Introduction to Data Science

Lecture 5; April 25th, 2016

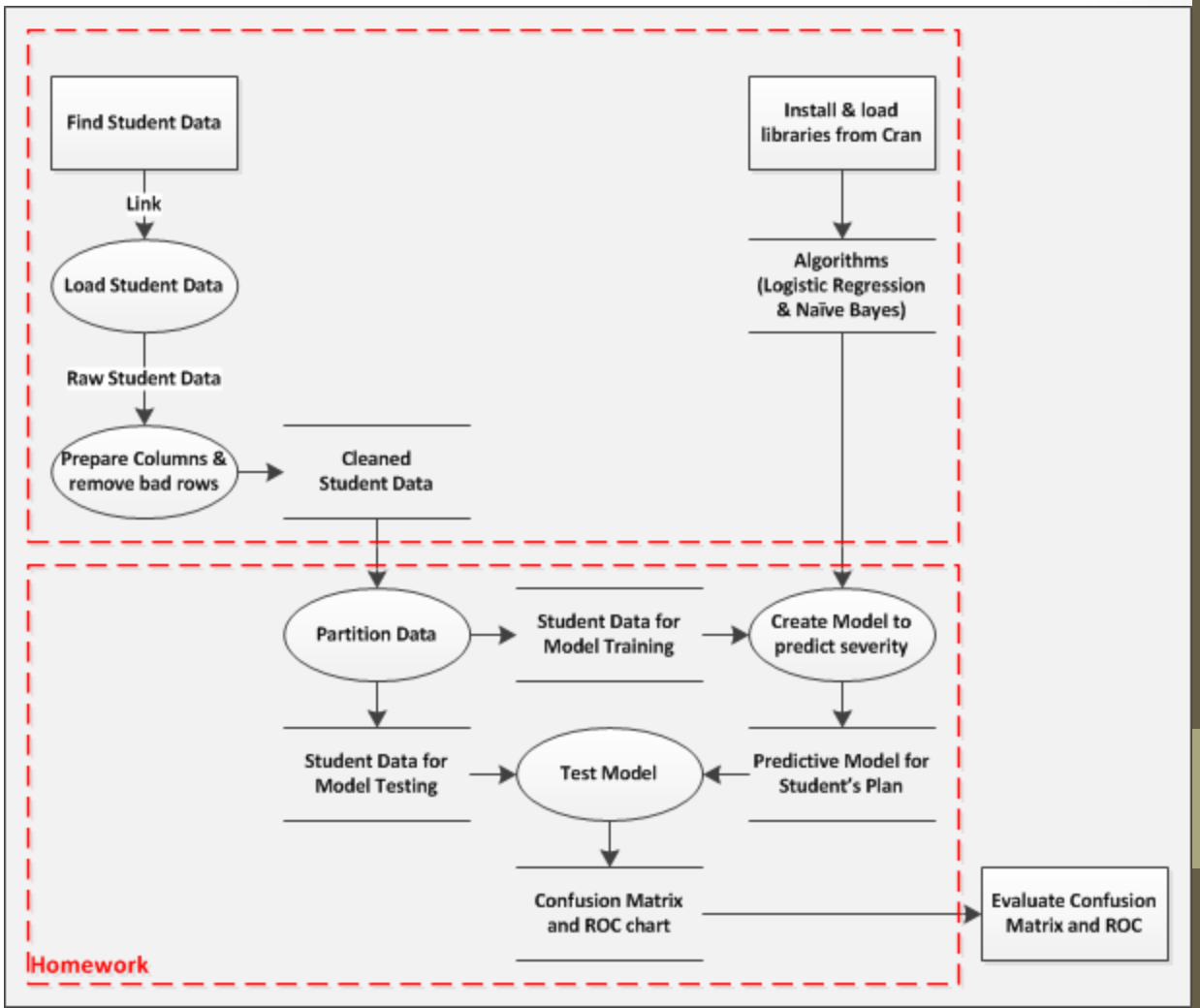
Ernst Henle
ErnstHe@UW.edu
Skype: ernst-henle

Agenda

- Announcements
  - Encourage Group Homework!
  - Ask questions on LinkedIn
  - Guest Lectures in May
    - Business side of Data Science by Marius Marcu on May 9th 2016
    - Data Visualization by Tatyana Yakushev on May 16th 2016
    - Building a Data Science Group by Sarmila Basu
- Review Classifications in R
- Quiz 05a Classifications
- Break
- Overfitting and Confusion Matrix
- Video and Break
- ROC Chart Demo
- Quiz 05b on Confusion Matrix
- How to make an ROC
- Predictive Analytics Iteration Trap (Time Permitting)
- Predictive Anecdotes (Time Permitting)
- Assignment (Complete all assignments items from all assignment slides)

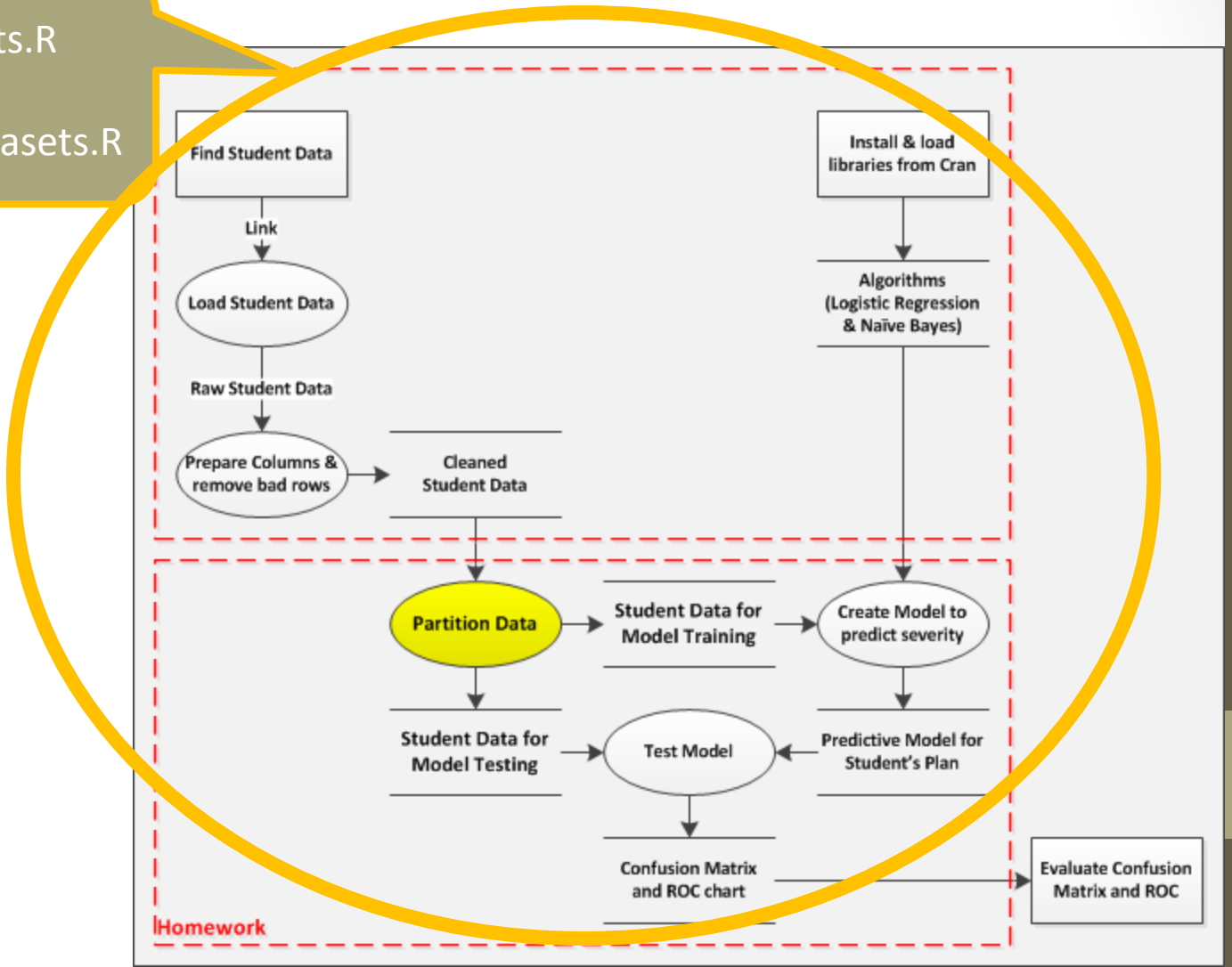
Homework Review: Classifications in R

Homework Review: Classifications in R



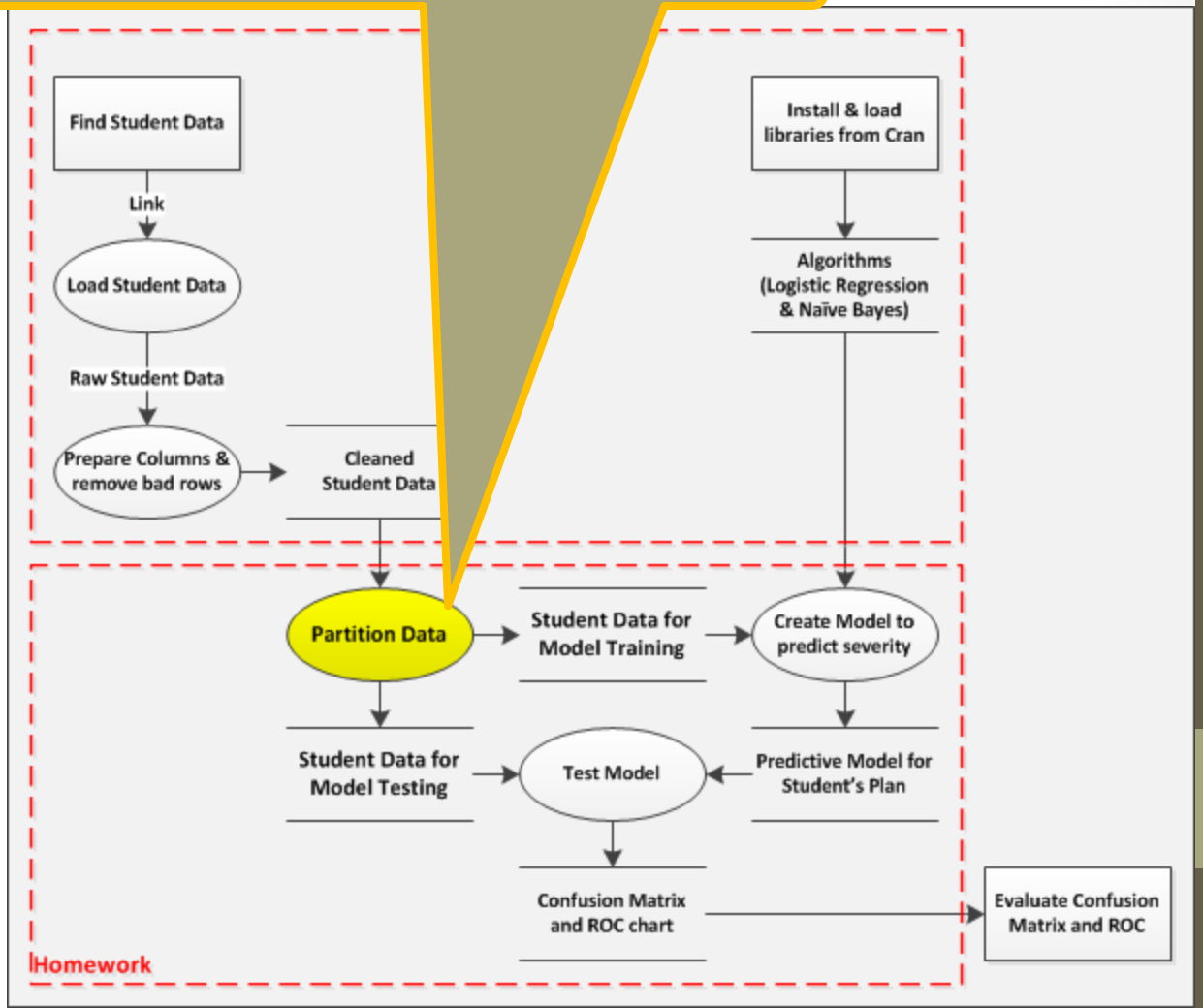
Homework Review: Classifications in R

ClassifyStudents.R
&
CollegeStudentDatasets.R



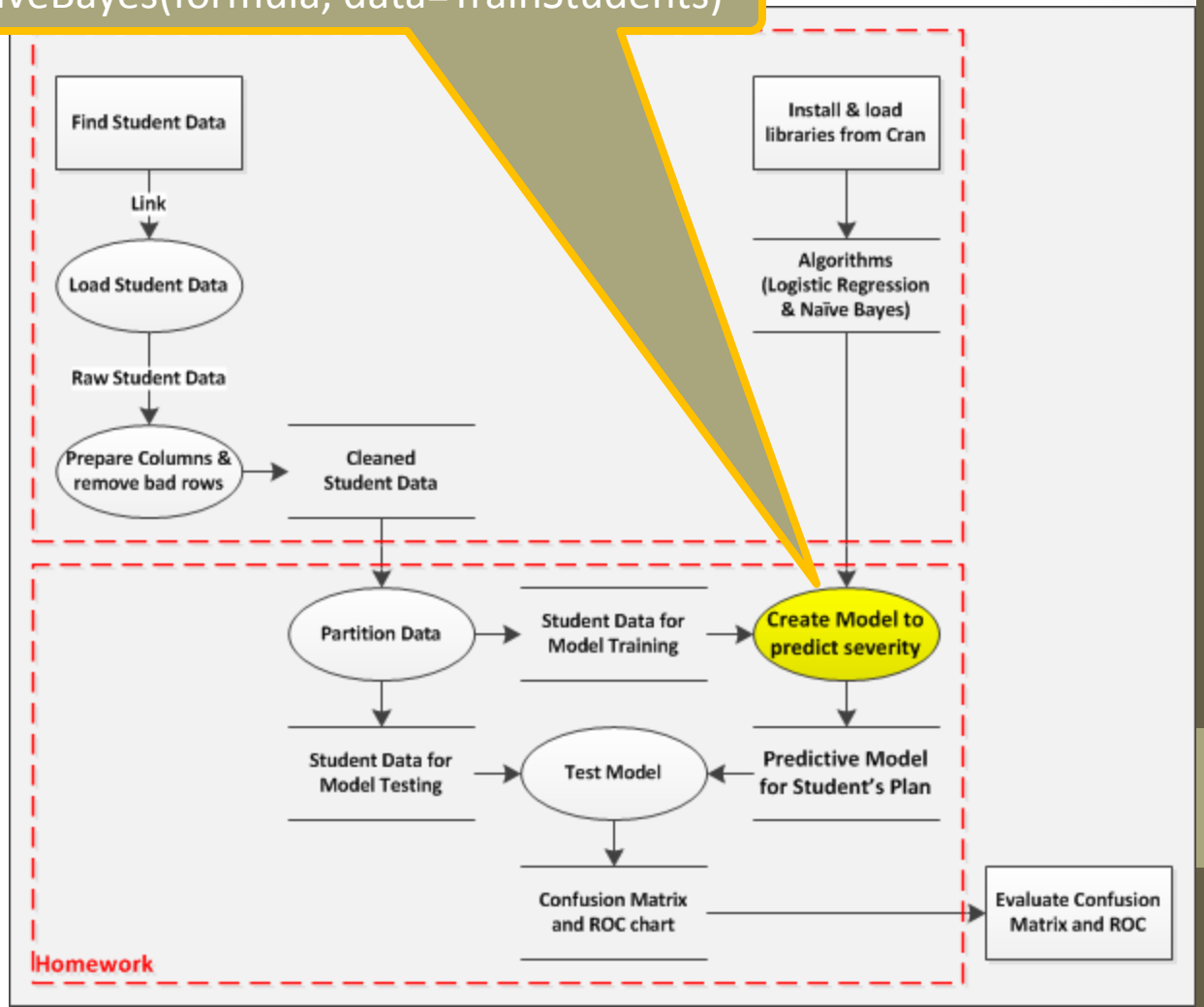
Homework Review: Classifications in R

PartitionFast(Students, fractionOfTest=0.4)



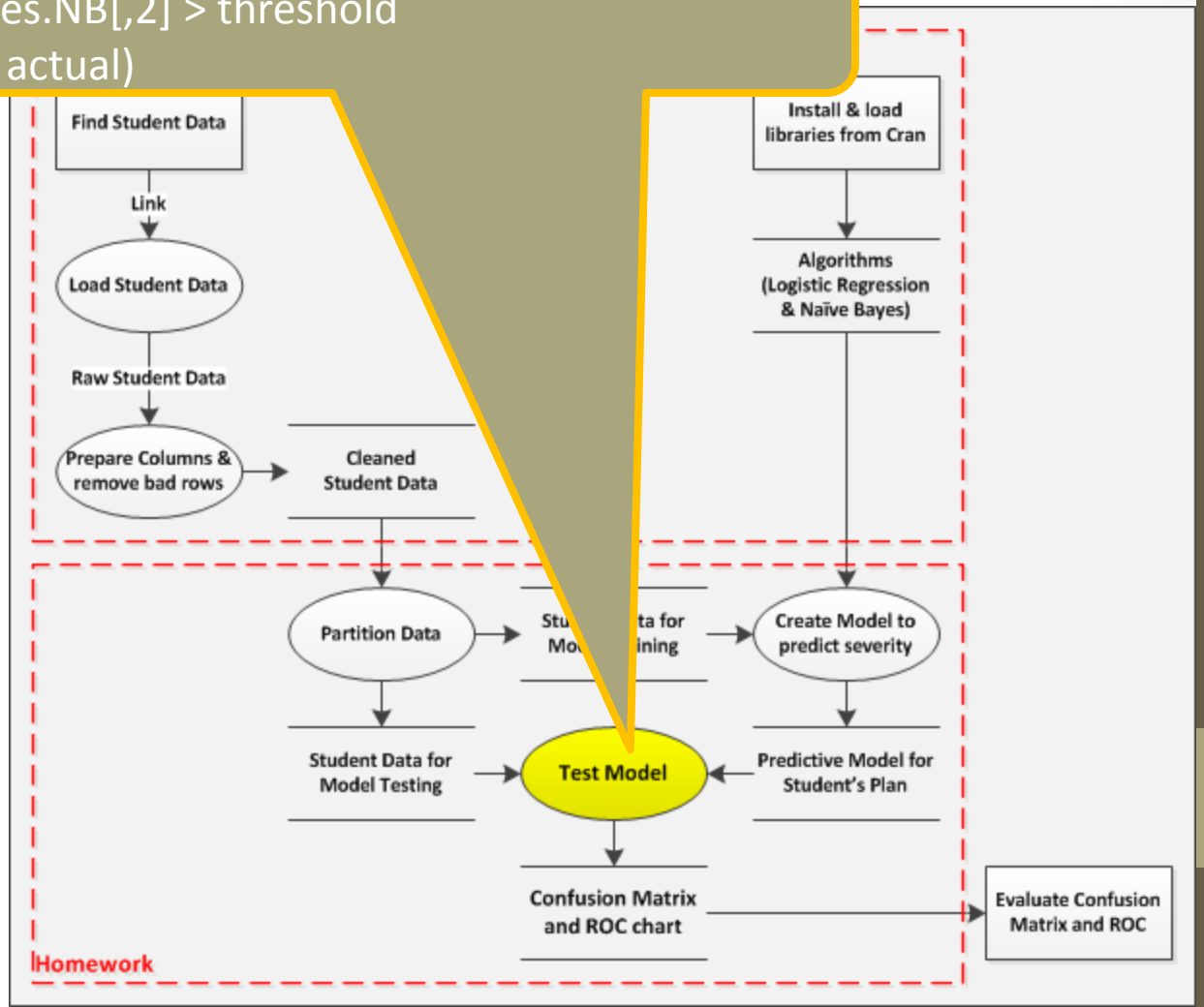
Homework Review: Classifications in R

naiveBayes(formula, data=TrainStudents)

