Random Forest and Applications

Sambit Panda

Question: What to do with all this data?

Mouse	Group	Locomotor Activity	Grip Strength
10	WT	24	0.1
7.2	Mutant	18	0.12
34	Mutant	17	0.22
20	WT	33	0.05
8.8	Mutant	15	0.3

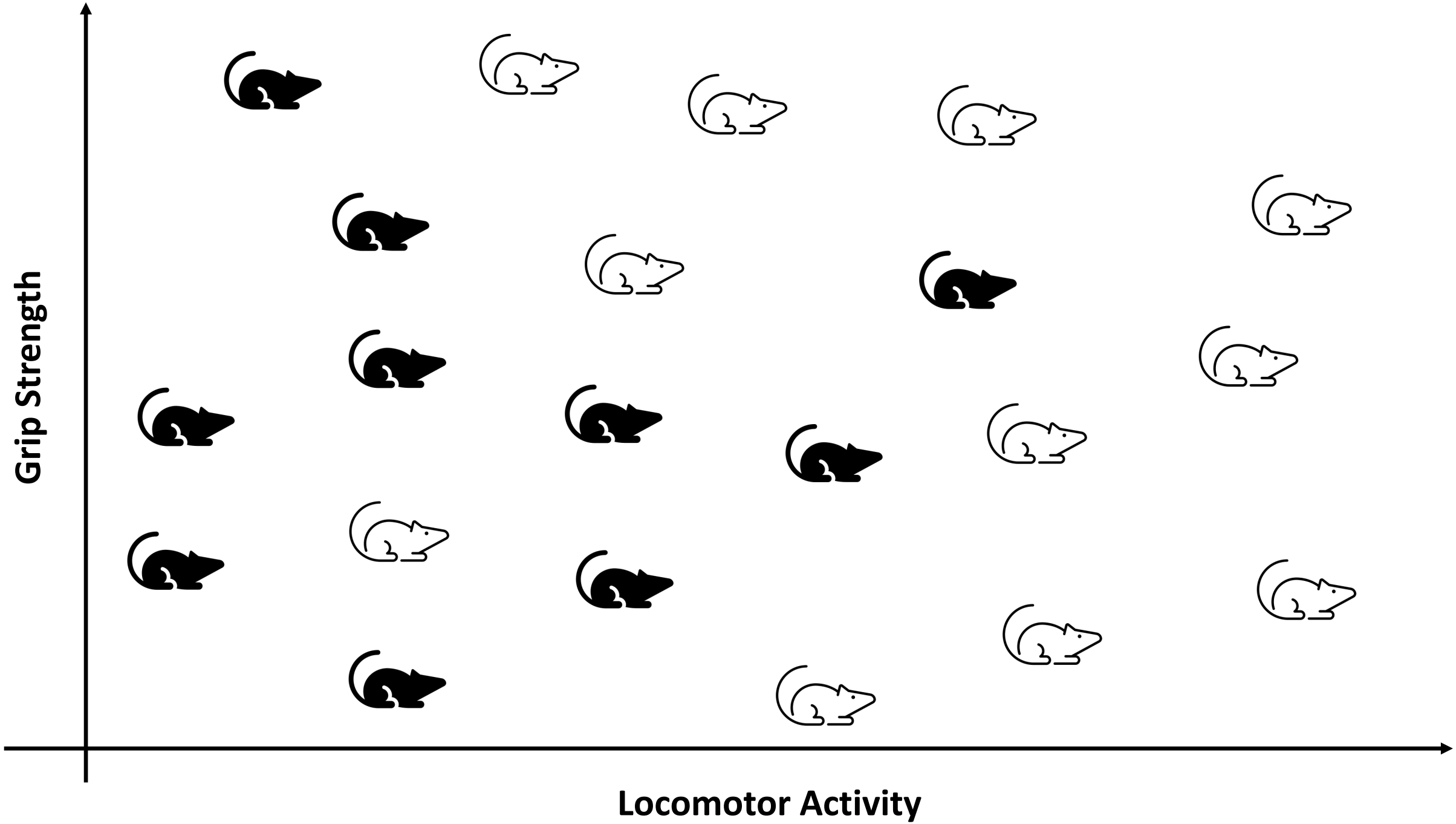
Mutant

WT

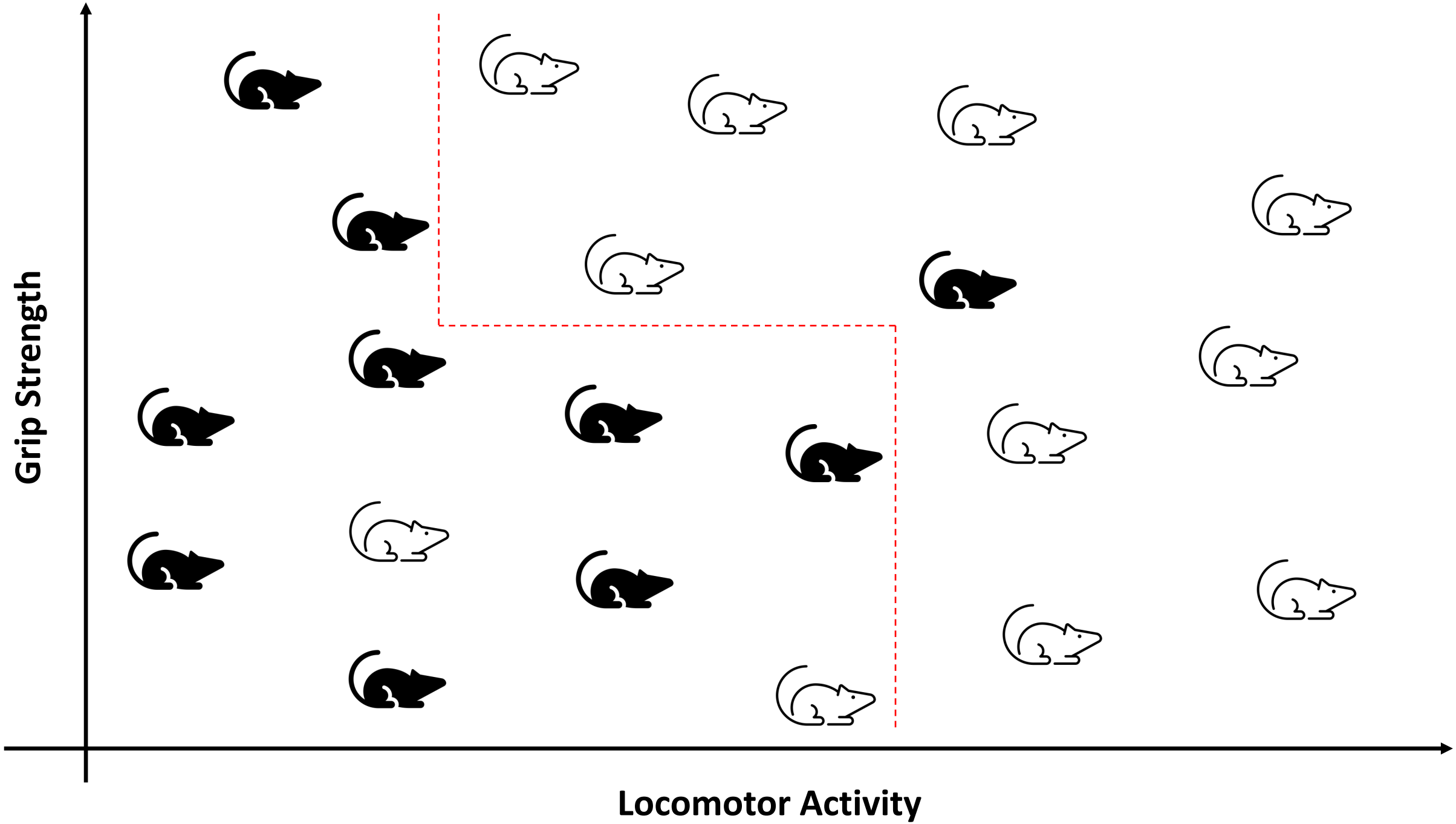
Columns = Features = Dimensions (p or d)

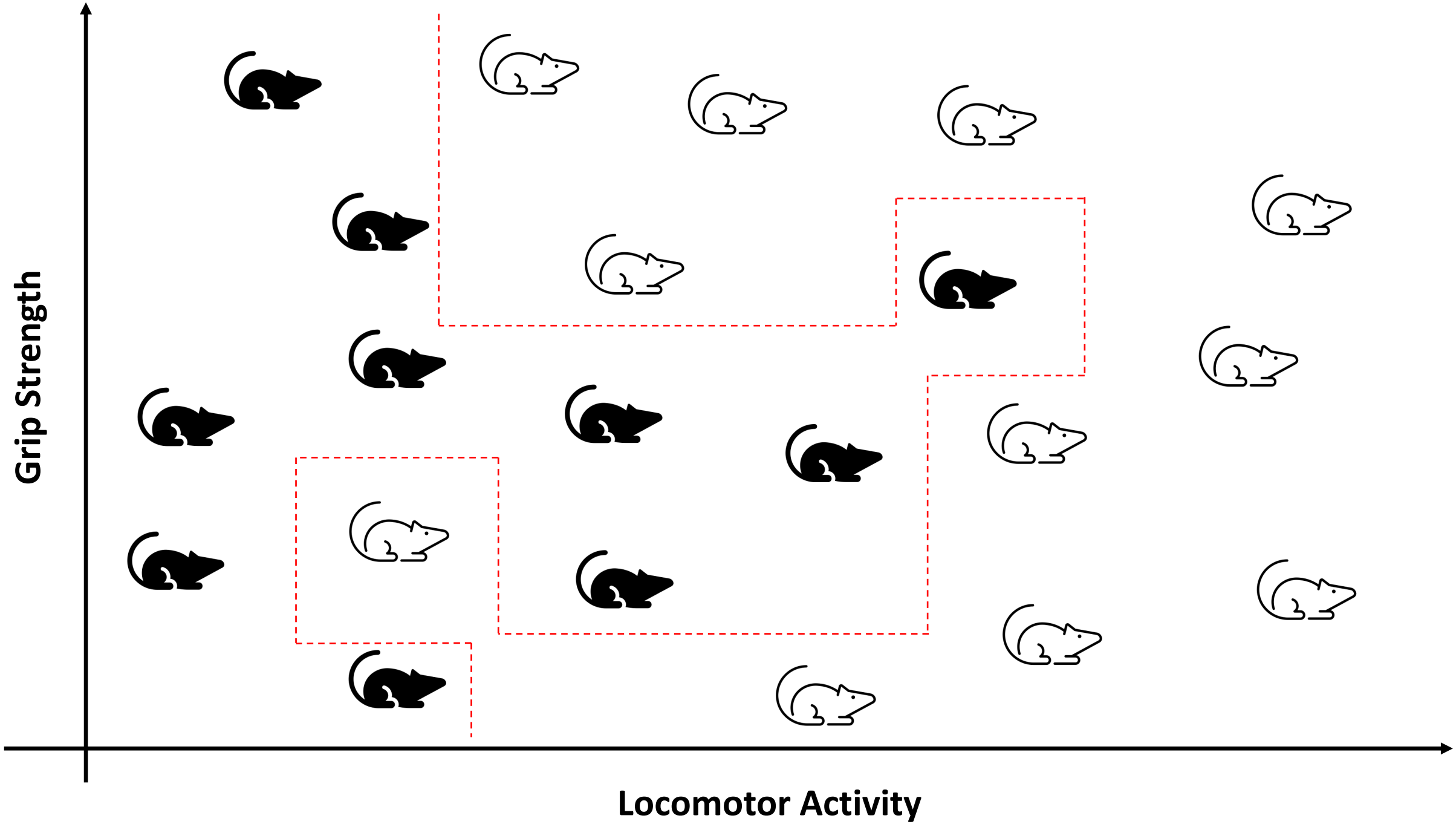
Rows = Subjects (n)

Mouse	Group	Locomotor Activity	Grip Strength
10	WT	24	0.1
7.2	Mutant	18	0.12
34	Mutant	17	0.22
20	WT	33	0.05
8.8	Mutant	15	0.3

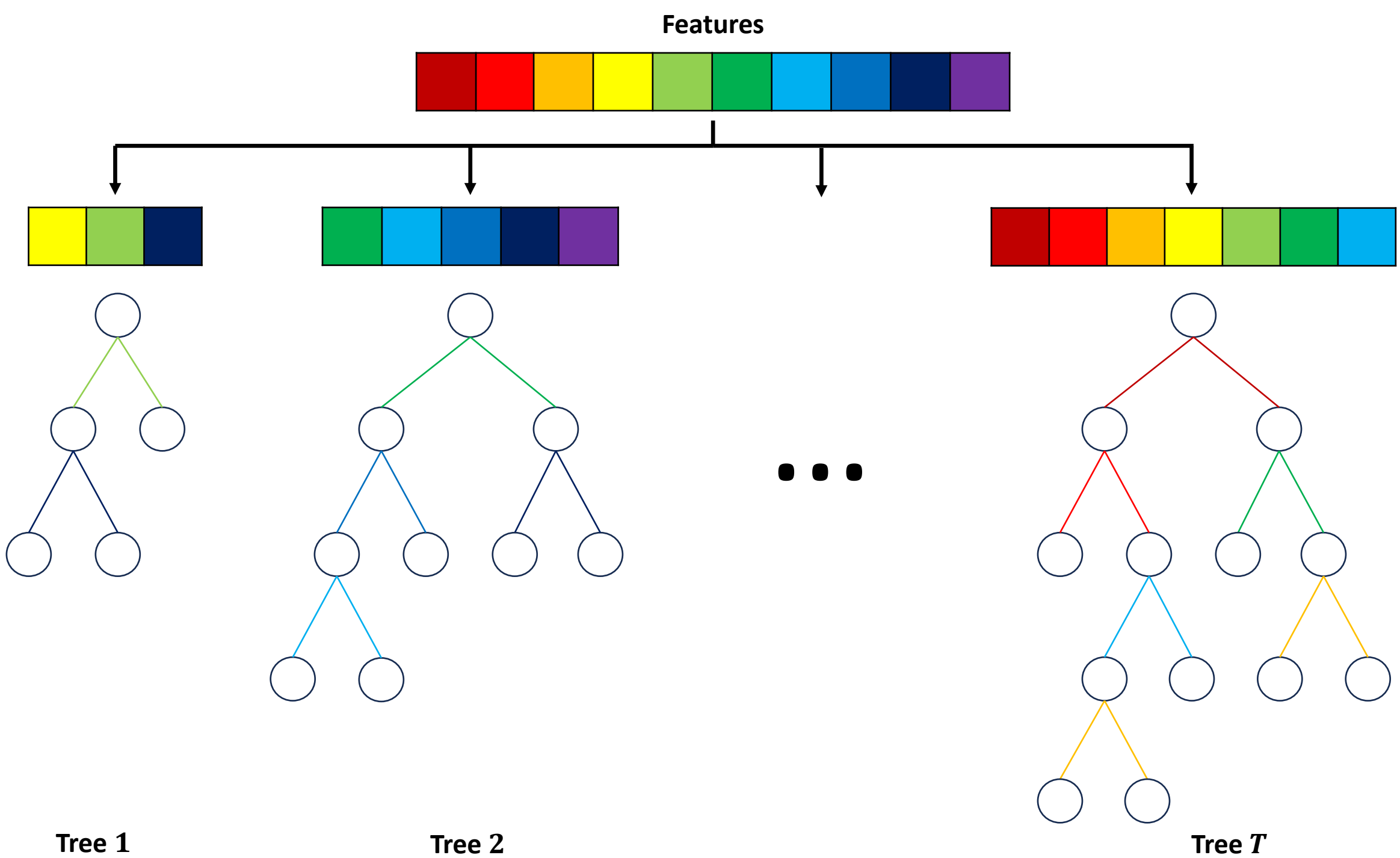

Grip Strength
Locomotor Activity

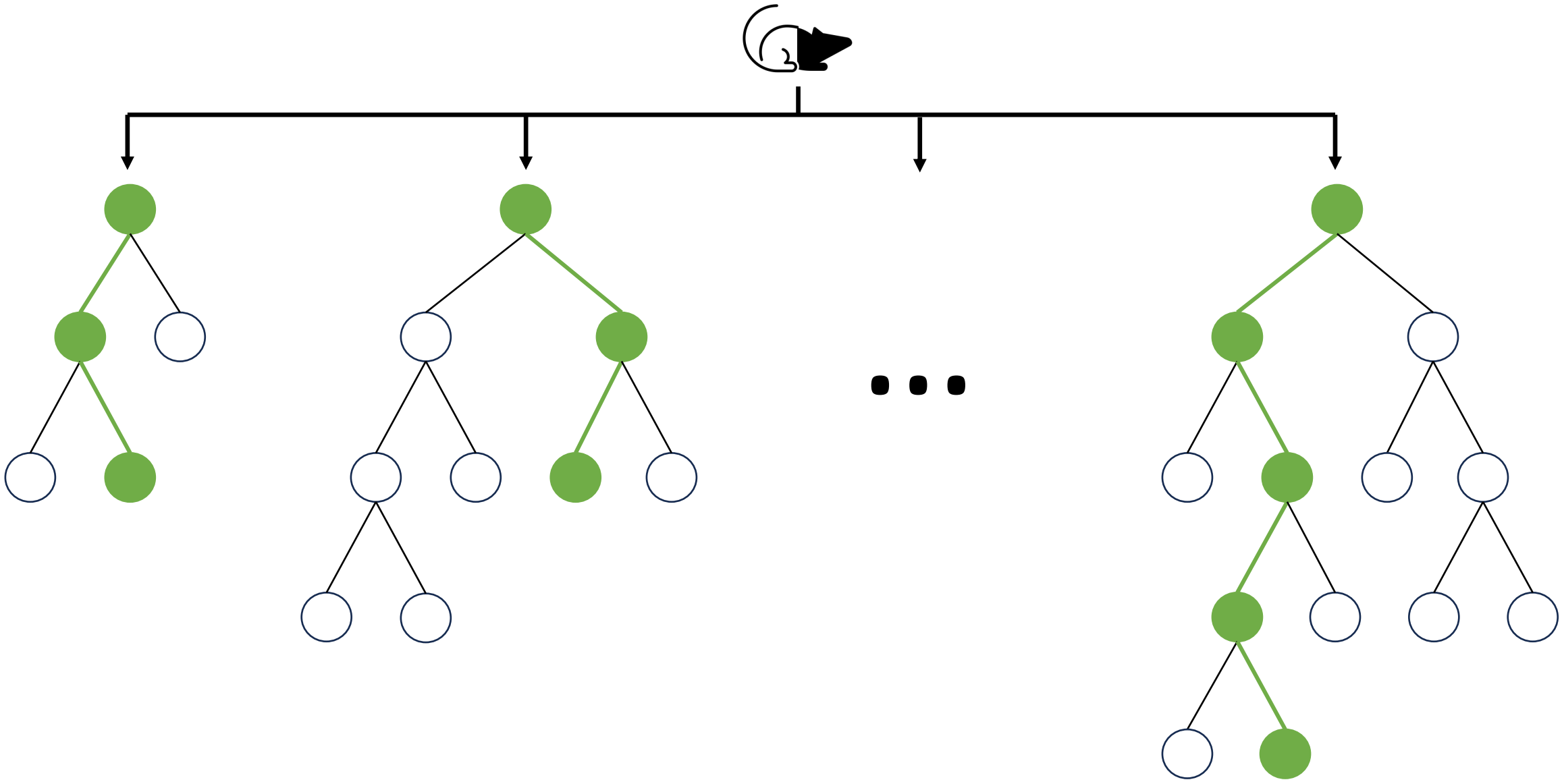
How Random Forest Works


Grip Strength
Locomotor Activity


Grip Strength
Locomotor Activity

Mouse	Group	Locomotor Activity	Grip Strength
10	WT	24	0.1
7.2	Mutant	18	0.12
34	Mutant	17	0.22
20	WT	33	0.05
8.8	Mutant	15	0.3


