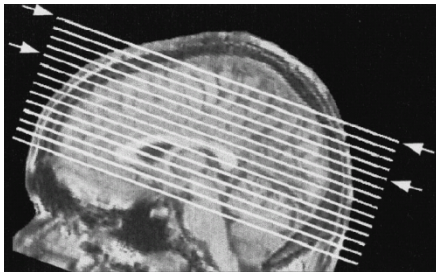
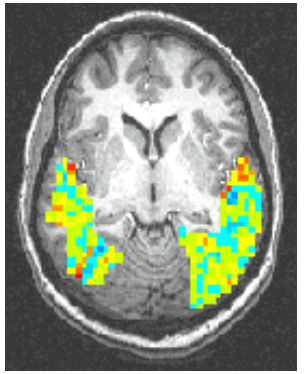
Brains, Meaning and Corpus Statistics



SIAM CSE 2009

Tom M. Mitchell

Machine Learning Department
Carnegie Mellon University



based in part on:

"Predicting Human Brain Activity Associated with the Meanings of Nouns," Mitchell, Shinkareva, Carlson, Chang, Malave, Mason, & Just, *Science*, 2008.

Computational Science: A Spectrum

Well-understood
component behaviors
and interconnections

Implement simulations,
run virtual experiments

Unknown
component behaviors
and interconnections

Discover models
from real experiments

Computational Science: A Spectrum

this talk: a case study in

Well-understood component behaviors and interconnections

Implement simulations, run virtual experiments

Unknown component behaviors and interconnections

Discover models from real experiments

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 the underlying principles of neural representations? (e.g., are representations built up from more primitive components?)

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

Functional MRI

functional Magnetic Resonance Imaging (fMRI)

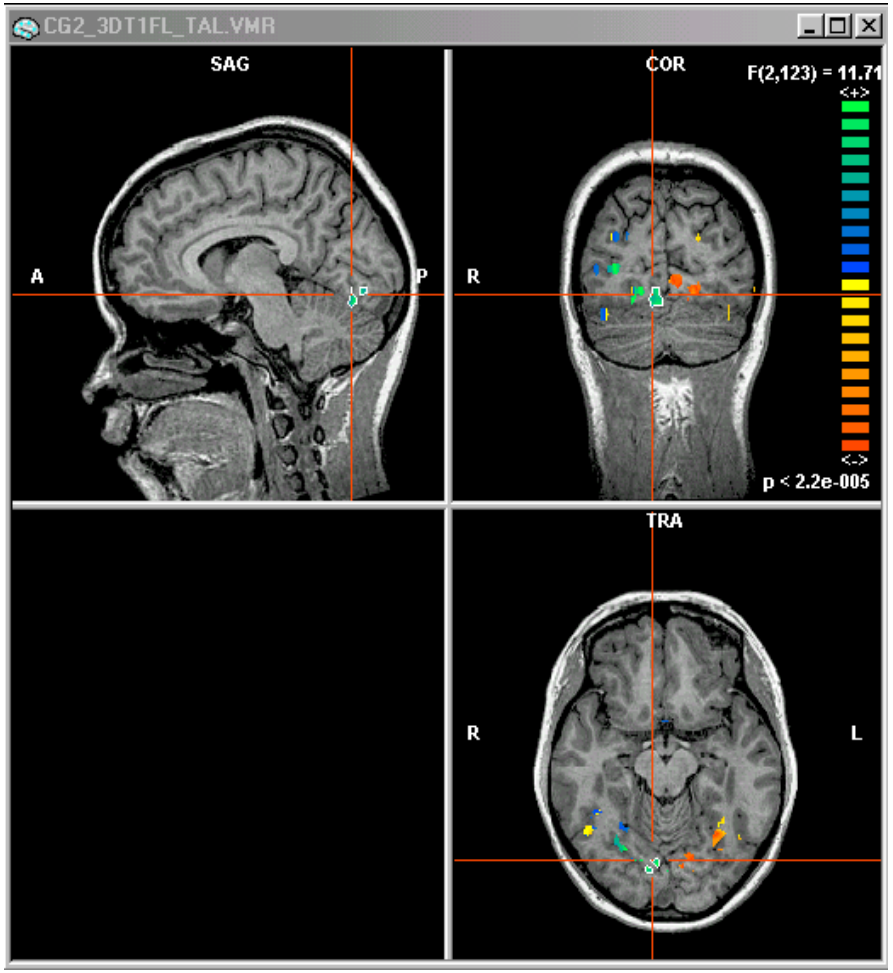
~1 mm resolution

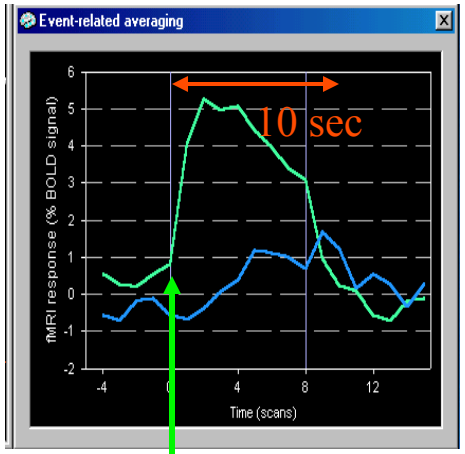
~1 image per sec.

20,000 voxels/image

safe, non-invasive

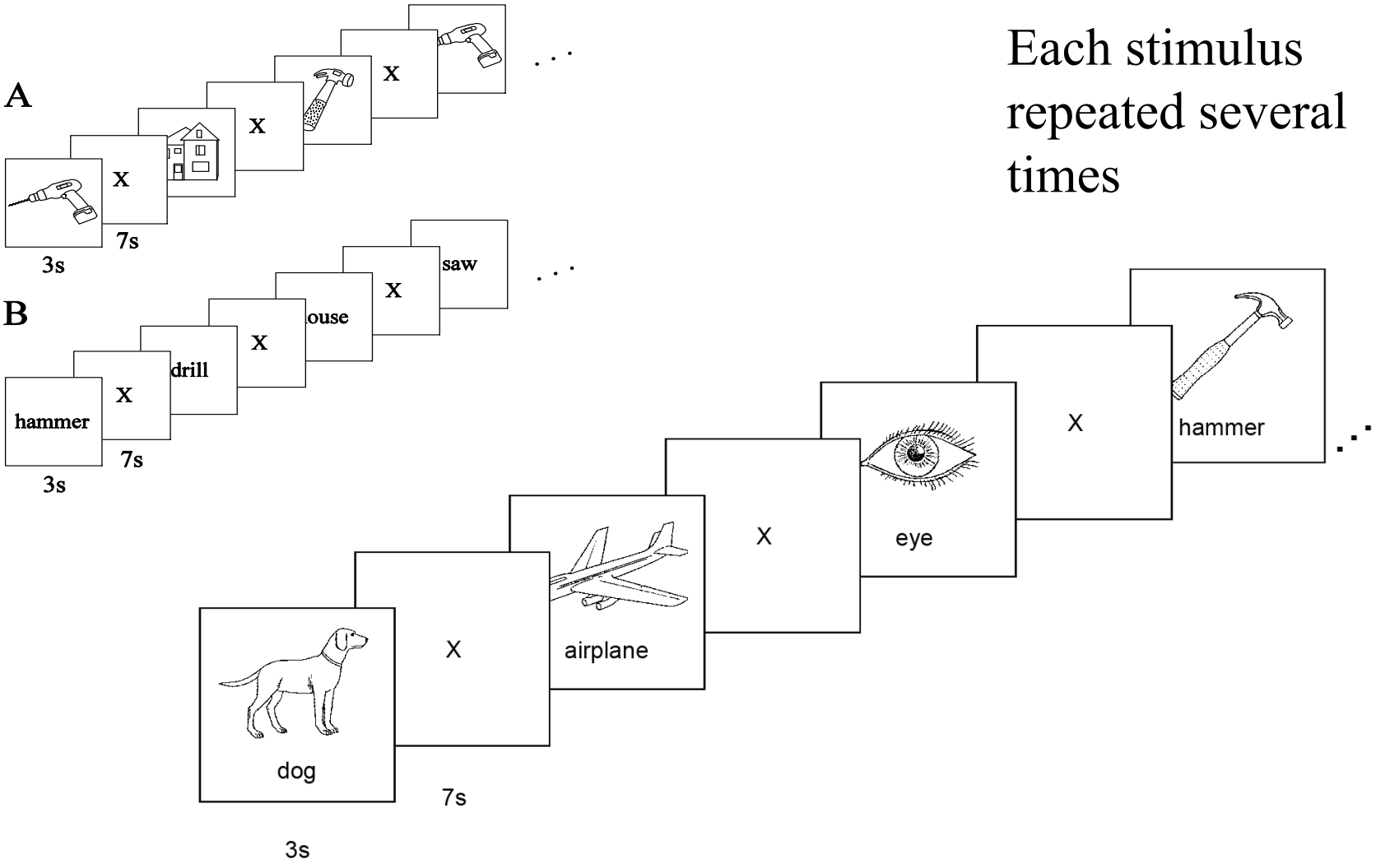
measures Blood Oxygen Level Dependent (BOLD) response





