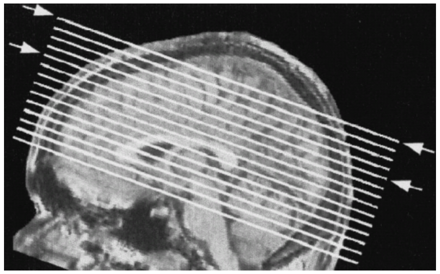
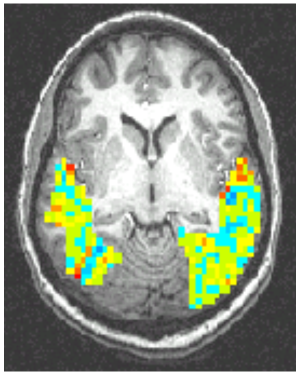
How Does the Brain Represent Word Meanings?



Tom M. Mitchell

Machine Learning Department
Carnegie Mellon University

September 2010



Neurosemantics Research Team

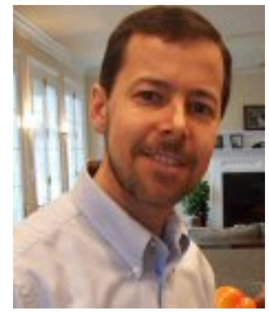
Tom Mitchell

Marcel Just

Postdoctoral Fellow

Rob Mason

Research Scientists



Dean Pommerleau

Vladimir Cherkassky

PhD Students

