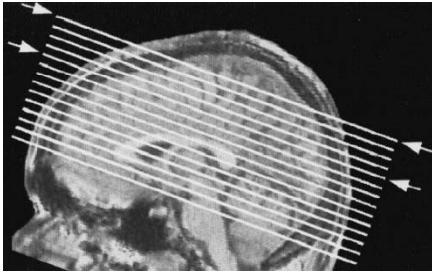
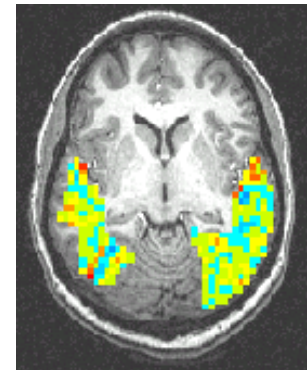


Brains, Meaning and Machine Learning



Tom Mitchell
and many collaborators



Machine Learning Department
Carnegie Mellon University

April, 2008

Neurosemantics Research Team

Postdoctoral Fellows



Svetlana Shinkareva



Rob Mason



Tom Mitchell



Marcel Just

Professional Staff



Vladimir
Cherkassky

PhD Students



Andy Carlson



Kai Min Chang



Rebecca Hutchinson



Mark Palatucci

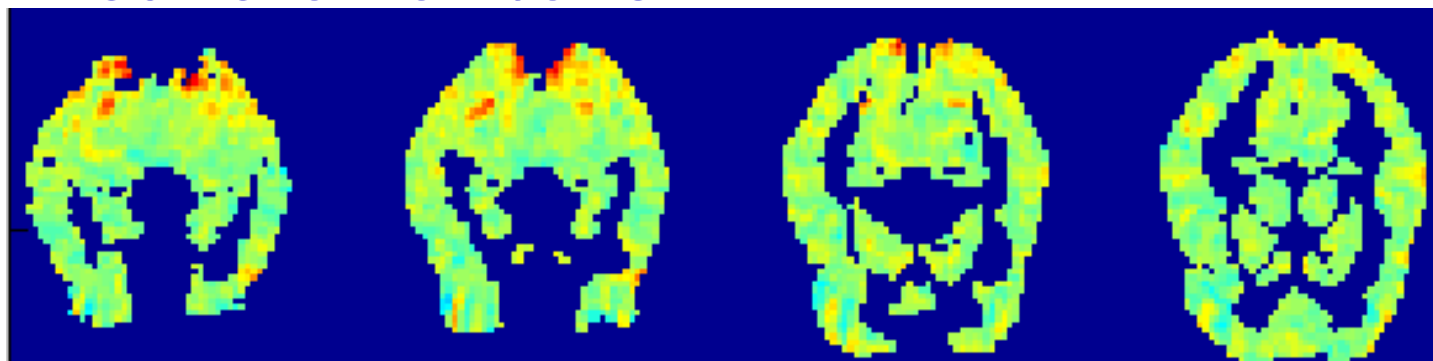


Indra Rustandi

