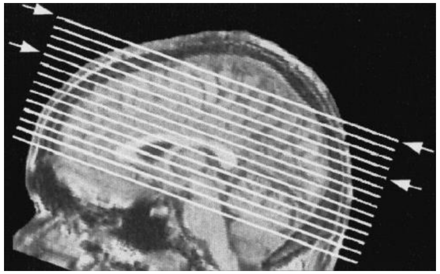
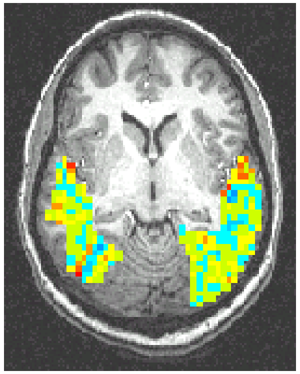
Using Machine Learning to Study the Neural Representations of Language Meanings

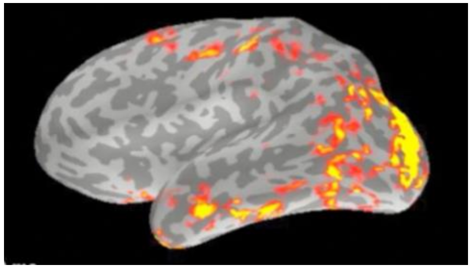


Tom M. Mitchell

Carnegie Mellon University

June 2017





How does neural activity encode word meanings?

How does neural activity encode word meanings?

How does brain combine word meanings into sentence meanings?

Neurosemantics Research Team

Tom Mitchell

Marcel Just

Research Scientists

Erika Laing

Research Scientists

Kai-Min Chang

Dan Howarth

Recent/Current PhD Students

Leila Wehbe

Dan Schwartz

Alona Fyshe

Mariya Toneva

Mark Palatucci

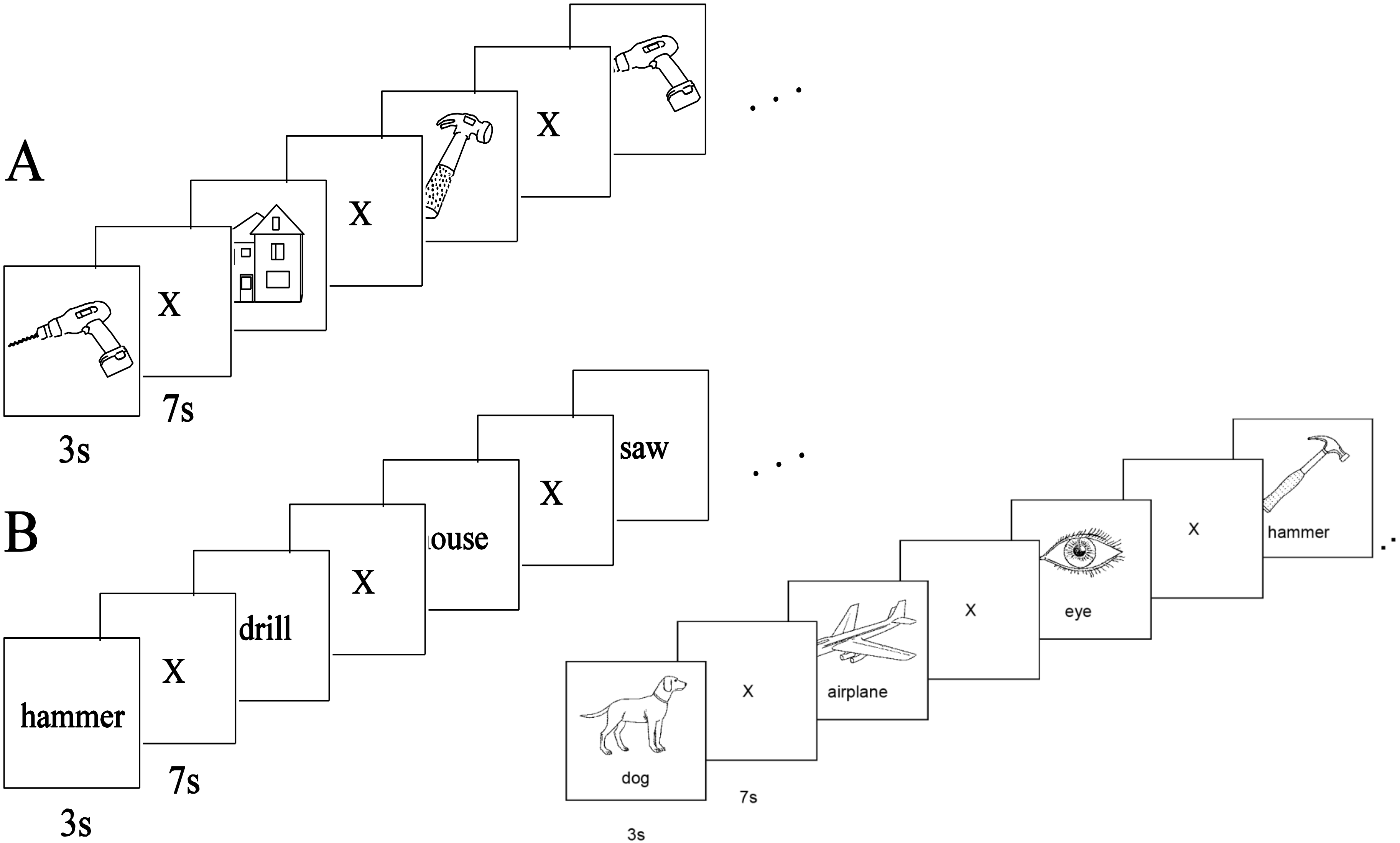
Gustavo Sudre

Nicole Rafidi

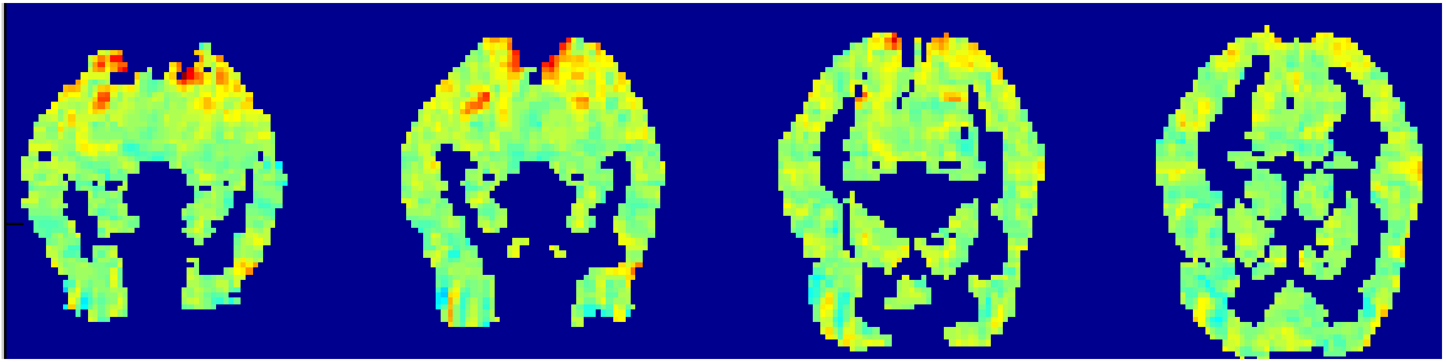
funding: NSF, NIH, IARPA, Keck

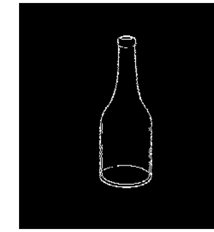
Functional MRI

Typical stimuli



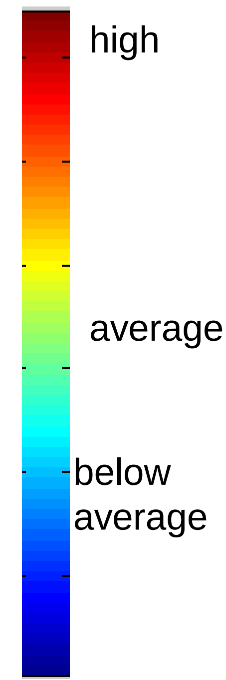
fMRI activation for "bottle":



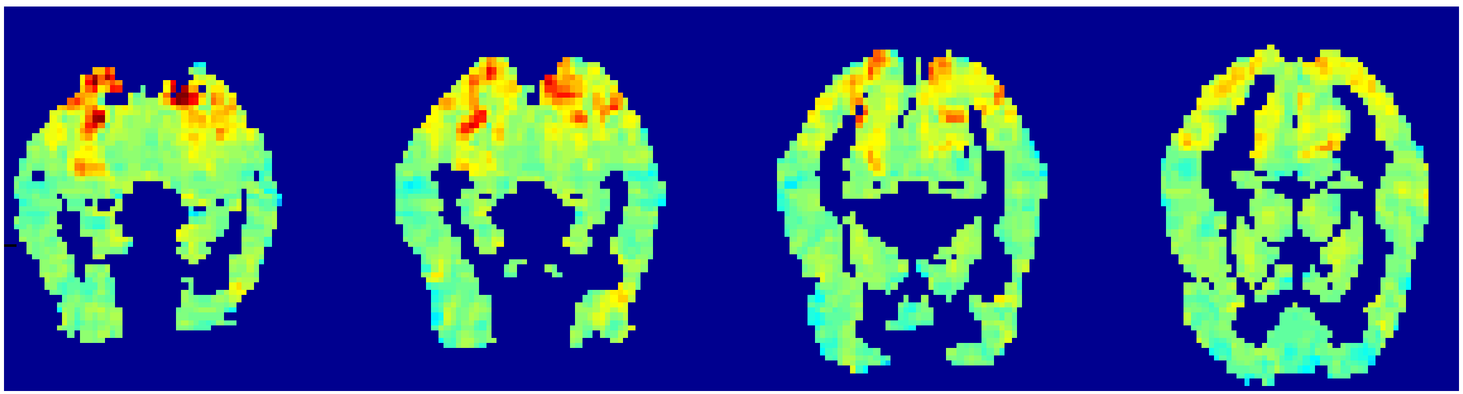


bottle

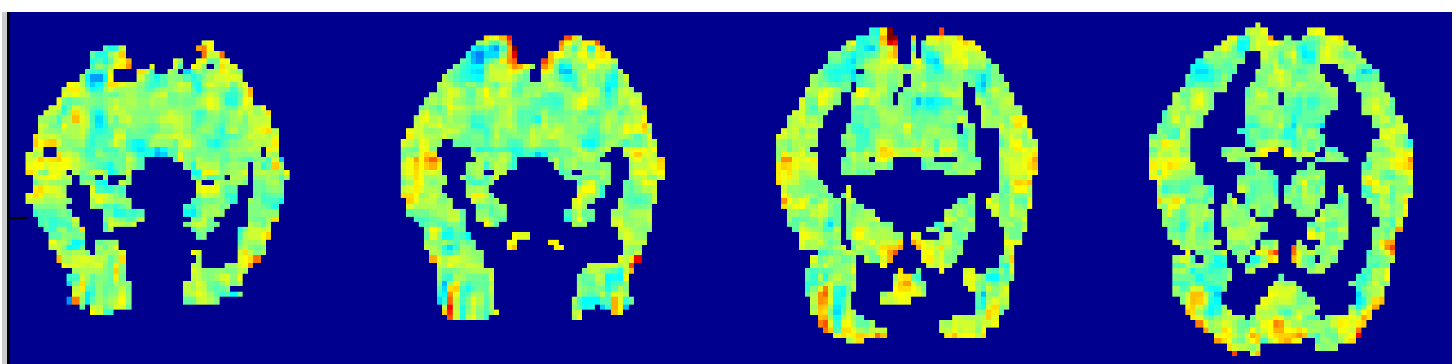
fMRI activation



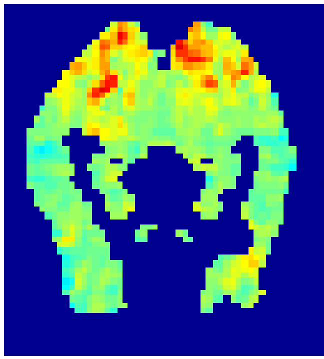
Mean activation averaged over 60 different stimuli:



"bottle" minus mean activation:



Classifiers trained to decode the stimulus word



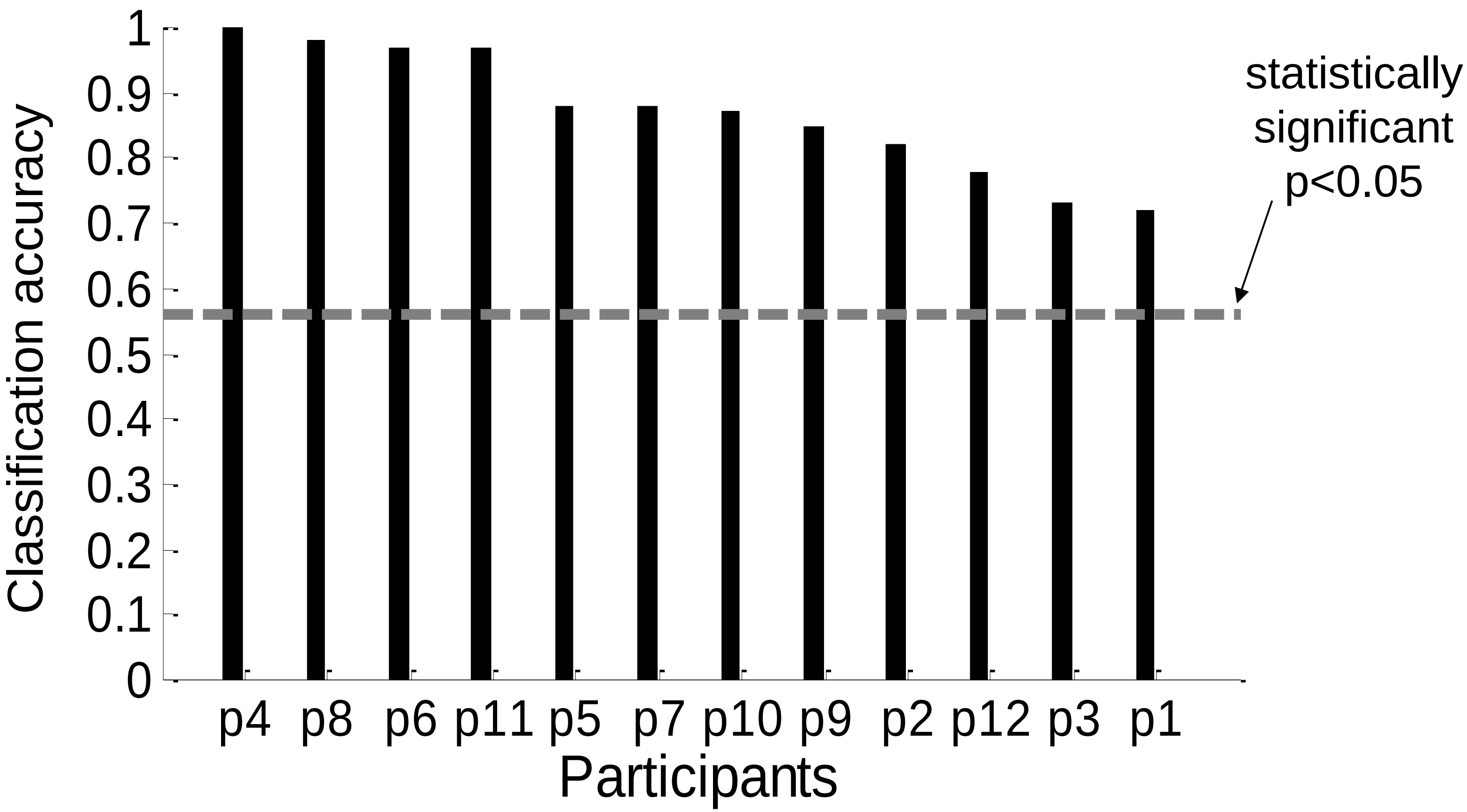
Trained Classifier

(SVM, Logistic regression, Deep net,Bayesian classifier ...)

~~Hammer~~
or
Bottle

(classifier as virtual sensor of mental state)

Classification task: is person viewing a “tool” or “building”?

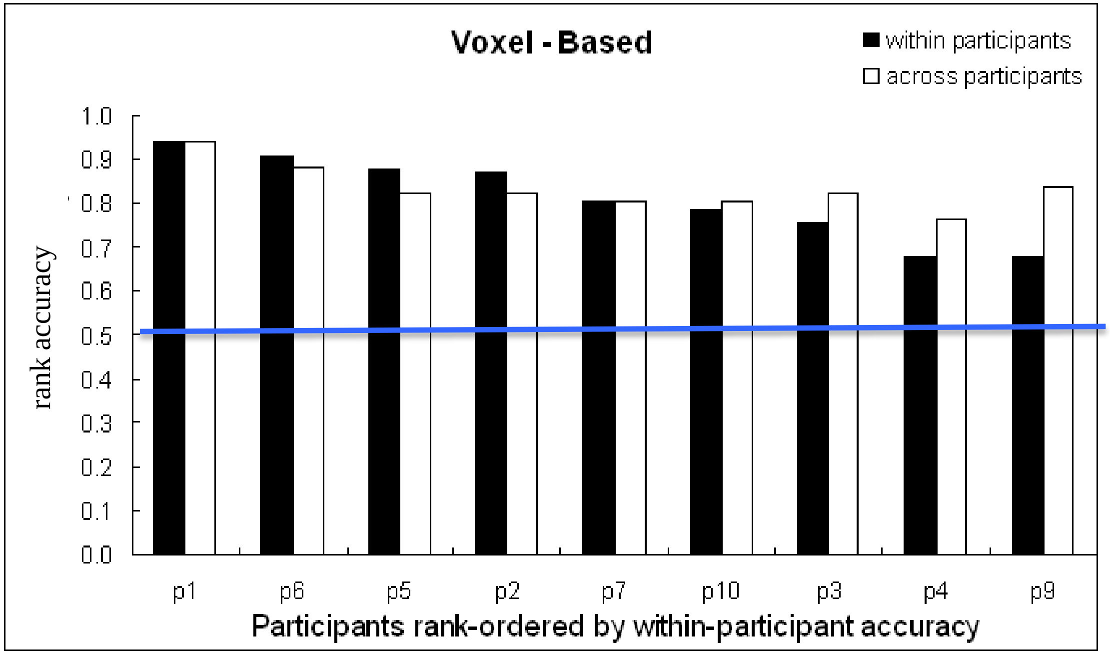


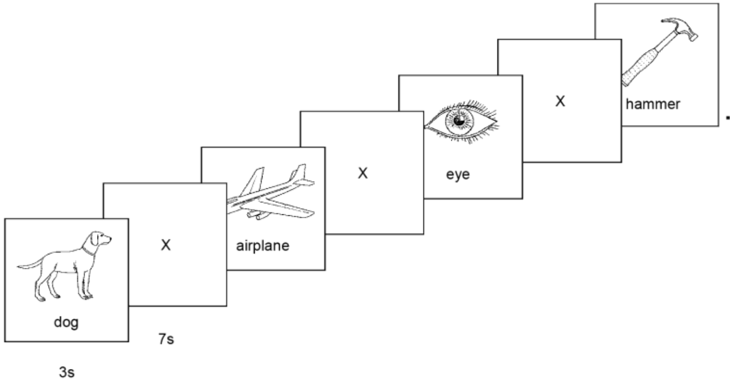
Are neural representations similar across people?

