ISMRM 27TH ANNUAL MEETING & EXHIBITION
Palais des congrès de Montréal · Montréal, QC, Canada · 11–16 May 2019

Best Practices & Pitfalls in Applying Machine Learning to Magnetic Resonance Imaging

Thomas Moreau

thomas.moreau@inria.fr

Declaration of
Financial Interests or Relationships

Speaker Name: Thomas Moreau

I have no financial interests or relationships to disclose with regard to the subject matter of this presentation.

Outline

1) Supervised Learning

2) Model selection and cross-validation

3) Weakly supervised learning

4) Large models computationnal tradeoff

Supervised Learning

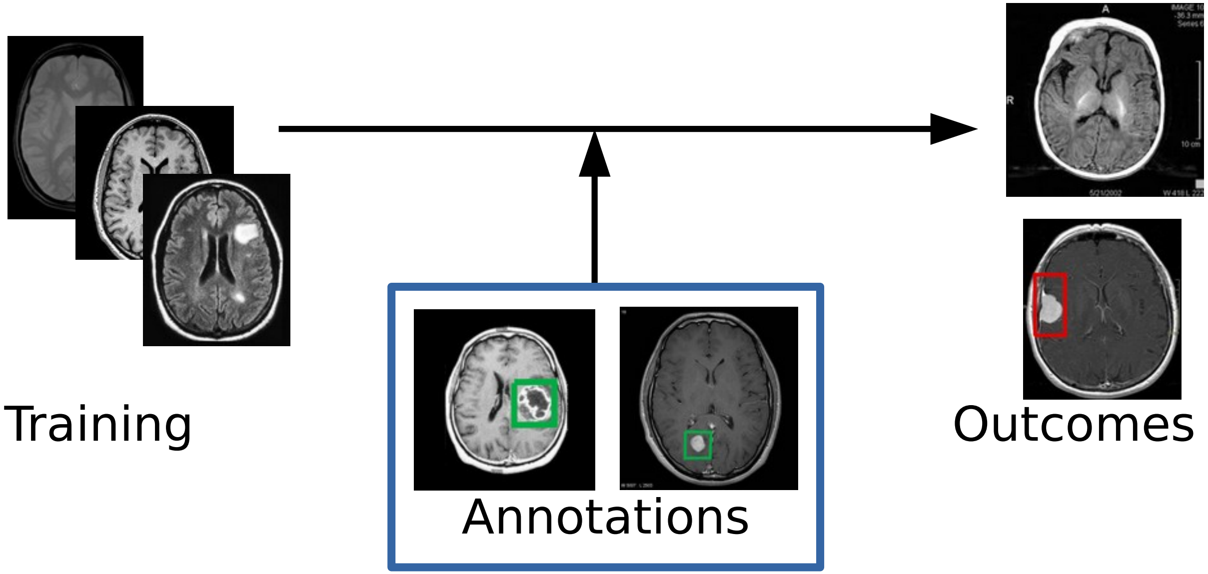
- Classical machine learning framework
  - Google, Facebook, amazon, …
  - Computer Vision, Speech processing, …
  - Millions/Billions of samples
  - Lots of annotations

Supervised Learning

- Classical machine learning framework
- From annotated data, predict an outcome

Supervised Learning

- From annotated data, predict an outcome



Empirical Risk Minimization

Data distribution: $X, y \sim \mathcal{P}$

Training set: $\{X_k, y_k\}_{k=1}^n$

Model: $\widehat{y} = f_\theta(X)$

Loss: $\ell(\widehat{y}, y)$

Risk minimization

$$\min_\theta E[\ell(f_\theta(X), y)]$$

Empirical Risk Minimization

Data distribution: $X, y \sim \mathcal{P}$

Training set: $\{X_k, y_k\}_{k=1}^n$

Model: $\widehat{y} = f_\theta(X)$

Loss: $\ell(\widehat{y}, y)$

Risk minimization

$$\min_\theta E[\ell(f_\theta(X), y)]$$

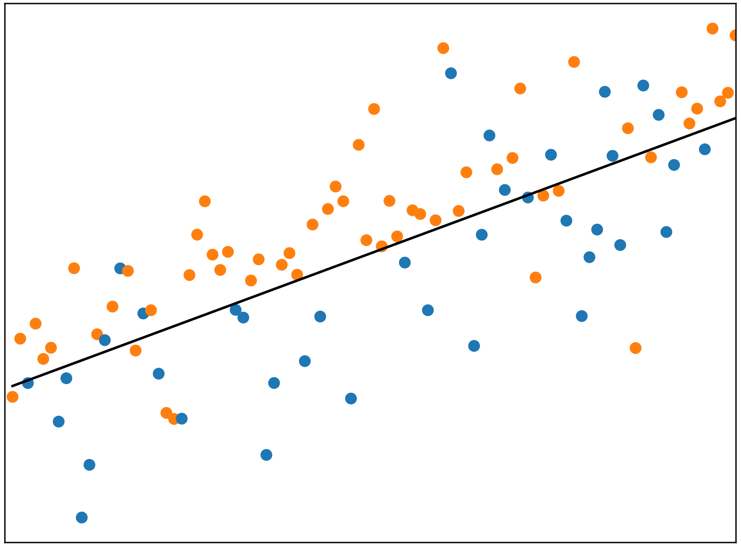
Empirical Risk Minimization

$$\min_\theta \frac{1}{n} \sum_{k=1}^n \ell(f_\theta(X_k), y_k)$$

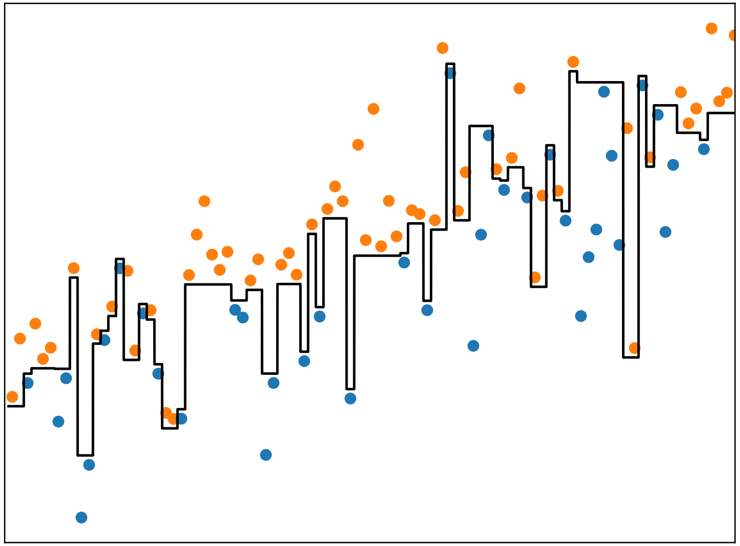
Model Selection

- Binary classification:

Linear model



Decision tree



Model Selection: Generalization

Data distribution: $X, y \sim \mathcal{P}$

Training set: $\{X_k, y_k\}_{k=1}^n$

Model: $\widehat{y} = f_\theta(X)$

Loss: $\ell(\widehat{y}, y)$

Risk minimization

Empirical Risk Minimization

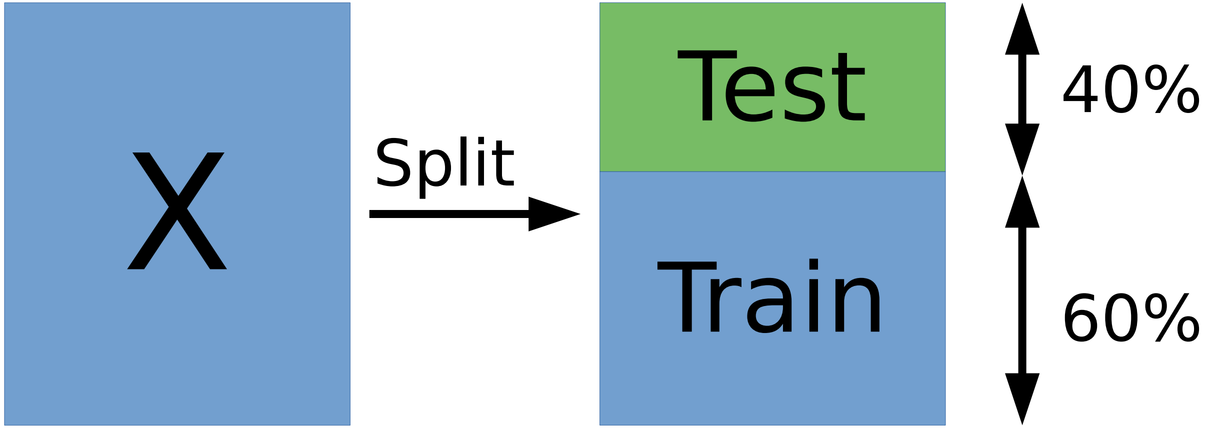
$$\min_\theta E[\ell(f_\theta(X), y)] \neq \min_\theta \frac{1}{n} \sum_{k=1}^{n} \ell(f_\theta(X_k), y_k)$$

Generalization: measure this discrepancy

Model Selection: Generalization

Measuring the generalization: **Test set**

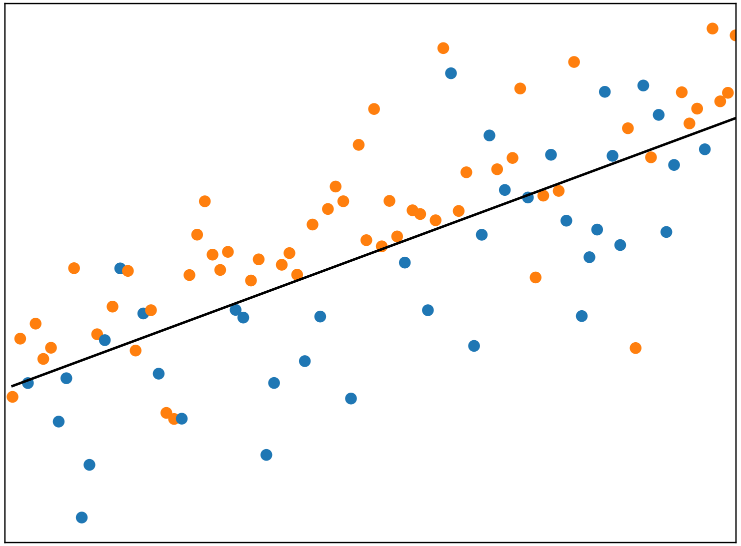
- Split the data in 2 parts:
  - Train the model on one part
  - Evaluate the model on **unseen** and **independent** data