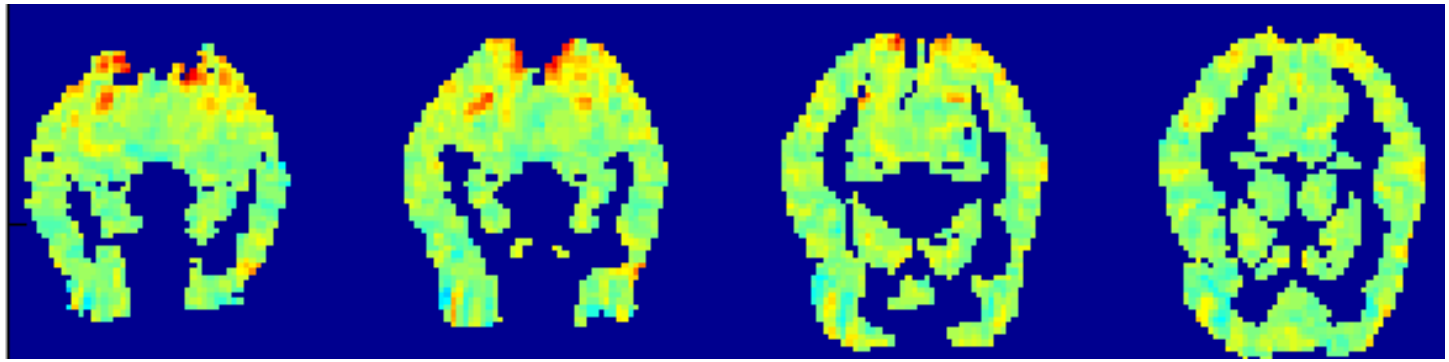
Typical fMRI response to impulse of neural activity

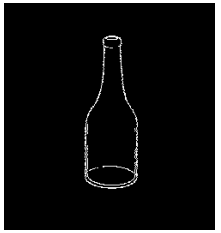
Typical stimuli



Each stimulus repeated several times

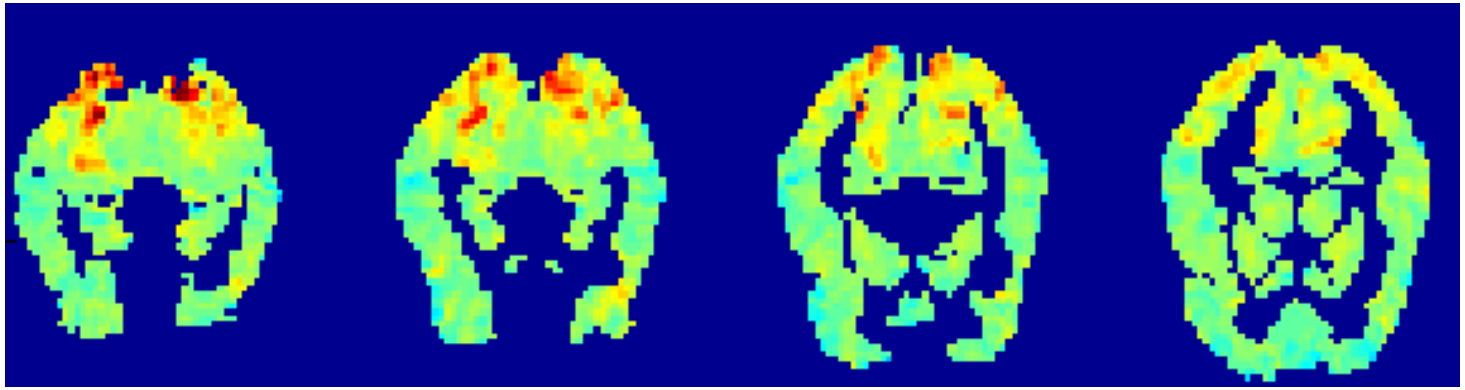
fMRI activation for “bottle”:



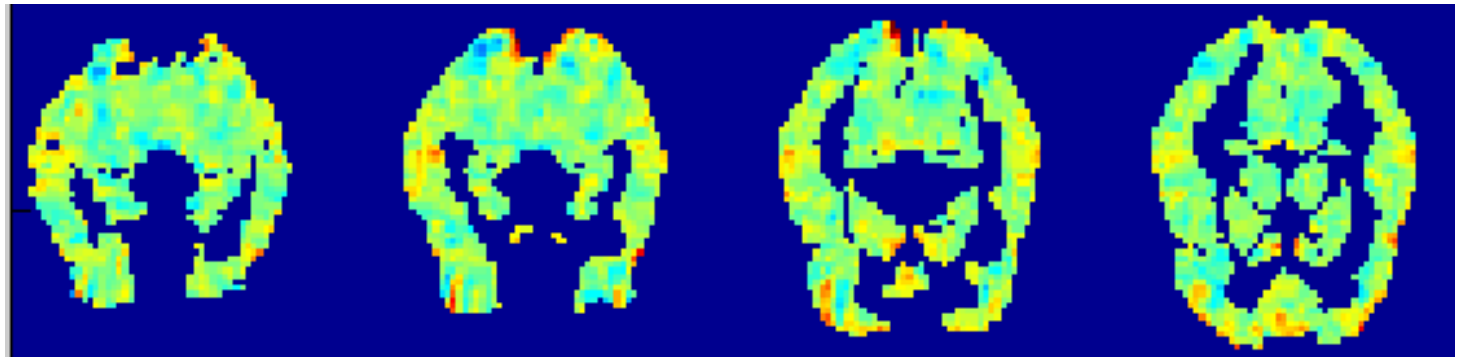


bottle

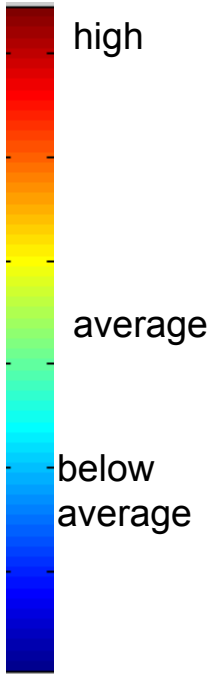
Mean activation averaged over 60 different stimuli:



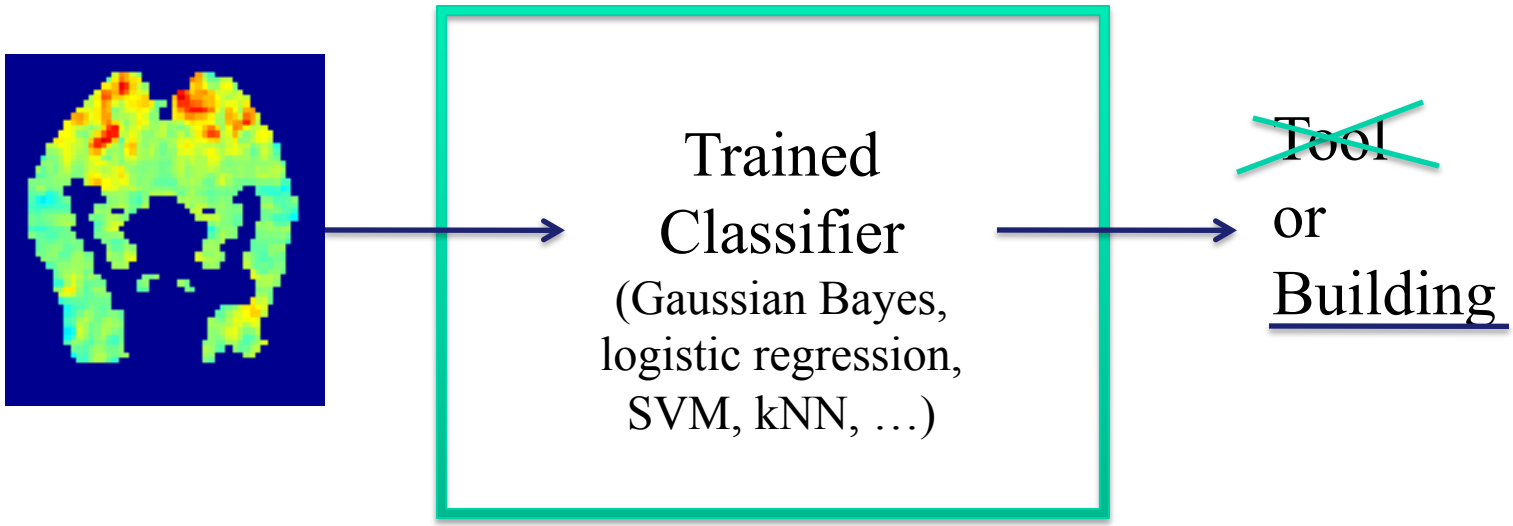
“bottle” minus mean activation:



fMRI 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

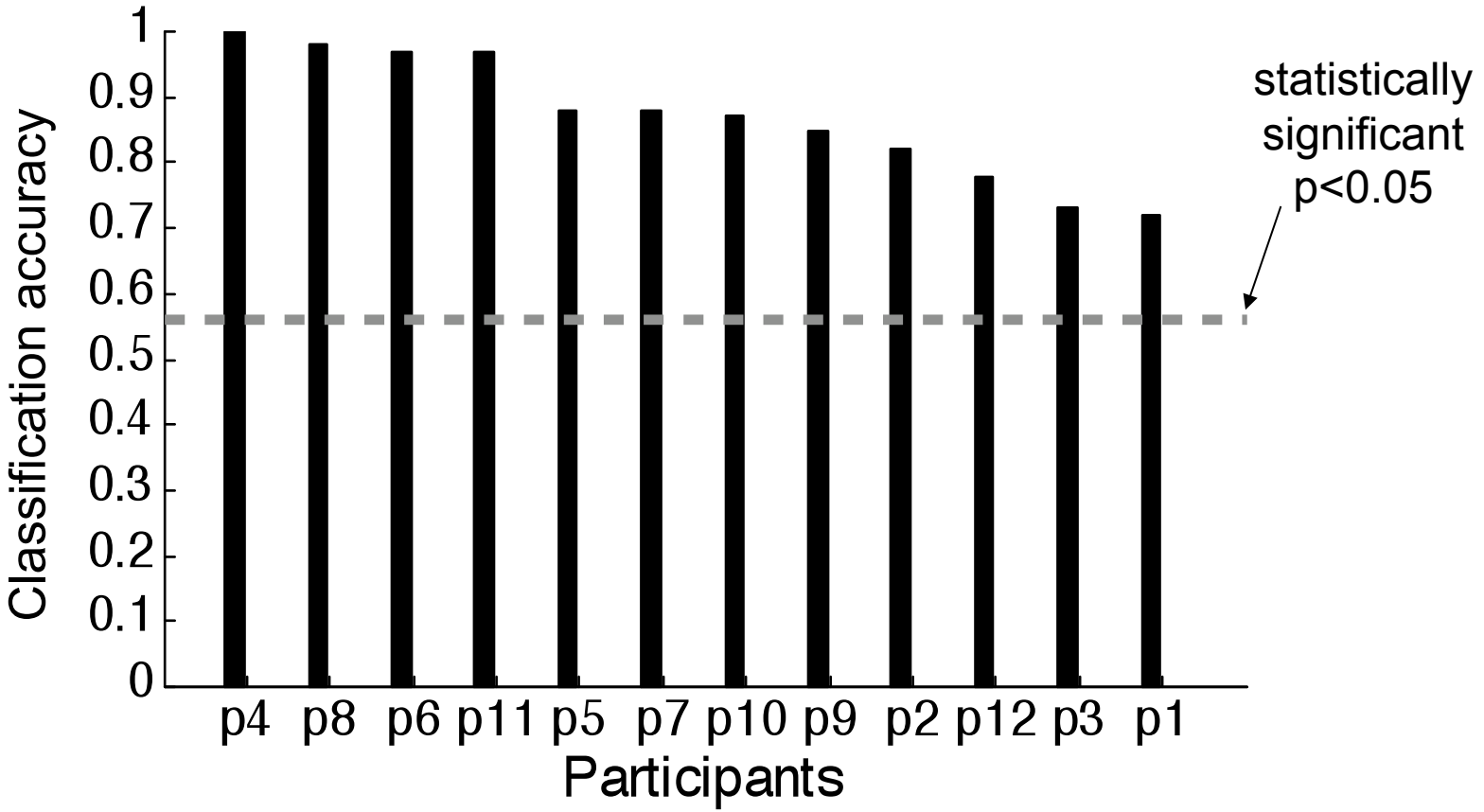
  Mean(fMRI(t+4), ...,fMRI(t+7)) → WordCategory
  - Leave one out cross validation over 84 word presentations

- 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

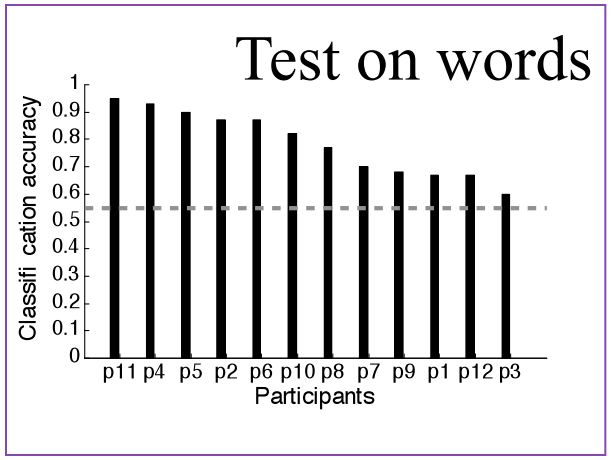
ML Case study: learning many related classifiers

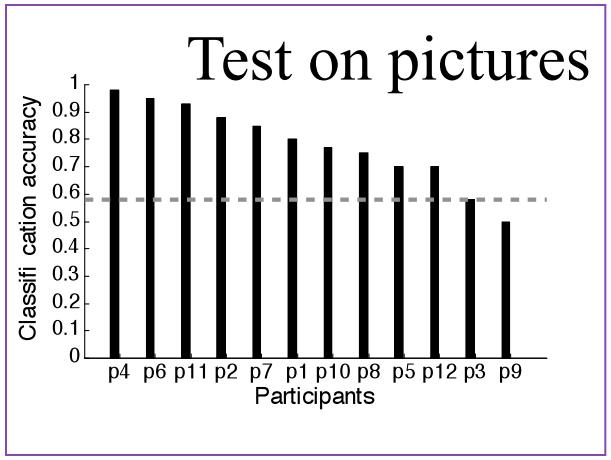
- “Training fMRI Classifiers to Discriminate Cognitive States across Multiple Subjects,” X. Wang, R. Hutchinson, T. Mitchell, *NIPS2003*
- "Classifying Multiple-Subject fMRI Data Using the Hierarchical Gaussian Naïve Bayes Classifier“, Indrayana Rustandi, *13th Conference on Human Brain Mapping.* June 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 topic

Question 2: Is our classifier capturing neural activity about 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



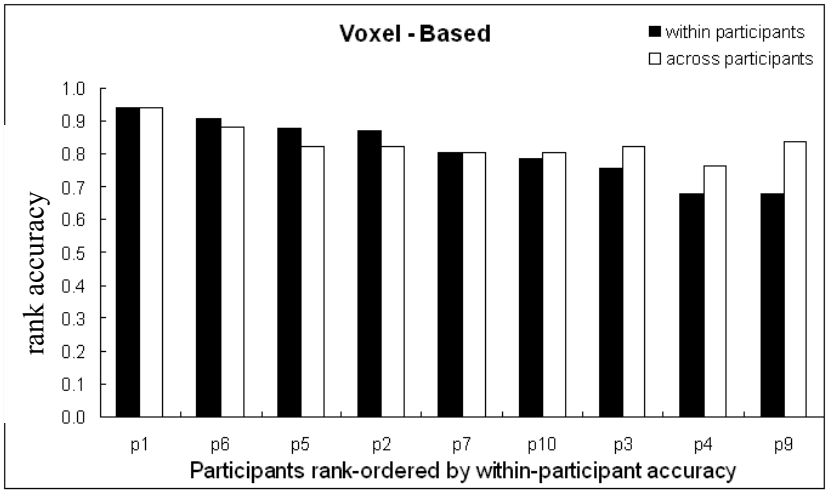


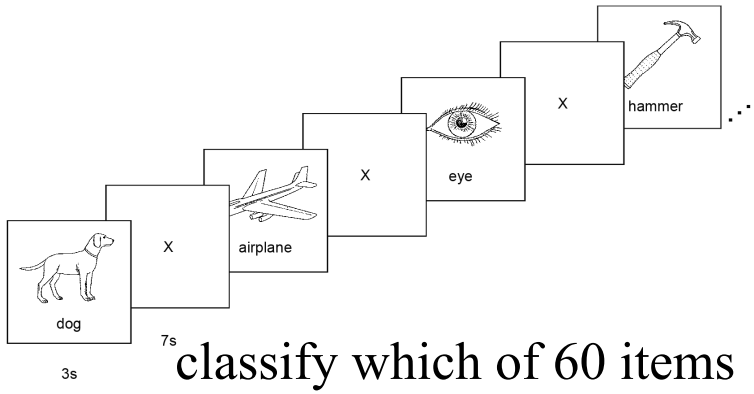
Therefore, the learned neural activation patterns must capture how the brain represents the meaning of input stimulus

