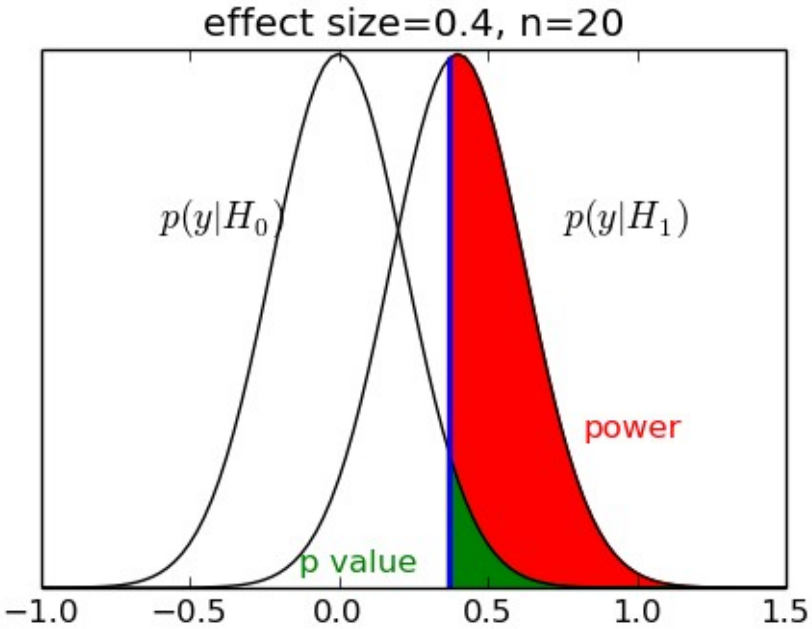
Large-scale analyses in functional brain Imaging

Bertrand Thirion
bertrand.thirion@inria.fr

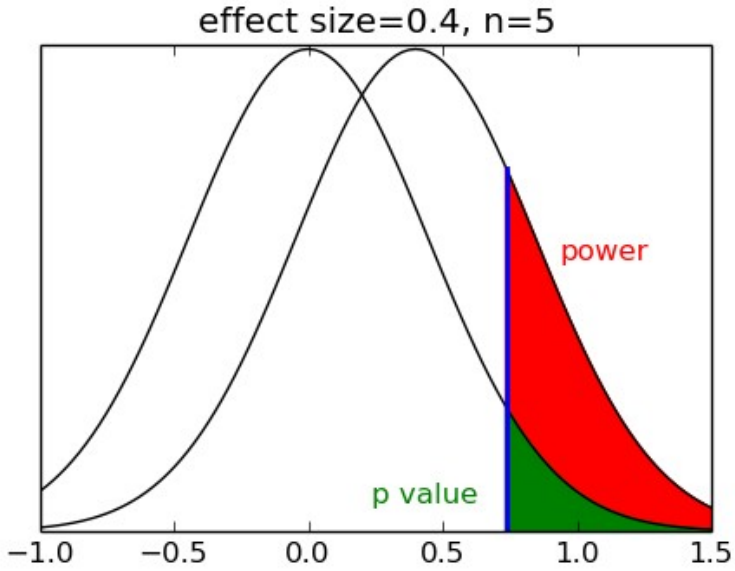
Outline

- The importance of using large(r) datasets in medical imaging
- Combining modalities to increase power
- Identify ensuing computational bottlenecks and algorithmic solutions

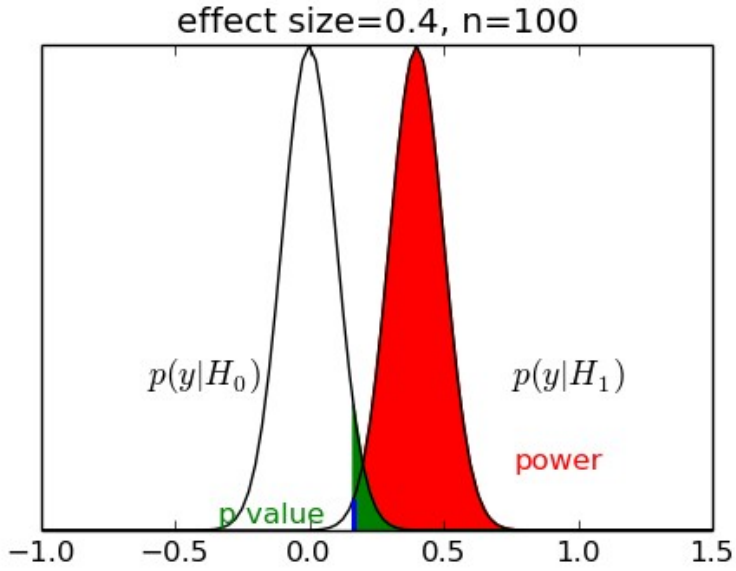
Increasing the number of samples: power



Less subjects

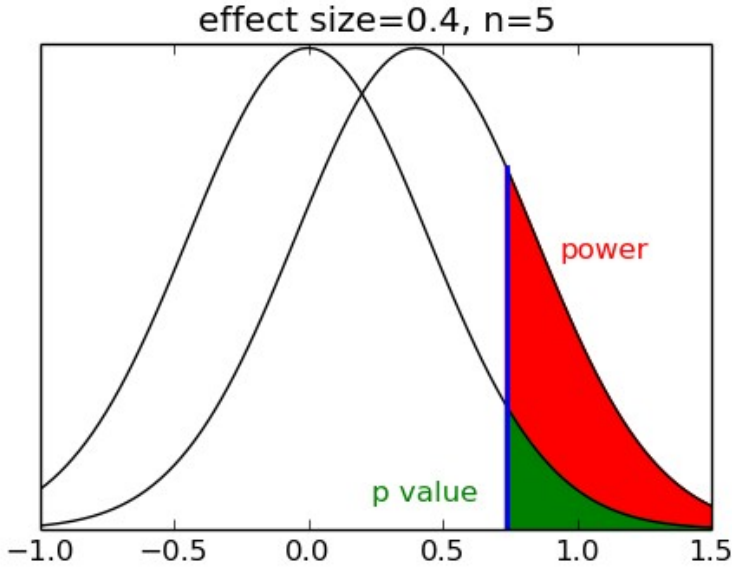


more subjects

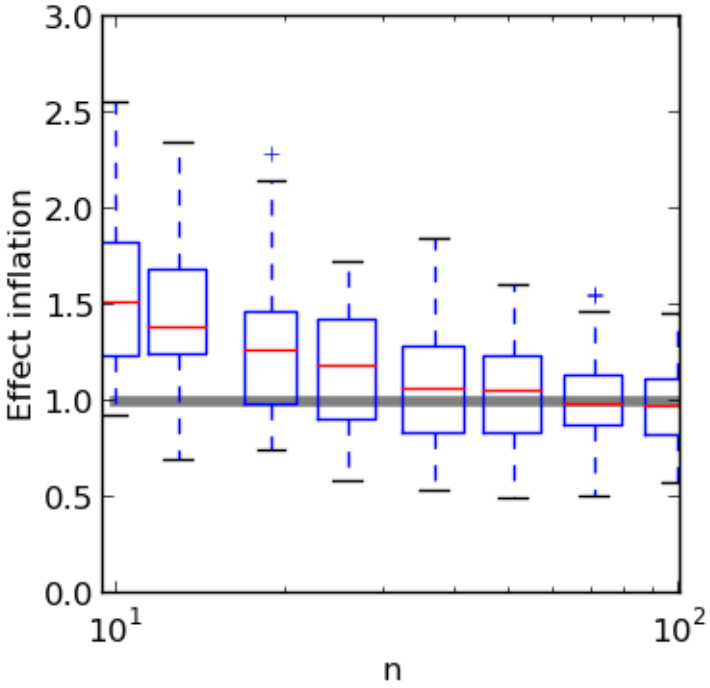


Imaging in the low power regime

Low power → unreliable findings



Low positive predictive value: low likelihood that a statistically significant finding is actually true



Exaggerated estimate of the magnitude of discovered effects "*winner's curse*"

Medical imaging in the big data era ?

- Neuroscientists reluctant to using big data : this is a burden
- Common belief : you should only focus on large effects
  → small datasets are good enough
- But...

