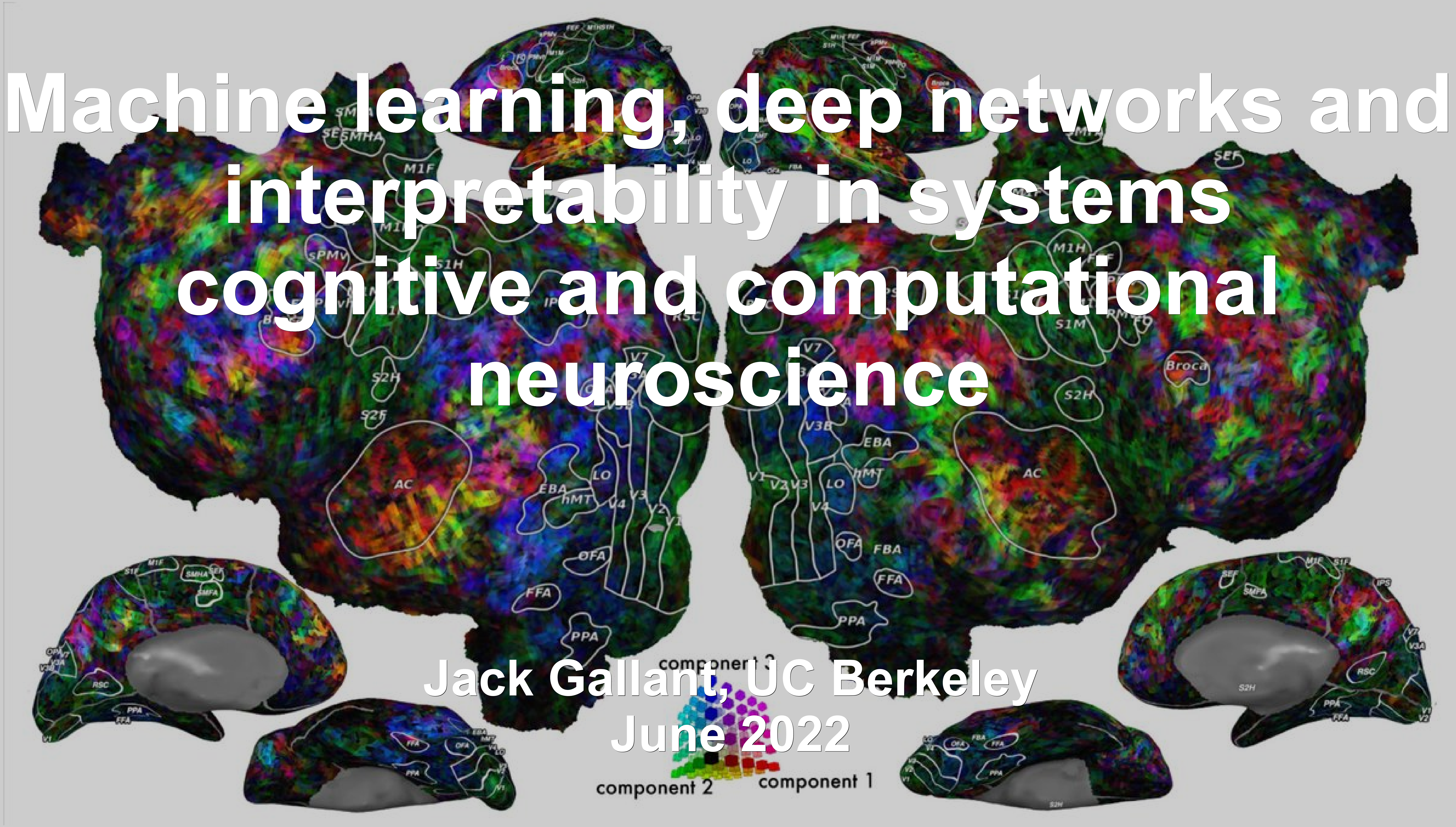
Machine learning, deep networks and interpretability in systems cognitive and computational neuroscience

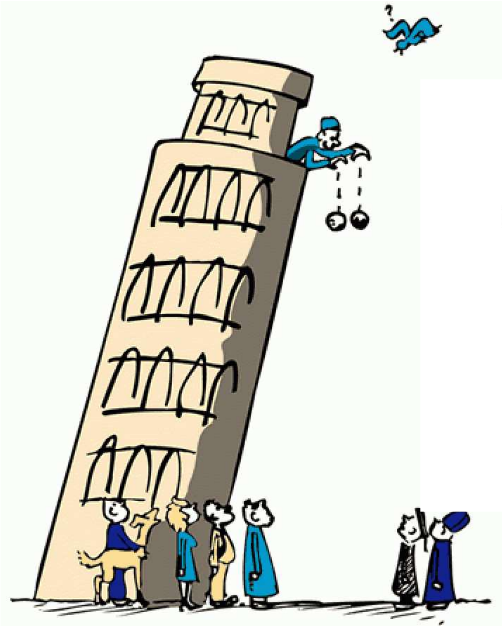
Jack Gallant, UC Berkeley

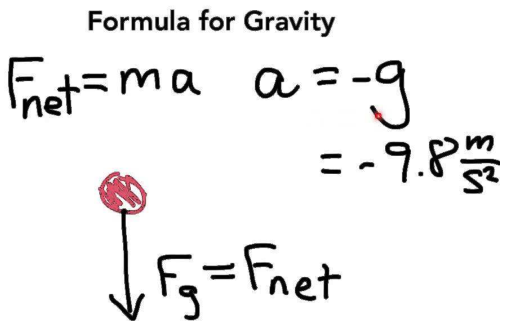
June 2022

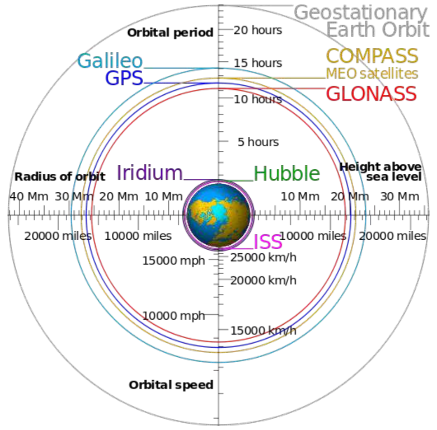


Science and engineering have different priorities and these affect our choices (especially regarding deep learning)

Measure → **Explain** → **Predict**

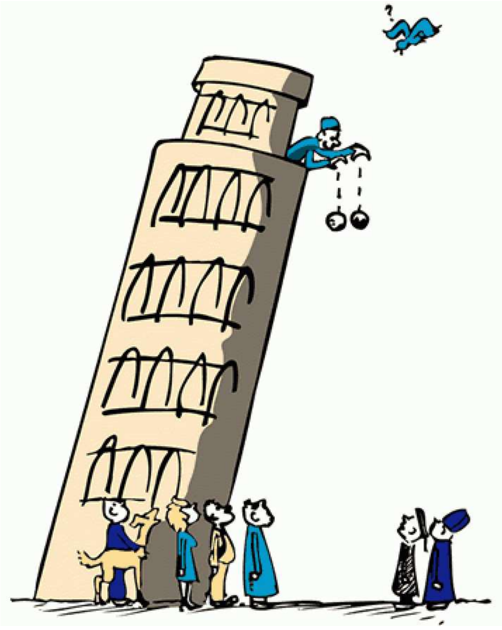


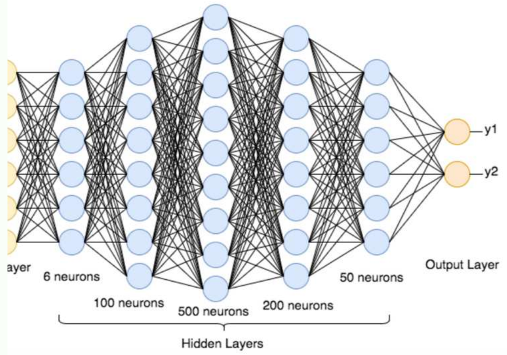


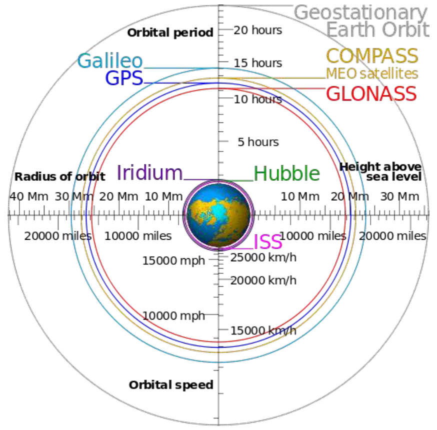


Science (and classical engineering)

Measure → **Train** → **Predict**

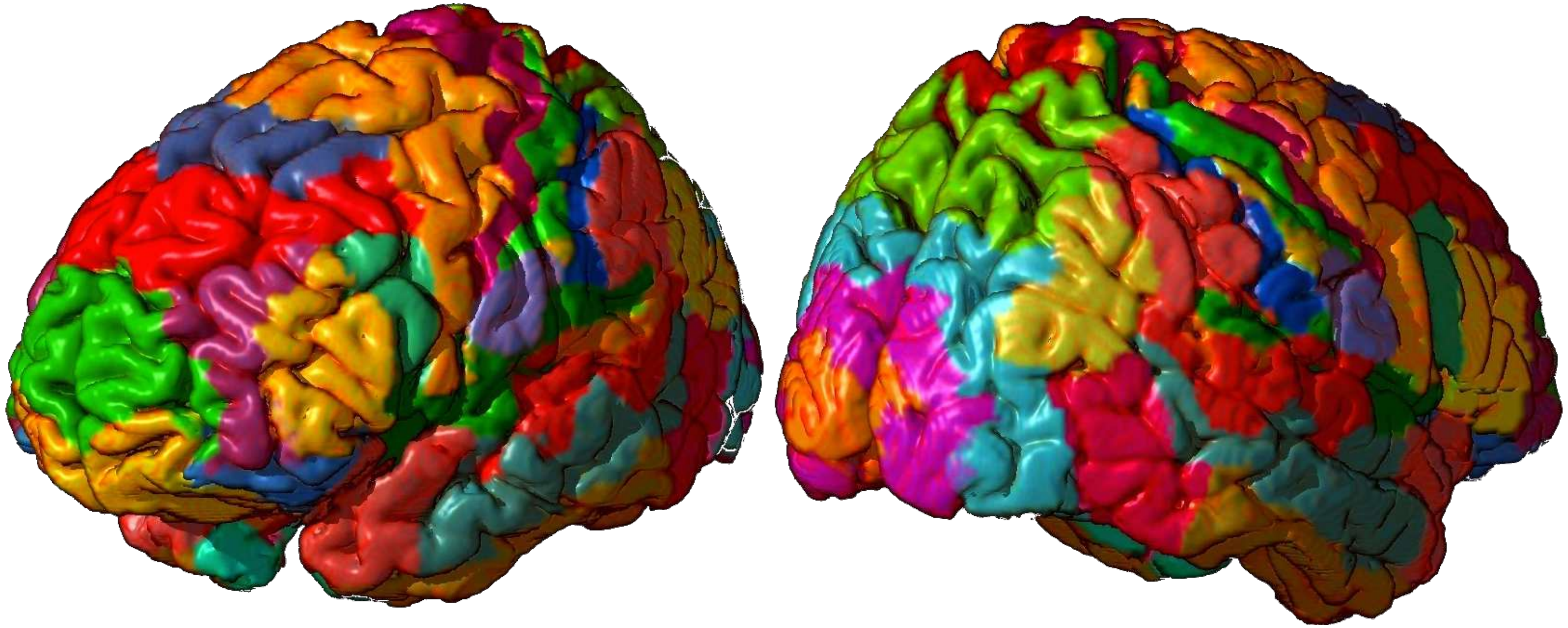


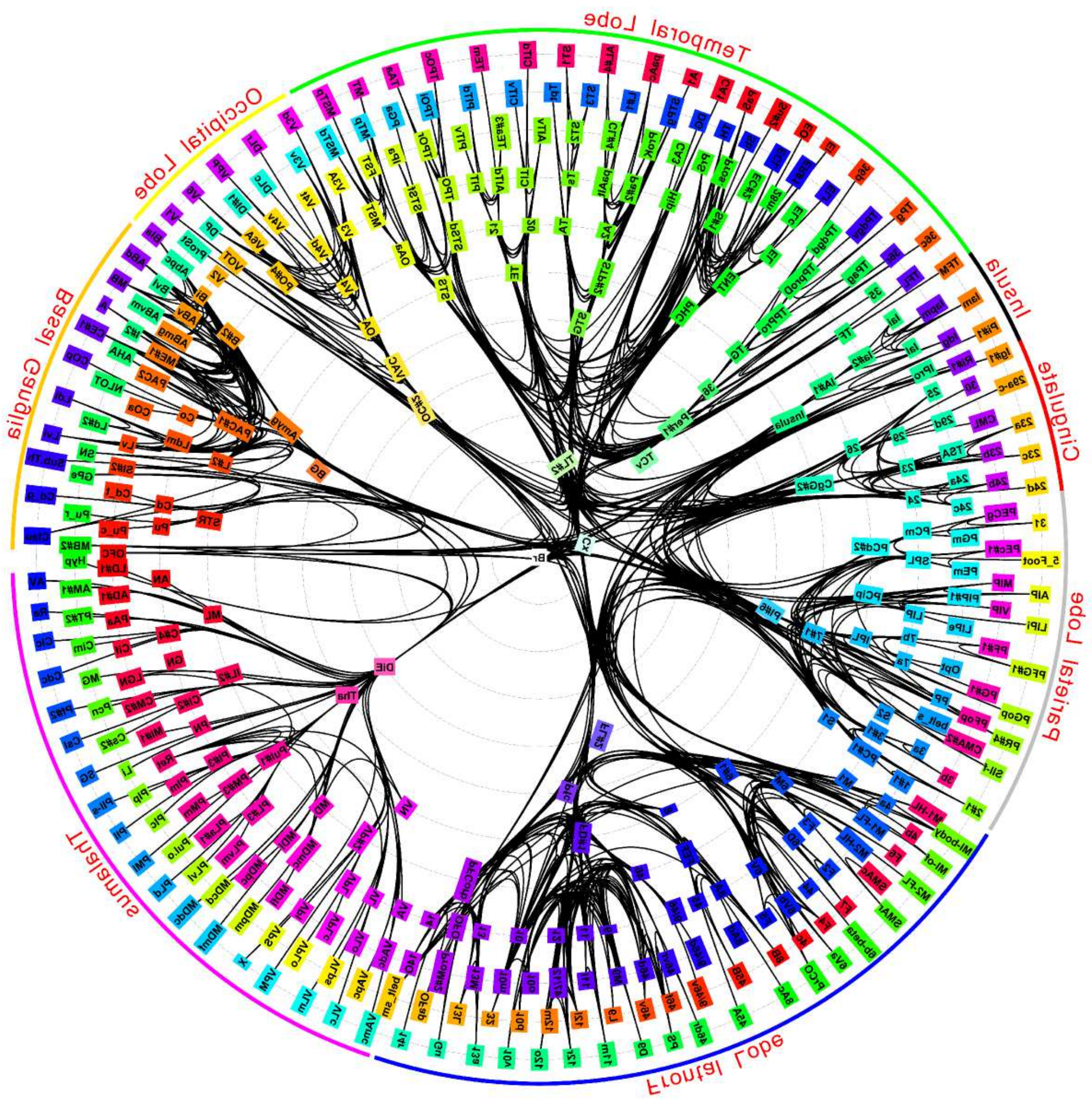




Deep learning

The brain is a highly interconnected, dynamic network organized hierarchically, in parallel, and at multiple scales



