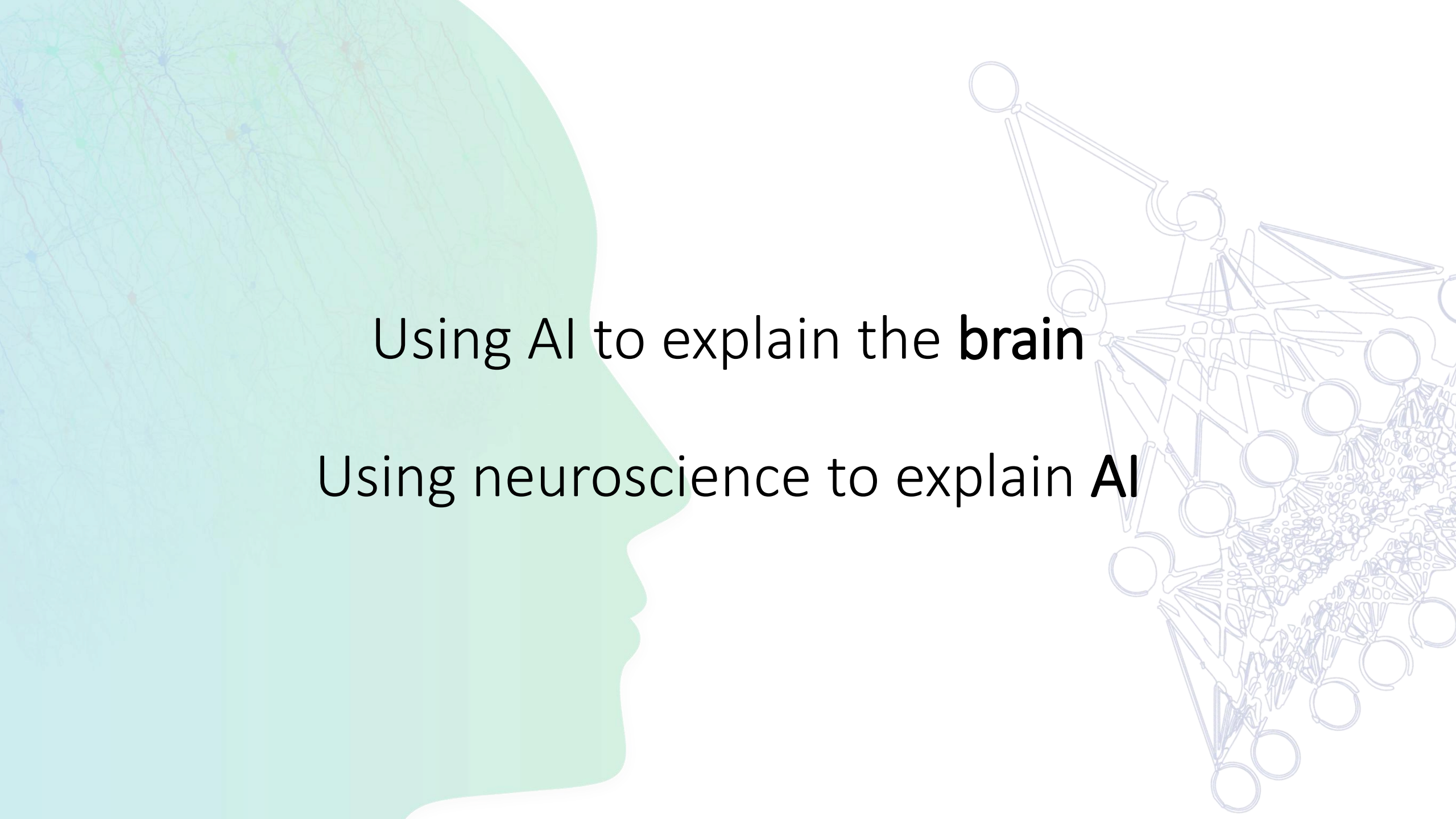
NeuroAI Lab

Martin Schrimpf

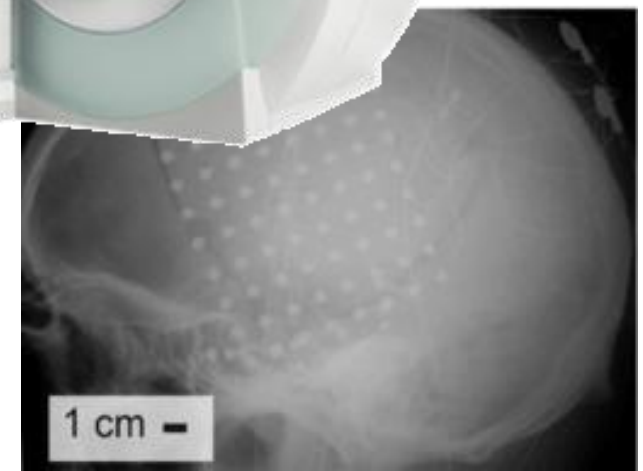
martin.schrimpf@epfl.ch

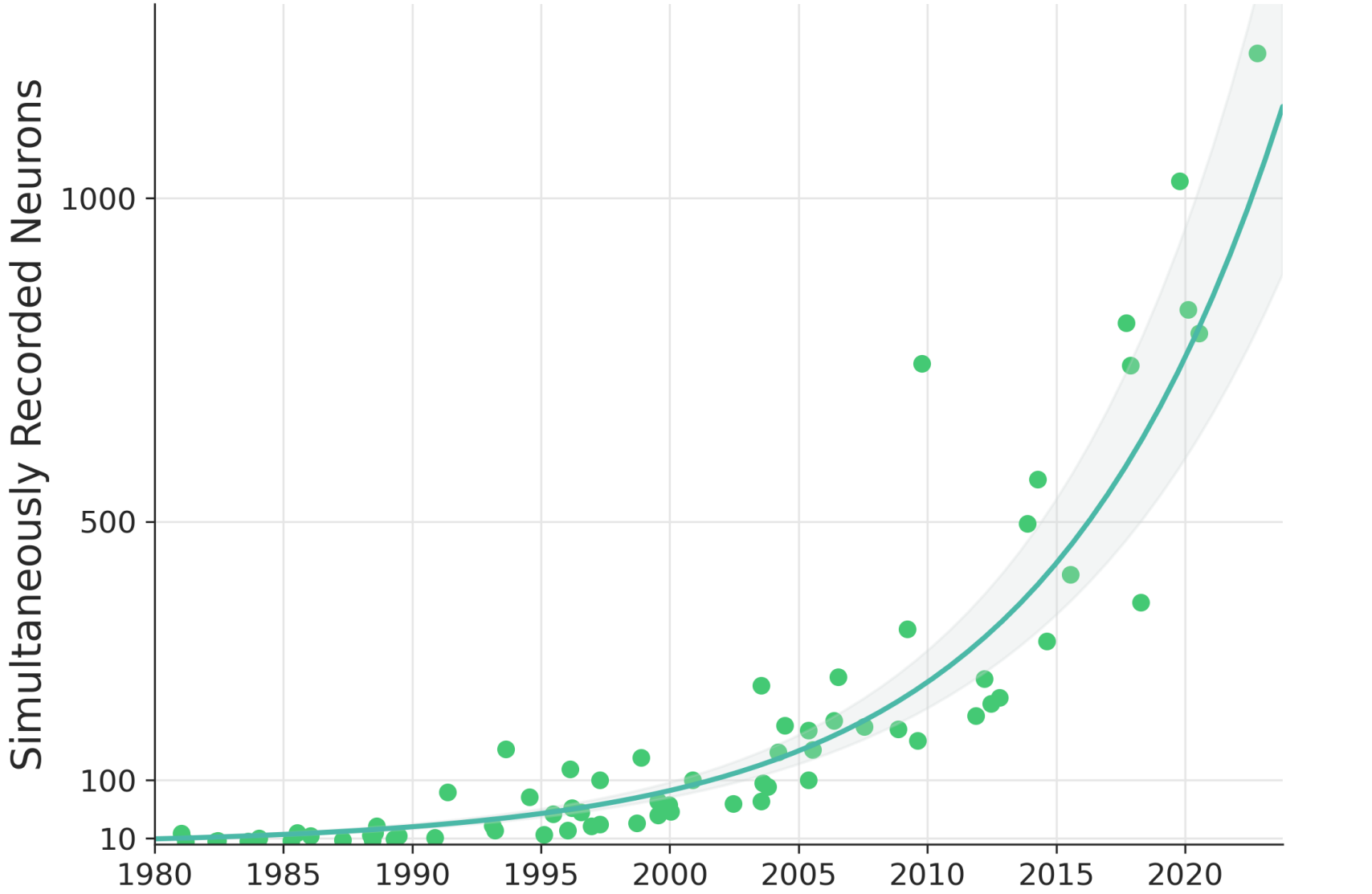
EPFL SV IC

Neuro X Institute


Using AI to explain the **brain**
Using neuroscience to explain **AI**

Access to neural data is increasing exponentially



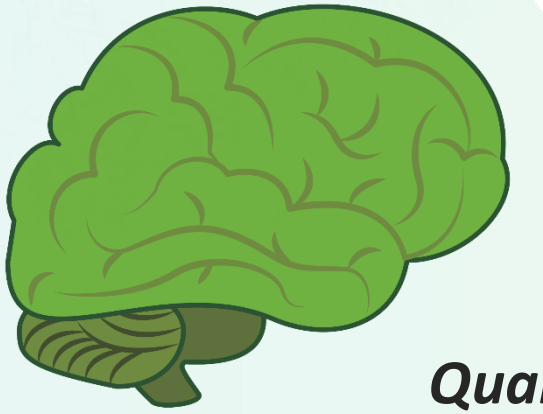


https://stevenson.lab.uconn.edu/scaling

Quantitative synergy between science + engineering

Brain and Cognitive Sciences

AI/ML Engineering

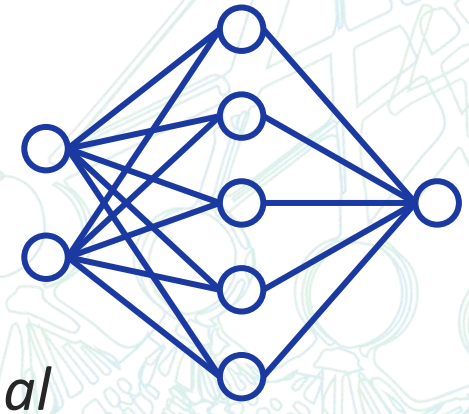


Quantitative *Measurements & Discoveries*

BUILD

Brain Models of **Natural Intelligence**

TEST



Computational ***Hypotheses*** *& Fine-grain* ***Predictions***

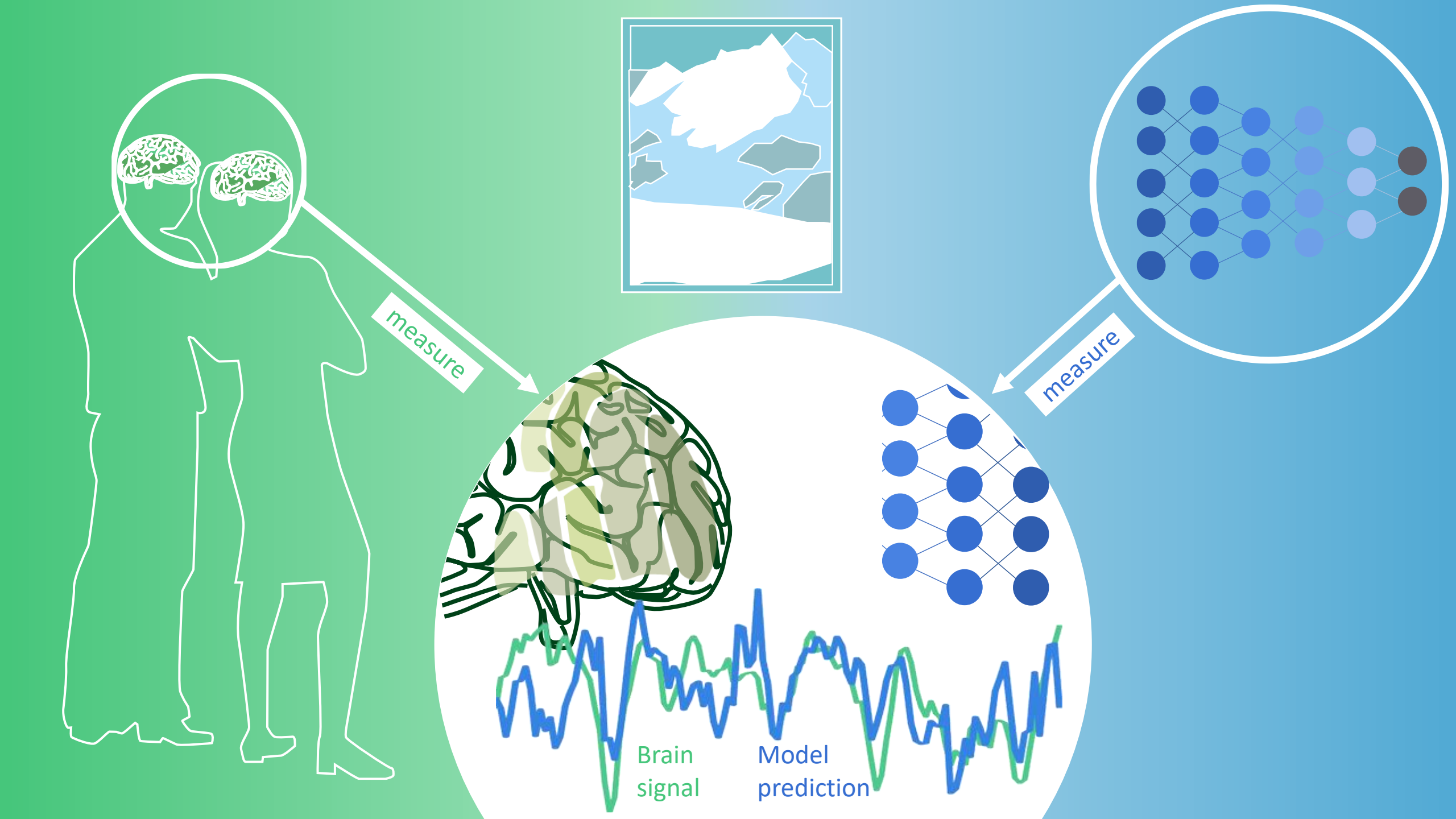
Key messages today

1. AI → Neuro: Particular neural network models are state-of-the-art at predicting brain function and human behavior.

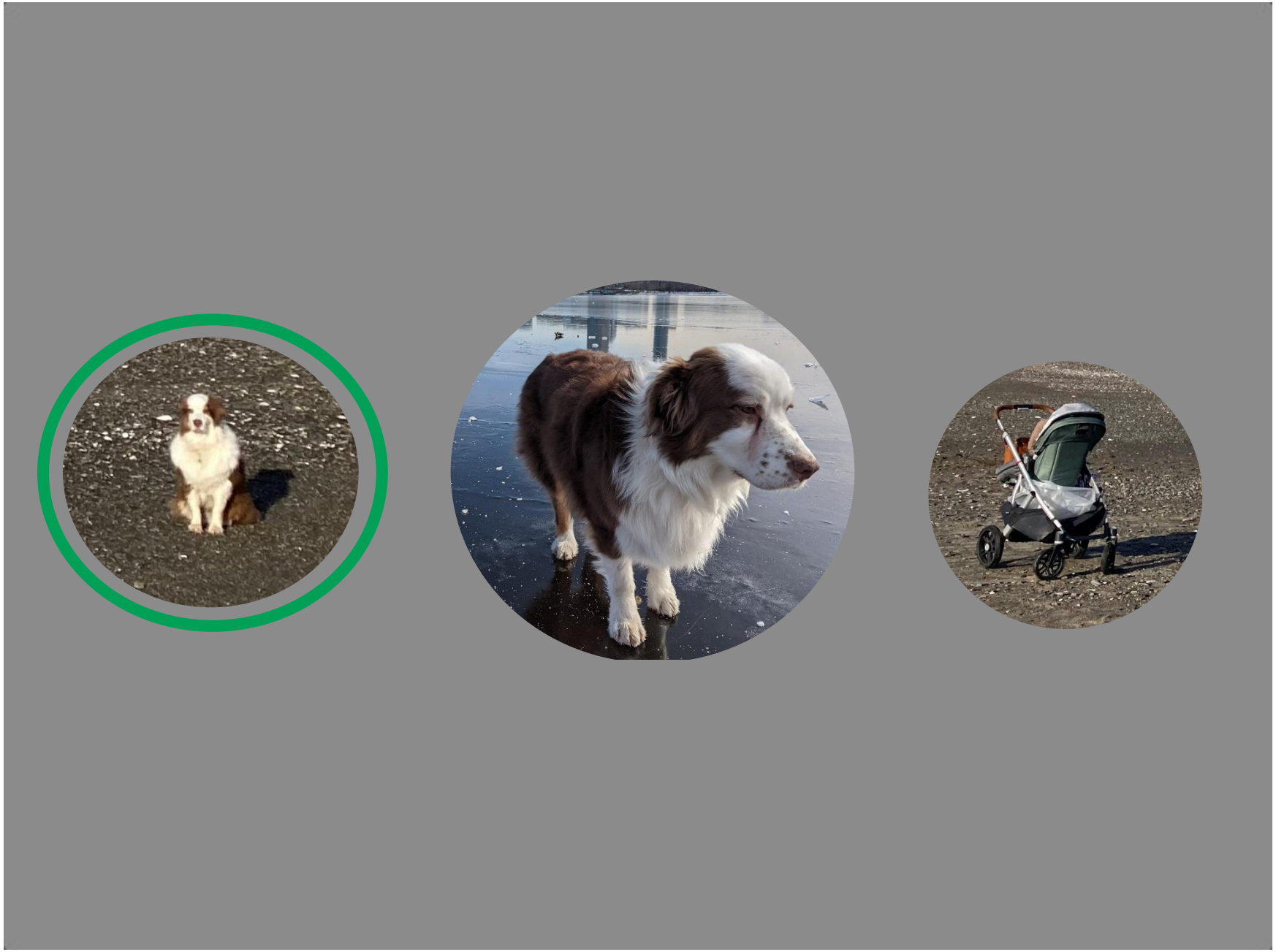
2. Neuro → AI: Neuroscience has found very effective ways of identifying network function which are applicable to AI.

