Sparse classification for significant anatomy detection in a group study

Dana Cobzas
University of Alberta

with
Linglong Kong   Math, U of Alberta
Mark Schmidt   CS, U of British Columbia
Alan Wilman   BME,U of Albefta

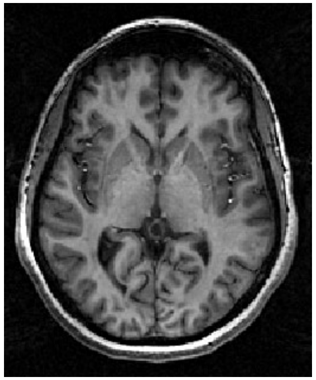
Healthy controls

Patients

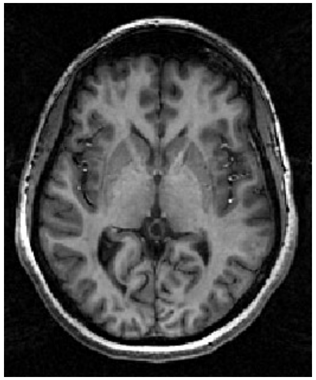
What are the areas of the brain significantly different between healthy subjects and patients ?

Voxel based analysis – data alignment

Scalar images

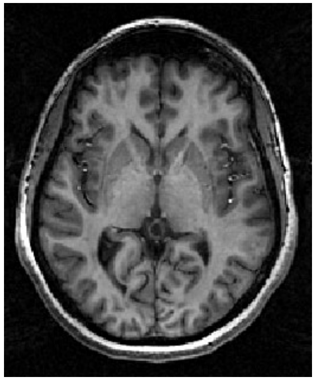


Subject 1



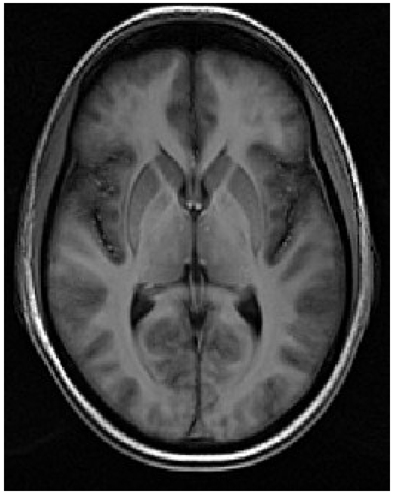
Subject 2

[...]



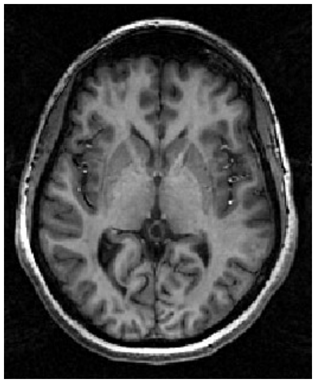
Subject n

Linear and nonlinear registration in atlas space

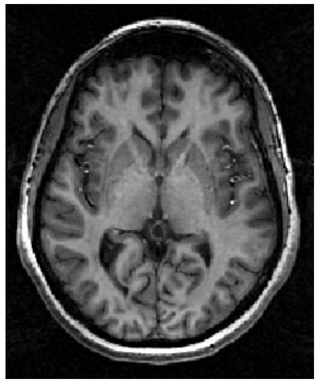


Voxel based analysis – data alignment

Scalar images

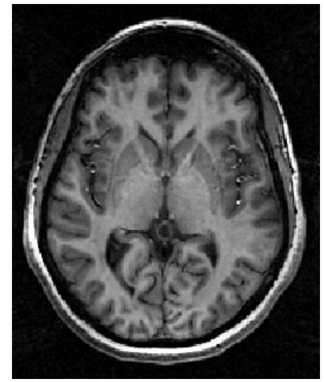


Subject 1



Subject 2

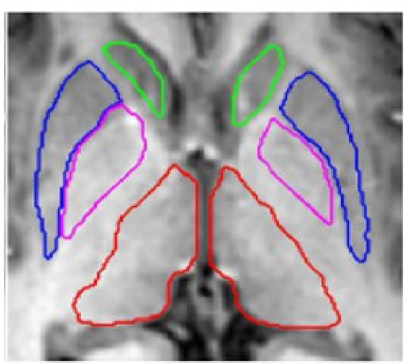
[...]



Subject n

Linear and nonlinear registration in atlas space

Flattened data:
Only relevant parts



Data

p voxels

Labels
Control/Patient

A_{nxp}

VBA – traditional approaches

Voxel based analysis:

- Voxel-based univariate approaches, permutations tests, FDR
- Advanced statistics to compensate for data correlation
  [SurfStat – Random Field Theory, Worsley, K. et al. 2008]

VBA – traditional approaches

Voxel based analysis:

- Voxel-based univariate approaches, permutations tests, FDR
- Advanced statistics to compensate for data correlation [SurfStat – Random Field Theory, Worsley, K. et al. 2008]

Reformulate the problem as a supervised dimensionality data reduction method such that

- The detected anatomy is discriminative
- Sparse
- Compact and interpretable from an anatomical viewpoint
- There is a principled way of finding an optimal model and testing its accuracy

Dimensionality reduction

Unsupervised :

- PCA, ICA : eigenvectors have global support and do not provide anatomical specificity

Dimensionality reduction

Unsupervised :

- PCA, ICA : eigenvectors have global support and do not provide anatomical specificity → Imposed **sparseness** of solution
- Sparse generative models “parts-based representations” ;
  But they do not explicitly optimize discrimination
  [Kandel, Avants et al. 2015][Lee, Seung 1999][Witten,Hastie 2009]

Dimensionality reduction

Unsupervised :

- PCA, ICA : eigenvectors have global support and do not provide anatomical specificity → Imposed **sparseness** of solution
- Sparse generative models "parts-based representations" ;
  But they do not explicitly optimize discrimination
  [Kandel, Avants et al. 2015][Lee, Seung 1999][Witten,Hastie 2009]

Supervised :

- Pattern classification methods – feature extraction and selection to achieve high classification accuracy; sparsity on feature selection;
- Goal : high classification accuracy with no focus on meaning and interpretability of selected regions; often data reduction decoupled from feature selection

[Batmanghelich et al. 2011][Sabuncu,Van Leemput 2012]
[Krishnapuram et al. 2005][Ryali et al. 2010]

Sparse classification

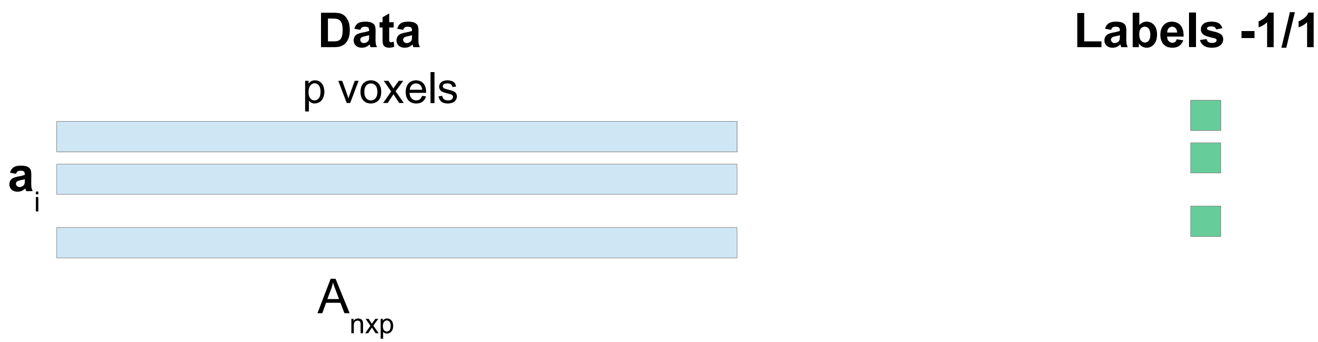
Discriminative method where **<u>relevant image regions</u>** are selected using an image-regularized sparse classification.

- The detected anatomy is discriminative
- Sparse
- Compact and interpretable from an anatomical viewpoint
- There is a principled way of finding an optimal model and testing its accuracy

- Logistic regression
- Sparseness constraint
- Image-based regularization
- Cross-validation

Similar to **[Kandel et al. 2013]** work on sparse regression

Formulation: Logistic regression



y_i follows a logistic regression distribution with location $\mathbf{a}_i\mathbf{x}+\mathbf{b}$

$$p(y_i|\mathbf{a}_i,\mathbf{x},b) = \frac{1}{1+\exp(-y_i(\mathbf{a}_i\mathbf{x}+b))}$$

image coefficients **x**
Scalar bias b

$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1+\exp(-y_i(\mathbf{a}_i\mathbf{x}+b))\right)$$

Optimal params **x**,b

Formulation: sparseness

Looking for a **sparse** solution x

$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{a}_i\mathbf{x} + b))\right)$$

$$\text{subject to} ||\mathbf{x}||_0 \leq s$$

Formulation: sparseness

Looking for a **sparse** solution x

$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{a}_i \mathbf{x} + b))\right)$$

subject to $||\mathbf{x}||_0 \leq s$

NP hard

Replace l_0 with l_1

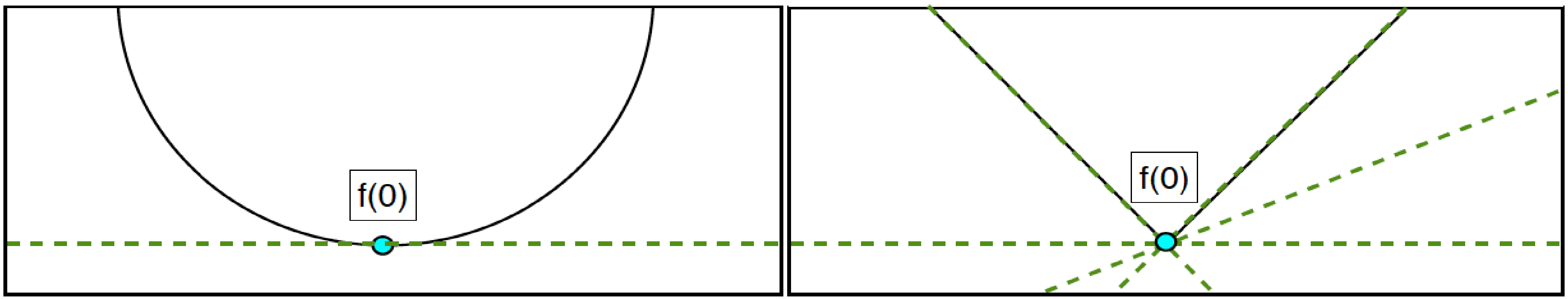
The two penalties give identical solution for many problems

$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{a}_i \mathbf{x} + b))\right) + \lambda_1 ||\mathbf{x}||_1$$

