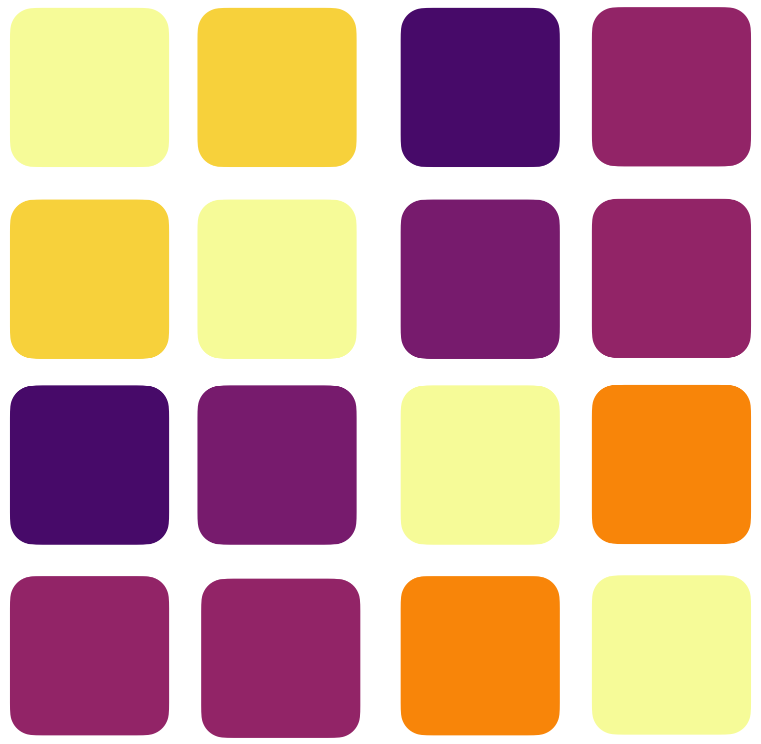
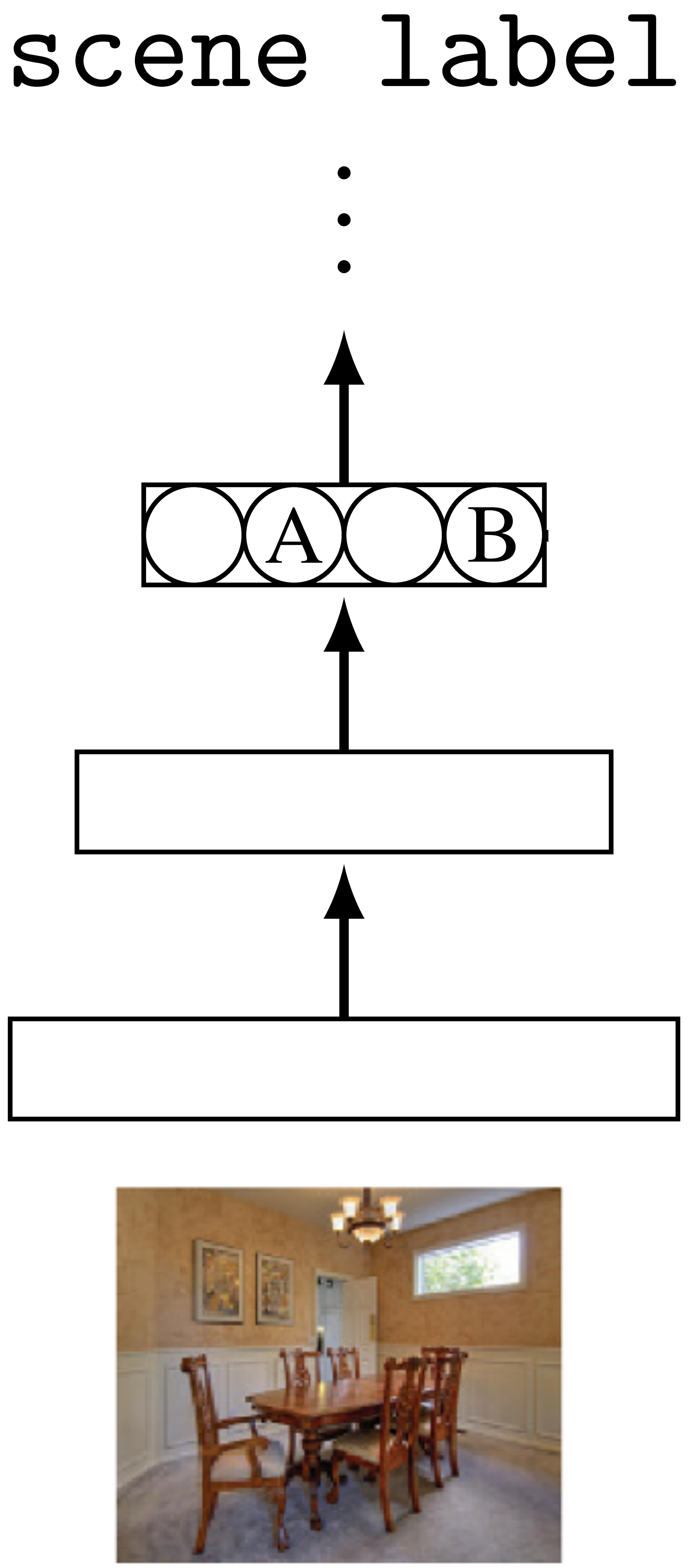


Understanding Representational Alignment in Neural Nets

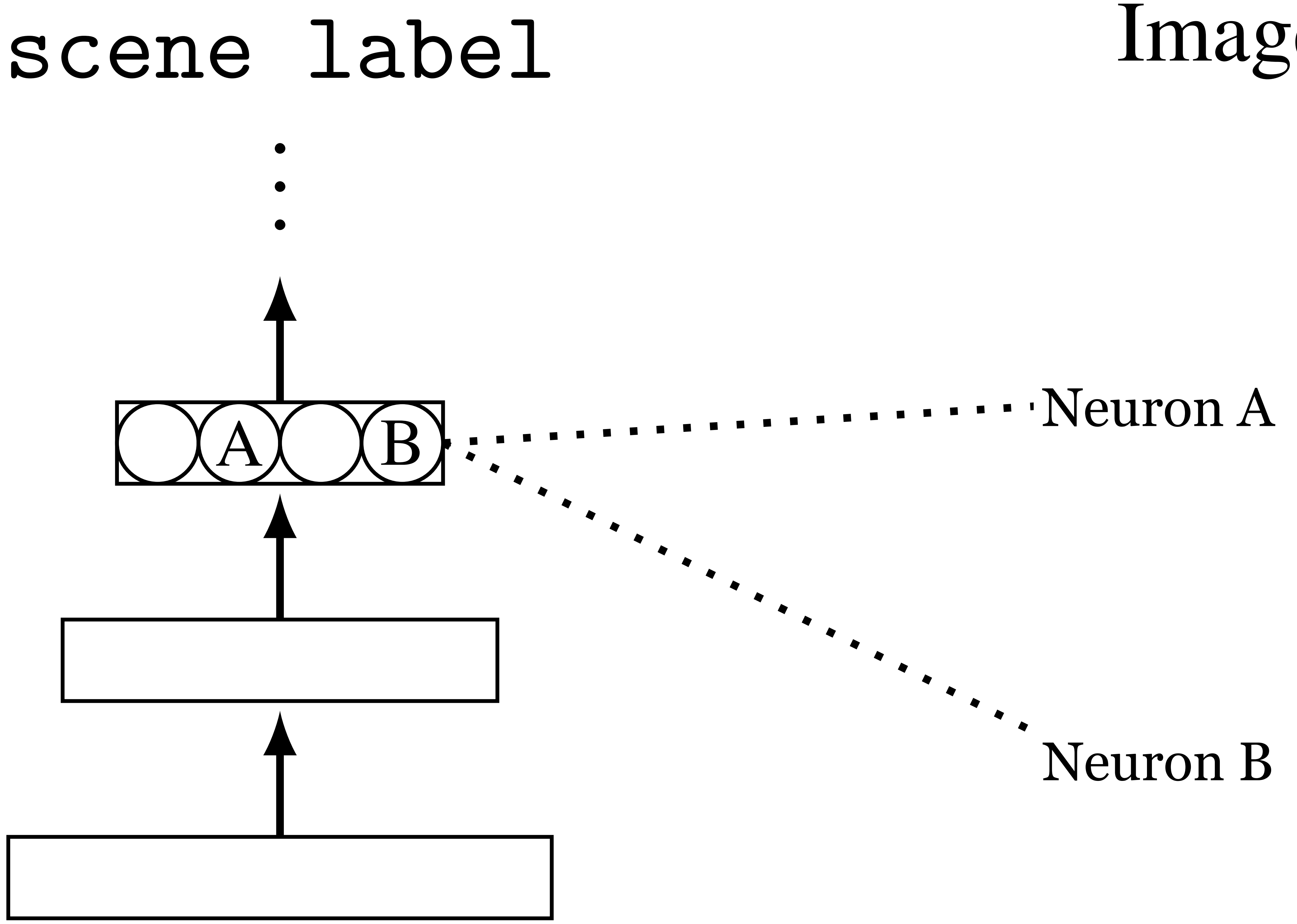
Phillip Isola, MIT
IN2346, TUM
July 15th , 2025

Object Detectors Emerge in Deep Scene CNNs



[Zhou, Khosla, Lapedriza, Oliva, Torralba 2015]

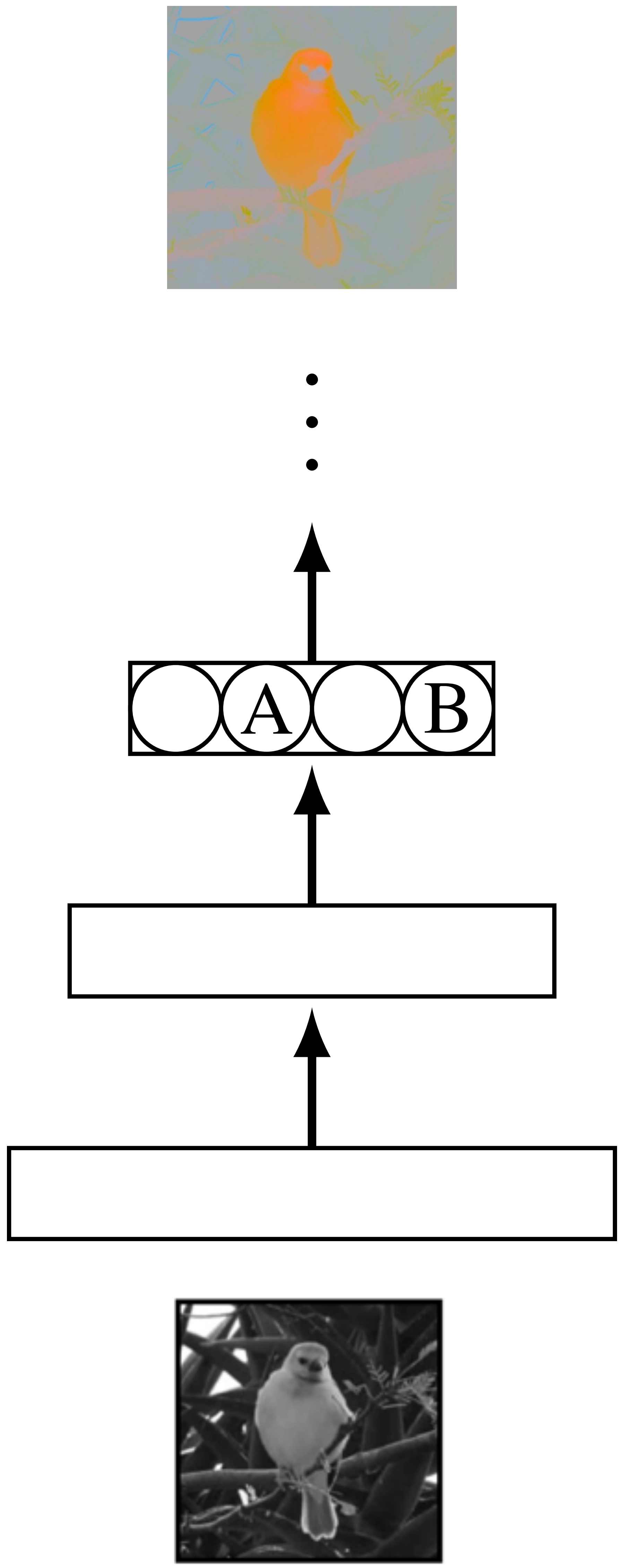
[fig modified from: Torralba, Isola, Freeman 2024]



Images that maximally activate these neurons

[Zhou, Khosla, Lapedriza, Oliva, Torralba 2015]

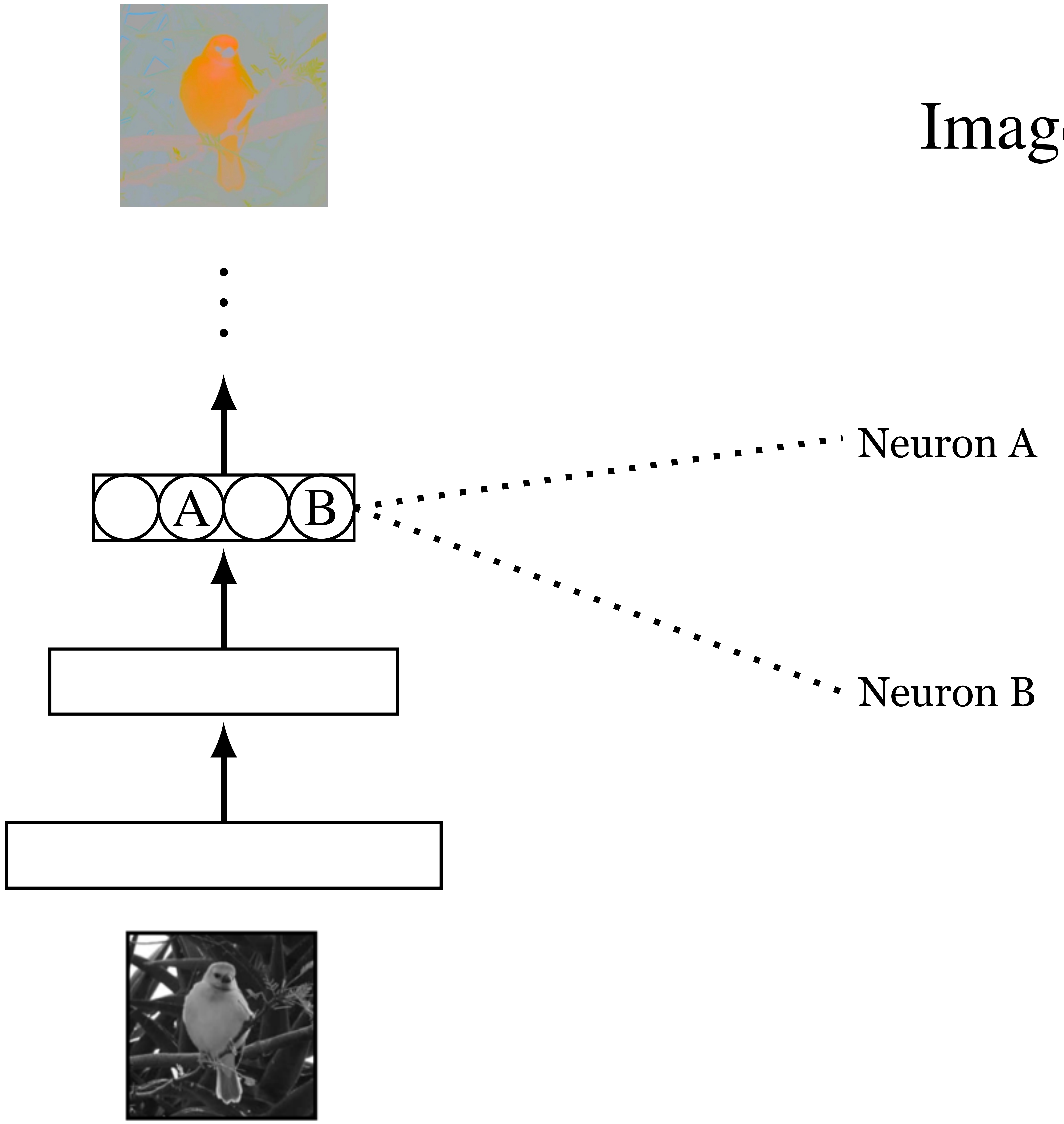
[fig modified from: Torralba, Isola, Freeman 2024]



[Zhang, Isola, Efros 2016]

[fig from: Torralba, Isola, Freeman 2024]

Images that maximally activate these neurons

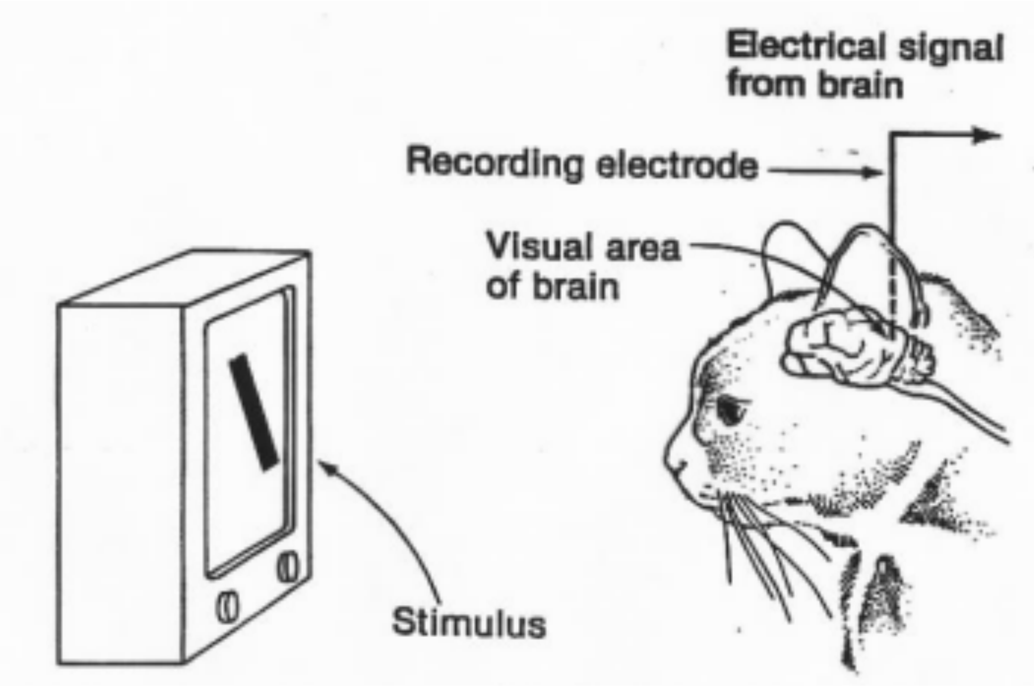


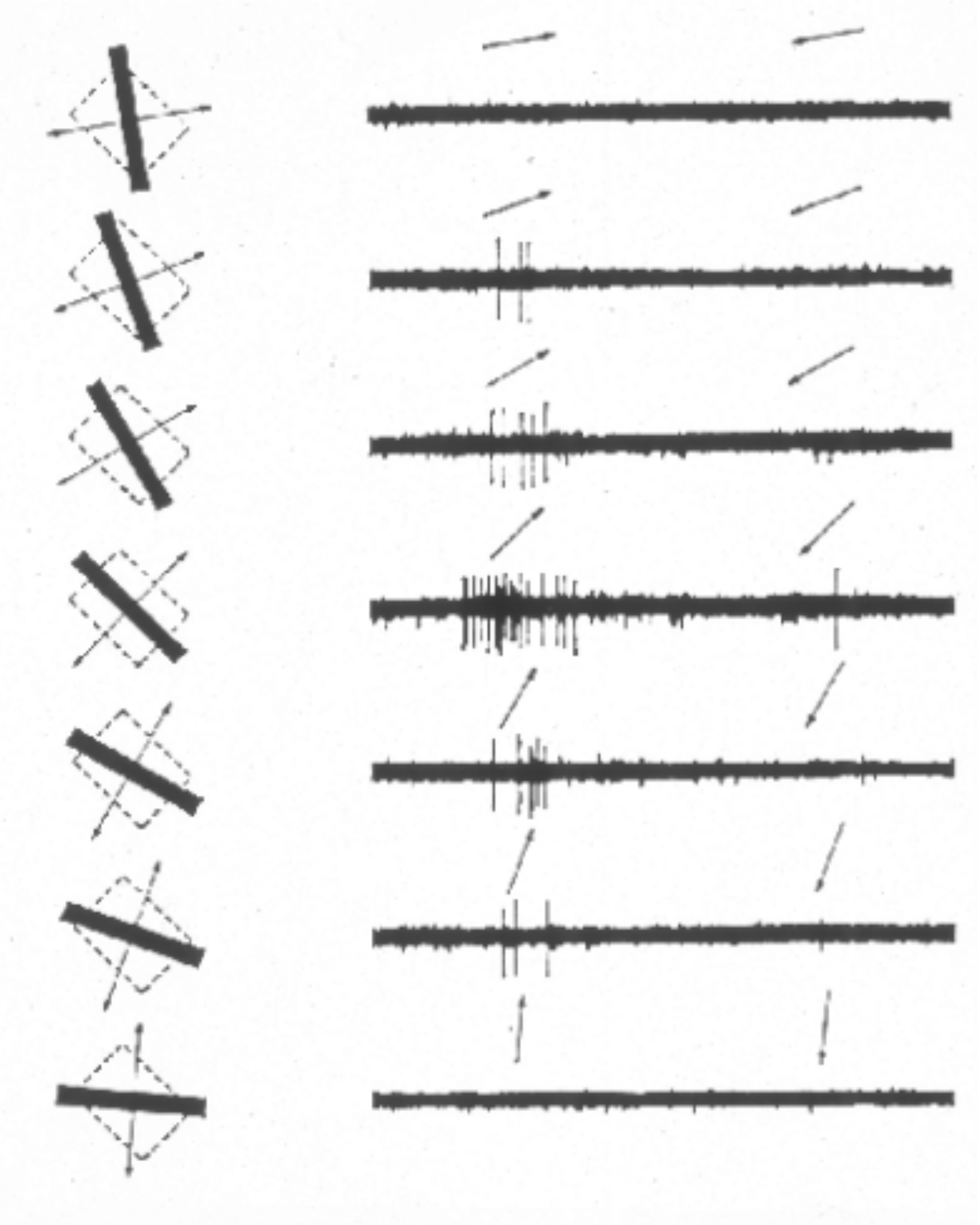
[Zhang, Isola, Efros 2016]

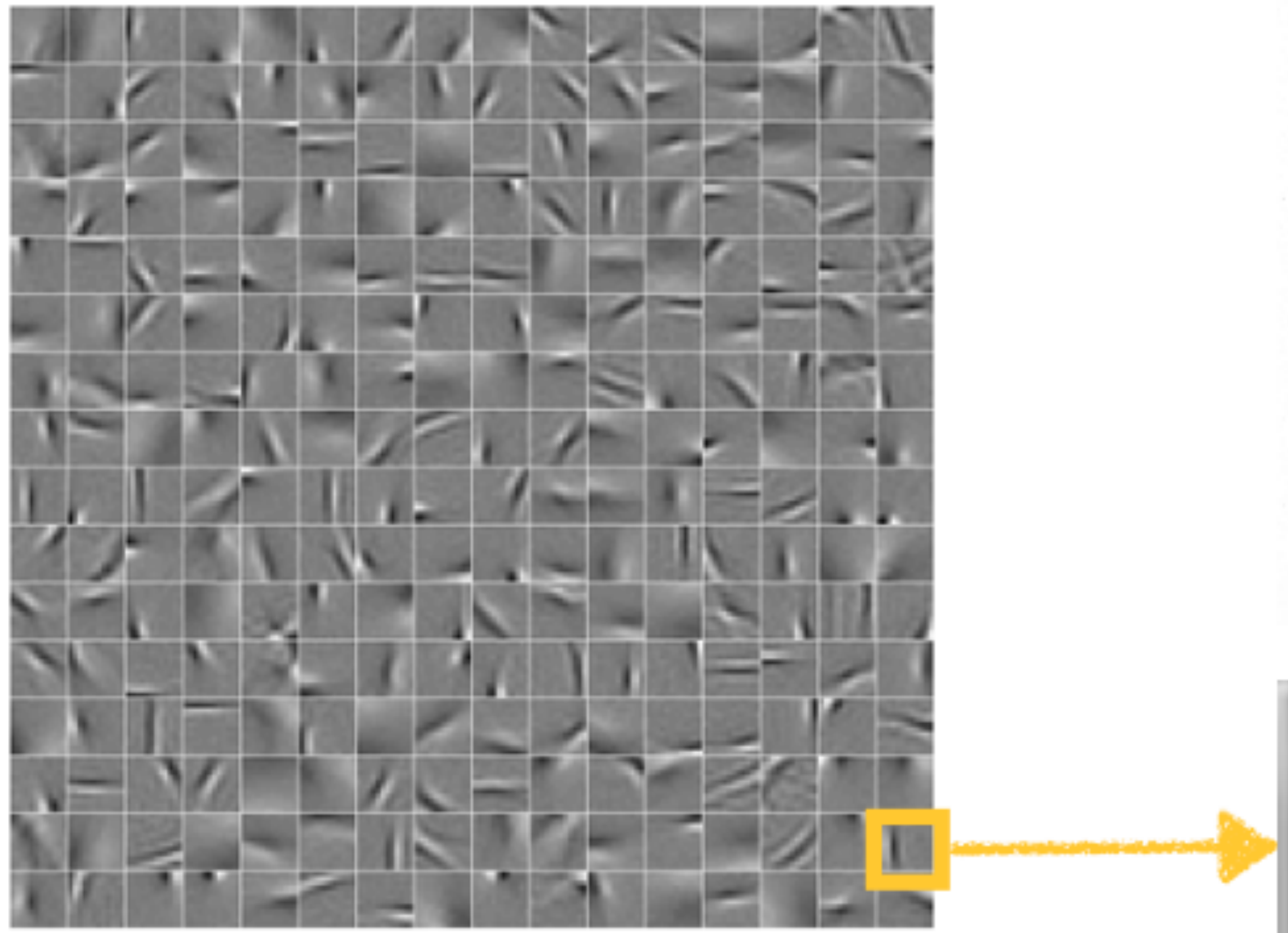
[fig from: Torralba, Isola, Freeman 2024]

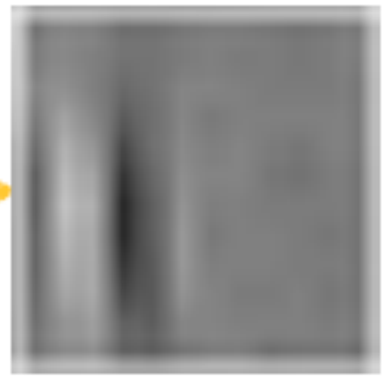
Common features

[Hubel and Wiesel 59]





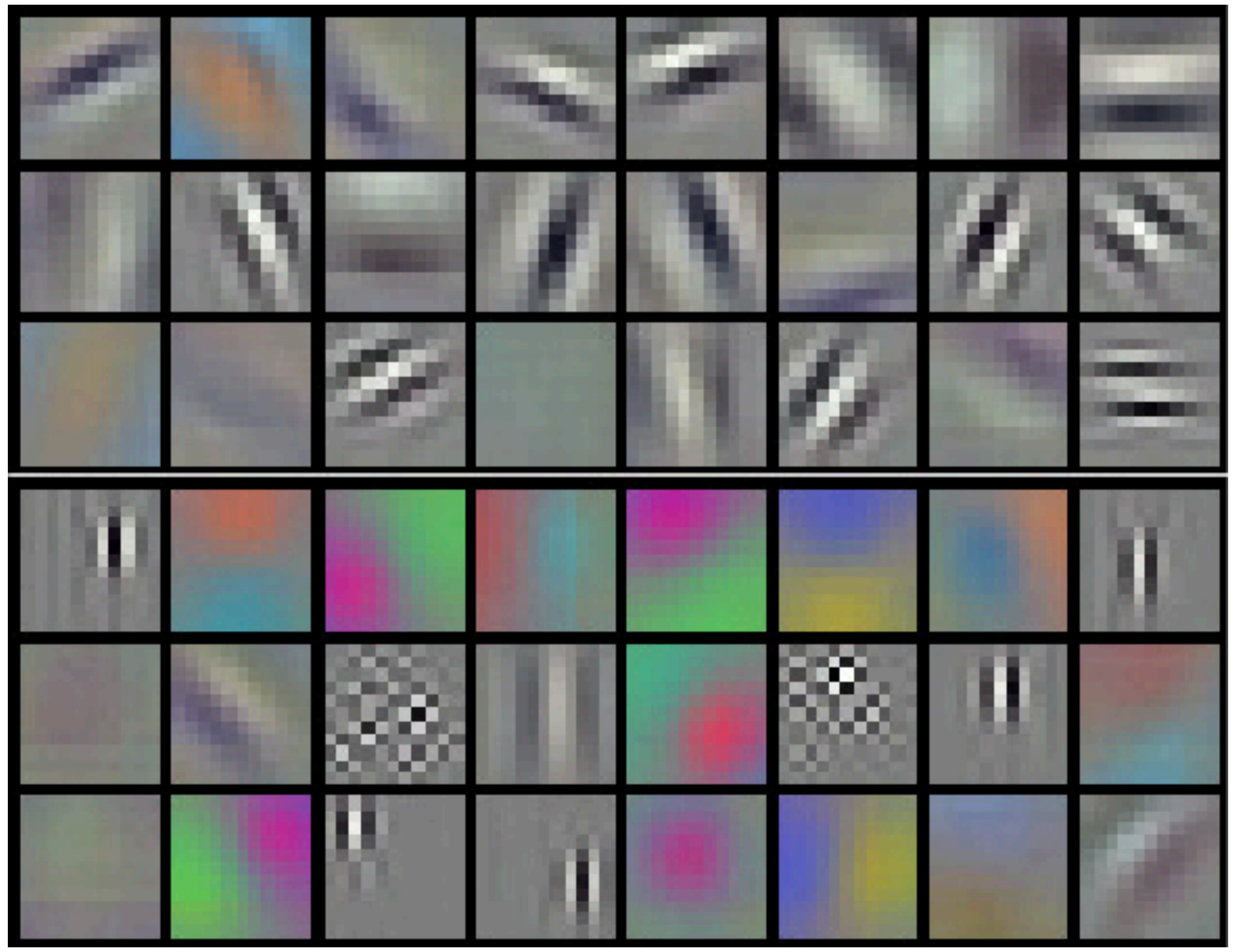




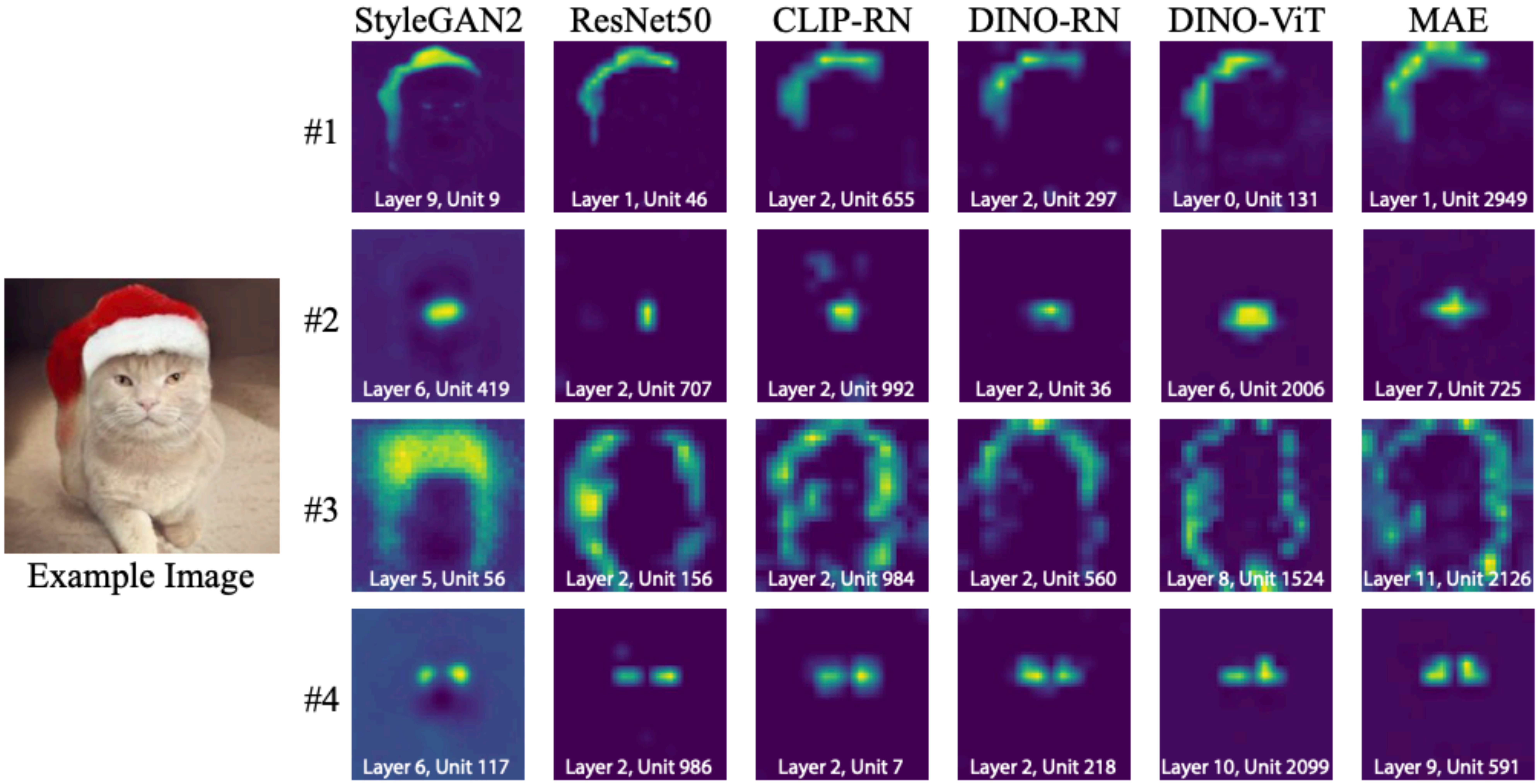
oriented filter

[fig from Andrea Vedaldi]

Filters in AlexNet



Rosetta Neurons



[Dravid*, Gandelsman*, et al. 2023]

Outline:

1. What's a representation?
2. How to measure representational similarity?
3. Which representations are similar and which are different?
4. What drives representational alignment?
5. Making representations more aligned

Outline:

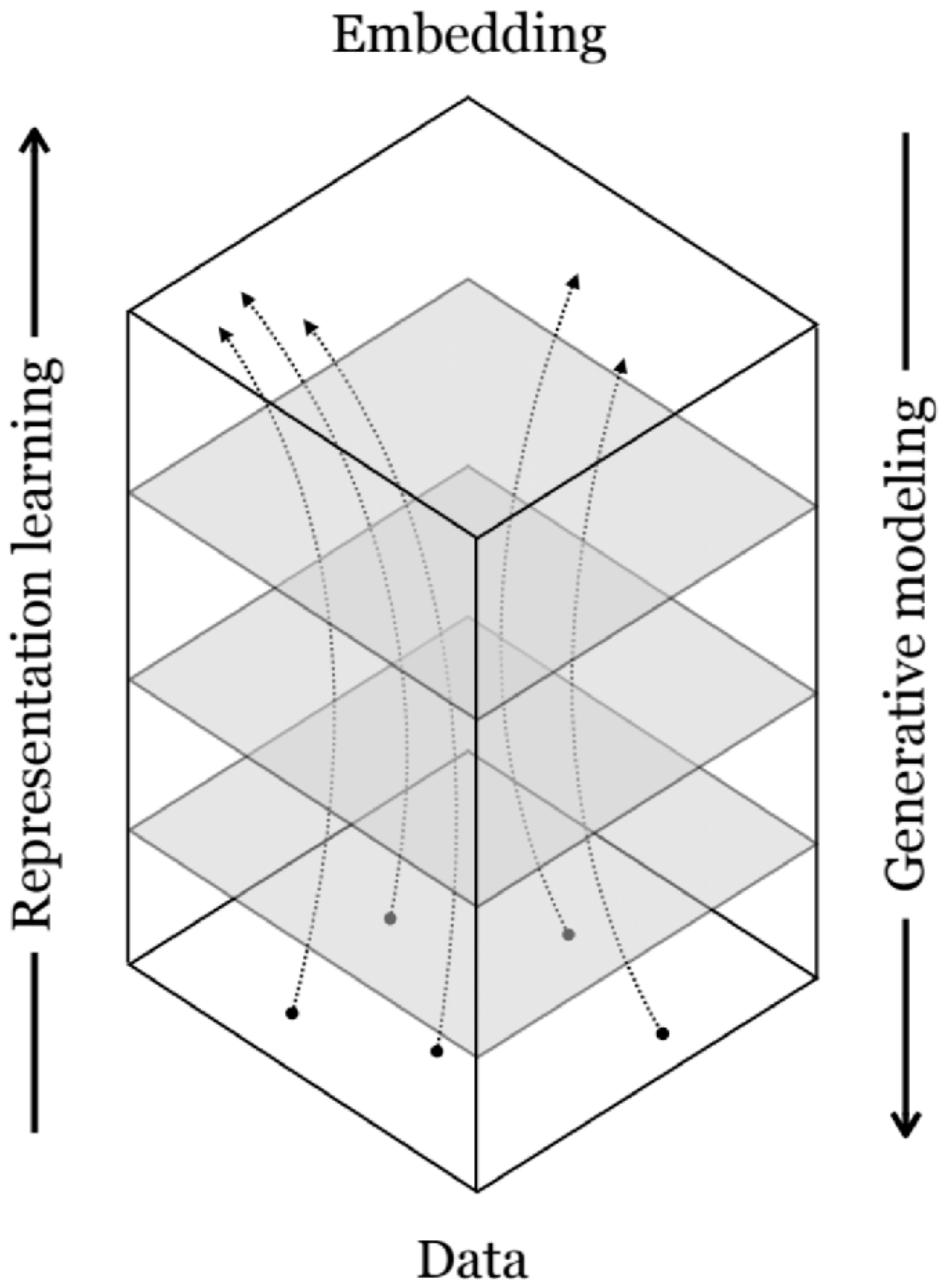
1. What's a representation?