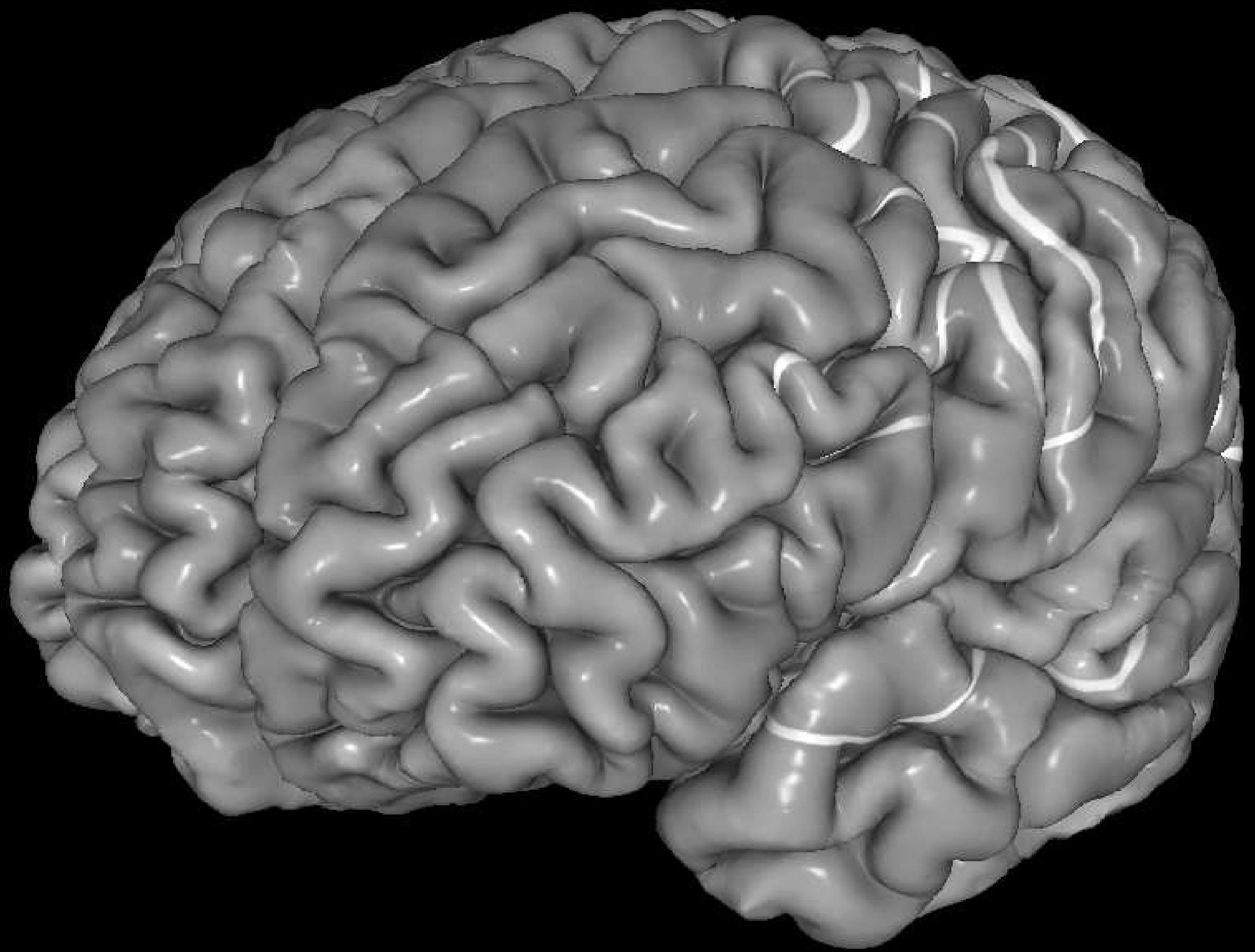


1. Multi-scale organization provides weak computational compartmentalization.
2. Brain connections are many-to-many and recurrent.
3. Brain representations are highly modulated by plans and goals.
4. The brain “learns” continuously and at multiple timescales.

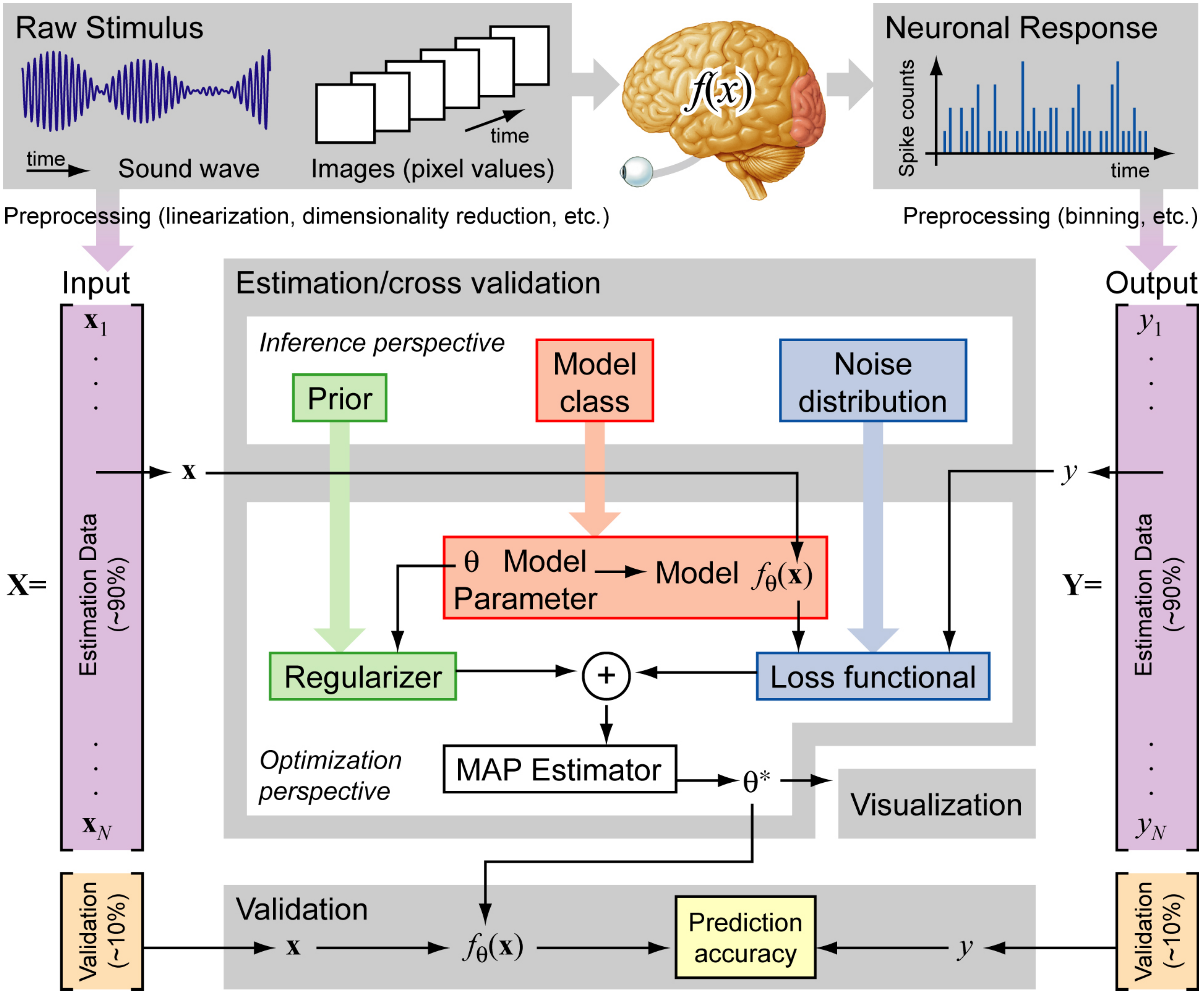
(right) Modha & Singh, 2010

In neuroscience we are always data-limited...

A central problem in systems and cognitive neuroscience is to model the representation of information across the brain

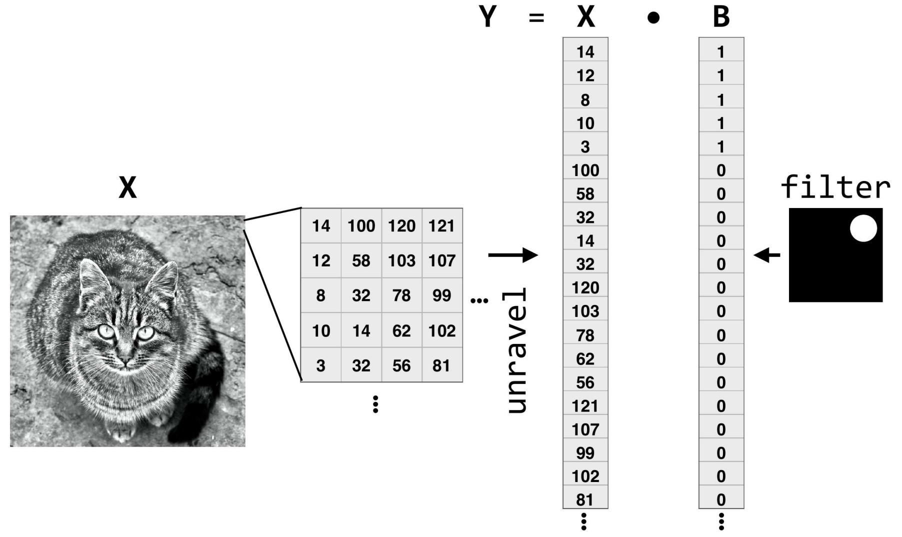


The best classical way to model representation in the brain is to use some form of the GLM



