

Brain-Score

Martin Schrimpf

EPFL
Neuro X Institute



martin.schrimpf
@epfl.ch



@martin_schrimpf



mastodon.social/
@mschrimpf



Katherine
Fairchild



David Tang



Mike Ferguson



Hannes
Mehrer



Ayu Marliawaty
I Gusti Bagus



Khai Loong
Aw



Badr
AlKhamissi



Franziska Geiger Jim DiCarlo



Greta Tuckute



Carina Kauf



Idan Blank



Ev Fedorenko



Anna Ivanova



Eghbal Hosseini



Nancy Kanwisher Josh Tenenbaum



Joel Dapello



Tiago Marques



Kohitij Kar



Chengxu Zhuang



Rishi Rajalingham



Michael Lee



Eshed
Margalit



Robert Ajemian



Jonas Kubilius



Paul McGrath



Corey Ziembra



Tony Movshon



Dan Yamins



Pouya Bashivan



Kailyn Schmidt Jon Prescott-Roy



Ratan Murty

Funding:

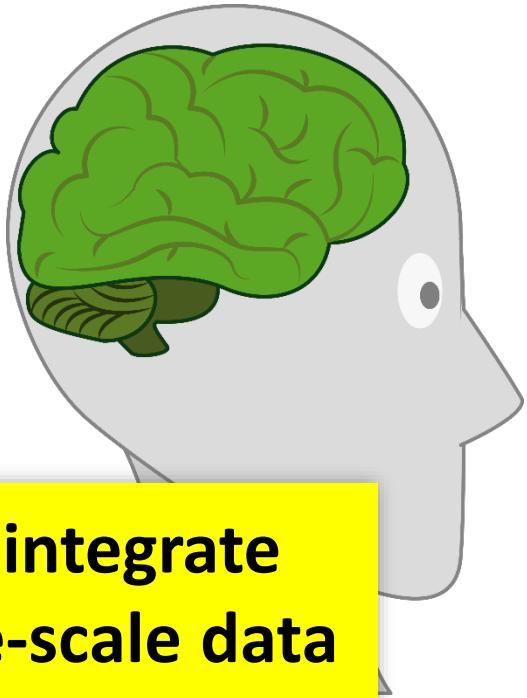


MCGOVERN INSTITUTE
FOR BRAIN RESEARCH



EPFL

Goal: Model Natural (Human) Intelligence and the Underlying Neural Mechanisms

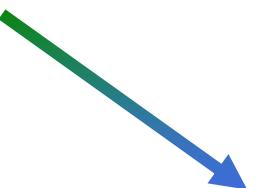
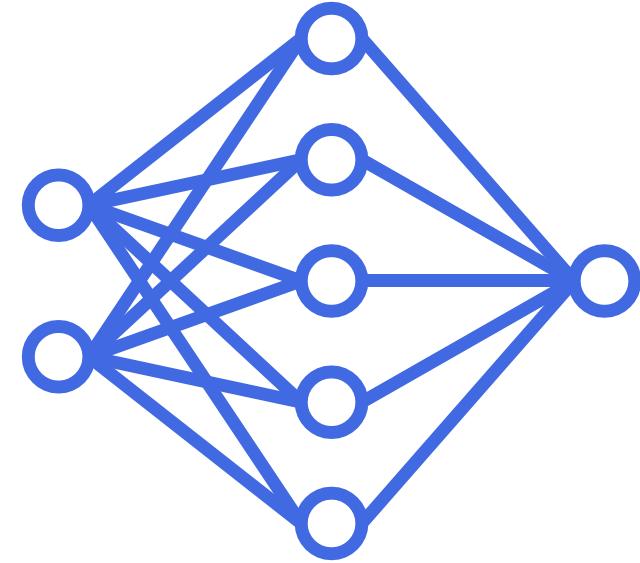


Models integrate over large-scale data from multiple sources

Computational understanding of human intelligence



Next-generation intelligence algorithms



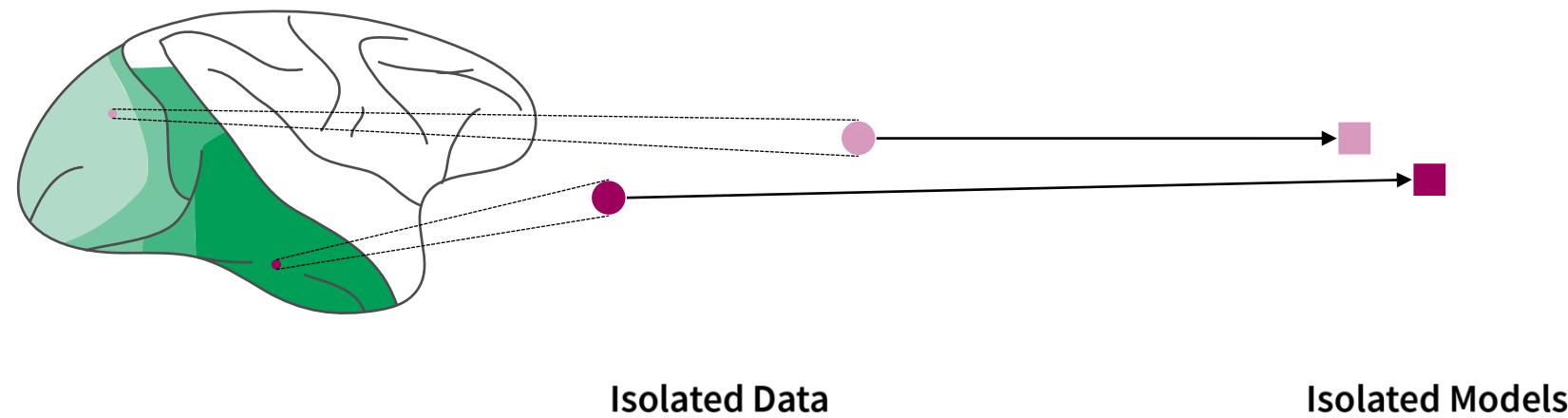
Future clinical applications



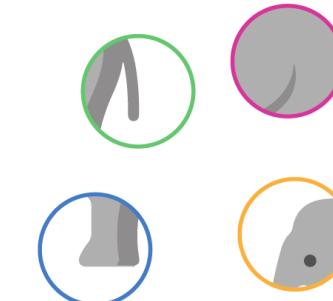
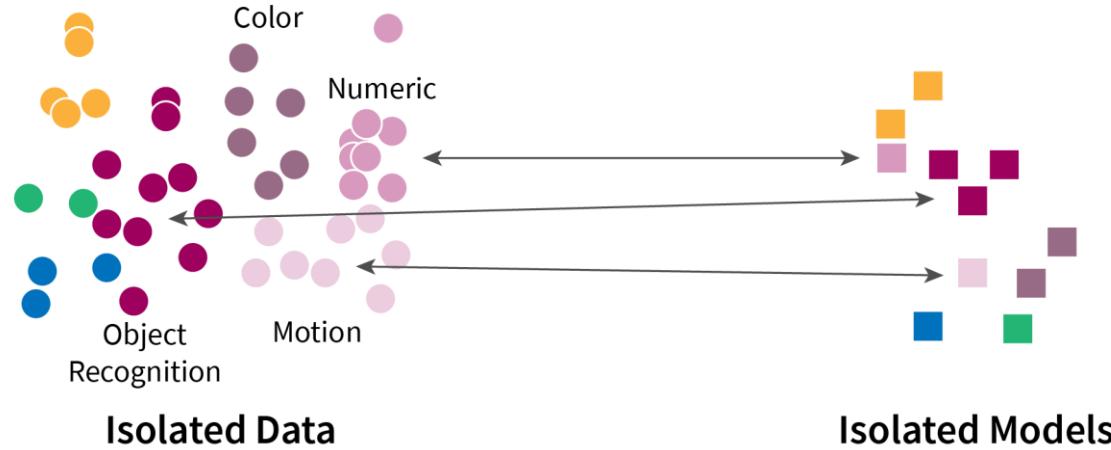
Key messages today

1. Data alone is not enough. We need **experimental benchmarks at scale** to model the brain. These make research more **efficient, and accessible** to newcomers.
2. Current models of human vision and language are **decent approximations of brain and behavior**. We can use these models to prototype experiments.

Piecewise Efforts

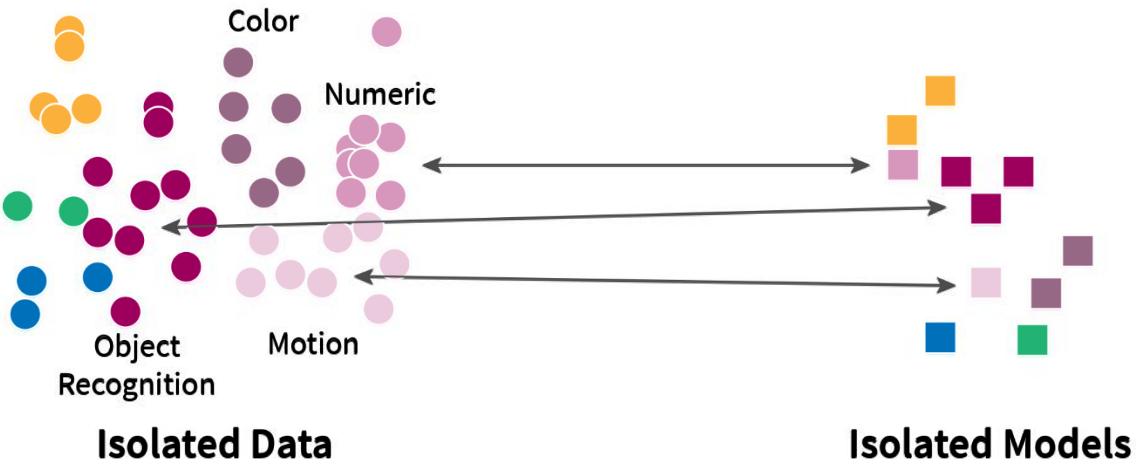


Piecewise Efforts



**These are necessary first steps!
But insufficient for a unified
model by themselves**

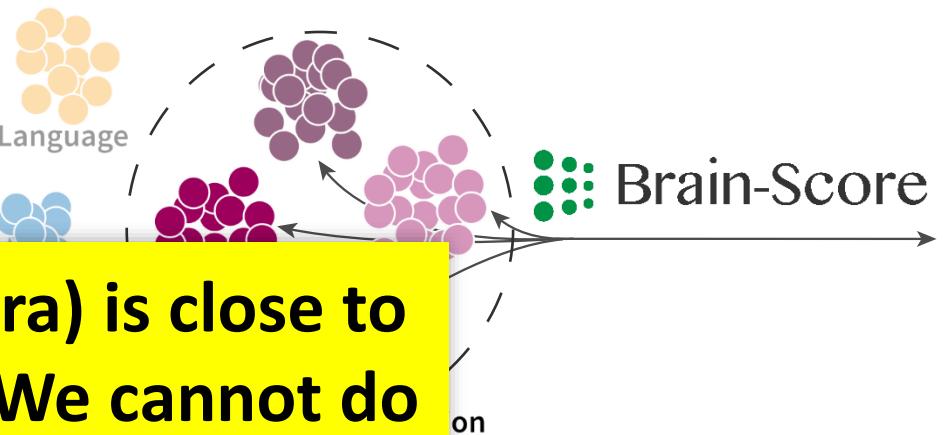
Piecewise Efforts



Isolated Models

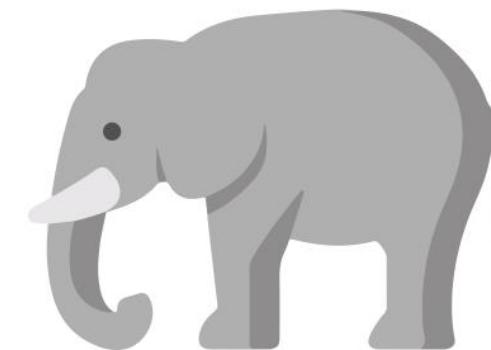


Integrative Benchmarking

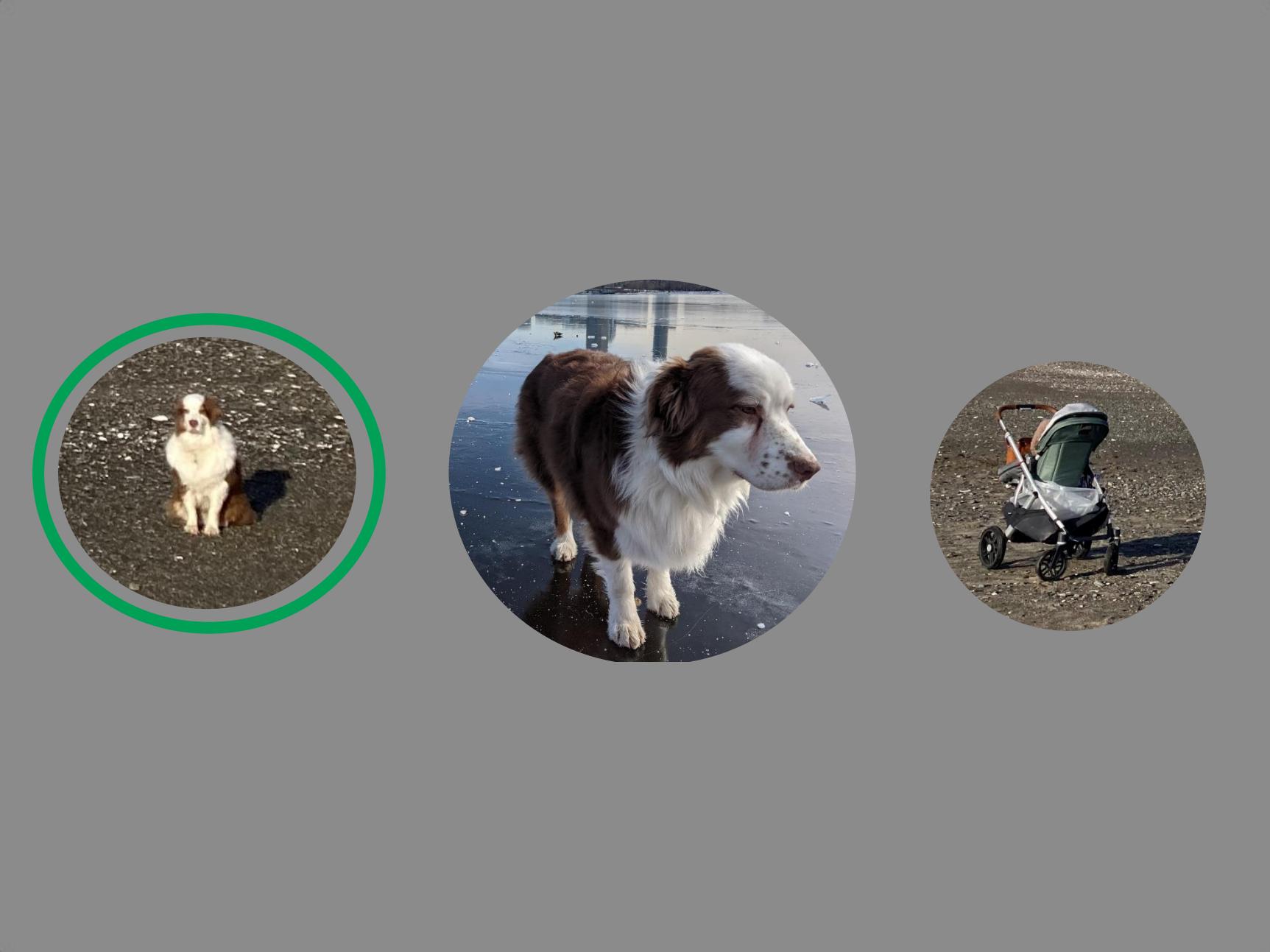


"Level 0" (Satra) is close to useless here. We cannot do anything without metadata

