DE LA RECHERCHE À L'INDUSTRIE

Machine Learning to identify neuroimaging biomarkers in psychiatry

Edouard Duchesnay, PhD, HDR
CEA, NeuroSpin – Université Paris-Saclay – France

Commissariat à l'énergie atomique et aux énergies alternatives
www.cea.fr

Outline

1) Introduction: Machine learning to identify prognostic signature in psychiatry

2) Fighting over-fitting: Feature selection, regularization and model selection

3) Spatial regularization to improve interpretability

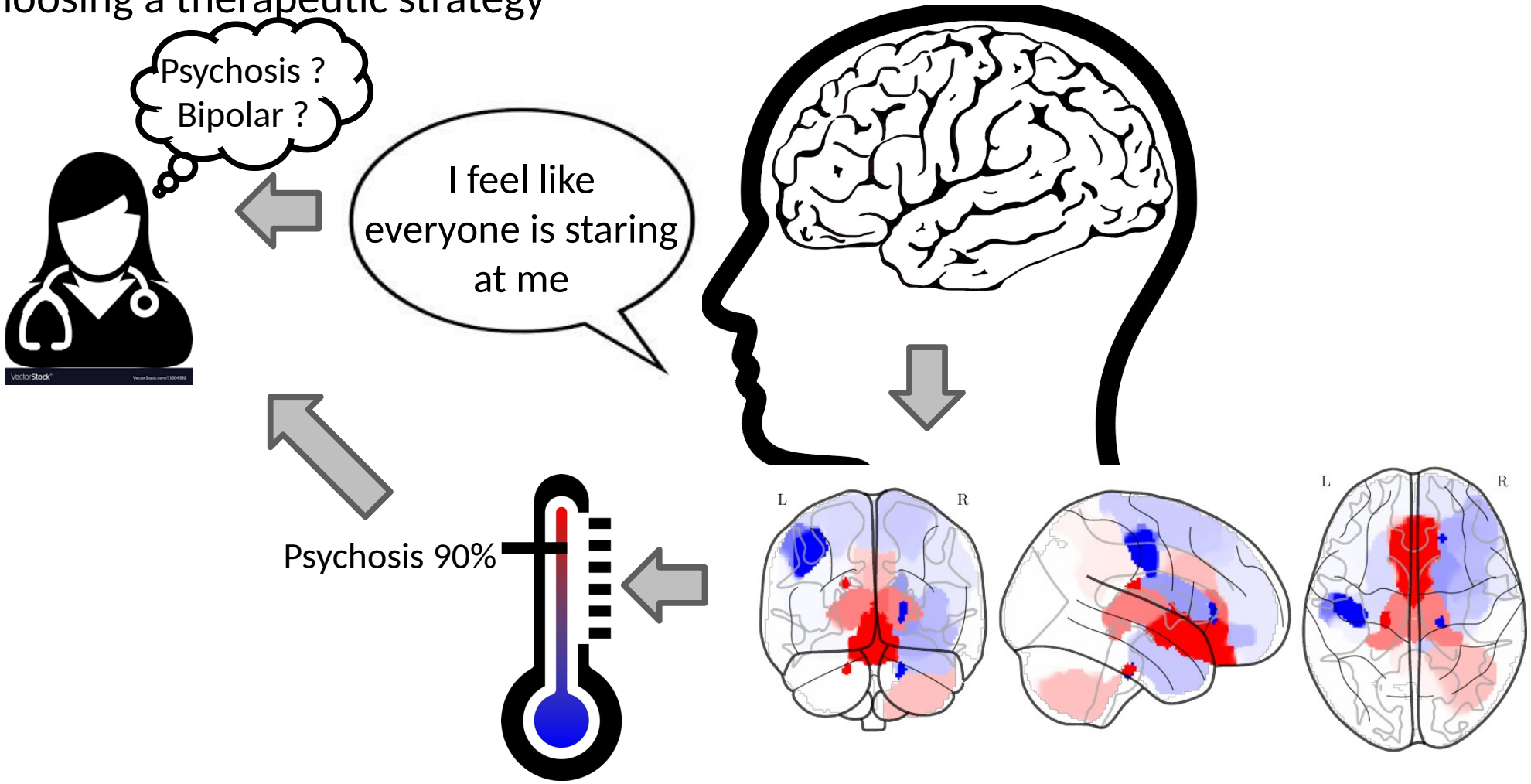
4) Applications to psychiatry

5) Deep learning for supervised task

6) Transfer learning and representation learning

Introduction: Machine learning to identify prognostic signature in psychiatry

Psychiatry lacks objective quantitative measures (such as blood dosage) to guide clinicians in choosing a therapeutic strategy

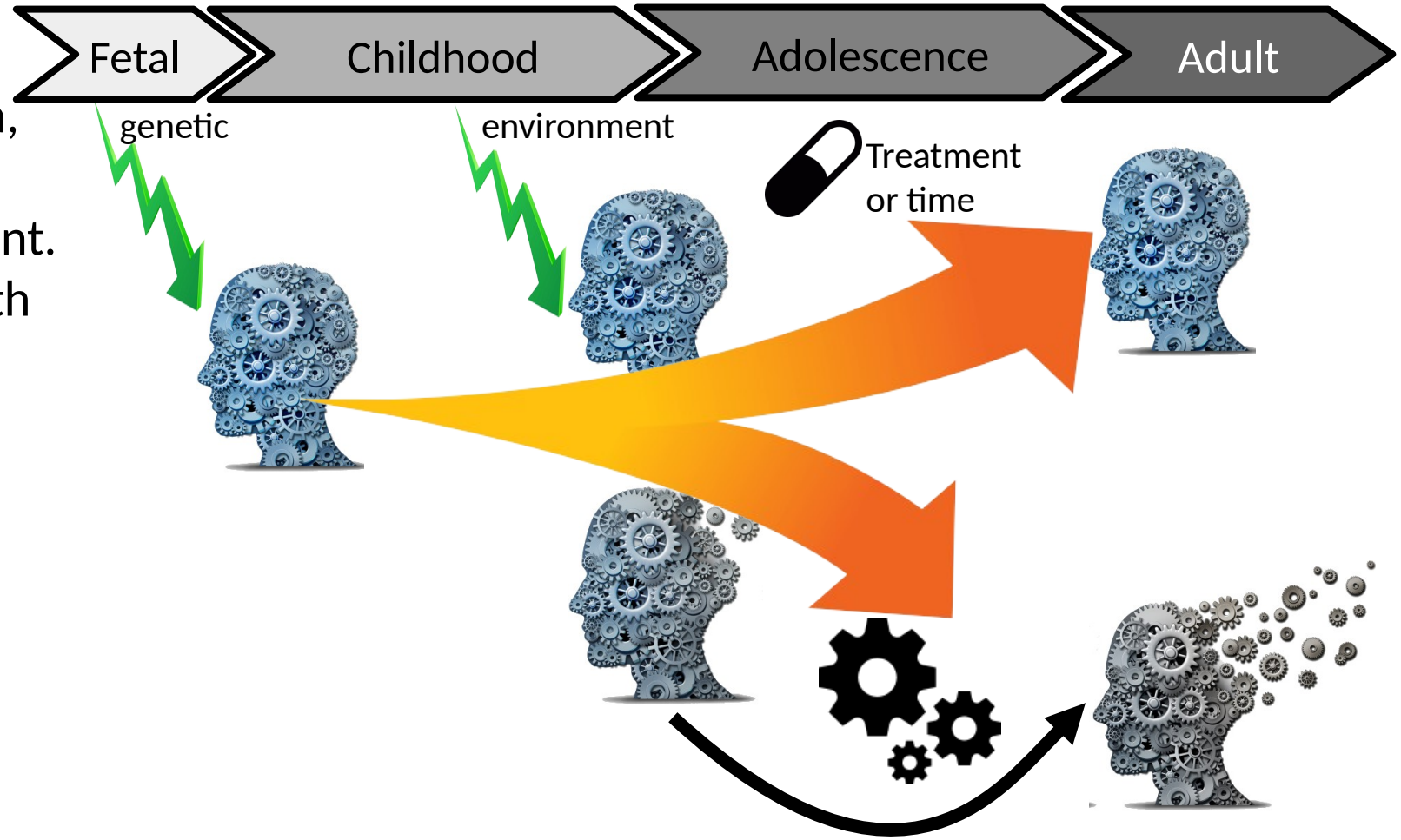


Prognostic brain signatures of clinical course or response to treatment for personalized medicine in psychiatry.

Neuroimaging: biological markers for personalized psychiatry

Rational

- Genetic and environment (trauma, stress, toxic, alcohol) modify the trajectory of the brain development.
- Subtle differences measurable with neuroimaging



Applications

- Transition to psychosis in subjects at risk
- Response to Lithium in Bipolar Disorder

Goal

- Capture those brain patterns with machine learning
- Learn to predict the outcome from past brain scans at early stage of the disorder progression

Outline

1) Introduction: Machine learning to identify prognostic signature in psychiatry

2) Fighting over-fitting: Feature selection, regularization and model selection

3) Spatial regularization to improve interpretability

4) Applications to psychiatry

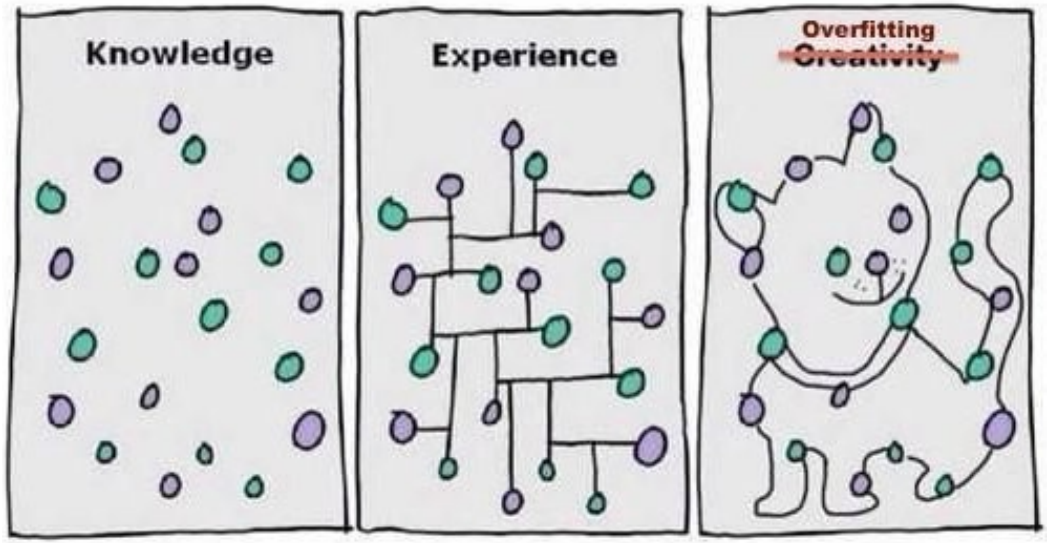
5) Deep learning for supervised task

6) Transfer learning and representation learning

Balance underfitting and overfitting

Too many (300,000) neuroimaging variables

- Avoid spurious correlations: *"With my red socks I get good results in my exams"*
- Find true correlations: *"By studying I get good results"*

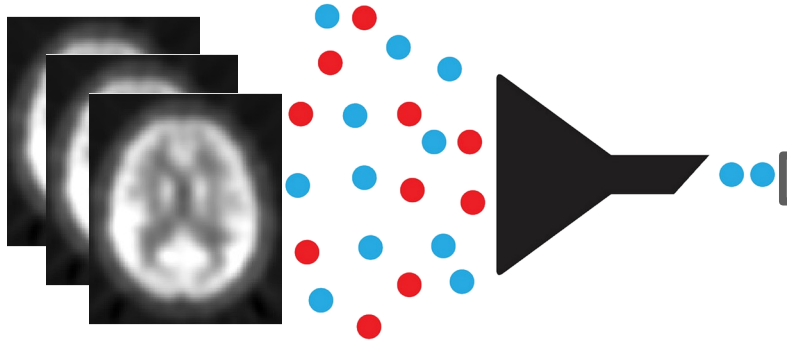


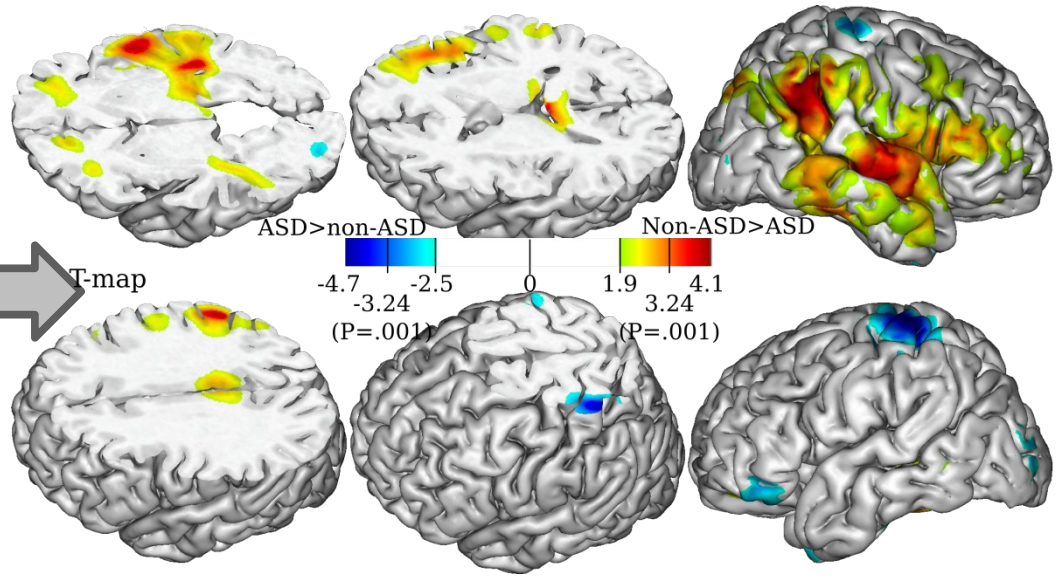
The old good time of *ad hoc* pipelines: variable selection (filtering) + regions extraction + model selection + classification

Application: Prediction of subjects with autism (ASD) on PET imaging

PET images

1) Univariate filter

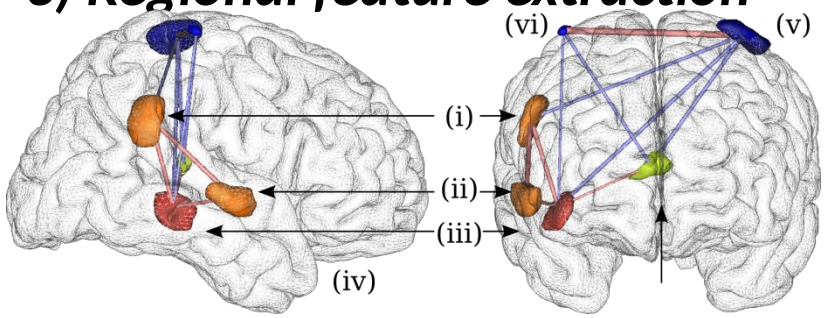




Classification
- ***All: 88%****
- ***ASD: 91% (41/45)****
- ***Non-ASD: 77% (10/13)****

