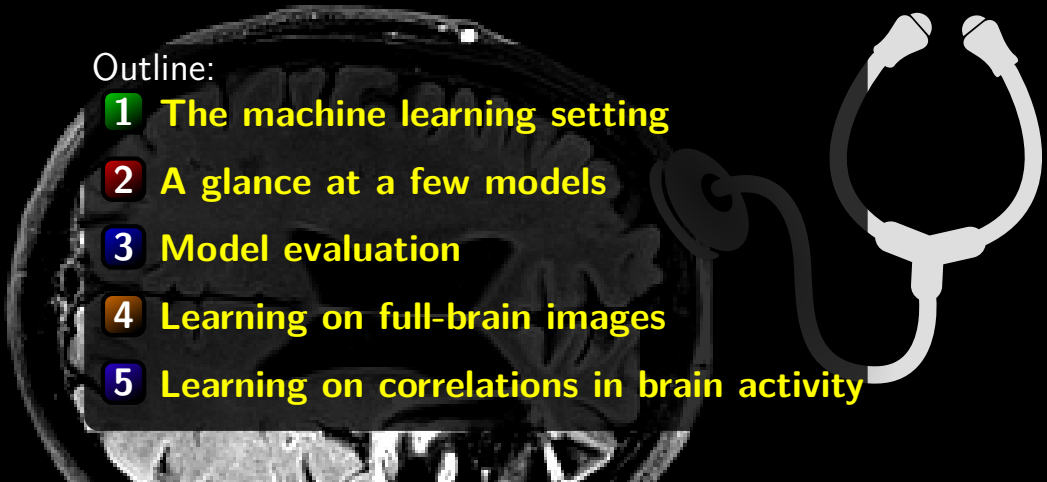
Introduction to Machine Learning in MR Imaging

Gaël Varoquaux

Introduction to Machine Learning in MR Imaging

Gaël Varoquaux



Outline:

1. The machine learning setting
2. A glance at a few models
3. Model evaluation
4. Learning on full-brain images
5. Learning on correlations in brain activity

JOINT ANNUAL MEETING
ISMRM–ESMRMB
16–21 June 2018

SMRT 27th Annual Meeting 15–18 June 2018
www.smrt.org

Paris Expo Porte de Versailles
Paris, France

Declaration of Financial Interests or Relationships

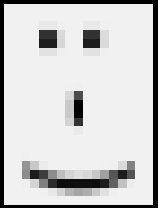
Speaker Name: Gaël Varoquaux

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

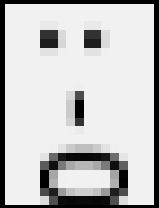
1 The machine learning setting

Adjusting models for prediction

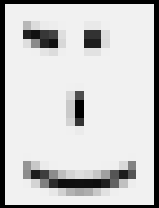
Face recognition



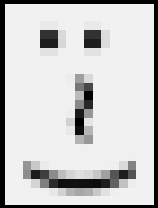
Andrew



Bill

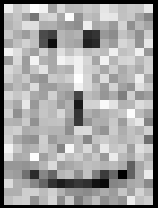


Charles

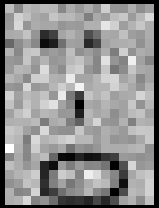


Dave

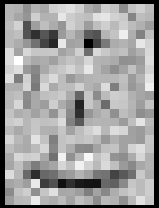
Face recognition



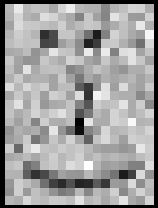
Andrew



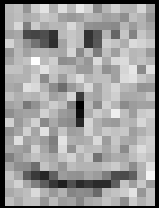
Bill



Charles



Dave



A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

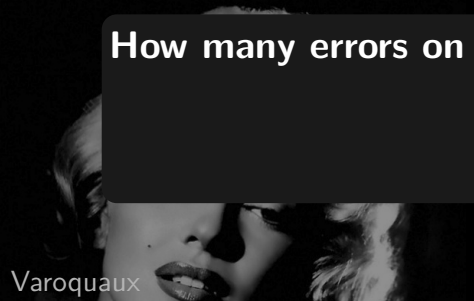
"Nearest neighbor" method

A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

"Nearest neighbor" method

How many errors on already-known images?



A simple method:

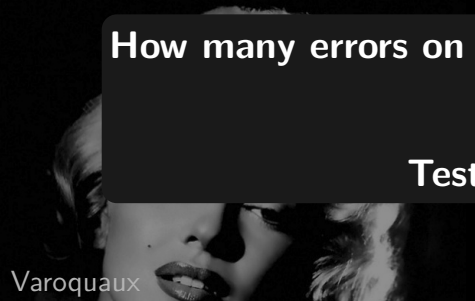
1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

"Nearest neighbor" method

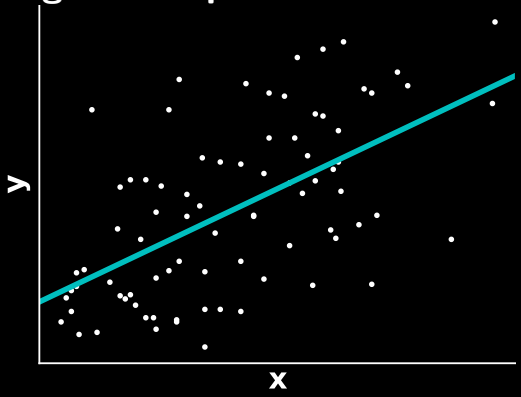
How many errors on already-known images?

... **0: no errors**

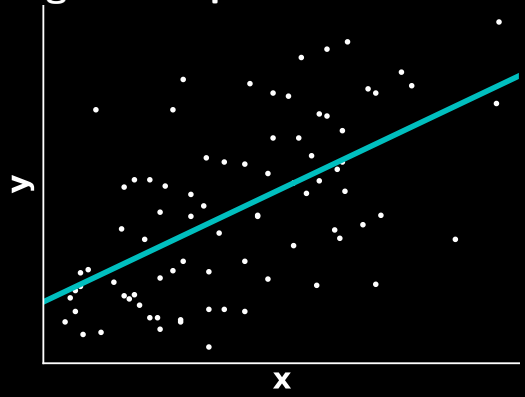
Test data $\neq$ Train data

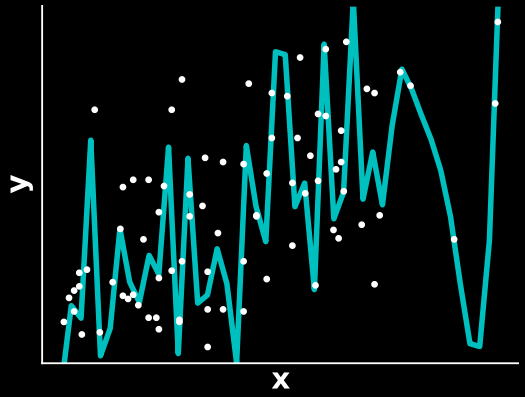


A single descriptor: 1 dimension



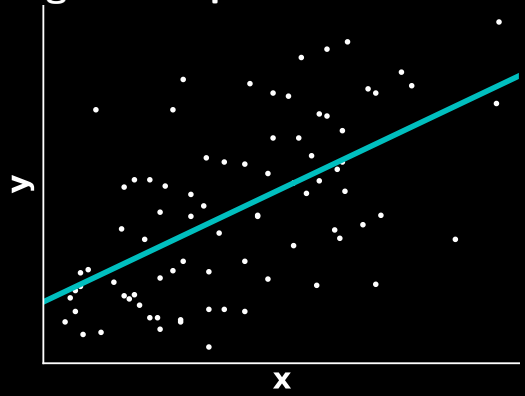
A single descriptor: 1 dimension

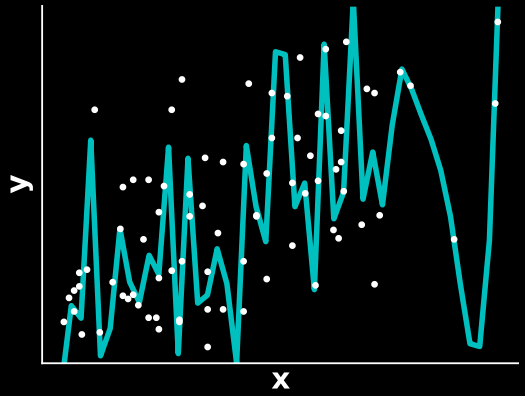




Which model to prefer?

A single descriptor: 1 dimension

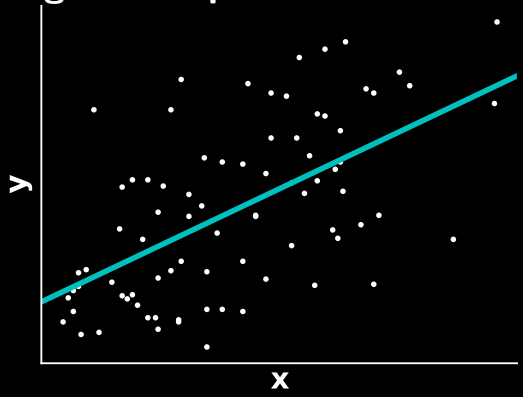


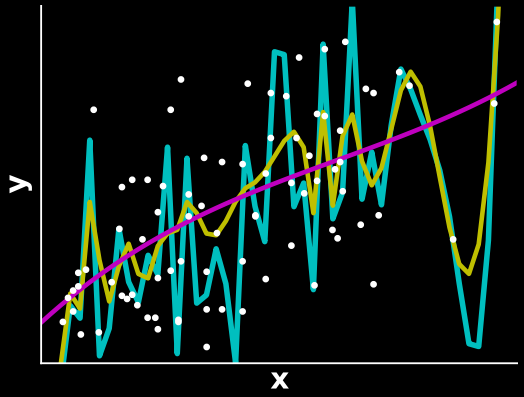


Problem of *"over-fitting"*

- Minimizing error is not always the best strategy (learning noise)
- Test data $\neq$ train data

A single descriptor: 1 dimension

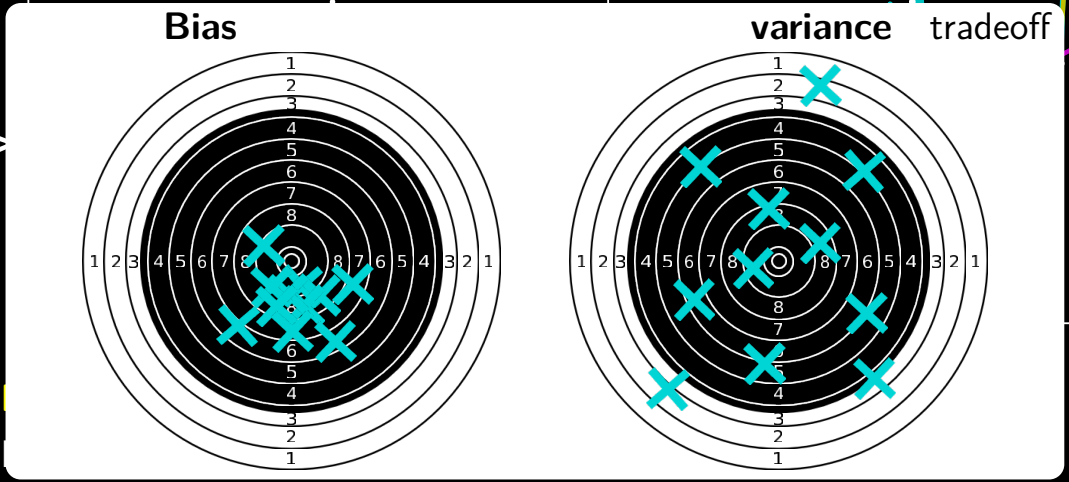




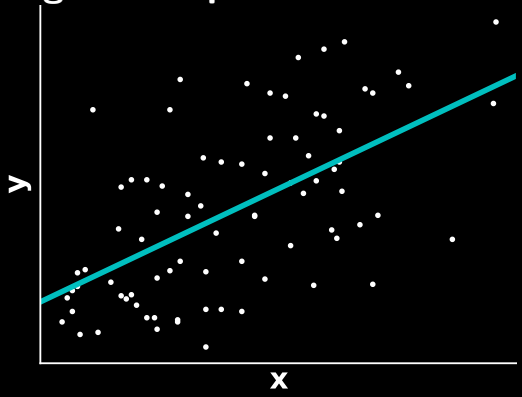
Prefer simple models = concept of *"regularization"*

