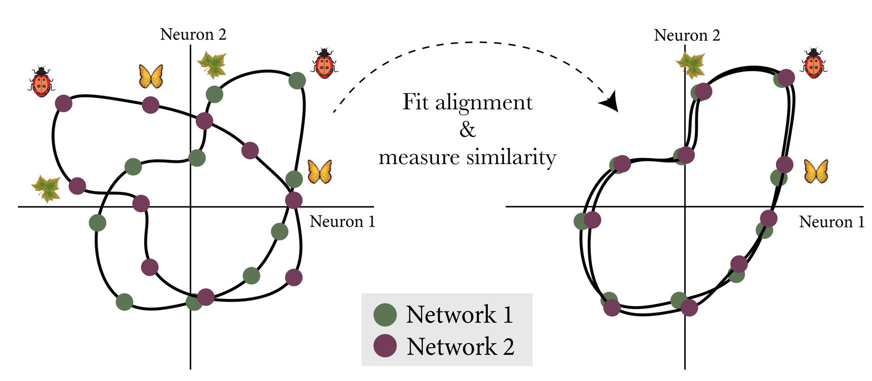
Comparative analyses are important tools for understanding complex systems

How do we quantify similarity between ***neural representations***?



Many Existing Approaches

(Dis)similarity measures that transform or align neural dimensions



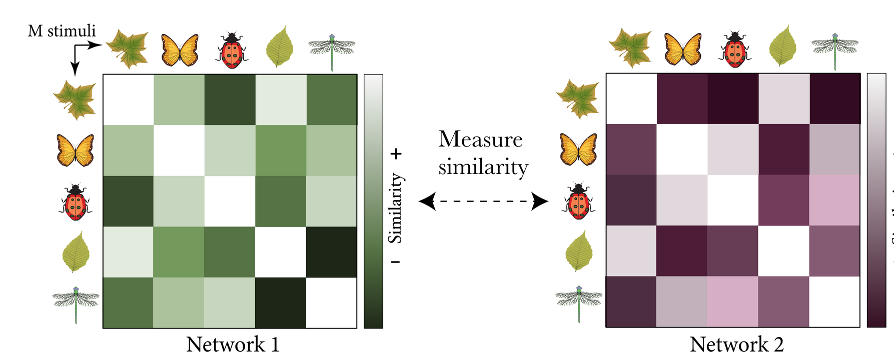
- Linear regression [1]
- Canonical correlations analysis [2]
- **Procrustes shape distances** [3]

$$\mathscr{P}(X,Y) = \min_{\mathbf{O}\in\mathcal{O}(N)} \|\mathbf{X}-\mathbf{YO}\|_F$$

Measure Euclidean distance between representations after rotational alignment

Takes values from 0 to ∞ (*distance*)

(Dis)similarity measures that quantify stimulus-by-stimulus relationships



- Representational Similarity Analysis (RSA) [4]
- Bures distance (Muzellec et. al. 2018) [6]
- **Centered Kernel Alignment (CKA)** [5]

