

# Understanding brain function with machine learning on large-scale data repositories



**Machine Learning for Life Science**

*15-17 Nov 2022 Montpellier (France)*



Bertrand Thirion, [bertrand.thirion@inria.fr](mailto:bertrand.thirion@inria.fr)

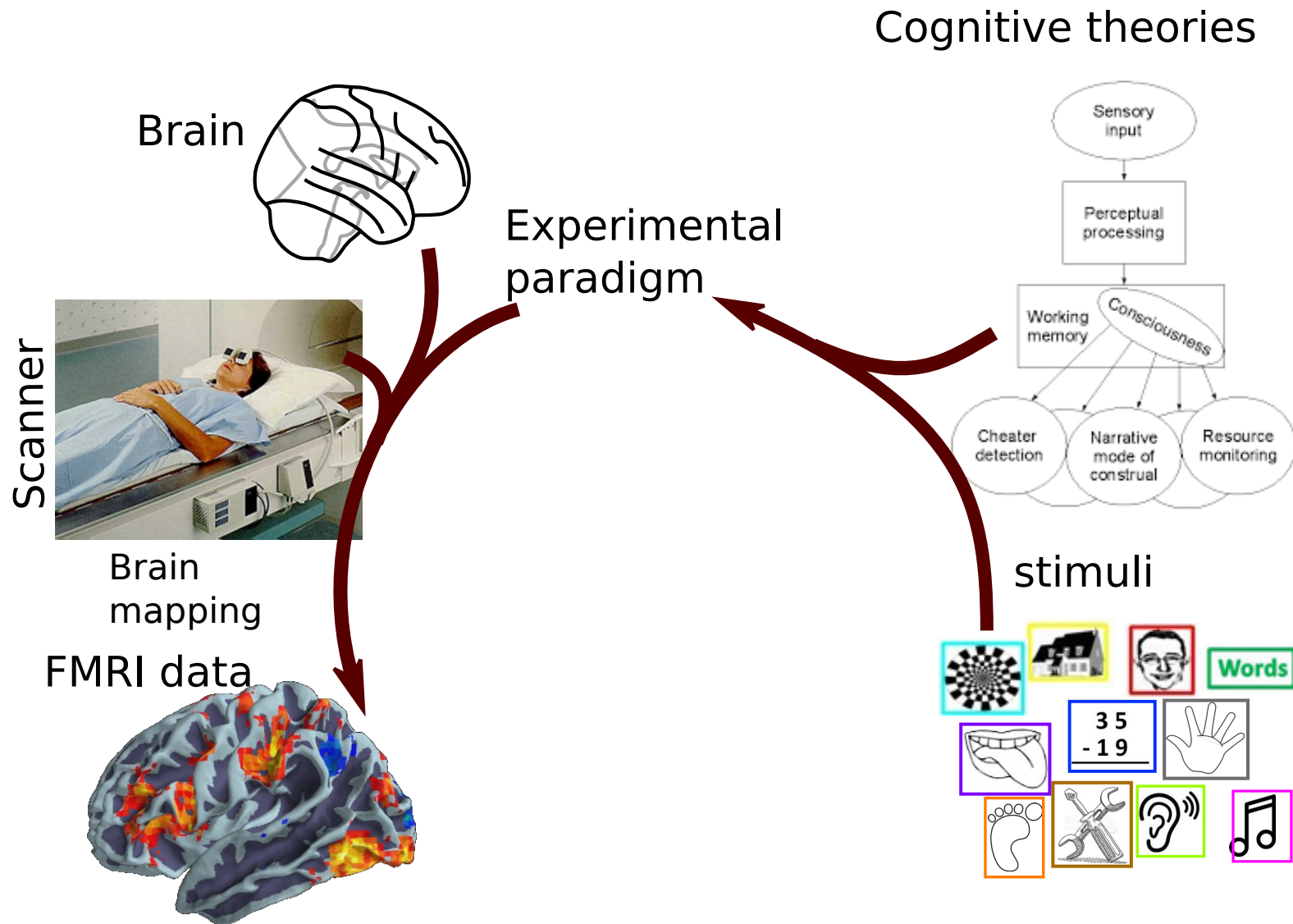


Bertrand Thirion

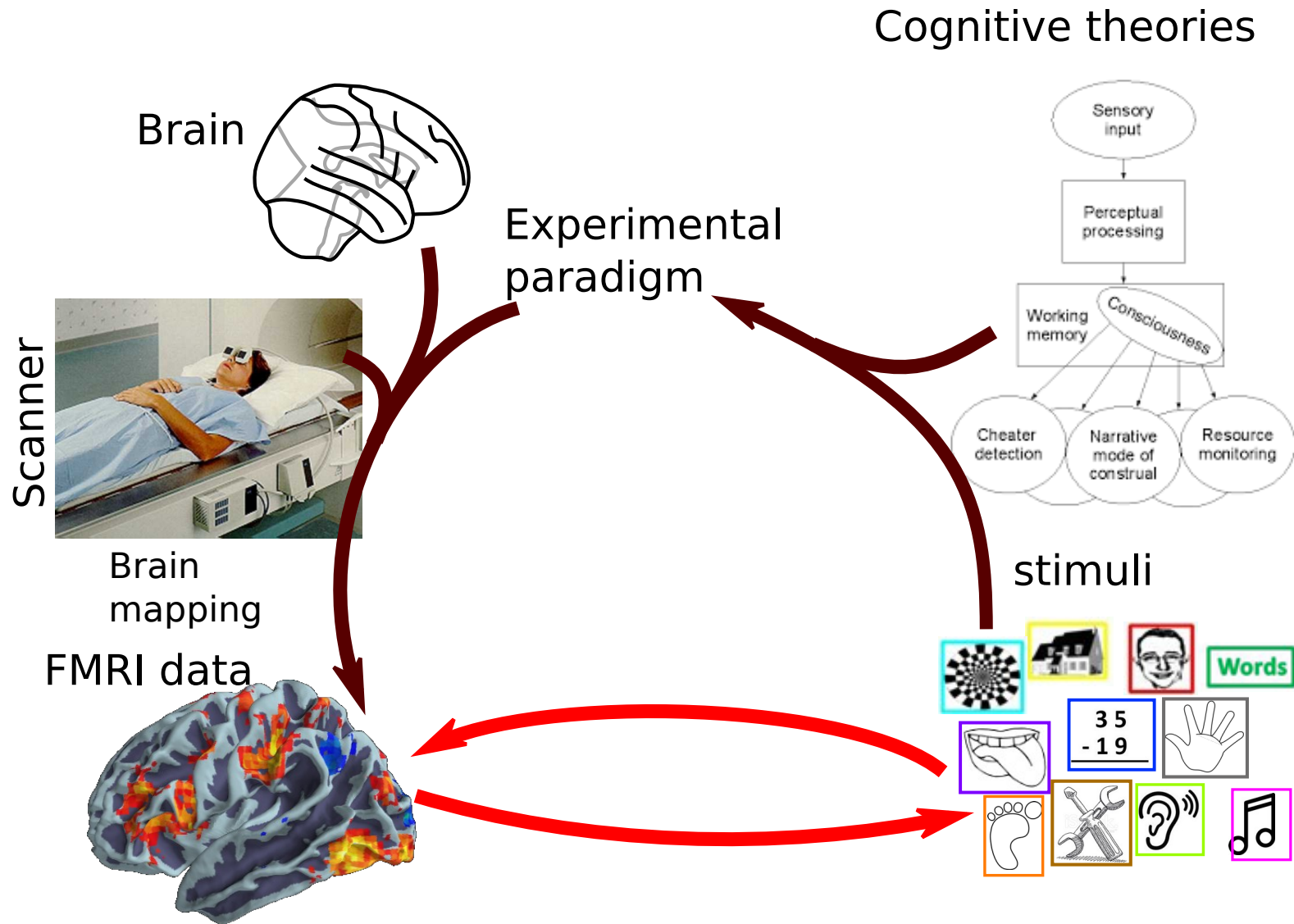
# Cognitive neuroscience

How are cognitive activities affected or controlled by neural circuits in the brain ?

# Cognitive neuroscience: From cognitive questions to data

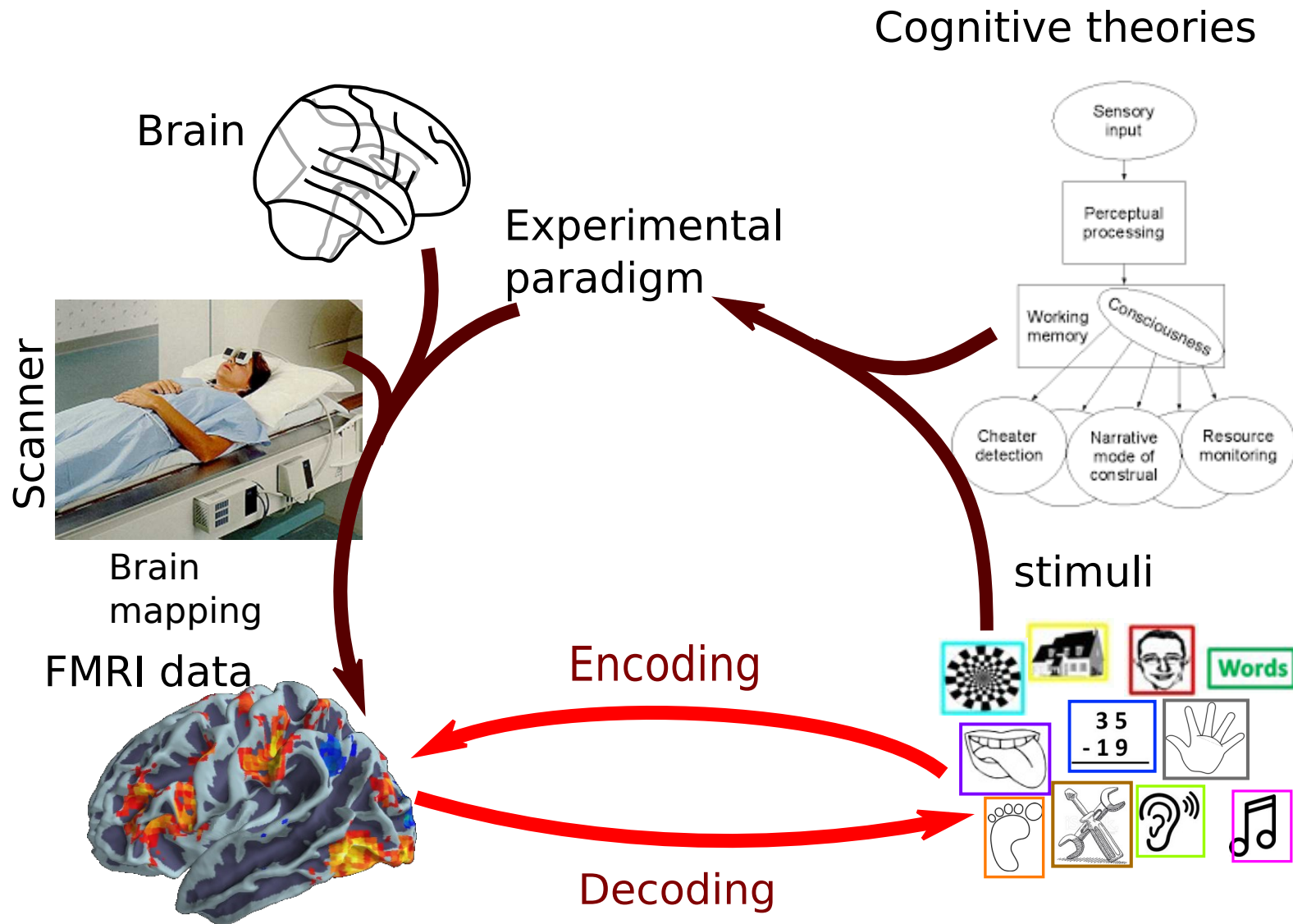


# Cognitive neuroscience: From cognitive questions to data





# Cognitive neuroscience: Brain activity *decoding*



# The big data revolution is ongoing – in neuroimaging also !

*Nature Reviews Neuroscience* | AOP, published online 10 April 2013; doi:10.1038/nrn3475



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

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

[https://en.wikipedia.org/wiki/Replication\\_crisis](https://en.wikipedia.org/wiki/Replication_crisis)

# The big data revolution is ongoing – in neuroimaging also !

*Nature Reviews Neuroscience* | AOP, published online 10 April 2013; doi:10.1038/nrn3475



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

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




**nature**  
REVIEWS

**NEUROSCIENCE**

Analysis | Published: 05 January 2017

## Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack , Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward Vul & Tal Yarkoni

# Problem: generalization across studies

“You cannot play 20 questions with nature and win”

[Newell A. Visual information processing; 1973]

[Poldrack & Yarkoni, Annu Rev Psycho 2016]

- **Joint analysis**: Use large studies to inform small studies (*transfer learning*)
  - Principle: leverage joint representations across datasets
- **Mega-analysis**: find *semantic* commonalities across studies
  - Difficulty: what common vocabulary across studies?