Features
Tree 1
Tree 2
Tree T

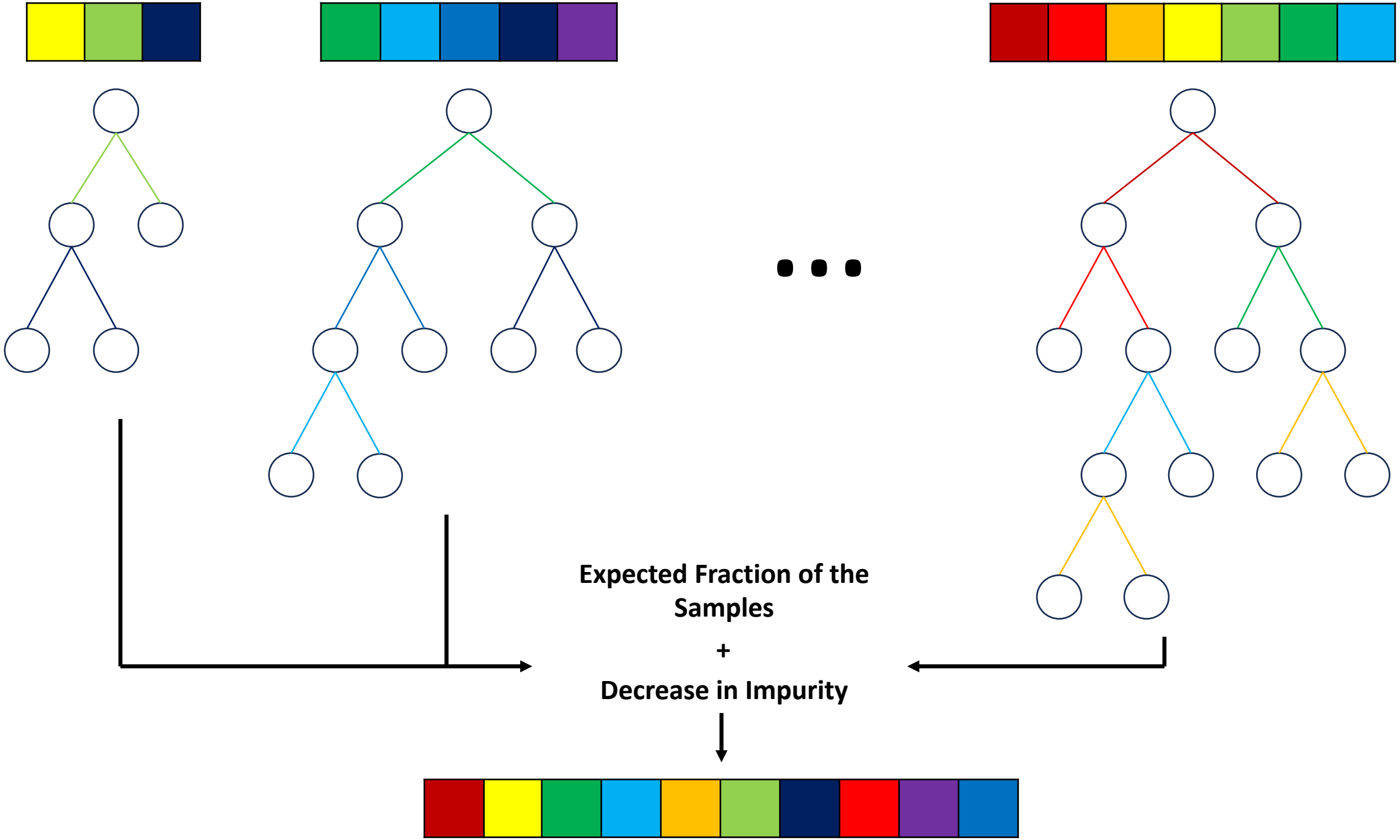


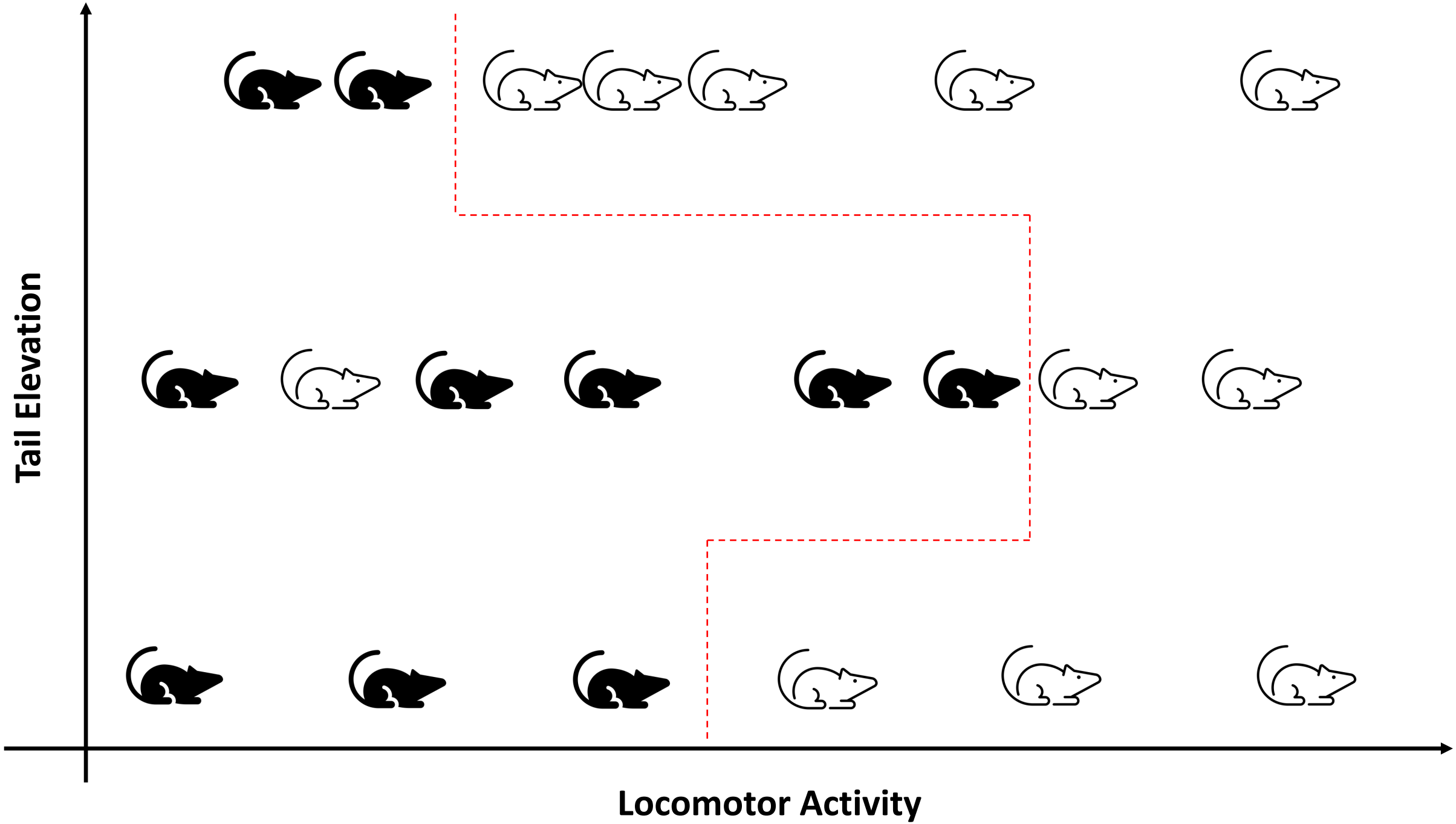
WT
$p_1 = 0.55$

Mutant
$p_2 = 0.76$

Mutant
$p_T = 0.82$

Majority: Mutant
$p = \text{Average } p_i$


Expected Fraction of the Samples
+
Decrease in Impurity


Tail Elevation
Locomotor Activity

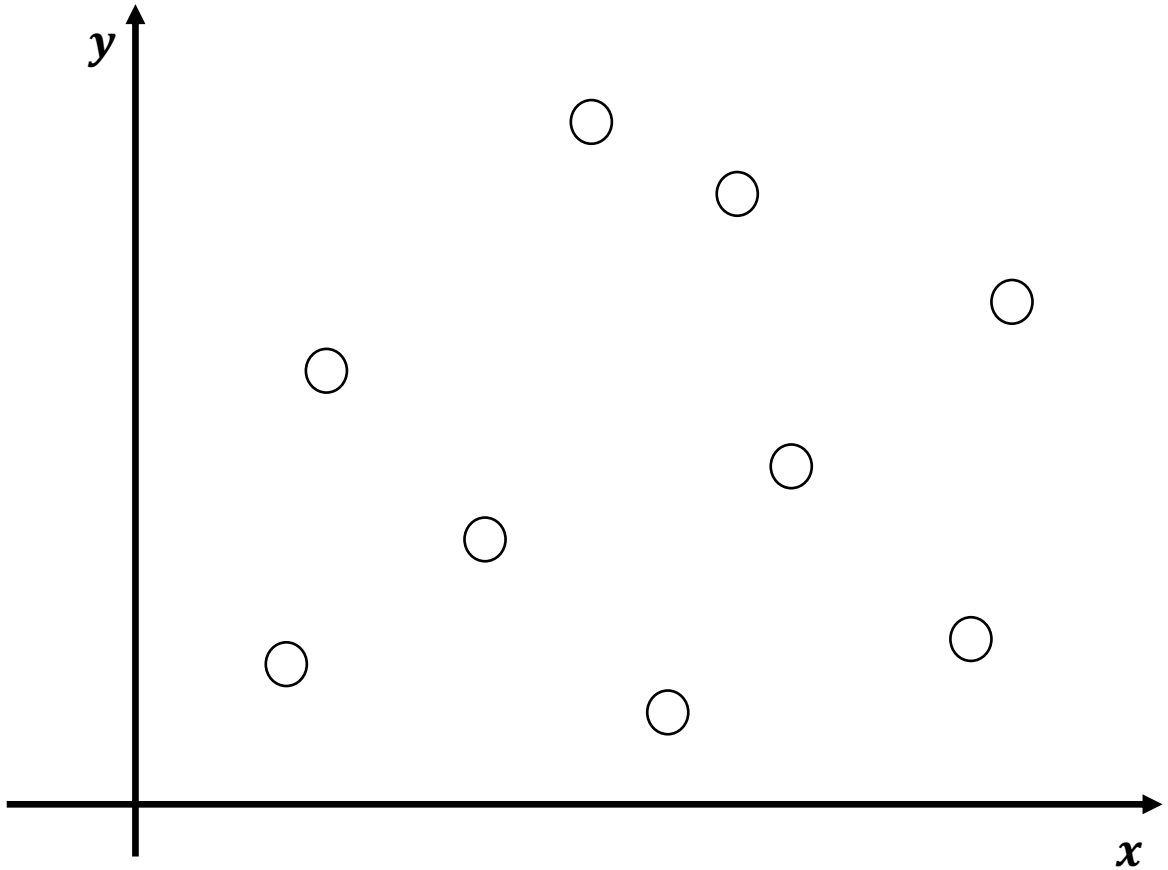
Summary

- Learned what tabular data is
- What splits are geometrically and within the algorithm
- How random forests trains on current data and tests on new points
- How random forest estimates feature importance

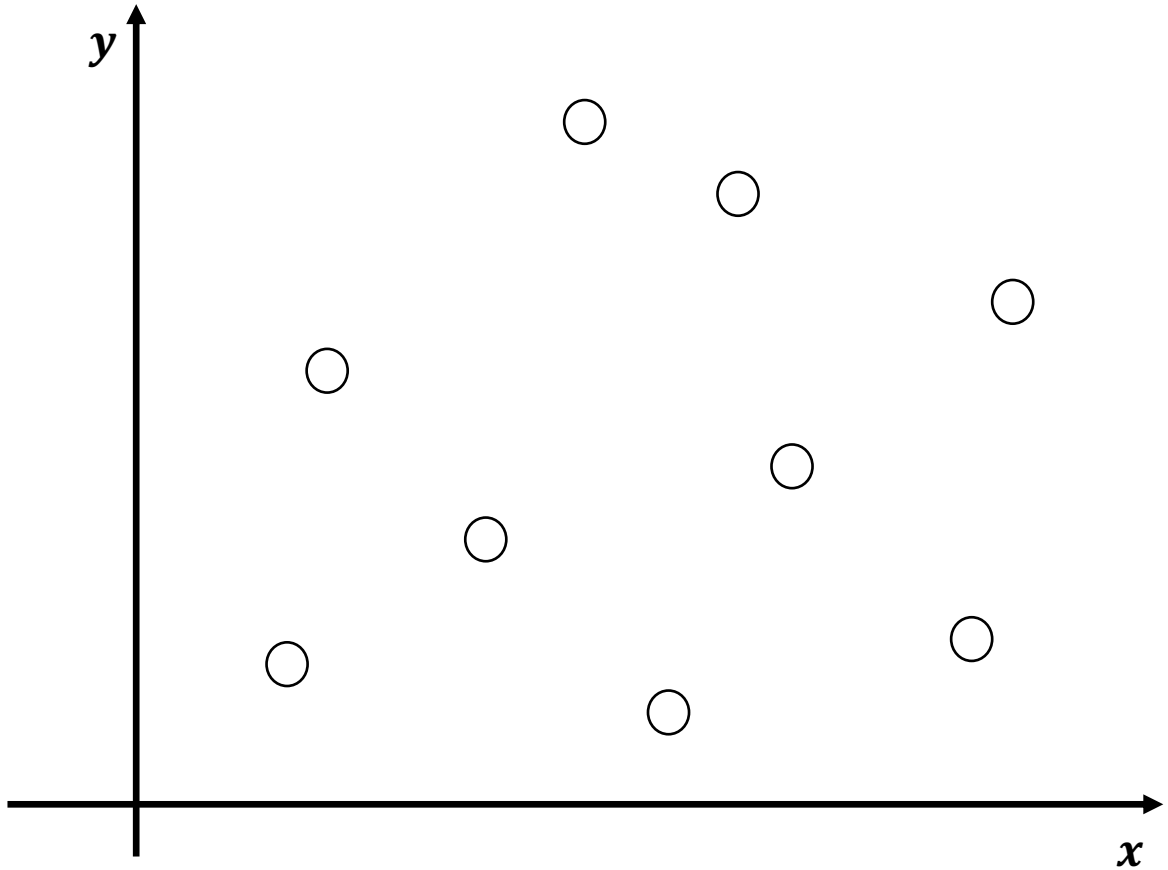
Independence Testing, K-Sample Testing and Random Forest

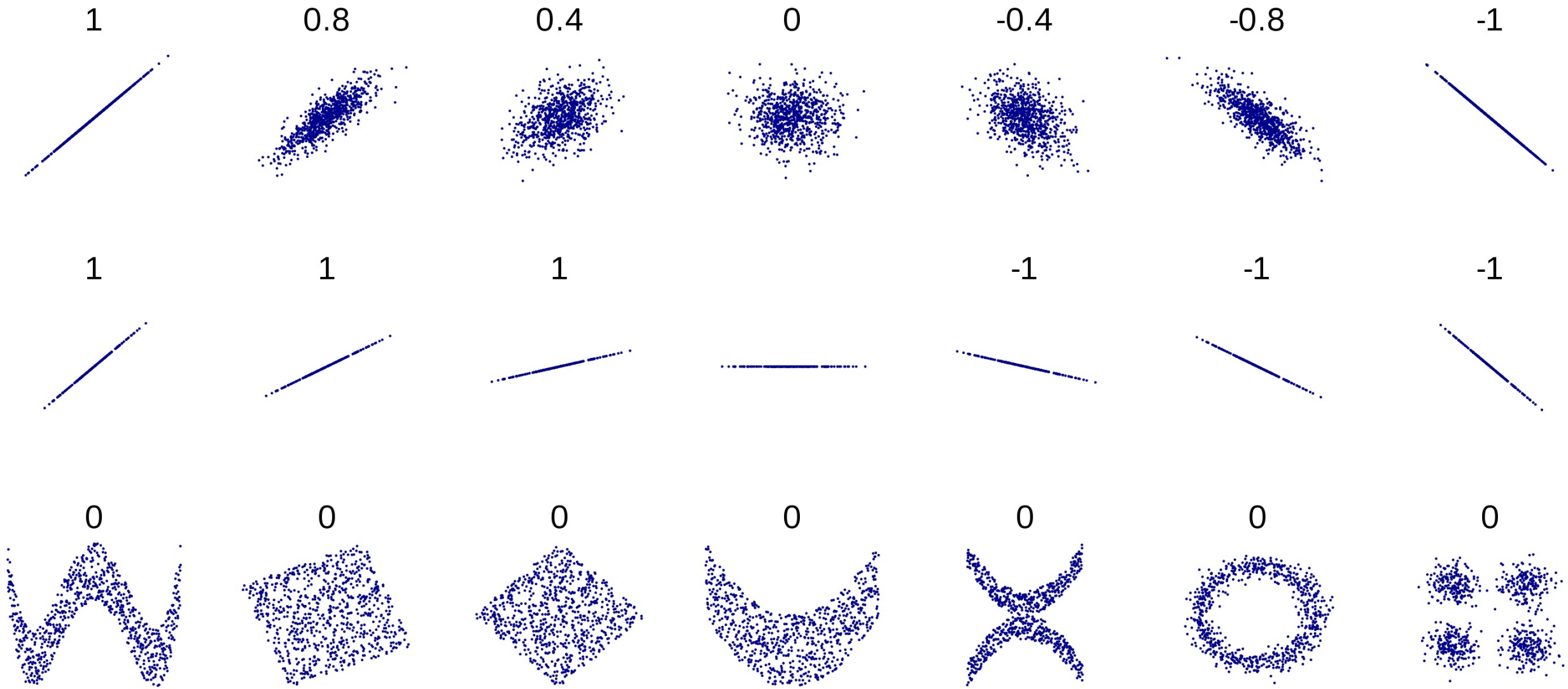
Question: Is there a relationship?