Can we train classifiers on one group of people, then decode from new person?

Are representations similar across people?

YES





classify which of 60 items

Lessons from fMRI Word Classification

Neural representations similar across

- people
- language
- word vs. picture

Easier to decode:

- concrete nouns
- emotion nouns

Harder to decode:

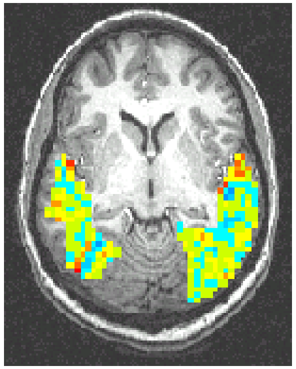
- abstract nouns
- verbs*

* except when placed in context

Predictive Model?

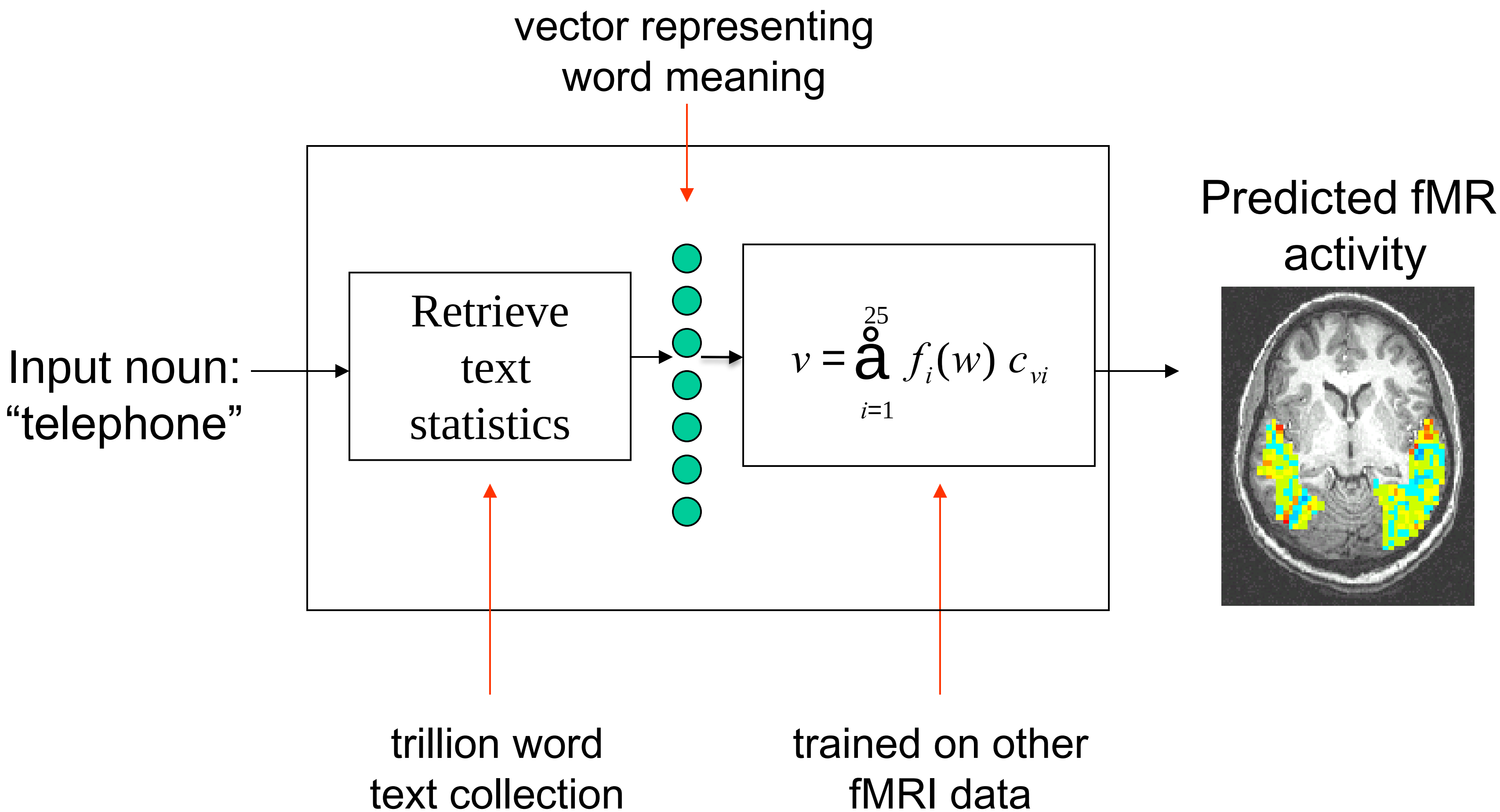
Arbitrary noun

Predicted fMRI activity



Predictive Model?

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



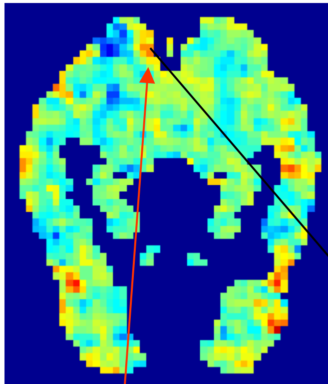
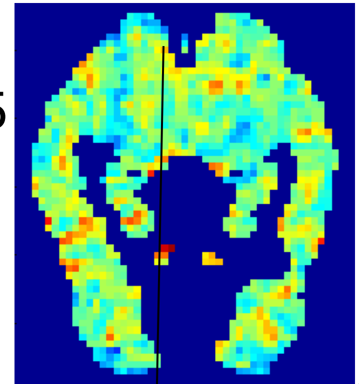
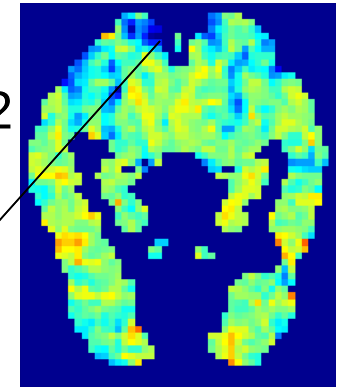
Represent stimulus noun by co-occurrences with 25 verbs*

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  + …

$f_{eat}(celery)$

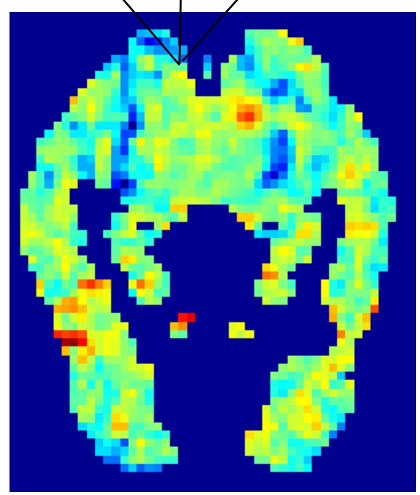
from corpus statistics

$c_{14382,eat}$

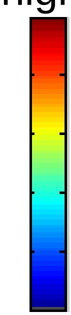
<u>learned</u>

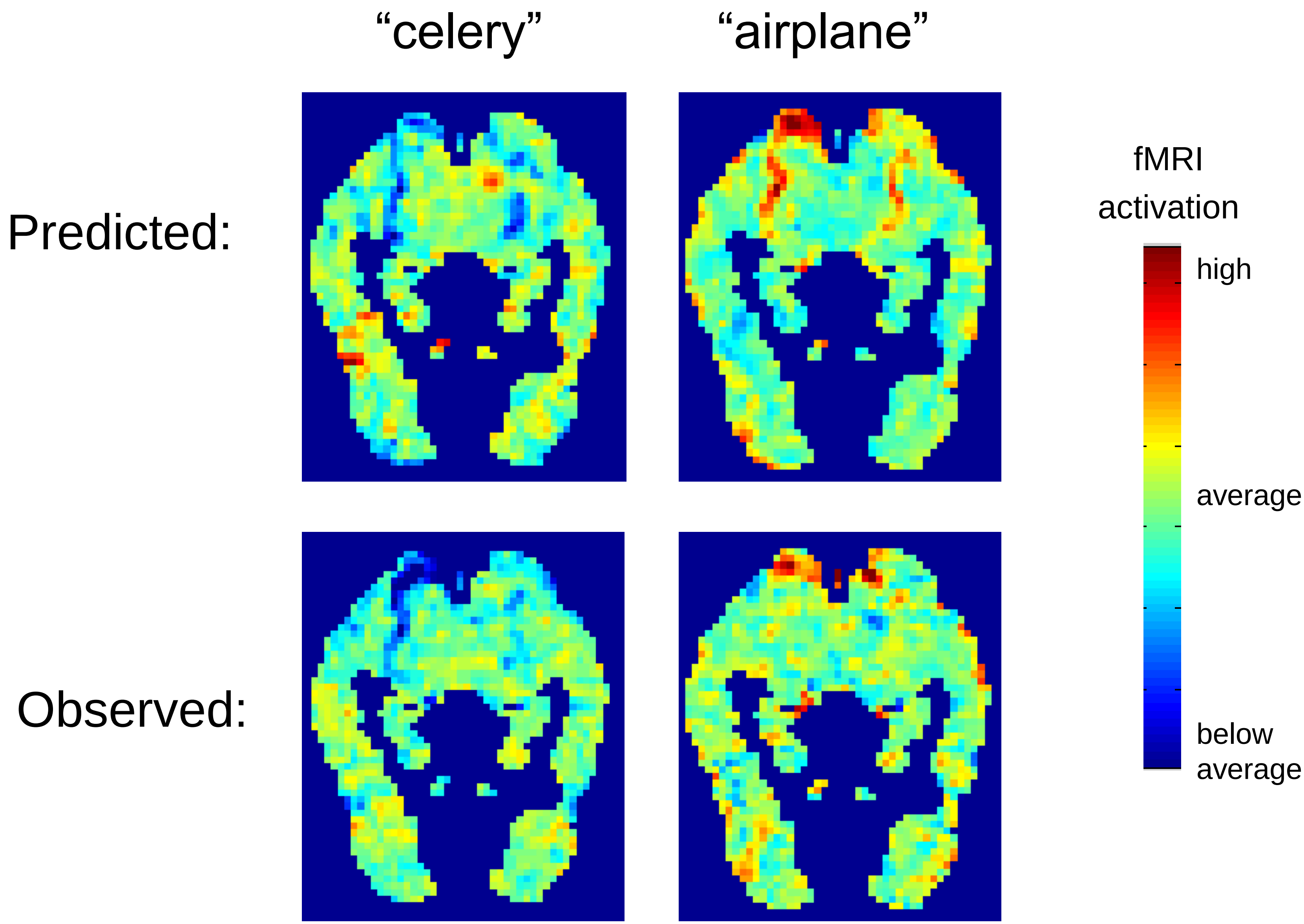
$$prediction_v = \sum_{i=1}^{25} f_i(w)\, c_{vi}$$

500,000 learned c_{vi} parameters



Predicted "Celery"



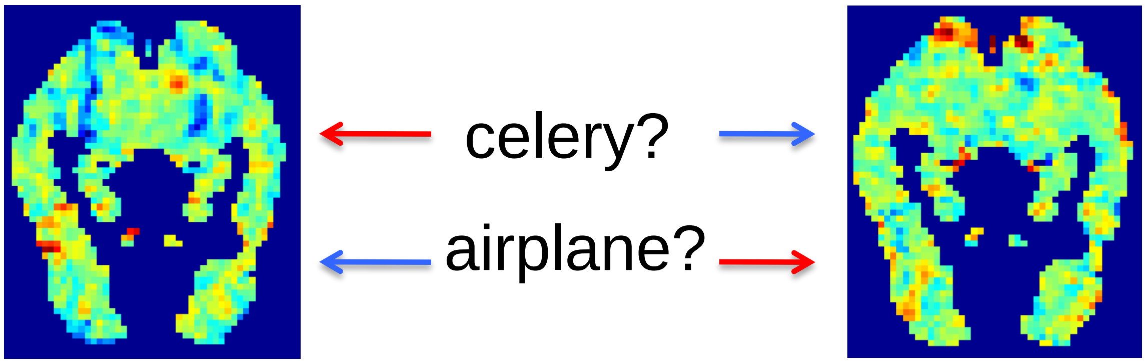


Predicted and observed fMRI images for "celery" and "airplane" after training on other nouns.

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

Evaluating the Computational Model

- Leave two words out during training

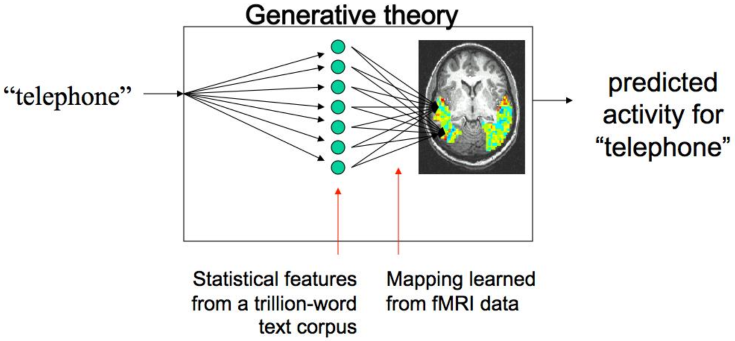


1770 test pairs in leave-2-out:

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

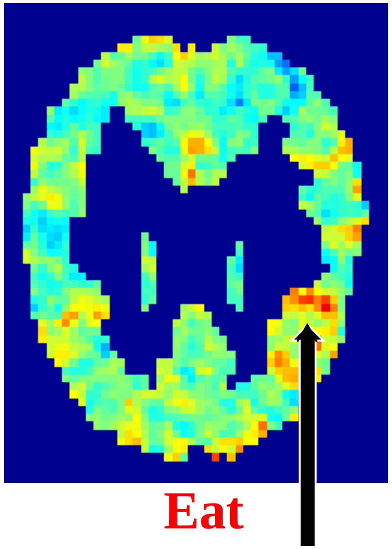
Mean accuracy over 9 subjects: 0.79

Learned activities associated with meaning components



Participant P1

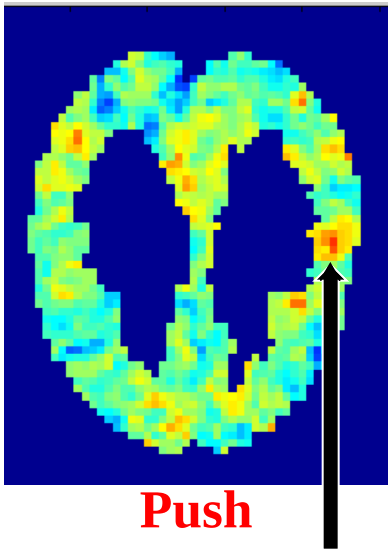
Semantic feature:



Eat

"Gustatory cortex"

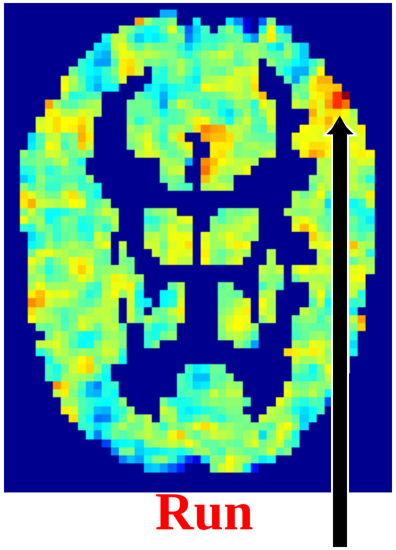
Pars opercularis (z=24mm)



Push

"somato-sensory"

Postcentral gyrus (z=30mm)



Run

"Biological motion"

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

Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78

Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
218 features collected using Mechanical Turk	**.83**

Is it heavy?
Is it flat?
Is it curved?
Is it colorful?
Is it hollow?
Is it smooth?
Is it fast?
Is it bigger than a car?
Is it usually outside?
Does it have corners?
Does it have moving parts?
Does it have seeds?

Can it break?
Can it swim?
Can it change shape?
Can you sit on it?
Can you pick it up?
Could you fit inside of it?
Does it roll?
Does it use electricity?
Does it make a sound?
Does it have a backbone?
Does it have roots?
Do you love it?
…

features authored by Dean Pomerleau.

feature values 1 to 5

features collected from at least three people

people provided by Amazon’s “Mechanical Turk”

Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
218 features collected using *Mechanical Turk**	**.83**
20 features discovered from the data**	**.86**

