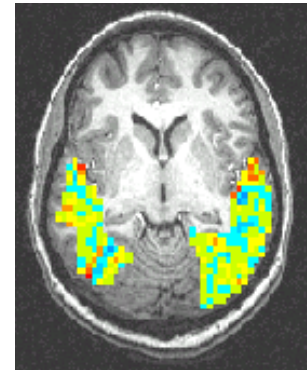
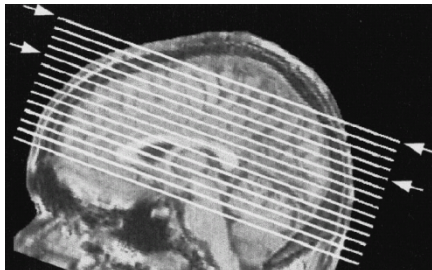


Brains, Meaning and Corpus Statistics

Tom M. Mitchell
and collaborators

Machine Learning Department
Carnegie Mellon University

May, 2009



based in part on:

[“Predicting Human Brain Activity Associated with the Meanings of Nouns,”](#)
Mitchell, Shinkareva, Carlson, Chang, Malave, Mason, & Just, *Science*, 2008.

Neurosemantics Research Team

Postdoctoral Fellows



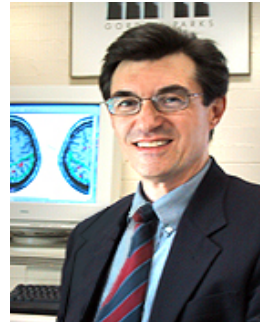
Svetlana Shinkareva



Rob Mason



Tom Mitchell



Marcel Just

Researchers

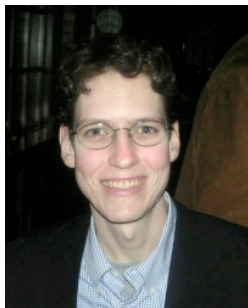


Dean Pommerleau



Vladimir Cherkassky

PhD Students



Andy Carlson



Kai Min Chang



Rebecca Hutchinson



Mark Palatucci



Indra Rustandi



Francisco Pereira

Neuroscience Research Questions

- Can we observe differences in neural activity as people think about different concepts?
- Is the neural activity that represents concepts localized or distributed?
- Are neural representations similar across people?
- Can we discover underlying principles of neural representations? (e.g., are representations built up from more primitive components?)

Functional MRI



functional Magnetic Resonance Imaging (fMRI)

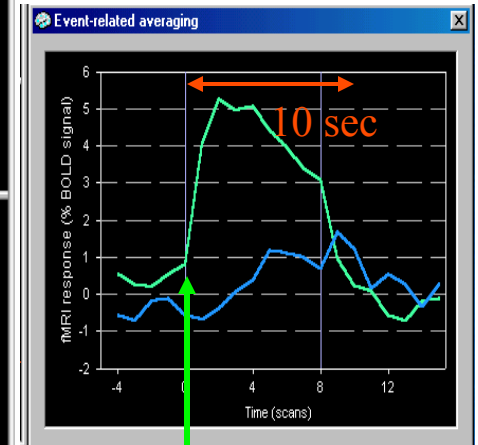
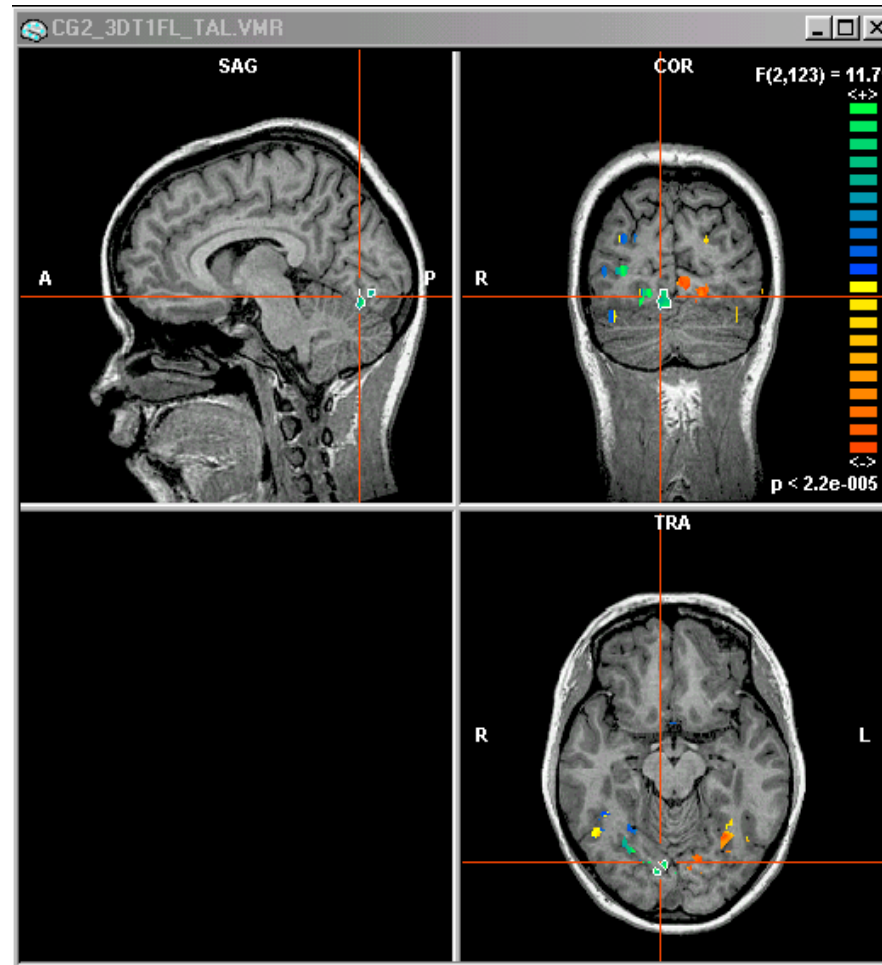
~1 mm resolution

~1 image per sec.

20,000 voxels/image

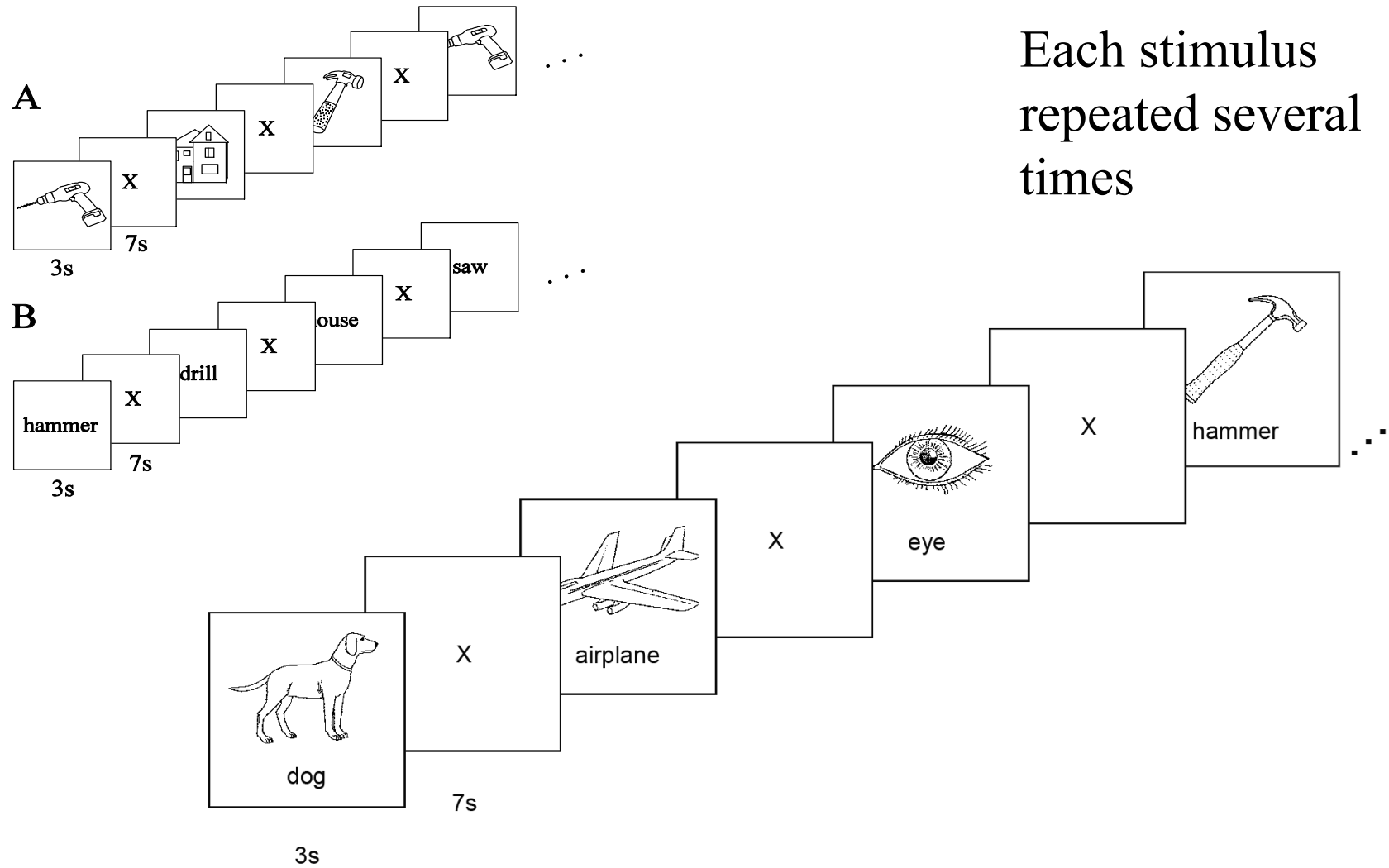
safe, non-invasive

measures Blood
Oxygen Level
Dependent (BOLD)
response

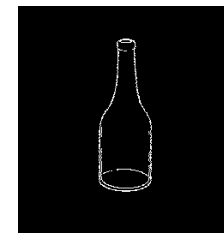
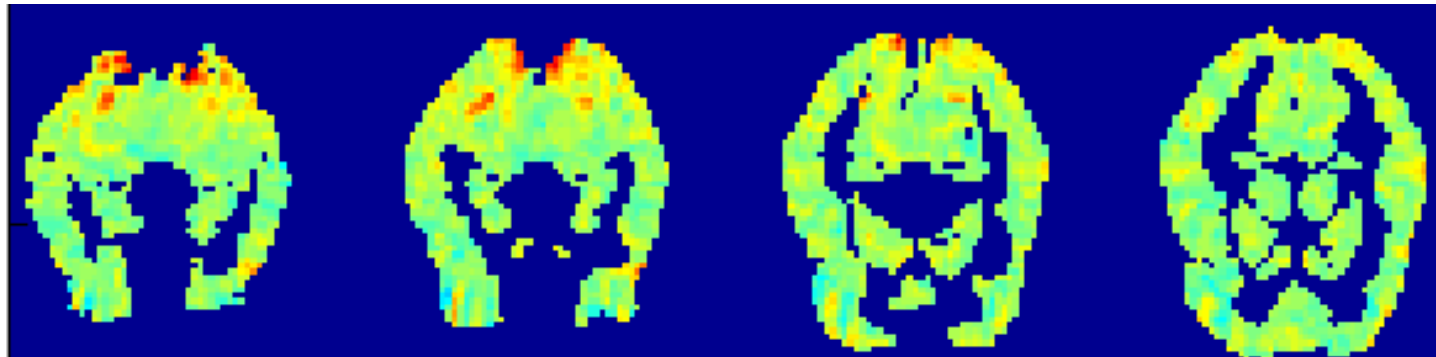