Non smooth at zero catches many solutions

L2-regularization

L1-regularization



Formulation: sparseness

Looking for a **sparse** solution x

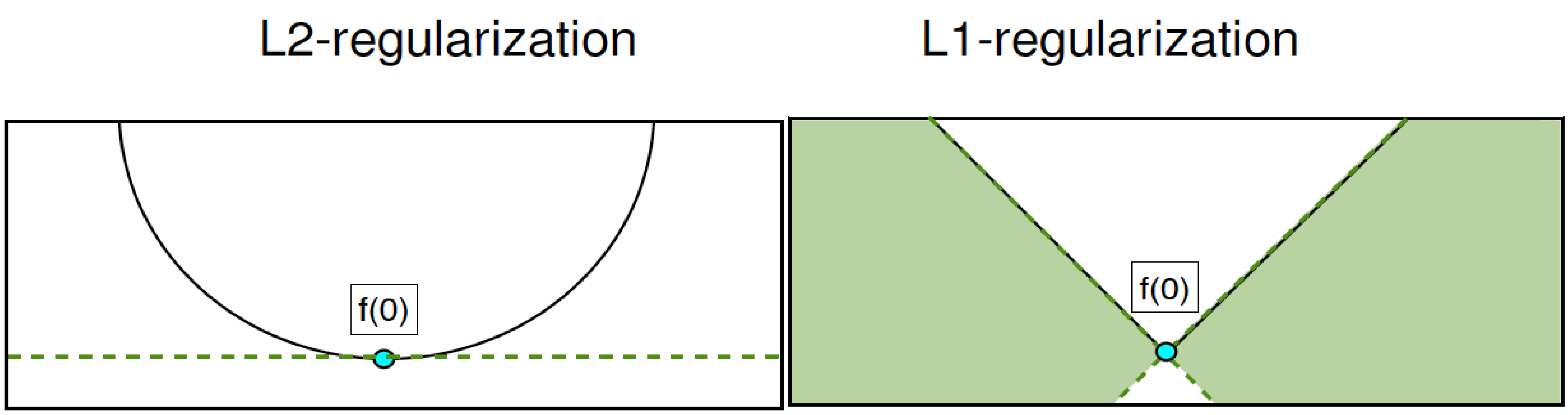
$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{a}_i\mathbf{x} + b))\right)$$

$$\text{subject to} ||\mathbf{x}||_0 \leq s$$

NP hard
Replace l_0 with l_1
The two penalties give identical solution for many problems

$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{a}_i\mathbf{x} + b))\right) + \lambda_1 ||\mathbf{x}||_1$$

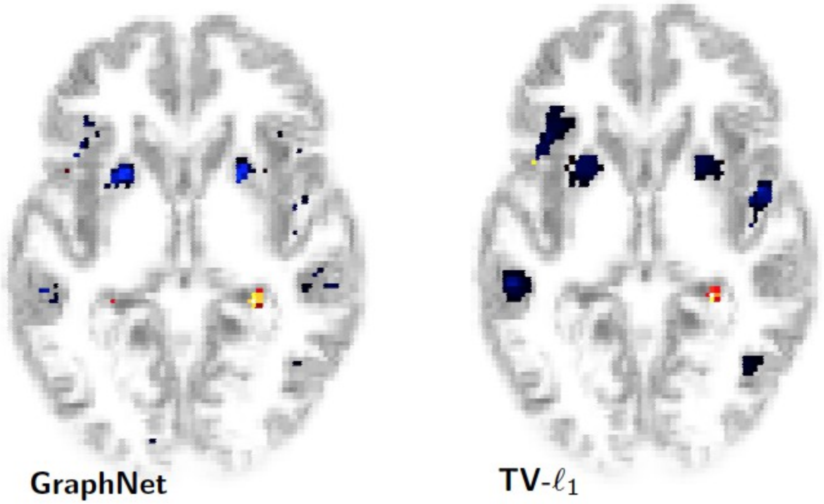
Non smooth at zero catches many solutions



Formulation: compactness

Problem:

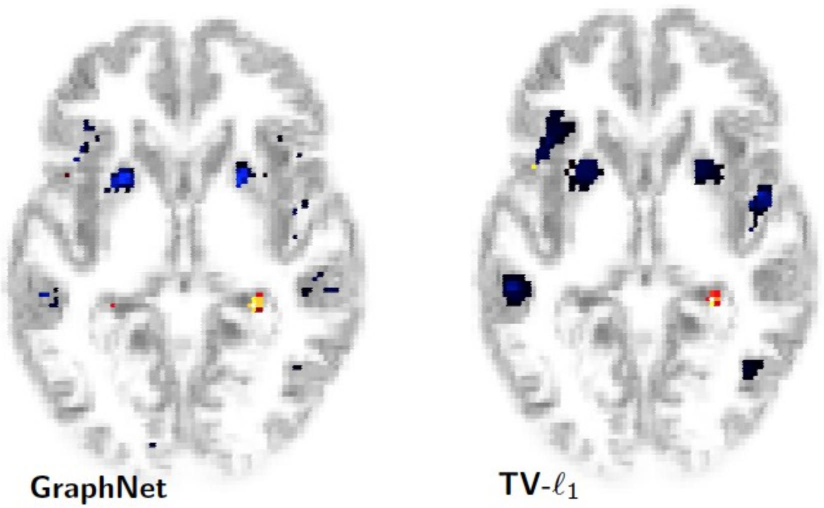
- We like the results to have an anatomical interpretation
- Add <u>image-based regularization terms</u>



Formulation: compactness

Problem:

- We like the results to have an anatomical interpretation
- Add image-based regularization terms



Diffusion based regularization

	Diffusion	*Regularization*
• Uniform diffusion	$\partial_t x = \Delta x$	$\|\nabla x\|_2^2$
• Total variation TV-l_1		$\|\nabla \mathbf{x}\|_{2,1} = \sqrt{\partial_x \mathbf{x}^2 + \partial_y \mathbf{x}^2 + \partial_z \mathbf{x}^2}$
Nonlinear isotropic diffusion [Weickert]	$\partial_t \mathbf{x} = \Psi'(\|\nabla \mathbf{x}\|^2)$	$\Psi(\|\nabla \mathbf{x}\|^2)$

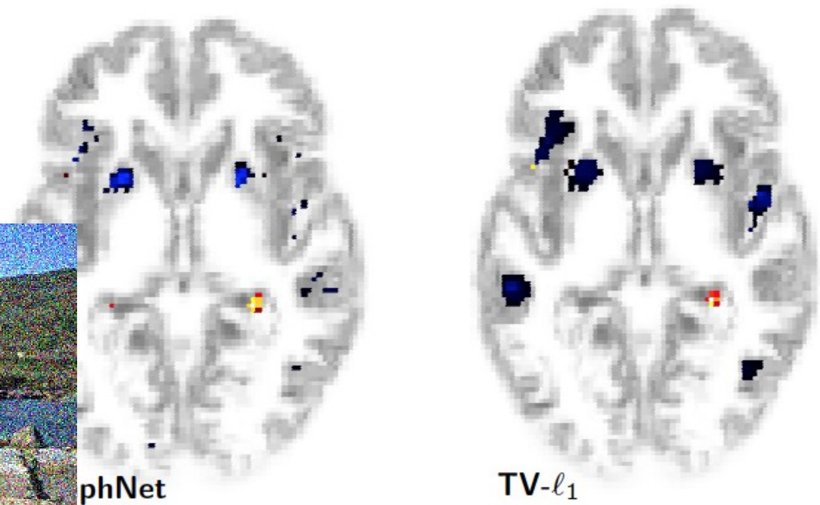
$$\Psi(s) = \sqrt{\beta^2 + s^2}$$

$$\Psi'(s) = \frac{1}{\sqrt{\beta^2 + s^2}}$$

Formulation: compactness

Problem:

- We like the results to have an anatomical interpretation
- Add <u>image-based regularization terms</u>



Diffusion

<u>Diffusion based regularization</u>

- Uniform diffusion

$$\partial_t x = \Delta x$$

$$\overline{\mathbf{x}^2 + \partial_y \mathbf{x}^2 + \partial_z \mathbf{x}^2}$$

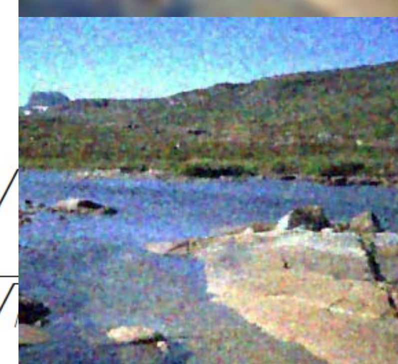
- Total variation TV-l_1

  Nonlinear isotropic diffusion [Weickert]

$$\partial_t \mathbf{x} = \Psi'(\|\nabla \mathbf{x}\|^2)$$

$$\Psi(s) = \sqrt{}$$

$$\Psi'(s) = \frac{}{\sqrt{}}$$



Optimization

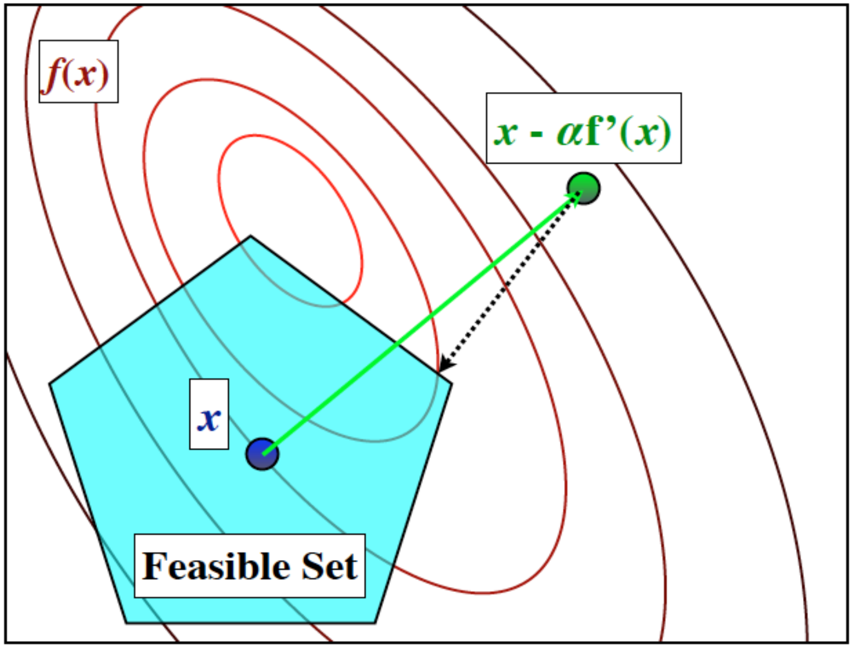
$$\min_{\mathbf{x},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{a}_i\mathbf{x} + b))\right) + \boxed{\lambda_1 \|\mathbf{x}\|_1} + \lambda_2 \|\nabla x\|_2^2$$

Solving this optimization is complicated due to non-differentiability of l_1 terms.
Use projected gradient methods

- choose an approximate steepest descent direction (pseudo-gradient) of the objective function
- take an approximate Newton step
- project solution (make x_k=0 if sign changed)

[Mark Schmidt UBC]

$$x^+ = \text{project}_{\mathcal{C}}[x - \alpha f'(x)],$$



Experiments

Compare **3 methods** :

- **Eigenanatomy** **[Avants et al.]**
  unsupervised detection of sparse regions using sparse PCA
  use projection of data onto eigenvectors in a logistic regression classification method
- **SurfStat** **[Worsley et al.]**
  compute significant regions from using RFT implemented in SurfStat
  use detected significant voxels in alogistic regression clasification
- **SparseClassification**

Each method returns – classification labels $\tilde{y}$ and sparse regions $\tilde{\mathbf{x}}$ that are compared with ground truth y and x.