Balance the number of parameters to learn with the amount of data

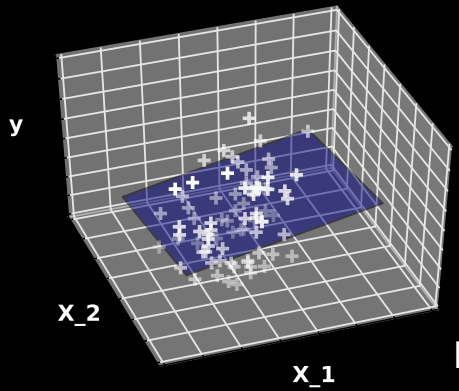
A single descriptor: 1 dimension



A single descriptor: 1 dimension

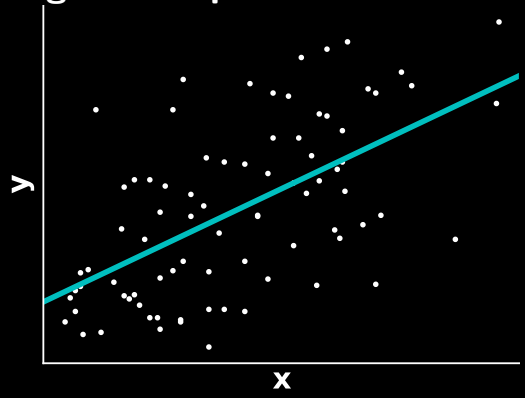


Two descriptors: 2 dimensions

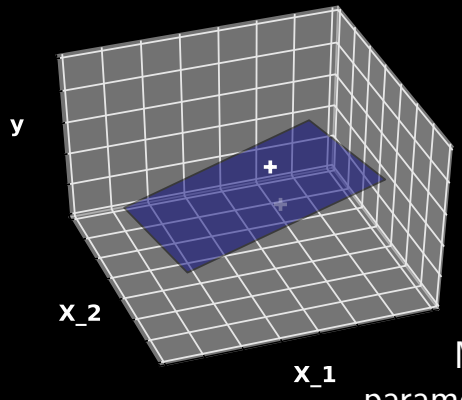


More parameters

A single descriptor: 1 dimension



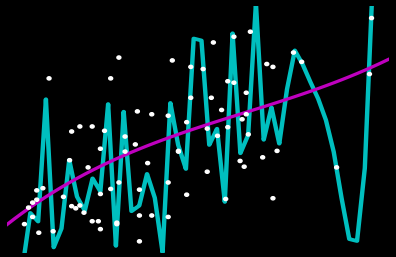
Two descriptors: 2 dimensions

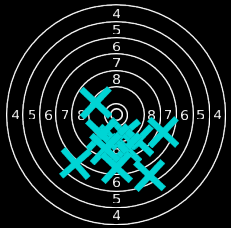


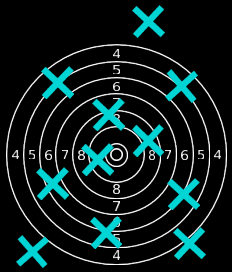
More parameters

⇒ Model with more parameters need much more data

"curse of dimensionality"

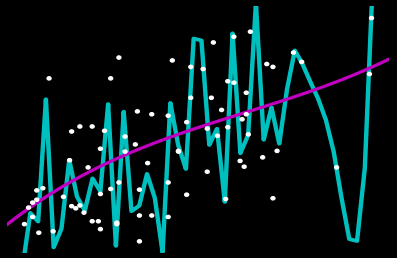


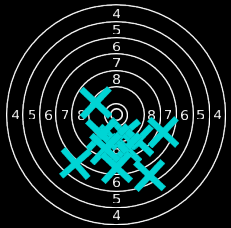


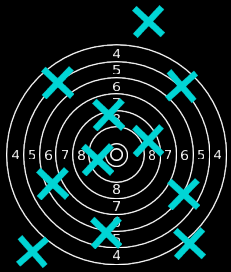


Settings: data $(\mathbf{X}, \mathbf{y})$, prediction $\mathbf{y} \sim f(\mathbf{X}, \mathbf{w})$

Our goal: $\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|$







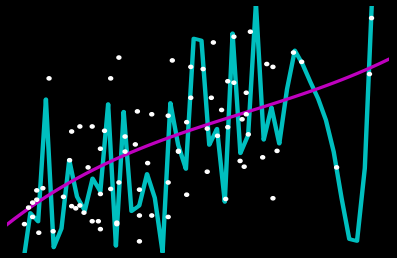
Settings: data $(\mathbf{X}, \mathbf{y})$, prediction $\mathbf{y} \sim f(\mathbf{X}, \mathbf{w})$

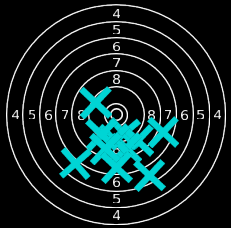
Our goal: $\underset{\mathbf{w}}{\text{minimize}}\ \mathbb{E}[\|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|\ ]$

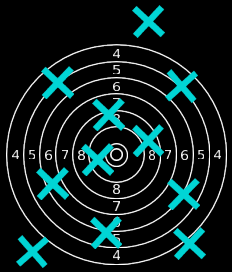
We only can measure $\|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|$

Prediction is very difficult, especially about the future.

Niels Bohr







Settings: data $(\mathbf{X}, \mathbf{y})$, prediction $\mathbf{y} \sim f(\mathbf{X}, \mathbf{w})$

Our goal: $\underset{\mathbf{w}}{\text{minimize}}\ \mathbb{E}[\|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|\ ]$

We only can measure $\|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\|$

Solution: bias $\mathbf{w}$ to push toward a plausible solution

In a minimization framework: $\underset{\mathbf{w}}{\text{minimize}}\ \|\mathbf{y} - f(\mathbf{X}, \mathbf{w})\| + p(\mathbf{w})$

$$\mathcal{P}(\mathbf{w}|\mathbf{x},\mathbf{y}) \propto \mathcal{P}(\mathbf{y},\mathbf{x}|\mathbf{w})\ \mathcal{P}(\mathbf{w}) \qquad (\star)$$

"Posterior"
Quantity of interest

Forward model

"Prior"
Expectations on $\mathbf{w}$

- Forward model: $\mathbf{y} = f(\mathbf{x},\mathbf{w}) + \mathbf{e}$, $\mathbf{e}$: noise
  $\Rightarrow \mathcal{P}(\mathbf{x},\mathbf{y}|\mathbf{w}) \propto \exp -\mathcal{L}(\mathbf{y} - f(\mathbf{x},\mathbf{w}))$
- Prior: $\mathcal{P}(\mathbf{w}) \propto \exp -p(\mathbf{w})$

Negated log of $(\star)$: $\mathcal{L}(\mathbf{y} - f(\mathbf{X},\mathbf{w})) + p(\mathbf{w})$

Minimization framework = maximum a posteriori

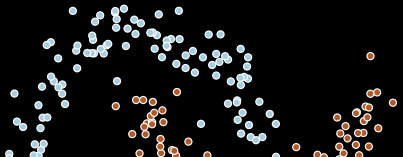
Note that the Bayesian interpretation may be more subtle [Gribonval 2011].

1 **Summary**: elements of a machine-learning method

- **A forward model**: $\mathbf{y}_{\text{pred}} = f(\mathbf{X}, \mathbf{w})$
  Numerical rules to go from $\mathbf{X}$ to $\mathbf{y}$

- **A loss**, or data fit
  A measure of error between $\mathbf{y}_{\text{true}}$ and $\mathbf{y}_{\text{pred}}$
  Can be given by a noise model

- **Regularization**:
  Any way of restricting model complexity
  - by choices in the model
  - via a penalty



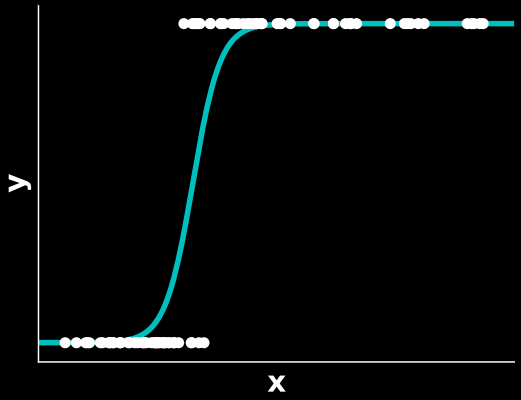
2 A glance at a few models

- Linear models
- Random forests
- Gradient boosted trees

y describes categories, say $(0, 1)$

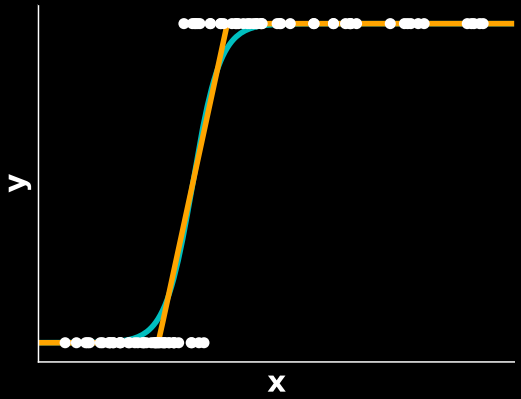
- Fit **y** by a straight line?

y describes categories, say $(0, 1)$



- Fit **y** by a straight line?
- A sigmoid is better suited
  $\Rightarrow$ Logistic regression

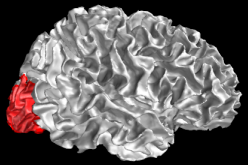
y describes categories, say $(0, 1)$

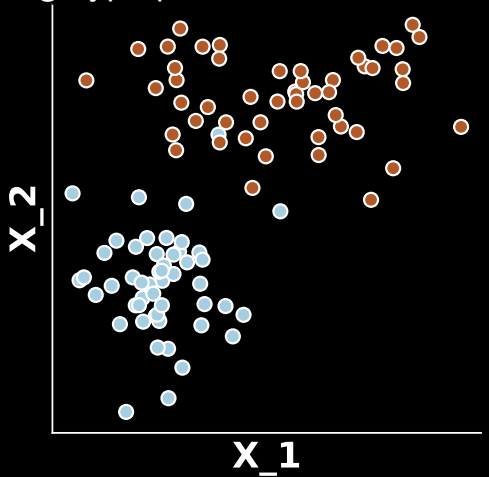


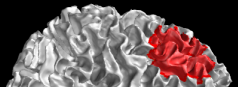
- Fit **y** by a straight line?
- A sigmoid is better suited
  $\Rightarrow$ Logistic regression
- SVM (Support Vector Machines) are similar

