STRATIO

DISTRIBUTED LOGISTIC MODEL TREES

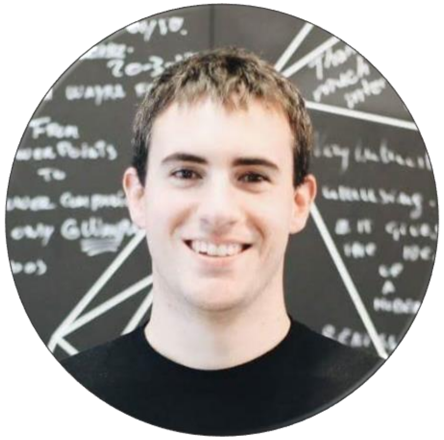
16 NOV 2016 @ APACHE BIG DATA EUROPE

Distributed Logistic Model Trees, Stratio Intelligence

Mateo Álvarez and **Antonio Soriano**

@StratioBD

PROFILE



MATEO ÁLVAREZ

Aerospace Engineer, MSc in Propulsion Systems (UPM), Master in Data Science (URJC). Working as data scientist and Big Data developer at Stratio Big Data in the data science department

in mateo-alvarez

SKILLS

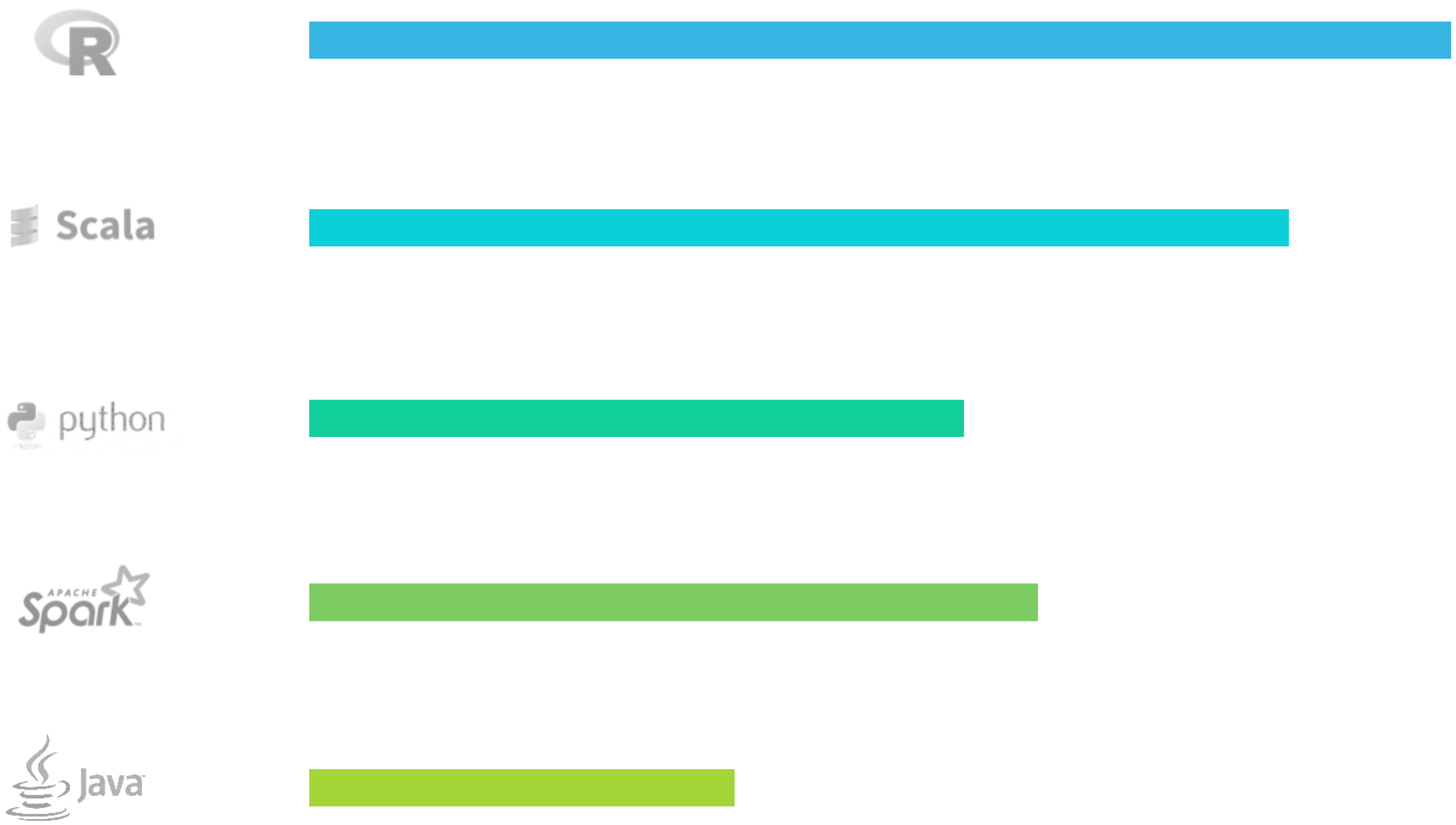
PROFILE

ANTONIO SORIANO

Ph.D. in Telecommunications, MSc in Electronic Systems Engineering and Telecommunication Technologies, Systems and Networks (UPV), and MSc “Big Data Expert” (UTAD).
Working as data scientist and Big Data developer at at Stratio Big Data in the data science department

@Phd_A_Soriano

SKILLS



INDEX

INTERPRETABLE ALGORITHMS

- Why use interpretable algorithms instead of “black boxes”
- Logistic Regression
- Decision Trees
- Variance-Bias tradeoff