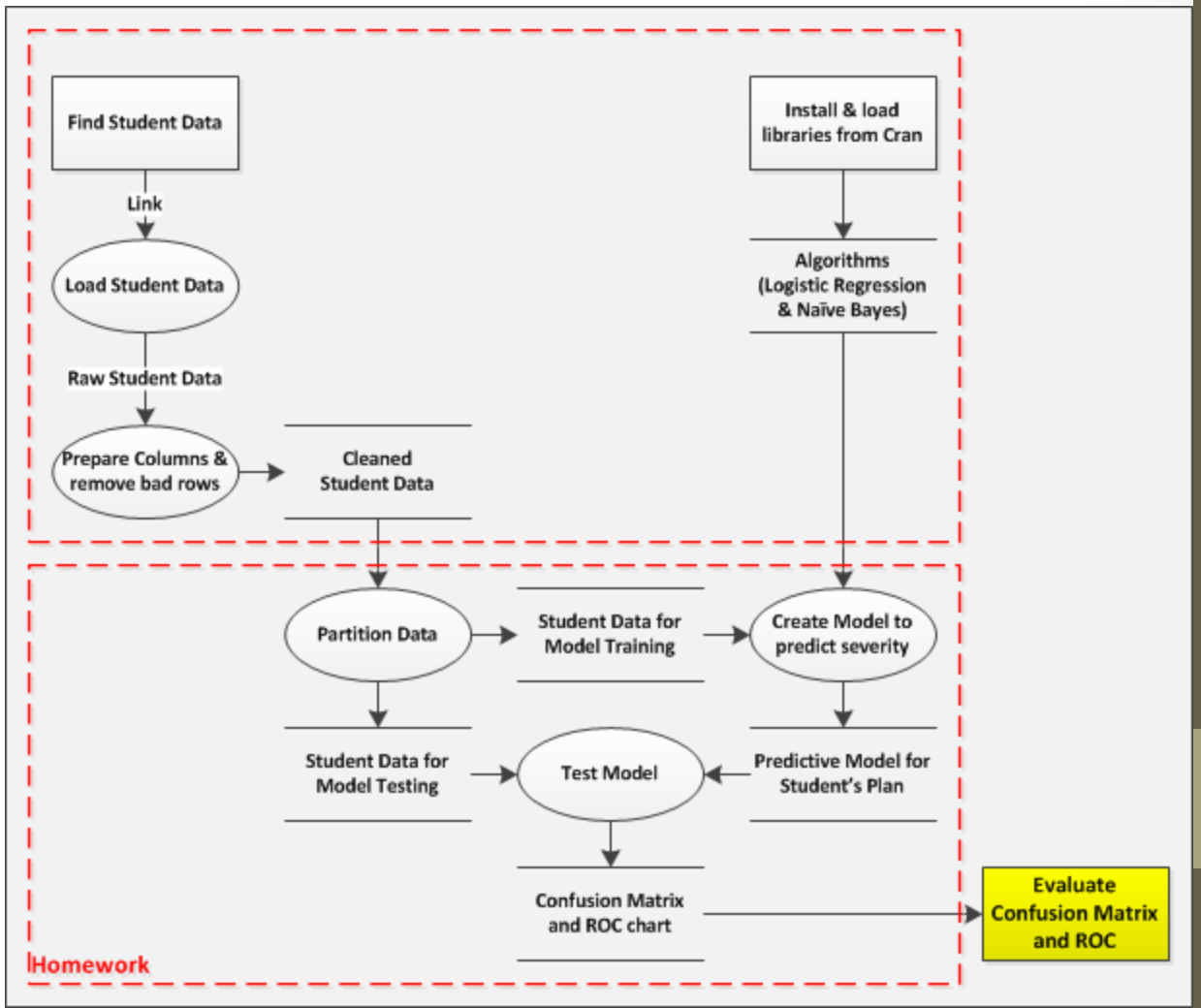


Homework Review: Classifications in R

```
predict(naiveBayesModel, newdata=TestStudents, type="raw")
predictedProbabilities.NB[,2] > threshold
table(predicted.NB, actual)
```



Homework Review: Classifications in R



Homework Review: Classifications in R

- See: today's versions of:
  - ClassifyStudents_complete.R
  - CollegeStudentsDatasets_complete.R

- Partitioning was tested with:
  - PartitionTestFunctions.R

Quiz 05a Classification

- For the last questions in this quiz you will need to download the R-script Quiz05a_Classification.R from Canvas. That R-script will download the required data. The data are available from Canvas, too.

- You can answer the first 6 questions without that R script or using R

Homework Review: Classifications in R

Break

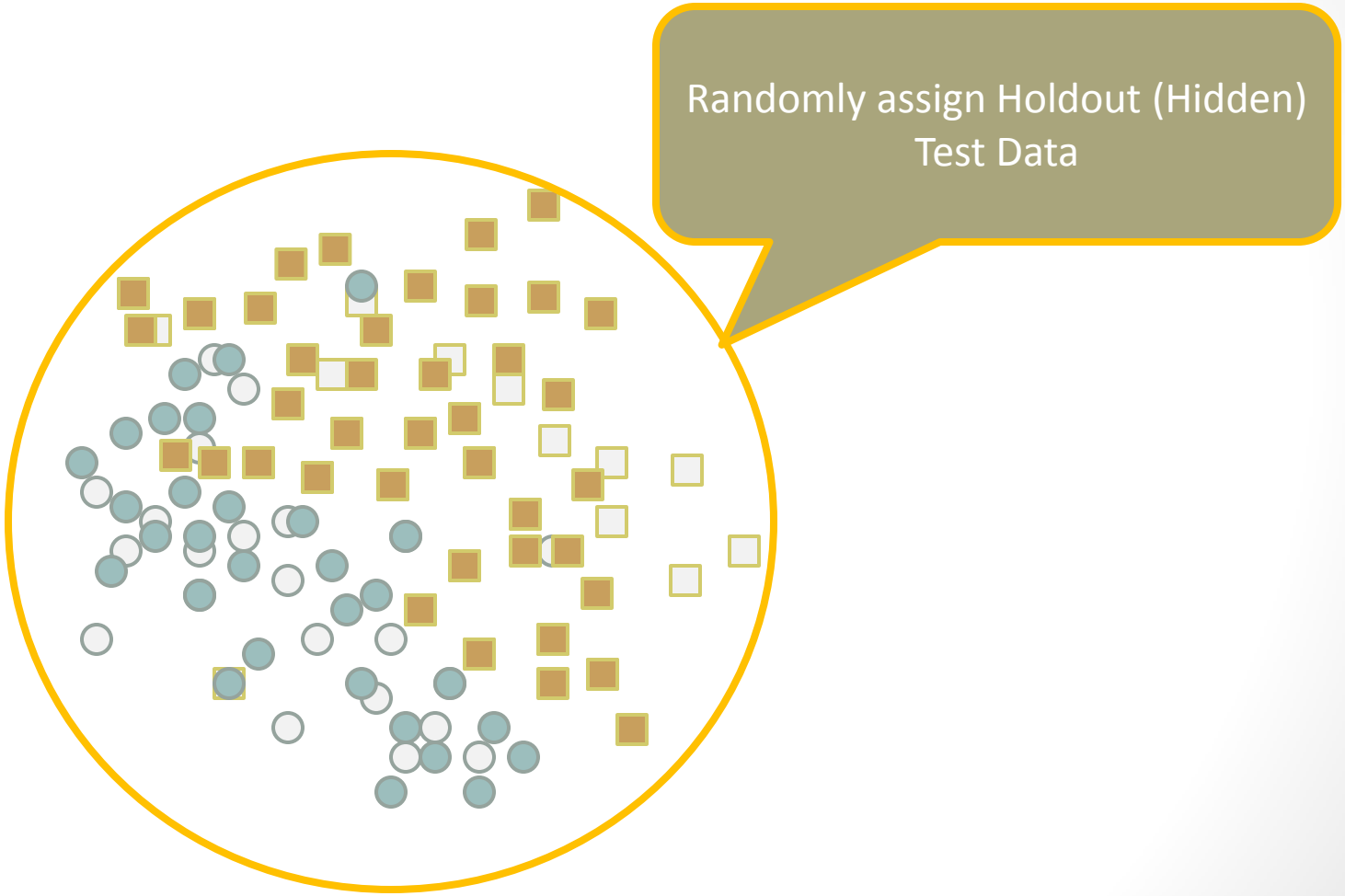
Over-fitting and Confusion Matrix

Evaluate Model

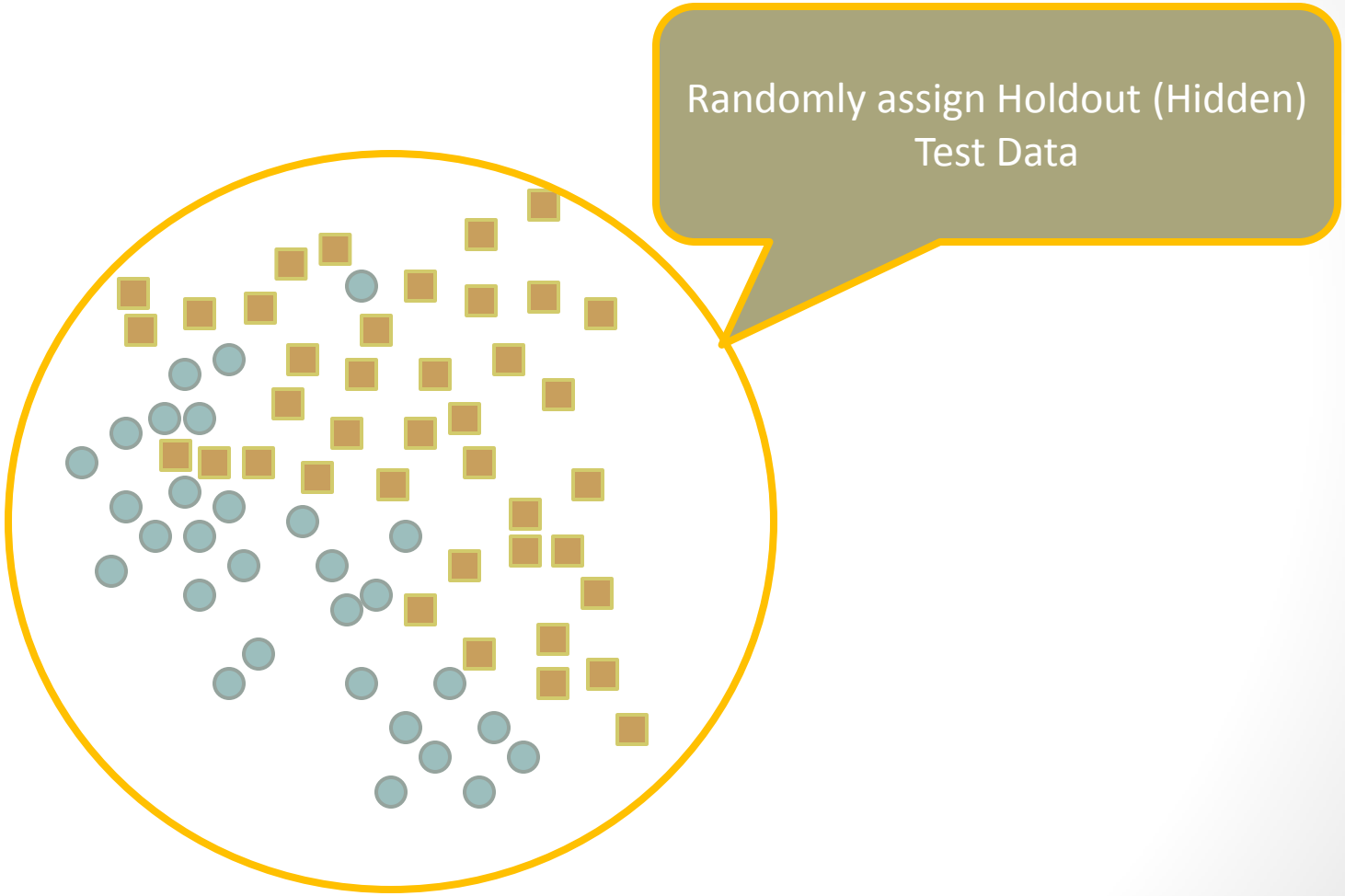
- The following segment will use an over-fitting example to explain the following concepts:
  - Modeling Data
    - Training Data
    - Test Data
  - Model (Hypothesis)
  - Over-fitting
  - Model Accuracy
  - Confusion Matrix (Classification Matrix)
    - True Positive
    - False Positive
    - True Negative
    - False Negative

