# What Representational Similarity Measures Imply about Decodable Information


FLATIRON INSTITUTE

Sarah Harvey[1,3] David Lipshutz[1,2] Alex Williams[1,3]

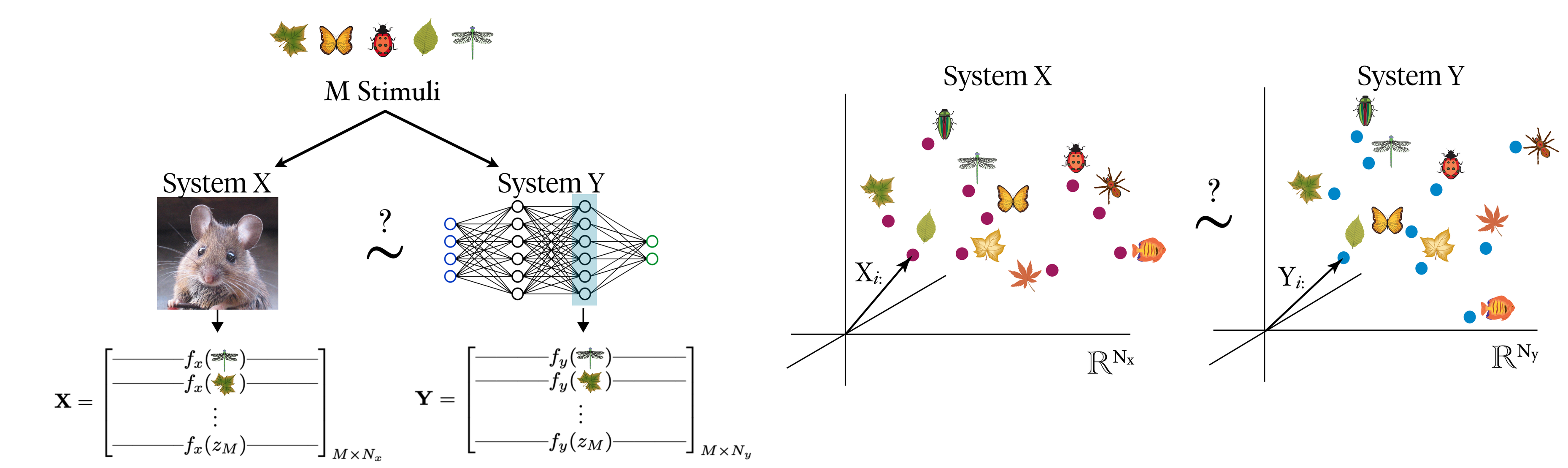
NYU NEUROSCIENCE

[1] Flatiron Institute, Center for Computational Neuroscience [2] Baylor College of Medicine [3] New York University, Center for Neural Science


## Measuring Representational Similarity

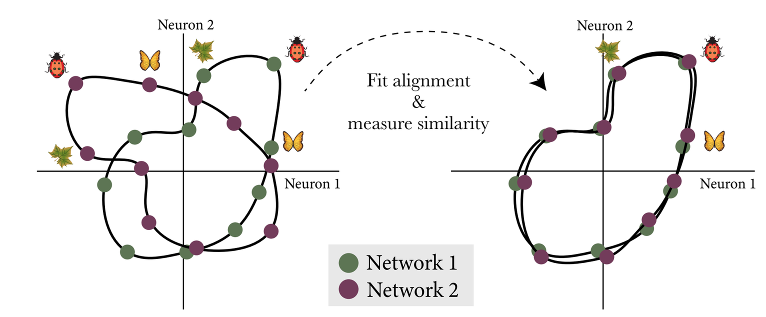
**Comparative analyses are important tools for understanding complex systems**

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




## Many Existing Approaches

**(Dis)similarity measures that transform or align neural dimensions**



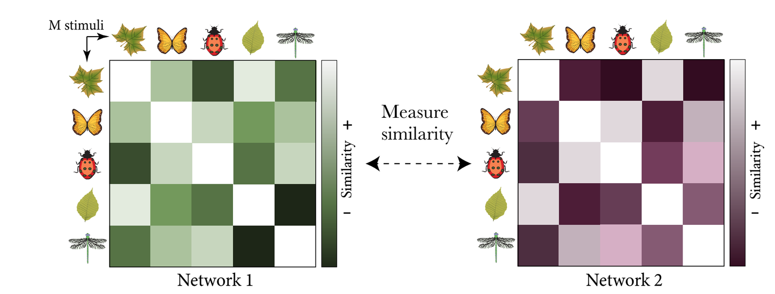

- Linear regression [1]
- Canonical correlations analysis [2]
- **Procrustes shape distances** [3]

