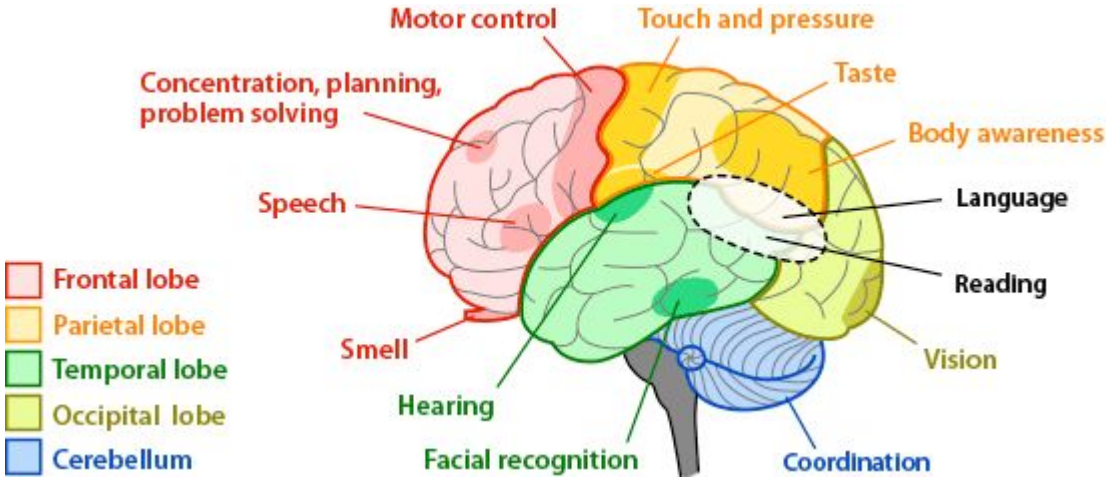
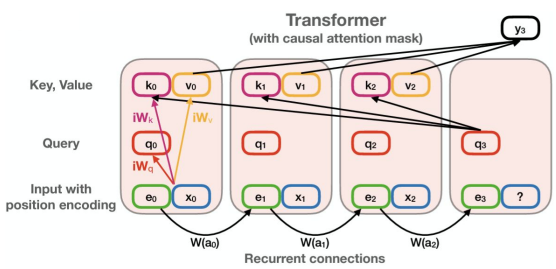
Bridging Theories:

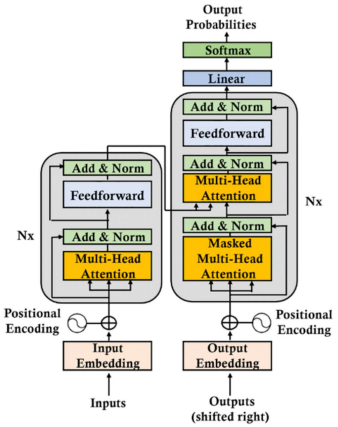
What Task-Optimized Models Reveal About Brain Computations

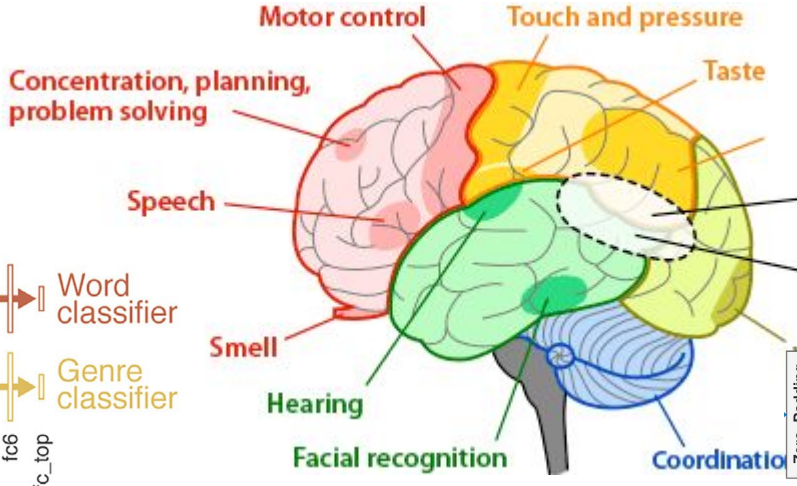
Tahereh Toosi

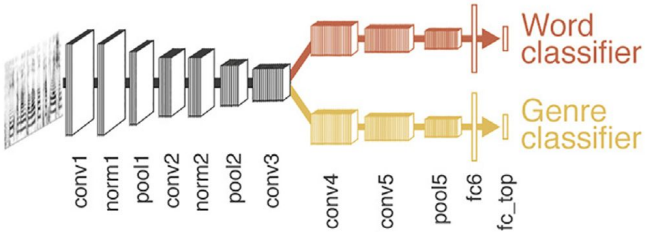
Advanced Topics in Theoretical Neuroscience
March 2025

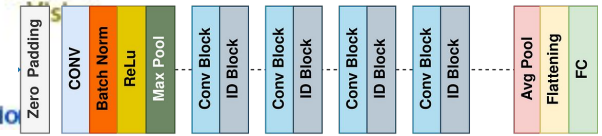

Motor control
Touch and pressure
Concentration, planning, problem solving
Taste
Body awareness
Speech
Language
Reading
Smell
Vision
Hearing
Facial recognition
Coordination
Frontal lobe
Parietal lobe
Temporal lobe
Occipital lobe
Cerebellum


Transformer
(with causal attention mask)
Key, Value
Query
Input with position encoding
Recurrent connections


Output Probabilities
Softmax
Linear
Add & Norm
Feedforward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

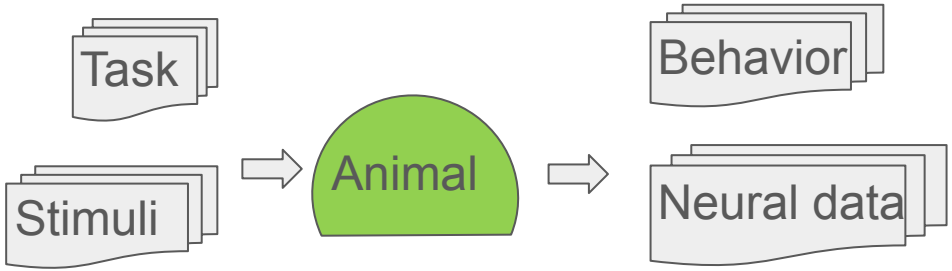

Motor control
Touch and pressure
Taste
Concentration, planning, problem solving
Speech
Language
Reading
Smell
Hearing
Facial recognition
Coordination


Word classifier
Genre classifier
conv1
norm1
pool1
conv2
norm2
pool2
conv3
conv4
conv5
pool5
fc6
fc_top

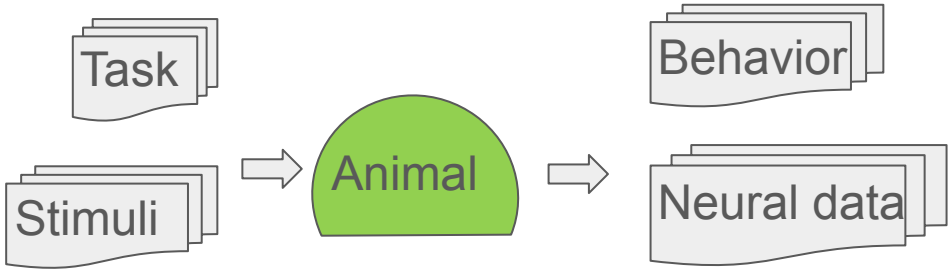

Zero Padding
CONV
Batch Norm
ReLu
Max Pool
Conv Block
ID Block
Avg Pool
Flattening
FC

There is more to the task-optimized modeling framework than to replace every brain region with a network!

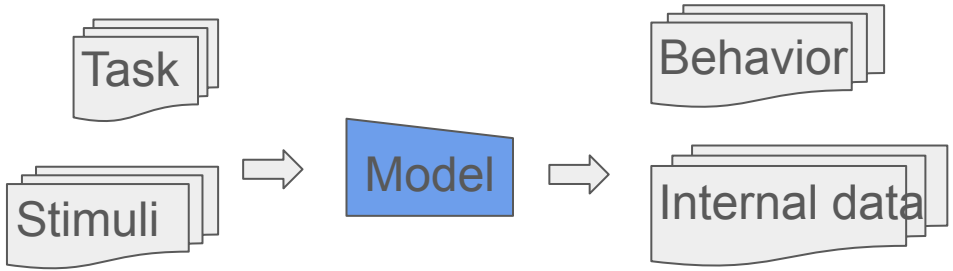
Systems Neuroscience



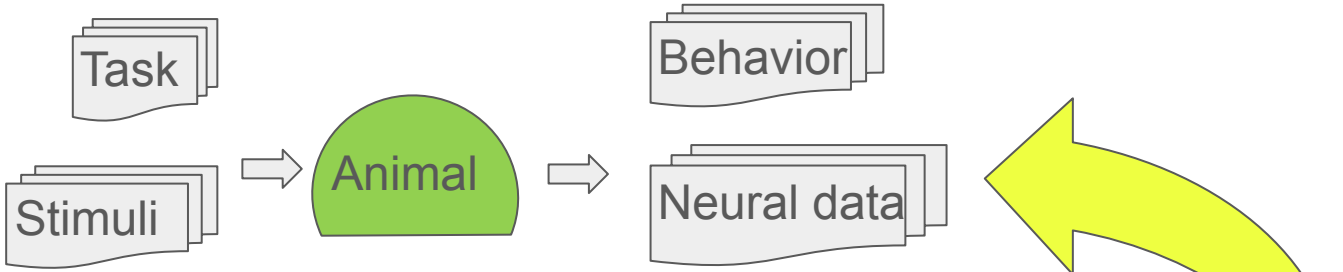
Systems Neuroscience



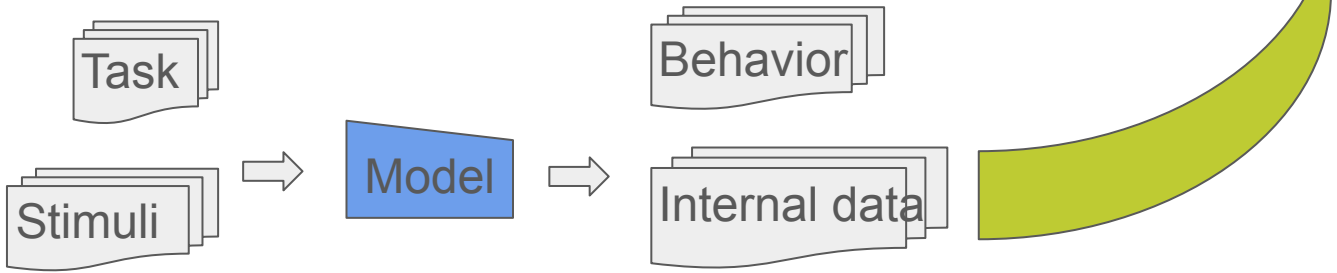
Task-optimized models



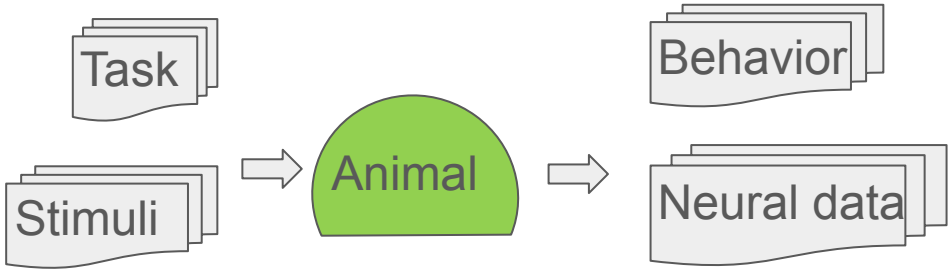
Systems Neuroscience



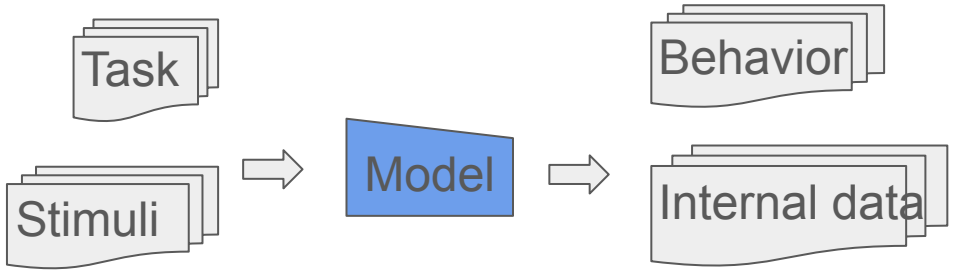
Task-optimized models



Systems Neuroscience



Task-optimized models



Mapping
Prediction
Representational similarity

Roadmap

- Comparing models to data
  - Compare model 1 to model 2: which one maps to data better? (**Case study 1**)
  - However, the platonic representation hypothesis!
- Closing the loop by synthesizing stimuli
  - Once a model predicts neural data well, it can be inverted, to generate new stimuli (**Case study 2**)
  - Still a very new field
- Computational tricks
  - Contrastive learning
  - Regularizations (**Case study 3**)
- Bridging theories (different perspectives, same math!)
  - Recognition | Generation
  - Learning | Attention

Roadmap

- Comparing models to data
  - Compare model 1 to model 2: which one maps to data better? (**Case study 1**)
  - However, the platonic representation hypothesis!
- Closing the loop by synthesizing stimuli
  - Once a model predicts well, it can be inverted, to give stimuli (**Case study 2**)
  - However, Mechanistic interpretability hasn't really been useful in deep learning
- Computational tricks
  - Contrastive learning
  - Regularizations (**Case study 3**)
- Bridging theories (different perspectives, same math!)
  - Attention mechanism to Memory
  - Regularization to Generativity (**Case study 4**)

Case study 1: the beginning

Performance-optimized hierarchical models predict neural responses in higher visual cortex

Daniel L. K. Yamins[a,1], Ha Hong[a,b,1], Charles F. Cadieu[a], Ethan A. Solomon[a], Darren Seibert[a], and James J. DiCarlo[a,2]

[a]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Harvard-MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved April 8, 2014 (received for review March 3, 2014)

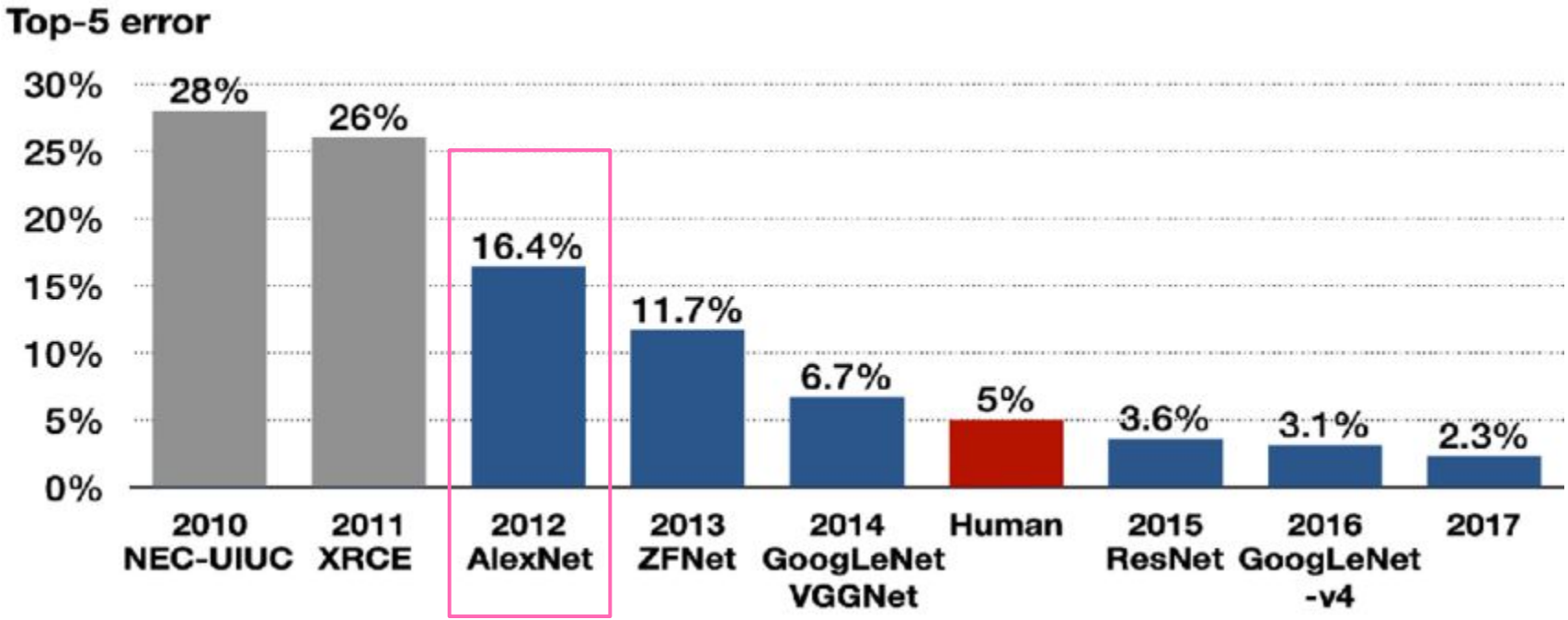
The ventral visual stream underlies key human visual object recognition abilities. However, neural encoding in the higher areas of the ventral stream remains poorly understood. Here, we describe

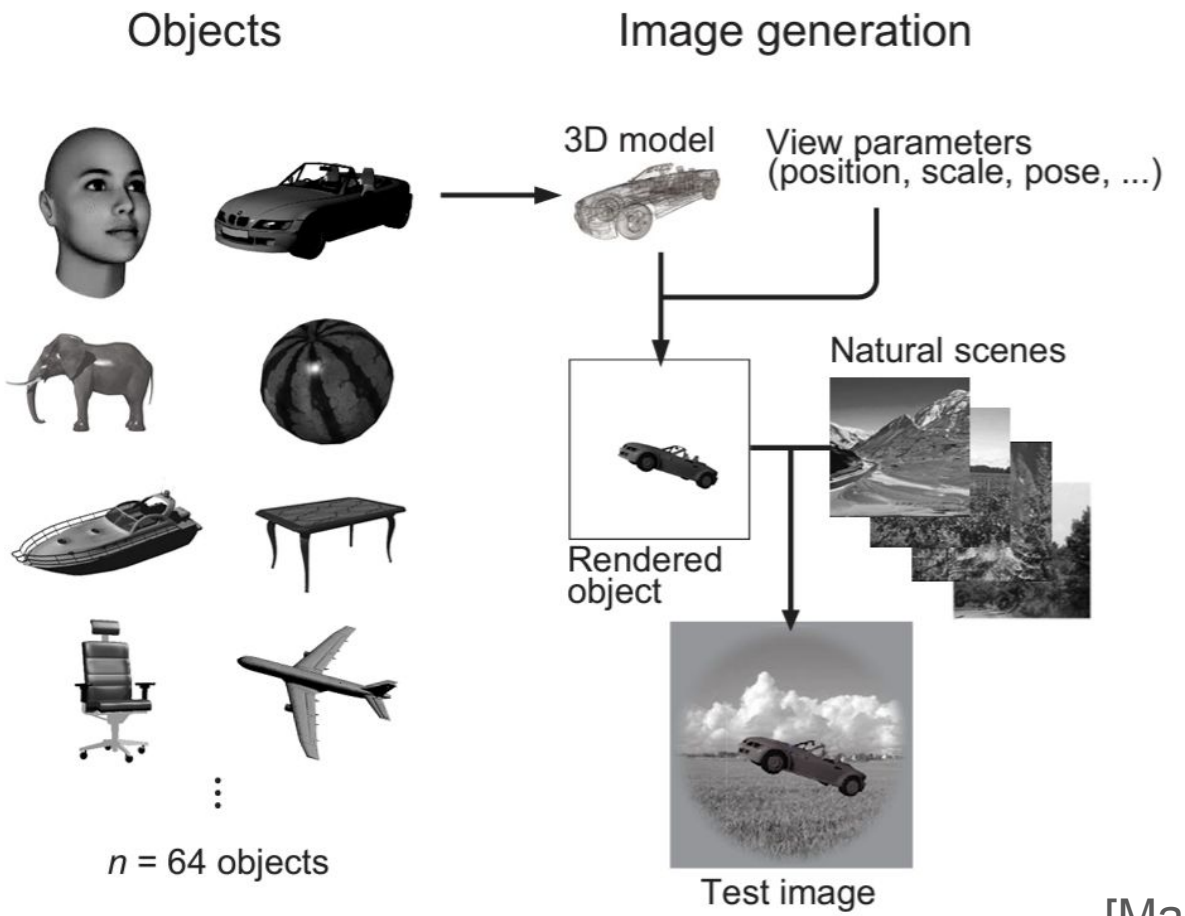
Explaining the neural encoding in these higher ventral areas thus remains a fundamental open question in systems neuroscience.