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

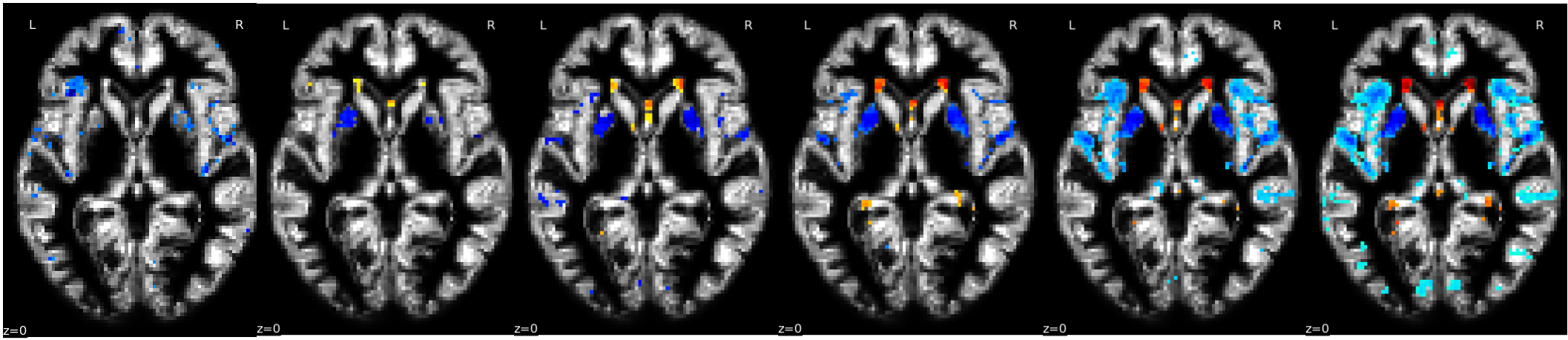
nature REVIEWS

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]

Example

OASIS dataset, effet of age on gray matter density

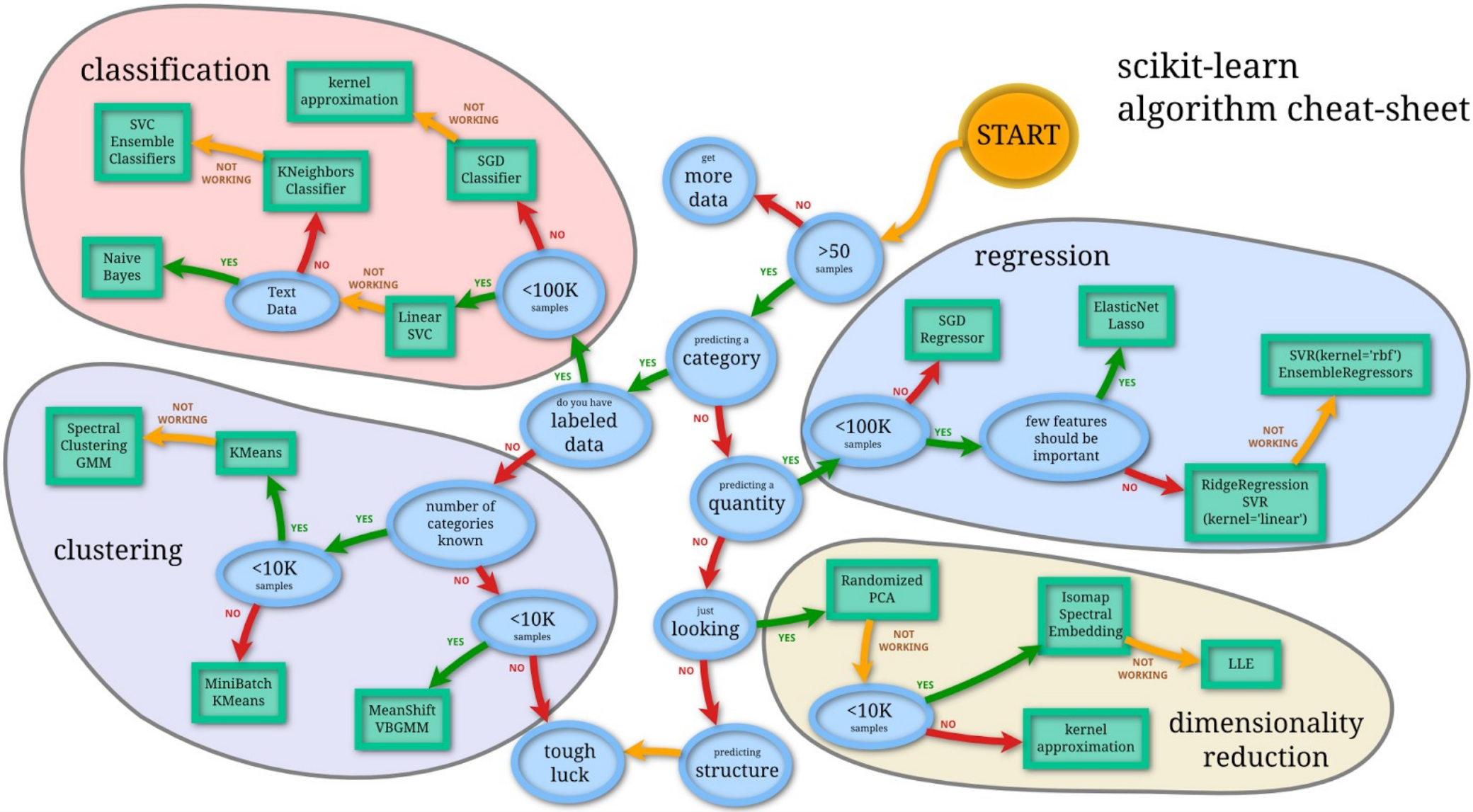


n=10 **n=20** **n=30** **n=50** **n=70** **n=100**

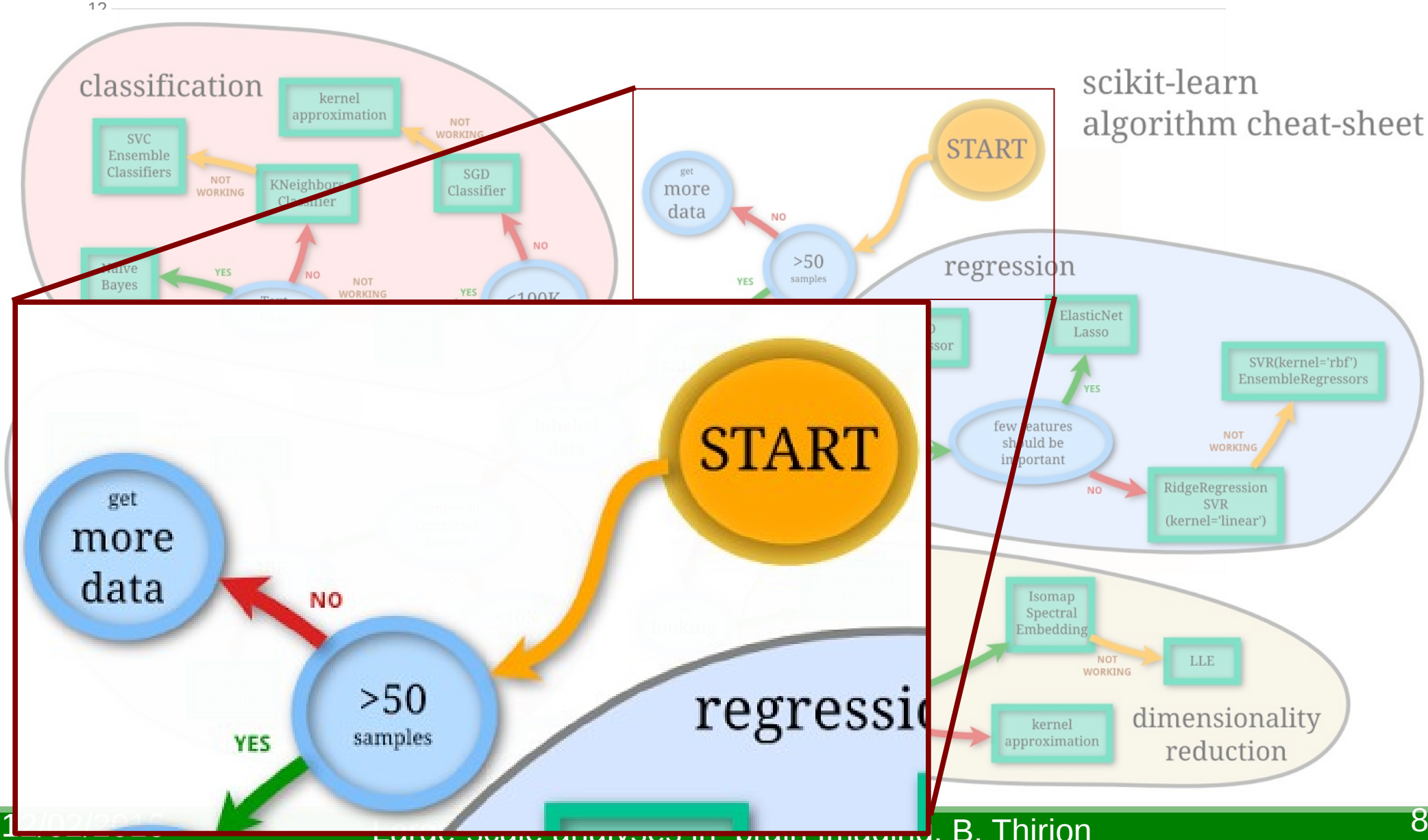
Reproducibility by bootstrap:

7% 19% 32% 53% 66% 75%

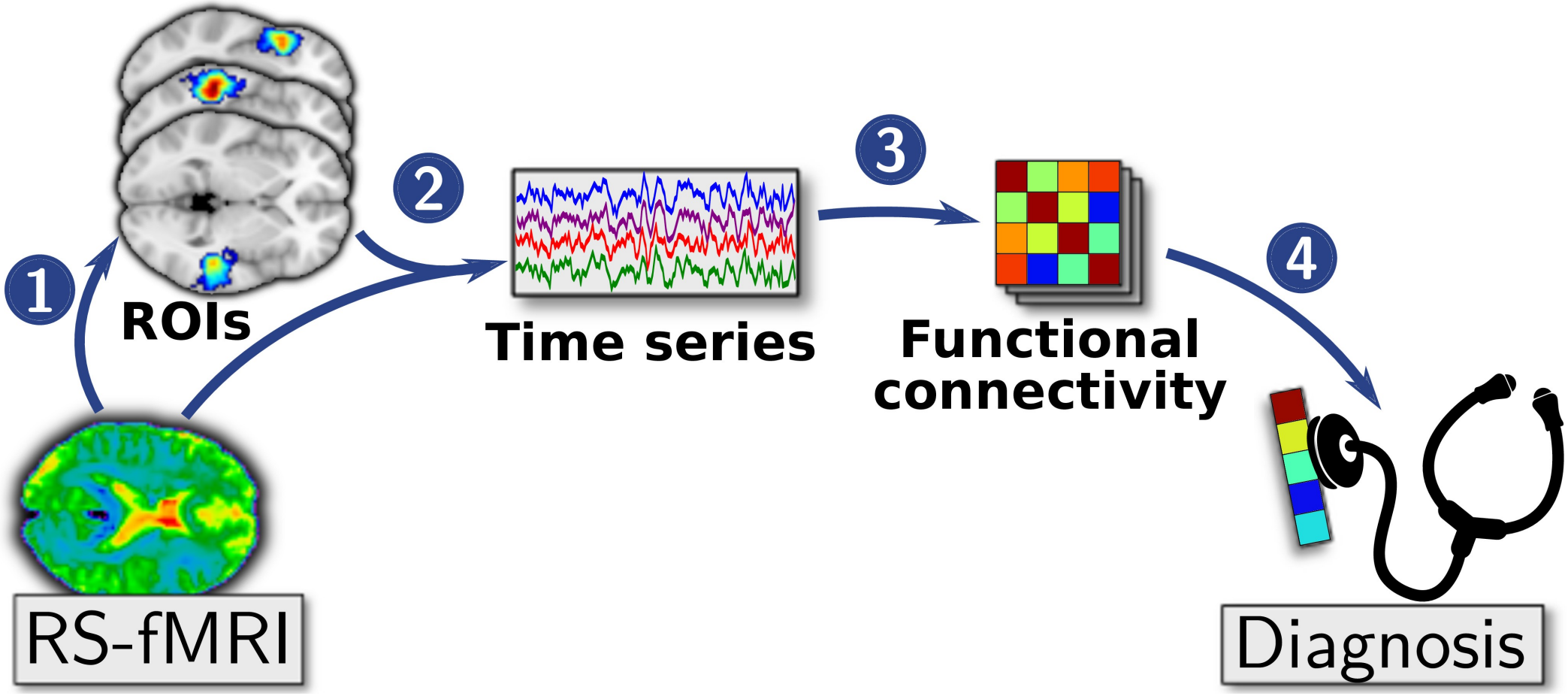
Multivariate analysis



Multivariate_analysis



Large-scale connectome analysis for disease prediction



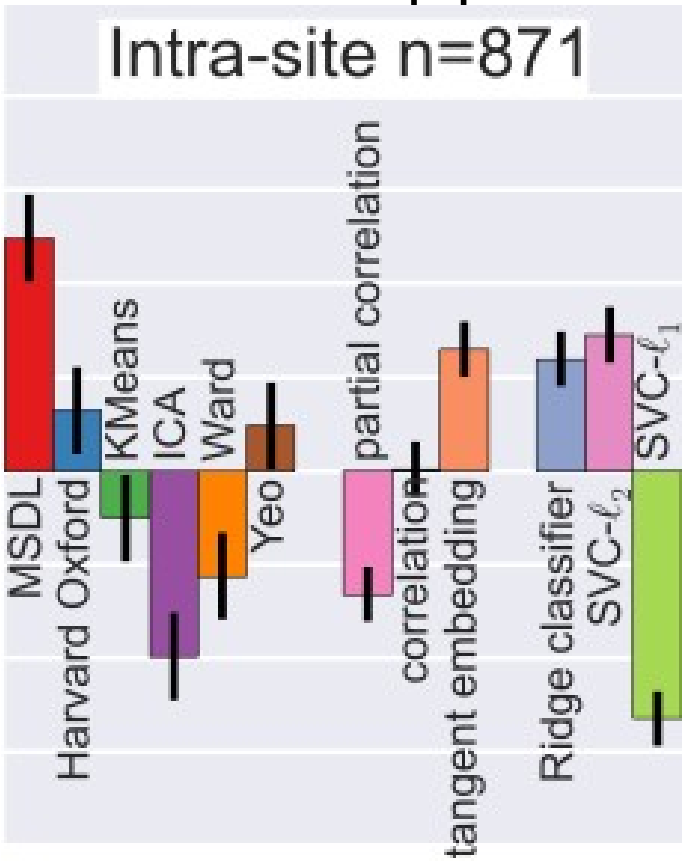
871 subjects, 300-400GB of data

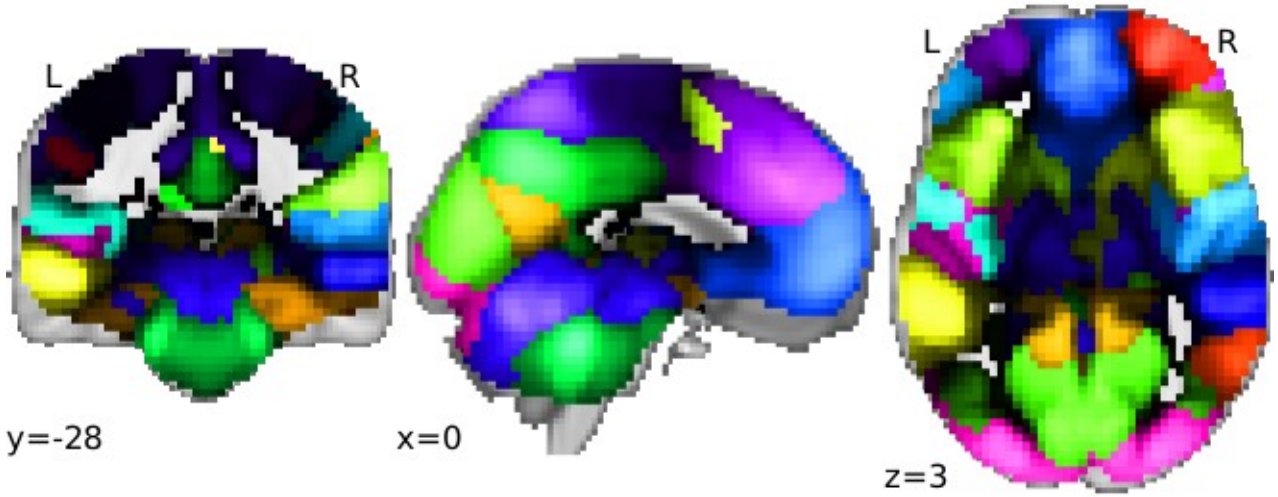
Large scale connectome analysis for disease prediction

- Classification of ASD patients
  - Public ABIDE dataset [Di Martino et al. Mol. Psych. 2014]
  - Large heterogeneity: multi-centric study
  - Functional connectivity: weakly sensitive marker
- 68% correct classification (65% when generalizing across sites)

Large scale connectome analysis for disease prediction

What are the most important factors of the pipeline ?





