Algorithm optimization for signature discovery

Case of Breast Cancers

Ahmed DEBIT

02/27/2018

Laboratory Context

- Diagnosis of breast cancer
- Breast cancer Treatment response

- Design of signatures
- Circulating miRNA

Why Breast Cancer ?

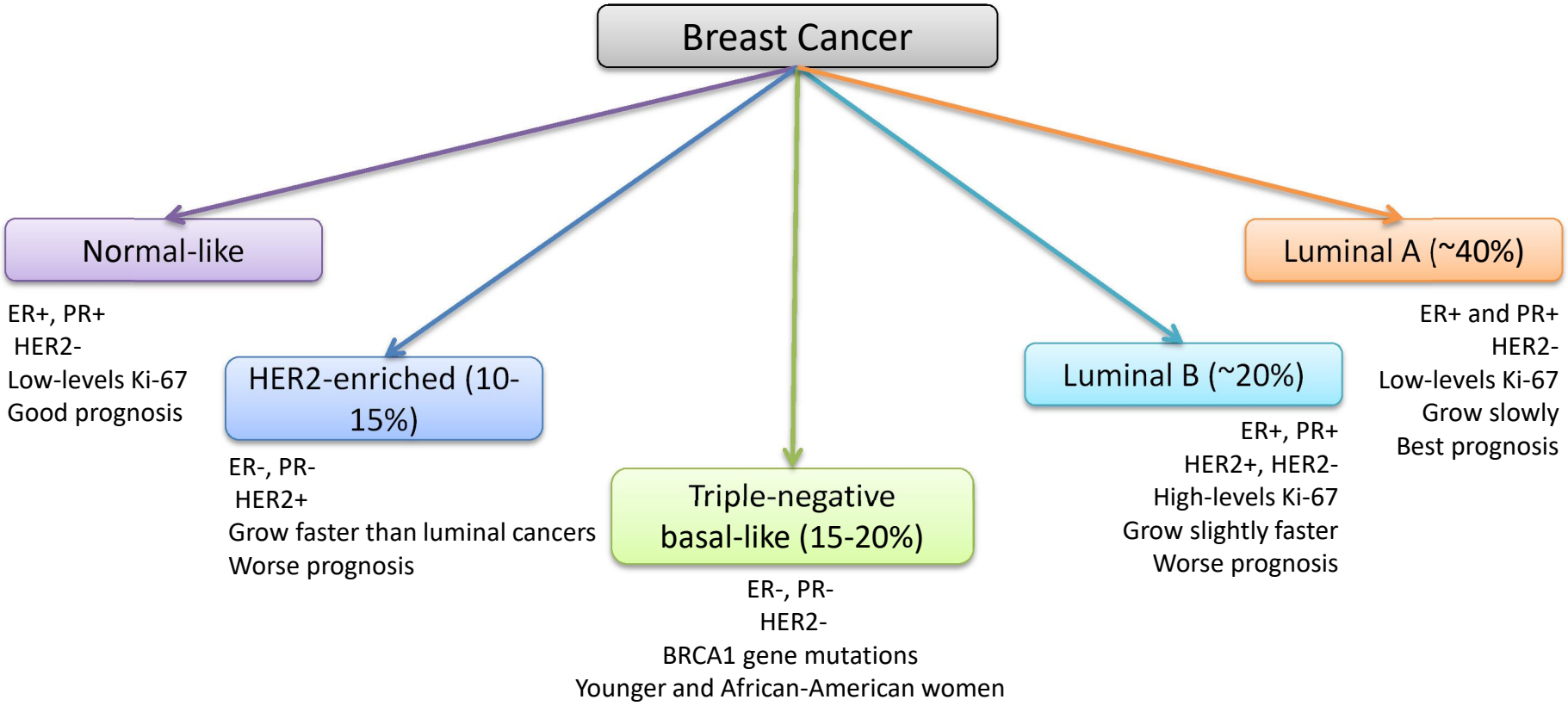
- The most common cancer in women
  - ~ 1.7 million new cases diagnosed in 2012

- The fifth most common cause of death in women

- **Belgium:** Number of women still alive five years after a breast cancer diagnosis 41,418 / 100K

Ferlay J. et al. 2015: Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012

Molecular classification of Breast Cancer



Source: BREASTCANCER.ORG

Why circulating microRNAs ?

- Blood is an accessible source of information
- miRNAs are stable in the blood stream
  - Potential good biomarkers

- <u>Def:</u>
  - Short, single-stranded RNA sequences ~19–23 nucleotides

Freres P. et al. 2014. Neoadjuvant Chemotherapy in Breast Cancer Patients Induces miR-34a and miR-122 Expression
Hamam R. et al. 2017: Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers

Signature Design Goal

- Selection of best markers
- Reduce amount and combine best markers
- Towards an easy clinical routine:
  - Affordable
  - Fast and simple assay

Machine learning can improve decision support

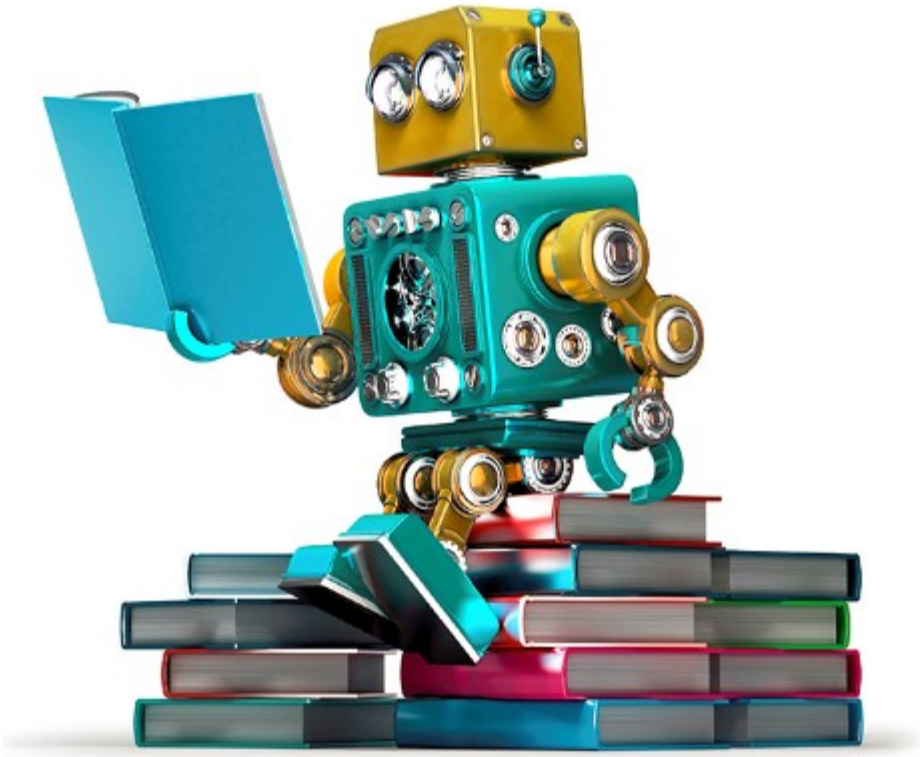


Image source: applikeysolutions.com

Random Forest

SVM

Deep Learning

Unsupervised clustering

Genetic Algorithm

Random Forest

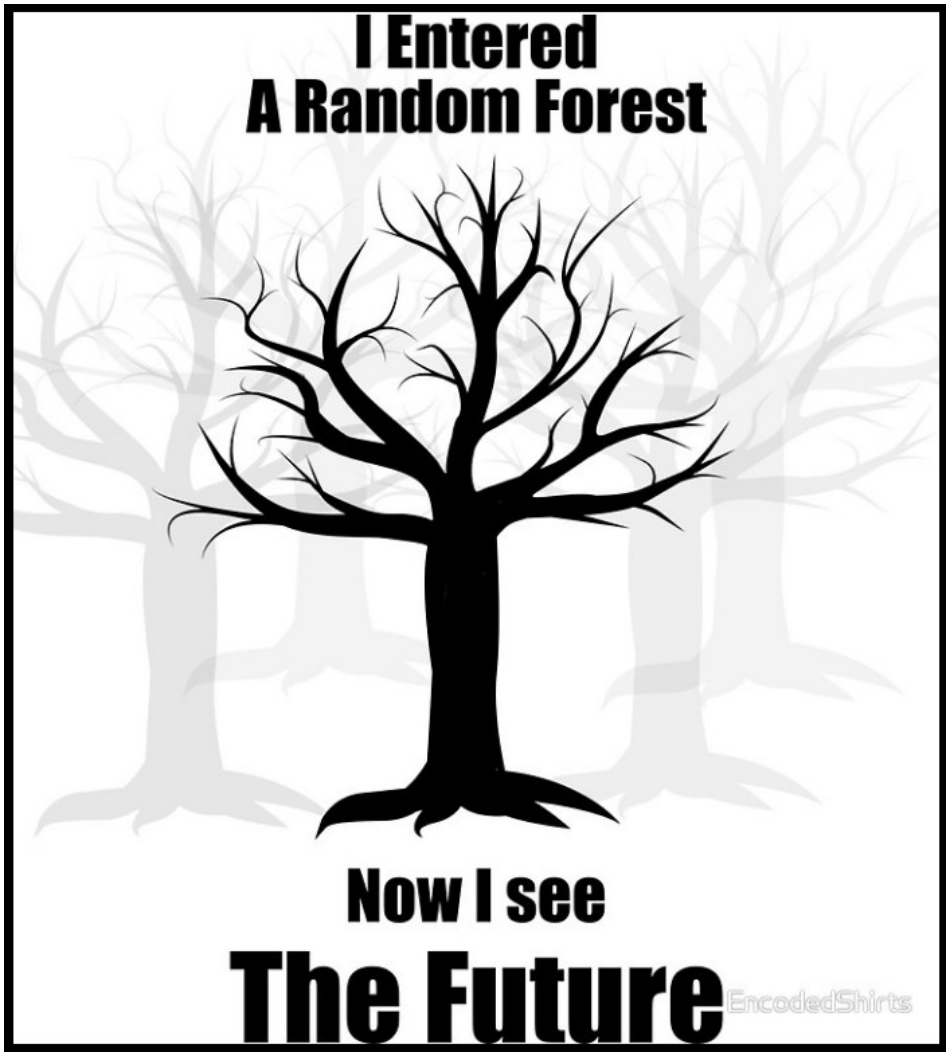
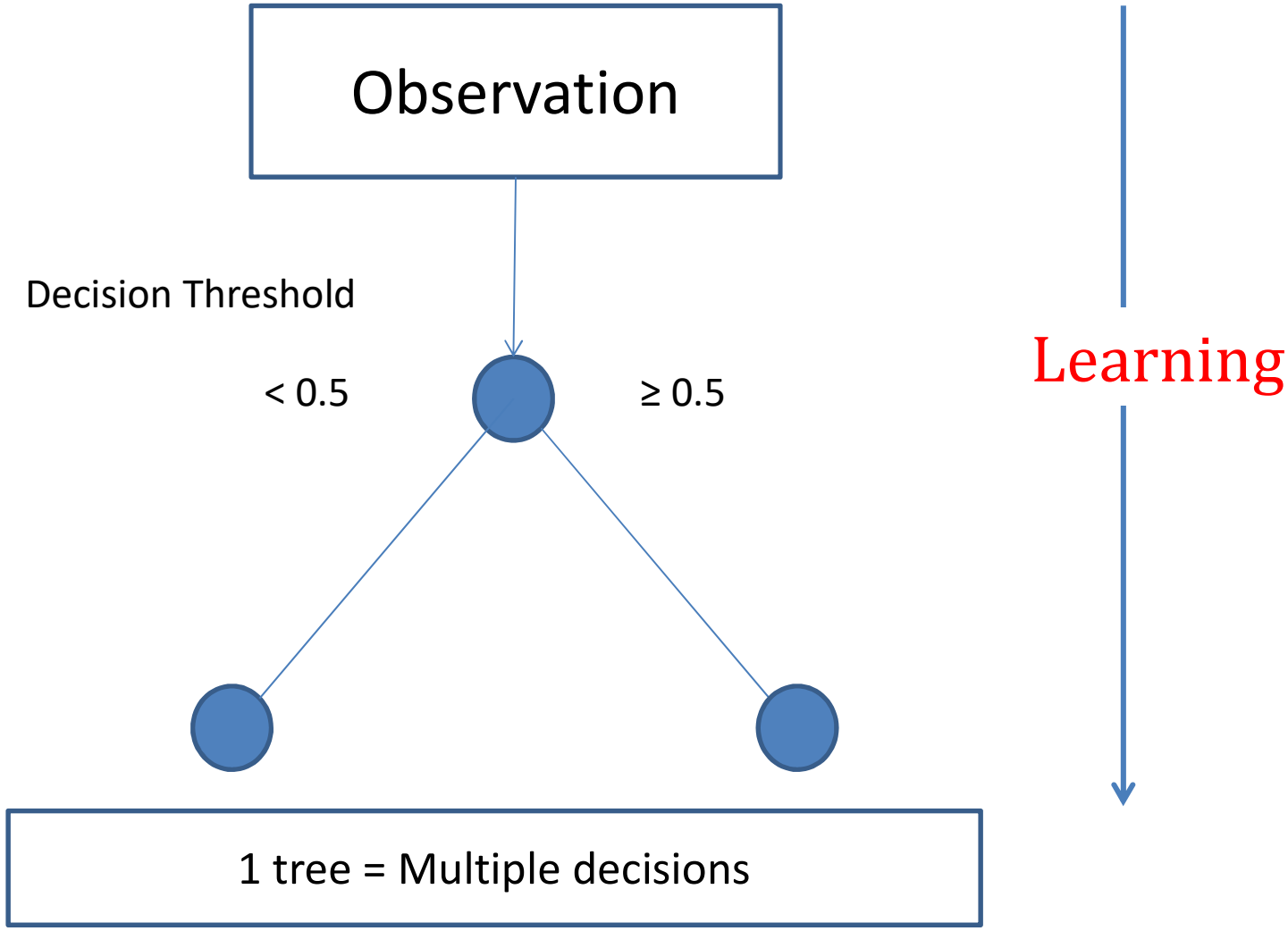