As with V1, models of higher ventral areas should be neurally

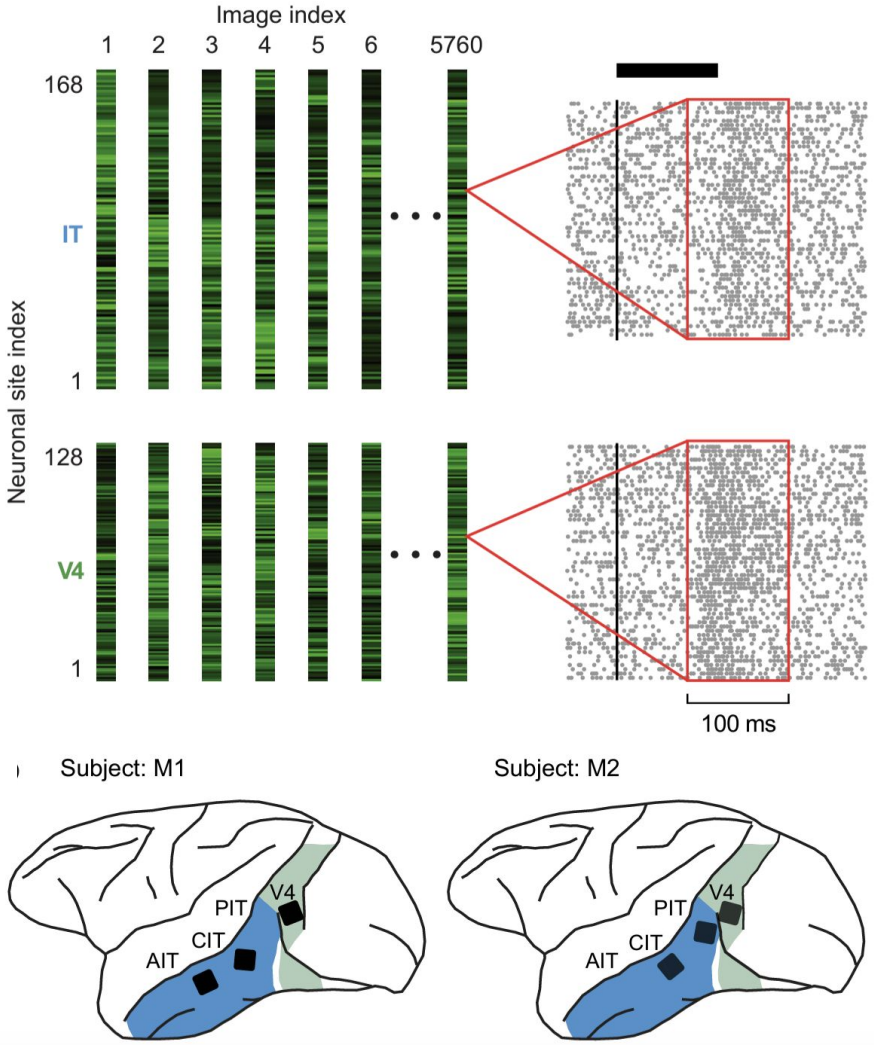
2012 Imagenet Challenge

2014 Neuroscience papers

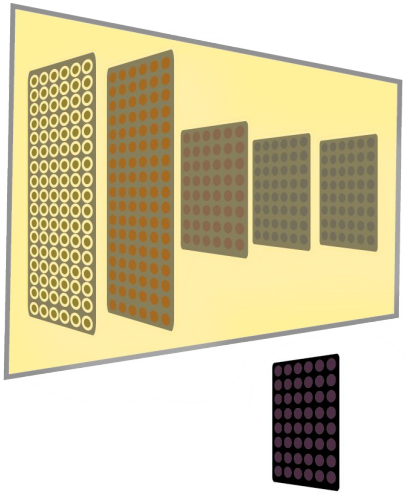


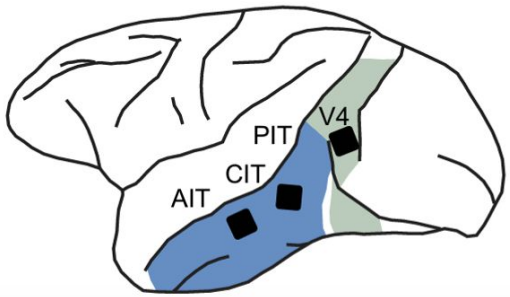


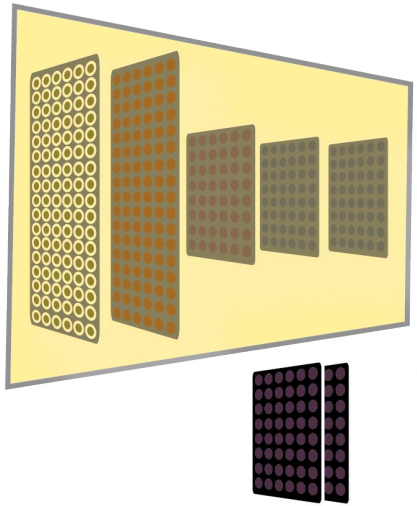
[Majaj et al 2015]

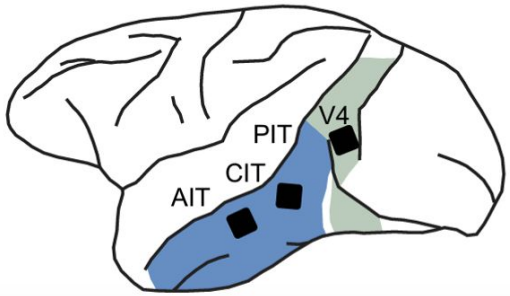


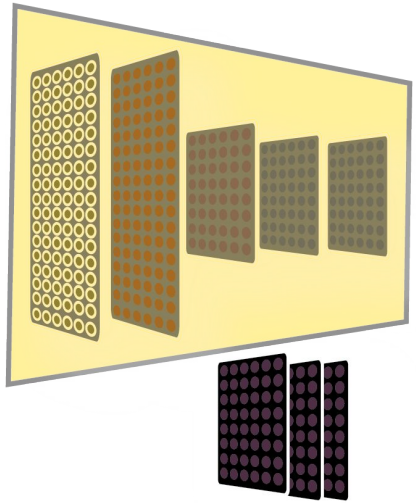
[Majaj et al 2015]

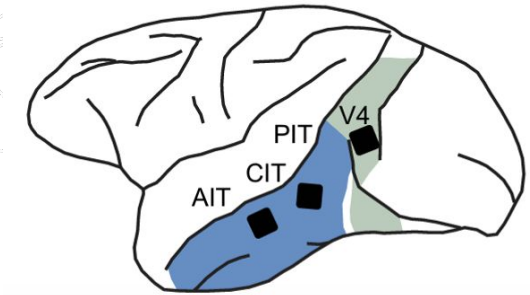


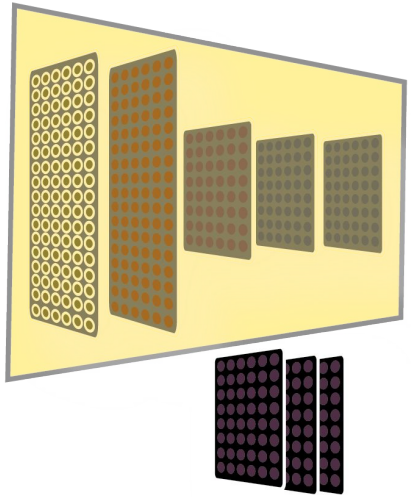

V4
PIT
CIT
AIT

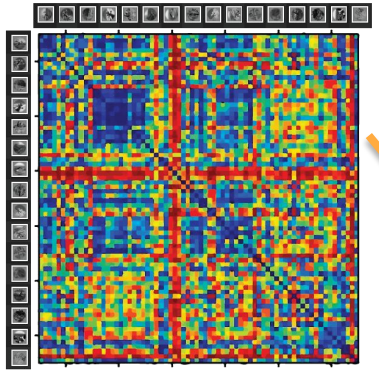



V4
PIT
CIT
AIT

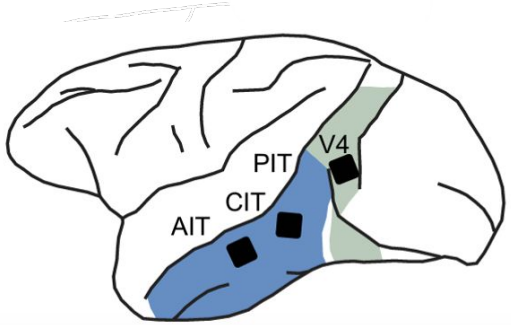


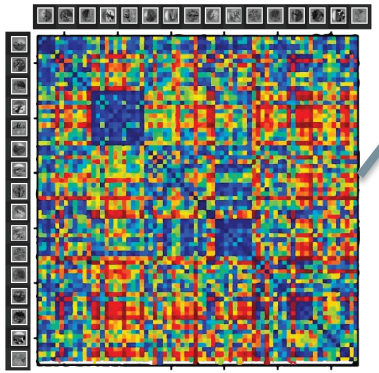

V4
PIT
CIT
AIT





Representational
Similarity

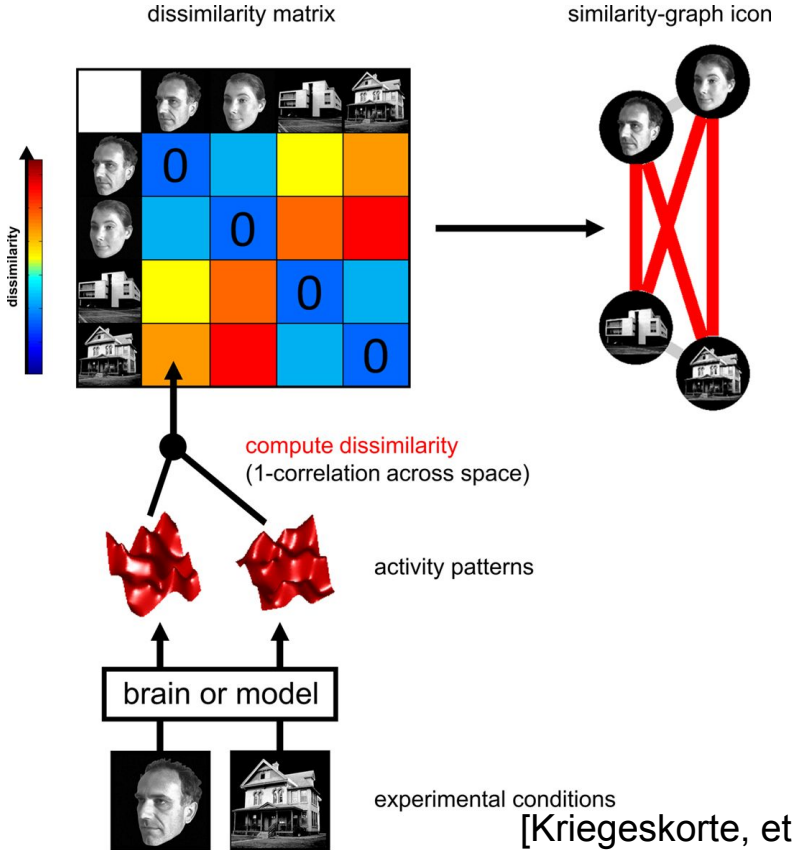




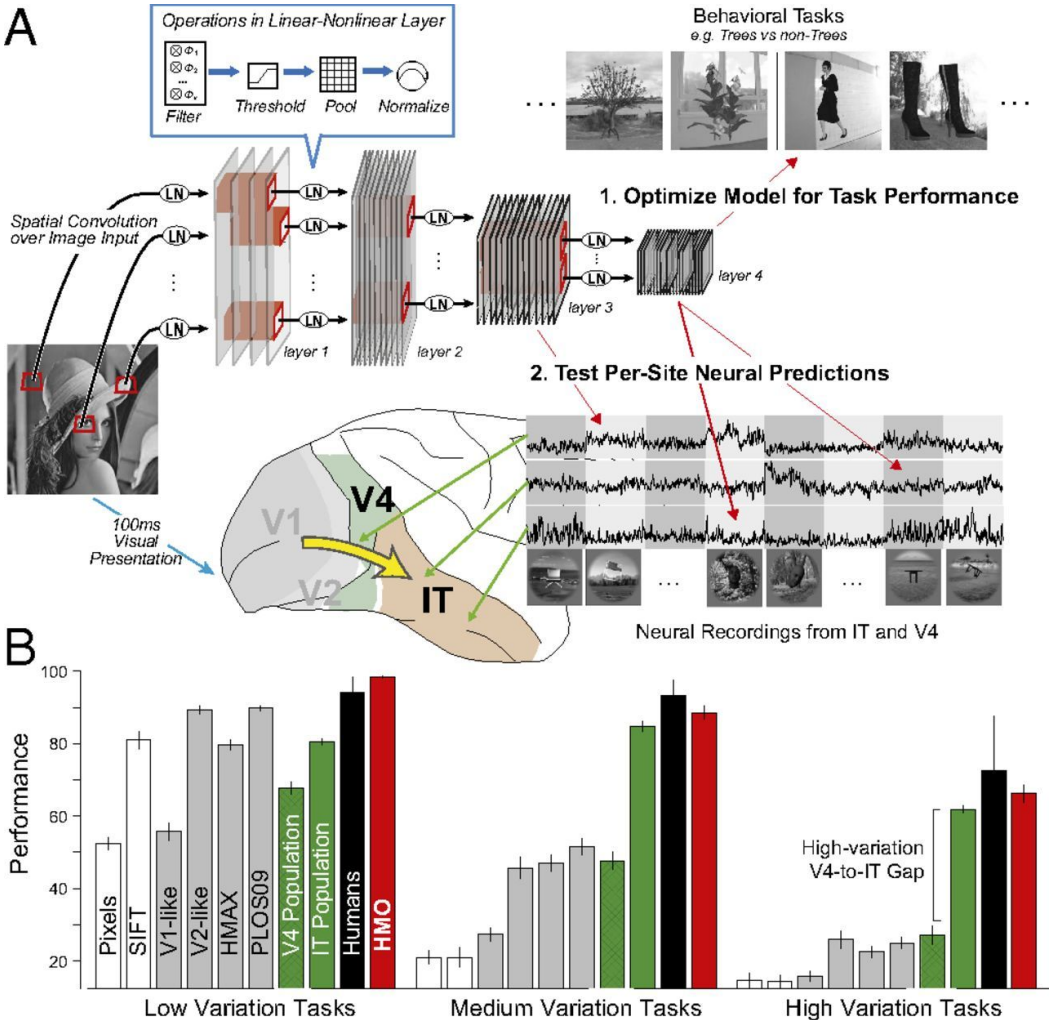
Evaluation of similarity between DNNs and Brain

Linear mapping

Representational (dis)similarity

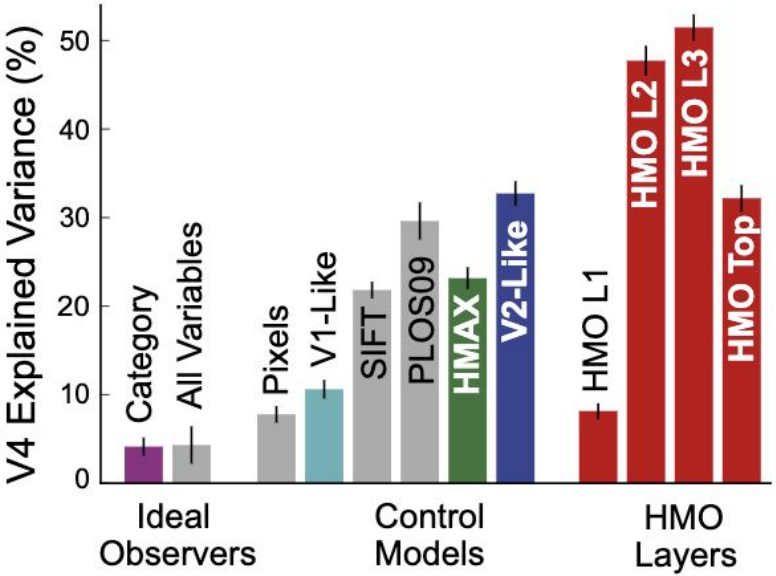


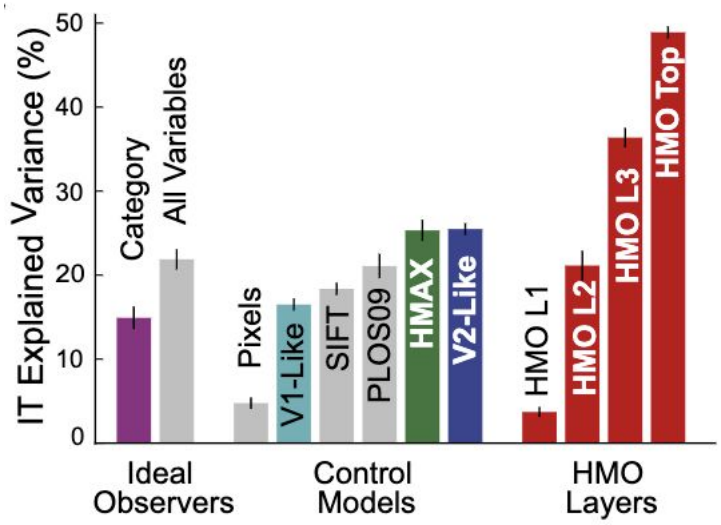
[Kriegeskorte, et al 2008]