1. AI → Neuro: Particular neural network models are state-of-the-art at predicting brain function and human behavior.

2. Neuro → AI: Neuroscience has found very effective ways of identifying network function which are applicable to AI.

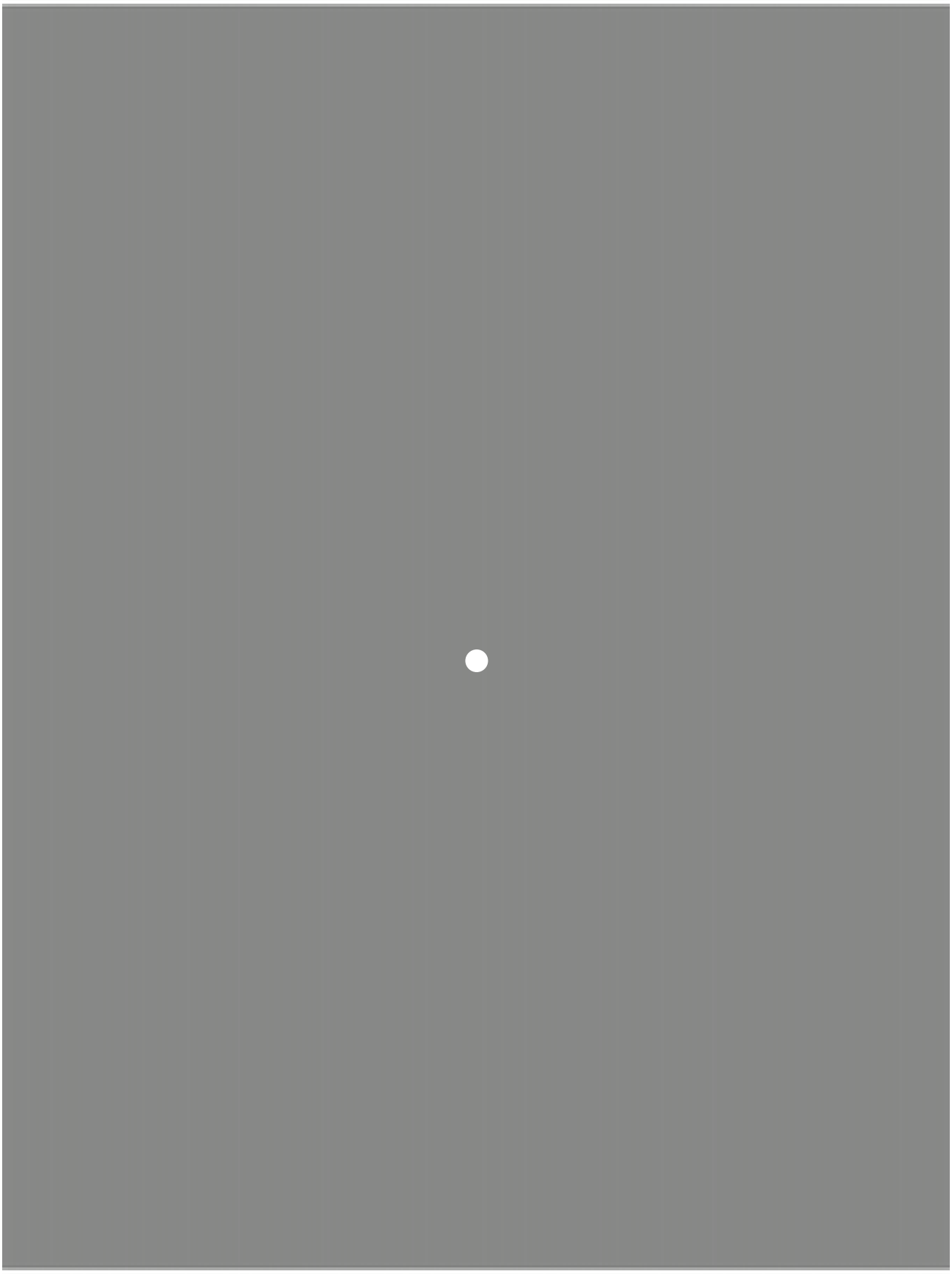

measure
measure
Brain
signal
Model
prediction

Behavioral experiment



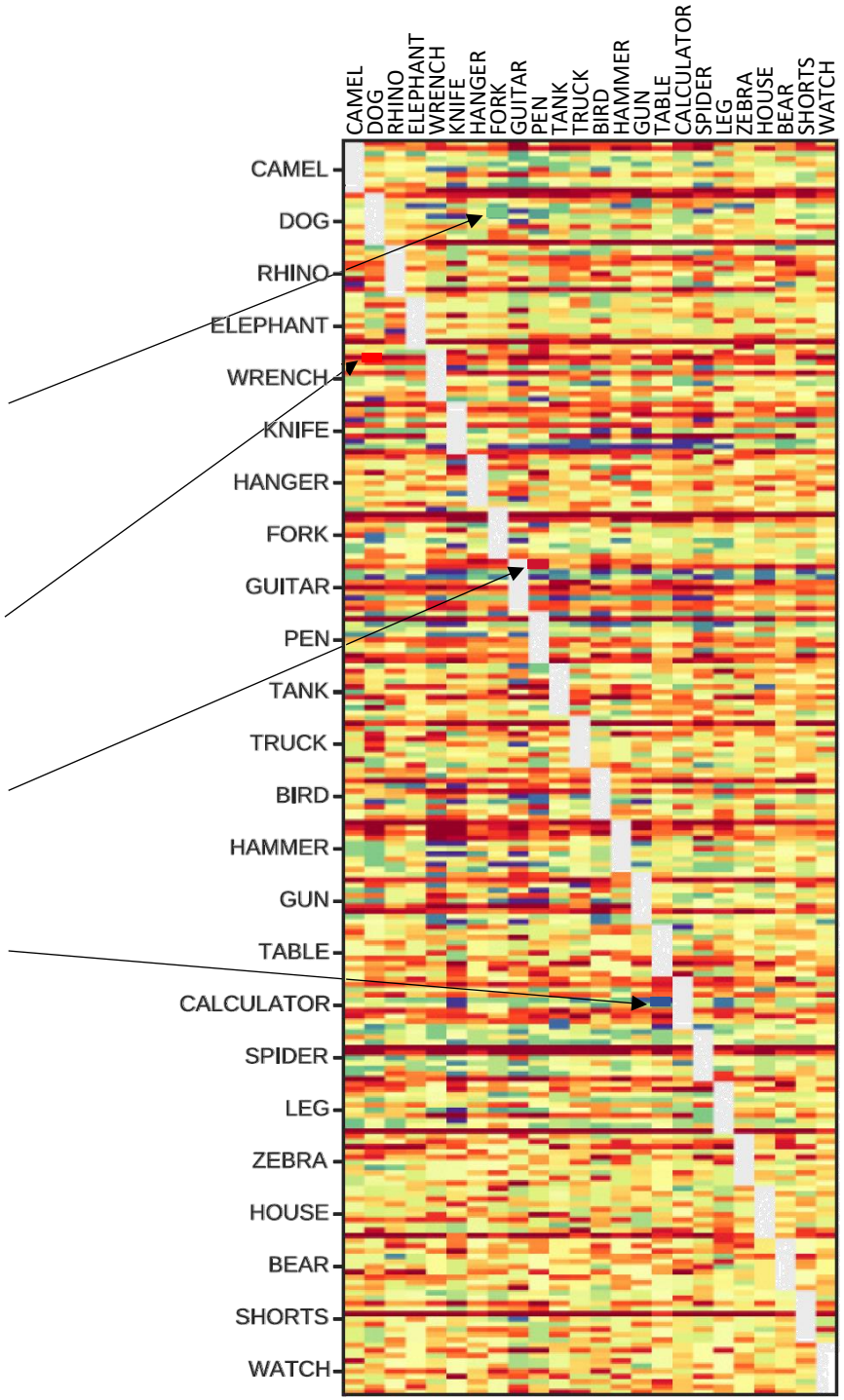
Rajalingham, Issa*, et al. (JNeuro 2018)*

