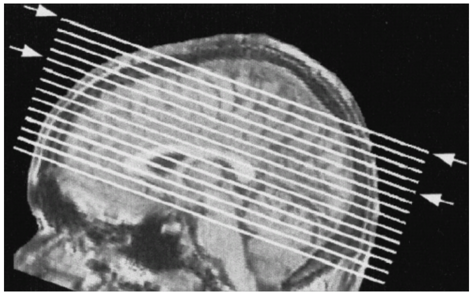
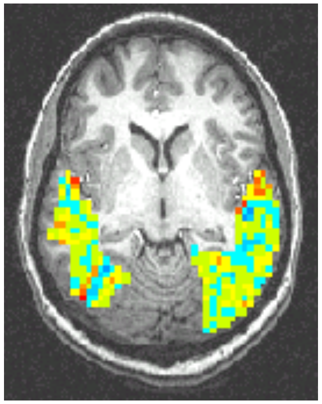
Understanding Neural Processes: Beyond Where and When, to How

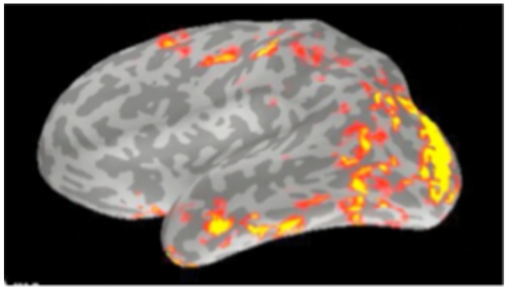


Tom M. Mitchell

Carnegie Mellon University

December 2019





Once we understand the brain, what will be the form of the answer?

Once we understand the brain, what will be the form of the answer?

It will have to include a description of the brain’s computational processes/algorithms.

Brain imaging studies have so far shown

- Where is neural activity that encodes information
- When this activity occurs during stimulus processing
- But not much about How the brain computes these

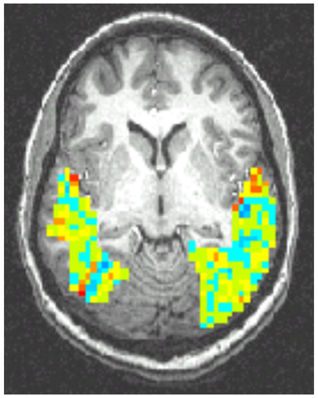
<u>Where</u>

does neural activity encode specific information?

Predicting fMRI Activity during Word Reading

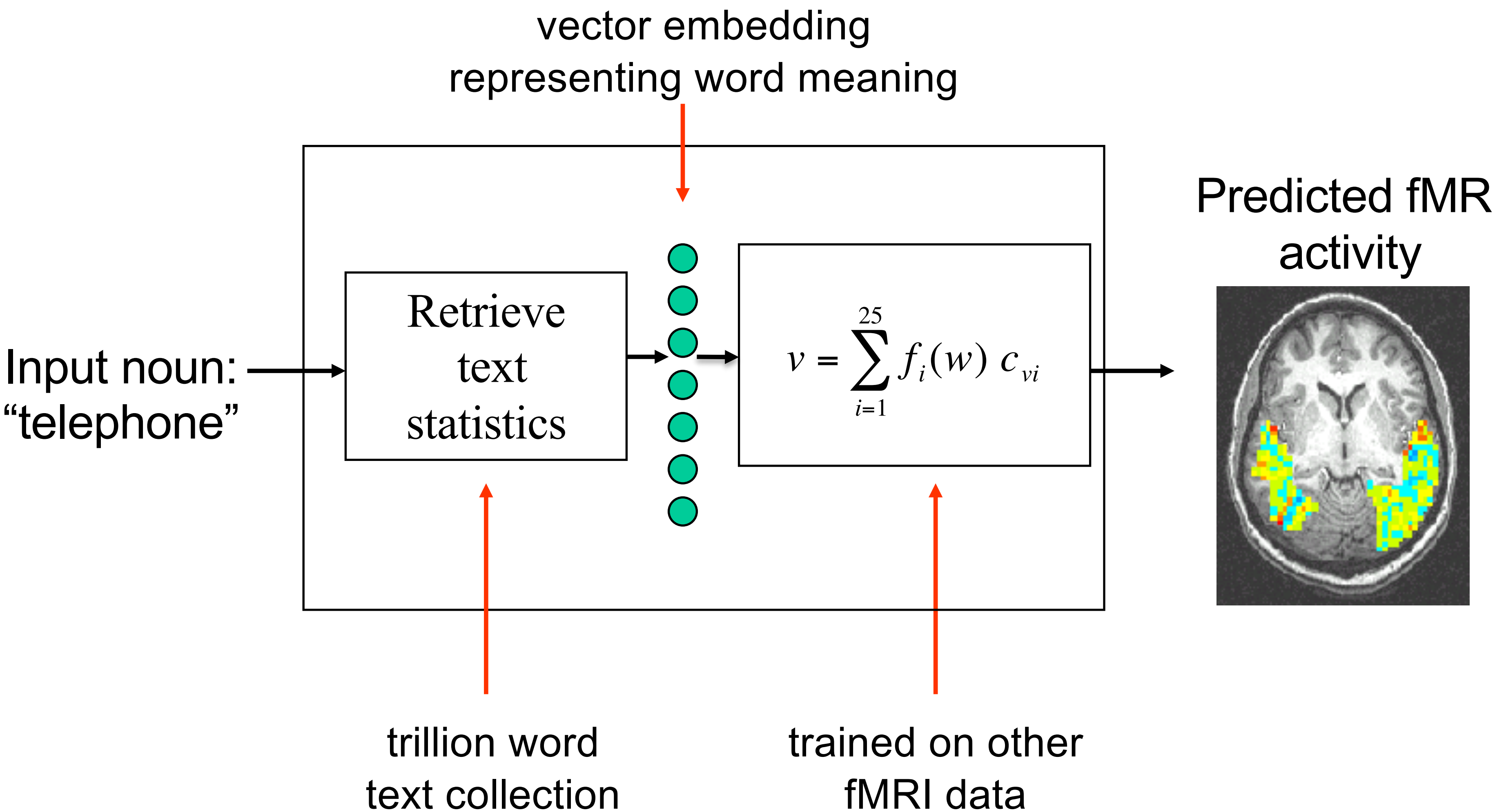
Arbitrary noun

Predicted fMRI activity



Predicting fMRI Activity during Word Reading

[Mitchell et al., *Science*, 2008]



Represent stimulus noun by co-occurrences with 25 verbs*

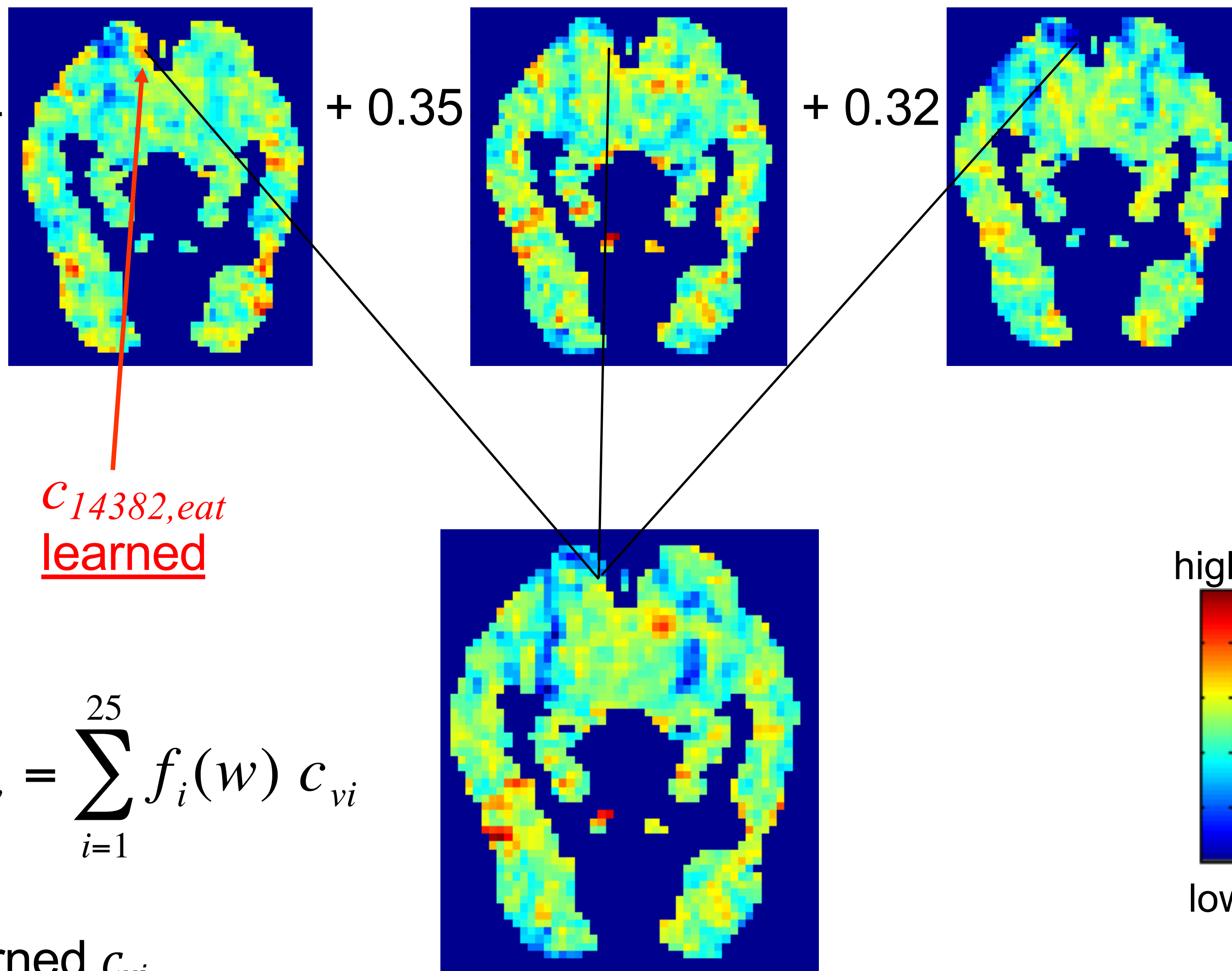
Semantic feature values: "**celery**"	Semantic feature values: "**airplane**"
0.8368, eat	0.8673, ride
0.3461, taste	0.2891, see
0.3153, fill	0.2851, say
0.2430, see	0.1689, near
0.1145, clean	0.1228, open
0.0600, open	0.0883, hear
0.0586, smell	0.0771, run
0.0286, touch	0.0749, lift
…	…
…	…
0.0000, drive	0.0049, smell
0.0000, wear	0.0010, wear
0.0000, lift	0.0000, taste
0.0000, break	0.0000, rub
0.0000, ride	0.0000, manipulate

* in a trillion word text collection

Predicted Activation is Sum of Feature Contributions

"eat" "taste" "fill"

Celery = 0.84 + 0.35 + 0.32 + . . .



Predicted "Celery"