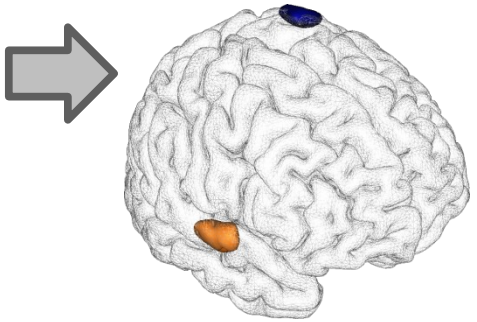
Discriminative pattern
- Hyper. in STS
- Hypo. in postcentral

3) Regional feature extraction



4) Feature subset ranking

5) Model selection

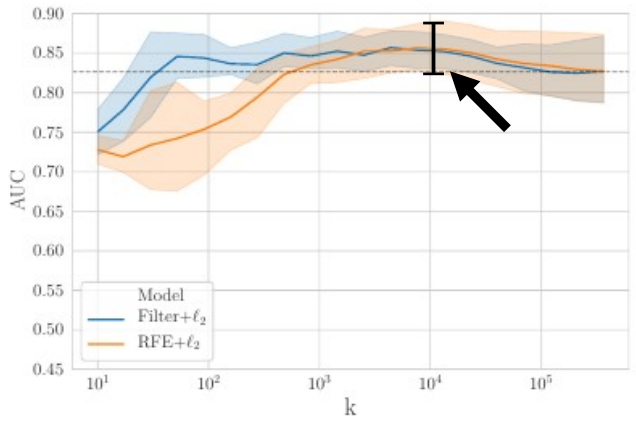


[Duchesnay et al. Neuroimage 2011, IEEE TMI 2007]

Feature selection with whole brain VBM (Gray Matter maps 360 000 voxels): Mixed results

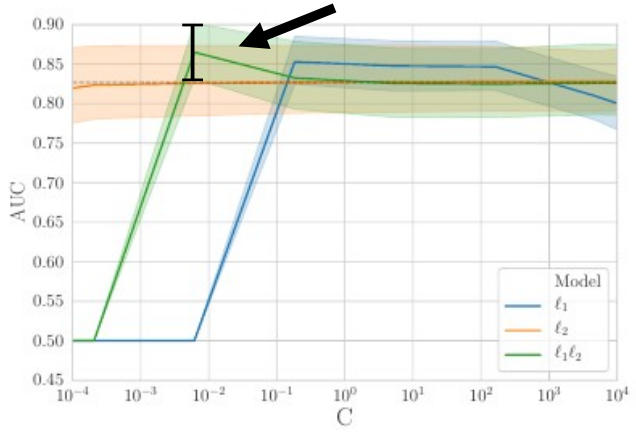
N=605 participants: 330 Controls vs 275 with patients with chronic schizophrenia.

(a) Filter vs. RFE + LR



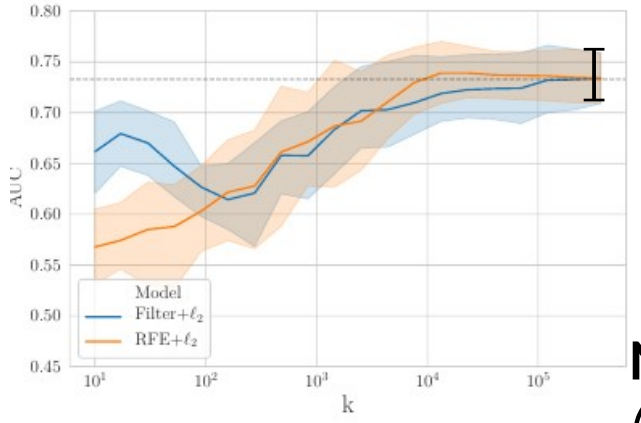
Improvement +5%(vs L2)
"Detectable" with CV

(c) ℓ_1, ℓ_2, ElasticNet LR



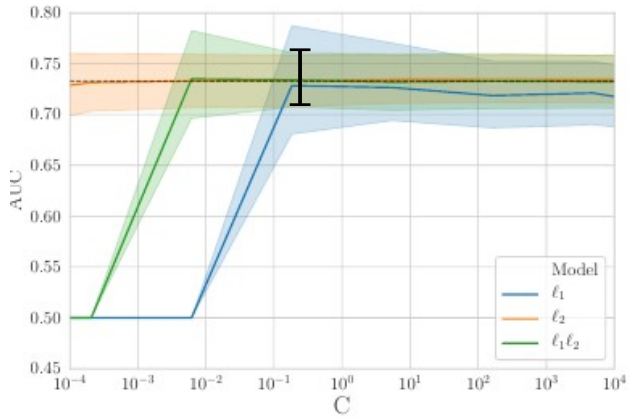
N=662 participants: 356 Controls vs 306 patients with Bipolar Disorder (BD).

(b) Filter vs. RFE + LR



No improvement (vs L2)

(d) ℓ_1, ℓ_2, ElasticNet LR

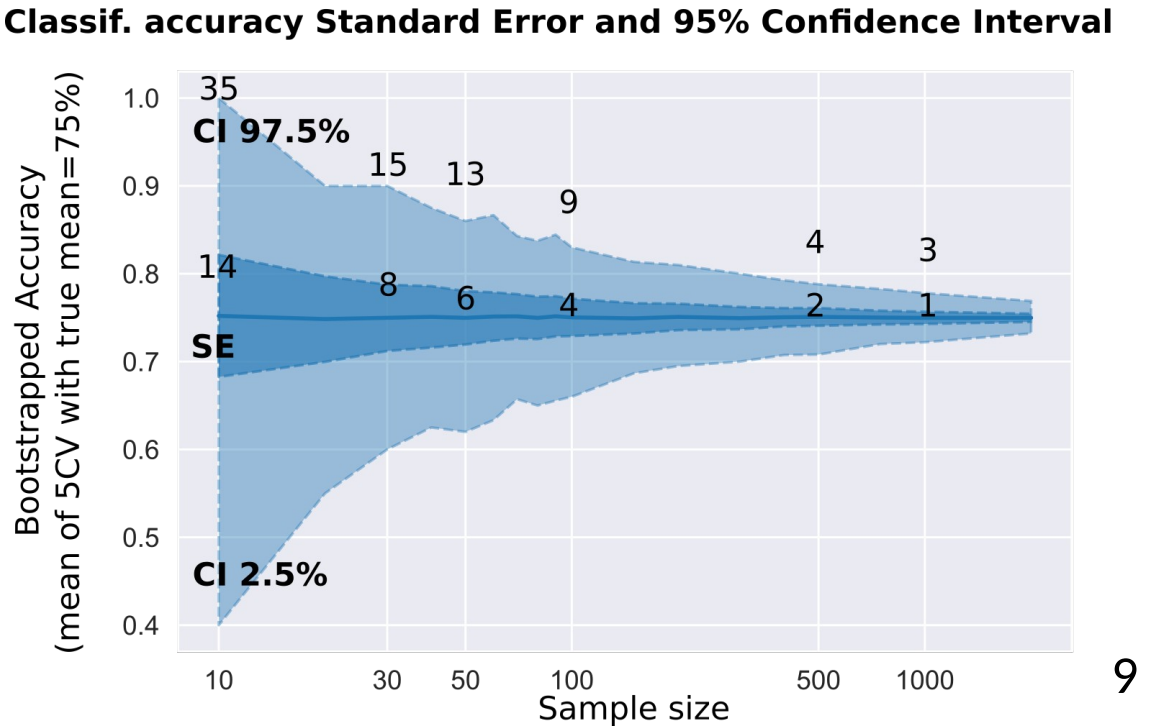


Conclusion on sparse models and feature selection

- "Venerable" L2-regularization provides a baseline performances.
- Basic univariate Filter/Lasso or ElasticNet can help
- **Model selection is the main issue**
- For large dataset N~500, Cross-Validation (CV) is efficient
- For small dataset <200
  - CV has a large variance
  - In-sample estimates provide promising results

N	$\mathrm{SE}(p)$
30	10%
100	5%
500	2%

Theoretical standard errors of classification rate



Outline

1) Introduction: Machine learning to identify prognostic signature in psychiatry

2) Fighting over-fitting: Feature selection, regularization and model selection

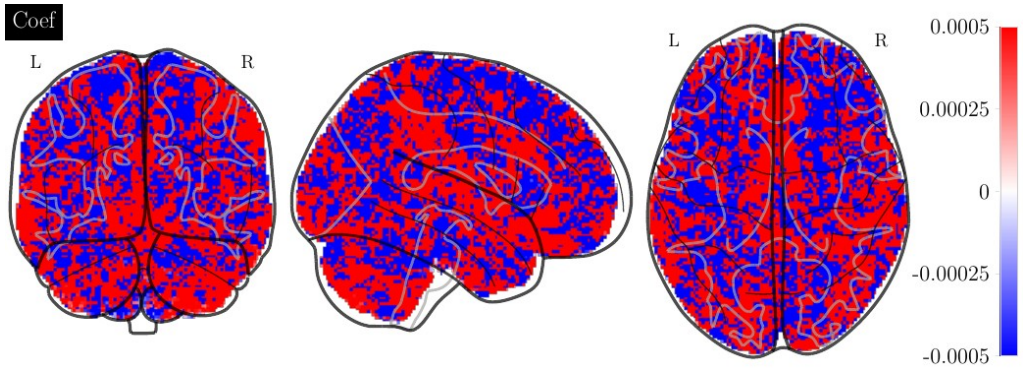
3) Spatial regularization to improve interpretability

4) Applications to psychiatry

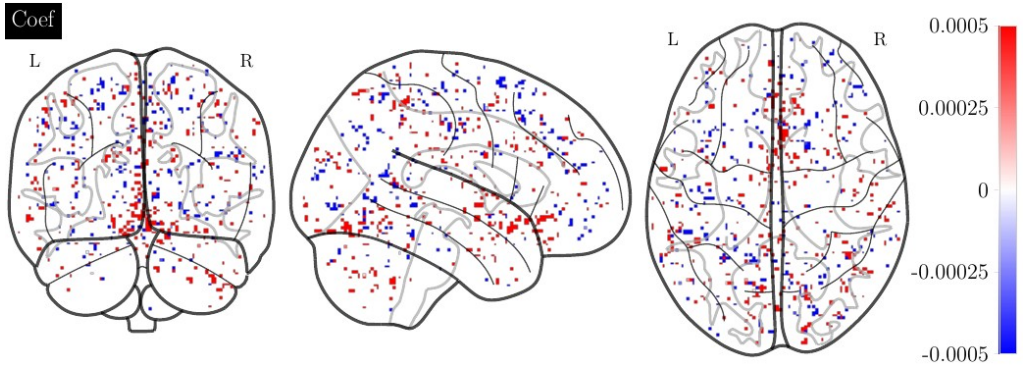
5) Deep learning for supervised task

6) Transfer learning and representation learning

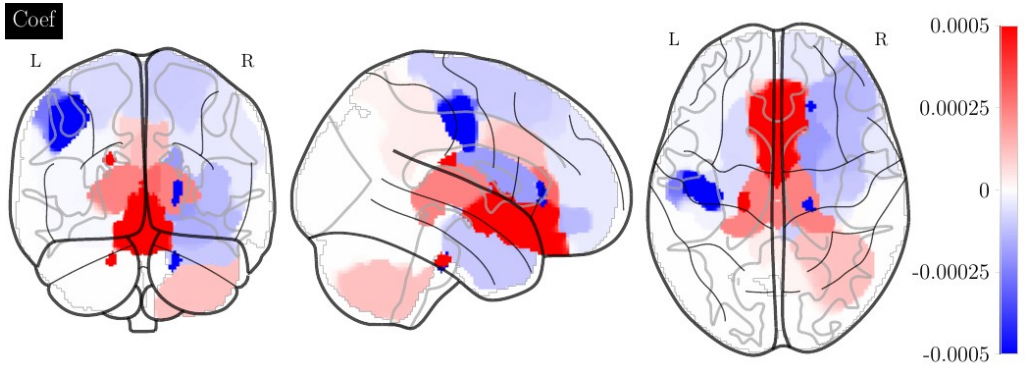
Spatial regularization



l2-regularization



l1-regularization



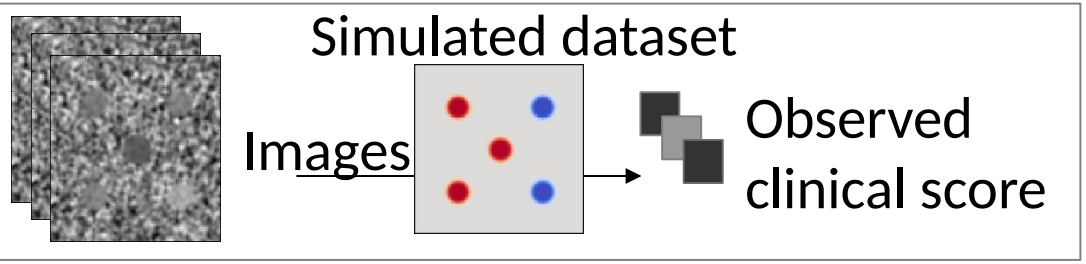
Predict transition to psychosis in at-risk adolescents
Collab. with MO Krebs and A. Iftimovici, Ste-Anne, Hosp. Paris [to be published]

Design of a solver for convex minimization problem with non-smooth terms:
– [Hadj-Selem 2018 IEEE TMI]
– Python library

Add spatial regularization:

– "spatial smoothness" or
– solution should be organized as regions

Effect of spatial regularization



Classical univariate associations
(GWAS, VBM) cor(X^j, y) for each locus j

Multiple regression with constraint (penalty): shrink **coefficients**
⇒ **Ridge regression (or SVM)**

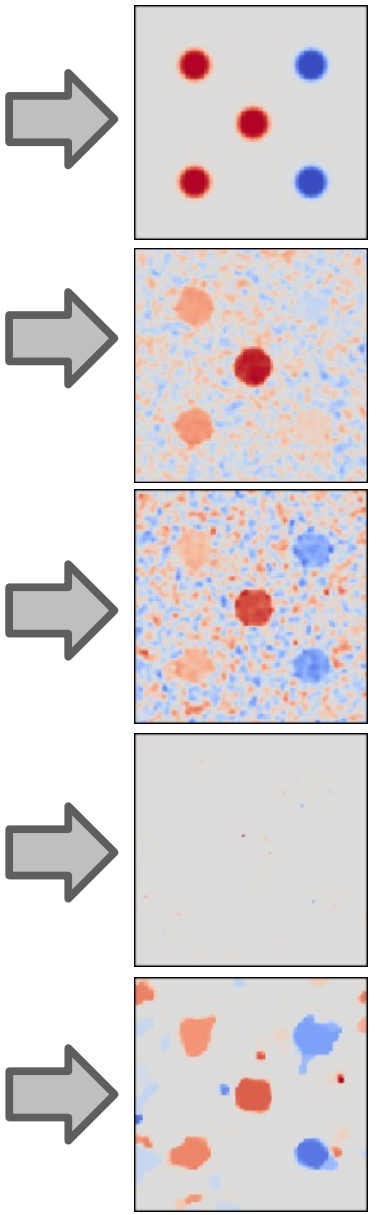
$$\mathcal{L}(\mathbf{w}) + ||\mathbf{w}||_2^2$$

