WORDS

WORDS-AS-FEATURES AS MODELS OF COGNITION. MITCHELL ET AL. 2008

CORE QUESTIONS: NEUROSCIENCE

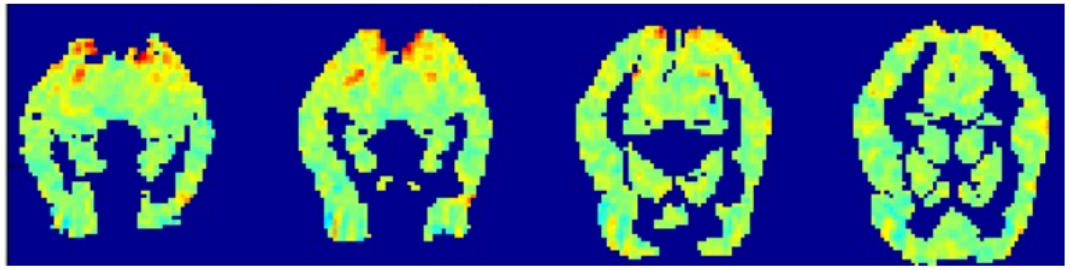
- Are there systematic differences in neural activity as people think about different concepts?
- Is the neural representation of concepts localized in specific brain areas or is it distributed across the entire cortex?
- How meaningful are individual differences, or is the representation of meaning similar across people

CORE QUESTIONS: AI / NLP

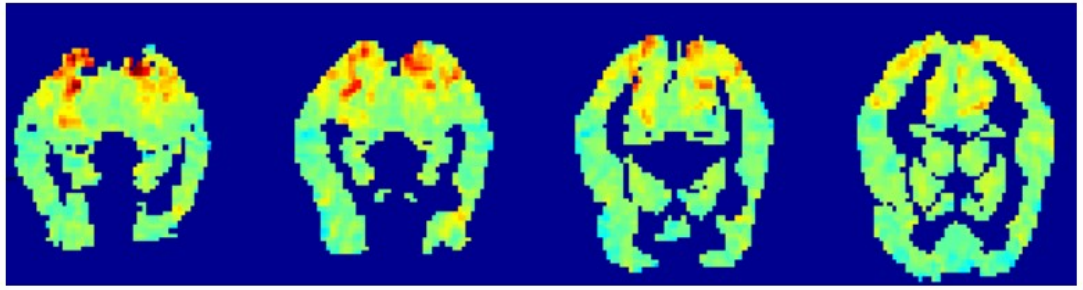
- Can fMRI and neuroscience allow us to ‘test’ or ‘understand’ what are the basis functions (the semantic feature space) that underlies the representation of words?
- If so: demonstration of “empirical NLP”; use behavior to guide construction of more cognitive real computational models.

ACTIVITY FOR SINGLE WORD

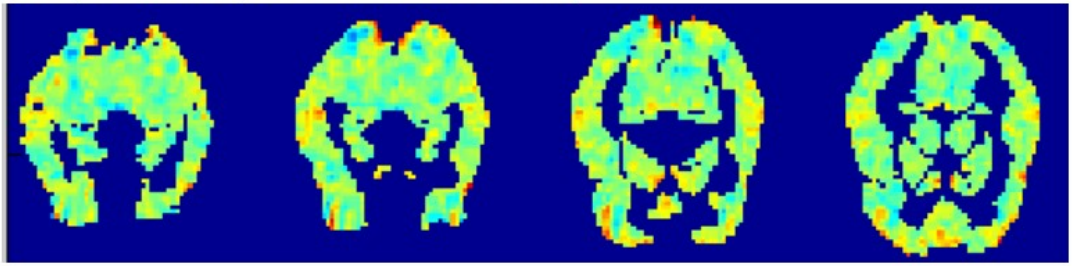
fMRI activation for "bottle":



Mean activation averaged over 60 different stimuli:

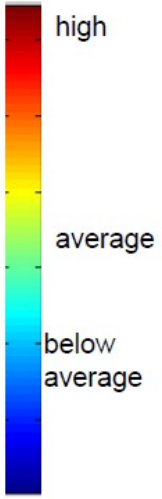


"bottle" minus mean activation:



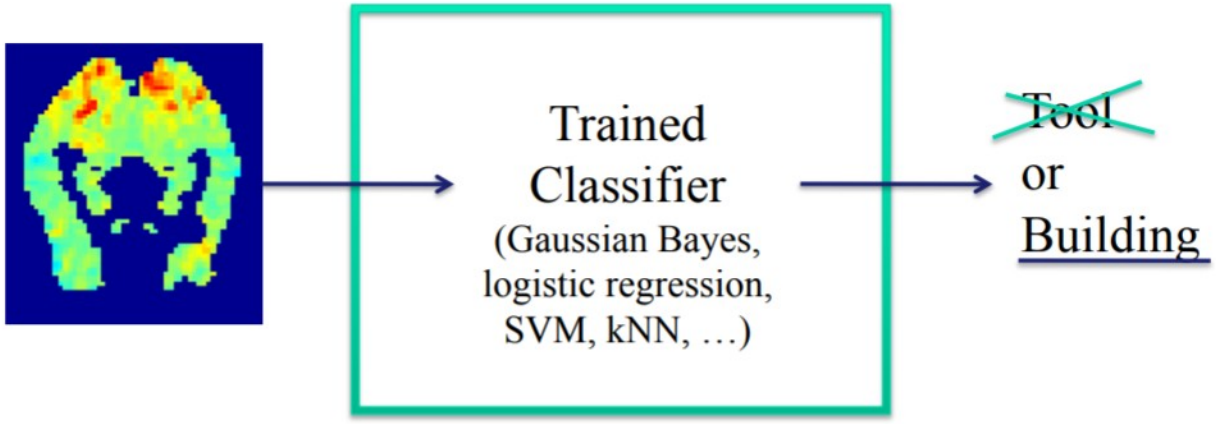
bottle

fMRI activation



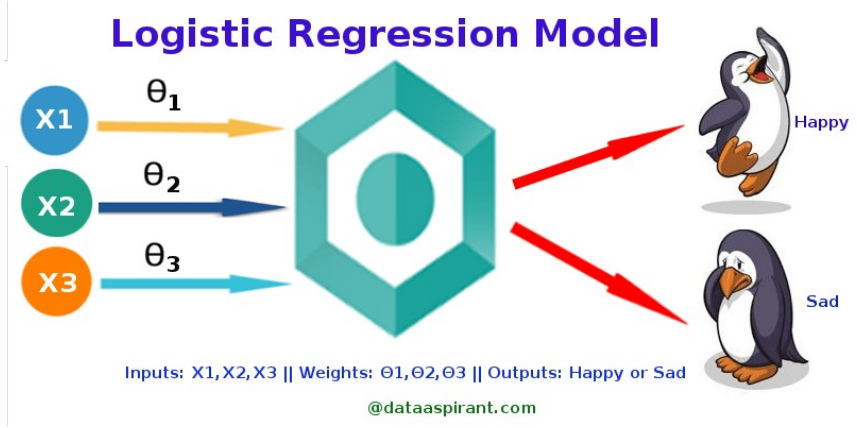
HISTORICAL APPROACH (NOT IN PAPER)

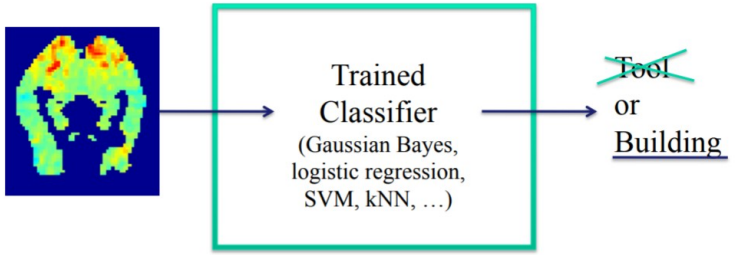
- Present multiple words sampled from several categories (e.g., Tools, Buildings).
- Train a classifier that predicts the class (Tool/ Building) from the brain images of the words
- The results of classifier can be used as a tool for studying the semantics in the brain. E.g., which brain area contain information about particular classes.