For each data A we computed classification results using **3 folds cross-validation** > data divided in 3 groups, 2 used for training and one for testing – all 3 combinations
Optimal params for each method are determined using cross-validation for each dataset A. Only results with optimal params are considered.

Reported measures

- Sparseness of detected regions sp
- Classification results (y vs $\tilde{y}$) : accuracy, sensitivity, specificity, AUC
- Accuracy and stability of regions ($\mathbf{x}$ vs $\tilde{\mathbf{x}}$) :
  dice score ; dice overlap for the 3 folds $\tilde{\mathbf{x}}_1$ $\tilde{\mathbf{x}}_2$ $\tilde{\mathbf{x}}_3$
- Significance of regions in a t-test : use mean data in sparse regions x in a t-test p-val

Experiments

Compare **3 methods** :

- **Eigenanatomy [Avants et al.]**
  unsupervised detection of sparse regions using sparse PCA
  use projection of data onto eigenvectors in a logistic regression classification method
- **SurfStat [Worsley et al.]**
  compute significant regions from using RFT implemented in SurfStat
  use detected significant voxels in alogistic regression clasification
- **SparseClassification**

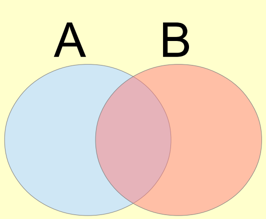
Each method returns – classification labels $\tilde{y}$ and sparse regions $\tilde{\mathbf{x}}$ that are compared with ground truth y and x.

For each data A we computed classification results using **3 folds cross-validation** > data divided in 3 groups, 2 used for training and one for testing – all 3 combinations
Optimal params for each method are determined using cross-validation for each dataset A. Only results with optimal params are considered.

Reported measures

- Sparseness of detected regions sp
- Classification results (y vs $\tilde{y}$) : accuracy, s
- Accuracy and stability of regions ($\mathbf{x}$ vs $\tilde{\mathbf{x}}$) :
  dice score ; dice overlap for the 3 folds $\tilde{\mathbf{x}}_1$ $\tilde{\mathbf{x}}$
- Significance of regions in a t-test : use mea

Dice Score:



$$\frac{2(A \cap B)}{(|A|+|B|)}$$

Synthetic data

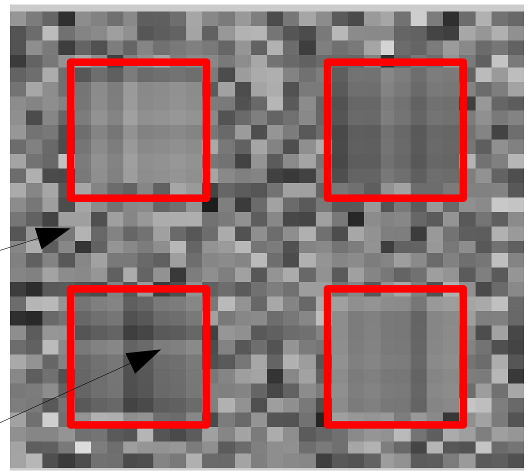
a_i:
32 x 32 x 8

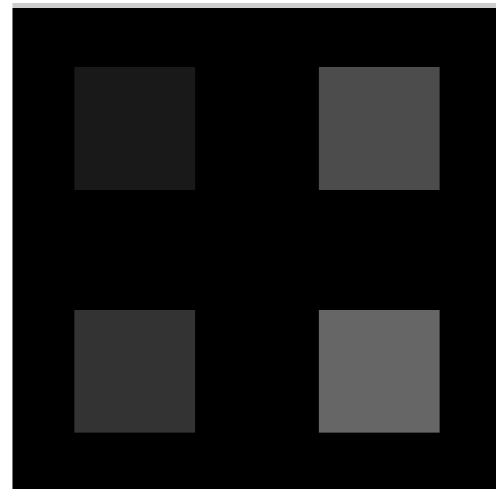
> a_i ~N(0,1)

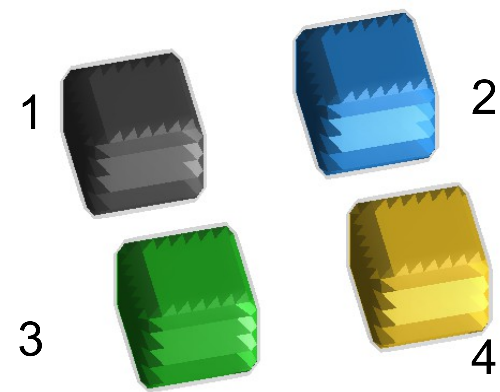
> 4 blocks of 8x8x4
Where data follows a Multinomial Normal Distribution with parameter ρ $N(0,\Sigma_\rho)$
Add coherence between voxels

<u>3 values for ρ</u> [0.1, 0.5, 0.9]
Bigger ρ = stronger coherence

100 data samples (subjects) in the matrix A







x:
Same structure as a_i

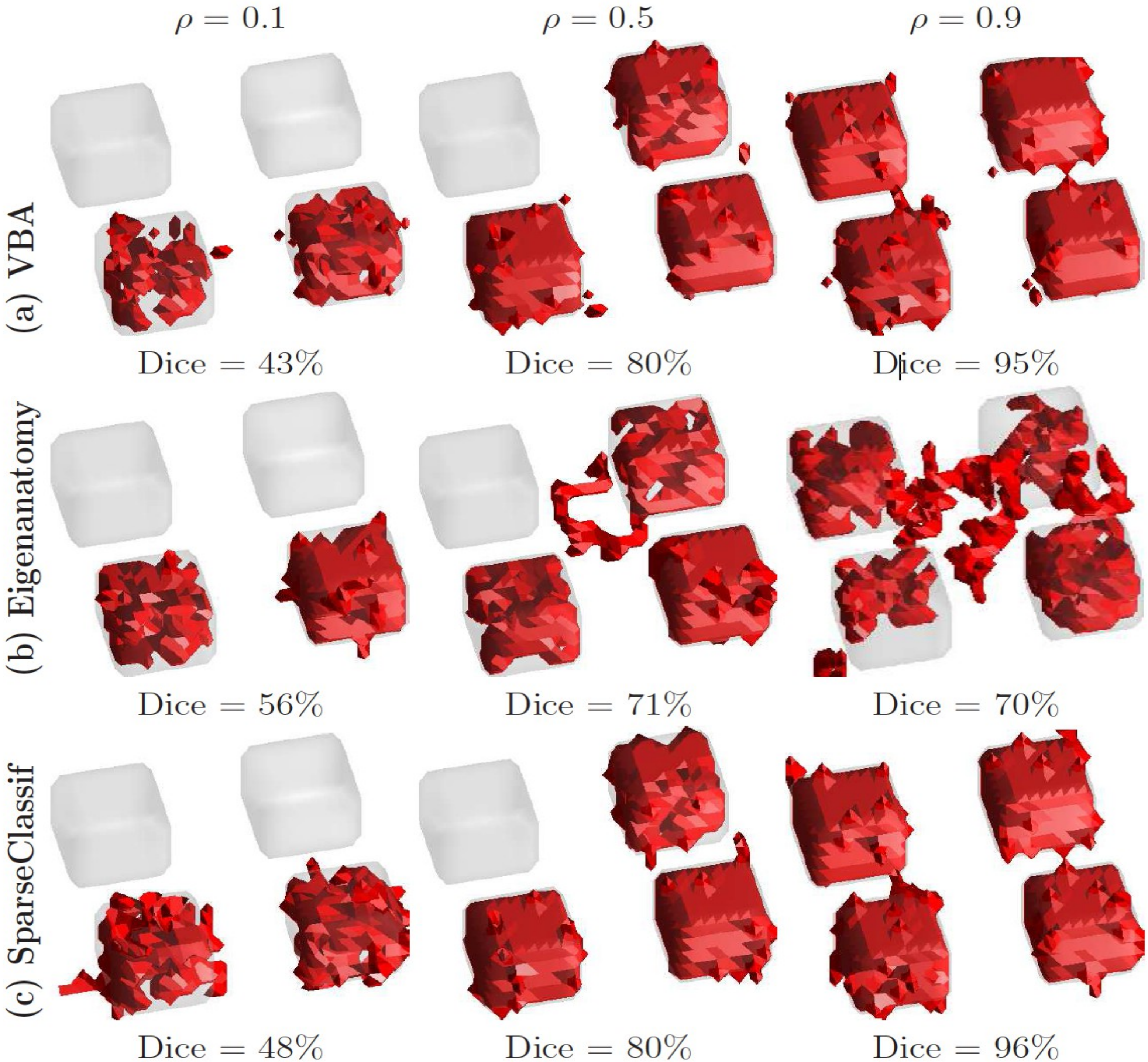
Each block has a different strength [0.1 0.2 0.3 0.4]

Will determine how much the signal in this region contributes to the data label y_i

