mscohen@g.ucla.edu

The Ballad of Sad Panda, or How I learned to (mostly) stop worrying and love meta-analysis

Tal Yarkoni
UT-Austin

The Sad Panda study

- My dissertation study
- It was going to be my magnum opus
- I never published it

The Sad Panda study

- Hypothesis: conserved neural circuitry supports representation of emotional expressions not only in humans, but also in other mammalian species
- n = 32 undergraduate subjects
- 3 experimental conditions: happy, neutral, sad pandas
- Prediction: the amygdala shows greater activation for happy and sad pandas than neutral pandas

sad pandas

happy pandas

neutral pandas

The Sad Panda study

- “Previous fMRI studies have demonstrated that specific brain circuits such as the amygdala are selectively implicated in representing human emotional expressions. The degree to which this finding generalizes to the rest of the animal kingdom is unknown. Here we investigated the (human) neural representation of emotional expression in the giant panda (*Ailuropoda melanoleuca*). We scanned 32 participants with fMRI as they viewed and rated previously normed images of pandas displaying happy, sad, and neutral emotional expressions. As predicted, emotional content robustly activated the amygdala relative to neutral content. A whole-brain search further identified selective activations in inferotemporal cortex, precuneus, and left inferior frontal gyrus. These results held when controlling for differences in a range of low-level stimulus properties (e.g., luminance and contrast) as well as trial-by-trial differences in response time (RT), suggesting that the representations we observed were likely to be affective in nature. Mediation analysis further indicated that individual differences in amygdala activation partially mediated the relationship between experimental condition and affective ratings of the stimuli. Finally, in cross-validated searchlight MVPA analyses, we were able to correctly classify panda expression 94% of the time using amygdala activation. Collectively, our results provide compelling evidence that the neural representation of basic affective expressions in humans is not unique to humans, and that similar circuitry supports recognition of emotion in pandas and possibly other mammals.”

A masterpiece of converging evidence

- Randomized, controlled experimentation
- Use of both hypothesis-driven (ROI) and exploratory (whole-brain) univariate analyses
- Explicit control of likely confounds
- Evaluation of plausible causal models
- Cross-validated evaluation of predictive performance
- Data-driven clustering analyses

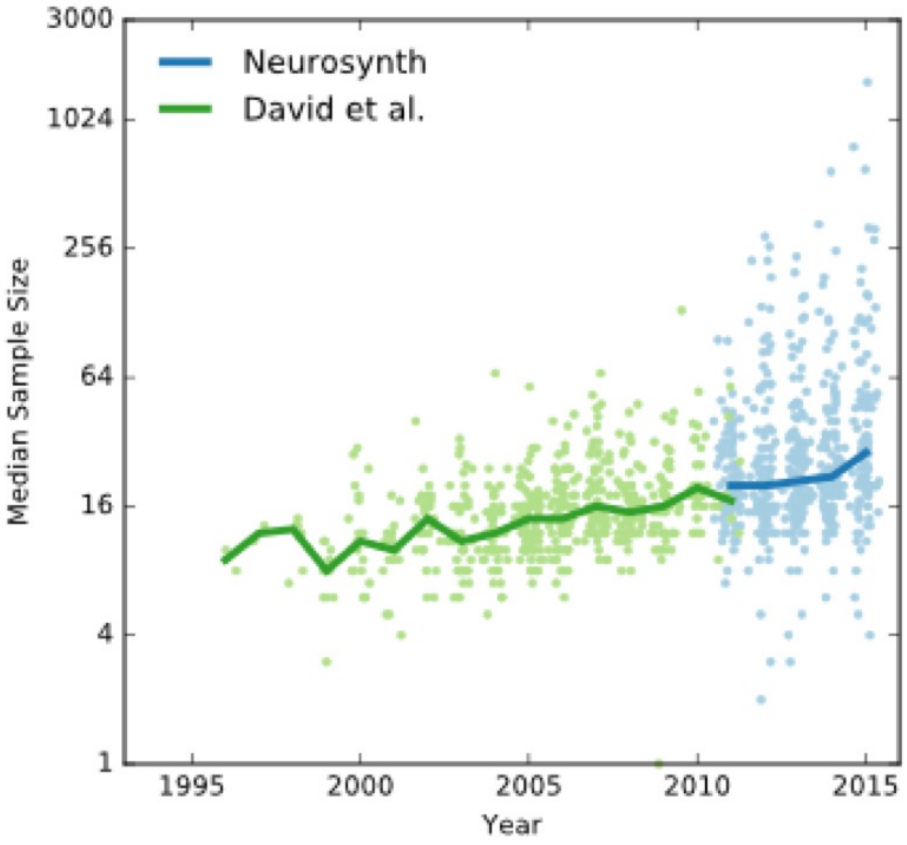
- And all of these converge to tell the same story!

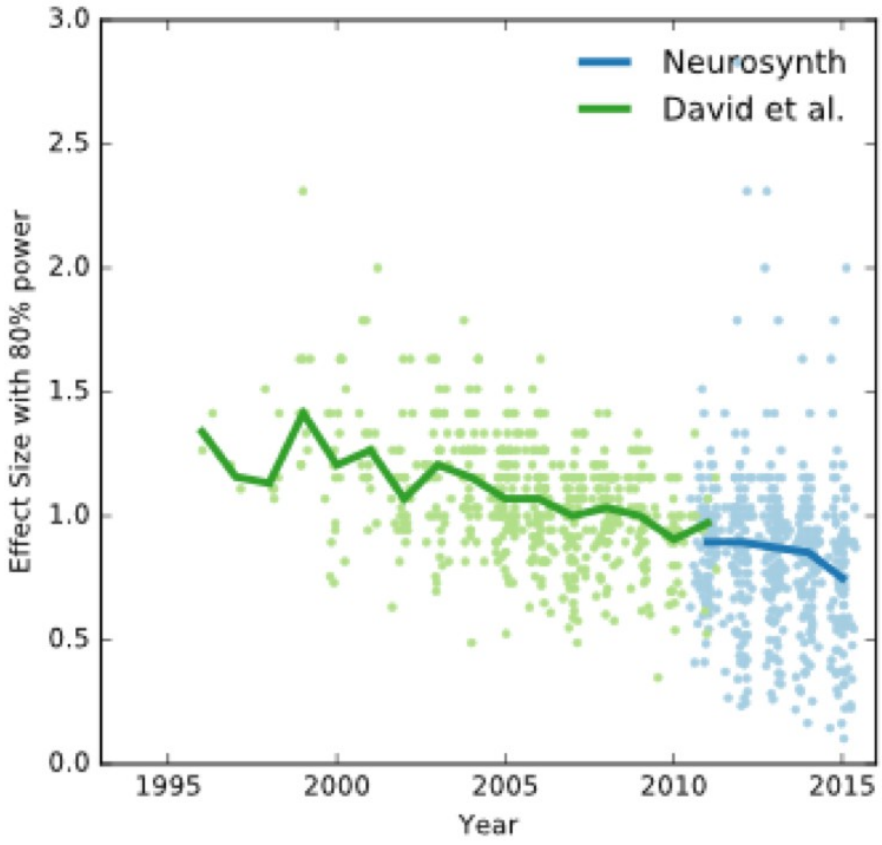
Why didn’t I publish this?

- The real reason: it doesn’t exist
- A reason that fits our StoryTime narrative: because there are good reasons to question most of the results—without even knowing what they are

The Sad Panda study

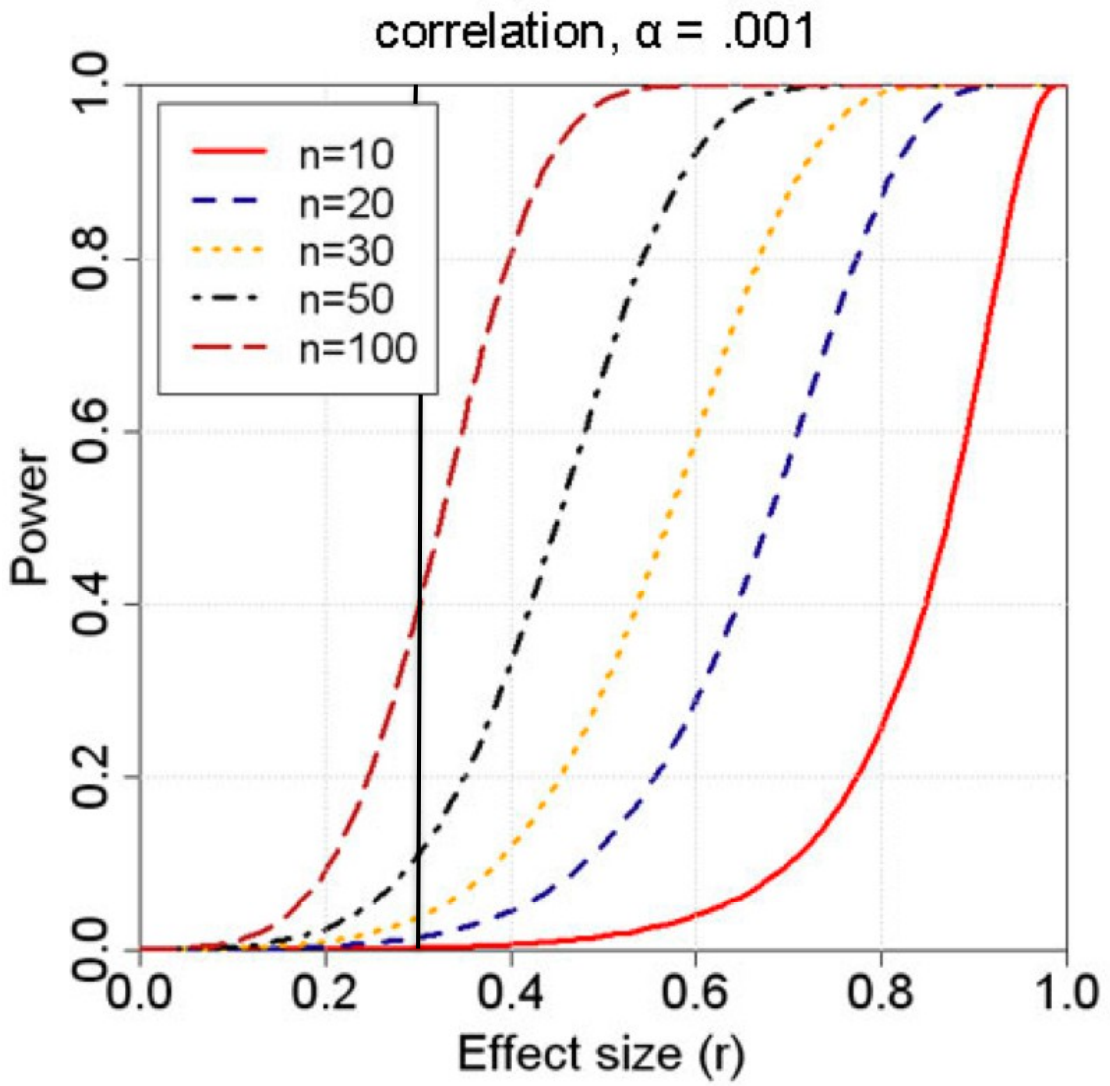
- "Previous fMRI studies have demonstrated that specific brain circuits such as the amygdala are selectively implicated in representing human emotional expressions. The degree to which this finding generalizes to the rest of the animal kingdom is unknown. Here we investigated the (human) neural representation of emotional expression in the giant panda (*Ailuropoda melanoleuca*). **We scanned 32 participants with fMRI** as they viewed and rated previously normed images of pandas displaying happy, sad, and neutral emotional expressions. As predicted, emotional content robustly activated the amygdala relative to neutral content. **A whole-brain search further identified selective activations in inferotemporal cortex, precuneus, and left inferior frontal gyrus.** These results held when controlling for differences in a range of low-level stimulus properties (e.g., luminance and contrast) as well as trial-by-trial differences in response time (RT), suggesting that the representations we observed were likely to be affective in nature. Mediation analysis further indicated that individual differences in amygdala activation partially mediated the relationship between experimental condition and affective ratings of the stimuli. Finally, in cross-validated searchlight MVPA analyses, we were able to correctly classify panda expression 94% of the time using amygdala activation. Collectively, our results provide compelling evidence that the neural representation of basic affective expressions in humans is not unique to humans, and that similar circuitry supports recognition of emotion in pandas and possibly other mammals."



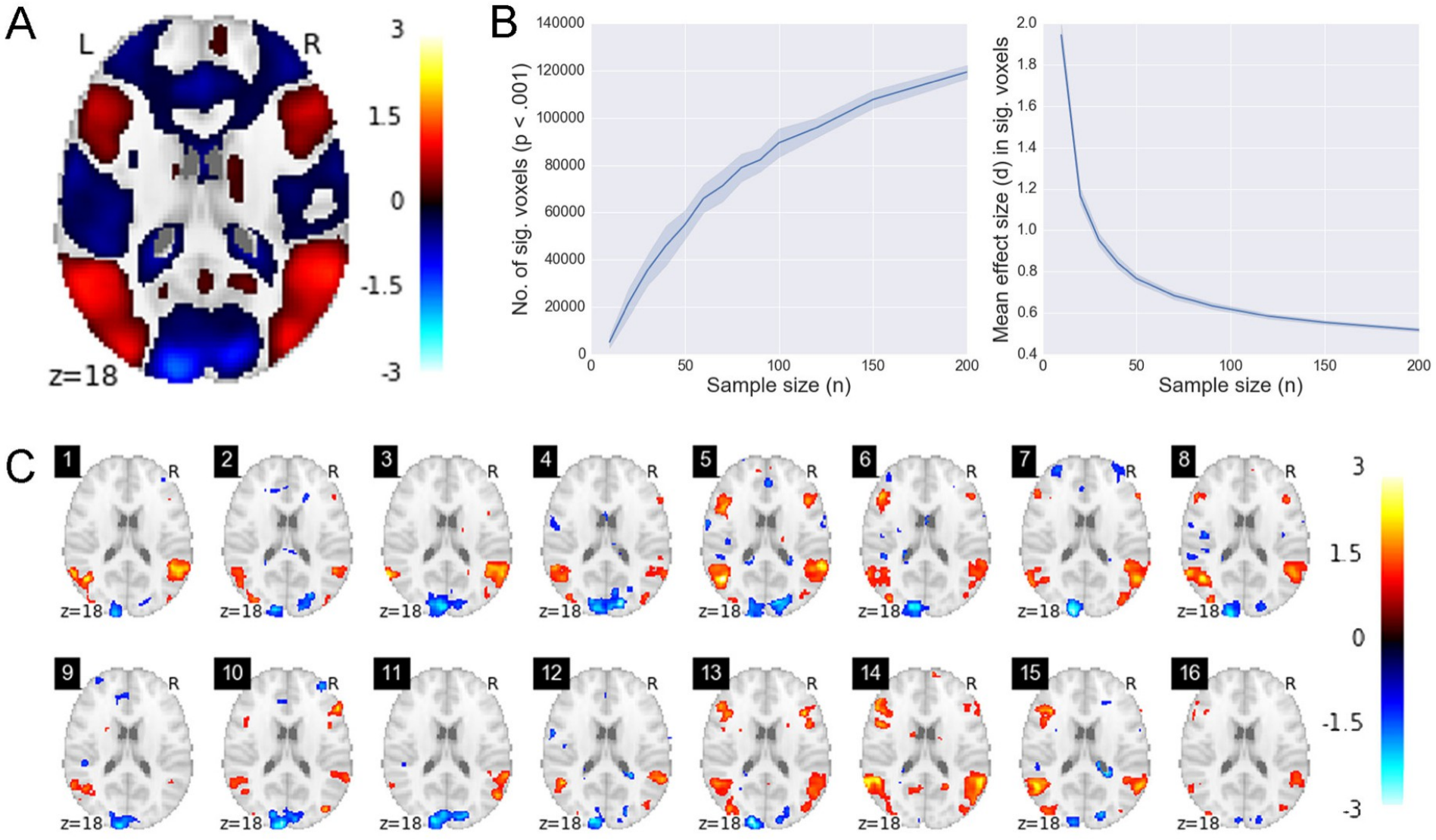


Poldrack et al. (submitted)

Statistical power and sampling error

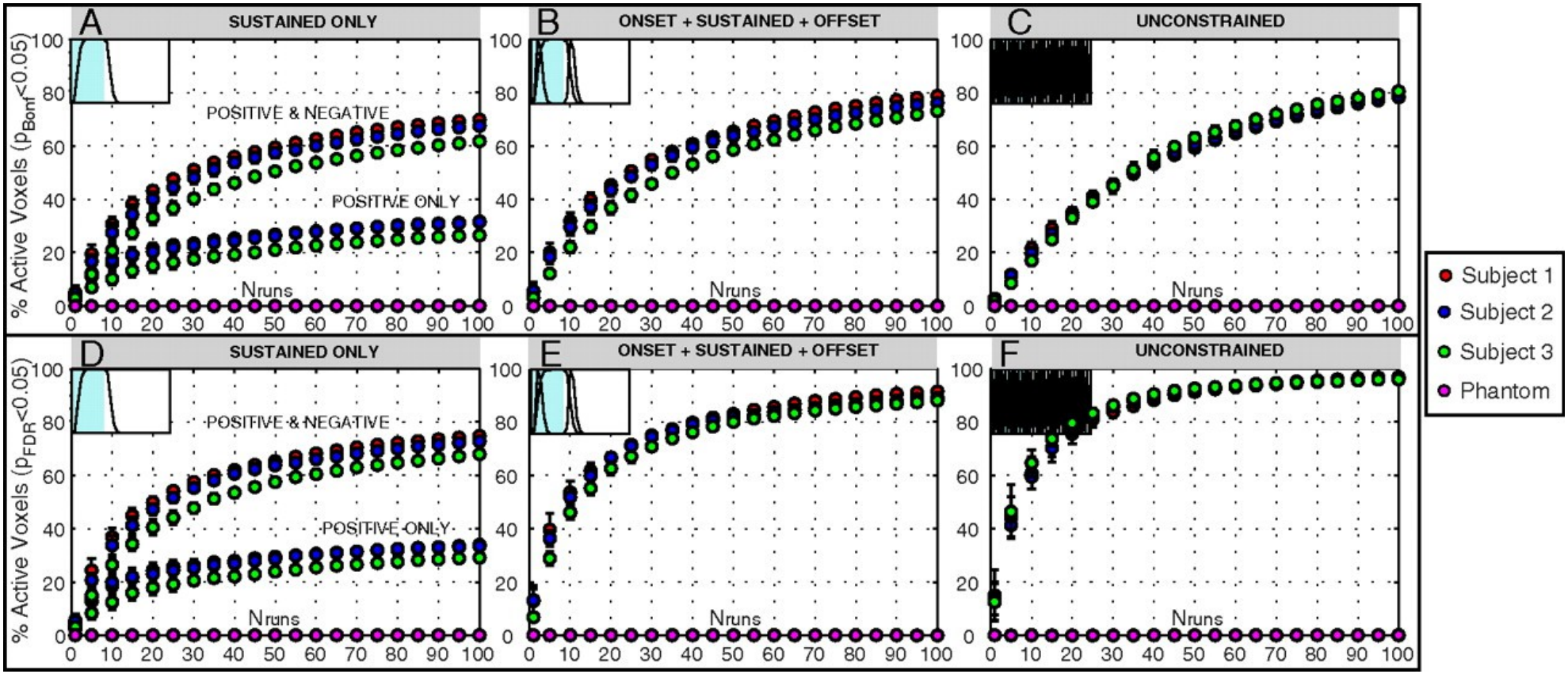


Sub-sampling the HCP social cognition task



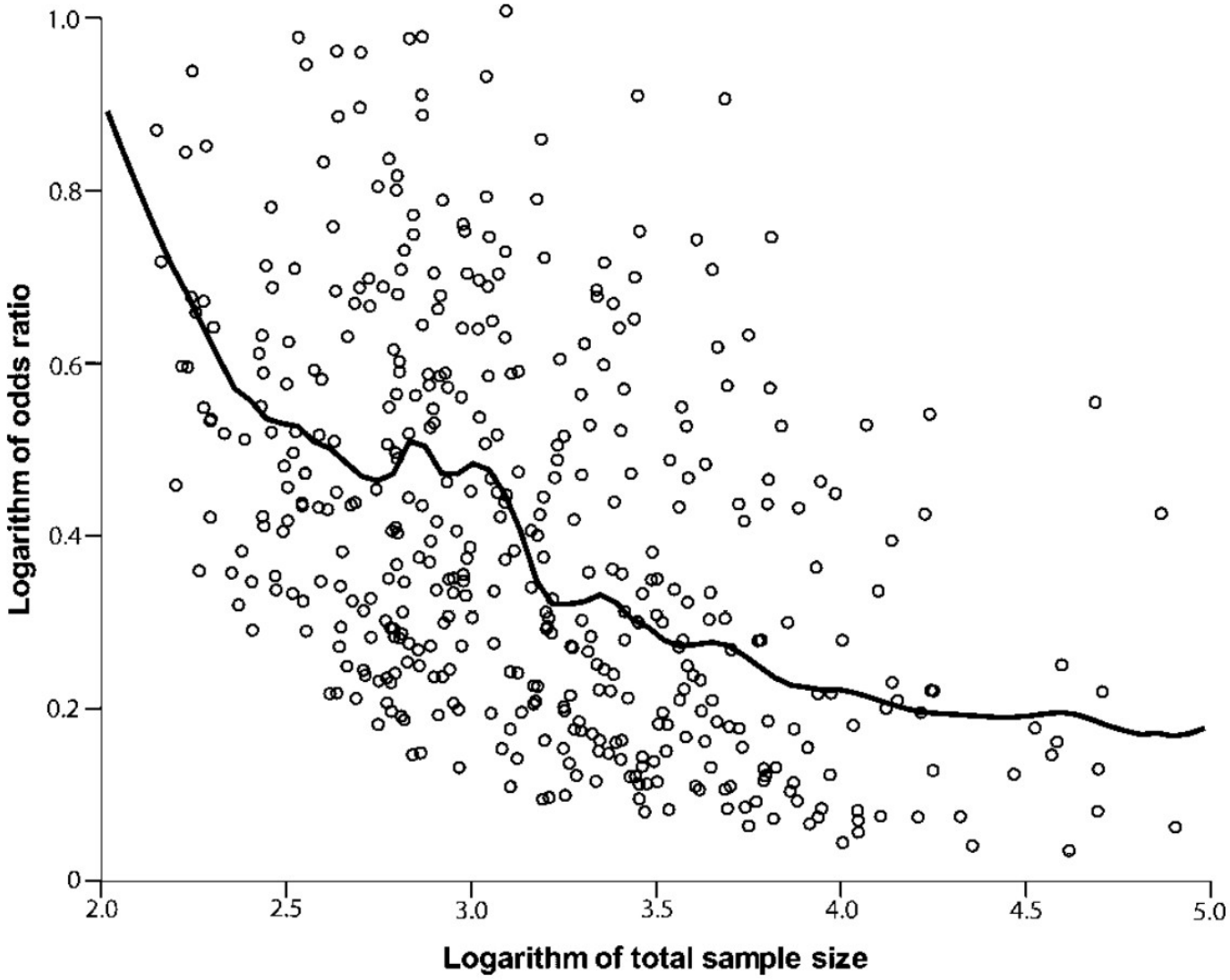
Cremers, Wager, & Yarkoni (in prep.)

The null is always* false, so at the limit…



Gonzalez-Castillo et al. (2012)

Not fMRI-specific; this is a universal problem



Ioannidis (2008): analysis of Cochrane meta-analyses

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

Interim conclusions

- Power is low
- There is rarely any support for claims of “selective” activation
- Type II error (if you believe in that sort of thing) is likely to be *very* high
- This changes interpretation!

The Sad Panda study

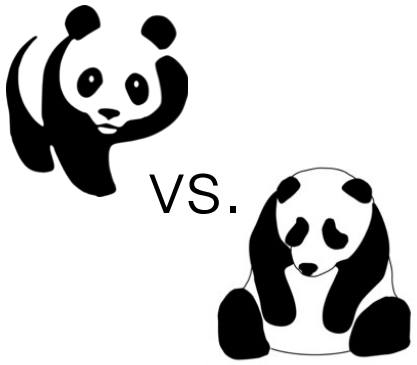
- “Previous fMRI studies have demonstrated that specific brain circuits such as the amygdala are selectively implicated in representing human emotional expressions. The degree to which this finding generalizes to the rest of the animal kingdom is unknown. Here we investigated the (human) neural representation of emotional expression in the giant panda (*Ailuropoda melanoleuca*). We scanned 32 participants with fMRI as they viewed and rated previously normed images of pandas displaying happy, sad, and neutral emotional expressions. **As predicted, emotional content robustly activated the amygdala relative to neutral content.** A whole-brain search further identified selective activations in inferotemporal cortex, precuneus, and left inferior frontal gyrus. These results held when controlling for differences in a range of low-level stimulus properties (e.g., luminance and contrast) as well as trial-by-trial differences in response time (RT), suggesting that the representations we observed were likely to be affective in nature. Mediation analysis further indicated that individual differences in amygdala activation partially mediated the relationship between experimental condition and affective ratings of the stimuli. Finally, in cross-validated searchlight MVPA analyses, we were able to correctly classify panda expression 94% of the time using amygdala activation. Collectively, our results provide compelling evidence that the neural representation of basic affective expressions in humans is not unique to humans, and that similar circuitry supports recognition of emotion in pandas and possibly other mammals.”

Causal inference

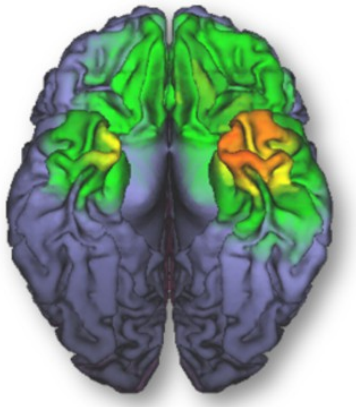
- Randomized assignment to different experimental conditions is the gold standard for a reason
- Given the design, we can safely conclude that the relationships between our manipulation and observed brain/behavior changes are causal
- Happy/sad/neutral panda images cause brain activity! Happy (and sad, and neutral) days!