More constraints: **many coefficient should be null:**
⇒ **Elasticnet**

$$\mathcal{L}(\mathbf{w}) + ||\mathbf{w}||_2^2 + ||\mathbf{w}||_1$$

Biological constraints: **solution should be smooth**
⇒ **Total Variation**

$$\mathcal{L}(\mathbf{w}) + ||\mathbf{w}||_2^2 + ||\mathbf{w}||_1 + TV(\mathbf{w})$$



True biomarkers

- No prediction
- Overlook information

+ Prediction
- Interpretation
(Signal everywhere)

+ Prediction
- Interpretation
(Too sparse)

+ Prediction
+ Interpretation

Spatial regularization GraphNet vs Total Variation

GraphNet: GN (Dohmatob 2015; Grosenick 2013)

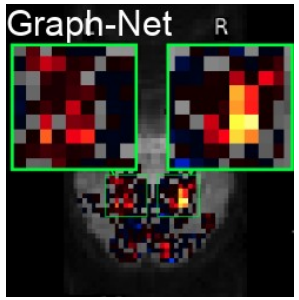
$$\sum_{i,j,k} ||\nabla(\mathbf{w}_{i,j,k})||_2^2 = ||\nabla(\mathbf{w})||_2^2$$

- Squared L2-norm on the gradient of the weight map
- **Promote smooth continuous change**
- Differentiable solved with any proximal gradient method

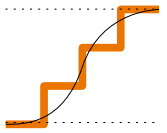
Total-Variation: TV

$$\sum_{i,j,k} ||\nabla(\mathbf{w}_{i,j,k})||_2 = ||\nabla(\mathbf{w})||_{2,1}$$

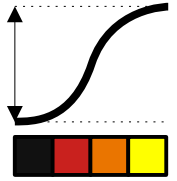
- L12 -norm norm on the gradient of the weight map
- **Segmenting the weight map** into spatially-contiguous parcels with almost constant values

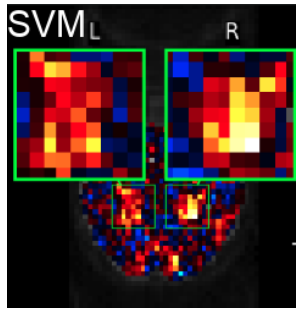


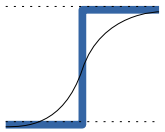
GN: 12=3*2²
TV: 6=3*2



3 steps of size 3







1 step of size 6

GN: 36=1*6²
TV: 6=1*6

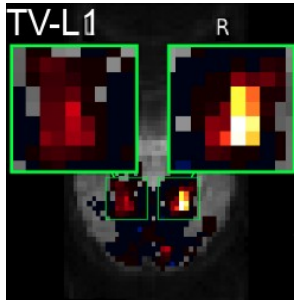


Image: Dohmatob 15

Spatial regularization GraphNet vs Total Variation

Datasets

NUSDAST dataset N = 236

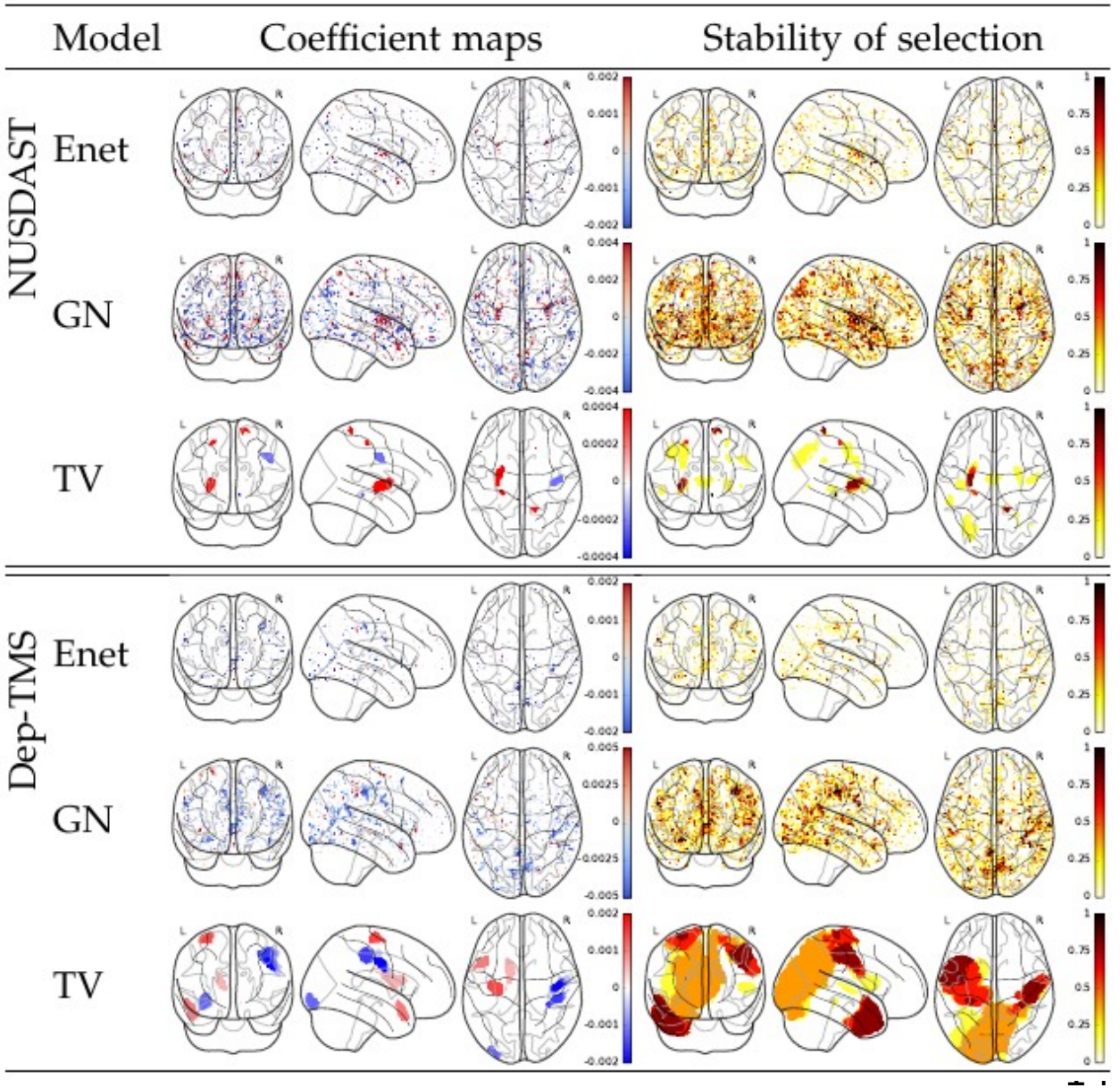
- 97 patients with schizophrenia vs 139 healthy controls.
- Predict clinical status

Dep-TMS dataset N=34

- Patients with treatment-resistant depression.
- 18 responders vs 16 non-responders
- Predict response to transcranial magnetic stimulation (TMS).

TV for spatial regularization

- Similar performances
- More interpretable models: identify regions
- **Improved stability of predictive support**



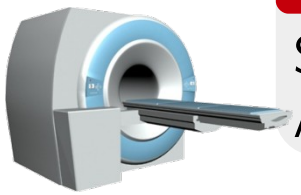
Spatial regularization on 3D images and meshes: improved stability

ADNI dataset N= 201

- 81 patients with with mild cognitive Impairments that converted to Alzheimer disease vs 120 healthy controls elderly subjects.
- VBM 3D GM maps
- Cortical thickness.

- Improved stability of coefficient map
- Similar performances
- We needed **fast solvers for 300 000 features working on meshes of cortical surface**

sMRI



Interpretable and stable predictors of brain diseases

Sparse map: moderate perf., low relevance & stability

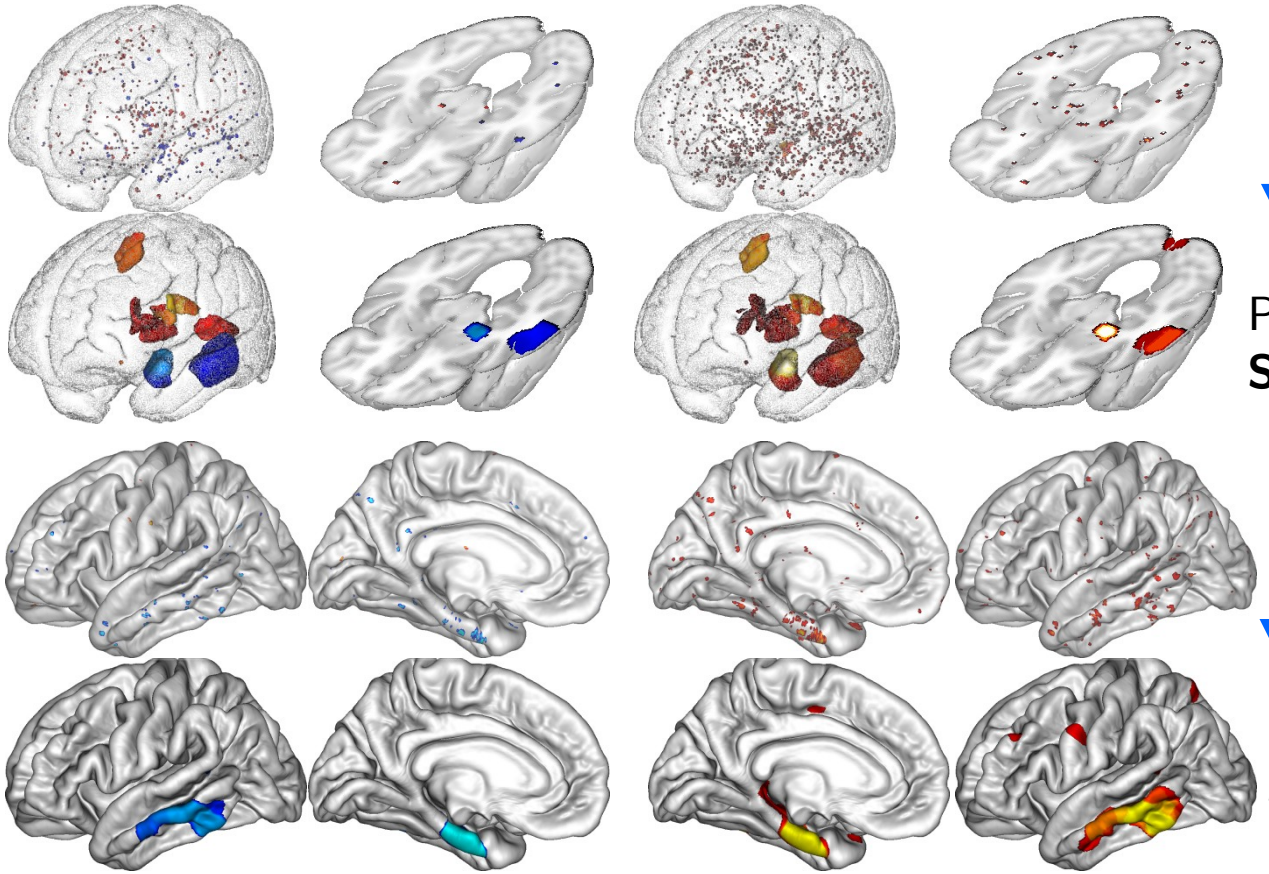
Add **spatial constraints**: improve perf., relevance & stability

Clinic

Predictive map | Stability of the map

3D images: Enet | Spatial(TV)

2D meshes: Spatial(TV) Enet



Application to Alzheimer's Disease aMRI

Perf.:+5%

Stability:+225%

Perf.: +8%

Stability: +510%

Reformulating TV as a linear operator

Reformulate the structured penalty as a product with a linear operator

$$\mathrm{TV}(\boldsymbol{w}) \equiv \sum_{i,j,k} \left\| \nabla \left(\boldsymbol{w}_{\phi(i,j,k)} \right) \right\|_2$$

$$\nabla \left(\boldsymbol{w}_{\phi(i,j,k)} \right) \equiv \underbrace{\begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{A}'_{\phi}} \underbrace{\begin{bmatrix} \boldsymbol{w}_{\phi(i,j,k)} \\ \boldsymbol{w}_{\phi(i+1,j,k)} \\ \boldsymbol{w}_{\phi(i,j+1,k)} \\ \boldsymbol{w}_{\phi(i,j,k+1)} \end{bmatrix}}_{\boldsymbol{w}'_{\phi(i,j,k)}}.$$

$$\mathrm{TV}(\boldsymbol{w}) = \sum_{i,j,k} \left\| \boldsymbol{A}_{\phi(i,j,k)} \boldsymbol{w} \right\|_2$$

$\boldsymbol{A}$ Is a huge *3 x P* sparse matrix

General and flexible formulation can be adapted:
– other penalties: group lasso
– other input feature meshes

ElasticNet with structured penalty (TV, Group lasso,)

$$\min_{w} f(w) = \overbrace{L_\varepsilon(w) + \lambda_2 \|w\|_2^2}^{\text{smooth}} + \overbrace{\lambda_1 \|w\|_1 + \lambda_s \sum_{g \in \mathcal{G}} \|A_g w\|_2}^{\text{non-smooth}},$$

Any differentiable Loss

Any function with known proximal operator

$s(w)$ structured penalty: TV, Group lasso, etc.
Unknown proximal operator
Any function formulated as a norm of product

Problem

The proximal operator of TV (several useful non-smooth penalties) are mostly not known or difficult to compute

Strategies

1)Approximate the proximal operator of TV [Beck and Teboulle 2009, Schmidt et al., 2011]

2)Smooth TV [Nesterov, 2005, Chen et al. 2012]

CONESTA solver: Nesterov smoothing with continuation steps guided by the duality gap

Nesterov smoothing of the complex term

We apply Nesterov smoothing technique to approximate $s(\boldsymbol{w})$, of parameter μ, such that:
$s_\mu(\boldsymbol{w}) \leq s(\boldsymbol{w}) \leq s_\mu(\boldsymbol{w}) + \mu P/2$, where P is the number of features.

Problem

$s_\mu(\boldsymbol{w})$ is smooth, the problem can be minimized with FISTA (Beck and Teboulle, 2009).

