LIÈGE université
GIGA

# Towards an accurate cancer diagnosis modelization: Comparison of Random Forest strategies

Ahmed DEBIT

11/30/2018

# Laboratory Context

- Diagnosis of Breast Cancer
- Breast Cancer treatment response
- Design of signatures

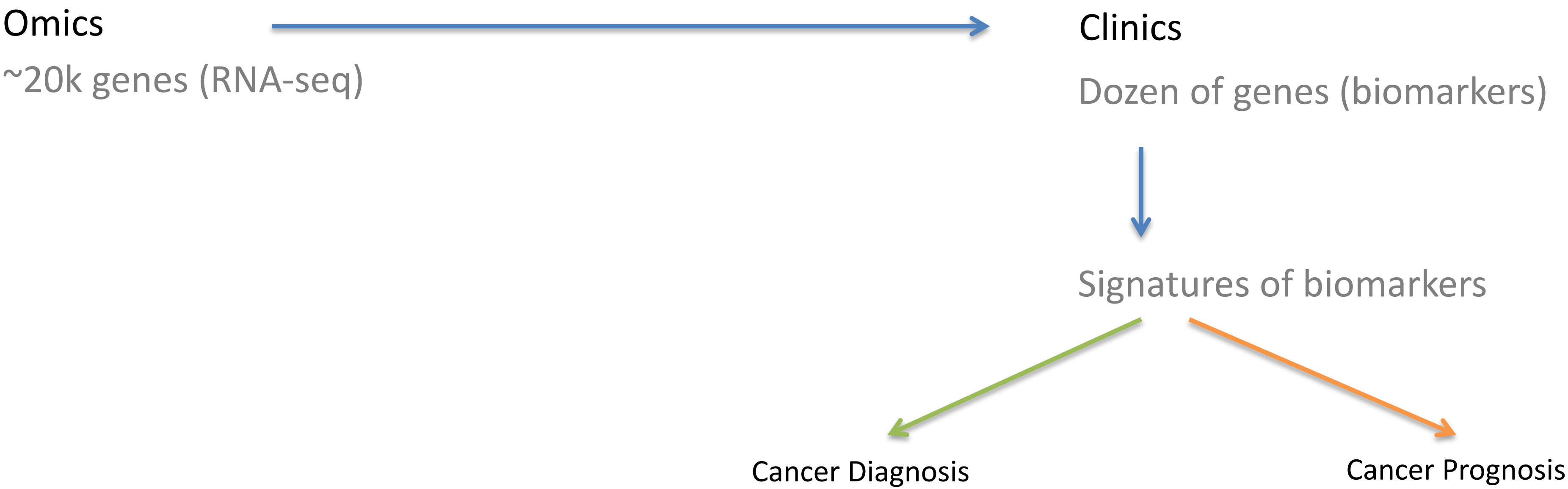
# From Omics to Clinics

Omics → Clinics

~20k genes (RNA-seq)

Dozen of genes (biomarkers)

# From Omics to Clinics



Omics

~20k genes (RNA-seq)

Clinics

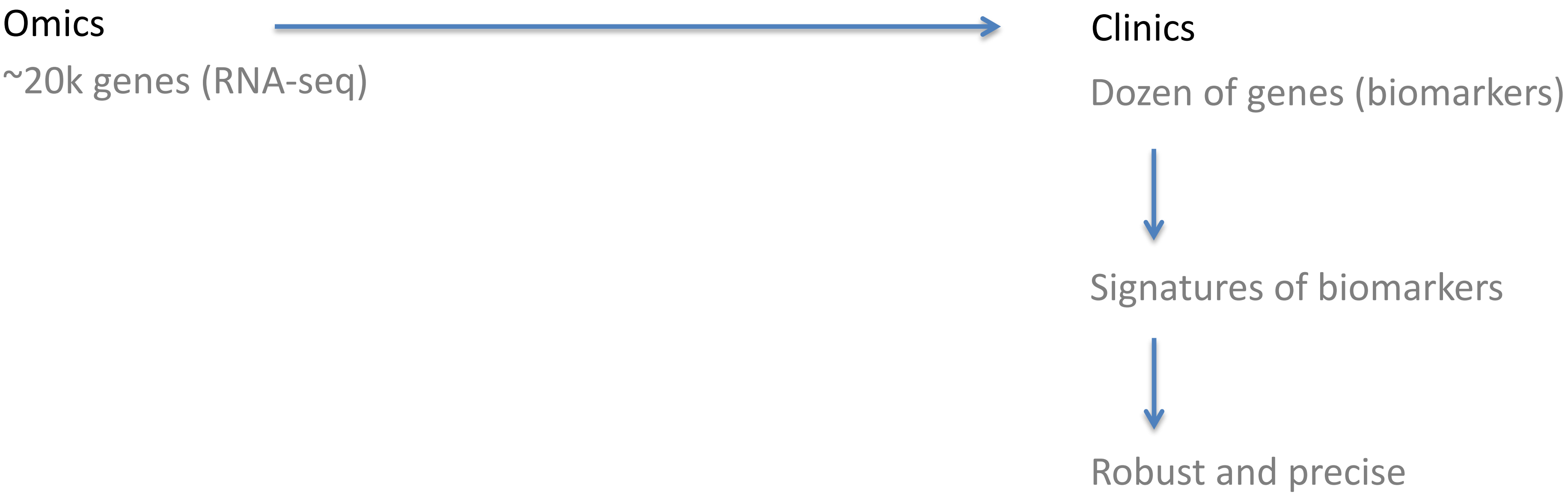
Dozen of genes (biomarkers)

Signatures of biomarkers

Cancer Diagnosis

Cancer Prognosis

# From Omics to Clinics



Omics

~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise

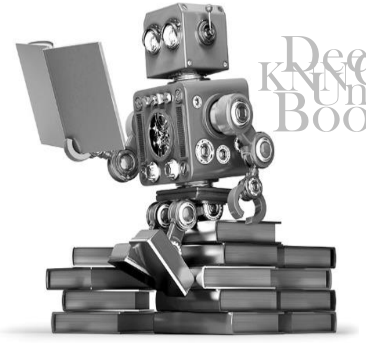
# From Omics to Clinics

Omics

~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)



SVM

Deep Learning

KNN

Genetic Algorithm

Unsupervised Clustering

Boosting

Neural Networks

**Random Forest**

Signatures of biomarkers

Robust and precise
Models and signatures

# Random Forest

Omics

~20k genes (RNA-seq)

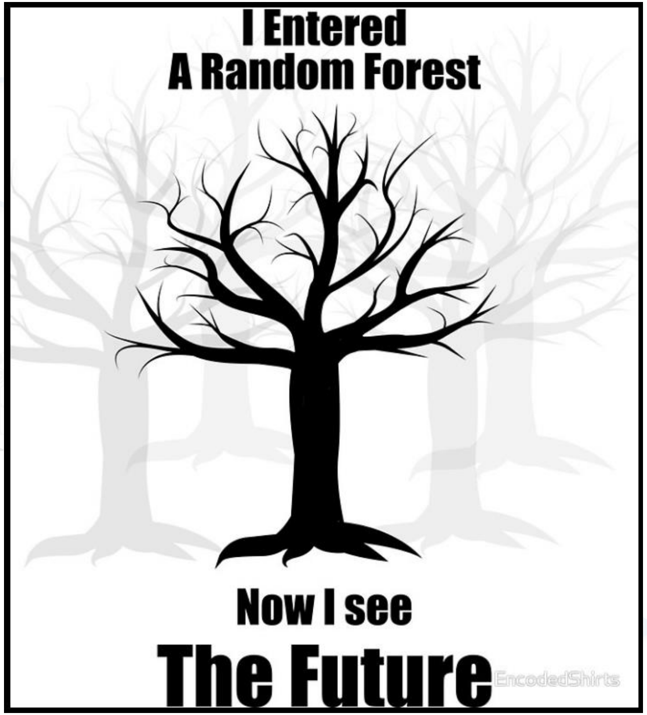


**Image source:** redbubble.com

Clinics

Dozen of genes (biomarkers)

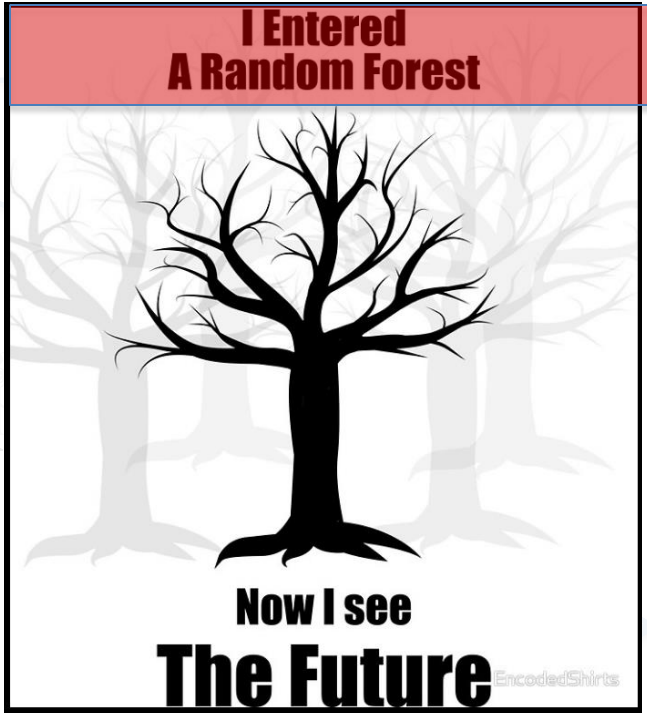
Signatures of biomarkers

Robust and precise
Models and signatures

# Random Forest

Omics

~20k genes (RNA-seq)



Training = modeling

**Image source:** redbubble.com

Clinics

Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise
Models and signatures

# Random Forest

Omics

~20k genes (RNA-seq)

Image source: redbubble.com

Training = modeling

Clinics

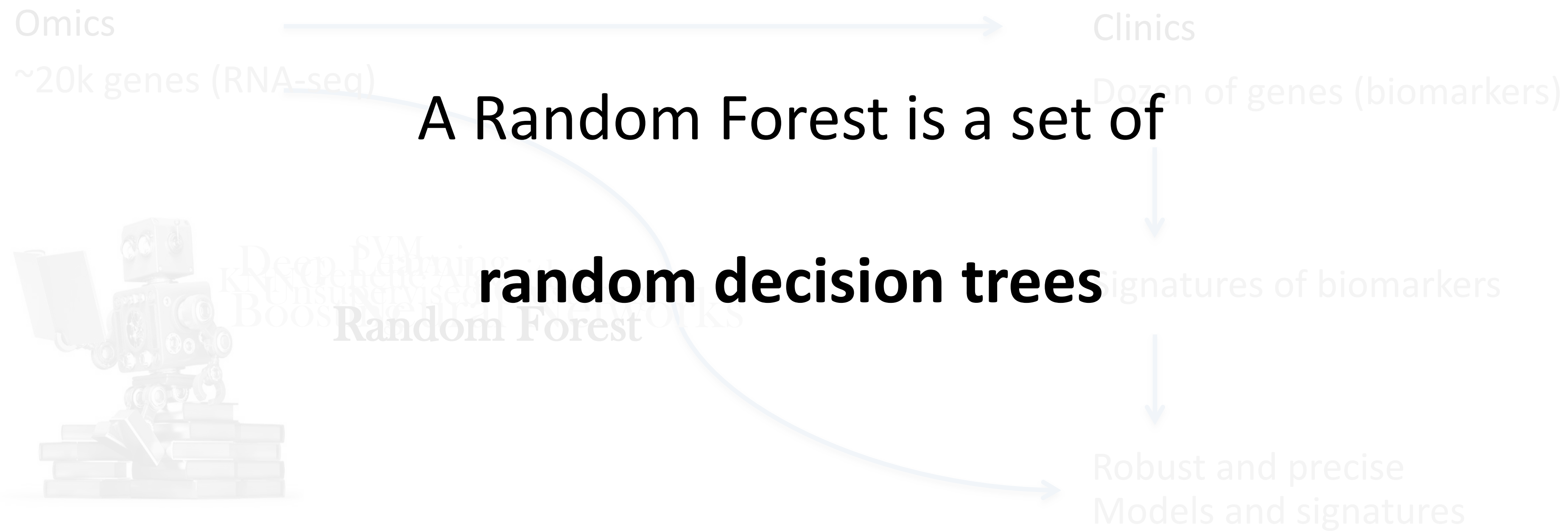
Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise Models and signatures