Causal pandas

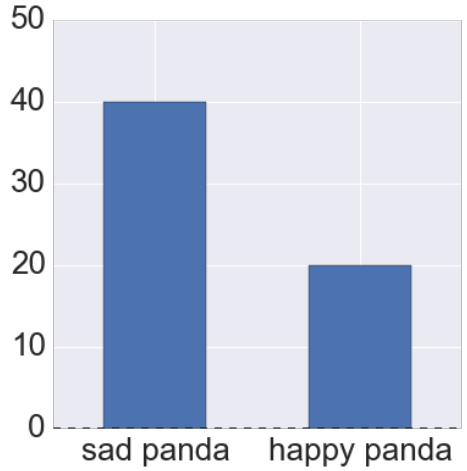
Experimental manipulation



Neural changes



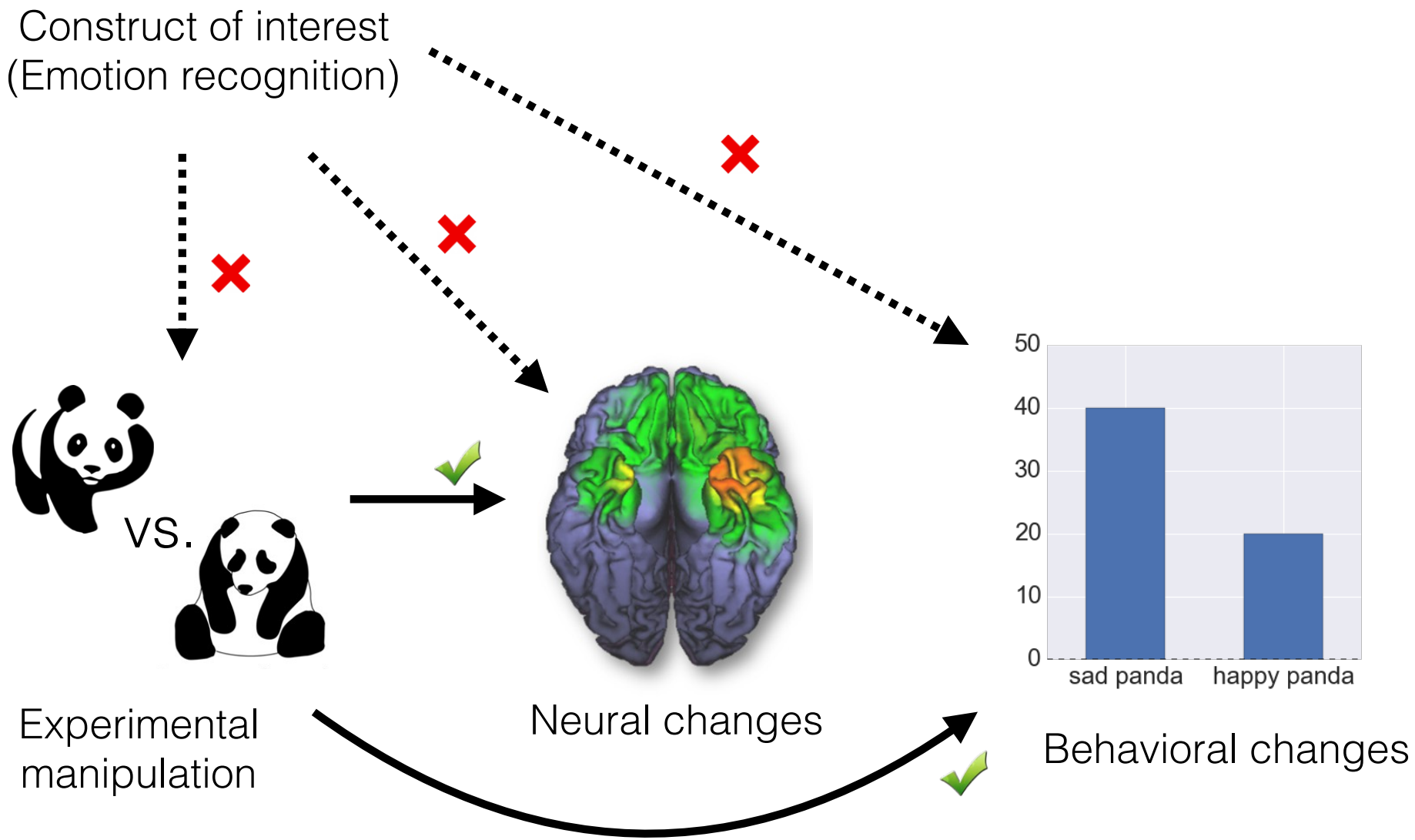
Behavioral changes



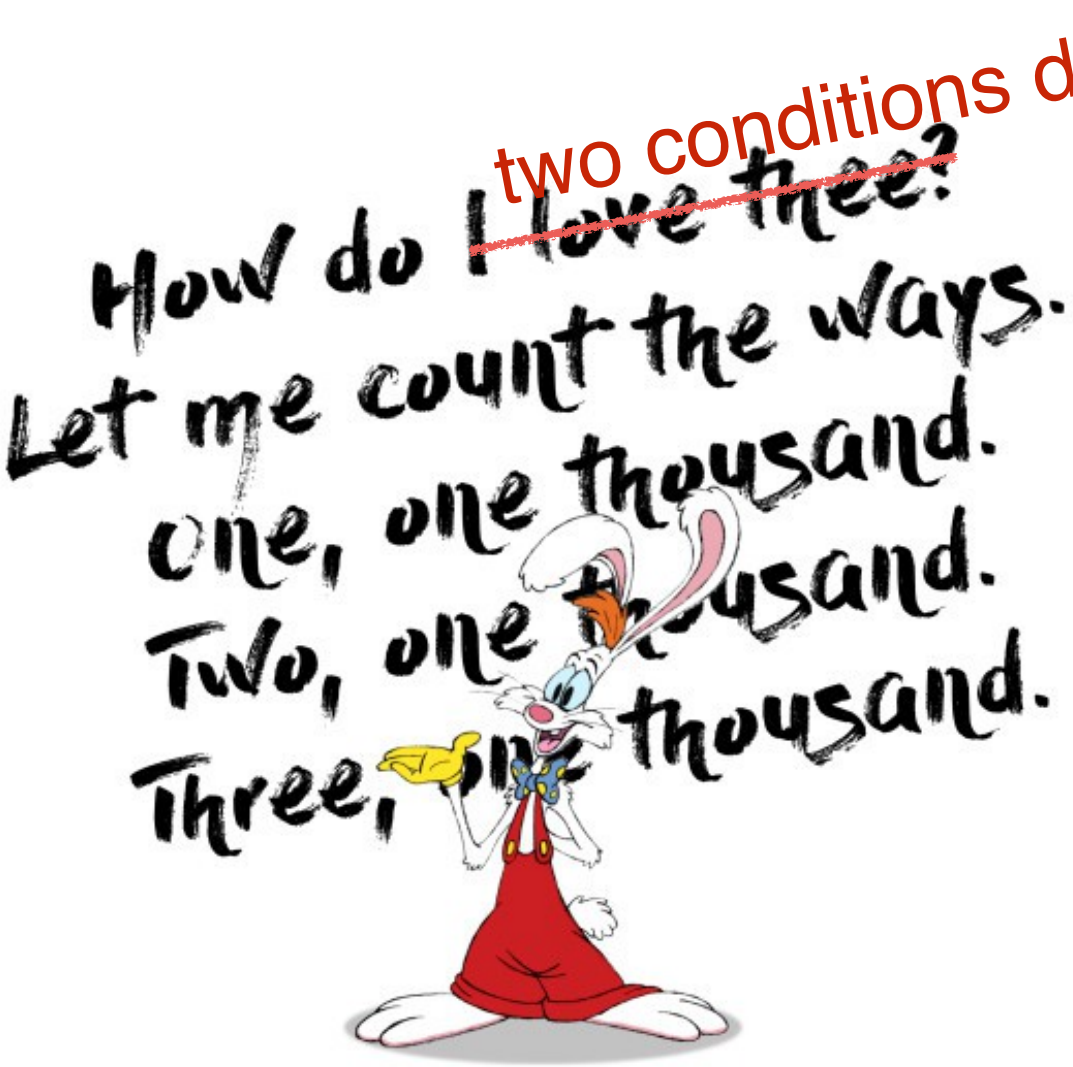
But…

- That's not what we claimed in our abstract!
- We said: "Comparison of happy, sad, and neutral panda expressions confirmed **that emotional content** activated the left amygdala relative to neutral content"
- We're implicitly assuming that our manipulation is reliably measuring the construct we care about (Cronbach & Meehl, 1955)
  - An empirical claim, not a statement of fact
  - The fact that we *intended* our manipulation to reflect emotion doesn't automatically make it so

Acausal pandas



No task is process-pure



It isn't just pandas!

- Some other common switcheroos:
  - N-back task —> working memory
  - Priming task —> Implicit processing
  - Listening to narrative text —> language/comprehension
- See also: *Neuroimage, Table of Contents of*

A useful exercise

- Next time you're tempted to label your contrast with a single construct…
- Spend 5 minutes listing all the things that could *plausibly* differ between the two conditions
- The logic of subtraction (Friston et al., 1996) requires that your list *not* be very long

About those pandas…

- Some possible differences between conditions:
  - Number of pandas in the image
  - Distance from camera
  - Gaze direction
  - Background setting
  - Ease of anthropomorphizing activity
  - Etc…
- What basis do we have for attributing activation/ behavior differences to emotion recognition?

Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 3, 615–622

0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.3.615

BRIEF REPORT

Working Memory, Attention Control, and the *N*-Back Task: A Question of Construct Validity

Michael J. Kane
University of North Carolina at Greensboro

Andrew R. A. Conway
Princeton University

Timothy K. Miura and Gregory J. H. Colflesh
University of Illinois at Chicago

The *n*-back task requires participants to decide whether each stimulus in a sequence matches the one that appeared *n* items ago. Although *n*-back has become a standard "executive" working memory (WM) measure in cognitive neuroscience, it has been subjected to few behavioral tests of construct validity. A combined experimental–correlational study tested the attention-control demands of verbal 2- and 3-back tasks by presenting $n - 1$ "lure" foils. Lures elicited more false alarms than control foils in both 2- and 3-back tasks, and lures caused more misses to targets that immediately followed them compared with control targets, but only in 3-back tasks. *N*-back thus challenges control over familiarity-based responding. Participants also completed a verbal WM span task (operation span task) and a marker test of general fluid intelligence (Gf; Ravens Advanced Progressive Matrices Test; J. C. Raven, J. E. Raven, & J. H. Court, 1998). *N*-back and WM span correlated weakly, suggesting they do not reflect primarily a single construct; moreover, both accounted for independent variance in Gf. *N*-back has face validity as a WM task, but it does not demonstrate convergent validity with at least 1 established WM measure.

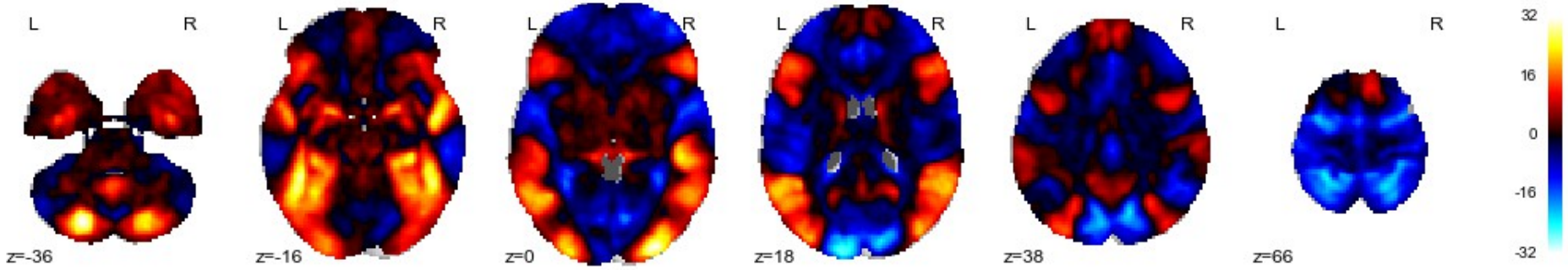
Keywords: working memory, memory span, *n*-back, intelligence, individual differences

Construct validity in the HCP

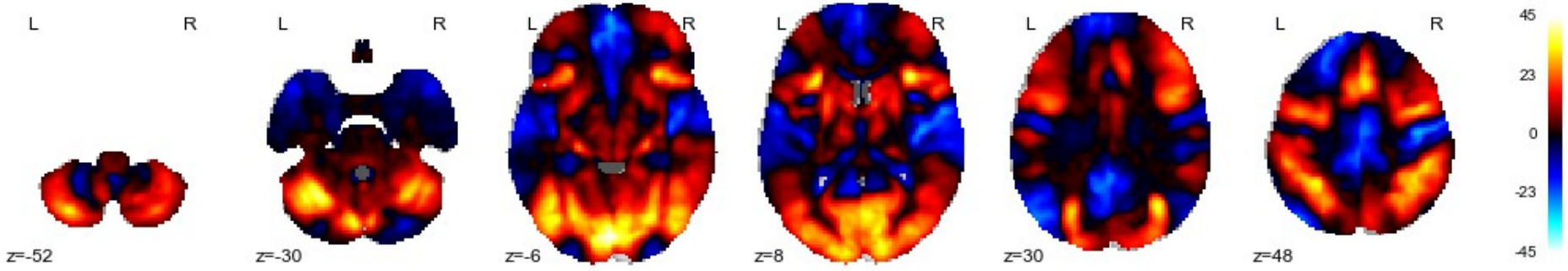
- The HCP contains tasks that putatively assess:
  - Working memory
  - Incentive processing
  - Language processing
  - Social cognition
  - Emotion processing
  - Motor processing
- What are these labels based on?
  - Usually, only authors' beliefs and expectations

Construct validity in the HCP

Biological > random motion (HCP “social cognition” task)



Reward > loss (HCP “incentive processing” task)



A question of construct validity

- The fallacy of affirming the consequent rears its head again
  - a.k.a. reverse inference
  - a.k.a. why many people misunderstand *p*-values
- We all know that P(Data|Hypothesis) != P(Hypothesis|Data)
- But it may be helpful to break this into *two* inequalities:
  - P(Effect|Data) != P(Data|Effect)
  - P(Theory|Effect) != P(Effect|Theory)
- Oh no! Now there are *two* reverse inference problems!
- Neuroimaging researchers talk about the former, but almost completely neglect the latter

How do we address this?

- Think about, and explicitly test, the validity of your measures/tasks
- “Control for” as many variables as you can*
- Use multiple measures of the same construct
- Use latent-variable models (e.g., SEM)
- Read the psychometric literature
  - Psychometricians have thought about these issues for ~100 years!
  - “Whatever you do, somebody in psychometrics already did it long before.” — Andrew Gelman

Interim conclusions

- Drawing measurement-level causal inferences can be easy with the right (experimental) design
- Drawing construct-level inferences is *much* harder —and randomized designs aren't sufficient!
- There is a huge literature on this stuff; we ignore it at our own peril in neuroimaging

The Sad Panda study

- "Previous fMRI studies have demonstrated that specific brain circuits such as the amygdala are selectively implicated in representing human emotional expressions. The degree to which this finding generalizes to the rest of the animal kingdom is unknown. Here we investigated the (human) neural representation of emotional expression in the giant panda (*Ailuropoda melanoleuca*). We scanned 32 participants with fMRI as they viewed and rated previously normed images of pandas displaying happy, sad, and neutral emotional expressions. As predicted, emotional content robustly activated the amygdala relative to neutral content. A whole-brain search further identified selective activations in inferotemporal cortex, precuneus, and left inferior frontal gyrus. **These results held when controlling for differences in a range of low-level stimulus properties (e.g., luminance and contrast) as well as trial-by-trial differences in response time (RT), suggesting that the representations we observed were likely to be affective in nature.** Mediation analysis further indicated that individual differences in amygdala activation partially mediated the relationship between experimental condition and affective ratings of the stimuli. Finally, in cross-validated searchlight MVPA analyses, we were able to correctly classify panda expression 94% of the time using amygdala activation. Collectively, our results provide compelling evidence that the neural representation of basic affective expressions in humans is not unique to humans, and that similar circuitry supports recognition of emotion in pandas and possibly other mammals."

Controlling for stuff

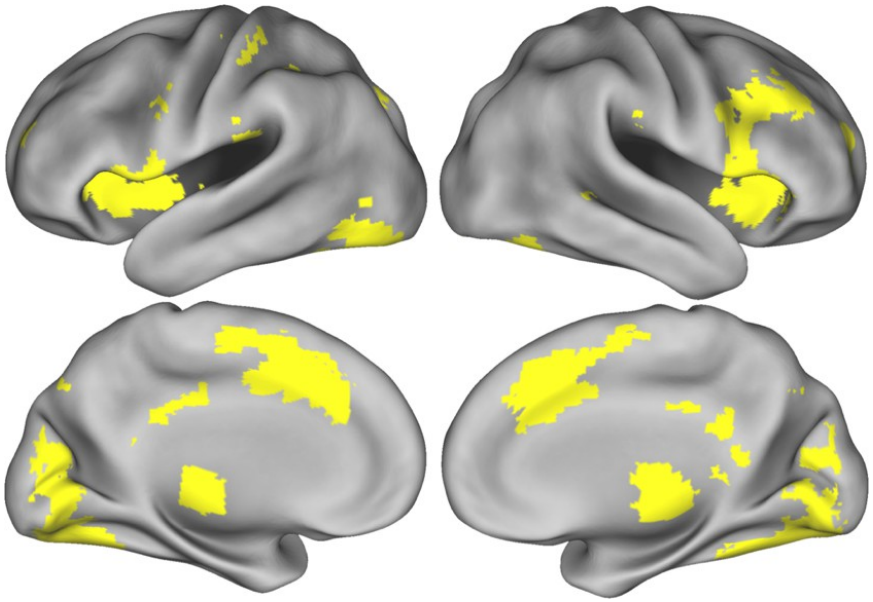
- Got confounds? No problem!
- Just add them to the model!
- Right?

The case of response time

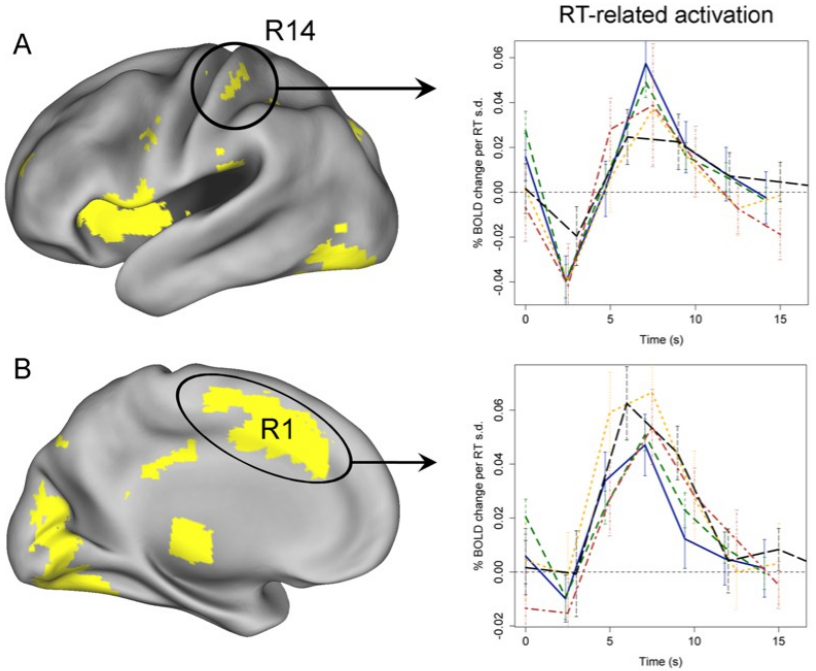
- People's response times vary when they do different tasks—both within and between individual
- Brain activity often varies systematically in proportion to trial-by-trial differences in RT (Grinband et al., 2008; Yarkoni et al., 2009)
- These effects are often very strong

Consistent trial-by-trial effects in 5 different paradigms (WM, emotion, decision-making…)

Positive correlations with RT



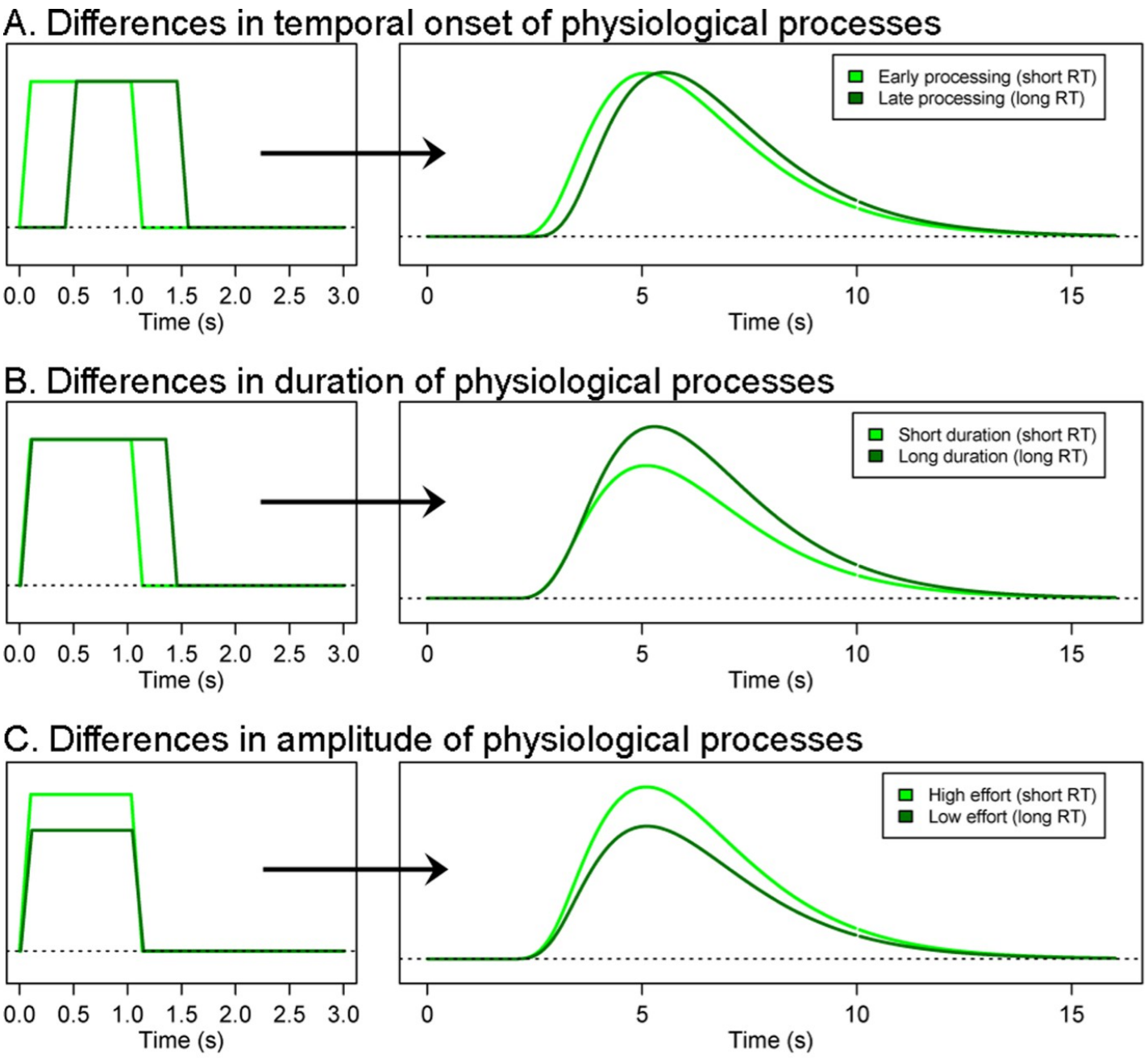
Time course of RT-related activation



Yarkoni et al. (2009)

A mysterious convergence?

- Why do RT-sensitive regions resemble the task-positive/salience/task set/goal-directed attention/ multiple-demand network?
- Is this a mysterious coincidence?
  - No! It's entailed by the linear, time-invariant nature of the HRF…



Yarkoni et al. (2009)

Implications

- If you don't examine results both with and without RT in the model, you don't know if activation differences reflects a quantitative or qualitative difference in processing
  - E.g., do schizophrenics show hyperactivation/inefficiency of PFC during WM tasks (e.g., Potkin et al., 2009) only because they respond more slowly, or are there residual differences?
- A serious concern (Yarkoni et al. 2009; Grinband et al., 2011), but very few task-related studies explicitly control for RT differences
- If the relationship between response (or inspection) time and BOLD activity is nonlinear, all bets are off
- But as long as things are (roughly) linear, we can control for things like this statistically, right? Right?

PLOS ONE

RESEARCH ARTICLE

Statistically Controlling for Confounding Constructs Is Harder than You Think

Jacob Westfall*, **Tal Yarkoni**

University of Texas at Austin, Austin, TX, United States of America

* jake.westfall@utexas.edu

Abstract

Social scientists often seek to demonstrate that a construct has *incremental validity* over and above other related constructs. However, these claims are typically supported by measurement-level models that fail to consider the effects of measurement (un)reliability. We use intuitive examples, Monte Carlo simulations, and a novel analytical framework to demonstrate that common strategies for establishing incremental construct validity using multiple regression analysis exhibit extremely high Type I error rates under parameter regimes common in many psychological domains. Counterintuitively, we find that error rates are highest—in some cases approaching 100%—when sample sizes are large and reliability is moderate. Our findings suggest that a potentially large proportion of incremental validity claims made in the literature are spurious. We present a web application (http://jakewestfall.org/ivy/) that readers can use to explore the statistical properties of these and other incremental validity arguments. We conclude by reviewing SEM-based statistical approaches that appropriately control the Type I error rate when attempting to establish incremental validity.

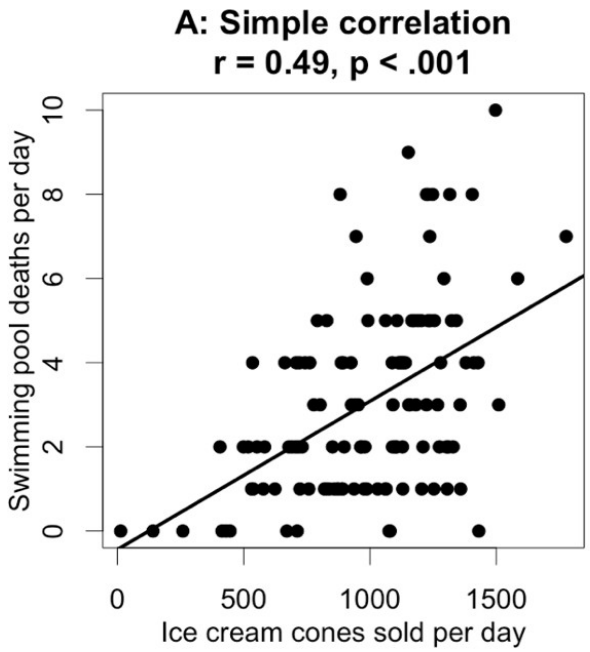
Citation: Westfall J, Yarkoni T (2016) Statistically Controlling for Confounding Constructs Is Harder than You Think. PLoS ONE 11(3): e0152719. doi:10.1371/journal.pone.0152719

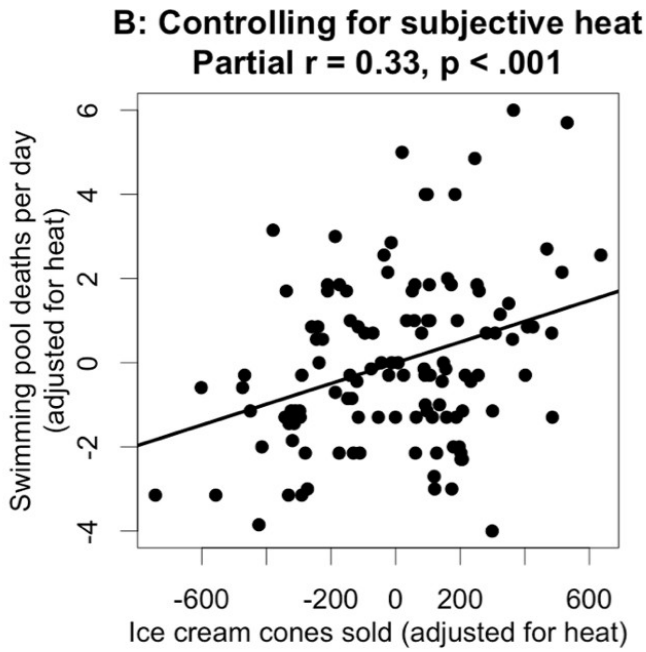
Editor: Ulrich S Tran, University of Vienna, School of Psychology, AUSTRIA

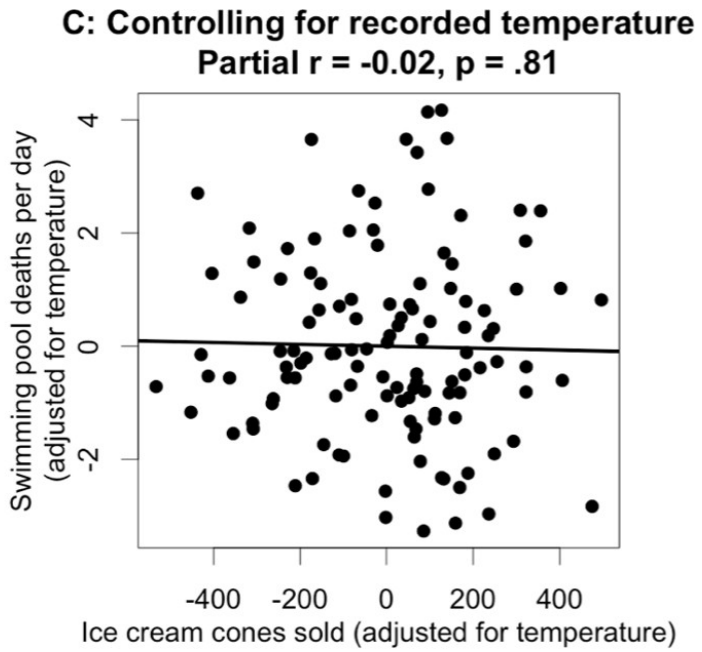
Measures vs. constructs

- Controlling for a measure does not mean you've removed the influence of the targeted *construct*
  - E.g., controlling for self-reported scores on the single item "I feel sleepy" is not the same as "controlling for fatigue"
- The two will only coincide when the measure perfectly captures the construct
  - I.e., virtually never

Mmm, ice cream

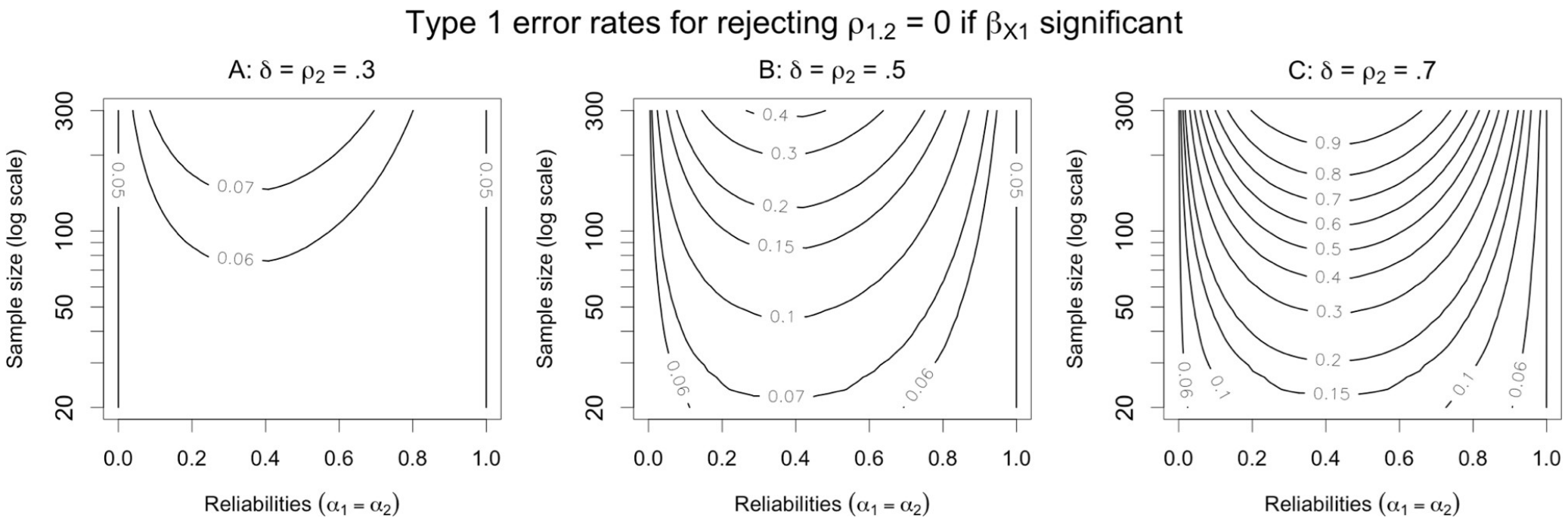






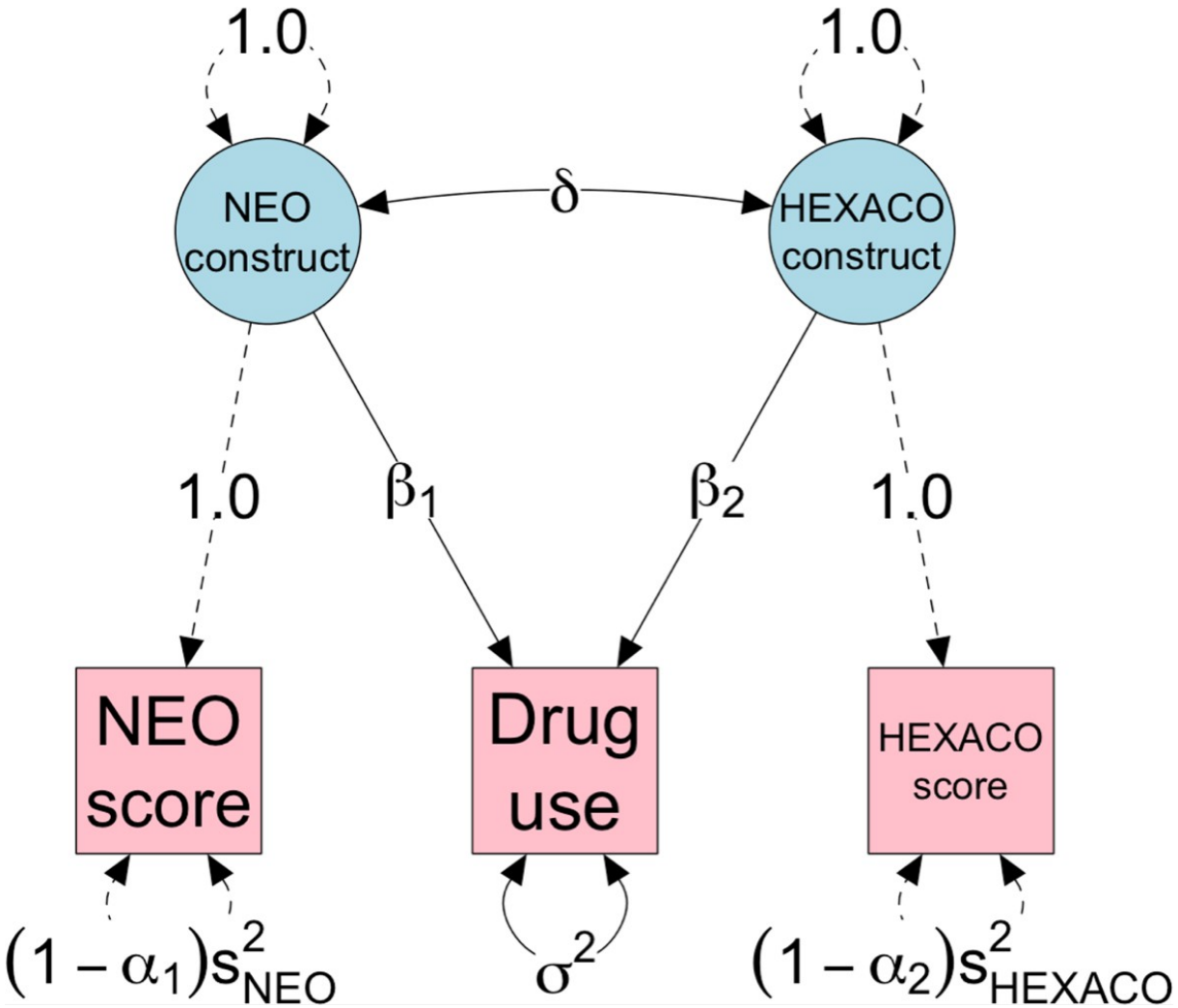
Westfall & Yarkoni (2016)