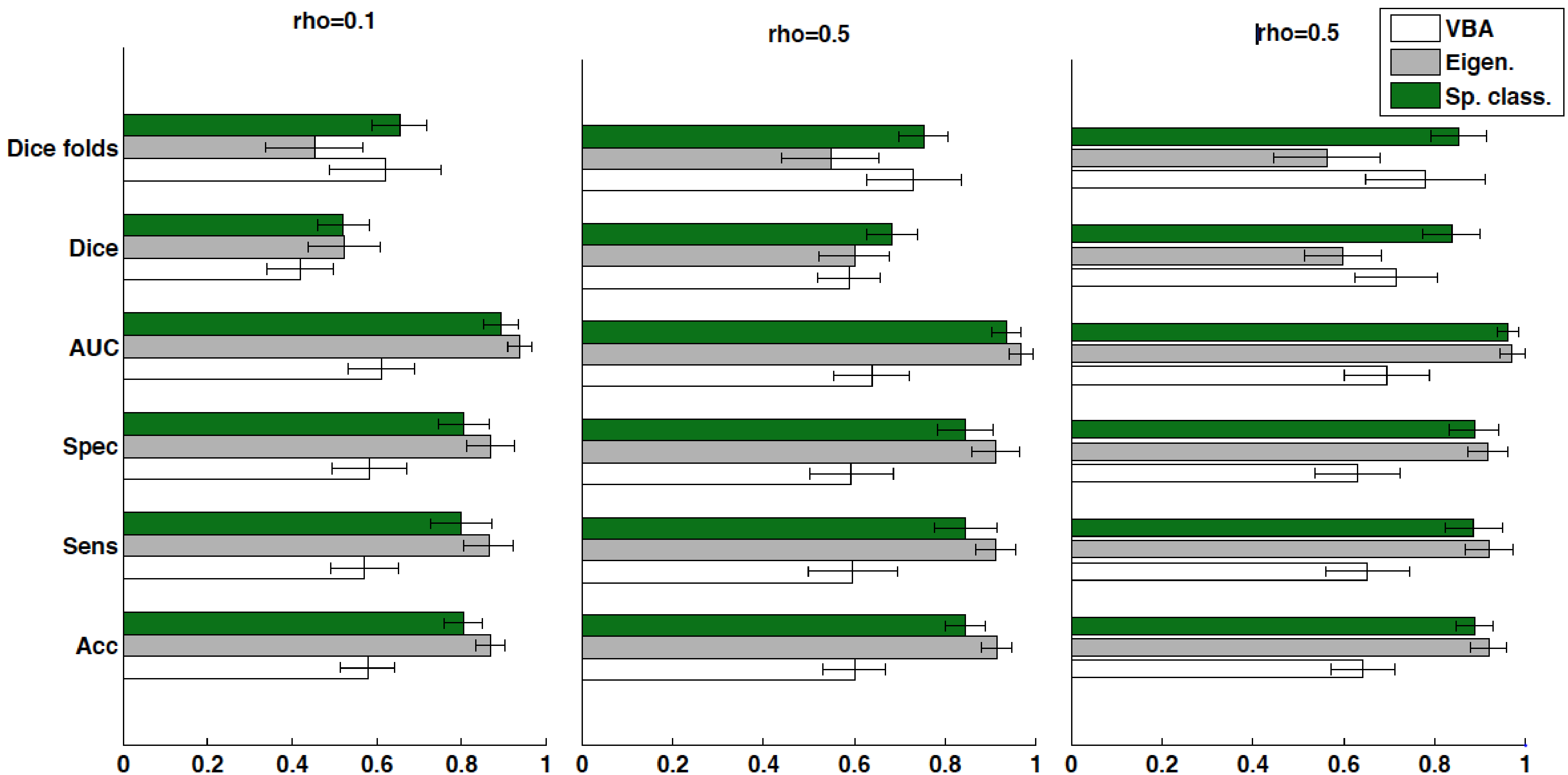
y:
Calculated using the value of the logistic regression distribution $p(y|a_i)$

+1/-1 labels using the Bernoulli distribution B(1,p)

$$p(y_i|\mathbf{a}_i,\mathbf{x},b) = \frac{1}{1+\exp(-y_i(\mathbf{a}_i\mathbf{x}+b))}$$



$\rho = 0.1$
$\rho = 0.5$
$\rho = 0.9$
(a) VBA
Dice = 43%
Dice = 80%
Dice = 95%
(b) Eigenanatomy
Dice = 56%
Dice = 71%
Dice = 70%
(c) SparseClassif
Dice = 48%
Dice = 80%
Dice = 96%



- **Accuracy of regions:** sparse classification is best in determining stable and accurate regions of difference
- **Classification accuracy**: Eigenanatomy slightly better results

Real MRI data

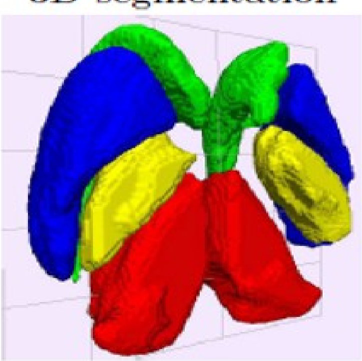
MS study : Investigating the role of iron and atrophy in MS

37 **RRMS** patients (6 males) and 37 matched controls
Age: RRMS 35.63 (std 9.2)
Controls 35.69 (std 9.0) pval .97

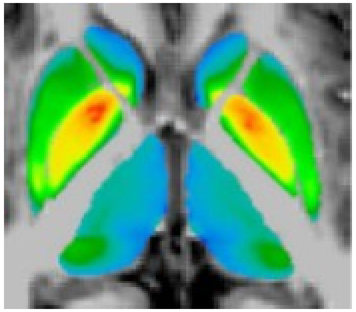
MRI data at 4.7T : T1w, R2*, QSM (284 x 222 x 84 at .9 x .9 x 2 mm)
All data is normalized to an in-house unbiased template (nonlinear registration on T1w and QSM using ANTs)

Target regions : **4 subcortical deep GM structures**
Only points inside this mask are considered in all methods

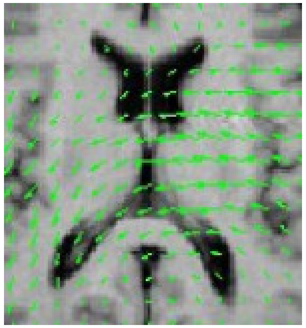
3D segmentation



"iron"



atrophy

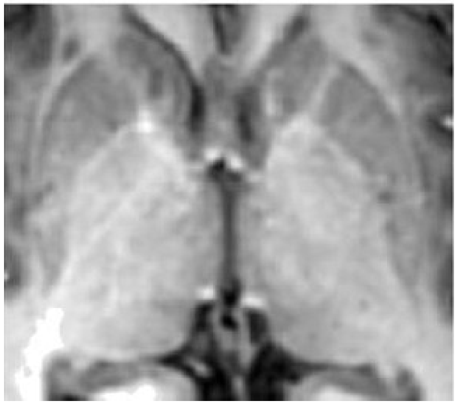


"Iron" : R2* data
Atrophy : logDetJac with respect to template

Note on data processing

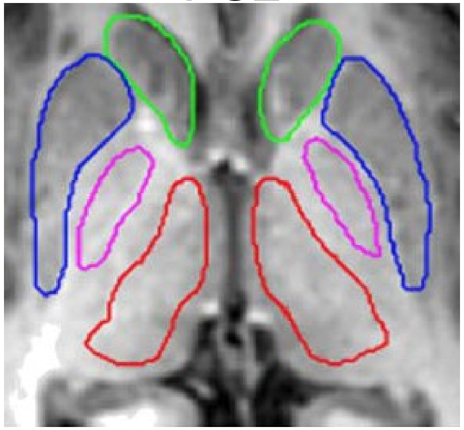
High field MRI at 4.7T

- Several imaging modalities
  low contrast T1w



> standard segmentation methods are suboptimal

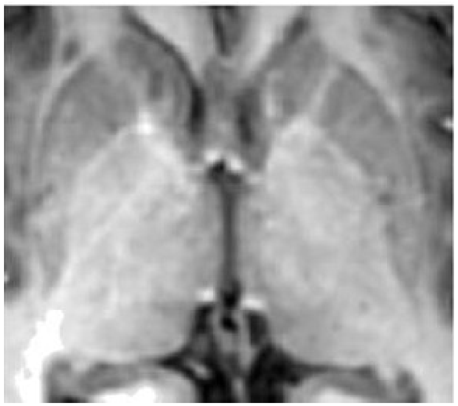
FSL



Note on data processing

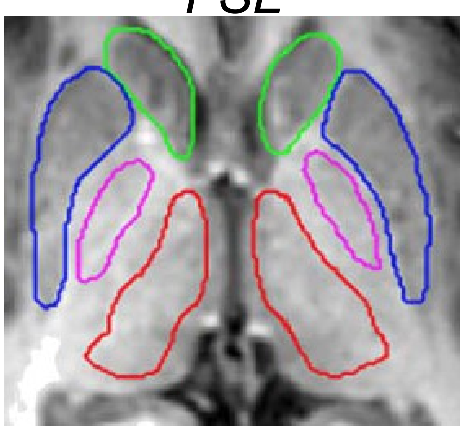
High field MRI at 4.7T

- Several imaging modalities
  low contrast T1w



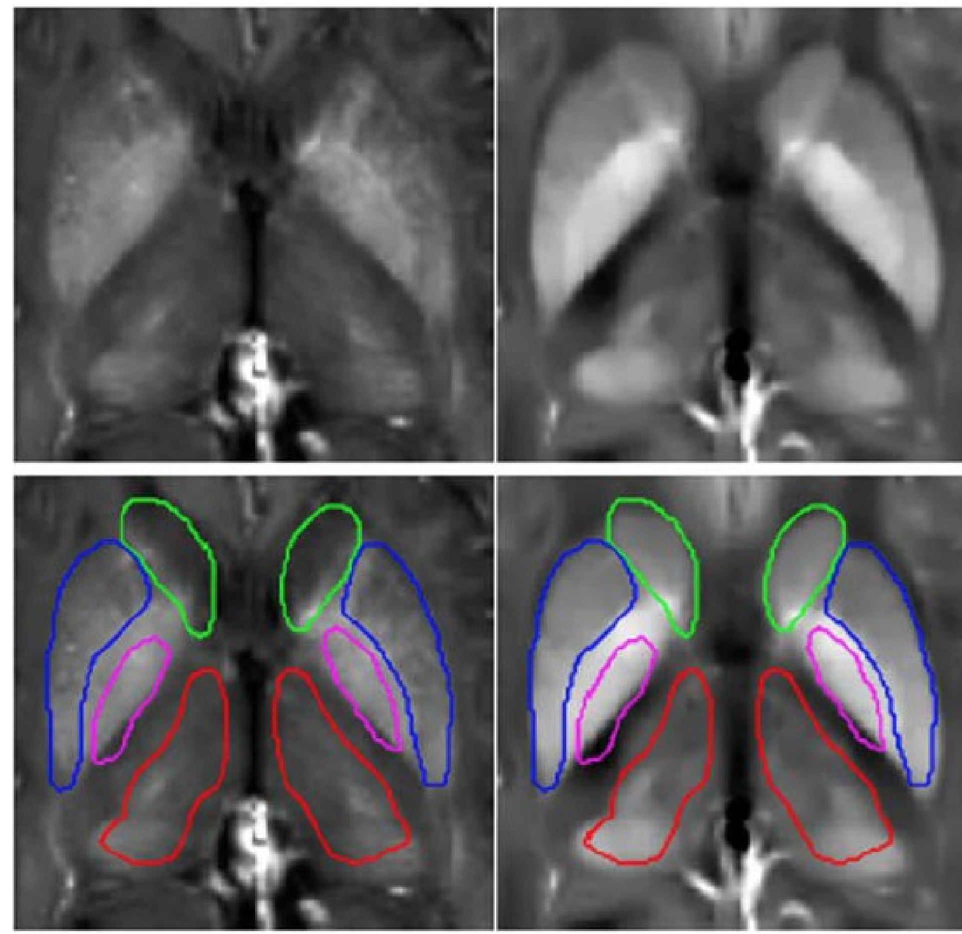
> standard segmentation methods are suboptimal

FSL



- Quantitative iron sensitive MRI :