$f_{eat}(celery)$

from corpus statistics

$c_{14382,eat}$

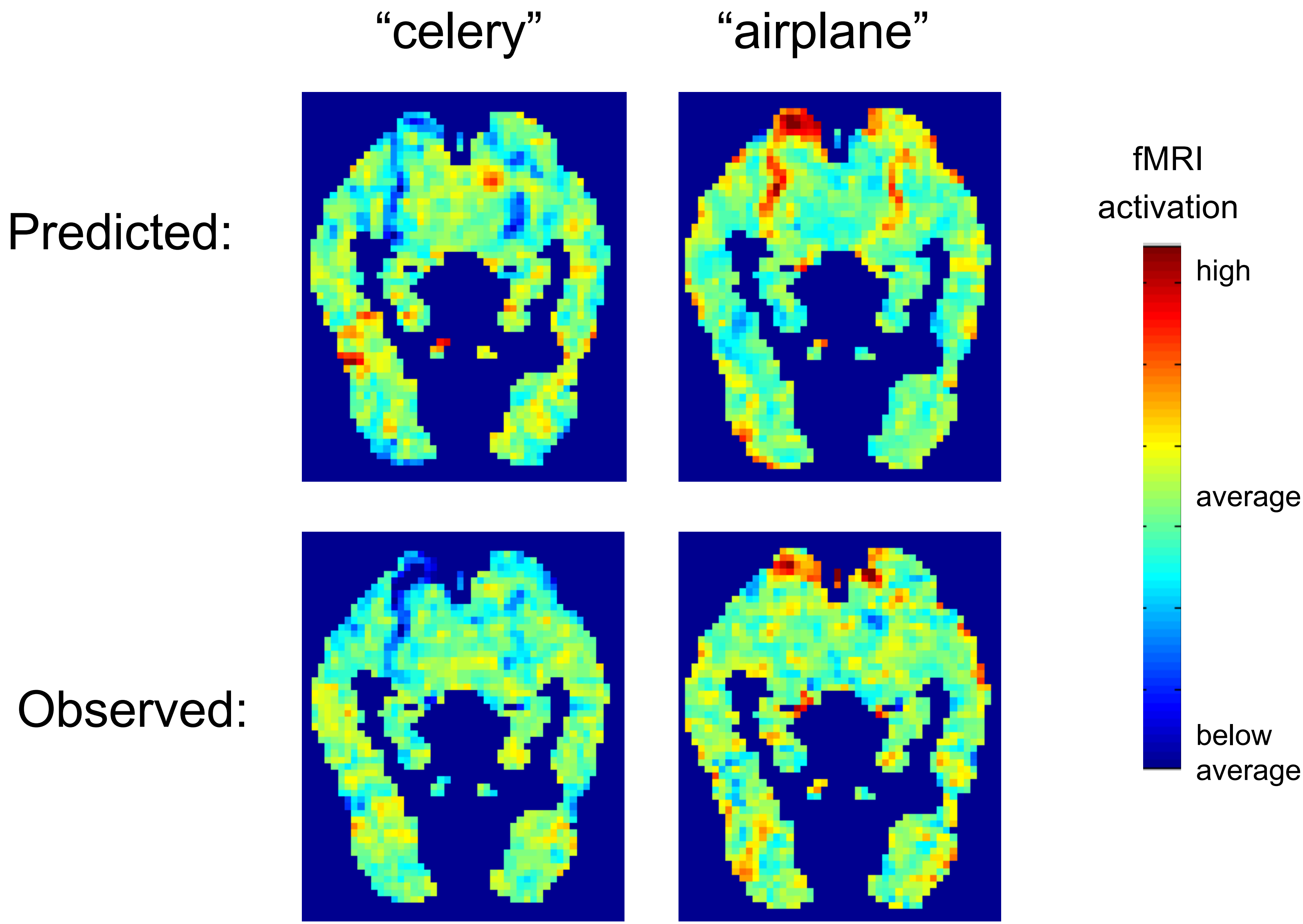
<u>learned</u>

$$prediction_v = \sum_{i=1}^{25} f_i(w)\, c_{vi}$$

500,000 learned c_{vi} parameters

high

low

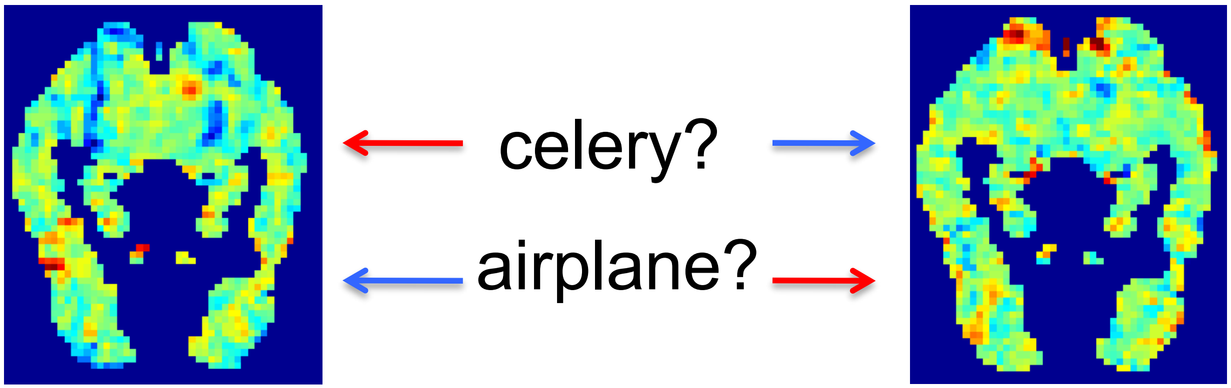


Predicted and observed fMRI images for “celery” and “airplane” after training on other nouns.

[Mitchell et al., *Science*, 2008]

Evaluating the Computational Model

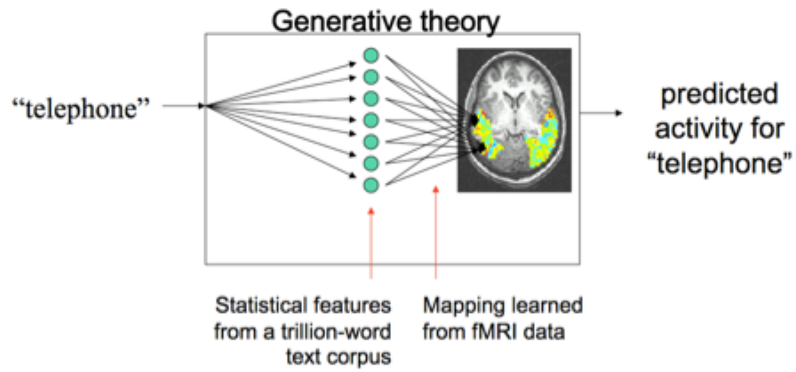
- Leave two words out during training



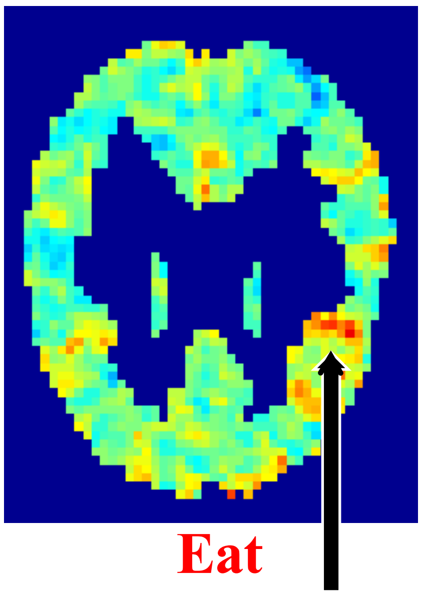
1770 test pairs in leave-2-out:

- Random guessing → 0.50 accuracy
- Accuracy above 0.61 is significant (p<0.05)

Mean accuracy over 9 subjects: 0.79



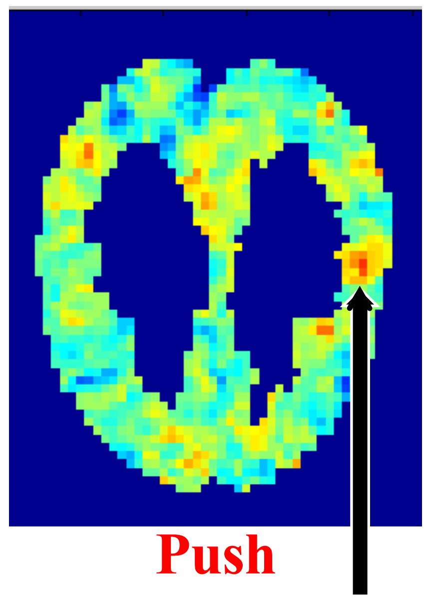
Participant
P1



Eat

"Gustatory cortex"

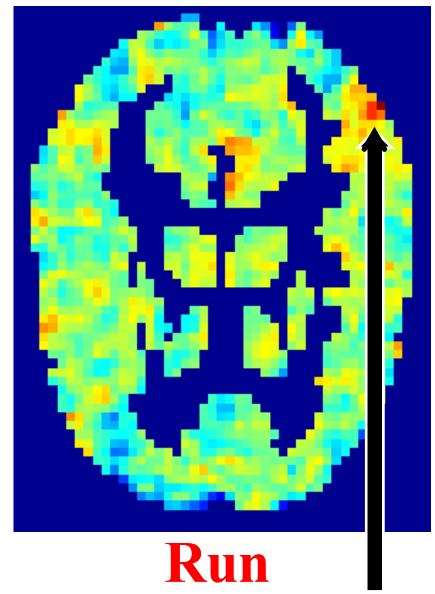
Pars opercularis
(z=24mm)



Push

"somato-sensory"

Postcentral gyrus
(z=30mm)



Run

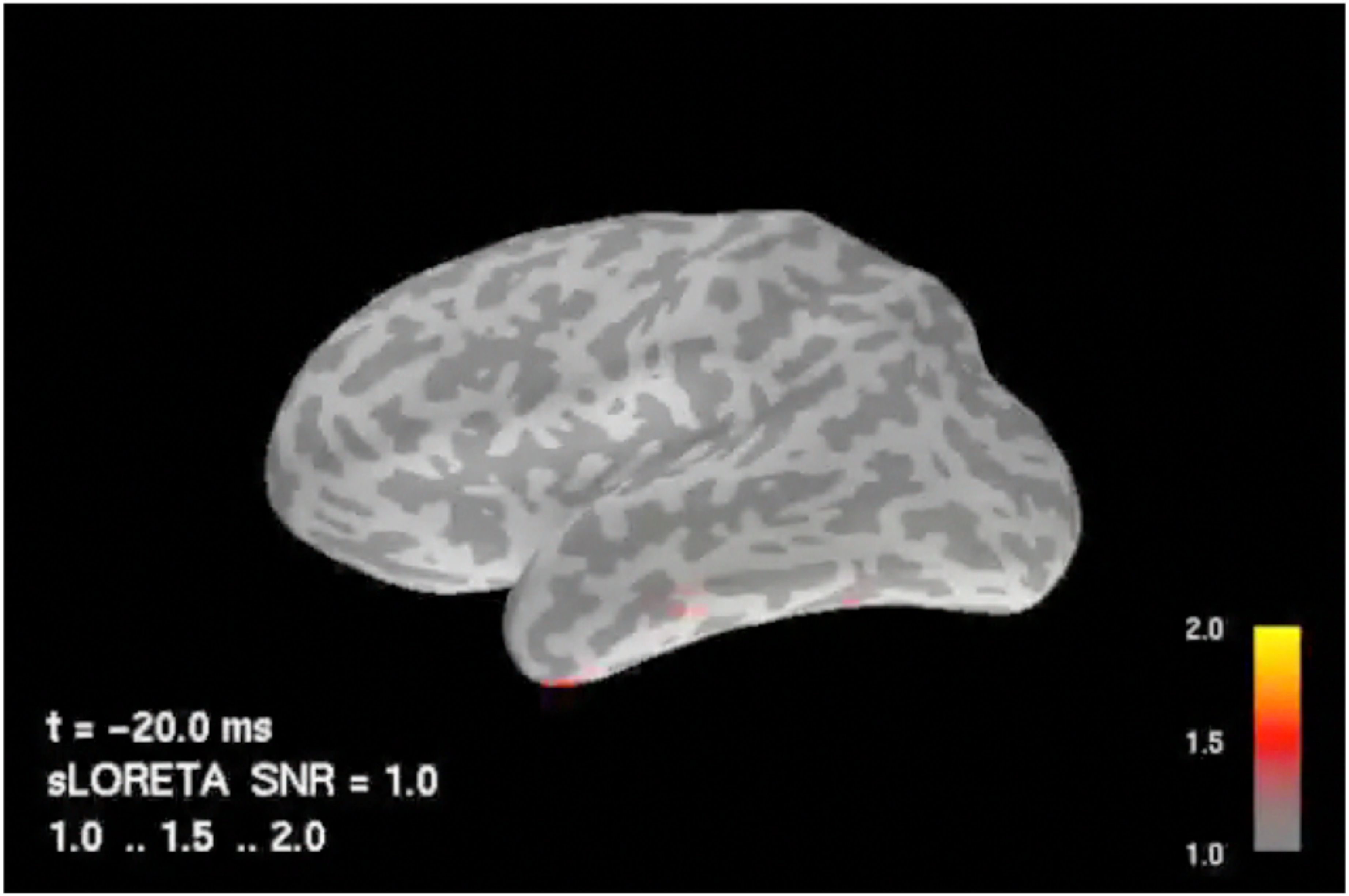
"Biological motion"

Superior temporal
sulcus (posterior)
(z=12mm)

When

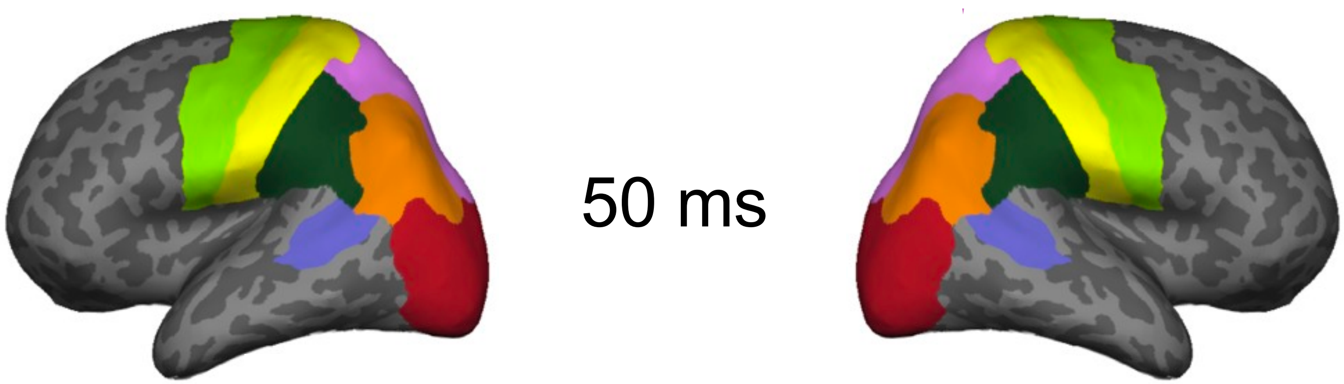
does neural activity encode specific information?