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

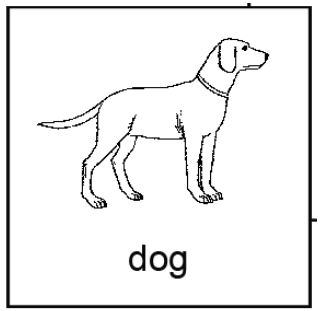




classify which of 60 items

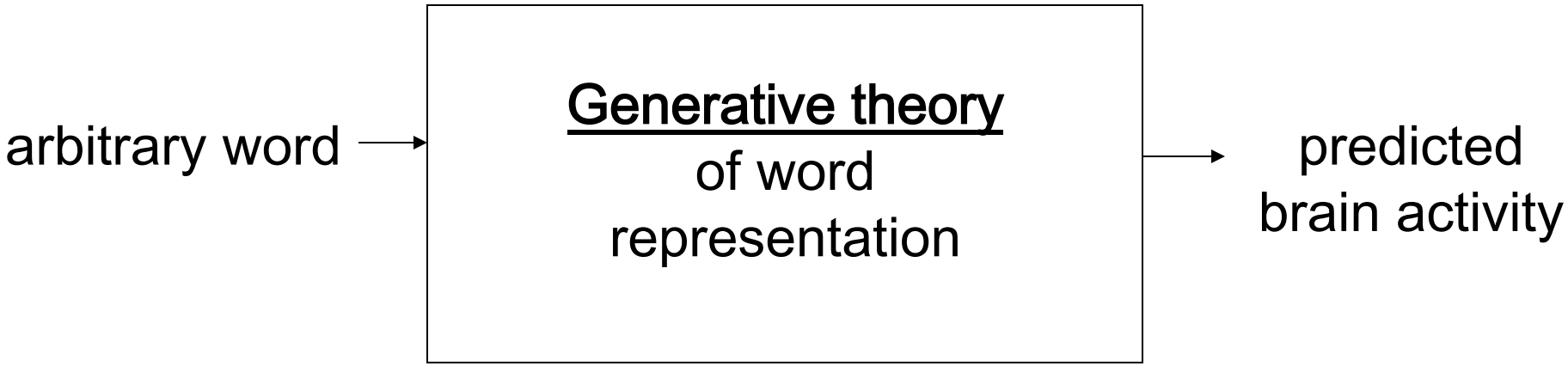
Therefore, seek a theory of neural representations common to all of us (and of how we vary)

60 exemplars



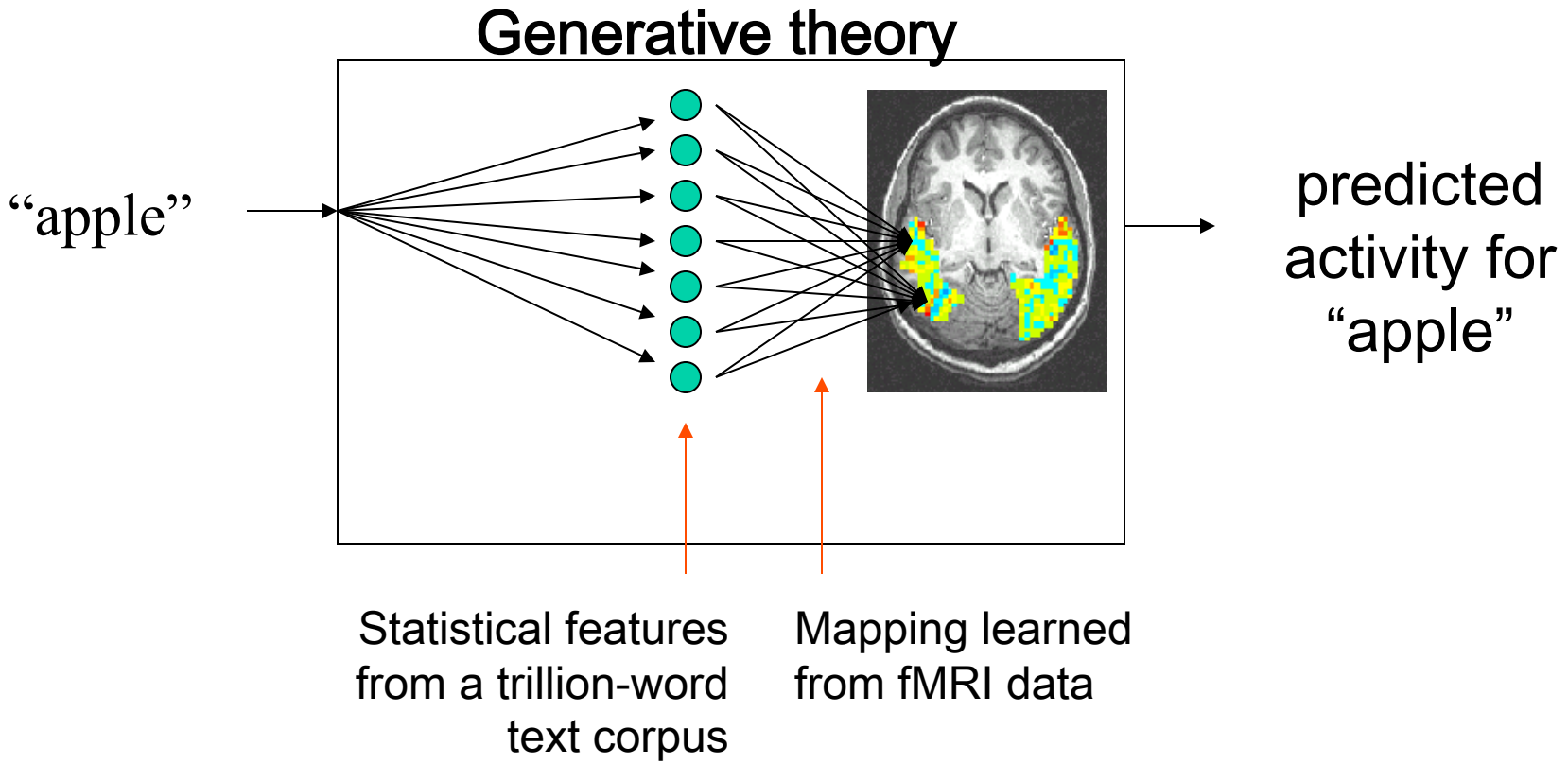
Categories	Exemplars				
BODY PARTS	leg	arm	eye	foot	hand
FURNITURE	chair	table	bed	desk	dresser
VEHICLES	car	airplane	train	truck	bicycle
ANIMALS	horse	dog	bear	cow	cat
KITCHEN UTENSILS	glass	knife	bottle	cup	spoon
TOOLS	chisel	hammer	screwdriver	pliers	saw
BUILDINGS	apartment	barn	house	church	igloo
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?

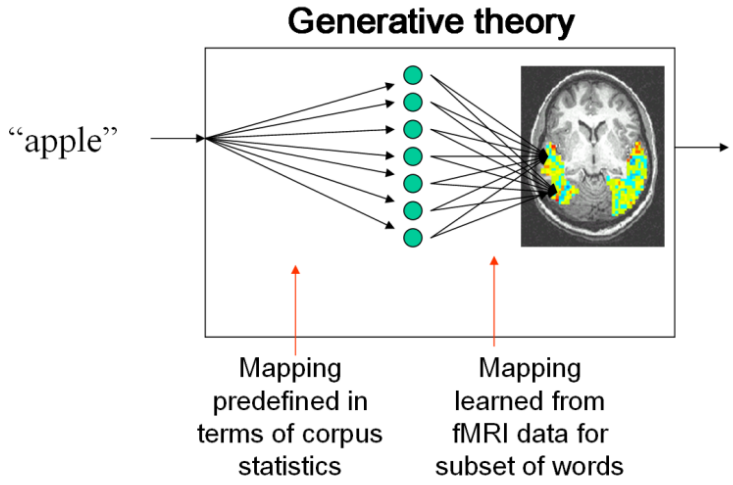


Idea: Predict neural activity from corpus statistics of stimulus word

[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?)