Prediction

# Random Forest



A Random Forest is a set of

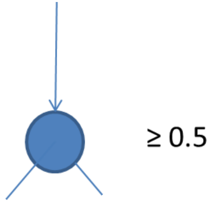
**random decision trees**

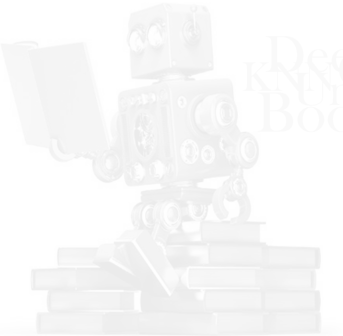
# A decision helps to stratify the data

Omics

~20k genes (RNA-seq)

es (biomarkers)



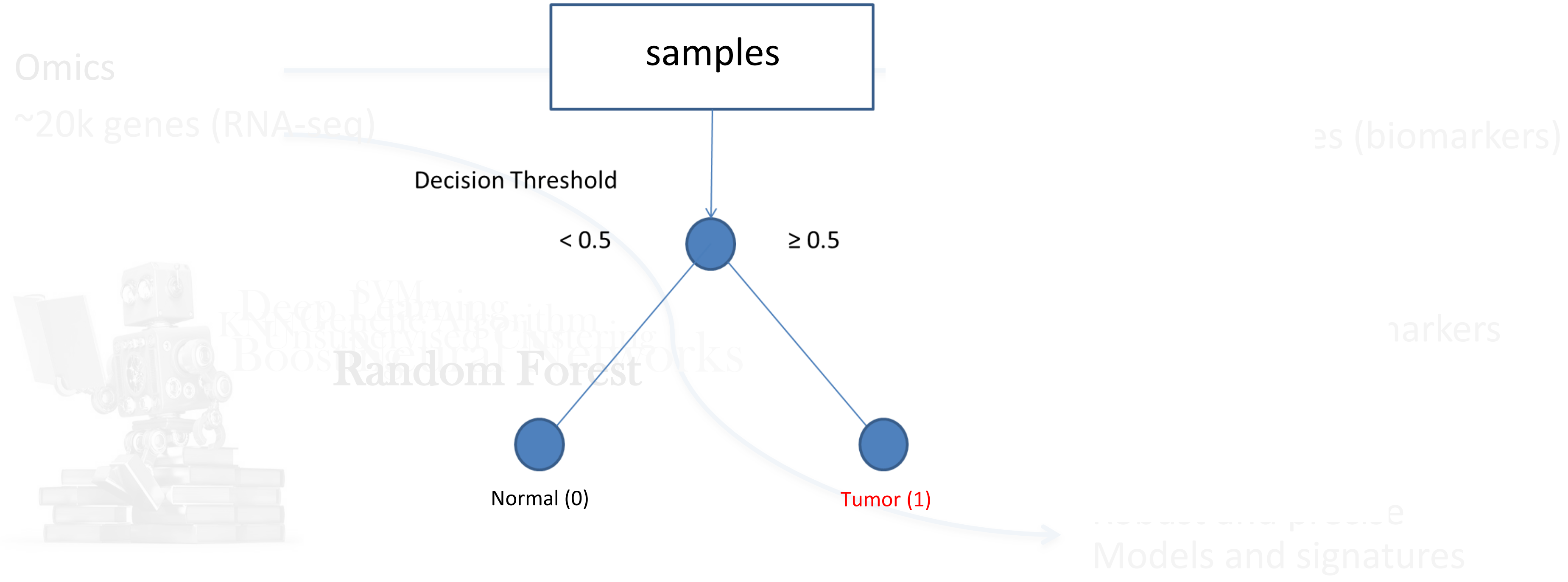


SVM
Deep Learning
KNN Genetic Algorithm
Unsupervised Cluste
Boosting Neural N
Random Fo

narkers

e
Models and signatures

# A decision helps to stratify the data



samples

Decision Threshold

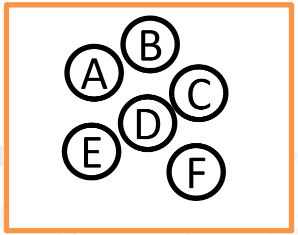
< 0.5

≥ 0.5

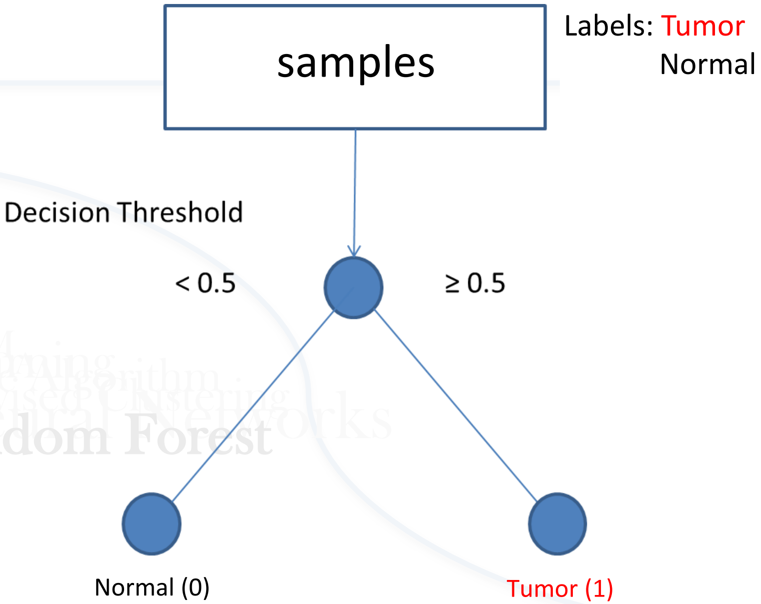
Normal (0)

Tumor (1)

