ARAMIS LAB
BRAIN DATA SCIENCE

Machine Learning to Predict Impulse Control Disorders in Parkinson's Disease

Johann Faouzi, PhD candidate
supervised by Olivier Colliot and Jean-Christophe Corvol

February 14, 2019

PRECISE-PD meeting

Introduction

- Impulse Control Disorders (ICDs) are much more common in the Parkinson's Disease (PD) population than in the general population.
- ICDs in PD represent an important Public Health problem because of their familial, social, economic or legal impact.
- There is a substantial amount of articles about this issue, most of them focusing on associating factors in cross-sectional or longitudinal studies.
- Almost no article about prediction of ICDs in PD!

Outline

Outline

Parkinson's Disease (PD)

- Parkinson's Disease is the second most common neurodegenerative disease with around 7 million affected people in the world.

- There is currently no cure and therapies aim at improving the quality of life.

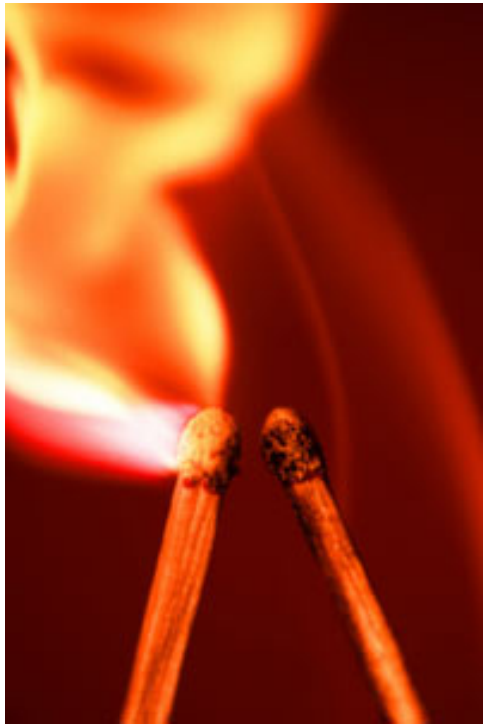
- Most common therapies are based on dopamine replacement, with the use of levodopa, dopamine agonists, inhibitors.

Impulse Control Disorders (ICD)

- **Impulse Control Disorders: class of psychiatric disorders characterized by impulsivity.**

- **Impulsivity: failure to resist a temptation, an urge, an impulse, or the inability to not speak on a thought.**

- **Examples of ICDs:**
  - Hypersexuality
  - Compulsive shopping
  - Pyromania
  - Intermittent explosive disorder
  - Kleptomania
  - Binge eating
  - Internet addiction
  - Pathological gambling



ICDs in PD

- **ICDs in PD are part of a more global term called "behavioral addictions" also including *dopamine dysregulation syndrome* and *punding*.**

- **Much higher prevalence than in the general population.**

- **Only a subset of ICDs are reported in the PD population:**
  - Hypersexuality
  - Compulsive shopping
  - Binge eating
  - Pathological gambling

ICDs in PD: Literature

- **Cross-sectional and longitudinal studies**
- **High focus on *covariates associated with ICDs*:**
  - Age at onset
  - Gender
  - Motor complications
  - Sleep disorders
  - Psychiatric symptoms (anxiety, depression)
  - Dopamine replacement therapy (specially dopamine agonist)
  - Genetics (SNPs)

- **Only two articles with a prediction task:**
  - Kraemmer et al. Clinical-genetic models predicts Impulse Control Disorders in Parkinson's Disease. J Neurol Neurosurg Psychiatry, 2016.
  - Erga et al. Dopaminergic and Opiod Pathways Associated with Impulse Control Disorders in Parkinson's Disease. Front Neurol, 2018.

Outline

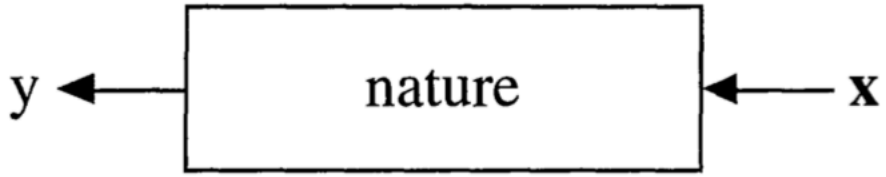
What is machine learning?

- Process of automatically extracting information from data
  - The user does NOT provide explicit rules
  - The user provides a class of rules (an algorithm), with the best rules (the parameters) being selected in a data-driven approach

- Most common goal: **generalization**
  - Generalization = making predictions on new, unseen data

Statistical Modeling: The Two Cultures

Data analysis



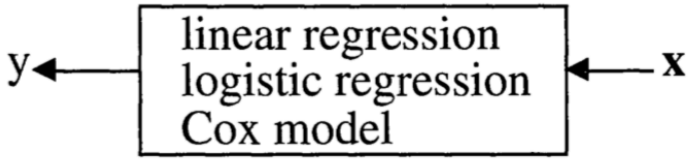
There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

Leo Breiman, "Statistical Modeling: The Two Cultures". Statistical Science, 2001, Vol. 16, No. 3, 199-231

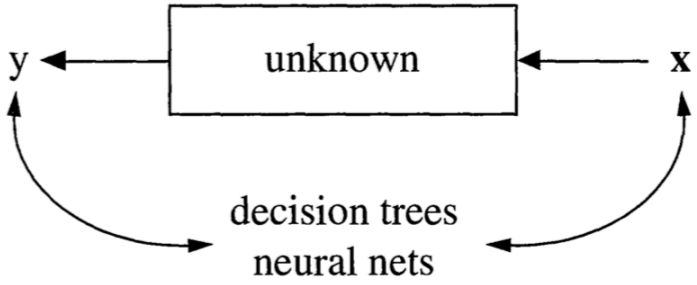
Statistical Modeling: The Two Cultures

The Data Modeling Culture



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Leo Breiman, "Statistical Modeling: The Two Cultures". Statistical Science, 2001, Vol. 16, No. 3, 199-231

Statistical Modeling: The Two Cultures

The Data Modeling Culture:

- **More focused on the model than the data**
- Model evaluation using goodness-of-fit tests

The Algorithmic Modeling Culture

- **More focused on the data than the model**
- Model evaluation using predictive accuracy on an independent dataset

Leo Breiman, "Statistical Modeling: The Two Cultures". Statistical Science, 2001, Vol. 16, No. 3, 199-231

Statistical Modeling: The Two Cultures

Interpreting a model is risky

(With great power comes great responsibility!)