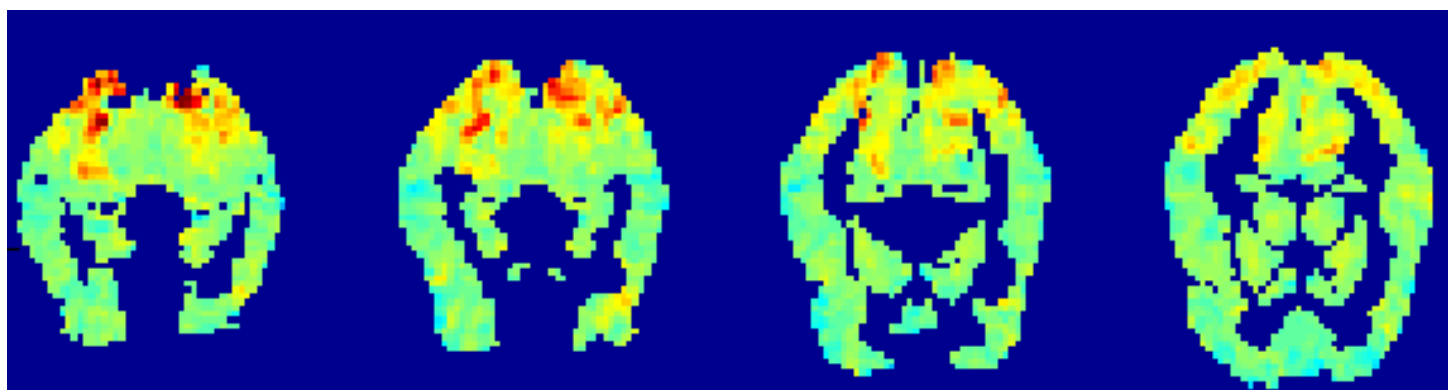


Francisco Pereira

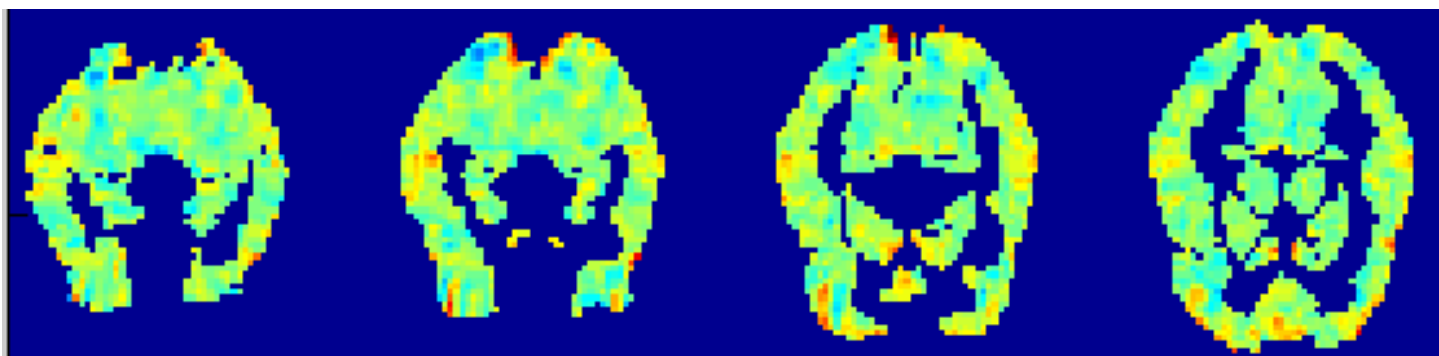
fMRI activation for “bottle”:



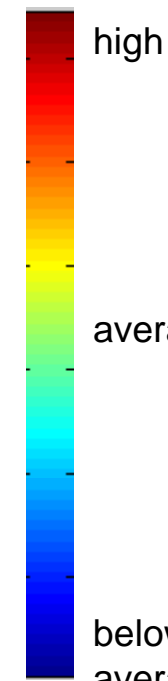
Mean activation averaged over 60 different stimuli:



“bottle” minus mean activation:



fMRI
activation

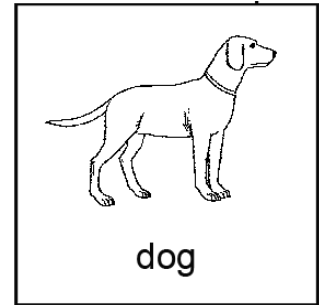


high

average

below
average

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 0:

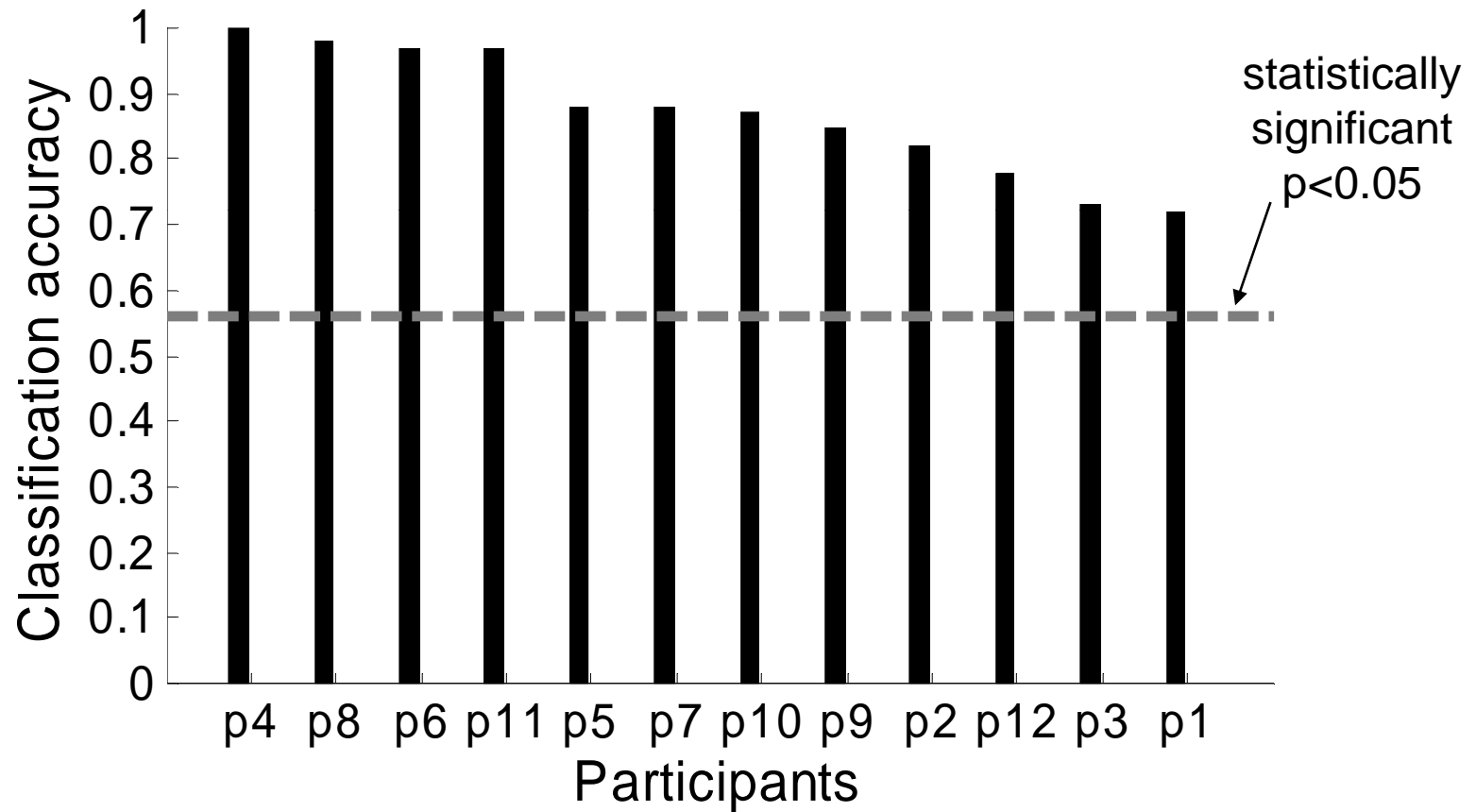
Using fMRI, can we observe brain activation representing the meaning of input stimulus?