Note that points far from threshold don't influence the fit

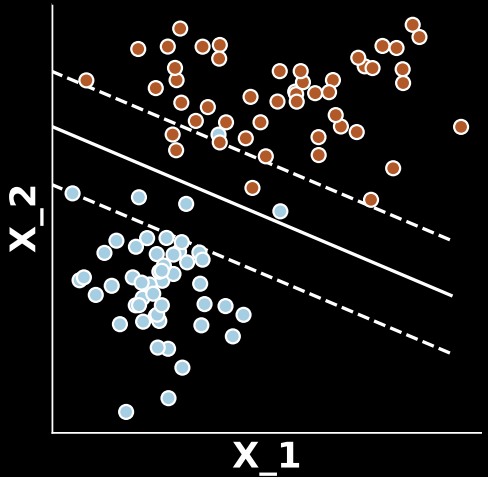
Choosing a separating hyperplane





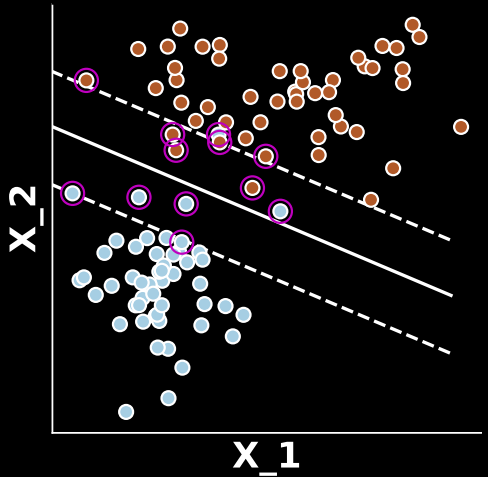


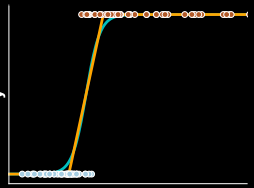
SVM: support vector machine



Discriminative method: choose hyperplane to maximize **margin**

SVM: support vector machine

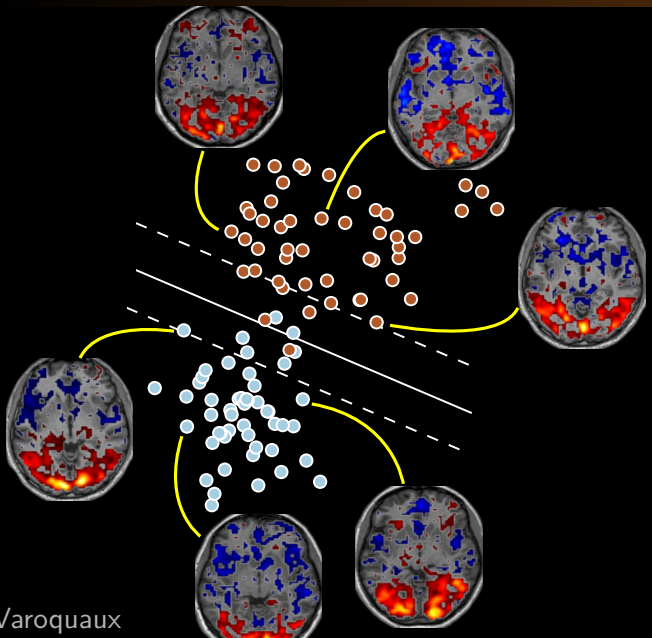




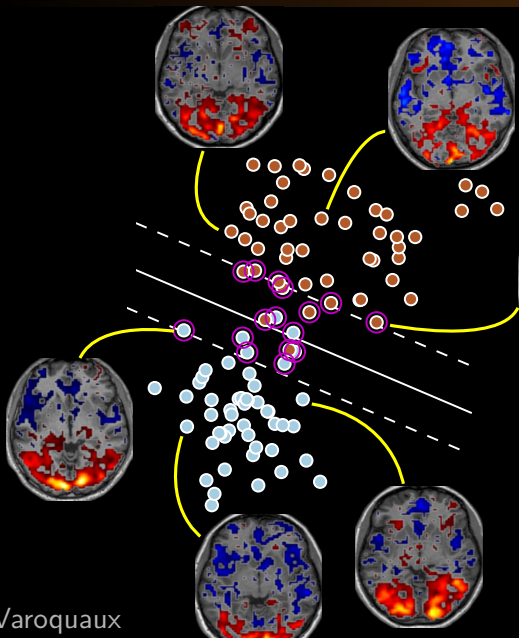
Discriminative method: choose hyperplane to maximize **margin**

use exemplars close to hyperplane (**support vectors**)

2 With brain images



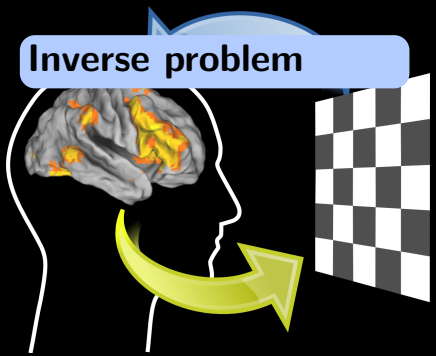
Find a separating hyperplane between images



Find a separating hyperplane between images

SVM builds it by combining exemplars

$\Rightarrow$ hyperplane looks like train images



- Minimize the error term:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w})$$

Ill-posed:

Many different $\mathbf{w}$ will give the same prediction error

- To choose one: inject prior with a penalty

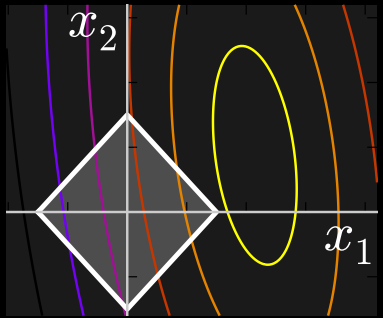
$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w}) + p(\mathbf{w})$$

- Lasso estimator

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \; + \; \lambda \, \ell_1(\mathbf{x})$$

Data fit

Penalization

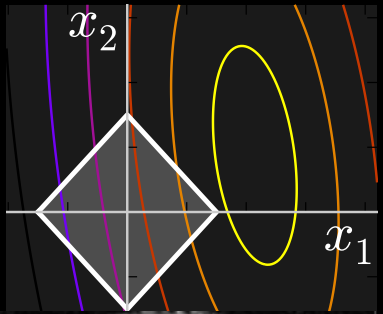


- Lasso estimator

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \ell_1(\mathbf{x})$$

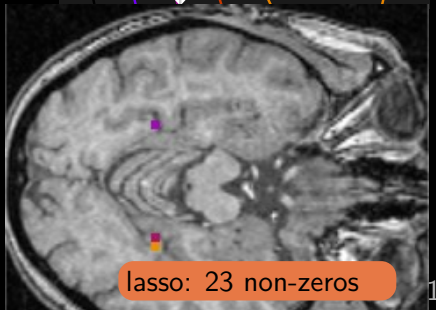
Data fit

Penalization

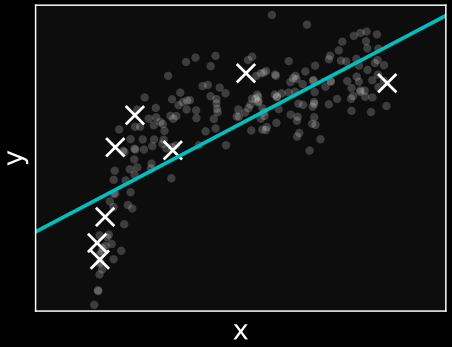


- Between correlated features, selects a random subset
  [Wainwright 2009, Varoquaux... 2012]