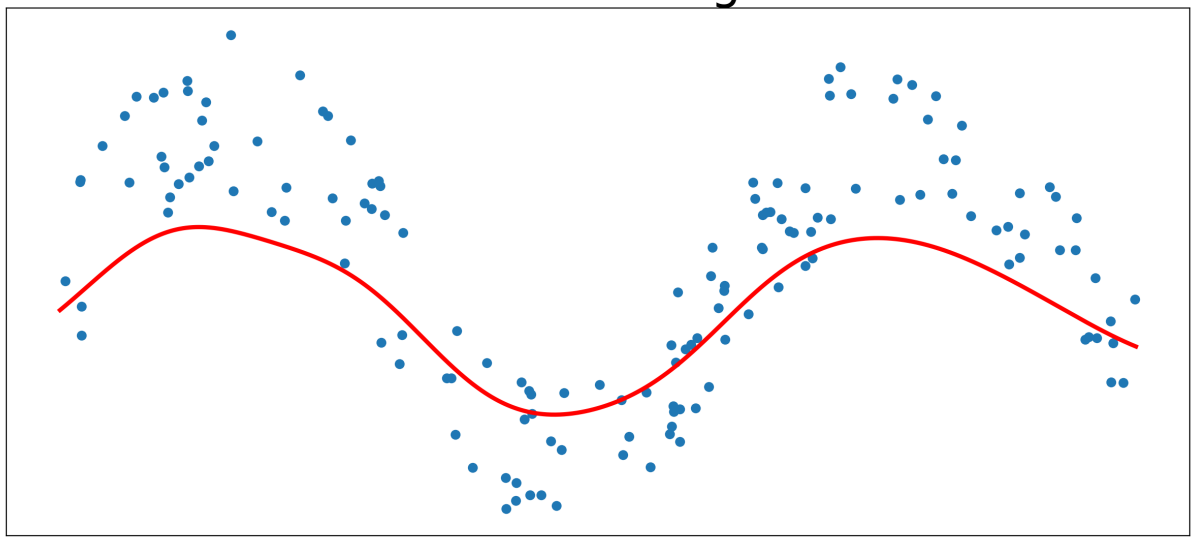
Be careful of:

- Overfitting (and underfitting)
- Numerical optimization
- Non-convex functions

Leo Breiman, "Statistical Modeling: The Two Cultures". Statistical Science, 2001, Vol. 16, No. 3, 199-231