Evaluate Model: All Data

Evaluate Model: Test Data

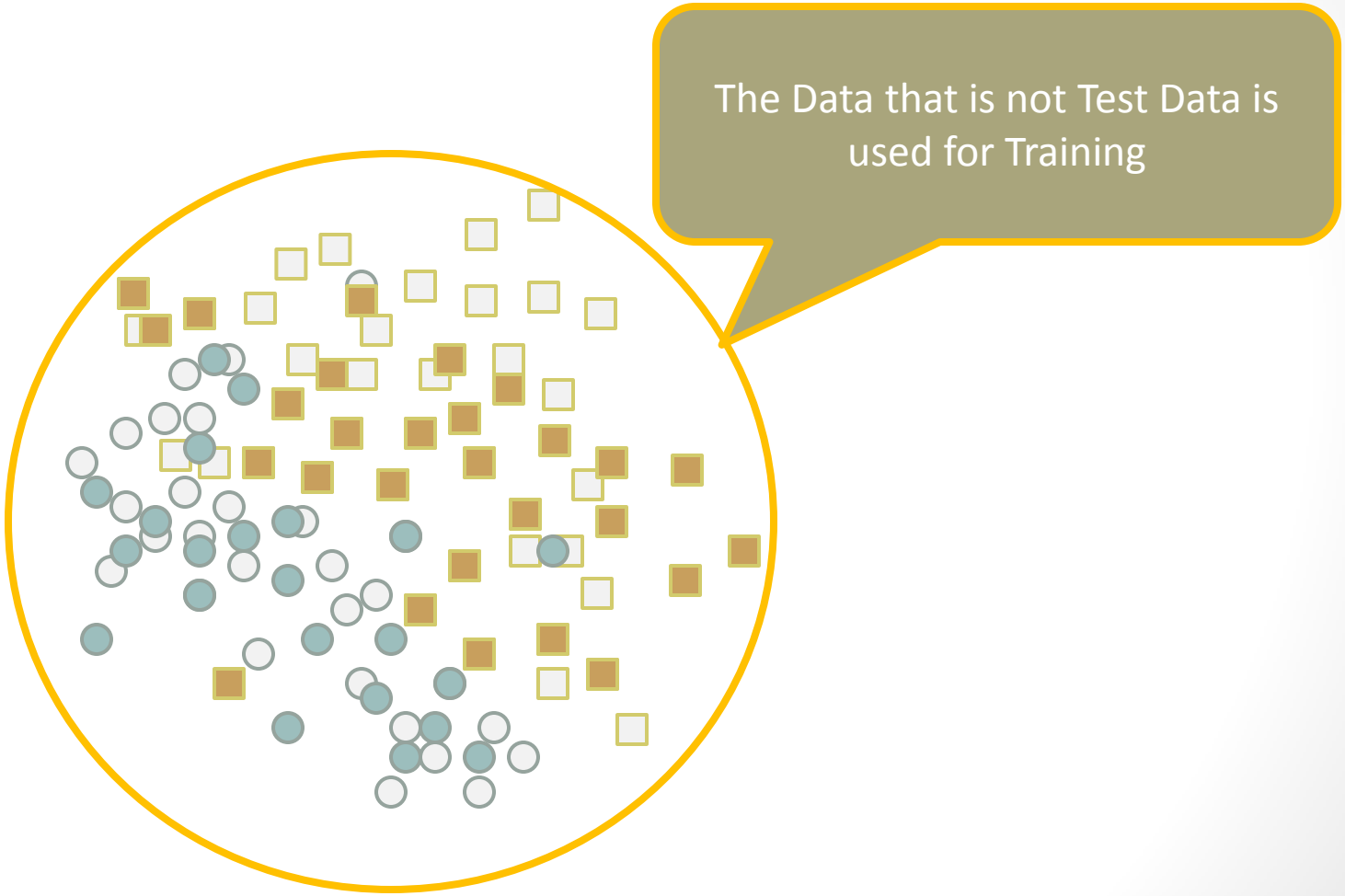


Evaluate Model: Test Data

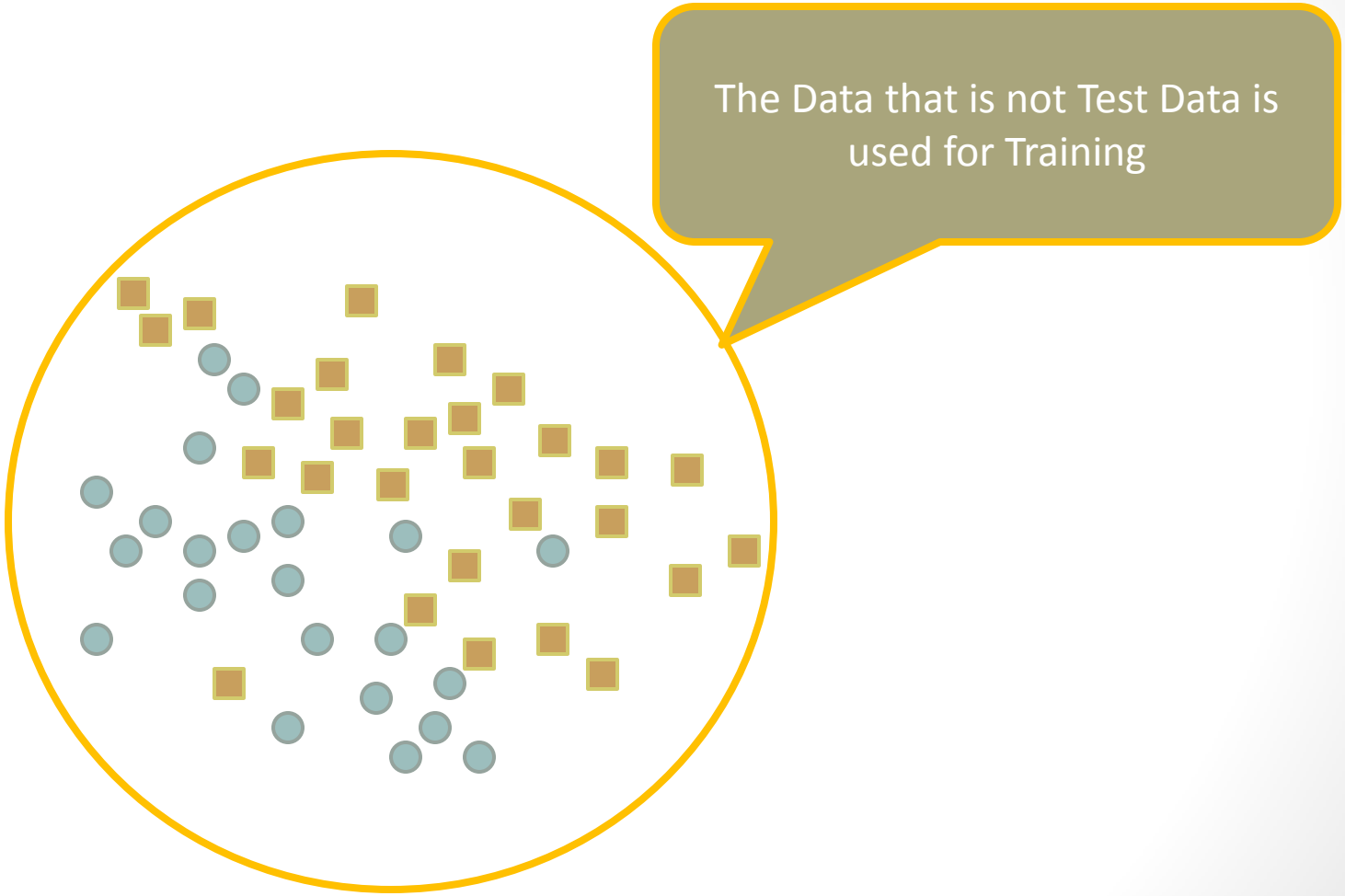


Evaluate Model: All Data

Evaluate Model: Training Data

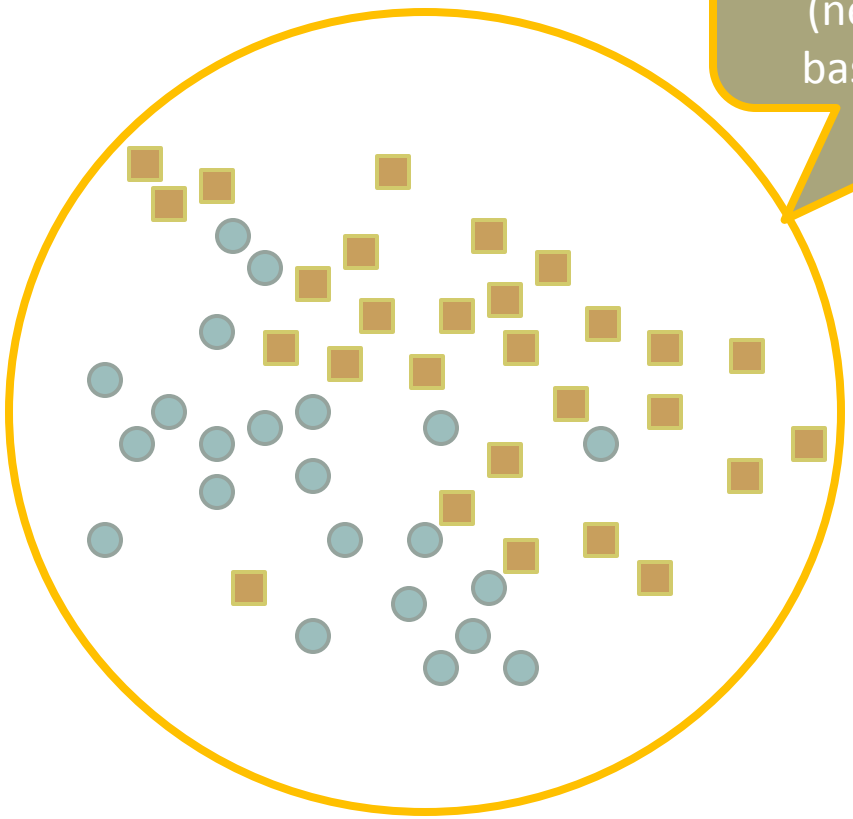


Evaluate Model: Training Data

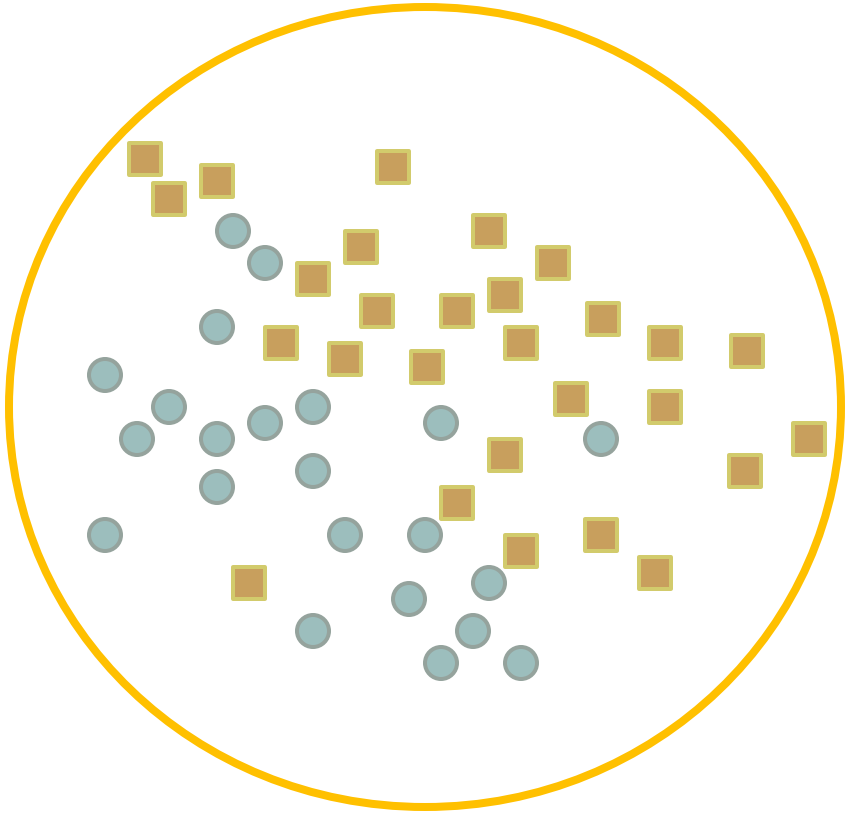


Evaluate Model: Training

I want to predict if a point is a square (positive) or a circle (negative). The prediction is based on the point's location



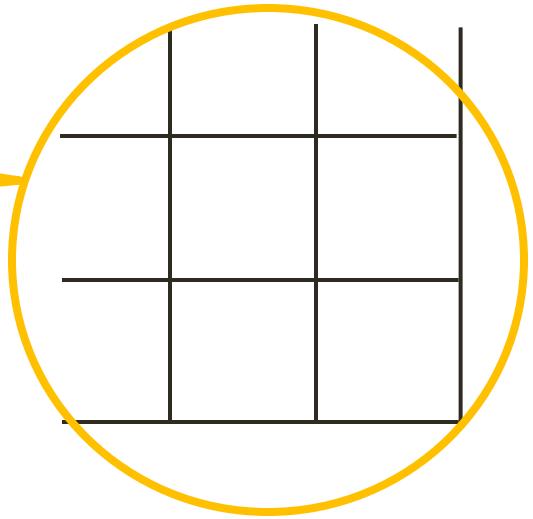
Evaluate Model: Training

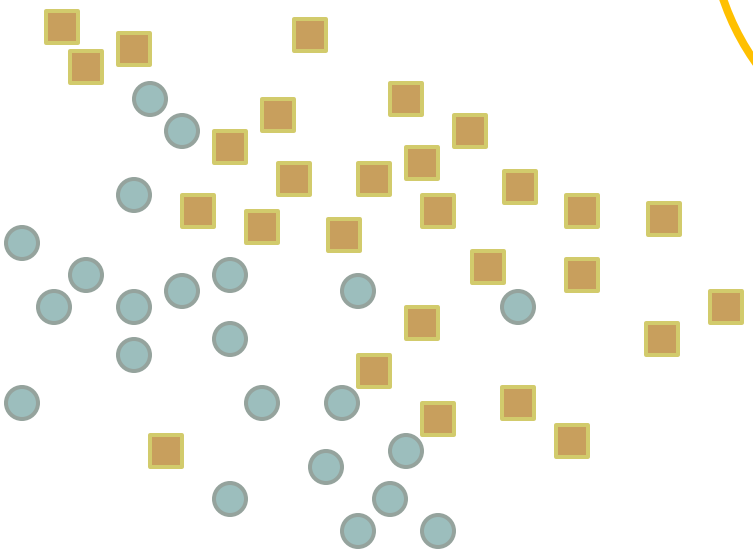


isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

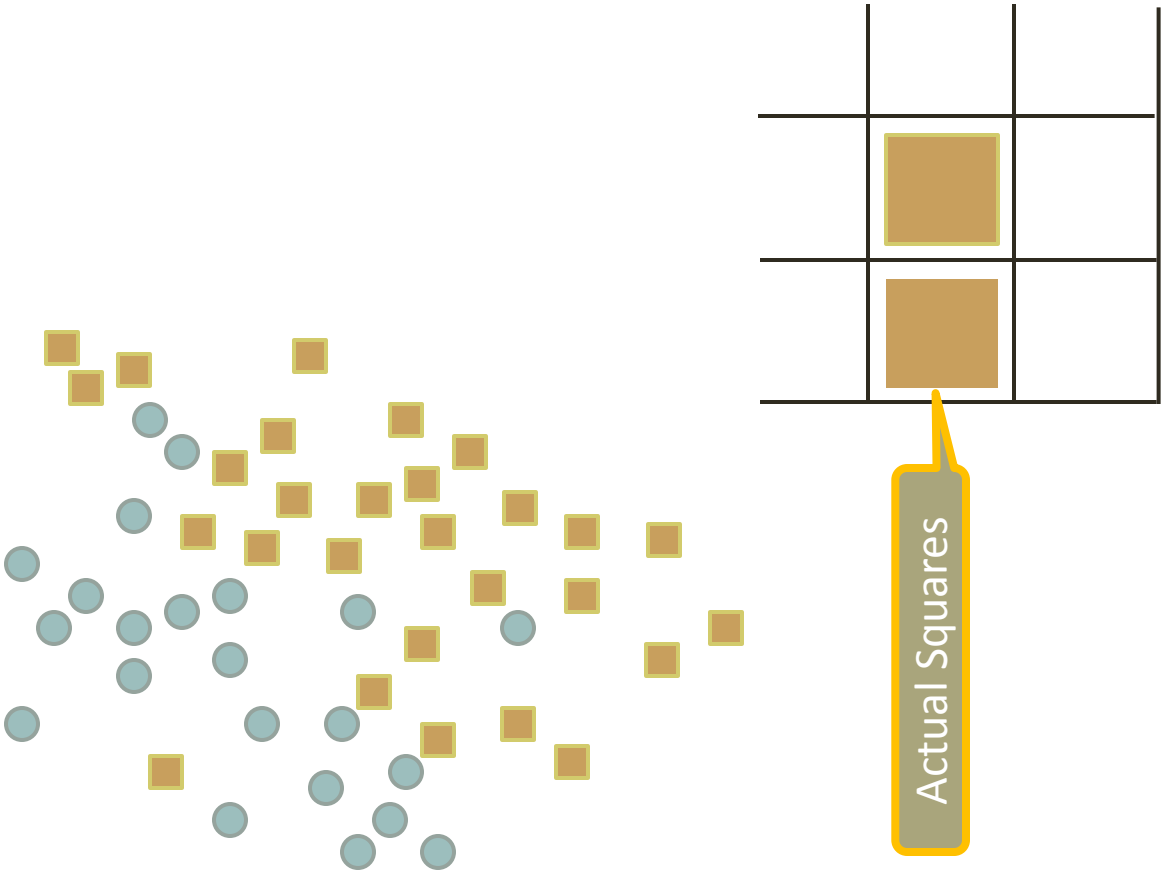
Confusion Matrix (Classification Matrix): Compare Squares and Circles with Predicted Squares and Circles





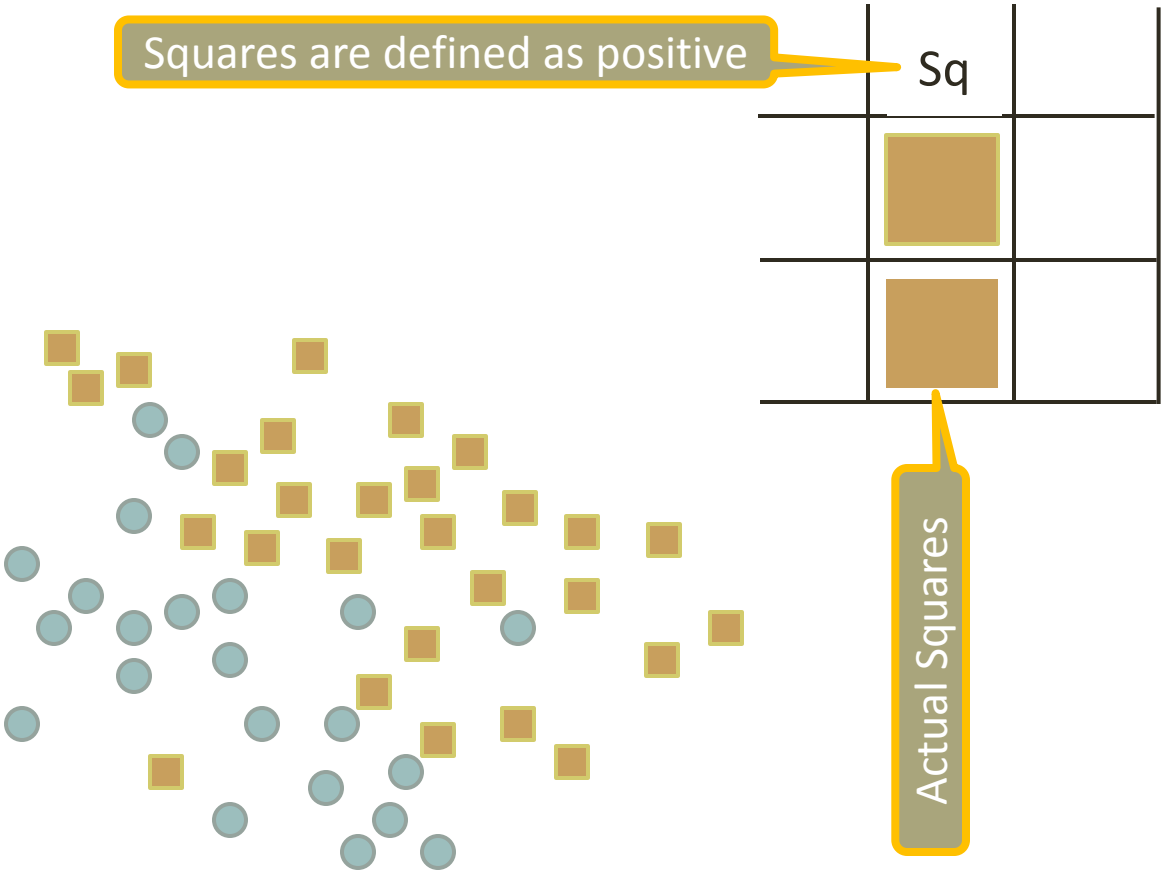
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