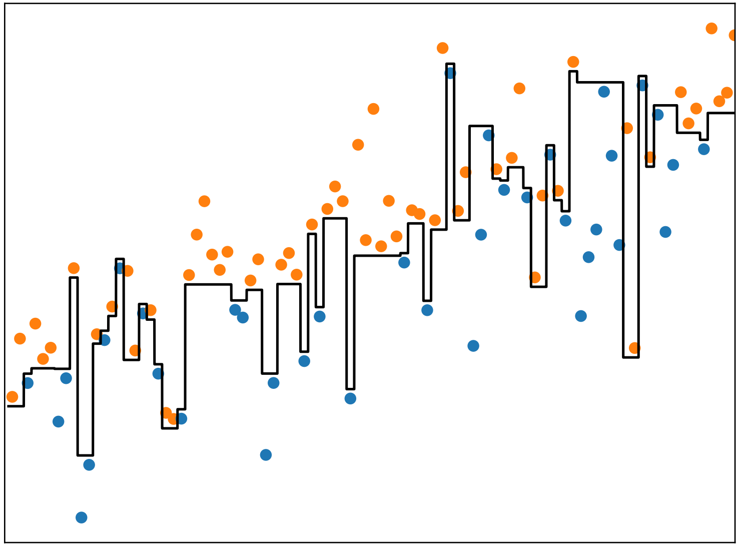
Model Selection

- Binary classification:

Linear model



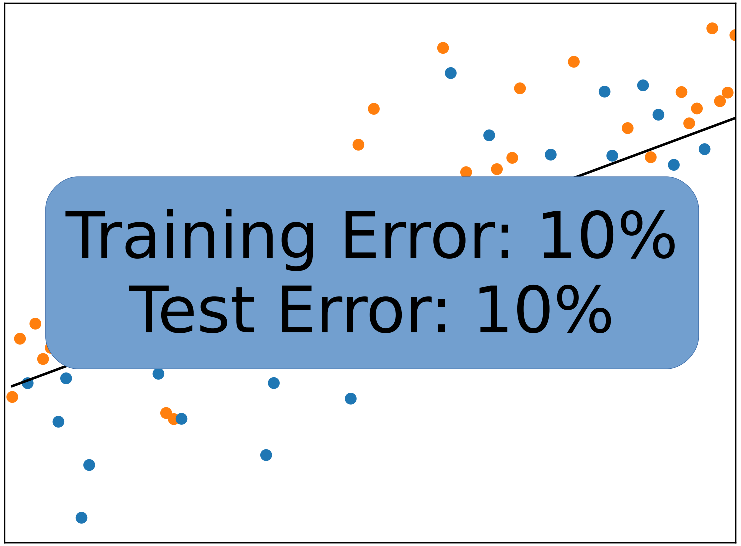
Decision tree



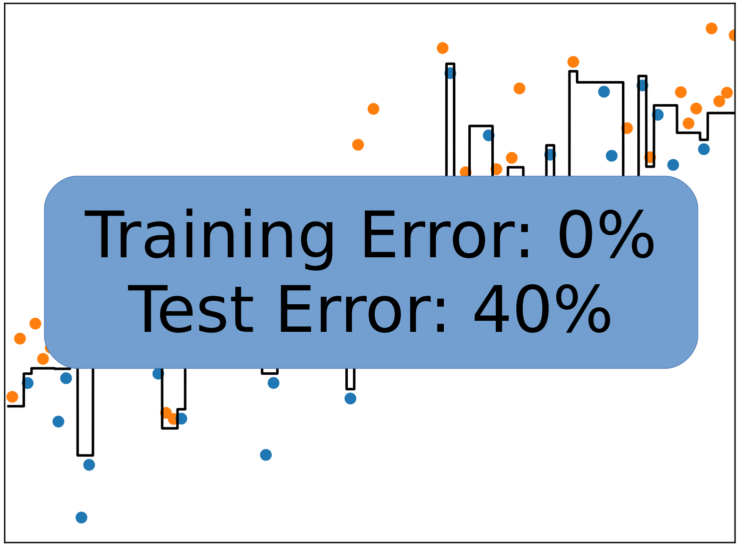
Model Selection

- Binary classification:

Linear model



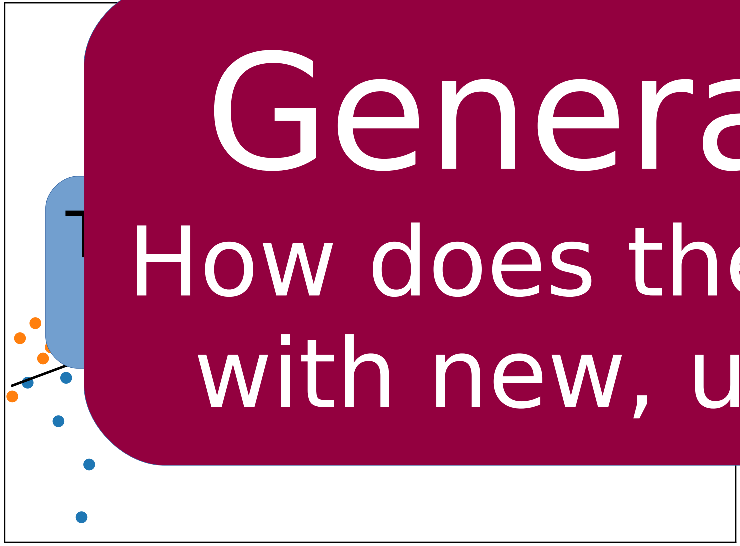
Decision tree



Model Selection

- Binary classification:

Linear model



Decision tree

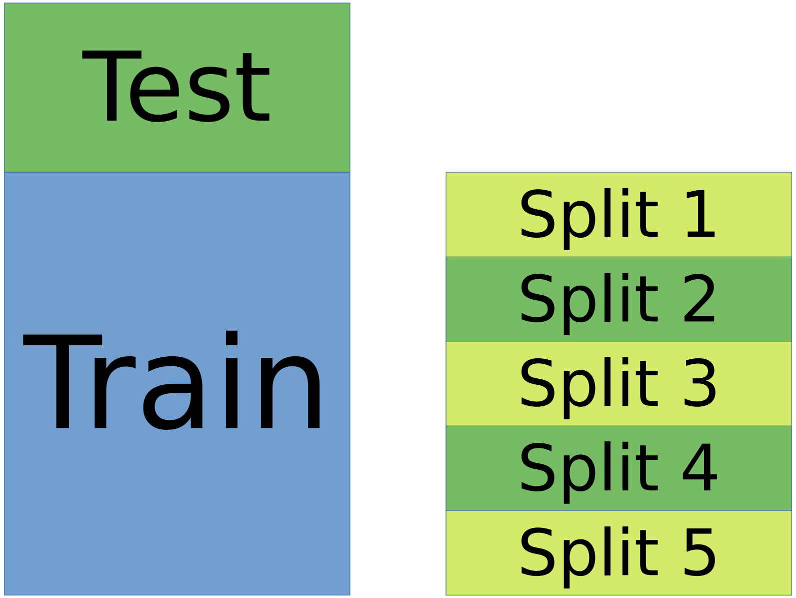
Generalization:

How does the model behave with new, unlabelled data

Model Selection: Generalization

Generalization for model selection:

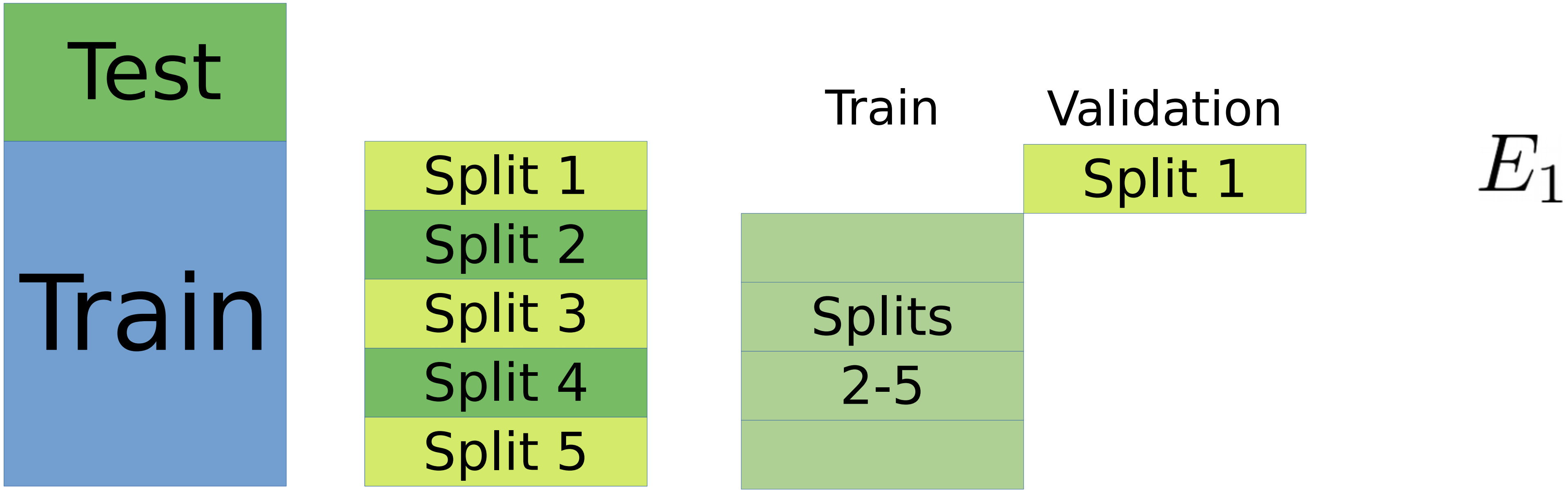
Cross validation



Model Selection: Generalization

Generalization for model selection:

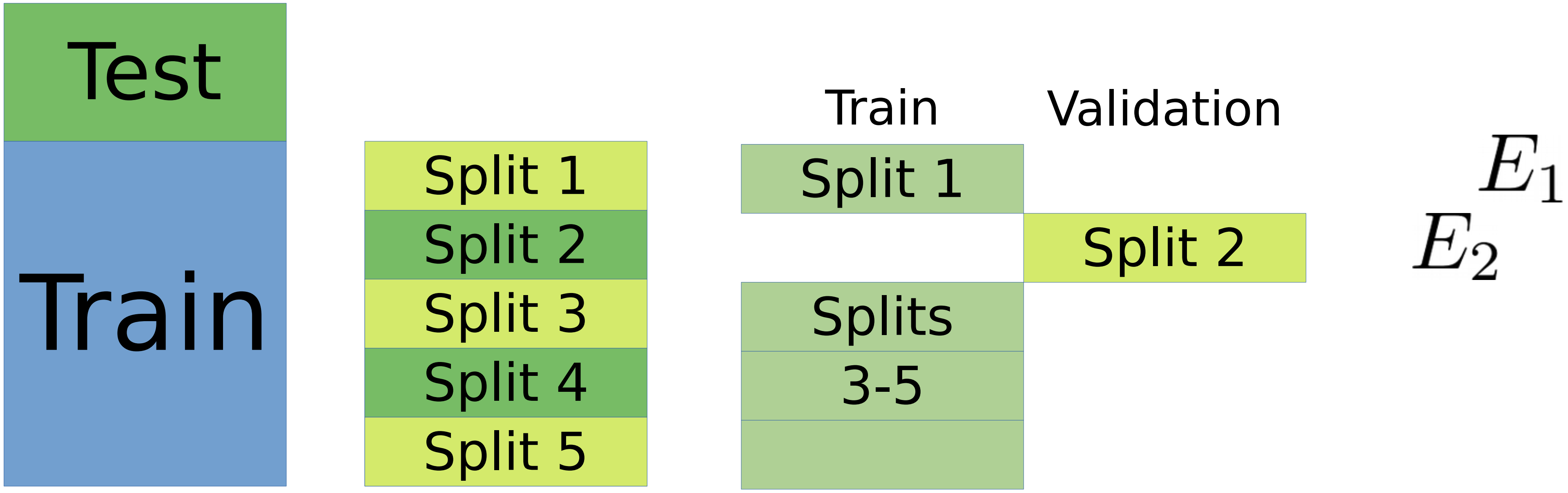
Cross validation



Model Selection: Generalization

Generalization for model selection:

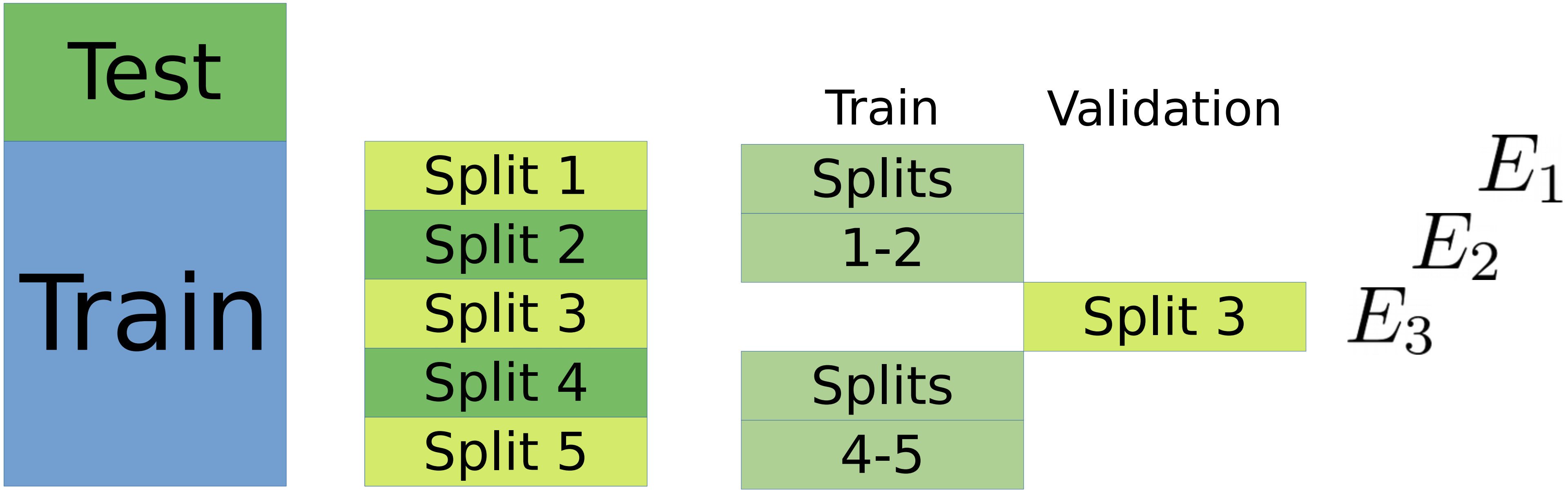
Cross validation



Model Selection: Generalization

Generalization for model selection:

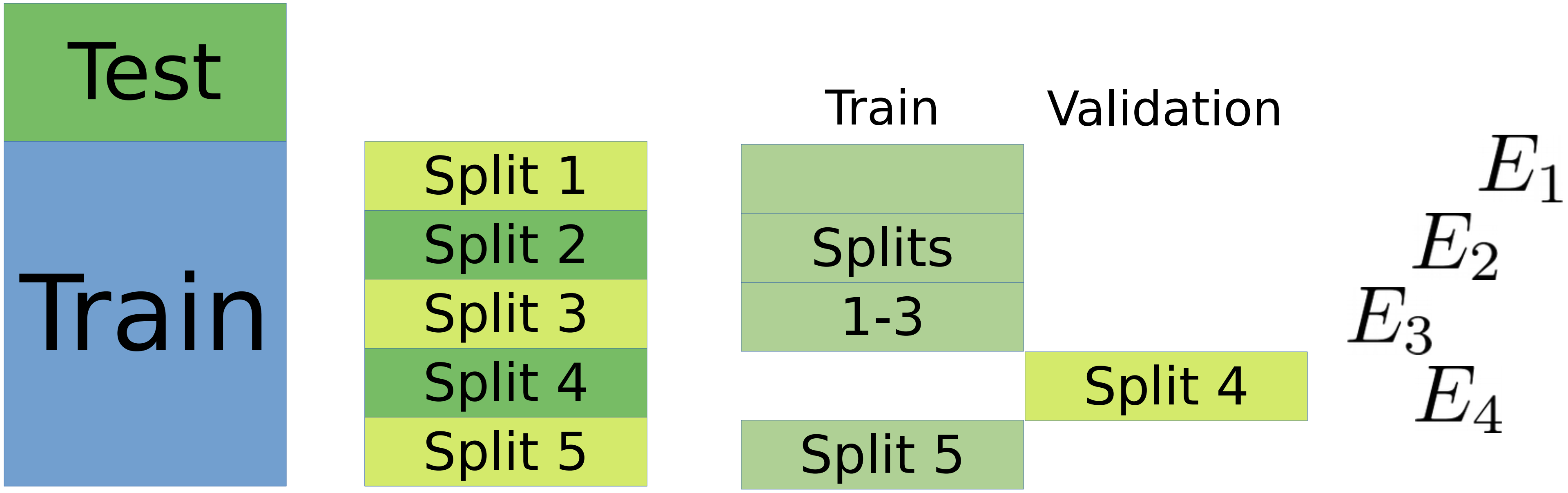
Cross validation



Model Selection: Generalization

Generalization for model selection:

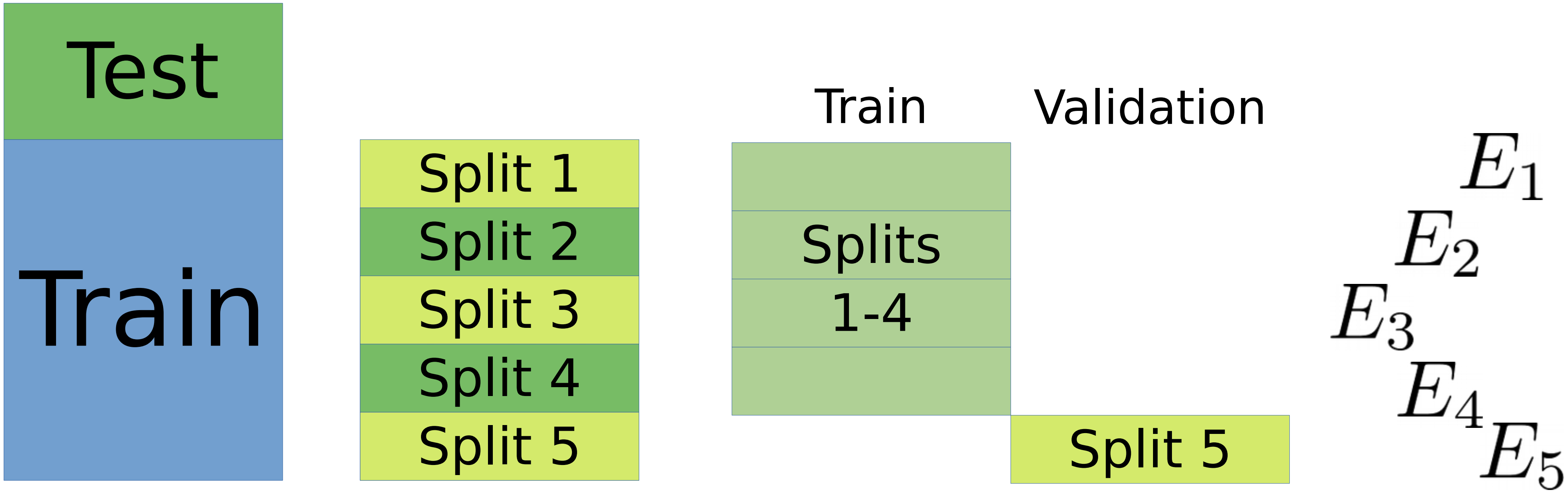
Cross validation



Model Selection: Generalization

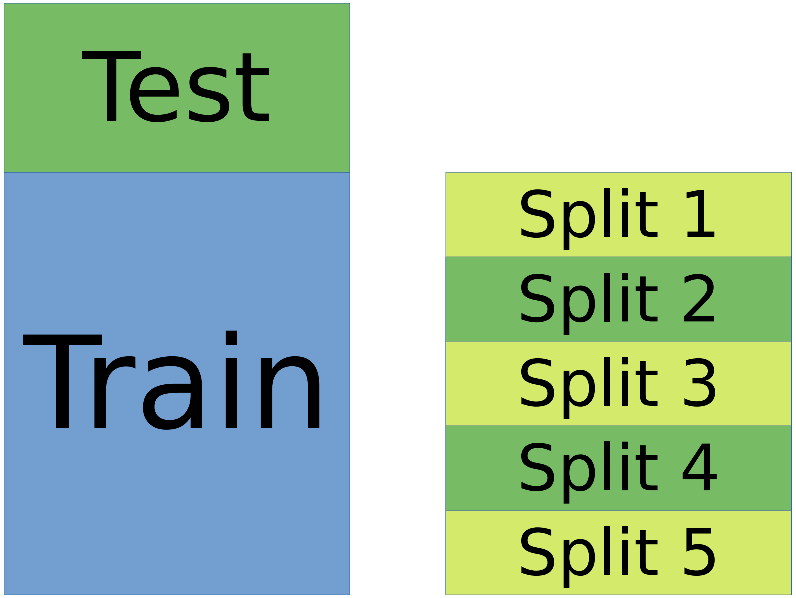
Generalization for model selection:

Cross validation



Model Selection: Generalization

Generalization for model selection:
Cross validation



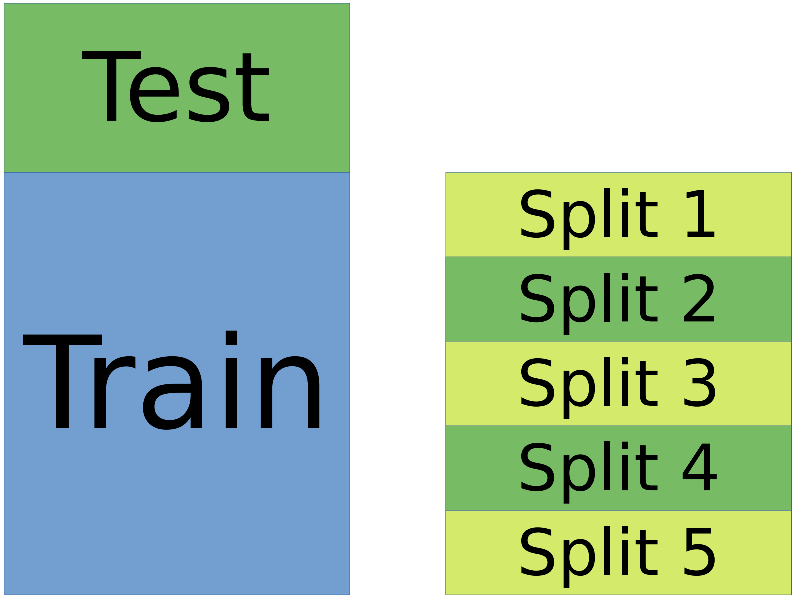
$$\hat{E}(\theta) = \frac{1}{5}\sum_{k=1}^{5} E_k(\theta)$$

E_1
E_2
E_3
E_4
E_5

Model Selection: Generalization

Generalization for model selection:

Cross validation



$$\hat{E}(\theta) = \frac{1}{5}\sum_{k=1}^{5} E_k(\theta)$$

$$\theta^* = \arg\min_{\theta} \hat{E}(\theta)$$

$$E_{test} = \frac{1}{N_{test}}\sum_{k=1}^{N_{test}} \ell(f_{\theta^*}(X_k), y_k)$$

Model Selection: Generalization

Generalization for model selection:

Cross validation

Evaluate the risk with left out data

Splitting strategies:

- Leave-one-out (LOO)
- Random splits
- Stratified

Model Selection: Sample size

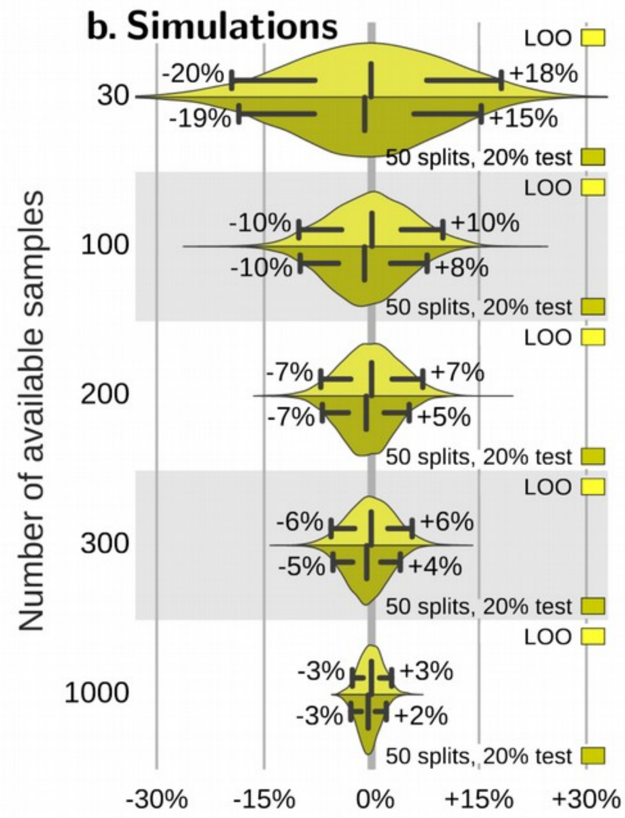
- Uncertainty of CV

  Sample size

- X drawn from 2 Gaussian
- Display the difference:

$$E(\theta) - E_{test}$$

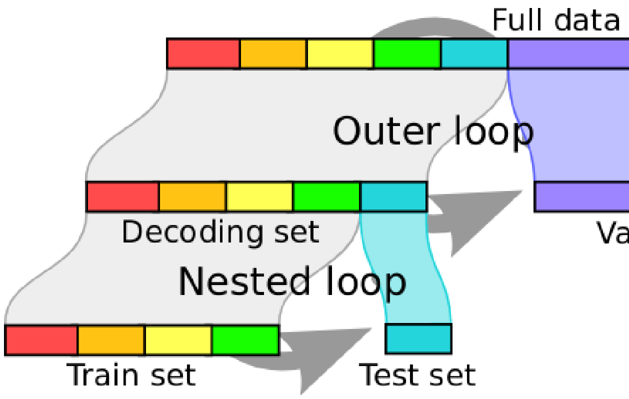
for large N_{test}=10000

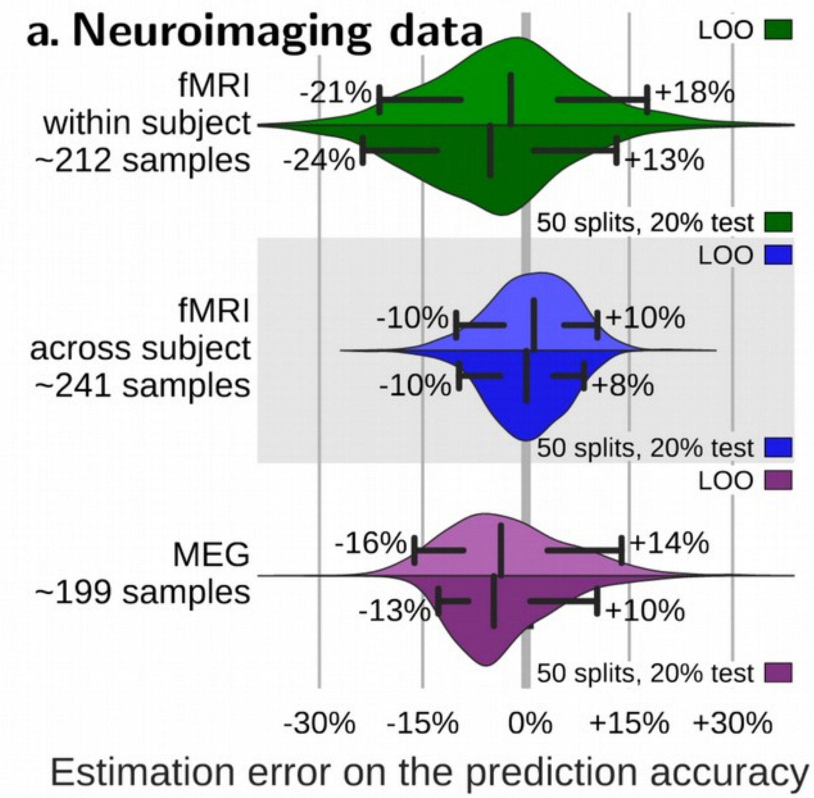


[Varoquaux, 2017]

Model Selection: Sample size

- CV and test error discrepancy
  - fMRI
  - MEG





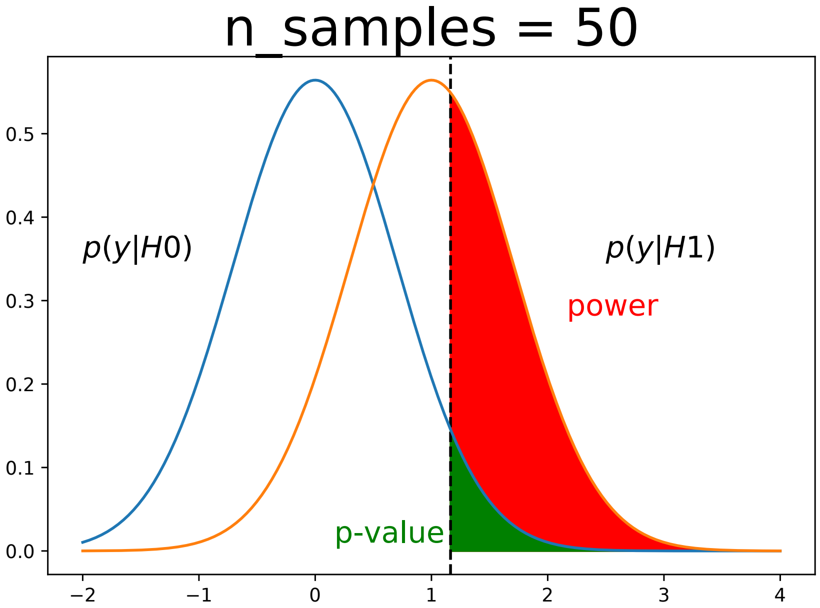
[Varoquaux, 2017]