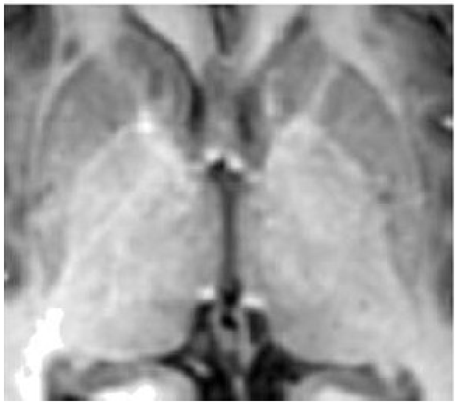
R2 mapping* *QSM*



Note on data processing

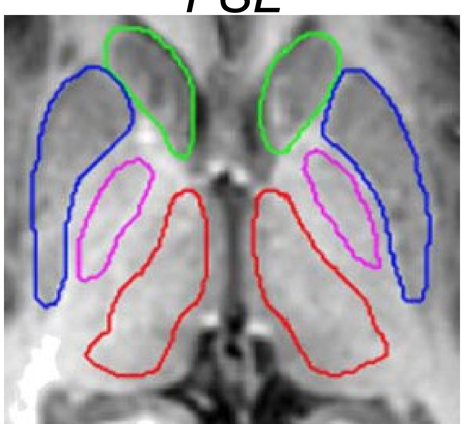
High field MRI at 4.7T

- Several imaging modalities
  low contrast T1w

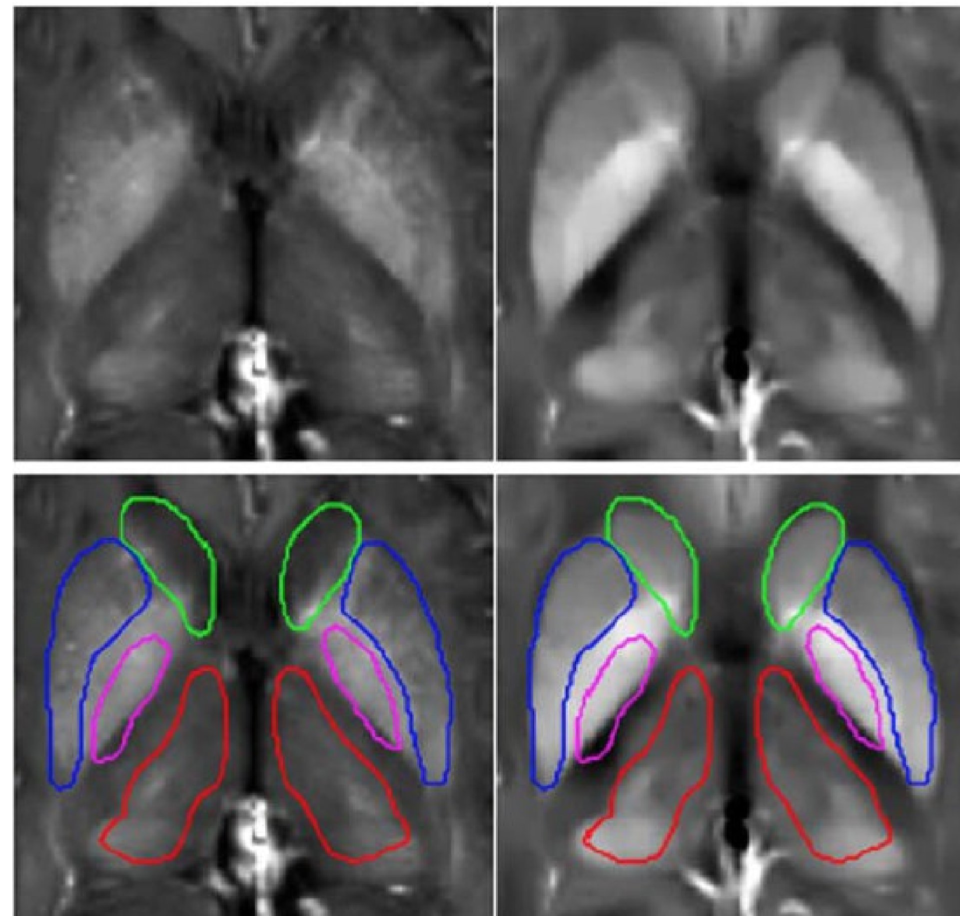


> *standard segmentation methods are suboptimal*

FSL



- Quantitative iron sensitive MRI :
  R2 mapping* *QSM*

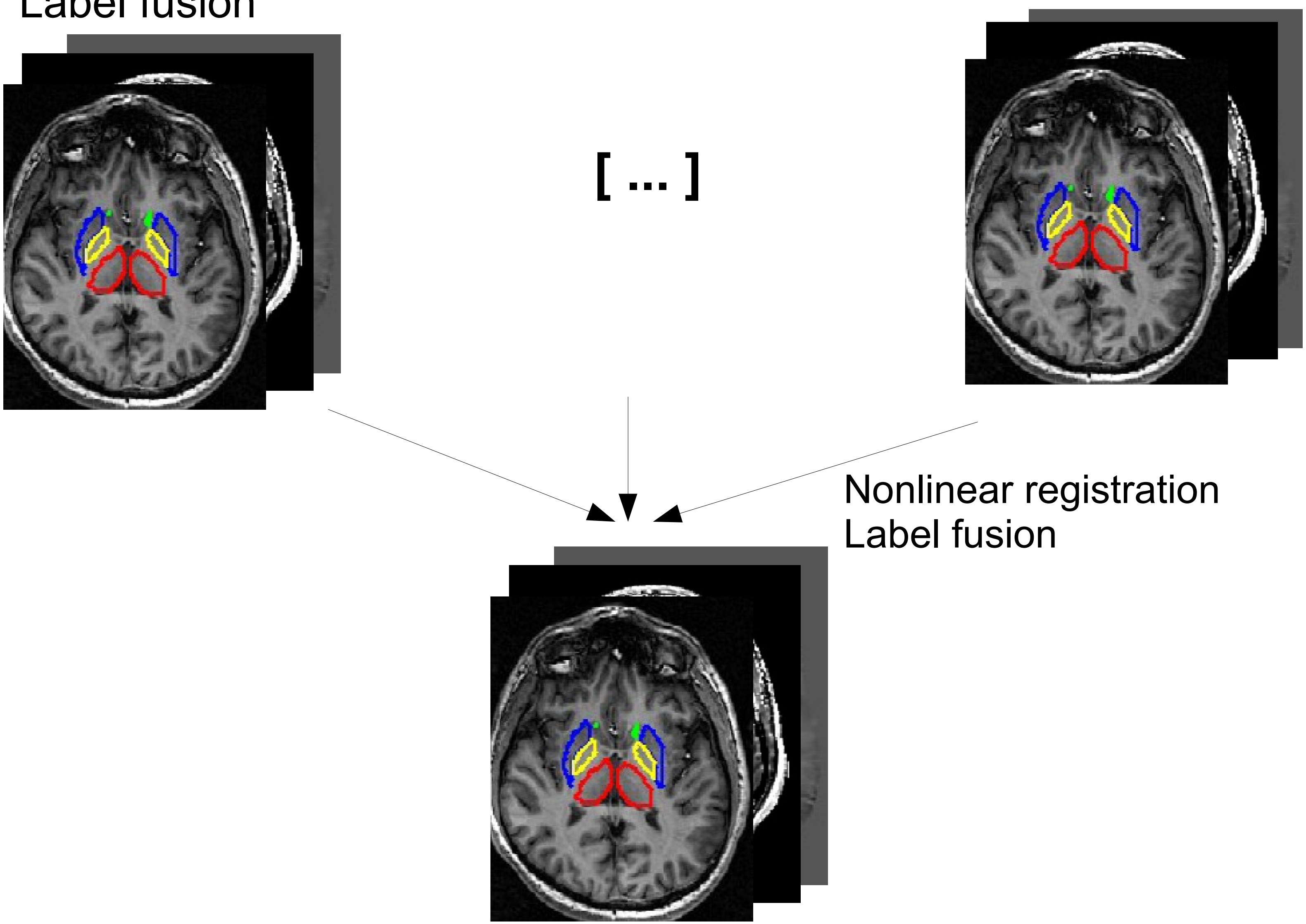


Our own pipeline

Multi-atlas segmentation

[Heckemann et al Neuriom 2005][...]

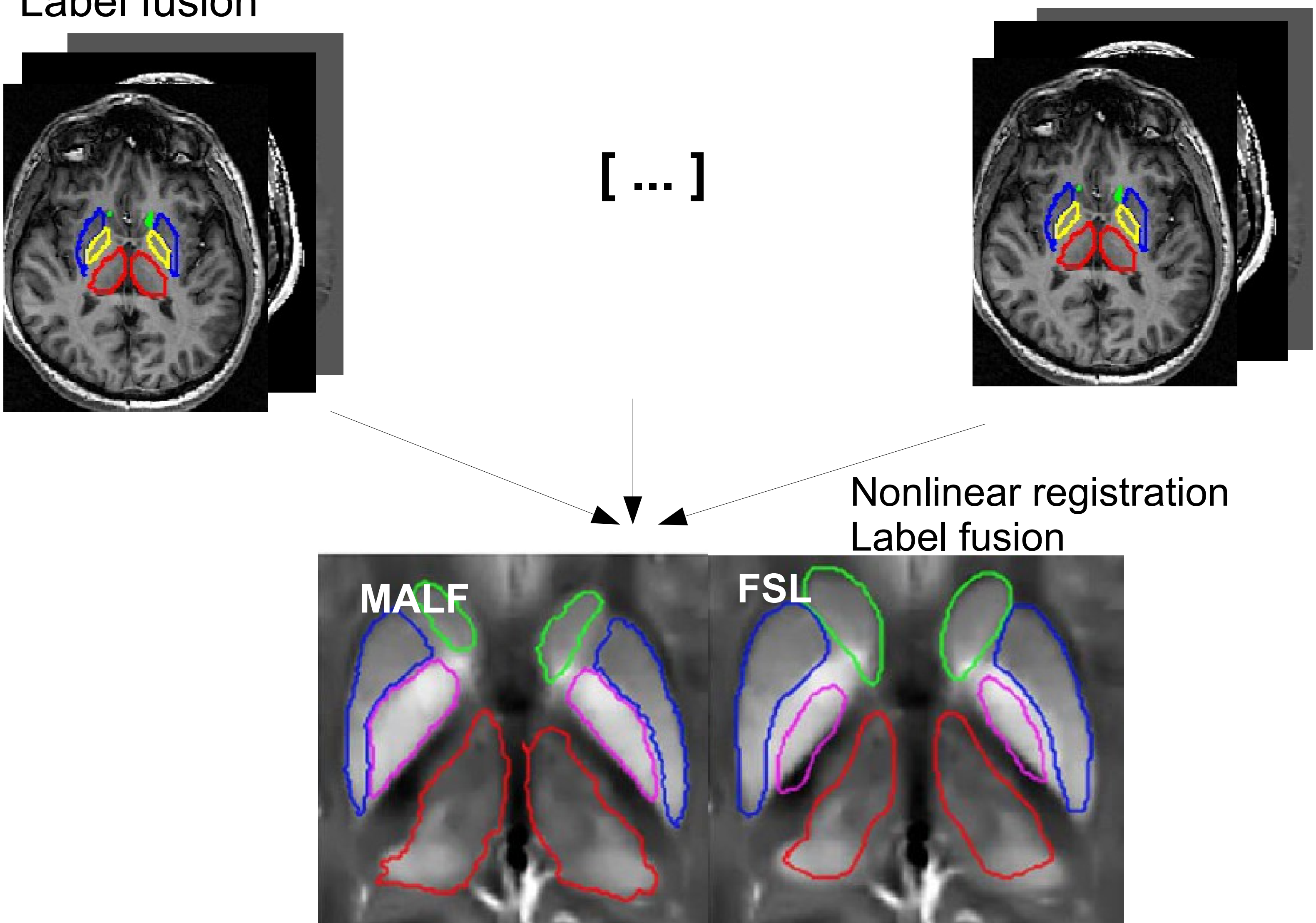
- 10 manually segmented controls
- Nonlinear registration based on T1w+R2*+QSM (SyN ANTS)
- Label fusion



