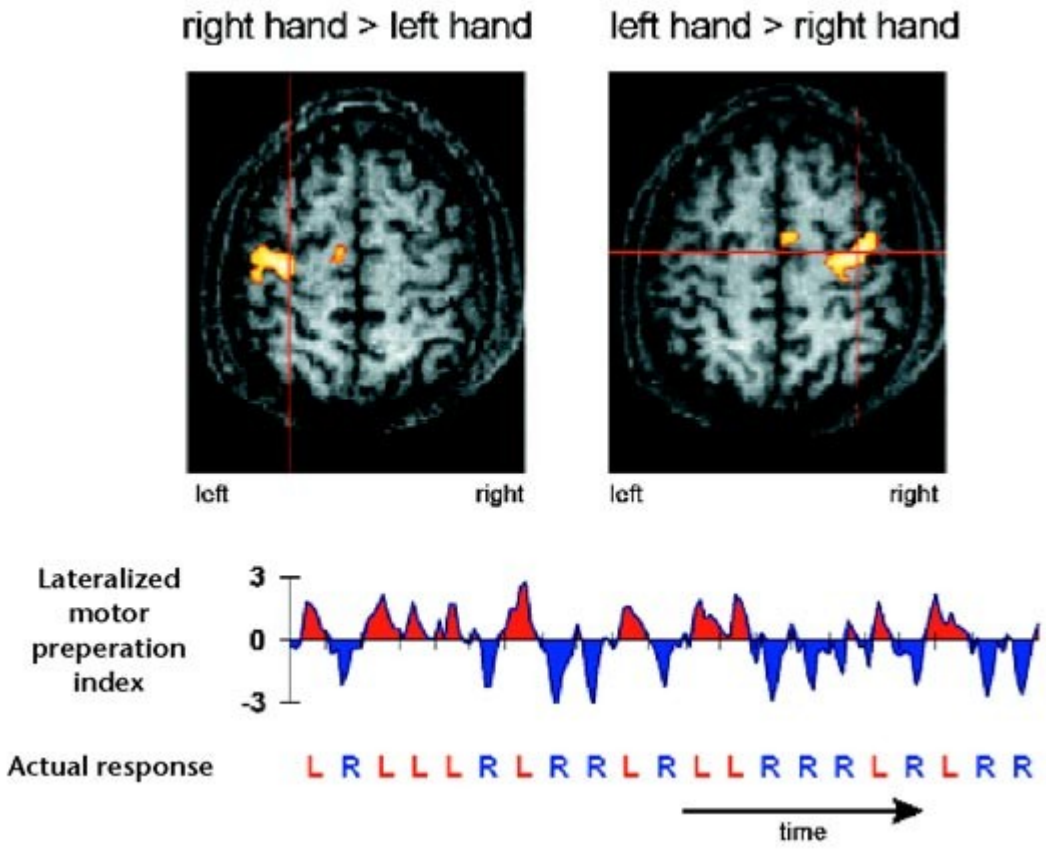
Classification and machine learning techniques for fMRI data analysis

Bertrand Thirion
INRIA Saclay–Ile de France, PARIETAL team, Neurospin
Bertrand.thirion@inria.fr

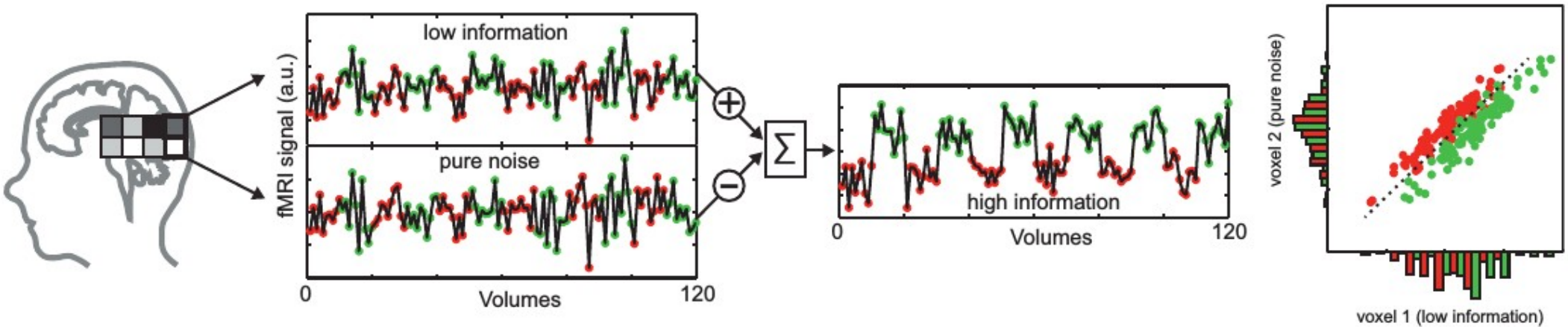
Inverse Inference and Brain reading

- Inverse inference consists in inferring something about the subject or the subject's activity/state from neuroimaging data

"Inferring behavior from functional brain images", Deahene et al., nat. neurosc. 1998



Idealized case for brain decoding



- Individual voxels corrupted by a noise source → weakly significant
- Their difference is strongly task related: accurate classification

  [Haufe et al. nimg 2013, Haynes neuron 2015]

Outline

- Identification issues
- Large-scale decoding
- Experiments on vision

Multivariate analysis and inference

Model: $\mathbf{y} = f(\mathbf{X}) + \varepsilon$

Aka predictive model, fit a target **y** with some explanatory variables **X**
p = # features
n = # samples
(Easy) question: minimize prediction error

Multivariate analysis and inference

Model: $\mathbf{y} = f(\mathbf{X}) + \varepsilon$

Hard question: what variables in **X** are important to predict **y** ? aka explanation, interpretation

Problems:

- whenever p>n, *any explanation* is possible (linear case)
- point estimates may simply reflect our prior (e.g. Ridge vs Lasso)

Explanations can be arbitrary, and come without guarantees

Multivariate analysis and inference

- What we want: **statistically valid** account of variables importance
  - p-value
  - confidence intervals
- How ?
  - Classical theory breaks in high dimension
  - A bunch of solutions have been proposed recently for $n \approx p$
  - $n \ll p$ case not addressed in the literature

Error rates and confidence intervals

- False positive:
  - Declare that $\mathbf{X}_i$ plays a role while it does not
- False positive rate $\frac{|i|:\mathbf{X}_i \text{ wrongly reported}}{|i|:\mathbf{X}_i \text{ not important}}$
- False discovery prop $\frac{|i|:\mathbf{X}_i \text{ wrongly reported}}{|i|:\mathbf{X}_i \text{ reported}}$
- Confidence intervals

$$\mathbf{w}_i \in [\hat{\mathbf{w}}_i - 2\sigma_i, \hat{\mathbf{w}}_i + 2\sigma_i]$$

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon,\ \sigma_i = std(\mathbf{w}_i)$$

Inference with linear models

- **n > p** → OLS

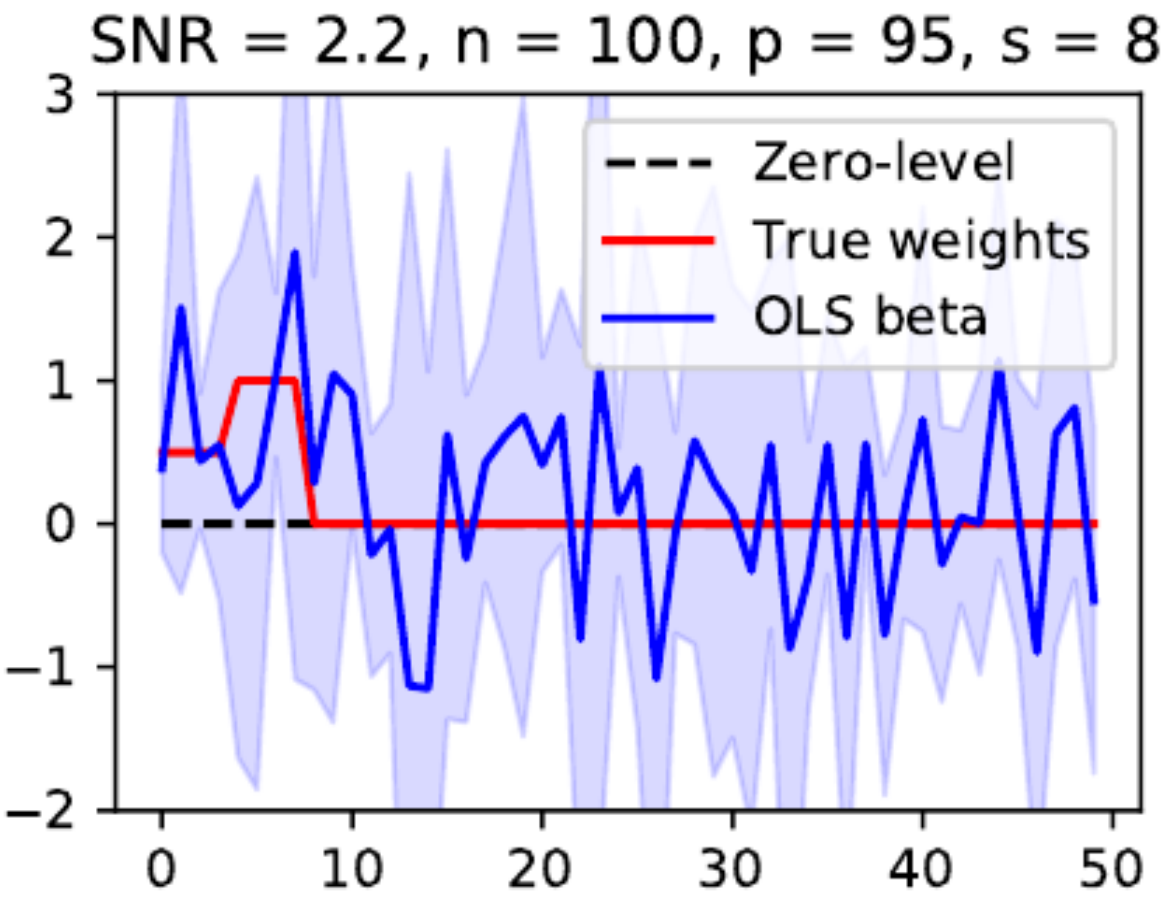
$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathrm{Cov}}(\mathbf{w}) = (\mathbf{X}^T\mathbf{X})^{-1}\frac{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2}{n}$$

$$t = \frac{\hat{\mathbf{w}}_i}{\sqrt{\hat{\mathrm{Cov}}(\mathbf{w})_{ii}}}$$

- As long as X is well-conditioned
- Multiple testing problem: non-independent stats

Classical inference breaks for p≅n

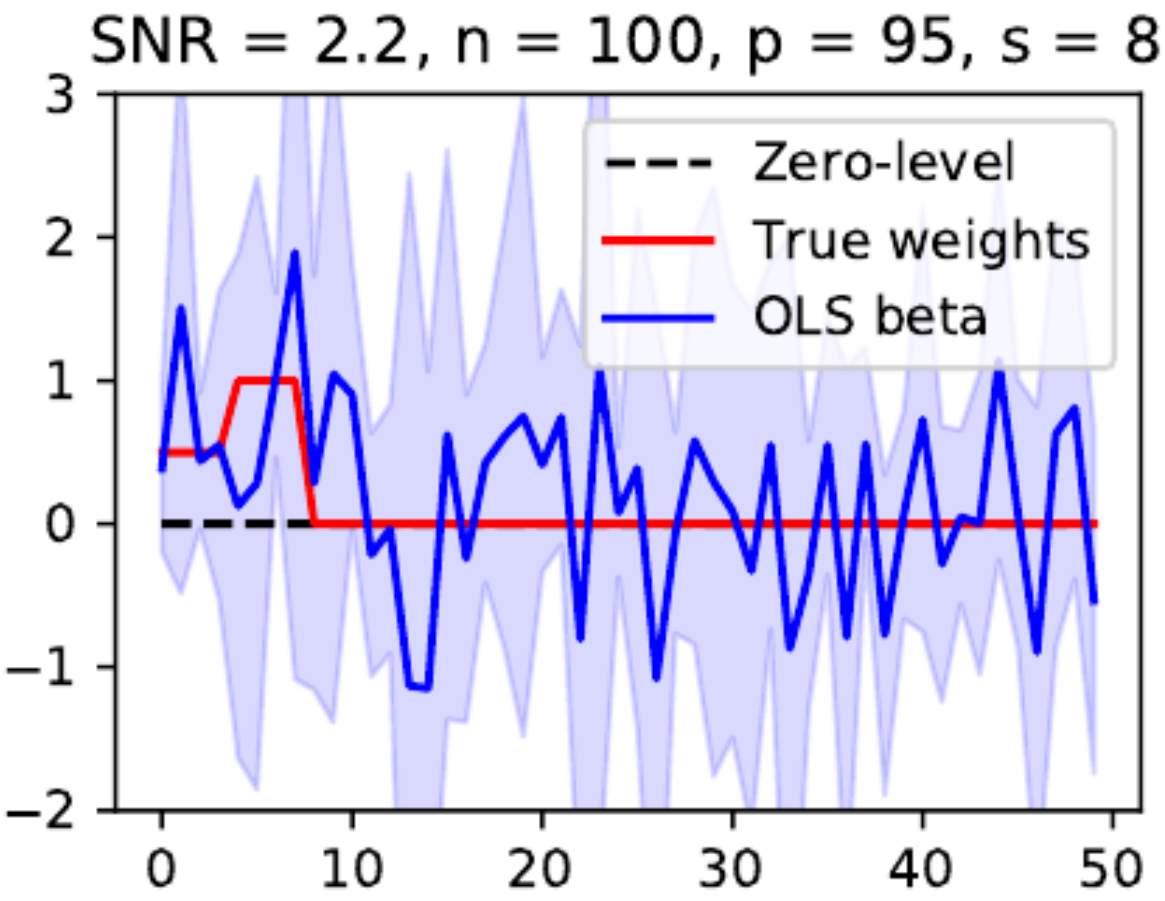


OLS regression when $p \approx n$

- Variance explosion,

  p>n non-defined solution

Classical inference breaks for p≅n



OLS regression when $p \approx n$

- Variance explosion,

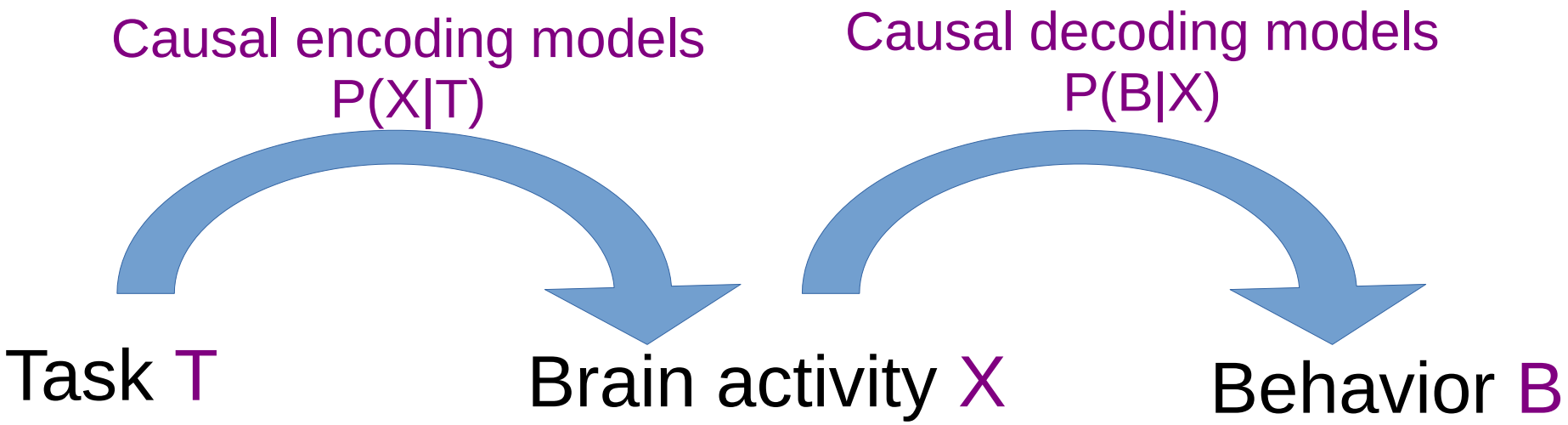
  p>n non-defined solution
- Estimator classically needs regularization (Ridge, Lasso)
- Harder statistical inference

Curse of dimensionality

- Ridge / Lasso estimators need a correction [Buhlmann Bernoulli 2013]
- Confidence intervals for Ridge, not Lasso
- Bootstrap Lasso fails [Bach NIPS 2008]
- Stability selection too conservative and costly

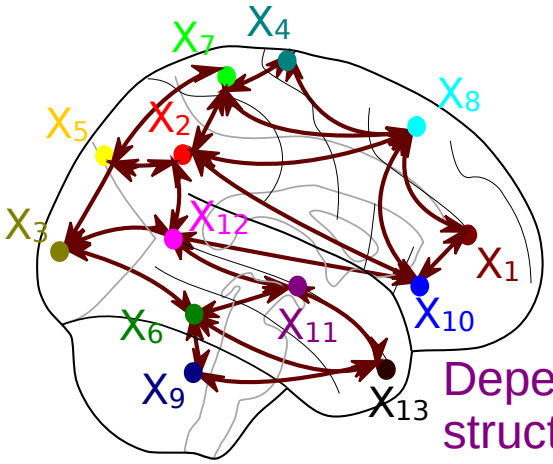
  [Meinshausen & Buhlmann, RSS 2010]
- ...

Causal reasoning on encoding/decoding

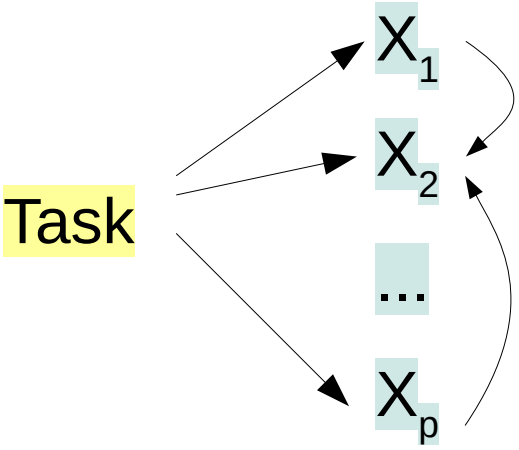


[Weichwald et al Nimg 2015]

Impact for statistical inference

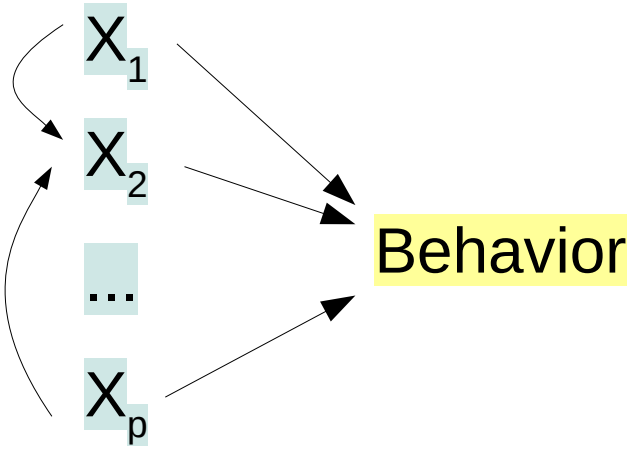


Dependence structure among brain regions



Encoding

$$\mathbf{X_i} \perp\!\!\!\perp \mathbf{T}$$

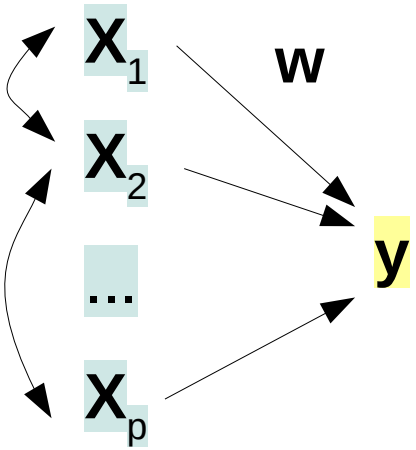


Decoding

$$\mathbf{X_i} \perp\!\!\!\perp \mathbf{B} | \left(\mathbf{X_j}, \mathbf{j} \neq \mathbf{i}\right)$$

[Weichwald et al. Nimg 2015]

Brain activity decoding

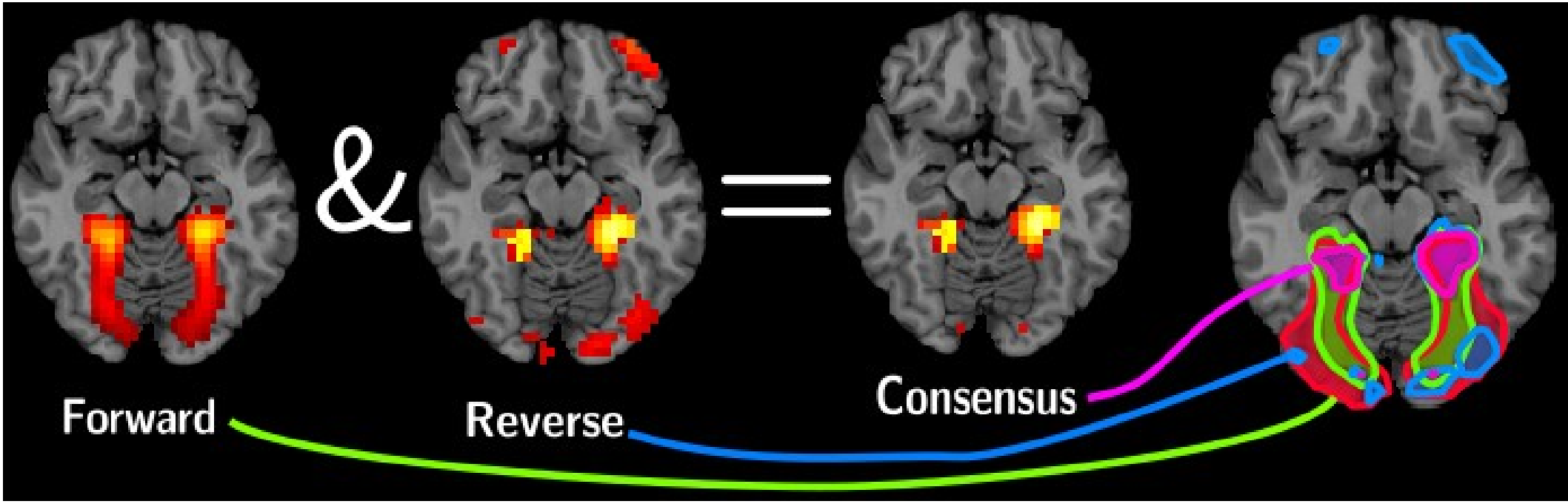


- behavior = f (brain activity)

$$\mathbf{y} = \mathbf{X}\boldsymbol{w}^* + \sigma_* \boldsymbol{\varepsilon}$$

- error vector: $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
- noise magnitude: $\sigma_* > 0$

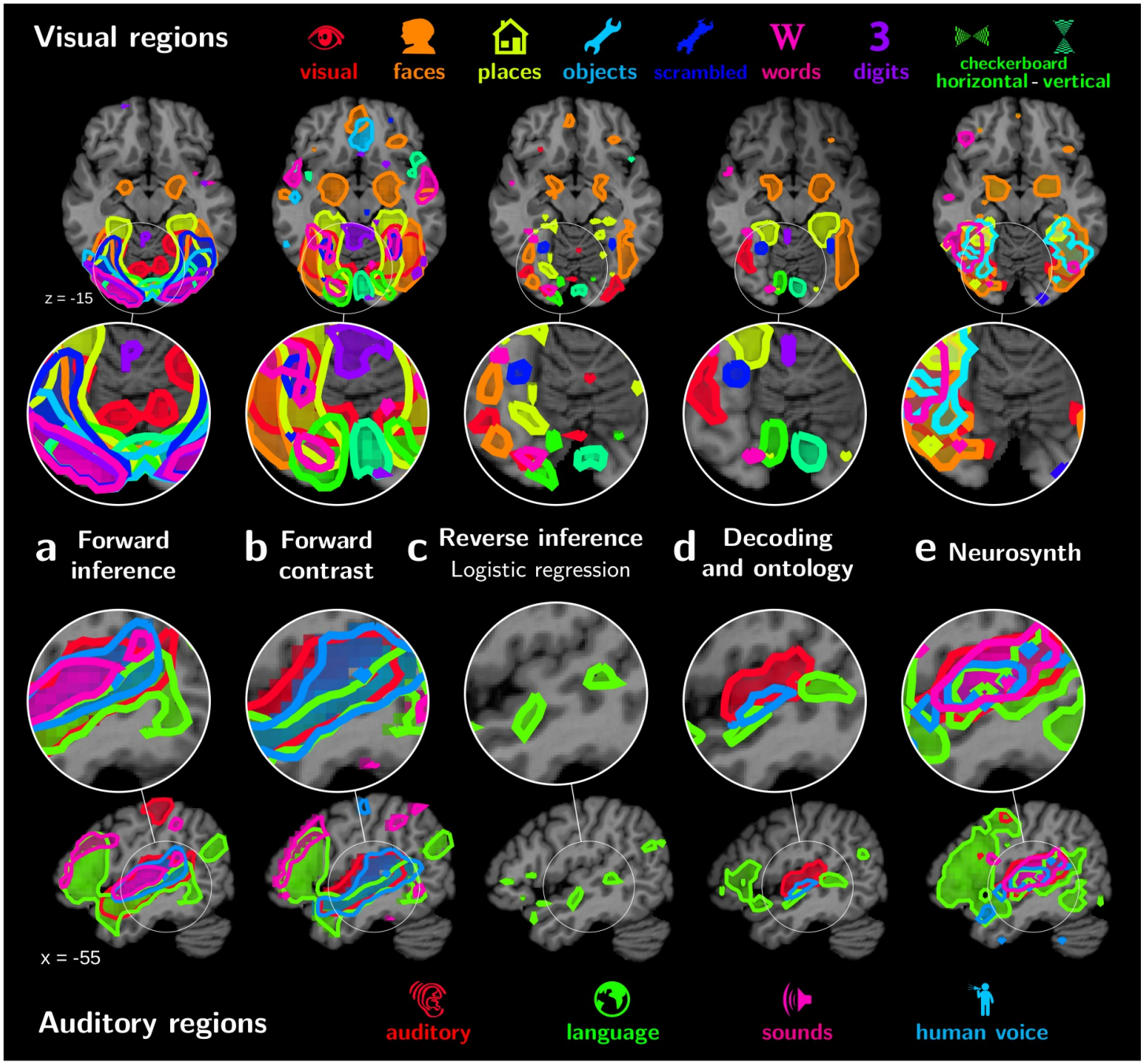
Joint encoding and decoding



“Encoding” “Decoding”

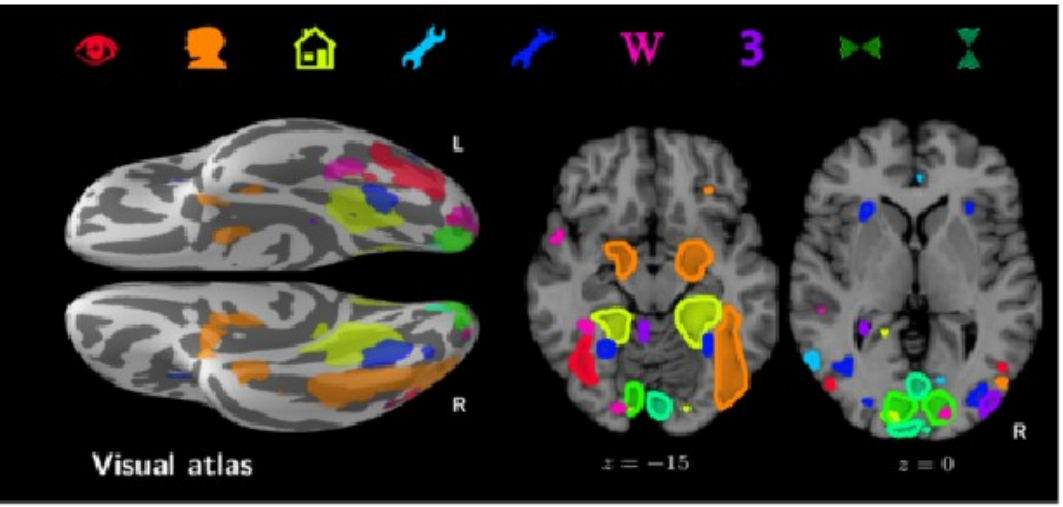
[Schwartz et al. NIPS 2013, Varoquaux et al. PCB 2018]

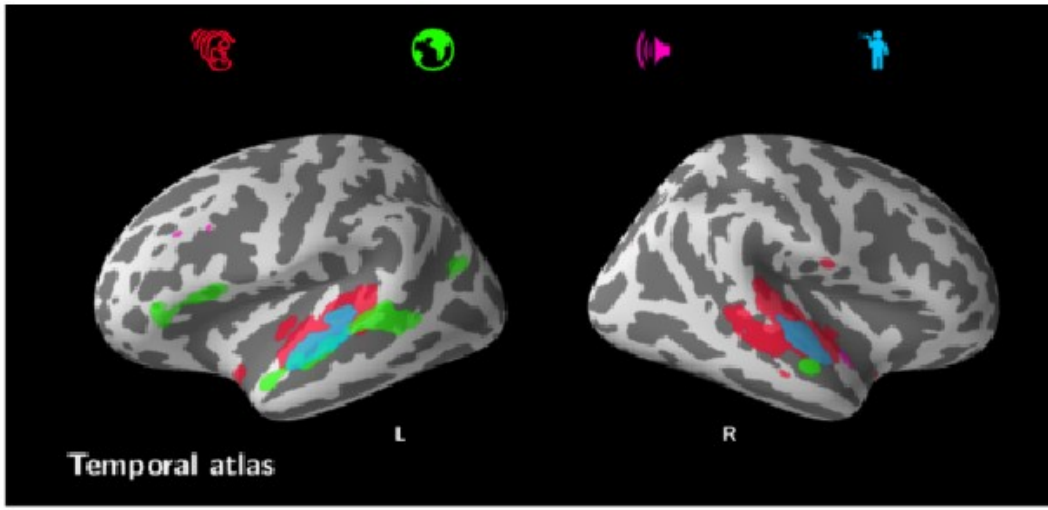
Decoding maps

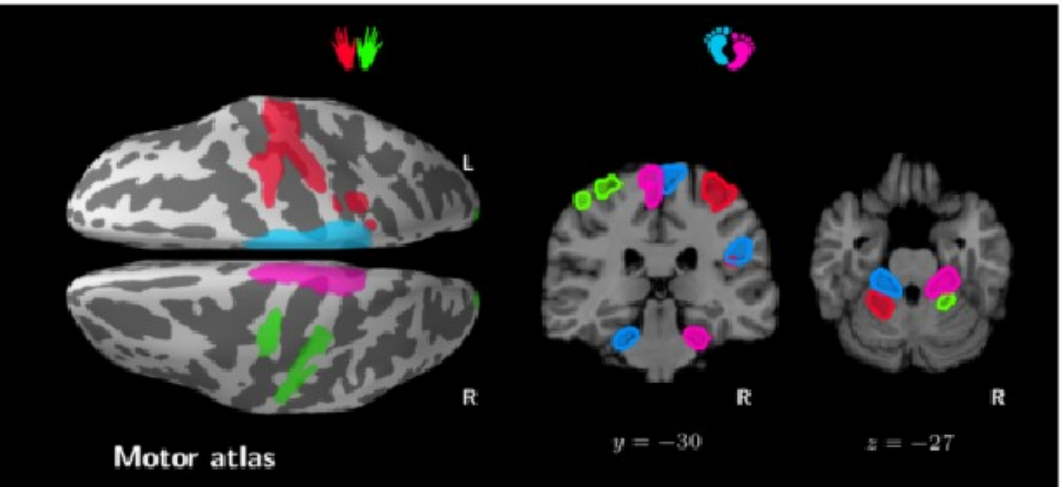


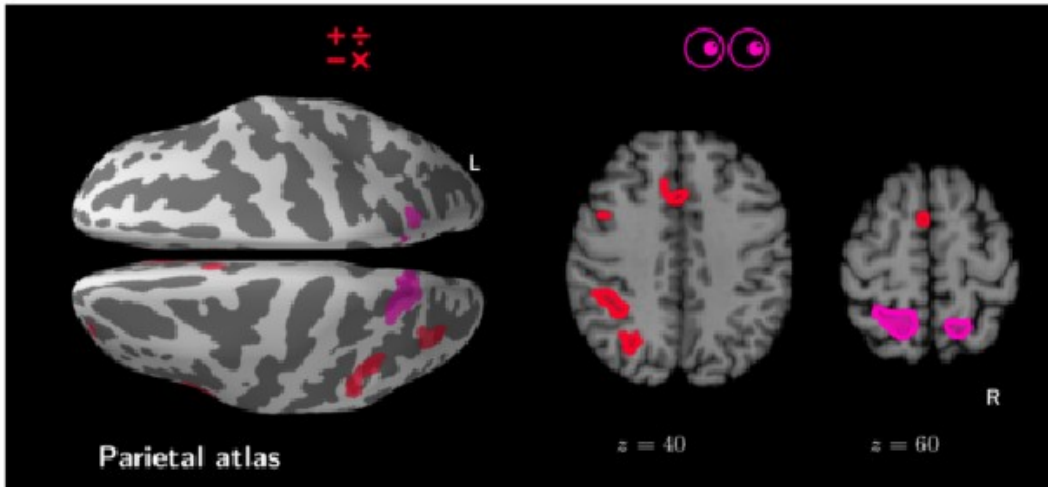
[Varoquaux et al. Plos Comp Bio 2018]