Can we train a classifier to decode the semantic category of stimulus?

Answer:

Yes, for categories such as “tools,” “buildings,” “foods,” “body parts,” “vehicles”, etc.

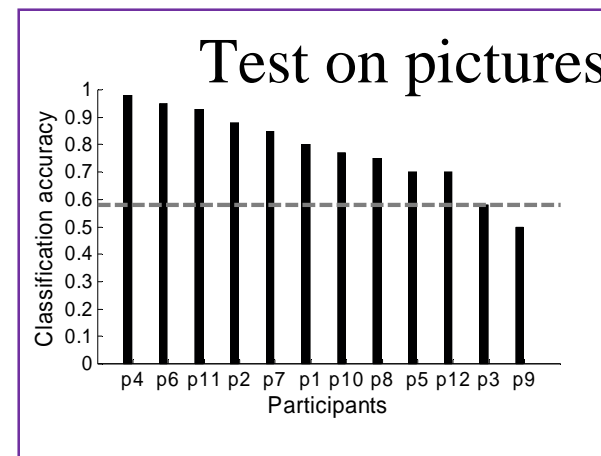
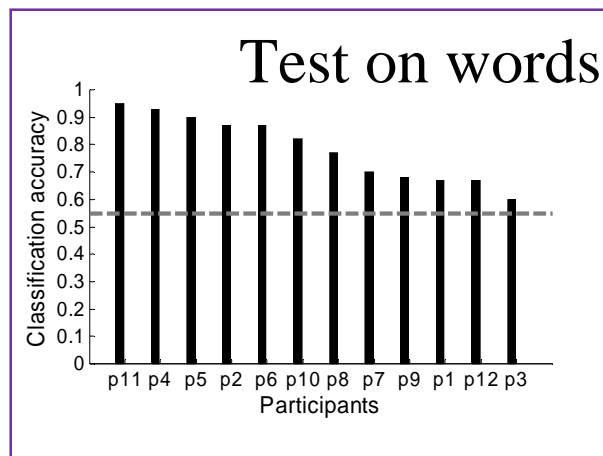
Classification task: is person viewing a “tool” or “building”?



Question 1: Is brain activity common across stimulus modality?

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

YES: Train on words, then:

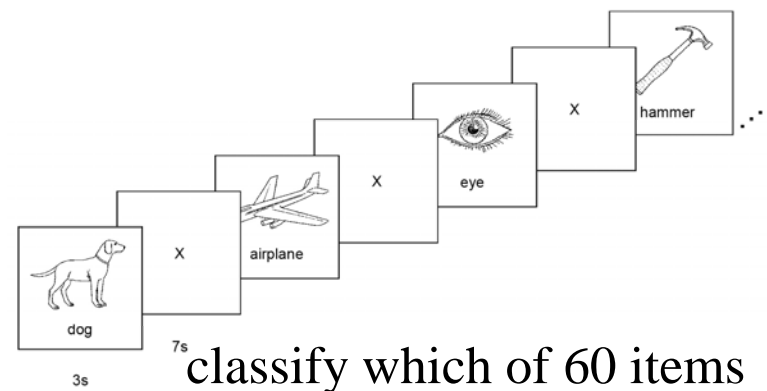
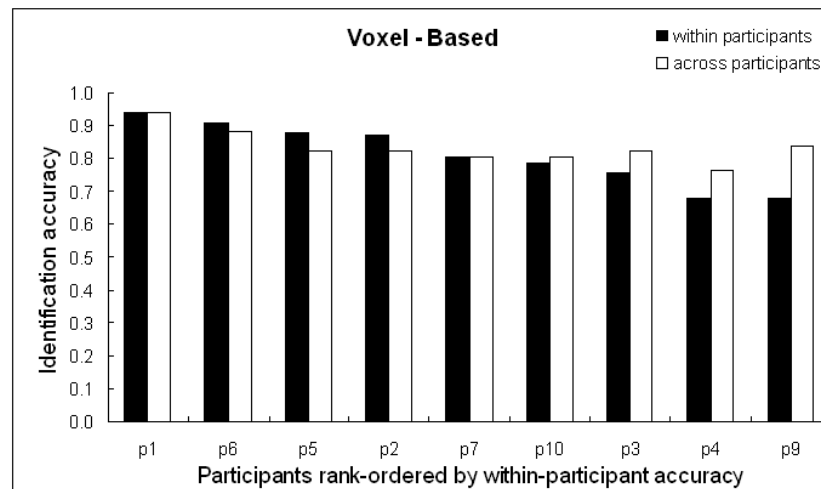


Therefore, the learned neural activation patterns must capture how the brain represents stimulus meaning

Question 2: Are representations similar across different people?