2. How to measure representational similarity?

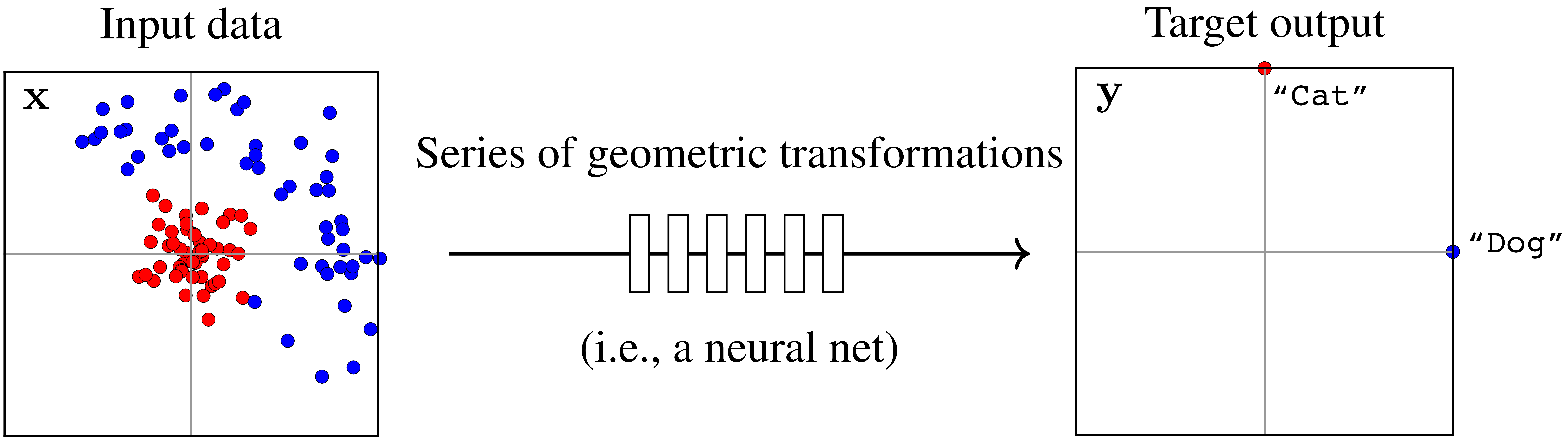
3. Which representations are similar and which are different?

4. What drives representational alignment?

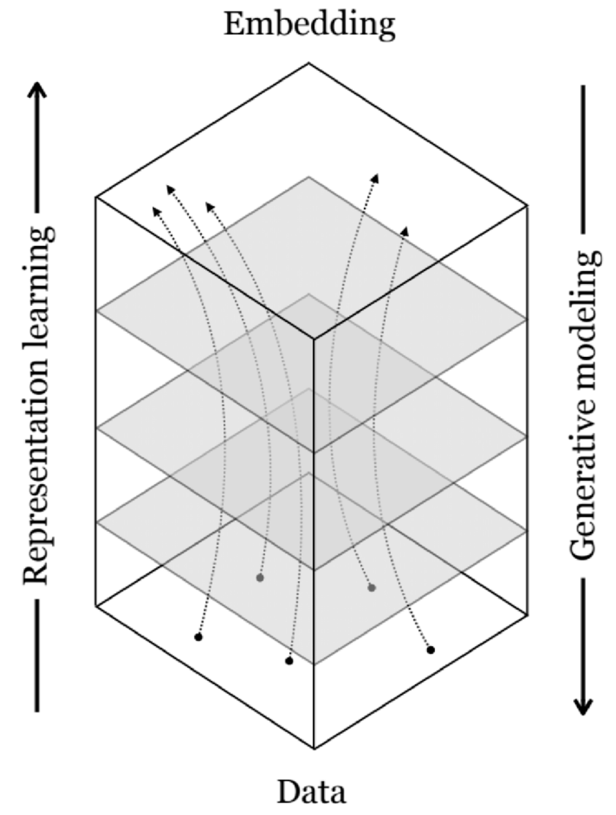
5. Making representations more aligned


Embedding
Representation learning
Generative modeling
Data

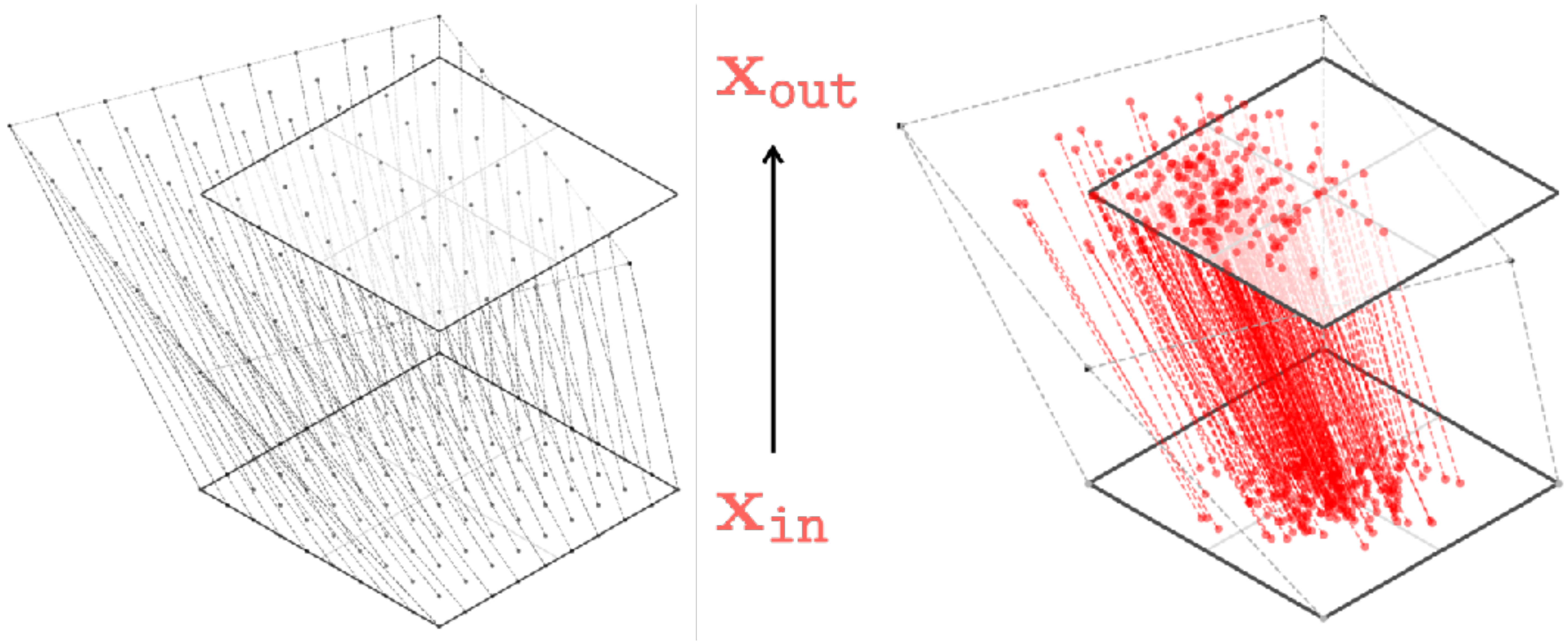
What does training a deep net classifier look like?



linear



$$\mathbf{x}_{\text{out}} = \mathbf{W}\mathbf{x}_{\text{in}} + \mathbf{b}$$



linear

$$\mathbf{x}_{\text{out}} = \mathbf{W}\mathbf{x}_{\text{in}} + \mathbf{b}$$

relu

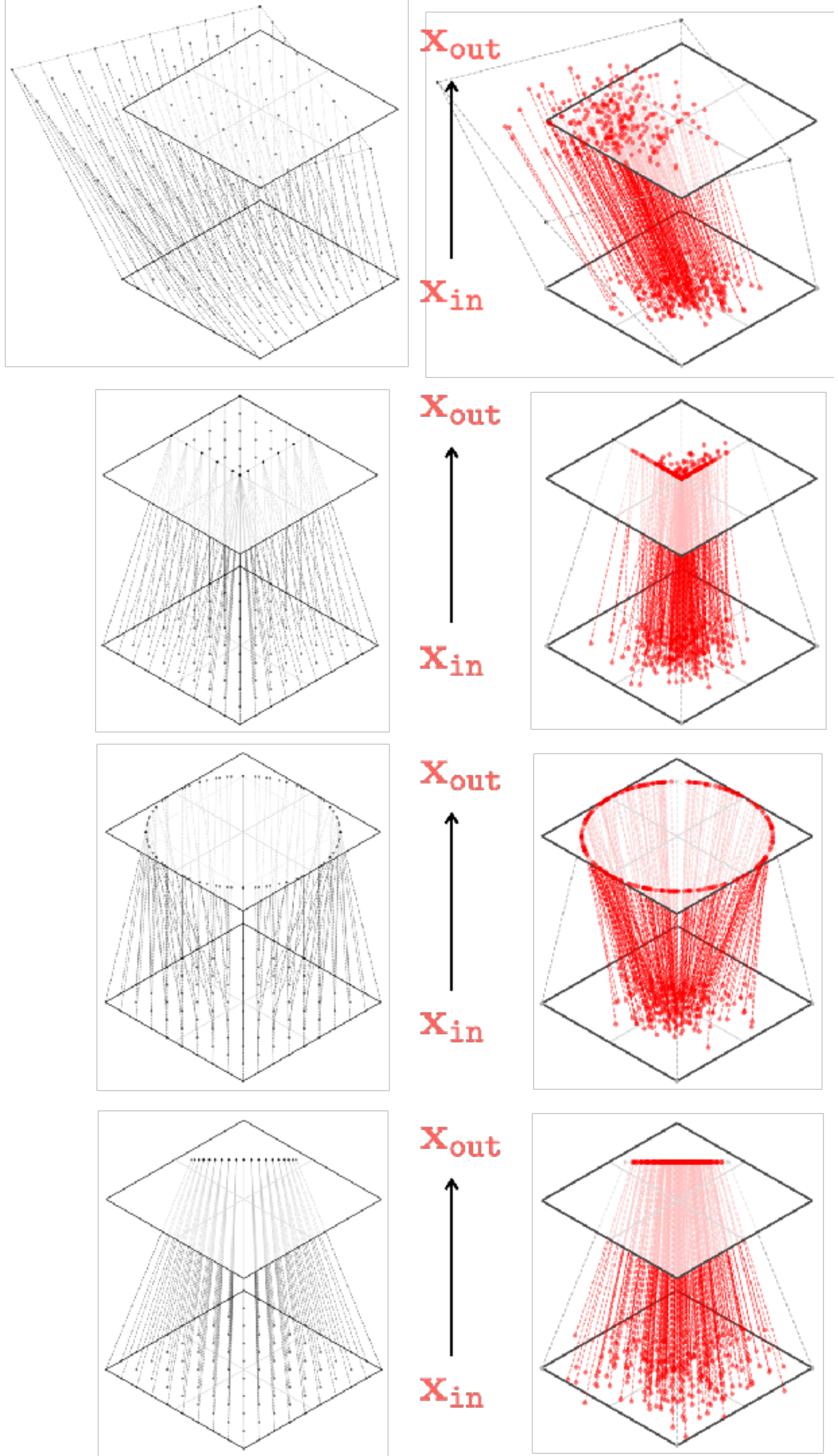
$$x_{\text{out}}[i] = \max(x_{\text{in}}[i], 0)$$

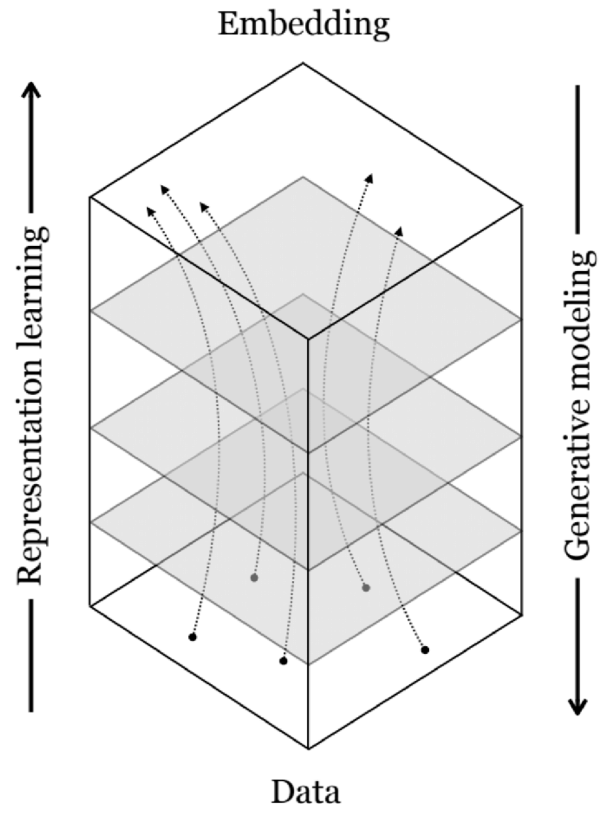
L2-norm

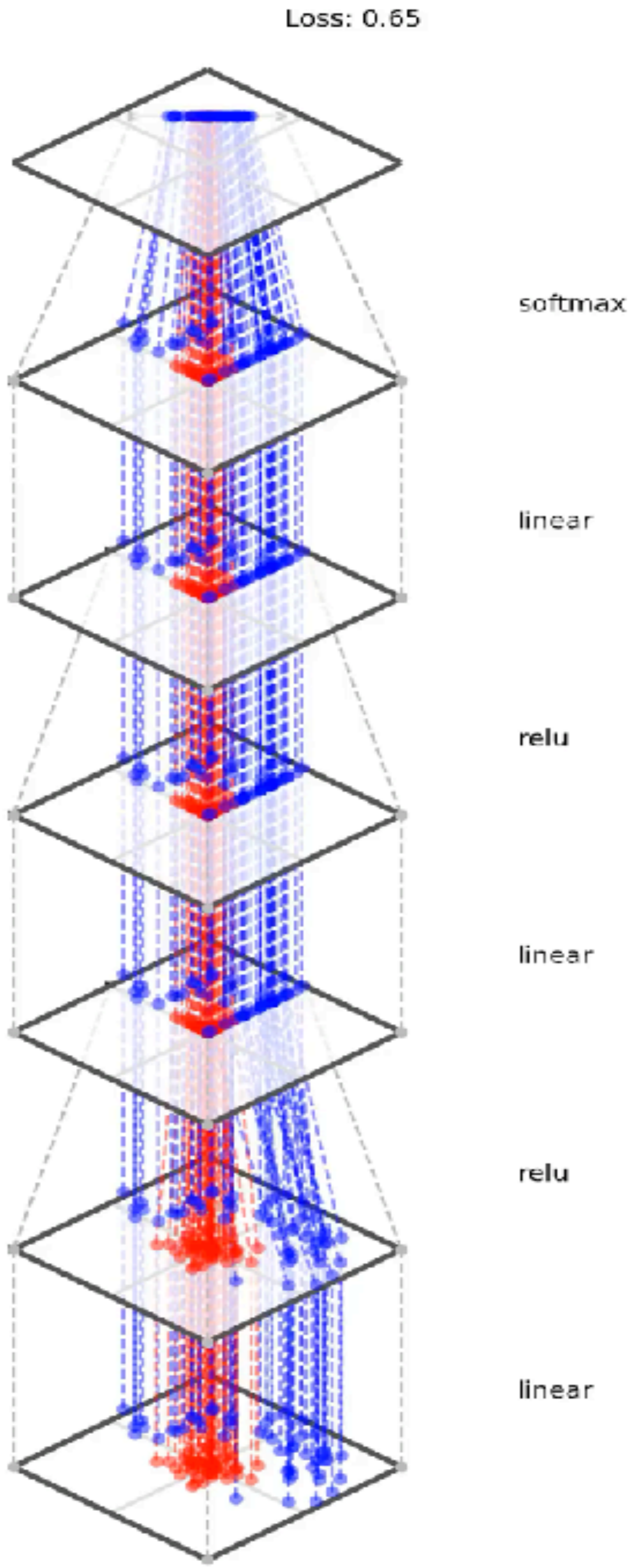
$$x_{\text{out}}[i] = \frac{x_{\text{in}}[i]}{\|\mathbf{x}_{\text{in}}\|_2}$$

softmax

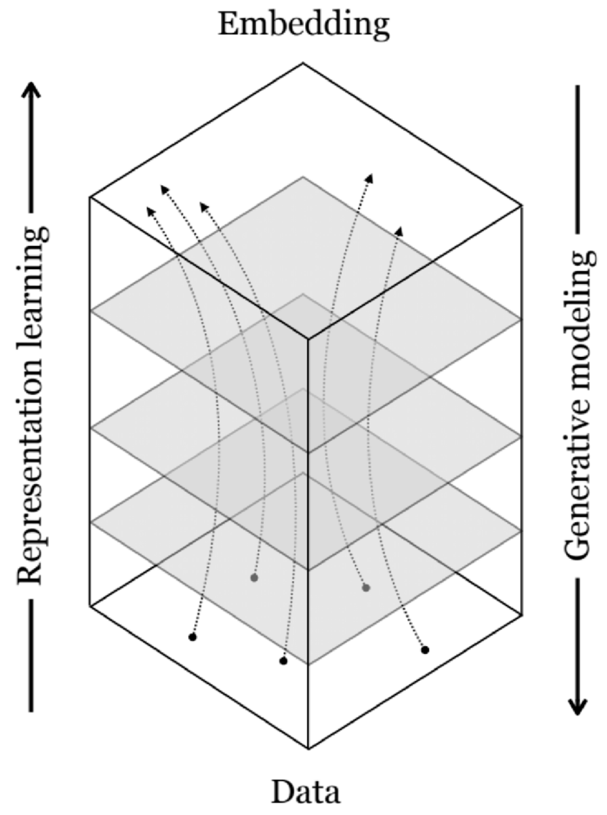
$$x_{\text{out}}[i] = \frac{e^{-\tau x_{\text{in}}[i]}}{\sum_{k=1}^{K} e^{-\tau x_{\text{in}}[k]}}$$





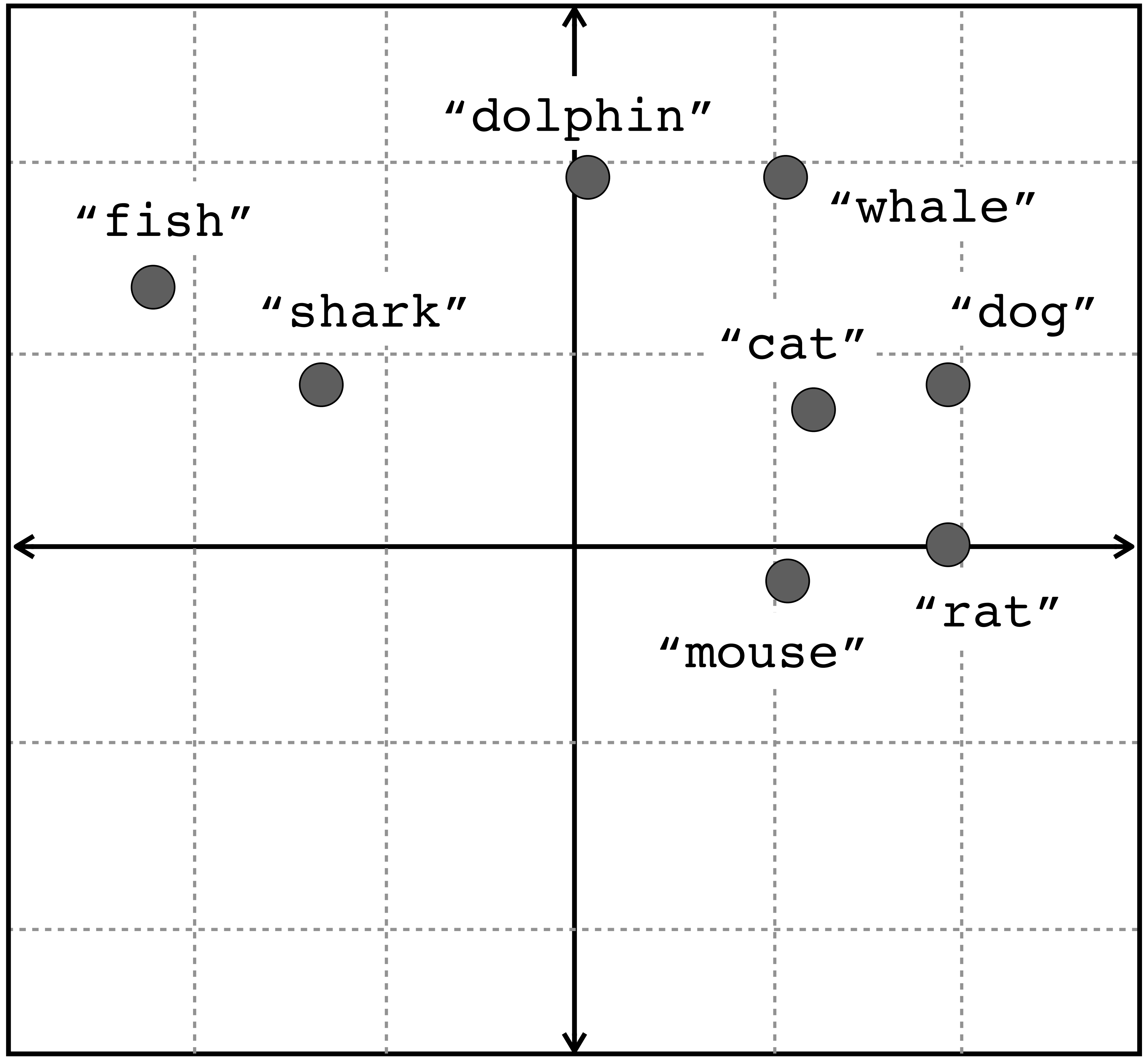


Each layer is a representation



A “representation” is an assignment datapoints to locations in some space

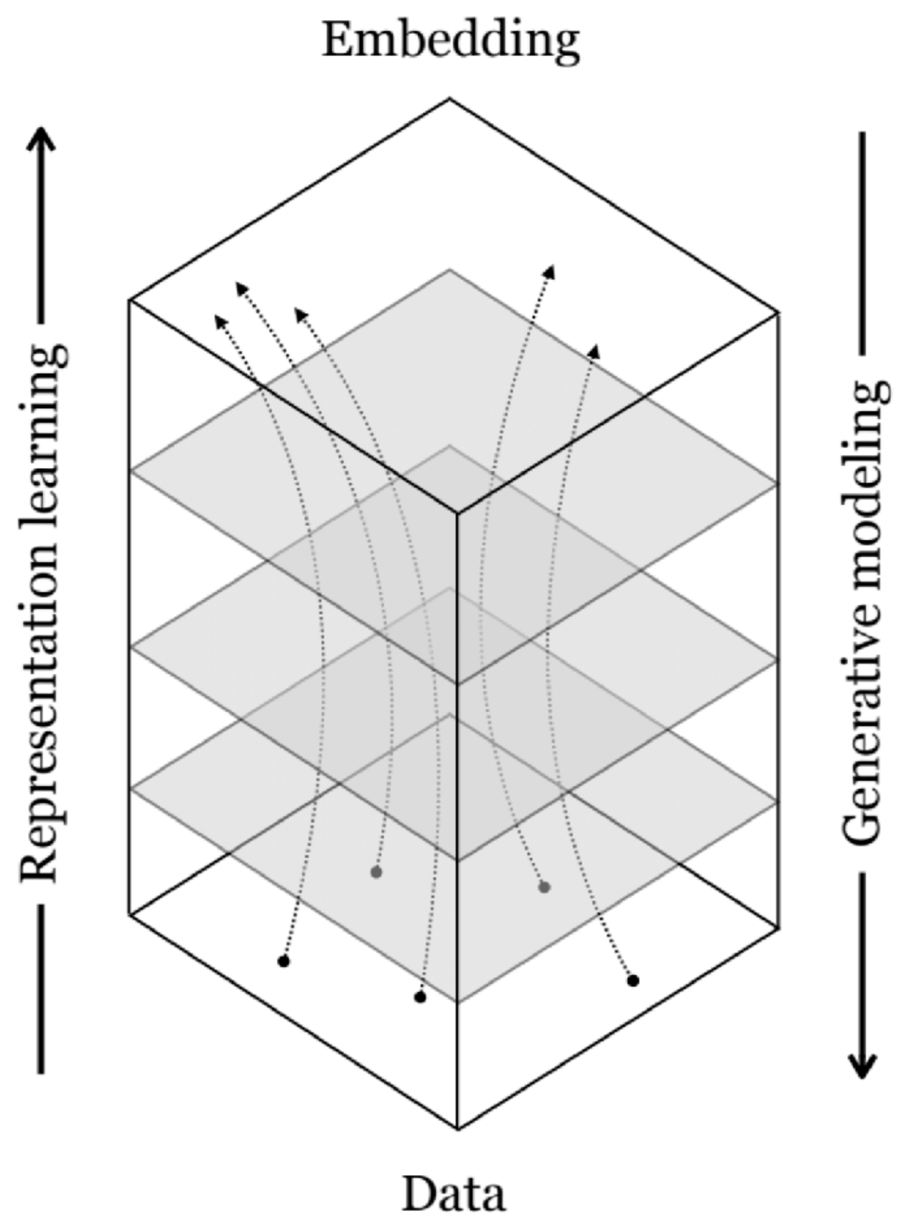
i.e. a labeled point cloud:

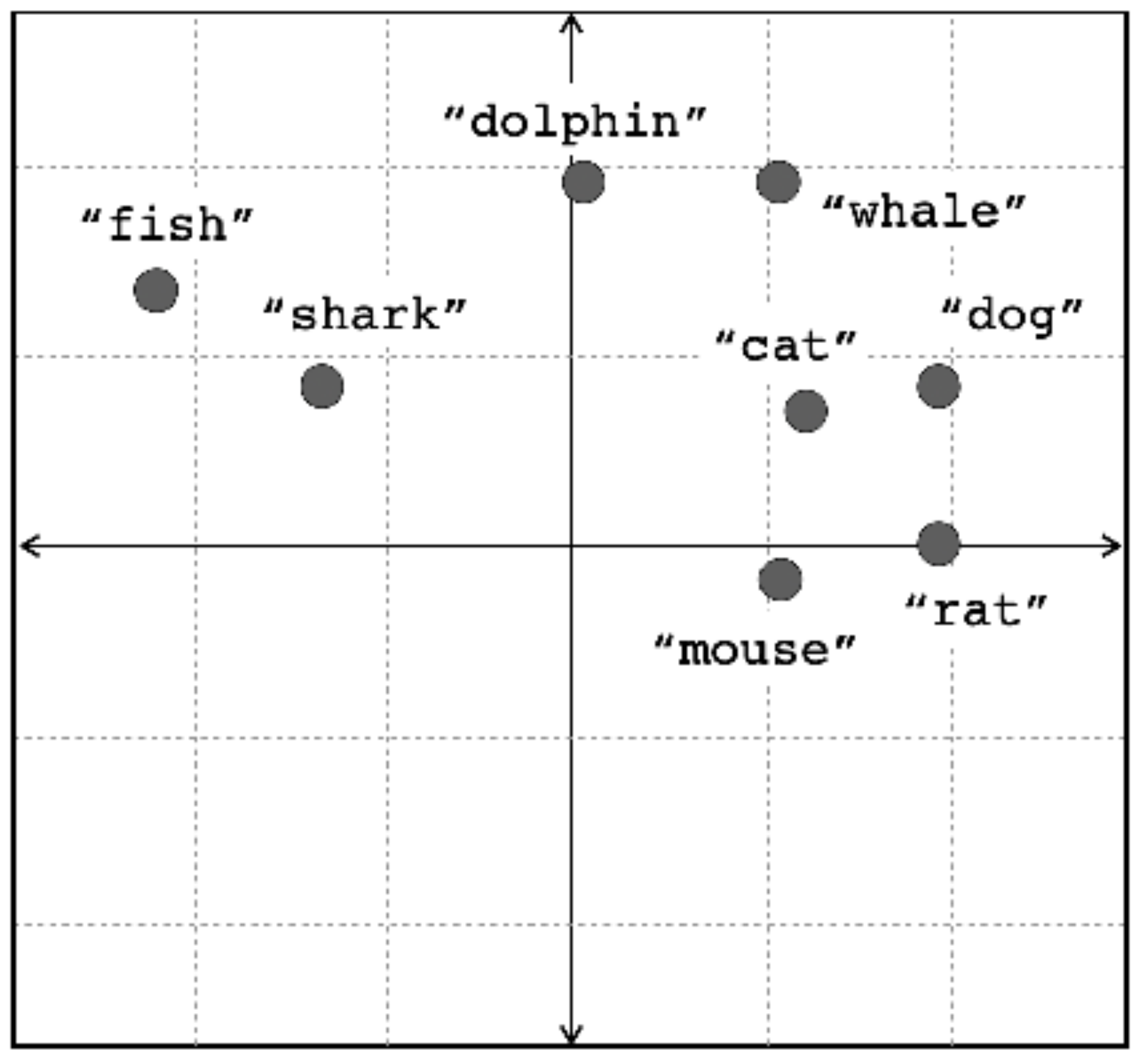


Definitions and notation

- A **representation** is a mapping $f : \mathscr{X} \to \mathscr{Z}$, where $x \in \mathscr{X}$ is data and $z \in \mathscr{Z}$ is some transformation of the data.

  - Typically we have $\mathscr{Z} = \mathbb{R}^d$, i.e. the representation maps data to vector embeddings.





Summary #1:

All layers are a representation, and so are the input data and the output beliefs.

Representations can be understood in terms of their geometry.

