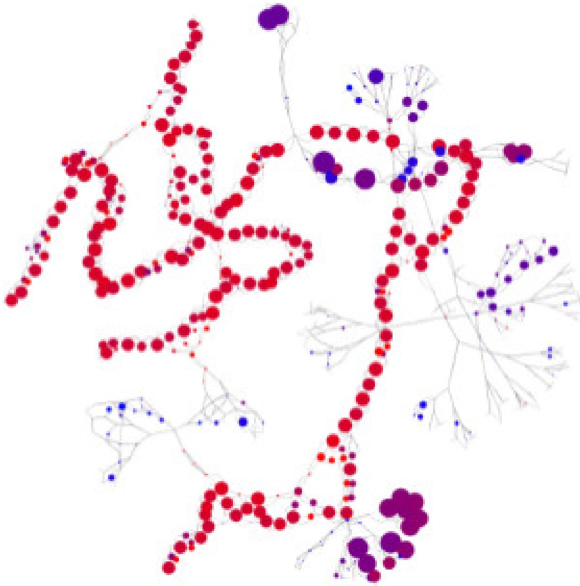
Toward rigorous e-sciences: High-dimensional statistical inference

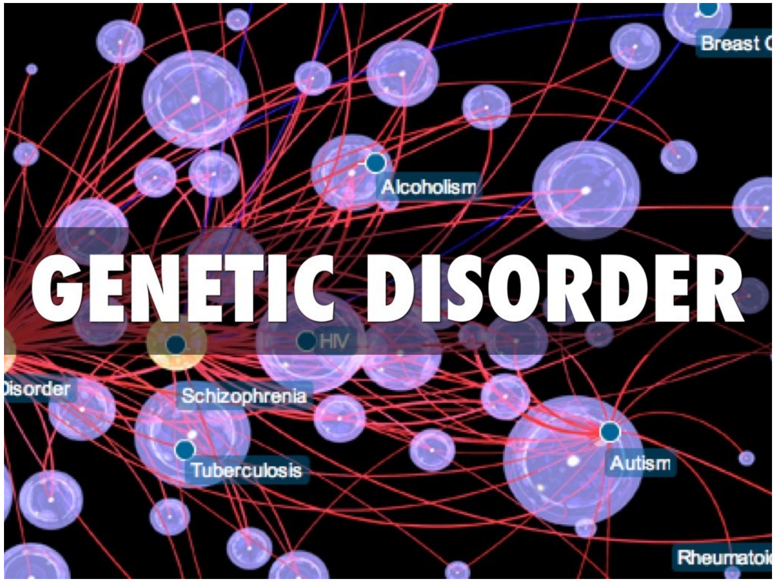
Bertrand Thirion
bertrand.thirion@inria.fr