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

*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]

$$CKA(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{Tr}\, \mathbf{X}\mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top}{\sqrt{\mathrm{Tr}\,(\mathbf{X}\mathbf{X}^\top)^2\, \mathrm{Tr}\,(\mathbf{Y}\mathbf{Y}^\top)^2}}$$

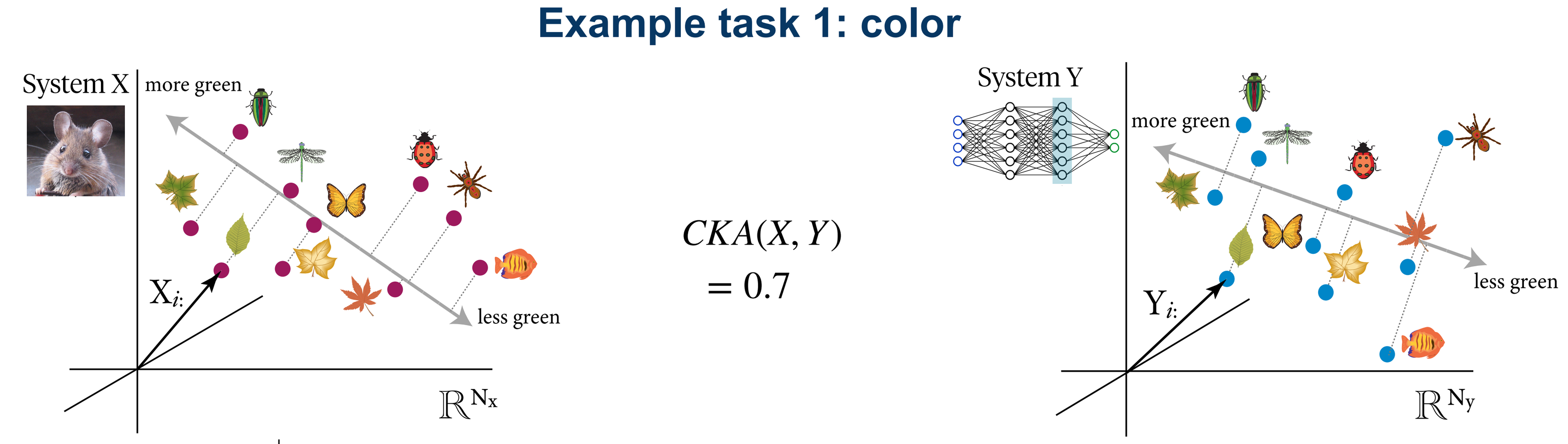
*(Linear kernels)*

*Compute inner product between centered representations and normalize*

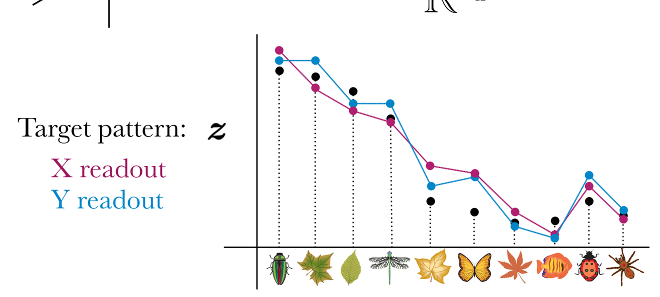
*Takes values from 0 to 1 (similarity)*

## Comparing representations with linear decoding

### Example task 1: color




$CKA(X,Y)$
$= 0.7$

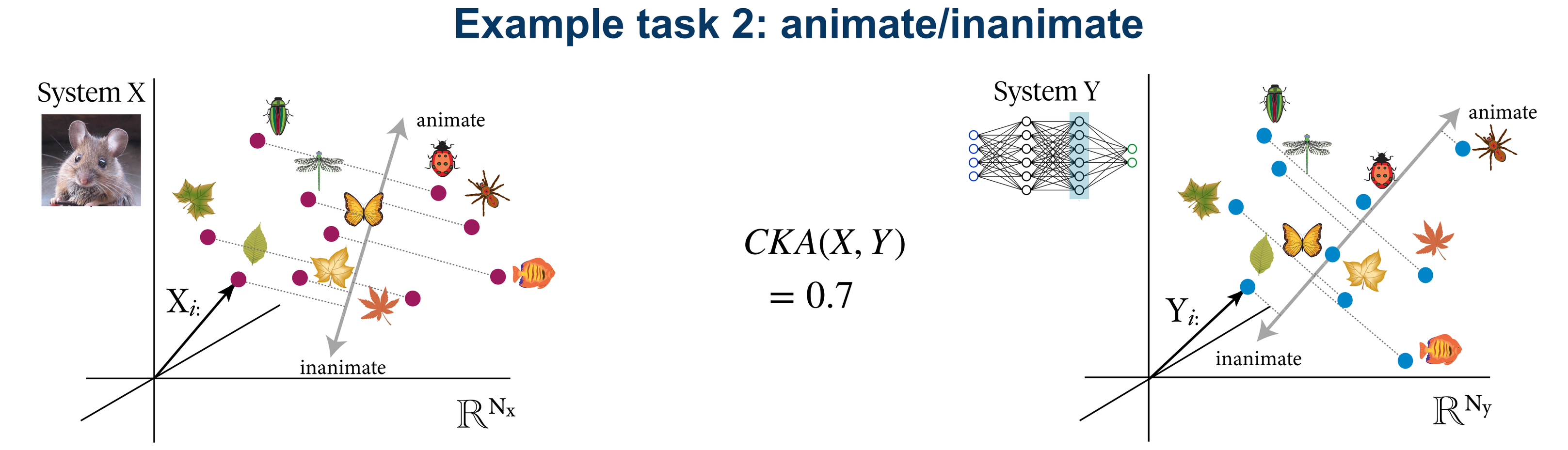



Compute optimal weights for system $X$
$\Longrightarrow$ Decoded pattern: $\boldsymbol{X}\boldsymbol{w}^*$

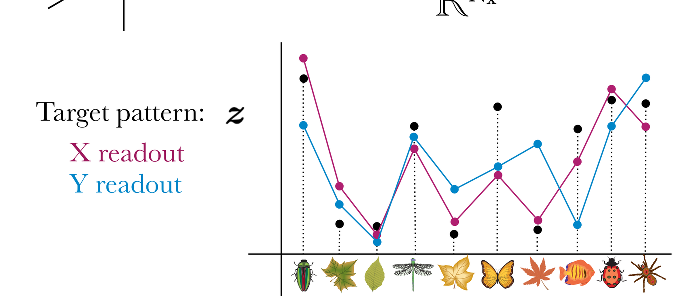
Compute optimal weights for system $Y$
$\Longrightarrow$ Decoded pattern: $\boldsymbol{Y}\boldsymbol{v}^*$

Measure similarity:
(Scaled to be unit vectors)
$\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle$
$= 0.82$

### Example task 2: animate/inanimate




$CKA(X,Y)$
$= 0.7$




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

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

Measure similarity:
(Scaled to be unit vectors)
$\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle$
$= 0.63$

## 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)*

$$\mathbb{E}_{z \sim P_z} \langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle$$

*average decoding distance (ADD)*

$$\mathbb{E}_{z \sim P_z} \|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2$$

To compute these, we must choose:

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

**We show:** certain choices here $\Longrightarrow$ average decoding similarity/distance = popular representational similarity/distance measures

## General mathematical framework

➡ Set up a family of linear decoding problems

Decoding optimization problem:

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

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 $\mathbb{R}^{M \times N} \to$ 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 $a = 1$, $b = \lambda$ ➡

$$\mathbf{w}^* = \text{argmin } \|\mathbf{X}\mathbf{w} - \mathbf{z}\|_2^2 + \lambda \mathbf{I}$$
$$\mathbf{v}^* = \text{argmin } \|\mathbf{Y}\mathbf{w} - \mathbf{z}\|_2^2 + \lambda \mathbf{I}$$

