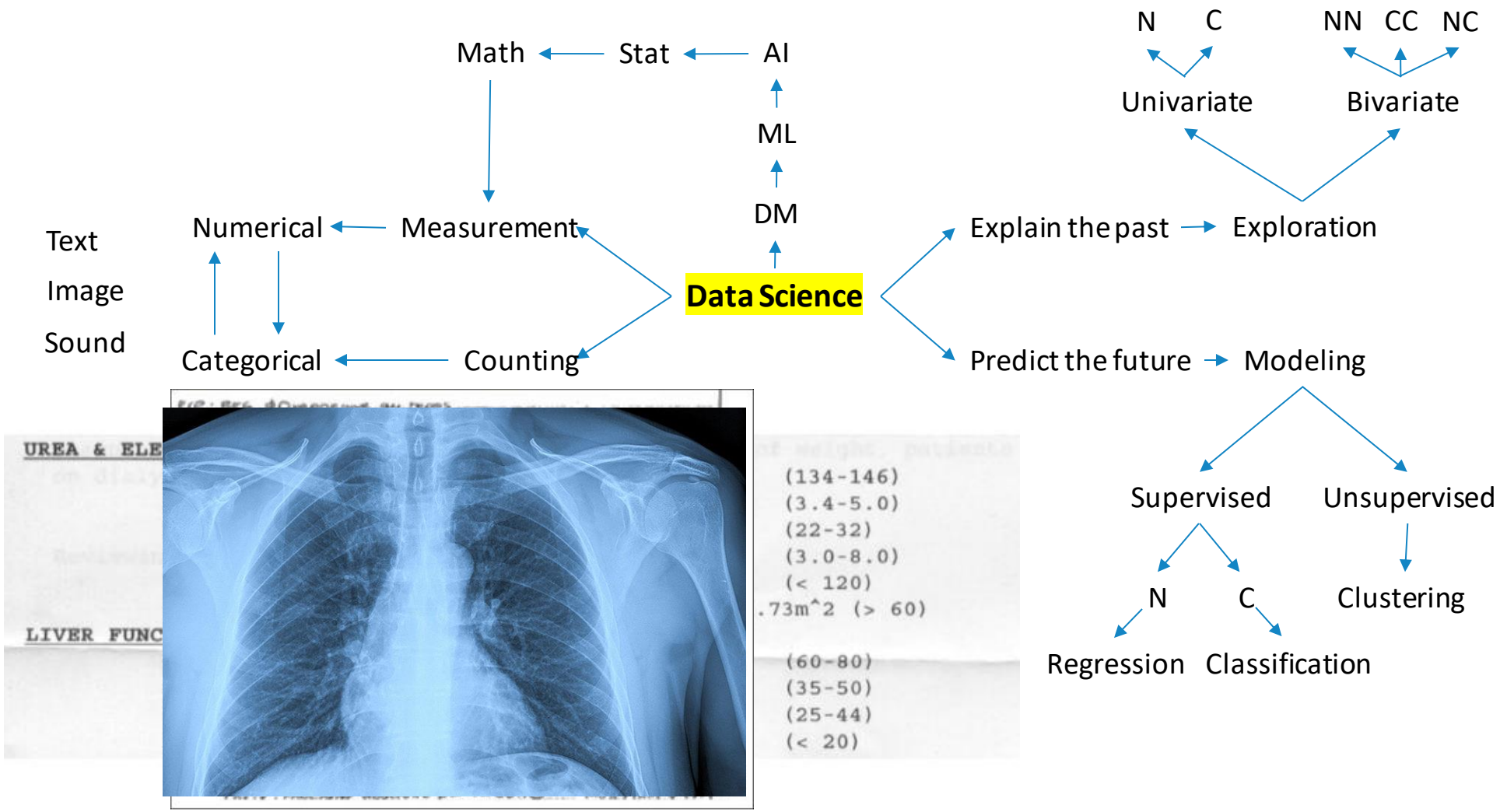
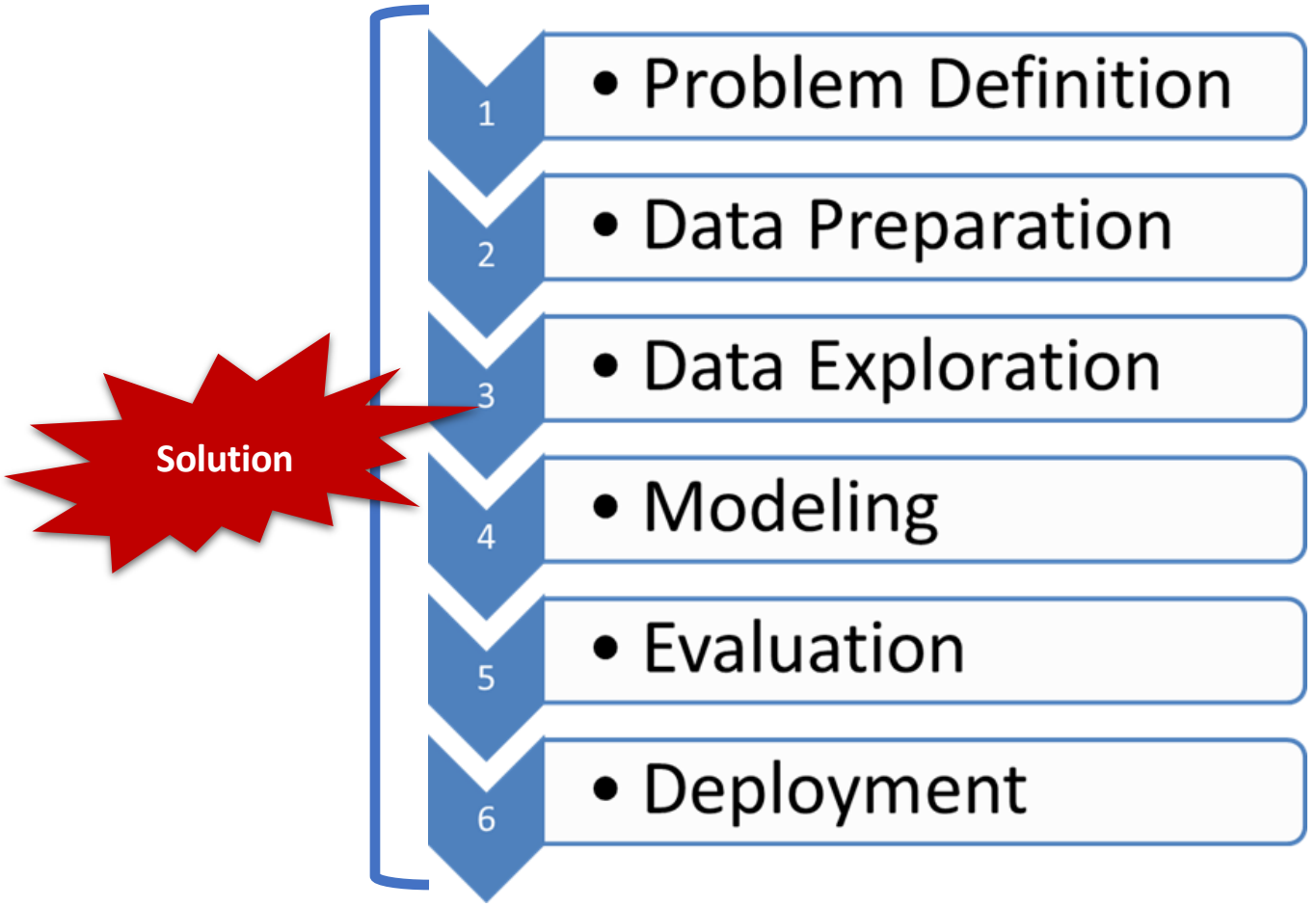
Data Science for Biomarkers Discovery