Behavioral experiment

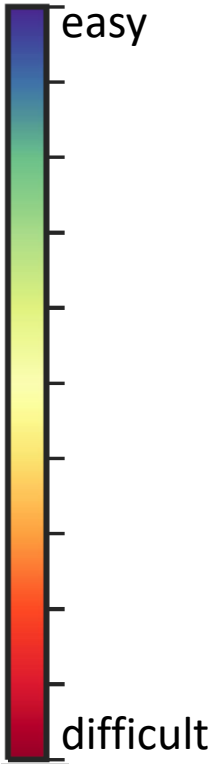


Rajalingham, Issa*, et al. (JNeuro 2018)*

Behavioral data

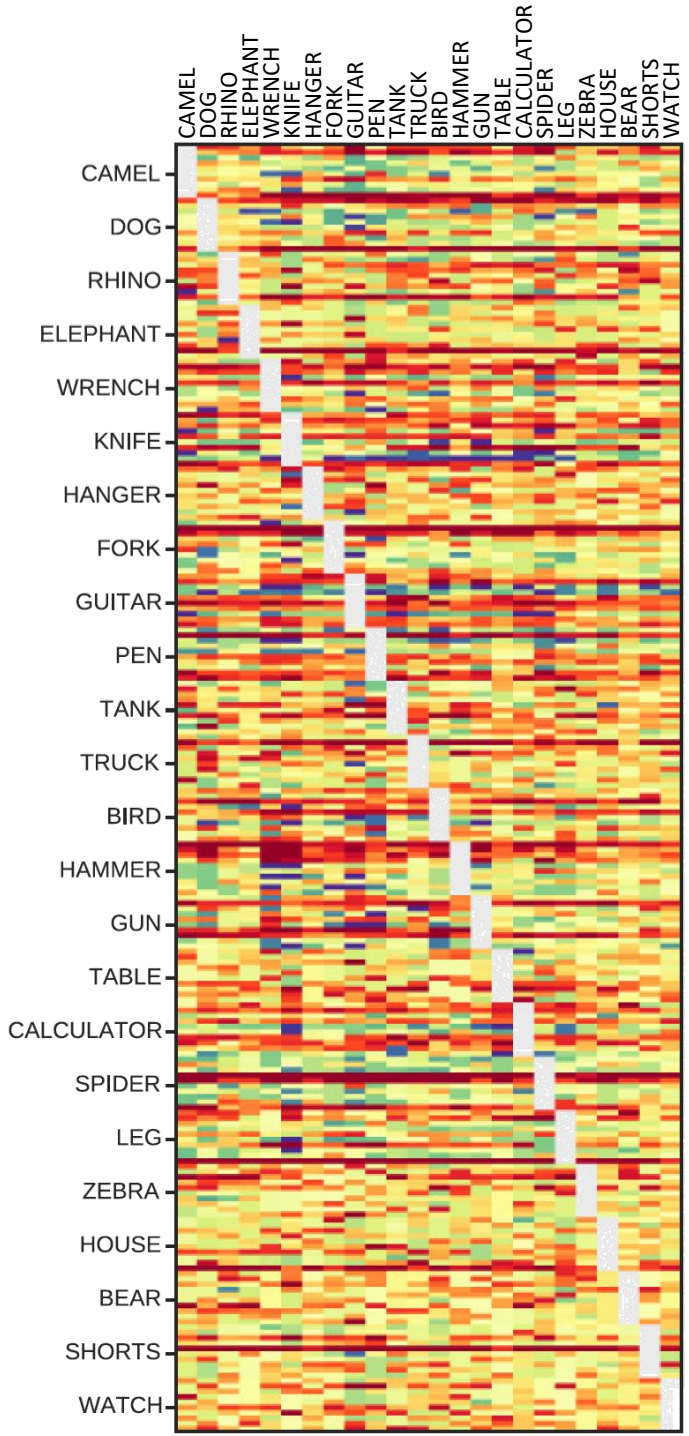


performance per binary categorization

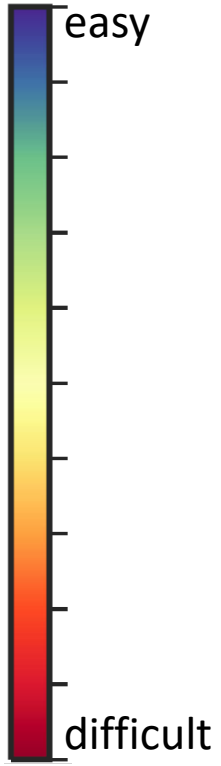


Rajalingham, Issa*, et al. (JNeuro 2018)*

Behavioral data

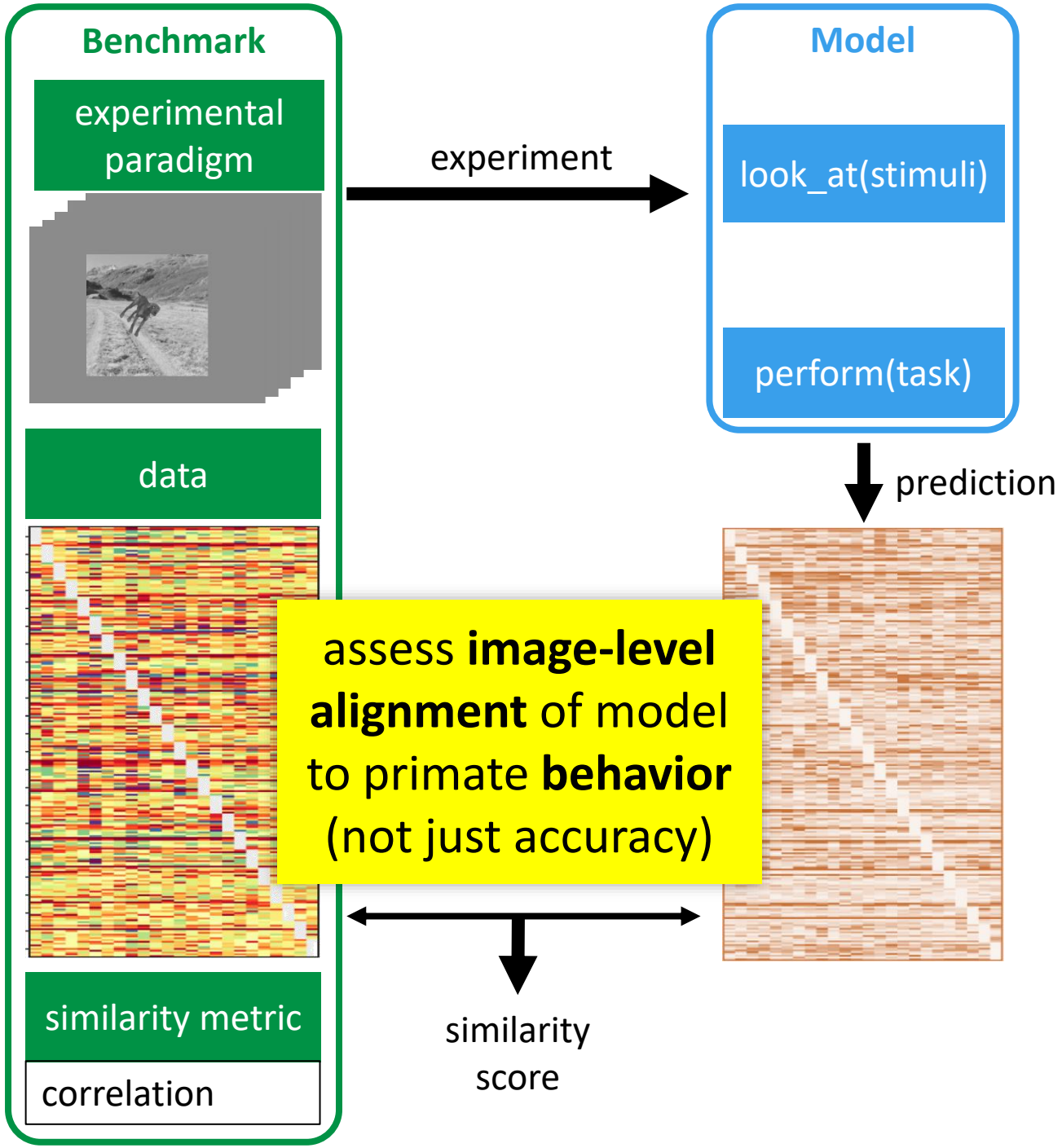


performance per binary categorization



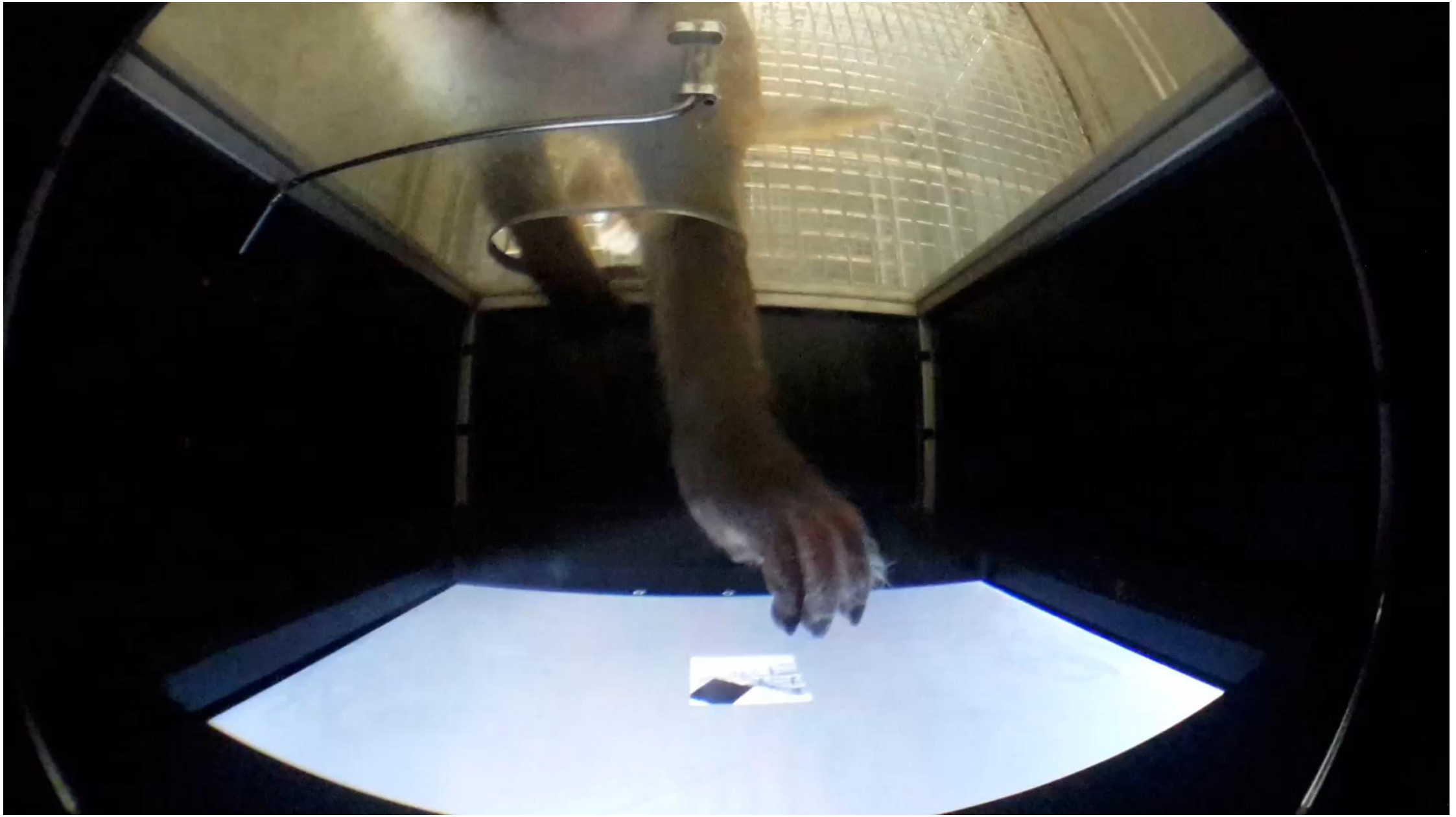
Rajalingham, Issa*, et al. (JNeuro 2018)*

Behavioral benchmark



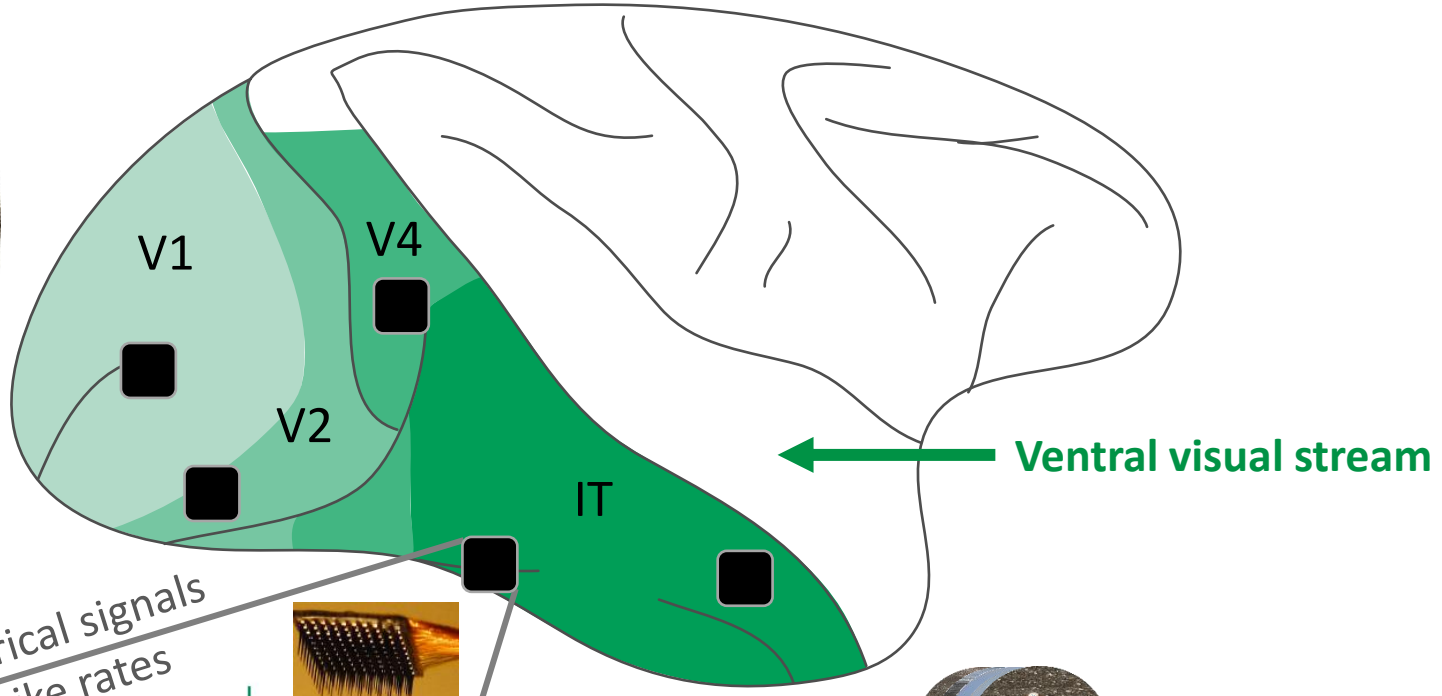
Neural data

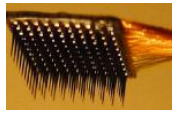
Neural data



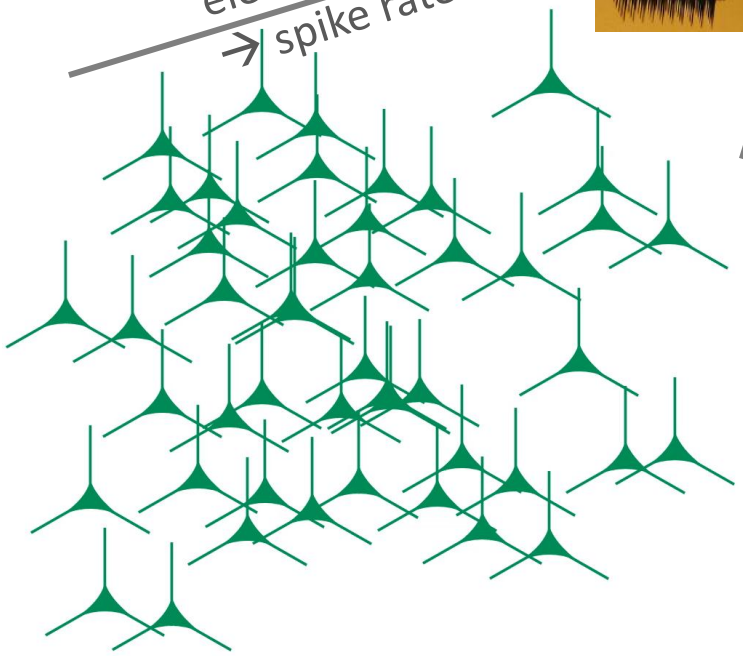
video courtesy of Kailyn Schmidt

Neural data





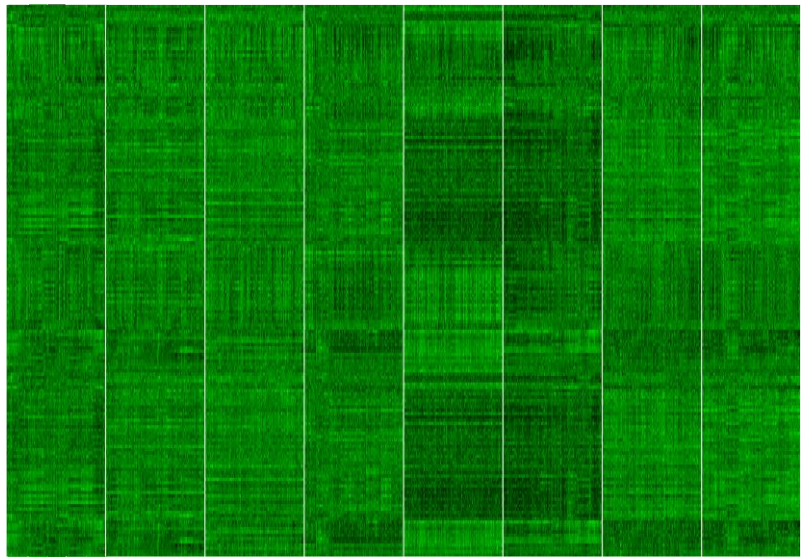
electrical signals
spike rates



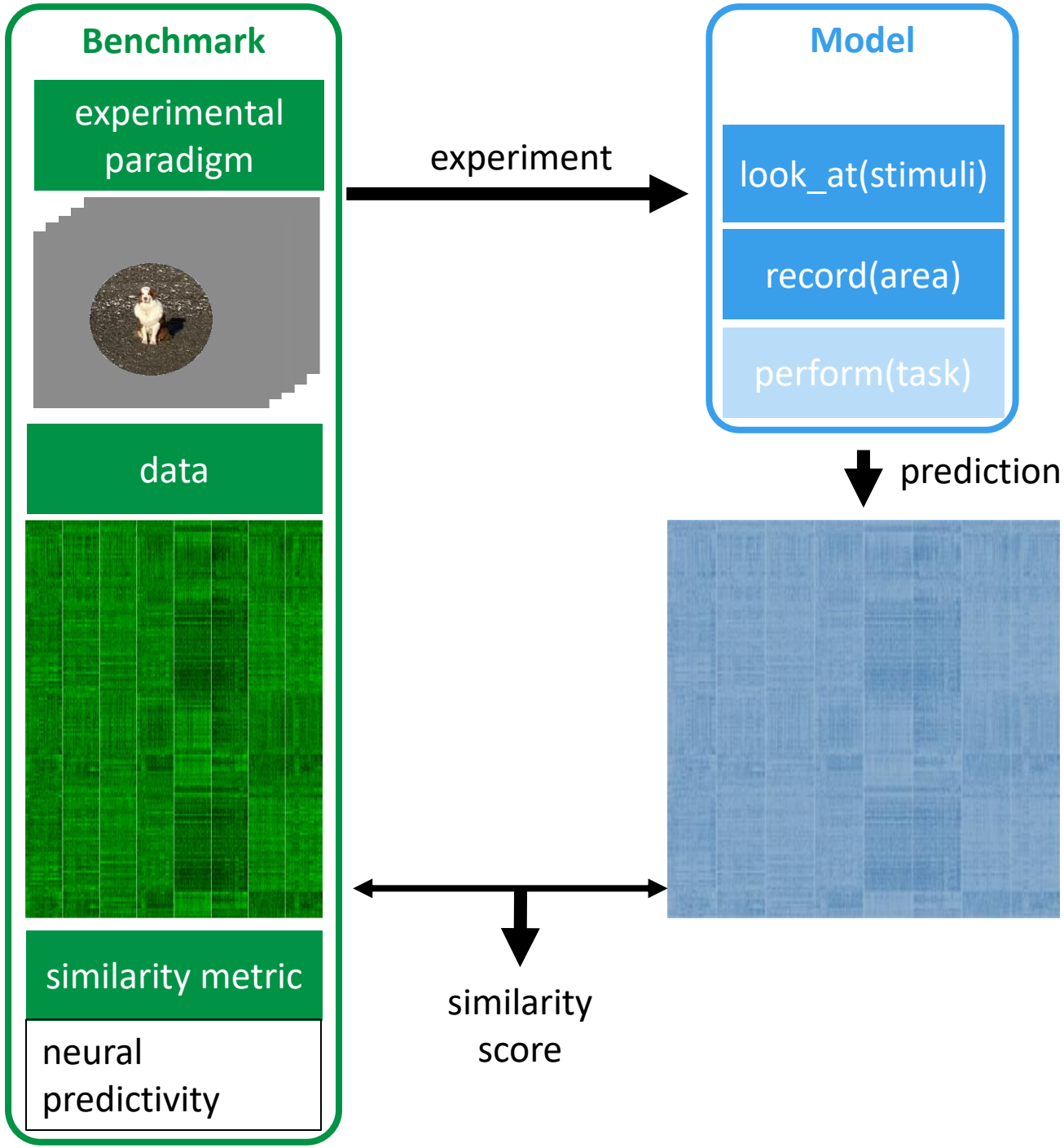
...