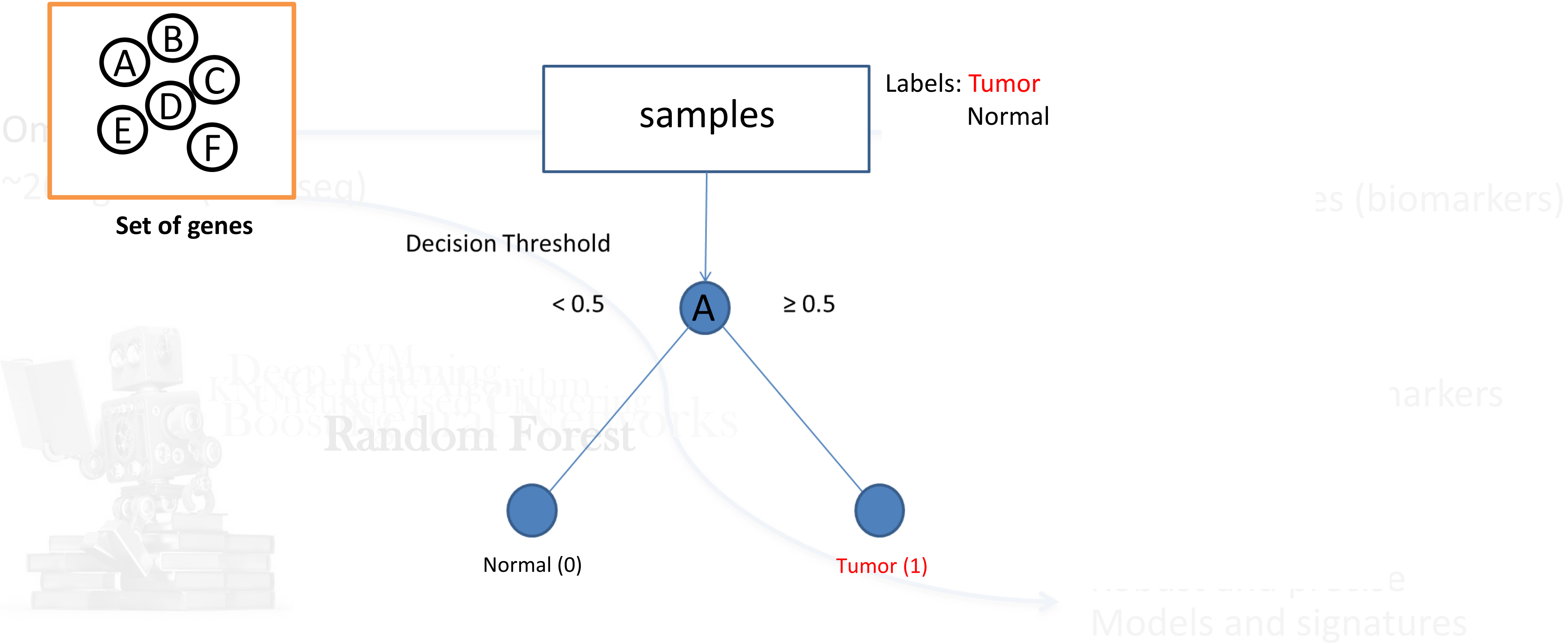
# A decision helps to stratify the data



**Set of genes**

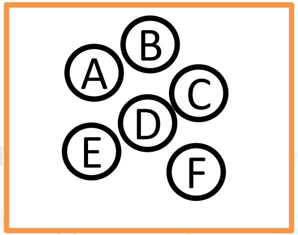


# A decision helps to stratify the data

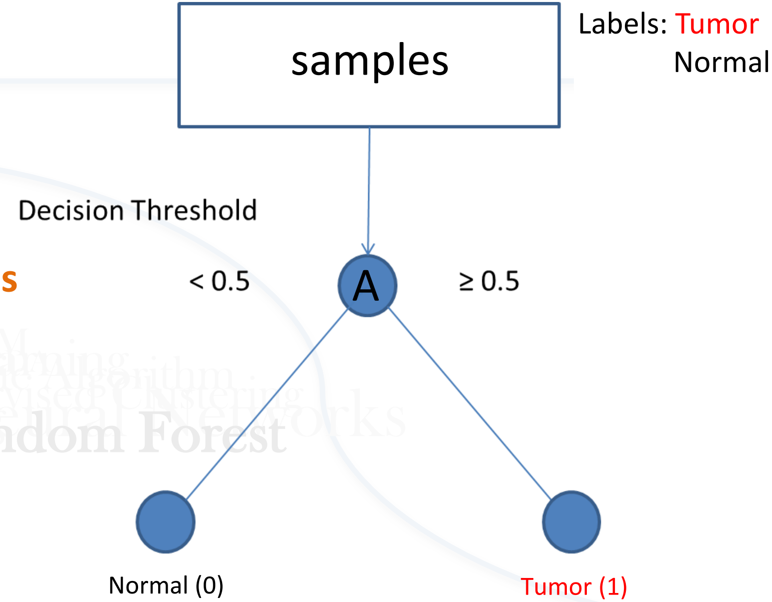


Set of genes

# A decision helps to stratify the data

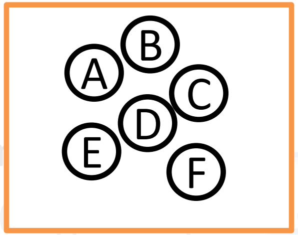


**Set of genes**



**Ability to discriminate groups of samples ➔ Importance**

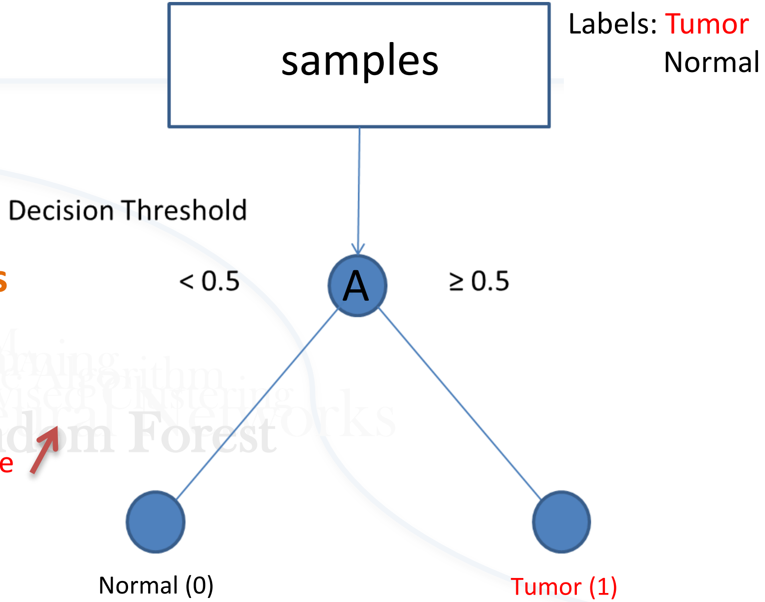
# A decision helps to stratify the data



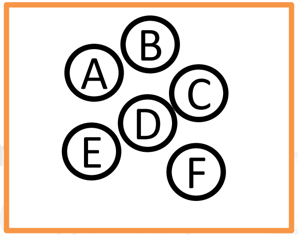
**Set of genes**

**Ability to discriminate groups of samples ➔ Importance**

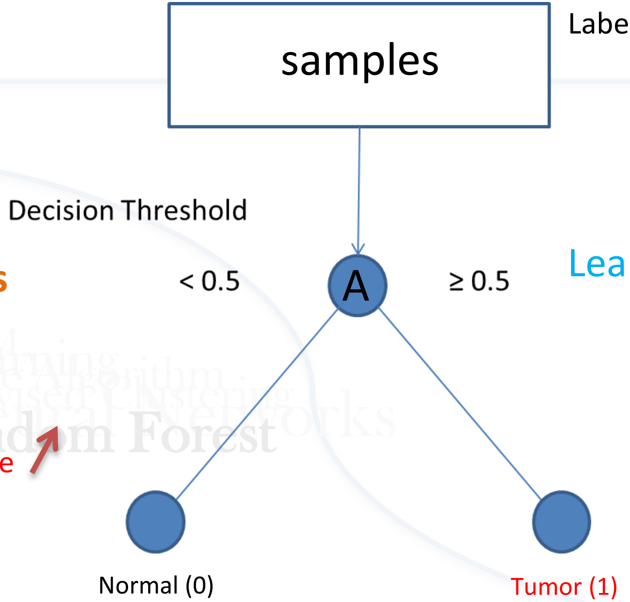
discriminative power ↗ ➔ importance ↗



# A decision helps to stratify the data



**Set of genes**



Labels: Tumor
Normal

Learning (training) process

Clinics

Dozen of genes (biomarkers)

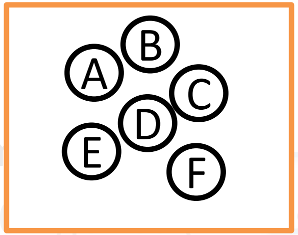
Signatures of biomarkers

Robust and precise
Models and signatures

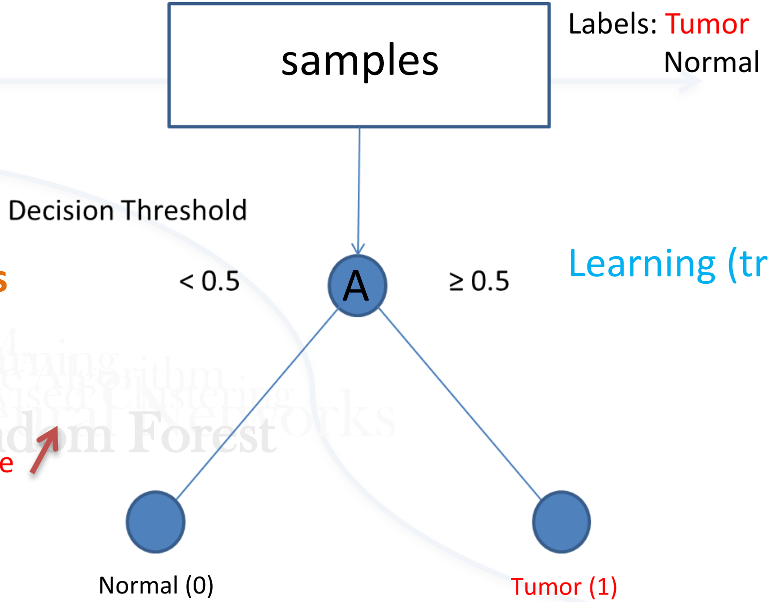
**Ability to discriminate groups of samples ➔ Importance**

discriminative power ↗ ➔ importance ↗

# A decision helps to stratify the data



Set of genes



Learning (training) process

Clinics

Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise
Models and signatures

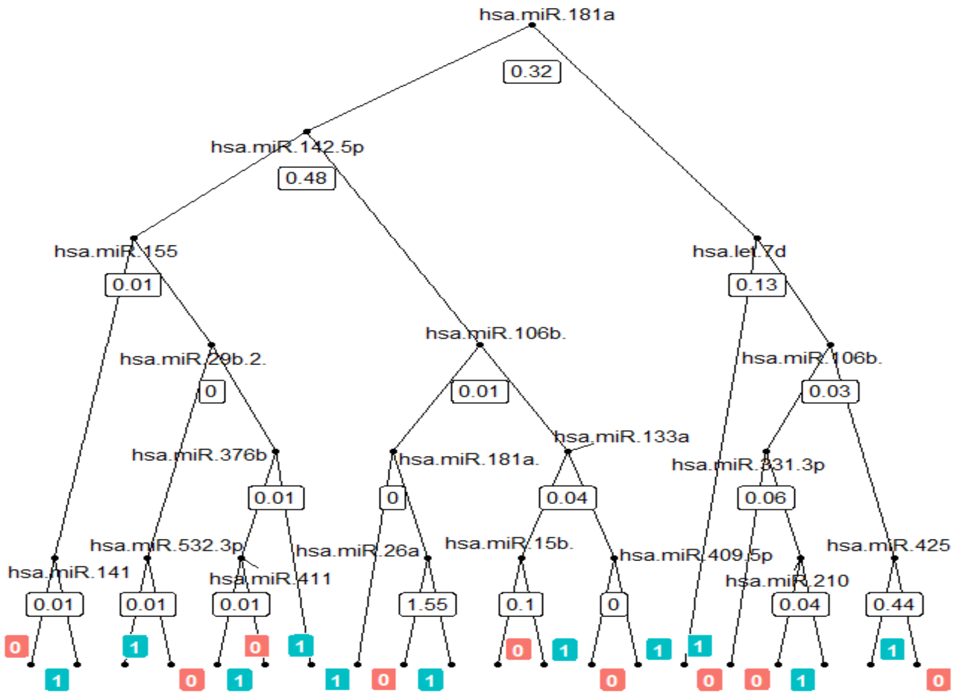
**Ability to discriminate groups of samples ➔ Importance**

discriminative power ↗ ➔ importance ↗

Random Forest

1 tree = Multiple decisions

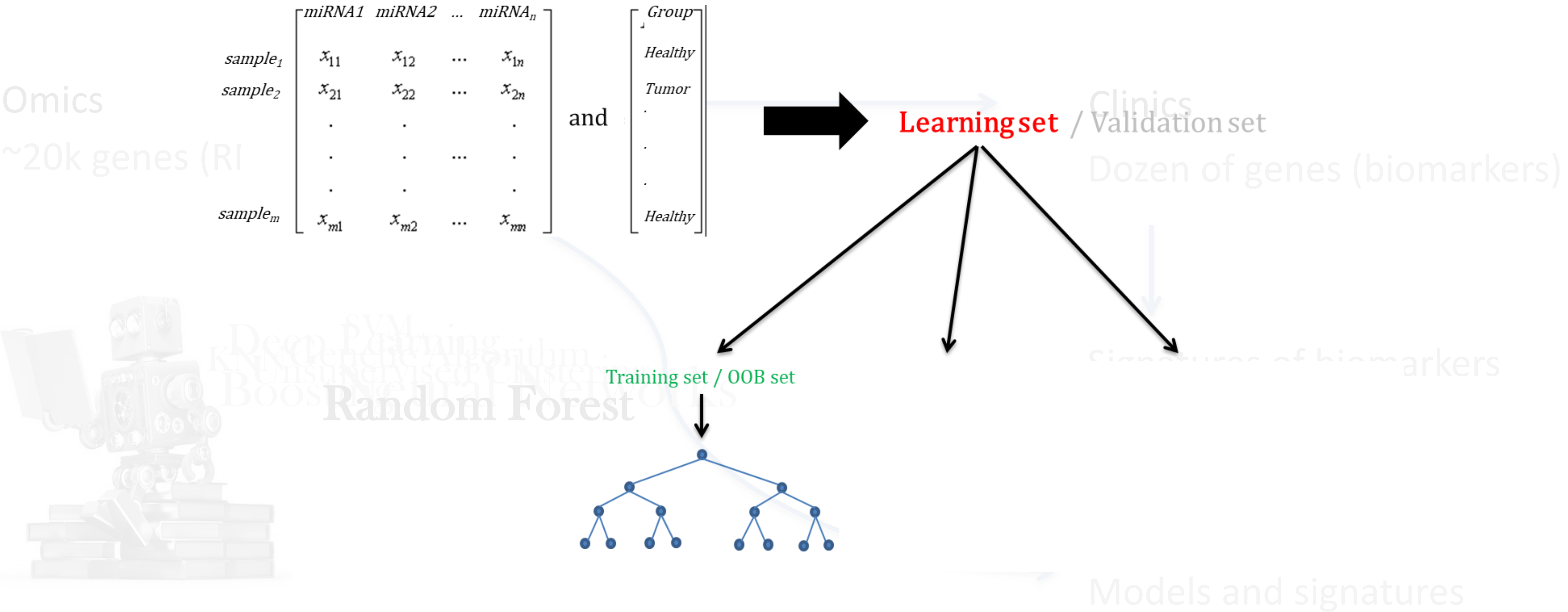
# Decision tree (real example)



Healthy: 0
Tumor: 1

1 tree = Multiple decisions

# Random Forest is a set of random trees

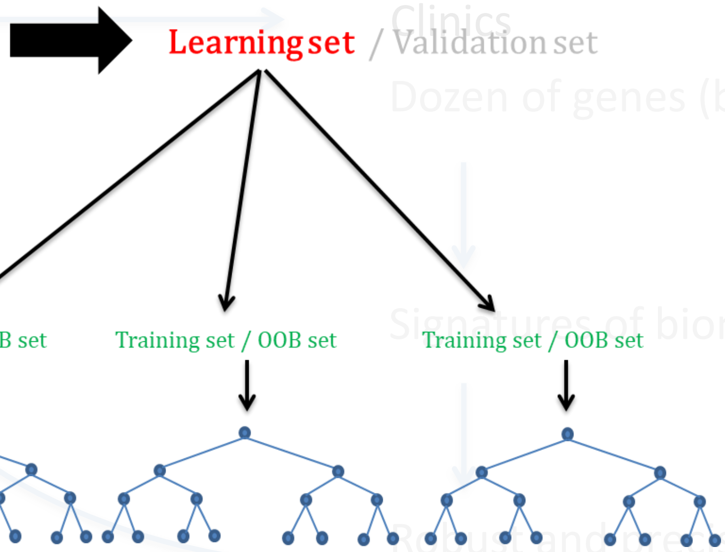


$$\begin{array}{c} \\ sample_1 \\ sample_2 \\ \\ \\ \\ sample_m \end{array} \begin{bmatrix} miRNA1 & miRNA2 & \dots & miRNA_n \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ . & . & & . \\ . & . & \cdots & . \\ . & . & & . \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \text{ and } \begin{bmatrix} Group \\ Healthy \\ Tumor \\ . \\ . \\ . \\ Healthy \end{bmatrix}$$