Saed Sayad M.D. Ph.D.
Bioada.com


Math
Stat
AI
ML
DM
Data Science
Text
Image
Sound
Numerical
Measurement
Categorical
Counting
N
C
Univariate
NN
CC
NC
Bivariate
Explain the past
Exploration
Predict the future
Modeling
Supervised
Unsupervised
N
C
Clustering
Regression
Classification
UREA & ELE
LIVER FUNC
(134-146)
(3.4-5.0)
(22-32)
(3.0-8.0)
(< 120)
.73m^2 (> 60)
(60-80)
(35-50)
(25-44)
(< 20)

Data Science 6-Step



Domain Expert

DBA and a dedicated DB server

Univariate and Bivariate analysis

Modeling using the right toolbox

Predefined evaluation metrics

Three levels of model testing

1- Problem Definition

- Building a **classification model** using a novel combination of serum microRNAs for detecting breast cancer in the early stage

- Success Criteria
  - ROC chart/AUC
  - Precision/Recall
  - FDR
  - Gain/Lift chart
  - K-S chart
  - Deciles chart

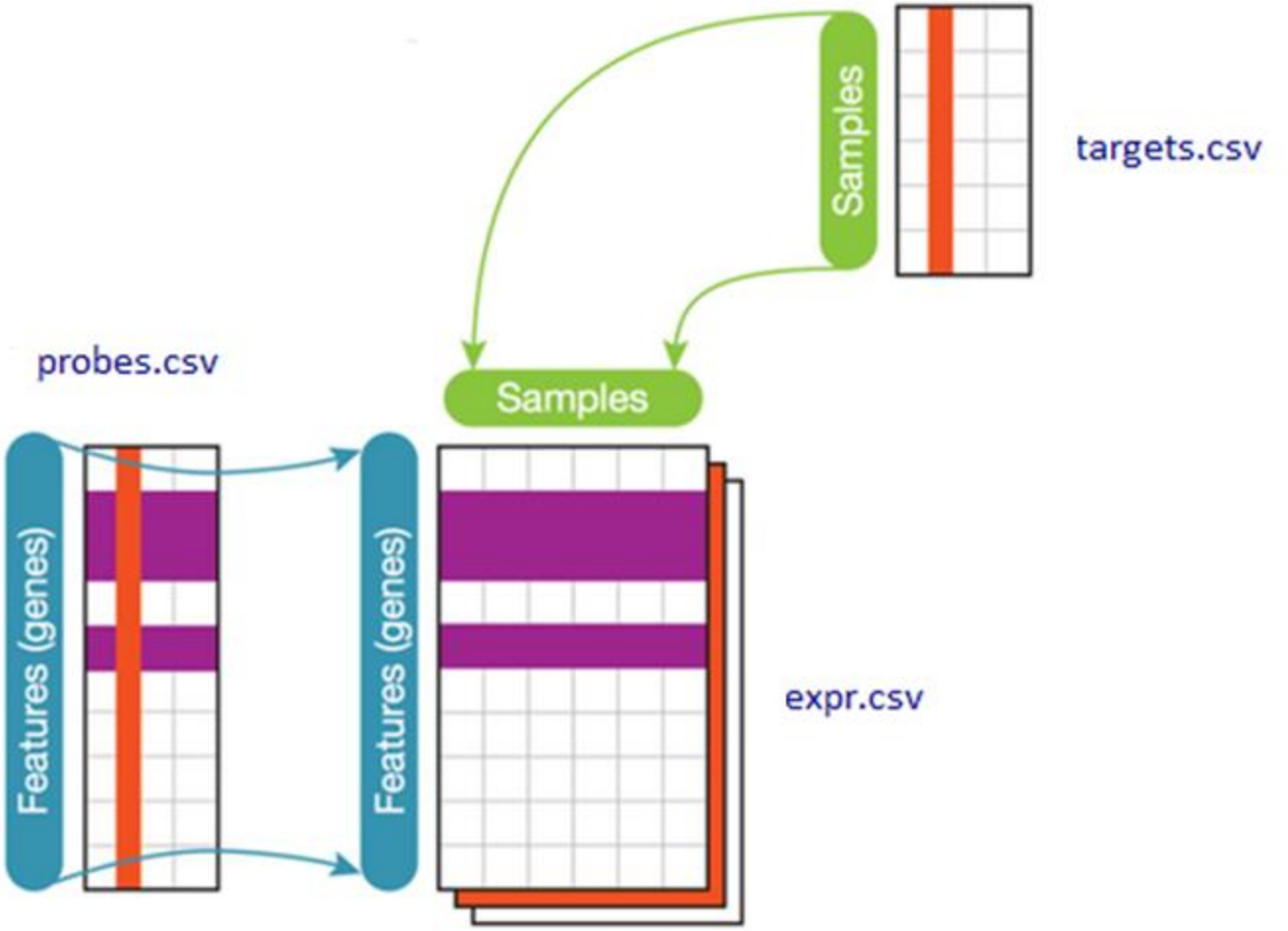
Data

- **GSE73002**

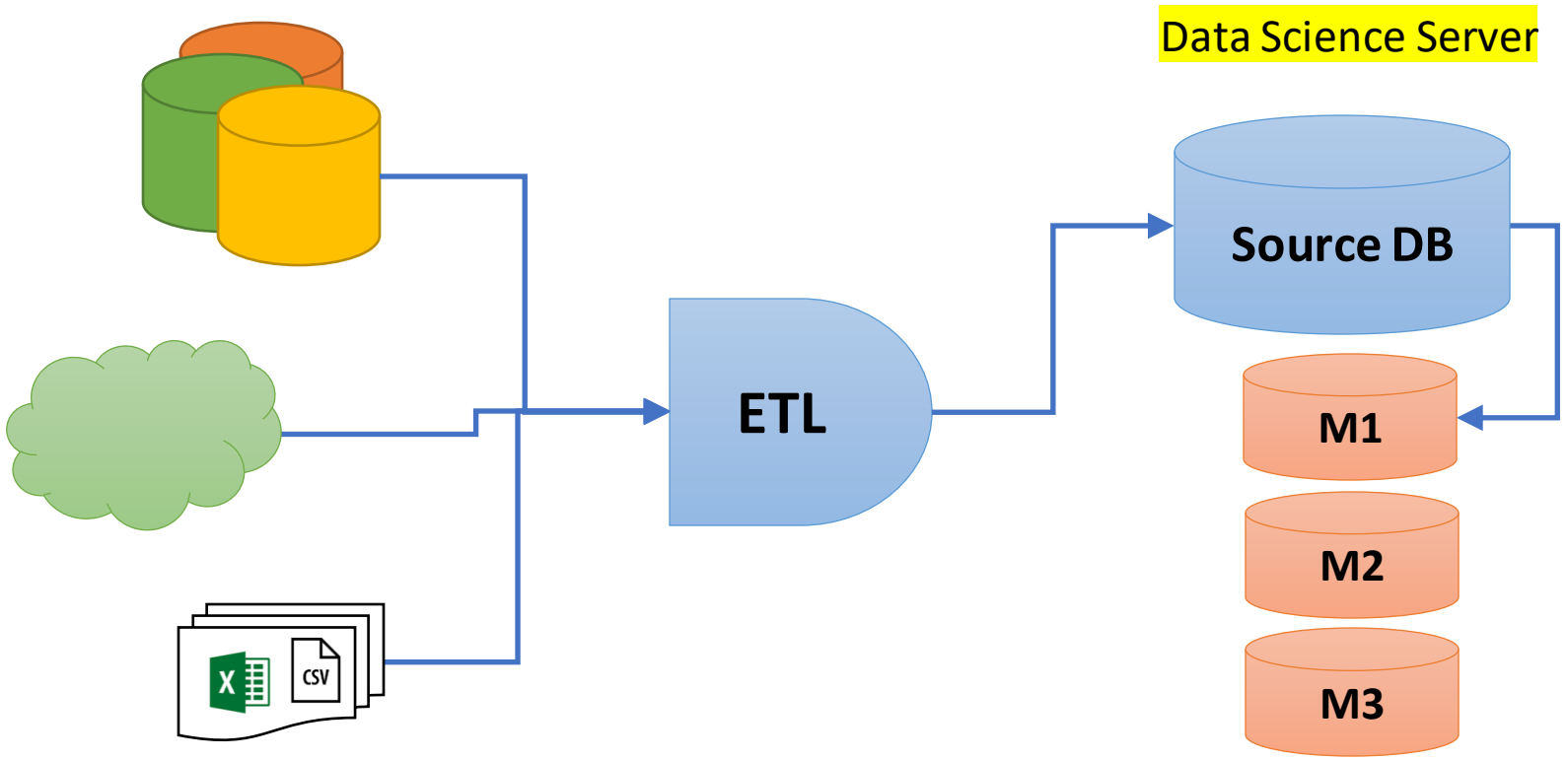
  https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73002
- **Title**