Image source: redbubble.com

A Random Forest
is a set of
random
decision trees

A decision helps to stratify the data

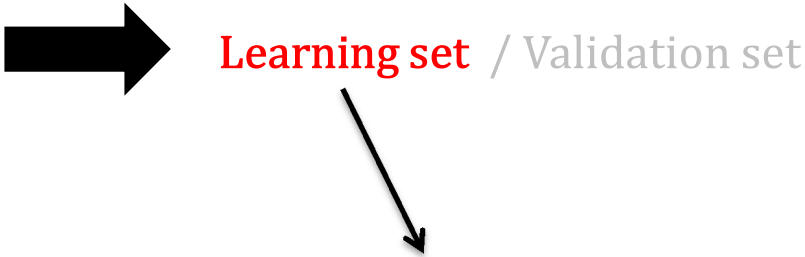


Complete Learning procedure

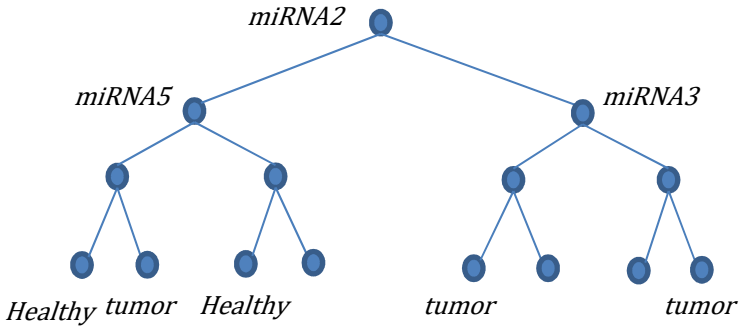
	miRNA1	miRNA2	...	$miRNA_n$
$sample_1$	x_{11}	x_{12}	...	x_{1n}
$sample_2$	x_{21}	x_{22}	...	x_{2n}
	.	.		.
	.	.	...	.
	.	.		.
$sample_m$	x_{m1}	x_{m2}	...	x_{mn}

and

Group
Healthy
Tumor
.
.
.
Healthy



Training set / OOB set



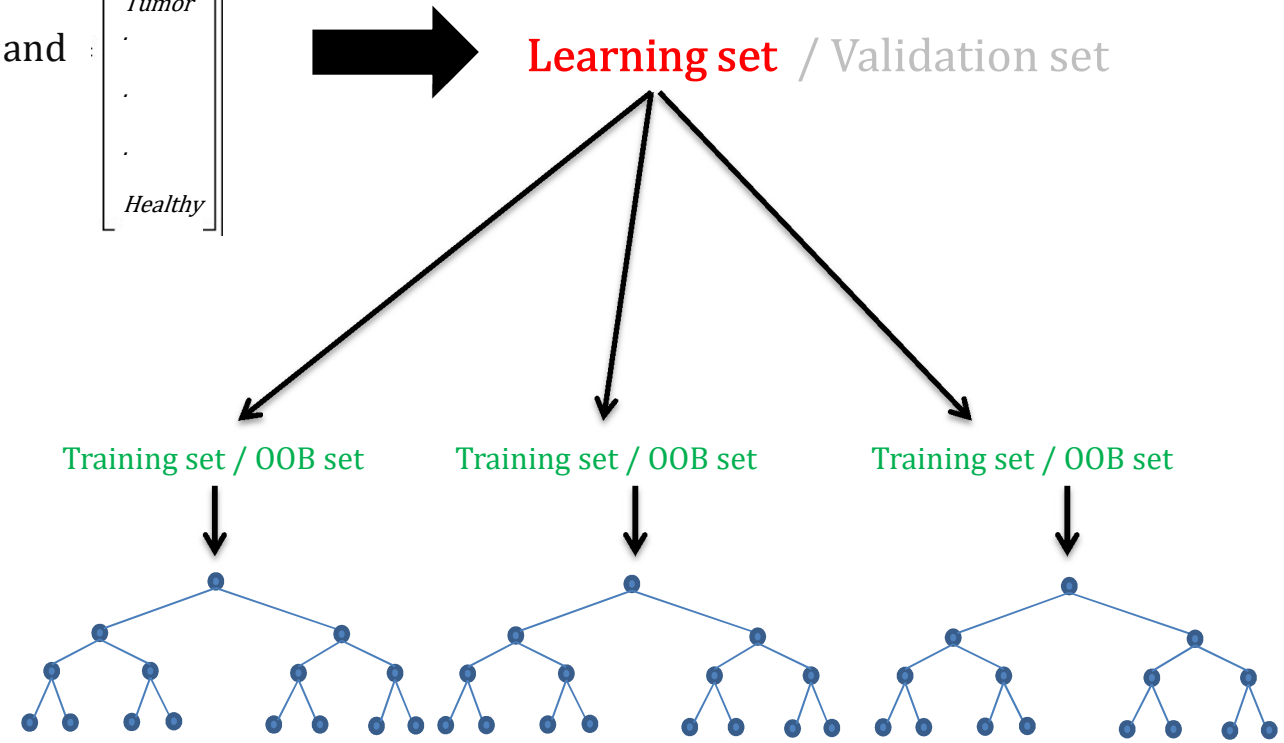
1 tree = Multiple decisions

Random Forest is a set of random trees

	$miRNA1$	$miRNA2$	...	$miRNA_n$
$sample_1$	x_{11}	x_{12}	...	x_{1n}
$sample_2$	x_{21}	x_{22}	...	x_{2n}
	.	.		.
	.	.	...	.
	.	.		.
$sample_m$	x_{m1}	x_{m2}	...	x_{mn}

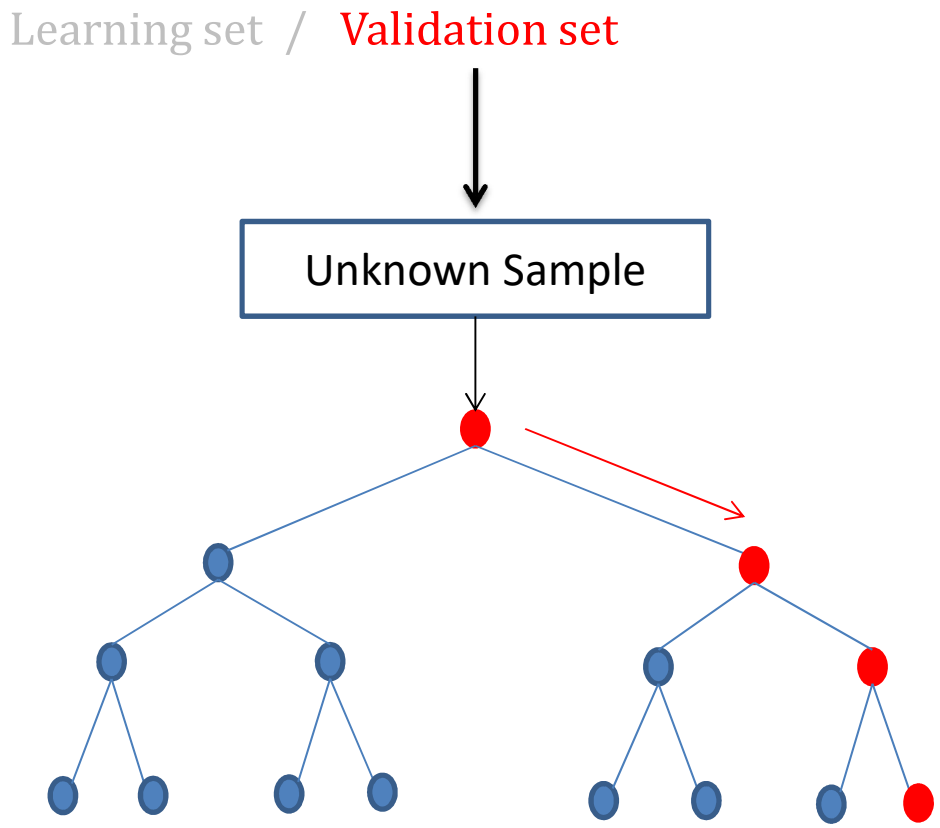
and

$Group$
Healthy
Tumor
.
.
.
Healthy



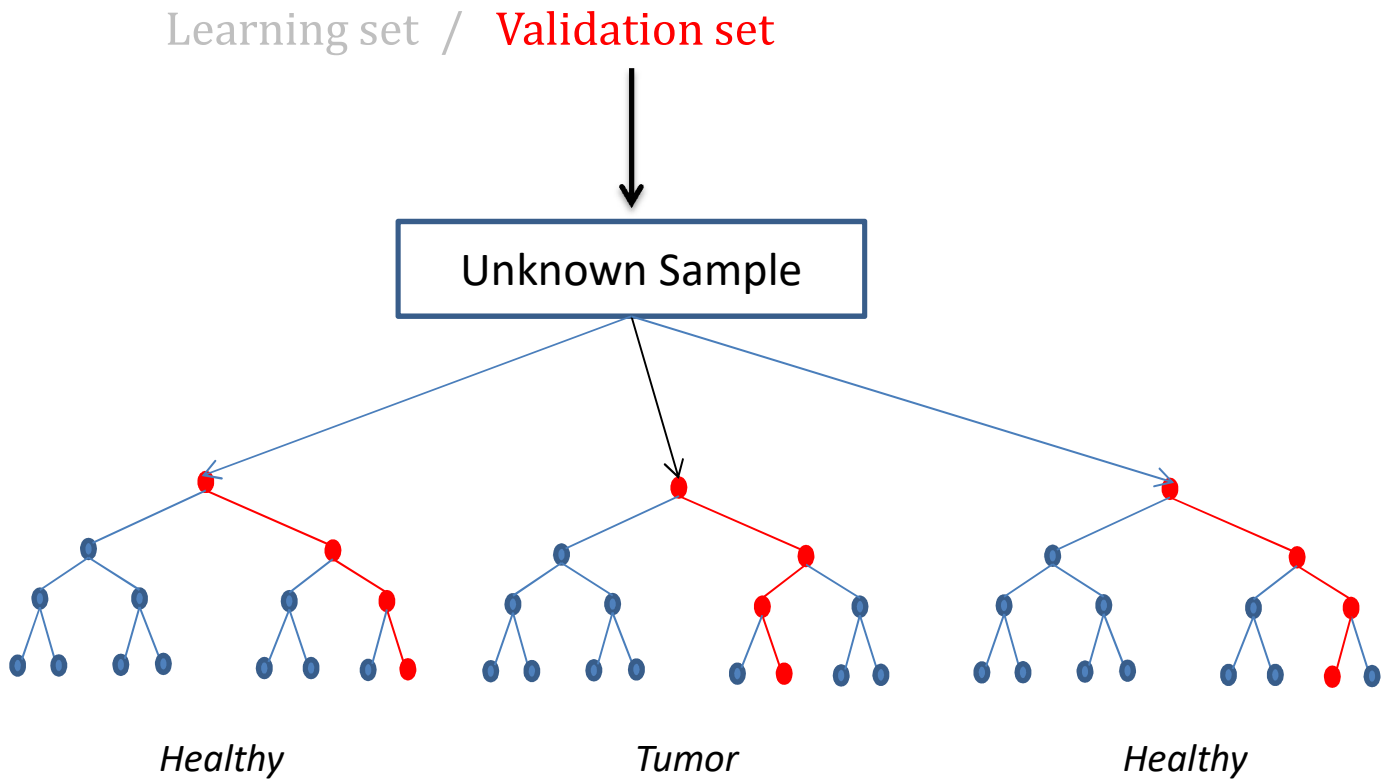
1 Forest = Multiple trees

Prediction procedure.



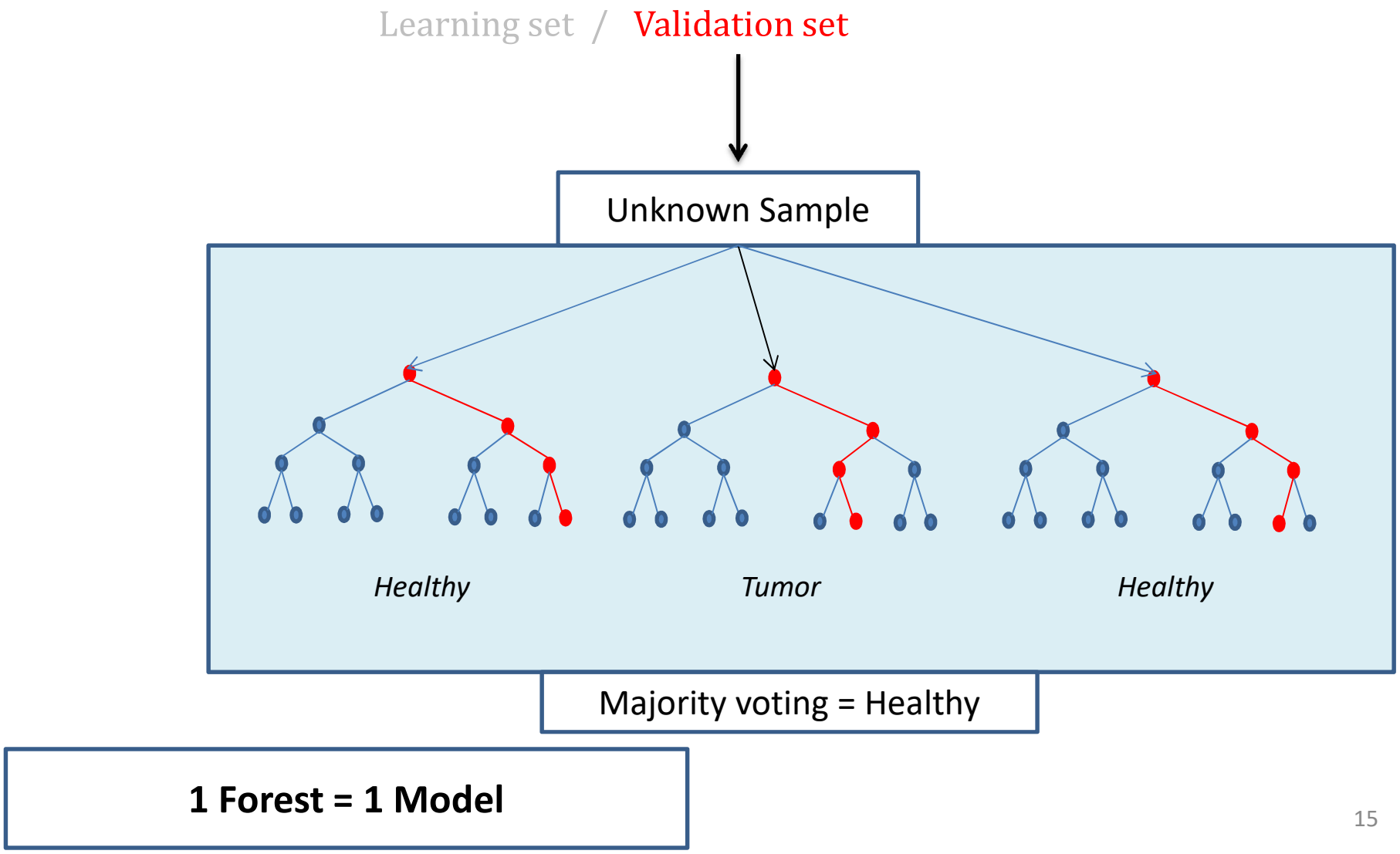
One Tree prediction procedure

Prediction procedure.