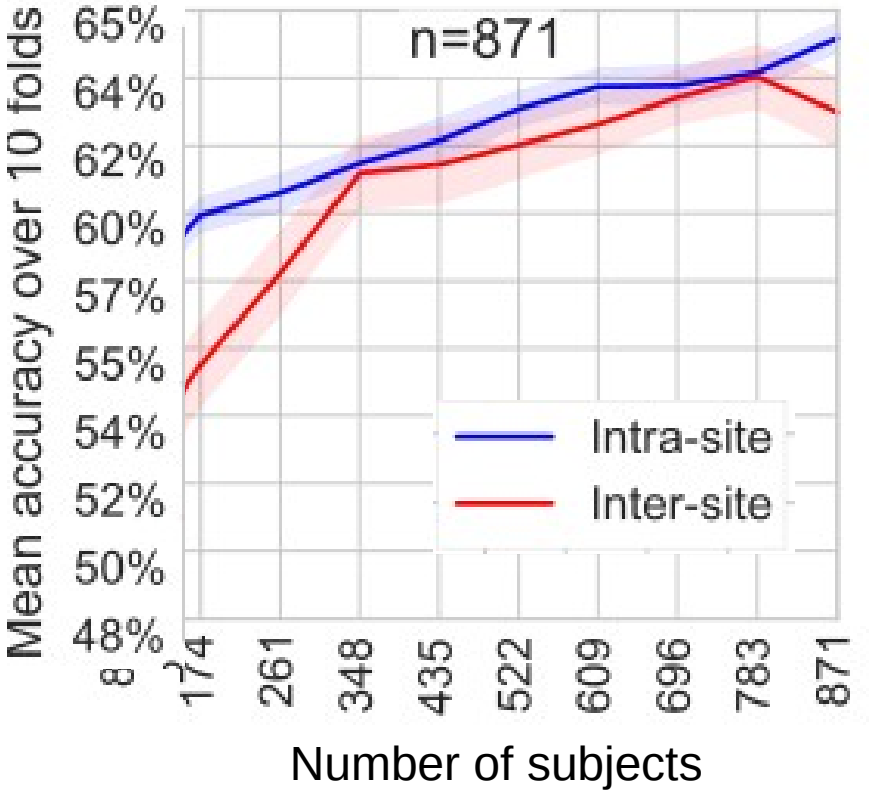
Response: the choice of the atlas.

[Abraham et al. Subm to NIMG]

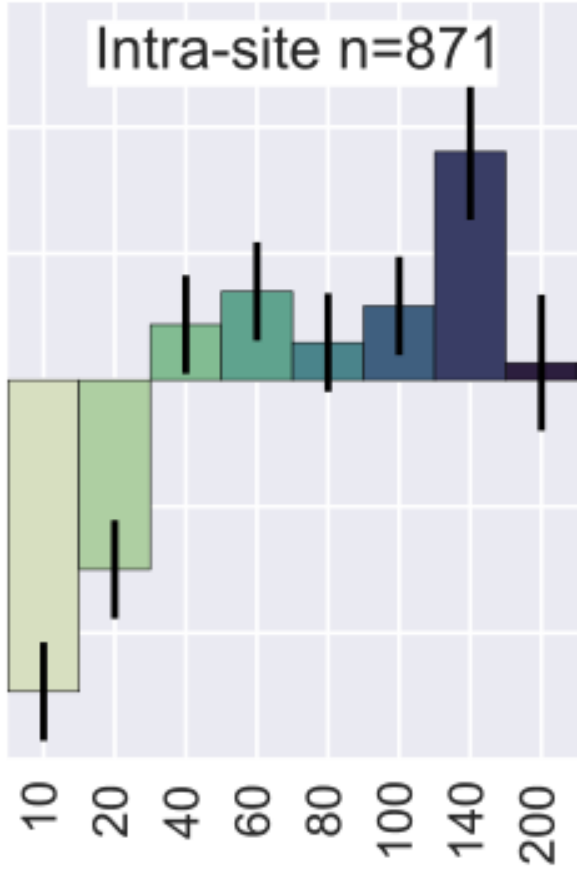
Large scale connectome analysis for disease prediction