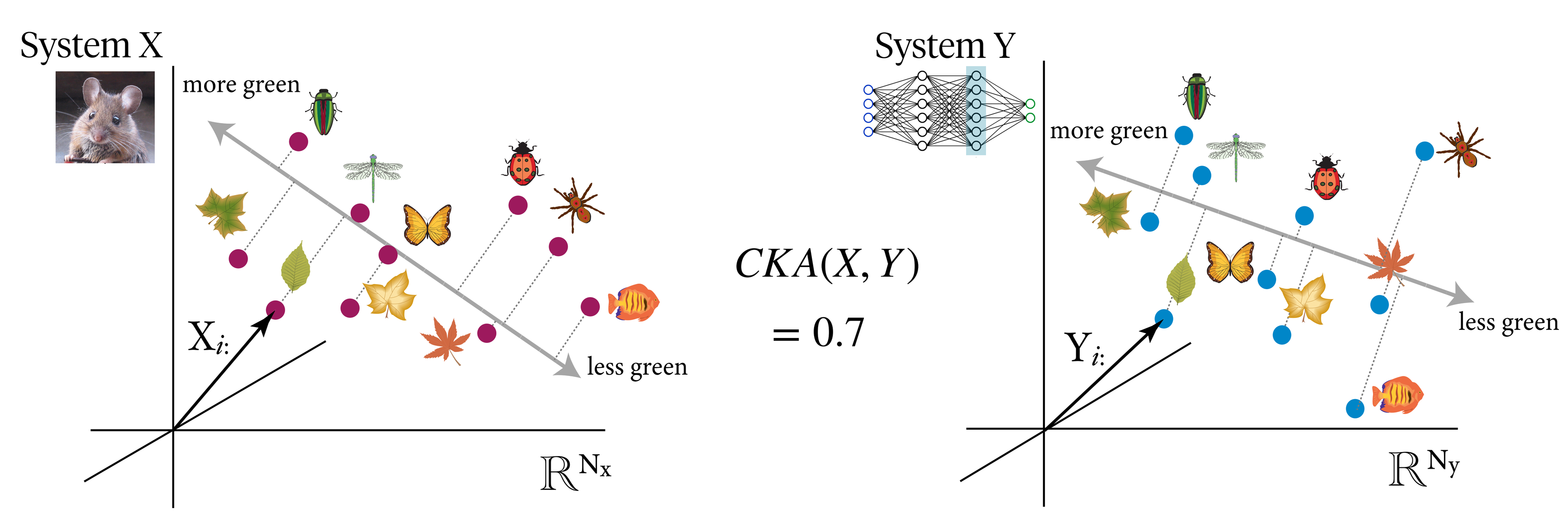
$$CKA(\mathbf{X},\mathbf{Y}) = \frac{\operatorname{Tr}\mathbf{XX}^\top\mathbf{YY}^\top}{\sqrt{\operatorname{Tr}(\mathbf{XX}^\top)^2\operatorname{Tr}(\mathbf{YY}^\top)^2}} \quad \textit{(Linear kernels)}$$

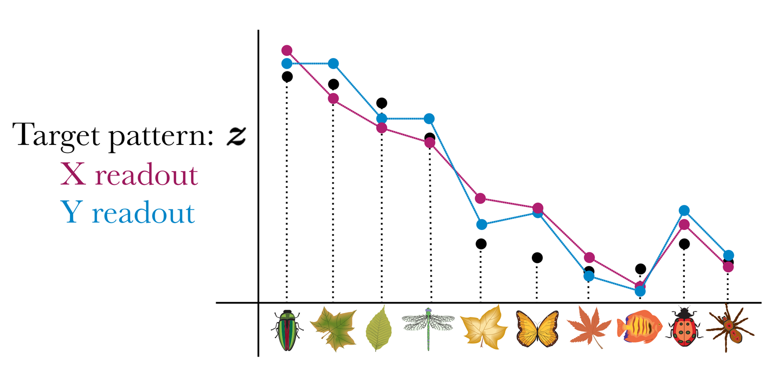
Compute inner product between centered representations and normalize

Takes values from 0 to 1 (*similarity*)