Joint encoding and decoding



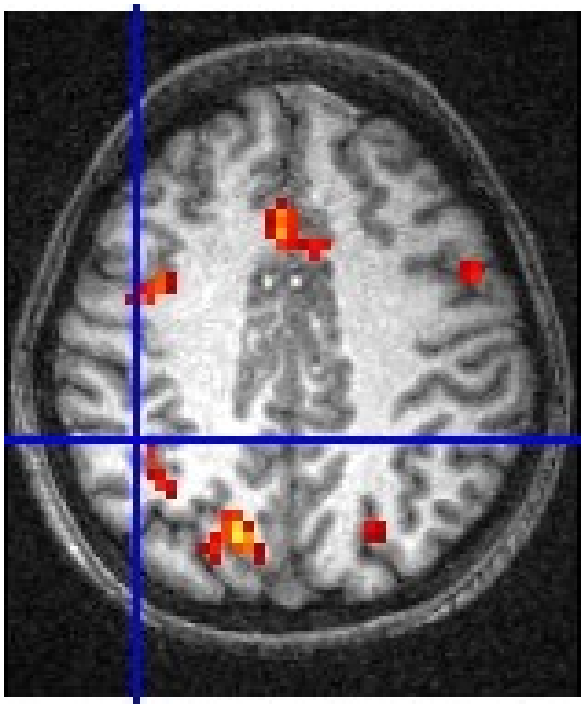


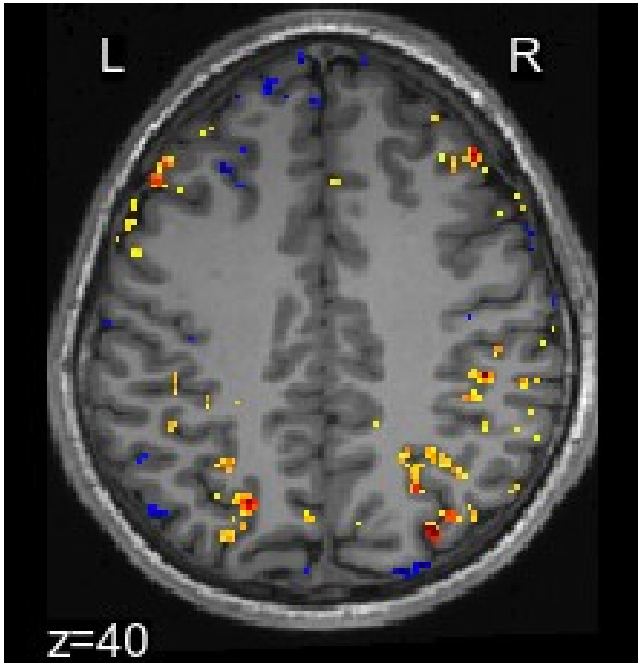




[Schwartz et al. NIPS 2013, Varoquaux et al. PCB 2018]

Resolution increases





2007:
3 mm

2014:
1.5 mm

2022:
0.5 mm ?

p = 50,000

p = 400,000

p = 10^7

Still working with linear models ?

- Brain activity decoding = small n, large p
- Low SNR
- No translation equivariance
- Linear (possibly multi-layer) models still outperform neural nets [He et al. Nimg 2019, Mensch et al. Plos Comp Biol 2021]

Statistical inference on w

- Inference: find $\{j: \mathbf{w}_j > 0\}$ with statistical guarantees (p-values / fdr control)

- Standard solutions for high-dimensional linear models ($p \cong n$)
  - knockoffs [Candès 2015+]
  - Desparsified Lasso [Zhang & Zhang 2014, Montanari 2014]
  - Conditional randomization tests [Candès 2018]
  - Multi-split [Meinshausen 2009], Corrected ridge [Bühlmann 2013]

Understanding desparsified Lasso

- **Ordinary Least Squares**:
  - when $n > p$ take $\mathbf{z}_j = $ OLS residual of $\mathbf{X}_{\cdot,j}$ against $\mathbf{X}^{(-j)}$
  - notice that $\frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} = \mathbf{w}_j^* + \frac{\mathbf{z}_j^\top \varepsilon}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}$ since $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$ if $j \neq k$
  - dismiss noise term:

$$\hat{\mathbf{w}}_j^{\mathrm{OLS}} = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}$$

- **Property**: denoting $\hat{\Sigma} = \mathbf{X}^\top \mathbf{X}$

$$\sigma_\varepsilon^{-1}(\hat{\Sigma}_{jj})^{-1/2}(\hat{\mathbf{w}}_j^{\mathrm{OLS}} - \mathbf{w}_j^*) \sim \mathcal{N}(0,1)$$

Understanding desparsified Lasso

- **Desparsified Lasso** [Zhang and Zhang, 2014]:
  - when $n < p$ take $\mathbf{z}_j$ = Lasso residual of $\mathbf{X}_{\cdot,j}$ against $\mathbf{X}^{(-j)}$
  - notice that $$\frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} = \mathbf{w}_j^* + \frac{\mathbf{z}_j^\top \varepsilon}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} + \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \mathbf{w}_k^*}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}$$
  - dismiss noise term and plug Lasso estimator in bias term:

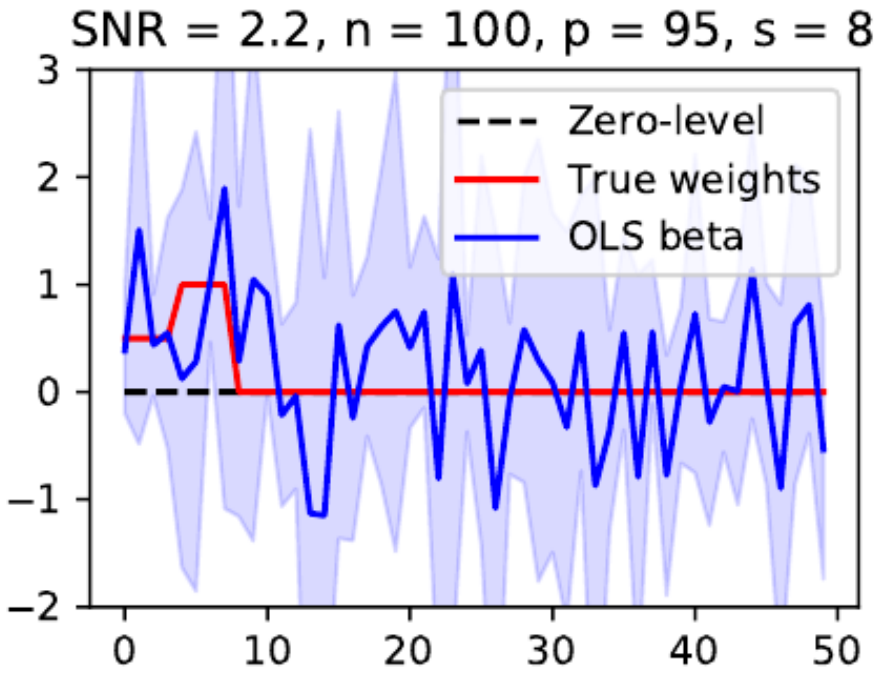
$$\hat{\mathbf{w}}_j^{\mathrm{DL}} = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \hat{\mathbf{w}}_k^{L(\lambda)}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}$$

- **Theorem**: asymptotically, under sparsity assumptions

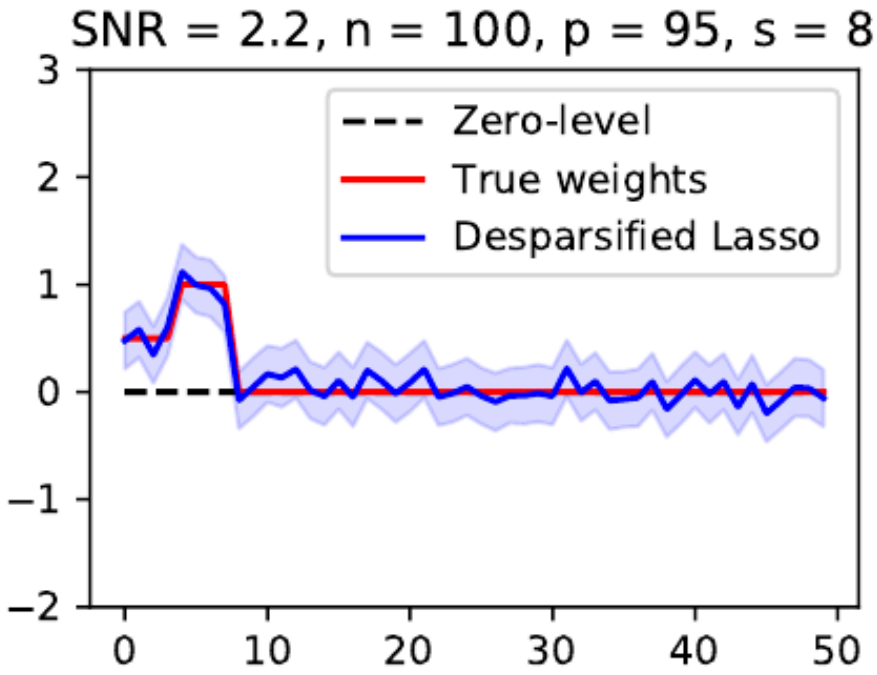
$$\sigma_\varepsilon^{-1} (\Omega_{jj})^{-1/2} (\hat{\mathbf{w}}_j^{\mathrm{DL}} - \mathbf{w}_j^*) \sim \mathcal{N}(0, 1)$$

Desparsified Lasso

- **Comparing OLS and Desparsified Lasso solutions:**



OLS regression when $p \approx n$



Desparsified Lasso when $p \approx n$

Desparsified logreg

Logistic Model $\mathbb{P}(y = 1 \mid \mathbf{x}_i) = \dfrac{1}{1 + \exp(-\mathbf{x}^T \boldsymbol{\beta}^0)}$

[Ning and Liu Annals stats 2017,
Van de Geer Annals stats2014]

Desparsified logreg

Logistic Model $\mathbb{P}(y = 1 \mid \mathbf{x}_i) = \dfrac{1}{1 + \exp(-\mathbf{x}^T \boldsymbol{\beta}^0)}$

Partial regressions

$$\hat{\boldsymbol{\beta}}^{d_{x_j}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right)}{\left[1 + \exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right)\right]^2} (x_{i,j} - \boldsymbol{\beta}^T \mathbf{X}_{-j})^2 + \lambda \|\boldsymbol{\beta}\|_1$$

$$\hat{\boldsymbol{\beta}}_j^{d_y} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p)$$

[Ning and Liu Annals stats 2017,
Van de Geer Annals stats2014]

Desparsified logreg

Logistic Model $\mathbb{P}(y = 1 \mid \mathbf{x}_i) = \dfrac{1}{1 + \exp(-\mathbf{x}^T \boldsymbol{\beta}^0)}$

Partial regressions

$$\hat{\boldsymbol{\beta}}^{d_{x_j}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right)}{\left[1 + \exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right)\right]^2} (x_{i,j} - \boldsymbol{\beta}^T \mathbf{X}_{-j})^2 + \lambda \|\boldsymbol{\beta}\|_1$$

$$\hat{\boldsymbol{\beta}}_j^{d_y} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p)$$

Decision statistic

$$T_j^{\text{decorr}} = -\frac{1}{\sqrt{n}} \hat{\mathbf{I}}_{j|-j}^{-1/2} \sum_{i=1}^{n} \left(y_i - \frac{1}{1 + \exp\left(-\mathbf{x}_{i,-j}^T \hat{\boldsymbol{\beta}}_j^{d_y}\right)} \right) \left(x_{i,j} - \mathbf{x}_{i,-j}^T \hat{\boldsymbol{\beta}}_j^{d_{x_j}}\right)$$

$$\hat{\mathbf{I}}_{j|-j}^{-1/2} = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\left(\hat{\boldsymbol{\beta}} \mathbf{x}_i\right)}{\left[1 + \exp\left(\hat{\boldsymbol{\beta}} \mathbf{x}_i\right)\right]^2} \left(x_{i,j} - \mathbf{x}_{i,-j}^T \, \hat{\boldsymbol{\beta}}^{d_{x_j}} \right)^2 \, x_{i,j}$$

[Ning and Liu Annals stats 2017,
Van de Geer Annals stats2014]

Desparsified logreg

Logistic Model $\mathbb{P}(y = 1 \mid \mathbf{x}_i) = \dfrac{1}{1 + \exp(-\mathbf{x}^T \boldsymbol{\beta}^0)}$

Partial regressions

$$\hat{\boldsymbol{\beta}}^{d_{x_j}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right)}{\left[1 + \exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i\right)\right]^2} (x_{i,j} - \boldsymbol{\beta}^T \mathbf{X}_{-j})^2 + \lambda \|\boldsymbol{\beta}\|_1$$

$$\hat{\boldsymbol{\beta}}_j^{d_y} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p)$$

Decision statistic

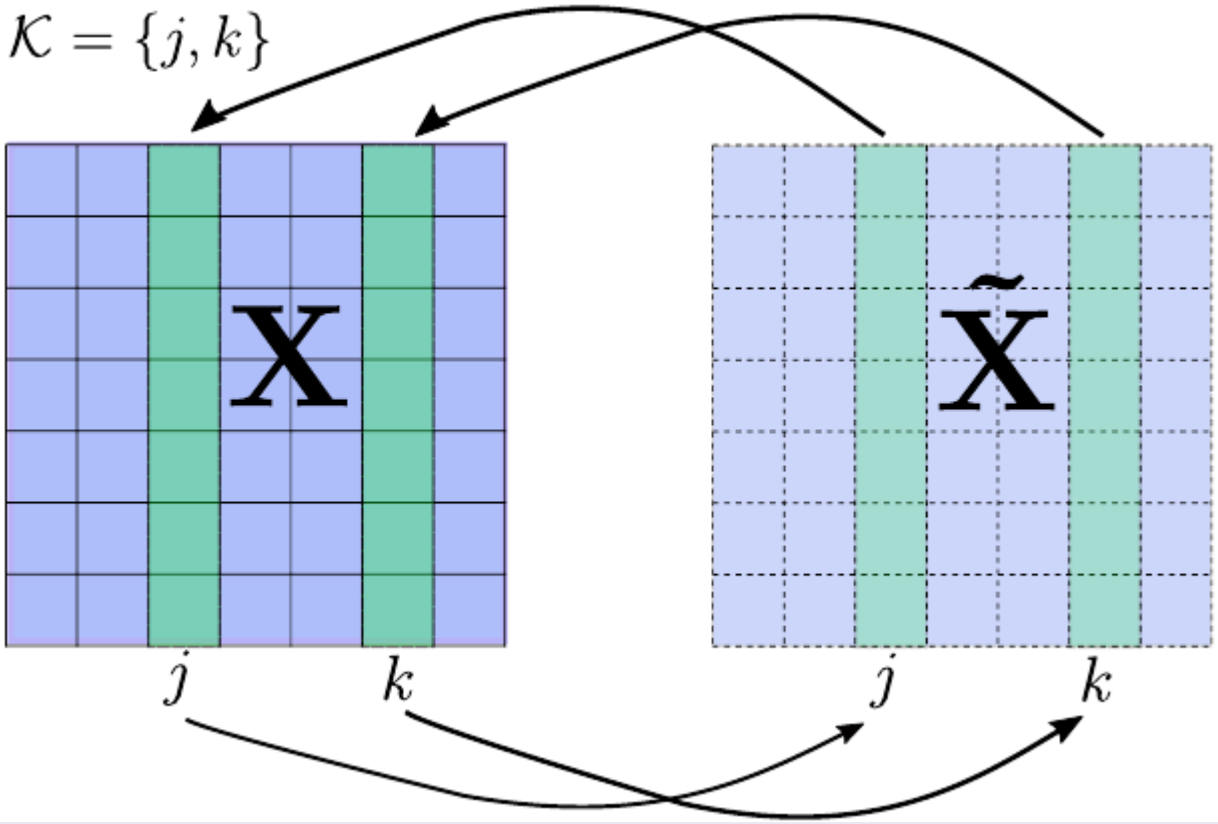
$$T_j^{\text{decorr}} = -\frac{1}{\sqrt{n}} \hat{\mathbf{I}}_{j|-j}^{-1/2} \sum_{i=1}^{n} \left(y_i - \frac{1}{1 + \exp\left(-\mathbf{x}_{i,-j}^T \hat{\boldsymbol{\beta}}_j^{d_y}\right)} \right) \left(x_{i,j} - \mathbf{x}_{i,-j}^T \hat{\boldsymbol{\beta}}_j^{d_{x_j}}\right)$$

$$\hat{\mathbf{I}}_{j|-j}^{-1/2} = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\left(\hat{\boldsymbol{\beta}} \mathbf{x}_i\right)}{\left[1 + \exp\left(\hat{\boldsymbol{\beta}} \mathbf{x}_i\right)\right]^2} \left(x_{i,j} - \mathbf{x}_{i,-j}^T \, \hat{\boldsymbol{\beta}}^{d_{x_j}}\right)^2 \, x_{i,j}$$

Core property $T_j^{\text{decorr}} \xrightarrow[n \to \infty]{\text{d}} \mathcal{N}(0, 1)$

[Ning and Liu Annals stats 2017, Van de Geer Annals stats2014]

Definition of Knockoff



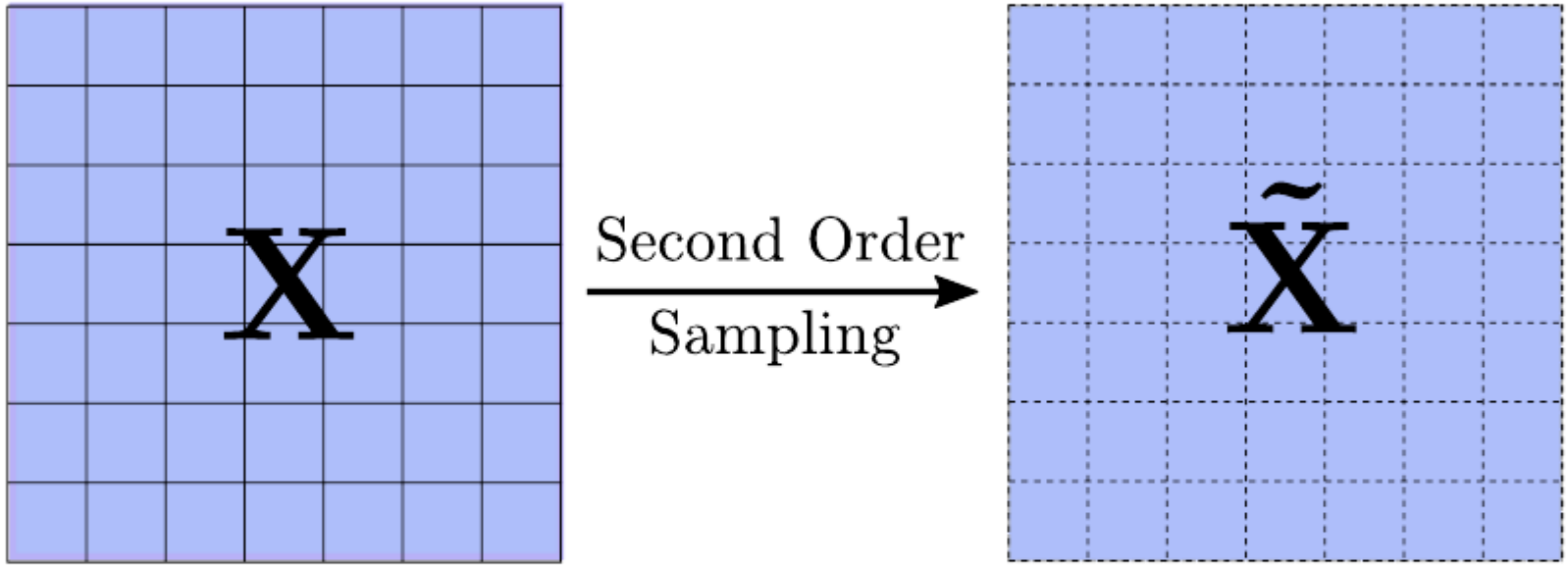
[Barber & Candès 2015]

$\tilde{\mathbf{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is model-X knockoffs of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ if and only if:

1. $\forall$ subset $\mathcal{K} \subset \{1, \ldots, p\}$: $(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(\mathcal{K})} \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})$
2. $\tilde{\mathbf{X}} \perp \mathbf{y} \mid \mathbf{X}$

Sampling Knockoffs

$$\text{cov}(\mathbf{X}, \tilde{\mathbf{X}}) = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diags}\{s\} \\ \boldsymbol{\Sigma} - \text{diags}\{s\} & \boldsymbol{\Sigma} \end{bmatrix}$$



Shares the same first 2 moments - mean and covariance:

$$\mathbb{E}[\tilde{\mathbf{X}}] = \mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}, \quad \mathbb{E}[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}] = \boldsymbol{\Sigma} \quad \text{and} \quad \mathbb{E}[\tilde{\mathbf{X}}^T\mathbf{X}] = \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}$$

Knockoff statistic

Step 1

Construct knockoff variables, concatenate $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$

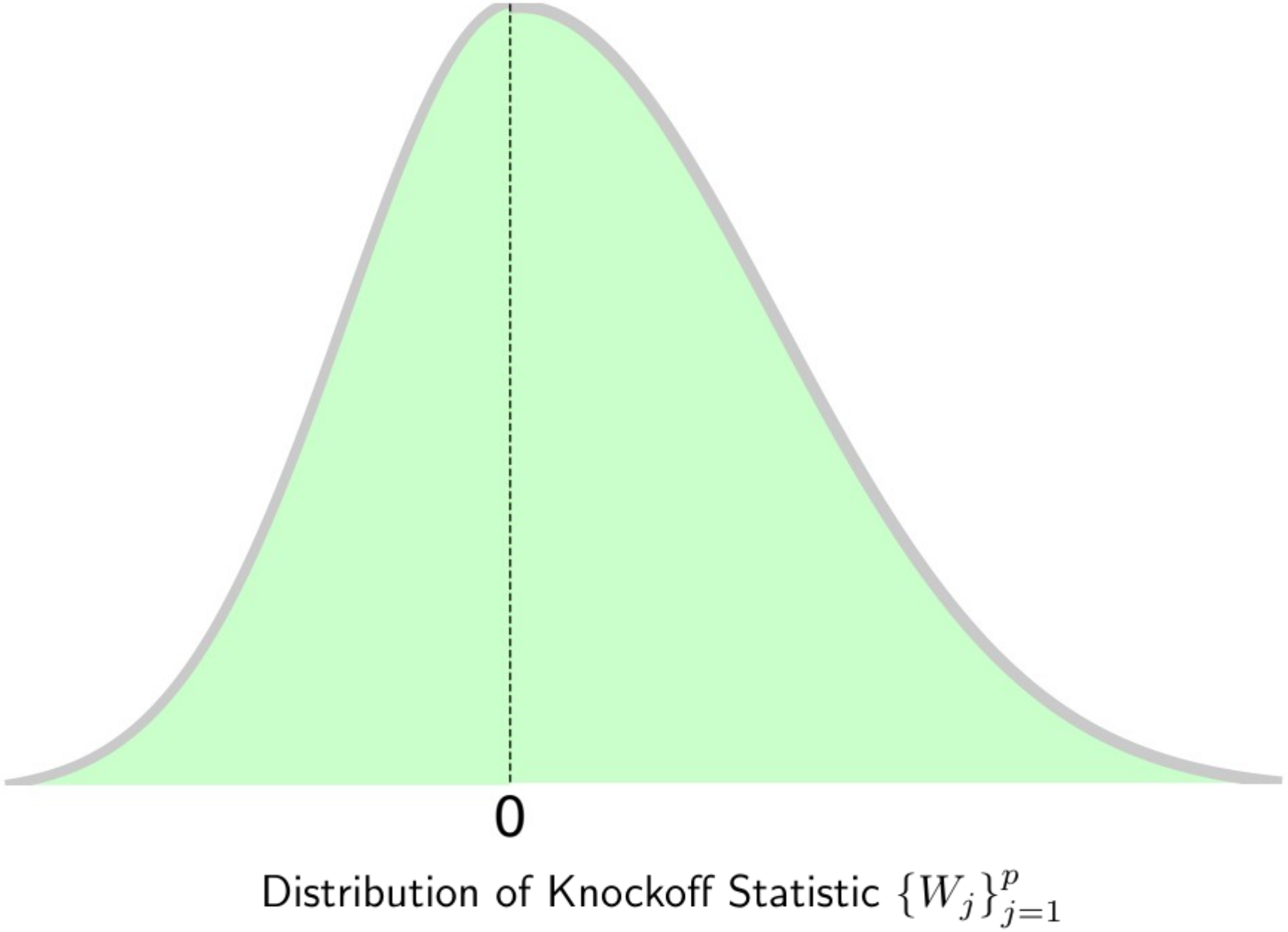
Step 2

Calculate knockoff test-statistics: *Lasso coefficient-difference*, obtain

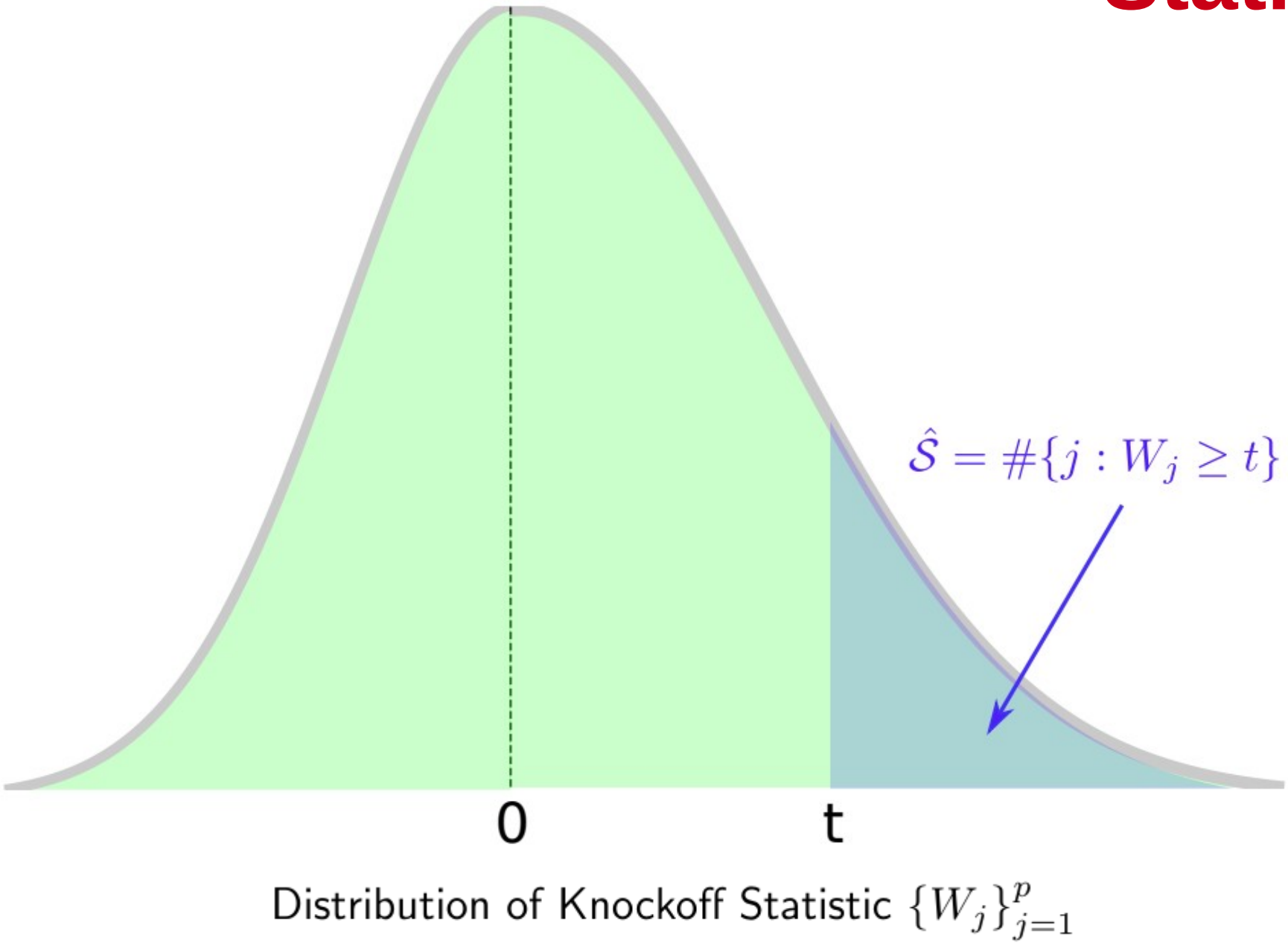
$$\hat{\boldsymbol{\beta}} = \min_{\mathbf{w} \in \mathbb{R}^{2p}} \frac{1}{2} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

then take the difference: $W_j = \left|\hat{\beta}_j(\lambda)\right| - \left|\hat{\beta}_{j+p}(\lambda)\right|$ for each j