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0$$

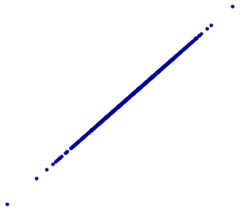


$$H_0 : F_{XY} = F_X F_Y$$
$$H_A : F_{XY} \neq F_X F_Y$$

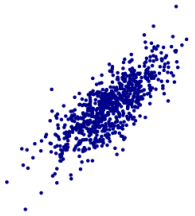



1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

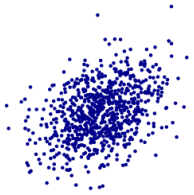
1



0.7

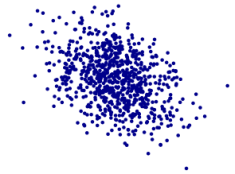


0.3

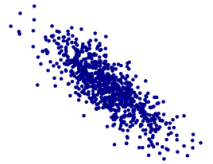


0.1

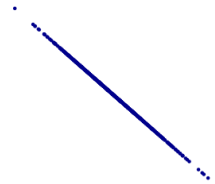
0.3



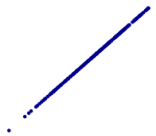
0.8



1



1



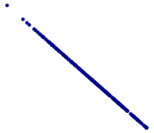
1

1

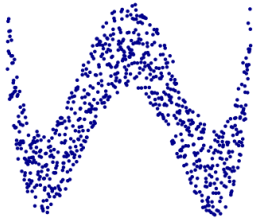
1

1

1



0.3



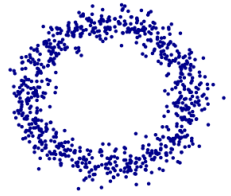
0.1

0.1

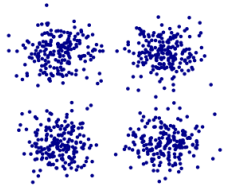
0.3

0.2

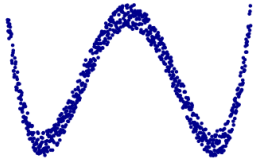
0.2



0

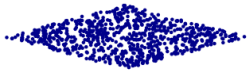


0.4



0.2

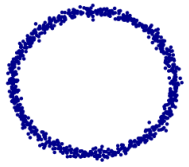
0.2



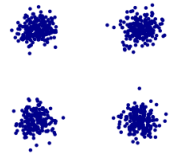
0.5

0.3

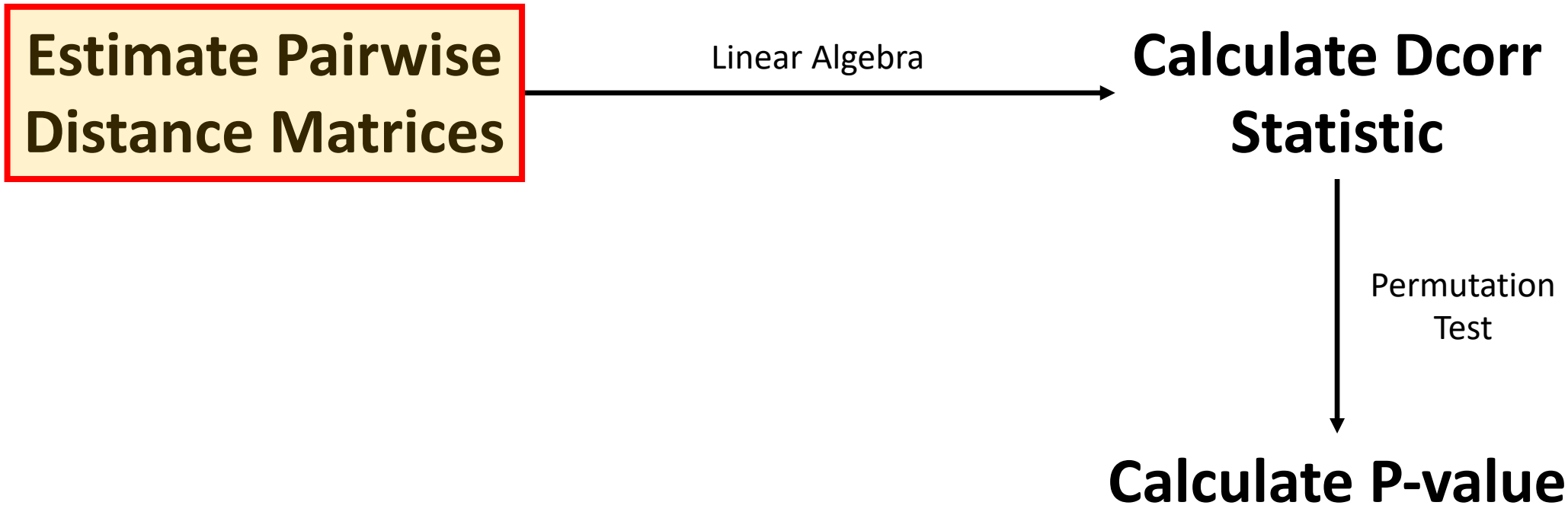
0.2

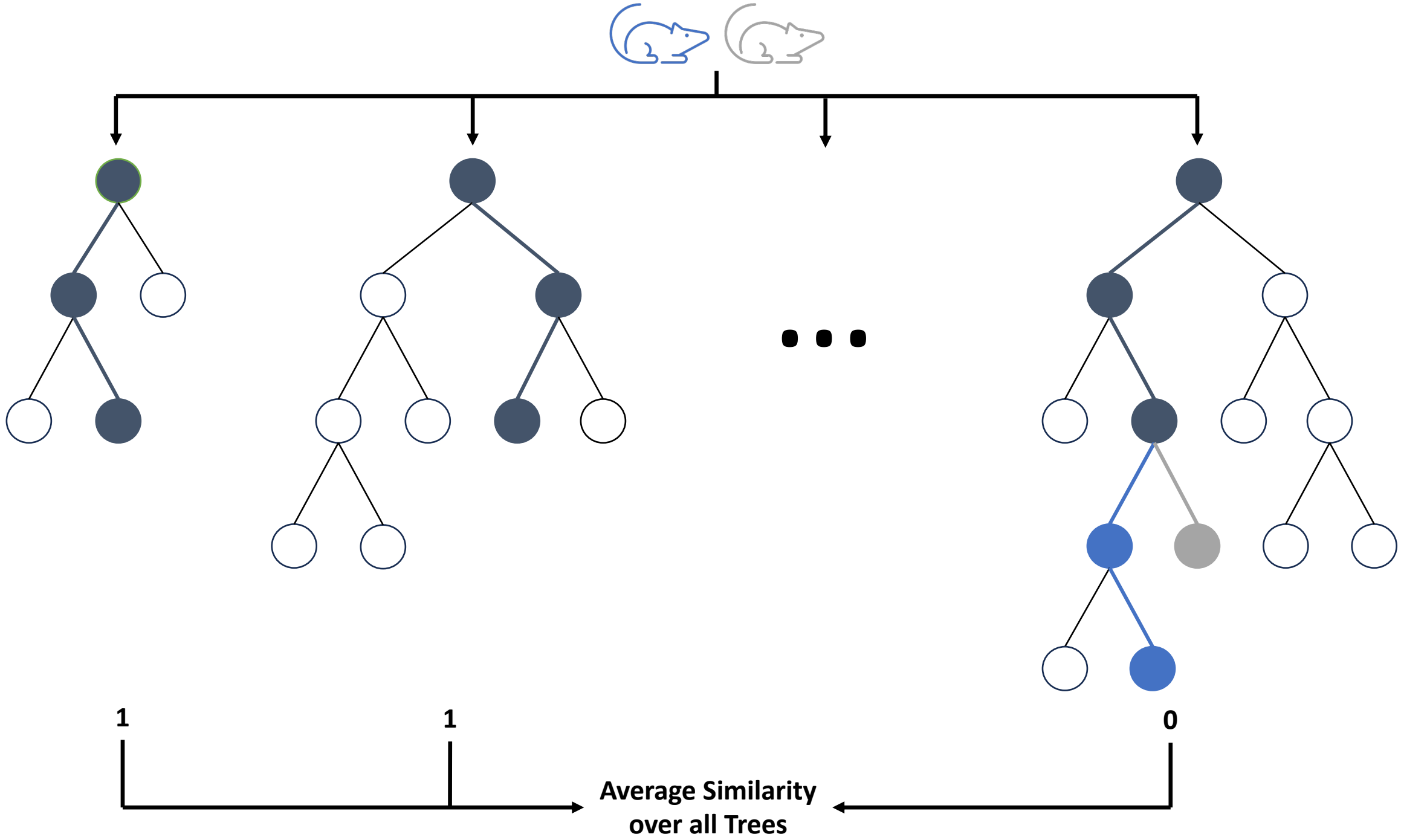


0

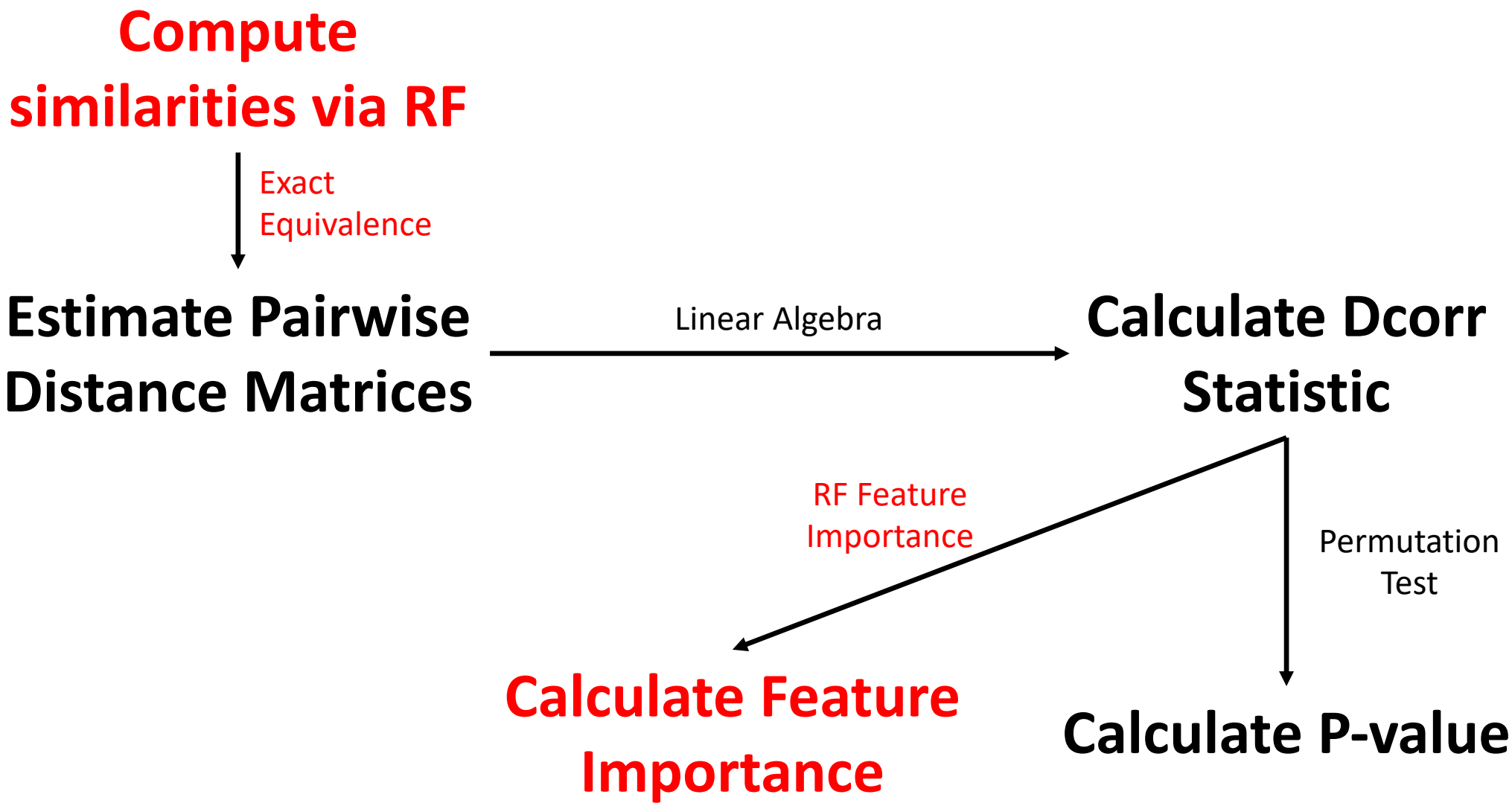


How to Compute Dcorr




1
1
0
Average Similarity
over all Trees

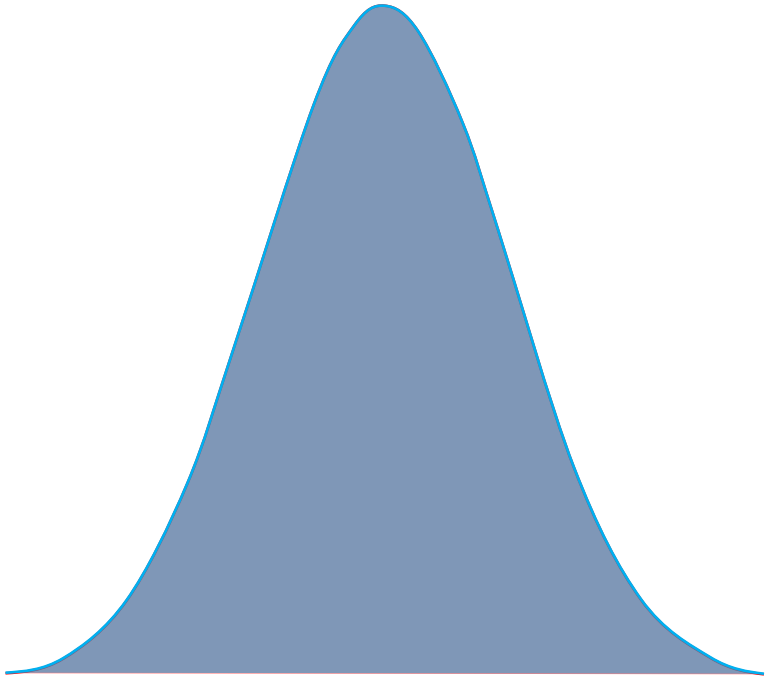
How to Compute KMERF

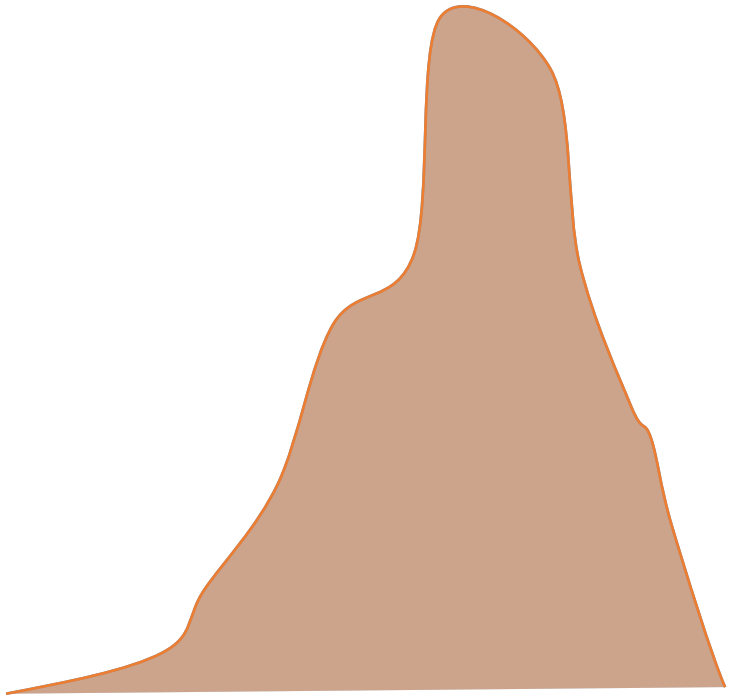


Question: Are they different?

$$H_0 : \mu_X = \mu_Y$$
$$H_A : \mu_X \neq \mu_Y$$

$$H_0 : F_X = F_Y$$
$$H_A : F_X \neq F_Y$$





Question: Is there a relationship?

=

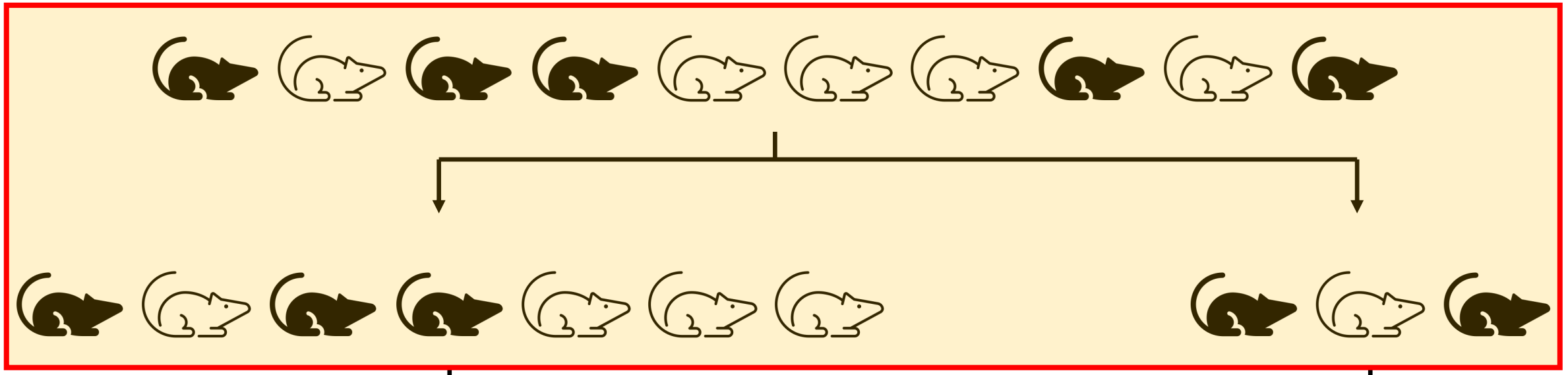
Question: Are they different?

(via a transformation of the data)

Summary

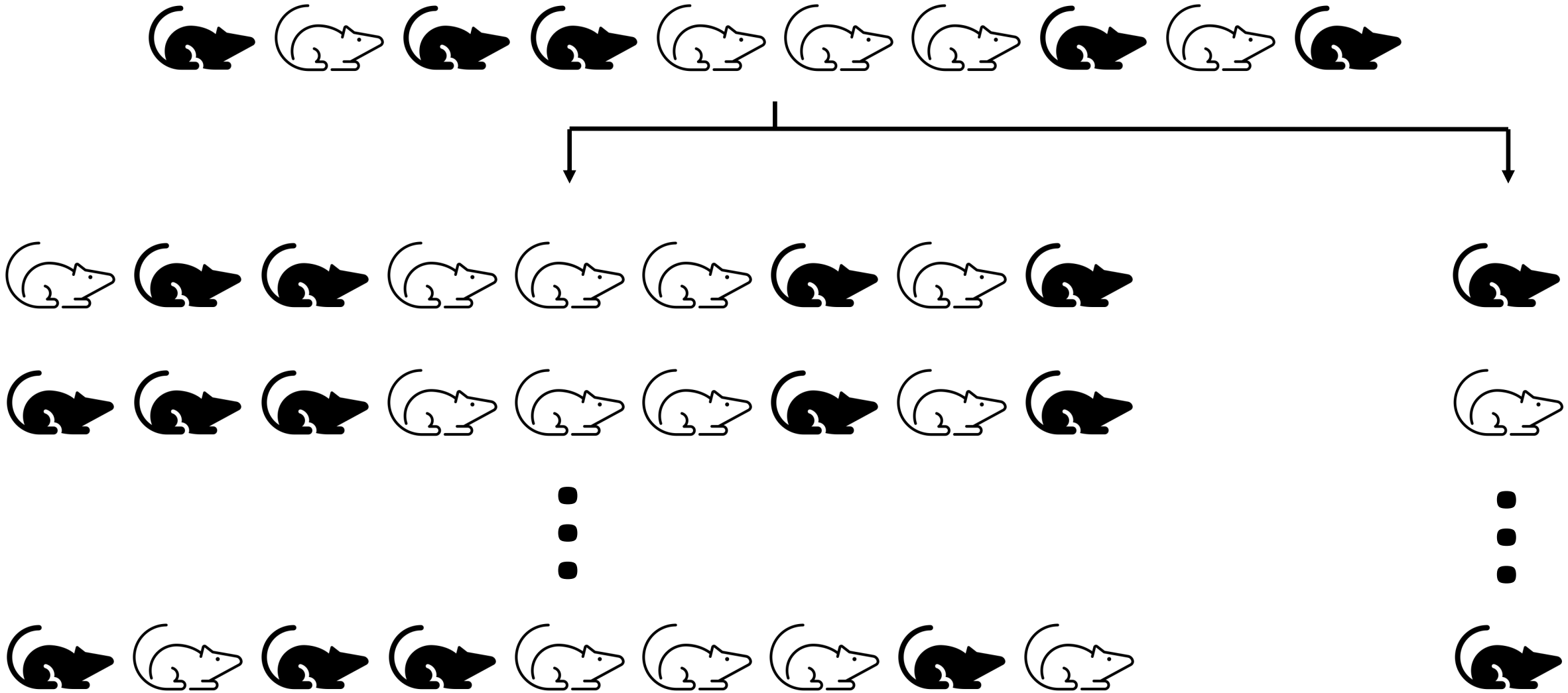
- Learned what independence and k-sample testing are
- Learned how Dcorr and KMERF works
- Found out what results are on our data