@StratioBD

INTERPRETABLE ALGORITHMS

Accuracy

VS

Explainability

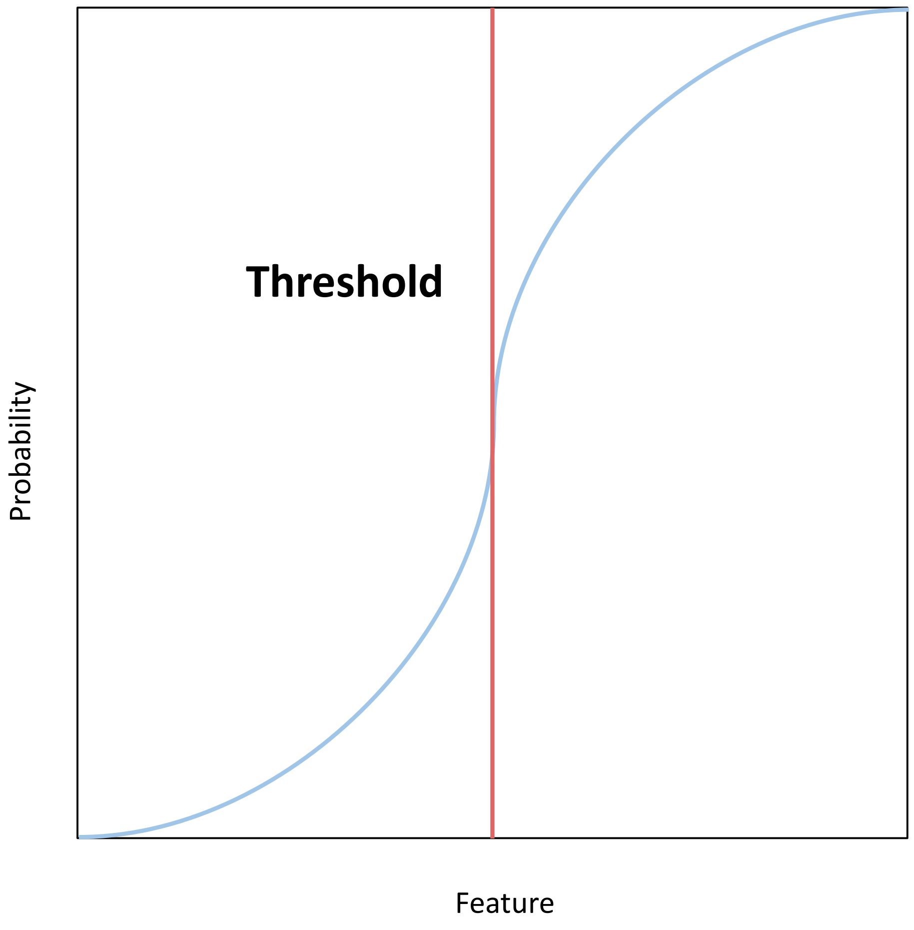
Medical Studies

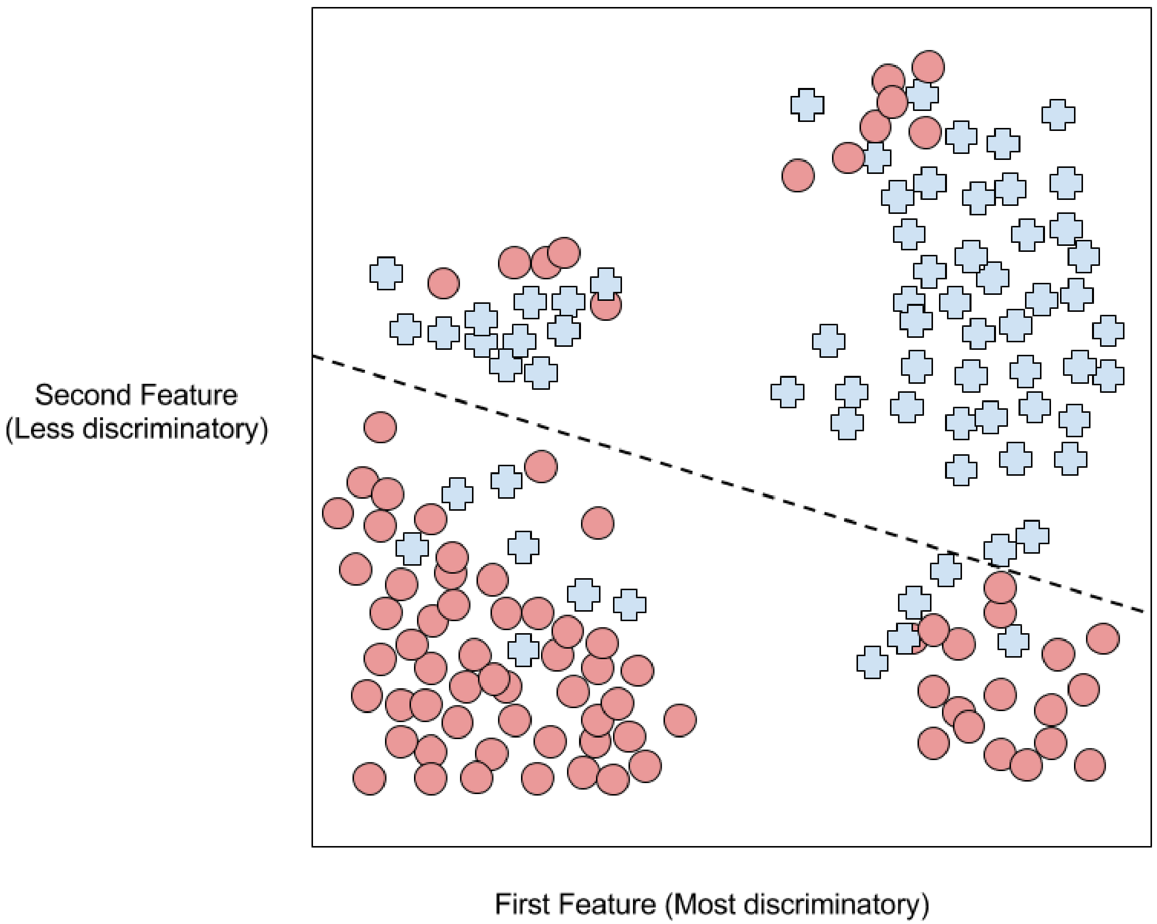
Power management

Financial environment

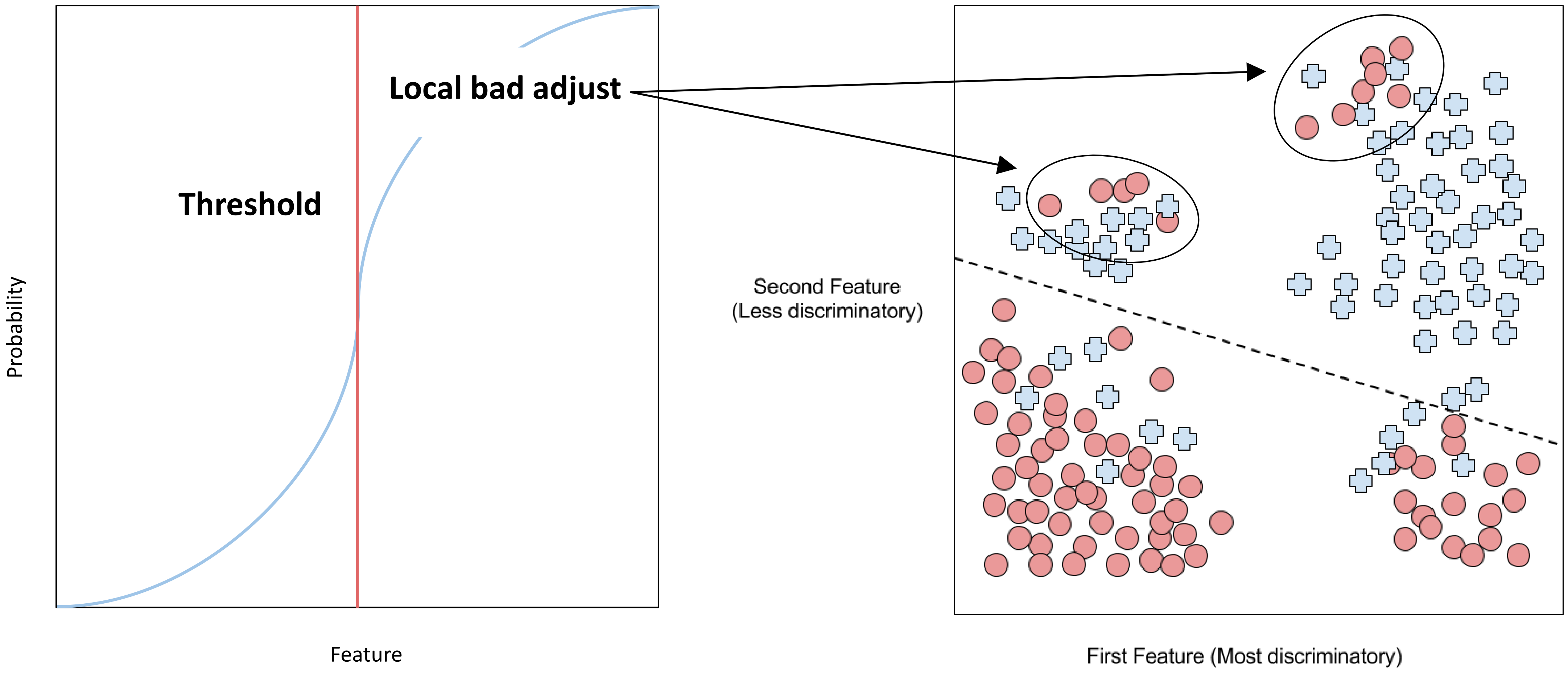
Criminal activity

LOGISTIC REGRESSION

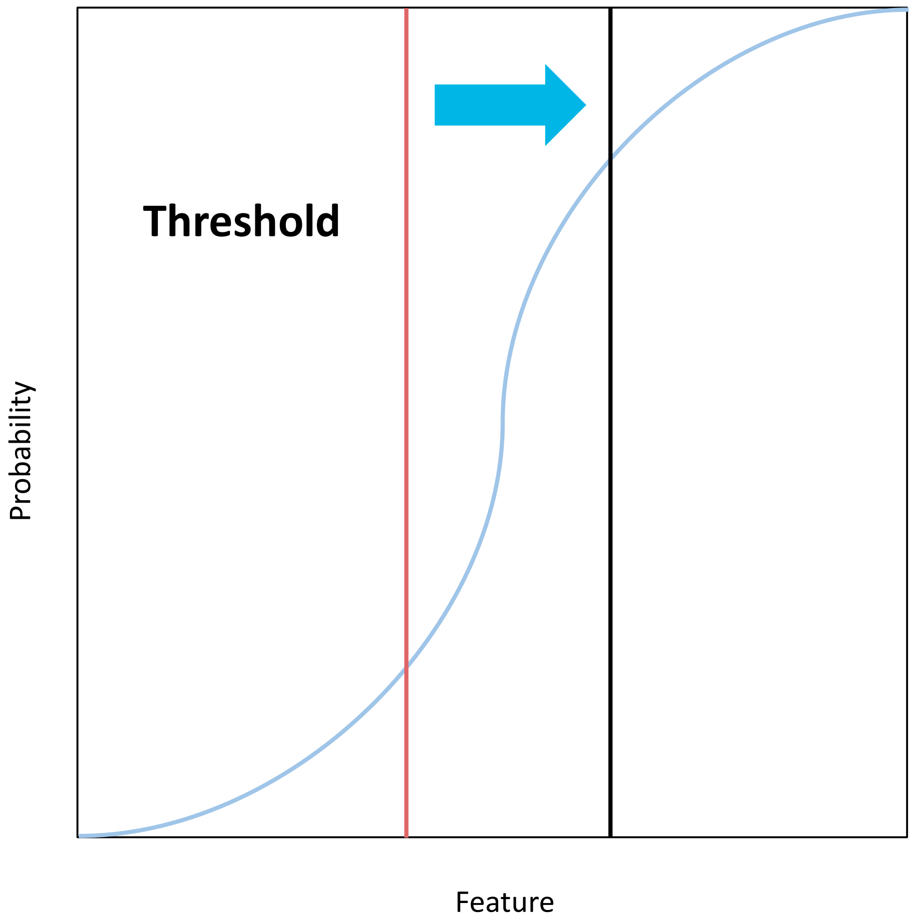


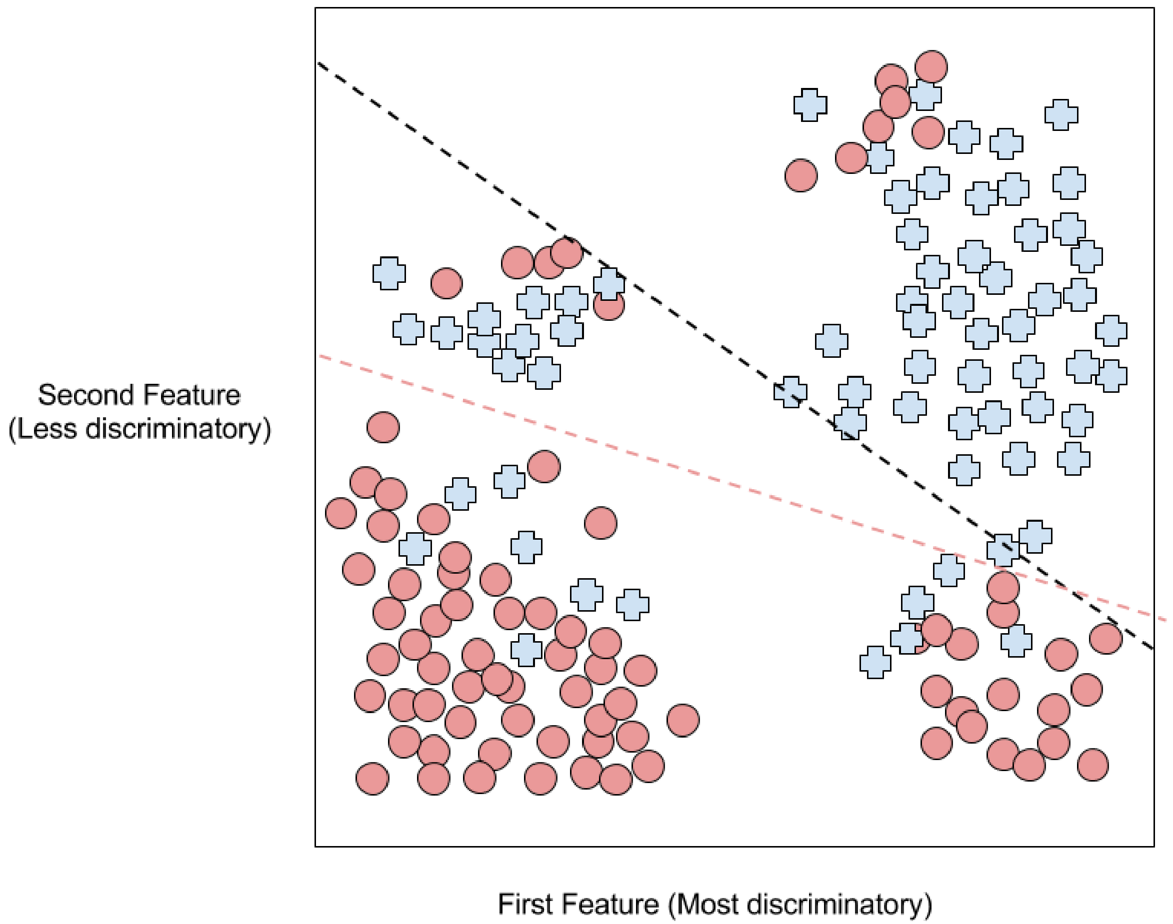


LOGISTIC REGRESSION

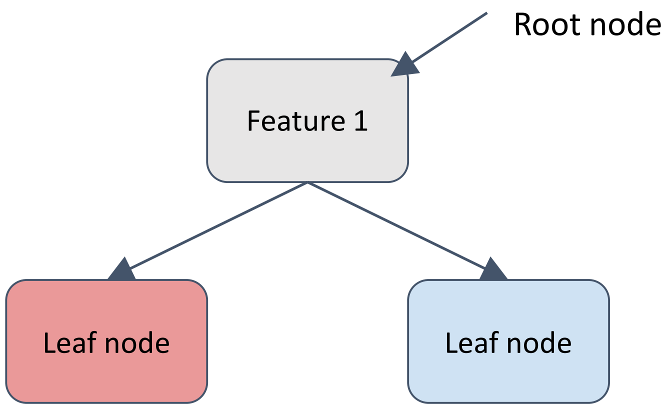


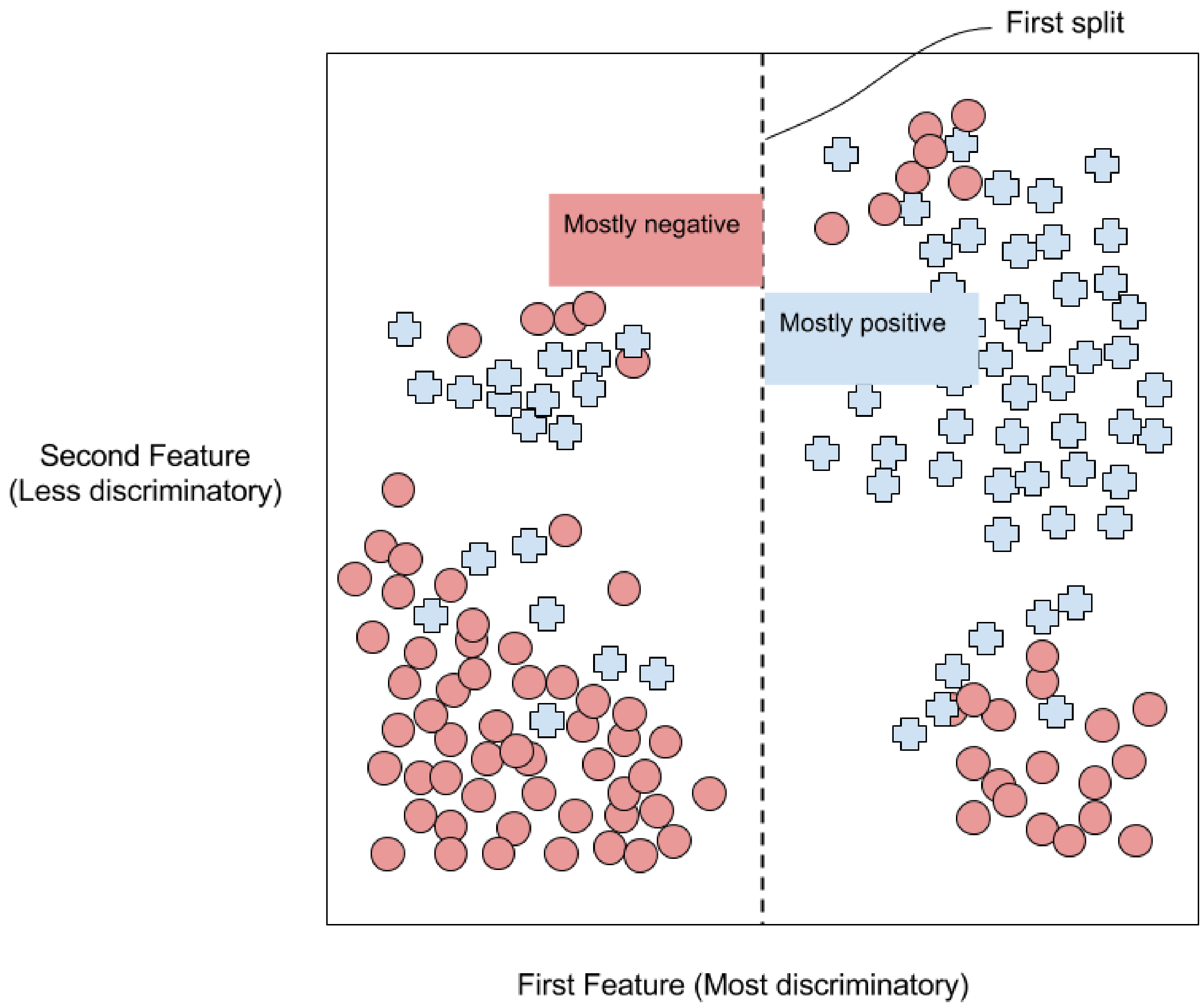
LOGISTIC REGRESSION



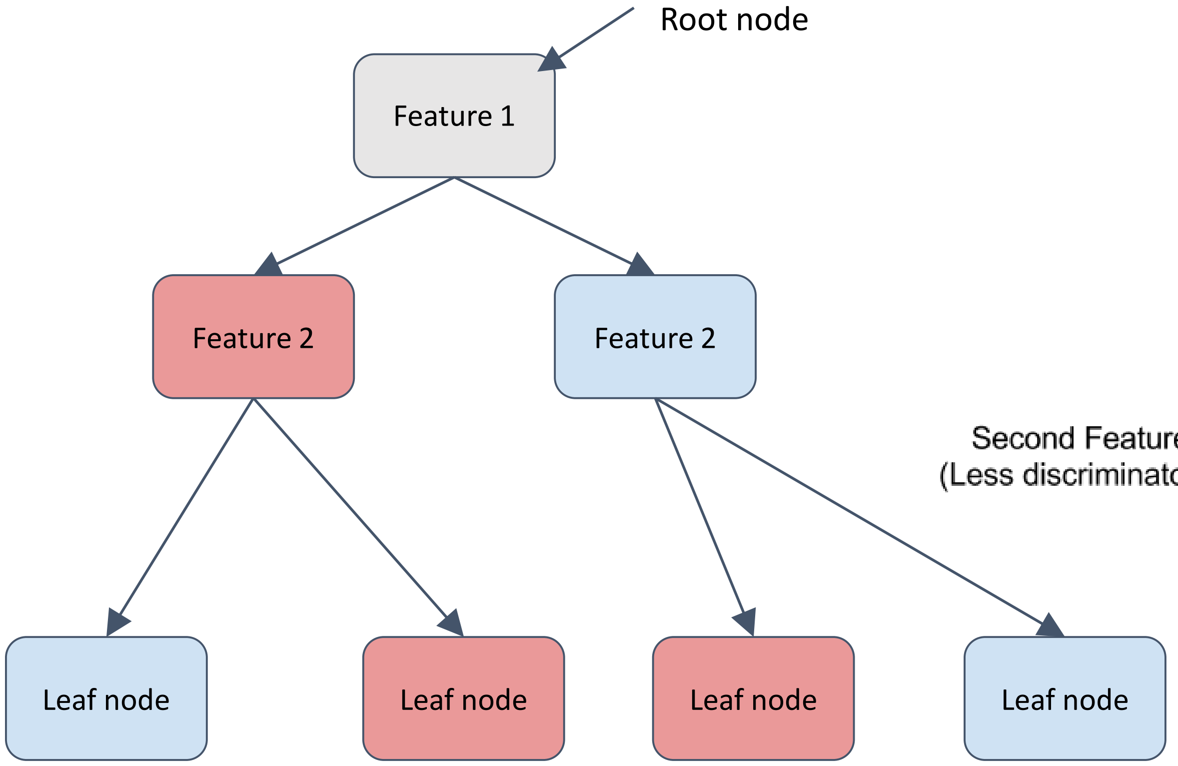


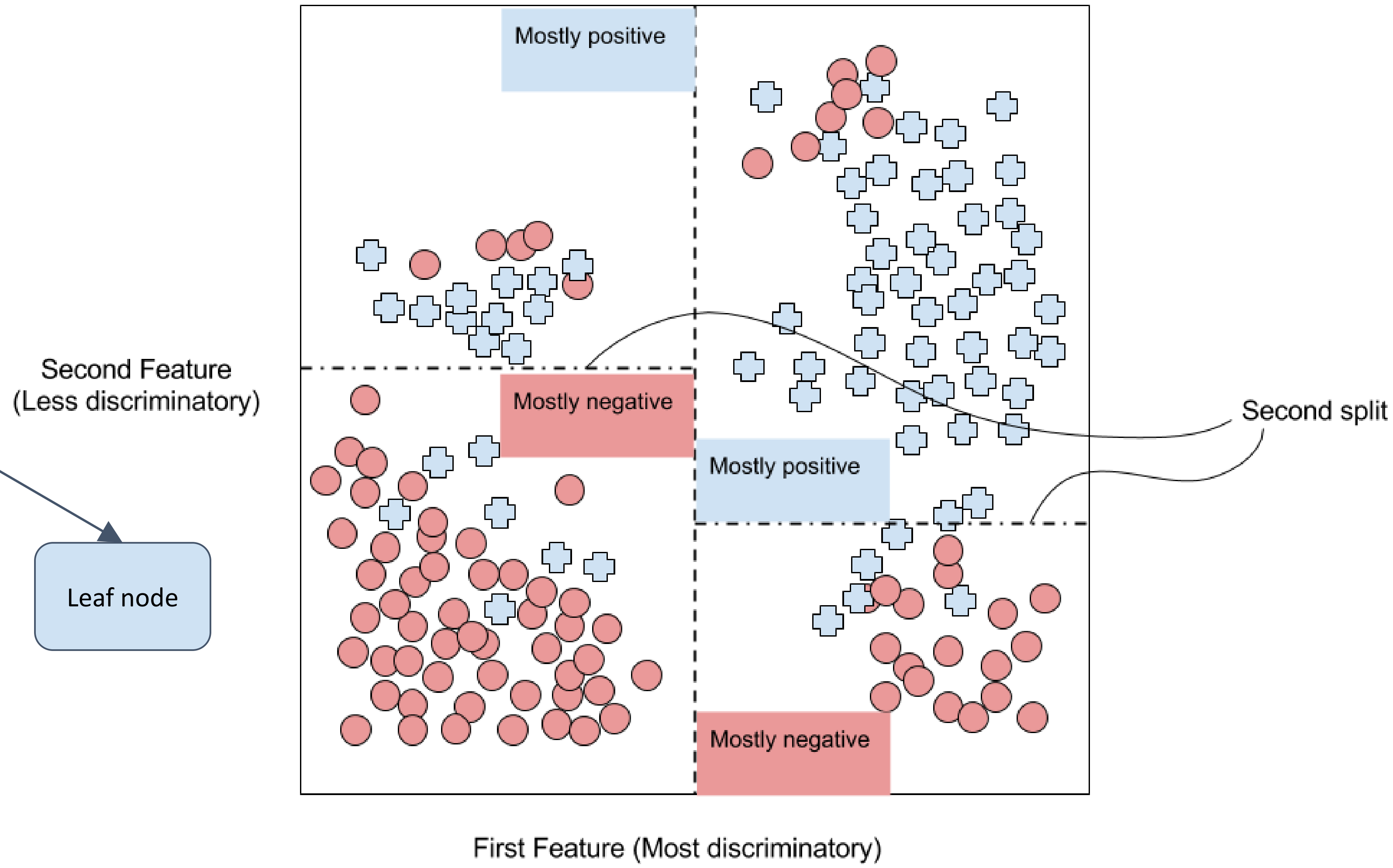
DECISION TREES



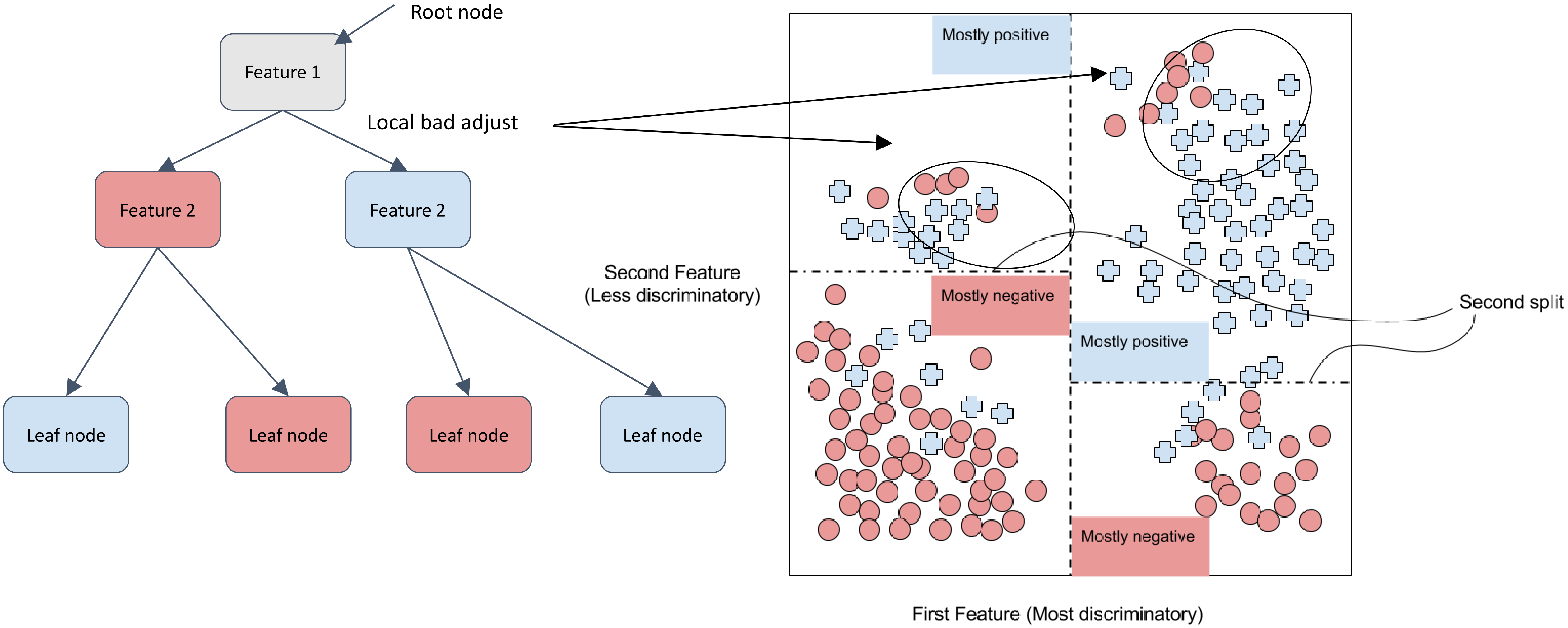