Outline:

1. What's a representation?

2. How to measure representational similarity?

3. Which representations are similar and which are different?

4. What drives representational alignment?

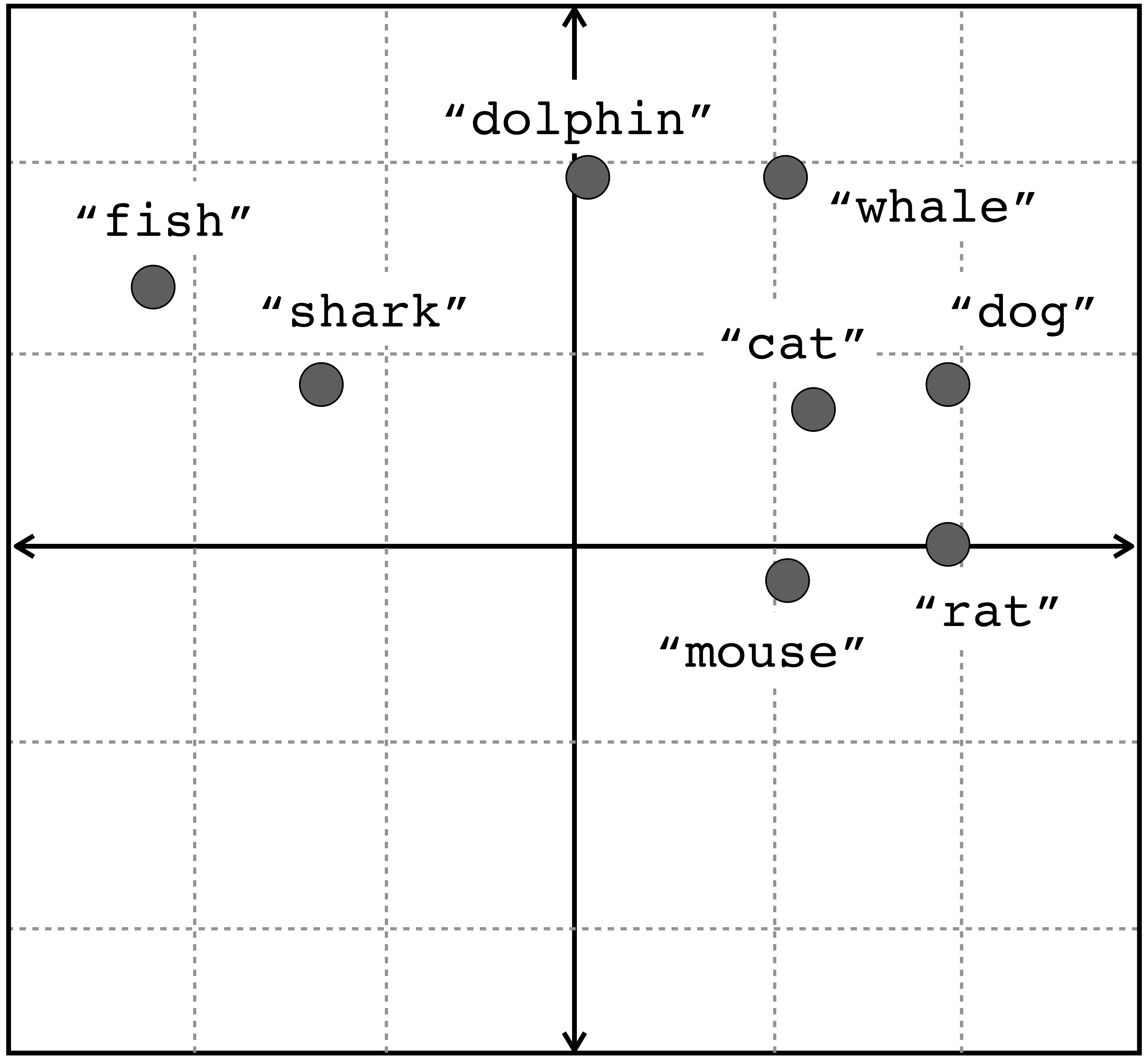
5. Making representations more aligned

Definitions and notation

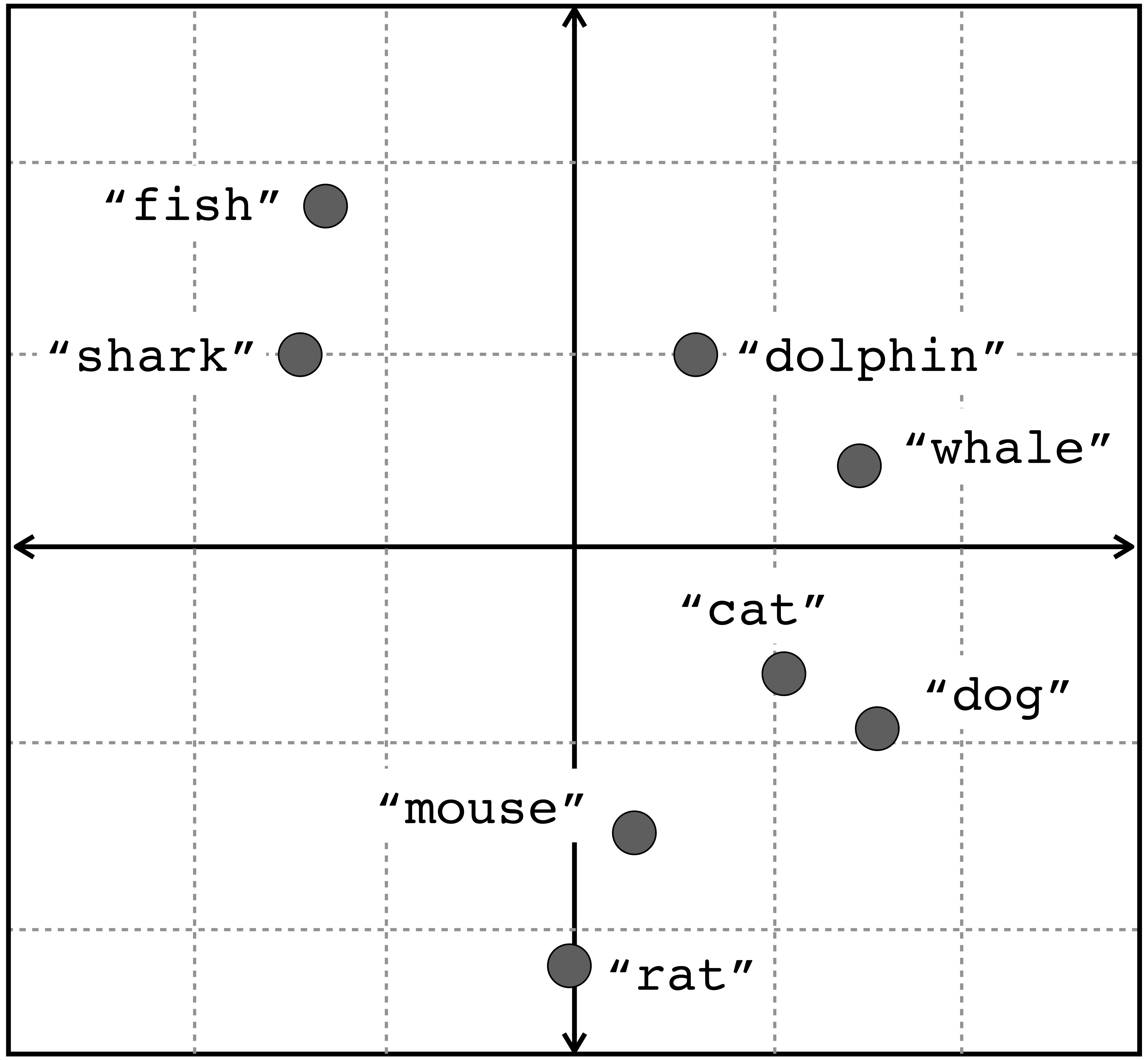
- **Representational similarity** is a measure $d : f_1 \times f_2 \rightarrow \mathbb{R}$
  - It takes two representations as input and outputs a number that is higher if the two representations are to be considered more alike.
  - Often we will measure d over a finite set of datapoints, $\mathbf{Z}_1 = \{f_1(x^{(i)})\}_{i=1}^{n}$, $\mathbf{Z}_2 = \{f_2(x^{(i)})\}_{i=1}^{n}$, with $d^z : \mathbf{Z}_1 \times \mathbf{Z}_2 \rightarrow \mathbb{R}$

The main question

Neural net 1's embeddings ($\mathbf{Z}_1$)



Neural net 2's embeddings ($\mathbf{Z}_2$)



How similar are these two point clouds?

Regression-based metrics

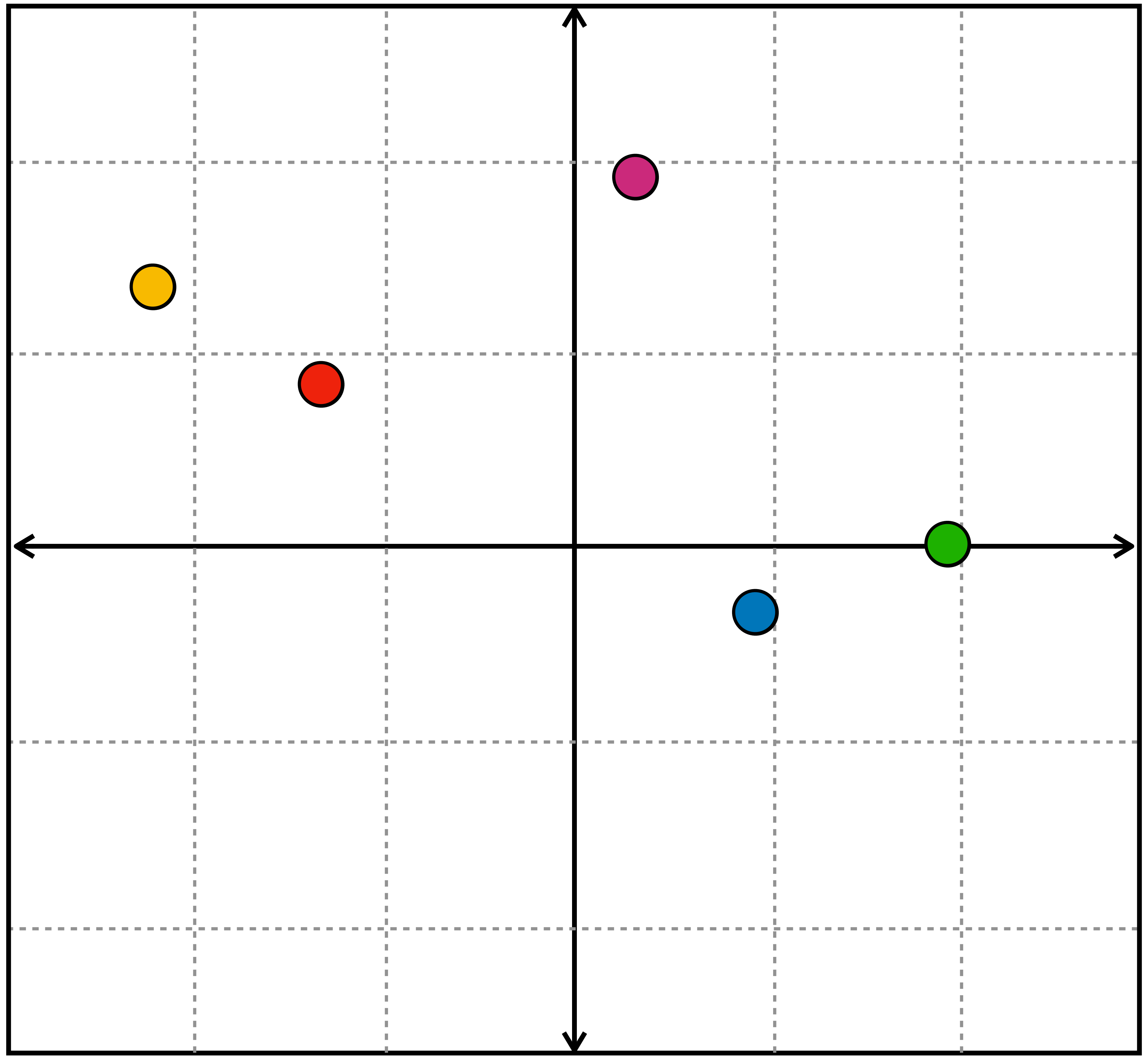
Neural net 1's embeddings ($\mathbf{Z}_1$) $\xrightarrow{h}$ Neural net 2's embeddings ($\mathbf{Z}_2$)

$$d(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{n} \sum_{i=1}^{n} ||h^*(z_1^{(i)}) - z_2^{(i)}||$$

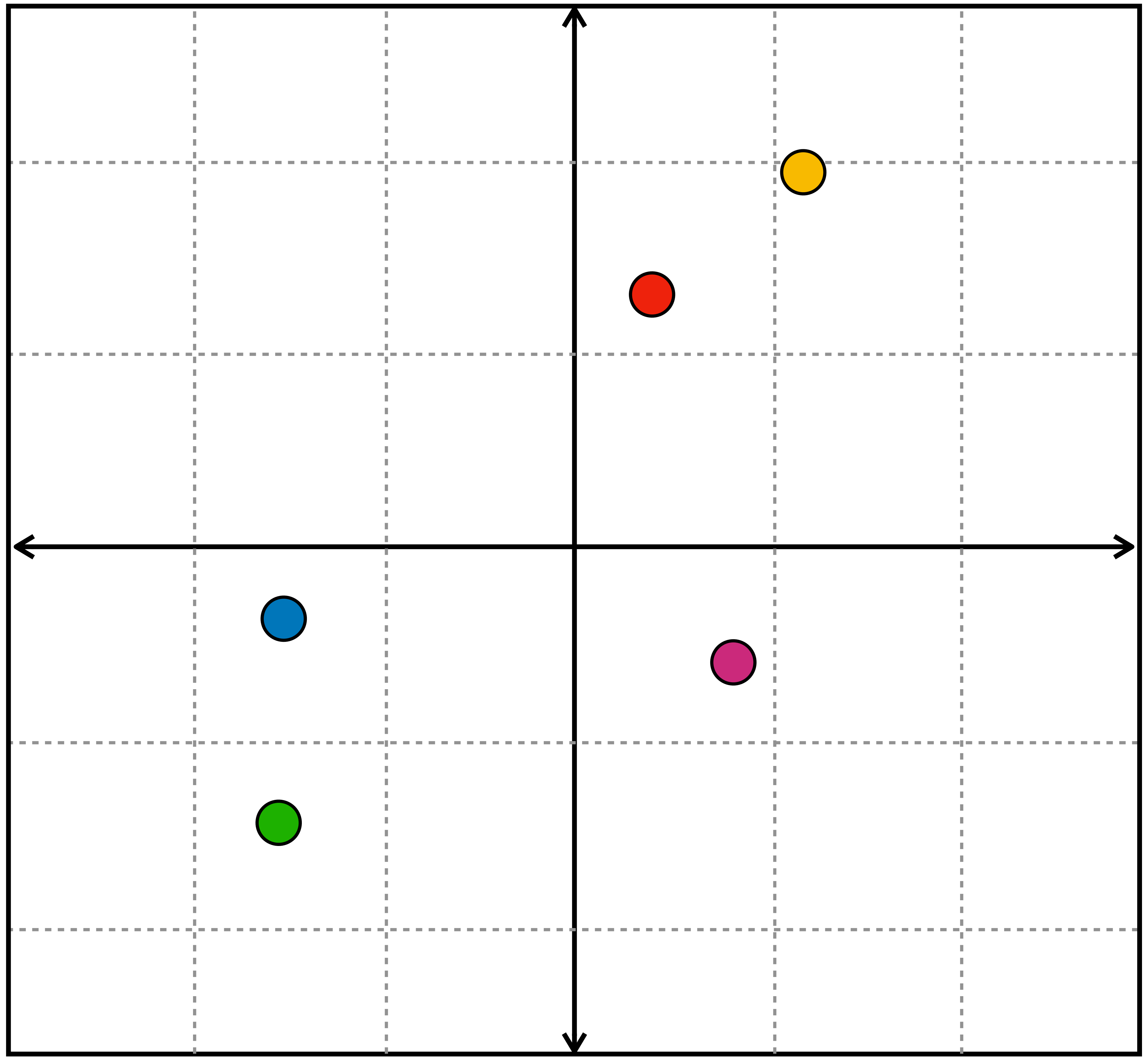
$$h^* = \arg\min_{h} \frac{1}{n} \sum_{i=1}^{n} ||h(z_1^{(i)}) - z_2^{(i)}||$$

Two equivalent representations under linear regression

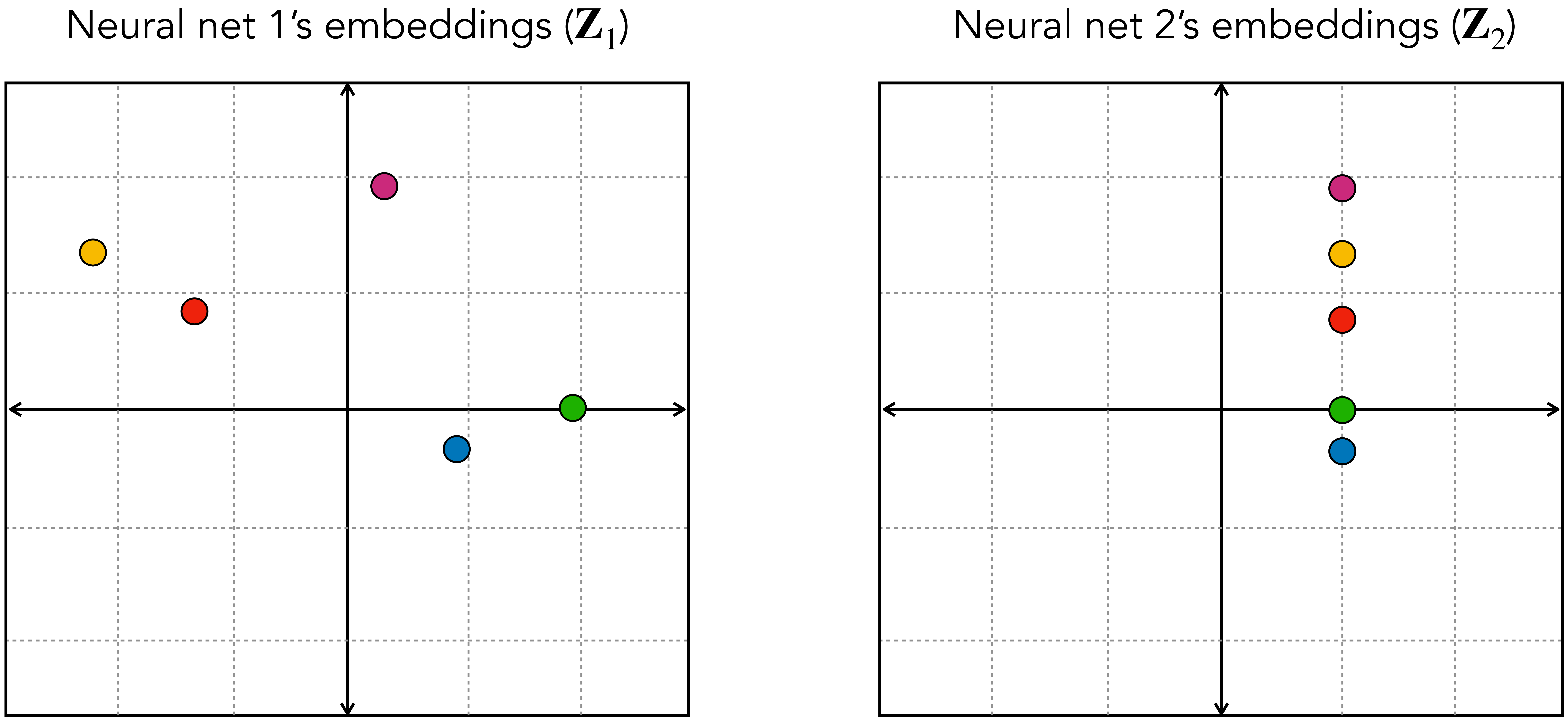
Neural net 1's embeddings ($\mathbf{Z}_1$)



Neural net 2's embeddings ($\mathbf{Z}_2$)

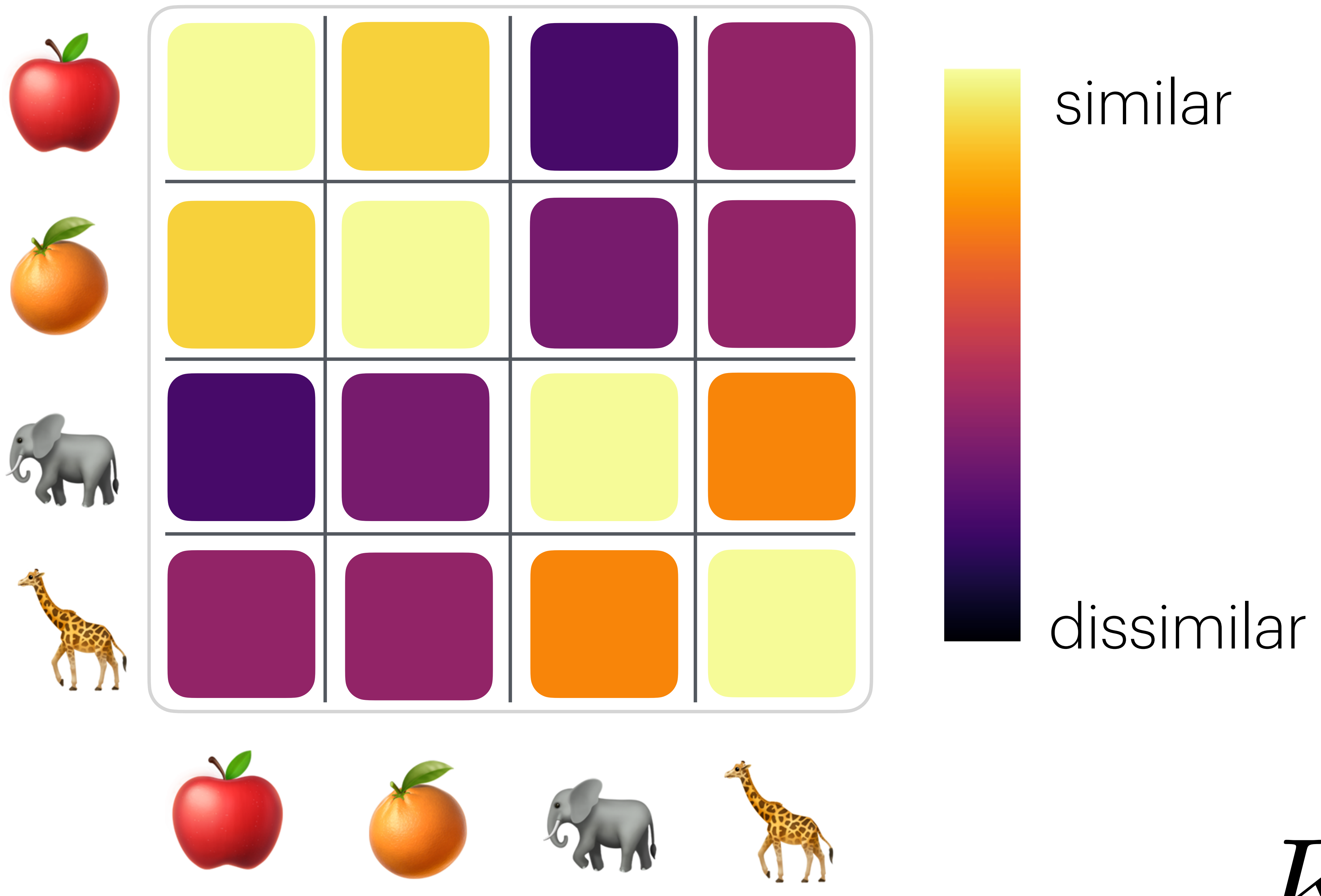


Two equivalent representations under linear regression



Kernel-alignment metrics

$\mathbf{K}_{\text{vision}}$



Restrict our attention to **vector embeddings**

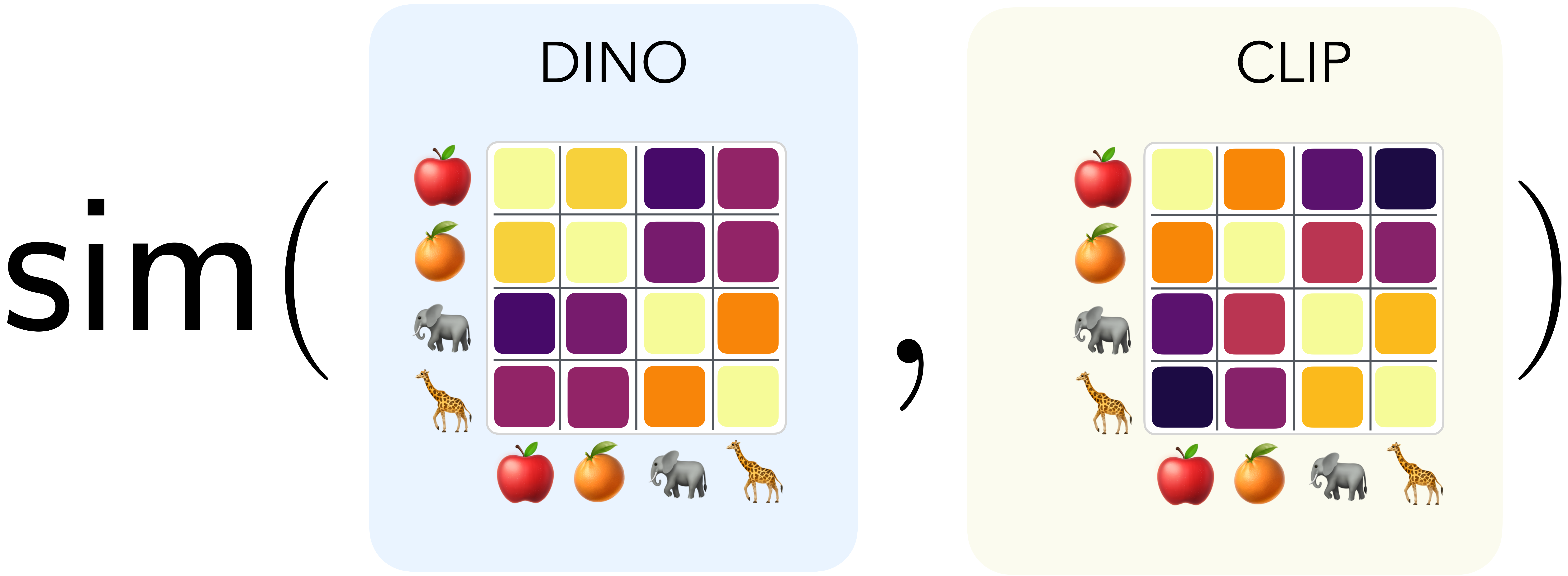
$$f : \mathcal{X} \to \mathbb{R}^n$$

Characterize a representation in terms of its **kernel**

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

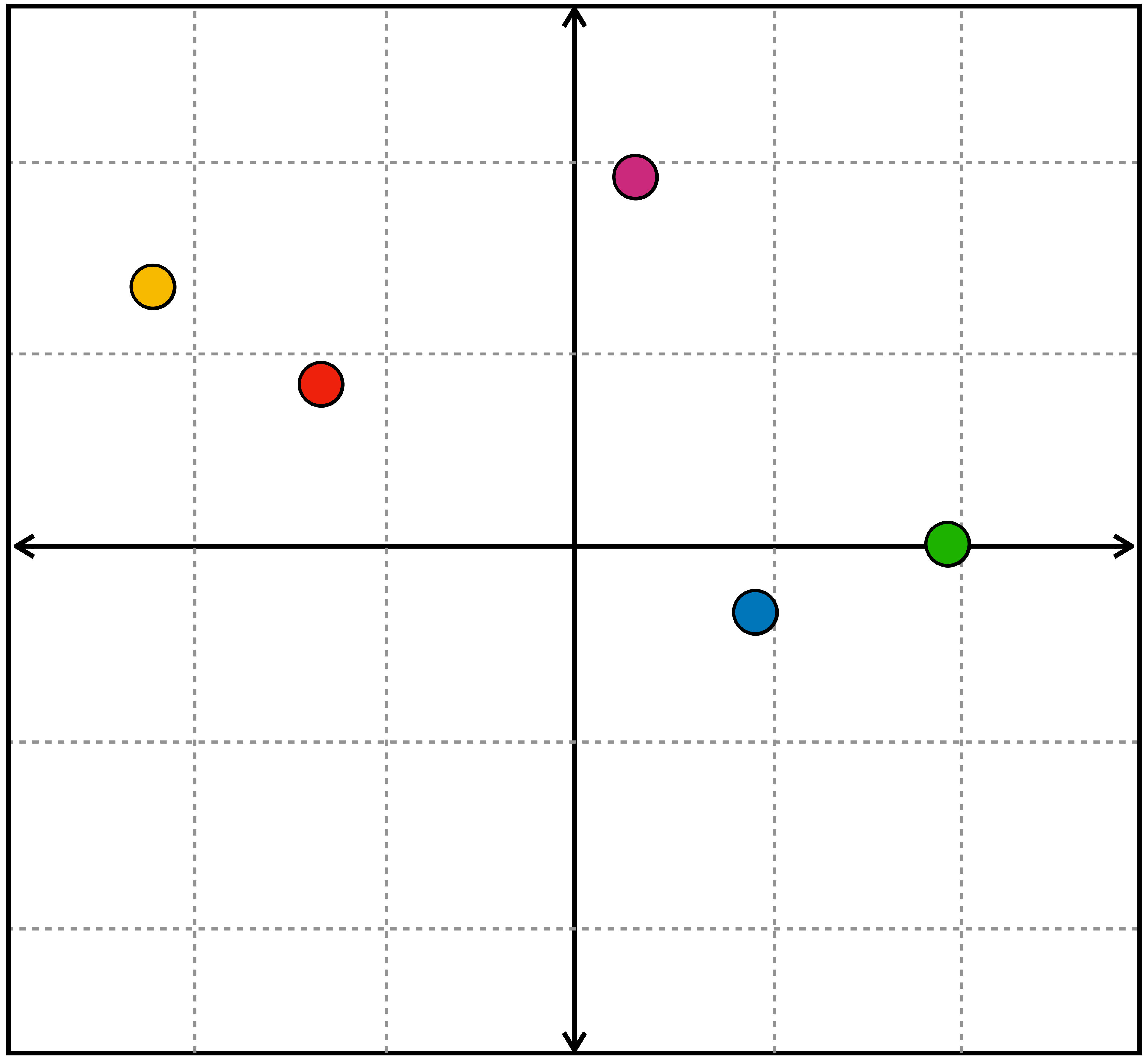
$$K(x_i, x_j) = \langle\, f(\text{🍎}), f(\text{🍊})\, \rangle$$

Kernel-alignment metrics

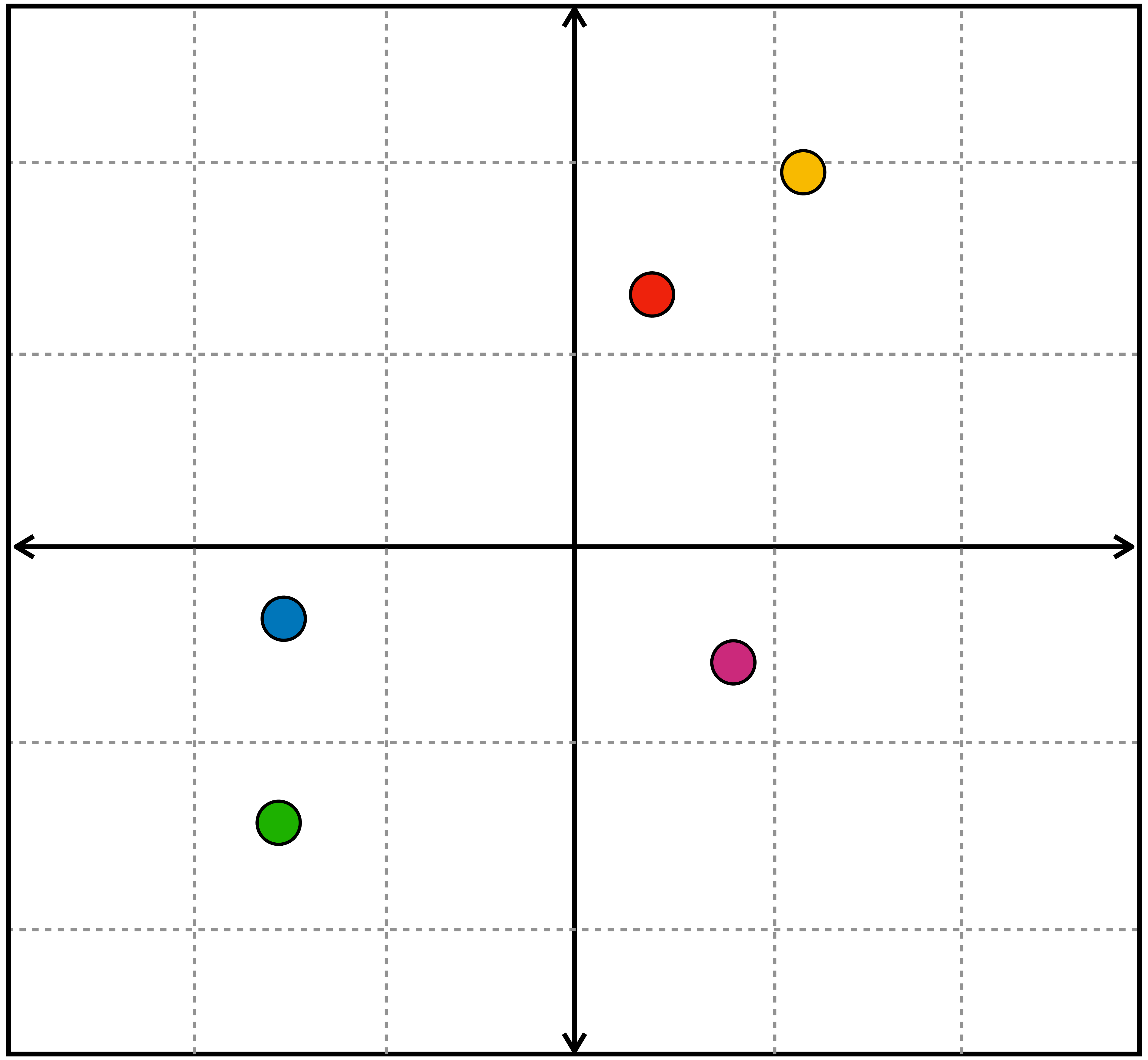


Two representations with equivalent kernels

Neural net 1's embeddings ($\mathbf{Z}_1$)



Neural net 2's embeddings ($\mathbf{Z}_2$)



Rigid transformations don't change distances

Centered Kernel Alignment (CKA)

- Kernel alignment metrics are invariant to isometries (i.e. rotation, translation, mirror flips, "glide reflections")
- CKA says: Let's also be invariant to isotropic scaling

kernel similarity

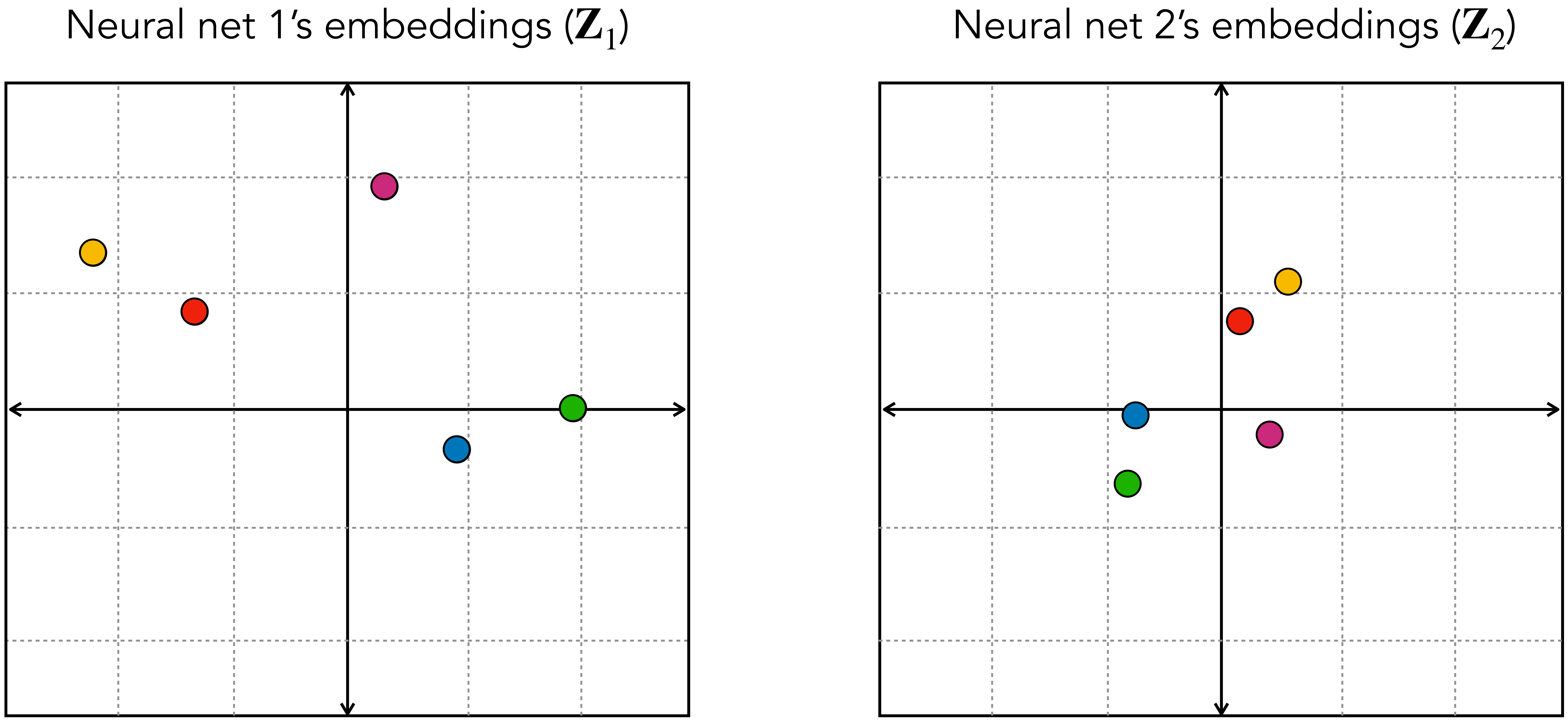
subtracts mean similarity

normalize by scale

$$\mathrm{CKA}(\mathbf{K}_1, \mathbf{K}_2) = \frac{\mathrm{tr}(\mathbf{K}_1 H \mathbf{K}_2 H)}{\sqrt{\mathrm{tr}(\mathbf{K}_1 H \mathbf{K}_1 H)\mathrm{tr}(\mathbf{K}_2 H \mathbf{K}_2 H)}}$$