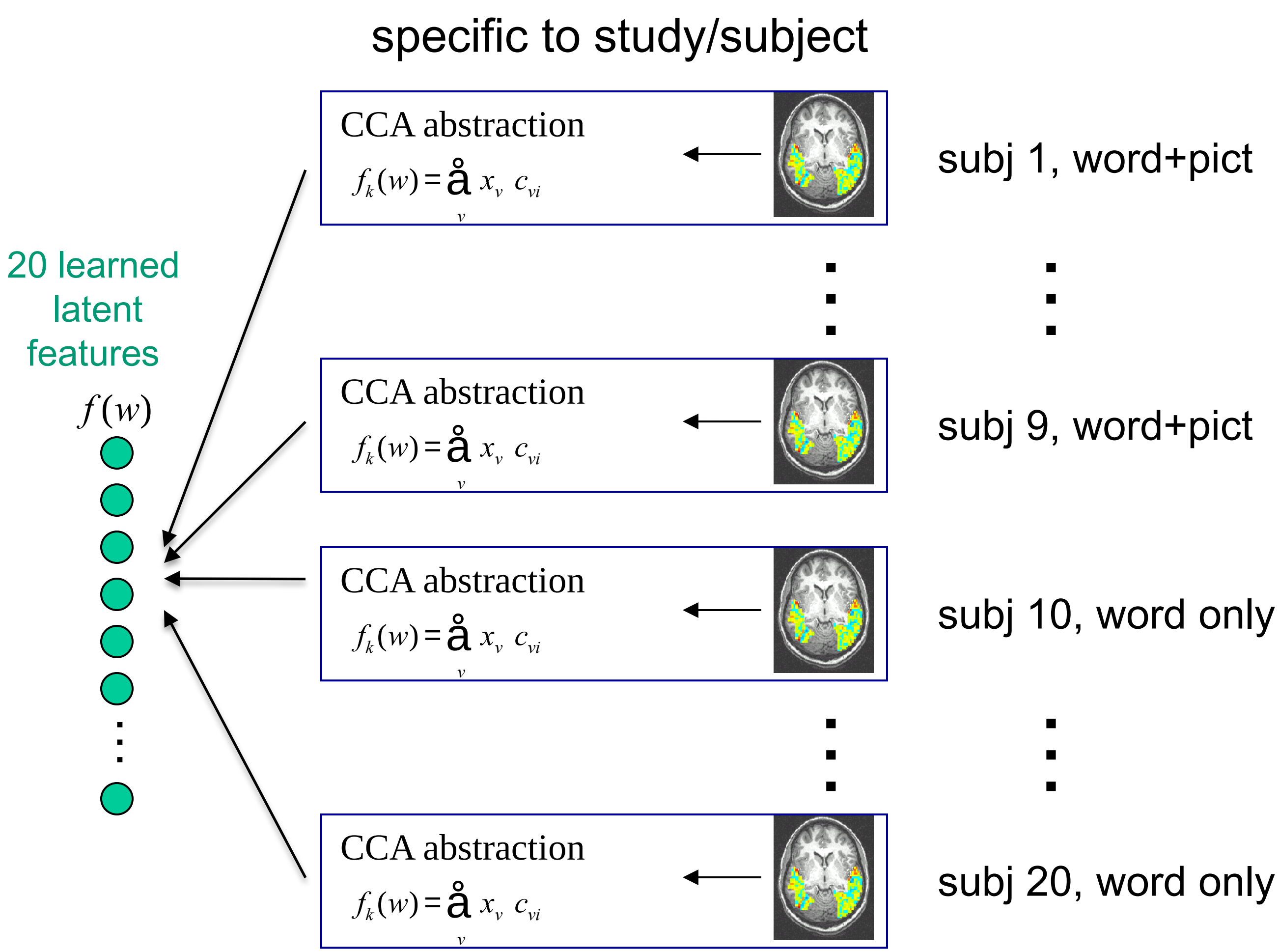
* developed by Dean Pommerleau

** developed by Indra Rustandi

Discovering shared semantic basis

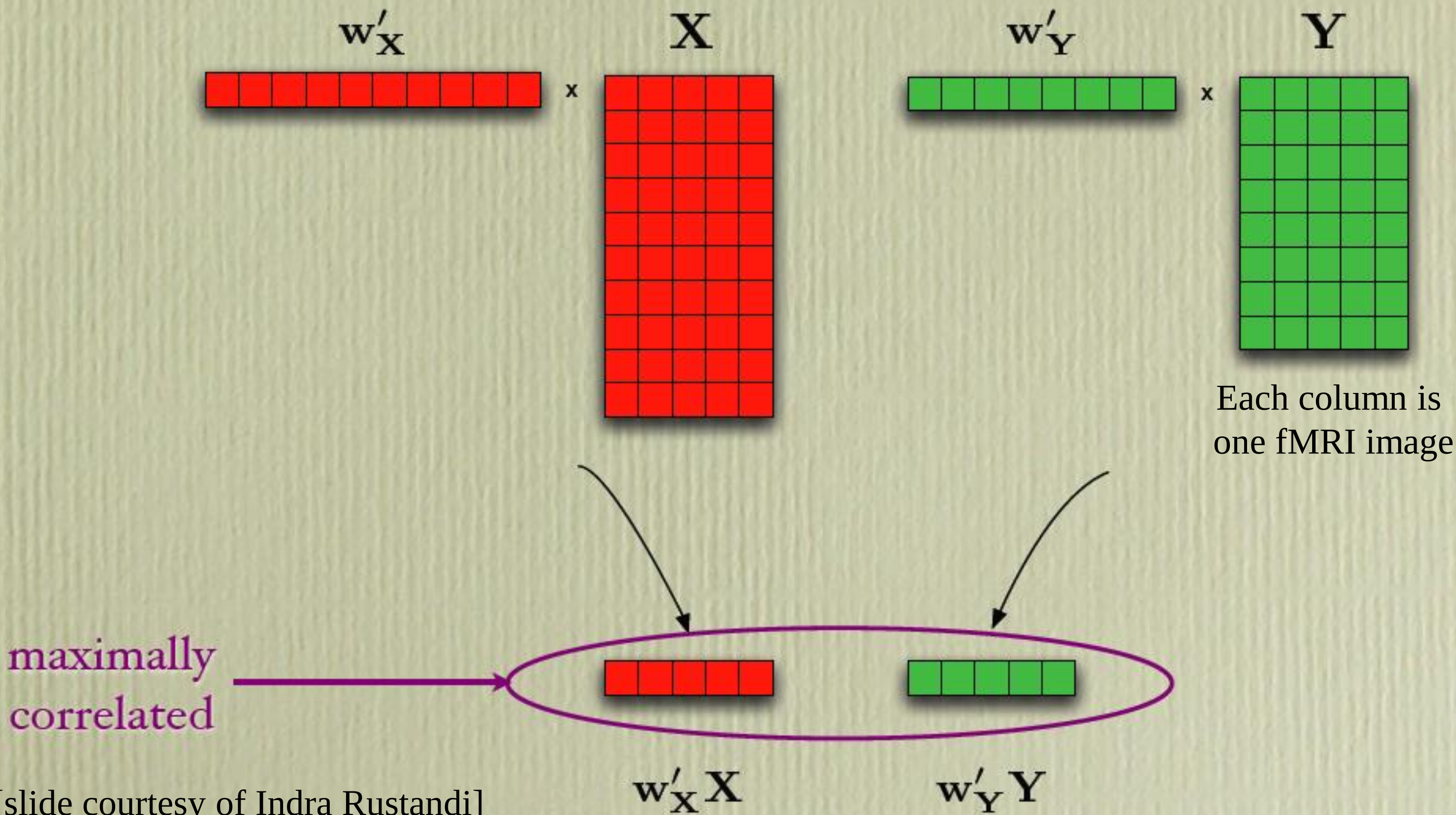
[Rustandi et al., 2009]

1. Use CCA to discover latent features across subjects



Canonical correlation analysis

$$Corr(A,B) = \frac{1}{N}\sum_{i=1}^{N}\frac{(A_i-\bar{A})}{\sigma_A}\ \frac{(B_i-\bar{B})}{\sigma_b}$$

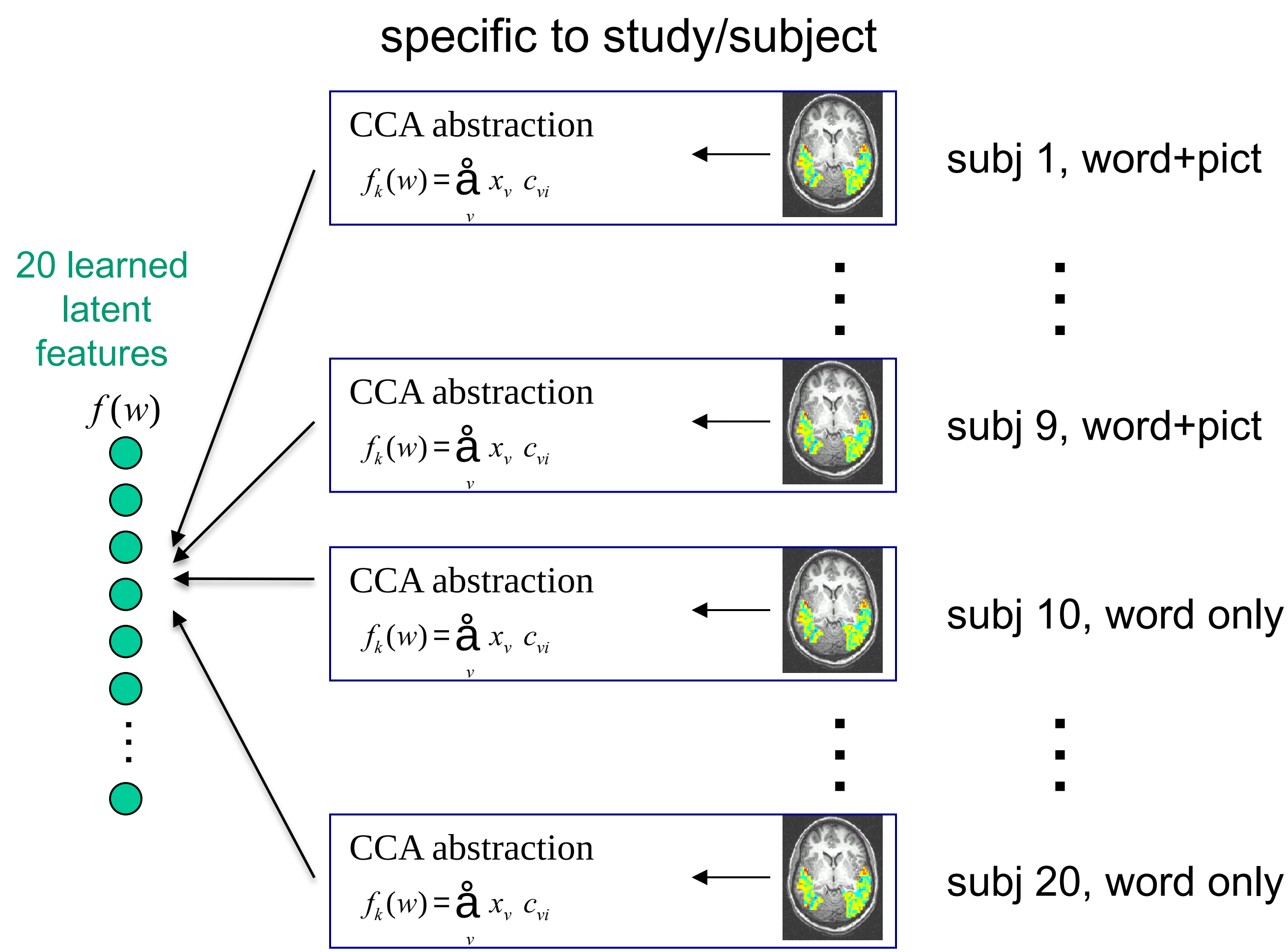


[slide courtesy of Indra Rustandi]

Discovering shared semantic basis

[Rustandi et al., 2009]

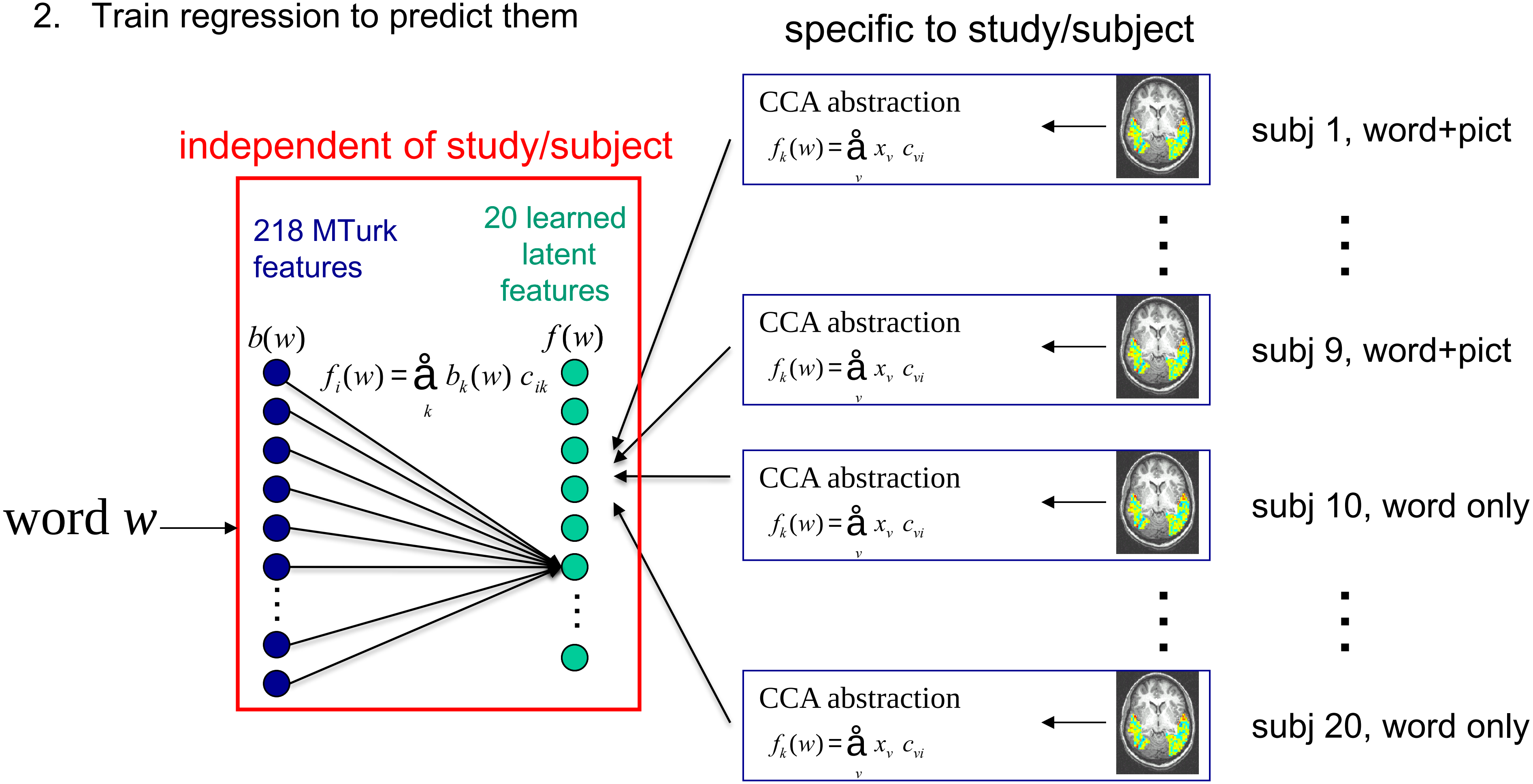
1. Use CCA to discover latent features



Discovering shared semantic basis

[Rustandi et al., 2009]

1. Use CCA to discover latent features
2. Train regression to predict them

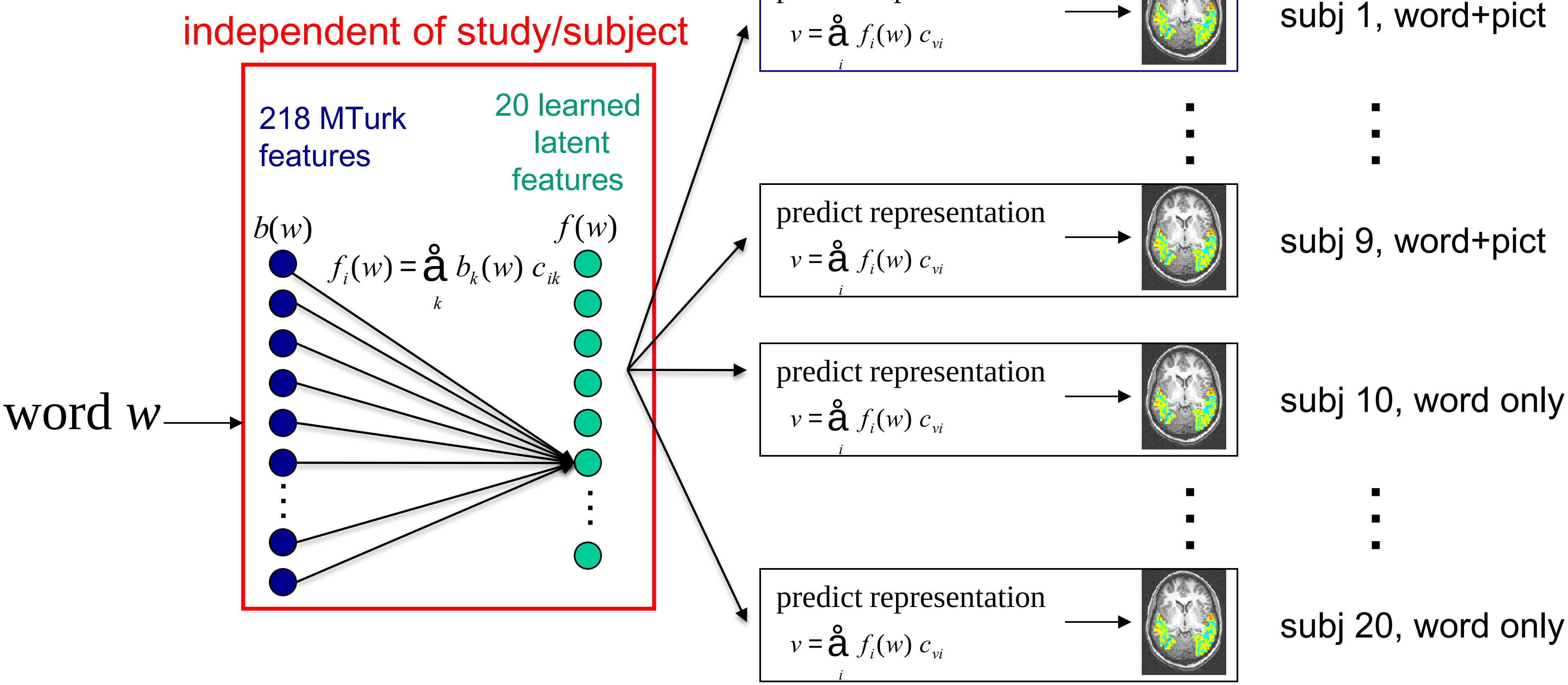


Discovering shared semantic basis

[Rustandi et al., 2009]

1. Use CCA to discover latent features
2. Train regression to predict them
3. Invert CCA mapping

specific to study/subject



CCA Components: Top Stimulus Words

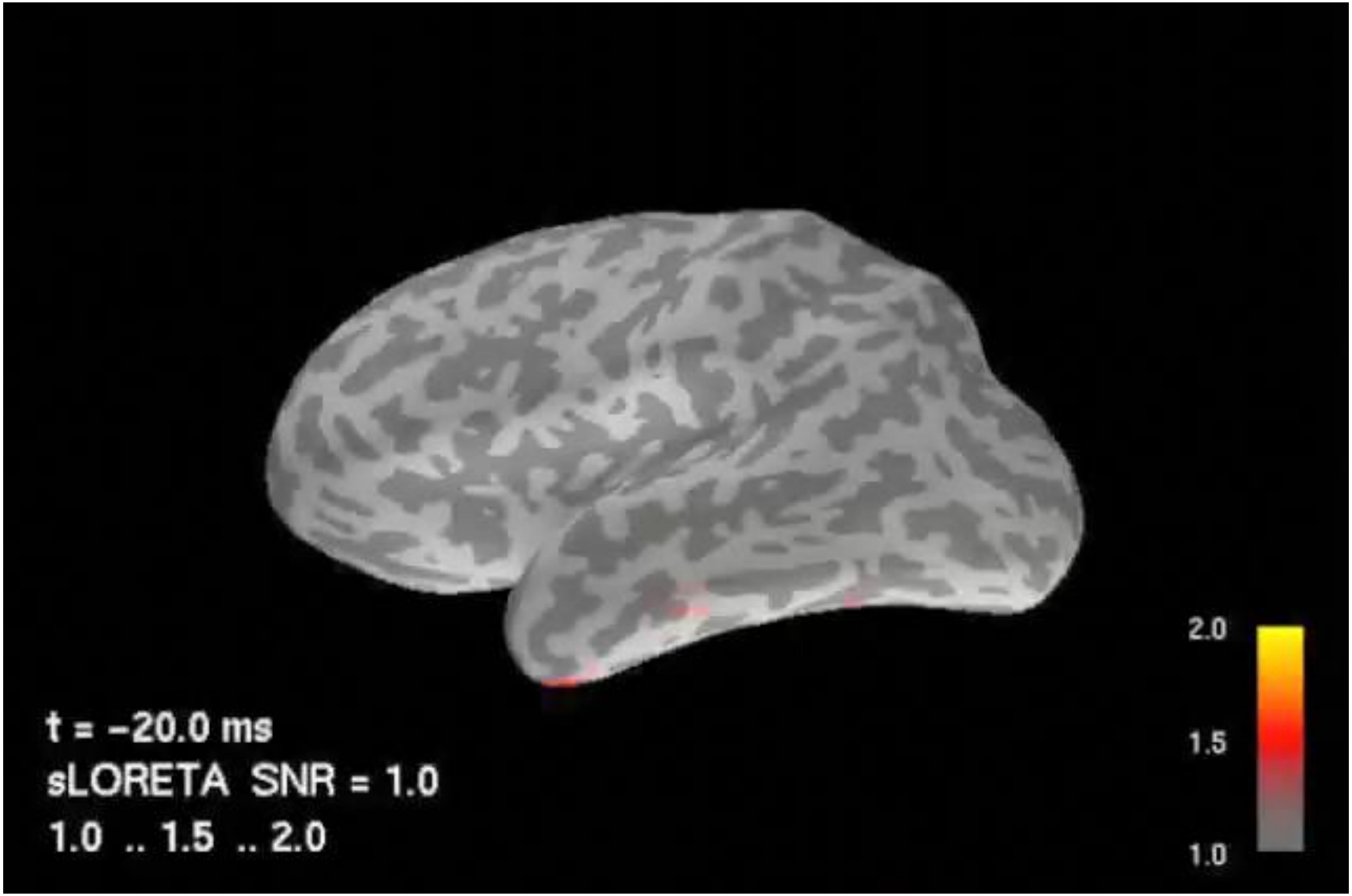
	component 1	component 2	component 3	component 4
Stimuli that most activate it	apartment church closet house barn	screwdriver pliers refrigerator knife hammer	telephone butterfly bicycle beetle dog	pants dress glass coat chair

shelter? manipulation?

things that touch my body?