System Models

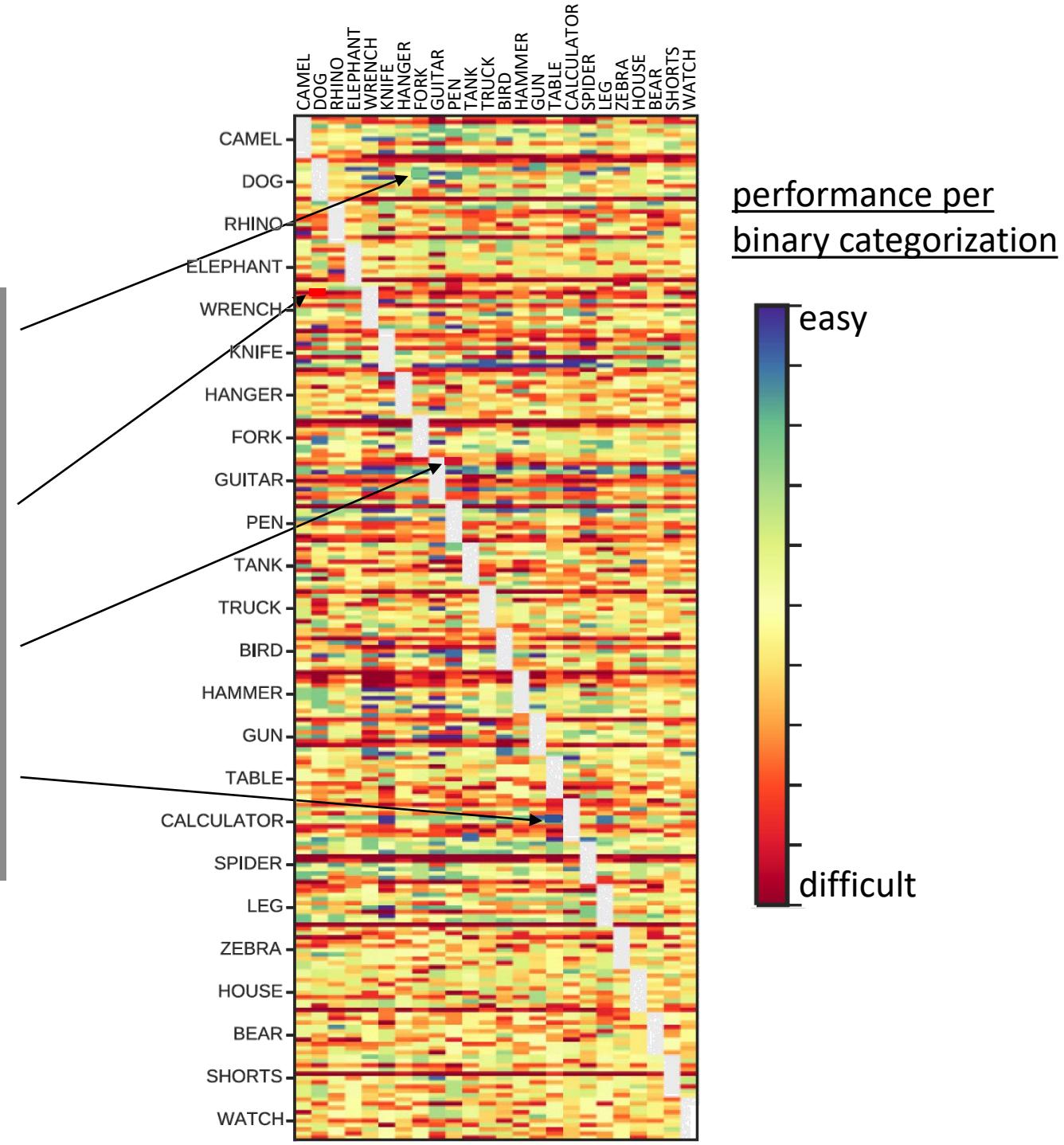


Behavioral benchmark

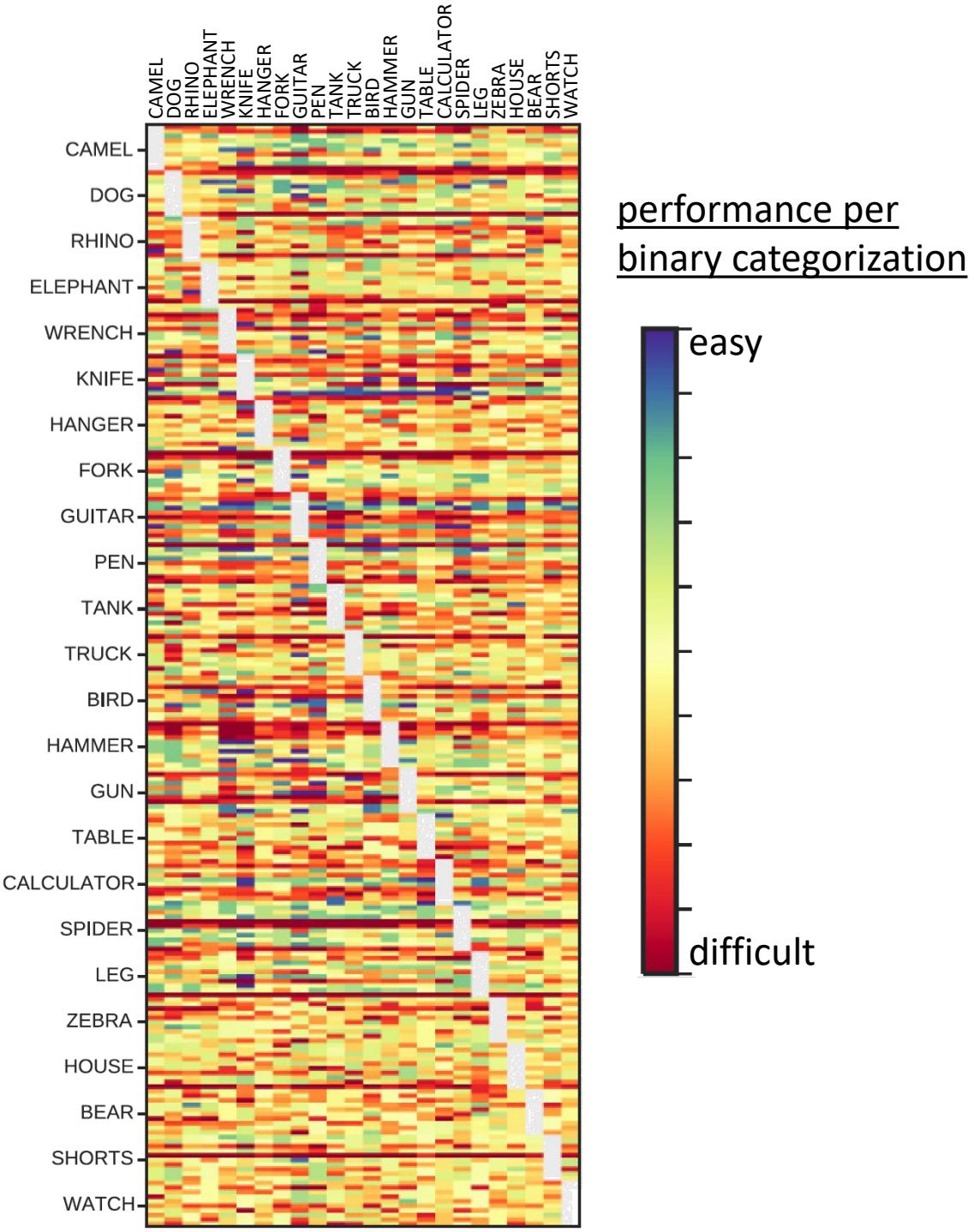


Behavioral benchmark

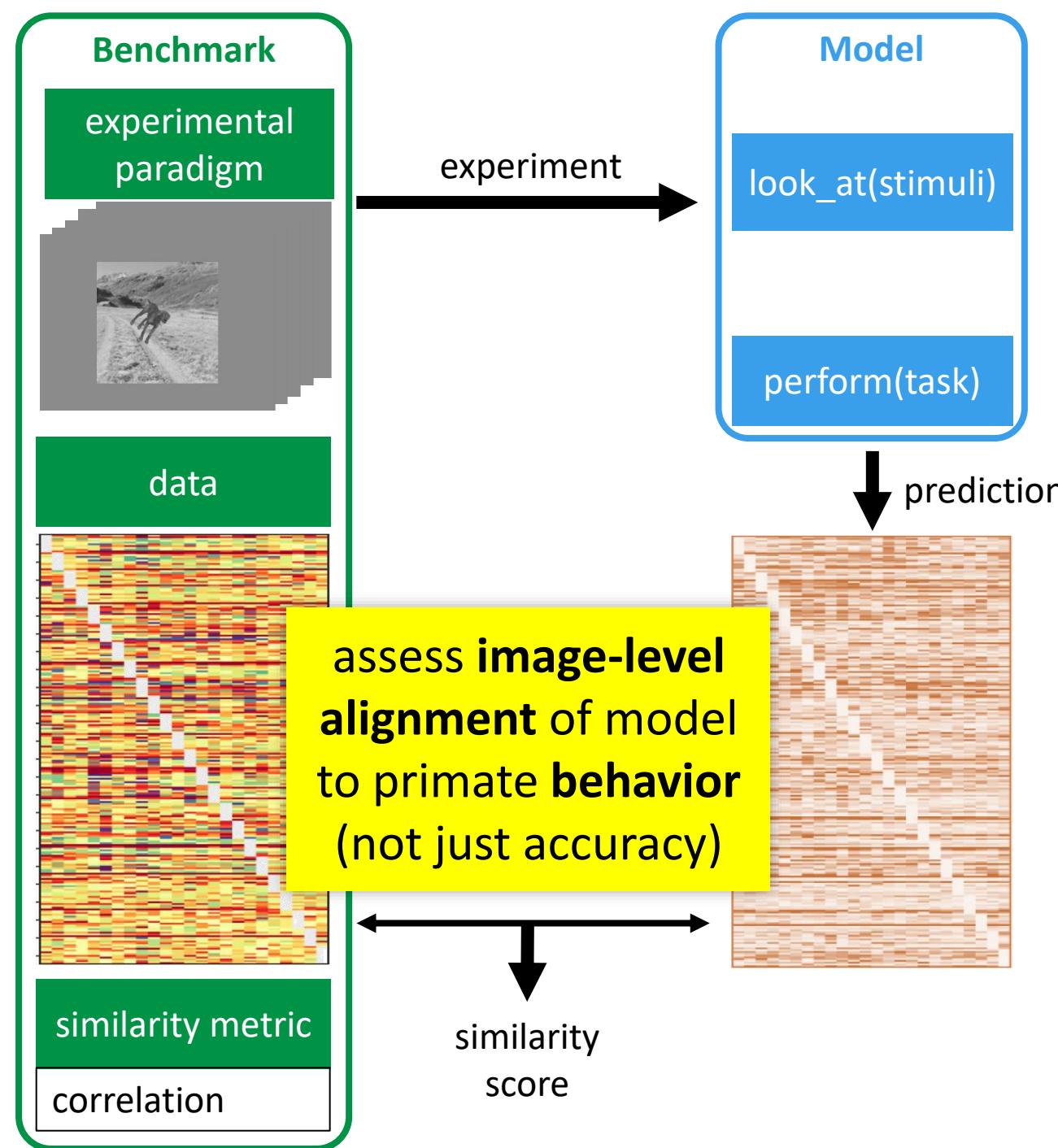
Behavioral benchmark



Behavioral benchmark

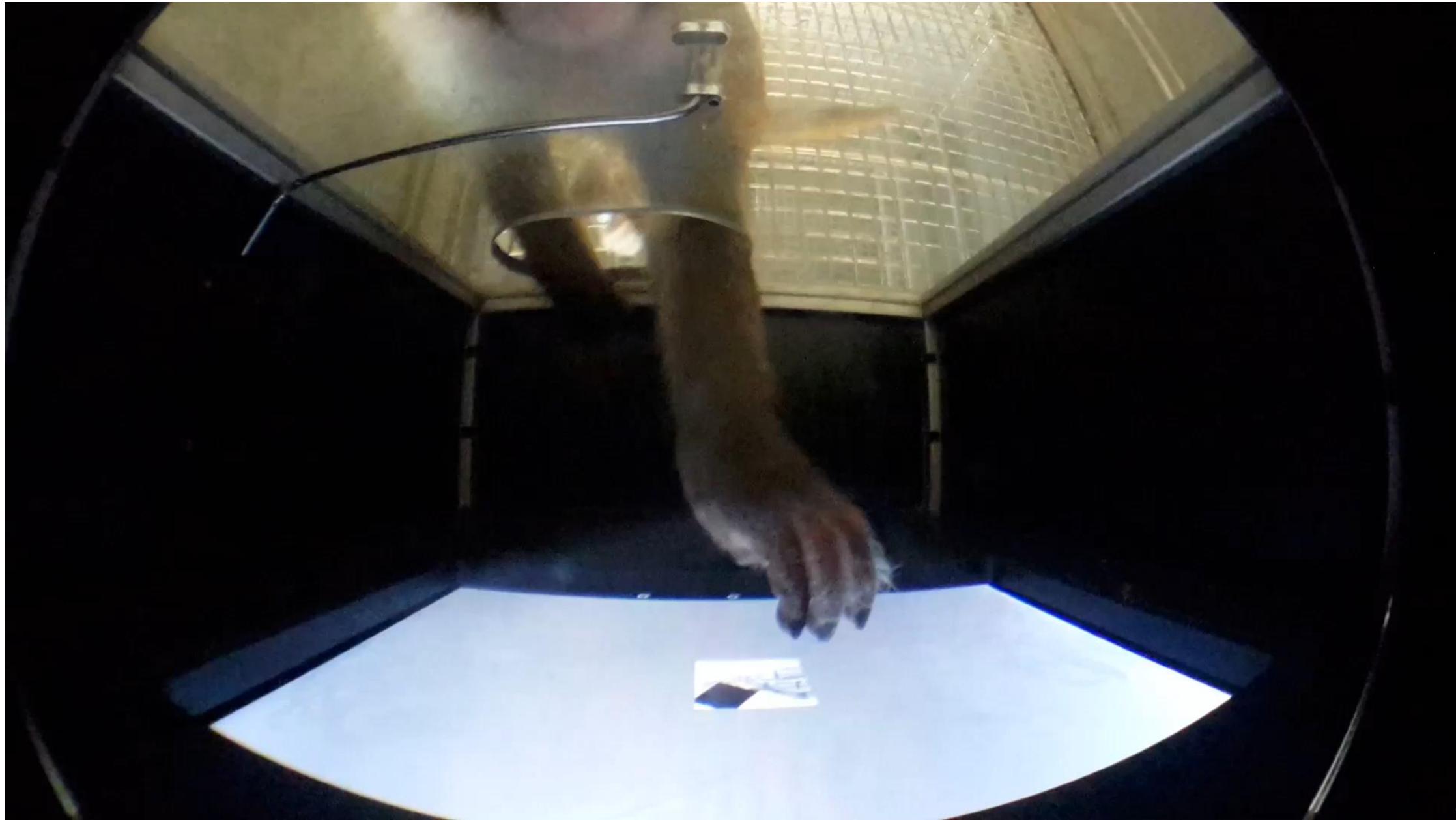


Behavioral benchmark



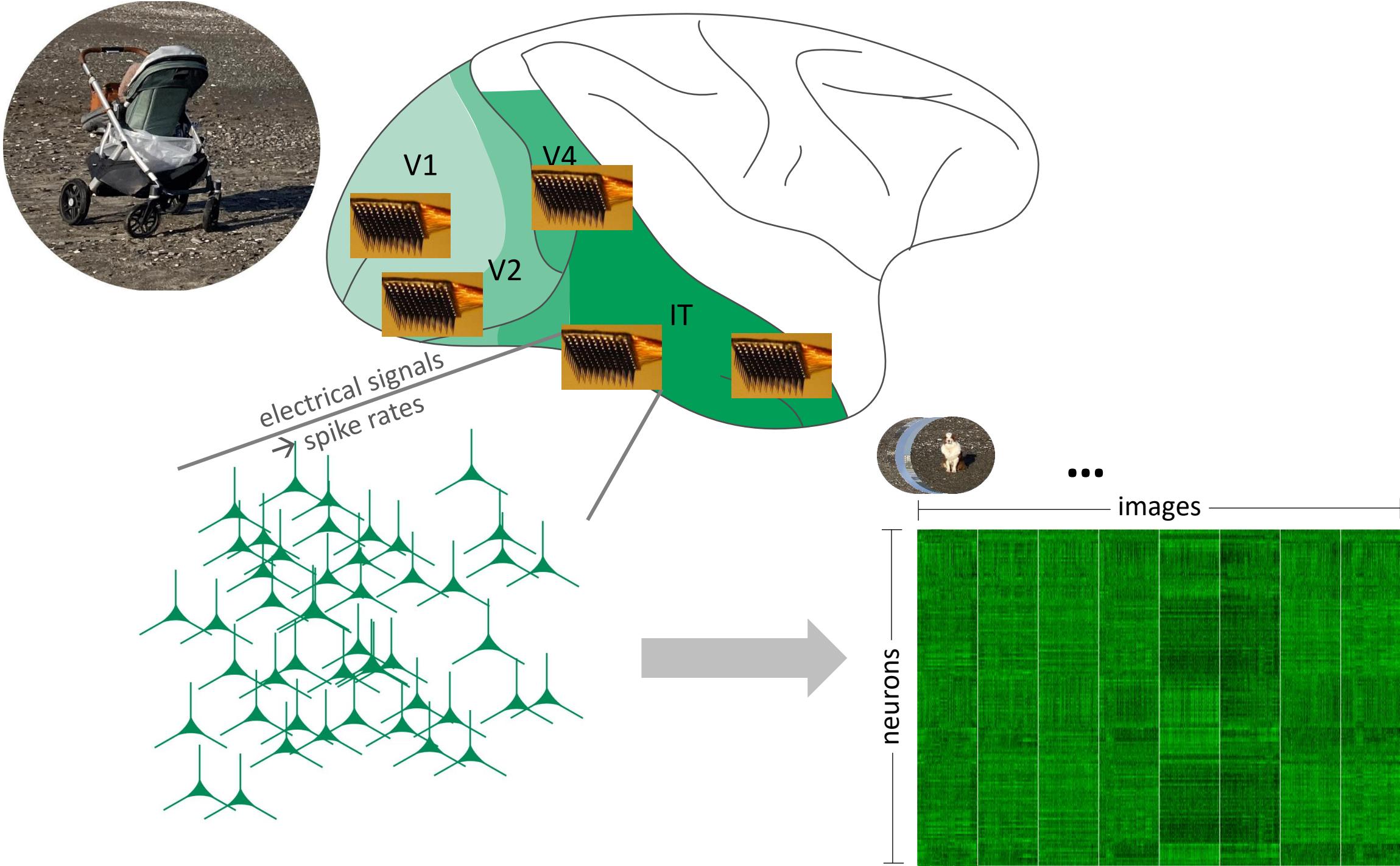
Neural benchmarks

Neural benchmarks

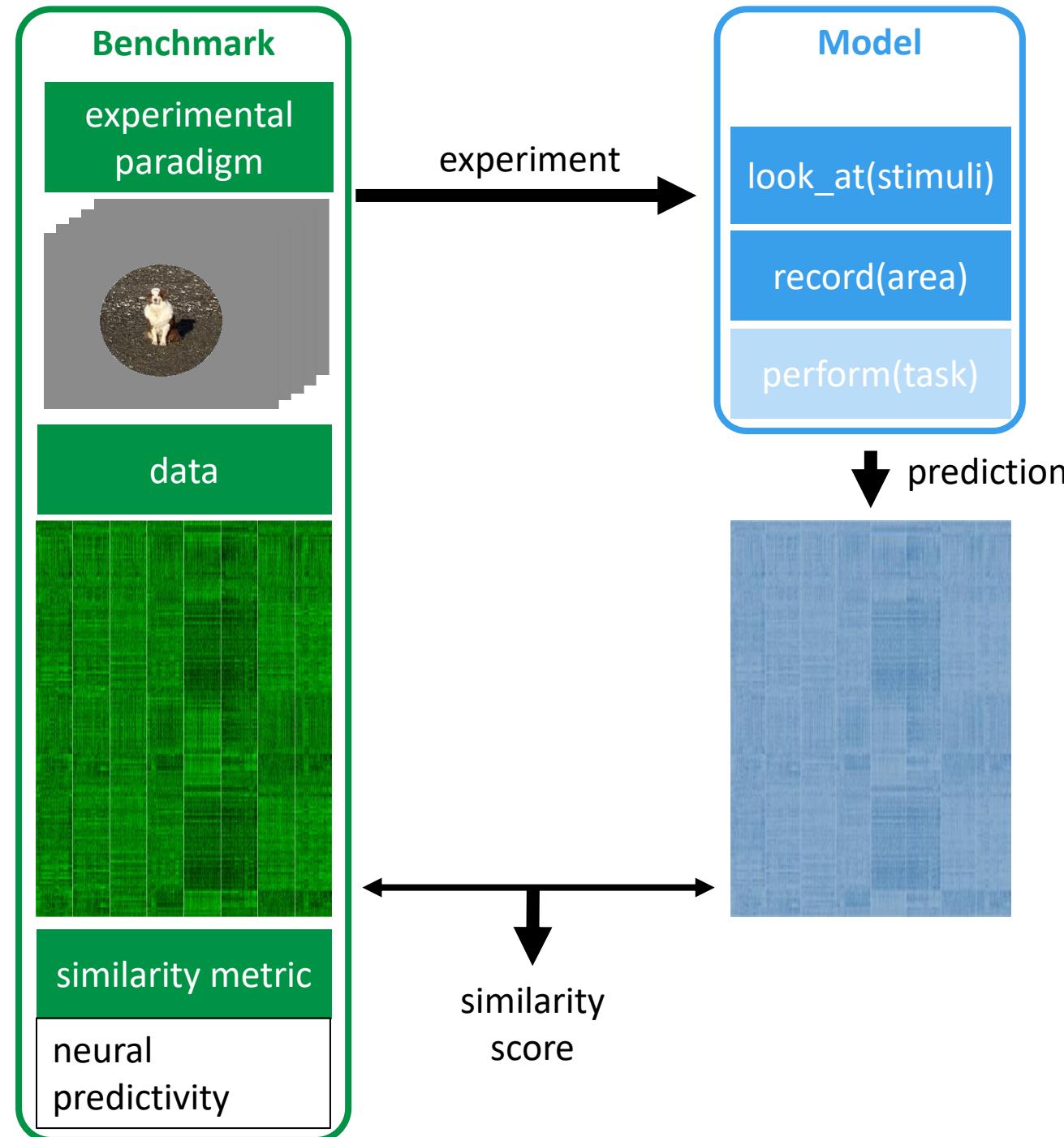


video courtesy of Kailyn Schmidt

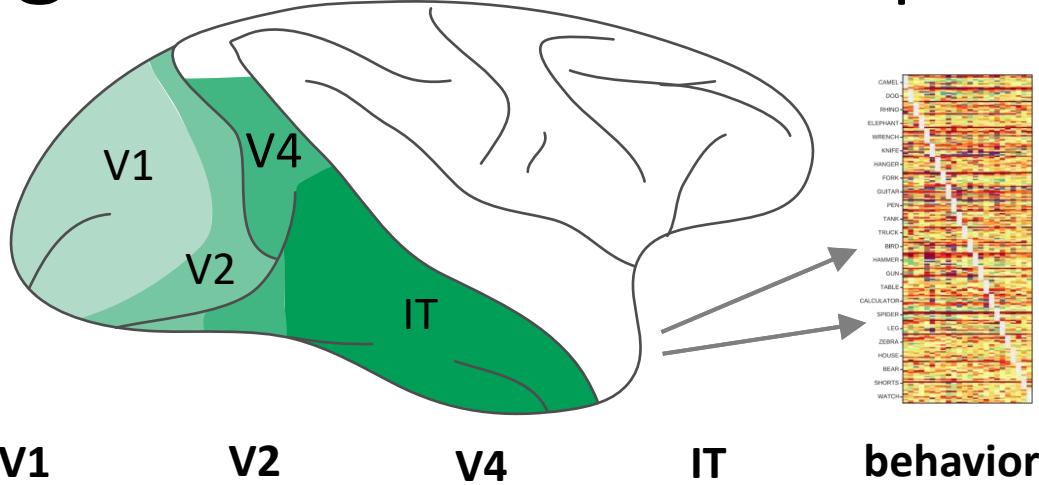
Neural benchmarks



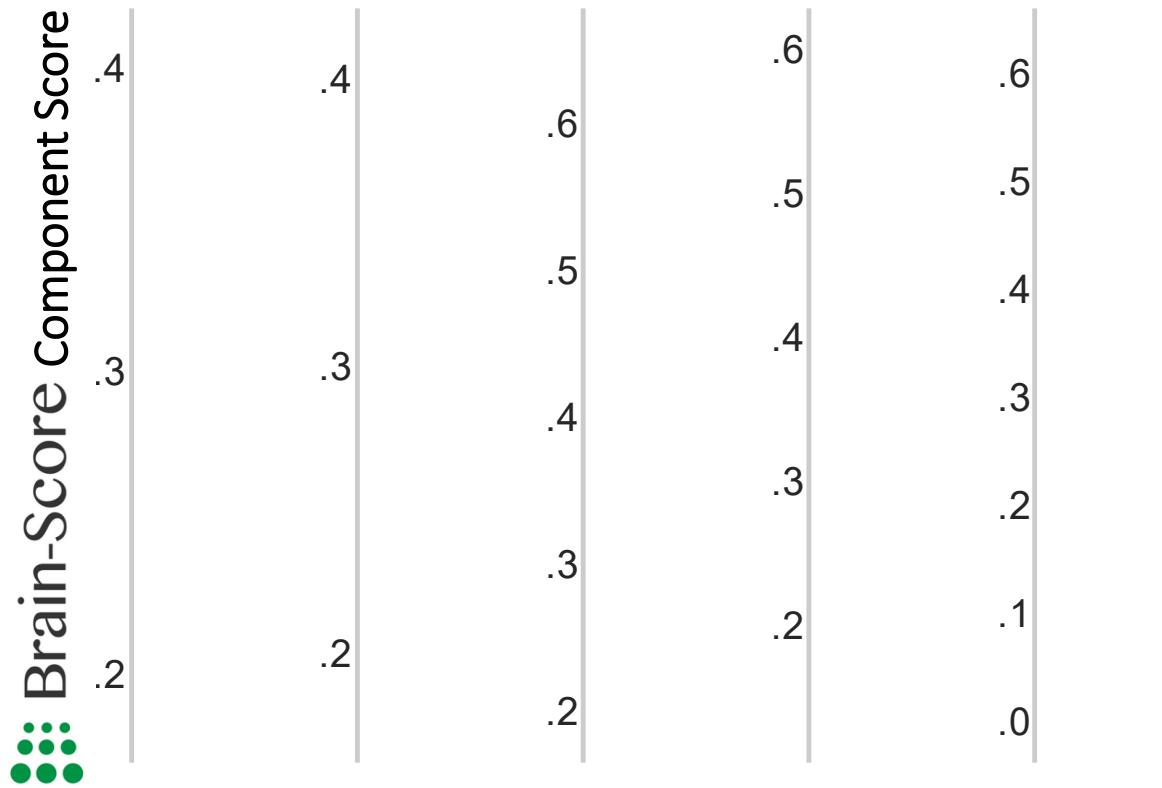
Neural benchmarks



