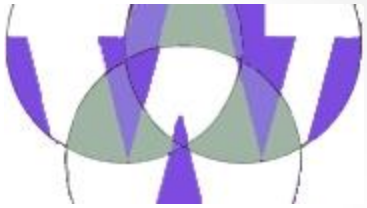
Introduction to Data Science

Lecture 4; October 26th, 2016

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

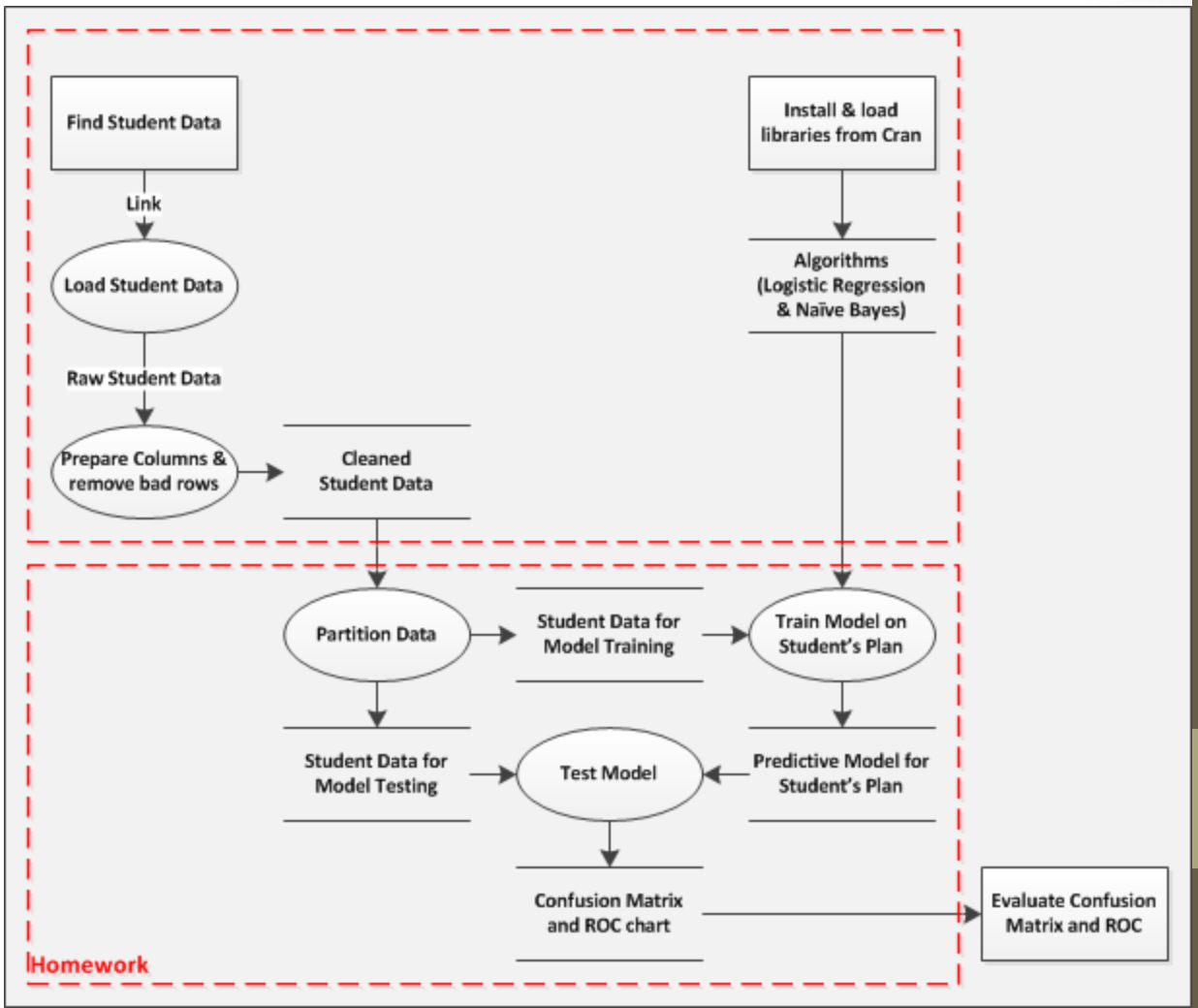
Agenda



- Announcements
  - The social component is a course requirement:
    - On LinkedIn, ask questions about the homework, start a discussion, or make a comment on an existing discussion.
    - Please collaborate on homework!
  - Guest Lecture in November
    - Business Side of Data Science by Marius Marcu on November 16th 2016
- Review Homework: Classifications in R
- Quiz on Classifications in R
- Overfitting and Confusion Matrix
- Break and optional Video
- ROC Chart Intro
- Quiz on intro to Confusion Matrix
- How to make an ROC
- Break and optional Video
- Data Structures (Homework Reading)
- Assignment. See assignment slides at the end of the deck. (Complete all assignments items from all assignment slides. Submit by Saturday 11:57 PM)

