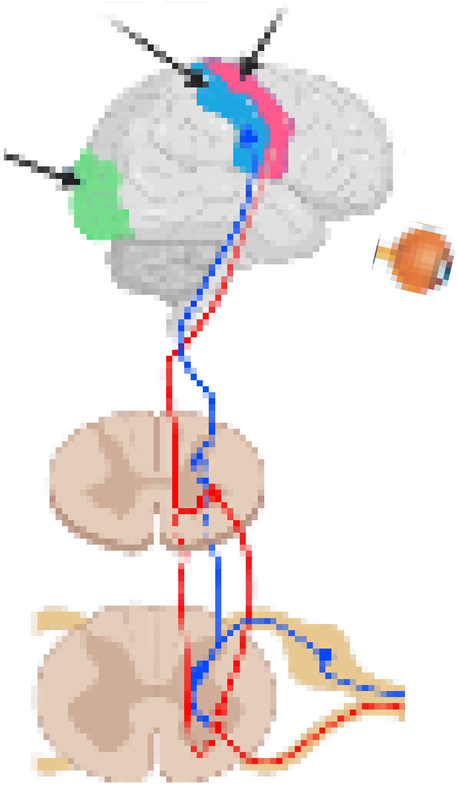
EPFL

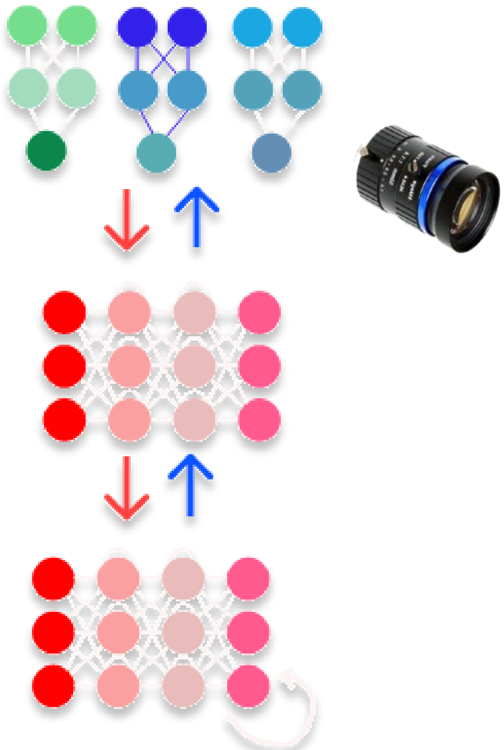
NX-414: Brain-like computation and intelligence

Martin Schrimpf

Lecture 6, March 26 2025

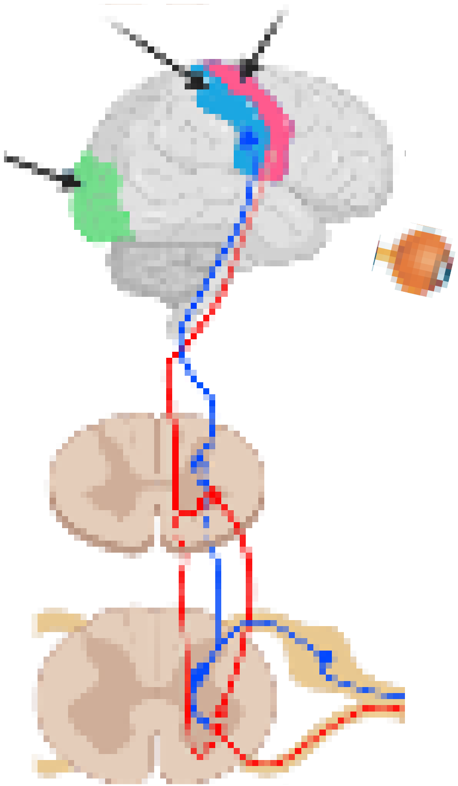
Biological Intelligence ⟷ Artificial Intelligence





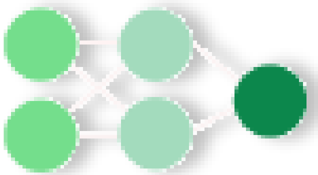
Hausmann & Marin-Vargas et al., 2021

Normative frameworks



Information theoretic

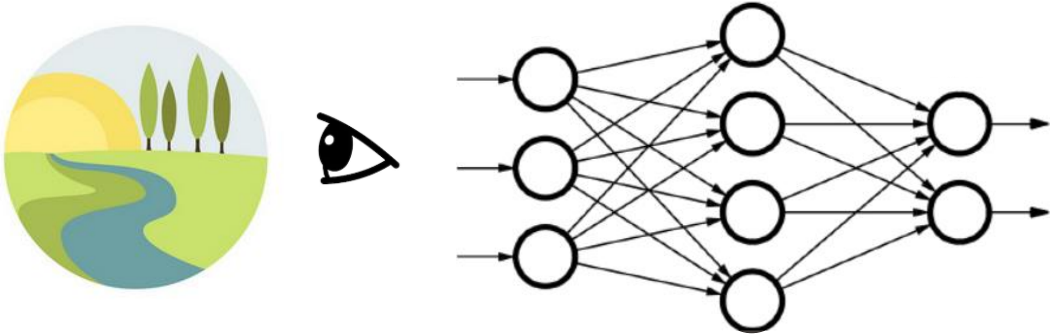
e.g. sparse coding,
redundancy reduction,
mutual information …

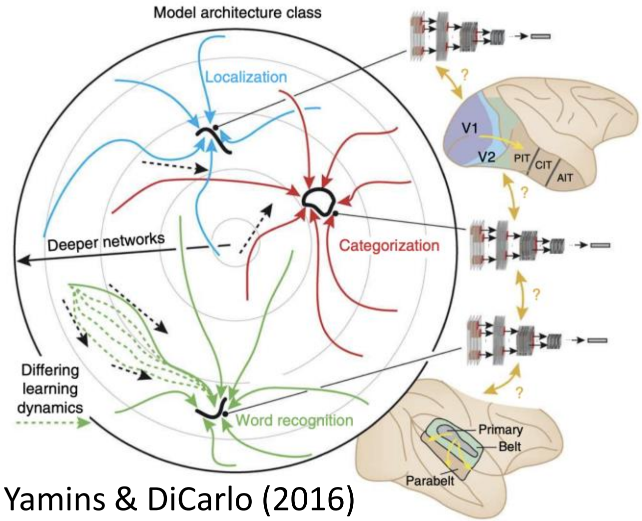


Utilitarian

e.g. **recognize objects**,
chase prey, navigate …

Using deep neural networks as goal-driven models of a system





Yamins & DiCarlo (2016)

Vision: object recognition.
Yamins & Hong et al. (2014), Schrimpf & Kubilius et al. (2018)

Audition: speech recognition, speaker & sound identification. Kell et al. (2018)

Somatosentation: shape recognition.
Zhuang et al. (2017)

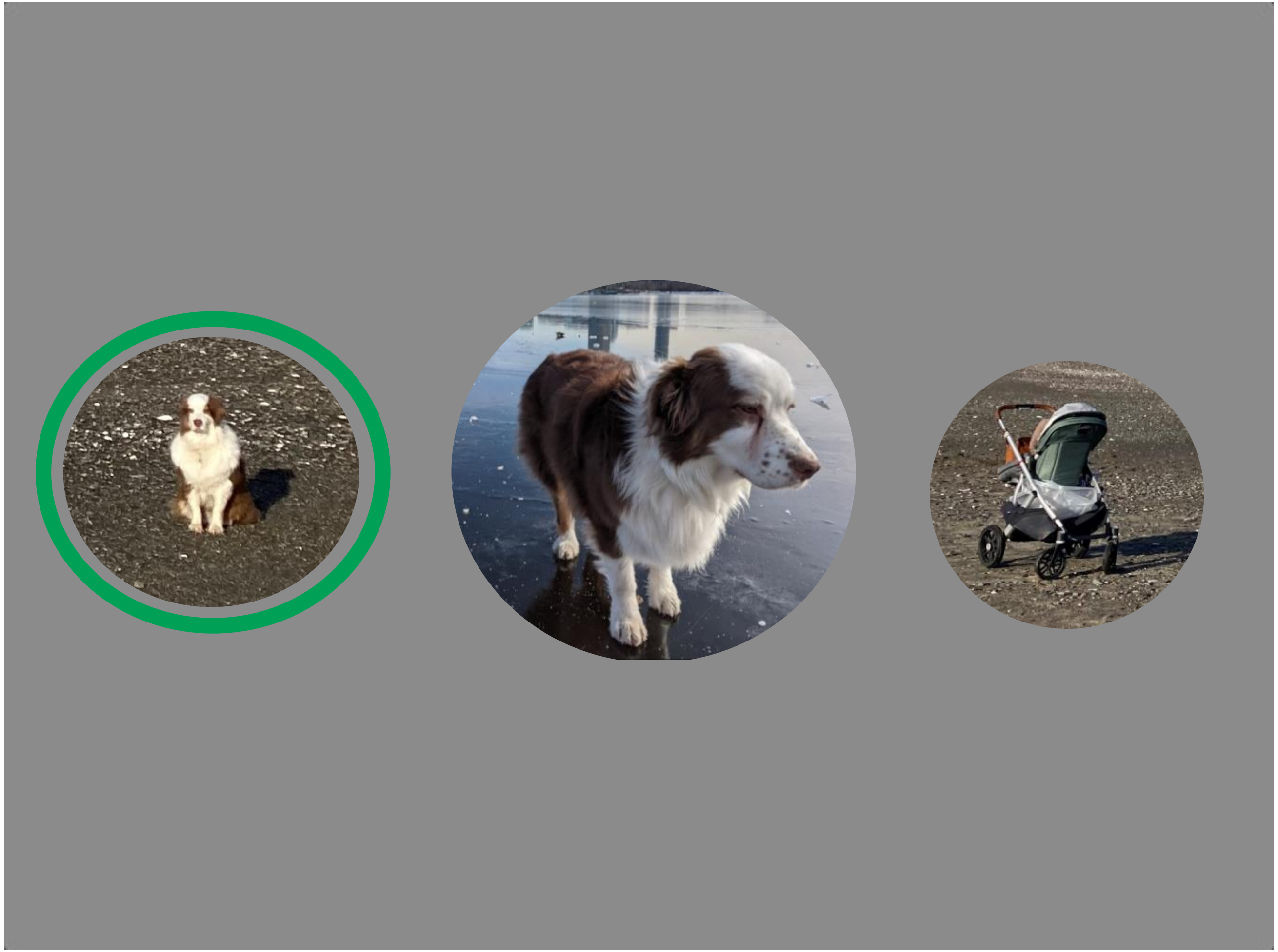
Language: next-word prediction.
Schrimpf et al. (2021)

Decision making: context-dependent choice. Mante & Sussilo et al. (2013)

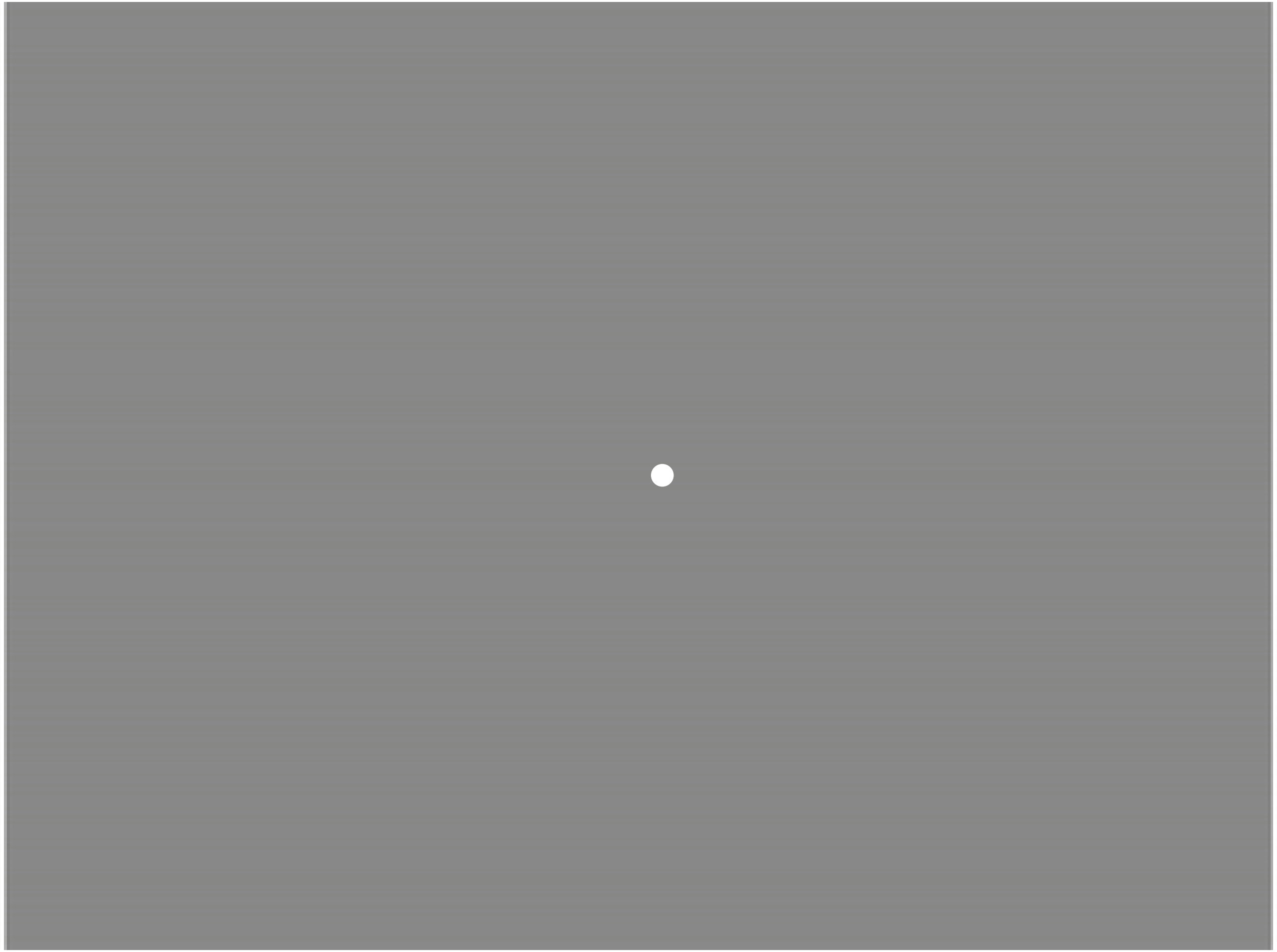
Proprioception: action recognition.
Sandbrink et al. (2023)