Central Limit Theorem

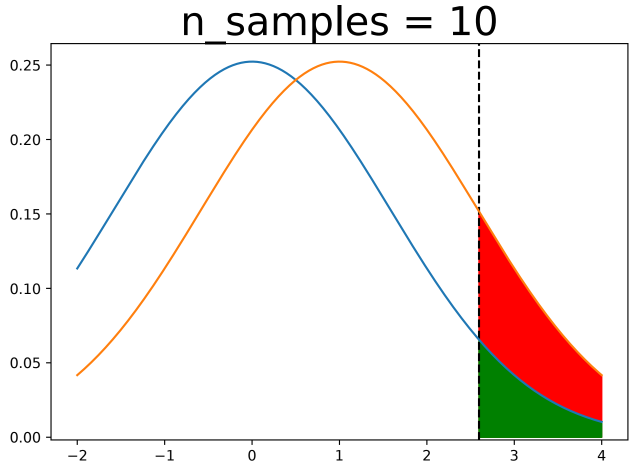
- For $\{X_1, \ldots, X_n\}$ i.i.d random variables
- If $E[X_1] = \mu$ and $E[(X_1 - \mu)^2] = \sigma^2$
- Then $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ verifies

$$\sqrt{n}(S_n - \mu) \xrightarrow[n\to\infty]{} \mathcal{N}(0, \sigma^2)$$

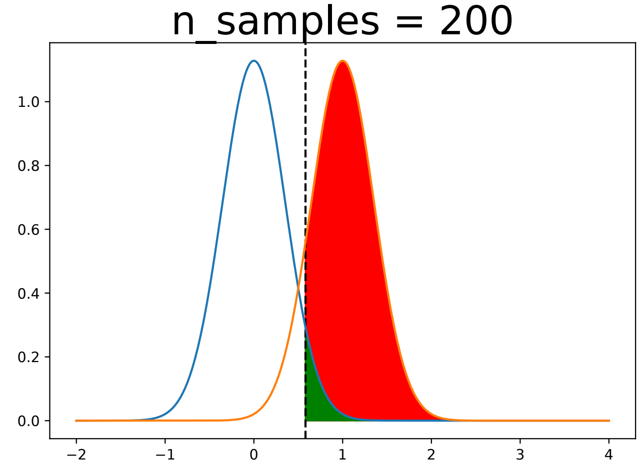
Sample size effect



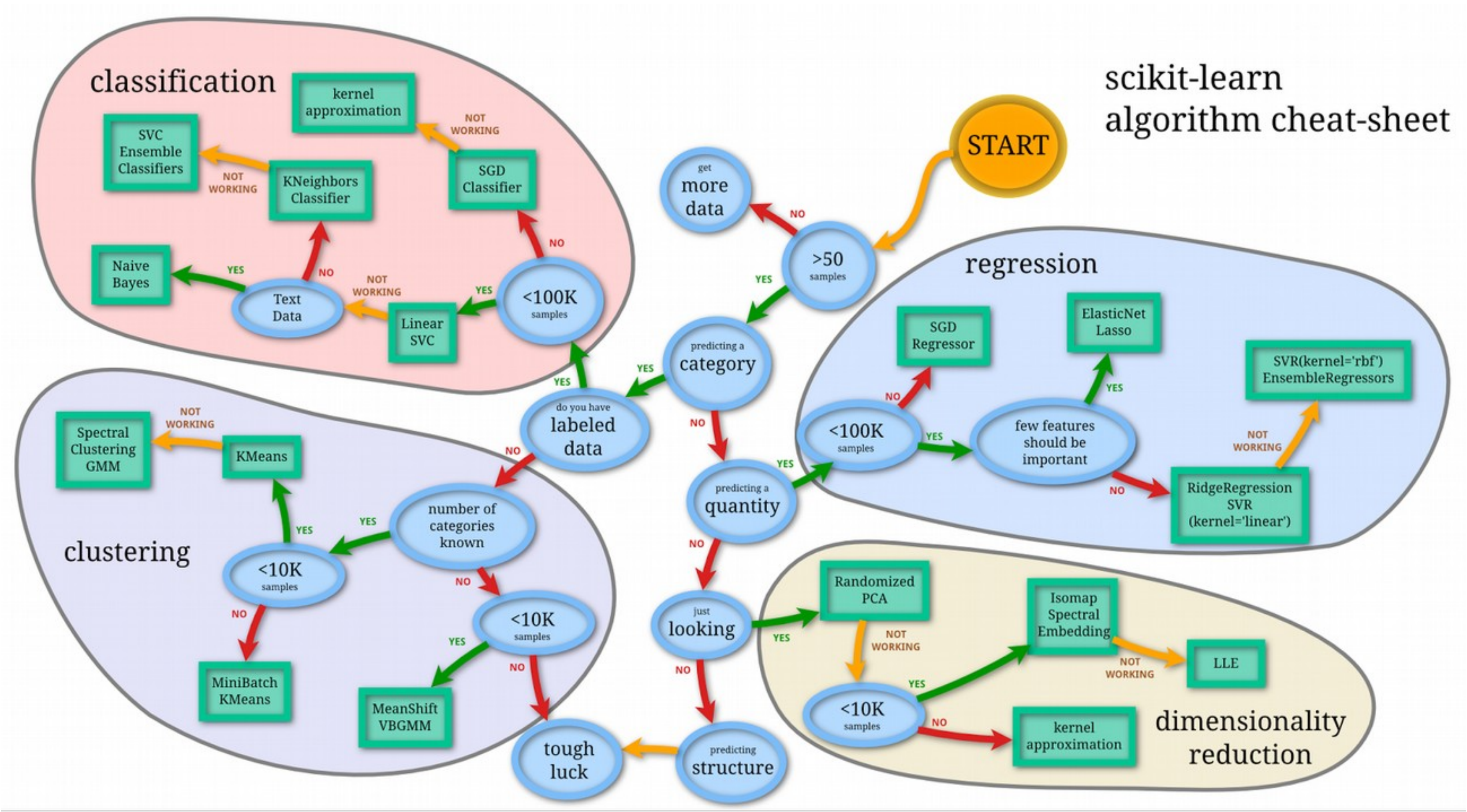
Less subjects



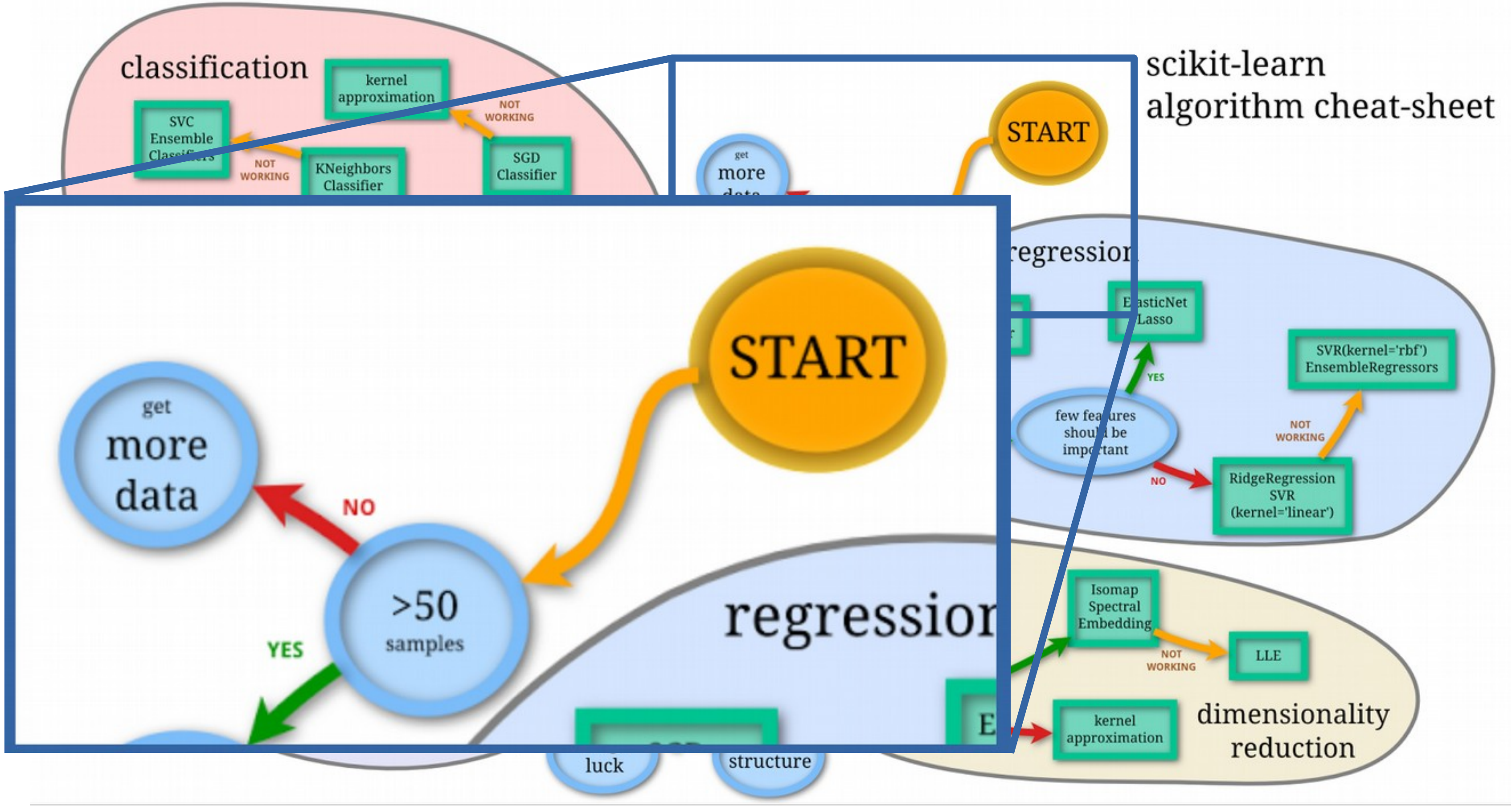
More subjects



Robust machine learning



Robust machine learning



Part 1: Take home messages

- **Test data** is necessary

Validate on independent data

- **Larger sample size** is needed

Reliable findings

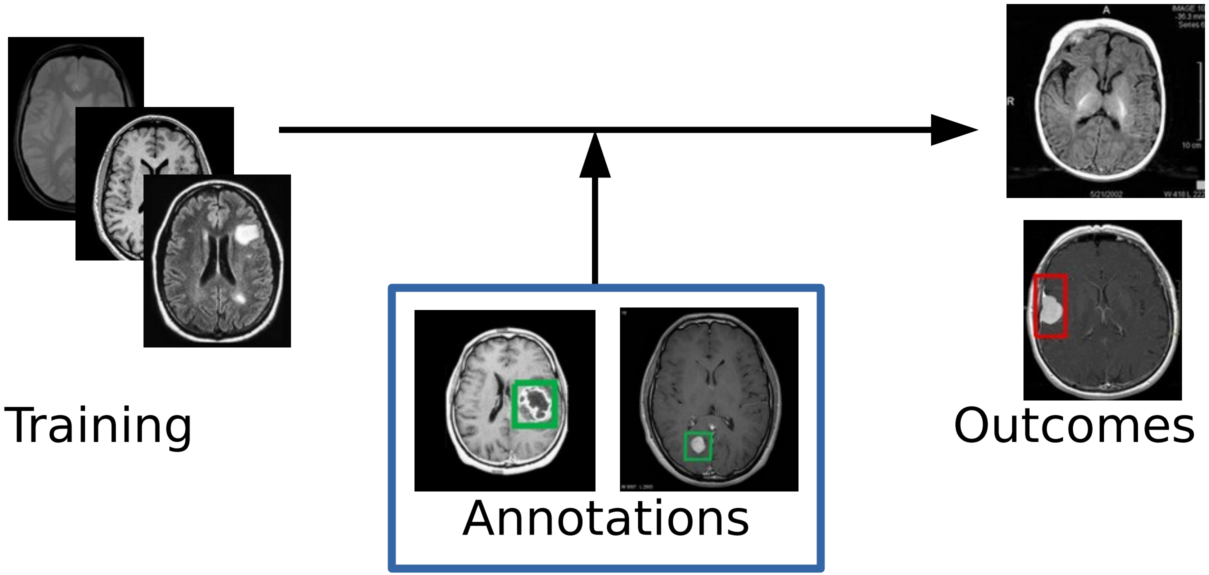
Reproducibility

Annotated data

- Humans are really good at classifying images, sounds...
- … but not that good for other data!
  - Vectorial data with more than 2D
  - fMRI: very high dimensional signals
- Not possible to have annotations of the brain functions and structures.

