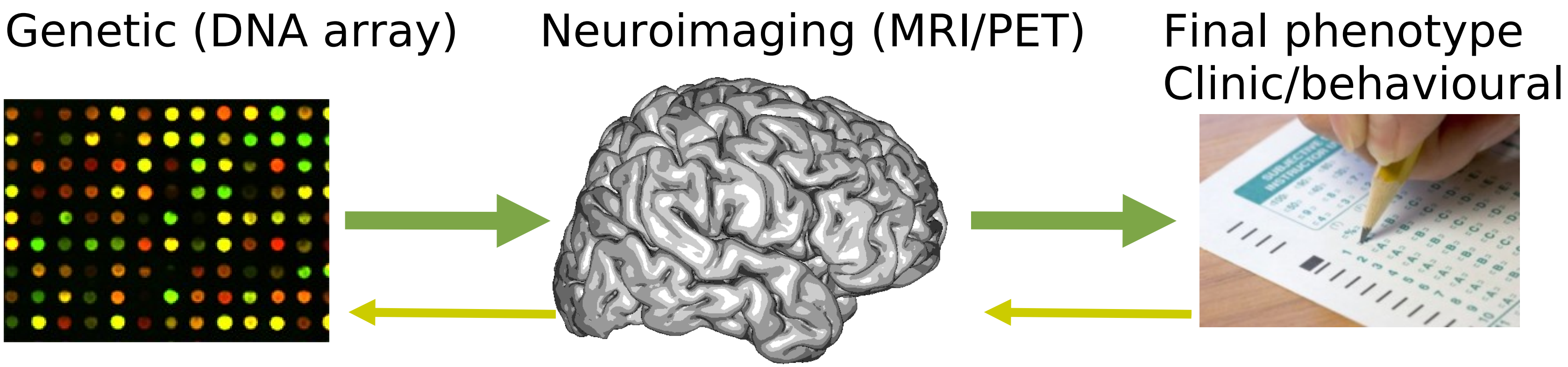
Neuroimaging as an intermediate phenotype to bridge the gap between clinic and genetic: Machine learning methods

Edouard Duchesnay

Edith Le Floch, Vincent Frouin, Bertrand Thirion, JB Poline

NeuroSpin, LNAO, CEA Saclay

Overview



Genetic (DNA array)

Neuroimaging (MRI/PET)

Final phenotype
Clinic/behavioural

Principle

Imagery as an **intermediate phenotype**

Example of exploratory data strategy

Step 1: Neuroimaging to clinic

→ identify neuroimaging-based intermediary phenotypes

Step 2: Genetic to neuroimaging-based intermediary phenotypes

→ identify genetic markers (link to pathways etc.)

Problem

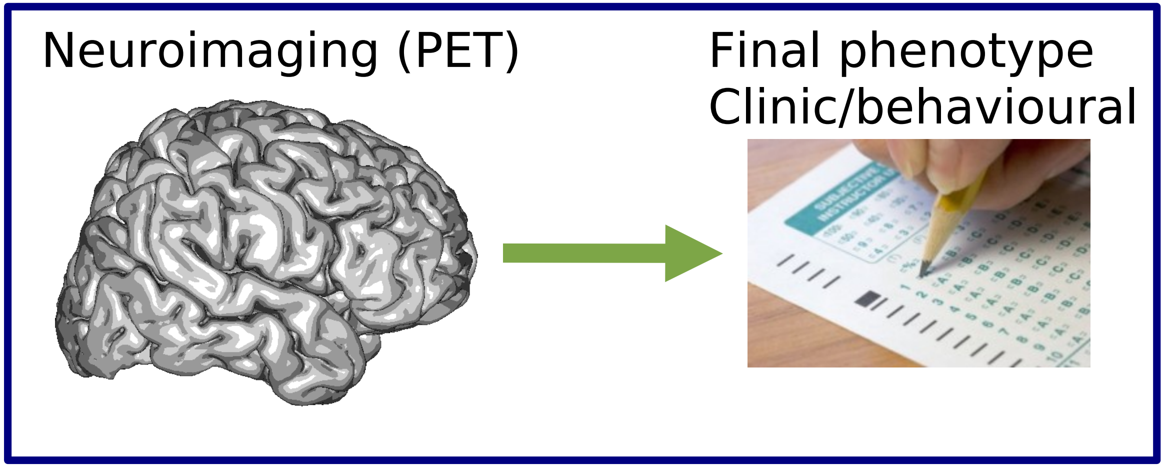
Vast amount of biological measurements:

– Neuroimaging (~10^6 voxels), DNA array (~10^6 SNPs)

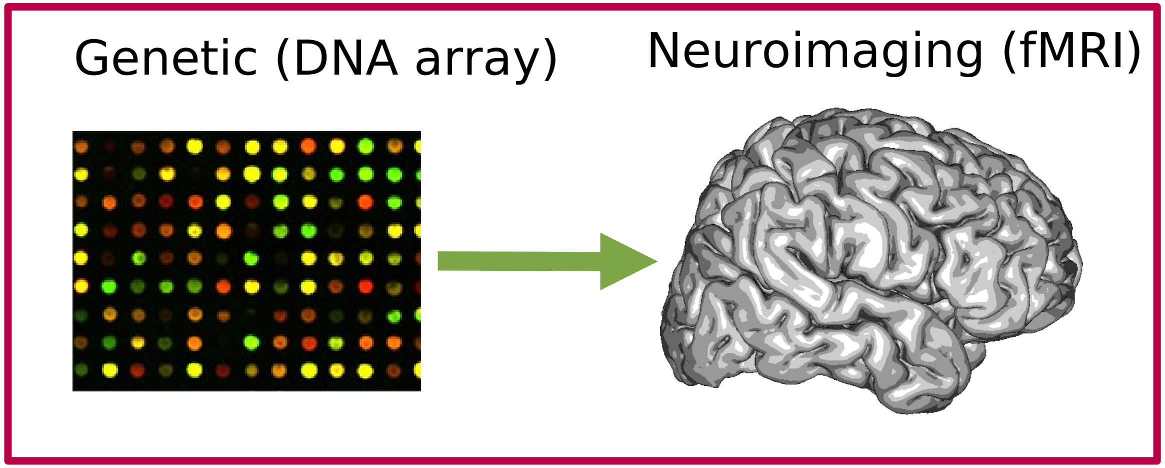
→ Spurious associations between: *genetic x imaging x clinic*

→ Poor reproducibility (multiple comparison/over-fitting issues)

Outline

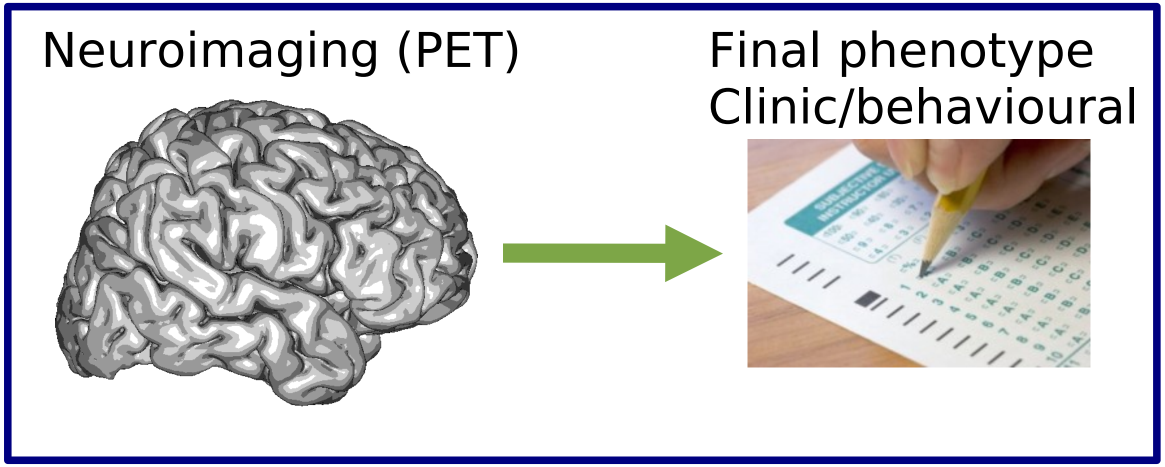


Application to autism



Application: asymetries in language processing

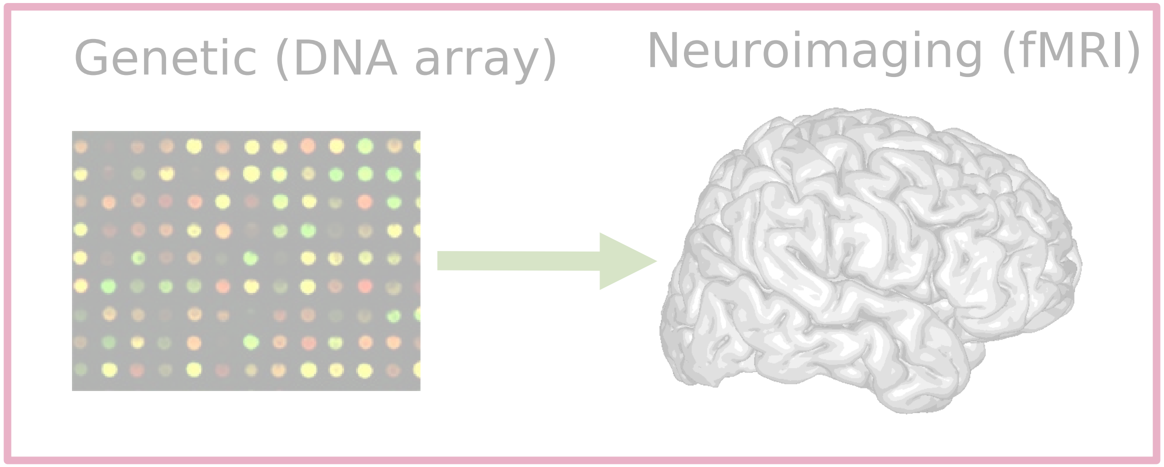
Neuroimaging to predict the clinical status



Neuroimaging (PET)

Final phenotype
Clinic/behavioural

Application to autism



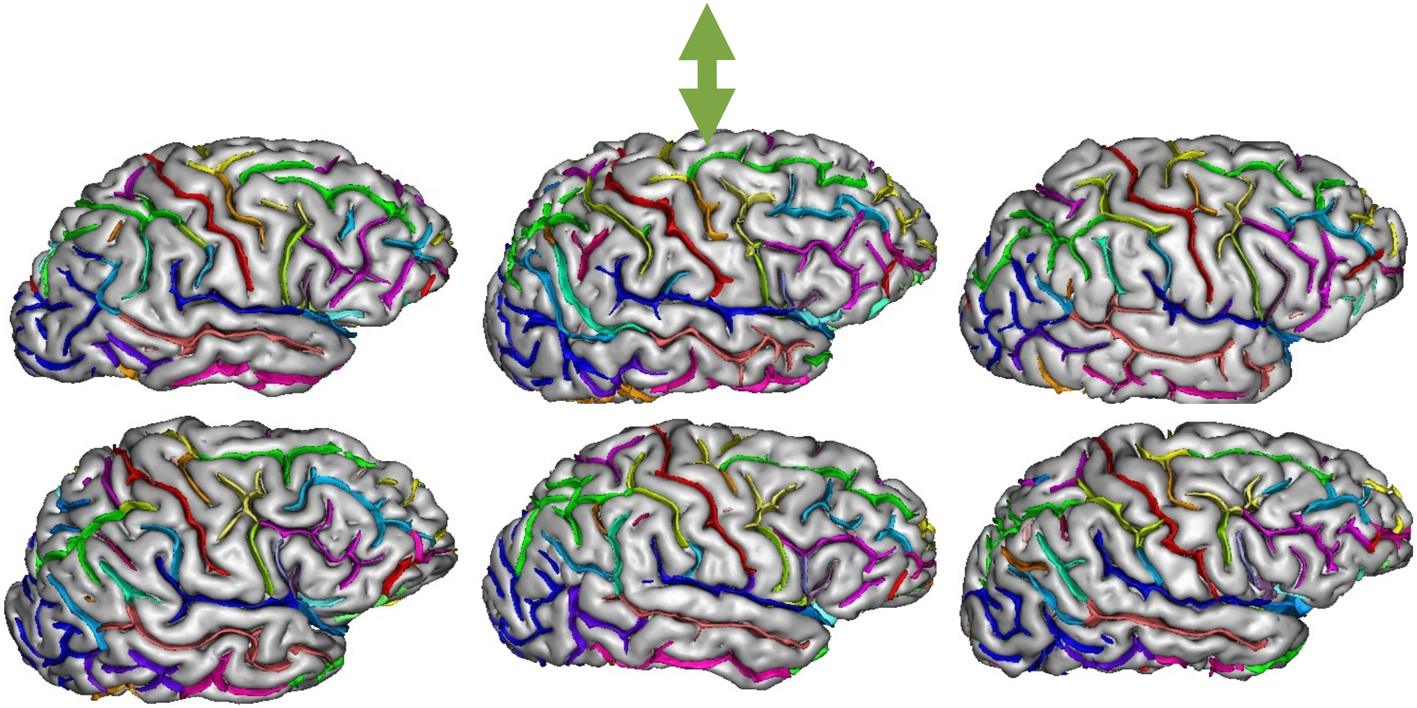
Genetic (DNA array)

Neuroimaging (fMRI)

Application: asymetries in language processing

Brains variability

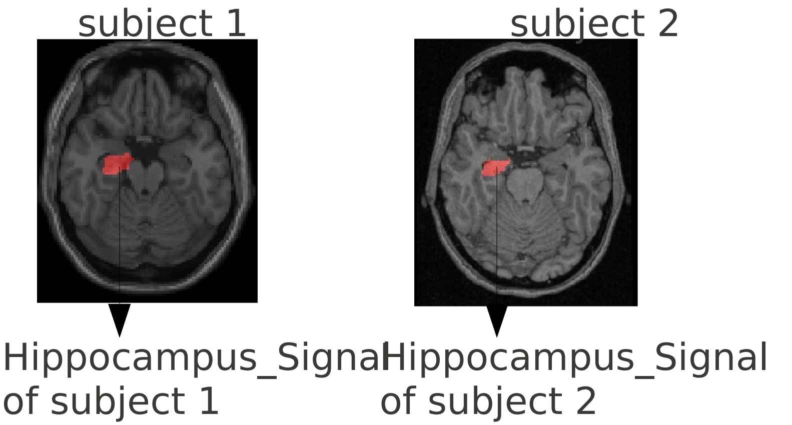
– Find the brain variability associated to a clinical trait
– How to compare brains?