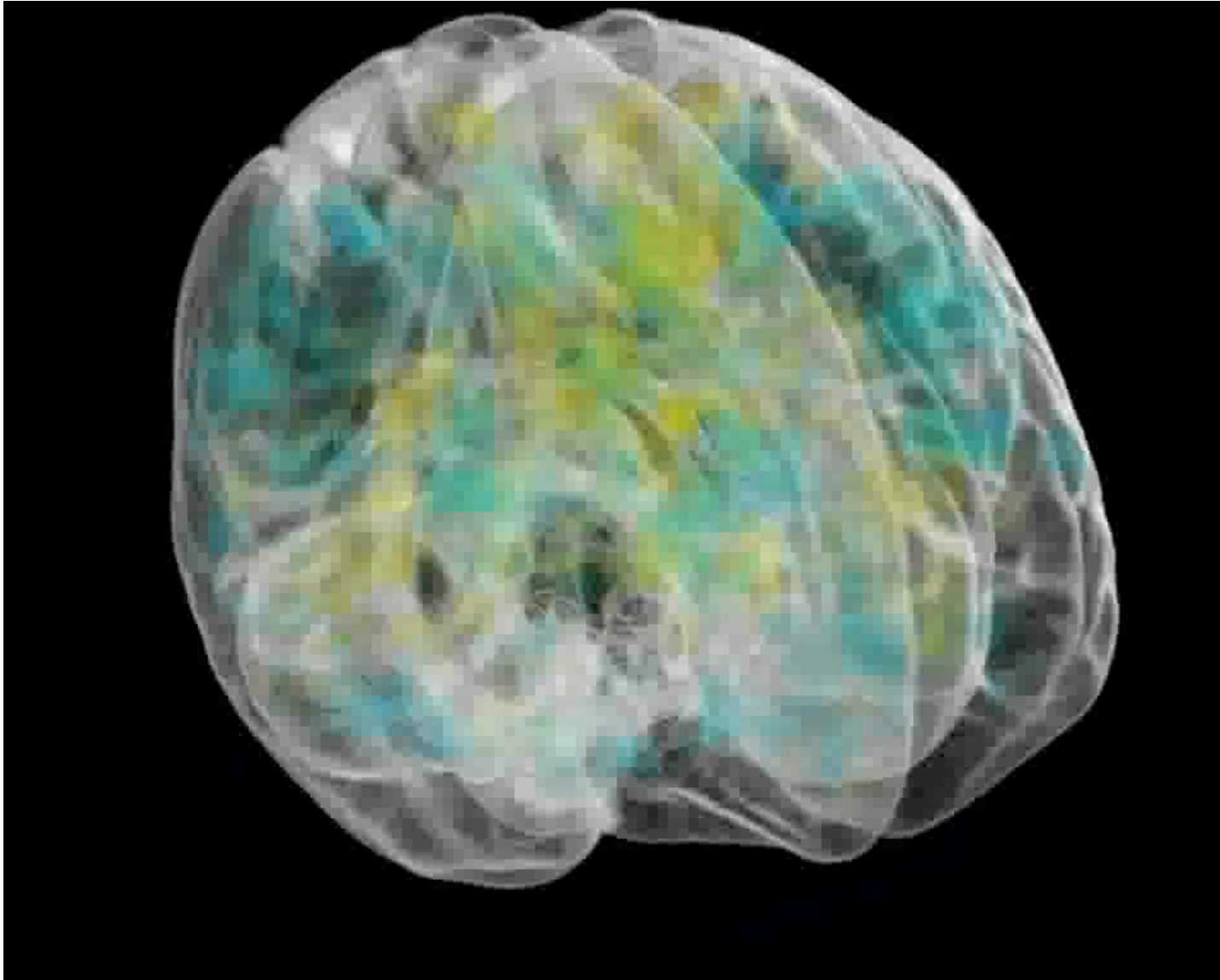
# Outline

- Learning good representations for brain images: unsupervised approach
- Learning good representations for brain images: supervised approach
- *In the wild* brain activity decoding
- Dealing with semantics and data labelling issues

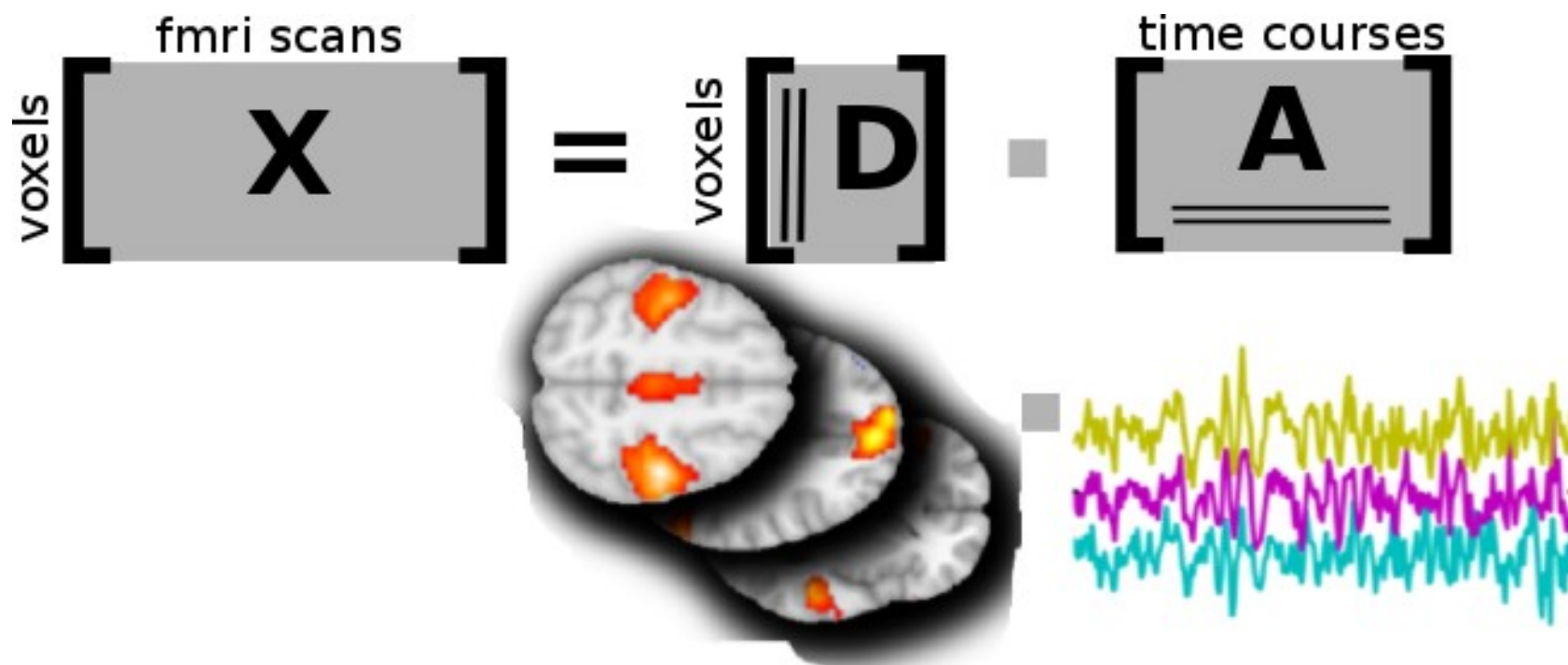
# Learning good representations for brain images: unsupervised approach



# Discovering structure in fMRI data



# Discovering structure in fMRI



$$\operatorname{argmin}_{A, D} \|X - DA\|^2 + \lambda \|D\|_1$$

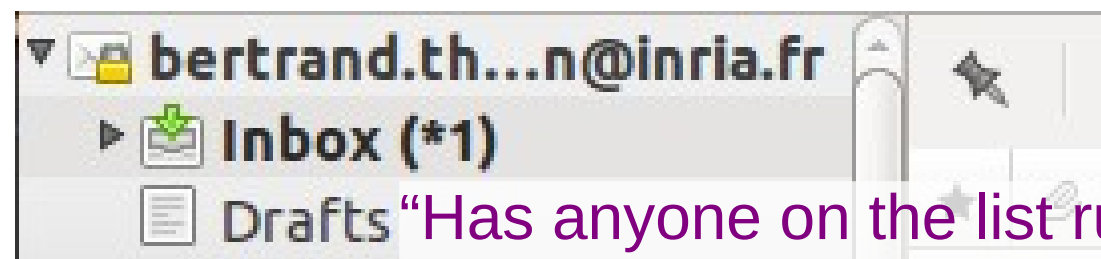
Can be captured by dictionary learning / sparse coding

[Olshausen Nature 1996]

→ Use of sparse PCA

# High-dimensional fMRI

- $n$  = number of samples,  $10^2$  to  $10^6$
- $p$  = number of voxels,  $10^5$ - $10^6$



“Has anyone on the list run group-wise analysis on the HCP resting state data, and if so what tools did you use?”

**I am having memory issues when running more than 10 subjects** and I was wondering if anyone has a way of getting around the large memory requirements when concatenating in time.”

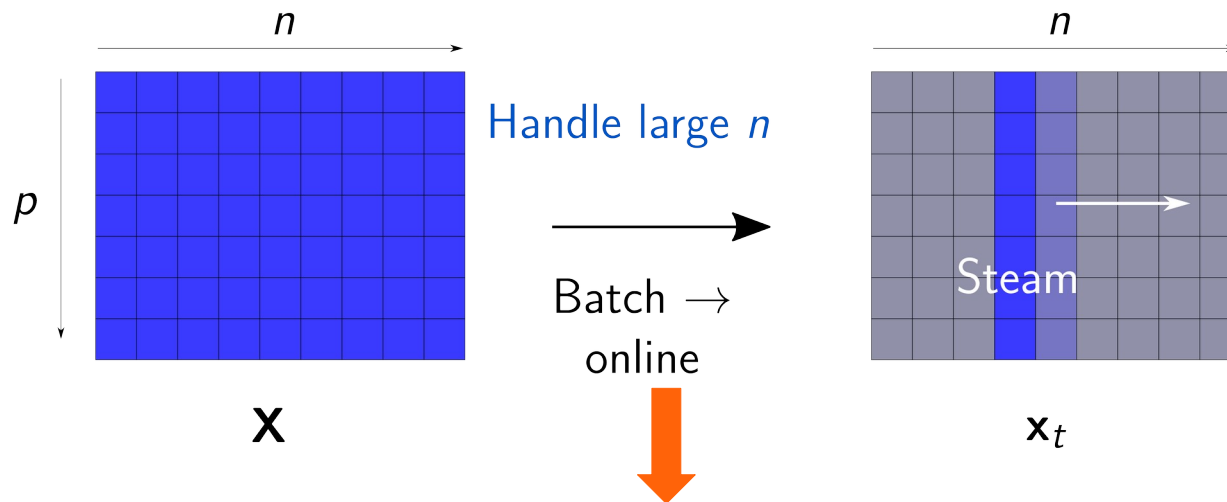
# Factorizing high-dimensional data

- Human Connectome project  $n=4.10^6$ ,  $p=2.10^5$ , 4**TB** of data
- Online dictionary learning [Mairal et al. ICML 2009]
- How to go faster ?
  - Work on batches of images **and** voxels
    - Online method in both samples and feature dimensions

[Mensch et al. ICML 2016, IEEE TSP 2018]

# Stochastic gradient approaches

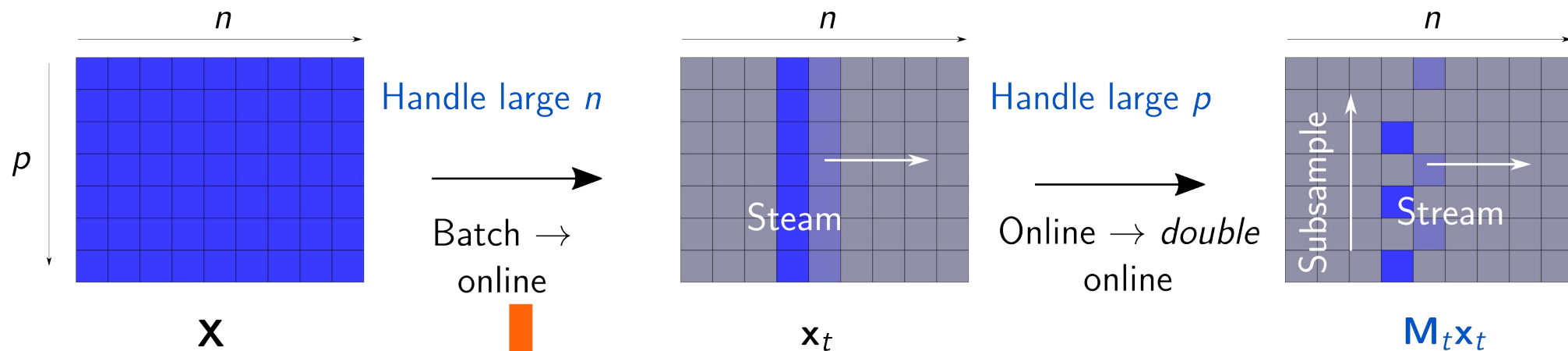
<http://amensch.fr/research/2016/06/10/modl.html>



$$\alpha_t(\mathbf{D}) = \operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{k \times n}} \|\mathbf{x}_t - \mathbf{D}_{t-1} \alpha_t\|_F^2 + \lambda \Omega(\alpha_t)$$
$$\mathbf{D}_t = \operatorname{argmin}_{\mathbf{D} \in \mathcal{C}} \sum_{i=1}^t \|\mathbf{x}_i - \mathbf{D} \alpha_i\|_F^2$$

# Stochastic gradient approaches

<http://amensch.fr/research/2016/06/10/modl.html>



$$\alpha_t(\mathbf{D}) = \operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{k \times n}} \|\mathbf{x}_t - \mathbf{D}_{t-1} \alpha_t\|_F^2 + \lambda \Omega(\alpha_t)$$

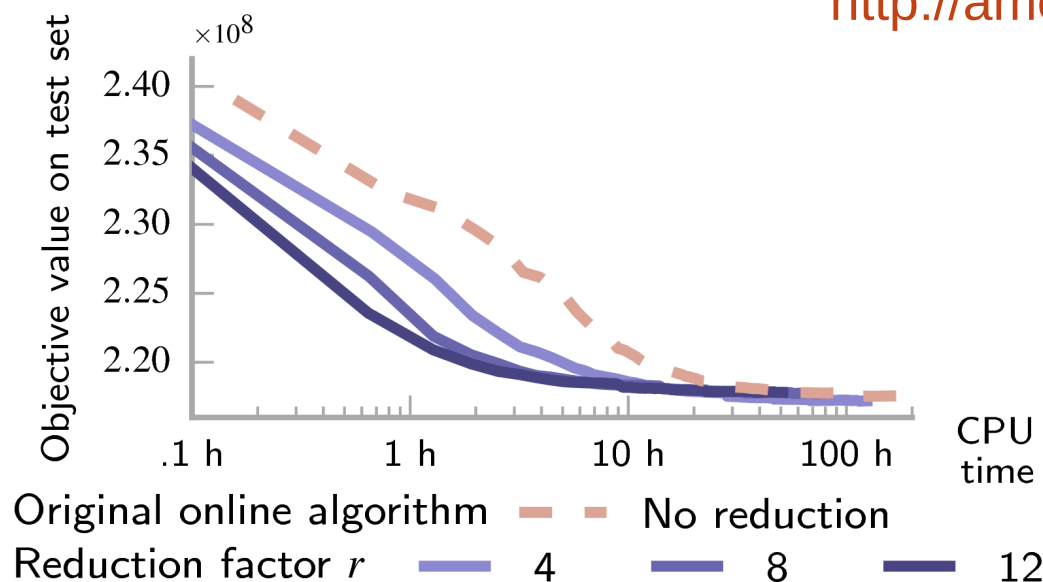
$$\mathbf{D}_t = \operatorname{argmin}_{\mathbf{D} \in \mathcal{C}} \sum_{i=1}^t \|\mathbf{x}_i - \mathbf{D} \alpha_i\|_F^2$$

$$\alpha_t(\mathbf{D}) = \operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{k \times n}} \|\mathbf{M}_t(\mathbf{x}_t - \mathbf{D}_{t-1} \alpha_t)\|_F^2 + \lambda \frac{s}{p} \Omega(\alpha)$$