Toward interpretable machine learning

E-sciences



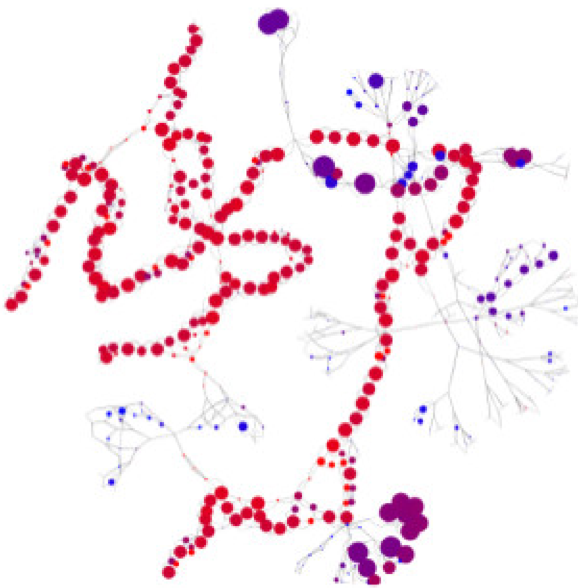
Complex phenomenon



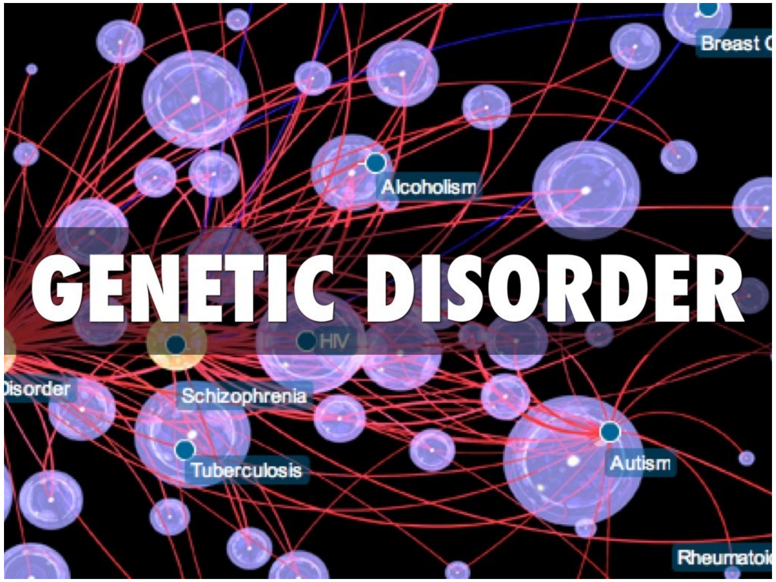
Outcome

Toward interpretable machine learning

E-sciences



Machine learning



Complex phenomenon

Outcome

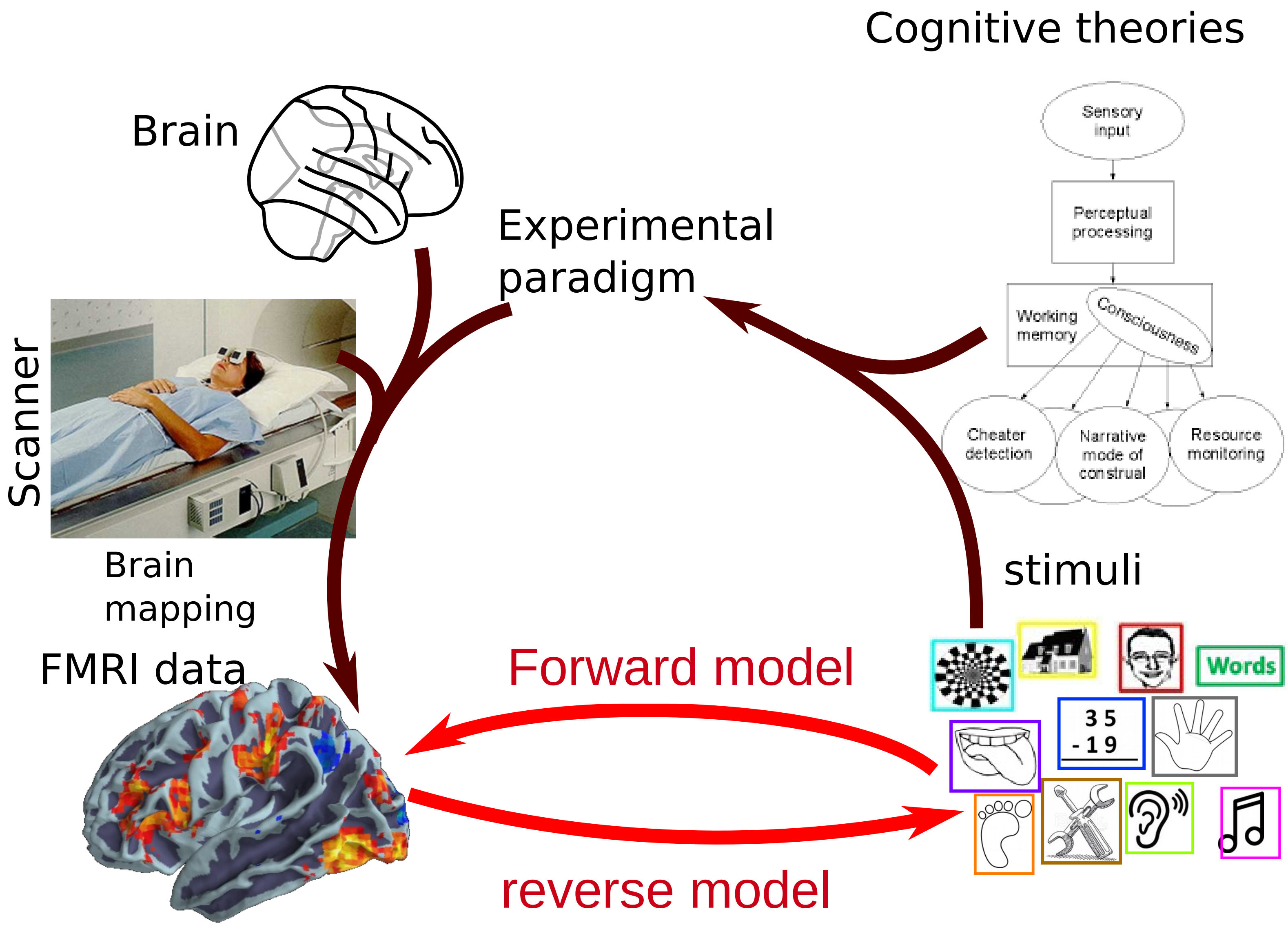
What are the **causes** of the outcome ?
Can one **localize** the effect ?

Problem (genetics, brain imaging):
high dimensional small samples problems

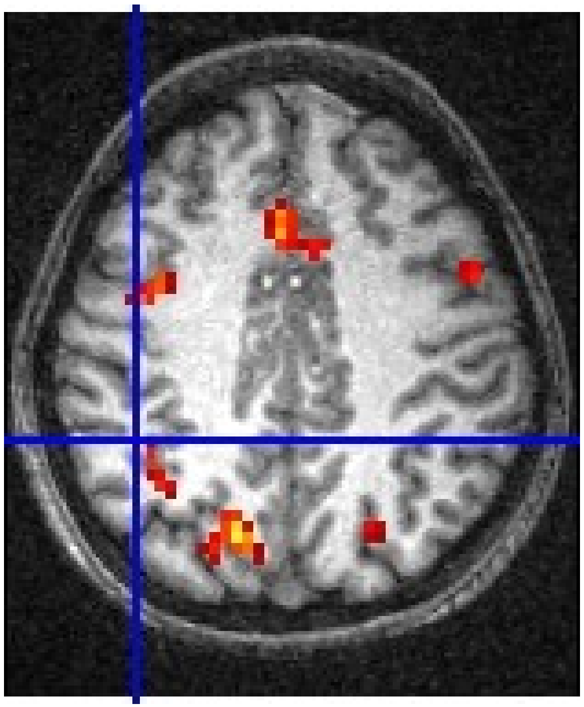
Cognitive neuroscience

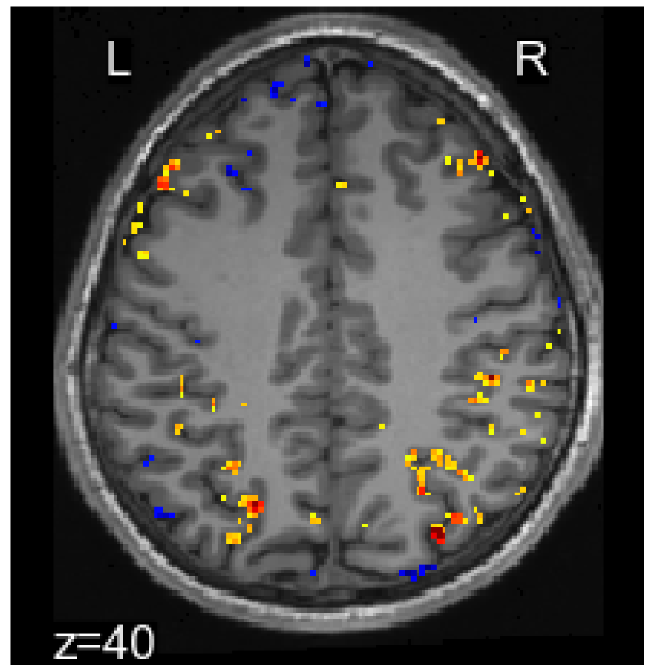
How are cognitive activities affected or controlled by neural circuits in the brain ?