DECISION TREES

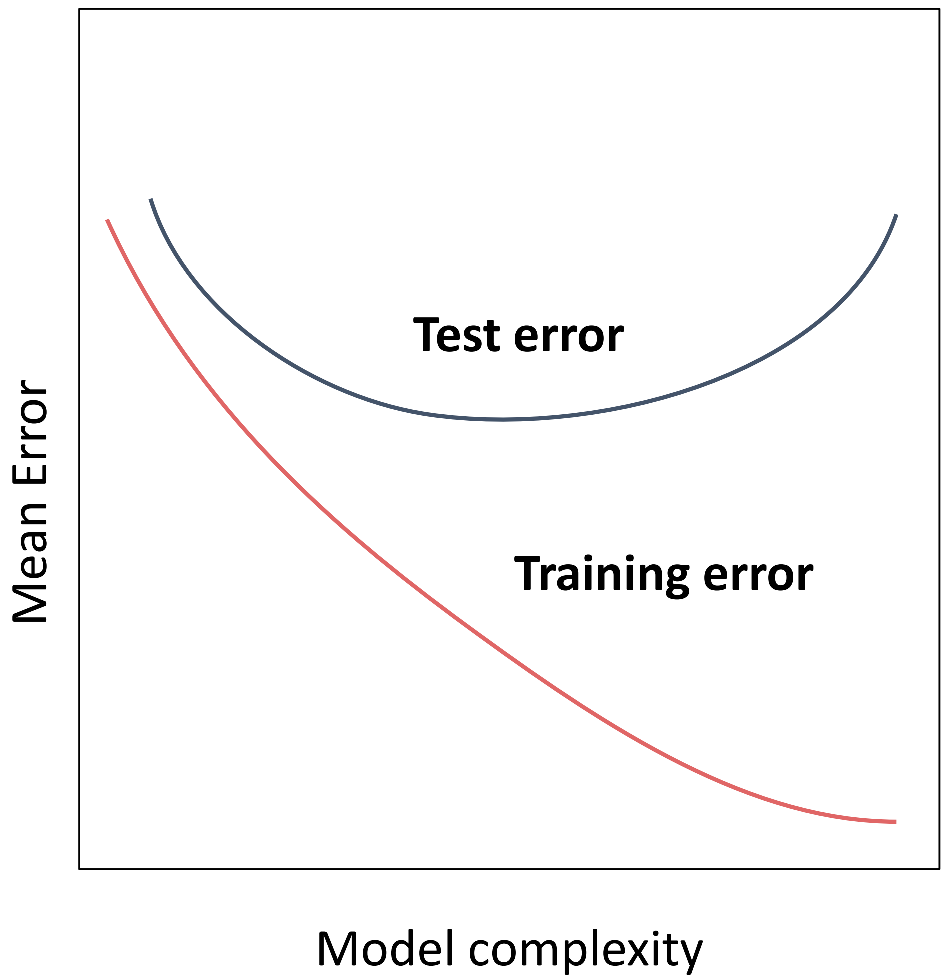


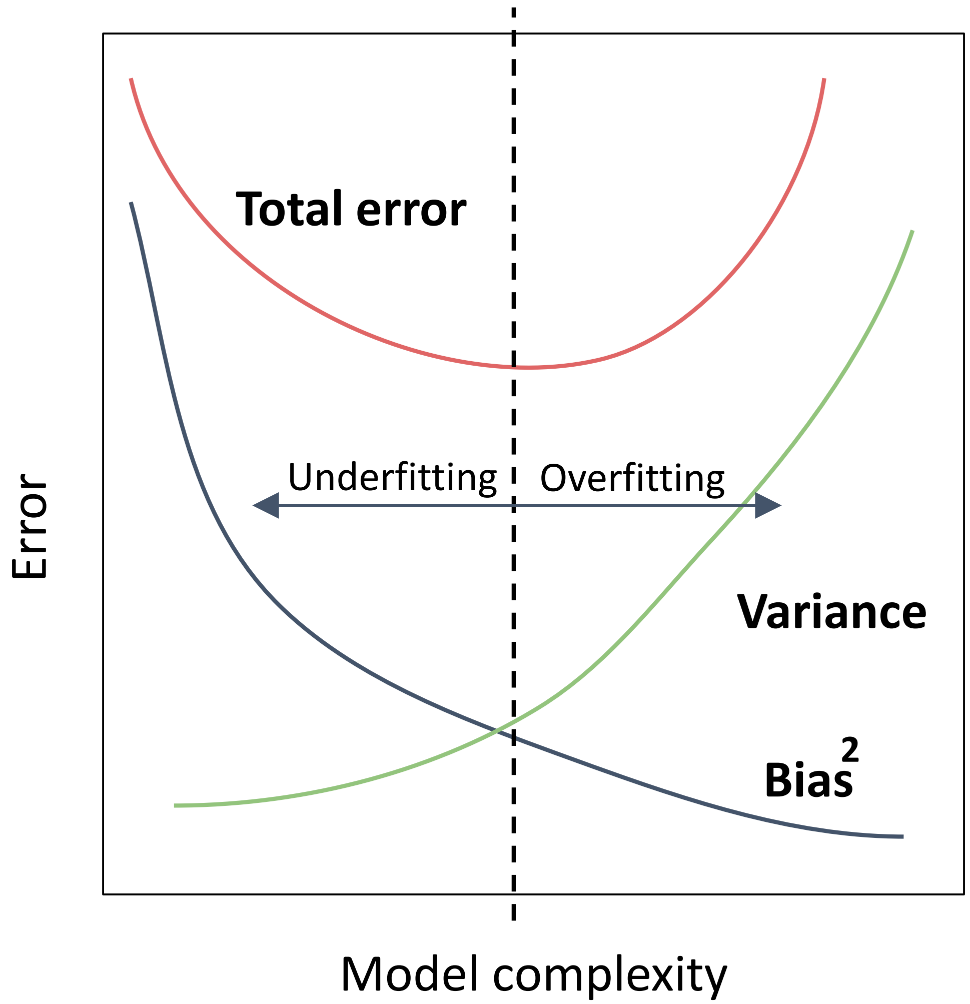


DECISION TREES



BIAS-VARIANCE TRADE-OFF





BIAS-VARIANCE TRADE-OFF

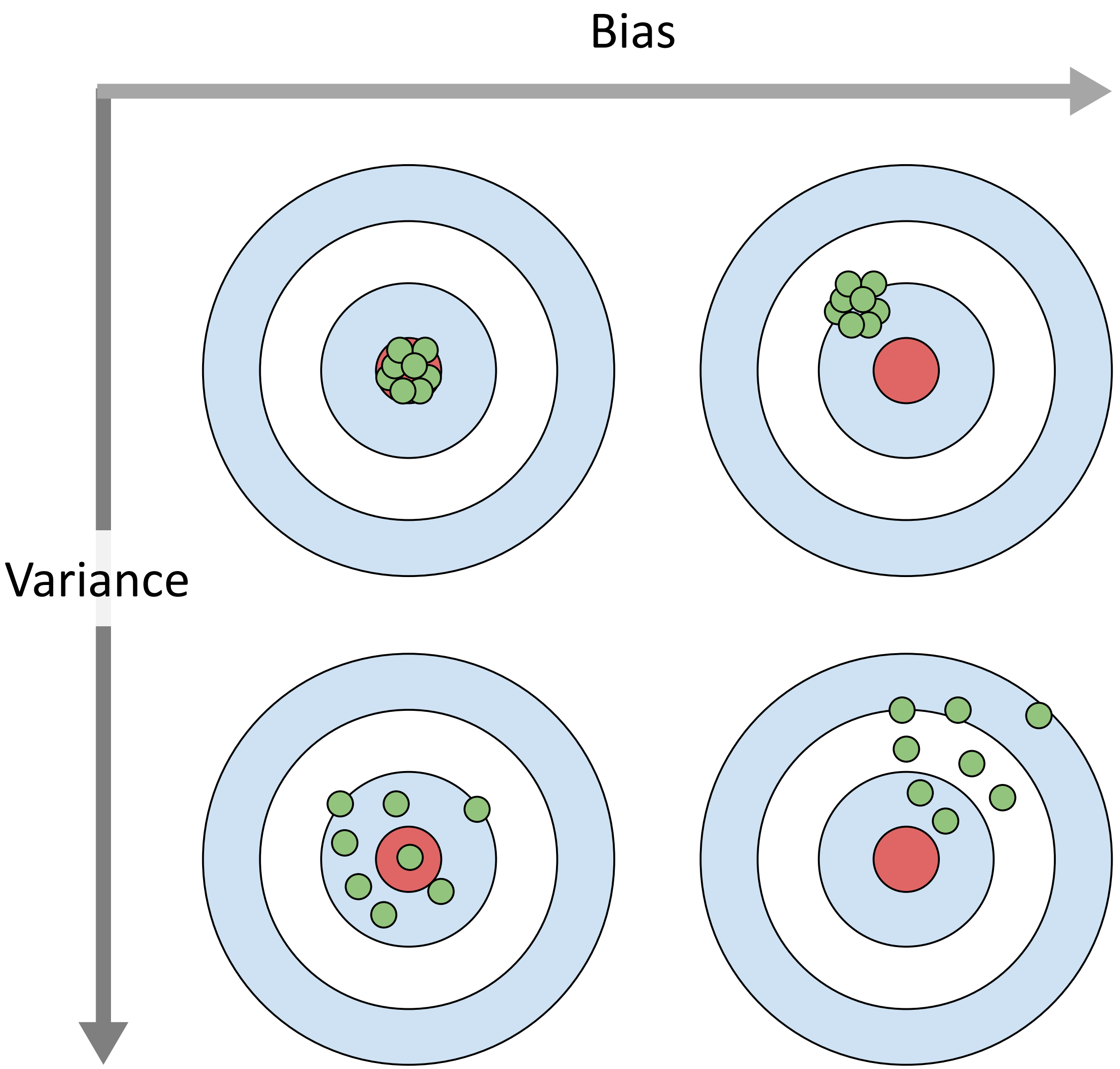
$$\mathrm{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Bias}\left[\hat{f}(x)\right]^2 + \mathrm{Var}\left[\hat{f}(x)\right] + \sigma^2$$

Where:

$$\mathrm{Bias}\left[\hat{f}(x)\right] = \mathrm{E}\left[\hat{f}(x) - f(x)\right]$$

and

$$\mathrm{Var}\left[\hat{f}(x)\right] = \mathrm{E}[\hat{f}(x)^2] - \mathrm{E}[\hat{f}(x)]^2$$



BIAS-VARIANCE TRADE-OFF

Missing important variables for the problem to make the predictions