# Stochastic gradient approaches

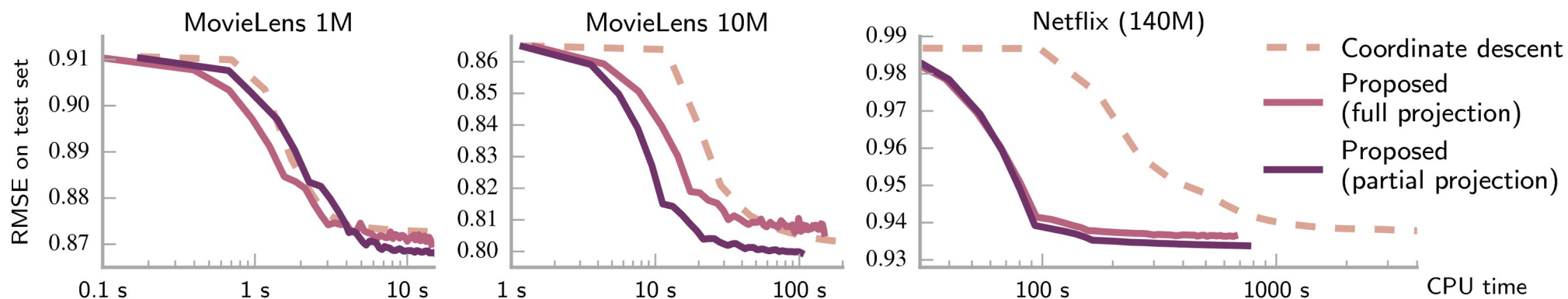
<http://amensch.fr/research/2016/06/10/modl.html>



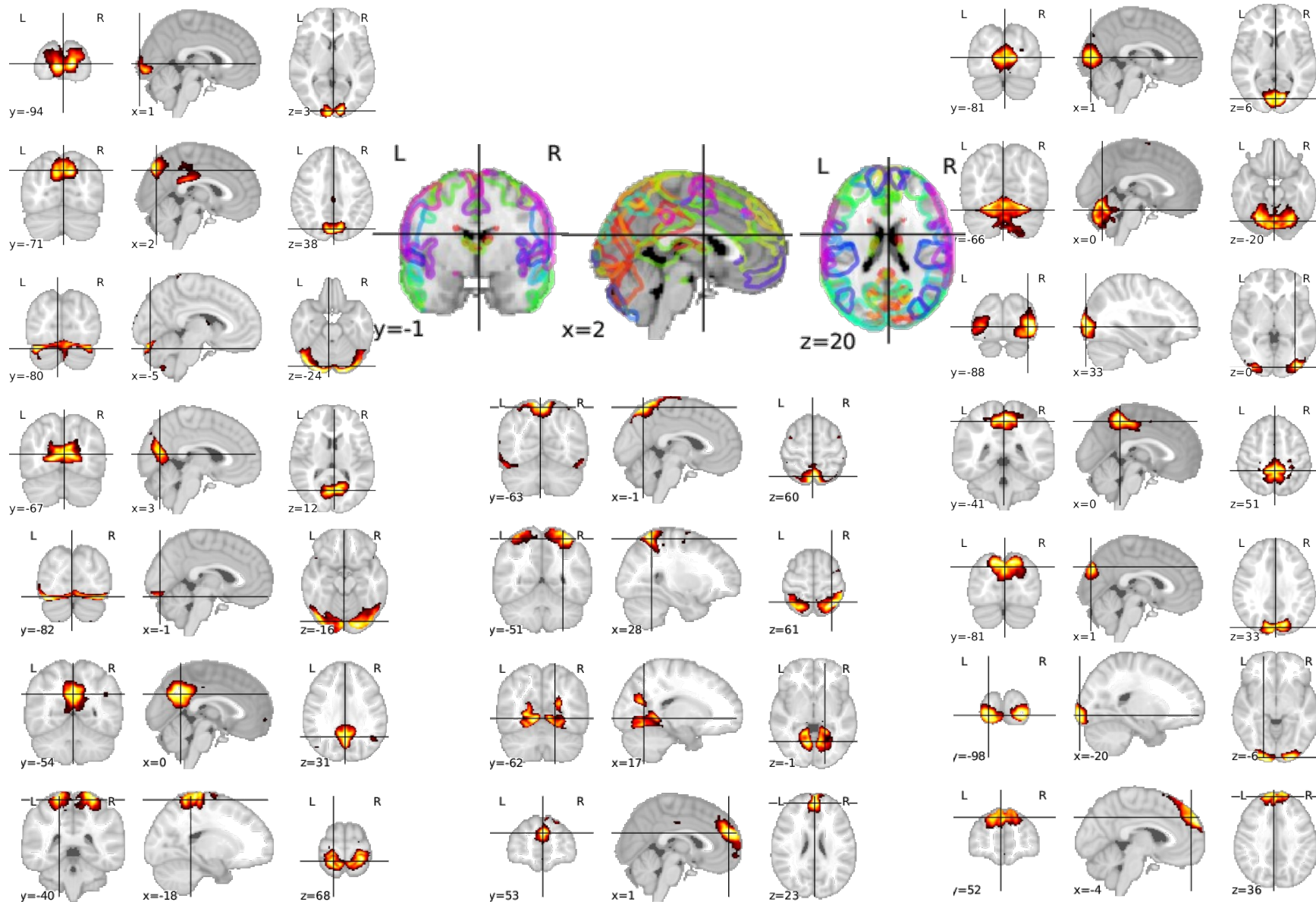
10-fold gain in CPU time  
without loss in accuracy

[Mensch et al. ICML 2016,  
IEEE TSP 2018]

Can be used for recommender systems



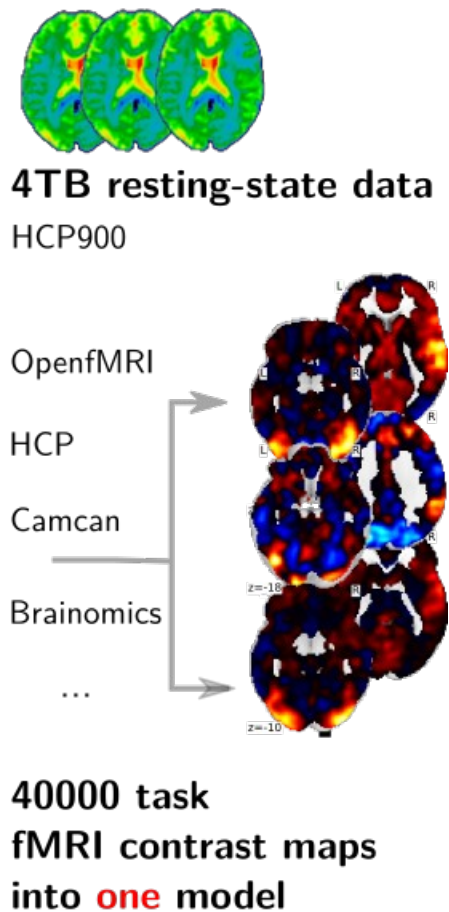
# Brain atlases



[Mensch et al. ICML 2016 IEEE TSP 2018, Dadi et al. Nimg 2020]

# Learning good representations for brain images: supervised approach

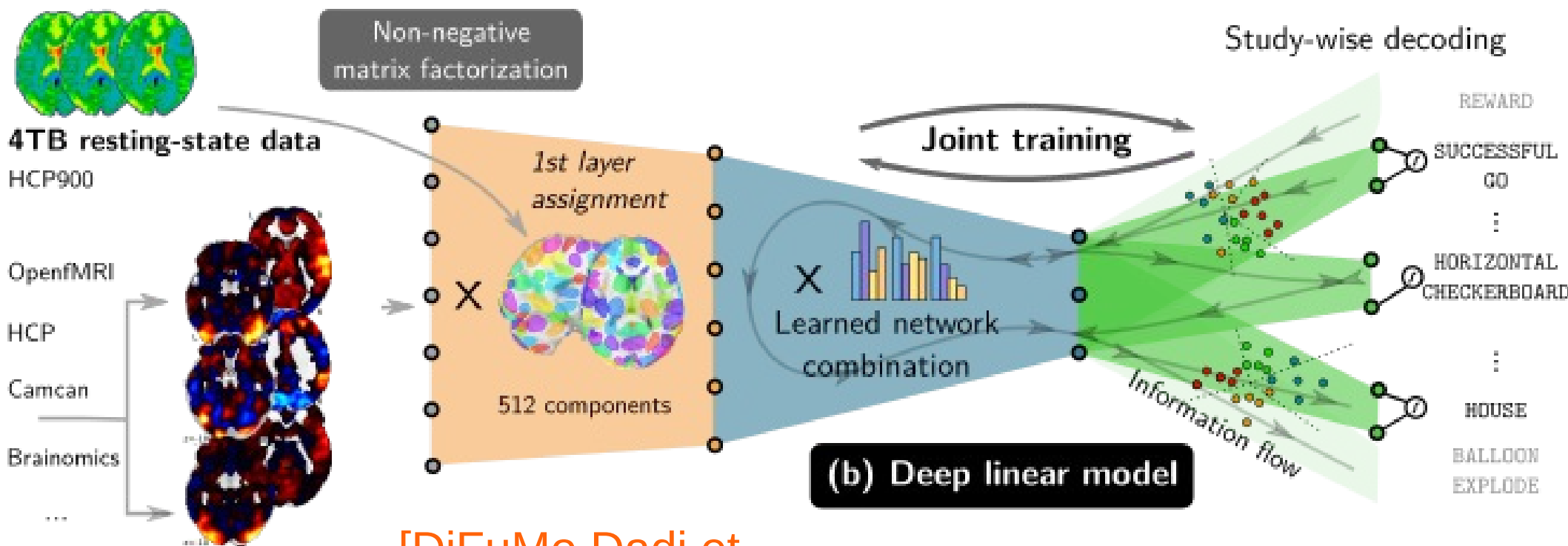
# Predictive modeling across datasets



(a) Aggregation  
from many  
fMRI studies

[Bzdok et al. Plos Comp Biol 2016, Mensch et al NIPS 2017, PCB 2021]

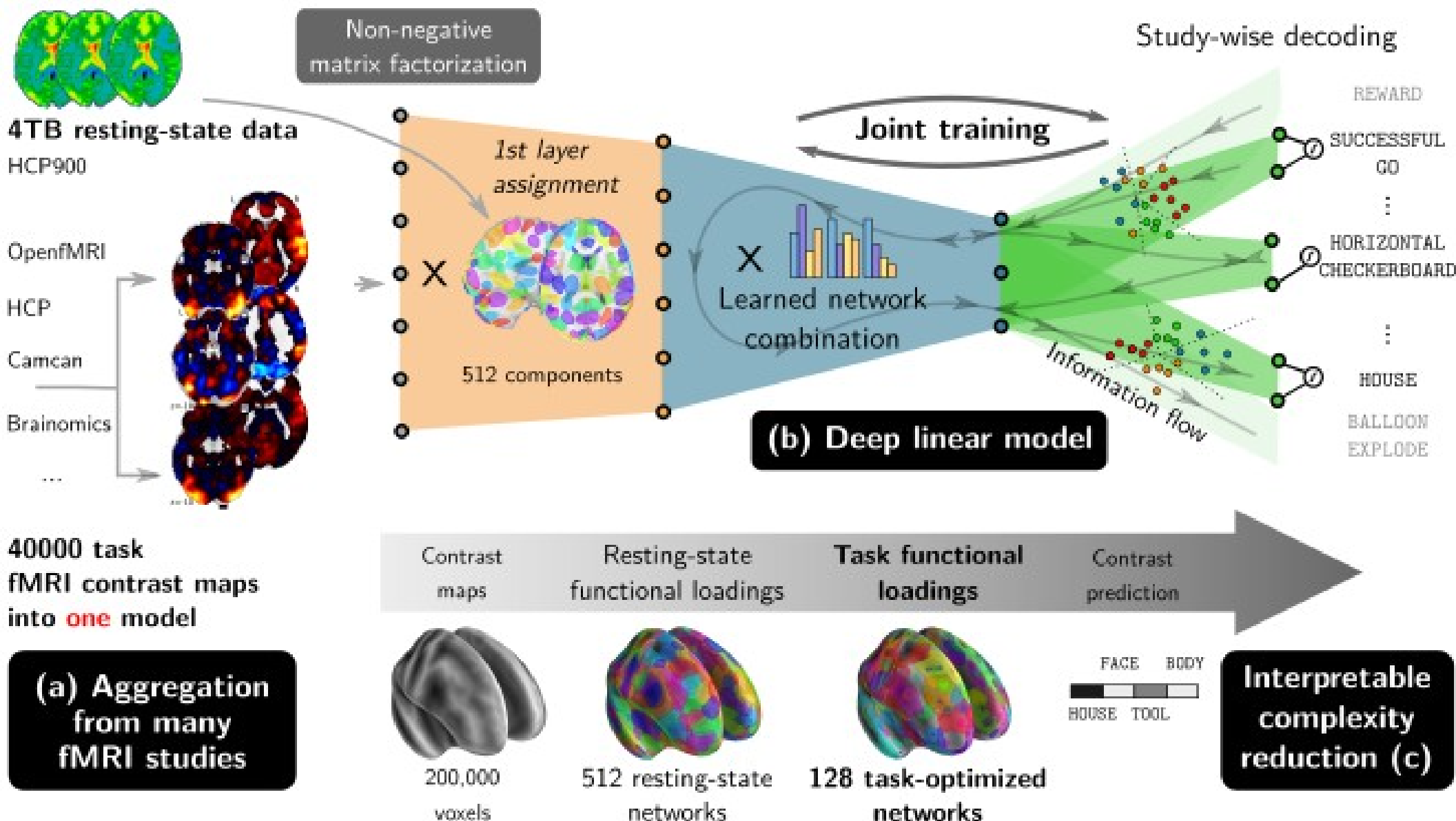
# Predictive modeling across datasets



[DiFuMo Dadi et al. Nimg 2020]

[Bzdok et al. Plos Comp Biol 2016, Mensch et al NIPS 2017 PCB 2021]