Methods to Evaluate Performance

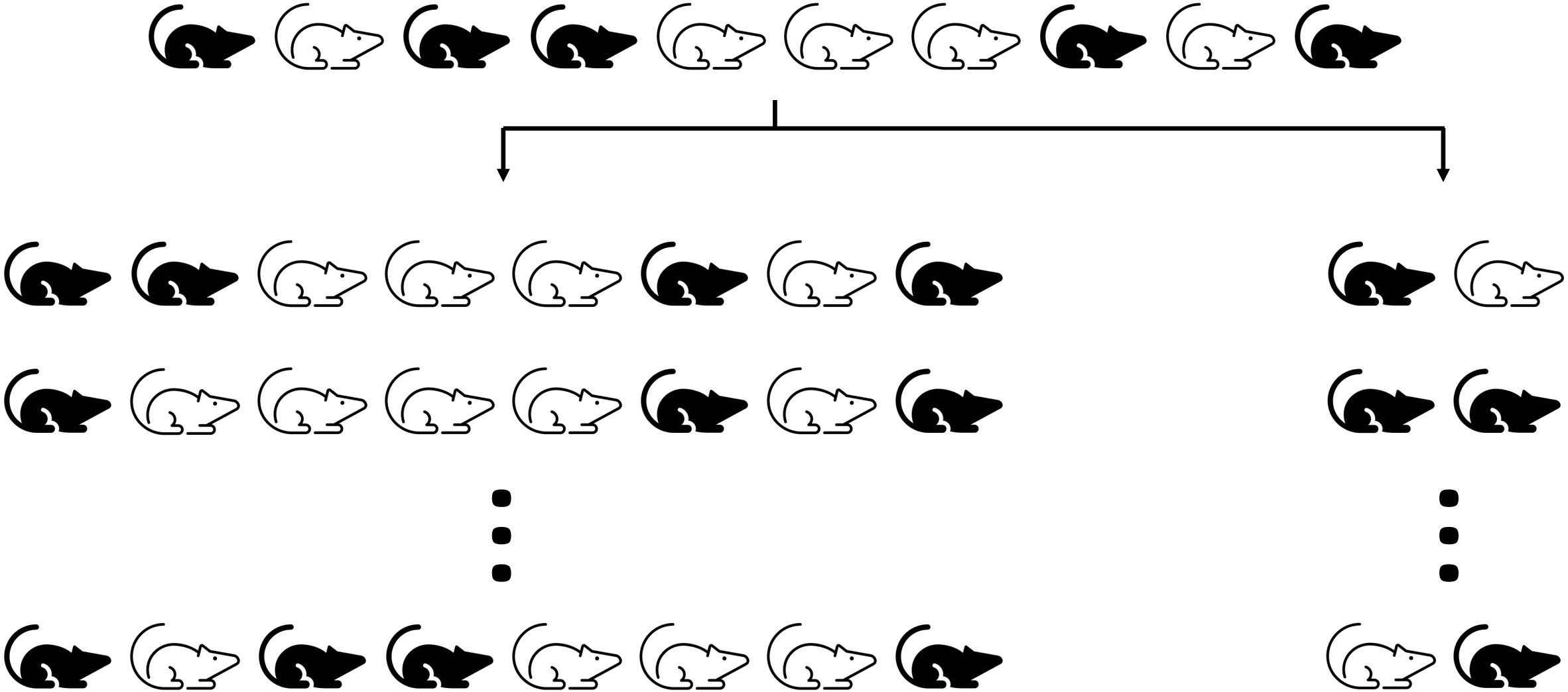


Train Classifier → Test Classifier

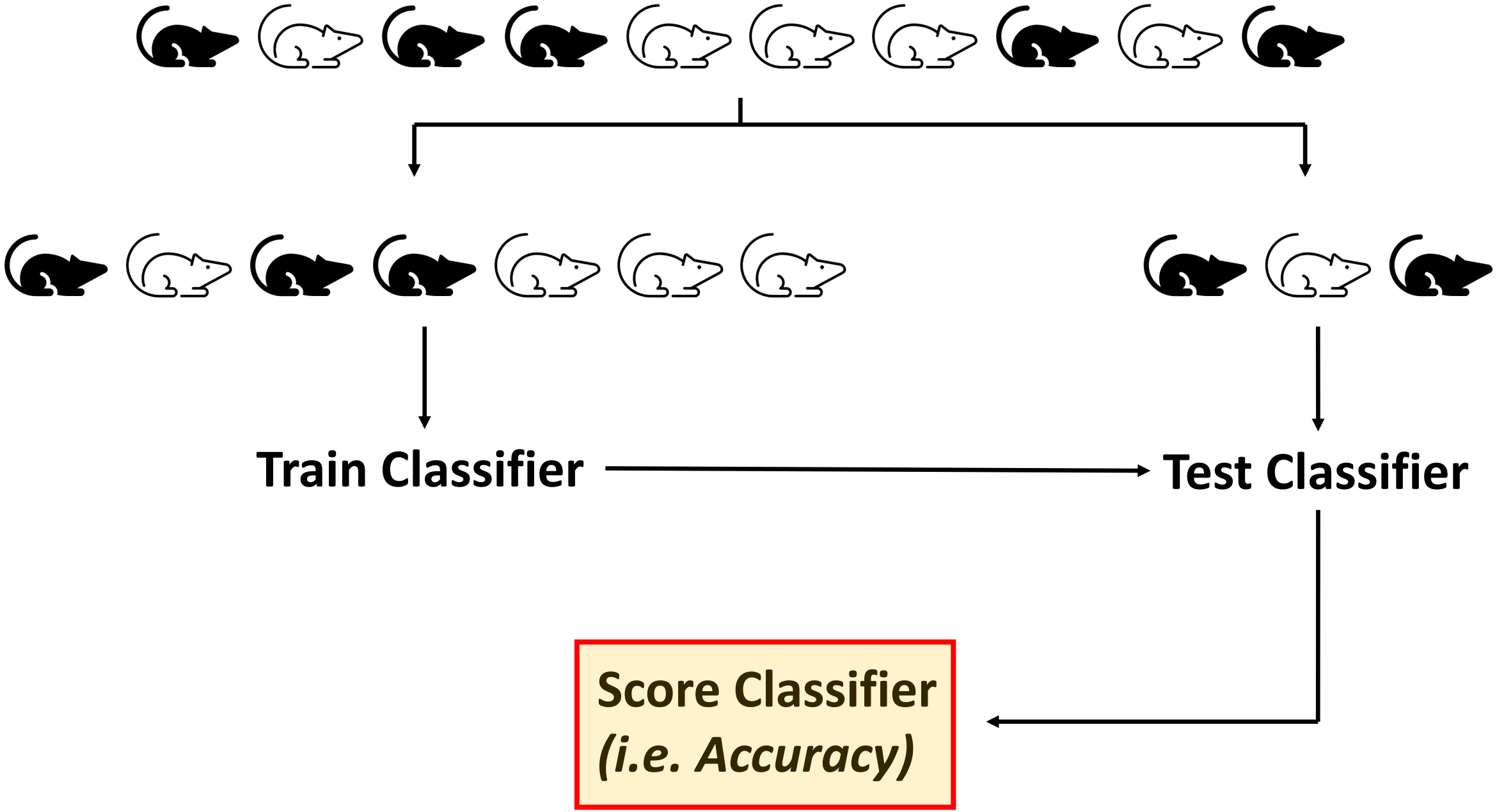
Score Classifier
(i.e. Accuracy)

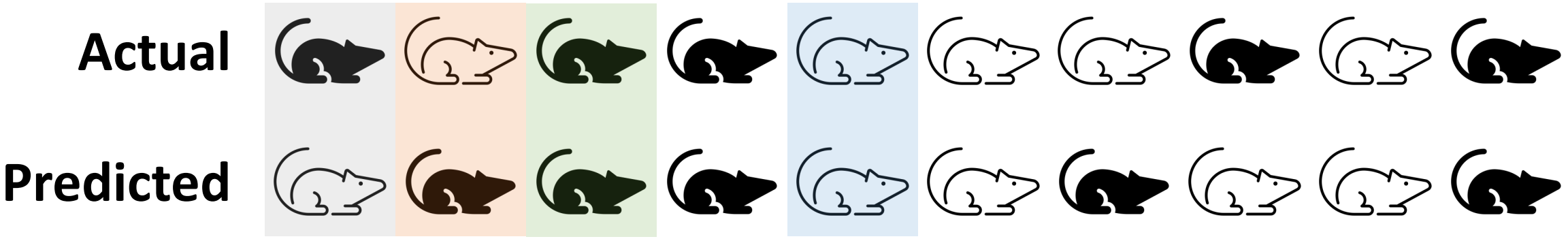


Leave One Out



K-Fold ($k = 5$)


Train Classifier
Test Classifier
Score Classifier
(i.e. Accuracy)



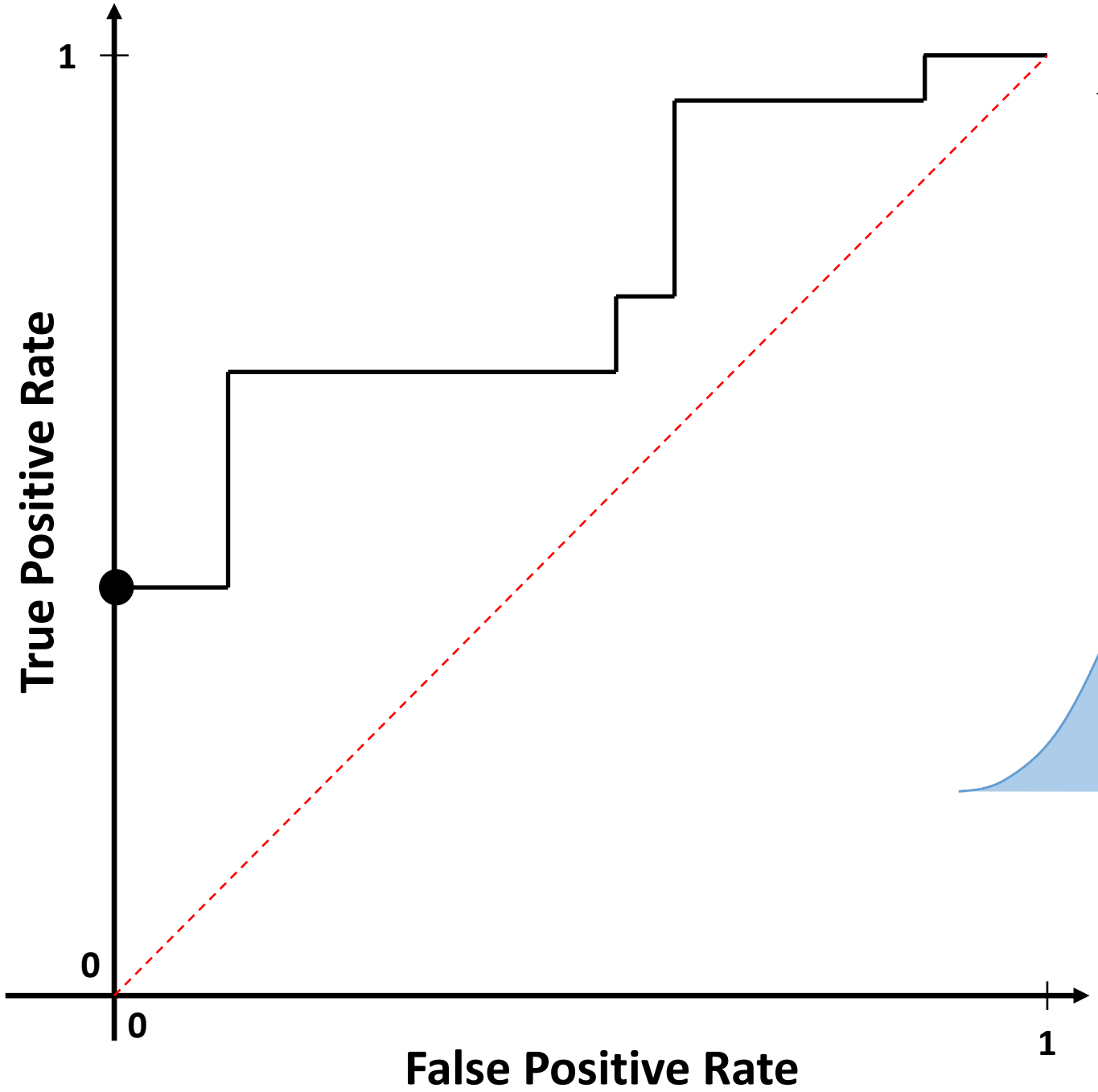
		Predicted: True (Pos)	Predicted: False (Neg)
Actual	True (Pos)	True Positive (TP)	False Negative (FN)
Actual	False (Neg)	False Positive (FP)	True Negative (TN)

False Positive Rate
(1 – Specificity)

$$\frac{FP}{TN + FP}$$

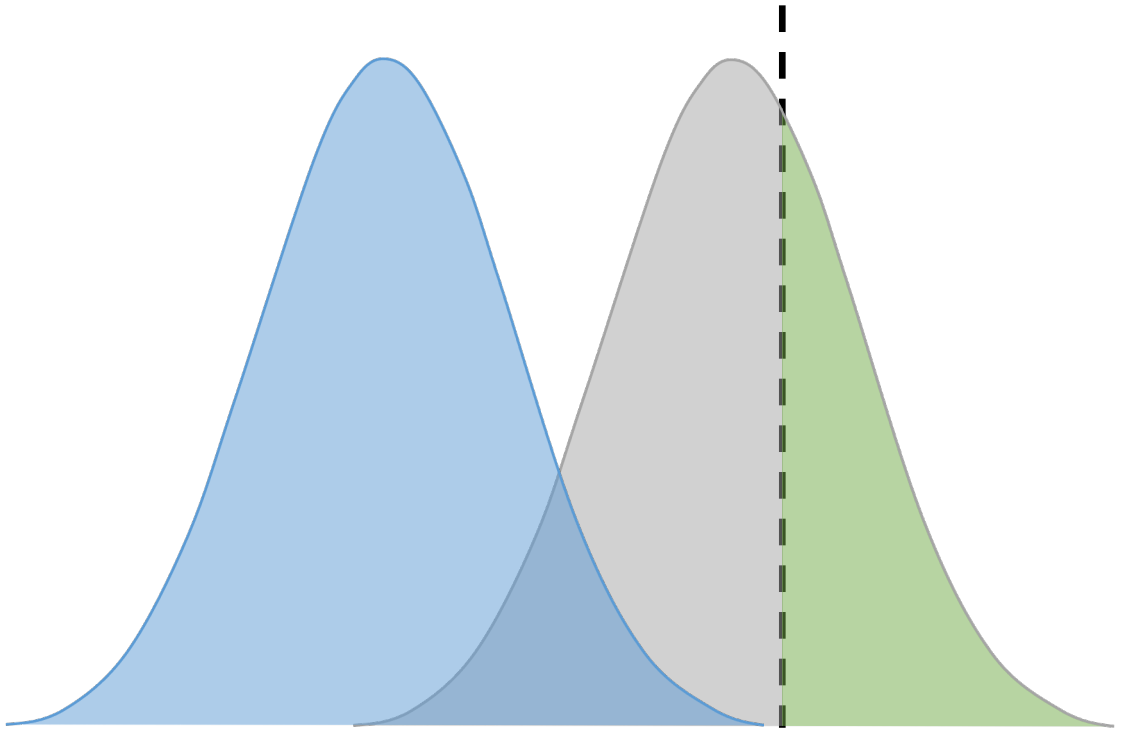
True Positive Rate
(Sensitivity)