Typical fMRI
response to
impulse of
neural activity

Typical stimuli

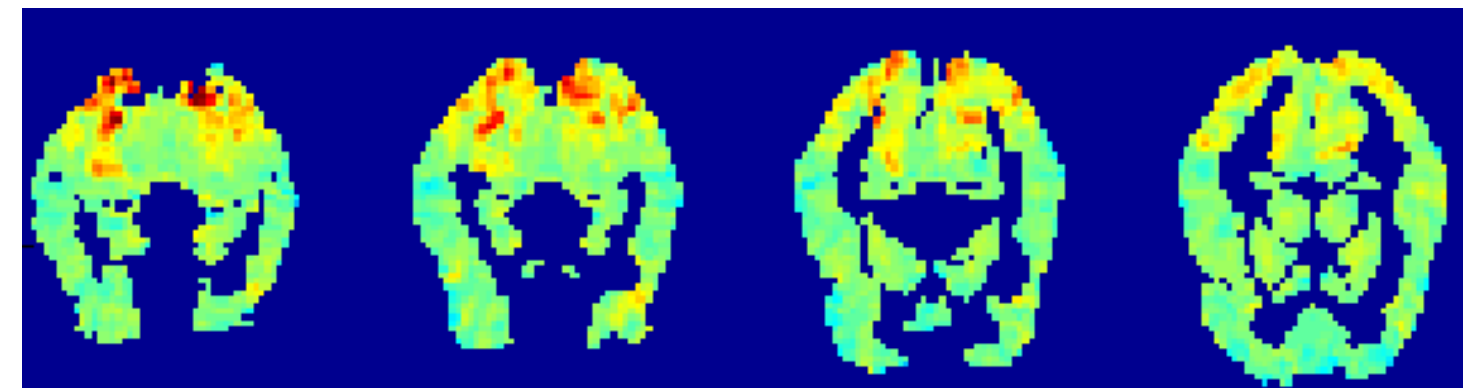


fMRI activation for “bottle”:

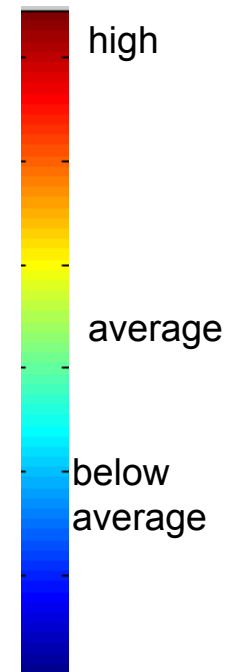


bottle

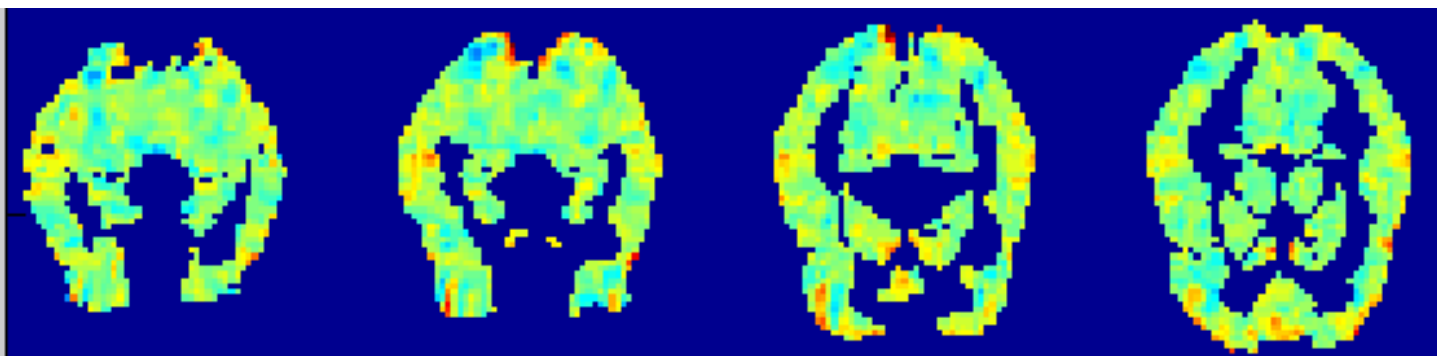
Mean activation averaged over 60 different stimuli:



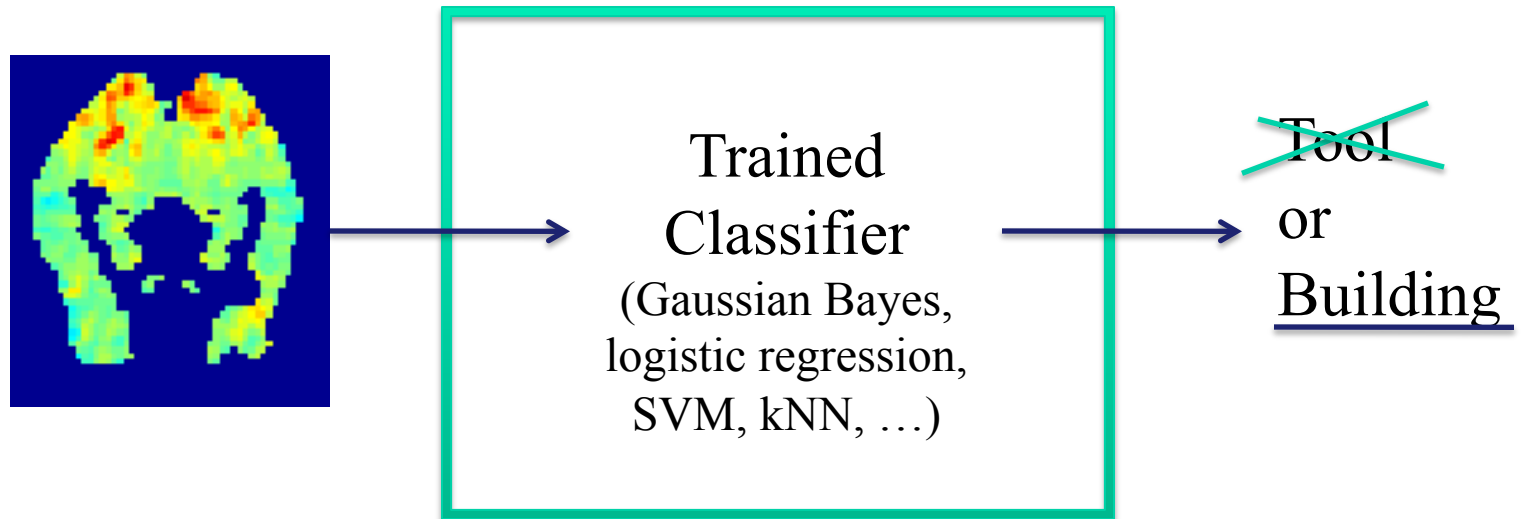
fMRI
activation



“bottle” minus mean activation:



Q1: Can one classify mental state from fMRI images?



(classifier as virtual sensor of mental state)

Training Classifiers over fMRI sequences

- Learn the classifier function

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

- Leave one out cross validation over 84 word presentations

- 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

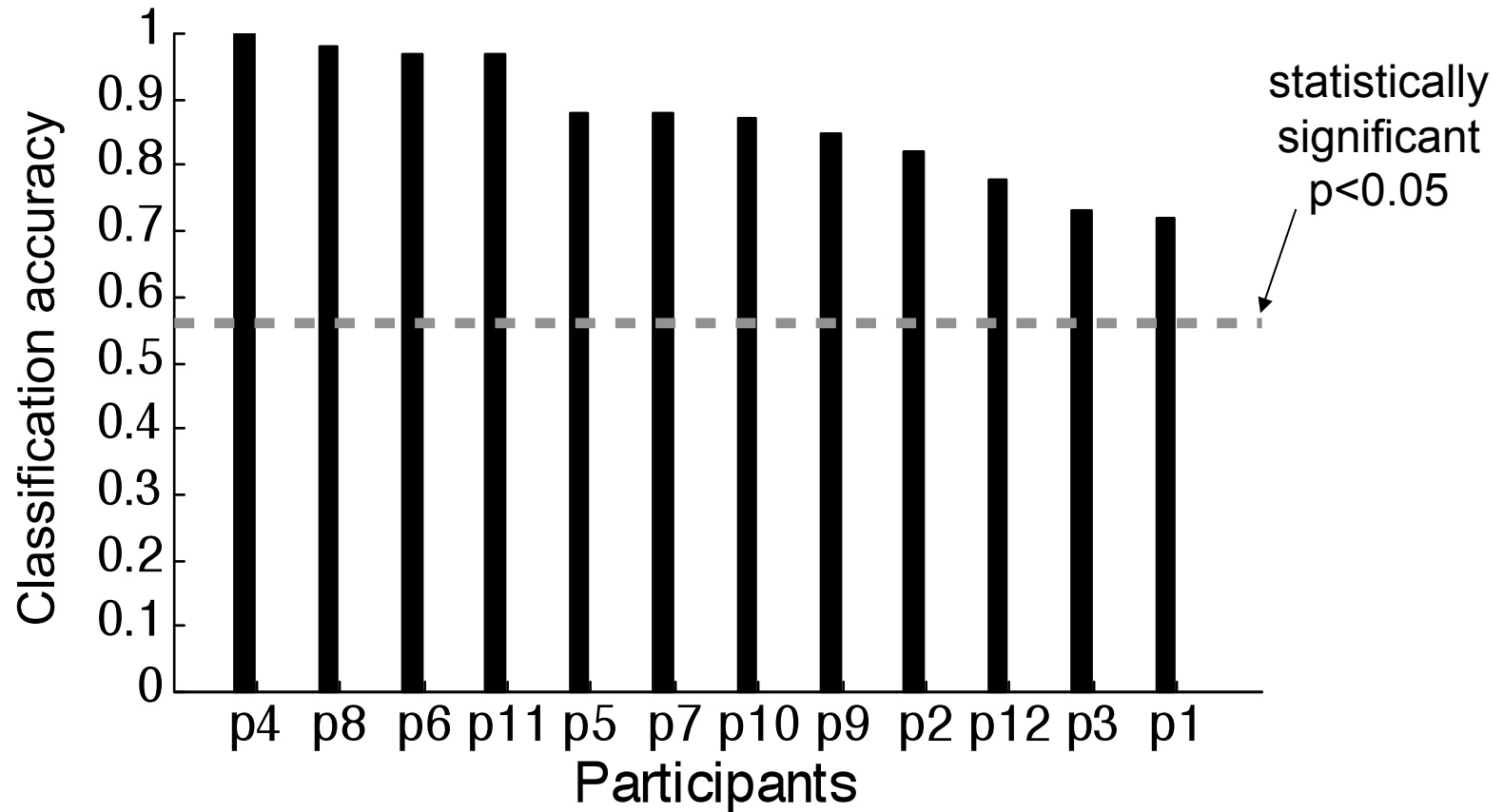
- Gaussian Naïve Bayes

- Regularized Logistic regression ← current favorite

- Feature selection methods tried:

- Logistic regression weights, voxel stability, activity relative to fixation,...