EPFL

Behavioral experiment



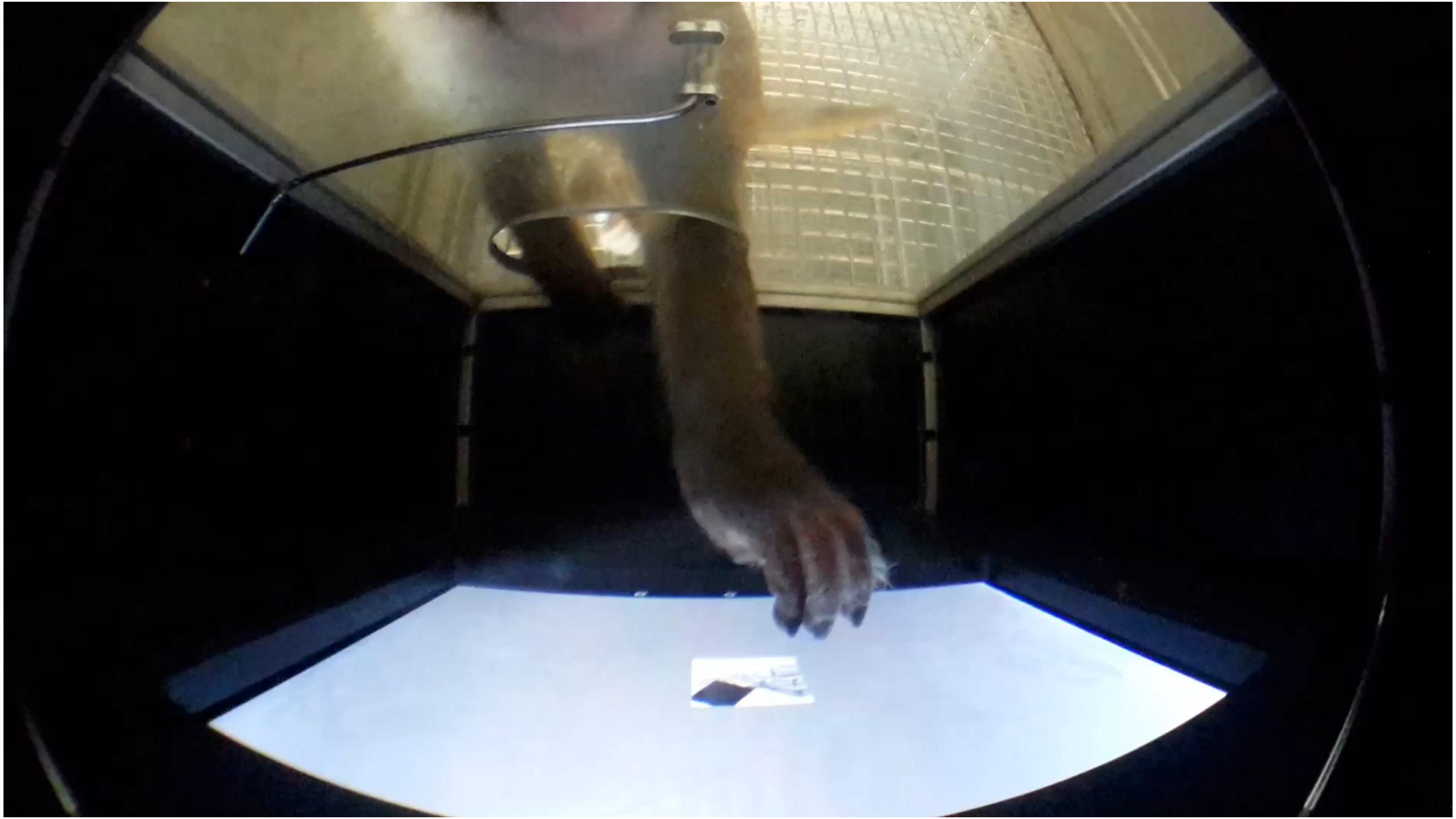
Behavioral experiment



Rajalingham, Issa*, et al. (JNeuro 2018)*

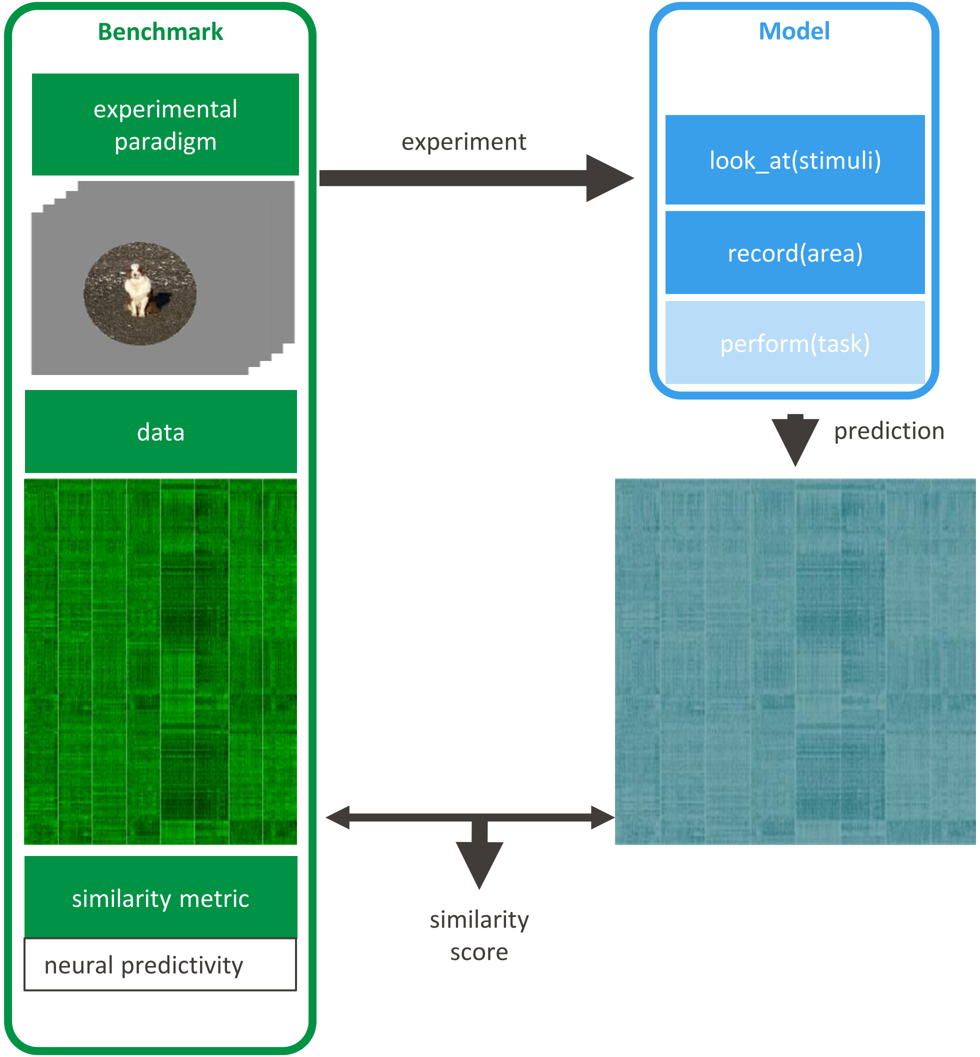
EPFL

Neural data



video courtesy of Kailyn Schmidt, MIT

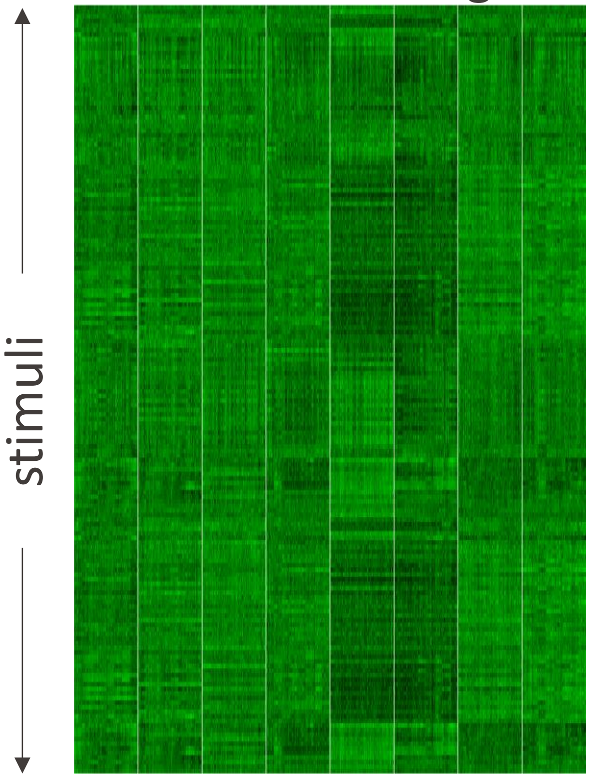
Neural benchmark



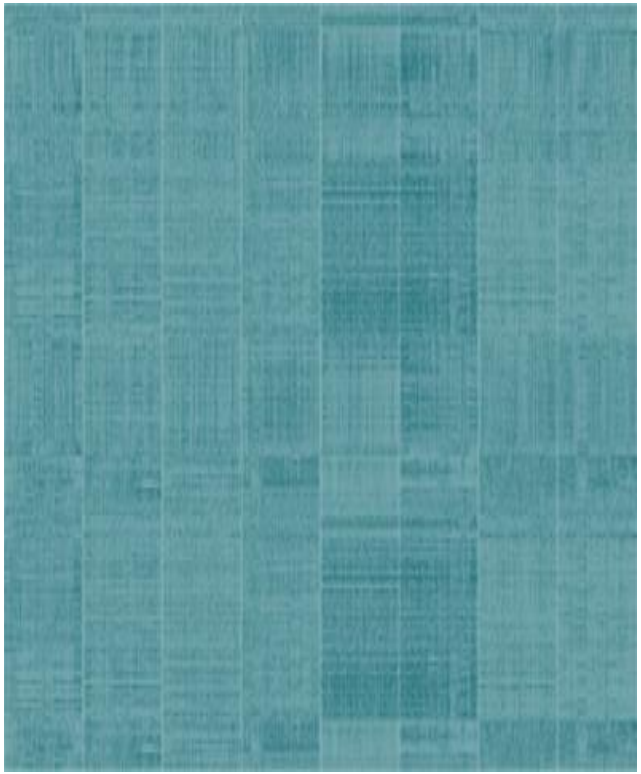
Neural alignment = alignment between stimulus-matched recordings

Neural benchmarks

Brain recordings



Model units



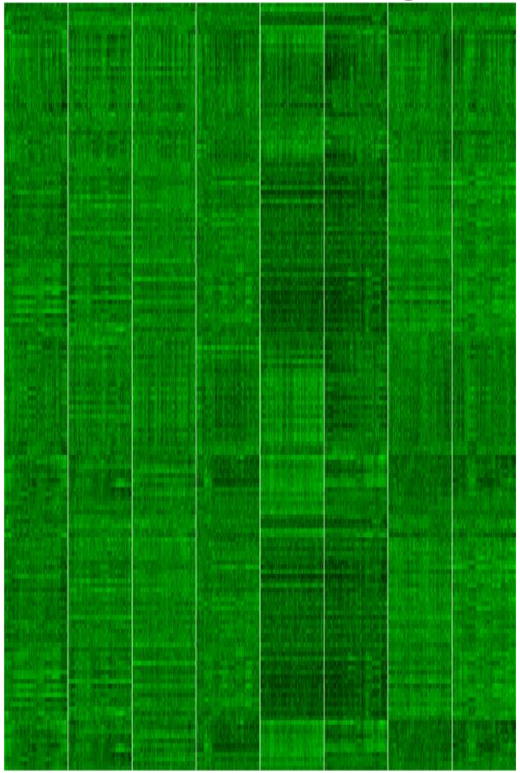
Neural benchmarks

similarity metric

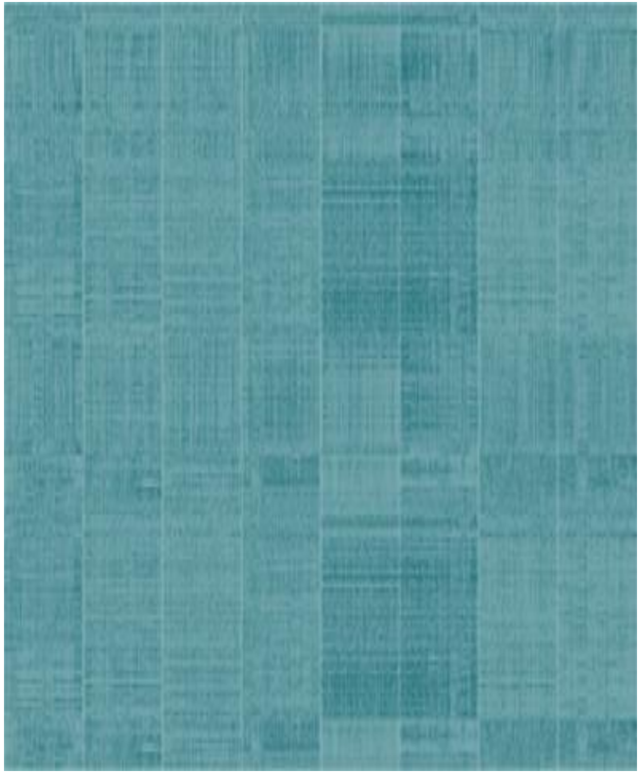
neural predictivity

Brain recordings

stimuli



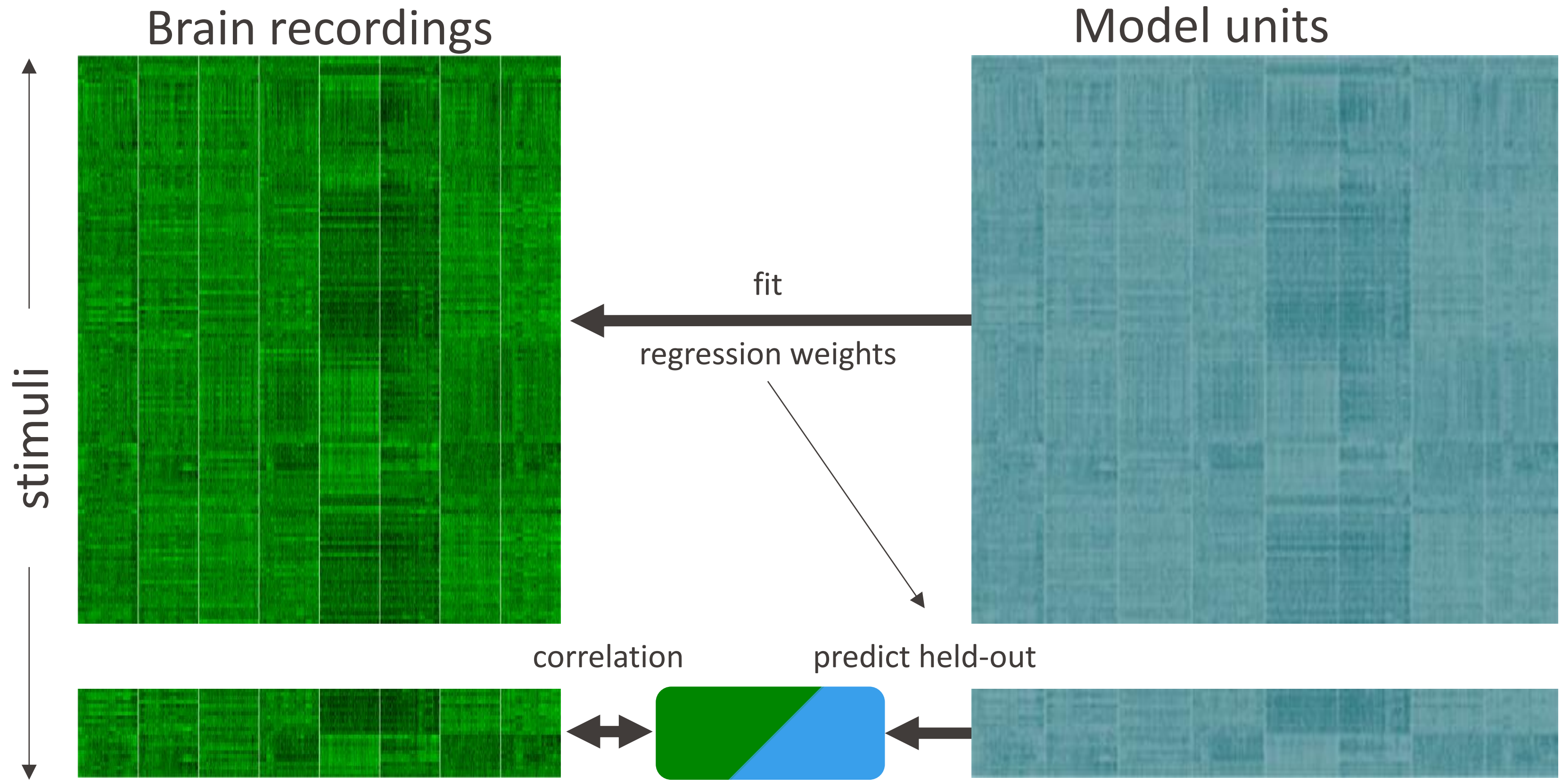
Model units



Neural benchmarks

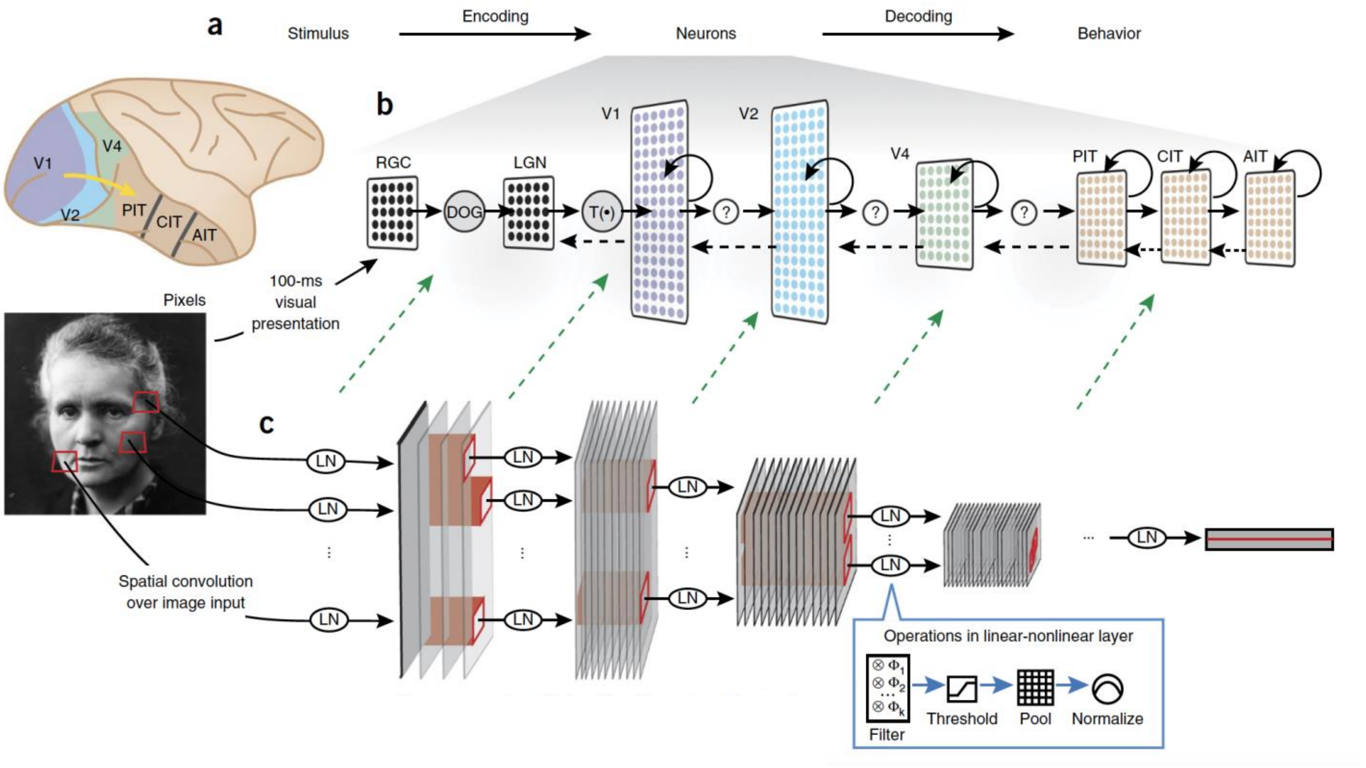
similarity metric

neural predictivity



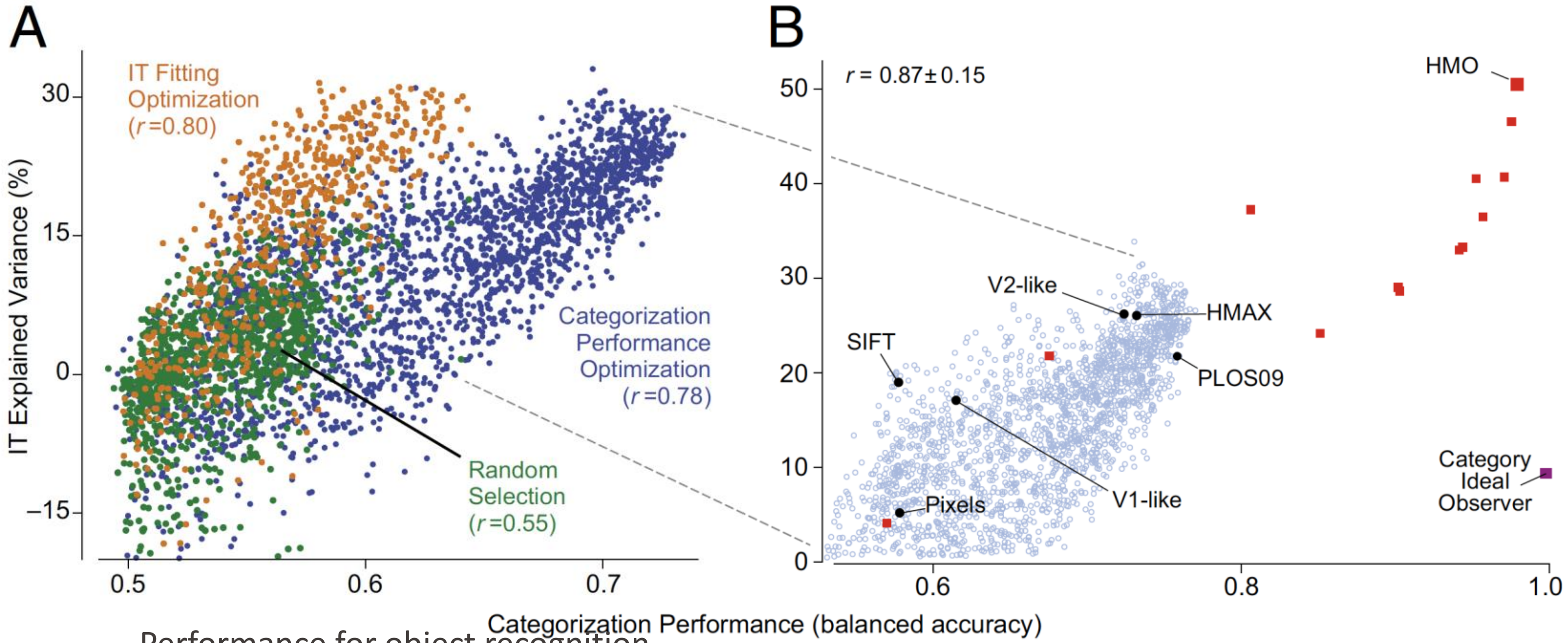
Yamins*, Hong*, et al. 2014   Schrimpf*, Kubilius*, et al. 2018