Multi-atlas segmentation

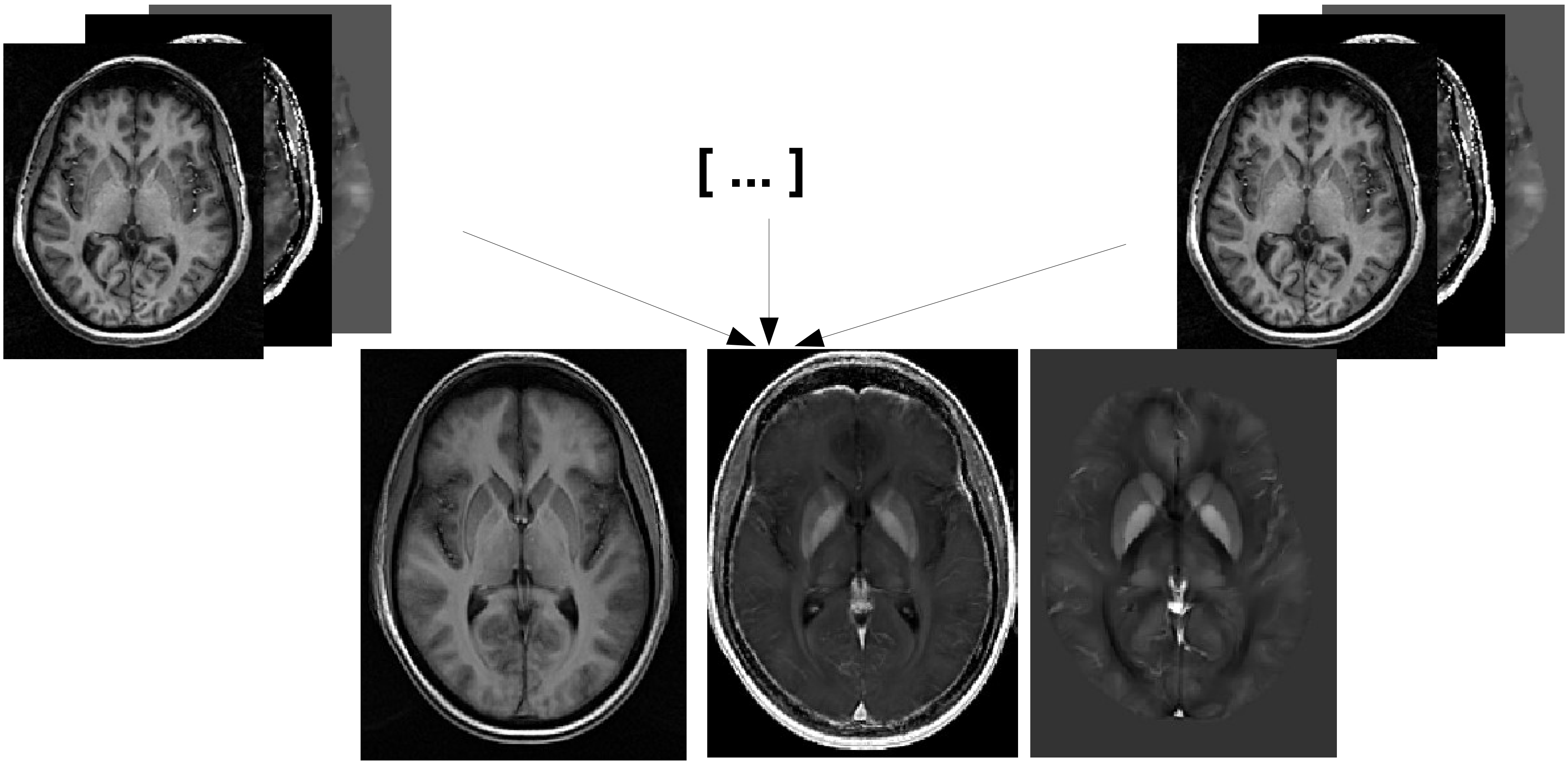
[Heckemann et al Neuriom 2005][...]

- 10 manually segmented controls
- Nonlinear registration based on T1w+R2*+QSM (SyN ANTS)
- Label fusion



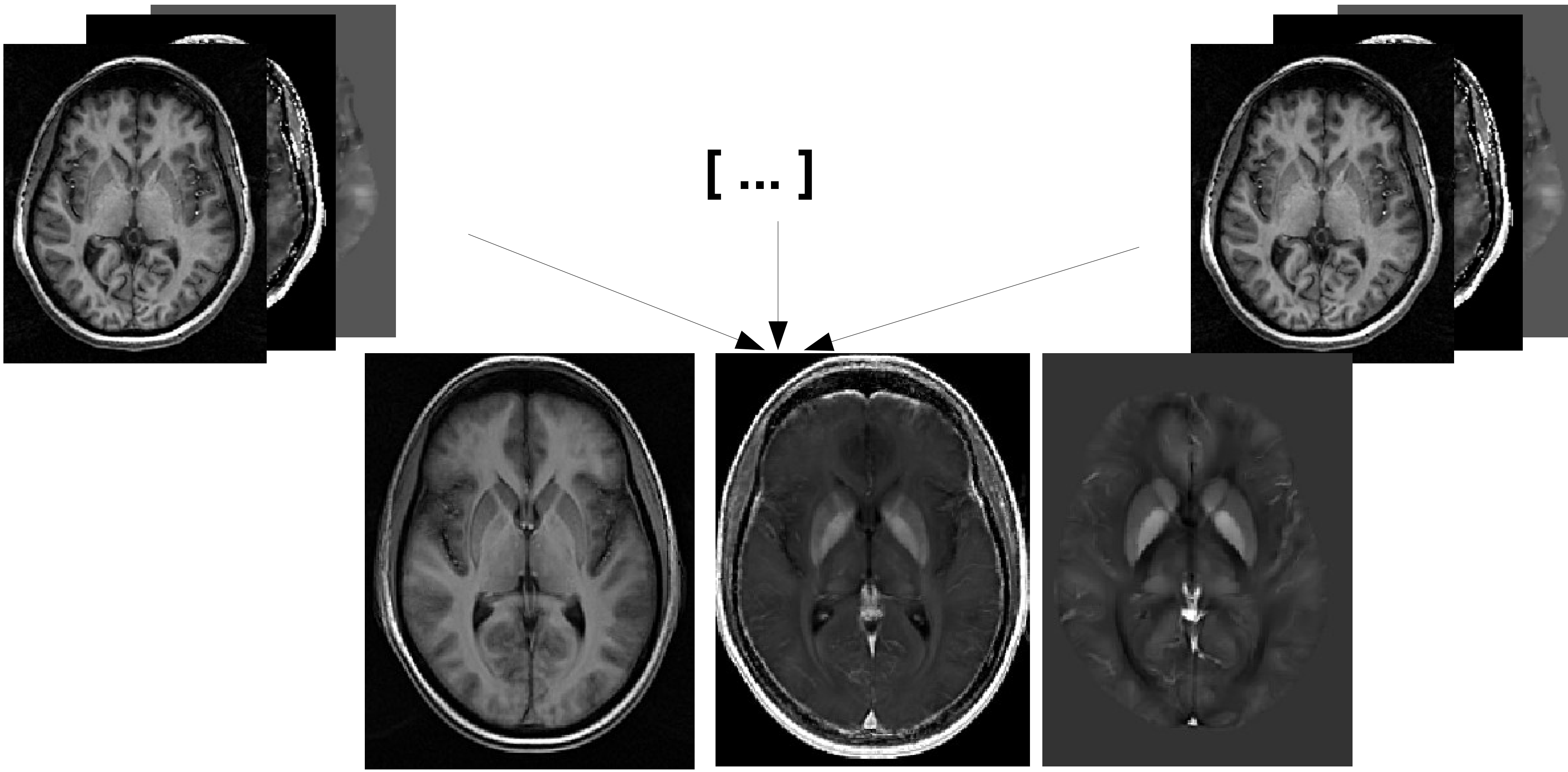
Unbiased atlas

Use the same 10 controls to create an unbiased template
Iterative method from **[Guimond et al. CVIU 2000]**



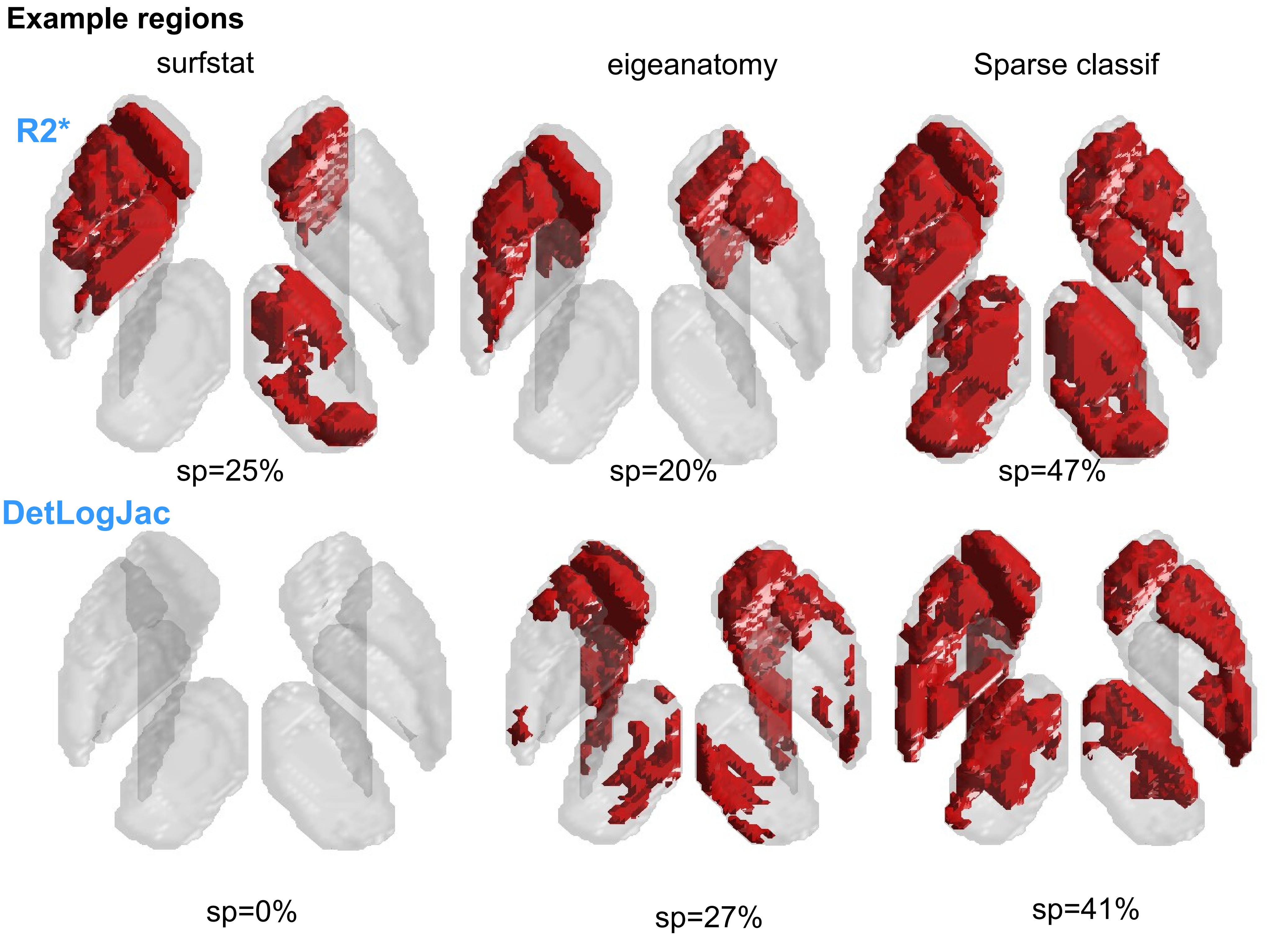
Unbiased atlas

Use the same 10 controls to create an unbiased template
Iterative method from **[Guimond et al. CVIU 2000]**