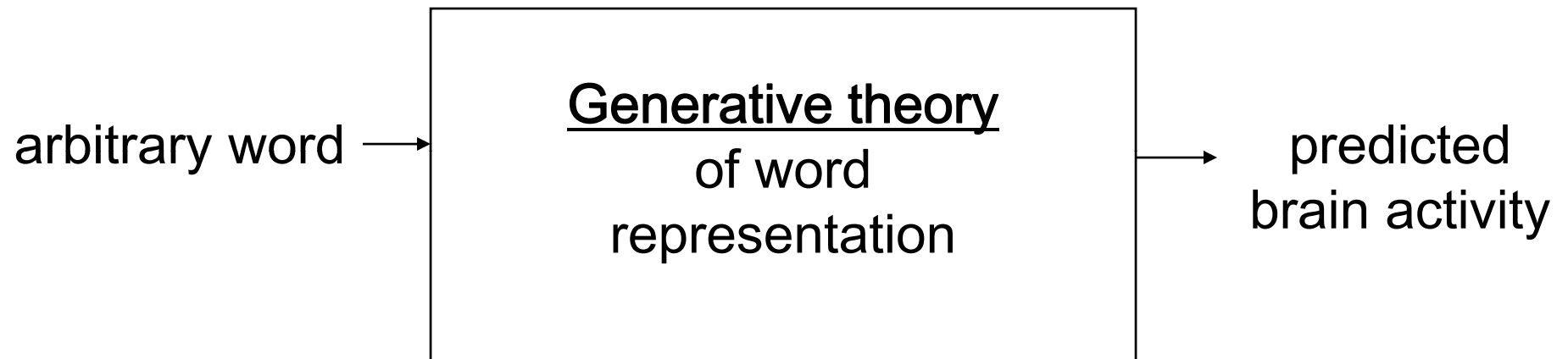
Can we train classifier on data from a collection of people, then decode stimuli for a new person?

YES: Train on one group of people, and classify fMRI images of new person



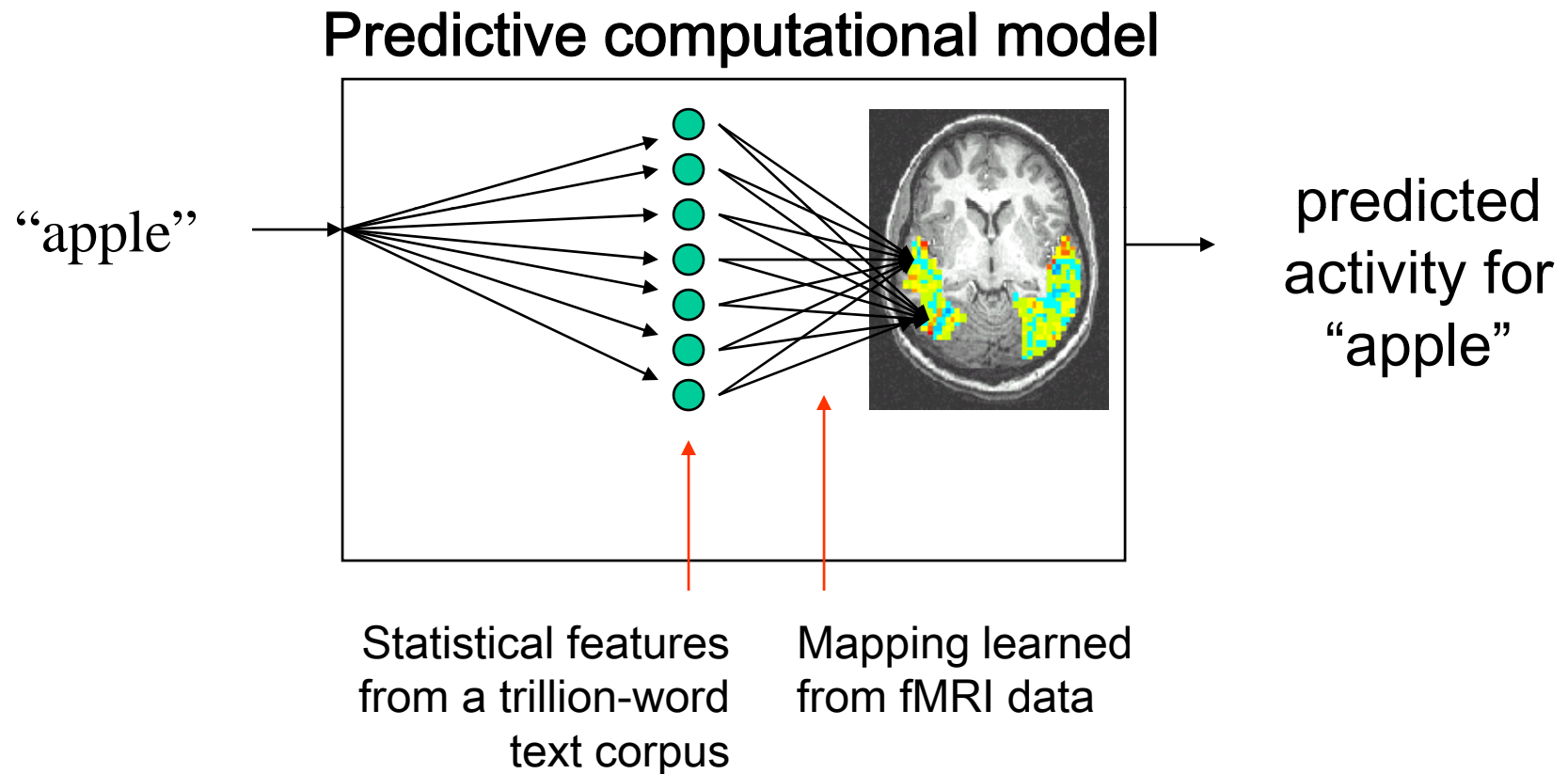
Therefore, we can seek a theory of neural representation common to all of us (and of how we vary)

What we really want: a generative theory

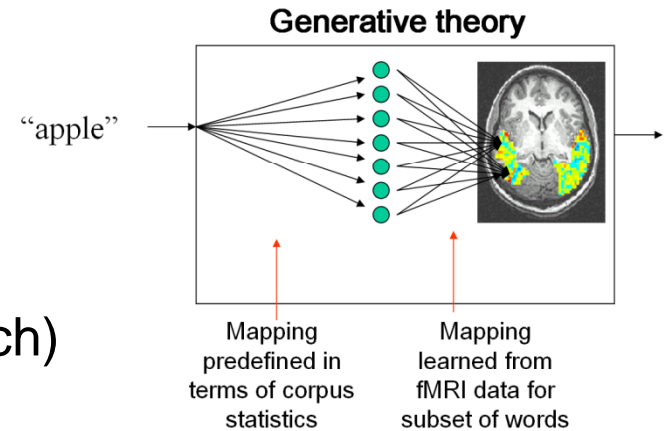


But how??