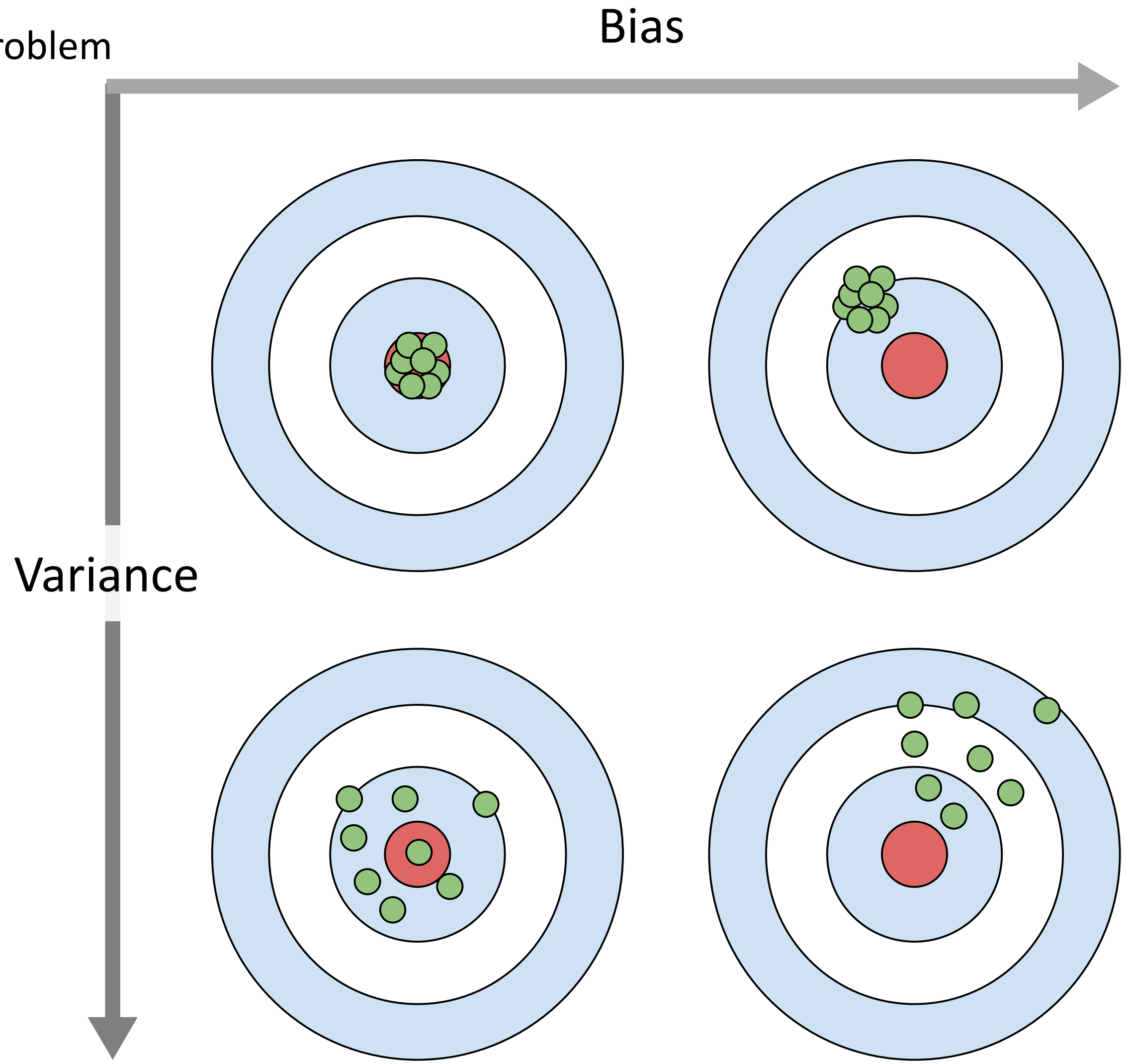
$$\mathrm{E}\Big[\big(y-\hat{f}(x)\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(x)\big]^2 + \mathrm{Var}\big[\hat{f}(x)\big] + \sigma^2$$

Where:

$$\mathrm{Bias}\big[\hat{f}(x)\big] = \mathrm{E}\big[\hat{f}(x) - f(x)\big]$$

and

$$\mathrm{Var}\big[\hat{f}(x)\big] = \mathrm{E}[\hat{f}(x)^2] - \mathrm{E}[\hat{f}(x)]^2$$



BIAS-VARIANCE TRADE-OFF

Overfitting to the sample/training data

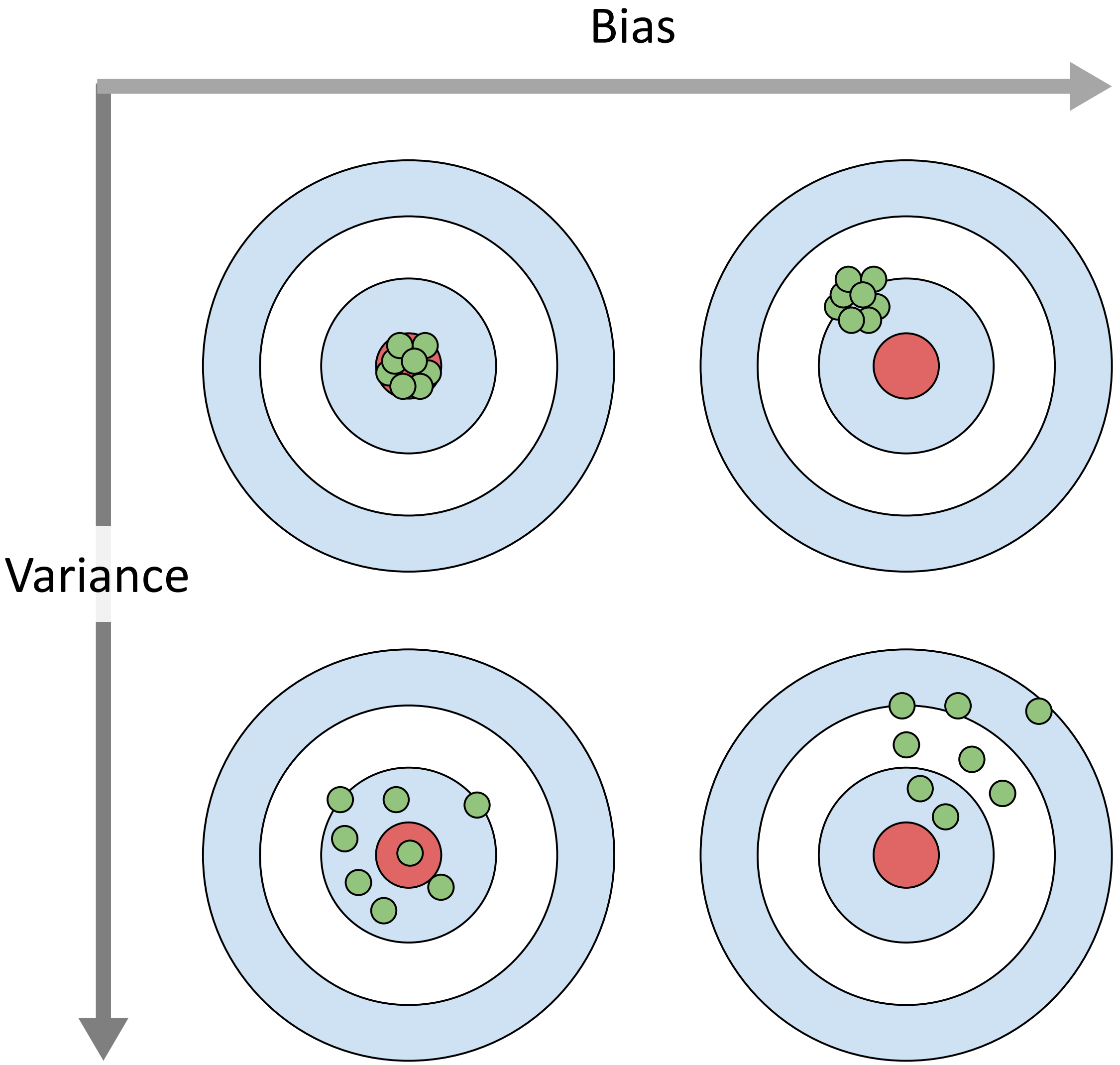
$$E\left[\left(y-\hat{f}(x)\right)^2\right] = \mathrm{Bias}\left[\hat{f}(x)\right]^2 + \mathrm{Var}\left[\hat{f}(x)\right] + \sigma^2$$