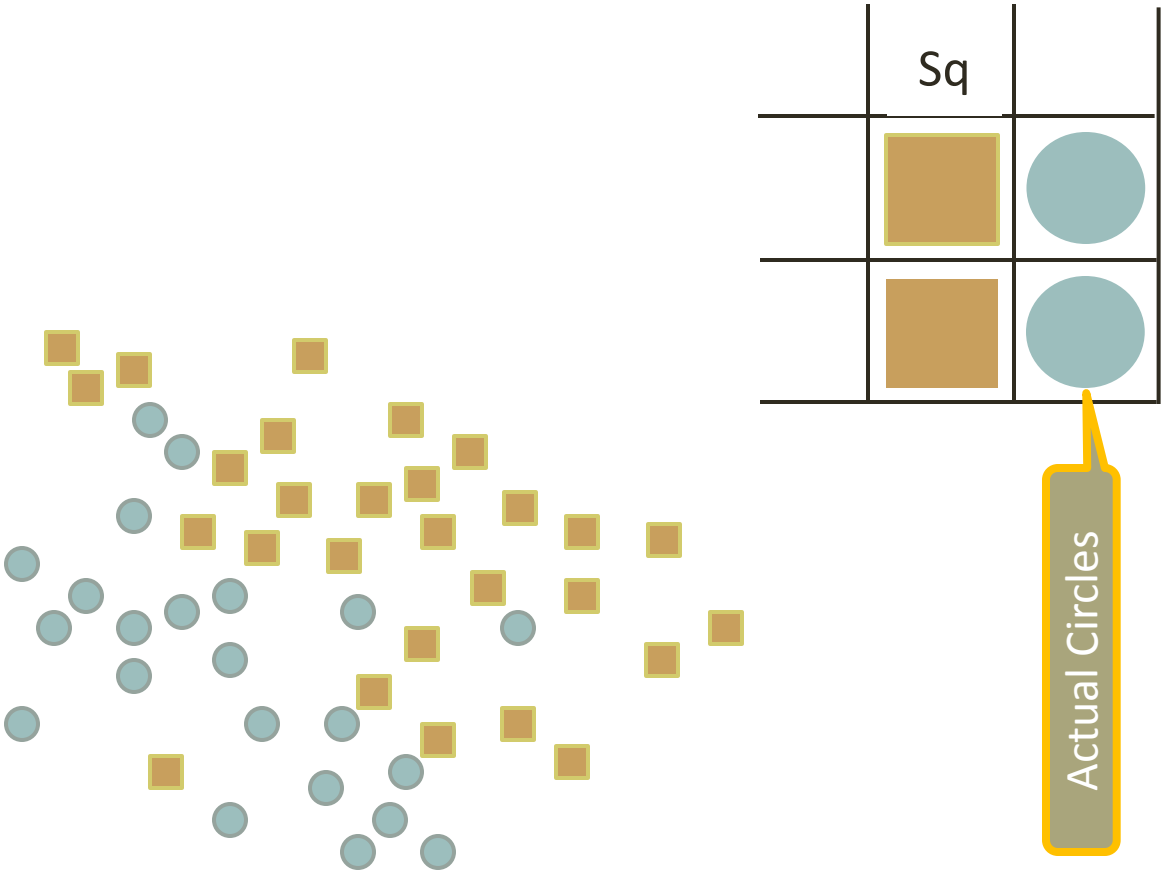
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



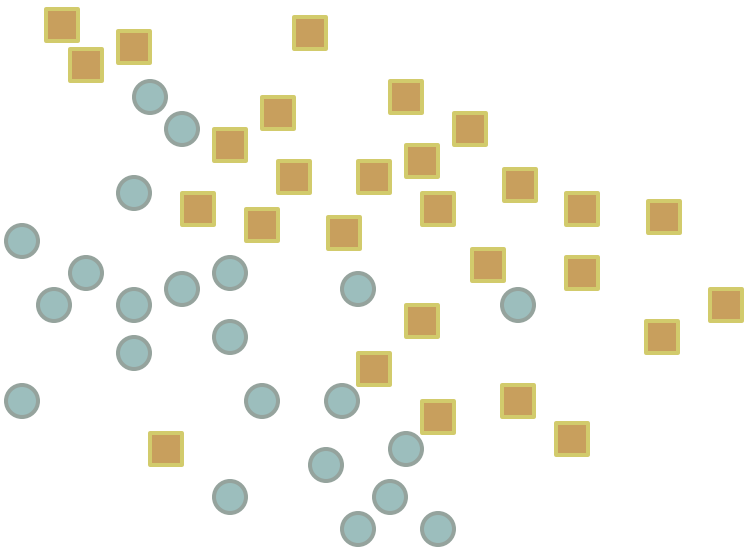
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



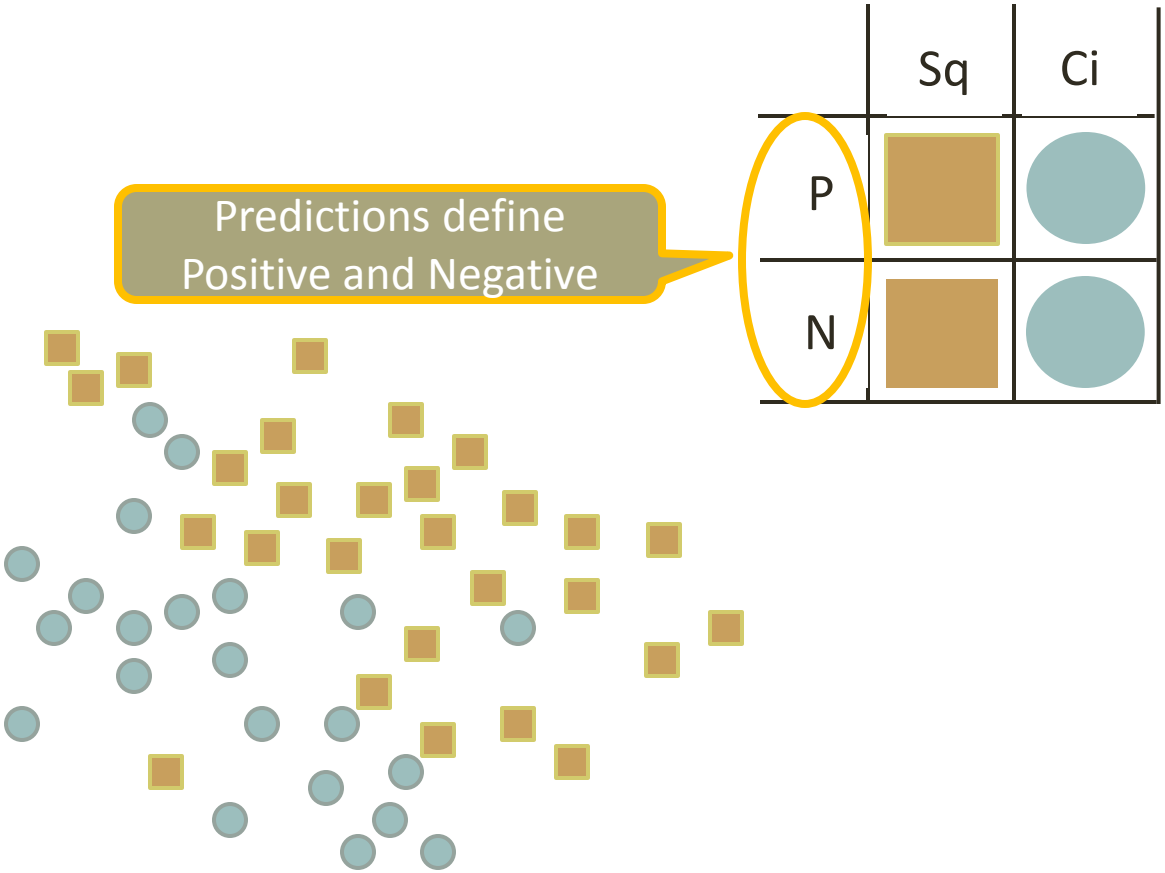
Circles are defined as negative

	Sq	Ci
	■	●
	■	●

Actual Circles

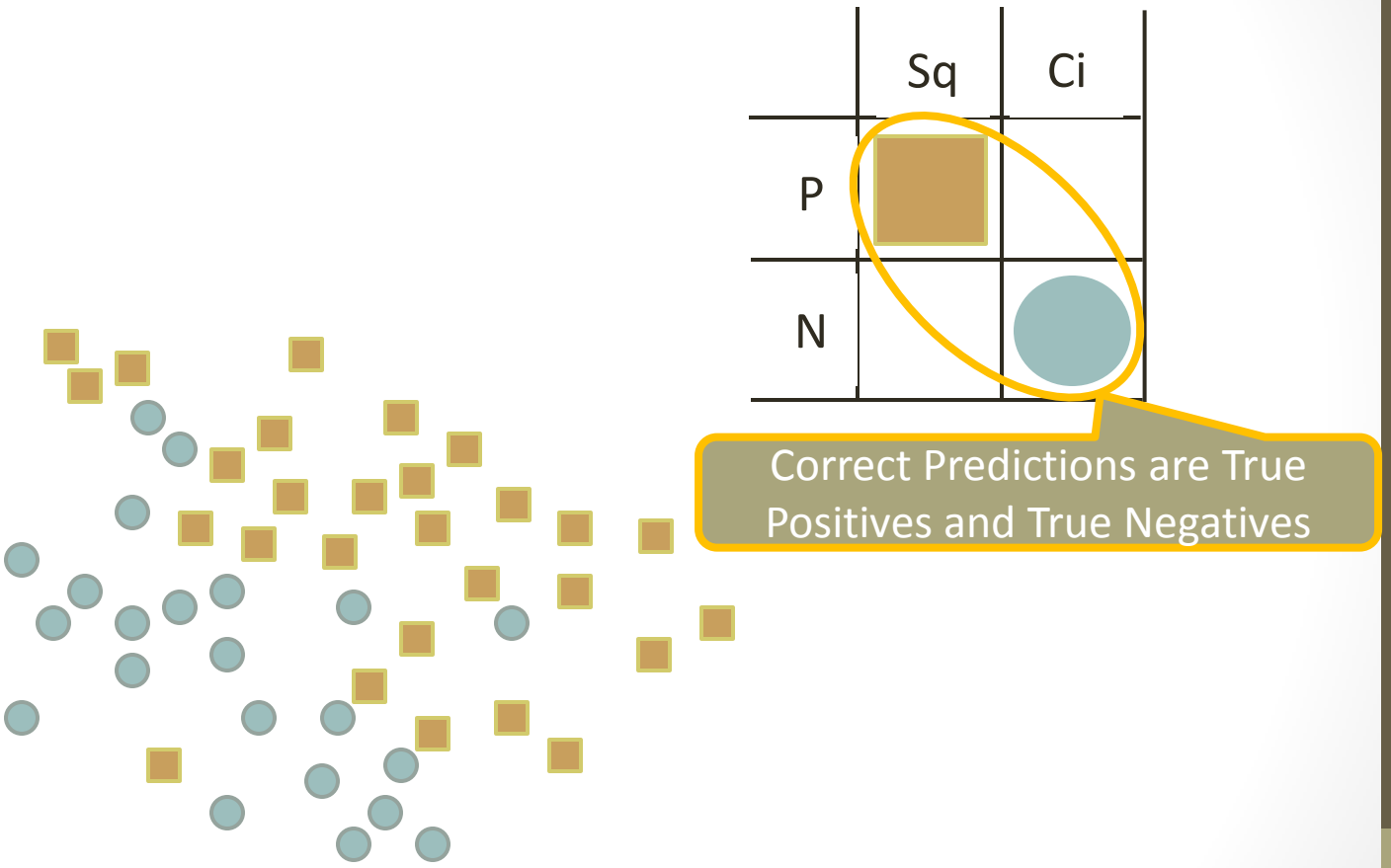
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



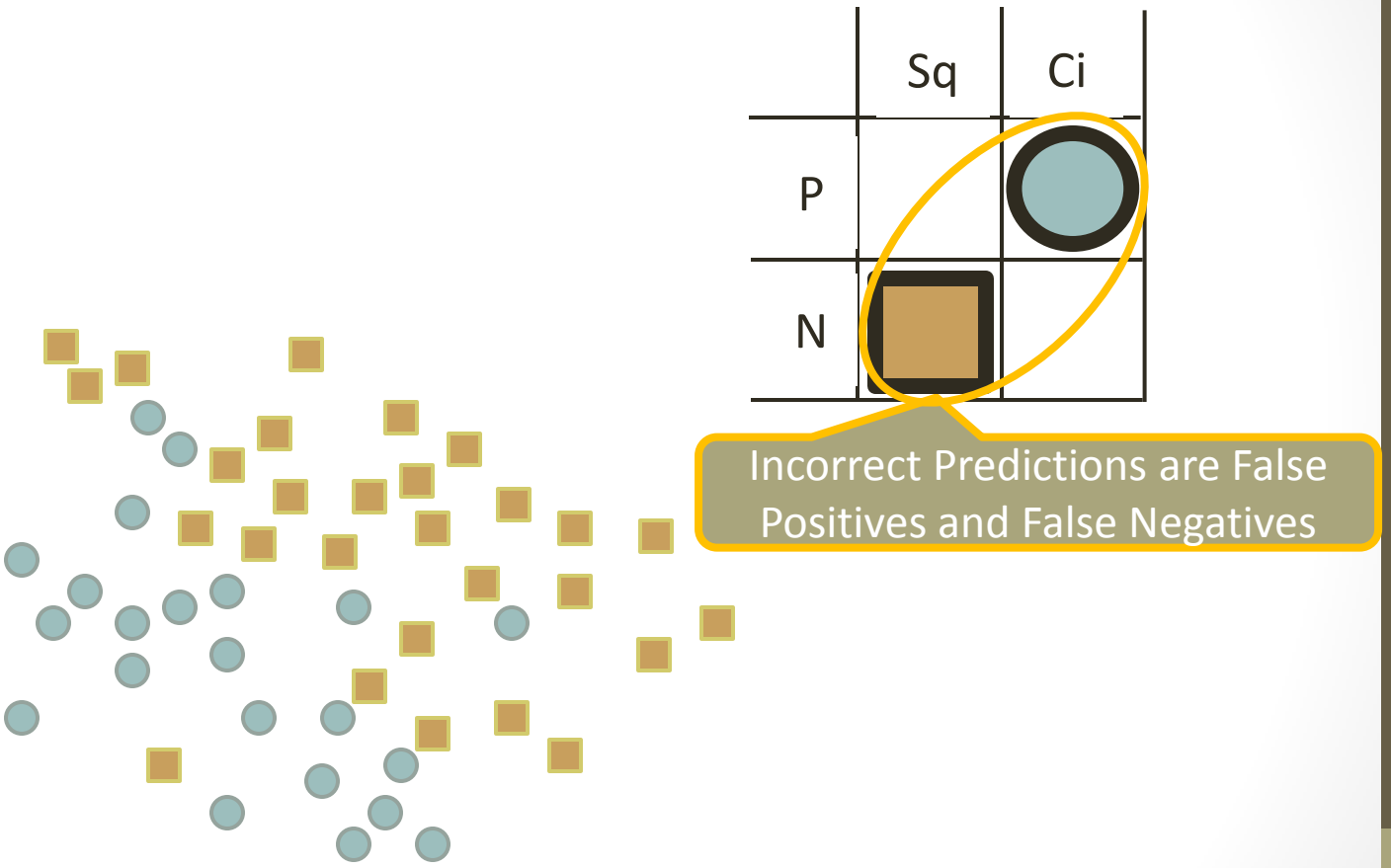
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

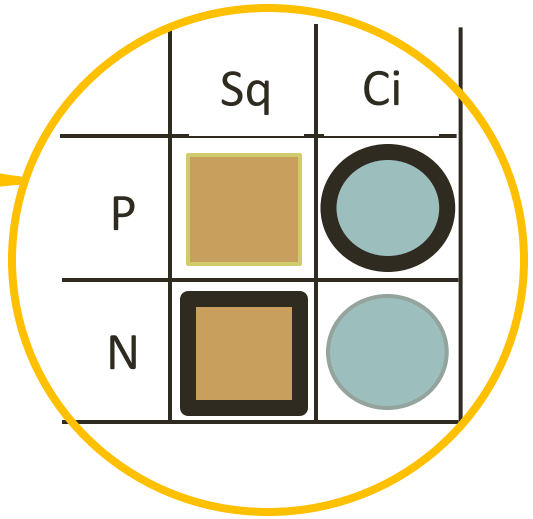
Evaluate Model: Confusion Matrix

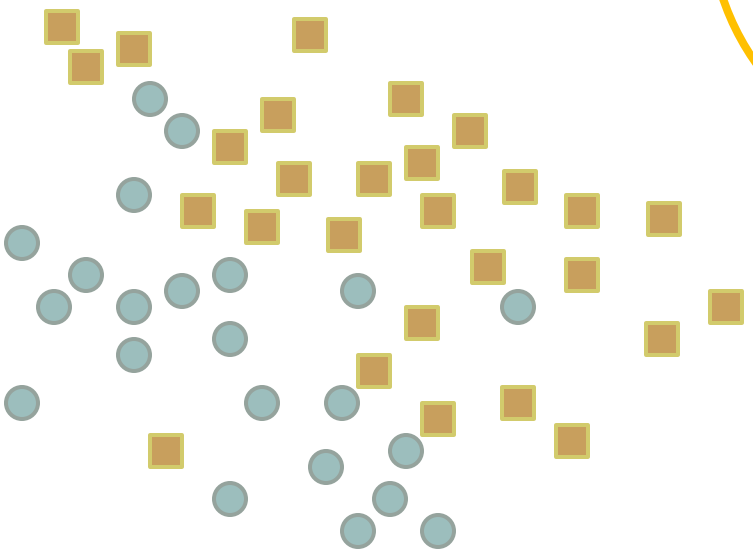


isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

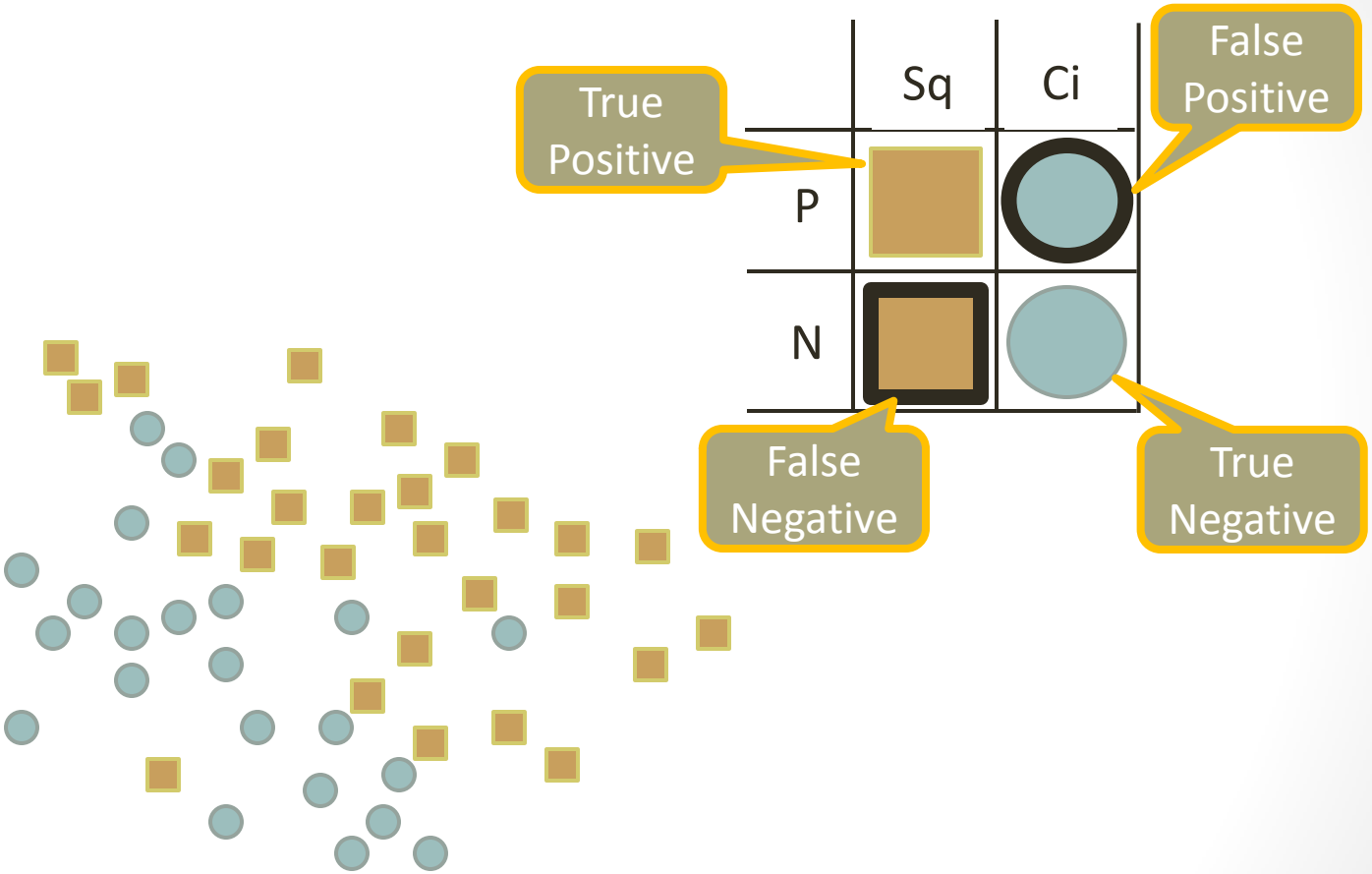
Confusion Matrix (Classification Matrix):
Vertical are actual classes
Horizontal are predicted classes