More subjects → higher accuracy

Asymptote not yet reached at n=871

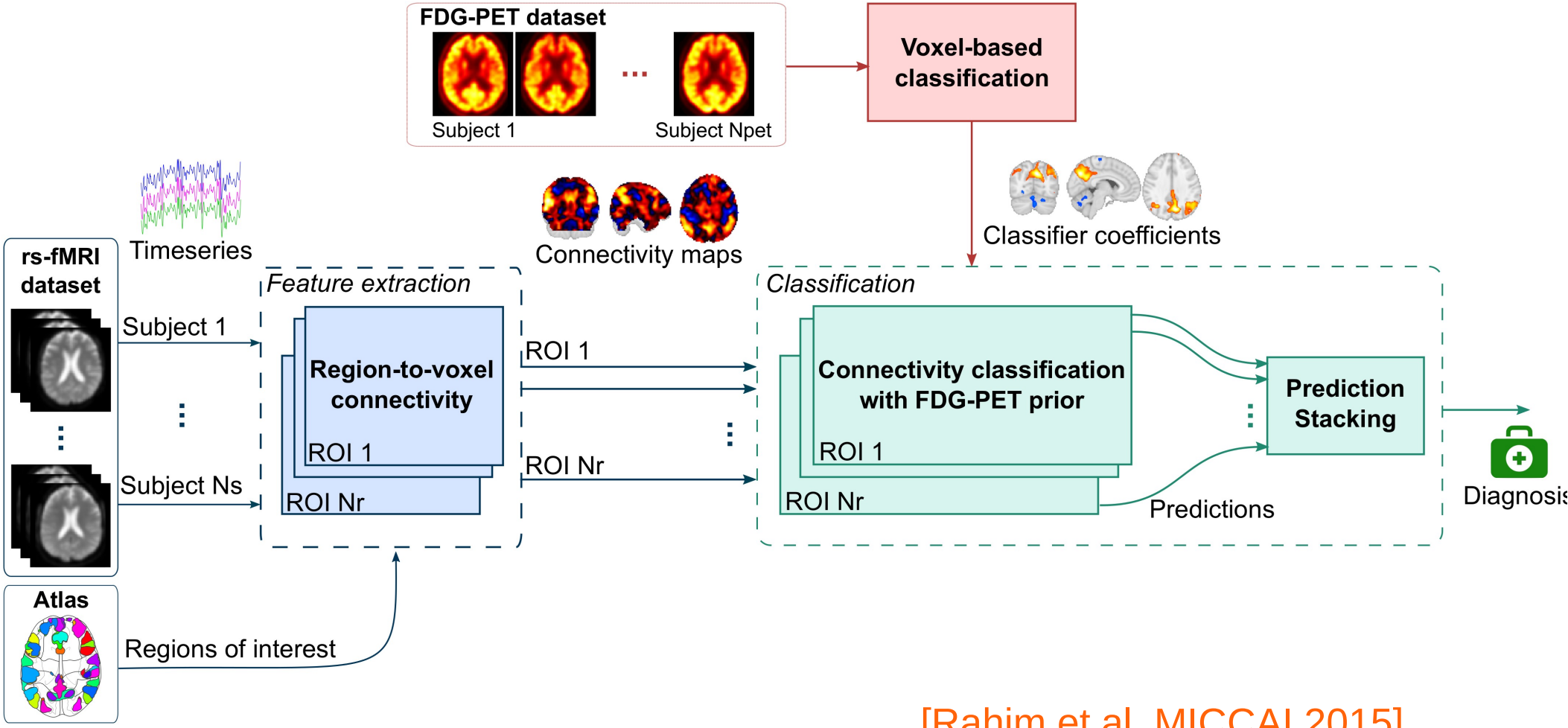


Impact of the number of regions



Leveraging multi-modal datasets

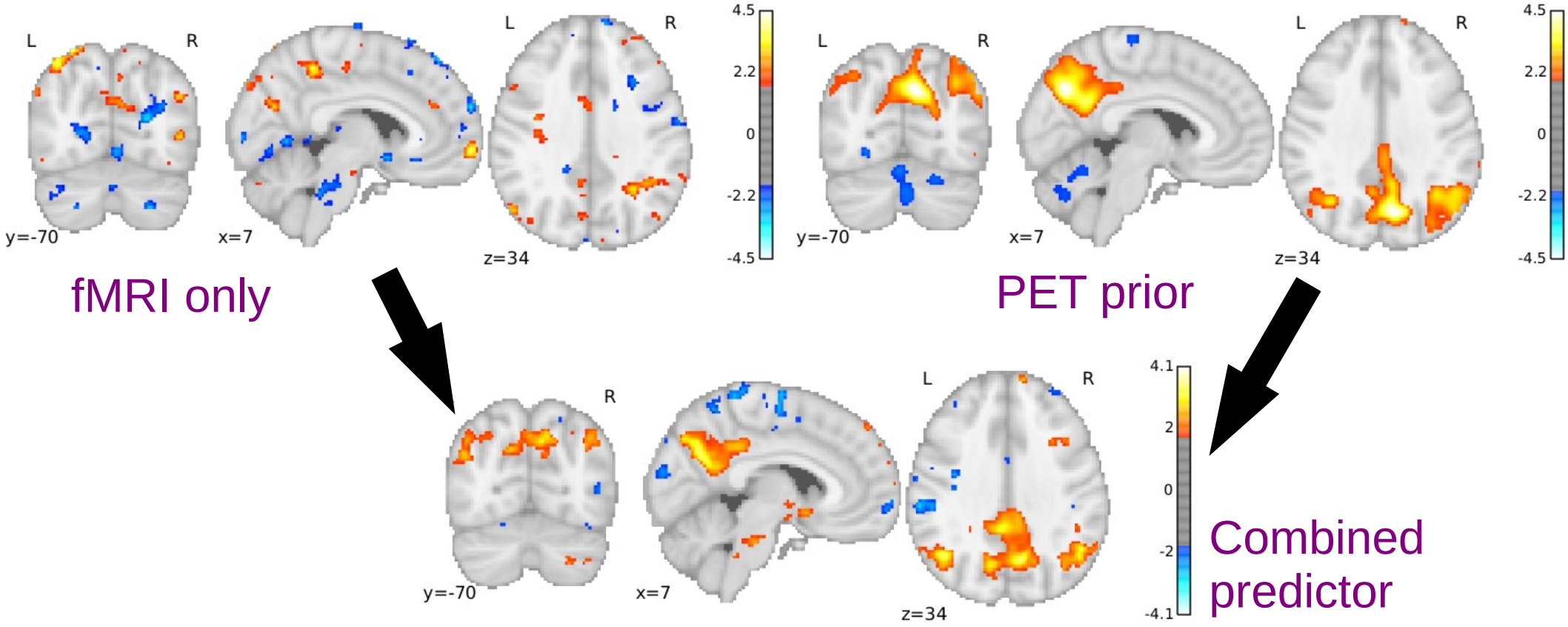
Combining images from Different modalities



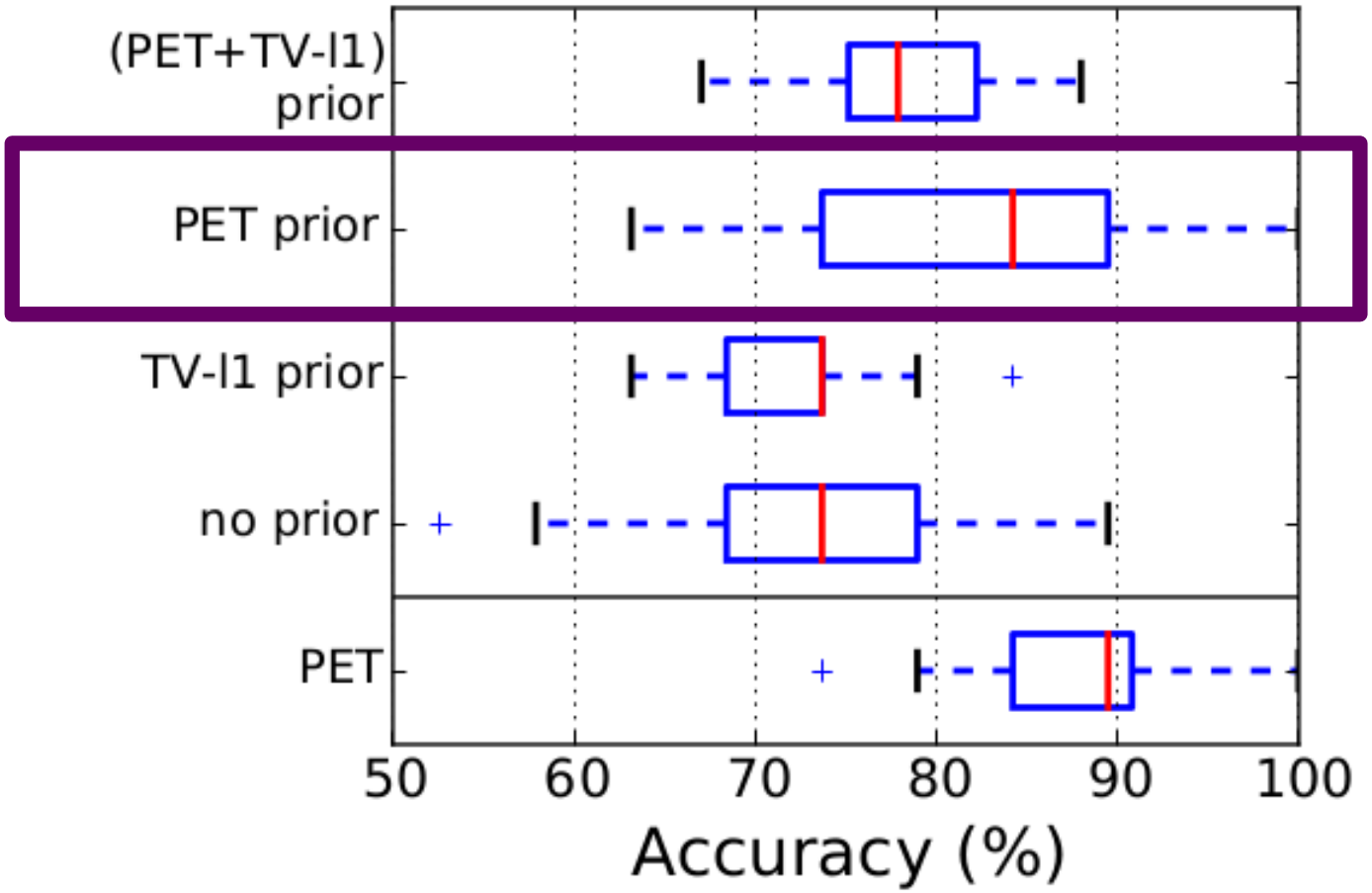
[Rahim et al. MICCAI 2015]

Trans-modal priors improve discriminative models

Diagnosis of Alzheimer disease (Alzheimer vs MCI) using resting-state fMRI and a PET prior



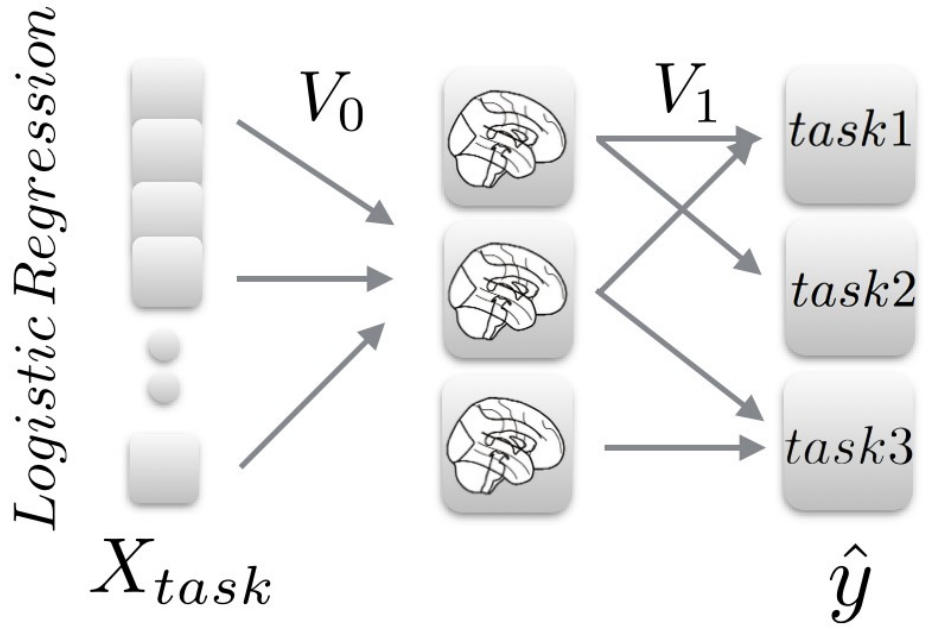
Trans-modal priors improve discriminative models



[Rahim et al. MICCAI 2015]