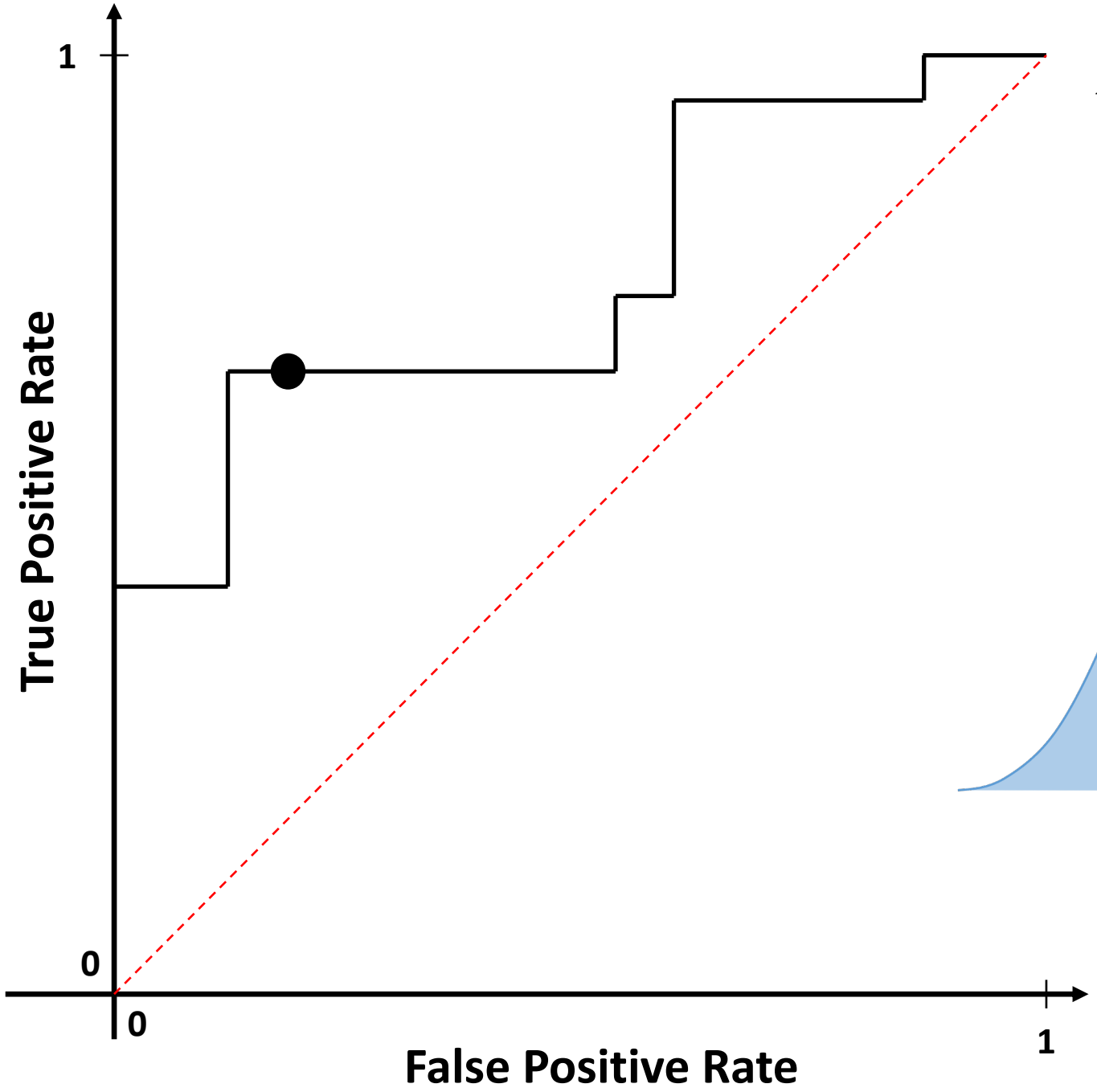
$$\frac{TP}{TP + FN}$$



$$FPR = \frac{FP}{TN + FP}$$

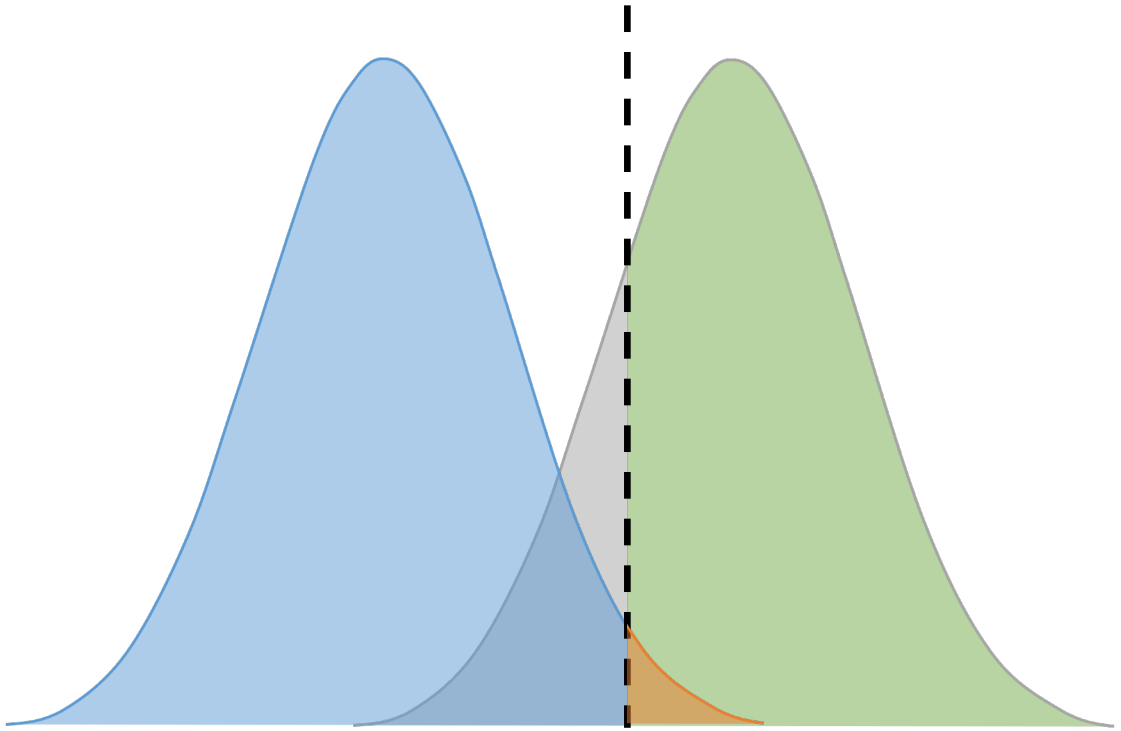
$$TPR = \frac{TP}{TP + FN}$$

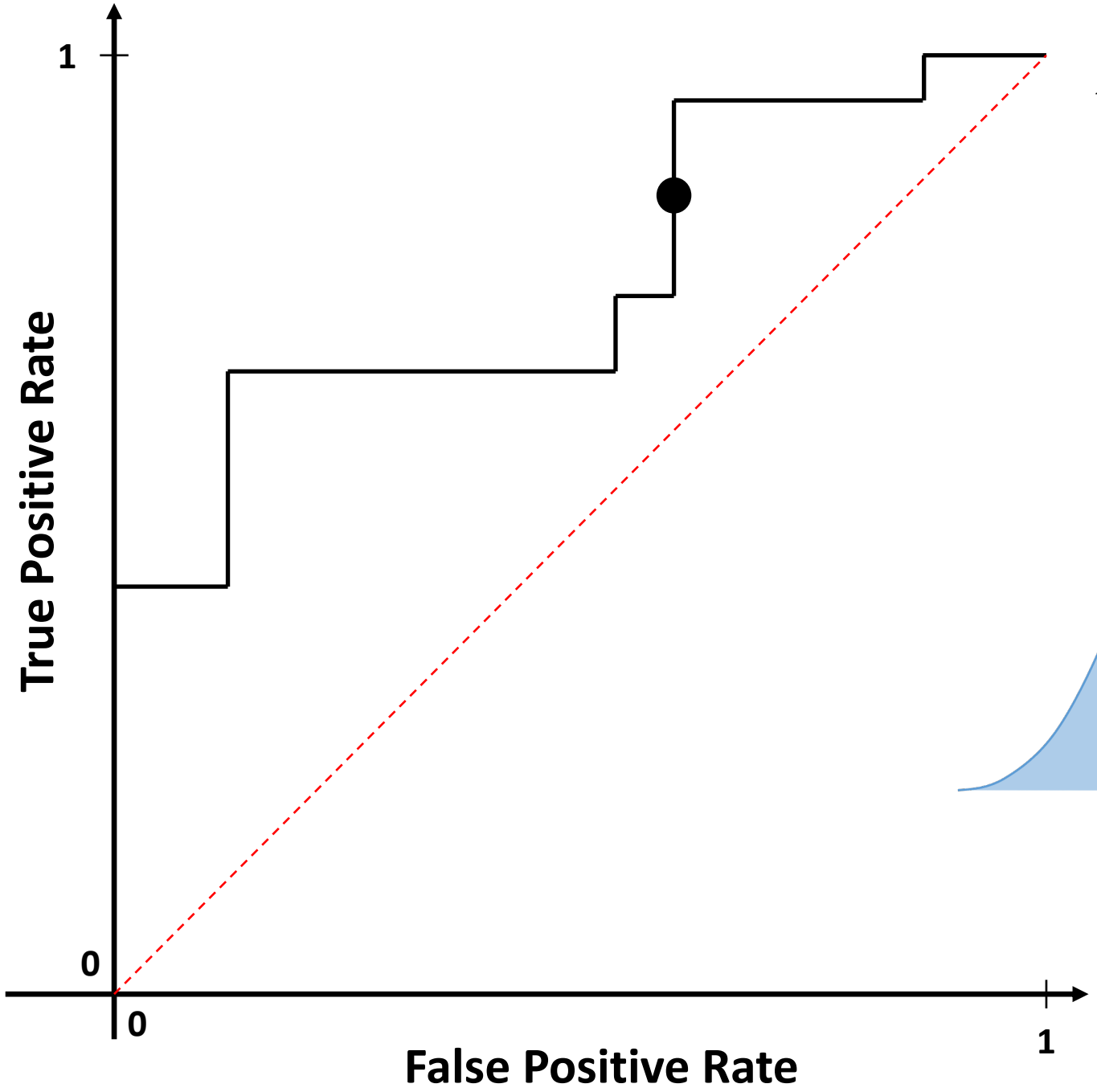




$$FPR = \frac{FP}{TN + FP}$$

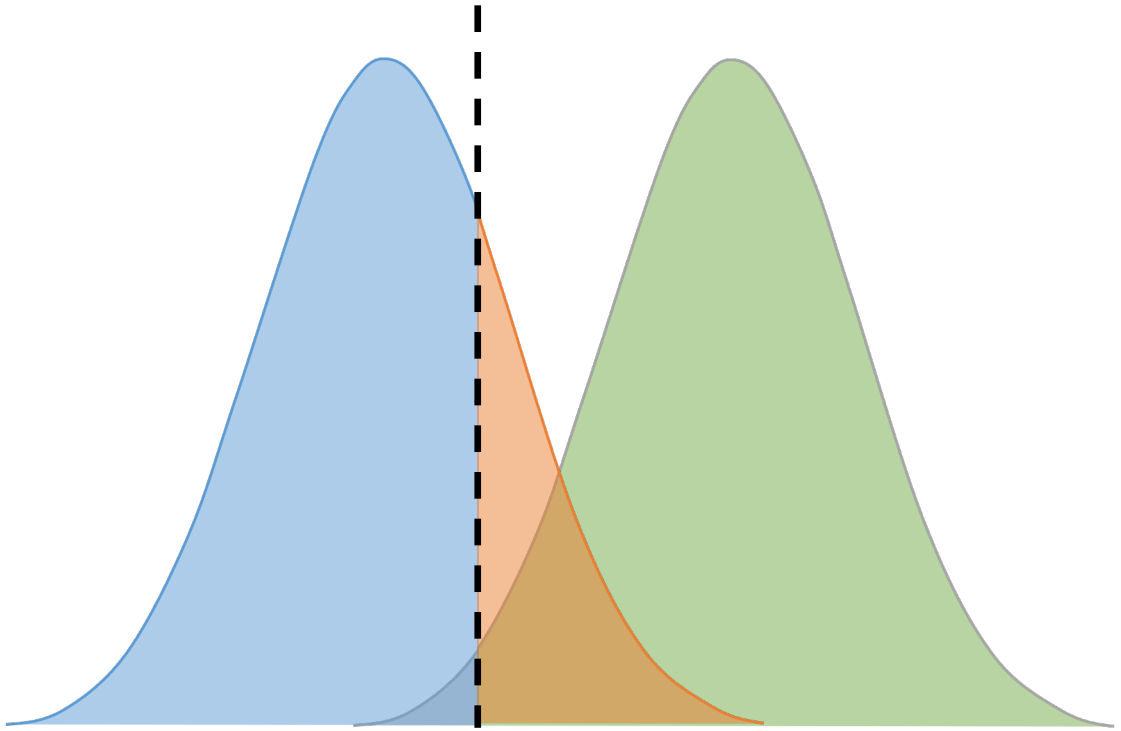
$$TPR = \frac{TP}{TP + FN}$$

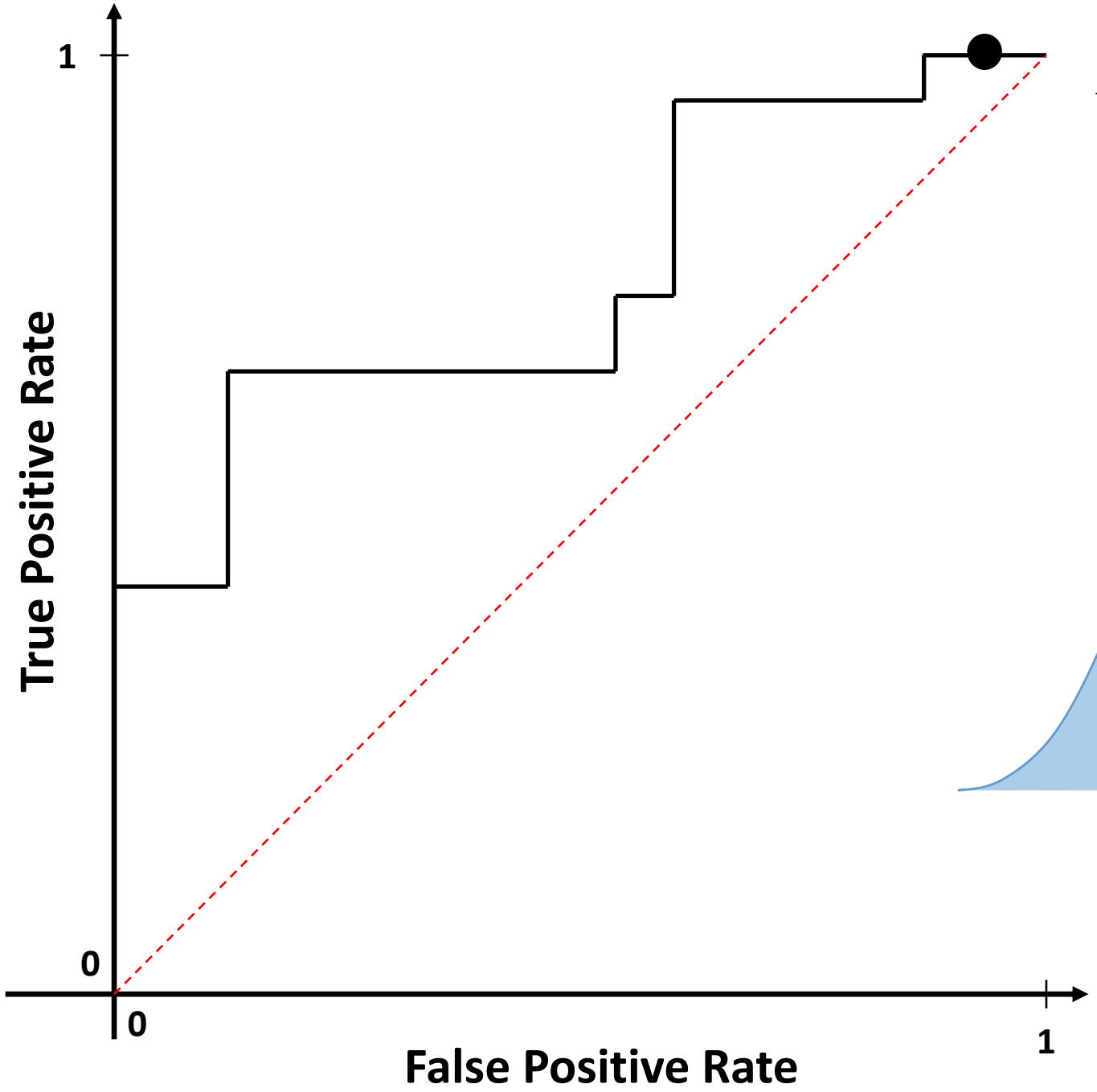




$$FPR = \frac{FP}{TN + FP}$$

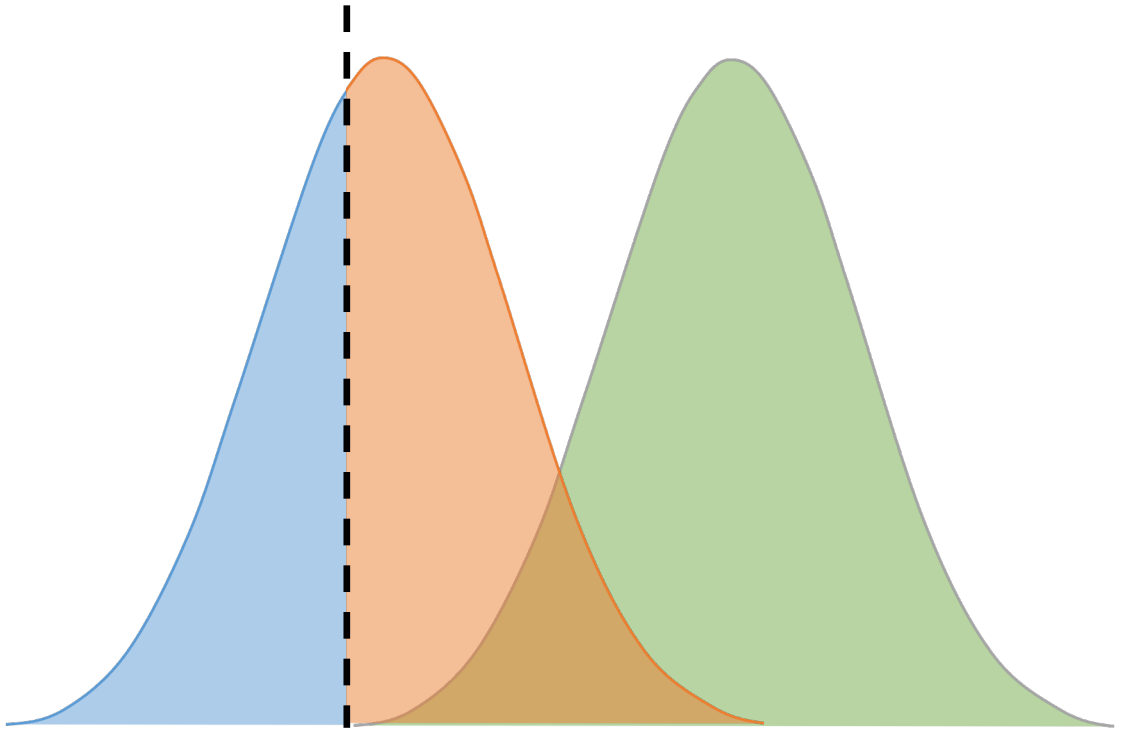
$$TPR = \frac{TP}{TP + FN}$$

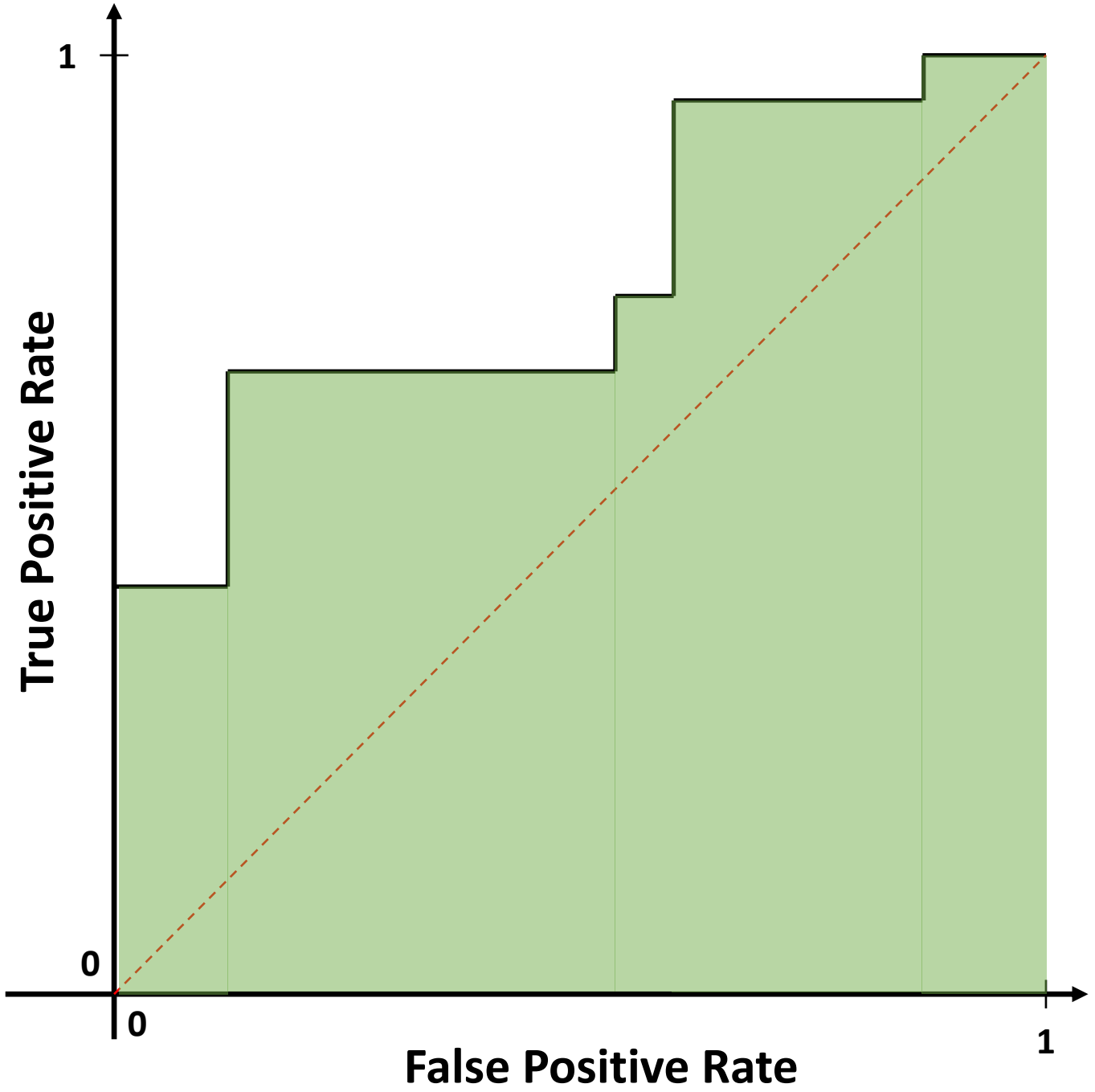


