The neural structure of human concepts: Applying machine learning and factor analysis to brain imaging data

Marcel Just

Carnegie Mellon

CENTER FOR COGNITIVE BRAIN IMAGING CARNEGIE MELLON

10:20a.m.-11:20a.m.

October 31, 2013

Center for Cognitive Brain Imaging Research Team

Tim Keller

Vlad Cherkassky

Rob Mason

Diane Williams

Nancy Minshew

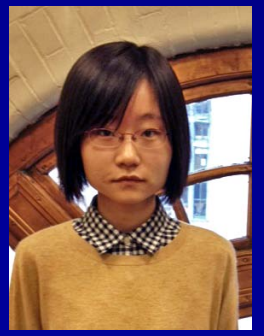
Tom Mitchell

Andrew Bauer

Kai-min Chang

Akiko Mizuno

Hide Komeda (Kyoto Hakubi)

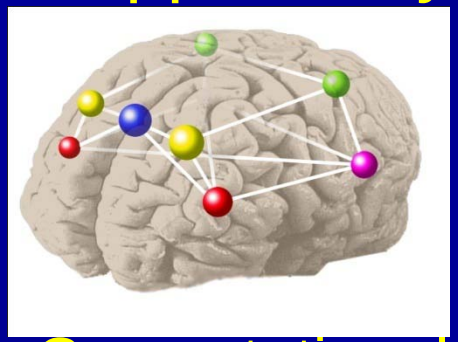


Jing Wang

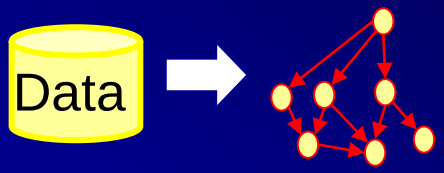
Outline of Mind-Reading talk

- Neural representation of concrete objects
- Neural representation of emotions
- Real-time mind-reading
- Glimpses of the future
  - Bilingualism
  - Common representations for words and pictures

An opportunity for cognitive neuroscience



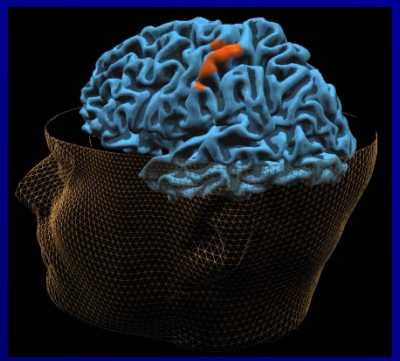
Computational cognitive models



Machine learning and dimension reduction for scientific discovery

*

Brain imaging



READING THOUGHTS OF CONCRETE OBJECTS

Issues that this research is addressing

1. Are cognitive states (thoughts) identifiable by applying machine learning and factor analytic techniques to fMRI brain activation patterns?

2. Are there some core dimensions of meaning that underpin the representation of concrete nouns/physical objects? Of interpersonal interactions? Of emotions?

3. In which part or parts of the brain is a concrete noun/physical object represented?

4. Is there a commonality across people of neural representations of such concepts?

How we represent concrete word meanings

A set of 60 words is presented six times in a random order.

Each presentation of a word lasts 3s, followed by a 7s fixation period.

60 Items

Categories					
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

Task

The participants' task was to actively think about the properties of the object to which the word referred.

Each participant was free to choose any properties for a given item, and there was no attempt to obtain consistency across participants in the choice of properties

11 participants, college age

Factor analysis

Find sets of voxels (3x3x5 mm volume elements) out of 20000 voxels that have similar preferences among the 60 words

- These sets of voxels constitute the factors or dimensions of meaning representation
- Focus on the most stable voxels, ones that respond similarly (reliably) to the set of 60 words each time the set is presented
- Measure a voxel's stability as the mean correlation between pairs of 60-element intensity vectors across pairs of presentations
- Select the 50 most stable voxels in each of the 4 lobes (plus fusiform)
  - These 50 voxels are central to the input to the Factor analysis

Factor analysis cont'd

- Compute the intercorrelations among the 50 most stable voxels in each "lobe" for 4 individual participants
  - i.e. 5 lobes x 4 participants
- Do 20 factor analyses on the 20 correlation matrices
- Finally converge on a total of 80 voxels per person (= .5% (half of one percent) of the brain
- Each voxel is about 50 mm^3 in our analyses
  - About the size of a peppercorn
- (Details of Methods in Just et al., 2010, PLOS One)

Four factors emerge that are common across participants

(Factor labels are our interpretation)

Shelter	Manipulation	Eating	Word length

Words with highest scores for the four common factors

Shelter	Manipulation	Eating	Word length
apartment	pliers	carrot	butterfly
church	saw	lettuce	screwdriver
train	screwdriver	tomato	telephone
house	hammer	celery	refrigerator

Voxel locations for each factor

- It is possible to trace the factors back to their root voxels, and determine where the voxels associated with a given factor are located
  - By finding voxels whose activation profile is correlated with the factor profile
  - Voxels were uniquely assigned to the 4 factors

- For each factor, the associated voxels tended to cluster in 3 to 5 different locations in the brain, distributed across more than a single lobe.
  - A sphere was defined with its center at the centroid of the cluster
  - radius is equal to mean distance of all voxels in cluster from centroid

- Each word has some factor score for each factor
  - e.g. *spoon* might be high on both *manipulation* and *eating*