Using rest data to better model task data

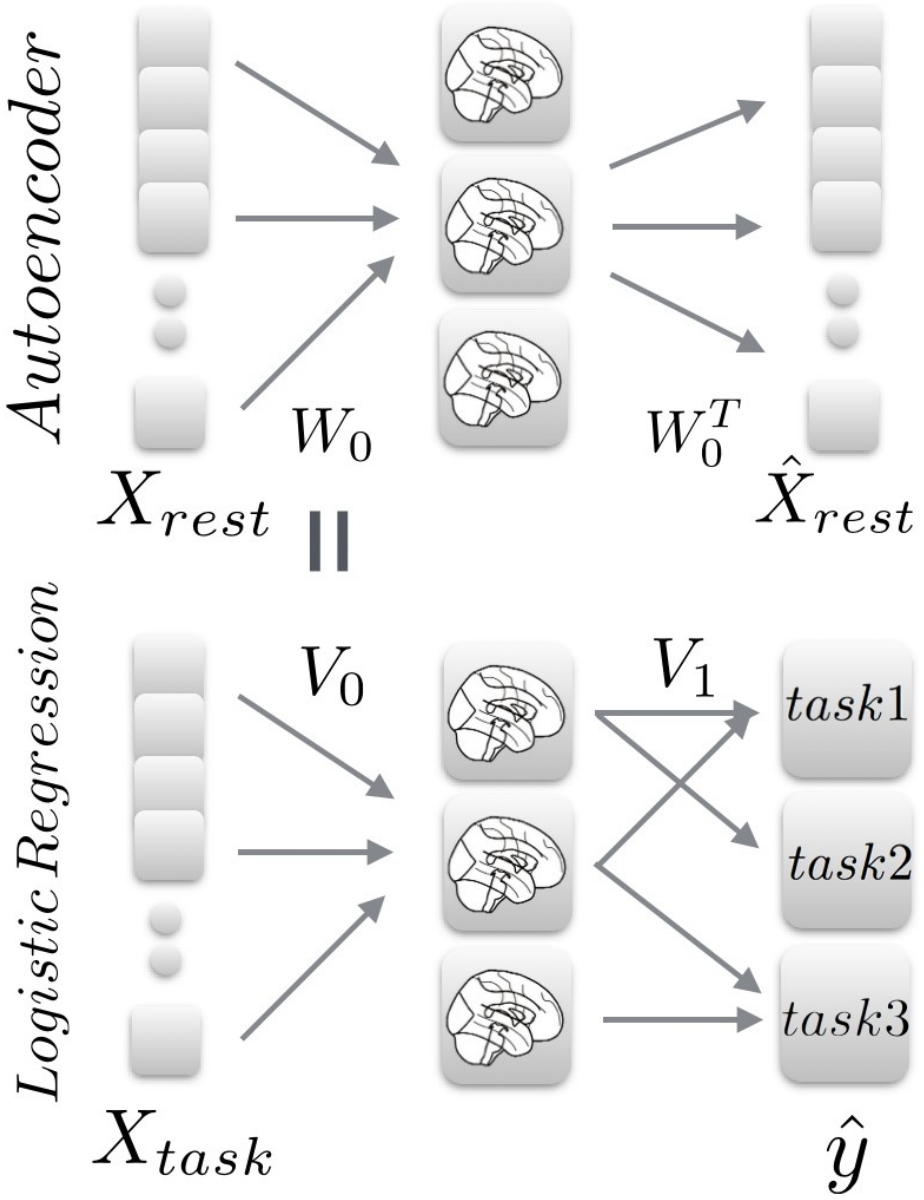
- A wealth of “resting-state” data available
- Little cognitive data with proper annotations
- Idea: use unlabelled data to improve the learning of discriminative models



[Bzdok et al. NIPS 2015]

Using rest data to better model task data

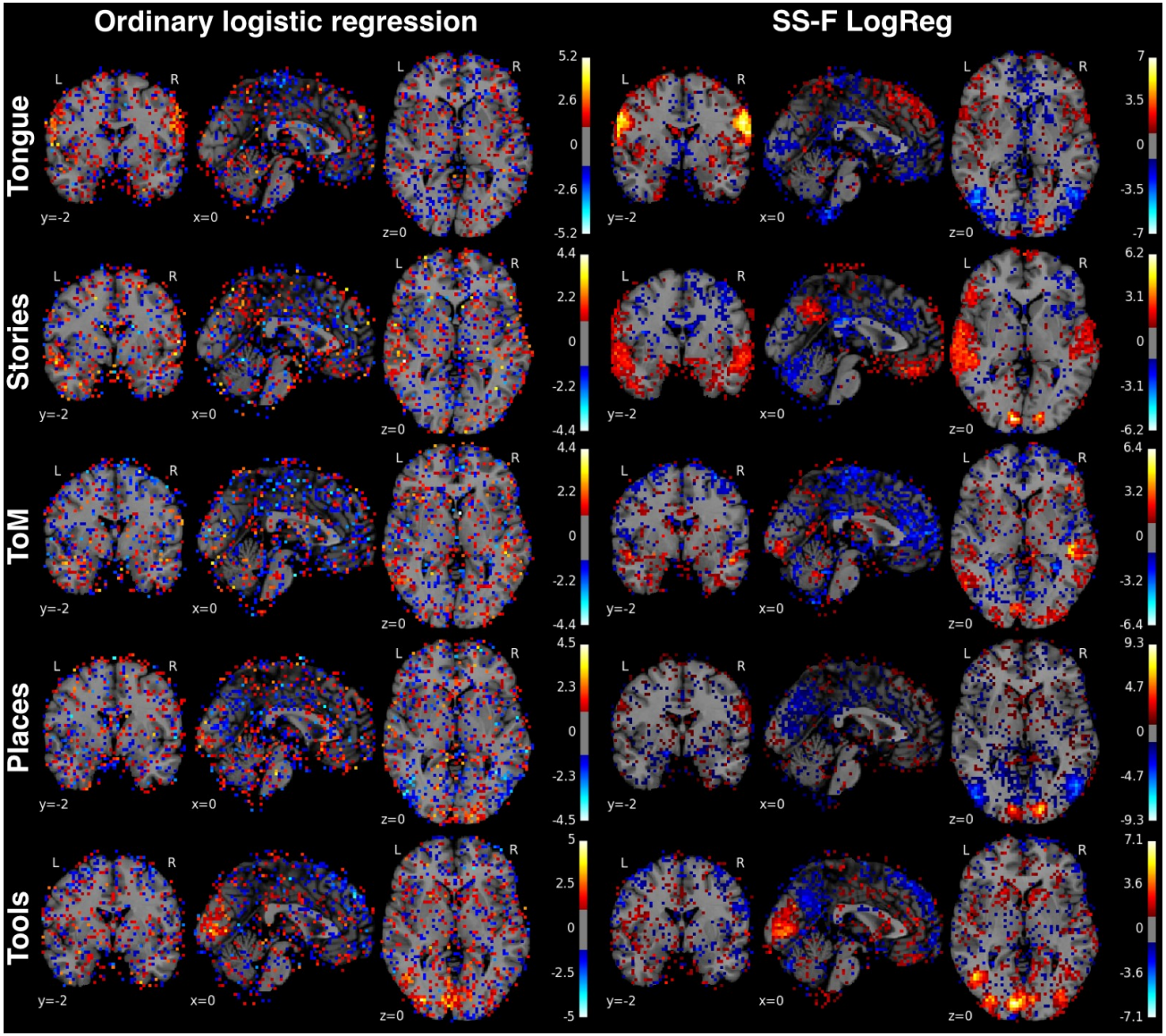
- A wealth of “resting-state” data available
- Little cognitive data with proper annotations
- Idea: use unlabelled data to improve the learning of discriminative models



[Bzdok et al. NIPS 2015]

Using rest data to better model task data

- The discriminative patterns for many task data is much cleaner thanks to the composite objective



[Bzdok et al. NIPS 2015]

Technical aspects of big data analysis

Big data in medical imaging ?

HCP mailing list, Jan 19th, 2015

"Has anyone on the 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."

