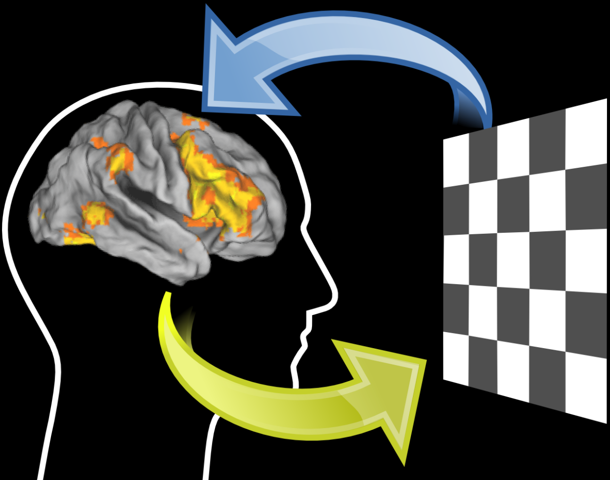
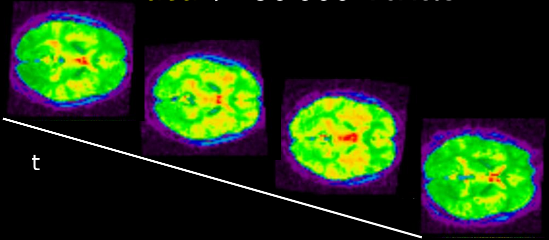
Brain maps from machine learning? Spatial regularizations

Gaël Varoquaux
Inria
PARIETAL



fMRI data > 50 000 voxels

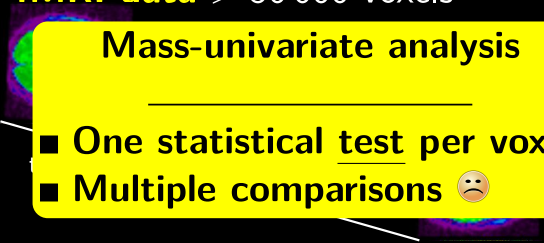


stimuli

Standard analysis

Detect voxels that correlate to the stimuli

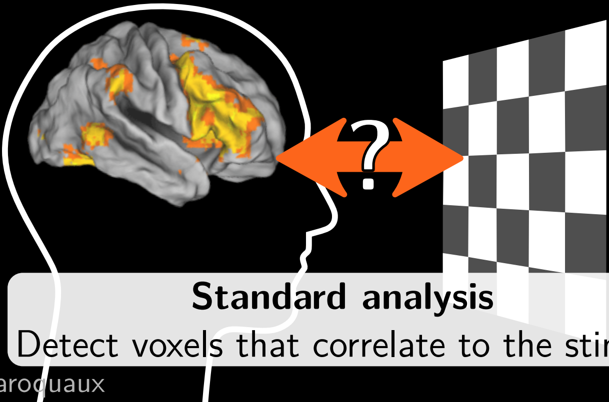
fMRI data $> 50\,000$ voxels



stimuli

Mass-univariate analysis

- One statistical test per voxel
- Multiple comparisons ☹



Standard analysis

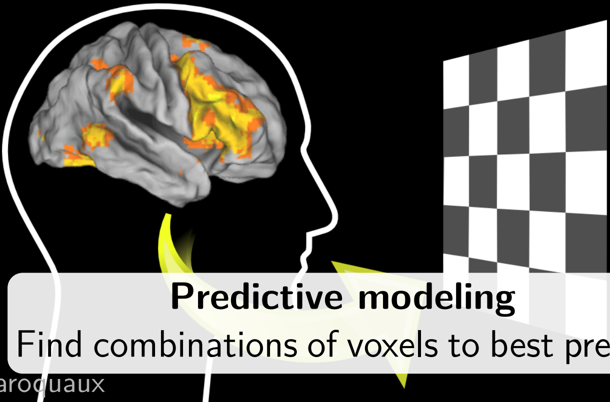
Detect voxels that correlate to the stimuli

Brain decoding

- Predicting stimulus / cognitive state

  [Haxby... 2001] *Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex*

Supervised learning task



Predictive modeling

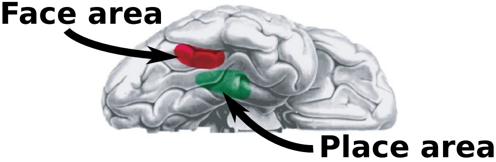
Find combinations of voxels to best predict

Brain decoding

- Predicting stimulus / cognitive state

  [Haxby... 2001] *Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex*

Take home message:
brain regions, not prediction

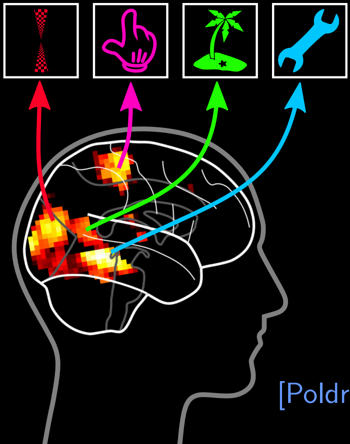


Predictive modeling

Find combinations of voxels to best predict

Inference in cognitive neuroimaging

- What is the neural support of a function?
- What is function of a given brain module?

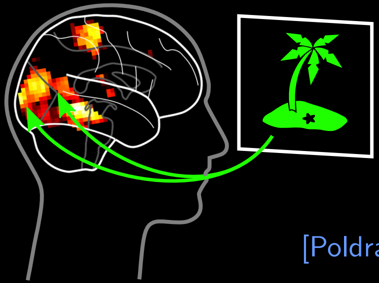


[Poldrack 2006, Henson 2006]

Inference in cognitive neuroimaging

- What is the neural support of a function?
- What is function of a given brain module?

Brain mapping = task-evoked activity



[Poldrack 2006, Henson 2006]

Inference in cognitive neuroimaging

- What is the neural support of a function?

- What is function of a given brain module?

Reverse inference ⚠

Brain mapping = task-evoked activity
+ crafting "contrasts" to isolate effects

[Poldrack 2006, Henson 2006]

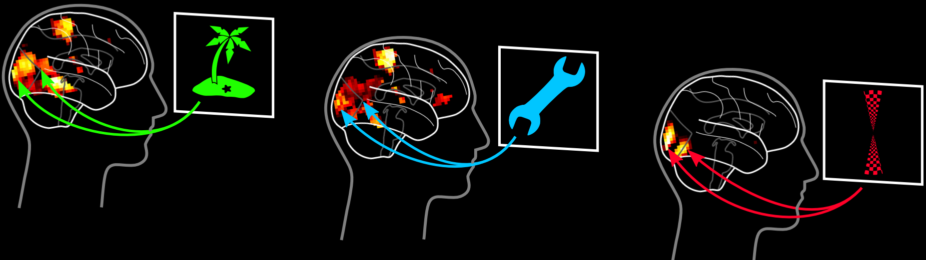
Inference in cognitive neuroimaging

- What is the neural support of a function?