Where:

$$\mathrm{Bias}\left[\hat{f}(x)\right] = E\left[\hat{f}(x) - f(x)\right]$$

and

$$\mathrm{Var}\left[\hat{f}(x)\right] = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2$$



BIAS-VARIANCE TRADE-OFF

Irreducible error on prediction

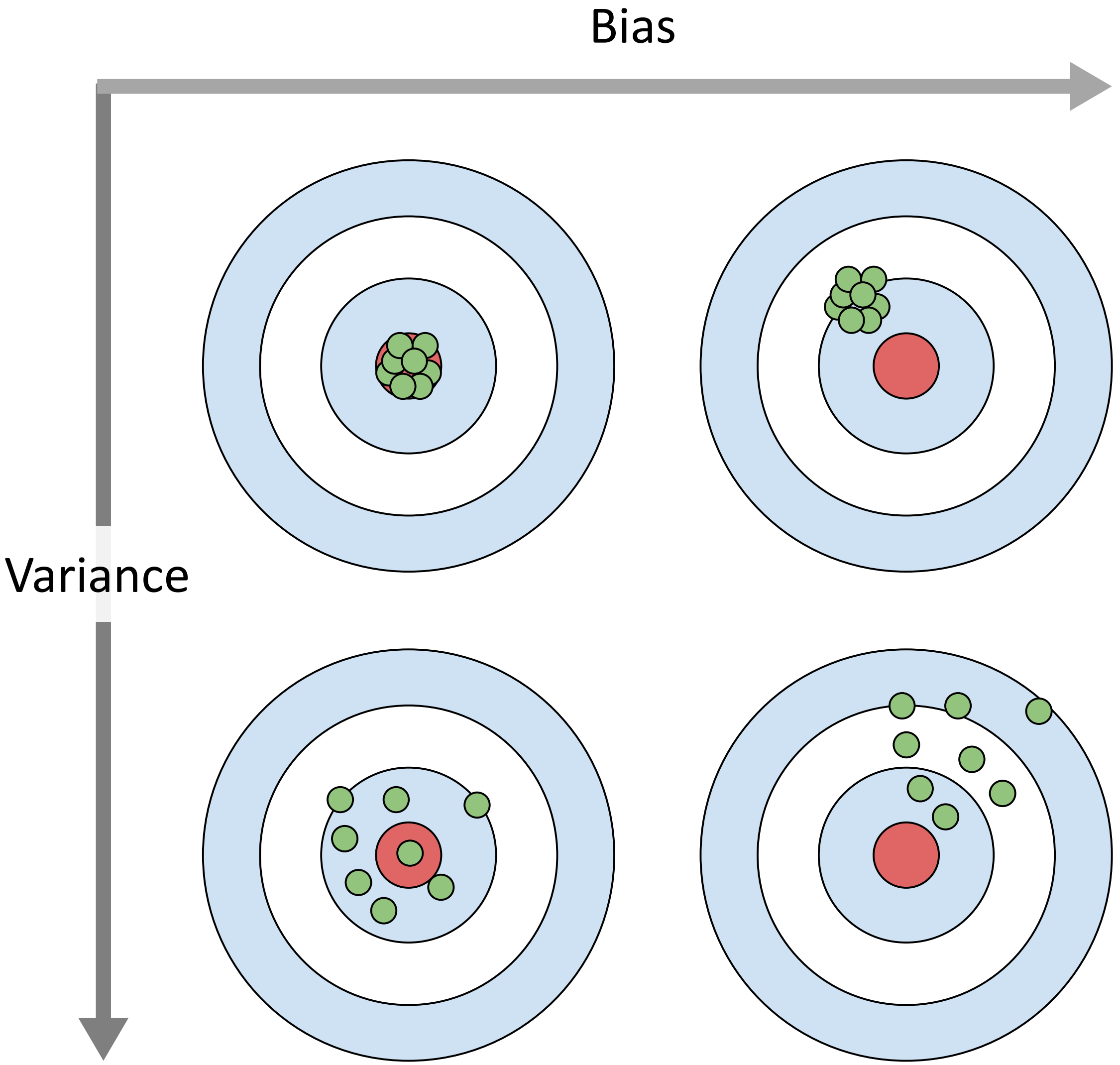
$$\mathrm{E}\left[\left(y-\hat{f}(x)\right)^2\right] = \mathrm{Bias}\left[\hat{f}(x)\right]^2 + \mathrm{Var}\left[\hat{f}(x)\right] + \sigma^2$$

Where:

$$\mathrm{Bias}\left[\hat{f}(x)\right] = \mathrm{E}\left[\hat{f}(x) - f(x)\right]$$

and

$$\mathrm{Var}\left[\hat{f}(x)\right] = \mathrm{E}[\hat{f}(x)^2] - \mathrm{E}[\hat{f}(x)]^2$$

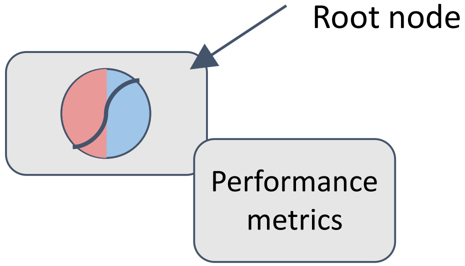


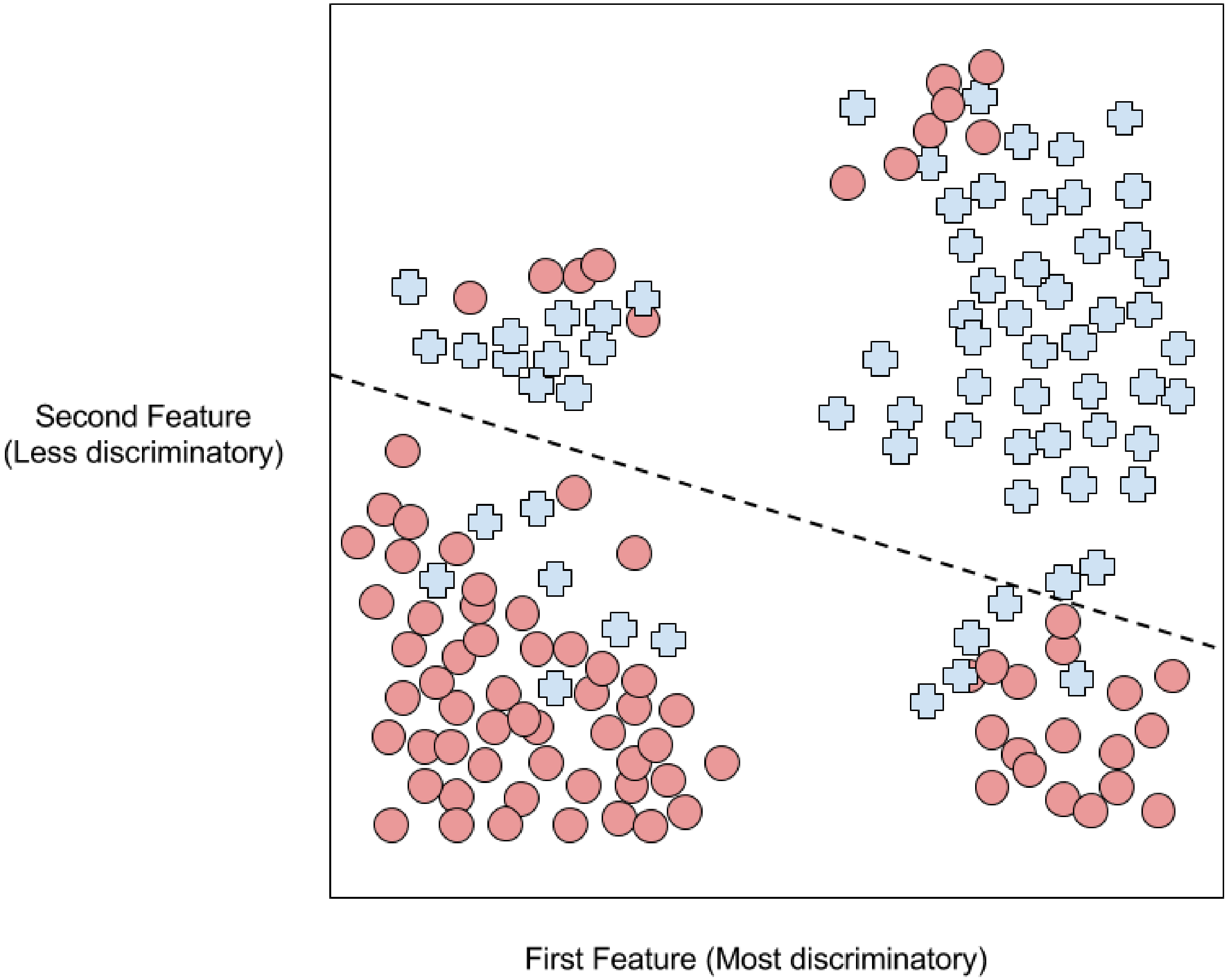
DISTRIBUTED LOGISTIC MODEL TREES

- Logistic Model Trees
- Distributed implementation
- Cost function & configuration parametres
- Demo