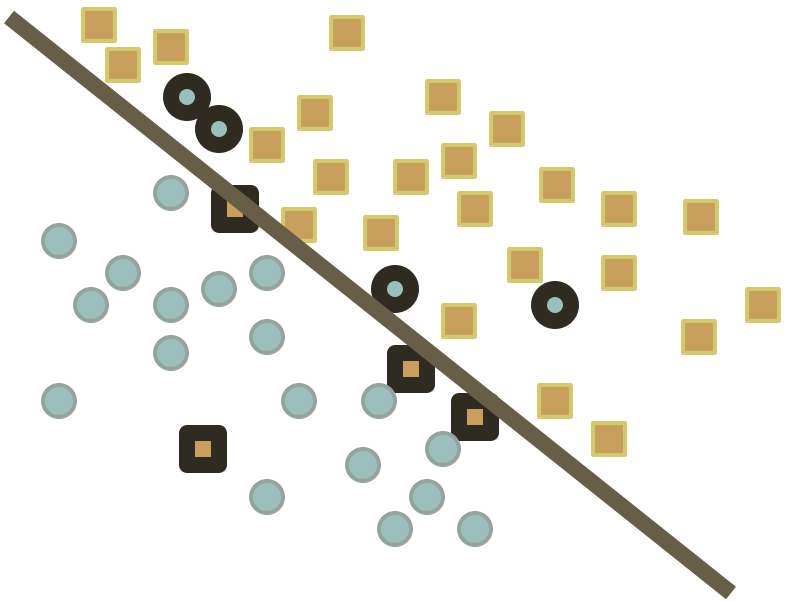
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

Evaluate Model : Train Model 1

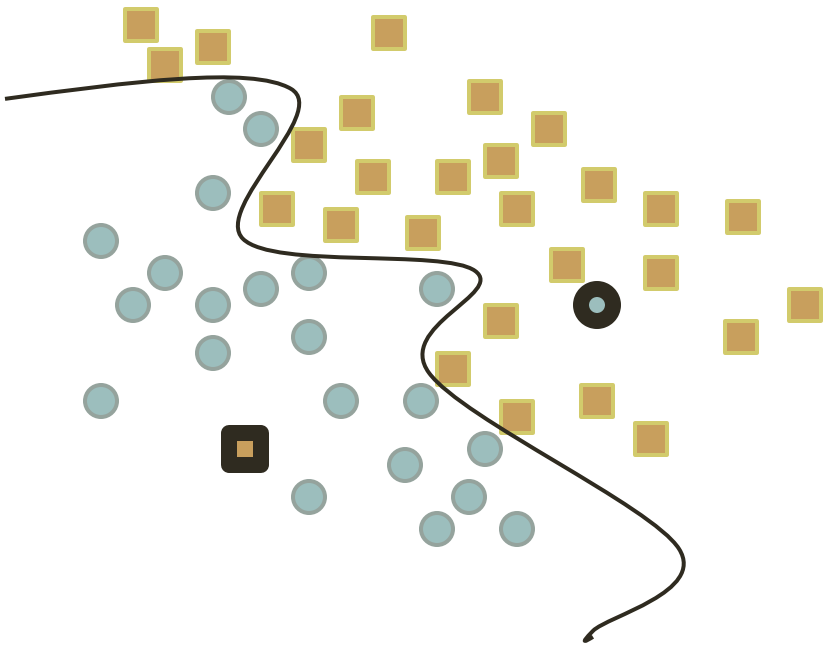


	Sq	Ci
P	36	4
N	4	26

isSquare ~ xLocation + yLocation

Evaluate Model : Train Model 2

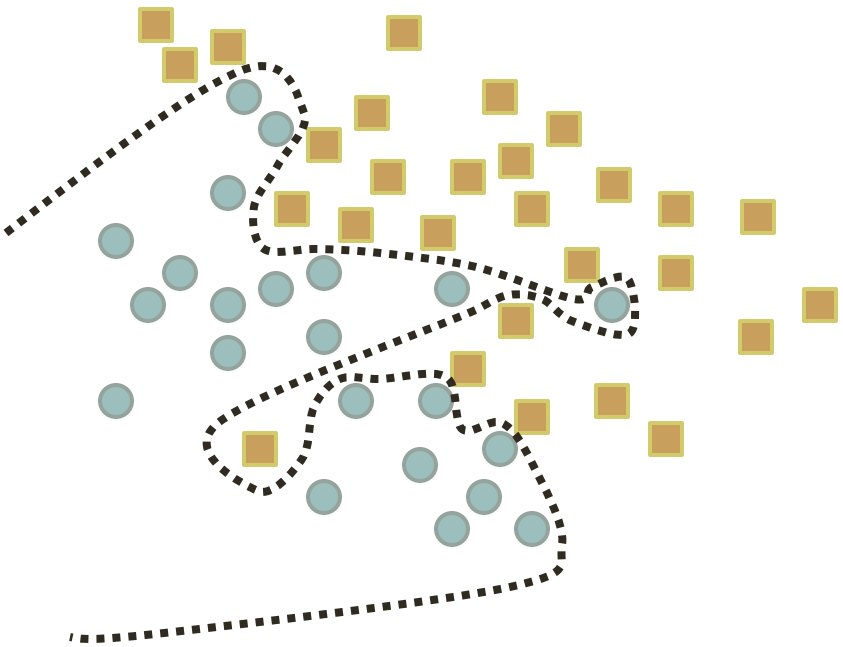
	Sq	Ci
P	39	1
N	1	29



isSquare ~ xLocation + yLocation

Evaluate Model : Train Model 3

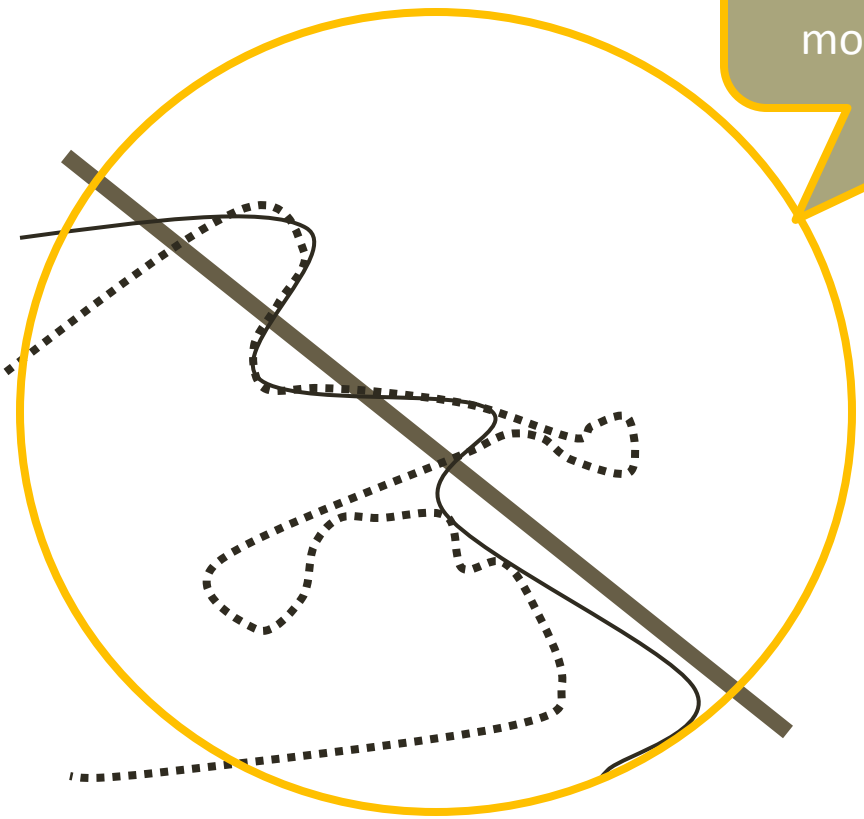
	Sq	Ci
P	40	0
N	0	30



isSquare ~ xLocation + yLocation

Evaluate Model : 3 Models

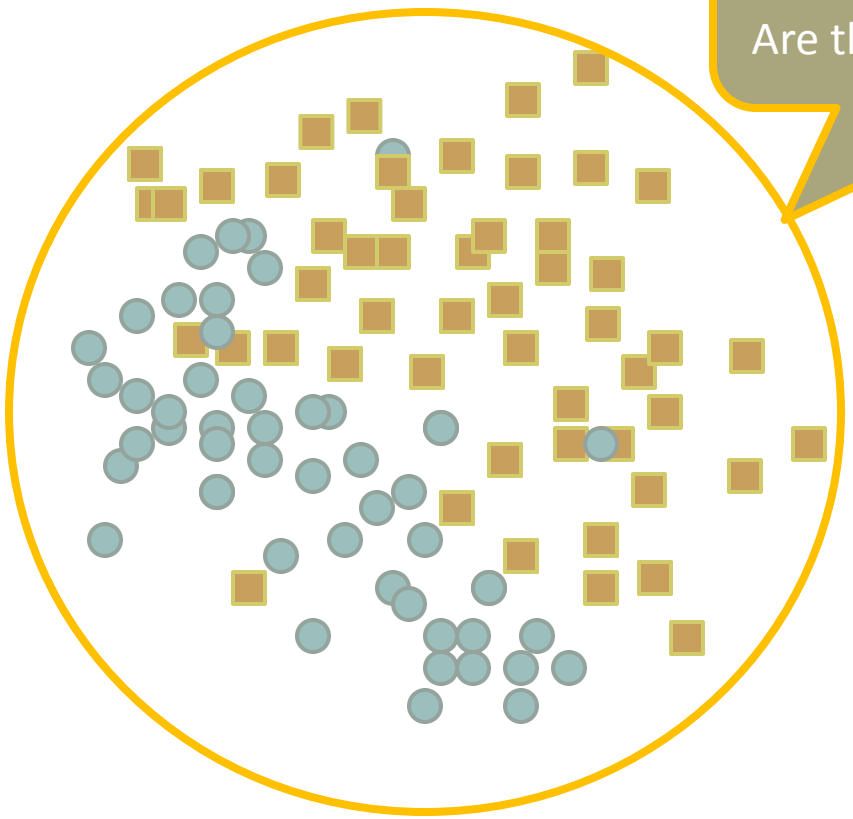
These models are based on training data. In these cases, models are called hypotheses.



isSquare ~ xLocation + yLocation

Evaluate Model : All Data

Training data overlaid on test data.
Visual comparison of data sets.
Are the distributions comparable?