- What is function of a given brain module?

Reverse inference

Is there a face area?



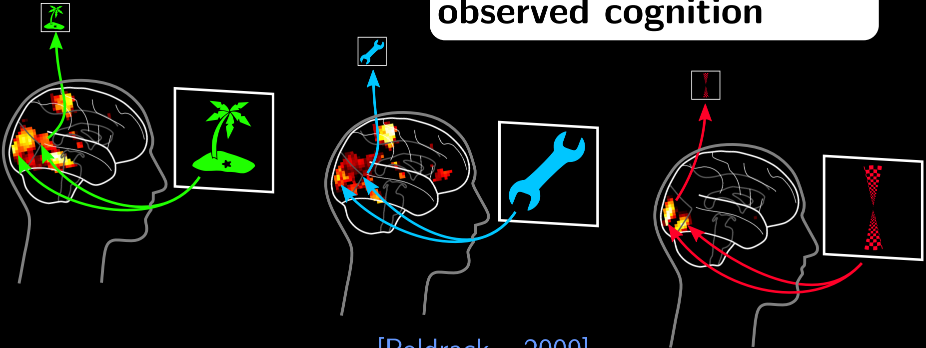
[Kanwisher... 1997, Gauthier... 2000, Hanson and Halchenko 2008]

Inference in cognitive neuroimaging

- What is the neural support of a function?
- What is function of a given brain module?

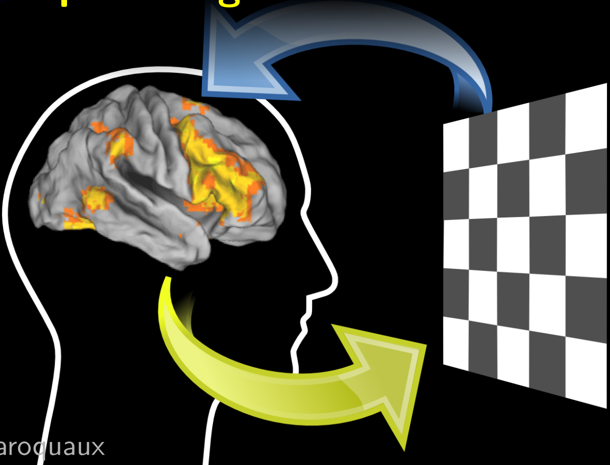
Reverse inference

Find regions that predict observed cognition

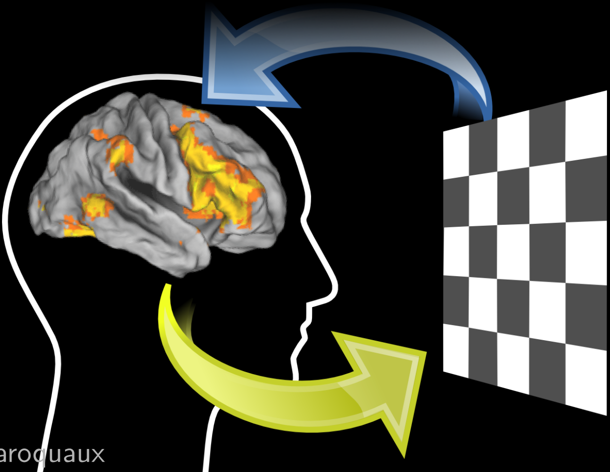


[Poldrack... 2009]

1. Recovering predictive regions is hard
2. Opening the black box of decoders
3. Spatial regularization



1 Recovering predictive regions is hard

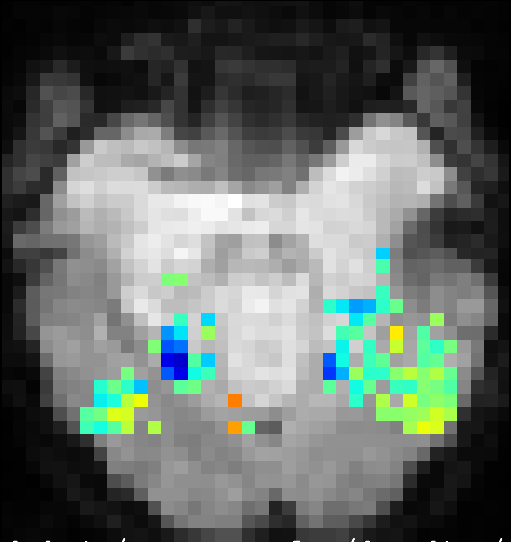


1 Brain decoding to recover predictive regions?

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

SVM

error: 26%



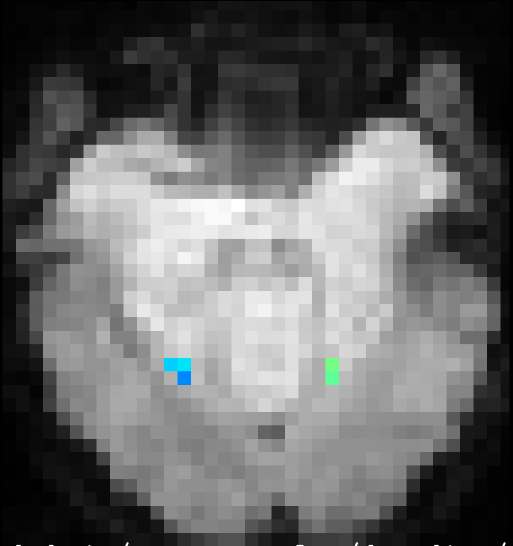
http://nilearn.github.io/auto_examples/decoding/plot_haxby_different_estimators.html

1 Brain decoding to recover predictive regions?

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

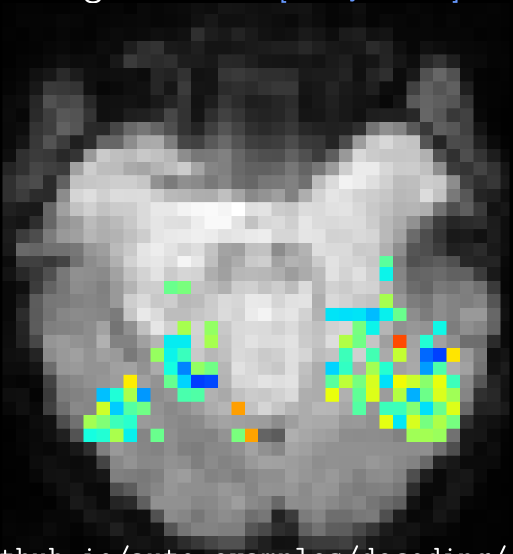
Sparse model

error: 19%



http://nilearn.github.io/auto_examples/decoding/plot_haxby_different_estimators.html