Model building



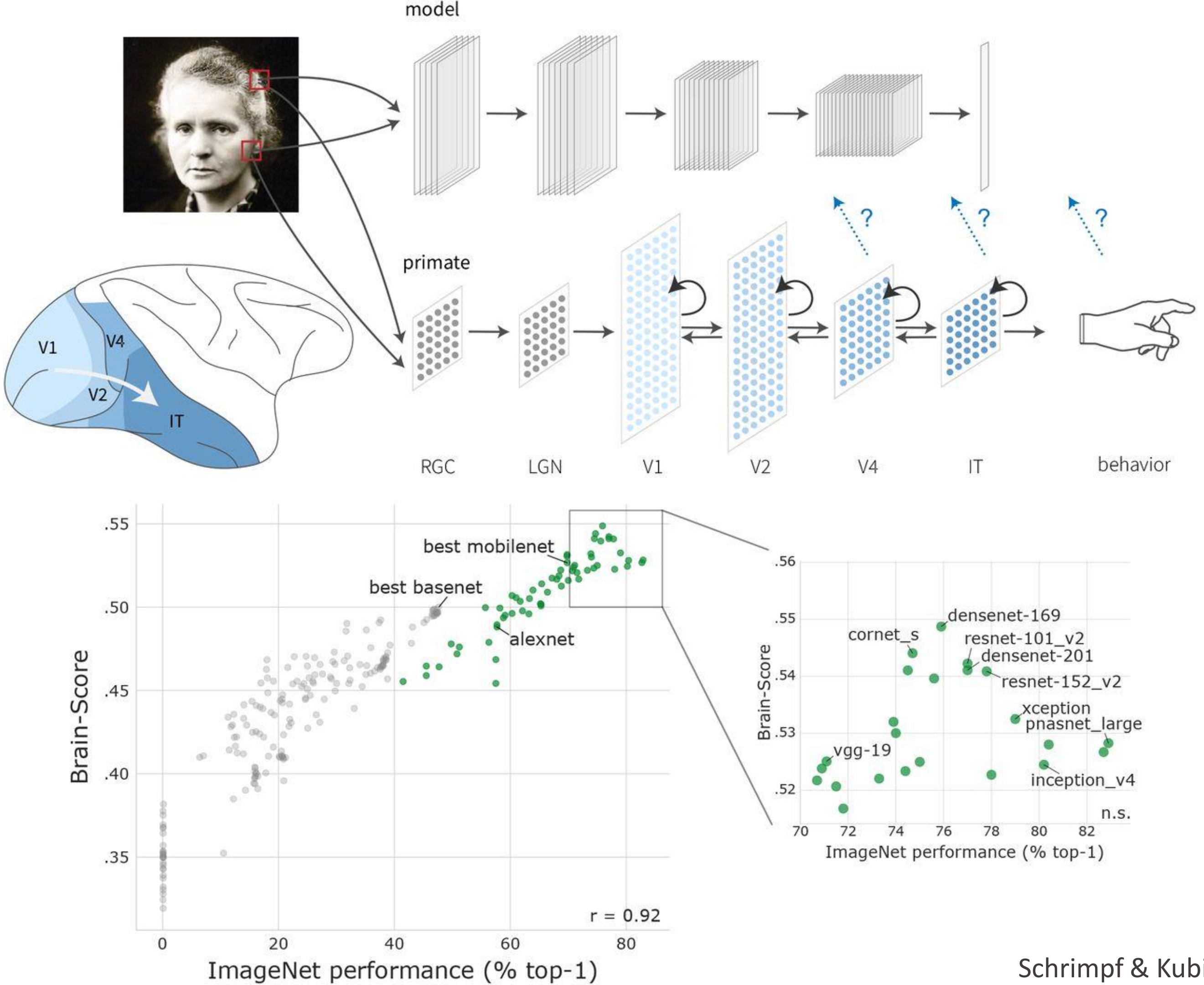
Yamins and DiCarlo, Nat Neuro 2016

Large scale architecture search and model comparison



Performance for object recognition

Yamins & Hong et al., PNAS 2014



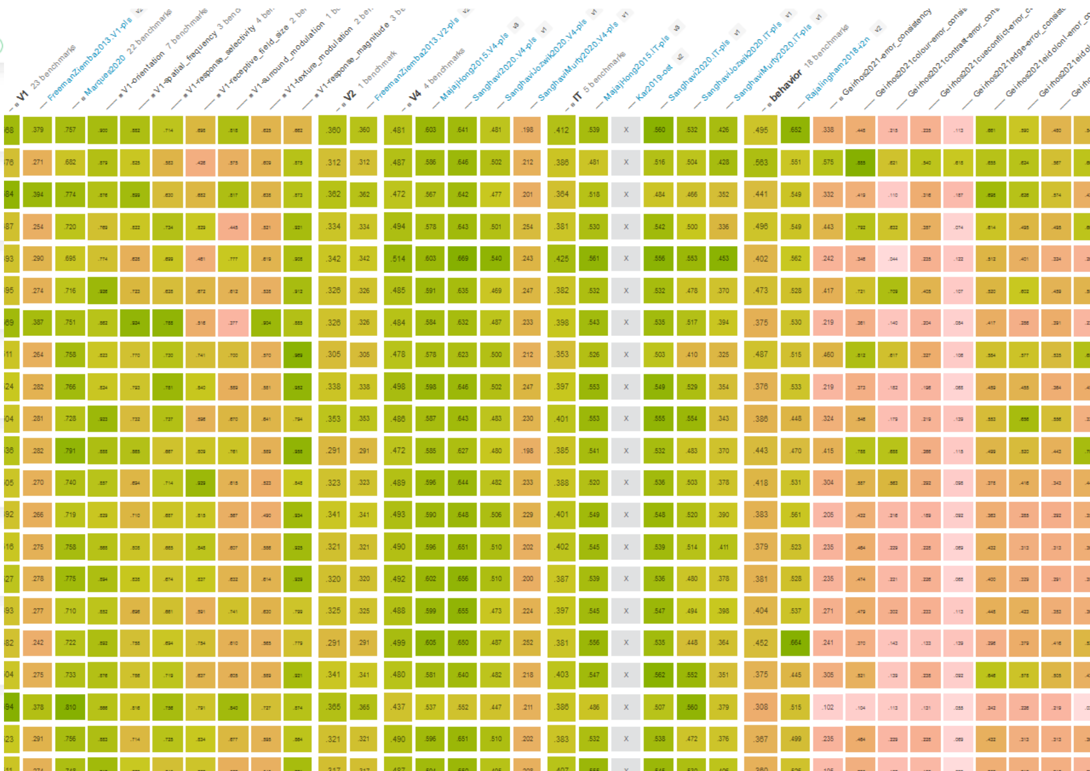
Schrimpf & Kubilius et al. 2018, 2020

Brain-Score 100+ brain & behavior benchmarks, 300+ models

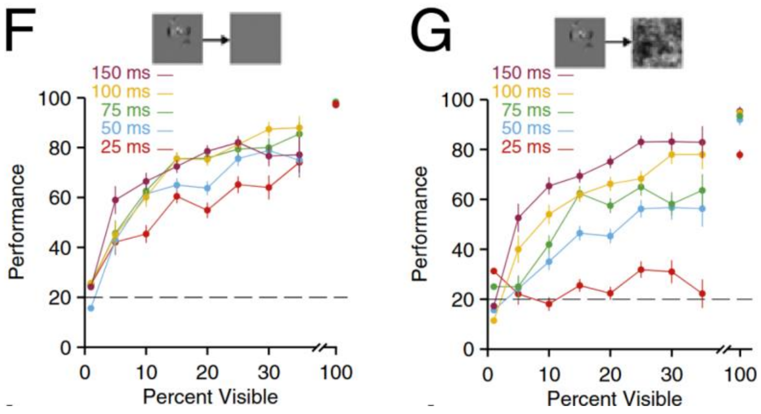
e.g. neural predictions for different image sets, distributional alignments such as spatial frequency, behavioral generalization, …

www.Brain-Score.org

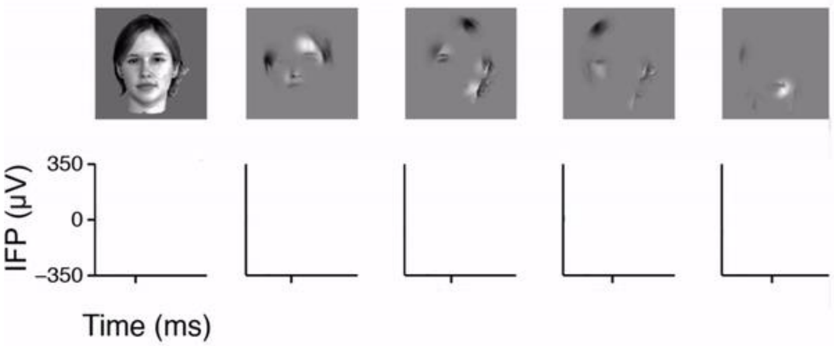
Rank	Model	average 81	neural 38	behavior 43
1	convnext_large_mlp:clip_laio	.467	.378	.555
2	convnext_xlarge:fb_in22k_ft_	.449	.338	.560
3	vit_base_patch16_clip_224:o|	.445	.343	.548
4	vit_large_patch14_clip_224:la	.445	.332	.559
5	vit_large_patch14_clip_224:o	.443	.341	.545
6	vit_base_patch16_clip_224:o|	.442	.352	.532
7	vit_relpos_base_patch16_cls|	.437	.381	.494
8	cvt_cvt-w24-384-in22k_finetuned-in1k_4	.430	.327	.533



Recurrent processing in the visual system

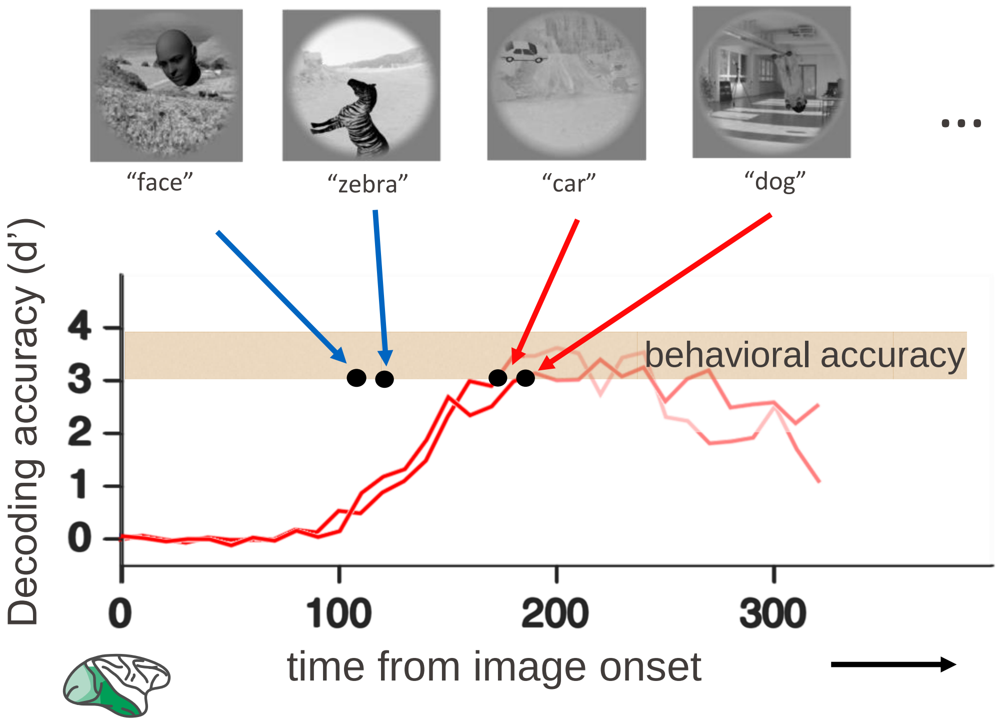


Electrode in left fusiform gyrus (face-selective)



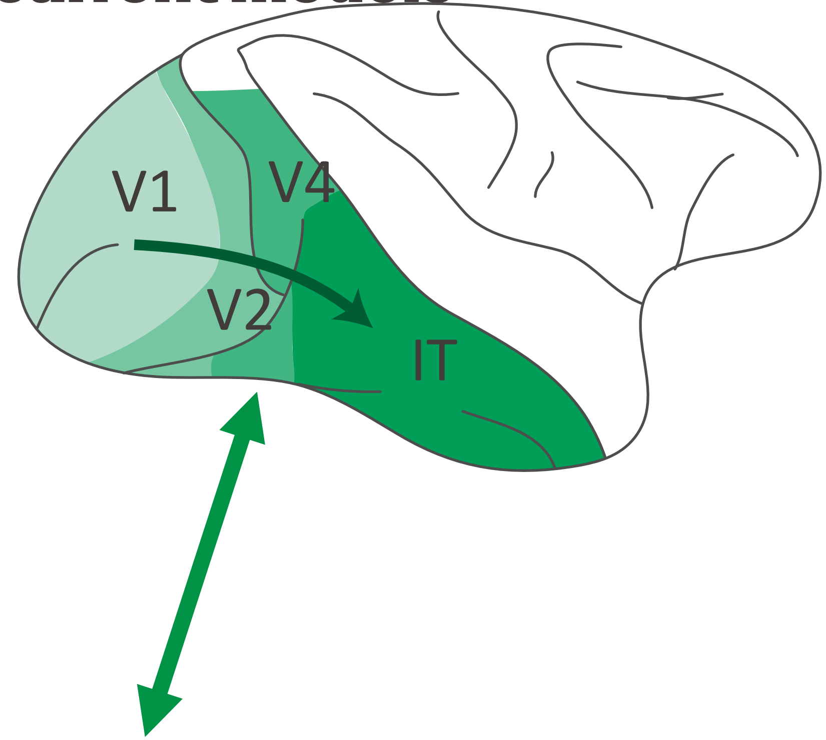
Tang & Schrimpf & Lotter et al. 2018

- Control images are solved quickly
- Challenge images require more processing



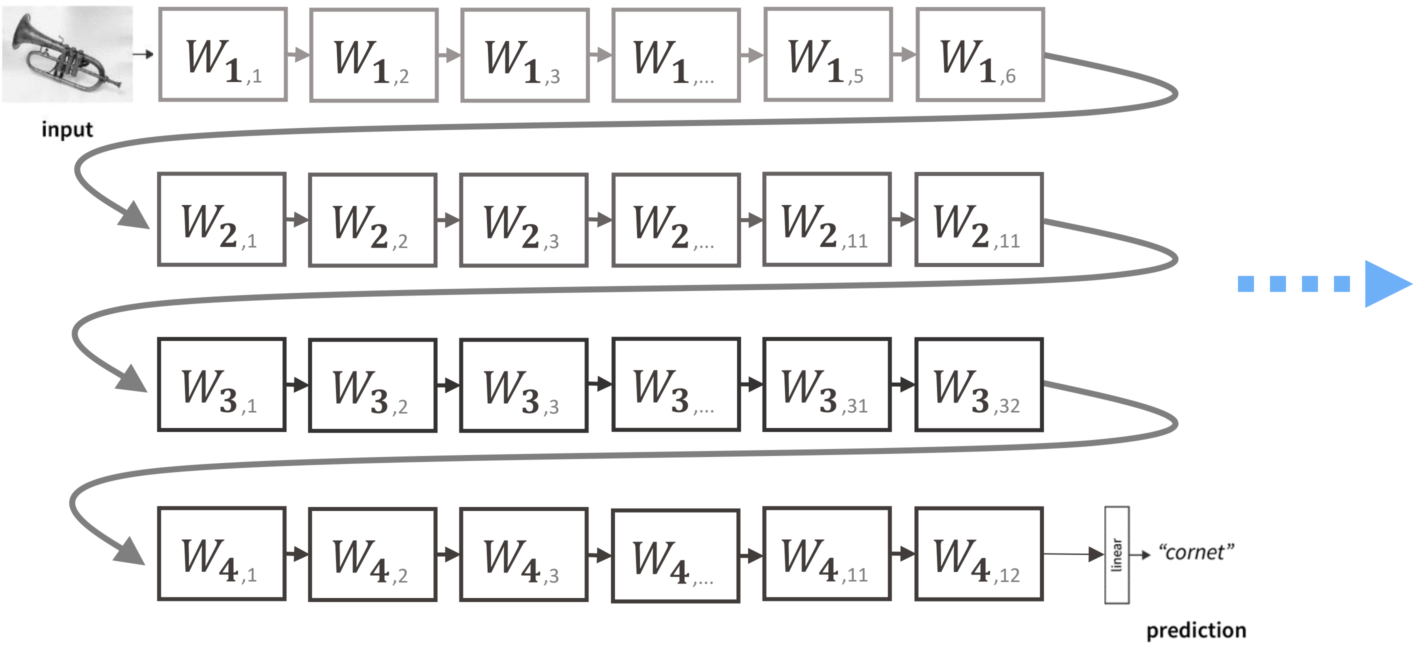
Kar et al. 2019

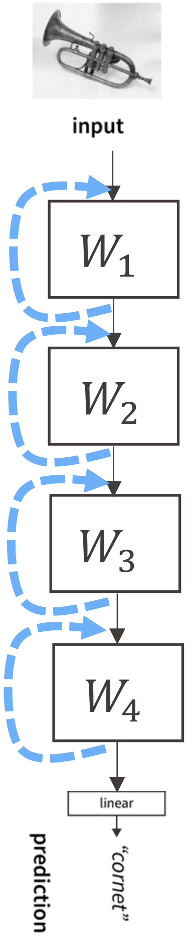
Modeling recurrence: transform feed-forward networks into recurrent models