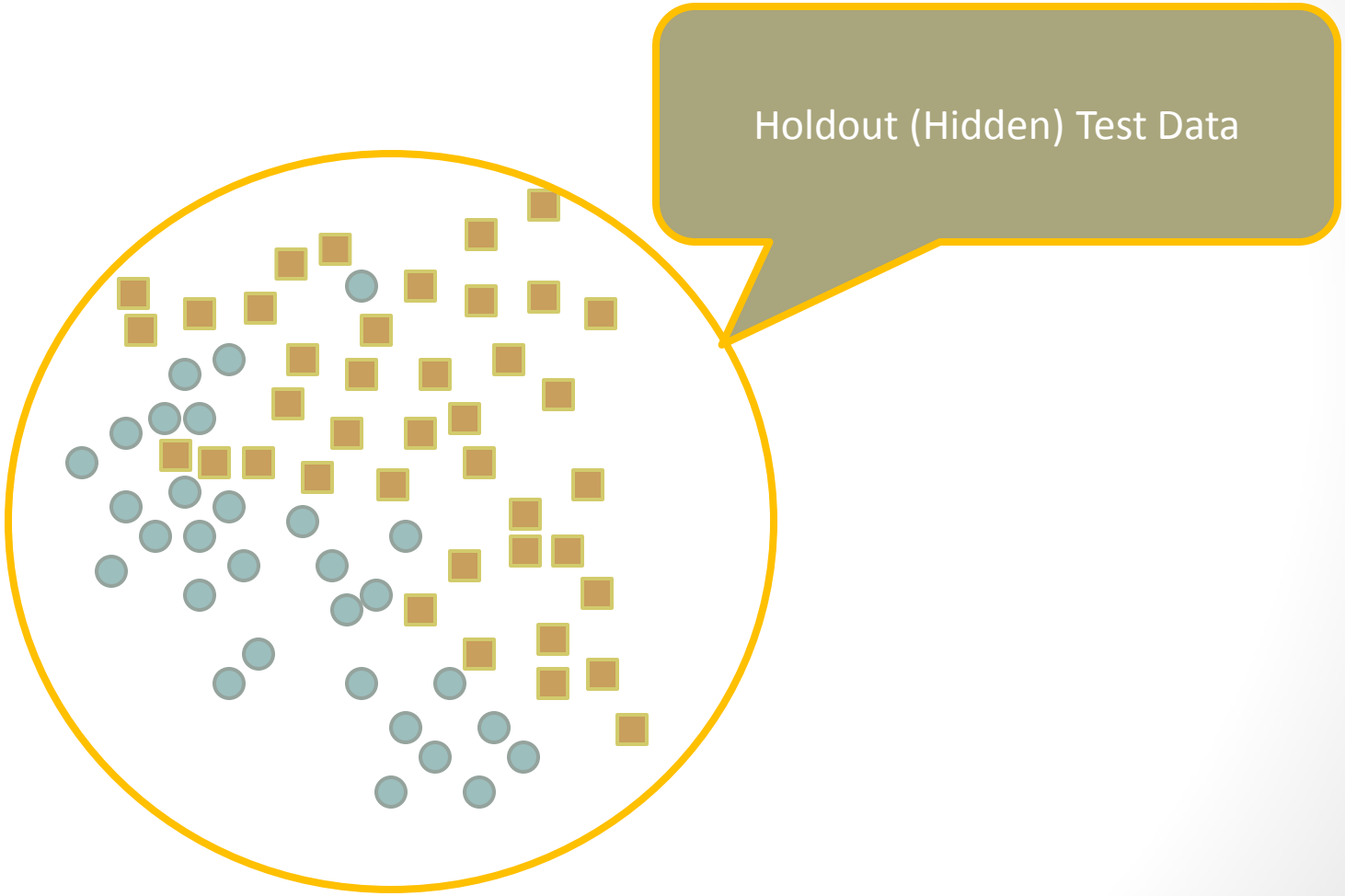


isSquare ~ xLocation + yLocation

Evaluate Model : Training Data

isSquare ~ xLocation + yLocation

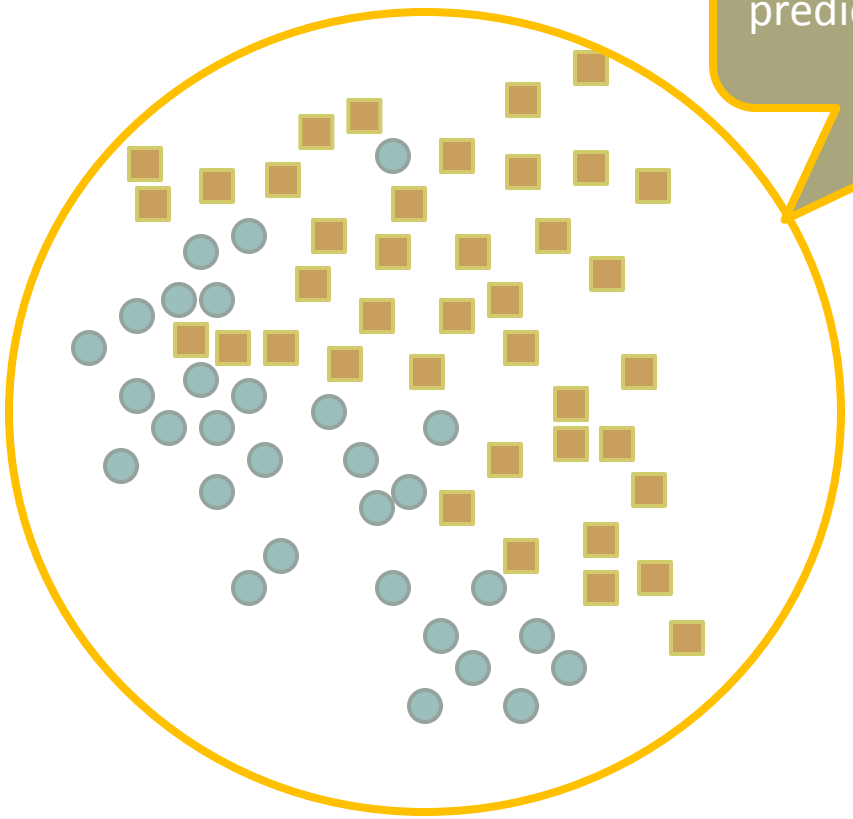
Evaluate Model : Test Data



isSquare ~ xLocation + yLocation

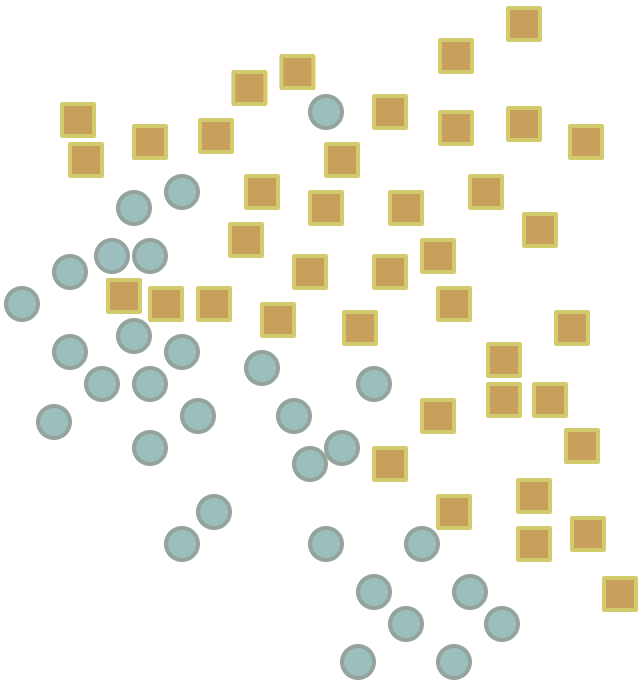
Evaluate Model : Test Data

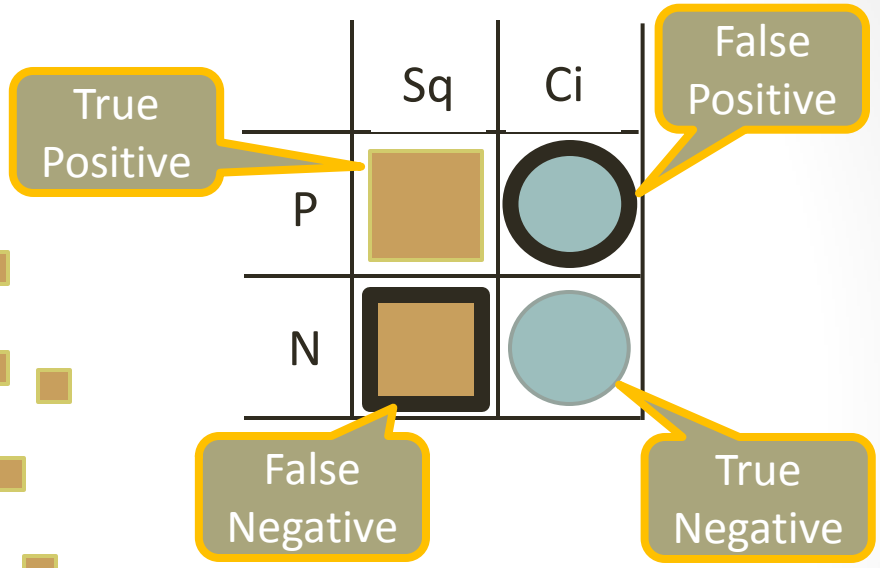
In the test data set:
I want to test if a square is predicted as positive and if a circle is predicted as negative



isSquare ~ xLocation + yLocation

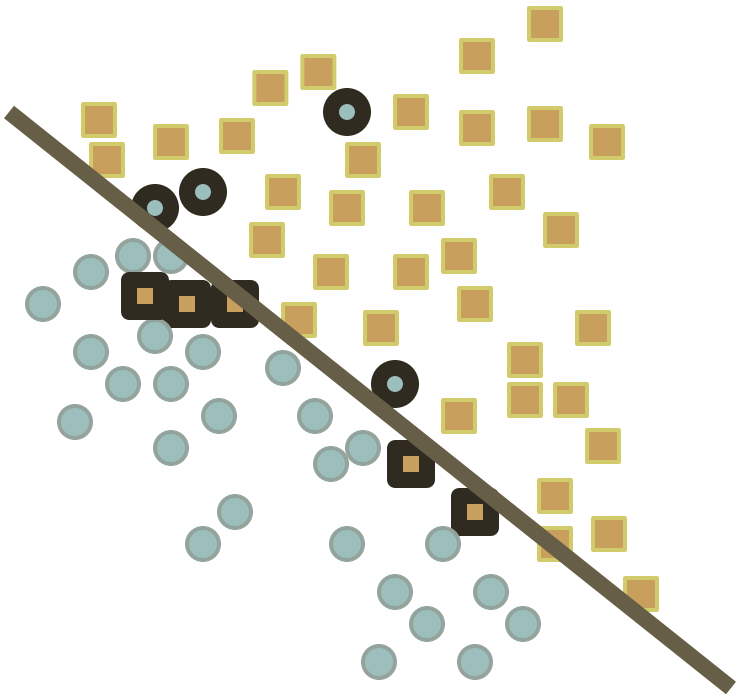
Evaluate Model : Test Data





isSquare ~ xLocation + yLocation

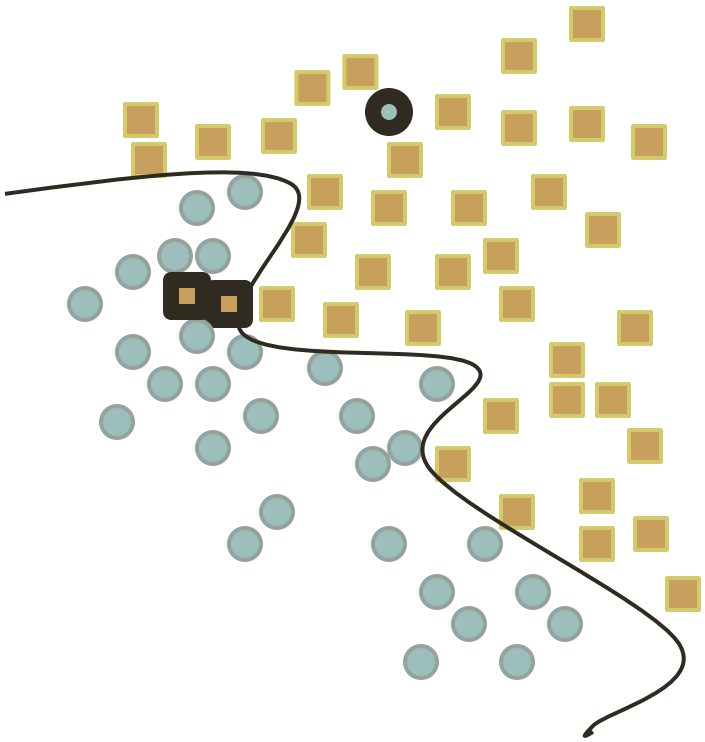
Evaluate Model : Test Model 1



	Sq	Ci
P	35	4
N	5	26

isSquare ~ xLocation + yLocation

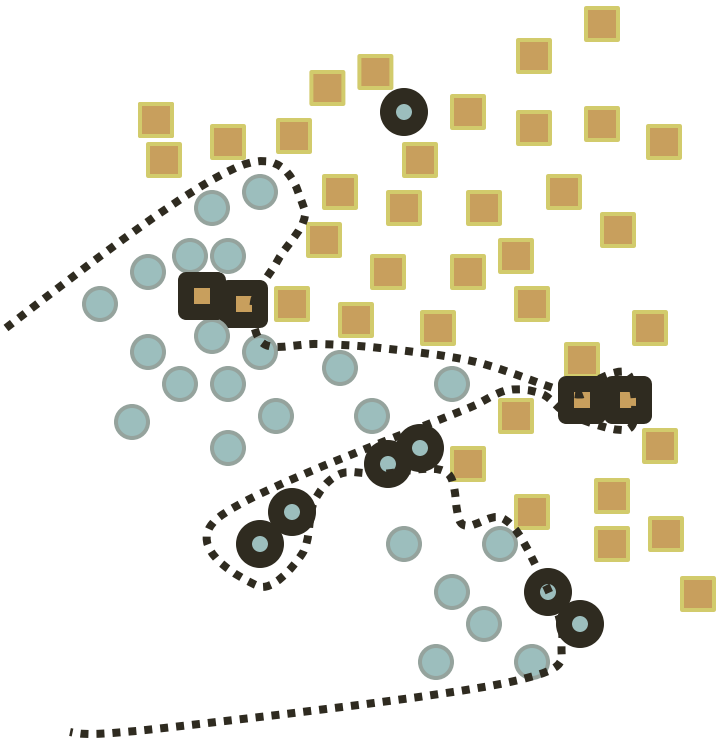
Evaluate Model : Test Model 2



	Sq	Ci
P	38	1
N	2	29

isSquare ~ xLocation + yLocation

Evaluate Model : Test Model 3



	Sq	Ci
P	36	7
N	4	23

isSquare ~ xLocation + yLocation

Relate a Confusion Matrix with an ROC chart

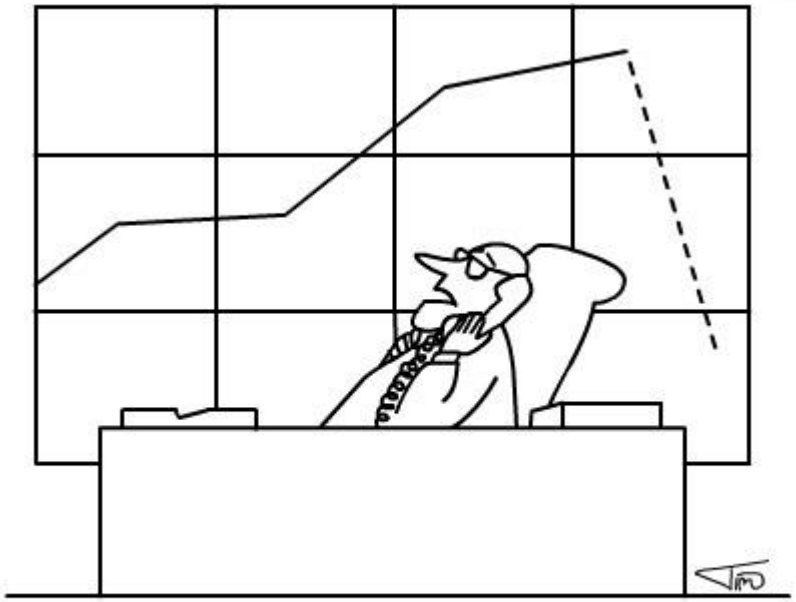
- Optional demo: Open up a synced Classification (Confusion) Matrix and ROC chart.
  - Set the threshold of the Classification Matrix to 0, 0.5, and 1. How do these thresholds compare to the FPR and TPR on the ROC chart?
  - Set the FPR on the ROC chart to 0, 0.5, and 1. What are the TPR on the ROC chart? How does the threshold of the classification matrix change?
  - Open up a cost chart. Set the readmission penalty to 3X the cost of the intervention cost. What is the optimal threshold? What is the FPR?

Over-fitting and Confusion Matrix

Video and Break

- Watch in class this advertisement for IBM's predictive analytics: https://www.youtube.com/watch?v=iY3WRvXVogo

- Another video on predictive policing:
  - https://www.youtube.com/watch?v=pkGhPSoH7Xk

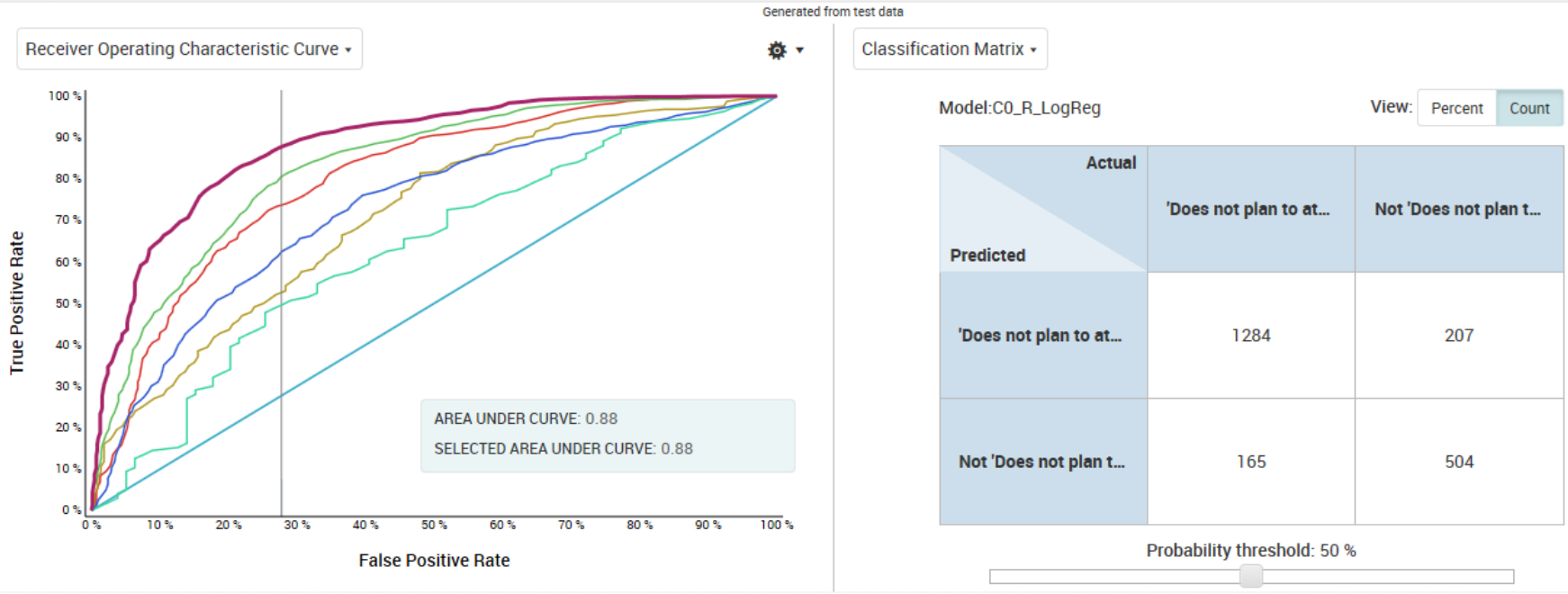


"BI tech support? The predictive analysis system is giving the wrong answer again—can you please fix it?... "

ROC Chart Demo

ROC Chart Demo

- Confusion Matrix and ROC Chart



ROC Chart Demo

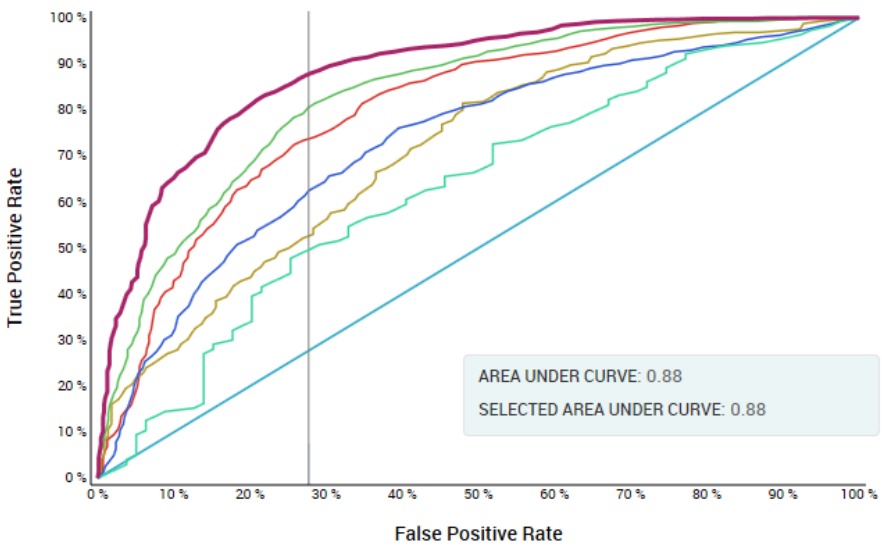
- Confusion Matrix and ROC Chart

Comparison of 6 ROC curves
Each curve is from a different model

The confusion matrix for one model at one threshold

Generated from test data

Receiver Operating Characteristic Curve



Classification Matrix

Model:C0_R_LogReg

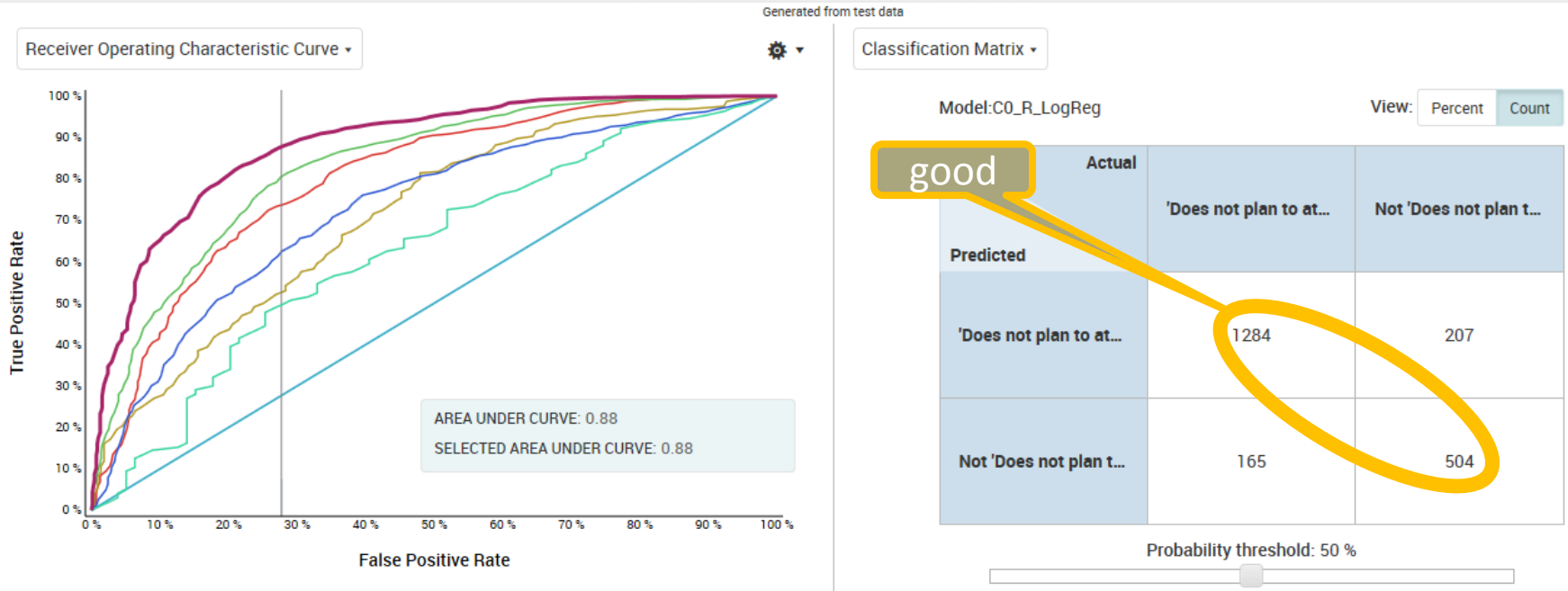
View: Percent Count

Predicted \ Actual	'Does not plan to at...	Not 'Does not plan t...
'Does not plan to at...	1284	207
Not 'Does not plan t...	165	504