*Ridge regression*

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

Take $a = 0$, $b = 1$ ➡

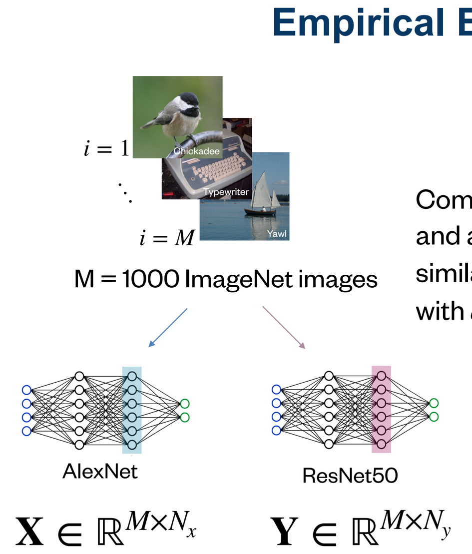
$$\frac{\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle}{\sqrt{\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{X}\boldsymbol{w}^* \rangle \mathbb{E}\langle \boldsymbol{Y}\boldsymbol{v}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle}} = \mathrm{CKA}(\boldsymbol{X},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

*(Normalized) Average decoding similarity*

**More in paper:**

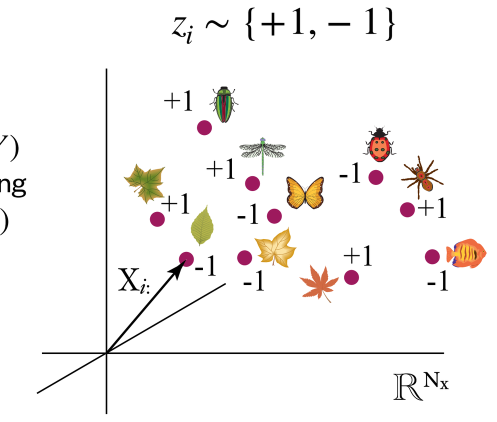
| Similarity measure | $a$ | $b$ |
|---|---|---|
| Linear CKA | 0 | $b$ |
| GULP | 1 | $\lambda$ |
| CCA | 1 | 0 |
| ENSD | 0 | $\frac{1}{M}\frac{\mathrm{Tr}[\boldsymbol{C}_X^2]}{\mathrm{Tr}[\boldsymbol{C}_X]}$ |

### Empirical Example: CKA

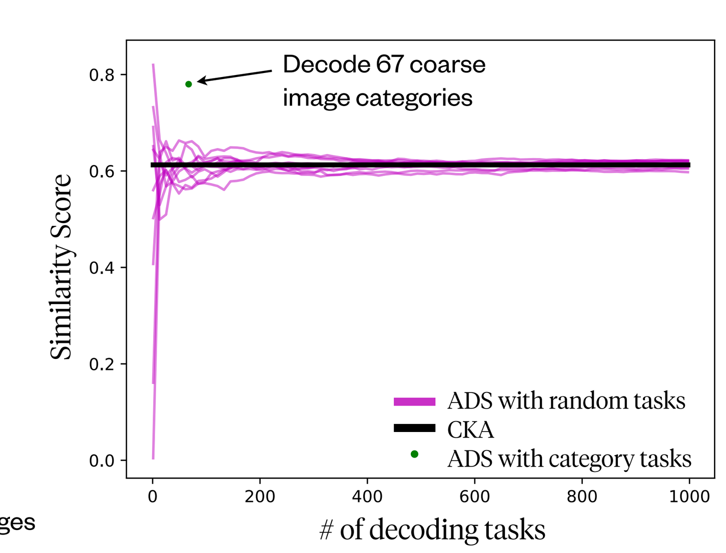



Compute $CKA(X,Y)$ and average decoding similarity $ADS(X,Y)$ with $a = 0, b = 1$