All data is normalized into the space of the atlas by nonlinear registration (SyN)
based on T1w, QSM and the MALF segmentation

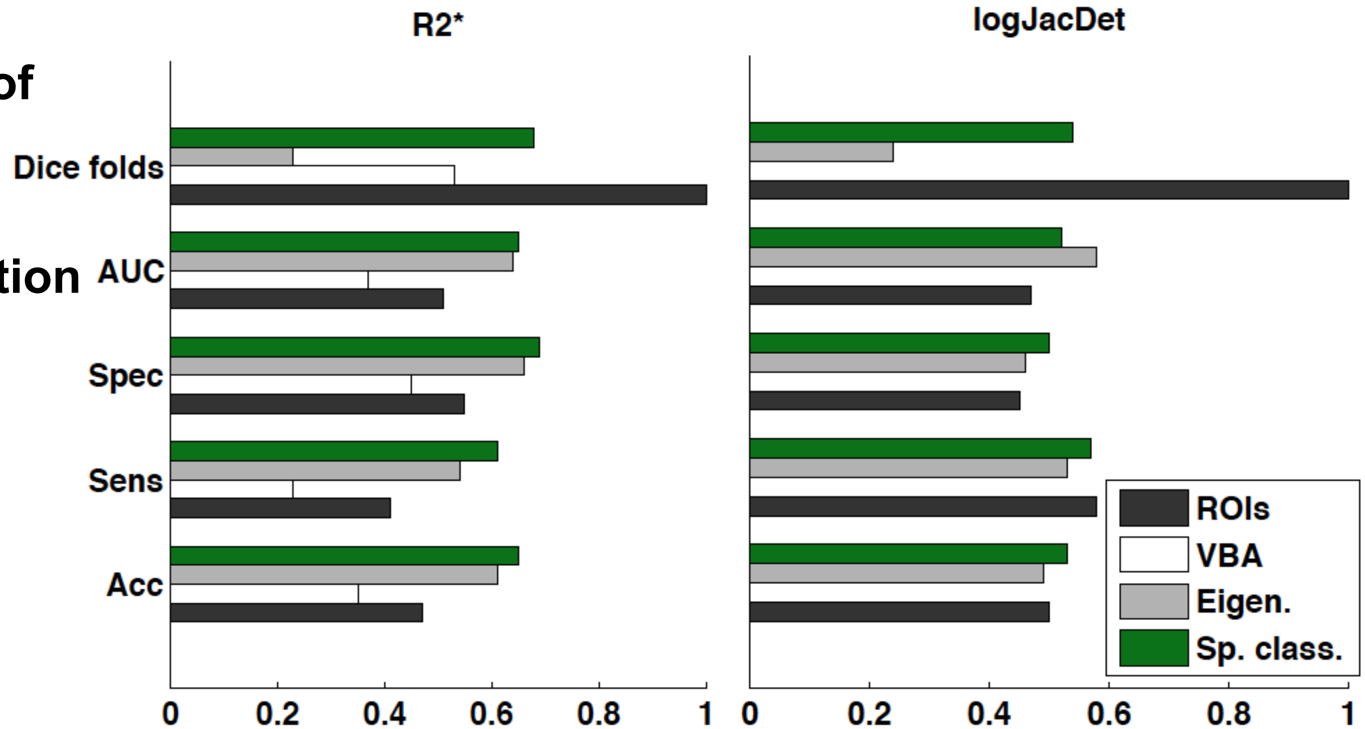
Look for regions that differentiate MS patients cs Control based on
Iron (R2*) measurements and **Atrophy (LogDetJac of deformations)**



Example regions
surfstat
eigeanatomy
Sparse classif
R2*
sp=25%
sp=20%
sp=47%
DetLogJac
sp=0%
sp=27%
sp=41%

Stability of regions

Classification accuracy



Significance of detected regions

	Group study (p-values)				
	All	Caudate	Putamen	Thalamus	GPallidus
*R2**					
ROIs	0.02	0.09	0.04	0.06	0.05
VBA	0.00001	0.0001	0.0004	0.003	0.004
eigenan	0.002	0.0005	0.002	0.02	-
sp. class.	0.0001	0.0001	0.0004	0.003	0.006
logJacDet					
ROIs	0.007	0.004	0.05	0.5	0.6
VBA	-	-	-	-	-
eigenan	0.003	0.005	0.006	0.3	0.3
sp. class.	0.00001	0.0004	0.0006	0.01	0.03

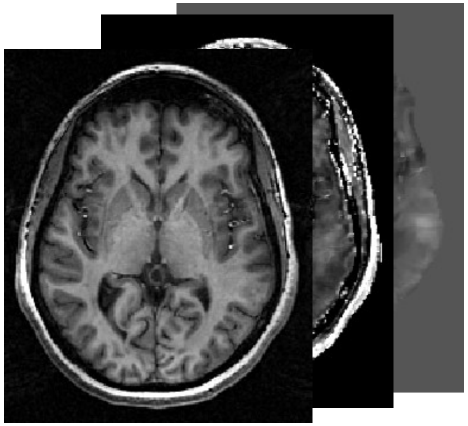
Regions : sparse classification detects most stable, accurate and significant regions

Classification accuracy:
Also best for real data

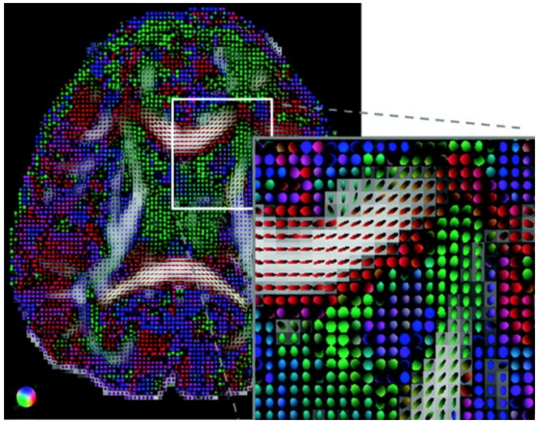
Extensions - data

Vector data

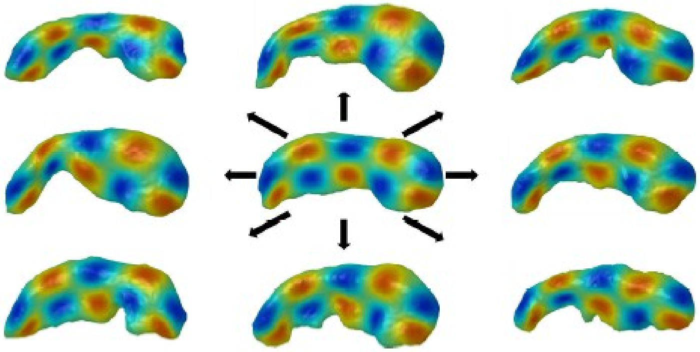
Multi modalities images



Tensor images
(log > vectors space)



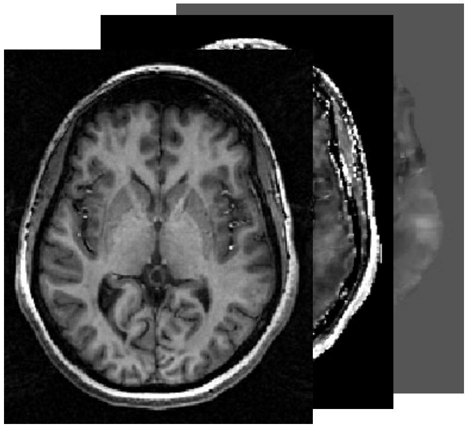
Shape data
as 3D points



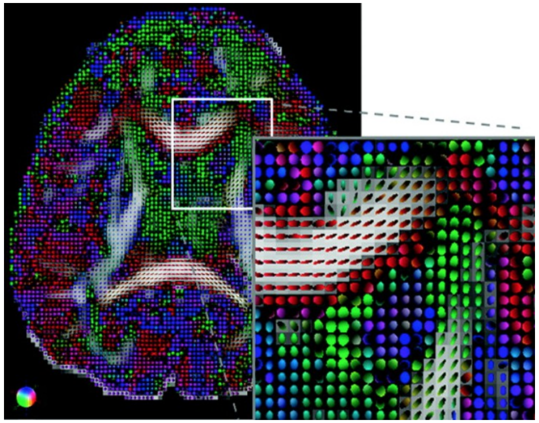
Extensions - data

Vector data

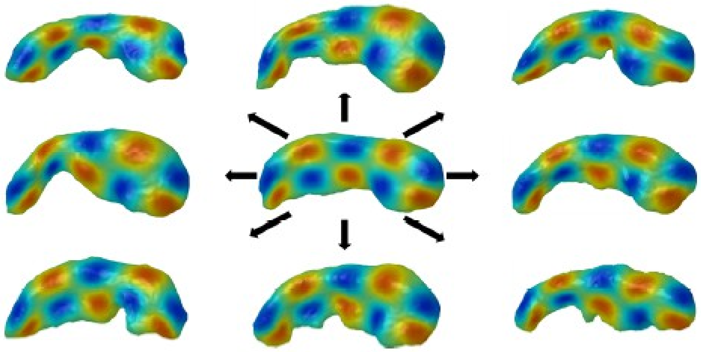
Multi modalities images

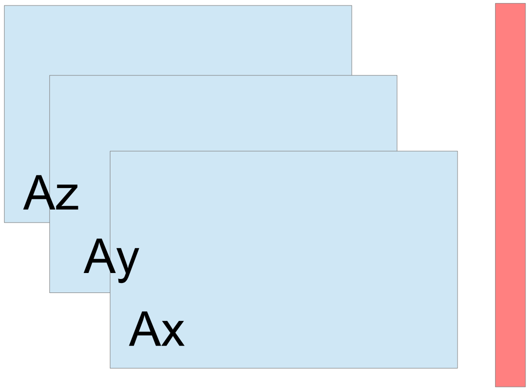


Tensor images
(logEuclidean
> vector space)