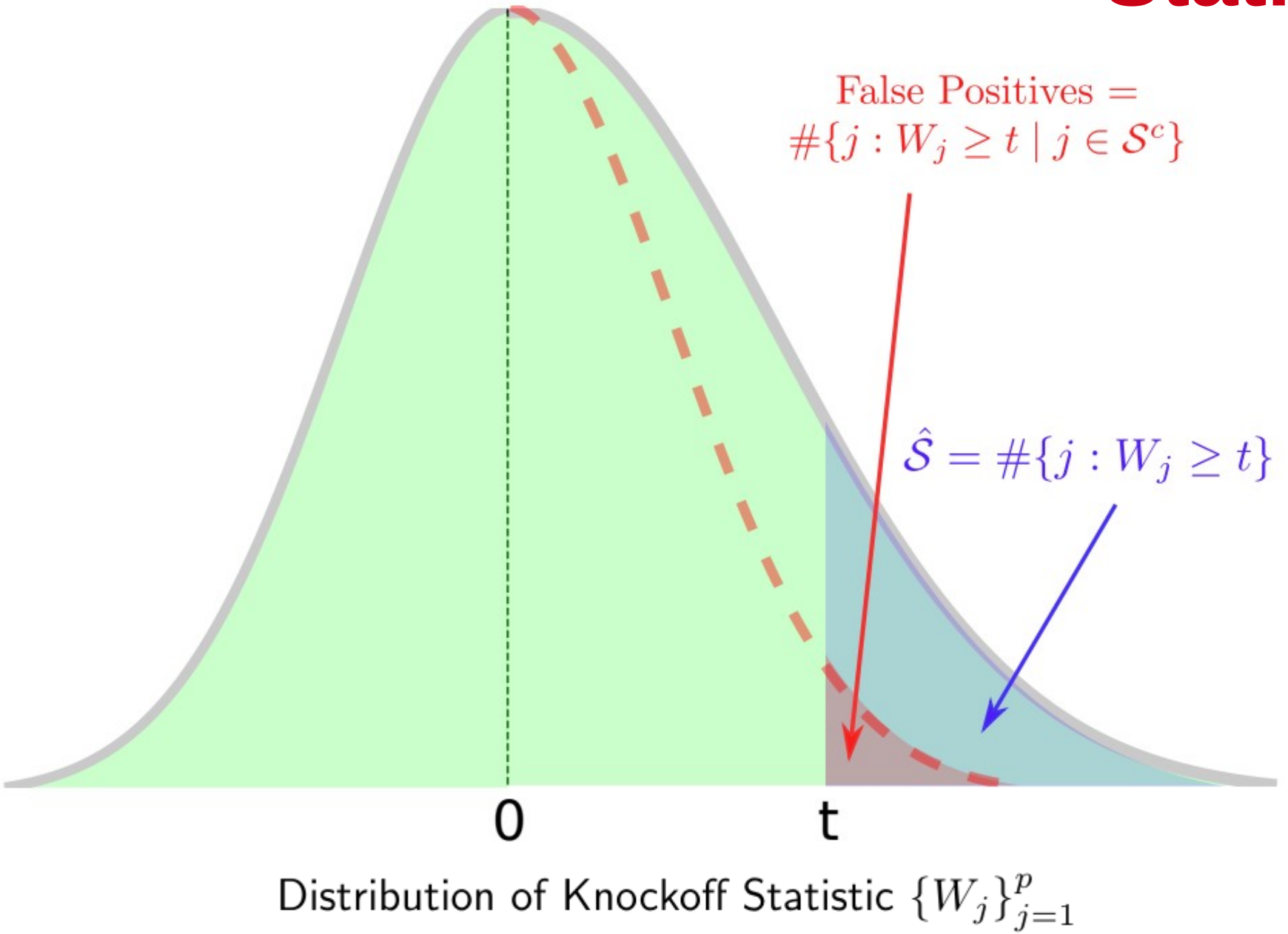
FDP estimation with Knockoff Statistic



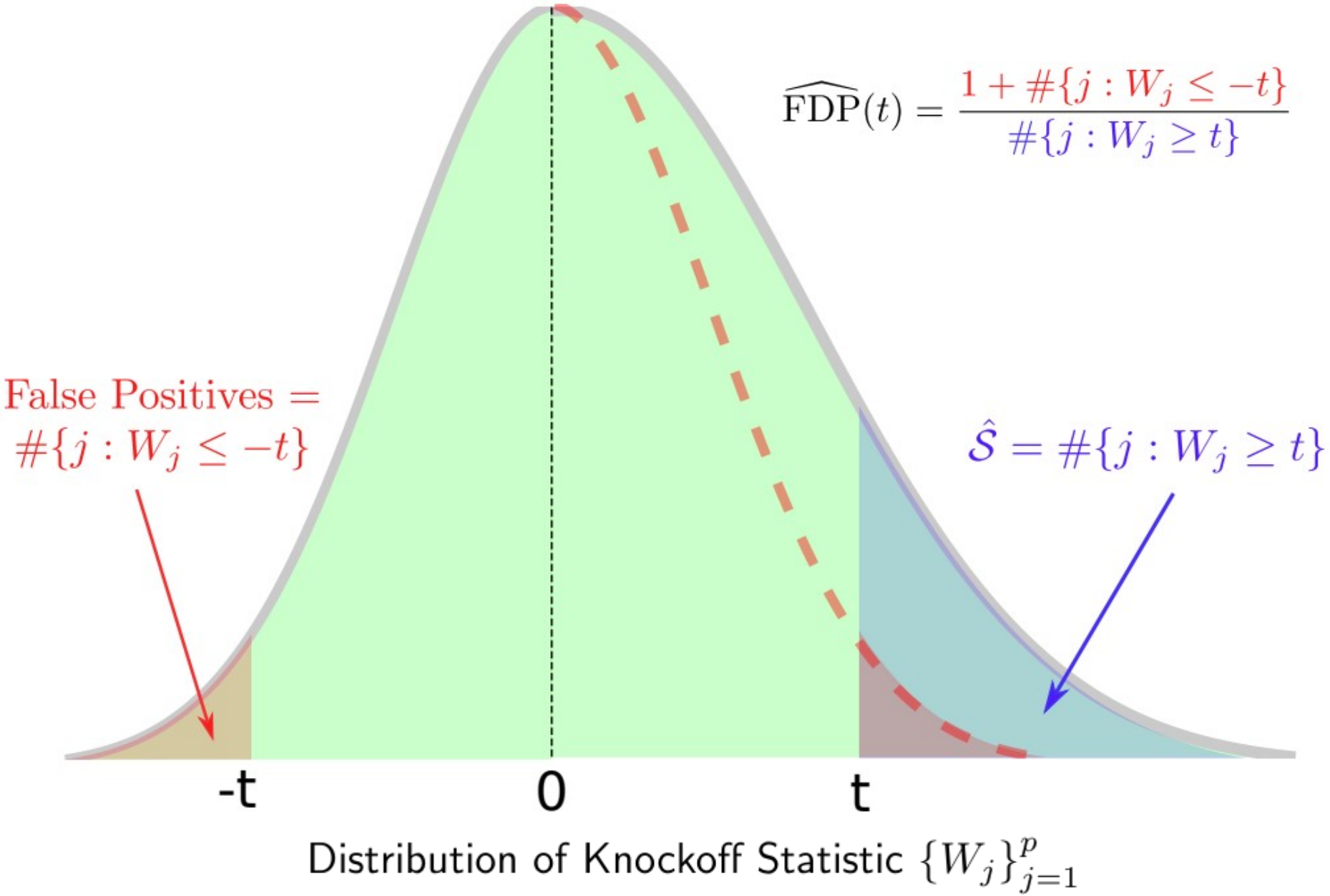
FDP estimation with Knockoff Statistic



FDP estimation with Knockoff Statistic



FDP estimation with Knockoff Statistic



Problem:
Instability of knockoff estimates

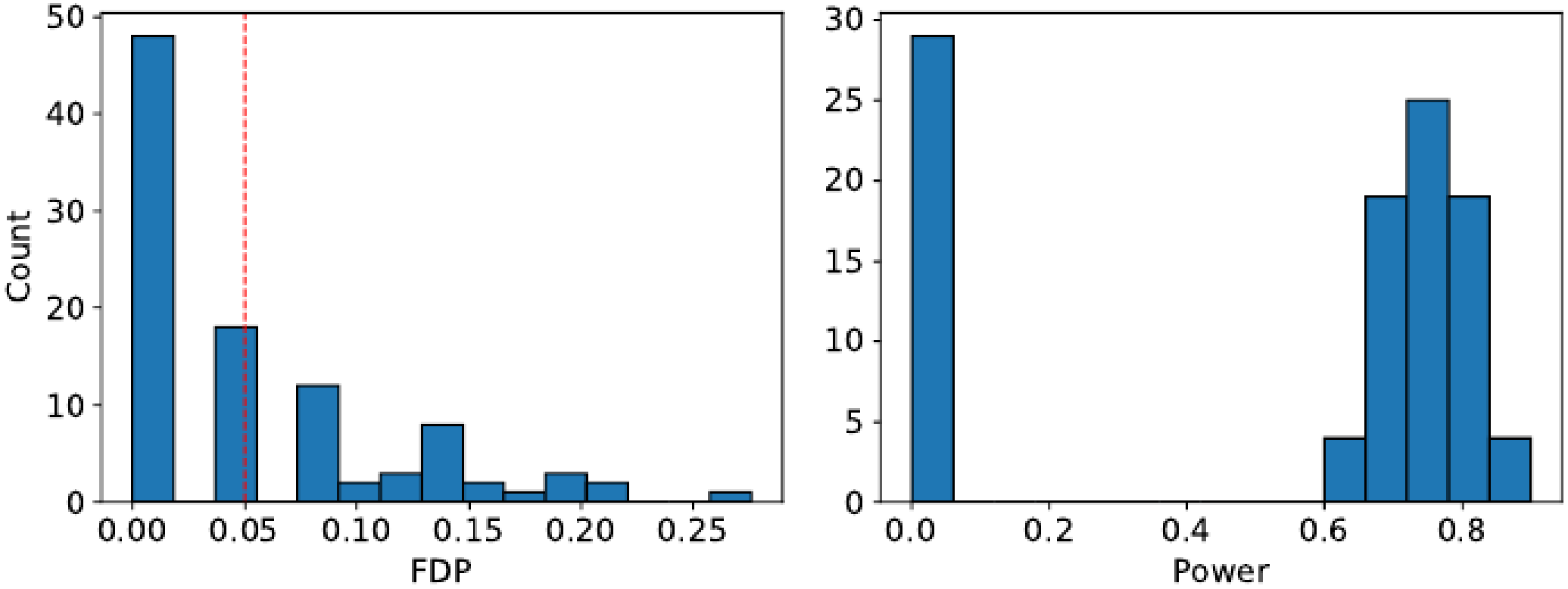
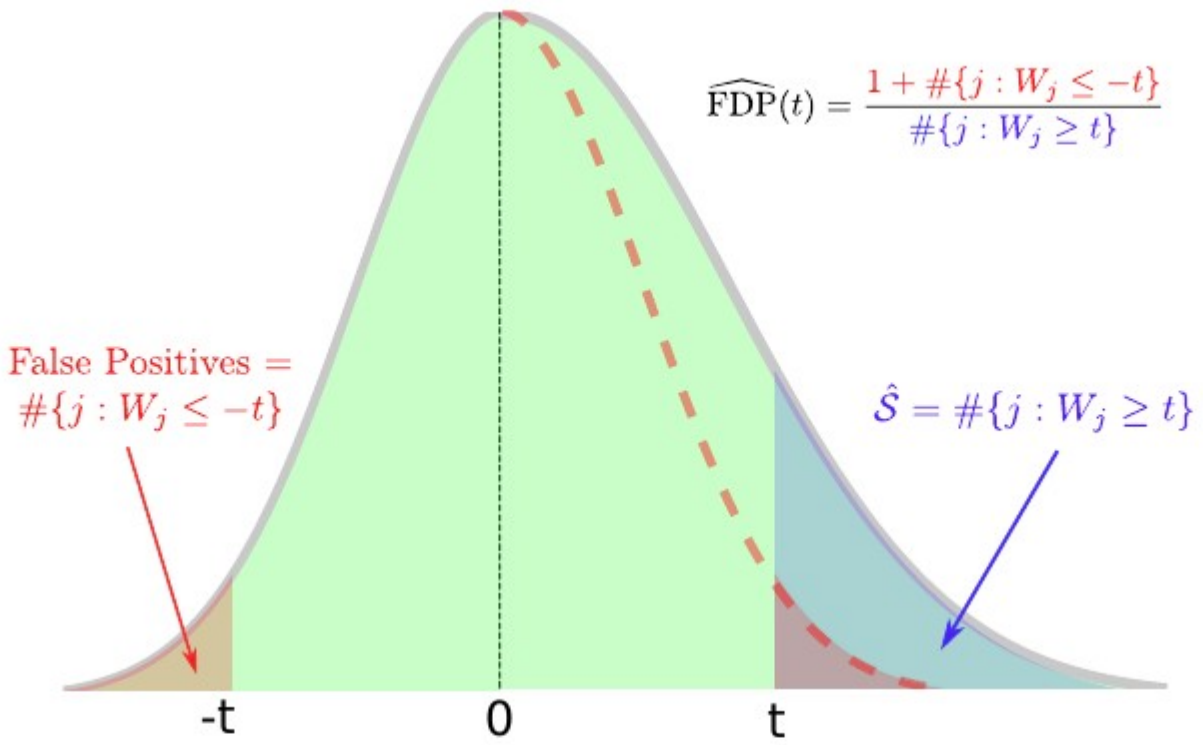


Figure: 100 runs of knockoff inference on the **same simulation**
n=500, p=1000, snr=3.0, $\rho = 0.7$, sparsity $= 0.06$

Intermediate p-values



Introduce the intermediate p-values: convert Knockoff statistic W_j to π_j:

$$\pi_j = \begin{cases} \dfrac{1+\#\{k : W_k \le -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \le 0 \end{cases}$$

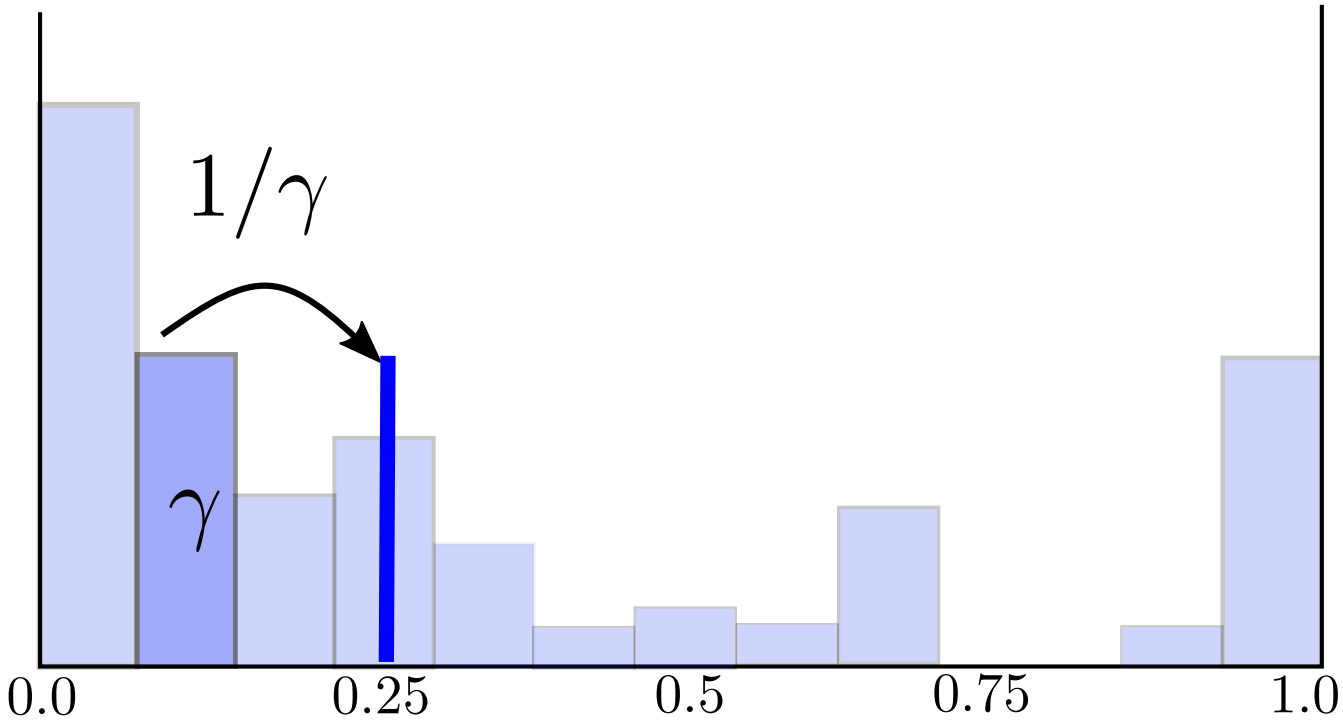
Knockoff aggregation

Step 1: For $b = 1, 2, \ldots, B$:

- Run knockoff sampling, calculate test statistic $\left\{W_j^{(b)}\right\}_{j\in[p]}$
- Convert the test statistic $W_j^{(b)}$ to $\pi_j^{(b)}$:

$$\pi_j^{(b)} = \begin{cases} \dfrac{1 + \#\{k : W_k^{(b)} \leq -W_j^{(b)}\}}{p} & \text{if} \quad W_j^{(b)} > 0 \\ 1 \quad \text{if} \quad W_j \leq 0 \end{cases}$$

Knockoff aggregation



Step 2 – Quantile Aggregation of p-values (Meinshausen et al., 2009)

$$\bar{\pi}_j = \min\left\{ \frac{q_\gamma(\pi_j^{(b)})}{\gamma}, 1 \right\} \quad \forall j \in [p]$$

For $\gamma \in (0, 1)$ with $q_\gamma(\cdot)$ the empirical γ-quantile function.

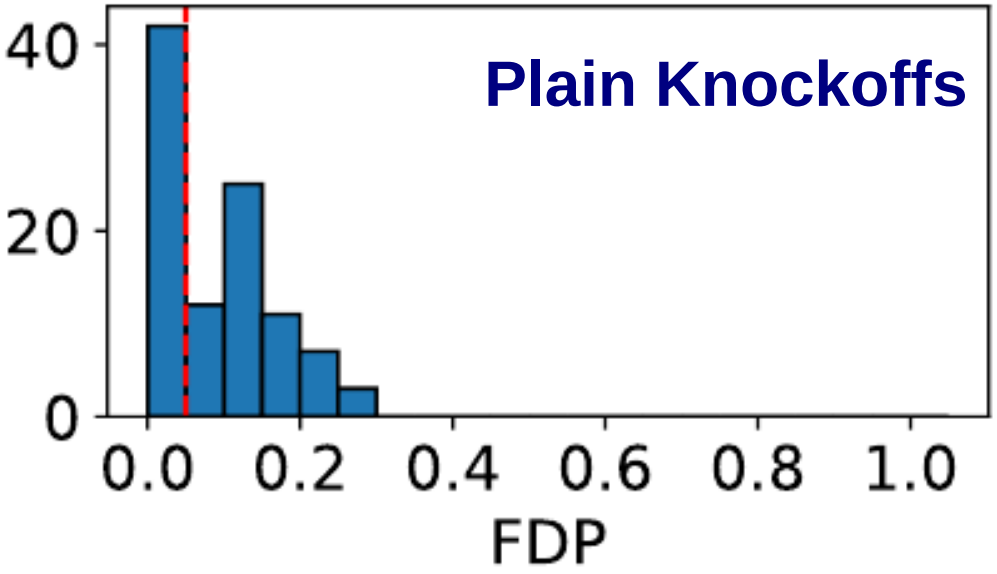
Knockoff aggregation

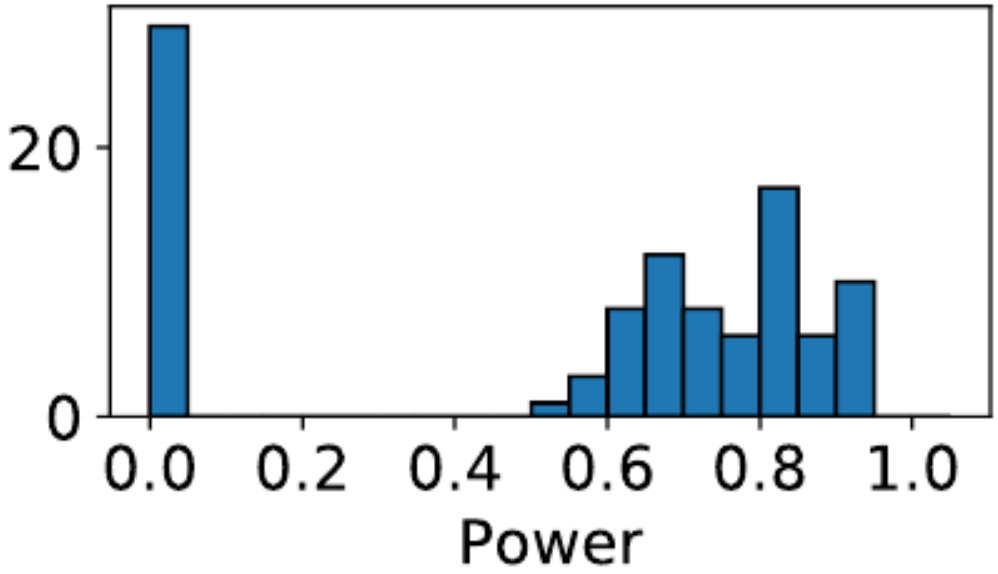
Step 3 – FDR control with $\bar{\pi}$

- Order $\bar{\pi}_j$ ascendingly: $\bar{\pi}_{(1)} < \bar{\pi}_{(2)} \cdots < \bar{\pi}_{(p)}$
- Given FDR control level $\alpha \in (0, 1)$, find largest k such that:
  - $\bar{\pi}_{(k)} \leq k\alpha/p$ (Benjamini and Hochberg, 1995), or
  - $\bar{\pi}_{(k)} \leq \dfrac{k\alpha}{p \sum_{i=1}^{p} 1/i}$ (Benjamini and Yekutieli, 2001)

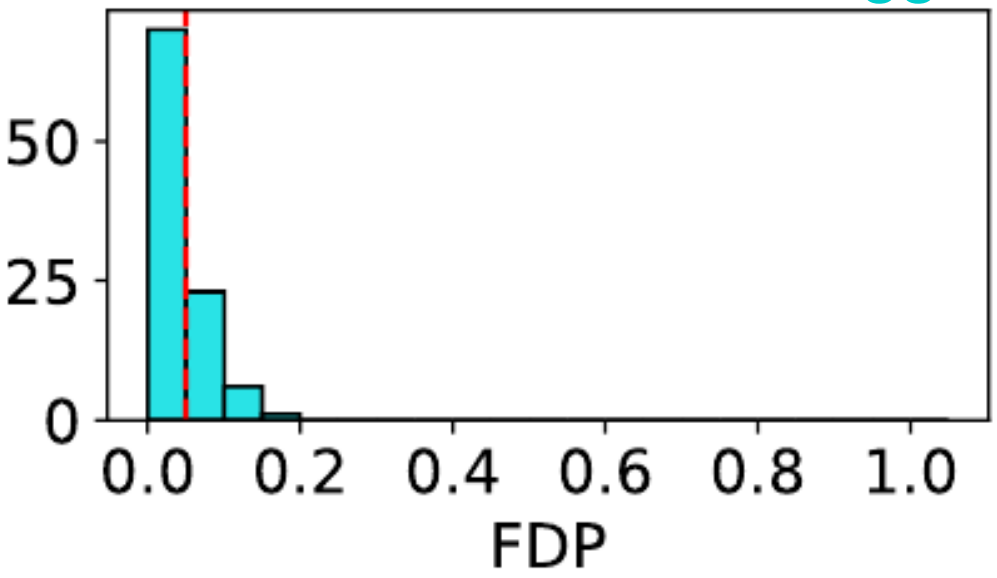
  $\longrightarrow$ FDR threshold: $\tau = \bar{\pi}_{(k)}$
- $\hat{\mathcal{S}}_{AKO} = \{j : \bar{\pi}_j \leq \tau \mid j \in [p]\}$

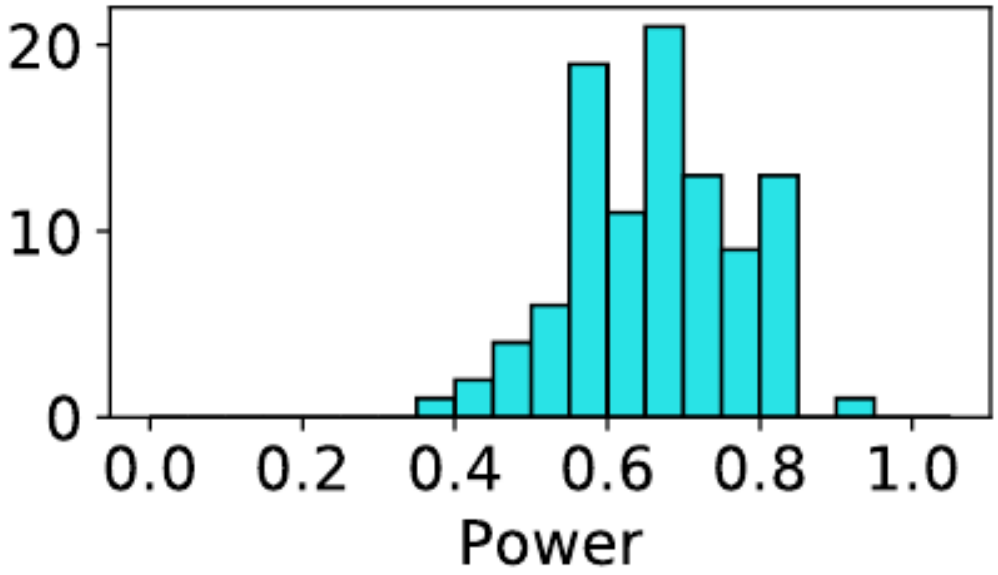
Empirical results: more stability





Aggregated Knockoffs





Brain activity decoding example

- Data: Human Connectome Project
- Objective: predict the experimental condition per task given brain activity
- $n = 900$ subjects, $p \approx 212000$
- Preprocessing: dimension reduction by clustering

$$p = 212000 \longrightarrow p = 1000$$

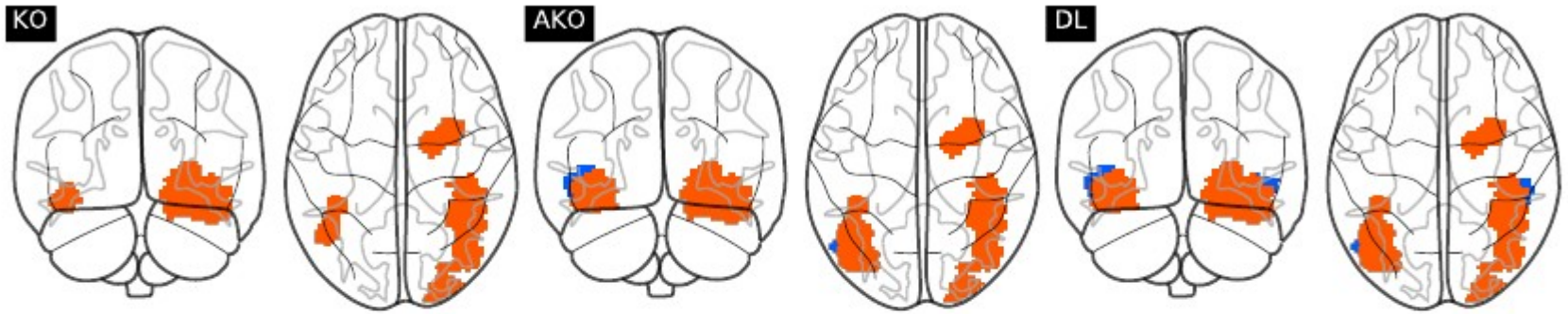
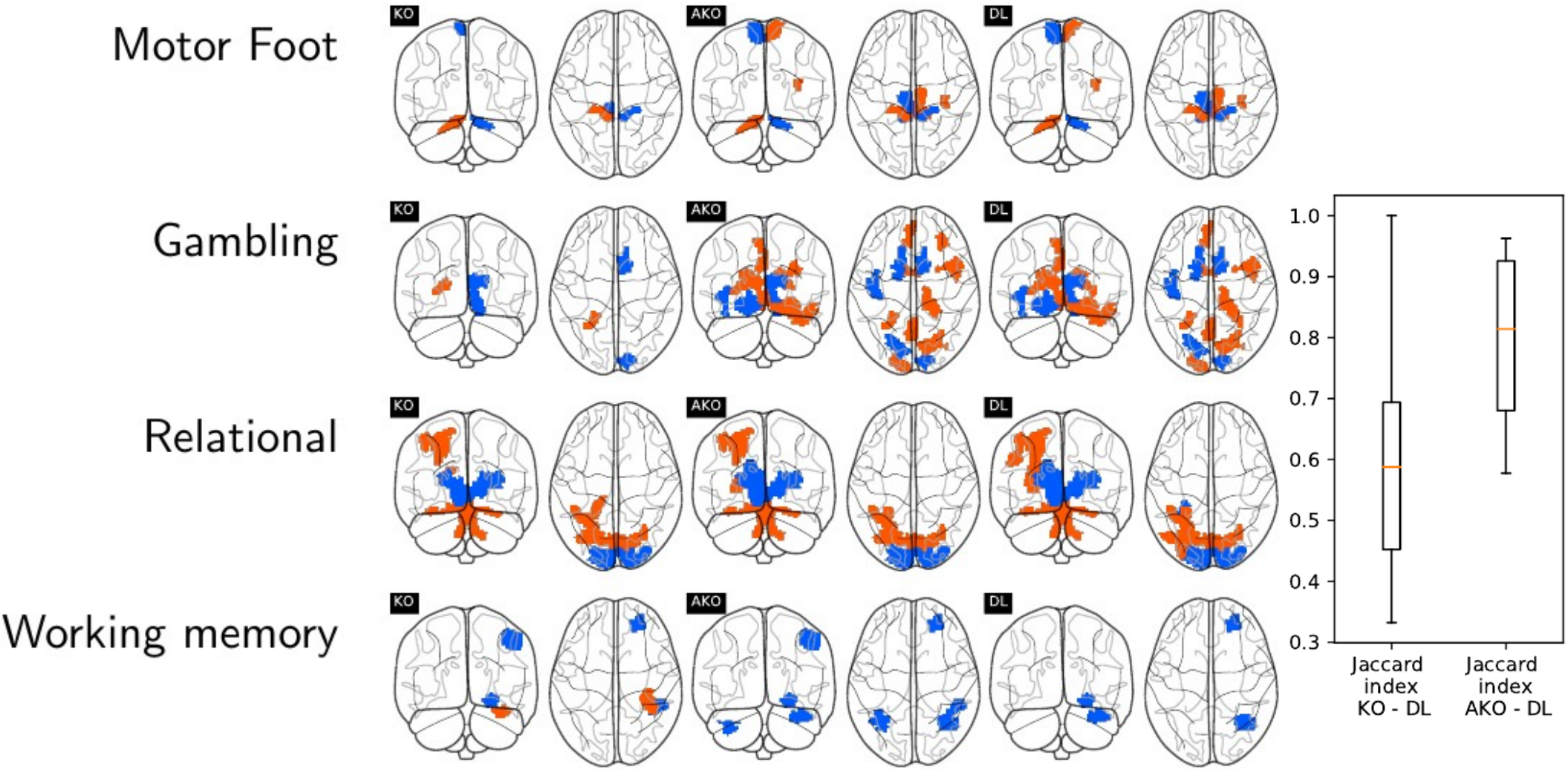


Figure: Detection of significant brain regions for HCP data (900 subjects)
Selected regions in a reaction with Emotion images task.
Orange: brain areas with positive sign activation.
Blue: brain areas with negative sign activation

Brain activity decoding example



Aggregation of Multiple Knockoffs

Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, Sylvain Arlot

Conditional randomization test

Algorithm 1: Conditional Randomization Test [CFJL18]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, number of sampling runs B, test statistic T_j, conditional distribution $P_{j|-j}$ for each $j = 1, \ldots, p$;

OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^p$;

for $j = 1, 2, \ldots, p$ **do**

  for $b = 1, 2, \ldots, B$ **do**

    1. Generate $\tilde{\mathbf{X}}_{*,j}^{(b)}$, a knockoff sample from $P_{j|-j}$;

    2. Compute test statistics T_j for original variable and $\tilde{T}_j^{(b)}$ for knockoff variables;

  end

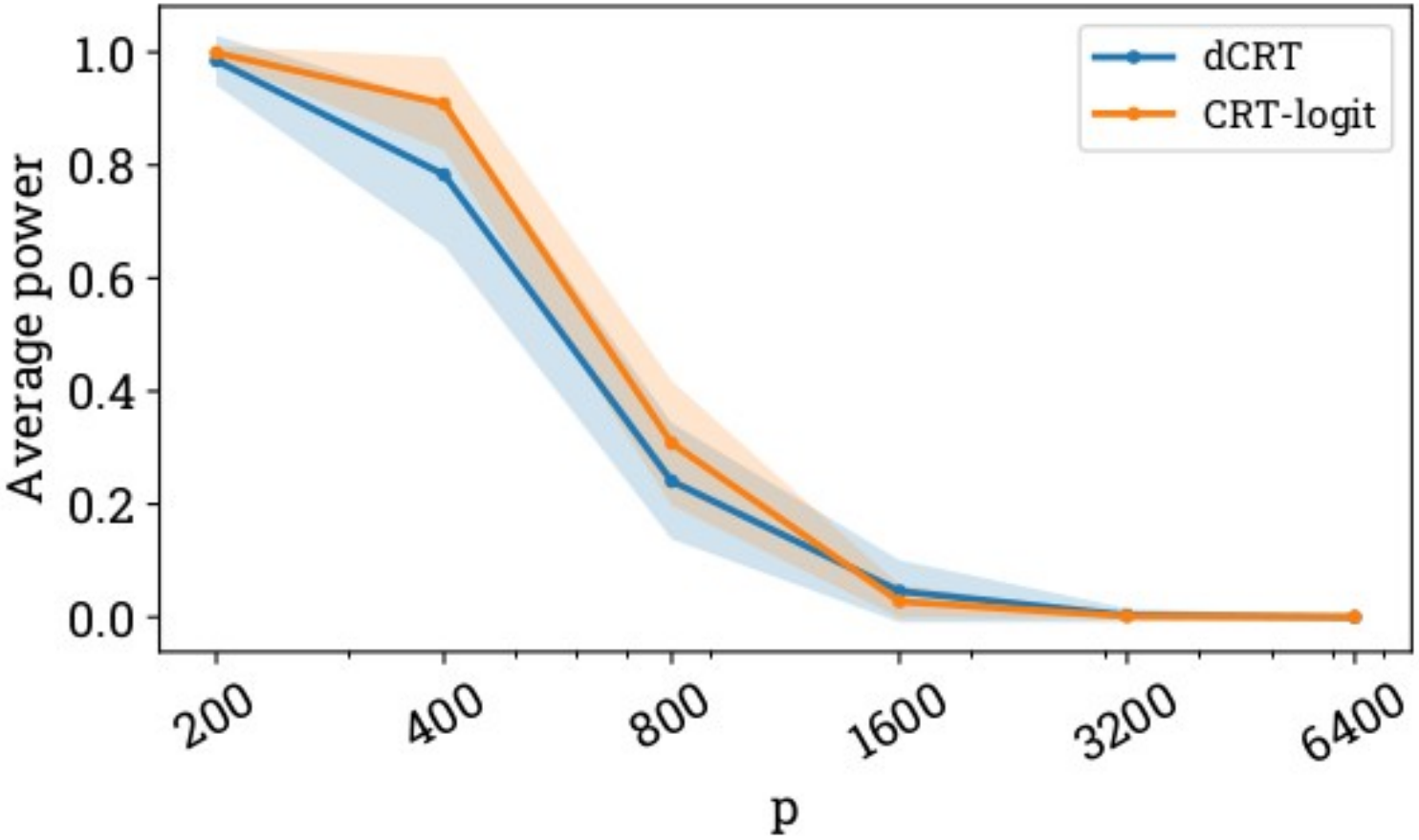
  Compute the empirical p-value

$$\widehat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B}$$

end

Upscaling model dimension

- Power collapses with increasing p: example with CRT



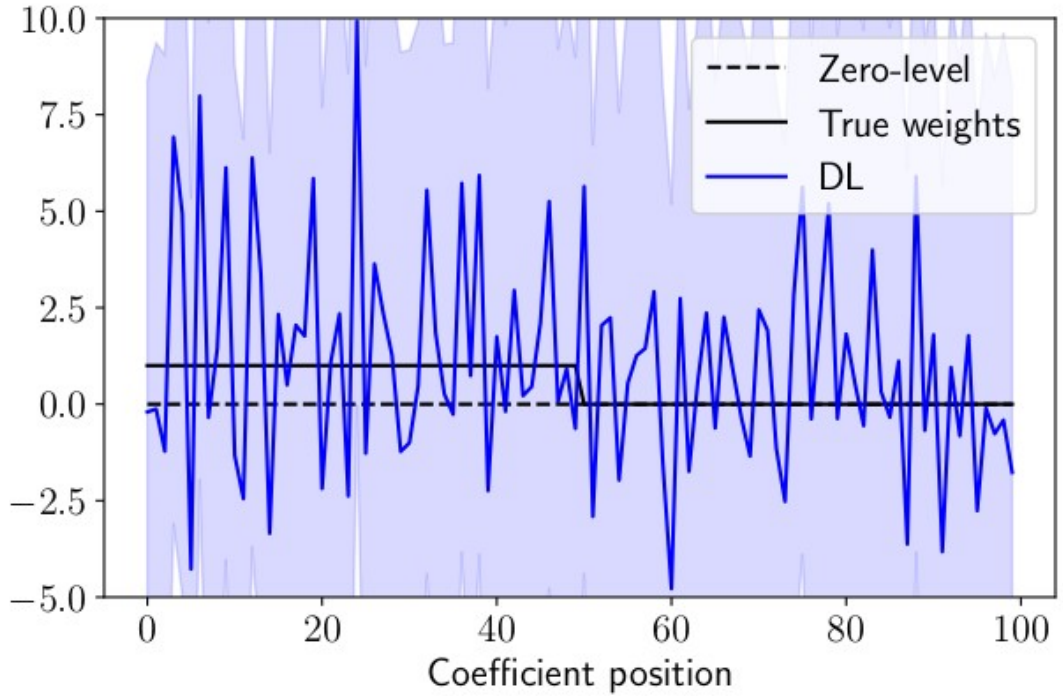
Power of 100 runs of simulations while varying the number of variables p and fixing the number of observations **n = 400**

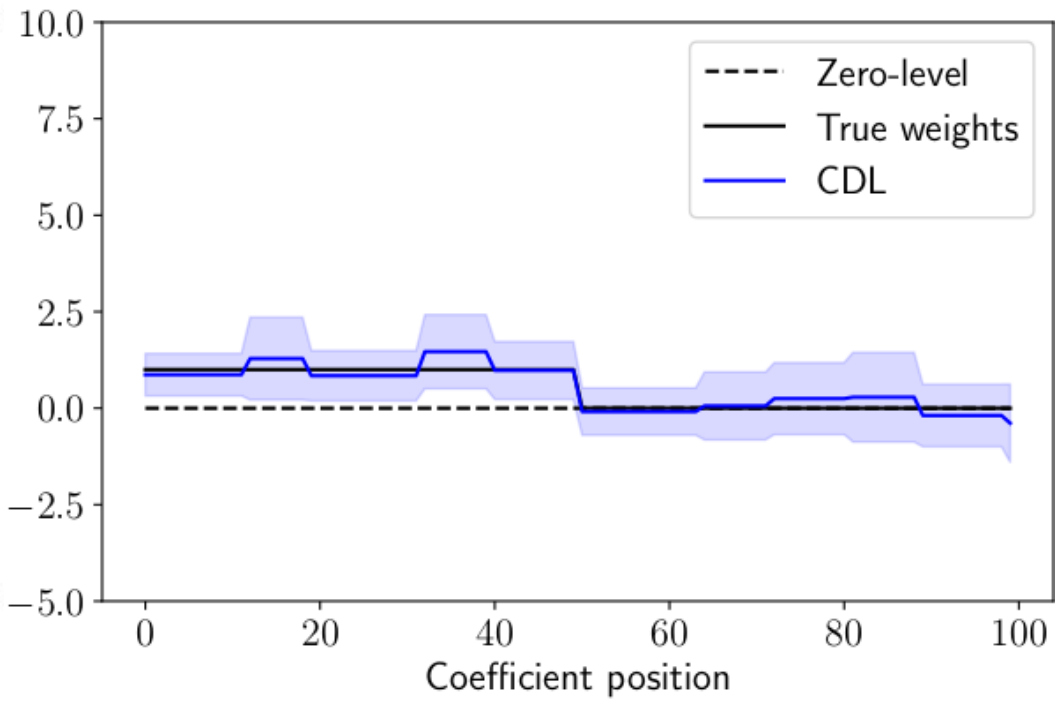
Other parameters:
SNR = 2.0,
ρ = 0.5,
κ = 0.04.
FDR control α = 0.1.

[Nguyen et al. In prep]

clustering-based dimension reduction

p=2000, n=100, inference with desparsified Lasso





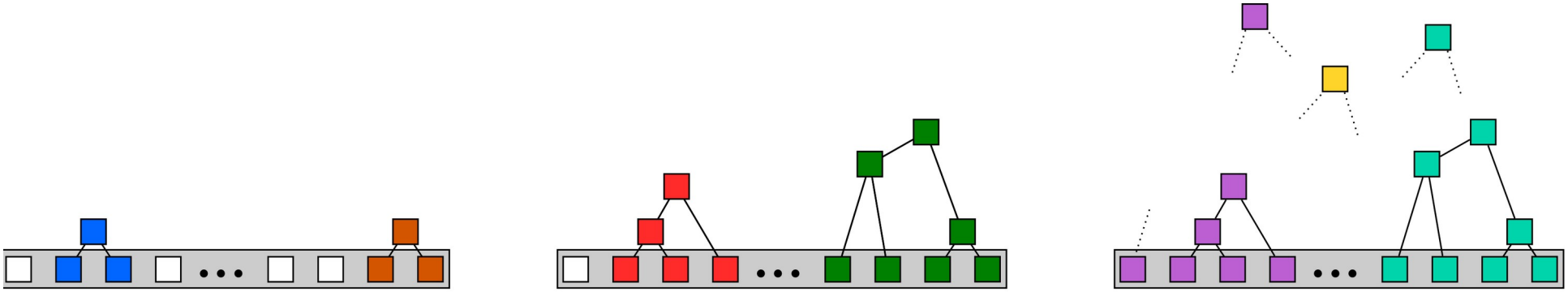
Large p kills statistical power

Clustering tames variance

[Chevalier et al. MICCAI 2018]

Adaptation to brain imaging

Clustering uses Ward's method → compact, homogeneous clusters



[Thirion et al. Front. 2014]

Regarding the **number of clusters**: Choose C = n by default
less clusters → more sensitivity
more clusters → higher spatial accuracy
[Chevalier et al. Nimg 2021]

What makes a good clustering ?

Consitions under which the reduced model does not create spurious associations

Proposition 4.1. *Considering the Gaussian linear model in (1) and assuming:*

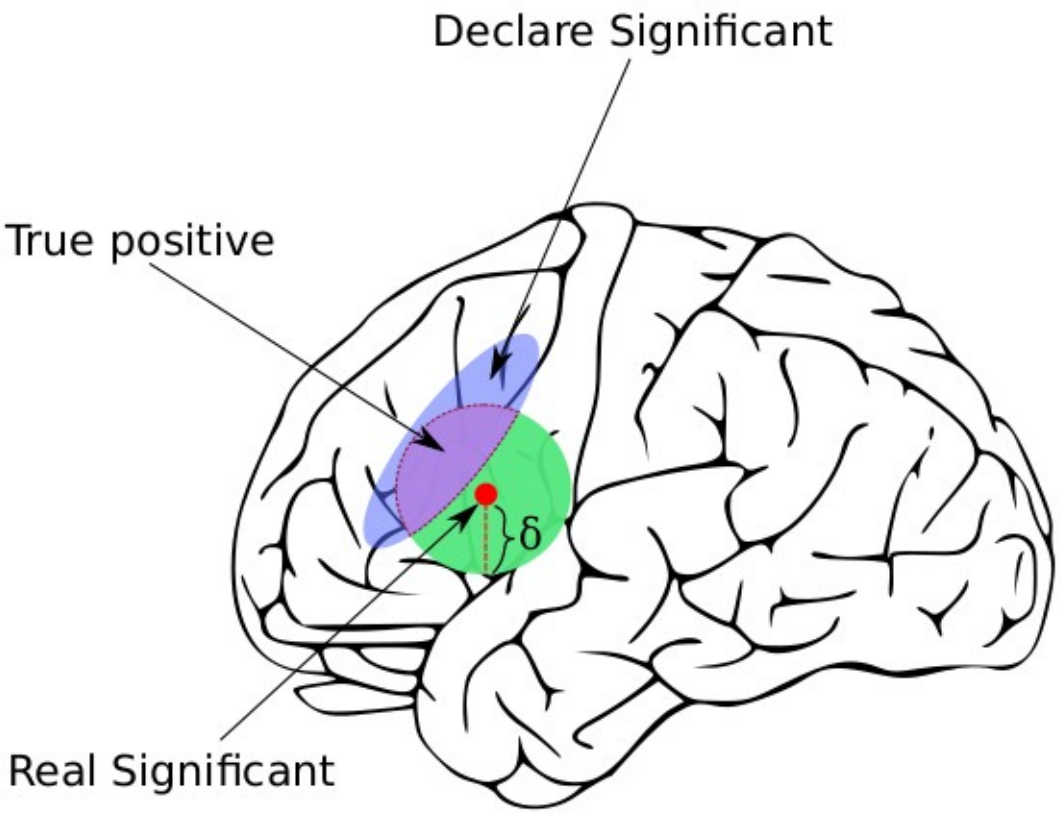
$$(i)\ \textit{for all } c \in [C],\ \textit{for all } (j,k) \in (G_c)^2,\ \boldsymbol{\Sigma}_{j,k} \geq 0\ ,$$
$$(ii)\ \textit{for all } c \in [C],\ \textit{for all } c' \in [C] \setminus \{c\},\ \boldsymbol{\Upsilon}_{c,c'} = 0\ ,$$
$$(iii)\ \textit{for all } c \in [C], \left(\boldsymbol{\beta}^*_j \geq 0 \textit{ for all } j \in G_c\right) \text{or} \left(\boldsymbol{\beta}^*_j \leq 0 \textit{ for all } j \in G_c\right)\ ,$$

then, in the compressed representation (3), for $c \in [C]$, $\boldsymbol{\theta}^*_c \neq 0$ *if and only if there exists* $j \in G_c$ *such that* $\boldsymbol{\beta}^*_j \neq 0$. *If such an index* j *exists then* $\text{sign}(\boldsymbol{\theta}^*_r) = \text{sign}(\boldsymbol{\beta}^*_j)$.

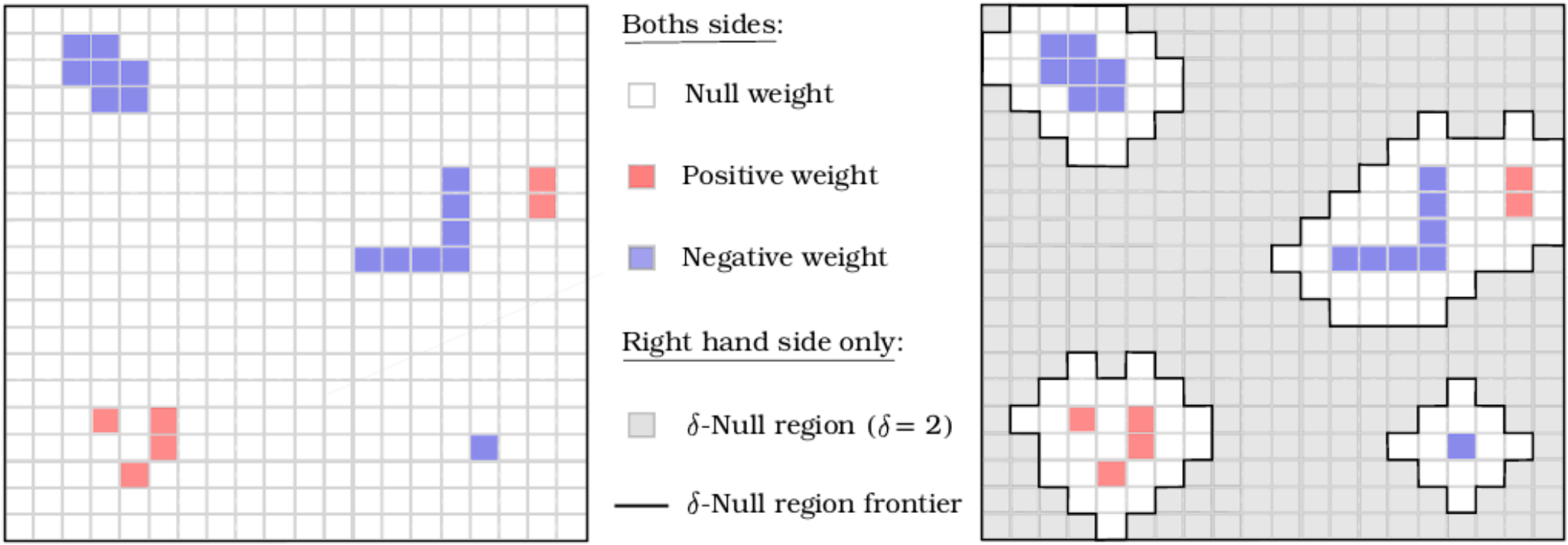
[Bühlman et al. 2013, Chevalier et al. STCO, in revision]

Insights on dimension reduction

- Inferring on a reduced model: accept that inference is δ-approximate → use of δ-FDR, δ-FWER



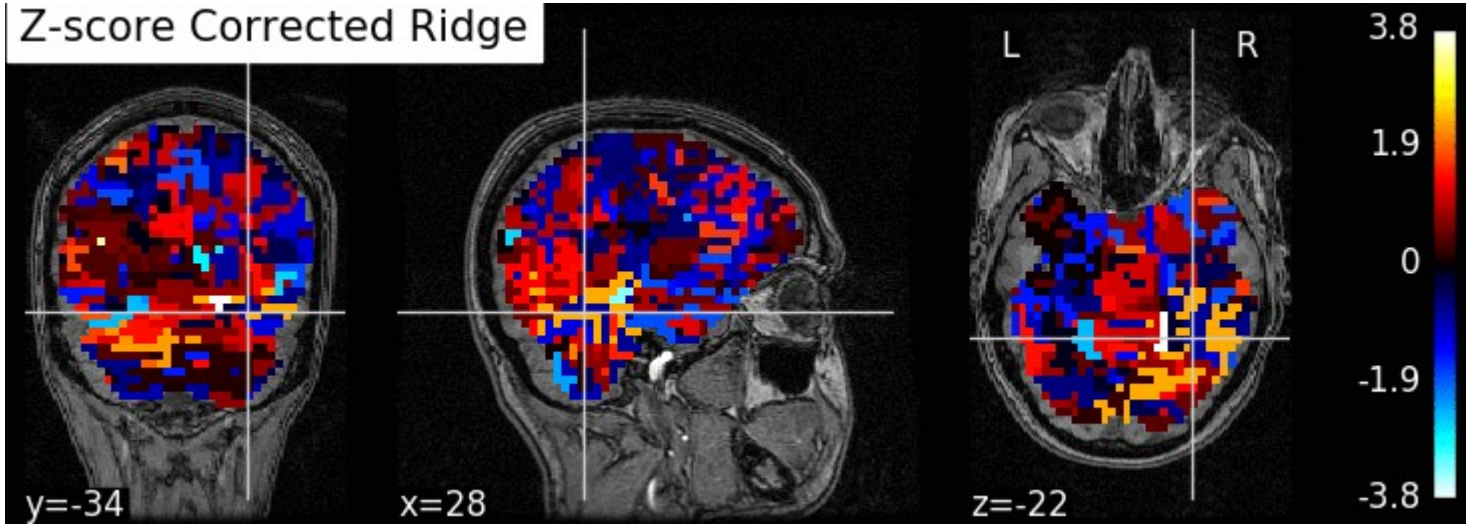
δ-FWER-control



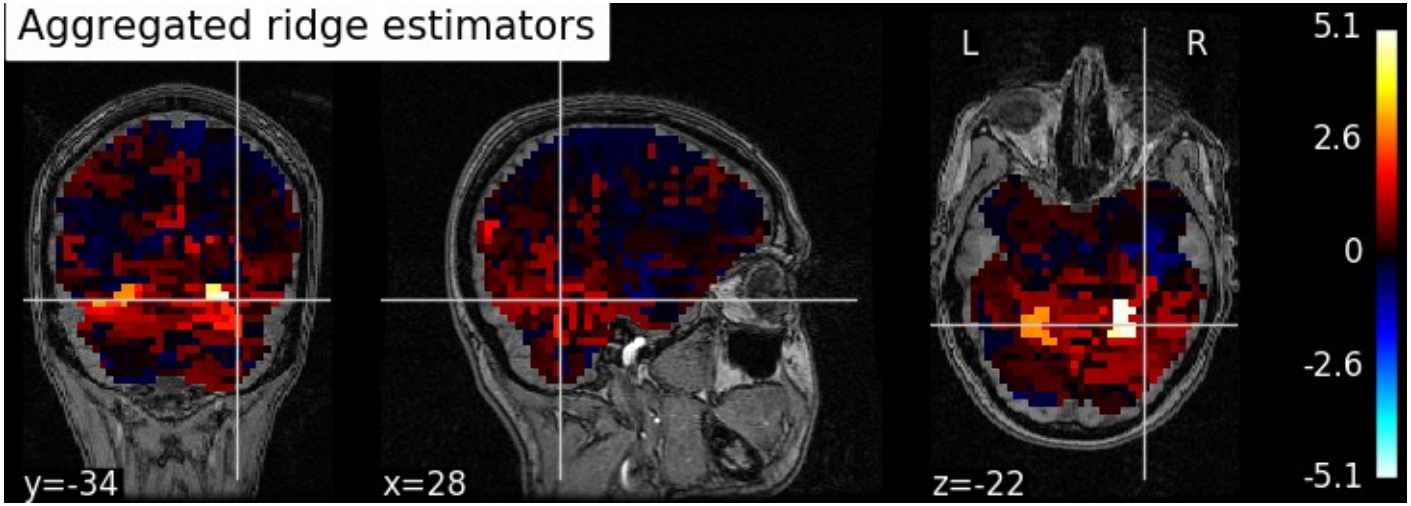
Spatial relaxation on FWER control

Why we need ensembling

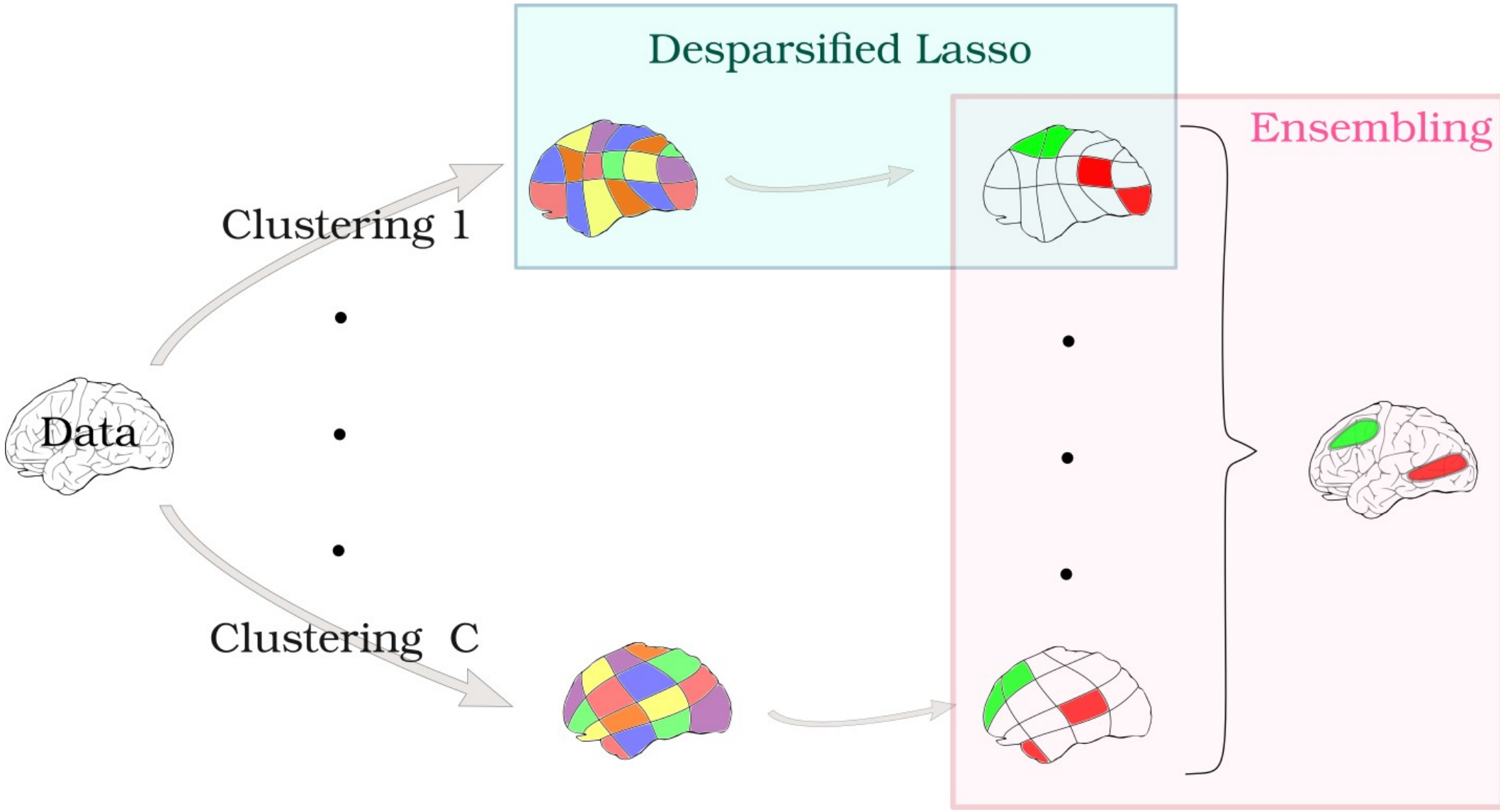
DL p-values from different clusterings



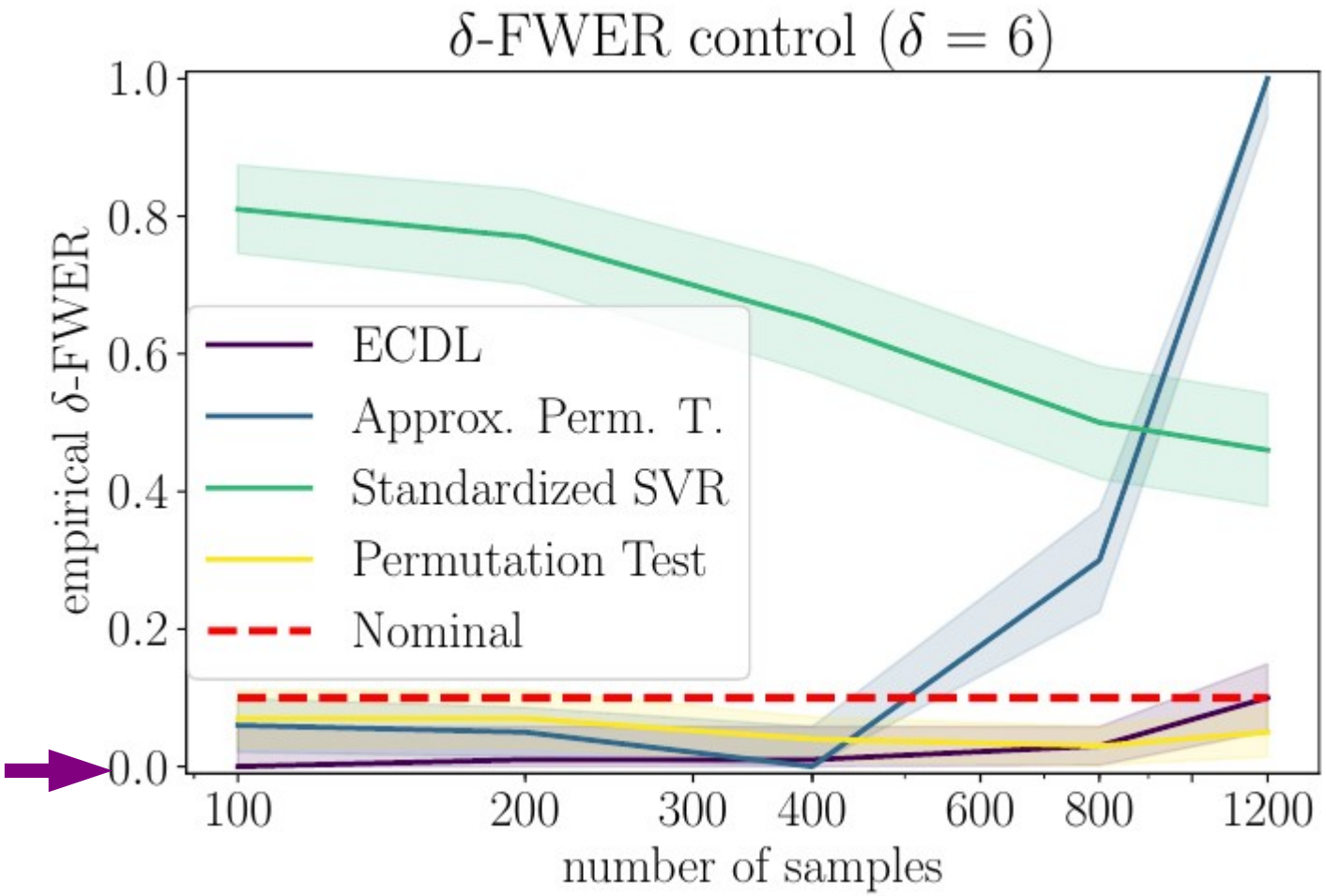
P-value aggregation
[Meinshausen 2009]



Ensemble of Clustered Desparsified Lasso for brain imaging



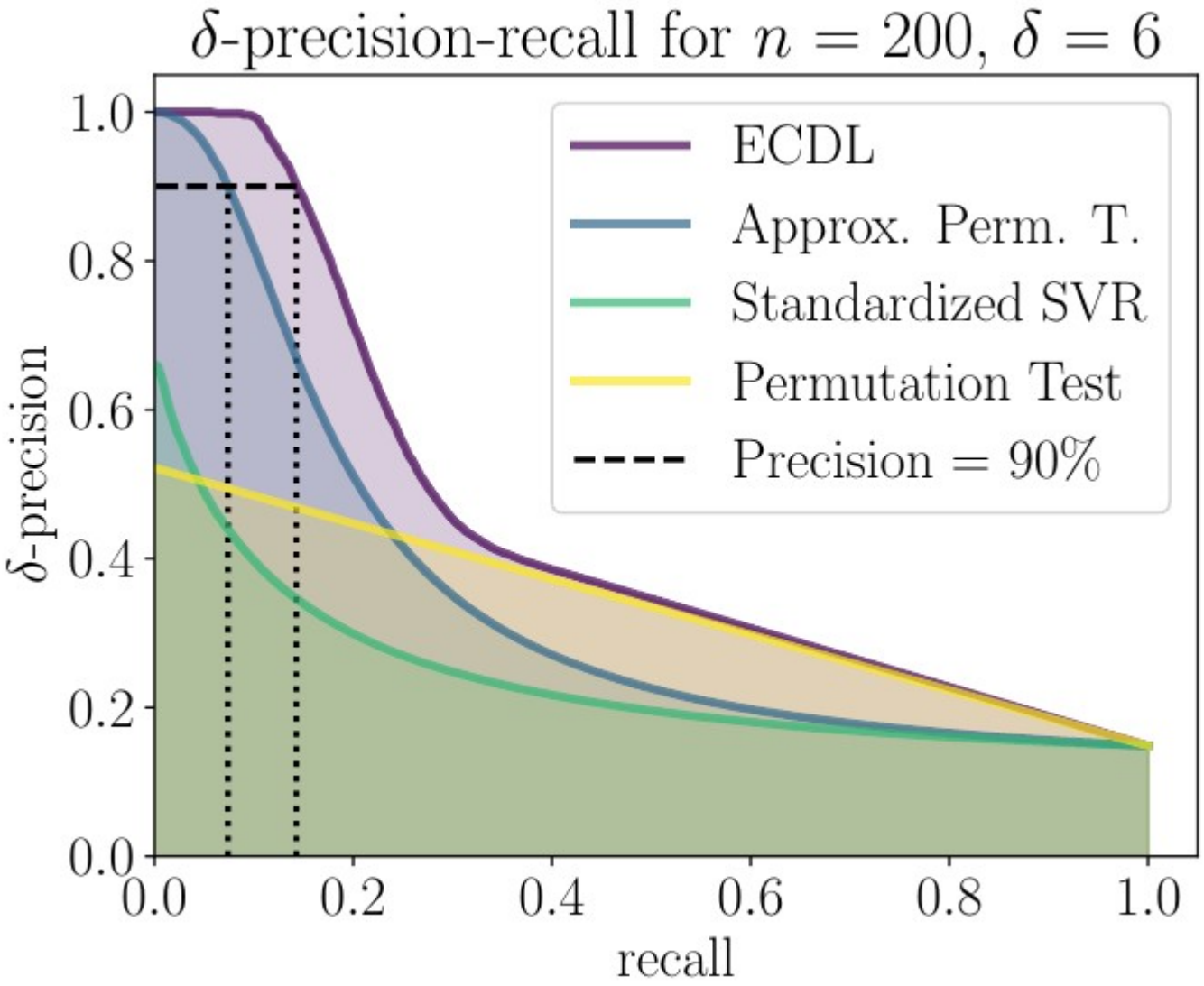
δ-FWER is controlled



δ-FWER control on semi-simulated data, obtained with 100 repetitions for every sample size.