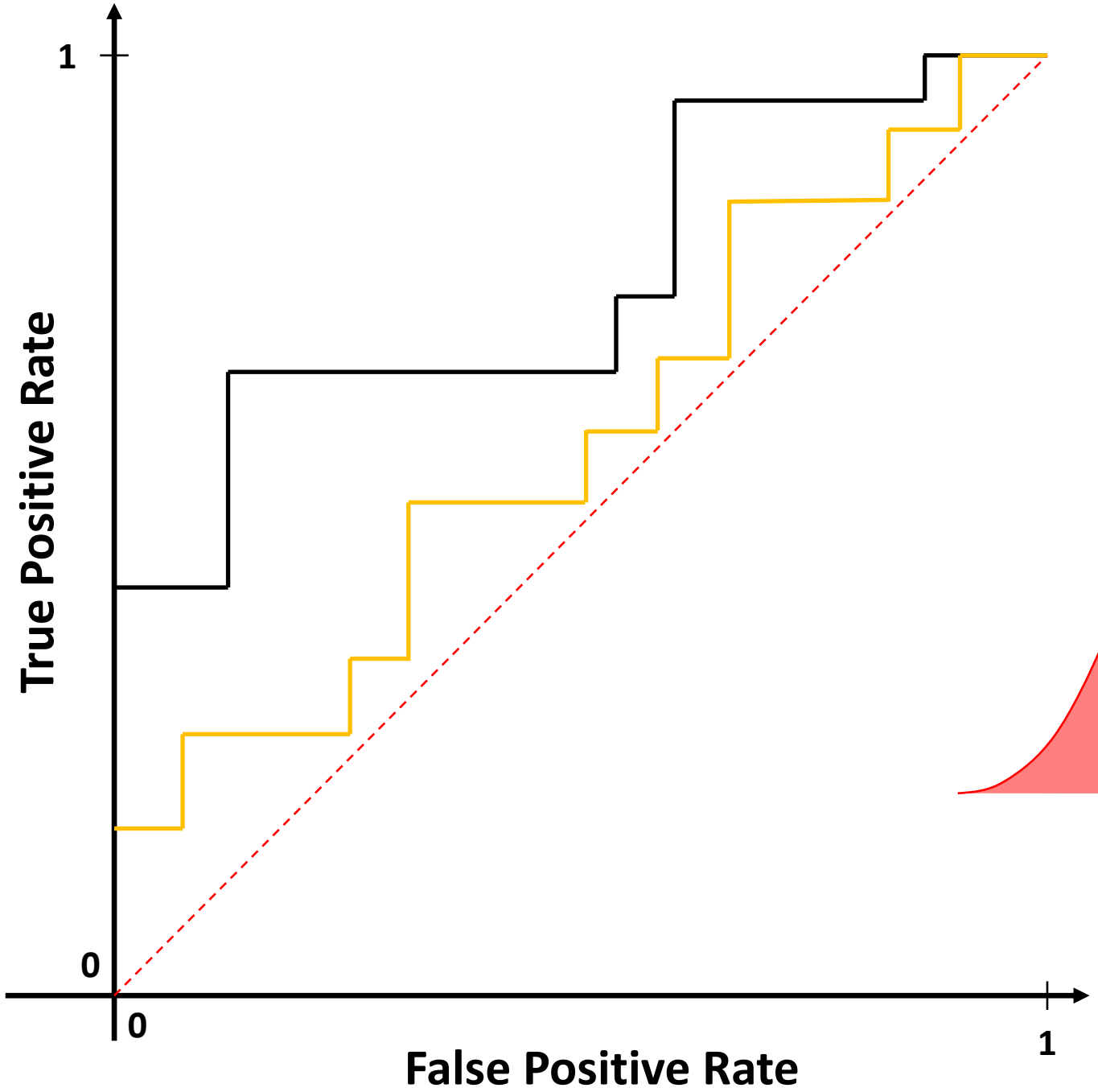


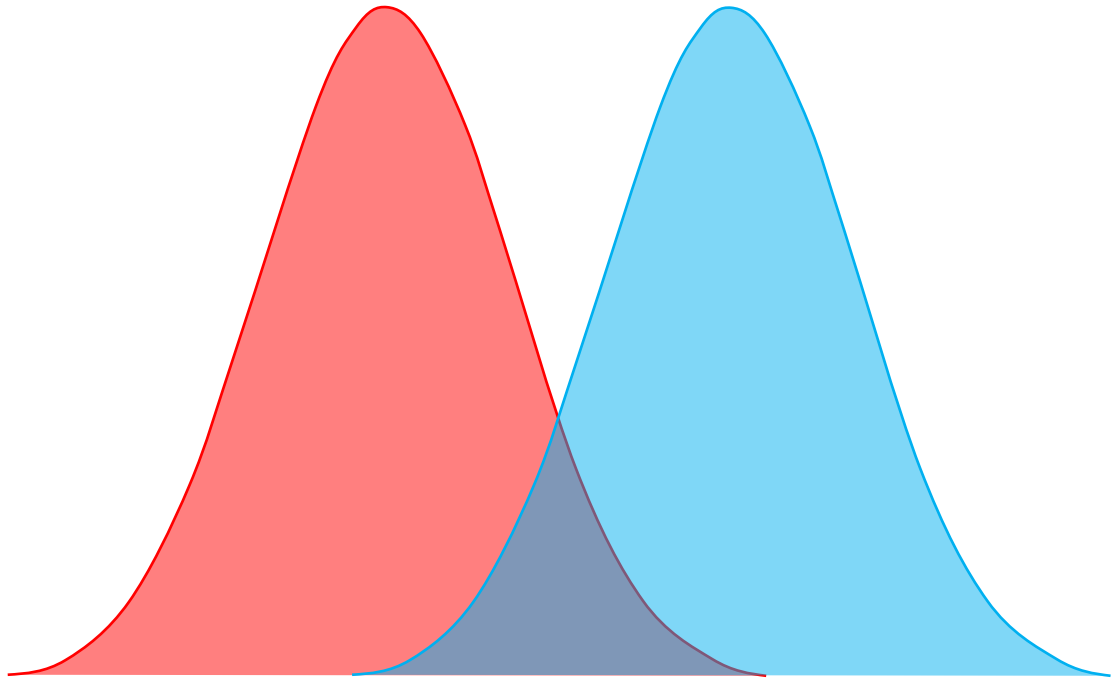
$$FPR = \frac{FP}{TN + FP}$$

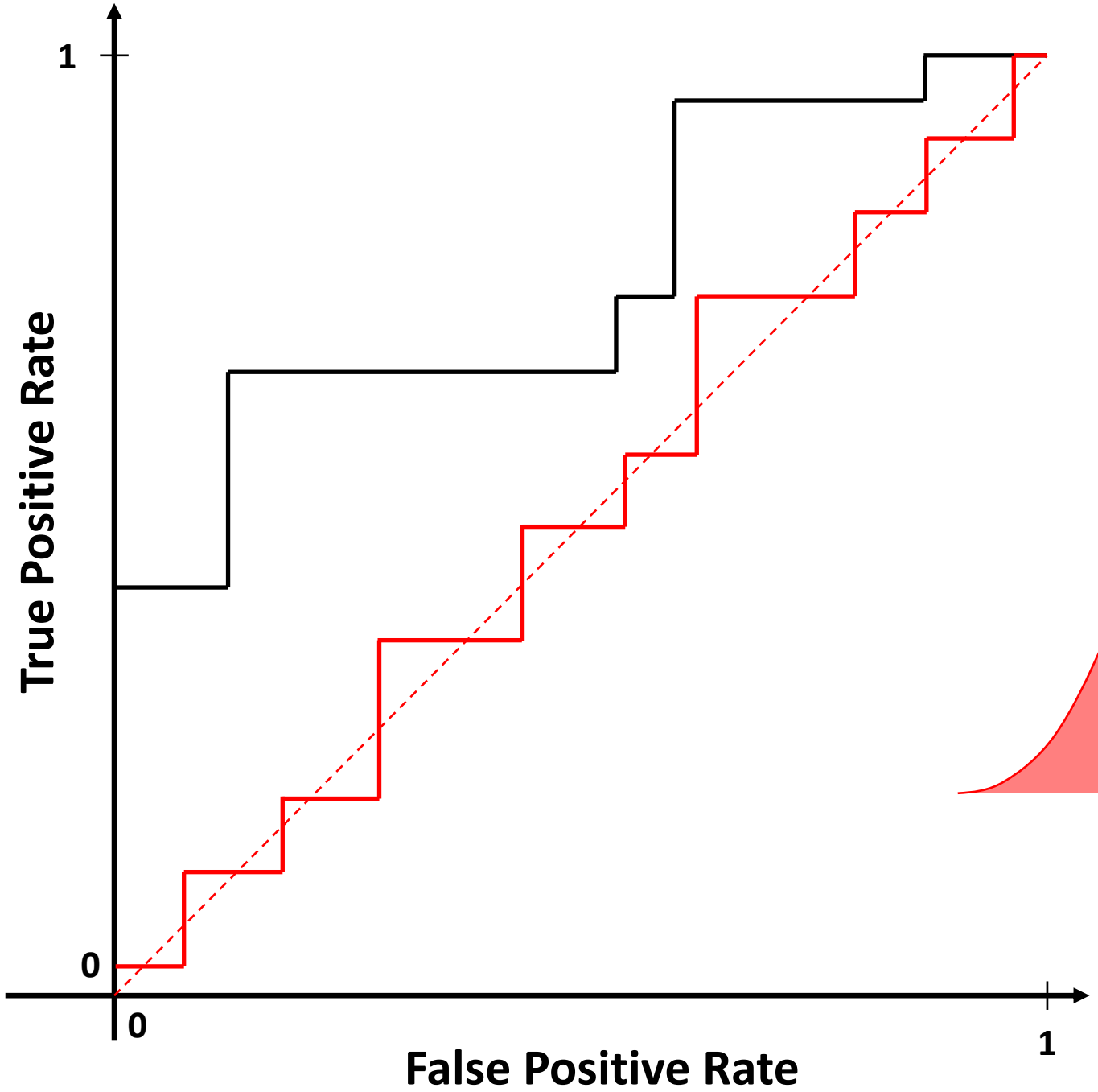
$$TPR = \frac{TP}{TP + FN}$$

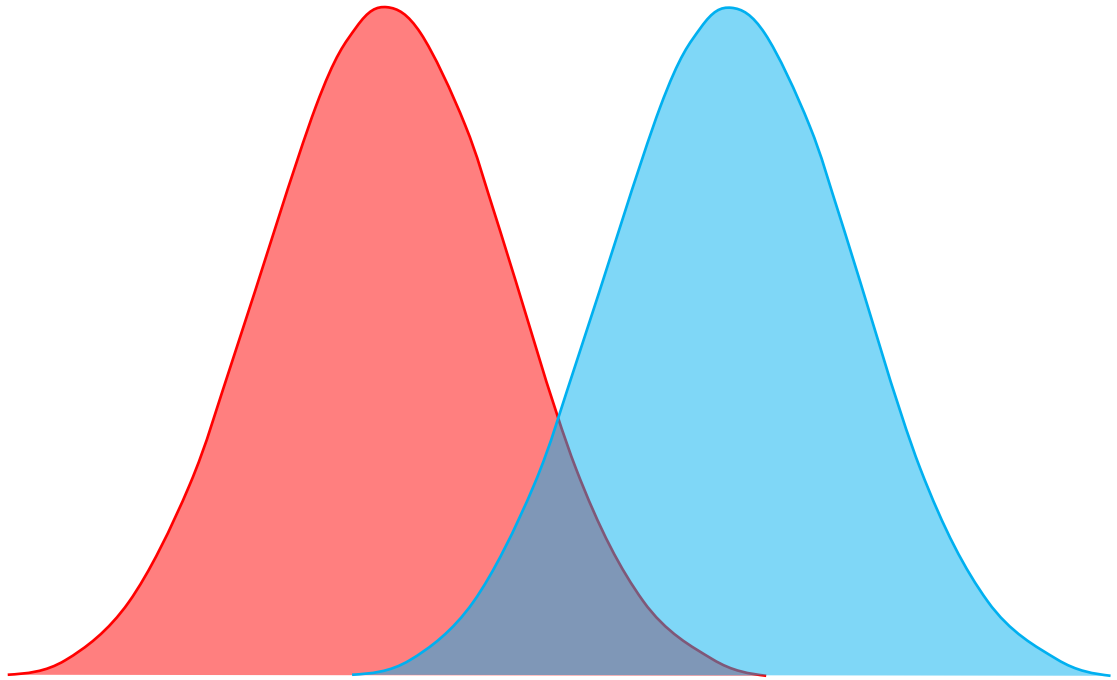


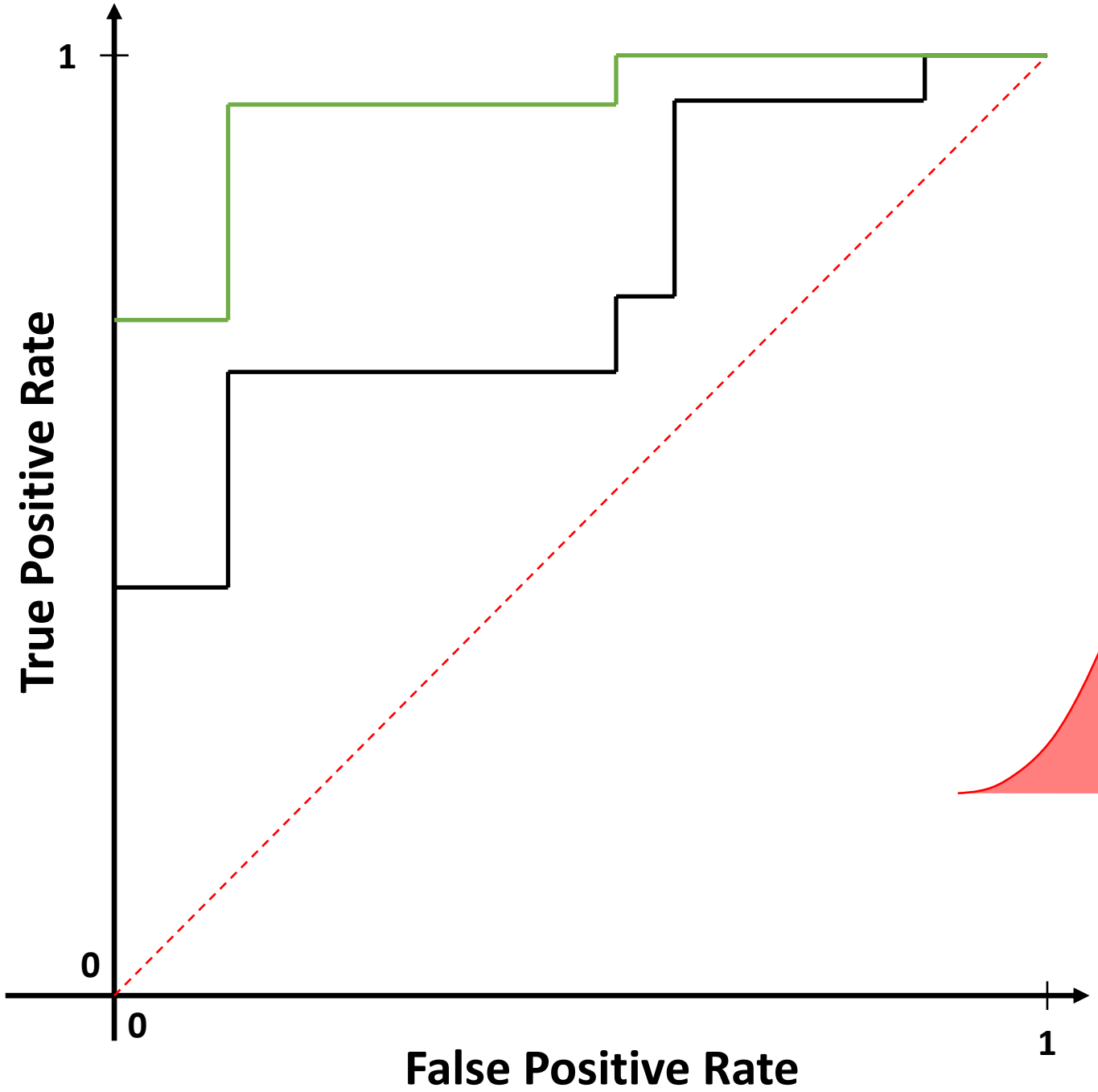

1
True Positive Rate
0
0
1
False Positive Rate

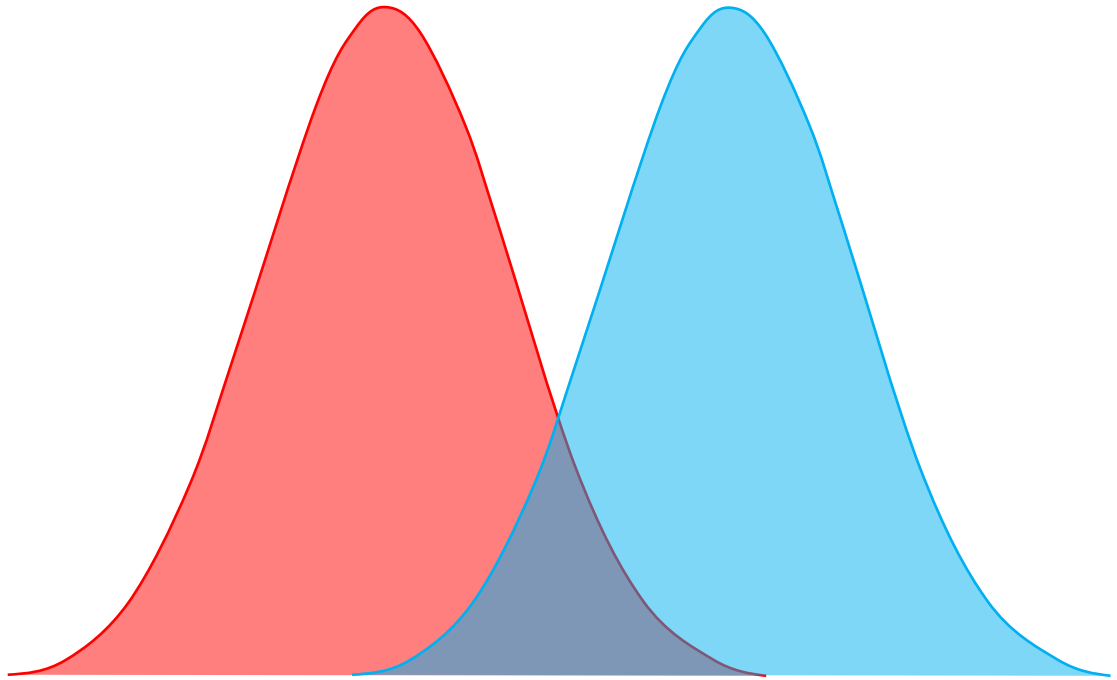

1
True Positive Rate
0
0
1
False Positive Rate

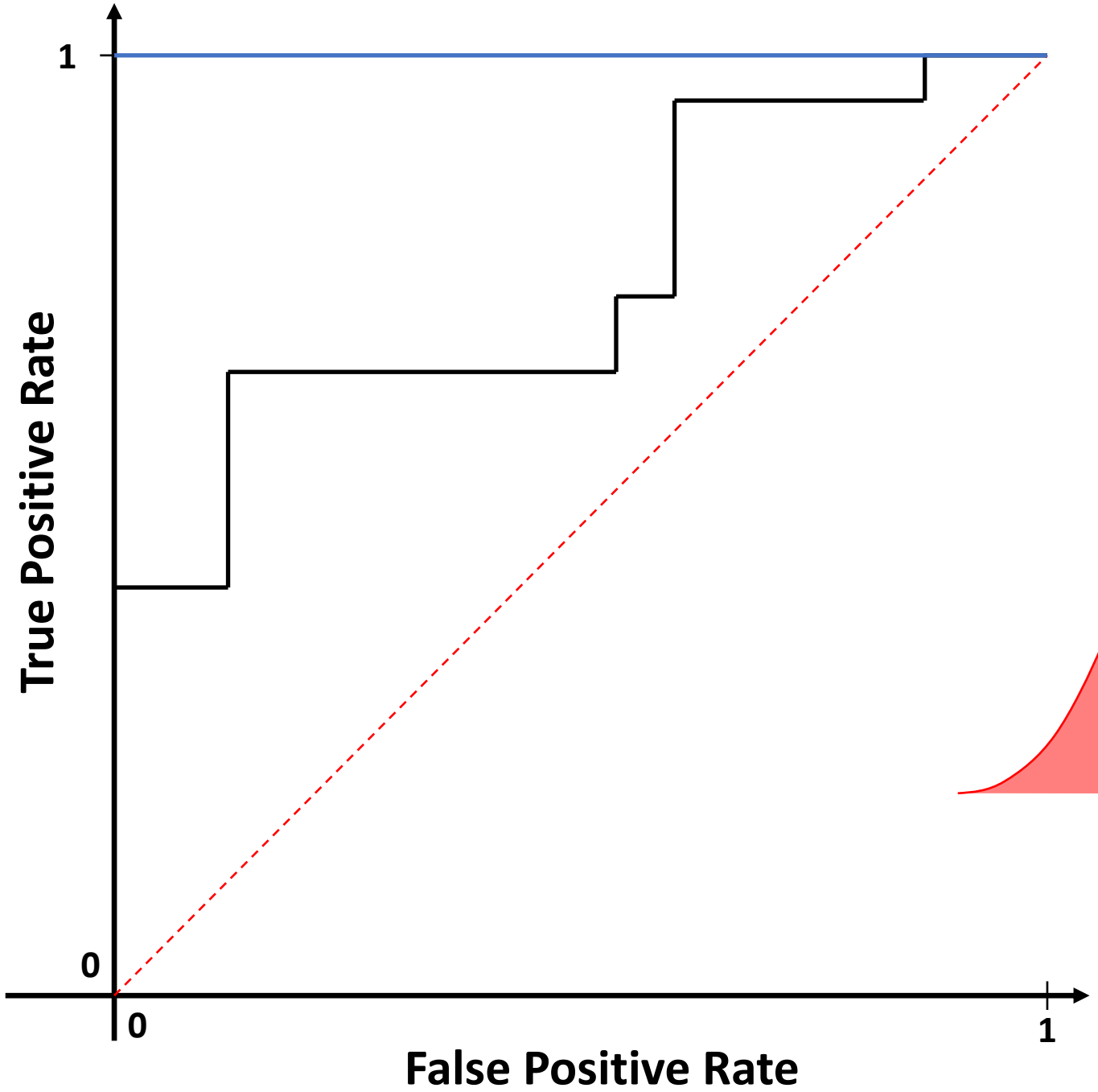


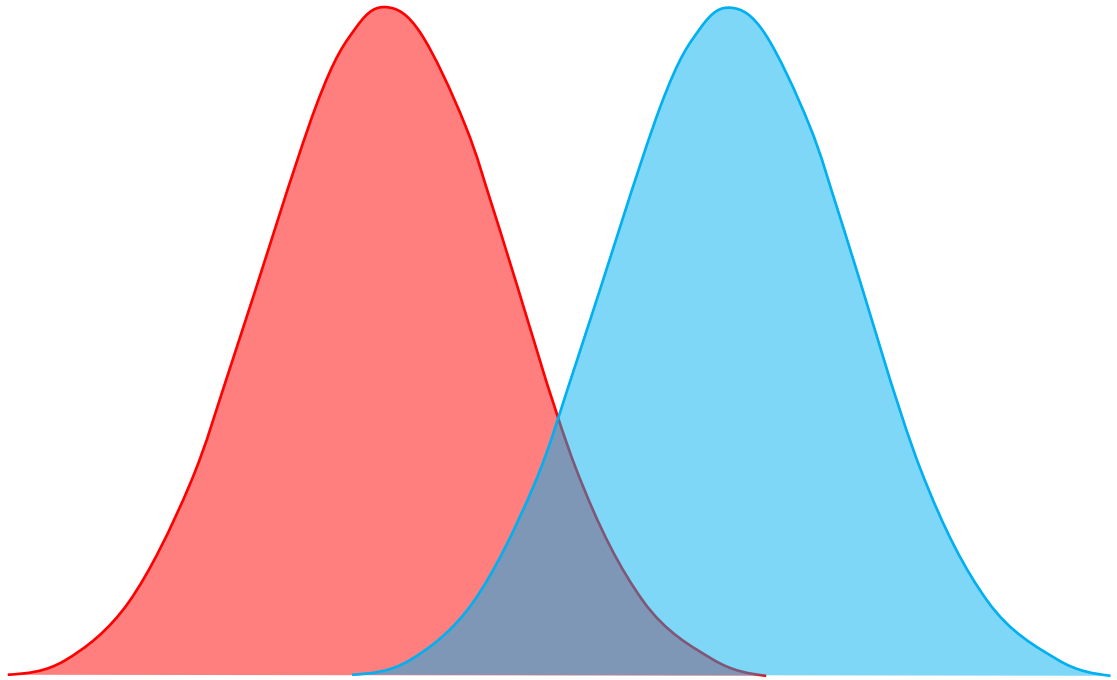

1
True Positive Rate