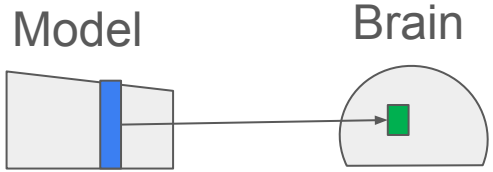
[Yamins, Hong, et al 2014]

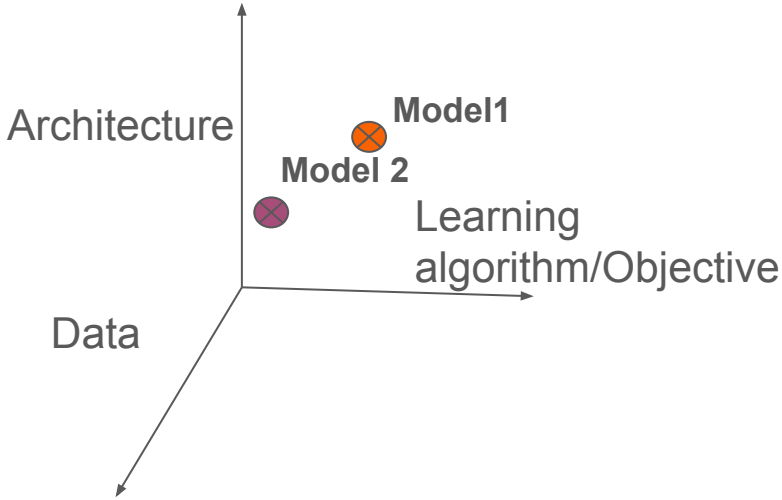
Mapping model units to neurons

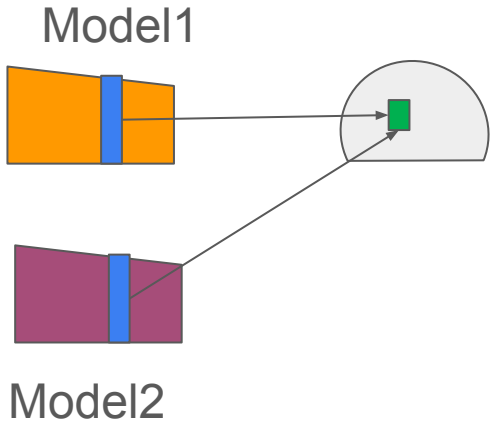




Model selection







score 1 > score 2?

Proc Natl Acad Sci U S A. 2021 Jan 11;118(3):e2014196118. doi: 10.1073/pnas.2014196118

Unsupervised neural network models of the ventral visual stream

Chengxu Zhuang [a,1], Siming Yan [b], Aran Nayebi [c], Martin Schrimpf [d], Michael C Frank [a], James J DiCarlo [d], Daniel L K Yamins [a,e,f]

PLoS Comput Biol. 2022 Jan 7;18(1):e1009739. doi: 10.1371/journal.pcbi.1009739

Increasing neural network robustness improves match to macaque V1 eigenspectrum, spatial frequency preference and predictivity

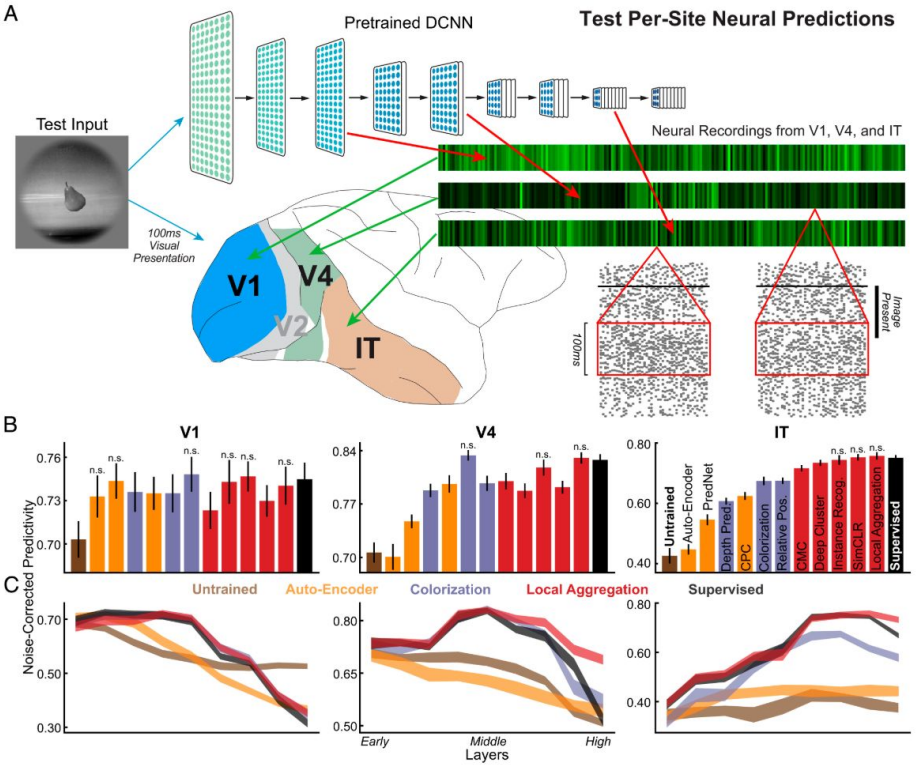
Nathan C L Kong [1,3,*], Eshed Margalit [2,3], Justin L Gardner [1,3], Anthony M Norcia [1,3]

Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic Scenes

Aran Nayebi[1,*], Rishi Rajalingham[1,4], Mehrdad Jazayeri[1,2], and Guangyu Robert Yang[1,2,3]

being explicitly trained to do so. Finally, we find that not all foundation model latents are equal. Notably, models that future predict in the latent space of video foundation models that are optimized to support a *diverse* range of egocentric sensorimotor tasks, reasonably match *both* human behavioral error patterns and neural dynamics across all environmental scenarios that we were able to test. Overall,

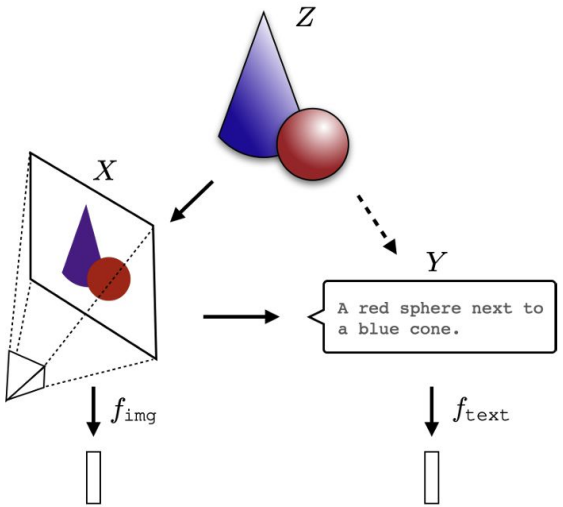
self-supervised models compared to supervised models



There is a catch!

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



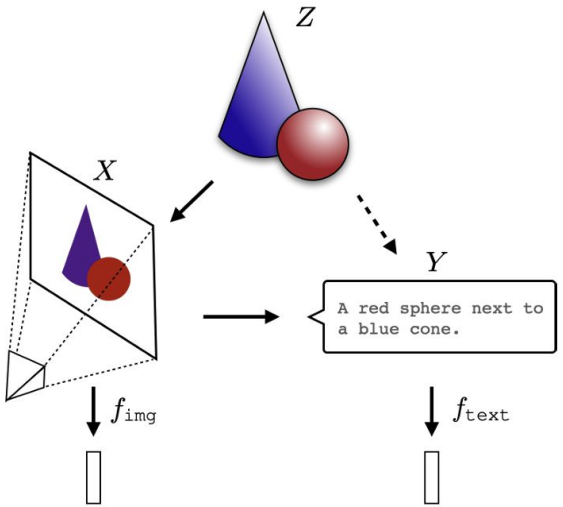
The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

There is a catch!

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



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.

The Capacity Hypothesis

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

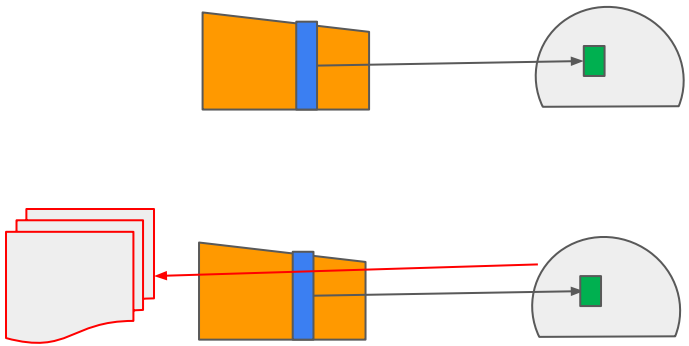
The Simplicity Bias Hypothesis

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.

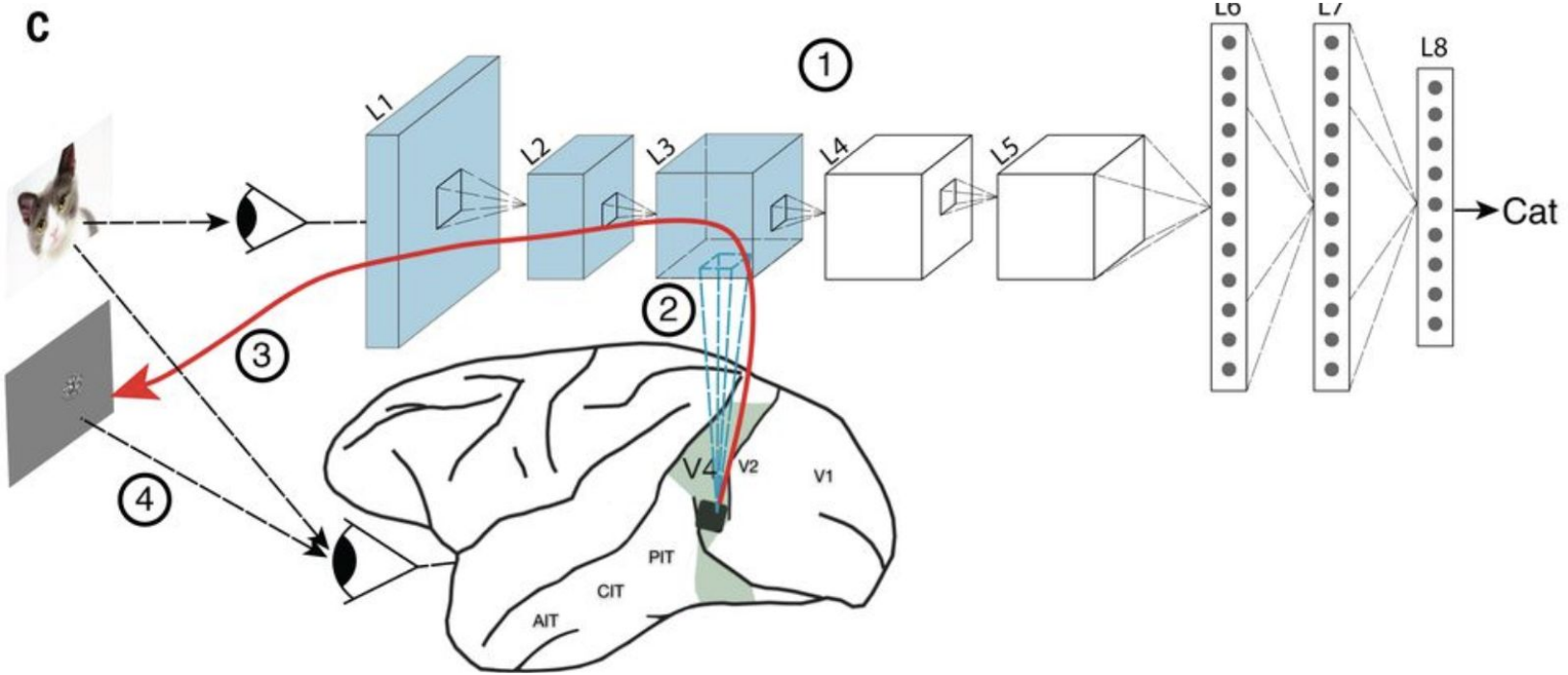
Roadmap

- Comparing models to data
  - Compare model 1 to model 2: which one maps to data better? (**Case study 1**)
  - However, the platonic representation hypothesis!
- Closing the loop by synthesizing stimuli
  - Once a model predicts neural data well, it can be inverted, to give new stimuli (**Case study 2**)
  - Still a very new field
- Computational tricks
  - Contrastive learning
  - Regularizations (**Case study 3**)
- Bridging theories (different perspectives, same math!)
  - Attention mechanism to Memory
  - Regularization to Generativity (**Case study 4**)

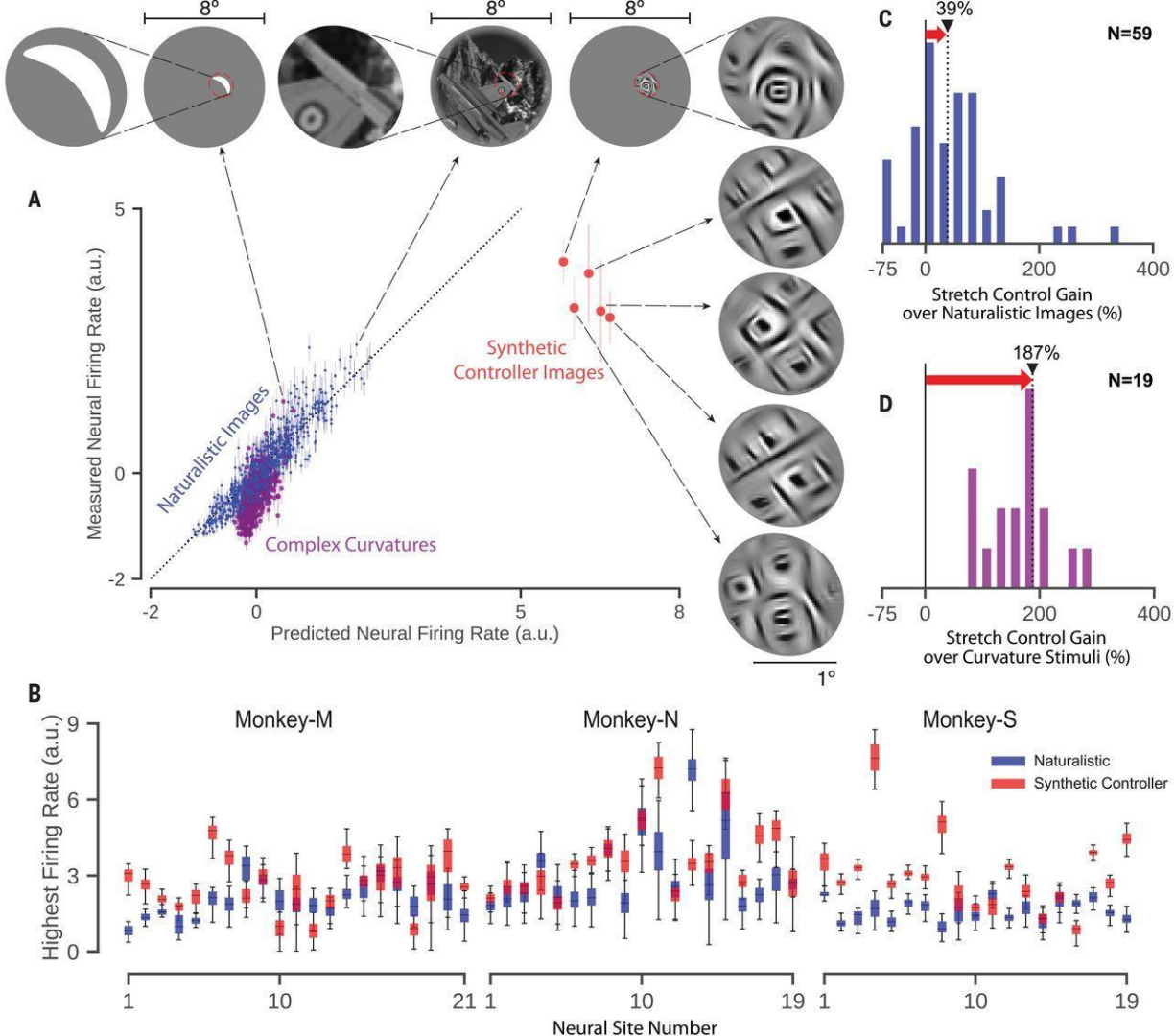
Super stimuli or Maximally exciting inputs (MEI)