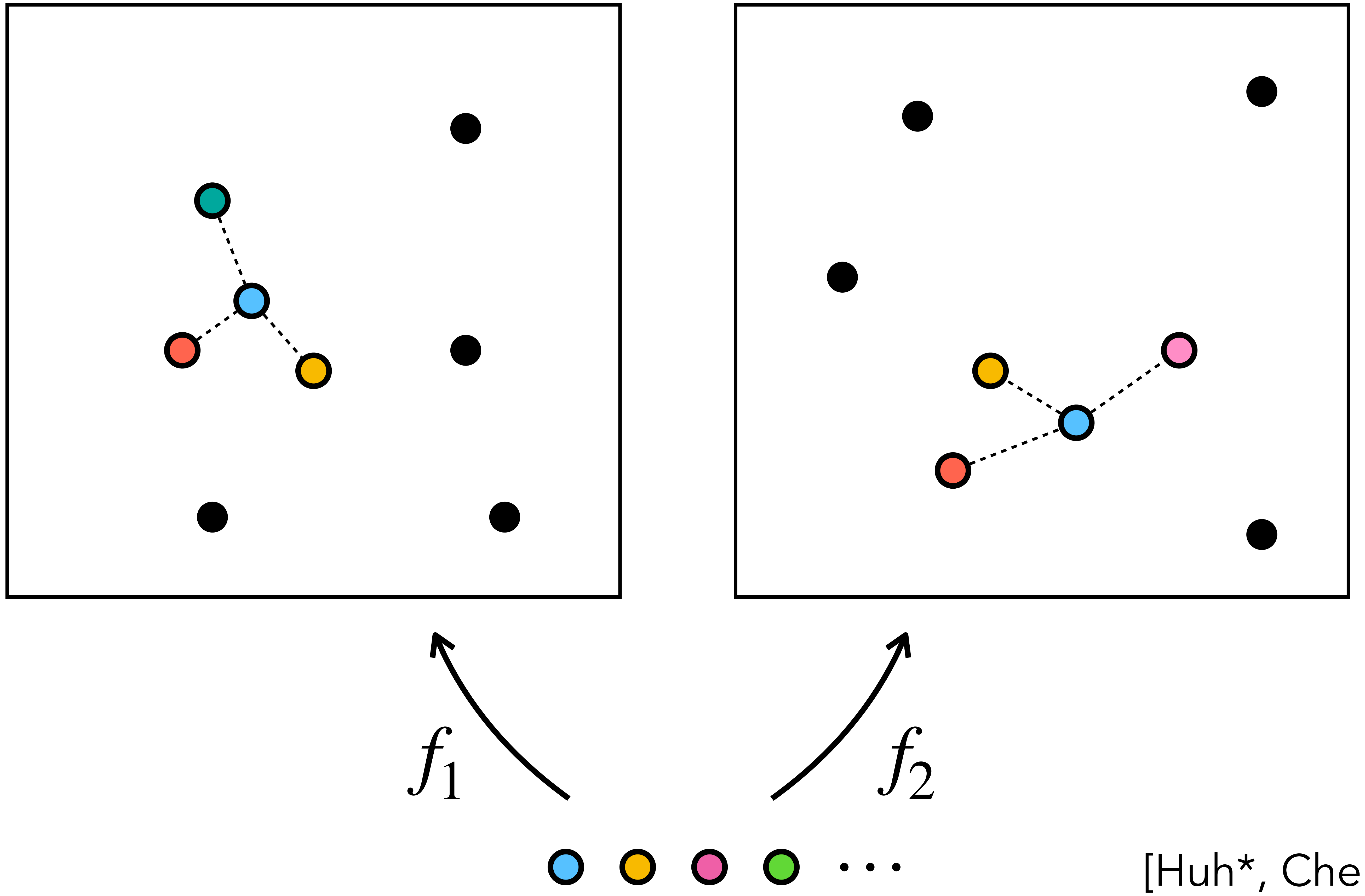
[Kornblith, Norouzi, Lee, Hinton, ICML 2019]

Two equivalent representations under CKA



Nearest-neighbor kernel-alignment metric

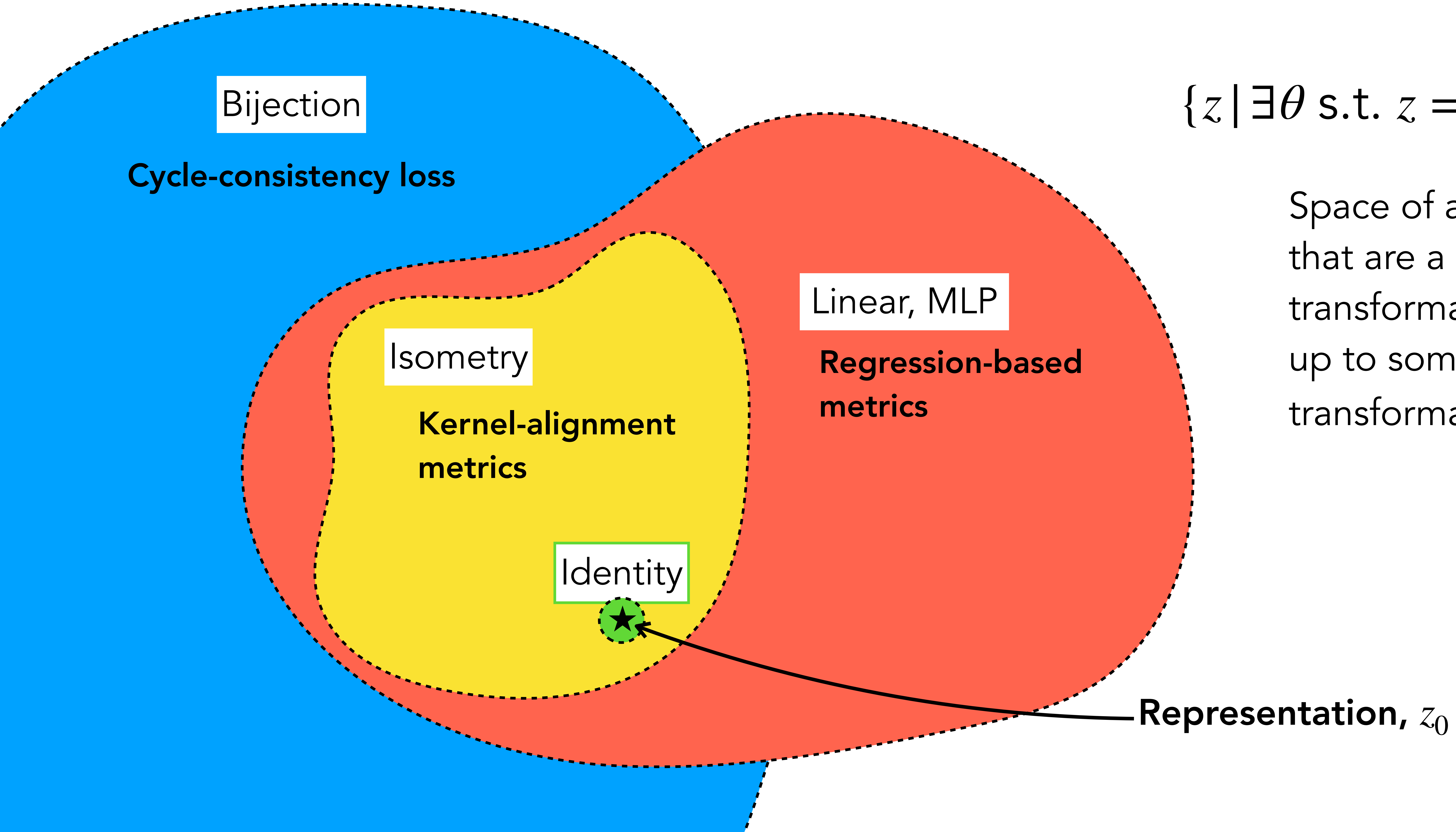
What percent of my nearest-neighbors under representation f are also my nearest neighbors under representation g?



[Huh*, Cheung*, Wang*, Isola, ICML 2024]

[Park et al. (2024), Klabunde et al. (2023) Oron et al. (2017)]

Metrics measure sameness up to a transformation T



$$\{z \,|\, \exists \theta \text{ s.t. } z = T(z_0, \theta)\}$$

Space of all points that are a transformation of z, up to some class of transformations T.

Which way of measuring is best?

- My opinion: **kernel alignment metrics**
- Why? Because *distance* is the thing that matters for most downstream tasks
  - Two representations that are related by an isometry are the same for most practical purposes
  - Linear isometry —> equivalence in: retrieval, k-NN classifier, min-norm linear regression, MLPs in the NTK regime, …
- (We could make this definitional: *a representation is a specificiation of* $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$)

[See more: "Getting Aligned on Representational Alignment," Sucholutsky*, Muttenthaler*, et al. arXiv 2024]

Summary #2:

Representations can be compared via distance functions.

Each distance yields different inferences you can make about how a representation will behave, and what you can do with it.

Outline:

1. What’s a representation?

2. How to measure representational similarity?

3. Which representations are similar and which are different?

4. What drives representational alignment?

5. Making representations more aligned

How different are these images?

$D($  ,  $)$

Fu*, Tamir*, Sundaram*, Chai, Zhang, Dekel, Isola. *DreamSim.* NeurIPS 2023.

Which image is more similar to the middle?

< Clap >

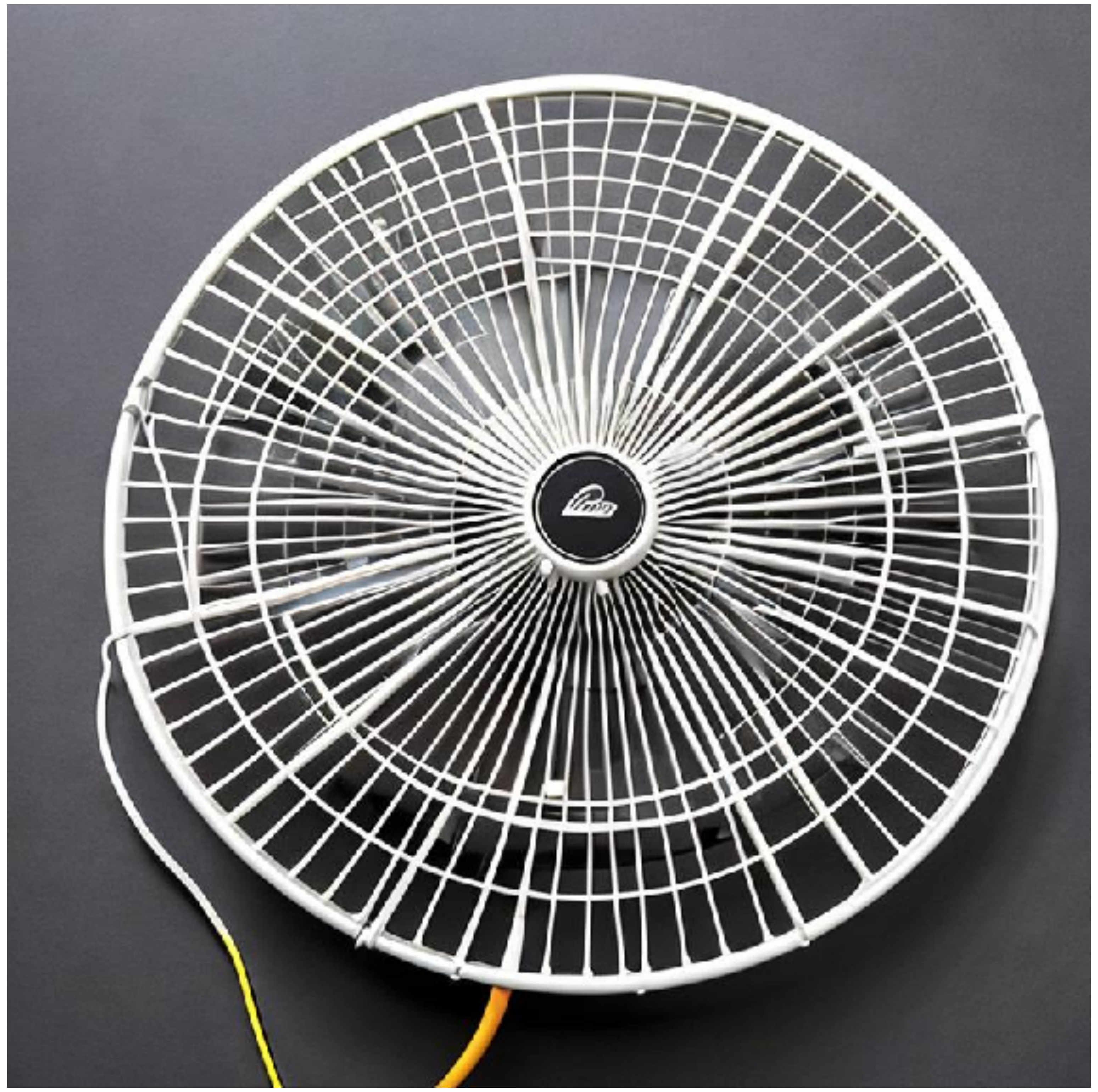
< DINO >

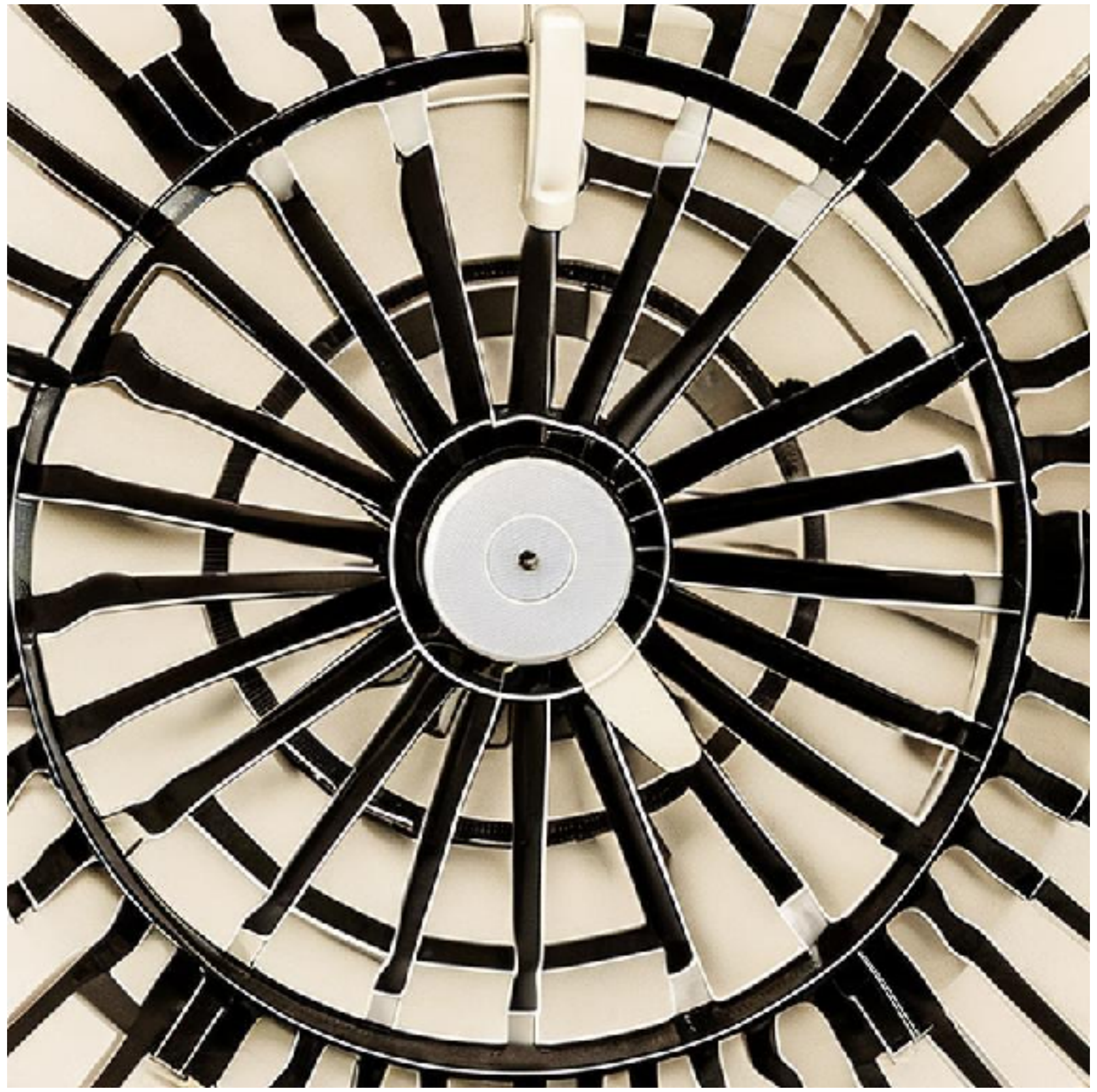
Which image is more similar to the middle?

< Clap >

< DINO >

Which image is more similar to the middle?

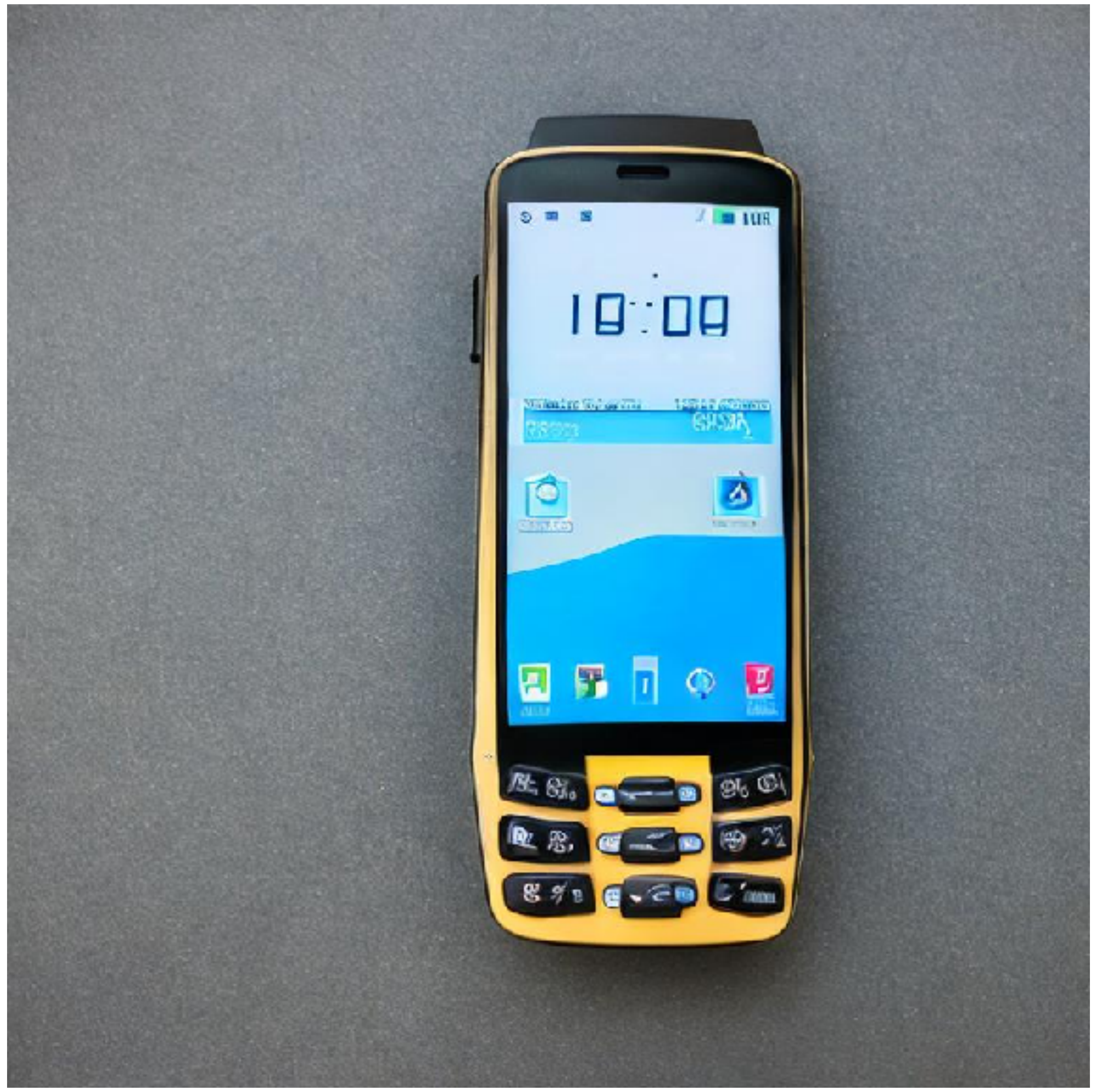


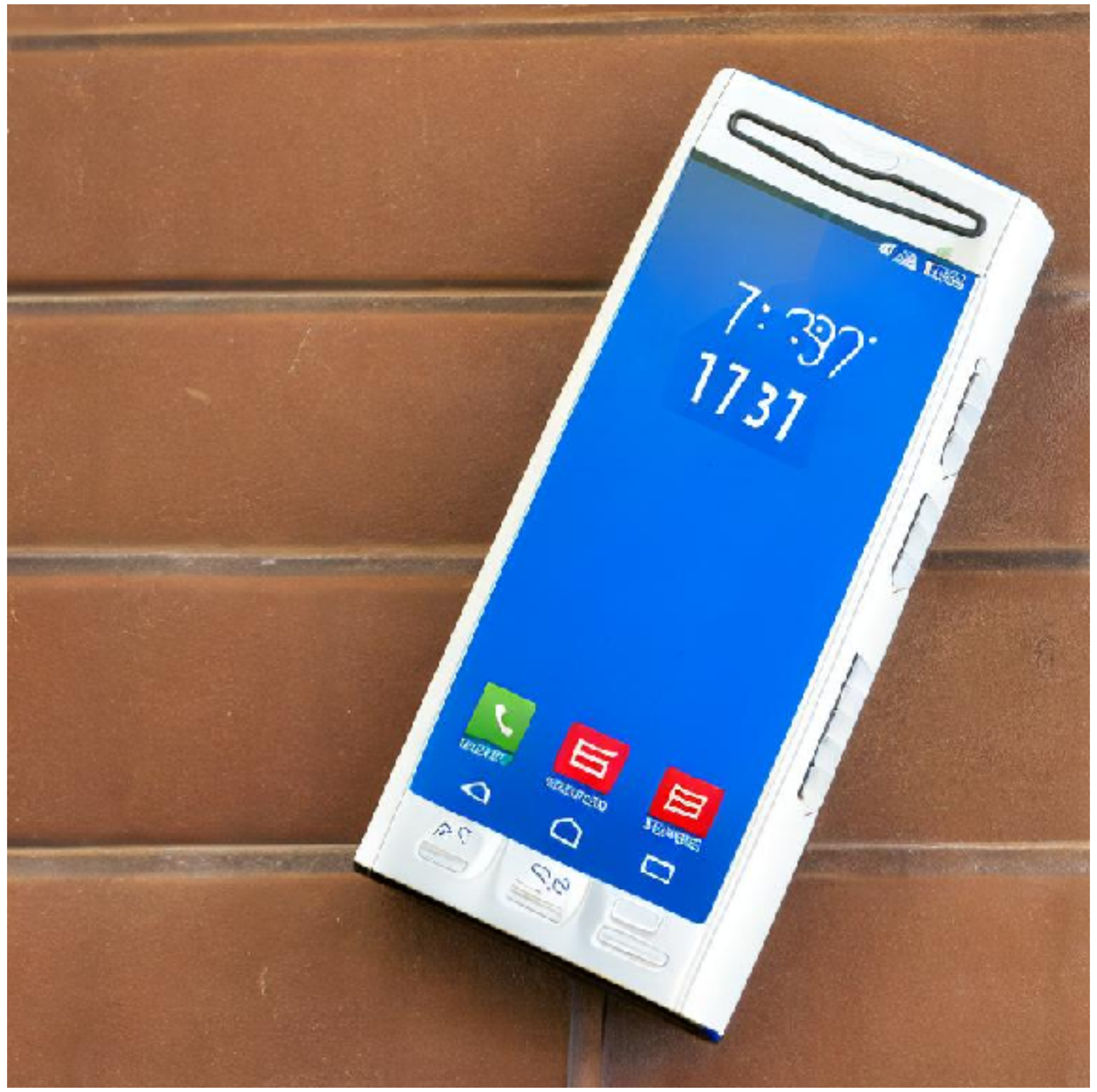


< Clap >

< Clap >

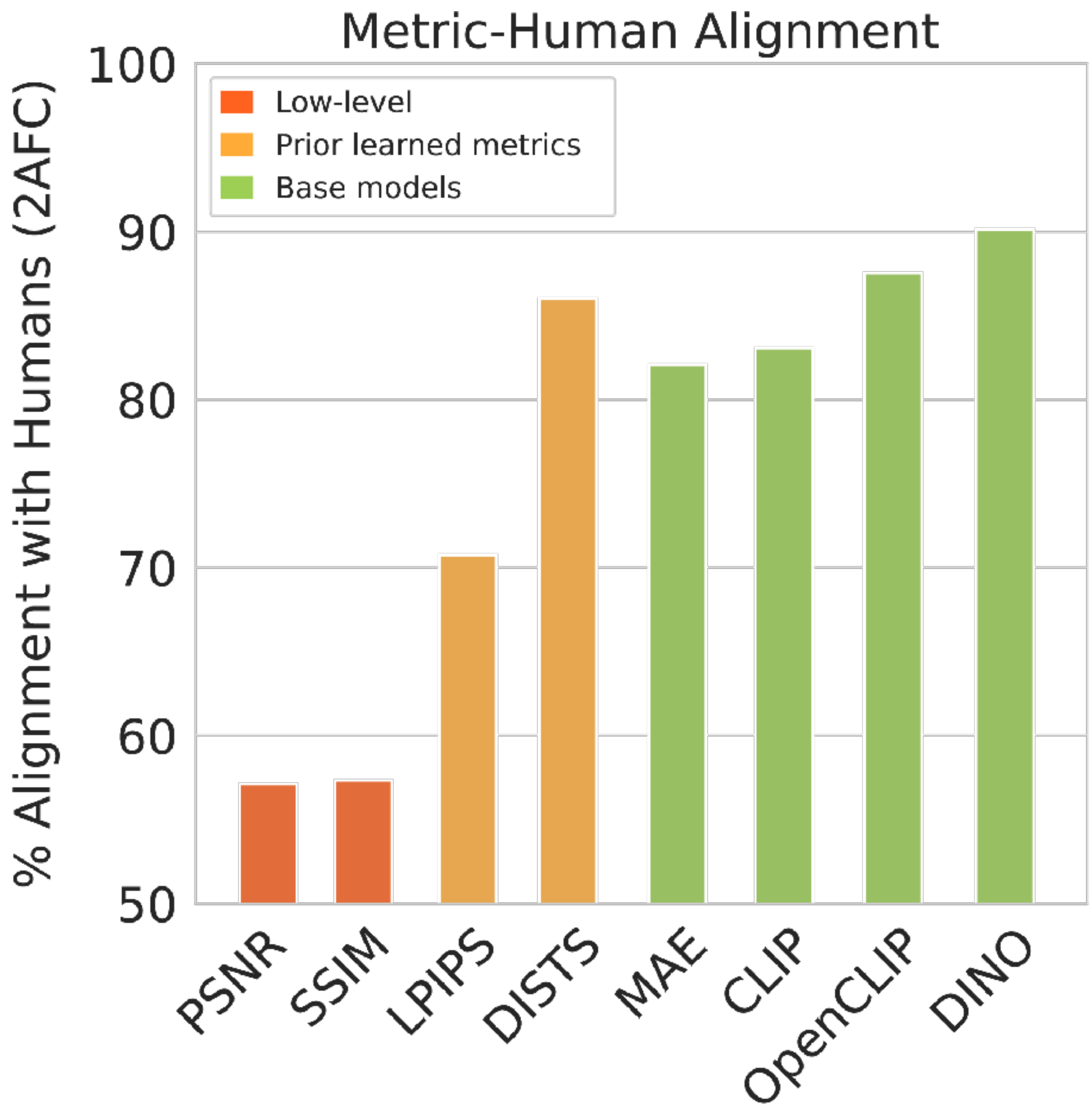
Which image is more similar to the middle?





< Clap >

< Clap >



Fu*, Tamir*, Sundaram*, Chai, Zhang, Dekel, Isola. *DreamSim.* NeurIPS 2023.

Investigating representations in the brain

How similar are these two images?

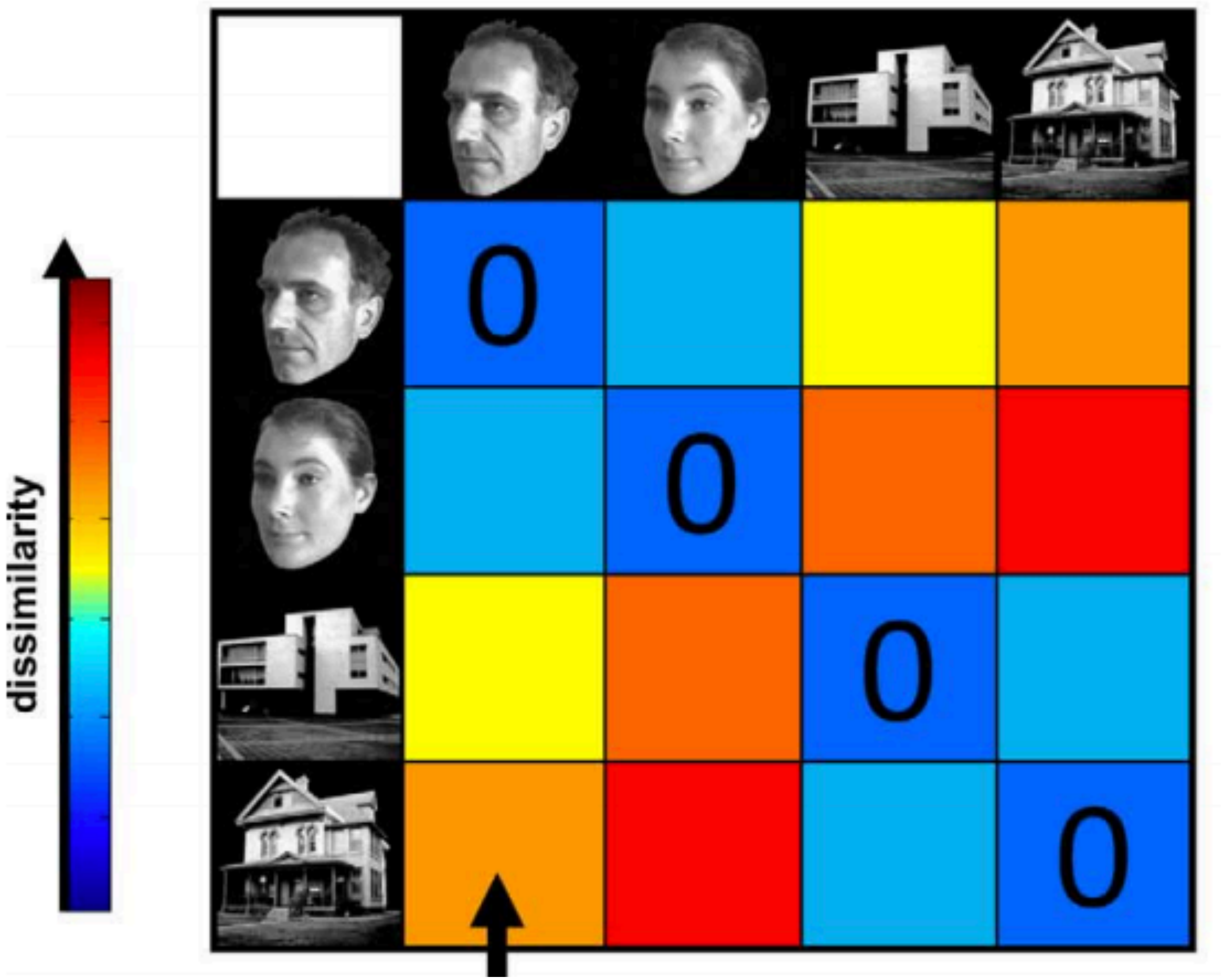
How about these two?