Gustavo Sudre

Kai Min Chang

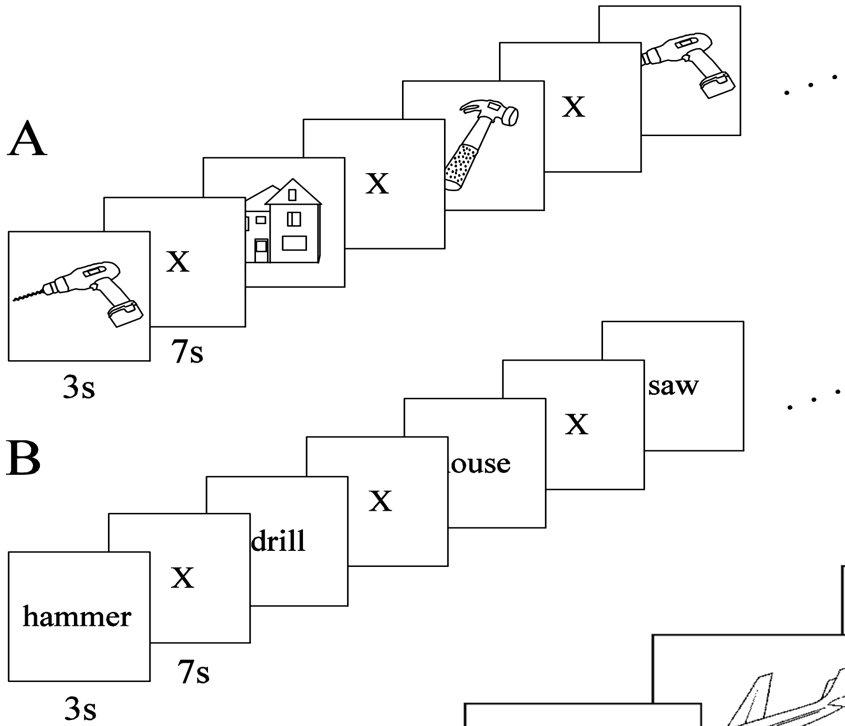
Leila Wehbe

Indra Rustandi

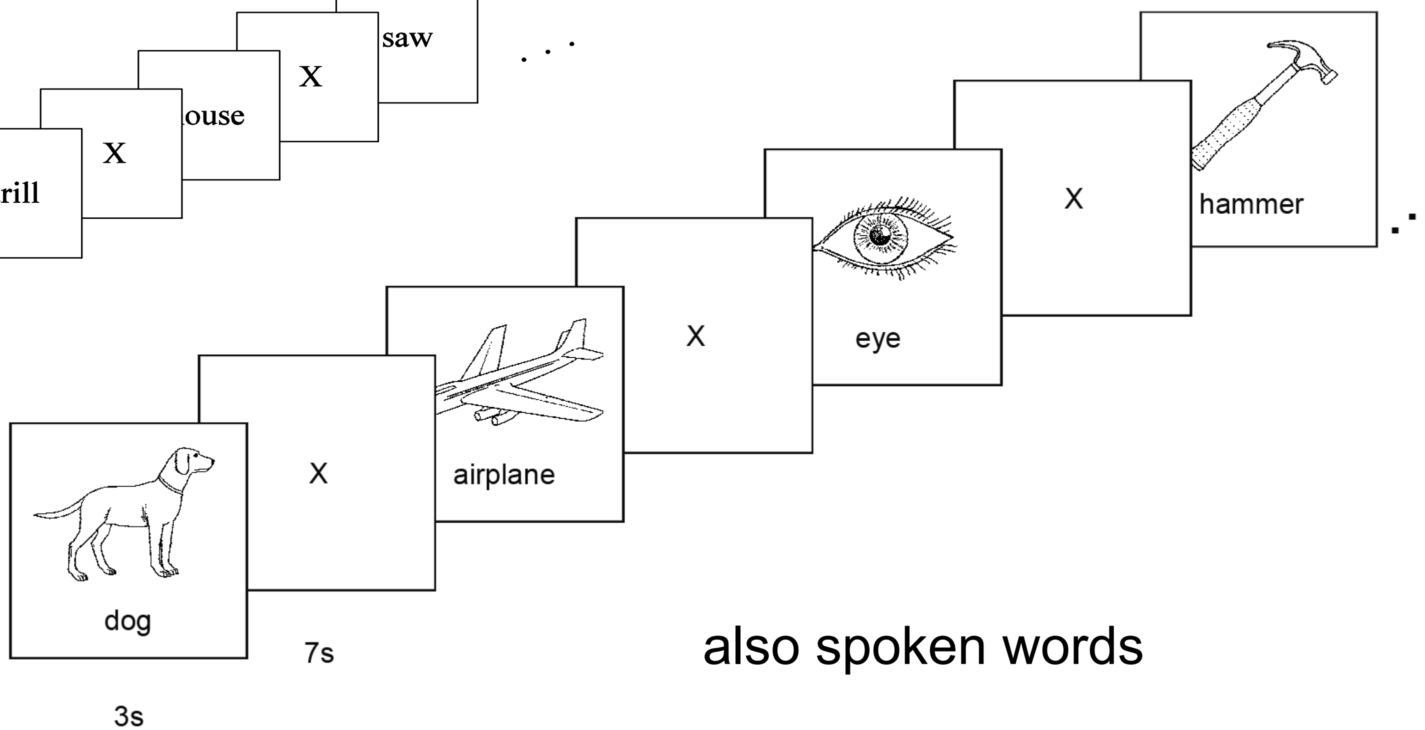
Mark Palatucci

Alona Fyshe

Typical stimuli



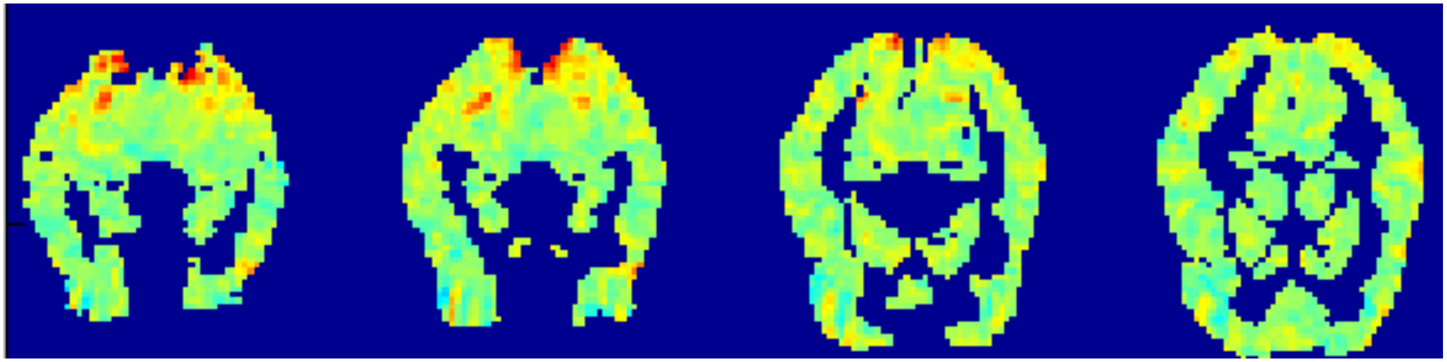
Each stimulus repeated several times

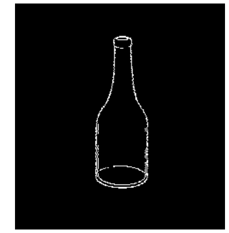


also spoken words

Functional MRI

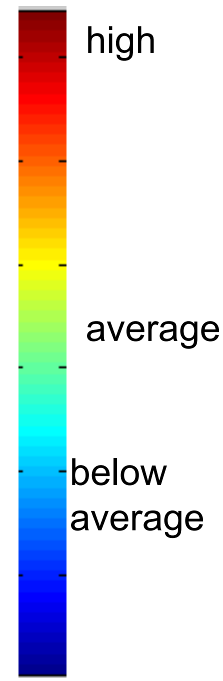
fMRI activation for "bottle":