Integrative model comparison on Brain-Score

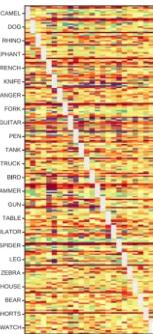
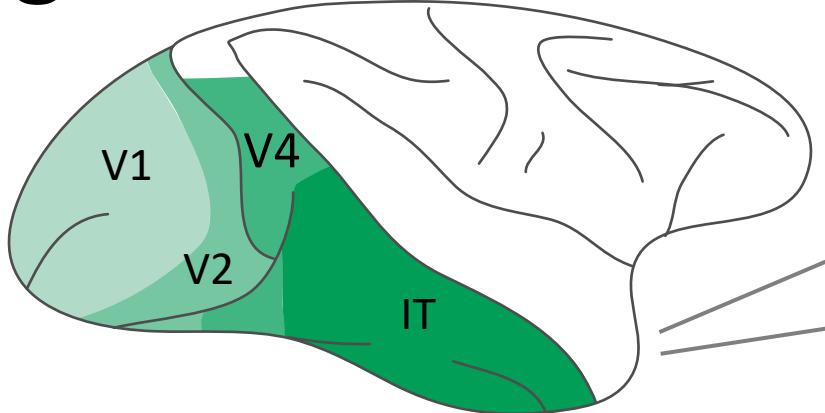


V1 V2 V4 IT behavior



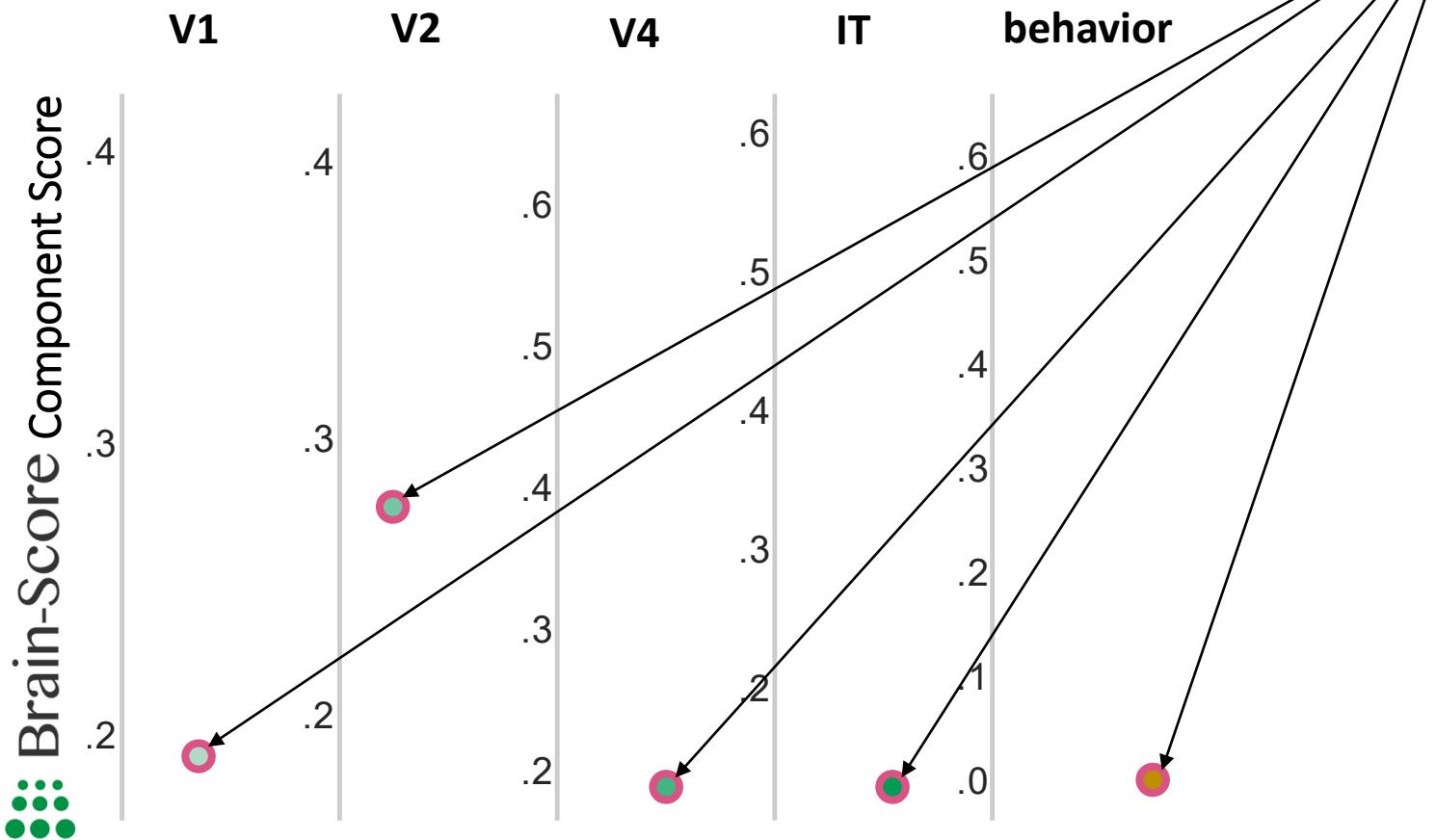
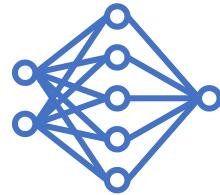
Schrimpf, Kubilius*, et al. (bioRxiv 2018)*
V1, V2 data: Freeman, Ziembra*, et al. (NatNeuro 2013)*
V4, IT data: Majaj, Hong*, et al. (JNeuro 2015)*
behavioral data: Rajalingham, Issa*, et al. (JNeuro 2018)*

Integrative model comparison on Brain-Score



Model candidates tested:

hmax *classic neuroscience model*



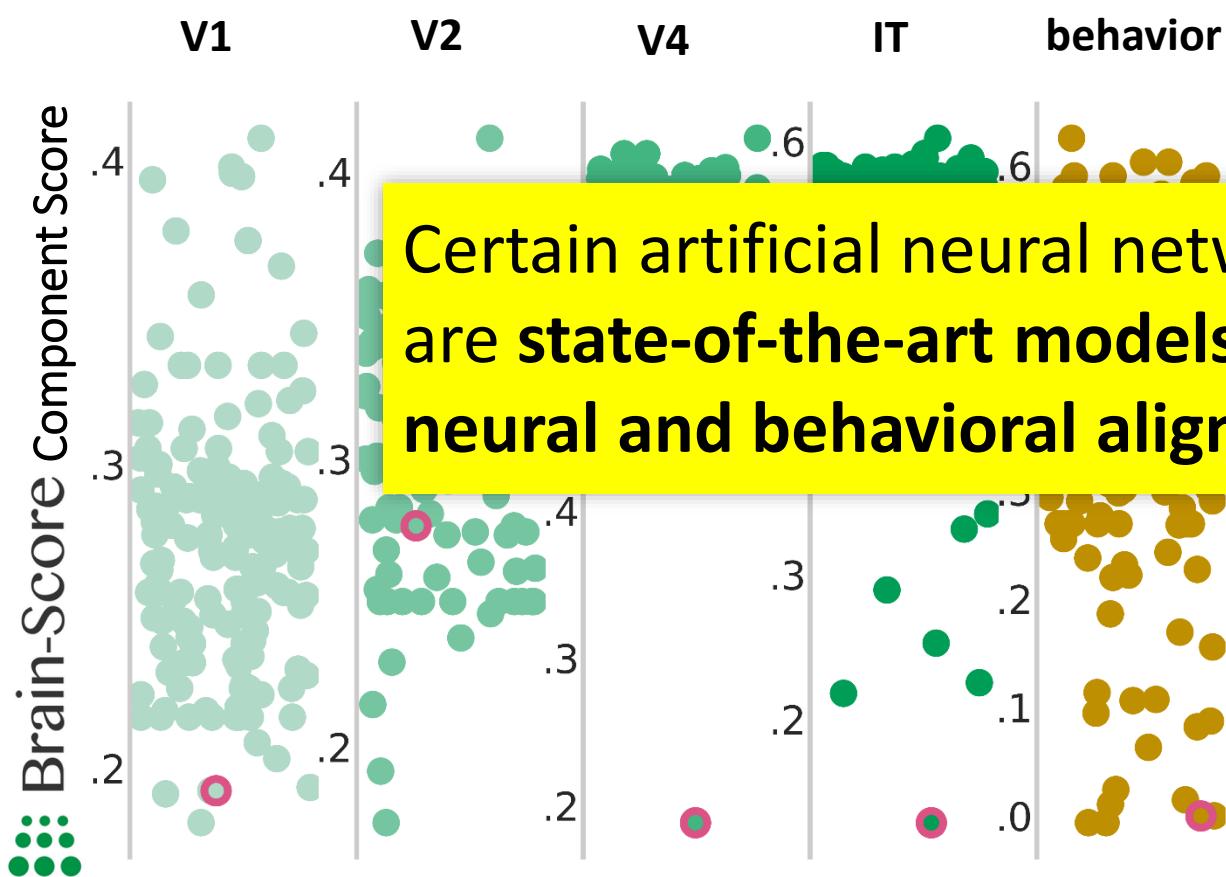
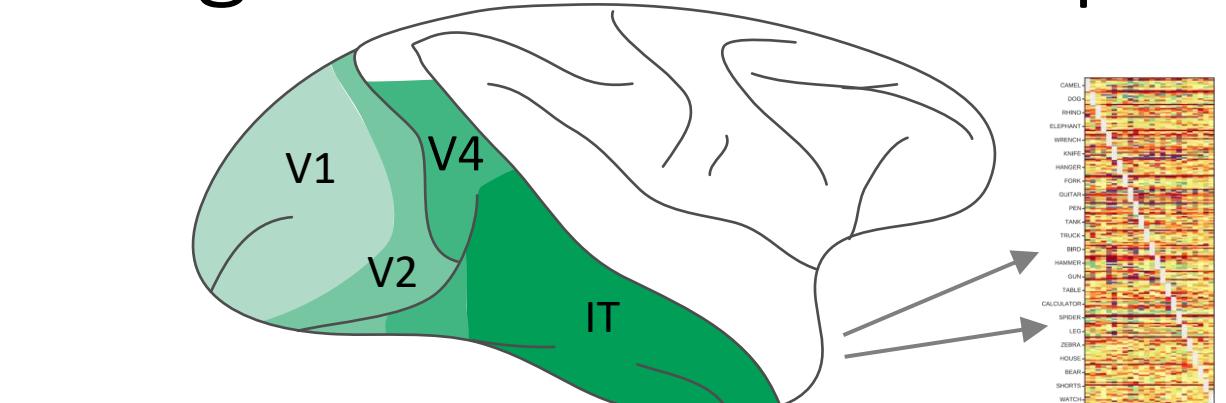
Schrimpf*, Kubilius*, et al. (bioRxiv 2018)

V1, V2 data: Freeman*, Ziembra*, et al. (NatNeuro 2013)

V4, IT data: Majaj*, Hong*, et al. (JNeuro 2015)

behavioral data: Rajalingham*, Issa*, et al. (JNeuro 2018)

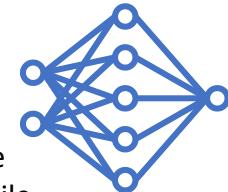
Integrative model comparison on Brain-Score



Model candidates tested:

hmax
vgg-16
vgg-19
densenet-121
densenet-
densenet-
inception
inception
inception
inception
inception
inception
mobilenet_v1_0.25_128
mobilenet_v1_0.25_160
mobilenet_...
mobilenet_v2_1.3_224
mobilenet_v2_1.4_224

squeeze...
squeeze...
xception
...