Homework Review: Classifications in R

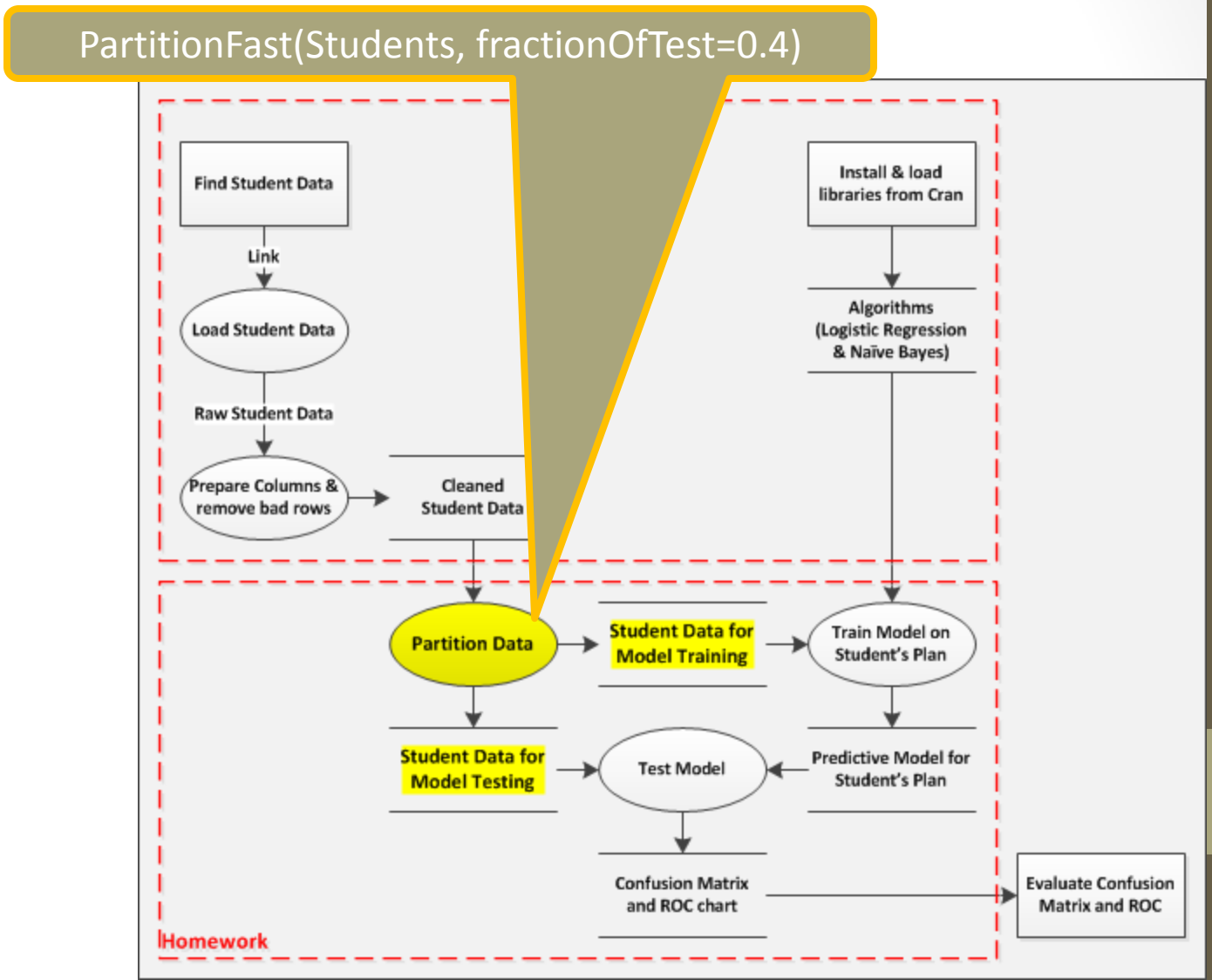
Homework Review: Classifications in R



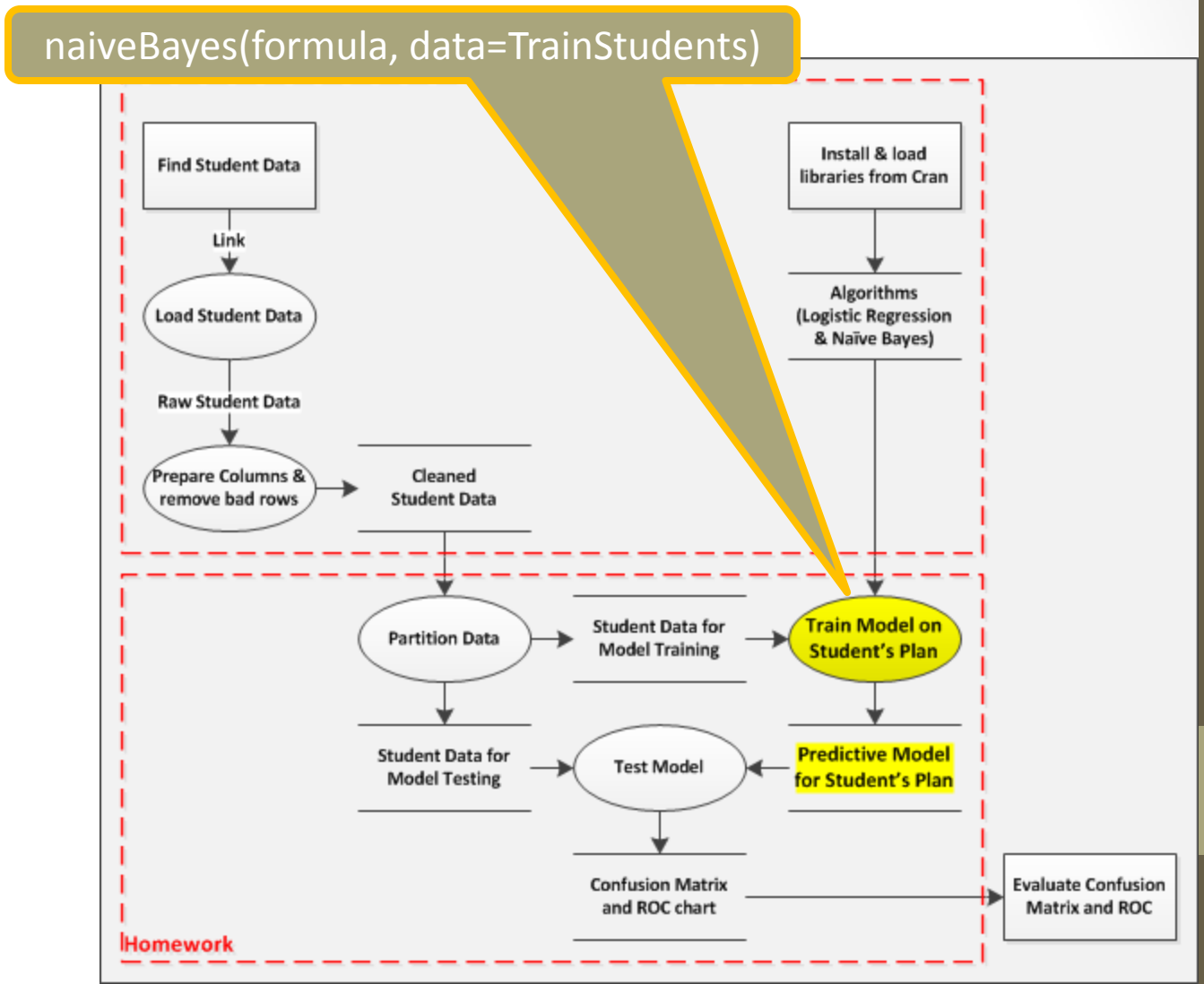
Homework Review: Classifications in R

ClassifyStudents.R
&
CollegeStudentDatasets.R

Homework Review: Classifications in R

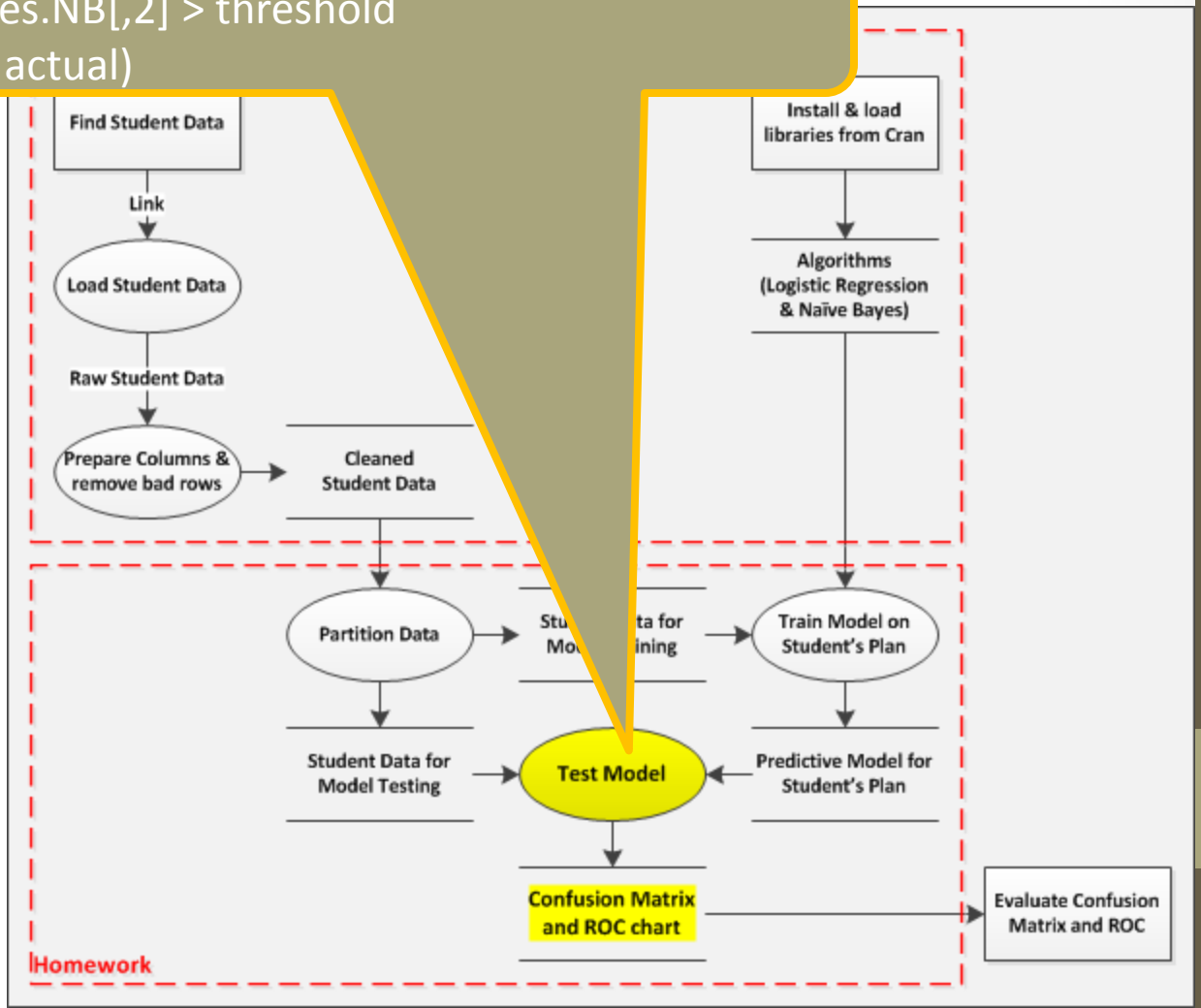


Homework Review: Classifications in R