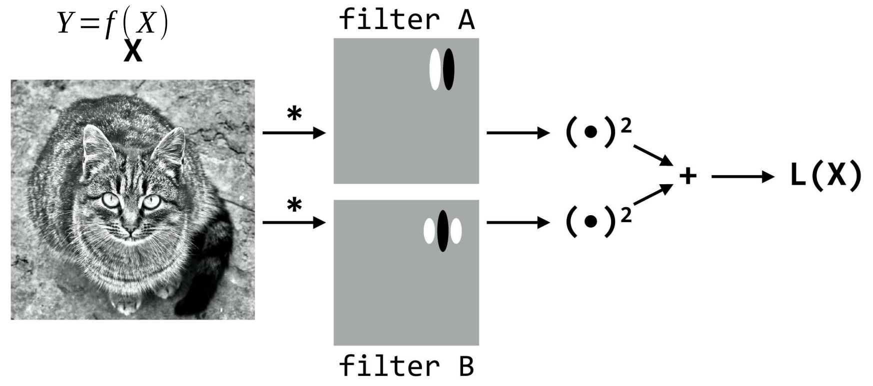
Linear model

$$Y = X\beta$$



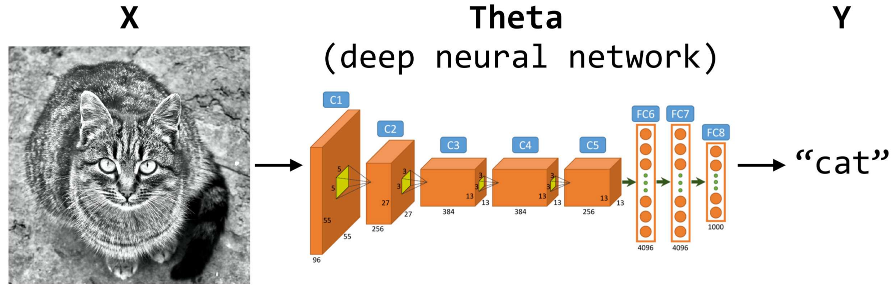
Linearized model

$$Y = L(X)\beta$$



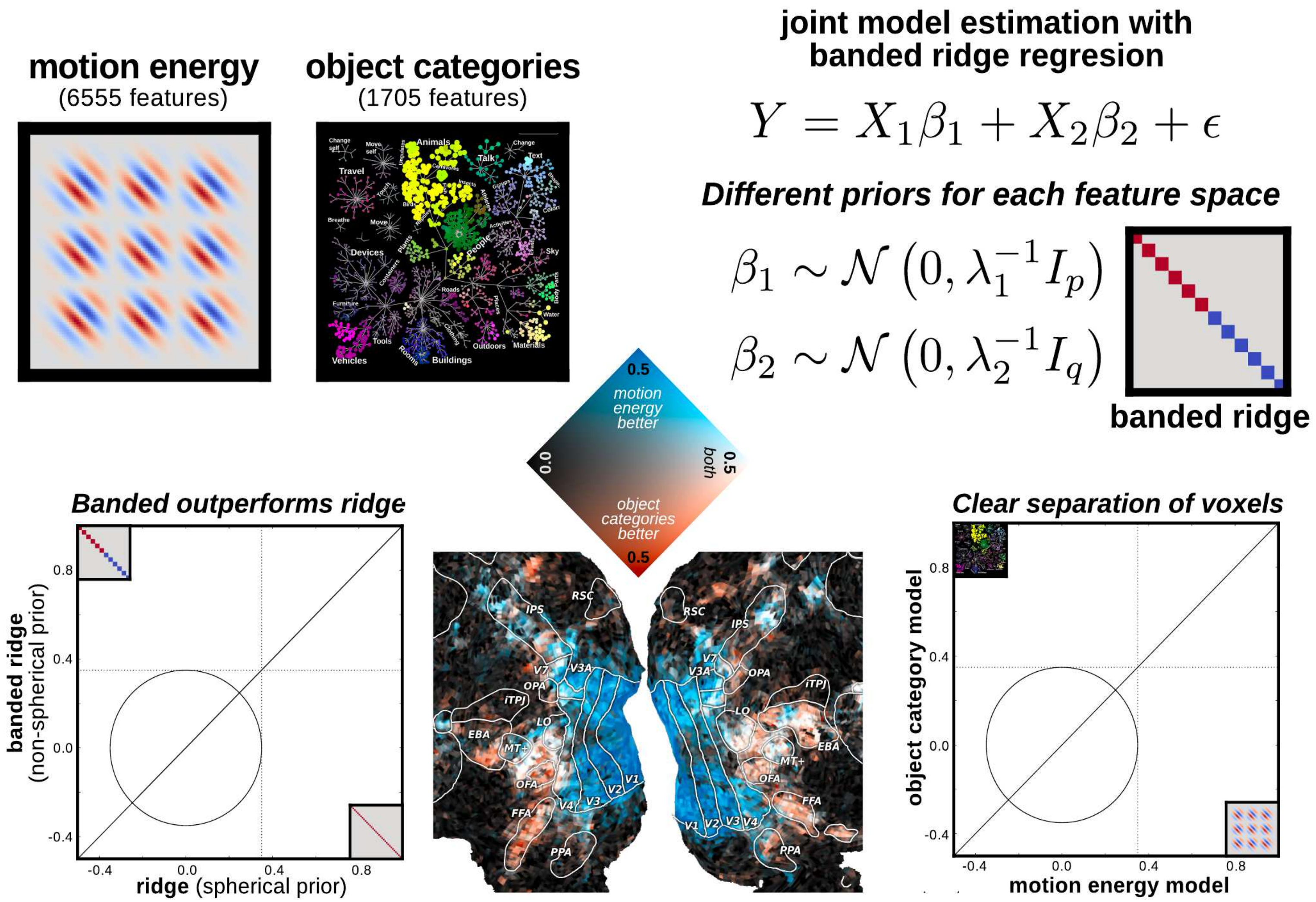
Nonlinear model

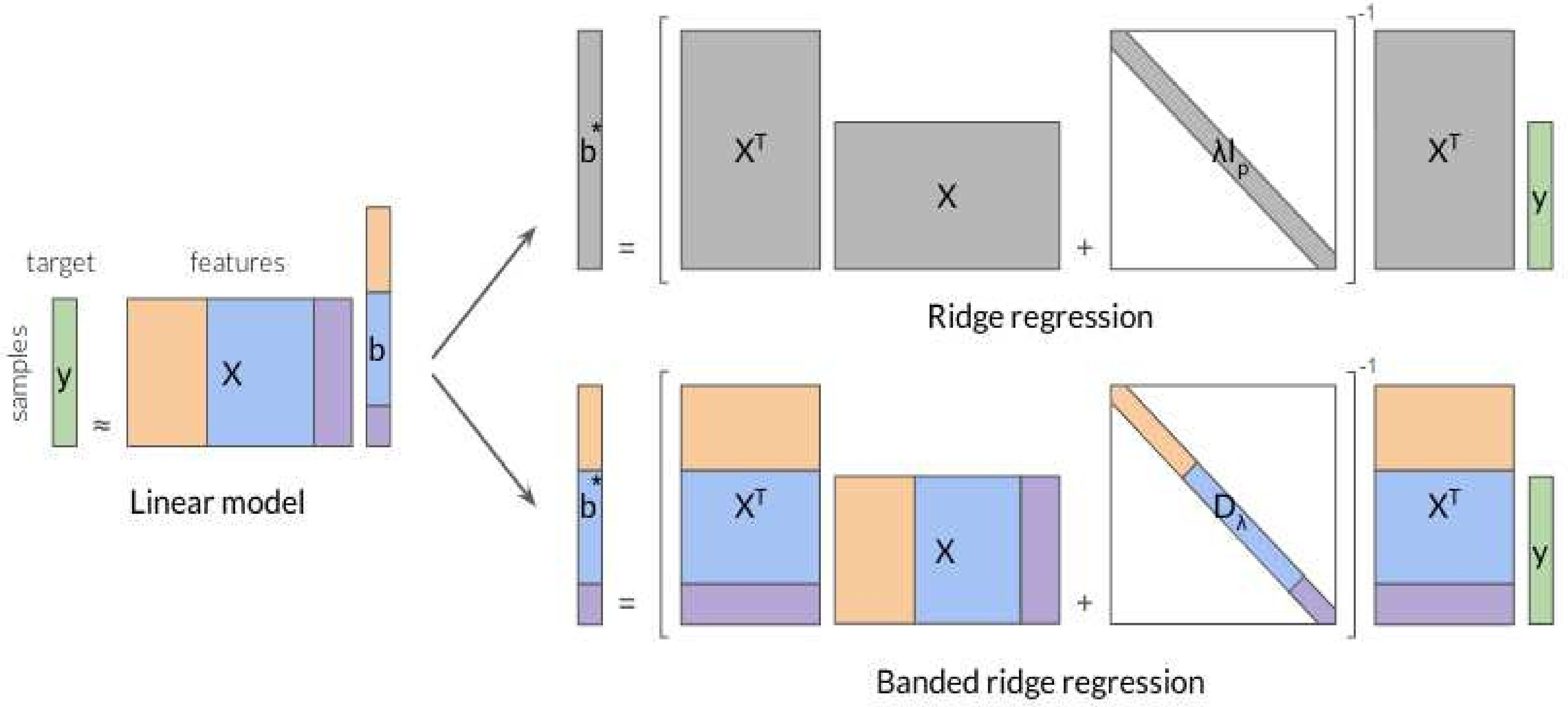
$$Y = \theta(X)\beta$$



Wu, David and Gallant, *Ann. Rev. Neuro.*, 2006

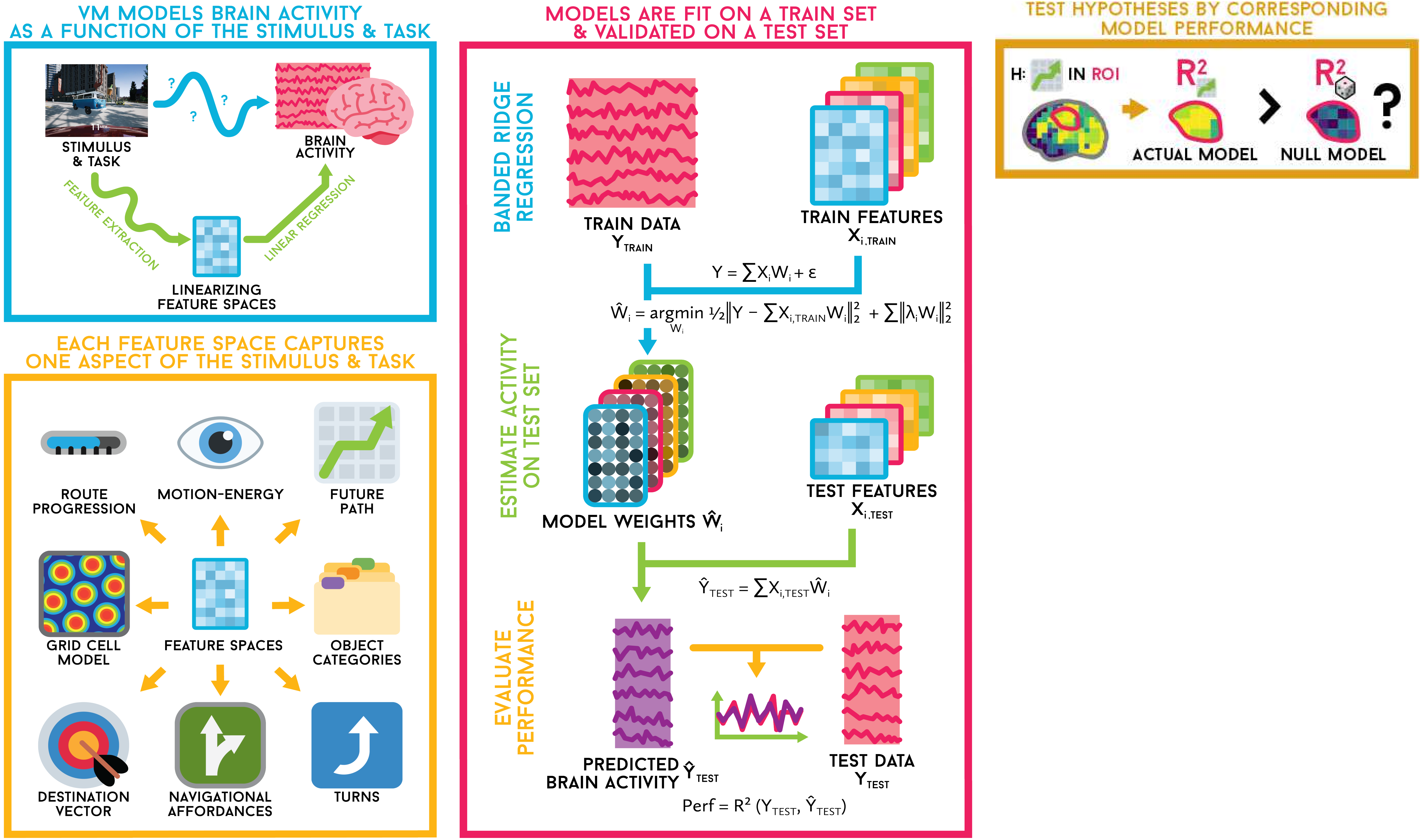
Because brain data have low SNR, regularization must be carefully managed across different feature spaces





Nunez-Elizalde, Huth & Gallant, *NeuroImage*, 2019

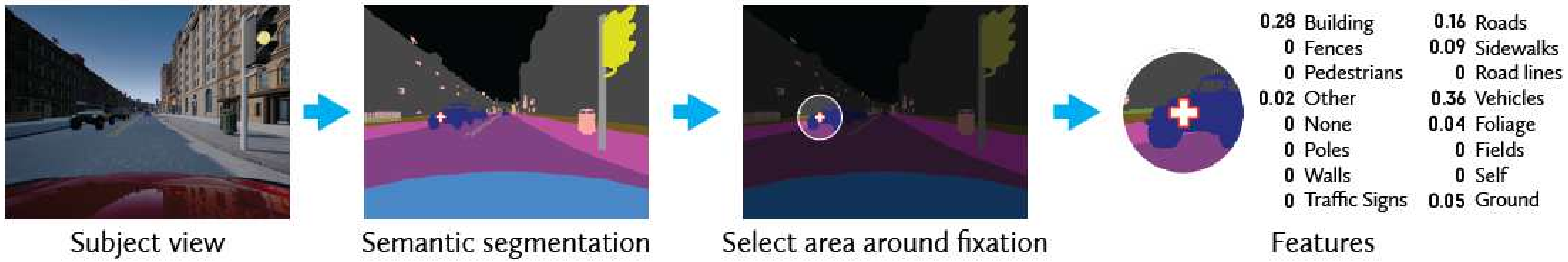
We use an encoding model approach to fit multiple navigation-related feature spaces to each voxel in each subject