Multiple Trees prediction procedure

Prediction procedure.



Towards the best model

- Multiple models are built for the same classification or prediction task.

- Prediction obtained on a model can assess it's classification power: the "AUC".
  - High AUC value = Better prediction.

Breast cancer screening tool

Profiling cohort (n=86)
41 Primary Breast Cancer PBC
45 Controls
Validation cohort (n=196)
108 PBC
88 Controls
 # miRNAs: 188

- 25 important miRNAs
- **#combinations** = 2^25 -1 ~ 33M

- **Best Signature**:

miR-16, let-7d, miR-103, miR-107, miR-148a, let-7i, miR-19b, and miR-22*

Best signature performances

AUC = 0.81
Sensitivity = 91%
Specificity = 49%

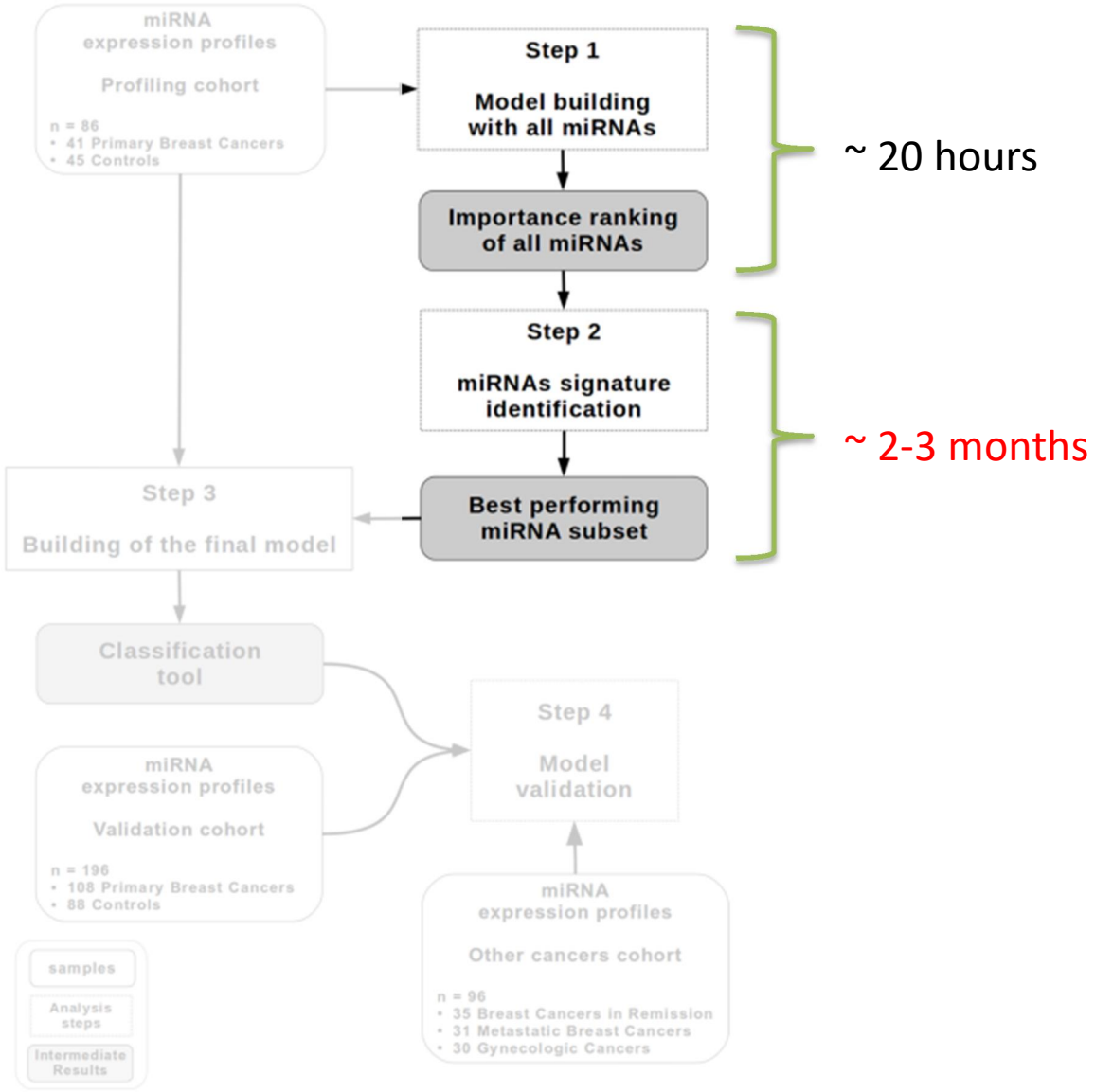
The screening tool can be used as a complementary tool to the mammography test to get a best diagnosis test of Breast Cancer

www.impactjournals.com/oncotarget/ Oncotarget, Vol. 7, No. 5

Circulating microRNA-based screening tool for breast cancer

Pierre Frères[1,2,*], Stéphane Wenric[2,*], Meriem Boukerroucha[2], Corinne Fasquelle[2], Jérôme Thiry[2], Nicolas Bovy[3], Ingrid Struman[3], Pierre Geurts[4], Joëlle Collignon[1], Hélène Schroeder[1], Frédéric Kridelka[5], Eric Lifrange[6], Véronique Jossa[7], Vincent Bours[2,*], Claire Josse[2,*], Guy Jerusalem[1,*]

Methodology used previously



[P. Freres et al. 2015]

Problematic ?

Runtime improvement

Best sensitivity

Best specificity

Best AUC value

Problematic ?

Runtime improvement

Best sensitivity

Best specificity

Best AUC value

Proposed solutions

- Decrease number of combinations

  - PCA-based filtering strategy ?
  - PRPE-based filtering strategy ?

PART I: Reducing of the total number of combinations

PCA-BASED FILTERING STRATEGY

Aims

- Design a method to filter signature combinations

Aims

- Design a method to filter signature combinations
- Idea: combinations with less variability are discarded

Formula

- For variable v

$$contribScore(v) = \sum_{j=1}^{K} a_j$$

a_j: loading of the variable v for PC_j

K: the number of PCs to be included

Formula

- For variable v

$$contribScore(v) = \sum_{j=1}^{K} a_j$$

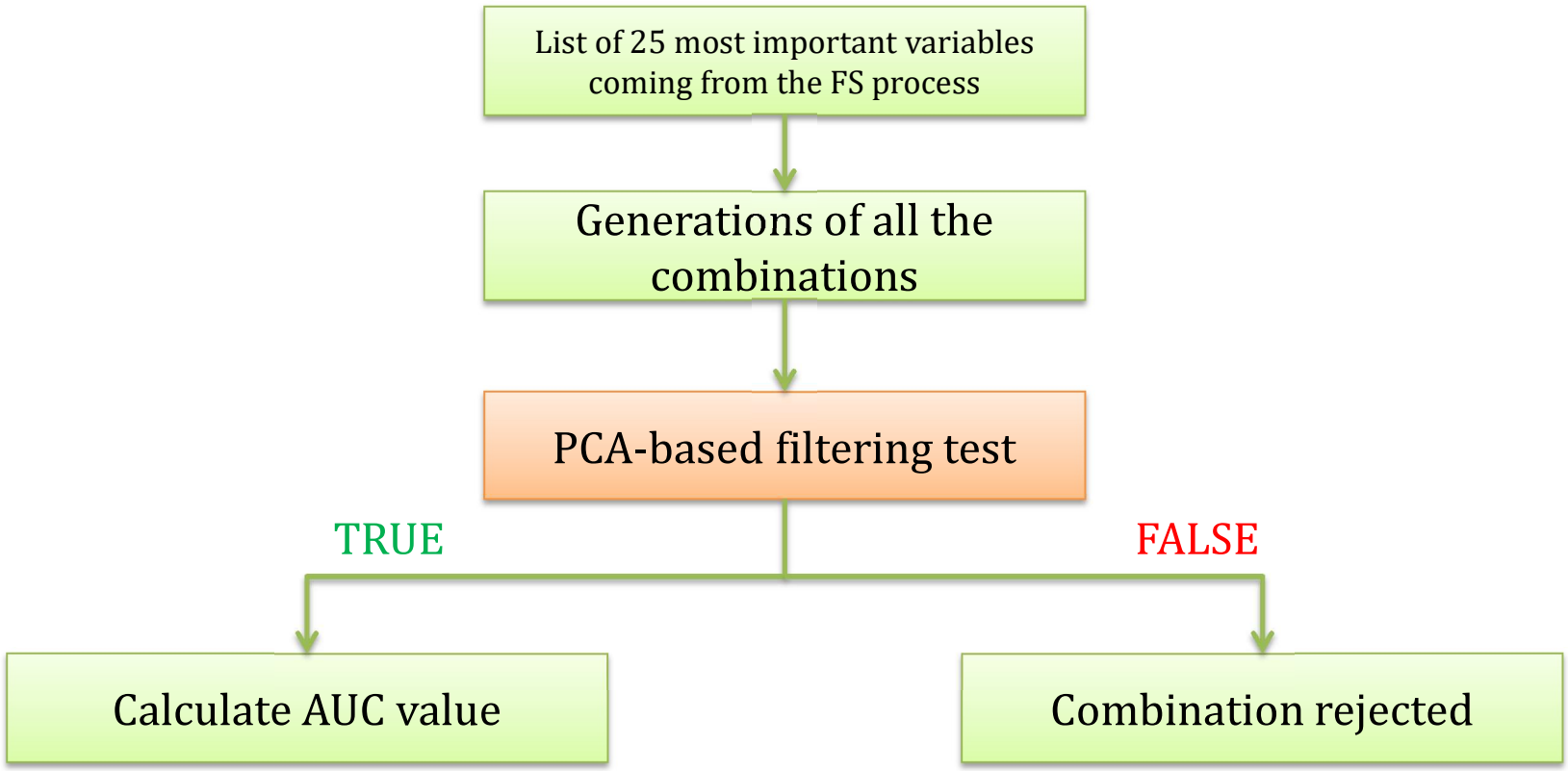
a_j: loading of the variable v for PC_j

K: the number of PCs to be included

- For combination C of x variables

$$contribScore(C) = \sum_{V \in C} contribScore(v)$$

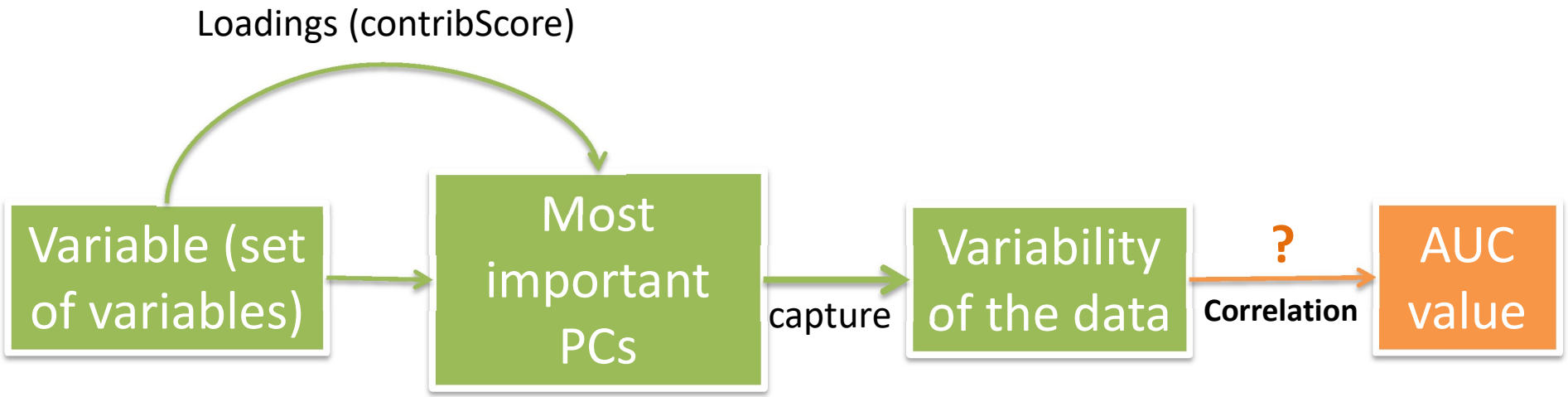
Pipeline


List of 25 most important variables coming from the FS process
Generations of all the combinations
PCA-based filtering test
TRUE
FALSE
Calculate AUC value
Combination rejected

Hypothesis

For one combination:

PCA-based contribution score is correlated with the end-point AUC



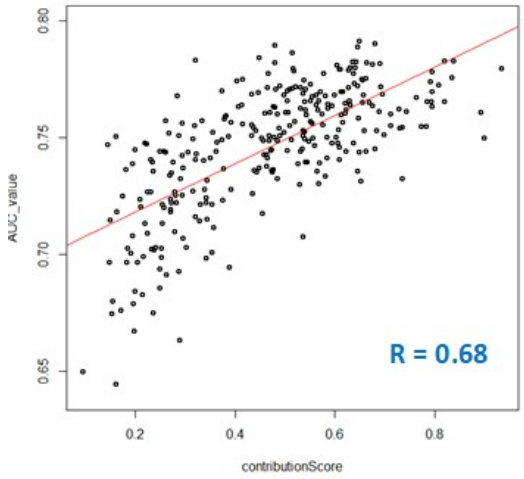
Method

- $N \in \{300, 5400\}$ combinations of different lengths
  - AUC values
  - PCA-based contribution scores

- Correlation between PCA-based score and AUC for different number of PCs?