Multiple brain locations for each semantic factor

Four Factors
Shelter
Manipulation
Eating
Word Length

Locations of voxel clusters corresponding to the factors

- 3-5 locations per factor
- *manipulation* and *eating* are left lateralized

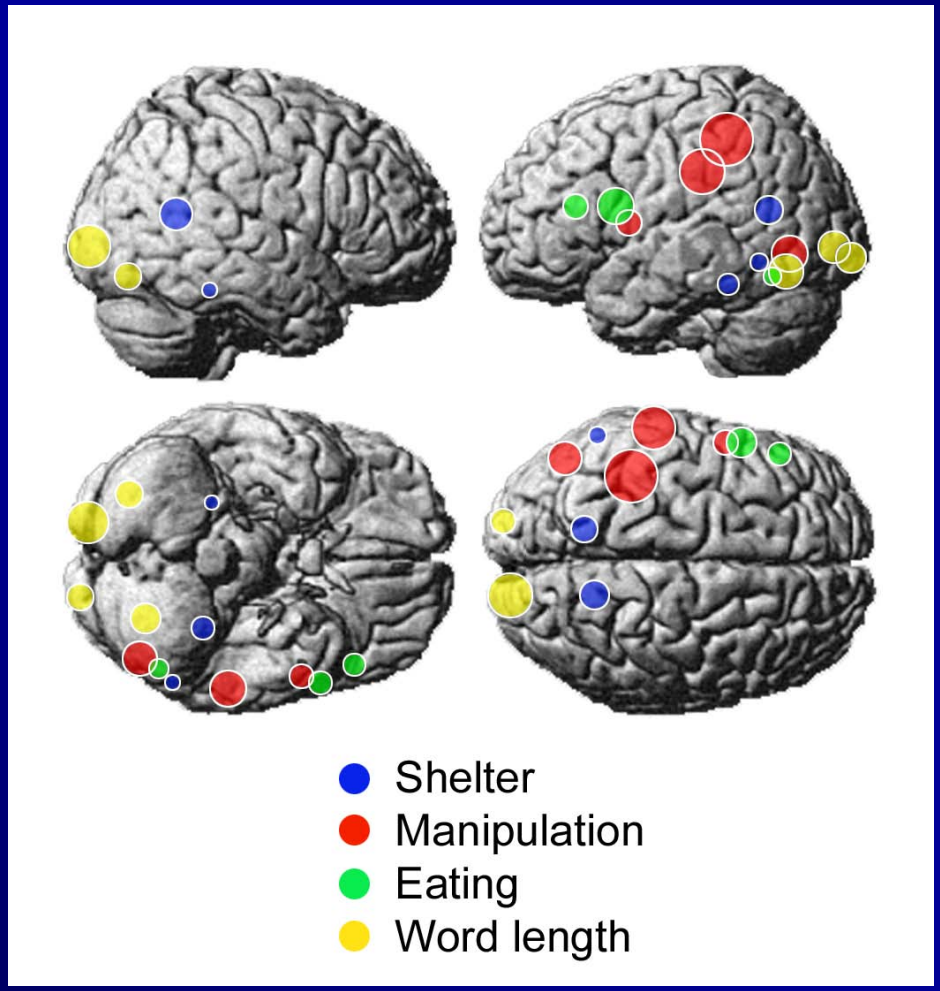
Factor	Location	x	y	z	No. of Voxels	Radius (mm)
Shelter	L Fusiform Gyrus/ Parahippocampal Gyrus (PPA)	-32	-42	-18	26	6
	R Fusiform Gyrus/ Parahippocampal Gyrus (PPA)	26	-38	-20	6	4
	L Precuneus*	-12	-60	16	40	8
	R Precuneus	16	-54	14	36	8
	L Inf Temporal Gyrus	-56	-56	-8	12	4
Manipulation	L Supramarginal Gyrus*	-60	-30	34	51	10
	L Postcentral/Supramarginal Gyri	-38	-40	48	21	12
	L Precentral Gyrus	-54	4	10	18	6
	L Inf Temporal Gyrus	-46	-70	-4	34	8
Eating	L Inf Frontal Gyrus*	-54	10	18	26	8
	L Mid/Inf Frontal Gyri	-48	28	18	10	6
	L Inf Temporal Gyrus	-52	-62	-14	7	4
Word Length	L Occipital Pole*	-18	-68	-12	20	8
	R Occipital Pole	16	-94	0	47	10
	L Lingual/Fusiform Gyri	-28	-68	-12	20	8
	R Lingual/Fusiform Gyri	30	-76	-14	14	6

Good convergence with previous imaging findings using visual stimuli instead of words

The locations associated with a factor have been previously characterized as being related to the factor in conventional fMRI research

- RH shelter Parahippocampal area corresponds (within 2.8 mm) to classic parahippocampal place area (PPA) obtained when people look at pictures of houses
  - 4 of the 5 shelter locations correspond to 4 areas activated when judging familiarity of pictures of places (participant's own office, house) (Sugiura et al., 2005)
- Also excellent correspondence for manipulation with study of actual and pantomimed tool use (Hermsdorfer , 2007)
  - All 4 of the manipulation locations from the factor analysis match up with the 4 activation loci in the tool use task (within about 5 mm)
- *eating* factor includes an L IFG cluster that is close to the location associated with face-related actions like with chewing or biting reported by Hauk et al. (2004)

Surface rendered locations of the voxel clusters associated with the 4 factors;
Not just “hot spots”; these are neurosemantic computational centers, common across people



Machine learning of fMRI word representations (aka “multivoxel pattern analysis”) to test this theory