The brain, the mind and the scanner



Resolution increases





2007:
3 mm

2014:
1.5 mm

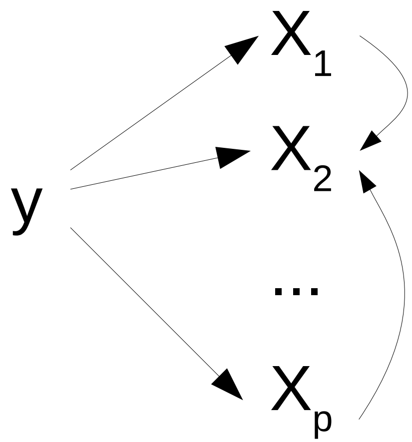
2020:
0.5 mm ?

p = 50,000

p = 400,000

p = 10^7

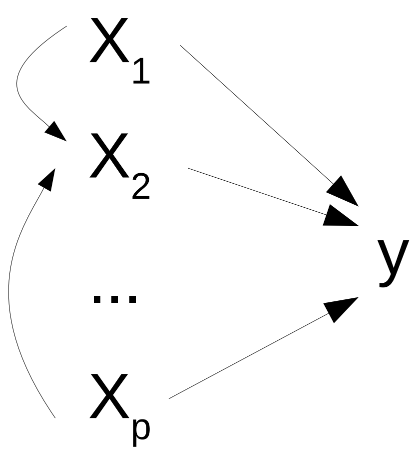
Statistical problem



$$\mathbf{X_i} \overset{?}{\perp\!\!\!\perp} \mathbf{y}$$

Forward model
(easy)

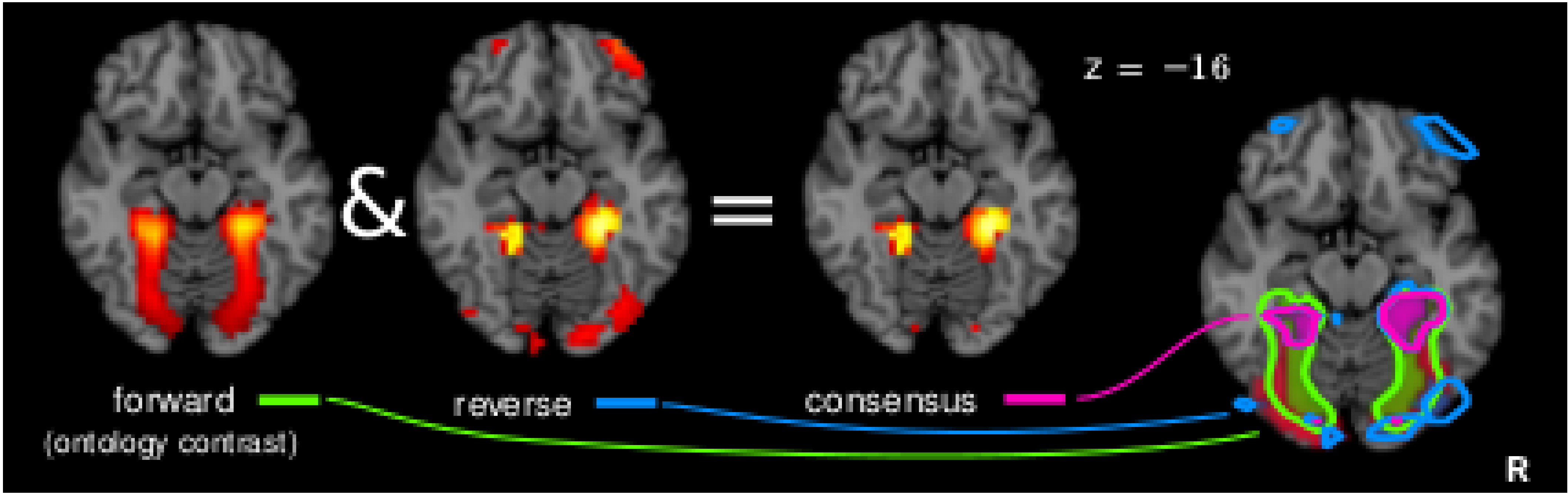
Reverse model (hard)



$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}$$

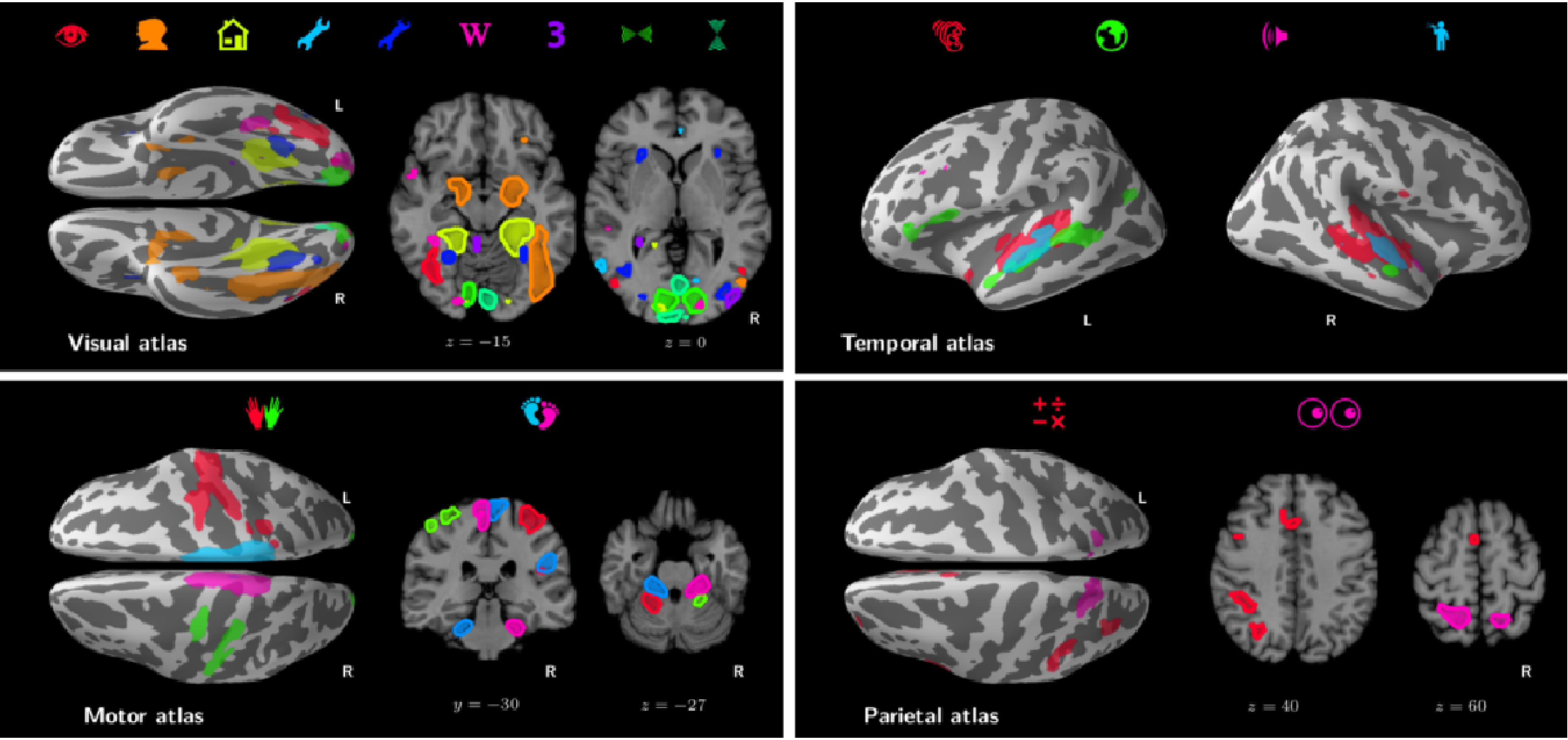
$$\mathbf{w}_i = 0 \iff \mathbf{X_i} \perp\!\!\!\perp \mathbf{y} | (\mathbf{X_j}, \mathbf{j} \neq \mathbf{i})$$