MEG: Stimulus “hand” (word plus line drawing)

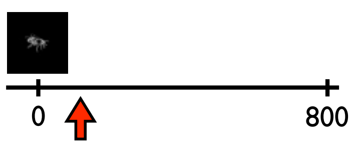


[Sudre et al., *NeuroImage* 2012]

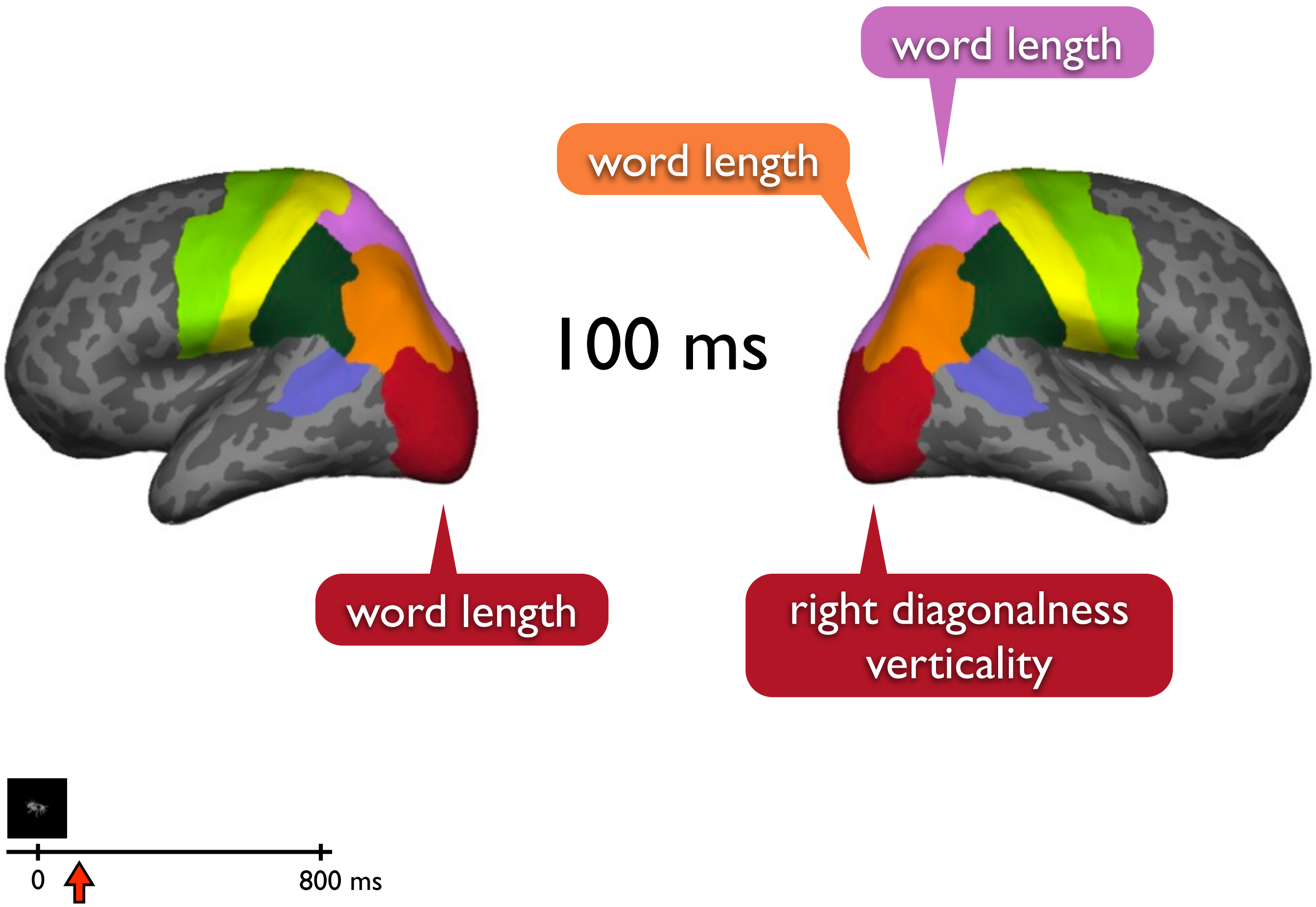
Gustavo Sudre



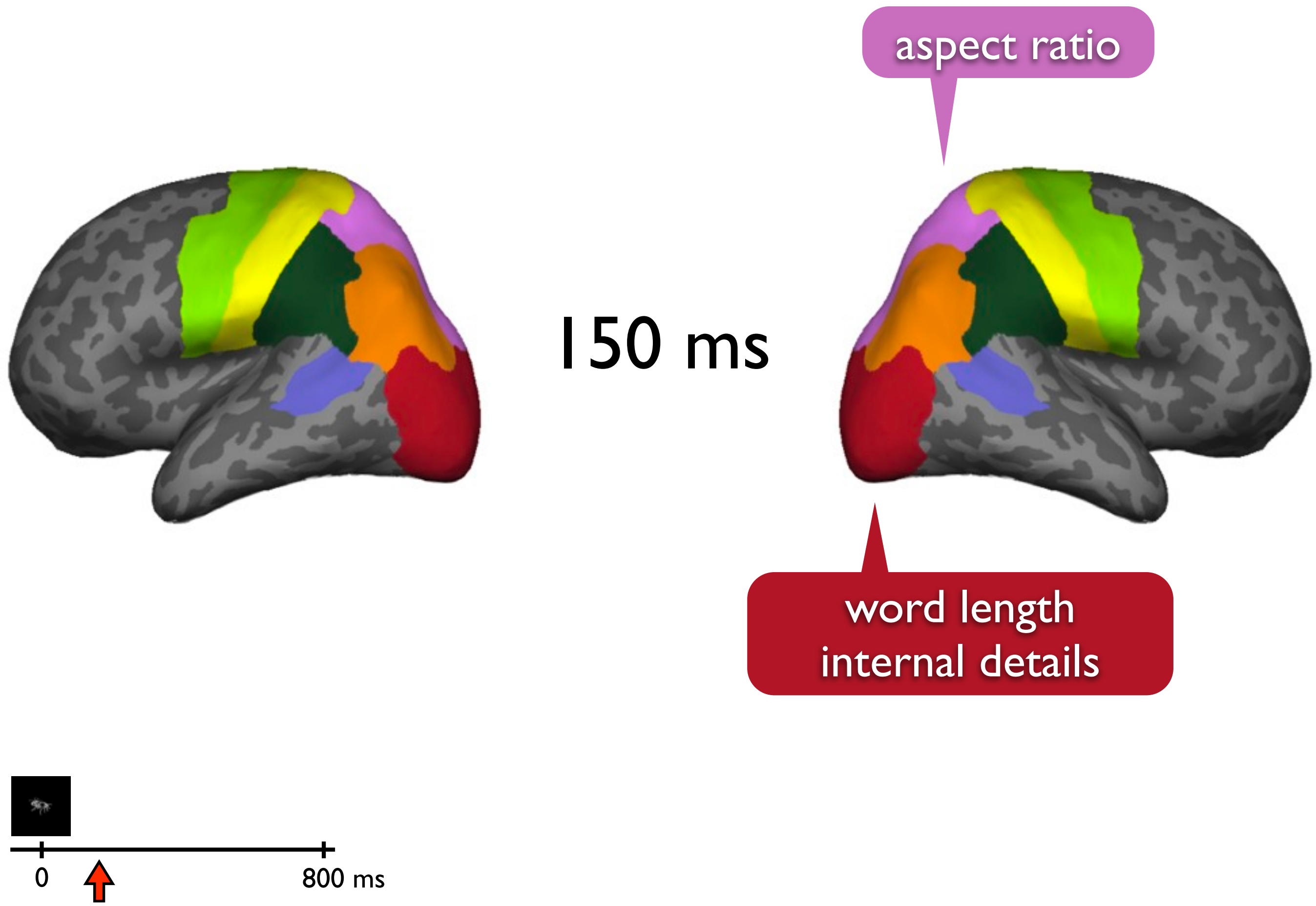
50 ms



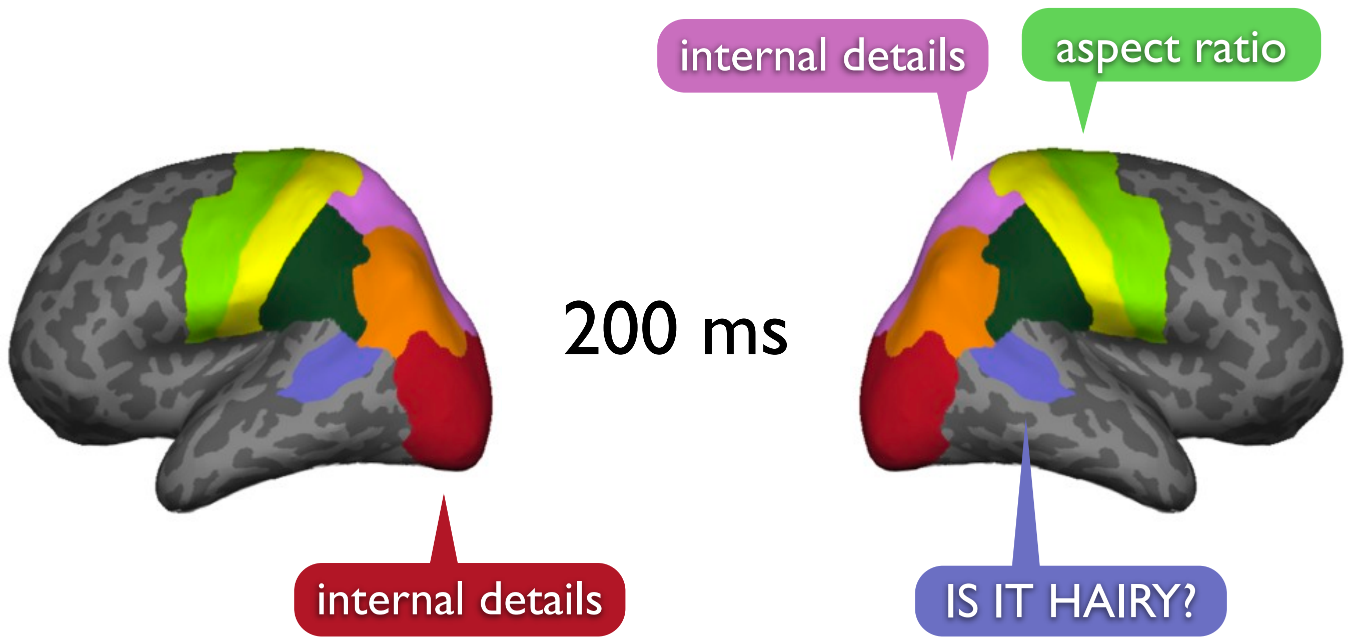
[Sudre et al., *NeuroImage* 2012]

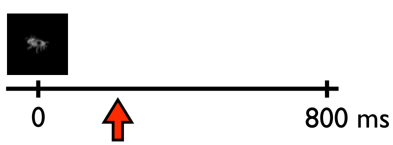


[Sudre et al., 2012]

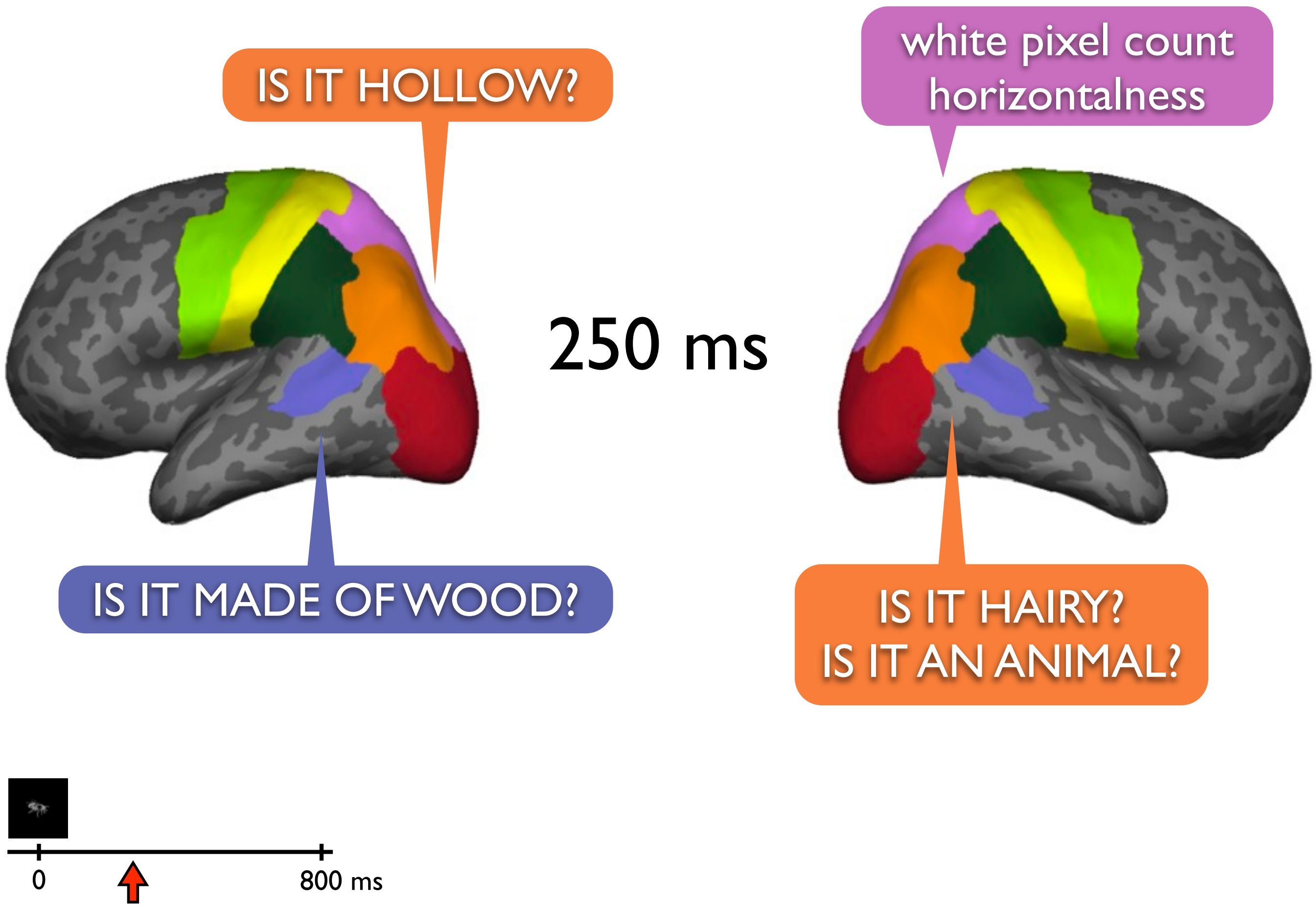


[Sudre et al., 2012]