Question 3: Can we develop a theory to predict neural encoding for any word?



Approach: Integrate corpus data and fMRI data



- Semantic feature i = English word i (e.g., touch)
- Value of feature i = co-occurrence frequency of stimulus with i
 - in trillion-word text collection (tera-word ngram database provided by Google)
- Which semantic features? First attempt: 25 sensory/action verbs:
 - Sensory actions: *see, hear, listen, taste, touch, smell, fear,*
 - Motor actions: *rub, lift, manipulate, run, push, move, say, eat,*
 - Abstract actions: *fill, open, ride, approach, near, enter, drive, wear, break, clean*

(why these 25?)

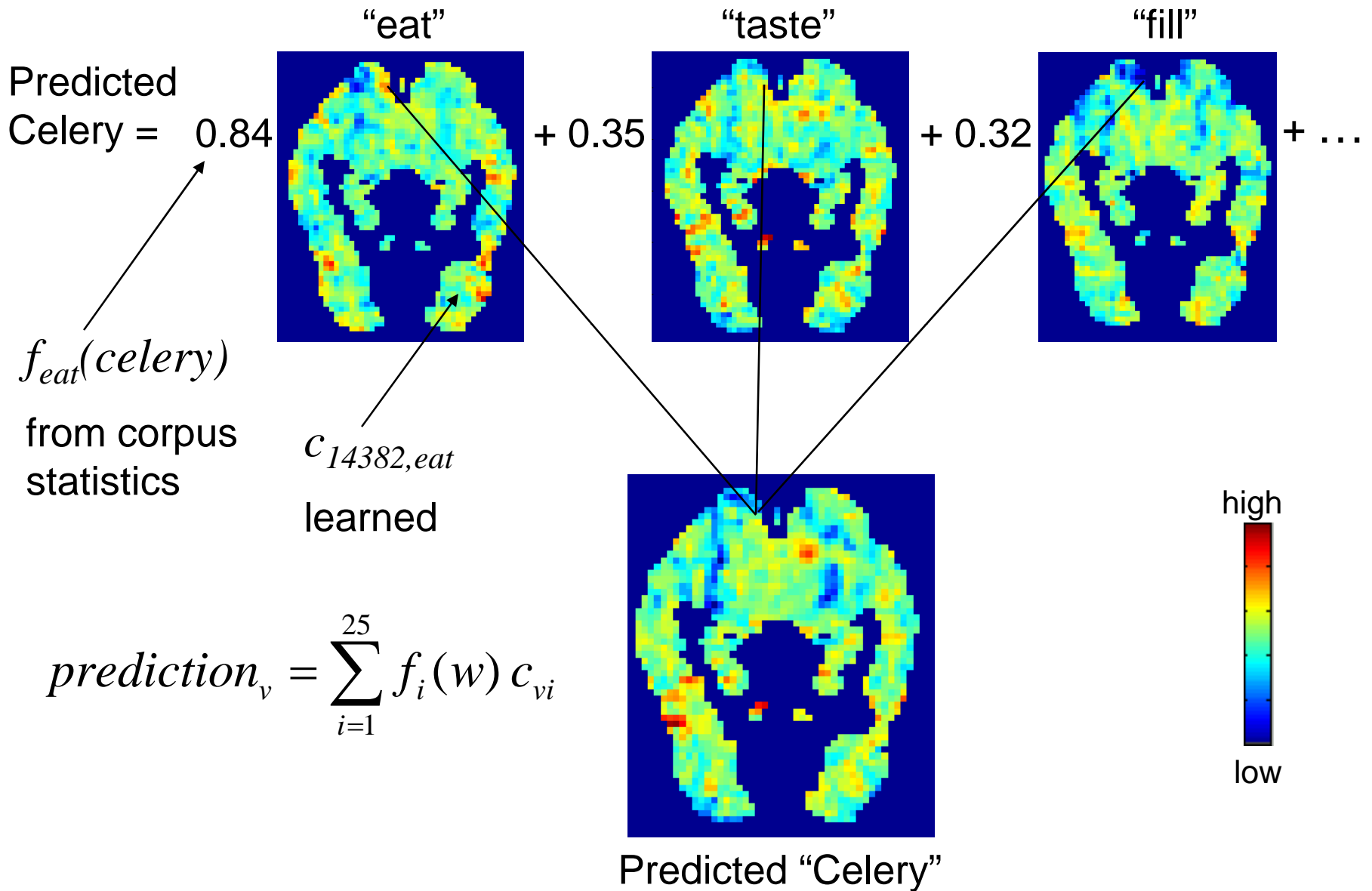
Semantic feature values: **“celery”**

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