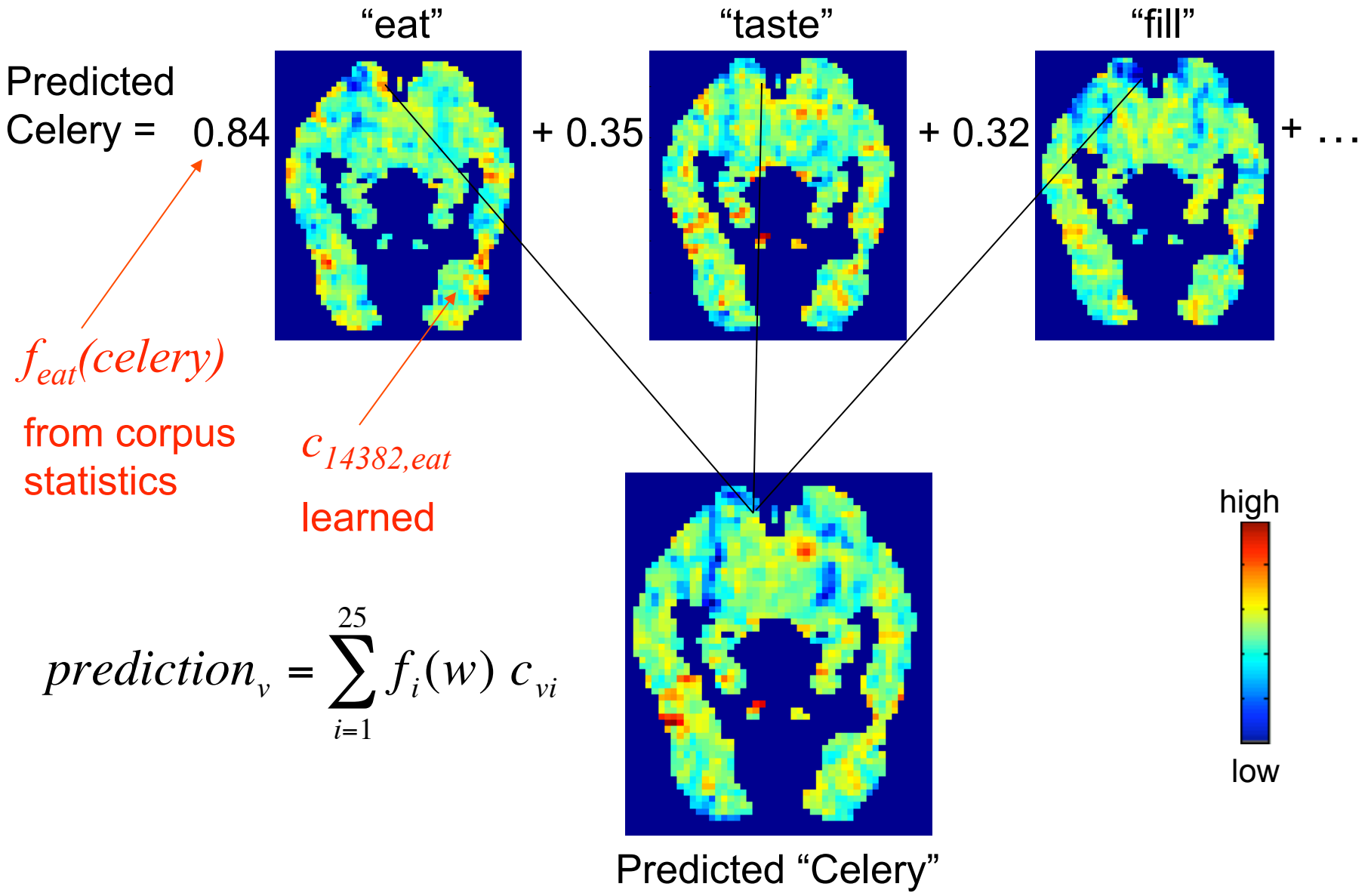
Semantic feature values: “**celery**”

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

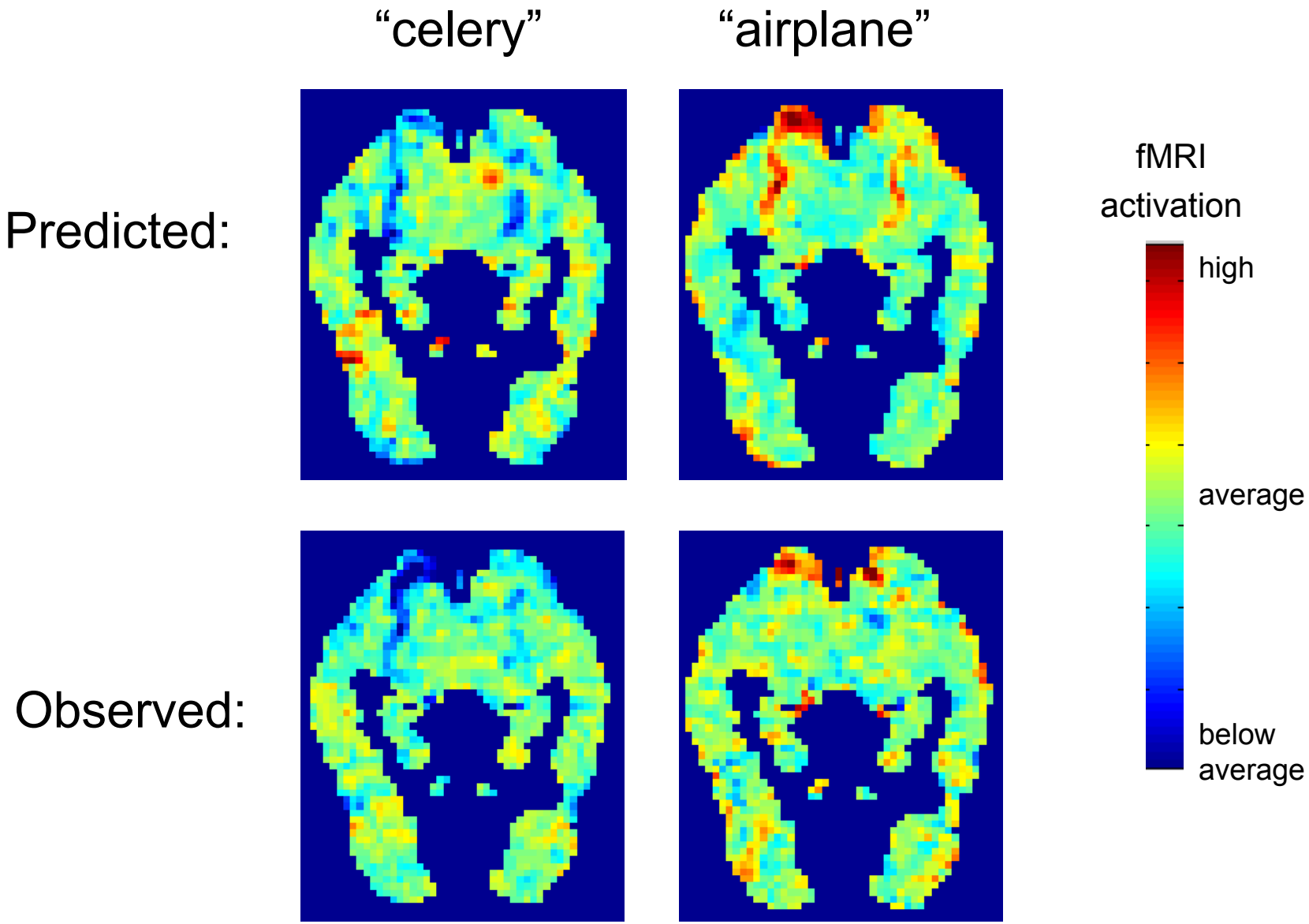
Semantic feature values: “**airplane**”

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



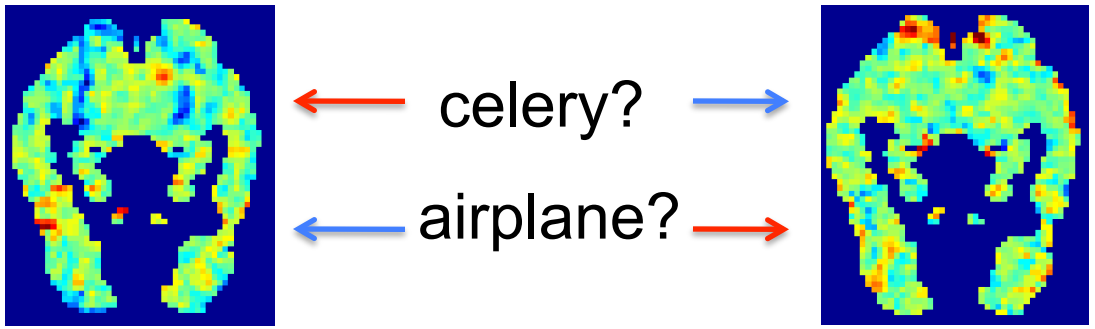
$$prediction_v = \sum_{i=1}^{25} f_i(w)\, c_{vi}$$