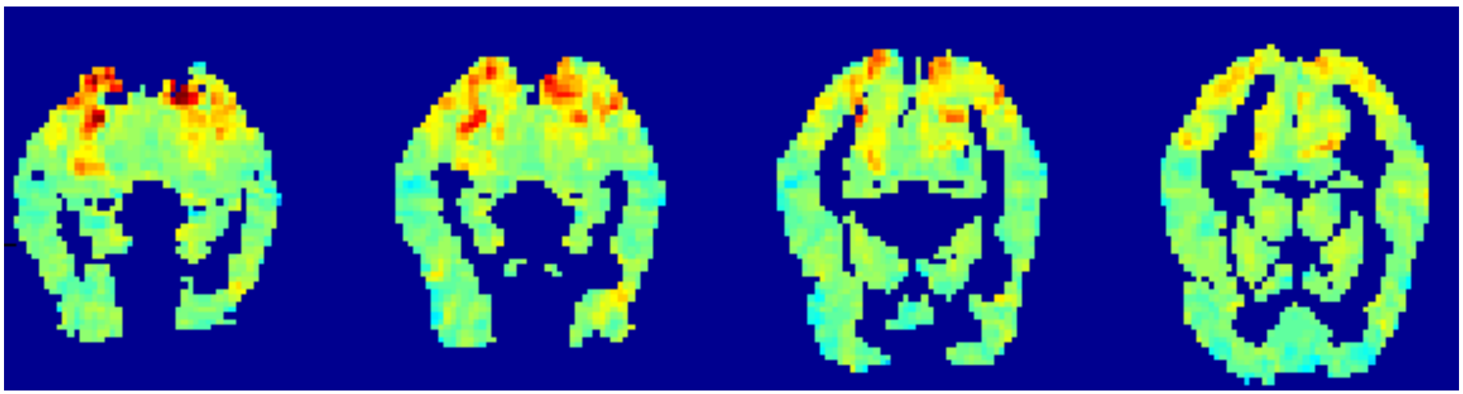


bottle

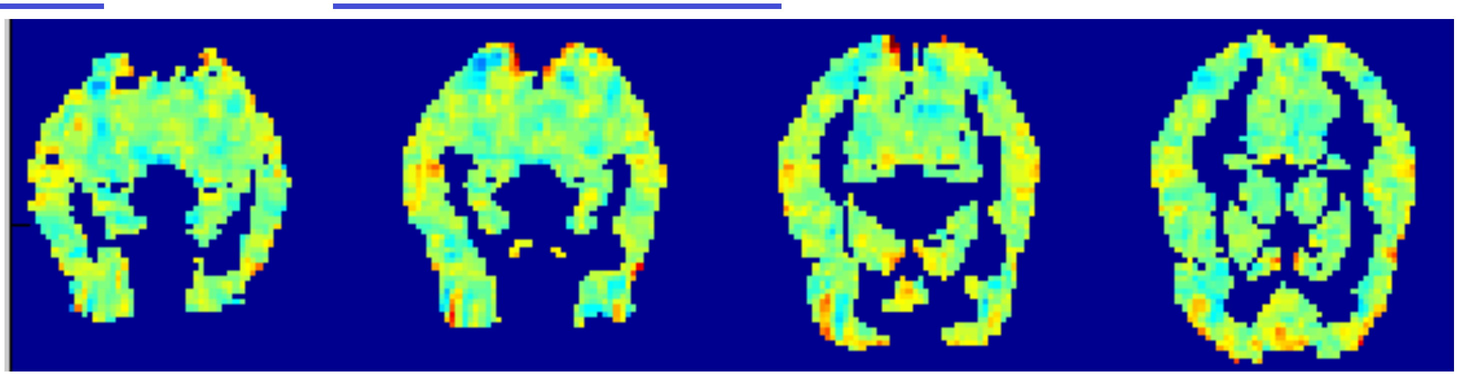
fMRI activation



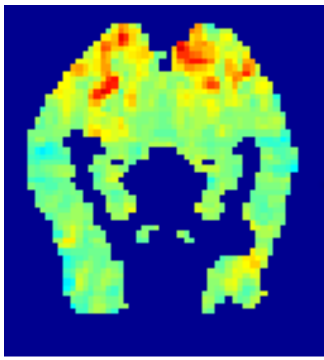
Mean activation averaged over 60 different stimuli:



"bottle" minus mean activation:



Q1: Can one distinguish which word you're thinking about based on fMRI?



Trained Classifier
(Bayes classifier, logistic regression, SVM, kNN, …)

~~Hammer~~
or
House

(classifier as virtual sensor of mental state)

Training Classifiers over fMRI sequences

- Train the classifier function

  Mean(fMRI(t+4), ...,fMRI(t+7)) → WordCategory

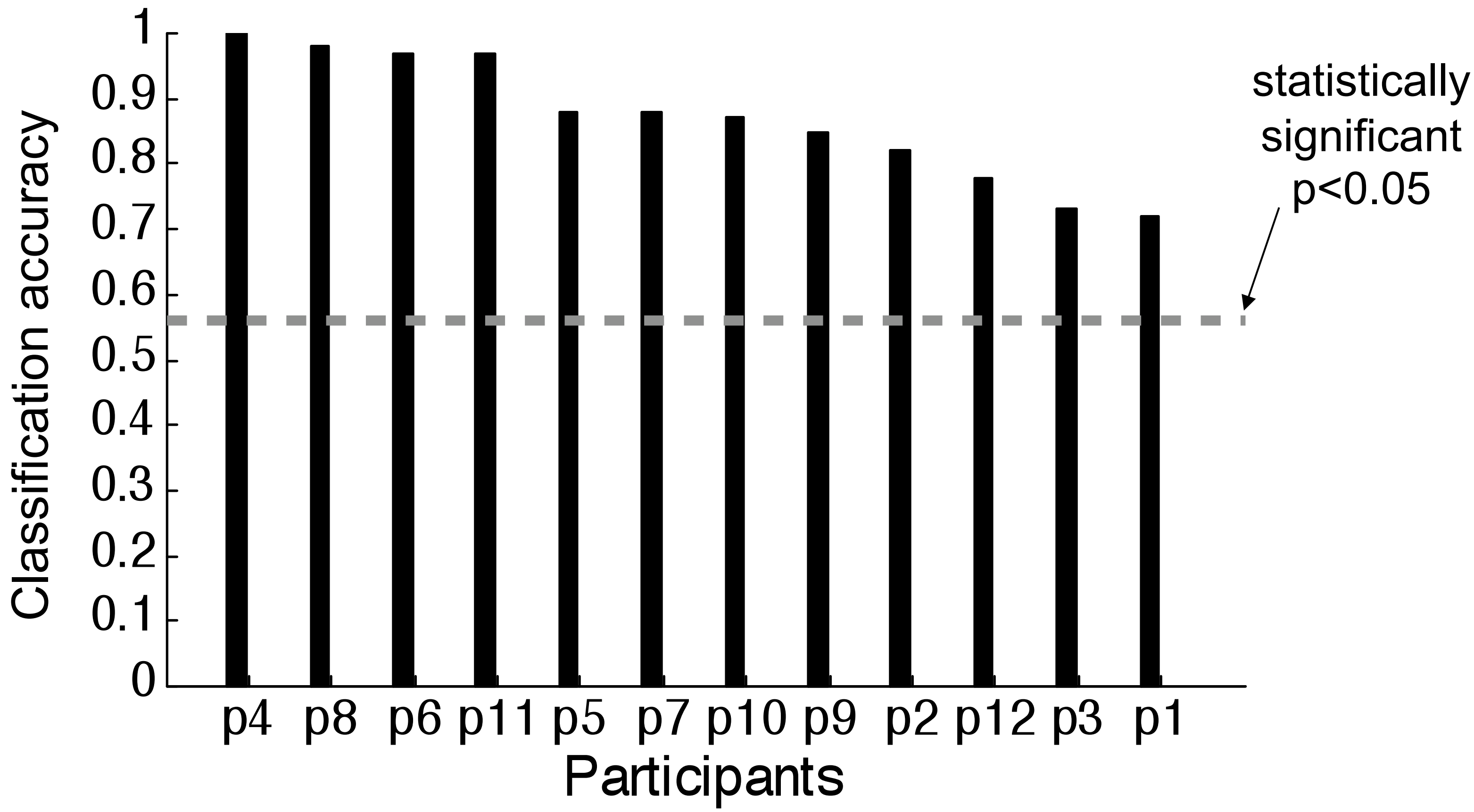
- Preprocessing:
  - Adjust for head motion
  - Convert each image x to standard normal image

$$x(i) \leftarrow \frac{x(i) - \mu_x}{\sigma_x}$$

- Learning algorithms tried:
  - kNN (spatial correlation)
  - SVM
  - SVDM
  - Gaussian Naïve Bayes
  - Regularized Logistic regression ← current favorite
  - …

- Feature selection methods tried:
  - Logistic regression weights, voxel stability, activity relative to fixation, regularization (L1, L2), ...

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



Q2: Are neural representations similar across people?

Can we train on one group of people, decode for new person?