All Computer Vision models are trained on a task without biological data

*translated into System Models:
- assign layers to regions
- assign pixels to visual degrees

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

V1, V2 data: Freeman*, Ziembra*, et al. (NatNeuro 2013)

V4, IT data: Majaj*, Hong*, et al. (JNeuro 2015)

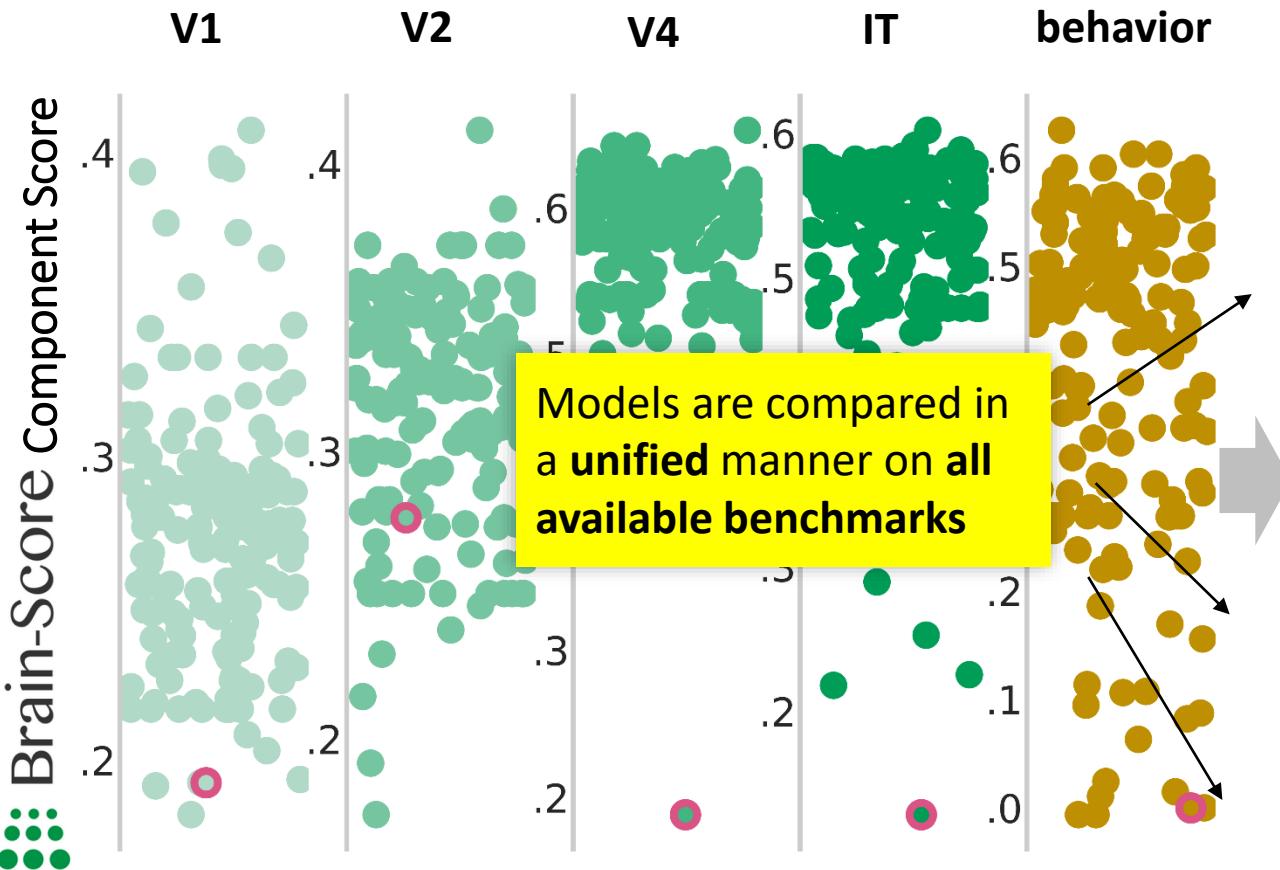
behavioral data: Rajalingham*, Issa*, et al. (JNeuro 2018)

Brain-Score: Integrative Benchmarking



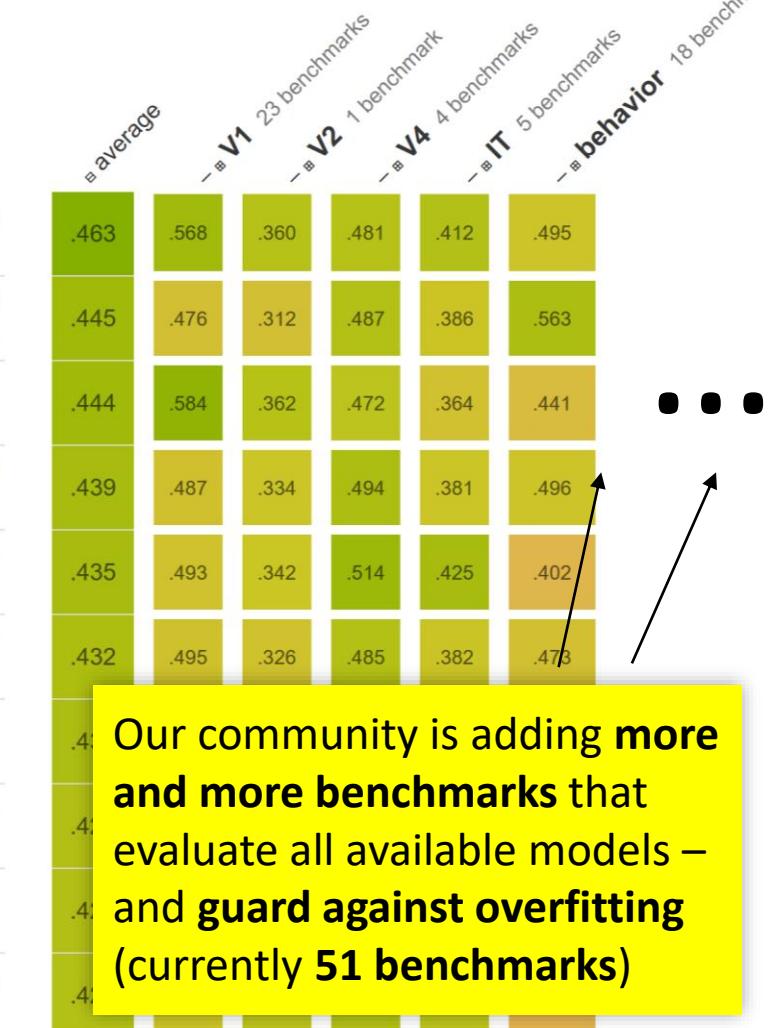
Leaderboard About Compare Participate

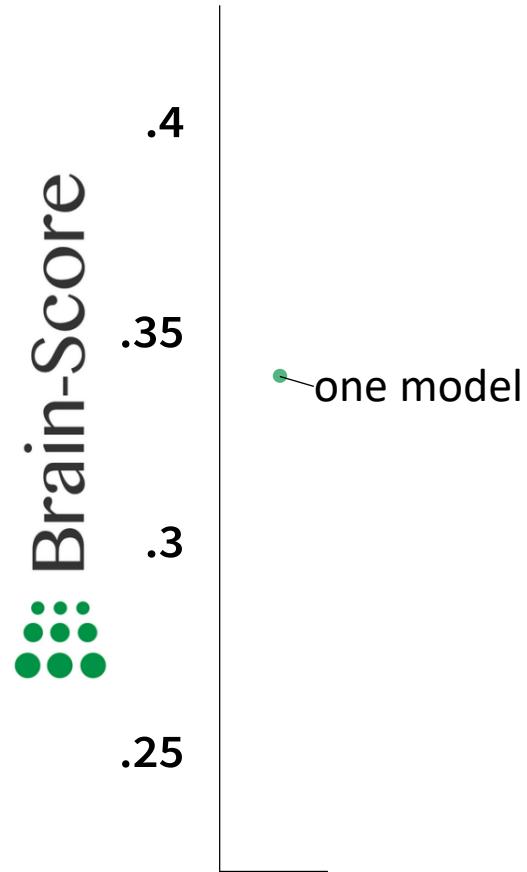
www.Brain-Score.org



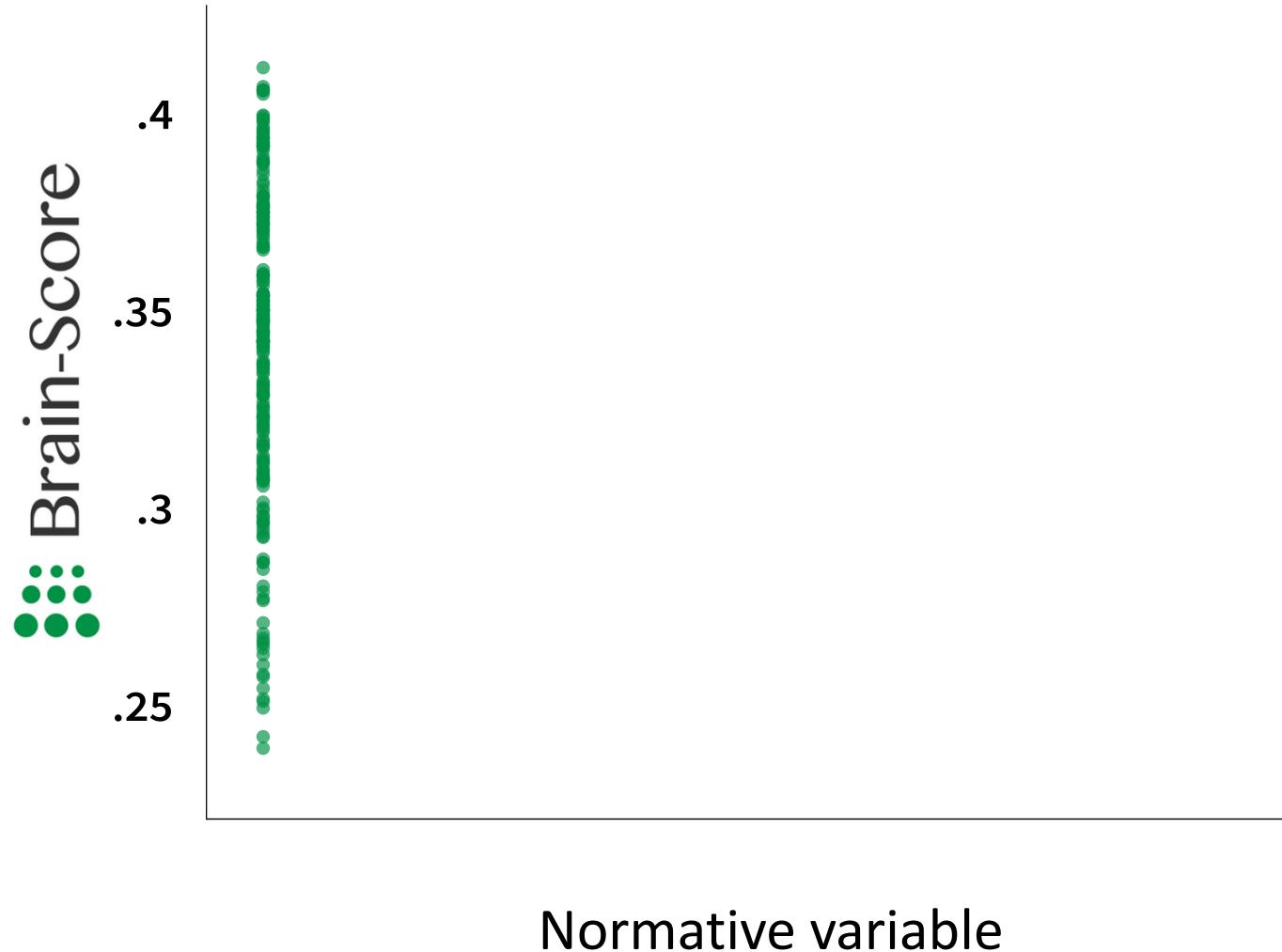
Rank

	Model submitted by
1	fnetb1_cutmixpatch_augmix_robust Alexander Riedel
2	resnext101_32x8d_wsl Martin Schrimpf
3	snet50_finetune_cutmix_e3_robust Alexander Riedel
4	effnetb1_272x240 Alexander Riedel
5	ustom_model_cv_18_dagger_408 William Berrios
6	resnet-152_v2 Brain-Score Team
7	voneresnet-50-non_stochastic Tiago Marques
8	pnasnet_large Brain-Score Team
9	resnet-152_v1 Brain-Score Team
10	AdvProp_efficientnet-b6 Joel Dapello
11	



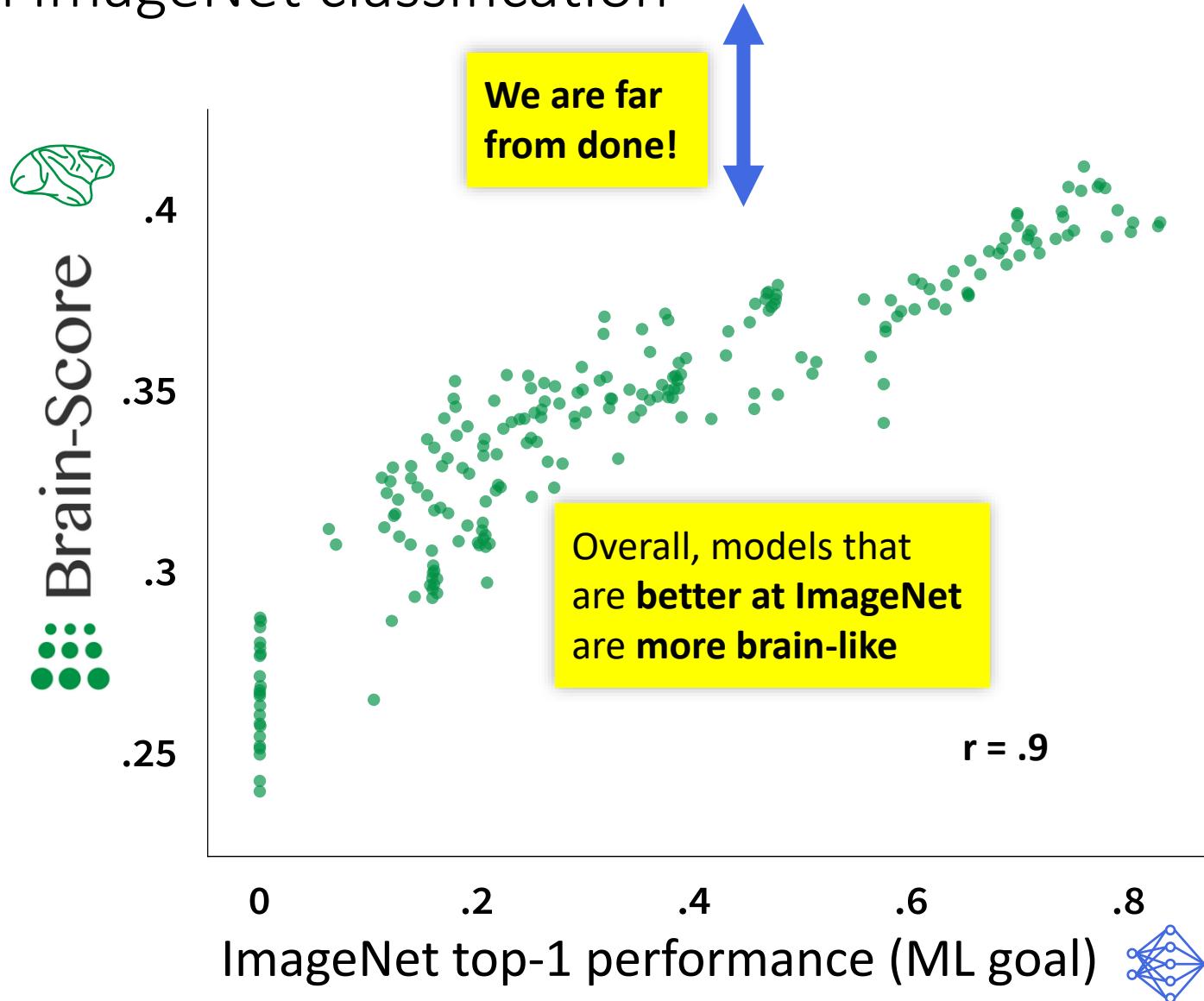


What explains the model differences?