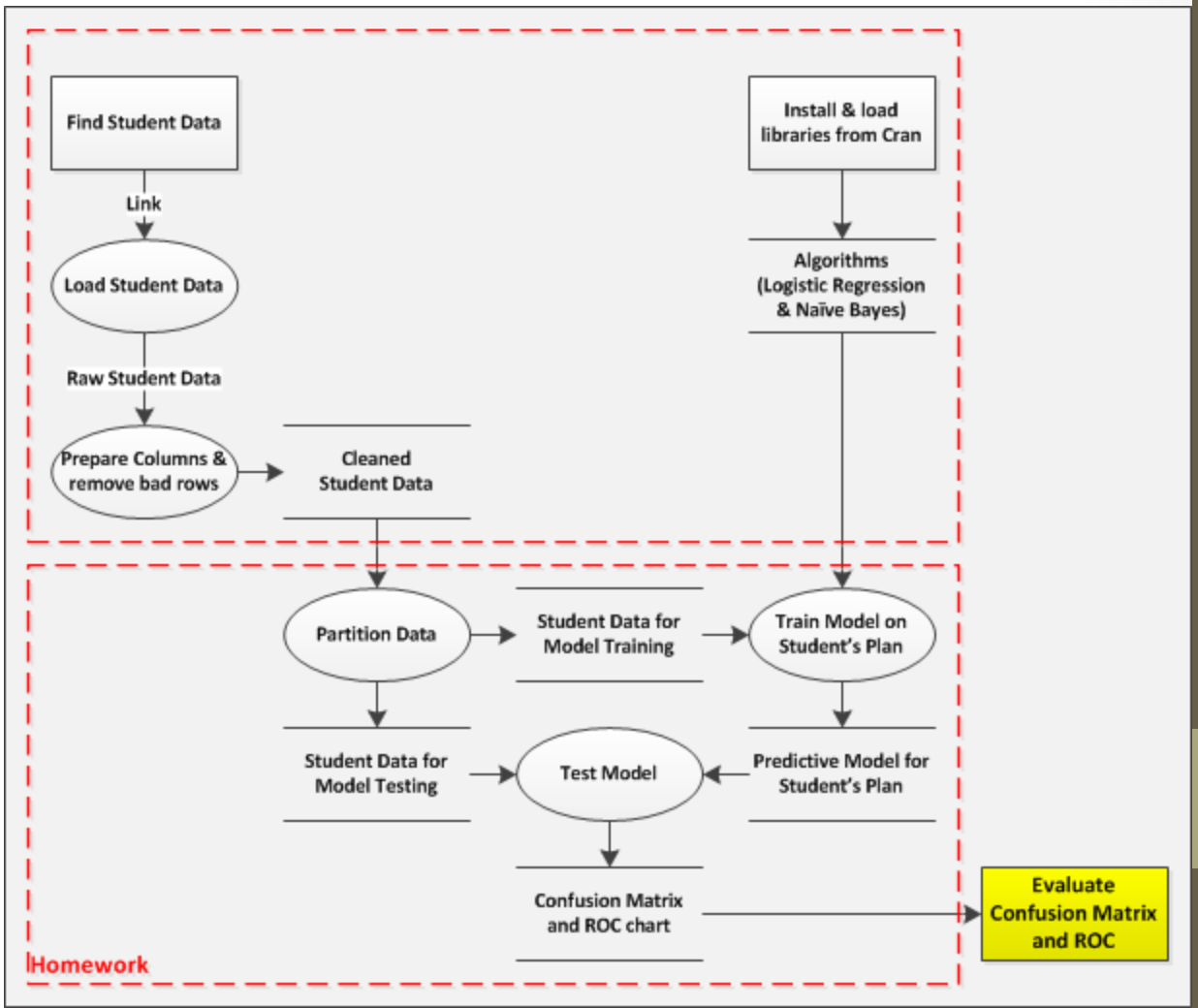


Homework Review: Classifications in R

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



Homework Review: Classifications in R



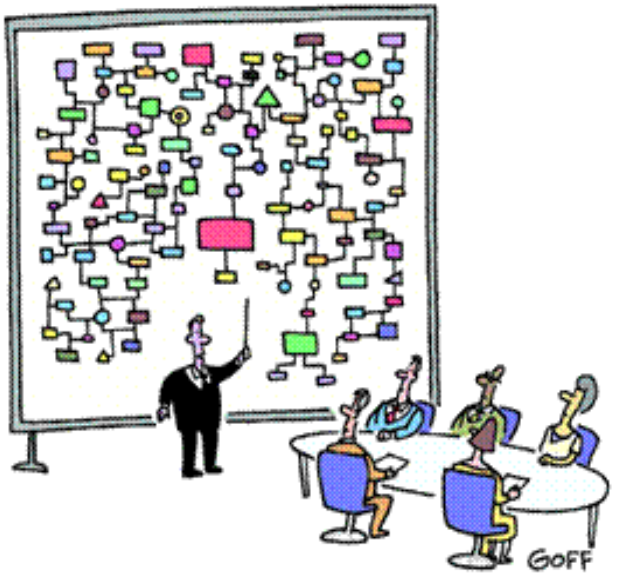
Homework Review: Classifications in R

- See: today's versions of:
  - ClassifyStudents.R
  - CollegeStudentsDatasets.R

- Partitioning was tested with:
  - PartitionTestFunctions.R

Quiz on Classification in R

- You can answer the first questions without R or an R script.
- For the last questions in this quiz you will need to download the R-script PatientReadmission.R from Canvas. That R-script will download the required data (PatientReadmission.csv) from dropbox. Or, you can get those data from Canvas.



"And that's why we need a computer."