Diagnosis and strategy

- Modern, multiple, datasets: 100 GB-TBs
  - Do not fit in memory. Need online learning
  - Memory access is the main bottleneck
  - Large p, large n
  - Better file formats would help, but compatibility with existing tool stacks.
- Agile approach: readable Python code, runs on your laptop
  - Online learning, SGD algorithms
  - compression

Caching & parallel comp.: joblib

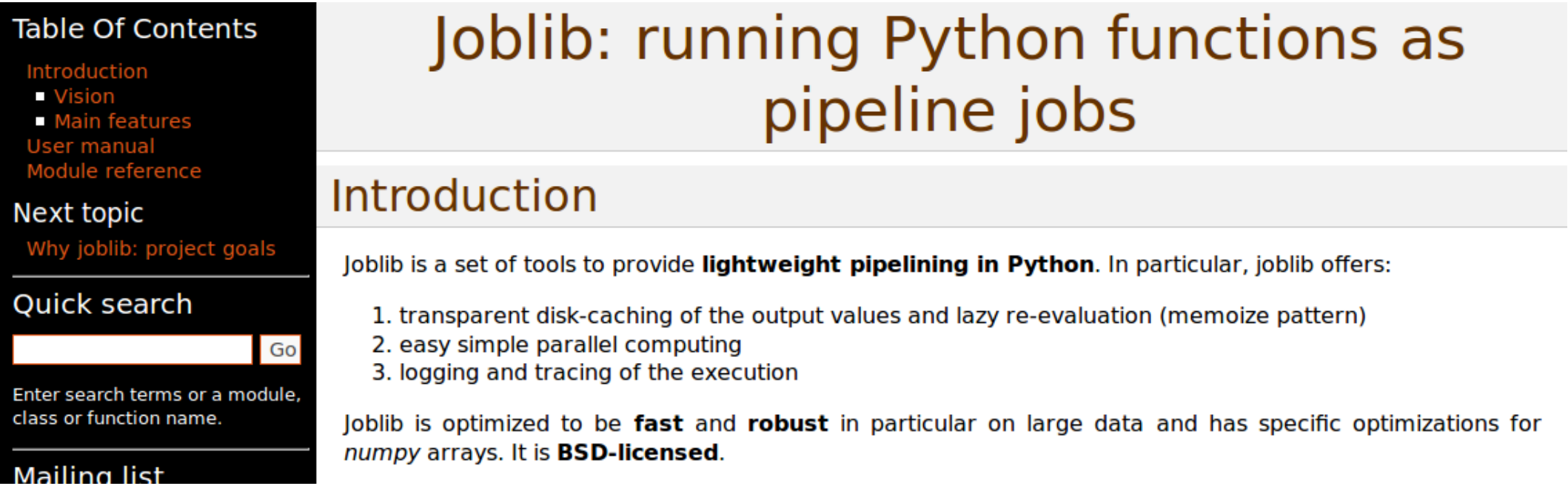


Table Of Contents

Introduction
- Vision
- Main features

User manual
Module reference

Next topic
Why joblib: project goals

Quick search
Go
Enter search terms or a module, class or function name.

Mailing list

Joblib: running Python functions as pipeline jobs

Introduction

Joblib is a set of tools to provide **lightweight pipelining in Python**. In particular, joblib offers:

1. transparent disk-caching of the output values and lazy re-evaluation (memoize pattern)
2. easy simple parallel computing
3. logging and tracing of the execution

Joblib is optimized to be **fast** and **robust** in particular on large data and has specific optimizations for *numpy* arrays. It is **BSD-licensed**.

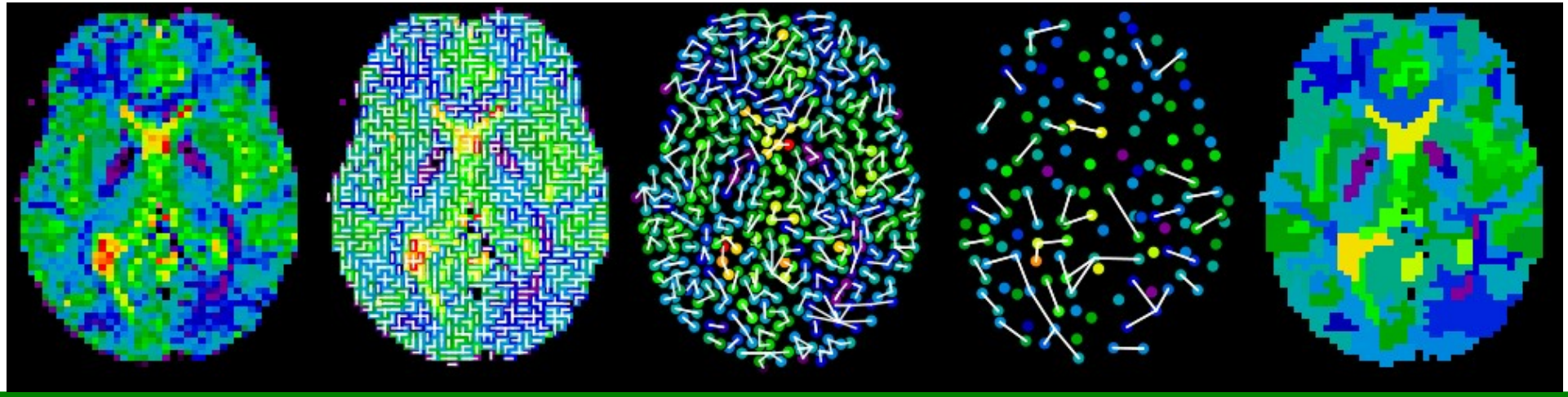
```
$ from nilearn.decomposition.dict_learning import DictLearning
$ dict_learning = DictLearning(n_components=n_components, smoothing_fwhm=6.,
                               memory="nilearn_cache", memory_level=5,
                               threshold=1., verbose=2, random_state=0,
                               n_jobs=8, n_init=1, alpha=3, n_iter=1000)
```

Fast image compression

- Cost of Logistic regression $\propto pn^2$

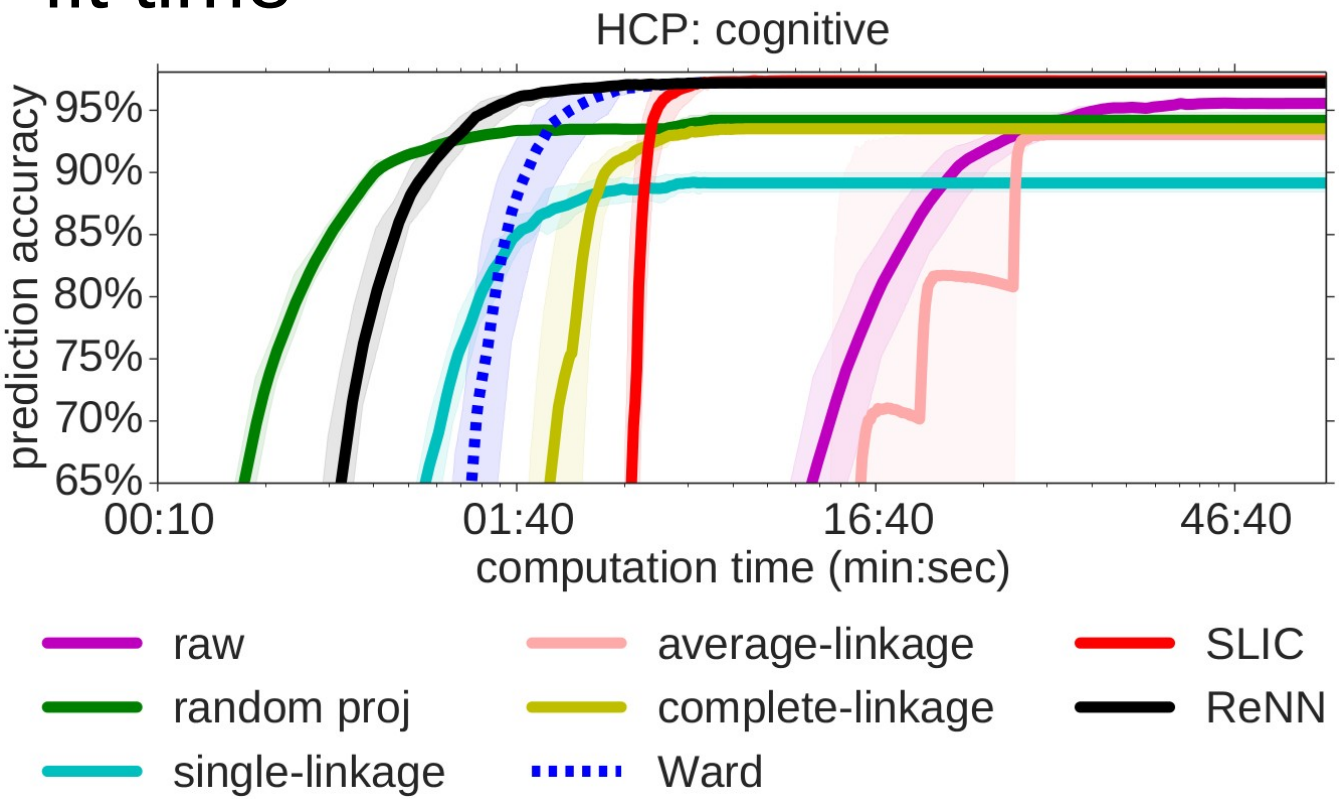
  + memory consumption $\propto pn$

- Solution: reduce p by compression
  - Random projections [Johnson, Lindenstrauss, AMS 1984]
    - Explicit control on precision
  - Clustering: informed by data [Thirion et al. Stamlins 2015]