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



Brain Imaging and Machine Learning

ML Case study: high dimensional, sparse data

20,000 features

dozens of examples

- "Learning to Decode Cognitive States from Brain Images," T.M. Mitchell, et al., *Machine Learning*, 57(1), pp. 145-175, 2004
- "The Support Vector Decomposition Machine" F. Pereira, G. Gordon, *ICML-2006*.
- "Classification in Very High Dimensional Problems with Handfuls of Examples", M. Palatucci and T. Mitchell, *ECML-2007*
- Francisco Pereira PhD (2007).

Brain Imaging and Machine Learning

ML Case study: complex time series generated by hidden processes

- “Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models,” Hutchinson, et al., *NeuroImage*, 2009 (to appear).
- "Hidden Process Models", Rebecca Hutchinson, T. Mitchell, I. Rustandi, *ICML-2006*.
- "Learning to Identify Overlapping and Hidden Cognitive Processes from fMRI Data," R. Hutchinson, T.M. Mitchell, I. Rustandi, *11th Conference on Human Brain Mapping*. 2005.
- Rebecca Hutchinson PhD thesis (2009)

Brain Imaging and Machine Learning

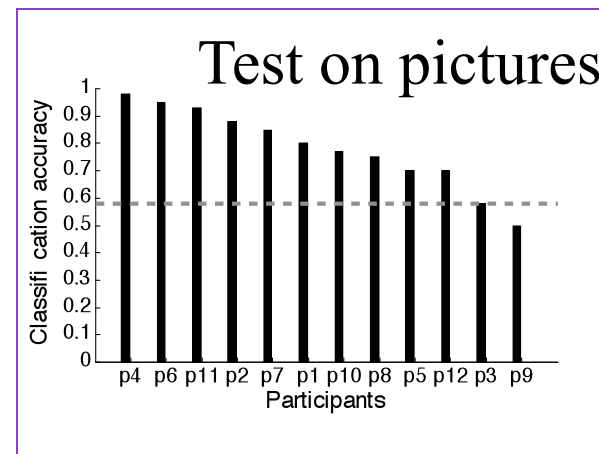
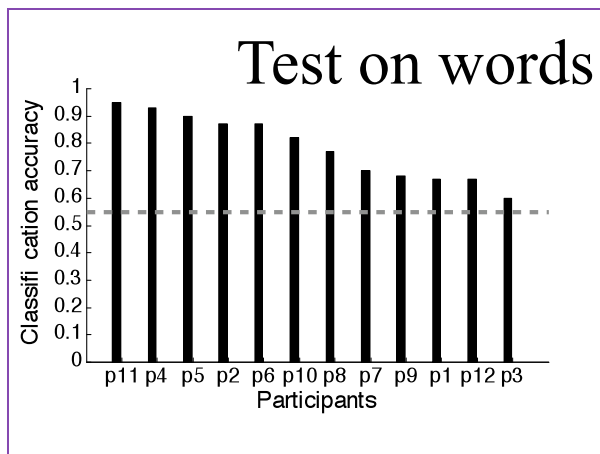
ML Case study: training many related classifiers

- "Training fMRI Classifiers to Discriminate Cognitive States across Multiple Subjects," X. Wang, R. Hutchinson, T. Mitchell, *NIPS2003*
- "Classifying Multiple-Subject fMRI Data Using the Hierarchical Gaussian Naïve Bayes Classifier", Indrayana Rustandi, *13th Conference on Human Brain Mapping*. June 2007.
- "Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings," S.V. Shinkareva, et al., *PLoS ONE* 3(1), January, 2008.
- Indra Rustandi PhD thesis (soon ...)

Question 2: Is our classifier capturing neural activity encoding stimulus meaning or appearance?

Can we train on word stimuli, then decode picture stimuli?

YES: We can train classifiers when presenting English words, then decode category of picture stimuli, or Portuguese words

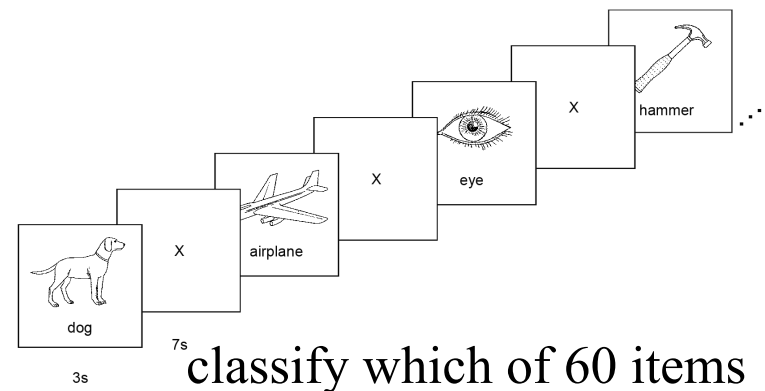
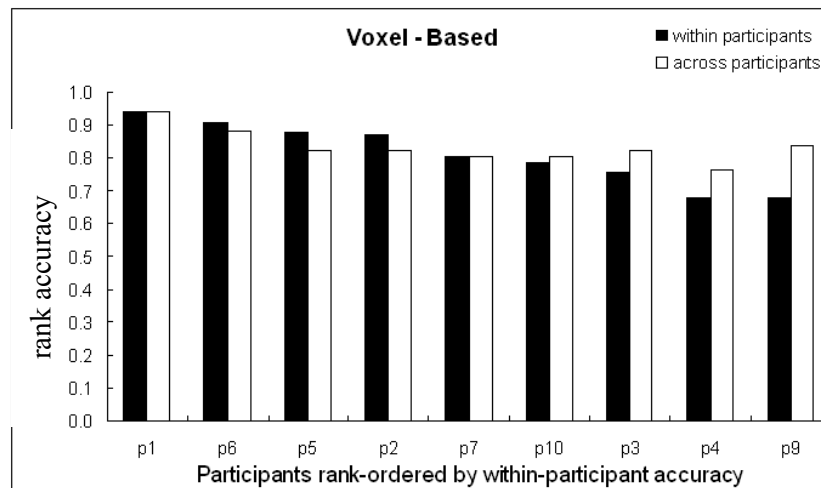


Therefore, the learned neural activation patterns must capture how the brain represents the meaning of input stimulus

Question 3: Are representations similar across people?

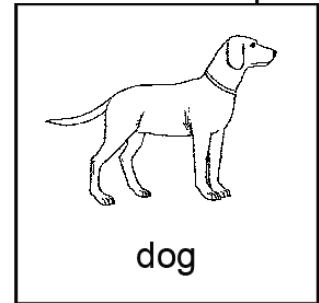
Can we train classifier on data from a collection of people, then decode stimuli for a new person?

YES: We can train on one group of people, and classify fMRI images of new person



Therefore, seek a theory of neural representations common to all of us (and of how we vary)