[Chevalier et al. Nimg 2021]

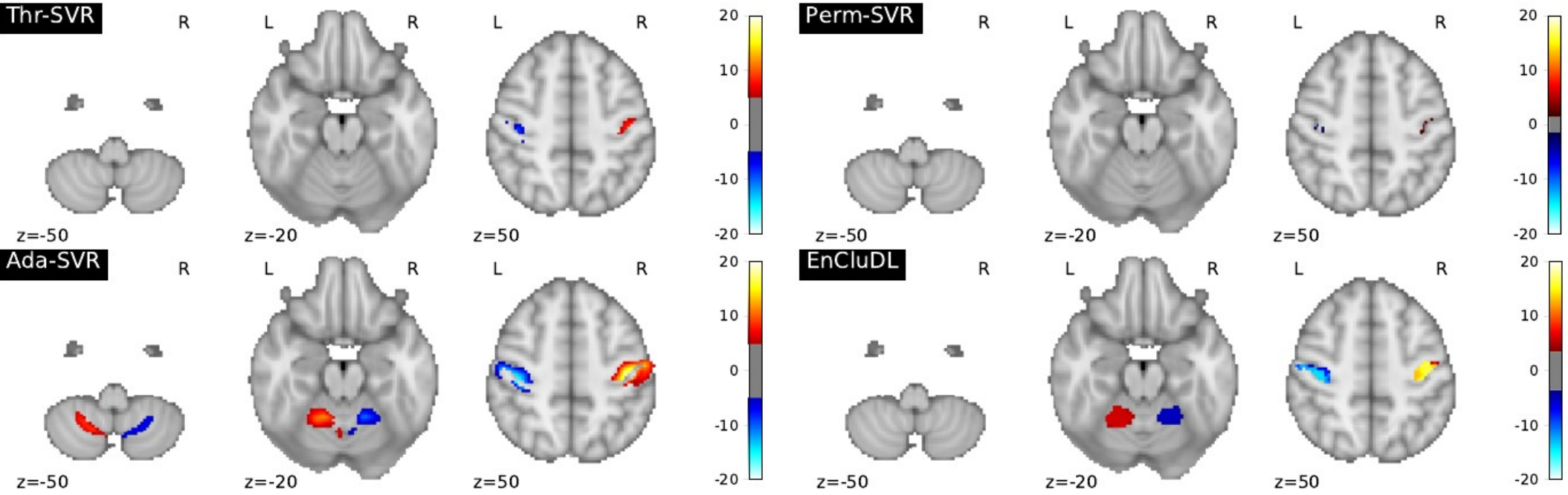
Results: higher PR curve



- On semi-simulated data, EDCL achieves better PR compromise

[Chevalier et al. Nimg 2020]

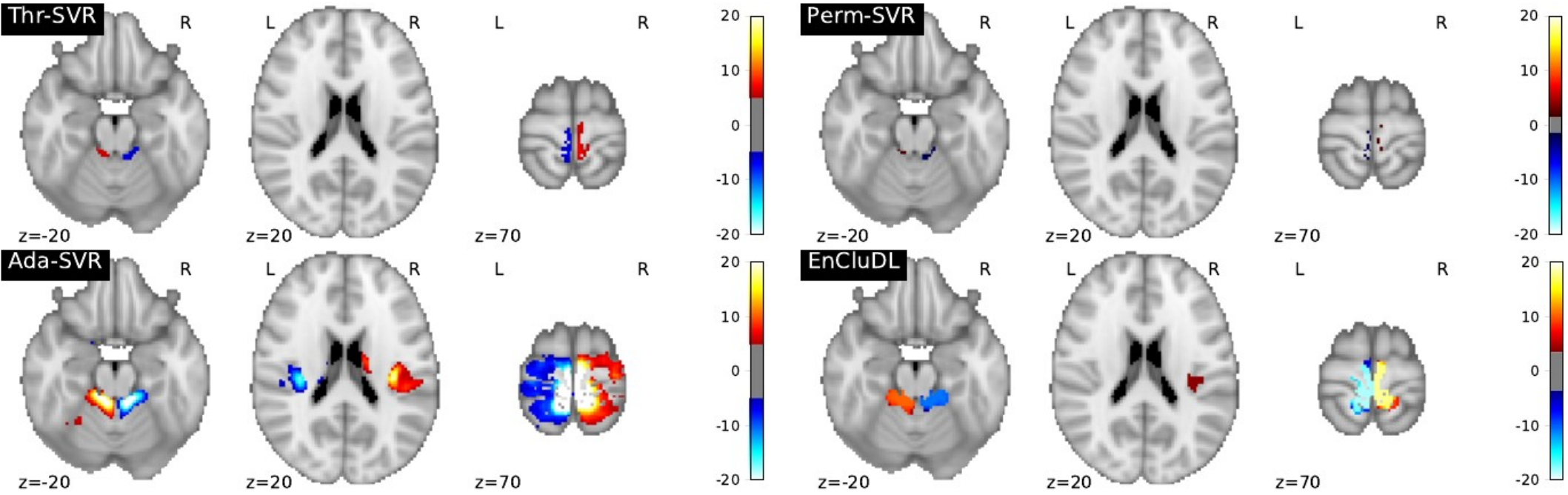
Face validity on real data



HCP dataset, n=800

[Nguyen et al. IPMI 2019, Chevalier et al. MICCAI 2018, Cevalier etal. Nimg 2021]

Face validity on real data

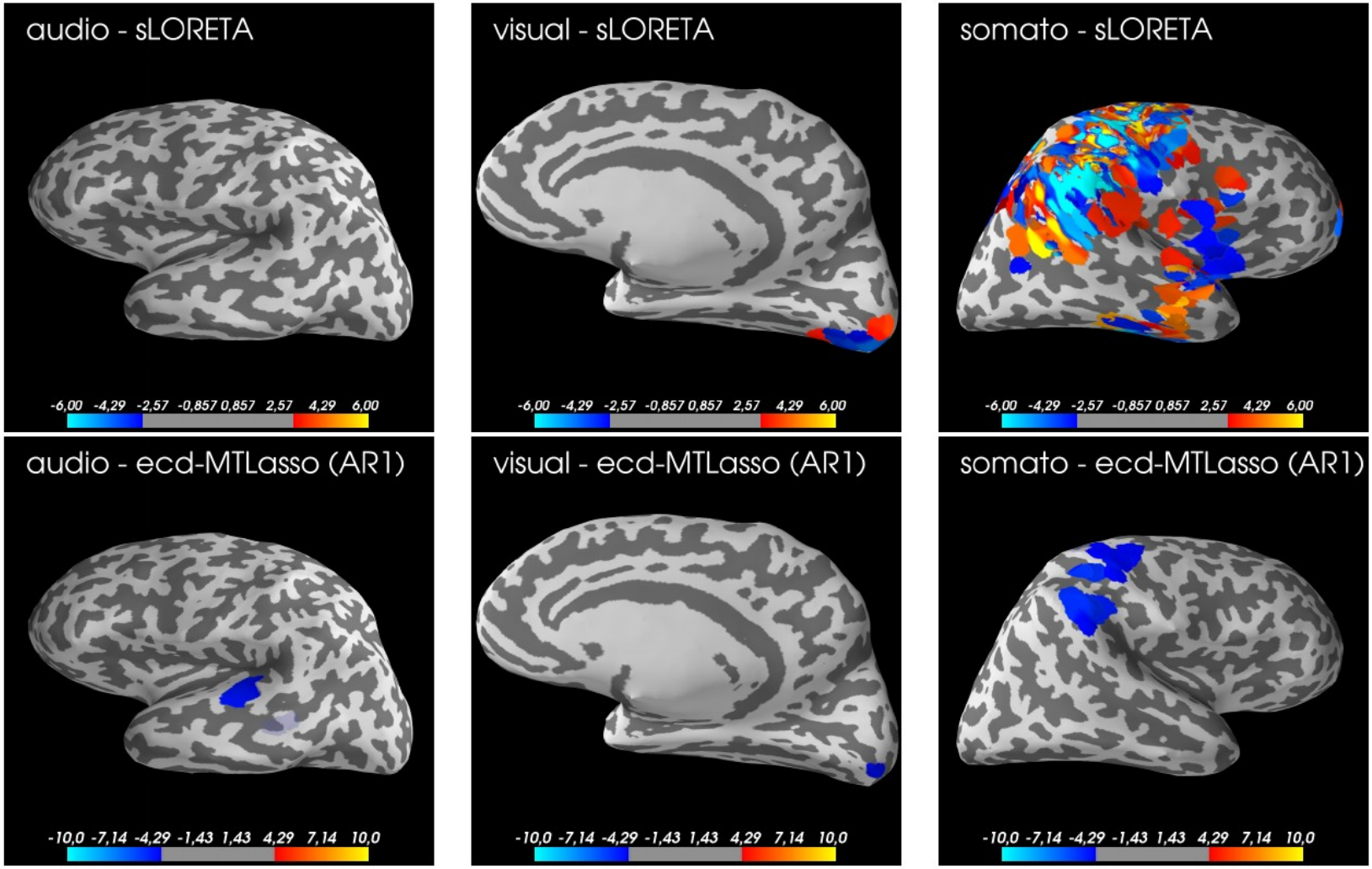


HCP dataset, n=800

[Nguyen et al. IPMI 2019, Chevalier et al. MICCAI 2018, Cevalier etal. Nimg 2021]

Solving the MEG inverse problem

- **sLORETA:** scaled Ridge estimator



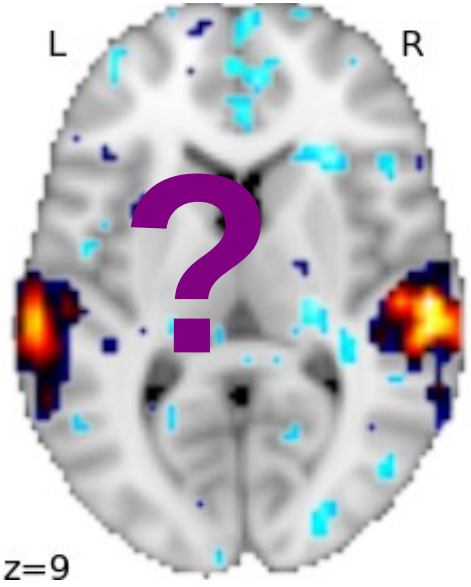
- **ecd-MTLasso offers a universal threshold**

[Chevalier et al NeurIPS 2020]

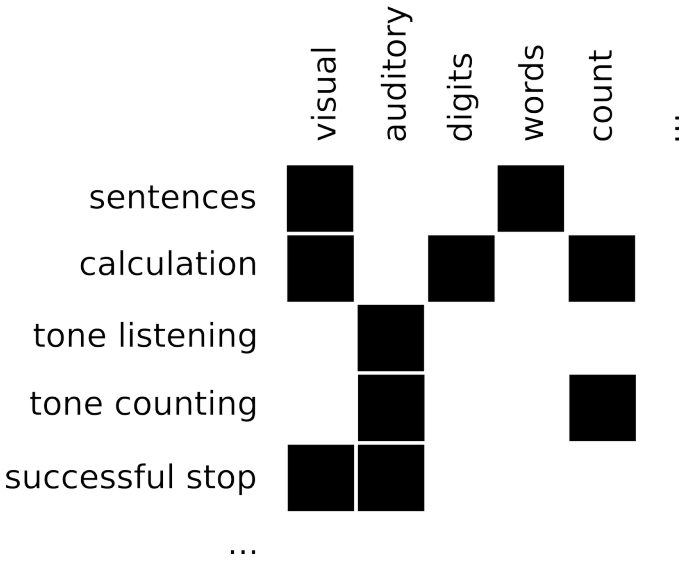
Outline

- Identification issues
- Large-scale decoding
- Experiments on vision

Decoding beyond pre-defined categories



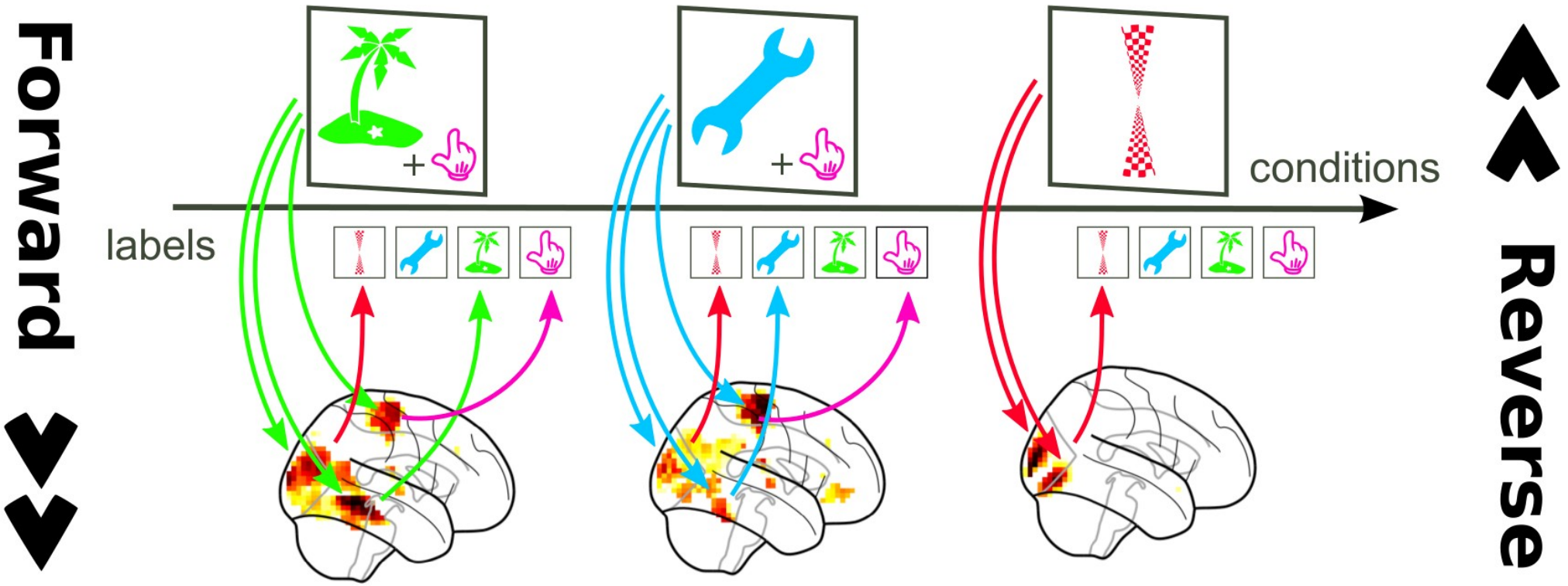
What is this brain doing?



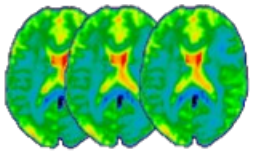
Cognitive tasks can be represented as sets of concepts

Identify the brain substrate of these concepts across experiments

Decoding beyond pre-defined categories



Predictive modeling across datasets



4TB resting-state data

HCP900

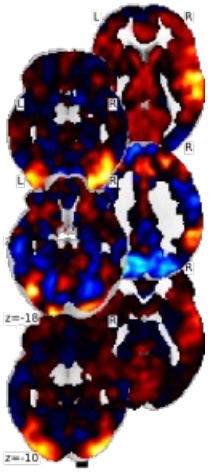
OpenfMRI

HCP

Camcan

Brainomics

...

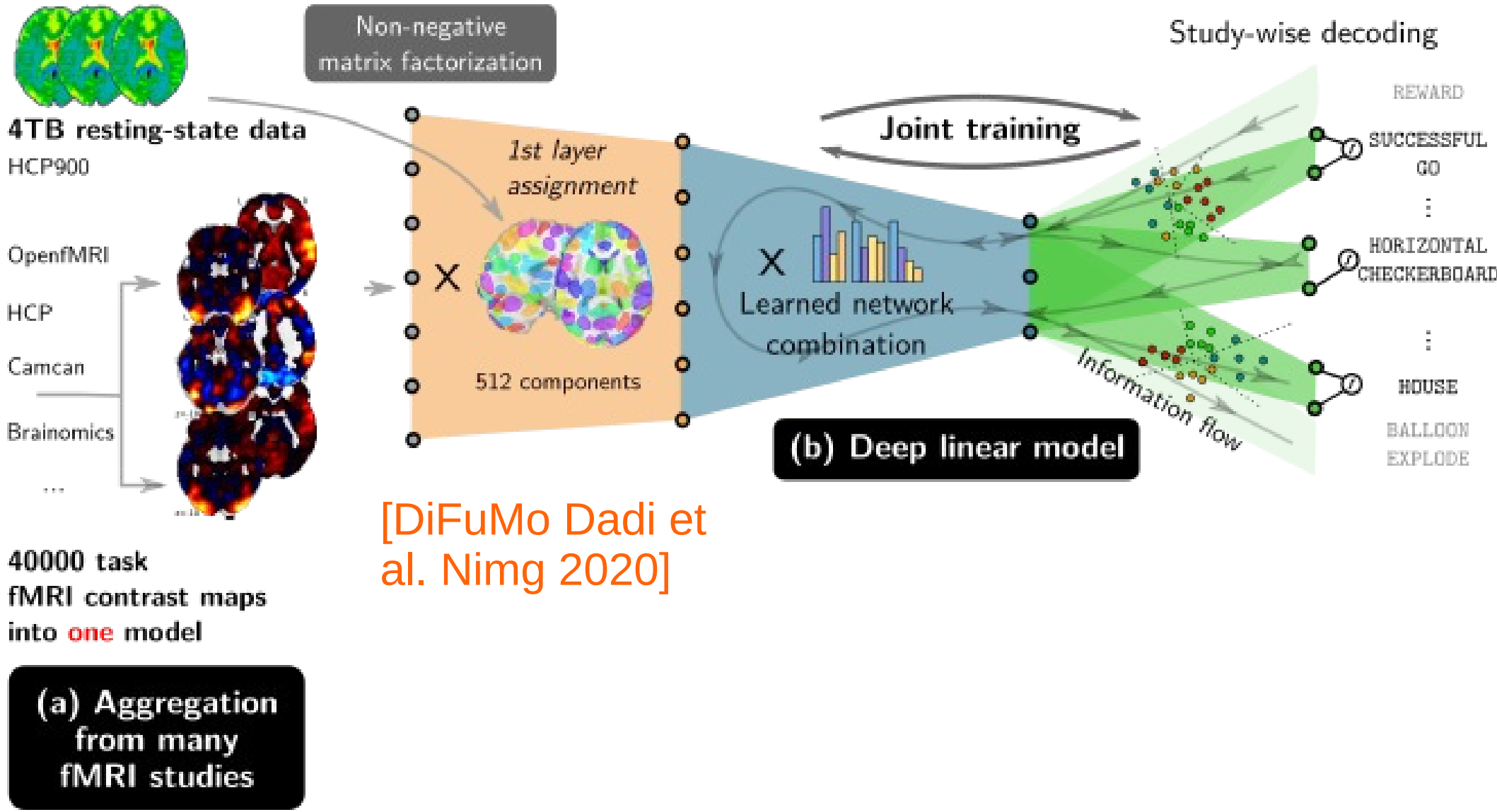


40000 task fMRI contrast maps into one model

(a) Aggregation from many fMRI studies

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

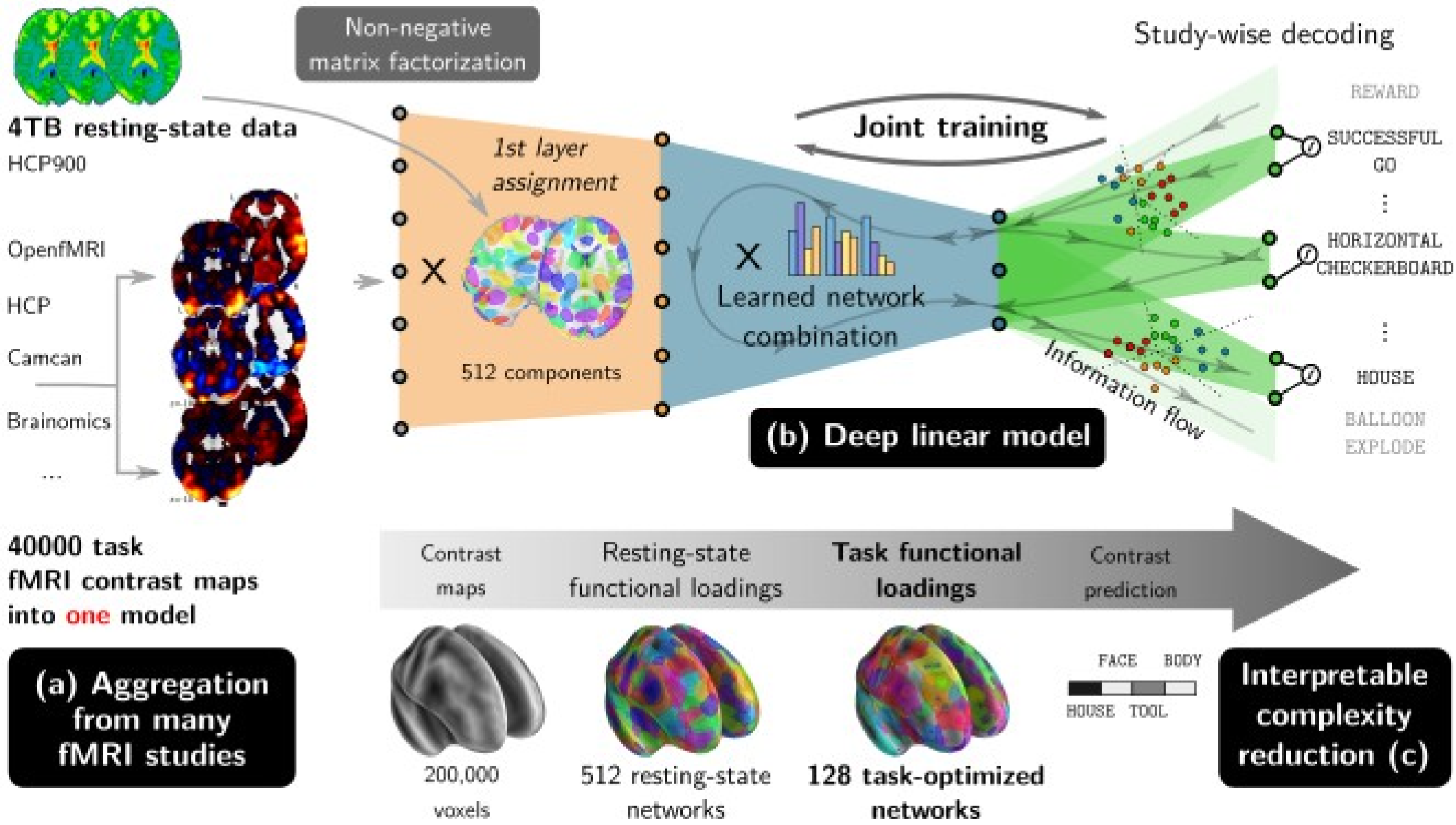
Predictive modeling across datasets



[DiFuMo Dadi et al. Nimg 2020]

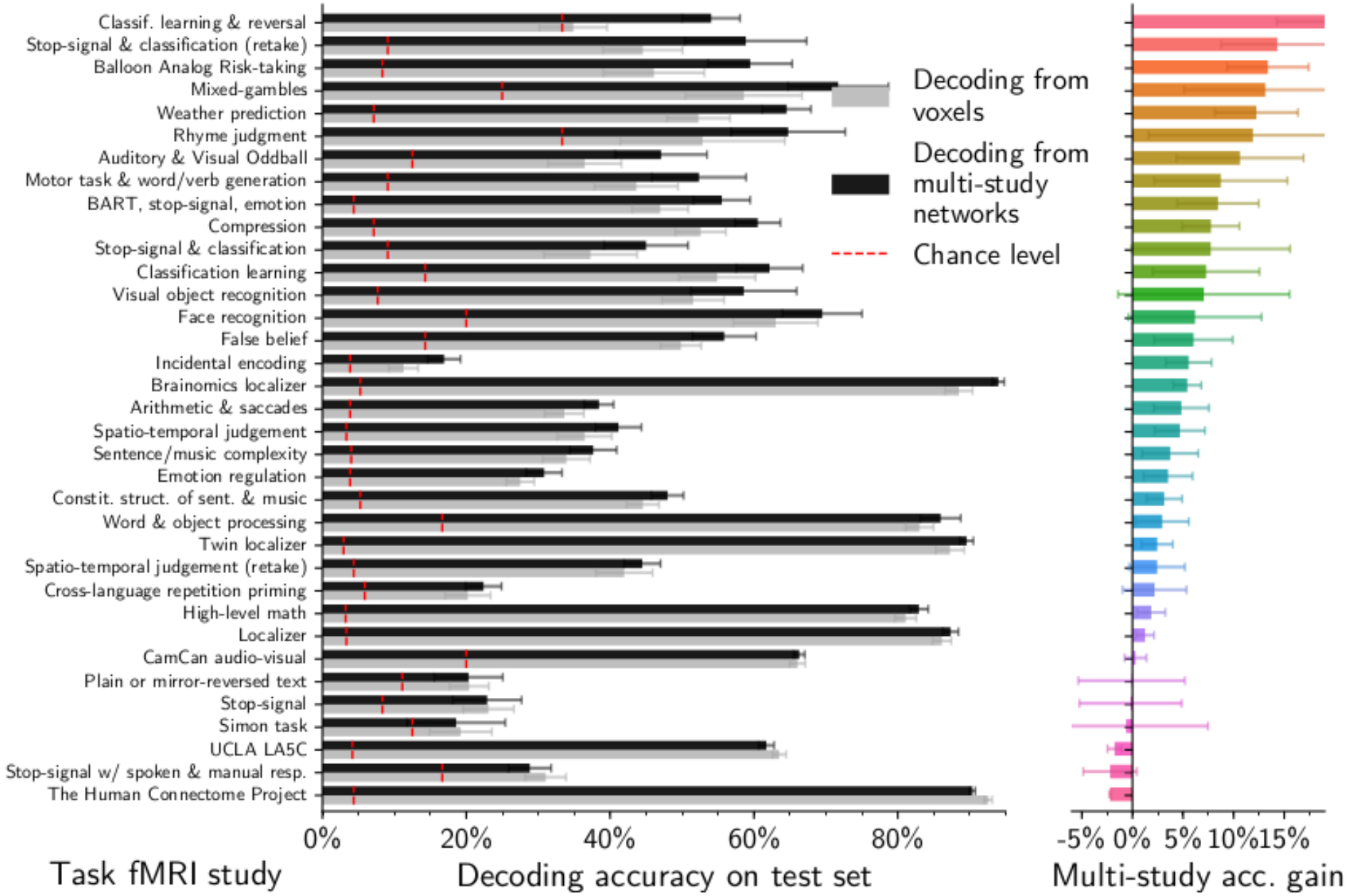
[Bzdok et al. Plos Comp Biol 2016, Mensch et al NIPS 2017 PCB 2021]

Predictive modeling across datasets

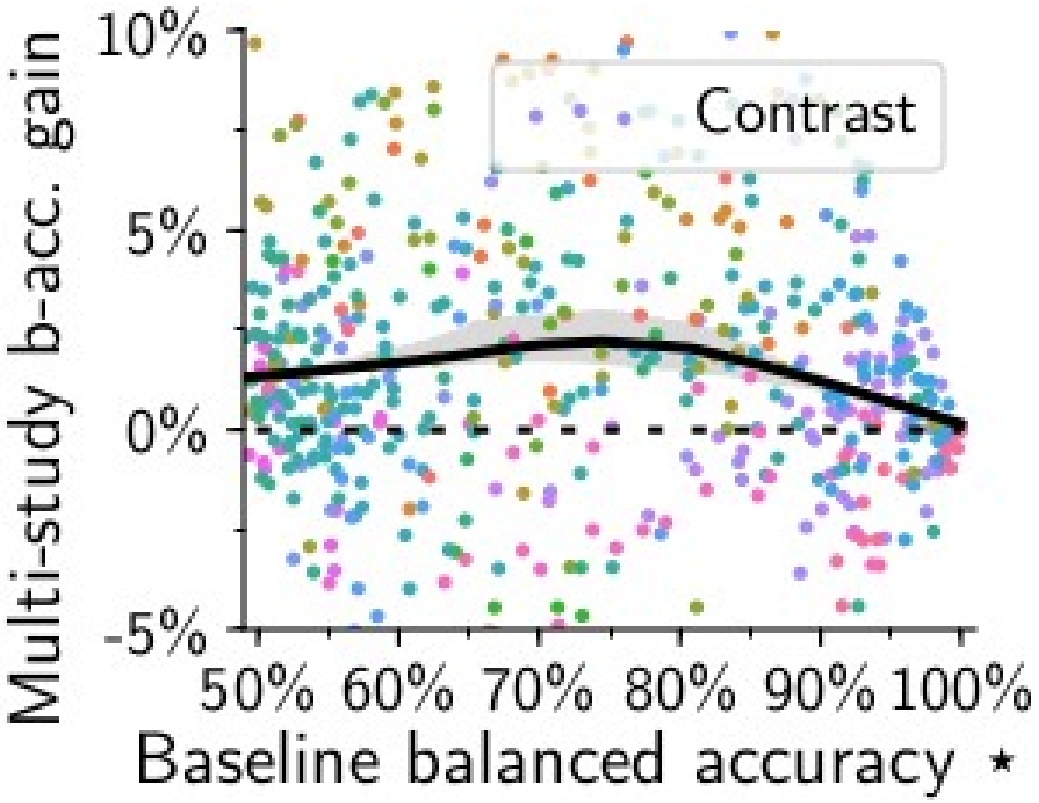


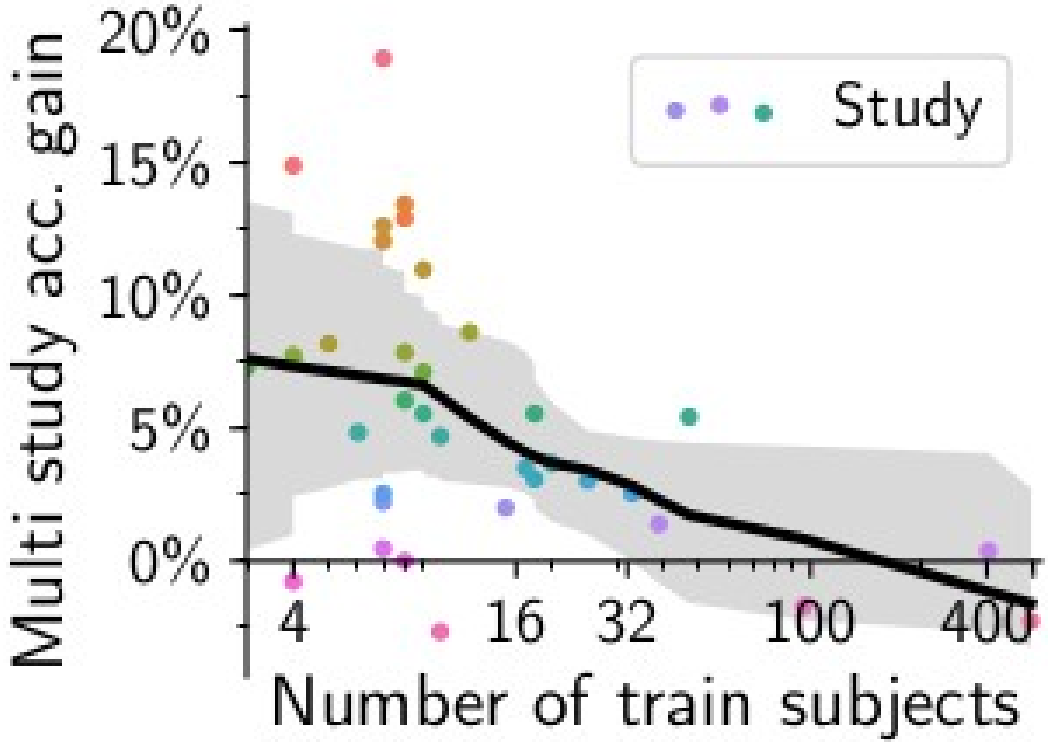
[Bzdok et al. Plos Comp Biol 2016, Mensch et al NIPS 2017, PCB 2021]