Comparing representations with linear decoding

Example task 1: color



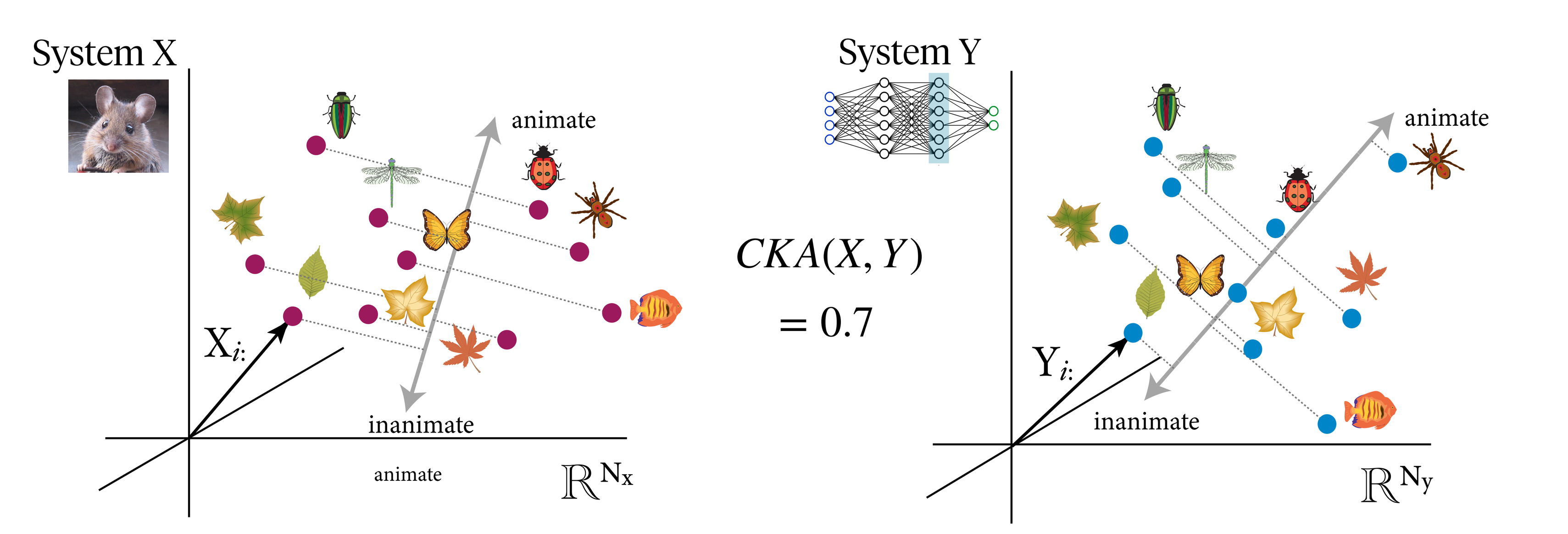


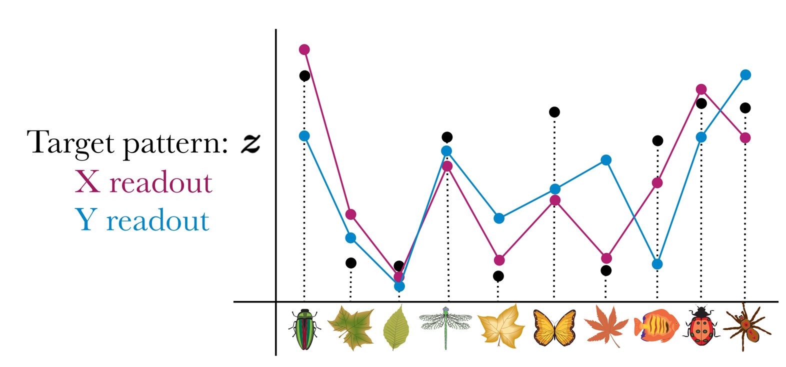
Compute optimal weights for system X ⟹ Decoded pattern: $\boldsymbol{Xw}^*$

Compute optimal weights for system Y ⟹ Decoded pattern: $\boldsymbol{Yv}^*$

Measure similarity: (Scaled to be unit vectors) $\langle \boldsymbol{Xw}^*, \boldsymbol{Yv}^*\rangle = 0.82$

Example task 2: animate/inanimate





Compute optimal weights for system X ⟹ Decoded pattern: $\boldsymbol{Xw}^*$

Compute optimal weights for system Y ⟹ Decoded pattern: $\boldsymbol{Yv}^*$

Measure similarity: (Scaled to be unit vectors) $\langle \boldsymbol{Xw}^*, \boldsymbol{Yv}^*\rangle = 0.63$

Average decoding similarity/distance

★ Similarity depends on the choice of decoding task

Idea: Measure similarity over an ensemble of decoding tasks

average decoding similarity (ADS)

$$\mathbb{E}_{z\sim P_z}\langle \boldsymbol{Xw}^*, \boldsymbol{Yv}^*\rangle$$

average decoding distance (ADD)

$$\mathbb{E}_{z\sim P_z}\|\boldsymbol{Xw}^* - \boldsymbol{Yv}^*\|_2^2$$

To compute these, we must choose:

1. Regression loss function
2. Ensemble of tasks to average over

We show: certain choices here ⟹ average decoding similarity/distance = popular representational similarity/distance measures

General mathematical framework

➡ Set up a family of linear decoding problems

Decoding optimization problem:

$$\underset{\boldsymbol{w}}{\text{maximize}} \quad \underbrace{\tfrac{1}{M}\boldsymbol{z}^\top\boldsymbol{Xw}}_{\substack{\textit{Maximize overlap} \\ \textit{between } \mathbf{Xw} \textit{ and } \mathbf{z}}} - \underbrace{\tfrac{1}{2}\boldsymbol{w}^\top\boldsymbol{G}(\boldsymbol{X})\boldsymbol{w}}_{\substack{\textit{Penalty on a norm of} \\ \textit{the weights } \mathbf{w}}}$$

This problem has a nice closed form solution: $\boldsymbol{w}^* = \frac{1}{M}\boldsymbol{G}(\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{z}$ *Optimal Decoding Weights*

$\mathbf{G}(\cdot)$ is a function mapping $\mathbb{R}^{M\times N} \rightarrow$ symmetric positive definite $N\times N$ matrices

Consider $\boldsymbol{G}(\boldsymbol{X}) = a\boldsymbol{C}_X + b\boldsymbol{I}$ with neuron-by-neuron covariance $\boldsymbol{C}_X := \frac{1}{M}\boldsymbol{X}^\top\boldsymbol{X}$

Relations to geometric similarity measures

Special cases

Take $a = 1$, $b = \lambda$ ➡

$$\mathbf{w}^* = \operatorname{argmin}\ \|\mathbf{Xw}-\mathbf{z}\|_2^2 + \lambda\mathbf{I}$$
$$\mathbf{v}^* = \operatorname{argmin}\ \|\mathbf{Yw}-\mathbf{z}\|_2^2 + \lambda\mathbf{I}$$

Ridge regression

If we also assume identity covariance structure of the task ensemble we are averaging over, i.e. $\mathbb{E}_z\left[\mathbf{zz}^\top\right] = \mathbf{I}$

Take $a = 0$, $b = 1$ ➡

$$\frac{\mathbb{E}\langle\boldsymbol{Xw}^*,\boldsymbol{Yv}^*\rangle}{\sqrt{\mathbb{E}\langle\boldsymbol{Xw}^*,\boldsymbol{Xw}^*\rangle\mathbb{E}\langle\boldsymbol{Yv}^*,\boldsymbol{Yv}^*\rangle}} = \mathrm{CKA}(\boldsymbol{X},\boldsymbol{Y})$$

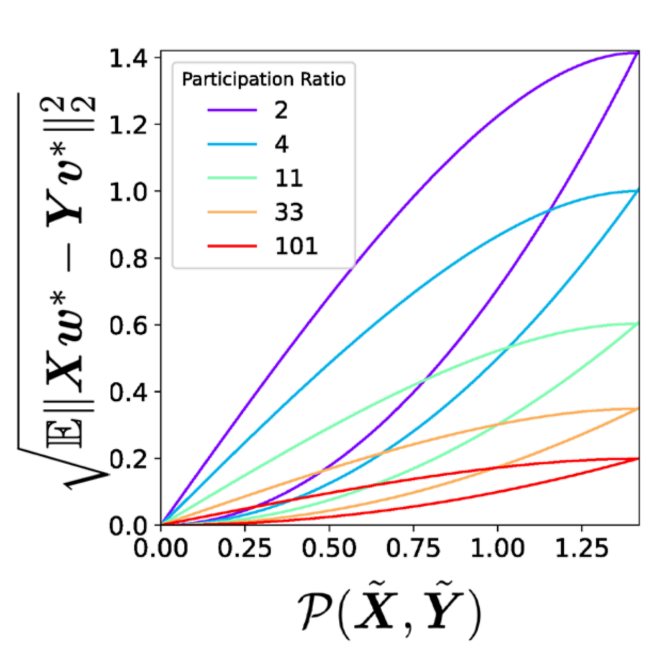
(Normalized) Average decoding similarity