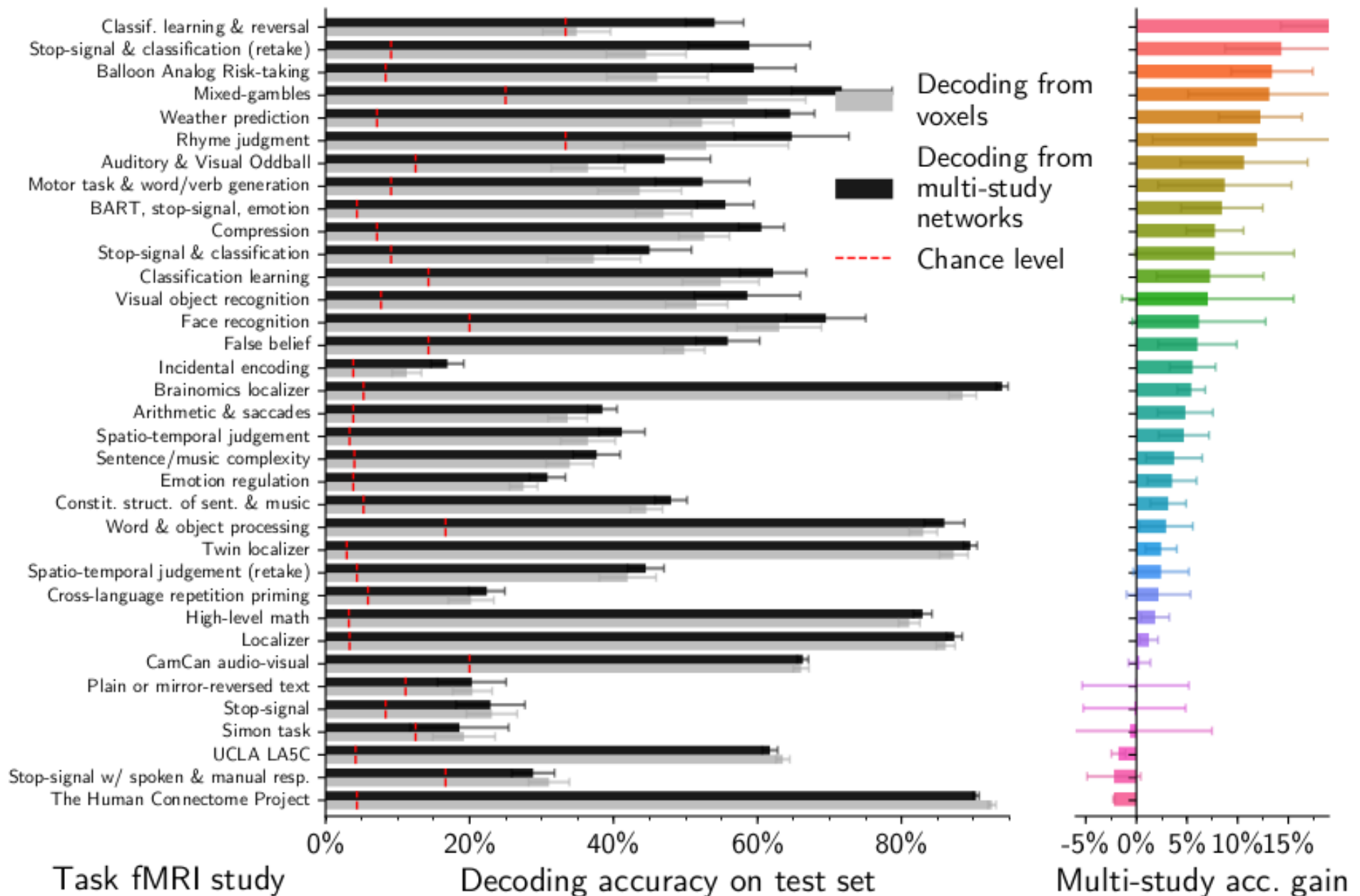
# Predictive modeling across datasets



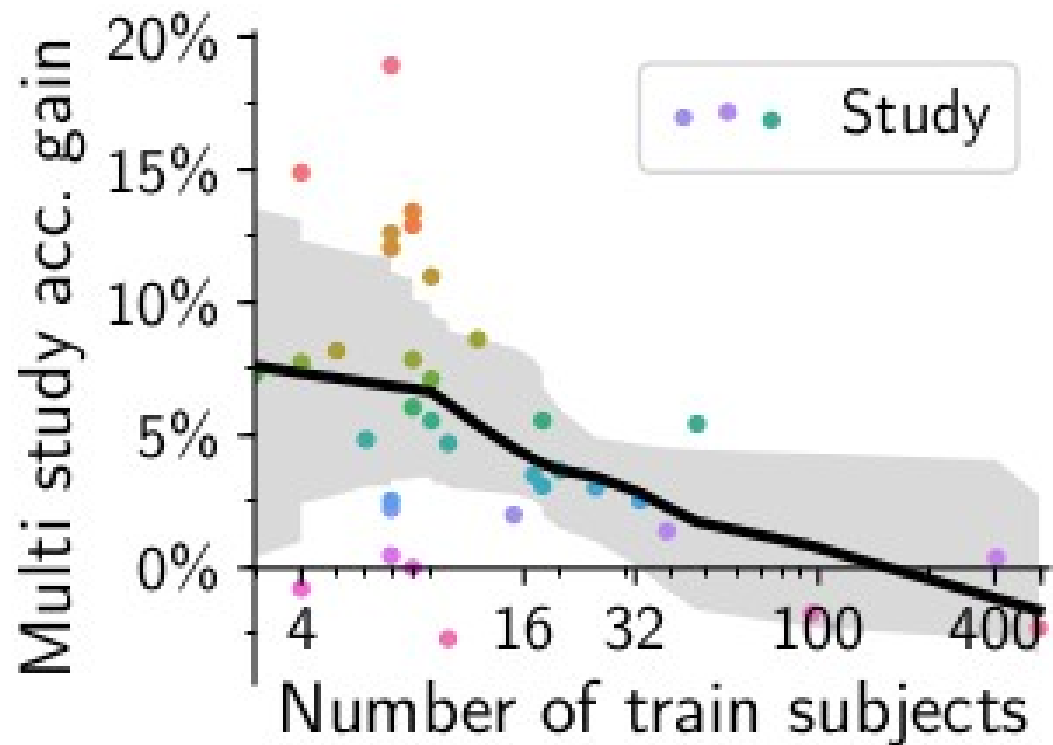
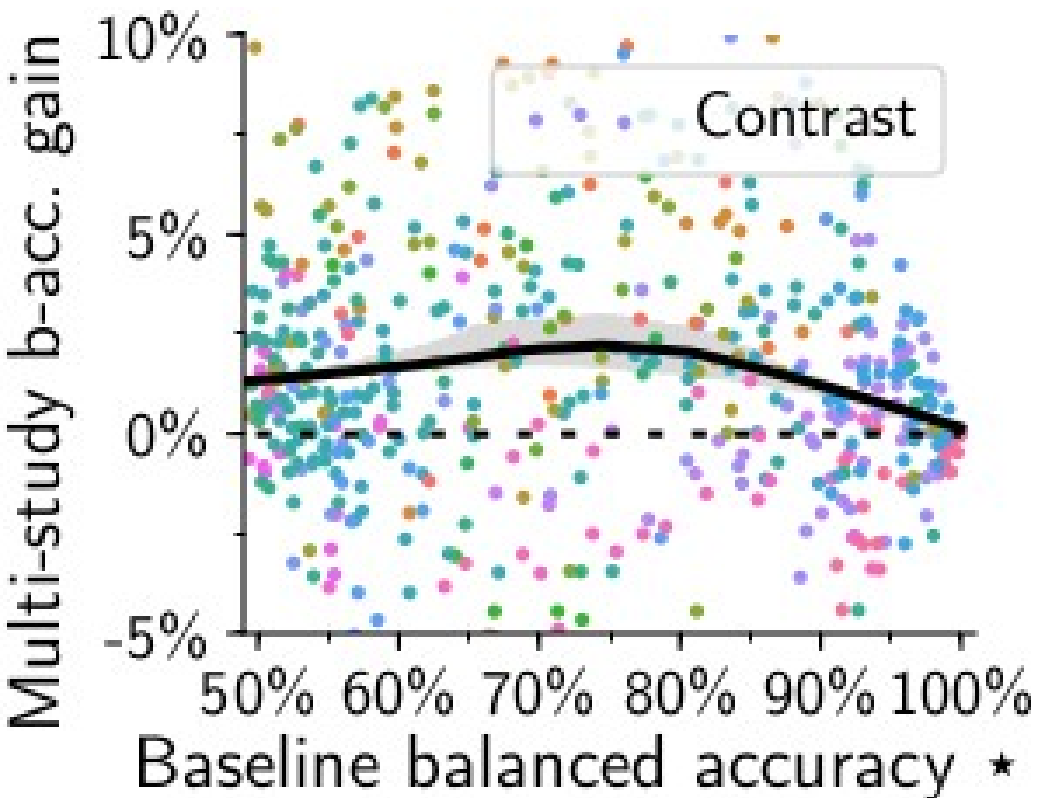
[Bzdok et al. Plos Comp Biol 2016, Mensch et al NIPS 2017, PCB 2021]



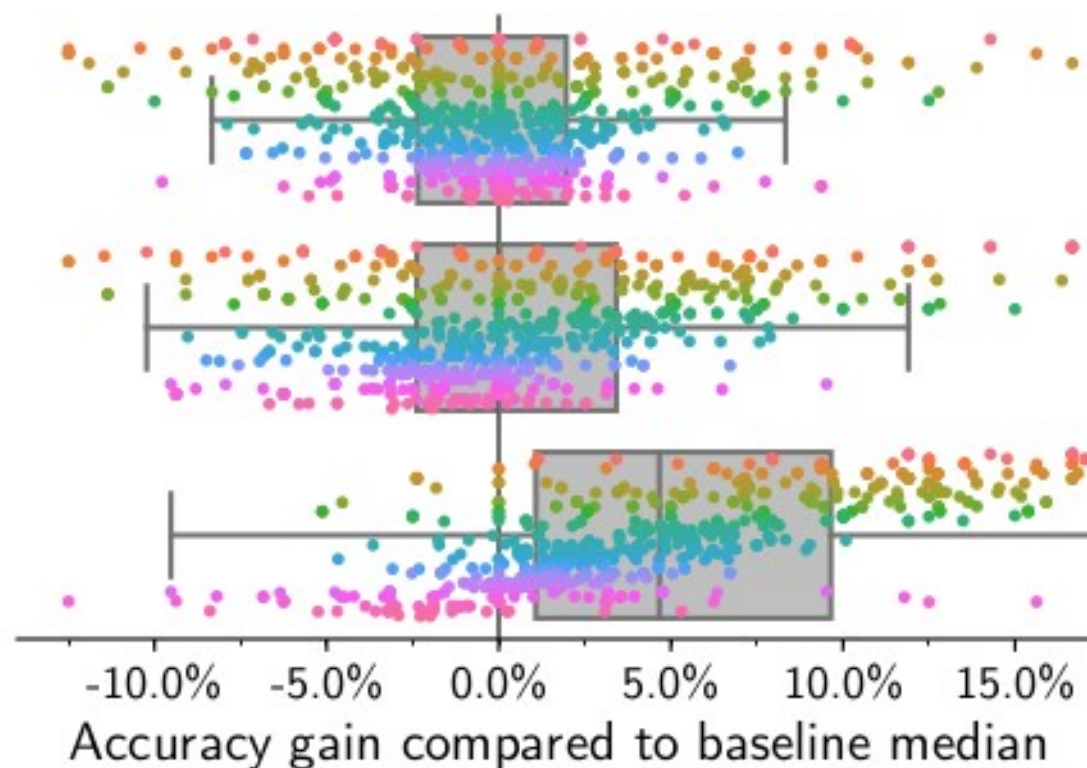
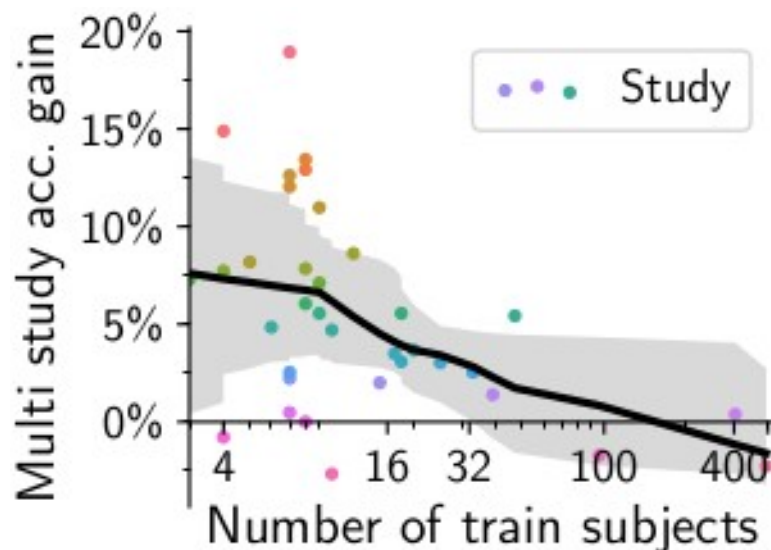
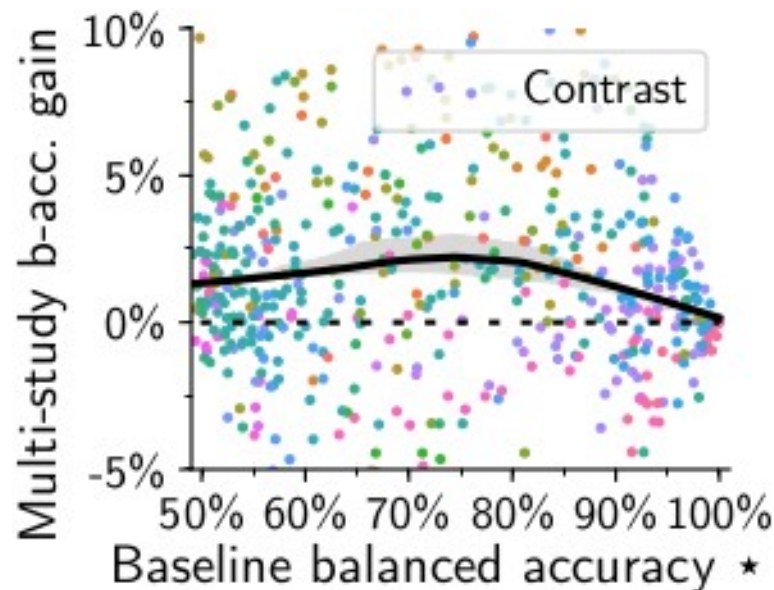
# Transfer learning