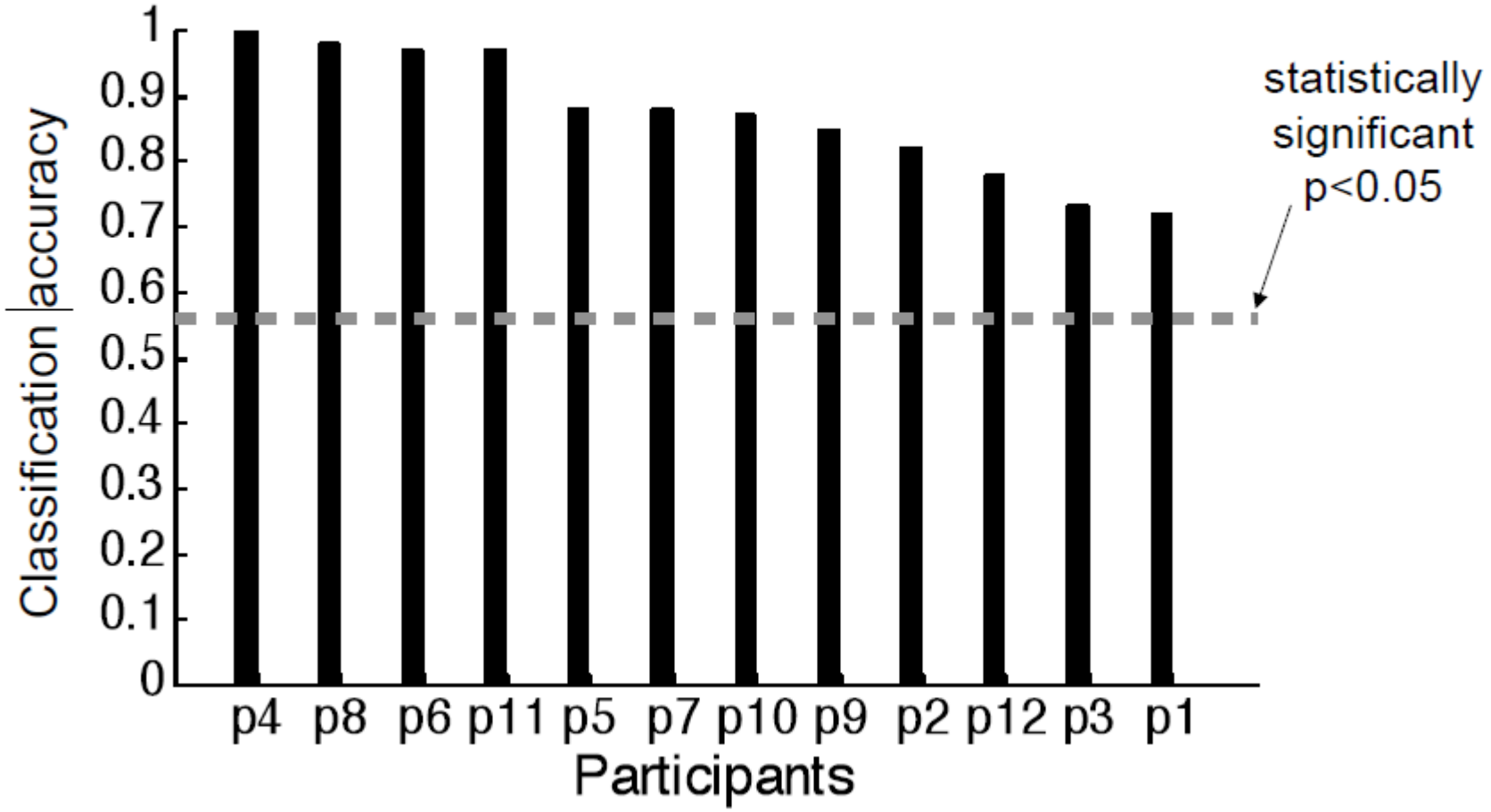
HISTORICAL APPROACH (NOT IN PAPER)

- Present multiple words sampled from several categories (e.g., Tools, Buildings).
- Train a classifier that predicts the class (Tool/ Building) from the brain images of the words
- The results of classifier can be used as a tool for studying the semantics in the brain: e.g., for logistic regression, which voxels have strong weights loading on 'building' response
- This is a pure 'decoder'; there is no domain-based knowledge that is applied to predict the brain response from more basic principles





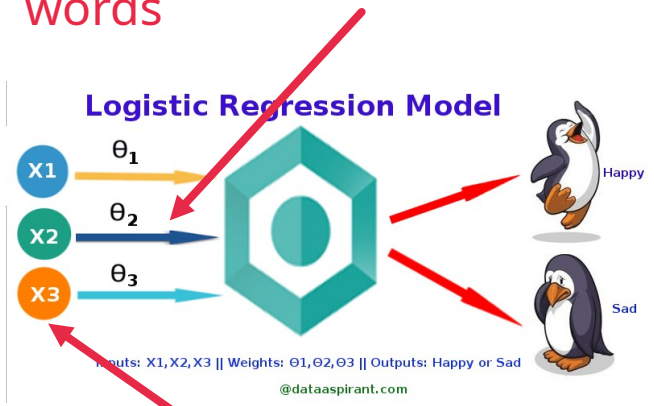
CLASSIFICATION/DECODING WORKS: TOOL OR BUILDING?



CLASSIFIERS CAPTURE SOME MEANING: CROSS-DOMAIN GENERALIZATION (TRAIN ON WORDS, GUESS CLASS OF IMAGE)

- They also have:
  - brain activity while Portuguese people watch the same words;
  - people watching images corresponding to the word-set
- They now take the results of the classifier trained to discriminate categories based on brain responses to **words presented in English** and input apply it to those other datasets.
- Both ‘Testing on pictures’ and ‘Testing on other language’ produce above chance accuracy: **semantics generalize beyond modality used**.