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

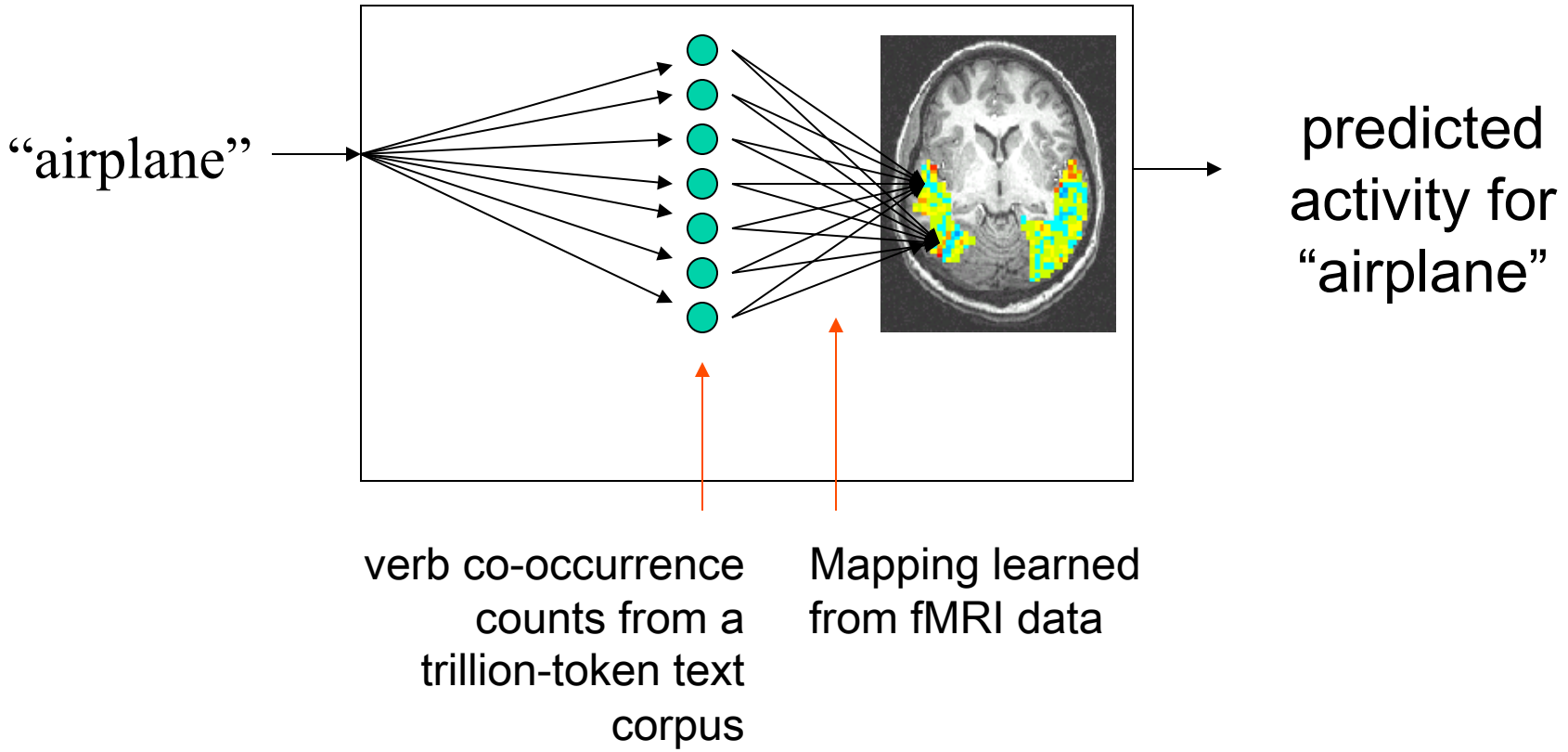


1770 test pairs in leave-2-out:

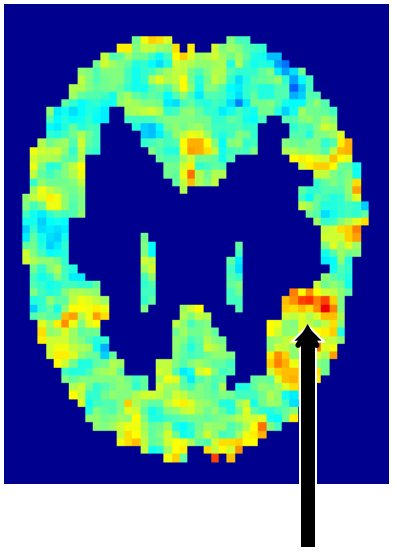
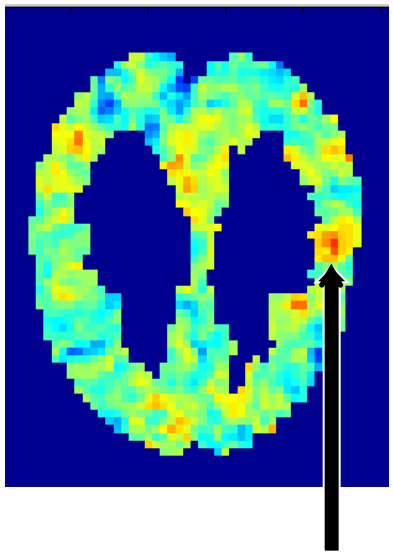
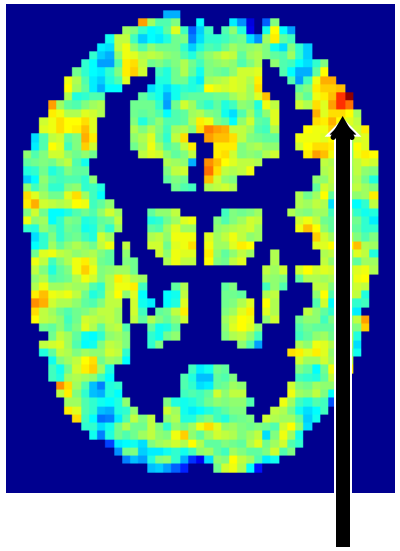
- Random guessing → 0.50 accuracy
- Accuracy above 0.61 is significant ($p<0.05$)

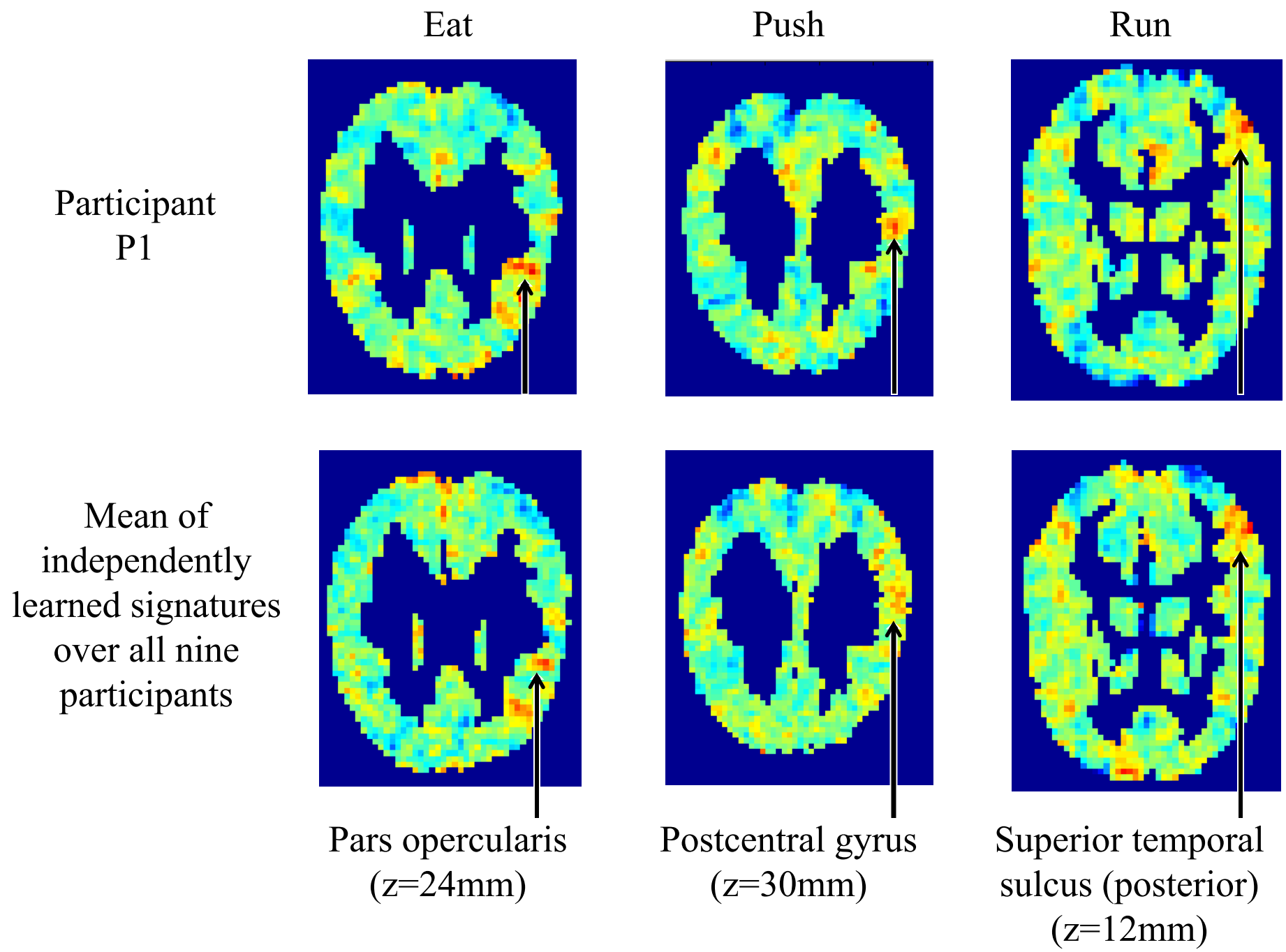
Mean accuracy over 9 subjects: 0.79

What are the learned semantic feature activations?