  A novel combination of **serum microRNAs** for detecting breast cancer in the early stage
- **1280** breast cancer and **2686** non-cancer control
- **Bioconductor** R packages can be used for data preparation

Data Model


Samples
targets.csv
probes.csv
Samples
Features (genes)
Features (genes)
expr.csv

Data Management

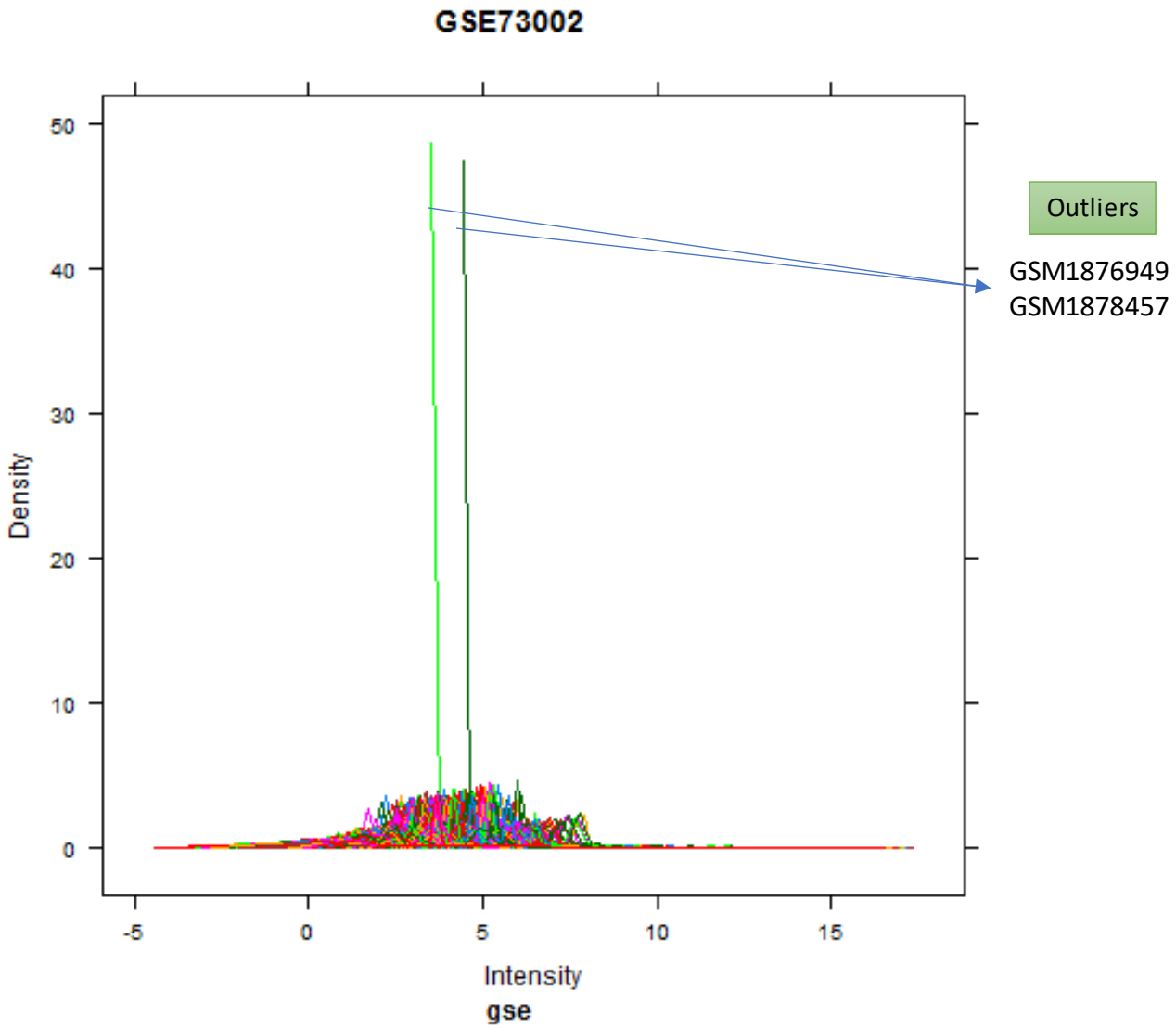


3- Data Exploration

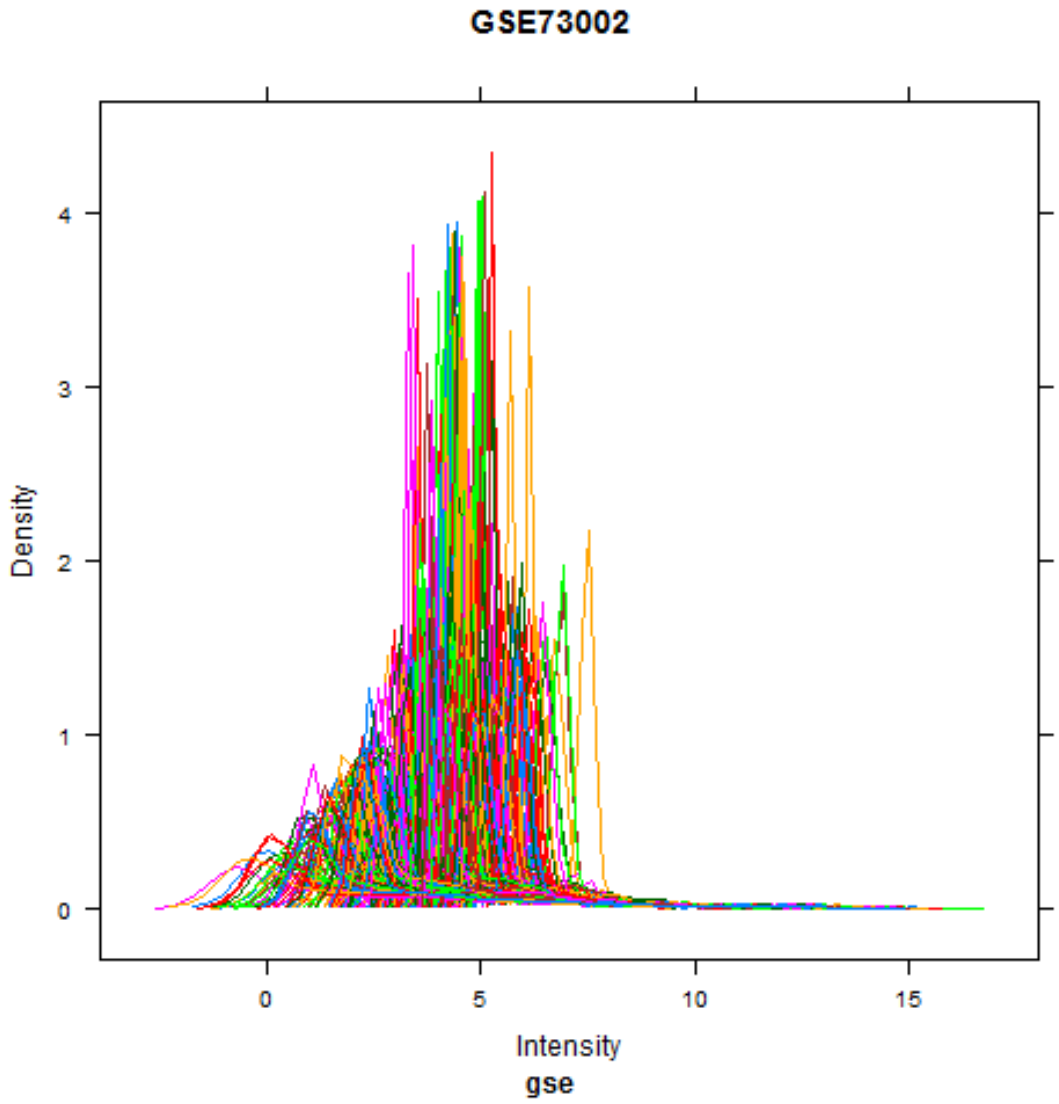
- **Univariate Analysis** on all variables/probes
- **Bivariate Analysis** at least against the target