More in paper:

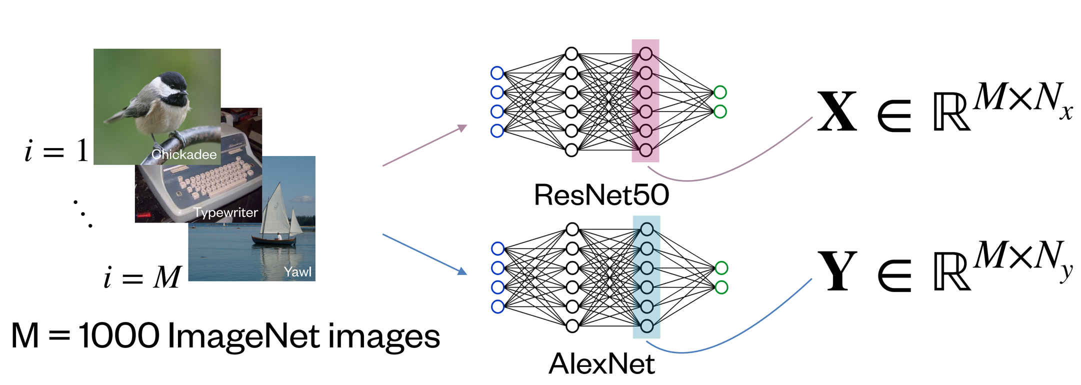
Similarity measure	a	b
Linear CKA	0	b
GULP	1	λ
CCA	1	0
ENSD	0	$\frac{1}{M}\frac{\mathrm{Tr}[\boldsymbol{C}_X^2]}{\mathrm{Tr}[\boldsymbol{C}_X]}$

We find saturated bounds relating *average decoding distance* to *Procrustes shape distance* between normalized representations

$$\tilde{\boldsymbol{X}} := \tfrac{1}{\sqrt{M}}\boldsymbol{XG}(\boldsymbol{X})^{-1/2} \quad \textit{and} \quad \tilde{\boldsymbol{Y}} := \tfrac{1}{\sqrt{M}}\boldsymbol{YG}(\boldsymbol{Y})^{-1/2}$$