Mind Reading
produced by
Shari Finkelstein

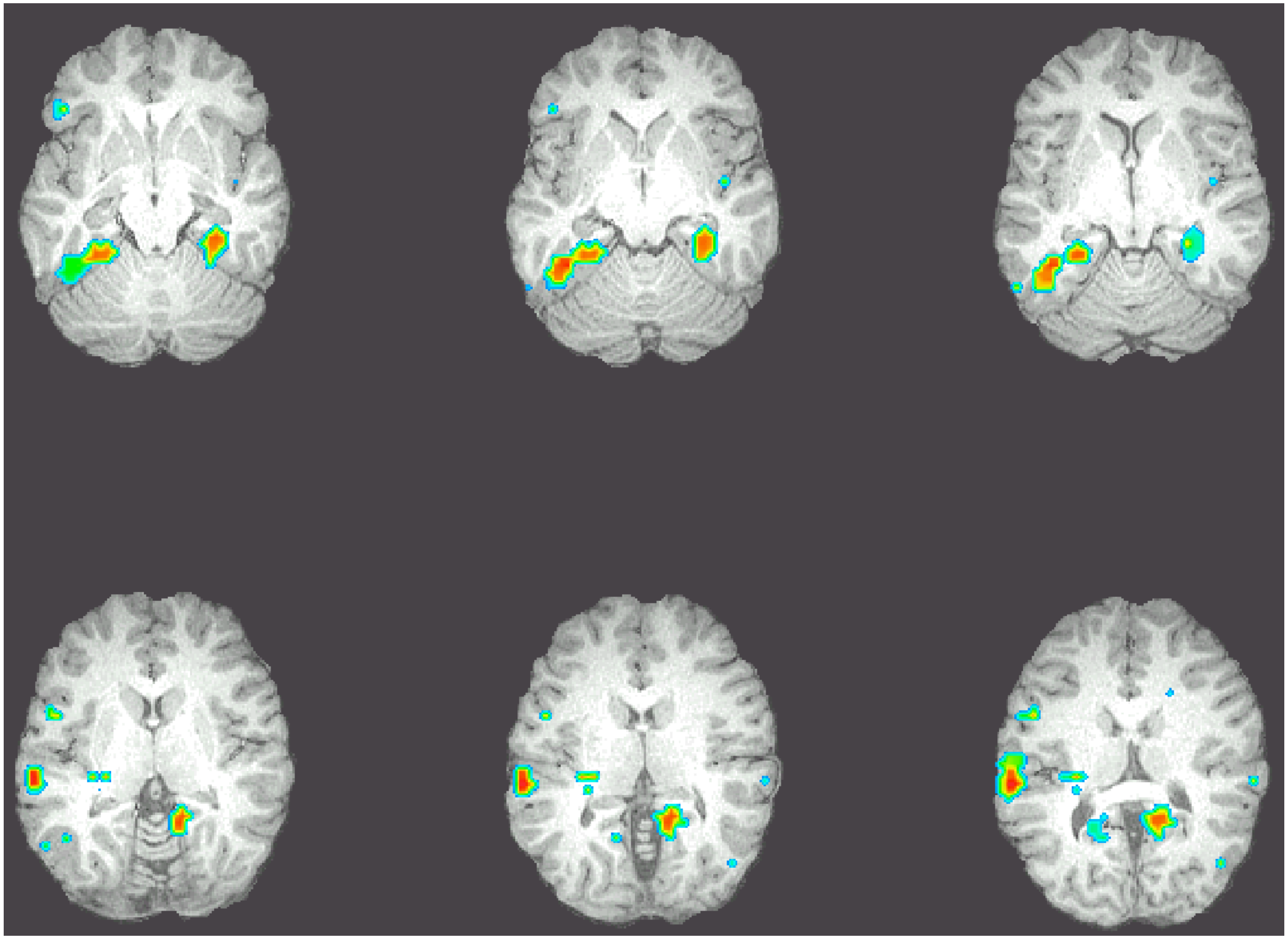
Local classifiers show where information is encoded

[F. Pereira]
spotlight classifiers [N. Kriegeskorte]

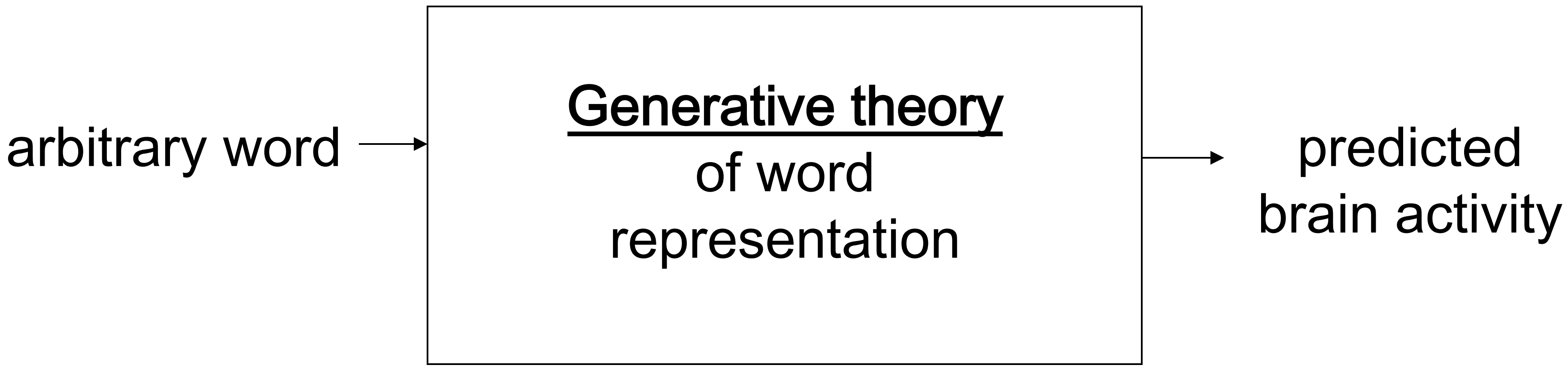
"tools" vs
"buildings"

Accuracies of
cubical 27-voxel
classifiers
centered at
each voxel

[0.7-0.8]