# Transfer learning



# Transfer learning

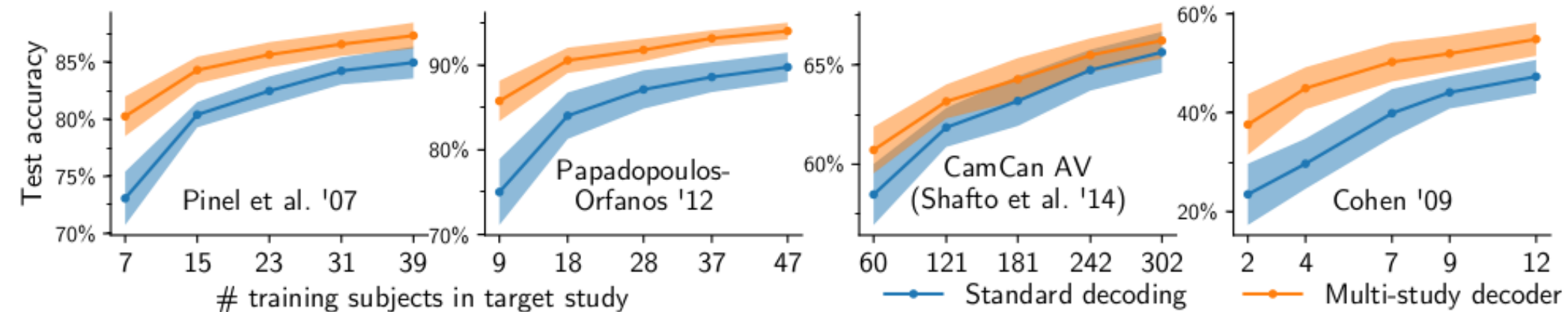


Standard decoding  
from voxels

Decoding from  
functional networks

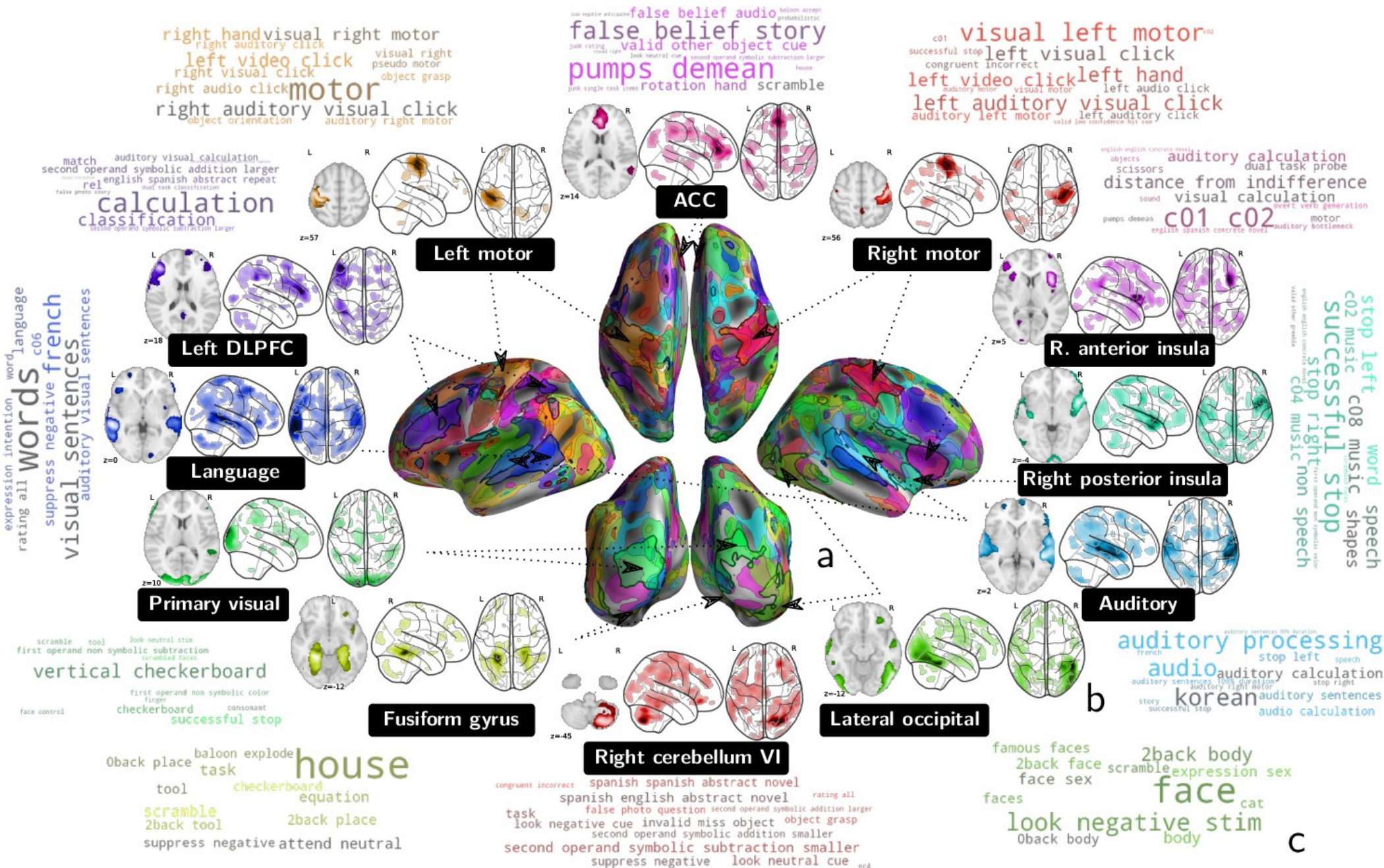
**Decoding from  
multi-study  
task-optimized  
networks**

# Small studies benefit more than large studies





# Resulting atlas





# Resulting atlas

Available at  
<https://github.com/arthurmensch/cogspaces>  
[Mensch et al. PCB 2021]

Nov 2022

Bertrand Thirion

28

# Resulting atlas

The resulting atlas displays various brain regions and their associated cognitive tasks, represented by word clouds:

- ACC**: false belief audio, valid other object cue, pumps demean, Rotation hand scramble.
- Left motor**: right handvisual right motor, left video click, right visual click, right audio clickmotor, right auditory visual click, object orientation auditory right motor.
- Right motor**: c01 visual left motor, successful stopleft visual click, congruent incorrect, left video clickleft hand, left auditory visual click, auditory left motor, left auditory click.
- Left DLPFC**: match auditory visual calculation, second operand symbolic addition larger, english spanish abstract repeat, dual task classification, false photo story, calculation, classification.
- R. anterior insula**: english english concrete novel, objects scissors, distance from indifference, sound visual calculation, pumps demean, c01 c02, oververb generation, auditory bottleneck.
- Primary visual**: expression intention word language, rating all words, suppress negative french, visual sentences, auditory visual sentences.
- Fusiform gyrus**: vertical checkerboard, first operand non symbolic subtraction, scrambled faces, face control, checkerboard, consonant, successful stop.
- Lateral occipital**: famous faces, 2back face, face sex, faces, look negative stim, body, 0back body.
- Right cerebellum VI**: congruent incorrect, spanish spanish abstract novel, rating all, spanish english abstract novel, false photo question, second operand symbolic addition larger, look negative cue invalid miss object object grasp, second operand symbolic addition smaller, second operand symbolic subtraction smaller, suppress negative, look neutral cue.
- Auditory**: auditory processing, stop left speech, c08 music shapes, successful stop, stop right non speech, c04 music.

Available at  
<https://github.com/arthurmensch/cogspaces>  
[Mensch et al. PCB 2021]

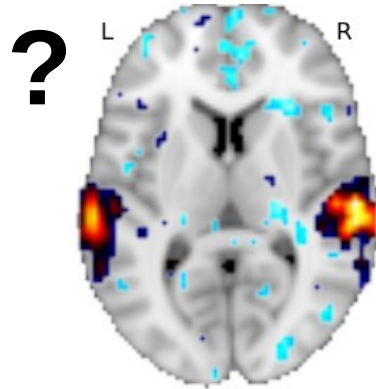
Nov 2022 Bertrand Thirion

[illegible]



# *In the wild* brain activity decoding

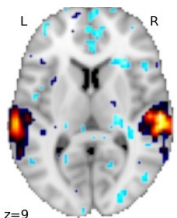
# Open-ended brain decoding



*What is  
this brain doing?*

**Which regions are predictive of tasks containing a given term?**

- **Multilabel** classification problem
  - more than one class may be associated with each sample
- Predict occurrence of frequent terms

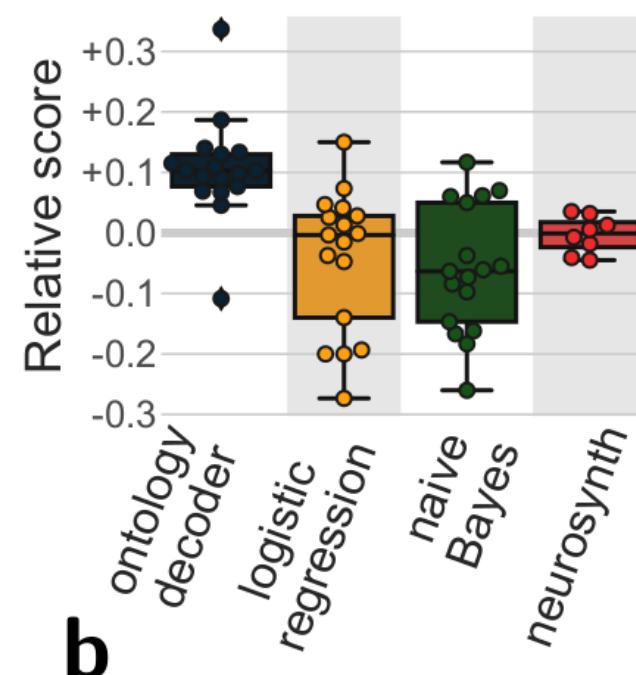
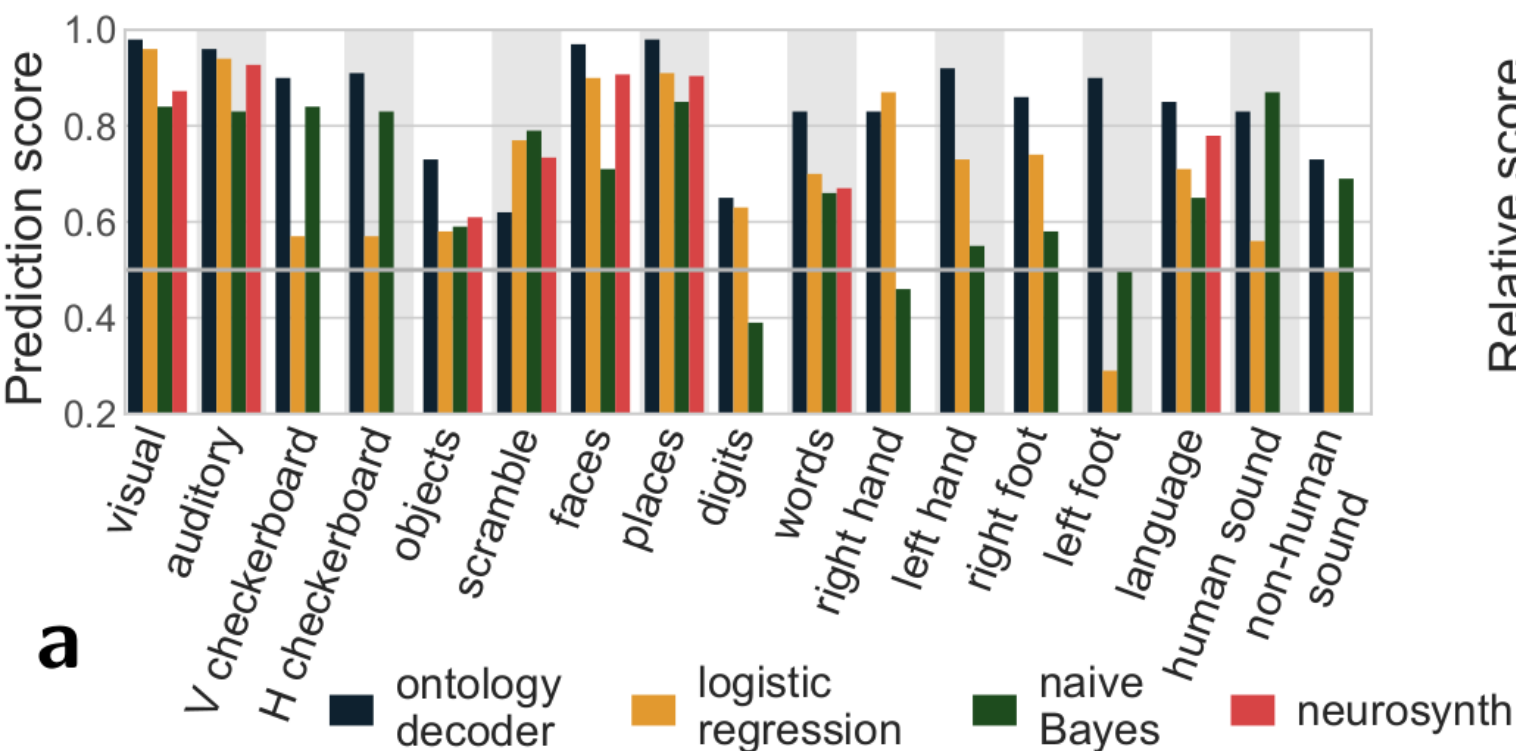


**Data: experimental  
condition images**

	visual	auditory	digits	words	count	...
sentences						
calculation						
tone listening						
tone counting						
successful stop						
...						

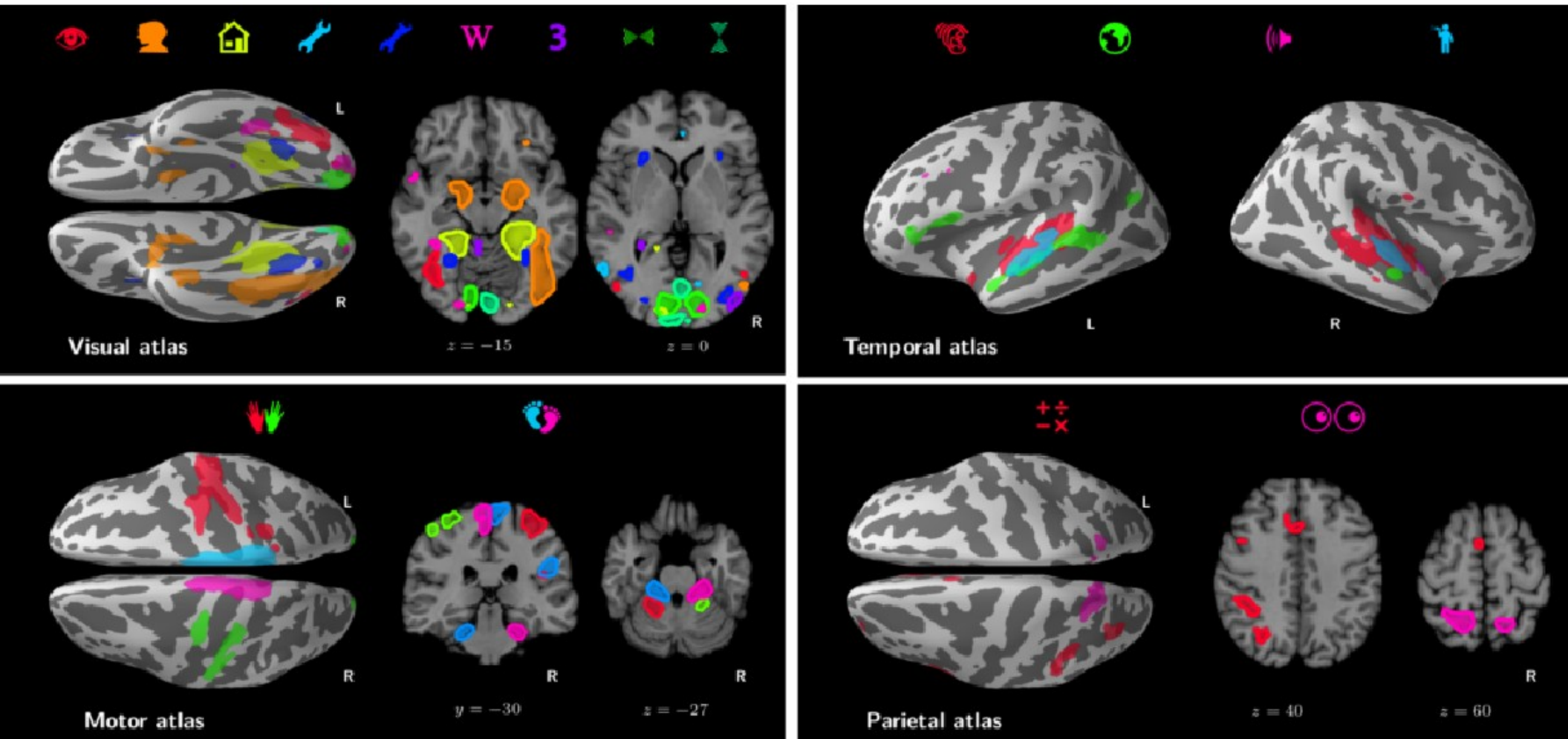
**Target**

# Classification results



[Schwartz et al. NIPS 2013, Varoquaux et al. PCB 2018]

# Discriminative patterns



[Schwartz et al. NIPS 2013, Varoquaux et al. PCB 2018.]