e.g. ResNet-101

Transform feed-forward networks into recurrent models

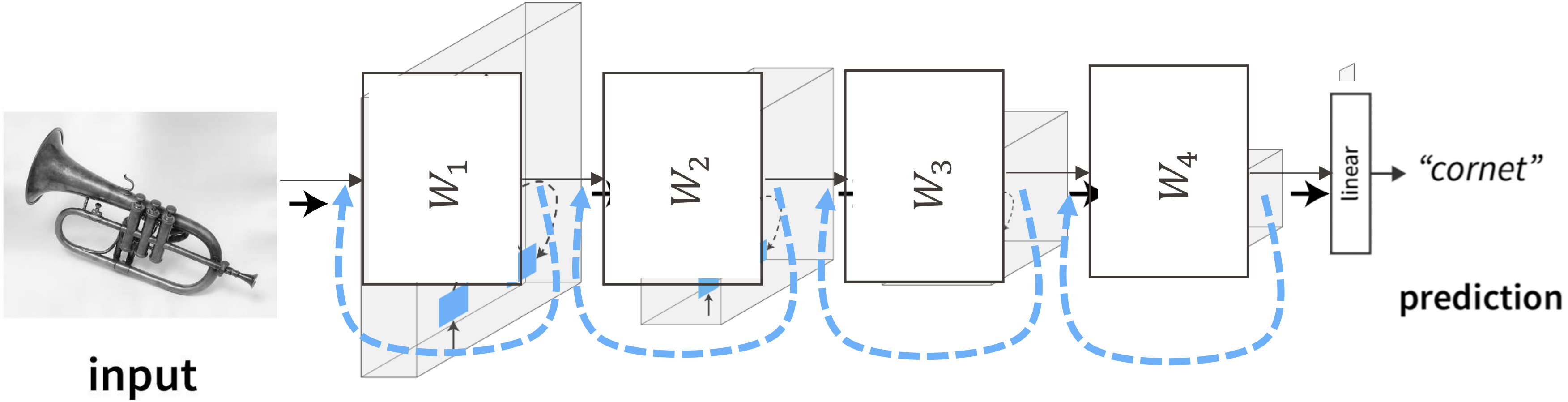




e.g. He, Zhang, Ren, Sun (CVPR 2016)
Huang, Liu, van der Maaten, Weinberger (CVPR 2017)

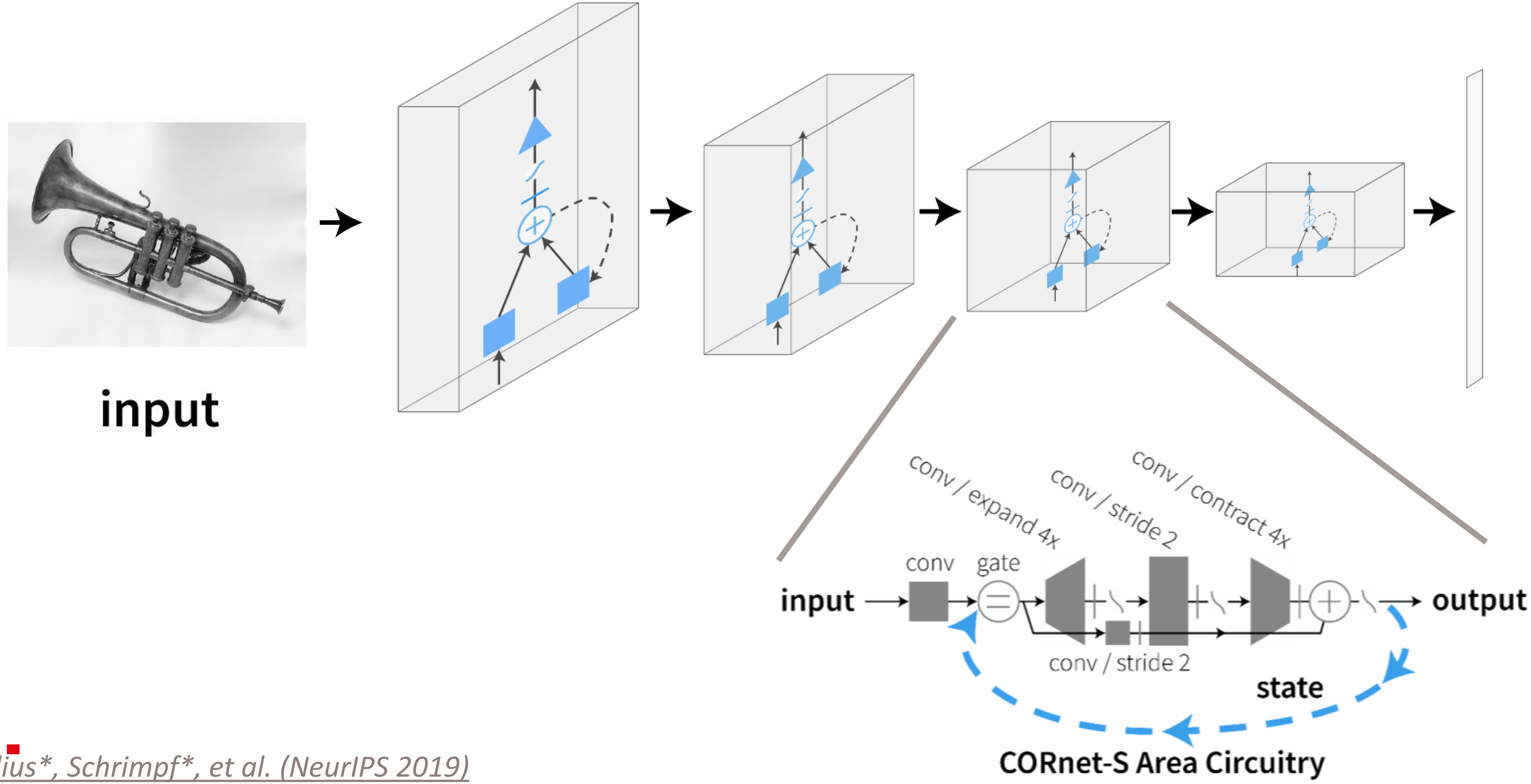
Liang & Hu (CVPR 2015) Liao & Poggio (arXiv 2016) Tang*, Schrimpf*, Lotter* et al. (PNAS 2018) Nayebi*, Bear*, Kubilius* et al. (NIPS 2018)

Modeling recurrence: transform feed-forward networks into recurrent models: CORnet



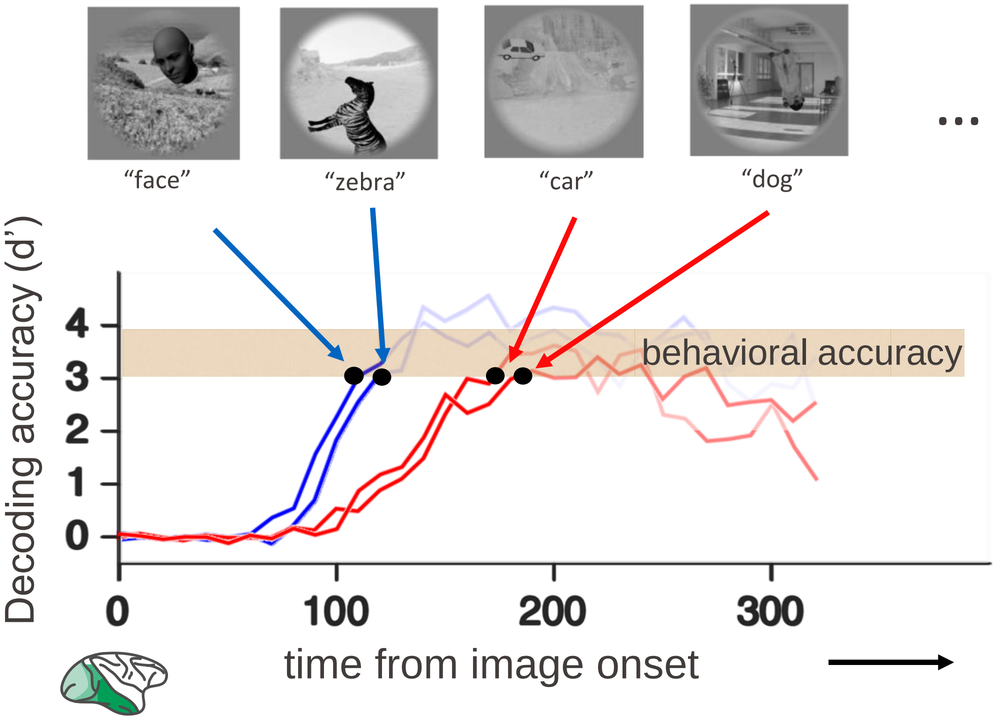
Kubilius*, Schrimpf*, et al. (NeurIPS 2019)

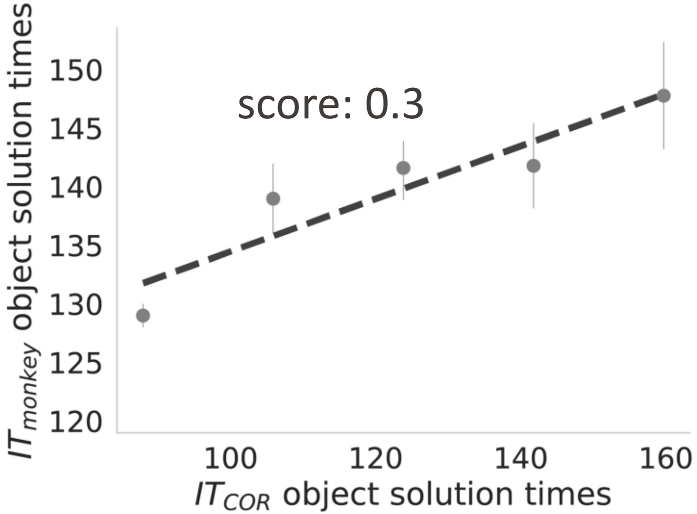
Recurrent CORnet model: compact architecture via recurrence



Kubilius*, Schrimpf*, et al. (NeurIPS 2019)

Recurrent model predicts temporal dynamics in IT





- Unlike feedforward models, CORnet-S can **predict neural responses over time**.
- i.e., when the **brain's IT is fast** to process images, **CORnet's IT-layer is also fast**

Kar, Kubilius, Schmidt, Issa, DiCarlo (Nature Neuroscience 2019)

Kubilius*, Schrimpf*, et al. (NeurIPS 2019)

Unsupervised learning with a contrastive loss

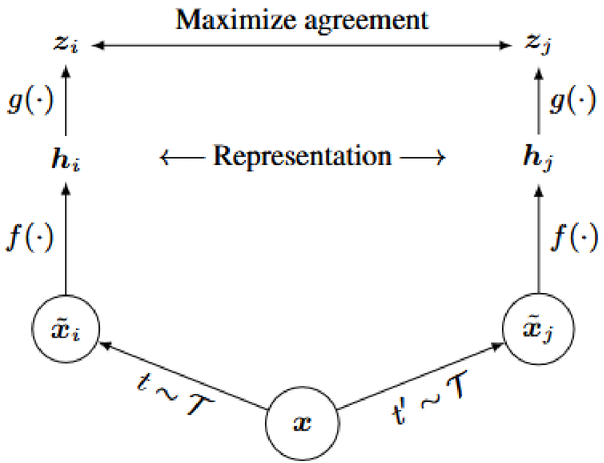
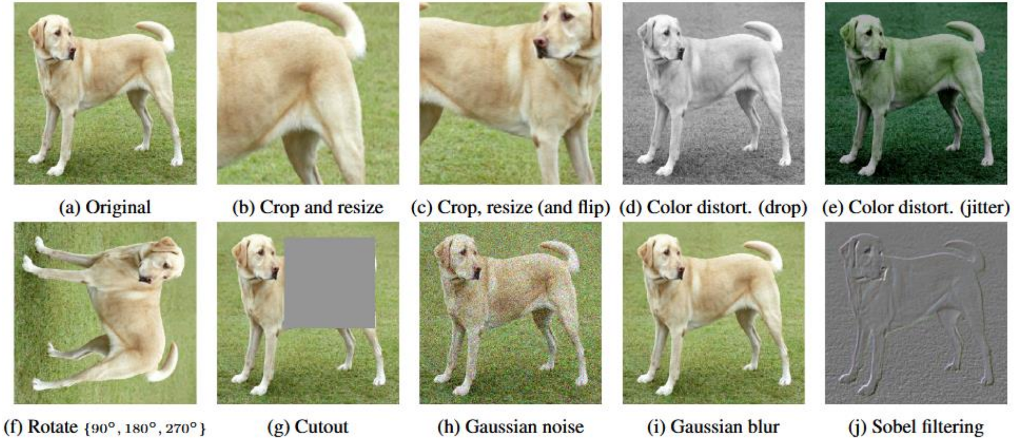
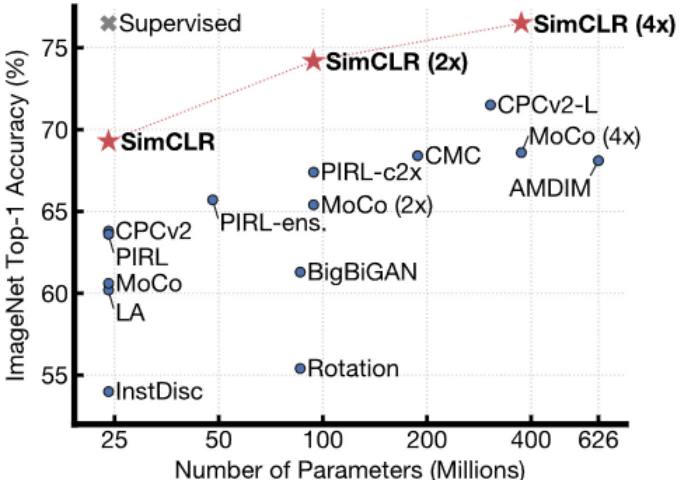


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation h for downstream tasks.

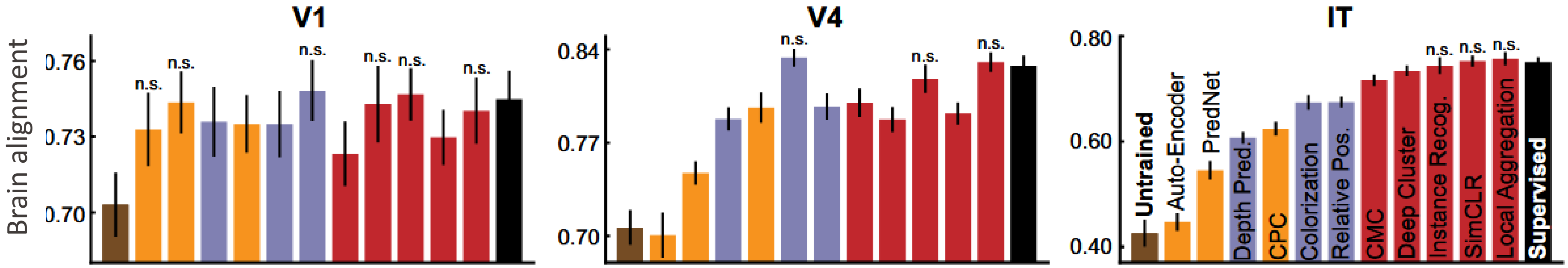




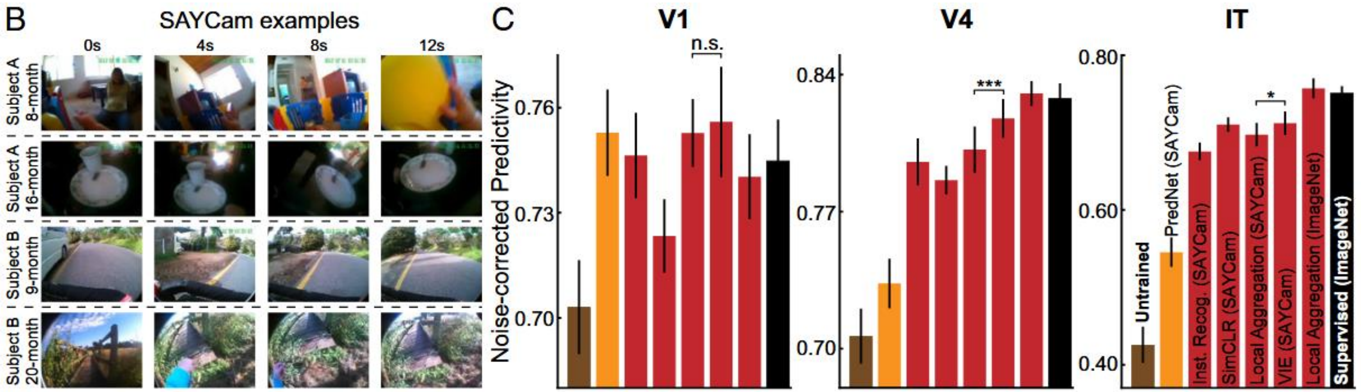
- Unsupervised approaches such as SimCLR encourage similar representations for similar inputs
- Performance rivals supervised learning