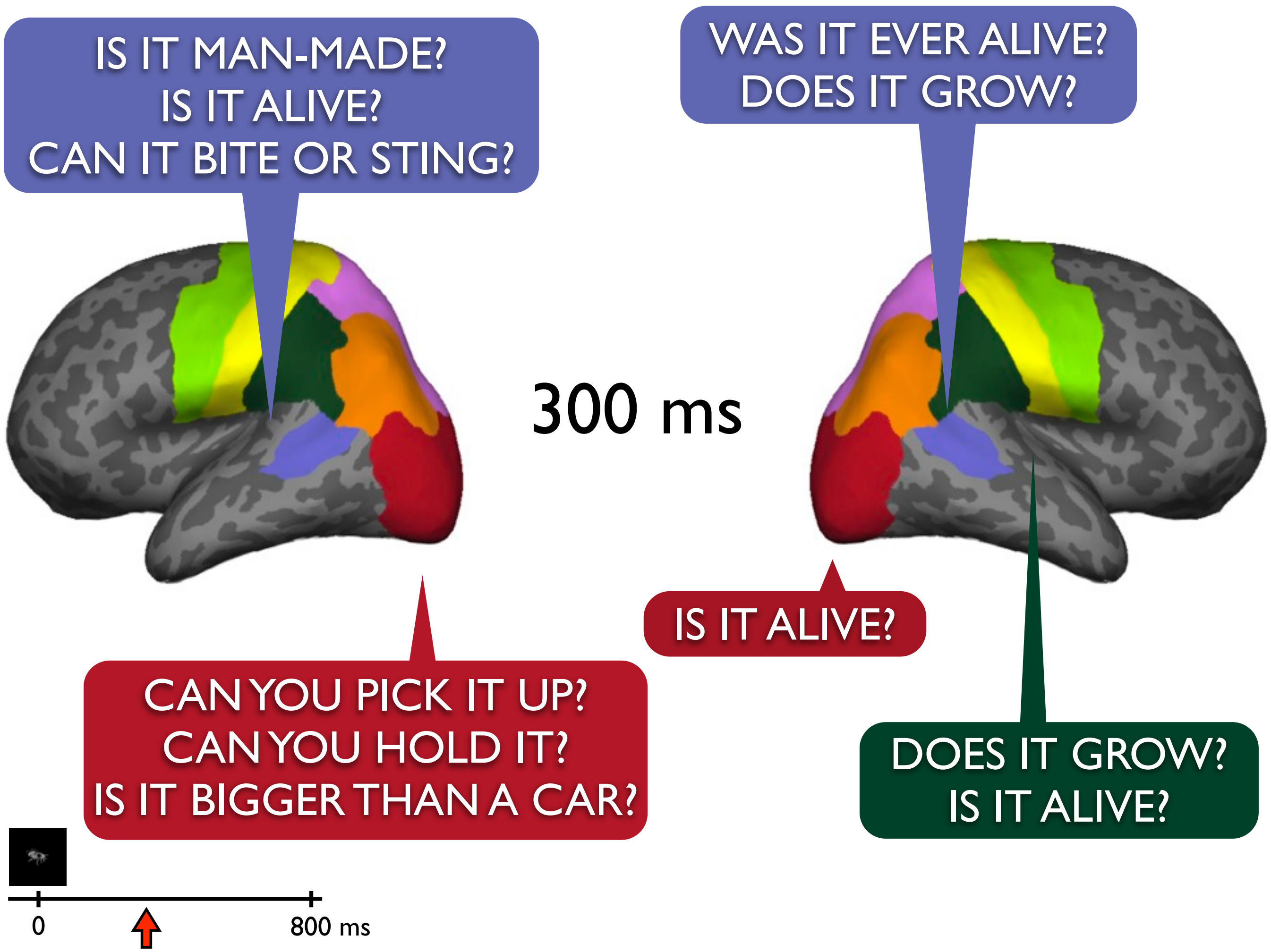




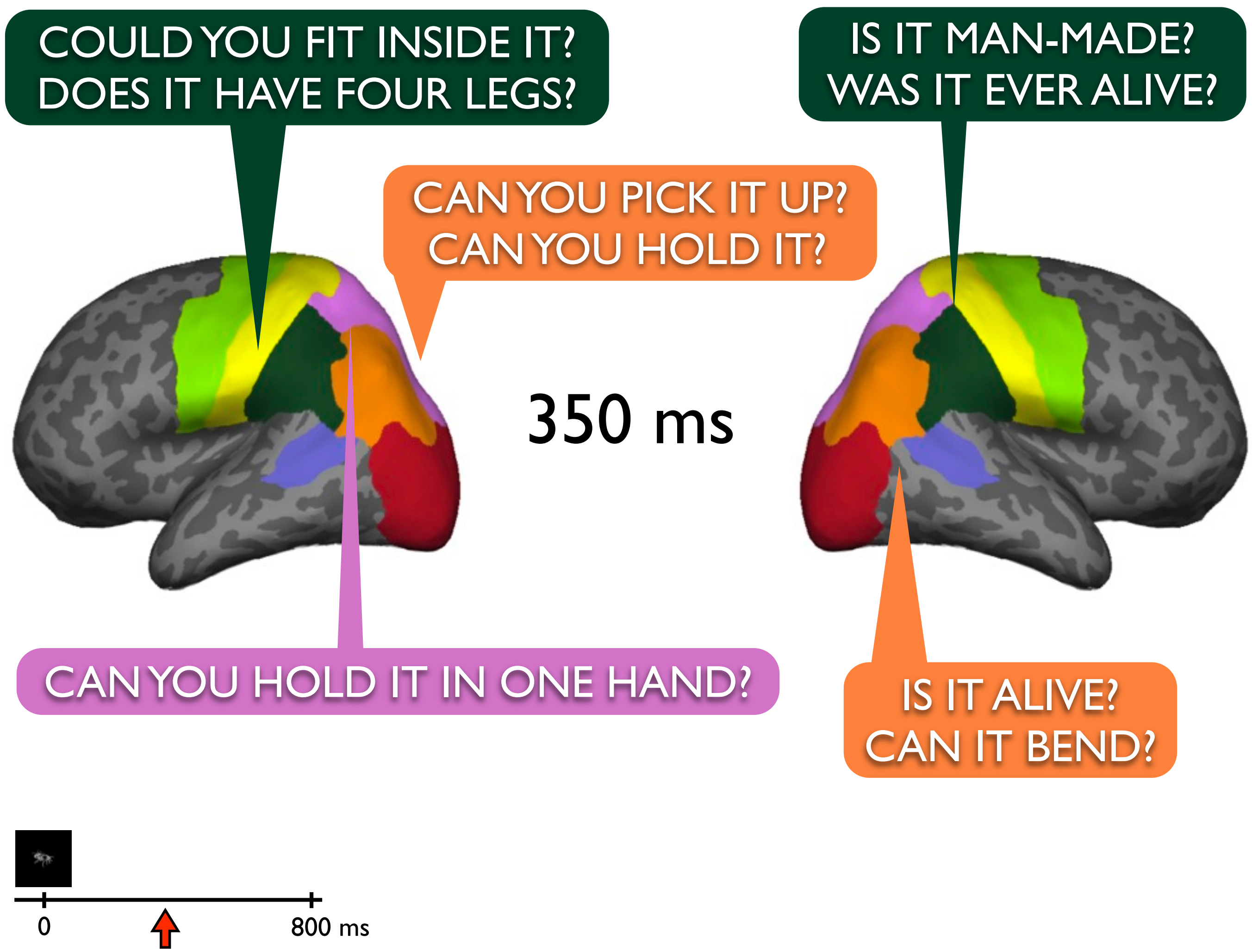
[Sudre et al., 2012]



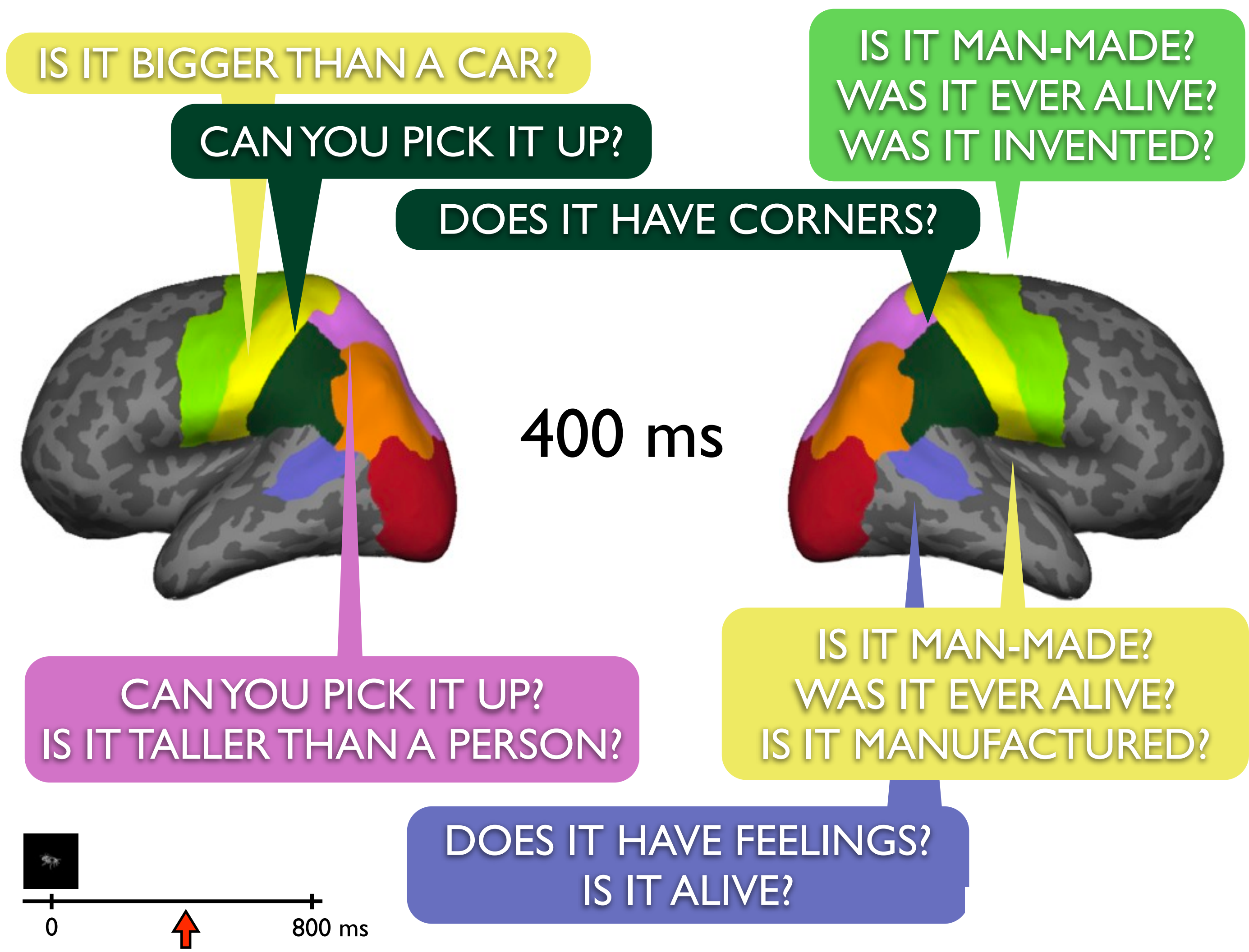
[Sudre et al., 2012]



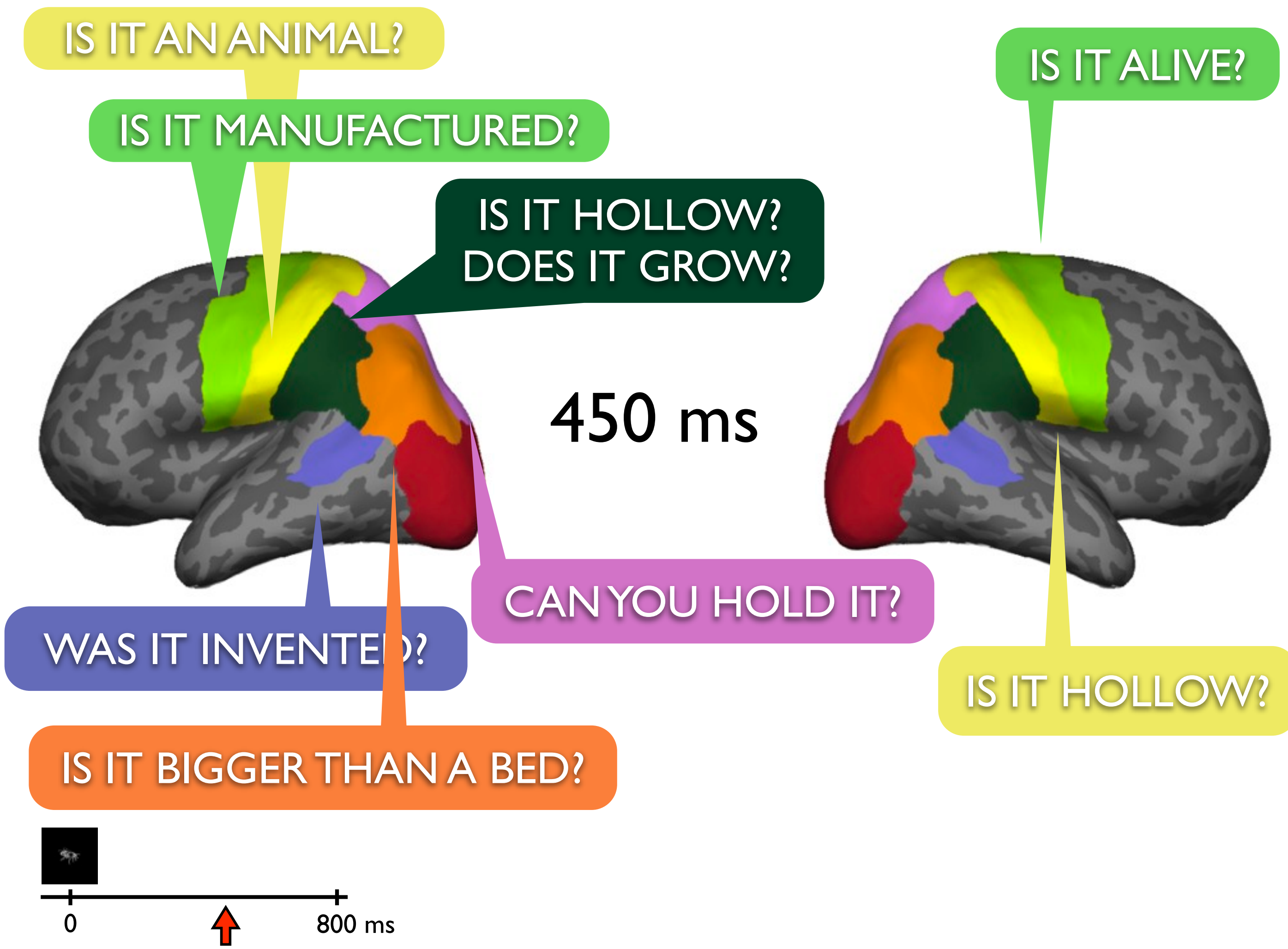
[Sudre et al., 2012]