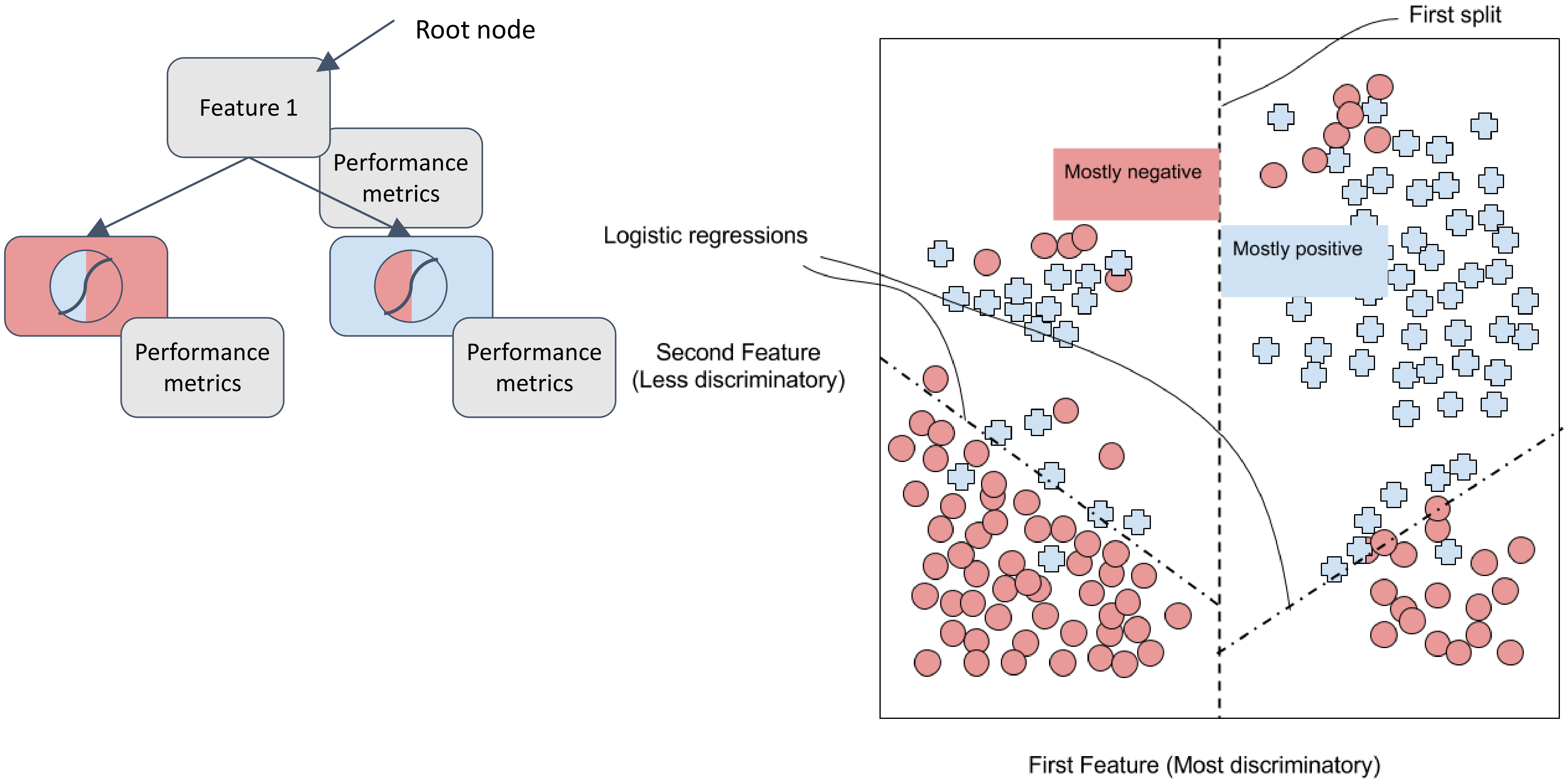
@StratioBD

LOGISTIC MODEL TREES

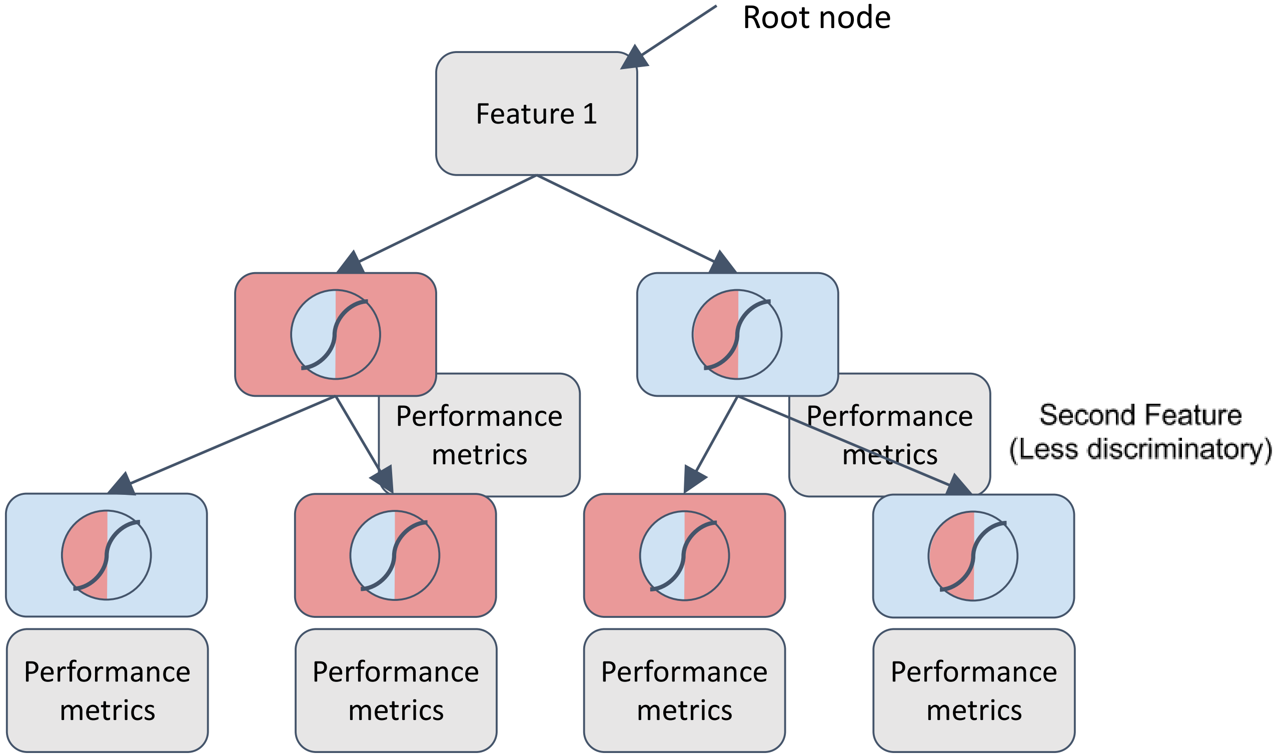


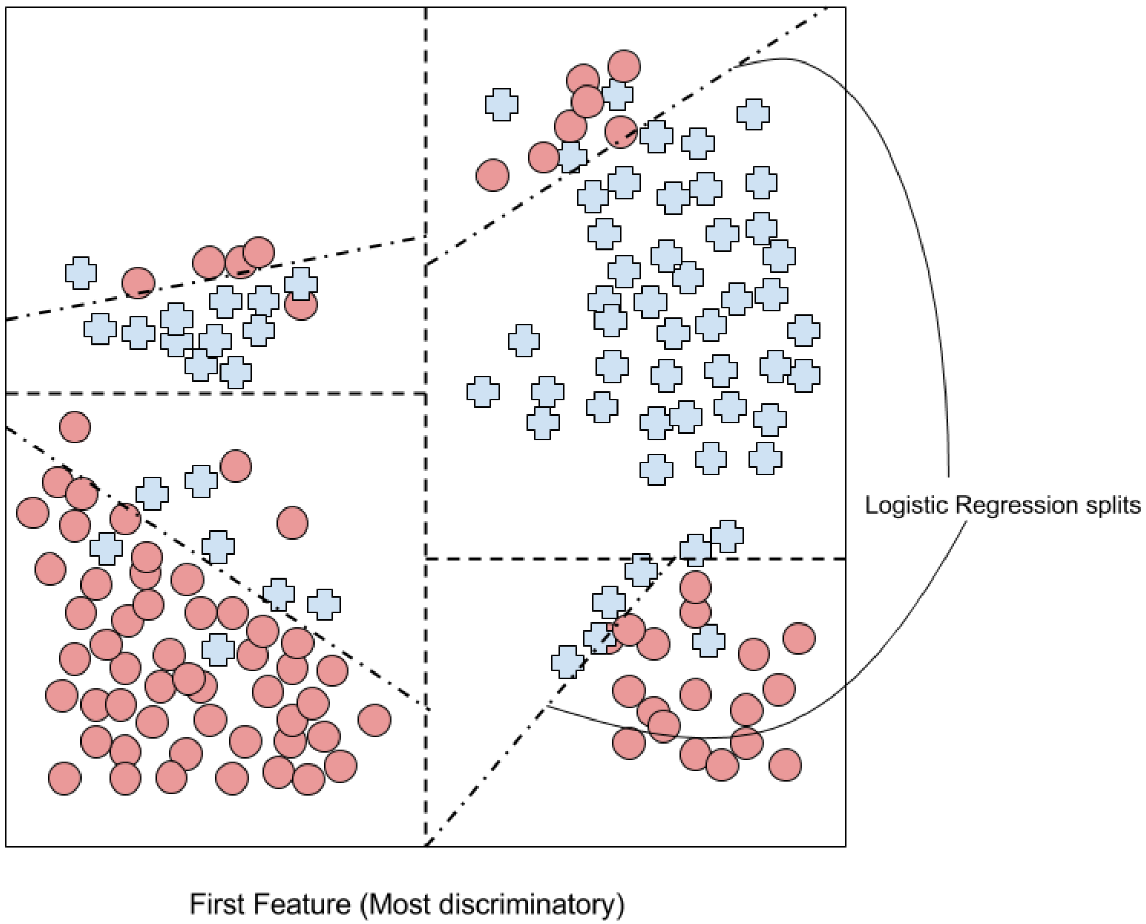


LOGISTIC MODEL TREES

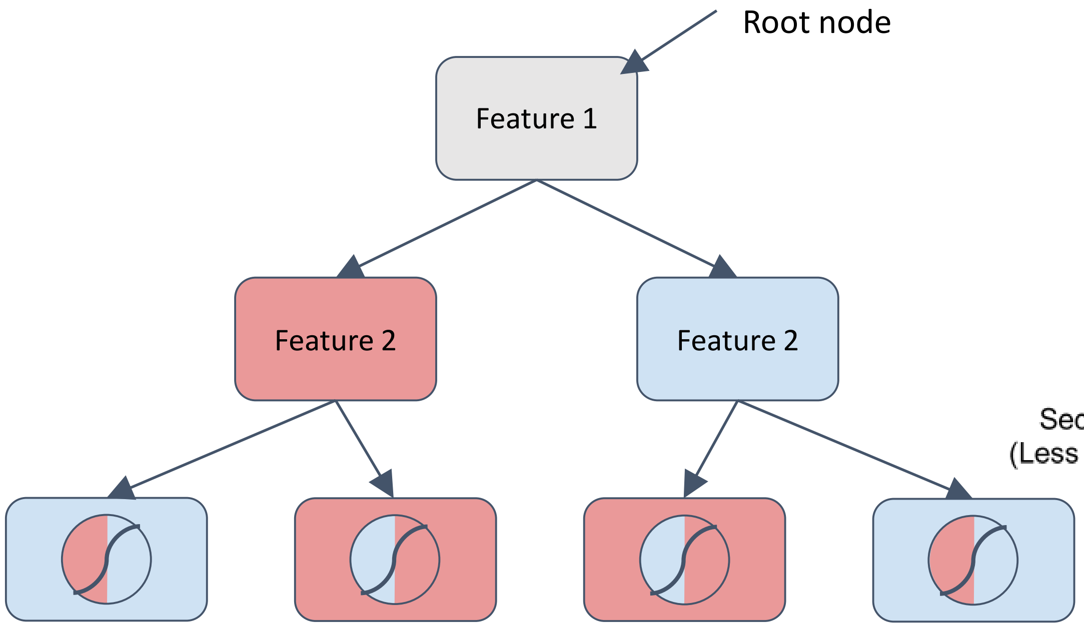


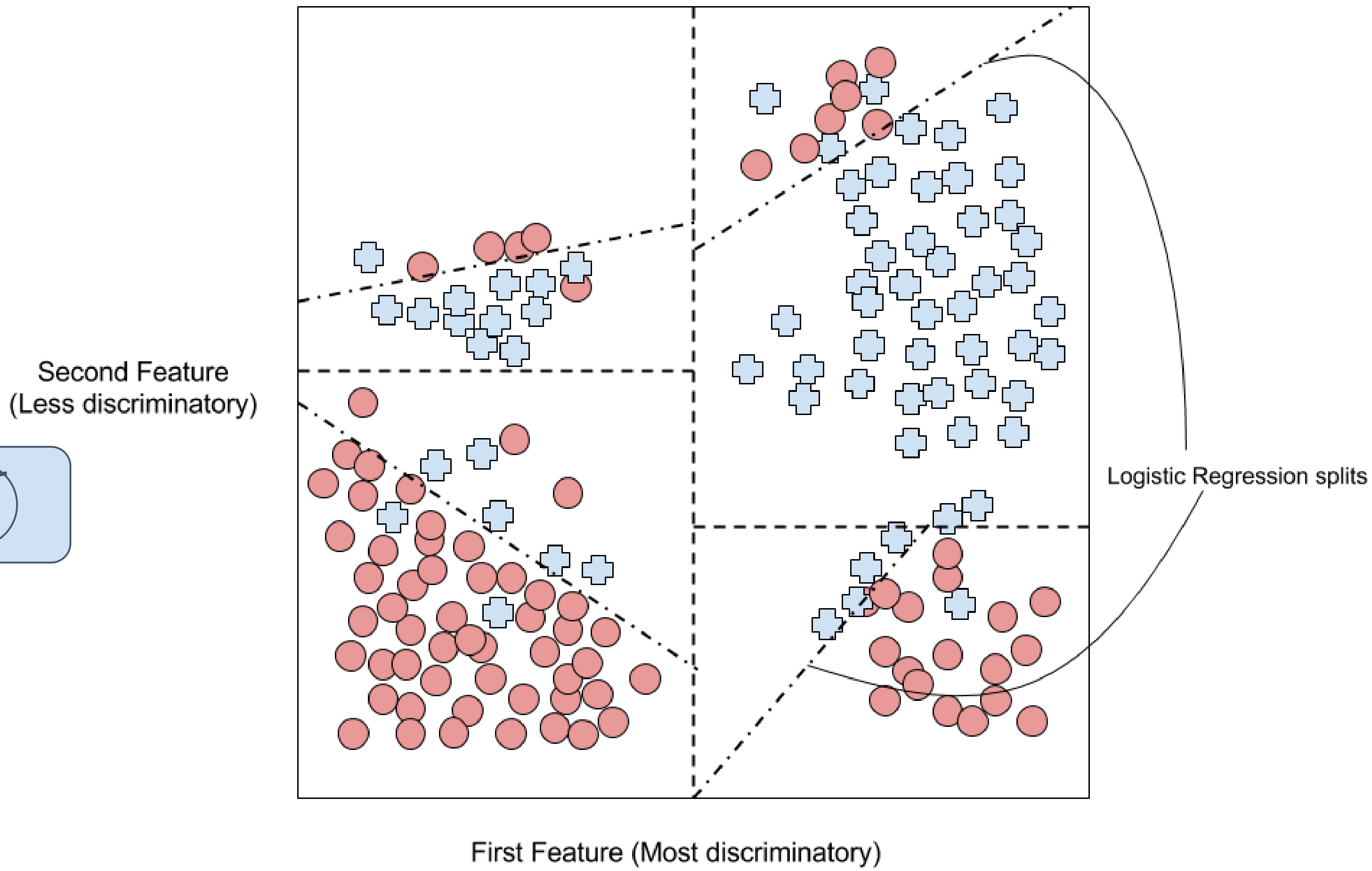
LOGISTIC MODEL TREES



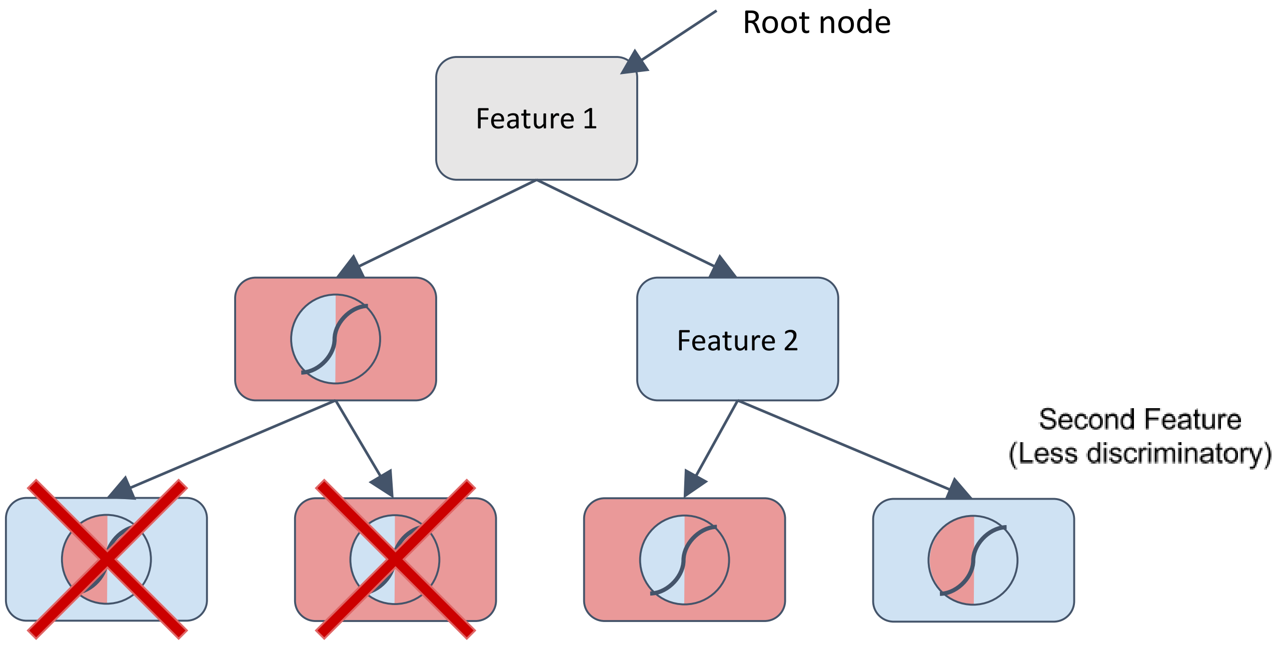


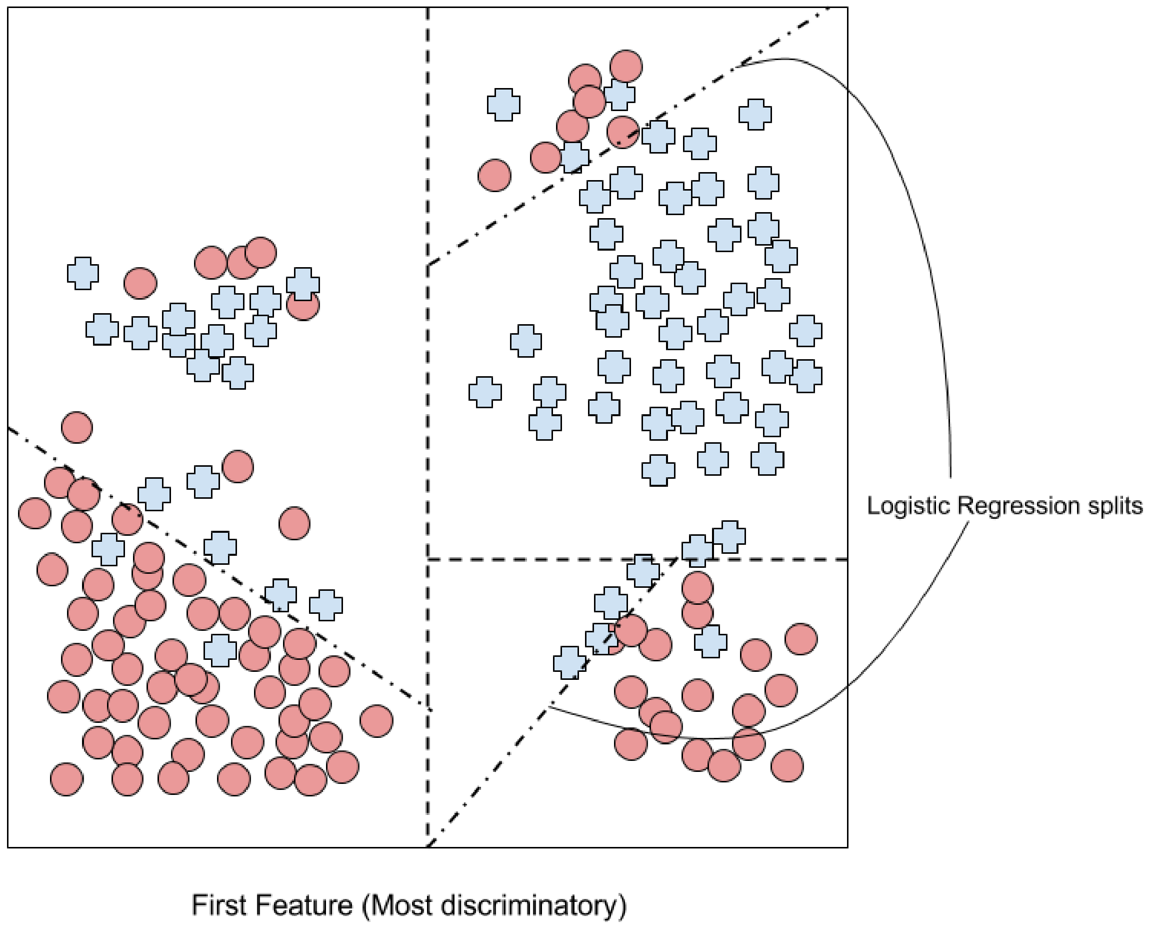
LOGISTIC MODEL TREES



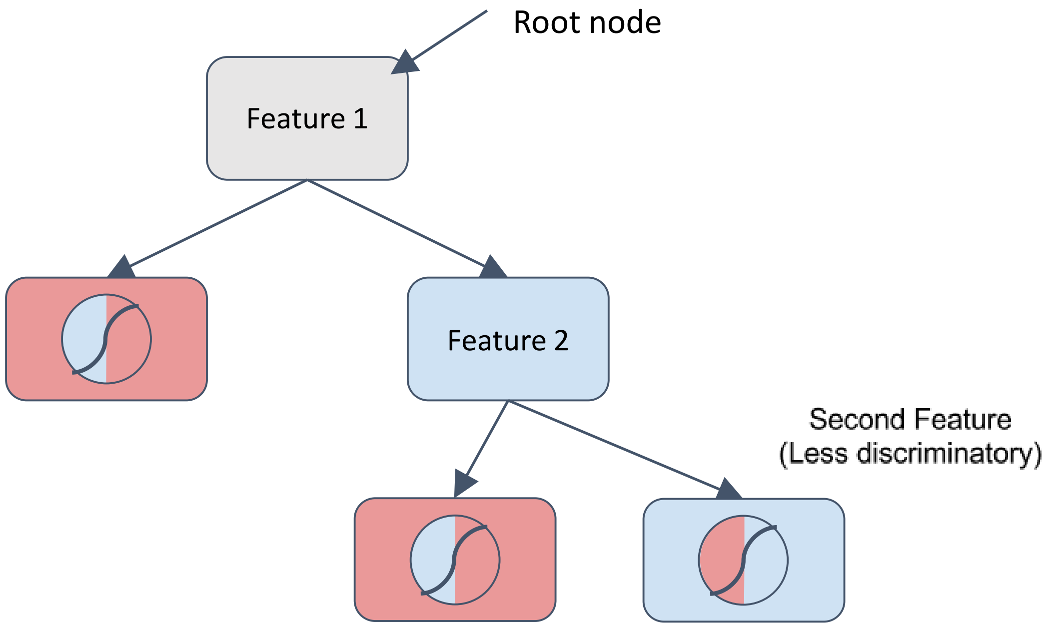


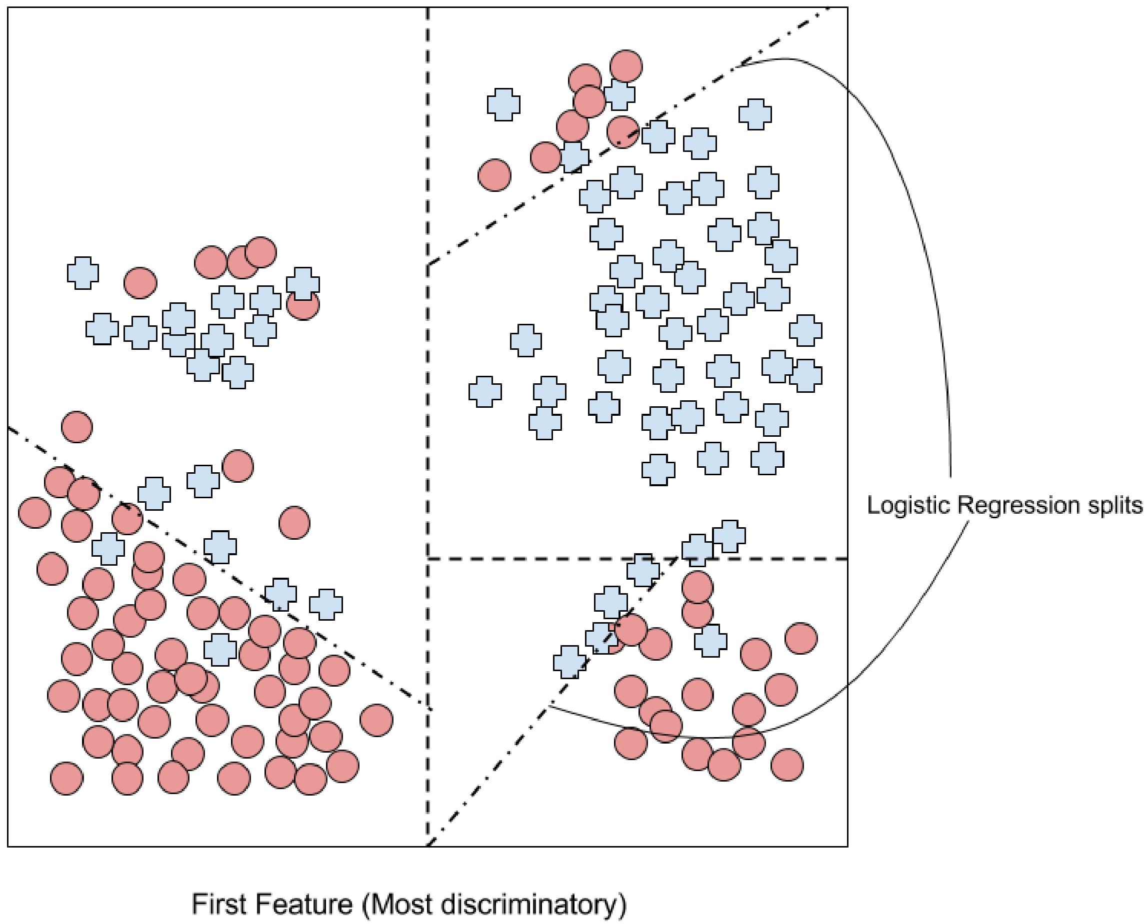
LOGISTIC MODEL TREES



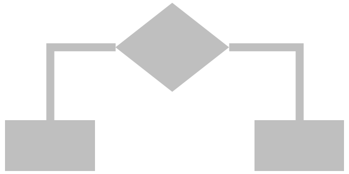