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



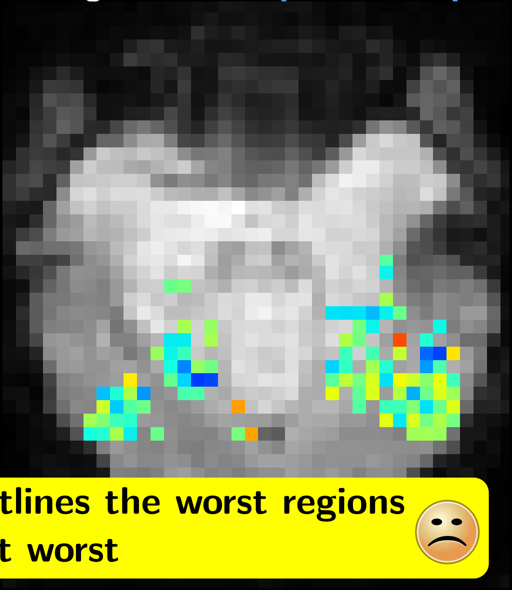
Ridge

error: 15%

http://nilearn.github.io/auto_examples/decoding/plot_haxby_different_estimators.html

Brain decoding to recover predictive regions?

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

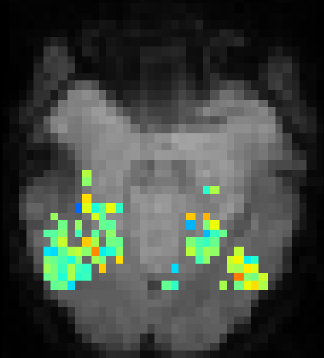


- Best predictor outlines the worst regions
- Best maps predict worst

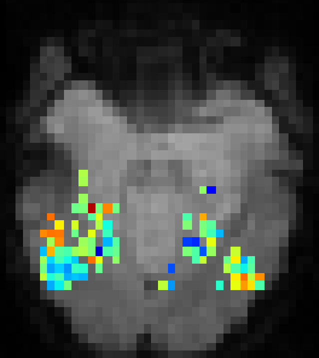
Brain decoding with local models

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

SVM error: 24%



First 6 sessions

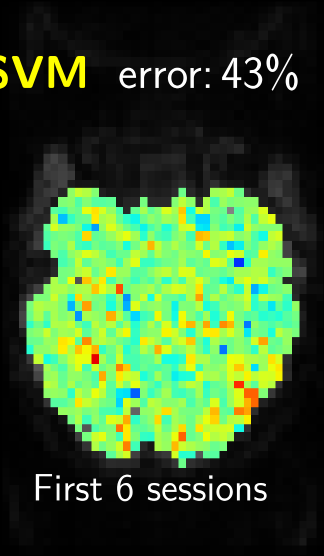


Last 6 sessions

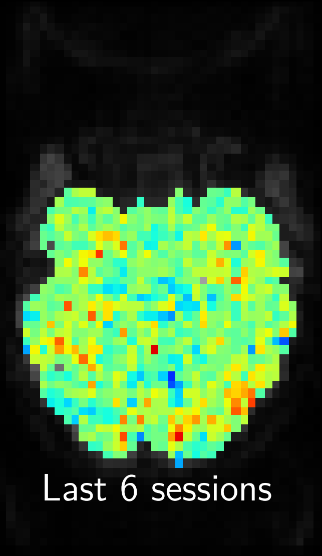
1 Brain decoding with local models

Face vs cat visual recognition **full brain**

SVM error: 43%



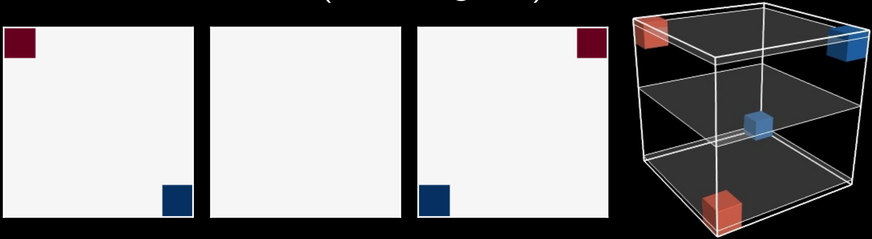
First 6 sessions



Last 6 sessions

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

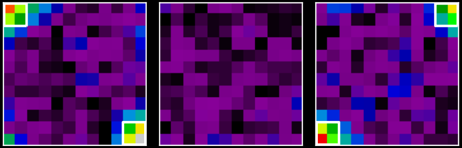
- **X**: observed fMRI images: spatially smooth
- **e**: noise
- **w**: true coefficients (brain regions)