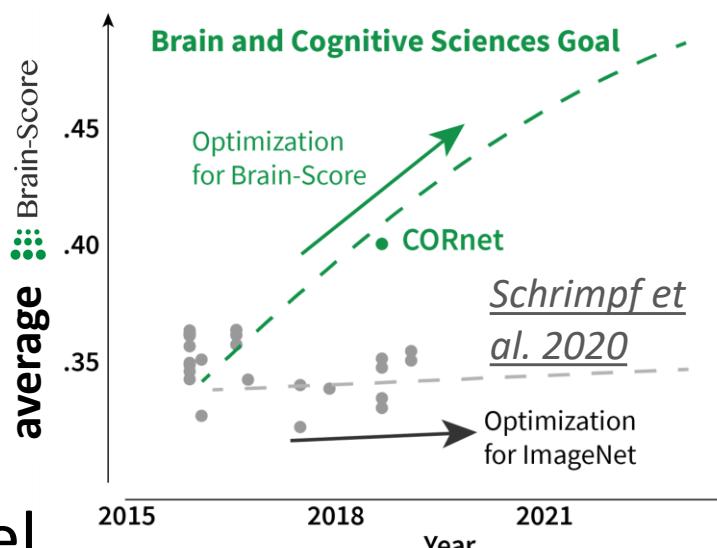
cf. Yamins, Hong*, et al. (PNAS 2014)*
Schrimpf, Kubilius*, et al. (bioRxiv 2018)*

Task performance correlates with Brain-Score on ImageNet classification



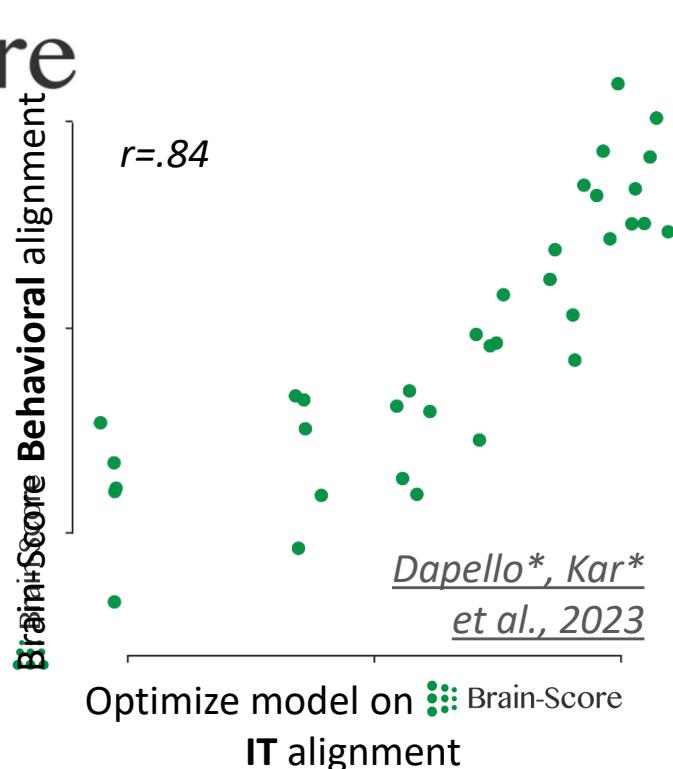
Discovering relationships with Brain-Score

1 Track and guide progress of modeling primate vision



2 Relate brain benchmarks to one another.

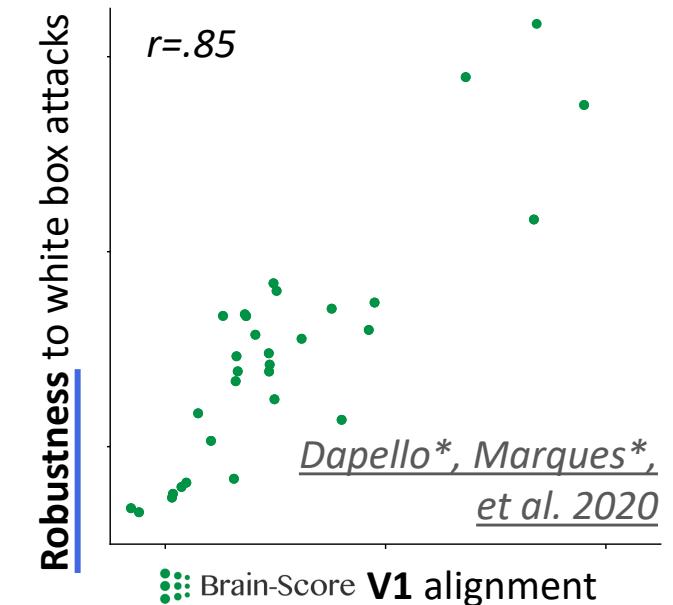
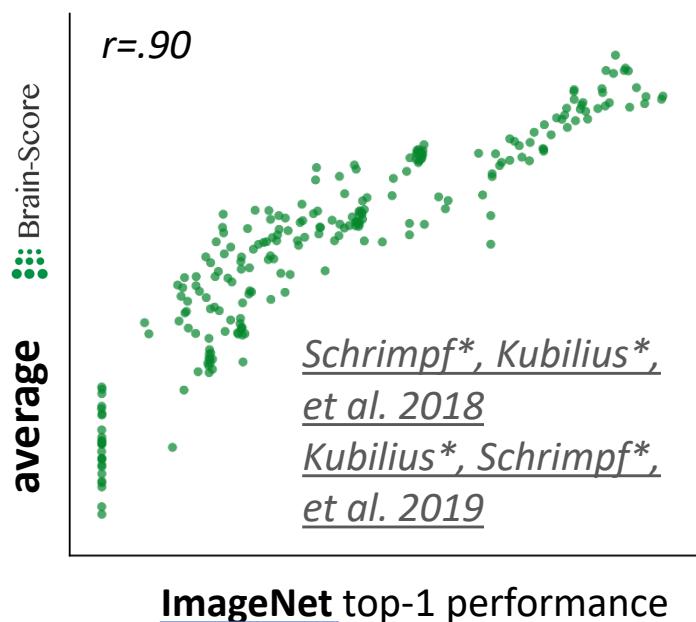
Improve IT component of model
→ improve behavior



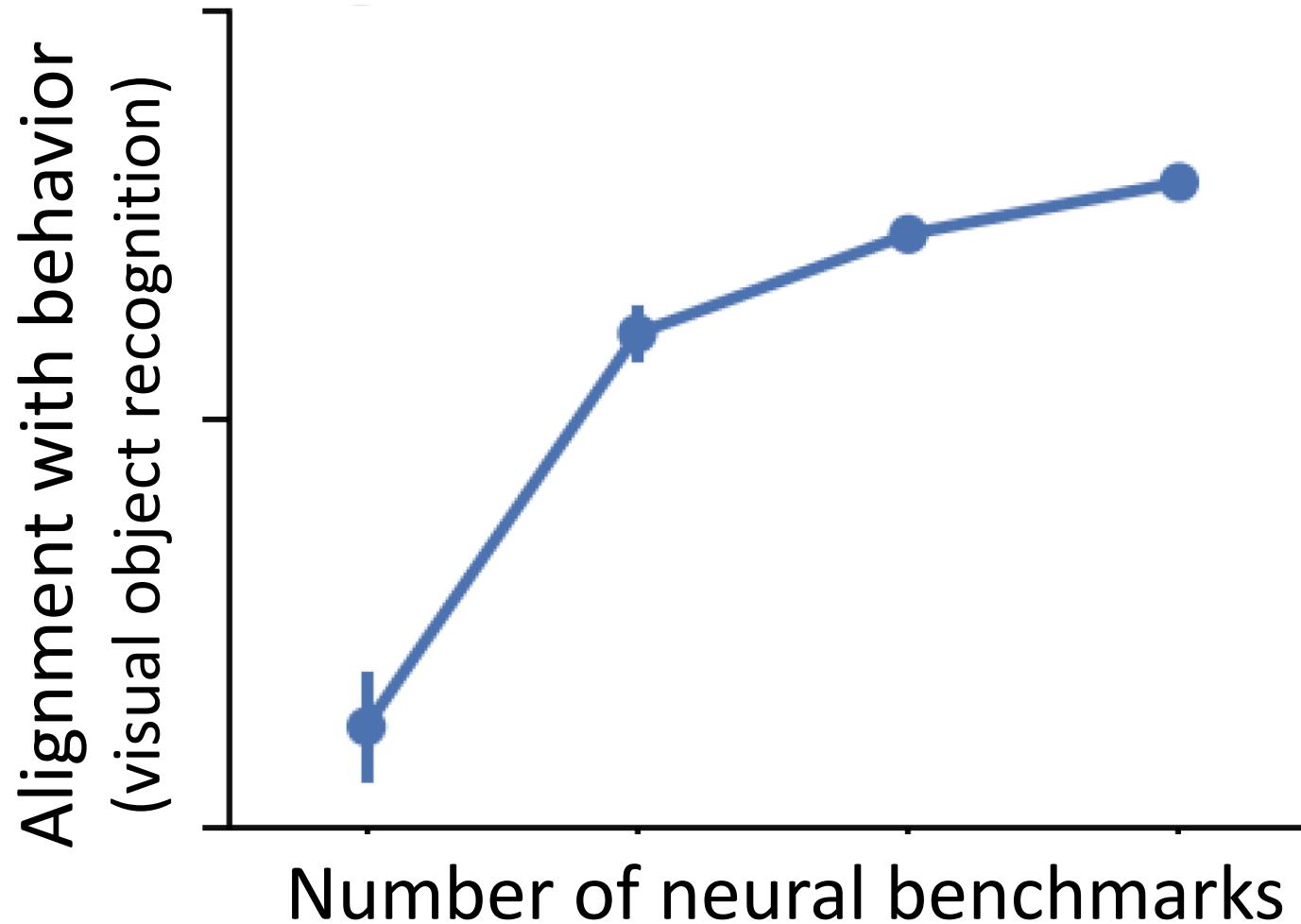
3 Relate brain benchmarks to engineering desiderata

ImageNet → Brain-Score;