60 exemplars

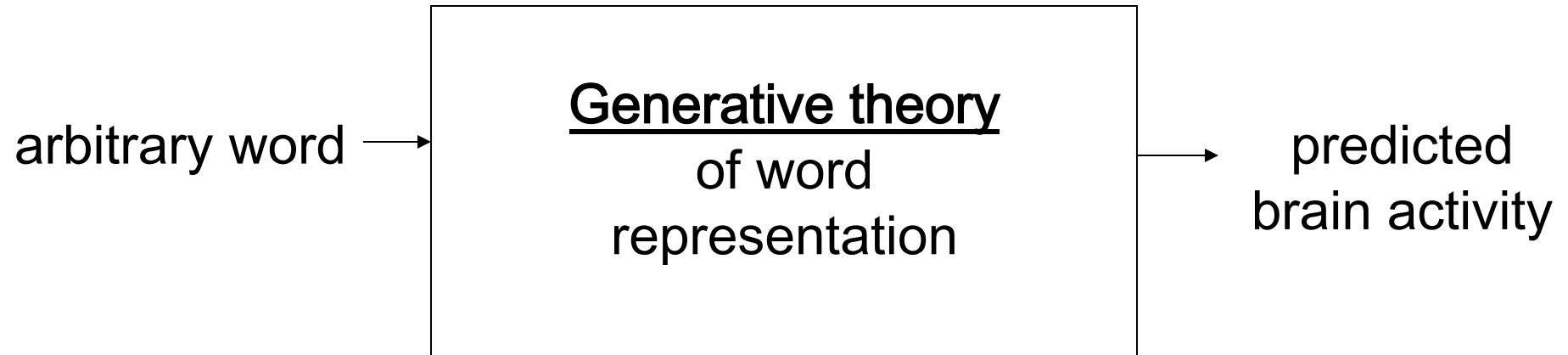


Categories

Exemplars

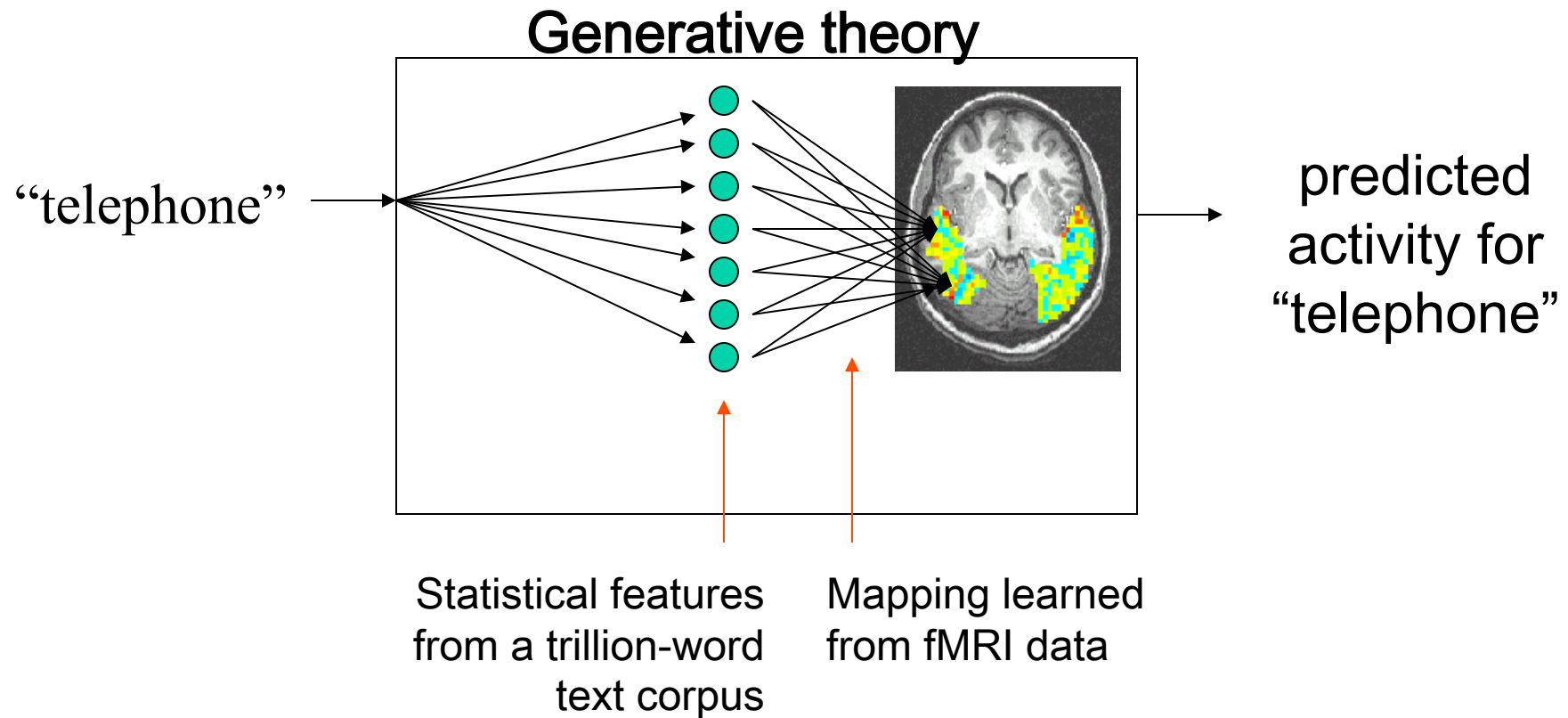
BODY PARTS	leg	arm	eye	foot	hand
FURNITURE	chair	table	bed	desk	dresser
VEHICLES	car	airplane	train	truck	bicycle
ANIMALS	horse	dog	bear	cow	cat
KITCHEN UTENSILS	glass	knife	bottle	cup	spoon
TOOLS	chisel	hammer	screwdriver	pliers	saw
BUILDINGS	apartment	barn	house	church	igloo
PART OF A BUILDING	window	door	chimney	closet	arch
CLOTHING	coat	dress	shirt	skirt	pants
INSECTS	fly	ant	bee	butterfly	beetle
VEGETABLES	lettuce	tomato	carrot	corn	celery
MAN MADE OBJECTS	refrigerator	key	telephone	watch	bell

Question 4: Can we discover underlying principles of neural representations?

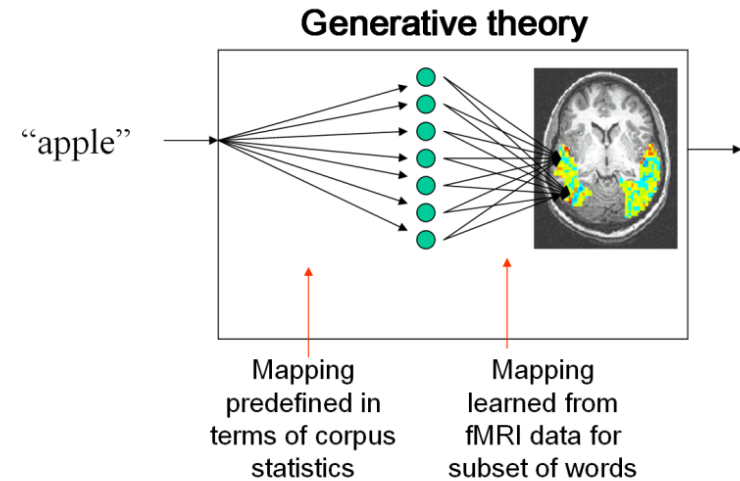


Idea: Predict neural activity from corpus statistics of stimulus word

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



Which corpus statistics?



- Feature i = co-occurrence frequency of stimulus noun with verb i
- The model uses 25 verbs:
 - Sensory: *see, hear, listen, taste, touch, smell, fear,*
 - Motor: *rub, lift, manipulate, run, push, move, say, eat,*
 - Abstract: *fill, open, ride, approach, near, enter, drive, wear, break, clean*

(why these 25?)

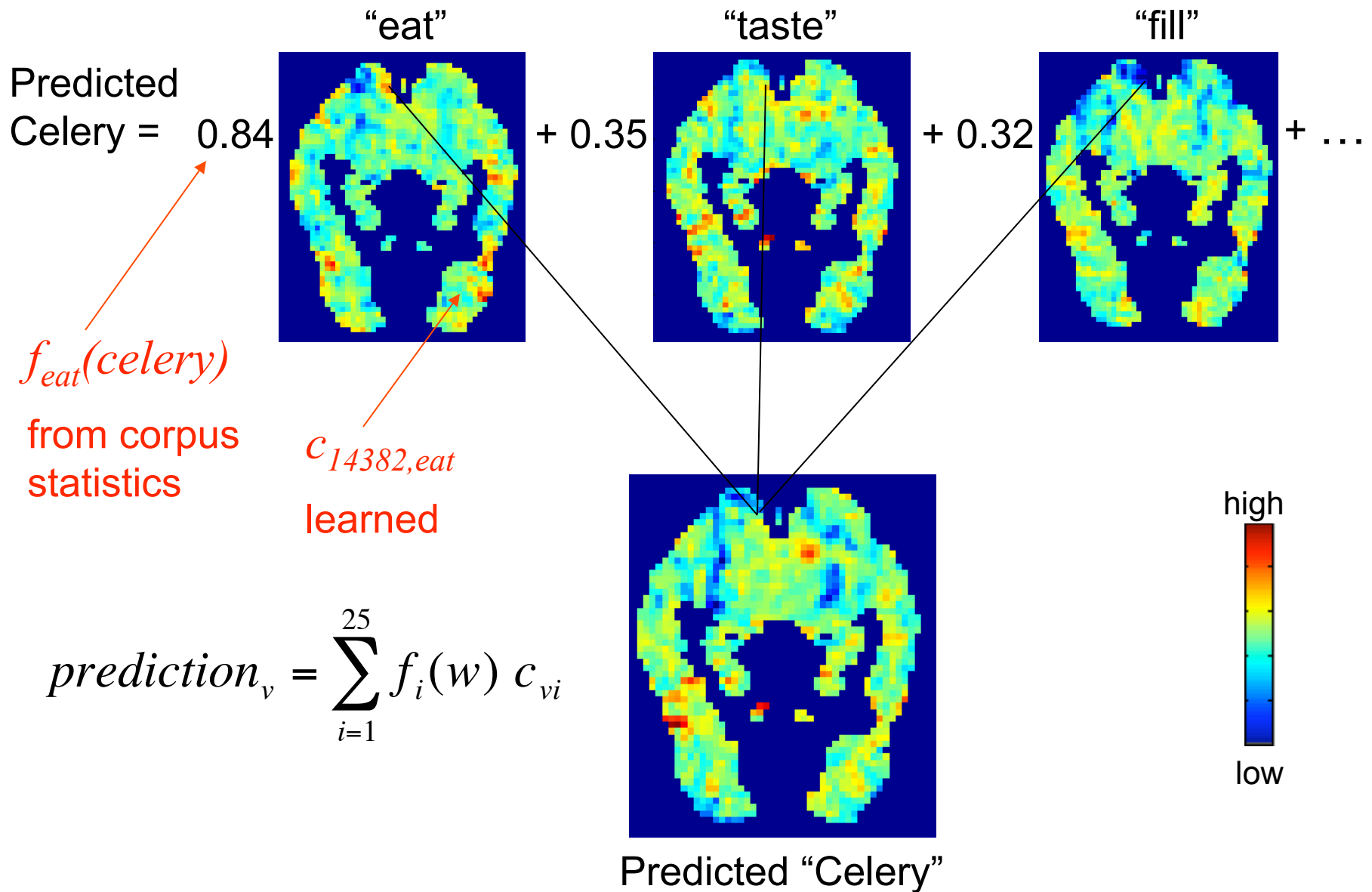
Semantic feature values: “**celery**”

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

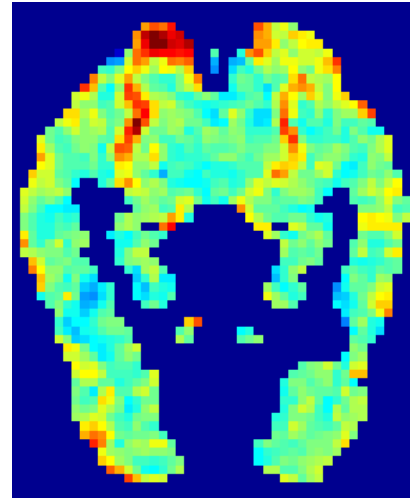
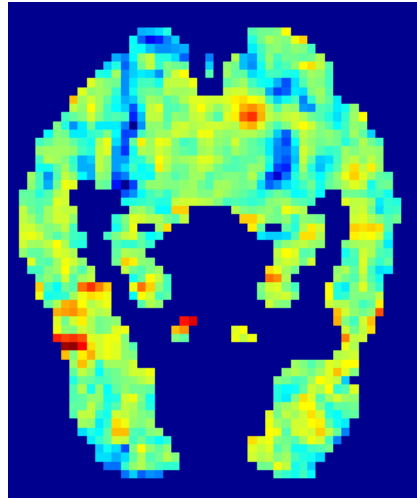
Predicted Activation is Sum of Feature Contributions



“celery”

“airplane”