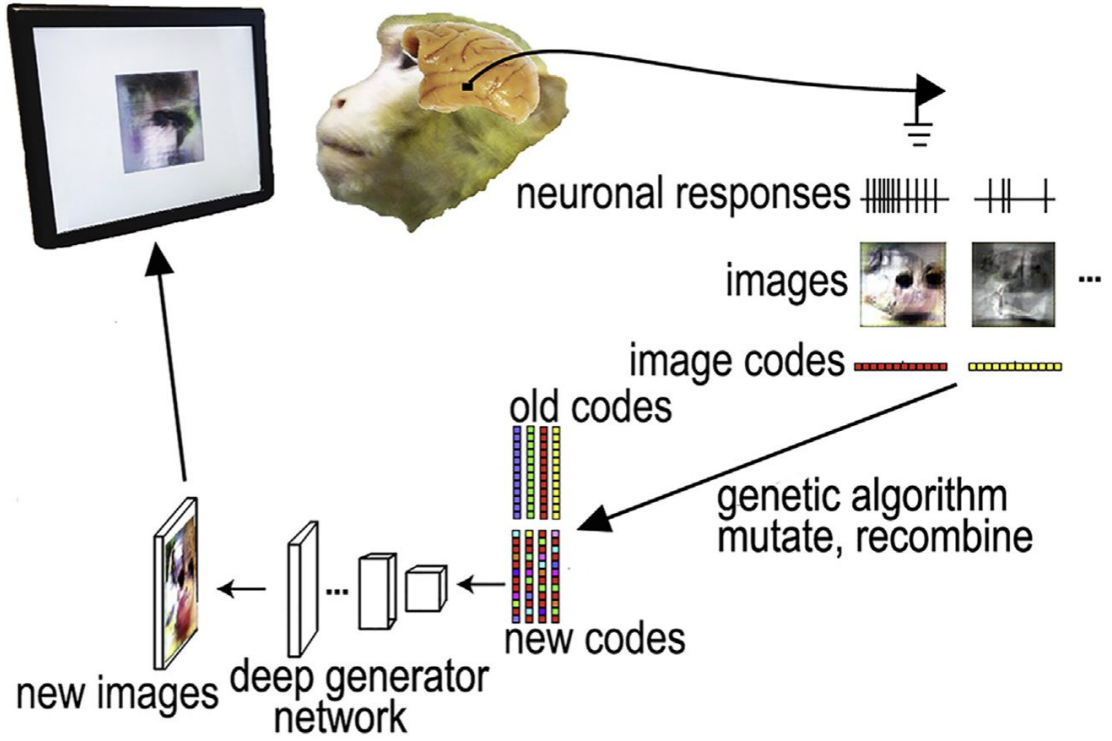


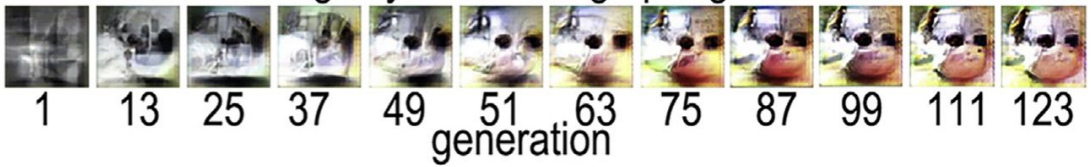
Maximally Exciting inputs



[Bashivan et al, 2019]

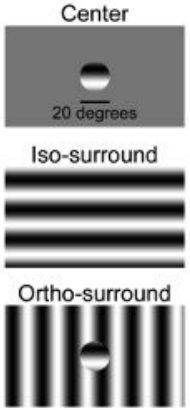

8°
8°
8°
A
Measured Neural Firing Rate (a.u.)
Predicted Neural Firing Rate (a.u.)
Naturalistic Images
Complex Curvatures
Synthetic
Controller Images
1°
C
39%
N=59
Stretch Control Gain
over Naturalistic Images (%)
D
187%
N=19
Stretch Control Gain
over Curvature Stimuli (%)
B
Monkey-M
Monkey-N
Monkey-S
Naturalistic
Synthetic Controller
Highest Firing Rate (a.u.)
Neural Site Number

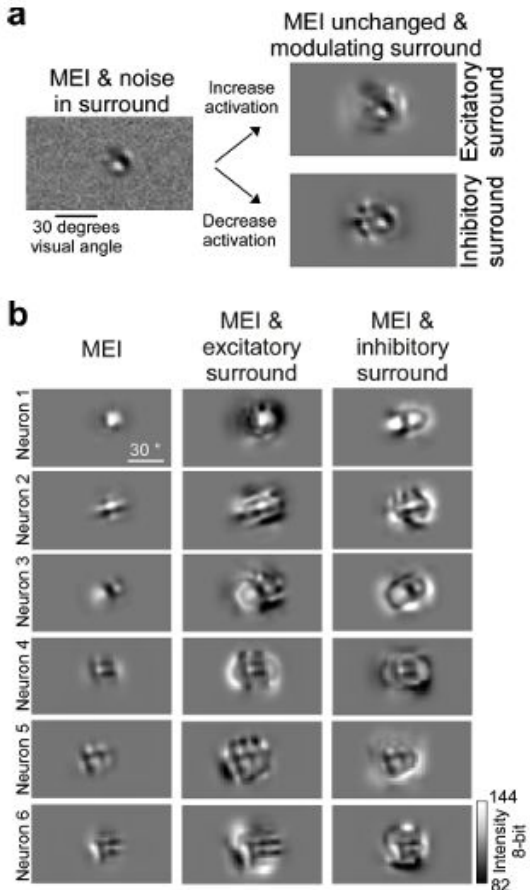




[Ponce et al, 2019]

What can be learned from MEIs?



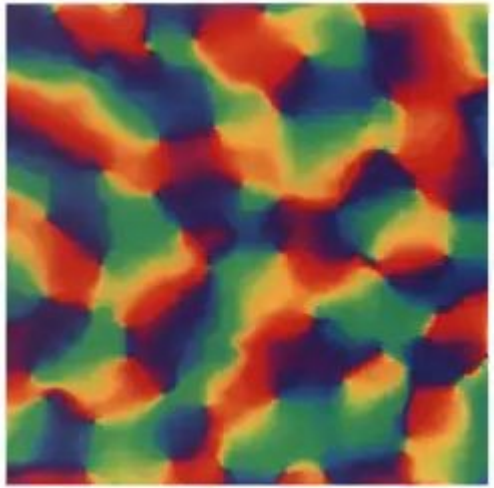


Pattern completion and disruption characterize contextual modulation in the visual cortex

Jiakun Fu[1], Suhas Shrinivasan[2,*], Luca Baroni[8,*], Zhuokun Ding[1,9-11], Paul G. Fahey[1,9-11], Paweł A. Pierzchlewicz[2], Kayla Ponder[1], Rachel Froebe[1,9-11], Lydia Ntanavara[1,9-11], Taliah Muhammad[1], Konstantin F. Willeke[2], Eric Wang[1,9-11], Zhiwei Ding[1,9-11], Dat T. Tran[1,9-11], Stelios Papadopoulos[1,9-11], Saumil Patel[1,9-11], Jacob Reimer[1], Alexander S. Ecker[2,3], Xaq Pitkow[1,6,7], Jan Antolik[8], Fabian H. Sinz[1,2,5], Ralf M. Häfner[4], Andreas S. Tolias[1,9-11,†✉], and Katrin Franke[1,9-12,†✉]

Topography matters?

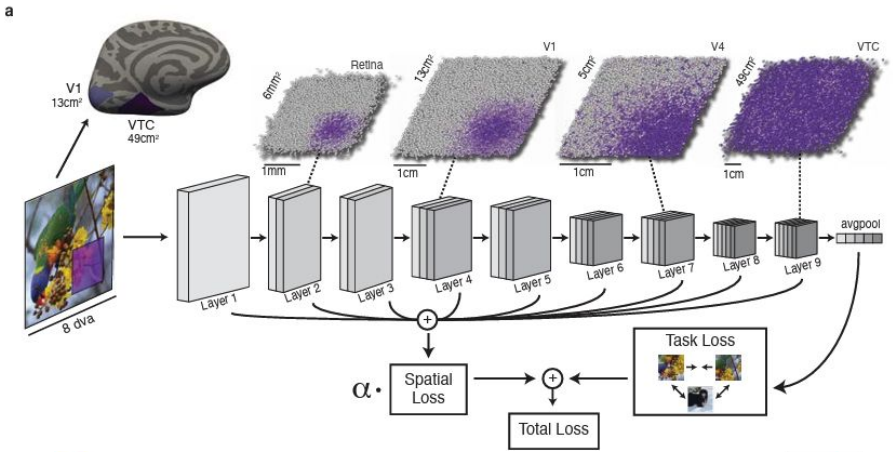
Beyond explained variance: cortical maps examples

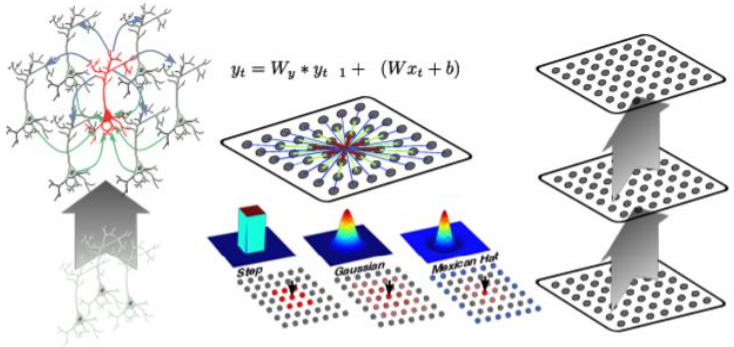


Low-level vision

High-level vision

Two modeling approaches: direct optimization for the effect | use of bio-inspired inductive biases





Neuron

Volume 112, Issue 14, 17 July 2024, Pages 2435-2451.e7

Article

A unifying framework for functional organization in early and higher ventral visual cortex

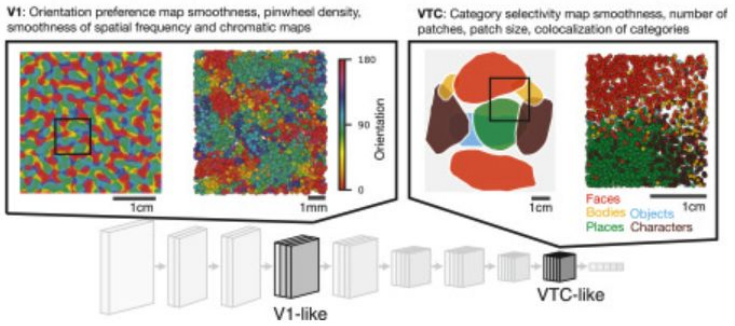
Eshed Margalit [1 9], Hyodong Lee [2], Dawn Finzi [3 4], James J. DiCarlo [2 5 6], Kalanit Grill-Spector [3 7 8], Daniel L.K. Yamins [3 4 7 8]

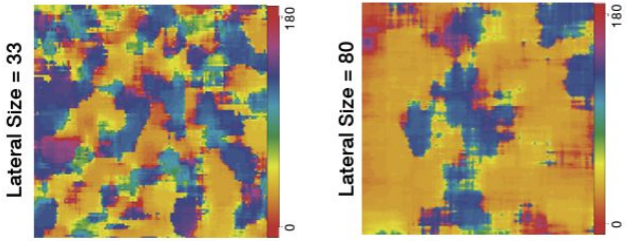
Local lateral connectivity is sufficient for replicating cortex-like topographical organization in deep neural networks

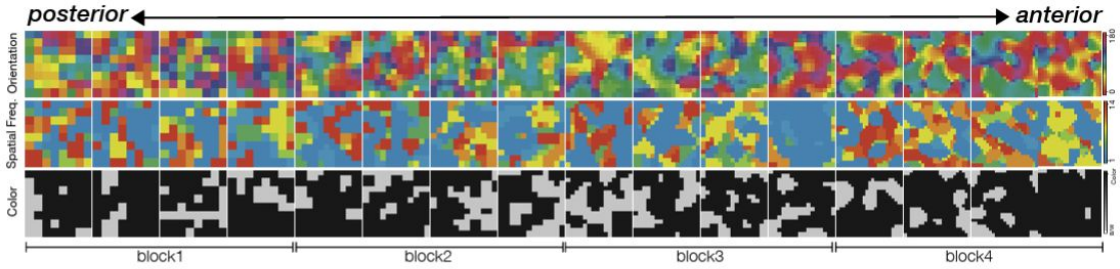
Xinyu Qian[1], Amir Ozhan Dehghani[2], Asa Borzabadi Farahani[1], Pouya Bashivan[1,2,3]
[1] Department of Computer Science, McGill University
[2] Department of Physiology, McGill University
[3] Mila, Université de Montréal

October 4, 2024

Both approaches work







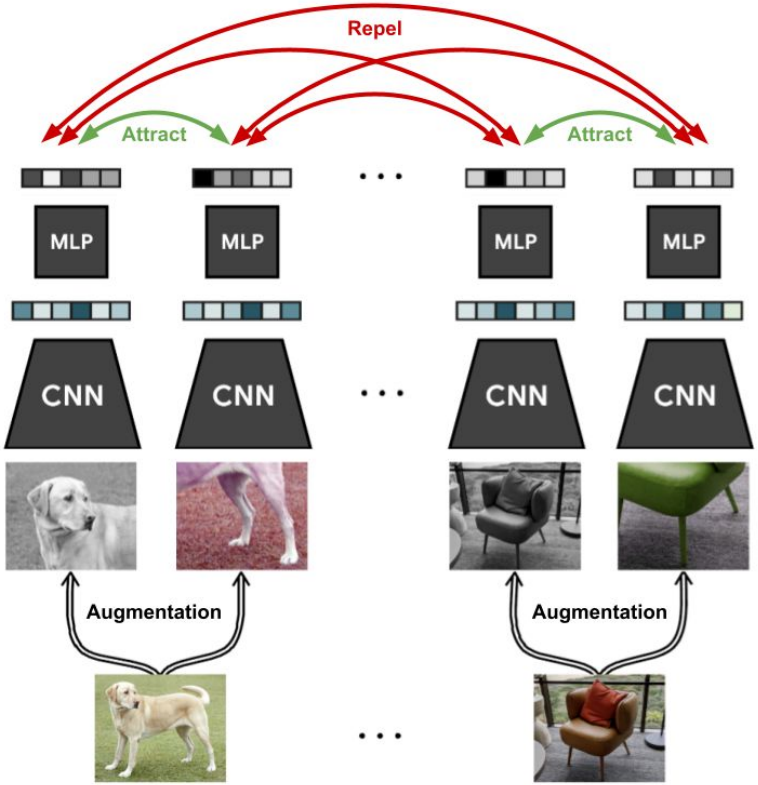
Roadmap

- Comparing models to data
  - Compare model 1 to model 2: which one maps to data better? (**Case study 1**)
  - However, the platonic representation hypothesis!
- Closing the loop by synthesizing stimuli
  - Once a model predicts neural data well, it can be inverted, to give new stimuli (**Case study 2**)
  - Still a very new field
- Computational tricks
  - Contrastive learning
  - Regularizations (**Case study 3**)
- Bridging theories (different perspectives, same math!)
  - Attention mechanism to Memory
  - Regularization to Generativity (**Case study 4**)

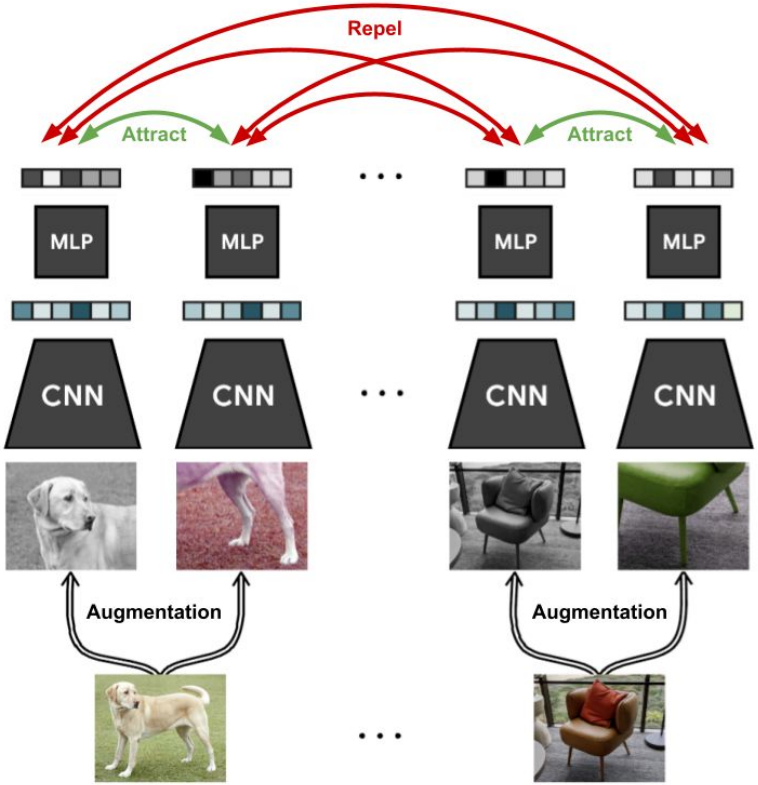
Useful Computational hacks neuroscience learned from deep learning framework

- Contrastive learning
  - In Tools: Cebra
  - In Plasticity rules: Hebbian+predictive
- Attention
  - 3-factor plasticity rule using astrocytes
- Regularization
  - Temporal prediction can be learned from robustness to noise

Contrastive learning



Contrastive learning



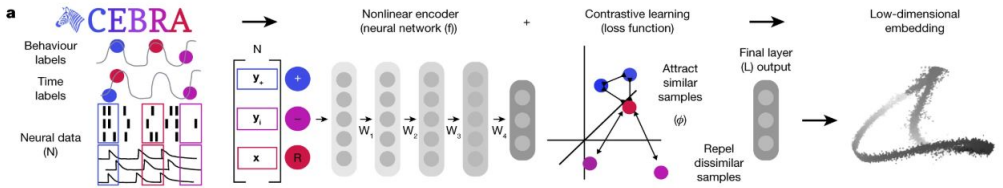
Regularization

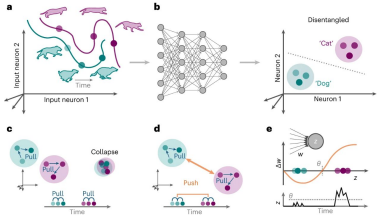
Optimization Goal

Task relevant	*Task irrelevant*	
Objective +	Regularization	**Primary visual cortex propeties**
e.g.	*e.g.*	
Reconstruction	*Sparse coding*	*Gabor-like receptive fileds*
Classification	*Temporal continuity*	*Translation invariance*
	Wiring minimization	*Ocular dominance map*
	noise robustness	*Temporal prediction*

Useful Computational hacks neuroscience learned from deep learning framework

- Contrastive learning
  - In Tools: Cebra
  - In Plasticity rules: Hebbian+predictive





nature neuroscience

Explore content · About the journal · Publish with us

nature › nature neuroscience › articles › article

Article | Open access | Published: 12 October 2023

The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks

Manu Srinath Halvagal & Friedemann Zenke

nature

Explore content · About the journal · Publish with us

nature › articles › article

Article | Open access | Published: 03 May 2023

Learnable latent embeddings for joint behavioural and neural analysis

Steffen Schneider, Jin Hwa Lee & Mackenzie Weygandt Mathis

- Attention
  - 3-factor plasticity rule using astrocytes

Building transformers from neurons and astrocytes

Leo Kozachkov, Ksenia V. Kastanenka, and Dmitry Krotov Authors Info & Affiliations