  Violates the restricted isometry property

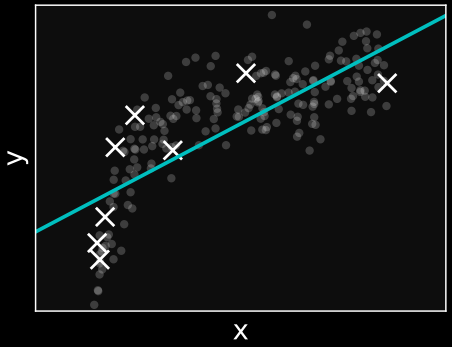


Polynome degree 1

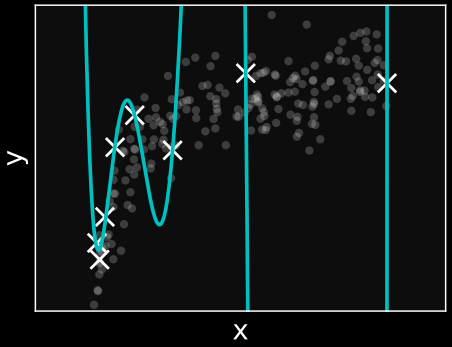


2 Underfit versus overfit

Polynome degree 1

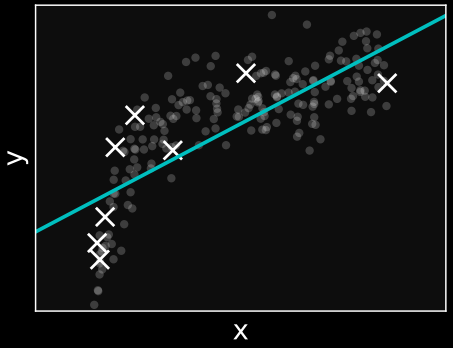


Polynome degree 9



2 Underfit versus overfit

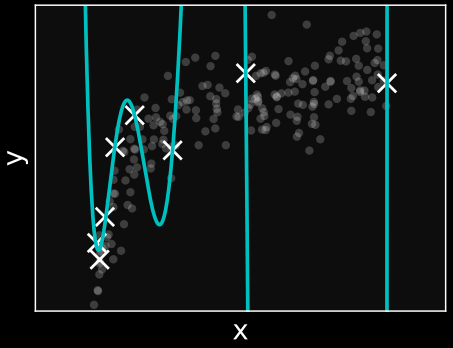
Polynome degree 1



Train explained variance: 0.56
Test explained variance: 0.56

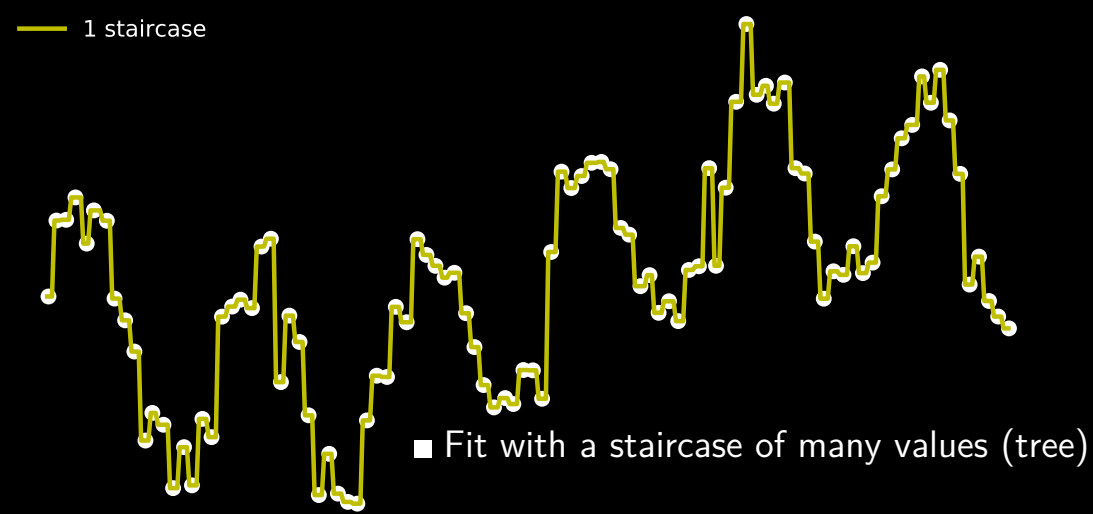
Underfit

Polynome degree 9

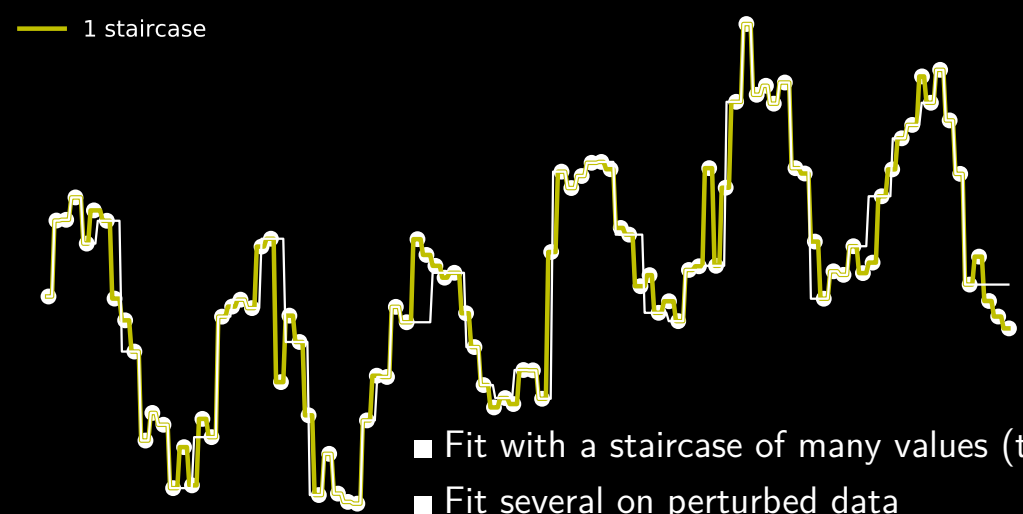


Train explained variance: 0.99
Test explained variance: -98000

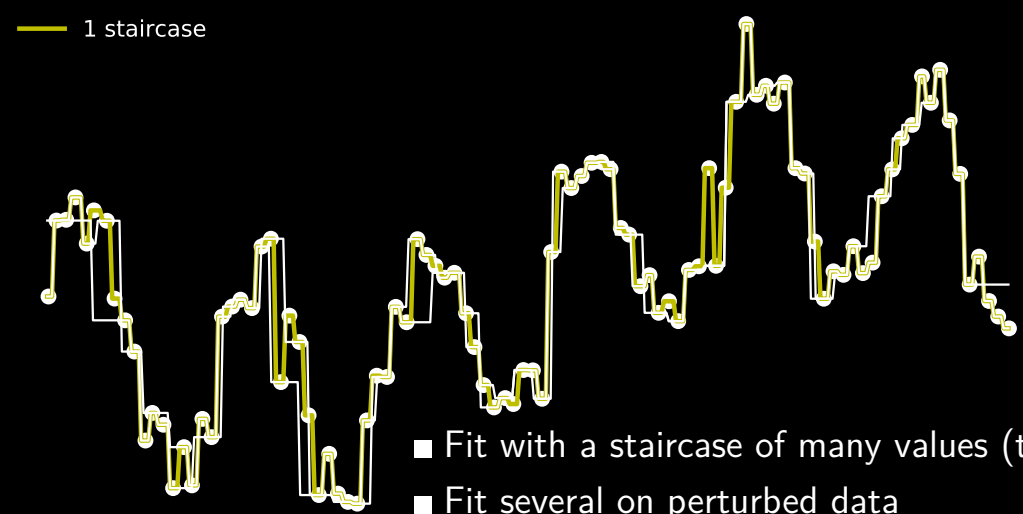
Overfit



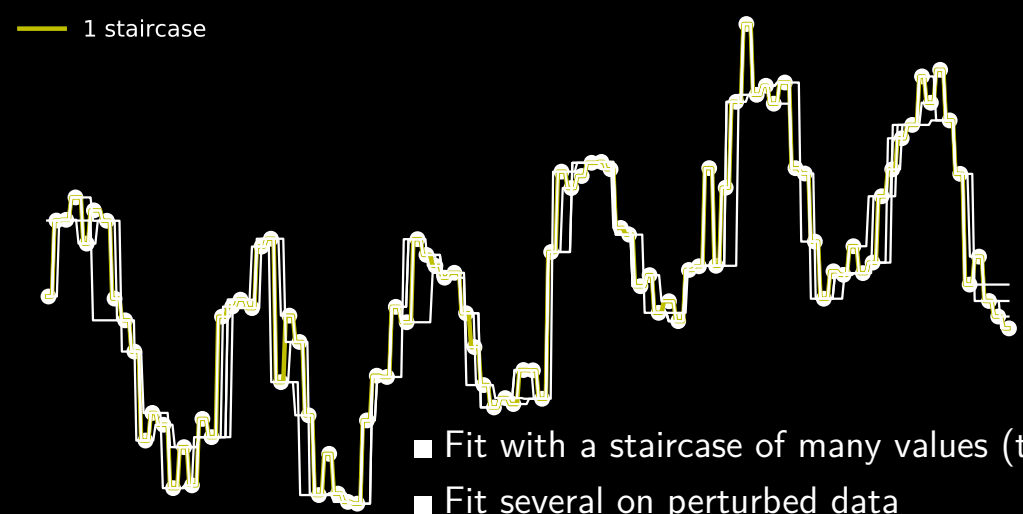
- Fit with a staircase of many values (tree)



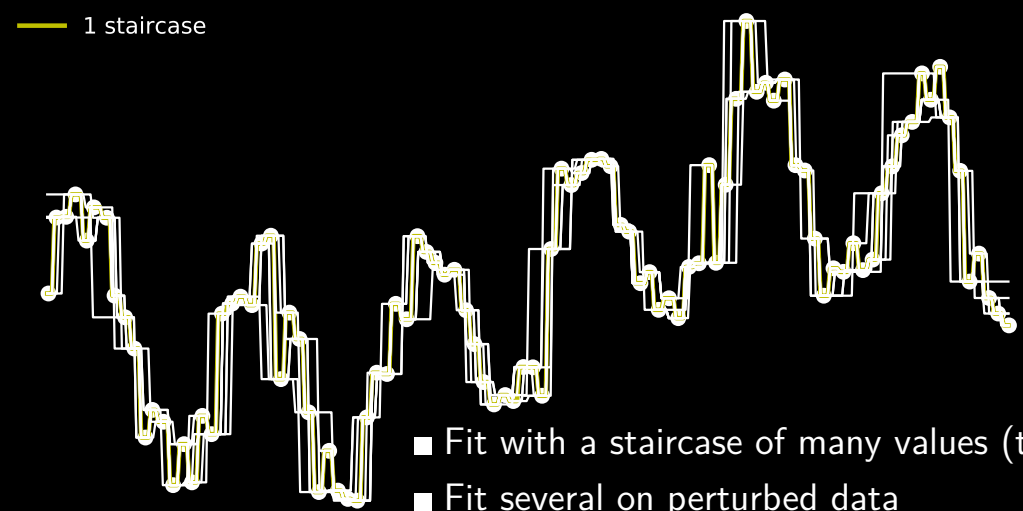
- Fit with a staircase of many values (tree)
- Fit several on perturbed data



- Fit with a staircase of many values (tree)
- Fit several on perturbed data

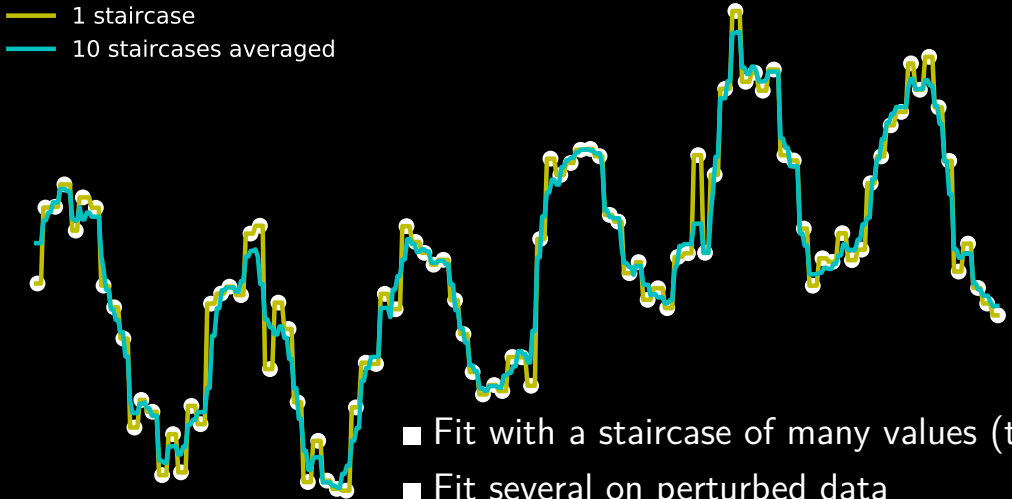


- Fit with a staircase of many values (tree)
- Fit several on perturbed data



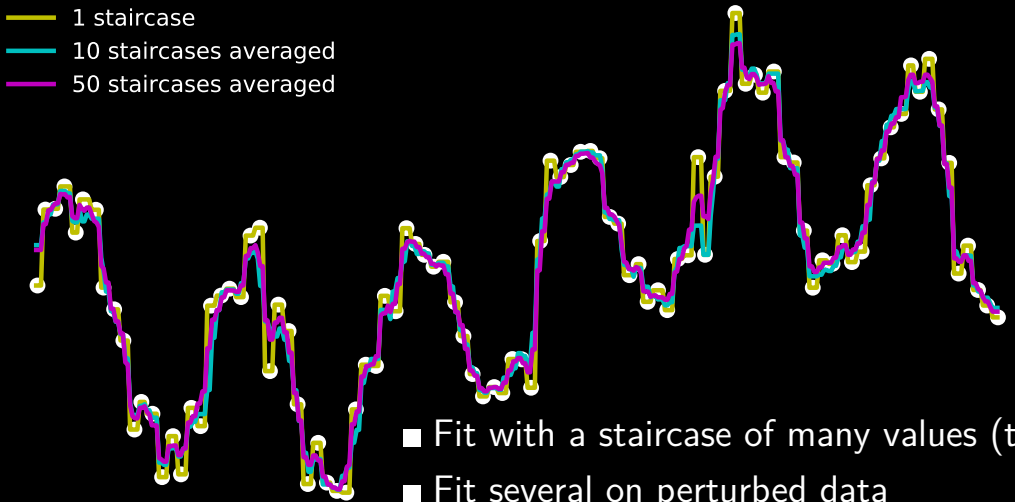
- Fit with a staircase of many values (tree)
- Fit several on perturbed data

2 Random forest



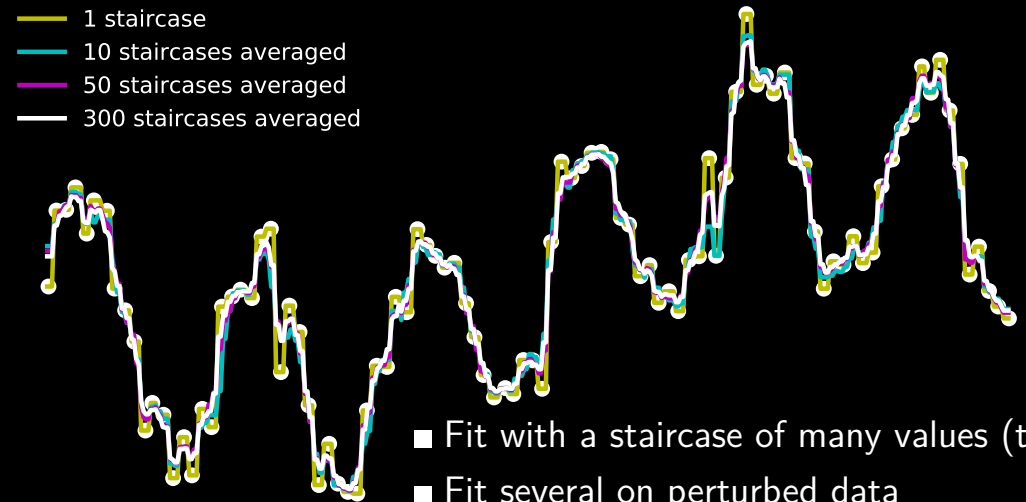
- Fit with a staircase of many values (tree)
- Fit several on perturbed data
- Average prediction

2 Random forest

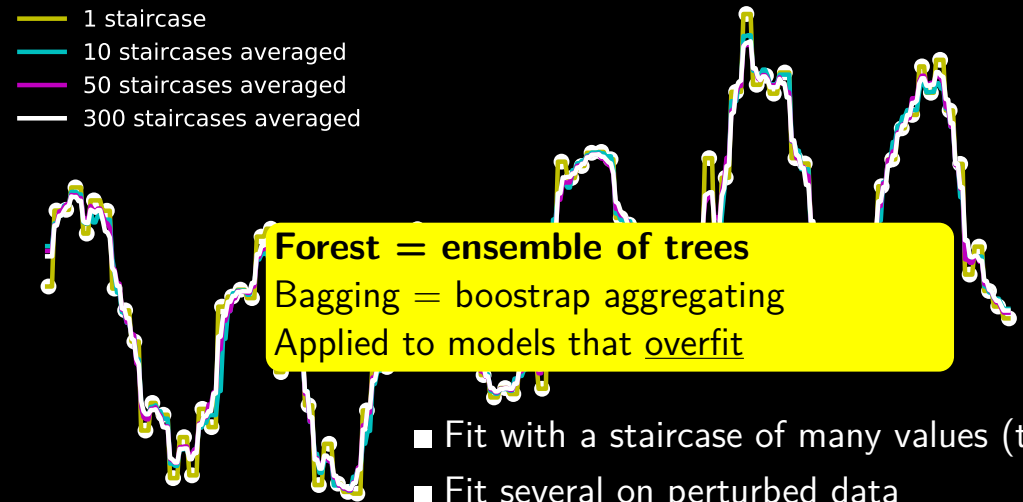


- Fit with a staircase of many values (tree)
- Fit several on perturbed data
- Average prediction

2 Random forest



- Fit with a staircase of many values (tree)
- Fit several on perturbed data
- Average prediction

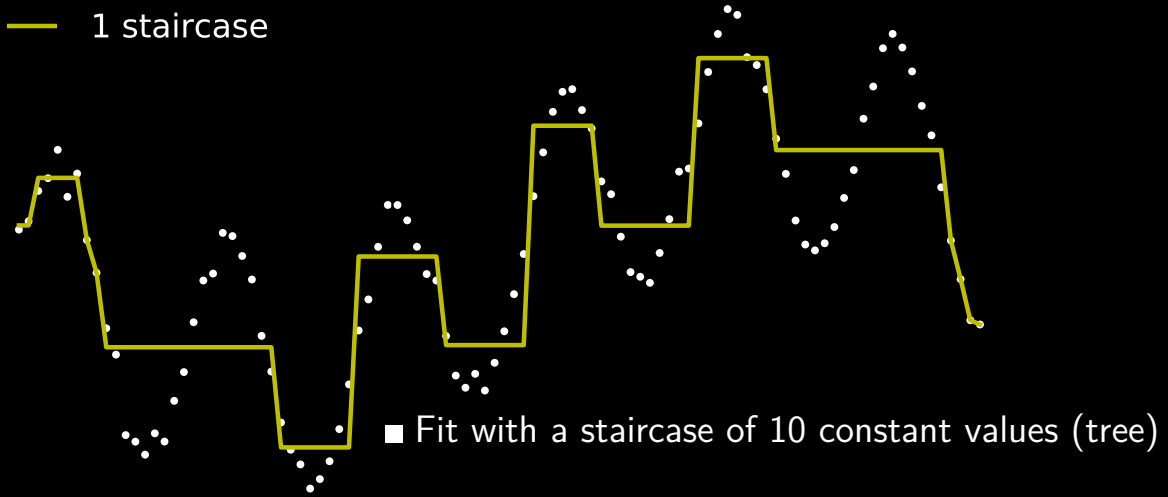


Forest = ensemble of trees

Bagging = boostrap aggregating

Applied to models that overfit

- Fit with a staircase of many values (tree)
- Fit several on perturbed data
- Average prediction