LOGISTIC MODEL TREES





DISTRIBUTED IMPLEMENTATION



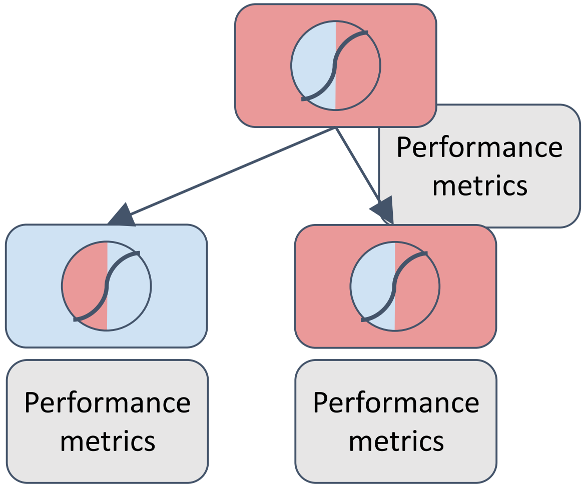
Spark's Decision Tree
(distributed implementation of random forests)

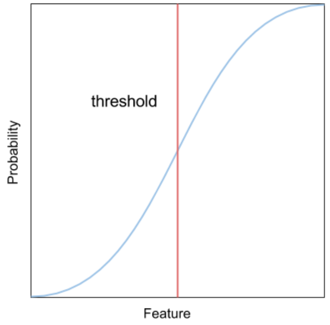
Spark's Logistic Regression / weka's Logistic Regression on the nodes

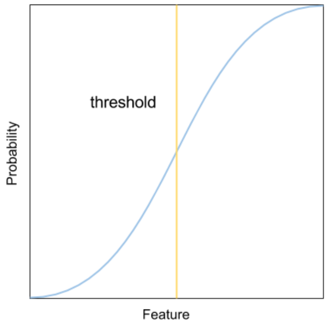
LMT Cost function to fix the logistic regression threshold

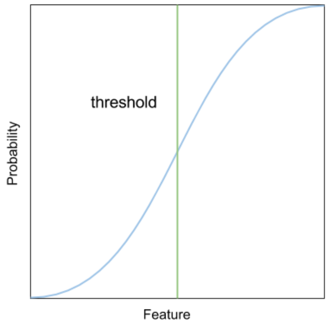
- AccuracyCostFunction
- ConfusionMatrix
- PrecisionCostFunction
- PrecisionRecallCostFunction
- RocCostFunction