Predicted:



fMRI
activation

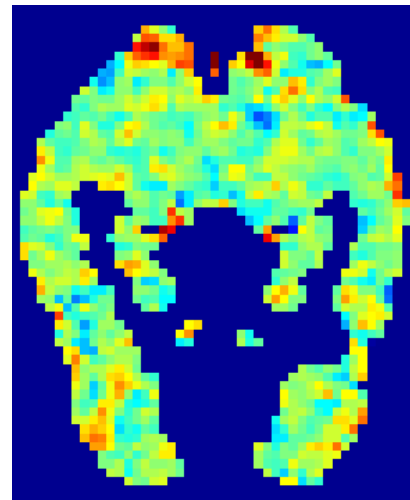
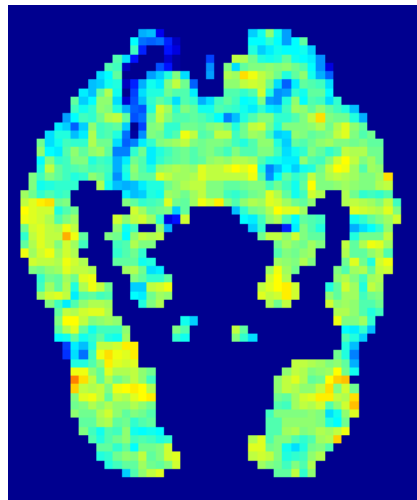


high

average

below
average

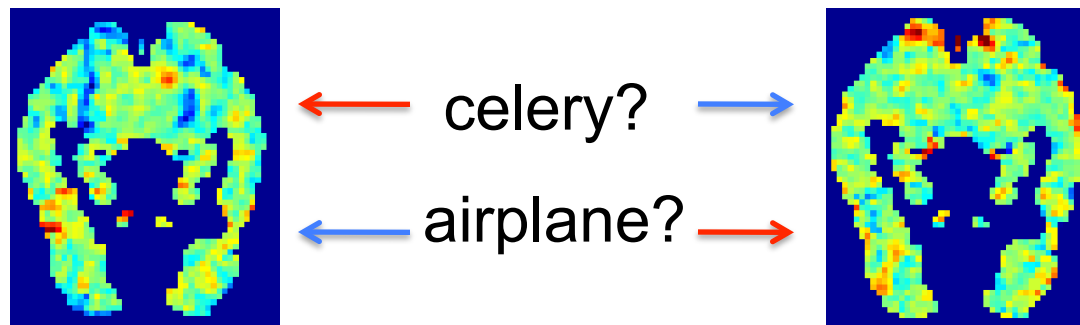
Observed:



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

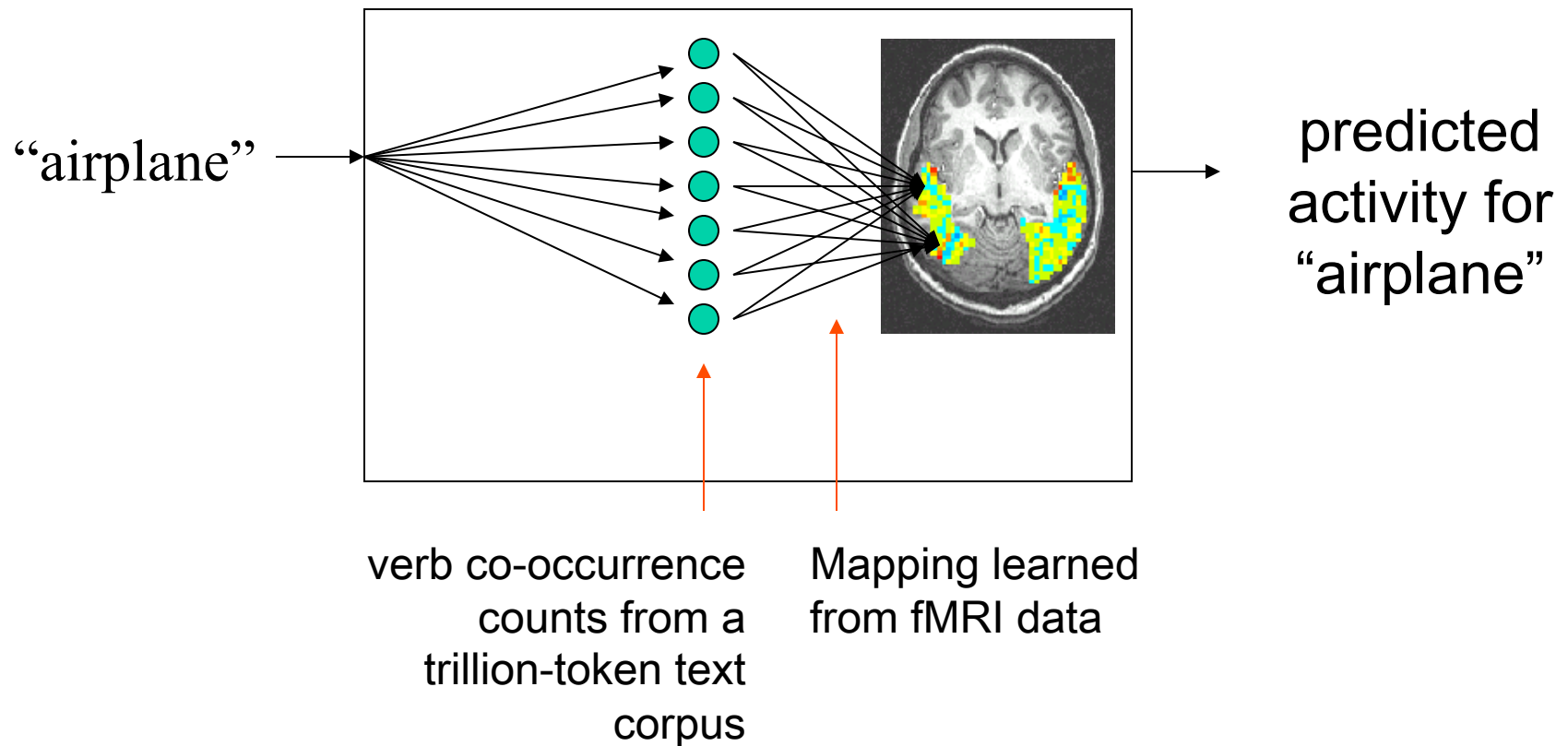


1770 test pairs in leave-2-out:

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

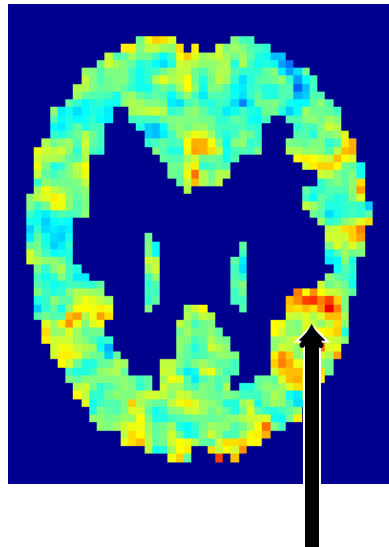
Mean accuracy over 9 subjects: 0.79

What are the learned semantic feature activations?



Participant
P1

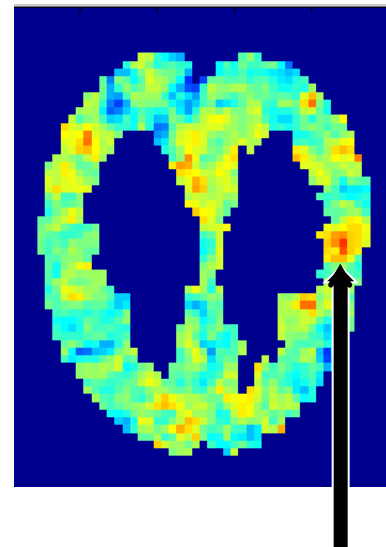
Eat



“Gustatory cortex”

Pars opercularis
(z=24mm)

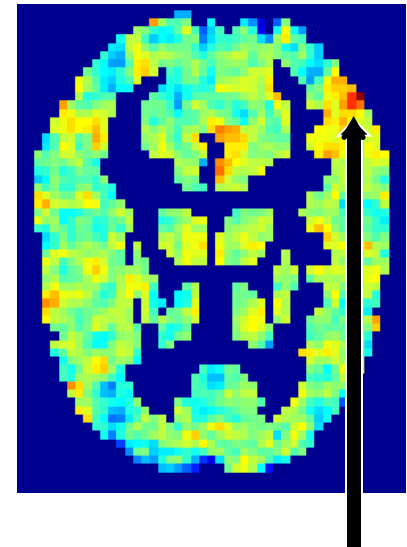
Push



“Planning motor
actions”

Postcentral gyrus
(z=30mm)

Run



“Body motion”

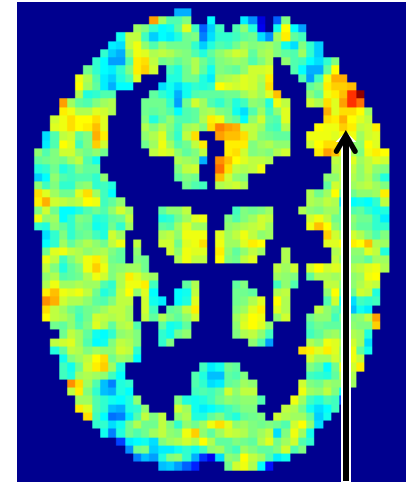
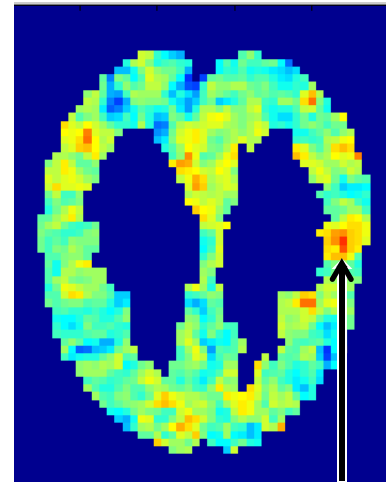
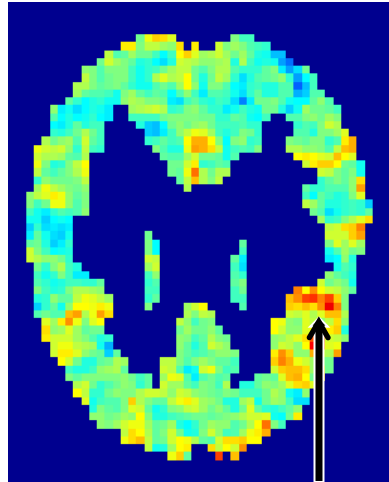
Superior temporal
sulcus (posterior)
(z=12mm)

Eat

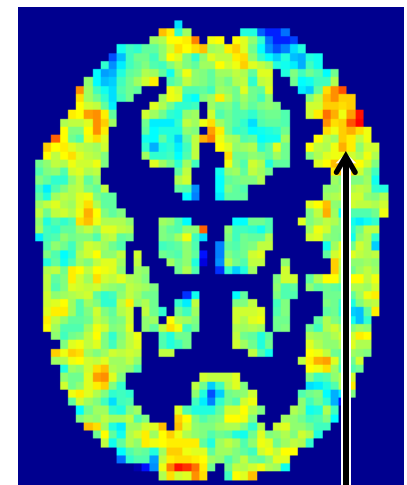
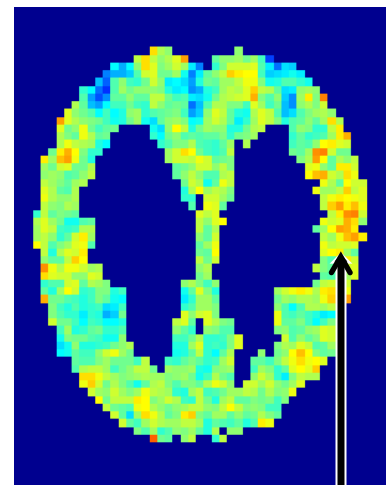
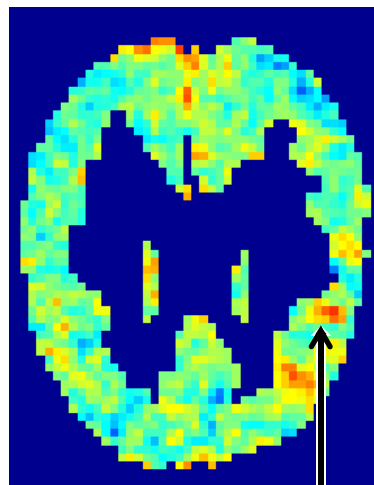
Push

Run

Participant
P1




Mean of
independently
learned signatures
over all nine
participants



Pars opercularis
(z=24mm)

Postcentral gyrus
(z=30mm)

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



Of the 10,000 most frequent English words, which noun is predicted to most activate each brain region?