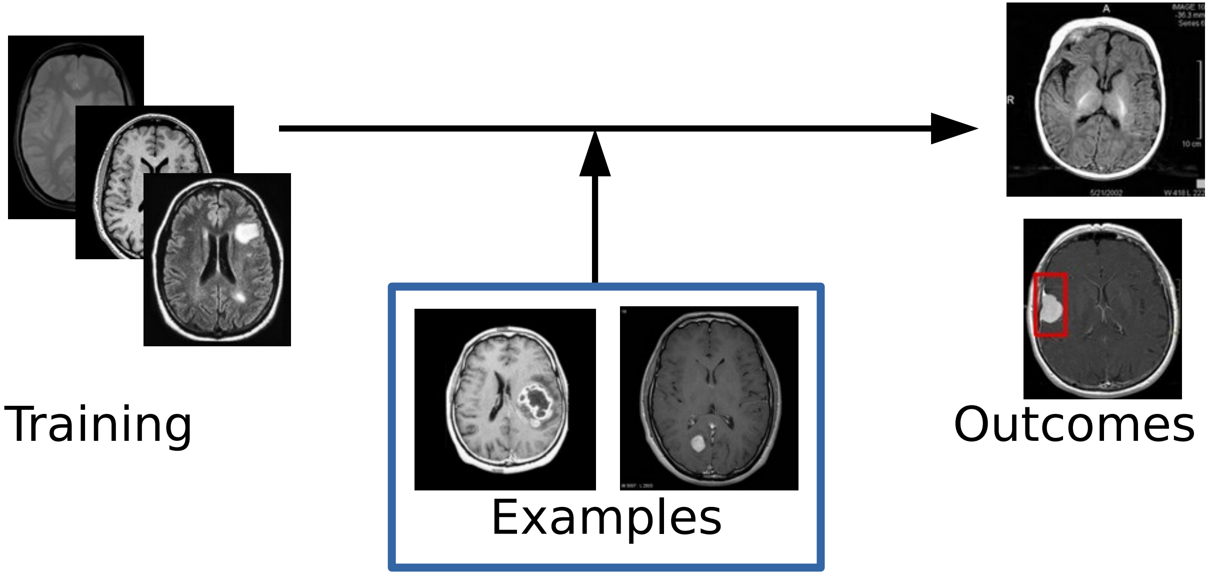
Weakly-supervised learning

- Labels are weakly related to the task we want to solve:



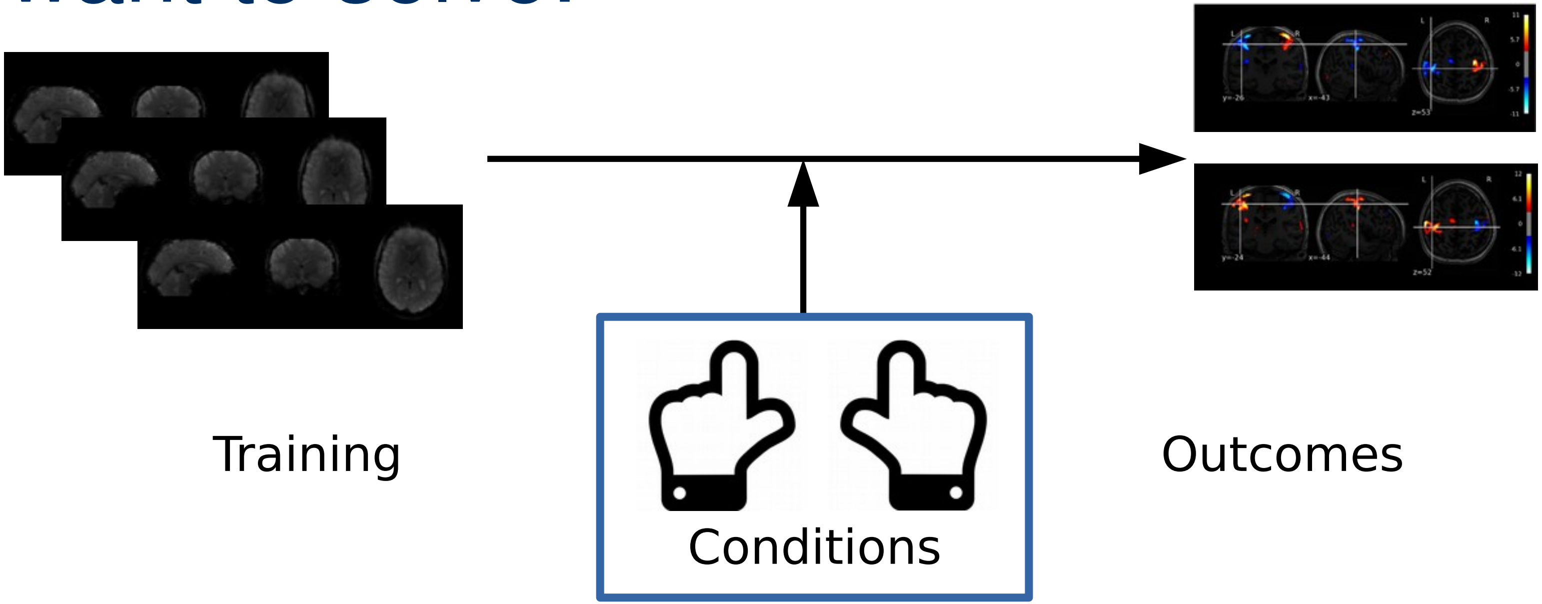
Weakly-supervised learning

- Labels are weakly related to the task we want to solve:



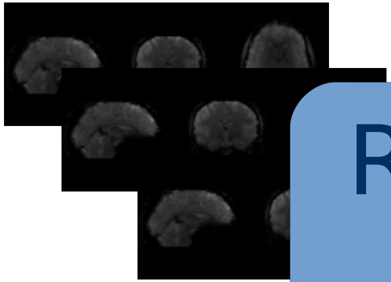
Weakly-supervised learning

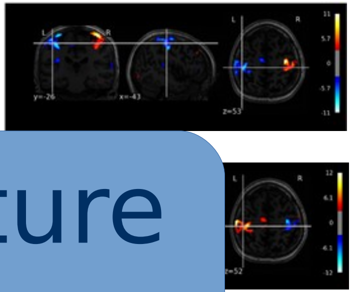
- Labels are weakly related to the task we want to solve:



Weakly-supervised learning

- Labels are weakly related to the task we want to solve:





Rely on the structure of the data!