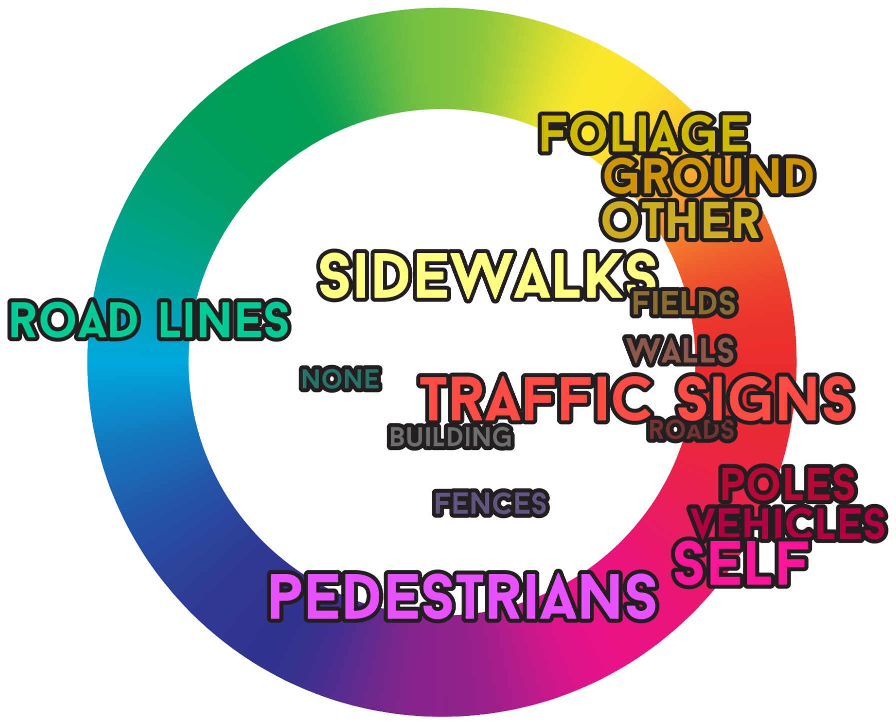


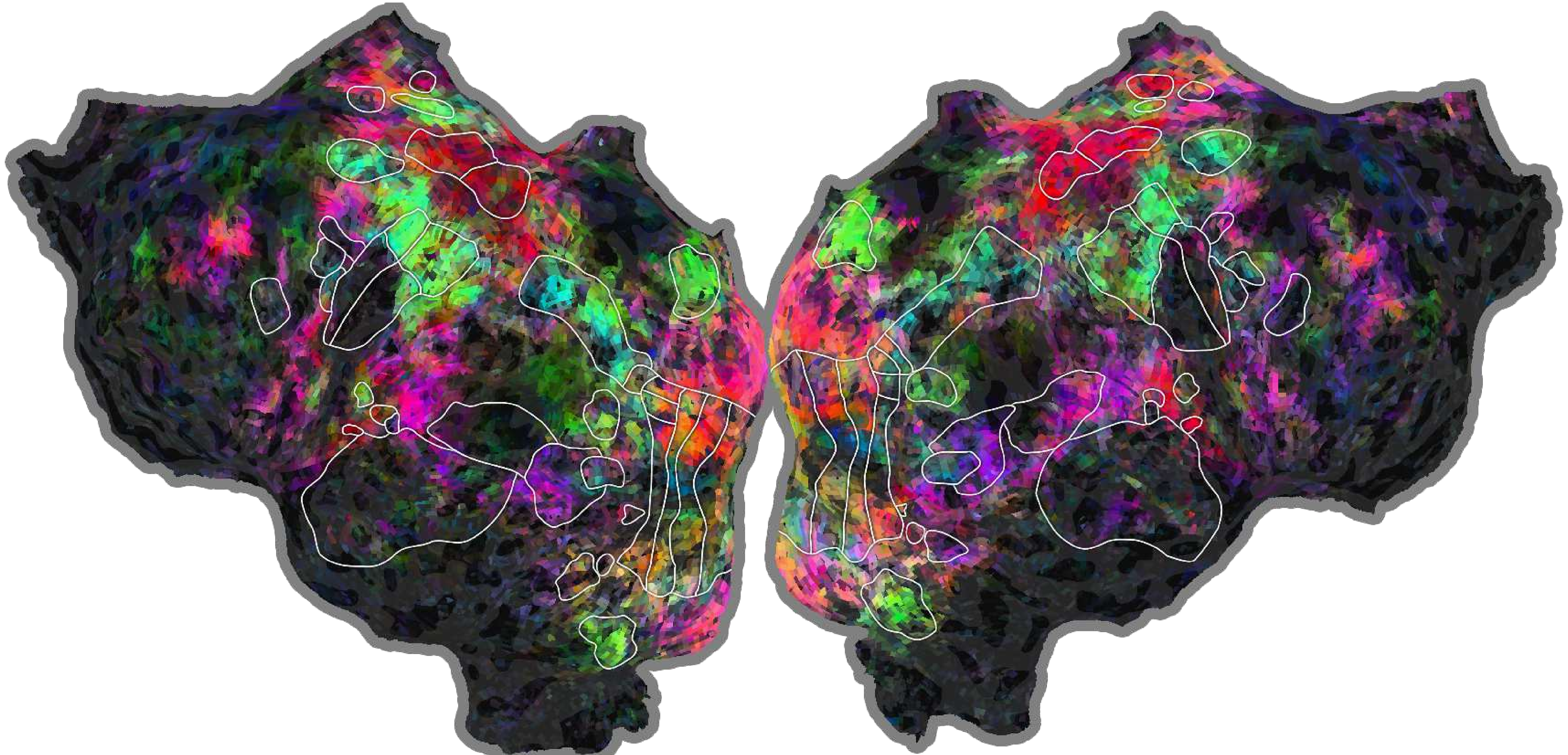
Tianjiao Zhang

Zhang & Gallant, in prep

We visualize the representation of navigation-related features by projecting voxelwise model weights onto the cortical surface







Zhang & Gallant, in prep

To visualize the general distribution of navigation-related representations we use a low-dimensional embedding

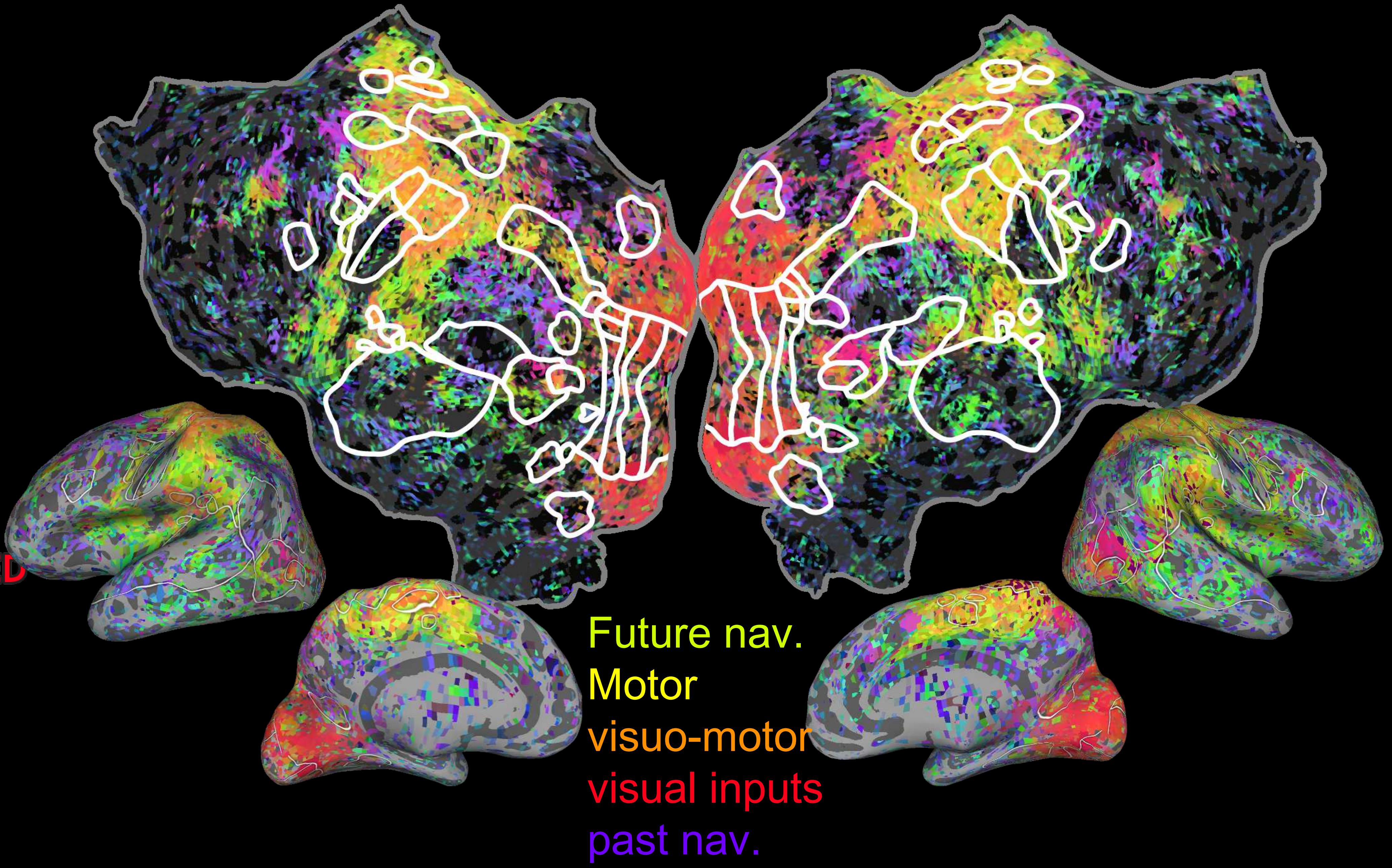
FUTURE NAVIGATION
GRAPH
PATH DISTANCE REMAINING
FUTURE PATH
GAZE DIRECTION
BEELINE DISTANCE REMAINING

MOTOR ACTIONS
CONTROLS
SCENE STRUCTURE
DESTINATION GRID REPRESENTATION
TURN SPACE PHASE
GAZE GRID

VISUO-MOTOR
MOTION-ENERGY RAW
DESTINATION ANCHORED VECTOR
TURN TIME PHASE
OTHER ENTITIES
GAZE SEMANTICS
AFFORDANCE
GRID CELLS
FRAME SEMANTICS
HEAD DIRECTION PHASE
EYETRACKING

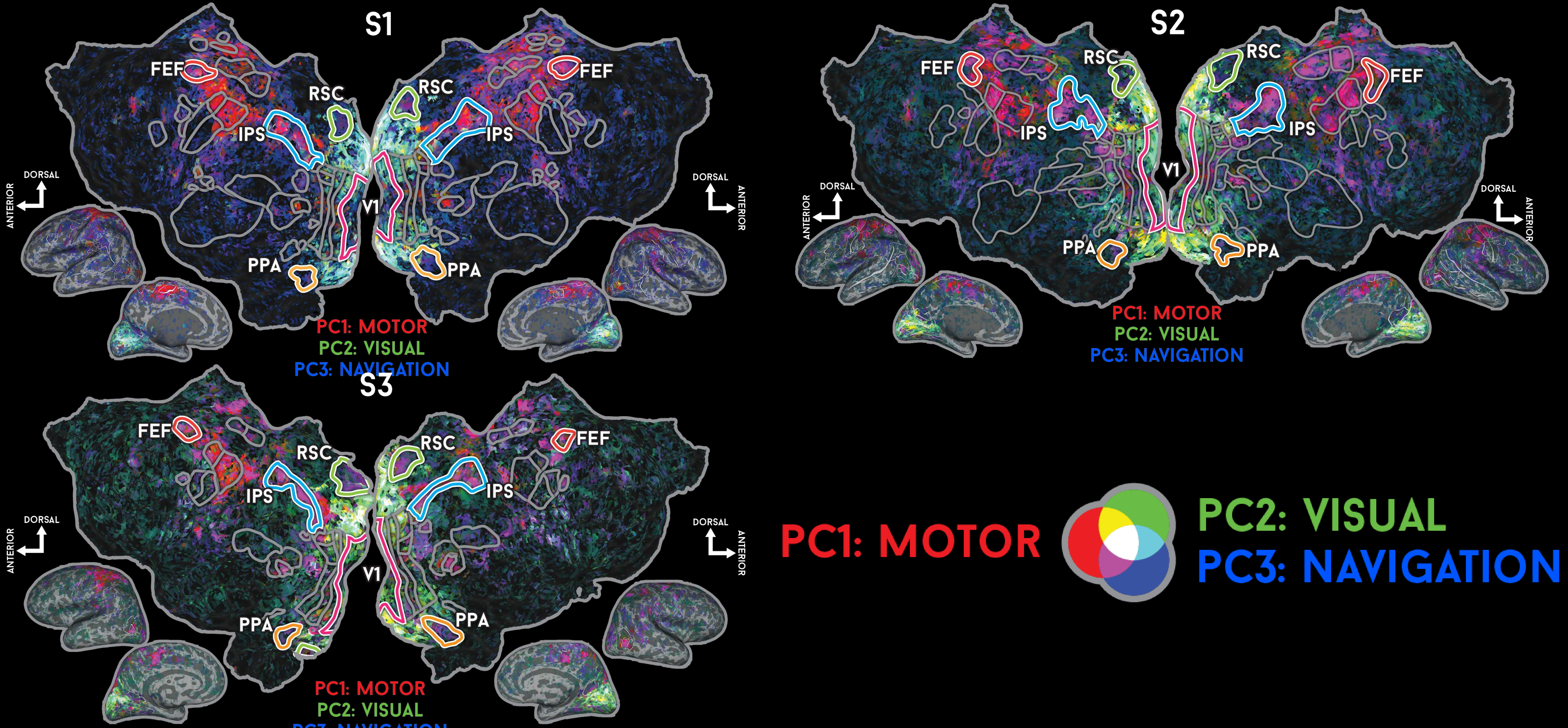
VISUAL INPUTS
MOTION-ENERGY RECENTERED
ROUTE SPACE PHASE
DEPTH
SPATIAL SEMANTICS
PATH DISTANCE ELAPSED
ROUTE TIME PHASE

PAST NAVIGATION
PATH INTEGRATION ALLOCENTRIC
DESTINATION VECTOR LOG
PATH INTEGRATION EGOCENTRIC
BEELINE DISTANCE ELAPSED



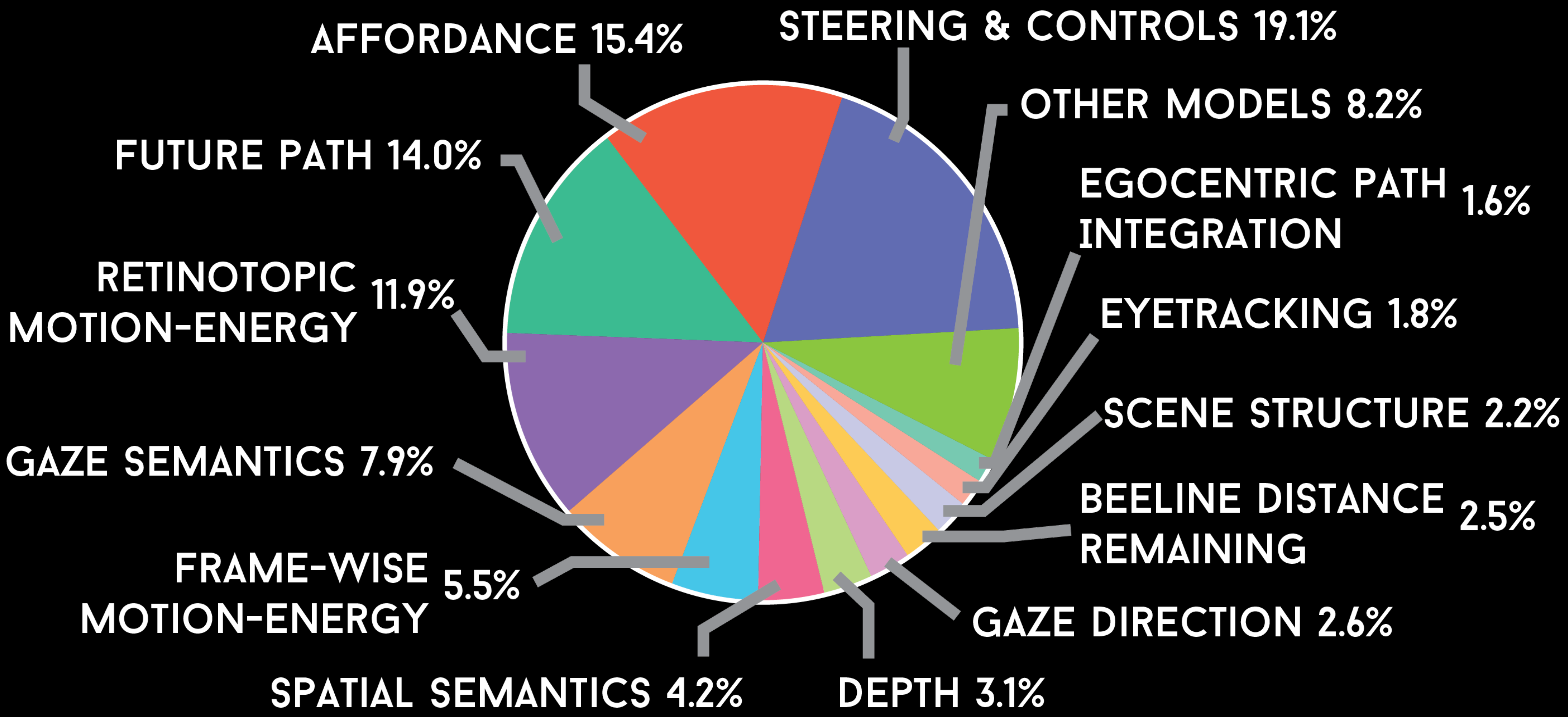
Zhang & Gallant, in prep

PCA of these data reveals that navigation-related networks are organized into three main functional classes