The same cost function for pruning criteria







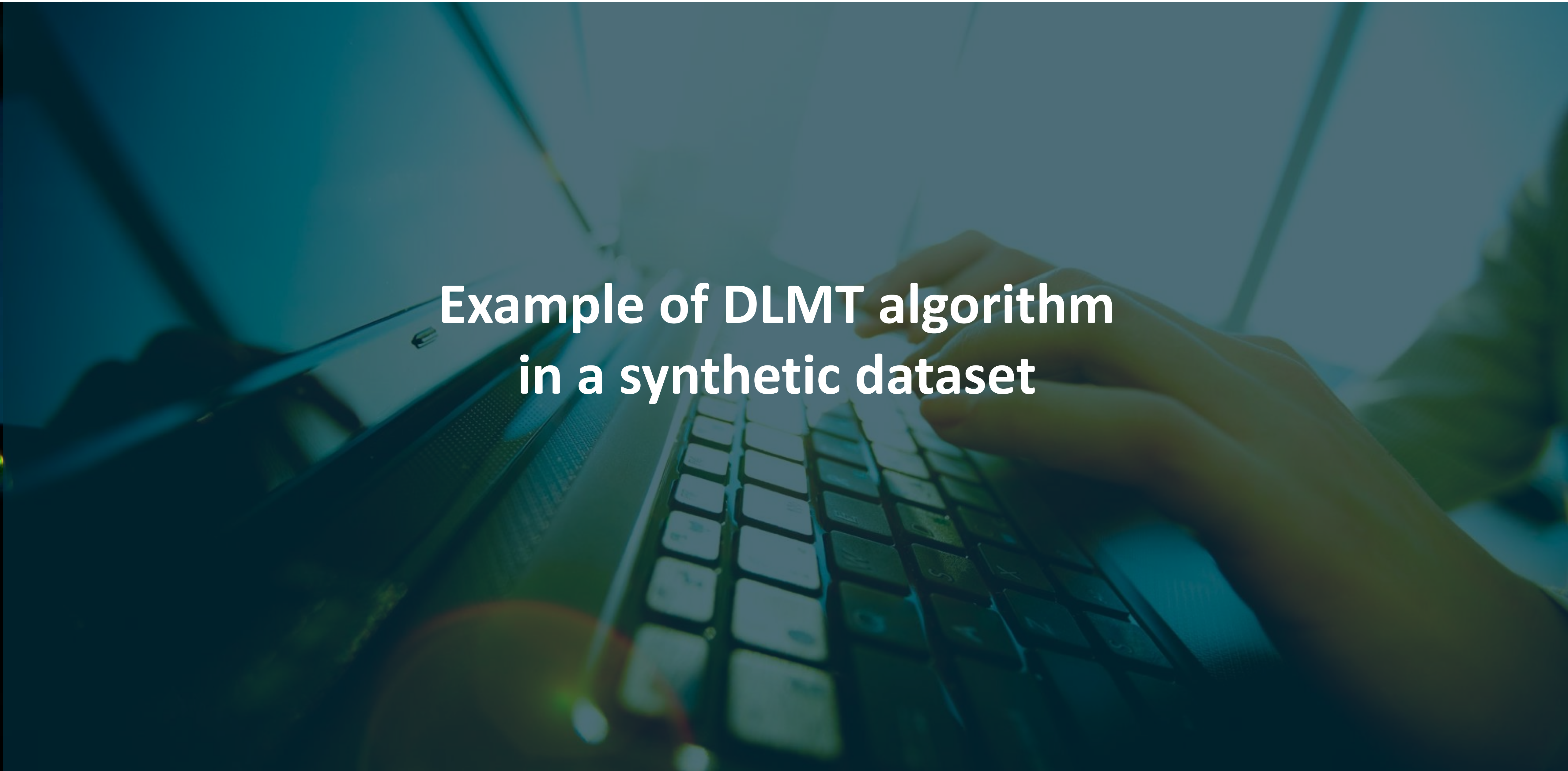


ADVANTAGES OF THIS IMPLEMENTATION

Big datasets
Power of spark to distribute building the tree and logistic regressions

Medium datasets
Distributed tree growth and weka's logistic regression

Small datasets
Although it can be slow to distribute the data for the decision tree, cost functions can be still used and specific optimization for particular cases



Example of DLMT algorithm in a synthetic dataset

AUTOMATED BENCHMARKING FRAMEWORK

- Metrics
- Demo

@StratioBD

METRICS

		PREDICTION		
		Positive	Negative	
TRUE CONDITION	Positive	True Positives	False Negatives	**True Positive Rate** (Recall)
	Negative	False Positives	True Negatives	**False Positive Rate**

Precission

TPR = TP/(TP+FN) Insensitive to unbalance

FPR = FP/(FP+TN) Insensitive to unbalance

Precision = TP/(TP+FP) Sensitive to unbalance

Accuracy = (TP+TN)/(TP+TN+FP+FN) Sensitive to unbalance

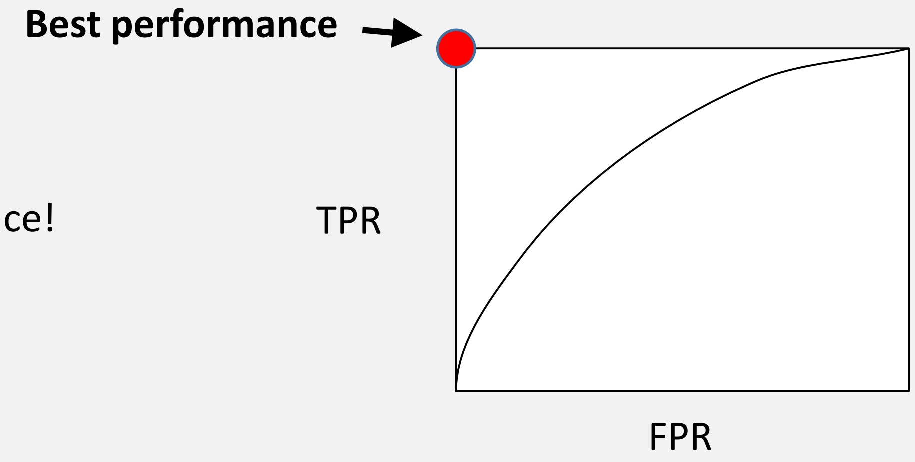
METRICS

		PREDICTION	
		Positive	Negative
TRUE CONDITION	Positive	True Positives	False Negatives
	Negative	False Positives	True Negatives

→ **True Positive Rate** (Recall)

→ **False Positive Rate**

AUROC (AUC): TPR/FPR -> Insensitive to unbalance!



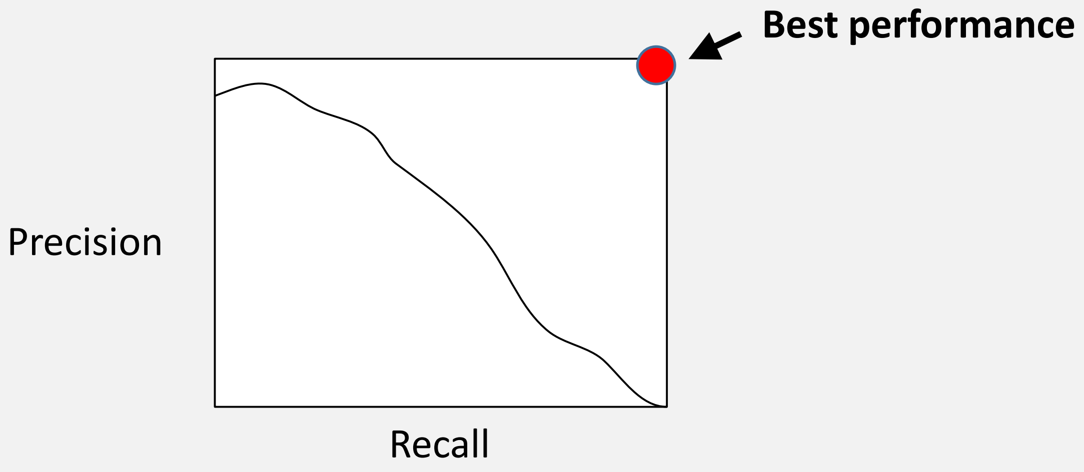
METRICS

		PREDICTION	
		Positive	Negative
TRUE CONDITION	Positive	True Positives	False Negatives
	Negative	False Positives	True Negatives

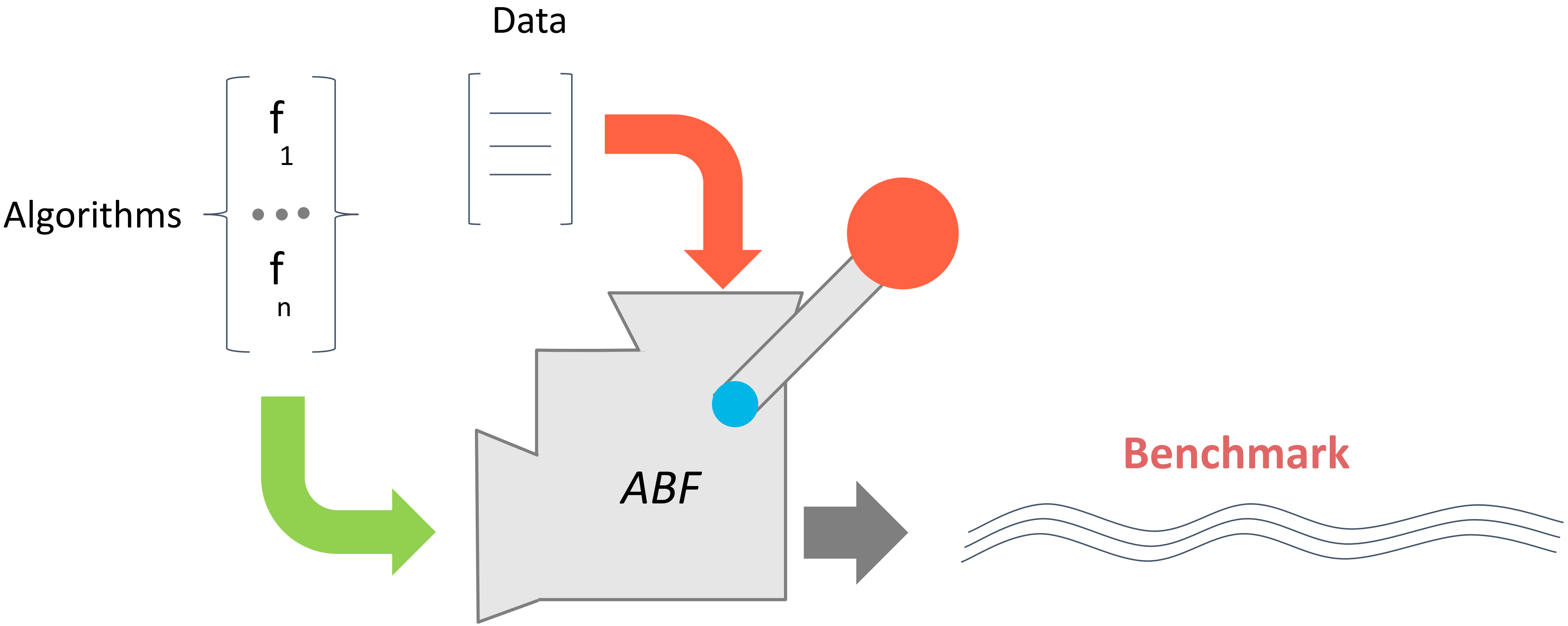
True Positive Rate (Recall)

Precission

AUPRC: Precision/TPR -> Sensitive to unbalance!



DEMO



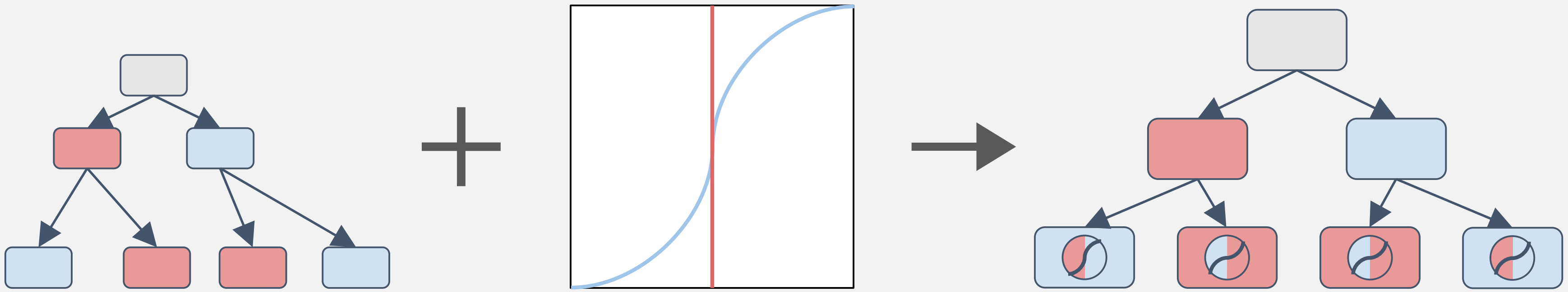
BENCHMARKING RESULTS

@StratioBD

SUMMARY

1

2

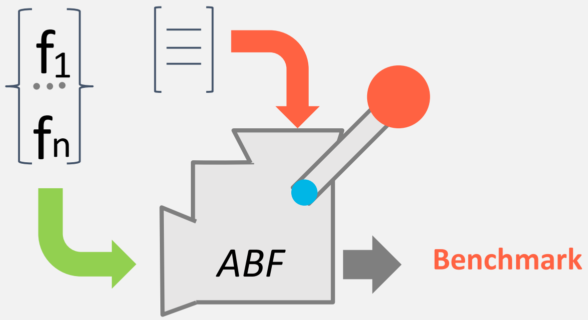


3

Performance Metrics:
AUROC, AUPRC, ACCURACY

4

Automatic Benchmarking Framework



THANK YOU

UNITED STATES
Tel: (+1) 408 5998830

EUROPE
Tel: (+34) 91 828 64 73

contact@stratio.com

www.stratio.com

@StratioBD

STRATIO®

WE ARE HIRING

people@stratio.com

@StratioBD