- Chen et al. 2020

Unsupervised models also explain visual brain activity

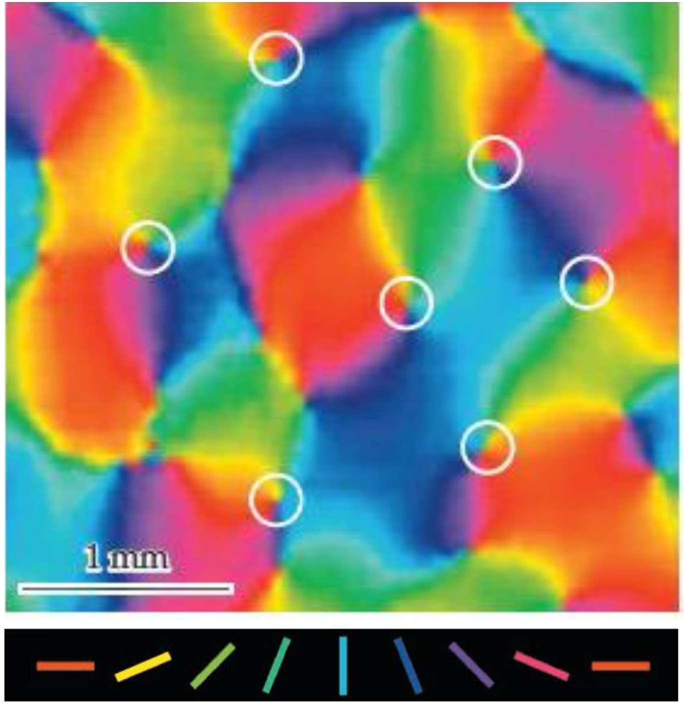


- When trained on regular computer vision datasets (top) or developmental data streams (below, SAYCam), unsupervised models develop brain-like visual representations

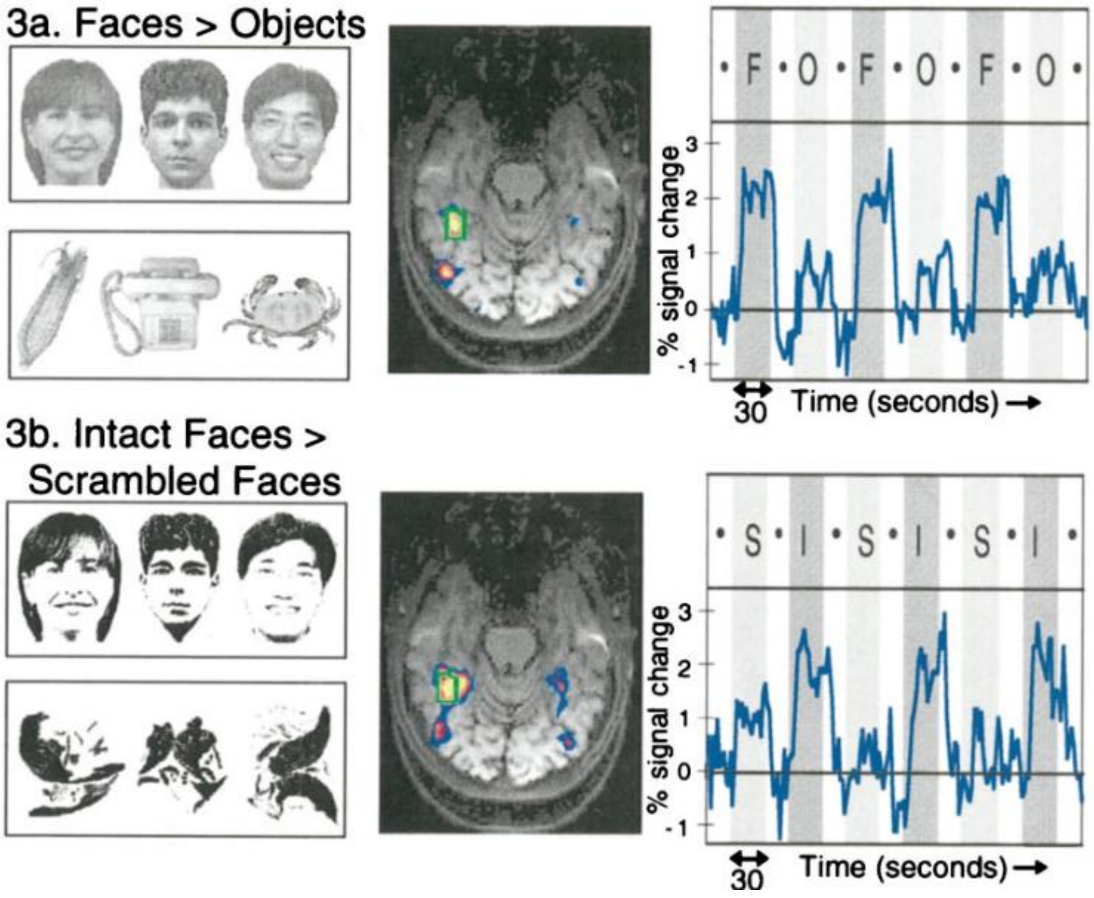


- Zhuang et al. 2021

Neurons in cortex are topographically organized



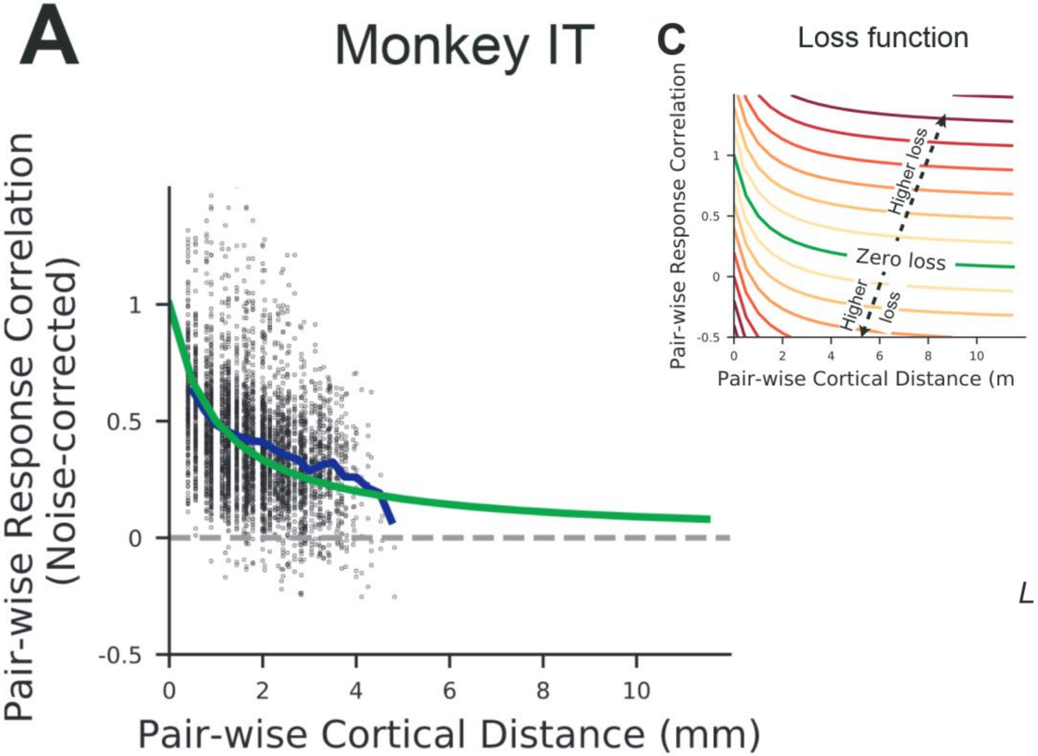
Purves, chapter 11

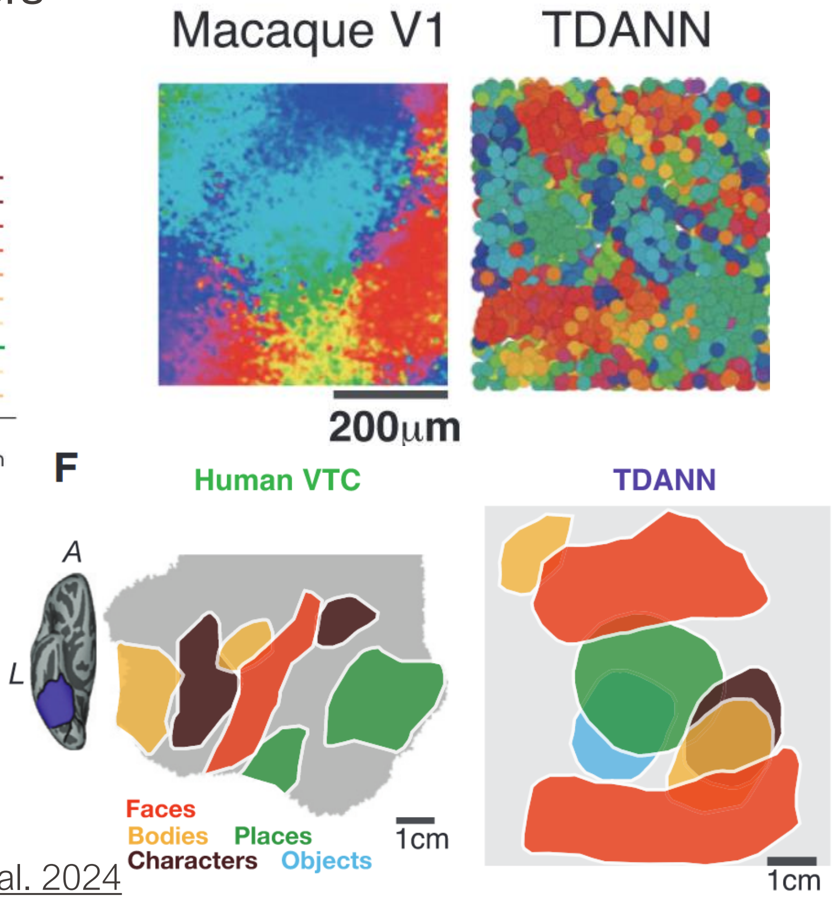


Kanwisher et al. 1997

Modeling spatial smoothness with a topographic loss

- A spatial loss term leads to brain-like clusters along the visual ventral stream (V1 to IT)

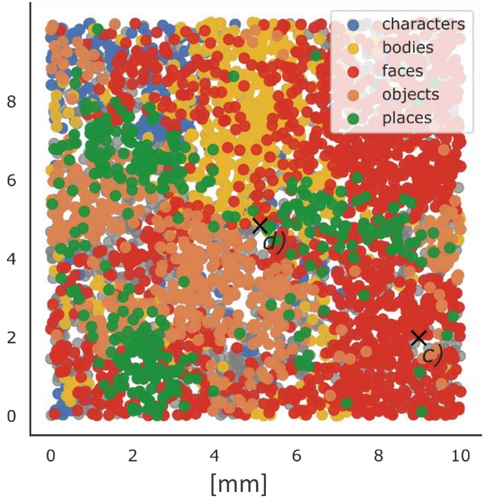


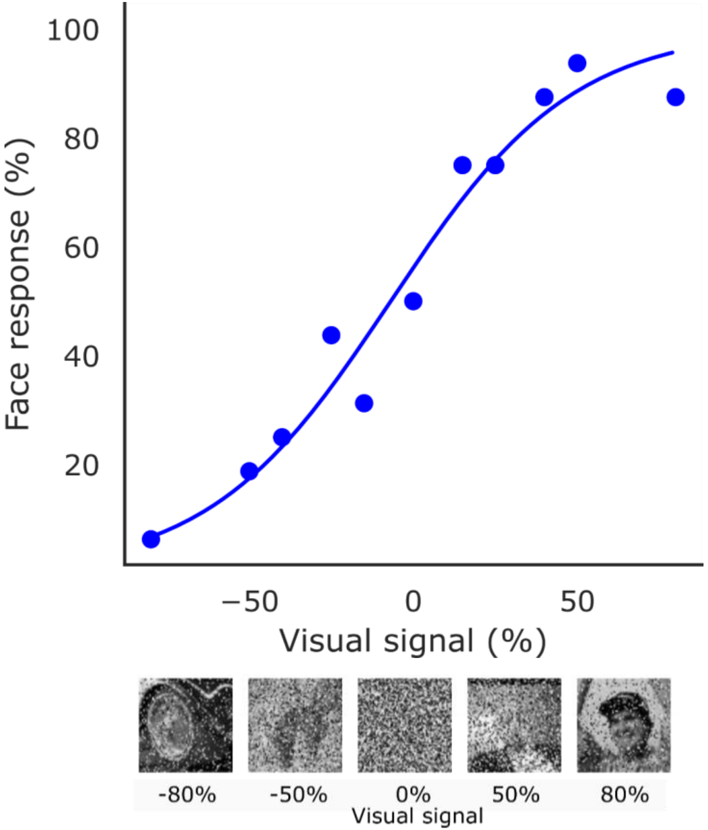


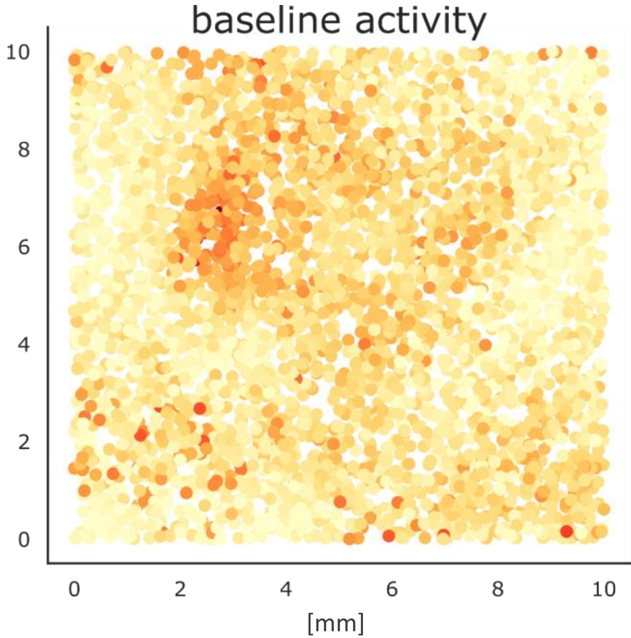
- [Lee et al. 2020](#)

[Margalit et al. 2024](#)

Topographic models enable the modeling of causal interventions, e.g. micro-stimulation

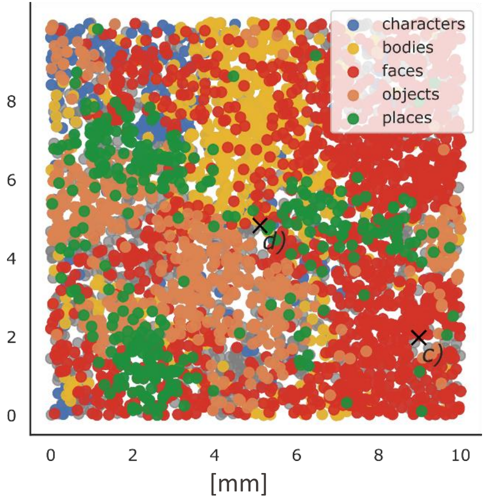


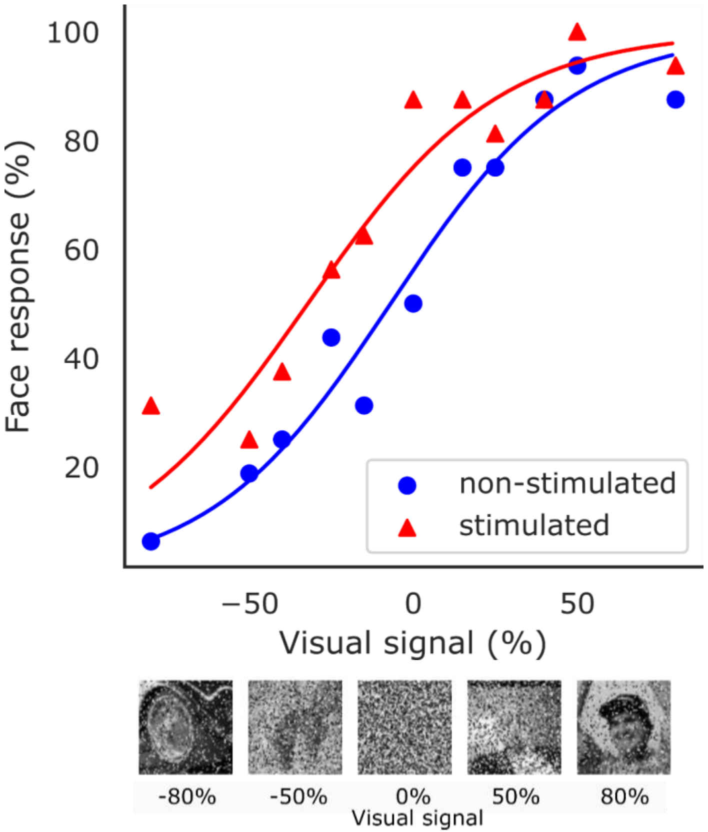


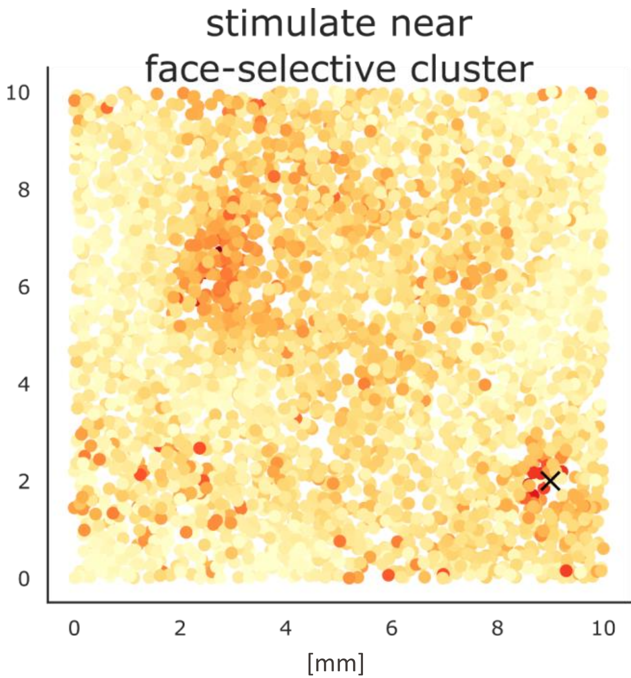


Schrimpf et al. 2024

Topographic models enable the modeling of causal interventions, e.g. micro-stimulation