0
0
1
False Positive Rate




1
True Positive Rate
0
0
1
False Positive Rate



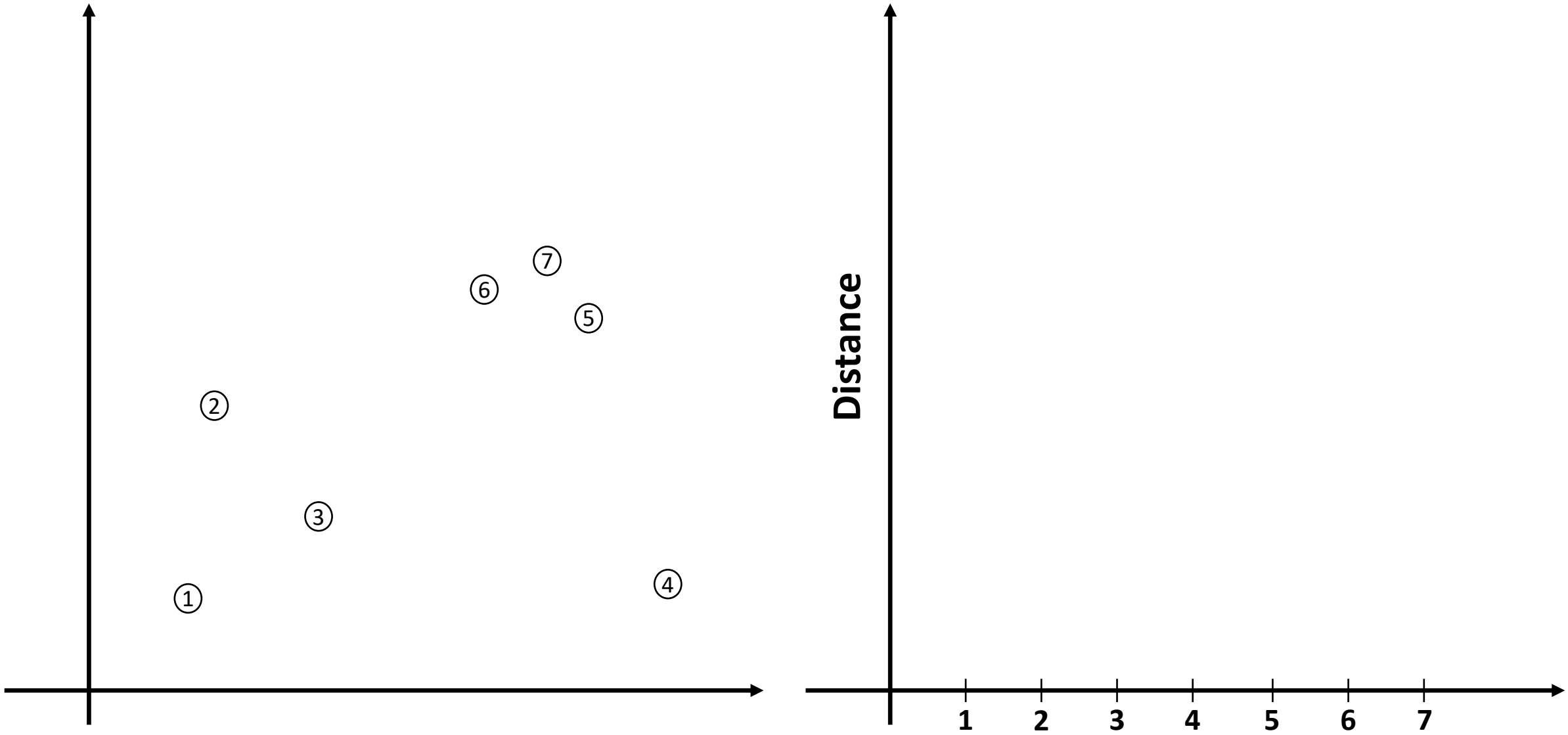

1
True Positive Rate
0
0
1
False Positive Rate

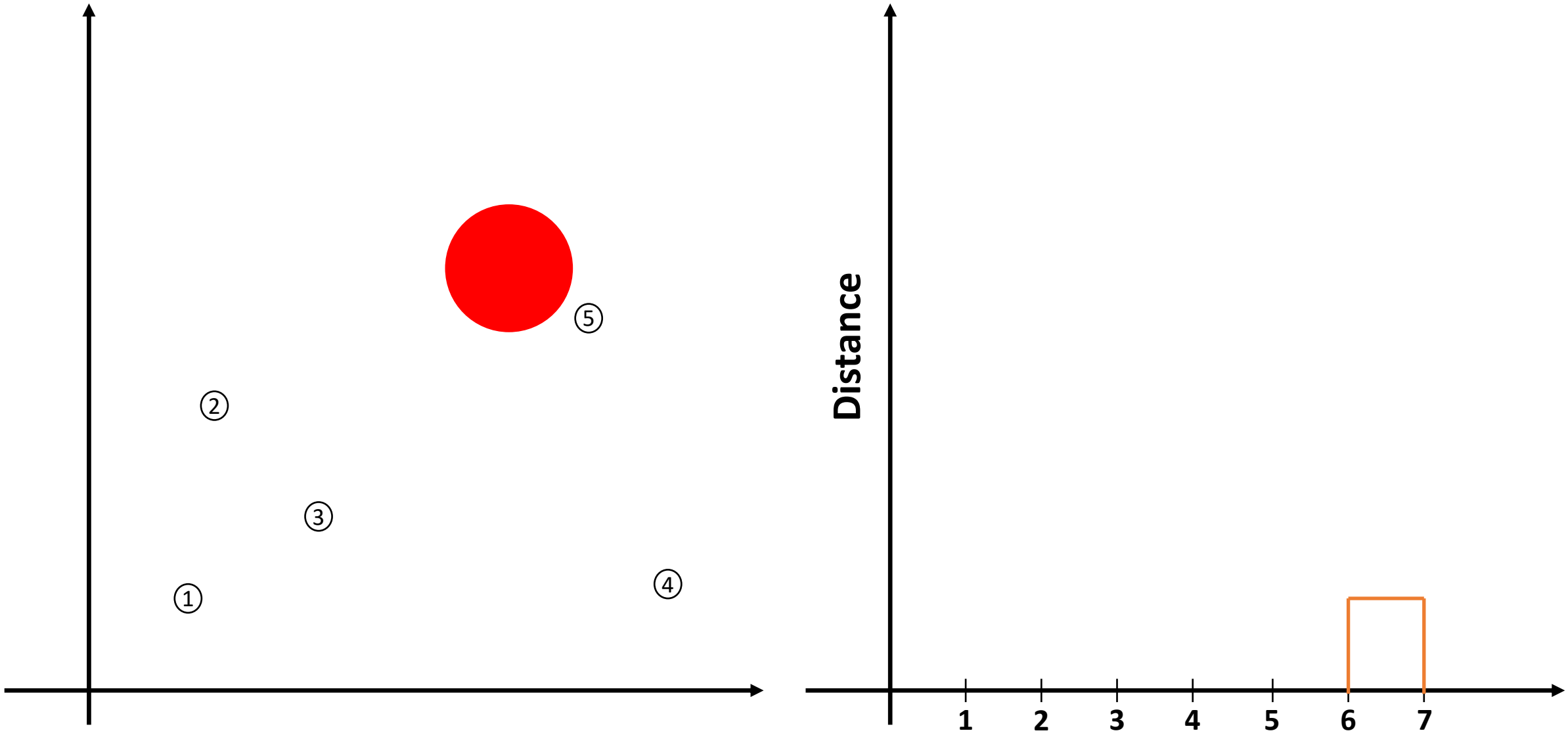


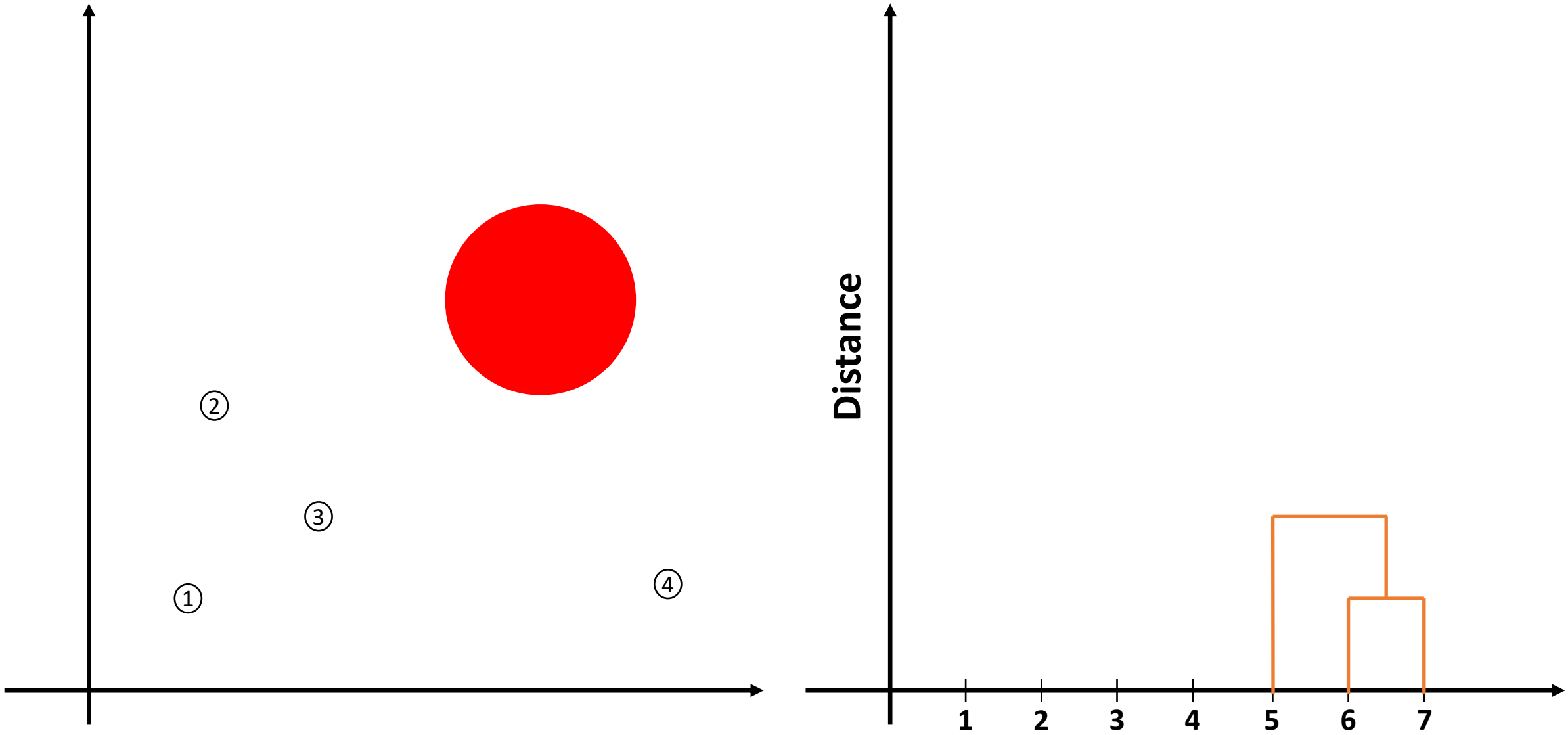
Summary

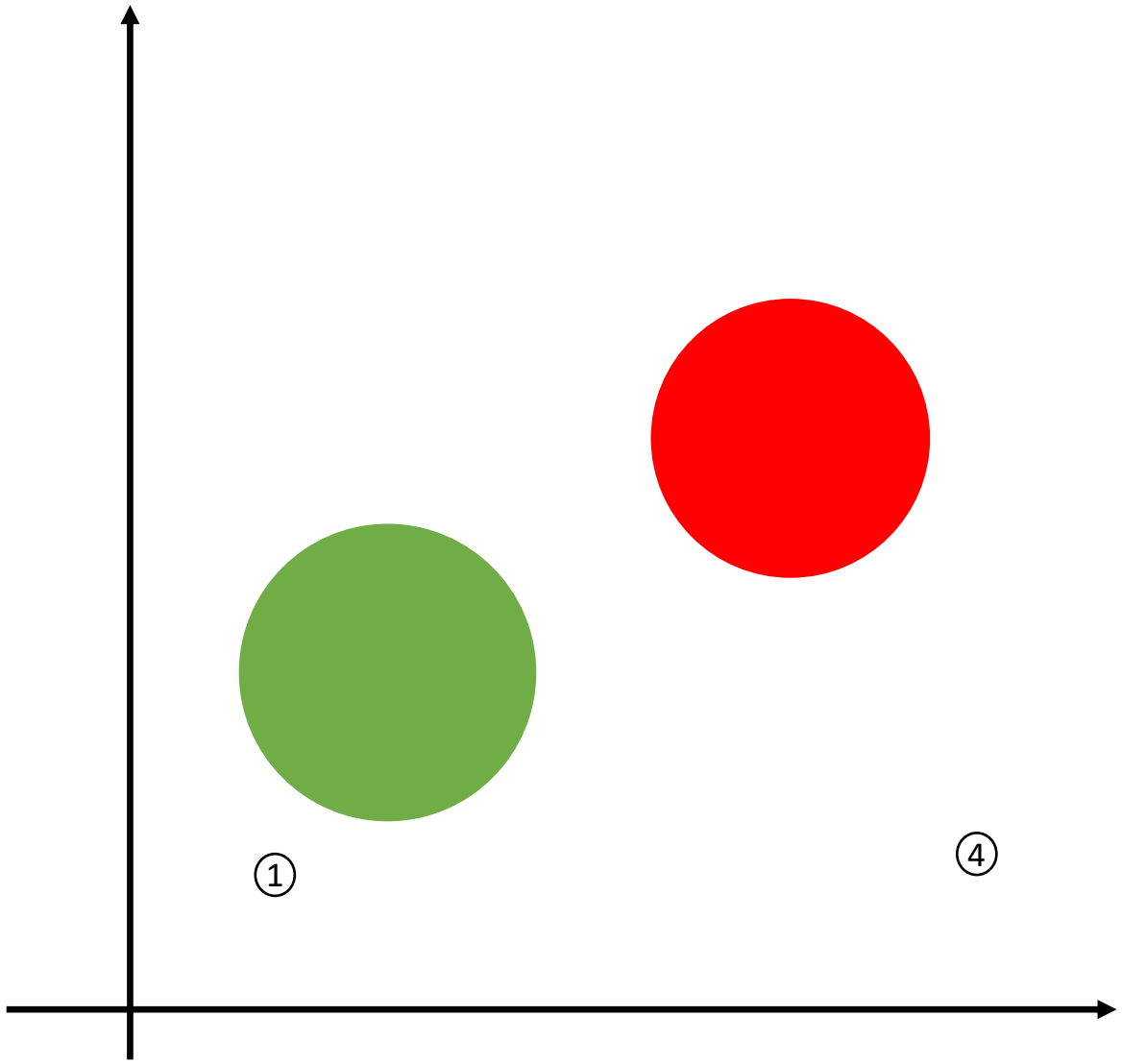
- Methods of performing cross-validation
- How to evaluate classifiers using classification accuracy and ROC
- What a good and bad classifier is with regards to AUC and ROC
- How this works with our data