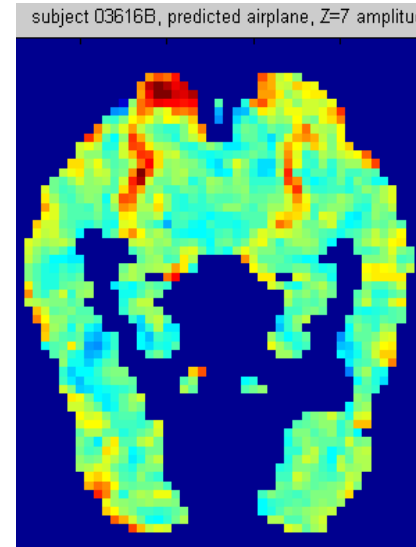
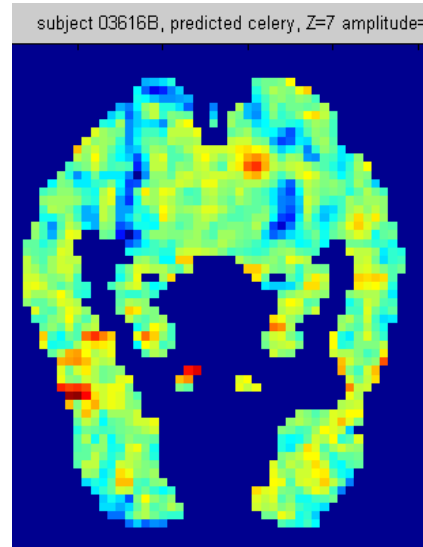
Predicted Activation is Sum of Feature Contributions



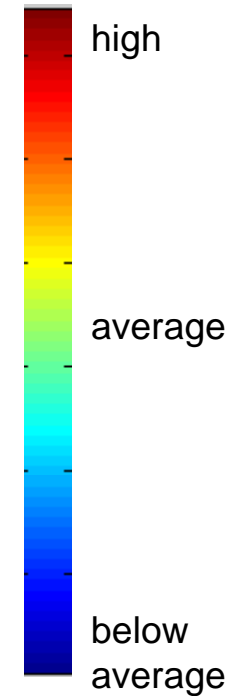
“celery”

“airplane”

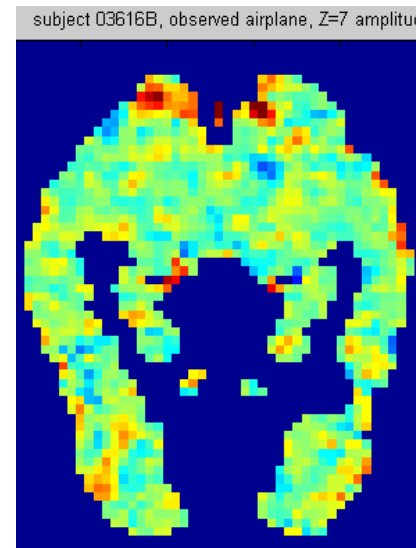
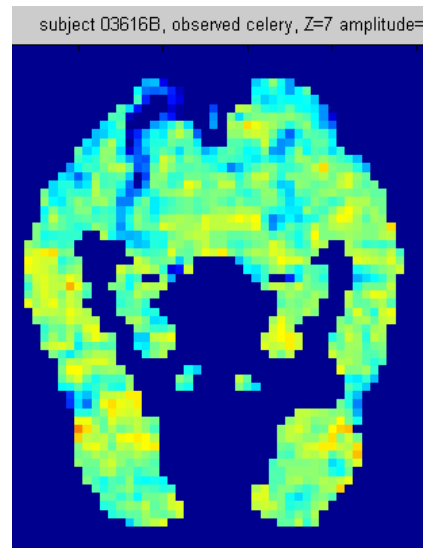
Predicted:



fMRI
activation



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
- Image similarity measured by cosine similarity using only the 500 most “stable” voxels (over training set)
- 1770 test pairs in leave-2-out
 - Random guessing \rightarrow 0.50 prediction accuracy
 - Accuracy above 0.61 is significant ($p < 0.05$) according empirical distribution

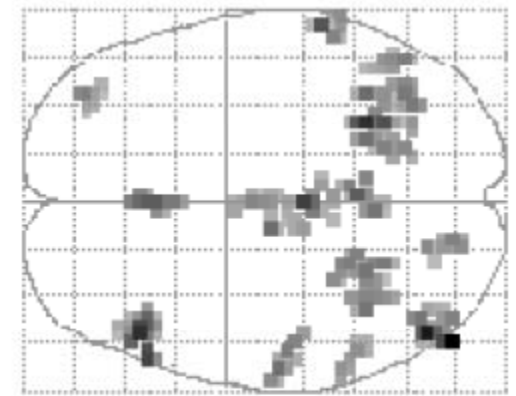
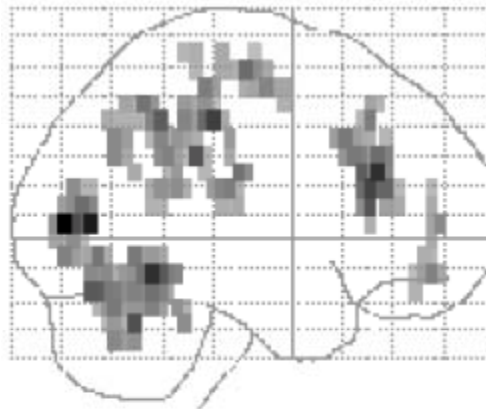
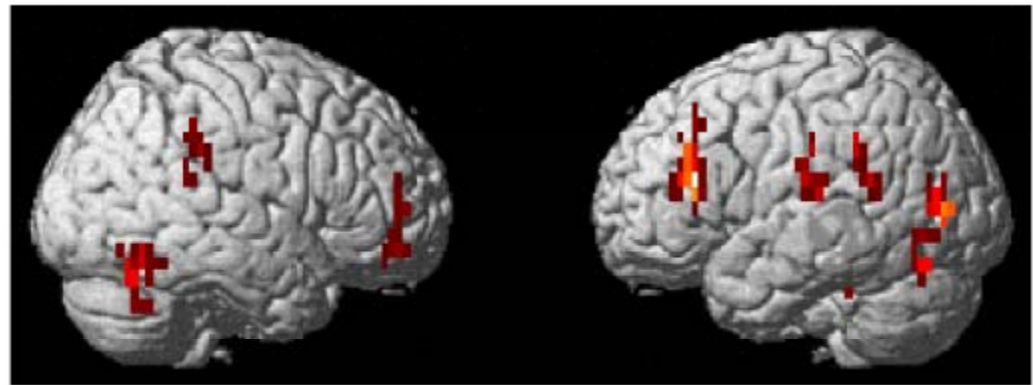
Accuracy predicting images for new words