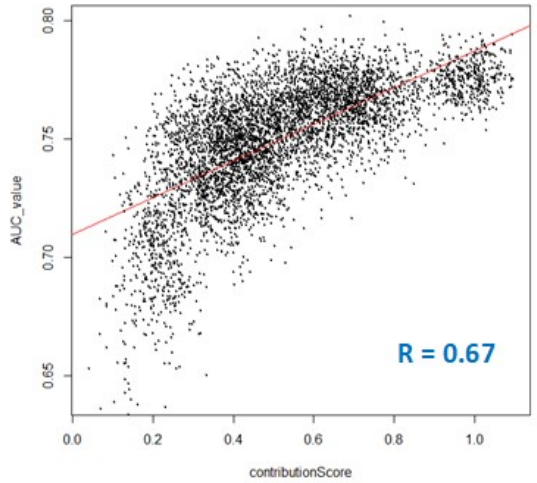
PC1 correlates with AUC

#300



PCs = 1

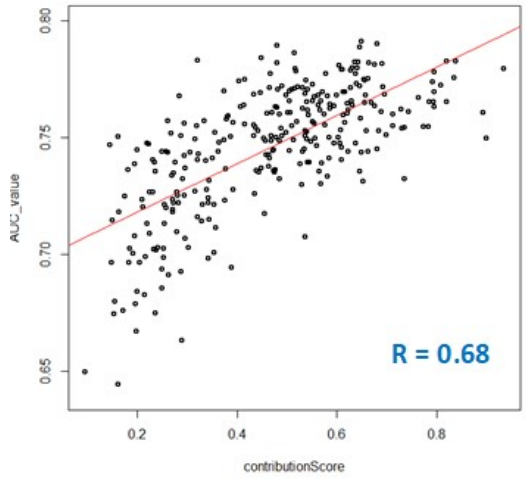
#5000



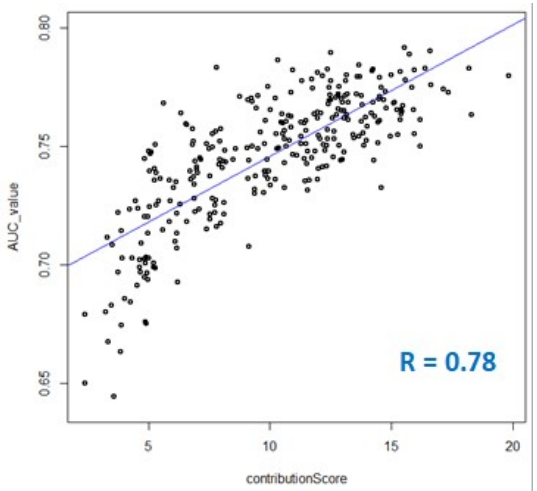
PCs = 1

First 21 PCs correlates better with AUC

#300

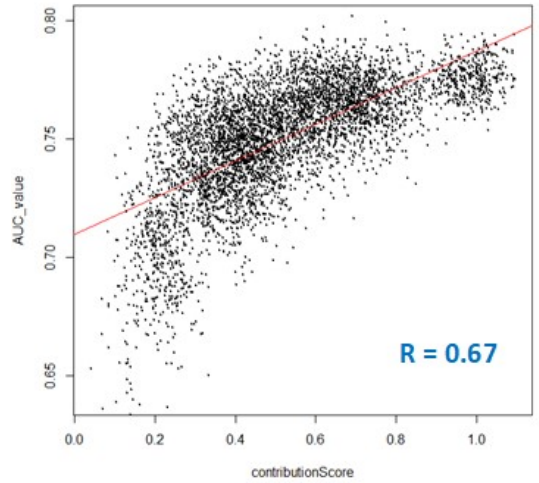


PCs = 1

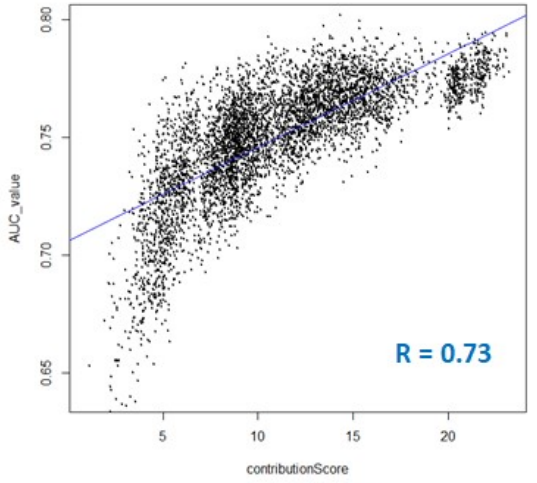


PCs = 21

#5000

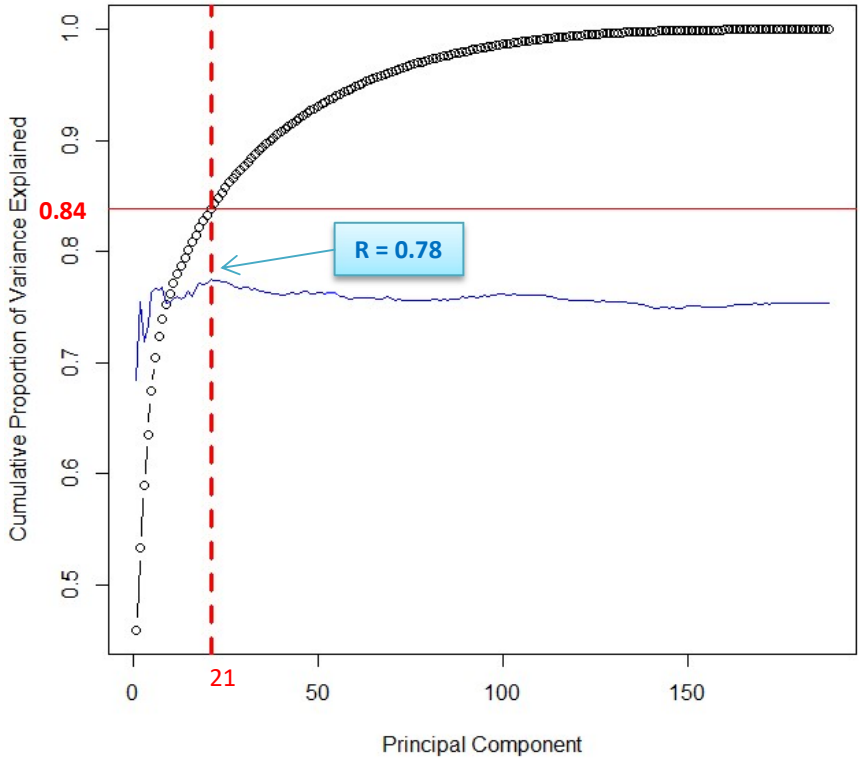


PCs = 1

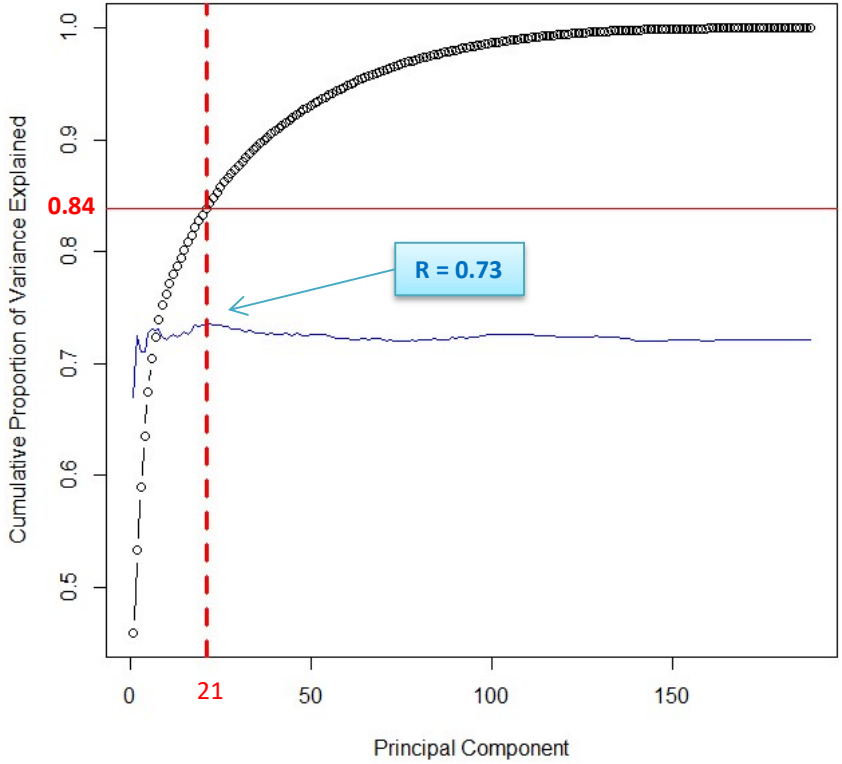


PCs = 21

21 PCs allow for correlation score maximization



#300



#5000

Conclusion

- 21 first PCs can be used for best AUC selection.
- The correlation score should be stabilized.

PART I: Reducing of the total number of combinations

PRPE-BASED FILTERING STRATEGY

Aims

- Design a method to filter signature combinations

Aims

- Design a method to filter signature combinations
- Idea: combinations with less variability are discarded

Proportion of Reduction in Prediction Error (PRPE)

- **PRPE of a variable v**
  - Reduction of prediction error when v is present along with all biomarkers.

Formula

- For variable v

$$prpe(v) = 1 - \frac{e_{all}}{e_v}$$

prpe ↑ → Importance ↑

e_{all}:
error with all variables

e_v:
error with all variables except v

Formula

- For variable v

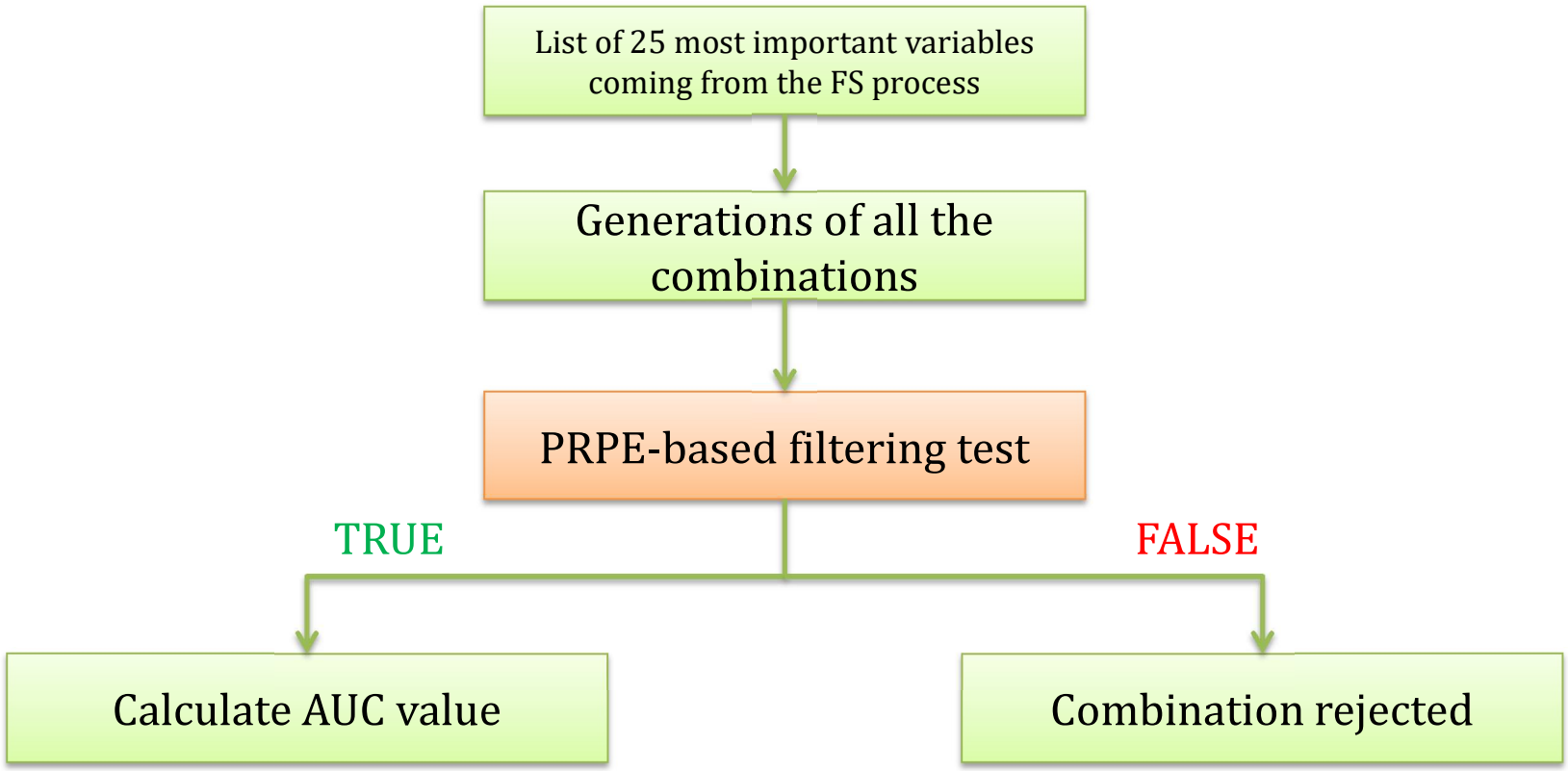
$$prpe(v) = 1 - \frac{e_{all}}{e_v}$$

prpe ↗ → Importance ↗

- For combination C of x variables

$$prpe(C) = \sum_{v \in C} prpe(v)$$

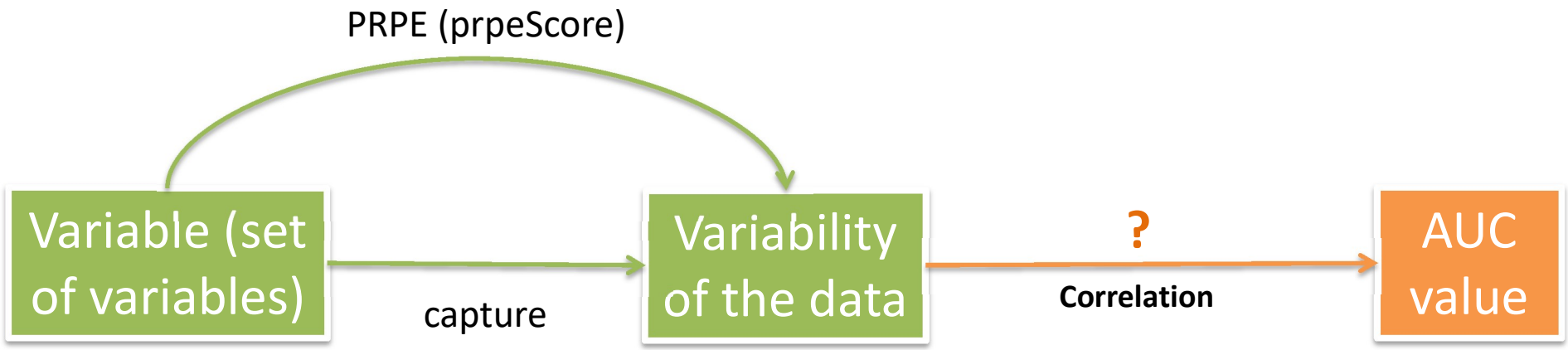
Pipeline


List of 25 most important variables coming from the FS process
Generations of all the combinations
PRPE-based filtering test
TRUE
FALSE
Calculate AUC value
Combination rejected