SVM

Prediction: 0.71

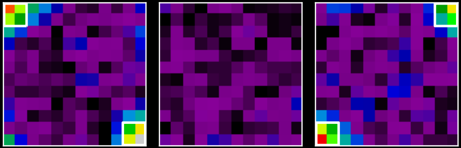
Recovery: 0.464



SVM

Prediction: 0.71

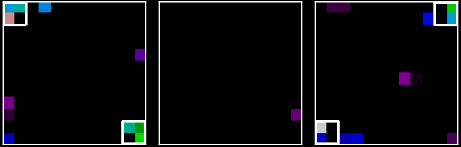
Recovery: 0.464



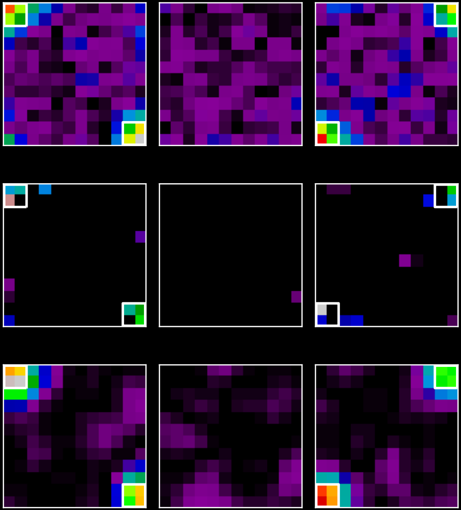
Sparse models

Prediction: 0.77

Recovery: 0.461



1 Simple simulations



SVM

Prediction: 0.71

Recovery: 0.464

Sparse models

Prediction: 0.77

Recovery: 0.461

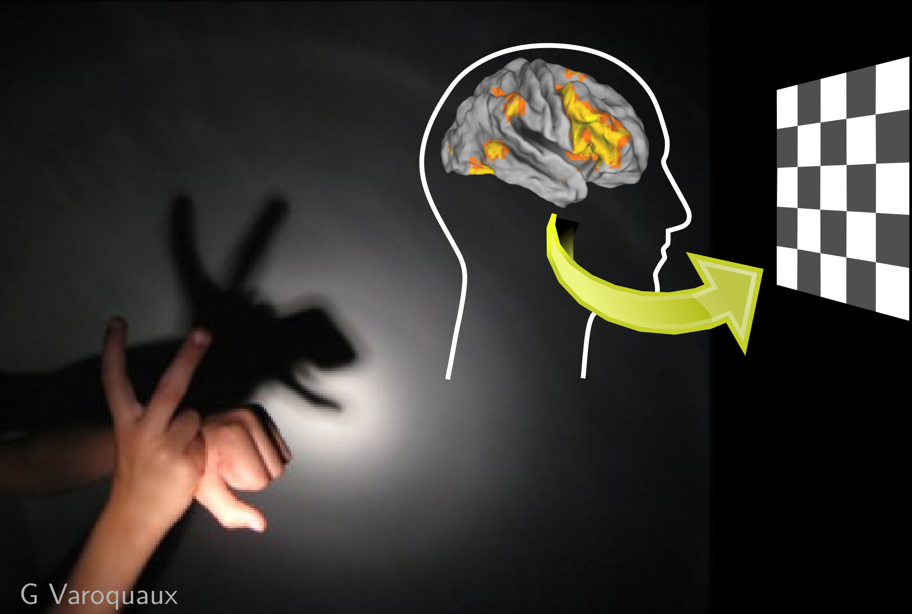
F-score

~~Prediction~~:

Recovery: 0.963

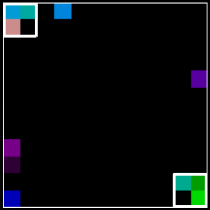
Good prediction $\neq$ good recovery

Recovering predictive regions is hard



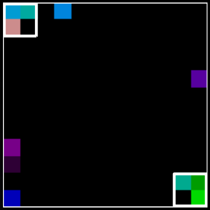
Recovering predictive regions is hard

- Not all informative features are need to predict

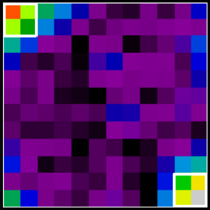


Recovering predictive regions is hard

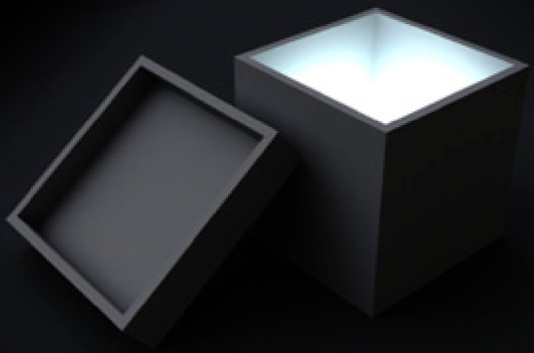
- Not all informative features are need to predict



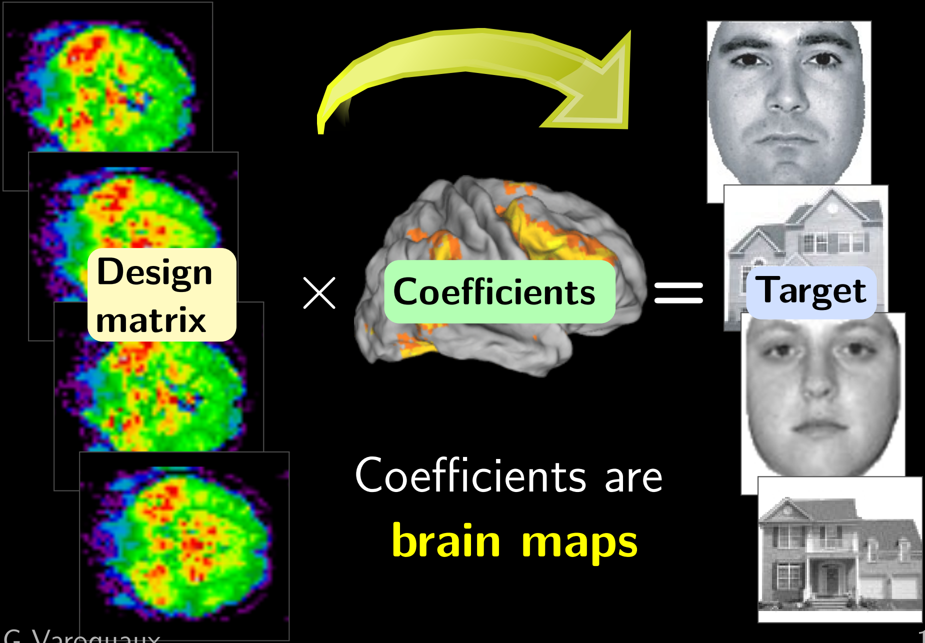
- Weight on uninformative features may not harm prediction



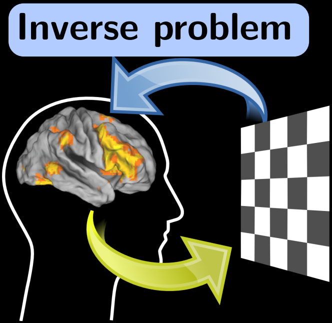
2 Opening the black box of decoders



2 Brain decoding with linear models



Coefficients are
brain maps



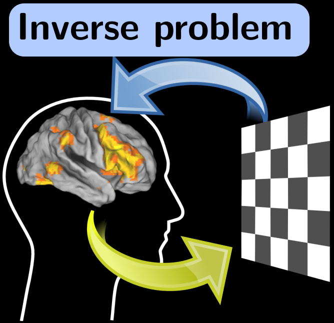
Find brain maps **w** to minimize the prediction error

Ill-posed:

Many different **w** will give the same prediction error

How to choose one?