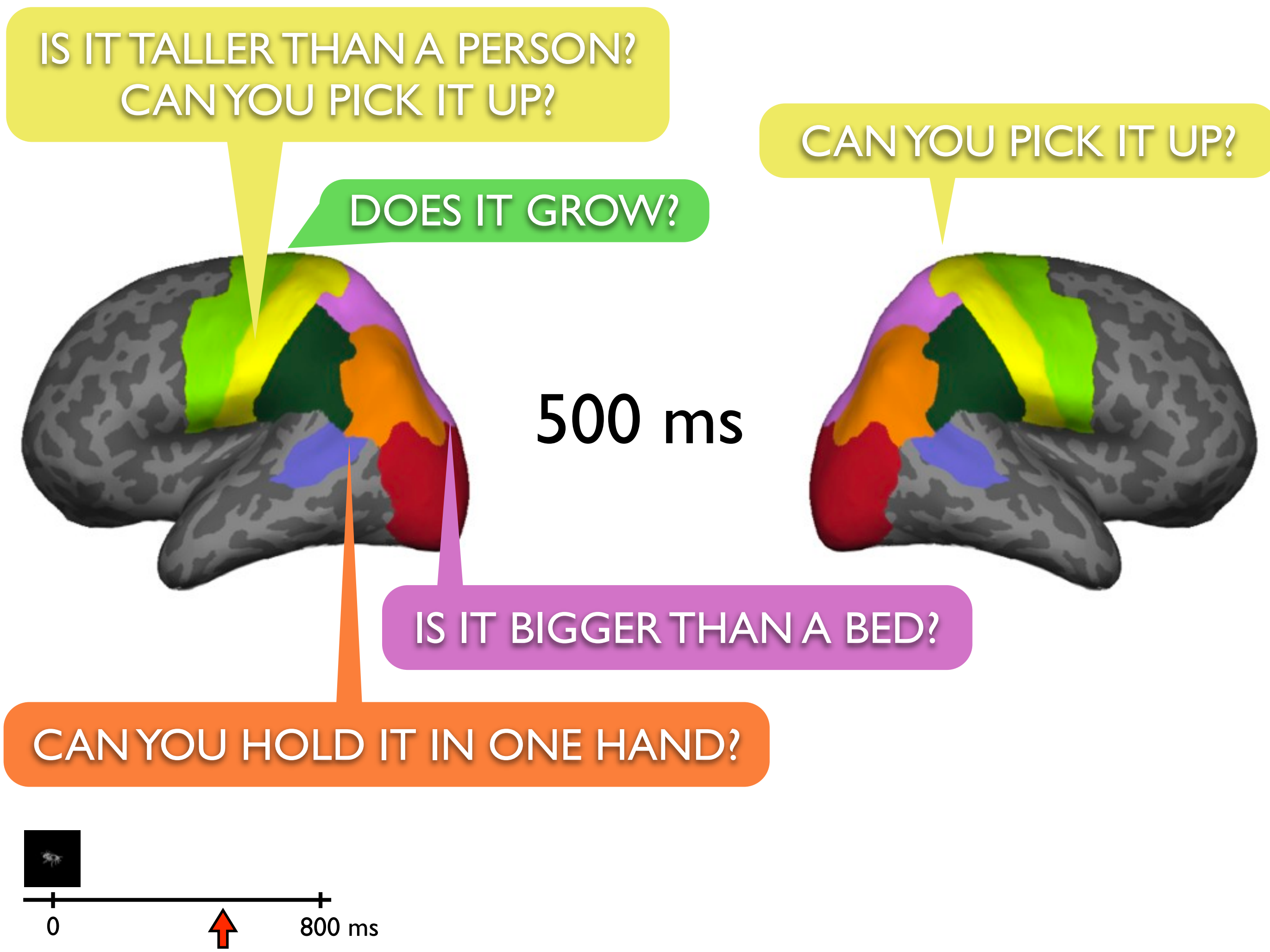
[Sudre et al., 2012]



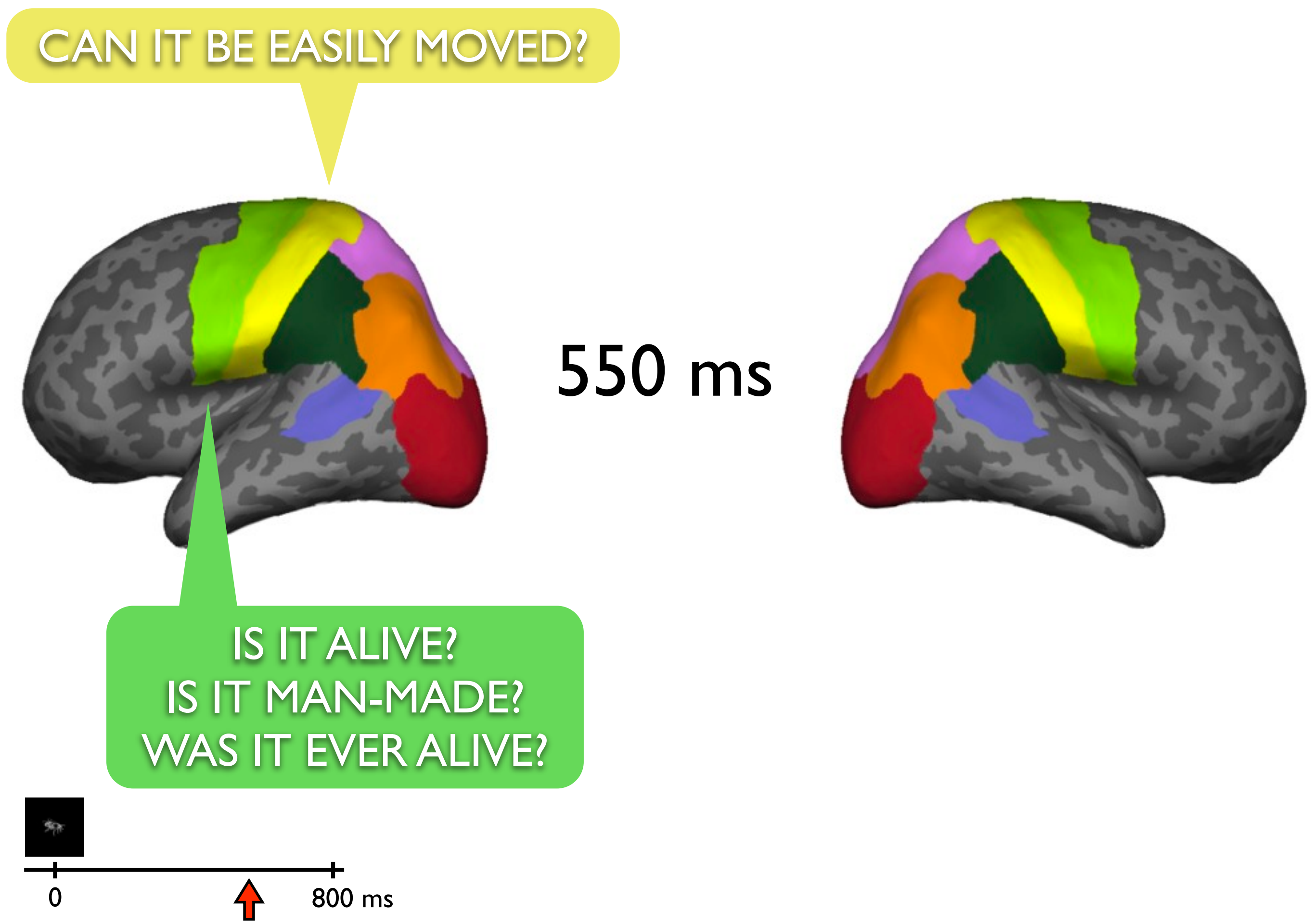
[Sudre et al., 2012]



[Sudre et al., 2012]



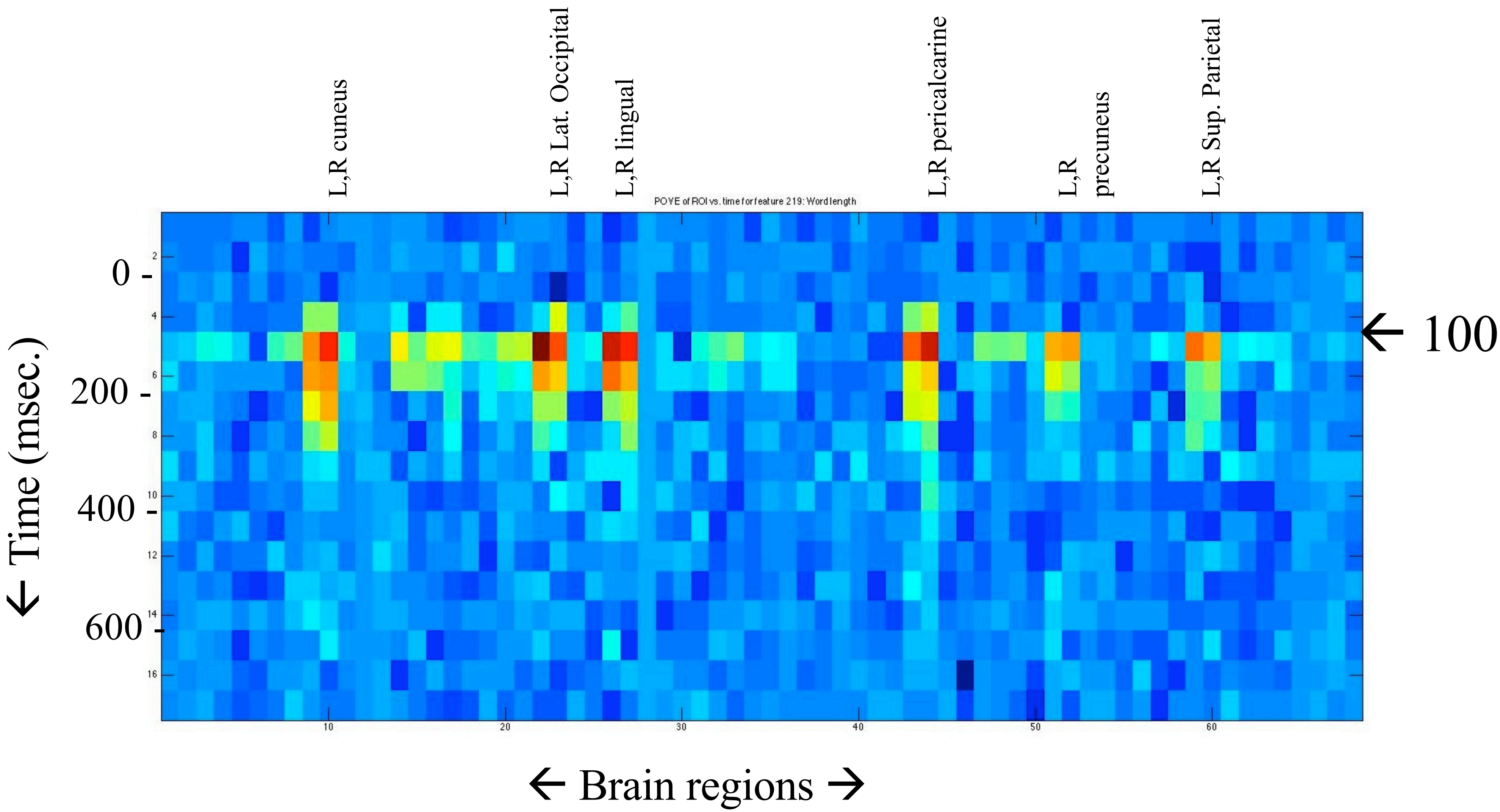
[Sudre et al., 2012]



[Sudre et al., 2012]

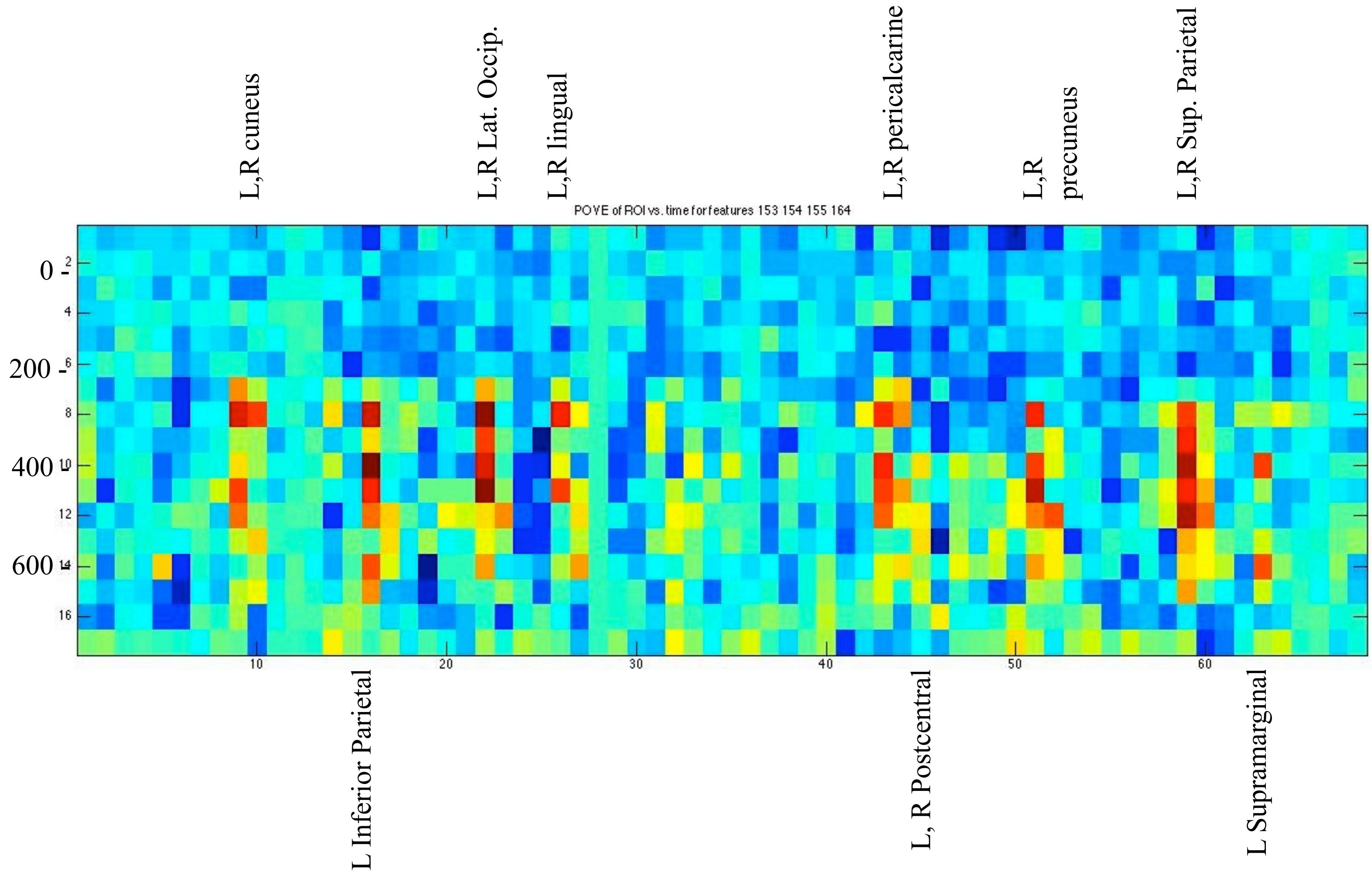
When and Where: Details

Color= decodability* of feature “wordlength” (peak decodability 100-150 msec)