Transfer learning

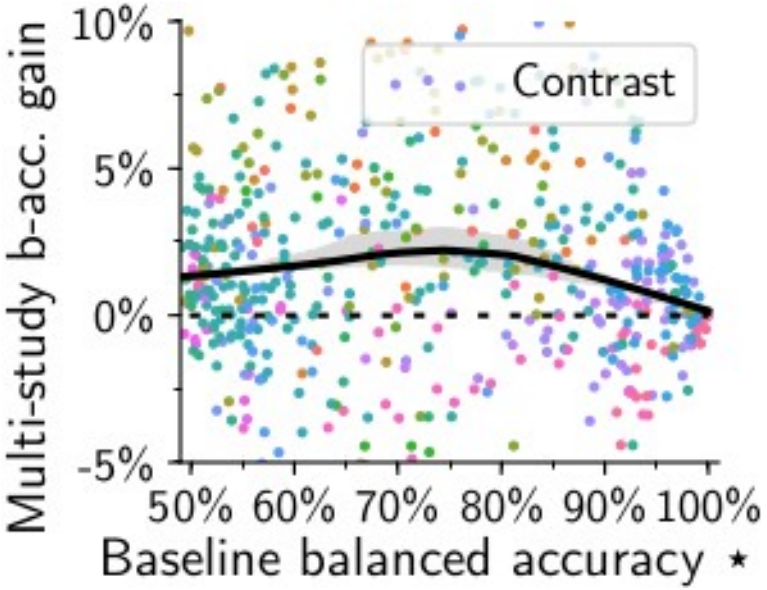


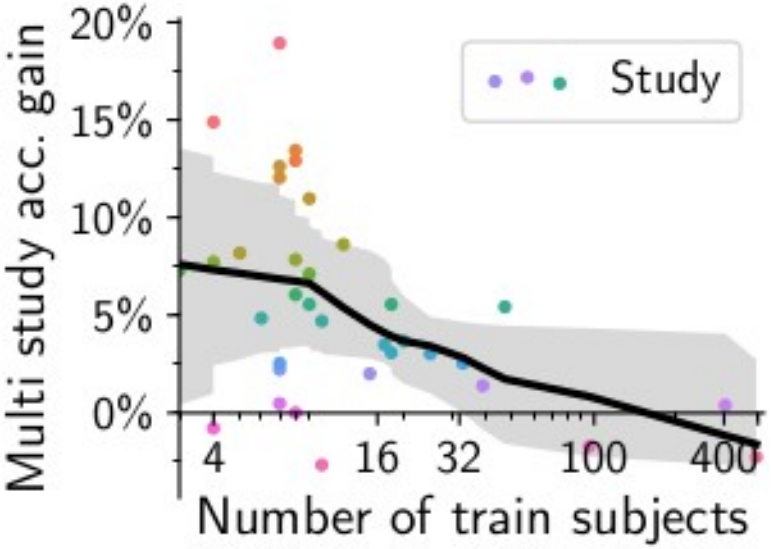
Transfer learning

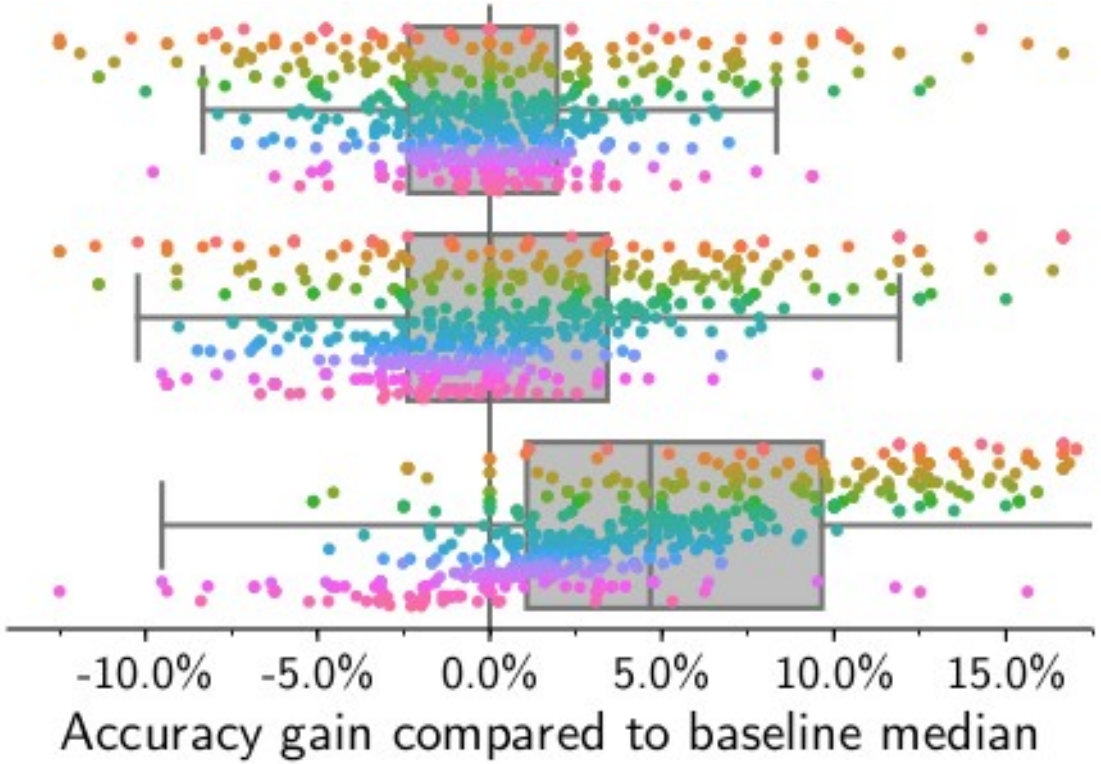




Transfer learning





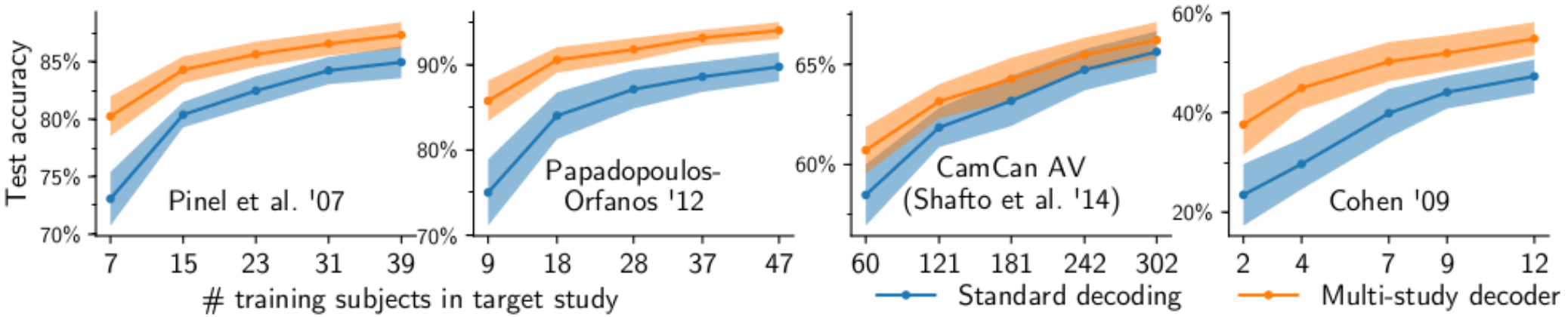


Standard decoding from voxels

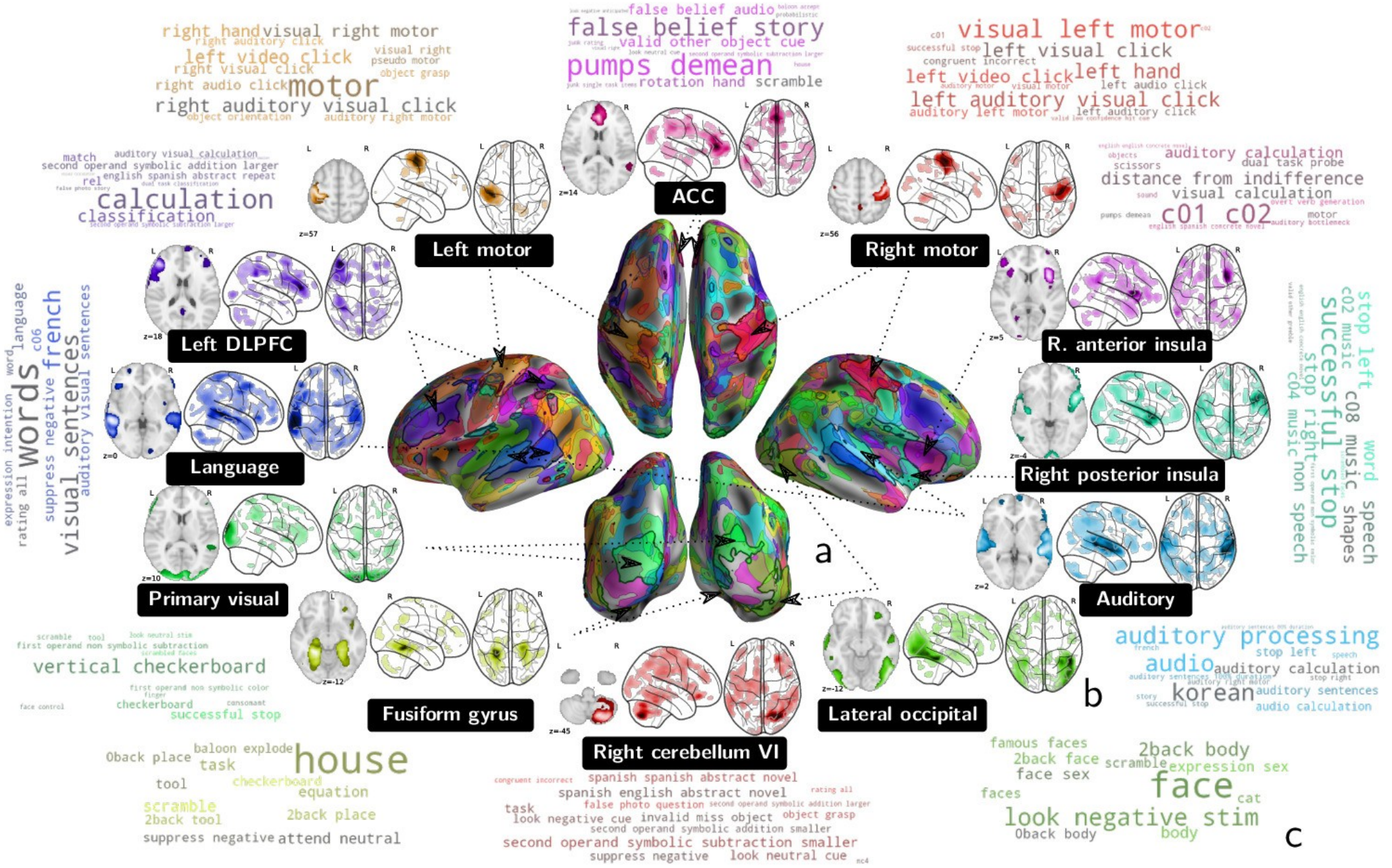
Decoding from functional networks

Decoding from multi-study task-optimized networks

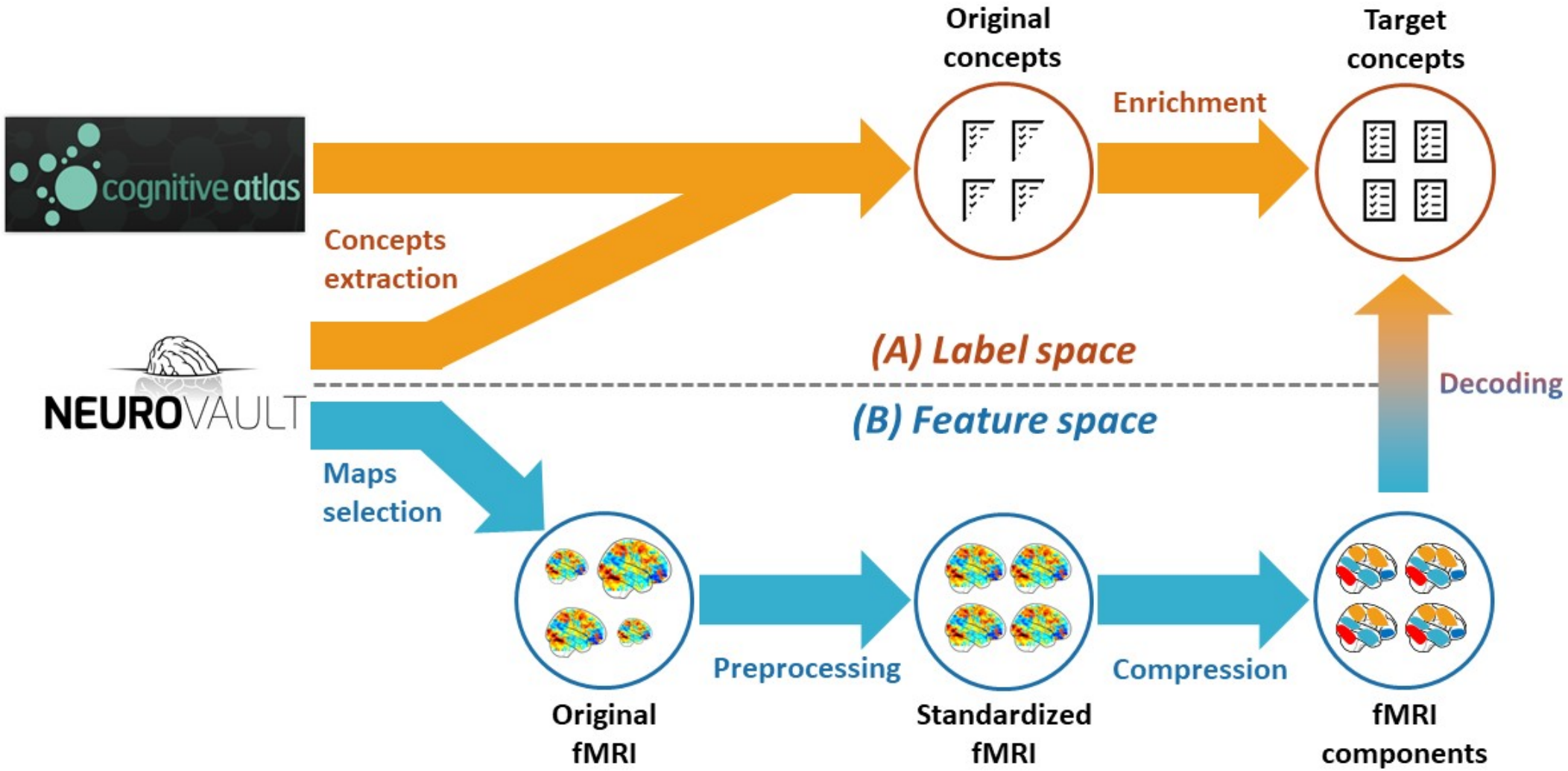
Small studies benefit more than large studies



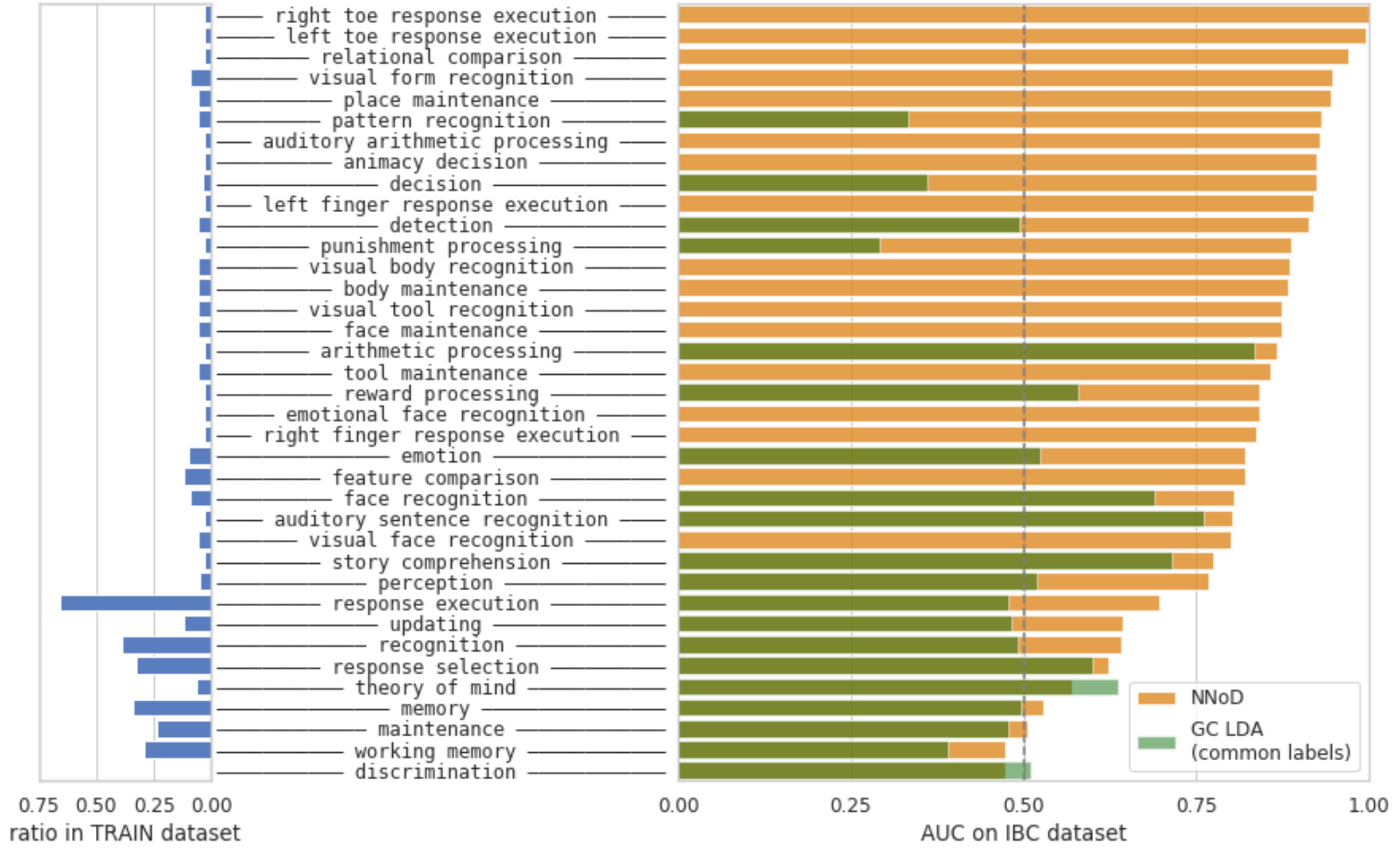
Resulting atlas



From multi-study to universal decoder



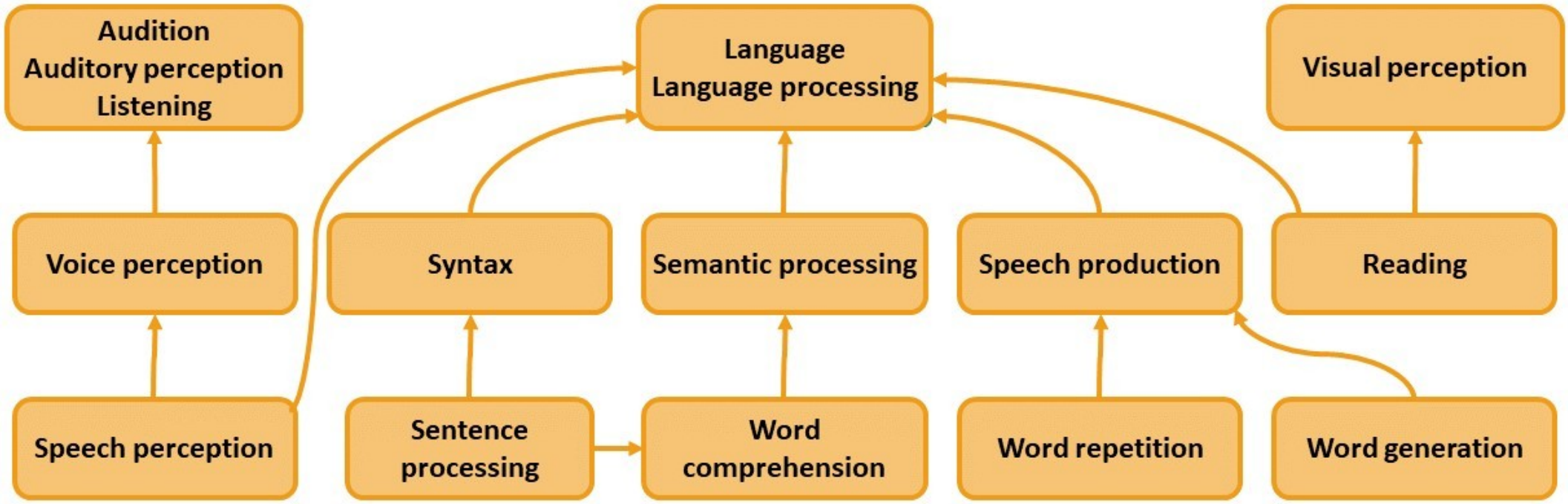
Results (naive approach)



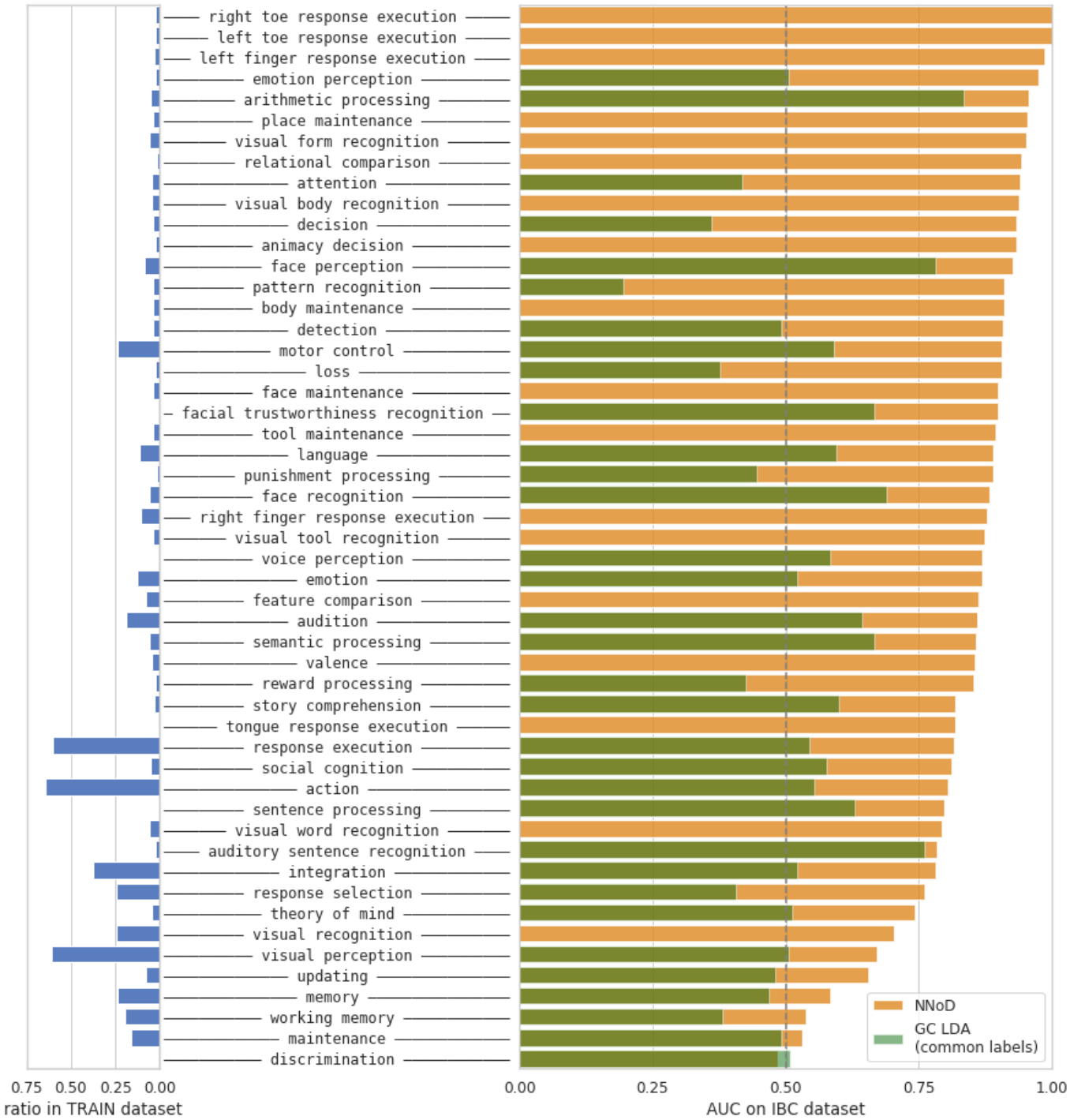
Fixing labels

Problem:
synonims, false negatives (missing annotations)

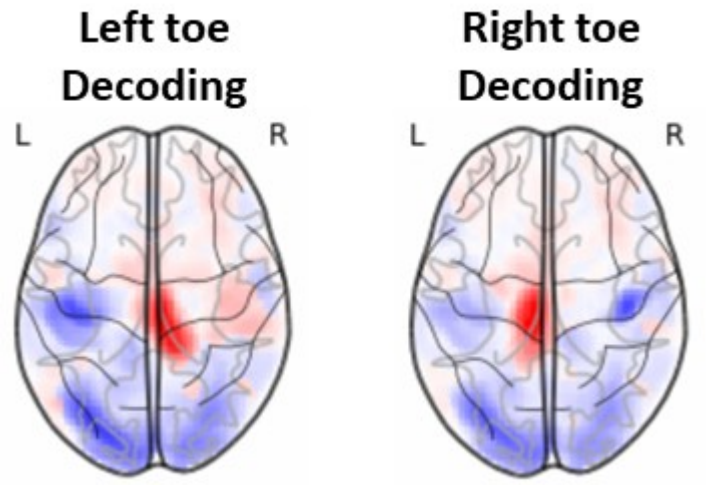
→ **Simple rules to impute labels:**



Results (2): label imputation boosts accuracy

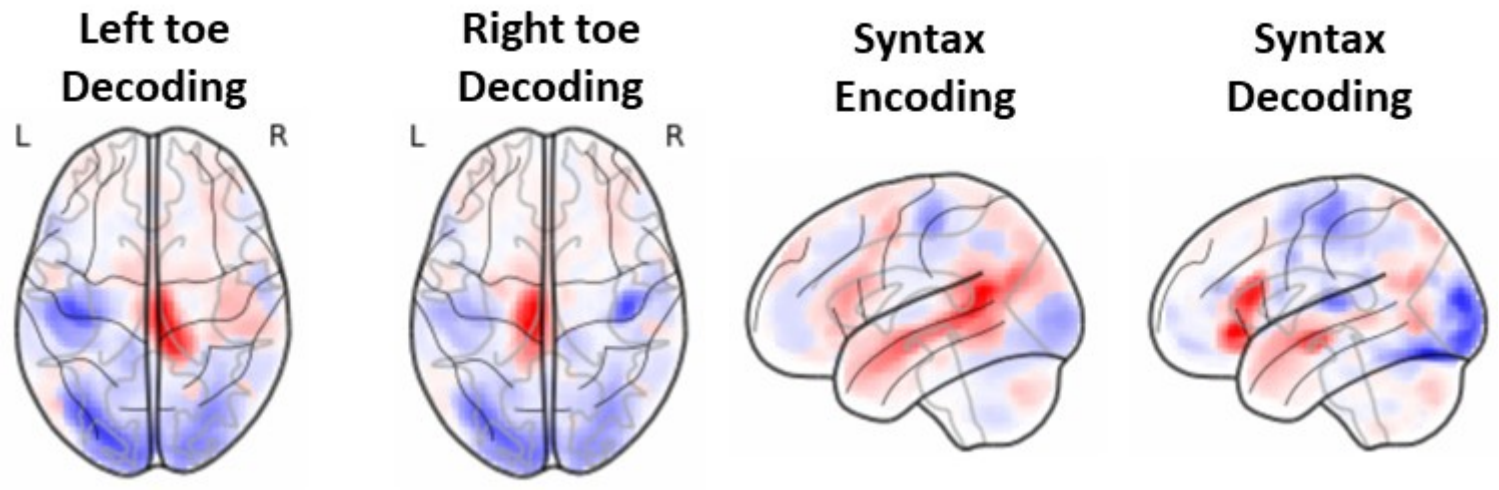


Open the box



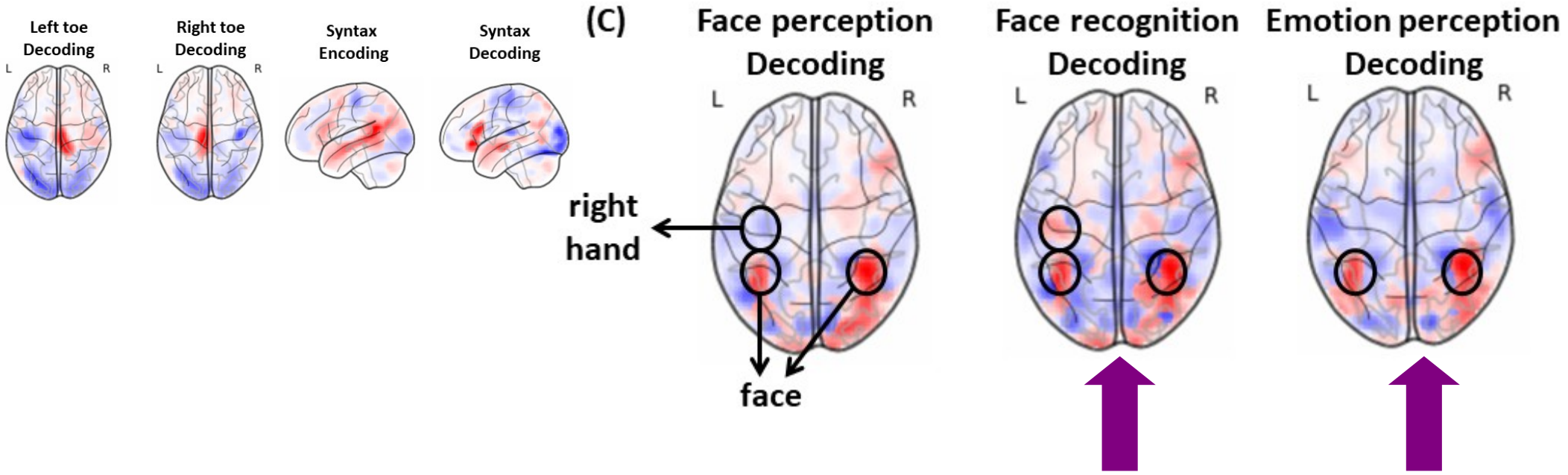
Non-controversial case

Open the box



decoding > encoding

Open the box



Experimental biases

[Menuet et al. In rev]