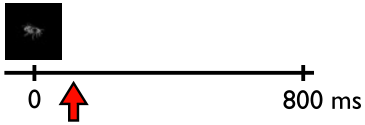
Timing?

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

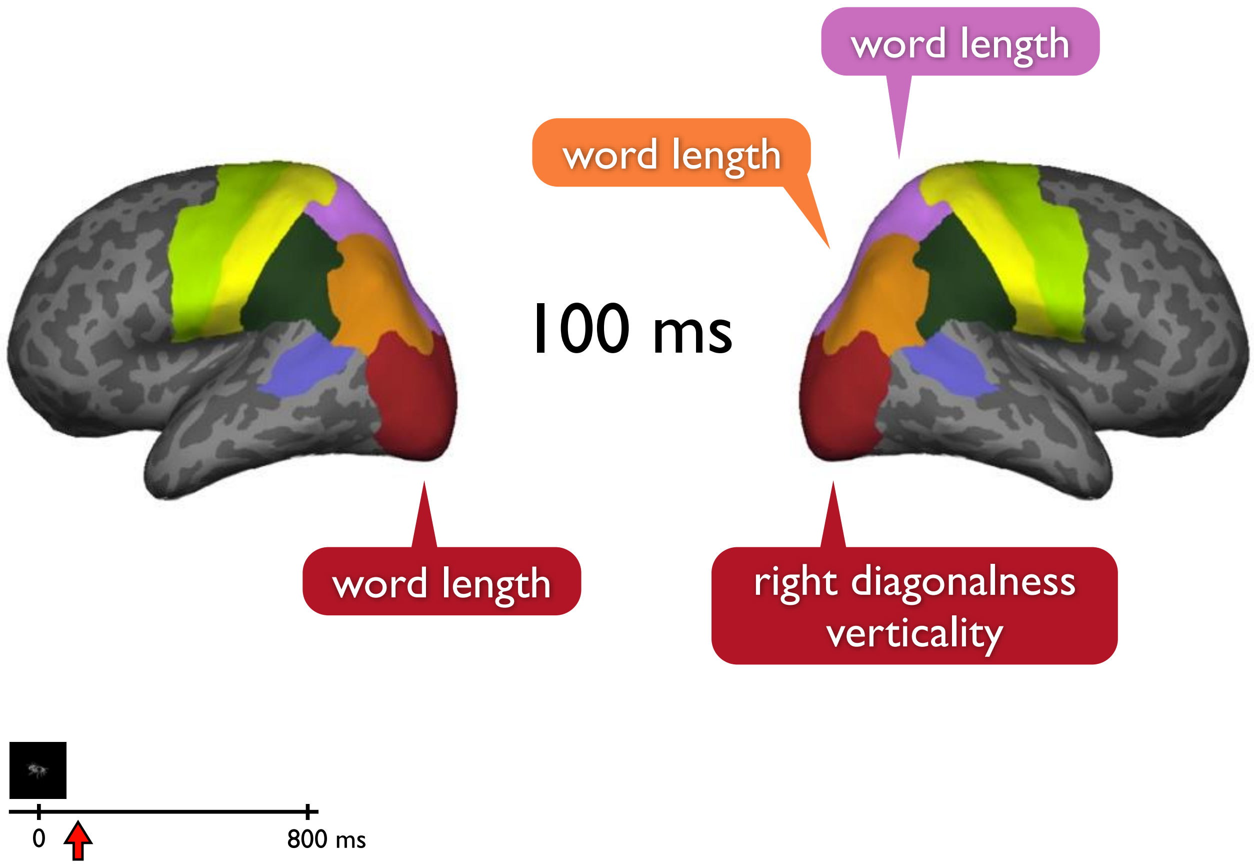


[Sudre et al., *NeuroImage* 2012]

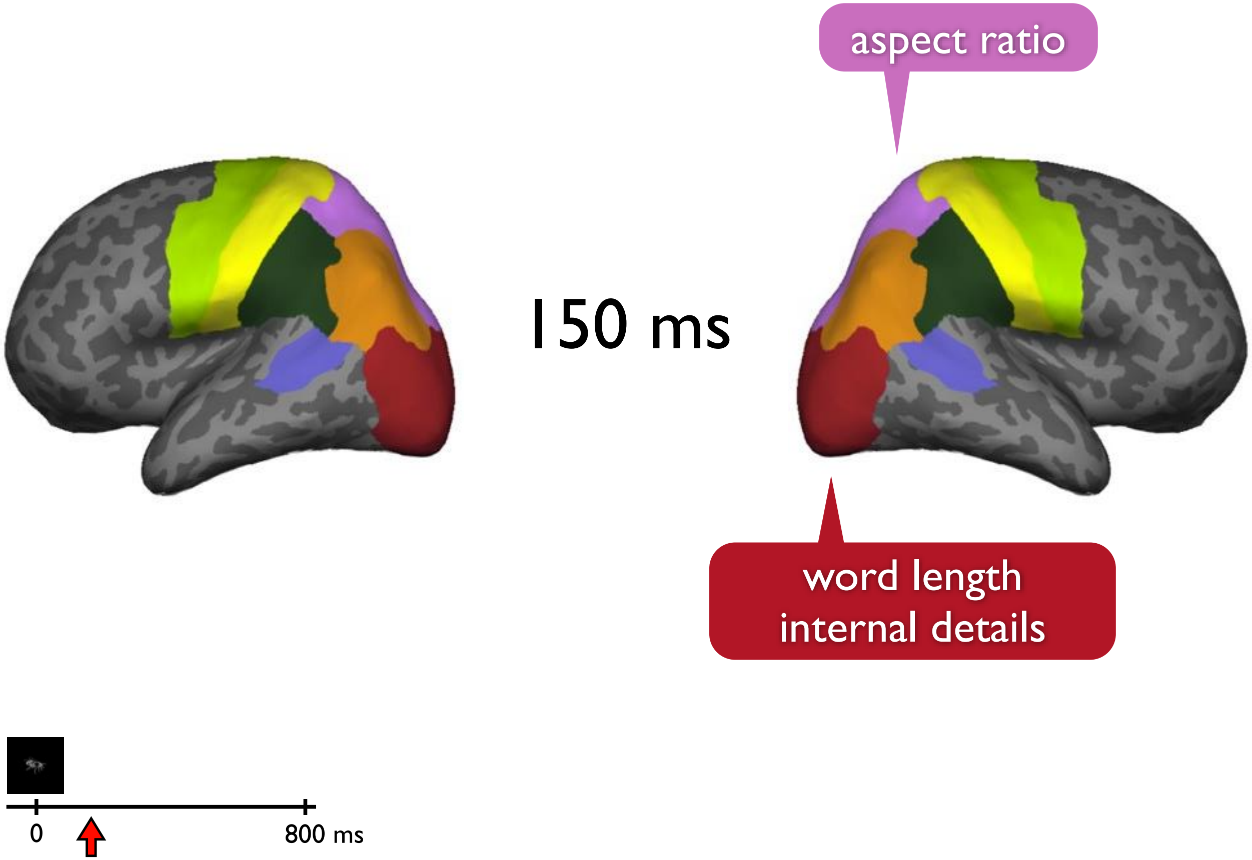
50 ms



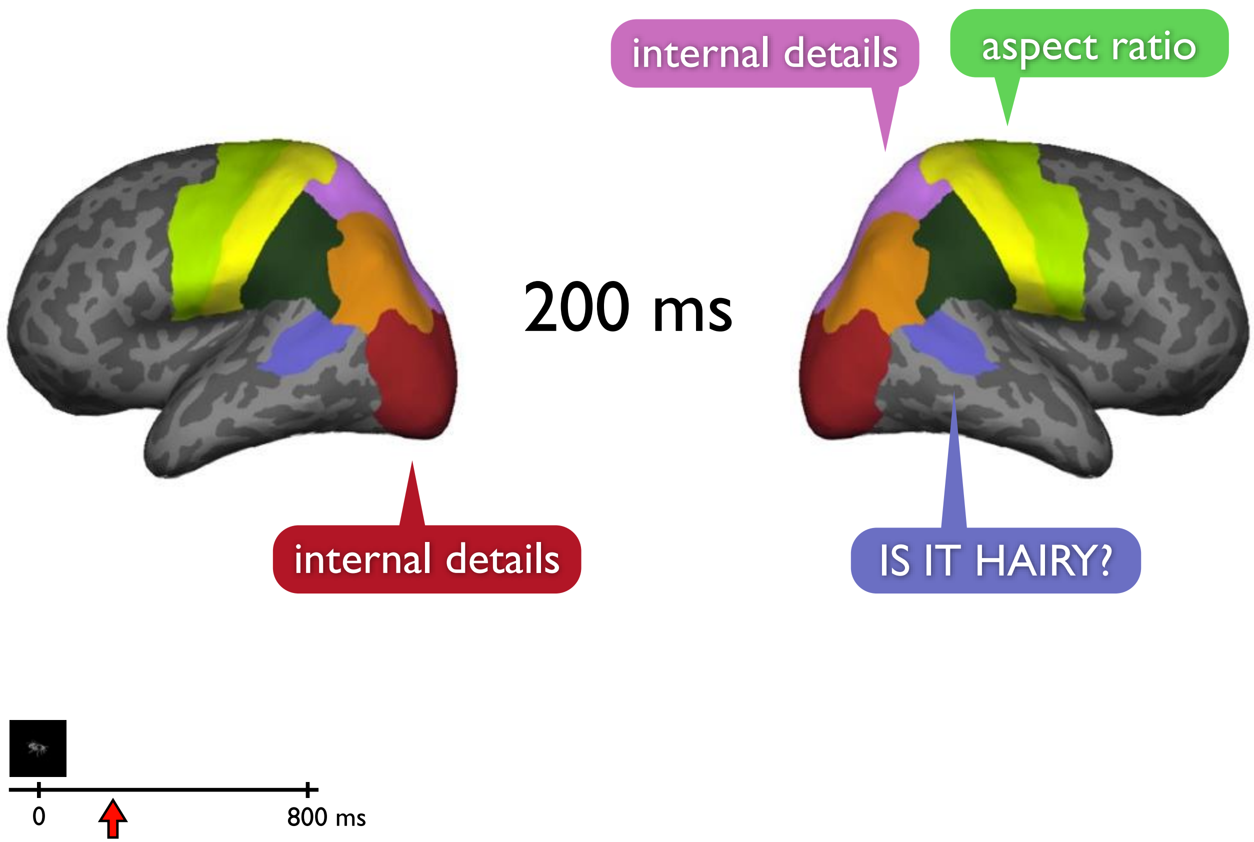
[Sudre et al., *NeuroImage* 2012]



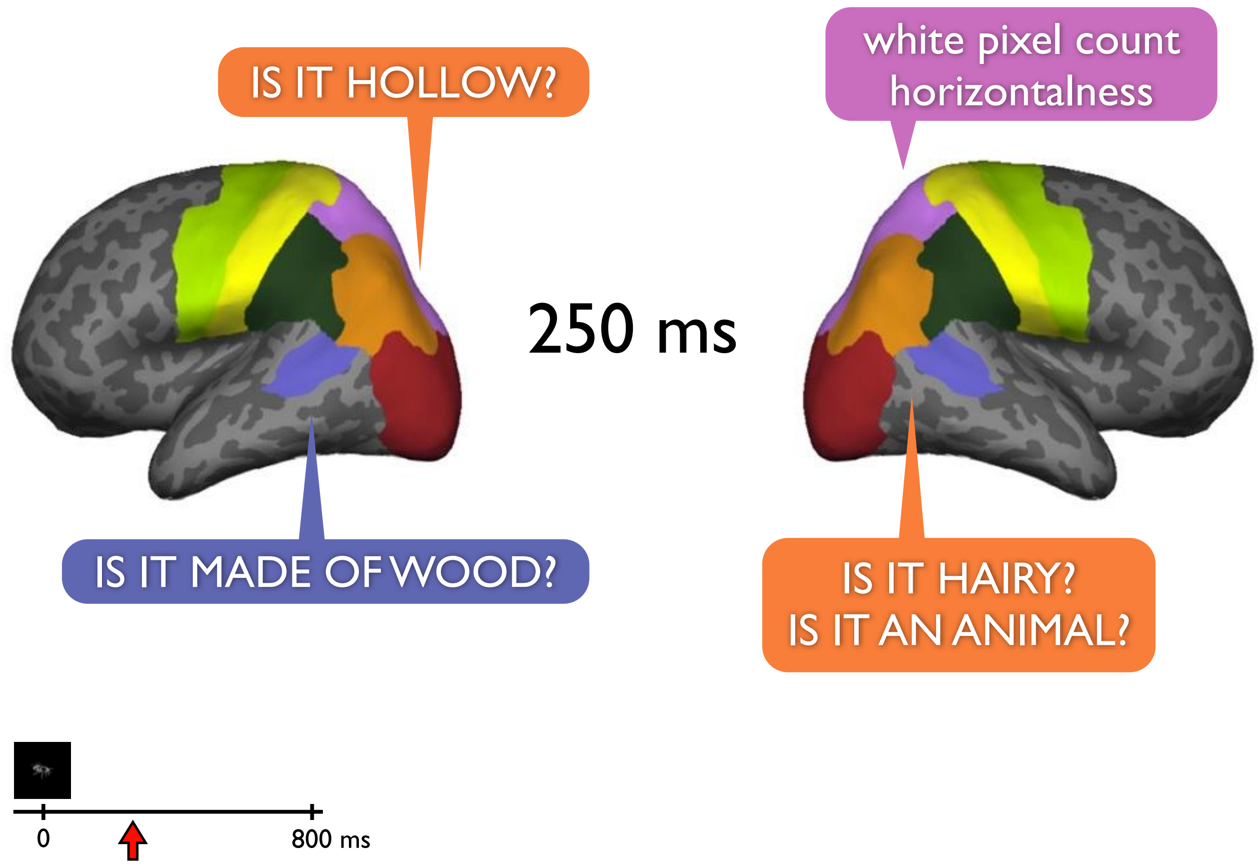
[Sudre et al., 2012]



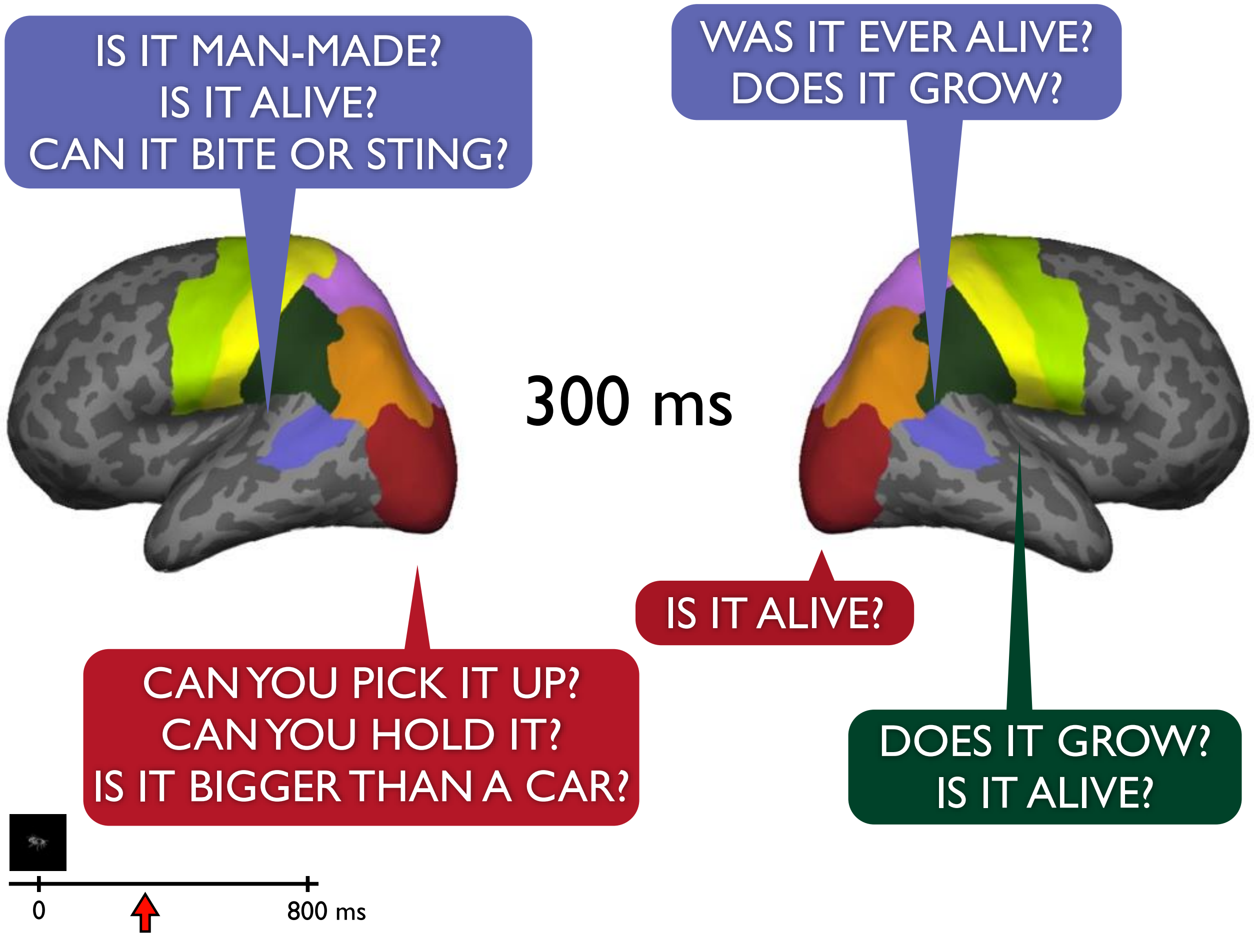
[Sudre et al., 2012]