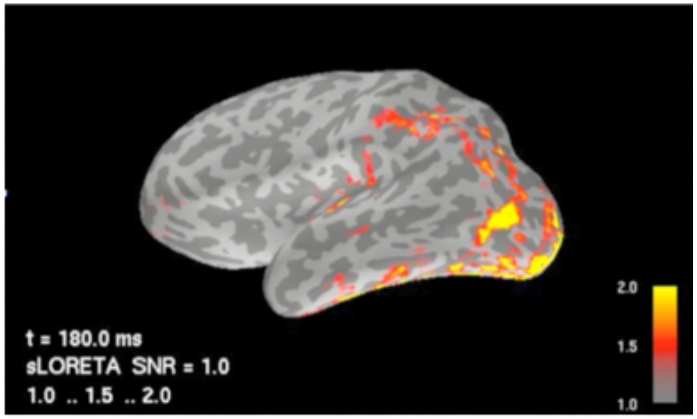
* % of feature variance predicted by MEG, mean across 9 subjects

Color= decodability of “grasping“ features (initial peak: 200-300 msec)



[Sudre et al., 2012]

Results: Timing



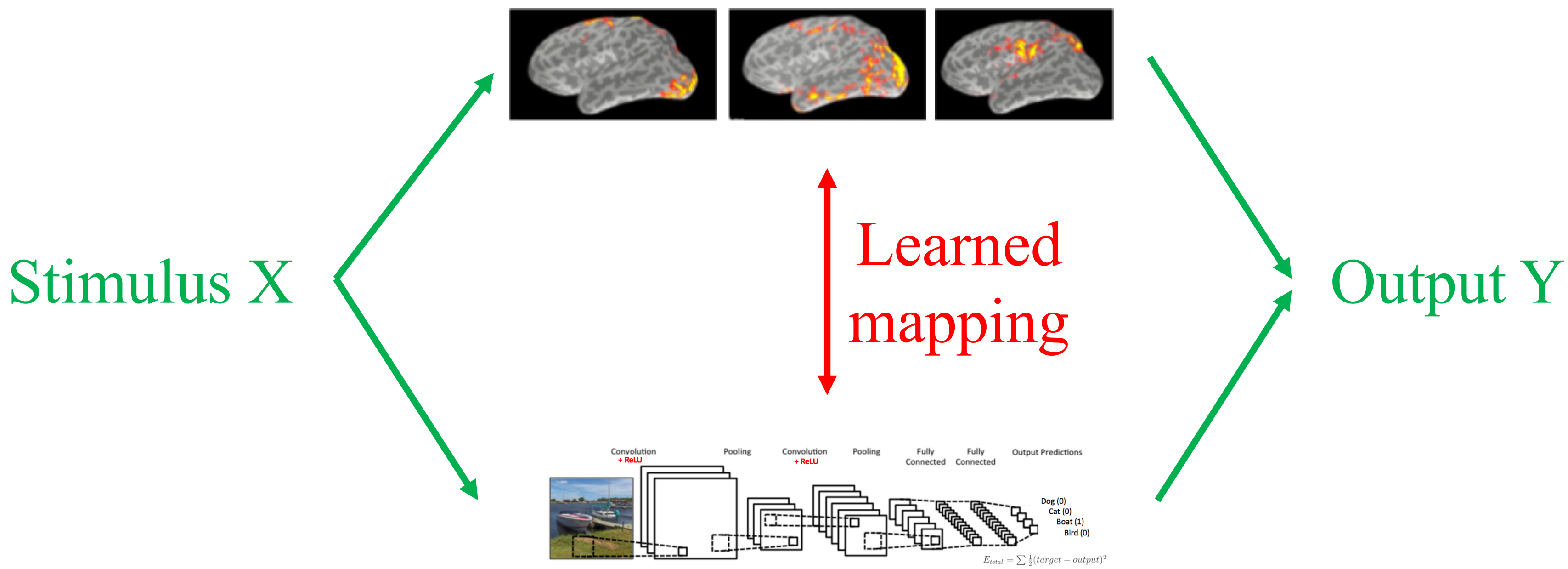
- Neural encodings of word meaning are most complete at 400 msec post onset

- But semantic features do not all appear at once, they trickle in over time, and endure through 400-500 msec

How

does the brain compute neural representations?

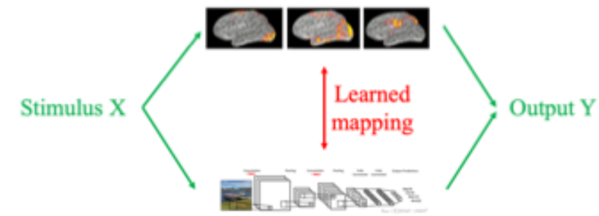
A Paradigm for Studying "How"

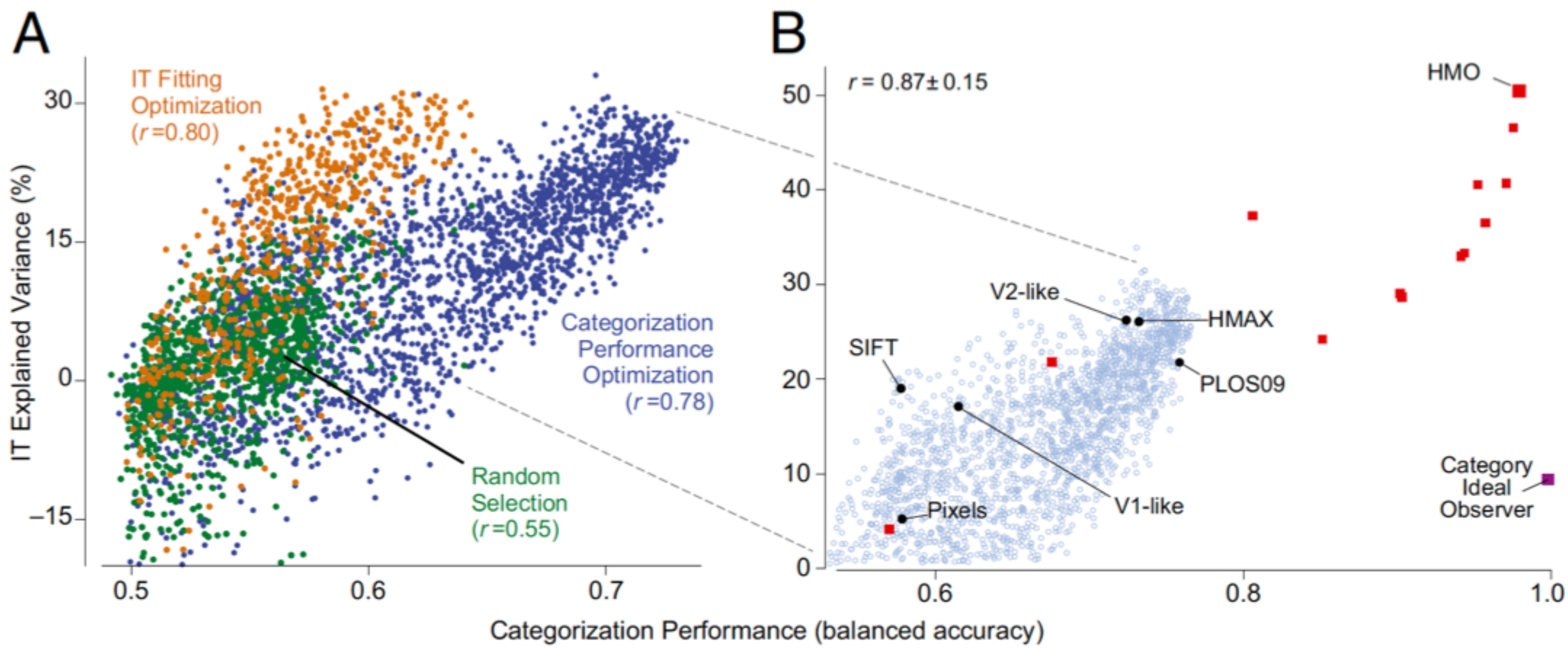


1. Create computer program f(x)=y as hypothesis of brain processing
2. Give same stimuli to computer and brain
3. Train mapping between sequence of brain activity, and intermediate states of computer program
4. Test ability to predict observed neural activity from these intermediate states

How?: Visual Processing

Network Accuracy correlates with IT Predictability



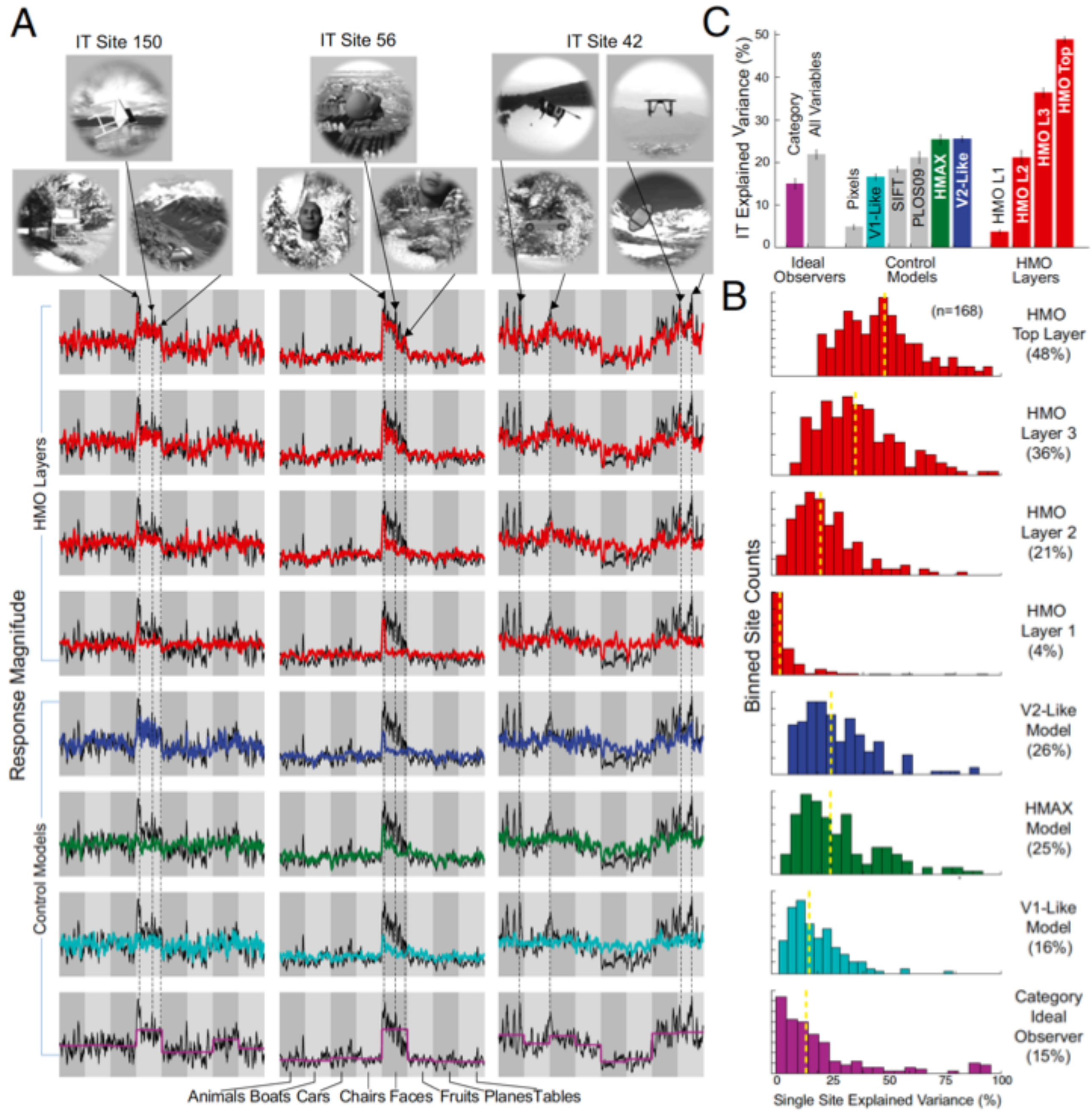


[Yamins et al., 2014]