# An image database



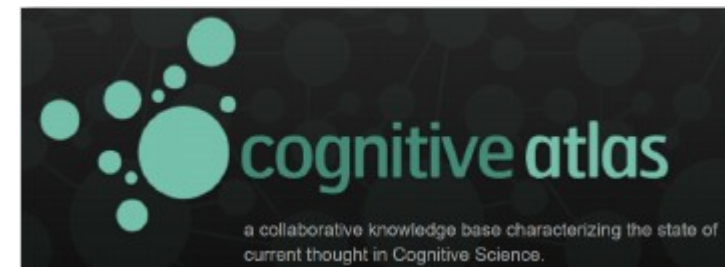
**NEURO**VAULT

Task fMRI repository  
[Gorgolewski et al. 2015]

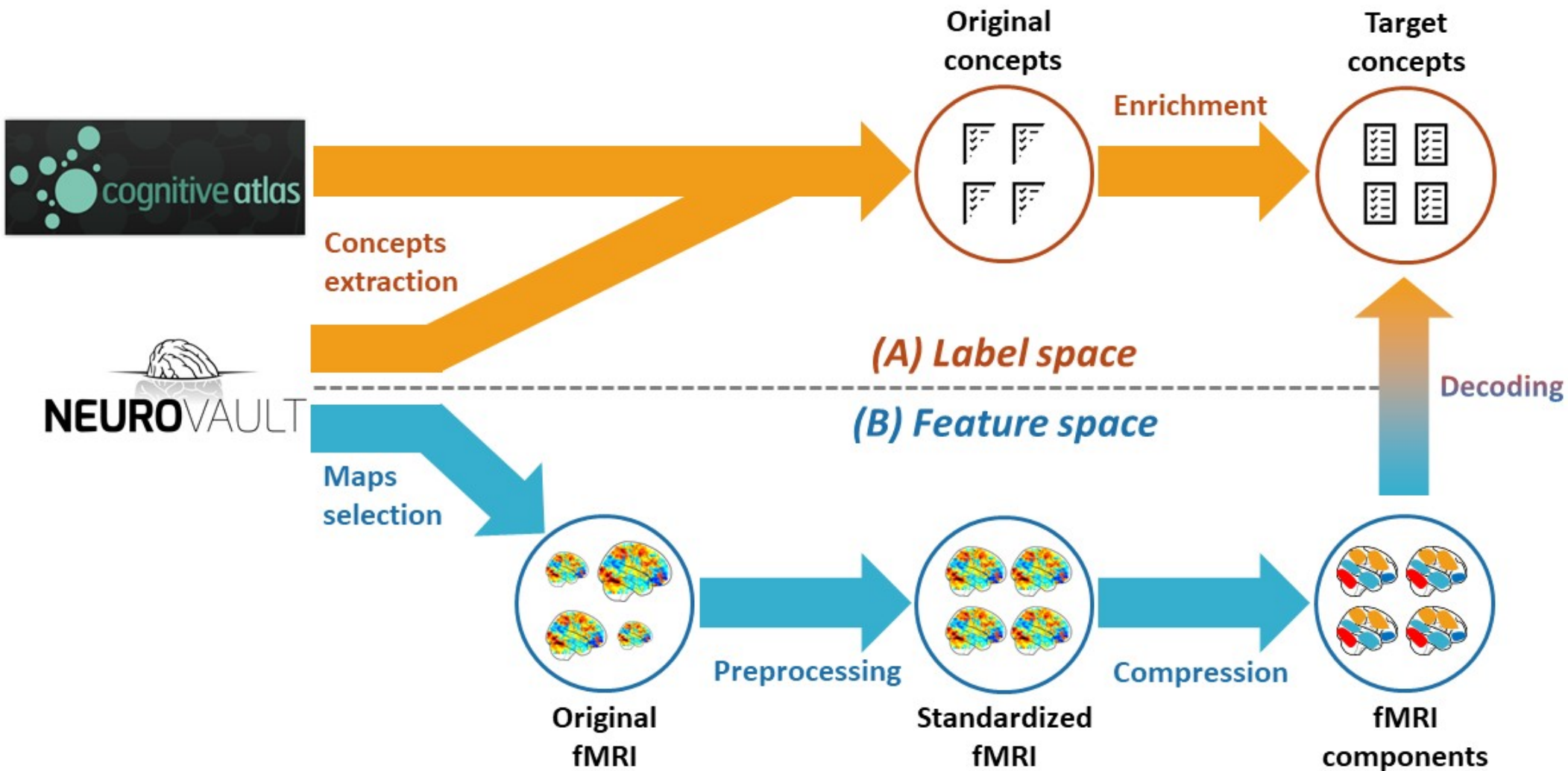
Currently 48k independent  
usable fMRIs

[Poldrack 2011], knowledge-base

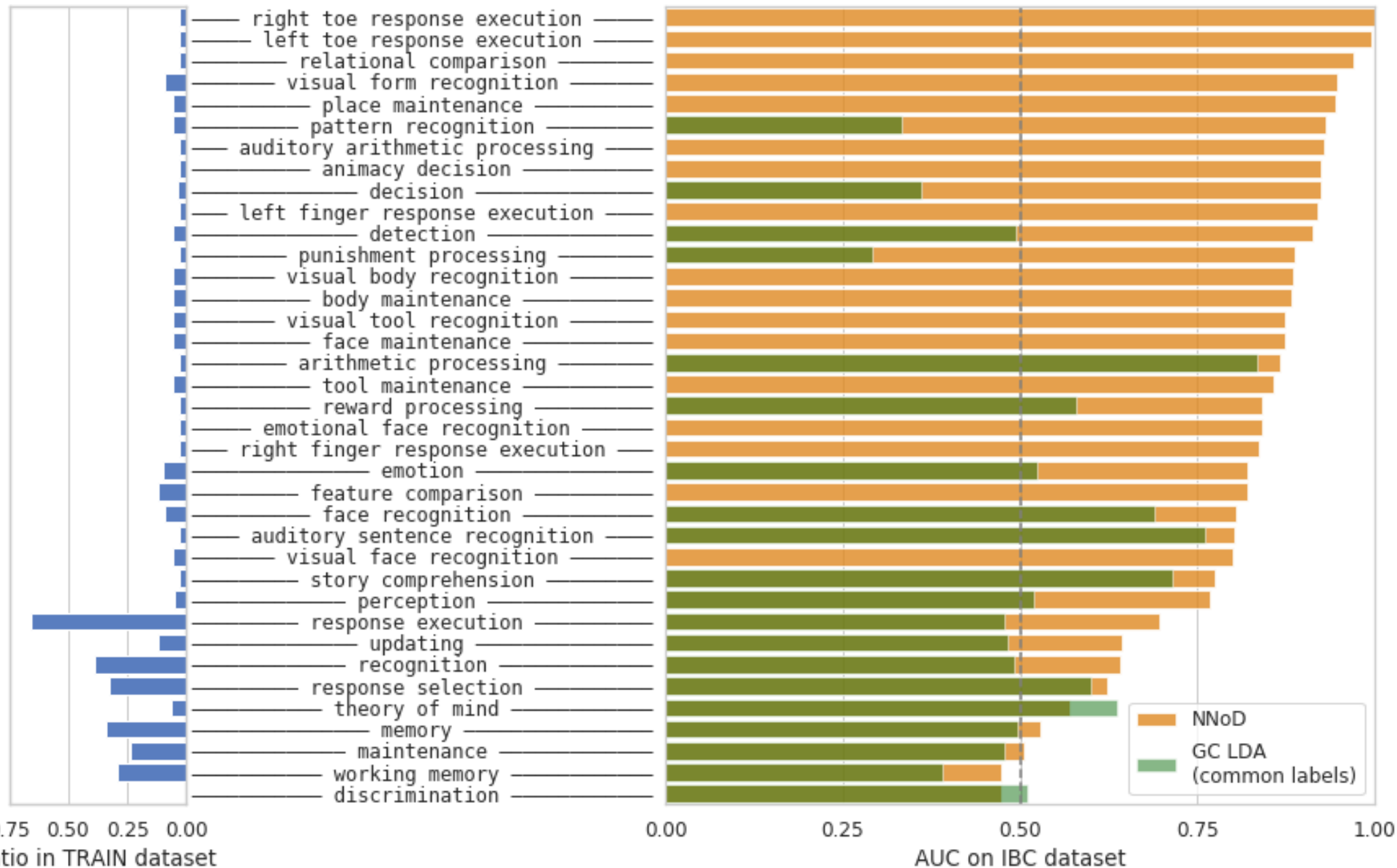
- *concepts*: cognitive activity/state (e.g. working memory)
- *tasks*: standard experiment to probe it (e.g. n-back task)



# From multi-study to universal decoder



# Results (naive approach)



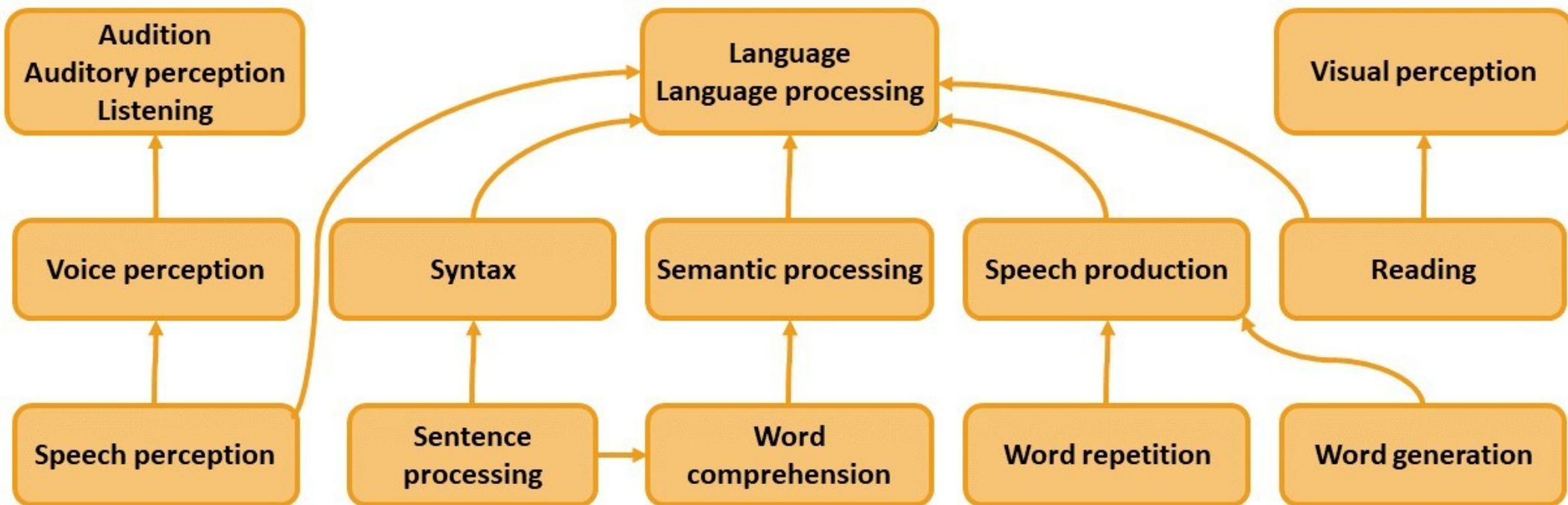


# Fixing labels

## Problem:

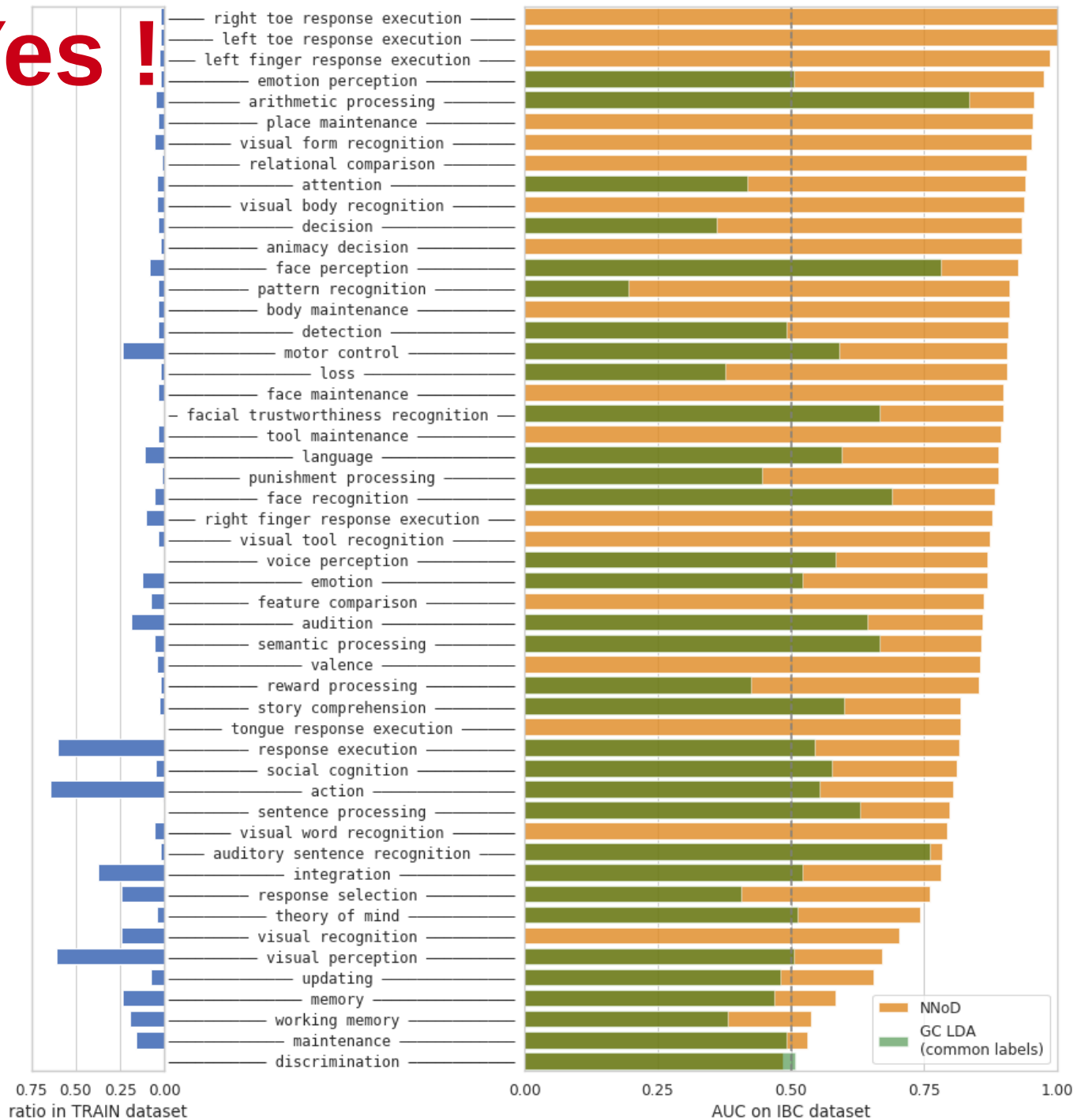
synonyms, false negatives (missing annotations)