Ideal scenario:
1) Neuroimaging measurements (anatomical/diffusion/functional-MRI)
2) Re-align brains: remove non-specific variability
3) Do Machine learning on the specific variability

Features extraction strategies

Volume of Interest (VOI or ROI)
- Manually defined
- Template-based


subject 1

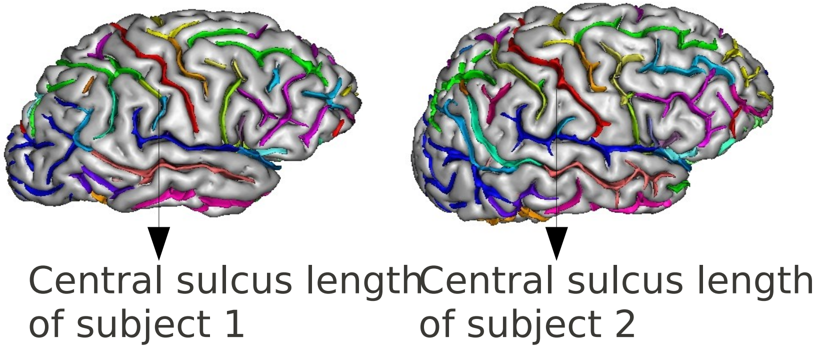
subject 2

Hippocampus_Signal of subject 1

Hippocampus_Signal of subject 2

Structure identification
- Automatically defined
- Manually defined

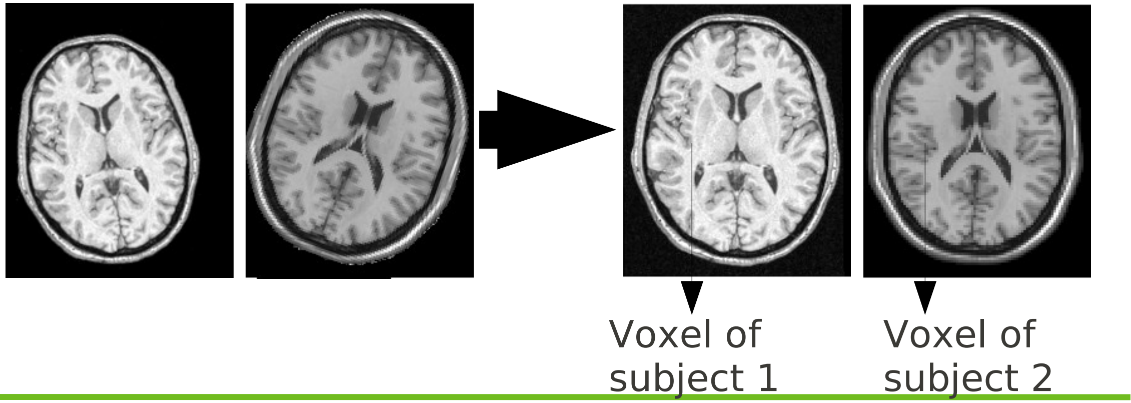
=> **Structural data**


Central sulcus length of subject 1

Central sulcus length of subject 2

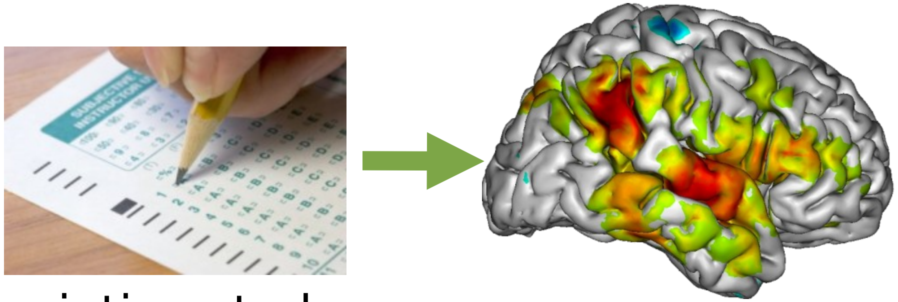
Brain warping
- Warp brain toward a common template (coordinate system)

=>**Iconic data**


Voxel of subject 1

Voxel of subject 2

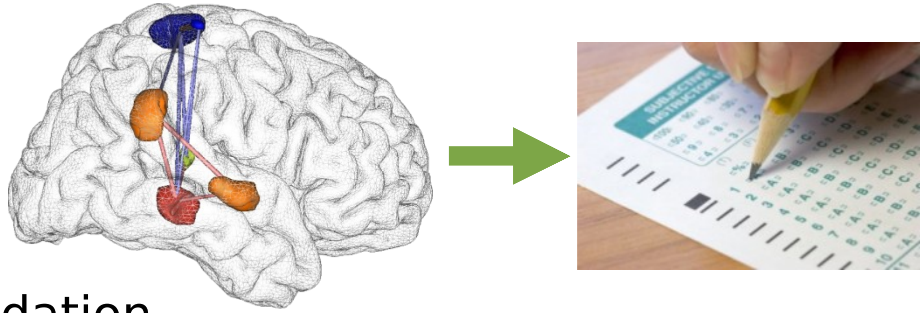
ML in neuroimaging



Classical voxel-wise analysis

Analyse each brain locus **Independently**: Measure the brain/clinic association

- Similar to (GWAS) Genome-wide association study
- **Overlook** brain-brain interaction
- **No individual** classification
- Multiple comparison issues



Multivariate classification

Find a global mapping from the brain to the clinic

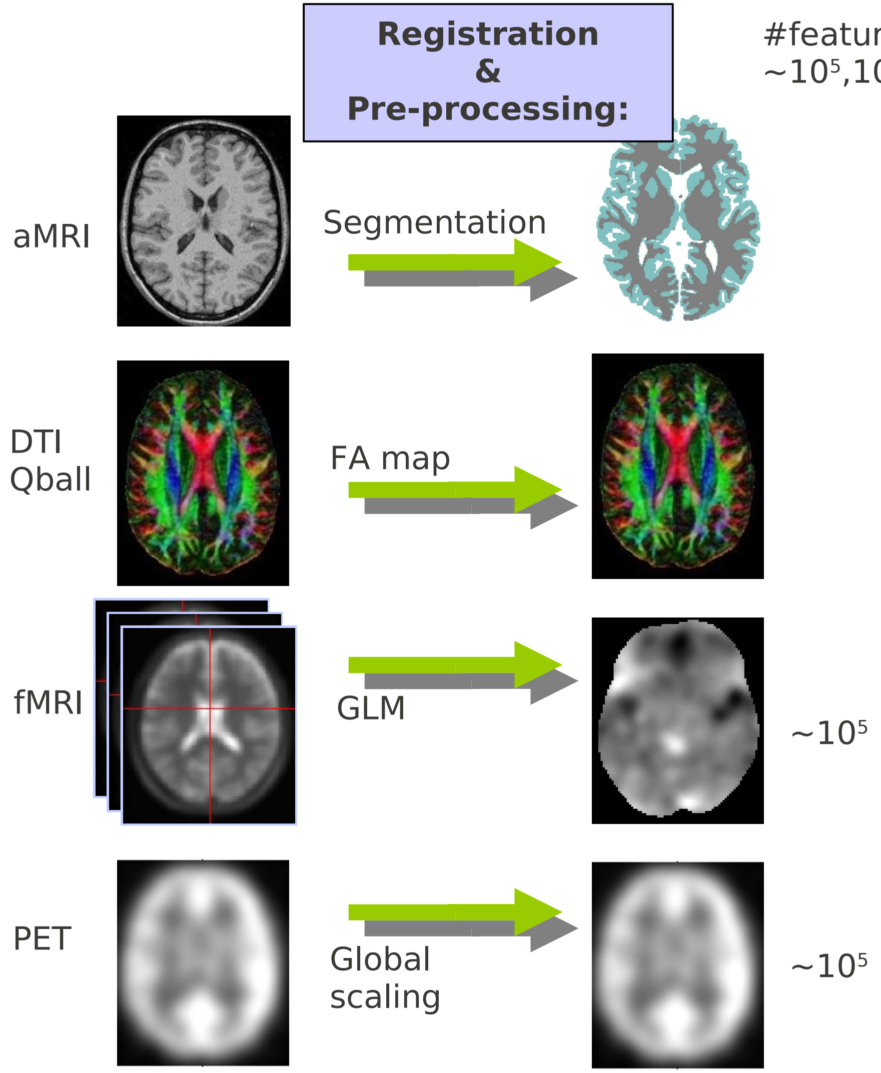
- Improved sensitivity
- Consider brain-brain interactions
- Over-fitting issue: need careful validation

Goal:

(1) **Imaging profile** that covariate with phenotype (clinical severity)
(2) Individual **computer aided diagnosis**
(3) Individual predictor of **response to treatment**

Data: Iconic / Structural

Iconic (image based) methods



Registration & Pre-processing:

#features ~10^5,10^6

aMRI

Segmentation

DTI Qball

FA map

fMRI

GLM

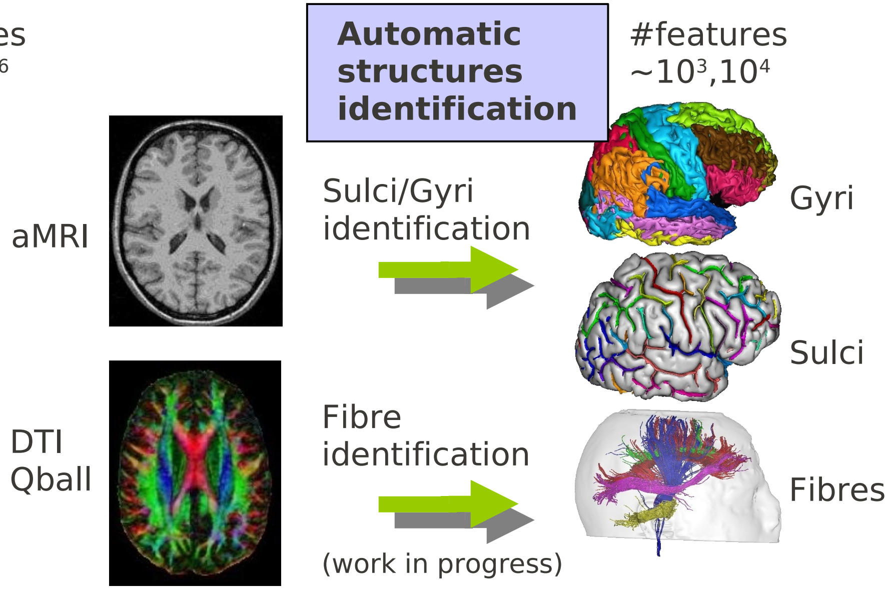
~10^5

PET

Global scaling

~10^5

Structural methods



Automatic structures identification

#features ~10^3,10^4

aMRI

Sulci/Gyri identification

Gyri

Sulci

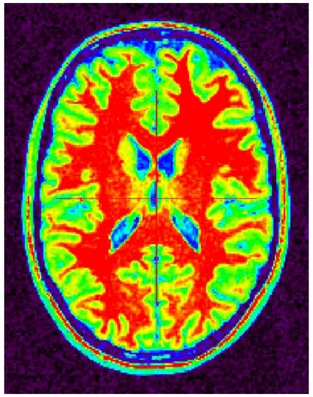
DTI Qball

Fibre identification

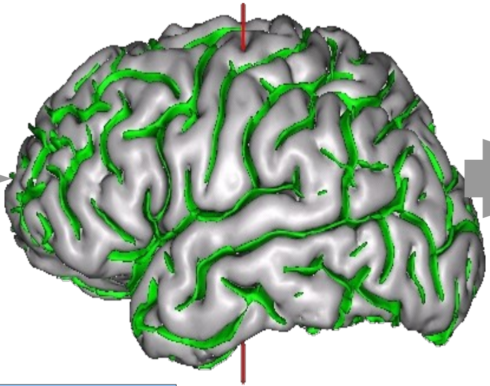
(work in progress)

Fibres

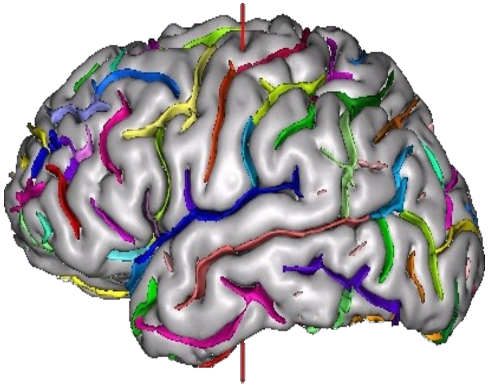
Structural Based Morphometry: summary



Grey/White segmentation

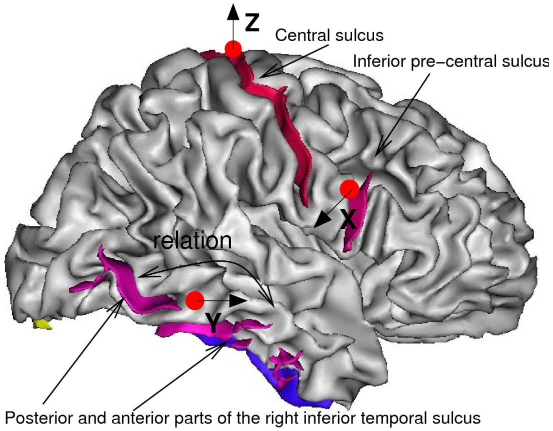


Sulci extraction



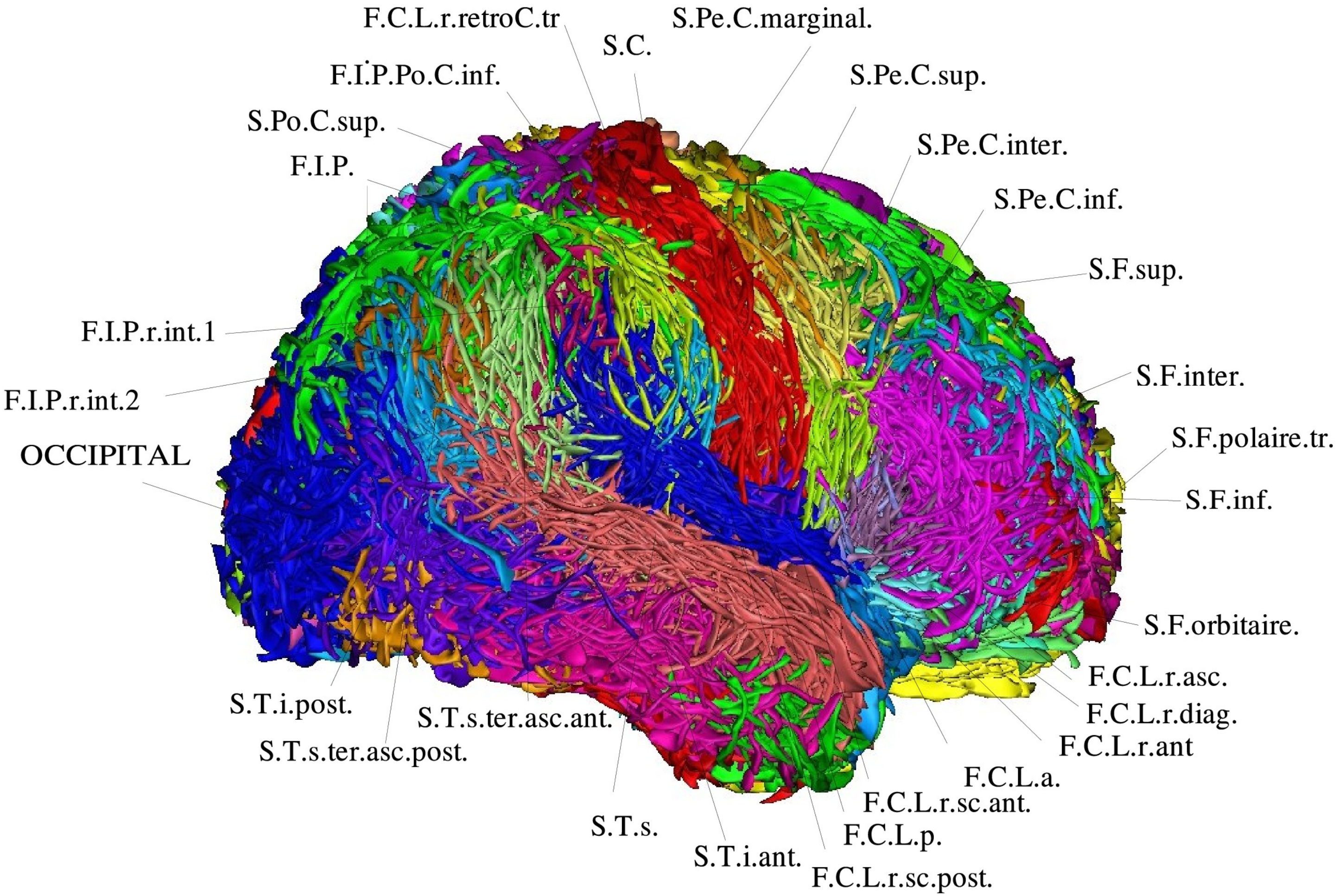
Automatic sulci identification

Machine learning


Z
Central sulcus
Inferior pre−central sulcus
X
relation
Y
Posterior and anterior parts of the right inferior temporal sulcus