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

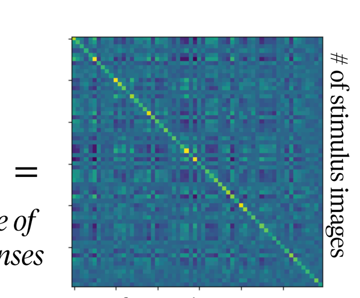


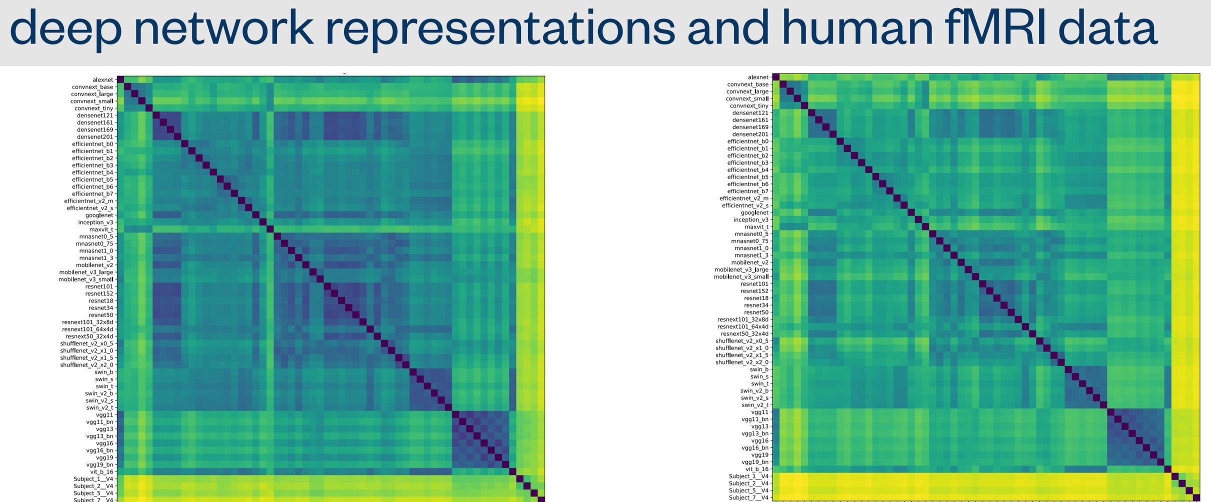

Ensemble of decoding tasks:
‣ Random binary partitions of input images




## Example: Comparing deep network representations and human fMRI data

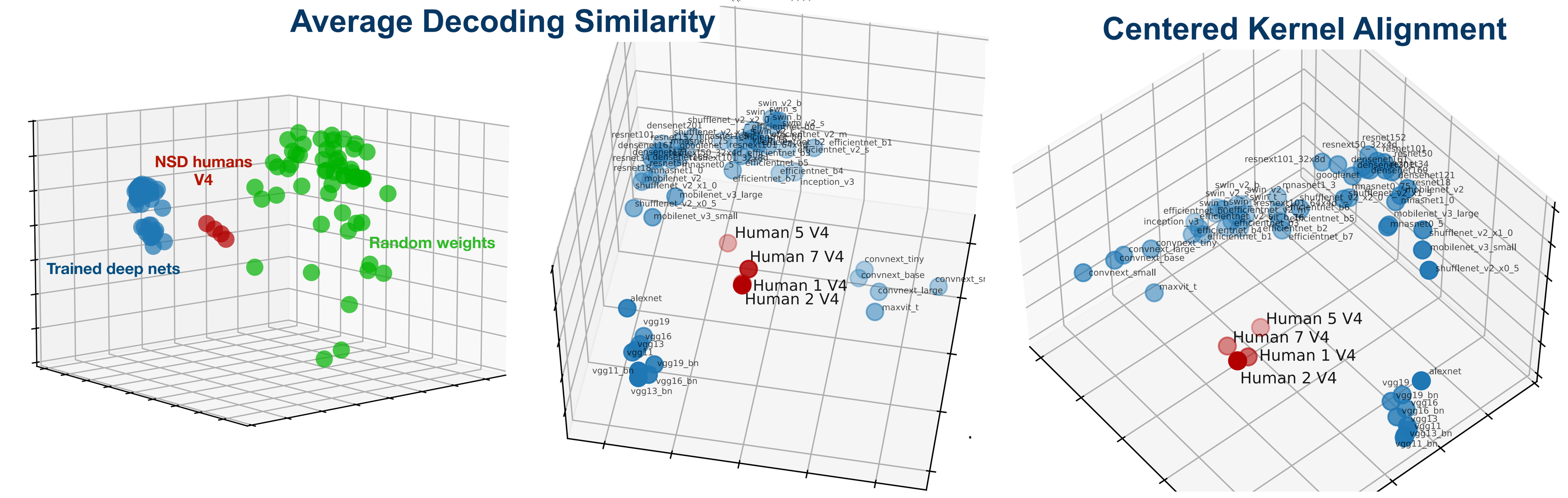
$\mathbb{E}_{\mathbf{z}}\left[\mathbf{z}\mathbf{z}^\top\right] =$
*Covariance of fMRI responses*



### Average Decoding Similarity

### Centered Kernel Alignment




## 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).