Forward *and* reverse brain imaging



[Schwartz et al. NIPS 2013, Varoquaux et al. Submitted to PCB]

Forward *and* reverse brain imaging



[Schwartz et al. NIPS 2013, Varoquaux et al. Submitted to PCB]

Statistical associations and causal reasoning

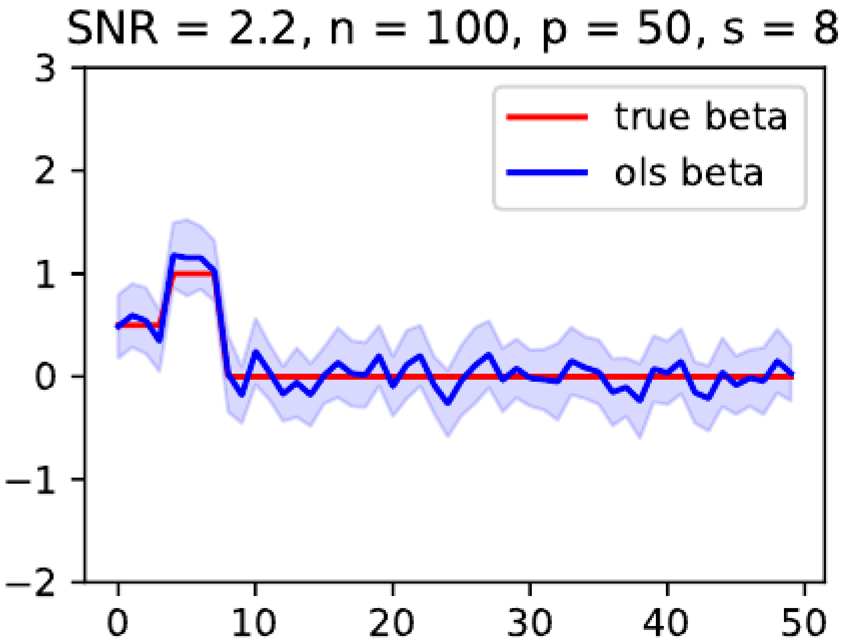
- Problems:
  - Establish non-independence based on finite datasets → statistical tests
  - **Large number of conditioning variables**

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}$$

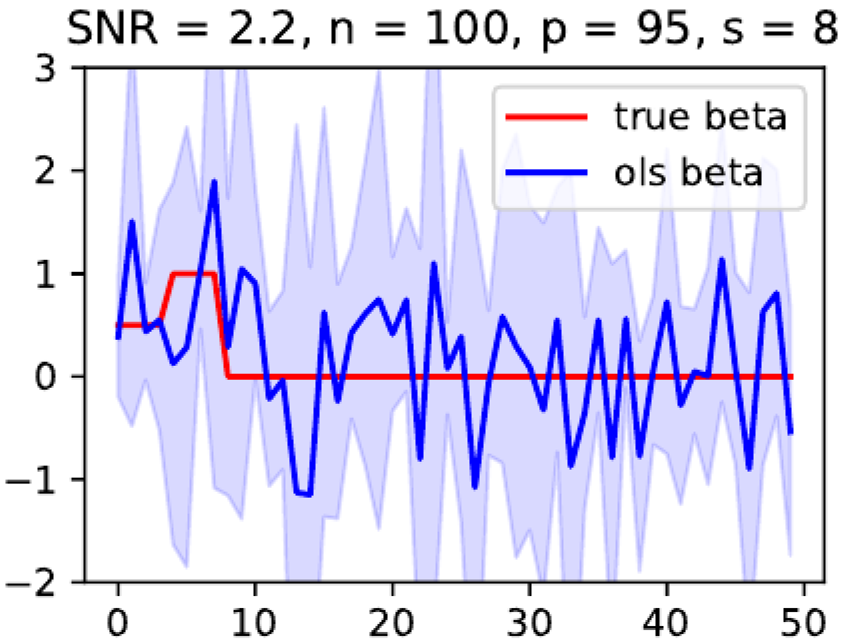
$$\mathbf{w}_i \overset{?}{\neq} 0$$

Statistical inference on w

- Inference: find {j: $\mathbf{w}_j > 0$} with some statistical guarantees
- Classical statistics **break** whenever **p close to n**



OLS regression for $p = \frac{n}{2}$



OLS regression for $p \approx n$

Desparsified Lasso

- **Desparsified Lasso estimator:** when $n < p$, $\mathbf{z}_j$ is the residual of a Lasso-CV regression of $\mathbf{x}_j$ vs $\mathbf{X}^{(-j)}$ and the debiased estimator is:

$$\hat{w}_j = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{x}_j} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k \hat{w}_k^{(init)}}{\mathbf{z}_j^\top \mathbf{x}_j} ,$$

  where $\hat{\boldsymbol{w}}^{(init)}$ is an initial non linear estimator of $\boldsymbol{w}^*$ (*e.g.*, Lasso)

- **Covariance:** the covariance matrix of this estimator is:

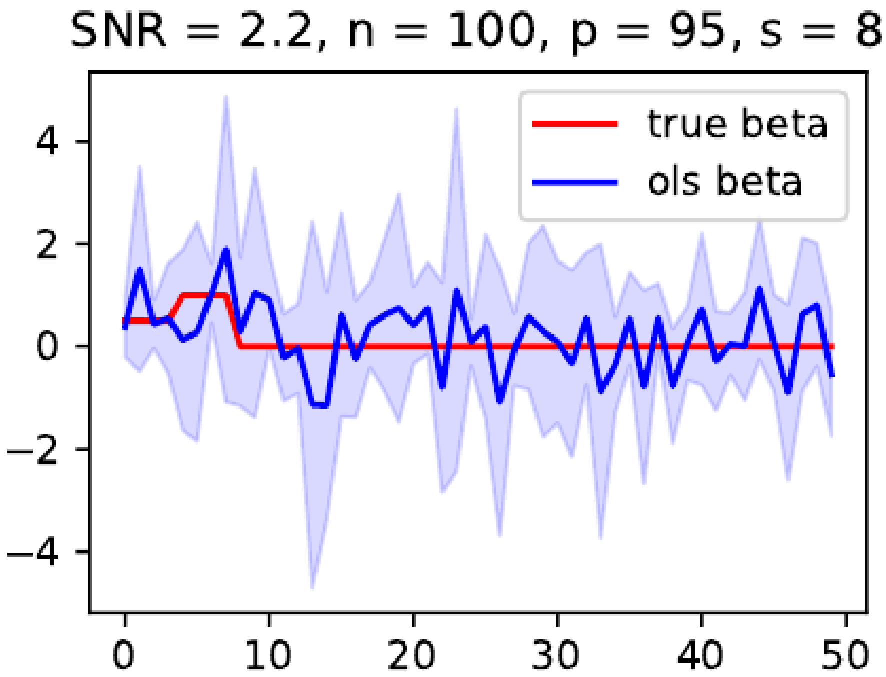
$$\Omega_{jk} = \frac{n\mathbf{z}_j^\top \mathbf{z}_k}{(\mathbf{z}_j^\top \mathbf{x}_j)(\mathbf{z}_k^\top \mathbf{x}_k)}$$

- **Confidence bounds:** under few assumptions (Dezeure et al. [2015]):

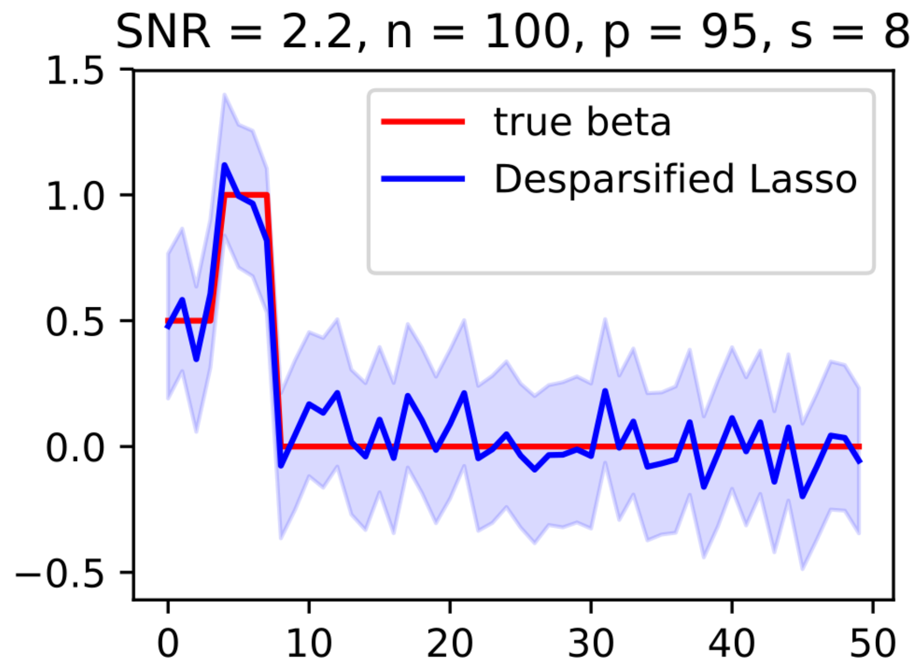
$$\sigma_*^{-1}(\Omega_{jj})^{-1/2}(\hat{w}_j - w_j^*) \sim \mathcal{N}(0, 1)$$

Preliminary assessment

- **Low dimension:** $n = 100$ and $p = 95$
- **OLS versus corrected Ridge and desparsified Lasso:**