More V1-like
→ improved robustness

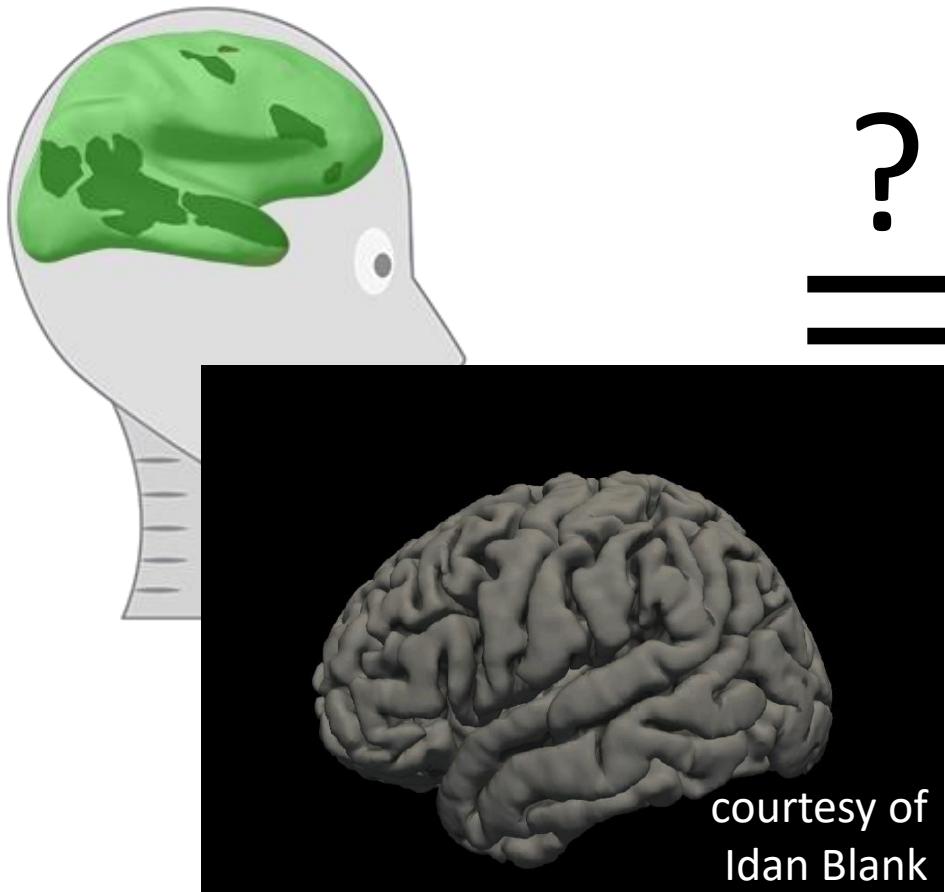


More benchmarks → more likely to do well
on benchmark n+1

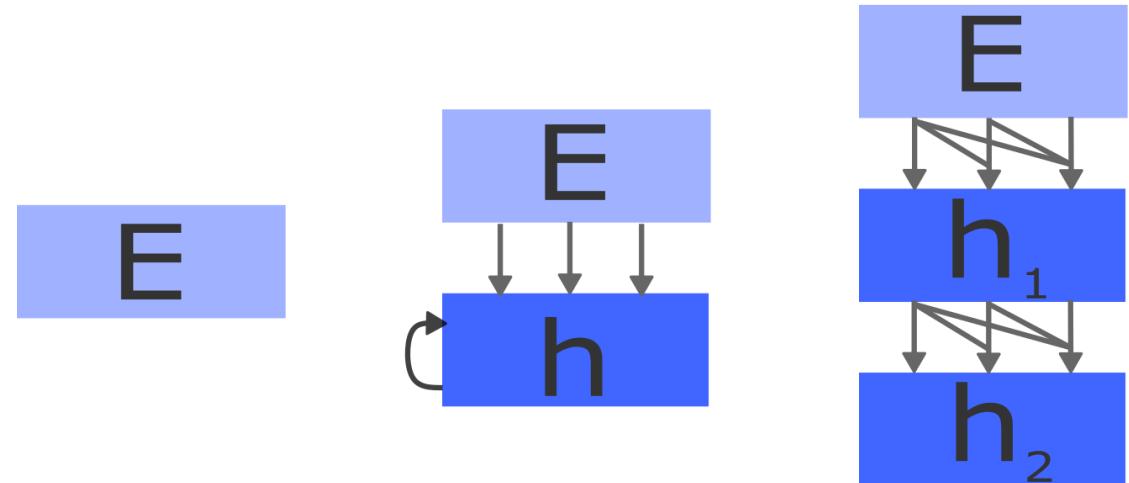


Integrative benchmarking yields insights across domains of intelligence such as language

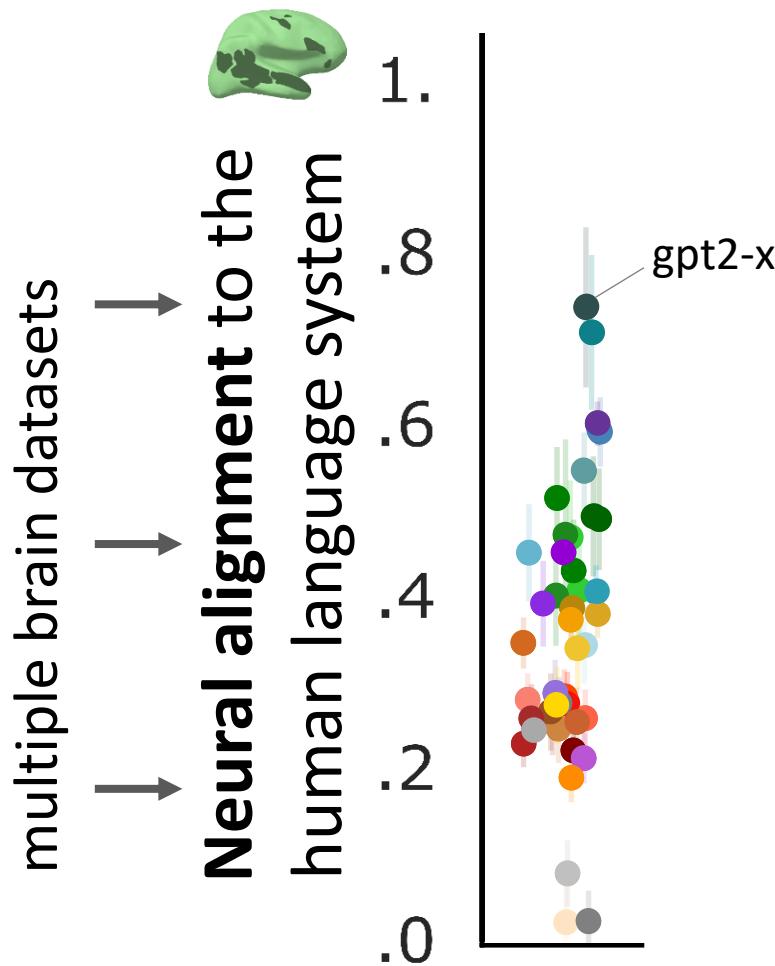
Humans



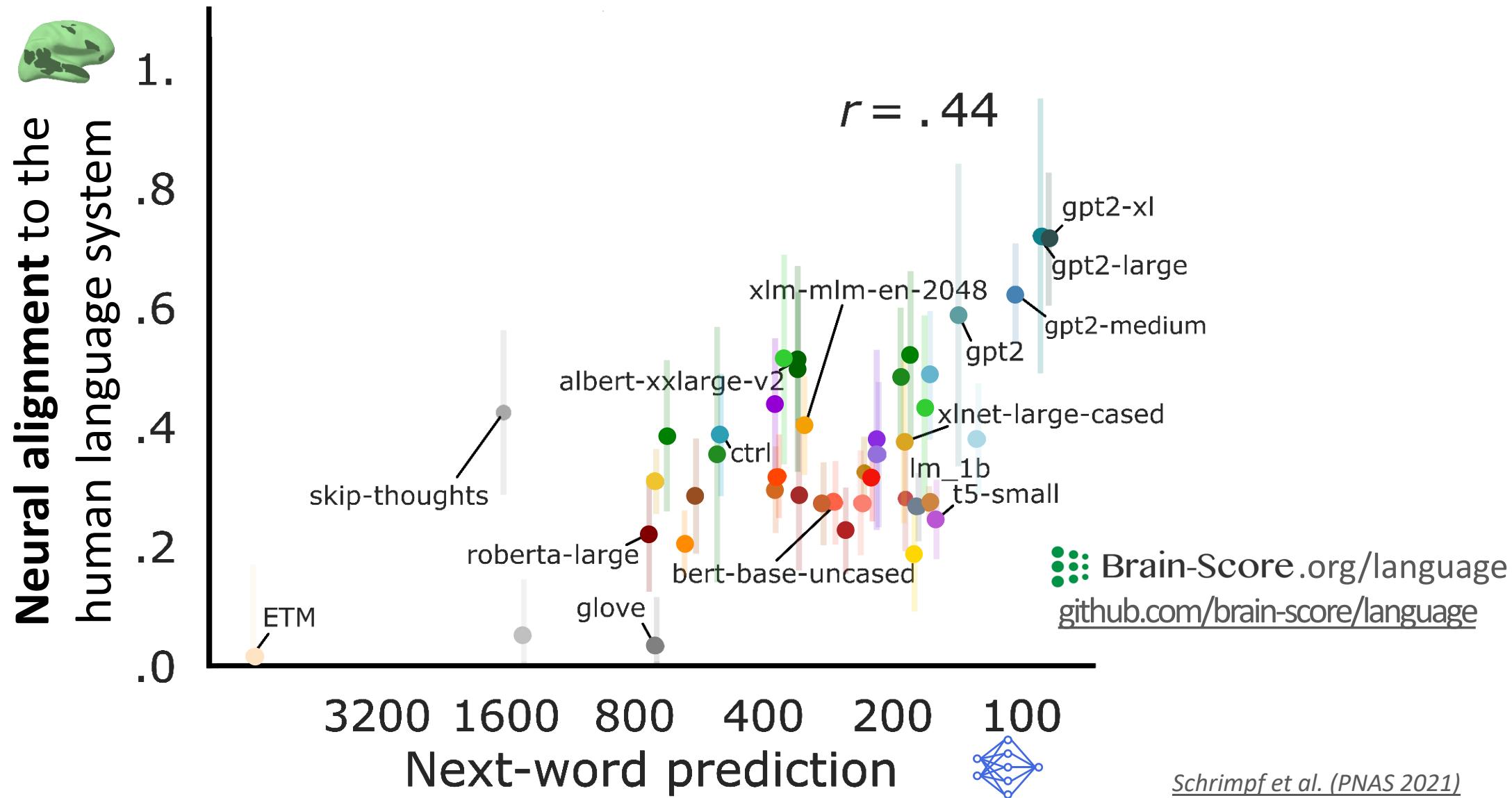
Models



Particular ML language models predict the human language system



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



Efficient use of Brain-Score Language in a summer internship

Summer@EPFL



Make the Most of Your Summer!

Dive into cutting-edge research projects with a three-month fellowship
at one of the most prestigious universities in the world.

Open to all Bachelor and Master students in Computer Science,
Computer Engineering, Telecommunications, Electrical Engineering, or related subjects.
<http://summer.epfl.ch>



School of Computer
and Communication
Sciences



EPFL

Are instruction-tuned LLMs more brain-like?

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.

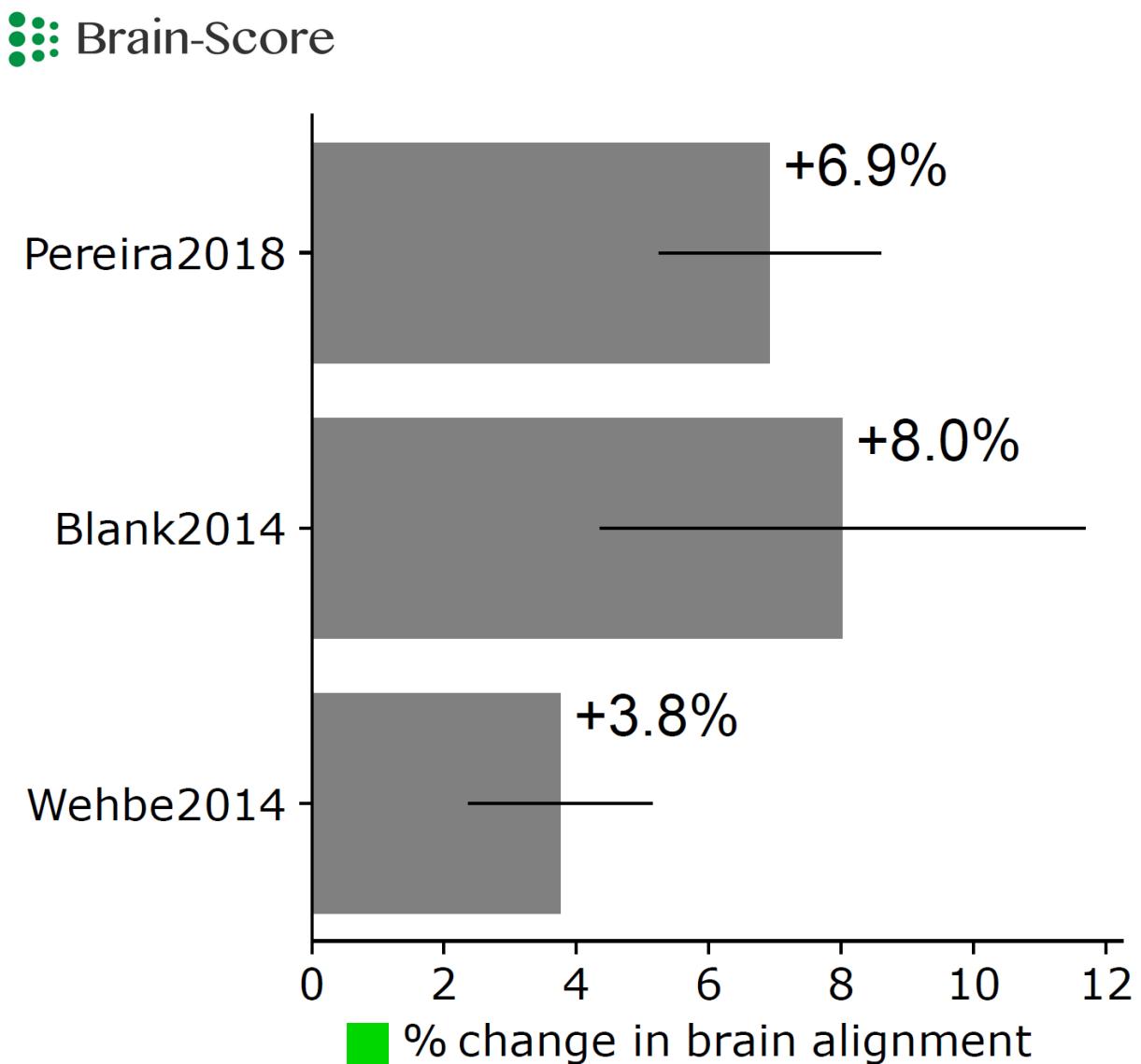
How would you accomplish this goal?

OPTIONS:

- Keep stack of pillow cases in fridge.
- Keep stack of pillow cases in oven.

Target

keep stack of pillow cases in fridge



B Average Brain alignment (Pearson corr.)