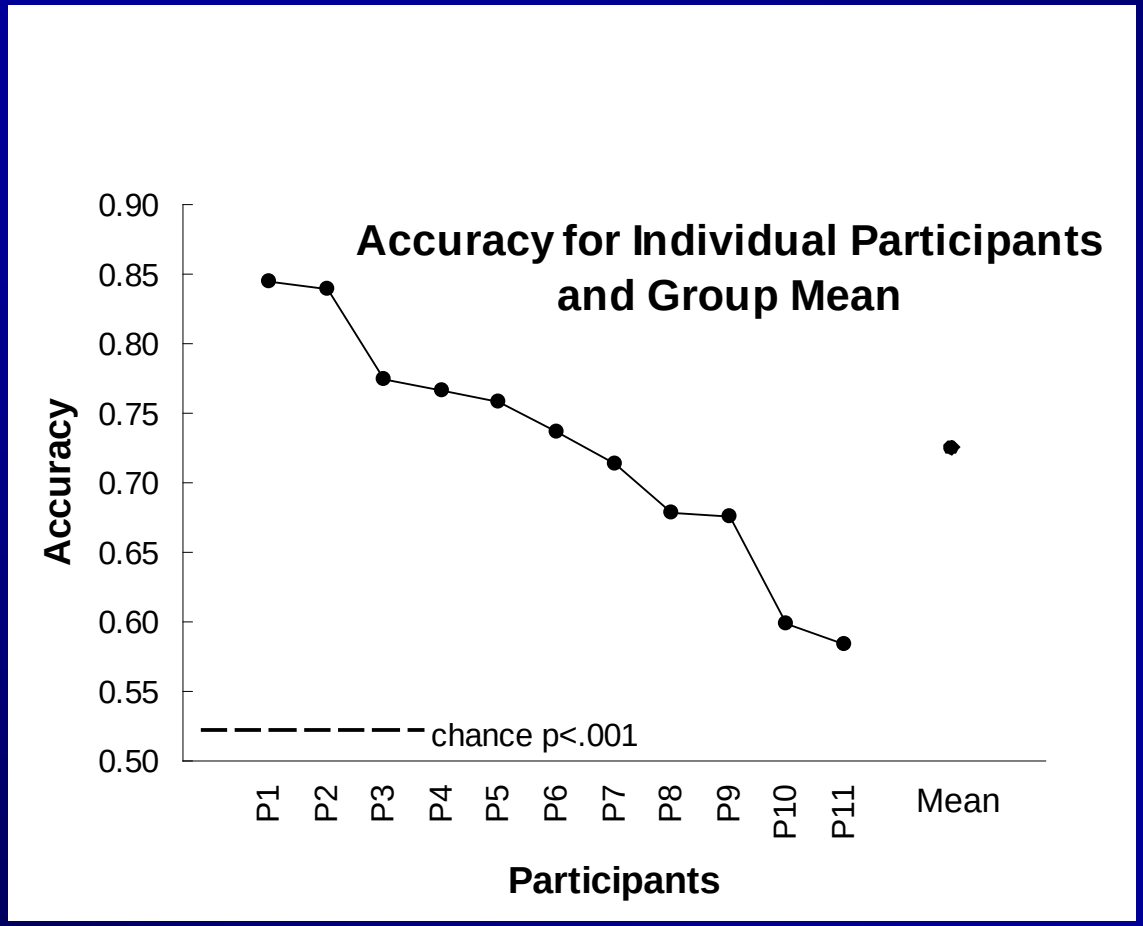
- Factor analysis is a powerful discovery tool, but often suffers from a lack of an independent method to assess the explanatory and predictive power of the analysis
- To assess how the 4 factors (their profiles and locations) reflect the properties of the 60 words, machine learning (ML) methods were used to construct and test the theoretical model (and compare several alternative models) of the activation
- We train a classifier to identify patterns associated with each of the 60 words (Gaussian Naïve Bayes, SVM)
- Cross validation: train classifier on 4 presentations and test it on the mean of the remaining 2 (to reduce noise)
- Select features (voxels) from training set based on their correspondence to factors in training set factor analysis

Rank Accuracy measure of classifier performance

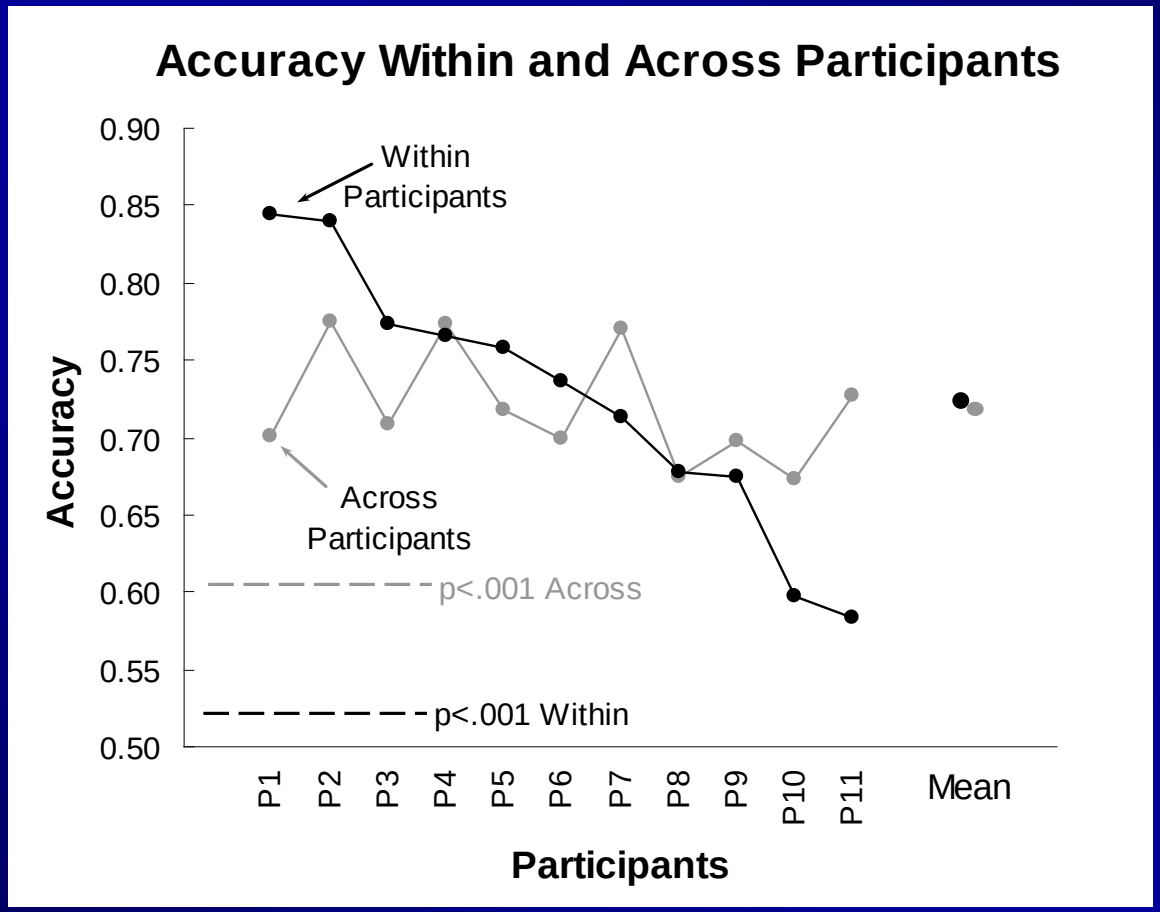
- The classifier takes many “guesses” at what each thought was, ordered from most likely to least likely, as many guesses as there are stimuli (probability-ranked list of all items)
- The dependent measure is the mean normalized rank of the correct guess (expressed as a percentile) – Rank accuracy
- Expect about .50 rank accuracy by chance