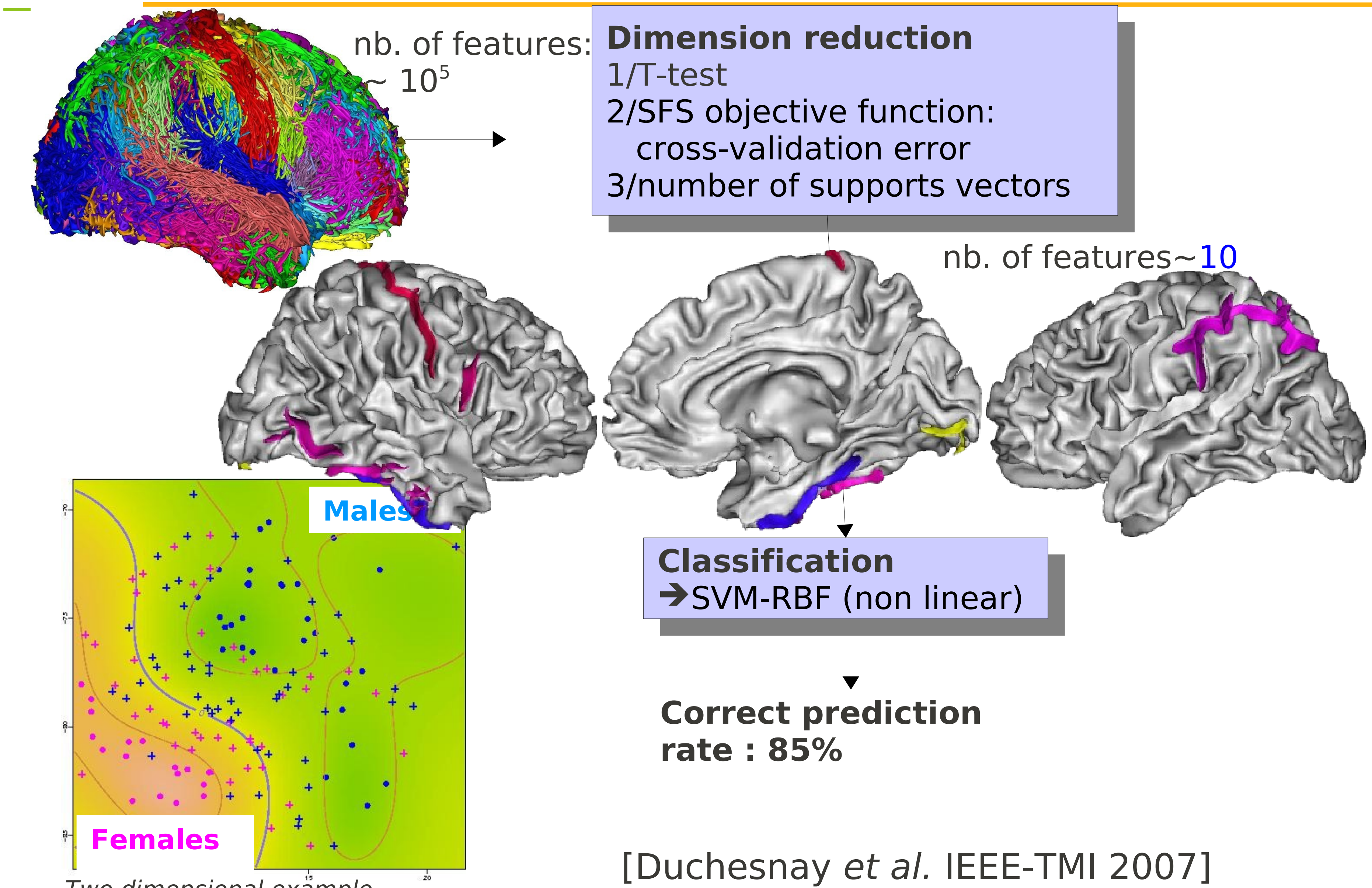
Anatomical MRI: compare sulci measurements

Superimposed sulci of 70 subjects



Compare the same anatomical structures (sulci)
across all subject without registration (just linear normalization)

Classification based on sulci: predict gender



nb. of features: ~ 10^5

Dimension reduction
1/T-test
2/SFS objective function:
cross-validation error
3/number of supports vectors

nb. of features~10

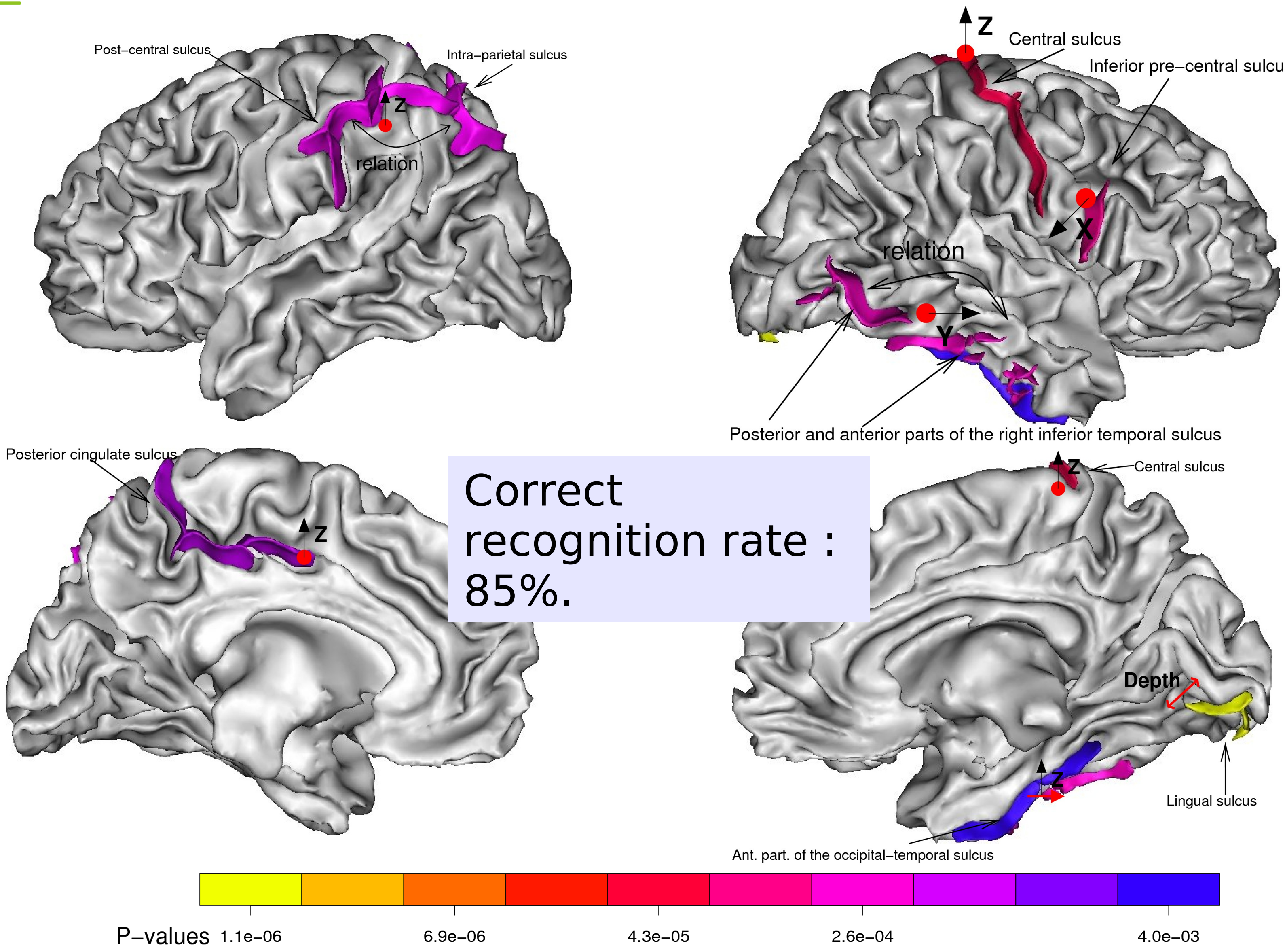
Classification
➔SVM-RBF (non linear)

Correct prediction rate : 85%

Two dimensional example

[Duchesnay *et al.* IEEE-TMI 2007]

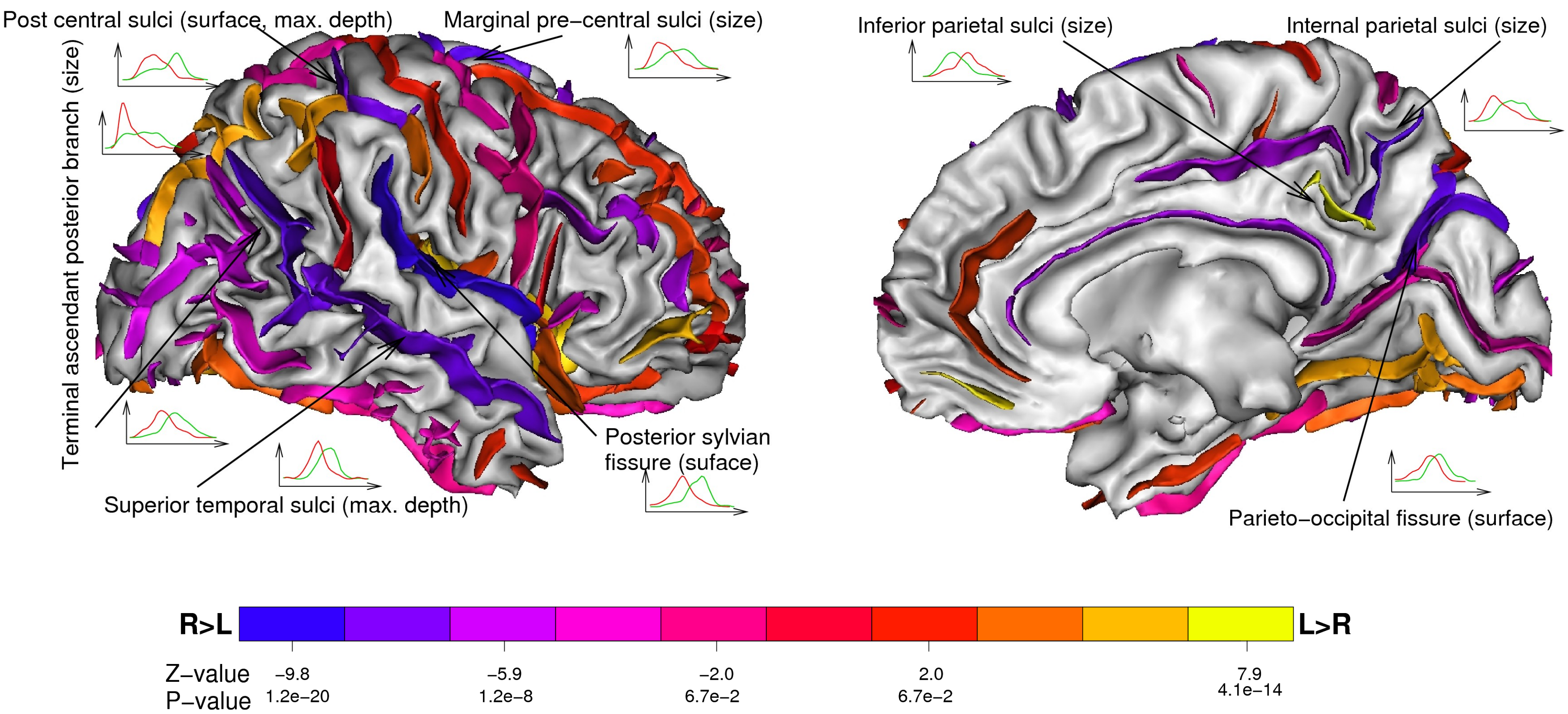
Guess the gender from sulci


Post−central sulcus
Intra−parietal sulcus
z
relation
Z
Central sulcus
Inferior pre−central sulcus
X
relation
Y
Posterior and anterior parts of the right inferior temporal sulcus
Posterior cingulate sulcus
z
Z
Central sulcus
Depth
Z
Lingual sulcus
Ant. part. of the occipital−temporal sulcus
P−values
1.1e−06
6.9e−06
4.3e−05
2.6e−04
4.0e−03

Correct recognition rate : 85%.

Guess left/right hemisphere from sulci

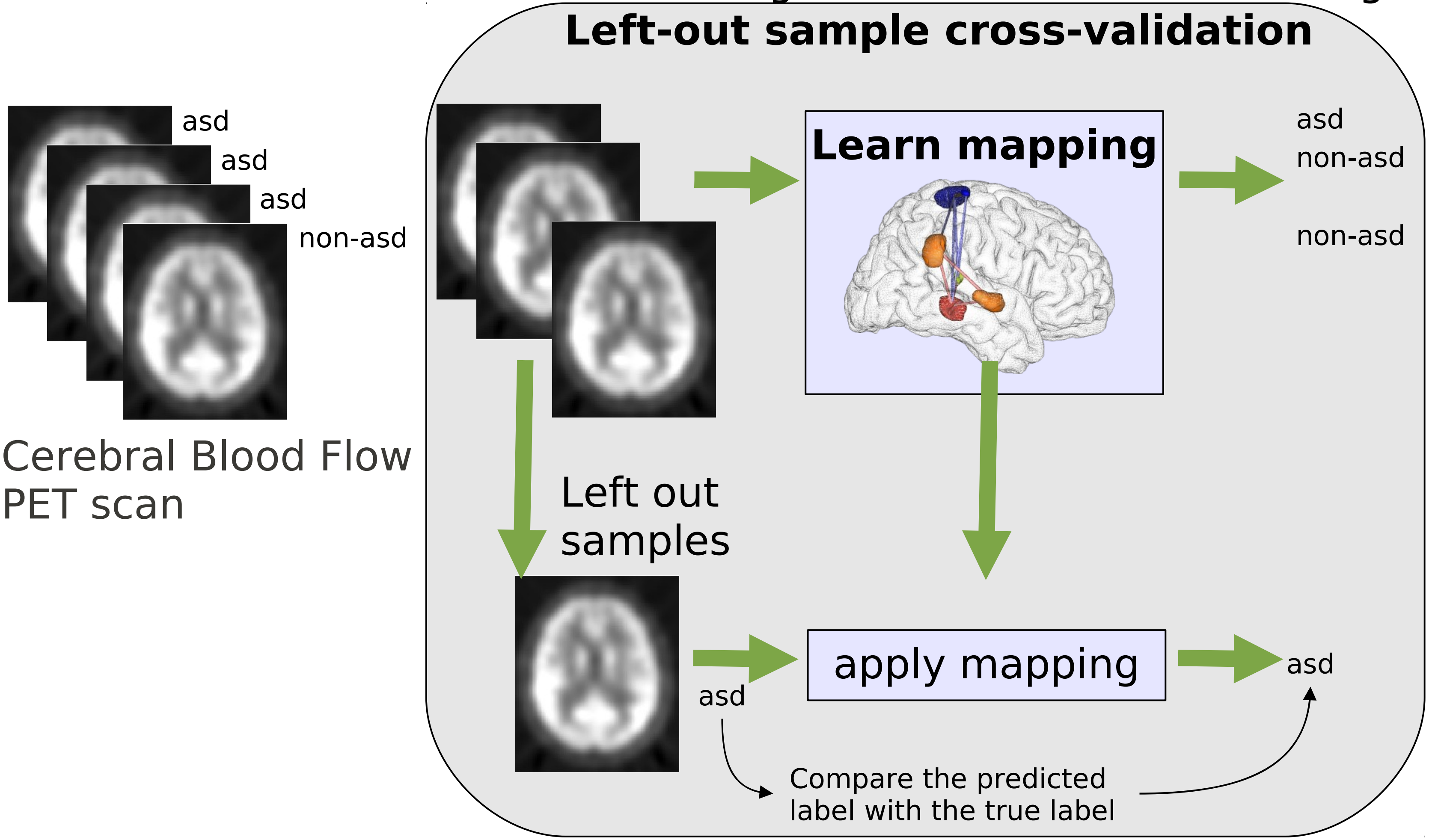
Use only “size” descriptors: sulcus length, depth, surface


Post central sulci (surface, max. depth)
Marginal pre−central sulci (size)
Terminal ascendant posterior branch (size)
Superior temporal sulci (max. depth)
Posterior sylvian fissure (suface)
Inferior parietal sulci (size)
Internal parietal sulci (size)
Parieto−occipital fissure (surface)
R>L
L>R
Z−value
P−value
−9.8
1.2e−20
−5.9
1.2e−8
−2.0
6.7e−2
2.0
6.7e−2
7.9
4.1e−14

Correct recognition rate : 96%.