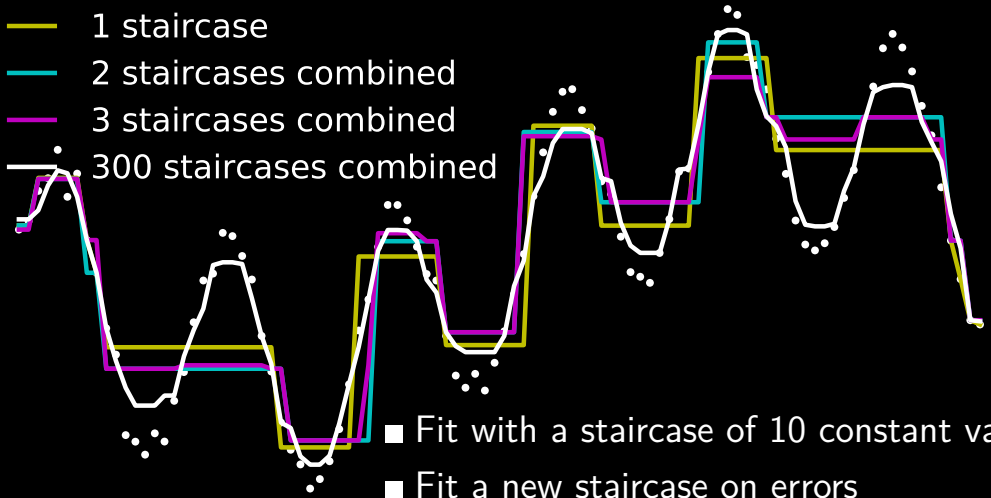

1 staircase
Fit with a staircase of 10 constant values (tree)

2 Gradient boosted trees

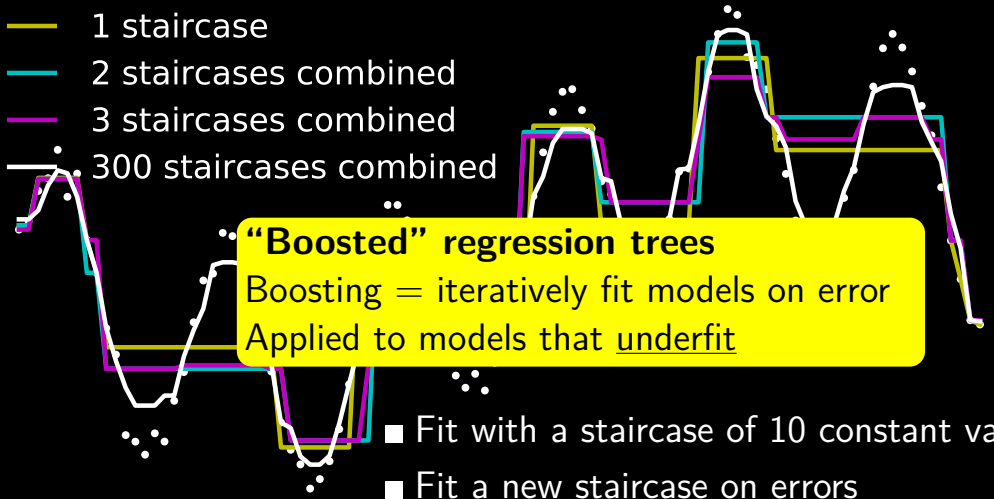
- Fit with a staircase of 10 constant values (tree)
- Fit a new staircase on errors

- Fit with a staircase of 10 constant values (tree)
- Fit a new staircase on errors
- Keep going

2 Gradient boosted trees



- Fit with a staircase of 10 constant values (tree)
- Fit a new staircase on errors
- Keep going

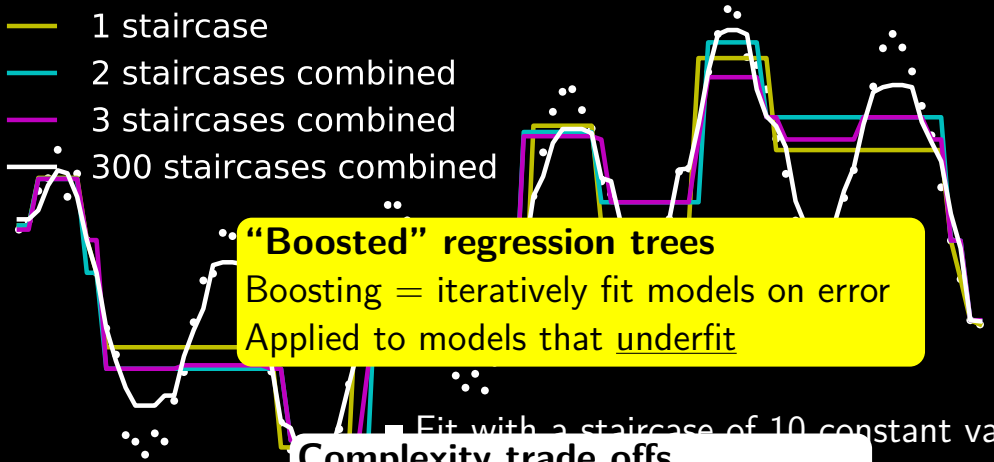


"Boosted" regression trees

Boosting = iteratively fit models on error

Applied to models that underfit

- Fit with a staircase of 10 constant values (tree)
- Fit a new staircase on errors
- Keep going



"Boosted" regression trees

Boosting = iteratively fit models on error

Applied to models that underfit

- Fit with a staircase of 10 constant values (tree)

Complexity trade offs

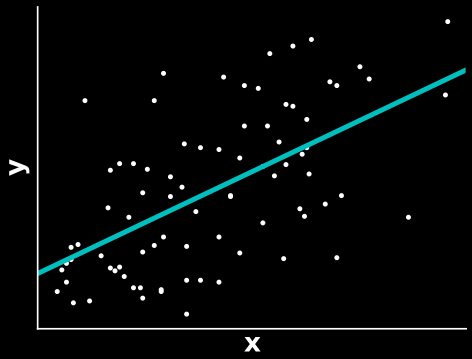
Computational + statistical

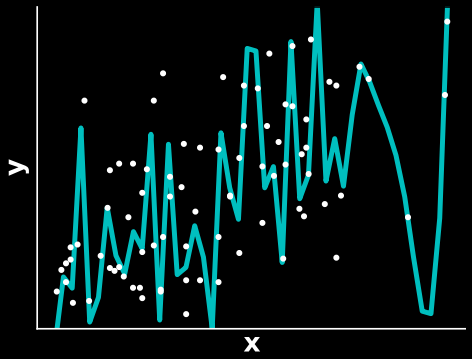
- Keep going

3 Model evaluation

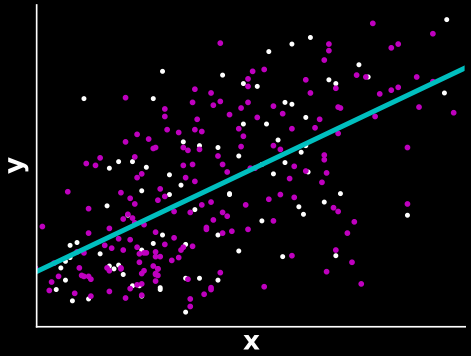
Does it predict?

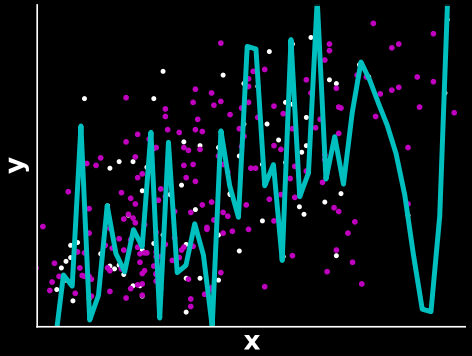
Generalization as a test: cross-validation





3 Generalization as a test: cross-validation



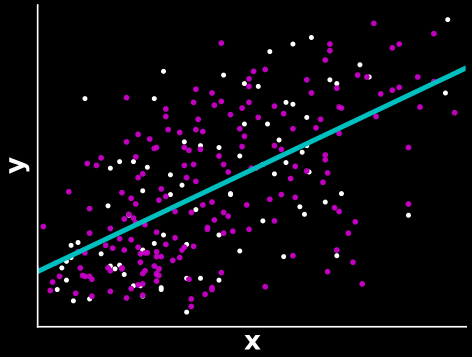


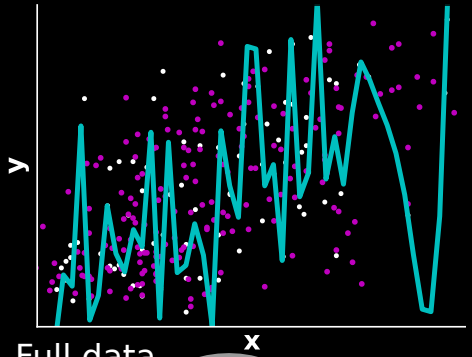
⇒ Need test on **independent** data, to control for model complexity

Measures prediction accuracy

[Varoquaux... 2017]

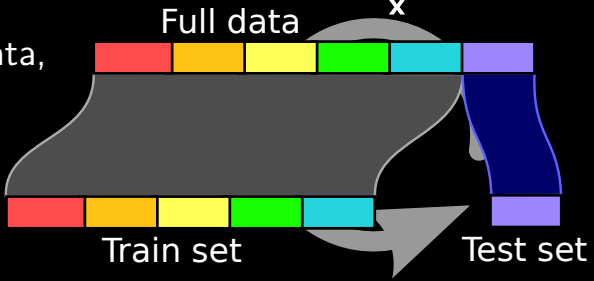
3 Generalization as a test: cross-validation



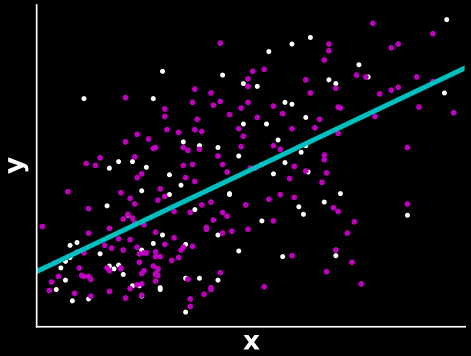


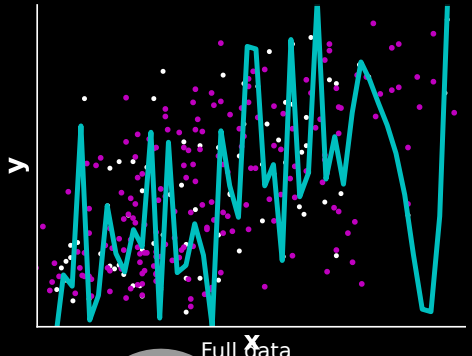
⇒ Need test on **independent** data,

Loop



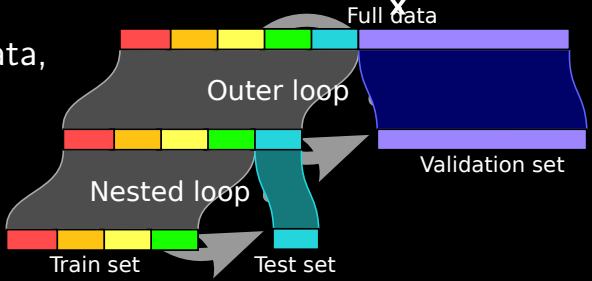
3 Generalization as a test: cross-validation