Accurate identification of which of the 60 words a participant was thinking about

Mean accuracy across participants = .72, very far above p < .001
Max accuracy =.84 for 2 participants



Classifier also works across participants; Train on 10 participants – test on 11th gray curve

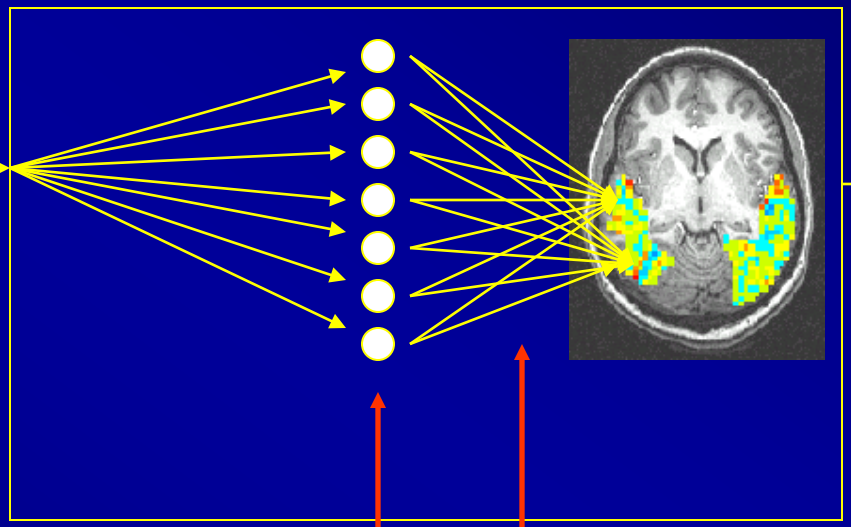


The neural representation of these words is common across people

60 Minutes demo


Mind Reading
produced by
Shari Finkelstein

Now instantiate the Model as a Predictive theory



New word "celery"

predicted activity for "celery"

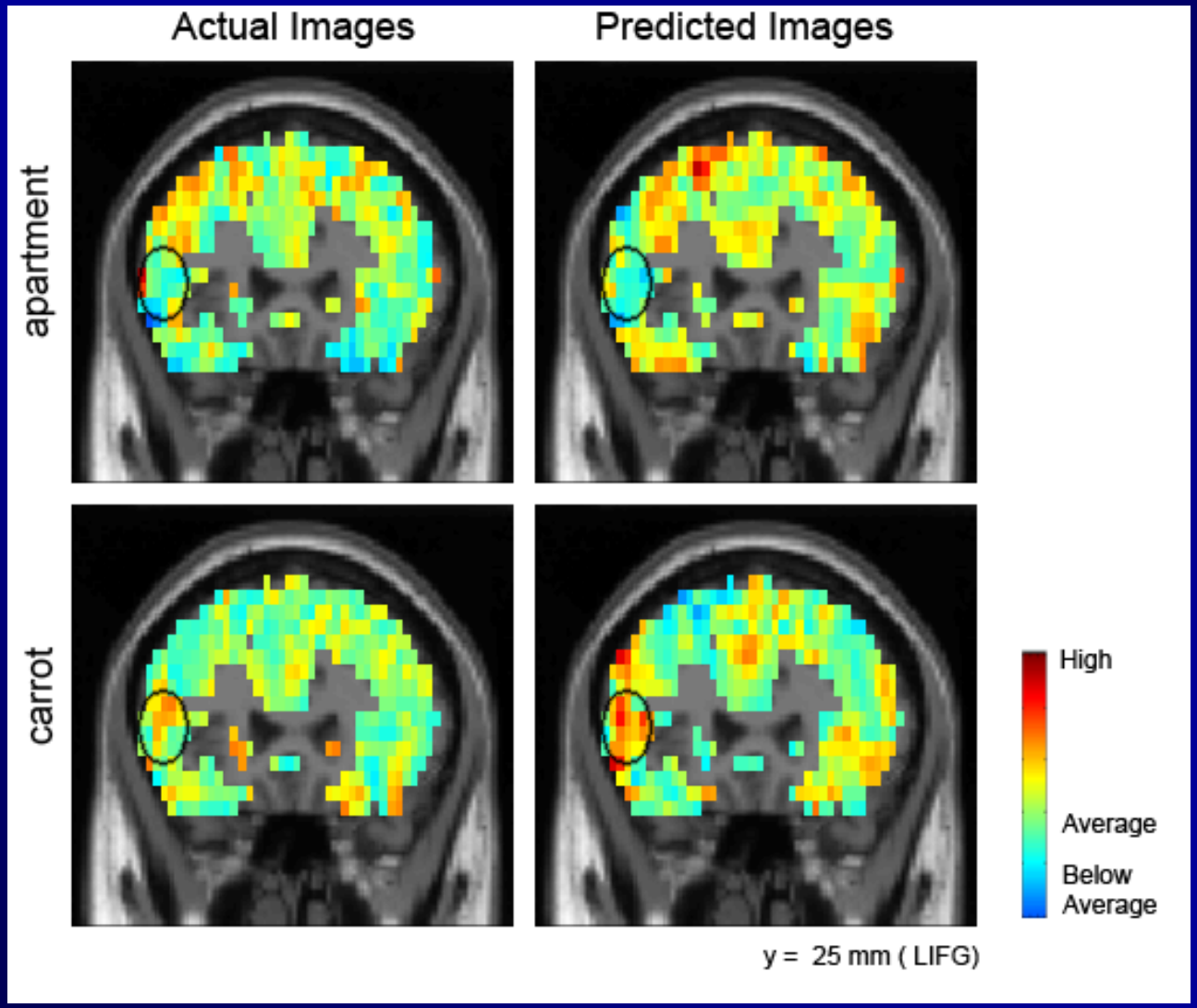
Intermediate semantic factors weighted by subject ratings: how much is the word related to *shelter* or *manipulation* or *eating*?