Expression data - density plot

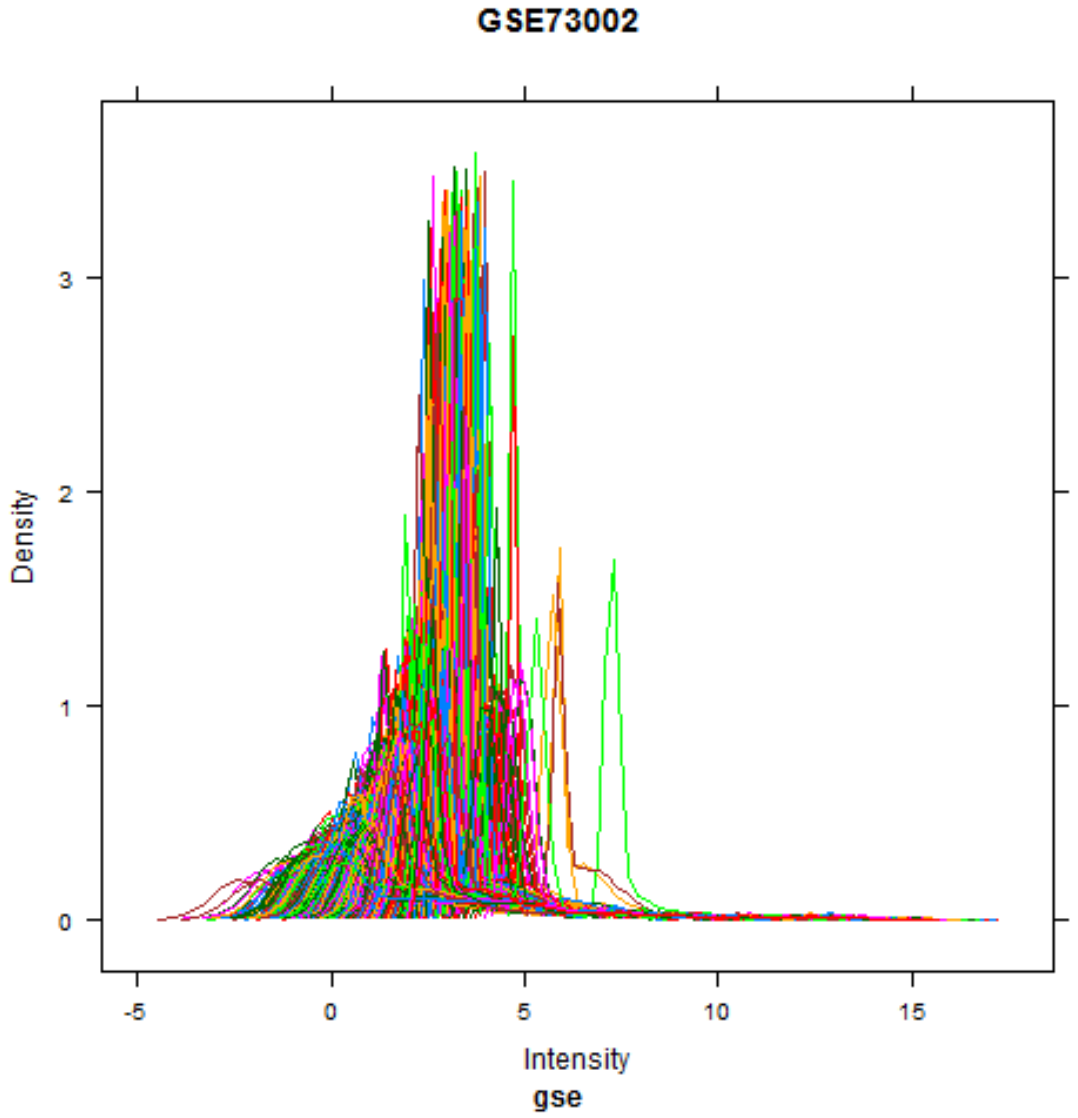
breast cancer	1280
non-cancer	2686



Exploration data - Breast Cancer



Exploration data – Normal Breast

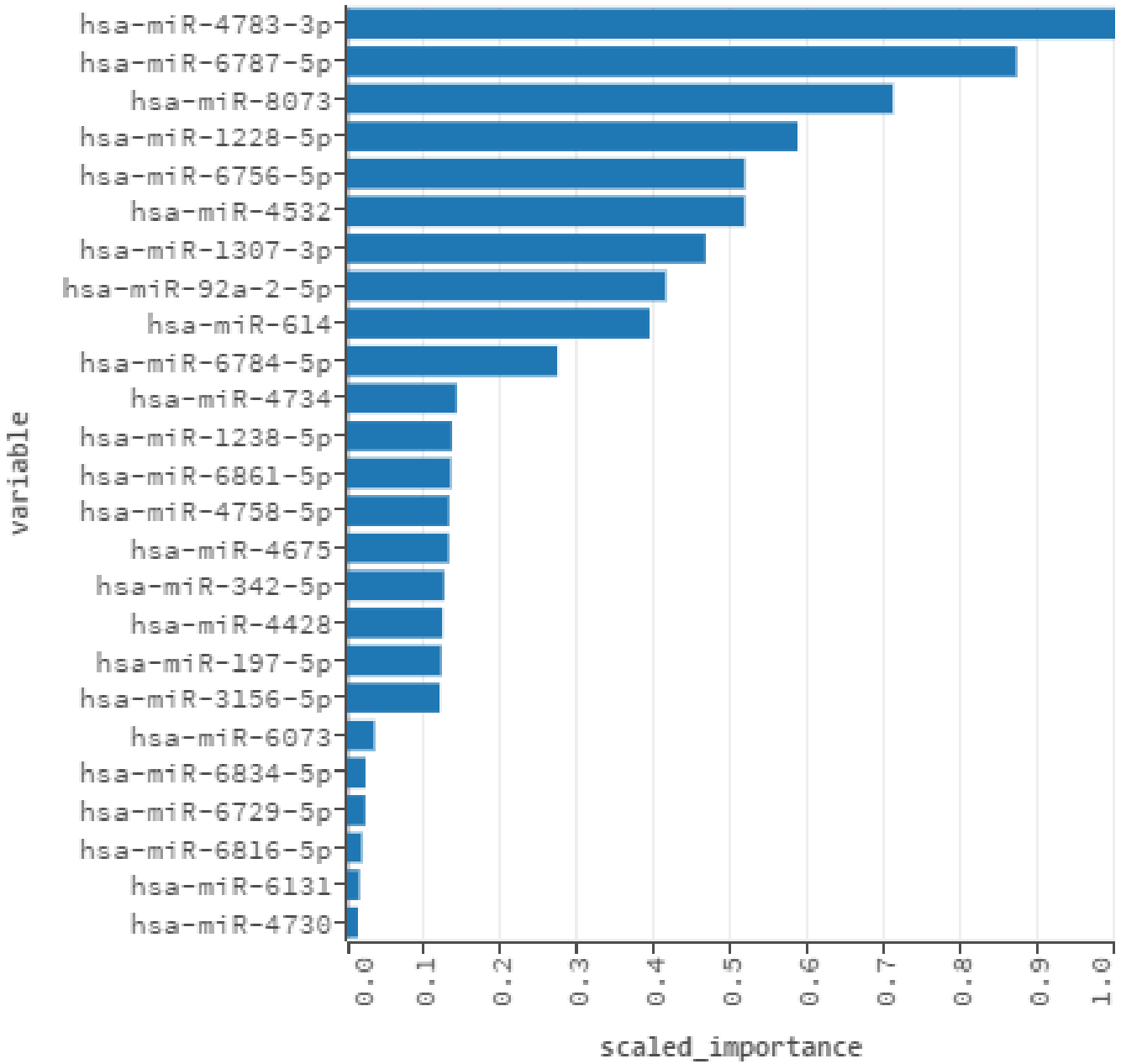


4- Modeling

1. Dividing data to **trainset** and **testset**
2. Selecting **top N probes** using trainset only and a proper statistical test. N is usually between 10-100
3. Building predictive models to choose the final set of probes (**biomarkers**)
   - Building a **classification model** to find the best subset of probes (biomarkers) to separate the breast cancers from the non-cancers cases