1 Forest = Multiple trees

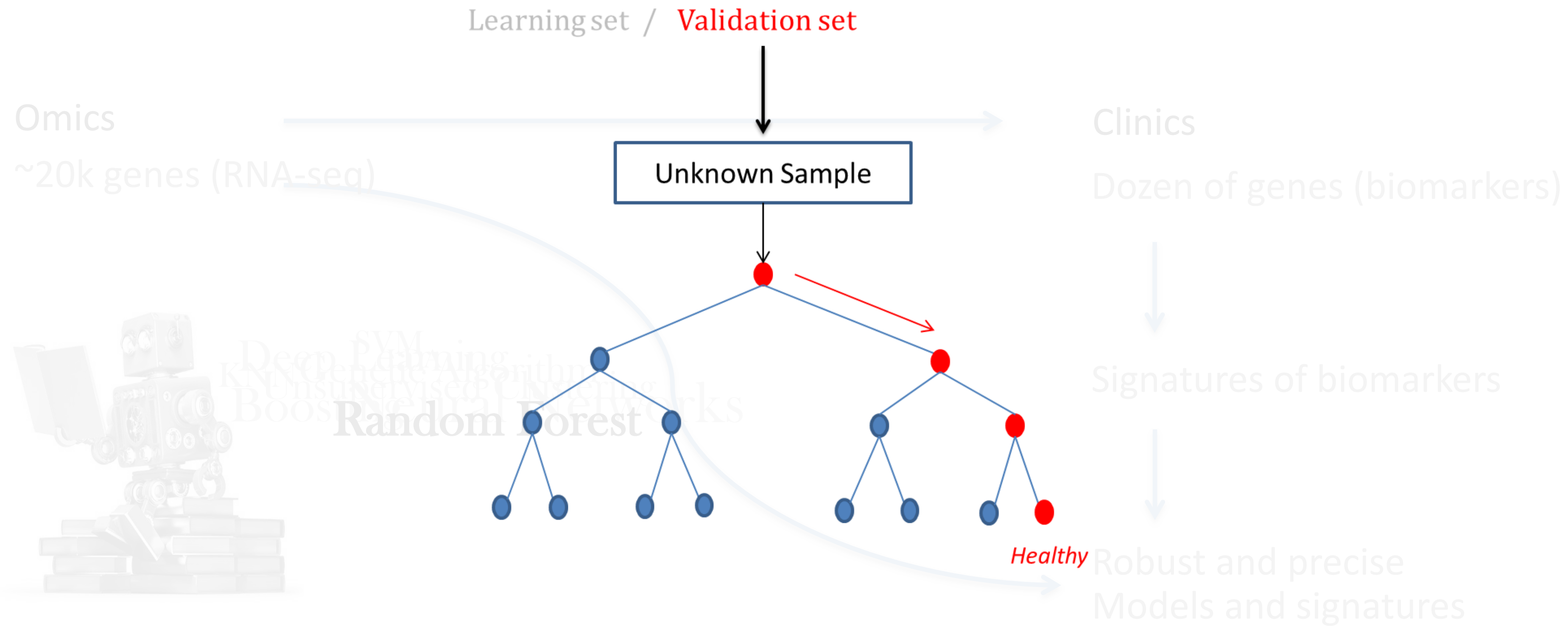
# Random Forest is a set of random trees

$$\begin{array}{c|cccc|} & miRNA1 & miRNA2 & \ldots & miRNA_n \\ sample_1 & x_{11} & x_{12} & \cdots & x_{1n} \\ sample_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ & . & . & & . \\ & . & . & \cdots & . \\ & . & . & & . \\ sample_m & x_{m1} & x_{m2} & \cdots & x_{mn} \end{array} \quad \text{and} \quad \begin{bmatrix} Group \\ Healthy \\ Tumor \\ . \\ . \\ . \\ Healthy \end{bmatrix}$$



1 Forest = Multiple trees

# Prediction procedure



**One Tree prediction procedure**

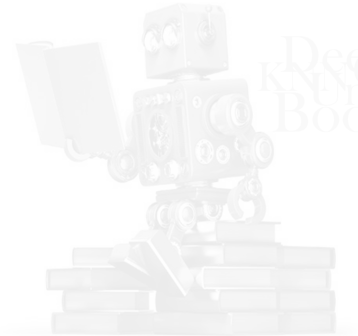
# Prediction procedure

Omics

~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)



SVM

Deep Learning

KNN

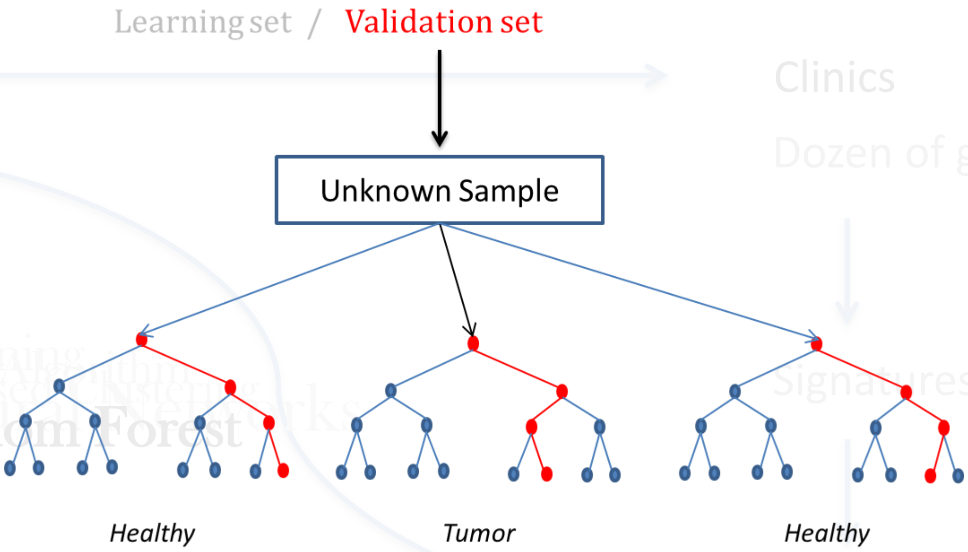
Genetic Algorithm

Unsupervised clustering

Boosting

Neural Networks

Random Forest



Signatures of biomarkers

Robust and precise
Models and signatures

**Multiple Trees prediction procedure**

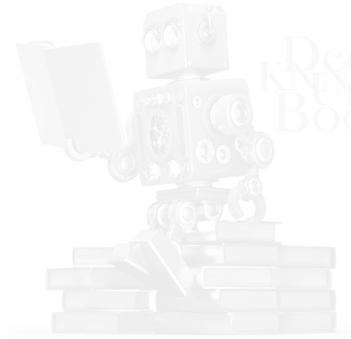
# Prediction procedure

Omics

~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)



Deep Learning

SVM

KNN

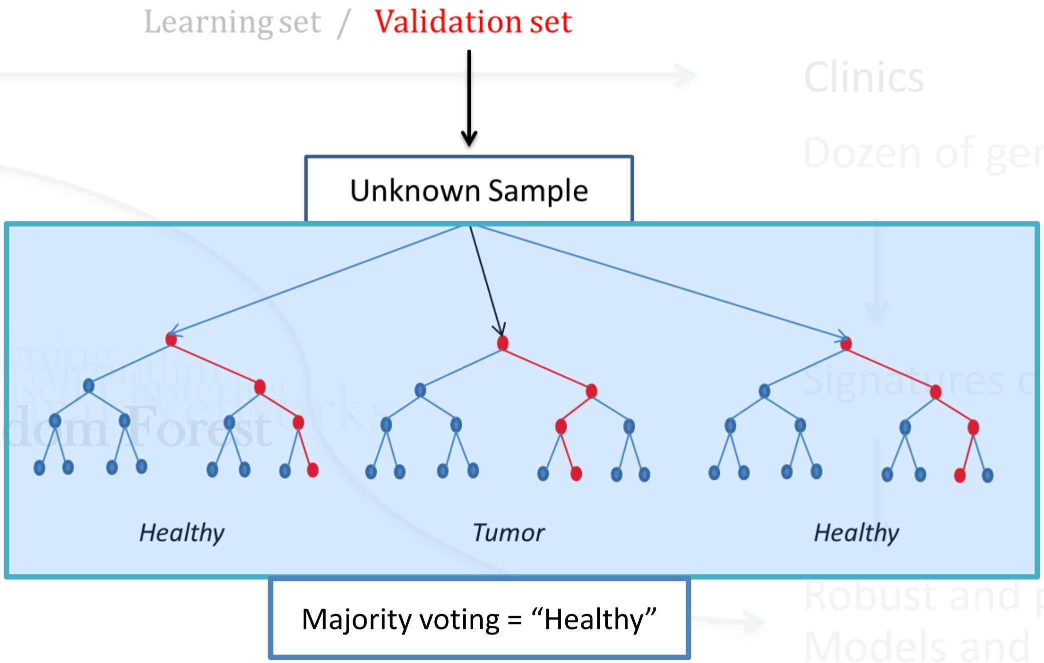
Genetic Algorithm

Unsupervised

Boosting

Neural Networks

Random Forest



Signatures of biomarkers

Robust and precise
Models and signatures

**1 Forest = 1 Model**

# From Omics to Clinics

Omics

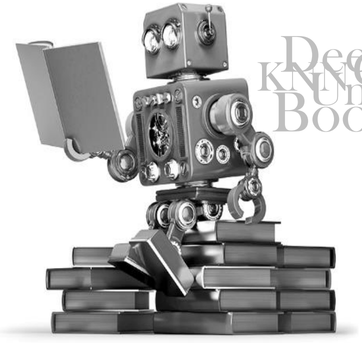
~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise
Models and signatures



SVM
Deep Learning
KNN Genetic Algorithm
Unsupervised Clustering
Boosting Neural Networks

**Random Forest**

**R** implementation**s** of
Random Forest algorithm
proposed by *[Breiman et al. 2001]*

# From Omics to Clinics

Omics

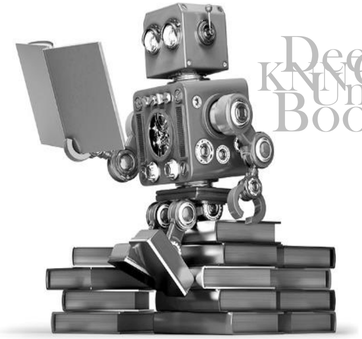
~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise
Models and signatures ???



SVM
Deep Learning
KNN
Genetic Algorithm
Unsupervised Clustering
Boosting
Neural Networks

**Random Forest**

**R** implementation**s** of
Random Forest algorithm
proposed by *[Breiman et al. 2001]*

# Objectives

**Toward a robust RF method for the Biological question asked**

Which method is suitable for which dataset (platform/technology) ?

# Objectives

- Empirical comparison of random forest based methods
- Differences/Similarities of RF methods ➔ groups of methods
- Designing a high stability score to rank RF methods

**Toward a robust RF method for the Biological question asked**

Which method is suitable for which dataset (platform/technology) ?