Mapping learned from fMRI data

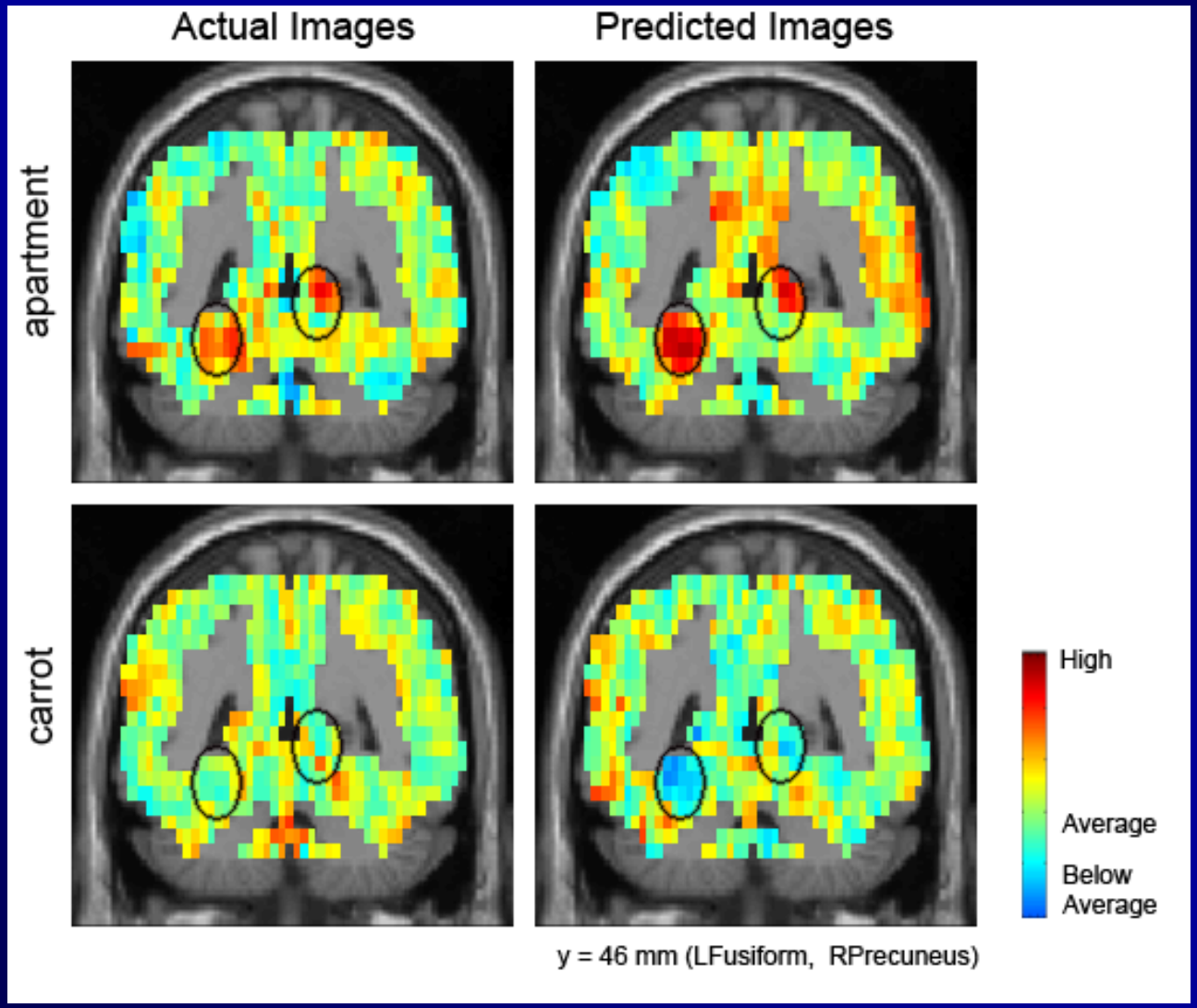
Generative modeling, extending to new words

- Regression model using a word's factor scores or subject ratings to predict activation
  - 4 independent variables (including word length)
- Predict activation level of 80 voxels
  - (5 voxels per factor location, ~4 locations/factor)
- Construct model on the basis of 58 of the 60 words
  - Attempt to match up the two left-out brain images with their words on the basis of the images' similarity to the predicted images
    - E.g. leave out *carrot* and *apartment*
    - Get independent ratings of the left out words with respect to the factors
    - Generate predicted images for *carrot* and *apartment*
    - Determine which predicted images are more similar to the two acquired images
- Do this for all possible sets of two left-out words
  - 1770 possibilities
- Repeat for each of the 11 participants
- Obtain .73 accuracy in matching the two "new" words to their images

Factor Based Prediction Model (LIFG)



Factor Based Prediction Model (LFusiform, RPrecuneus)