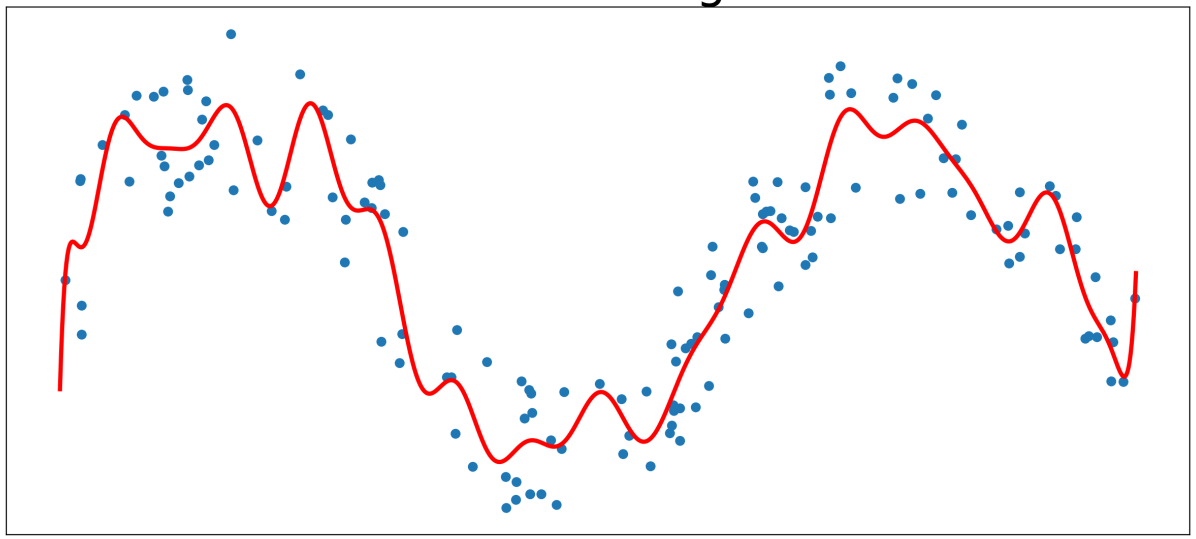
Fitting

Underfitting



Bad

Overfitting



Worse

Fitting - cross-validation

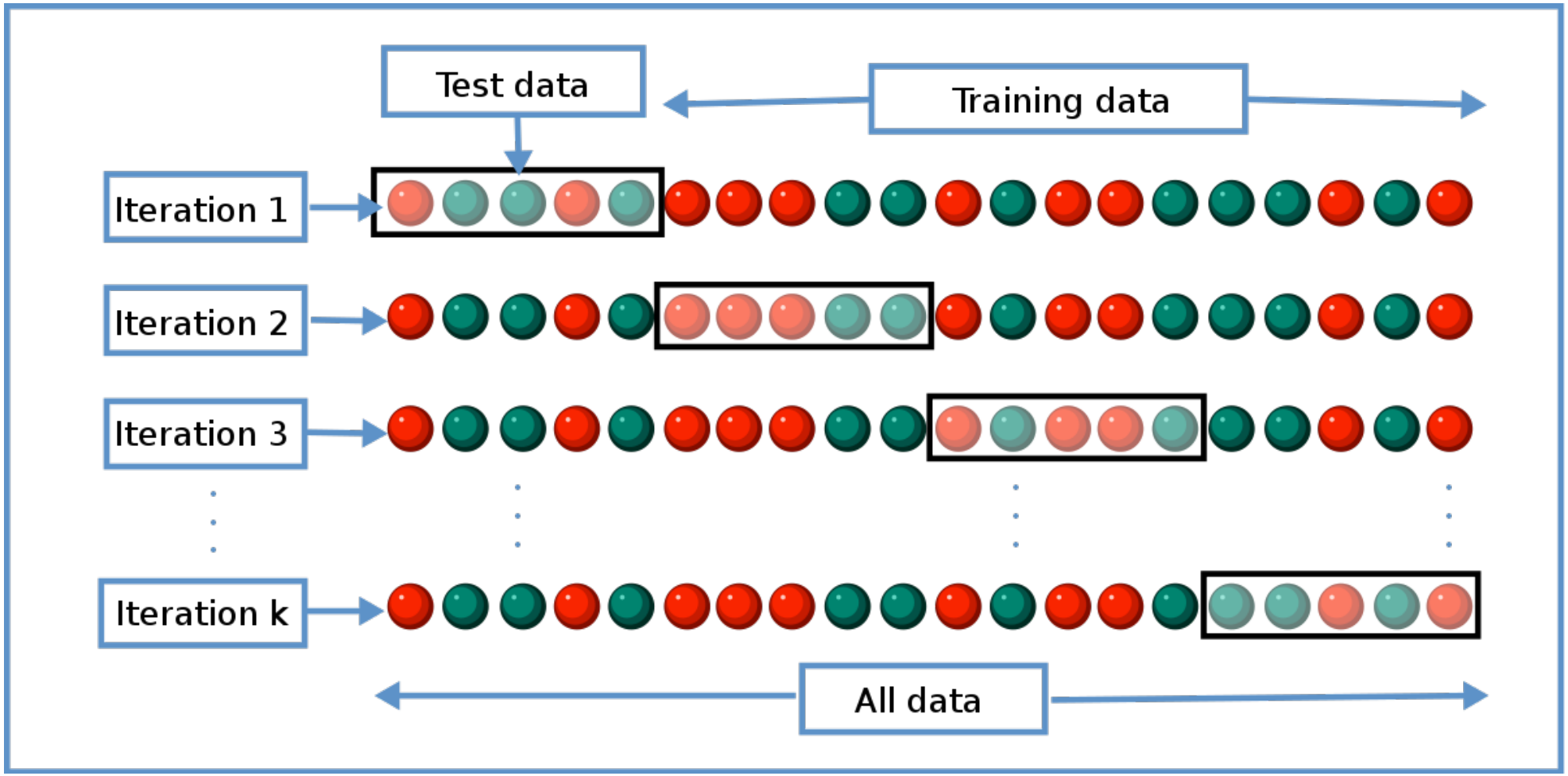
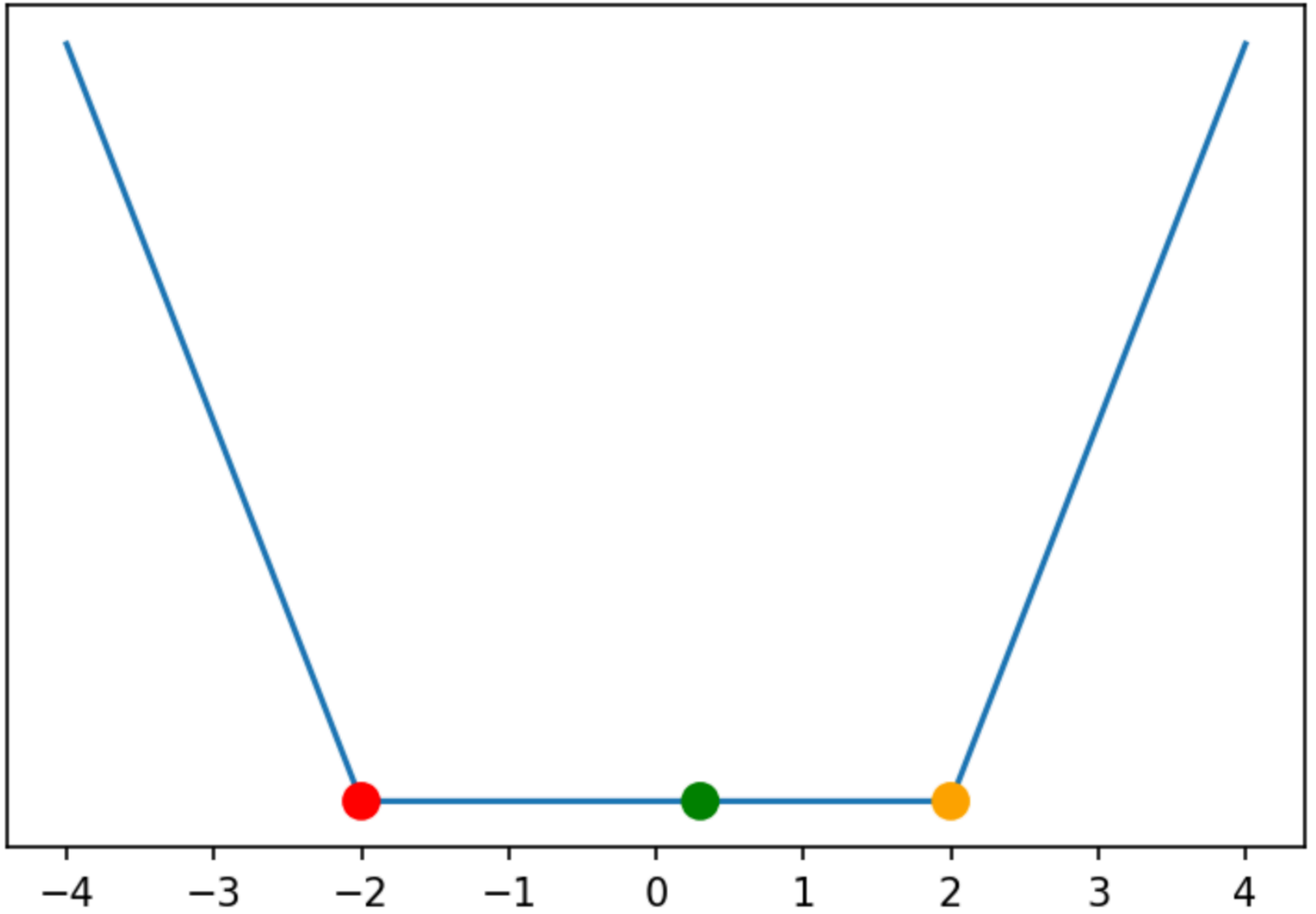
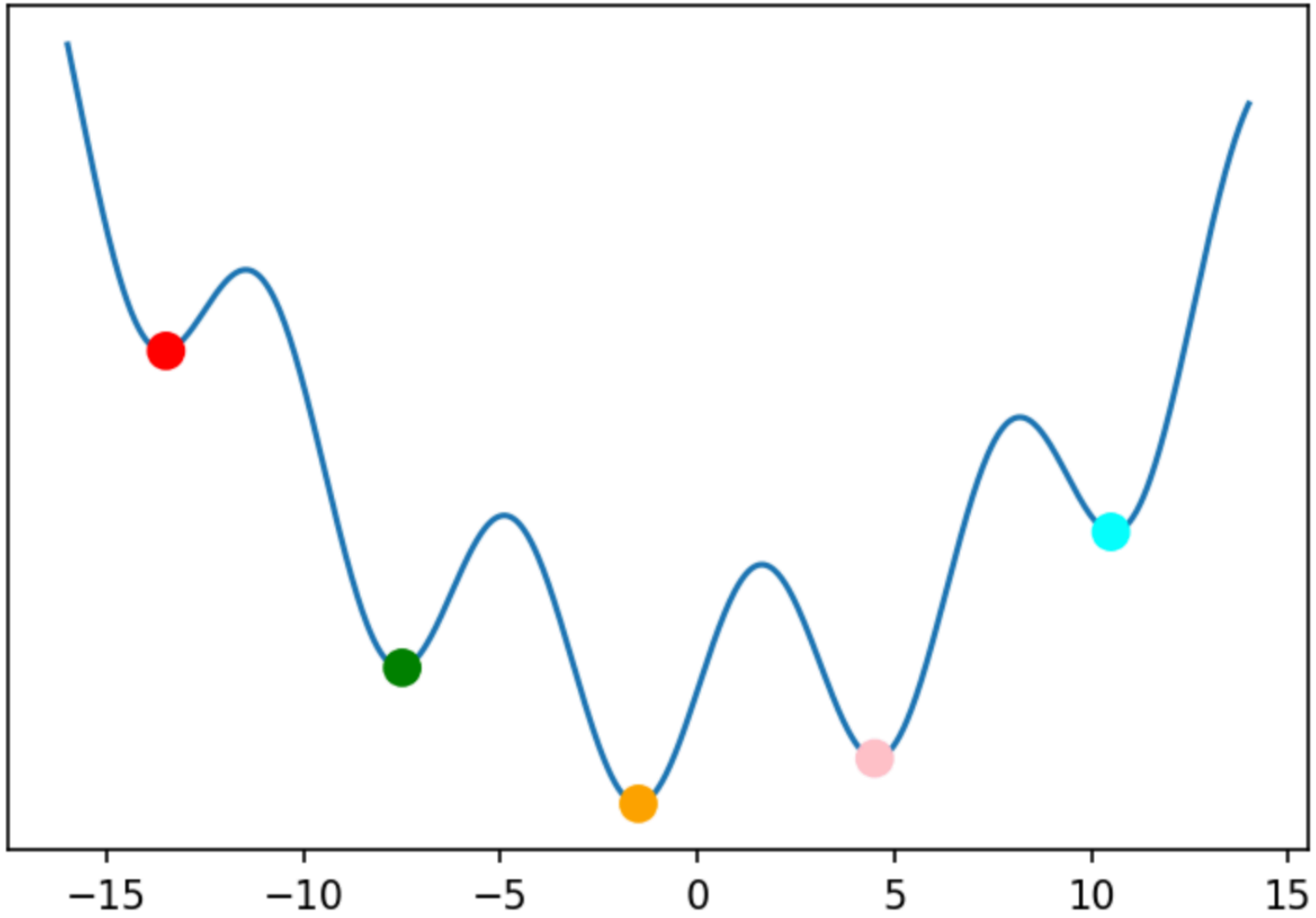