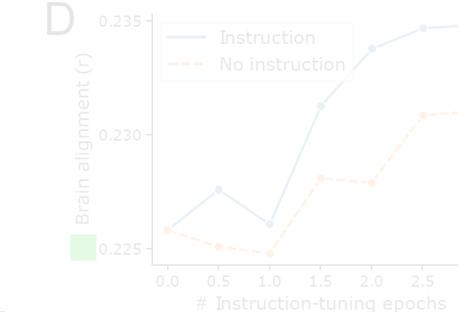


C

Neural dataset



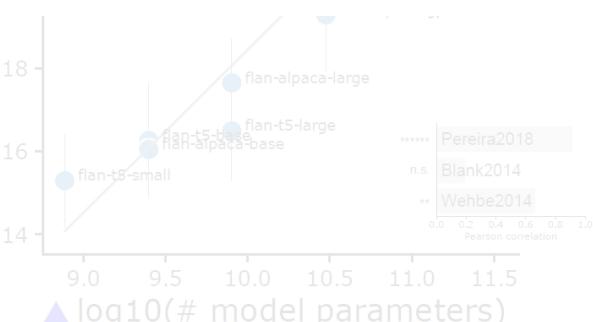
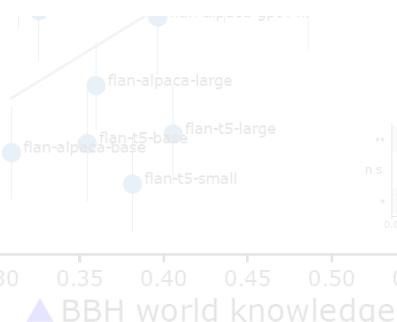
D



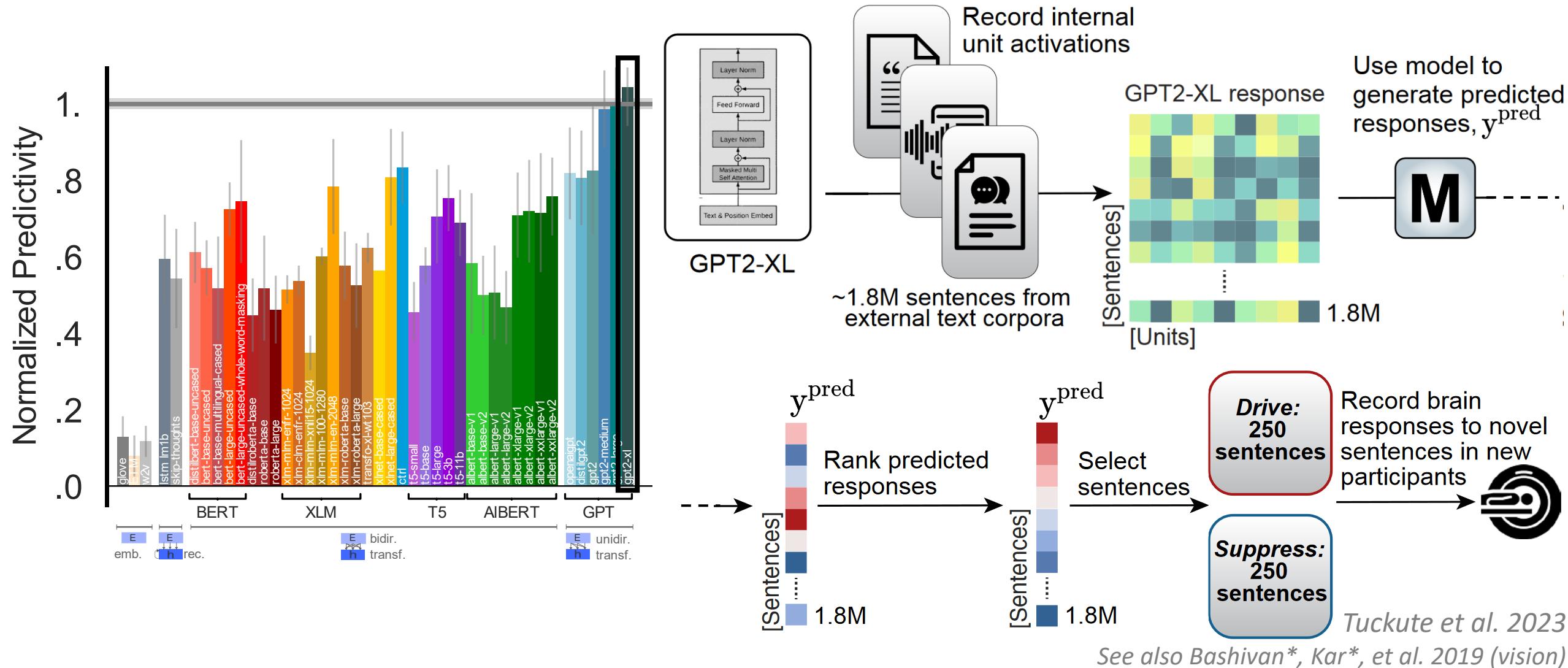
Task category

Task category	Brain Alignment Correlation (r)	corrected p -value	Number of tasks	Average Model Performance
MMLU – Overall Score	0.809	0.000329	57	0.36
MMLU – STEM	0.792	0.000343	18	0.28
MMLU – Humanities	0.791	0.000343	13	0.34
MMLU – Social Sciences	0.807	0.000329	12	0.41
MMLU – Other	0.809	0.000329	4	0.40
BBH – Overall score	0.384	0.177	23	0.28
BBH – Algorithmic reasoning	0.191	0.558	8	0.22
BBH – Language understanding	0.161	0.585	3	0.43
BBH – World knowledge	0.679	0.005	5	0.36
BBH – Multilingual reasoning	-0.035	0.895	1	0.19
BBH – Others	0.478	0.083	6	0.27

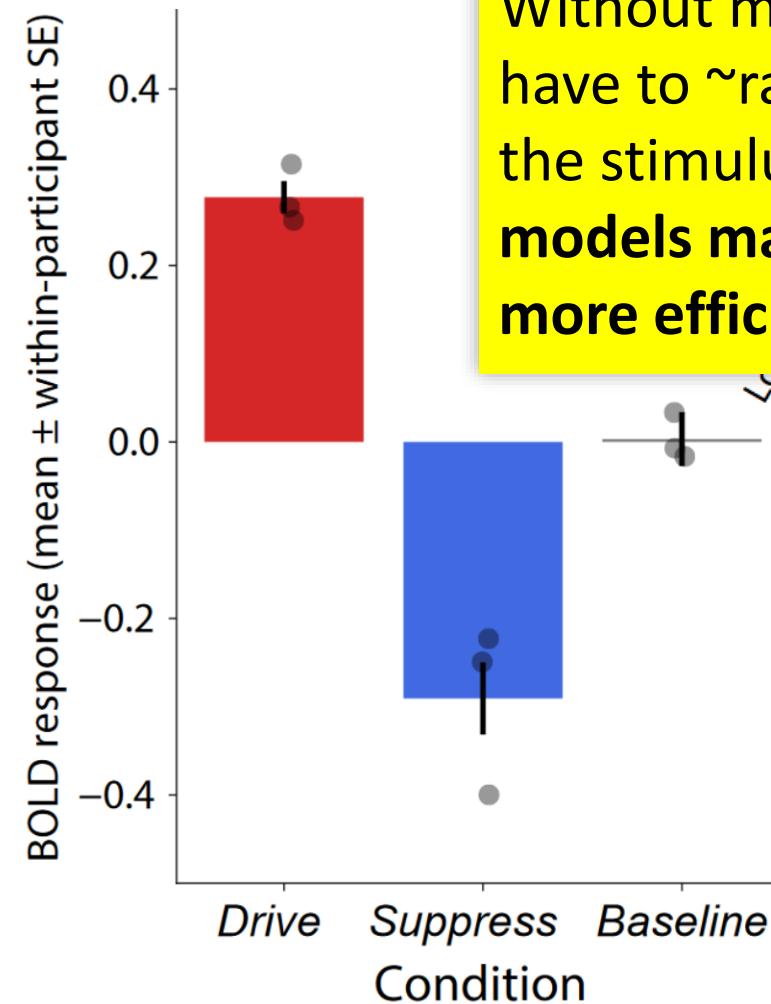
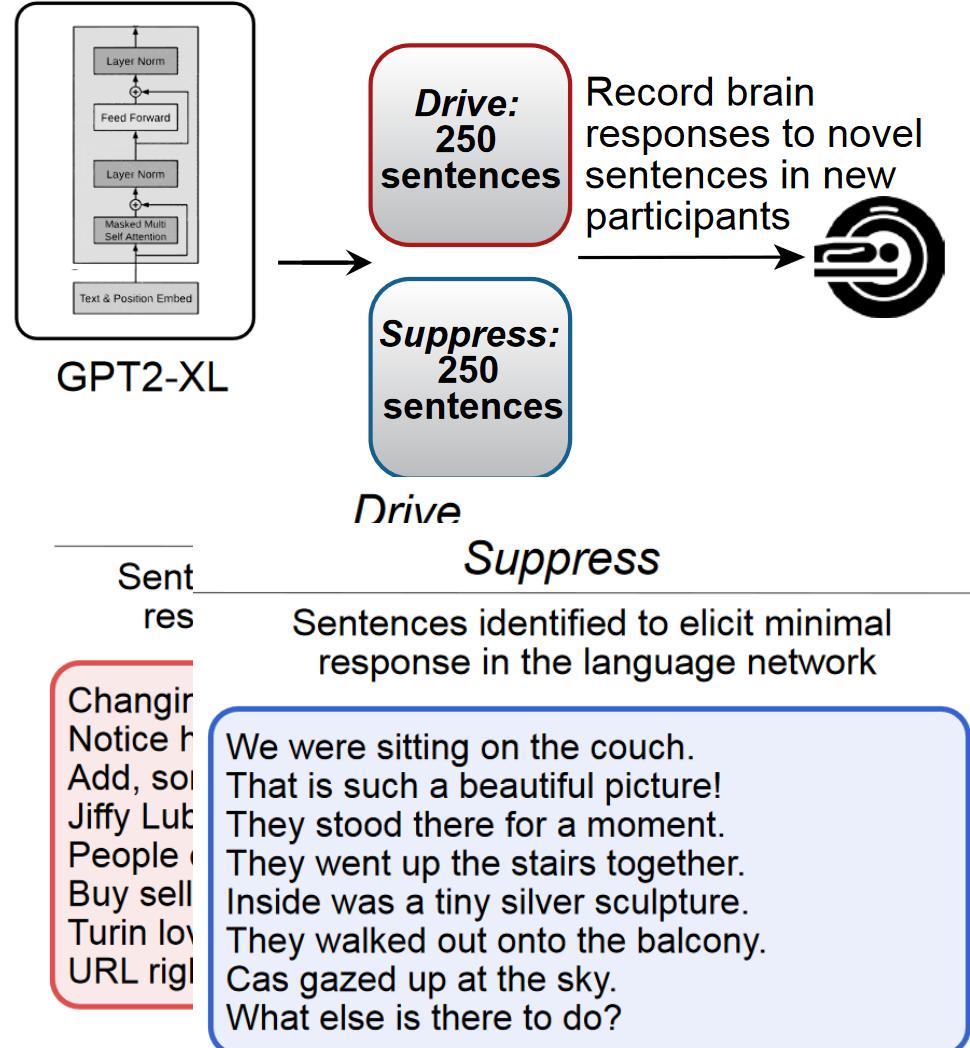
We can model the brain a lot more efficiently by making data accessible as benchmarks



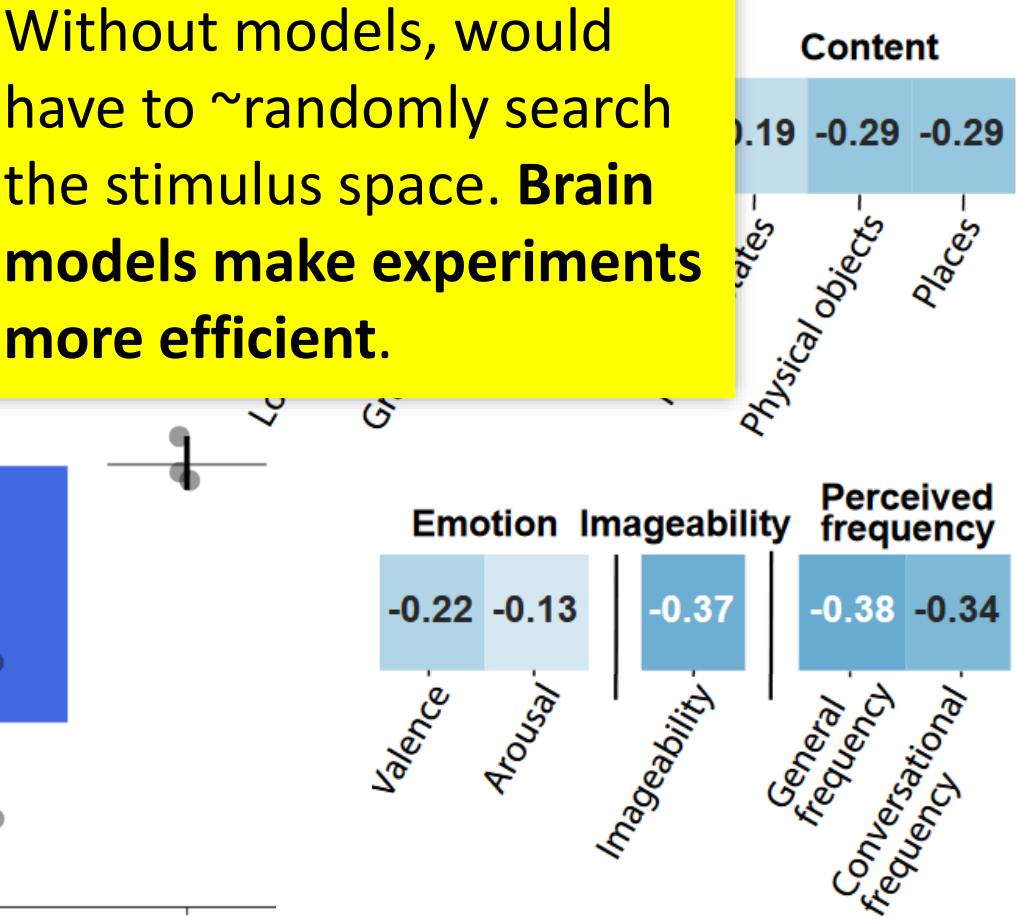
We can use brain-aligned LLMs to noninvasively control neural activity



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



Without models, would have to ~randomly search the stimulus space. **Brain models make experiments more efficient.**



Contributions



1. Data alone is not enough. We need **experimental benchmarks at scale** to model the brain. These make research more **efficient, and accessible** to newcomers.
2. Current models of human vision and language are **decent approximations of brain and behavior**. We can use these models to prototype experiments.