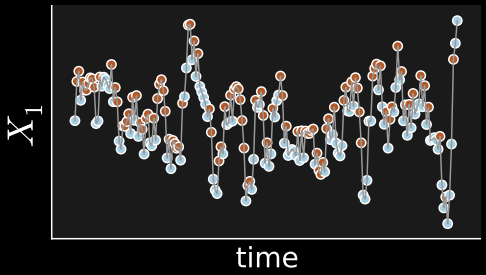


$\Rightarrow$ Need test on **independent** data,

Nested loop

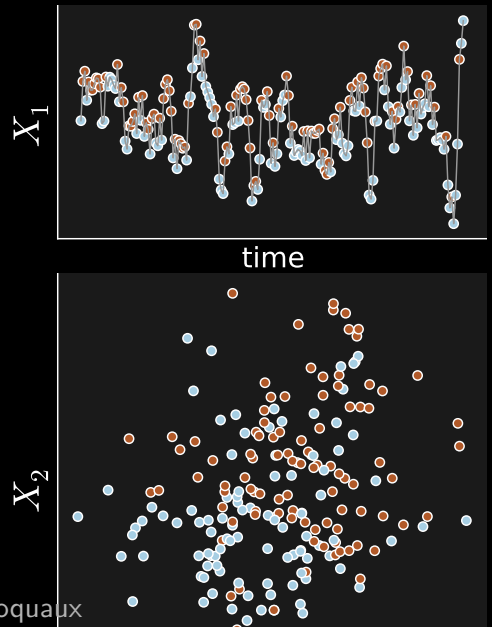


3 Confounds & dependencies



- Prediction is easy
  if you look at the last time point

3 Confounds & dependencies

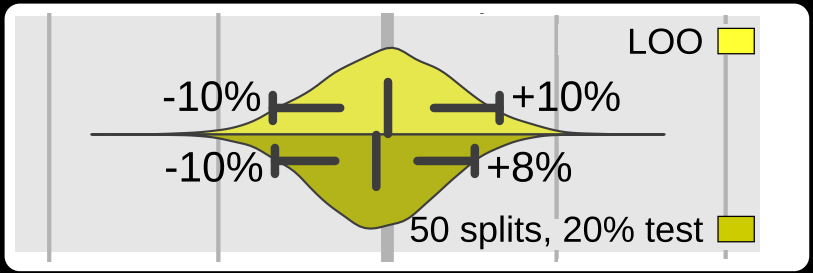


- Prediction is easy
  if you look at the last time point
- Auto-correlated noise
  Neighboring data points
  not independent
  No evidence of generalization
  No evidence of useful prediction

Validate on independent data

Not 2 images of the same subject
Unless it's a longitudinal study

[Varoquaux... 2017, Little... 2017]



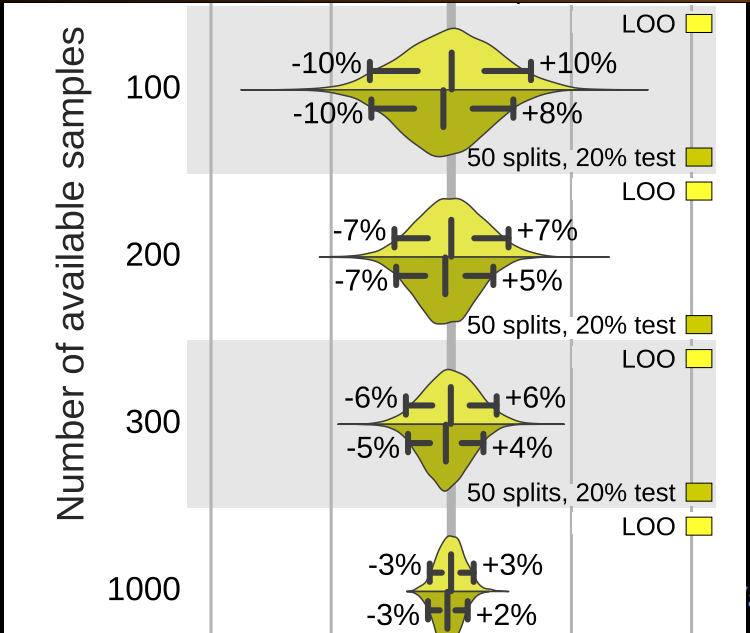
Distribution of prediction accuracies across 50 folds

- Should be seen as a posterior distribution (or bootstrap)
- $\neq$ folds are **not independent** measures

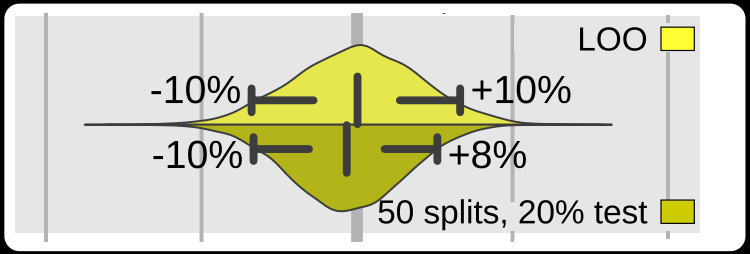
No statistical tests (eg T tests) on folds

[Varoquaux 2017]

3 Uncertainty of cross-validation



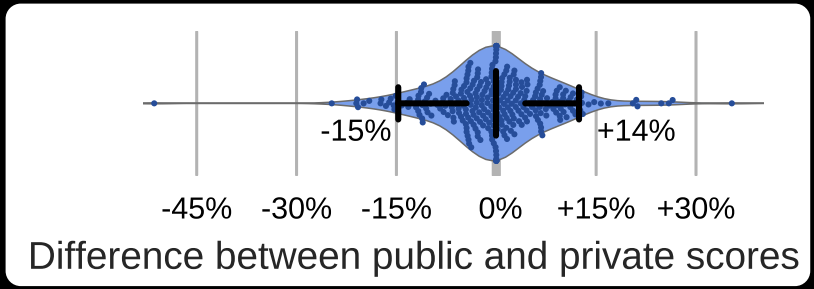
[aroquaux 2017]



Sample size	Confidence interval (binomial model)
100	18.0%
1 000	5.7%
10 000	1.8%
100 000	0.568%

Sampling noise – Decreases slowly with n [Varoquaux 2017]

3 Uncertainty of cross-validation



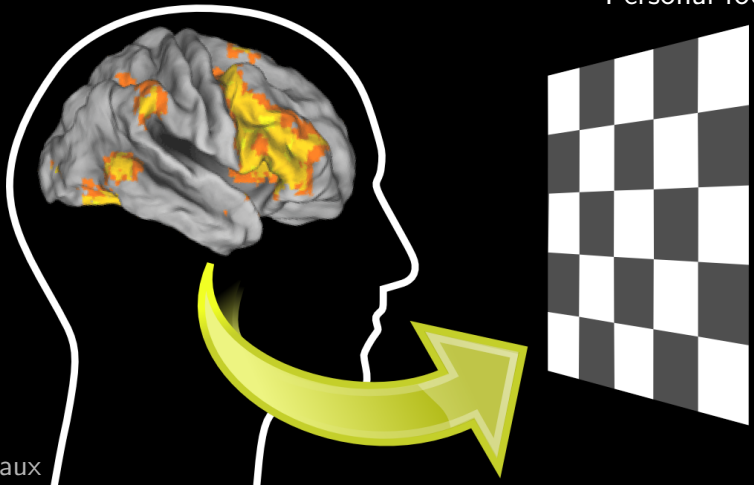
Kaggle competition on r-fMRI for Schizophrenia
2 different test sets: size 30 and 28

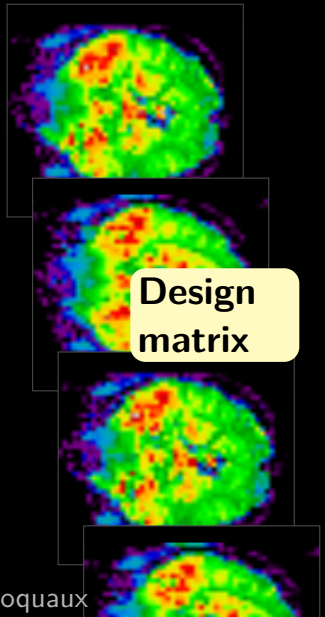
Winner of competition: first code written

[Varoquaux 2017]

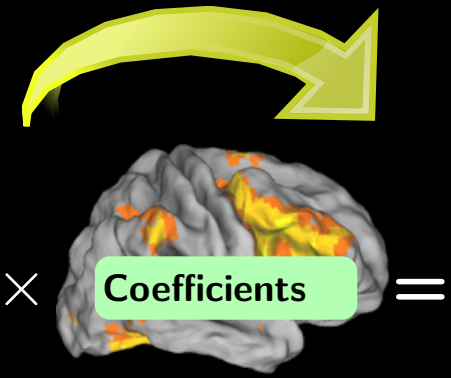
4 Learning on full-brain images

Personal focus on functional imaging





×



=

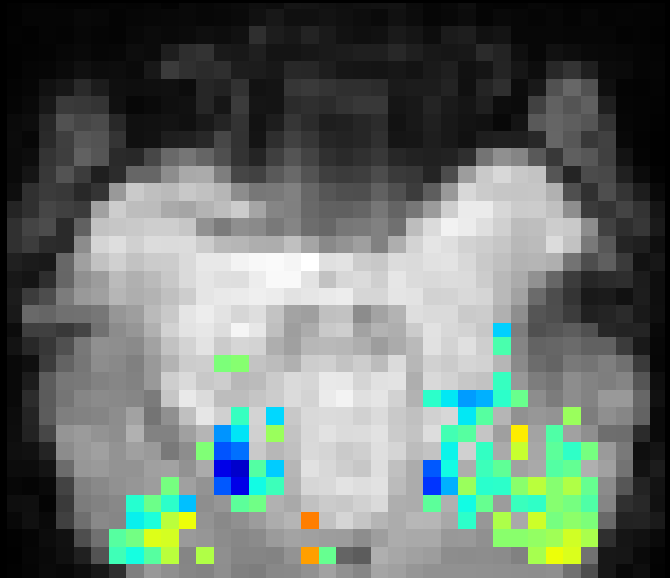
Coefficients are

brain maps

Face vs house visual recognition [Haxby... 2001]

SVM

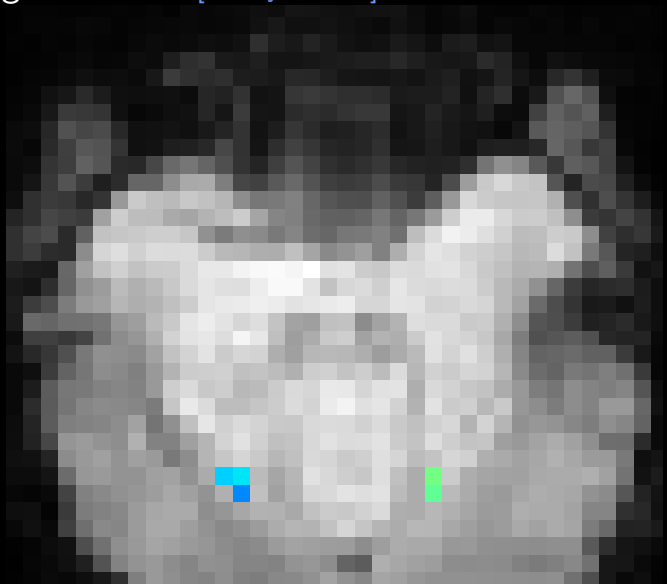
error: 26%



Face vs house visual recognition [Haxby... 2001]

Sparse model

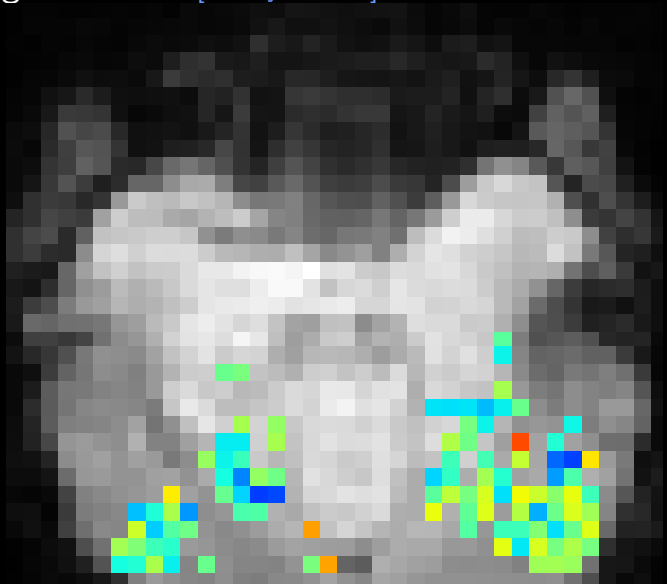
error: 19%

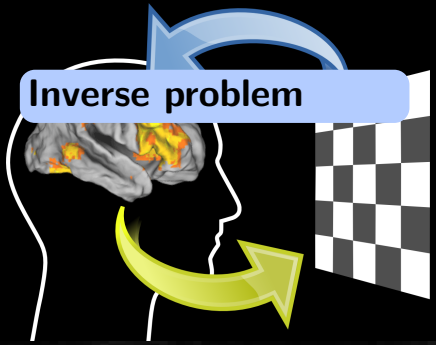


Face vs house visual recognition [Haxby... 2001]