Outline

- Identification issues
- Large-scale decoding
- Experiments on vision: encoding and decoding visual stimuli

Encoding visual stimuli

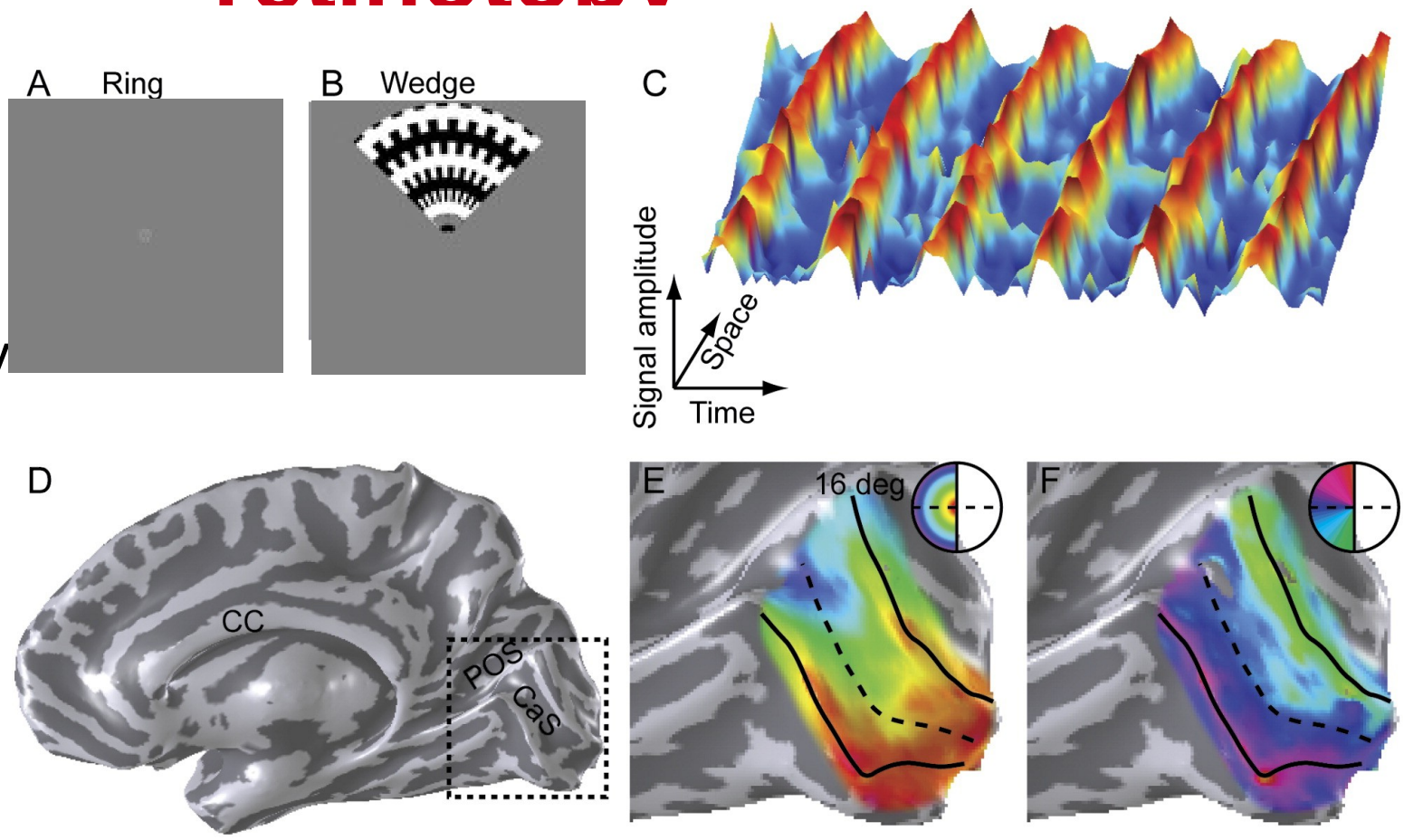
- Brain activation encodes some information about the stimuli
- Concept of receptive field
  - Location-specific (depends on where you are on the cortex)
  - Feature-related (position, contrast, speed, color...)

$$\mathbf{y} = \phi(\mathbf{X})\beta + \varepsilon$$ e.g. φ = wavelet transform

  - Extension of brain mapping to non-linear features
  - Hypothesis-testing framework

Accessing visual encoding w/ fMRI: retinotopy

- Brain voxels respond to location-specific contrasts in early visual areas
- Traveling-wave stimuli are used to represent the mapping of the visual field to visual cortex



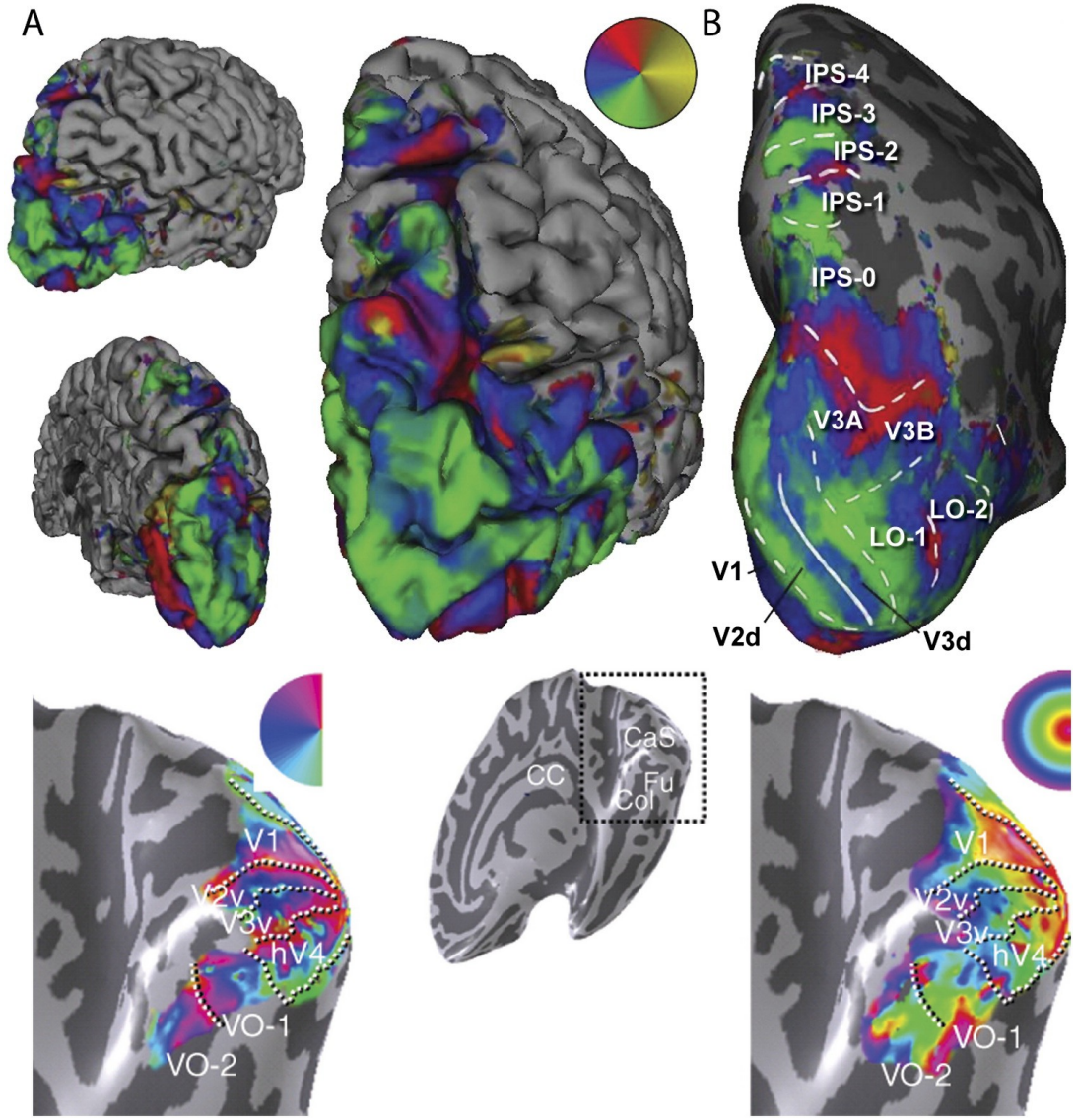
- The time (phase) of the BOLD signal peak varies smoothly across the cortical surface (space)

[Sereno et al. 1994,...,Wandell et al. 2007]

Accessing visual encoding w/ fMRI: retinotopy

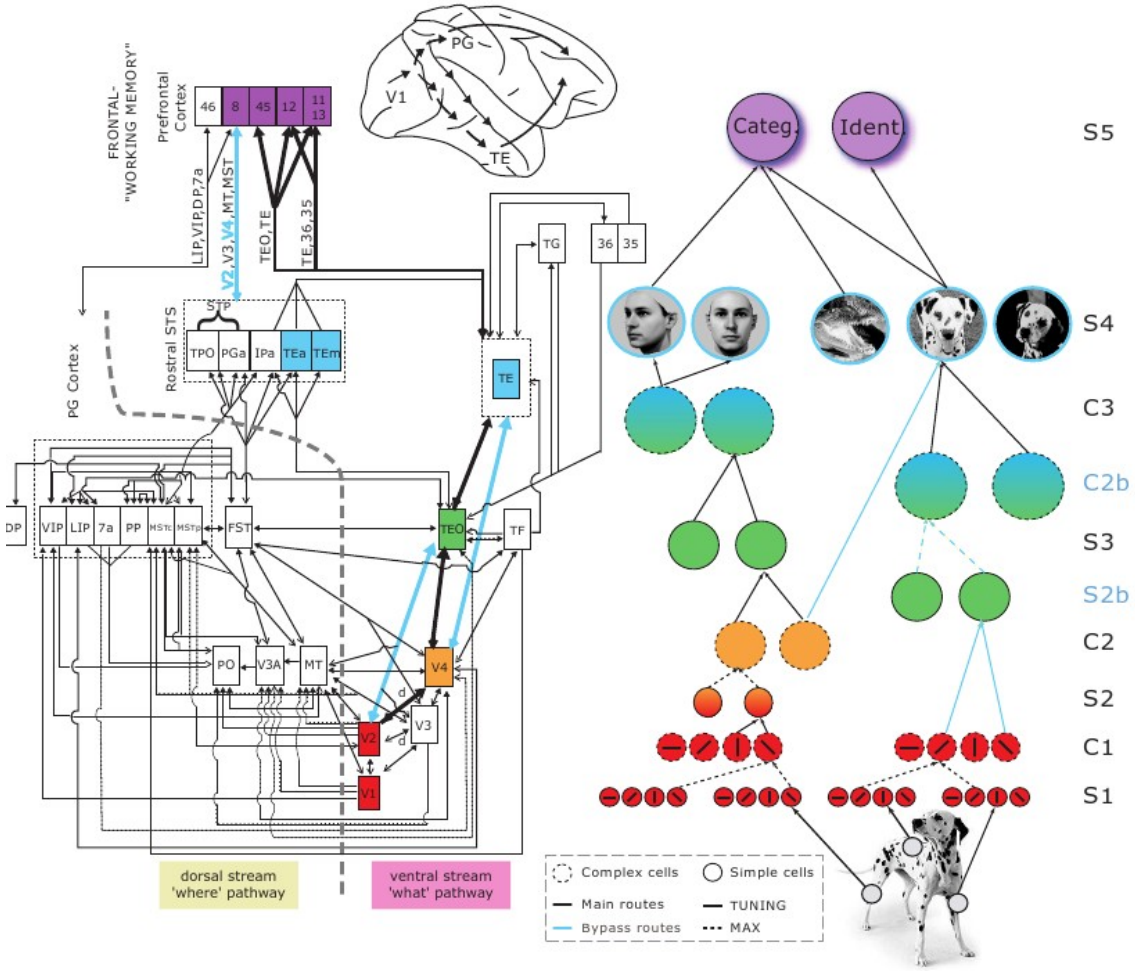
- Retinotopy provides the main access to the definition of visual areas in the cortex

  [Sereno et al. 1994, Wandell et al. 2007...]



Neurocomputational models of vision

- **Hmax** [Riesenhuber and Poggio, 1999]
- Bio-inspired but validated as computer vision tools [Serres et al. 2007]
- See also: Convolutional networks, scattering transform

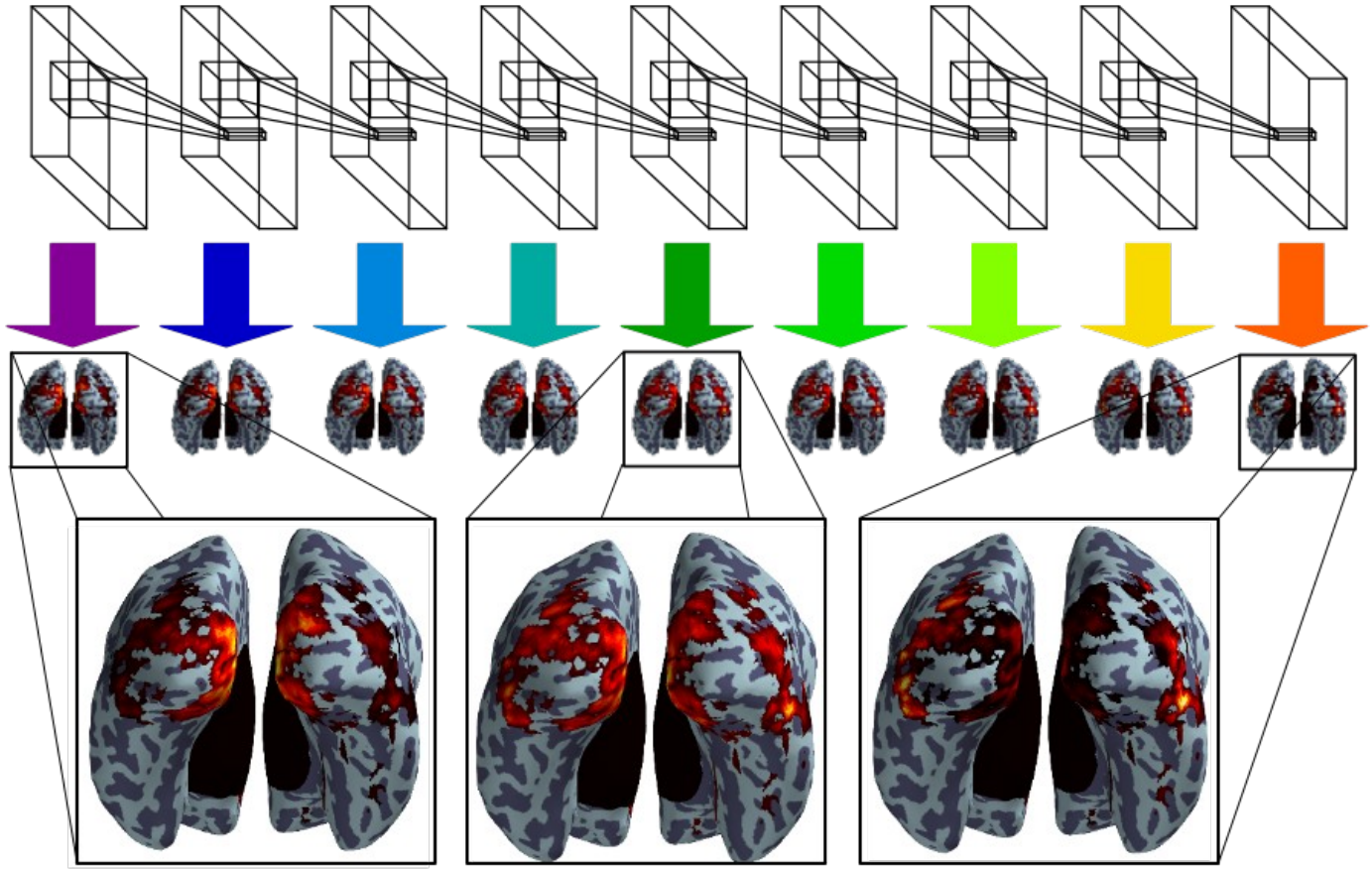


Encoding visual stimuli

Create Features

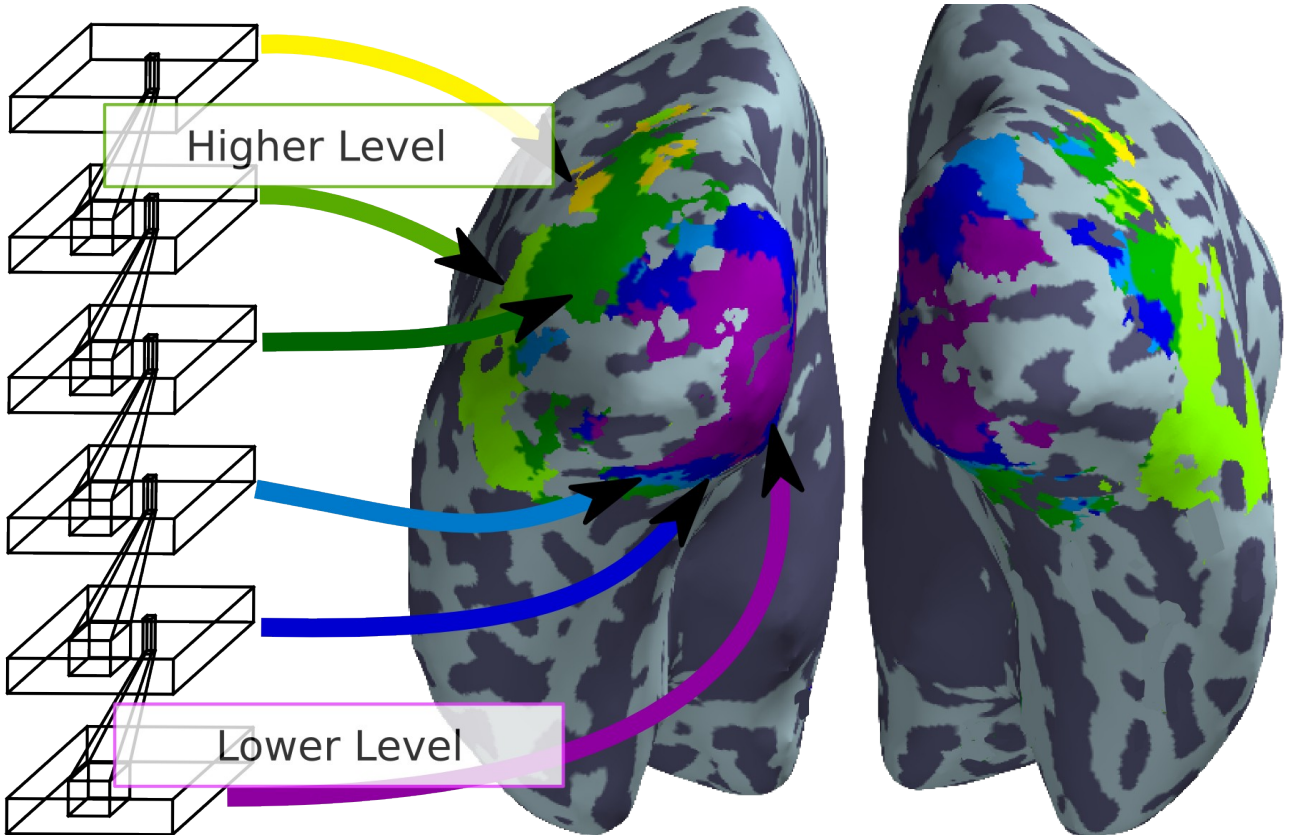
Convolution model: Feedfoward model of vision for object recognition

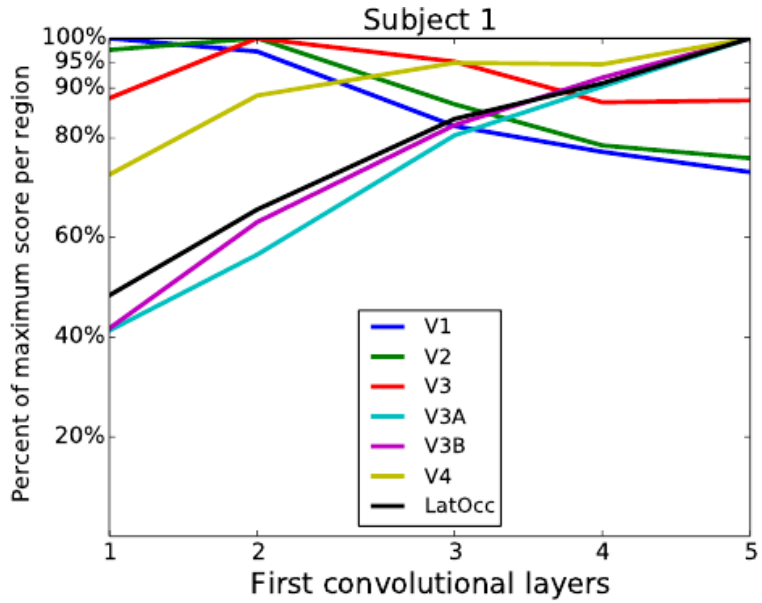
Comparison with fMRI
How well are brain regions explained by the features ?



[Eickenberg et al. Nimg 2017]

Mapping convnet layers to the cortex





The convolutional network reproduces the cortex structure !

[Cadieu et al. 2014, diCarlo et al. 2014, Güçlü et al. 2015, Eickenberg et al. Nimg 2017.]

Decoding visual stimuli

- Infer stimulus information from brain maps: inverse problem of the encoding.

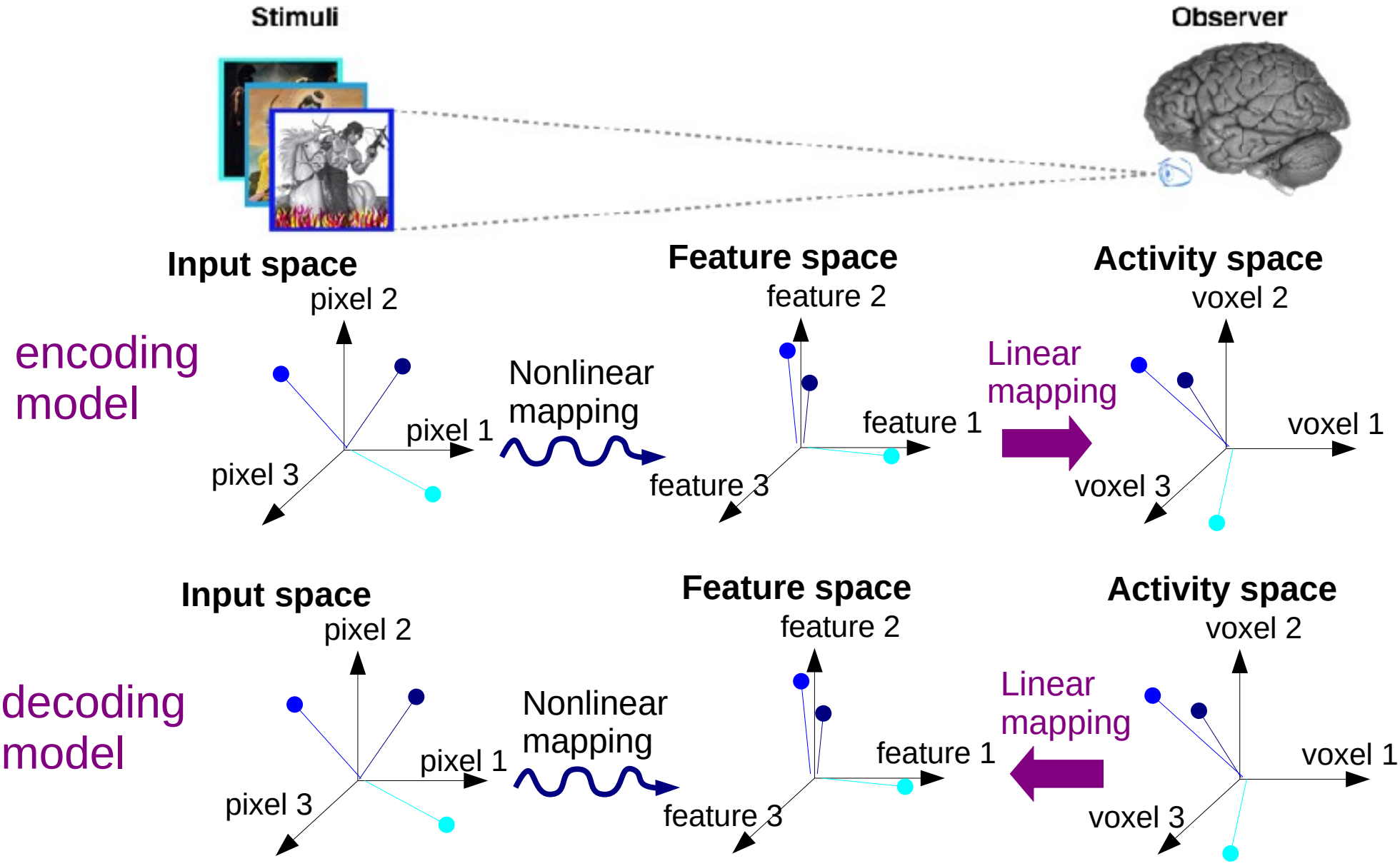
  ≈ BCI

- Detect distributed effects. Not relevant for localizing feature representations
- Requires high-dimensional linear models

$$\phi(\mathbf{X}) = \mathcal{L}(\mathbf{Y}\mathbf{w}) + \eta$$

$\mathcal{L}$ Possible nonlinearity (e.g. sign function)

  - Suffers from curse of dimensionality



[Naselaris et al. NeuroImage 2009]

Reconstructing a perceived stimulus

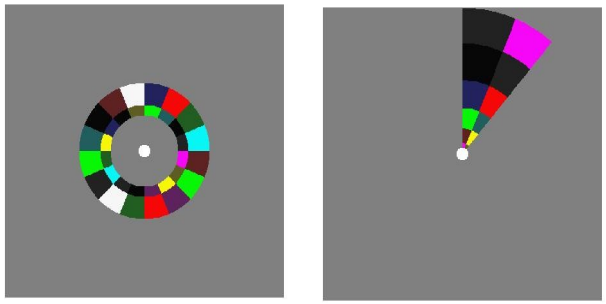
$$\phi(\mathbf{X}) = \mathcal{L}(\mathbf{Y}\mathbf{w}) + \eta$$

- Two possible approaches:
  - Inverse problem: assume that ϕ the identity / invertible
  - Identification among a finite set of samples

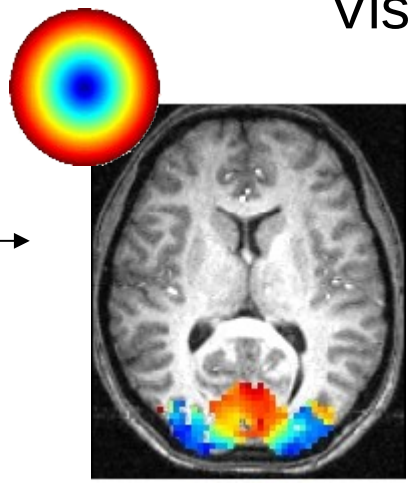
$$\hat{i}(\mathbf{y}) = \mathrm{argmin}_{i \in [1..n]} \| \phi(\mathbf{x_i}) - \mathcal{L}(\mathbf{y^T w}) \|^2$$

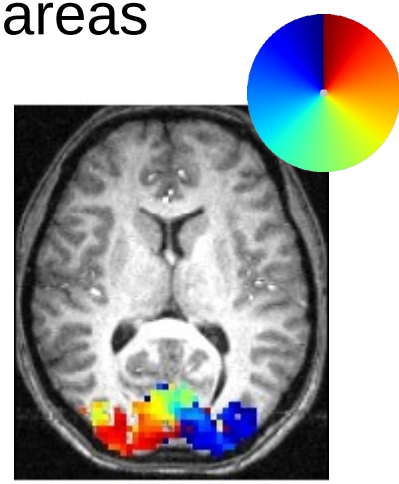
Testing vision models with fMRI: inverse retinotopy

Retina



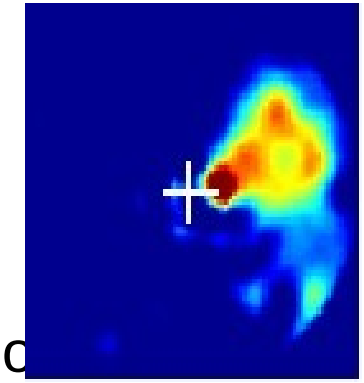
Visual areas





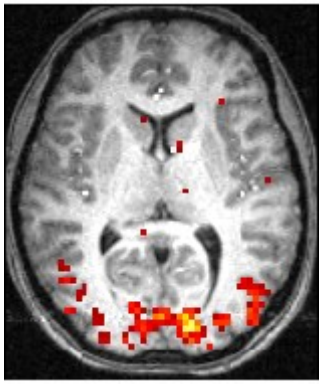
learning

Retinotopy experiment



Retinal representatio of the map

inverse inference: 'brain reading'