Step 1: Train weights from responses to English words



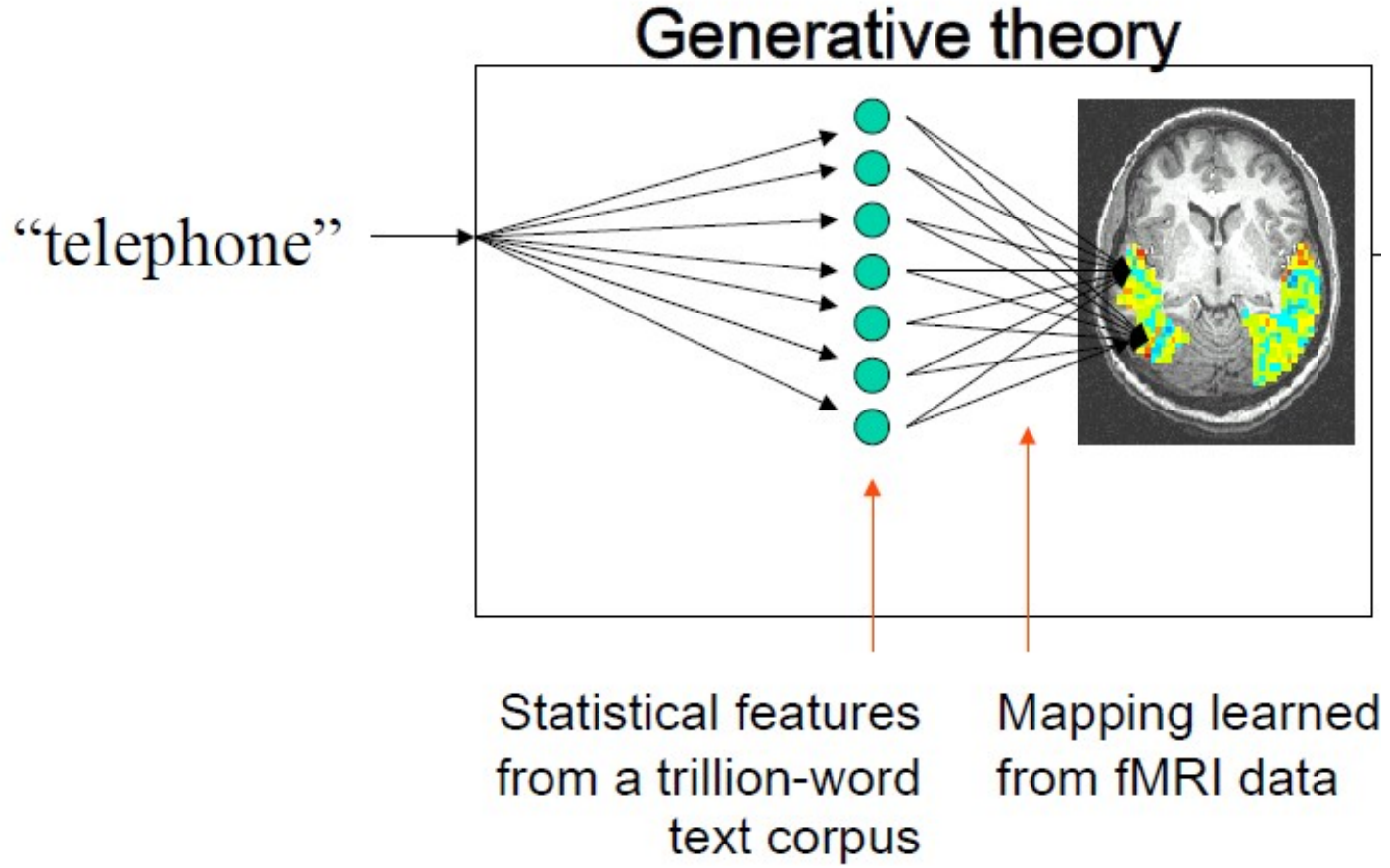
Step2: Test on brain activity from other domains (untrained)

DECODING WORKS, BUT HAS A PROBLEM

- Data are highly dimensional : over 20K features (voxels) per word. But at the same time sparse (only few examples of brain activity per category)
  - This is difficult from a regression perspective and solutions to this (regularization) are mathematically valid but may lose information about the brain
- Requires specialized solutions, “Classification in Very High Dimensional Problems with Handfuls of Examples", M. Palatucci and T. Mitchell, ECML-200

THE ALTERNATIVE, A 'GENERATIVE' ENCODING MODEL

- Come up with a theory of word meaning and see whether the theory predicts brain responses
- They capture 'word meaning' from corpus statistics.
- Challenge: find a 'mapping function' from word meaning to brain activity