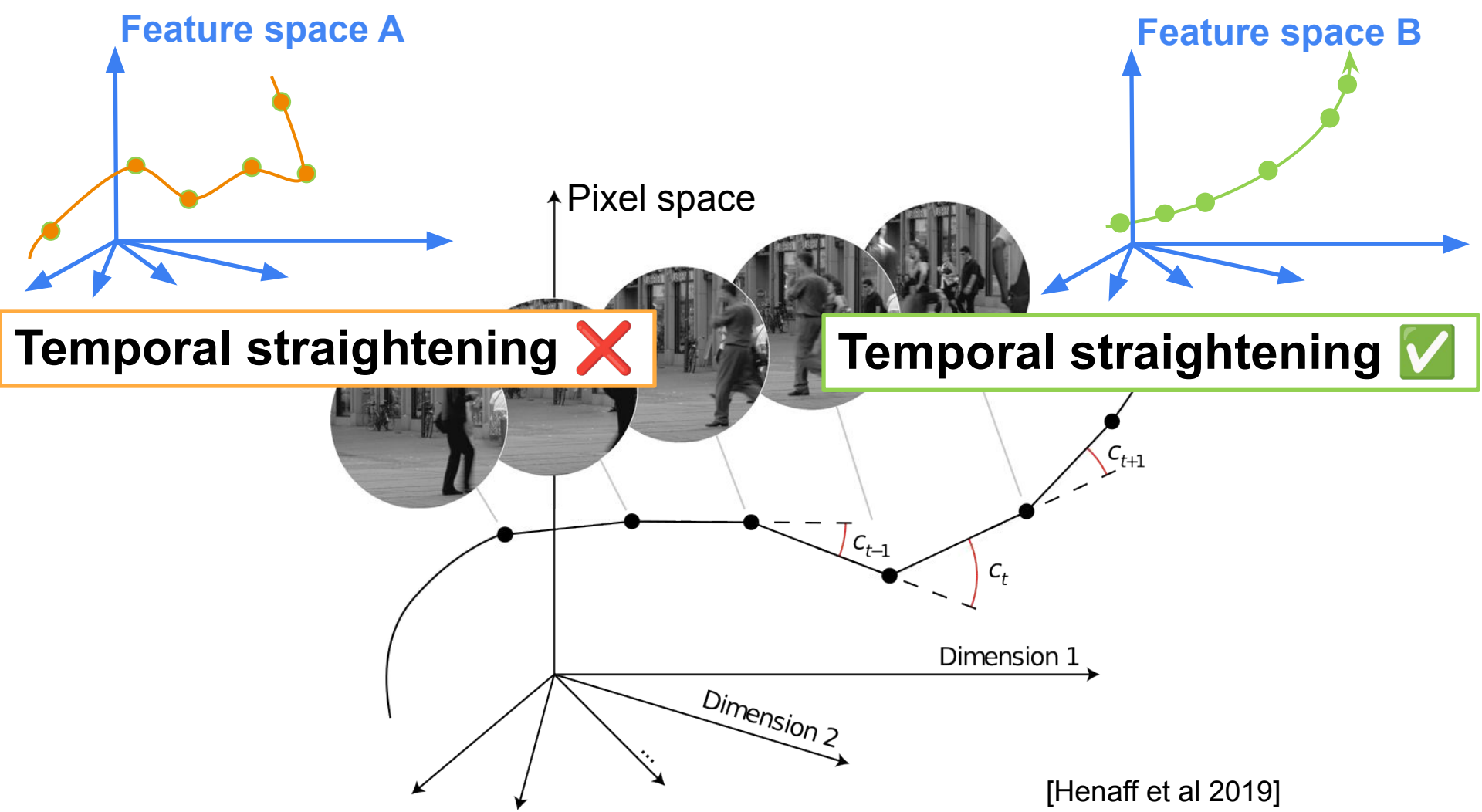
Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received November 9, 2022; accepted June 22, 2023

- Regularization
  - Temporal prediction can be learned from robustness to noise

Case study 3:
Temporal prediction can be achieved by regularized learning

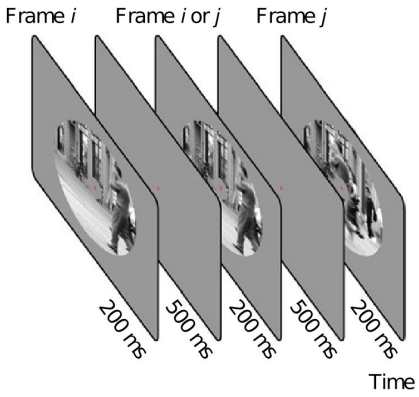
How to quantify the ability to predict over time?

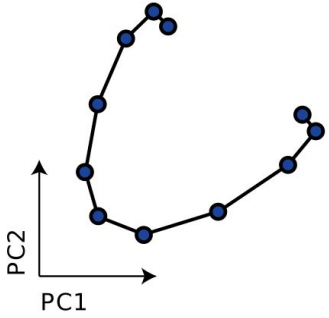
[Henaff et al 2019]



[Henaff et al 2019]

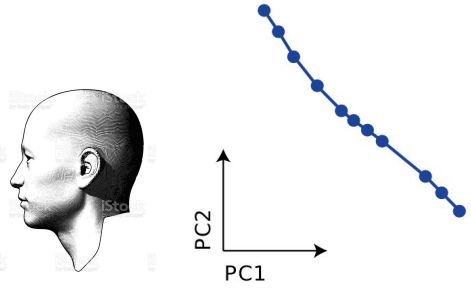
Pixel space





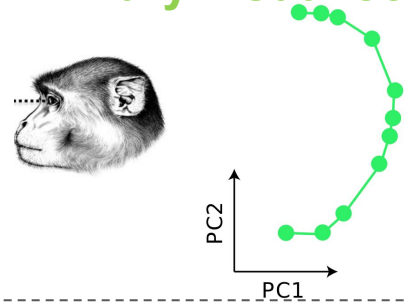
[Henaff et al 2019, 2021]

Perceptual space of humans



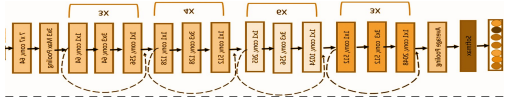
~30°

Primary visual cortex of macaque

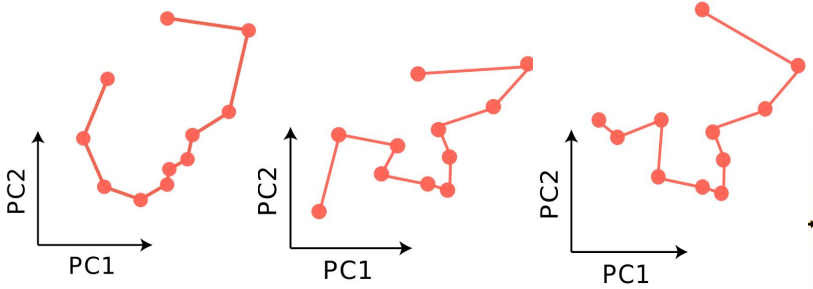


~10°

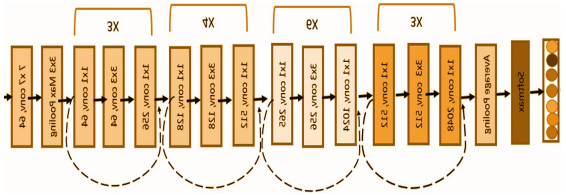
Deep Neural Networks



?



Deep Neural Networks



More straightening

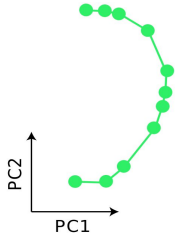
Curvature change

0

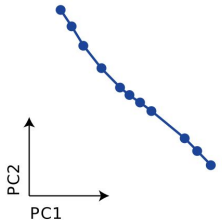
-10°

-30°

Primary visual cortex



Perceptual space



[Henaff, et al 2019, 2021]

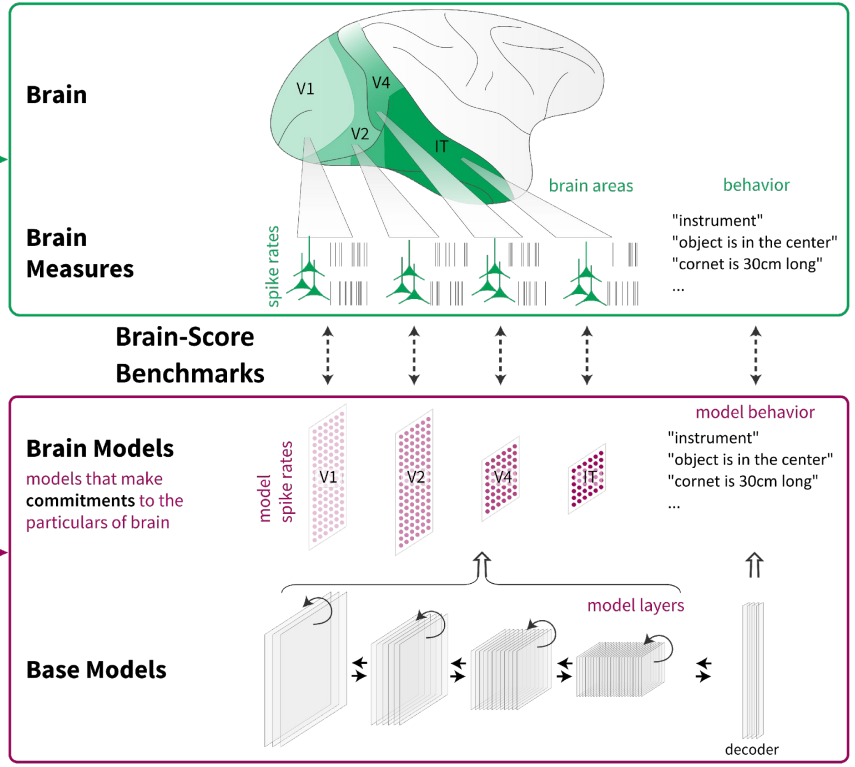
Puzzling: Why deep neural net models of vision don't exhibit temporal straightening?

Brain-like

Deep Nets explain variance in neural data ✅

Deep Nets do not exhibit temporal straightening ❌

input

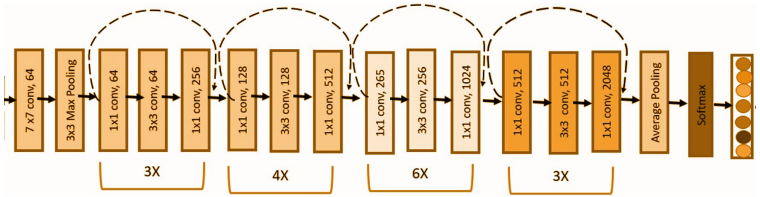


Previous conjecture: DNNs needs to be trained to predict the next frame to be able to straightened the temporal trajectories

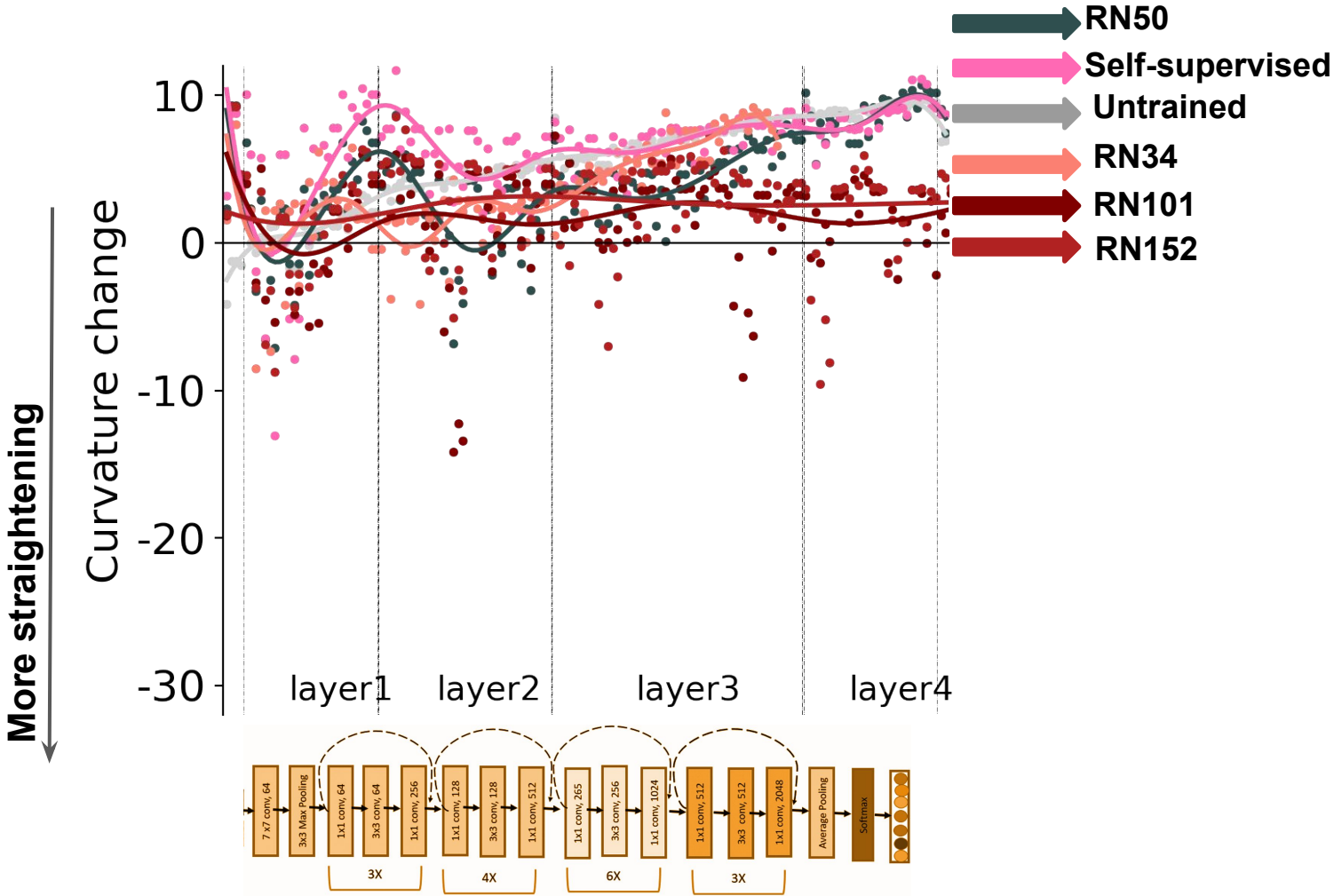
[Henaff, et al 2019, 2021]

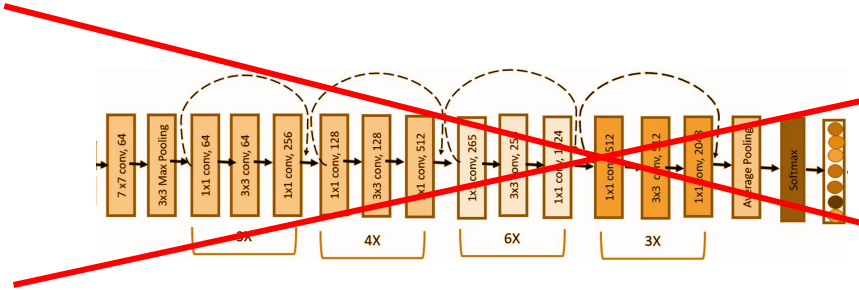
[Yamins & Hong, et al, 2014, Schrimpf et al. 2020]

Can any members of the feedforward DNN class straighten the trajectory of natural movies?



Classification?
Self-Supervised?
Deeper?
Untrained?
….

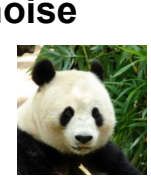

RN50
Self-supervised
Untrained
RN34
RN101
RN152
10
0
-10
-20
-30
Curvature change
More straightening
layer1
layer2
layer3
layer4
7x7 conv, 64
3x3 Max Pooling
1x1 conv, 64
3x3 conv, 64
1x1 conv, 256
3X
1x1 conv, 128
3x3 conv, 128
1x1 conv, 512
4X
1x1 conv, 265
3x3 conv, 256
1x1 conv, 1024
6X
1x1 conv, 512
3x3 conv, 512
1x1 conv, 2048
3X
Average Pooling
Softmax

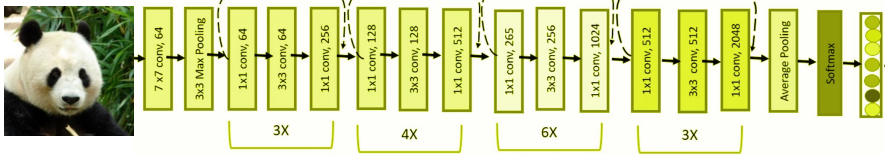


~~Classification?~~
~~Self-Supervised?~~
~~Deeper?~~
~~Untrained?~~

Next, we considered how ***training for noise robustness*** may affect manifold straightness in feedforward DNNs

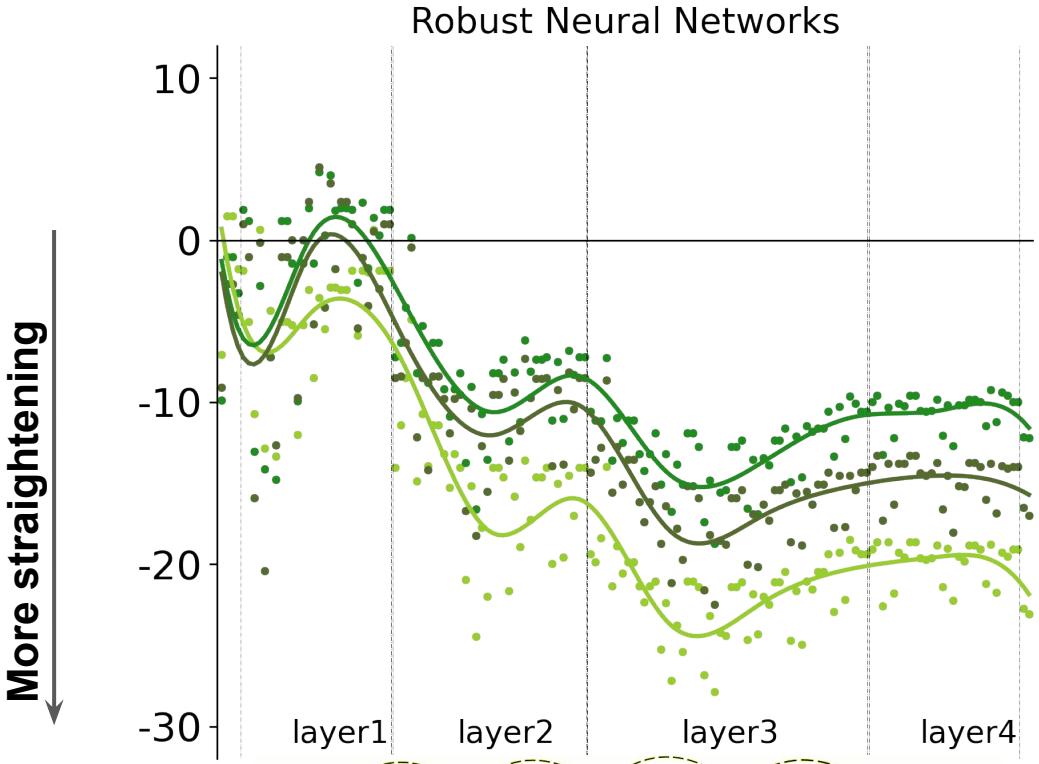
+ Imperceptible engineered noise





Classification +
Robustness to adversarial noise

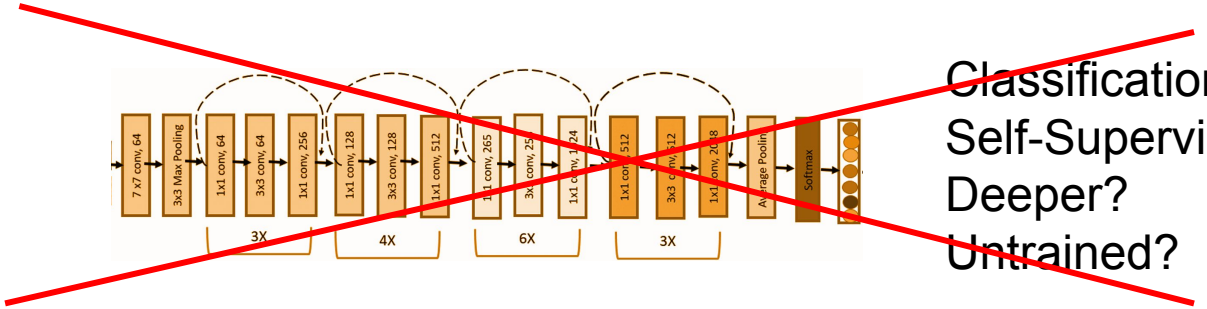
Robustness to adversarial noise give rise to temporal straightening



Robust to adversarial noise

\+ Imperceptible engineered noise

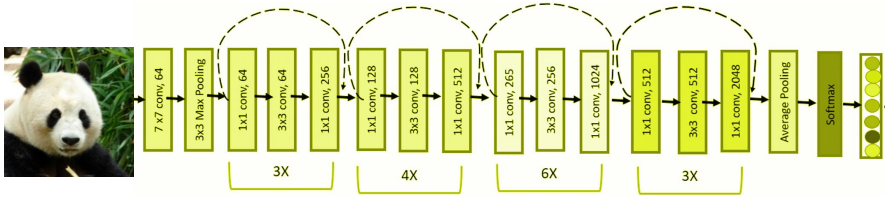
Classification +
Robustness to adversarial noise



Classification?
Self-Supervised?
Deeper?
Untrained?