ASD subject prediction based on PET-scans

rCBF PET scans of**45** low-functioning ASD and **13** low-functioning non-ASD



Cerebral Blood Flow
PET scan

Left-out sample cross-validation

Learn mapping

Left out
samples

apply mapping

Compare the predicted
label with the true label

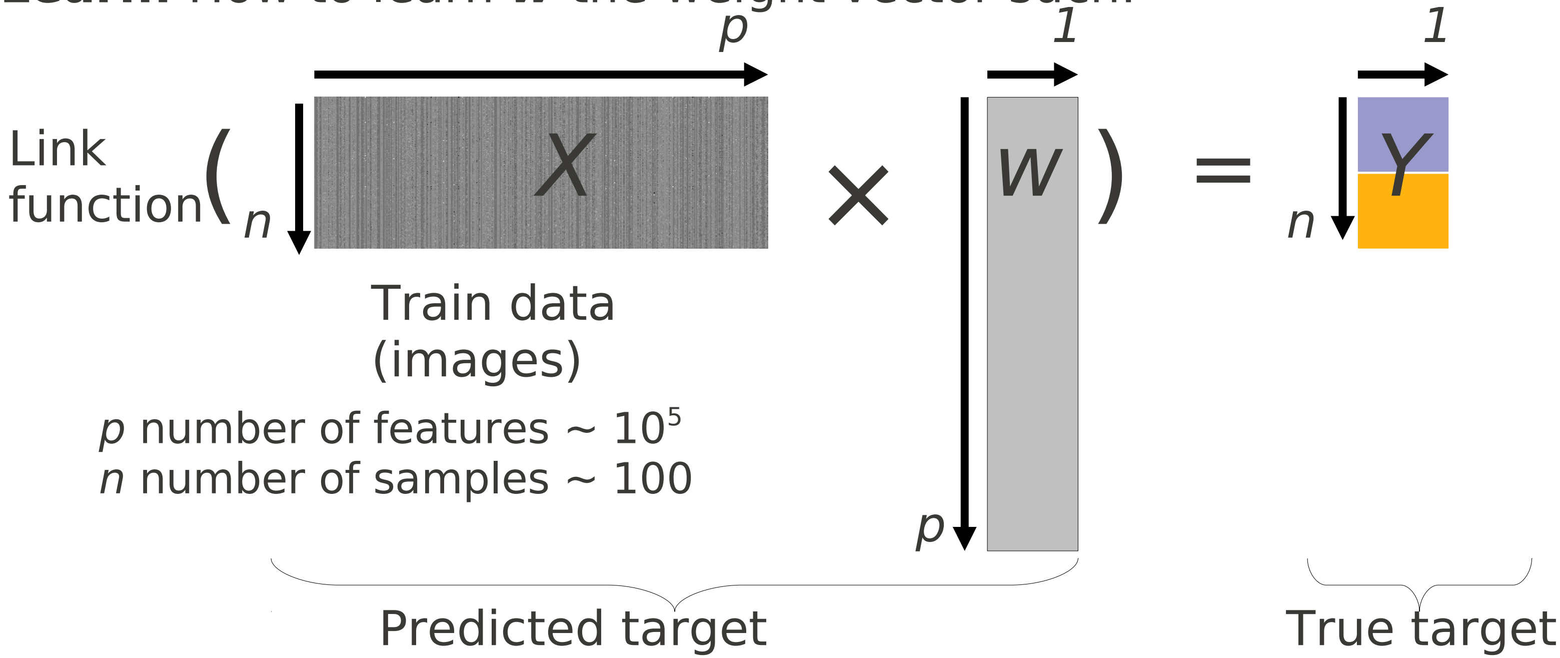
Assess the generalisation power of the learning algorithm on **independent data** (toward reproducibility)

Linear classifier

Prediction rule of linear discriminant classifier (**combine** features):

$$\text{Link function}\left(\boxed{\text{weight 0}} + \boxed{\text{weight 1}} * \boxed{\text{feat. 1}} + \boxed{\text{weight 2}} * \boxed{\text{feat. 2}} + \boxed{\text{...}} * \boxed{\text{...}} + \boxed{\text{weight P}} * \boxed{\text{feat. P}}\right) = \text{predicted target}$$

Learn: How to learn ***w*** the weight vector such:



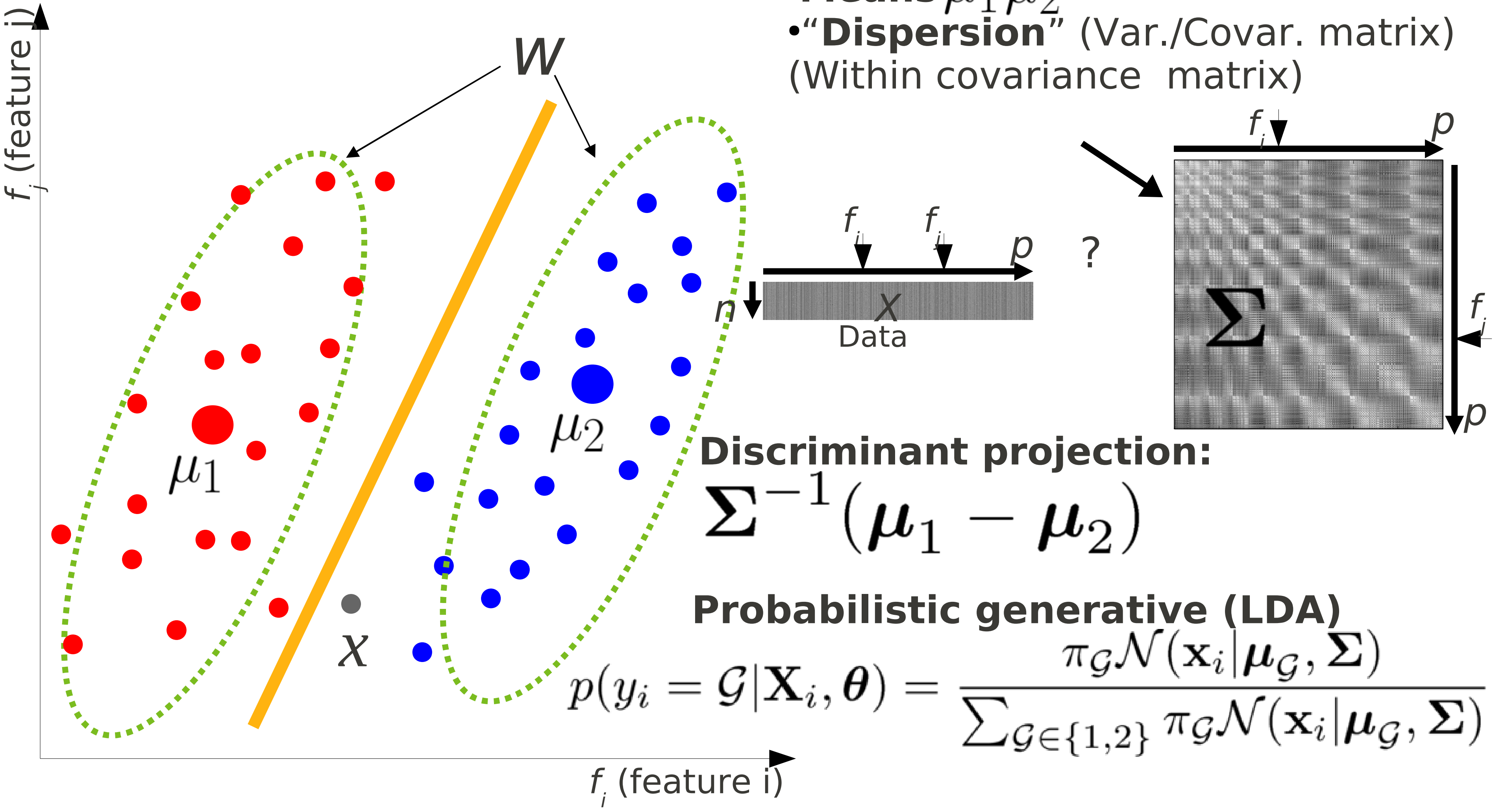
p number of features ~ 10^5

n number of samples ~ 100

Linear classifier: Parametric method

Estimate model parameters:

- **Means** $\mu_1\, \mu_2$
- "**Dispersion**" (Var./Covar. matrix) (Within covariance matrix)



Discriminant projection:

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Probabilistic generative (LDA)

$$p(y_i = \mathcal{G}|\mathbf{X}_i, \boldsymbol{\theta}) = \frac{\pi_{\mathcal{G}}\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{\mathcal{G}}, \boldsymbol{\Sigma})}{\sum_{\mathcal{G}\in\{1,2\}} \pi_{\mathcal{G}}\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{\mathcal{G}}, \boldsymbol{\Sigma})}$$

Linear classifier: discriminative approaches

Find ***w*** that minimize a prediction error on training data:

$$\underbrace{\|\mathbf{w}\|}_{\text{Penalization}}{}_L + C \underbrace{\sum_i \xi(\mathbf{w}; \mathbf{x}_i, y_i)}_{\text{Loss (Error) function between prediction And true label}}$$

L2 penalization: SVM

– *L* = 2

– Hinge loss: $\xi(\mathbf{w}; \mathbf{x}_i, y_i) = \max(1 - y_i \mathbf{x}_i' \mathbf{w}, 0)$

L1 penalization: Lasso Logistic Regression

– *L* = 1

– Logistic loss: $\xi(\mathbf{w}; \mathbf{x}_i, y_i) = \log(1 + e^{-y_i \mathbf{x}_i' \mathbf{w}})$

→ **Minimisation of misclassification: favour most numerous class**

→ **Poor specificity**

Group size imbalance problem

1) Samples re-weighting simple for both SVM and Lasso Logistic Regression
→ Good sensitivity (detection of the mos numerous class)
→ **Poor specificity** (detection of the least numerous class)

2) Sub-sampling of the most numerous class: can afford to drop some of the few 45 samples of ASD group

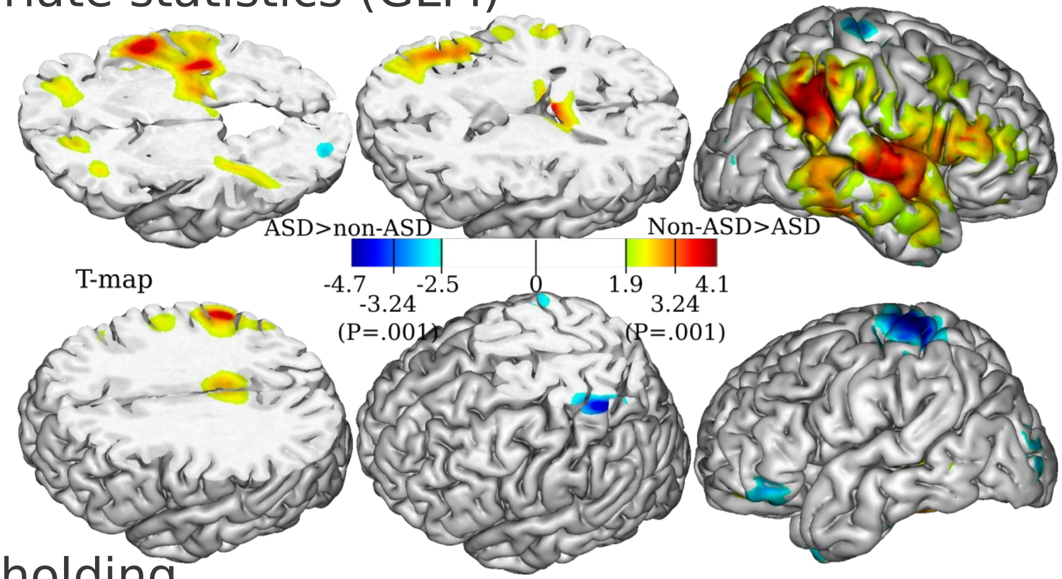
3) Two separate one class learning
~ **generative methods** ie.: learn the conditionals $p(\boldsymbol{x}_i|y_i)$.
The predictive = conditional * explicit priors
→ Linear Discriminant Analysis (**LDA**)

Pb.: Overfitting P(P+5)/2 estimated parameters (intraclass variance)

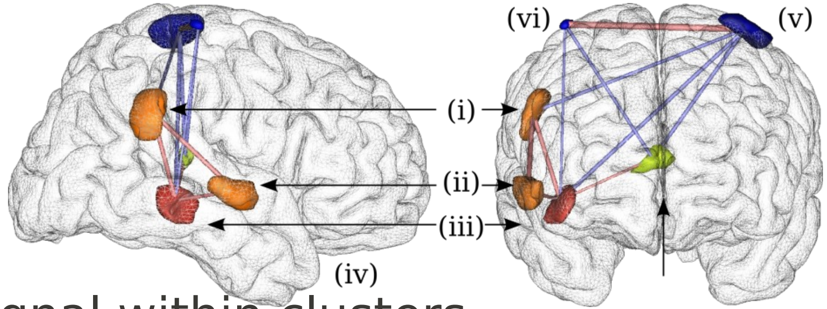
→ **Dimension reduction**

Dimension reduction: 1/2 regional features

Goal driven regional feature extraction:
Univariate statistics (GLM)

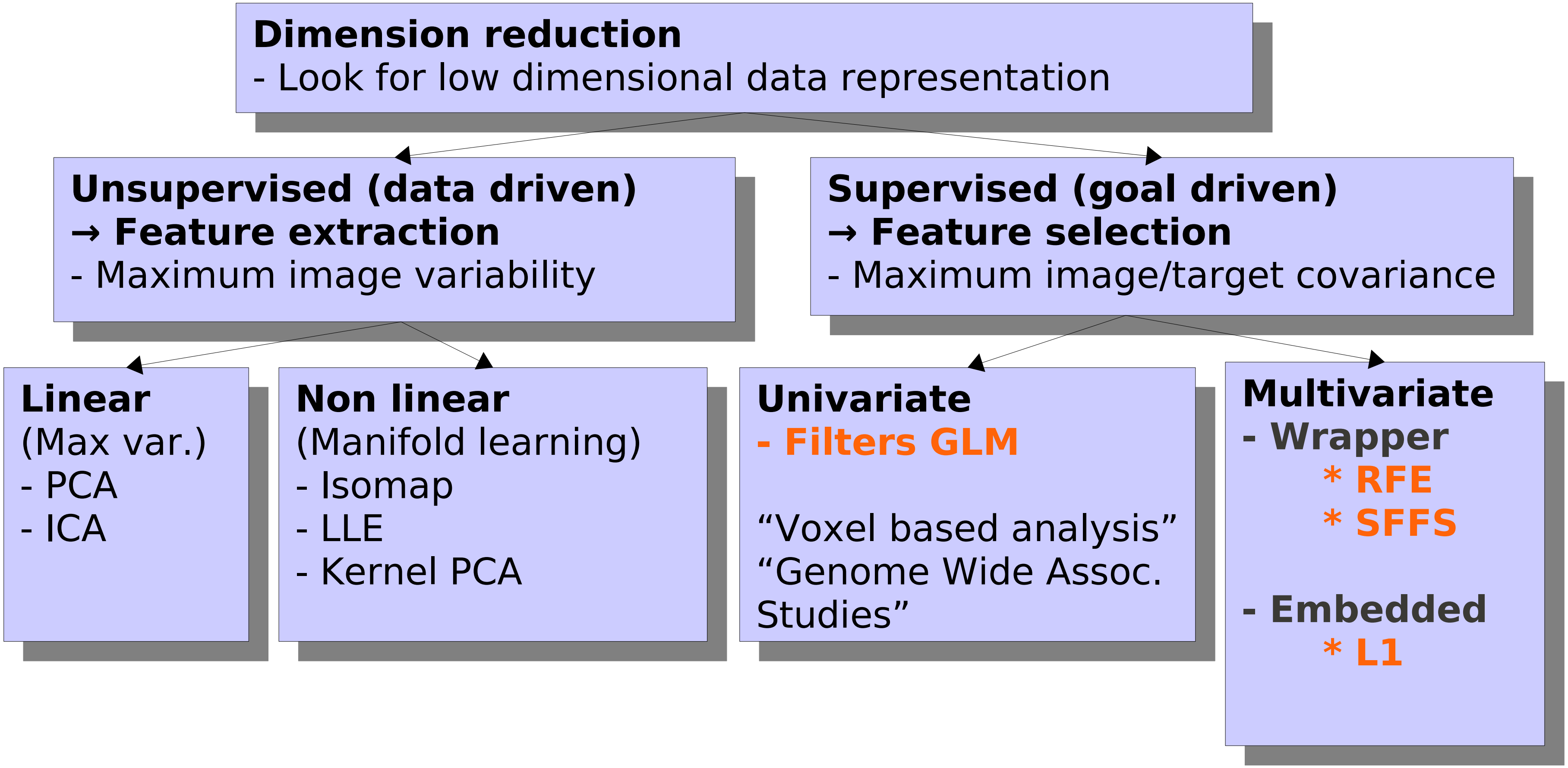


Thresholding



Average signal within clusters

Dimension reduction: 1/2 methods



Dimension reduction
- Look for low dimensional data representation

Unsupervised (data driven)
→ Feature extraction
- Maximum image variability

Supervised (goal driven)
→ Feature selection
- Maximum image/target covariance

Linear
(Max var.)
- PCA
- ICA

Non linear
(Manifold learning)
- Isomap
- LLE
- Kernel PCA

Univariate
- Filters GLM

“Voxel based analysis”
“Genome Wide Assoc. Studies”