Mmm, Type I error



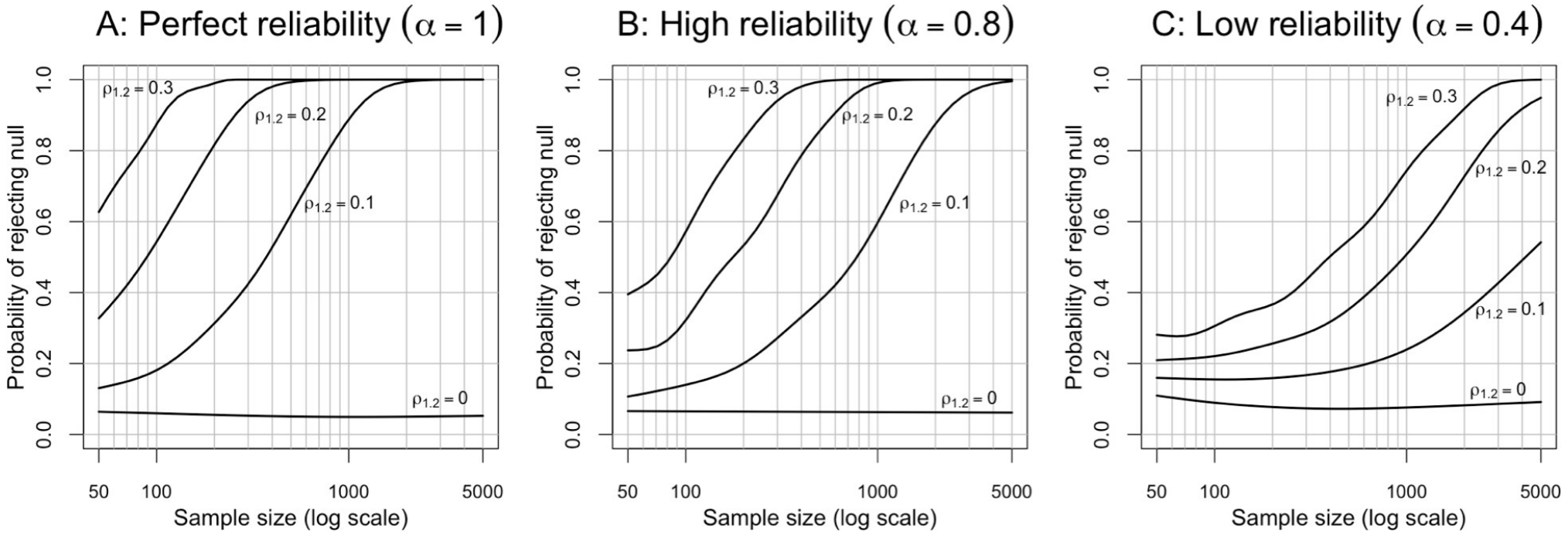
Westfall & Yarkoni (2016)

Mmm, SEM



Westfall & Yarkoni (2016)

Power to detect incremental validity at the latent-variable level



Westfall & Yarkoni (2016)

Conclusions

- Statistical control of potential confounds is hard
- Total latent variable-level control may be impossible under realistic parameter regimes in fMRI
  - Problem gets worse the more variables you attempt to control for
- This doesn't mean we shouldn't include covariates
  - It does mean you probably shouldn't say things like "we controlled for SES, mood, and fluid intelligence"
  - Gender, age, etc. may be fine (why?)

General conclusions

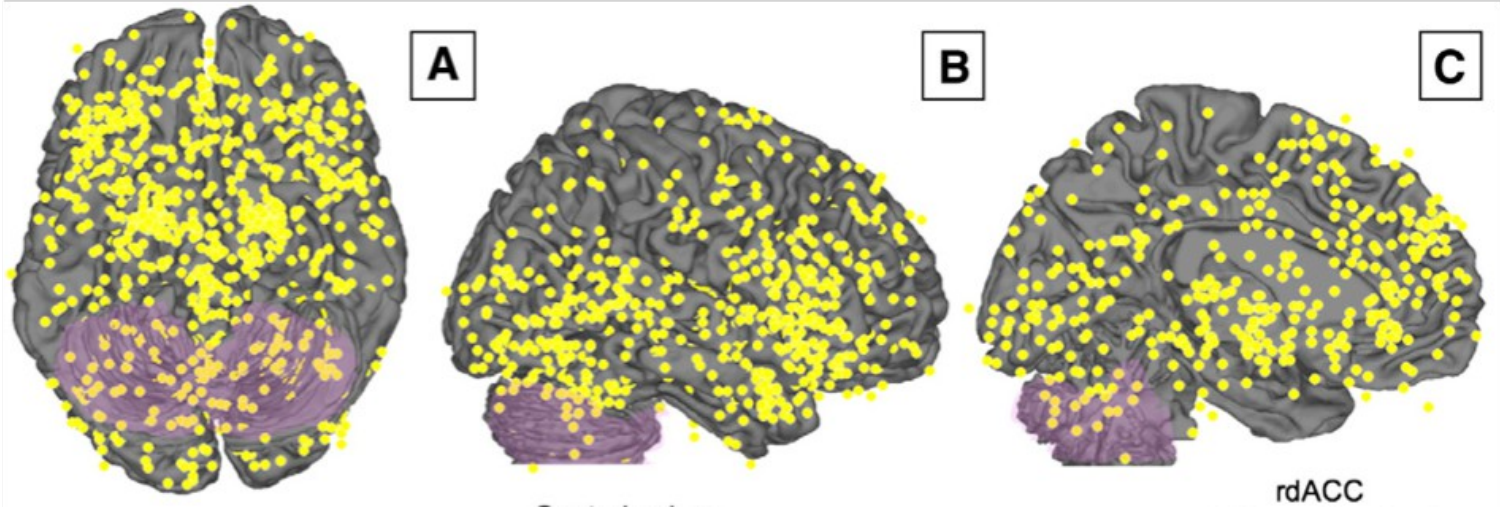
- Learning important new things is quite hard
- Most individual studies can't teach us very much
  - Not necessarily *wrong*
  - The generalizability of results is likely to be low
- What should we do about this?

Meta-analysis

- One study tells us very little… maybe several hundred studies can tell us less little?

A model of clarity

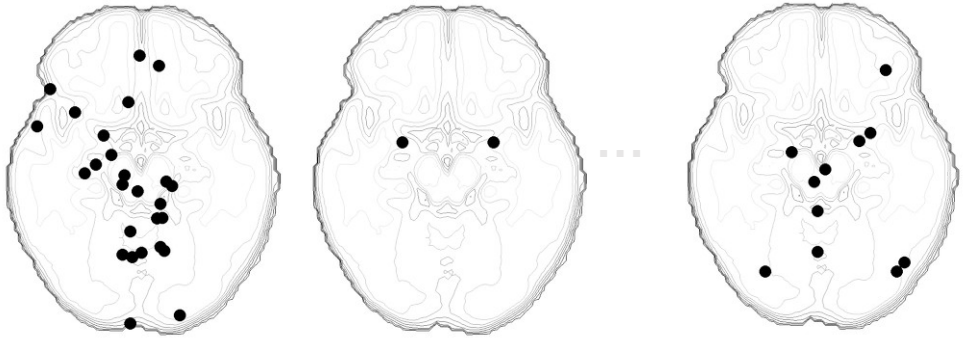
Activations from 437 emotion contrasts



Kober et al. (2008)

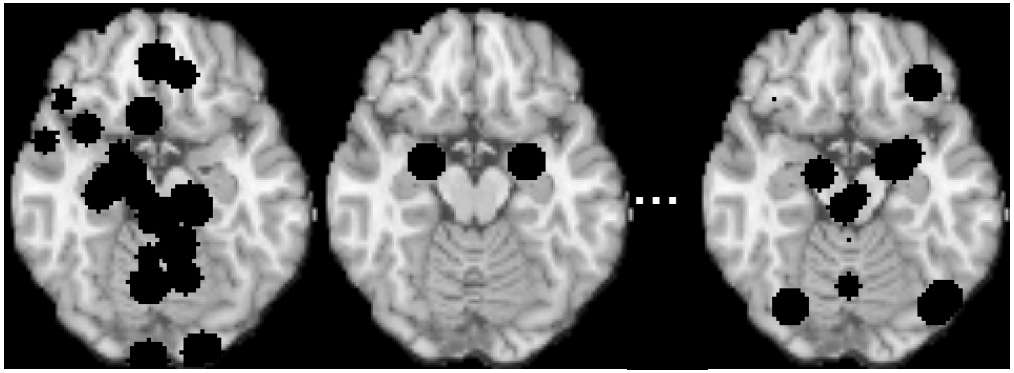
MKDA: Multilevel kernel density analysis

Peak coordinate locations (437 maps)



Damasio, 2000 Liberzon, 2000 Wicker, 2003

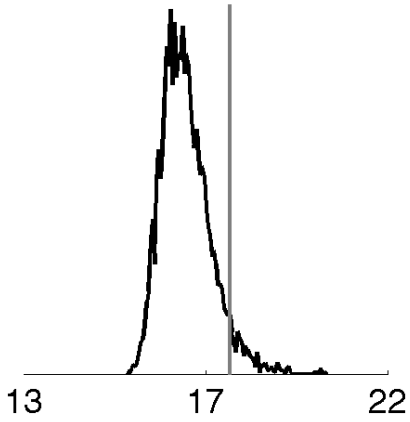
Kernel convolution



Comparison indicator maps

Weighted average

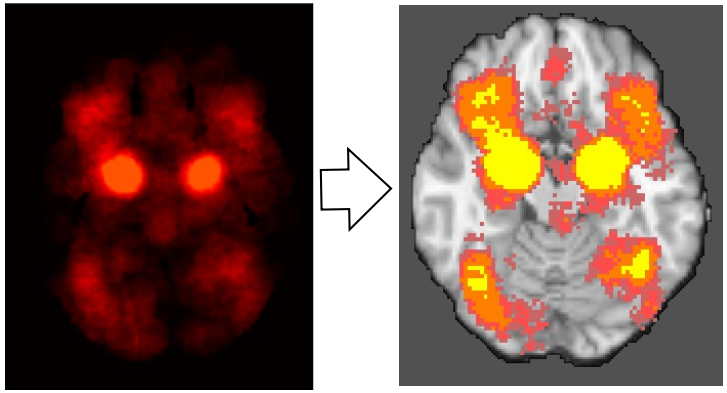
Monte Carlo:
Expected maximum proportion
Under the null hypothesis



Apply threshold

Randomize locations of blobs within study maps

Significant regions



Proportion of activated Comparisons map

Wager, Lindquist, & Kaplan, 2007; others

Slide courtesy of Tor Wager

Weighting by sample size and quality

MKDA analysis weights by sqrt(sample size) and study quality (including fixed/random effects)

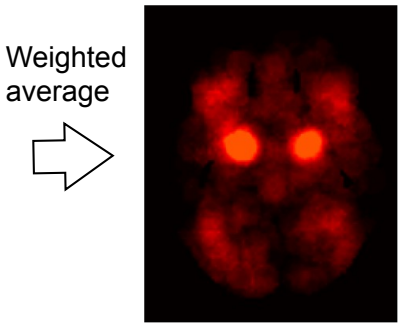
Weighted proportion of activating studies

$$P = \sum_c CIM_c \left(\frac{\delta_c \sqrt{N_c}}{\sum_c \delta_c \sqrt{N_c}} \right)$$

Study quality weight

Sample size for map *c*

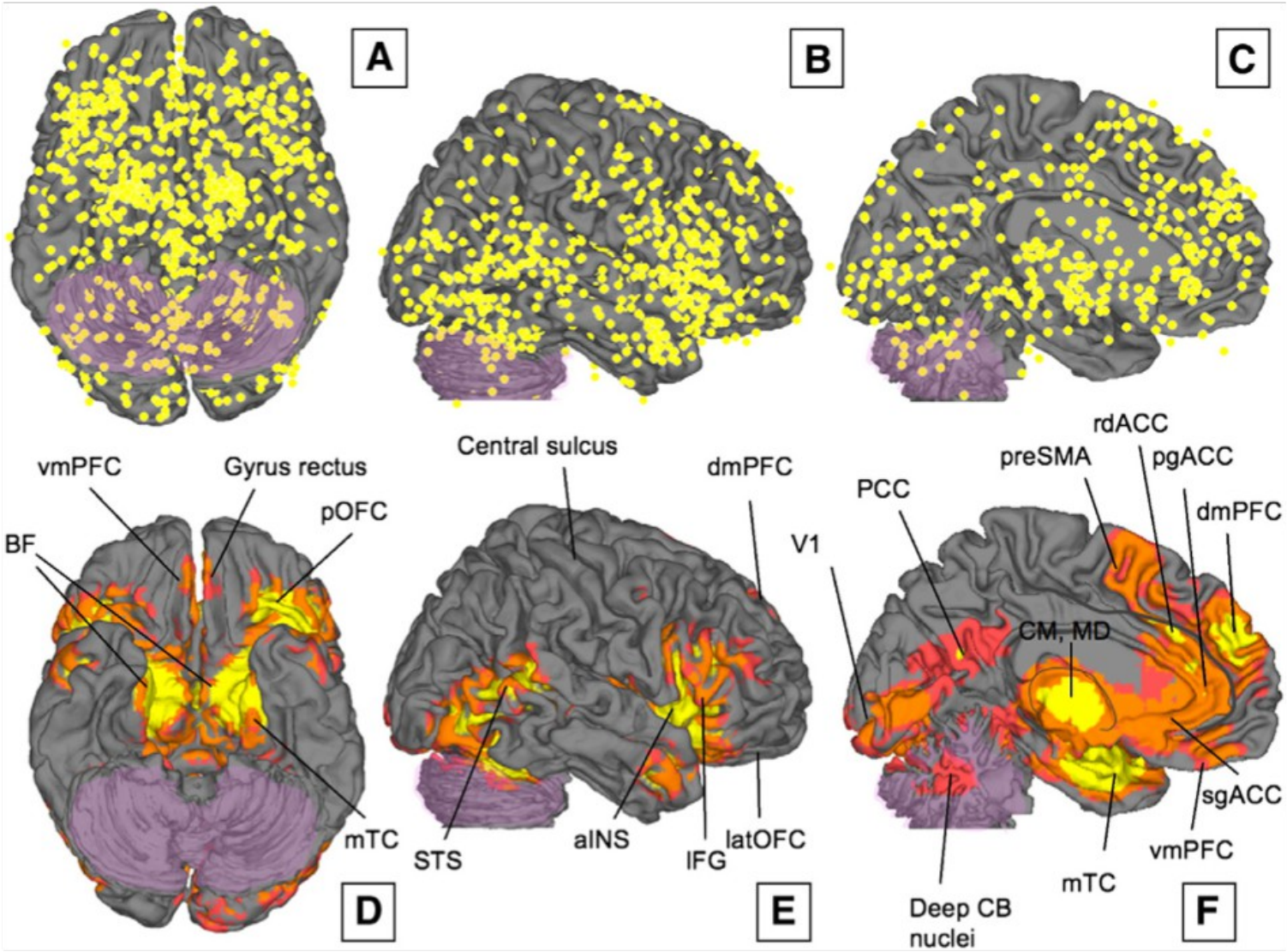
Activation indicator (1 or 0) for map *c*



$\delta_c = 1$ *Fixed effects*

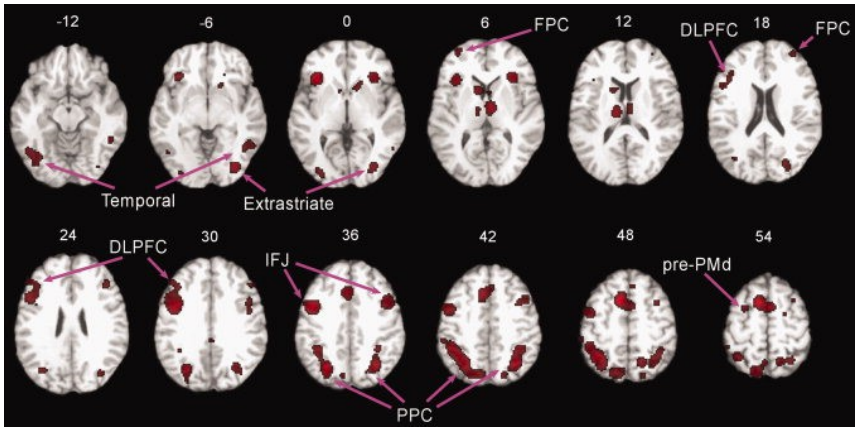
$\delta_c = 0.75$ *Random effects*

It works!

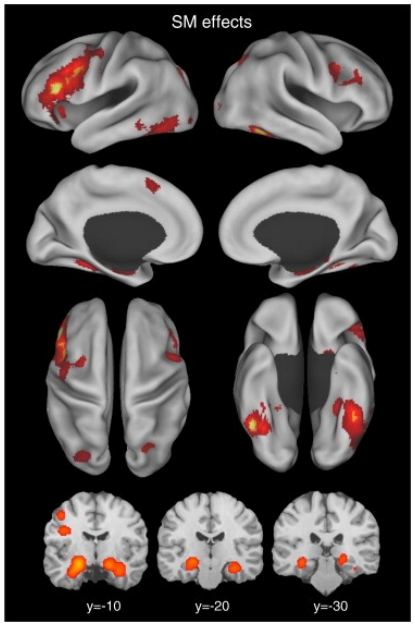


Kober et al. (2008)

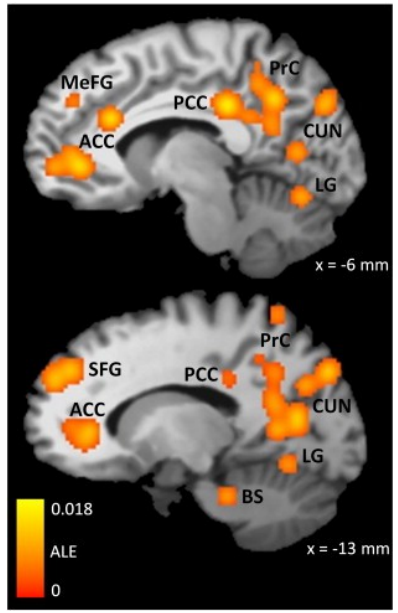
Meta-analyze all the things!



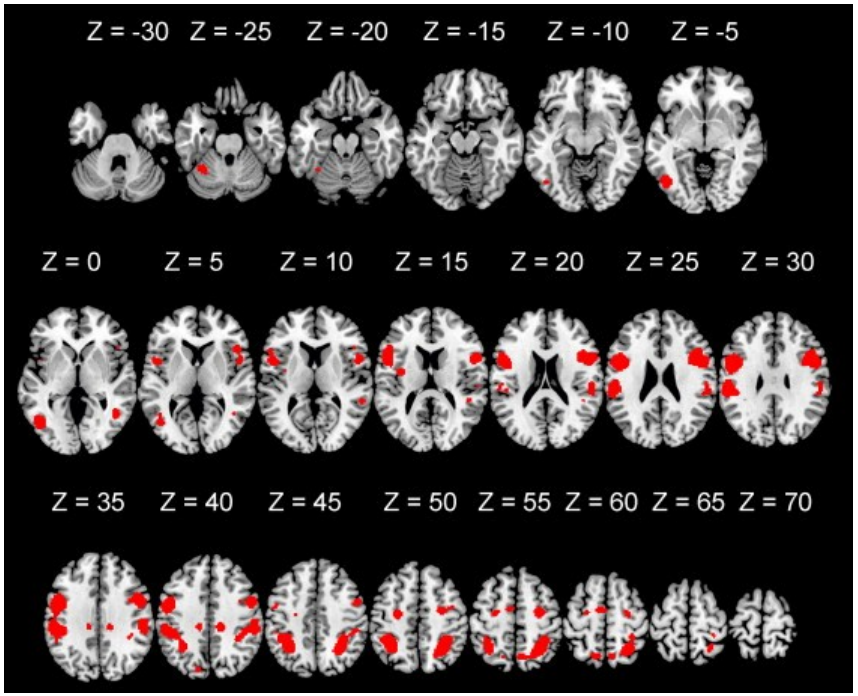
Kim et al. (2011) - Task-switching



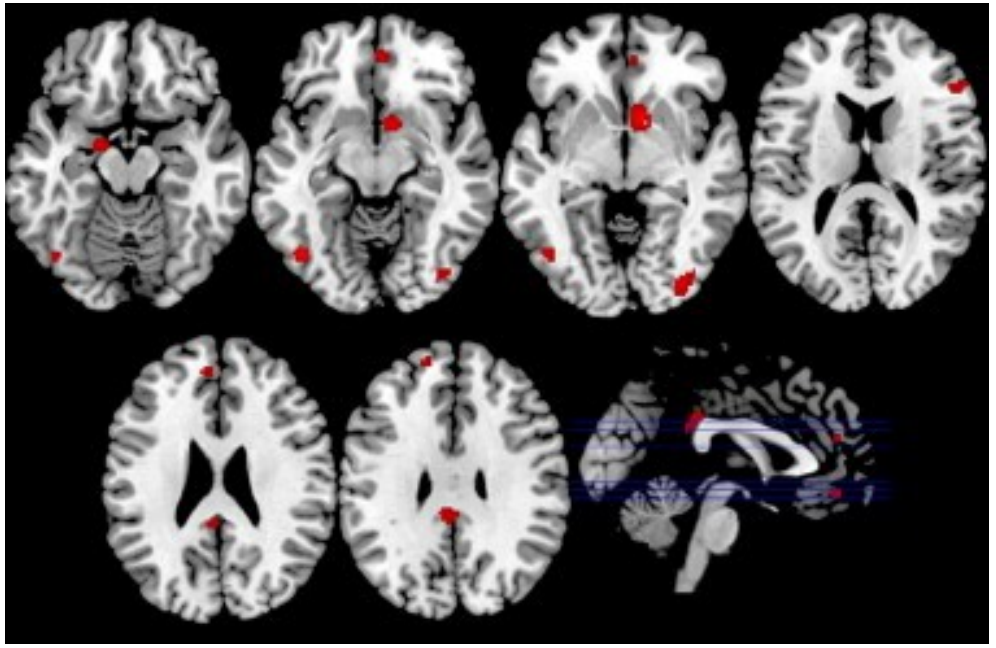
Kim (2011)
Subsequent memory



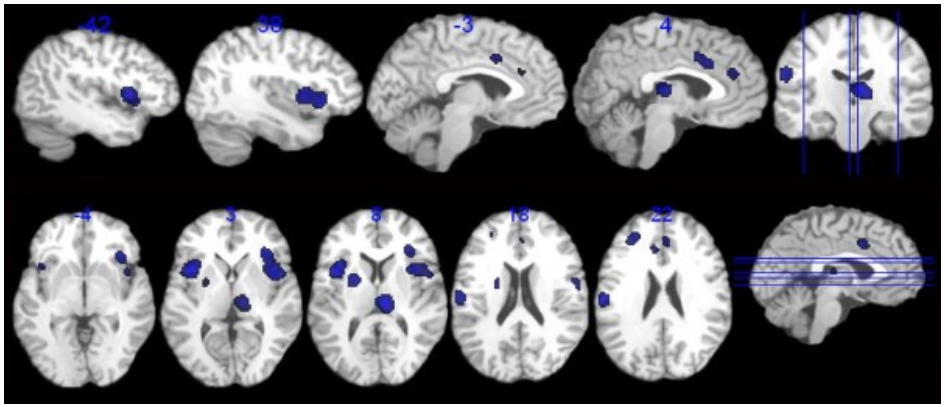
Engelmann et al. (2011)
Smoking cue reactivity



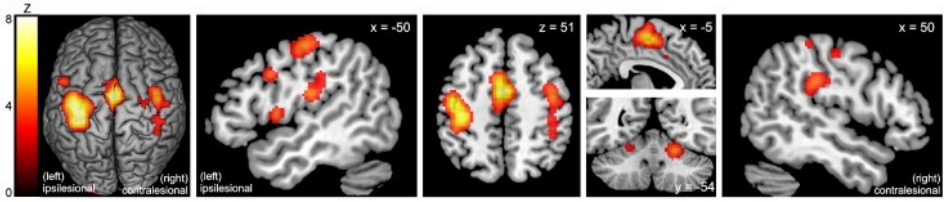
Molenberghs et al (2011) - Mirror system



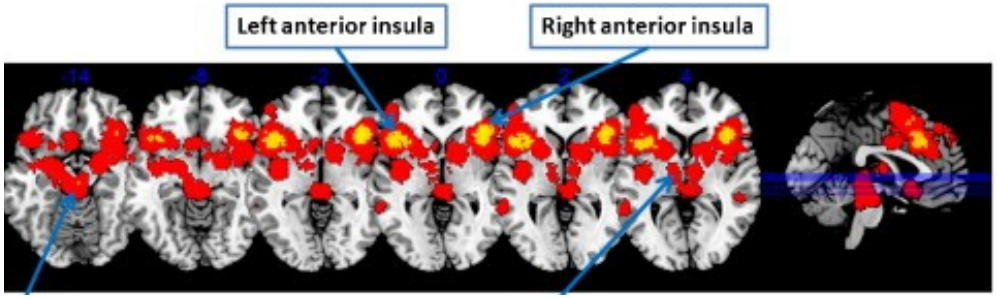
Chase et al. (2011) - Drug craving



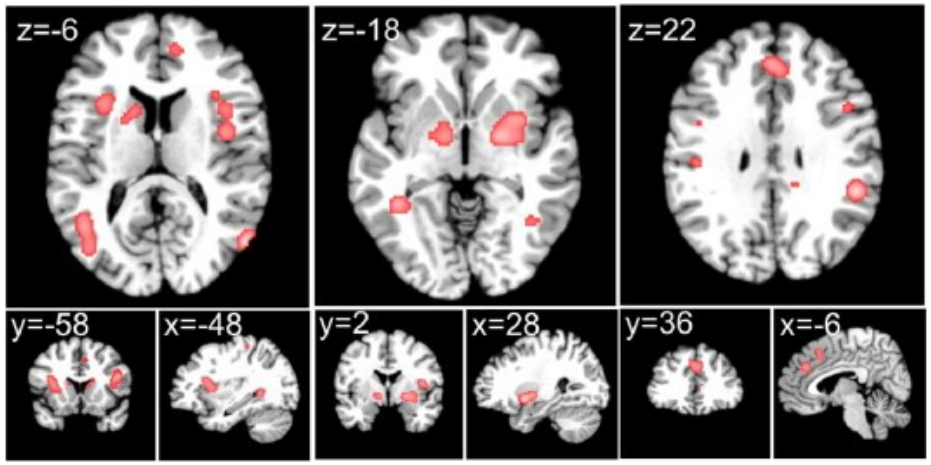
Tillisch et al. (2011) - Rectal distension