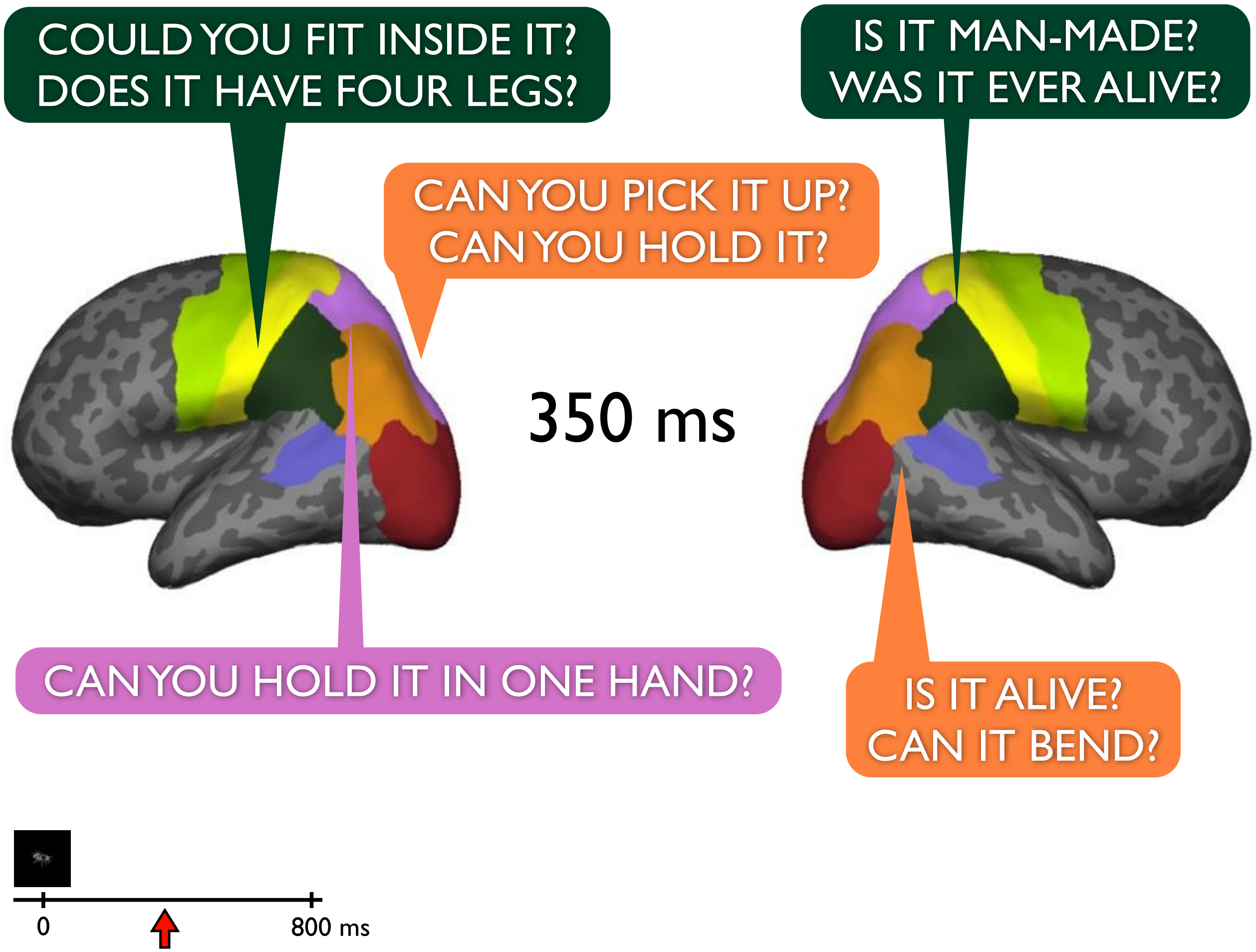
[Sudre et al., 2012]



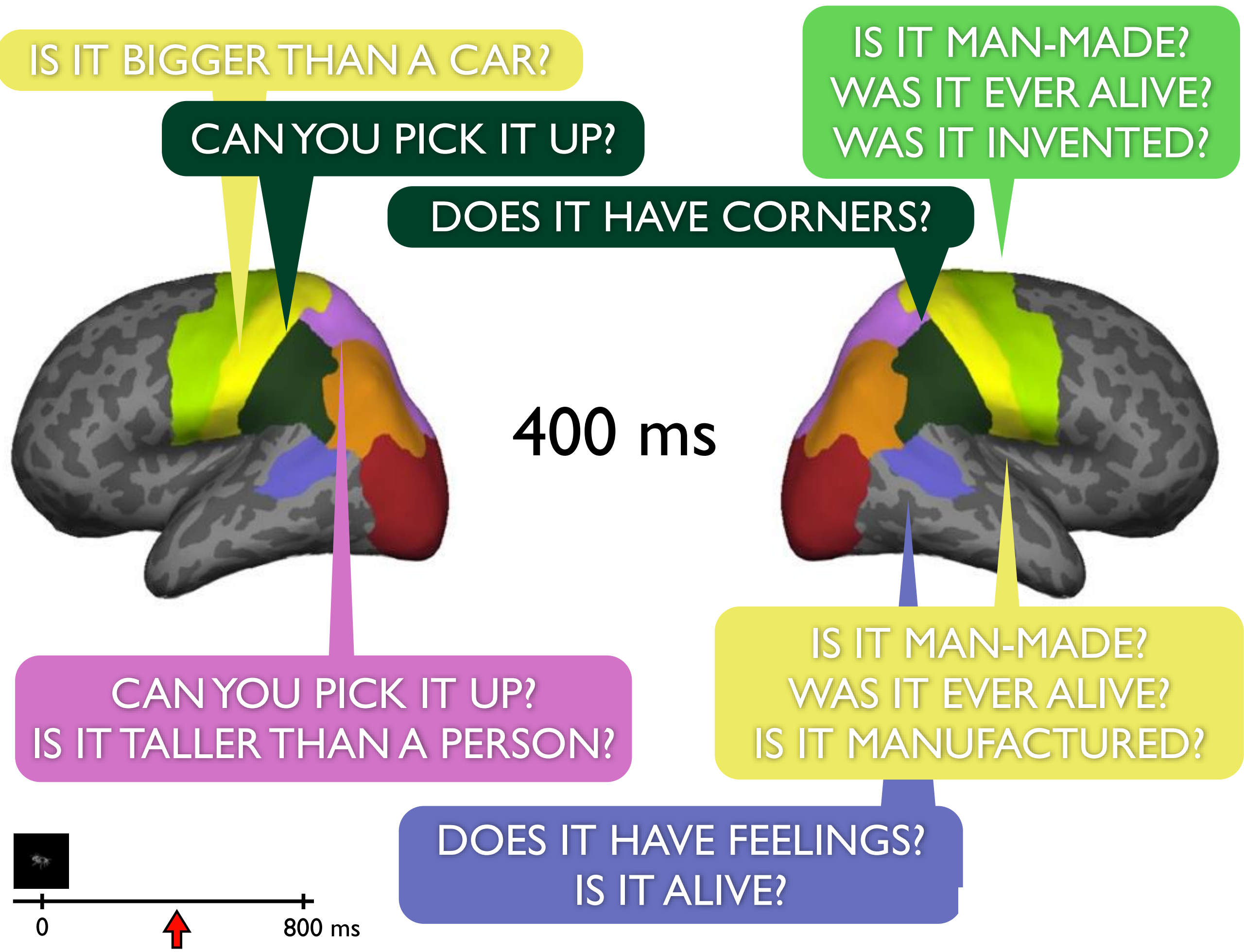
[Sudre et al., 2012]



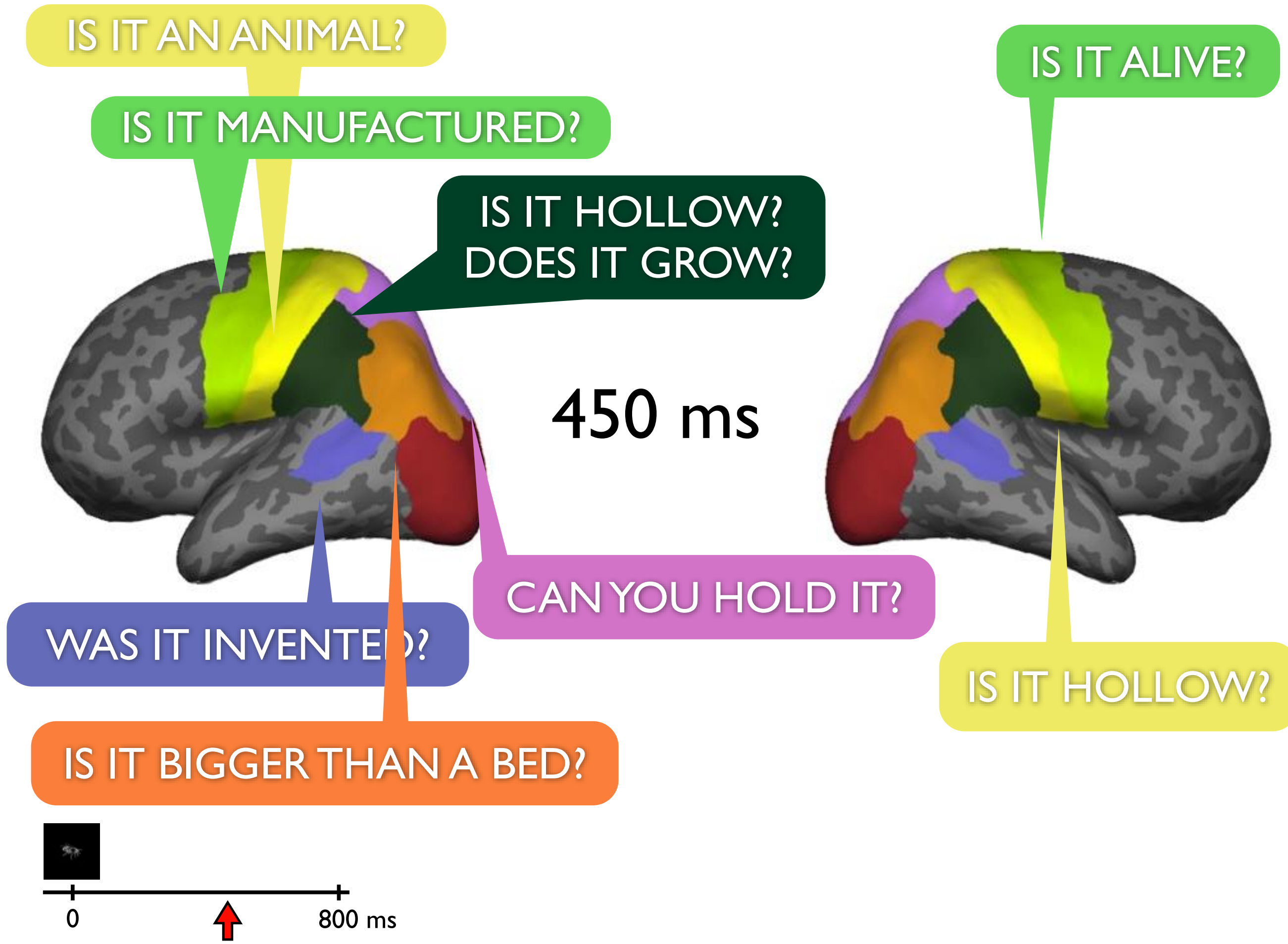
[Sudre et al., 2012]



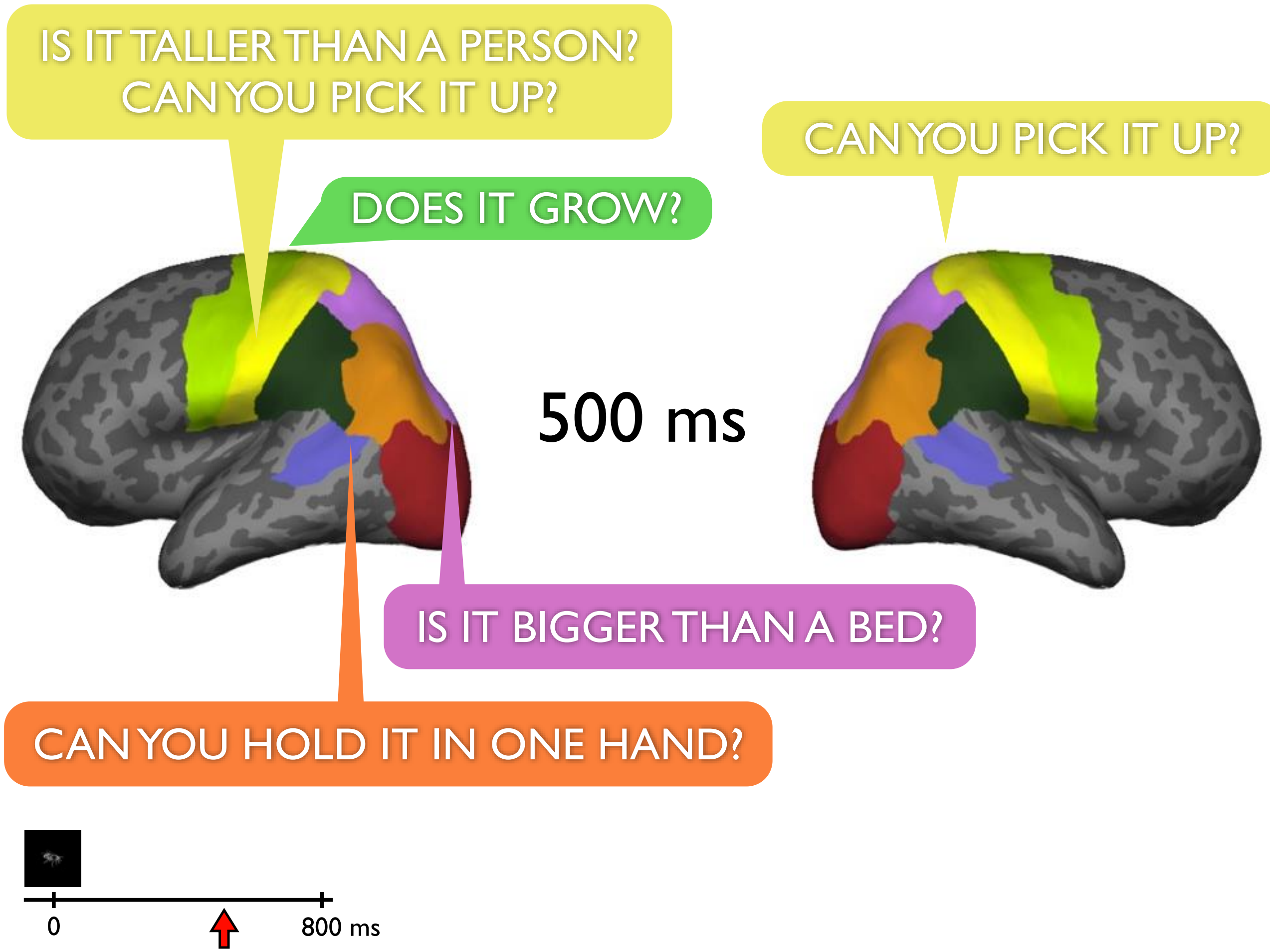
[Sudre et al., 2012]



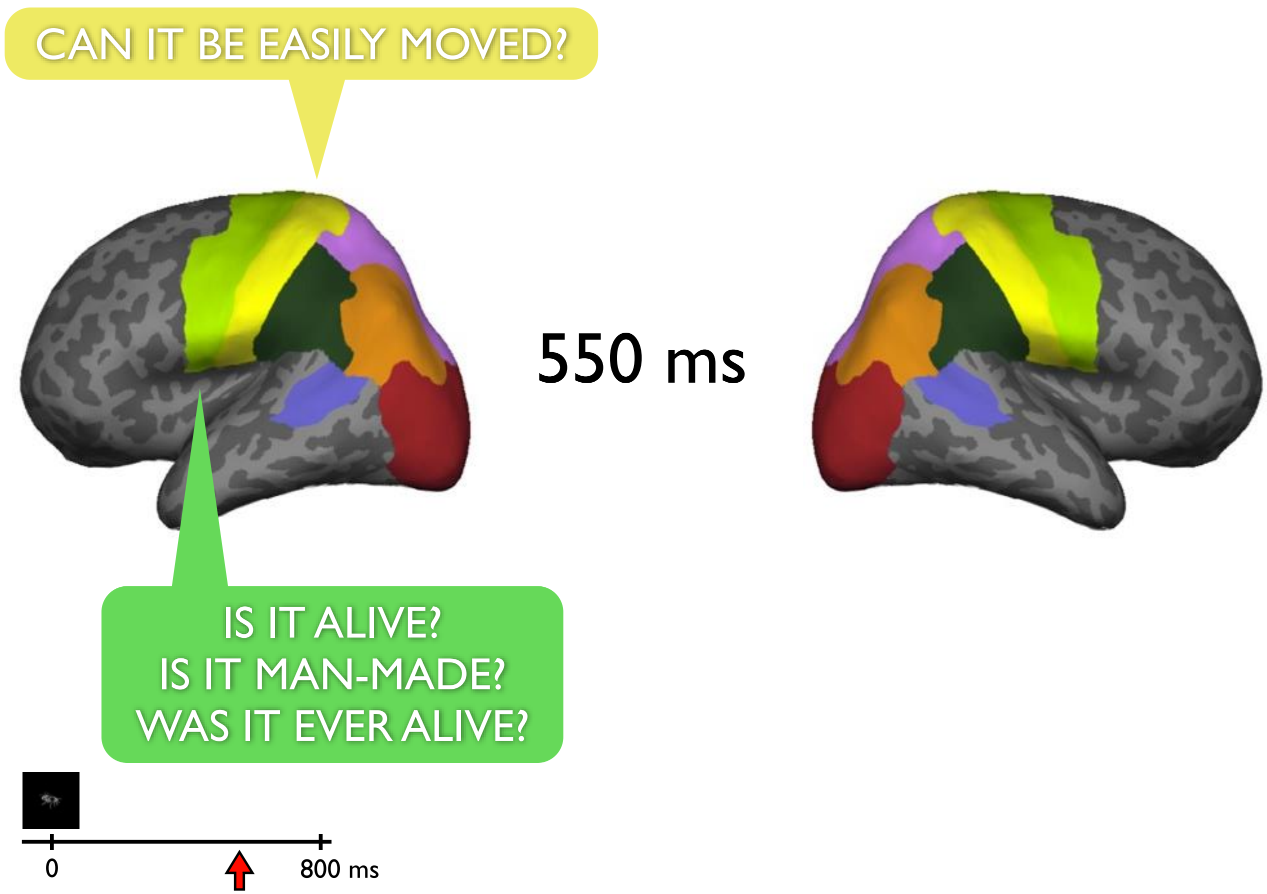
[Sudre et al., 2012]



[Sudre et al., 2012]