Homework Review: Classifications in R

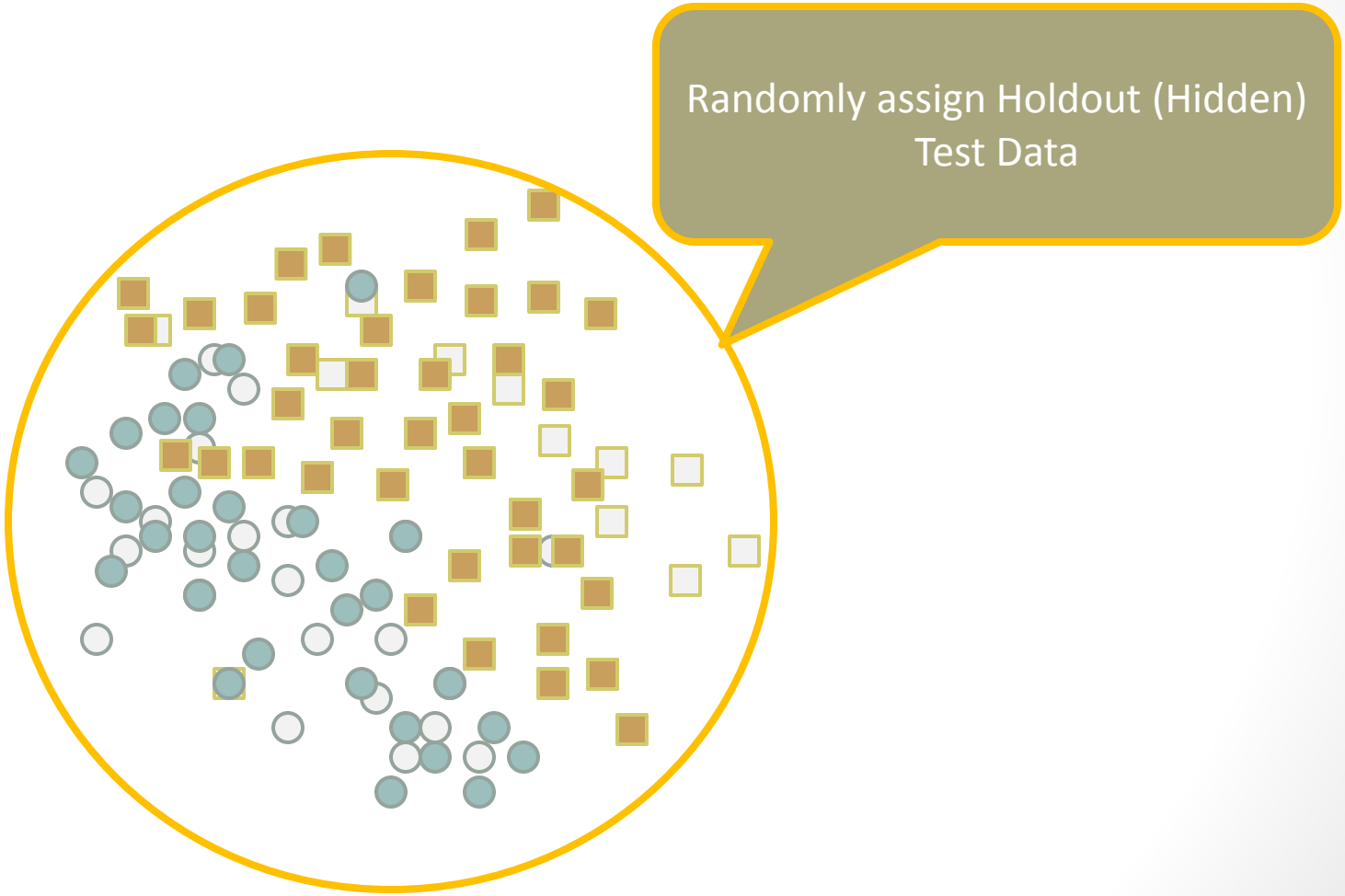
Over-fitting and Confusion Matrix

Evaluate Model

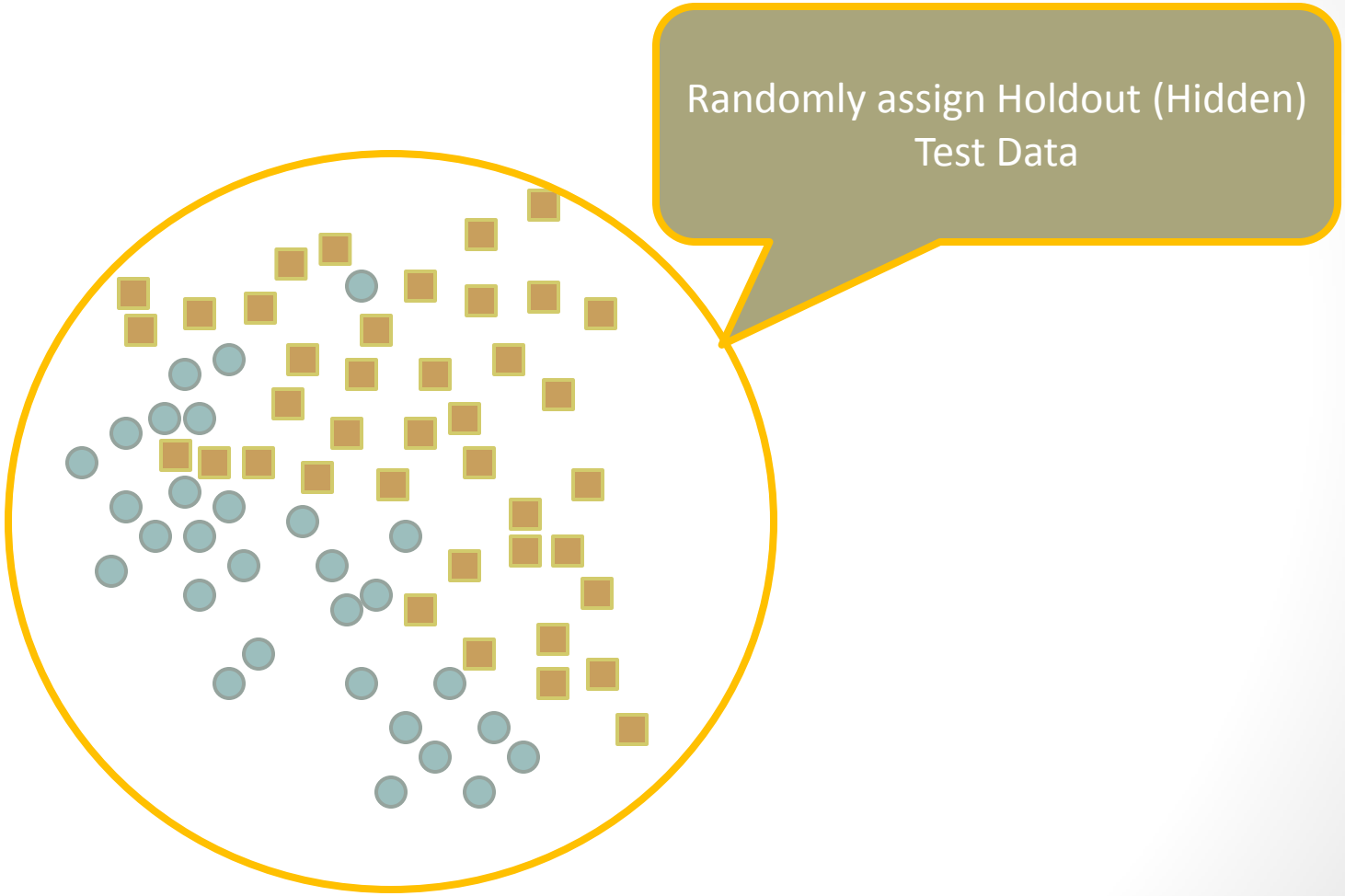
- The following segment will use an over-fitting example to explain the following concepts:
  - Modeling Data
    - Training Data
    - Test Data
  - Model (Hypothesis)
  - Over-fitting
  - Model Accuracy
  - Confusion Matrix (Classification Matrix)
    - True Positive
    - False Positive
    - True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