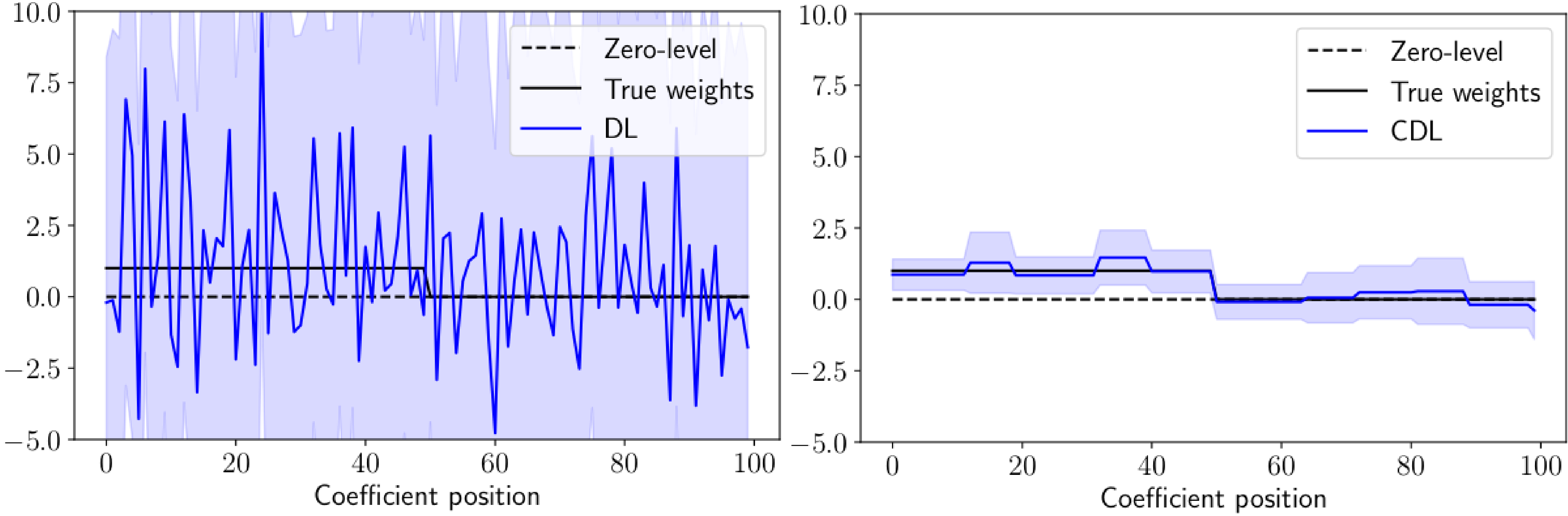
OLS regression when $p \approx n$



Corrected Ridge and **Desparsified Lasso** when $p \approx n$

Large p → need dimension reduction

p=2000, n=100

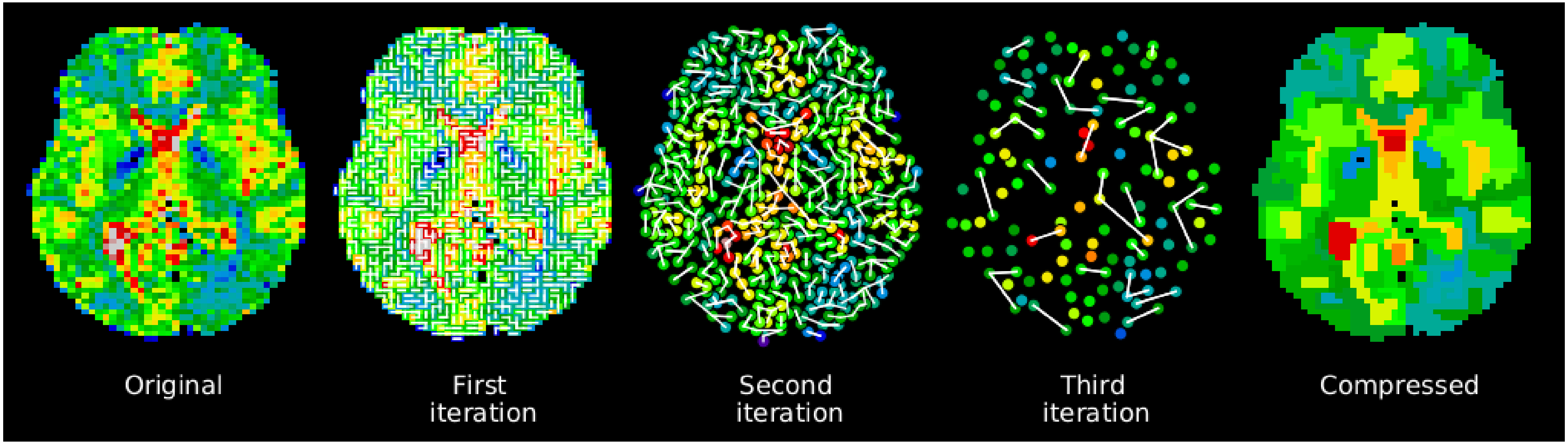


Large p kills statistical power

CDL tames variance

[Chevalier et al. MICCAI 2018]

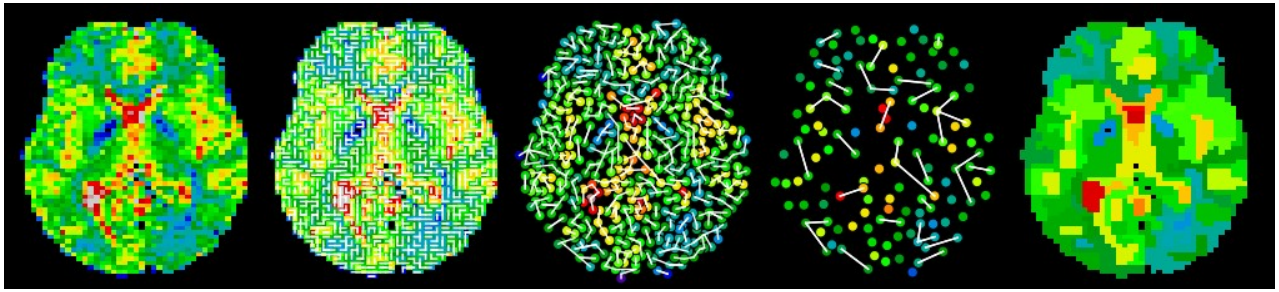
(Very) fast clustering for images



[Thirion et al. Stamlins 2015, Idrobo et al. PAMI 2018]

Adaptation to brain imaging

Step 1: compression by clustering



Step 2: inference on compressed representations

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

C*lustered*
D*esparsified*
L*asso*

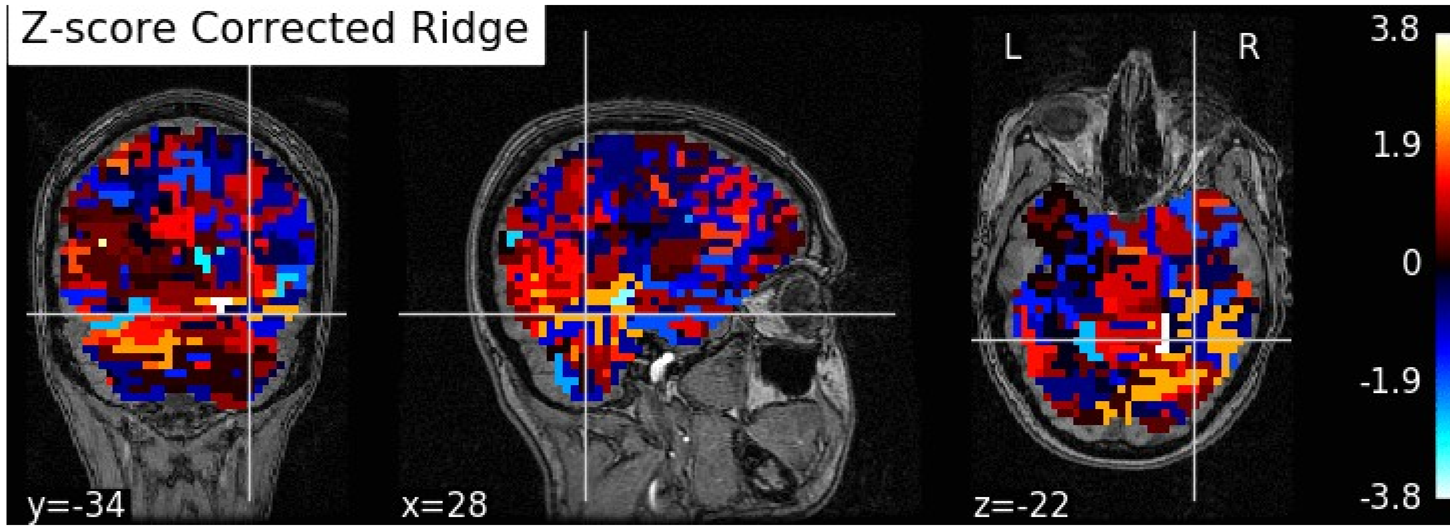
Step 3: ensembling iterate with different parcellations

→ **aggregate p-values** (FReM-like approach)