Probability threshold: 50 %

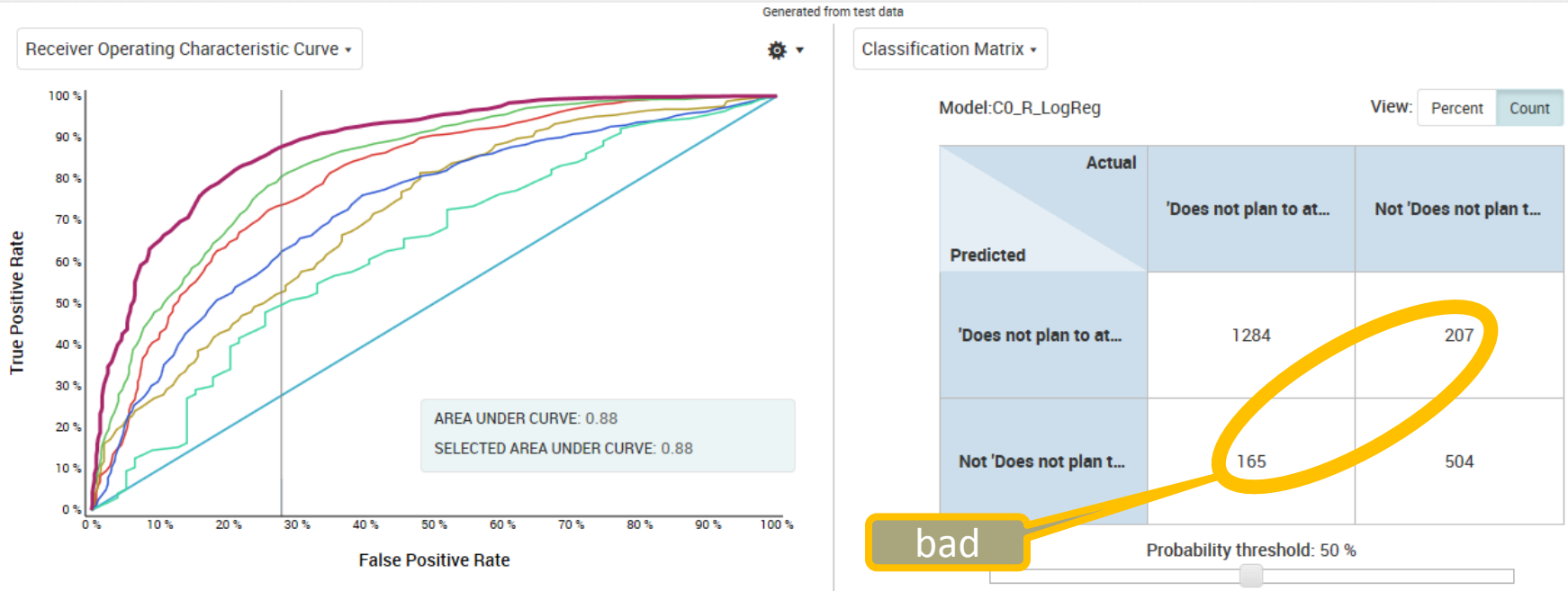
ROC Chart Demo

- Confusion Matrix and ROC Chart



ROC Chart Demo

- Confusion Matrix and ROC Chart

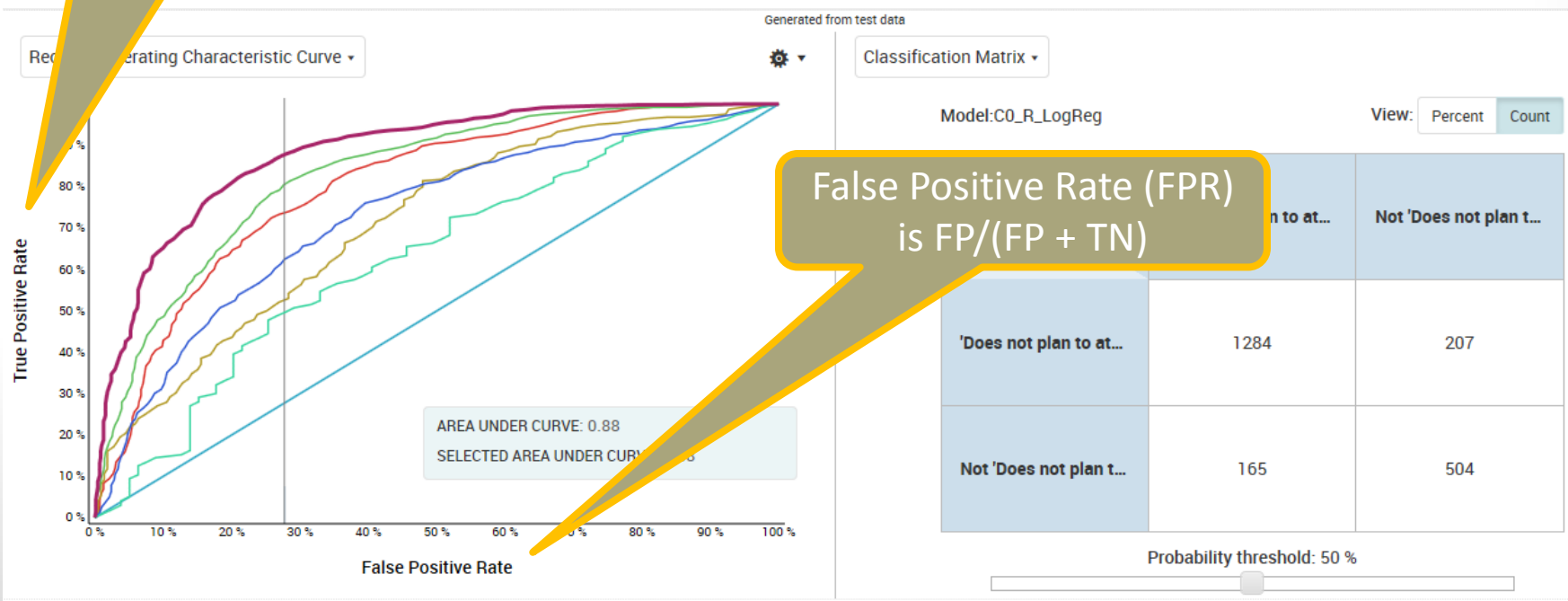


ROC Chart Demo

- Confusion Matrix and ROC Chart

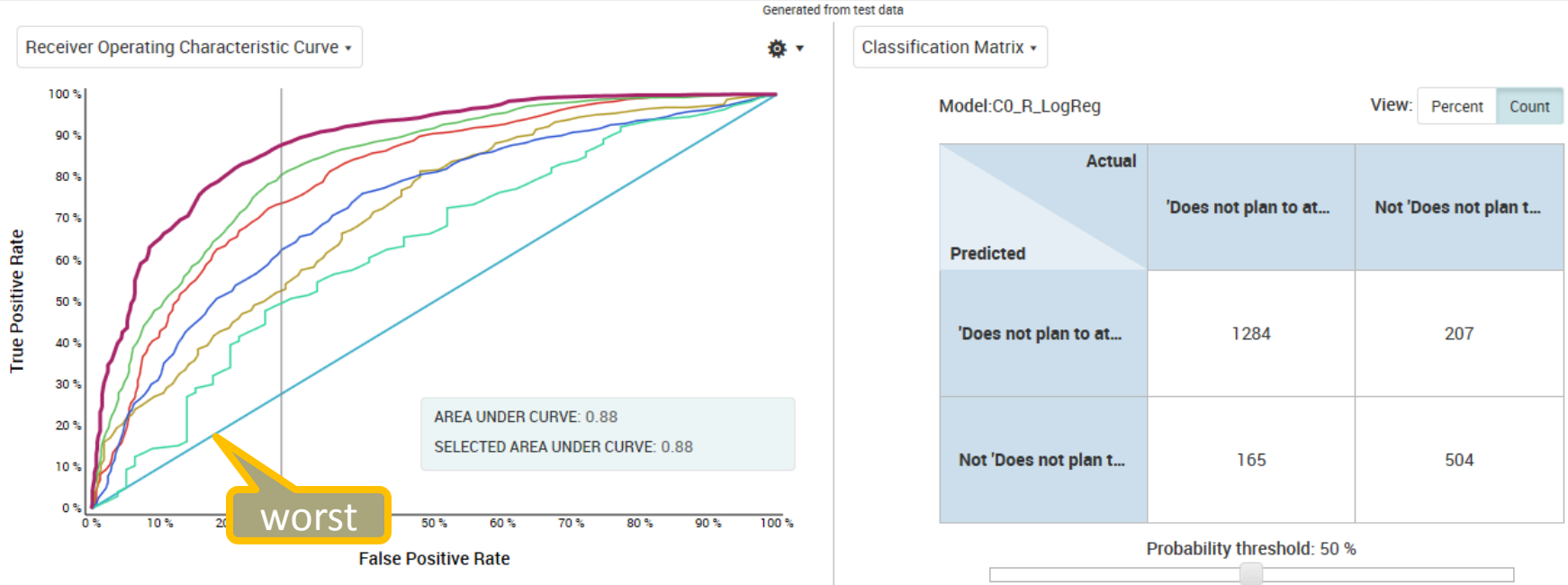
True Positive Rate (TPR) is TP/(TP + FN)

False Positive Rate (FPR) is FP/(FP + TN)