Evaluate Model: All Data

Evaluate Model: Test Data

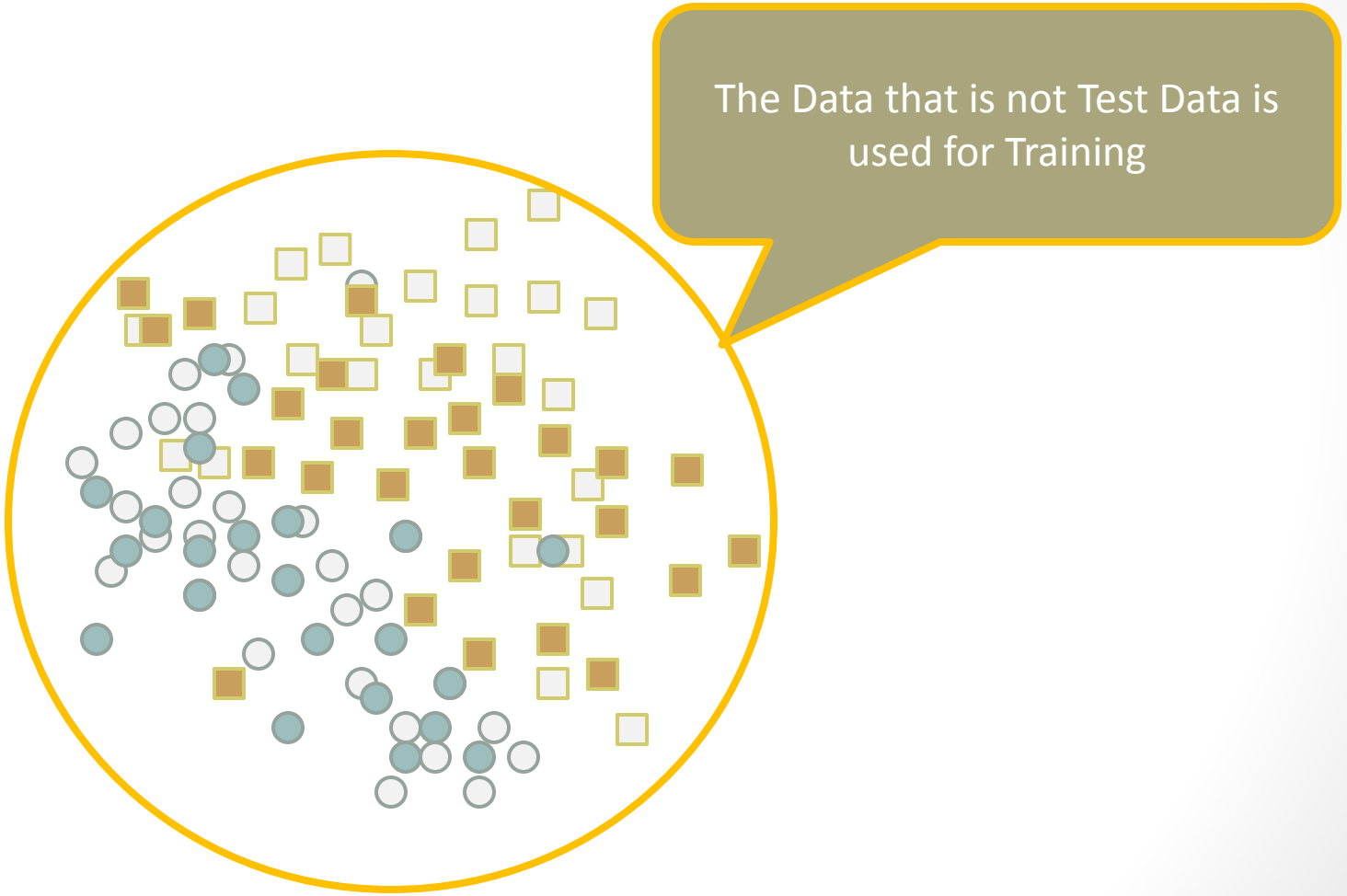


Evaluate Model: Test Data