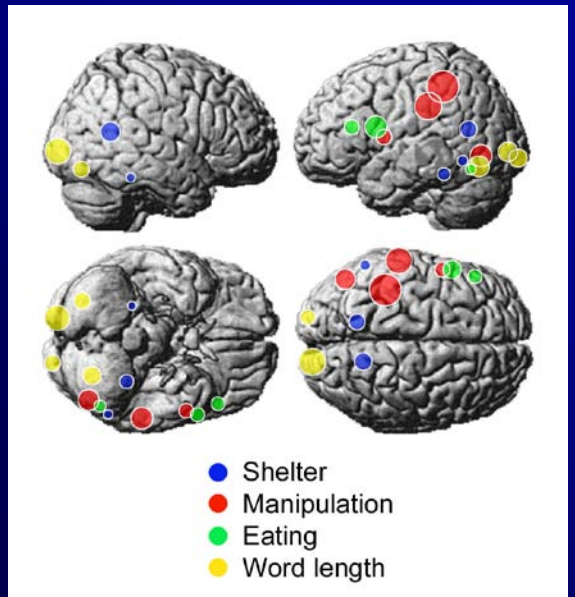


Relation to GLM/SPM univariate analyses

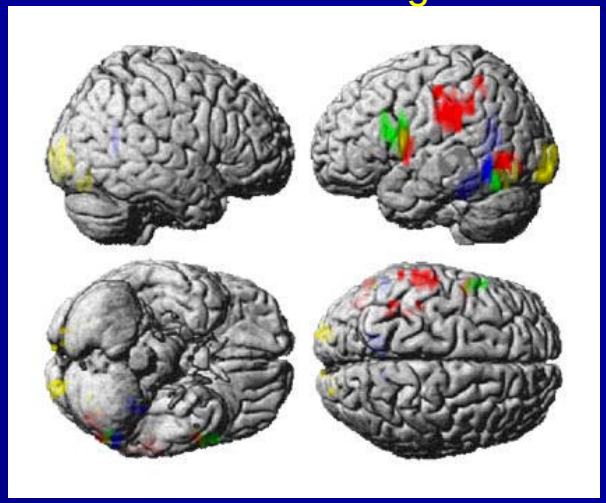
GLM/SPM analyses and contrasts can provide some category-specific location information for a few categories that is broadly consistent with the factor analysis outcomes, AFTER you program in the discovered factors; but conventional analyses provide:

- no discovery procedure
- no exemplar-specific information (factor scores)
- no clustering of distributed voxels that have similar responses over stimuli
- no theoretical basis for generalizing to new items

Machine Learning/Factor Analysis



SPM with the benefit of the factors discovered with Machine Learning/Factor Analysis



Summary of findings

- 3 main neurosemantic factors emerge for representing concrete nouns/physical objects
- Each factor represented in 3-5 cortical locations, corresponding to separable brain subsystems
- The theory is generative, extending to new words
- The neural representation is essentially common across people
- We can use this approach to build a testable theory of the neural representations (thoughts) of people as they process more abstract and more complex thoughts
- (Science 2008;PLoS One 2010)

READING EMOTIONS

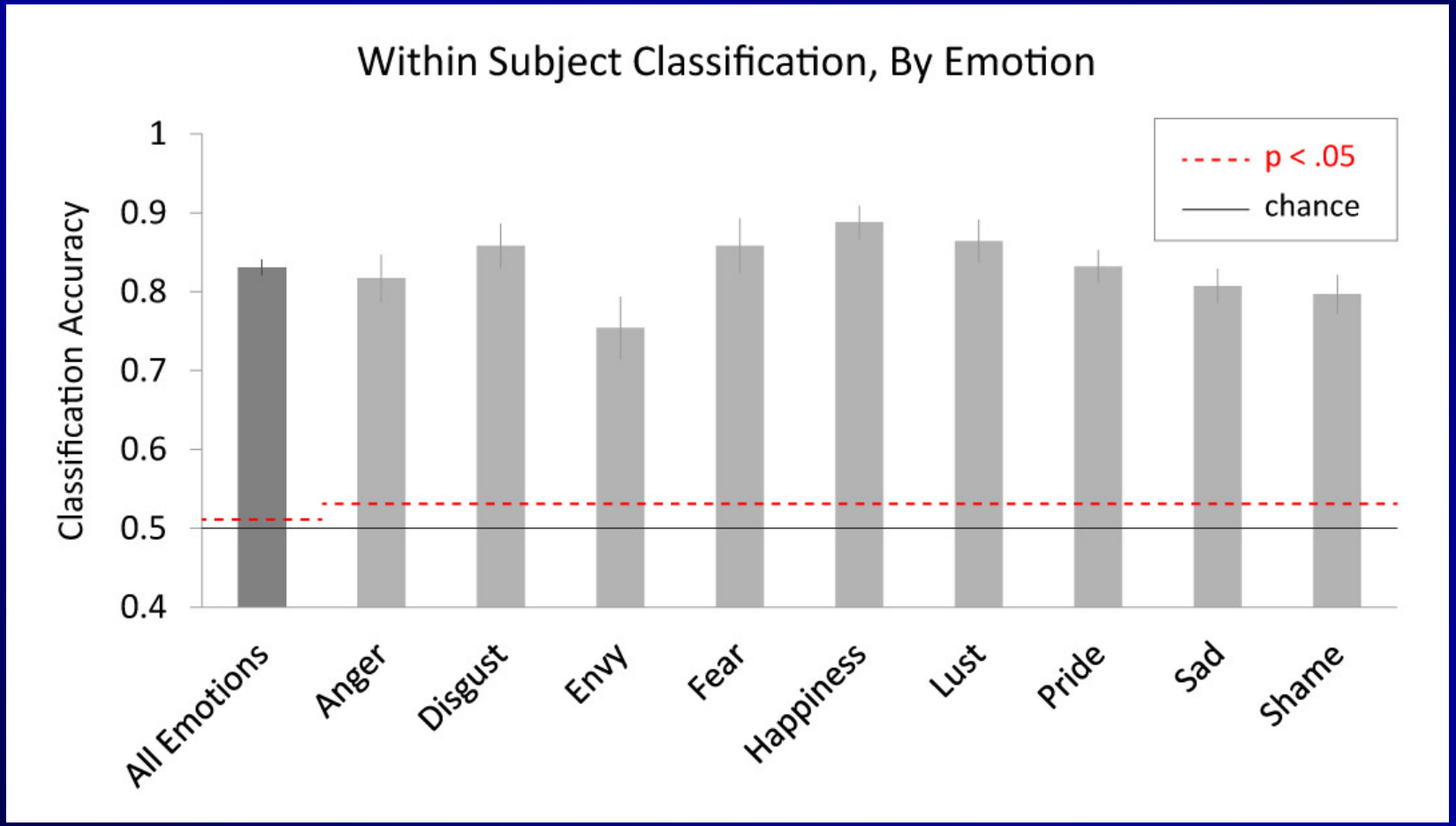
Identifying emotions from their fMRI patterns