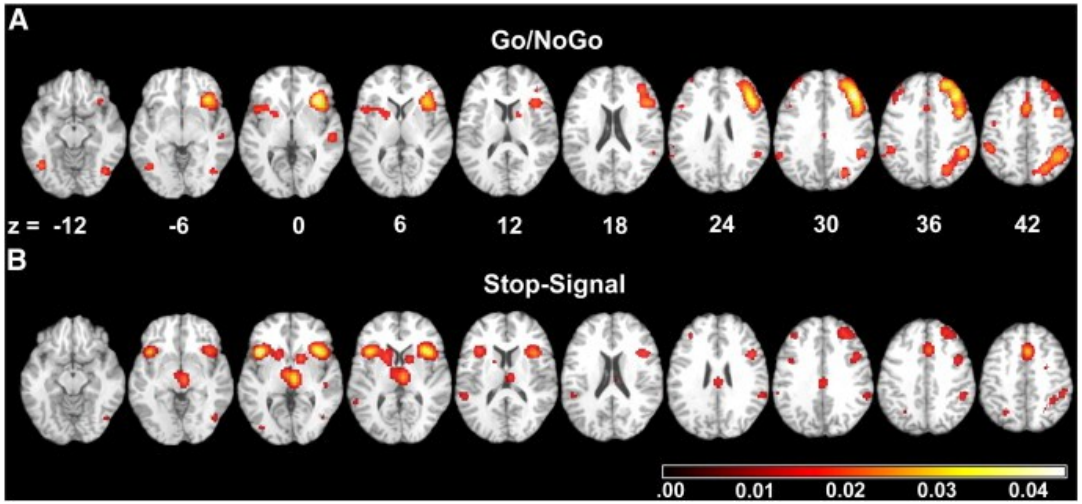
Rehme et al. (2012) - Movement after stroke



Fan et al. (2011) - Empathy



Brooks et al. (2011) Subliminal arousing stimuli



Swick et al. (2011) Go/NoGo & Stop-Signal

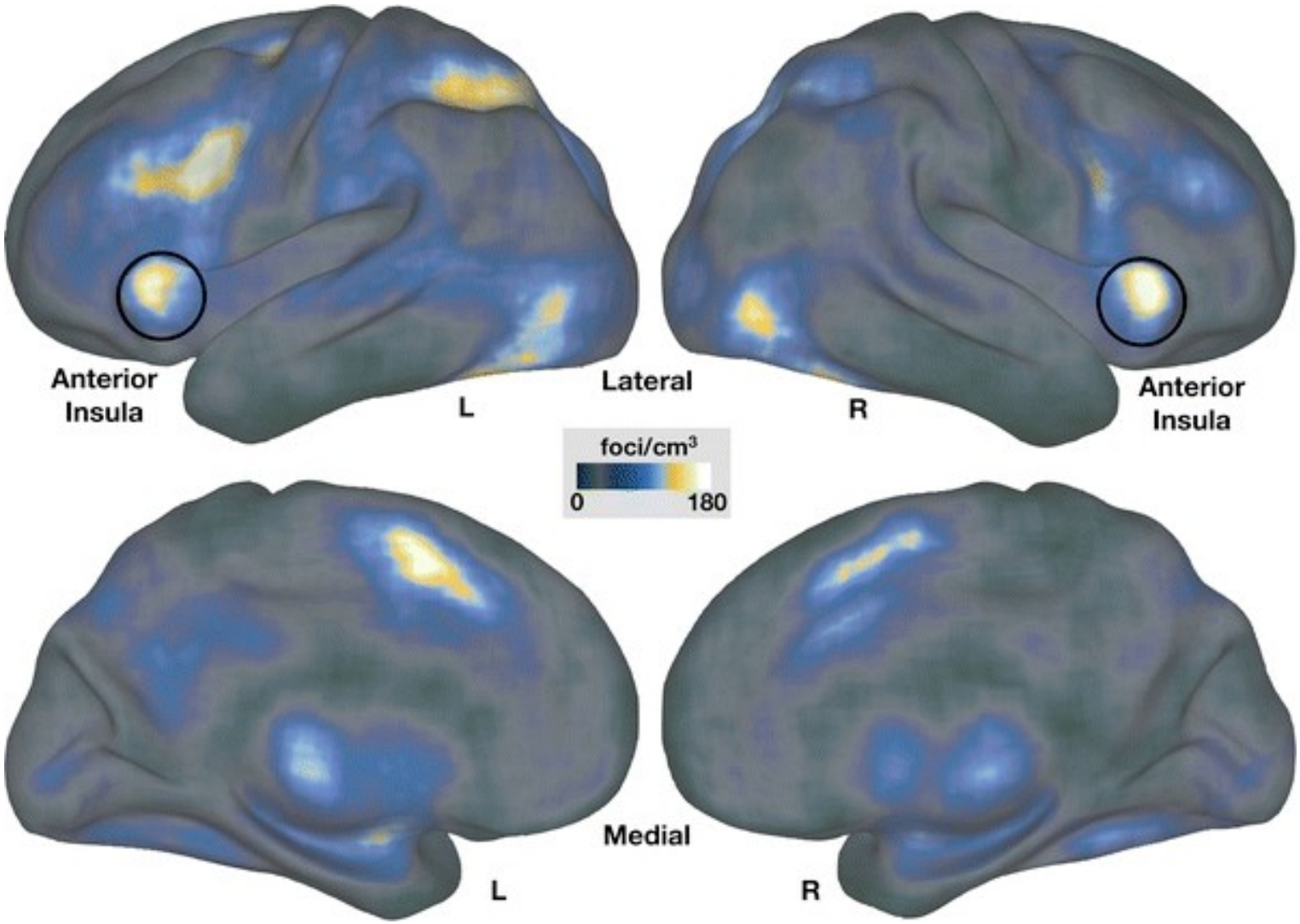
It works!

- Standard fMRI meta-analyses can solve some of our big problems
- Increases sensitivity, overcomes sample size issues
- Establishes the generalizability of a result across experimental designs, stimuli, labs, scanners, populations, etc…

But there's a catch!

- Notice anything odd about the distribution of regions in the previous meta-analyses?
- Some of them show up repeatedly…
- Standard meta-analysis approaches like MKDE and ALE *don't* solve the problem of reverse inference
  - They tell us what regions are *consistently* activated by a task or process
  - But not what regions are *preferentially* activated

All regions are not equal(ly active)



Nelson et al. (2010)

E.g., anterior insula is “selectively” activated by...

- Disgust
- Pain
- Empathy
- Risk
- Interoceptive awareness
- Autonomic arousal
- Self-reflection
- Task-switching
- Conscious error perception
- Response inhibition
- Speech production
- Sustained attention
- Etc...

Towards *wide* datasets

- Not enough to have a lot of subjects doing one task
- Need to compare one task/process with *many* others
- We want to ask: what tasks/processes are *most* likely to activate a given region/network?
- So we need to assemble a really diverse dataset
  - Also, we're very lazy
- Wouldn't it be nice if we could automate the process of assembling, and meta-analyzing, a giant database of fMRI studies?


GOOD NEWS, EVERYONE!
WWW.MEMEGASMS.COM

Towards an automated meta-analysis platform

Two challenges

- Data standardization
- Semantic annotation

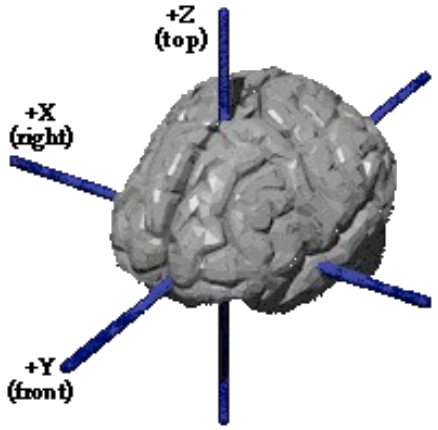
- Our approach: ignore both problems

Assumption 1: If it looks like a duck...

- Anything that looks like brain activation *is* brain activation
- Parser looks for x/y/z-like numbers in sequence within HTML tables

Table 1
Regions that showed a condition × time interaction in the ANOVA analysis

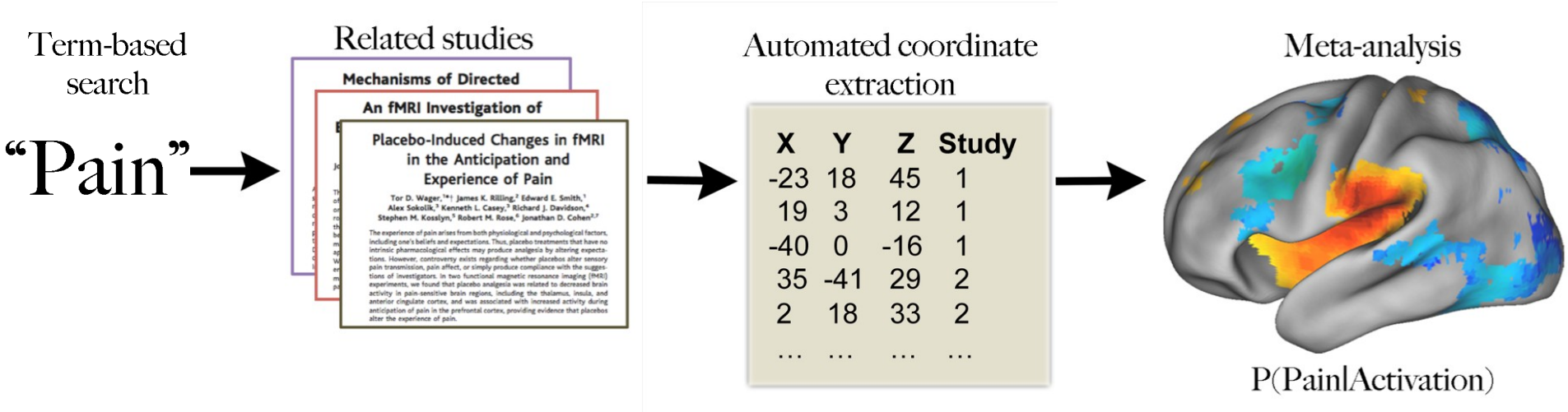
No.	Region	Hemisphere	BA	*x*	*y*	*z*	mm^3
1	Middle/superior temporal gyrus	L	21/22/37	−52	−54	9	13257
2	Inferior frontal gyrus	L	45/46/9	−49	26	6	2781
3	Posterior cerebellum	L		−19	−79	−38	2214
4	Dorsomedial PFC	L	9/8	−11	42	47	3051
5	Left anterior PFC	L	10	−37	49	15	2025
6	Inferior parietal cortex	L	40/7	−42	−58	47	3132
7	Dorsal premotor cortex	L	6	−43	0	50	1485
8	Lingual gyrus	L	17	−10	−95	−2	378
9	Middle /superior temporal gyrus	R	21/22/37	52	−40	5	16470
10	Inferior frontal gyrus	R	45/46	51	28	6	2241
11	Posterior cerebellum	R		23	−78	−34	2808
12	Dorsomedial PFC	R	9	5	53	29	405
13	Right anterior PFC	R	10	38	42	21	5022
14	Inferior parietal cortex	R	40/7	42	−53	48	9963
15	Superior frontal gyrus	R	6/8	10	28	60	297
16	Anterior cingulate cortex	M	32	0	26	35	5076
17	Posterior cingulate cortex	M	23/31/7	0	−35	31	9612
18	Precuneus	M	7/19	1	−76	36	10044



Assumption 2: A bag of words

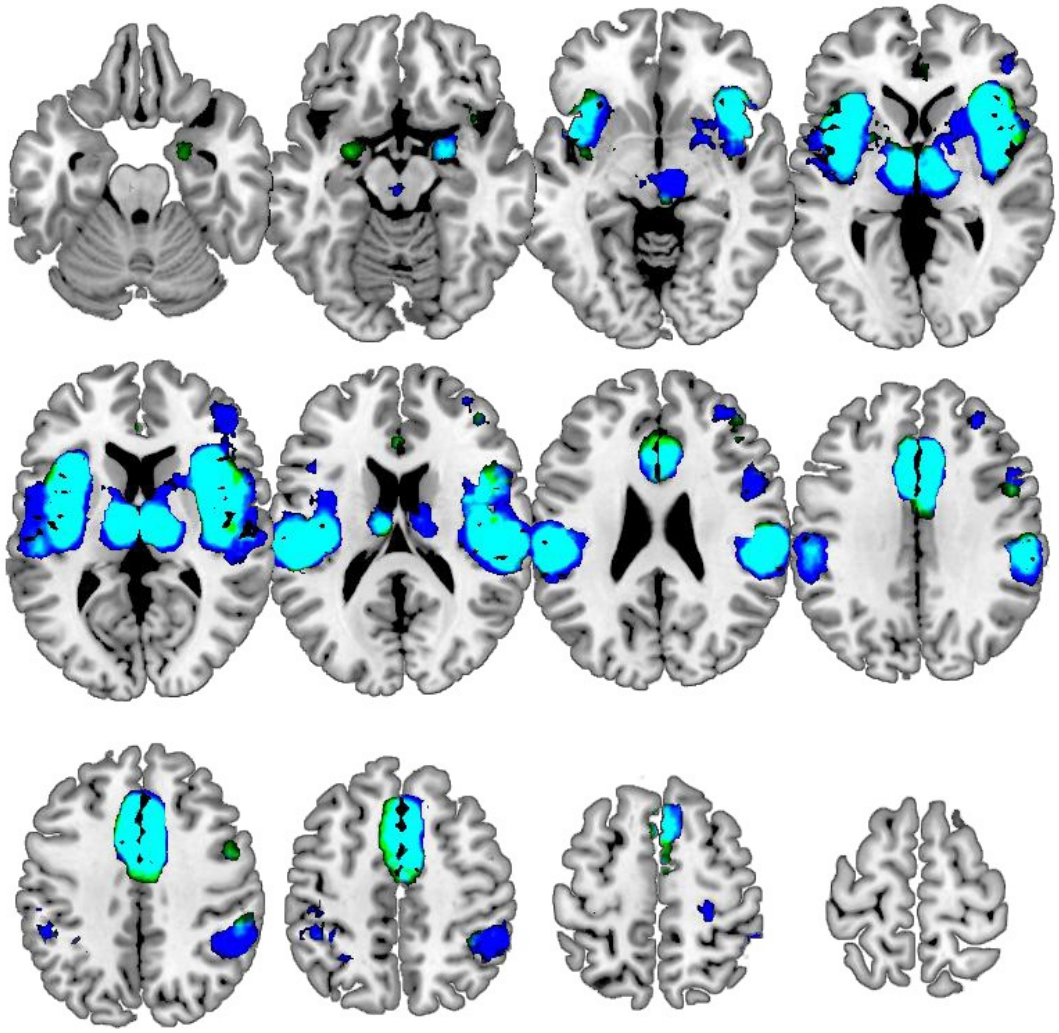
- Each published fMRI article is an unordered list of words
- A study is *about* the words it uses with high frequency

The approach



Yarkoni et al (2011)

Pain vs. pain



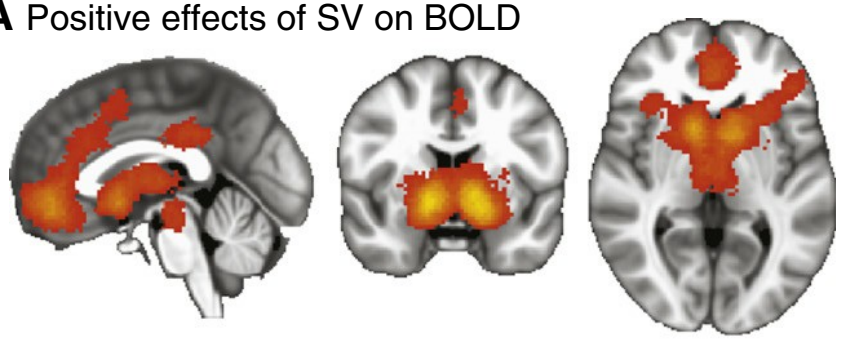
Blue: manual Green: automated

Yarkoni et al (2011)

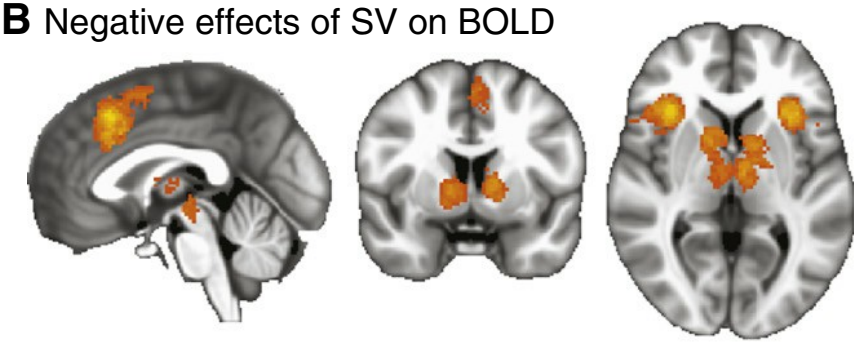
Reward vs. reward

Manual meta-analysis

A Positive effects of SV on BOLD

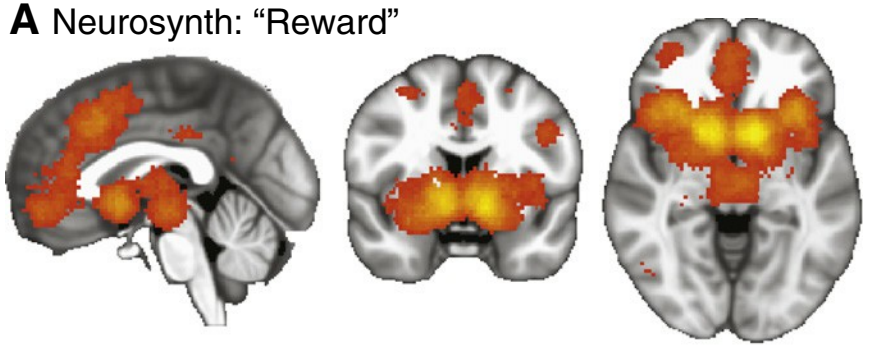


B Negative effects of SV on BOLD

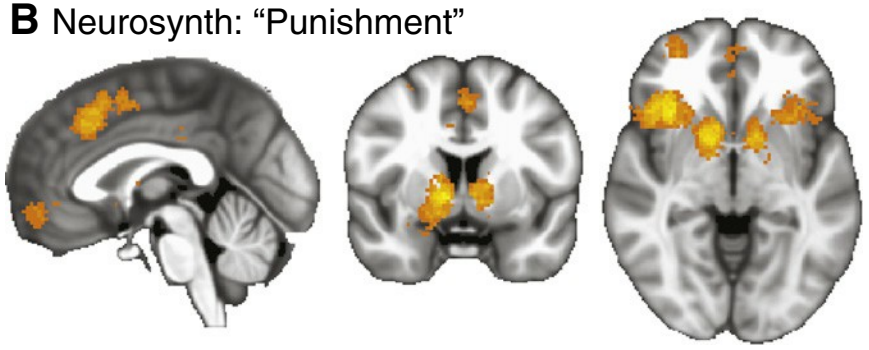


Neurosynth

A Neurosynth: “Reward”

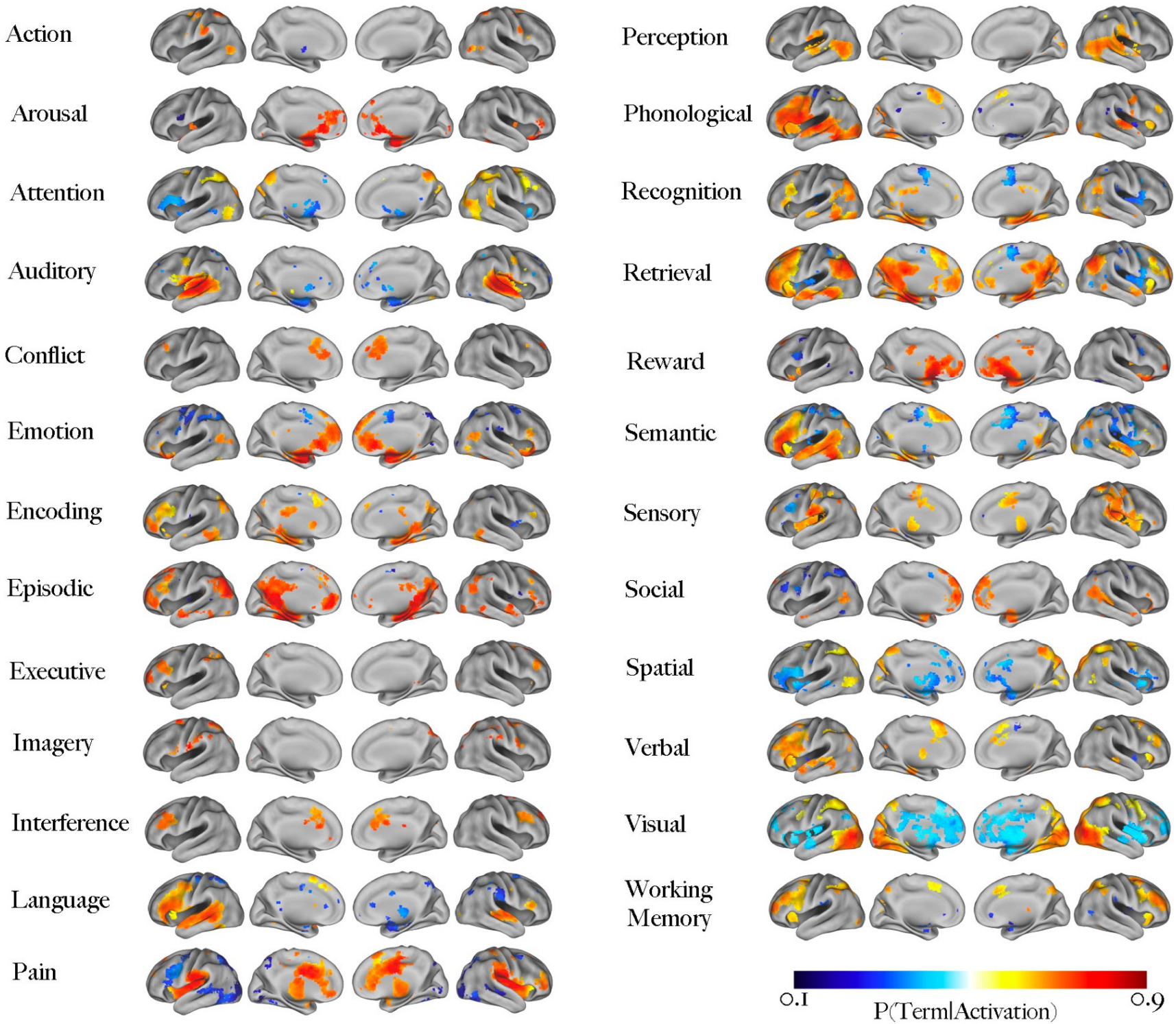


B Neurosynth: “Punishment”



x = 0 *y* = 4 *z* = -4

Bartra, McGuire, & Kable (2013)


Action
Arousal
Attention
Auditory
Conflict
Emotion
Encoding
Episodic
Executive
Imagery
Interference
Language
Pain
Perception
Phonological
Recognition
Retrieval
Reward
Semantic
Sensory
Social
Spatial
Verbal
Visual
Working Memory
0.1
0.9
P(Term|Activation)

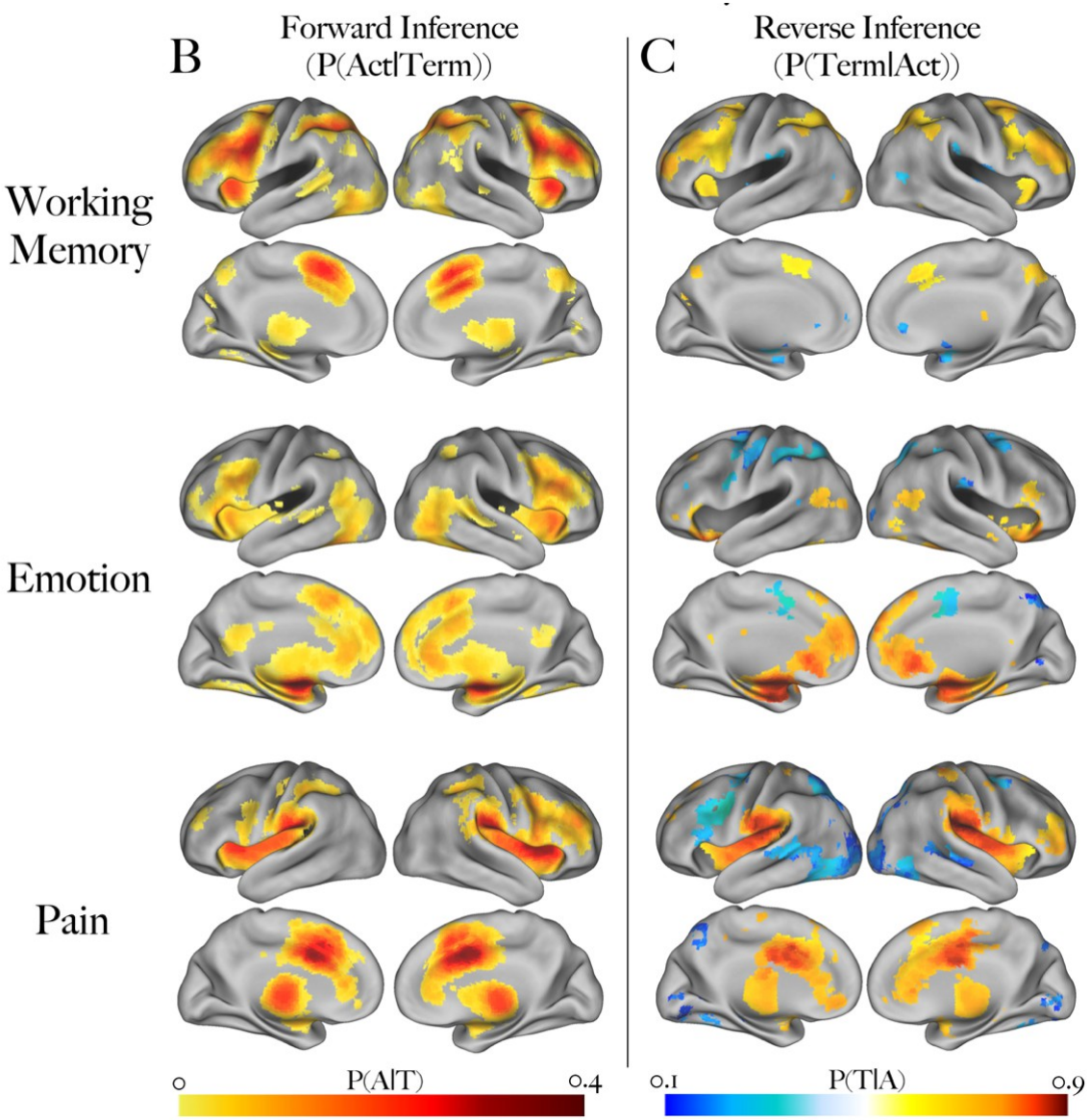
Neurosynth as infrastructure

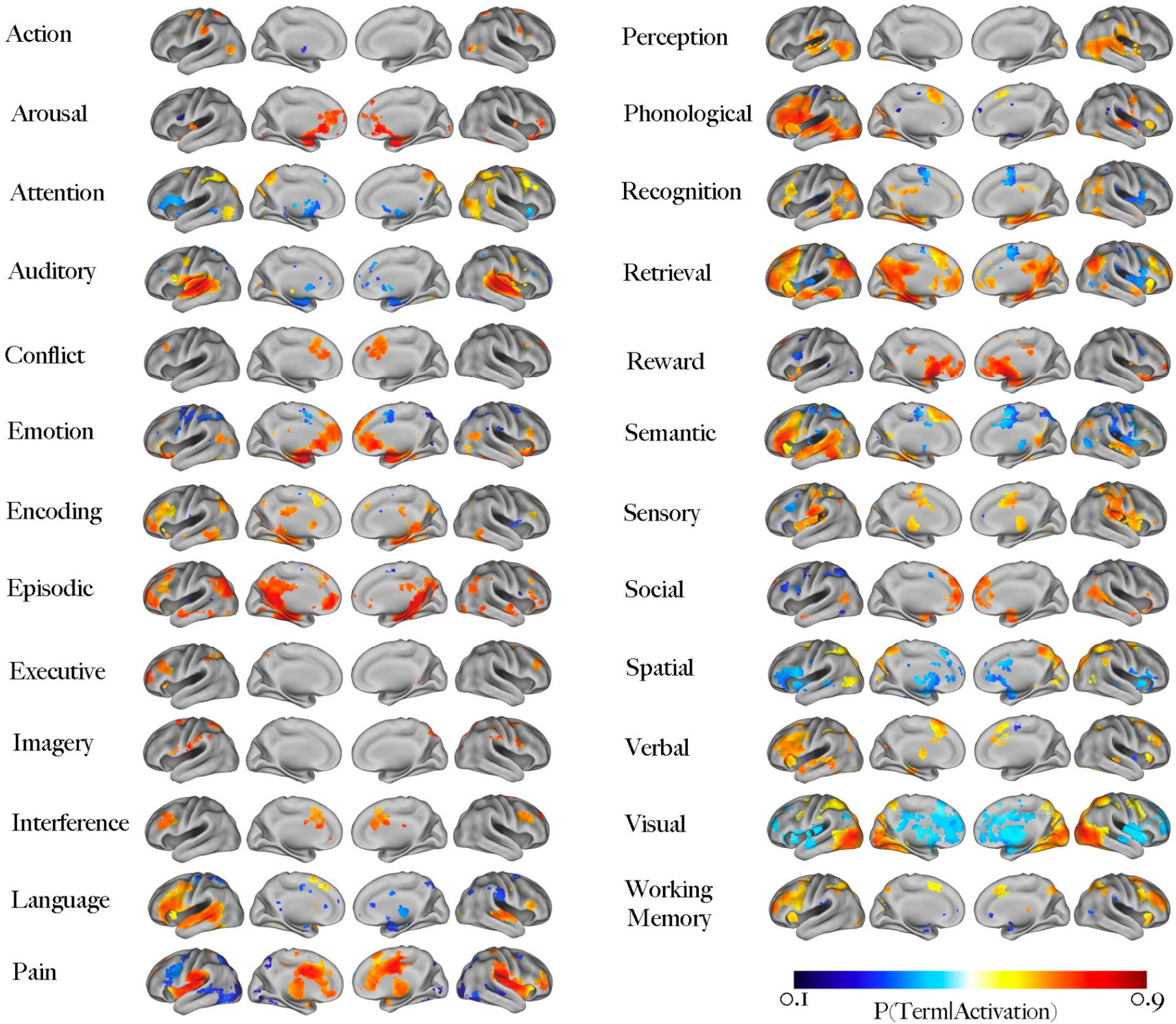
- >11,000 fMRI studies, 400,000 reported activations
- Represents almost any psychological state that can be indexed with words
- Automated and easy to scale
  - ~1,500 new studies per year
- Database is open
  - http://github.com/neurosynth/neurosynth-data
- Open-source software for manipulating data, performing common analyses
  - http://github.com/neurosynth/neurosynth
- Closely integrated with an interactive web interface
  - http://neurosynth.org

Limitations

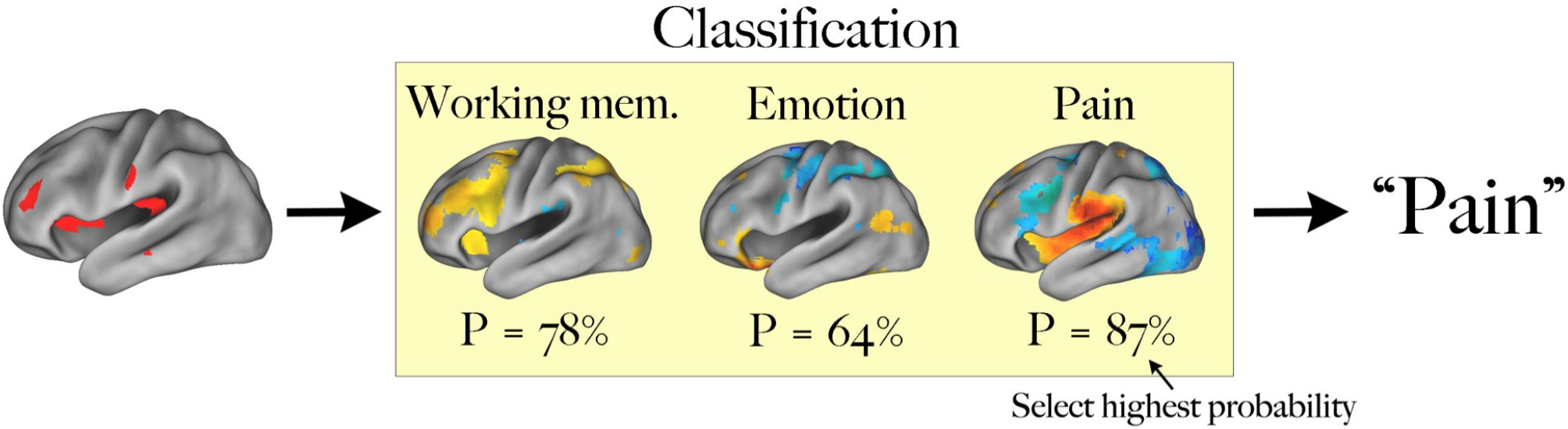
- Data quality issues
  - No concept of contrasts, deactivations, populations, sample size, statistics, etc.
- No ontology
  - Pure bag-of-words assumption
- Based on text rather than actual images
- Still… it works!