2 Estimation: an ill-posed problem



Find brain maps **w** to minimize the prediction error

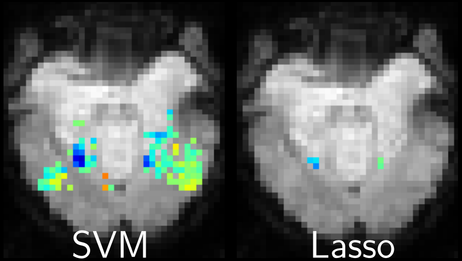
Ill-posed:

Many different **w** will give the same prediction error

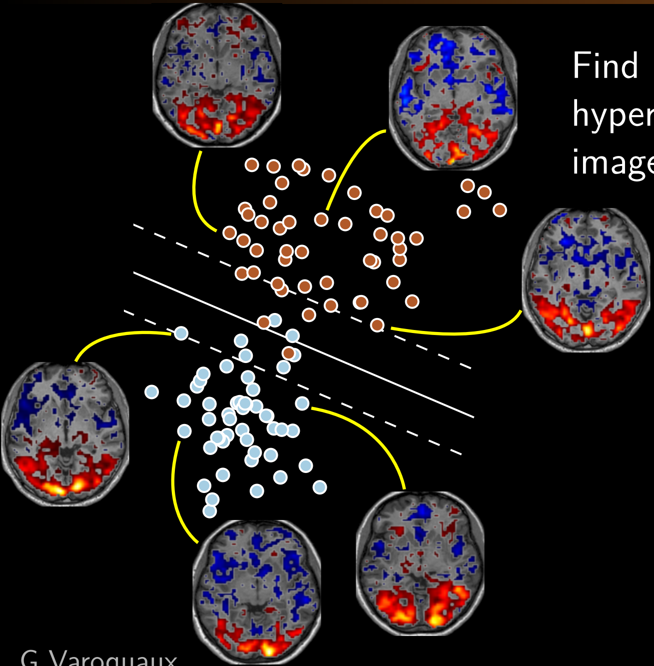
How to choose one?

- Somewhat arbitrary choice
  Not informed by the data

- Different methods
  $\Rightarrow$ different outputs

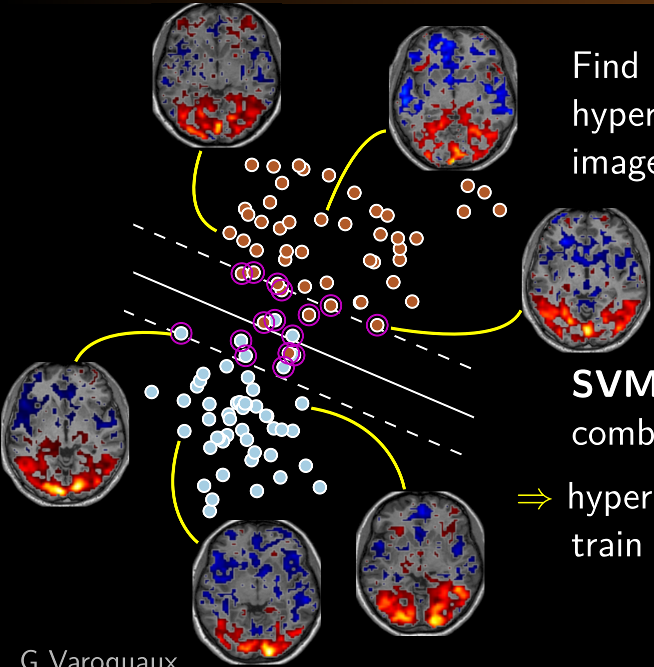


SVM: exemplar-based classifier



Find a separating hyperplane between images

2 SVM: exemplar-based classifier

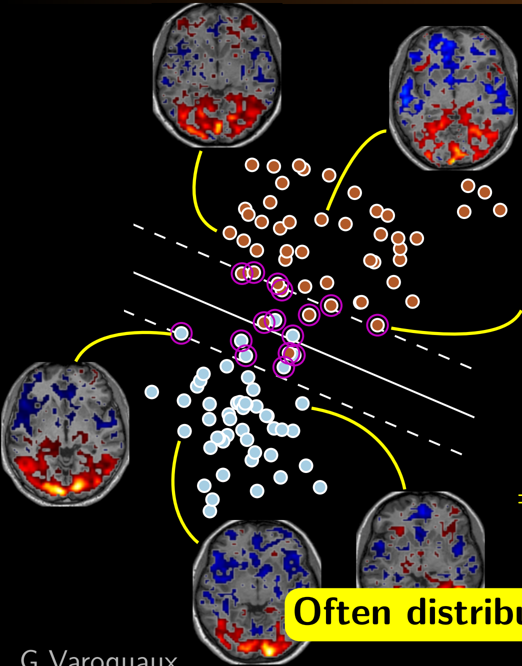


Find a separating hyperplane between images

SVM builds it by combining exemplars

$\Rightarrow$ hyperplane looks like train images

2 SVM: exemplar-based classifier

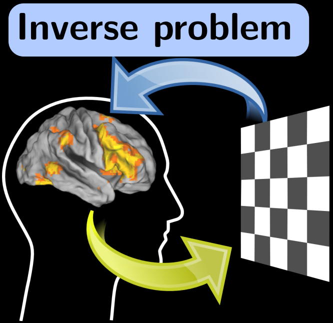


Find a separating hyperplane between images

SVM builds it by combining exemplars

$\Rightarrow$ hyperplane looks like train images

Often distributed by construction



- Minimize the error term:

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

Ill-posed:

Many different **w** will give the same prediction error

- To choose one: inject prior with a penalty

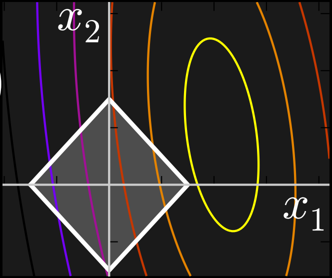
$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} l(\mathbf{y} - \mathbf{X}\,\mathbf{w}) + p(\mathbf{w})$$

- Lasso estimator

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

Data fit

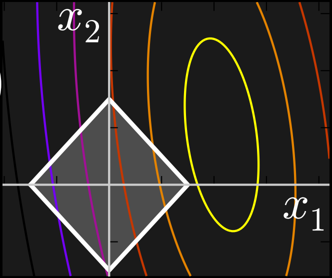
Penalization



Sparse models: selecting predictive voxels?

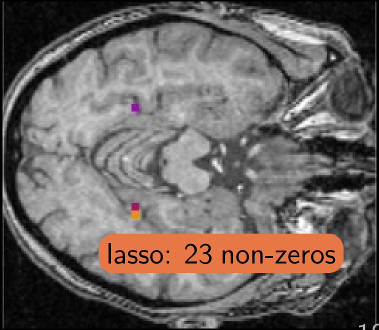
- Lasso estimator

$$\min_{\mathbf{x}} \underbrace{\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2}_{\text{Data fit}} + \lambda \underbrace{\ell_1(\mathbf{x})}_{\text{Penalization}}$$



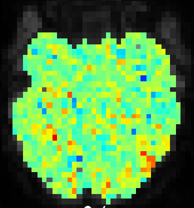
- Between correlated features, selects a random subset
  [Wainwright 2009, Varoquaux... 2012]
  [Rish... 2012]

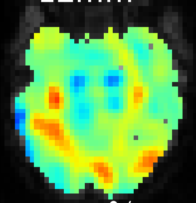
  Violates the restricted isometry property



Spatial smoothing

12mm

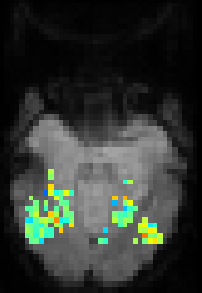




Face vs cat error rate: 43% 31%

Giving up on resolution

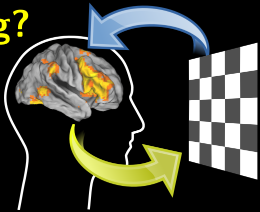
Feature selection



23%

Giving up on multivariate

From prediction to mapping?