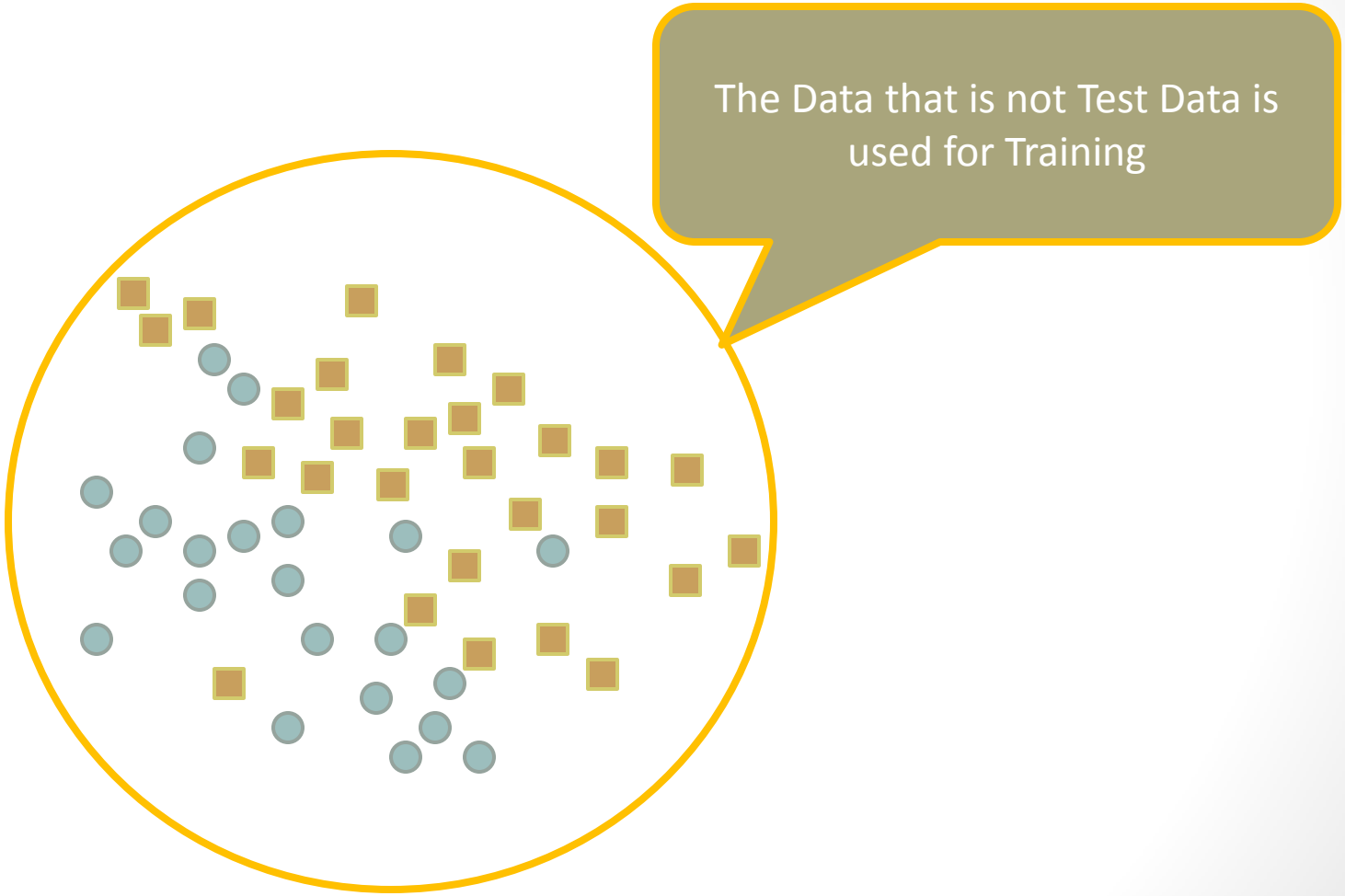


Evaluate Model: All Data

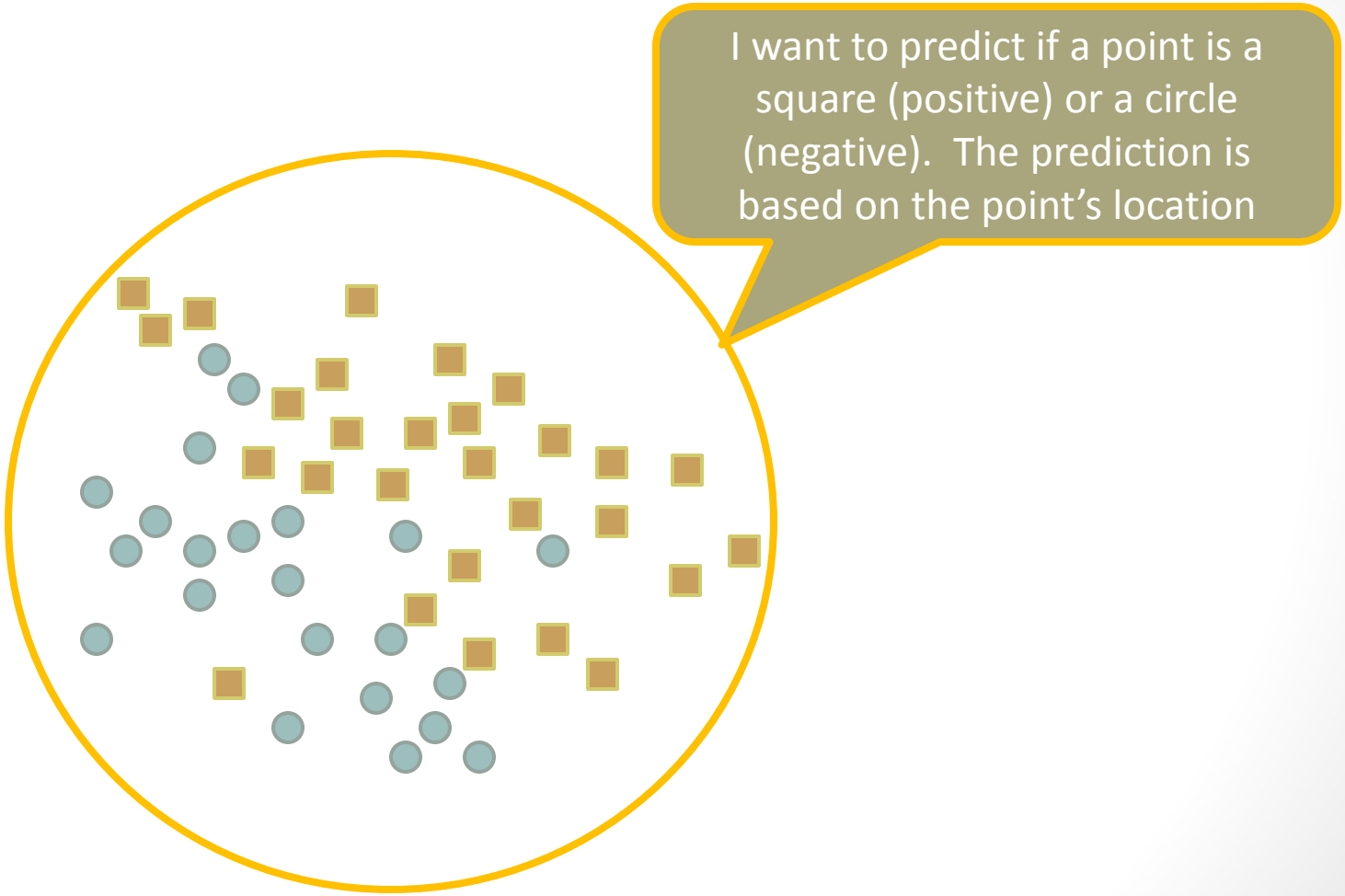