- Accuracy of independently trained models for nine participants (using 500 most stable voxels over the 58 training words):
 - .85, .83, .82, .78, .78, .76, .73, .72, .68
 - Mean: .77
- Accuracy extrapolating to new categories (when testing on “celery” vs “airplane”, leave out all training foods and vehicles)
 - .78, .78, .74, .69, .69, .68, .67, .64, .64
 - Mean: .70

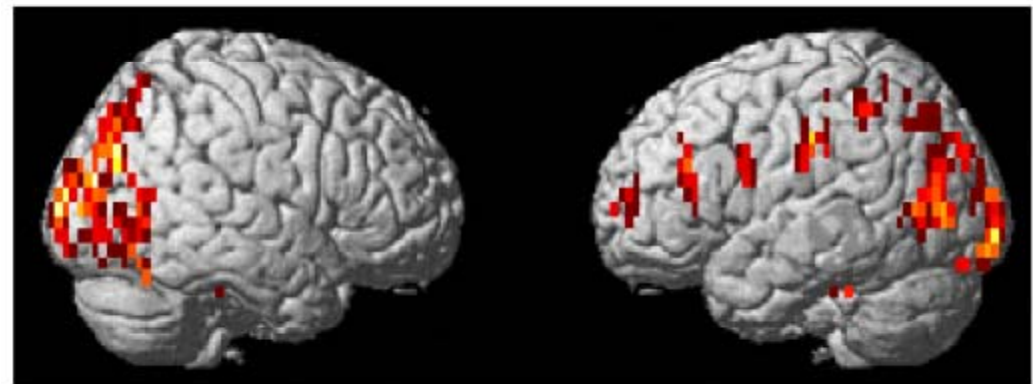
Accuracy Map

left hemisphere
inferior
temporal,
fusiform,
motor cortex,
intraparietal
sulcus,
inferior frontal,
orbital frontal,
occipital cortex