Diagram of k-fold cross-validation. Wikipedia entry for « Cross-validation (statistics) »

Numerical optimization



Non-convex functions



Outline

Assessing ICDs accurately

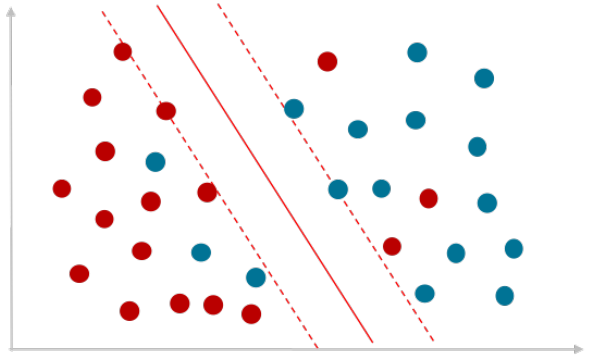
- **Several questionnaires to assess ICDs:**
  - QUIP (Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease)
  - ECMP (Evaluation Comportementale de la Maladie de Parkinson)
  - MDS-UPDRS 1.6

- **Heterogeneity in the diagnosis:**
  - Subjectivity in the interpretation of the answers
  - Cultural differences

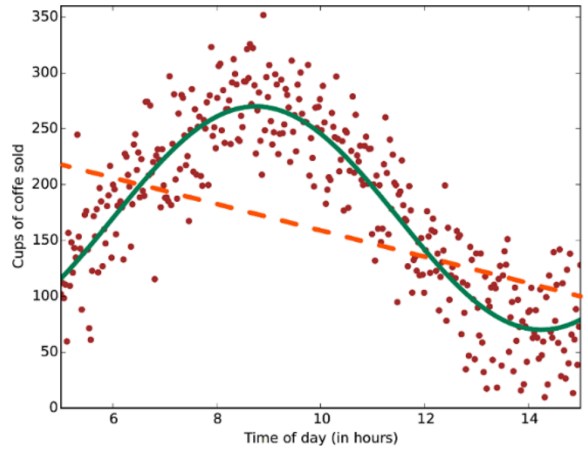
- **Wrong diagnosis of ICDs because of:**
  - Lack of awareness
  - Shame

Formulations

- **Predicting ICDs: what kind of predictions?**
  - Dates/numbers: First onset of ICDs
  - Binary prediction:
    - Ever or never ICDs during their (currently available) follow-up
    - Ever or never ICDs during the first N years of follow-up
    - Presence or absence of ICDs for each (patient, visit) pair



- **What kind of machine learning tasks and models**
  - Tasks: Regression, Classification
  - Predictors: “Cross-sectional”, Longitudinal



Outline

Dataset: PPMI

- **Parkinson's Progression Markers Initiative (PPMI): landmark observational clinical study taking place at clinical sites in the United States, Europe, Israel, and Australia.**

- **Available data:**
  - Clinical data (Parkinson's disease specific scales, psychological tests, etc.)
  - Imaging data (DaTSCAN, structural MRI)
  - Genetic data (genotype)

Variables included in the models

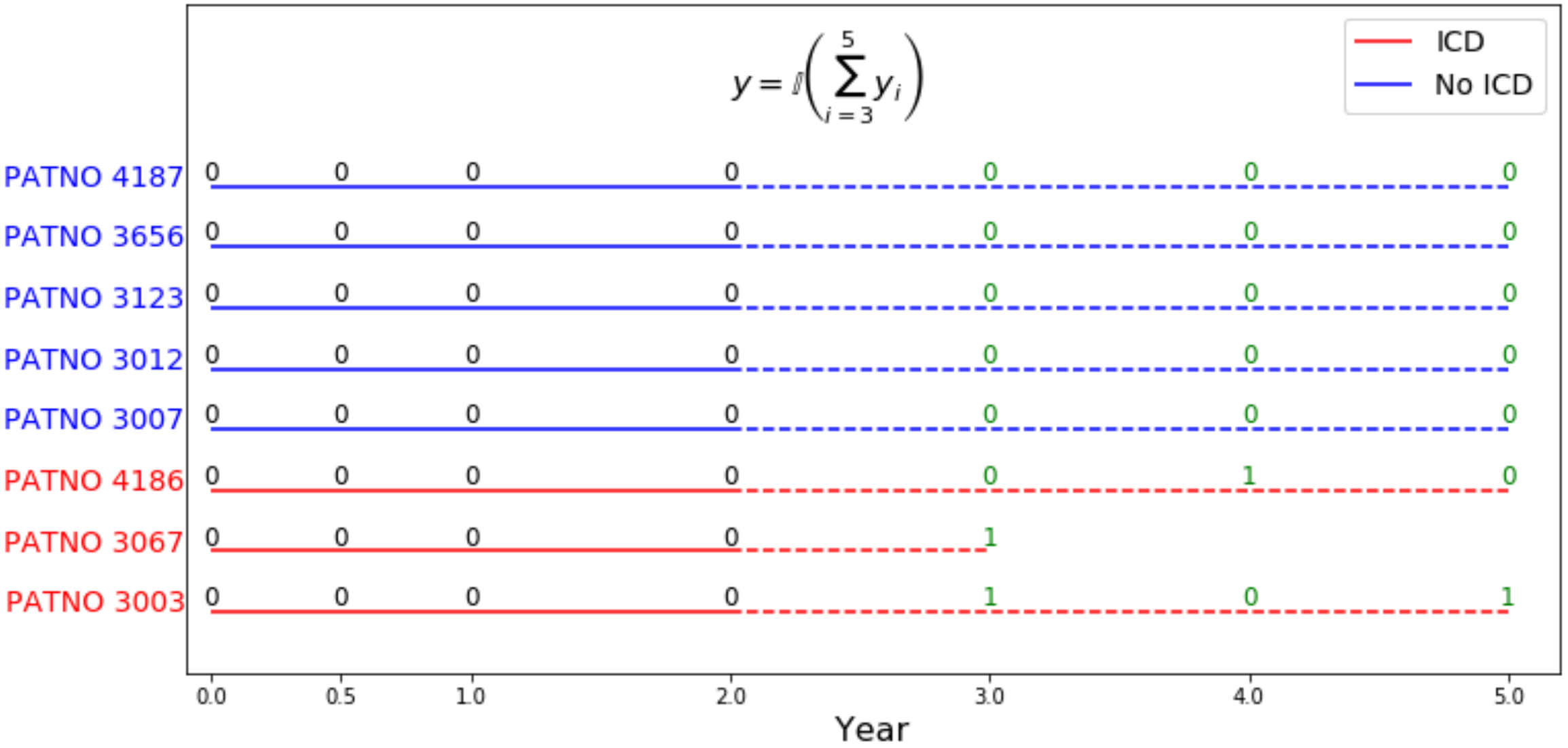
- **Age at onset**
- **Gender**
- **SNPs**
- **REM Sleep Behavior Disorders**
- **State and Trait Anxiety Inventory**
- **Geriatric Depression Scale**
- **Type of PD medication taken (dopamine agonists, levodopa, others)**
- **Unified Parkinson's Disease Rating Scale – Part III (motor exam)**
- **For dopamine agonists:**
  - mean daily dose
  - cumulative duration
  - total daily dose