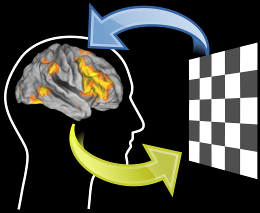


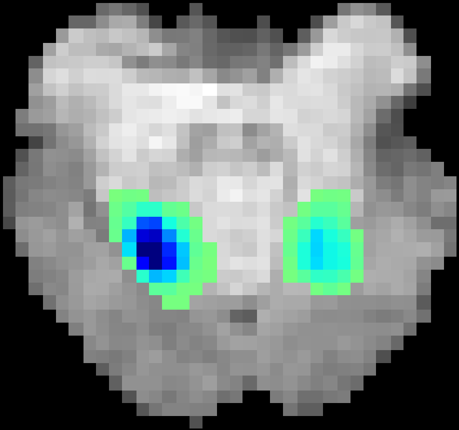
Inverse problem intractable in general

Need suitable simplifying hypothesis

- Spatial structure / contiguity (as Random Field Theory)
  $\Rightarrow$ Smoothed SVM work better
- Only a fraction of the brain useful to predict
  $\Rightarrow$ Feature-selection + Sparsity
- Neighbooring voxel multi-colinear
  $\Rightarrow$ Break multivariate estimators

3 Spatial regularization



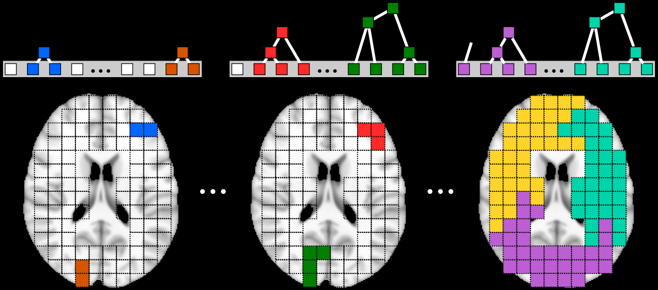


Clustering to group similar voxels

Hierarchical clustering:

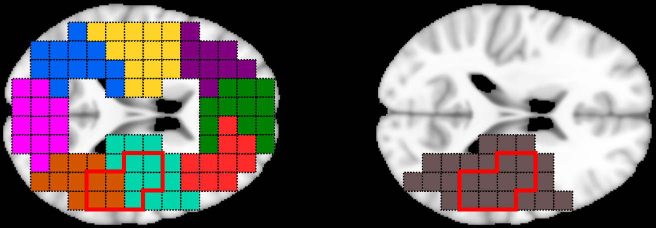
merge voxels with similar behavior

⇒ Good conditions for sparse models



[Varoquaux... 2012]

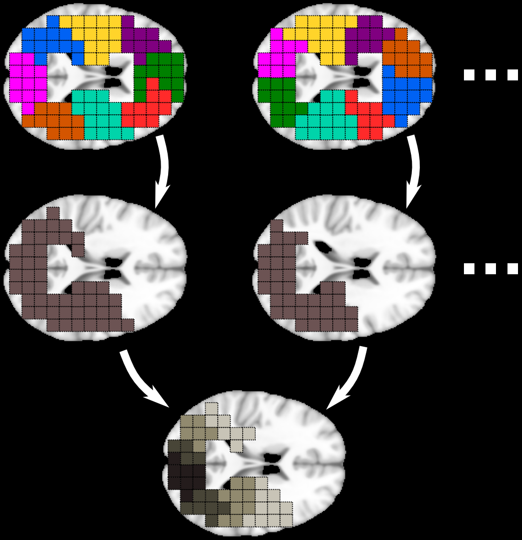
Sparse model on brain parcellation



Mapping is much easier on a parcellation [Filippone... 2012]

Parcellation / clustering is not always perfect 

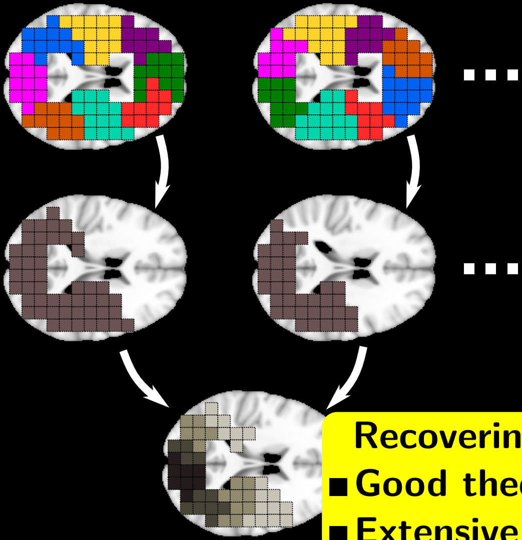
3 Randomized parcellations + sparsity



- Cluster sub-sampled data
- Average the results

[Varoquaux... 2012]

3 Randomized parcellations + sparsity



- Cluster sub-sampled data
- Average the results

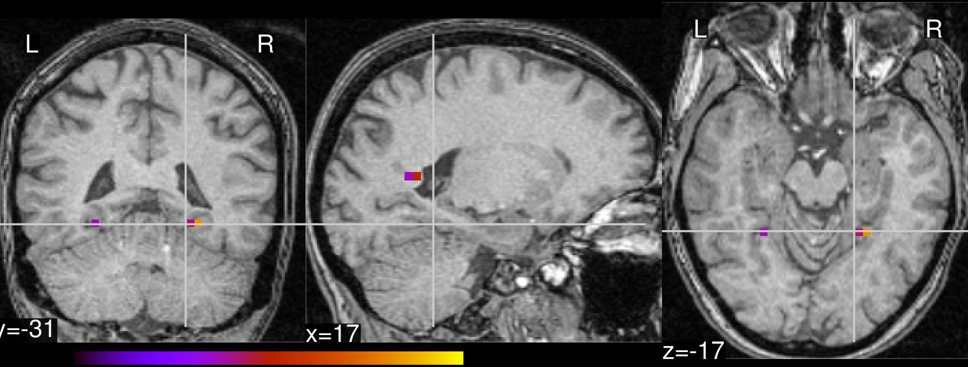
Recovering predictive regions

- **Good theoretical argument**
- **Extensive simulations**

[Varoquaux... 2012]