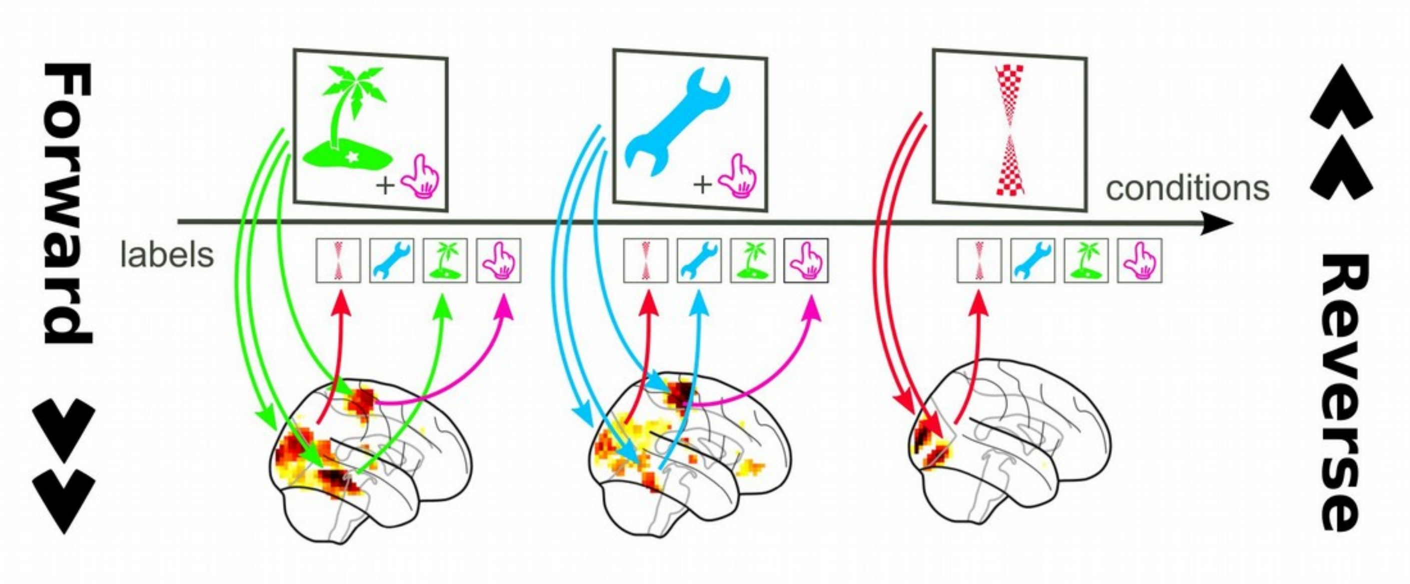
Training

Outcomes

Conditions

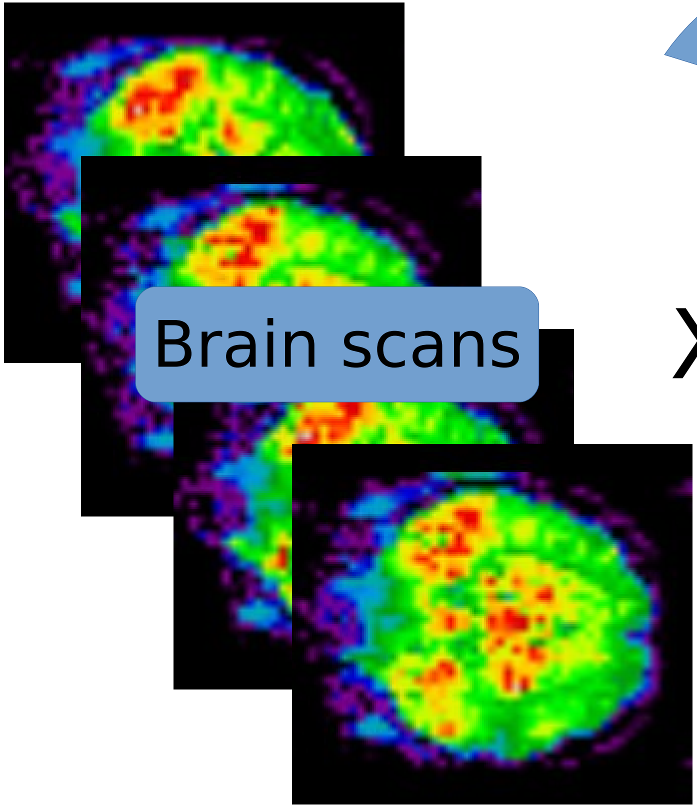
Large Scale Brain Mapping

- Predict conditions based on brain scan

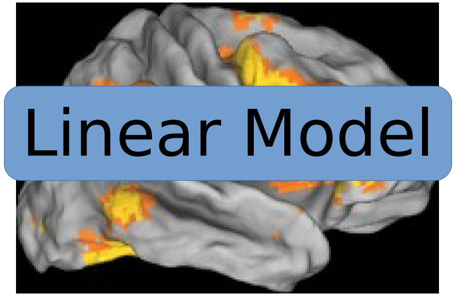


- Highlight functional maps

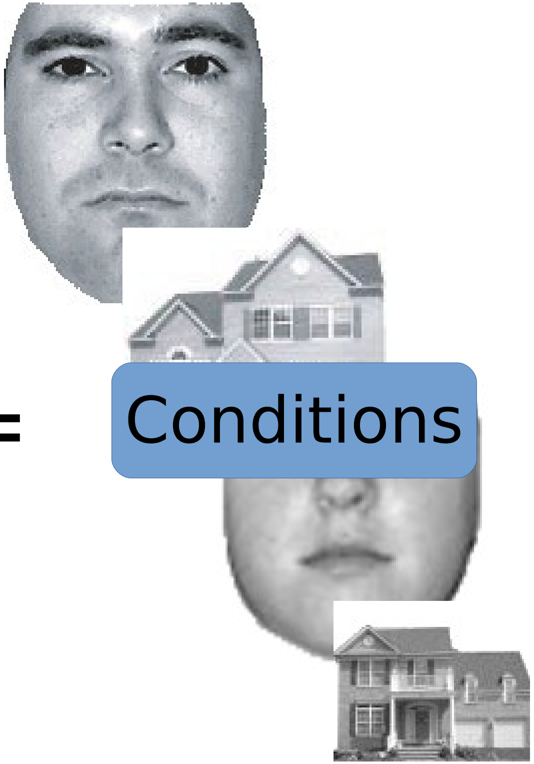
Large Scale Brain Mapping



X



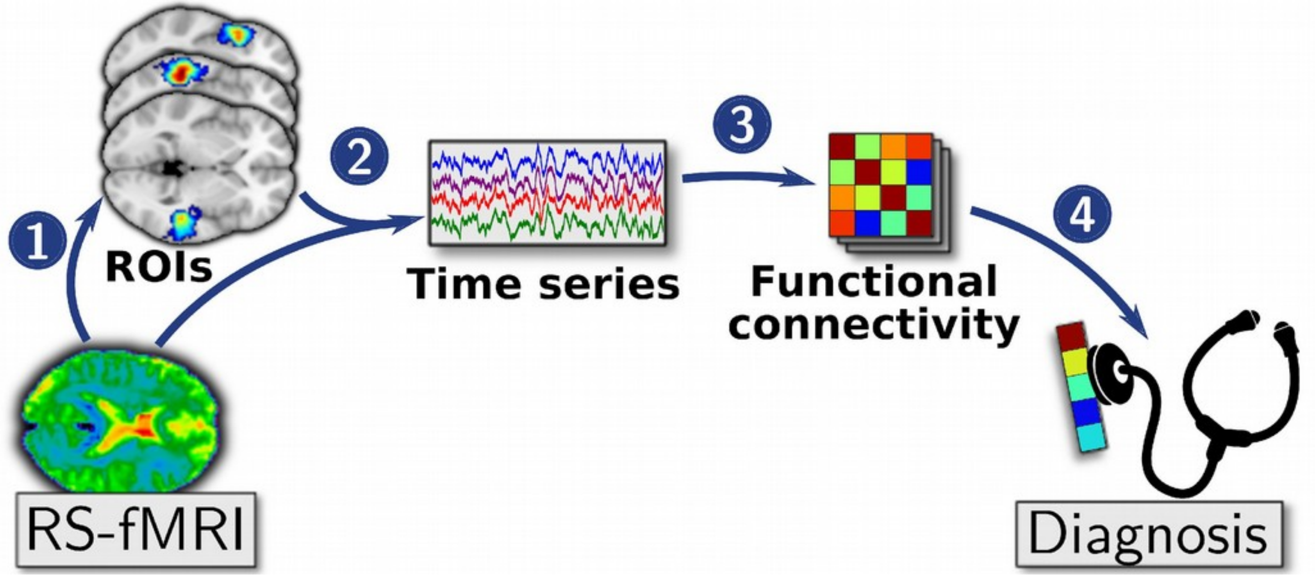
=



The linear model is
a **brain map**

Unsupervised Learning

- Resting State recordings
- Functional connectivity studies



Unsupervised Learning

- Resting State recordings
- Functional connectivity studies

ROI extraction relies on the structure of the data!

Big Data: Technical challenge

- New datasets provides larger sample size

  Camcam (650subjects), HCP (1,200 subjects),

  UKBB (5,000 subjects), …

- Very complex data
  - Large images: 10^5 to 10^6 voxels
  - Low SNR, structured noise, inter-subject variability,...

Dimension Reduction

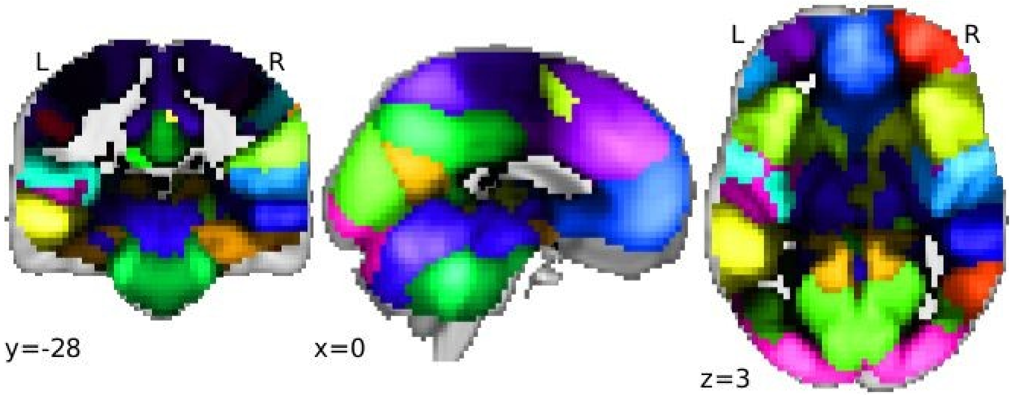
- Large n lead to exploding memory
- Computational bottleneck → memory
- Need to reduce dimension, *i.e.* # voxels
- Without losing too much info!

Reduce resolution

- Spatial averaging
  → averaging activity on regions
- Against the trend to go with larger resolution…
- **Smart** selection of related voxels?

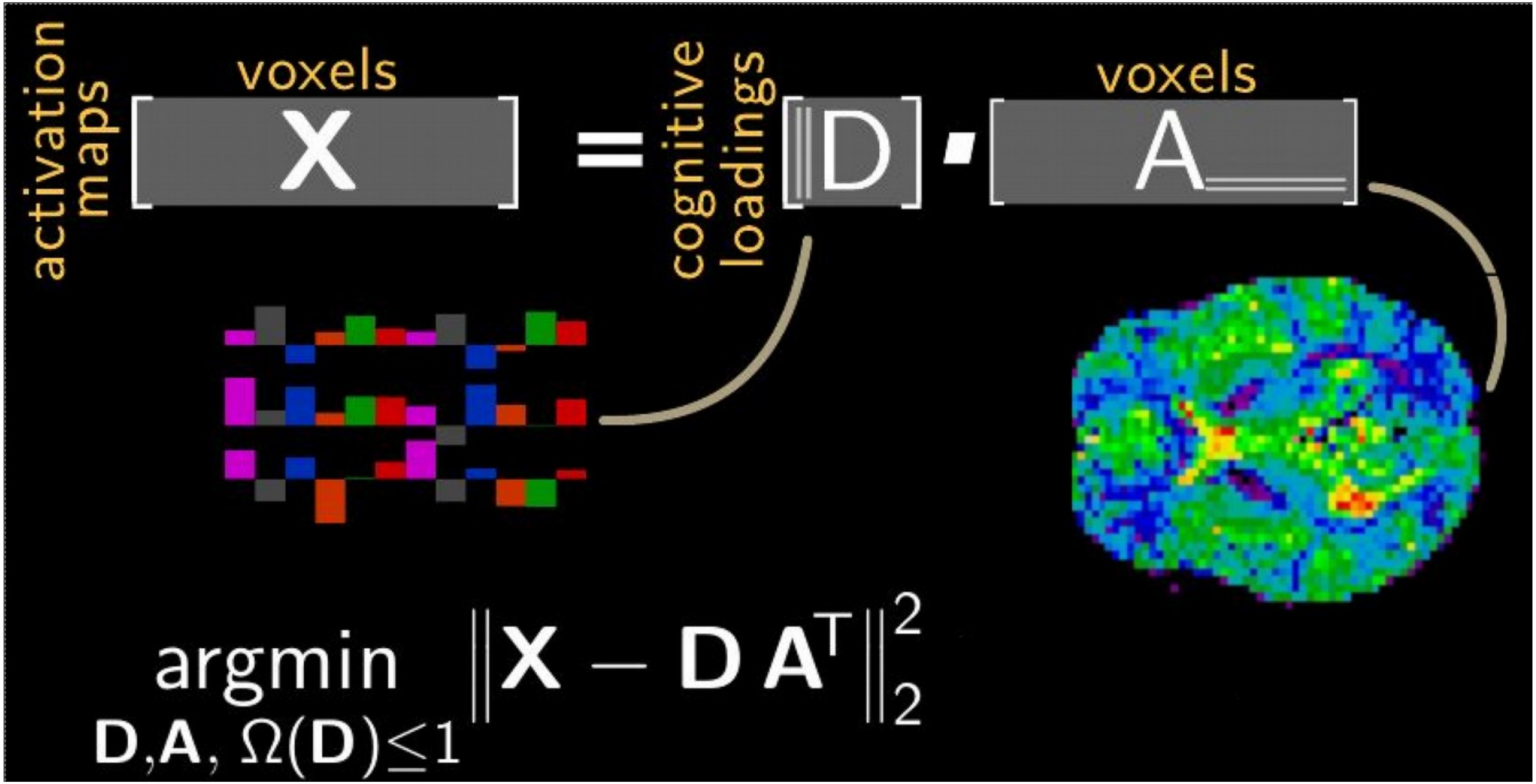
Fixed parcellation: Atlas

- Choose an Atlas
  - Destrieux 2009
  - Yeo 2011
  - MSDL (Varoquaux et al 2011)
  - Cradock 2012