Formulation

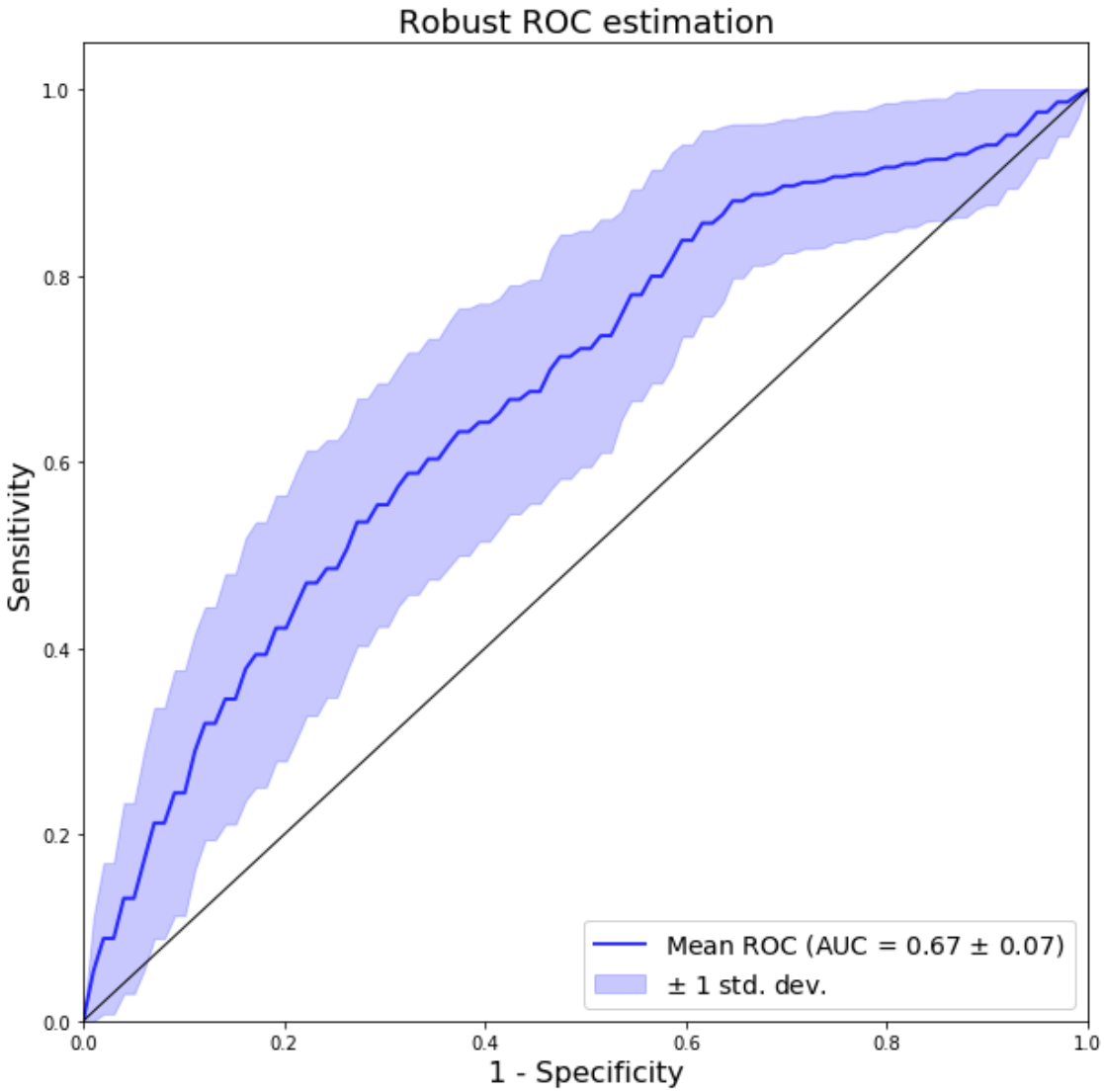
Given a patient who did not develop ICDs during the first M year, will this patient have ICDs in the next X years?

Binary classification in a cross-sectional model.

Will a patient without ICDs until year 2 develop ICDs between year 3 and year 5?



Will a patient without ICDs before year X develop ICDs between year 3 and year 5?



- 238 subjects:
  - 24 with ICDs
  - 196 without ICDs
- Not-so-great performance (ROCAUC ≈ 0.67)

Outline

Other formulation

Predict the presence or absence of ICDs at the next given information from the past:

$$y_{t+1} = f(x_1, \ldots, x_t, s)$$

$\boldsymbol{y}$ = presence or absence of ICDs (binary variable)

$\boldsymbol{x}$ = dynamic variables (clinical variables collected at each visit)

$\boldsymbol{s}$ = static variables (gender, age at onset, genetic data, …)

Genetics: SNP data

- **SNP chips : number of minor alleles : {0, 1, 2}**
- **Imputed SNP: number between 0 and 2**

- **We try 3 sets of SNP data:**
  1) No SNP data at all
  2) 13 SNPs that are known to be associated with ICDs from the literature
  3) 31 SNPs:
     - 13 SNPs that are known to be associated with ICDs from the literature
     - 18 exploratory SNPs from at most 10 genes

Genetics: SNP data

Known SNPs

Gene	SNP	Gene	SNP
DBH	rs1108580	COMT	rs4680
TPH2	rs1352250	BDNF	rs6265
DBH	rs1611115	DRD2	rs6277
OPRM1	rs1799971	DRD3	rs6280
ANKK1	rs1800497	HTR1B	rs6296
TPH1	rs1800532	TPH2	rs6582078
GRIN2B	rs1806201		

Exploratory SNPs

Gene	SNP	Gene	SNP
ARC	rs10097505	FOSB	rs2282695
CA12	rs1043239	MOSC1	rs2984657
CA12	rs1043256	CA12	rs4984241
FOSB	rs1049739	C8B	rs591730
CA12	rs1075456	C8B	rs617283
MOSC1	rs1109103	CA12	rs7166946
CA12	rs16946963	C8B	rs725330
CA12	rs2046484	CCRN4L	rs938836
FOSB	rs2276469	CA12	rs9989288

Cormier-Dequaire et al. *Suggestive association between OPRM1 and impulse control disorders in Parkinson's disease*. Movement Disorders, December 2018.

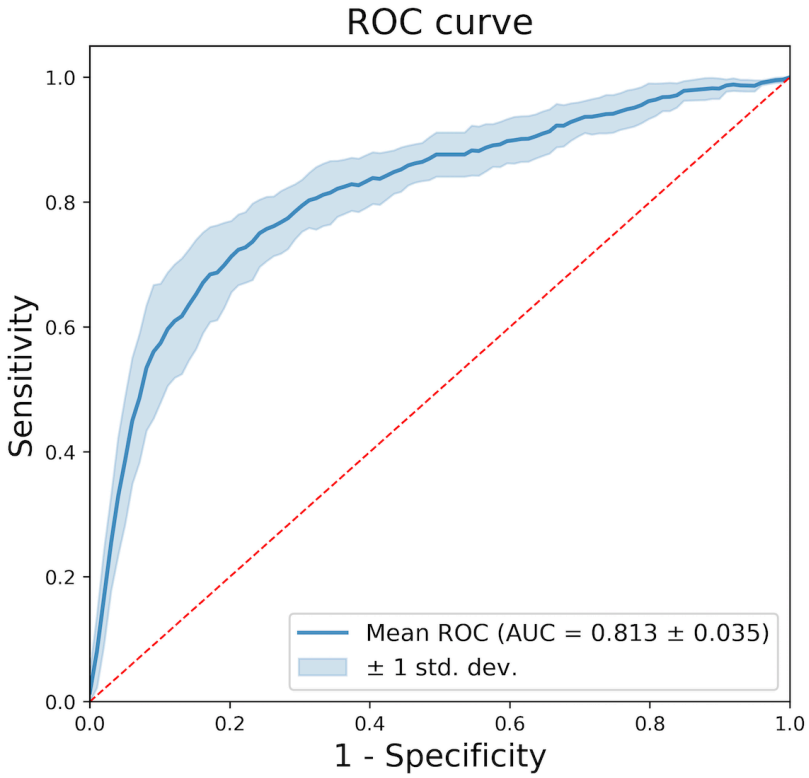
Machine learning

- **Challenge**: Standard algorithms (like logistic regression) cannot handle a varying number of visits.
- **Idea**: Merging all the previous visits into one "summary visit" using a linear combination:

Reduction	Weights
Baseline visit	$[1, 0, \ldots, 0]$
Previous visit	$[0, \ldots, 0, 1]$
Mean over all the previous visits	$[\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}]$
More generally	$[w_1, \ldots, w_N]$ with $w_i = \frac{f(t_i)}{\sum f(t_i)}$

Machine learning - Results on PPMI

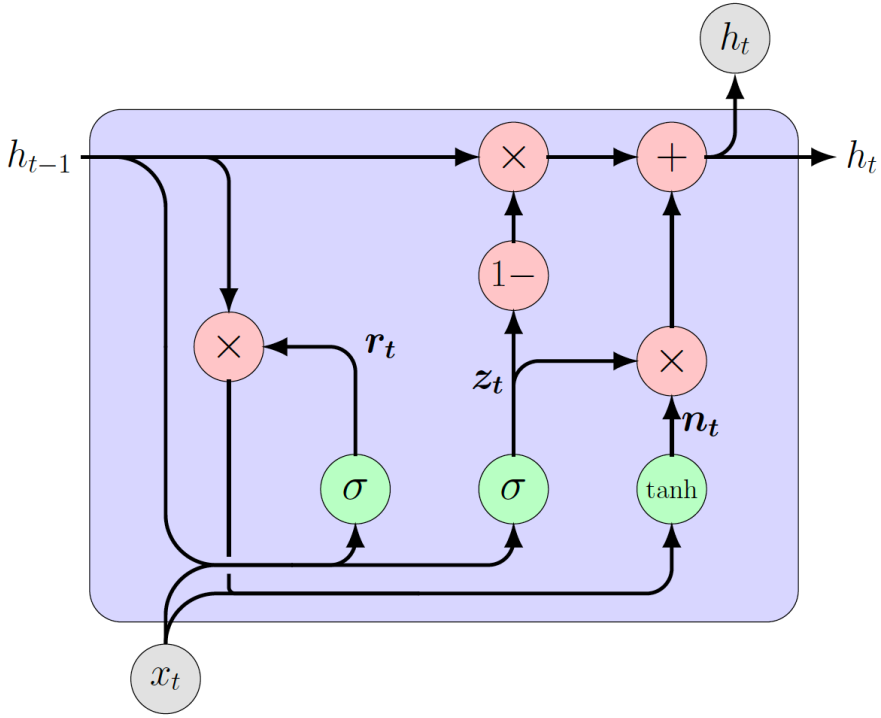
Reduction	Algorithm	ROC AUC	Average precision
”first visit”	LinearSVC	0.726 (0.061)	0.343 (0.092)
	XGBoost	0.703 (0.047)	0.303 (0.089)
“previous visit”	LinearSVC	0.773 (0.045)	0.408 (0.089)
	XGBoost	0.772 (0.044)	0.384 (0.085)
“mean”	LinearSVC	**0.813 (0.035)**	**0.426 (0.088)**
	XGBoost	**0.804 (0.035)**	**0.449 (0.101)**



- Using the **“mean” visit** leads to **better results.**
- Could we use algorithms that could **better take into account the previous visits** than using an arbitrary function (like the mean)?

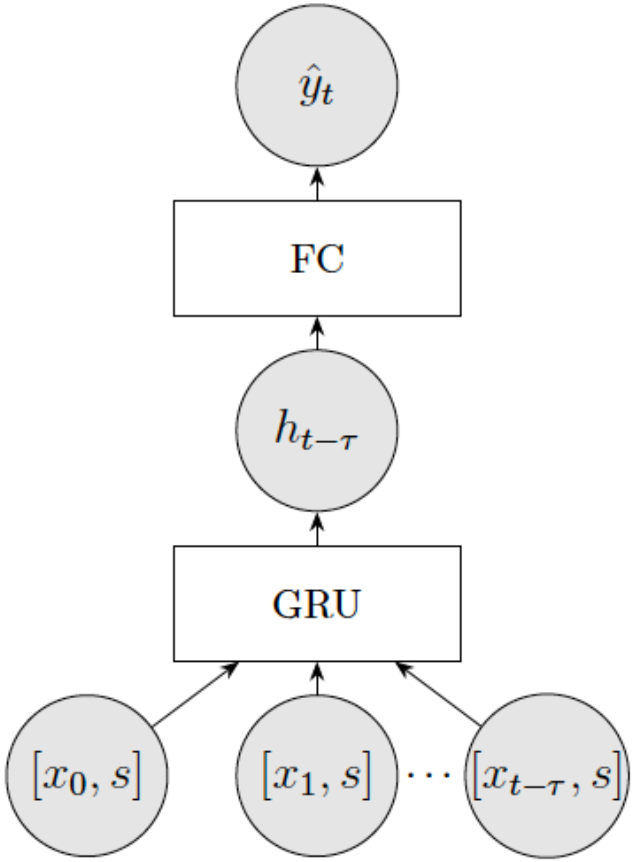
Deep learning

- **Idea**: Using a recurrent neural network (RNN) since longitudinal data is a particular case of sequential data
- **Challenge**: Integrating static data in a recurrent neural network:
  - Treating static data as **dynamic** data?
  - Putting static data **after** the RNN? It means that the RNN will extract information from the previous visits without knowing the static data (gender, SNP)?
  - Putting static data **before** the RNN?
  - **Init**ializing the RNN with static data?