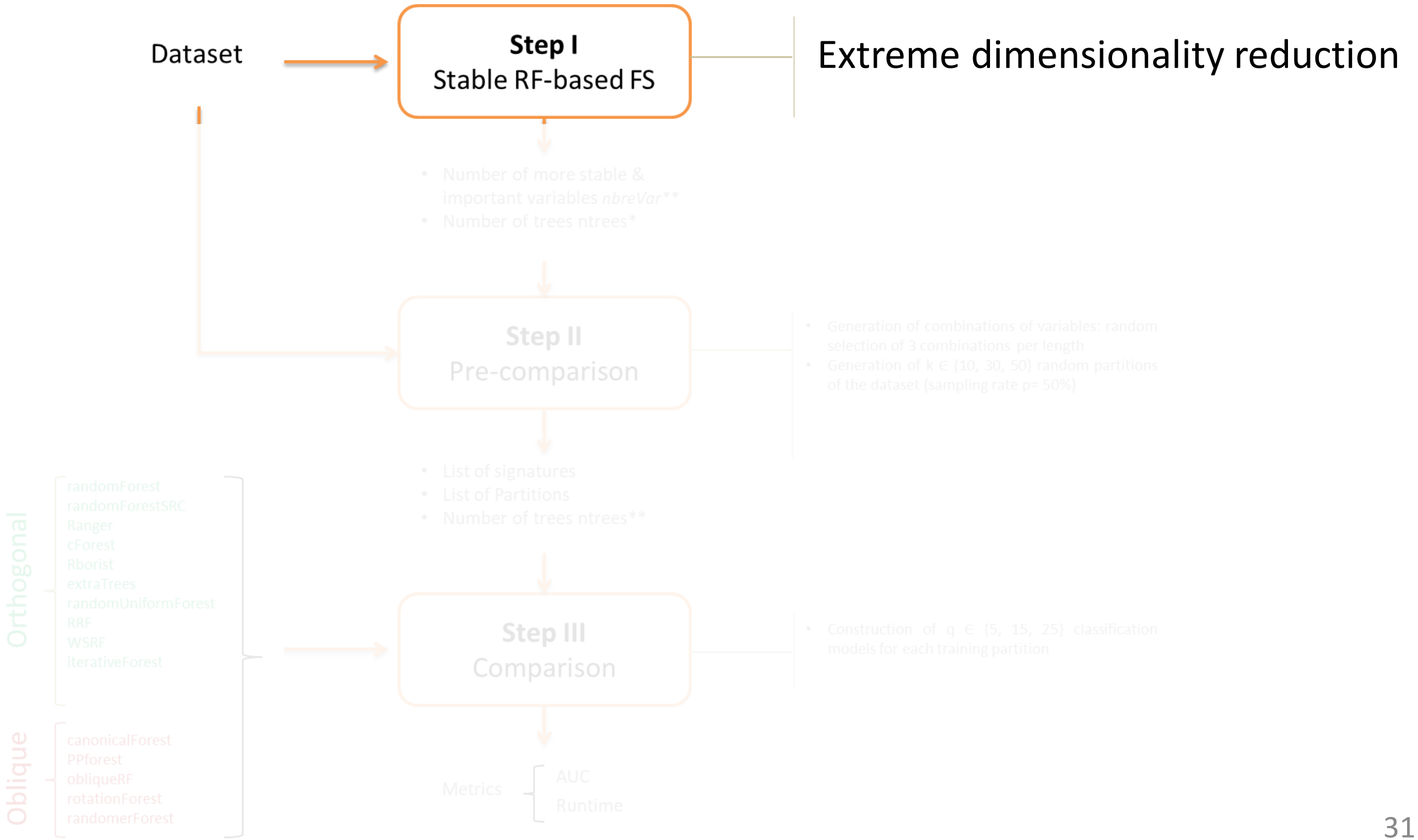
# Materials and Methods

- **Datasets (Perfectly balanced)**
  - TCGA-BRCA (RPKM): 182 samples x 9560 genes
  - TCGA-LUSC (RPKM): 96 samples x 9262 genes
  - TCGA-THCA (RPKM): 98 samples x 9353 genes

- **Main classification question**

The difference between paired **Tumor / Normal** samples will be used as a **strong classification** parameter, allowing for **strong** modeling only

# Overview of the method



Dataset

**Step I**
Stable RF-based FS

Extreme dimensionality reduction

- Number of more stable & important variables *nbreVar***
- Number of trees ntrees*

**Step II**
Pre-comparison

- Generation of combinations of variables: random selection of 3 combinations per length
- Generation of k ∈ {10, 30, 50} random partitions of the dataset (sampling rate p= 50%)

- List of signatures
- List of Partitions
- Number of trees ntrees**

Orthogonal

randomForest
randomForestSRC
Ranger
cForest
Rborist
extraTrees
randomUniformForest
RRF
WSRF
iterativeForest

Oblique

canonicalForest
PPforest
obliqueRF
rotationForest
randomerForest

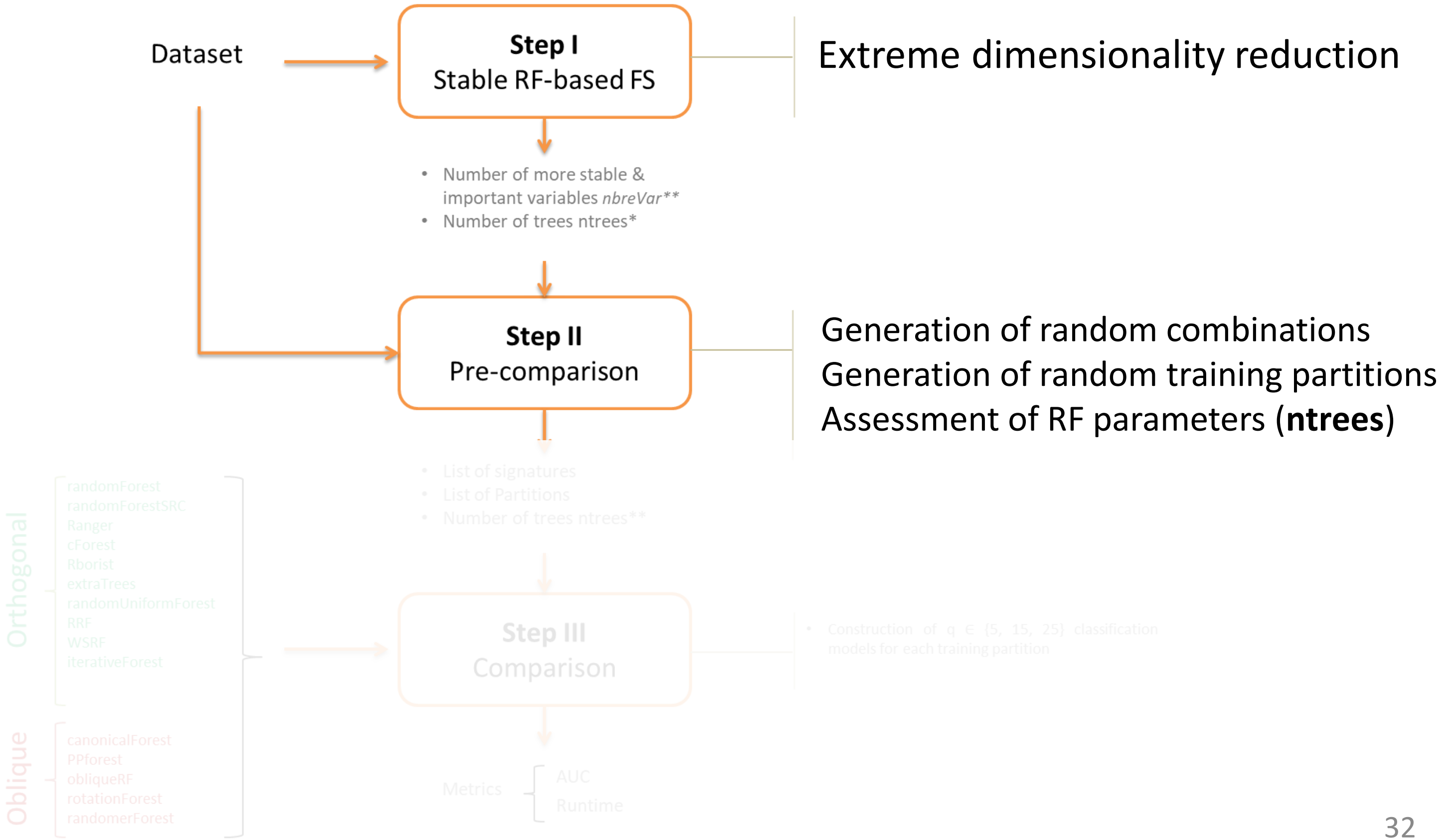
**Step III**
Comparison

- Construction of q ∈ {5, 15, 25} classification models for each training partition

Metrics

AUC
Runtime

# Overview of the method



Dataset

**Step I**
Stable RF-based FS

- Number of more stable & important variables *nbreVar***
- Number of trees ntrees*

Extreme dimensionality reduction

**Step II**
Pre-comparison

Generation of random combinations
Generation of random training partitions
Assessment of RF parameters (**ntrees**)

- List of signatures
- List of Partitions
- Number of trees ntrees**

Orthogonal

randomForest
randomForestSRC
Ranger
cForest
Rborist
extraTrees
randomUniformForest
RRF
WSRF
iterativeForest

Oblique

canonicalForest
PPforest
obliqueRF
rotationForest
randomerForest

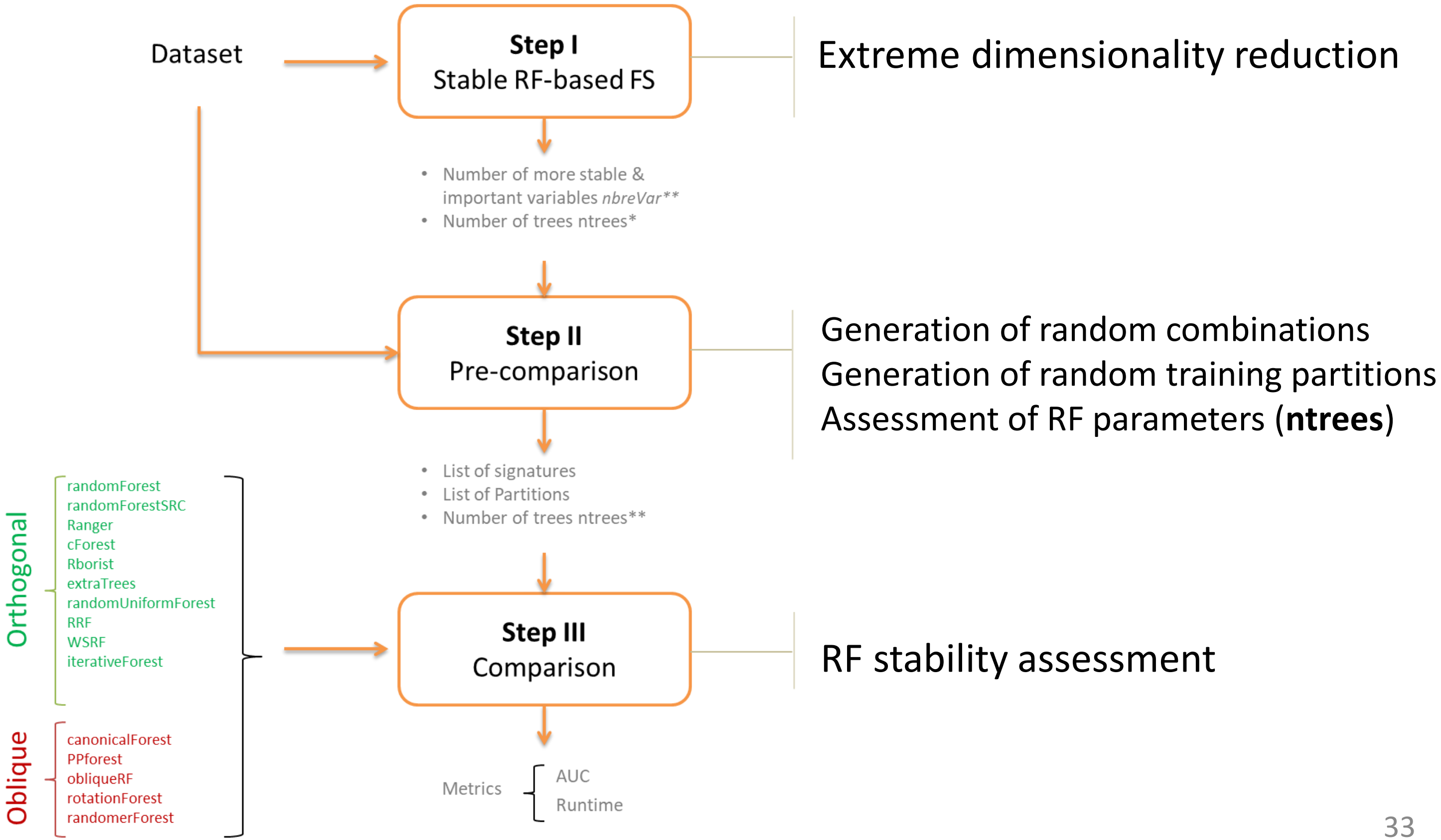
**Step III**
Comparison

- Construction of $q \in \{5, 15, 25\}$ classification models for each training partition

Metrics

AUC
Runtime

# Overview of the method



Dataset

**Step I**
Stable RF-based FS

Extreme dimensionality reduction

- Number of more stable & important variables *nbreVar***
- Number of trees ntrees*