Which nouns are *predicted to most activate**

Right Opercularis?

- wheat, beans, fruit, meat, paxil, pie, mills, bread, homework, eve, potatoes, drink
- **gustatory cortex [Kobayakawa, 2005]**

Right Superior Posterior Temporal lobe?

- sticks, fingers, chicken, foot, tongue, rope, sauce, nose, breasts, neck, hand, rail
- **associated with biological motion [Saxe et al., 2004]**

Left Anterior Cingulate?

- poison, lovers, galaxy, harvest, sin, hindu, rays, thai, tragedy, danger, chaos, mortality
- **associated with processing emotional stimuli [Gotlib et al, 2005]**

** for participant P1*

Which nouns are *predicted to most activate*

Left Superior Extrastriate ?

- madrid, berlin, plains, countryside, savannah, barcelona, shanghai, navigator, roma, stockholm, francisco, munich

Left Fusiform gyrus ?

- areas, forests, pool, bathrooms, surface, outlet, lodging, luxembourg, facilities, parks, sheffield

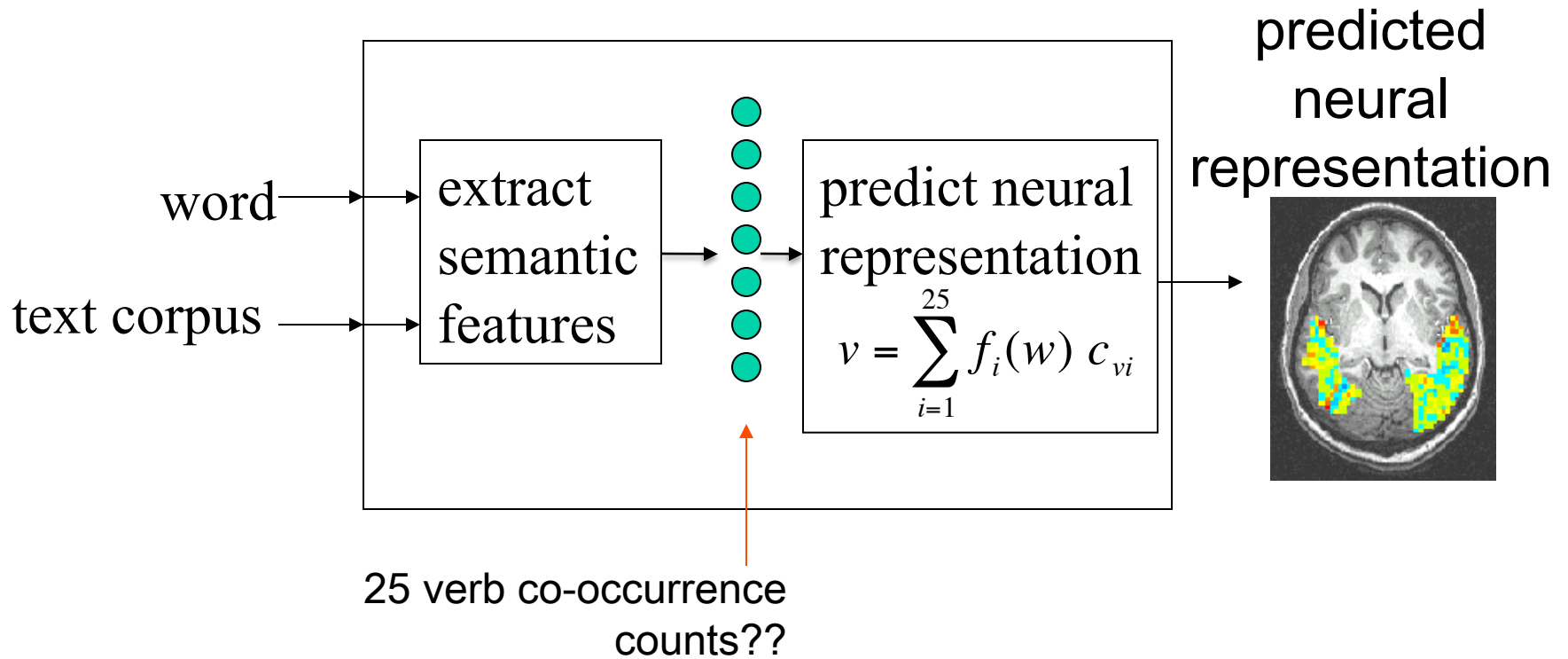
Left Inferior Posterior Temporal cortex ?

- thong, foot, skirt, neck, pantyhose, skirts, thongs, sexy, fetish, thumbs, skin, marks

- inferior temporal regions are associated with sexual arousal

[Stoleru 1999; Ferretti 2005]

Q: What is the semantic basis 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-occurrences	.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-occurrences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
207 features collected using <i>Mechanical Turk</i>	.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 Pommerleau.

feature values 1 to 5

features collected from
at least three people

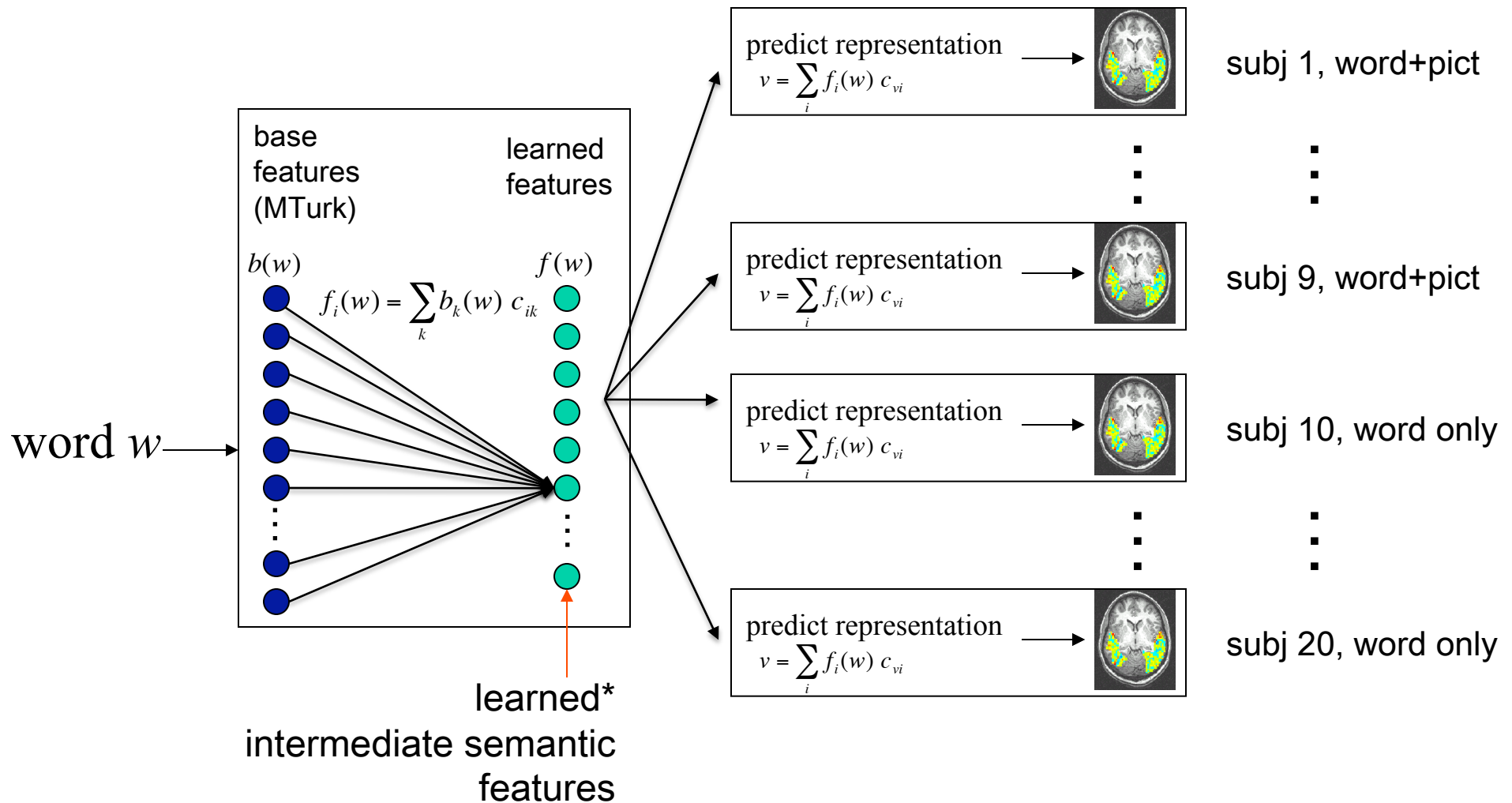
people provided by
Yahoo's "Mechanical
Turk"

Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurrences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
207 features collected using <i>Mechanical Turk</i>	.83
20 features discovered from the data	.88