- **Large smoothing fast convergence but low precision**
- **Small smoothing slow convergence with high precision**

Contributions

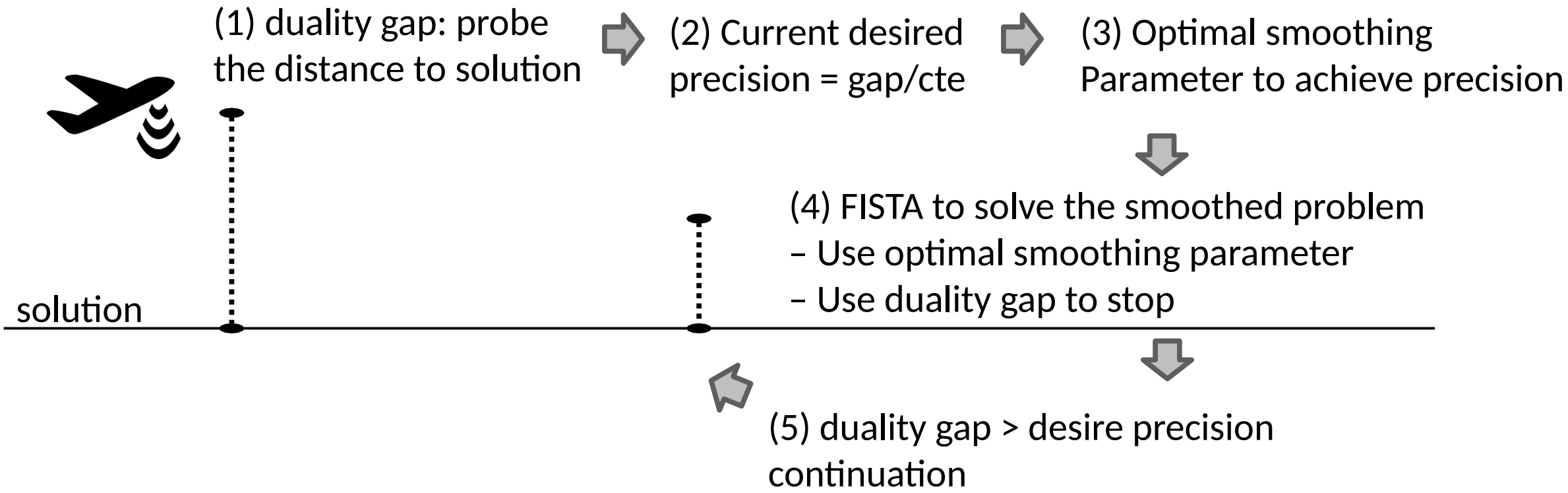
1) **Continuation procedure**: Start with large smoothing and decrease while approaching the solution
2) Duality gap to estimate the **distance to the solution**
3) Expression of **optimal smoothing parameter given a prescribed precision**

CONESTA solver: the algorithm

Principles: smooth touchdown procedure

- Solve a smoothed problem with dynamic reduction of the smoothing.
- Duality gap to probe the distance to the ground (global optimum).
- Dynamically adapts its speed (the smoothing parameter) according to this distance.

Algorithm



Solve the problem with **300k features in 30 minutes**

Outline

1) Introduction: Machine learning to identify prognostic signature in psychiatry

2) Fighting over-fitting: Feature selection, regularization and model selection

3) Spatial regularization to improve interpretability

4) Applications to psychiatry

5) Deep learning for supervised task

6) Transfer learning and representation learning

Identifying a Neuroanatomical Signature of Schizophrenia

Learn a predictive signature that generalize across sites and stages of the disorder

90 Controls

330 Controls

43 first episode of psychosis

Disorder progression

272 patients with Chronic SCZ

76% of prediction accuracy

68% of inter-site accuracy

Learn predictive signature

Predictive signature

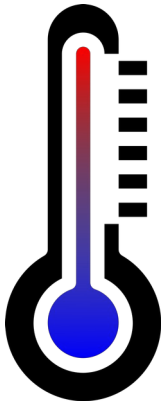
Apply to new patients

Brain score

[De Pierrefeu Acta Psychiatrica Scandinavica, 2018]

Collab with

- CHU Lille (R. Jardri)
- SHU Ste-Anne Paris (M.O. Krebs)
- CHU Créteil (J. Houenou)



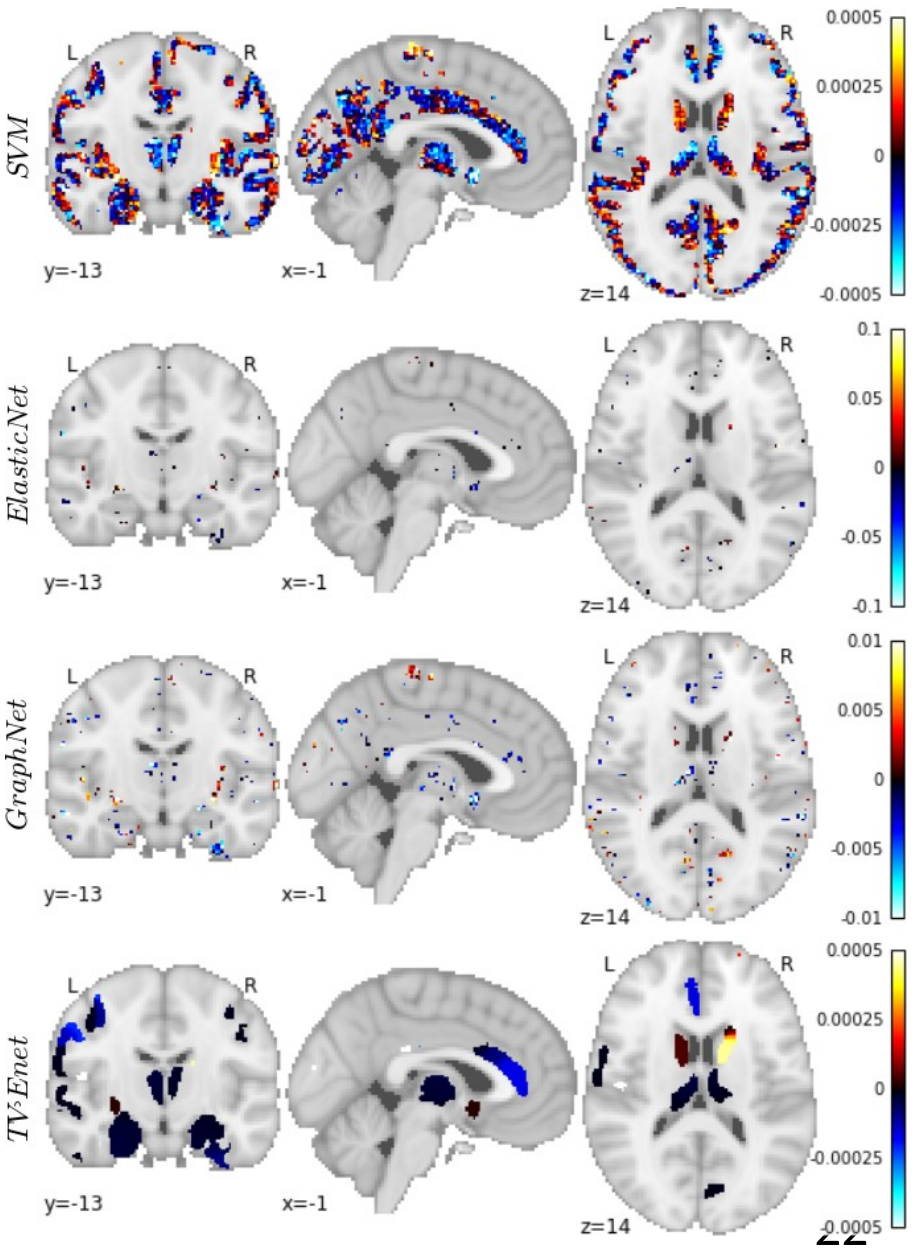
Identifying a Neuroanatomical Signature of Schizophrenia

Prediction performances on an independent cohort of controls and patients with first-episode psychosis.

Features	Model	AUC	bAcc	r_w
Gray Matter VBM	SVM	0.78	0.71	-
	Enet	0.78	0.73	0.34
	GraphNet	0.79	**0.76**	0.42
	TV-Enet	**0.80**	**0.76**	0.74
Vertex based cortical thickness	SVM	0.68	0.64	-
	Enet	0.65	0.62	0.09
	GraphNet	0.63	0.60	0.19
	TV-Enet	0.67	0.62	**0.76**
ROIs based volume	SVM	0.72	0.66	-

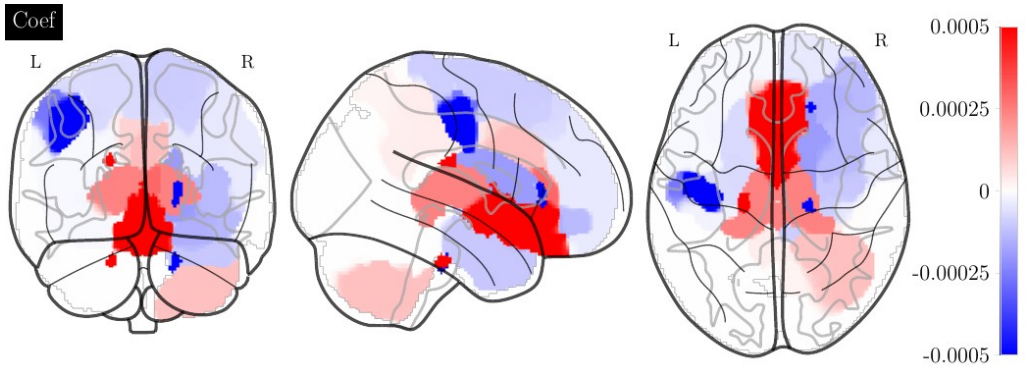
- **VBM outperforms** ROI and cortical thickness.
- All classifiers provide **similar performance.**
- **TV-Enet**: sparse and **stable** solution organized as **regions.**
- Signature of SCZ, shared by a majority of patients **across** different **sites** and already present at the **early stage** of the disorder.
- Signature provides a brain score is associated with the symptoms severity and the amount of cognitive deficit.

[De Pierrefeu Acta Psychiatrica Scandinavica, 2018]

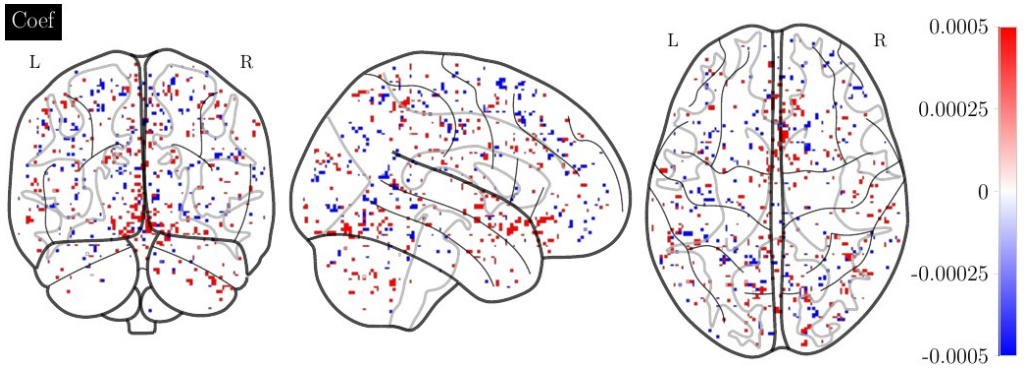


Preliminary experiments: Predict psychotic transition using anatomical imaging in prodromal (Ultra High Risk, UHR) subjects