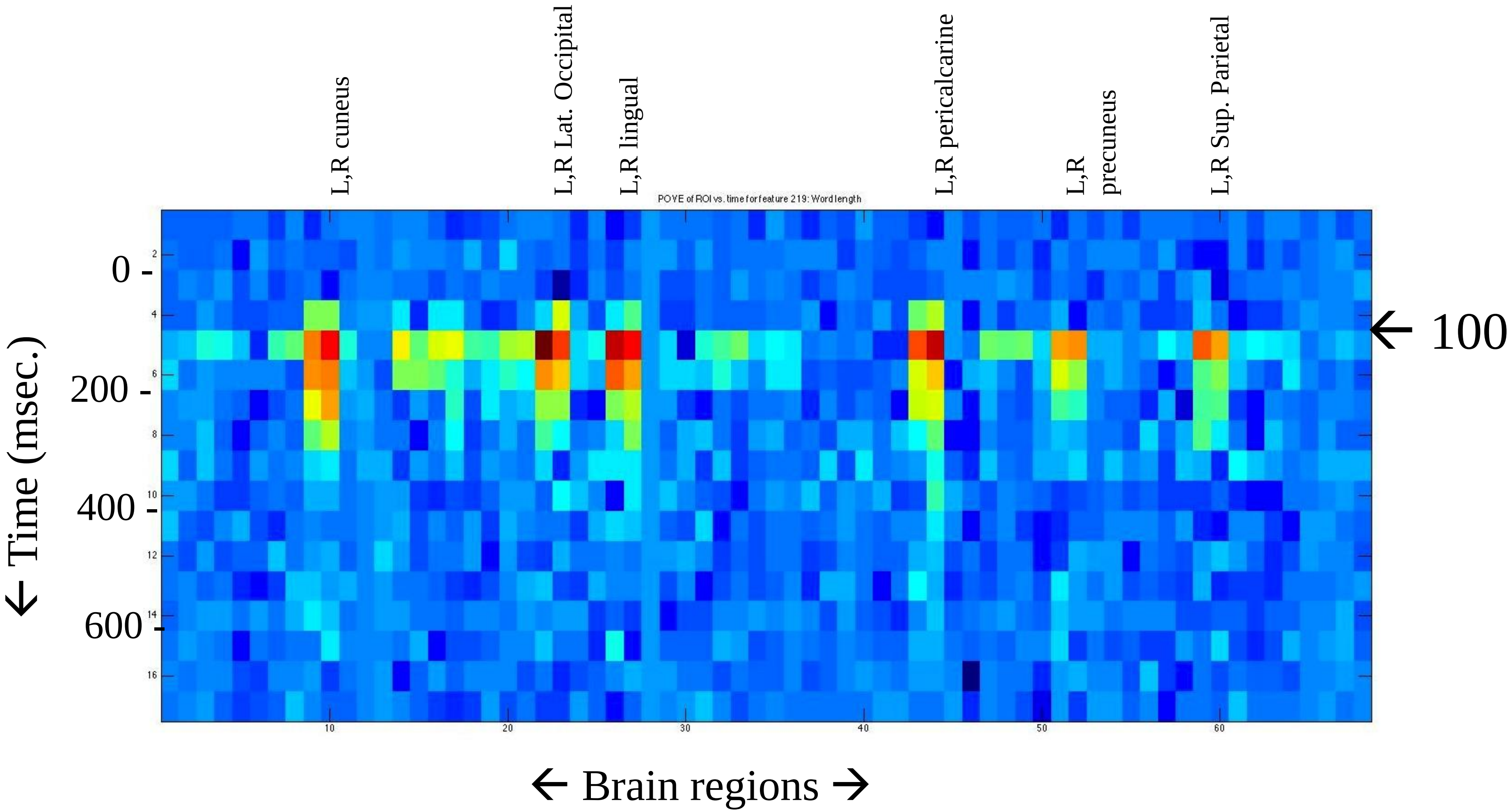
[Sudre et al., 2012]



[Sudre et al., 2012]

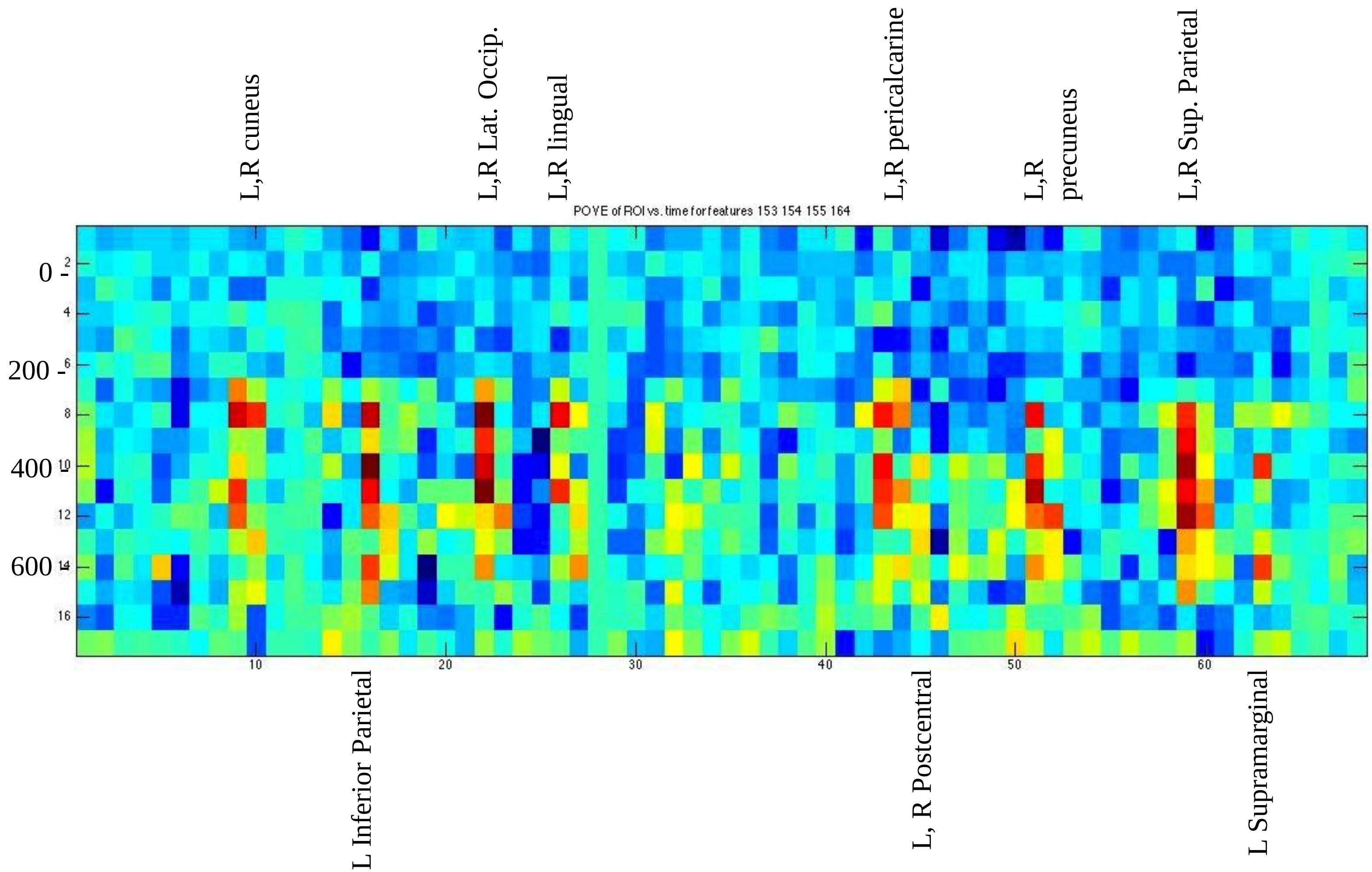
Details

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



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

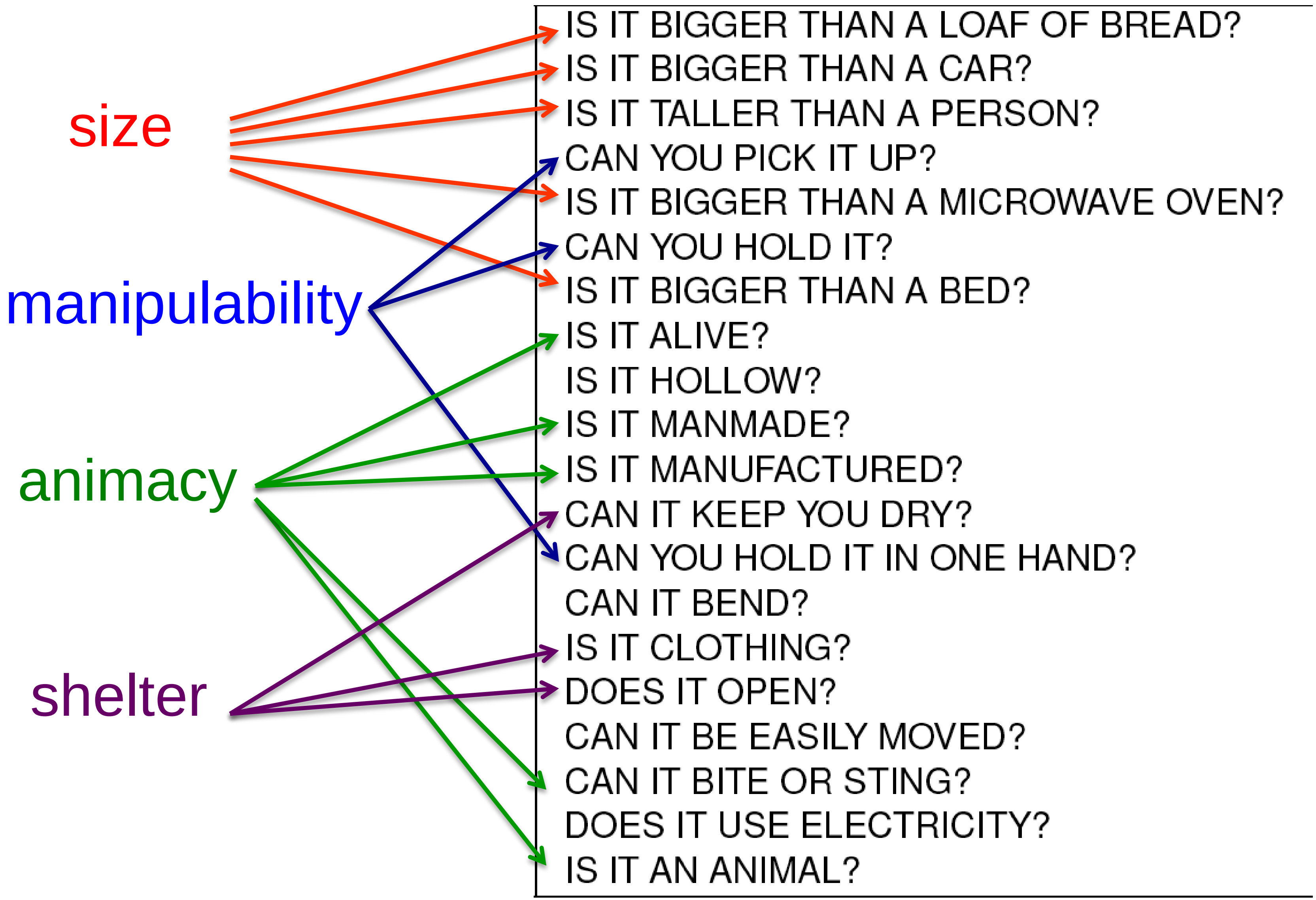
Color= decodability of “grasping“ features (initial peak: 200-300 msec)



[Sudre et al., 2012]

20 most accurately decoded semantic features out of 218

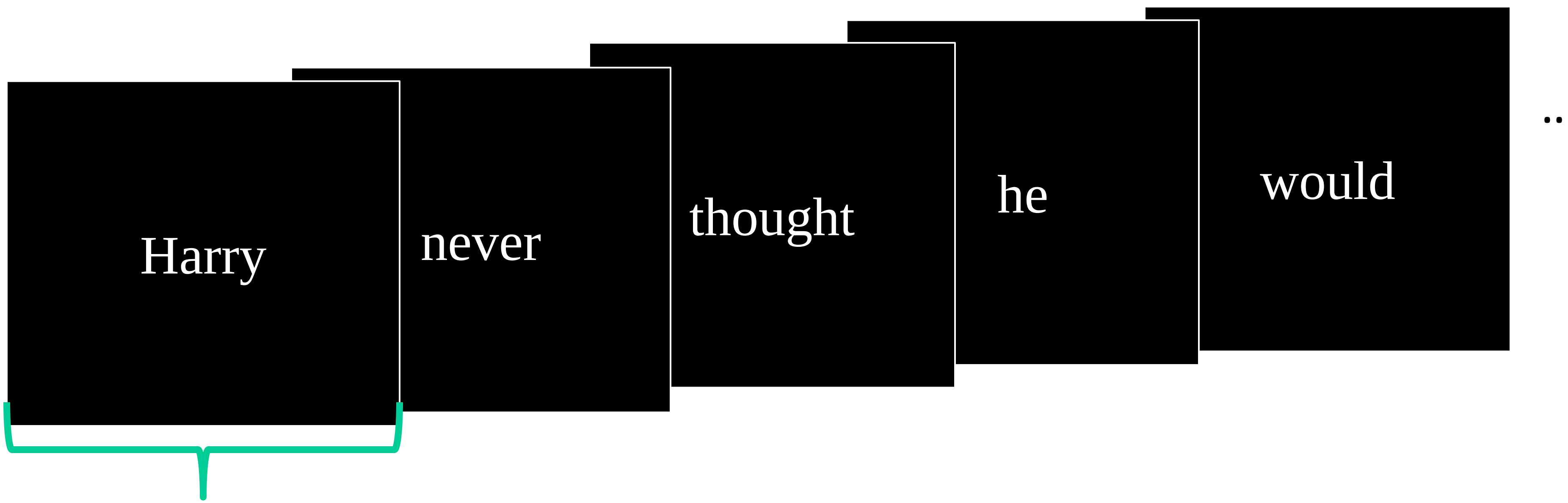
[G. Sudre et al., 2012]



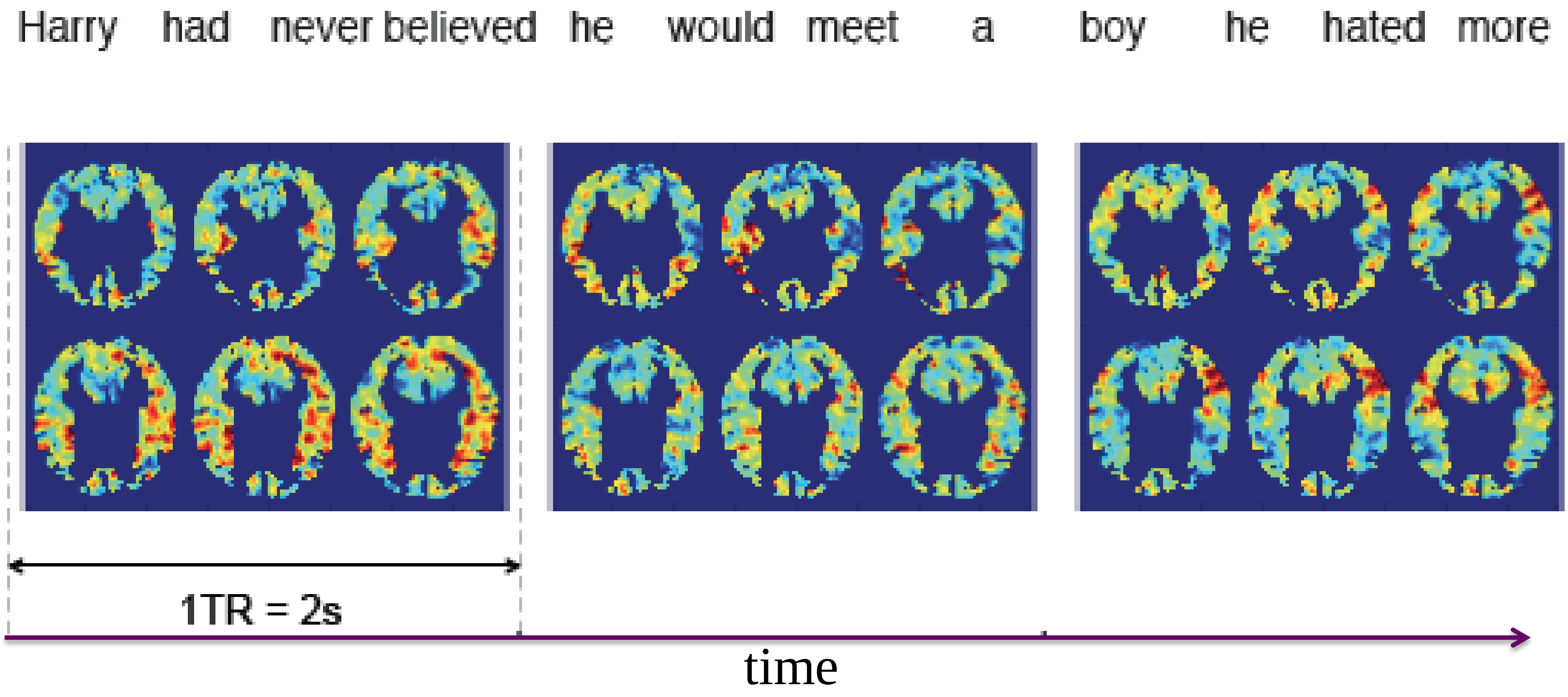
Story reading

Leila Wehbe

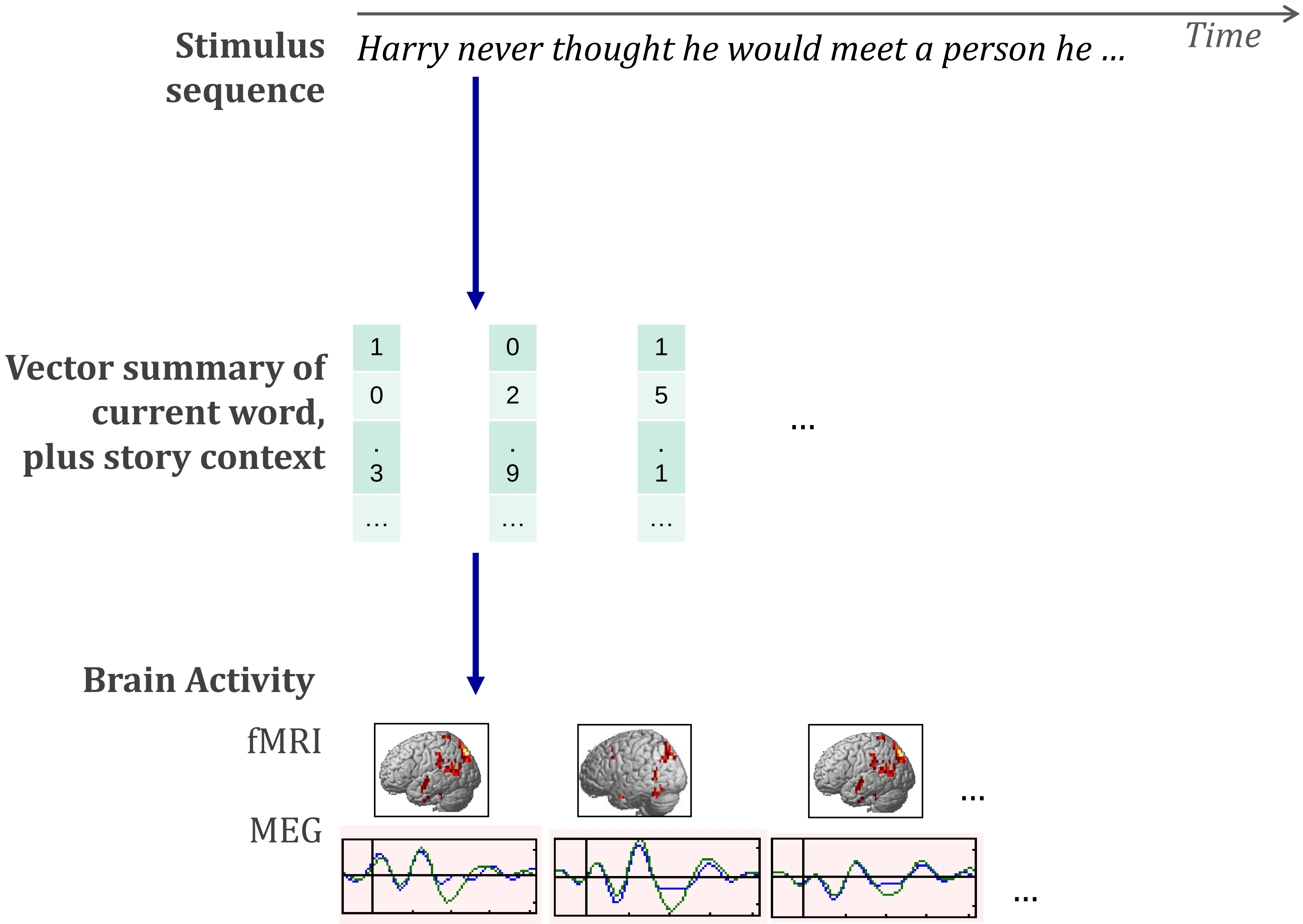
Reading Harry Potter, one word at a time…