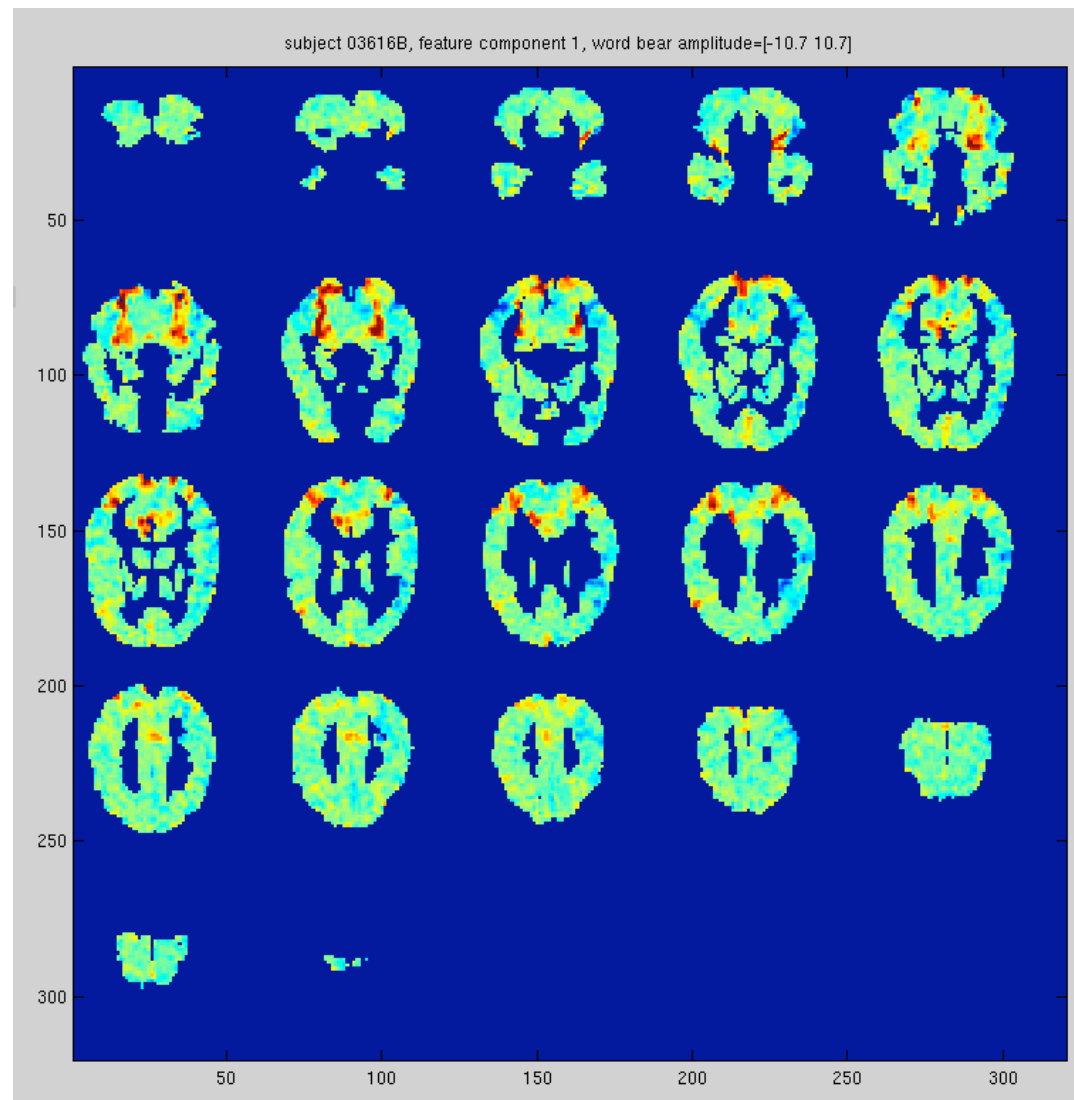
Discovering semantic basis shared across subjects and stimulus modality

[Rustandi, 2009]



* trained using Canonical Correlation Analysis

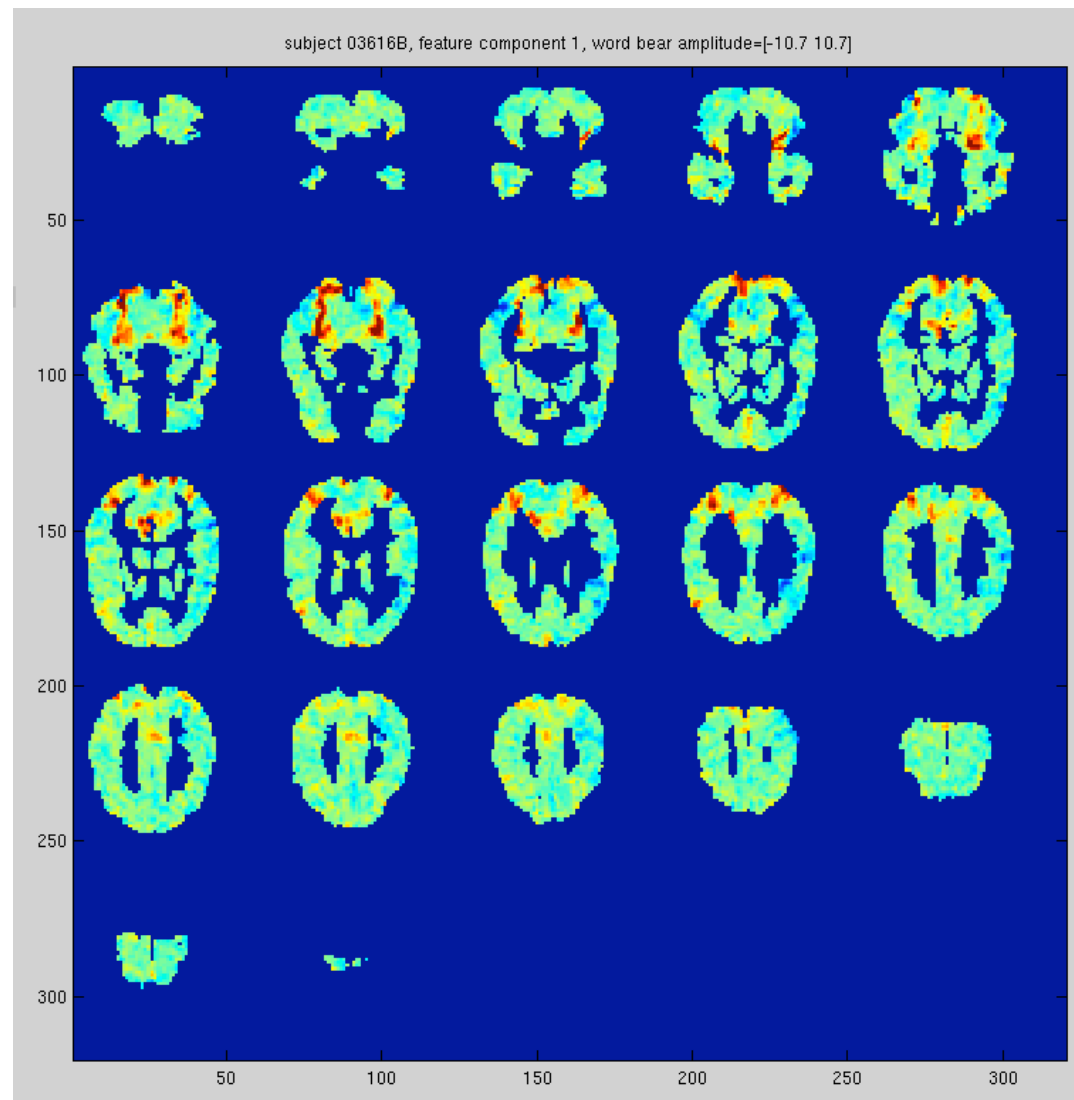
Subject 1 (Word-Picture stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1



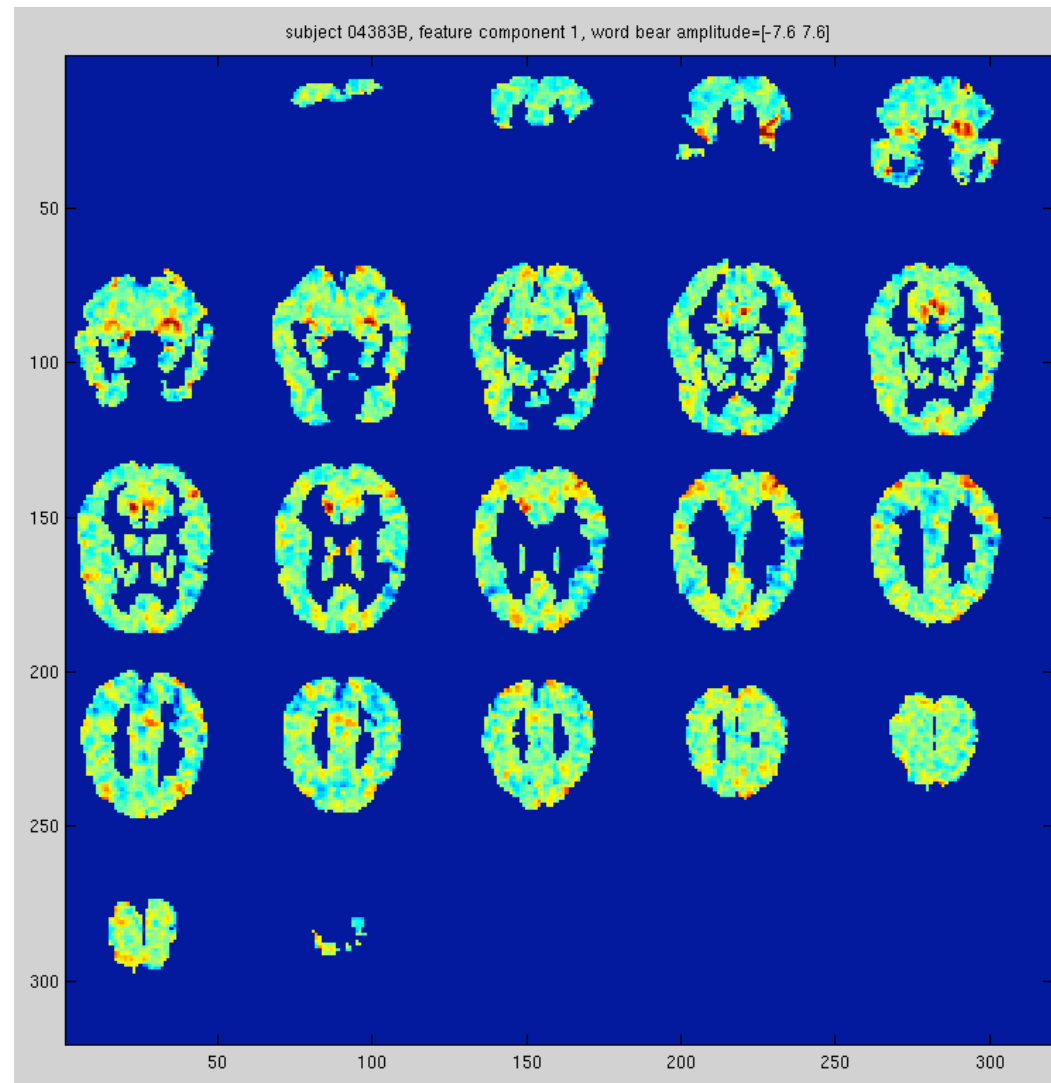
Multi-study (WP+WO) Multi-subject (9+11) CCA
 Top Stimulus Words

	component 1	component 2	component 3	component 4	component 5
positive	apartment church closet house barn	screwdriver pliers refrigerator knife hammer	telephone butterfly bicycle beetle dog	pants dress glass coat chair	corn igloo key cup eye
negative	knife cat spoon key pliers	cat car eye dog fly	leg key foot arm chair	car spoon screwdriver saw carrot	arm foot hand horse airplane