CNN-IT Alignment

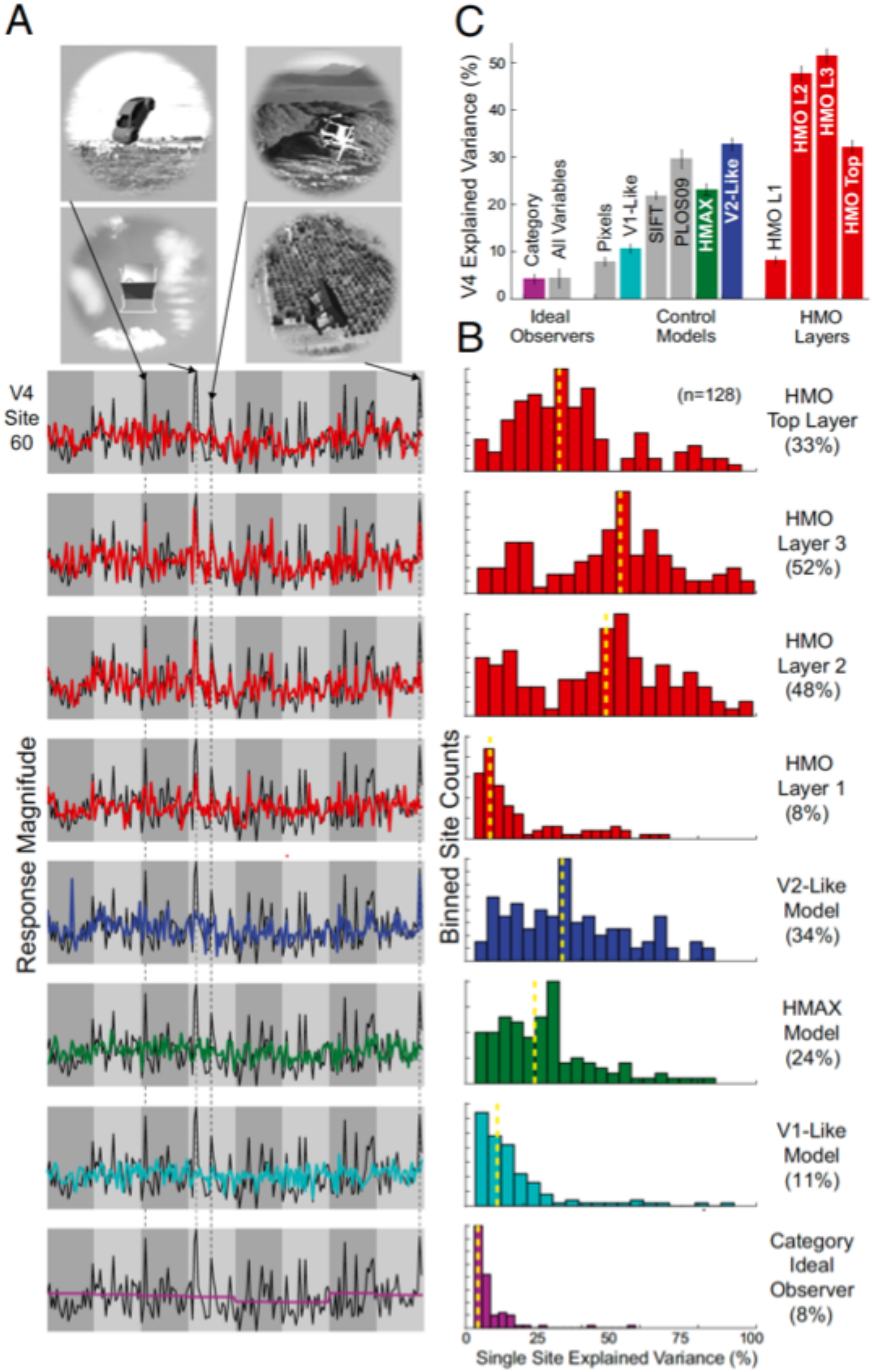
3 IT sites

[Yamins et al., 2014]



Computational Model

CNN-V4 Alignment



Computational Model

[Yamins et al., 2014]

Summary

- Object recognition accuracy of deep net correlates with ability to predict IT neural activity
- Output layer best predicts IT activity
- Penultimate layer best predicts V4 activity

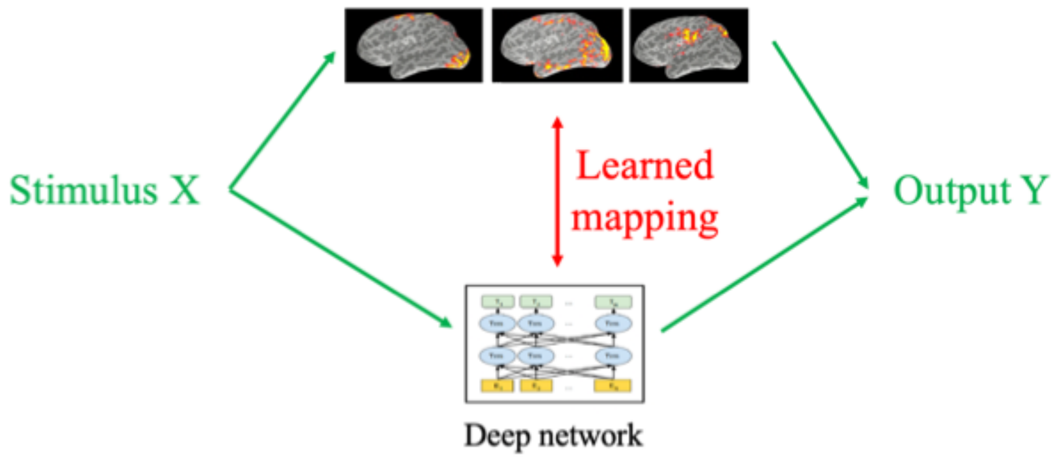
How?: Language Processing

How?: Language Processing

Sentence reading

[Jat, Hao, Talukdar, Mitchell. *ACL 2019*]

Question: How does brain process sentences?



Hypothesis 1: BERT
2: ELMo
3: Bi-Directional LSTM
4: sum of GloVe embeddings

Learned mapping Deep net to Brain activity: Ridge regression

A

student

found

the

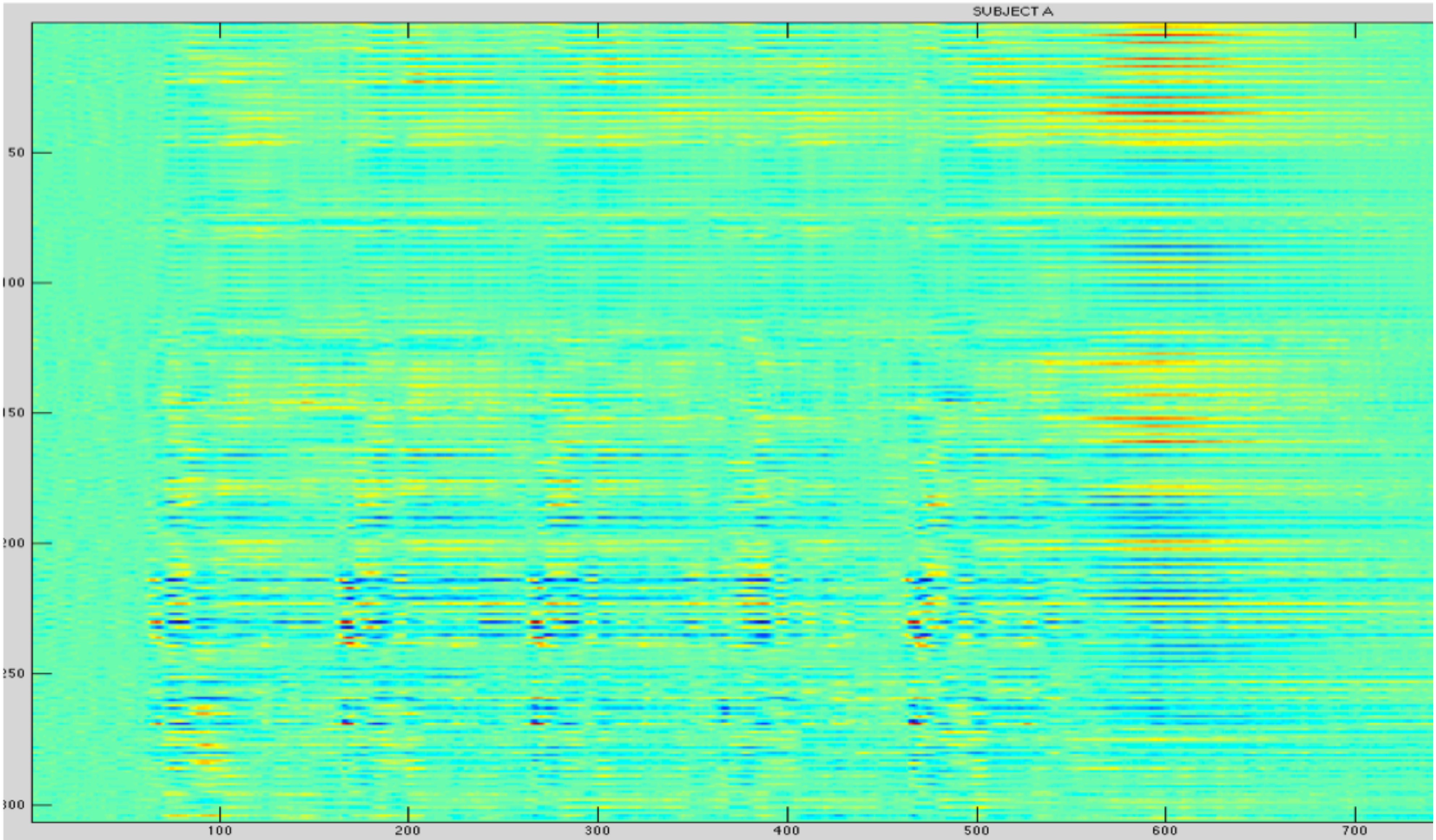
hammer.

Sentence mean MEG activity:

time →

| A | student | found | the | hammer | + |

brain locations / MEG sensors



| A | student | found | the | hammer | + |

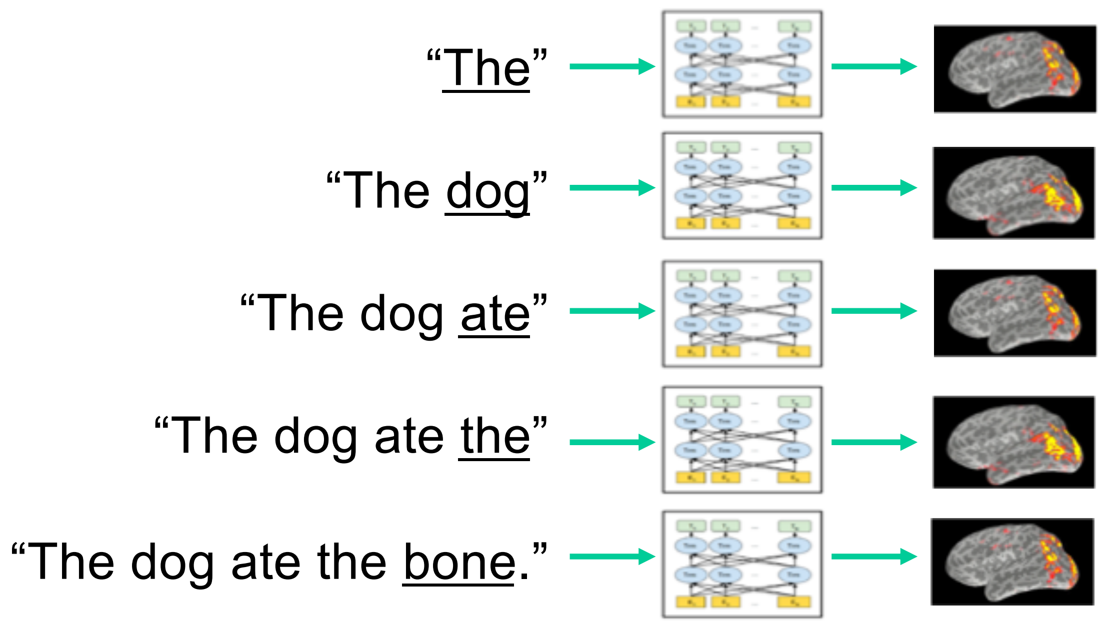
Data

- Collected MEG brain activity during simple sentence reading
  - Passive: “The dog ate the bone.”
  - Active: “The bone was eaten by the dog.”

Dataset	#Sentences	Voice	Repetition
PassAct1	32	P+A	10
PassAct2	32	P+A	10
PassAct3	120	A	10

Modeling sequential reading

- Give each prefix of sentence as separate input to deep network, to predict 500 msec of brain activity for each word position

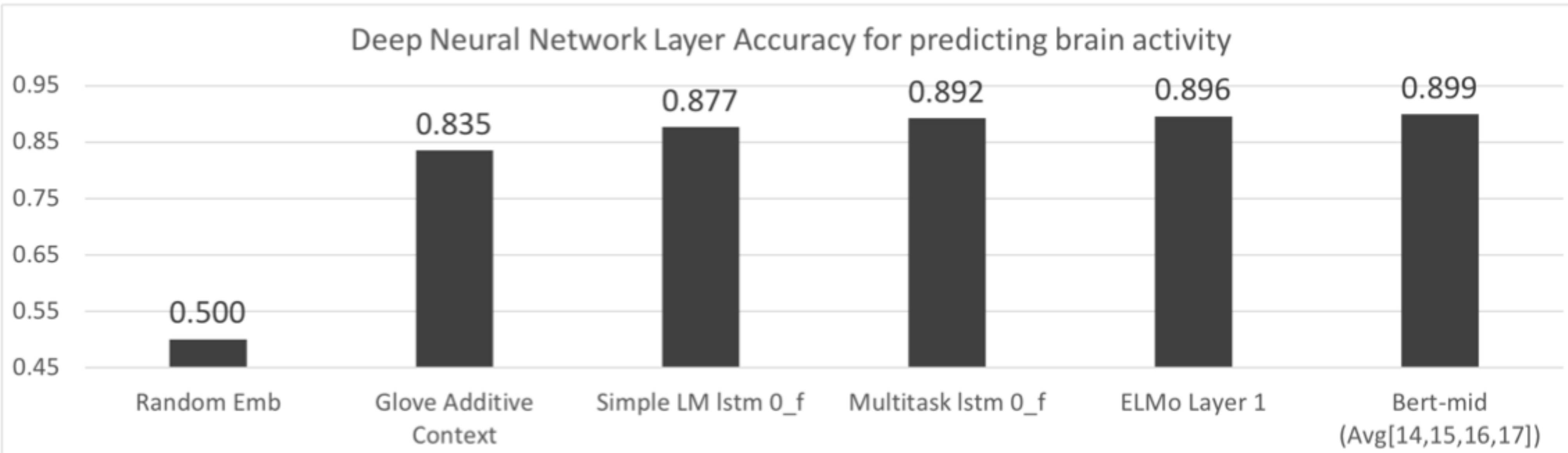


Question: How does brain process sentence?