Shape data
as 3D points



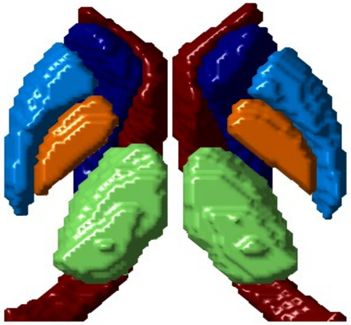


> x_x x_y x_z three sets of coefficients

> same entries should be zero > **group sparsity**

Extensions - regularization

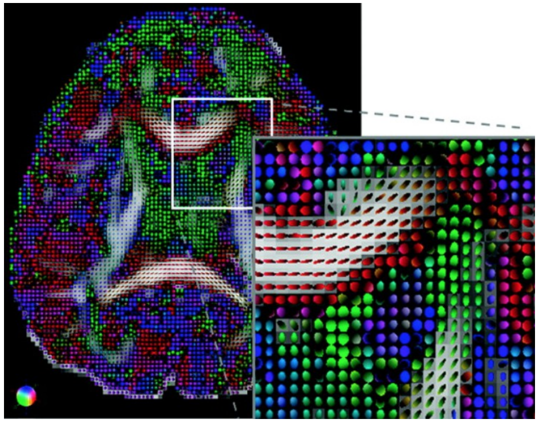
Vector data



$$\min_{\mathbf{x}} ||A\mathbf{x} - \mathbf{y}||_2^2 + \lambda_1 ||\mathbf{x}||_1 + \lambda_2 \boxed{||\nabla x||_2^2}$$

Discretized on the surface mesh

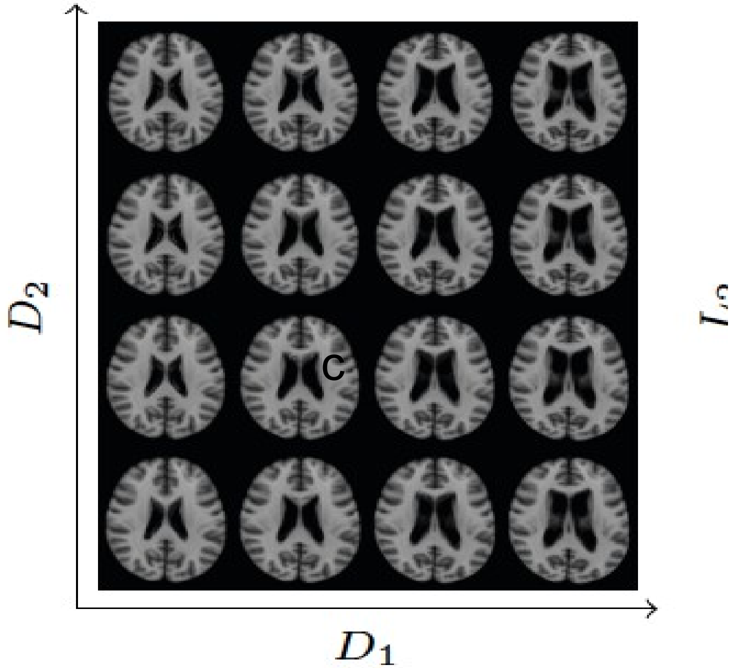
More general graph-based constraints for the image-regularization ?
ex. diffusion tensors give regularization along brain fibres



Extensions – formulation

- Other type of discriminative energies : ex SVMs
- Deep learning ? Convolutional nets ?

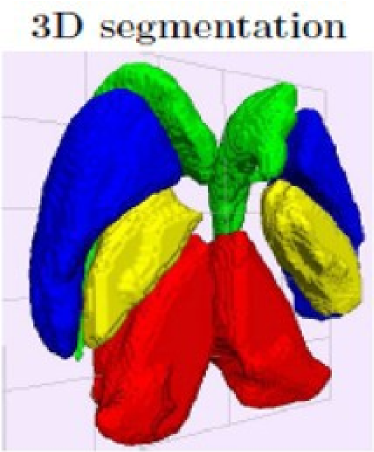
ex. **[Brosh et al MICCAI 2014]** Deep belief network used for generative learning of brain atrophy manifold
> how do we impose image-based regularization (compactness of features) Is convolutional enough ?

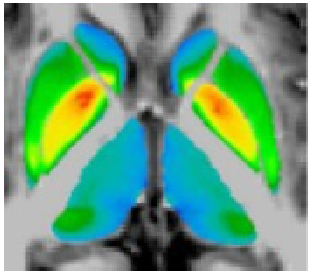


(a) Morphology manifold

MS data – big picture

Gray matter:



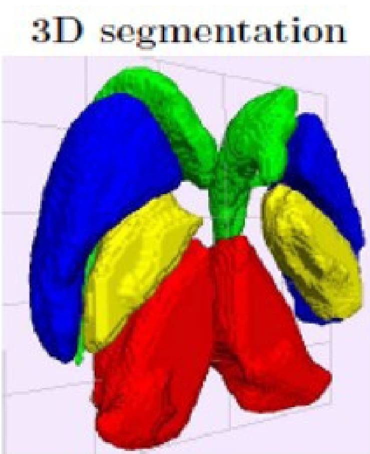


Atrophy defined on shapes

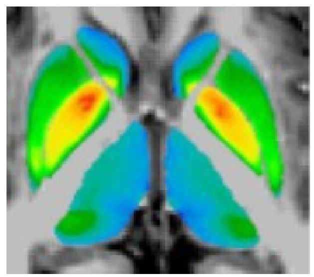
“iron” as voxel-based functional data

MS data – big picture

Gray matter:

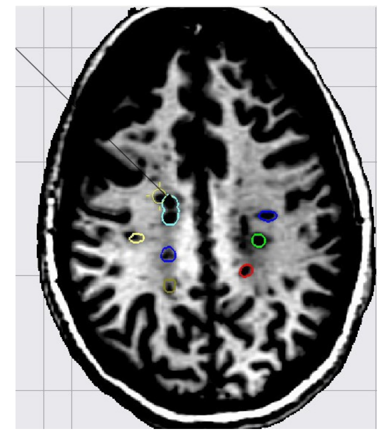


Atrophy defined on shapes

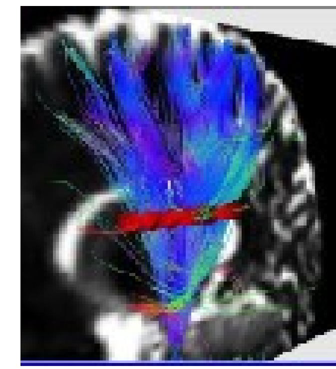


“iron” as voxel-based functional data

White matter:



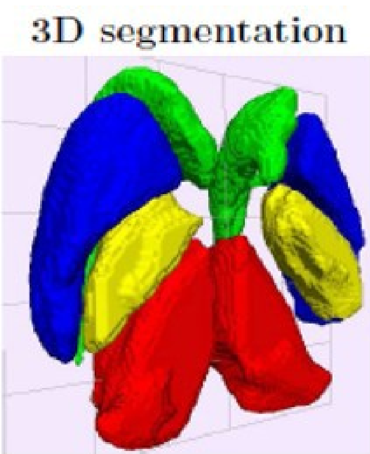
lesions



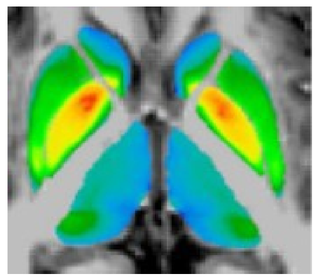
Degradation of fibres (DTI data)

MS data – big picture

Gray matter:

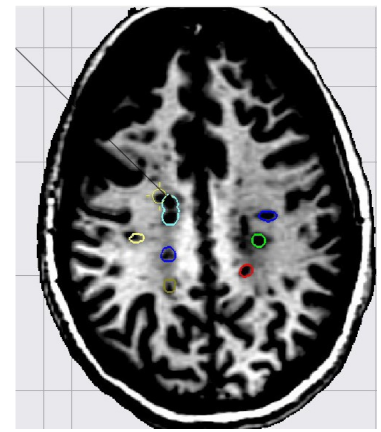


Atrophy defined on shapes

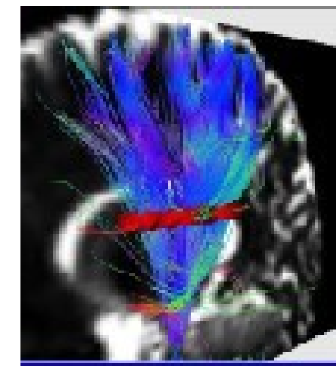


“iron” as voxel-based functional data

White matter:



lesions



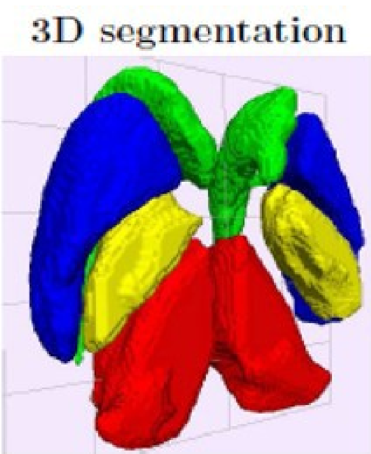
Degradation of fibres (DTI data)

How are all these parallel processes interacting ?
How are they related to disease (group study, relate to disease duration, disease severity) ?

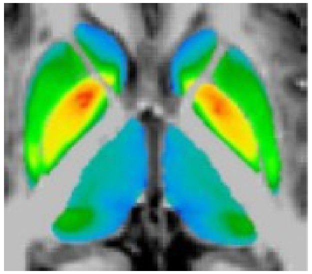
All this data is nonlinearly related to aging

MS data – big picture

Gray matter:

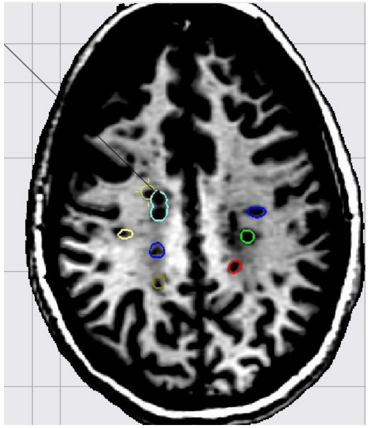


Atrophy defined on shapes

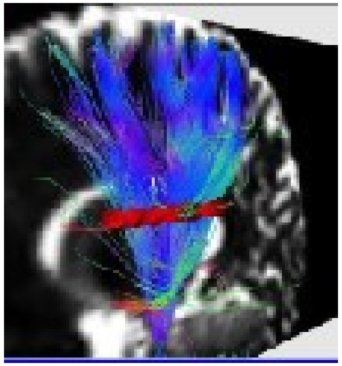


“iron” as voxel-based functional data

White matter:



lesions



Degradation of fibres (DTI data)

How are all these parallel processes interacting ?
How are they related to disease (group study, relate to disease duration, disease severity) ?

All this data is nonlinearly related to aging

THANK YOU