- Collaboration with
  - Karim Kassam
  - George Loewenstein
  - Amanda Markey
  - Vlad Cherkassky
- Actor subjects, students in CMU drama department
- Trained to evoke 18 different emotions

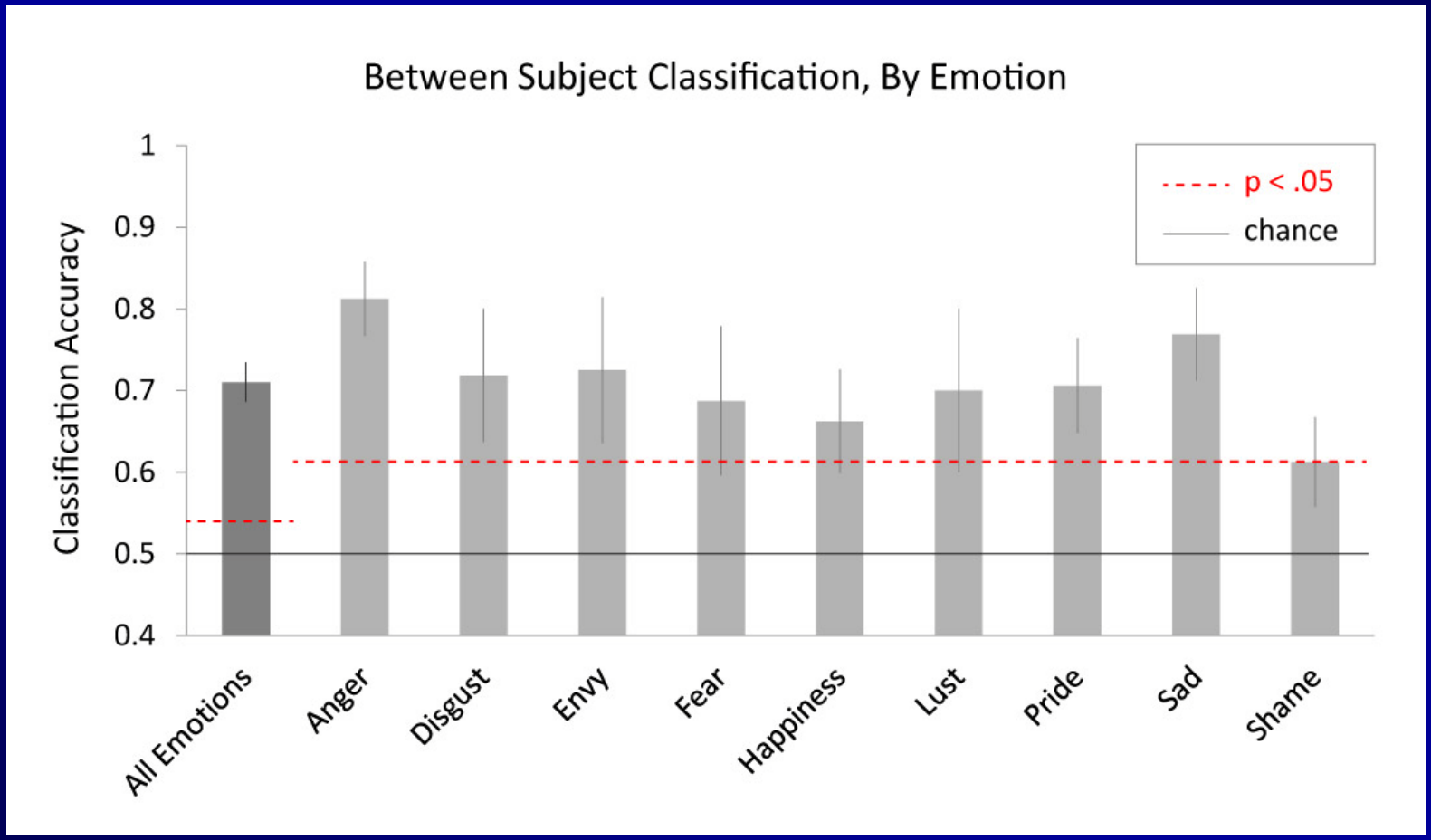
18 words grouped into 9 emotion categories: 2 emotion words per category

- Anger: angry, enraged
- Disgust: disgusted, revulsed
- Envy: envious, jealous
- Fear: afraid, frightened
- Happiness: happy, joyous
- Lust: lustful, horny
- Pride: proud, admirable
- Sadness: sad, gloomy
- Shame: ashamed, embarrassed

Accurate identification of emotions: mean = .82



Commonality across people; train classifier on data of 9 people, test on 10th



Validity check- are the activation patterns observed in our study the same as those that occur when people experience externally-induced emotion

- 12 pictures evoking disgust (from IAPS) were shown to the participants at the end of the study (interspersed with 12 neutral pictures)
  - E.g. human mutilation, rotting food
  - Presumably evokes externally-induced emotion
- The classifier was given the resulting activation to classify based on its training on the disgust items in the study
- The classification accuracy of the disgust pictures as disgust was very high -- .91
- →The activation pattern associated with the self-induced emotion must be very similar to the activation induced by the IAPS pictures

Neural factors underlying emotions

- Factor analysis of emotion activation
- Seeking subsets of voxels that similar patterns of response to the 18 emotions
- Interpreting the resulting patterns (factors)