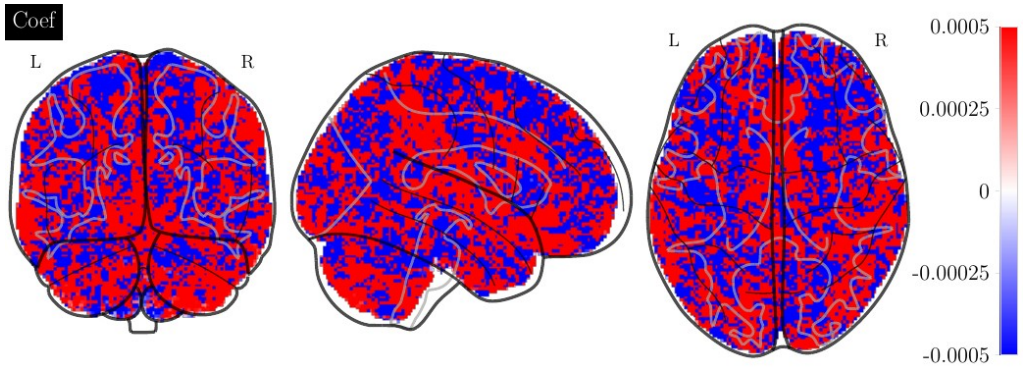
TV-L1L2
AUC:
0.79



L1
AUC:
0.7

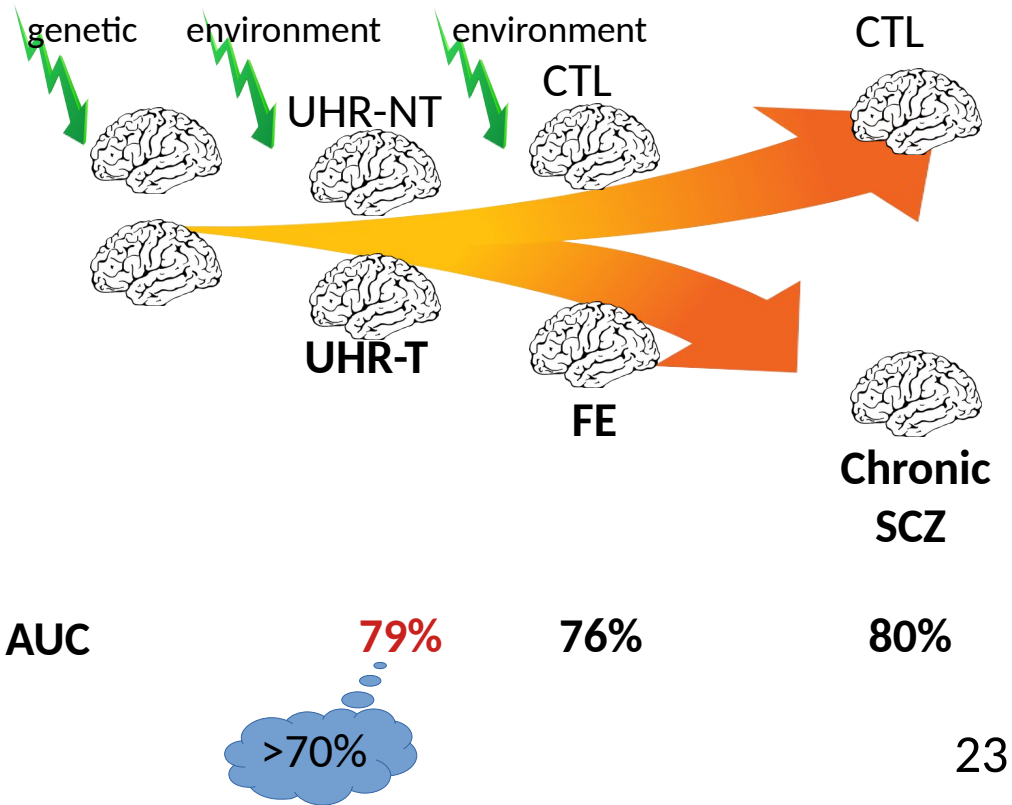


L2
AUC:
0.66



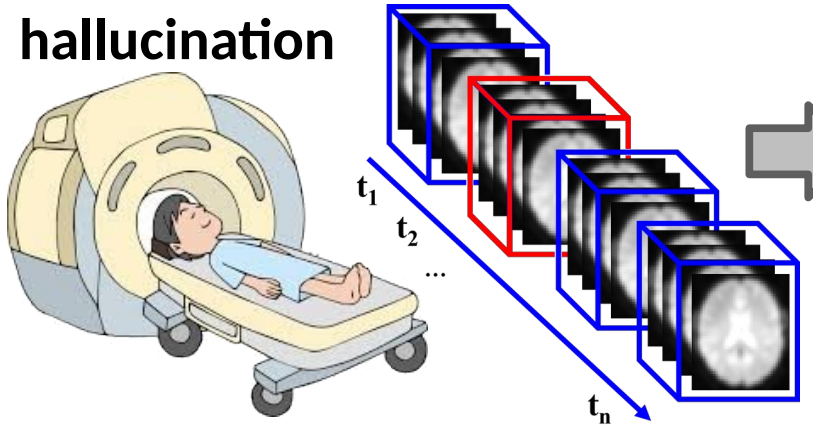
82 UHR participants
- NT=53
- T=29

Collab. Ste-Anne:
Anton Iftimovici
MO Krebs

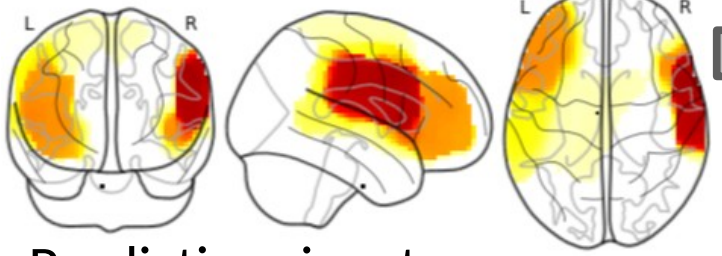


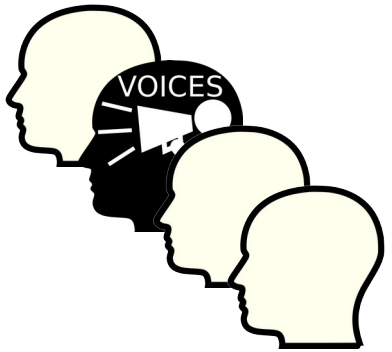
Functional MRI activations predict hallucinations in Schizophrenia

1) Rest functional MRI during hallucination



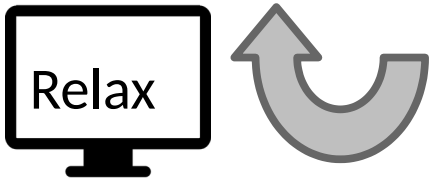
2) Learn functional signature of hallucination





Predictive signature:
Broca's area and its right homologue

3) Decoding
4) Neuro-feedback



[De Pierrefeu Hum. Brain Map., 2018]
Collab with CHU Lille (R. Jardri)

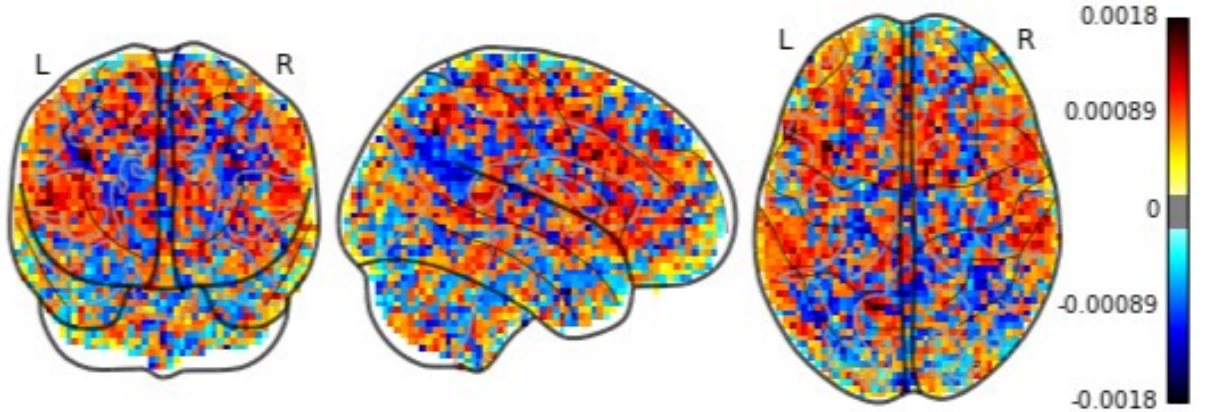
Functional MRI Activation Patterns to Predict hallucinations in Schizophrenia for neurofeedback

Inter-subject decoding performances

Model	AUC	bAcc	Spe	Sen
SVM	0.73*	0.73	0.78	0.67
TV-Enet	0.79*	0.74	0.76	0.71

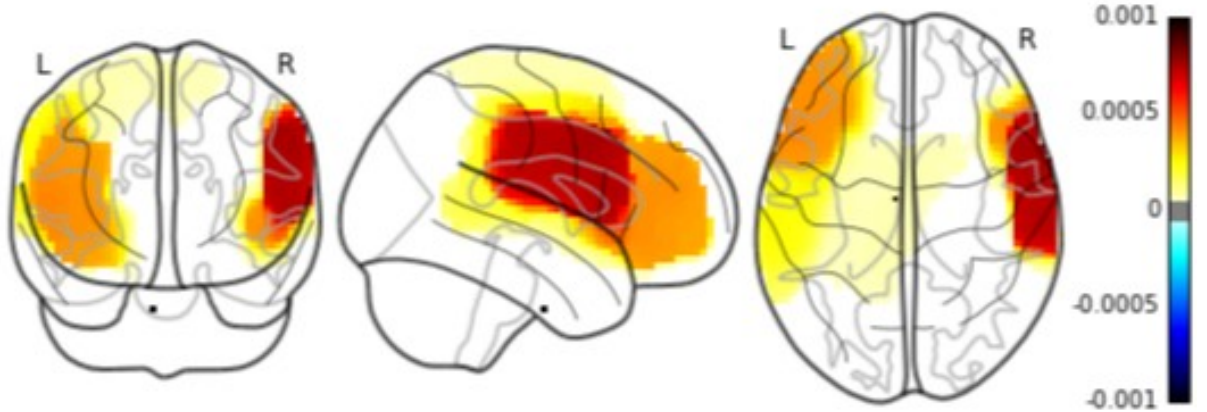
Linear SVM

A



TV-Enet

B



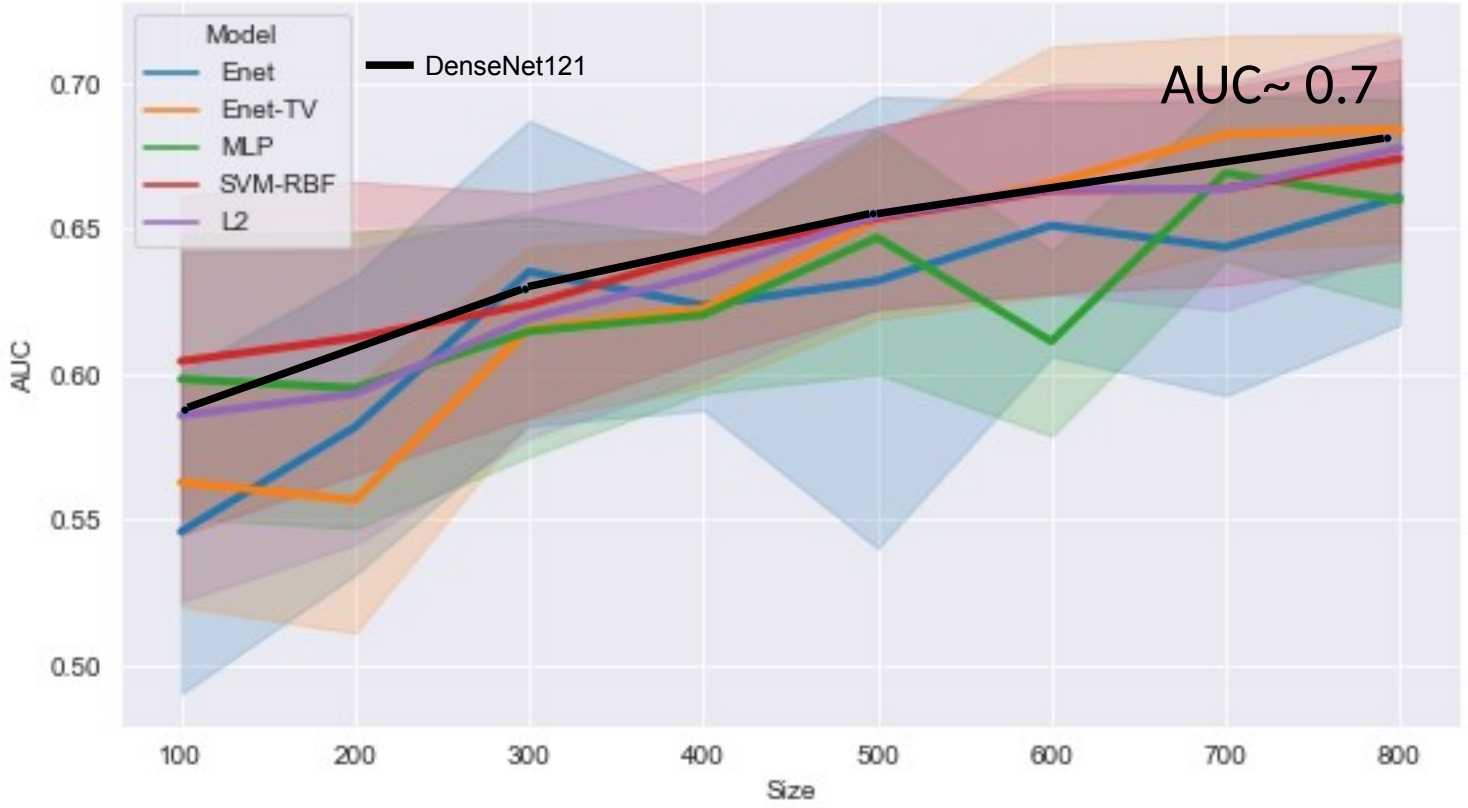
- Predictive signature: right homologue of Broca's area
- Perspective with near-infrared spectroscopy (NIRS)

[De Pierrefeu Hum. Brain Map., 2018]
Collab with CHU Lille (R. Jardri)

Prediction of Bibolar Disorder on 978 participants (423 BD vs 555 HC) from 13 sites

Learning curve: Leave-One-Site-Out CV

Collab with J. Houenou



- Performances still increase with sample size
- Non-linear and linear models provided similar performances (including DenseNet). Differences < Standard errors
  - Possibility to capture non-linear pattern?
  - Possibility to capture subgroups of patients/controls?

Interpretability of predictive models

Interpretability is not always a key issue (see "The Great AI Debate" – NIPS2017). In most situations, one would prefer a complex "black-box" that outperform an interpretable model.

The situation:

1) **Complex models perform like linear models**.

2) **Interpretability** is essential in the context of **clinical (psychiatric) applications**:

- Underpinning physiopathological process.
- Drug development (brain region targets).
- Brain stimulation (brain region targets).
- ML aims to guide treatment strategy, this choice must be justified by biological evidence.

Interpretability revisited:

3) Identify stable predictive regions.

4) Characterize the function of the regions:

- Directly linked (causal or consequence) with the disorder.
- Proxy of some important information (treatment, disease duration, etc.)
- "Suppressor variable" that adjusts the model for latent variables (ex. Age, sex,).

Interpretability of predictive pattern 1/2: Identify stable regions

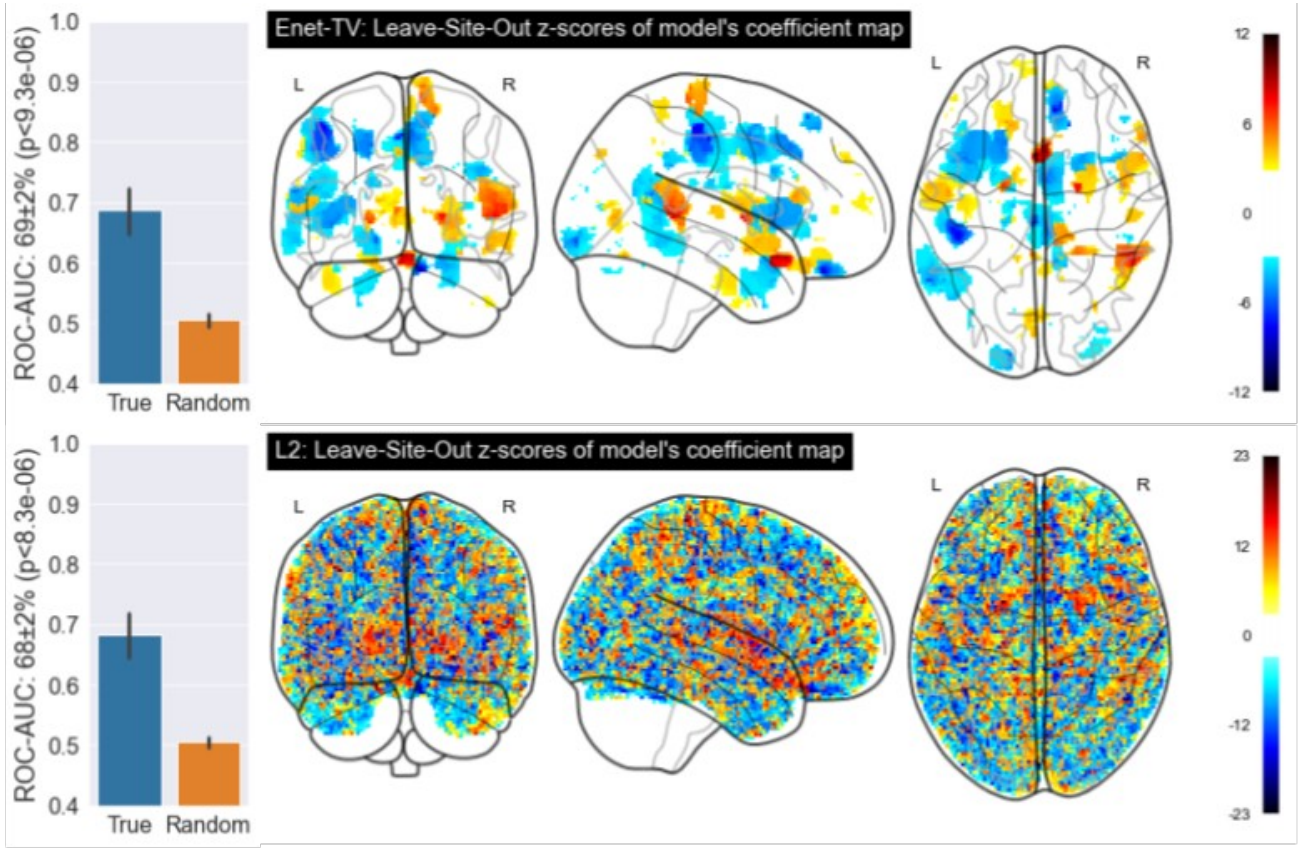
Dataset: Prediction of Bipolar disorder with 978 participants in 13 recruitment sites.