Multivariate
- Wrapper
*** RFE**
*** SFFS**

- Embedded
*** L1**

Feature selection = Feature subset ranking + model selection

Feature subset ranking produce sets of features (F_k) of increasing size k

– Filter and RFE: nested sets are nested
– Lasso, SFFS: eventually non-nested sets

Model selection

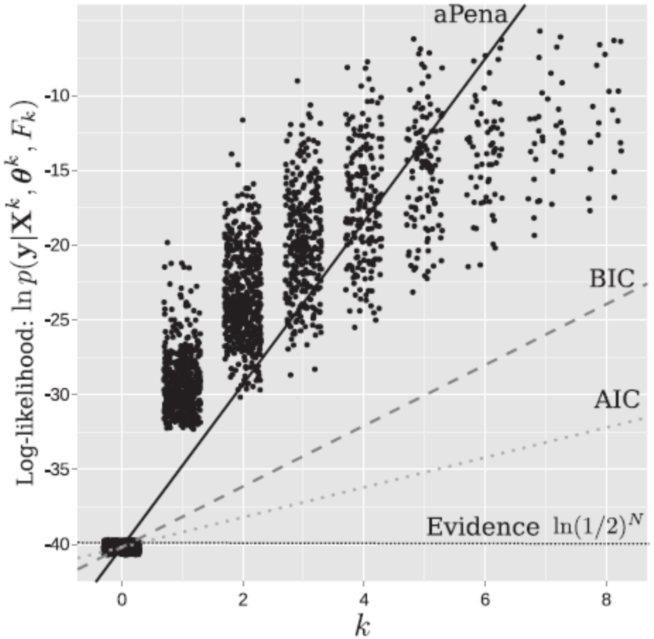
Select F_k that maximizes some criteria

Here Choose feature subset F_k made of k (regional) features

1) CV → Computational issue: 3 levels of nested sampling loop

2) Penalized likelihood

$$\underbrace{\ln p(\mathbf{y}|\mathbf{X}^k, F_k)}_{\text{Evidence}} \simeq \underbrace{\ln p(\mathbf{y}|\mathbf{X}^k, \boldsymbol{\theta}^k, F_k)}_{\text{Log likelihood}} - a\underbrace{\tfrac{1}{2}k\ln N}_{\text{Penalisation}}$$



Many fixed penalty criteria BIC, AIC, etc.

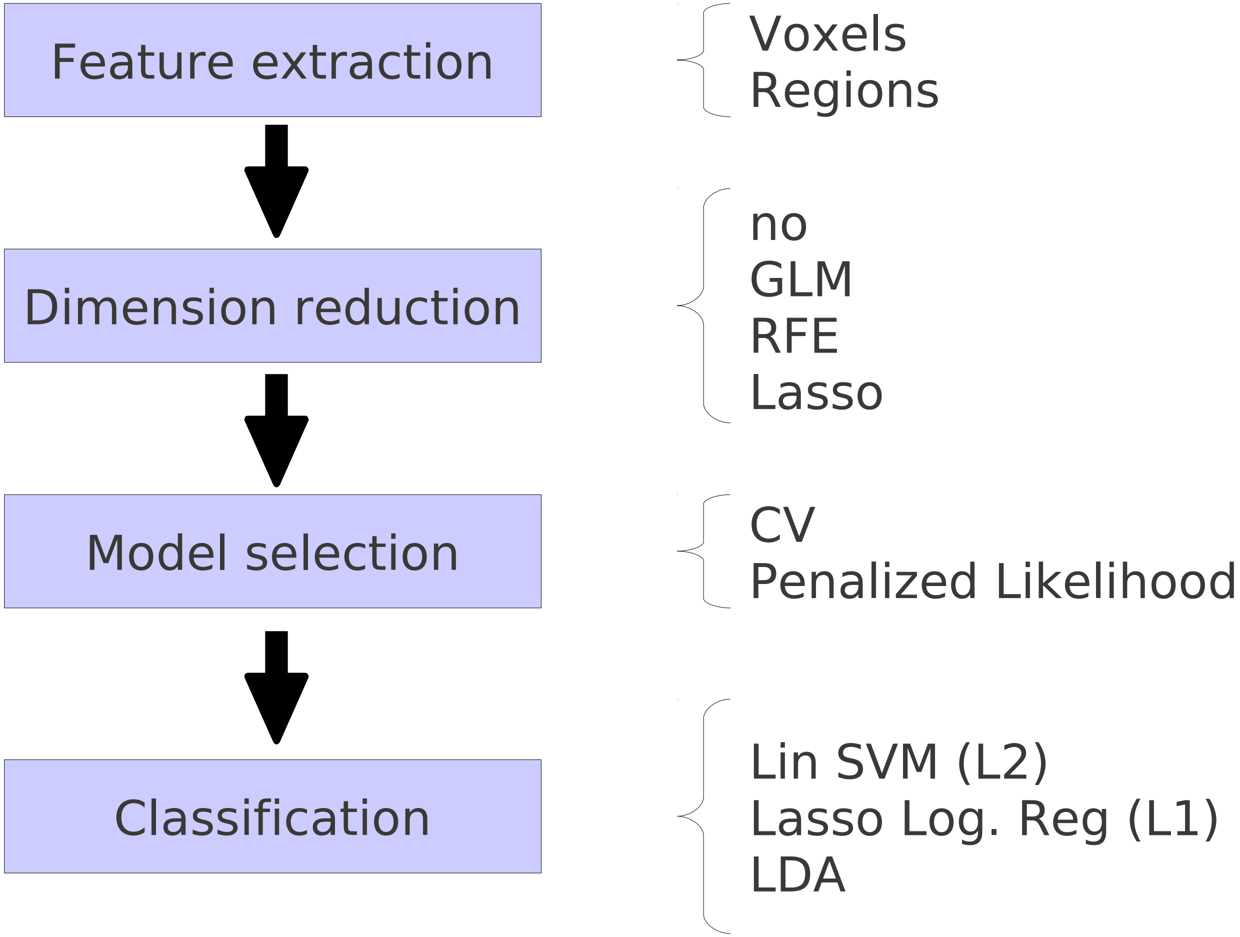
Under penalization (ignore feature selection)

Data driven calibration of the penalty [Birgé 07]

Add a free parameter "a"

Calibrated with random permutation

Comparison methodology



Feature extraction

- Voxels
- Regions

Dimension reduction

- no
- GLM
- RFE
- Lasso

Model selection

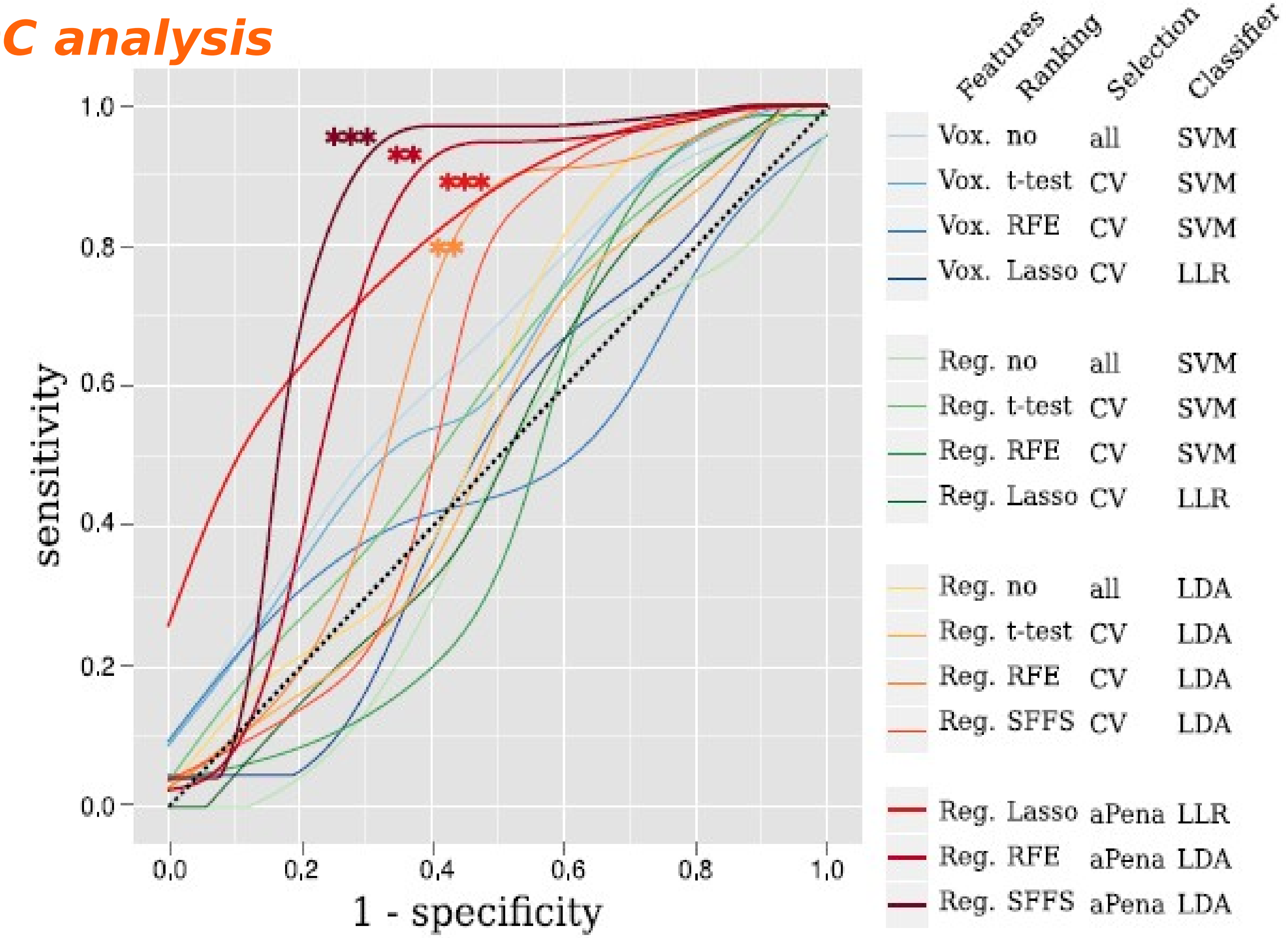
- CV
- Penalized Likelihood

Classification

- Lin SVM (L2)
- Lasso Log. Reg (L1)
- LDA

Results (Validation)

ROC analysis

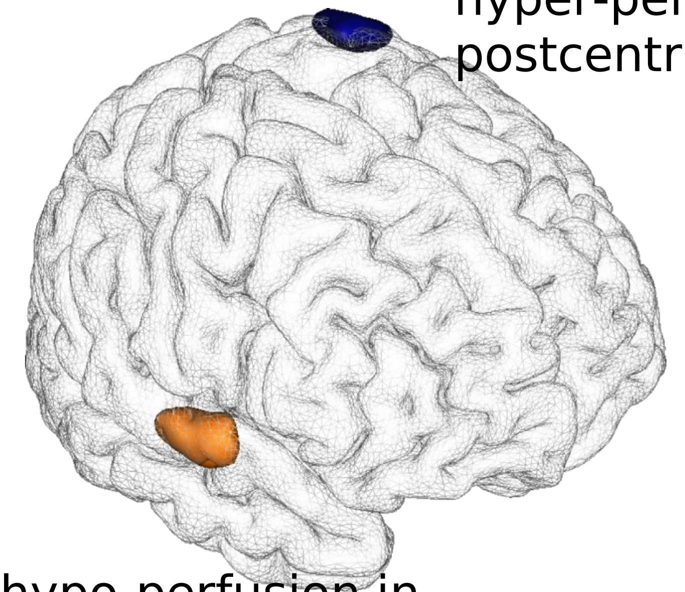


Regional features + Multivariate feature selection + Generative

Leave-One Out Cross validation

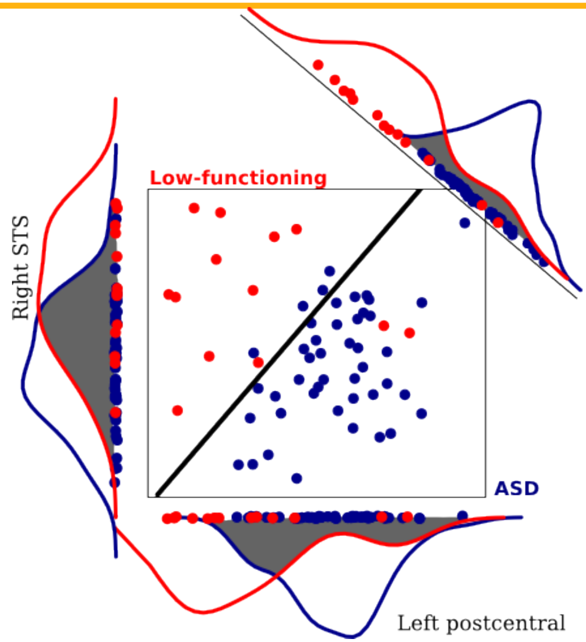
Accuracy 87%***, Sensi. 91%***, Speci. 77%*
Significance calibrated with permutation (*** p<0.001, * p<0.05)

Discriminative pattern



hyper-perfusion in
postcentral area (in ASD)

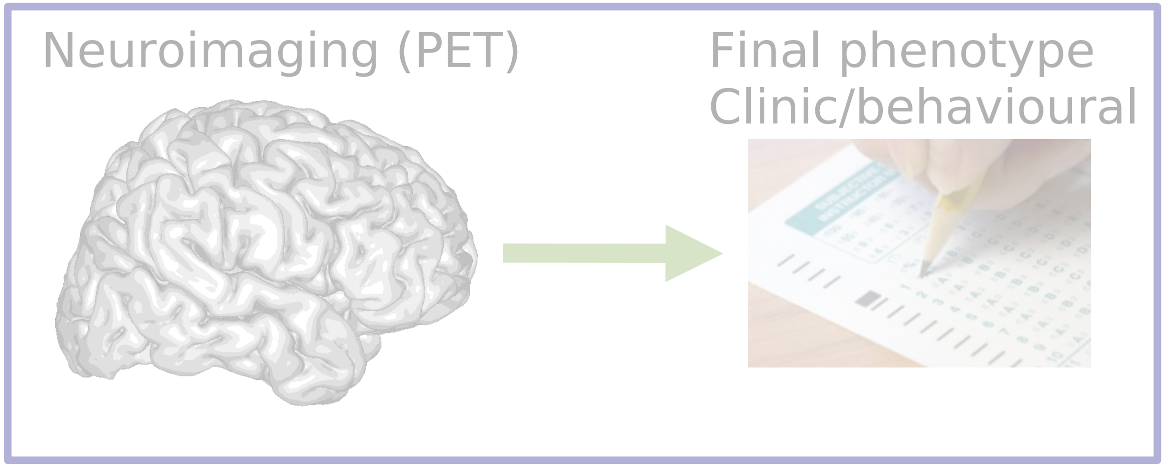
hypo-perfusion in
Superior Temporal Sulcus (in ASD)



The combination of the signal in the two regions enable a clear separation of ASD vs controls

– **Good stability**: same pattern is selected across all re-sampling
– **Shared** pattern that discriminates all ASD from controls
– Multiple etiologies of ASD + numerous neuroimaging findings suggests that **several others brain patterns** may exist across the autistic **spectrum**
– Next step: look for the more specific multiple patterns associated with the multiple etiologies