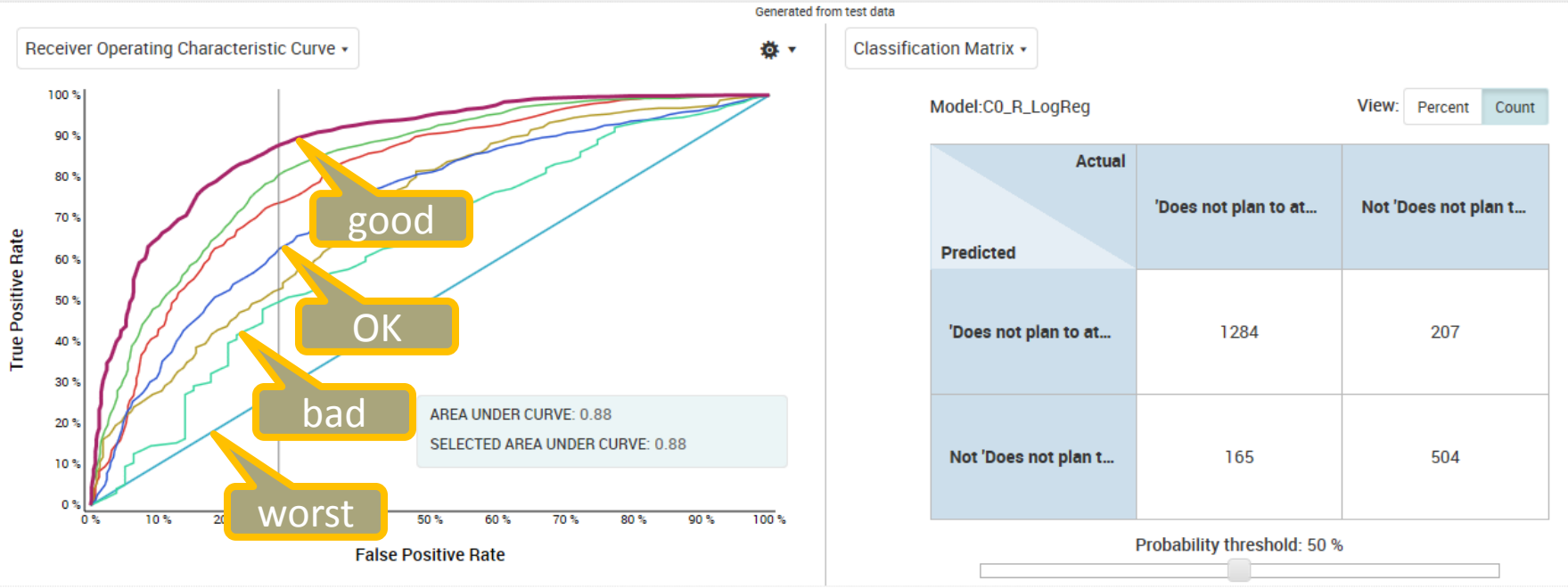
ROC Chart Demo

- Confusion Matrix and ROC Chart



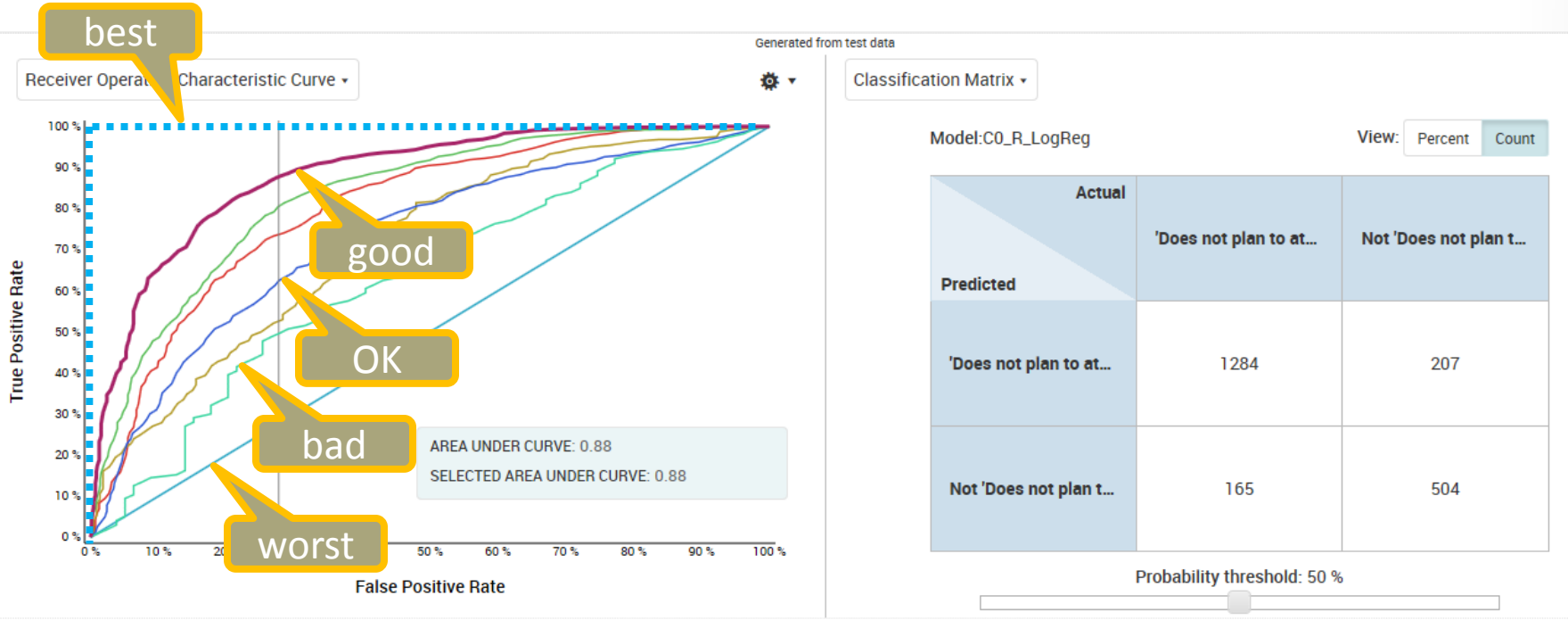
ROC Chart Demo

- Confusion Matrix and ROC Chart



ROC Chart Demo

- Confusion Matrix and ROC Chart



ROC Chart Demo

Quiz 05b

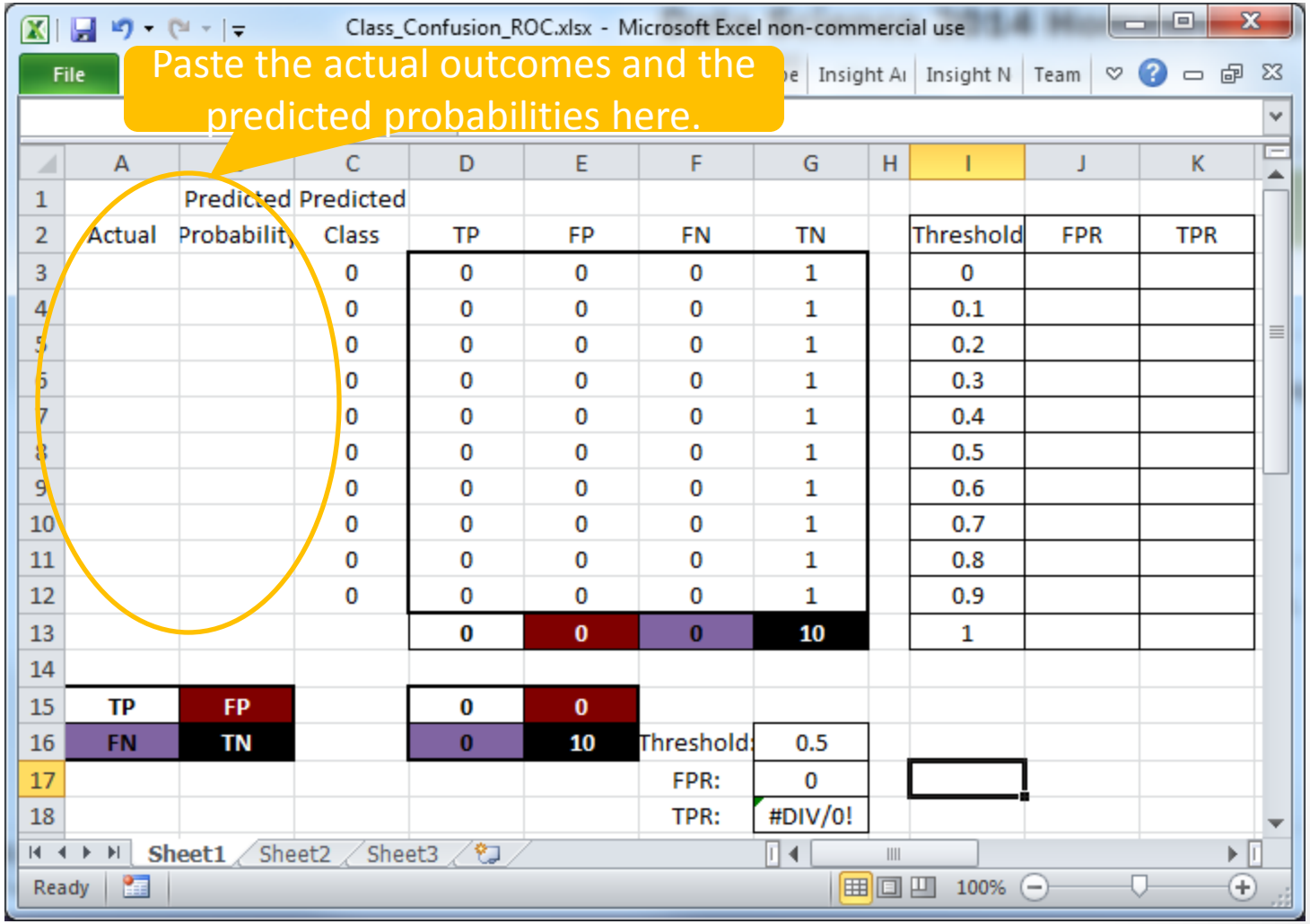
- Test Measures Intro (Confusion Matrix and ROC)

How to make an ROC

How to make an ROC (0)

- From Probabilities to ROC:
- Probabilities -> Threshold -> Predictions -> Confusion Matrix -> ROC
- Get Excel workbook: HowToMakeAnROC.xls
- Note that at the bottom of the worksheet are the actual outcomes and the predicted probabilities.

Exercise: Threshold → Confusion Matrix → ROC (1)



	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3			0	0	0	0	1		0		
4			0	0	0	0	1		0.1		
5			0	0	0	0	1		0.2		
6			0	0	0	0	1		0.3		
7			0	0	0	0	1		0.4		
8			0	0	0	0	1		0.5		
9			0	0	0	0	1		0.6		
10			0	0	0	0	1		0.7		
11			0	0	0	0	1		0.8		
12			0	0	0	0	1		0.9		
13				0	0	0	10		1		
14											
15	TP	FP		0	0						
16	FN	TN		0	10	Threshold:	0.5				
17						FPR:	0				
18						TPR:	#DIV/0!				

Exercise: Threshold → Confusion Matrix → ROC (2)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

Paste the actual outcomes and the predicted probabilities here

	A		C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (3)

The Predicted Probabilities need a threshold

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Exercise: Threshold → Confusion Matrix → ROC (4)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Set the threshold for the Predicted Probabilities

Exercise: Threshold → Confusion Matrix → ROC (5)

Class = Probability > Threshold

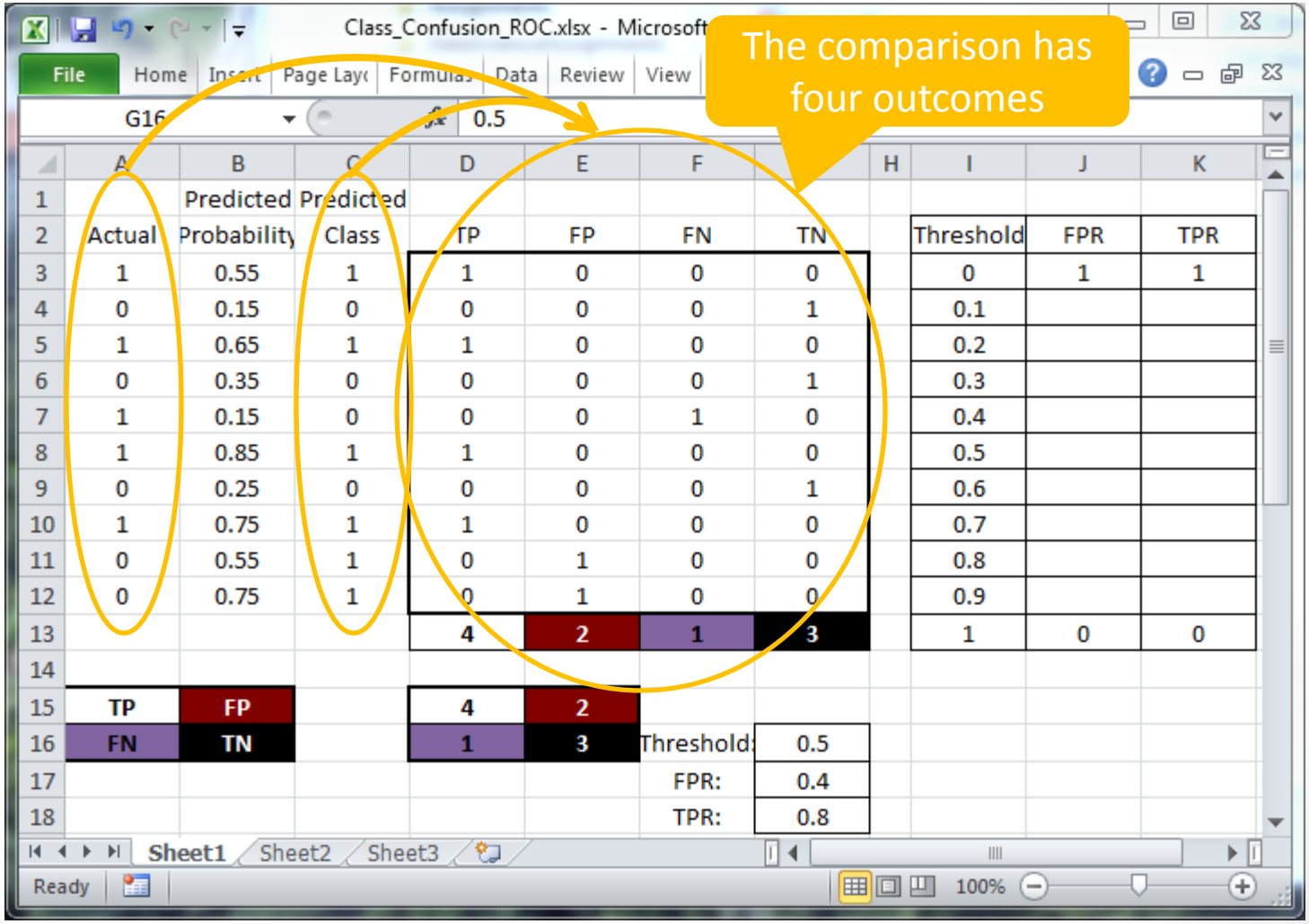
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Exercise: Threshold → Confusion Matrix → ROC (6)

Compare the predicted Class to the Actual Values

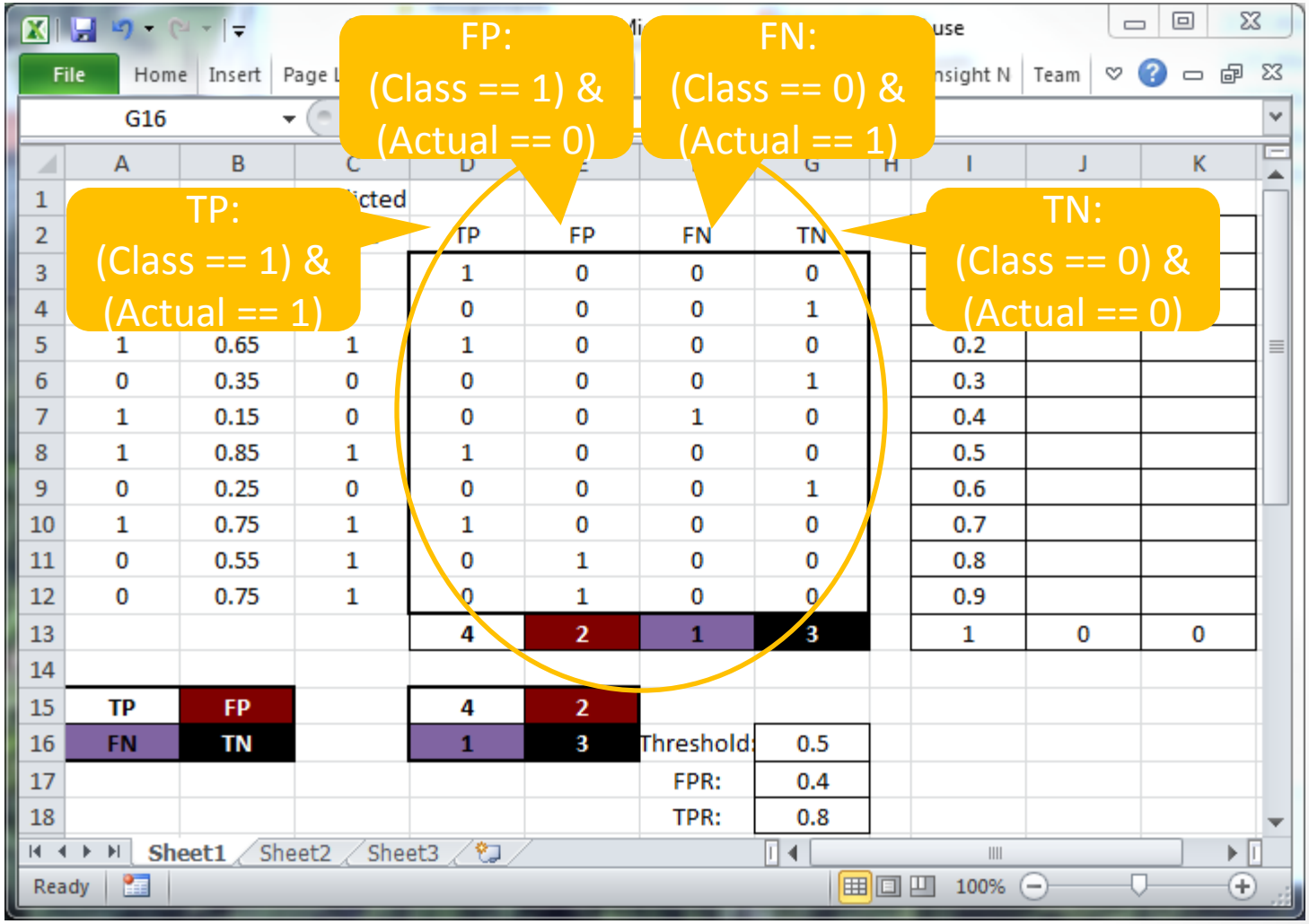
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Exercise: Threshold → Confusion Matrix → ROC (7)



	A	B	C	D	E	F		H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Exercise: Threshold → Confusion Matrix → ROC (8)



Exercise: Threshold → Confusion Matrix → ROC (9)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9			0	0	0	0	1		0.6		
10			1	1	0	0	0		0.7		
11			1	0	1	0	0		0.8		
12			1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Sum(TP)
Sum(FP)
Sum(FN)
Sum(TN)

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (10)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
		5	1	0	1	0	0		0.8		
		5	1	0	1	0	0		0.9		
				4	2	1	3		1	0	0

15	TP	FP		4	2	
16	FN	TN		1	3	Threshold: 0.5
17						FPR: 0.4
18						TPR: 0.8

Sheet1 Sheet2 Sheet3

Ready 100%

Optional: Organize sums into Confusion Matrix

Exercise: Threshold → Confusion Matrix → ROC (11)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3				
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

FPR = FP/(FP + TN)

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (12)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

TPR = TP/(TP + FN)

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (13)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (14)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

G16 | 0.5

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (15)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

K8 | 0.8

	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0						
5	1	0.65	1	1	0						
6	0	0.35	0	0	0	0	1				
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5	0.4	0.8
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

Finally! A point in the ROC Chart.

Sheet1 Sheet2 Sheet3

Ready 100%

Exercise: Threshold → Confusion Matrix → ROC (16)

Class_Confusion_ROC.xlsx - Microsoft Excel non-commercial use

K9 | 0.6

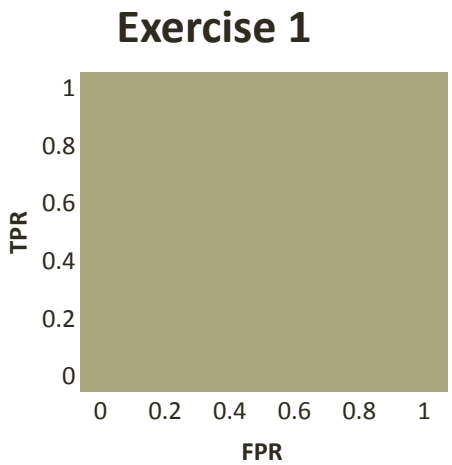
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	0	0	0	1	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5	0.4	0.8
9	0	0.25	0	0	0	0	1		0.6	0.2	0.6
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	0	0	0	0	1		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				3	1	2	4		1	0	0
14											
15	TP	FP		3	1						
16	FN	TN		2	4	Threshold:	0.6				
17						FPR:	0.2				
18						TPR:	0.6				

Repeat the process for all thresholds

Exercise: Threshold → Confusion Matrix → ROC (17)

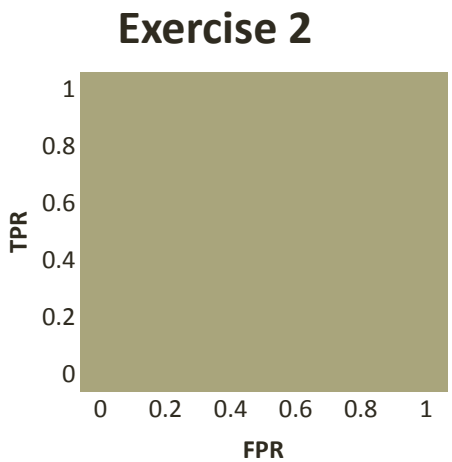
Actual	Predicted Probability
1	0.55
0	0.15
1	0.65
0	0.35
1	0.15
1	0.85
0	0.25
1	0.75
0	0.55
0	0.75

FPR	TPR
1	1
0	0



Actual	Predicted Probability
0	0.15
0	0.25
0	0.35
1	0.45
0	0.45
1	0.55
0	0.65
1	0.75
0	0.85
1	0.95

FPR	TPR
1	1
0	0



How to make an ROC

Assignment (1)

1. Training vs Test Data
   a) Why are performance metrics better on training data than on test data?
   b) Given modeling data, how do you determine which of this data will become training data and which data will become test data?
   c) Given a dataset that was either test or training data, how can you determine if this dataset was training data or test data?
2. Beware, this problem contains irrelevant data while some important numbers are not explicitly presented. A model was trained on *300* individuals where *149* had the cold and *151* were healthy. The model was tested on **100** individuals where **10** were actually ill. The model correctly predicted that **85** of the healthy individuals were indeed healthy and correctly predicted that **7** of the ill individuals were indeed ill. The other predictions were incorrect. Consult Wikipedia: http://en.wikipedia.org/wiki/Precision_and_recall and construct a confusion matrix and then calculate the following:
   a) Sensitivity
   b) Specificity
   c) Accuracy
   d) Precision
   e) Recall

Assignment (2)

3. The probability threshold for a classification varies in an ROC chart from 0 to 1.
   a) What point of the graph corresponds to a threshold of zero?
   b) What point of the graph corresponds to a threshold of one?
   c) What point of the graph corresponds to a threshold of 0.5? (trick question)
4. A Classification is tested on 1000 cases. In the approximate middle of its ROC chart there is a point where the false positive rate is 0.4, the true positive rate is 0.8, and the accuracy is 0.7.
   a) What does the confusion matrix look like?
   b) What can you say about the probability threshold at that point? (trick question)
5. In HowToMakeAnROC.xls, complete the Exercises 1 and 2 and graph both of these ROC charts in the same Excel file. Examples A and B are examples of how to do Exercises 1 and 2..

Assignment (3)

6. Get SetupVirtualMachine.pdf from Canvas and follow directions.
   - Download VM from this link:
     - https://www.dropbox.com/s/znw47w4lh9zcbrx/Cloudera-Training-VM-4.2.1.p-vmware_prist2.zip?dl=0
   - Install and setup the Hadoop VM according to SetupVirtualMachine.pdf". After you entered "hadoop fs -ls" (without quotes), enter your name into the console and then take a screen shot of the whole virtual machine. Submit the screenshot to Canvas.
7. Submit answers to items 1 through 4 in a text file. If you used R, then submit the R file, too. Submit the completed Excel file from item 5. Submit the screenshot from item 6. Submission deadline is Saturday 11:54 PM.
8. Look through the Preview section in Canvas. Read:
   - Google file system: http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf
   - MapReduce: http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf

Introduction to Data Science