Aims: find stable regions

- Z-score maps across leave-site-out cross-validation highlight stable regions identified by TV

Spatial (Enet-TV) regularization

l2-regularization



Interpretability of predictive pattern
2/2: Group regions into patterns and feature importances

Aims: group regions into more global patterns and evaluate importance/redundancy/complementary

- Ablation study (Loss of ROC-AUC)

Whole pattern Loss of AUC		Sum of AUC losses
-0.16	**>**	**-6**

Empowerment among regions: **the whole is bigger than the sum of its parts**.

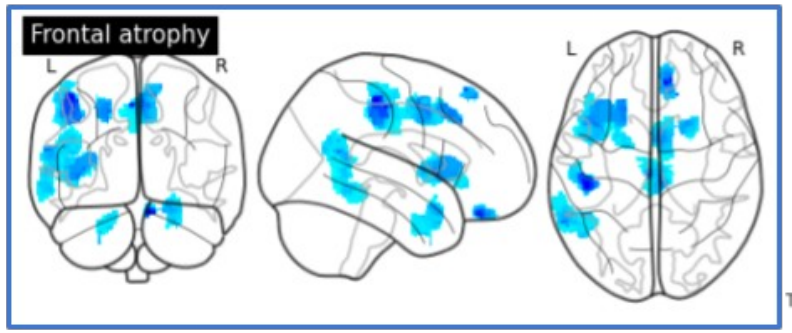
-0.03	=	-0.03

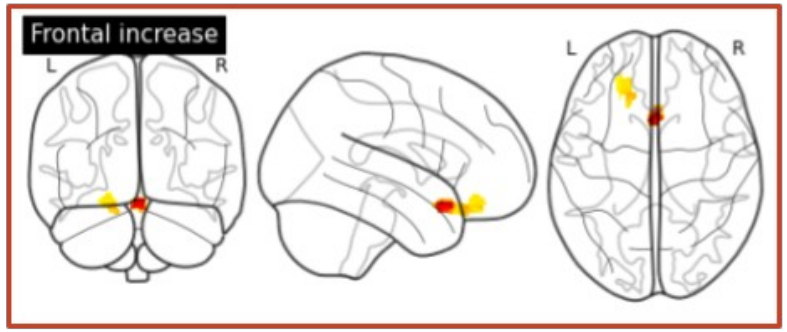
Simple additive effect

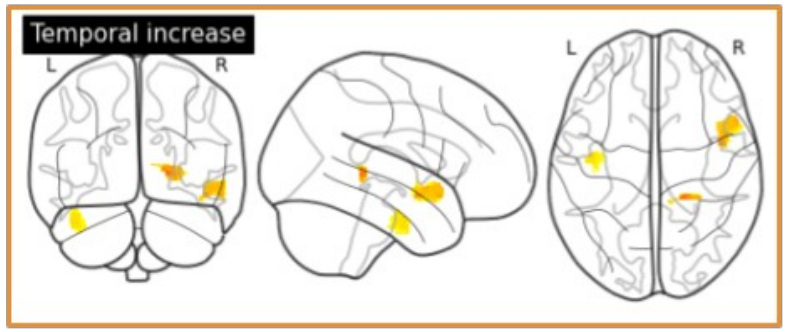
-0.07	<	-0.09

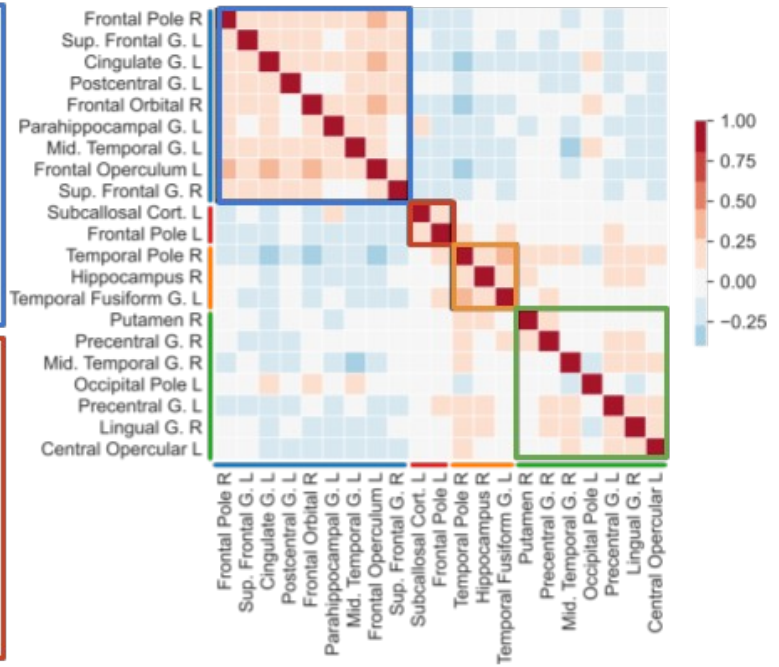
Redundancy

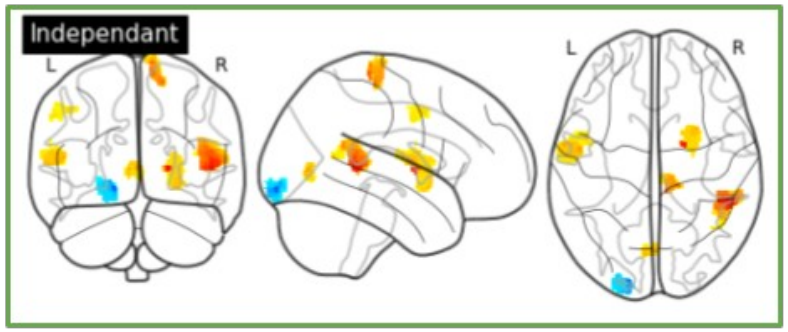
Explored other tracks, LARS-paths, etc.











-0.05	=	-0.05

Simple additive effect

Interpretability of multivariate models, ie, do methodologists design useful models?

Multivariate models are worth: “the whole is bigger than the sum of its parts”.

Ok but, to be useful (for clinical practice) models should address the “role” of a brain region given multivariate model?

Tools: Feature importance, Bootstrapping, ablation study, GradCam, etc.

Remains an open challenge: Indeed proposed solution have unsatisfactory methodological soundness: *a posteriori* exploration of solution will introduce bias

- Validation on an independent data set (need more data!)
- In-depth exploration with clinical conditions (medications, age, disorder severity, dimensions)
- In-depth exploration of the role of brain regions:
  - Directly (or proxy) associated with the disorder
  - “Suppressor variable” that adjusts the model for latent variable (ex. Age)

In conclusion, it is a laborious task, full of potential bias, for which good practices must be defined

Outline

1) Introduction: Machine learning to identify prognostic signature in psychiatry

2) Fighting over-fitting: Feature selection, regularization and model selection

3) Spatial regularization to improve interpretability

4) Applications to psychiatry

5) Deep learning for supervised task

6) Transfer learning and representation learning

Deep learning?

General idea

1) Deep learning does not require feature extraction: compare minimally processed “quasi-raw” data traditional feature extraction (here VBM).
2) Deep learning should outperformed other models.
   - Bad arguments: Because it's trendy and I like it… ;). CNN deals with translation, homothety, rotation.
   - Good ones: Because world is compositional, ie, multi-scale, non linearity, subgroups of subjects, etc.

Controversial findings

1. DL outperformed linear models.
2. Feature extraction is not necessary.

- Abrol… Calhoun (2021). Nature communications.
- Cole… Montana, G. (2017) NeuroImage.
- …

VS.

1. DL and linear models produce similar performances.
2. Feature extraction is necessary.

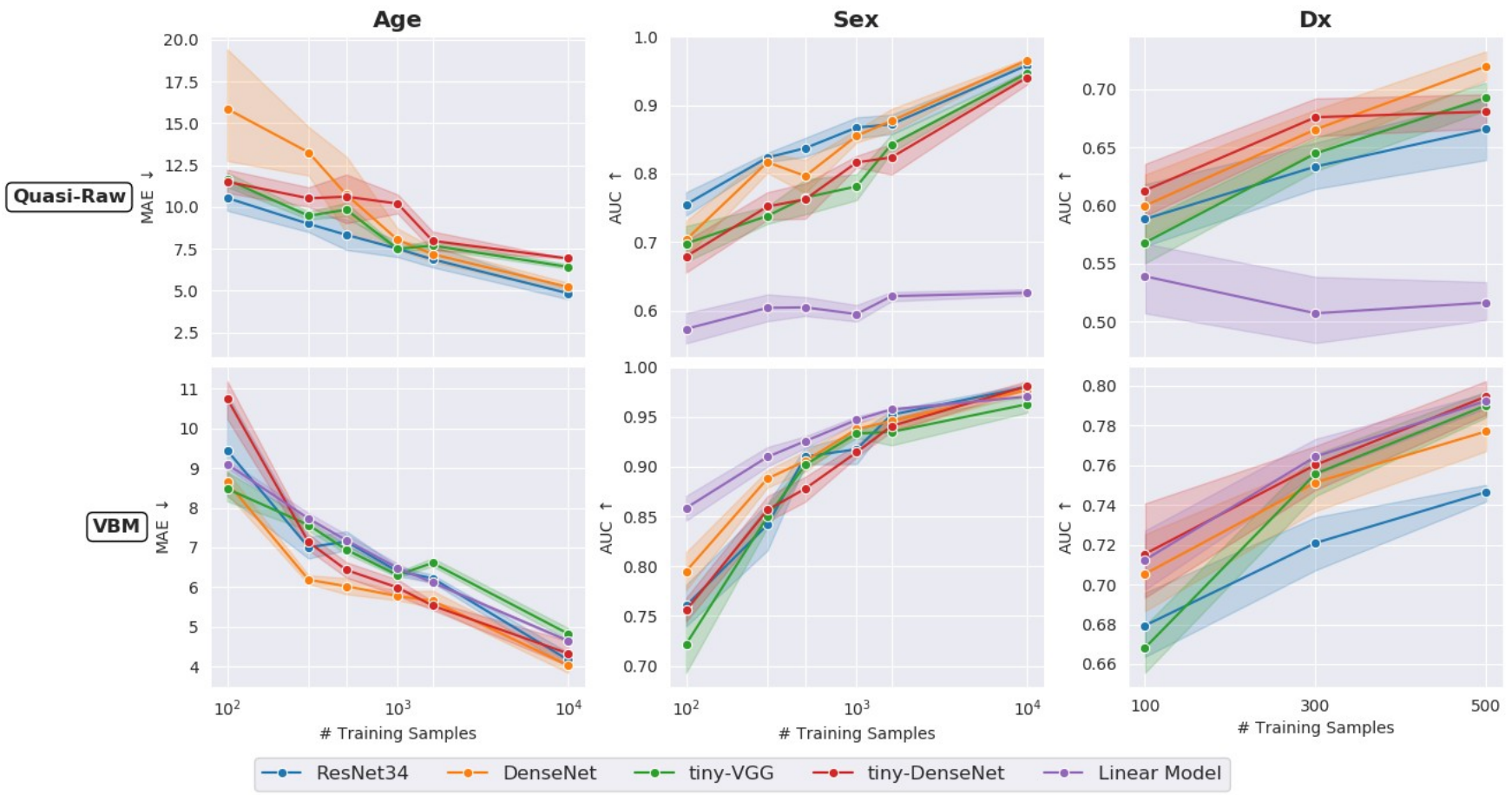
- Schulz (2020). Nature communications.
- Wen … Colliot (2020). MedIA.
- …

Aspects not addressed yet

- Choice of an architecture (no need to deal with invariances).
- Relevance of data augmentation.
- Ensemble learning and dropout.

- Dufumier … Duchesnay (2021) submitted to NeuroImage.

Our benchmark with 10k training samples from dozens of sites on 3 tasks: Age, Sex and Schizophrenia prediction

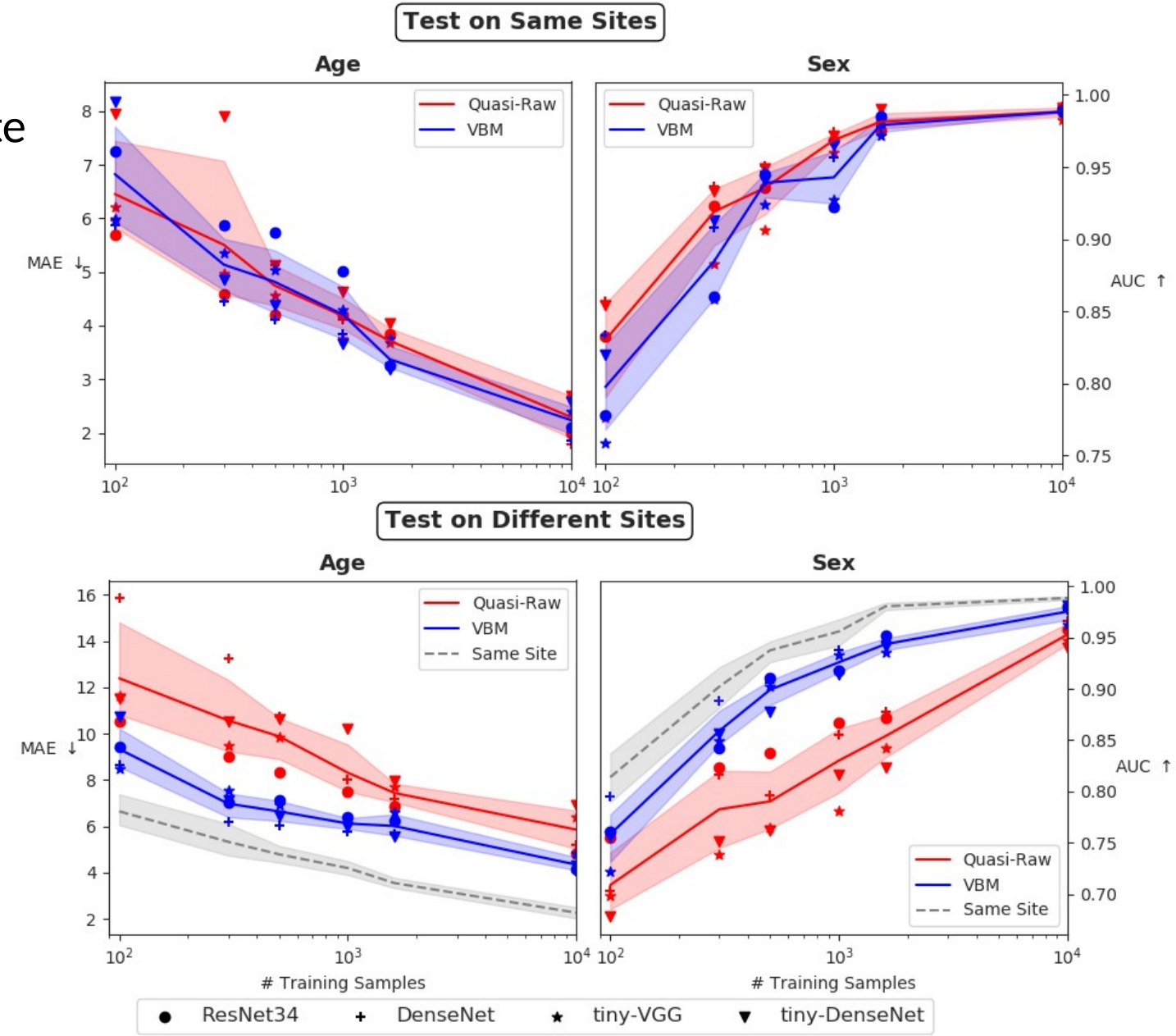


- Traditional feature extraction (VBM) outperforms “quasi-raw” data.
- Linear models equal DL.
- DenseNet is a good default choice.