Ridge

error: 15%





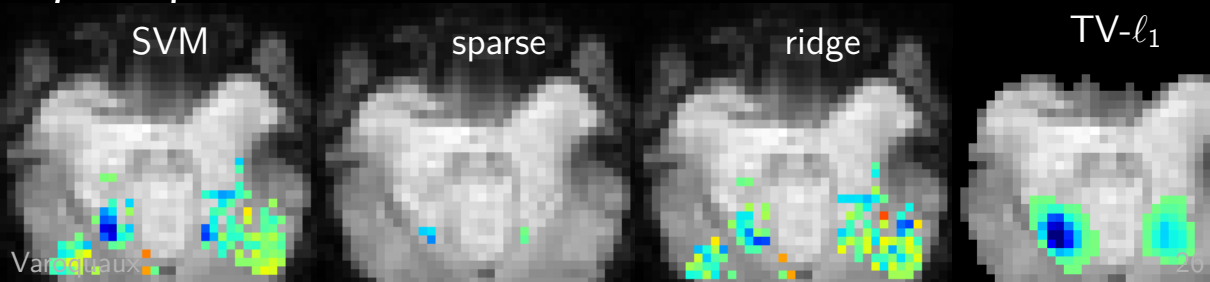
- Minimize the error term:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \, l(\mathbf{y} - \mathbf{X}\,\mathbf{w})$$

Ill-posed:

Many different $\mathbf{w}$ will give the same prediction error

Choice driven by (implicit) priors



[Gramfort... 2013]

- Minimize the error term:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w})$$

Ill-posed:

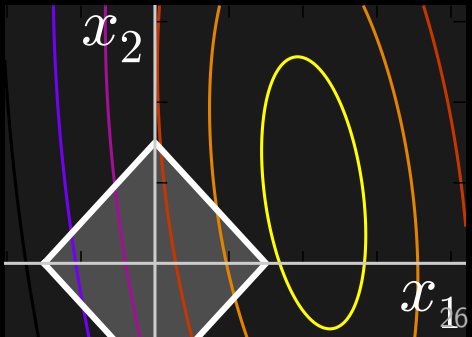
Many different $\mathbf{w}$ will give the same prediction error

Sparse estimators (Lasso)

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\, \ell_1(\mathbf{x})$$

Data fit

Penalization

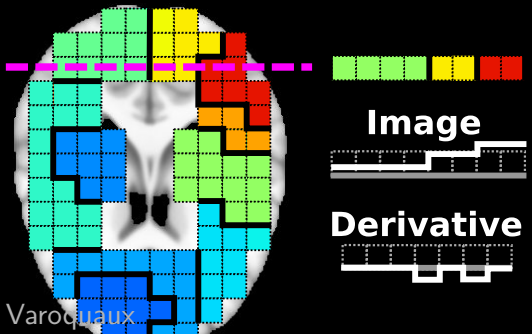


- Minimize the error term:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w})$$

Ill-posed:

Many different $\mathbf{w}$ will give the same prediction error

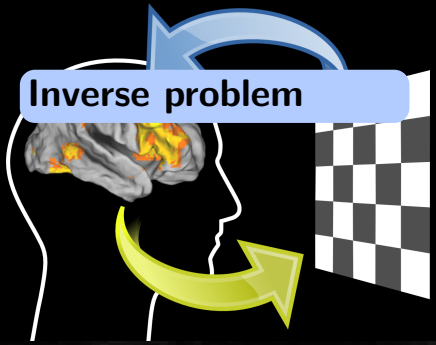


Total-variation penalization

Impose sparsity on the gradient of the image:

$$p(\mathbf{w}) = \ell_1(\nabla \mathbf{w})$$

In fMRI: [Michel... 2011]



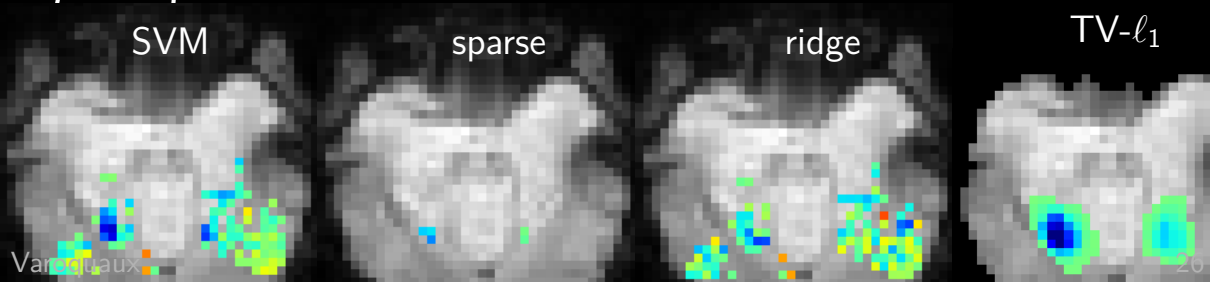
- Minimize the error term:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \, l(\mathbf{y} - \mathbf{X}\,\mathbf{w})$$

Ill-posed:

Many different $\mathbf{w}$ will give the same prediction error

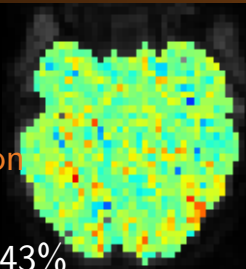
Choice driven by (implicit) priors



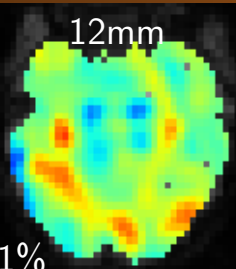
Spatial smoothing

Giving up on resolution

error rate: 43%



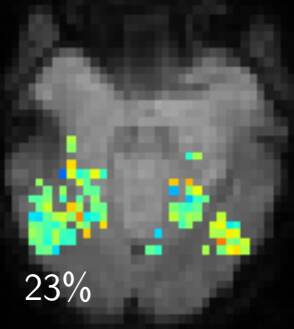
12mm



31%

Feature selection

Giving up on multivariate

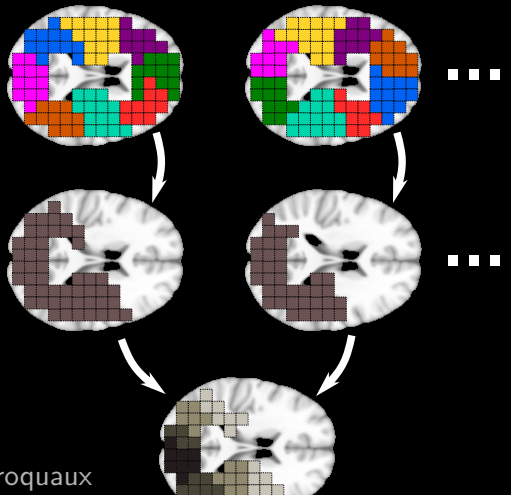


23%

4 **FREM**: Ensemble models [Hoyos-Idrobo... 2017]

- Very fast sub optimal models
- Average many of them

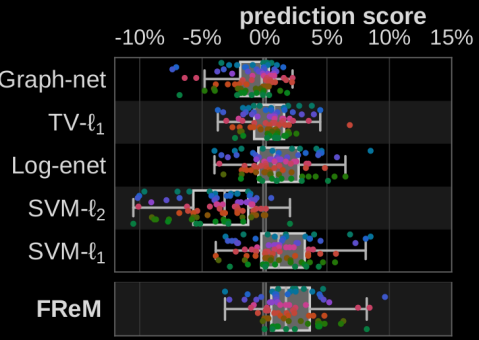
- Very fast sub optimal models
- Average many of them



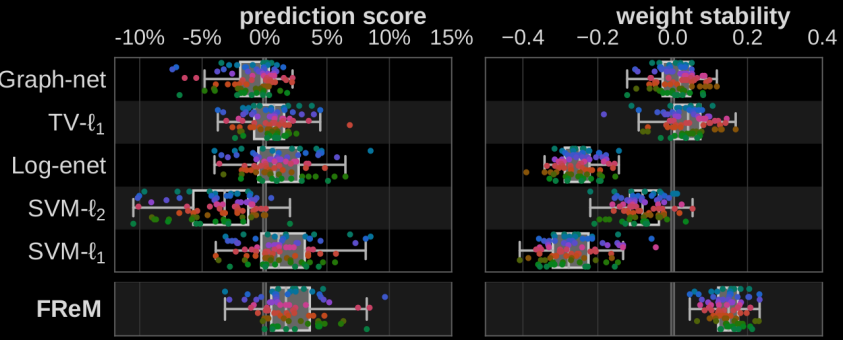
- Learn parcellation on perturbed data
- Estimate linear models

Average the results

- Very fast sub optimal models
- Average many of them



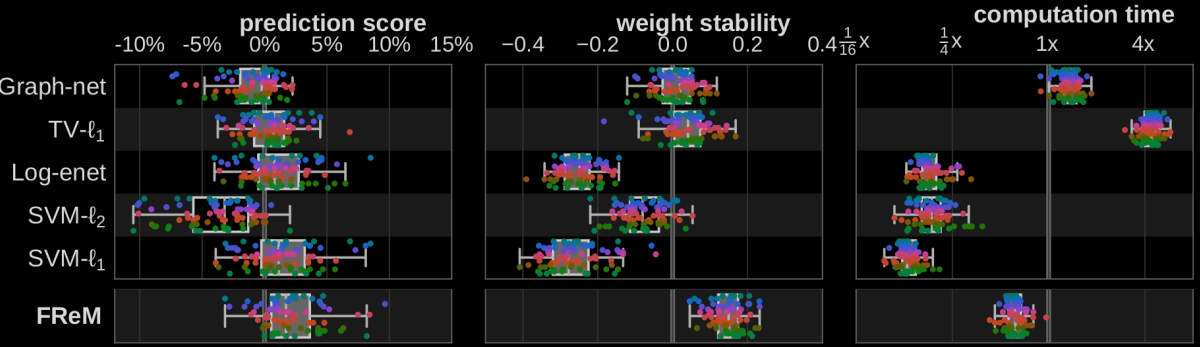
- Very fast sub optimal models
- Average many of them



4 FREM: Ensemble models

[Hoyos-Idrobo... 2017]

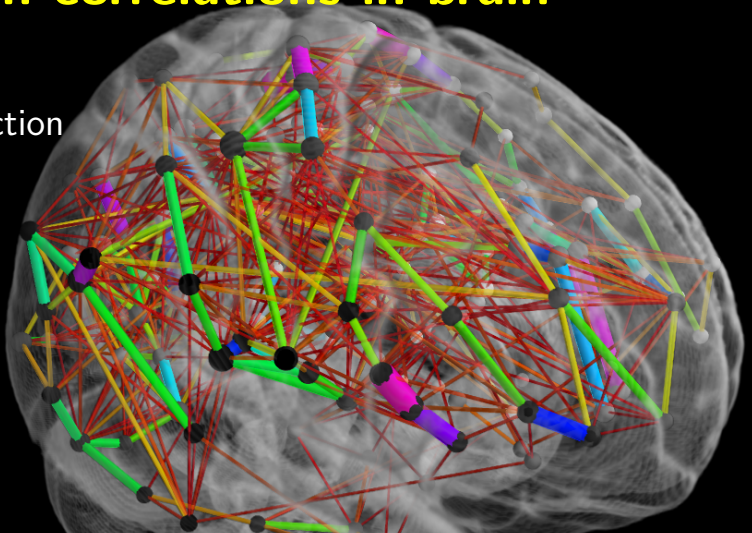
- Very fast sub optimal models
- Average many of them