Hypothesis

For one combination:

PRPE score is correlated with the end-point AUC

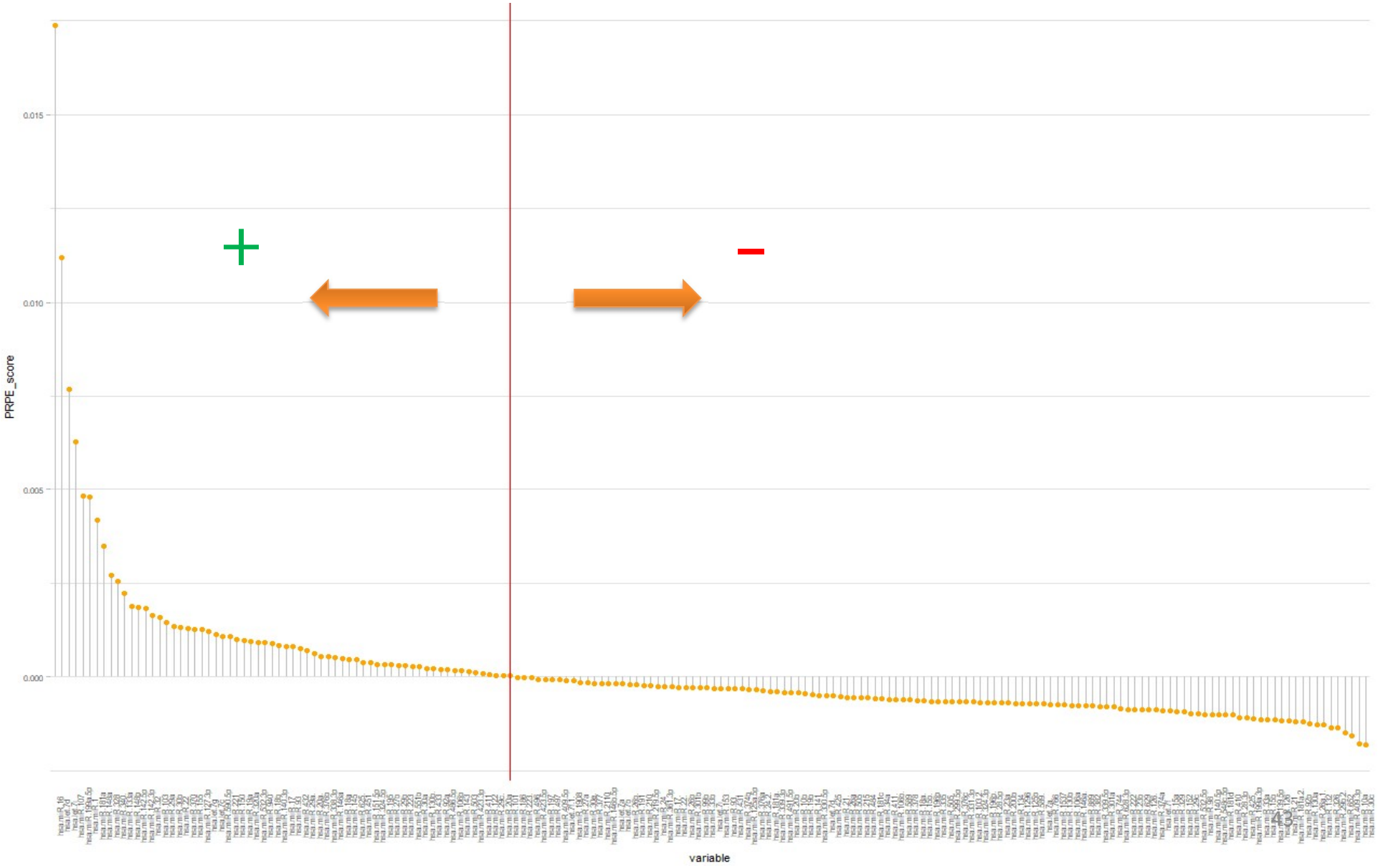


Method

- 5400 combinations of different lengths
  - AUC values
  - PRPE scores

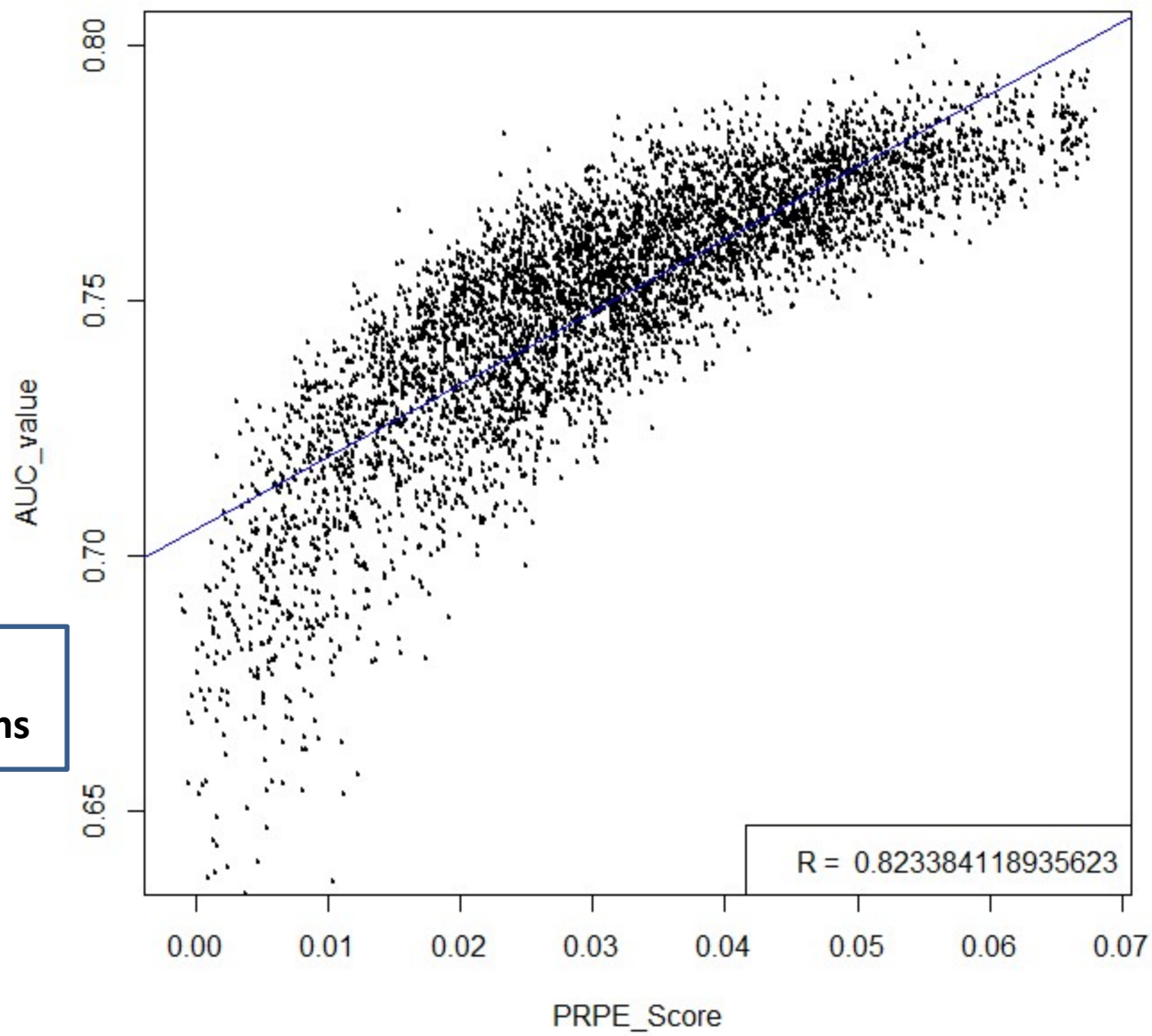
- Correlation between PRPE and AUC?