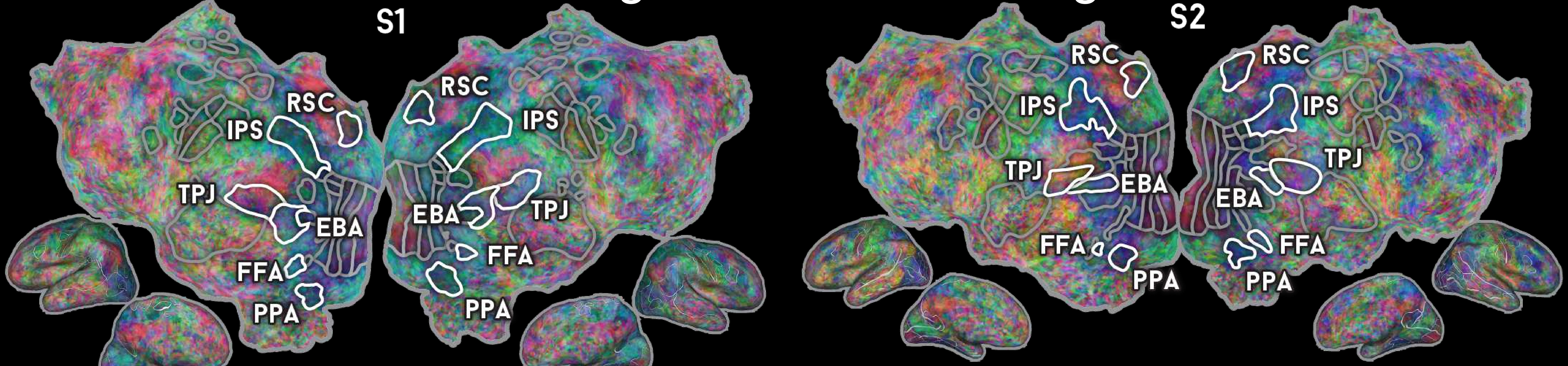
Zhang & Gallant, in prep

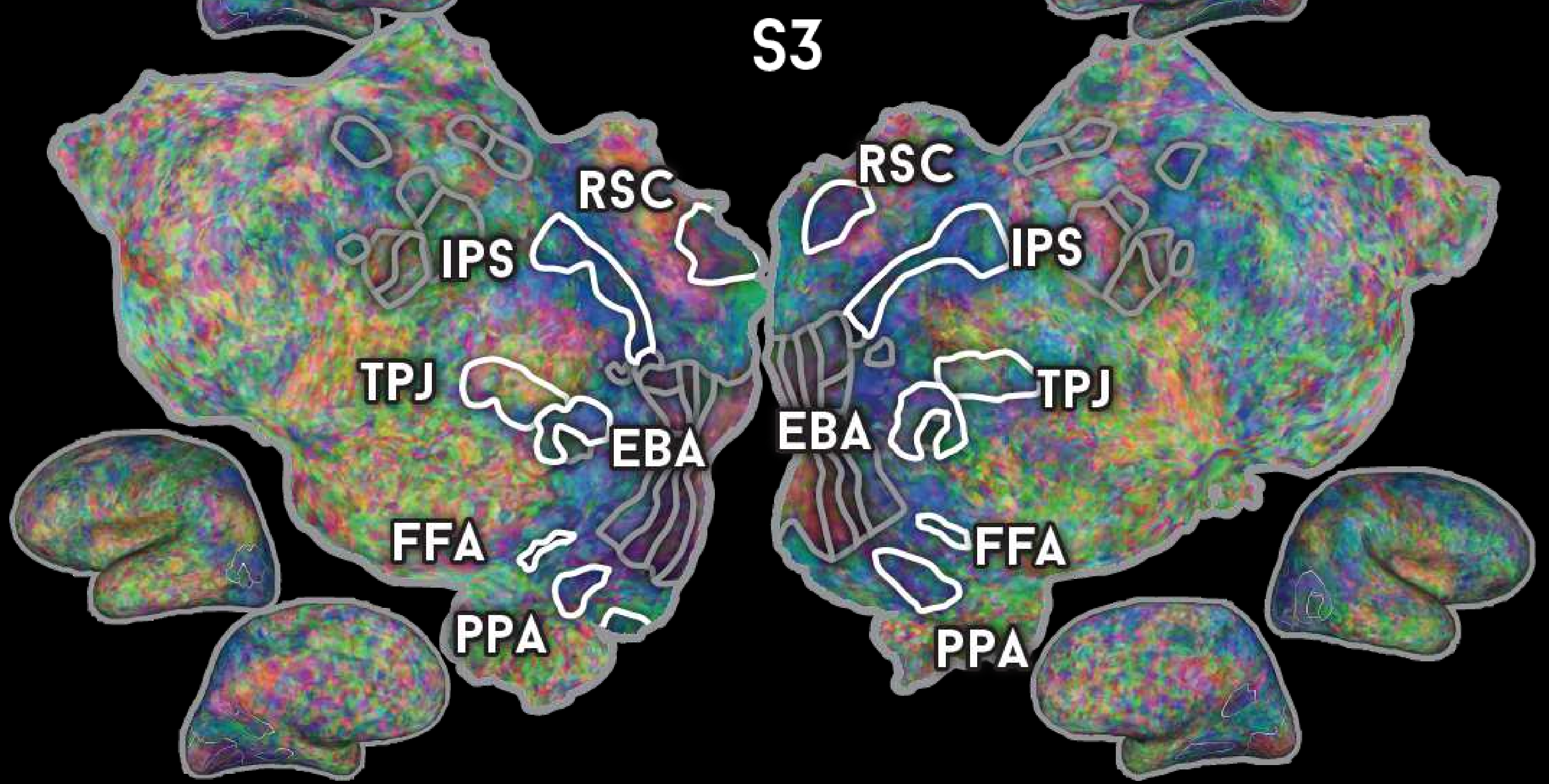
Most of the variance in brain activity is explained by variables related to perceptual, motor and goal-directed behavior