Hypothesis 1: BERT
2: ELMo
3: LSTM
4: sum of GloVe embeddings

Mapping Deep net to Brain activity: Ridge regression

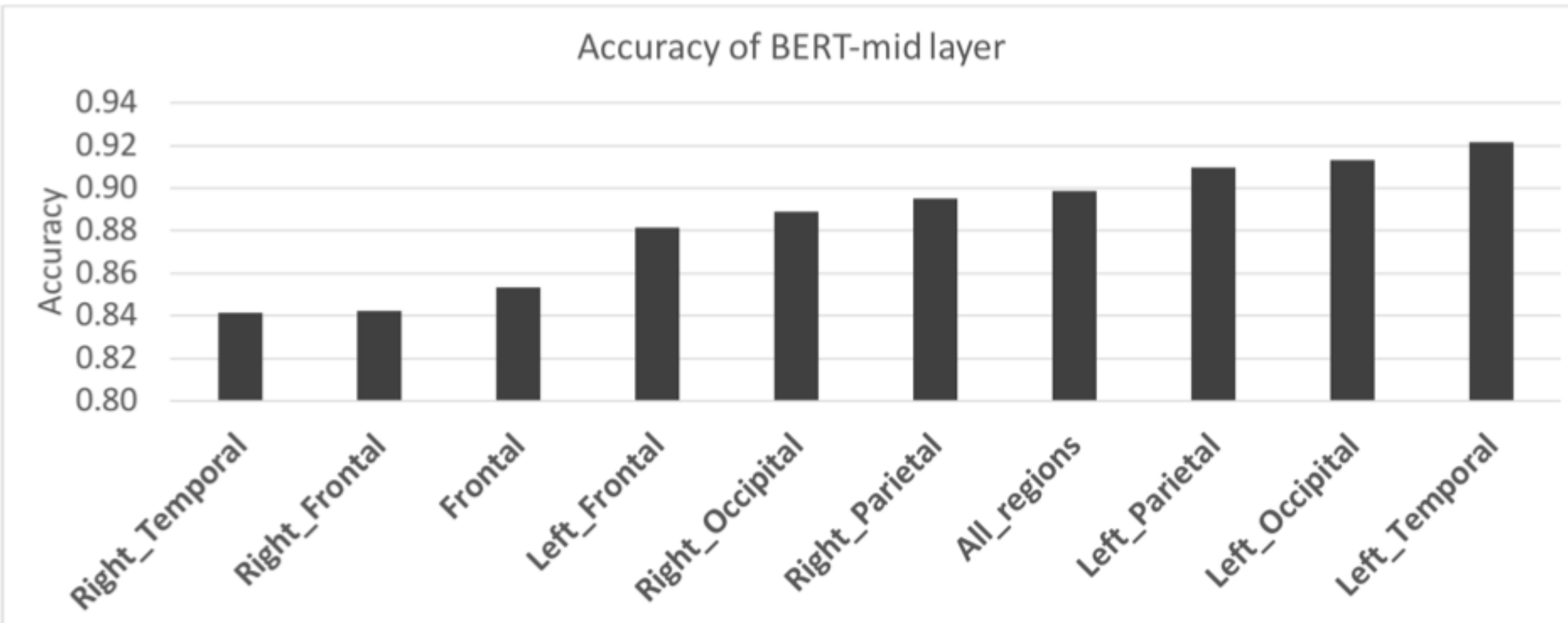
Brain activity prediction accuracy*



BERT-mid is most effective at predicting Brain activations

* 2x2 classification accuracy

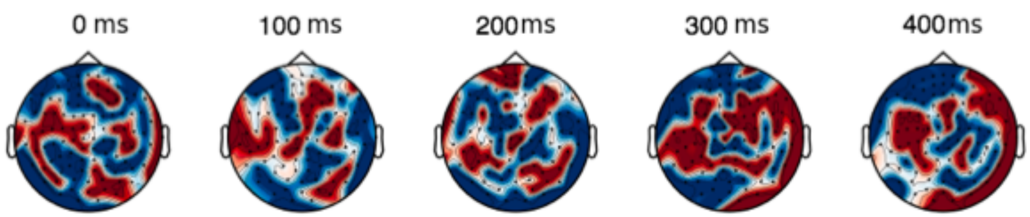
Brain activity prediction



Left temporal brain region is predicted with highest accuracy

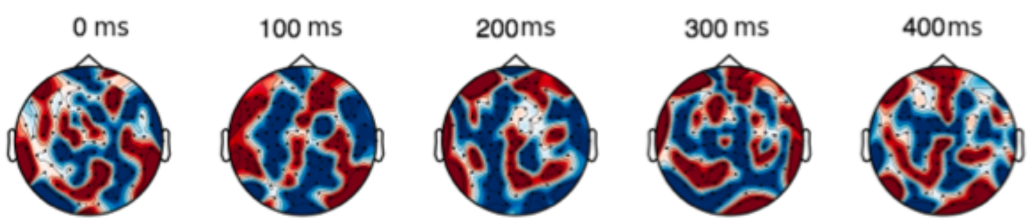
Results when text differs by one earlier word:

- E.g., Vary noun at t-2, classify it based on time t model-brain alignment:
  - "The <u>dog</u> ate the"
  - "The <u>girl</u> ate the"
  - Accuracy: 0.92



Red and blue show areas of correctly predicted positive and negative activity

- Vary verb at t-1, predict it at time t:
  - "The dog <u>saw</u> the"
  - "The dog <u>ate</u> the"
  - Accuracy: 0.92



Experiment: predict earlier noun, earlier verb

NOUN	“the dog ate the” vs “the girl ate the”	Most DNN layers retain Noun info	ELMO_{mid}(0.92)
VERB	“the dog ate the” vs “the dog saw the”	Most DNN layers retain Verb info	ELMO_{mid}(0.92)

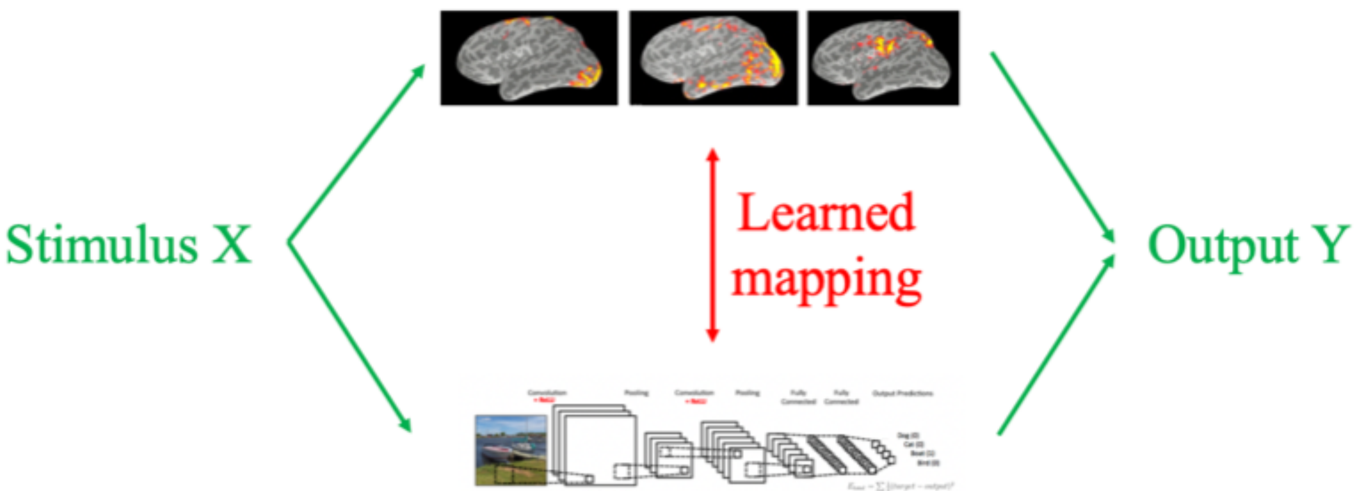
Experiment: predict earlier adj, earlier determiner

ADJECTIVE	"the happy child" vs "the child"	Middle DNN layers retain Adj info	Multitask LSTM layer1 (0.89)
FIRST DETERMINER	"the dog" vs "a dog"	Shallow DNN layers retain info better	BERT layer 3 (0.82) BERT layer 18 (0.78)

Summary

- Different deep nets have different abilities to predict neural activity
- BERT mid-layers predict most accurately overall
- Predicts word-by-word time neural activity
- Left temporal lobe (language related) is best predicted
- Deep net models predict influence of earlier words on later brain activity

Will this research paradigm really work?



Will this paradigm really work?

Supporting evidence: existing demonstrations

- Vision: CNN’s modeling aspects of visual cortex
- Language: State of art deep nets align with neural activity

DeepNets give us a number of relevant architectures

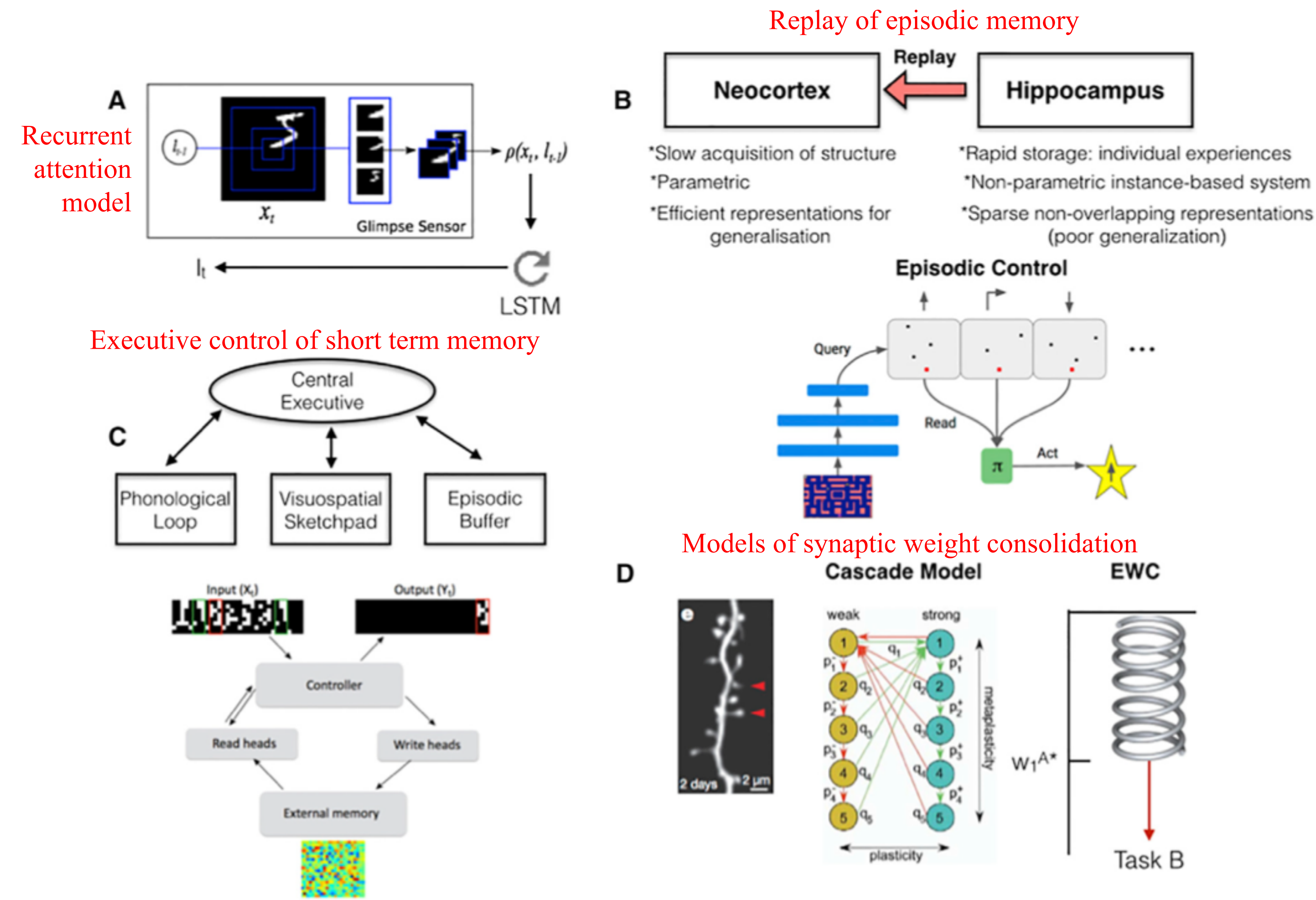


Figure 1. Parallels between AI Systems and Neural Models of Behavior

[Hassabis, et al, 2017]

Will this paradigm really work?

Limits:

- Mismatch of sequential computer processing vs. oscillatory, parallel neural activity
- Mismatch of constant activity in deep net units vs. spiking biological neurons
- Mismatch of brain image signal (e.g., blood oxygen fluctuations, magnetic fields) and actual neural activity

Will this paradigm really work?

Important questions:

- Does observed neural activity represent neural data representations, or *processes that alter* neural representations? (e.g., predictive coding: activity reflects prediction errors)
- Are brains truly performing the same task as the computer? What task *is* the brain performing?
- How do context and current physiological state of person influence neural activity? Can programs model these?

Will this paradigm really work?

Questions (continued):

- Should we care if we model only part of what the brain is doing? (e.g., BERT doesn't model word *perception*)

- If we can't interpret representations in deep nets, does it help to explain brain activity in terms of these?

- Given limited resolution and coverage of imaging methods, which computations predicted by computational models are observable in neural activity?

Will this paradigm really work?

My (current) humble opinion

- Like other paradigms, it is imperfect but helpful
- It allows writing candidate theories about How,
  - Whose different predictions can be easily identified
  - and directly tested
- Has similar upsides/downsides to cognitive modeling that attempts to fit observed behavioral data (response times, error rates)
  - And computational models should be evaluated on *all* this data
- Paradigm will grow in importance over coming decade

thank you!