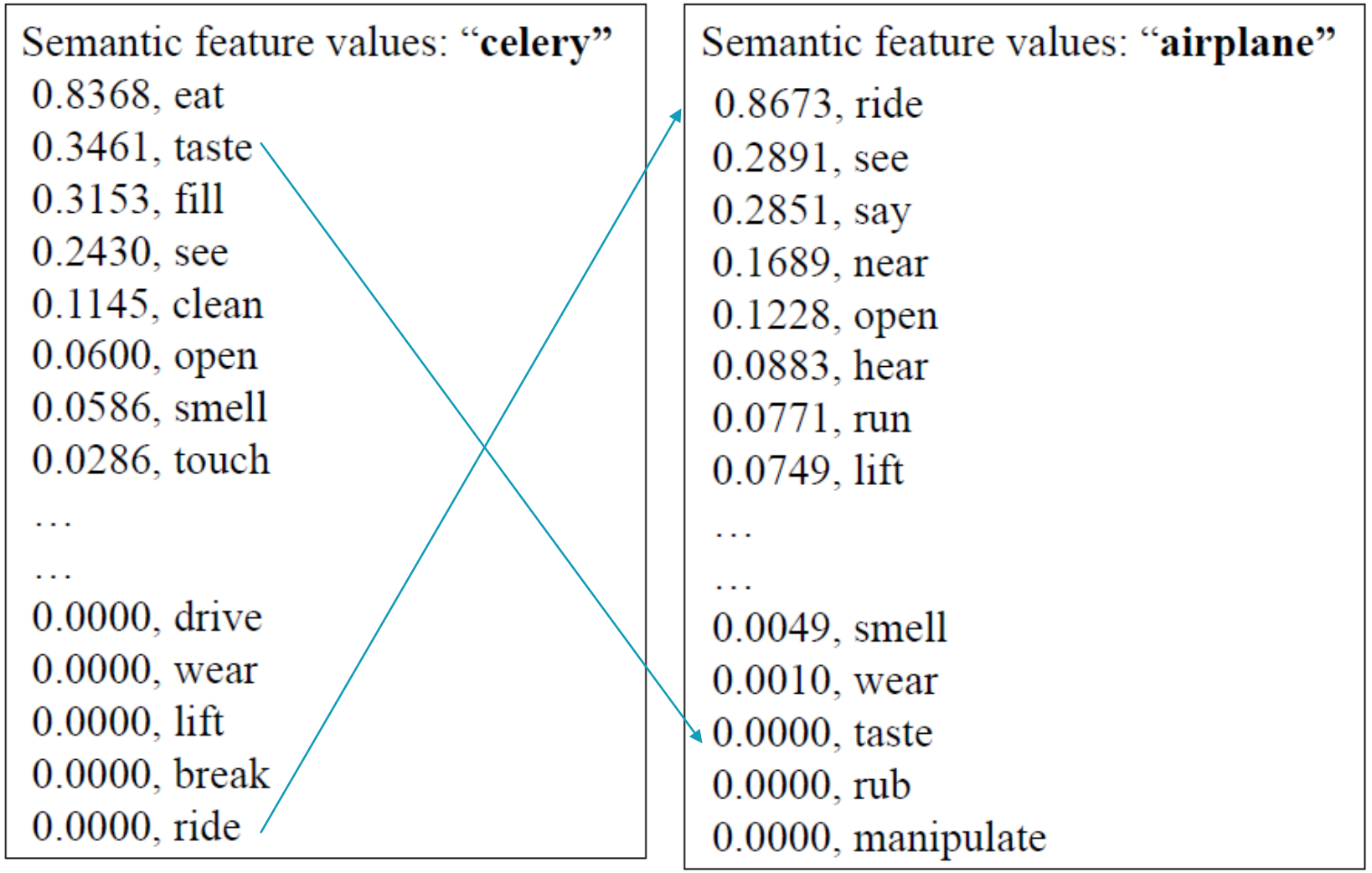
STEP 1: DECIDE ON HOW TO DESCRIBE THESE NOUNS FROM CORPUS STATISTICS.

dog

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

STEP 1: DECIDE ON HOW TO DESCRIBE THESE NOUNS FROM CORPUS STATISTICS.

- Each of their Nouns j will be described using 25 features. i= 1:25
- Feature i = co-occurrence frequency of stimulus noun with verb j
- 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


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

A SINGLE-VOXEL ANALYSIS

- For each voxel in the brain they learn the relation between voxel activity values for the 60 nouns, and the semantic features of the 60 nouns
- “The second step predicts the neural fMRI activation at every voxel location in the brain, as a weighted sum of neural activations contributed by each of the intermediate semantic features.”

A SINGLE-VOXEL ANALYSIS

In matrix notation the multiple regression model is: $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

where

$$\mathbf{Y}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_\kappa \end{bmatrix}, \quad \mathbf{X}=\begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & & X_{nk} \end{bmatrix}$$

Activity for 58 nouns in the voxel

A SINGLE-VOXEL ANALYSIS

In matrix notation the multiple regression model is: $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

where

Set of 25 feature-value per noun

$$\mathbf{Y}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_\kappa \end{bmatrix}, \quad \mathbf{X}=\begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & & X_{nk} \end{bmatrix}$$

Activity for 58 nouns in the voxel

A SINGLE-VOXEL ANALYSIS

In matrix notation the multiple regression model is: $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

where

25 Weights to be fit

Set of 25 feature-value per noun

$$\mathbf{Y}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_\kappa \end{bmatrix}, \quad \mathbf{X}=\begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & & X_{nk} \end{bmatrix}$$

Activity for 58 nouns in the voxel

WHAT DOES THE SINGLE VOXEL ANALYSIS TELL US?