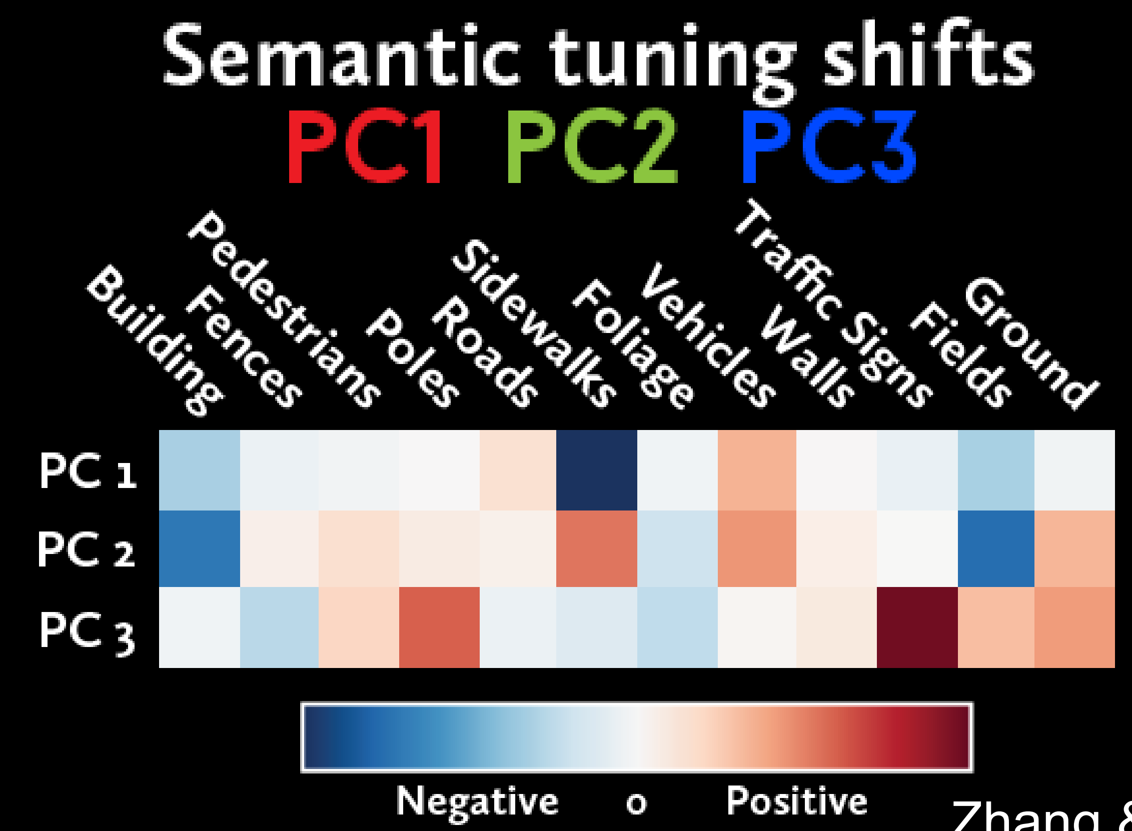


Zhang & Gallant, in prep

Active navigation shifts semantic tuning toward navigation-relevant categories

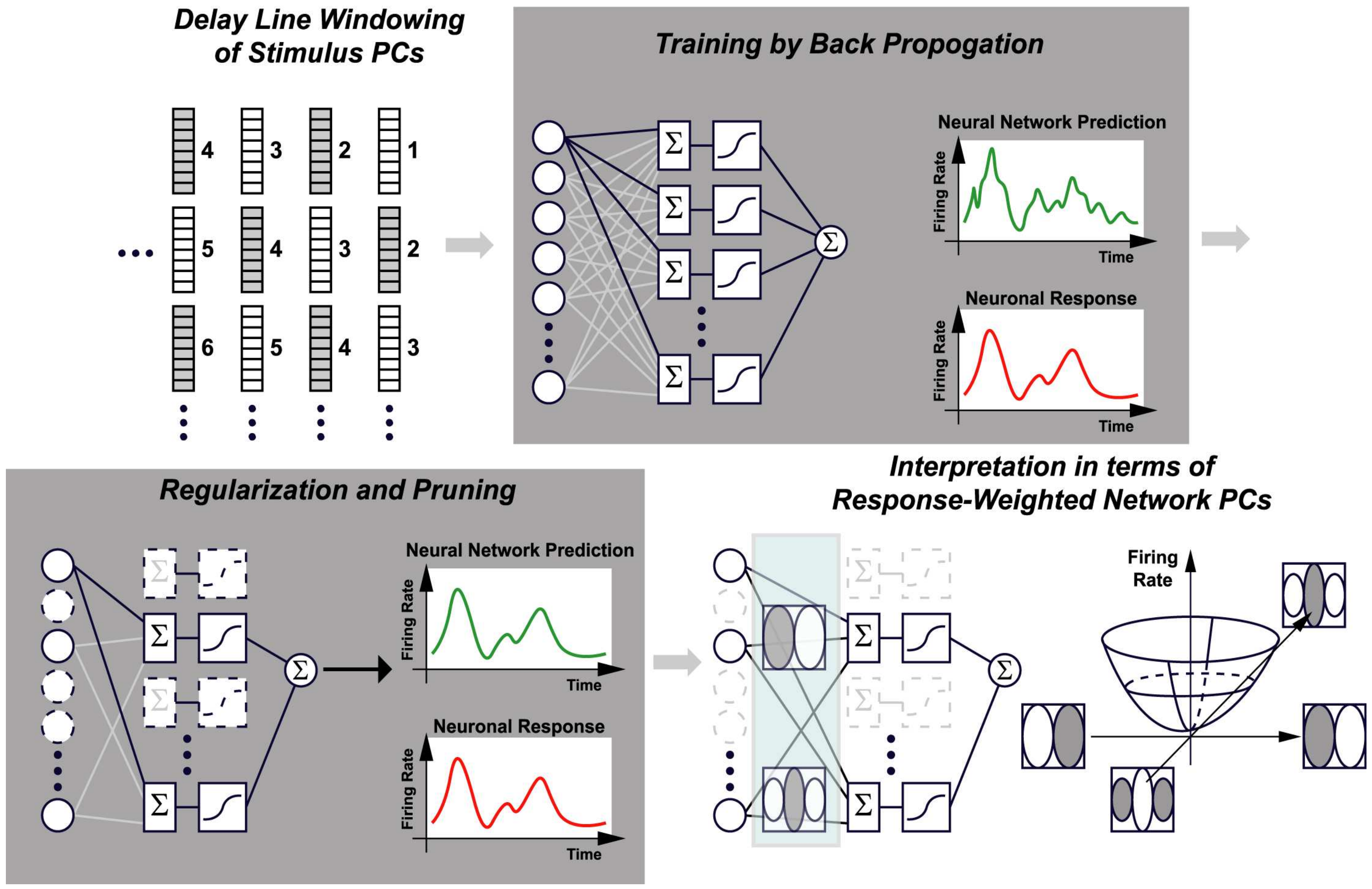


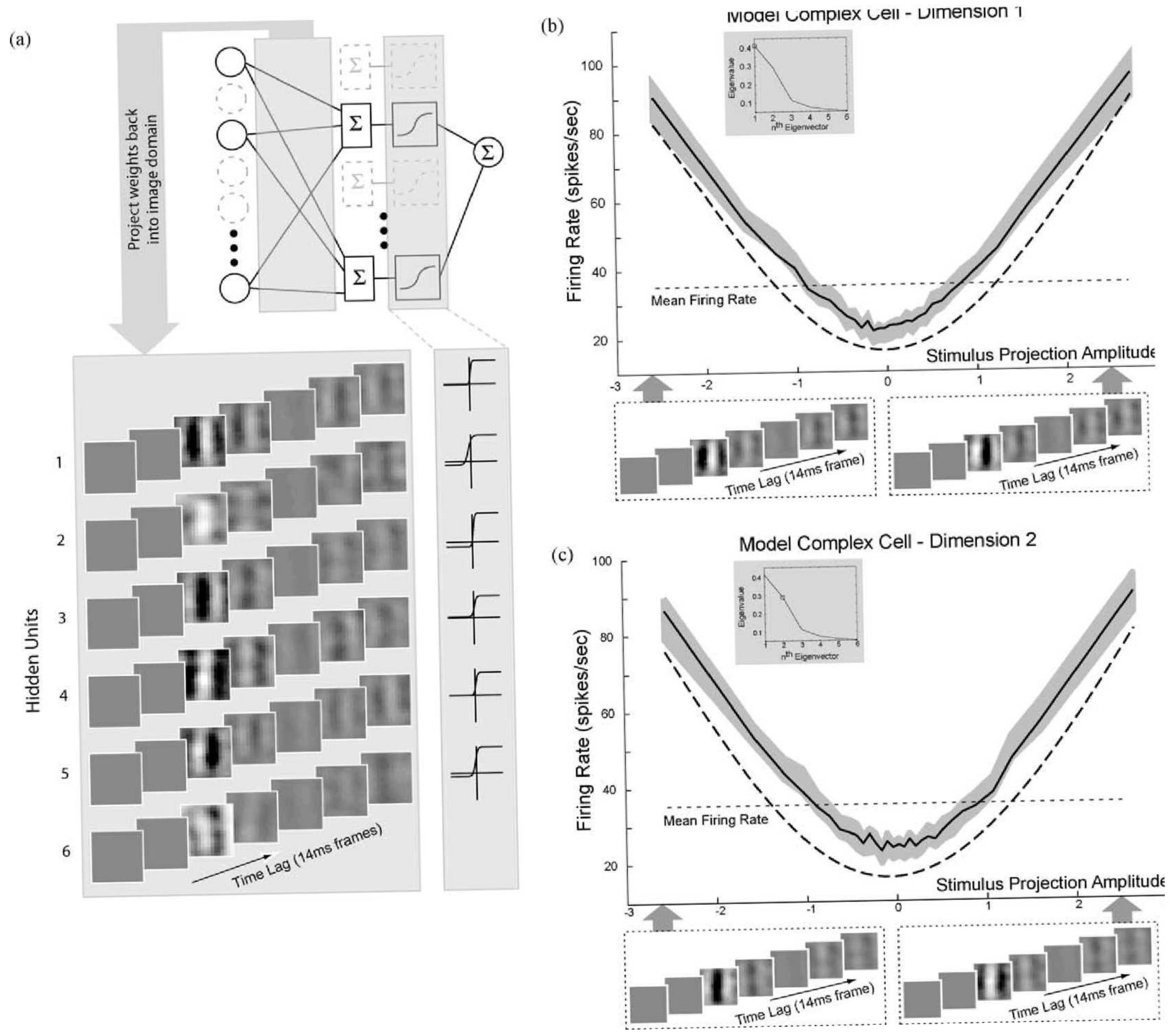




Zhang & Gallant, in prep

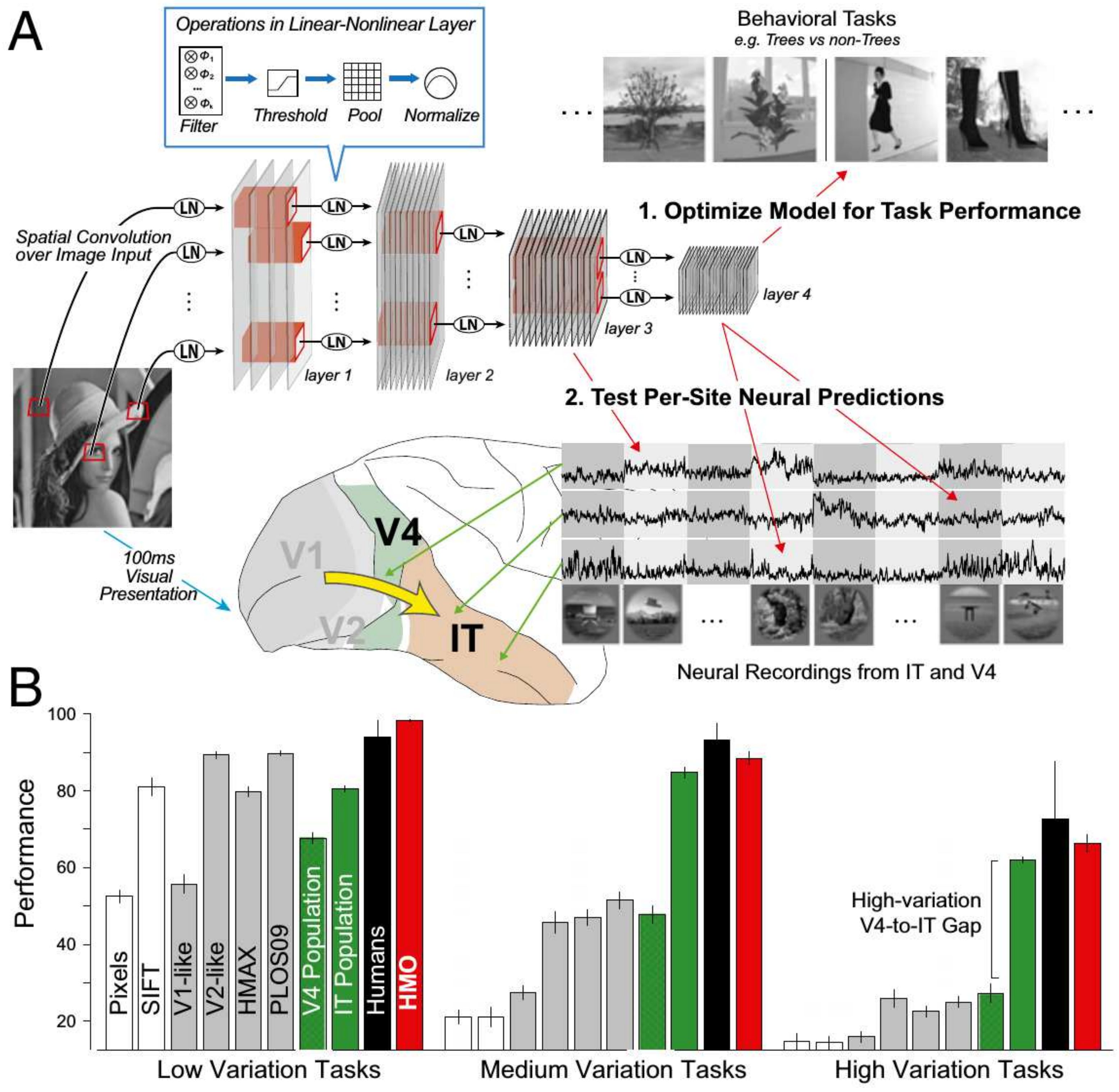
Can we use deep networks to model brain data directly?





Prenger, Wu, David and Gallant, N*eural Networks*, 2004
see also Lau, Stanley & Dan, *PNAS*, 2002

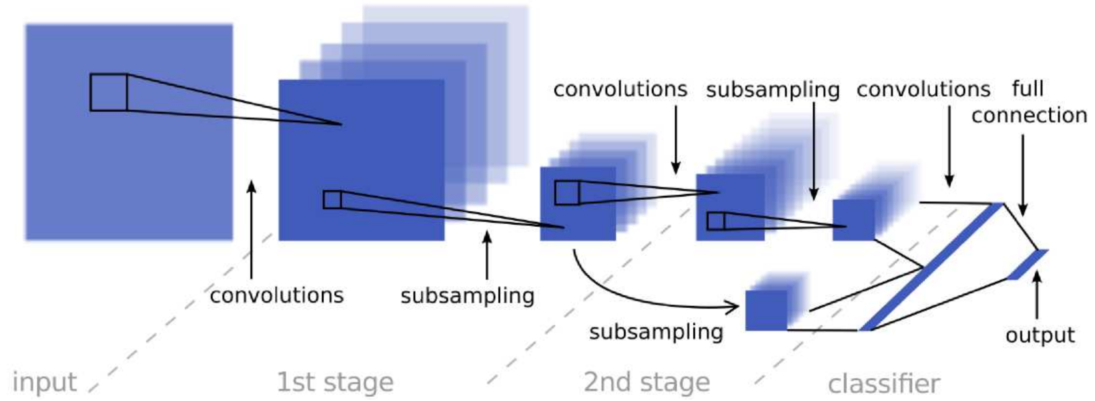
Can we use the learned weights of pre-trained deep network as a source of features for the GLM?

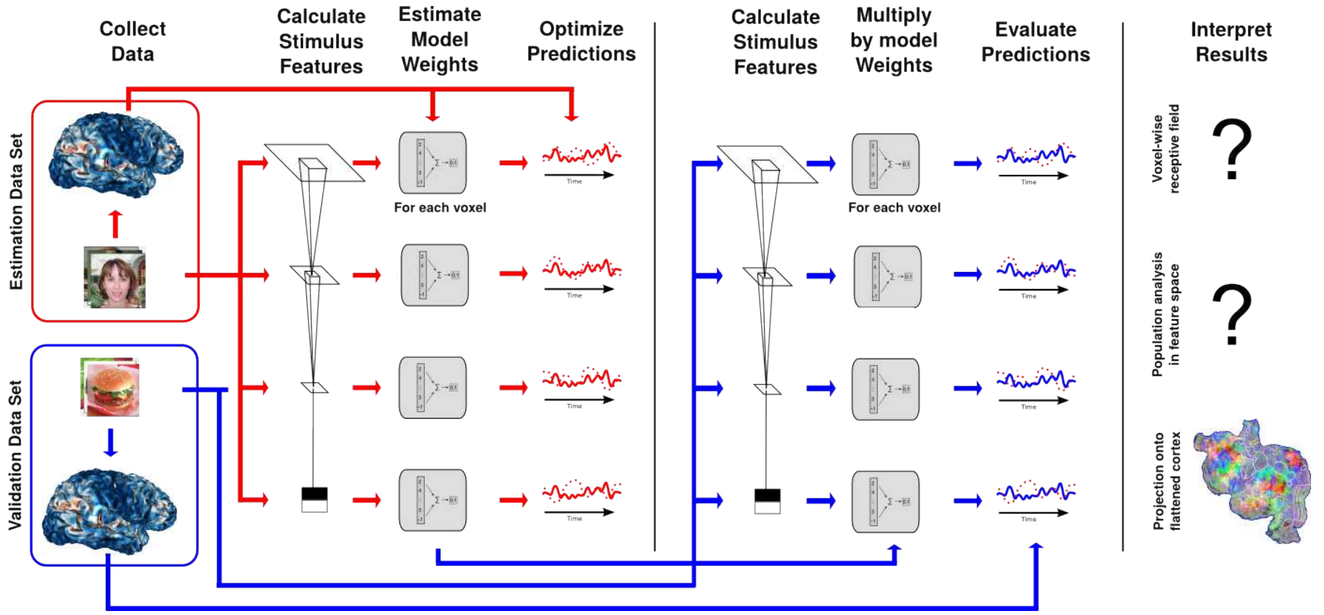


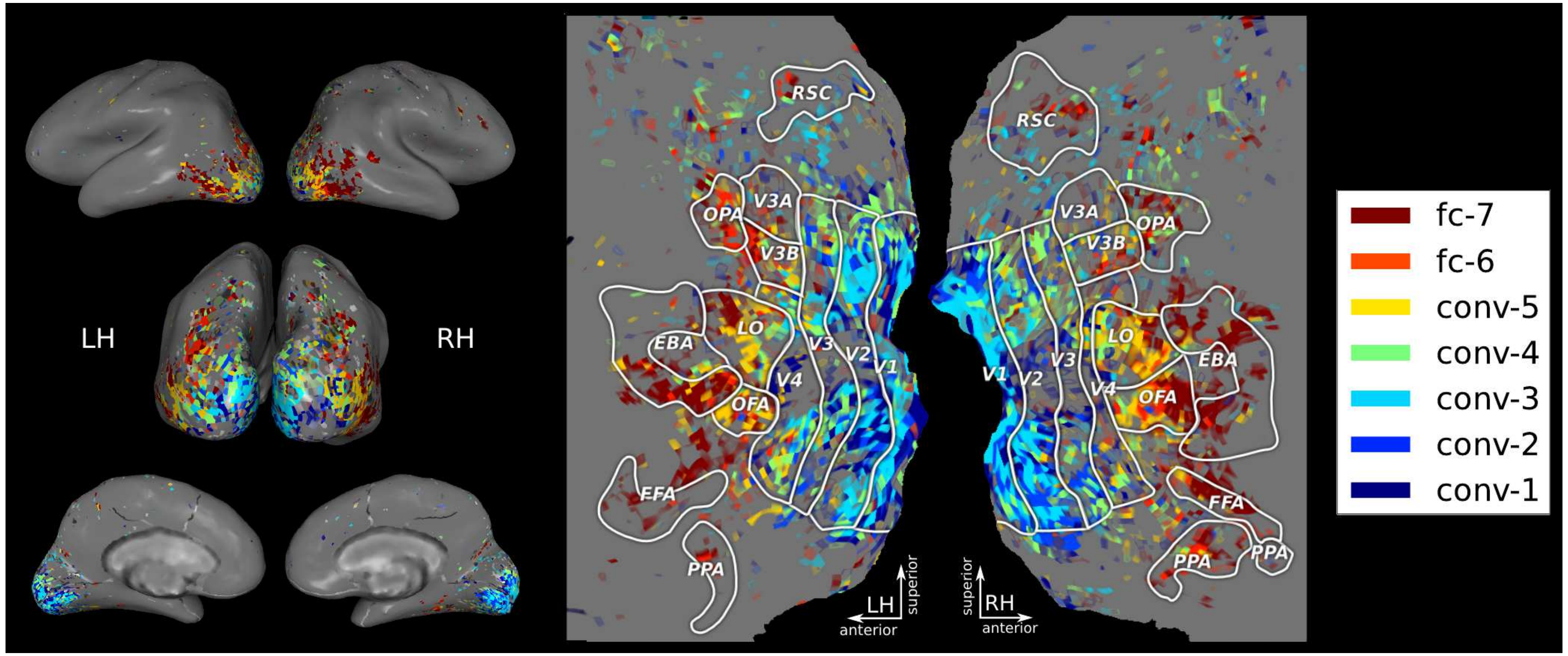
Yamins, Hong, Cadieu, Solomon, Seibert & DiCarlo, *PNAS*, 2014

Can we use the learned weights of pre-trained deep network as a source of features for the GLM?