images

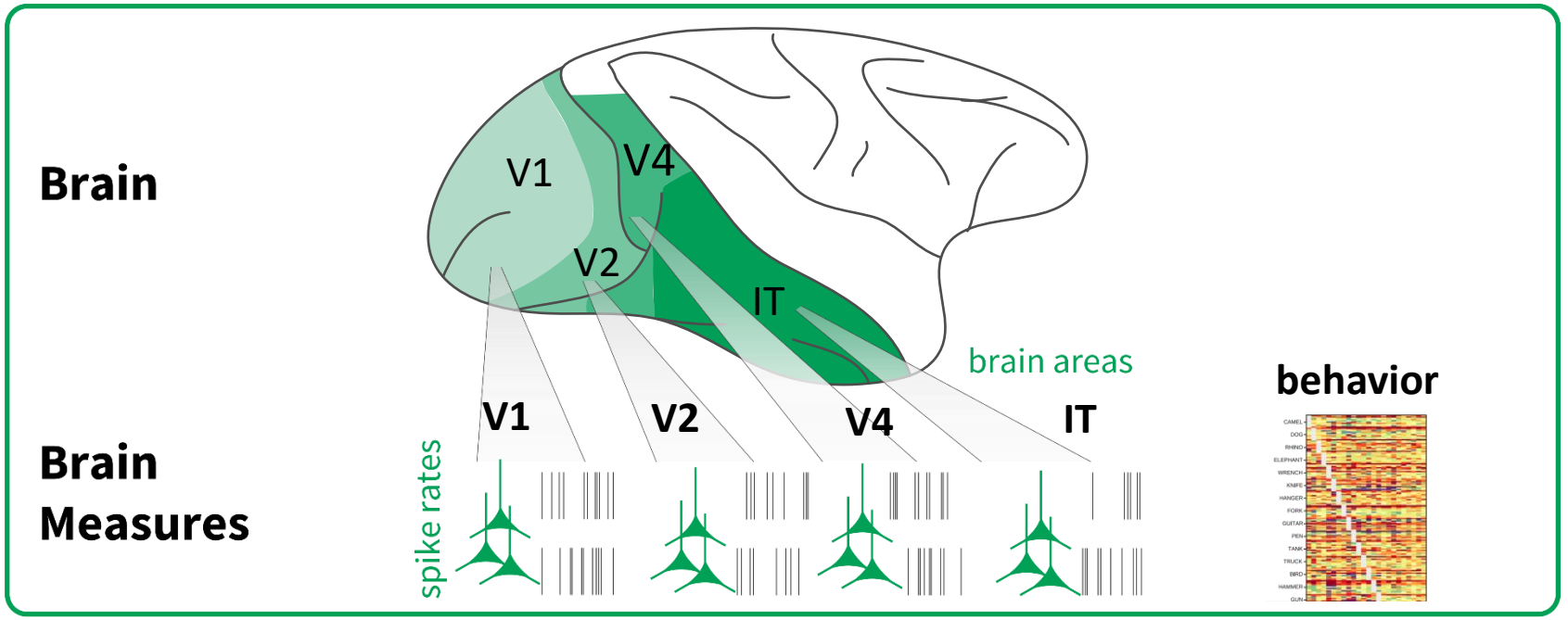
neurons



Neural benchmark

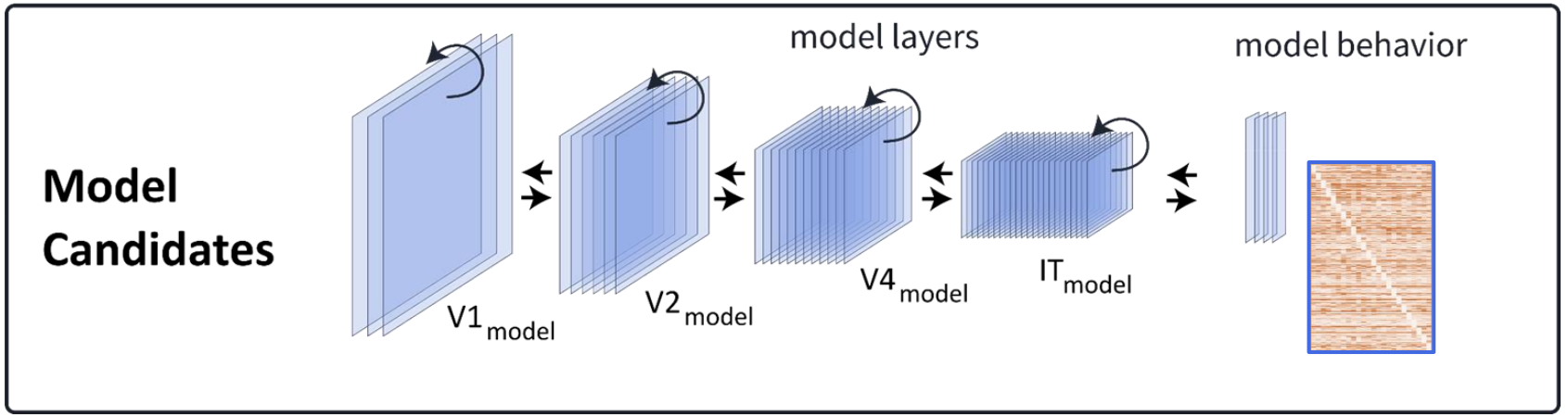


Integrative testing of models on brain + behavioral data



Brain-Score

Benchmarks

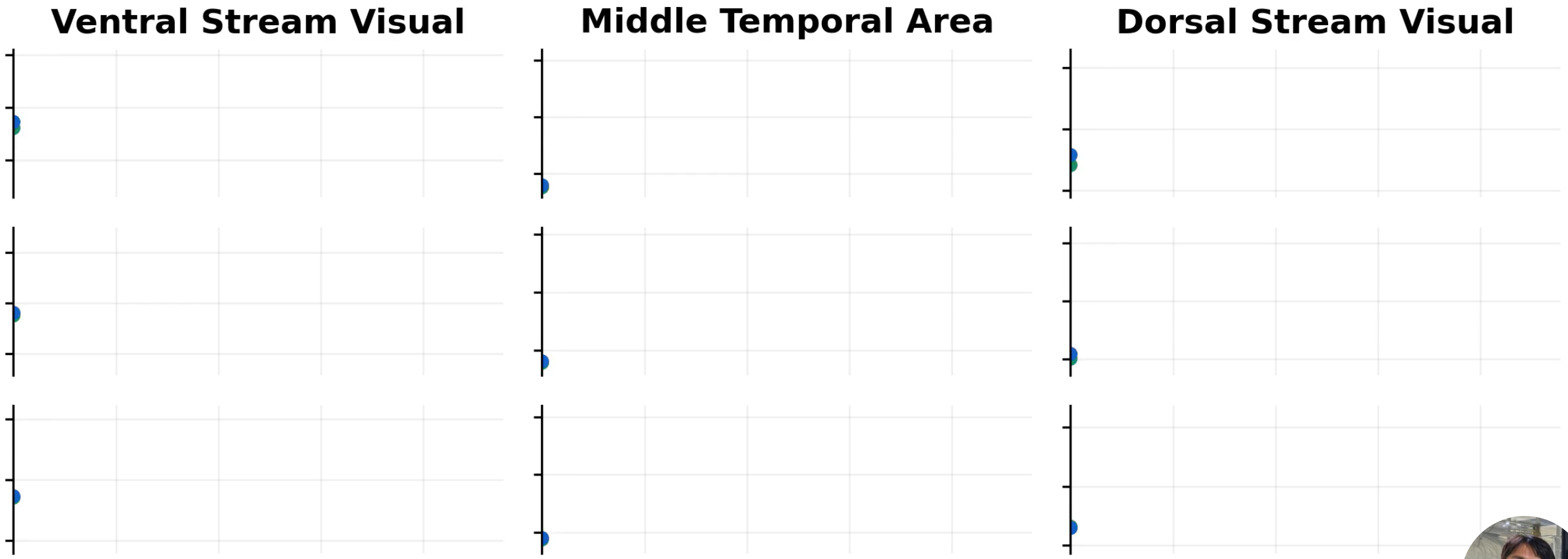


One model to predict *all* data
(not piecewise efforts)

Artificial neural network models

- Trained for computational task, weights optimized via backprop *(not trained on brain data)*
- Internal processing stages (hidden layers, “deep” learning)
- Accept any new input (pixels)

SOTA ML models predict brain activity with high fidelity

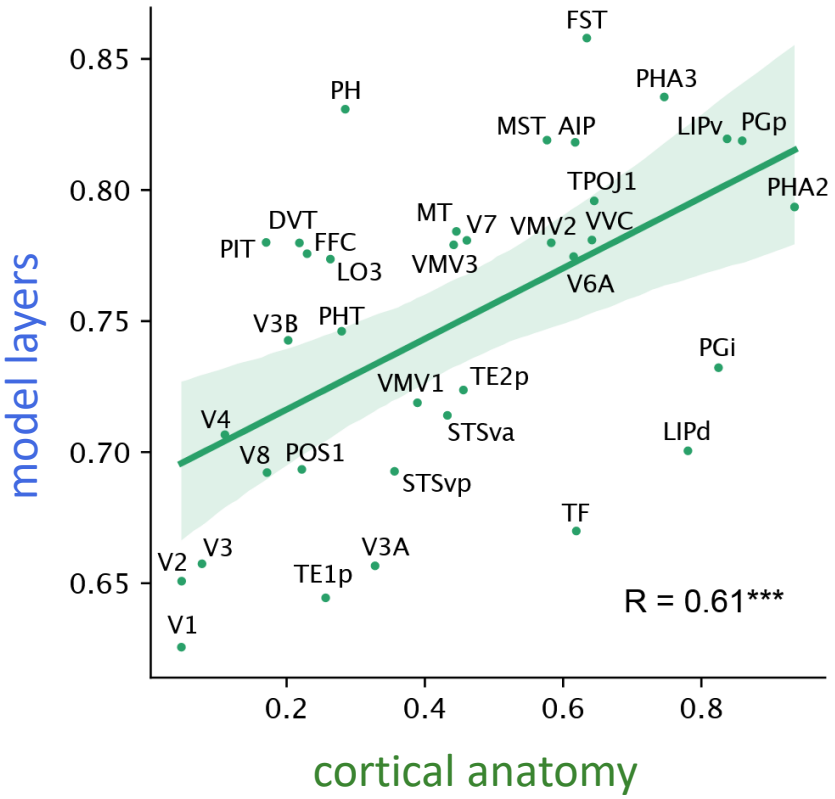


Brain V-JEPA (brainified)