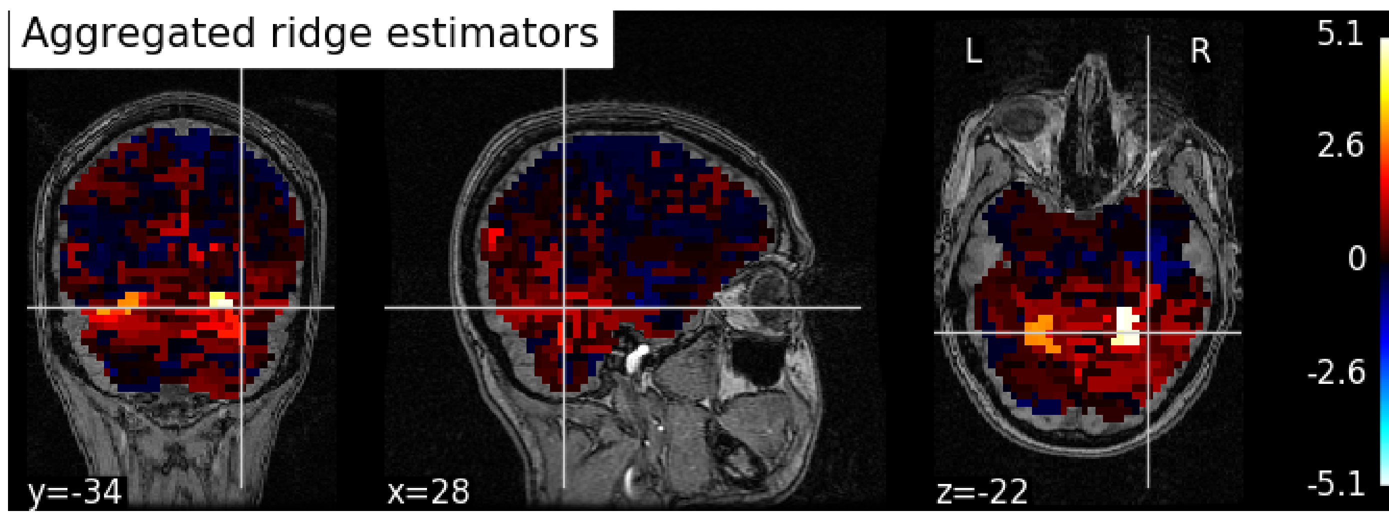
E*nsemble of*
C*lustered*
D*esparsified*
L*asso*

Ensemble of clustered desp. lasso

DL p-values from different clusterings



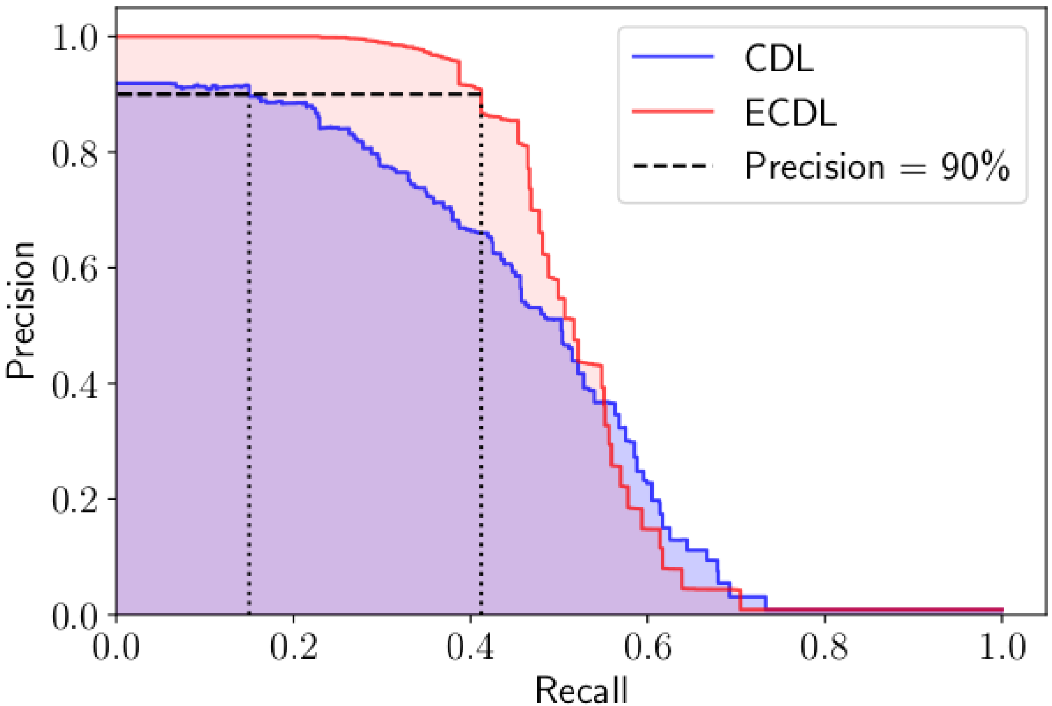
aggregation

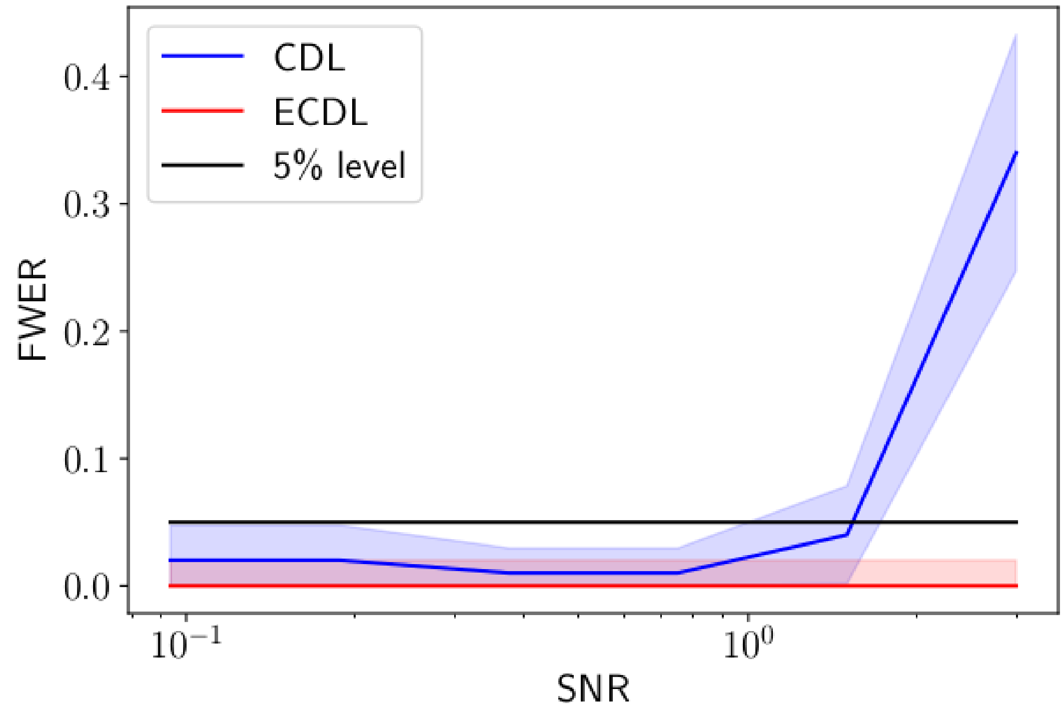


Experiments: PR and FWER control

$$\text{Recall} = \frac{\text{Number of true positive}}{\text{Size of the active set}} \qquad \text{Precision} = \frac{\text{Number of true positive}}{\text{Number of discoveries}}$$