→ **Simple rules to impute labels:**

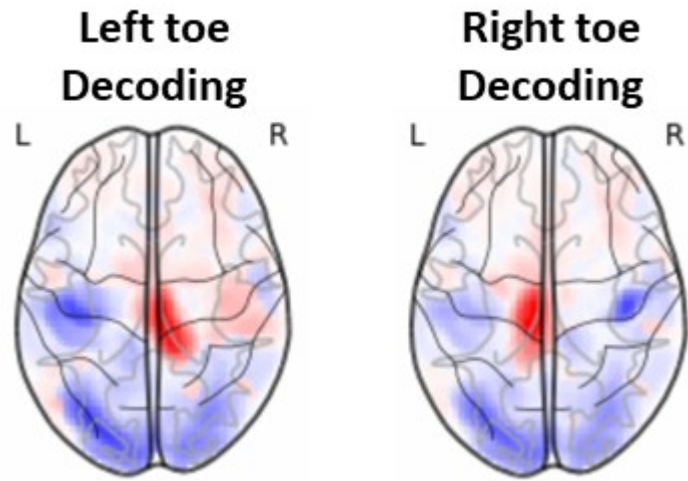




# Results (2): Yes !

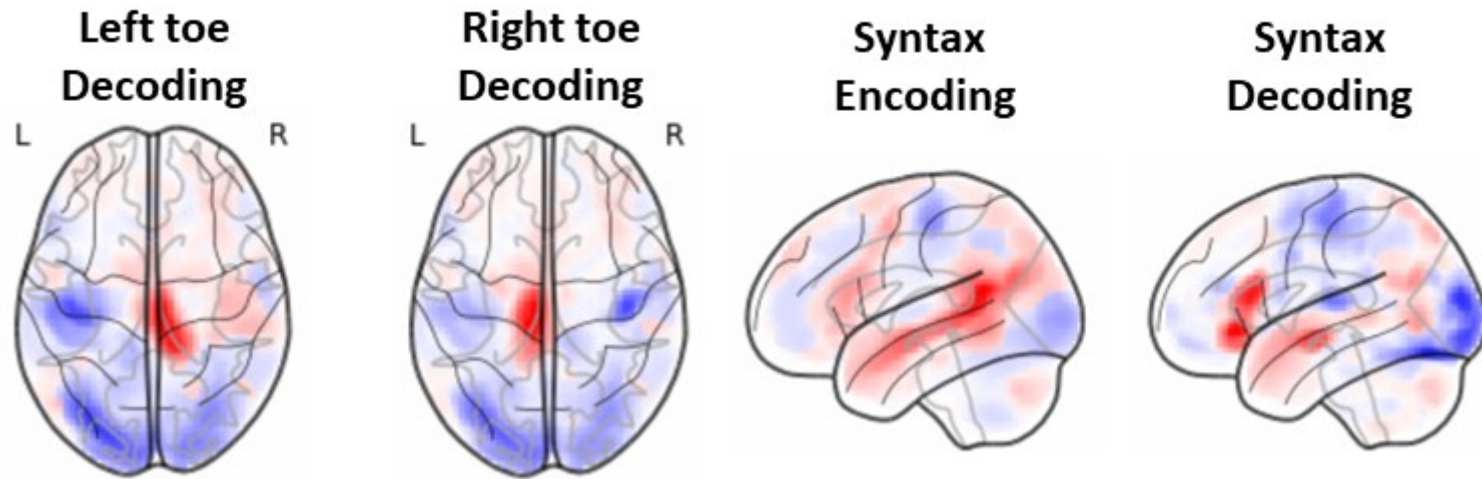


# Open the box



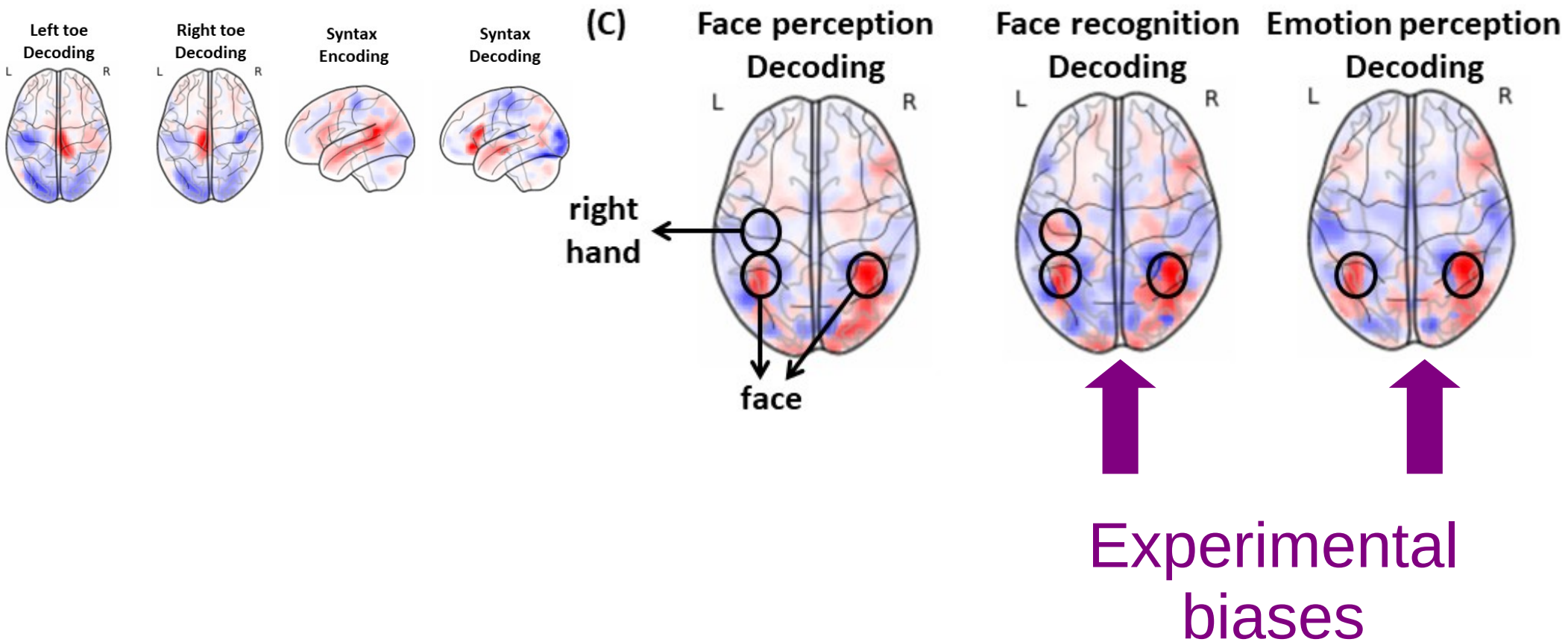
Non-controversial case

# Open the box



decoding > encoding

# Open the box



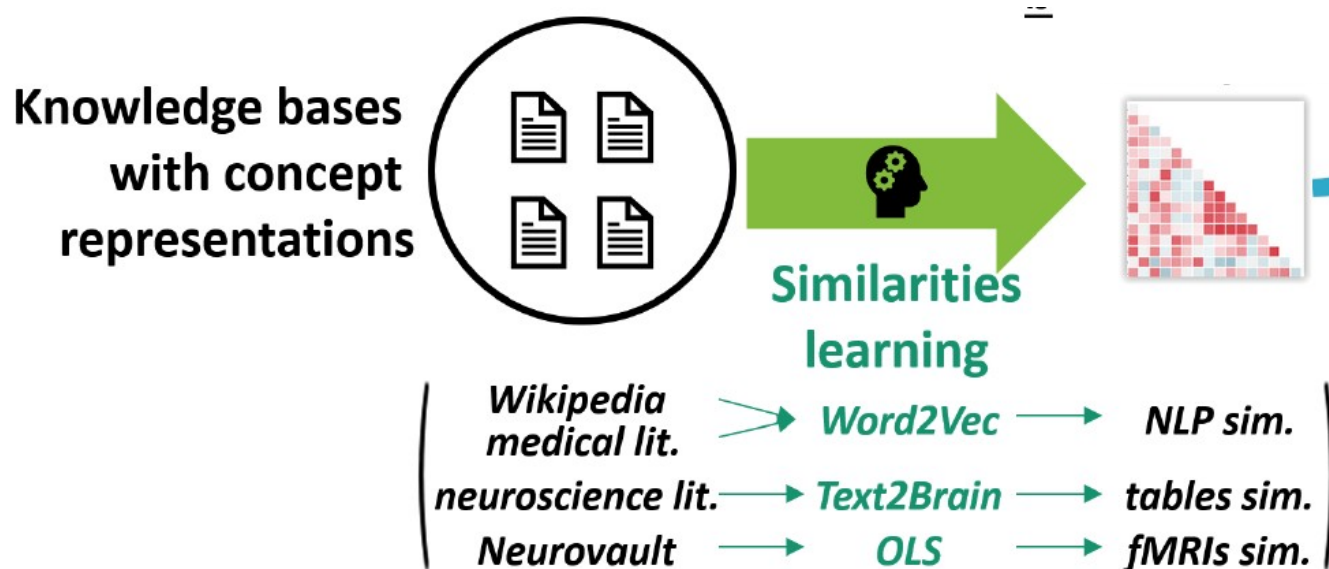
[Menuet et al. Scientific Reports 2022]

# Dealing with semantics and data labelling issues

... by mining the neuroscientific literature

# Need curated annotations

- Current ontology incomplete
  - Bigger limitation = lack of consistent vocabulary
- [Poldrack & Yarkoni, Annu Rev Psycho 2016]
- How to get those ?



# Mining neuroimaging literature

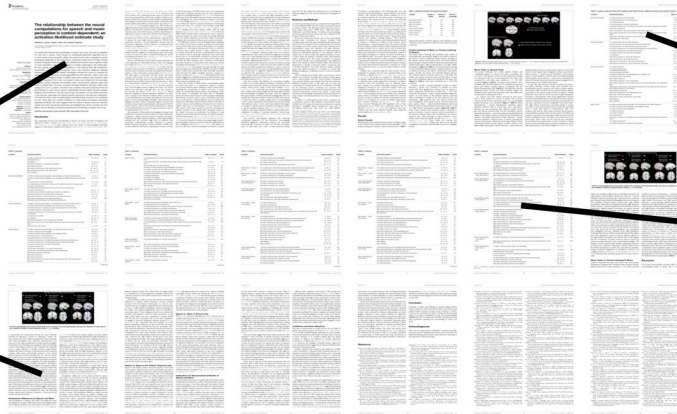
- Neuroimaging observations often stored in text.
- e.g “[...] in the anterolateral temporal cortex, especially the temporal pole and inferior and middle temporal gyri”
- Objectives:
  - transform neuroimaging publications into brain maps
  - meta-analysis of text-only corpora

# Neuroquery

## Extract

Are visual texture-selective areas recruited during haptic texture discrimination?

Shape and texture provide cues to object identity, both when objects are explored using vision and via touch (haptics). Visual shape information is processed within the lateral occipital complex (LOC), while texture is processed in medial regions of the collateral sulcus (CoS). Evidence indicates that the LOC is recruited during both visual and haptic shape processing. Here we used functional magnetic resonance imaging (fMRI) to examine whether 'visual' texture-selective areas are similarly recruited when observers discriminate texture via touch. We used a blocked design in which participants discriminated either the texture or shape of unfamiliar 3-dimensional (3D) objects, via vision or touch. We observed significant haptic texture-selective fMRI responses in medial occipitotemporal cortex within areas adjacent to, but not overlapping, those recruited during visual texture discrimination. Although areas of ventromedial temporal cortex are recruited during visual and haptic texture perception, these areas appear to be spatially distinct and modality-specific.