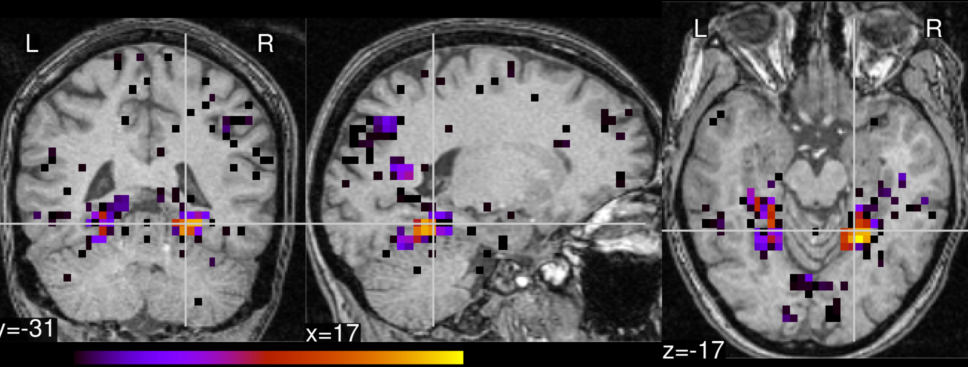
3 fMRI: face vs house discrimination [Haxby... 2001]

Sparse model



[Varoquaux... 2012]

Randomized Clustered ℓ_1 Logistic



[Varoquaux... 2012]

3 better prediction scores [Hoyos-Idrobo... 2015]

dataset		sparse	sparse + clustered
ds105 (haxby)	bottle/scramble	0.591	**0.626**
	cat/chair	0.558	**0.612**
	cat/house	0.698	**0.963**
	chair/house	0.668	**0.734**
	chair/scramble	0.700	**0.743**
	face/house	**0.766**	0.742
	tools/scramble	0.666	**0.743**
ds107	consonant/scramble	0.886	**0.897**
	objects/consonant	0.855	**0.901**
	objects/scramble	0.863	**0.898**
	objects/words	0.689	**0.708**
	words/scramble	0.782	**0.841**
ds108	negative cue / neutral cue	0.444	**0.497**
	negative rating / neutral rating	0.520	**0.537**
	negative stim / neutral stim	0.734	**0.743**
ds 109	false picture / false belief	0.664	**0.675**
oasis (vbm)	gender discrimination	0.617	**0.655**

3 better prediction scores [Hoyos-Idrobo... 2015]

dataset		sparse	sparse + clustered
ds105 (haxby)	bottle/scramble	0.591	**0.626**
	cat/chair	0.558	**0.612**
	cat/house	0.698	**0.963**
	chair/house	0.668	**0.734**
	chair/scramble	0.700	**0.743**
	face/house	**0.766**	0.742
			43
			97
			01
ds10			**98**
			08
			41
ds108	negative cue / neutral cue	[illegible]	**[illegible]497**
	negative rating / neutral rating	0.520	**0.537**
	negative stim / neutral stim	0.734	**0.743**
ds 109	false picture / false belief	0.664	**0.675**
oasis (vbm)	gender discrimination	0.617	**0.655**

Sparse-clustered: faster, better

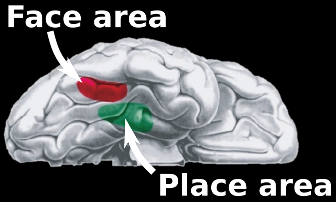
Faster: 10x speed up, due to smaller models

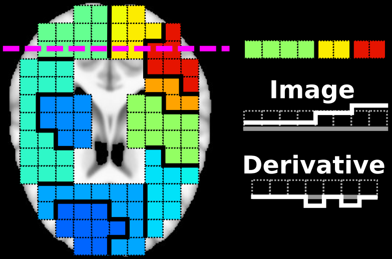
More stable: bagging effect

Averaging many estimates

Total variation: a penalty for regions

- Sparsity is useful
- Want to recover brain regions





Total-variation penalization

Impose sparsity on the gradient of the image:

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

In fMRI: [Michel... 2011]

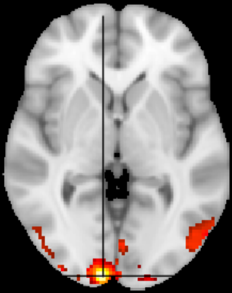
Related to GraphNet [Grosenick... 2013]

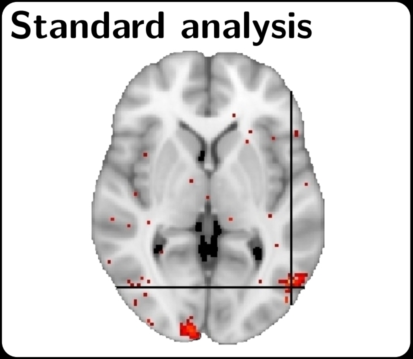
Decoding prediction performance:

SVM	0.77
Sparse regression	0.78
Total Variation	**0.84**

(explained variance)

[Michel... 2011]





3 Penalty engineering: TV-ℓ_1

Sparsity + **regions** [Baldassarre... 2012, Gramfort... 2013]

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\; l(\mathbf{y} - \mathbf{X}\,\mathbf{w}) + \lambda\Big(\rho\,\ell_1(\mathbf{w}) + (1-\rho)\,TV(\mathbf{w})\Big)$$

l: data-fit term

Poster 3980, Thursday

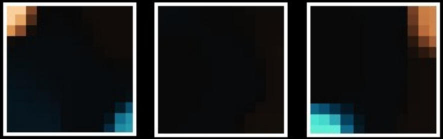
Sparsity + regions [Baldassarre... 2012, Gramfort... 2013]