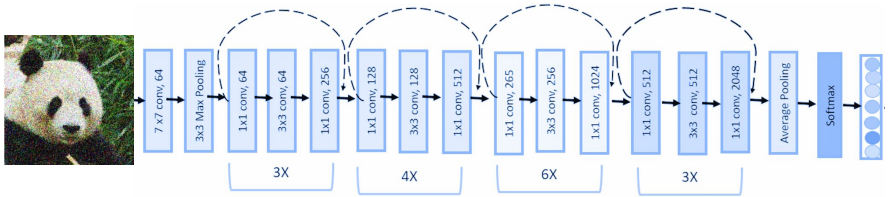
Next, we considered how ***<u>training for noise robustness</u>*** may affect manifold straightness in feedforward DNNs

+ Imperceptible engineered noise



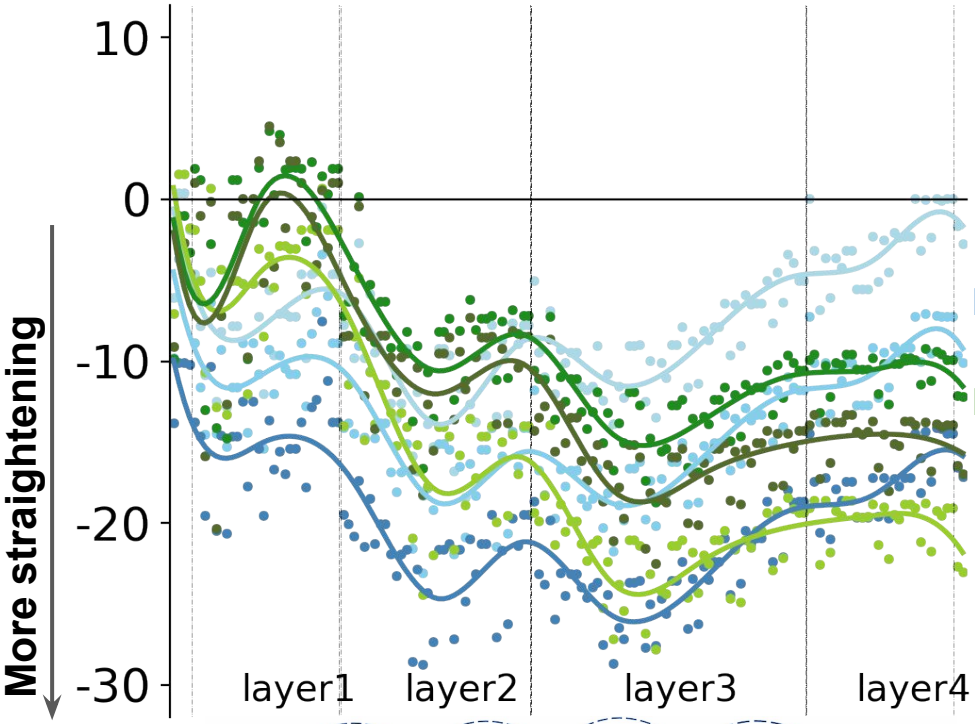
Classification +
Robustness to adversarial noise

+ Random Gaussian noise



Classification +
Robustness to Gaussian noise

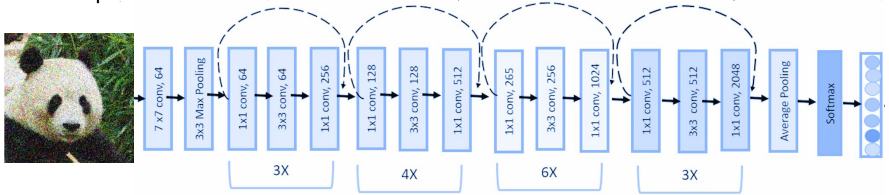
Robustness to noise give rise to temporal straightening



Robust to Gaussian noise

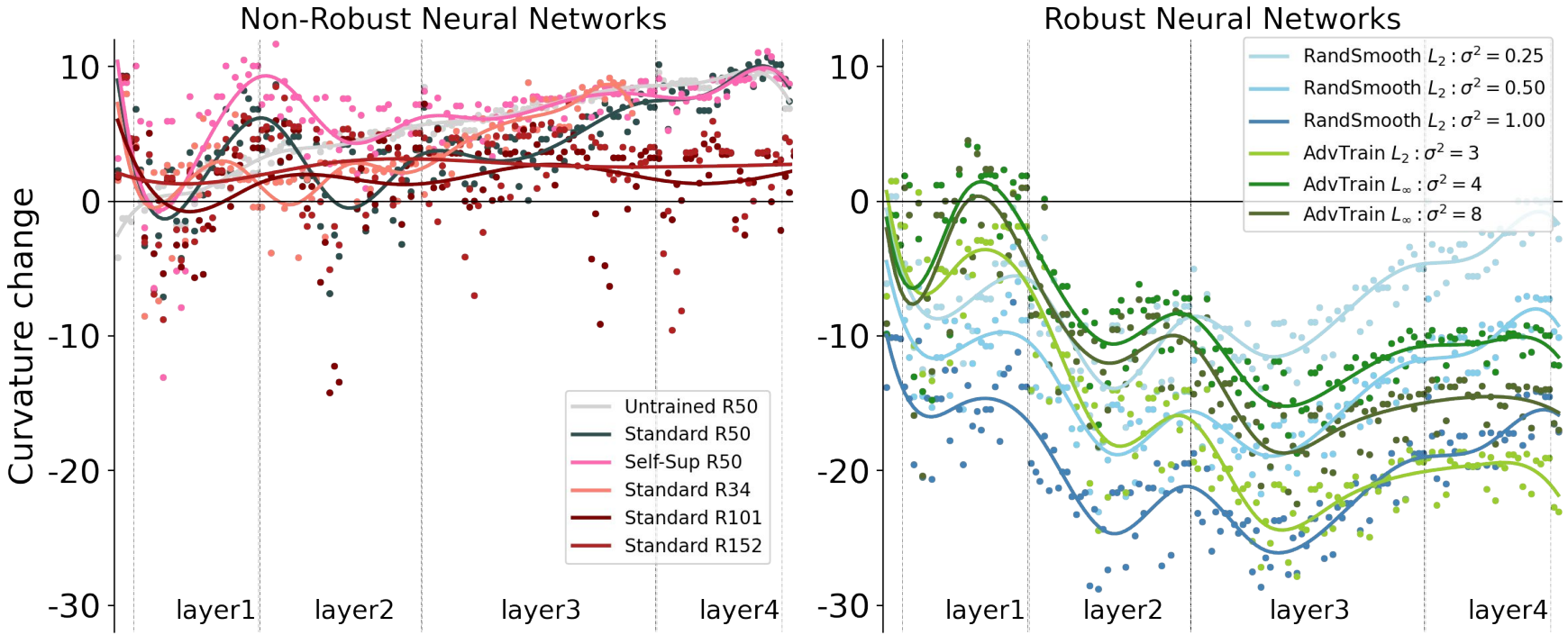
Robust to adversarial noise

\+ **Random Gaussian noise**



Classification +
Robustness to Gaussian noise

Feedforward neural networks can exhibit straightening…



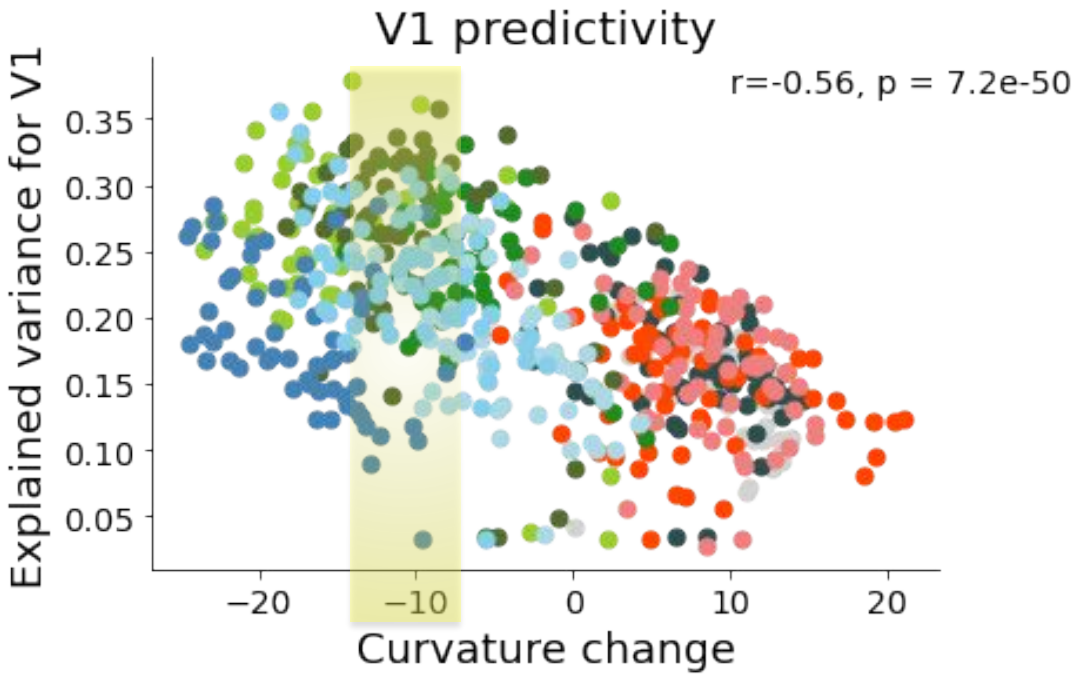
If they were trained to be robust to input noise!

[Toosi and Issa, ICLR, 2023]

How is straightening related to neural predictivity?

	Standard Deep Nets	Robust Deep Nets
Predict neural data	✅	?
temporal straightening	❌	✅

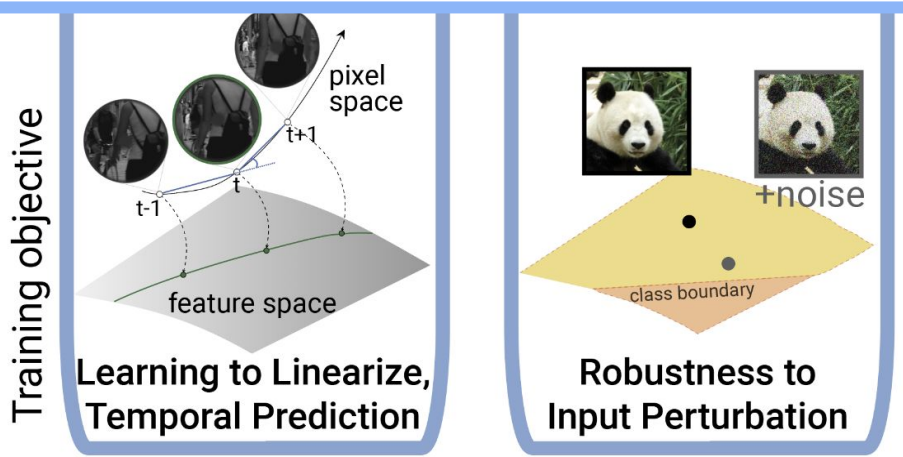
How is straightening related to neural predictivity?



[Toosi and Issa, ICLR, 2023]

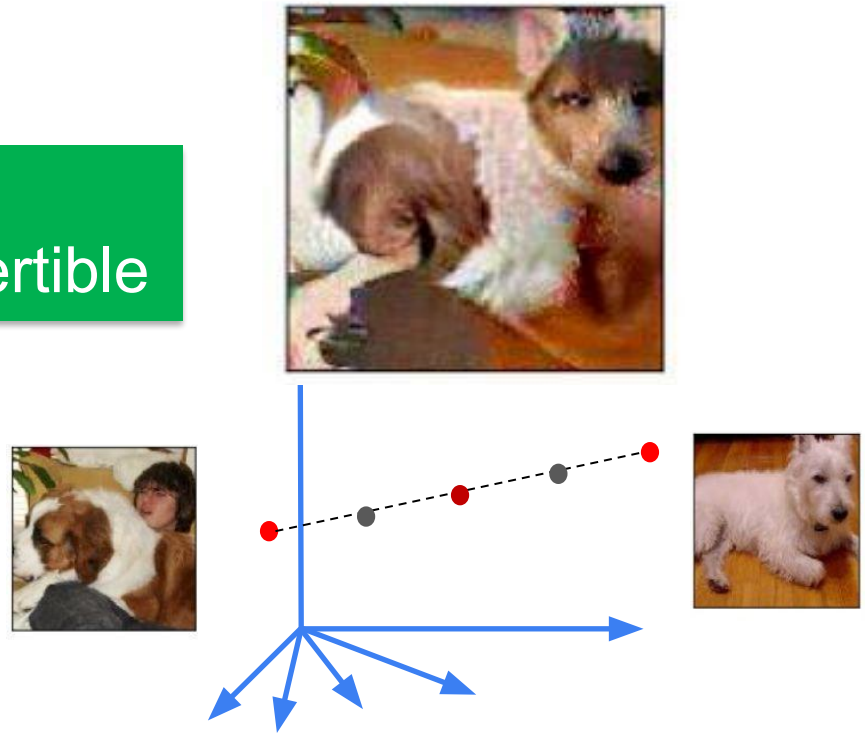
What is the intuition (and theory) behind the link?

temporal straightening & robustness to noise



Invertibility: Linear interpolation between the representations in the feature space corresponds to natural features

Invertible feature space

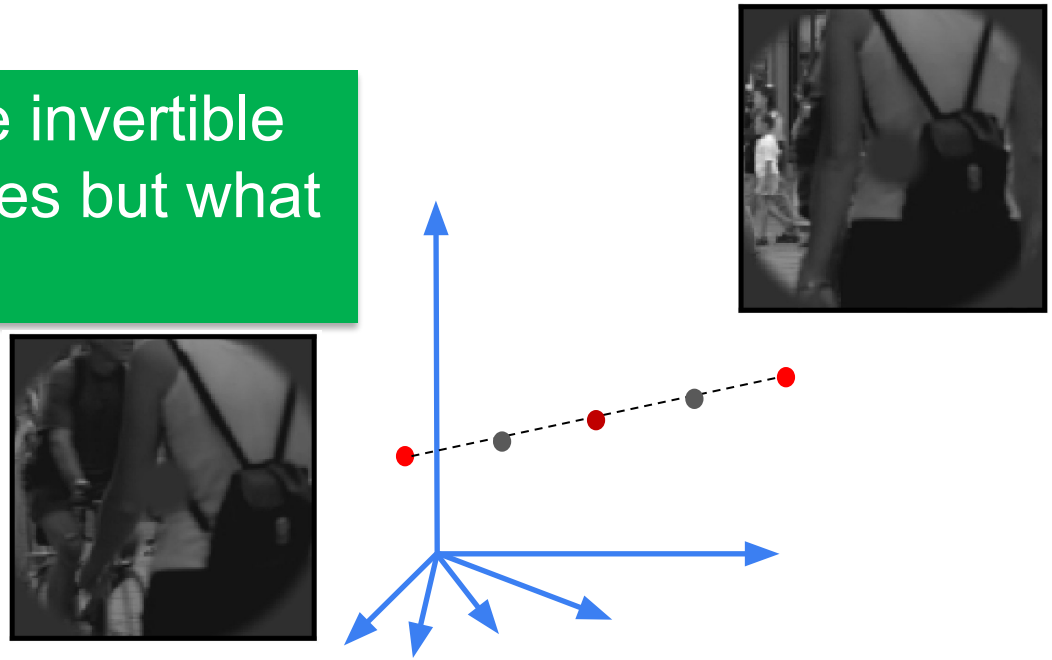


Prior finding:
Robust neural networks are invertible

[Engstrom et al, 2018]

Invertible feature space

Robust neural networks are invertible when interpolating categories but what about in natural movies?