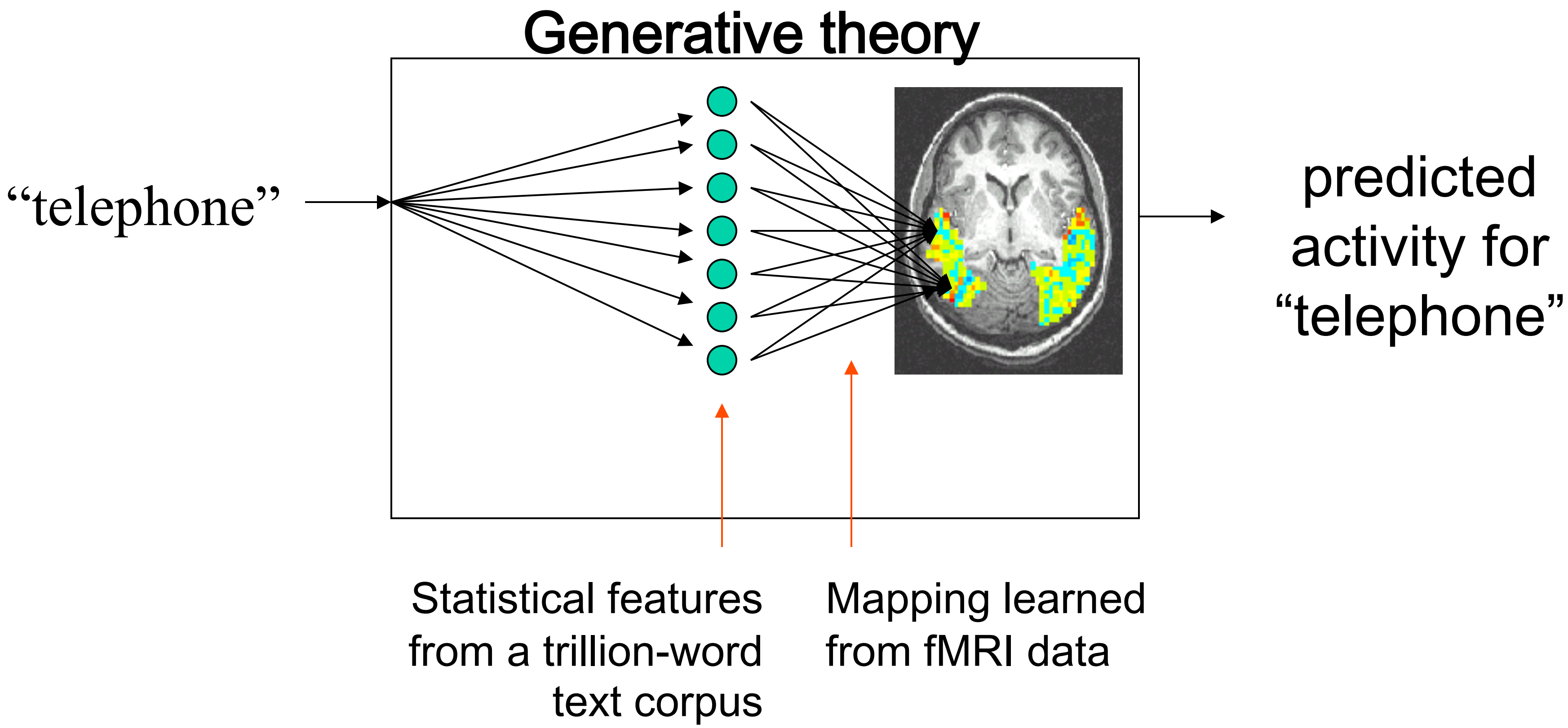


Q3: Can we discover underlying principles of neural encodings?



Idea: Predict neural activity from corpus statistics of stimulus word

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



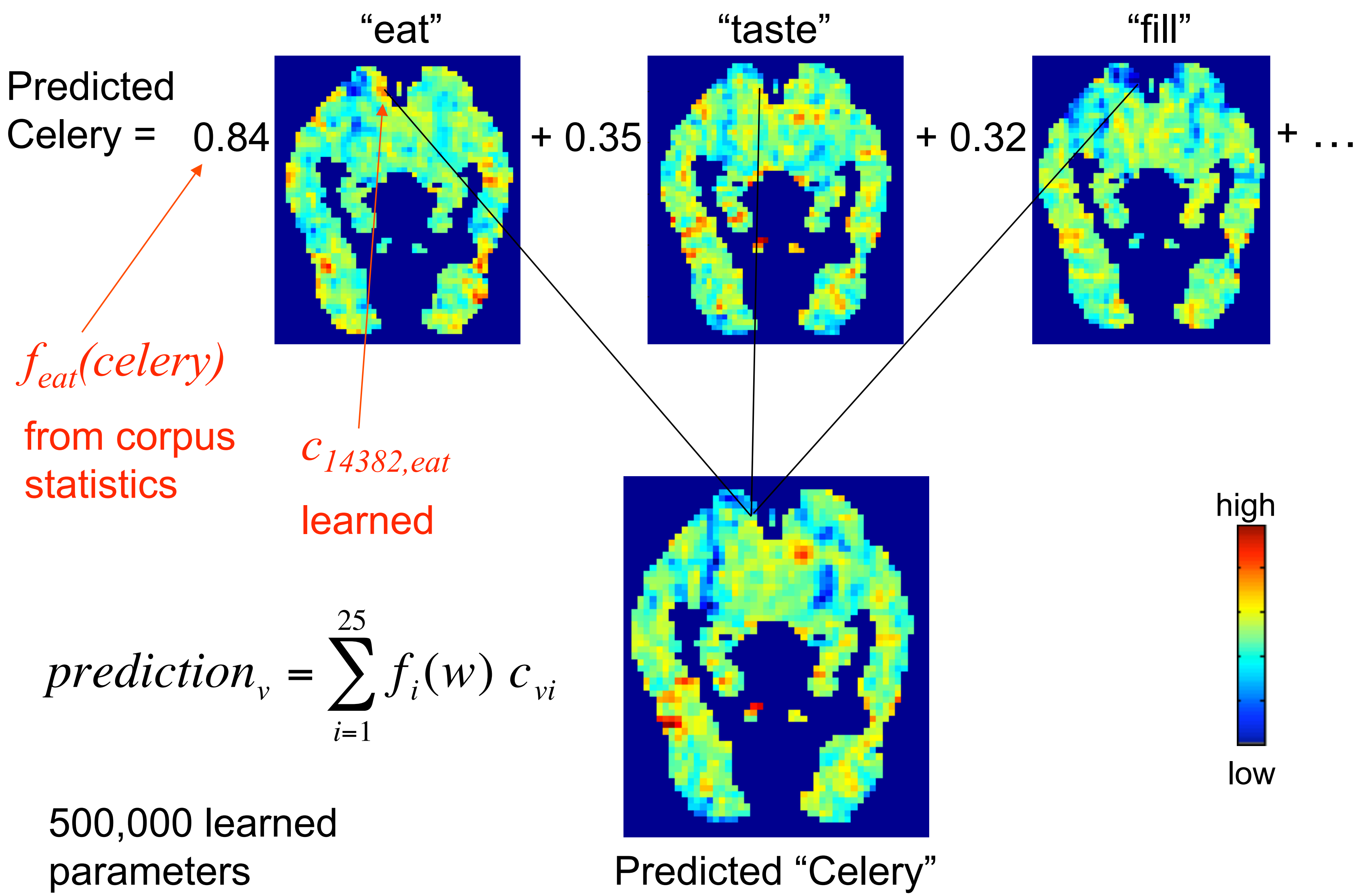
Semantic feature values: "**celery**"