5 Learning on correlations in brain activity

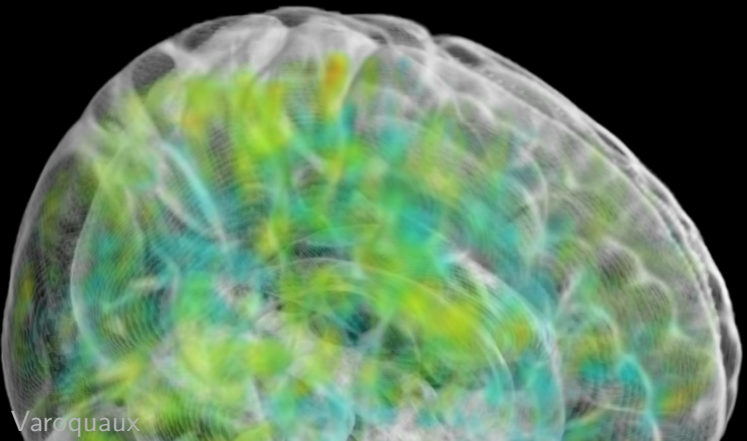
"connectome" prediction

Brain "at rest"
= easier to scan



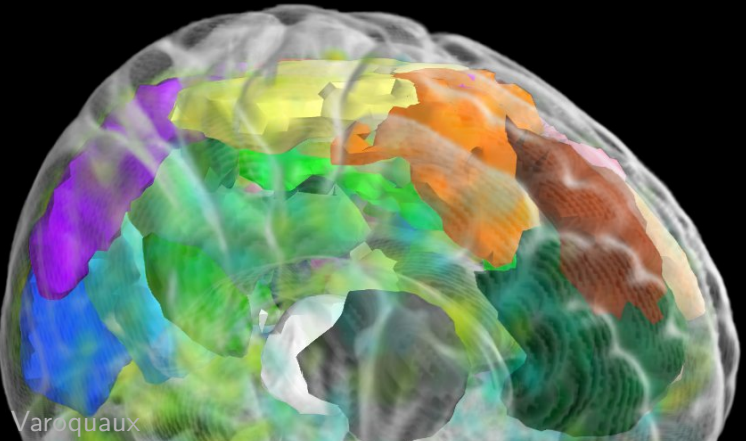
From rest-fMRI to biomarkers

No salient features in rest fMRI



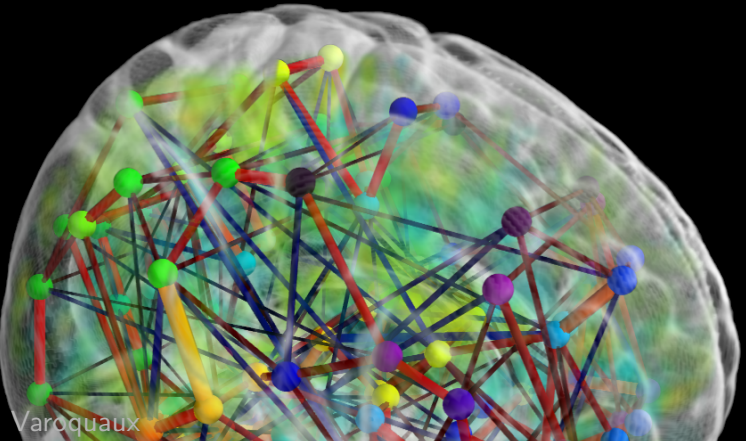
From rest-fMRI to biomarkers

- Define functional regions



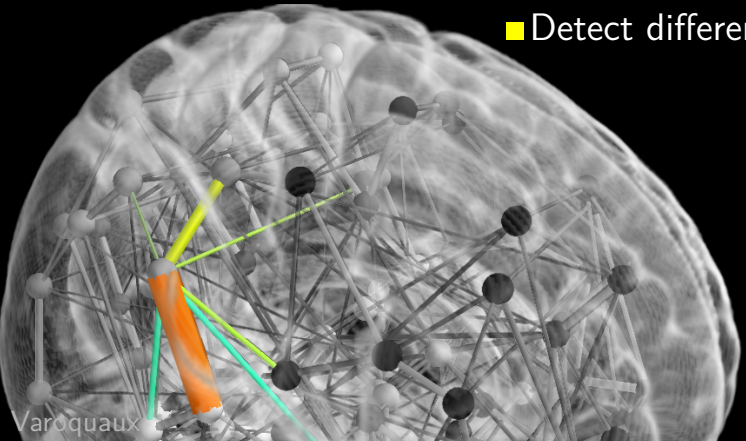
From rest-fMRI to biomarkers

- Define functional regions
- Learn interactions

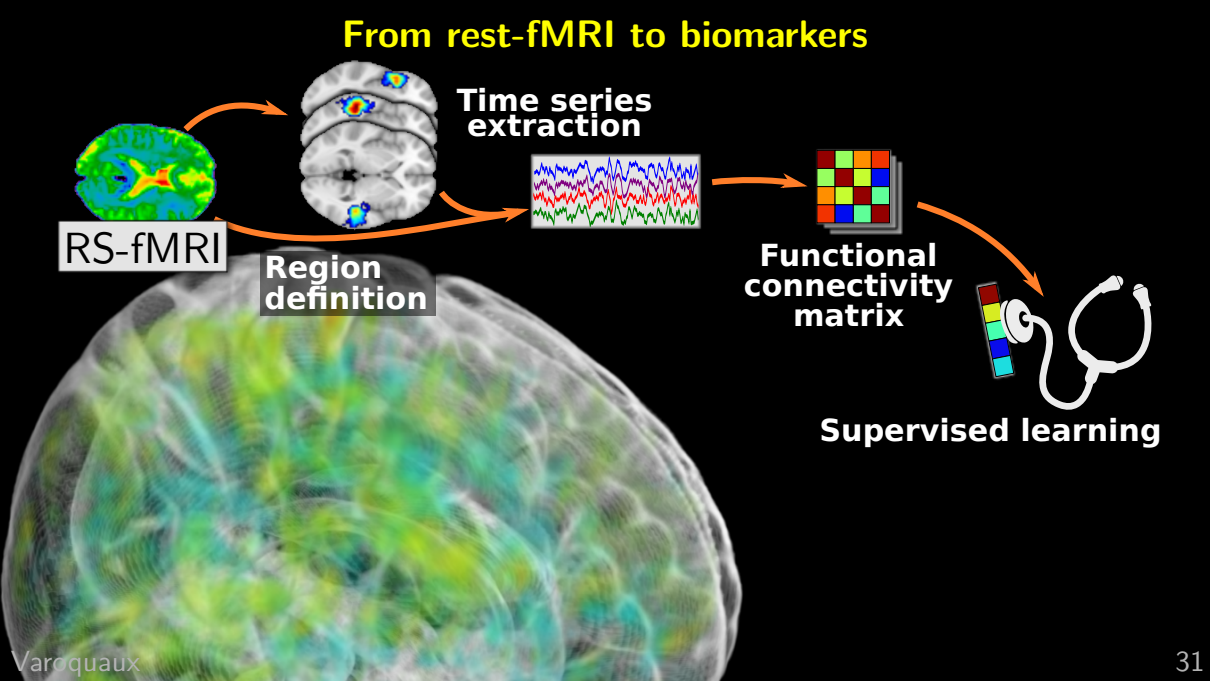


From rest-fMRI to biomarkers

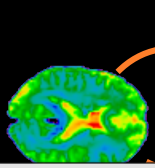
- Define functional regions
- Learn interactions
- Detect differences

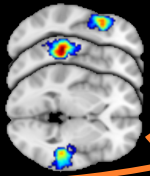


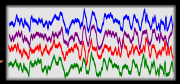
From rest-fMRI to biomarkers

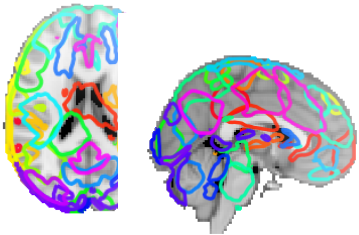


From rest-fMRI to biomarkers



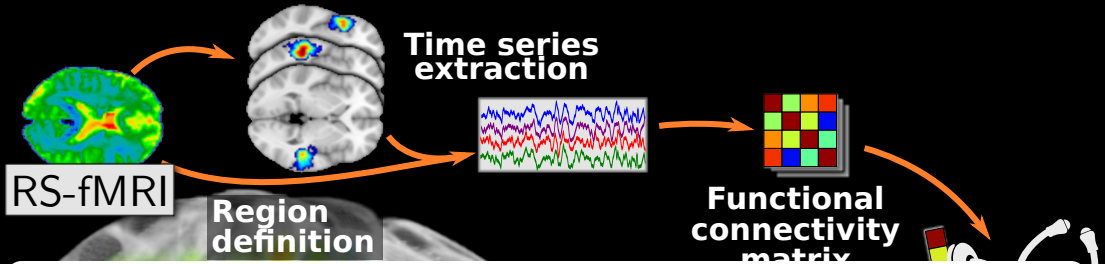


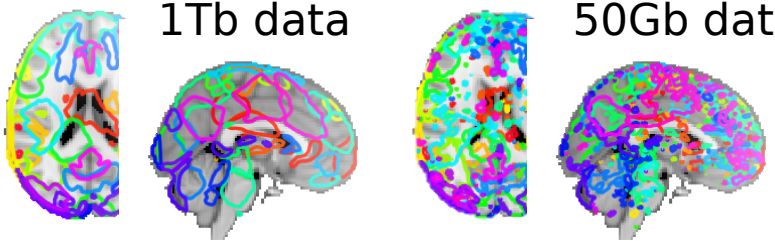




- Unsupervised learning $\Rightarrow$ adapted representations dictionary learning

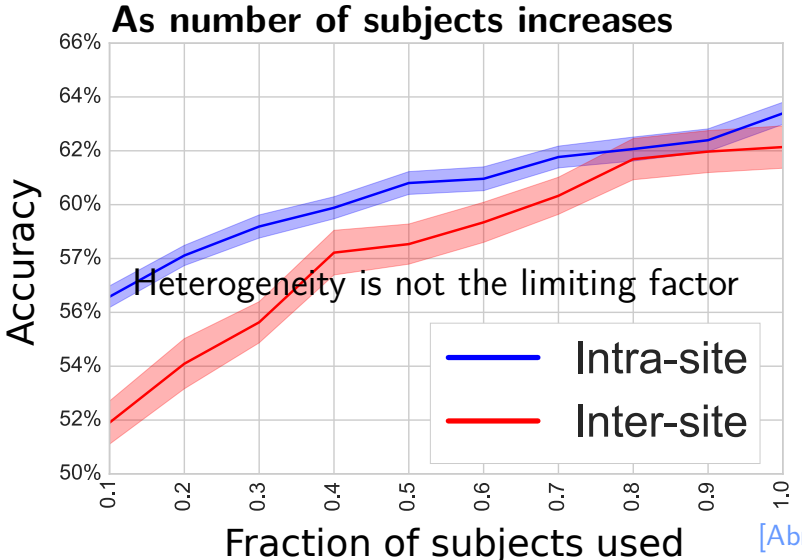
From rest-fMRI to biomarkers





- Unsupervised learning ⇒ adapted representations dictionary learning
- More data is always better computational cost [Mensch... 2016]

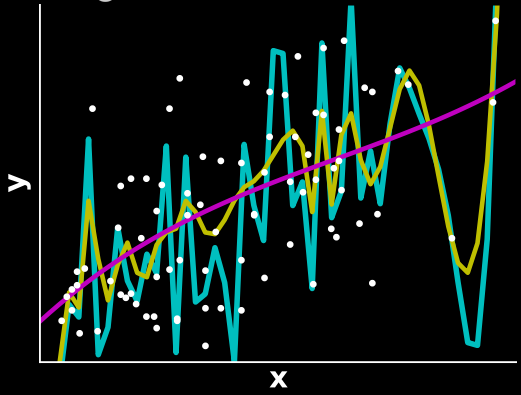
- Ill-defined pathology
- Heterogeneous dataset (ABIDE)



[Abraham... 2017]

An introduction to Machine Learning for MR imaging

- Overfit versus underfit
  - The best fit is not the best performer
  - Minimizing an observed error is the wrong solution



An introduction to Machine Learning for MR imaging

- Overfit versus underfit
- Evaluate on **independent** data

Cross-validation error bars

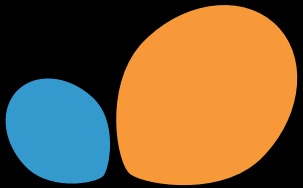
An introduction to Machine Learning for MR imaging

- Overfit versus underfit
- Evaluate on **independent** data
- More data trumps being clever

An introduction to Machine Learning for MR imaging

- Overfit versus underfit
- Evaluate on **independent** data
- More data trumps being clever
- Software: scikit-learn & nilearn

http://scikit-learn.org



In Python

http://nilearn.github.io

@GaelVaroquaux

References I

A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, page 17, 2013.

R. Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

J. V. Haxby, I. M. Gobbini, M. L. Furey, ... Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425, 2001.

A. Hoyos-Idrobo, G. Varoquaux, Y. Schwartz, and B. Thirion. Frem – scalable and stable decoding with fast regularized ensemble of models. *NeuroImage*, 2017.

References II

M. A. Little, G. Varoquaux, S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording. Using and understanding cross-validation strategies. perspectives on saeb et al. *GigaScience*, 6(5):1–6, 2017.

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Dictionary learning for massive matrix factorization. In *International Conference on Machine Learning*, pages 1737–1746, 2016.

V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fMRI-based prediction of behavior. *Medical Imaging, IEEE Transactions on*, 30:1328, 2011.

S. M. Smith, T. E. Nichols, D. Vidaurre, A. M. Winkler, T. E. Behrens, M. F. Glasser, K. Ugurbil, D. M. Barch, D. C. Van Essen, and K. L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565–1567, 2015.

G. Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage*, 2017.

References III

G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3:28, 2014.

G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *ICML*, page 1375, 2012.

G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *Trans Inf Theory*, 55:2183, 2009.