Participant P1

Eat	Push	Run
		
“Gustatory cortex”	“Planning motor actions”	“Body motion”
Pars opercularis (z=24mm)	Postcentral gyrus (z=30mm)	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)

Can we train the inverse model? Predict word from fMRI image?

[Palatucci, Hinton]

- Train a regularized linear regression model to predict co-occurrence frequencies from fMRI image

- Given a test fMRI image,
  - Predict its co-occurrence vector (text statistics)
  - Compare it to the vectors for the 1000 most frequent English words (skipping the 300 most frequent), plus the correct word
  - Observe percentile rank of correct word in this list of 1001

Top 5 Predictions out of 1001 candidate words (P1)

Truck (.998)	**Knife (1.0)**	**Desk (.977)**	**Arm (.998)**
auto	*knife*	table	hand
insurance	tool	bed	table
truck	seen	furniture	*arm*
airport	let	rental	turn
rental	hardware	parts	him

Skirt (1.0)	**Spoon (.975)**	**Lettuce (.988)**	**Airplane (.996)**
skirt	hand	blue	parts
men	tool	green	auto
blue	seen	red	rental
bed	hardware	brown	road
shoes	bed	color	*airplane*
			airport

Experimental Accuracy – participant P1

Mean Rank Accuracy: **0.868 (133/1001)**
Median Rank Accuracy: **0.980 (21/1001)**
averaged over all subjs: **0.865**

Accuracy: **12% (0.1% Random)**

Perfect-Rank 1

pants
skirt
knife
hammer
screwdriver
carrot
celery

Excellent - Rank 2-11

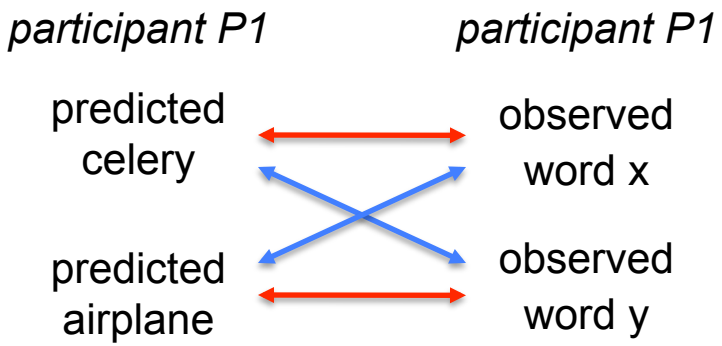
dog
arm
foot
apartment
horse
barn
dress
glass
pliers
tomato
airplane
car
truck

Worst – (Rank > 400)

igloo
arch
chimney
bee
watch
corn
eye

What is the source of the error in fMRI predictions?

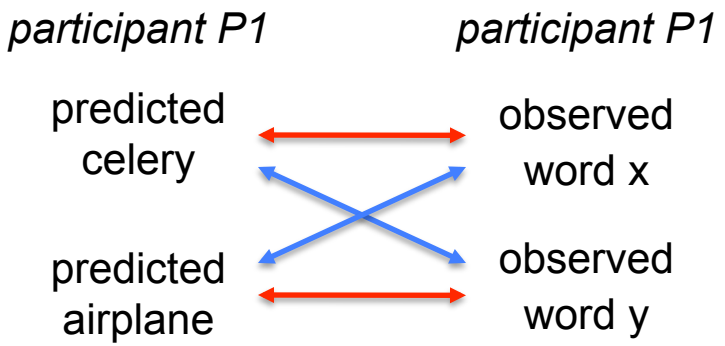
- Accuracy 0.79 → error 0.21
- Insufficient semantic features and model?
- Noise in fMRI data?



What is the source of the error in predictions?

- Insufficient semantic features?
- Noise in fMRI data?

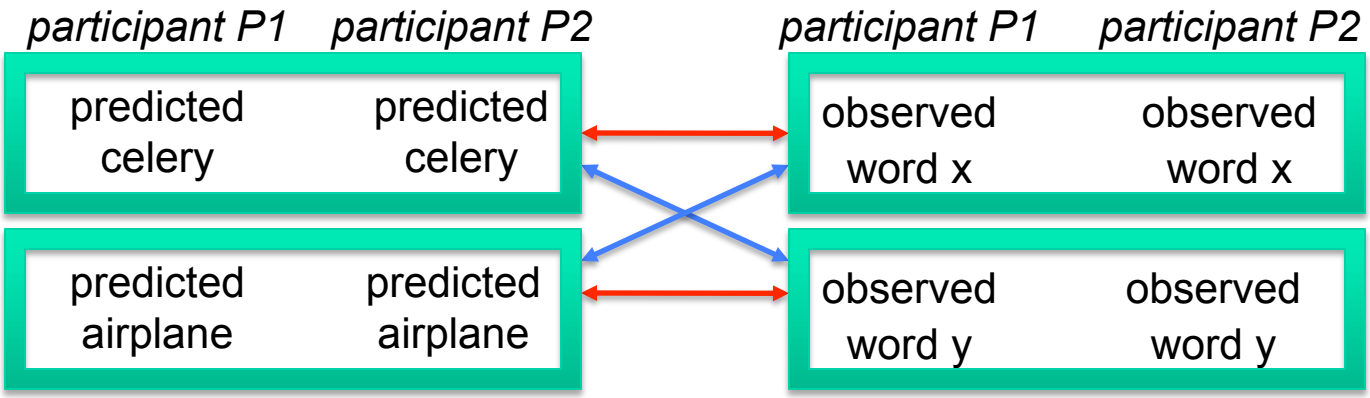
***Idea**: concatenate predicted and observed images for multiple participants*



What is the source of the error in predictions?

- Insufficient semantic features?
- Noise in fMRI data?

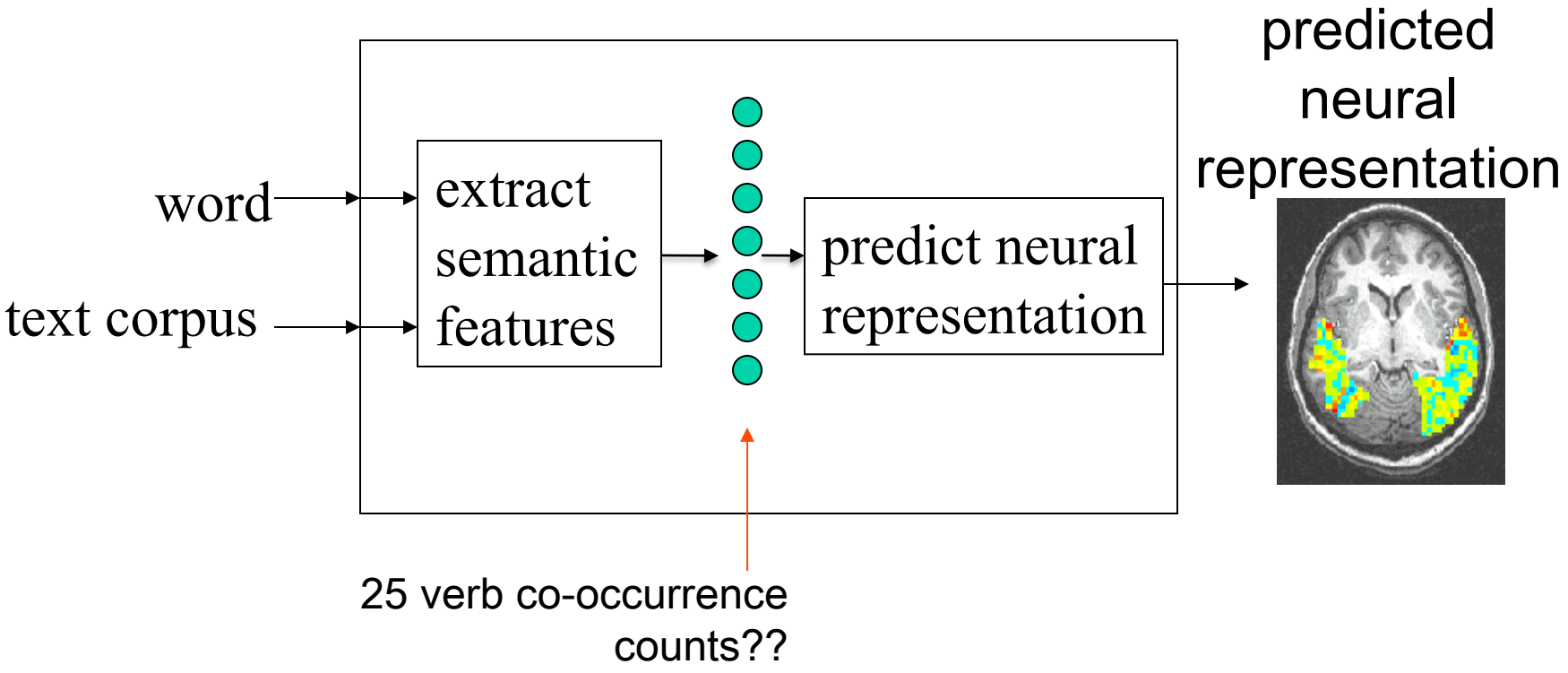
***Idea**: concatenate predicted and observed images for multiple participants*