[Kriegeskorte, Mur, Ruff, et al. 2008]

Investigating a representation via similarity analysis

Representational Dissimilarity Matrix



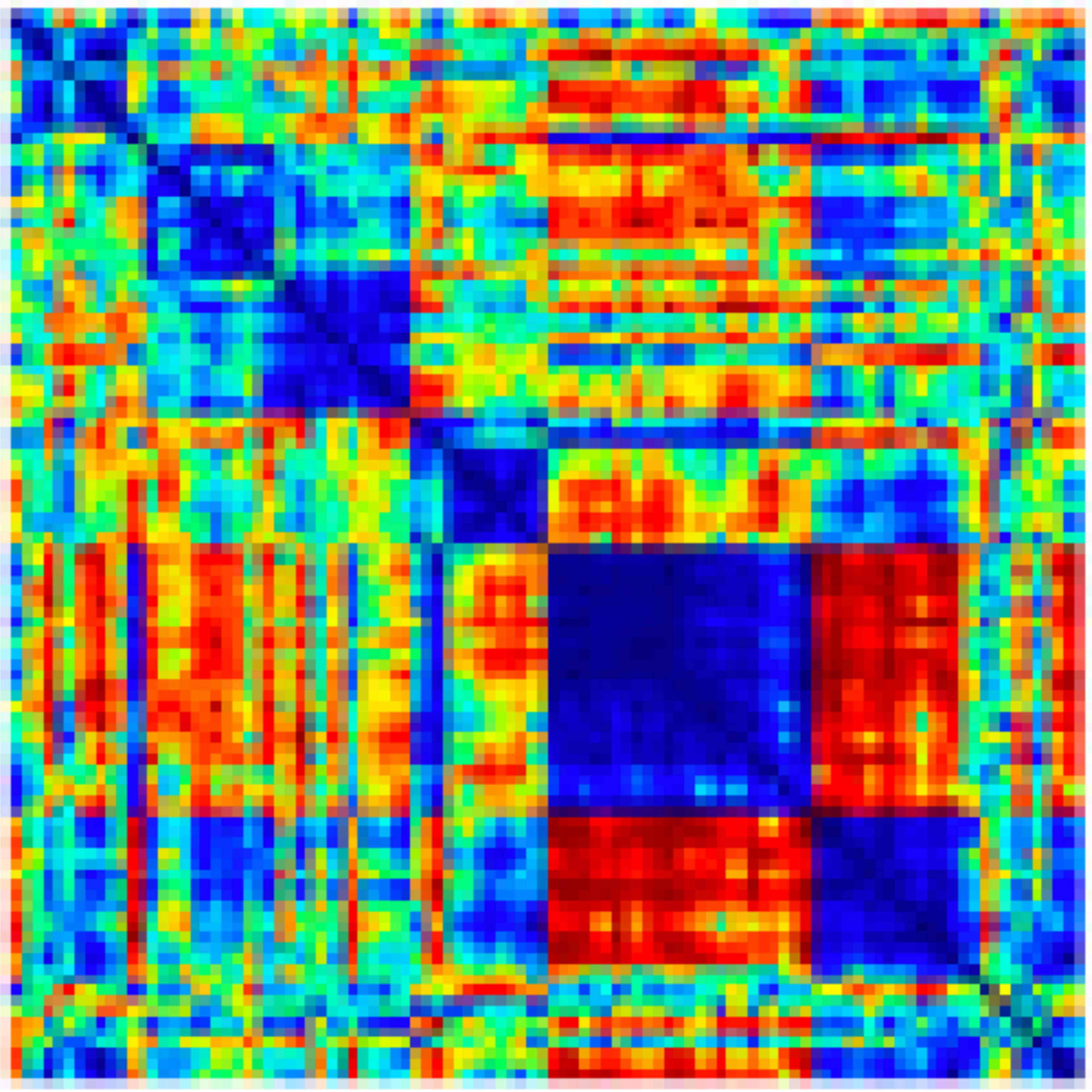
$$||z^{(i)} - z^{(j)}||$$

Neural activation vector

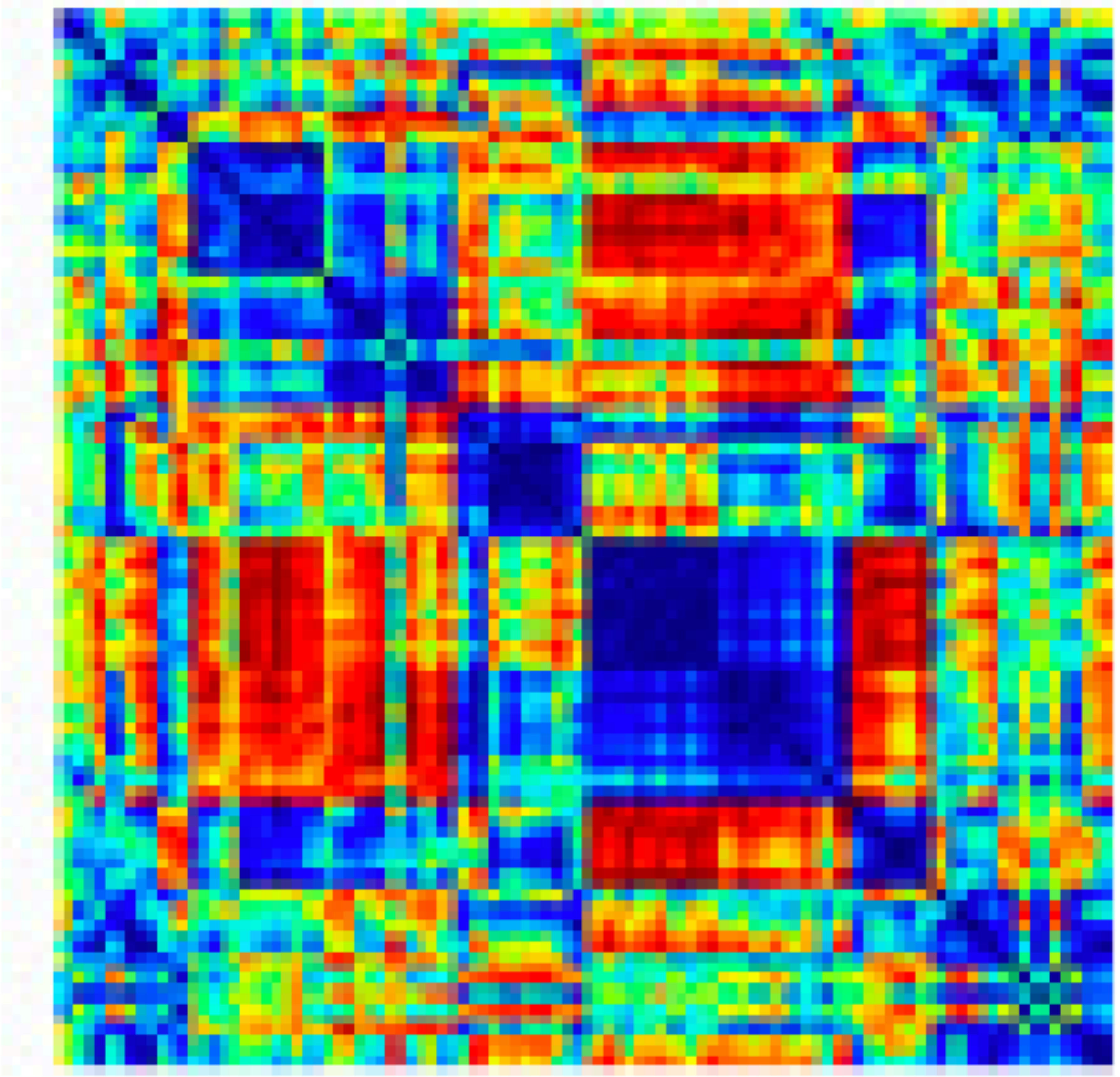
[Kriegeskorte, Mur, Ruff, et al. 2008]

Investigating a representation via similarity analysis

IT Neuronal Units



Deep net (in paricular, HMO)



[Yamins, Hong, Cadieu, Solomon, Seibert, DiCarlo, PNAS 2014]

What's the color space in which a language (model) sees?

Color space: a mapping from a spectral power distribution to 3 numbers

- Camera CCD: RGB color space
- Human vision: Lab color space

How did we determine this for humans?

- Ask them which colors are similar and which are different
- Find a 3D projection that best preserves distances

What's the color space in which a language (model) sees?

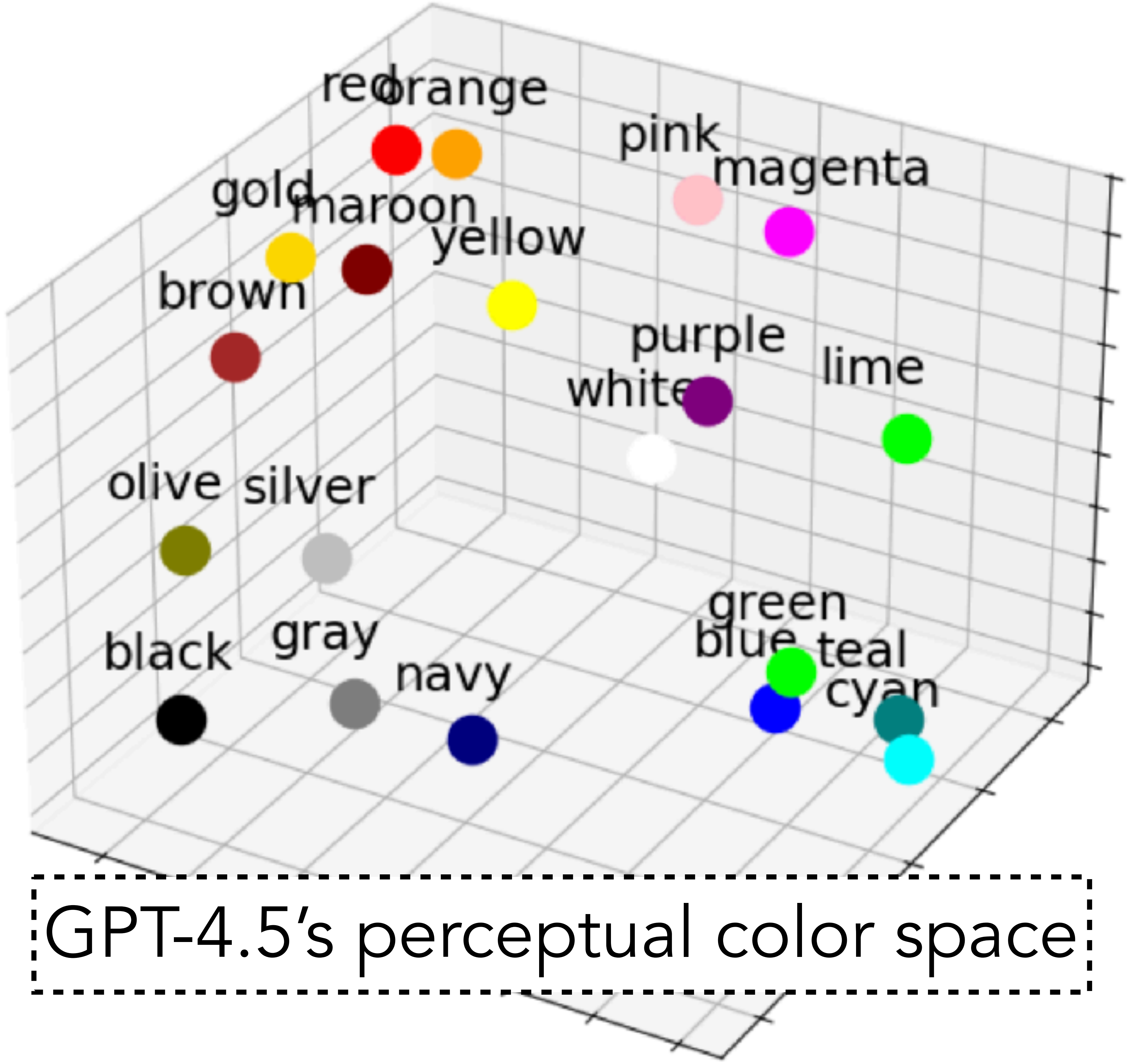
- Ask an LLM which colors are similar and which are different
- Find a 3D projection that best preserves distances

How similar is red to orange. Output a single number between 0 and 1.

0.8

What's the color space in which a language (model) sees?

- Ask an LLM which colors are similar and which are different
- Find a 3D projection that best preserves distances

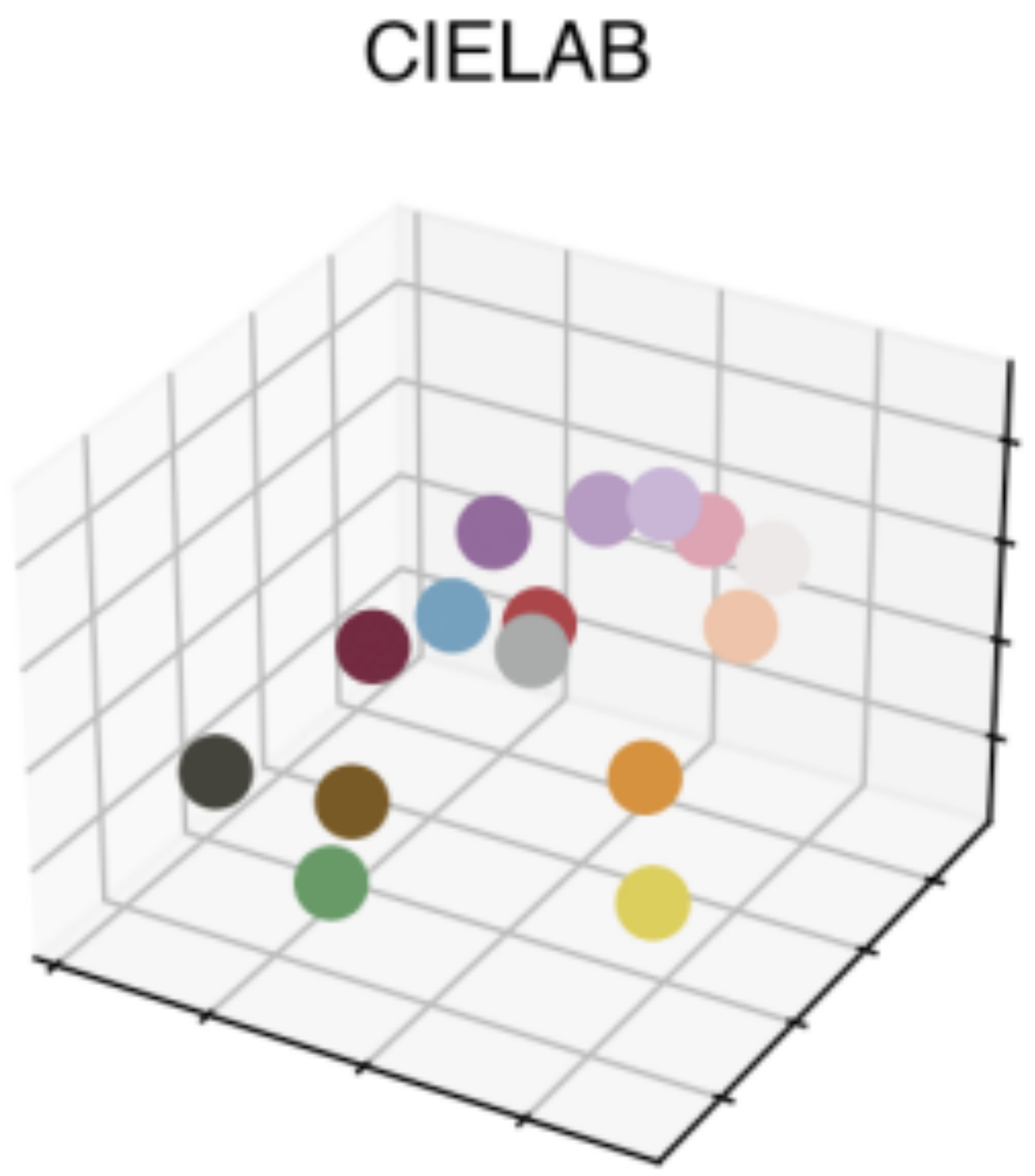


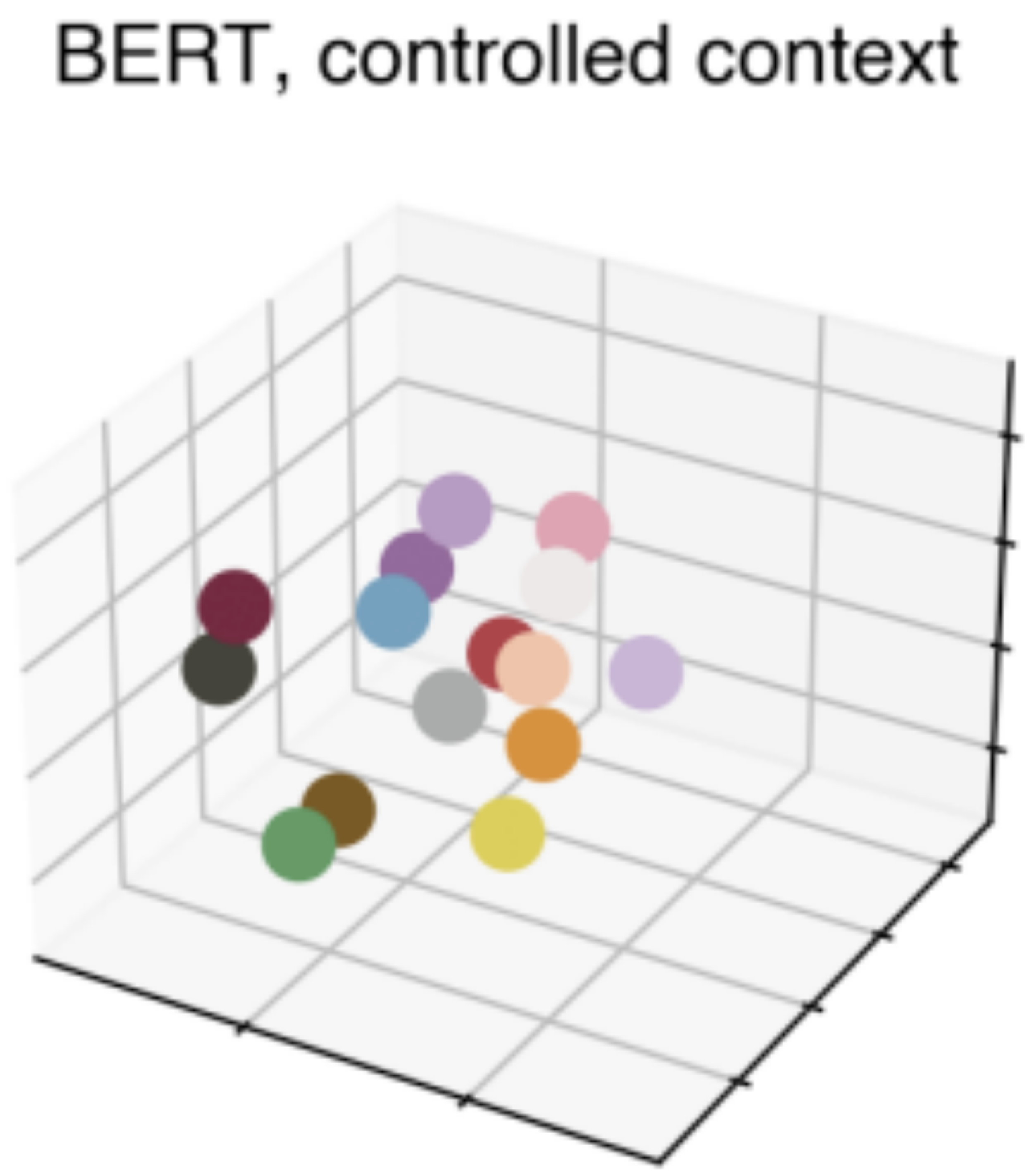
GPT-4.5's perceptual color space

but maybe it lied…

What's the color space in which a language (model) sees?

- Measure distance between LLM embeddings of different color words
- Find a 3D projection that best preserves distances





[Abdou et al. 2021]

Brains vs Machines

Deep nets and the human/primate brain both learn similar metric spaces.

Deep nets organize visual information similarly to how our brains do!

[Yamins, Hong, Cadieu, Solomon, Seibert, DiCarlo, PNAS 2014]

Alignment between different computer vision systems

[These slides from: Huh*, Cheung*, Wang*, Isola*, ICML 2024]

Experiment: Is alignment between vision models increasing as vision systems become stronger?

Hypothesis 1:
There are many different ways one can represent the visual world, and each can be highly effective.

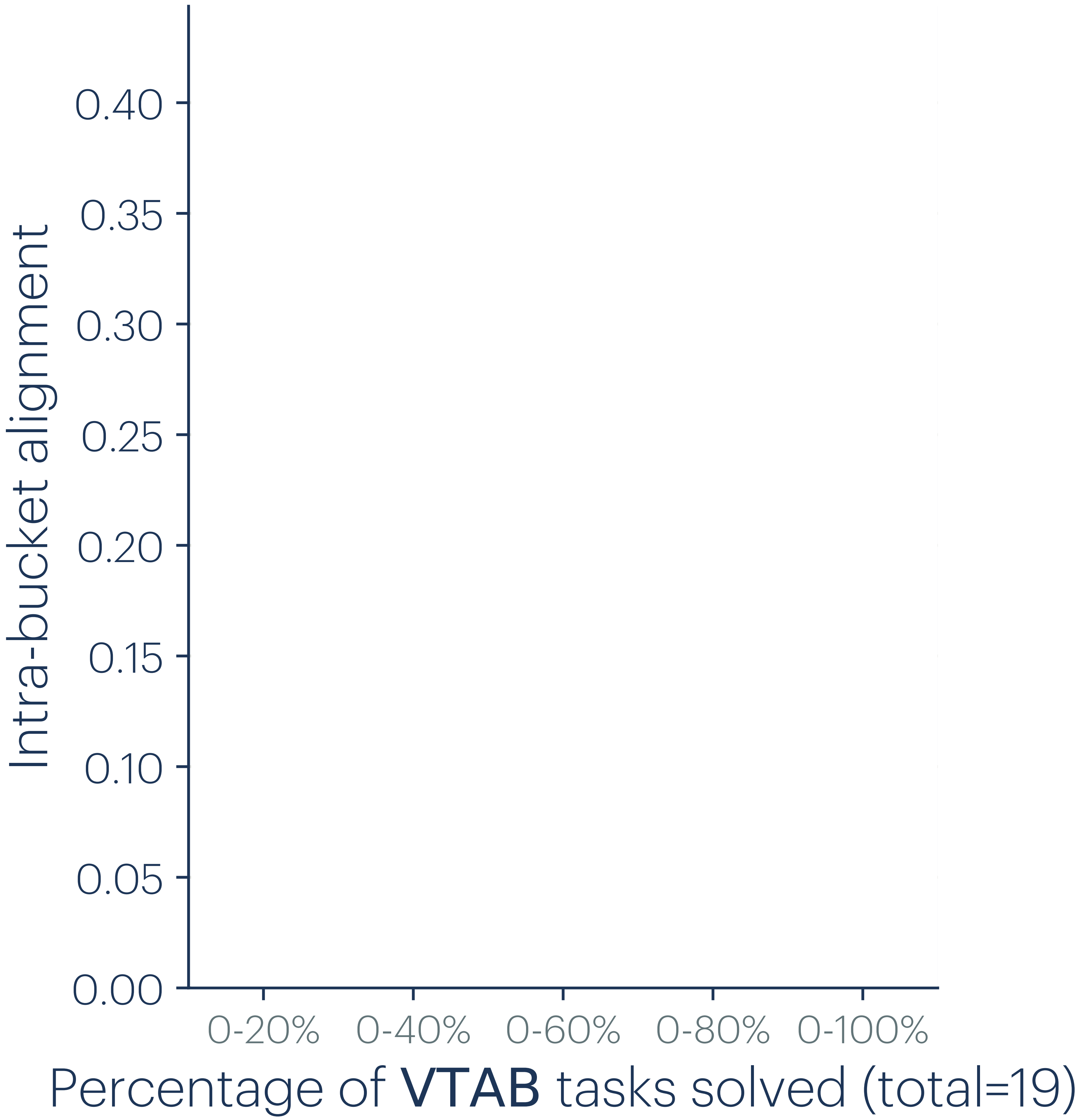
Hypothesis 2:
All strong visual representations are alike.

["Anna Karenina scenario," Bansal et al. 2021]

Experiment: Is alignment between vision models increasing as vision systems become stronger?

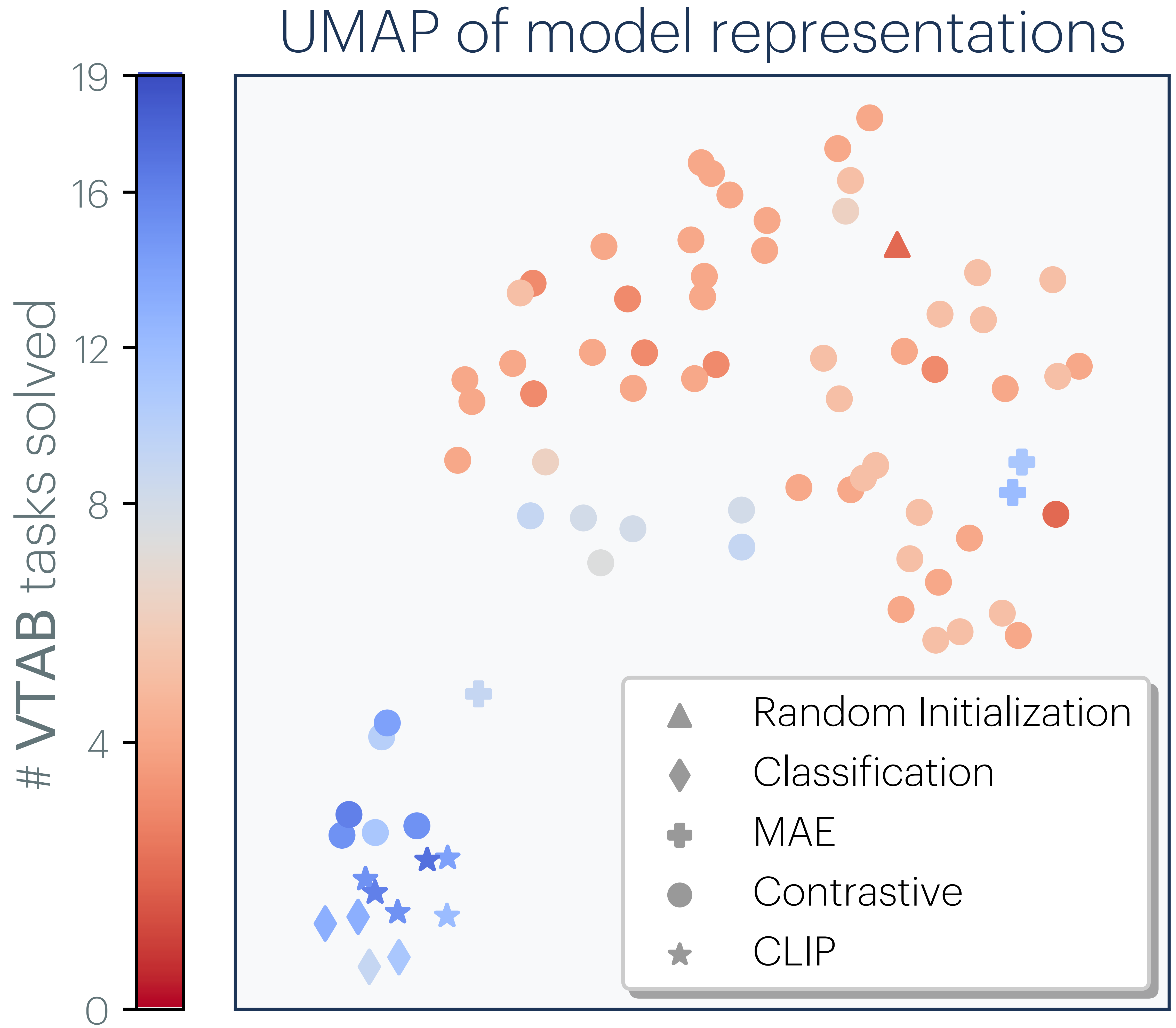
- 78 vision models: different architectures, objectives, training data distributions.

- Group models by performance on VTAB, and measure representational similarity within each group.



Experiment: Is alignment between vision models increasing as vision systems become stronger?

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



Alignment between different modalities

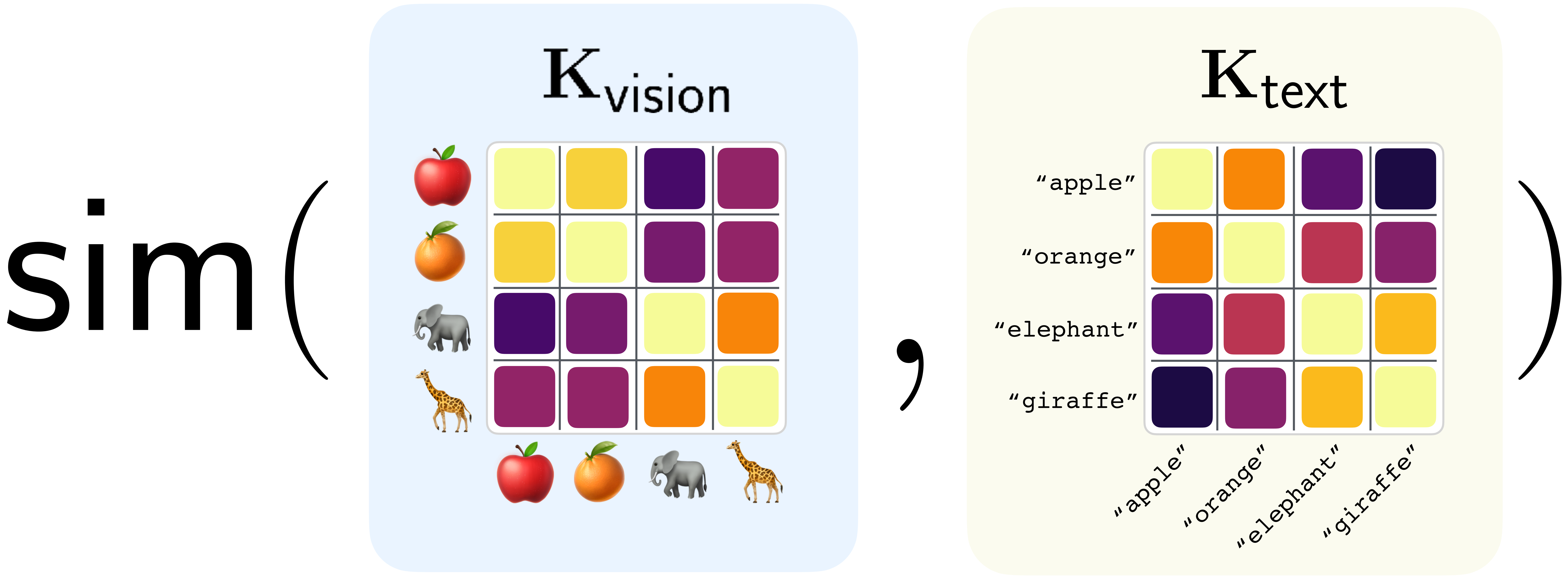
[These slides from: Huh*, Cheung*, Wang*, Isola*, ICML 2024]

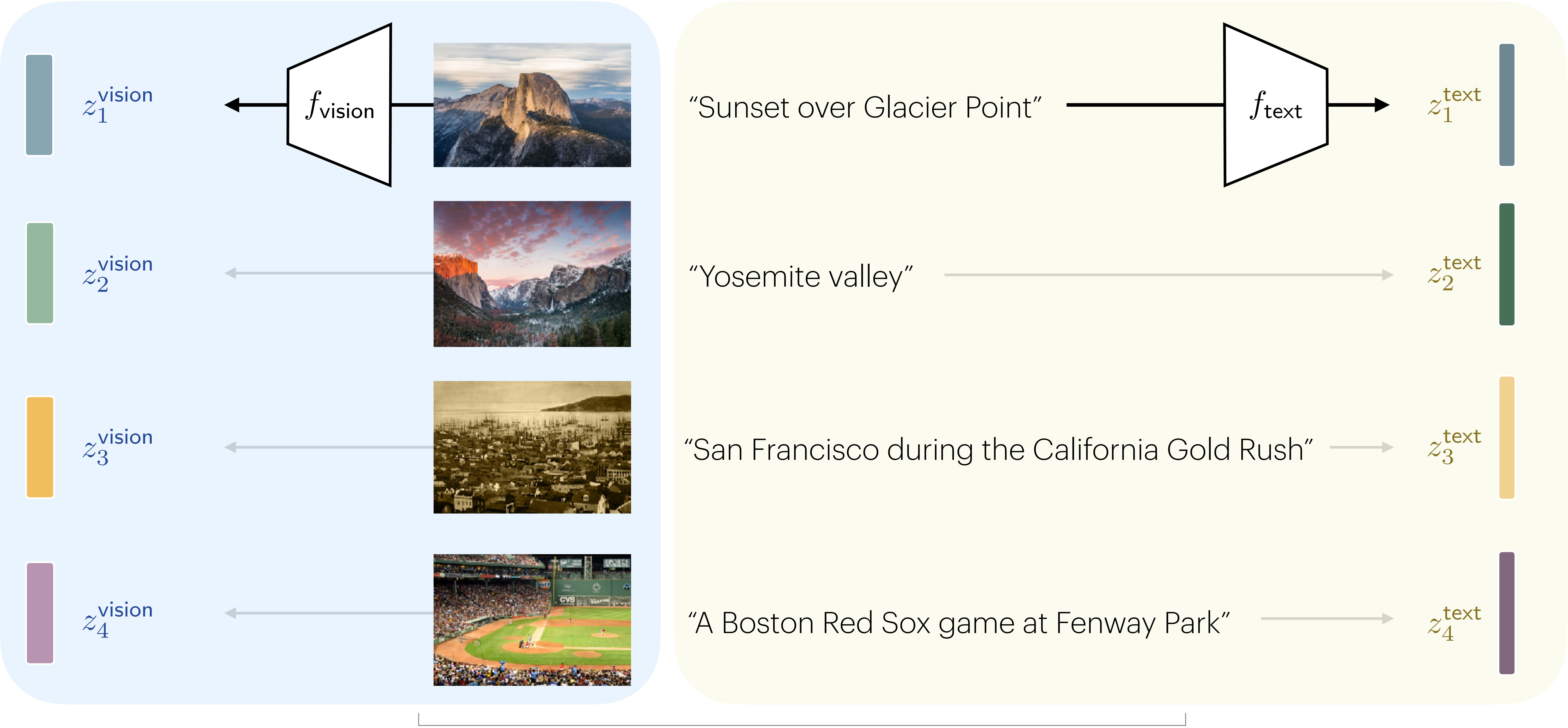
Experiment: Is language-vision alignment increasing?

Hypothesis 1:
As language models get better and better, they will become more and more specific to language, and start being less generally useful for vision.

Hypothesis 2:
Better language models are better vision models.

Hypothesis 2+:
The best language model is the best vision model. They converge to the *same* representation.

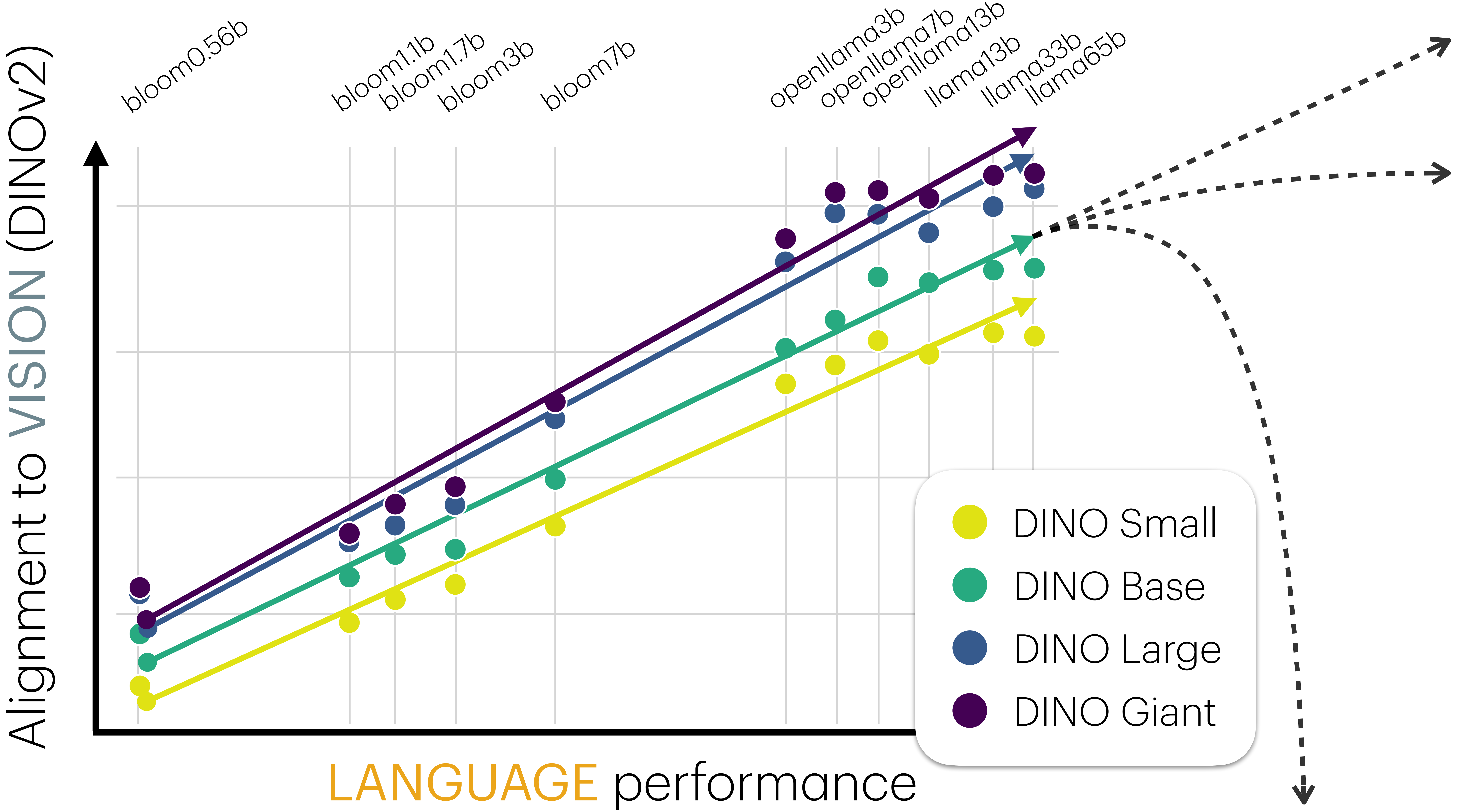

sim
K_{vision}
K_{text}
“apple”
“orange”
“elephant”
“giraffe”
“apple”
“orange”
“elephant”
“giraffe”



Wikipedia Image Text Dataset

[Srinivasan, Raman, Chen, Bendersky, Najork 2021]

Strong models converge in representation



Ongoing science!