Efficient discriminative models

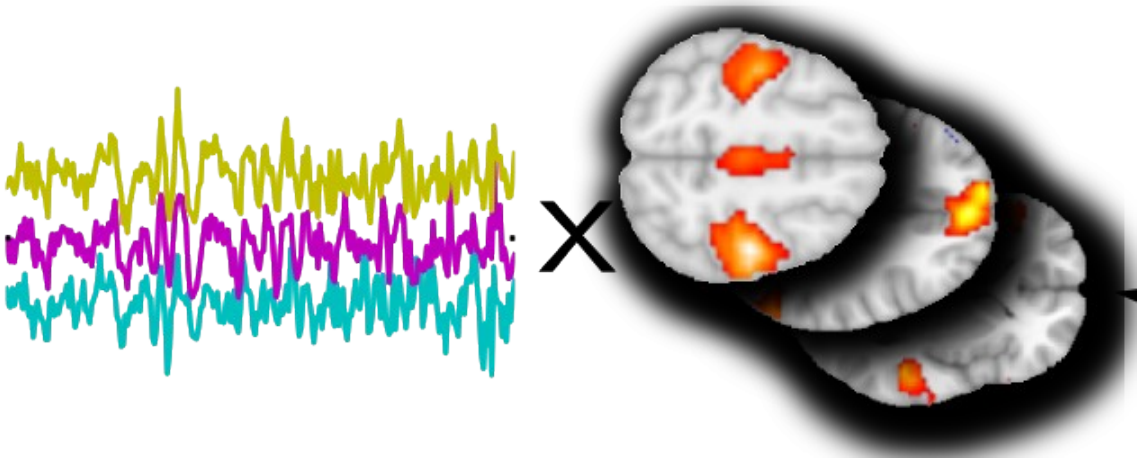
Example: HCP dataset, multi-class prediction (k=18), clustering + fit time



- Time savings are more than linear wrt data volume
- Data compression also benefits to accuracy

(Sparse) Principal Components Analysis

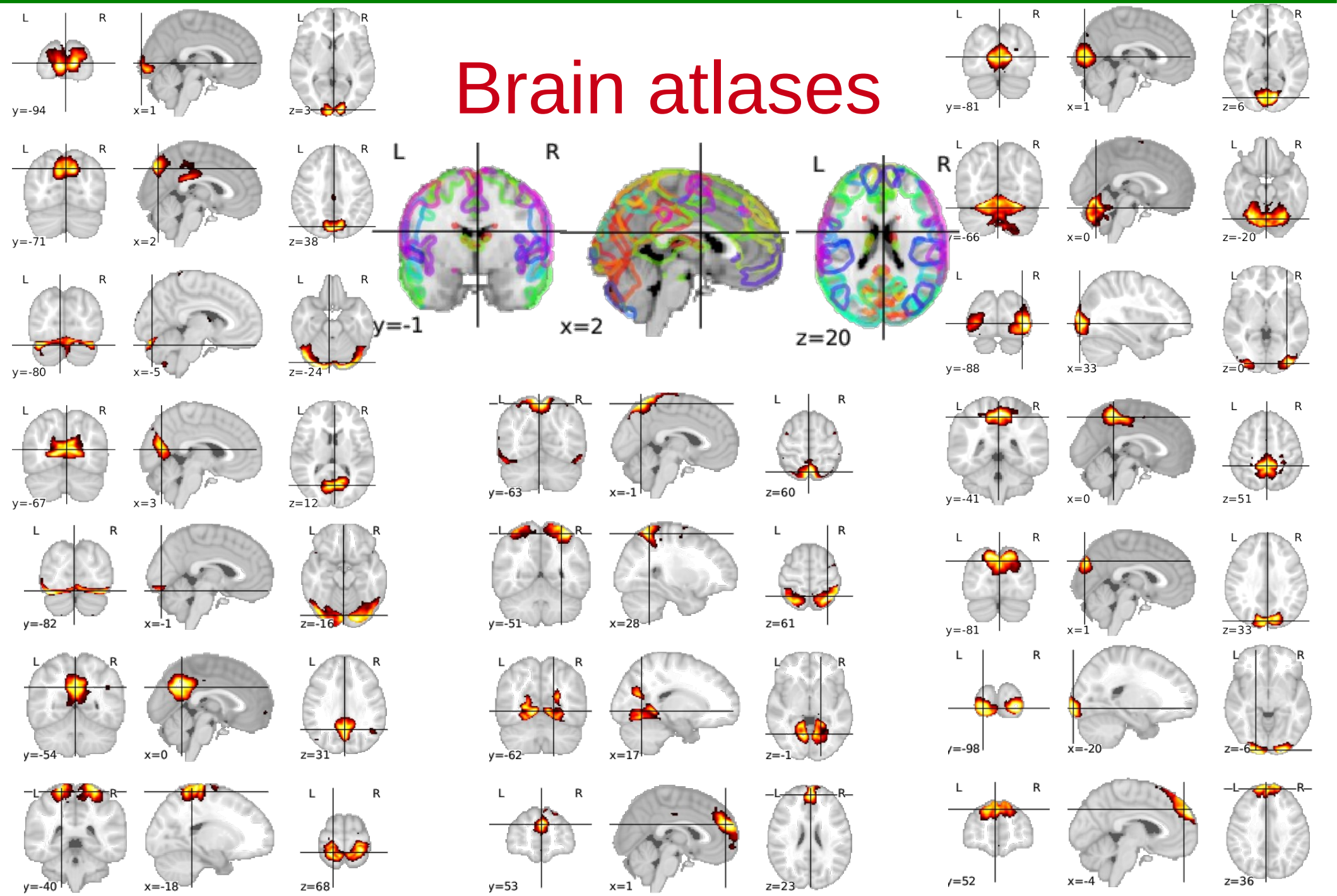
- Identify structure in data with *dictionary learning / sparse principal components analysis*



- Memory cost: n.p → scales poorly with large n (large number of subjects): 2TB of data on HCP
  - Run this on computers with 8GB of RAM
  - Possible with stochastic (online) methods !

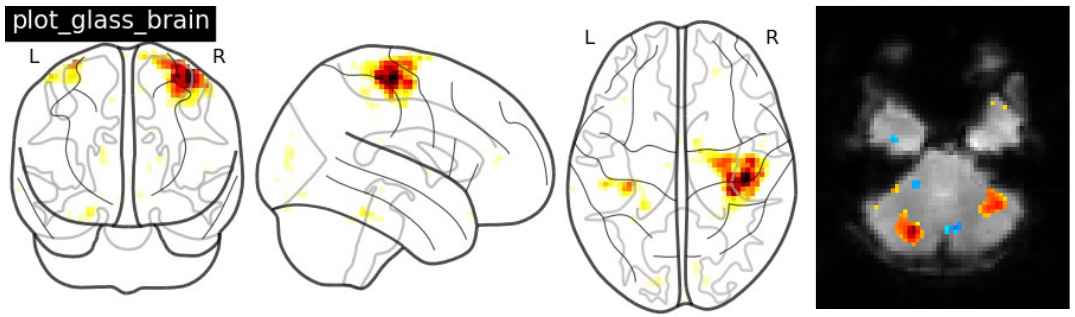
    [Mensch et al. ISBI 2016, ICML 2016]

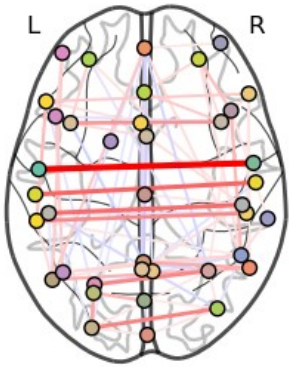
Brain atlases



Summary

- Large(r) sample sizes to make imagine findings more reliable
  - Even though less homogeneous data
- Combining multi-modal data to improve statistical analysis
- Improve software to support large data analysis
  - Reduce memory requirements: online methods
  - Leverage parallel computers





Acknowledgements

Parietal
G. Varoquaux,
P. Ciuciu
A. Gramfort,
O. Grisel,
L. Estève,
Y. Schwartz,
F. Pedregosa,
E. Dohmatob,
A. H. Idrobo,
V. Fritsch,
M. Eickenberg,
R. Bricquet
A. Abraham
D. Bzdok
A. Frau
N. Chauffert

Other collaborators
(thanks for the data)
S. Dehaene
E. Eger,
R. Poldrack,
K. Jimura,
J. Haxby
C. F. Gorgolevski
JB. Poline

Brainpedia project

Microsoft Research - Inria
JOINT CENTRE

Human Brain Project