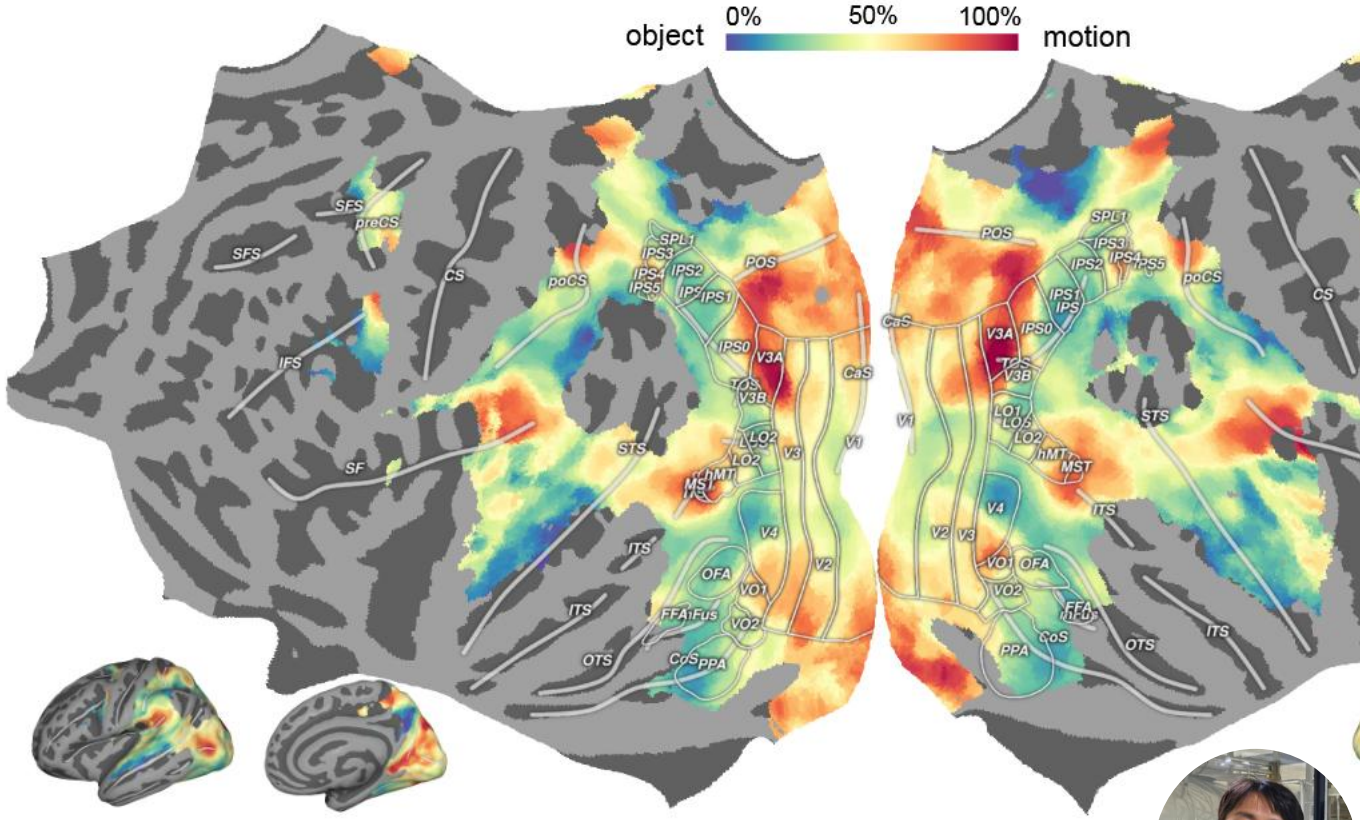
Tang et al. 2025

SOTA ML models predict brain activity with high fidelity

Anatomical hierarchy



Functional organization

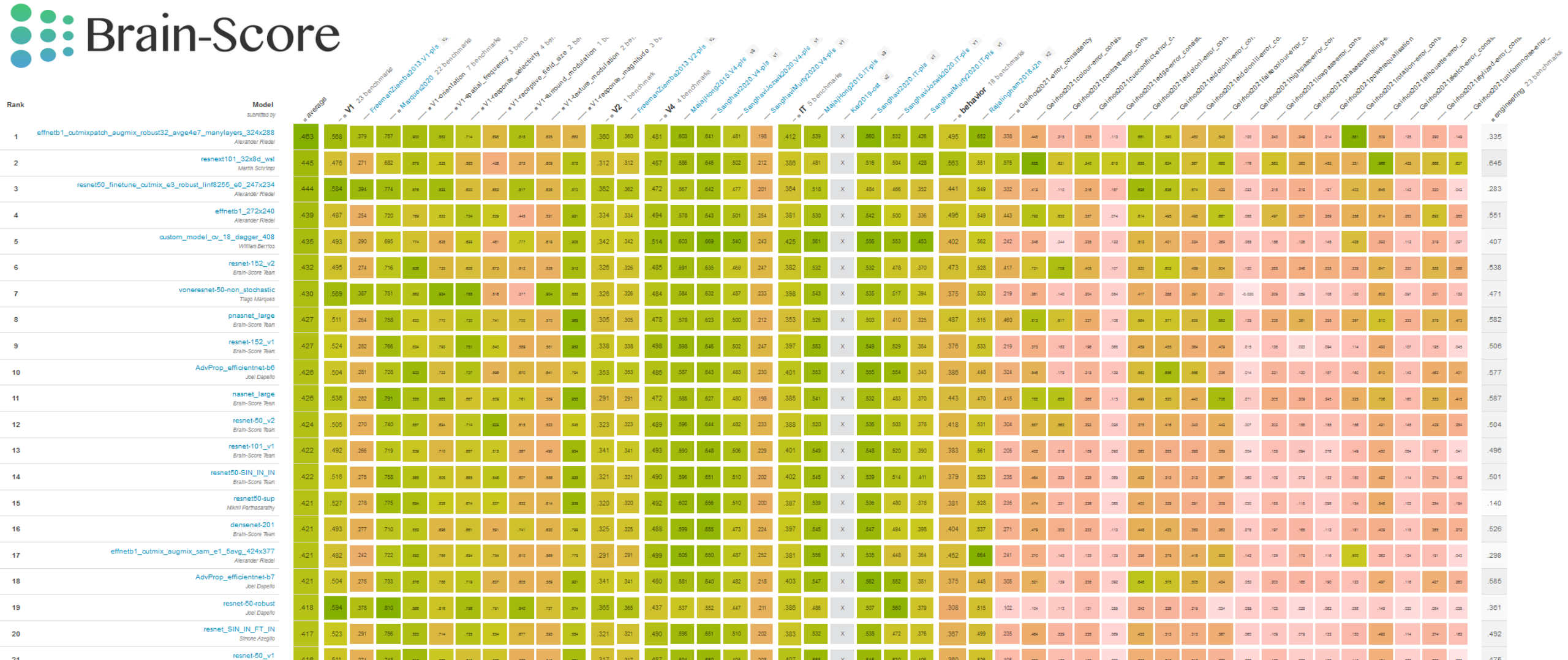


Tang et al. 2025

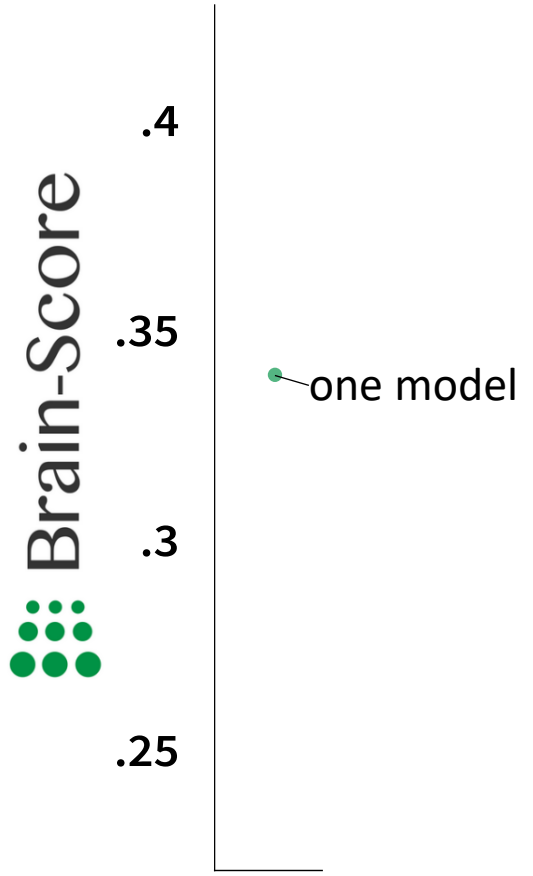
Integrative Benchmarking on Brain-Score

100+ brain & behavior benchmarks

Brain-Score

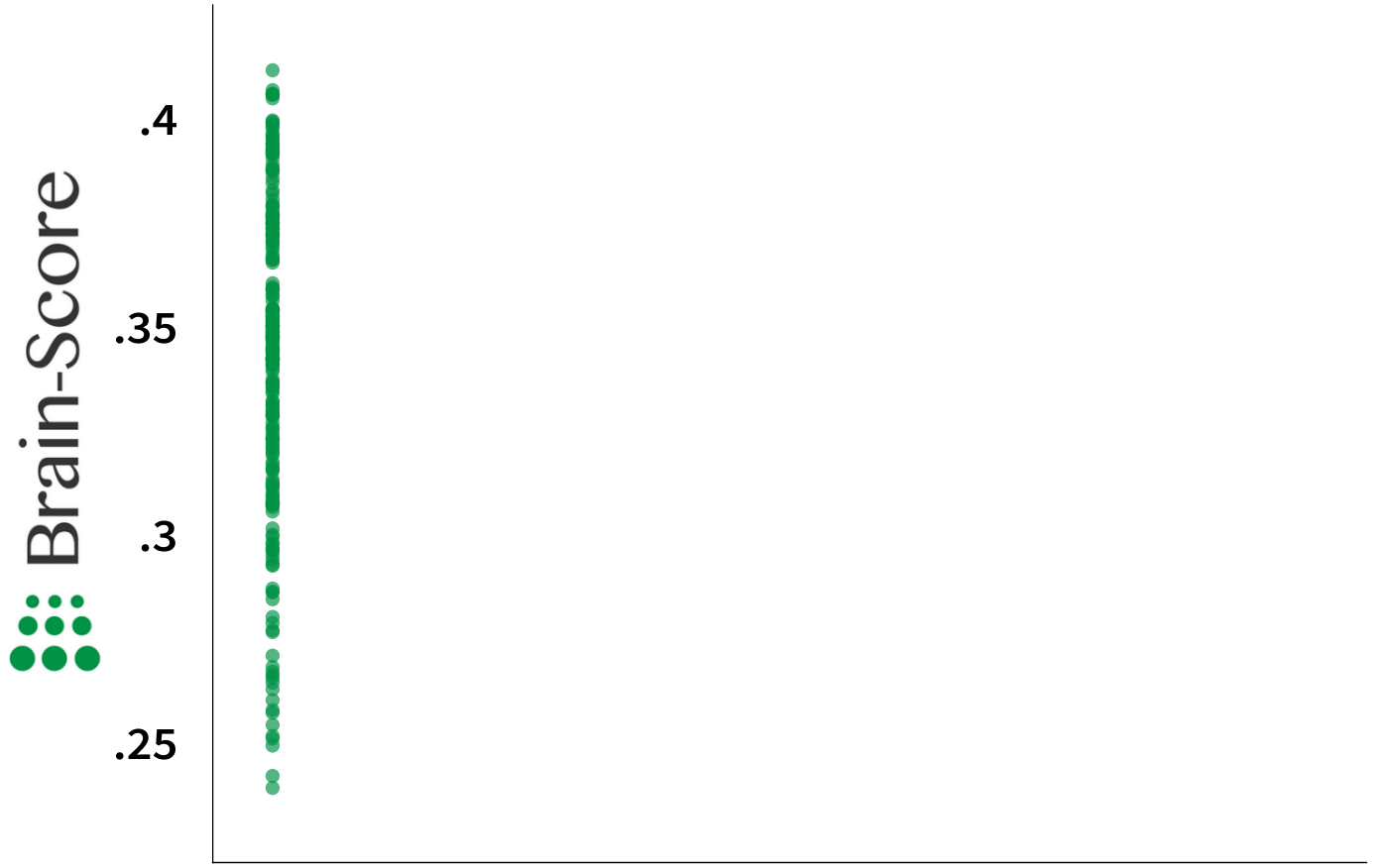


What explains the model differences?



Schrimpf, Kubilius*, et al. 2018*

What explains the model differences?

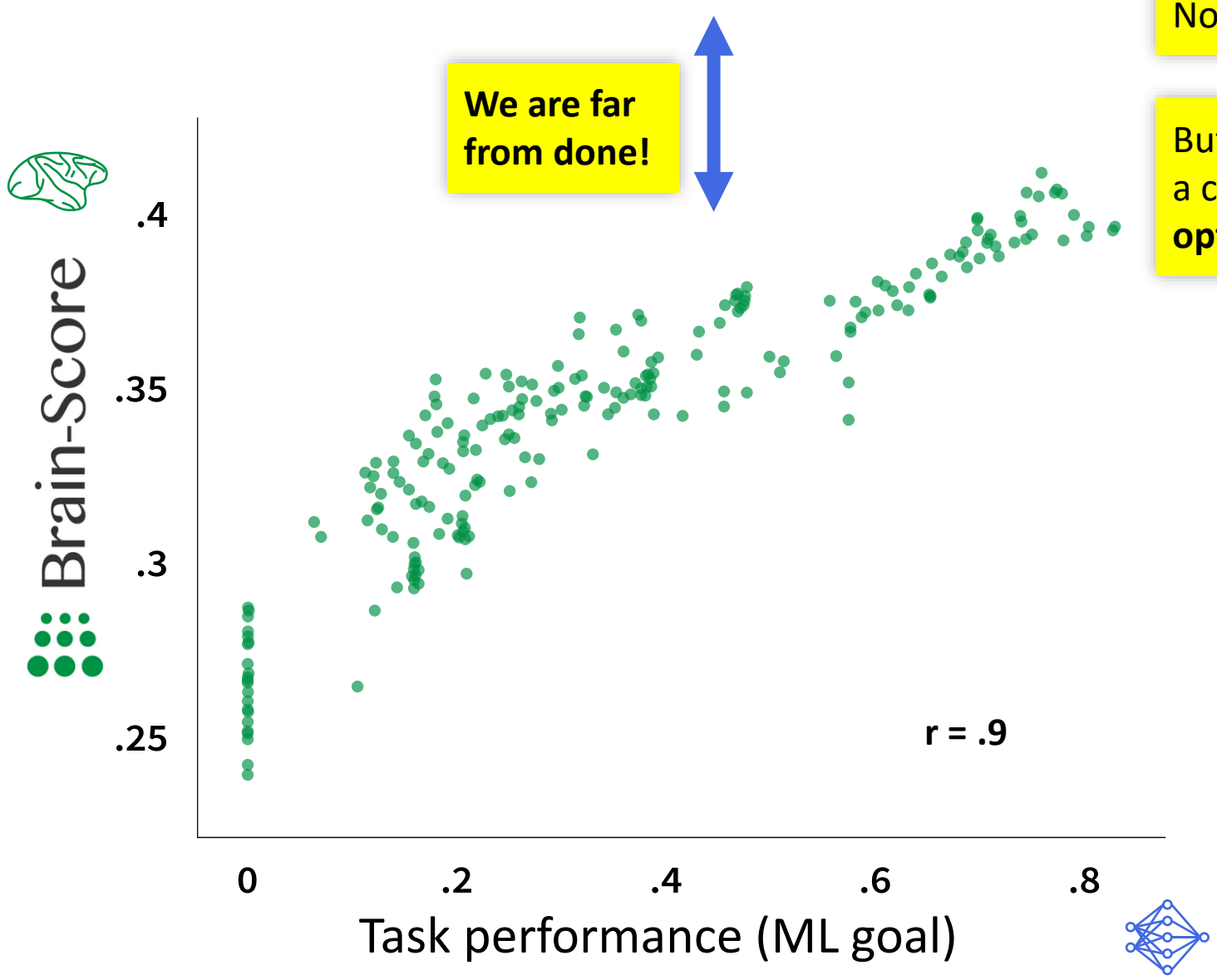


cf. Yamins, Hong*, et al. 2014*
Schrimpf, Kubilius*, et al. 2018*

What explains the model differences?

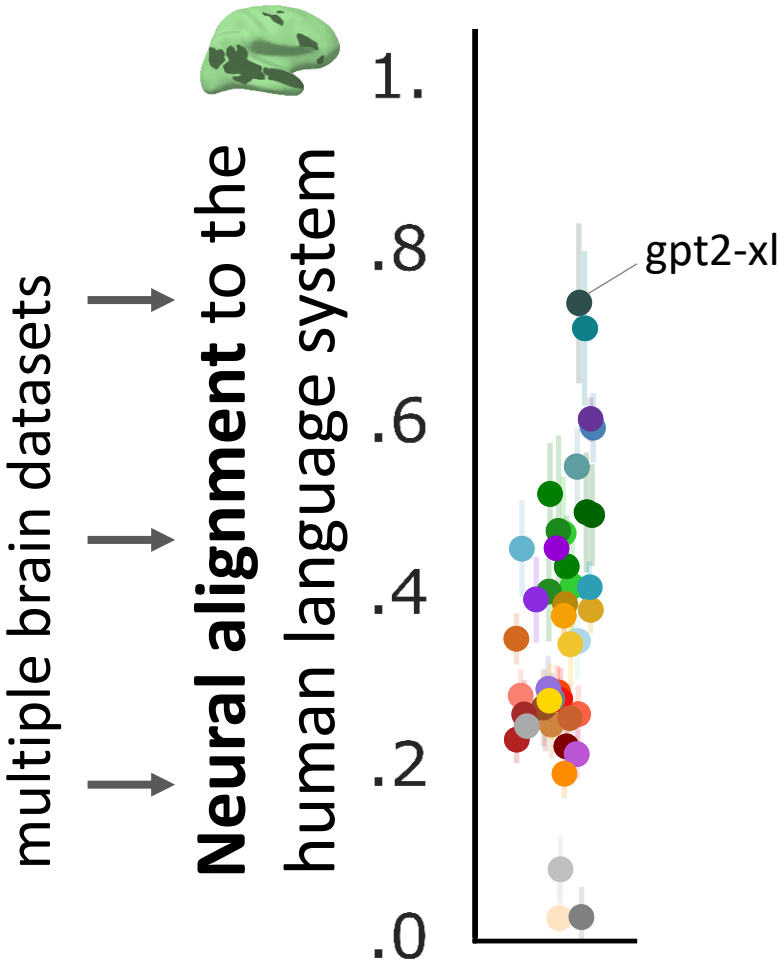
Not the small-scale circuits

But rather **task performance**, as a combination of **architecture + optimization**