PRPE-based variable importance

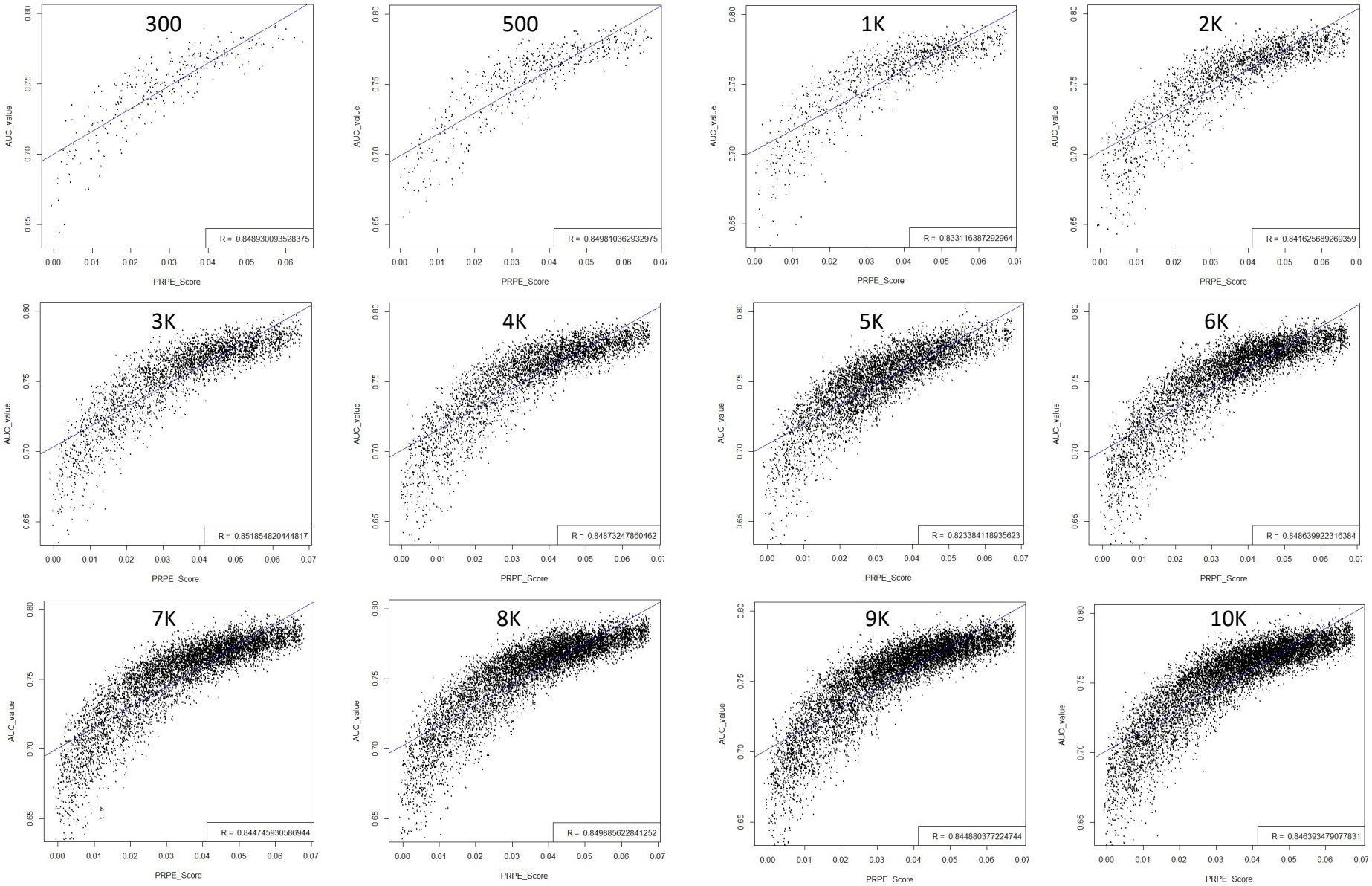


PRPE correlates with AUC

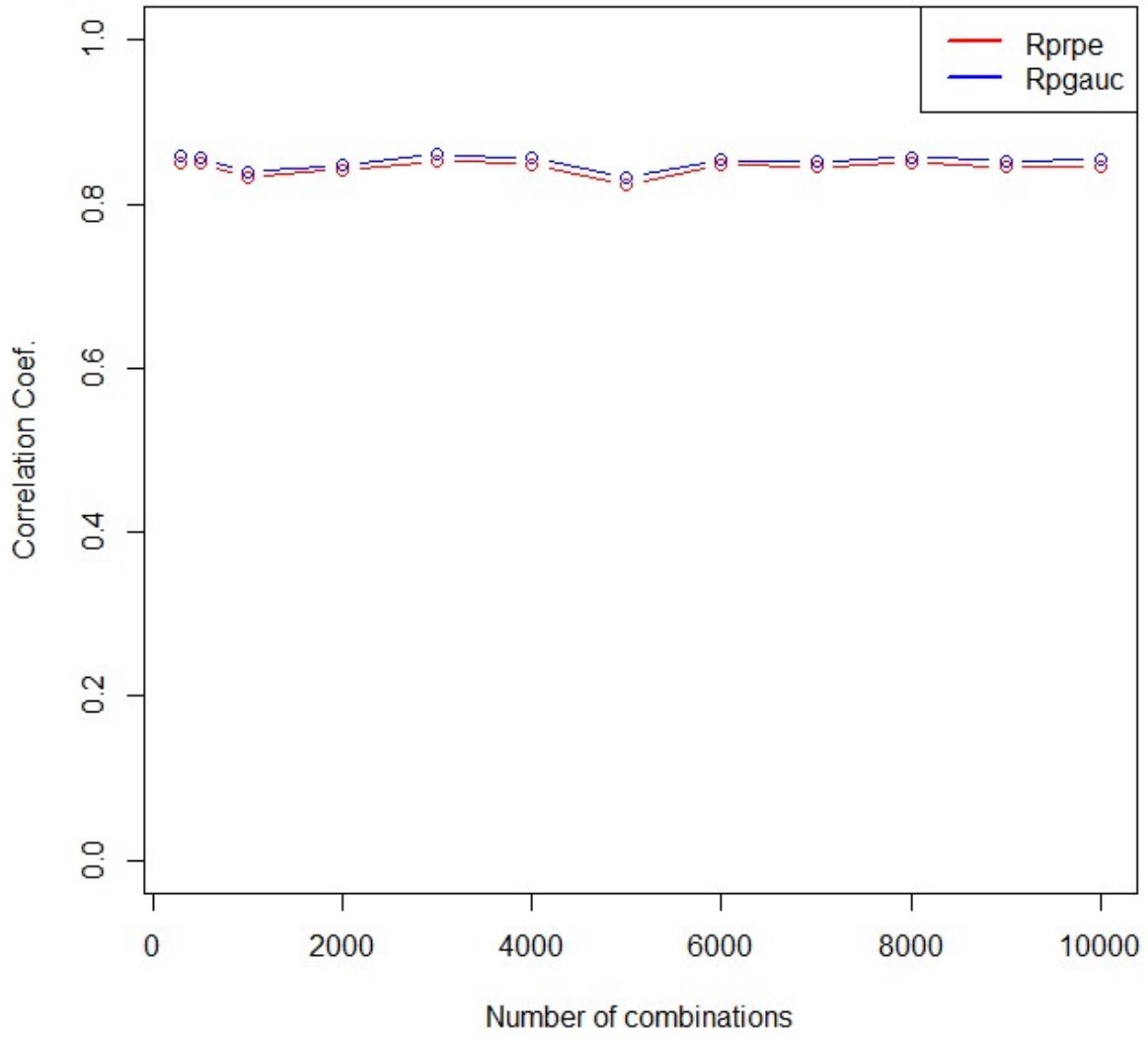
5400 combinations



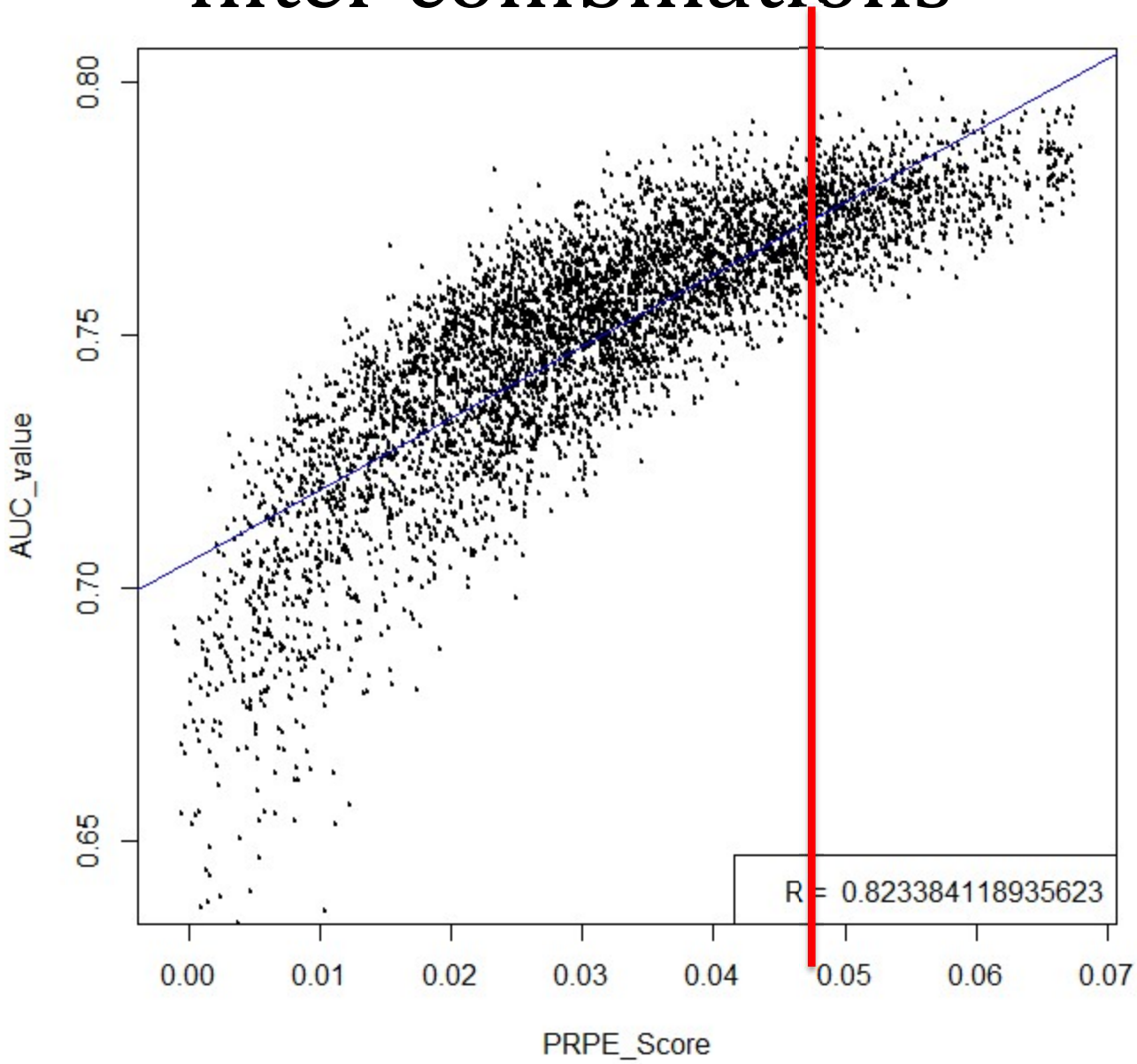
PRPE correlates with AUC



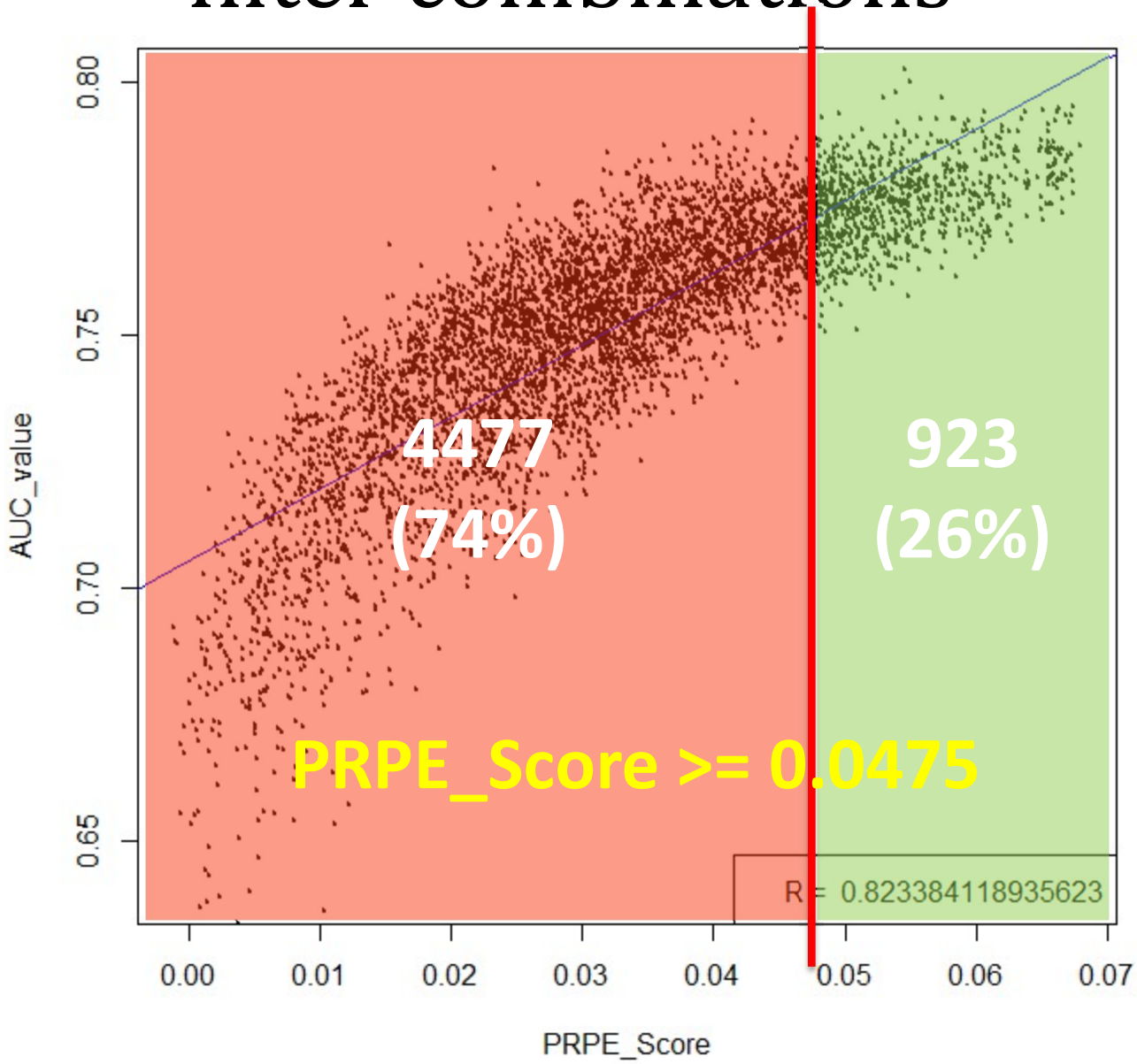
Correlation coefficient is stable



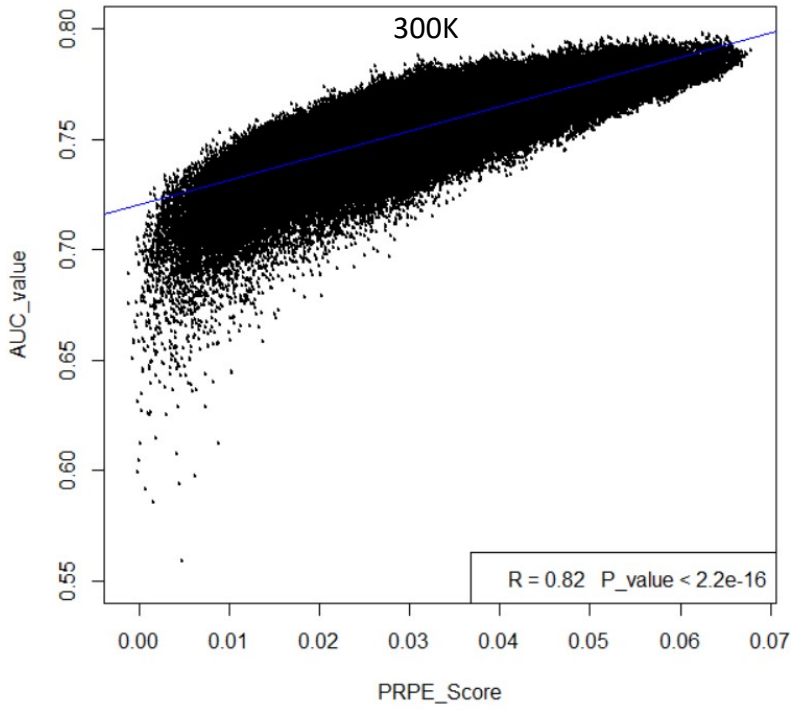
PRPE can be used as threshold to filter combinations



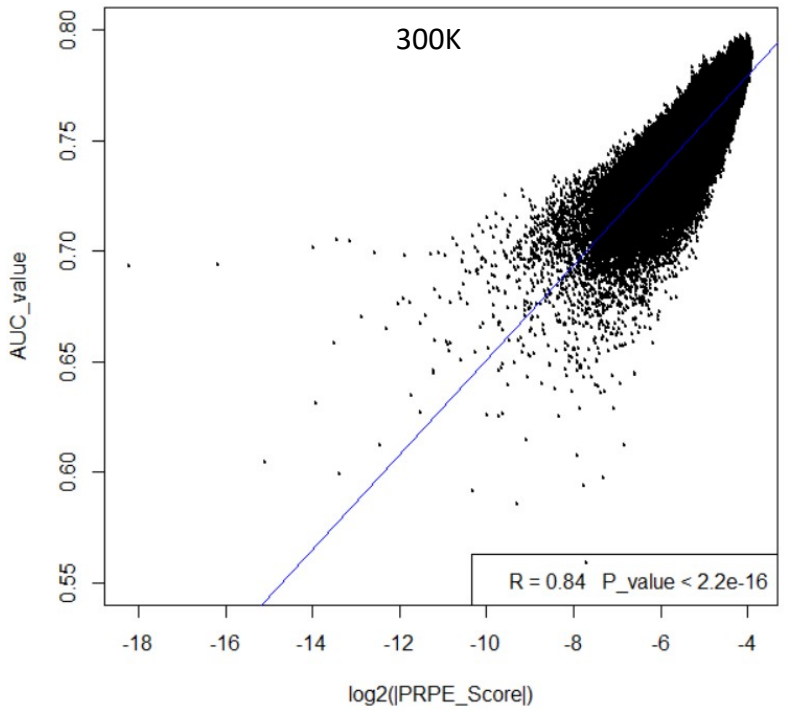
PRPE can be used as threshold to filter combinations



… and for 300K = 1% total number of combinations ?