Large-scale quantitative reverse inference




Action
Arousal
Attention
Auditory
Conflict
Emotion
Encoding
Episodic
Executive
Imagery
Interference
Language
Pain
Perception
Phonological
Recognition
Retrieval
Reward
Semantic
Sensory
Social
Spatial
Verbal
Visual
Working Memory
0.1
0.9
P(Term|Activation)

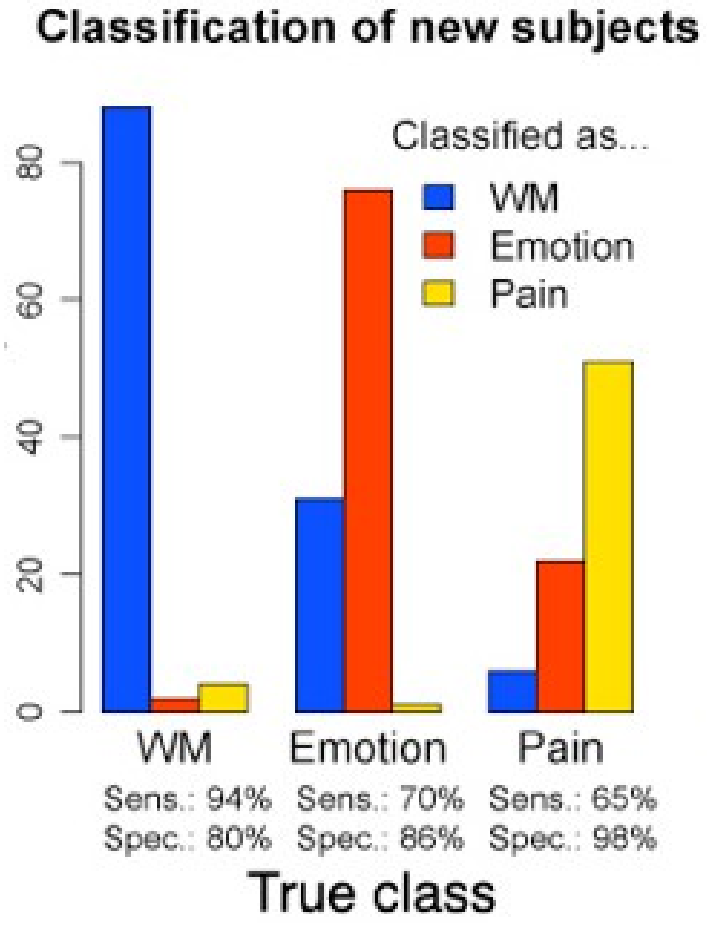
Classification of cognitive states



Yarkoni et al (2011)

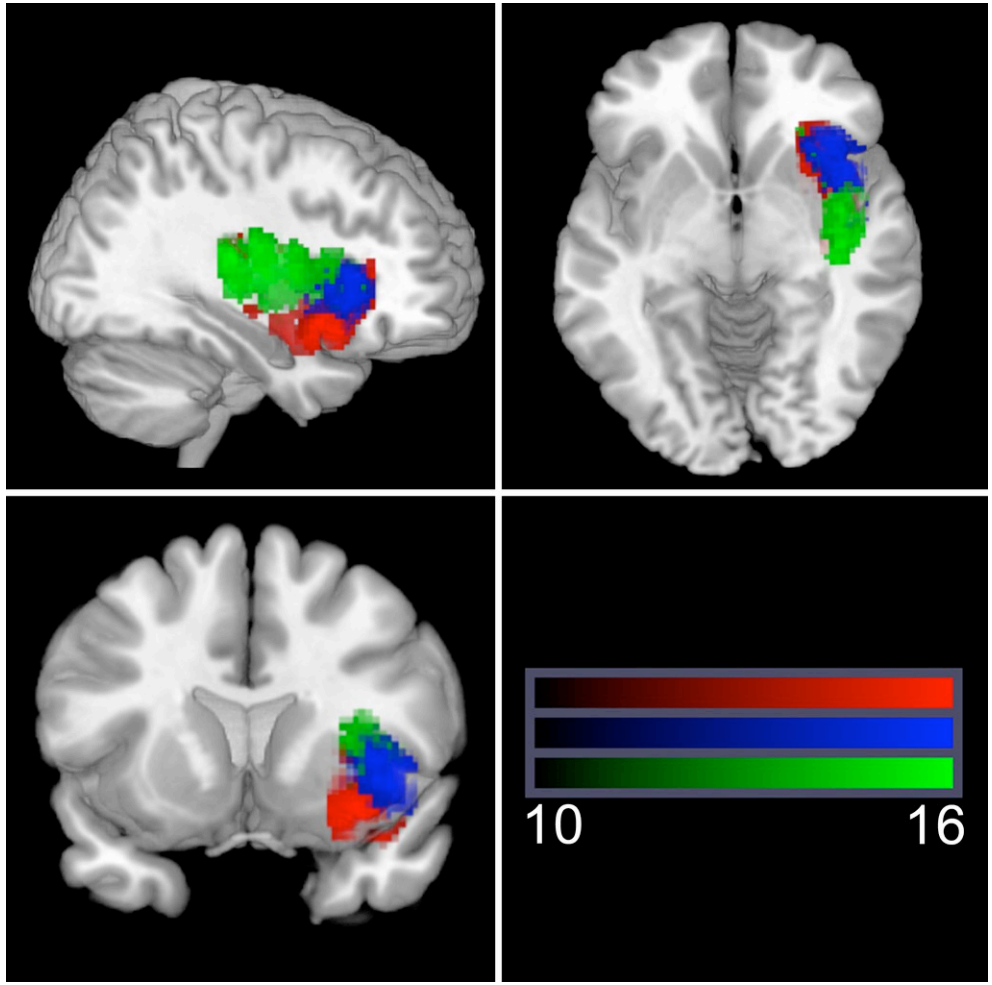
Decoding mental states in single subjects

- Can we identify cognitive states in individual subjects?
- Test ability to classify working memory, emotion, and pain
- Remember, this is all based on text…



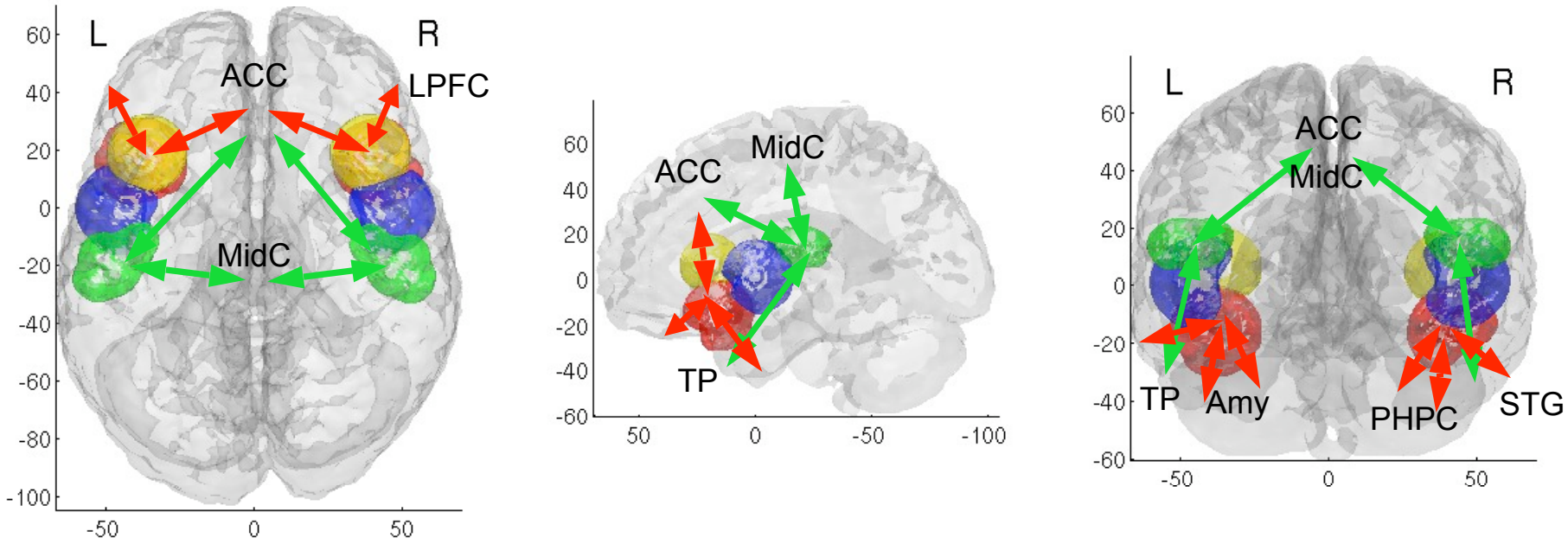
Yarkoni et al (2011)

Decoding insular function



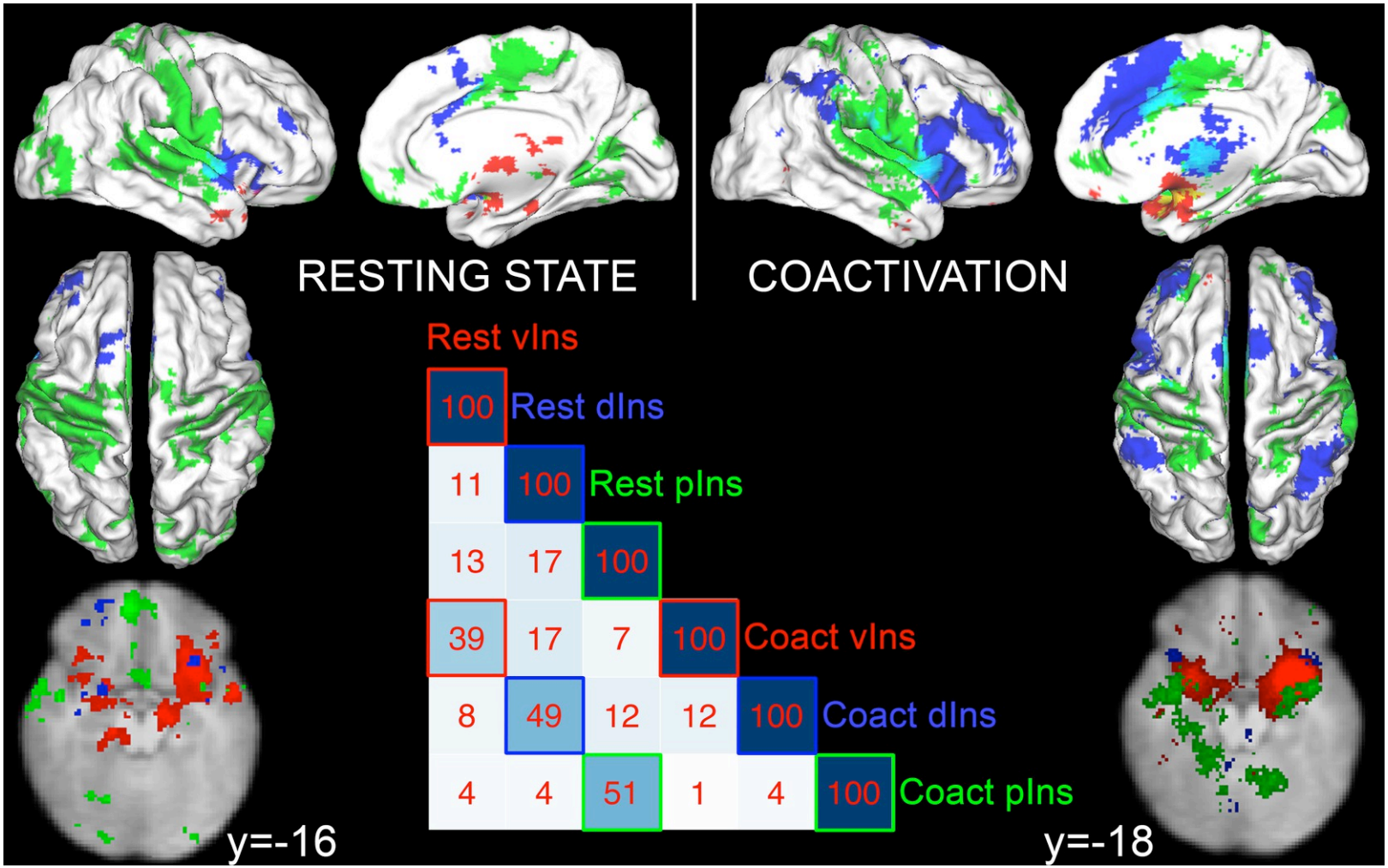
Chang, Yarkoni, Khaw & Sanfey (2013)

Figure 1. Anatomical subdivision of insular cortex



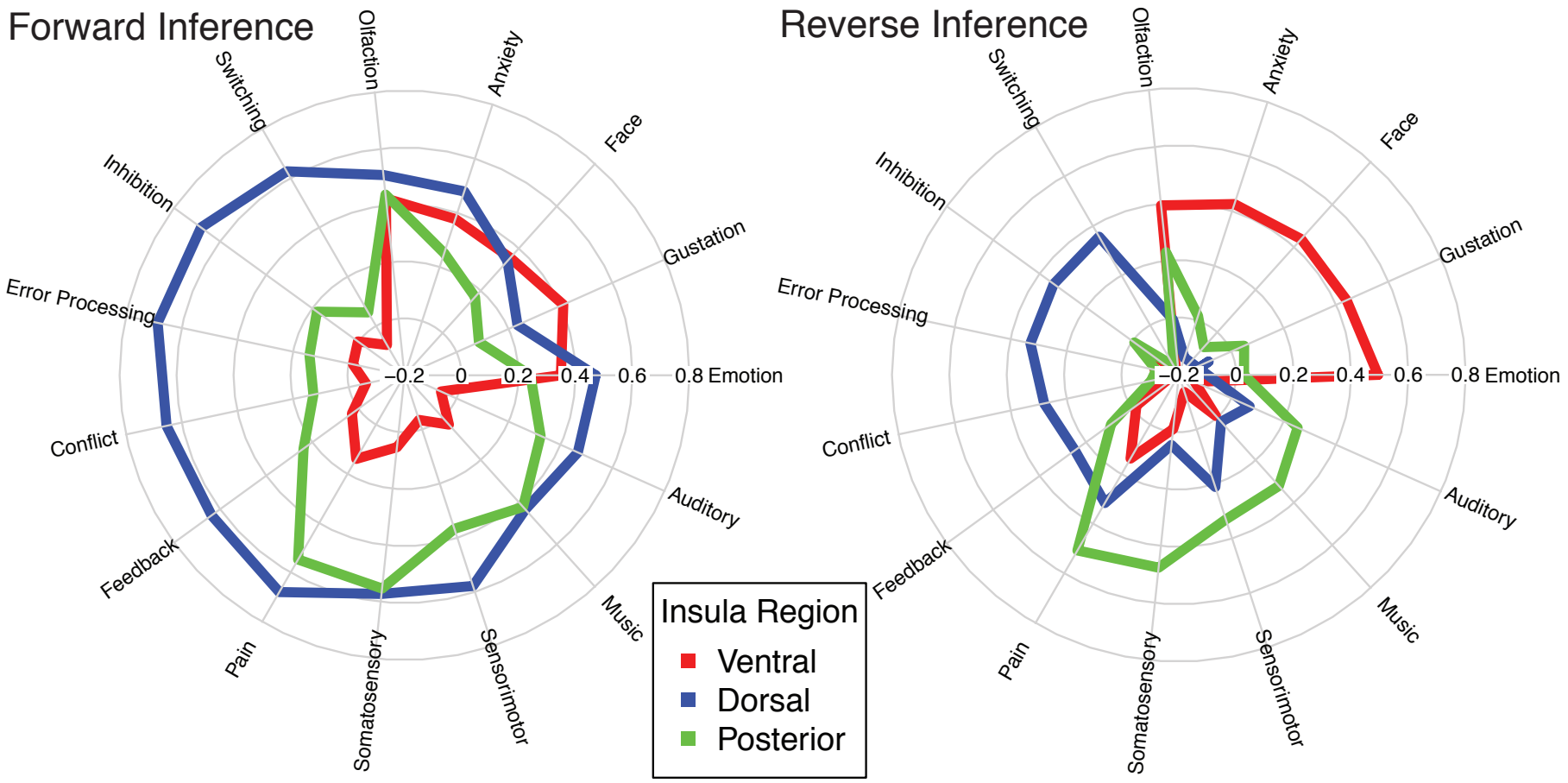
Wager & Feldman-Barrett (2004)

Insula parcellation



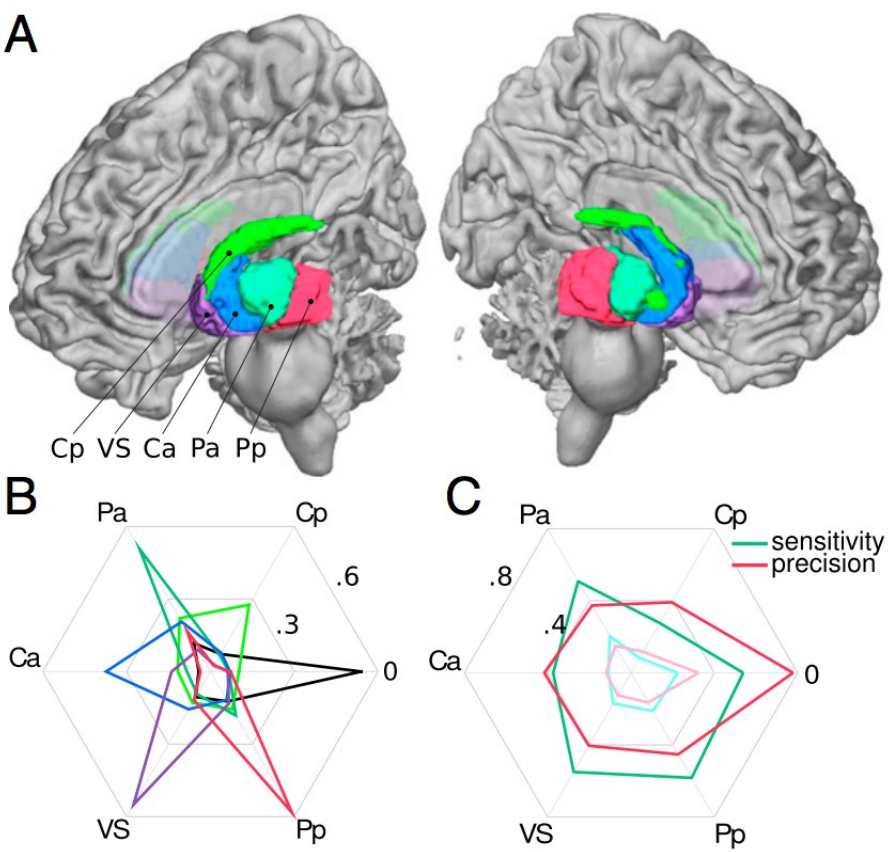
Chang, Yarkoni, Khaw & Sanfey (2013)

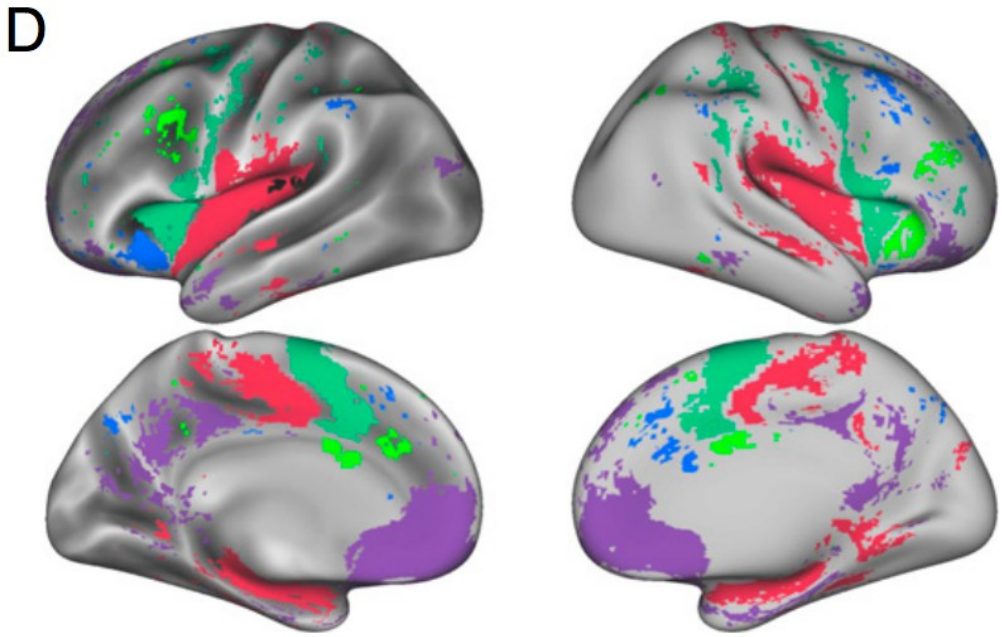
A tale of two inferences



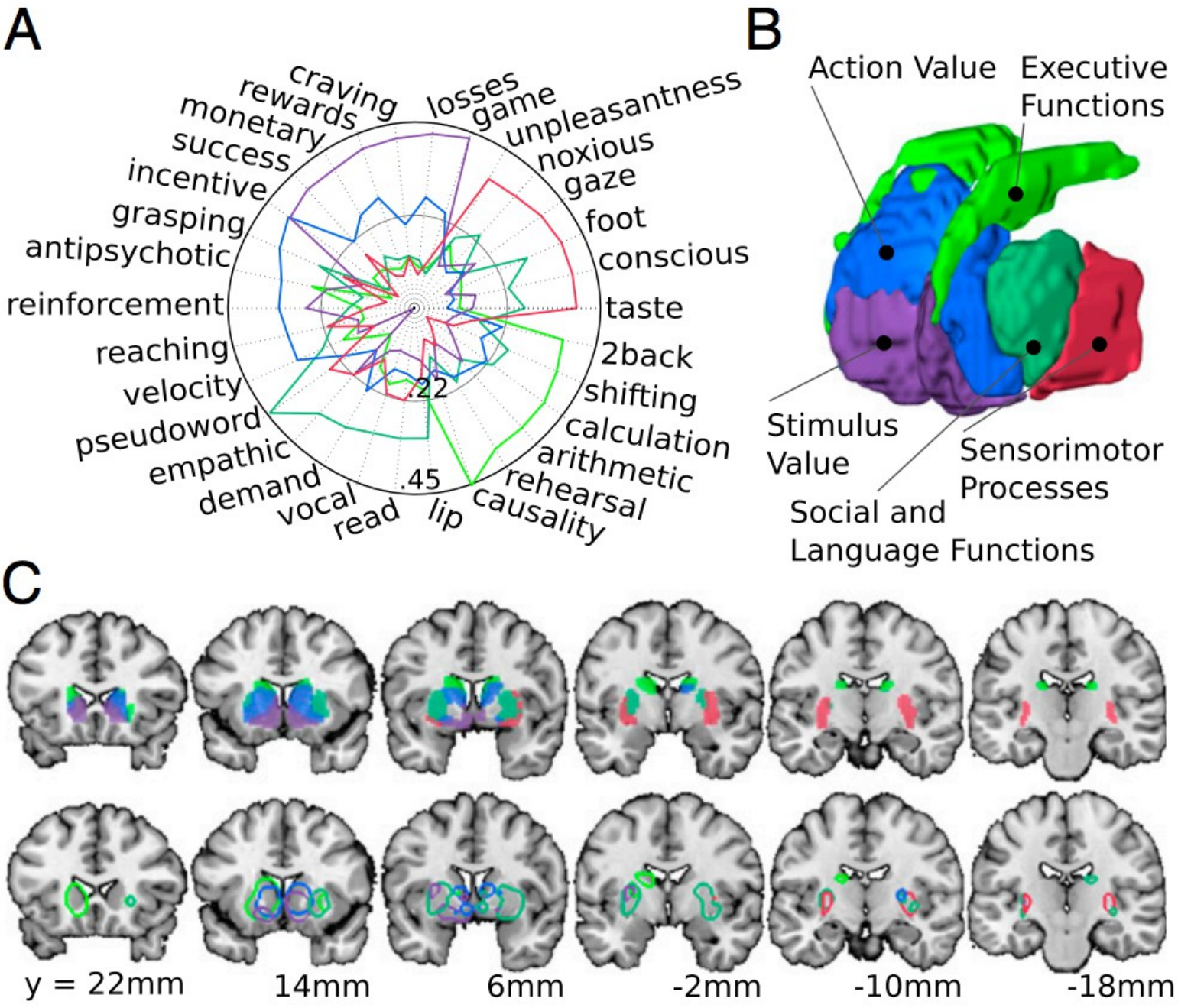
Chang, Yarkoni, Khaw & Sanfey (2013)

Parcellating the human striatum





Pauli, O'Reilly, Yarkoni, & Wager (2016)



Pauli, O'Reilly, Yarkoni, & Wager (2016)