Summary #3:

Humans and deep nets both measure distances between images in similar ways.

Different vision models, and language models, seem to be converging in how they measure distanes.

Outline:

1. What’s a representation?

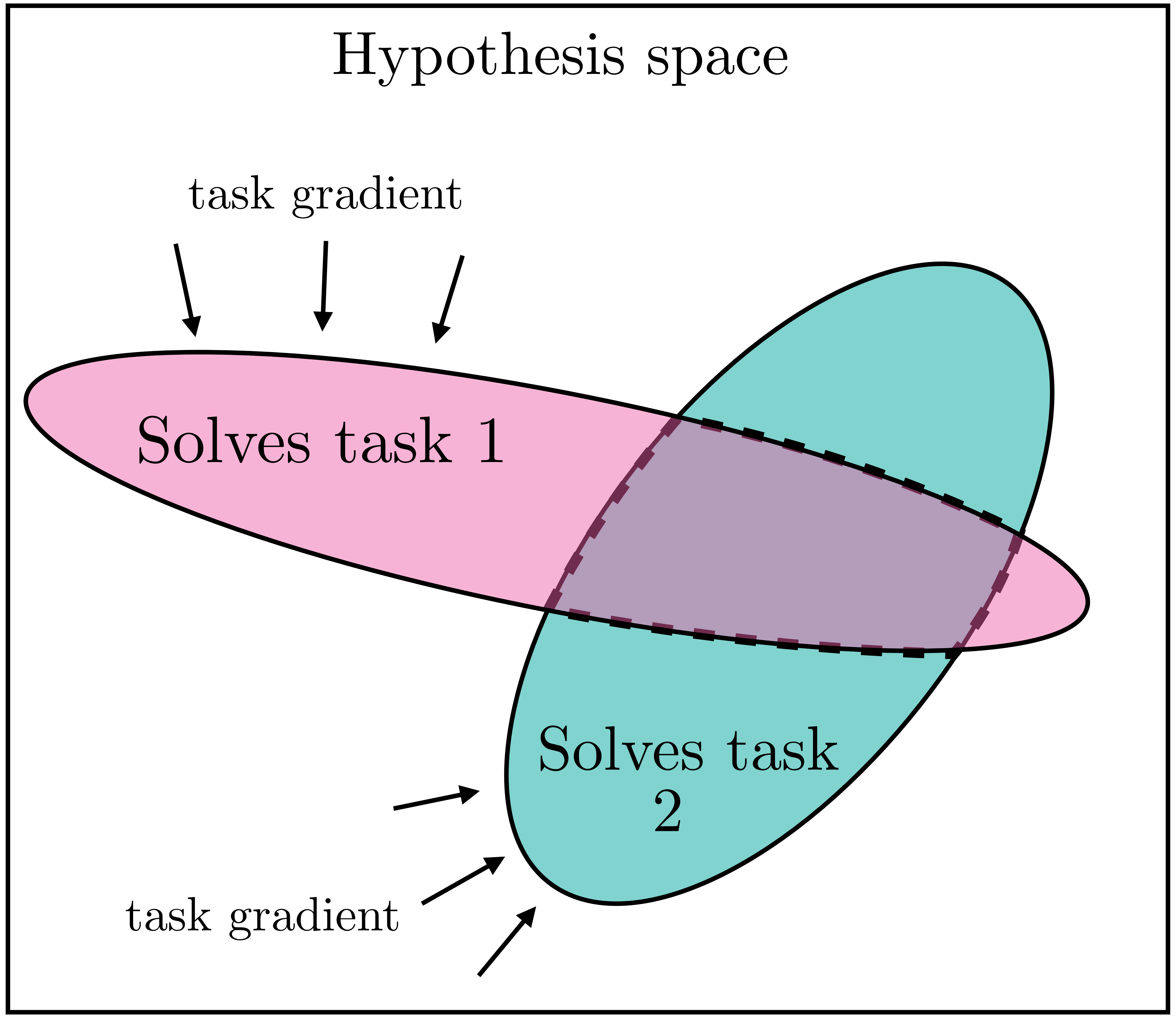
2. How to measure representational similarity?

3. Which representations are similar and which are different?

4. What drives representational alignment?

5. Making representations more aligned

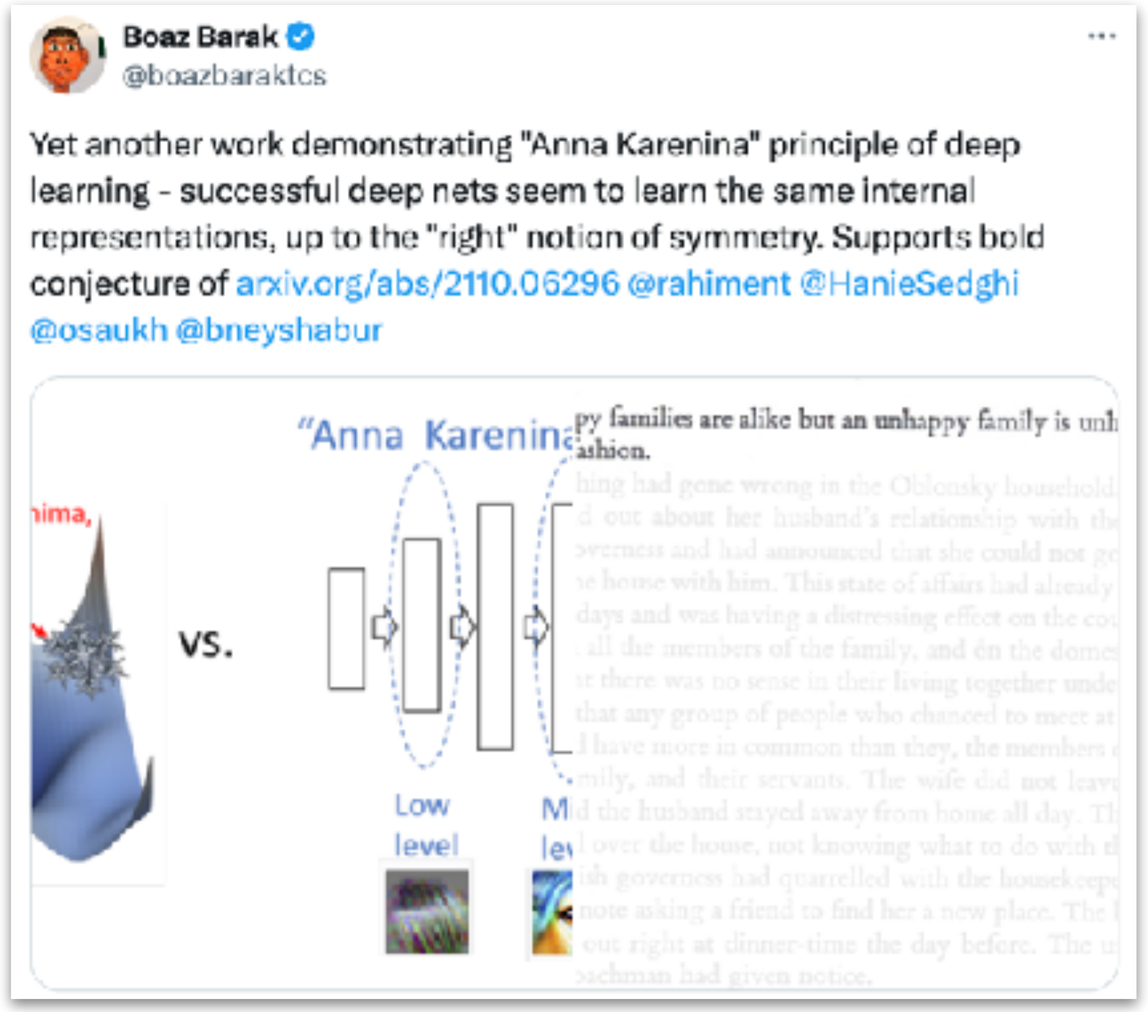
The Multitask Scaling Hypothesis



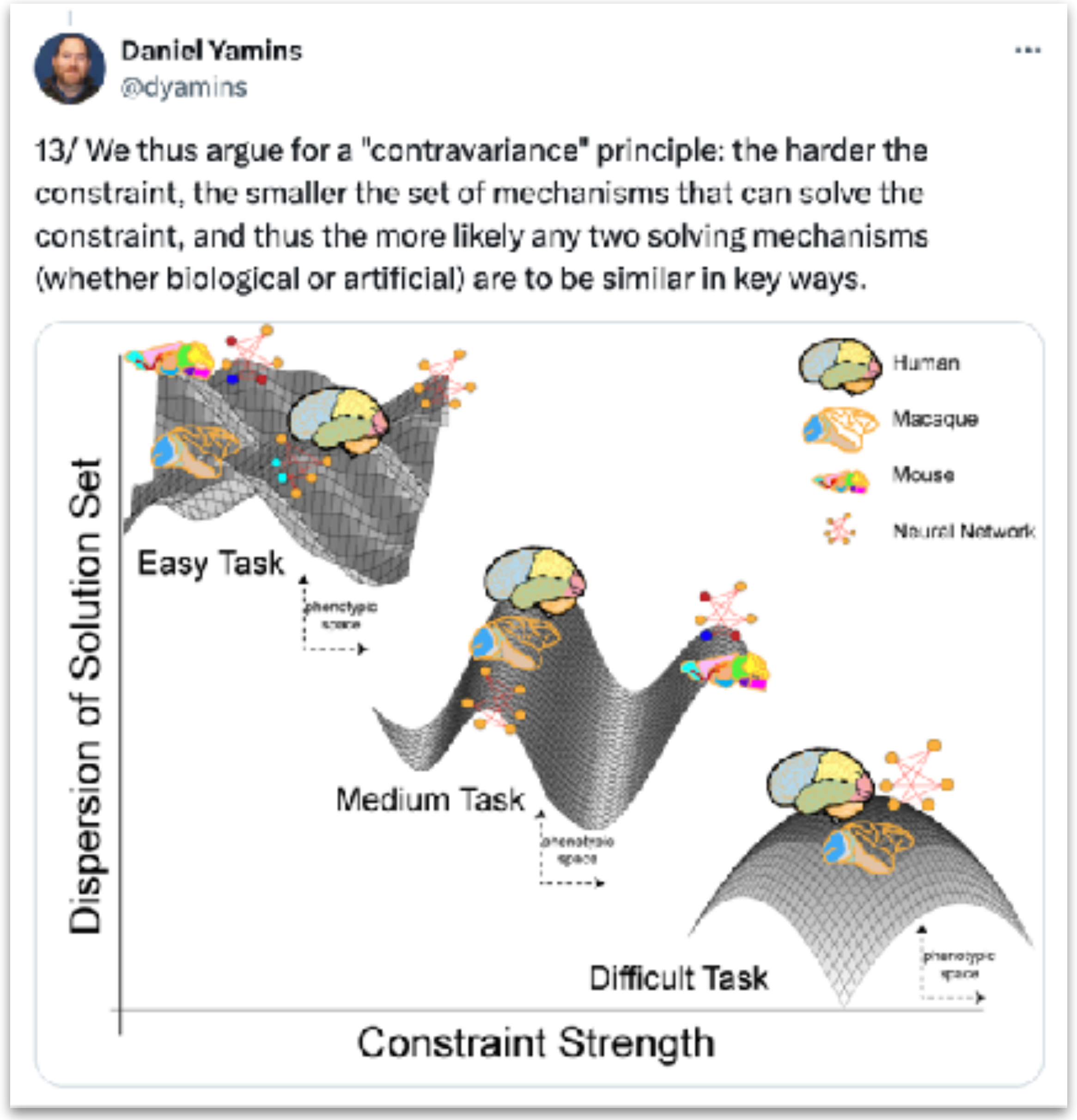
The Multitask Scaling Hypothesis

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.

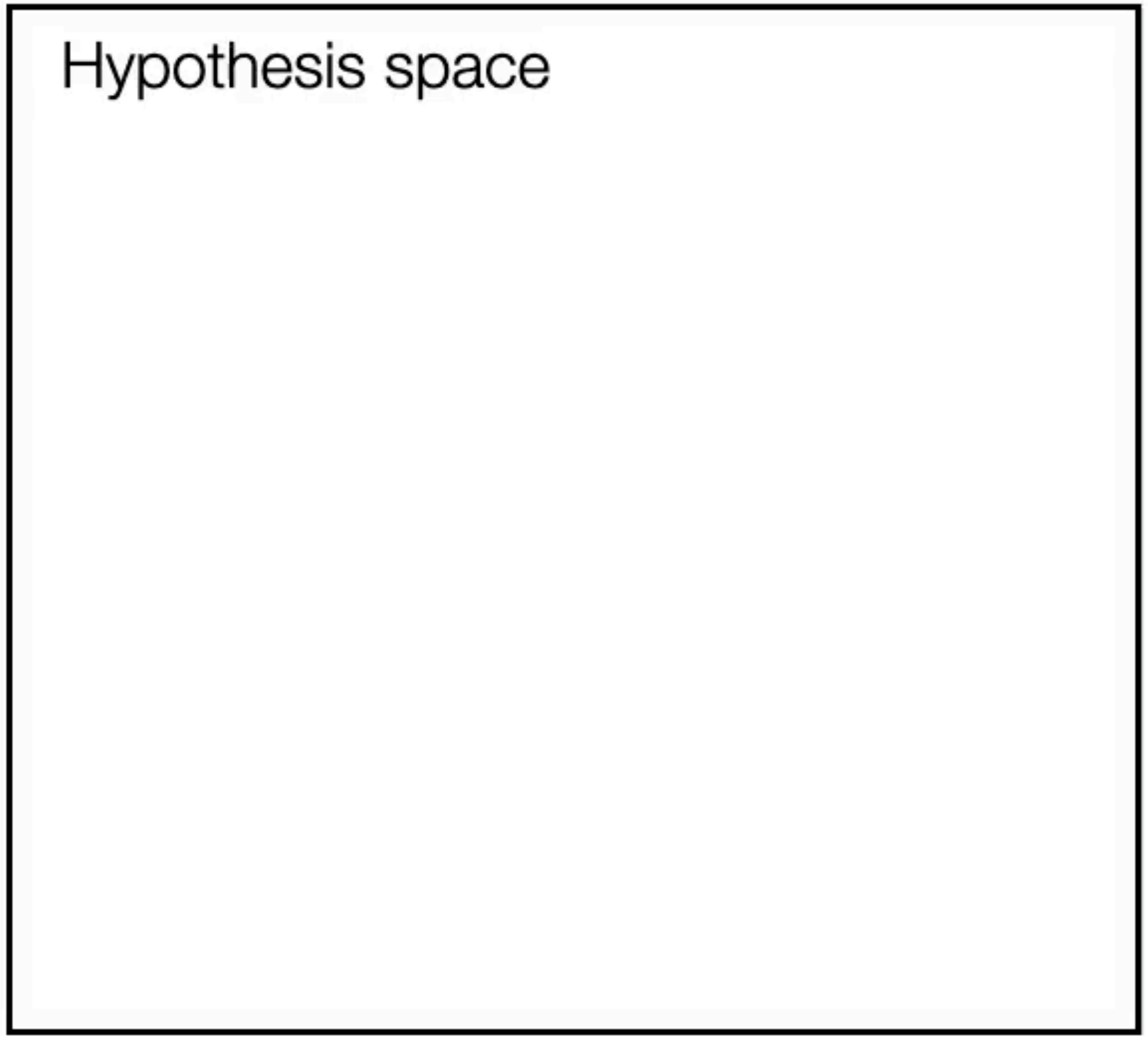
"Anna Karenina principle"



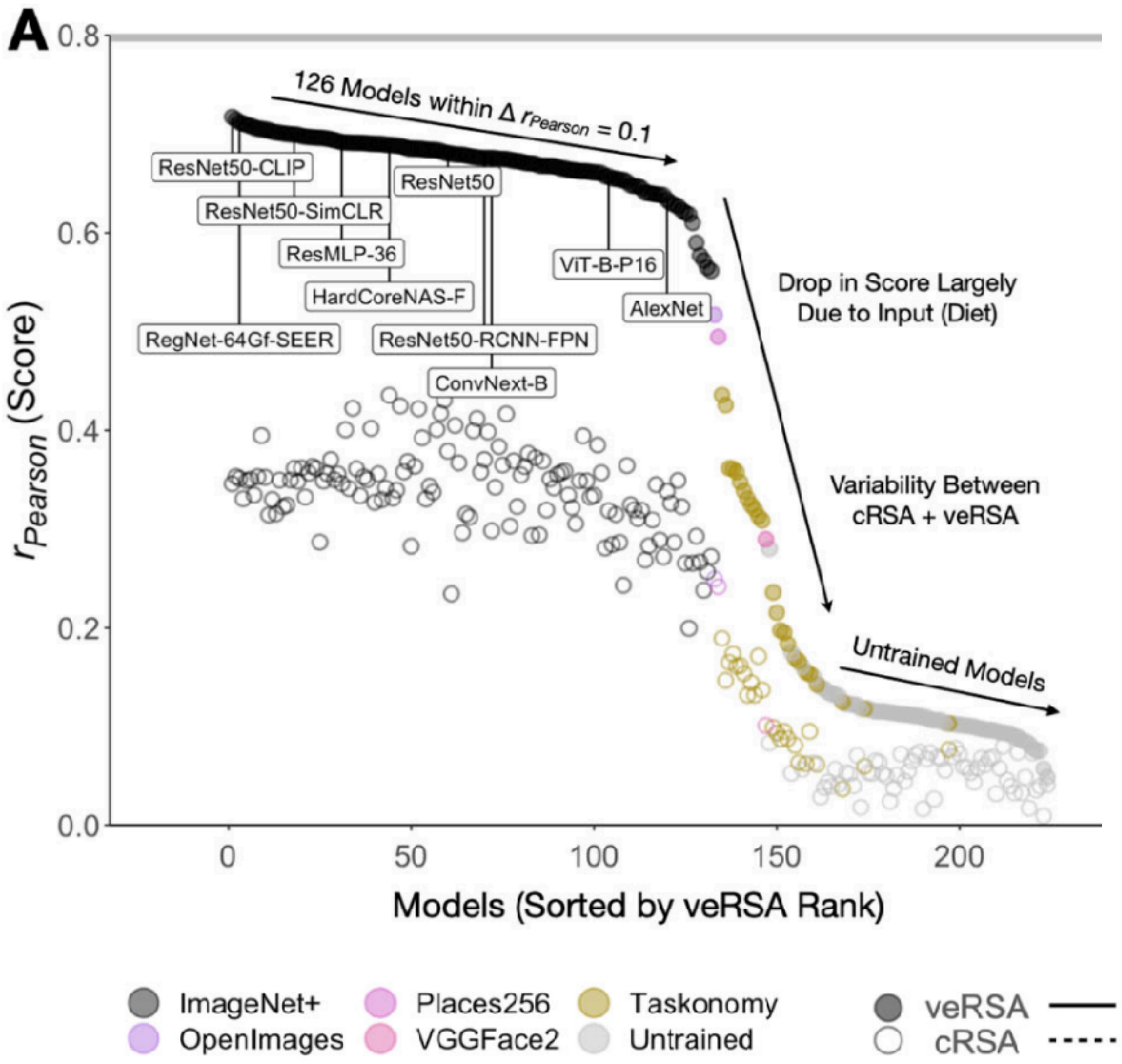
"Contravariance Principal"



The Multitask Scaling Hypothesis



Corollary: more data —> more convergence



Conwell et al. 2024 found that, of the factors they tested, **data diet** plays the greatest role in determining brain-machine alignment.

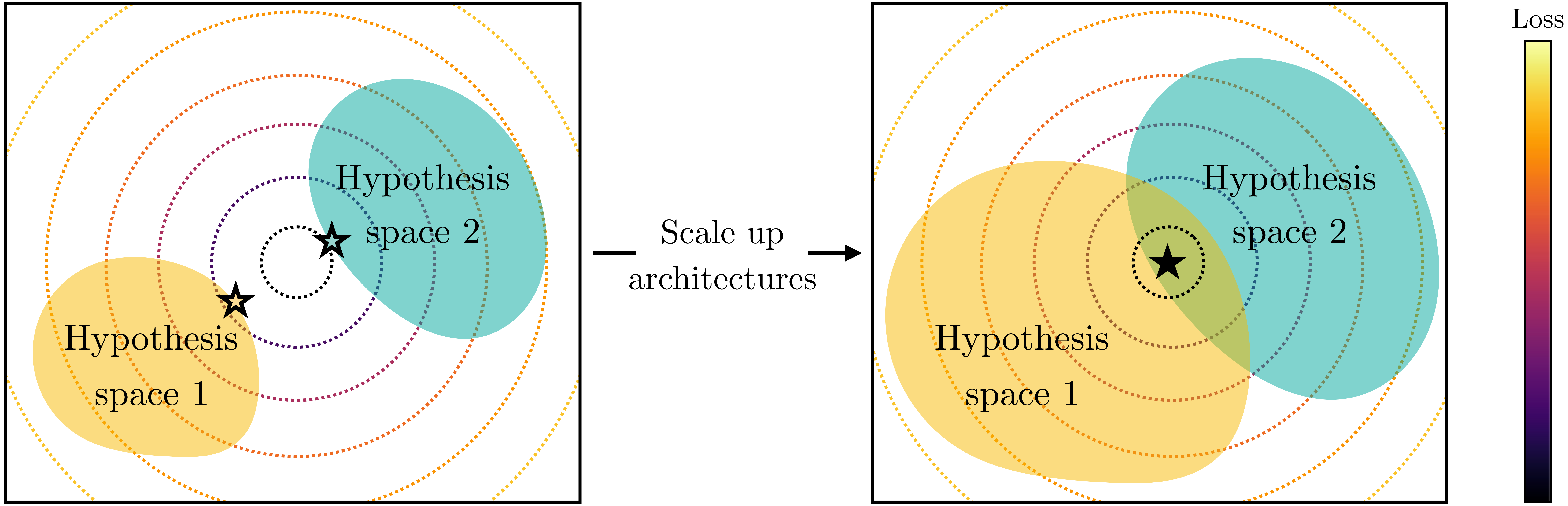
Models trained on more data are more aligned with the brain.

[Conwell, Prince, Kay, Alvarez, Konkle, Nature Communications 2024]

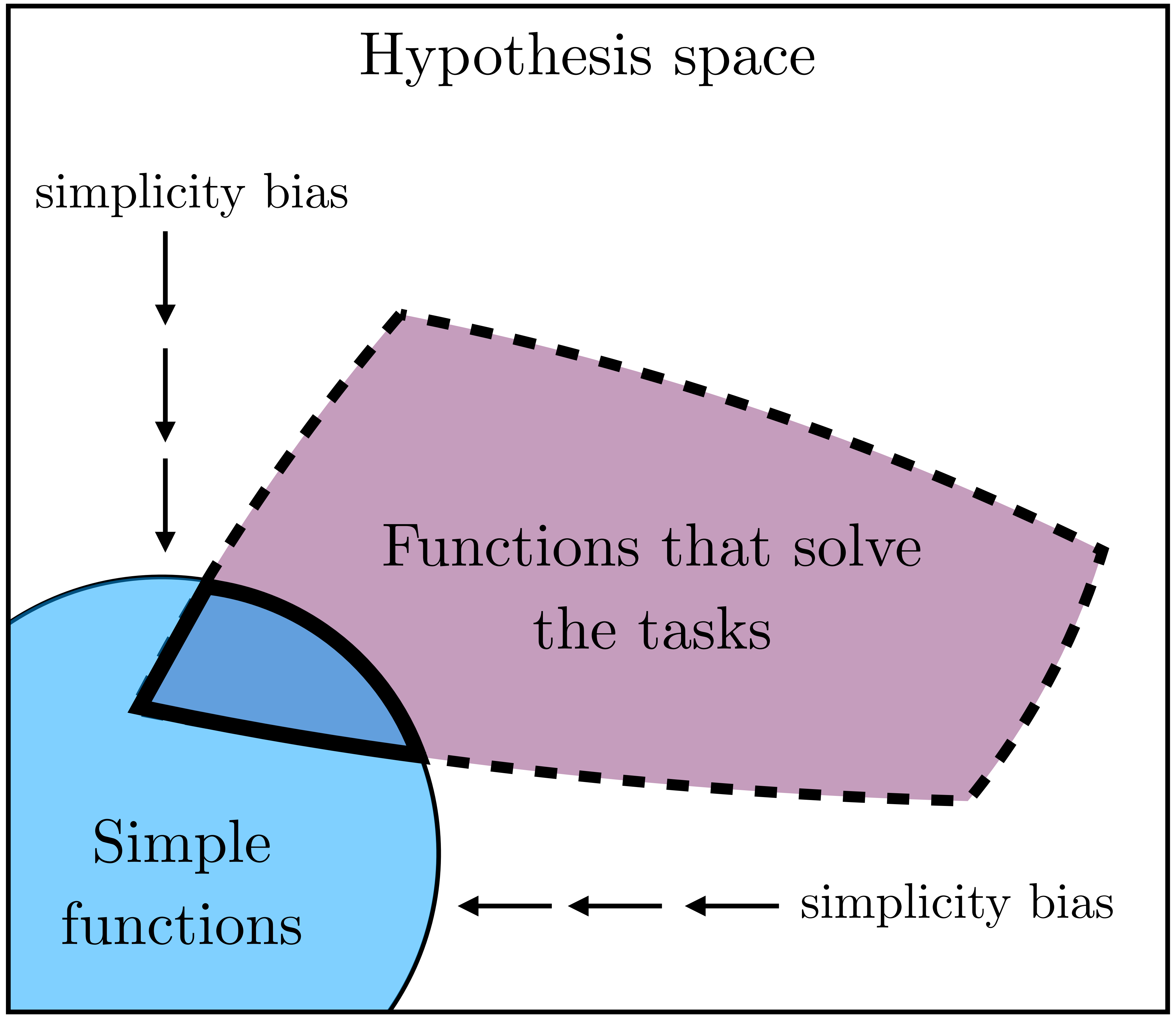
The Capacity Hypothesis

The Capacity Hypothesis

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

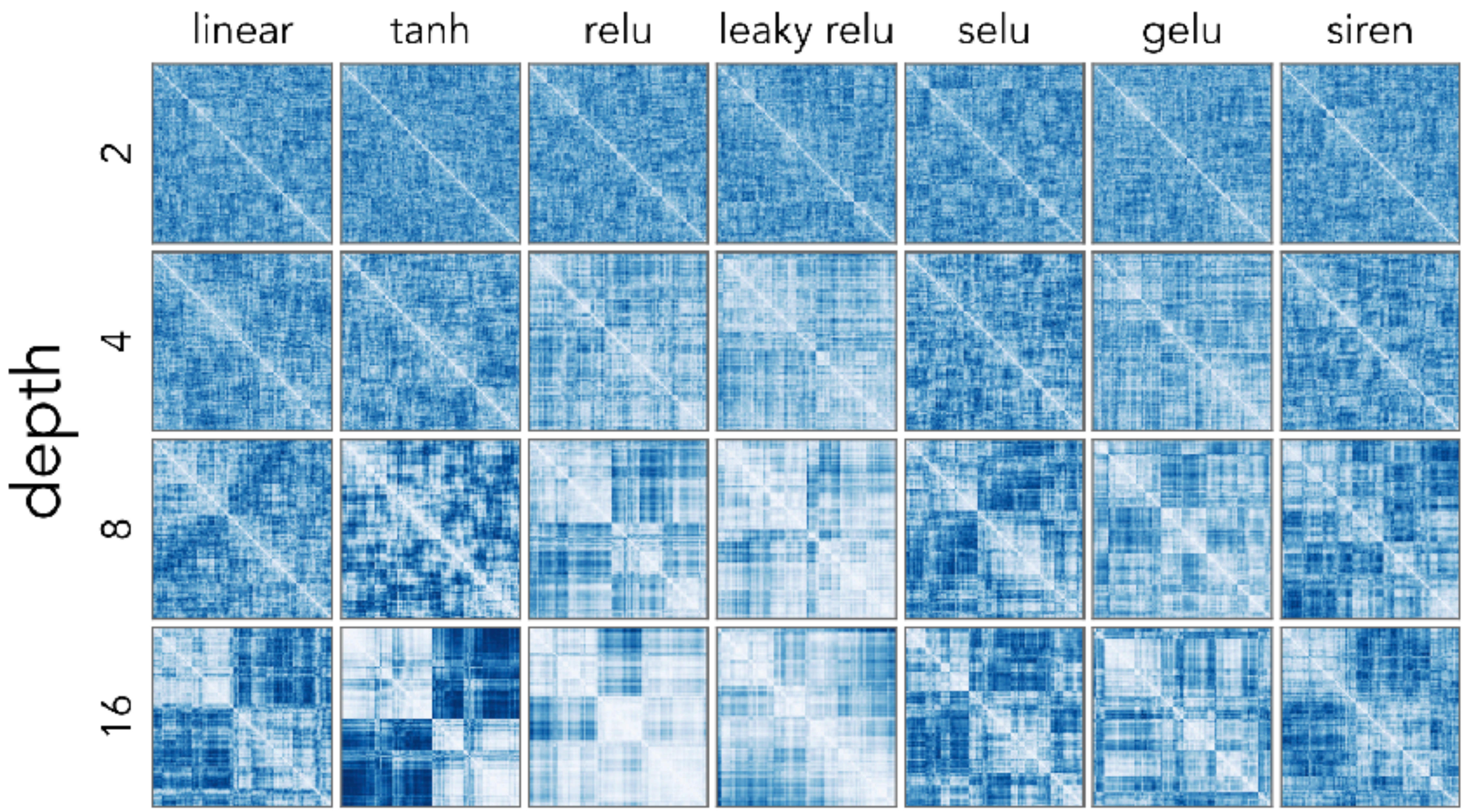


The Simplicity Bias Hypothesis



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.



[Gunasekar et al. 2018, Arora et al. 2019, Valle-Perez et al. 2019., Huh et al. 2023, etc]

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

Task/data pressures	Regularization	Model size

"Contravariance Principal" [Cao & Yamins 2024]

LVX VENIT IN MVNDVN ET DILEXERVNT HOMINES MAGIS TENEBRAS QVAM LVCEM. Io. 3. 19

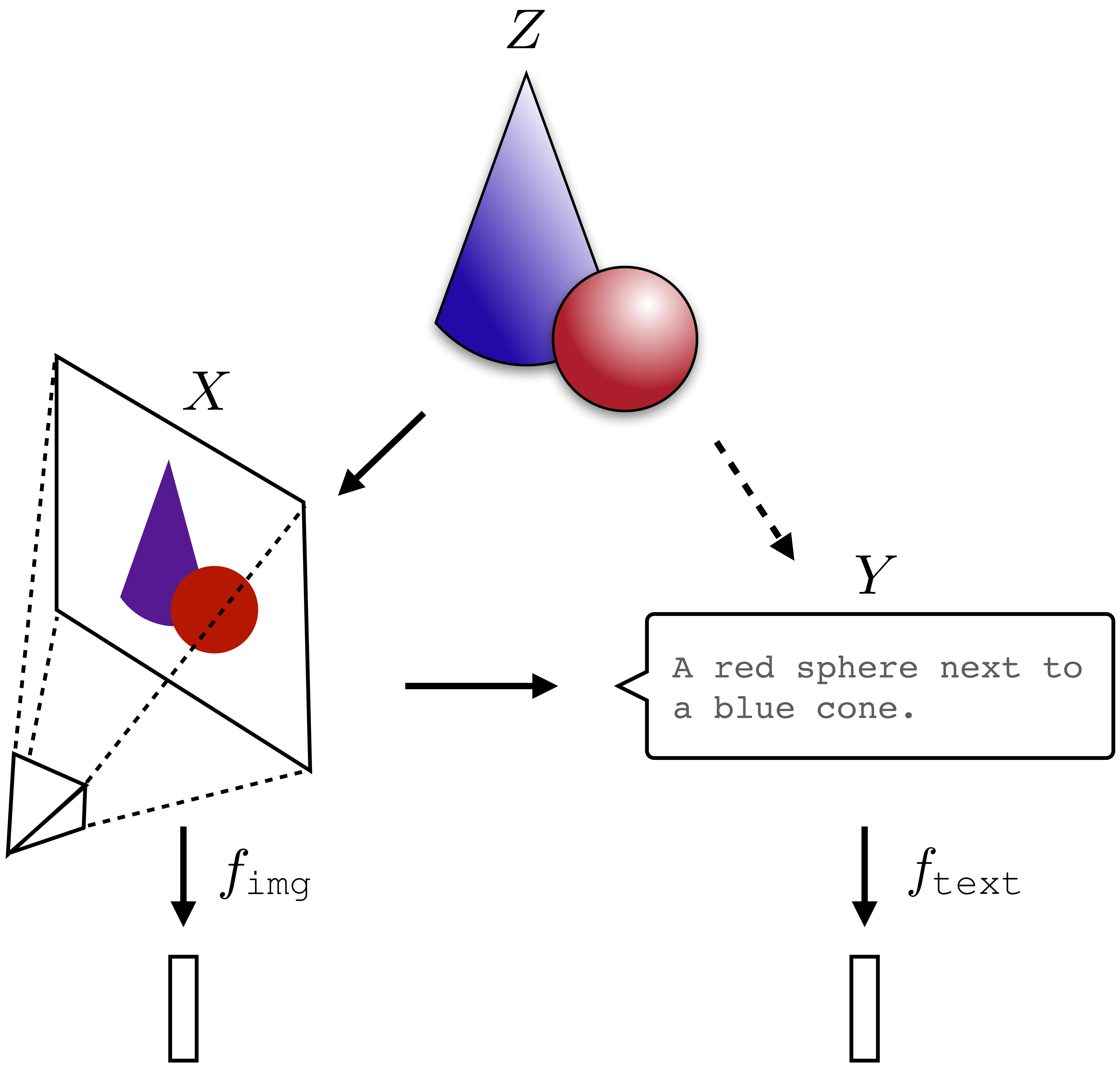
Maxima pars hominum cecis immersa tenebris
Volvitur assiduè, et studio letatur inani:
Adspice ut obrectis obtutuis inhereat umbris,
Vt VERI simulacra omnes mirentur amentes,

Et stolidi vanâ ludantur imagine rerum.
Quam pauci meliore luto, qui in lumine puro
Secreti â stolidâ turbâ, ludibria cernunt
Rerum umbras rectaque expendunt omnia lance:

Hi positâ erroris nebulâ dignoscere possunt
Vera bona, atque alios cecâ sub nocte latentes
Extrahere in claram lucem conantur, at illis
Nullus amor lucis, tanta est rationis egestas.