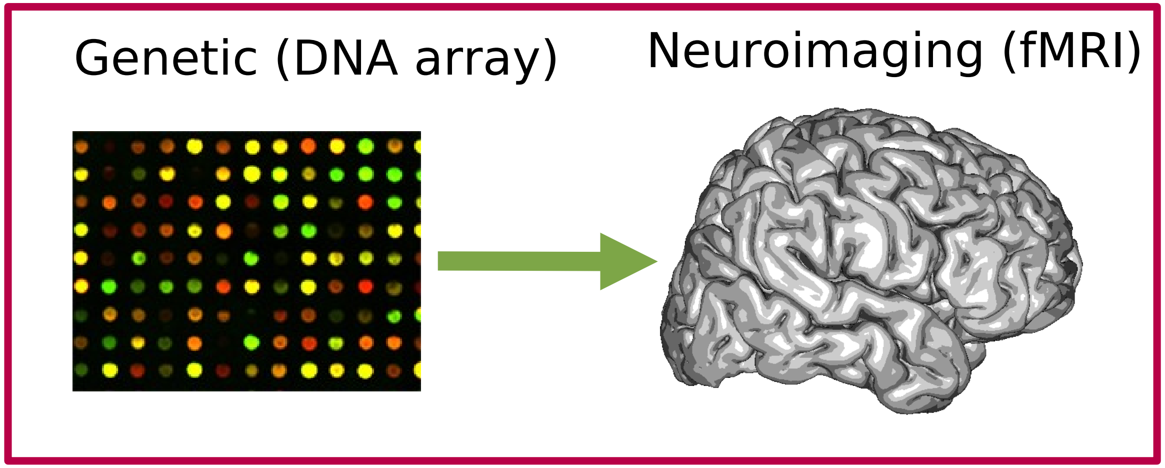
Imaging-genetics



Neuroimaging (PET)

Final phenotype
Clinic/behavioural

Application to autism



Genetic (DNA array)

Neuroimaging (fMRI)

Application: asymetries in language processing

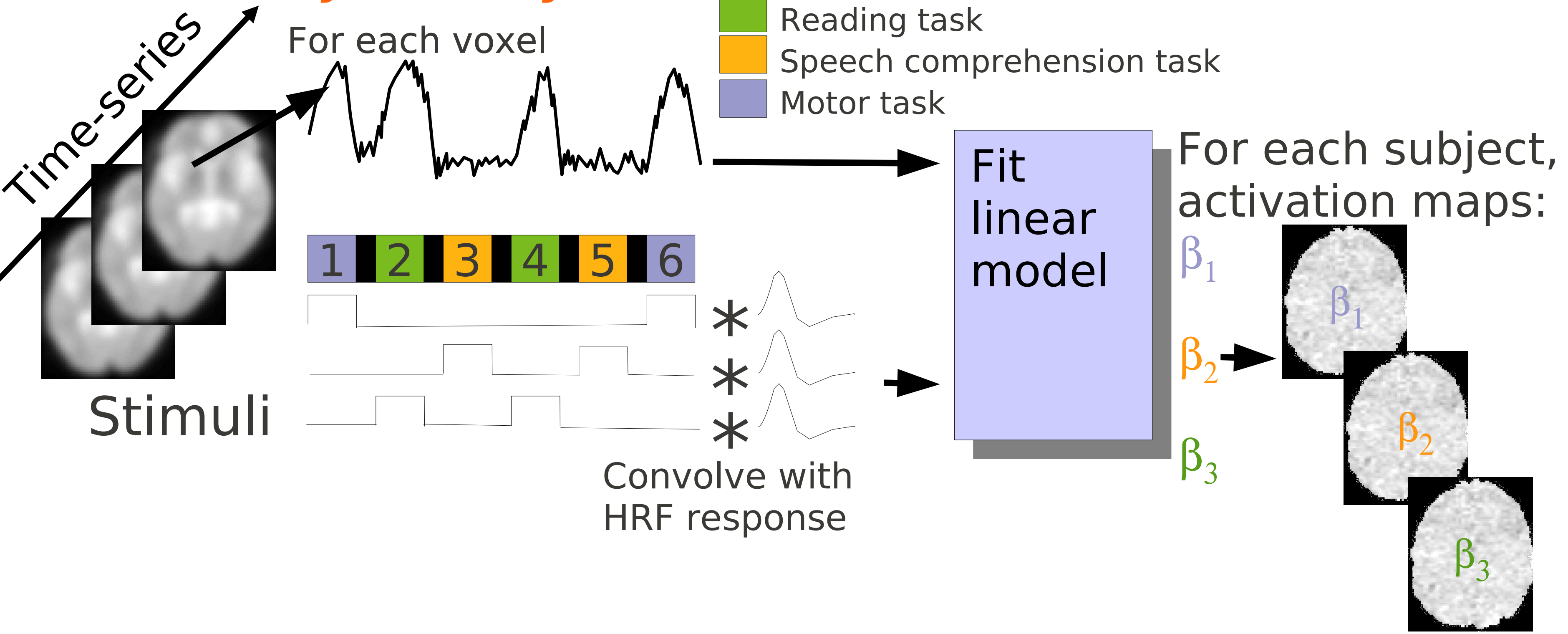
Functional MRI data of the experimental dataset

1) Inter-subject normalization

2) Intra-subject analysis

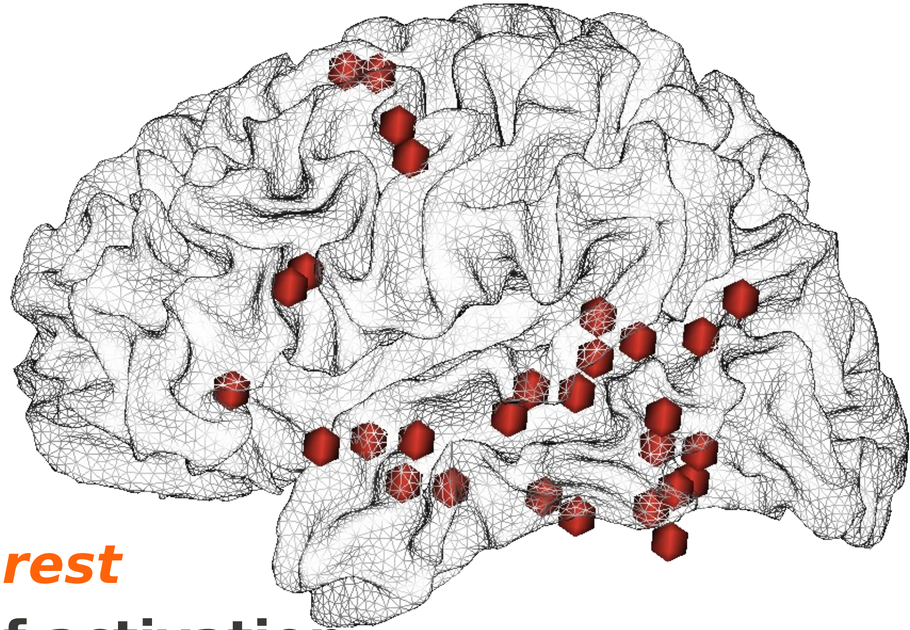
▮ No stimulus
🟩 Reading task
🟧 Speech comprehension task
🟪 Motor task

[Pinel et al., 2007]



Extraction of contrast maps for:
- a **reading** task
- a **speech comprehension** task

Functional MRI data of the experimental dataset

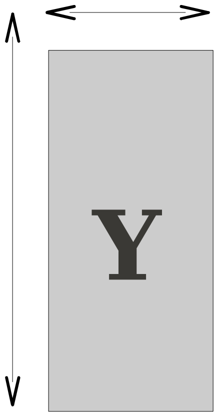


3) Choice of brain regions of interest

- According to the data: **maxima of activation**
- According to the literature: **involved in dyslexia and language networks**

4) Computation of 34 lateralization indexes

$$\hat{\beta}_s^{\text{index}} = \frac{|\hat{\beta}_s^{\text{left}} - \hat{\beta}_s^{\text{right}}|}{\sqrt{(\hat{\beta}_s^{\text{left}})^2 + (\hat{\beta}_s^{\text{right}})^2}}$$



Q=34 imaging phenotypes

N=94 subjects

Genetic data of the experimental dataset

DNA microarray (Illumina)

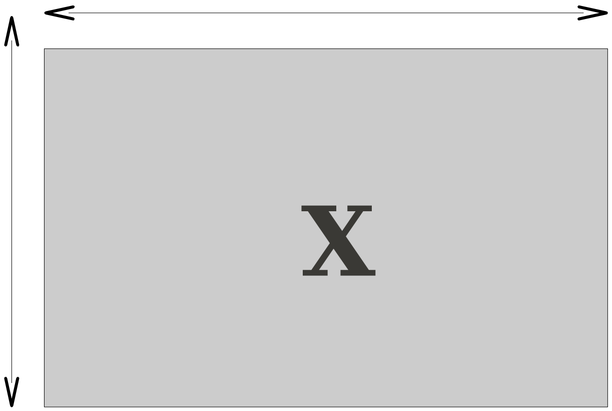
1,054,068 **Single Nucleotide Polymorphisms** (SNPs)

SNPs: Most variable nucleotides across the genome

For each SNP 3 possible values: AA, AB, BB

Pre-processing:

- **Filtering** : (1) Minor Allele Frequency (MAF) at least 10%
  (2) call rate at least 95%
  (3) Hardy-Weinberg test not significant at 0.005
- **Coding** : for each SNP, number of minor alleles {0,1,2}
- **Missing SNP** data were imputed with their corresponding median



p=622, 534 SNPs

N=94 subjects

[Pinel et al., 2007]

Simulated dataset

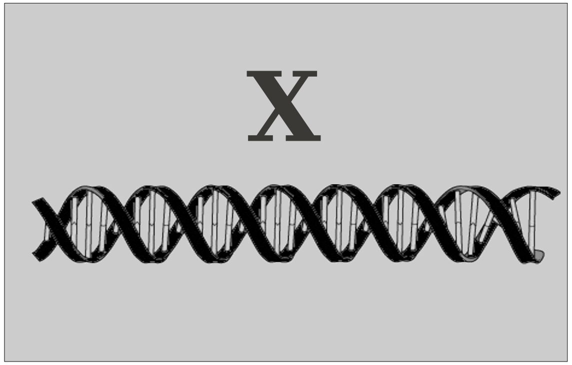
Genetic data
Simulate realistic genetic data
gs algo. [Li & Chen 08]
Hapmap CEU Panel
n=500 unrelated subjects
p=85,772 SNPs (chromosome 1)

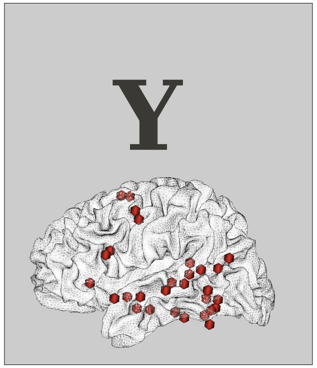
Imaging data
Sample from multivariate ()
Parameters $\mu_1 \mu_2 \mathbf{\Sigma}$ estimated
on experimental data
n=500, *q*=34

Genetic effect (additive model)

– Randomly select 10 SNPs with MAF=0.2
– Two causal patterns each involves 5SNPs → 4 ROIs
– For each causal patterns i in 1...2
   * Average the 5 SNPs $\bar{x}_i$
   * For each ROI *j* in 1...4: y_{ij}
      – $y^{\star}_{ij} = y_{ij} + \beta_{ij}\bar{x}_i$
      β_{ij} Controls for explained variance of $\bar{x}_i$ on $y^{\star}_{ij}$
– SNPs in high LD (R2>0.8) with true causal are considered causal (56 SPNs)
– Strip of haplotype blocks in the causal SNPs neighbourhood (198 SNPs) and move them at the beginning of the dataset

What is currently done in imaging genetics?





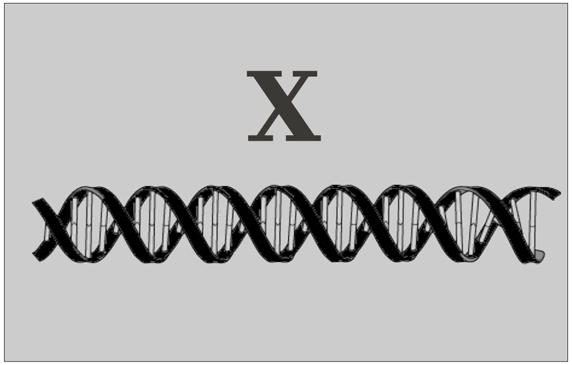
Genome Wide Association Studies (GWAS)

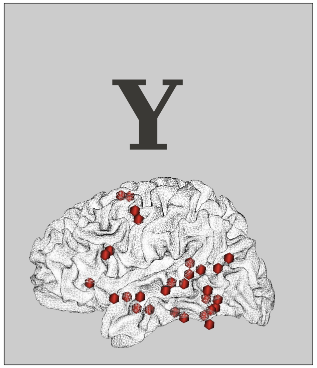
Massive univariate testing of each SNP versus each imaging phenotype independently **(simple linear regression)**

Problems:

- no SNP/phenotype association survived **multiple comparisons**
- **Multivariate nature** of the imaging/genetics link not taken into account

Multivariate methods in imaging genetics





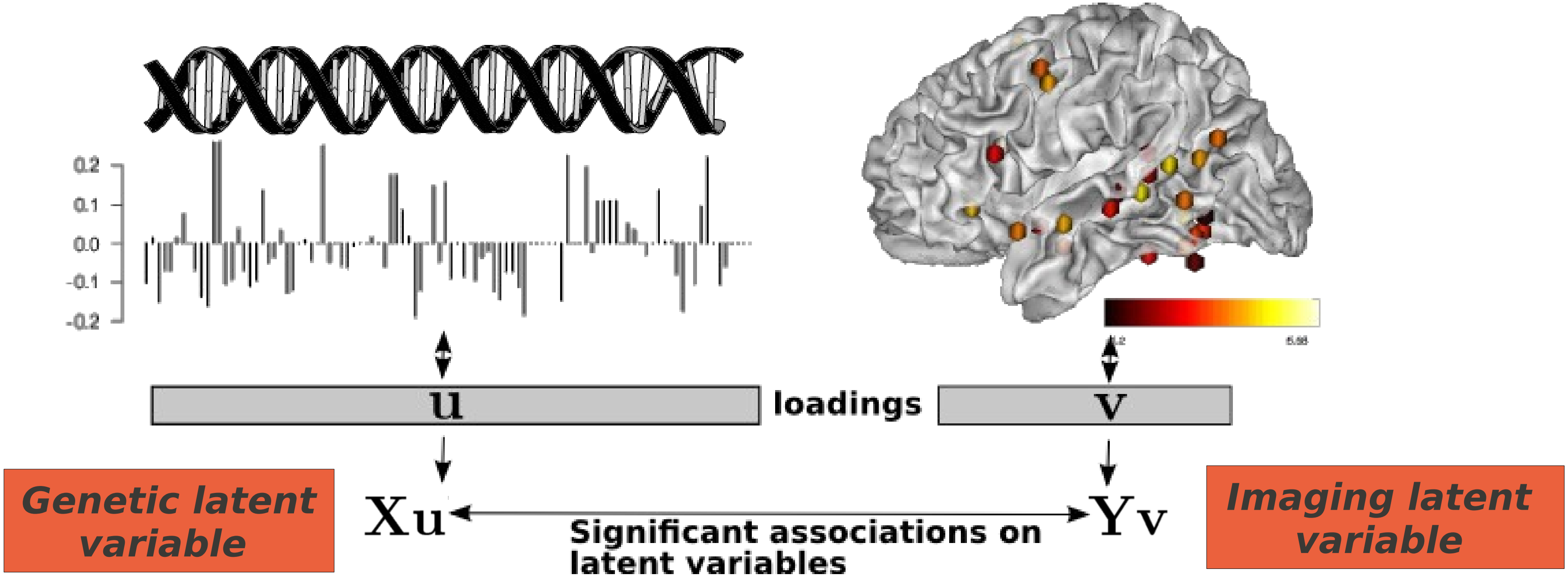
Goal

Study the link between genetic and neuroimaging data by taking into account the **interactions between genes** and the **interactions between brain regions**

→ looking for **associations** between **two co-varying networks** of genes and brain regions