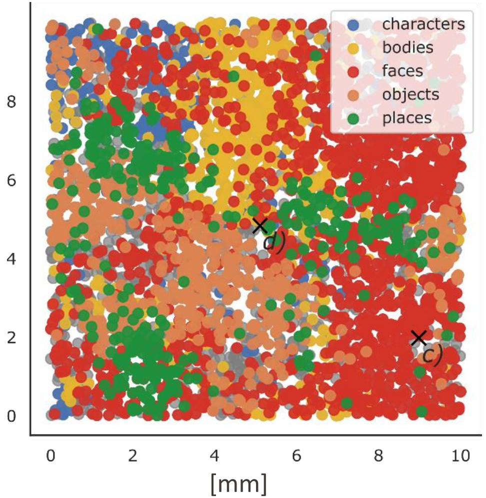




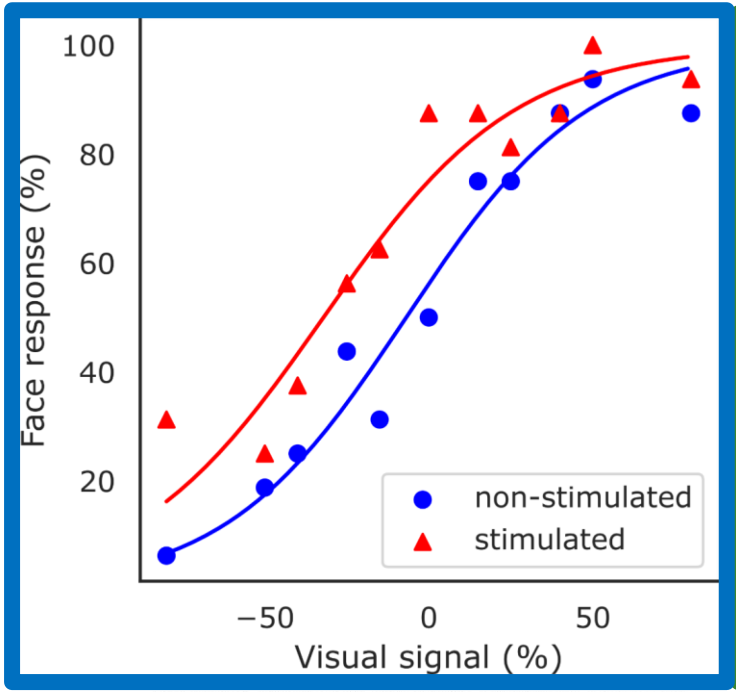


Schrimpf et al. 2024

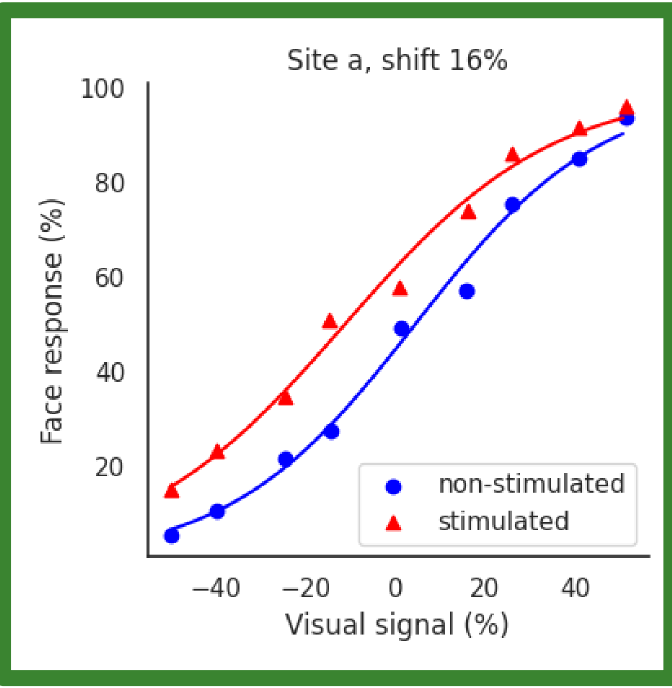
Topographic models enable the modeling of causal interventions, e.g. micro-stimulation

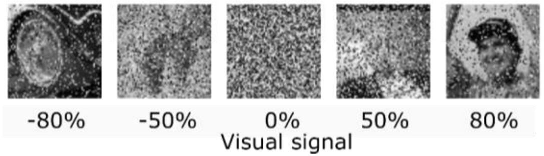


MODEL PREDICTION



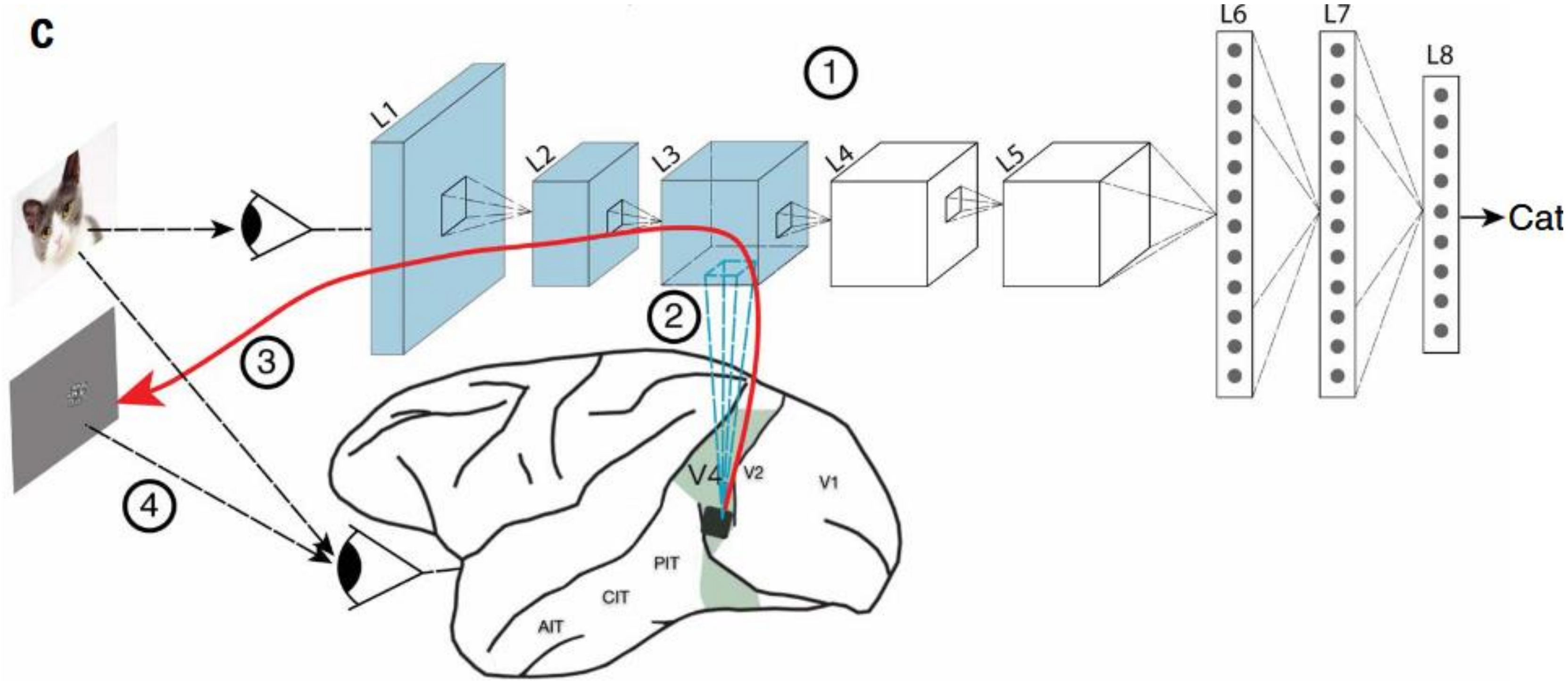
BIOLOGICAL DATA (Afraz et al. 2006)





Schrimpf et al. 2024

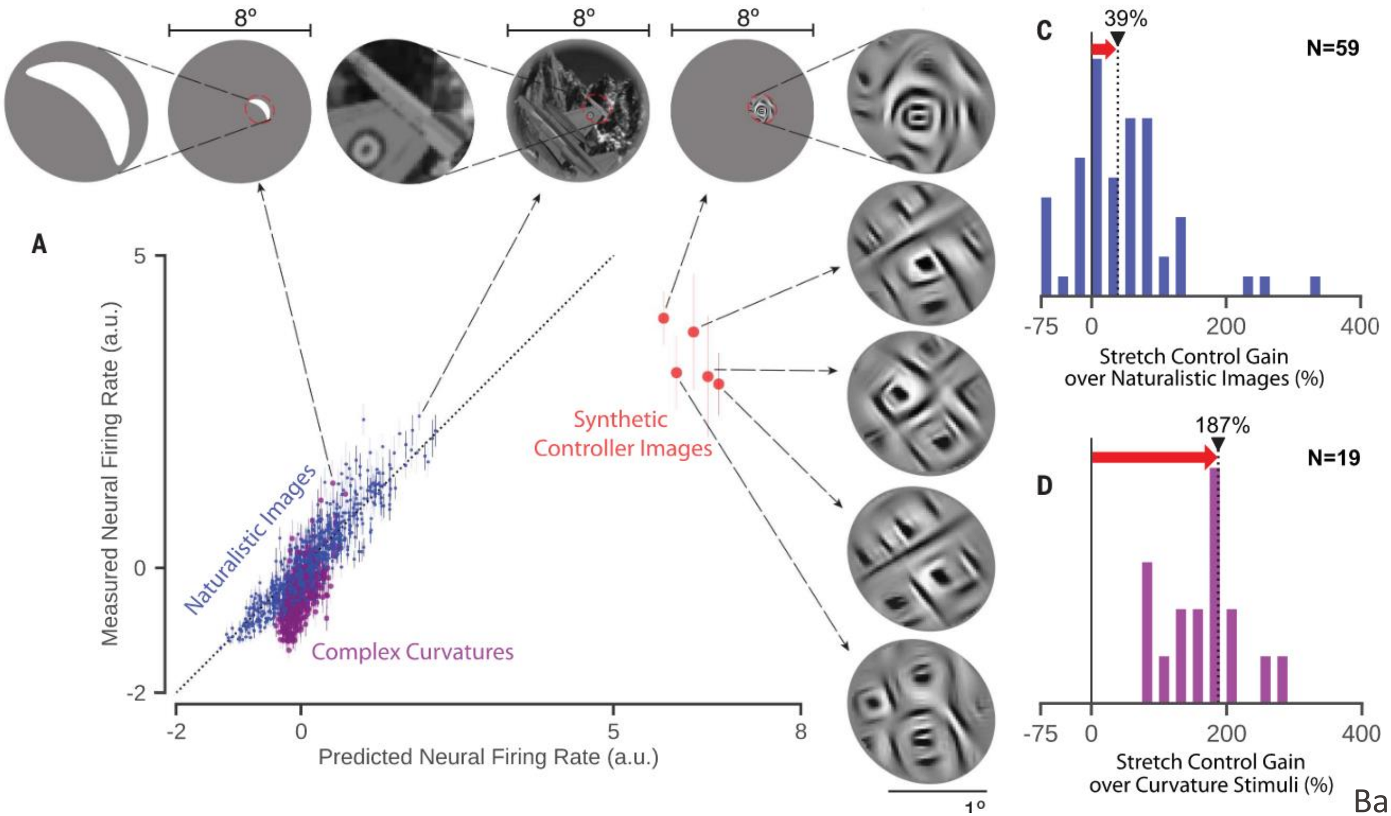
Synthesis of stimuli for neural population control



- Idea: model is fully differentiable, so we can set a desired target neural activity and update pixels in a way that they elicit the target state (according to model predictions)

Bashivan*, Kar*, DiCarlo 2019

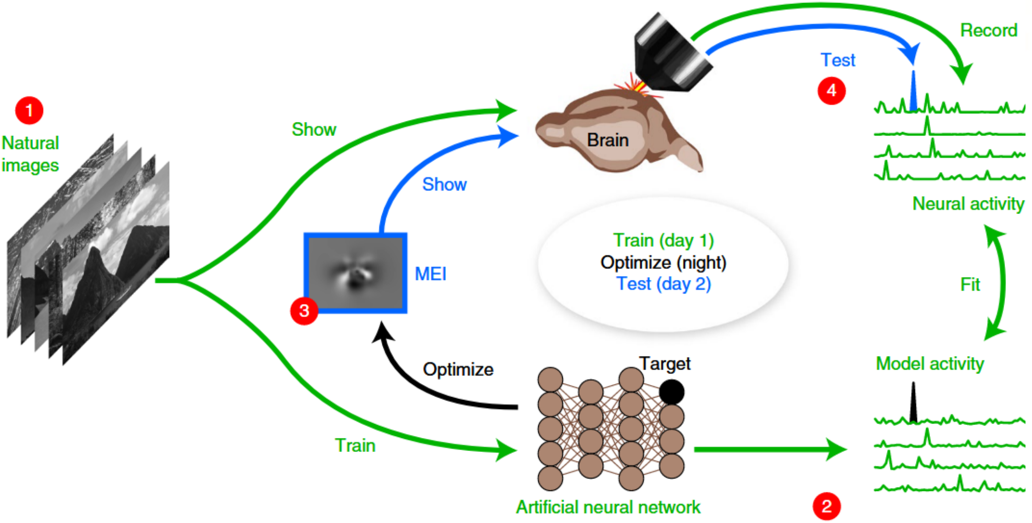
Model-guided synthesis non-invasively controls neural activity

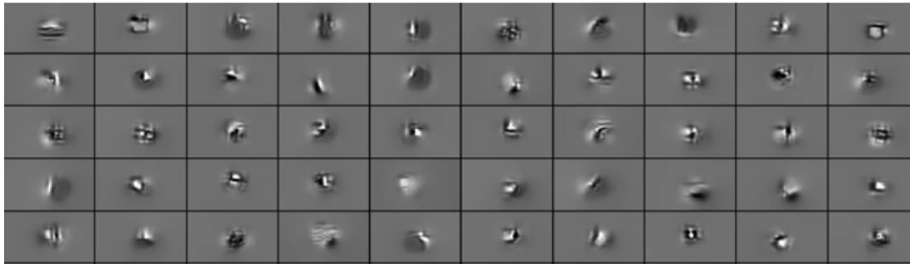


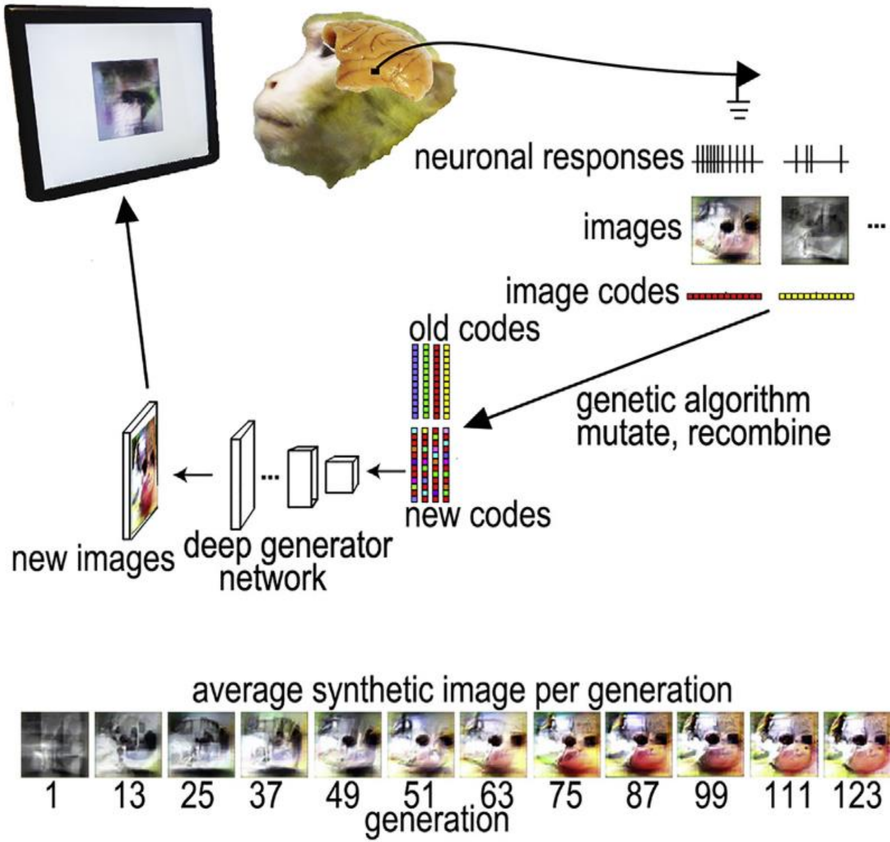
- This procedure works!
- We can generate stimuli that drive neural activity beyond the typical range
- This is a non-invasive control procedure

Bashivan*, Kar*, DiCarlo 2019

Generating "exciting" stimuli without a pre-trained encoding model







- Walker et al. 2019

Ponce & Xiao & Schade et al. 2019

Adversarial attacks in computer vision

- Models are fooled by small, imperceptible perturbations (white box adversarial attacks)
- Protection technique: train on adversarial images "adversarial training" (very costly)