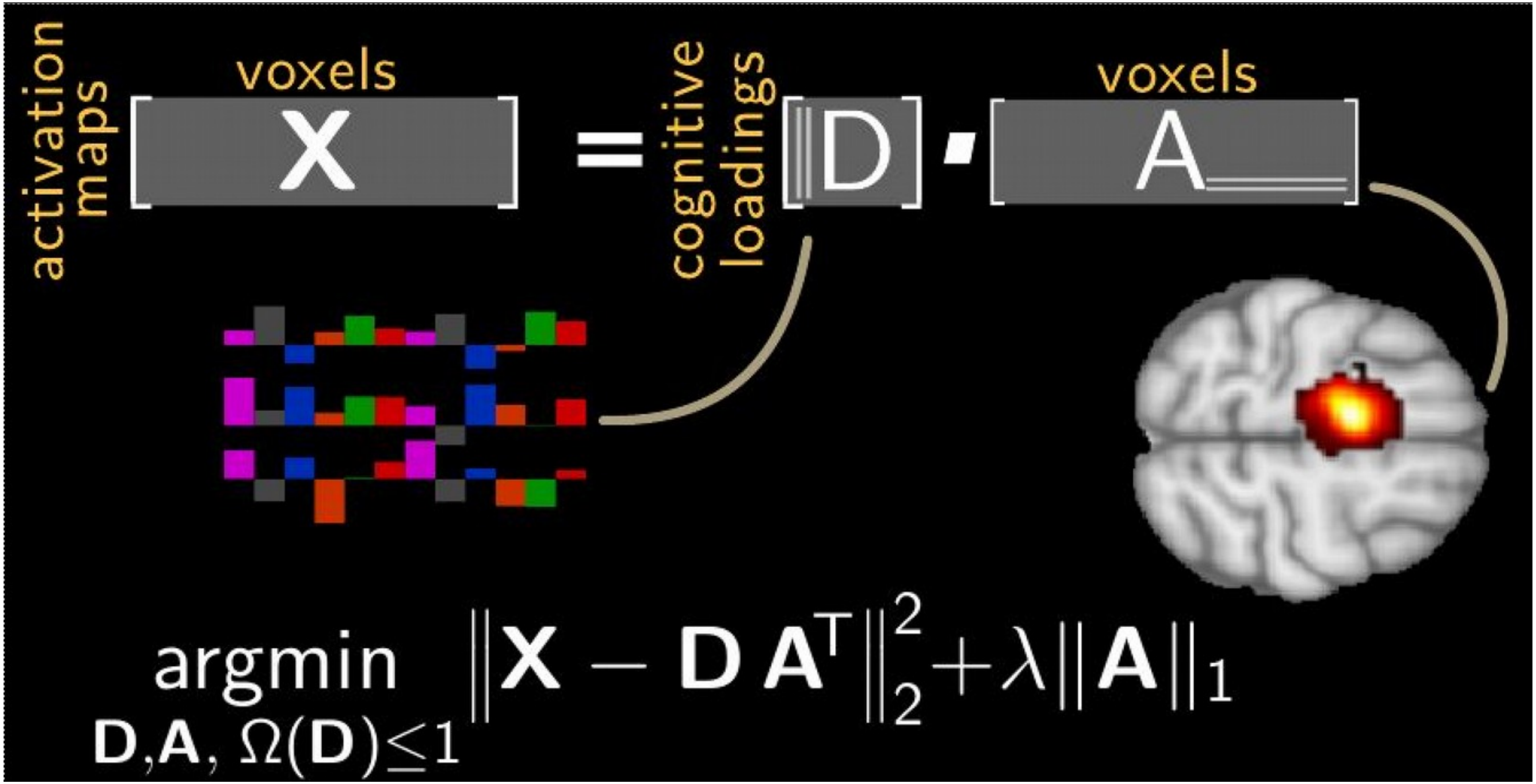


- Average over the parcels

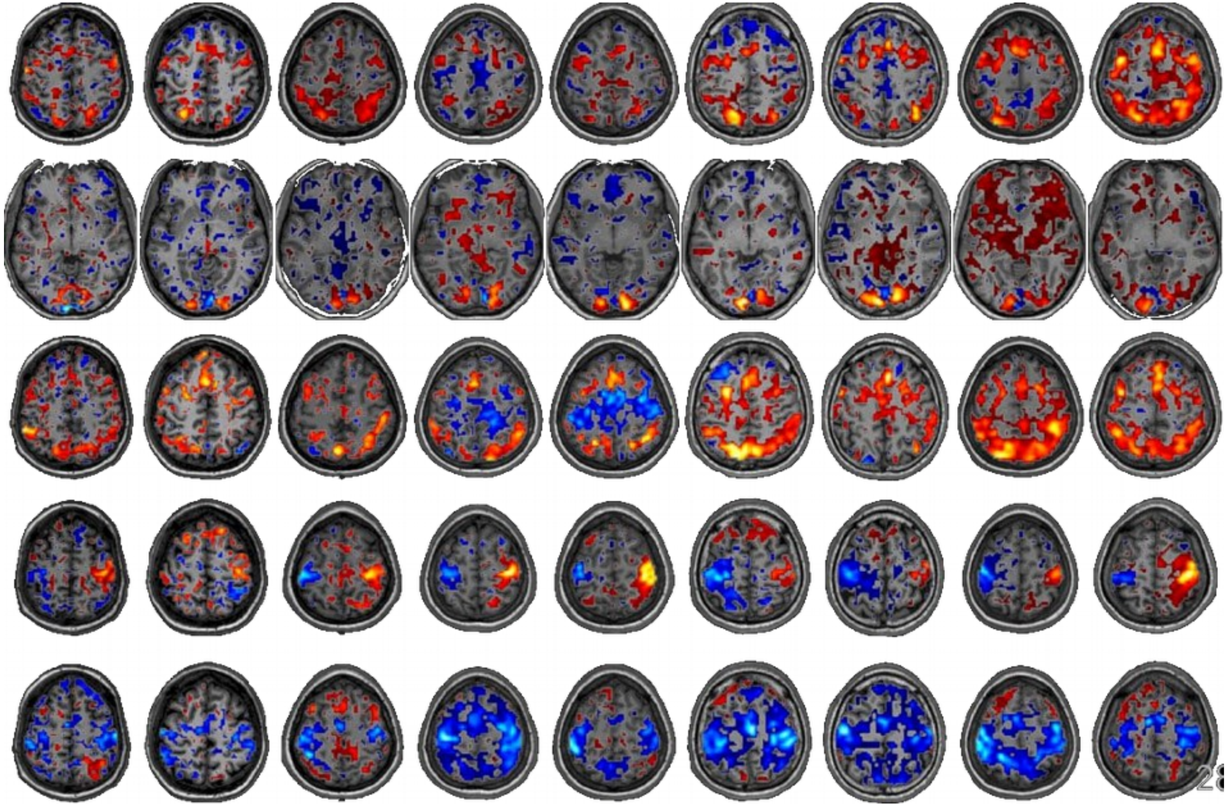
Principal Component Analysis



Sparse PCA


activation maps
voxels
X
=
cognitive loadings
D
voxels
A
argmin
D,A, Ω(D)≤1
$\|\mathbf{X} - \mathbf{D}\,\mathbf{A}^T\|_2^2 + \lambda \|\mathbf{A}\|_1$

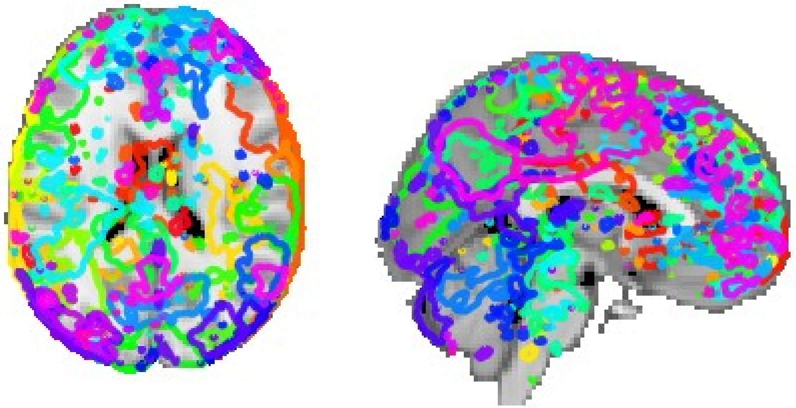
Sparse PCA



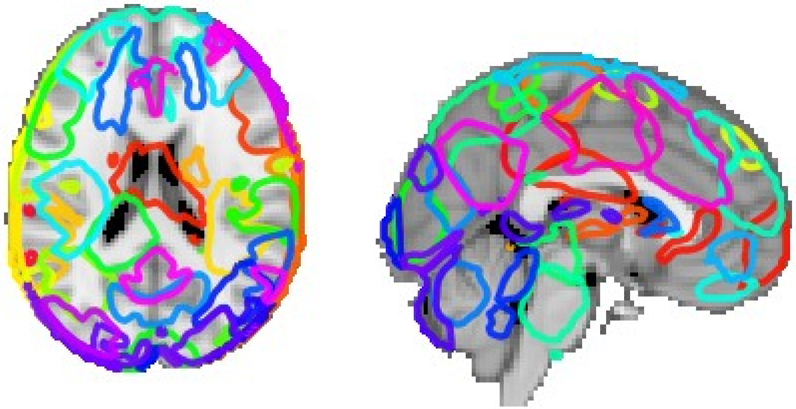
[Pinel et al. 2007]

Scaling to large datasets

50Gb data



1Tb data



Use more data to get better parcellation

[Mensch et al 2016]: Use stochastic updates to scale

Part 2: Take home messages

- Weakly-learning and unsupervised learning

Use data structure

- Reduce dimension of the data

Atlas

Learned Parcellation

Conclusion

- Use independent data to evaluate models
- Use a large number of samples
- Rely on the data structure

Conclusion

- Need more public data

OpenfMRI

NeuroVault

A public repository of unthresholded brain activation maps

- Need more open source Software

ni learn

scikit learn