To generalize to independent sites, traditional feature extraction (VBM) outperformed quasi-raw in generalization to independent sites

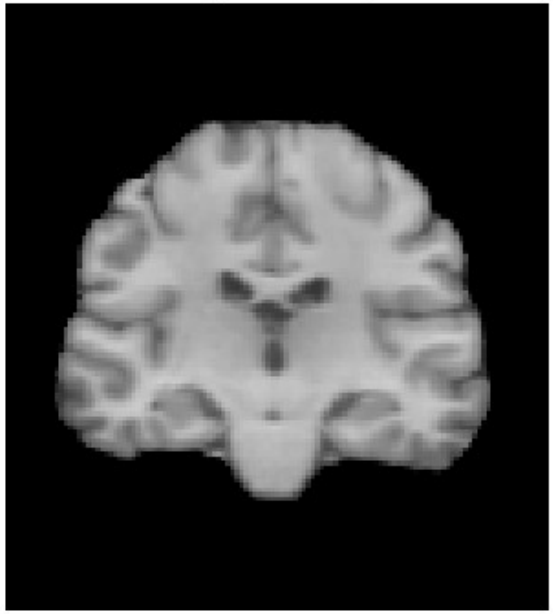
5-CV stratified for site
VBM = Quasi-Raw

Leave-One-Site-Out
VBM > Quasi-Raw

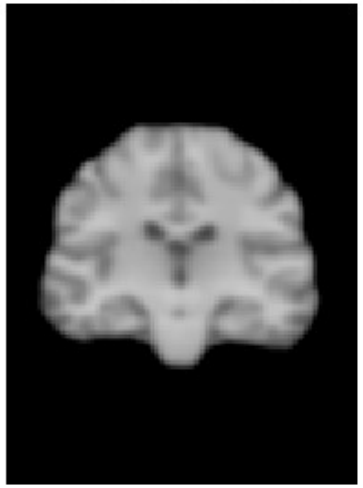


Data augmentation?

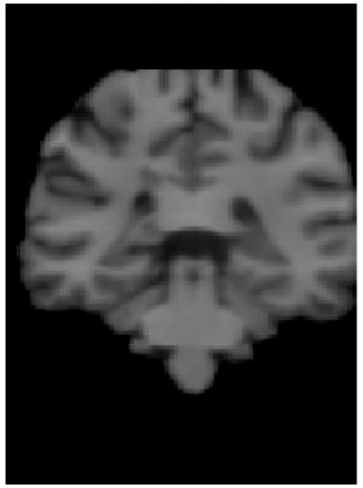
Original image



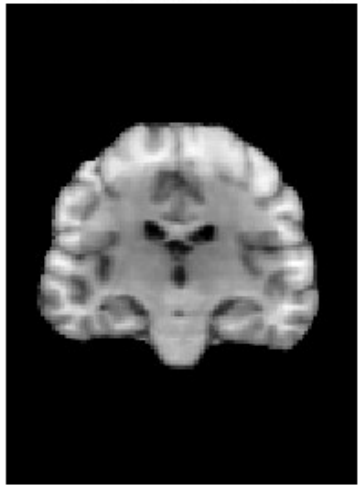
Gaussian Blur



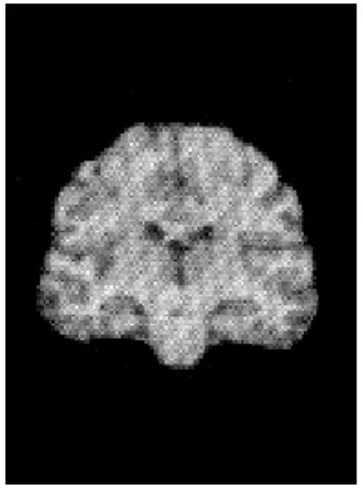
Crop



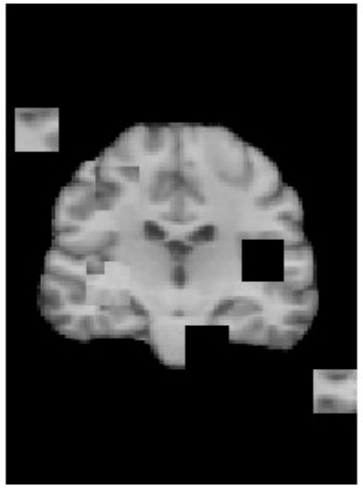
Ghosting Artefact



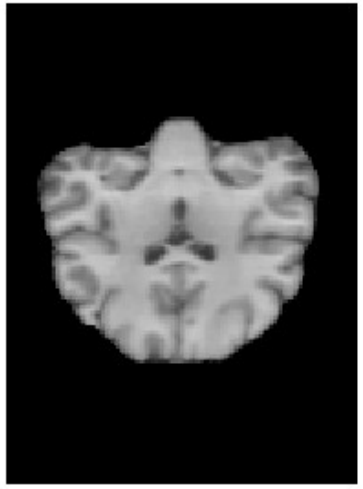
Spike Artefact



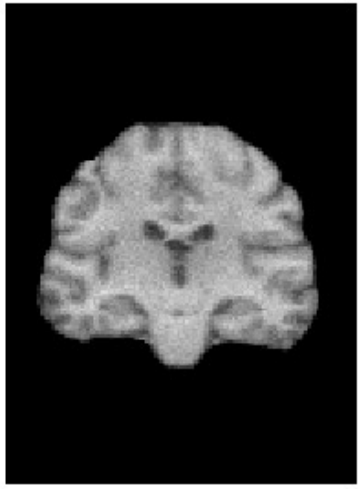
Random Swap



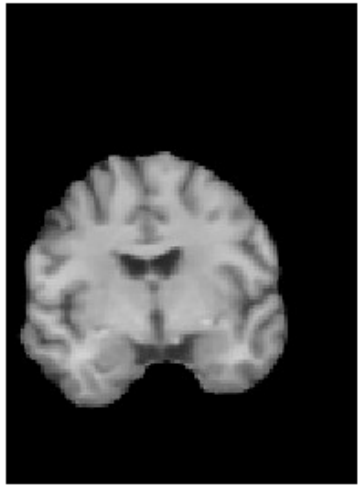
Flip



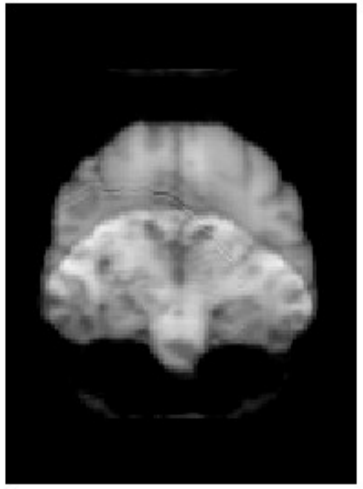
Gaussian Noise



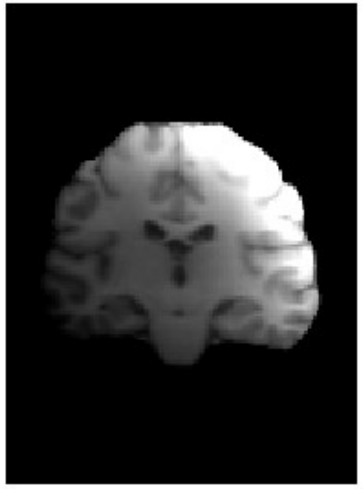
Affine transformation



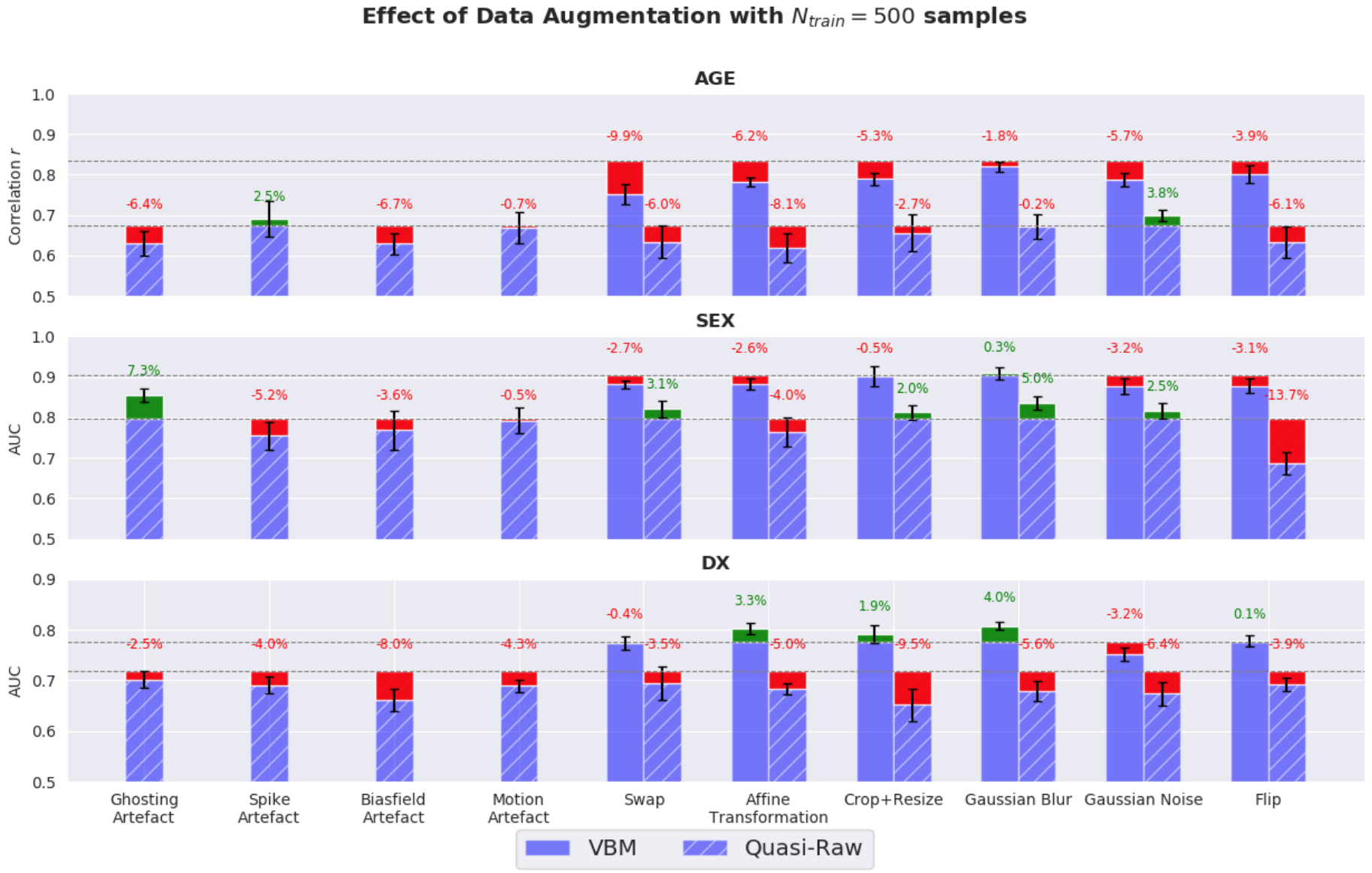
Motion Artefact



Biasfield Artefact



Data augmentation: no clear improvement



Increase
Decrease

Outline

1) Introduction: Machine learning to identify prognostic signature in psychiatry

2) Fighting over-fitting: Feature selection, regularization and model selection

3) Spatial regularization to improve interpretability

4) Applications to psychiatry

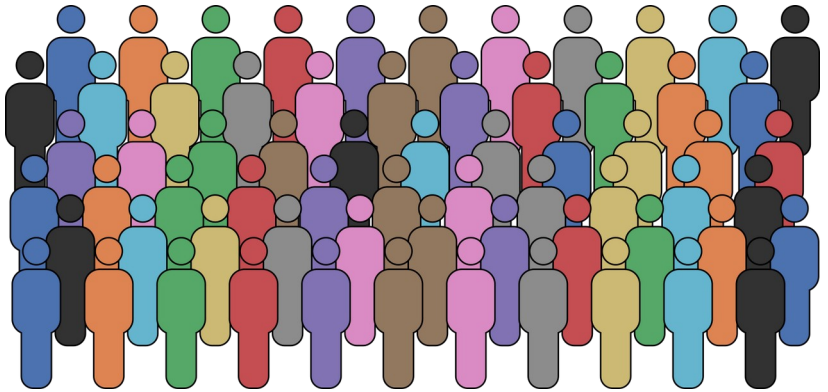
5) Deep learning for supervised task

6) Transfer learning and representation learning

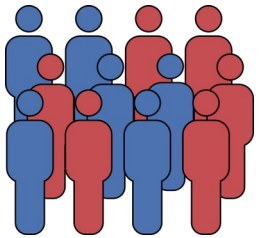
Transfer learning from big “dirty” datasets to small quality datasets: “learn reference curves”

Learn the general variability on large datasets

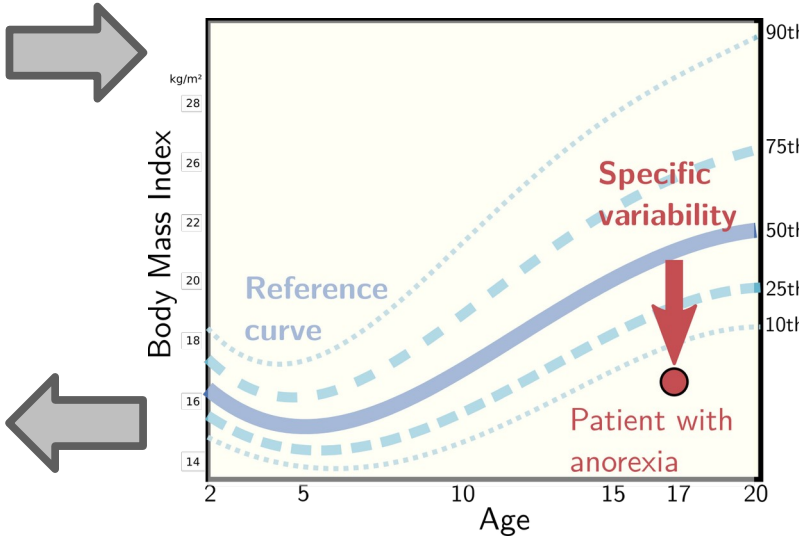
– Large “open” datasets
– Heterogeneous “dirty”
– Cross-sectional
– Transdiagnostic
– 10k subjects
– Psychiatry requires “young” participants (UK-BioBank > 40 years)



Small homogeneous datasets:
<1000 subjects, 10k€/subject.
10k subjects would cost 100M€ !