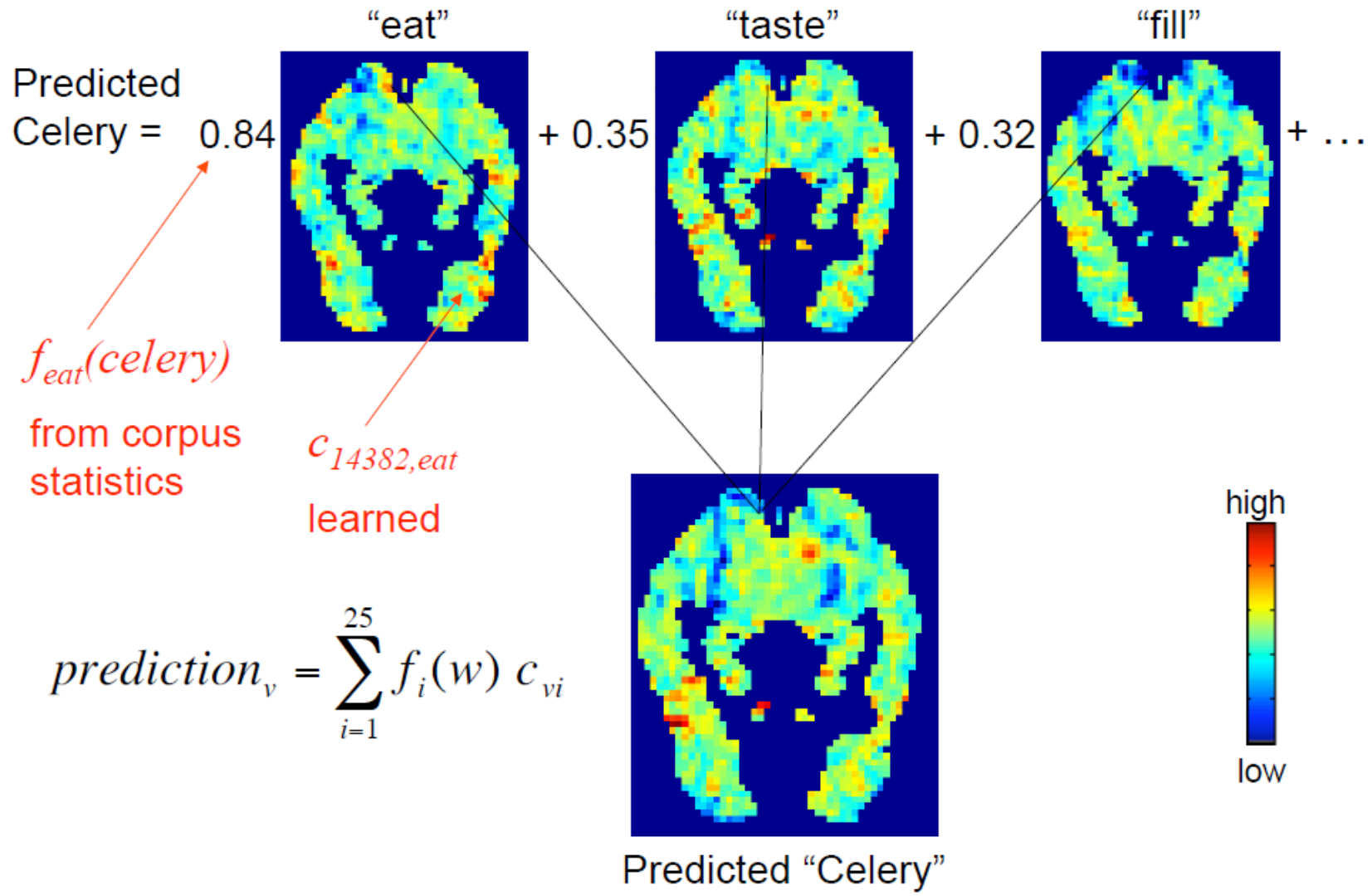
For each voxel, what is the relative importance of each of the 25 features!

We can learn this from 59 words and then ‘generate’ a predicted value of voxel activity for the 60th word.