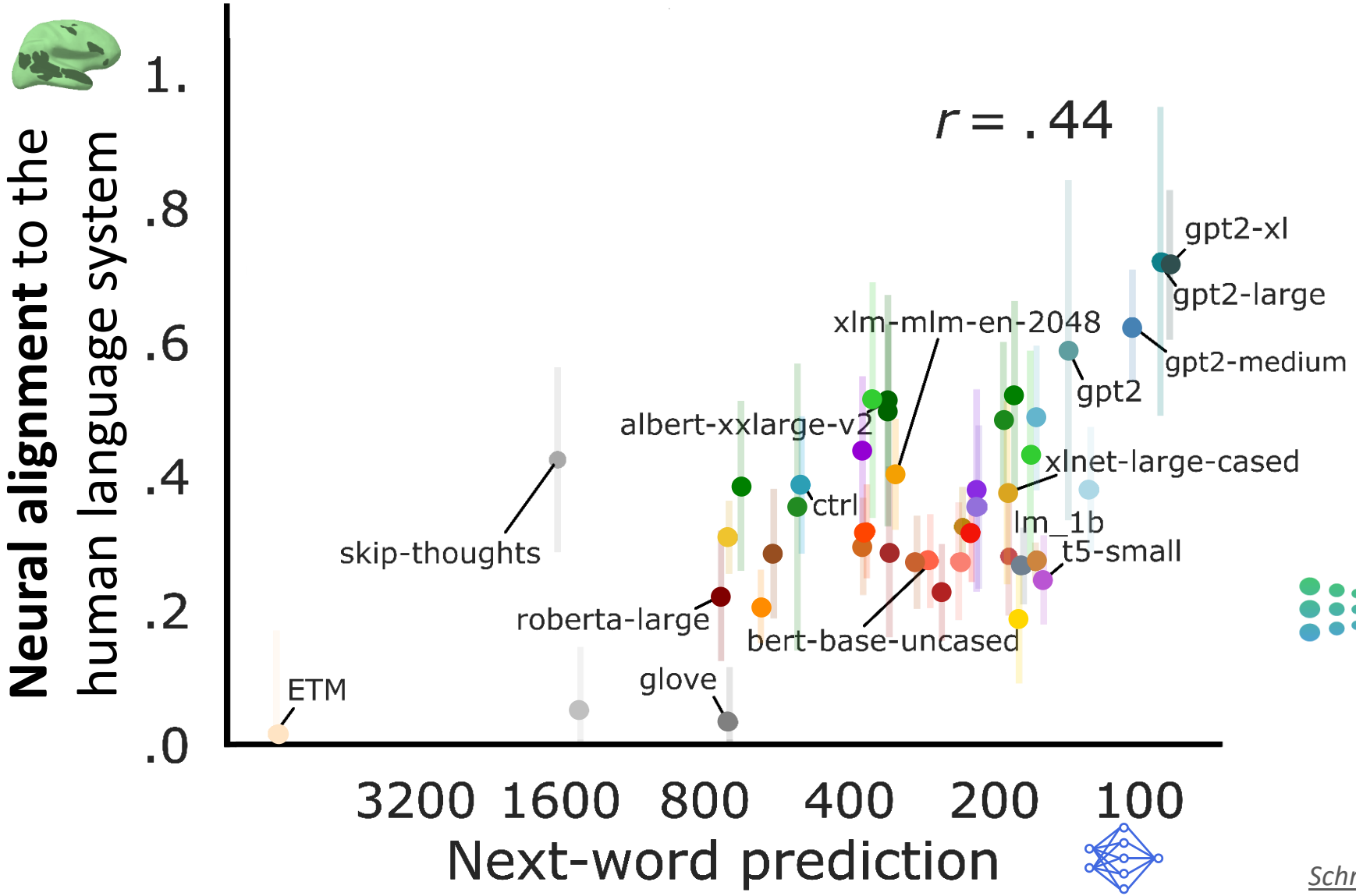
Schrimpf, Kubilius*, et al. (2018)*
Kubilius, Schrimpf*, et al. (2019)*

Particular ML language models predict the human language system



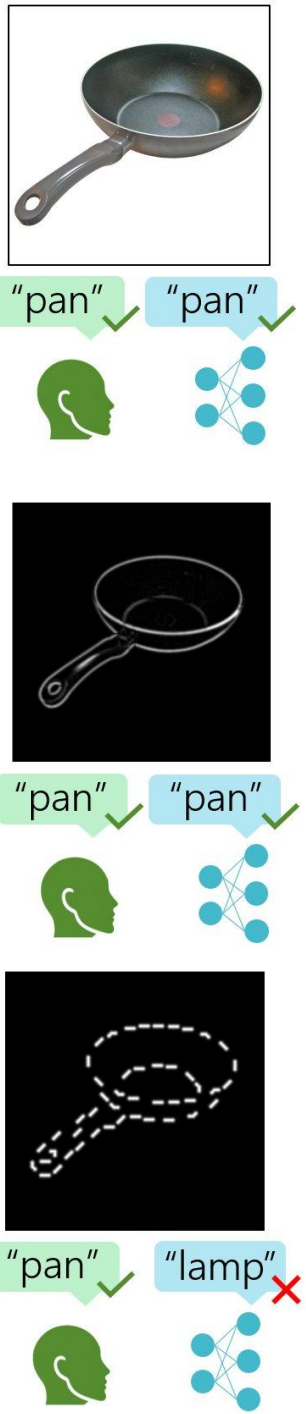
Schrimpf et al. (PNAS 2021)

The better models can predict the next word, the more brain-like they are

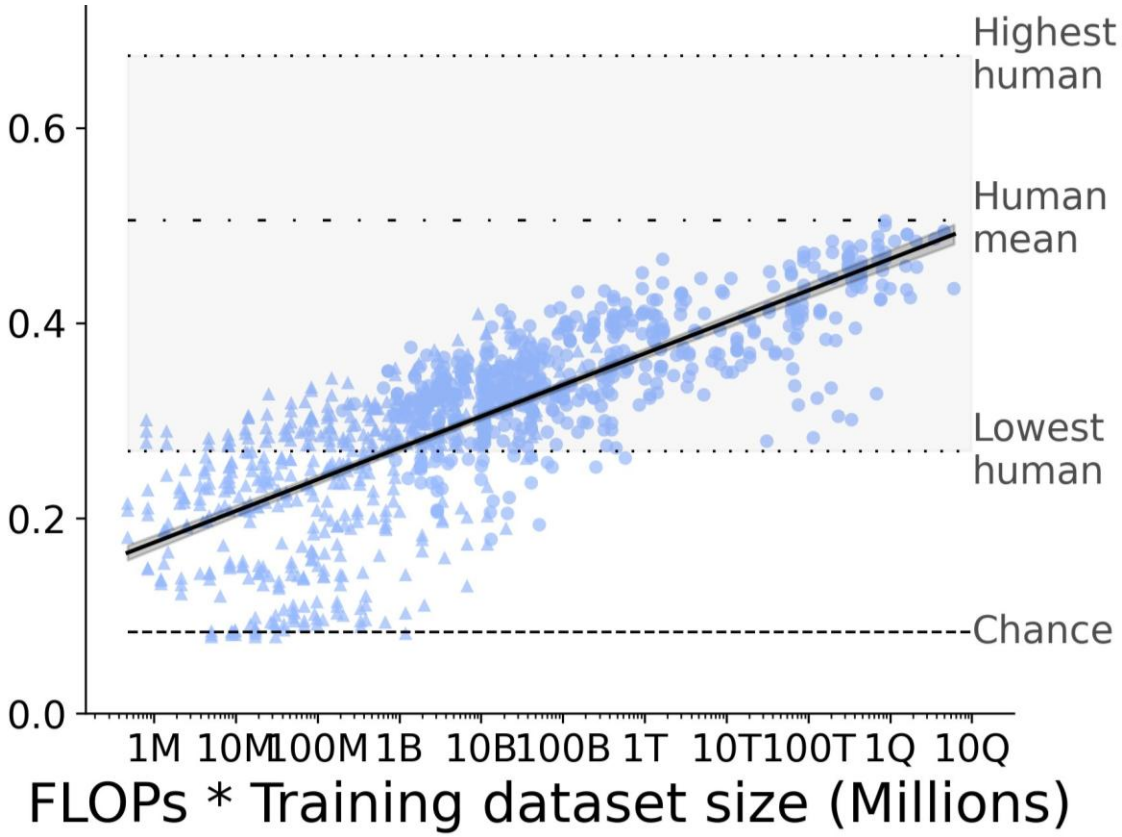


Brain-Score
.org/language

Schrimpf et al. (PNAS 2021)



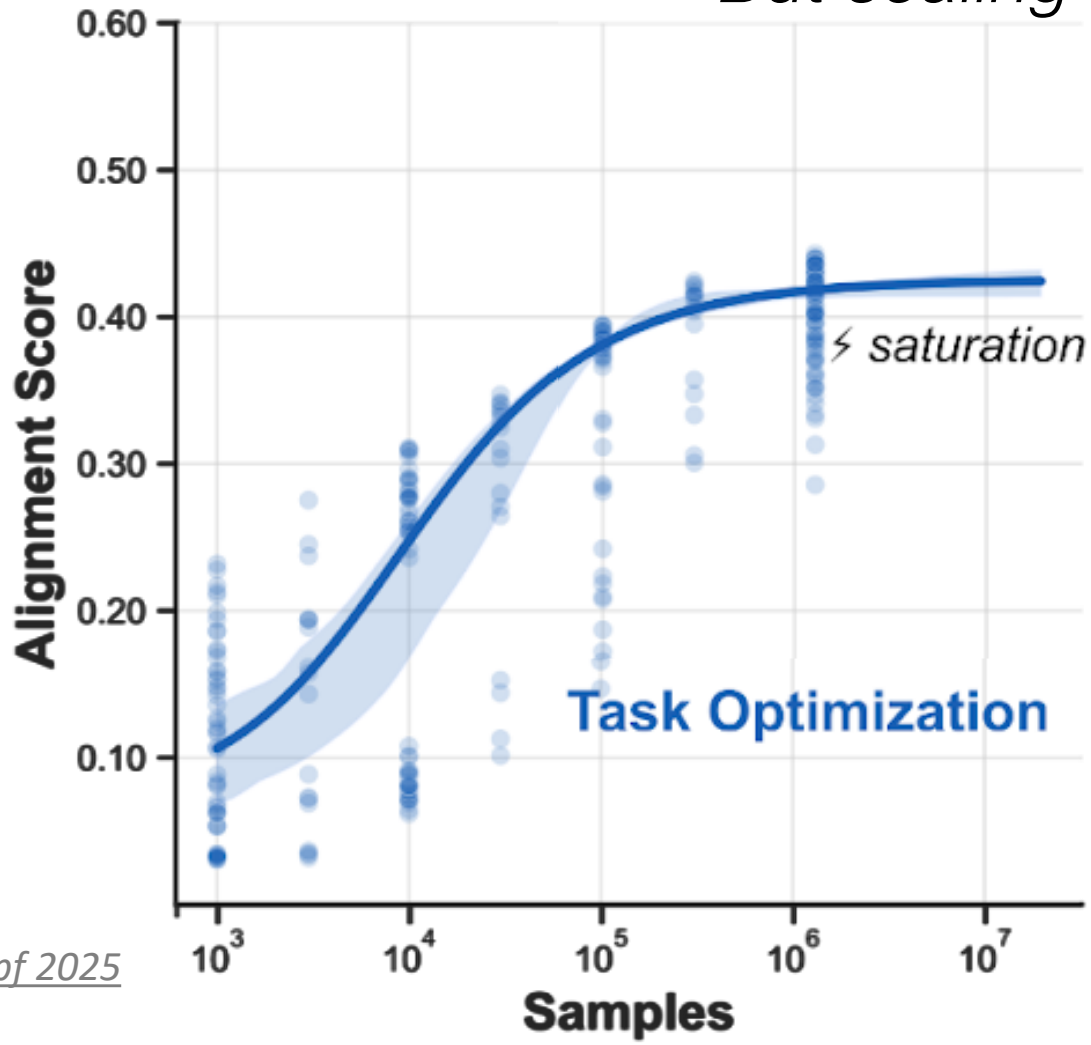
Most models are unable to recognize fragmented images. But increased scale improves behavioral alignment



Lonnqvist et al. 2025 (ICML)
see also Muttenthaler et al. 2023

Is ML-like scaling the solution for brain models?

No. *But scaling brain data seems promising.*

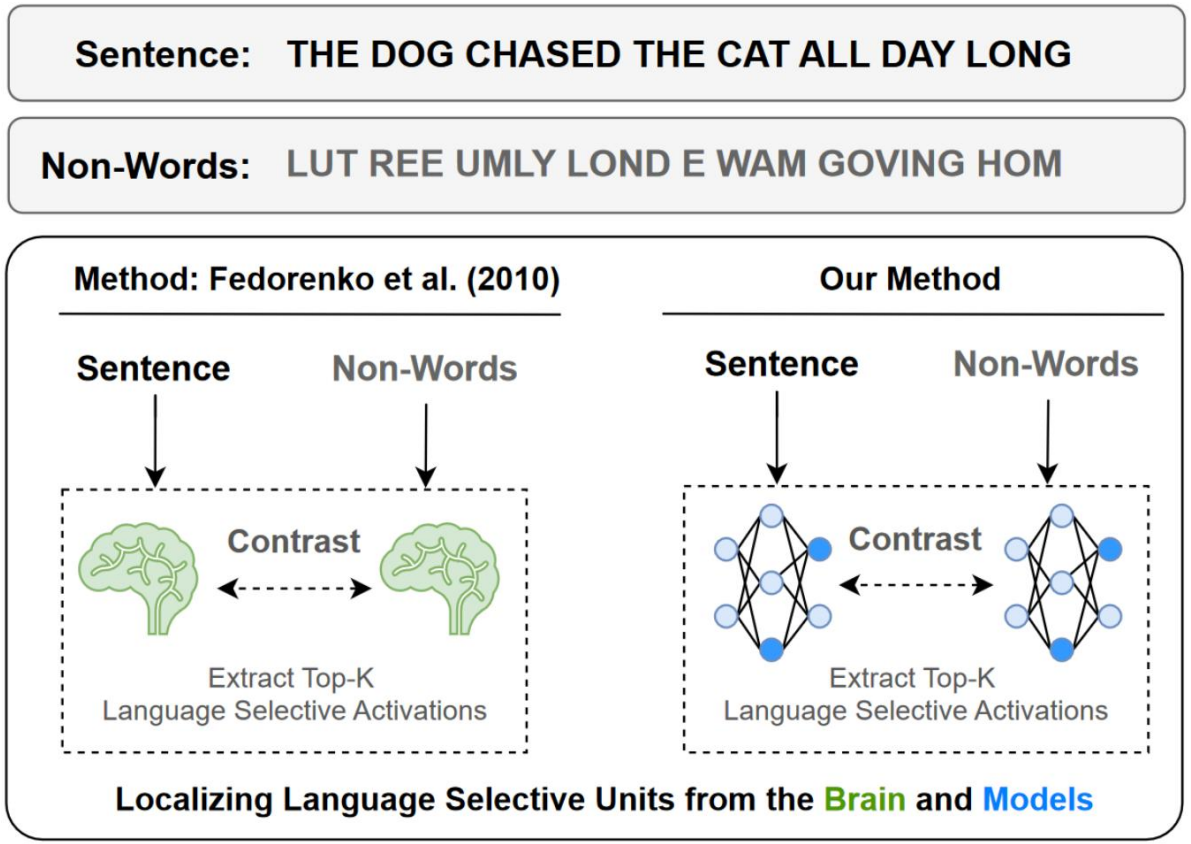


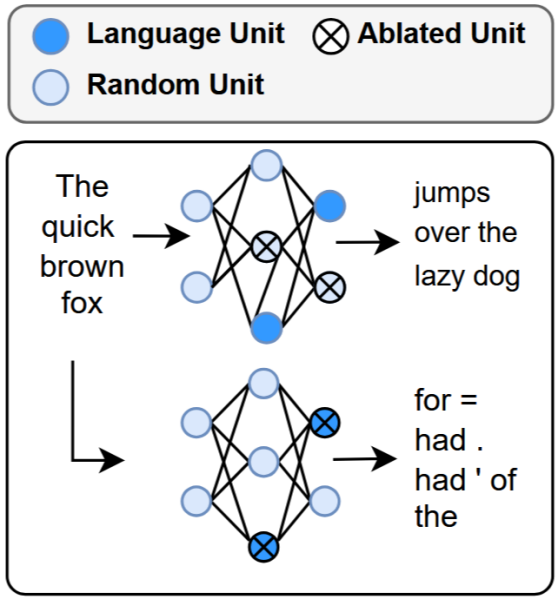
Gokce & Schrimpf 2025 (ICML Spotlight)

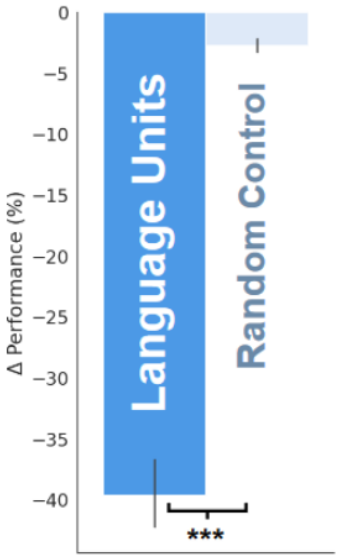
1. AI → Neuro: Particular neural network models are state-of-the-art at predicting brain function and human behavior.

2. Neuro → AI: Neuroscience has found very effective ways of identifying network function which are applicable to AI.

Neuroscience tools identify LLM mechanisms







"The quick brown fox..."

1% of units!