Example: CKA

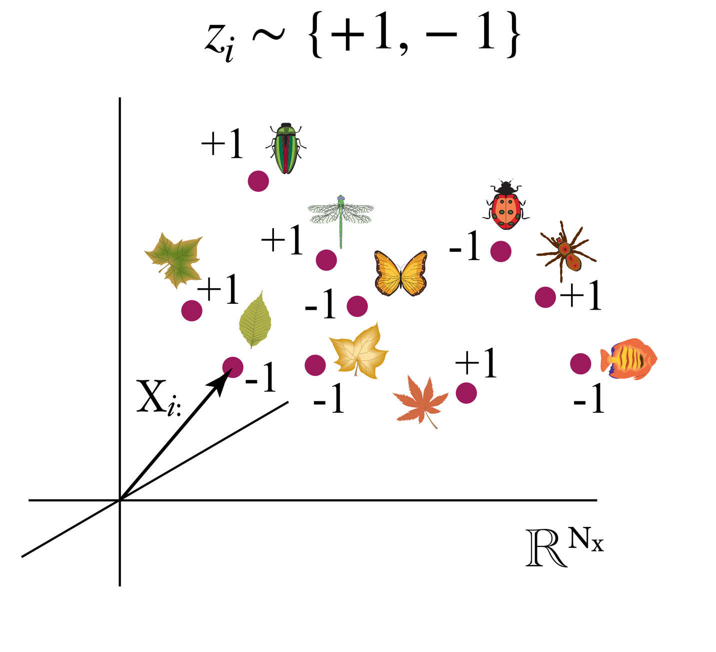


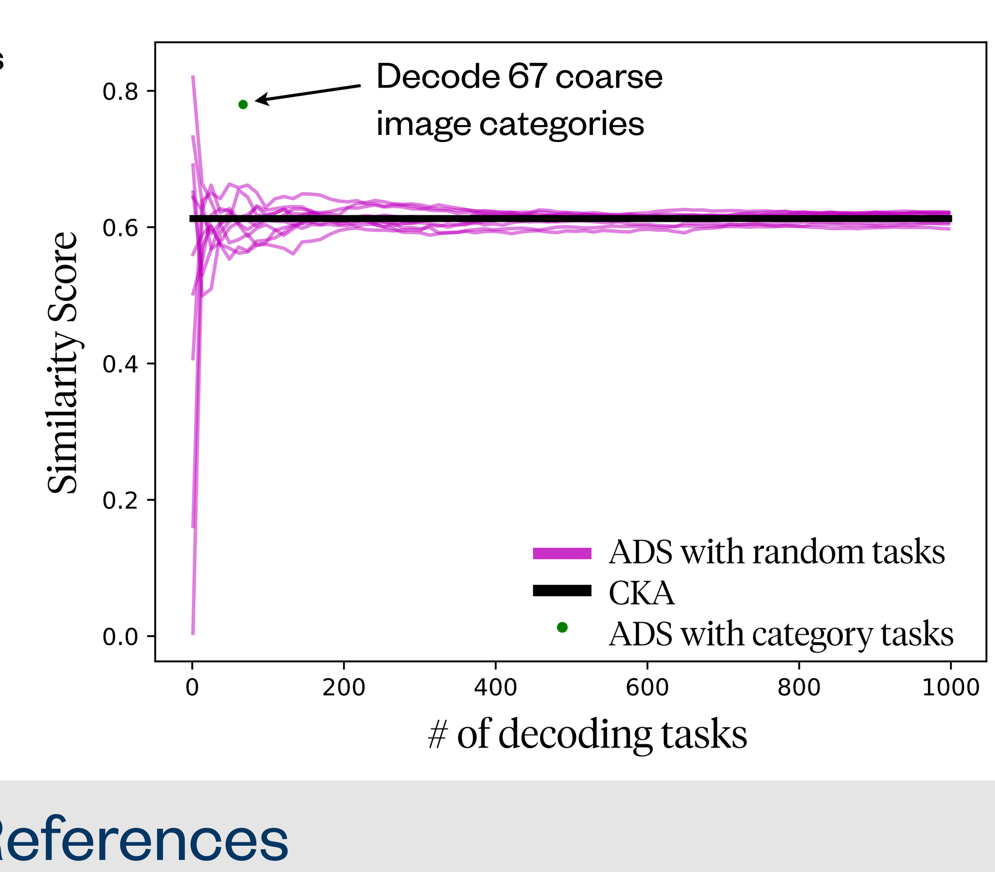
Compute $CKA(X,Y)$ and average decoding similarity $ADS(X,Y)$ with $a = 0, b = 1$

Ensemble of decoding tasks:

‣ Random binary partitions of input images

$$z_i \sim \{+1, -1\}$$





Links

Paper: arXiv:2411.08197

sharvey@flatironinstitute.org

Authors:
@sarah-harvey.bsky.social
@lipshutz.bsky.social
@itsneuronal.bsky.social

References

1. Yamins, D., DiCarlo, J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* 19, 356–365 (2016).
2. Raghu, M., et. al. "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability". NeurIPS 30 (2017).
3. Williams, A. H., et. al.. "Generalized Shape Metrics on Neural Representations". NeurIPS. Vol. 34. (2021).
4. Kriegeskorte N, et. al.. Representational similarity analysis-connecting the branches of systems neuroscience. Front Syst Neurosci. 2008; 2:4.
5. Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. "Similarity of Neural Network Representations Revisited".ICML. Vol. 97. (2019).
6. Muzellec, B., and Cuturi, M. "Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions". NeurIPS (2018).