Pulkit Agrawal

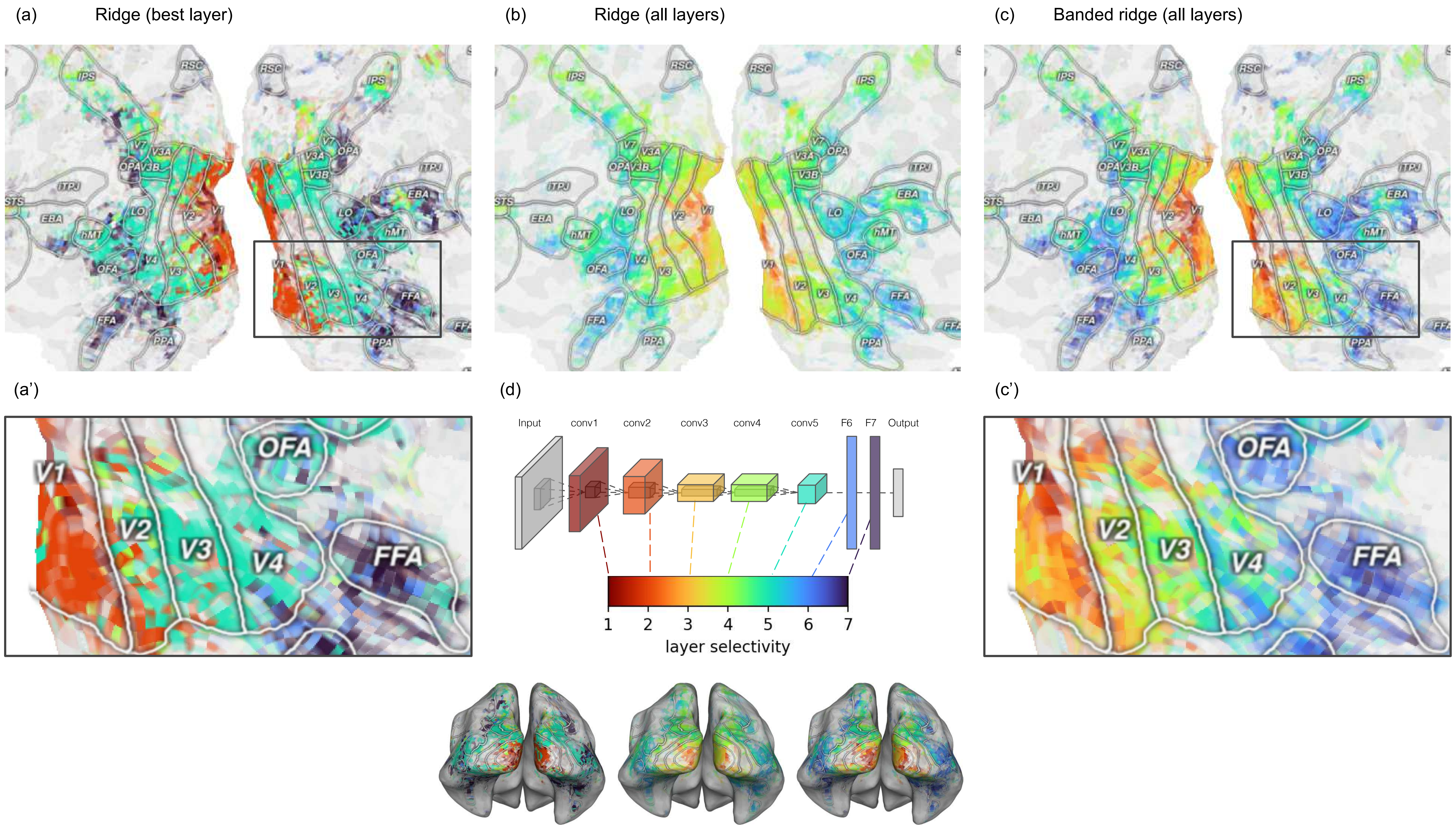






Agrawal, Stansbury, Malik and Gallant, arXiv, 2014

One problem with pre-trained deep networks is that their features are often correlated across layers



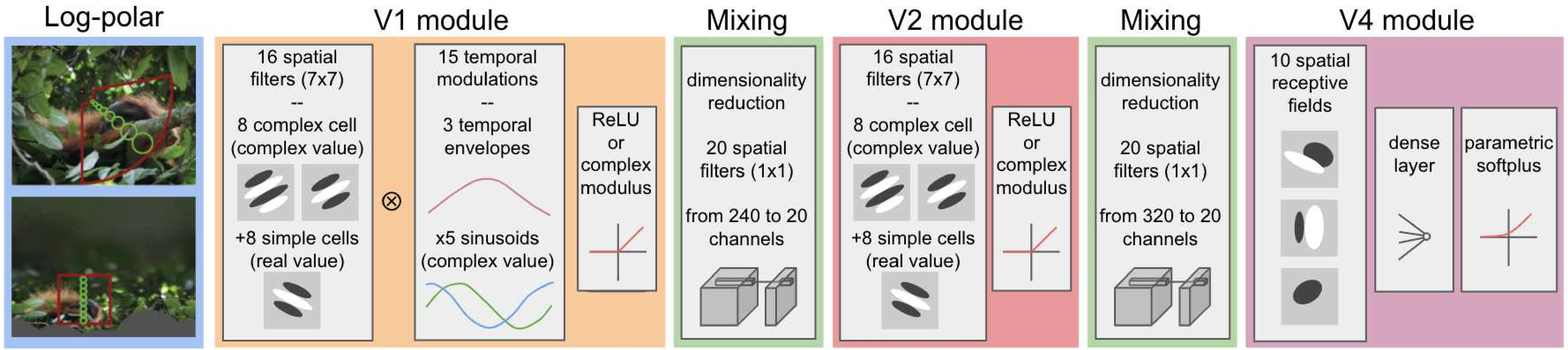
Tom
Dupre la Tour

Dupre la tour and Gallant, in prep

Can we develop a hybrid modeling scheme that uses deep learning to fit an explicit model?

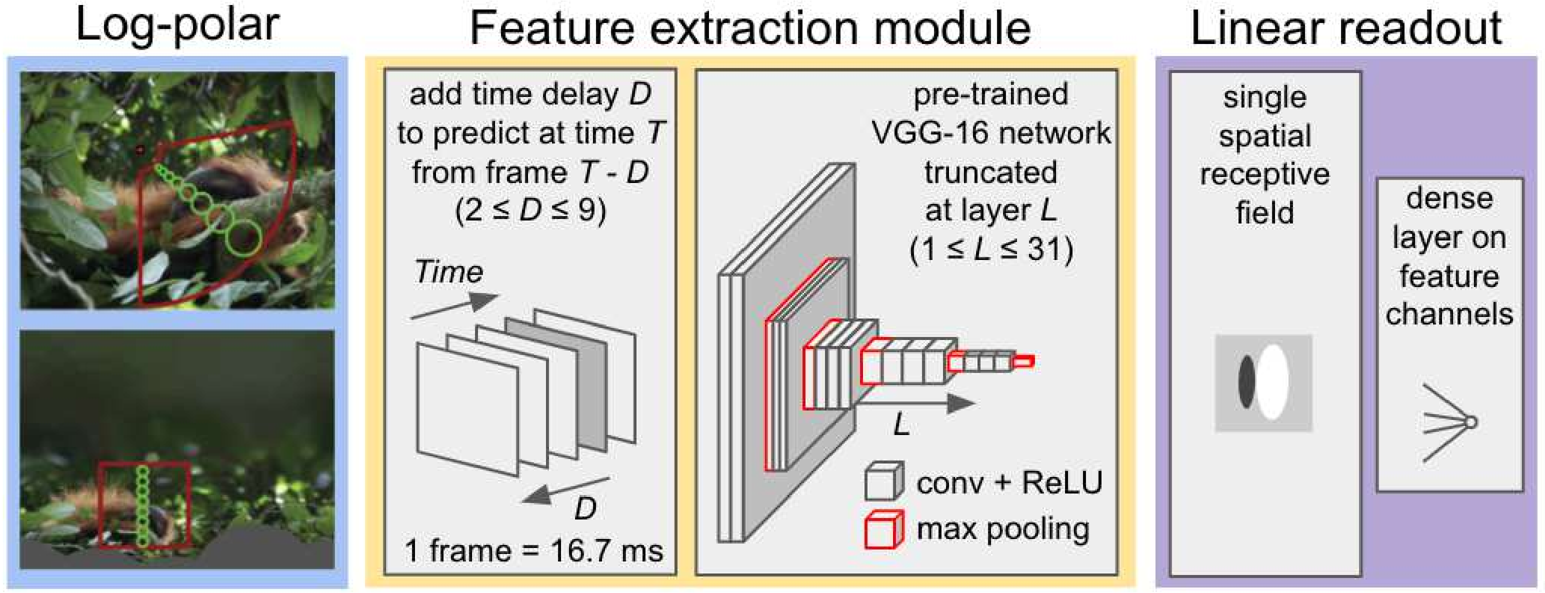
(a) Hierarchical convolutional energy (HCE) model

20,620 total parameters



(b) VGG Features (VGG-F) model

5,827 - 14,715,216 total parameters



(c)

HCE model layer	Module	Output shape	Parameters
Log-polar transform	Log-polar	(3, 64, 64)	0
V1 module	V1	(15, 16, 29, 29)	4767
Coupled Gaussian dropout	Mixing	(15, 16, 29, 29)	0
1x1 convolution	Mixing	(20, 29, 29)	4800
V2 module	V2	(320, 12, 12)	1192
Coupled Gaussian dropout	Mixing	(320, 12, 12)	0
1x1 convolution	Mixing	(20, 12, 12)	6400
Spatial filters	V4	(20, 10)	1440
Coupled Gaussian dropout	V4	(200)	0
Fully connected	V4	(10)	2010
Coupled Gaussian dropout	V4	(10)	0
Fully connected	V4	(1)	11
Parametric soft-plus	V4	(1)	2