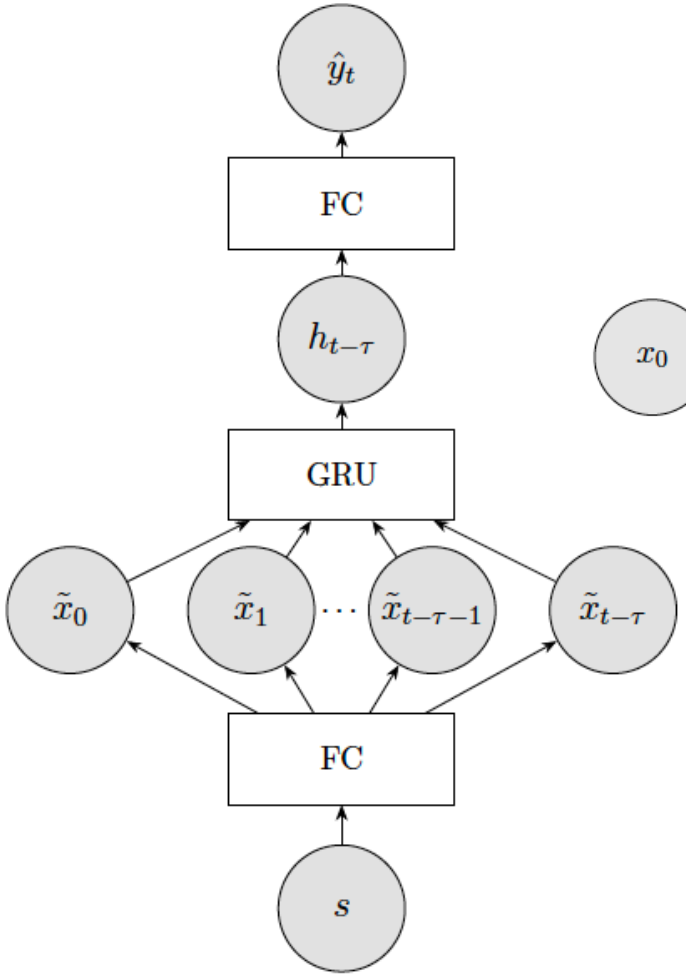


Deep learning

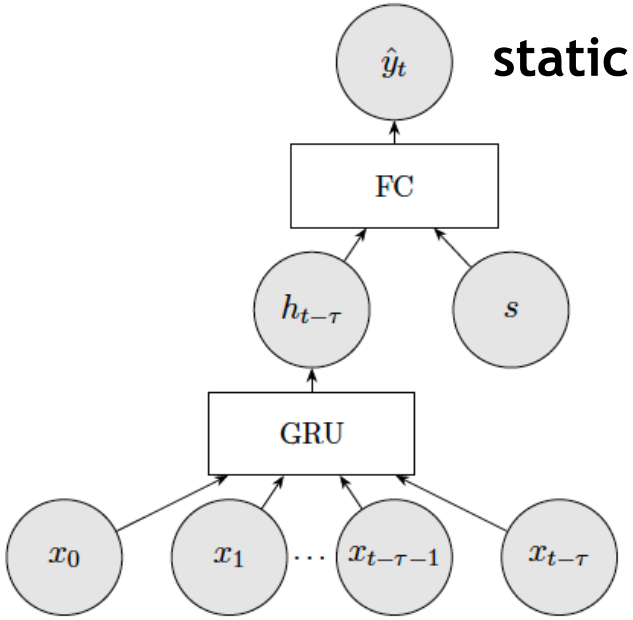
static = “dynamic”



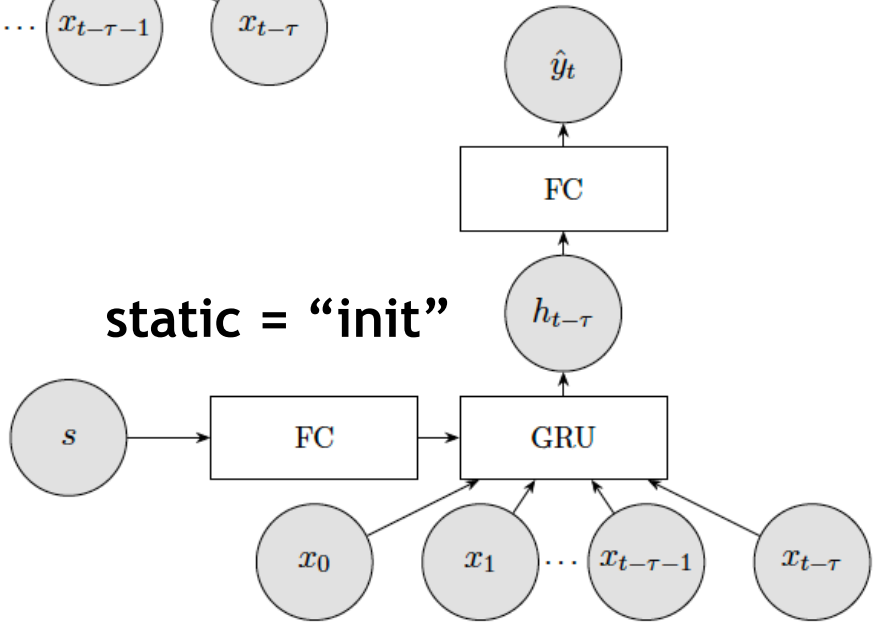
static = “before”



static = “after”



static = “init”



Metrics

- **We need metrics to evaluate and compare models.**
- **Two curves:**
  - **ROC curve**: sensitivity vs specificity
  - **Precision-recall** curve:
    - Precision = PPV
    - Recall = sensitivity

	True condition		
Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$

Results - No SNP

		ROC AUC		Average precision	
Reduction	**Algorithm**	**PPMI**	**DIG-PD**	**PPMI**	**DIG-PD**
”first visit”	LinearSVC	0.726 (0.061)	0.672 (0.010)	0.343 (0.092)	0.424 (0.006)
	XGBoost	0.703 (0.047)	0.658 (0.028)	0.303 (0.089)	0.437 (0.023)
“previous visit”	LinearSVC	0.773 (0.045)	0.783 (0.004)	0.408 (0.089)	0.619 (0.007)
	XGBoost	0.772 (0.044)	**0.790 (0.005)**	0.384 (0.085)	**0.644 (0.009)**
“mean”	LinearSVC	**0.813 (0.035)**	0.774 (0.012)	**0.426 (0.088)**	0.571 (0.017)
	XGBoost	**0.804 (0.035)**	0.772 (0.008)	**0.449 (0.101)**	0.549 (0.014)

	ROC AUC		Average precision	
Static data	**PPMI**	**DIG-PD**	**PPMI**	**DIG-PD**
”dynamic”	**0.817 (0.035)**	**0.802 (0.003)**	**0.470 (0.083)**	**0.628 (0.007)**
“before”	0.809 (0.035)	0.745 (0.019)	0.445 (0.078)	0.562 (0.032)
”after”	**0.815 (0.035)**	**0.800 (0.004)**	**0.474 (0.080)**	**0.624 (0.009)**
“init”	0.814 (0.036)	0.797 (0.005)	0.471 (0.089)	0.606 (0.018)

Results - Known SNP

		ROC AUC		Average precision	
Reduction	**Algorithm**	**PPMI**	**DIG-PD**	**PPMI**	**DIG-PD**
”first visit”	LinearSVC	0.746 (0.032)	0.680 (0.008)	0.374 (0.075)	0.419 (0.004)
	XGBoost	0.690 (0.042)	0.600 (0.037)	0.312 (0.071)	0.379 (0.030)
“previous visit”	LinearSVC	0.775 (0.036)	**0.790 (0.016)**	0.466 (0.048)	**0.624 (0.026)**
	XGBoost	0.774 (0.036)	0.774 (0.020)	0.449 (0.061)	0.619 (0.036)
“mean”	LinearSVC	**0.824 (0.015)**	0.784 (0.009)	**0.507 (0.062)**	0.591 (0.009)
	XGBoost	0.813 (0.016)	0.769 (0.013)	0.463 (0.052)	0.550 (0.023)

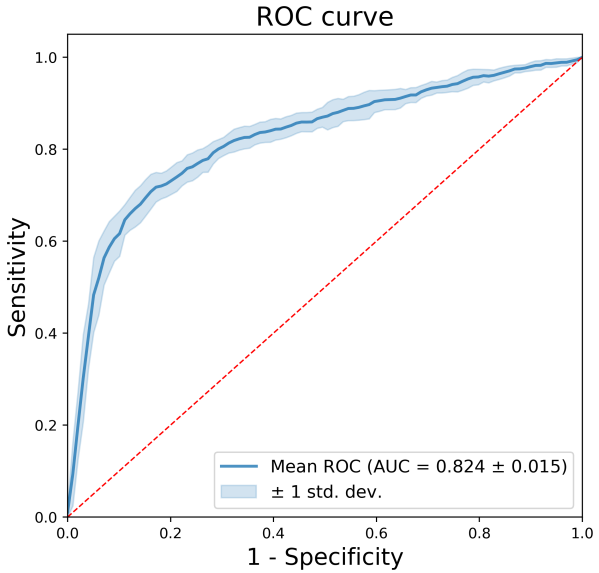
	ROC AUC		Average precision	
Static data	**PPMI**	**DIG-PD**	**PPMI**	**DIG-PD**
”dynamic”	**0.832 (0.021)**	0.788 (0.006)	**0.536 (0.041)**	0.602 (0.008)
“before”	0.831 (0.016)	**0.794 (0.006)**	0.532 (0.057)	**0.619 (0.013)**
”after”	**0.840 (0.022)**	0.790 (0.007)	**0.548 (0.049)**	0.605 (0.011)
“init”	0.830 (0.026)	0.782 (0.020)	0.542 (0.063)	0.594 (0.031)