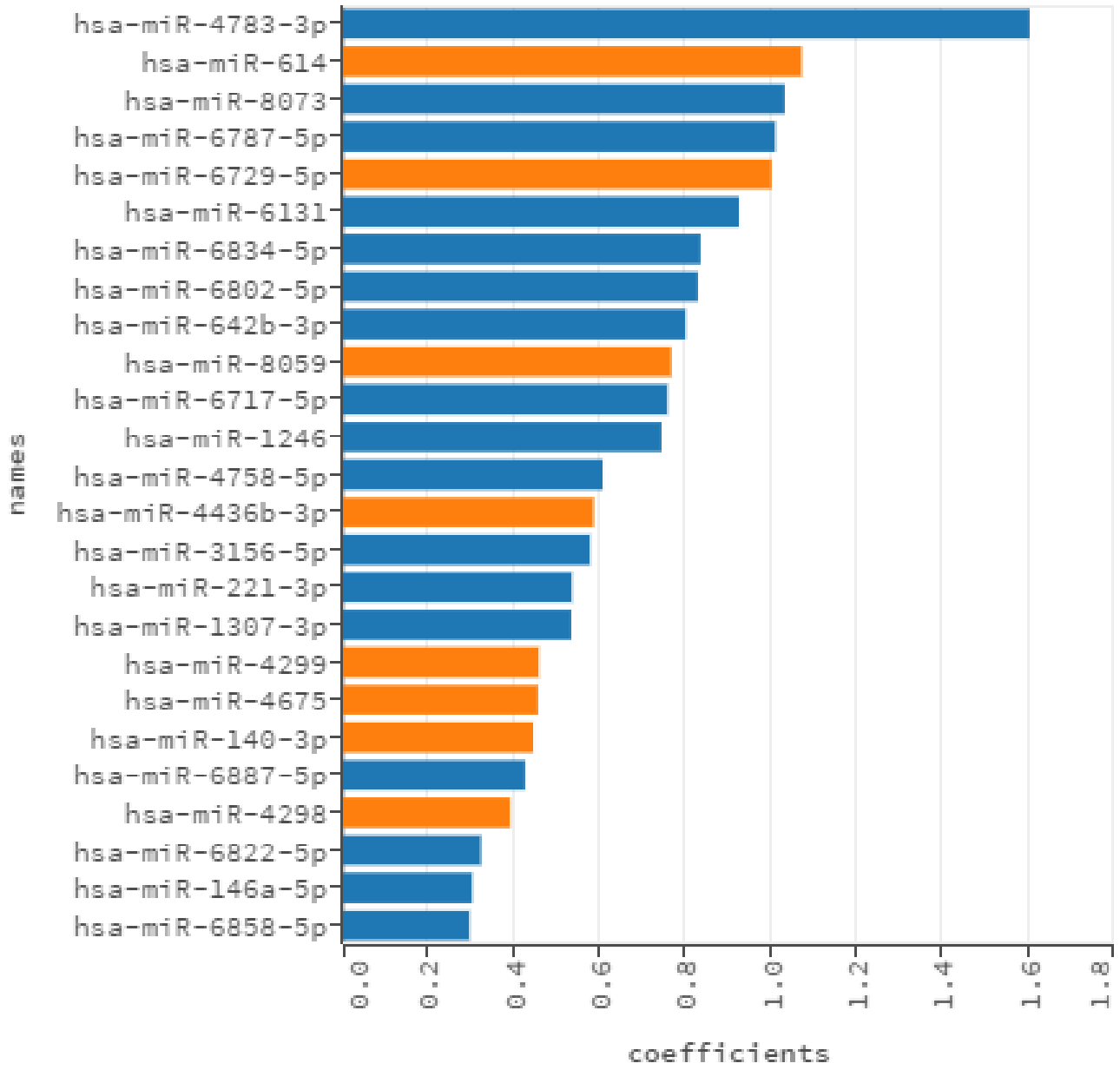
Modeling - RandomForest

Accuracy 98%



Modeling - GBM

Accuracy 99%



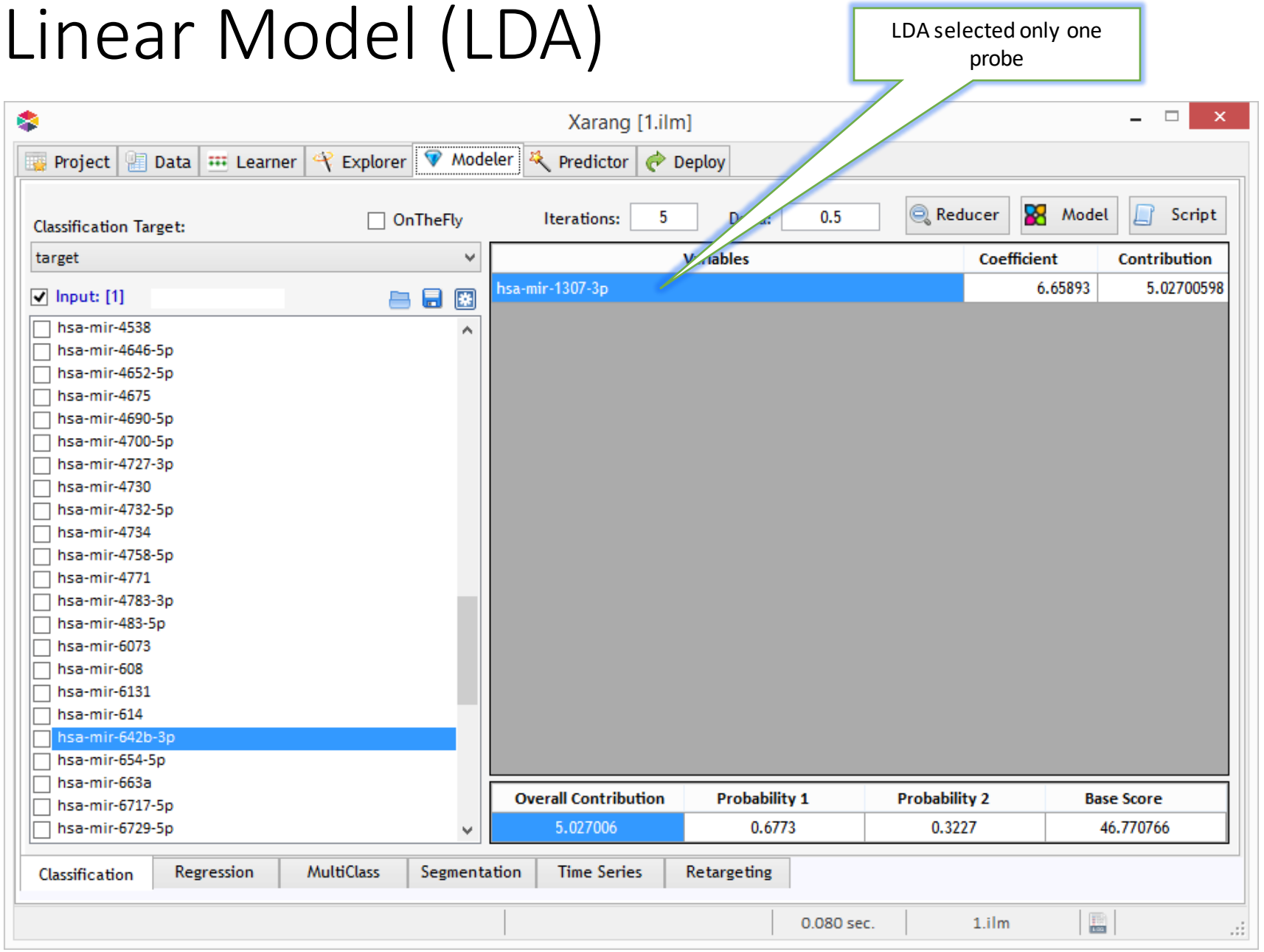
Modeling – Deep Learning

Accuracy 99%

DeepLearning does not have a straightforward variable selection capability.

Modeling – Linear Model (LDA)

Accuracy 99%



5- Model Evaluation

	RandomForest	GLM	DeepLearning	LDA
Accuracy	✔	✔	✔	✔
Probe Importance	✔	✔	✖	✔
Federated Learning	✖	✖	✖	✔

6- Model Deployment

- **3 levels of model testing**

1. Testing models on the **test platform** by the modeling team
2. Testing models by the **QA team**
3. A/B test on the **production platform**

Summary

- Hire a **Solution Architect**
- Implement **Data Science 6-Step** and be religious about it
- Have a **clear understanding** of **the question/problem** and how to **measure success**
- **Own the data!** Data is your reserve currency. If you do not want to become bankrupt always make sure you have full control of your data
- Always do **univariate** and **bivariate analysis**
- **A/B Test!**
- Choose the **right ML toolbox**

Thank You!

saed@bioada.com