Linear model
Pearson correlation



Linear model
Pearson correlation

Log transformation of the absolute value of PRPE score improve the correlation coefficient

Application of the PRPE-based filtering strategy

Using of the same 300k combinations and the same random partitions

- Without applying the PRPE-based filter

aucMean	sd	signature
0.7974817	0.02043621	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7970967	0.02020037	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7970063	0.02576832	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.let.7i, hsa.miR.19a, hsa.miR.1, hsa.miR.148a, hsa.miR.19b, hsa.miR.199a.5p, hsa.miR.590.5p
0.7969309	0.02226417	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.miR.93, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7967929	0.02230963	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.miR.93, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7961371	0.02329629	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.let.7f.1., hsa.miR.30b, hsa.miR.199a.5p, hsa.miR.590.5p

- Using the PRPE-based filter

aucMean	sd	signature
0.7975348	0.02027471	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7971715	0.02217562	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.miR.93, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7971576	0.02020627	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7971211	0.02188631	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.miR.107, hsa.miR.93, hsa.let.7i, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p
0.7970352	0.02540237	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.let.7i, hsa.miR.19a, hsa.miR.1, hsa.miR.148a, hsa.miR.19b, hsa.miR.199a.5p, hsa.miR.590.5p
0.7960926	0.02174152	hsa.miR.16, hsa.let.7d, hsa.miR.181a, hsa.miR.103, hsa.let.7i, hsa.miR.19a, hsa.miR.1, hsa.miR.148a, hsa.miR.199a.5p, hsa.miR.590.5p

PRPE-based filtering method finds the same most performant combinations with the same AUC values

Application of the PRPE-based filtering strategy

Method	Total #combinations	#processed combinations	#eliminated combinations	Processing time† (min)
Exhaustive*	300k	300k	0	4578
PRPE-based	300k	42074	**257926**	**675**

* The method used in [P. Freres et al. 2015]

† Using 1000 jobs launched as parallel tasks on Lemaitre2 (CECI's cluster), 50 jobs are allowed to be run in parallel. Each single job deals with 300 combinations.

PRPE-based method reduces drastically the processing time

Conclusion

- PRPE is a filtering tool for best AUC combinations.
  - Stable correlation score.

- Validation on 10% combinations (~ 3M)?

Conclusion

- PRPE filtering is better than PCA to :
  - Get best AUC signatures
  - Shorten run-time processes

Perspectives

- Further run-time and AUC improvements:
  - Compare Random Forests methodologies
    - Efficacy
    - Precision
    - Run-time
  - Improve best model selection

Acknowledgment

Human Genetics (GIGA)

- Prof. Vincent Bours, MD, PhD
- Christophe Poulet PhD
- Stephane Wenric, Ir, PhD,
- Corinne Fasquelle, Ir

Oncology (CHU)

- Prof. Guy Jerusalem, MD
- Dr. Claire Josse
- Pierre Freres, MD, PhD
- Aurelie Poncin, MD
- Jerome Thiry

BIO3 Unit (GIGA)