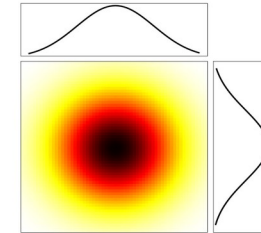
x	y	z
-34.0	-59.0	-9.0
-31.0	-78.0	-9.0
-17.0	-68.0	2.0
-33.0	13.0	15.0
25.0	-53.0	-9.0
24.0	-73.0	-9.0
5.0	19.0	37.0
37.0	19.0	4.0
27.0	41.0	32.0
-56.0	-43.0	26.0
-51.0	-56.0	17.0
-41.0	-72.0	14.0
-58.0	-23.0	22.0
-12.0	-88.0	17.0
-43.0	-56.0	6.0
-50.0	-67.0	2.0
52.0	-31.0	22.0
48.0	-67.0	2.0
47.0	-38.0	44.0

## Transform

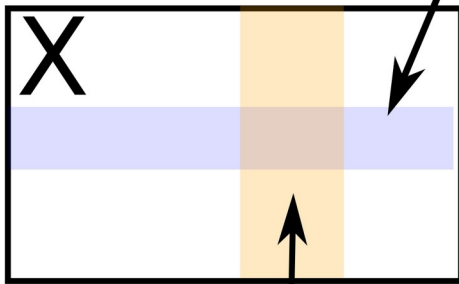


$X_i$

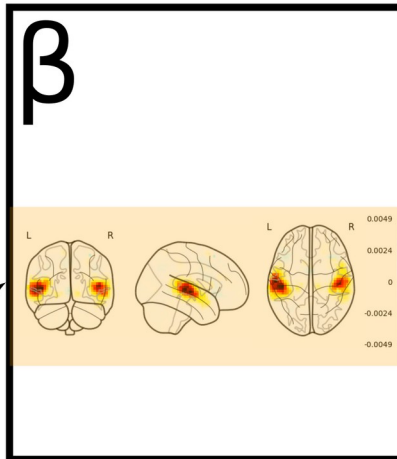
$y_i$



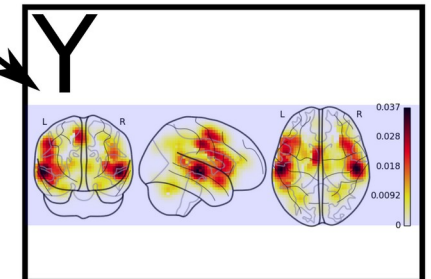
## Fit



"primary auditory cortex"

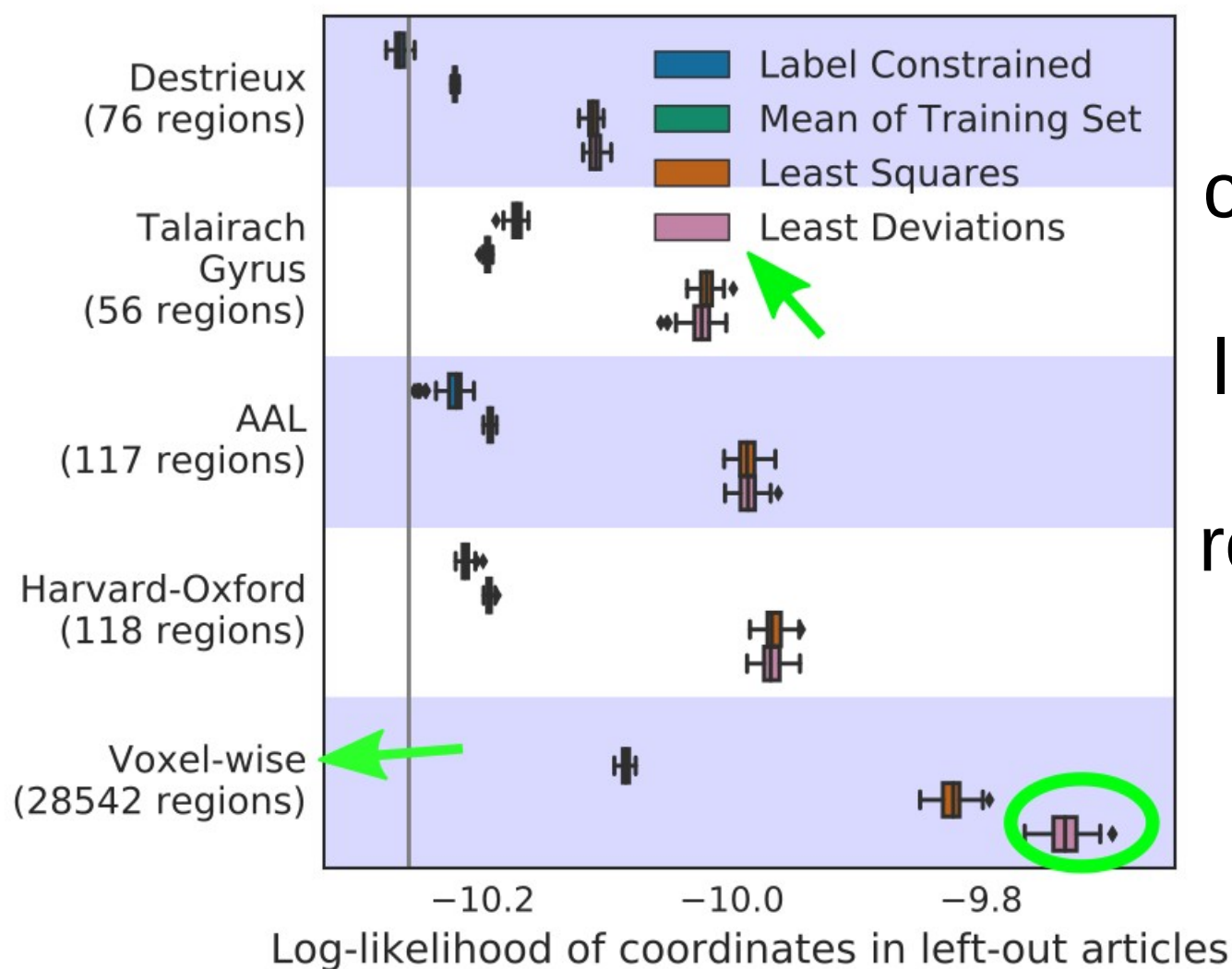


+ E =





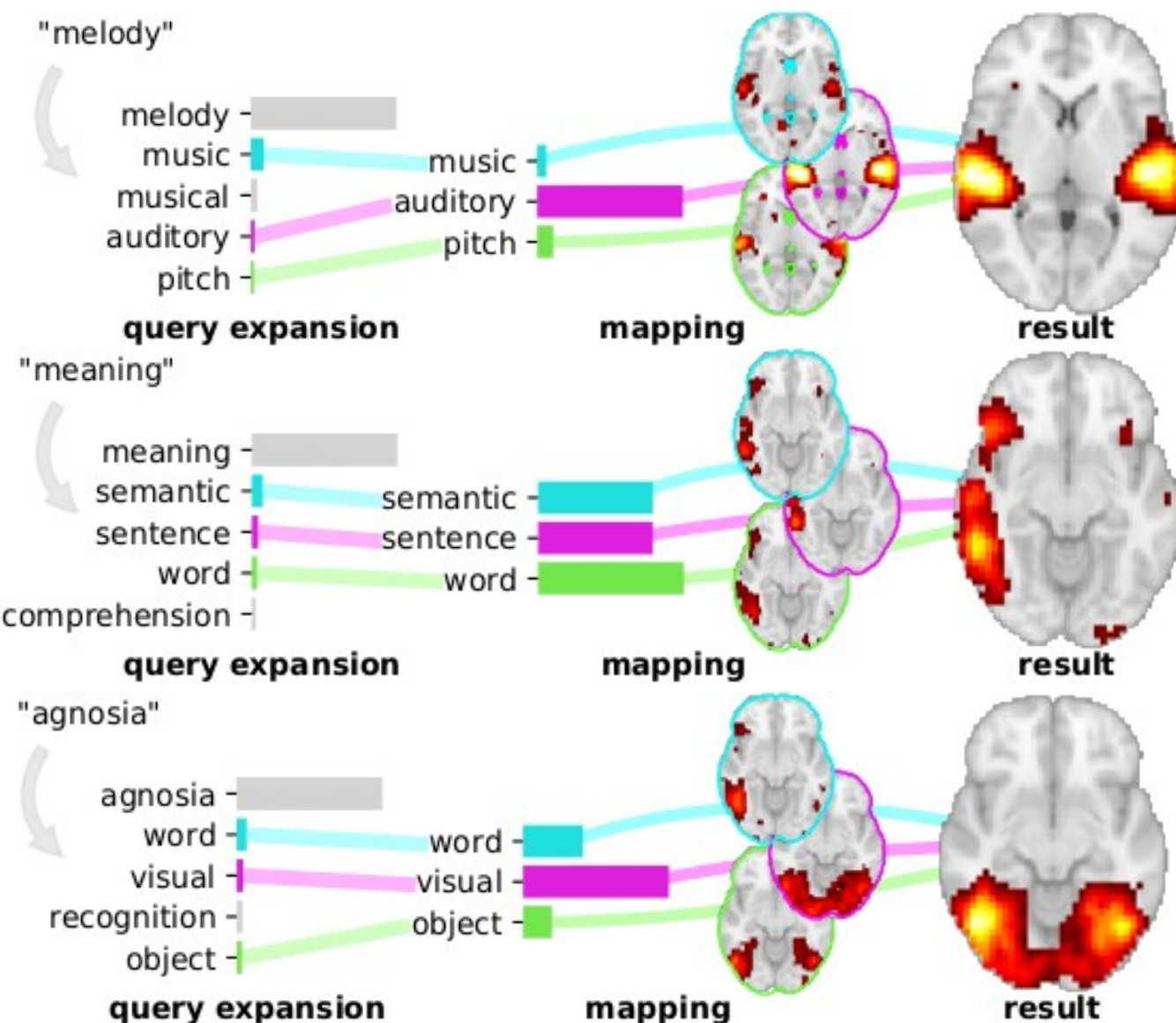
# Empirical evaluation of representations



Learning statistical correspondences across the literature is more effective than relying on atlases !

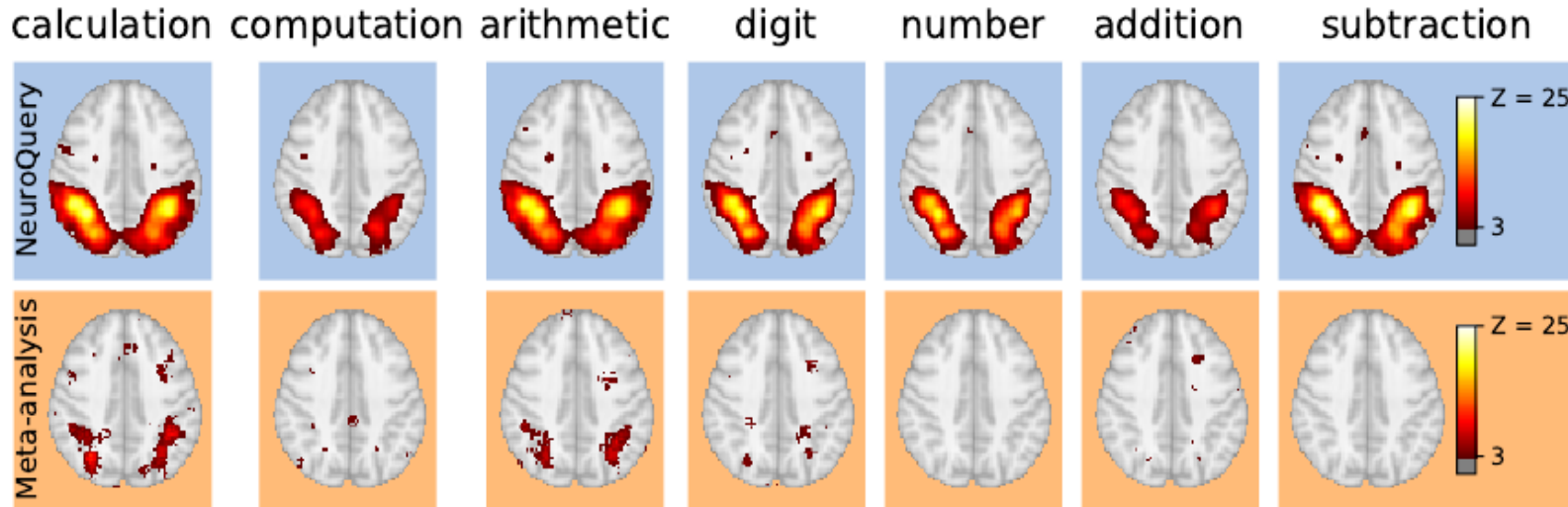
[Dockès et al. MICCAI 2018]

# Leveraging semantics for better encoding



Semantic structure  
→ map concepts  
with few/no data

# Neuroquery



<https://neuroquery.org>

[Dockès et al. *elife* 2020]



# Conclusion

- Large-p data bring challenges:
  - Computation cost
  - Difficulty of statistical inference
- Solutions: compression, subsampling, ensembling
- Finding **commonalities** across cognitive studies is hard
- Big data approach:
  - Extract **weak signals** from huge amounts of data
  - Common representation across datasets (*bottleneck*)

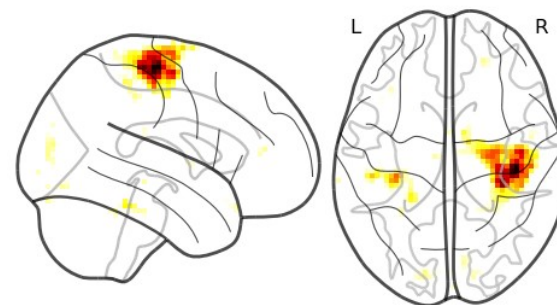
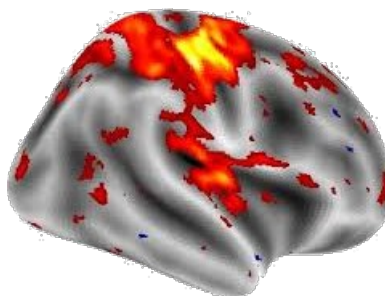
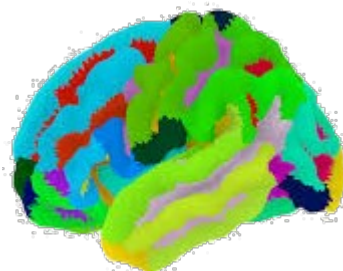
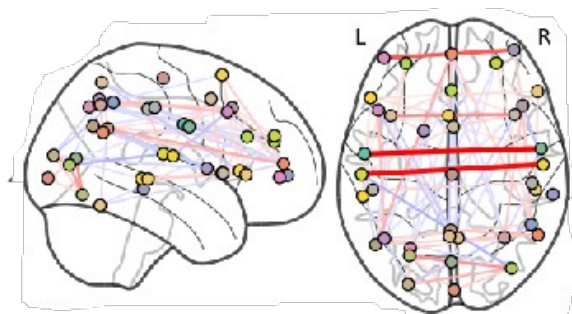


Image processing  
may not be the hard  
part !

# From good ideas to good practices: software



- Machine learning in Python
- Machine learning for neuroimaging  
<http://nilearn.github.io>
- BSD, Python, OSS
  - Classification of (neuroimaging) data
  - Network analysis





# Parietal/Mind

**G. Varoquaux,**  
A. Gramfort,  
P. Ciuciu,  
D. Wassermann,  
D. Engemann,  
A. Manoel,  
D. Chyzyhyk  
A.L. Grilo Pinho,  
E. Dohmatob,  
**A. Mensch,**  
J.A. Chevalier,  
A. Hoyos idrobo,  
D. Bzdok,  
**J. Dockès,**  
K.Dadi,  
P. Cerda,  
C. Lazarus  
D. La Rocca  
G. Lemaitre  
L. El Gueddari  
O. Grisel  
M. Massias  
P. Ablin  
H. Janati  
J. Massich  
C. Petitot  
JJ Torres



J. Abécassis,  
A. Chamma,  
R.Meudec,  
N.Gensollen,  
A.Pasquiou,  
A.Thual,  
T. Bazeille,  
B. Nguyen,  
T.Chapalain,  
H.Cherkaoui,  
H. Aggarwal,  
S. Shankar  
A.Blain



Human Brain Project

# Acknowledgements

## Other collaborators

J. Salmon  
S. Arlot  
M. Lerasle  
P. Neuvial  
( & thanks for the data):  
S. Dehaene  
R. Poldrack,  
J. Haxby  
C. F. Gorgolevski  
T.Yarkoni  
G. Gautier



université  
PARIS-SACLAY

AGENCE NATIONALE DE LA RECHERCHE  
ANR