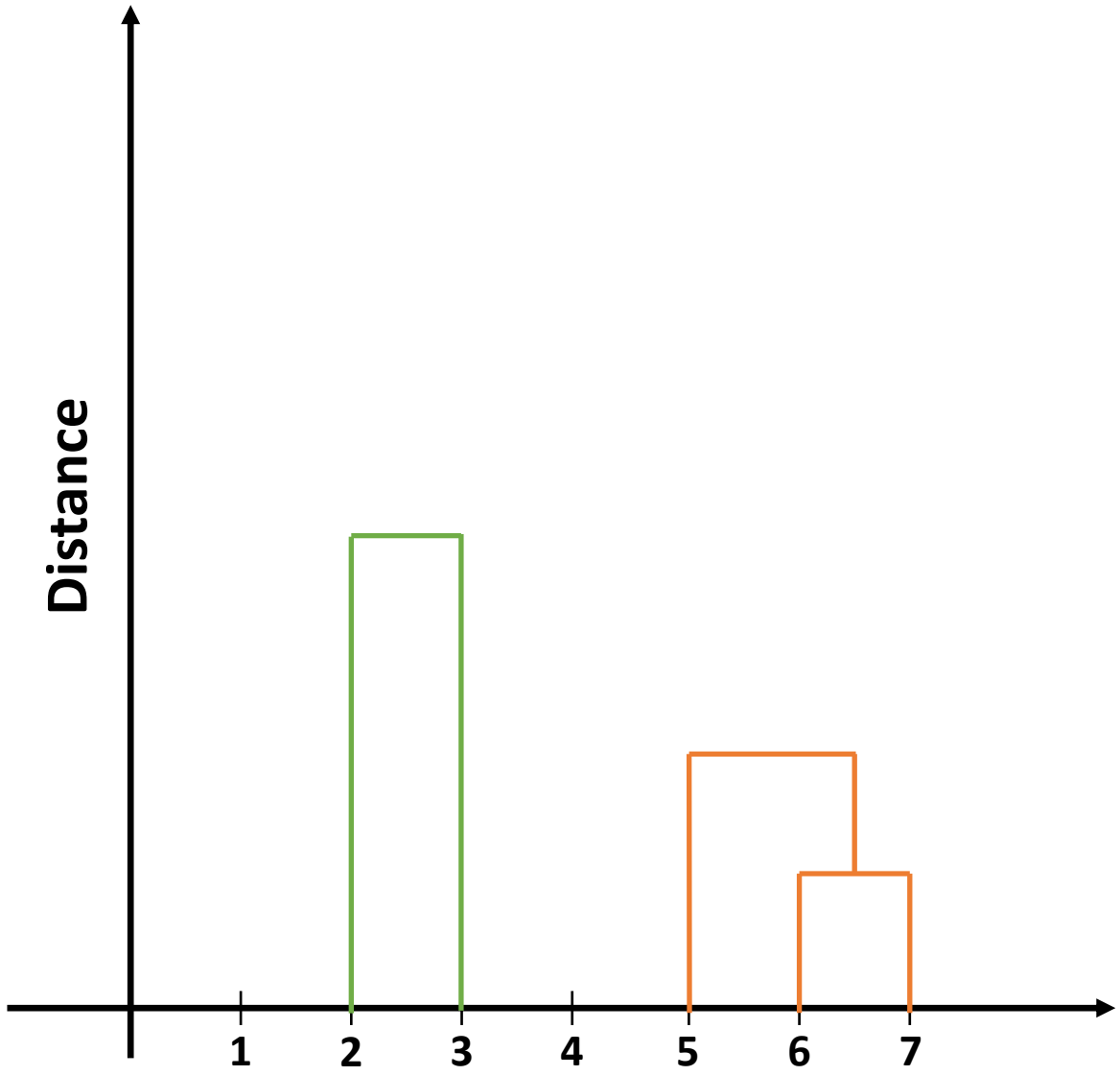
Hierarchical Clustering

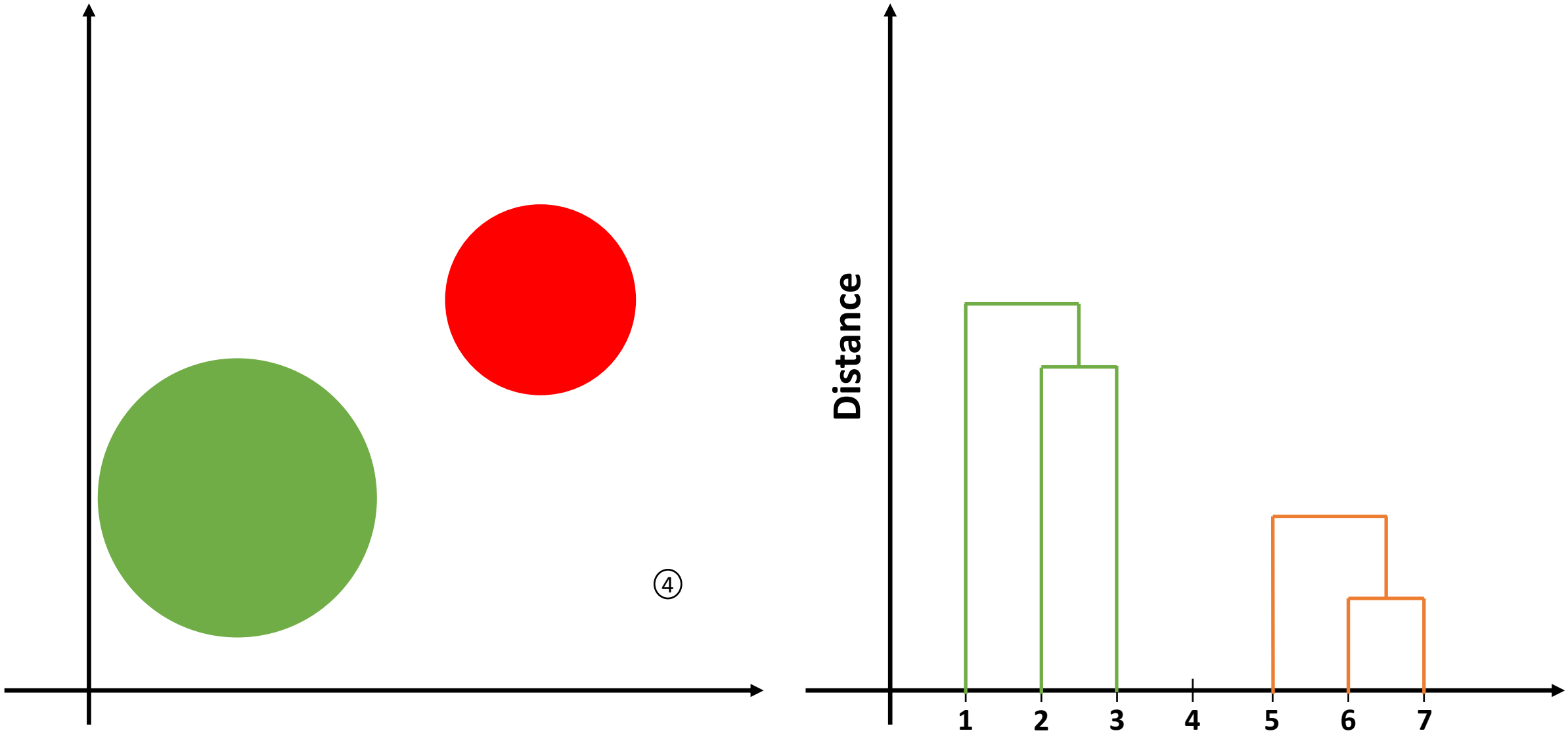

7
6
5
2
3
1
4
Distance
1
2
3
4
5
6
7

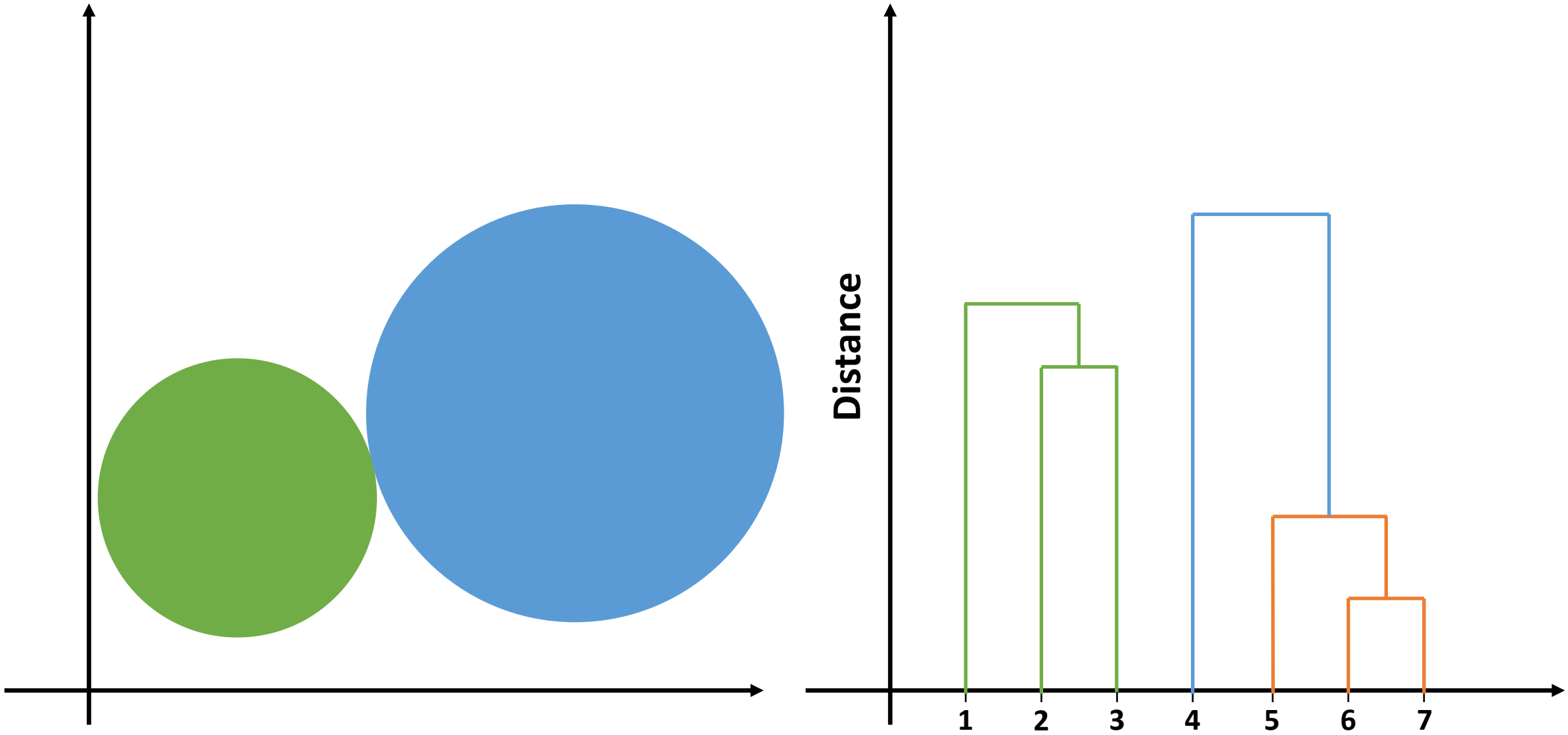

①
②
③
④
⑤
Distance
1
2
3
4
5
6
7

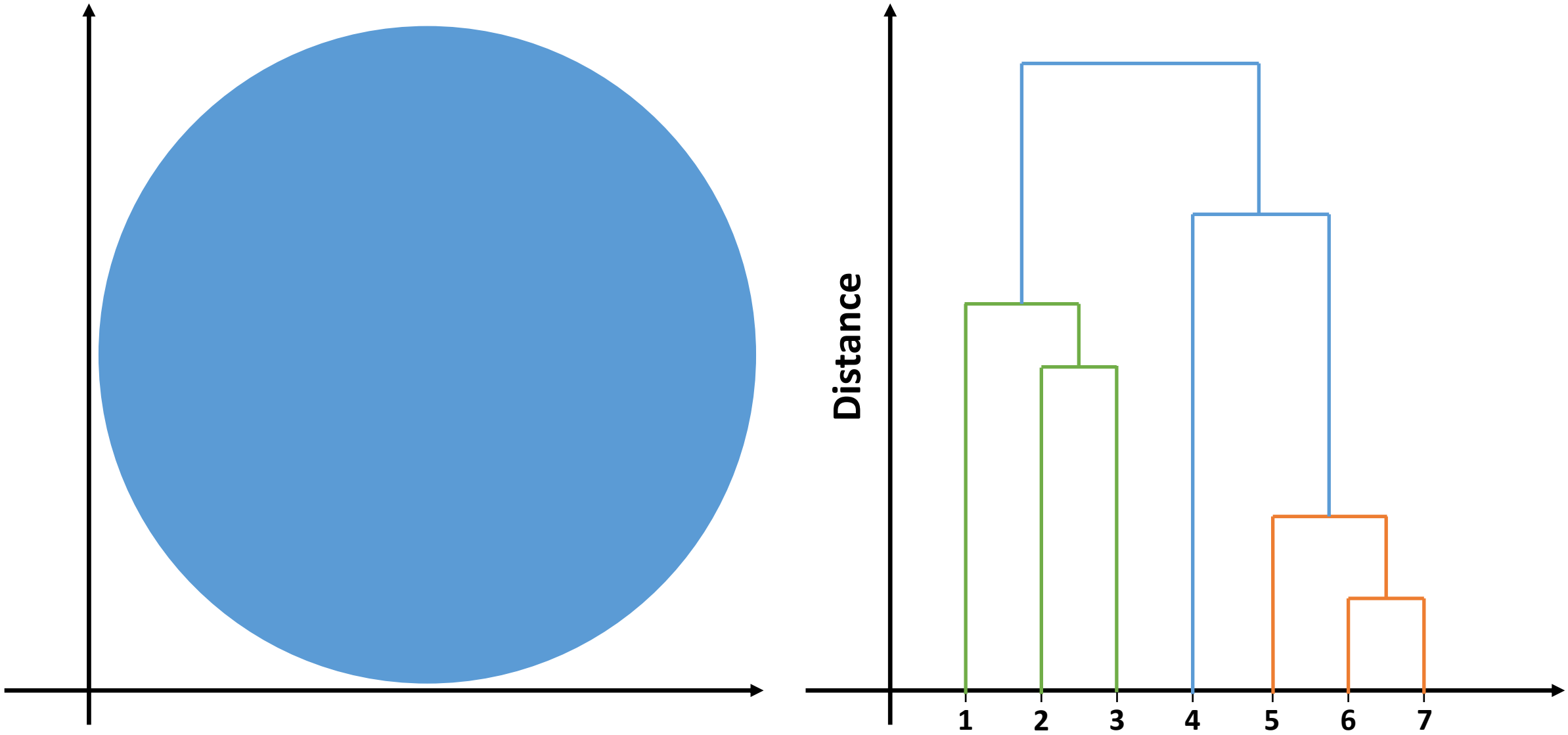

②
③
①
④
Distance
1
2
3
4
5
6
7

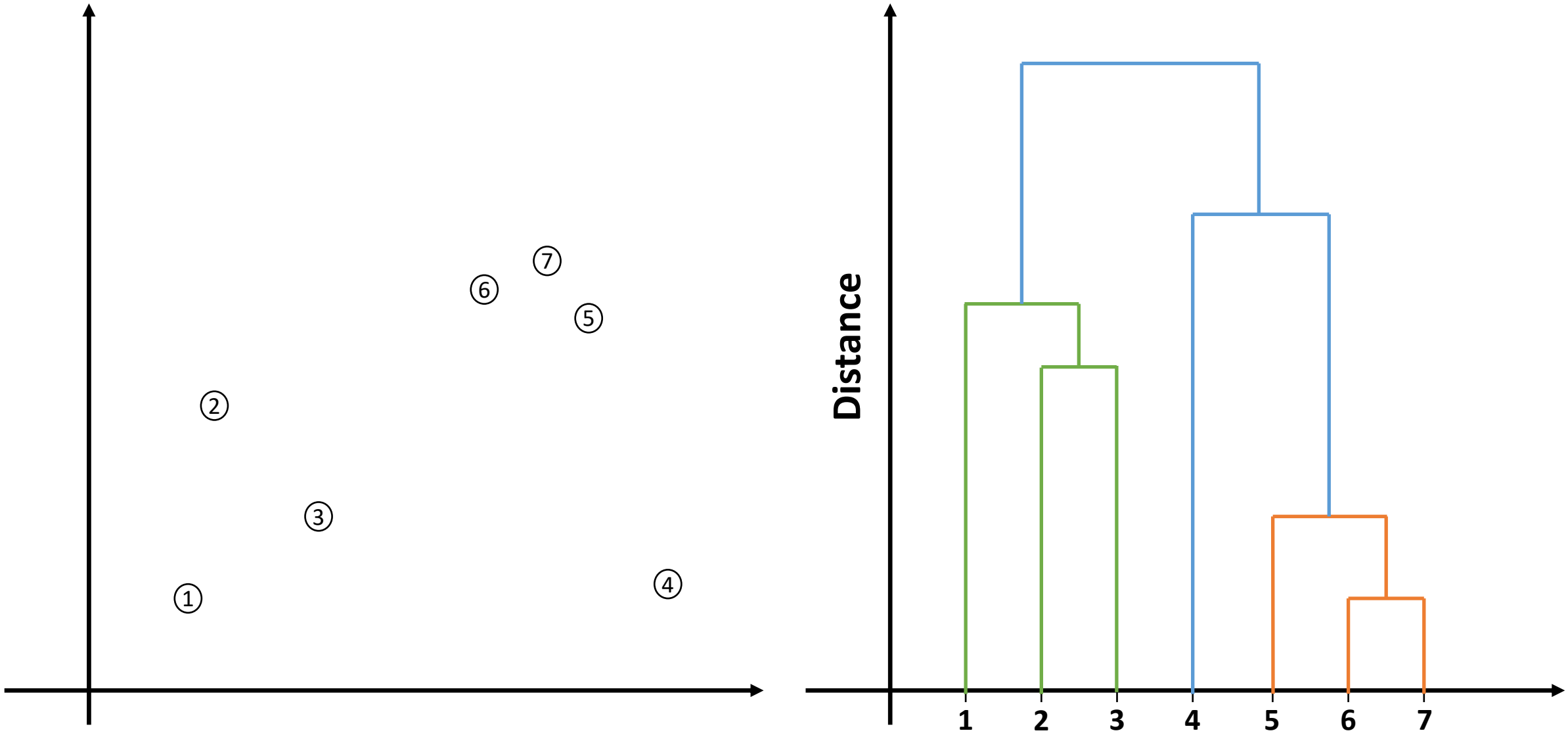

①
④


Distance
1
2
3
4
5
6
7


④
Distance
1
2
3
4
5
6
7


Distance
1
2
3
4
5
6
7


Distance
1
2
3
4
5
6
7


Distance
1
2
3
4
5
6
7

Summary

- Learned how to do hierarchical clustering

Acknowledgements

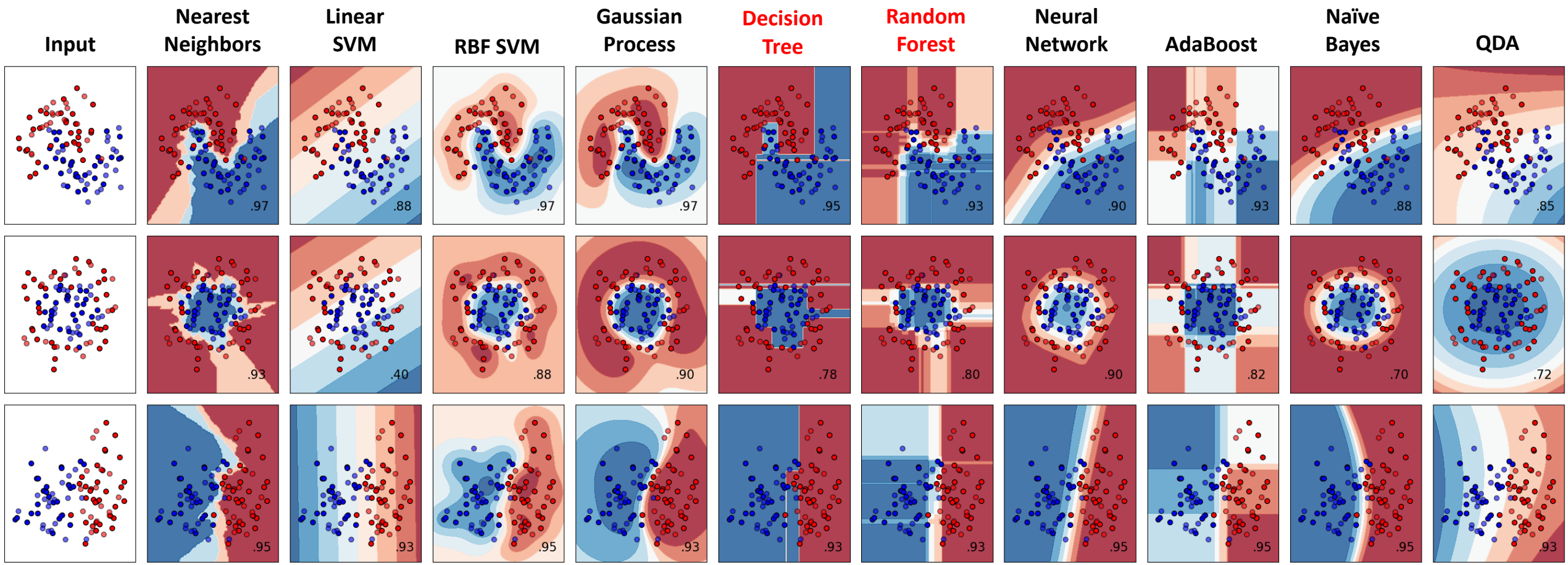
JHU

- Joshua Vogelstein
- Cencheng Shen

NIEHS

- Jesse Cushman
- Dalisa Kendricks
- Leslie Wilson
- Jariatu Stallone
- Sydney Fry
- DaNashia Thomas

Questions?


Input
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
Random Forest
Neural Network
AdaBoost
Naïve Bayes
QDA
.97
.88
.97
.97
.95
.93
.90
.93
.88
.85
.93
.40
.88
.90
.78
.80
.90
.82
.70
.72
.95
.93
.95
.93
.93
.93
.95
.95
.95
.93