original accuracy: 0.79

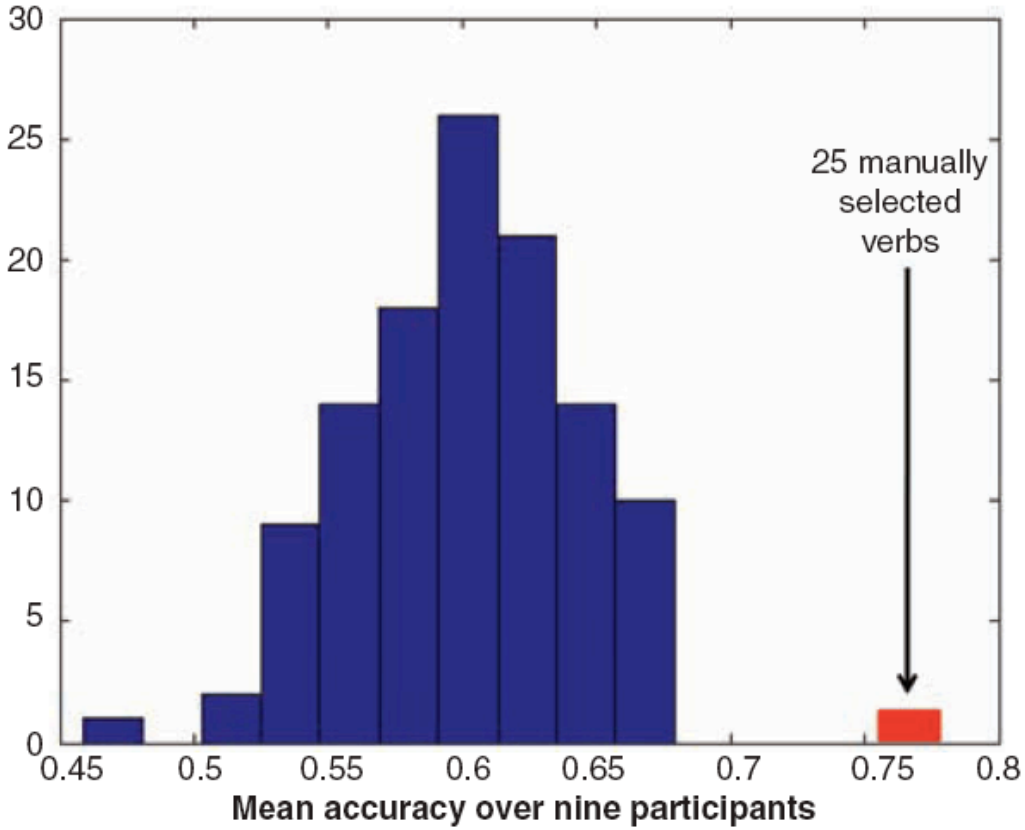
concatenating all 9 particpants: 0.87

Q: What is the semantic basis from which neural encodings are composed?



How Unique is our set of 25 verb features?

Empirical distribution of accuracy for 115 random feature sets



Features were drawn uniformly at random without replacement from the 5000 most frequent words, omitting the 500 most frequent.

Alternative semantic feature sets

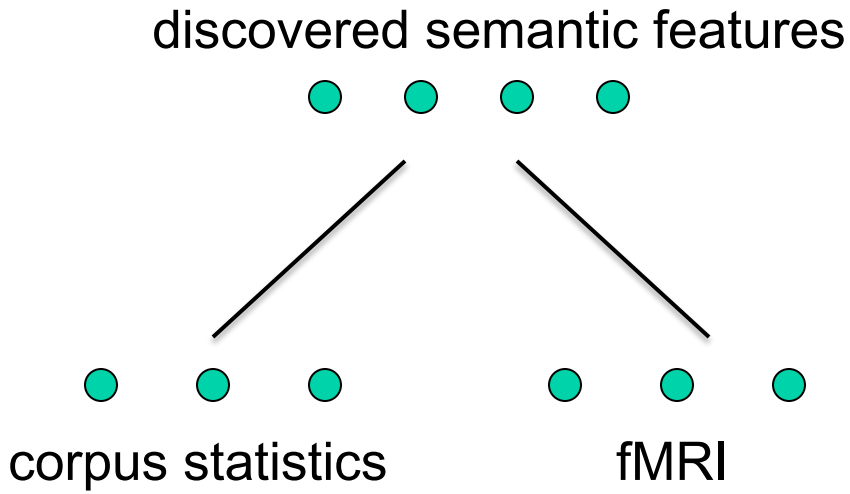
PREDEFINED corpus features	Mean Acc.	Acc. cat9
25 verb co-occurrences	.79	.87
486 verb co-occurrences	.79	**.89**
50,000 word co-occurences	.76	.86
300 Latent Semantic Analysis features	.73	.81
50 corpus features from Collobert&Weston ICML08	.78	--

Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.	Acc. cat9
25 verb co-occurrences	.79	.87
486 verb co-occurrences	.79	**.89**
50,000 word co-occurences	.76	.86
300 Latent Semantic Analysis features	.73	.81
50 corpus features from Collobert&Weston ICML08	.78	--

DISCOVER features
(projected space shared
by fMRI, corpus statistics)

Algs: GSVD, CCA



Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.	Acc. cat9
25 verb co-occurrences	.79	.87
486 verb co-occurrences	.79	**.89**
50,000 word co-occurences	.76	.86
300 Latent Semantic Analysis features	.73	.81
50 corpus features from Collobert&Weston ICML08	.78	--

DISCOVERED corpus features (~58)	Mean Acc.	Acc. cat9
GSVD features jt. analysis of 25/486/50k wds, fMRI	.79 / .74 / .69	.87 / .85 / --
CCA features joint analysis of 25 vb and fMRI	.78	
CCA features, jt. analysis of 25/486 verbs, fMRI	**.81** / .79	.88 / **.89**
sparse CCA, jt. analysis of 25/486 verbs, fMRI	.78 / **.81**	.86 / **.89**
CCA **top 10** features, jt. analysis of 50k wds, fMRI	.77	**.92**

[Indra Rustandi]

Summary

[Indra Rustandi]

Data set	Word-Picture stimuli (9 subjs)	Word-only stimuli (11 subjs)
25 verb co-occurrences	.87	.85
CCA top 10 features, jt. analysis of 50k wds, fMRI	**.92**	**.93**

What next for Machine Learning challenges?

- ML: discover optimal features to replace the 25 verbs
  - discover low-dimensional manifold for both corpus and fMRI
- ML: algorithm to learn cumulatively, from multiple studies with different words, people
  - must discover latent features with geometric biases
- ML: train using fMRI (1 mm) and MEG (1 msec)
  - *fuse data sources* and *train classifier, predictor*
- ML: study corpus statistics to generate conjectures about neural representations
  - especially to study representations of multi-word phrases (e.g., *fast rabbit* versus *hungry rabbit*)

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 results: successful (74%) classifier for “food” vs. “body parts”

Case Study of Computational Science

- Train classifiers as virtual sensors of system state
  - determine which part(s) of system contain information
  - classifiers plus changing input to system allow studying its properties

- Predictive models
  - characterize regularities even if not causality
  - based on cross-domain data (text, fMRI)

- Latent variable models
  - discover latent structure of system model

  (e.g., semantic features that underlie neural code)

thank you