**Step II**
Pre-comparison

Generation of random combinations
Generation of random training partitions
Assessment of RF parameters (**ntrees**)

- List of signatures
- List of Partitions
- Number of trees ntrees**

Orthogonal
randomForest
randomForestSRC
Ranger
cForest
Rborist
extraTrees
randomUniformForest
RRF
WSRF
iterativeForest

Oblique
canonicalForest
PPforest
obliqueRF
rotationForest
randomerForest

**Step III**
Comparison

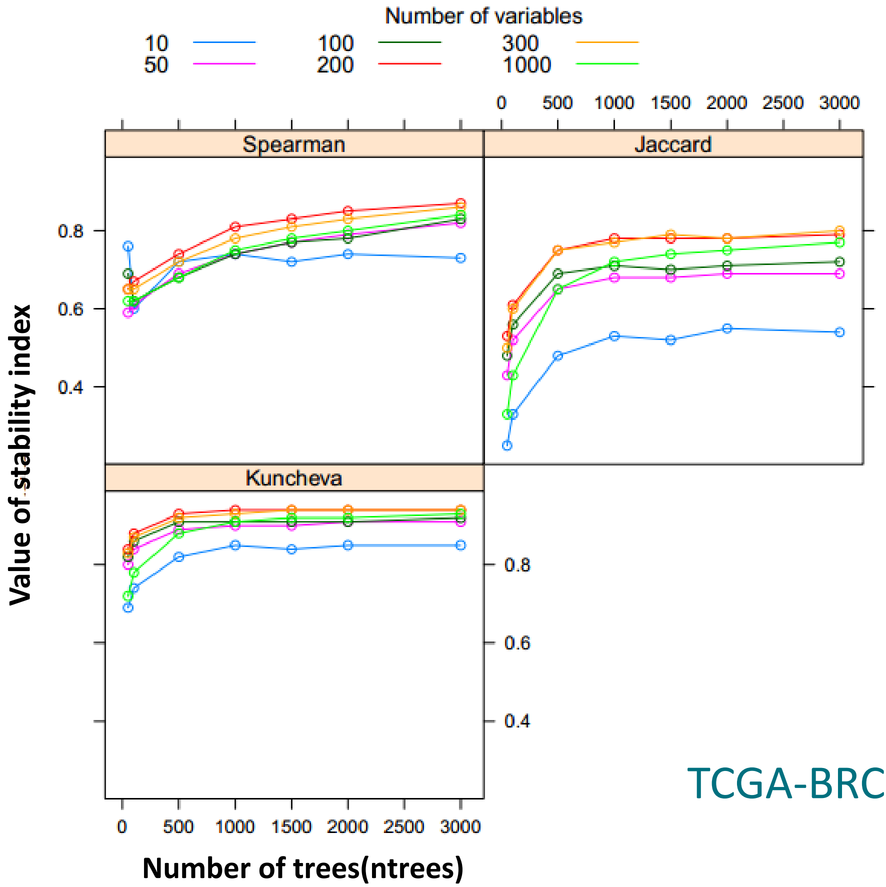
RF stability assessment

Metrics
AUC
Runtime

# Stable Feature Selection

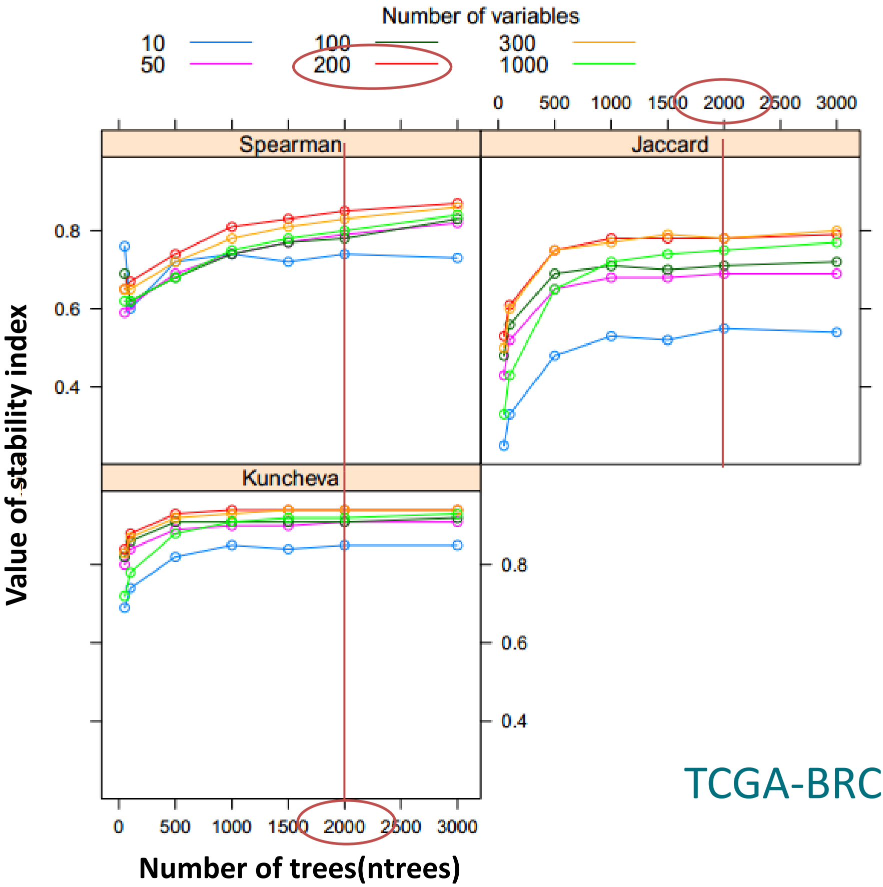
Extreme Dimensionality reduction

# First pass Feature Selection  results



TCGA-BRCA dataset

# First pass Feature Selection  results

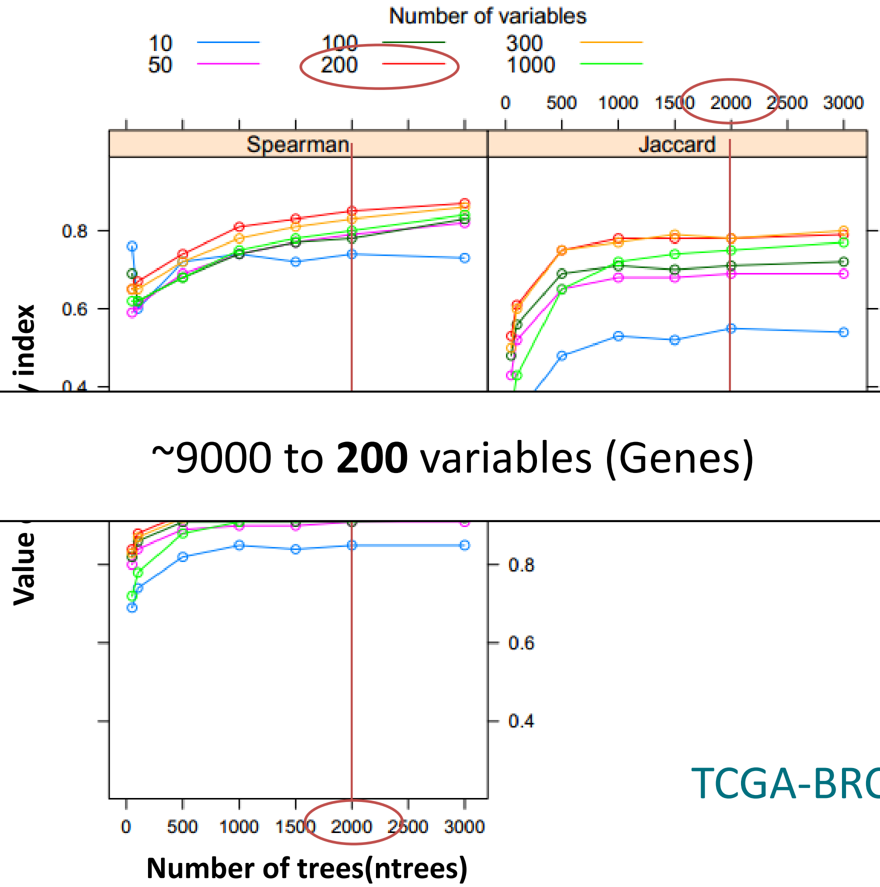


$ntrees^*$ = 2000

$nVar^*$ = 200

TCGA-BRCA dataset

# First pass Feature Selection  results



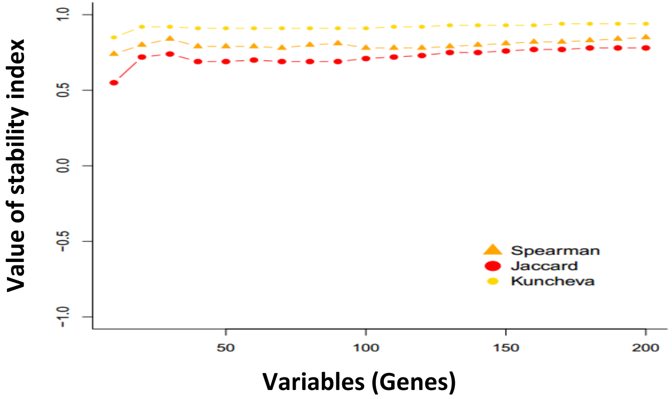
~9000 to **200** variables (Genes)