Problem:
Curse of dimensionality: multivariate methods overfit in high dimensional settings (find associations just by chance)

Multivariate methods based on latent variables



Genetic latent variable

u

loadings

v

Xu

Yv

Significant associations on latent variables

Imaging latent variable

– Canonical Correlation Analysis
– Partial Least Squares

Other two blocs methods in imaging-genetic context
→ Parallel ICA [Calhoun 09]

Multivariate methods based on latent variables

Partial Least Squares (PLS) Regression

Maximizes the covariance between the two latent variables:

$$\max_{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1} \mathbf{u}'_h \mathbf{X}'_{h-1} \mathbf{Y}_{h-1} \mathbf{v}_h$$

→ Solved by an iterative algorithm (NIPALS)
→ Further pairs of components obtained after deflation of X and Y

Canonical Correlation Analysis (CCA)

Maximizes the correlation between the two latent variables:

$$\max_{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1} \frac{\mathbf{u}'_h \mathbf{X}'_{h-1} \mathbf{Y}_{h-1} \mathbf{v}_h}{\sqrt{\mathbf{u}'_h \mathbf{X}'\mathbf{X}\mathbf{u}_h}\sqrt{\mathbf{v}'_h \mathbf{Y}'\mathbf{Y}\mathbf{v}_h}}$$

Numerical issues: dual formulation of CCA: Kernel CCA (KCCA)

Regularisation

L2 regularised CCA (rCCA)

Add diagonal term to intra-block scatter matrices

$$\mathbf{X}'\mathbf{X} \to \mathbf{X}'\mathbf{X} + \lambda_1 \mathbf{I} \quad \mathbf{Y}'\mathbf{Y} \to \mathbf{Y}'\mathbf{Y} + \lambda_2 \mathbf{I}$$

Extreme case of regularisation, scatter matrices → $\mathbf{I}$ ↔ rCCA ~ PLS

L2 regularised CCA (rCCA)

- [Waaijenborg08] Elastic Net: lasso and ridge $\mathbf{X'X=Y'Y+}\lambda\mathbf{I}$
- [Parkhomenko09] Soft-thresholding, $\mathbf{X'X}=$ diag($\mathbf{X'X}$)
- [Witten09] $\mathbf{X'X=Y'Y=I}$

→ such extreme regularization on $\mathbf{X'X}$ makes CCA ~ PLS

L1 regularised PLS: sparse PLS (sPLS)

Add L1 penalisation on the SNPs weights ($\mathbf{u}$)

$$\min_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_{1X}\|\mathbf{u}\|_1$$

– [Chun & Keles07] (regression)

– [LeCao08]

  * **Soft thresholding** within PLS iterations

Sparse PLS: NIPALS revisited

Bi-convex in ($\mathbf{u}$) and ($\mathbf{v}$), add soft-thresholding within the NIPALS loop

```
**NIPALS_soft_thresholding** (X, Y)
$\mathbf{X}_0 = \mathbf{X}, \mathbf{Y}_0 = \mathbf{Y}$
Iterate over components ($h$ in 0...H):
    1. Initialize $\mathbf{u}$ and $\mathbf{v}$ using for instance the first pair of singular vectors of
       the matrix $\mathbf{X}'\mathbf{Y}$ and normalize them.
    2. Until convergence of $\mathbf{u}$ and $\mathbf{v}$:
        (a) For fixed $\mathbf{v}$, find $\hat{\mathbf{u}} = \arg\min_{\|\mathbf{u}\|_2=1} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_{1X}\|\mathbf{u}\|_1$
            • $\hat{\mathbf{u}} = g_{\lambda_{1X}}(\mathbf{X}'\mathbf{Y}\mathbf{v})$;    $\mathbf{u} = \hat{\mathbf{u}}/\|\hat{\mathbf{u}}\|_2$
              Where $g_\lambda(y) = \text{sign}(y)(|y|-\lambda)_+$ is the soft-thresholding function.
        (b) For fixed $\mathbf{u}$, find $\hat{\mathbf{v}} = \arg\min_{\|\mathbf{v}\|_2=1} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v}$
            • $\hat{\mathbf{v}} = \mathbf{Y}'\mathbf{X}\mathbf{u}$;    $\mathbf{v} = \hat{\mathbf{v}}/\|\hat{\mathbf{v}}\|_2$
    3. Compute latent variables, loadings and deflation
        • $\mathbf{x}^* = \mathbf{X}_h\mathbf{u}$;    $\mathbf{x}_{load} = \mathbf{X}'_h\mathbf{x}^*/\|\mathbf{x}^*\|$;    $\mathbf{X}_{h+1} = \mathbf{X}_h - \mathbf{x}^*\mathbf{x}'_{load}$
        • $\mathbf{y}^* = \mathbf{Y}_h\mathbf{v}$;    $\mathbf{y}_{load} = \mathbf{Y}'_h\mathbf{y}^*/\|\mathbf{y}^*\|$;    $\mathbf{Y}_{h+1} = \mathbf{Y}_h - \mathbf{y}^*\mathbf{y}'_{load}$
**return** ($\mathbf{x}^*$, $\mathbf{y}^*$)
```

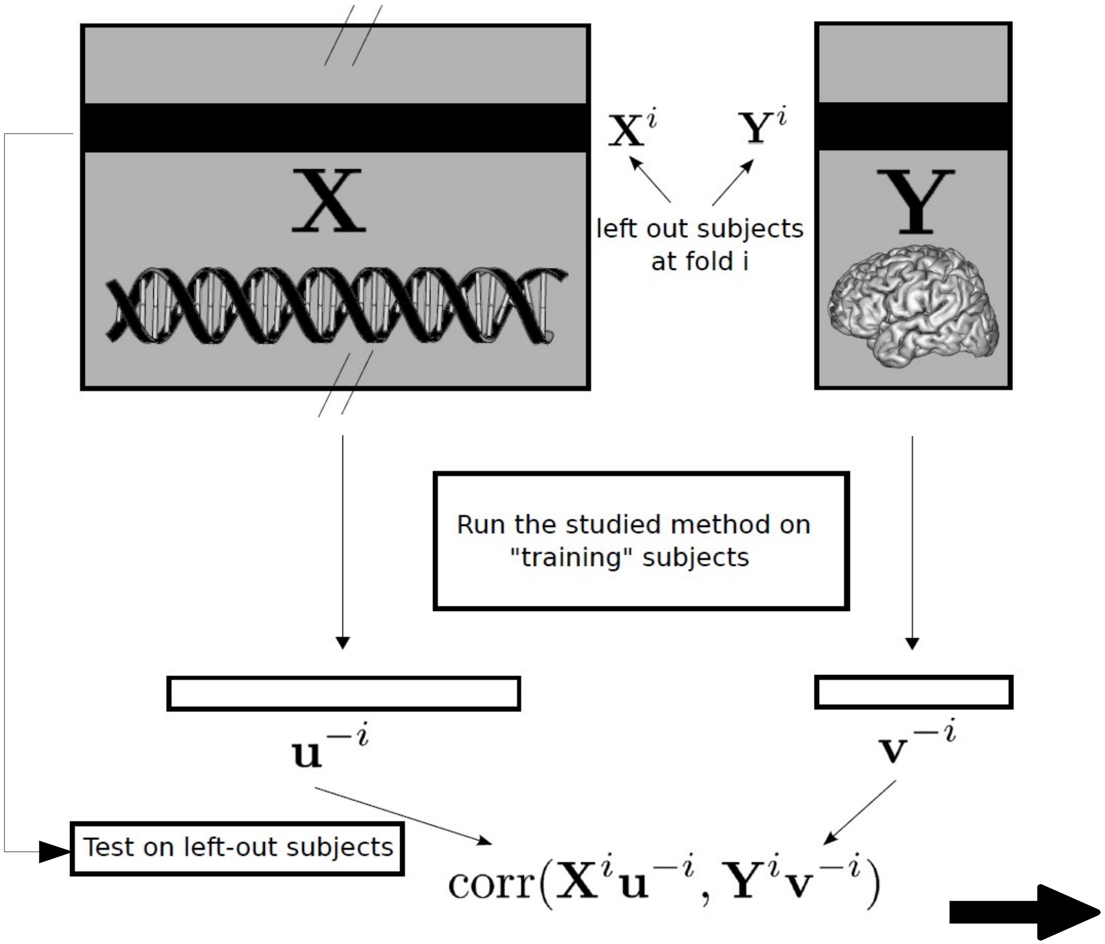
Dimension reduction

– We are interested in inter-blocks correlation → CCA
– However CCA overfitt: poor estimation of intra-bloc “variance” (scatter) matrix
– Sparse PLS show promising results on simulated but failed on experiential data
– Add dimension reduction to remove unwilling intra-bloc variance?

→ PCA
→ Feature selection based on filtering

Cross-validation (CV) and significance assessment

For each fold *i* of the 10-fold CV:



Run the studied method on "training" subjects

Test on left-out subjects

$$\mathrm{corr}(\mathbf{X}^i\mathbf{u}^{-i}, \mathbf{Y}^i\mathbf{v}^{-i})$$

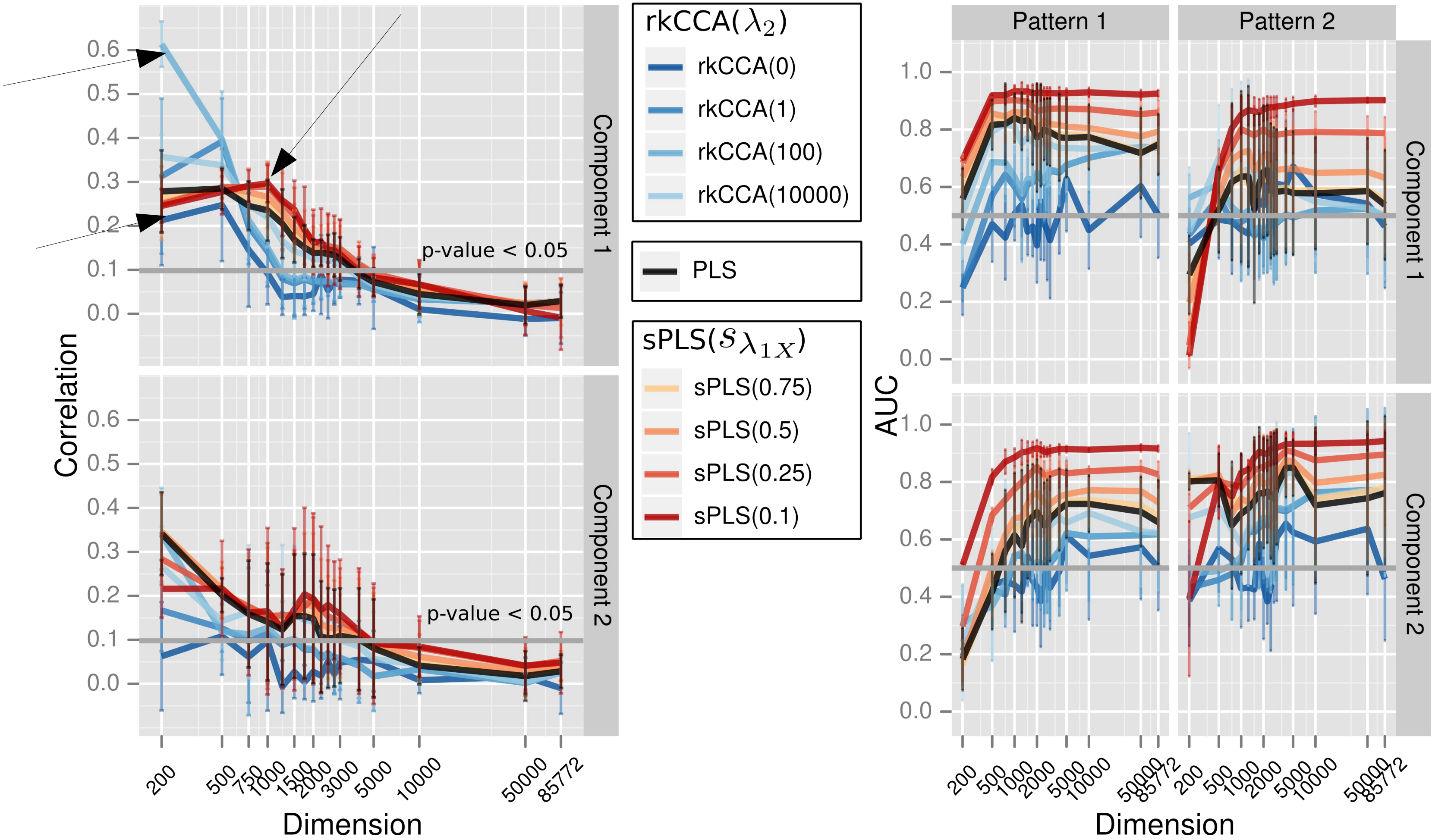
Out-of-sample correlation between latent variables averaged over folds

1000 permutations → **p-value** of the average out-of-sample correlation

Results on simulated dataset

Comparison of penalisation strategies

SPLS: sparse (L1 regularised) SPLS
rkCCA: L2 regularised kernelized CCA

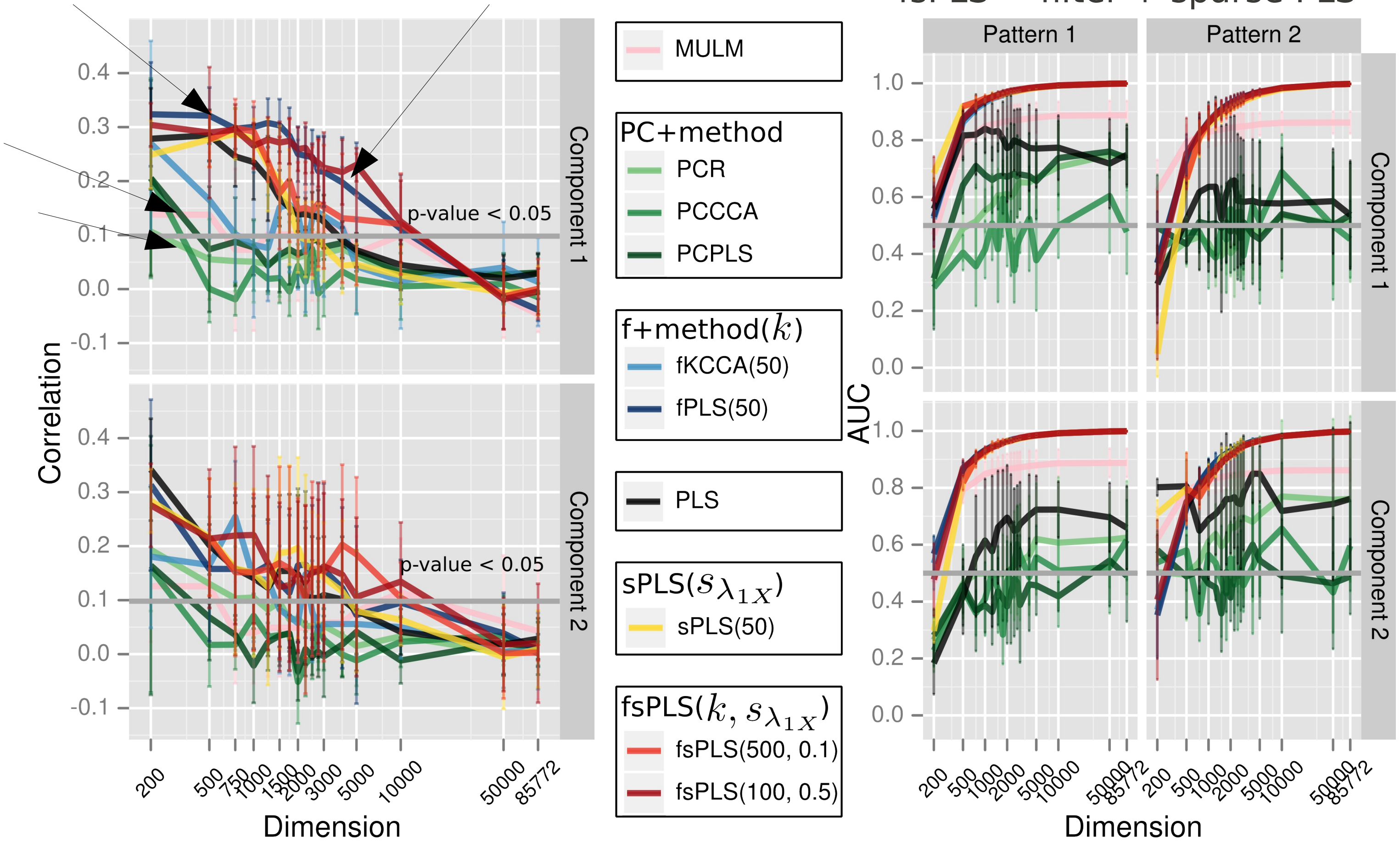


→ Sparse PLS outperformed other methods

Results on simulated dataset

Comparison of dimension reduction strategies

MULM: Massive Univariate Linear Model
PC = PCA, f = filter
fsPLS = filter + sparse PLS