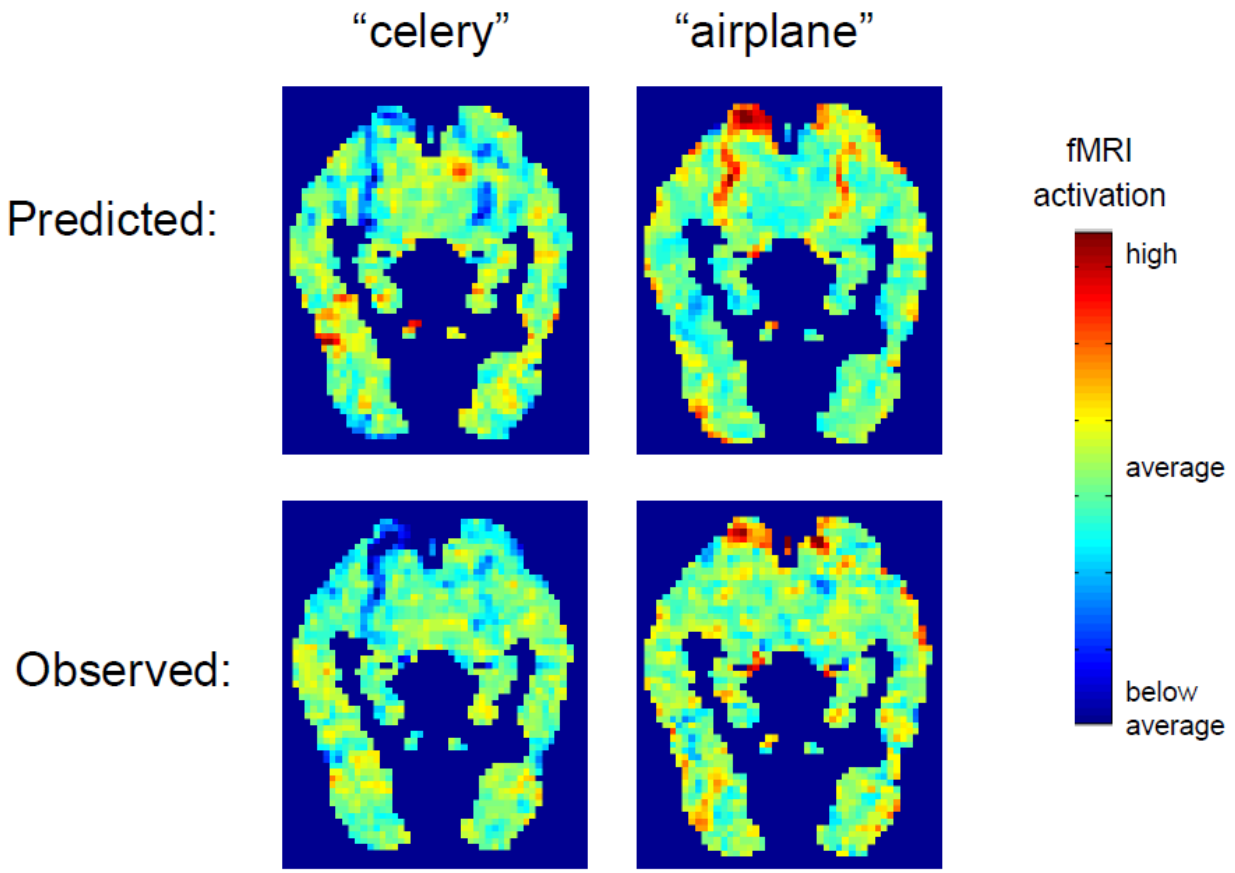
PREDICTING WORD ACTIVITY **IN EACH VOXEL** JUST MEANS MULTIPLYING ITS SEMANTIC FEATURE VALUES BY LEARNED WEIGHTS

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


"eat"
"taste"
"fill"
Predicted Celery =
0.84
+ 0.35
+ 0.32
+ . . .
$f_{eat}(celery)$
from corpus statistics
$c_{14382,eat}$
learned
$prediction_v = \sum_{i=1}^{25} f_i(w)\ c_{vi}$
Predicted "Celery"
high
low

HOW WELL DOES THIS WORK?



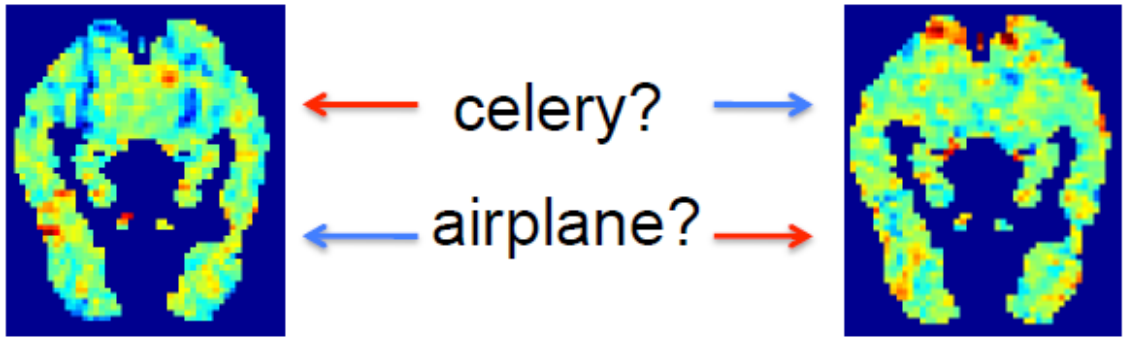
Predicted and observed fMRI images for "celery" and "airplane" after training on 58 other words.