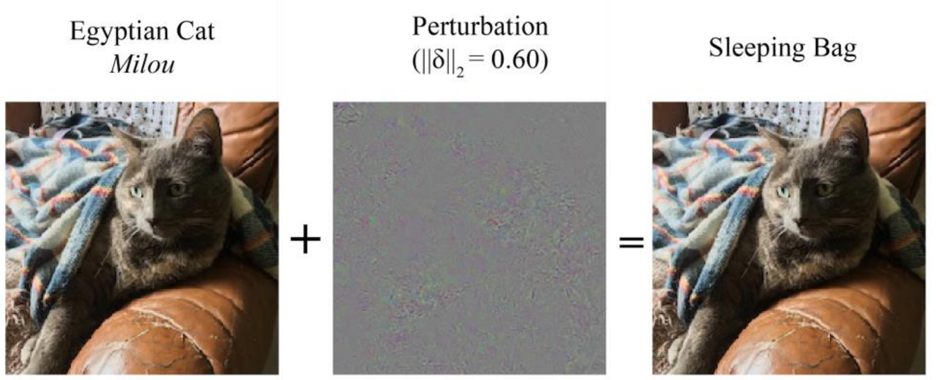


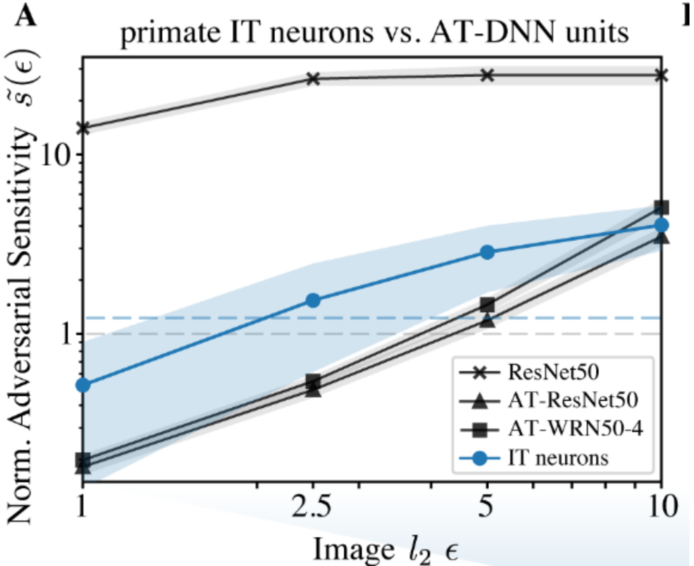
Image from Dapello*, Marques*, et al. (NeurIPS 2021)

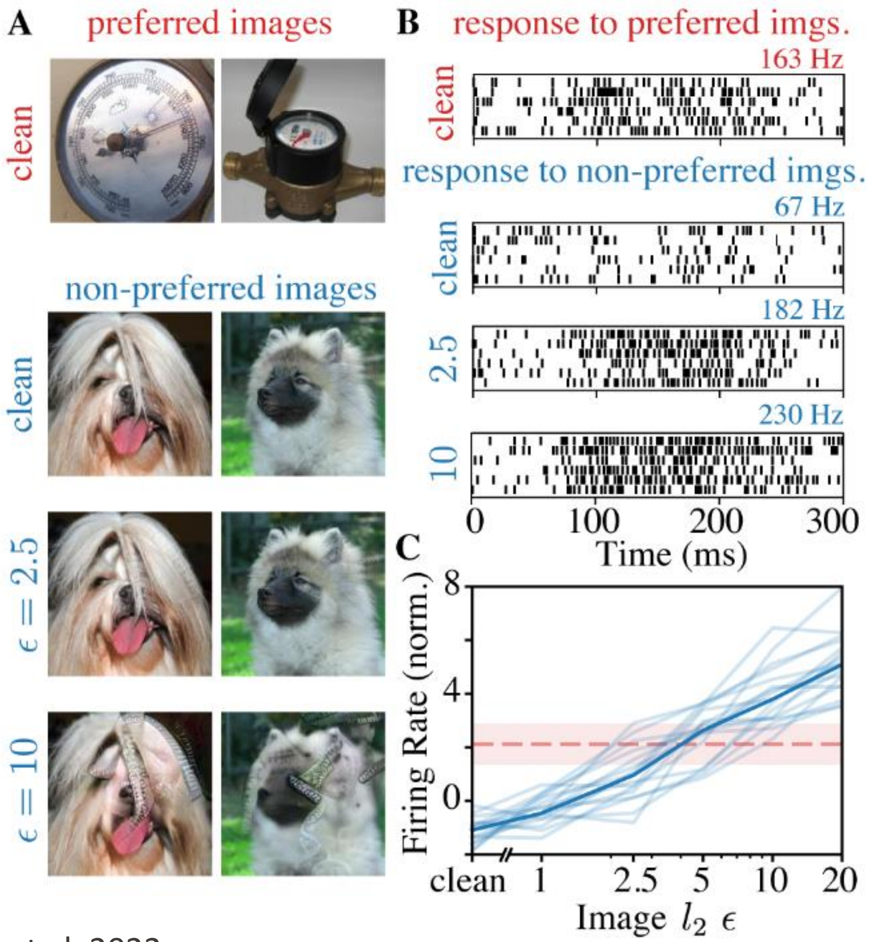
Graffiti

Szegedy et al. (ICLR 2014)
Eykholt*, Evtimov*, et al. (CVPR 2018)

Adversarial attacks on the brain

- Prevalent view: only computational models are susceptible to adversarial attacks
- But: can synthesize images that also fool IT neurons

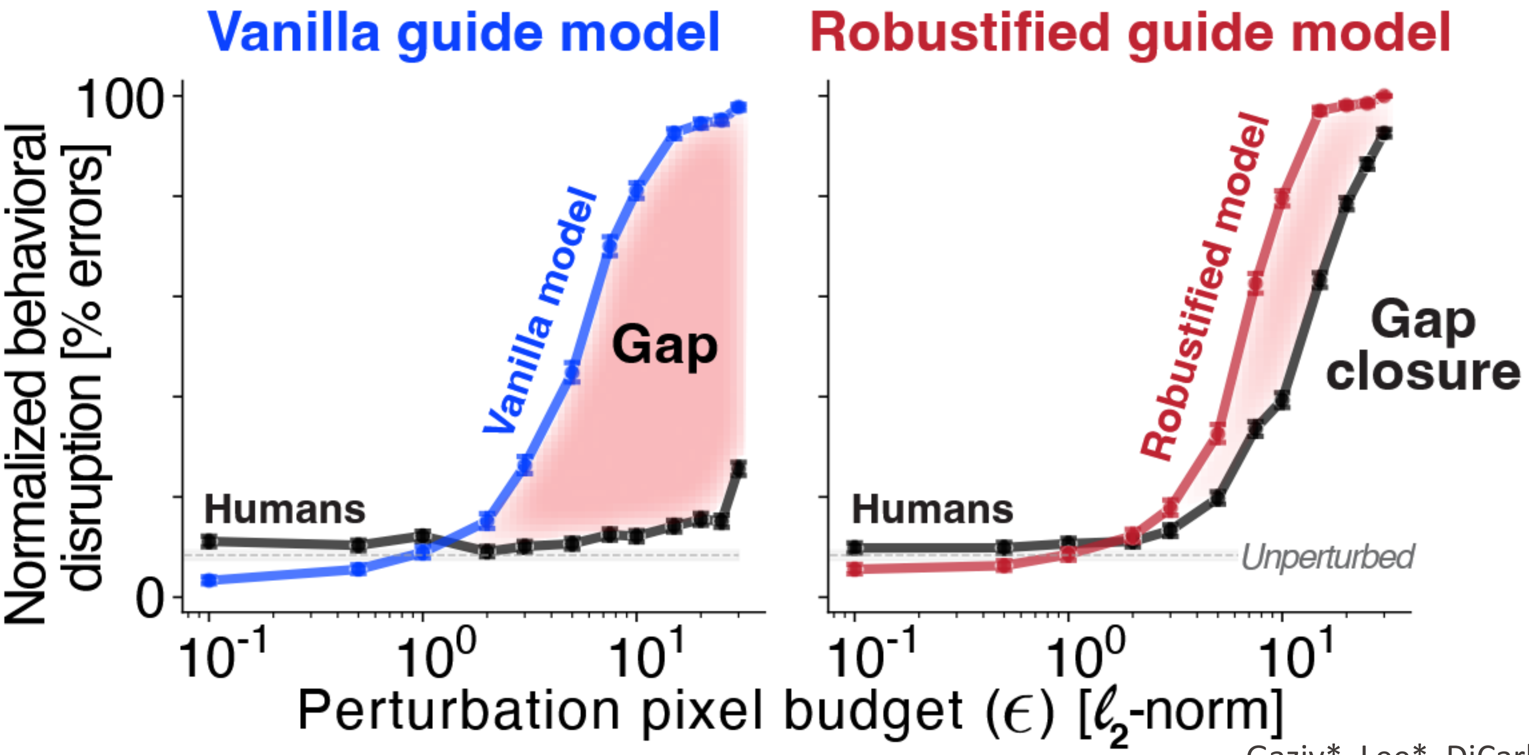




Guo et al. 2022

Adversarial attacks on behavior

- Using a robustified model (**trained with adversarial attacks**), can change images in a way that change the decision of humans



Gaziv*, Lee*, DiCarlo 2023

EPFL

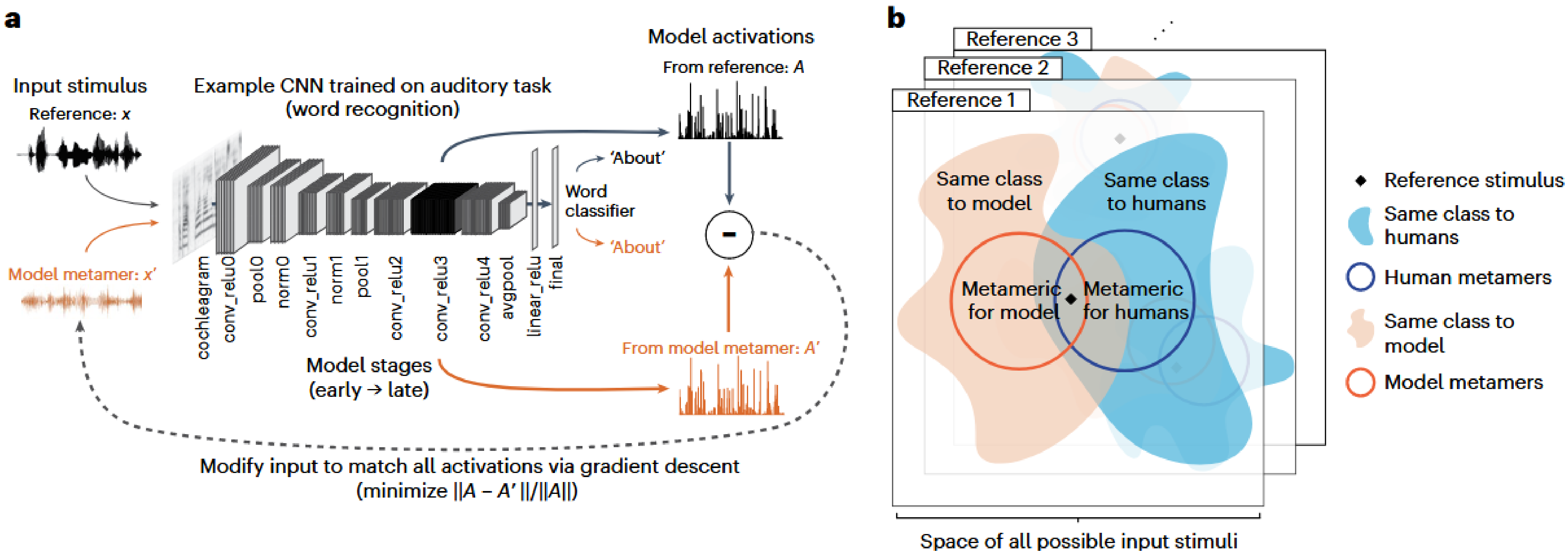
Adversarial attacks on behavior

start images | **Example target categories**
'insect' 'primate' 'fish'

Gaziv*, Lee*, DiCarlo 2023

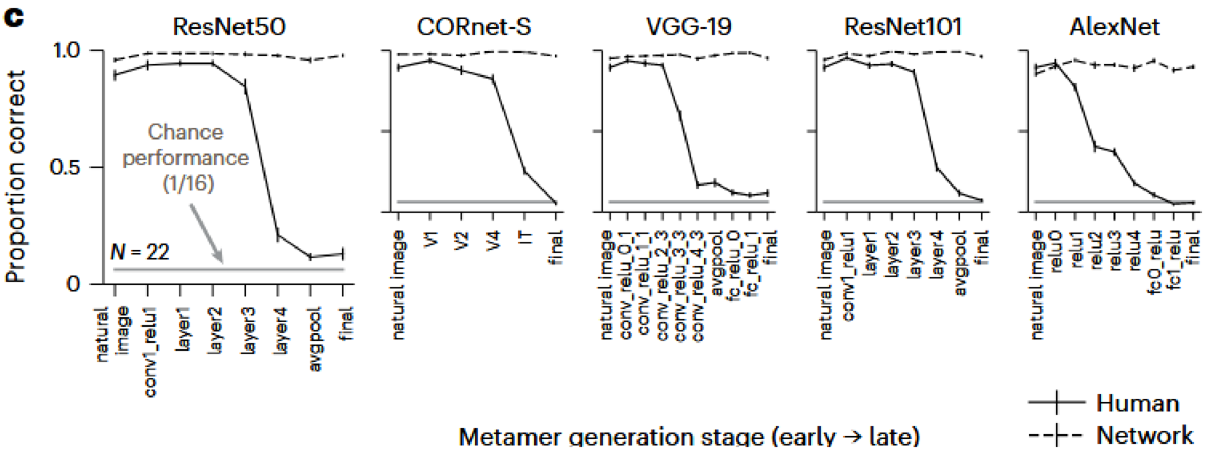
Metamers

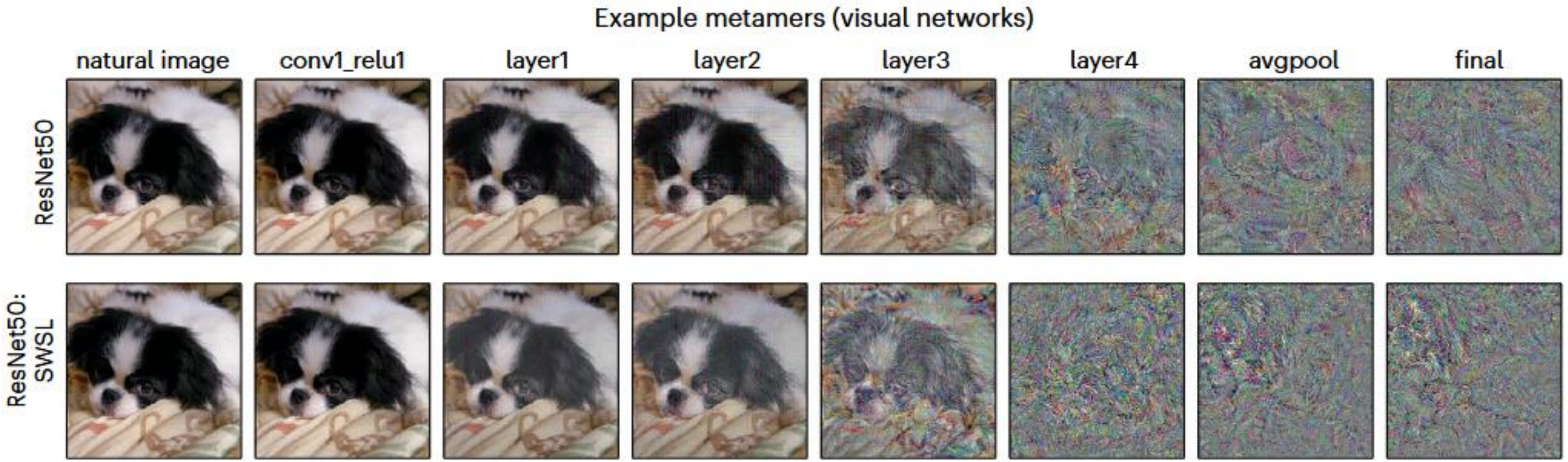
- Model metamers: “stimuli whose activations within a model stage are matched to those of a natural stimulus”



Feather et al. 2023

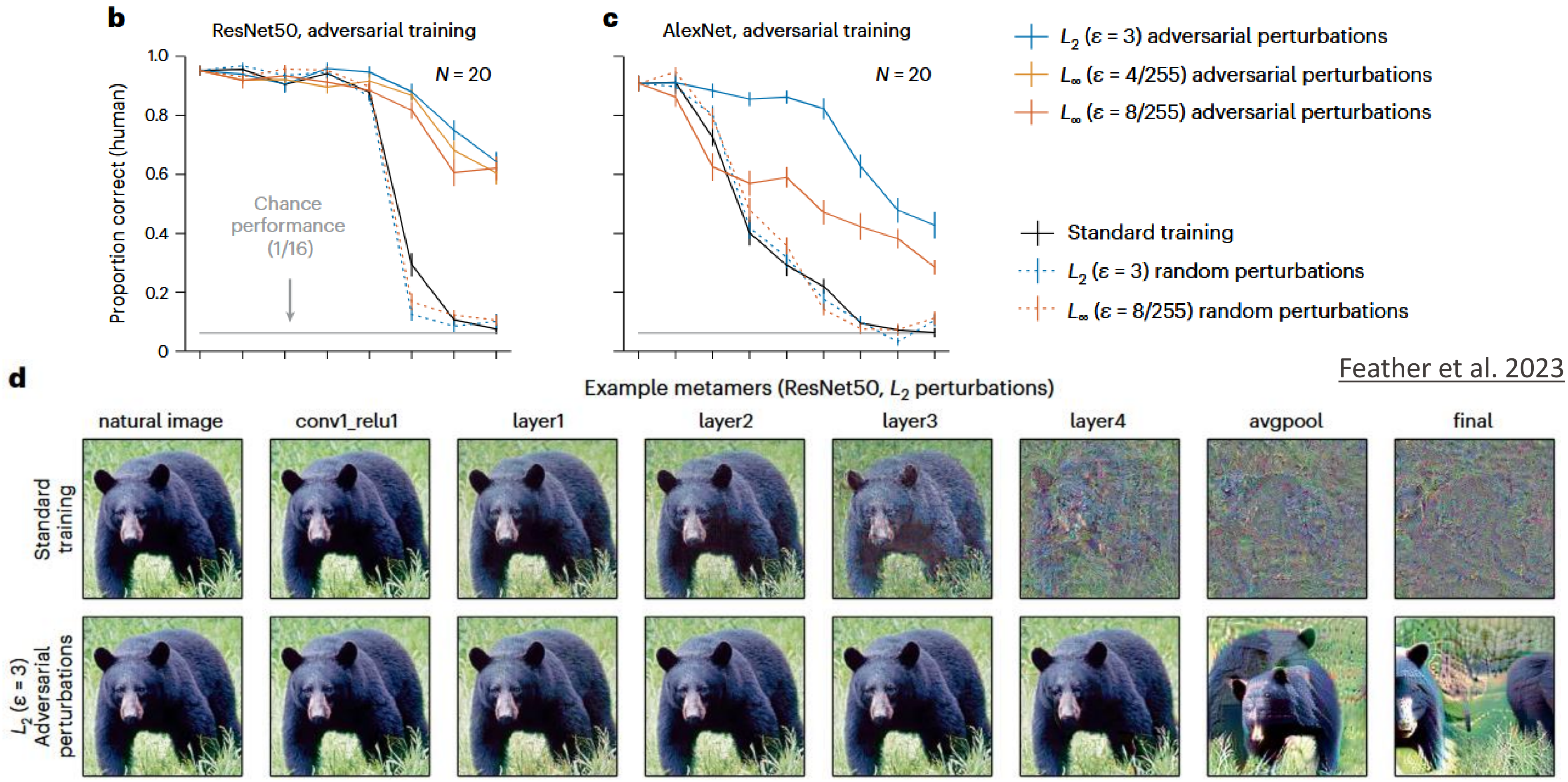
Metamers of standard models are not recognizable by humans