→ **Combined filter + sparse PLS outperformed other methods**

Results on experimental dataset

Classical univariate analysis

→ **no significant** SNP/phenotype associations after correction for **multiple comparisons**

Multivariate methods

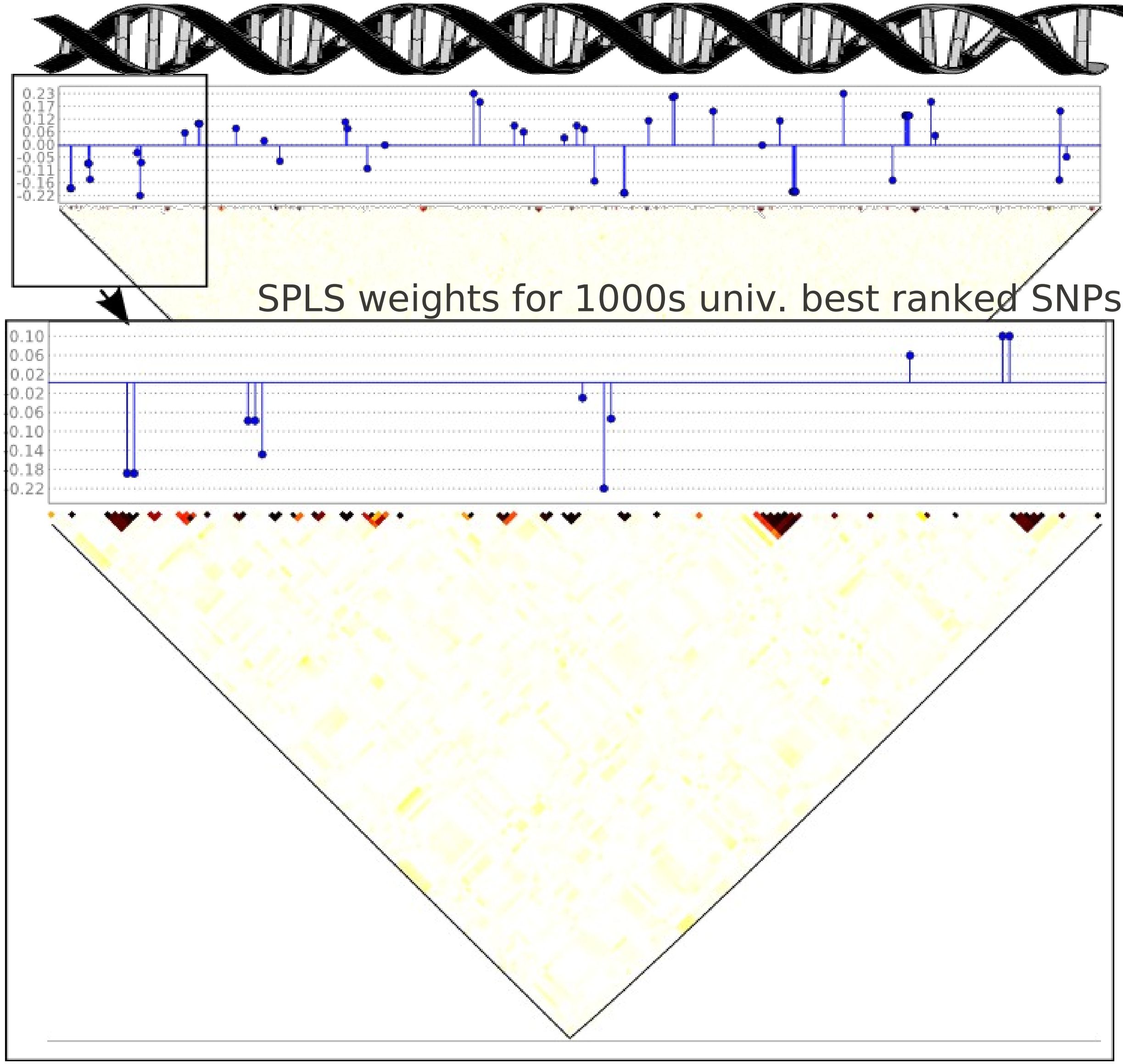
→ **Average out-of-sample correlation** between latent variables:

PLS	-0.09
Sparse PLS	0.19
Filtering + Sparse PLS	**0.43***

* **significant** p-value (p<0.05), computed with permutation correction for the multiple experiments using maxT

→ Gain in sensitivity compared to univariate analysis
→ Both filtering and sparsity seem necessary

Results: Weights assigned to selected SNPs



SPLS weights for 1000s univ. best ranked SNPs

→ 50 SNPs selected on **all chromosomes**

→ Only **14 Genes**

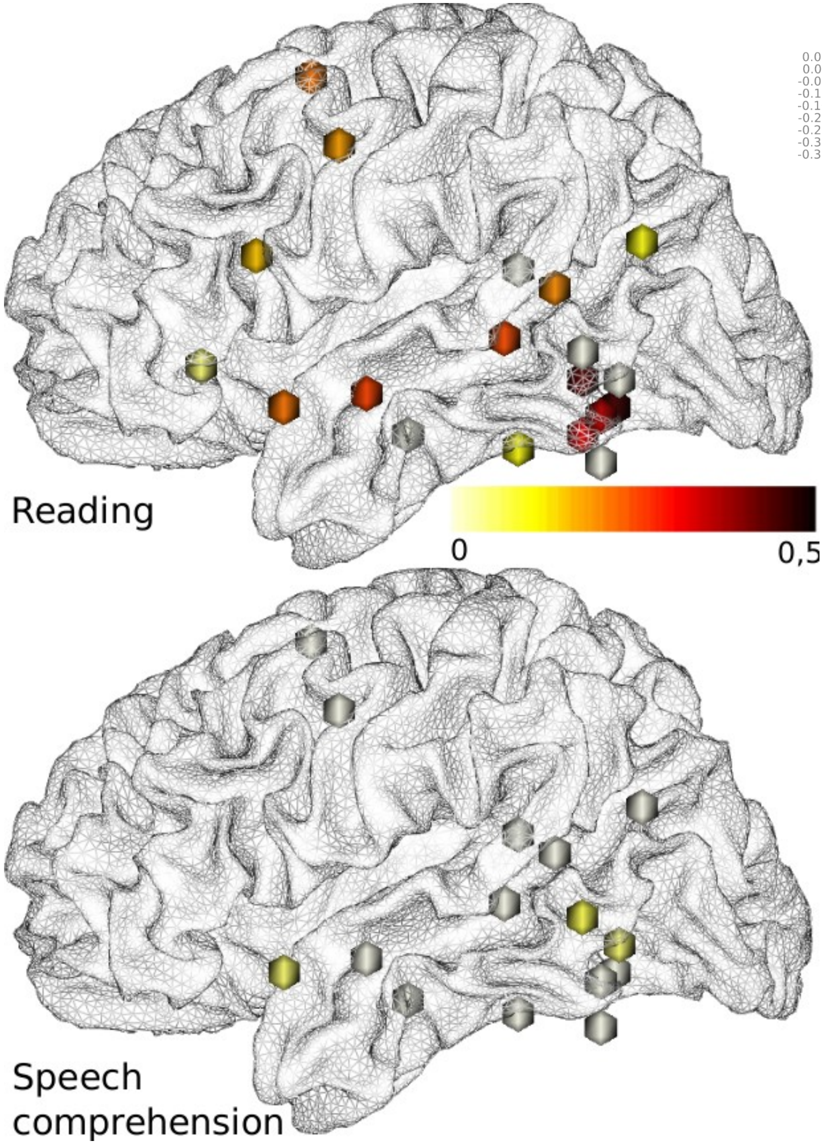
→ PLS weights != univ. ranking

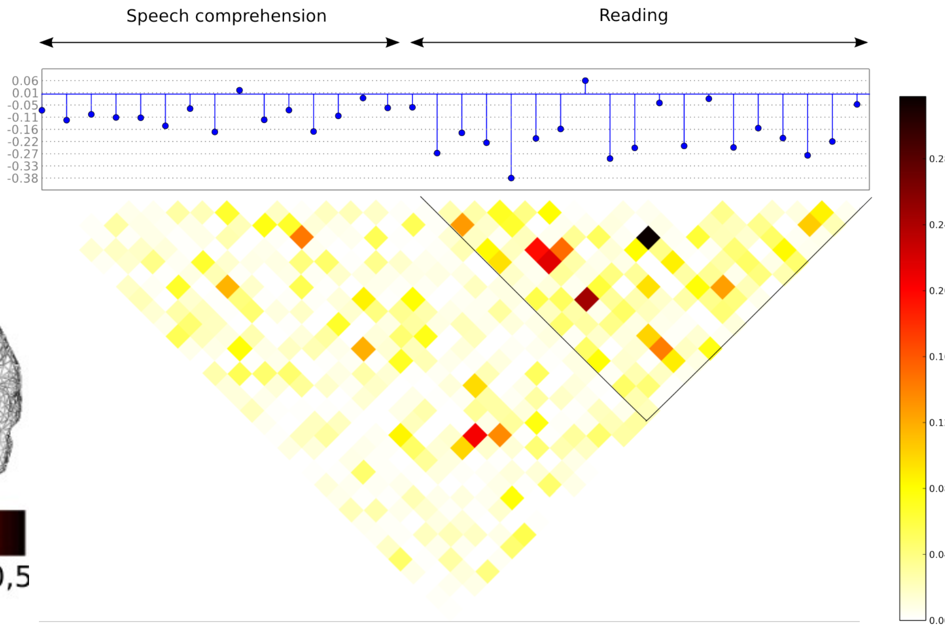
→ Some (rare) *correlated neighboring SNPs (in* ***linkage disequilibrium****)* *selected together*

→ *PPP2R2B and RBFOX1 ataxia and a poor coordination of* speech and body movement

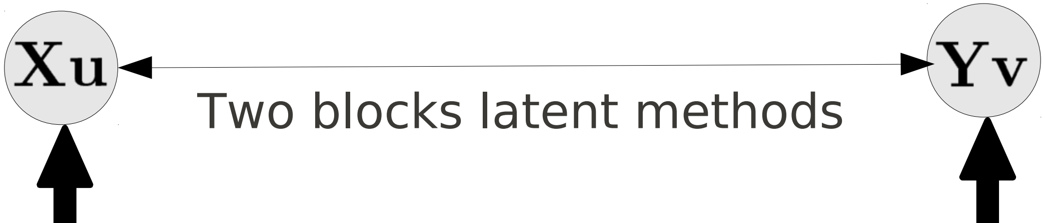
→ Poor stability: difficult to asses

Results: Weights assigned to selected phenotypes

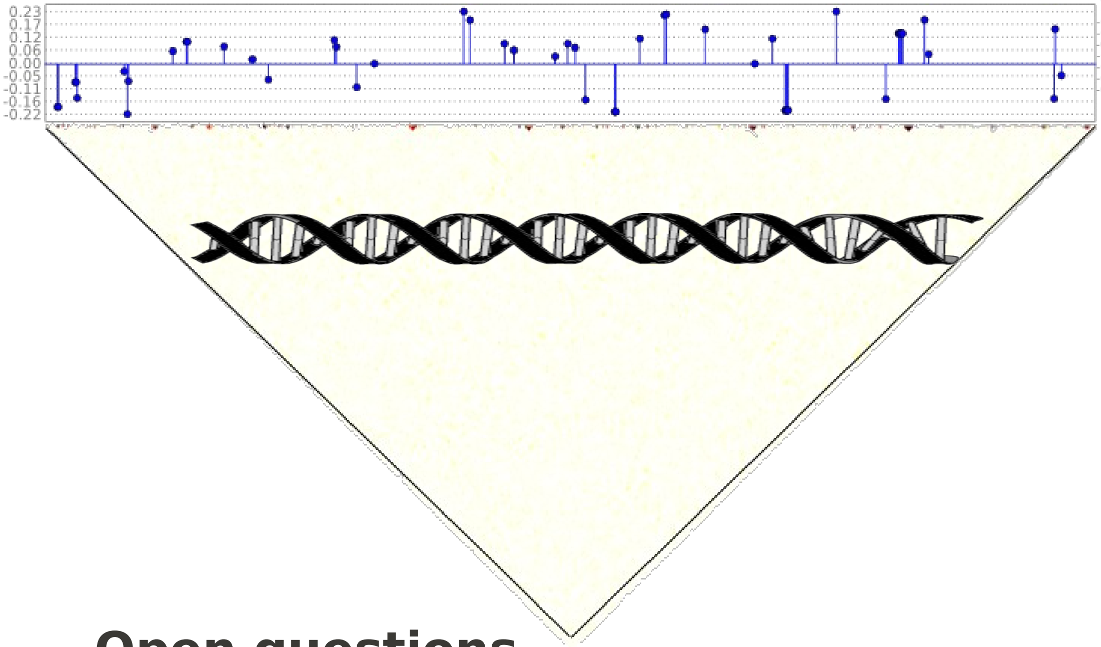


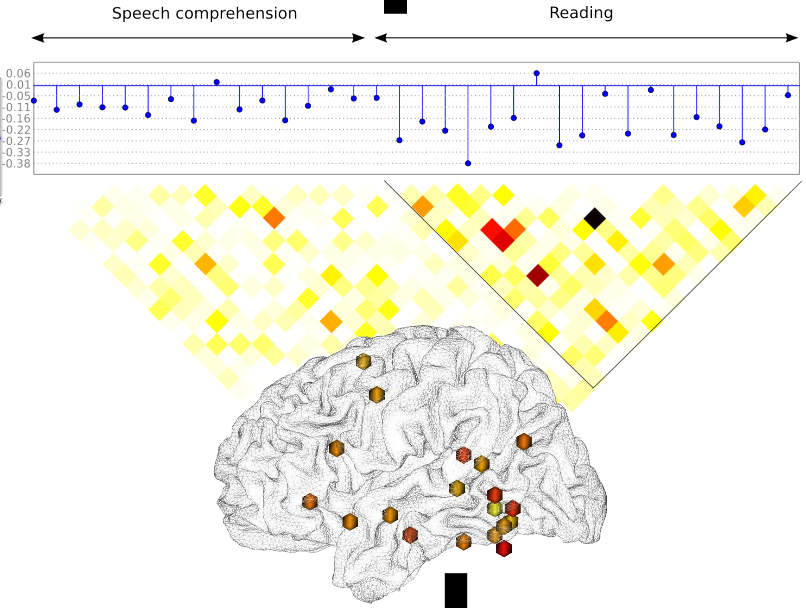


→ *17 selected lateralization phenotypes mainly from the* ***reading*** *task*

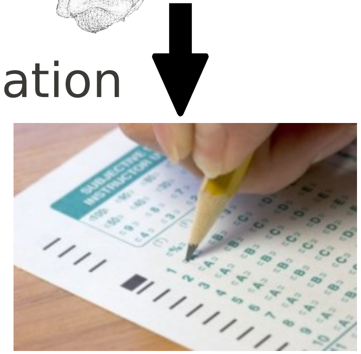


Two blocks latent methods





Classification



Open questions
– Causality
– Structure (gene/ima.)
– Multi-blocks gene.>Ima.>Clinic
...

Thanks

LNAO & Genim program @ NeuroSpin
– Edith Le floch
– Vincent Frouin
– Bertrand Thirion
– JB Poline
– Alexis Barbot
– Denis Rivière

Unicog @ NeuroSpin
– Philippe Pinel
– Stanislas Dehaene

Supelec
– Arthur Tenenhaus
– Laura Trinchera

St-Anne (AP-HP, Descartes)
– Arnaud Cachia

Necker (INSERM, CEA)
– Monica Zibovicius