500ms per word


Harry had never believed he would meet a boy he hated more
1TR = 2s
time

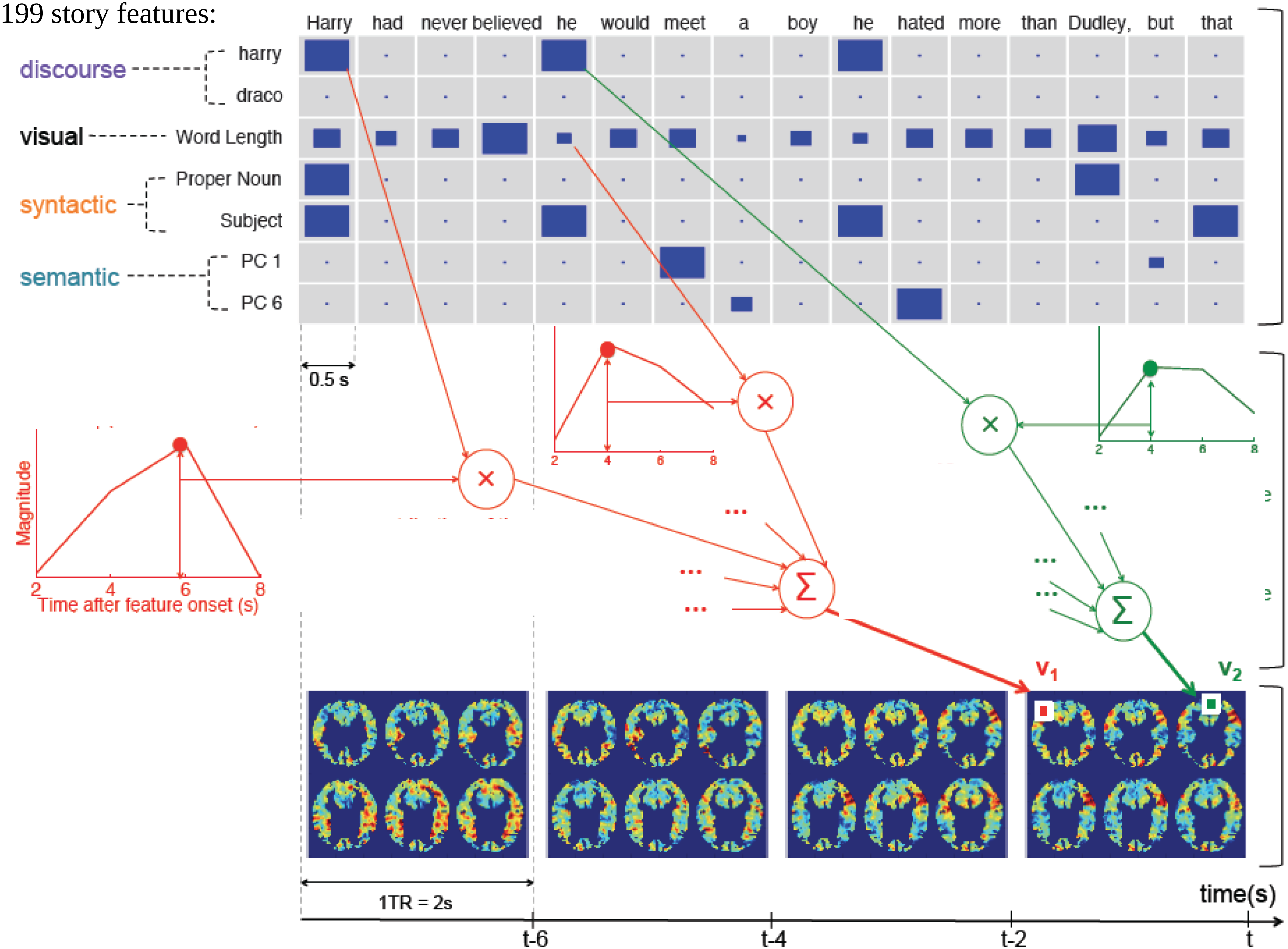
General Framework



Semantics	1...100	Syntax	150 Sentence Length
Speech	101 speak - sticky	-parts of speech	151 ,
	102 speak - puntual		152 .
Motion	103 fly - sticky		153 :
	104 manipulate - sticky		154 Coordinating conjunction
	105 move - sticky		155 Cardinal number
	106 collide physically - sticky		156 Determiner
	107 fly - punctual		157 Preposition / sub. conjunctio
	108 manipulate - punctual		158 Adjective
	109 move - puntual		159 Modal
Emotion	110 annoyed - puntual		160 Noun, singular or mass
	111 commanding - puntual		161 Noun, plural
	112 dislike - puntual		162 Proper noun, singular
	113 fear - puntual		163 Proper noun, plural
	114 like - punctual		164 Personal pronoun
	115 nervousness - puntual		165 Possessive pronoun
	116 questioning - punctual		166 Adverb
	117 wonder - punctual		167 Particle
	118 annoyed - sticky		168 to
	119 commanding - sticky		169 Interjection
	120 cynical - sticky		170 Verb, base form
	121 dislike - sticky		171 Verb, past tense
	122 fear - sticky		172 Verb, gerung or present part.
	123 mental hurting - sticky		173 Verb, past part.
	124 physical hurting - sticky		174 Verb, non-3rd person sing. pre
	125 like - sticky		175 Verb, 3rd person sing. present
	126 nervoussness - sticky		176 Wh-determiner
	127 pleading - sticky		177 Wh-pronoun
	128 praising - sticky		178 Wh-adverb

	128 praising - sticky		178 Wh-adverb
	129 pride - sticky	-dependency roles	179 Unclassified adverbial
	130 questioning - sticky		180 Modifier or adjective or adverb
	131 relief - sticky		181 Coordination
	132 wonder - sticky		182 Coordination
Verbs	133 be		183 Other dependent (default label
	134 hear		184 Indirect object
	135 know		185 Modifier of noun
	136 see		186 Object
	137 tell		187 Punctuation
Characters	138 Draco		188 Modifier of preposition
	139 Filch		189 Predicative complement
	140 Harry		190 Parenthetical
	141 Hermione		191 Particle
	142 Mrs. Hooch		192 Root
	143 Mrs. McGonagall		193 Subject
	144 Neville		194 Verb chain
	145 Peeves		195 Modifier of verb
	146 Ron		
	147 Wood		
Visual	148 Average Word Length		
	149 Variance of Word Length		

199 story features:



Test the model on new text passages

story passages
(4 TRs = 16 words)

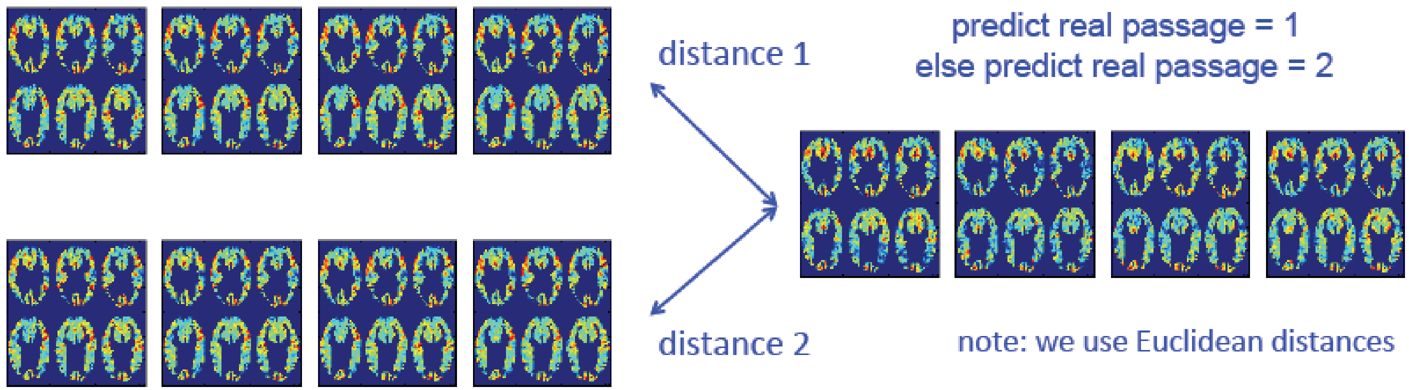
predicted segment of fMRI activity

real held out 4 TRs fMRI segment

if distance 1 < distance 2
predict real passage = 1
else predict real passage = 2

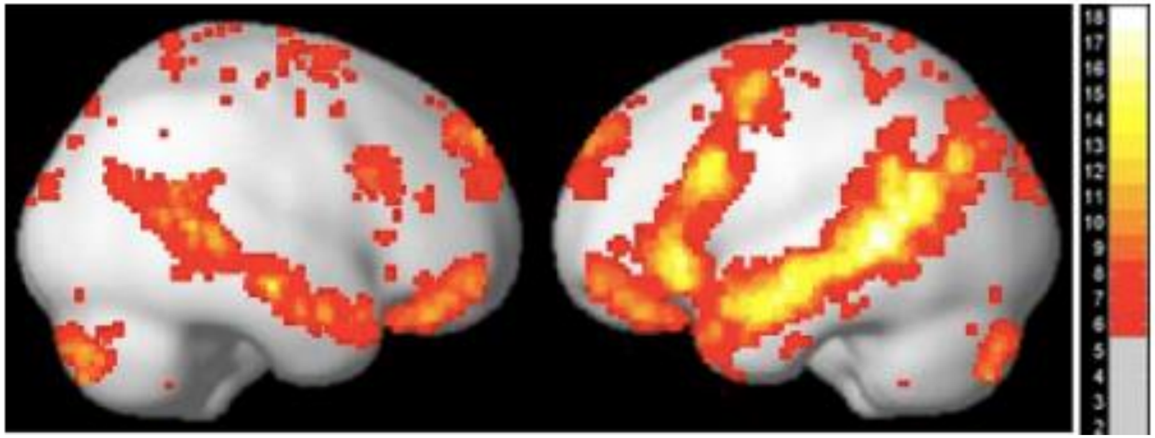
1 … They were half hoping for a reason to fight Malfoy, but Professor McGonagall, who could spot …

2 … Harry had heard Fred and George Weasley complain about the school brooms, saying that some of…



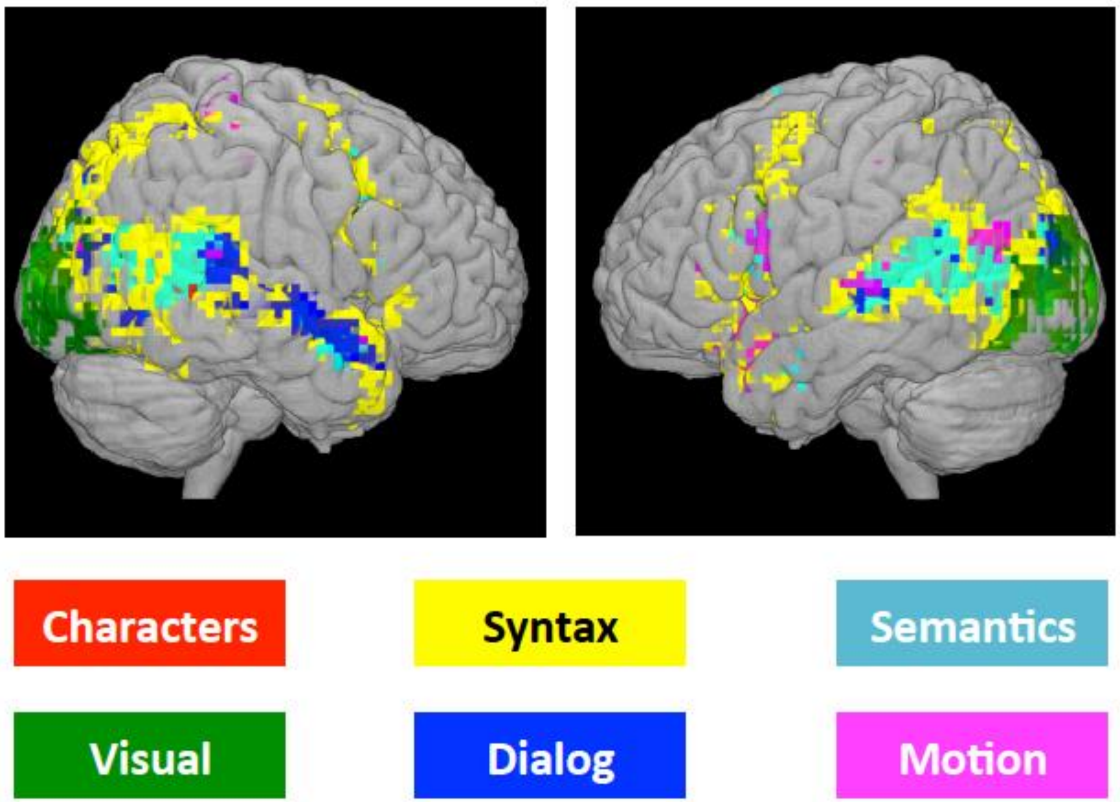
note: we use Euclidean distances

accuracy: 75%



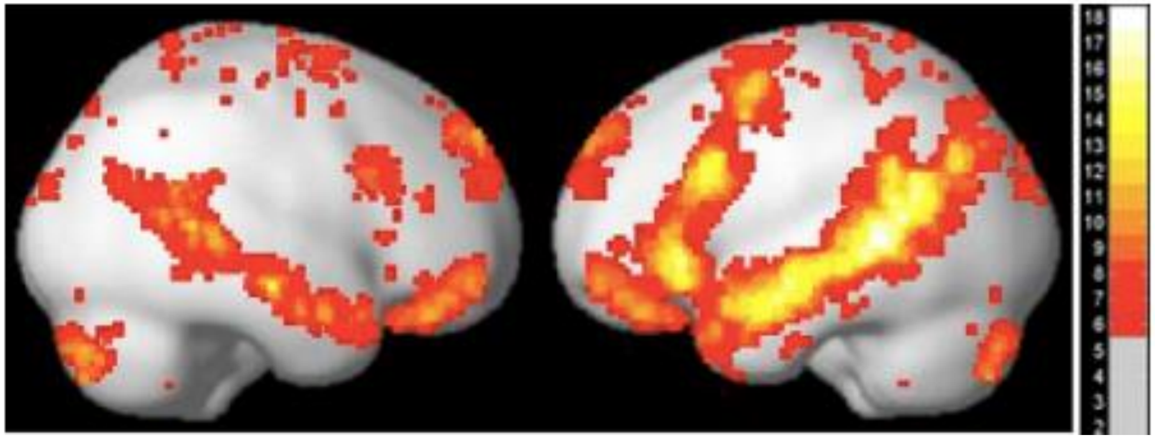
previous work:
where does reading generate activity?

Fedorenko et al.,
Neuropsychologia 2012



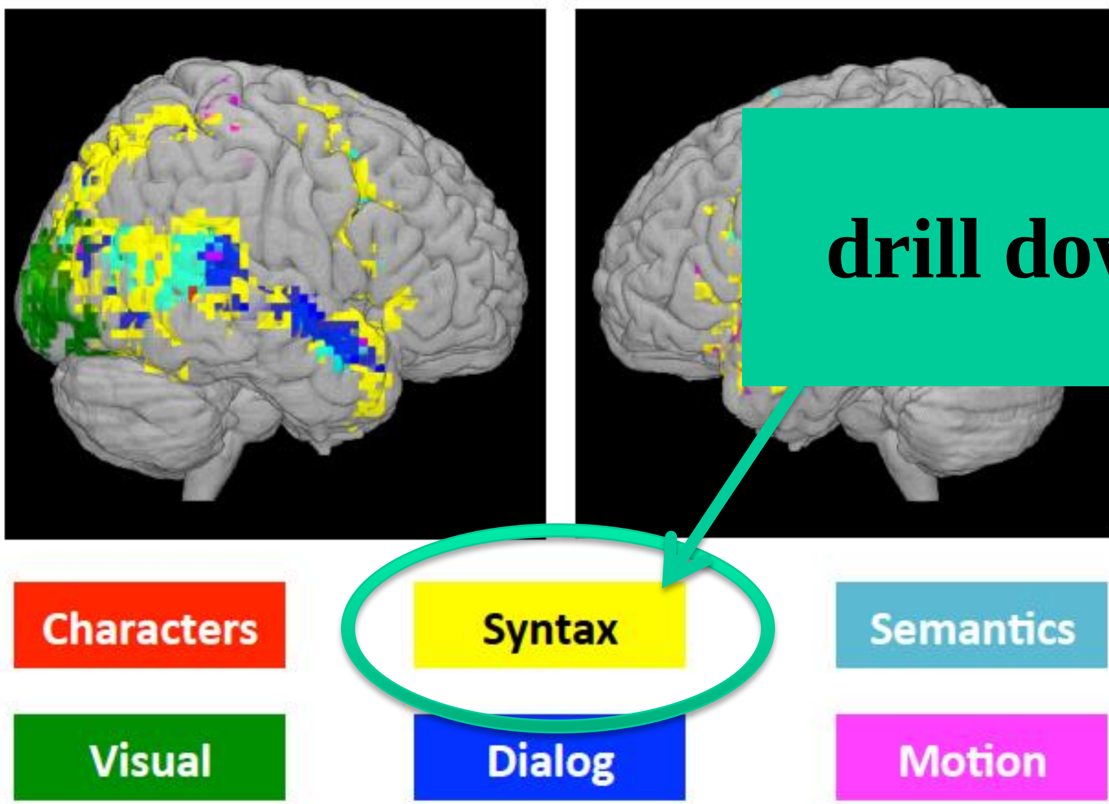
our work:
*where is **story information encoded**?*

Wehbe et al.,
PLoS One 2014



previous work:
where does reading generate activity?