Feather et al. 2023

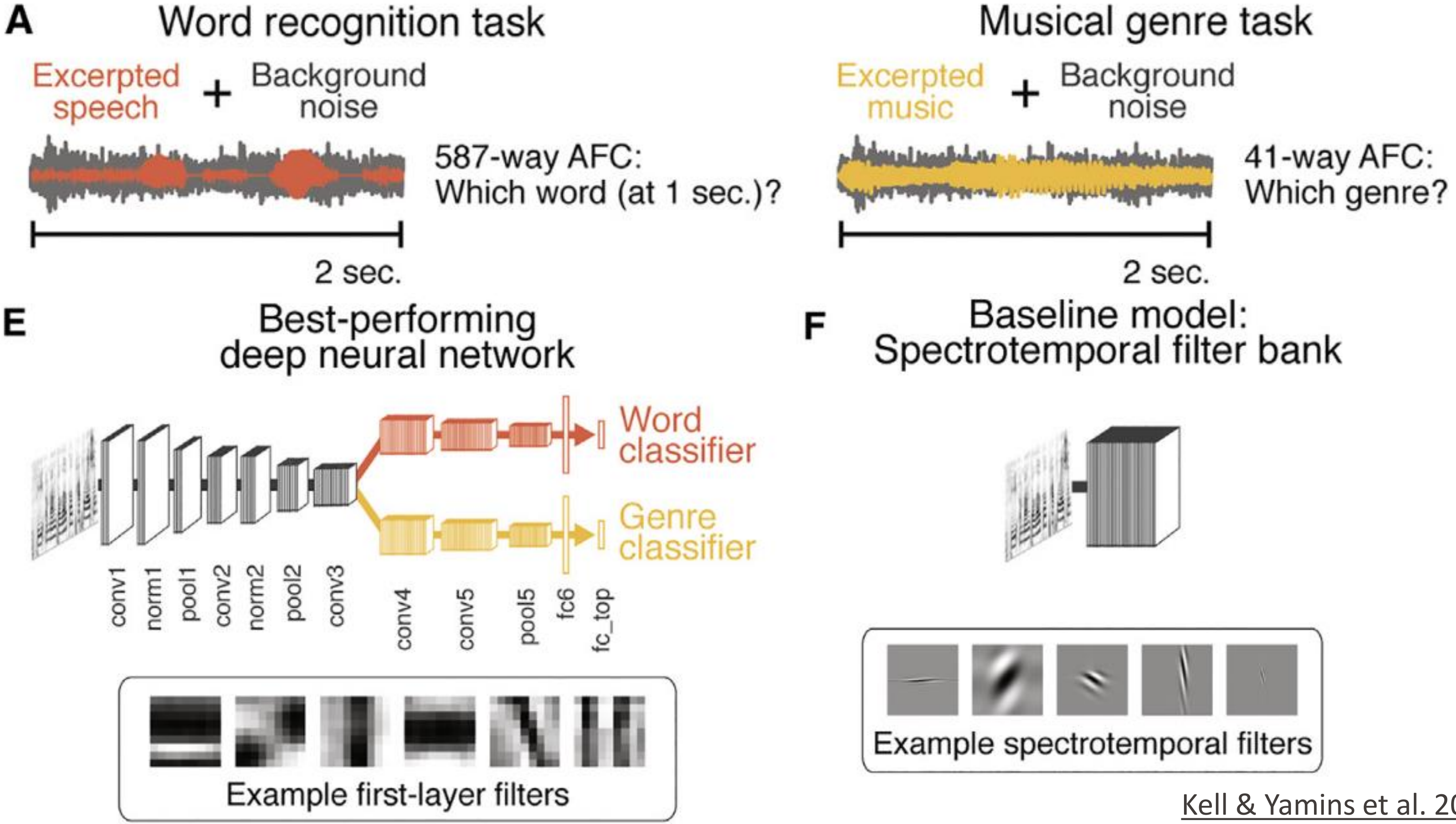
Adversarial training makes metamers human-recognizable



Feather et al. 2023

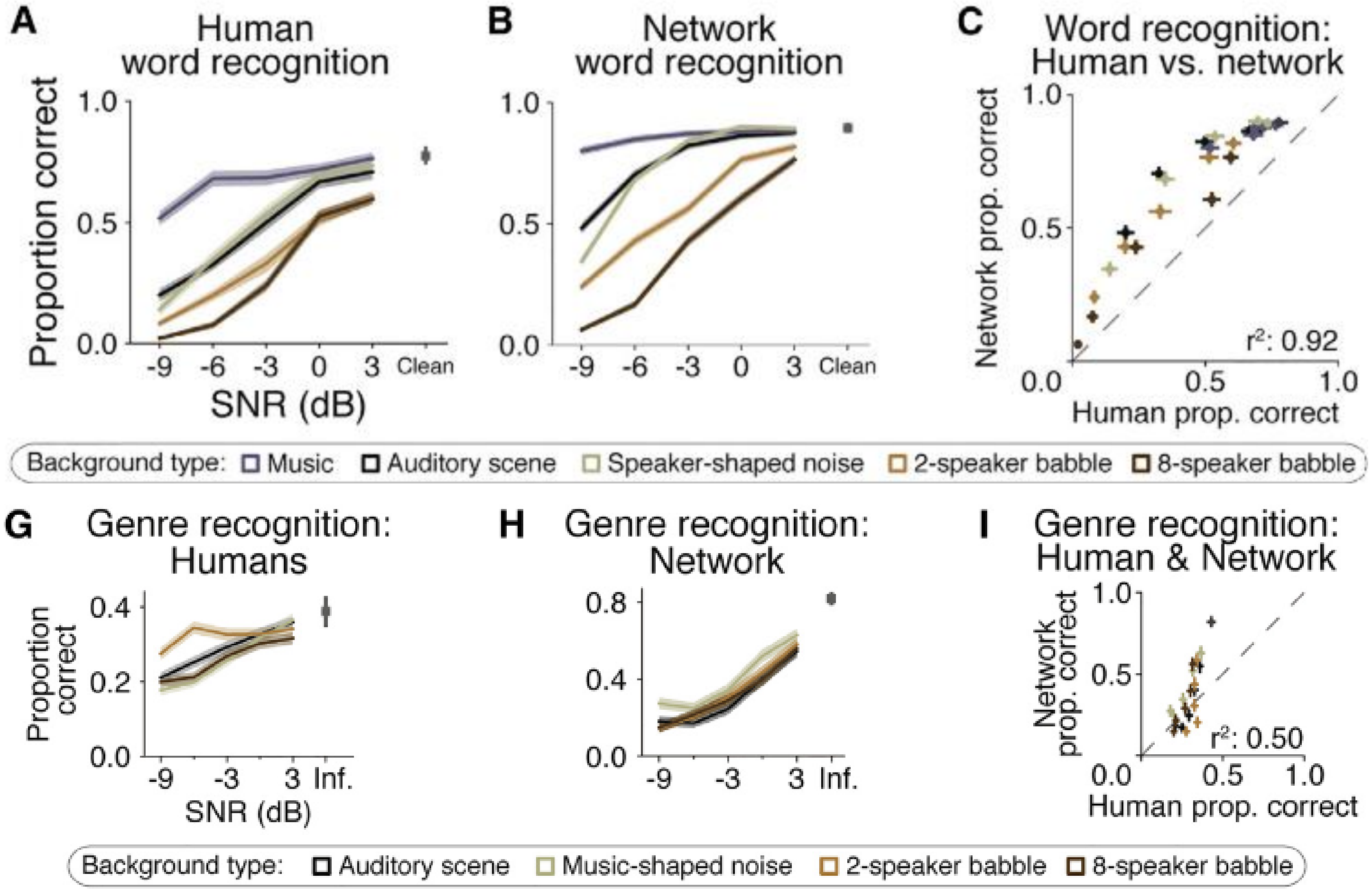
Models of auditory processing

- Jointly optimize CNN for word + genre recognition tasks



Kell & Yamins et al. 2018

Task-optimized model exhibits human-like behavior

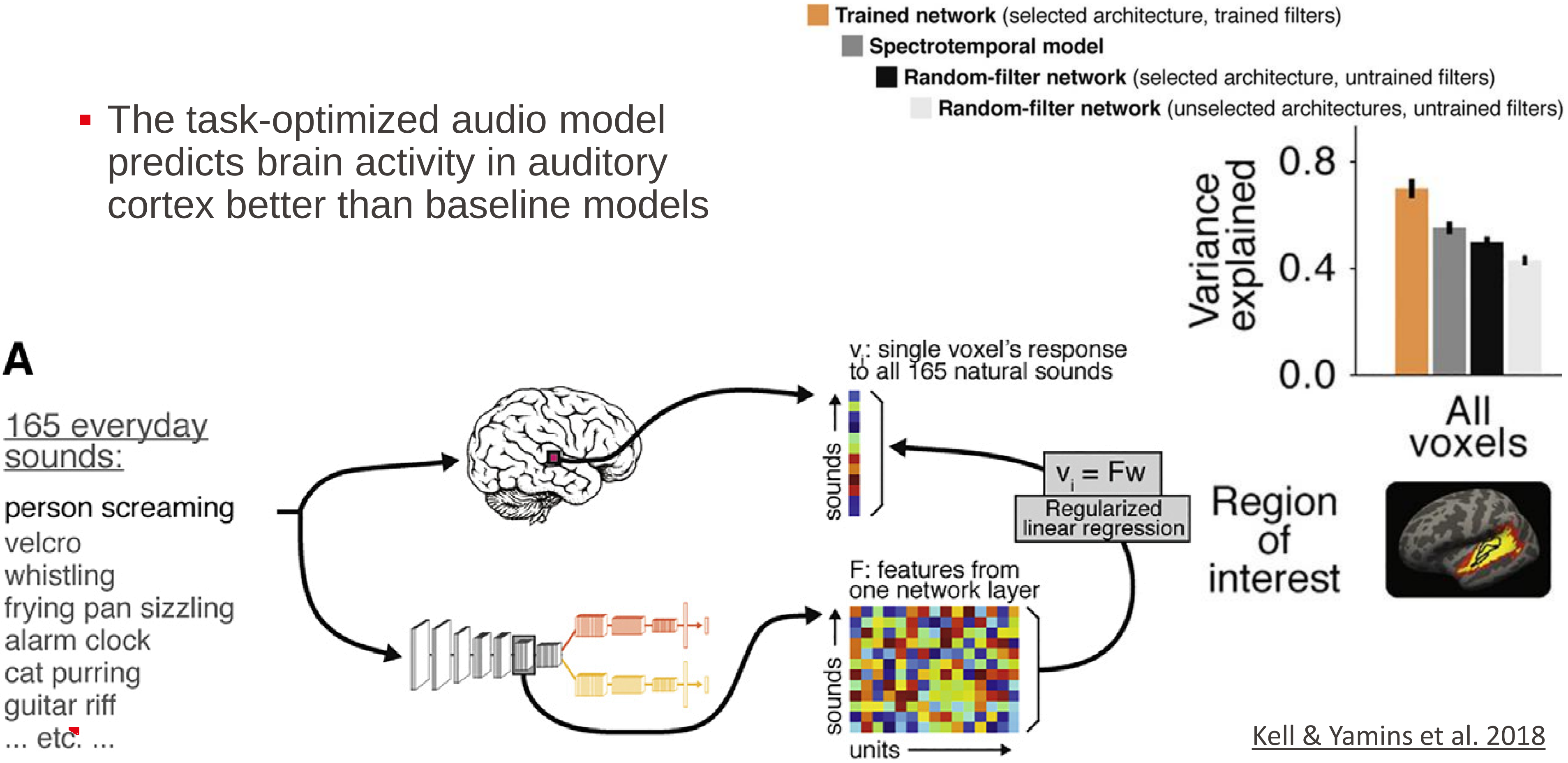


- Model closely predicts human performance patterns, especially for word recognition tasks
- Less behaviorally-aligned for genre recognition

Kell & Yamins et al. 2018

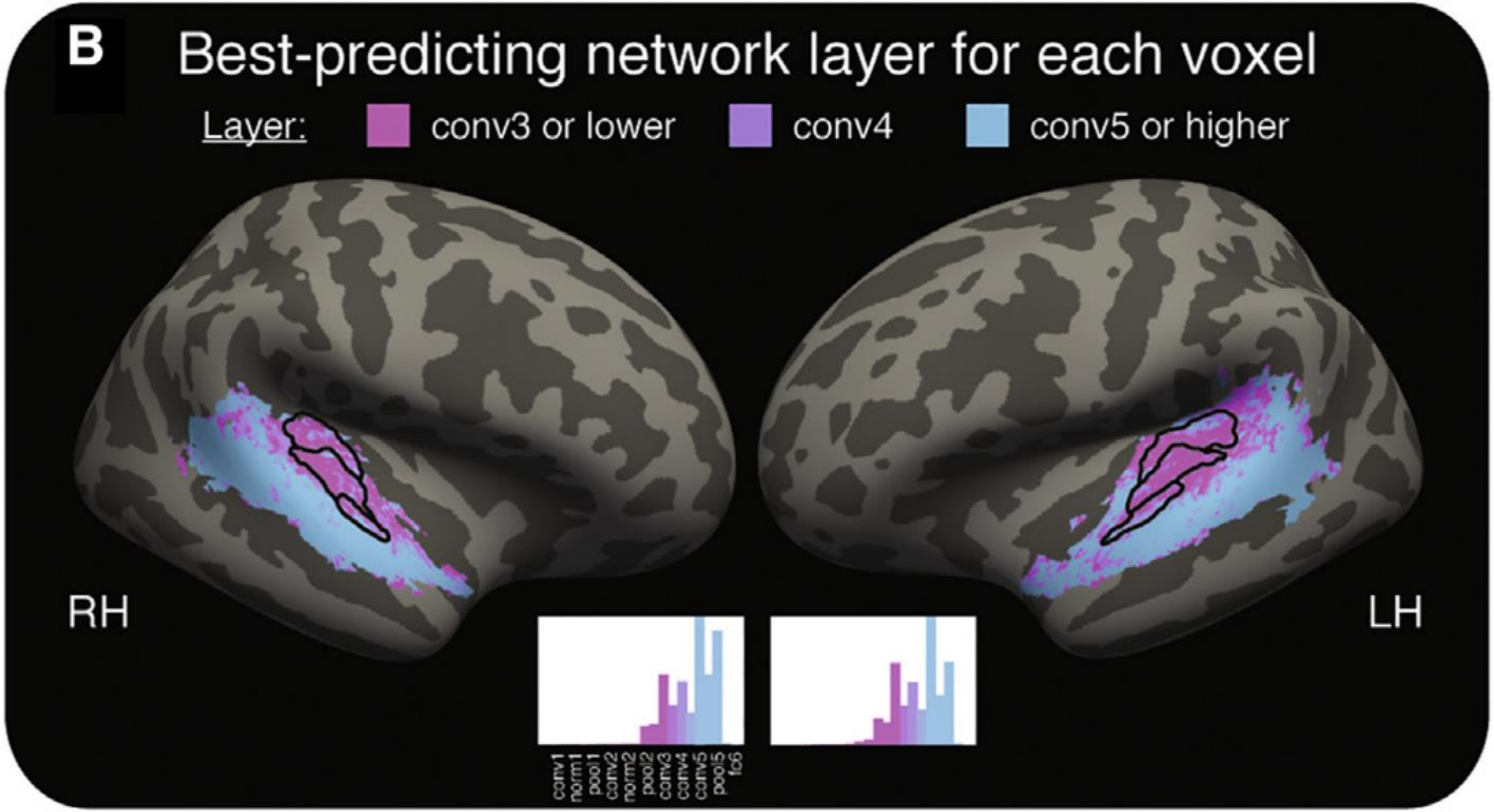
Task-optimized audio model predicts fMRI responses

- The task-optimized audio model predicts brain activity in auditory cortex better than baseline models



Kell & Yamins et al. 2018

Model Predicts Hierarchical Organization in Human Auditory Cortex



- Black outline: sub-divisions of primary auditory cortex
- Primary auditory cortex best explained by earlier layers
- Later layers best explain non-primary areas

Kell & Yamins et al. 2018

Take-home messages

- Unsupervised training yields brain-like representations
- Including a spatial loss term leads to topographic models that reproduce the spatio-functional organization in the brain. These models can predict the behavioral effects of neural interventions
- Encoding models of brain function allow for the synthesis of images to control neural activity. Can create images akin to adversarial attacks
- Yet, models often view things as identical that appear very different to humans (metamers)
- Very similar ideas from vision work for models of auditory processing