$ntrees^*$ = 2000

$nVar^*$ = 200

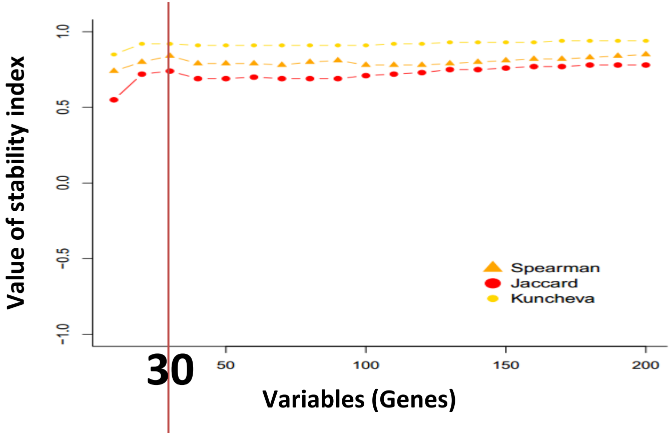
TCGA-BRCA dataset

# Second pass Feature Selection results



TCGA-BRCA dataset

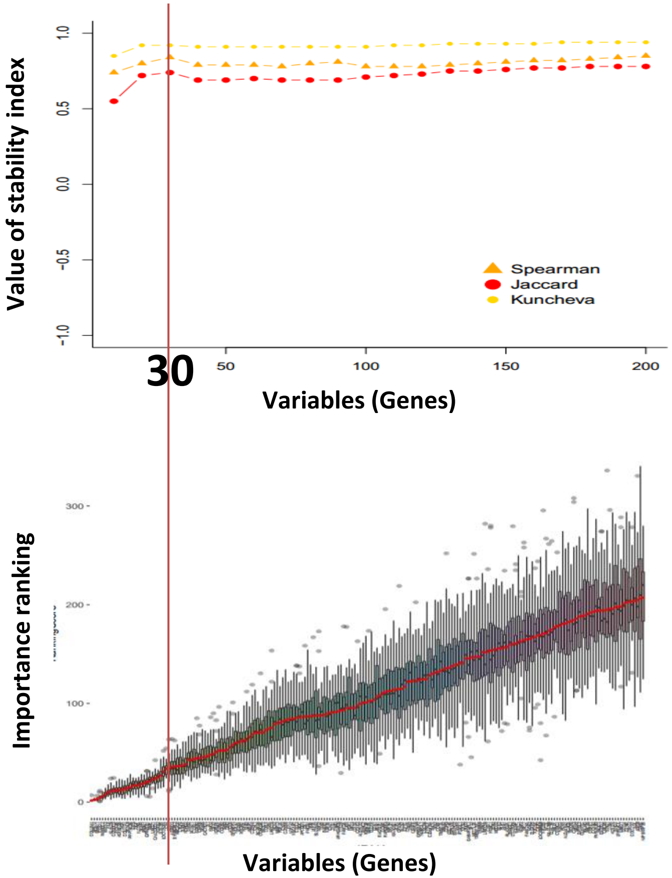
# Second pass Feature Selection results



$$nVar^{**} = 30$$

TCGA-BRCA dataset

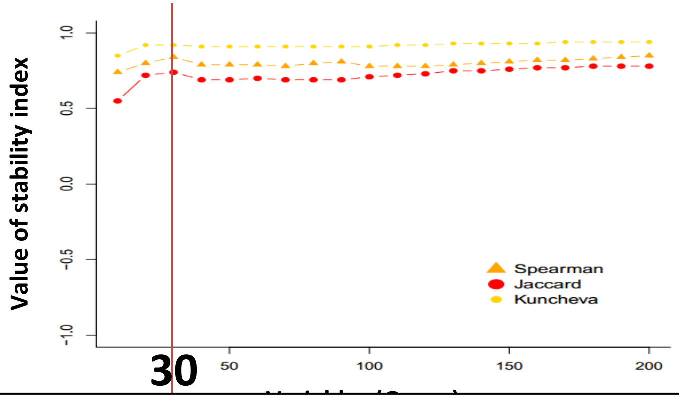
# Second pass Feature Selection results



$$nVar^{**} = 30$$

TCGA-BRCA dataset

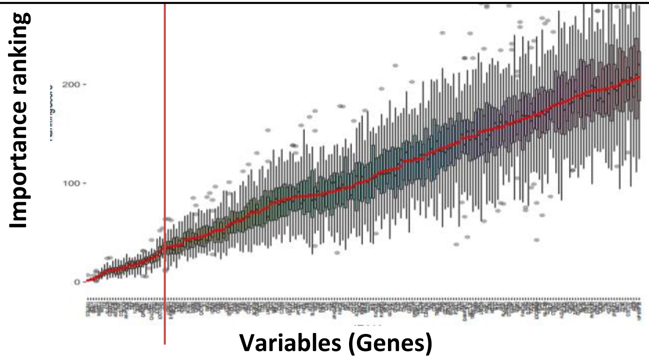
# Second pass Feature Selection results



30

$$nVar^{**} = 30$$

~200 to **30** variables (Genes)



TCGA-BRCA dataset

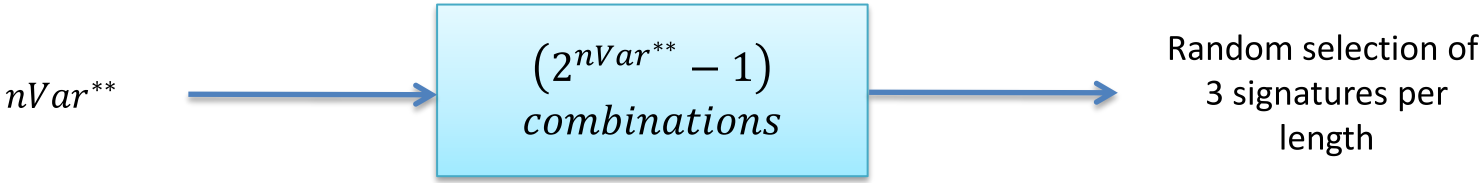
# Pre-Comparison

Assessment of RF parameters & Generation of random combinations

# Pre-comparison

## I- Generation of random combinations (Cancer signatures)

- Multiple predictive models using combinations of different lengths



$nVar^{**}$

$(2^{nVar^{**}} - 1)$ $combinations$

Random selection of 3 signatures per length

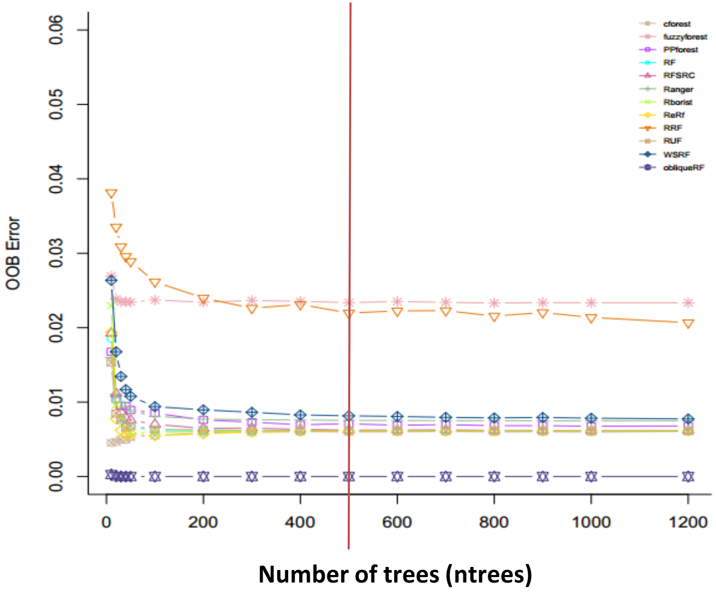
## II- Generation of random training partitions

- 50 random training partitions

training partition = a set of samples used to construct a model

# Pre-comparison

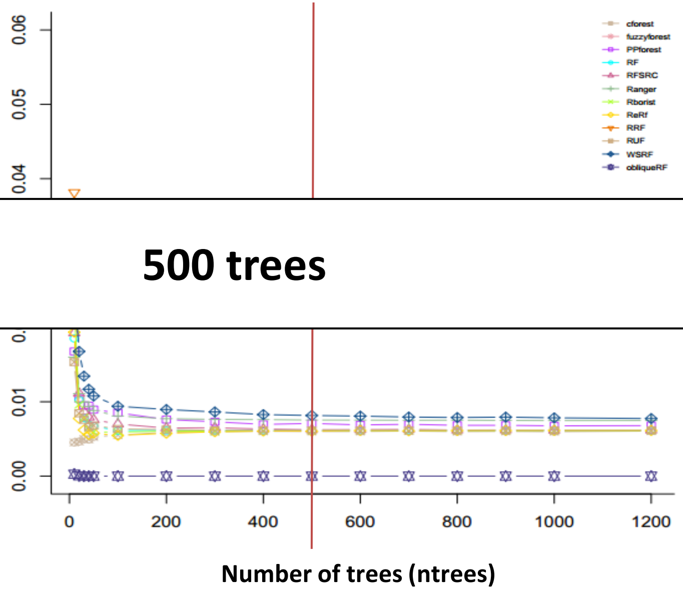
## III- Tuning the parameter $ntrees$ for each RF method



TCGA-BRCA dataset

# Pre-comparison

## III- Tuning the parameter $ntrees$ for each RF method



**500 trees**

TCGA-BRCA dataset

# Summary of step I + step II

| Dataset | $nVar$ | $nVar^{*}$ | $nVar^{**}$ | $ntrees^{*}$ | $ntrees^{**}$ | Nbre combinations |
|---|---|---|---|---|---|---|
| **TCGA-BRCA** | 9560 | 200 | 30 | 2000 | 500 | 78 |
| **TCGA-LUSC** | 9262 | 200 | 10 | 2000 | 500 | 21 |
| **TCGA-THCA** | 9353 | 200 | 40 | 2000 | 500 | 108 |

# Comparison

Random Forest stability assessment

# Random Forest Method Comparison

- Comparison of RF methods under perfect conditions
- Using **same** random training partitions
- Assessing the **same** signatures
- On computational cores of **same** characteristics

# Random Forest Method Comparison

- For each signature, we'll focus on:
    - **50 resampling** to build the Training and the Validation set.
    - **25 modeling and validations.**

- Analysis of:
    - **Coefficient of Variation of 1,250 models & AUCs**

**1,250 models & AUCs**

# Random Forest Method Comparison

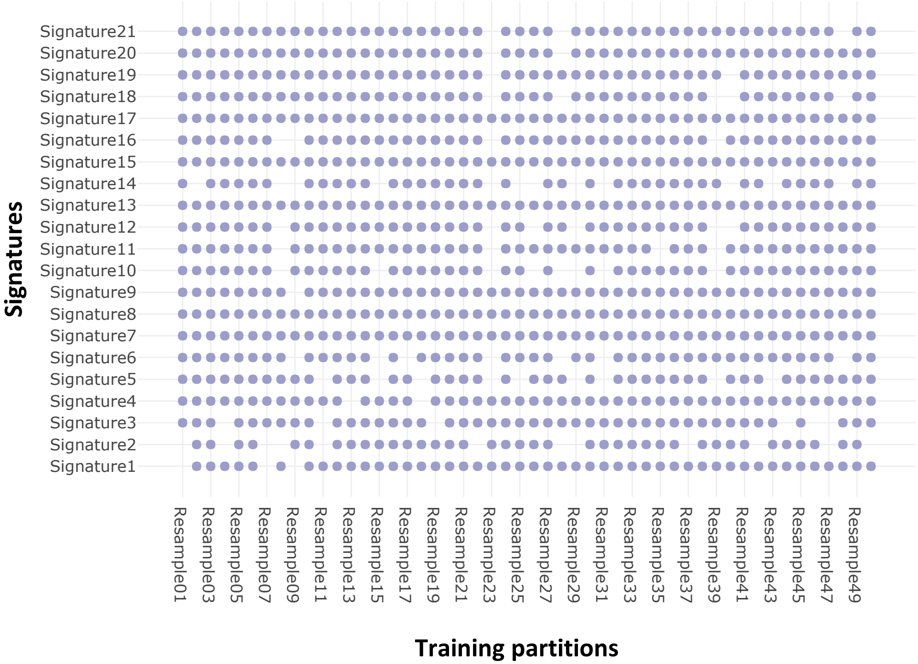
- For each signature, we'll focus on:
  - **50 resampling** to build the Training and the Validation set.
  - **25 modeling and validations.**

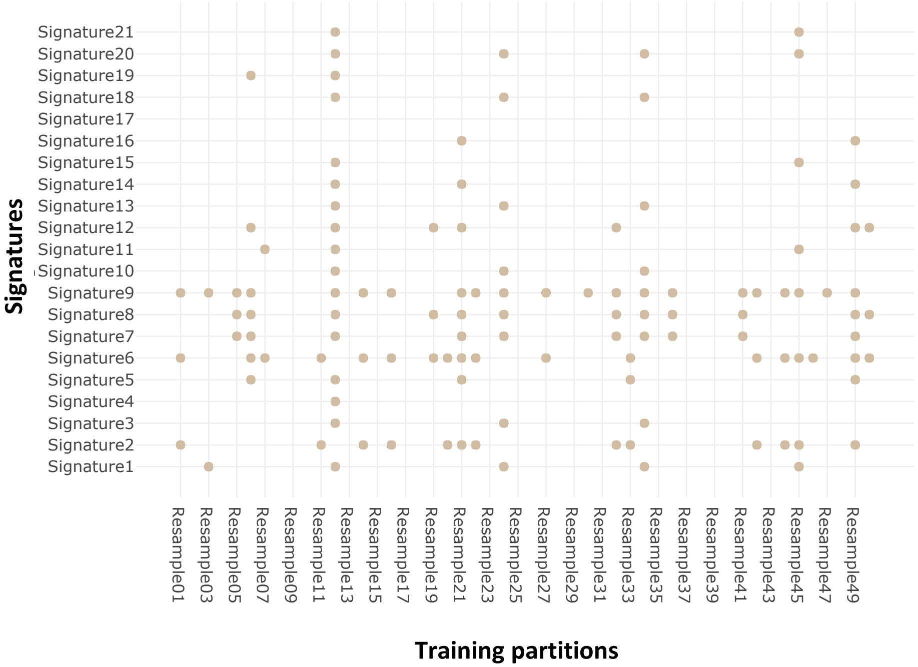
- Analysis of:
  - **Coefficient of Variation of 1,250 models & AUCs**