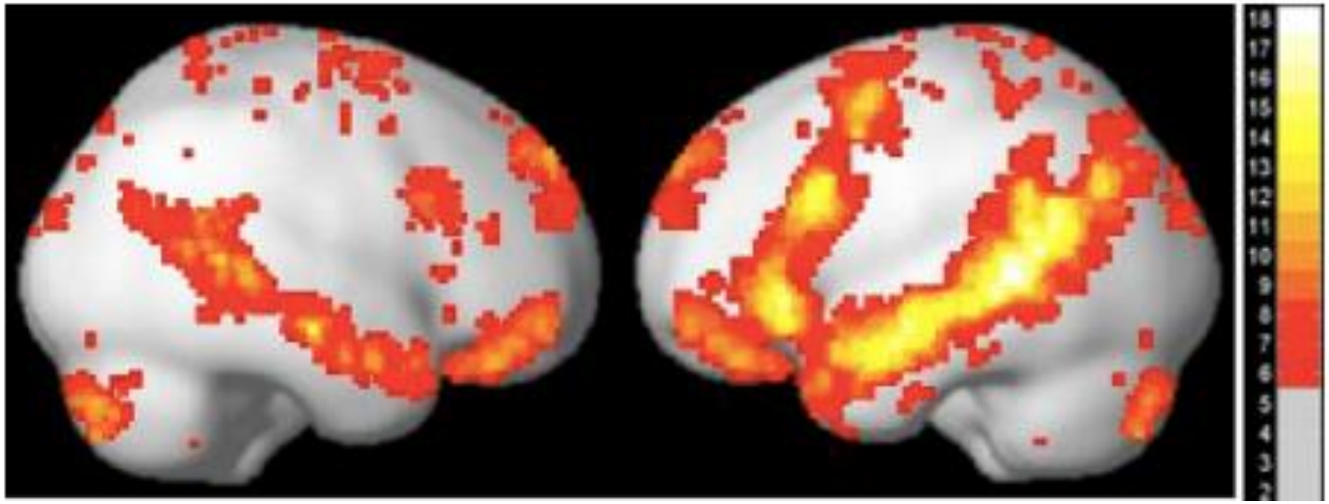
Fedorenko et al.,
Neuropsychologia 2012



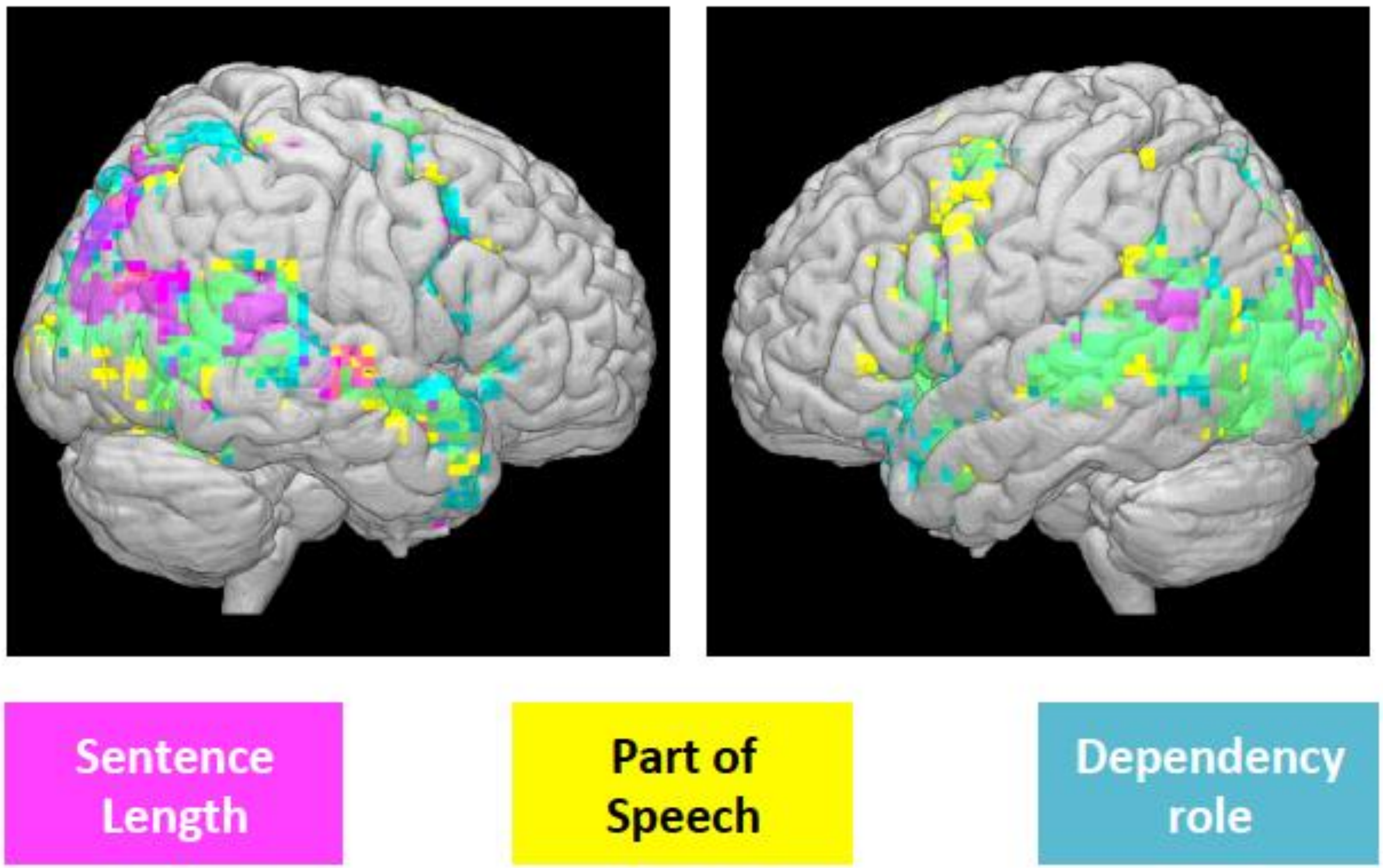
drill down

ur work:
here is ***story*** ***formation*** ***encoded****?*

Wehbe et al.,
PLoS One 2014



[Fedorenko et al. 2012]



[Wehbe et al., 2014]

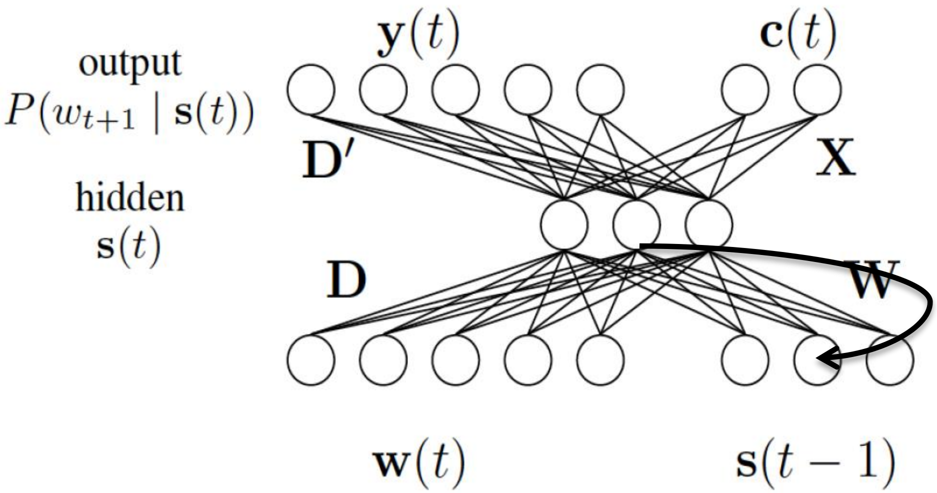
[Wehbe et al., EMNLP 2014]

Q: Can we observe neural encoding of story content?

Modeling context: Recurrent Network

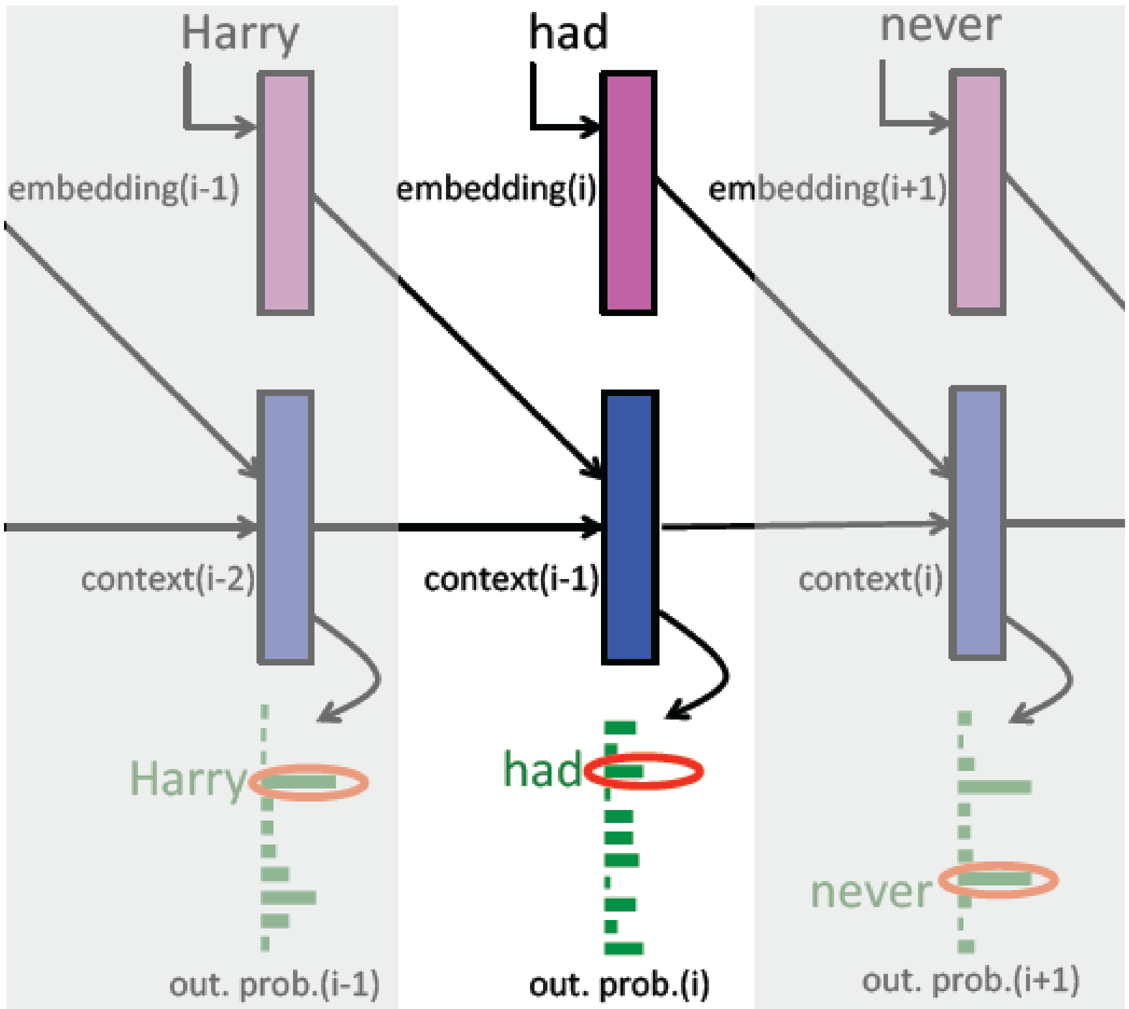
[Wehbe et al., *EMNLP14*]

1. MEG subjects read chapter of Harry Potter
2. Train recurrent network language model on 67M words of Harry Potter fan fiction



3. Use learned representation of context s(t-1), current word w(t), current word probability y(t),c(t), to decode* current word from 100 msec windows of neural activity

* concatenate 20 random words per example, 2x2



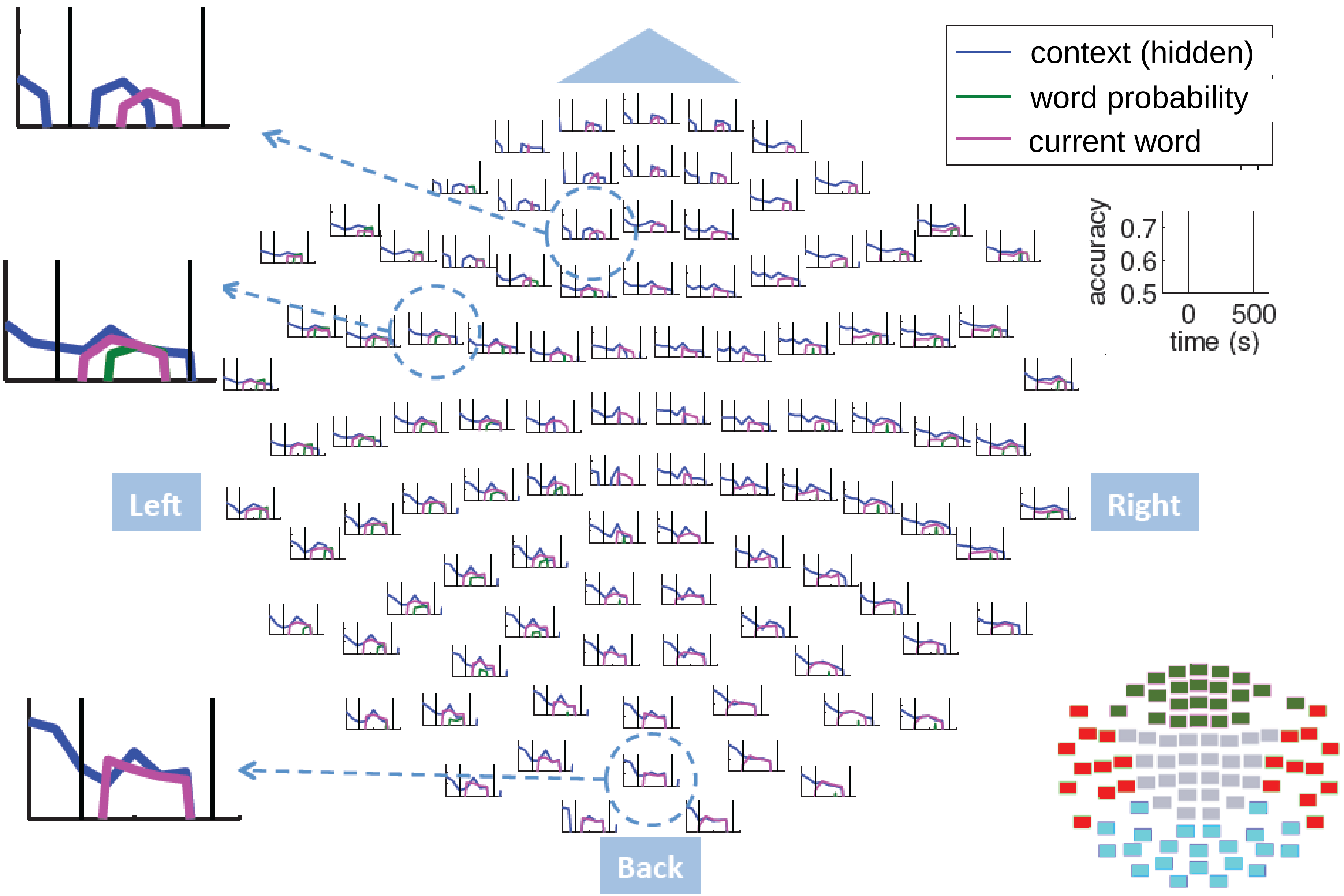
MEG classification accuracy:

- 0.80 current word (embedding)
- 0.93 context (recurrent hidden)
- 0.60 Predicted probability of current word

* concatenate MEG for 20 random words per example, 2x2

Results

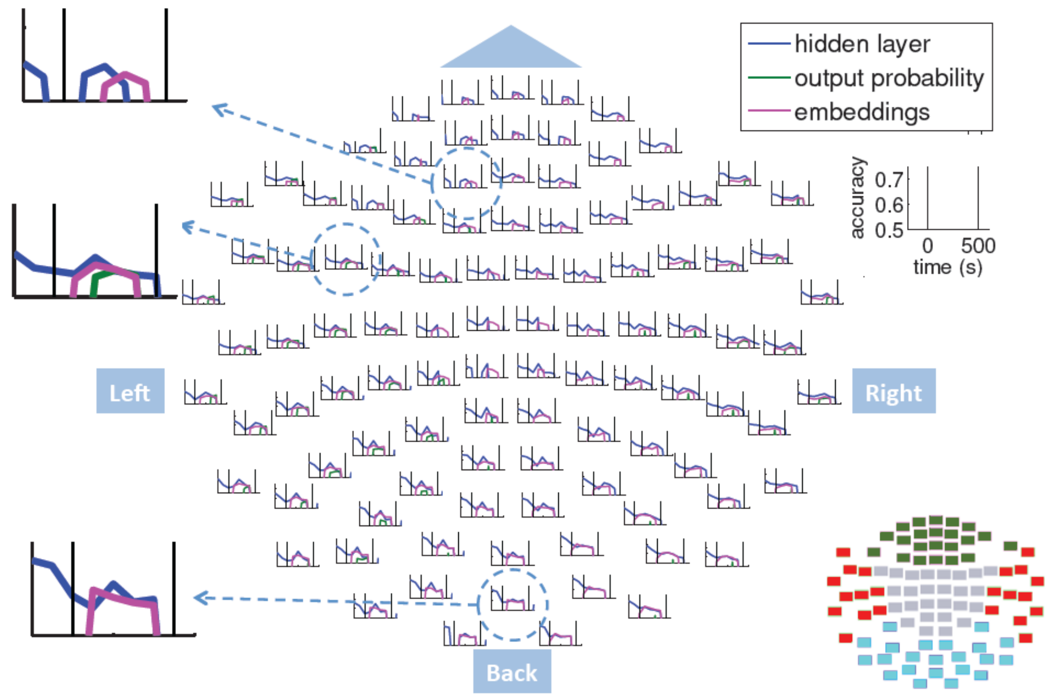
[Wehbe et al., *EMNLP14*]



Implications

[Wehbe et al., *EMNLP14*]

- Much activity encodes **context**
  - decoding based on context > based on current word
- **context** most salient 200-250 msec post word onset
- current word **probability** most salient in left hemisphere, at 200-400 msec



Lessons

Neuroscience:

- Neural code for word meanings distributed across the brain
- Your neural code and mine are very similar
- Neural code is built up from more primitive semantic features
- Neural code evolves over 400 msec after word onset
- During story reading, diverse information encoded brain-wide

Lessons

Neuroscience:

- Neural code for word meanings distributed across the brain
- Your neural code and mine are very similar
- Neural code is built up from more primitive semantic features
- Neural code evolves over 400 msec after word onset
- During story reading, diverse information encoded brain-wide

Methodology

- Key role of machine learning
  - classifiers, regression, latent representation discovery, language modeling, …
- Big opportunity 1: jointly analyze data from many experiments
- Big opportunity 2: build a program that understands sentences, and as a result predicts neural activity

thank you!