C.C. Harlemensis Inv.
Sanredam Sculpsit
Henr. Hondius excudit.
1604.

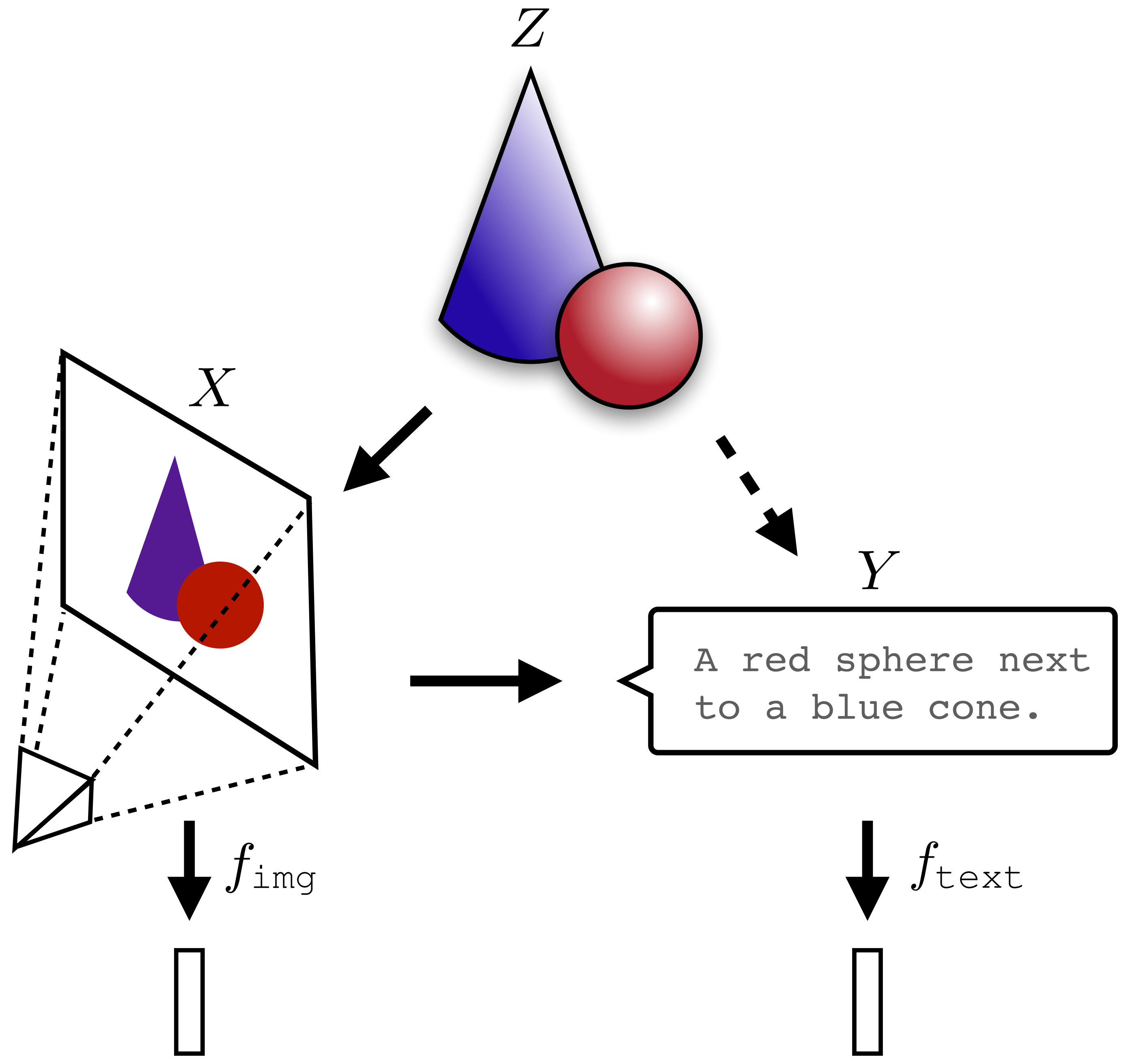
H.L. SPIEGEL FIGVRARI ET SCVLPI CVRAVIT. AC DOCTISS.° ORNATISS.°QZ D. PET. PAAW IN LVGDVN. ACAD. PROFESSORI MEDICO D.D.



Z
X
Y
A red sphere next to
a blue cone.
$f_{\mathtt{img}}$
$f_{\mathtt{text}}$

- World consists of a sequence of T discrete events $\mathbf{Z} \triangleq [z_1, \ldots, z_T]$
- Sampled from unknown $\mathbb{P}(\mathbf{Z})$
- All data is mediated via observation functions obs : $\mathcal{Z} \rightarrow \cdot$
- In this world, we will model **cooccurrences**:

$$P_{\text{coor}}(x_a, x_b) \propto \sum_{(t,t'):\,|t-t'|\leq T_{\text{window}}} \mathbb{P}(X_t = x_a, X_{t'} = x_b)$$



- Positives: two observations that cooccur; Negatives: two samples from marginals

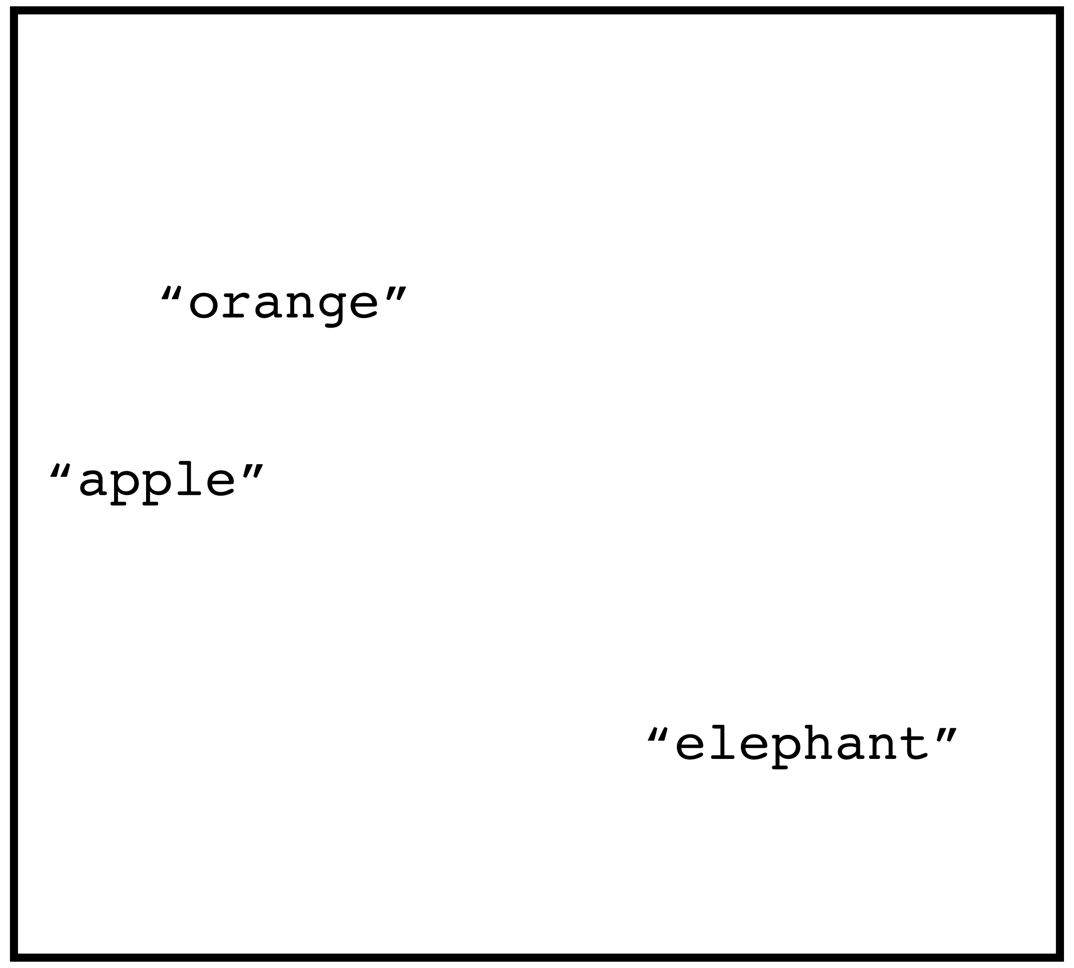
$\{$  ,  $\}$

`I parked the` **`car`** `in a nearby` **`street`**`.`

- Contrastive learner, with NCE objective converges to PMI:

$$\langle f_X(x_a), f_X(x_b)\rangle \approx \log \frac{P_{\text{coor}}(x_a, x_b)}{P_{\text{coor}}(x_a)P_{\text{coor}}(x_b)}$$

- An embedding in which similarly = (normalized) cooccurrence rate.



- For bijective, discrete `obs` functions, PMI over `obs` equals PMI over events, which implies that different `obs` converge to same kernel.

Summary #4:

Scaling up task/data/model can drive convergence.

Certain contrastive learners converge to kernel = rate at which events co-occur in nature.

Outline:

1. What’s a representation?

2. How to measure representational similarity?

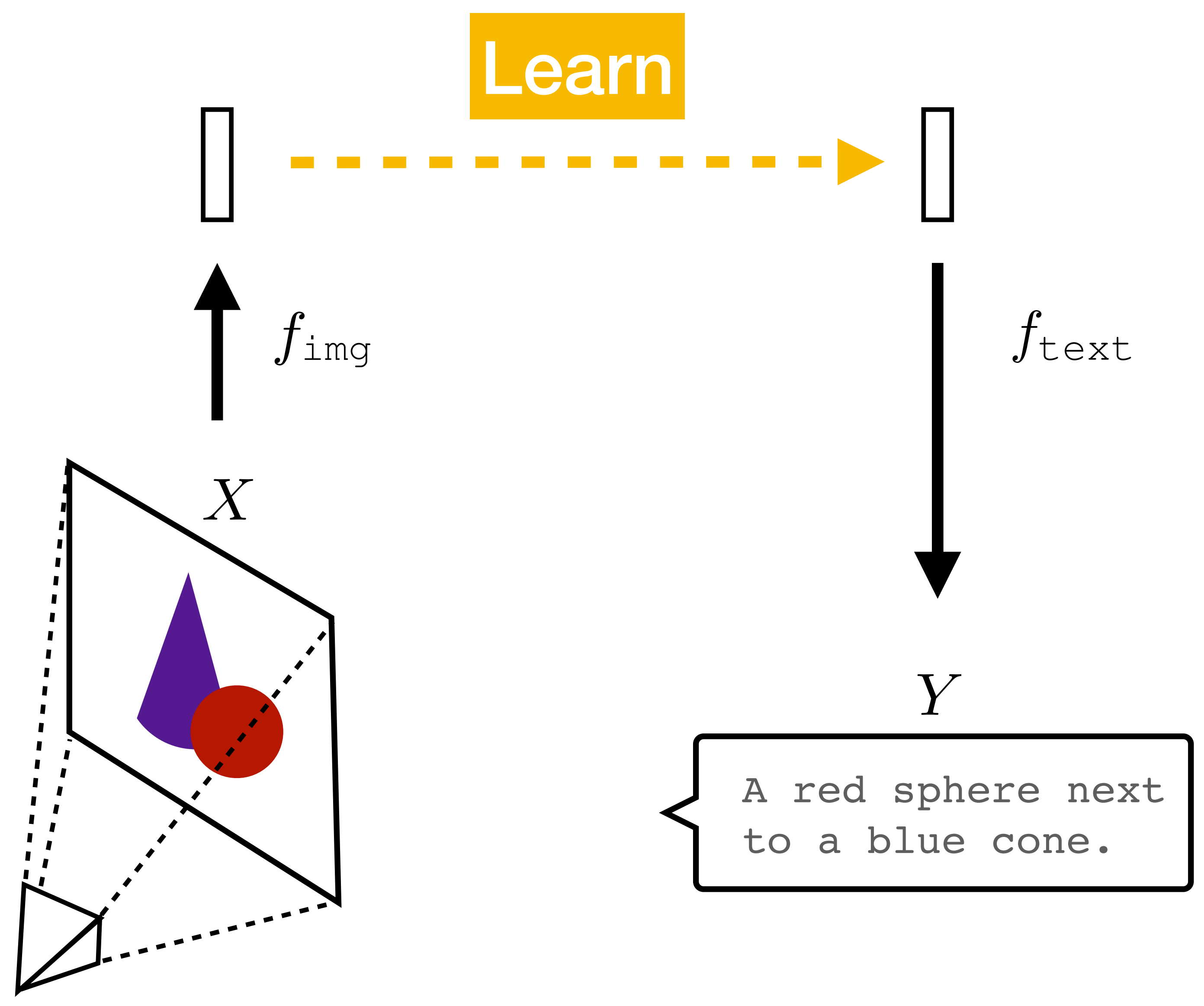
3. Which representations are similar and which are different?

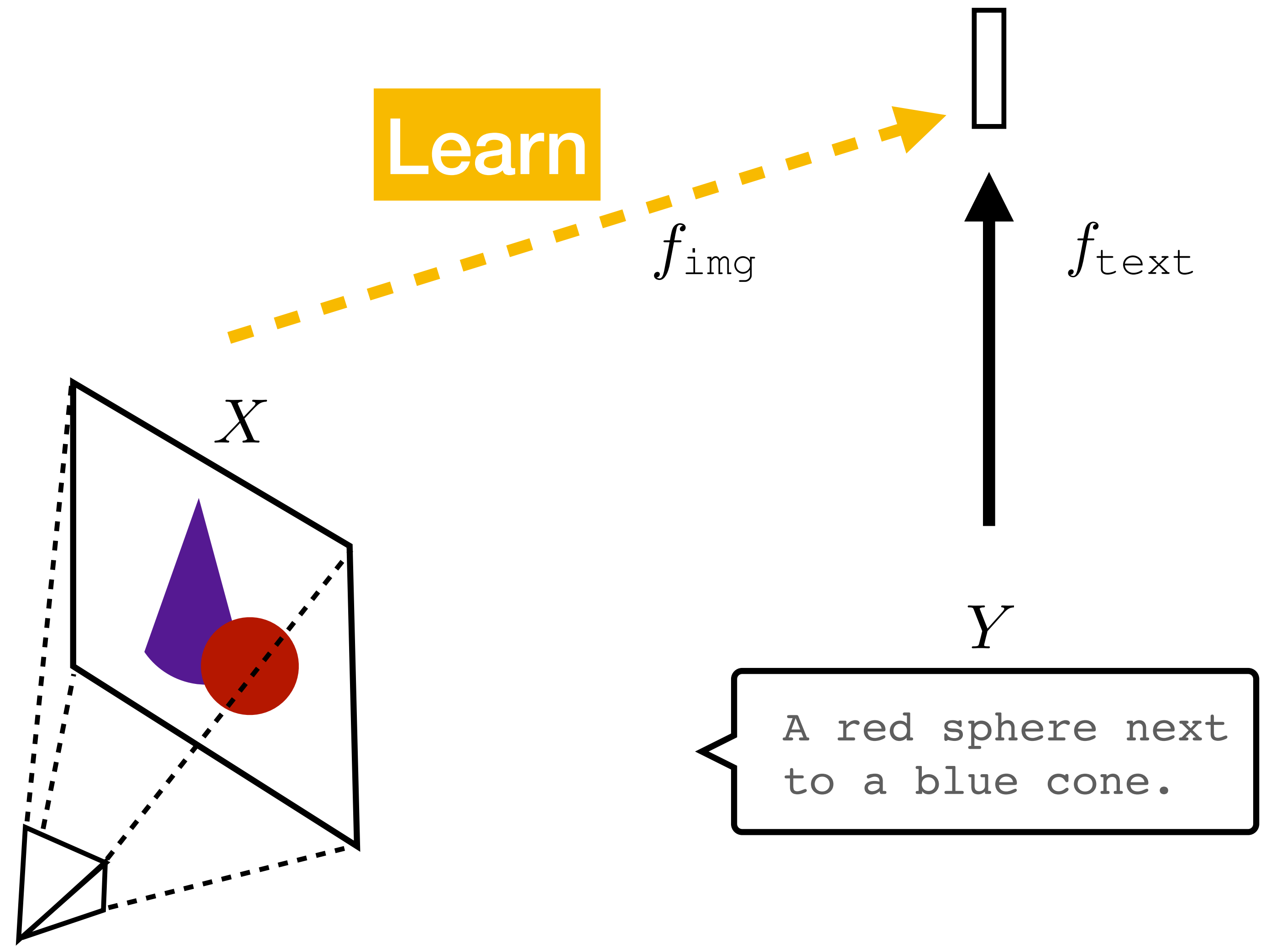
4. What drives representational alignment?

5. Making representations more aligned

Benefits of alignment

- Can share data/supervision betwen modalities
- A common representation can serve as a bridge for translation
- Can scaffold new models onto existing representations

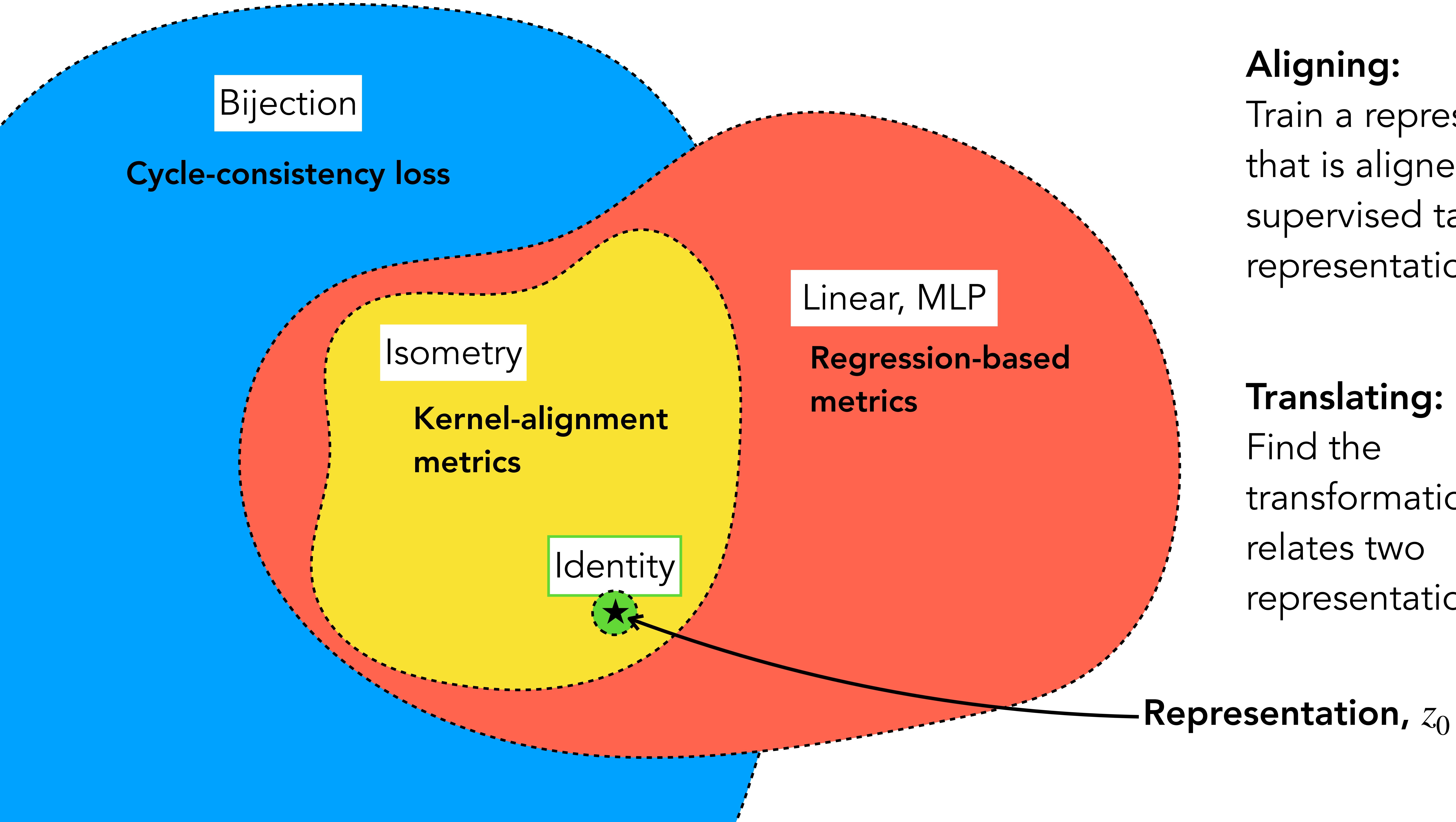




Detriments of alignment

- Lack of diversity in the population of models.
- Sometimes one modality has access to qualitatively different information than another, and this information can be useful; alignment will remove this information.
- There might not be a single best representation for all problems. (And in theory there isn't.)

Aligning and translating representations

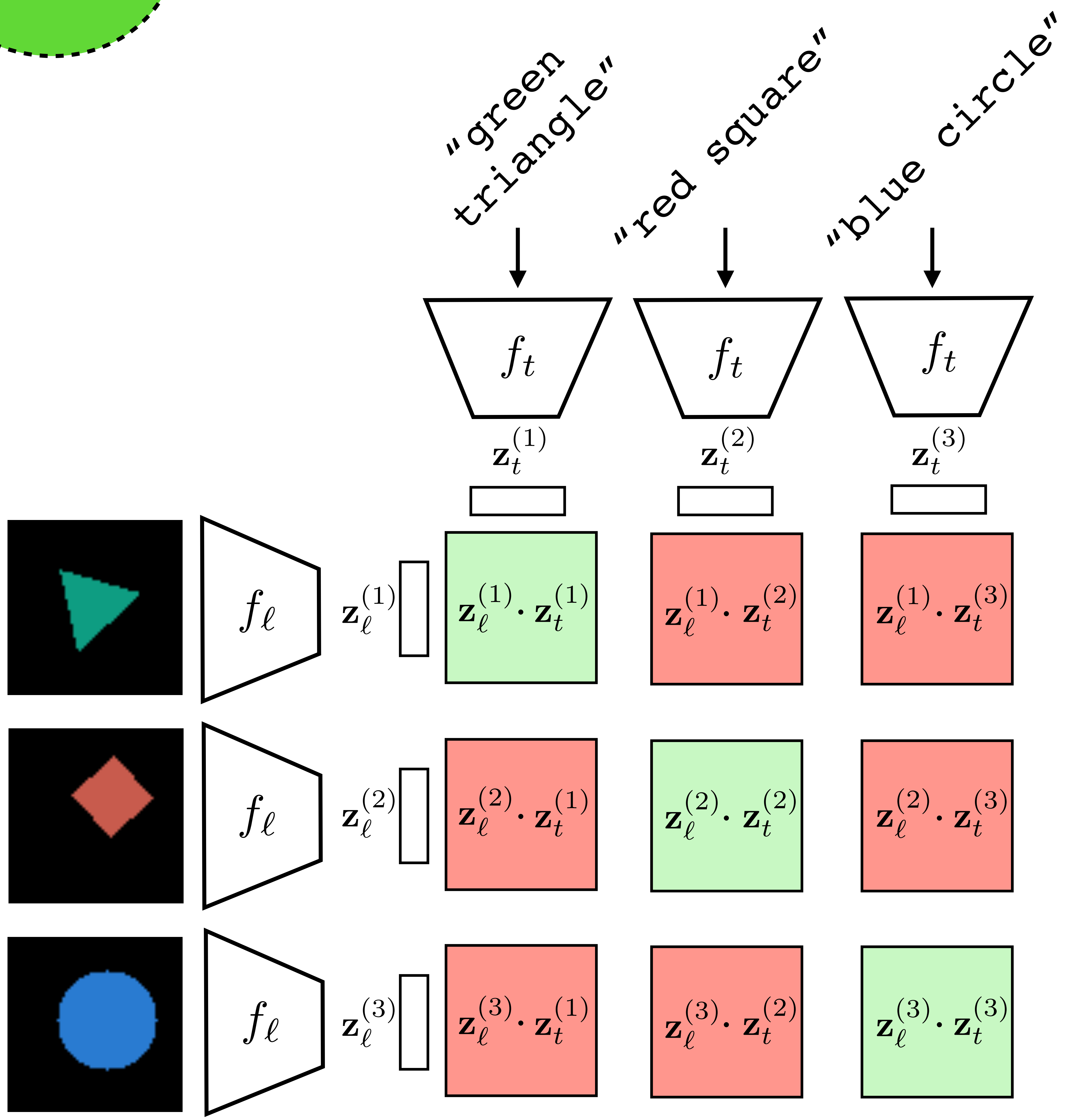


Aligning:
Train a representation that is aligned with a supervised target representation, up to T

Translating:
Find the transformation T that relates two representations

Identity

Training to align representations up to identity transformation



Contrastive Language-Image Pre-training (CLIP)

- Tries to find a representation in which an image and its caption are assigned identical embeddings.

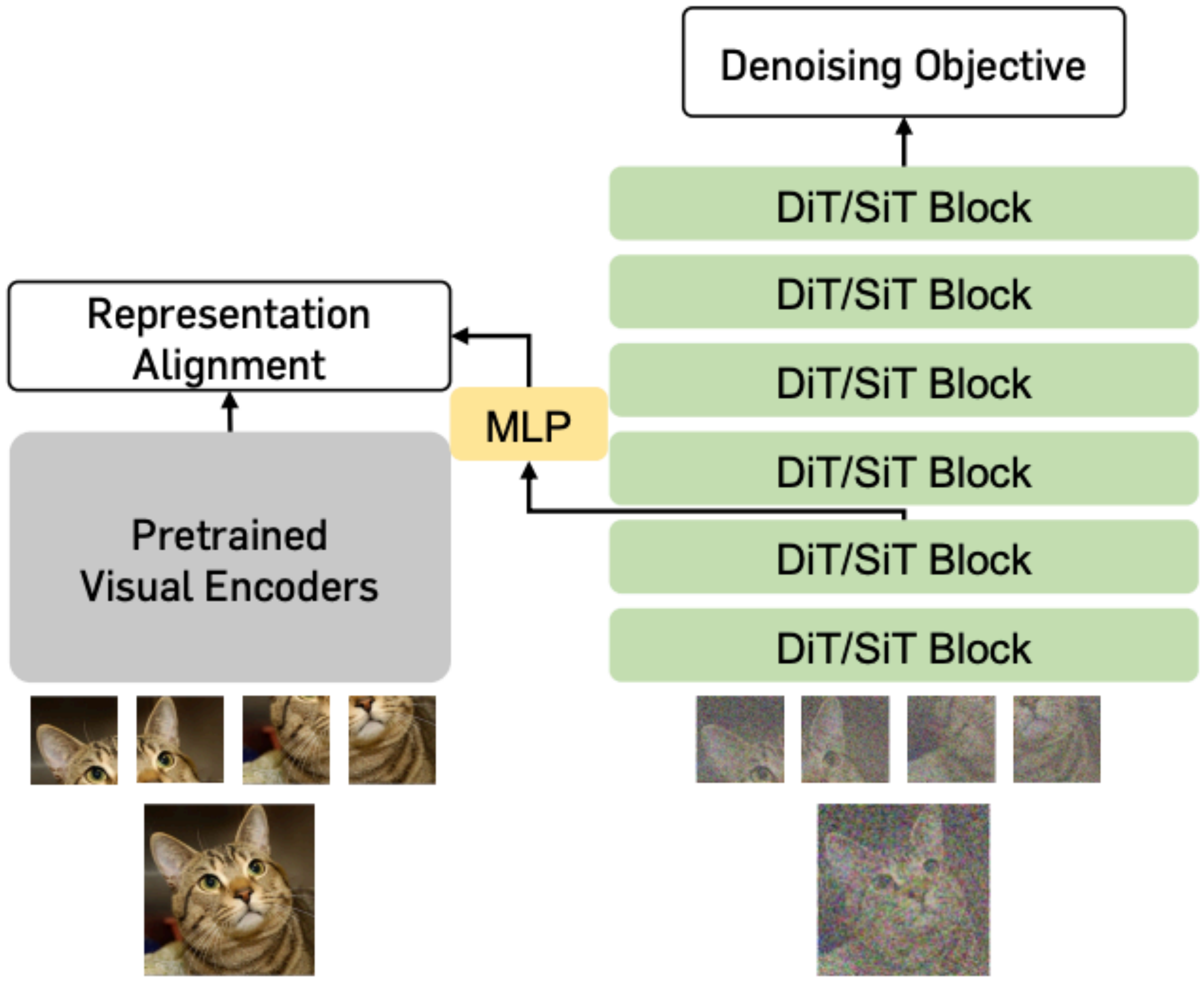
[Radford*, Kim* et al., ICML 2021]

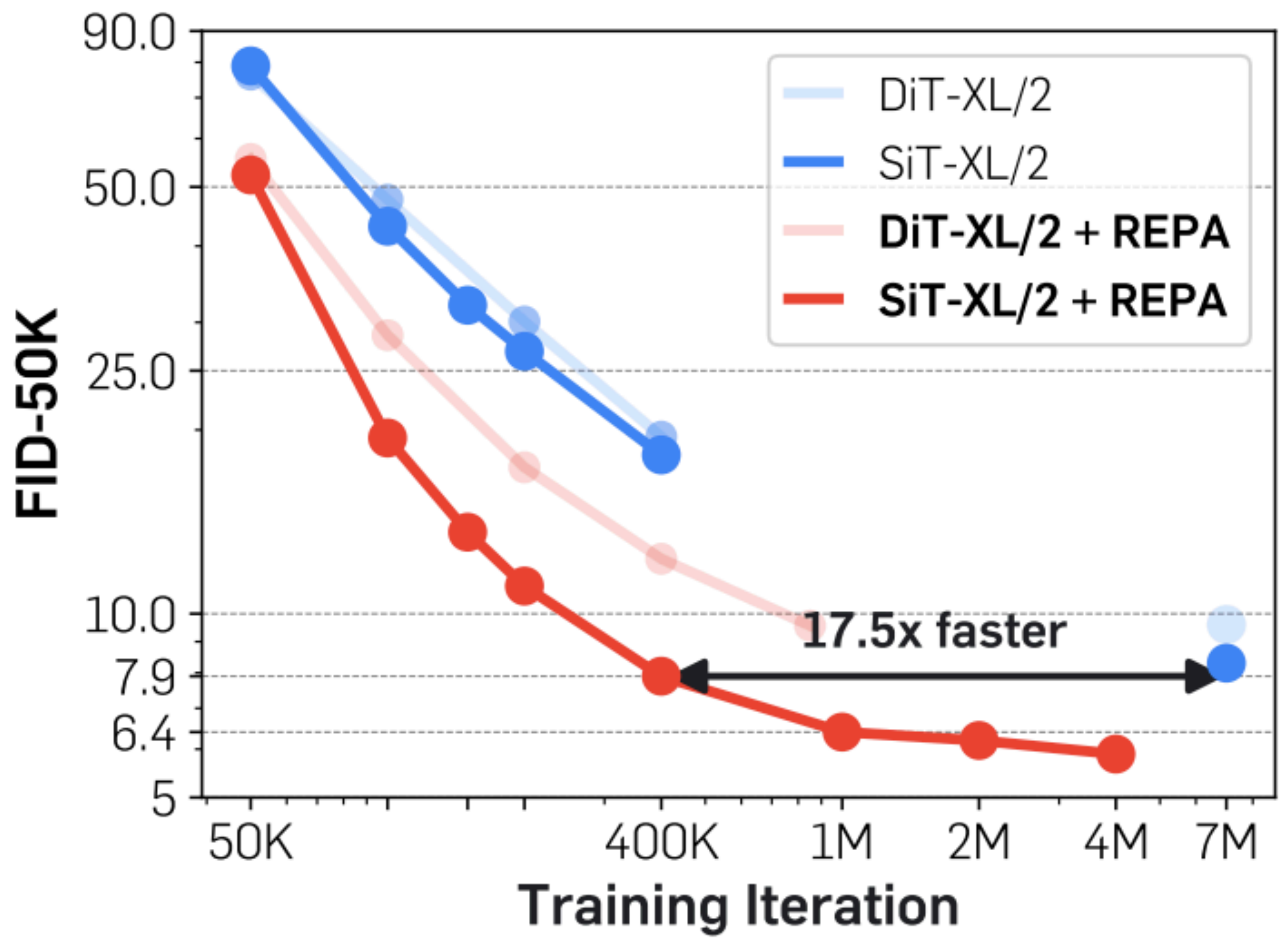
MLP

Training to align representation up to MLP transformation

Representation Alignment for Generation (REPA)

[Yu, Kwak, Jang, Jeong, Huang, Shin*, Xie*, ICLR 2025]

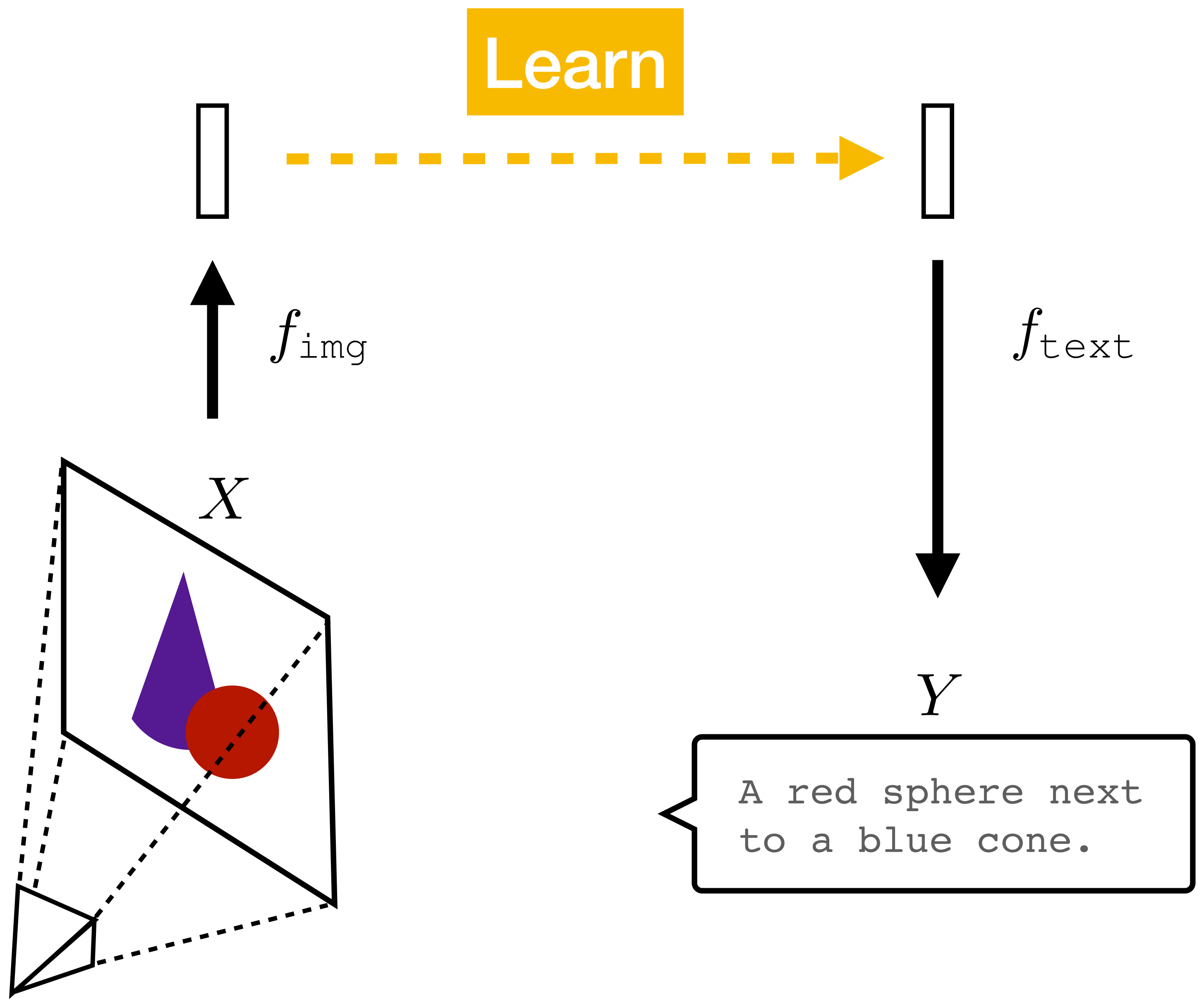




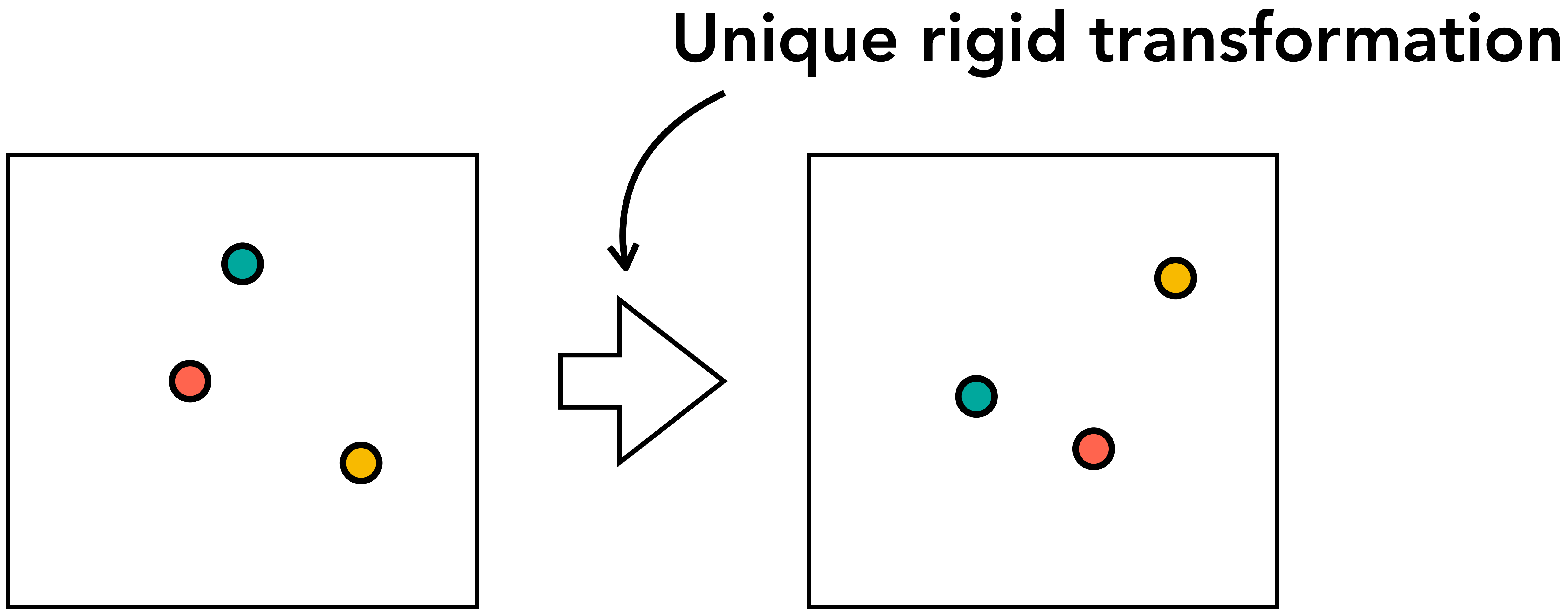
Isometry

Translating between representations

- Find an isometric transformation that relates $\mathbf{Z}_1$ to $\mathbf{Z}_2$.