Model	Ablate Language Units	Ablate Random Units
Gemma-2B	11 liquido _ sota(.)uggoon3	jumped over the lazy lamb.
Phi-3.5-Mini-Instruct	AME.AME and:ough.. MAR	jumps over the lazy dog.
Falcon-7b	SomeSReadWhenISearchSome	jumps over the lazy dog.
Mistral-7B-v0.3	foxfool foolfoolfoolfool	jumps over the lazy dog.
LLaMA-3.1-8B-Instruct	_ of_An_O_of_An_O_of	jumps over the lazy dog.

AlKhamissi et al. 2024 (NAACL Oral)

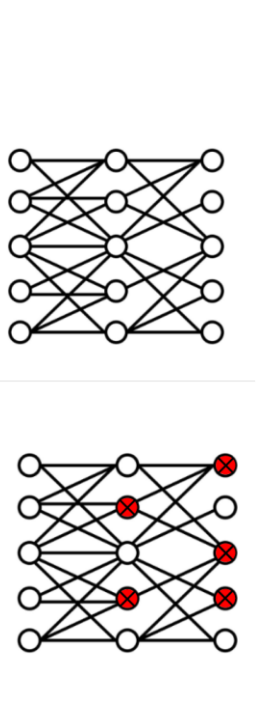
Localize causal mechanisms of brain disorders (e.g., dyslexia)

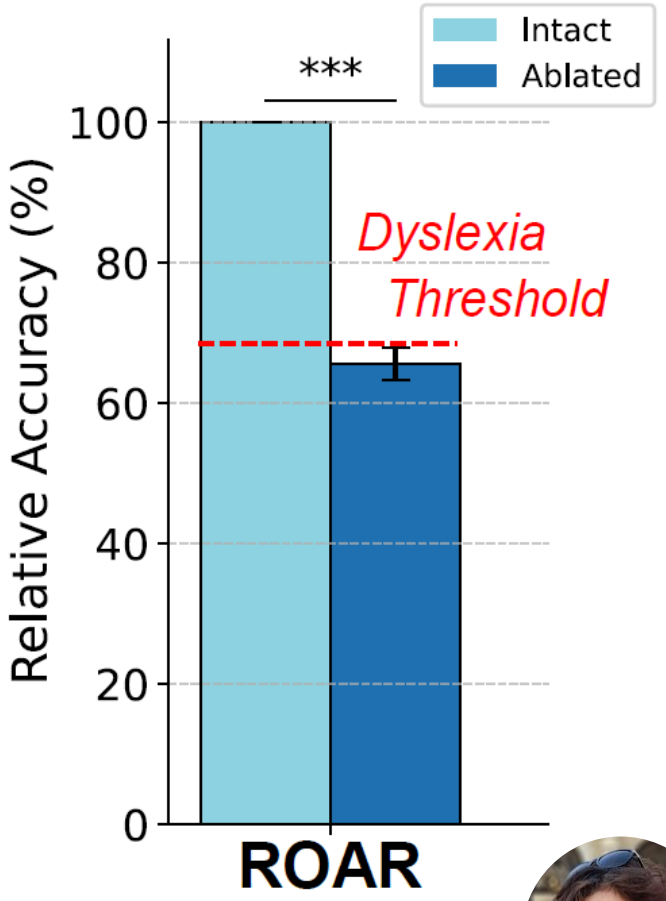
Is this a real word?

Ablood

The word in the image is a pseudo word. "Ablood" is not a real English word with a specific meaning.

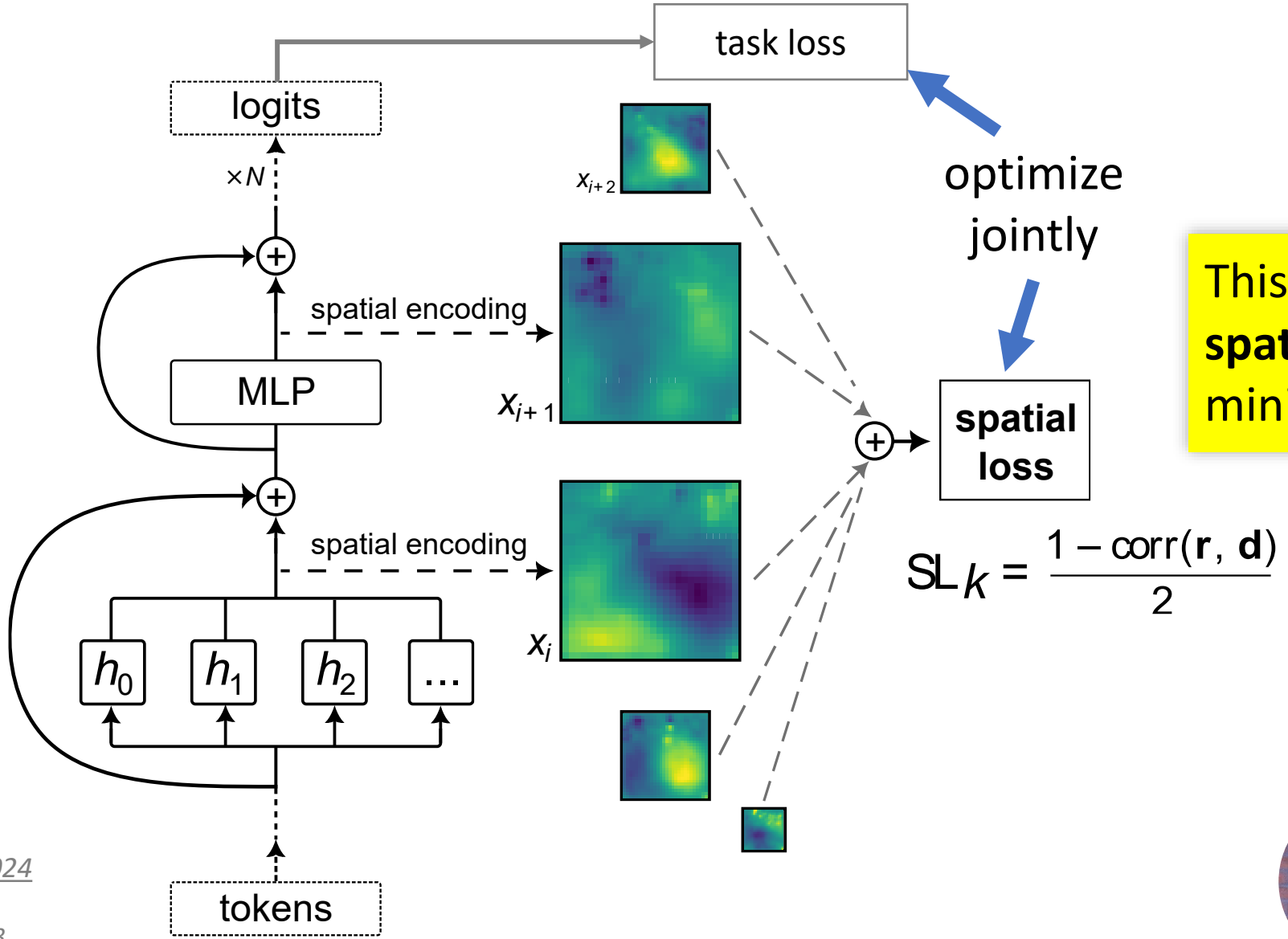
I cannot determine whether "Ablood" is a real word or a pseudo word without further context.





Honarmand et al. (in prep)

Topographic language models (TopoLM)

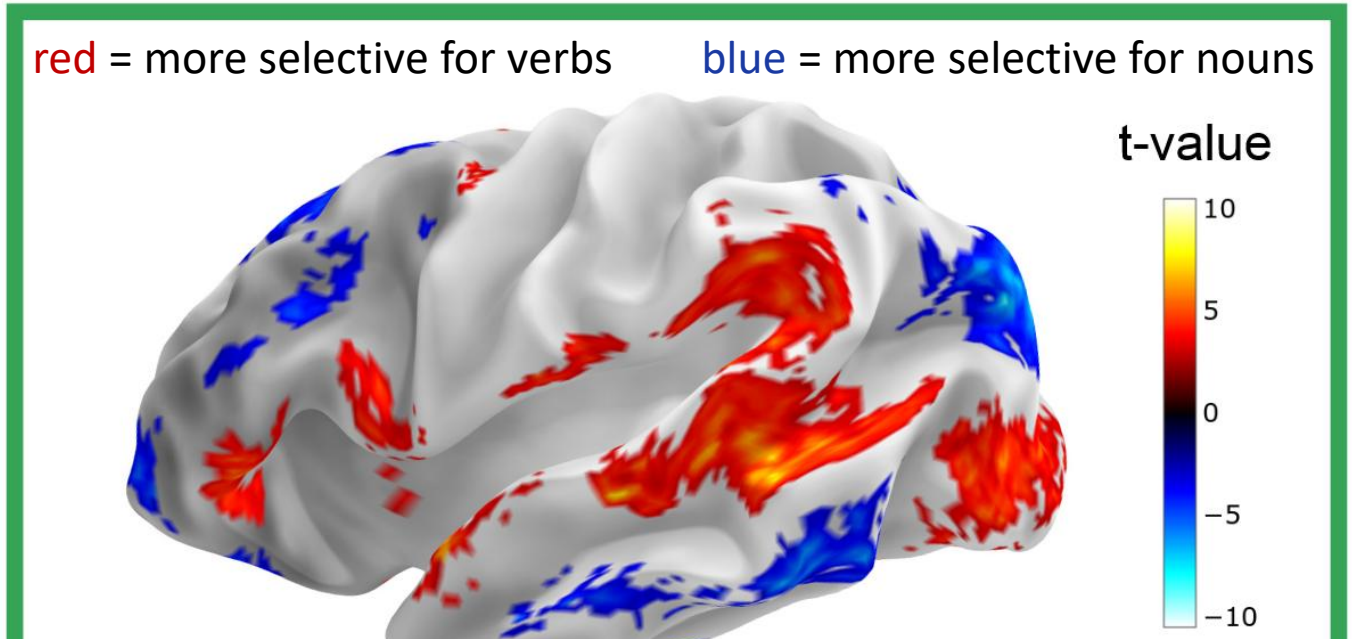


This loss encourages **spatial clustering** → minimizes wiring length

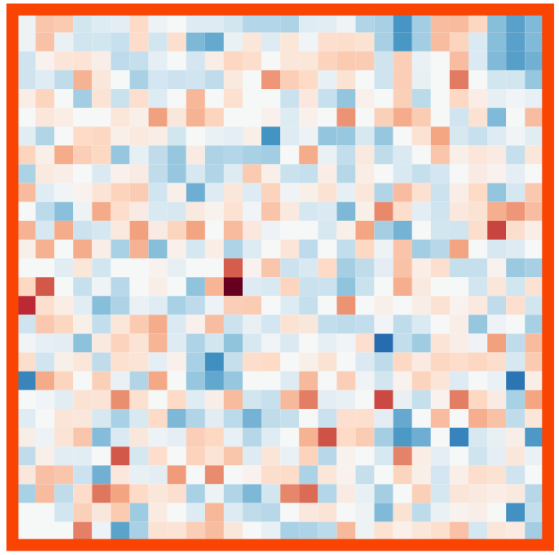
Rathi & Mehrer et al. 2024 (ICLR Oral)
See also Margalit et al. 2023

TopoLM exhibits brain-like clustering of verbs and nouns

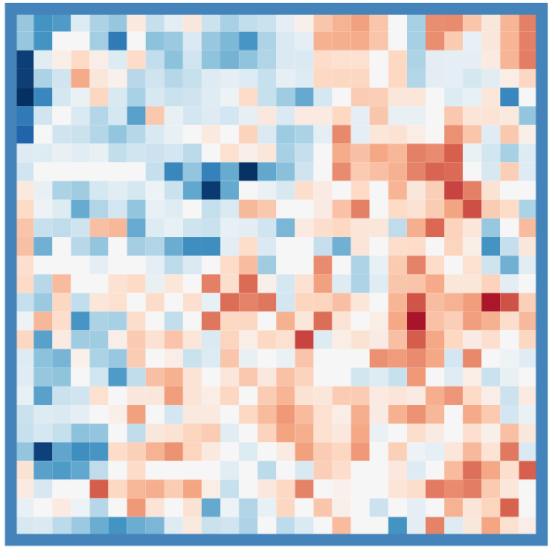
Neural Data — Hauptman et al. 2024



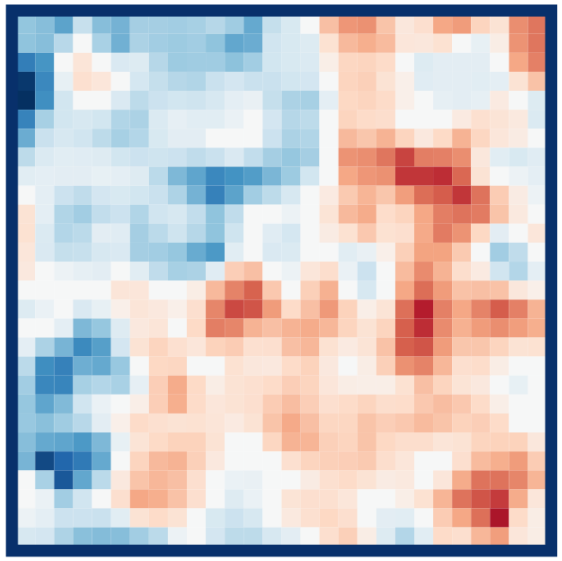
Non-topographic
no fMRI readout sampling