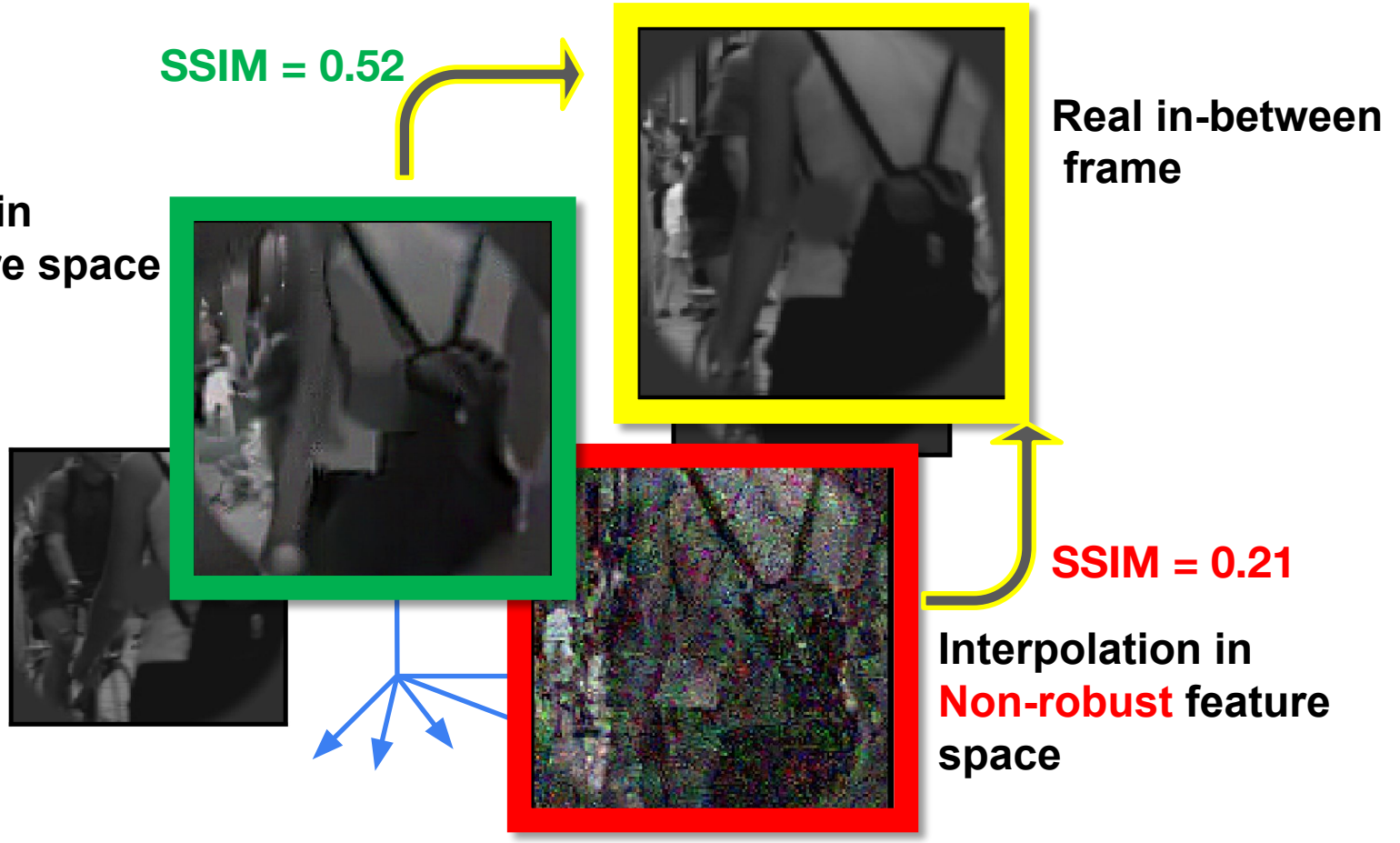


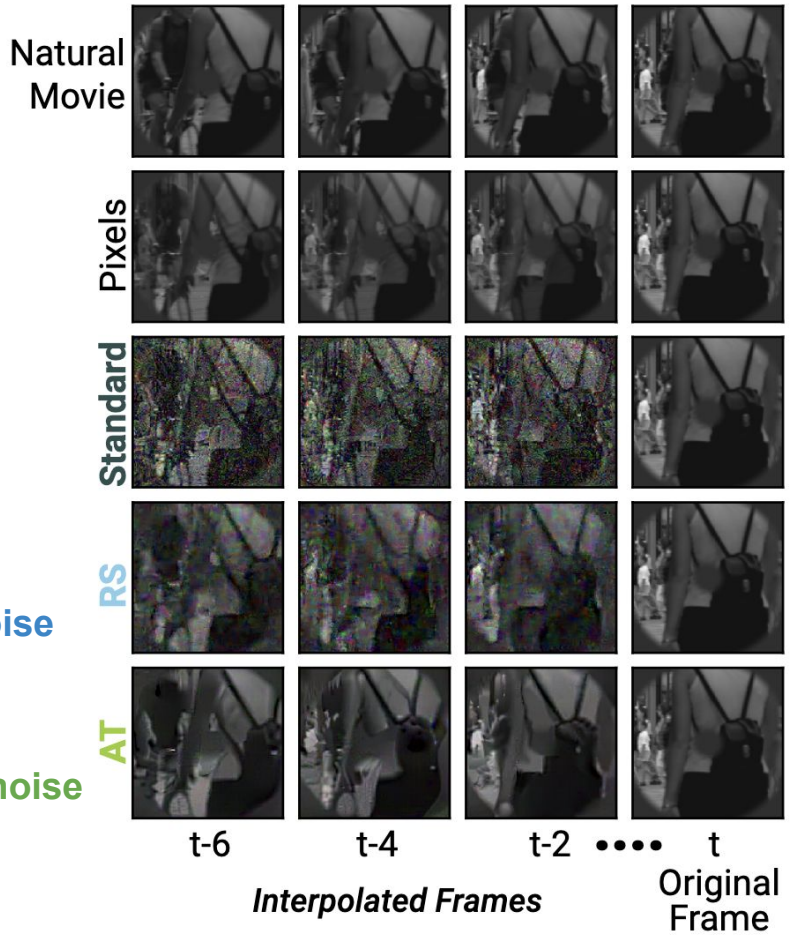
Inverting robust vs. non-robust features of natural movies

Interpolation in Robust feature space

Interpolation in Non-robust feature space

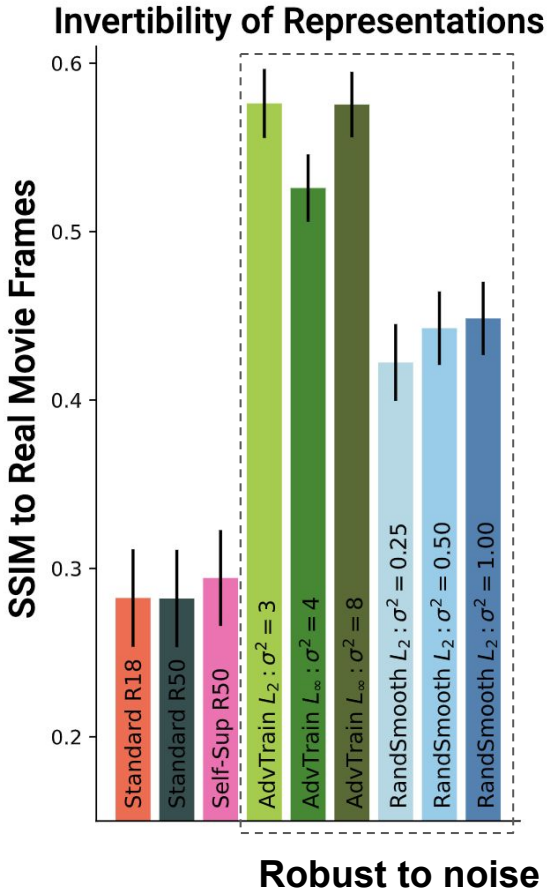
Quantitative comparison of interpolated vs. actual frame





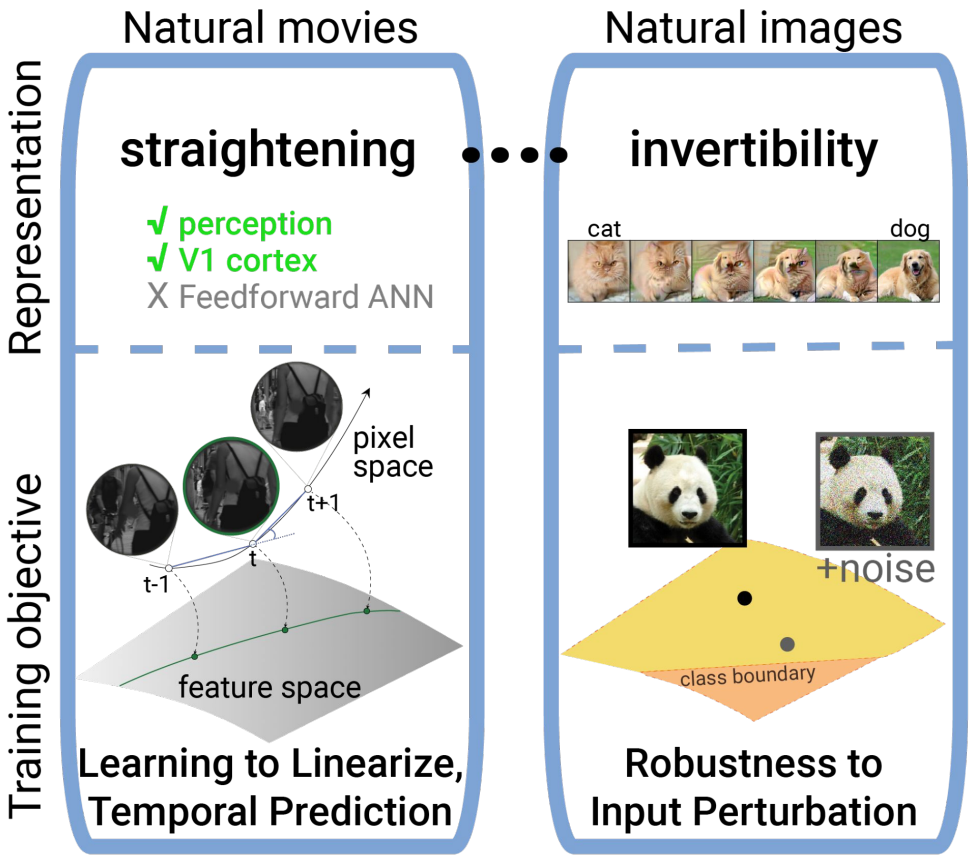
Robust to Gaussian noise

Robust to adversarial noise

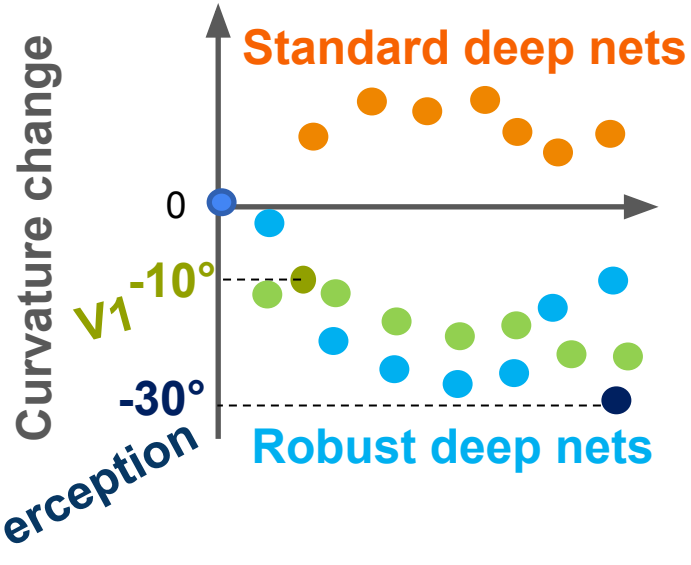


[Toosi and Issa, ICLR, 2023]

Summary



Robustly trained deep nets demonstrate natural movie straightening in their representations, *without directly training on movies, or for prediction*



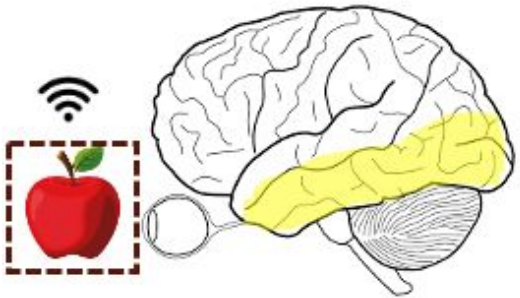
Roadmap

- Comparing models to data
  - Compare model 1 to model 2: which one maps to data better? (**Case study 1**)
  - However, the platonic representation hypothesis!
- Closing the loop by synthesizing stimuli
  - Once a model predicts neural data well, it can be inverted, to give new stimuli (**Case study 2**)
  - Still a very new field
- Computational tricks
  - Contrastive learning
  - Regularizations (**Case study 3**)
- Bridging theories (different perspectives, same math!)
  - Recognition | Generation
  - Learning | Attention

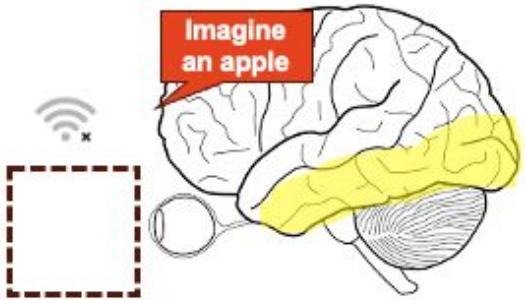
Theory bridges seemingly different computations

- How does the same neural circuitry support classification and generation?
  - (Regularized) discriminative models (i.e. classifiers) → Generative models
- Plasticity ←→ Memory
  - Key-value attention in transformers → Memory storage and retrieval