**1,250 models & AUCs**

**Clinics → Hyper Stability : CV == 0**

# Hyper Stability discriminates RF methods

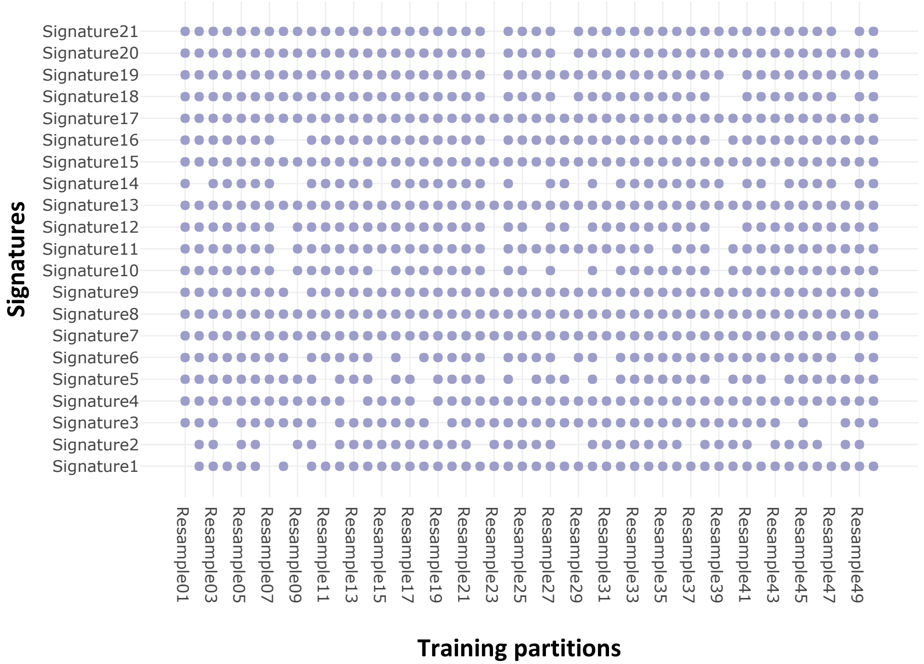




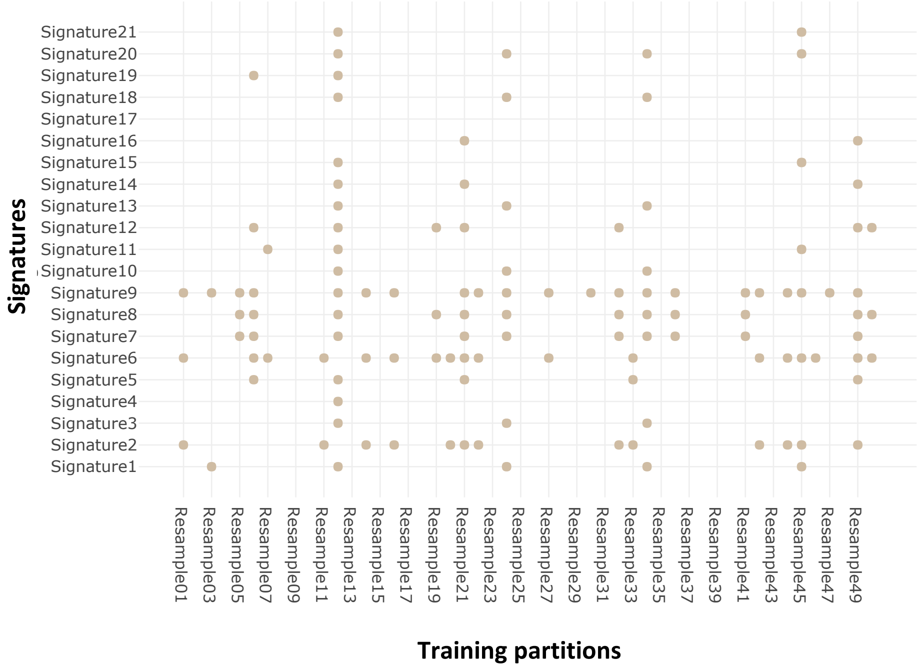
TCGA-LUSC dataset

# Hyper Stability discriminates RF methods

**Best Methods**

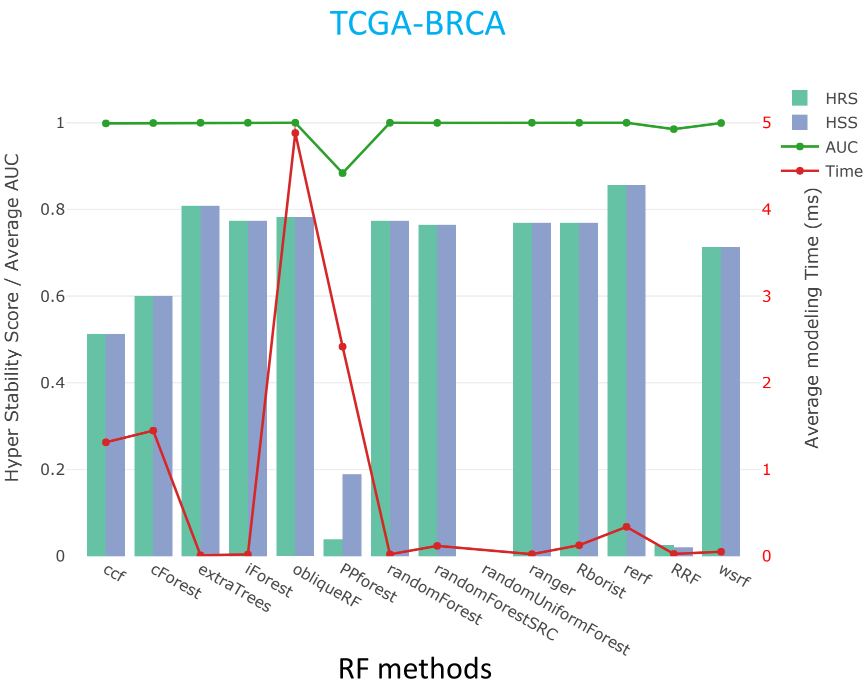


**Worst Methods**

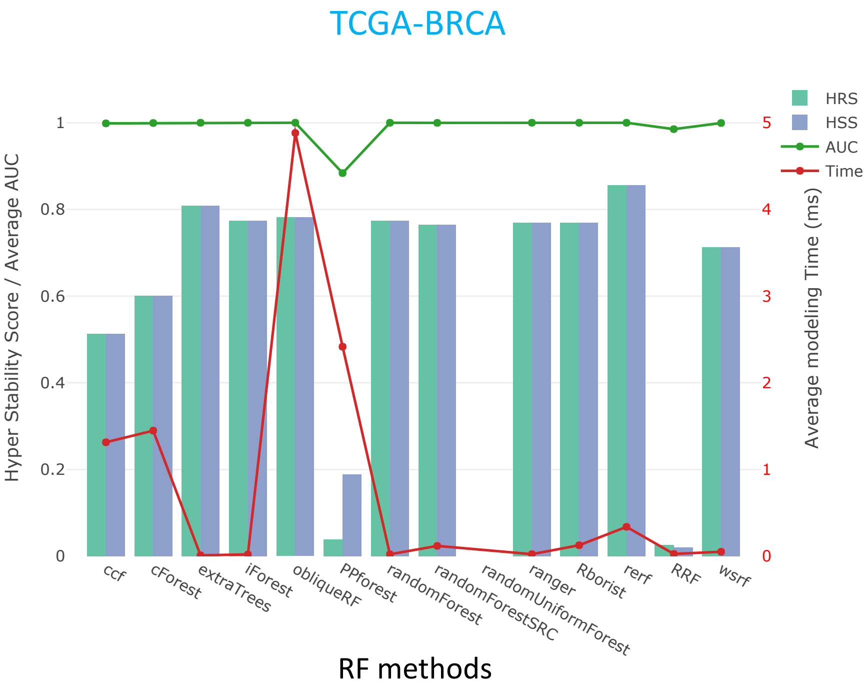


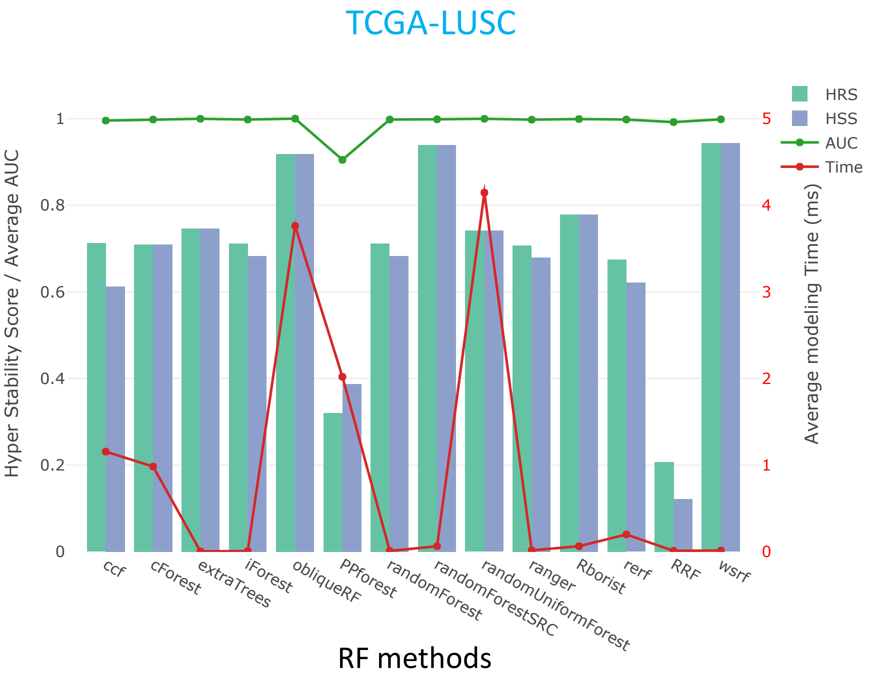
TCGA-LUSC dataset

# Hyper Stability Score helps finding the best method(s)



# Hyper Stability Score is dataset dependent





# Conclusions

- The AUC precision is dataset dependent
  - The Methods are dataset dependent.

- Trade-off:
  - AUC precision (hyper-stability)
  - Average AUC value
  - Modeling Time

Classification of classification methods

Towards robust signatures and predictions

# Acknowledgment

### Human Genetics (GIGA)

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

### BIO3 Unit (GIGA)

- Prof. Kristel Van Steen

### Oncology (CHU-Liege)

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

### CBIO, Mines ParisTech, Institut Curie

- Chloe-Agathe Azencott, PhD

LIÈGE université
GIGA

RÉGION WALLONNE

fnrs
LA LIBERTÉ DE CHERCHER

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

LIÈGE université
GIGA

Thank you for your attention

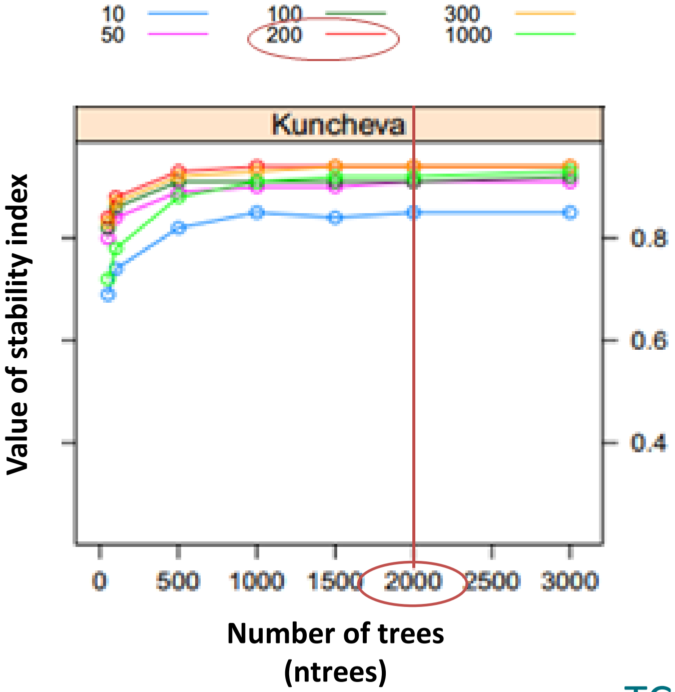
LIÈGE université

GIGA

LIÈGE université

GIGA

# First pass Feature Selection  results



$ntrees^*$ = 2000

$nVar^*$ = 200

TCGA-BRCA dataset