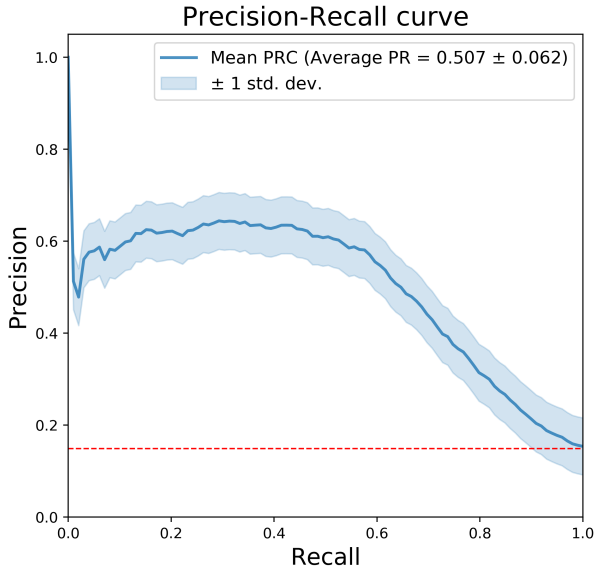
Results - Known and exploratory SNPs

		ROC AUC		Average precision	
Reduction	**Algorithm**	**PPMI**	**DIG-PD**	**PPMI**	**DIG-PD**
”first visit”	LinearSVC	0.715 (0.071)	0.666 (0.038)	0.365 (0.094)	0.425 (0.031)
	XGBoost	0.691 (0.041)	0.625 (0.017)	0.338 (0.081)	0.402 (0.022)
“previous visit”	LinearSVC	0.779 (0.033)	0.792 (0.022)	0.468 (0.048)	0.624 (0.035)
	XGBoost	0.782 (0.029)	**0.791 (0.006)**	0.455 (0.054)	**0.645 (0.020)**
“mean”	LinearSVC	**0.821 (0.015)**	0.784 (0.020)	**0.506 (0.071)**	0.597 (0.026)
	XGBoost	0.815 (0.017)	0.780 (0.010)	0.477 (0.053)	0.567 (0.026)

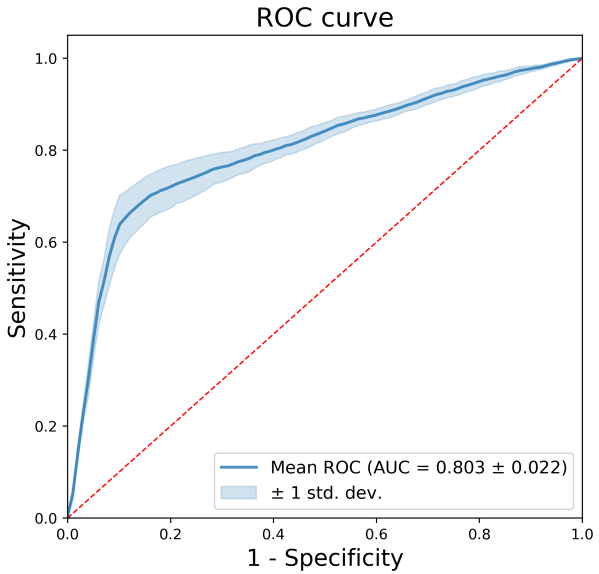
	ROC AUC		Average precision	
Static data	**PPMI**	**DIG-PD**	**PPMI**	**DIG-PD**
”dynamic”	0.828 (0.024)	0.790 (0.008)	0.531 (0.055)	0.621 (0.013)
“before”	**0.841 (0.018)**	**0.793 (0.007)**	**0.547 (0.057)**	**0.624 (0.013)**
”after”	**0.838 (0.024)**	0.794 (0.008)	**0.551 (0.055)**	0.609 (0.008)
“init”	0.832 (0.023)	0.792 (0.008)	0.545 (0.066)	0.611 (0.013)

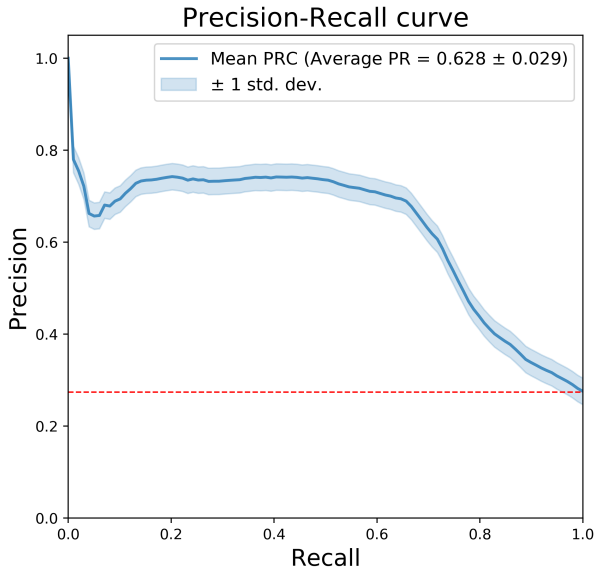
Results - Known SNPs





PPMI





DIG-PD

Results - Known SNPs

PPMI

Time (years)	0.5	1	2	3	4	5	6	7	8
ROC AUC	0.732 (0.164)	0.767 (0.087)	0.807 (0.084)	0.858 (0.082)	0.886 (0.067)	0.836 (0.064)	0.857 (0.090)	0.859 (0.124)	0.675 (0.202)
Average PR	0.477 (0.196)	0.412 (0.155)	0.564 (0.189)	0.584 (0.161)	0.756 (0.146)	0.722 (0.111)	0.642 (0.199)	0.690 (0.181)	0.473 (0.208)

DIG-PD

Time (years)		1	2	3	4	5	6	7	8
ROC AUC		0.731 (0.024)	0.756 (0.012)	0.827 (0.006)	0.861 (0.007)	0.778 (0.008)	0.746 (0.011)	0.794 (0.007)	1.000 (0.000)
Average PR		0.497 (0.012)	0.552 (0.015)	0.707 (0.032)	0.703 (0.007)	0.557 (0.014)	0.514 (0.018)	0.783 (0.014)	1.000 (0.000)

Outline

Conclusion

- **Longitudinal approaches are easier to address:**
  - More data (several time points for each patient)
  - The presence or absence of ICDs in previous visits is useful
  - There is information in several previous visits.

- Deep learning models perform ever so slightly better than cross-sectional models.

- **Replication** on DIG-PD is relatively **good**, **but:**
  - The previous visit is more informative than the mean over all the past visits.
  - Adding genetic data does not improve the predictive performance.

Future work

- Better understand why **genetics** does not help improve predictive performance on **DIG-PD**.

- Better understand what are the **most important features** in the model:
  - Coefficients for relevant models
  - Permutation feature importance

- Make the model available online via a web app.

- Try to **predict other phenotypes** using similar models.

Thanks to…

- My supervisors **Olivier Colliot** and **Jean-Christophe Corvol**
- **Samir Bekadar**
- **ARAMIS-Lab** and **Corti-Corvol group**