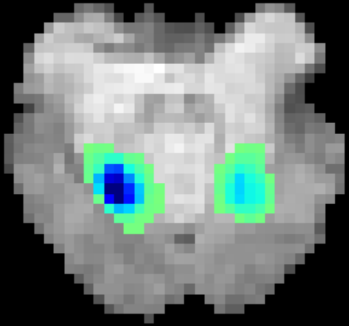
$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w}) + \lambda\Big(\rho\,\ell_1(\mathbf{w}) + (1-\rho)\,TV(\mathbf{w})\Big)$$

l: data-fit term

Poster 3980, Thursday

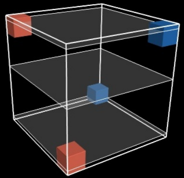
Experiment results

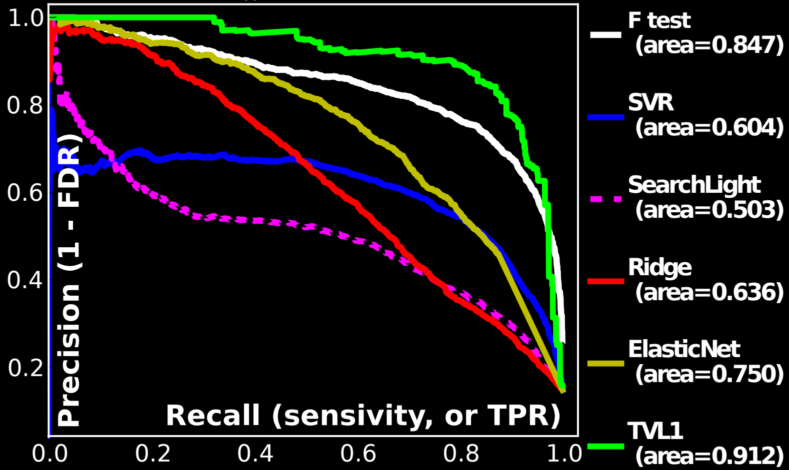




Vary threshold, and count false positive/false negatives

- Precision ∼ FDR
- Recall ∼ # detections

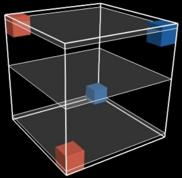




3 Prediction on simulations [Gramfort... 2013]

Prediction error

SNR	2.5	5.0	7.5	10.0
SVR	28.4	22.6	18.1	14.5
Ridge	27.8	21.6	17.0	13.4
ElasticNet	**26.6**	20.7	16.4	13.1
TV-ℓ_1	**25.7**	**20.1**	**15.8**	**12.5**

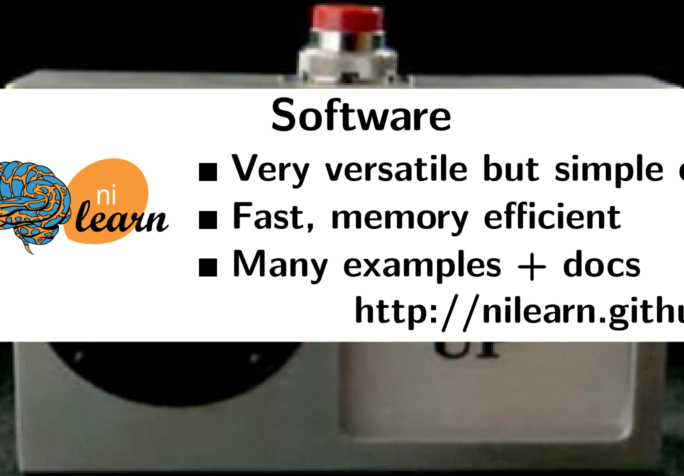


Prediction is easy, region recovery is hard

☺

Wrapping up

Wrapping up



Software

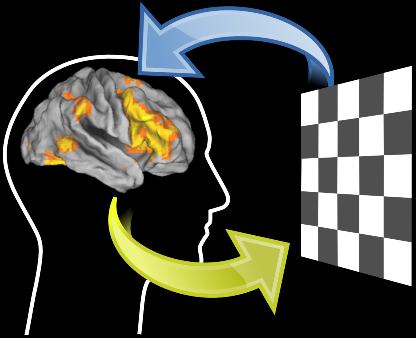
- Very versatile but simple code
- Fast, memory efficient
- Many examples + docs

http://nilearn.github.io

Decoding for reverse-inference mapping

- Standard decoders do not retrieve good maps

  Infinite number of maps predict as well



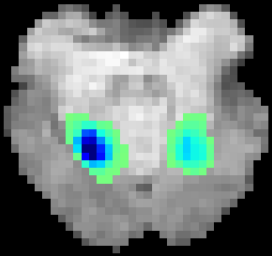
@GaelVaroquaux

Decoding for reverse-inference mapping

- Standard decoders do not retrieve good maps
- Spatial models to select predictive regions

TV-ℓ_1 = Sparsity + regions

Randomized clustering + sparsity



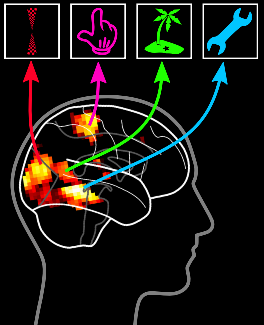
@GaelVaroquaux

Decoding for reverse-inference mapping

- Standard decoders do not retrieve good maps
- Spatial models to select predictive regions
- Decoding + mega-analysis $\Rightarrow$ reverse-inference atlas

Multi-label prediction

Predicting on new studies



@GaelVaroquaux

Decoding for reverse-inference mapping

- Standard decoders do not retrieve good maps
- Spatial models to select predictive regions
- Decoding + mega-analysis $\Rightarrow$ reverse-inference atlas
- Software: nilearn

In Python
http://nilearn.github.io

[Varoquaux and Thirion 2014]
How machine learning is
shaping cognitive neuroimaging

@GaelVaroquaux

References I

L. Baldassarre, J. Mourao-Miranda, and M. Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, page 5, 2012.

M. Filippone, A. F. Marquand, C. R. Blain, S. C. Williams, J. Mourão-Miranda, and M. Girolami. Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. *The annals of applied statistics*, 6 (4):1883, 2012.

I. Gauthier, M. J. Tarr, J. Moylan, P. Skudlarski, J. C. Gore, and A. W. Anderson. The fusiform "face area" is part of a network that processes faces at the individual level. *J cognitive neuroscience*, 12:495, 2000.

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

References II

L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor. Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, 72:304, 2013.

S. J. Hanson and Y. O. Halchenko. Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. *Neural Computation*, 20:486, 2008.

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.

R. Henson. Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in cognitive sciences*, 10:64, 2006.

A. Hoyos-Idrobo, Y. Schwartz, B. Thirion, and G. Varoquaux. Improving sparse recovery on structured images with bagged clustering. In *PRNI*. 2015.

References III

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neuroscience*, 17:4302, 1997.

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.

R. Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, 10:59, 2006.

R. A. Poldrack, Y. O. Halchenko, and S. J. Hanson. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, 20:1364, 2009.

References IV

I. Rish, G. A. Cecchi, K. Heuton, M. N. Baliki, and A. V. Apkarian. Sparse regression analysis of task-relevant information distribution in the brain. In *SPIE Medical Imaging*, pages 831412–831412. International Society for Optics and Photonics, 2012.

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.

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

Decoding object recognition

8 objects multi-class prediction

Categories: face, chair, srambledpix, scissors, house, bottle, shoe, cat

Difficult!

Lasso	69.4%
TV-ℓ_1	**75.3%**

Distinguishing scissors in the dorsal stream,
faces and cats in frontal areas
...

[Eickenberg Submitted]