0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
…
…
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

Semantic feature values: "**airplane**"

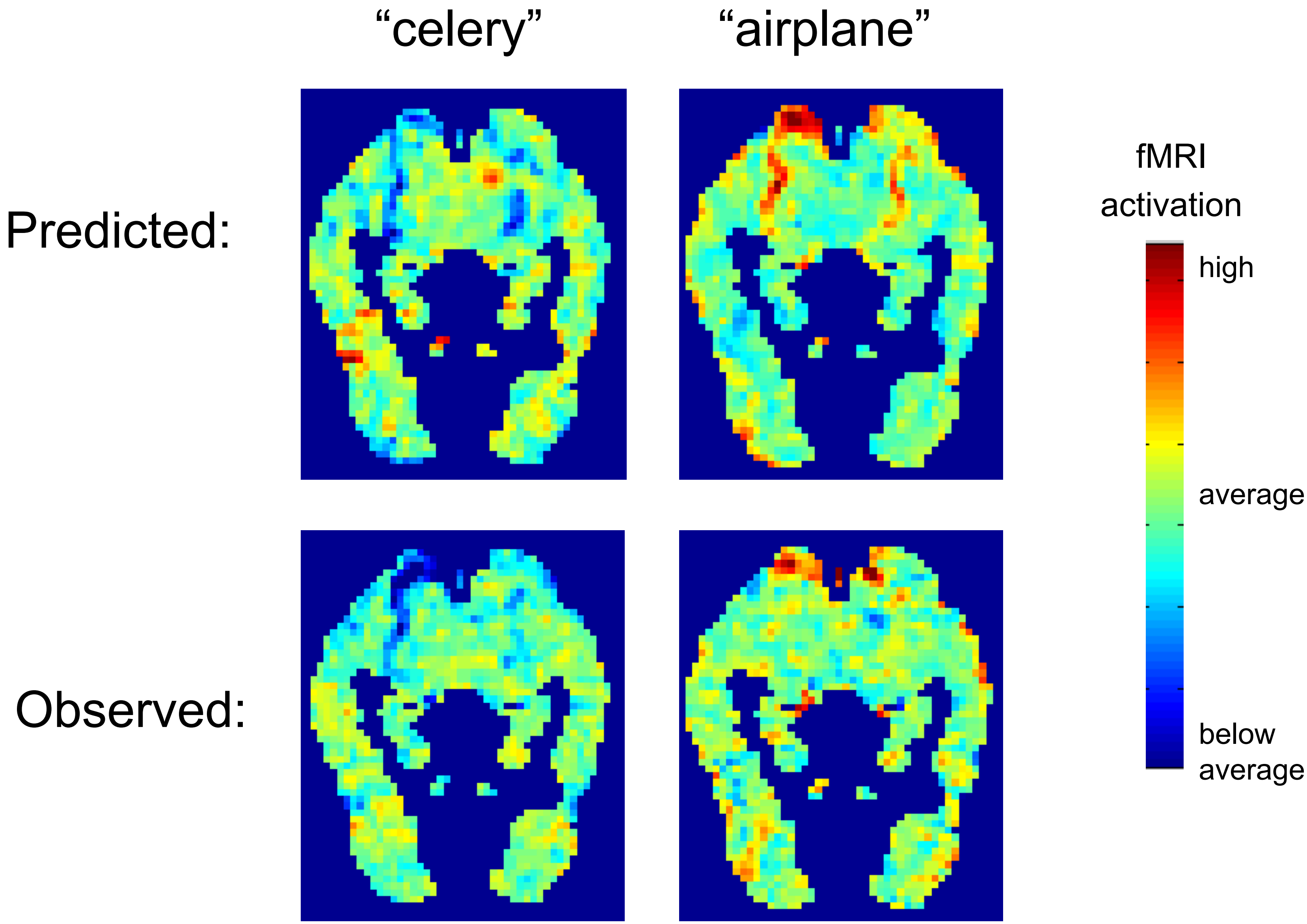
0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
…
…
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate

Predicted Activation is Sum of Feature Contributions



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

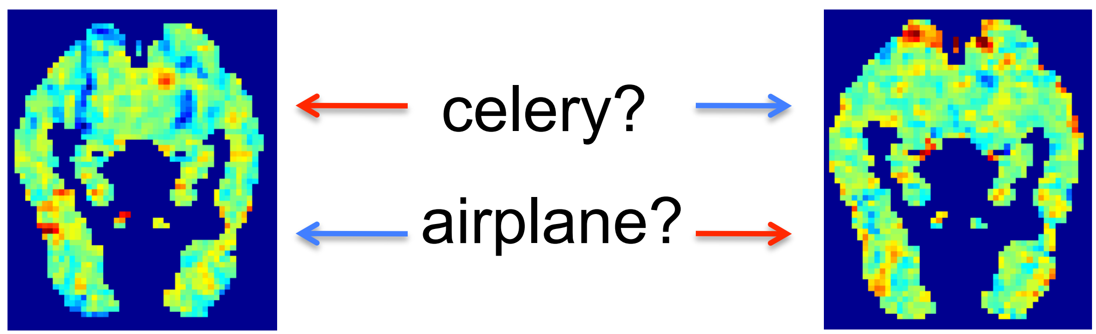
500,000 learned parameters



Predicted and observed fMRI images for "celery" and "airplane" after training on 58 other words.