1) Learn reference curve



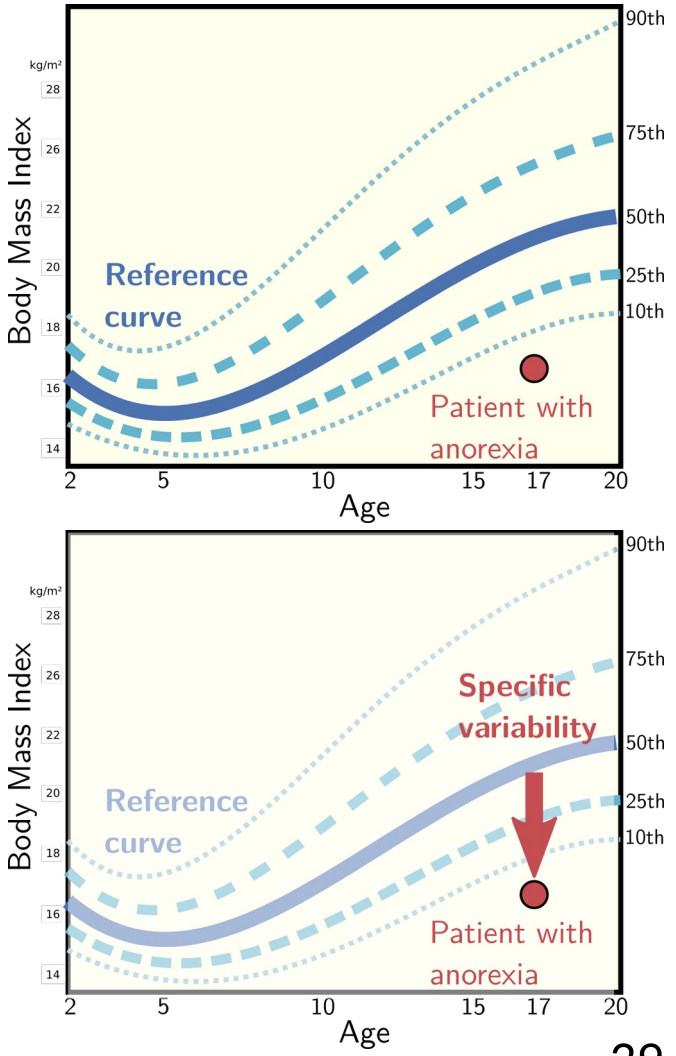
2) Transfer Learning from Big data to Small Data
Detect deviation from reference curve

Transfer learning from big “dirty” datasets to small quality datasets: “learn reference curves” details

Large cohorts of general population
- N ≥ 10,000, transnosographic heterogeneous
- (HBN, ABCD, UK Biobank, other open-datasets)
- Learn the general variability

Disorder-specific retrospective and heterogeneous cohorts
- N < 10,000, cross-sectional, case-control, “dirty”, limited clinical value
- (Schizconnect (SCZ), abide (ASD), BIOBD/BSNIP (BD)
- Focus on specific variability

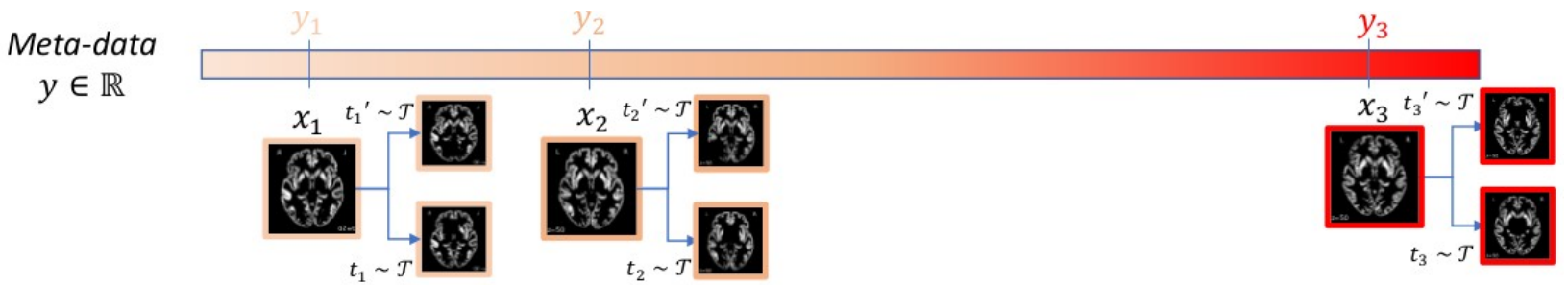
Disorder-specific longitudinal and homogeneous cohorts
- N < 2,000, longitudinal and homogeneous cohorts, high clinical value
- PsyCARE (RHU), R-LiNK, H2020, EU-AIMS
- Learn useful predictive signature

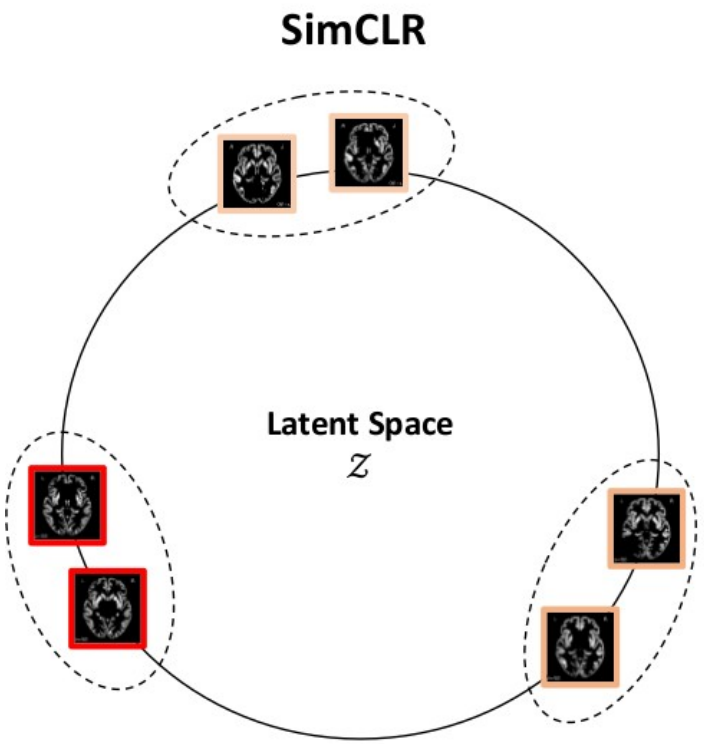


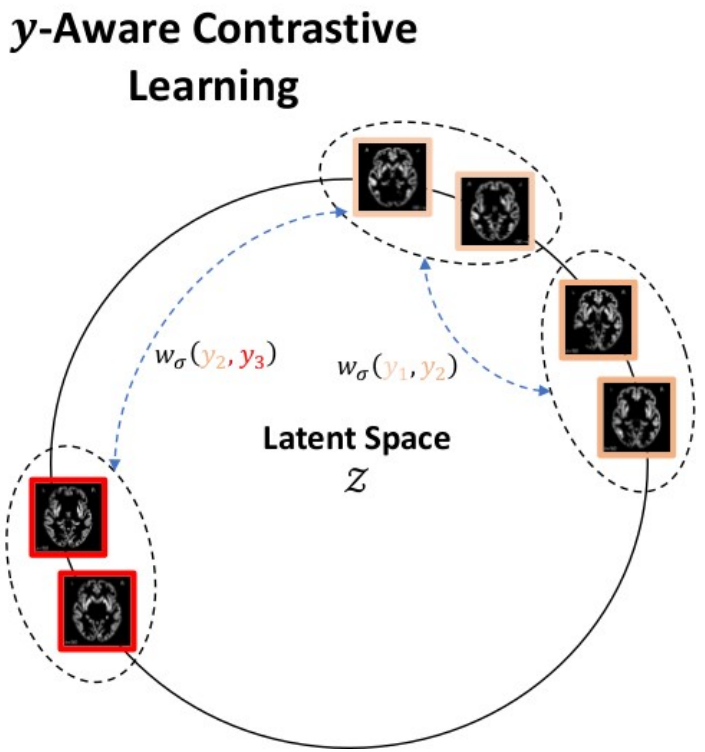
Transfer learning: Self-supervised contrastive learning integrating age

Encoders for representation learning of the general variability

– Learn on large 10k dataset
– Add contrastive learning
– Add prior on age



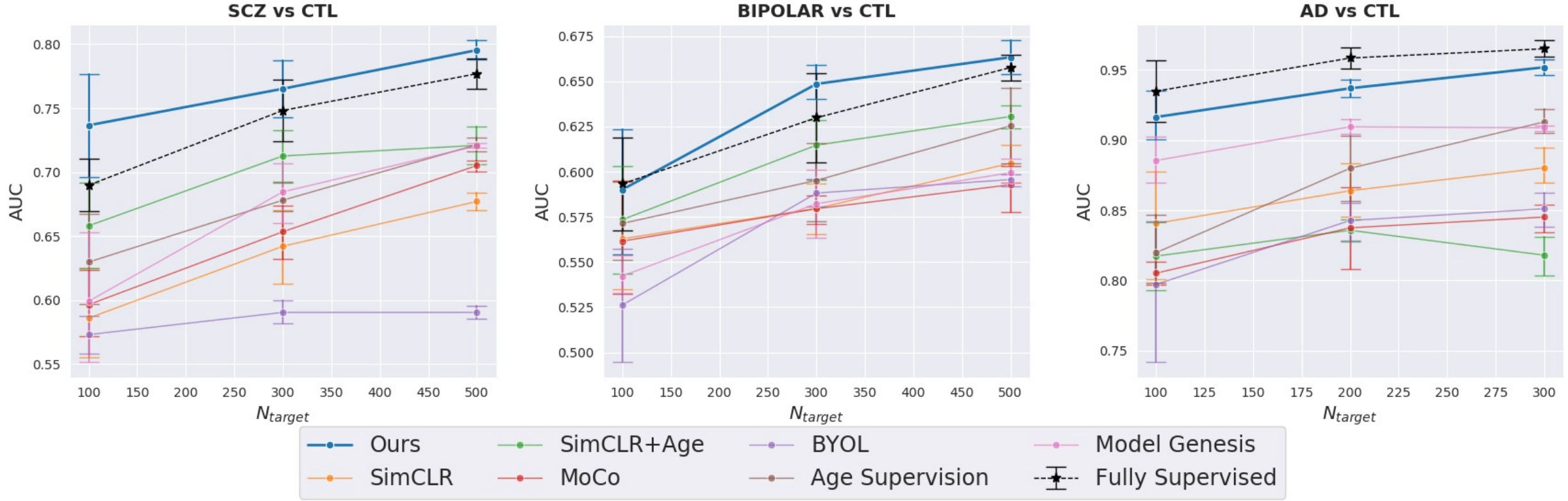




Dufumier (2021)
submitted to MICCAI

Representation learning with age-aware contrastive learning

Representation learning: learn encoder on 10k samples (upstream task).
Validation (downtream task): Learn linear layer on top of the pre-trained frozen encoders
Compare with different strategies and a fully supervised fine-tuned model (gold standard)



Conclusion

- Simple cutout transformation is efficient
- Age-aware CL pre-training significantly outperformed all other strategies: pre-training on age, classical SimCLR, etc.
- *Fully fine-tuning of pre-trained age-aware CL significantly outperformed linear…*

Dufumier (2021) submitted to MICCAI

Summary

Regularized linear models

1) Feature selection/sparsity may help but it raises the problem of model selection.
2) Most models provide similar performances, always try the venerable l2-regularization as baseline.
3) Interpretability is a key issue to consider a transfer to the clinic.

Deep learning

- We found that DL is not better than traditional feature extraction with linear models on supervised task
- However, DL opens perspectives for transfer learning from large heterogeneous datasets to small homogeneous datasets
- Representation learning of the general variability (remember the reference curve of BMI). This idea is similar to the "normative modeling" of André Marquand.

The data issue

- Most of existing (open) datasets are of low value for useful clinical application:
  - Either cross-sectional (case/control) or heterogeneous trans-diagnostic (HBN, UKB, ABCD, etc.)
- Need for large longitudinal datasets: data → treatment → outcome. Learn(data) → outcome
- Problem: cost is 10K€/sample → **100k€ for 10k subjects** !!!
- The best we will get for next 10 years is ~1000 subjects per clinical problem.
- That's why we are exploring transfer learning, any other idea?

Thanks!

Anton Iftimovici
MD Psychiatrist
PhD@NS

Benoit Dufumier
PhD@NS

Robin Louiset
PhD@NS

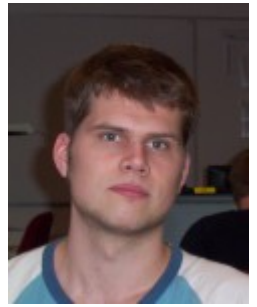
Julie Victor
Engineer@NS

Antoine Grigis
Engineer@NS

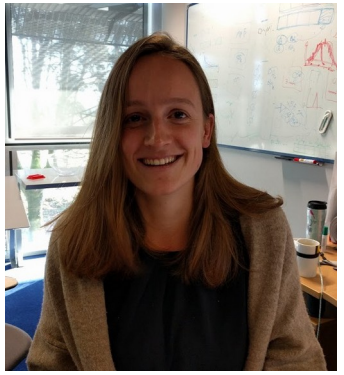
Pietro Gori
Researcher
@Télécom

Josselin Houenou
Pr in Psychiatry

Fouad Hadj-Selem
Researcher
@Vedecom



Tommy Loftstead
Researcher
@UMEA, Sweden



Amicie de Pierrefeu
Former PhD