Parcellating the medial frontal cortex

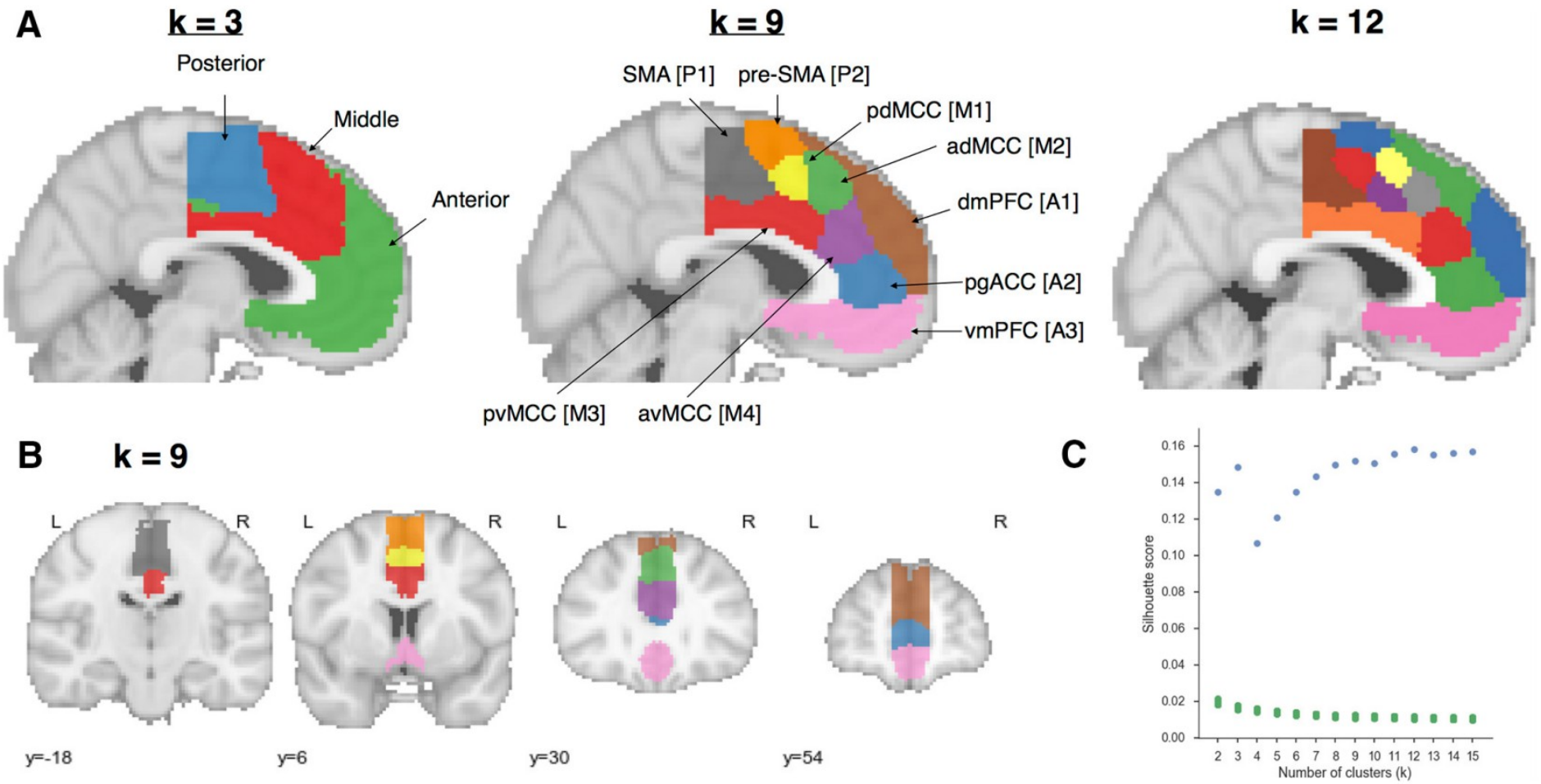
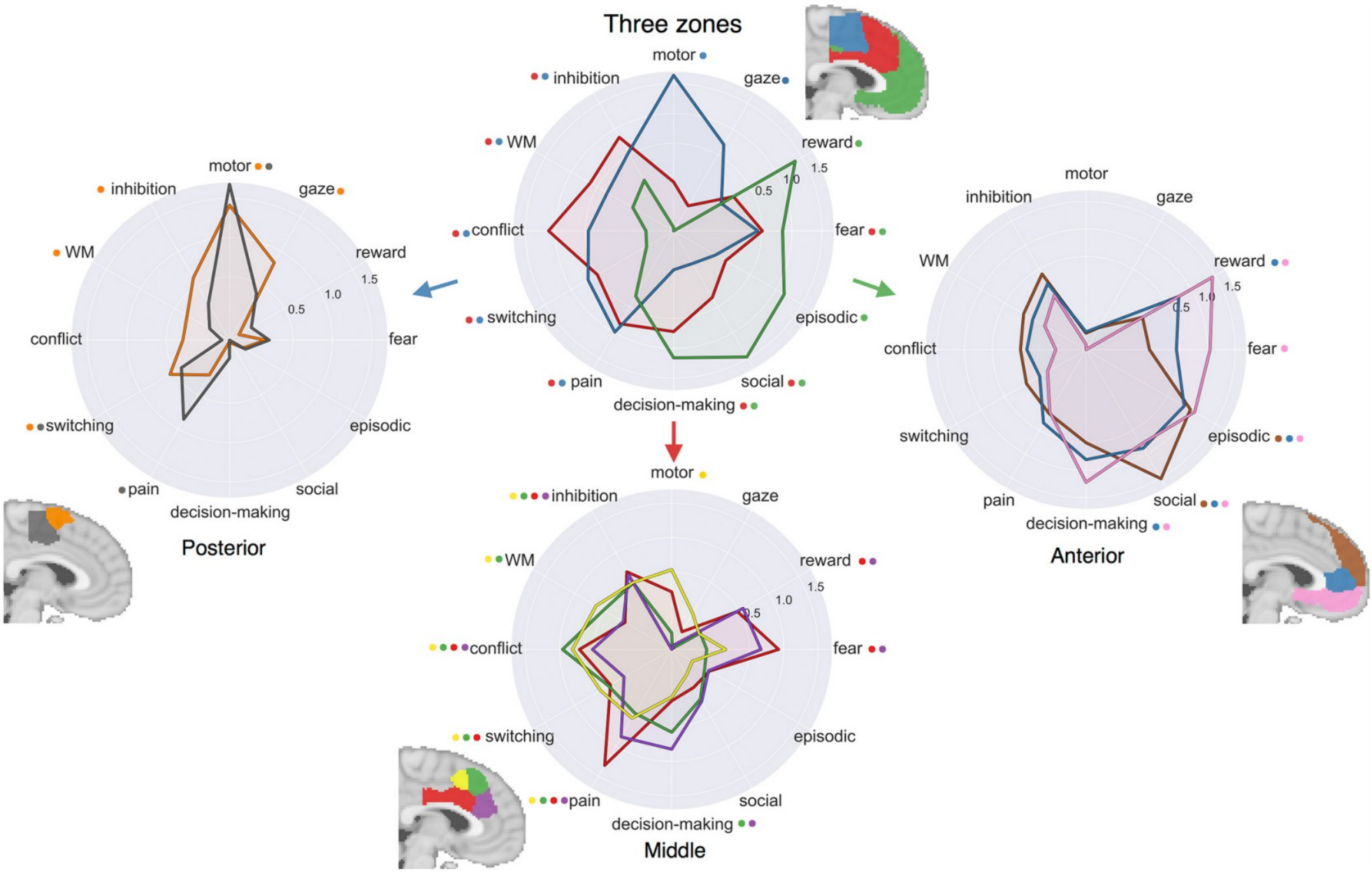


Figure 2. Coactivation-based clustering of MFC results. ***A***, Mid-sagittal view at three levels at granularity: three broad zones, nine and 12 subregions. Clusters in nine subregion solution are given both descriptive and alphanumeric names for reference. ***B***, Axial view of nine subregions. ***C***, Silhouette scores of real (green) and permuted (blue) clustering solutions. Clustering was performed on permuted data 1000 times for each *k* to compute a null distribution (*p* values for all clusters < 0.001). Silhouette scores reached local maxima at 3 regions and plateaued after 9.

De La Vega, Chang, Banich, Wager, & Yarkoni (2016)

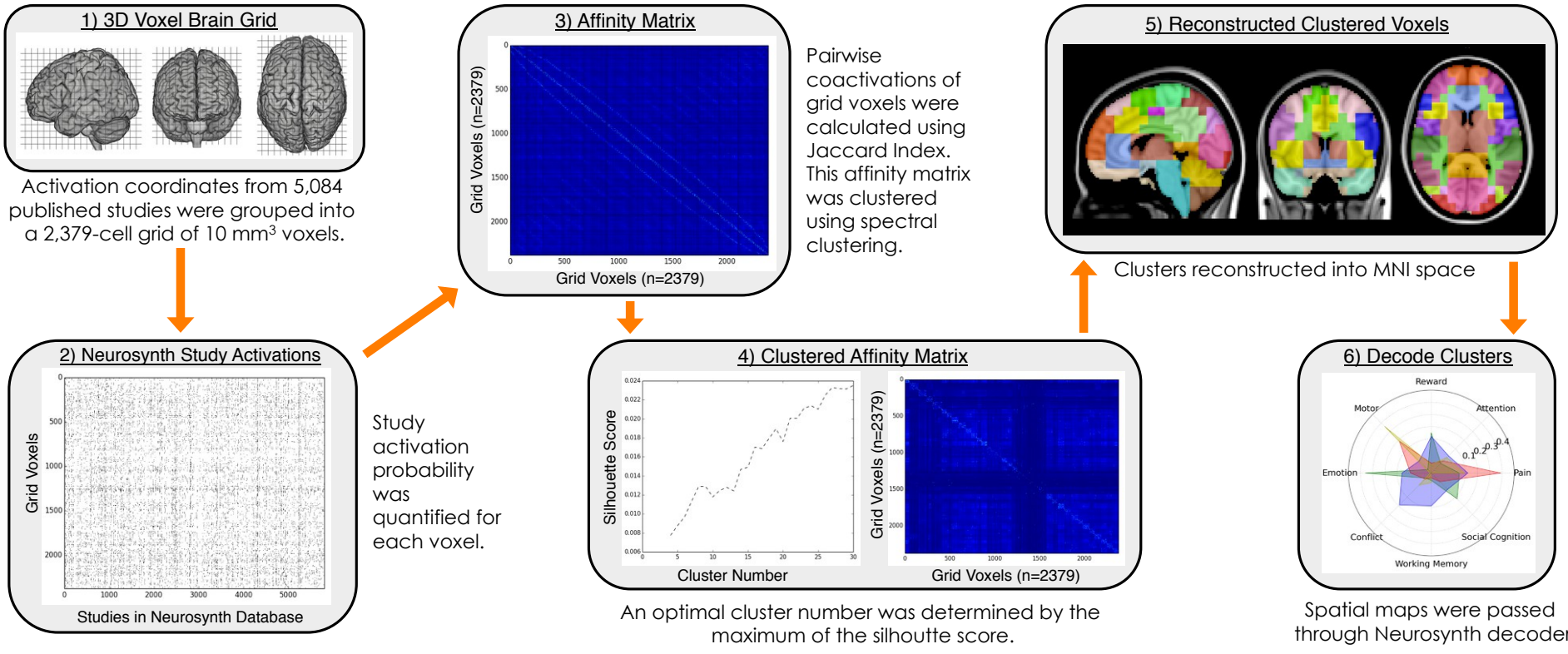


De La Vega, Chang, Banich, Wager, & Yarkoni (2016)

And so on and so forth…

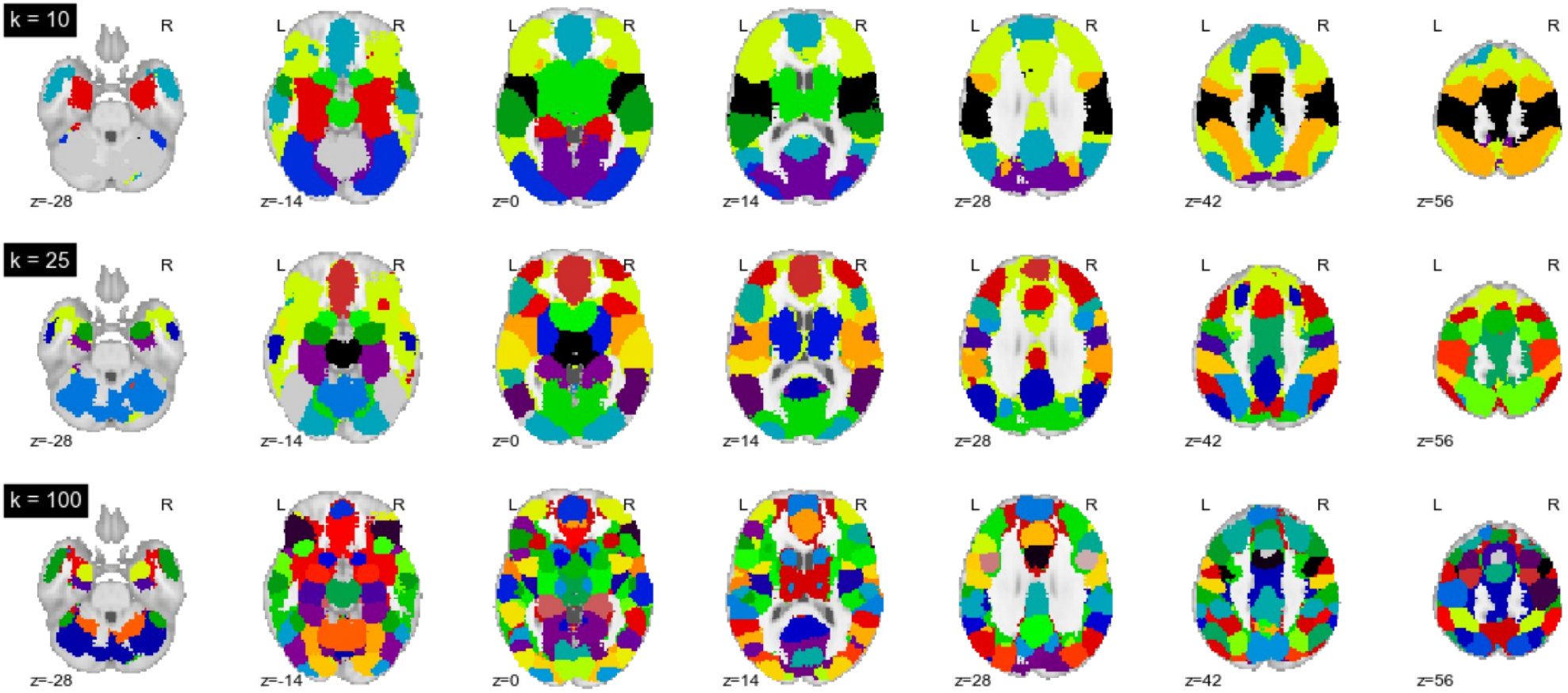
- You can do this for every part of the brain!
- But don't bother—they've all already been done!
  - Bzdok et al. (2013; TPJ), Cieslick et al. (2012; DLPFC); Clos et al. (2013; IFG); Eickhoff et al. (2014; DMPFC); Bzdok et al. (2013; amygdala); Zald et al. (2014; OFC); Riedel et al. (2015; cerebellum); etc. etc…
- Processing steps are fairly consistent
  - We should probably just hurry up and automate this

Automated parcellation and decoding



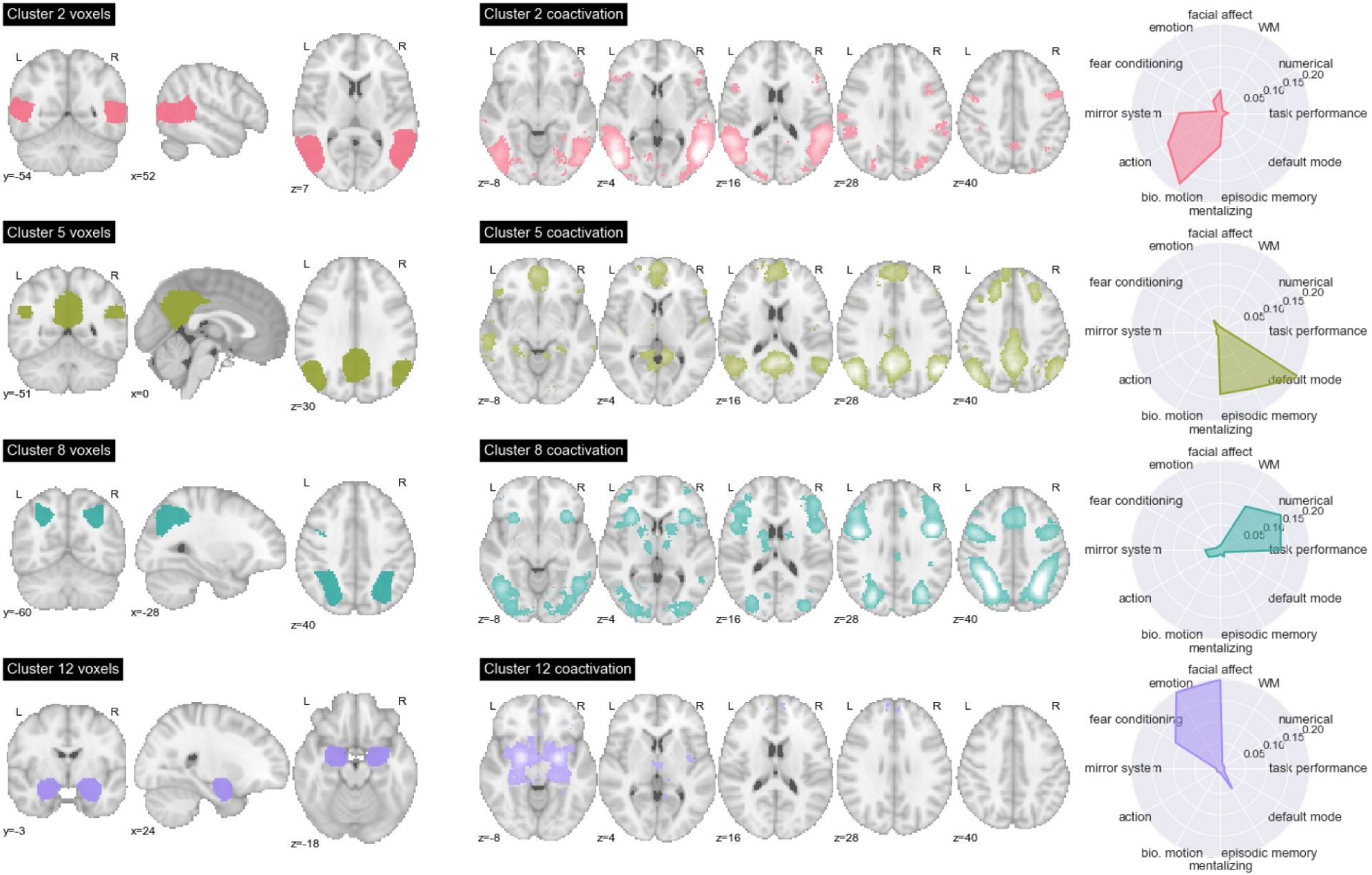
Yarkoni, De La Vega, & Chang (in preparation)

Whole-brain clustering of all Neurosynth data



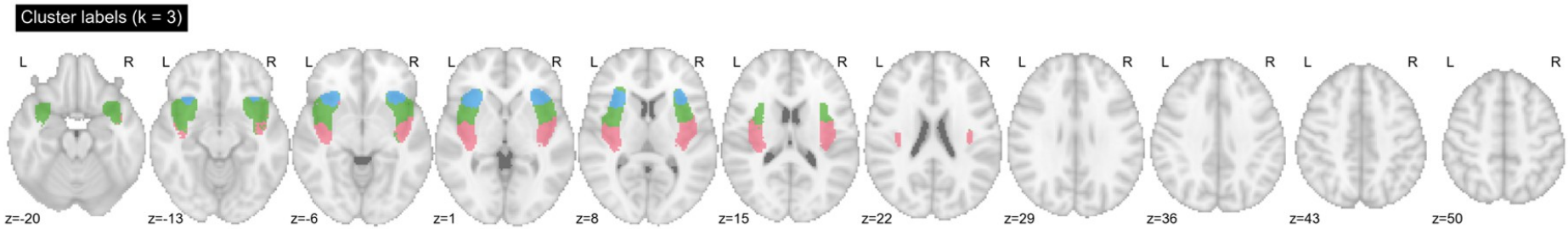
Yarkoni, De La Vega, & Chang (in prep.)

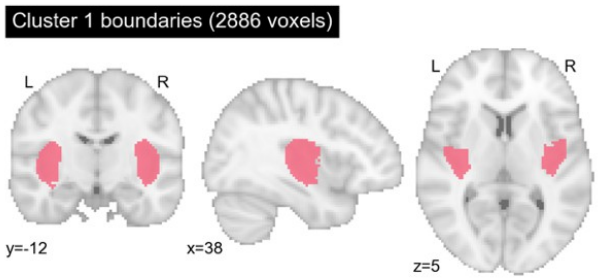
Selected clusters (k = 25)

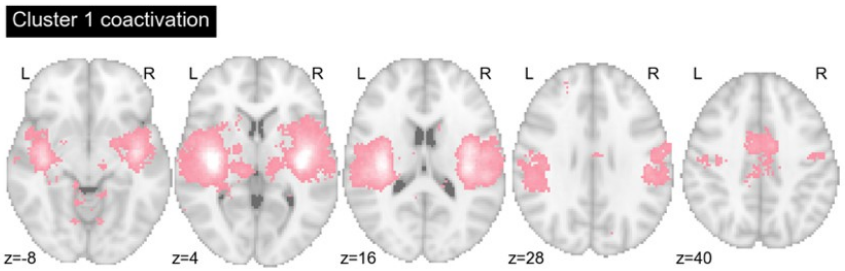


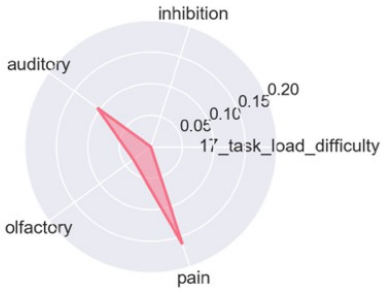
Yarkoni, De La Vega, & Chang (in prep.)

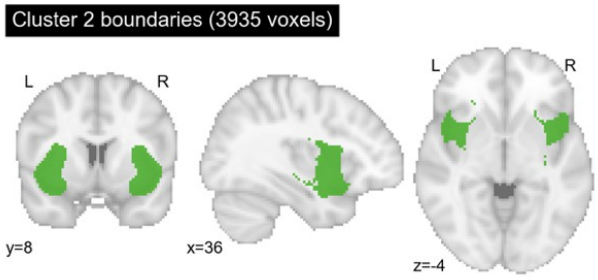
Automated clustering and decoding of the insula (k = 3)

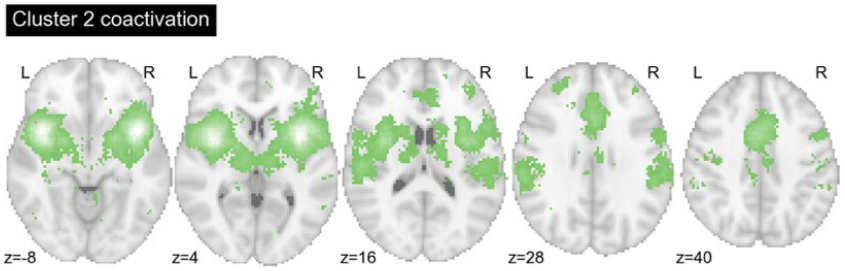


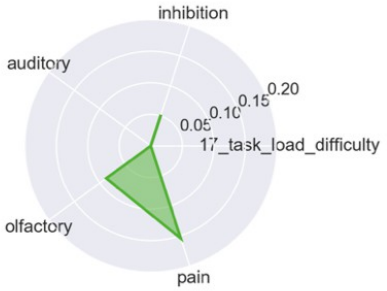


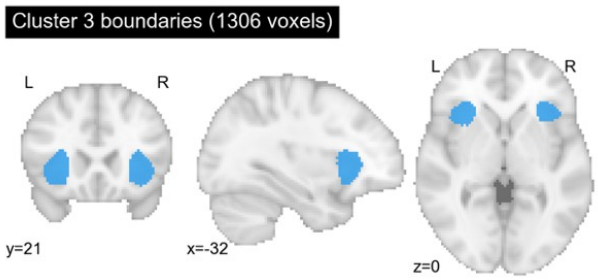


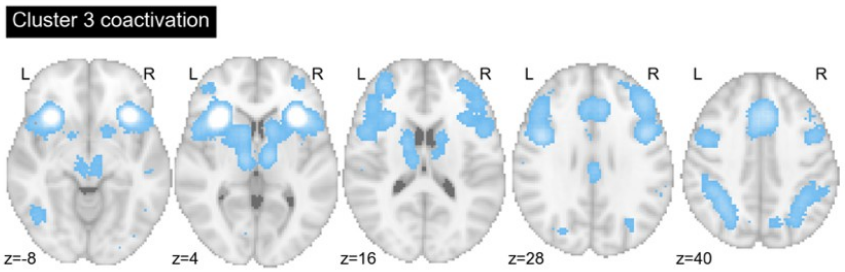


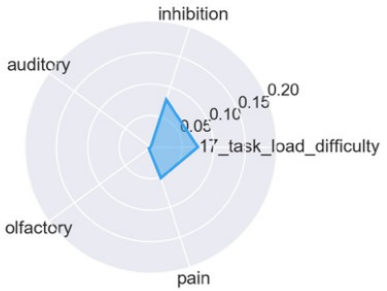




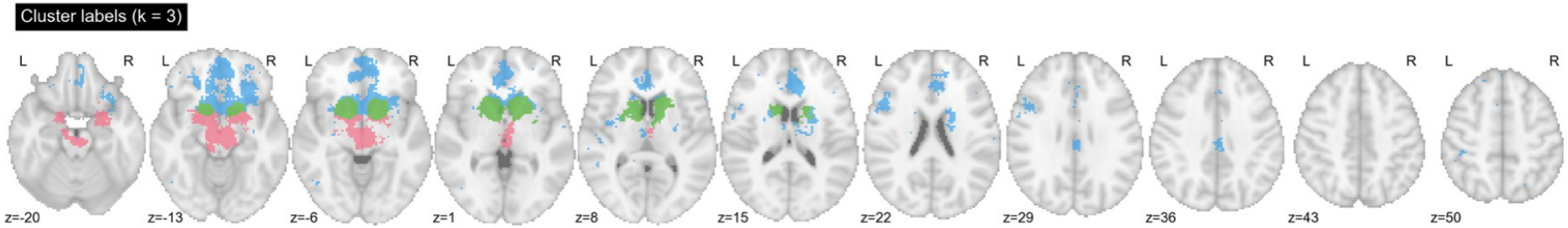


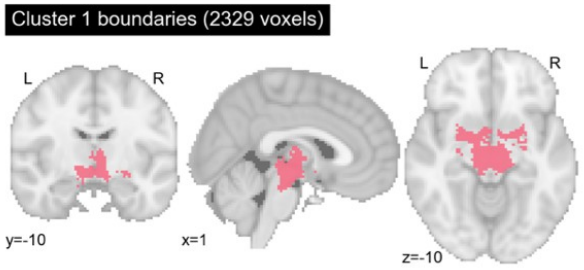


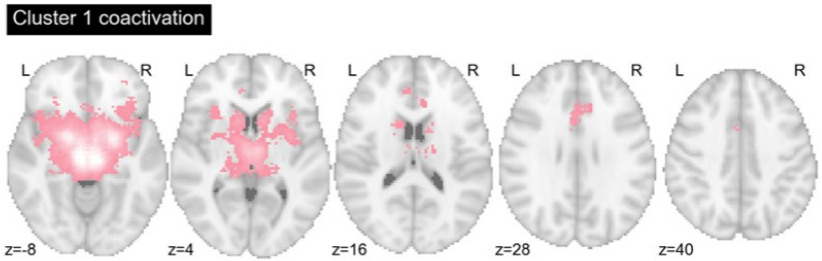


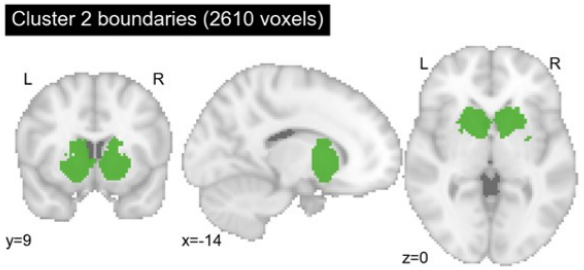


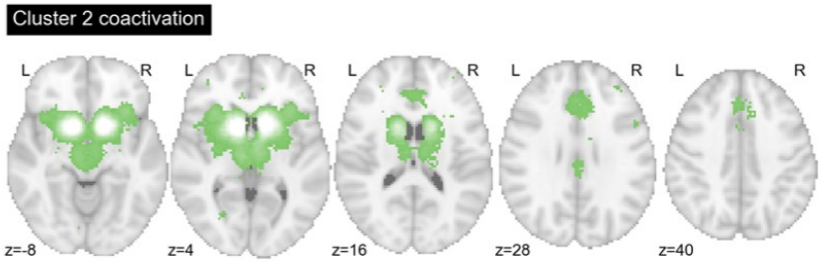
Automated clustering and decoding of Neurosynth “reward” map (k = 3)