Evaluating the Computational Model

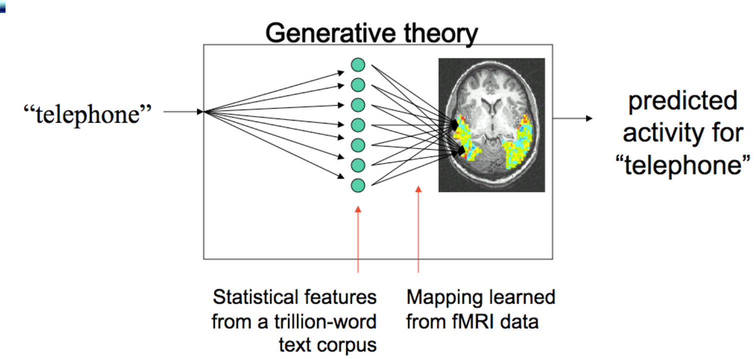
- Train it using 58 of the 60 word stimuli
- Apply it to predict fMRI images for other 2 words
- Test: show it the observed images for the 2 held-out, and make it predict which is which

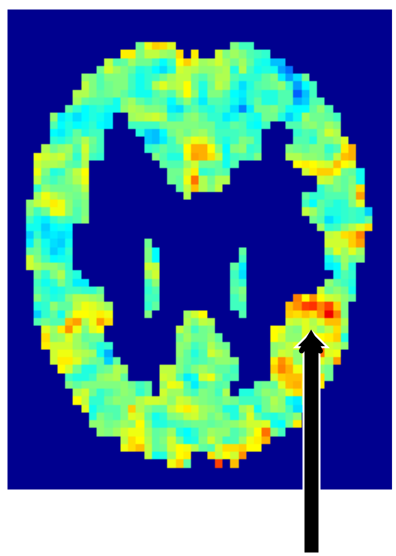
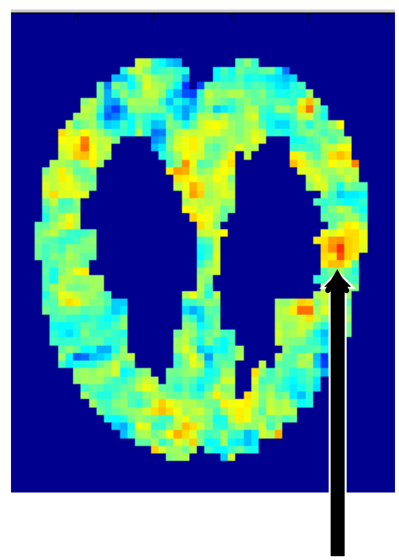
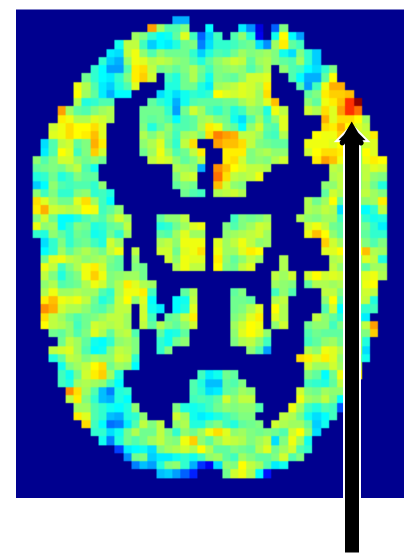


1770 test pairs in leave-2-out:

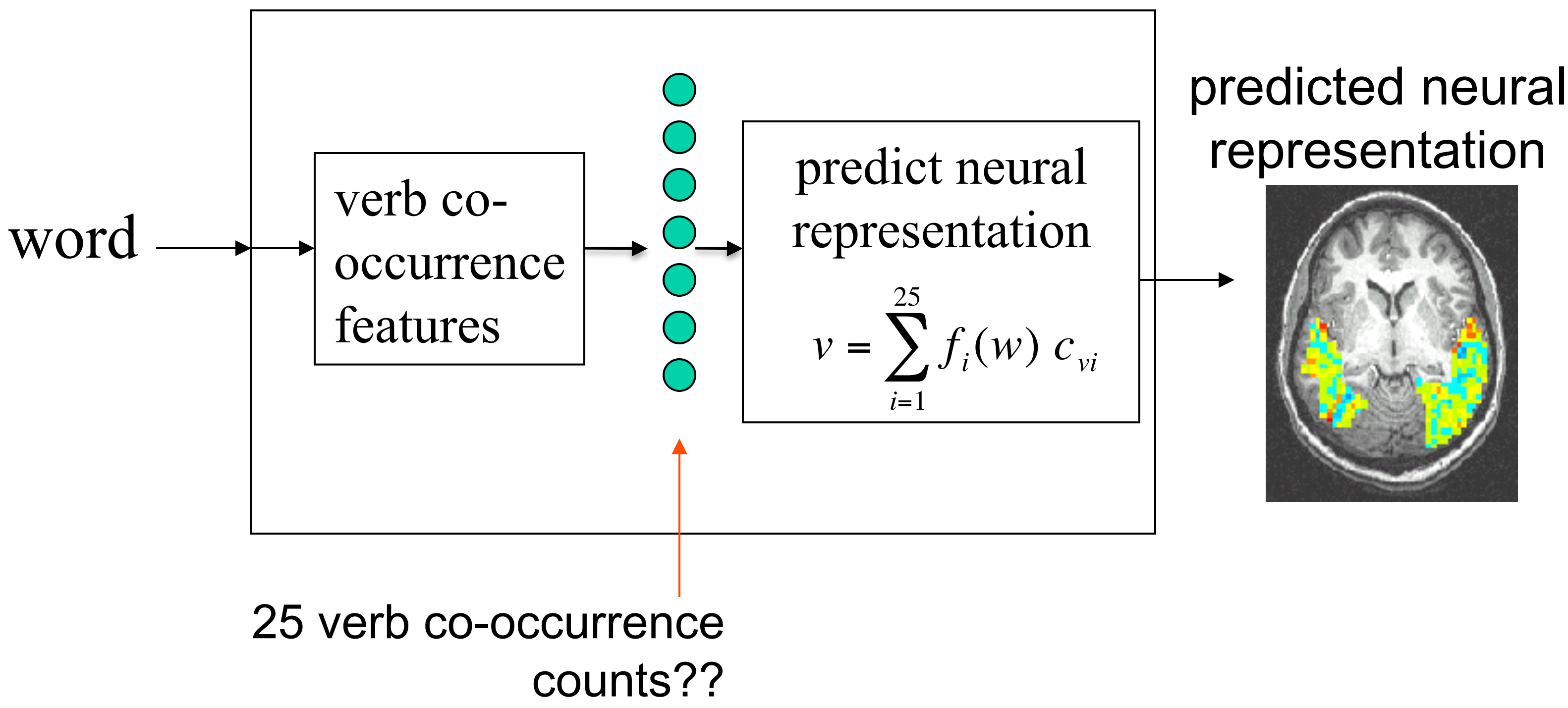
- Random guessing → 0.50 accuracy
- Accuracy above 0.61 is significant ($p<0.05$)