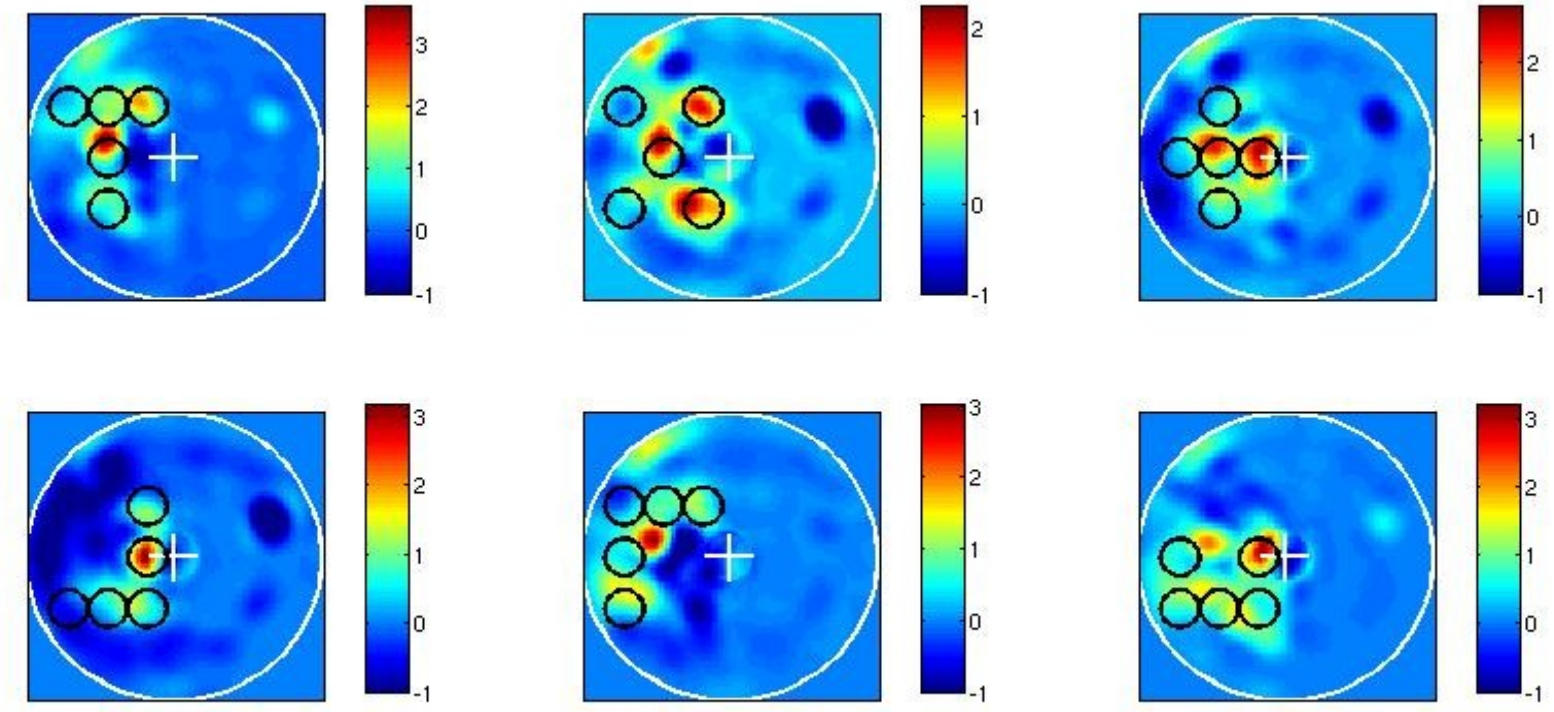


Any candidate activation map

[Thirion et al. 2006]

Inverse retinotopy : results

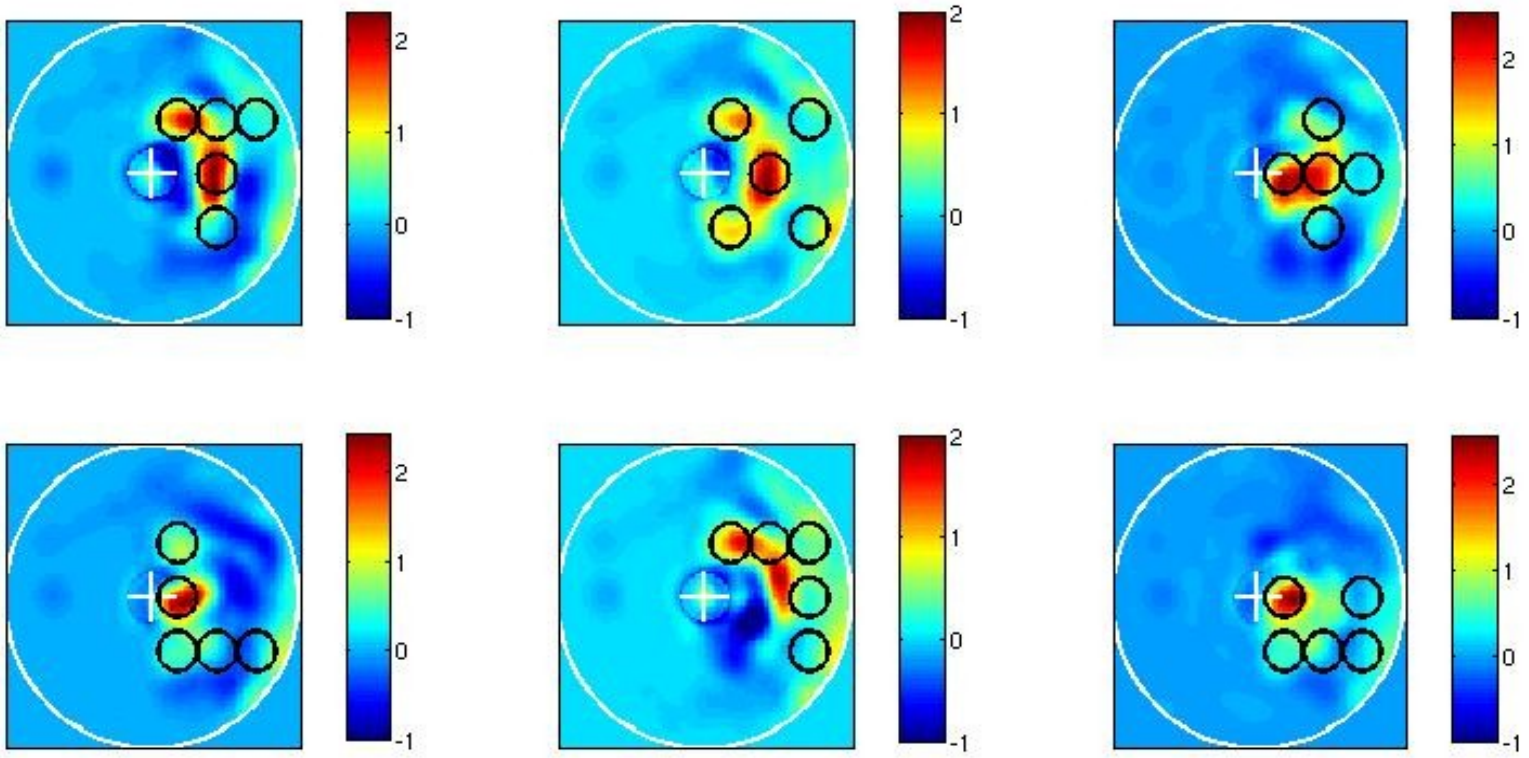
Average response for arbitrary visual patterns presented by their outline



The reconstruction ifs good enough to allow the identification of the stimuli on a trial-by-trial basis

Inverse retinotopy : results

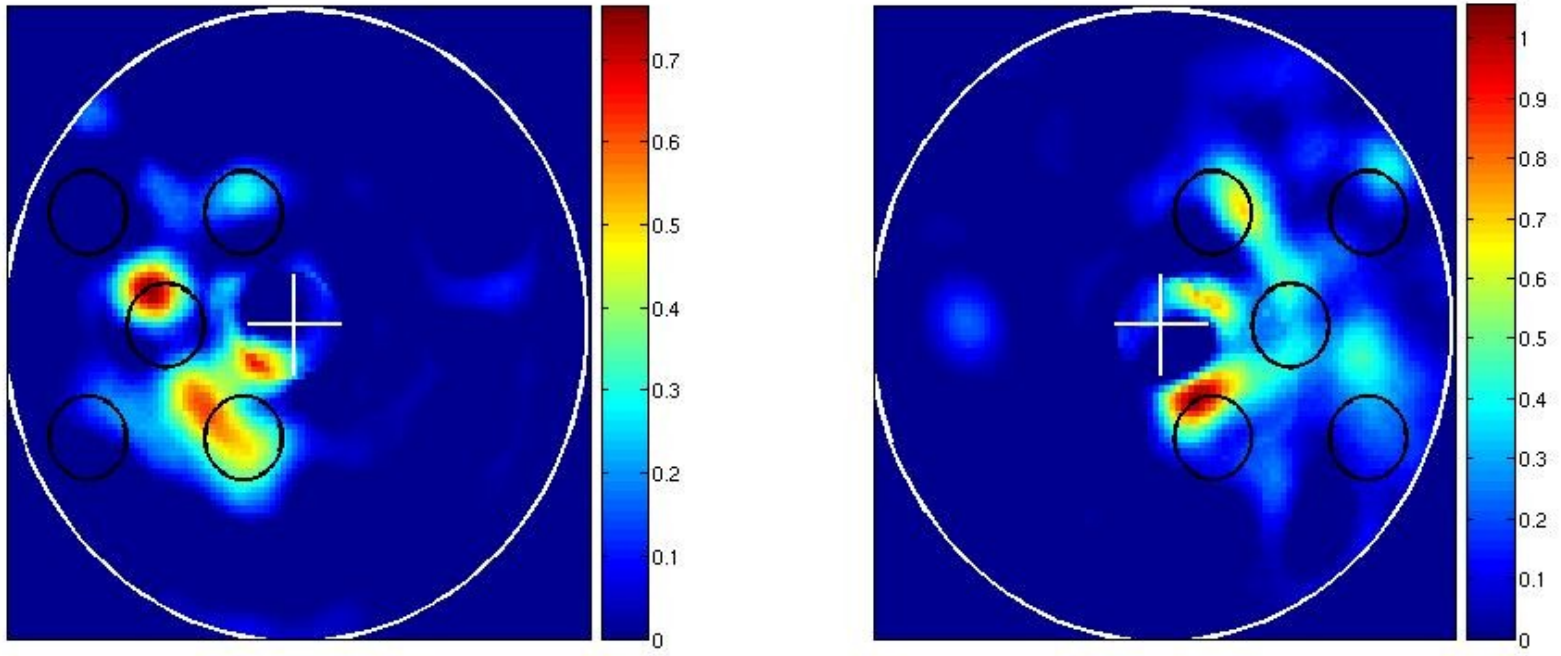
Average response for arbitrary visual patterns presented by their outline



The reconstruction ifs good enough to allow the identification of the stimuli on a trial-by-trial basis

Can we make inference about mental imagery ?

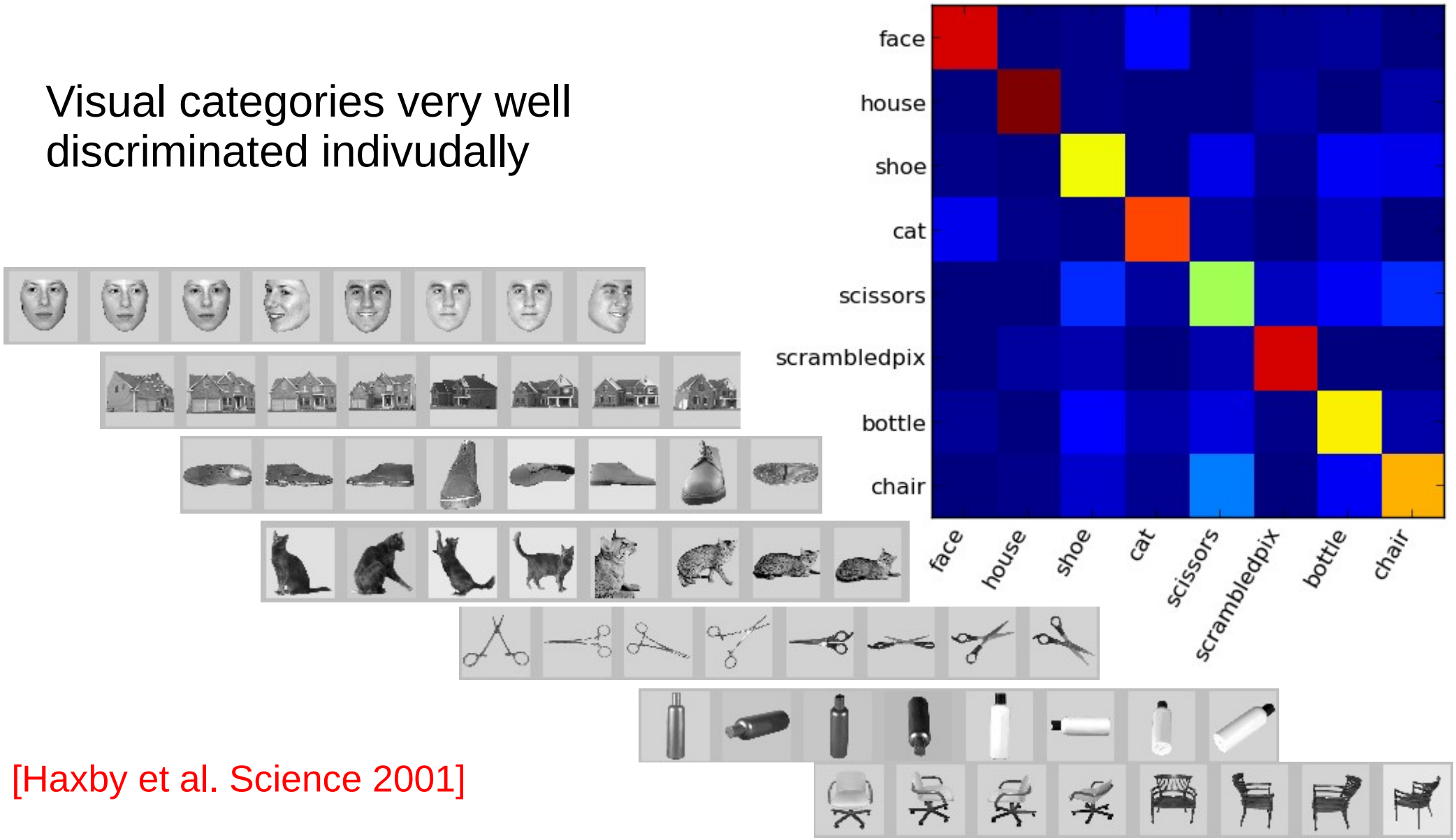
Reconstruction of a pattern that was imagined by a subject
Note: the imagined pattern was was disclosed after the scanning session



The true pattern was predicted in 5/16 hemispheres (P<0.05)
Trial-by-trial classification was successful in 5/16 cases.

Decoding visual categories

Visual categories very well discriminated indivudally

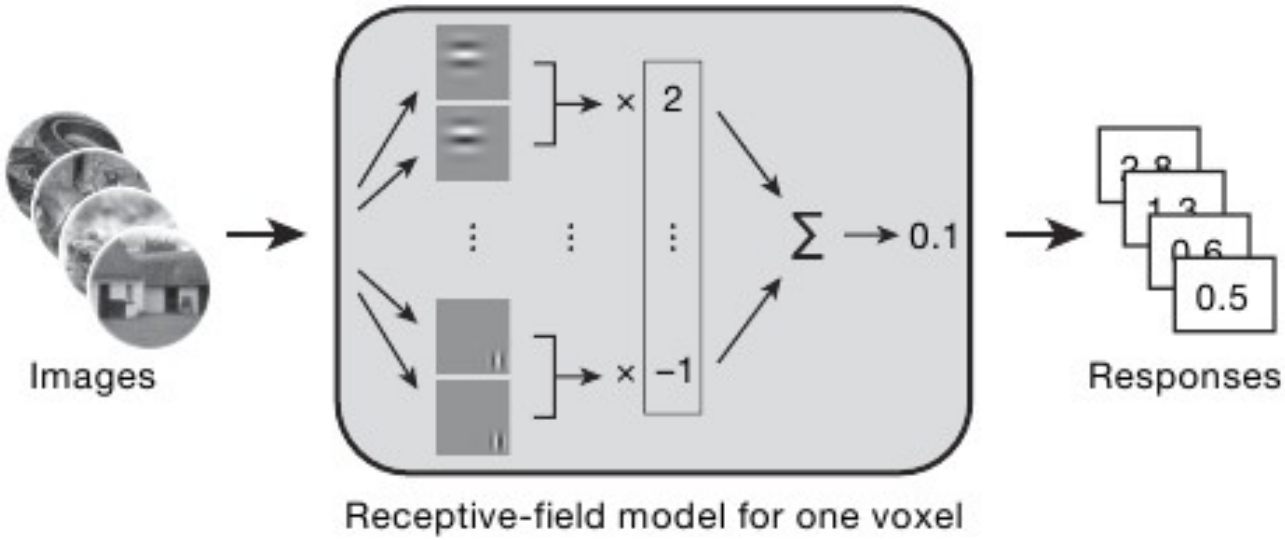


[Haxby et al. Science 2001]

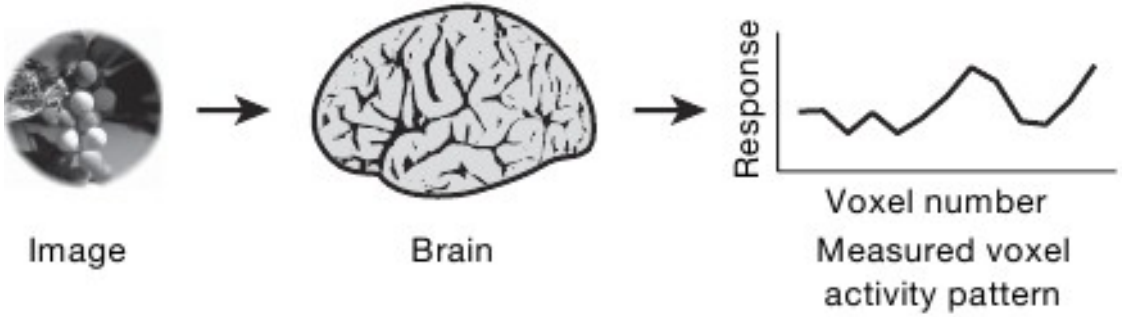
Combining position and orientation

[Kay et al. Nature Neuroscience 2008]
predict which image has been observed by the subject

Step 1: model the response in each voxel based on a training set (1750 images)



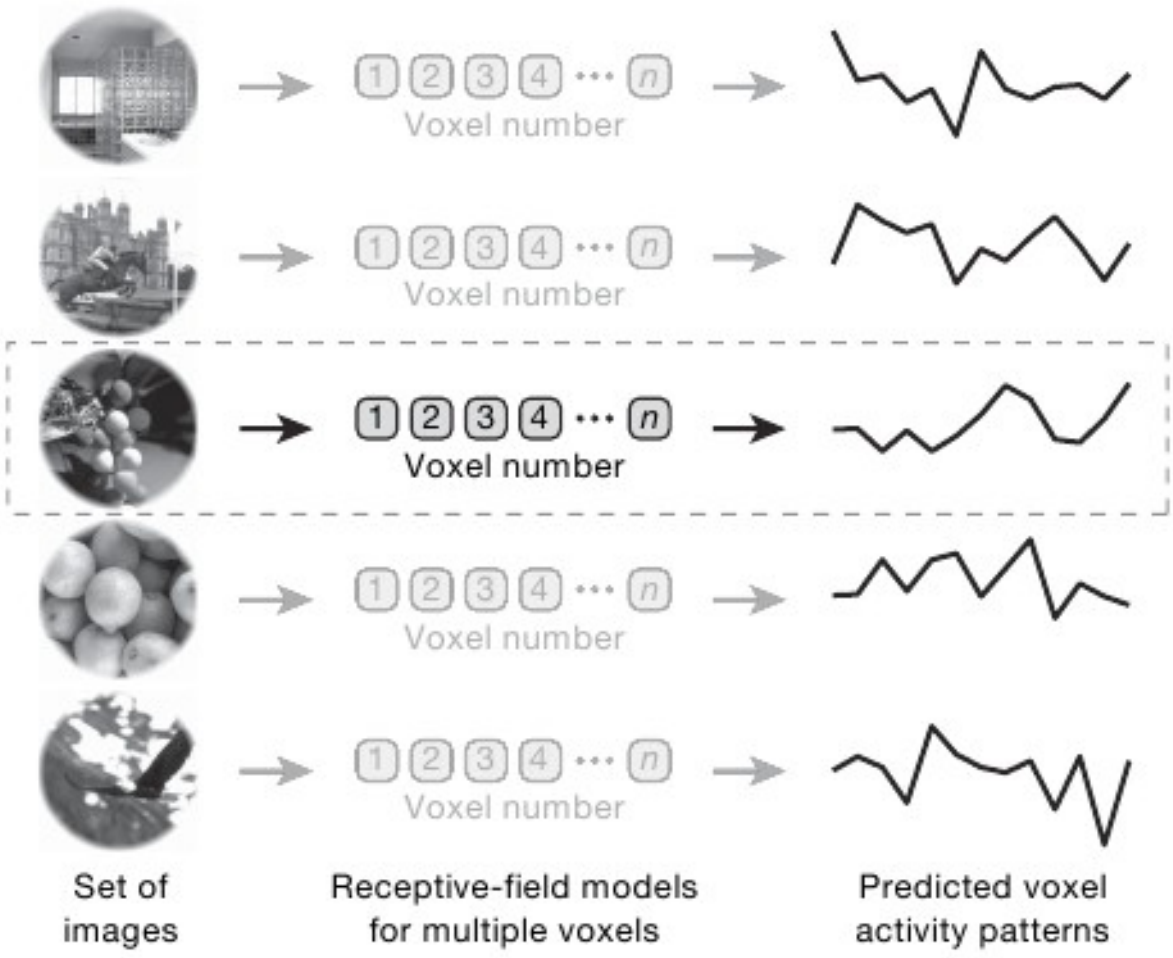
Step 2: measure brain activity for a test image



Combining position and orientation

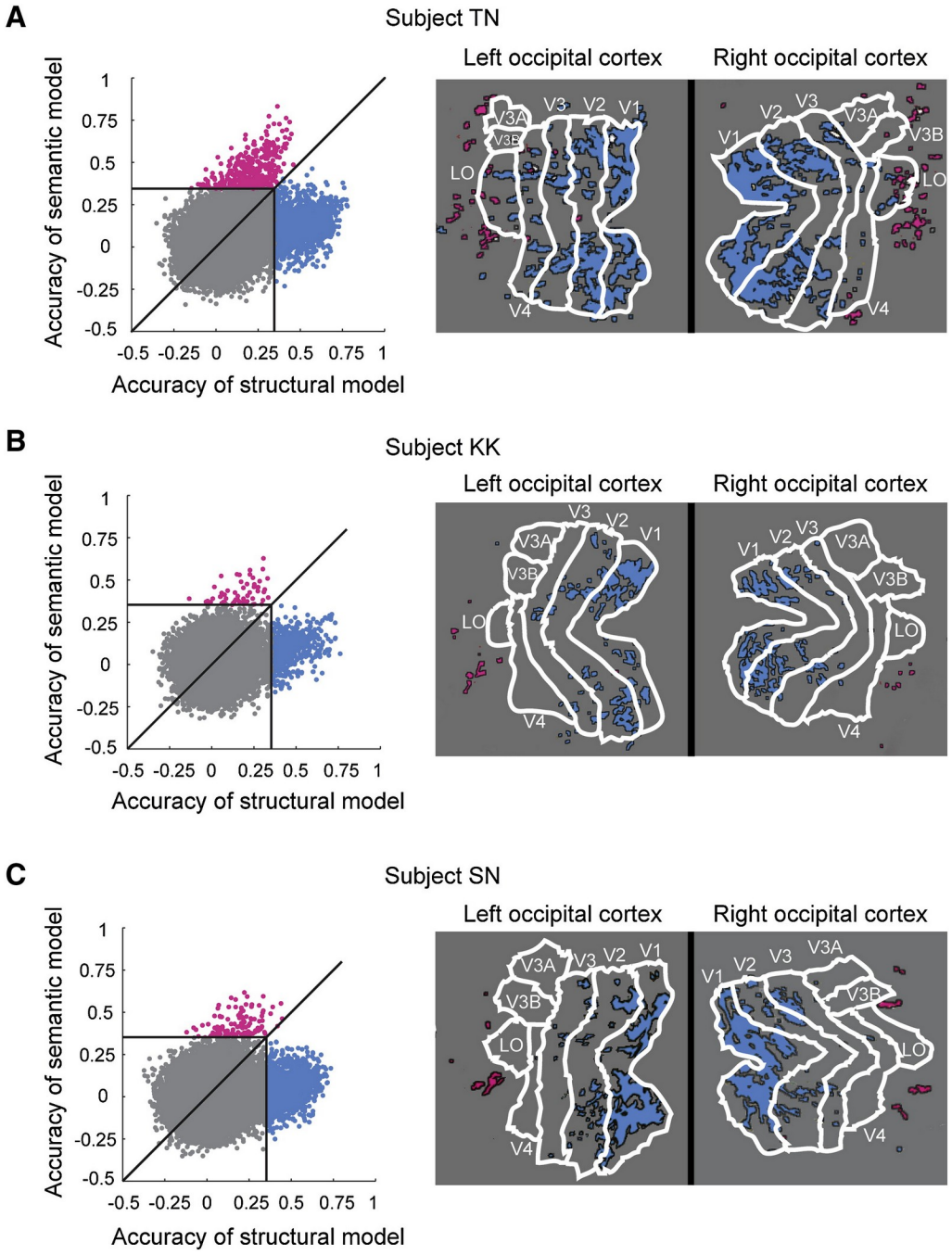
Step 3: among 100 possible images, select the image whose predicted brain activity is most similar to the measured brain activity

accuracy: 92%



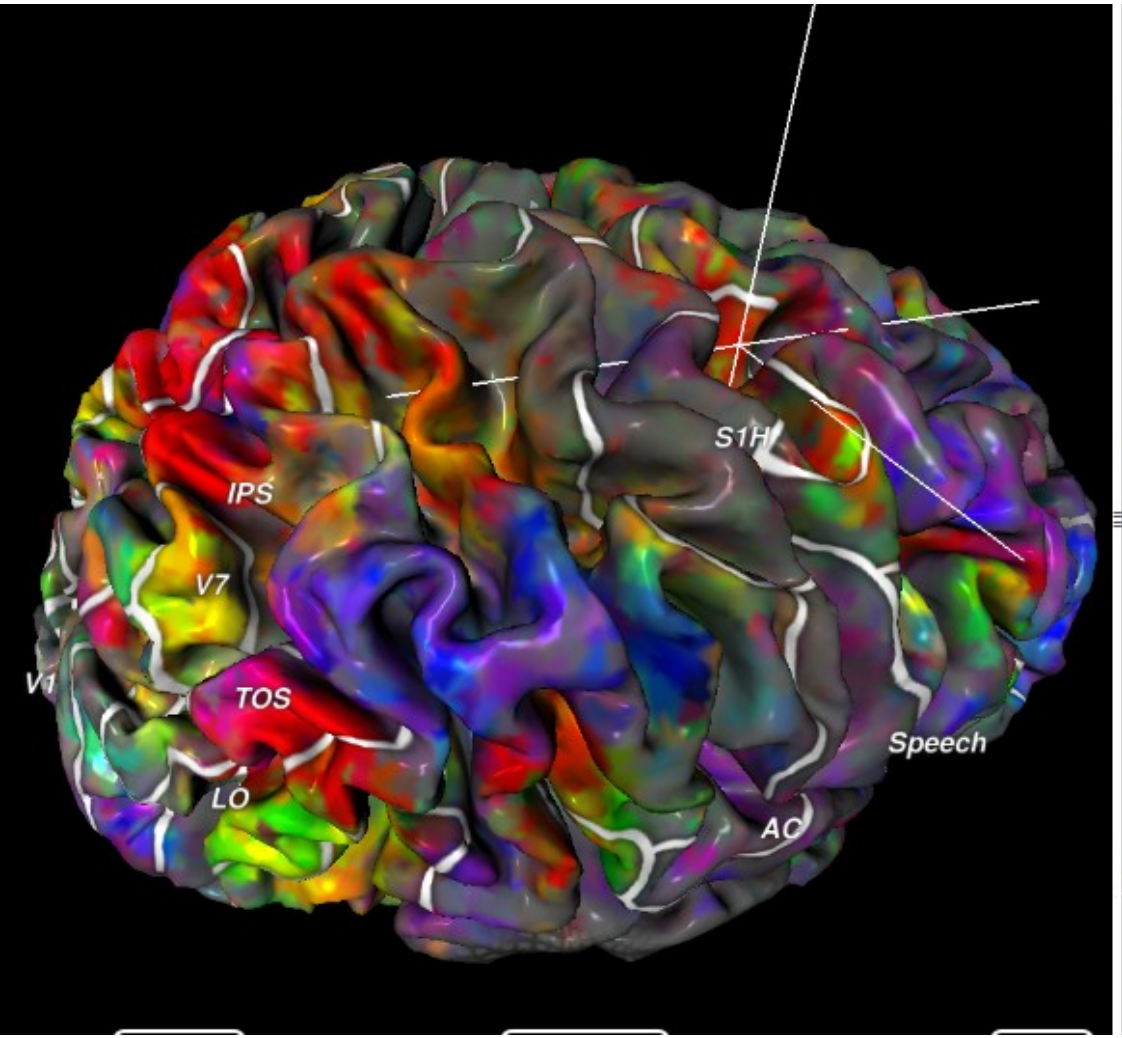
- Parallel semantic encoding increases prediction accuracy
- "obvious" mistakes are avoided
- Relative weighting of low-level / high-level features requires careful design

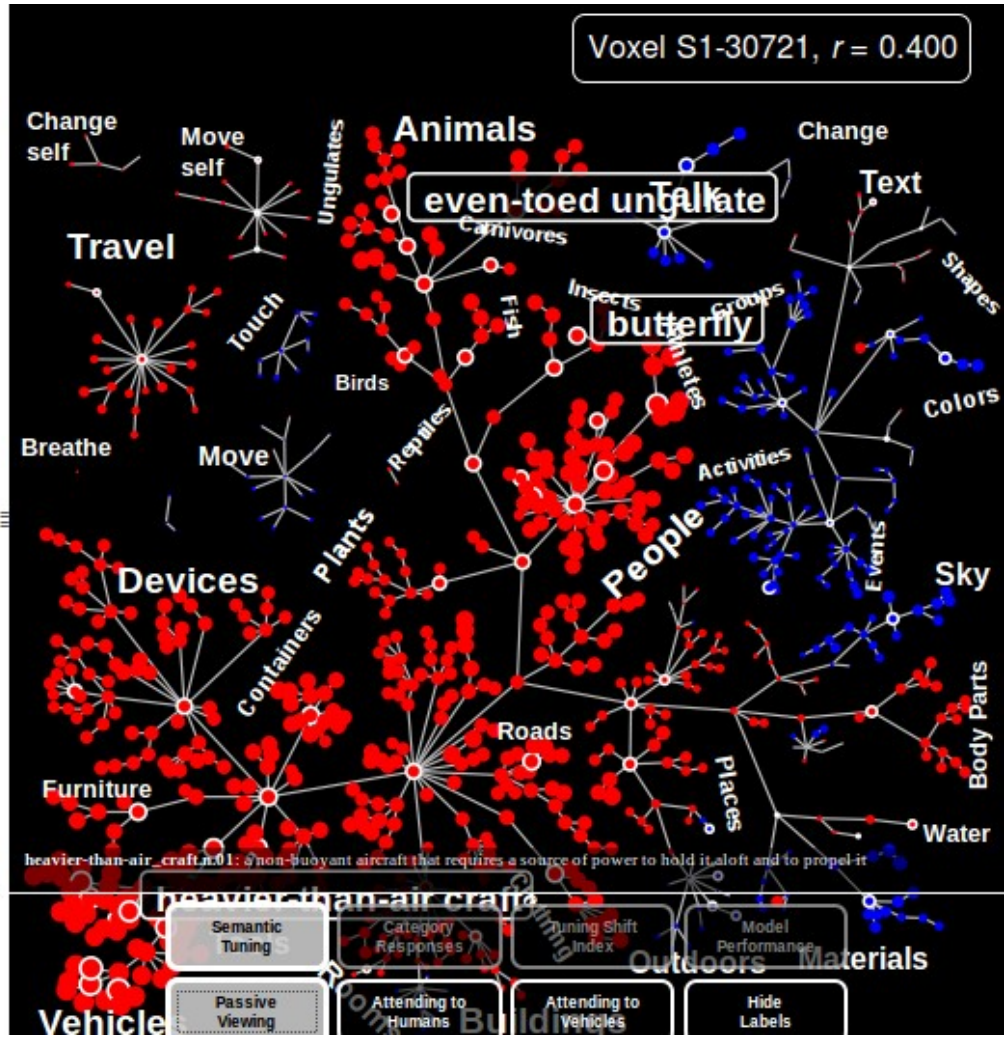
[Naselaris et al. 2009]



- Not surprisingly, the low level and high level features are encoded in non-overlapping regions of the visual pathway

Mapping the semantic space





http://gallantlab.org/brainviewer/cukuretal2013/

Associating brain regions with categories; modulation of attention

Do it yourself !

http://nilearn.github.io/