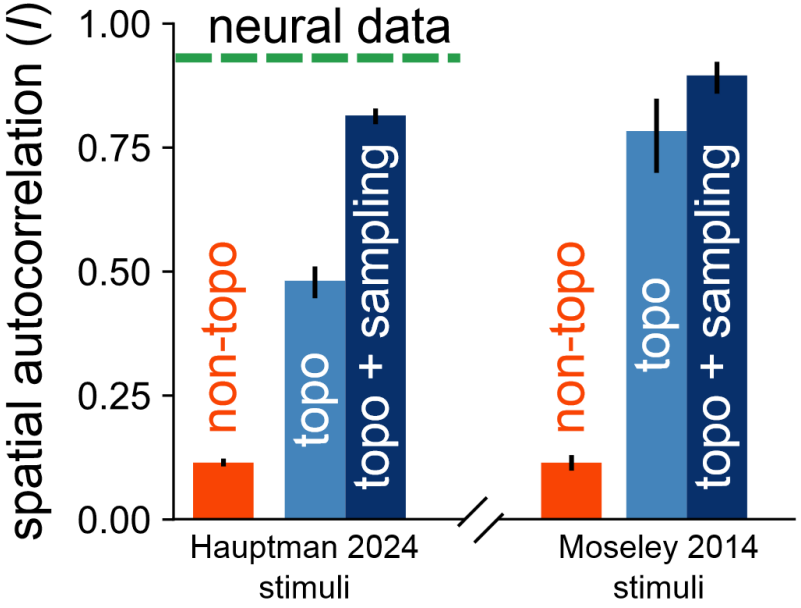


Topographic
no fMRI readout sampling



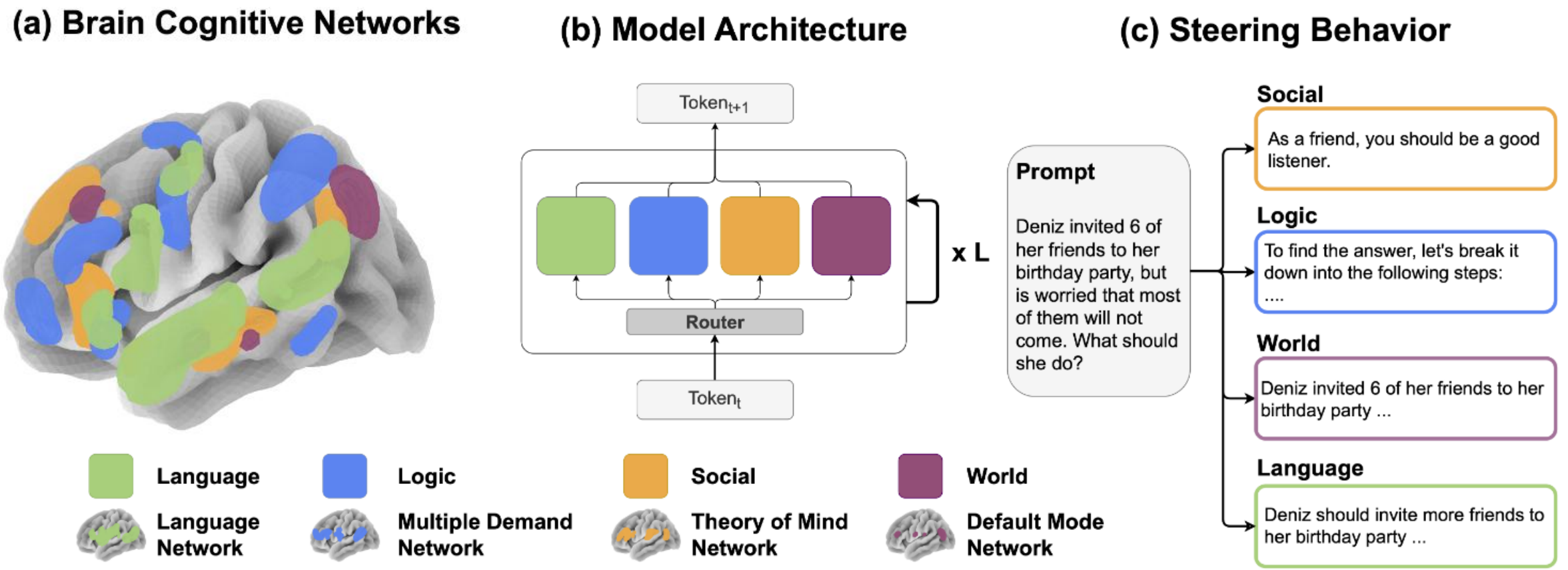
Topographic
+ fMRI readout sampling





Rathi & Mehrer et al. 2024

Brain-Like Cognitive Reasoning



Fedorenko et al. 2024

AlKhamissi et al. 2025

1. AI → Neuro: Particular neural network models are state-of-the-art at predicting brain function and human behavior.

2. Neuro → AI: Neuroscience has found very effective ways of identifying network function which are applicable to AI.

3. NeuroAI → Clinic: Digital twins to help treat brain disorders.

We can use brain-aligned LLMs to noninvasively control neural activity

Neural alignment to the human language system

1.
.8
.6
.4
.2
.0

Can a good model of the brain tell us how to control neural activity?

gpt2-xl

***Drive:* 250 sentences**

***Suppress:* 250 sentences**

Record brain responses to novel sentences in new participants

Drive

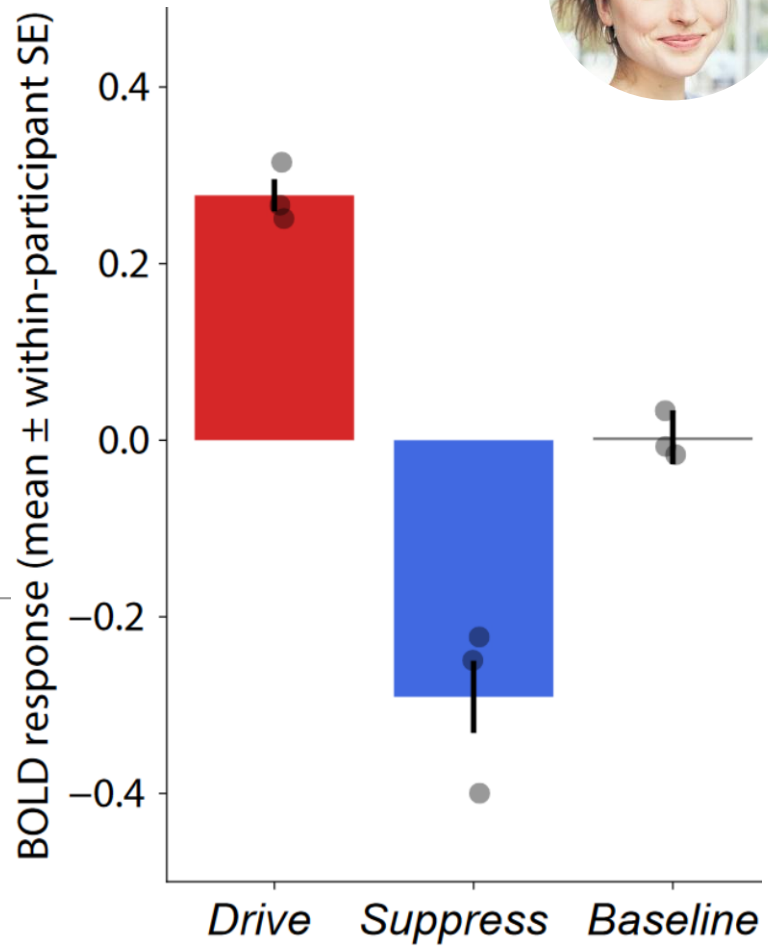
Sentences identified t
response in the lang

Changing PhD group: Yes
Notice how you reacted to
Add, some feminists are c
Jiffy Lube of -- of therapies
People on Insta Be Like, "
Buy sell signals remains a
Turin loves me not, nor wi
URL right, or report review

Suppress

Sentences identified to elicit minimal response in the language network

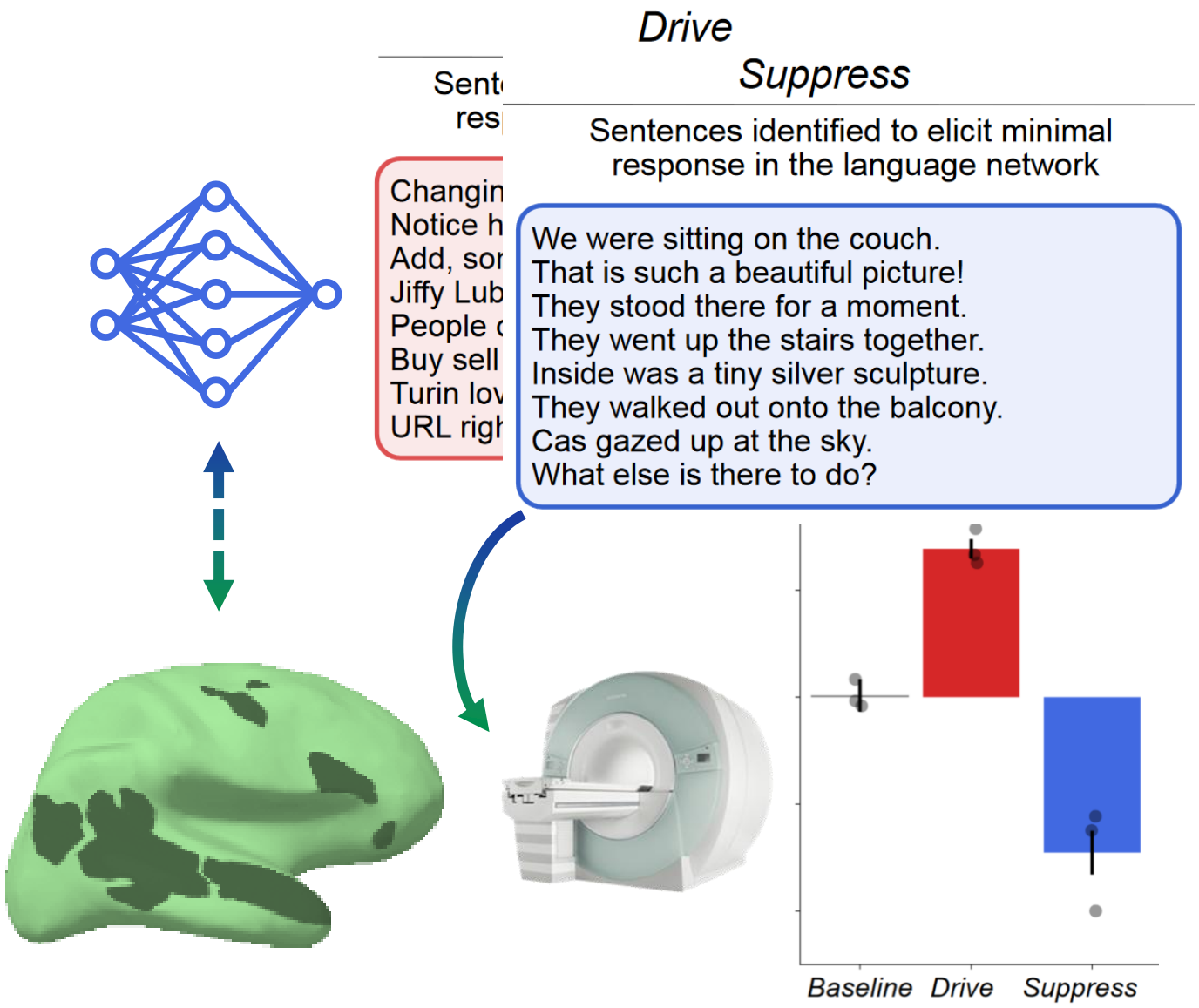
We were sitting on the couch.
That is such a beautiful picture!
They stood there for a moment.
They went up the stairs together.
Inside was a tiny silver sculpture.
They walked out onto the balcony.
Cas gazed up at the sky.
What else is there to do?

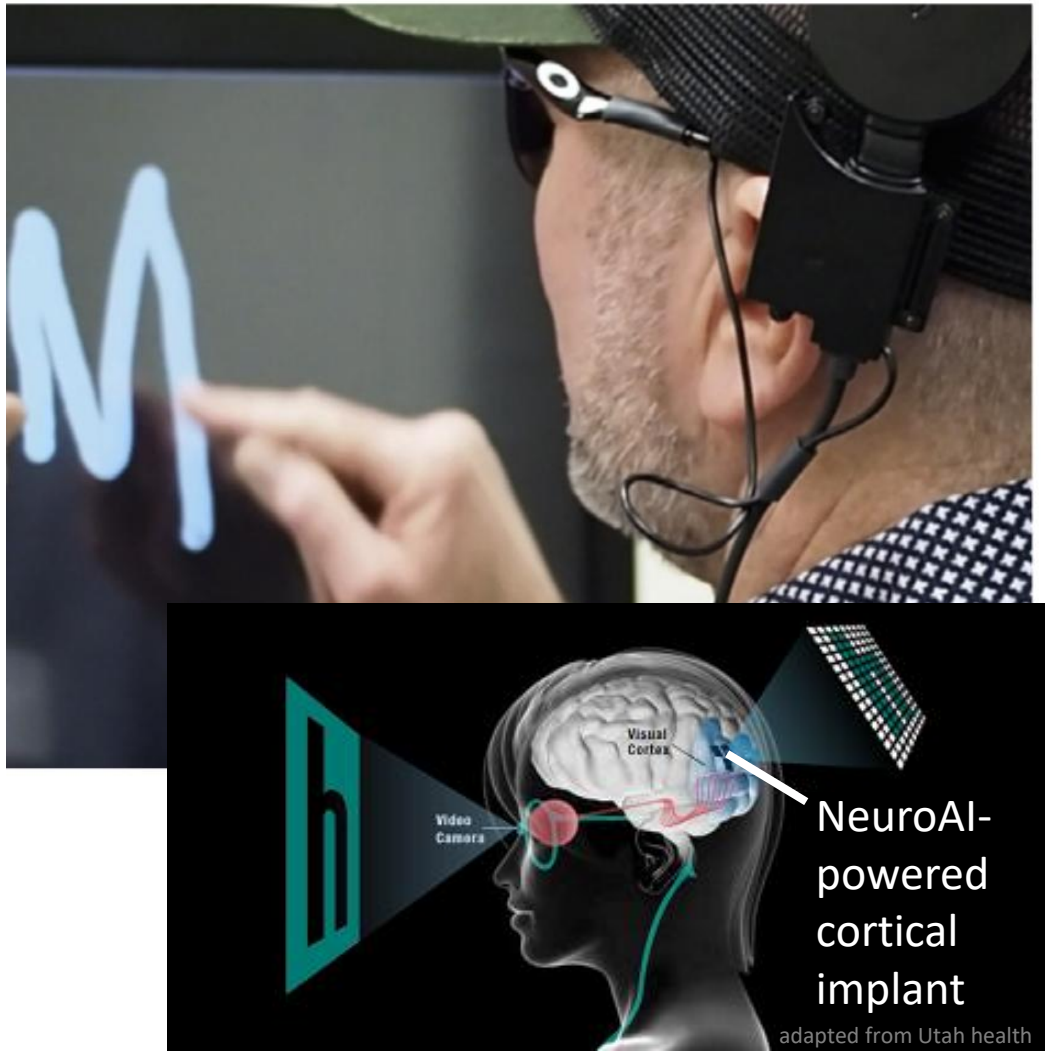


Tuckute et al. (Nature Human Behavior 2023)

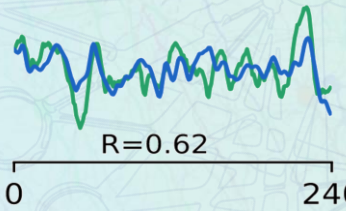
See also Bashivan*, Kar*, et al. (Science 2019)

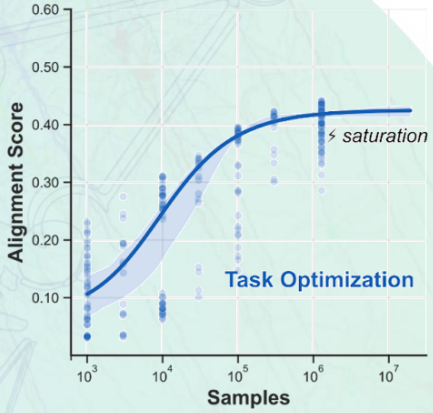
NeuroAI models can control brain activity

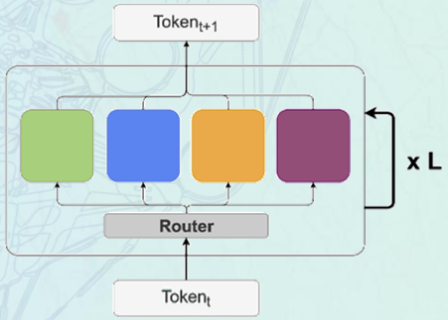


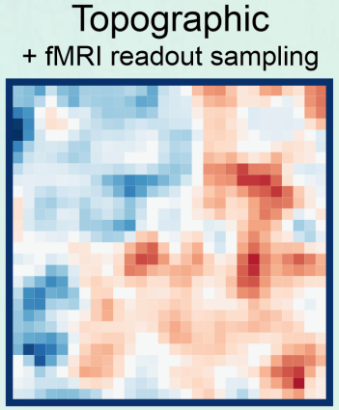


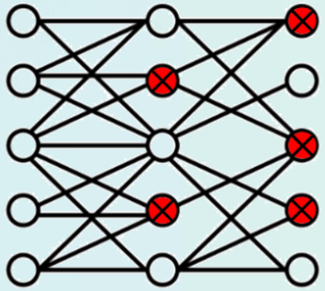
Tuckute et al. (2023)
Beauchamp et al. (2020) Chen et al. (2021) Schrimpf et al. (2024)

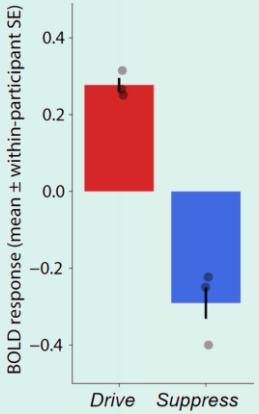












NeuroAI Lab

1. AI → Neuro: Particular neural network models are state-of-the-art at predicting brain function and human behavior.

Brain-Score

2. Neuro → AI: Neuroscience has found very effective ways of identifying network function which are applicable to AI.

3. NeuroAI → Clinic: Digital twins to help treat brain disorders.