$$\text{FWER} = \text{Prob}(\text{Number of false positive} \geq 1)$$





Better PR with ECDL

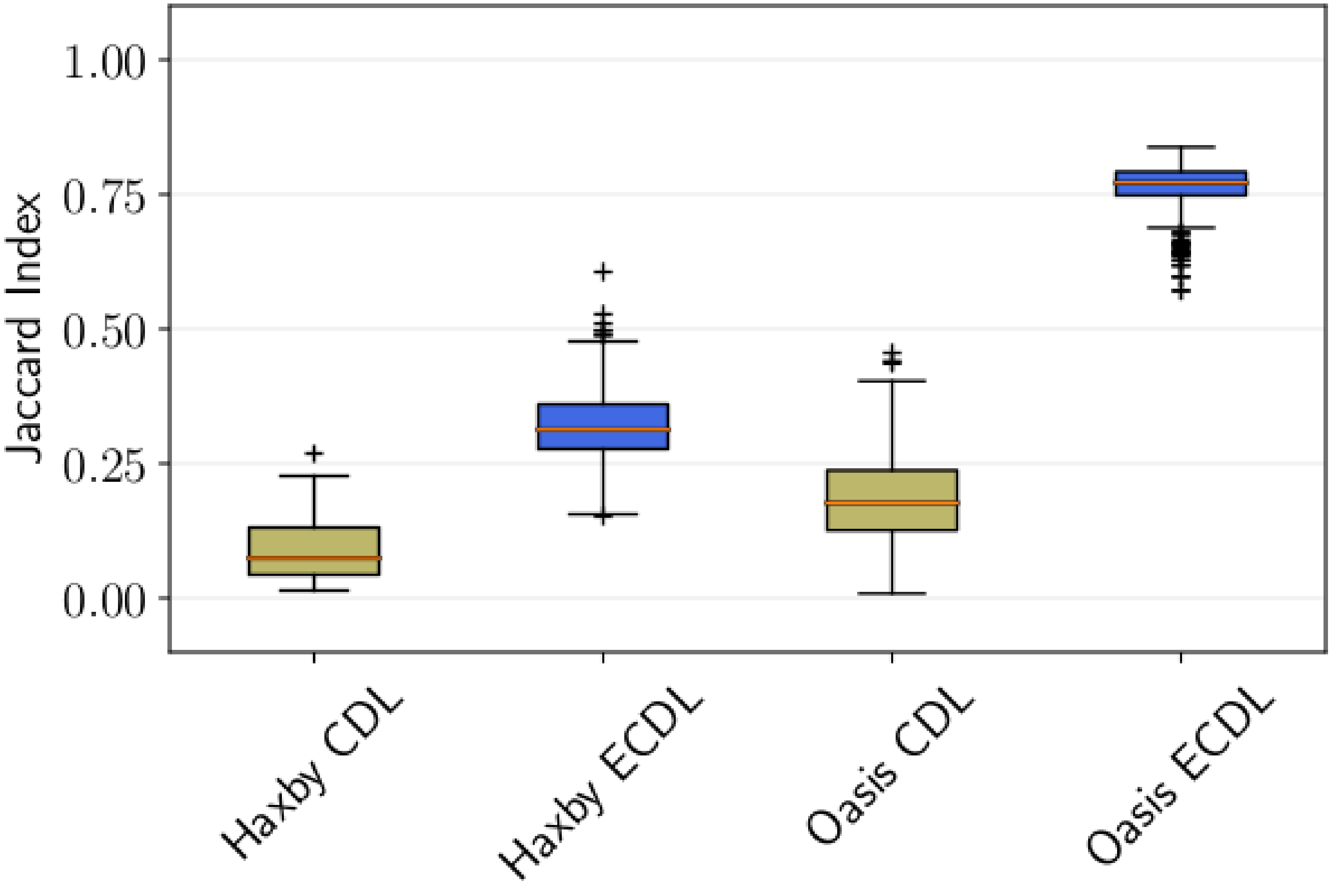
+ More accurate FWER control

[Chevalier et al. MICCAI 2018]

Stability gains on real data

Similarity across bootstrap replications of the inference

(same result with other metrics)



On two datasets, ECDL improves reproducibility

[Chevalier et al. MICCAI 2018]

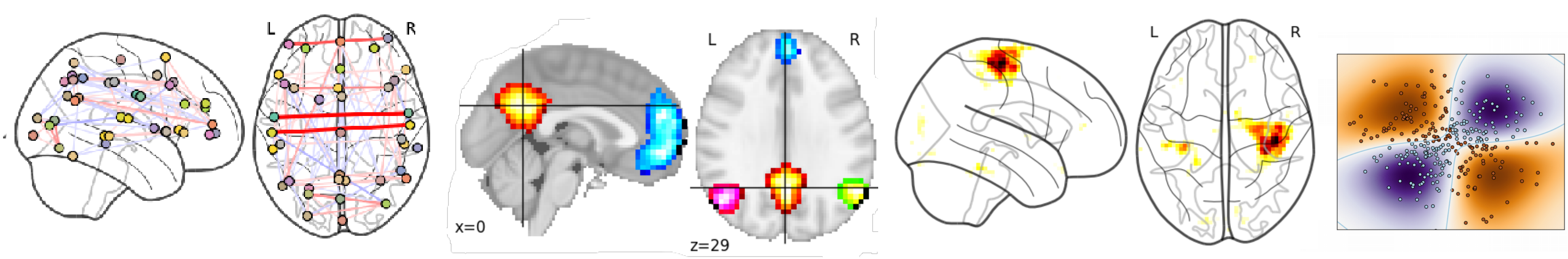
Conclusion

- Causal reasoning with **conditional association analysis**
- Large-p data bring **challenges**:
  - Computation cost
  - Difficulty of statistical inference
- Solutions: compression and esembling

High dimensional inference is now possible !

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



Acknowledgements

Parietal

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

Other collaborators
(thanks for the data)
S. Dehaene
R. Poldrack,
J. Haxby
C. F. Gorgolevski
J. Salmon

Human Brain Project

université PARIS-SACLAY

AGENCE NATIONALE DE LA RECHERCHE ANR