HOW DO YOU TEST THE PREDICTION EMPIRICALLY?

A SIMPLE TWO CHOICE TEST

- Predict activity for 2 left out words
- From the ‘database’ pull the ‘true’ activity for those two words
- See if the predicted activity for word X is more similar to true activity of word X than it is to the other word

- 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



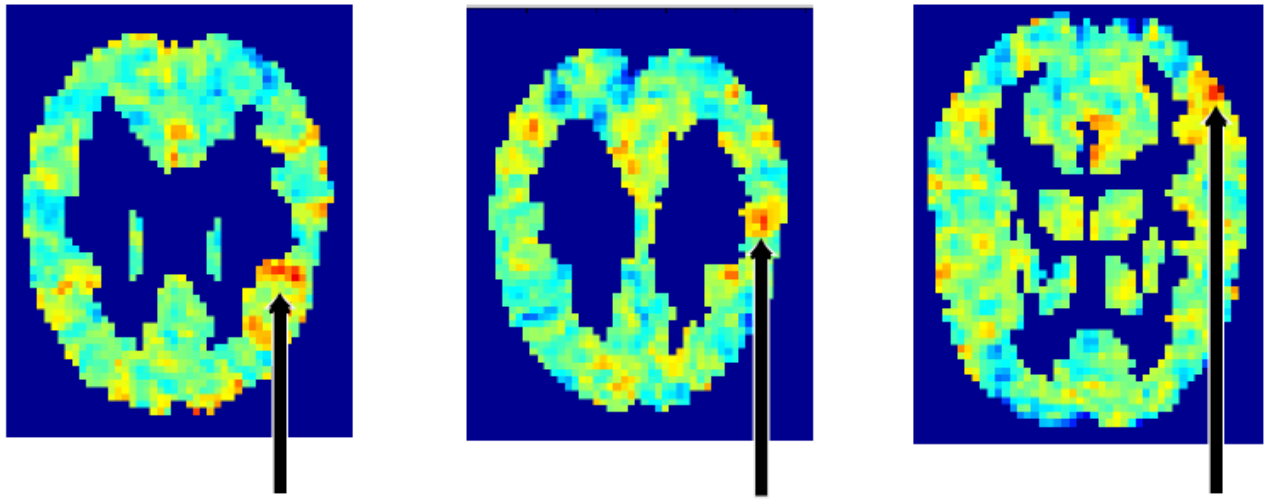
1770 test pairs in leave-2-out:

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

Mean accuracy over 9 subjects: 0.79

ADDITIONAL RESULTS

FOR EACH OF 25 FEATURES, EXAMINE IMPORTANCE ACROSS BRAIN



"Gustatory cortex"

"Planning motor actions"

"Body motion"

Pars opercularis (z=24mm)

Postcentral gyrus (z=30mm)

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

FOR A GIVEN PARTICIPANT, WHICH WORDS WOULD BE 'MOST ACTIVATING'?

- We can split the brain into 'areas' and now sweep through a very large number of words (10K) to see which word would maximally activate that region (or, if we wanted, the entire brain)
  - This is based on the model trained on 60 nouns.
  - Ideas for improvement? Are we sure this generalizes well? Note: regression typically works well for observations within boundaries of training set.
- This is a 'by participant' analysis and allows us to account for inter-individual differences.

FOR A GIVEN PARTICIPANT, WHICH WORDS WOULD BE ‘MOST ACTIVATING’?

- 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 (**biological motion** [Saxe et al., 2004])
- Left Anterior Cingulate
  - poison, lovers, galaxy, harvest, sin, hindu, rays, thai, tragedy, danger, chaos, mortality (**emotional stimuli** [Gotlib et al, 2005])

ALTERNATIVES TO THE 25-FEATURE BASIS SET

Would any random basis set work equally well? NO. Choosing 25 words randomly as basis sets always does worse than their manually chosen set.

Fig. 5. Accuracies of models based on alternative intermediate semantic feature sets. The accuracy of computational models that use 115 different randomly selected sets of intermediate semantic features is shown in the blue histogram. Each feature set is based on 25 words chosen at random from the 5000 most frequent words, excluding the 500 most frequent words and the stimulus words. The accuracy of the feature set based on manually chosen sensory-motor verbs is shown in red. The accuracy of each feature set is the average accuracy obtained when it was used to train models for each of the nine participants.