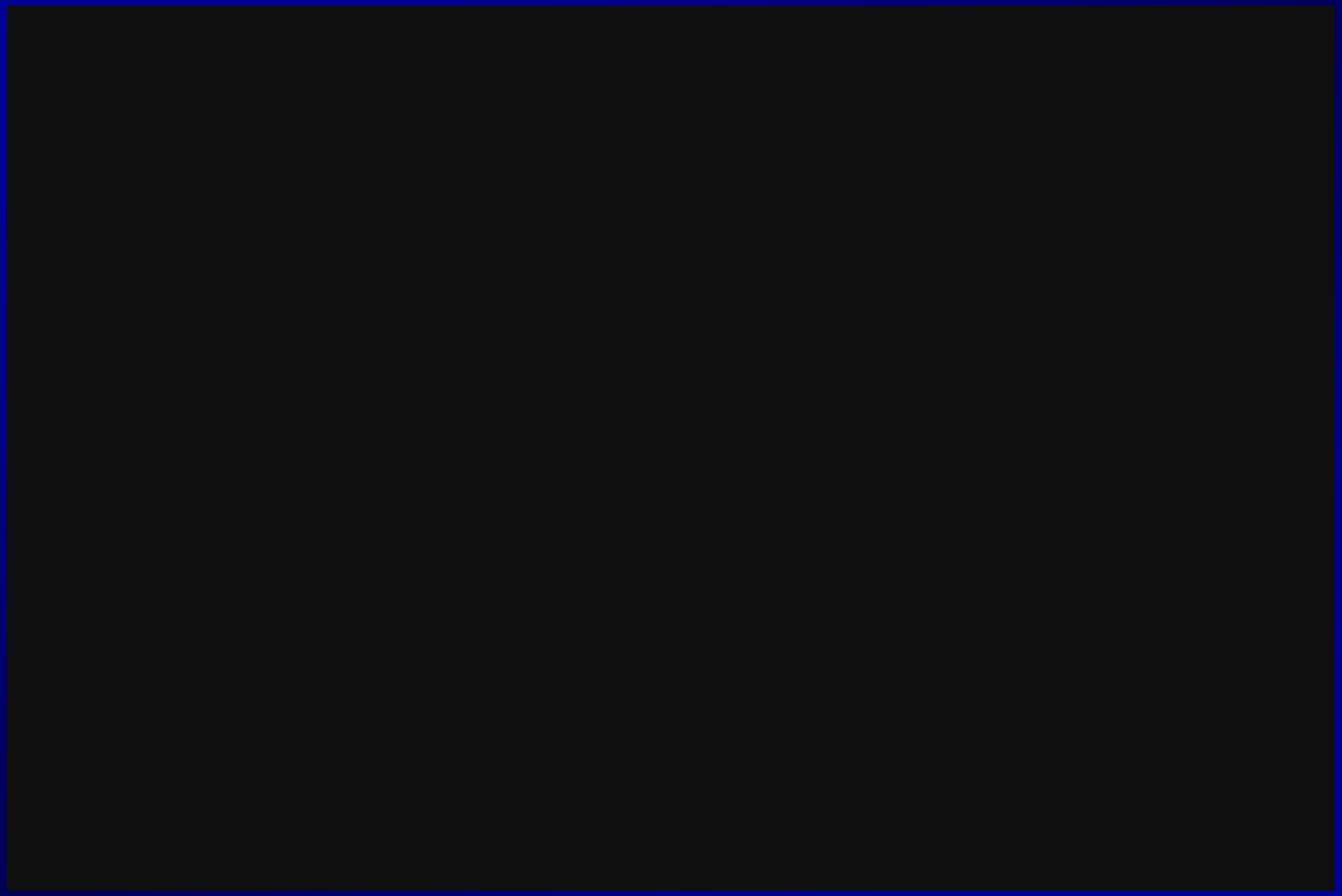
Neural factors underlying emotions

- *Valence* -- the goodness or badness of the emotional situation
  - Factor loadings correlate nearly perfectly *r* = 0.96 with ratings of pleasantness of the emotion scenarios
  - Neural regions: medial frontal (core affect and emotion regulation) and orbital frontal (affective value computation)
- *Arousal* or preparation for action.
  - Factor loadings correlate with ratings of arousal r = 0.49; highest: anger, fear and lust; lowest: sadness, shame, pride
  - Neural regions: Basal Ganglia and Precentral Gyrus – consistent with arousal
- *Social* (another person involved?).
  - jealousy, envy, lust had the highest loadings; disgust and revulsion had the lowest loadings
  - Neural regions: anterior and posterior areas of the cingulate cortex -- self perception and other perception
- *Lust*
  - separate from other emotion categories.
  - Neural regions: fusiform gyrus and inferior frontal areas implicated in face processing and areas involved in the processing of sexual stimuli
- *Word length* in the occipital cortex. Note that successful classification did not depend on occipital cortex.

Provides a new perspective on the structure of emotion (PLoS One, 2013)

Real-time Mind-reading

Real-time thought identification



Demo of Identification of a single presentation of each item; Classifier was trained on a previous dataset from this participant

Actually Presented Letter	Classifier's First Guess	First Guess score	Classifier's Second Guess	Second Guess Score
W	W	√		
G	W	X	G	√
E	E	√		
O	O	√		
C	C	√		
L	L	√		
H	M	X	H	√
D	D	√		
S	S	√		
A	A	√		
M	H	√		
N	N	√		
J	J	√		
I	I	√		
F	E	X	F	√

Other applications of neurosemantic research

- Cross-classification of words and pictures
- Cross-classification of a bilingual's two neural representations of a concept

A new frontier of brain science

- Understanding the structure of human thought
  - Knowing the brain's building blocks
  - Learning how to put the blocks together
- Brain-computer interfaces that can access thoughts
- Assessing alterations of thought in psychiatric diseases
  - Improving ability to treat the thought disorder
- Imaging not just the brain, but the human mind

CONTACT INFO & REPRINTS

Marcel Just
Center for Cognitive Brain Imaging
Psychology Department
Carnegie Mellon University
Pittsburgh PA 15213
email: just@cmu.edu
URL: www.ccbi.cmu.edu

Reprints downloadable