Subject 1 (Word-Picture stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1



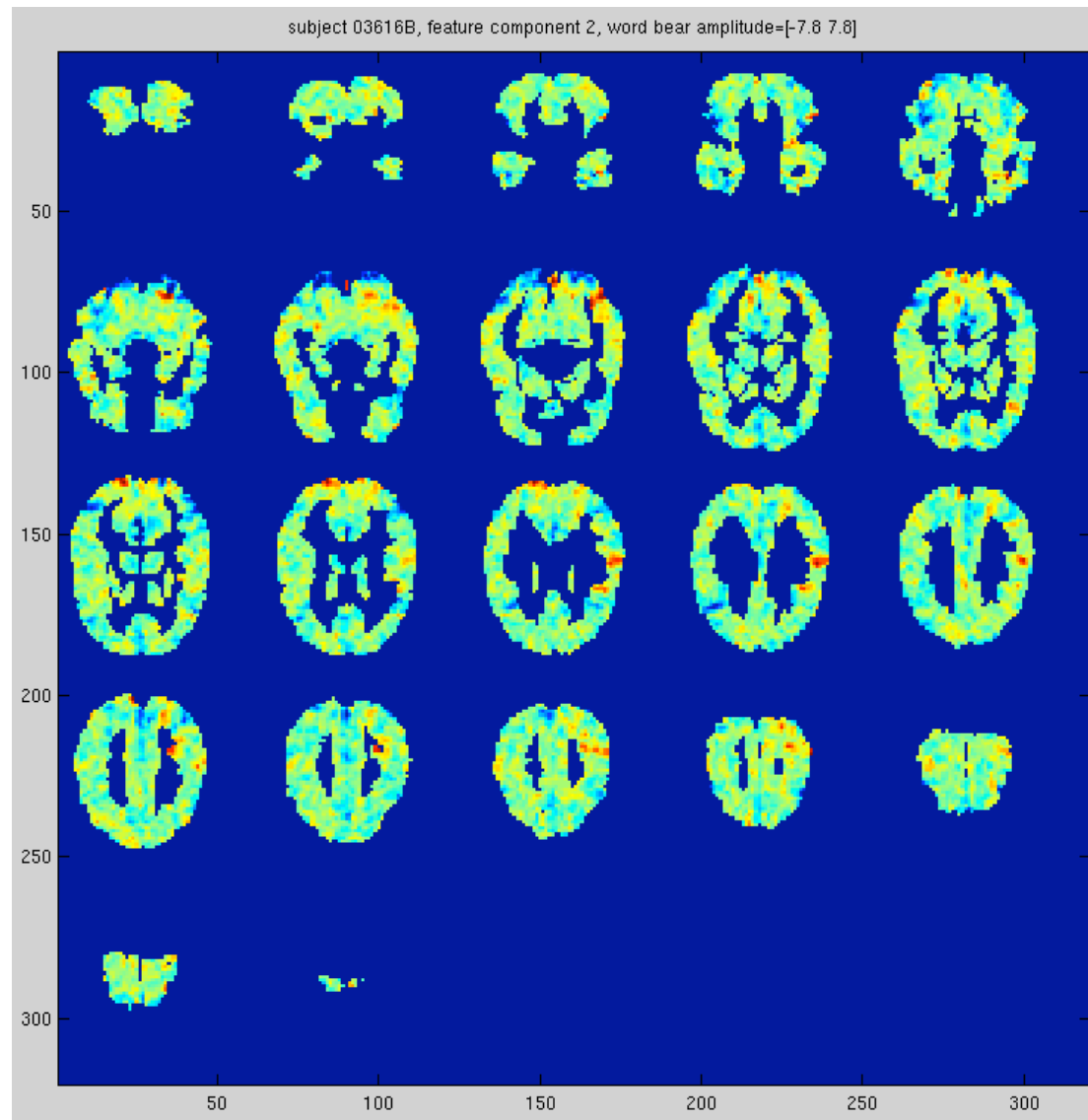
Subject 1 (Word-ONLY stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1



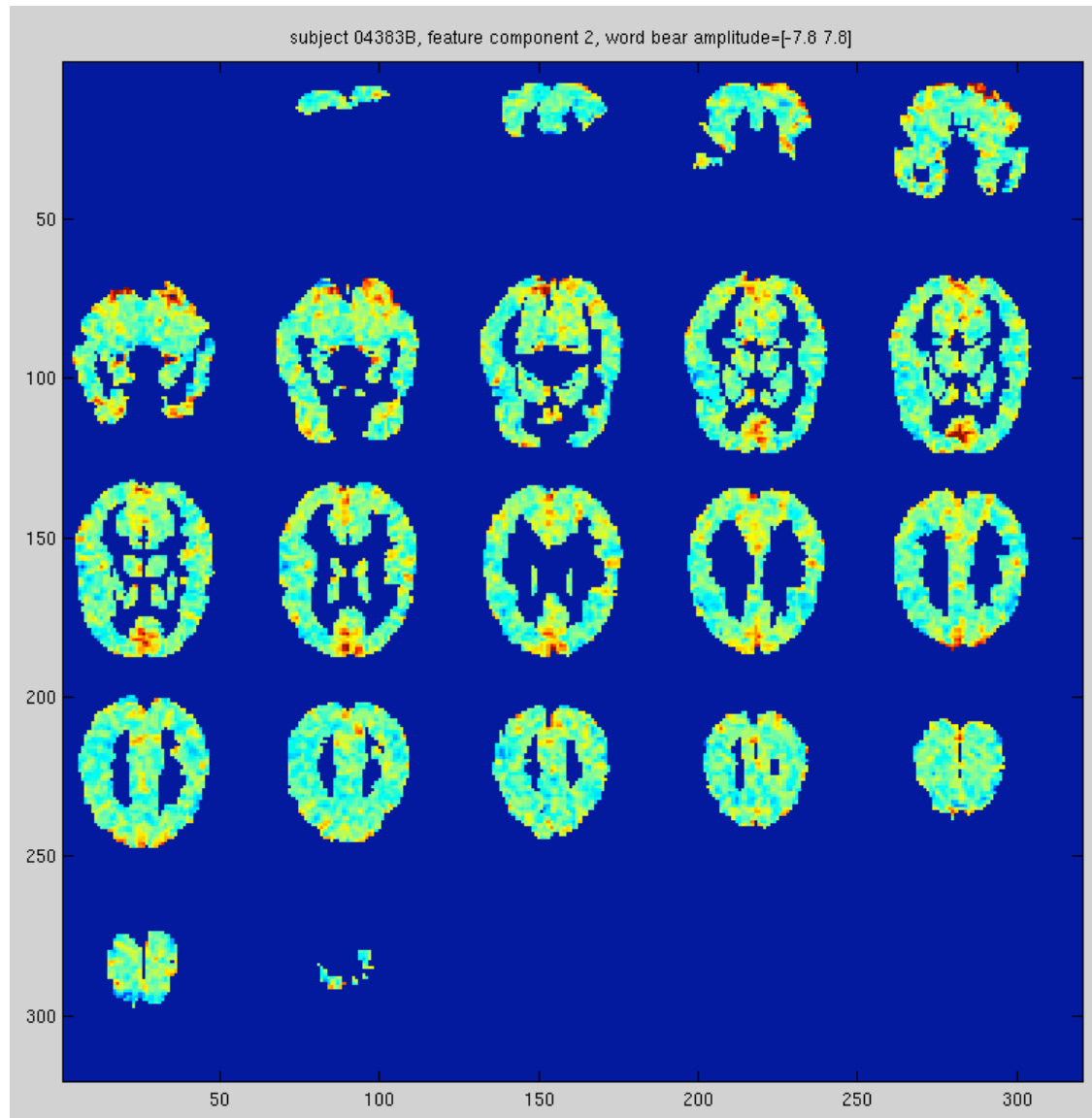
Multi-study (WP+WO) Multi-subject (9+11) CCA
 Top Stimulus Words

	component 1	component 2	component 3	component 4	component 5
positive	apartment church closet house barn	screwdriver pliers refrigerator knife hammer	telephone butterfly bicycle beetle dog	pants dress glass coat chair	corn igloo key cup eye
negative	knife cat spoon key pliers	cat car eye dog fly	leg key foot arm chair	car spoon screwdriver saw carrot	arm foot hand horse airplane

Subject 1 (Word-Picture stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 2



Subject 1 (Word-ONLY stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 2





How are neural representations of phrase meanings related to representations of component words?

Representing meaning of phrases

[K. Chang, 2009]

Have a learned $f(c(w)) = \text{fMRI}(w)$

- $c(w)$ = corpus statistics of w
- f is learned from data

$$f(c(\text{bear})) \rightarrow \text{fMRI}(\text{bear})$$

$$f(g(c(\text{soft}), c(\text{bear}))) \rightarrow \text{fMRI}(\text{“soft bear”})$$

but what is $g(x,y)$??

Experiment:

- present individual nouns
- present adjective-noun pairs

Analysis:

- train predictive model $f(c(w))=fMRI(w)$ using just five semantic features:
see, hear, smell, eat, touch
- for phrases, consider multiple ways (multiple functions “g”) to construct features from component words
 - $f(g(c(adj),c(noun))) = fMRI(adj,noun)$

Adjective	Noun	Category
Soft	Bear	Animal
Large	Cat	Animal
Strong	Dog	Animal
Plastic	Bottle	Utensil
Small	Cup	Utensil
Sharp	Knife	Utensil
Hard	Carrot	Vegetable
Cut	Corn	Vegetable
Firm	Tomato	Vegetable
Paper	Airplane	Vehicle
Model	Train	Vehicle
Toy	Truck	Vehicle

Table 1. Word stimuli.

Multiplicative Composition of Features Outperforms Alternatives

[K. Chang, 2009]

	See	Hear	Smell	Eat	Touch
Soft	0.07	0.07	0.01	0.02	0.83
Bear	0.57	0.16	0.02	0.17	0.08

Representations for “soft” and “bear”

	See	Hear	Smell	Eat	Touch
Adj	0.07	0.07	0.01	0.02	0.83
Noun	0.57	0.16	0.02	0.17	0.08
Add	0.64	0.23	0.03	0.19	0.91
Multi	0.04	0.01	0.00	0.00	0.06

Four possible representations for “soft bear”

	R^2
Adjective	0.36
Noun	0.38
Additive	0.35
Multiplicative	0.47

Multiplicative model yields best fMRI prediction

[J. Mitchell & M. Lapata, 2008] found multi. model ~predicted human similarity judgements

What next for Machine Learning challenges?

- ML: discover optimal features to replace the 25 verbs
 - discover low-dimensional semantic basis for both corpus and fMRI
 - (and for behavioral data such as subjective similarity judgments)
- ML: algorithm to learn cumulatively, from multiple studies with different words, people
 - must discover latent features
- ML: train using fMRI (1 mm) and MEG (1 msec)
 - *fuse data sources and train classifier, predictor*

What next for imaging experiments?

- Stimuli: 40 abstract nouns
 - love, democracy, anxiety, justice, ...
 - preliminary results: model can predict activation if retrained using 485 verbs
- Stimuli: adjective-noun pairs
 - ‘fast rabbit’ vs ‘hungry rabbit’ vs ‘cuddly rabbit’
 - study how brain combines representations of single words into representation of phrase meaning
- Collect new MEG, EEG, ECoG data with 1 msec temporal resolution
 - goal: combine 1mm fMRI spatial res, 1msec MEG temporal res
 - preliminary MEG results: successful (75% on average) classifier for category discrimination (“foods” vs. “body parts”)



thank you!