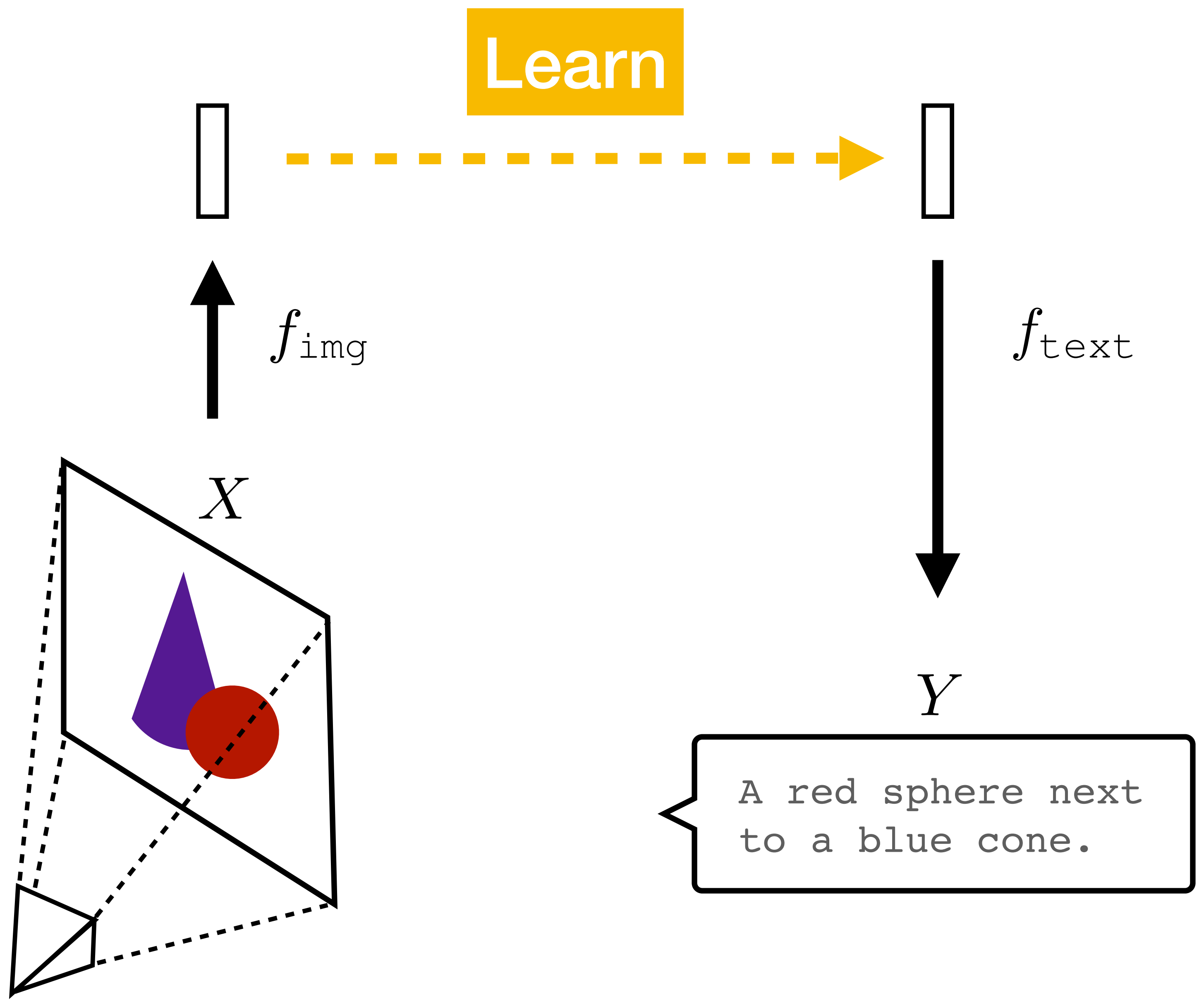
- With sufficiently non-degenerate data, You can translate between two representations related by an isometry with **zero** paired examples.

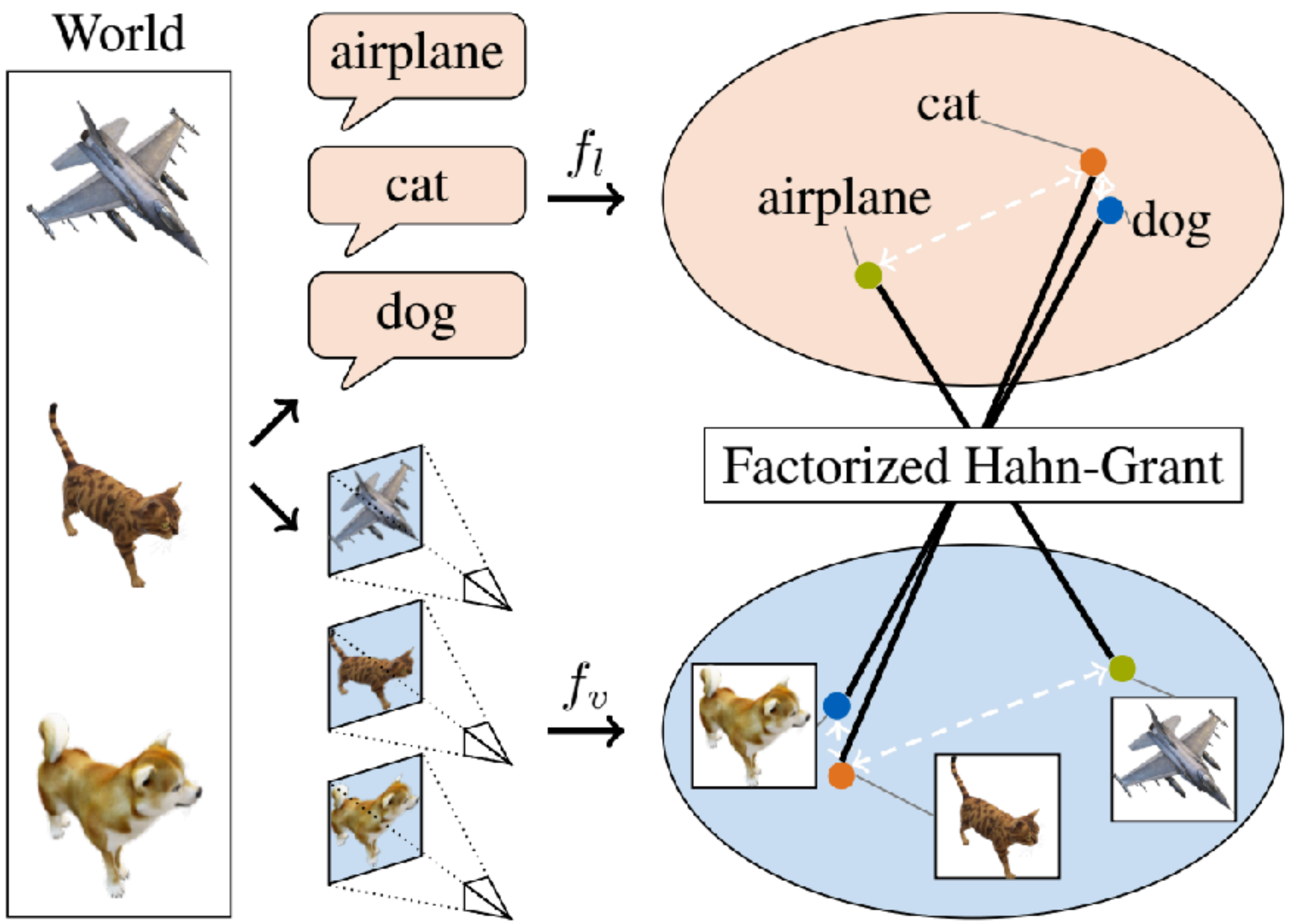


Isometry

Translating between representations

- Find an isometric transformation that relates $\mathbf{Z}_1$ to $\mathbf{Z}_2$.





[Schnaus, Araslanov, Cremers, arXiv 2025]

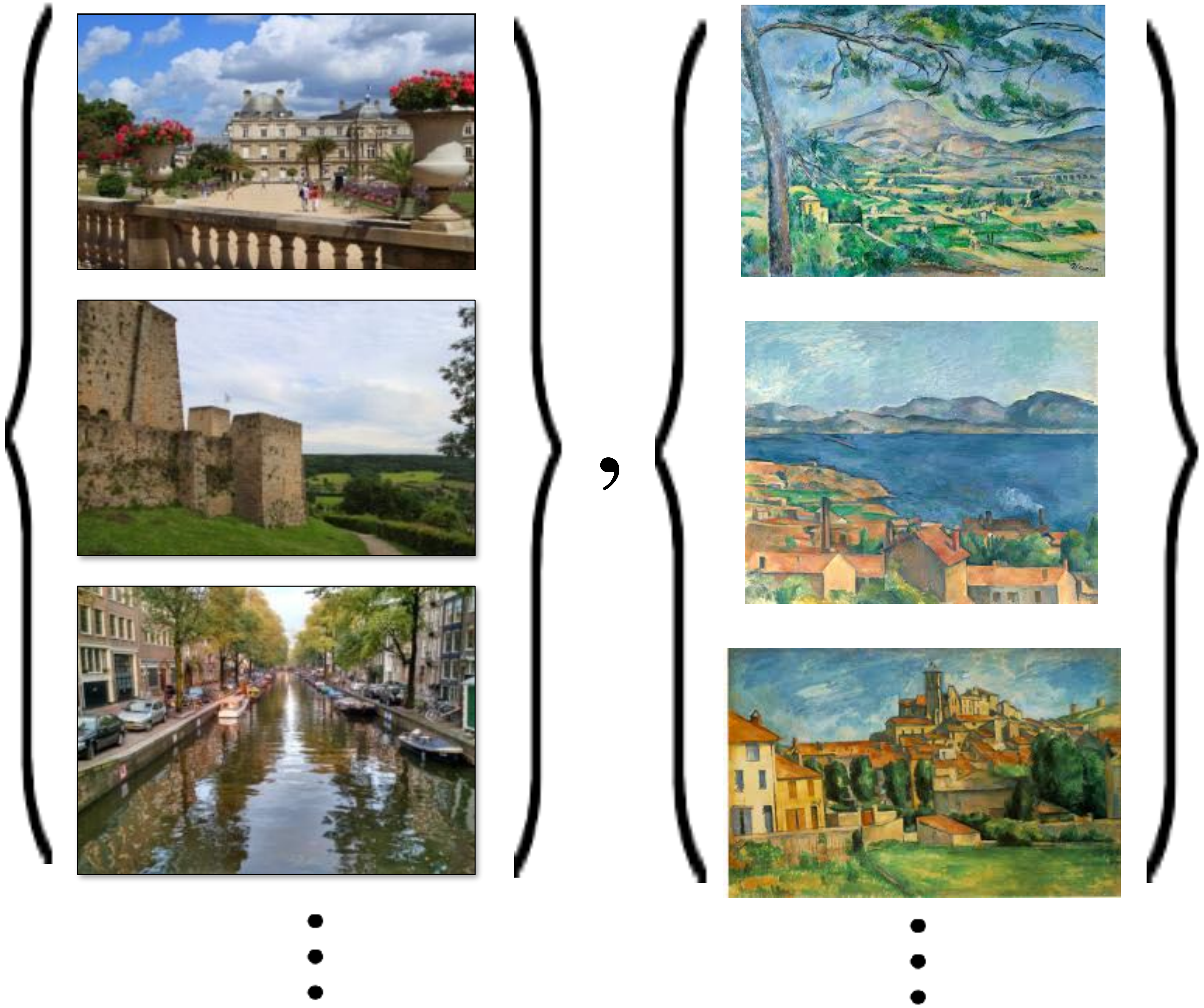
see also: Sorscher, Ganguli, Sompolinsky, PNAS 2022; Lazaridou, Bruni, Baroni, ACL 2014

Bijection

CycleGAN

[Zhu*, Park*, Isola, Efros, ICCV 2017]

$\mathbf{z}_1$, $\mathbf{z}_2$



[Zhu*, Park* et al. 2017], [Yi et al. 2017], [Kim et al. 2017]

Bijection

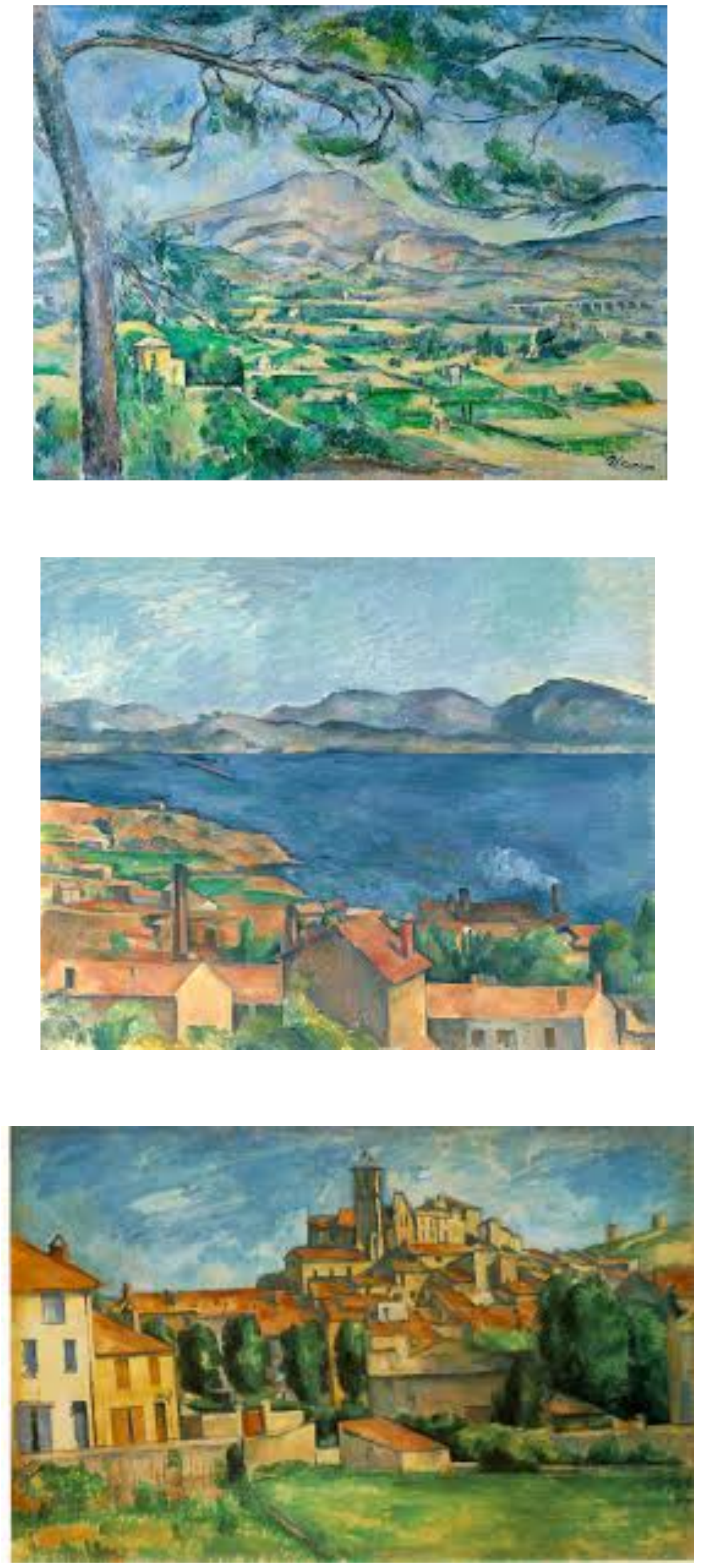
CycleGAN

[Zhu*, Park*, Isola, Efros, ICCV 2017]

$\mathbf{Z}_1$

$$\mathbf{Z}_1 \rightleftarrows \mathbf{Z}_2$$

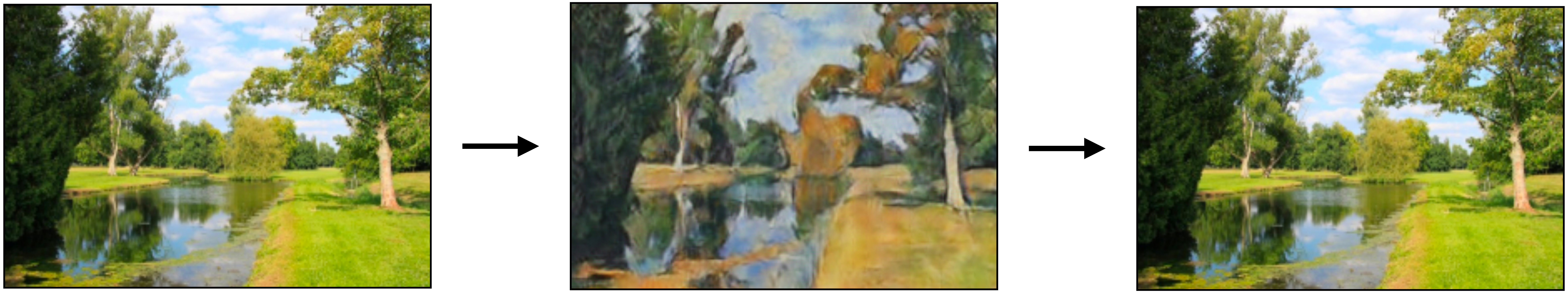
$\mathbf{Z}_2$

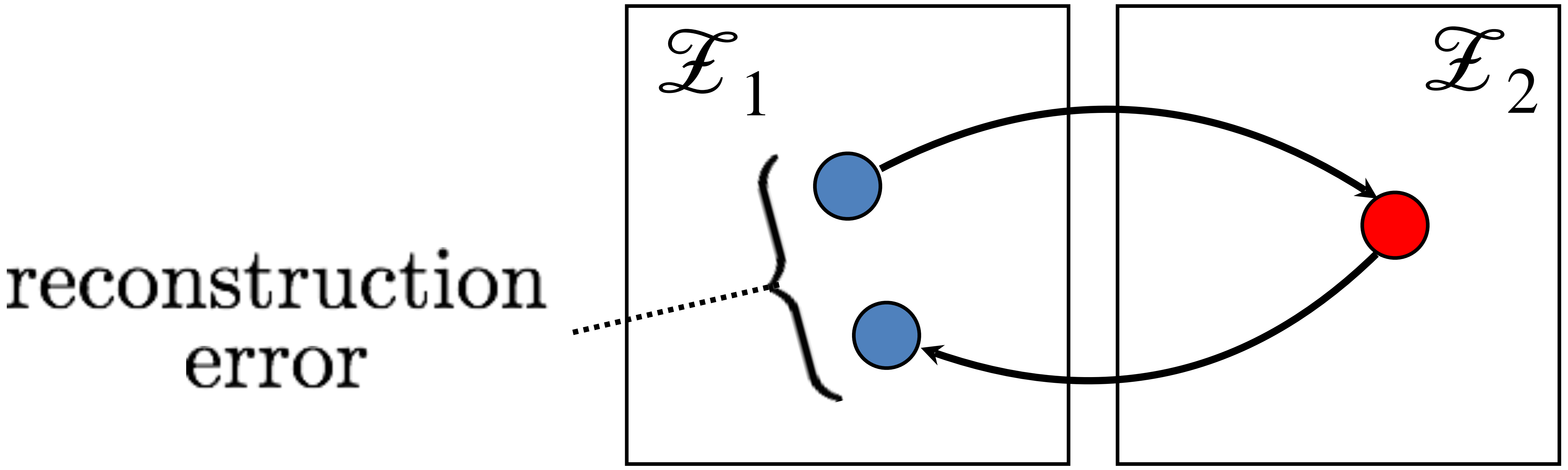


[Zhu*, Park* et al. 2017], [Yi et al. 2017], [Kim et al. 2017]

CycleGAN

[Zhu*, Park*, Isola, Efros, ICCV 2017]



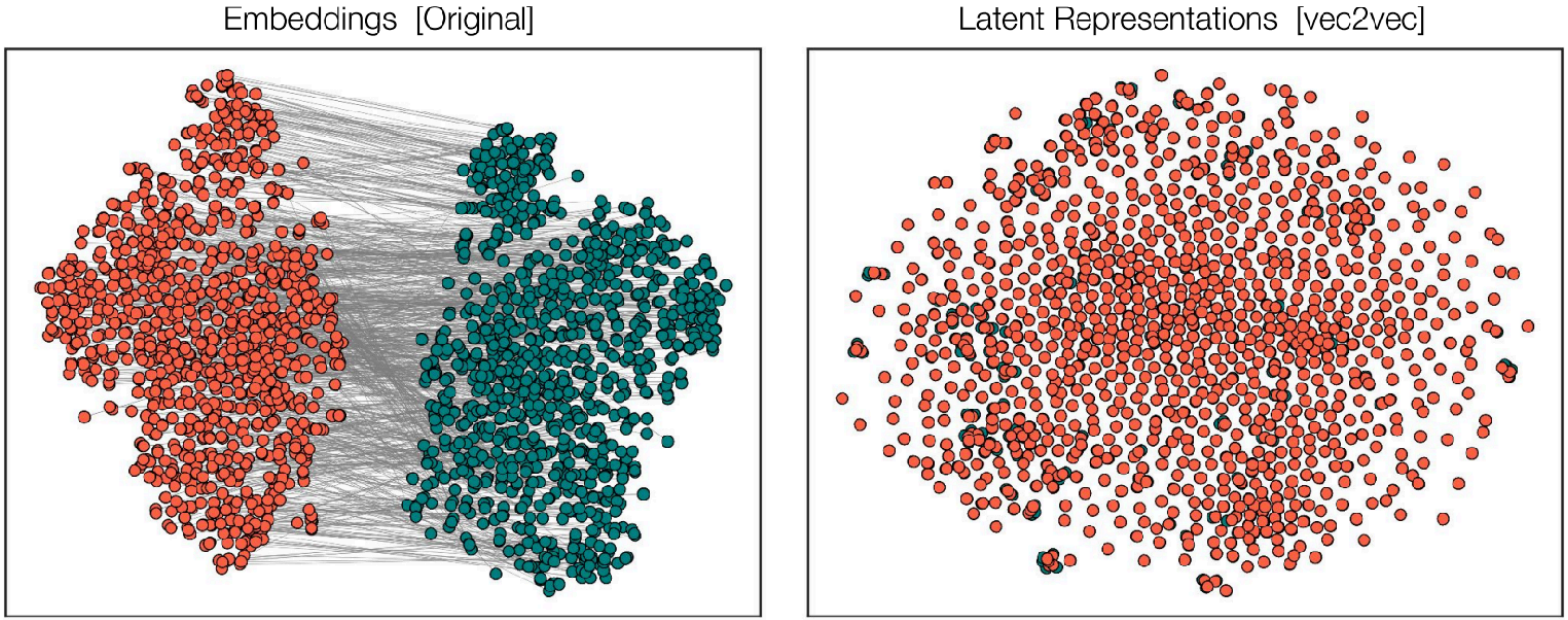


Bijection

Isometry

vec2vec

[Jha, Zhang, Shmatikov, Morris, arXiv 2025]



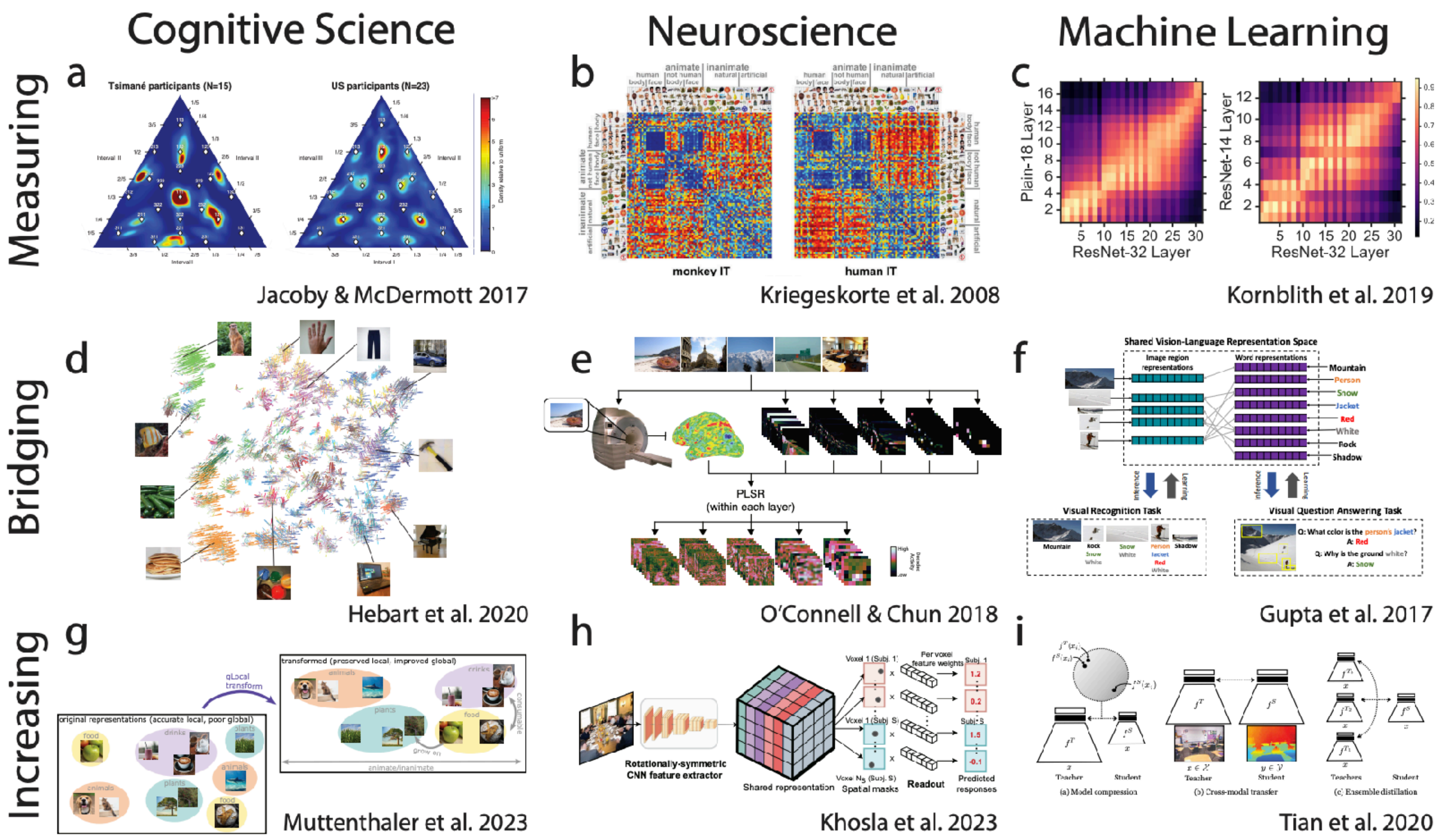
Method: GAN + cycle consistency loss + kernel matching loss

See also: [Conneau, Lample, Ranzato, Denoyer, Jégou, ICLR 2018]

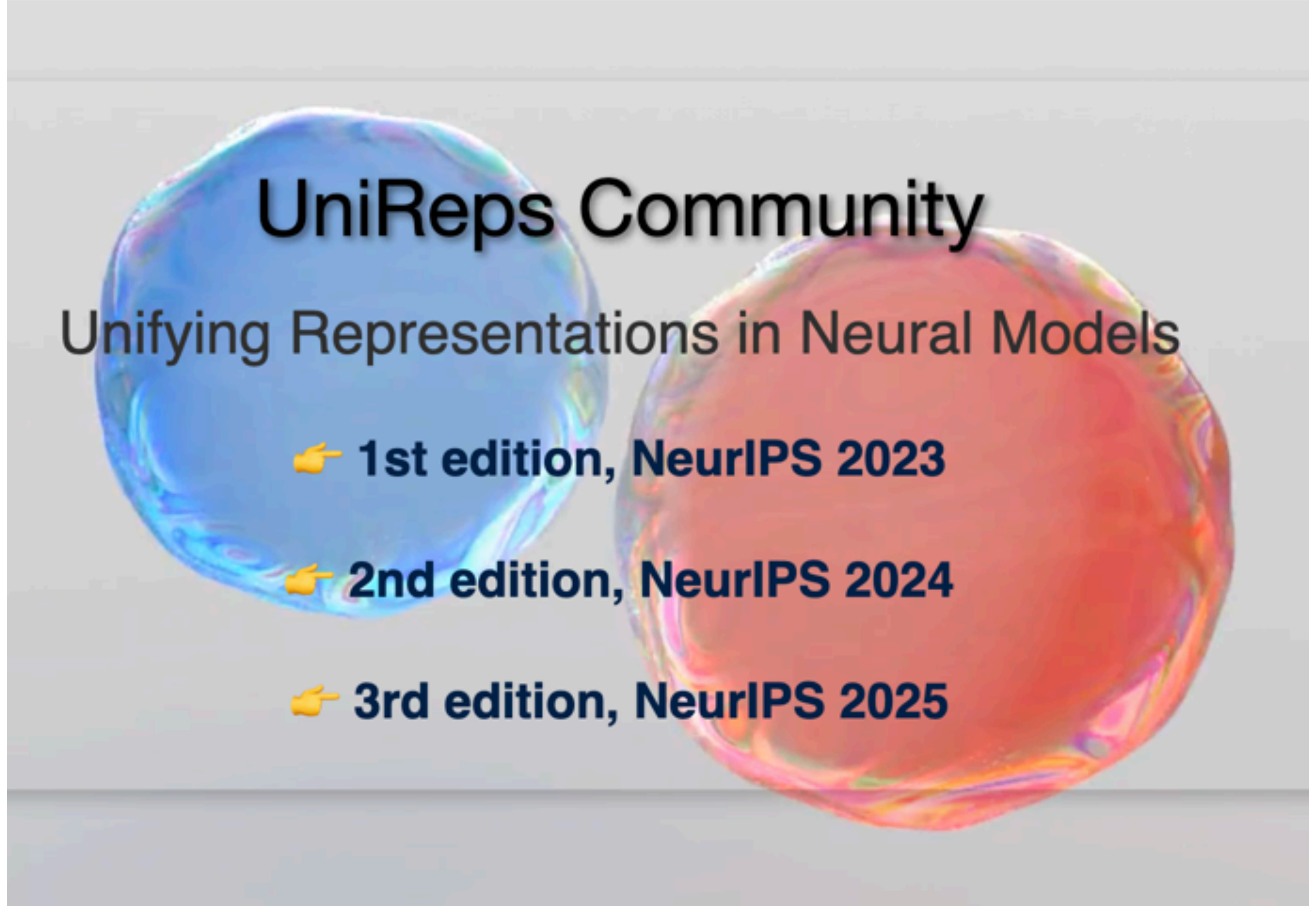
Summary #5:

Many important problems involve aligning or translating between representations.

You don't necessarily need paired data to do so.



[See more: "Getting Aligned on Representational Alignment," Sucholutsky*, Muttenthaler*, et al. arXiv 2024]



CCN

Home Account Attend Events Presentations Subm

Community Event

Wednesday, August 13, 10:00 am – 12:00 pm, Room C1.03

Universality and Idiosyncrasy of Perceptual Representations

Thanks!