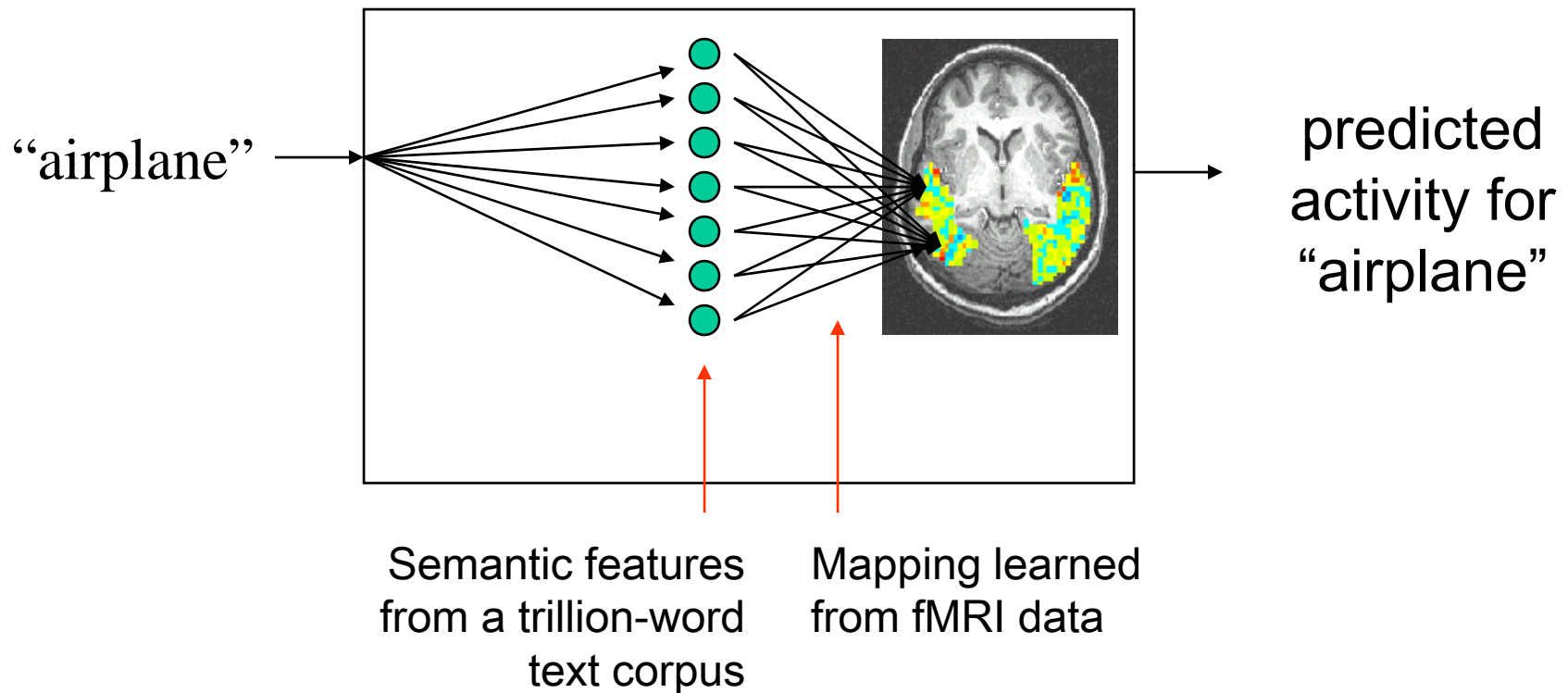
Subj
0399B



9 subj
mean

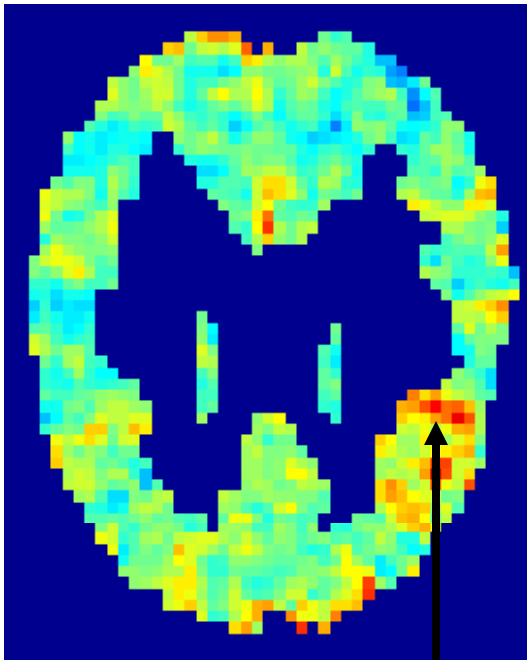


What are the learned semantic feature activations?



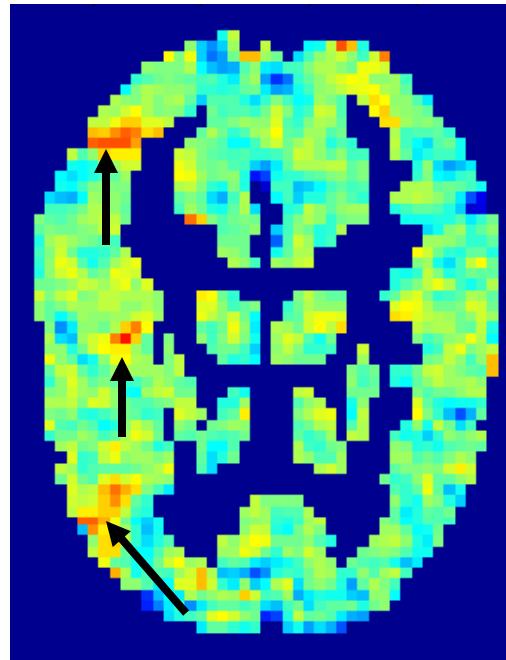
Learned Semantic Feature Signatures (P1)

'eat'



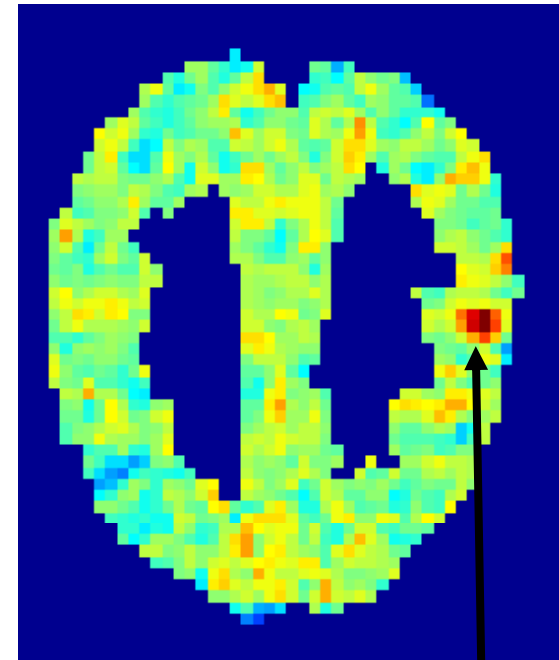
gustatory cortex

'listen'




regions of
language
processing and
audition

'touch'



primary
somatosensory
cortex



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 body 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]

Many next steps...

- Discover optimal semantic basis features
 - Alternatives to word cooccurrence
- Alternative Bayesian models
- Study individual differences/commonalities
 - Including different groups (e.g., autistics)
- Abstract nouns and verbs
 - Truth, beauty, justice. Think, complain, wonder, relax
- Processing multiple word phrases
 - “fast rabbit” vs. “sickly rabbit” vs. “white rabbit”
- Predict fMRI from both word and image stimuli

Brain Imaging and Machine Learning

ML Case study: high dimensional, sparse data

- "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

- "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 topic (2008?)

Brain Imaging and Machine Learning

ML Case study: learning models of individuals, of the population, and of individual variation

- "Training fMRI Classifiers to Discriminate Cognitive States across Multiple Subjects", X. Wang, R. Hutchinson, and T. M. Mitchell, *NIPS 2003*.
- "Classifying Multiple-Subject fMRI Data Using the Hierarchical Gaussian Naïve Bayes Classifier", Indrayana Rustandi, *13th Conference on Human Brain Mapping*. June 2007.
- Indra Rustandi PhD thesis topic (2008?)