Mean accuracy over 9 subjects: 0.79



	Eat	Push	Run
Participant P1			
	"Gustatory cortex"	"sensory motor"	"Biological motion"
	Pars opercularis (z=24mm)	Postcentral gyrus (z=30mm)	Superior temporal sulcus (posterior) (z=12mm)

Q4: What are the actual semantic primitives from which neural encodings are composed?



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

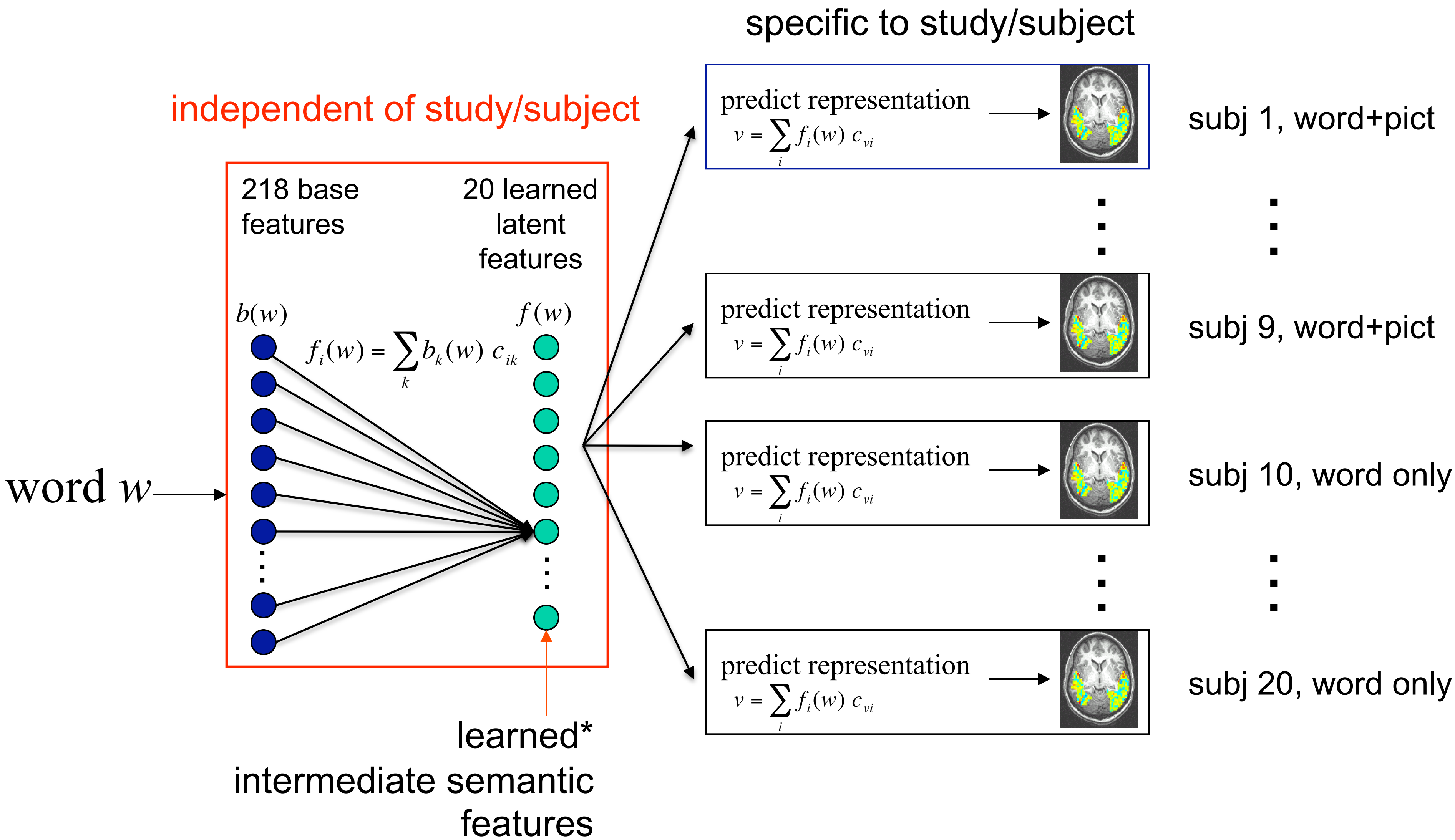
PREDEFINED corpus features	Mean Acc.
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
218 features collected using *Mechanical Turk**	**.83**
20 features discovered from the data**	**.87**

* developed by Dean Pommerleau

** developed by Indra Rustandi

Discovering shared semantic basis

[Rustandi et al., 2009]



* trained using Canonical Correlation Analysis

Multi-study (WP+WO) Multi-subject (9+11) CCA
Top Stimulus Words

	component 1	component 2	component 3	component 4
most positive stimuli	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 me?

Additional Directions

- Model for abstract words (love, justice, anxiety,…)
  – preliminary: accuracies similar to those for concrete nouns

- Model phrases (“firm tomato”)
  – [Chang et al., ACL2009]: composing corpus statistics for <adjective> and <noun> predicts fMRI for <adjective noun>

- MEG imaging (1 msec time resolution)
  – preliminary results: can train classifiers to detect both <u>where</u> and <u>when</u> neural activity codes word meanings, and stimulus percepts

- ML algorithms that build cumulative models from many (100’s of) data sets

Where Next?

- What will a “theory” of the brain (or the cell) look like?

- Set of architectural organizing principles,
- and a detailed computational model that follows them

- How will we learn it?

- Current approaches are data-starved
- Need algorithms that learn cumulatively from
  - many experiments
  - priors gleaned from research literature
  - priors that express researcher’s hypotheses
  - optimal planning of next experiment

thank you!