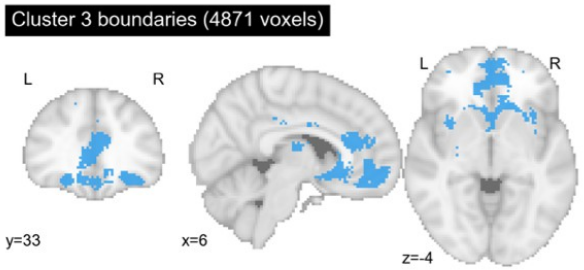


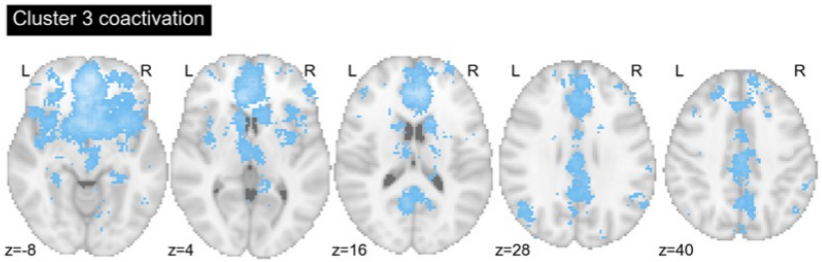




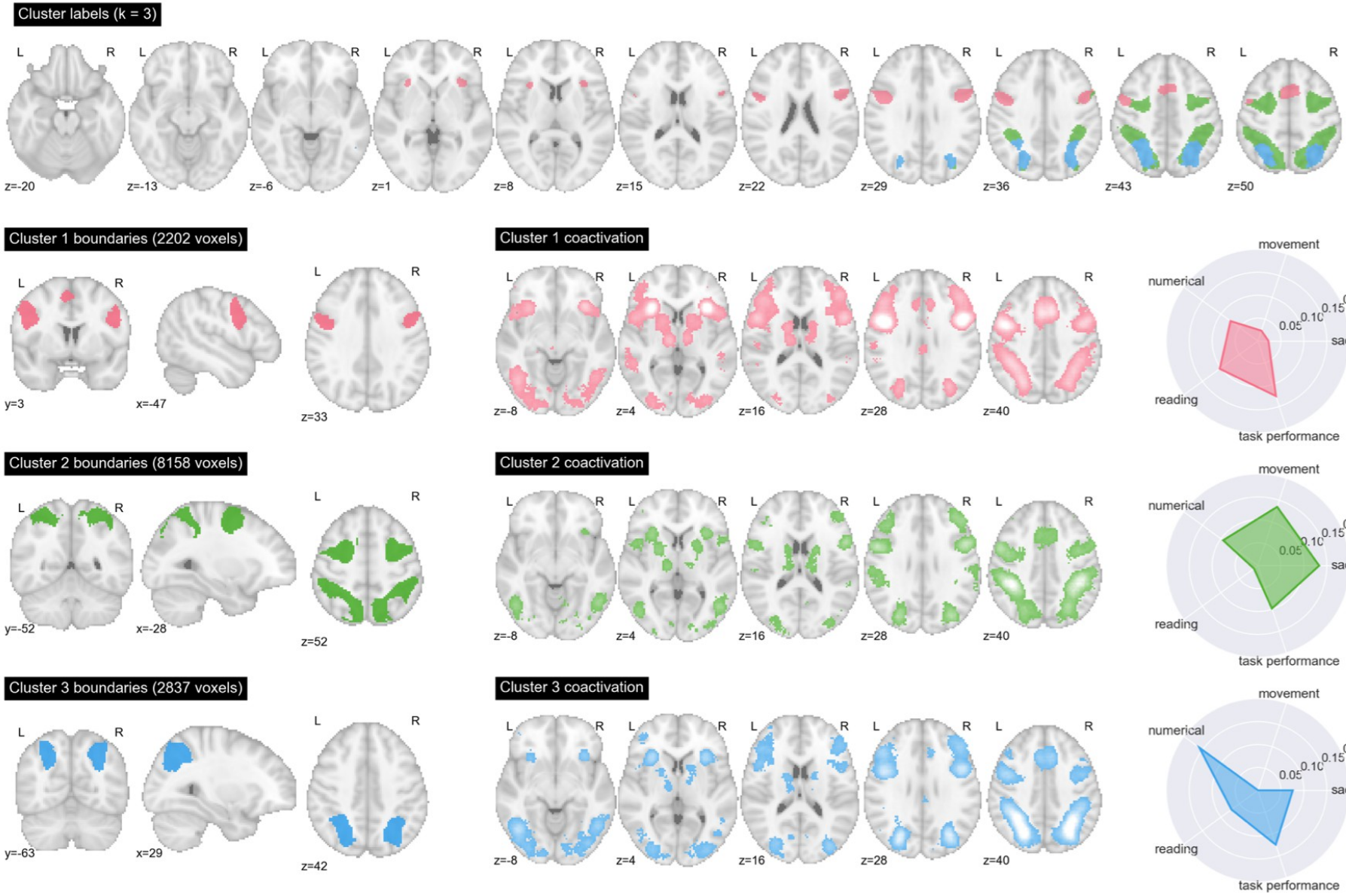








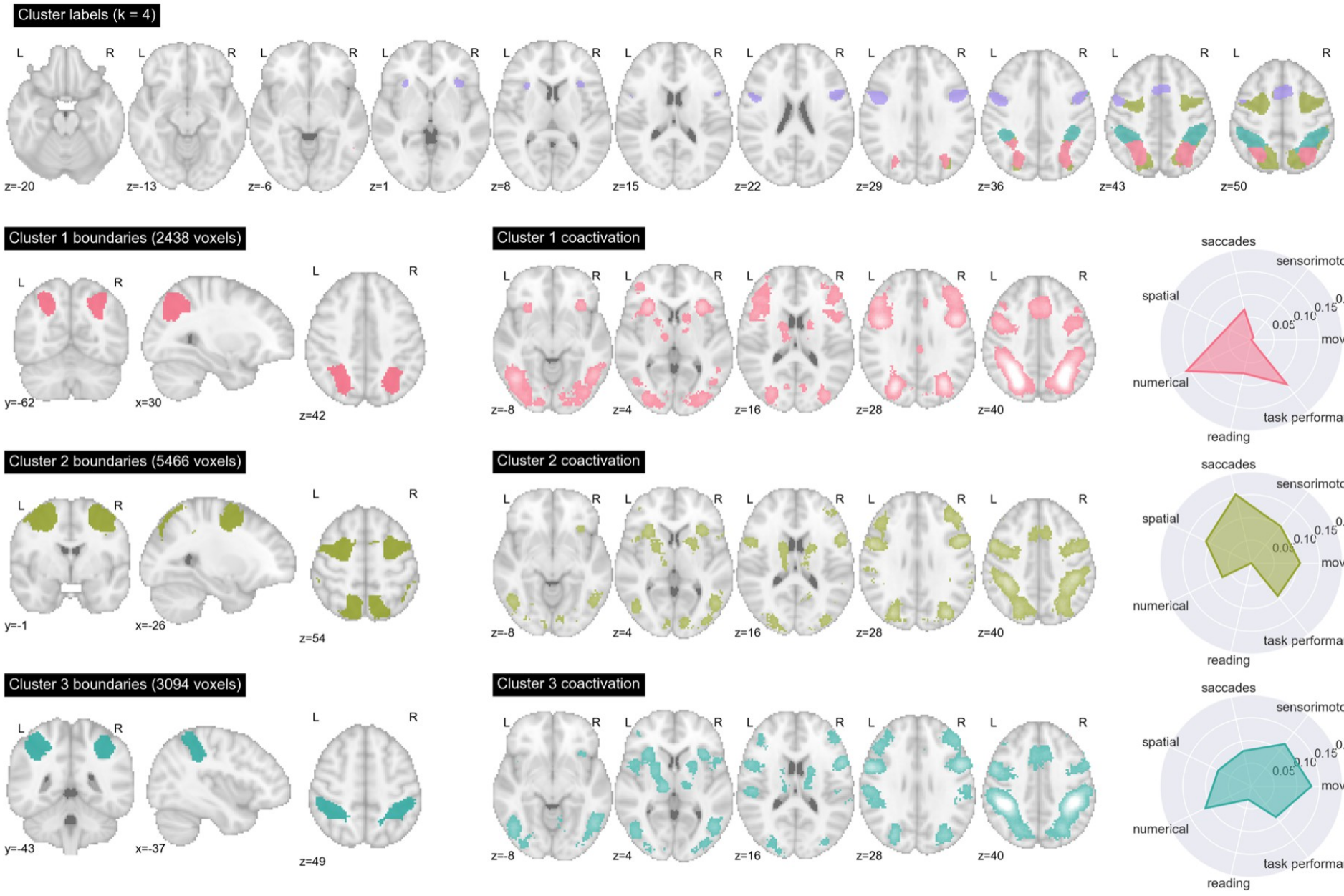
Automated clustering and decoding of dorsal frontoparietal network (k = 3)



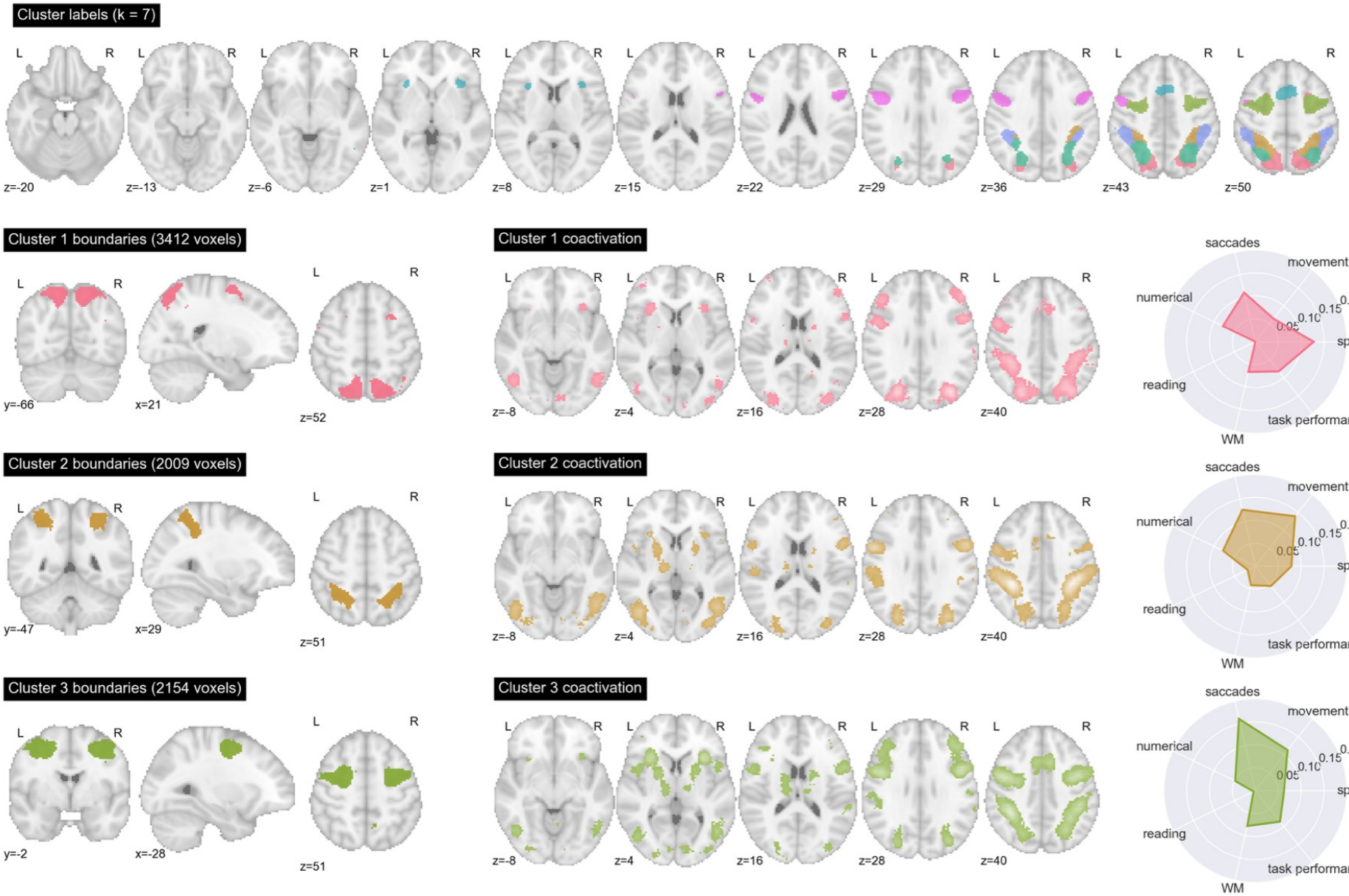
Wait, why k = 3?

- No reason

Automated clustering and decoding of dorsal frontoparietal network (k = 4)



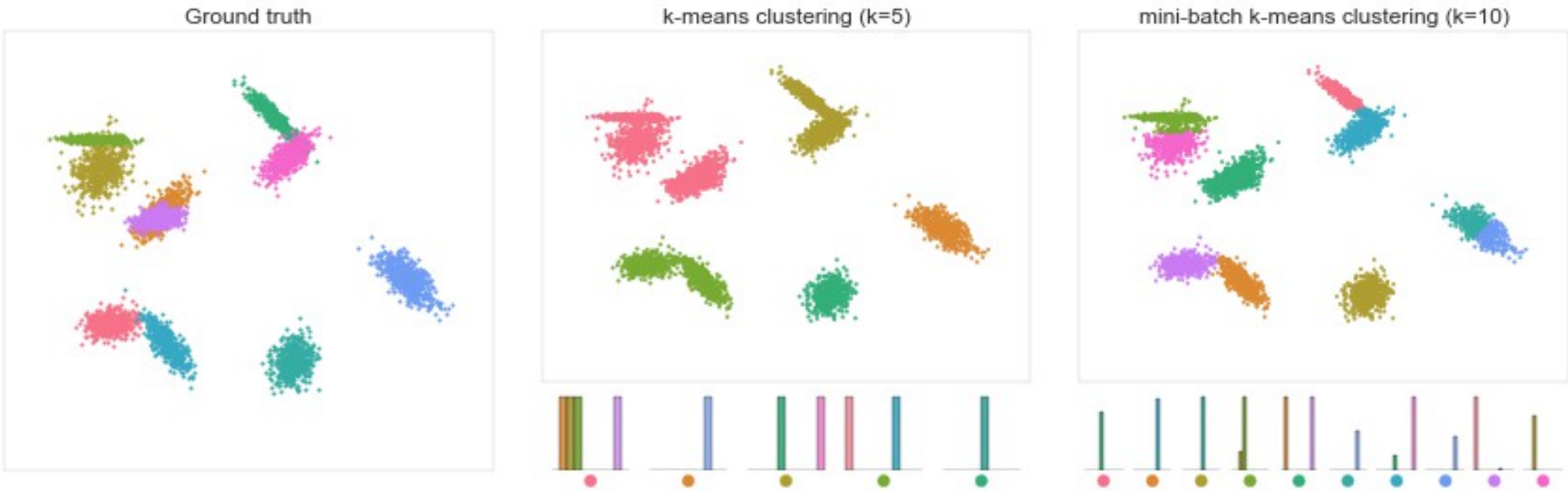
Automated clustering and decoding of dorsal frontoparietal network (k = 7)



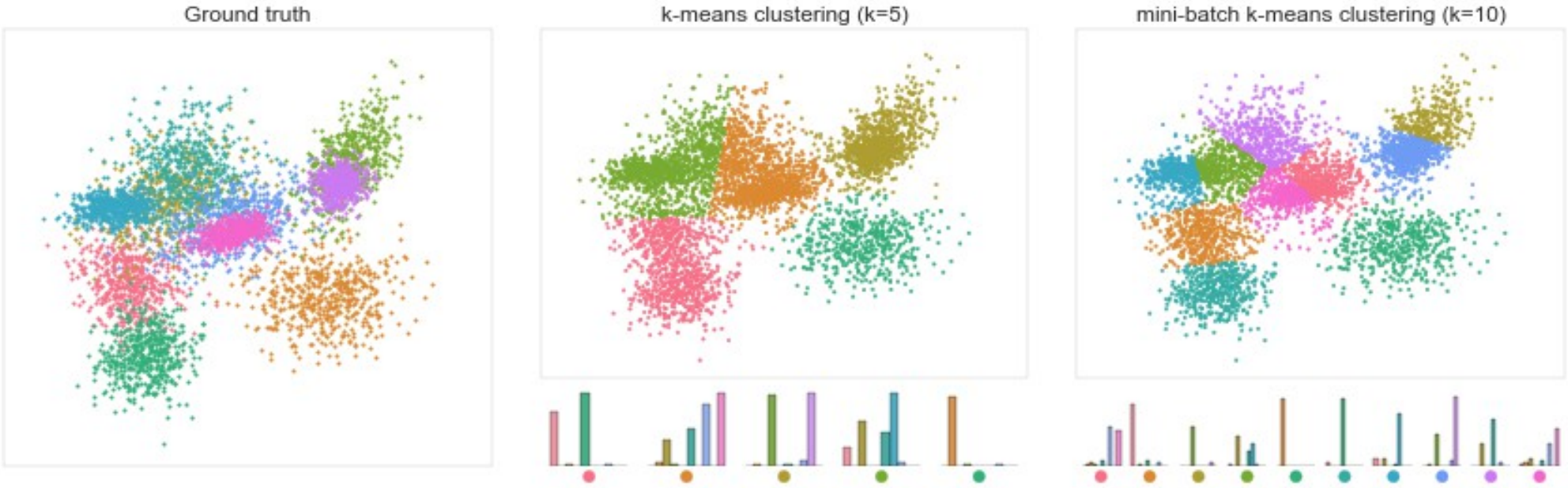
What have we learned?

- In general, we seem to get “nice” answers out of Neurosynth no matter what/how we cluster
- Clustering is a useful technique for *describing* one’s data in a low-dimensional way
  - Can help us discern interesting distinctions and associations
- But it’s very unlikely to recapture the “true” data-generating process
  - Hard assignment of each voxel to a single cluster is biologically implausible
  - The actual structure of human cognition is likely to be extremely high-dimensional, doesn’t necessarily sit on a low-D manifold
- Which of the following situations are we in?

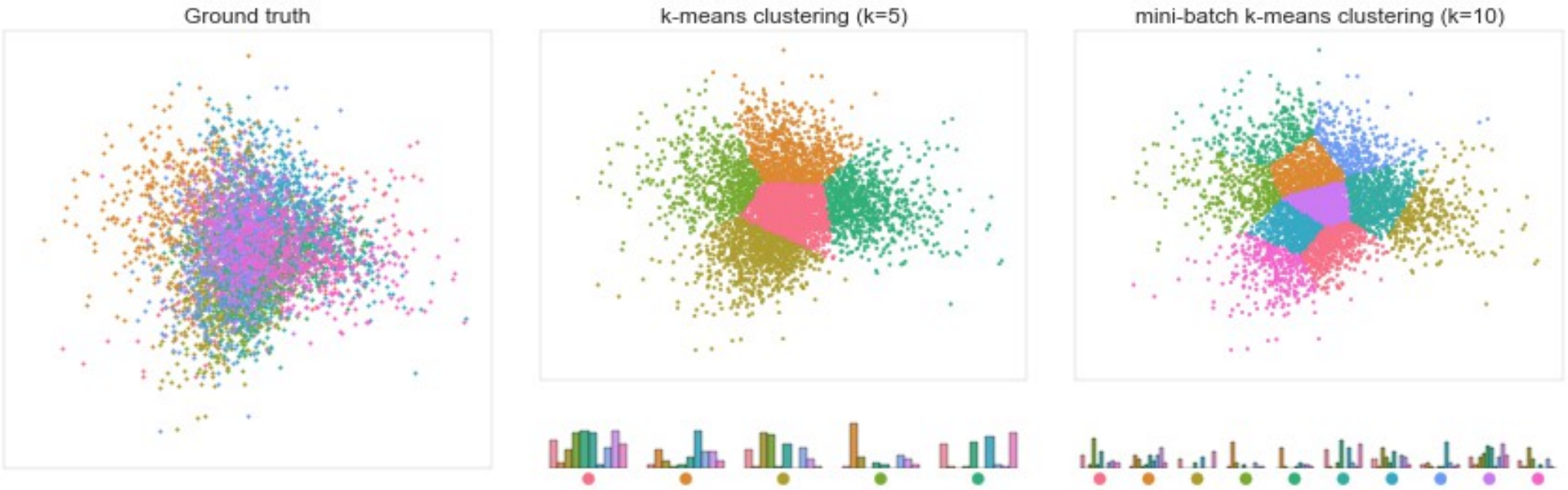
Good

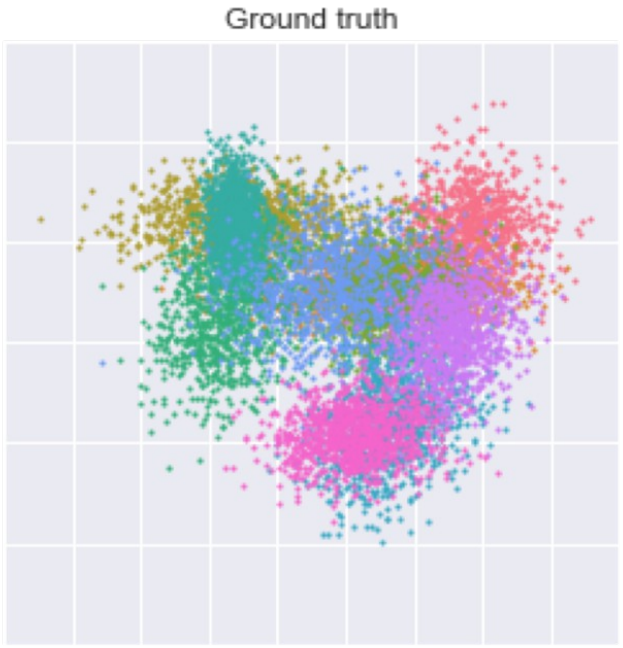

Ground truth
k-means clustering (k=5)
mini-batch k-means clustering (k=10)

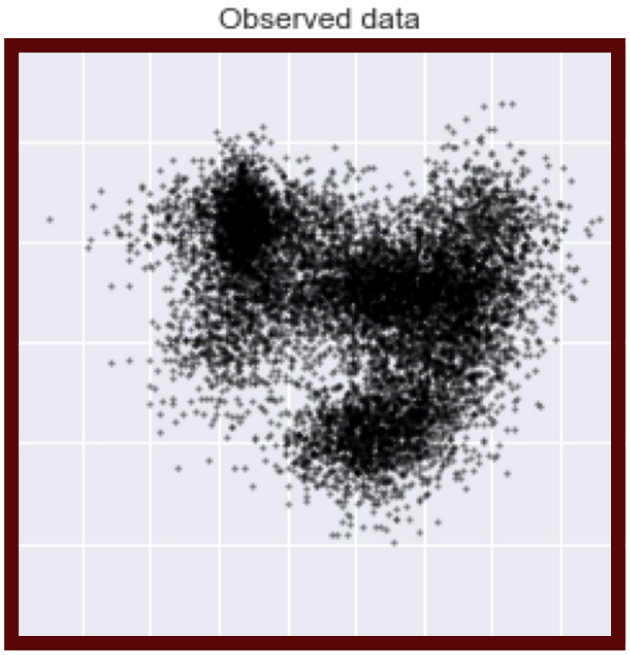
Bad


Ground truth
k-means clustering (k=5)
mini-batch k-means clustering (k=10)

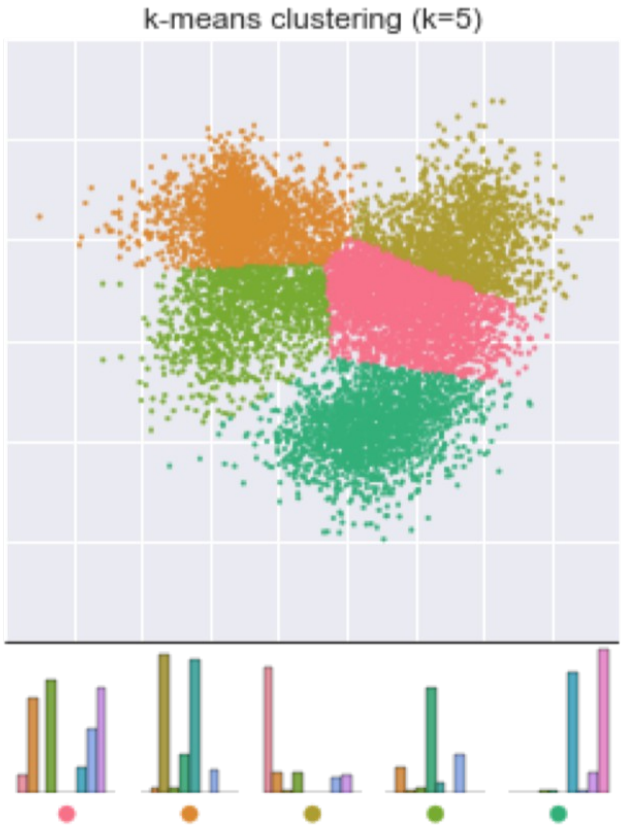
Ugly

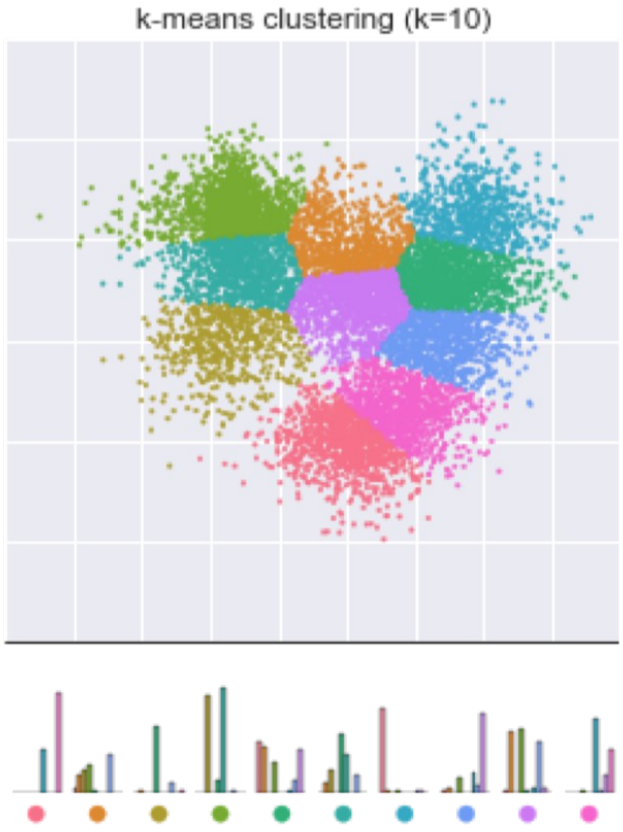

Ground truth
k-means clustering (k=5)
mini-batch k-means clustering (k=10)





This is what reality probably looks like most of the time with fMRI

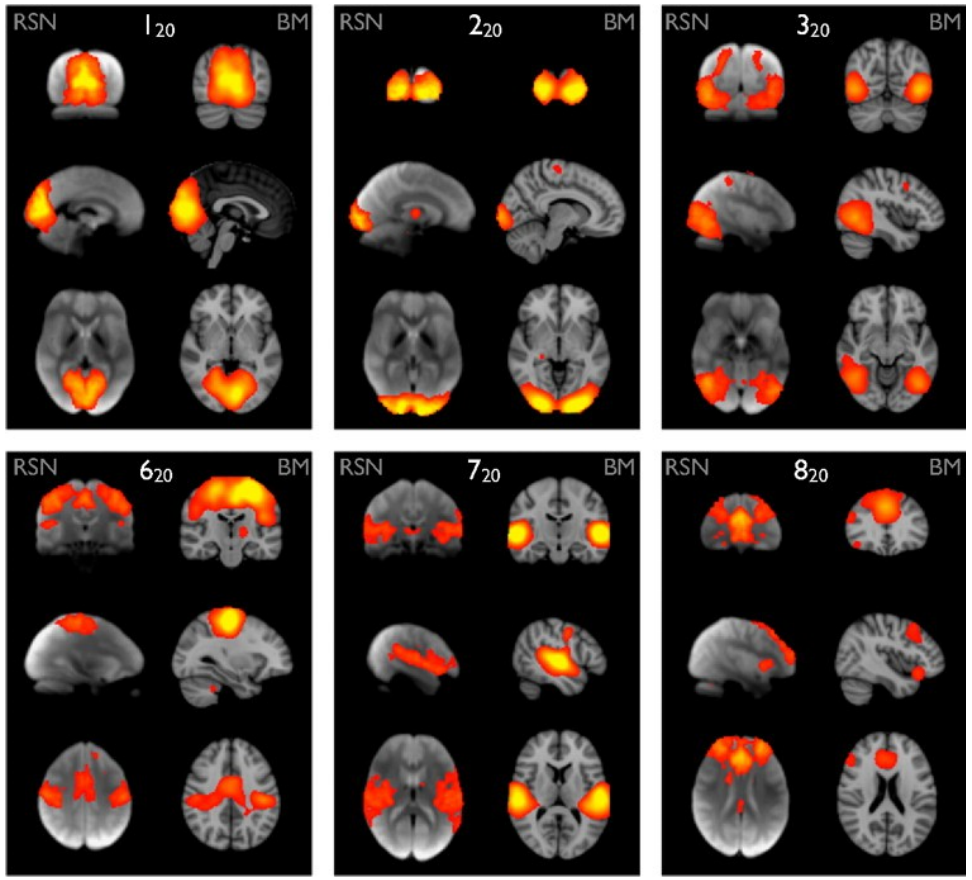


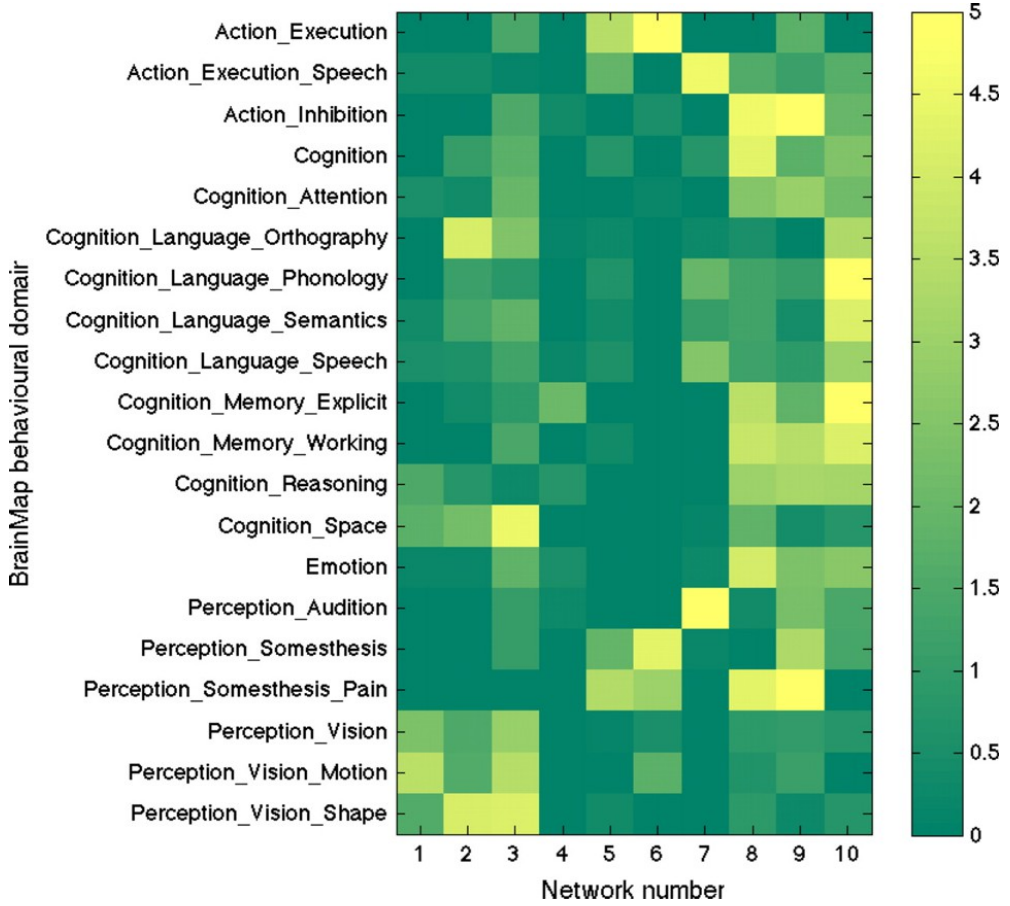


Alternative approaches

- Clustering approaches have very low plausibility as models of data-generating processes
  - This doesn't mean we shouldn’t use them—just be aware of their limitations
- There are many other decomposition approaches
  - PCA, ICA, dictionary learning, LDA, NMF, etc…
  - Each has advantages and disadvantages