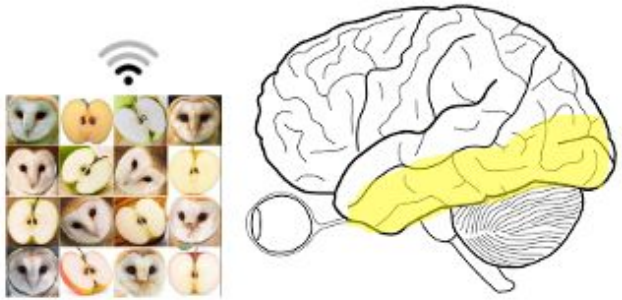
Sensory driven



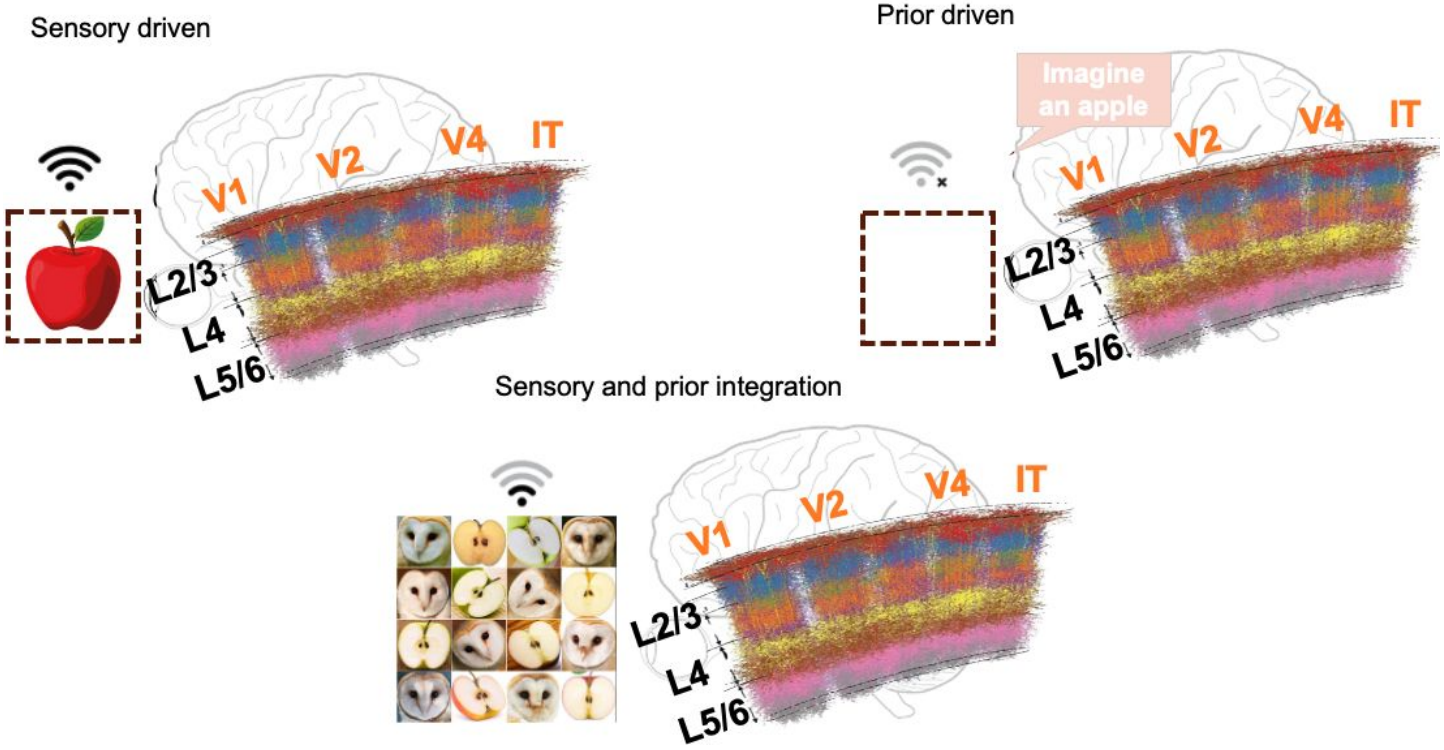
Prior driven



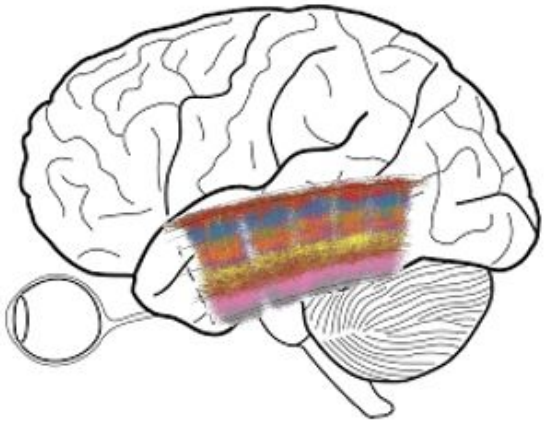
Sensory and prior integration



Same visual system supports both classification and generation

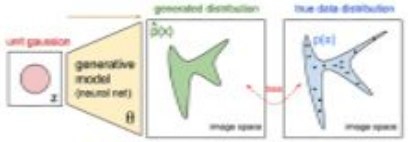


But different classes of models for each capability…

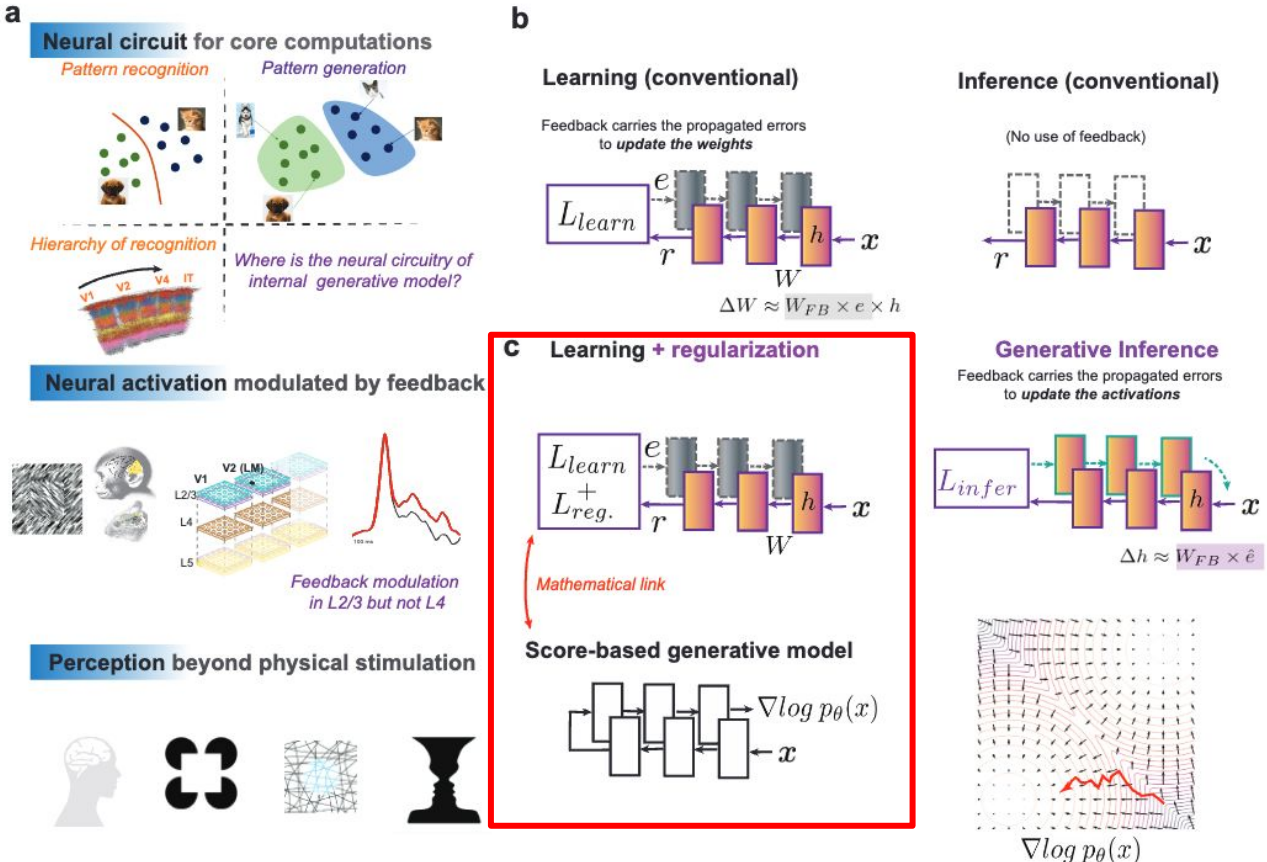


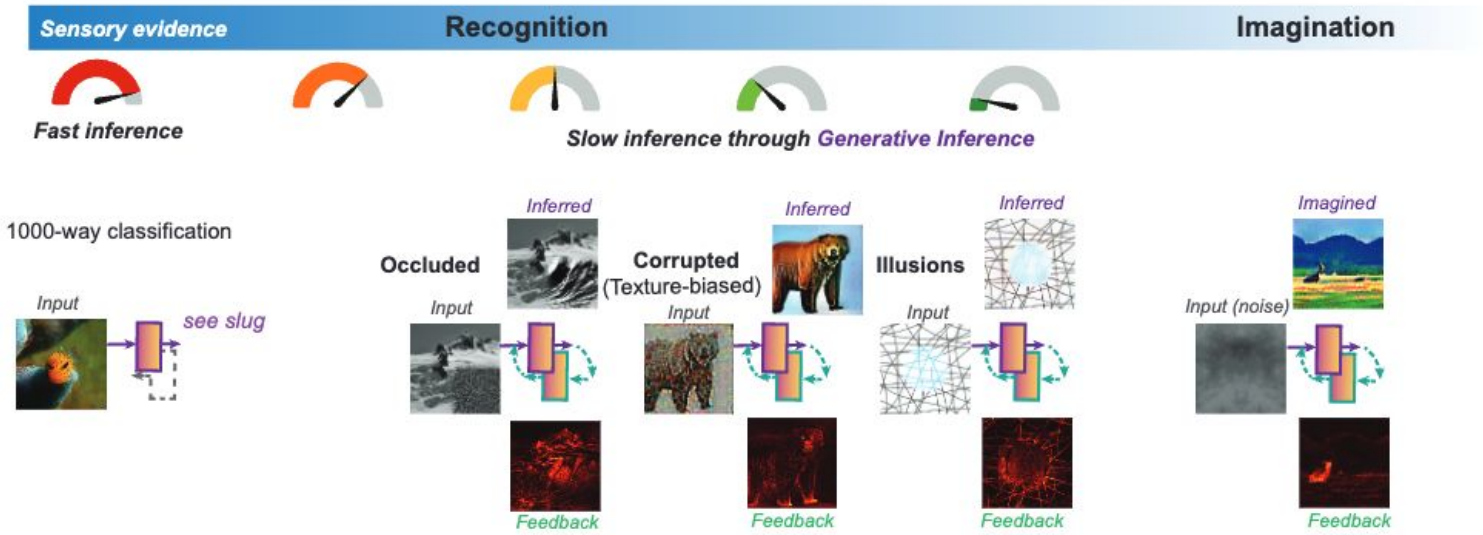
Discriminative models

Generative models



Recognition ⟷ Generation





Duality of discrimination and generation we established using deep learning framework, helped to understand the puzzling role of feedback in vision.

Generative perceptual inference in deep neural network models of object recognition induces illusory contours and shapes

Tahereh Toosi (tahereh.toosi@columbia.edu)
Kenneth D. Miller (ken@neurotheory.columbia.edu)

Kazuki Irie*, Róbert Csordás*, Jürgen Schmidhuber

The Dual Form of Neural Networks Revisited

ICML 2022, https://arxiv.org/abs/2202.05798

Irie, Csordás, Schmidhuber ICML 2022

These systems are "equivalent"

Arbitrary v and k

Linear layer

$$\boldsymbol{W} = \sum_{t=1}^{T} \boldsymbol{v}_t \otimes \boldsymbol{k}_t$$

$$S_1(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$$

Key-value memory attention layer

$$\boldsymbol{K} = (\boldsymbol{k}_1, ..., \boldsymbol{k}_T) \in \mathbb{R}^{d_{\text{in}} \times T}$$

$$\boldsymbol{V} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_T) \in \mathbb{R}^{d_{\text{out}} \times T}$$

$$S_2(\boldsymbol{x}) = \text{Attention}(\boldsymbol{K}, \boldsymbol{V}, \boldsymbol{x})$$

NB: Unnormalised dot attention $\text{Attention}(\boldsymbol{K}, \boldsymbol{V}, \boldsymbol{q}) = \boldsymbol{V}\boldsymbol{K}^\intercal\boldsymbol{q}$

Standard attention: $\boldsymbol{V}\text{softmax}(\boldsymbol{K}^\intercal\boldsymbol{q})$

[Slide from Katzuki Irie]

Application to a Linear Layer Trained by Gradient Descent:

Forward computation:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$$

Backward computation (**gradient descent**) to **update W**:

outer product

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t \underbrace{-\eta_t (\nabla_{\boldsymbol{y}} \mathcal{L})_t}_{\boldsymbol{e}_t} \otimes \boldsymbol{x}_t$$

t now denotes training iteration!

for some error function $\mathcal{L}$ learning rate η_t at step $t \in \mathbb{N}$

We can directly apply the duality from the previous slide to:

$$\boldsymbol{W} = \boldsymbol{W}_0 + \sum_{t=1}^{T} \boldsymbol{e}_t \otimes \boldsymbol{x}_t$$

[Slide from Katzuki Irie]

Linear layer trained by gradient descent

Store: $$\boldsymbol{W} = \boldsymbol{W}_0 + \sum_{t=1}^{T} \boldsymbol{e}_t \otimes \boldsymbol{x}_t$$

Compute: $$S_1(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$$

Key/value-attention memory storing entire training experience

$$\boldsymbol{E} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_T) \in \mathbb{R}^{d_{\text{out}} \times T}$$

$$\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T) \in \mathbb{R}^{d_{\text{in}} \times T}$$

$$S_2(\boldsymbol{x}) = \boldsymbol{W}_0\boldsymbol{x} + \text{Attention}(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{x})$$

[Slide from Katzuki Irie]

Linear layer trained by gradient descent

Store: $$\boldsymbol{W} = \boldsymbol{W}_0 + \sum_{t=1}^{T} \boldsymbol{e}_t \otimes \boldsymbol{x}_t$$

Key/value-attention memory storing entire training experience

$$\boldsymbol{E} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_T) \in \mathbb{R}^{d_{\text{out}} \times T}$$

$$\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T) \in \mathbb{R}^{d_{\text{in}} \times T}$$

$$S_2(\boldsymbol{x}) = \boldsymbol{W}_0 \boldsymbol{x} + \text{Attention}(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{x})$$

Key-value memory in the brain

Samuel J. Gershman[1,2,3,*], Ila Fiete[4], and Kazuki Irie[2]
[1]Department of Psychology
[2]Center for Brain Science
[3]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA
[4]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
*Corresponding author: gershman@fas.harvard.edu

January 7, 2025

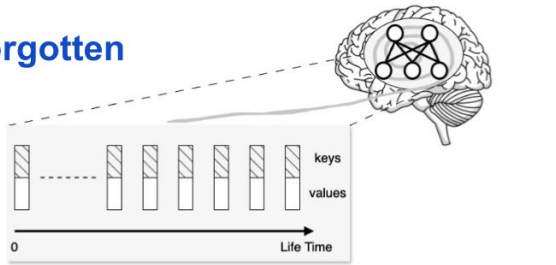
[Slide from Katzuki Irie]

Reminiscent of some biological findings?

In the model above, key-value pairs are never forgotten

BUT some of them become inaccessible as the "query" signals to activate them are lost.

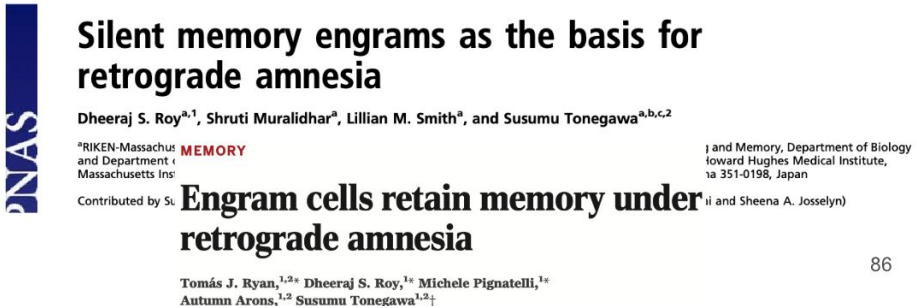
Forgetting = "Failure to remember"



"Unfolded brain"

Key-value pairs ~= "Memory engrams"

→ "Silent engram"?



PNAS

Silent memory engrams as the basis for retrograde amnesia

Dheeraj S. Roy[a,1], Shruti Muralidhar[a], Lillian M. Smith[a], and Susumu Tonegawa[a,b,c,2]

[a]RIKEN-Massachus and Department Massachusetts Ins

MEMORY

and Memory, Department of Biology oward Hughes Medical Institute, a 351-0198, Japan

Contributed by S ii and Sheena A. Josselyn)

Engram cells retain memory under retrograde amnesia

Tomás J. Ryan,[1,2*] Dheeraj S. Roy,[1*] Michele Pignatelli,[1*] Autumn Arons,[1,2] Susumu Tonegawa[1,2†]

[Slide from Katzuki Irie]

Pointers to other insightful dualities

Generative diffusion models ~ Modern hopfield networks

In search of dispersed memories: Generative diffusion models are associative memory networks

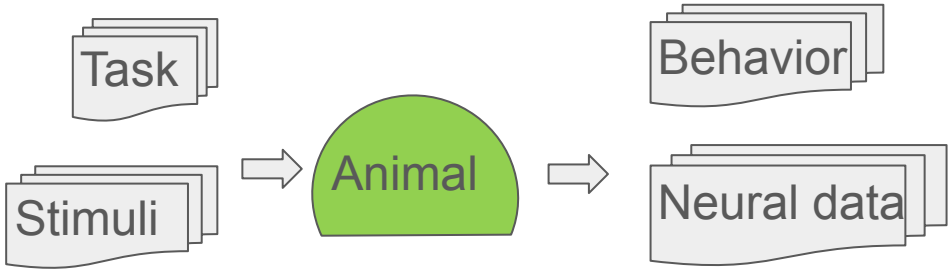
Luca Ambrogioni
Radboud University, Donders Institute for Brain, Cognition and Behaviour
luca.ambrogioni@donders.ru.nl

Variational autoencoders ~ Generative diffusion models

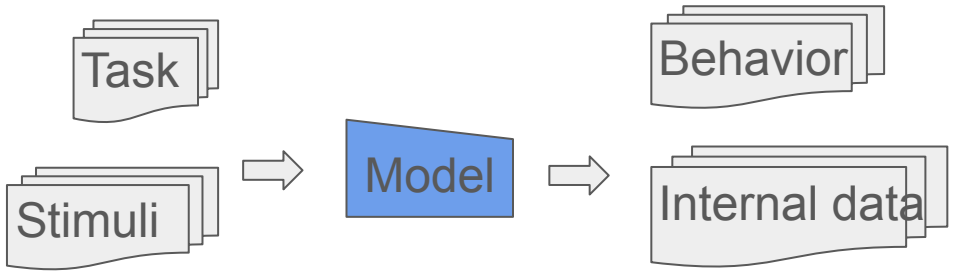
A Variational Perspective on Diffusion-Based Generative Models and Score Matching

Chin-Wei Huang, Jae Hyun Lim, Aaron Courville
University of Montreal & Mila
{chin-wei.huang, jae.hyun.lim, aaron.courville}@umontreal.ca

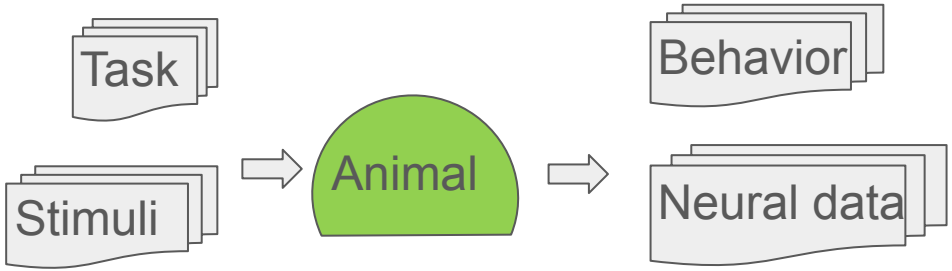
Systems Neuroscience



Task-optimized models

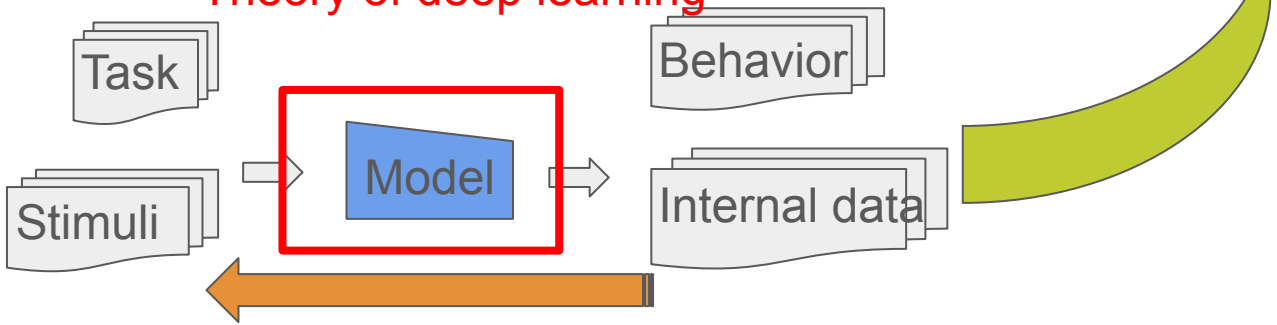


Systems Neuroscience

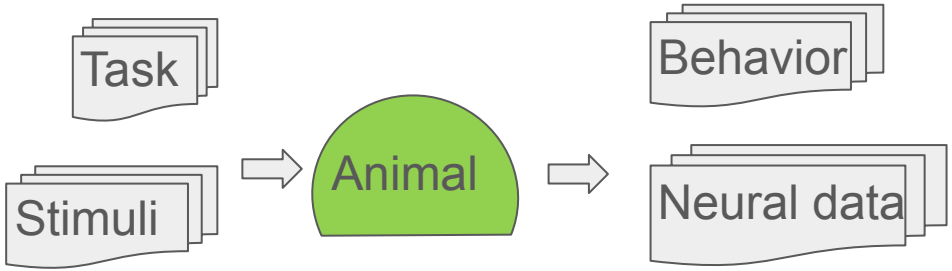


Task-optimized models

Theory of deep learning



Systems Neuroscience

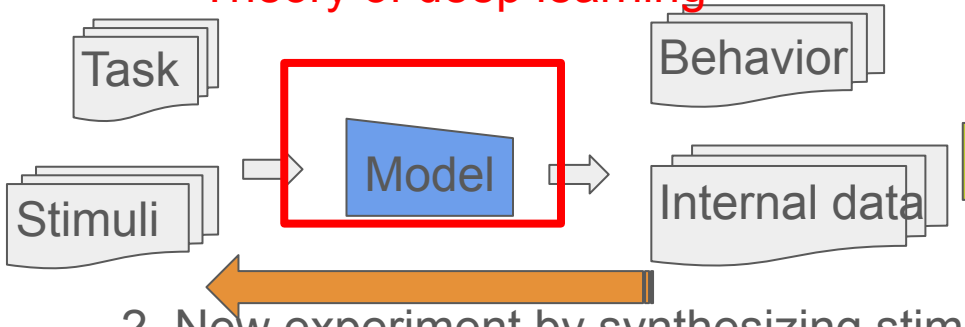


3. Learning core computations from models

Task-optimized models

Theory of deep learning

1. Model selection by mapping/scoring



2. New experiment by synthesizing stimuli