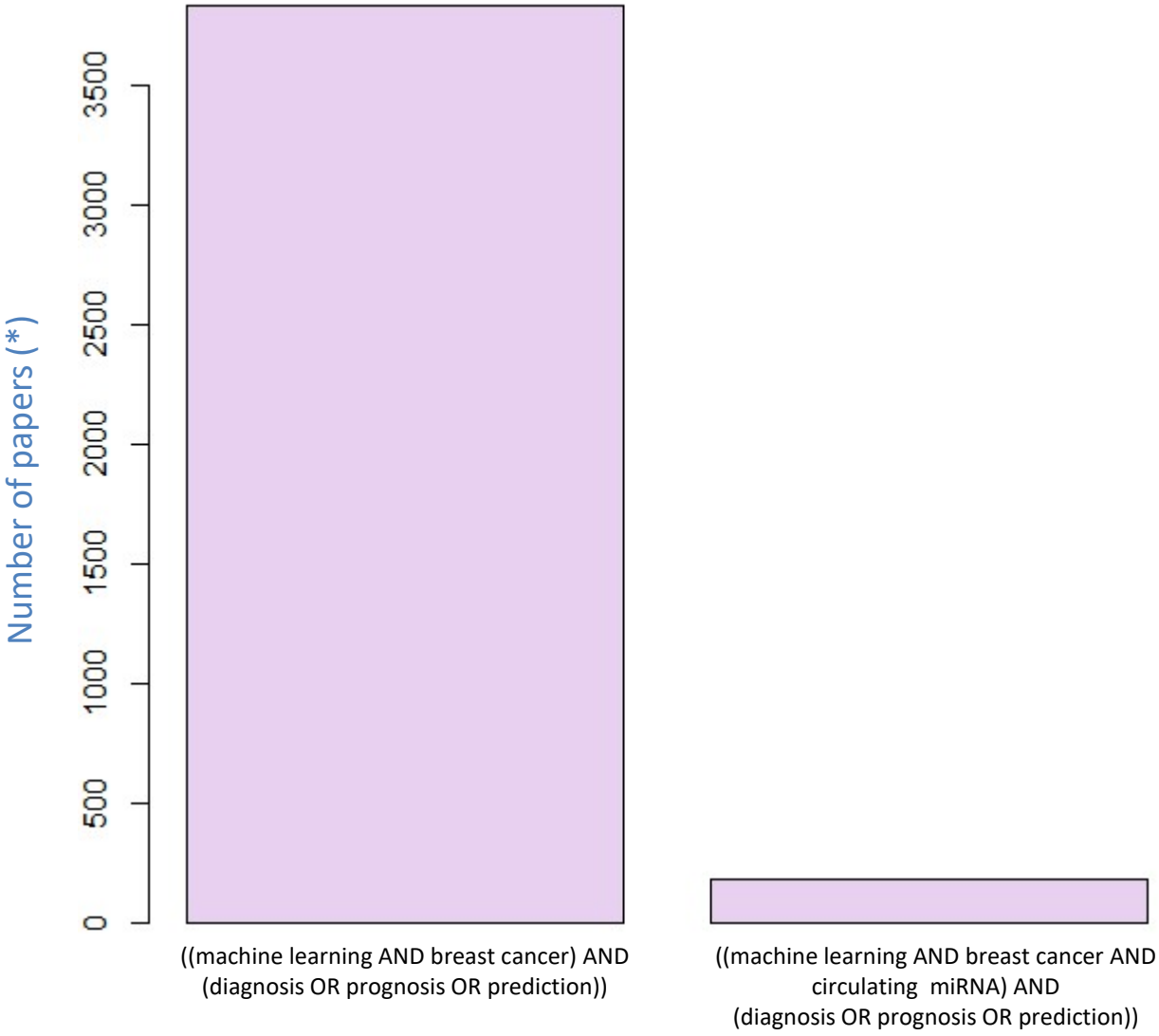
- Prof. Kristel Van Steen

CBIO, Mines ParisTech, Institut Curie

- Chloe-Agathe Azencott PhD

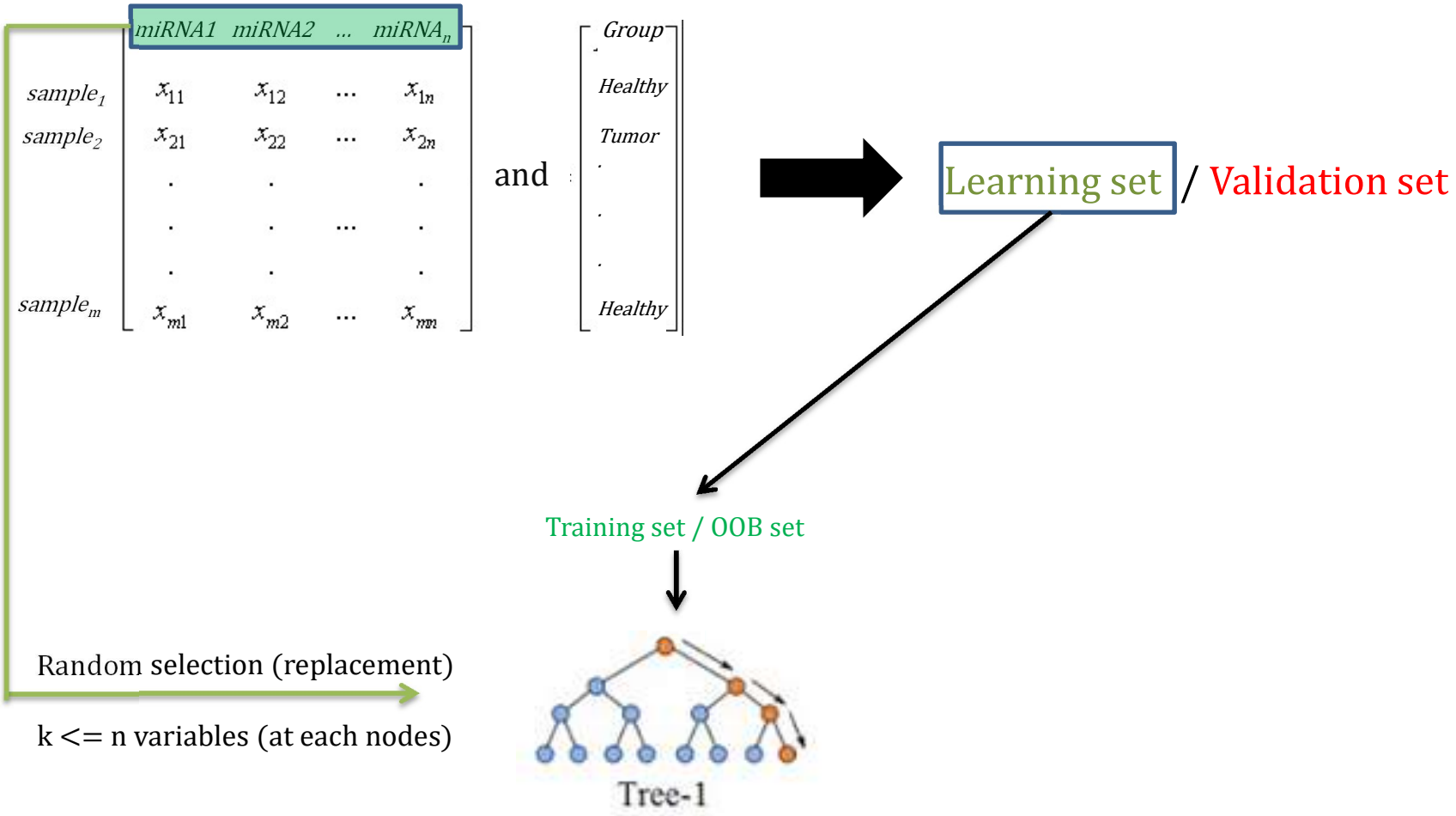
Thanks for your attention

Machine learning researches on Breast Cancer

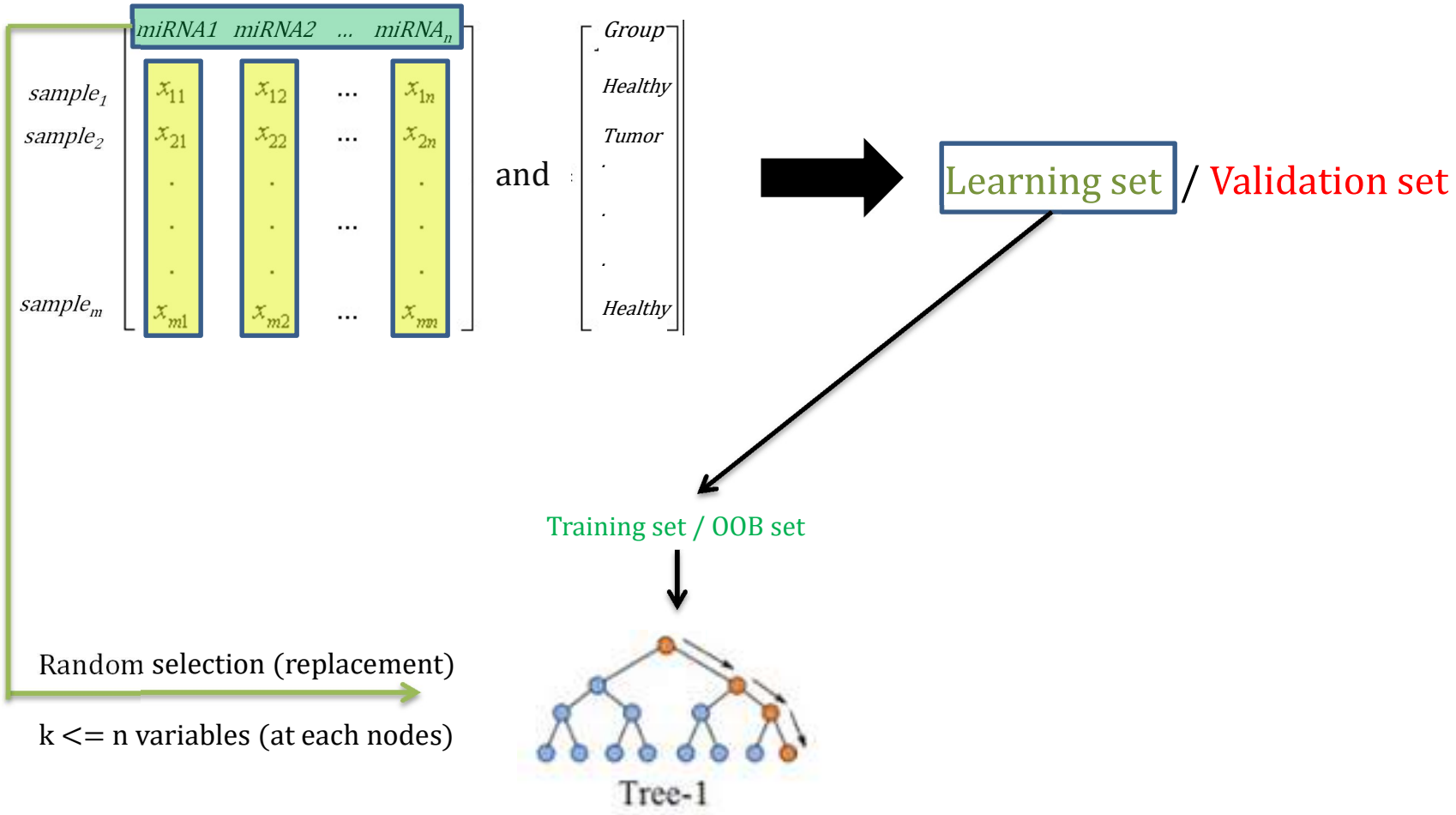


* PMC Published in the last 5 years

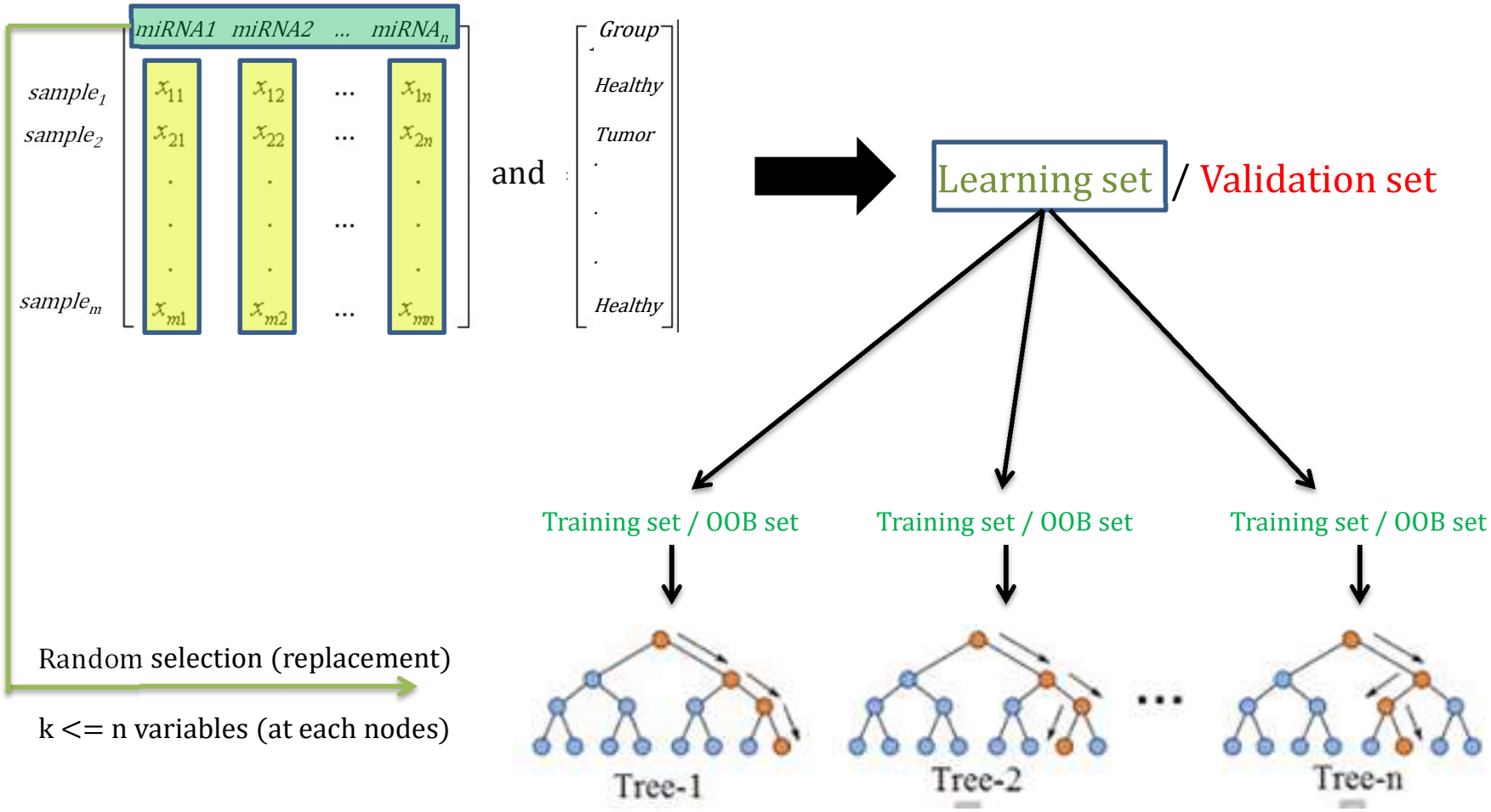
Random Forest is a set of random trees



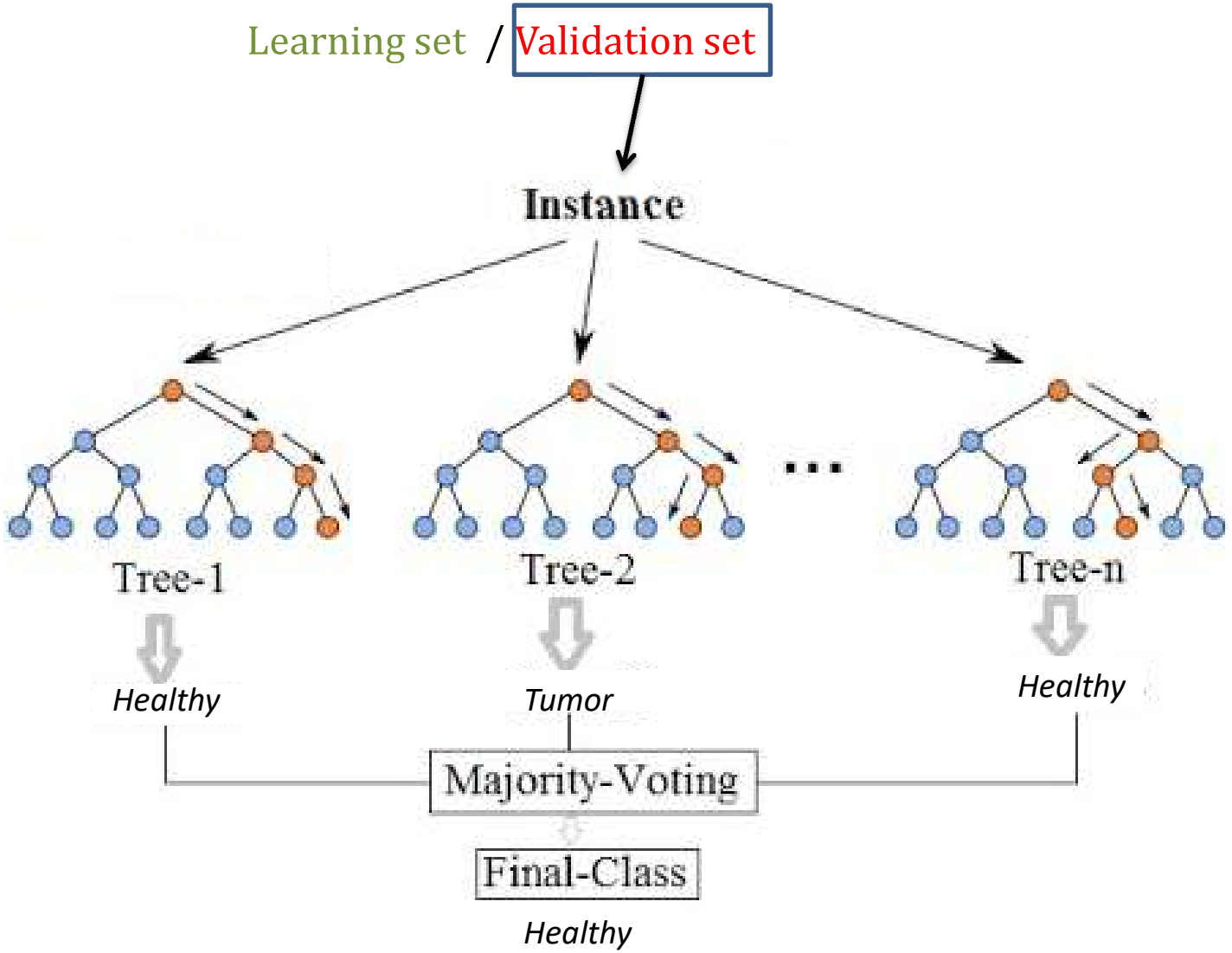
Random Forest is a set of random trees



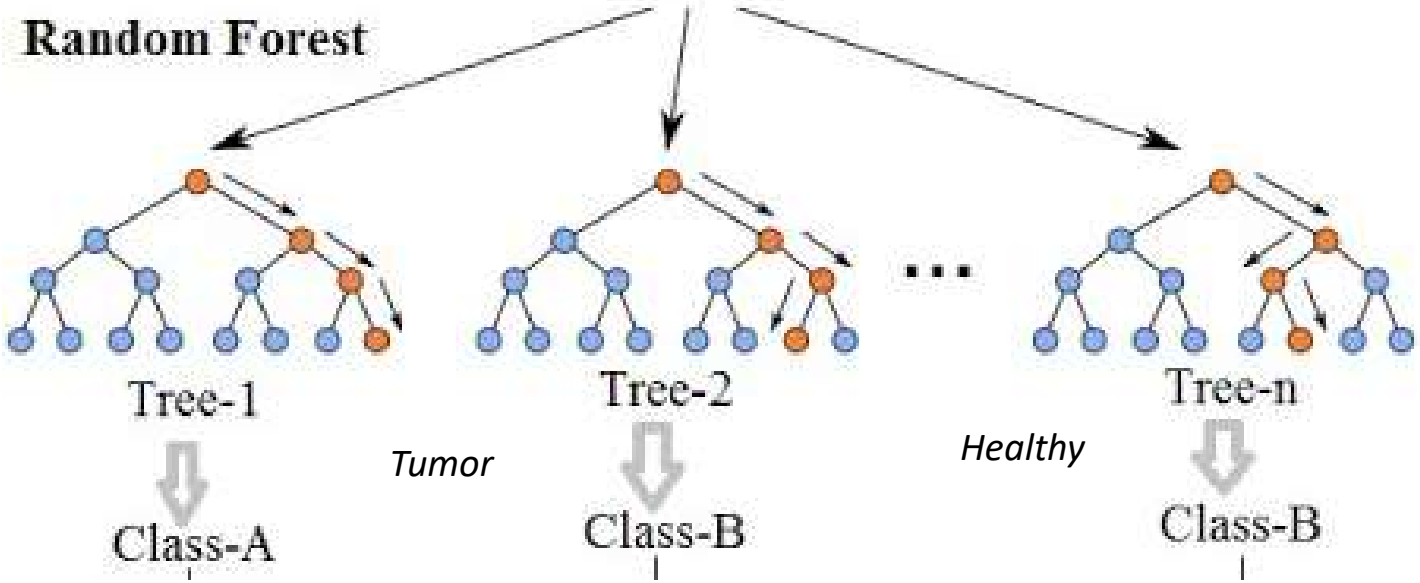
Random Forest is a set of random trees



Random Forest: prediction ?



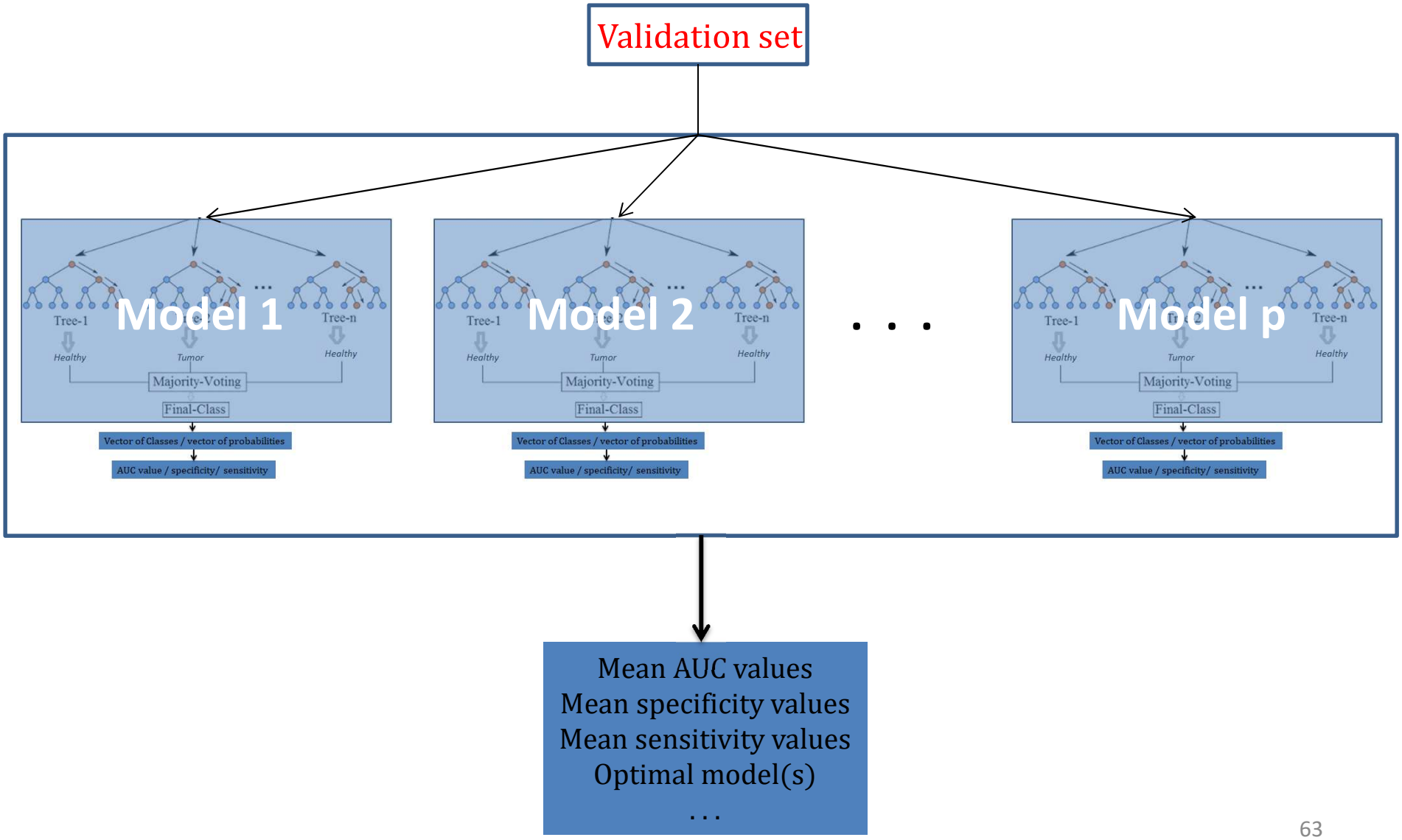
Prediction procedure



Vector of Classes / vector of probabilities

AUC value / specificity/ sensitivity

Random Forest: prediction using a set of models?



Variable Importance

- Explain the GINI index and the Information Gain, and the MDA and MDG
- Some method of feature selection