ICA applied to resting state and BrainMap data

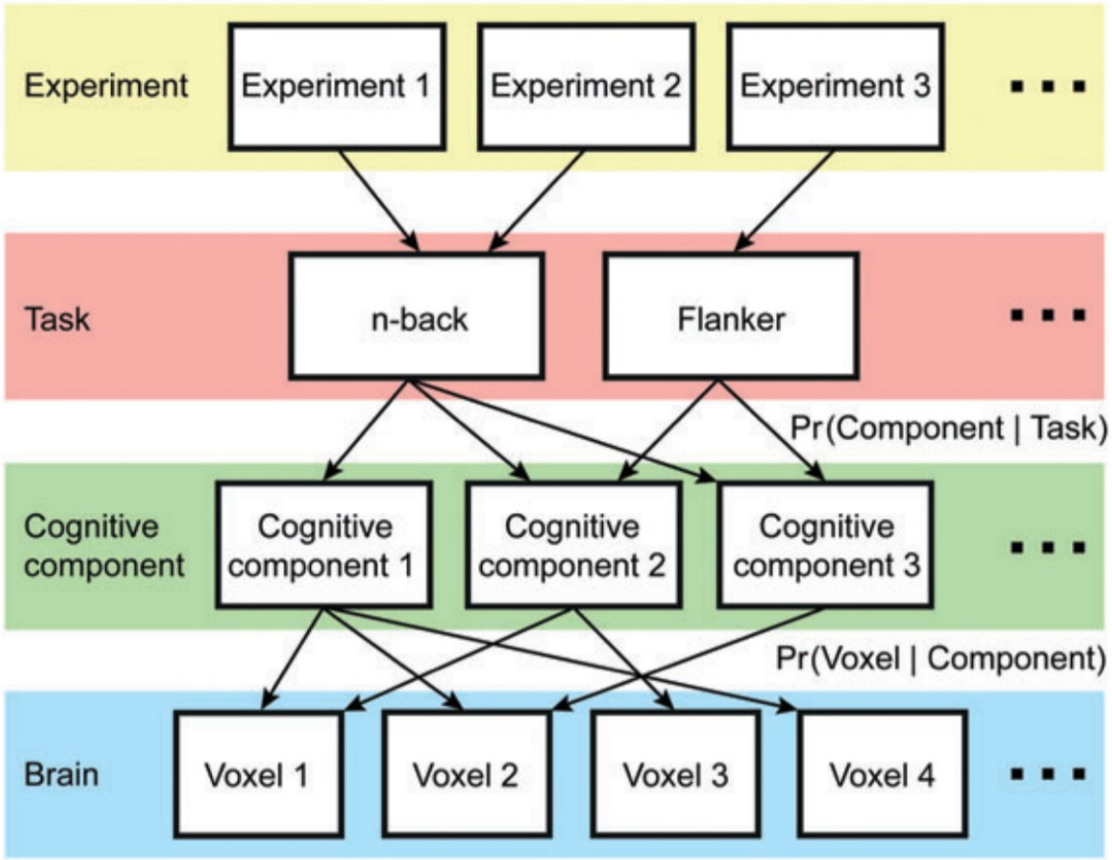




Smith et al. (2009)

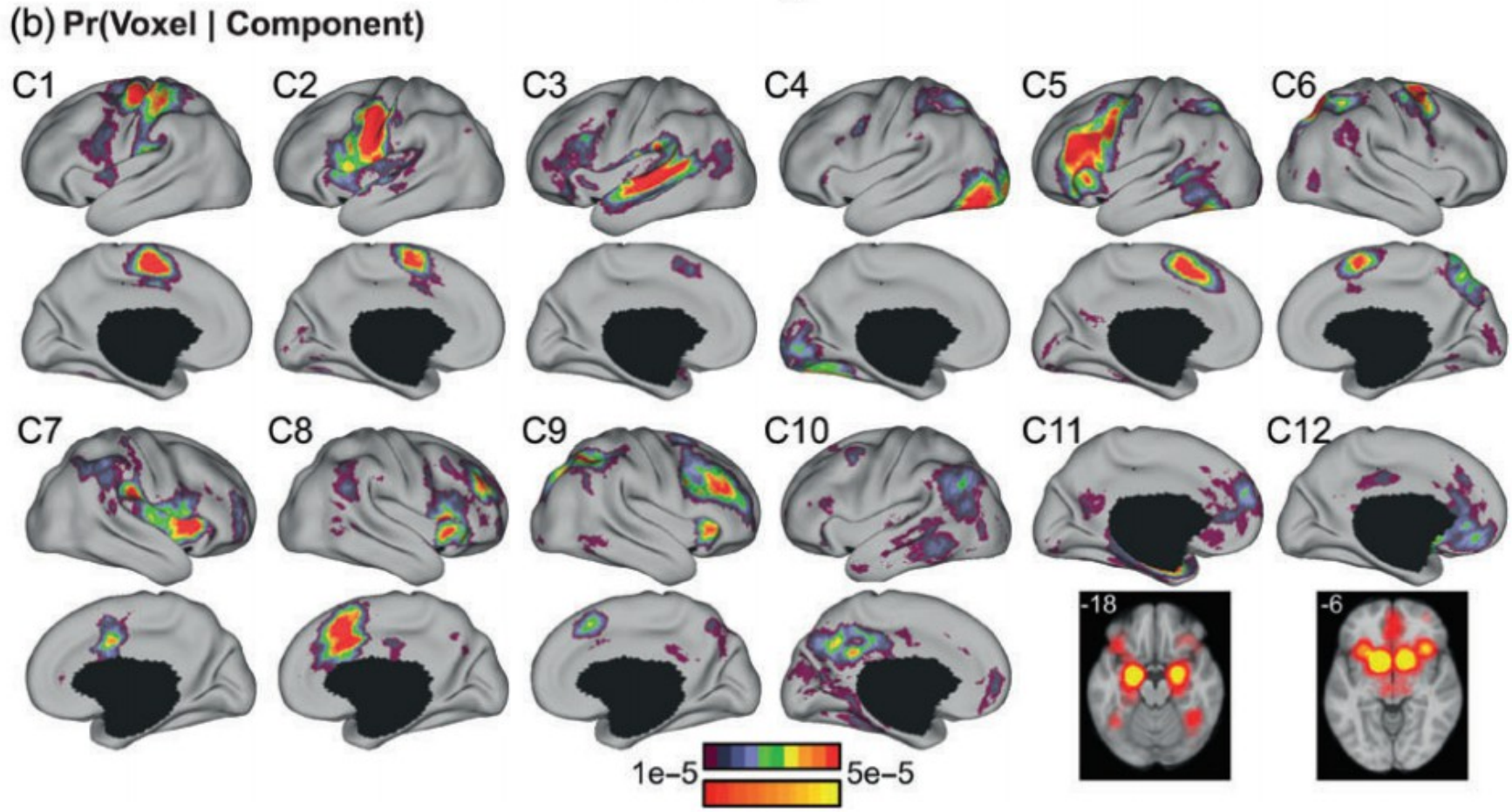
Adding structure

- Most approaches decompose brain activity directly
- Mapping to cognition/ task space is unclear
- Can we explicitly constrain our models to respect that structure?

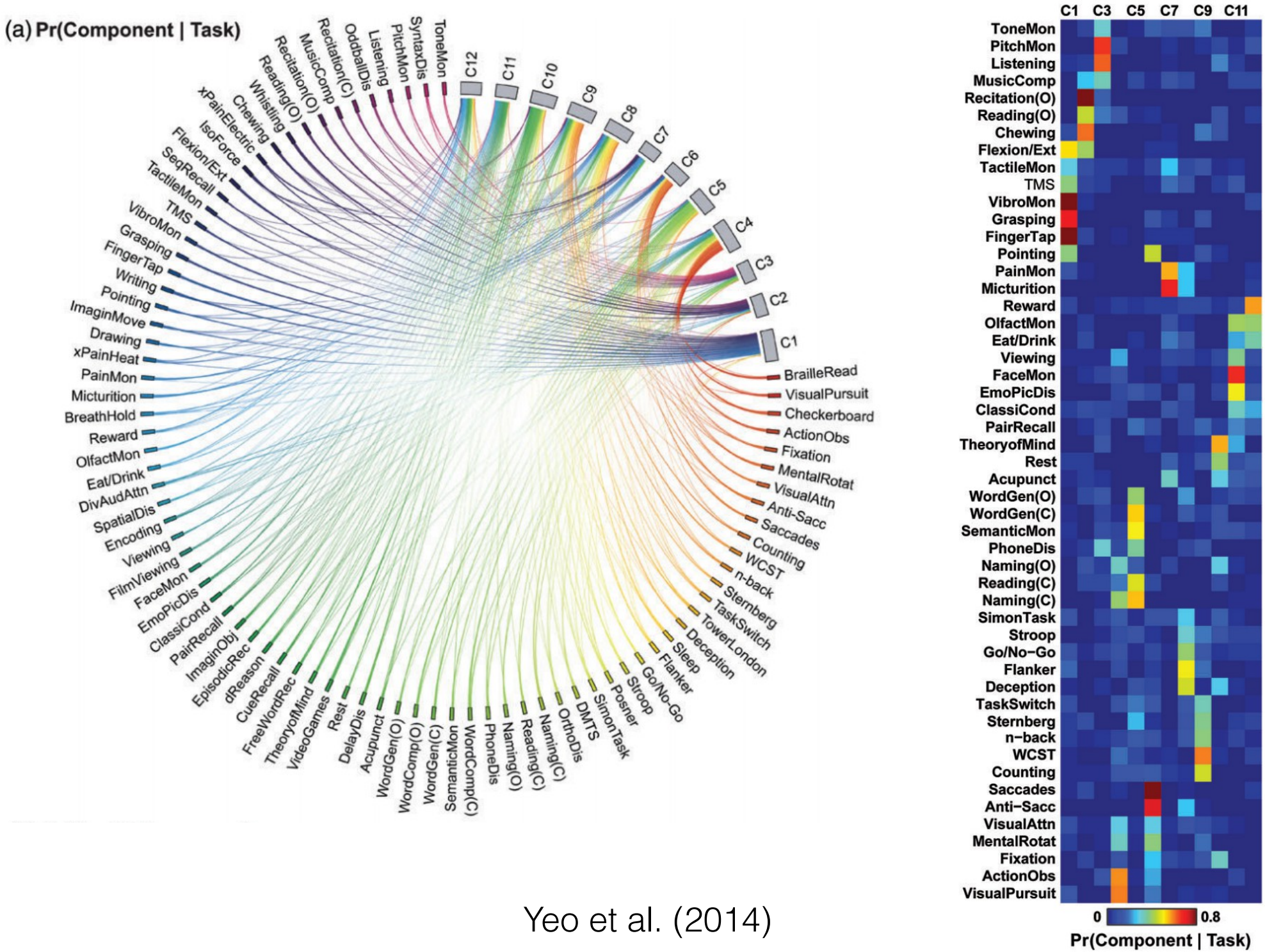


Yeo et al. (2014)

“Cognitive components”



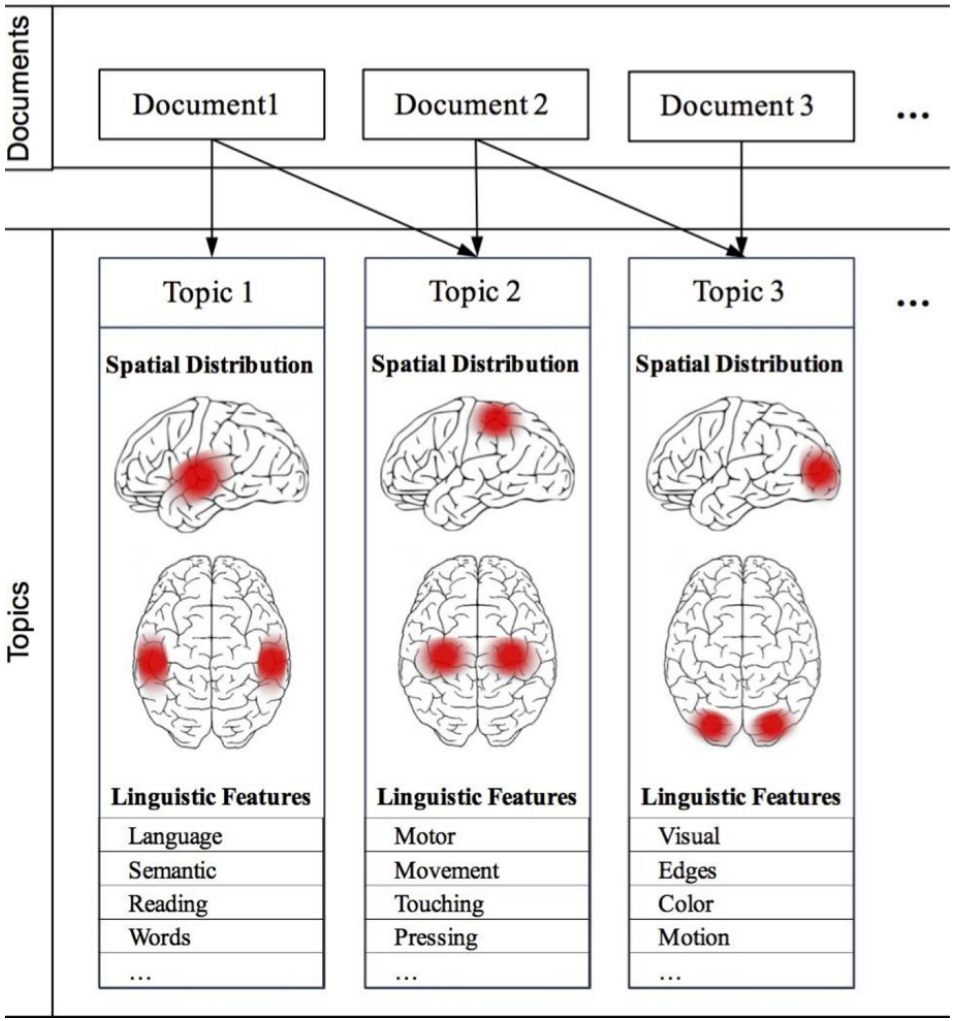
Yeo et al. (2014)



Yeo et al. (2014)

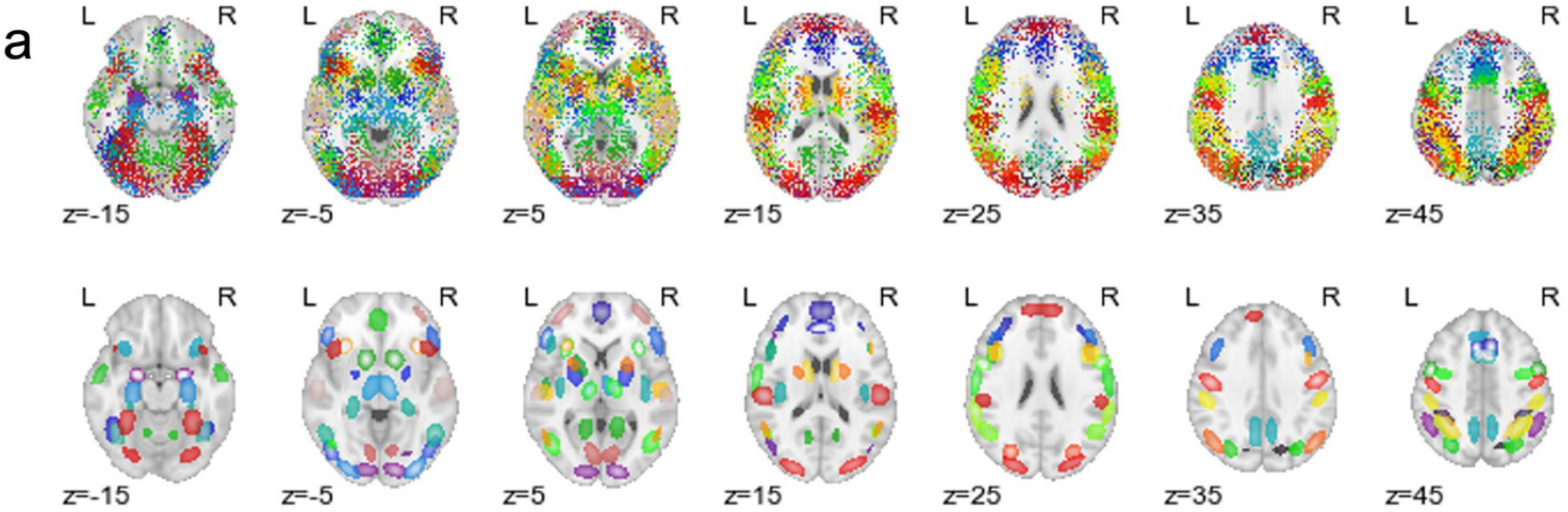
Generalized Correspondence LDA

- A generalization of correspondence-LDA (Blei, Ng, & Jordan, 2003)
- Each latent topic is associated with both a spatial distribution and a set of word tokens
- Explicitly designed to produce *region-like* topics
- We apply a symmetric constraint to produce bilaterally distributed topics



Rubin et al. (submitted)

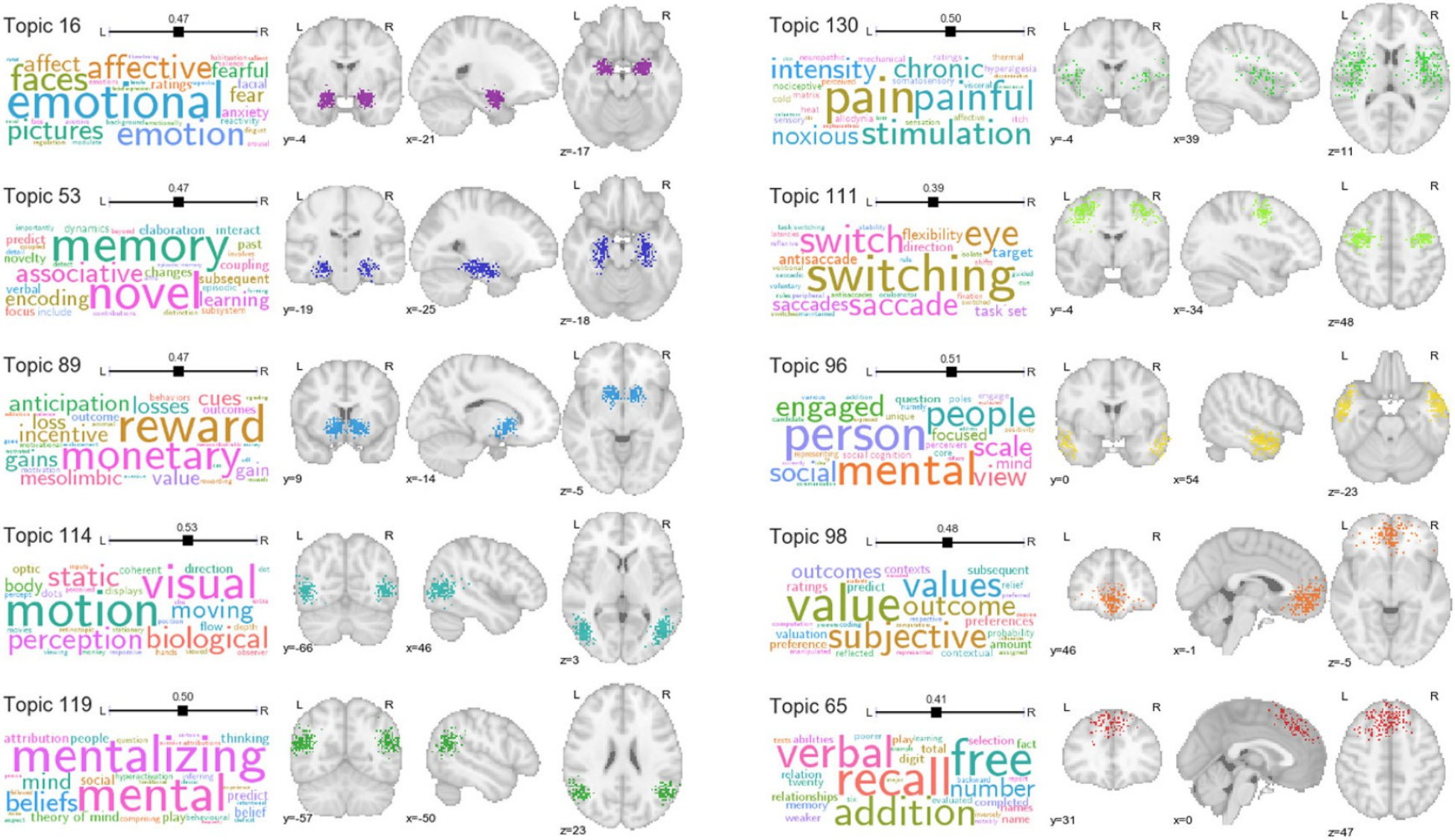
Neurosynth GC-LDA topics



Rubin et al. (submitted)

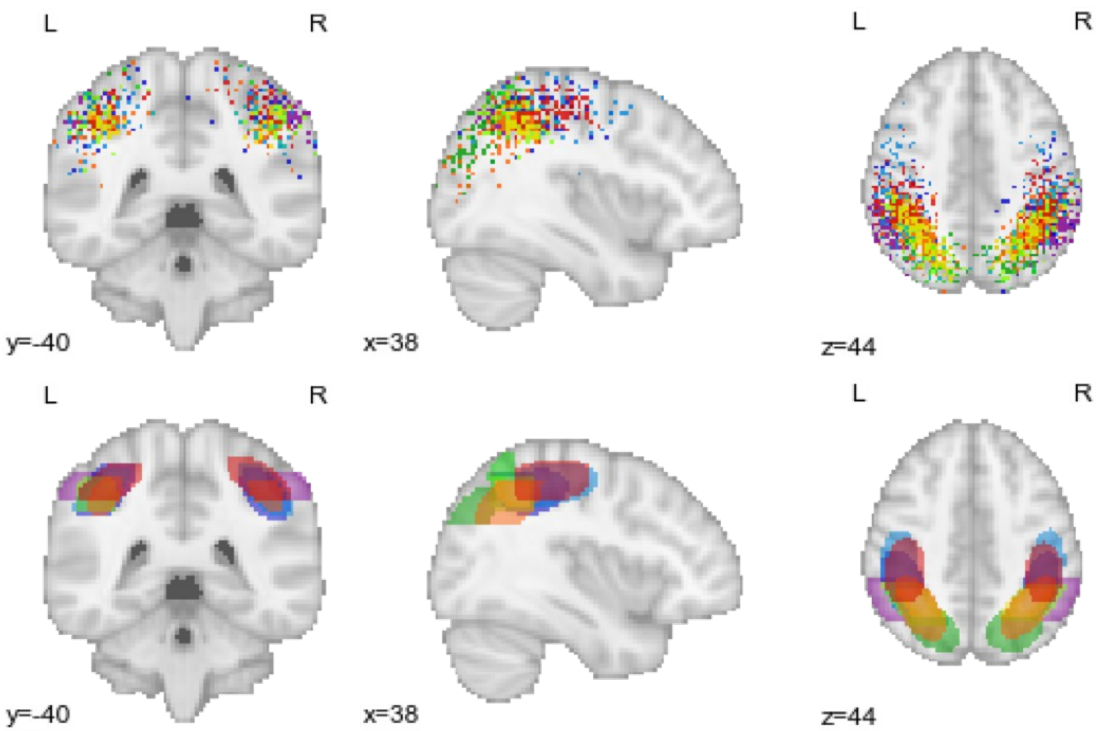
Sample topics

b



Rubin et al. (submitted)

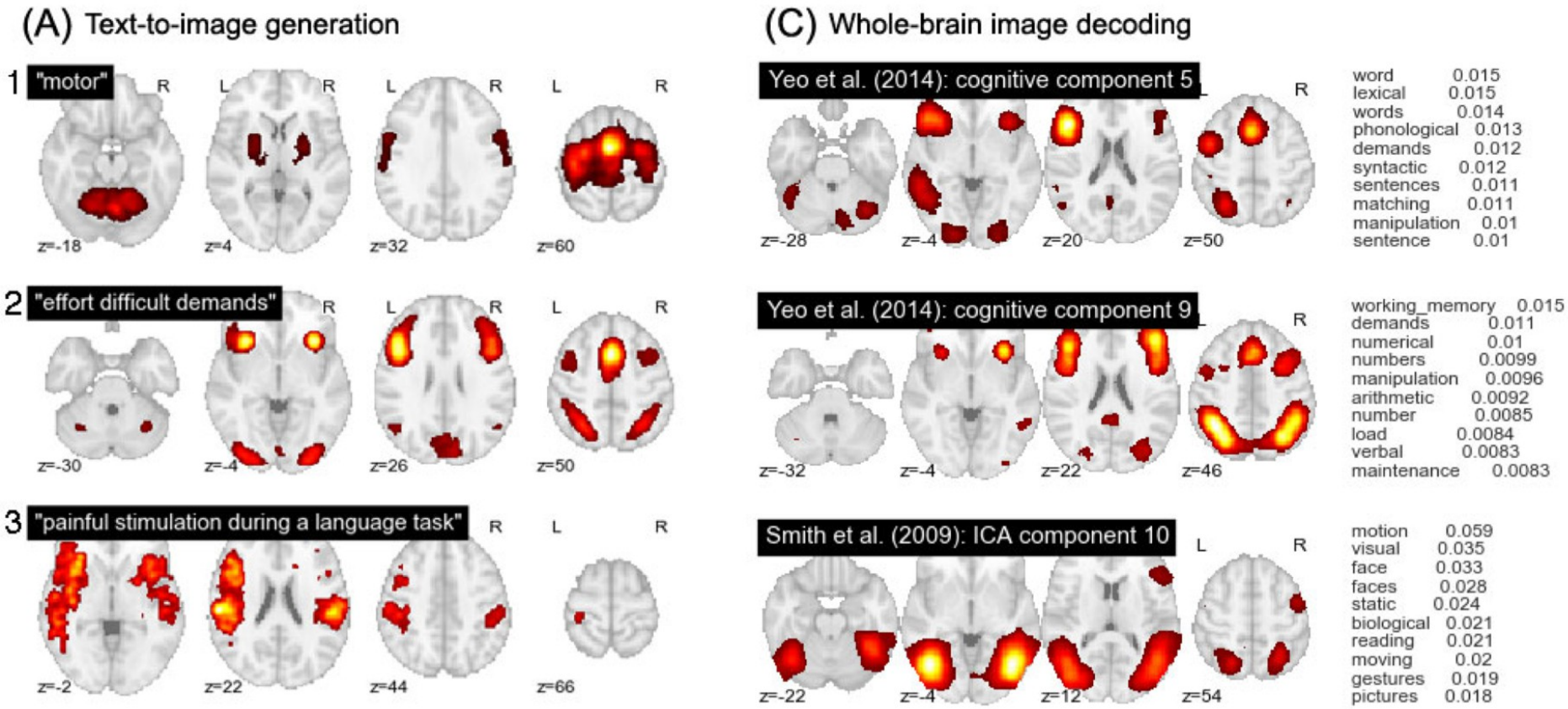
Topics have probabilistic distributions



Rubin et al. (submitted)

Bidirectional decoding

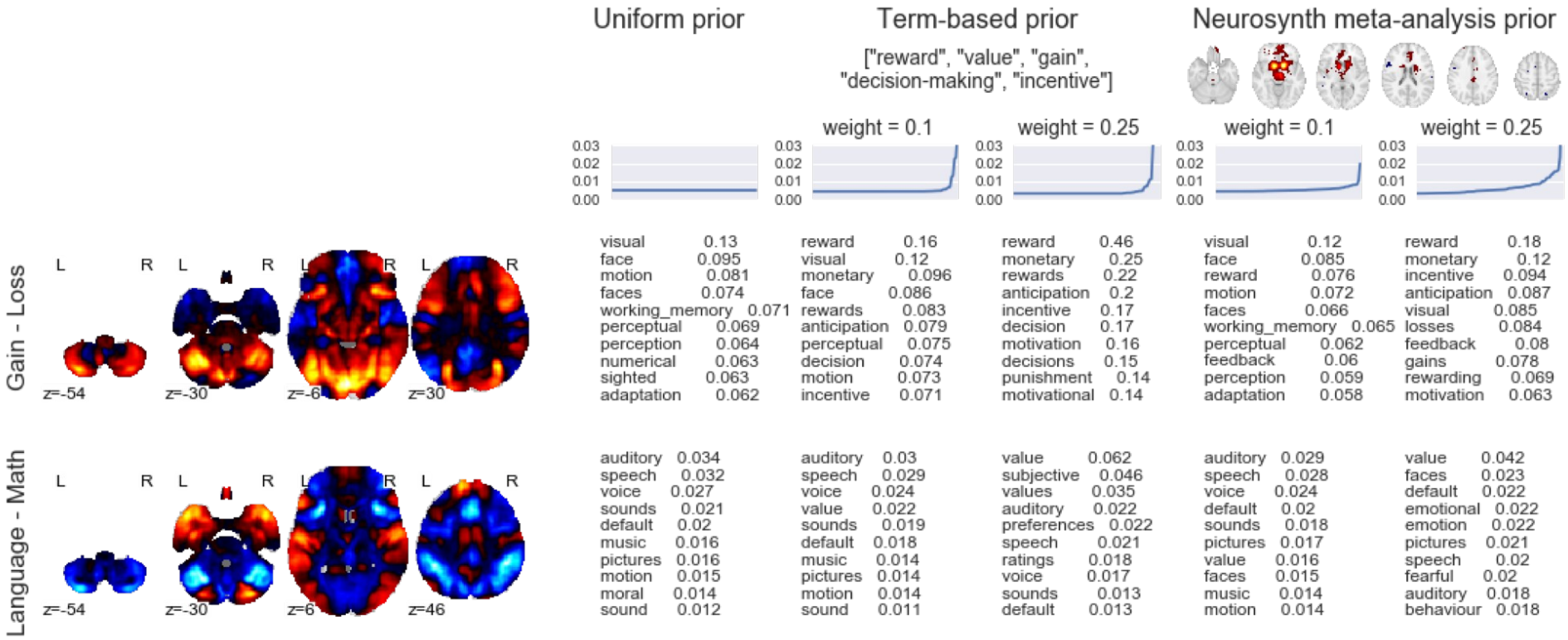
- We can decode both text from images/coordinates and images from text



Rubin et al. (submitted)

Context-sensitive decoding

- The framework is Bayesian; we can “seed” the model with priors derived from either text or images



Rubin et al. (submitted)

Grand, sweeping conclusions

- fMRI is great
- But no study is an island
  - There are principled limitations that make it hard to learn very much from any individual study
- Many kinds of inference require large-scale approaches
- Meta-analysis provides…
  - High sensitivity
  - More generalizable inferences
  - Estimation of relative specificity of activation
  - Ability to model brain-cognition mappings more comprehensively