Michael Oliver

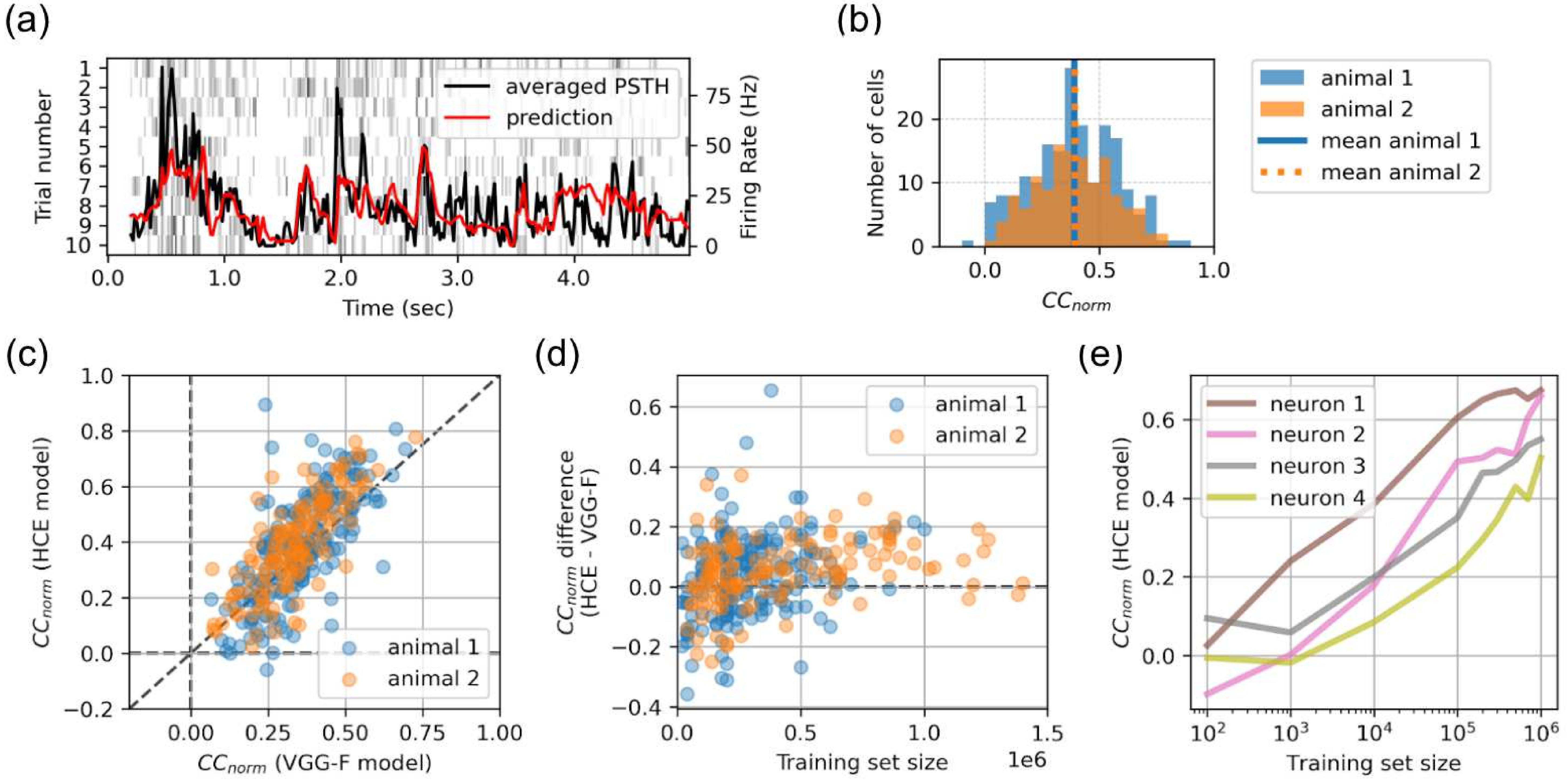
Michele Winter

Tom Dupre la Tour

Michael Eickenberg

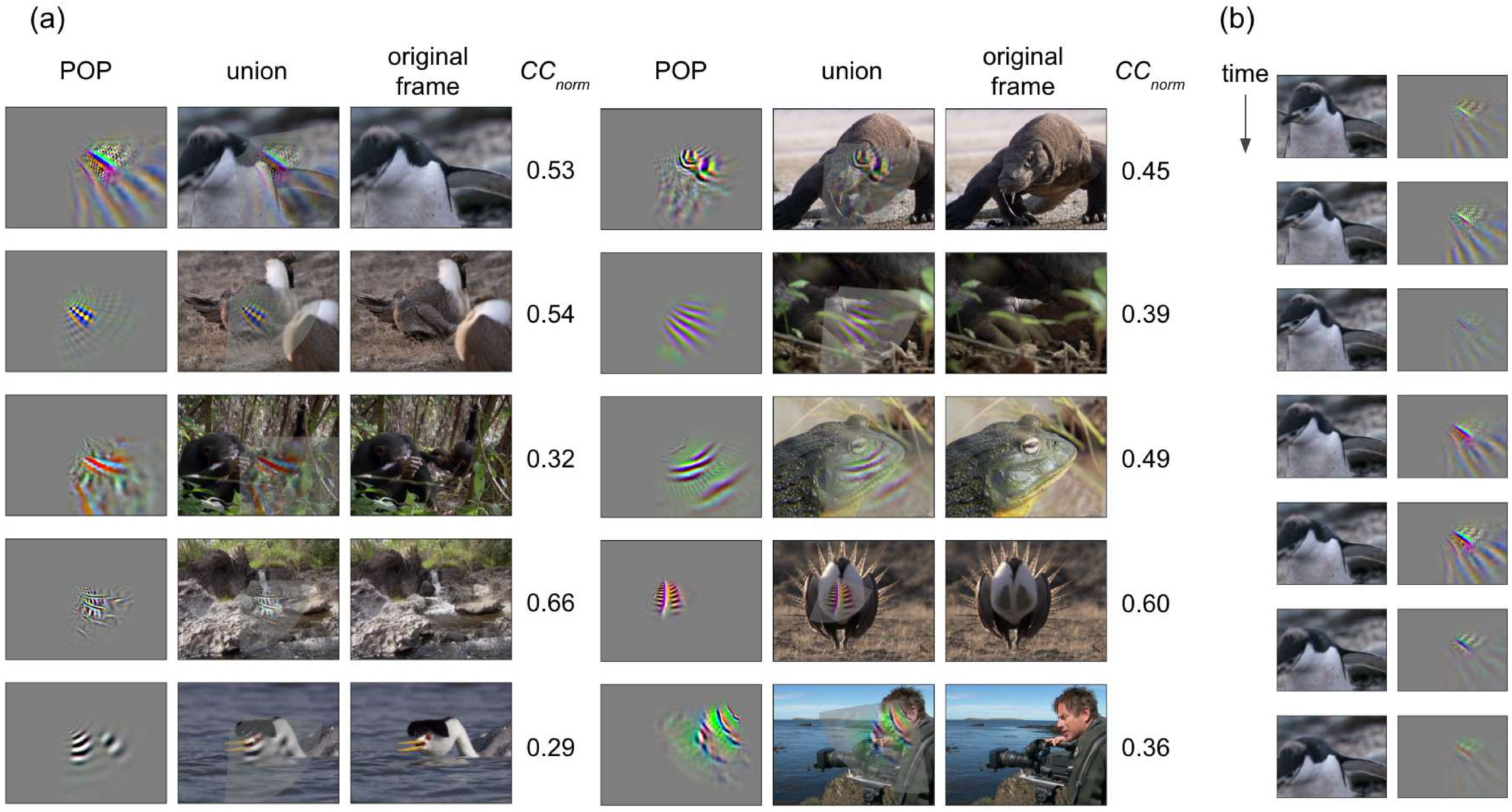
Oliver, Winter, Dupre la Tour, Eickenberg & Gallant, in prep

To test this idea we analyzed and modeled long-term recordings from 302 area V4 neurons



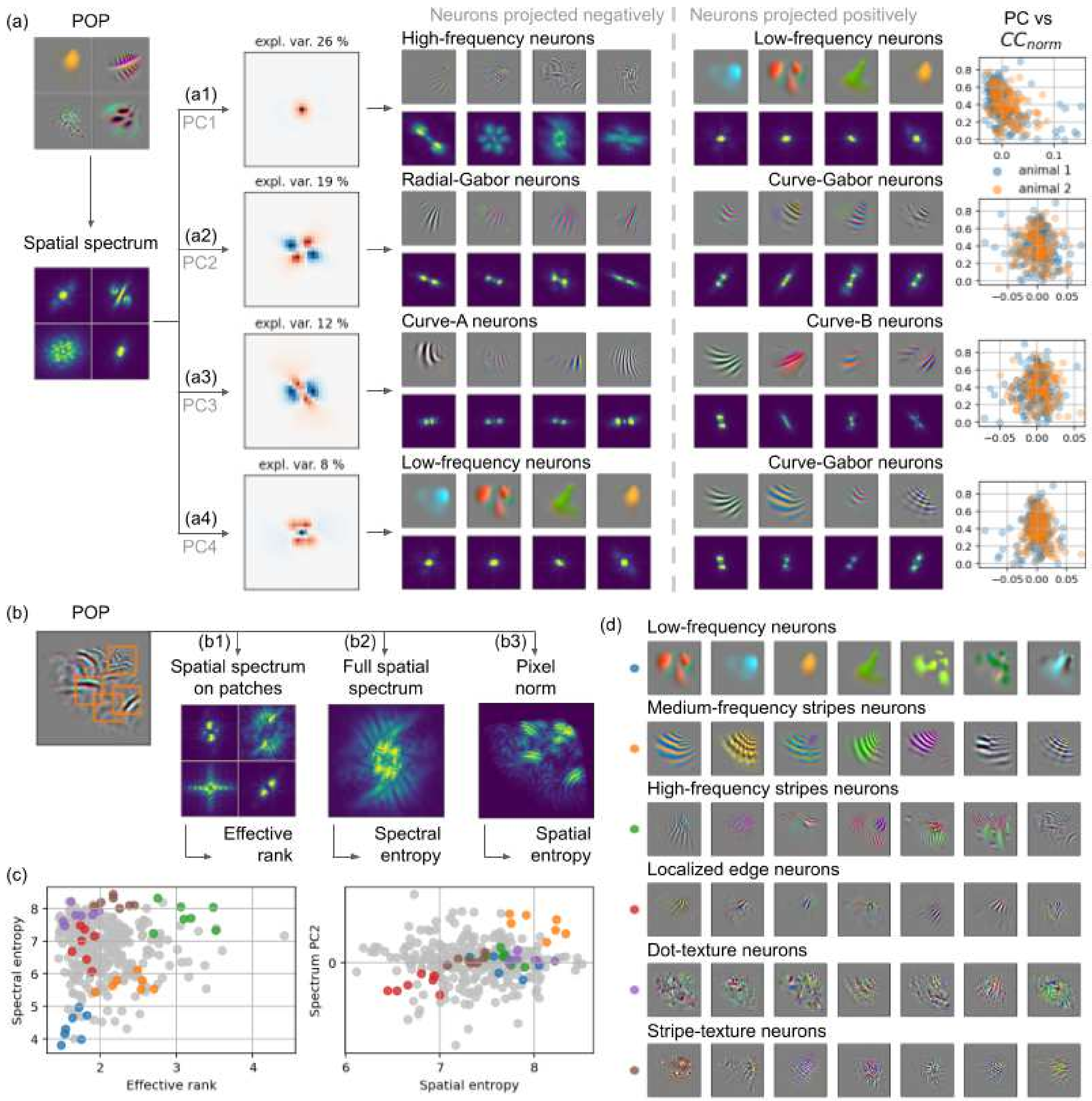
Oliver, Winter, Dupre la Tour, Eickenberg & Gallant, in prep

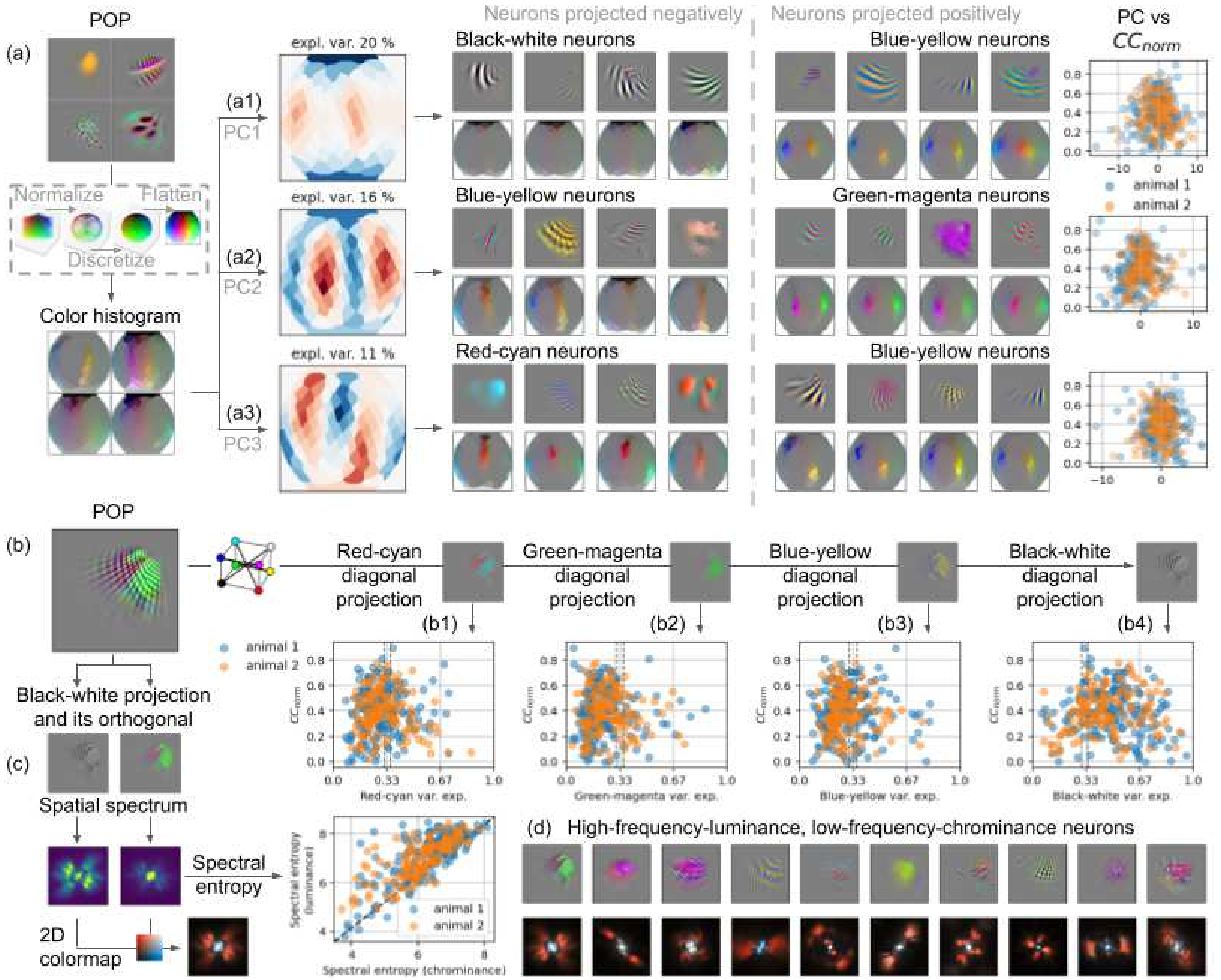
To better understand visual representation in area V4 we analyzed the predicted optimal patterns (POPs) for each cell



Oliver, Winter, Dupre la Tour, Eickenberg & Gallant, in prep

Spatial, chromatic and temporal tuning of the V4 sample can be recovered from feature-specific embeddings





Oliver, Winter, Dupre la Tour, Eickenberg & Gallant, in prep

Summary

The goals of science and engineering are somewhat different. Scientists tend to prioritize explanatory elegance, while engineers tend to prioritize utility (i.e., prediction accuracy and generalization).

The mammalian brain is a complex deep network that is organized hierarchically and in parallel, and which has complex dynamics. Measurement presents the most important current obstacle to understanding this system.

Modern methods of regression and data science provide the infrastructure necessary for fitting complex computational models to neuroscience data. When the model is described in terms of explicit transformations of measured stimulus-, task-, or behavior-related variables, the resulting models are directly interpretable.

When sufficient data are available, deep networks can be used in place of classical regression algorithms, by means of either supervised or unsupervised methods. However, the resulting networks are not directly interpretable.

Pre-trained deep networks can also be used as a source of features for classical regression algorithms. However, once again the resulting networks are not directly interpretable.

One little used approach is to leverage the infrastructure for training deep networks to fix explicit hierarchical computational models to brain data. The components of these models can be interpreted directly in terms of their basic computational properties. However, the function of the model as a whole may still be difficult to interpret.