Evaluate Model: Training Data



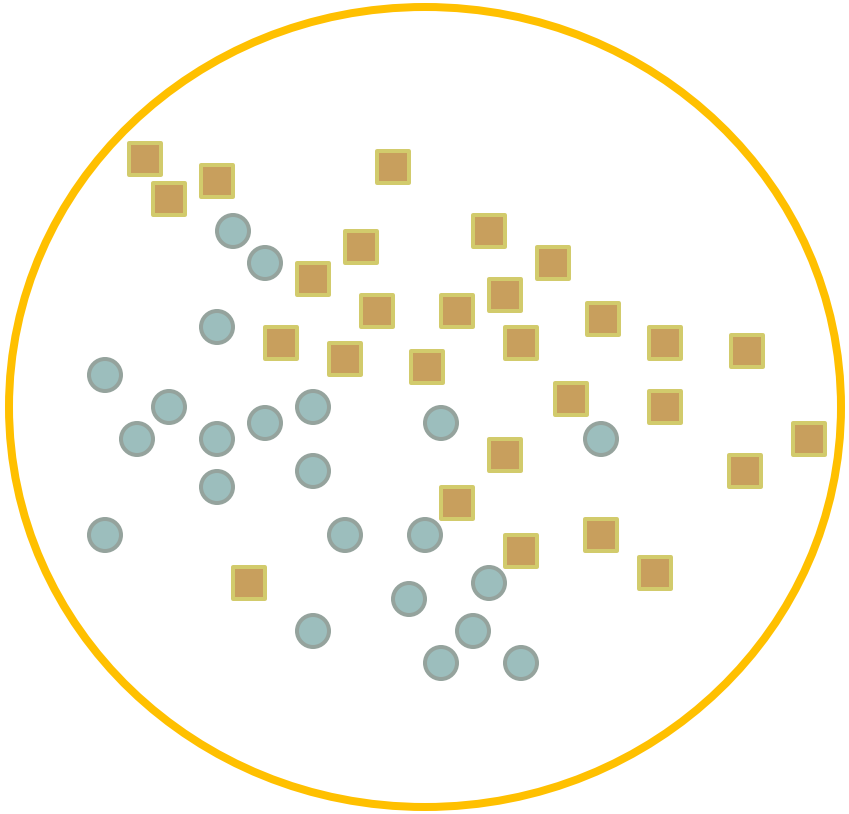
Evaluate Model: Training Data



Evaluate Model: Training



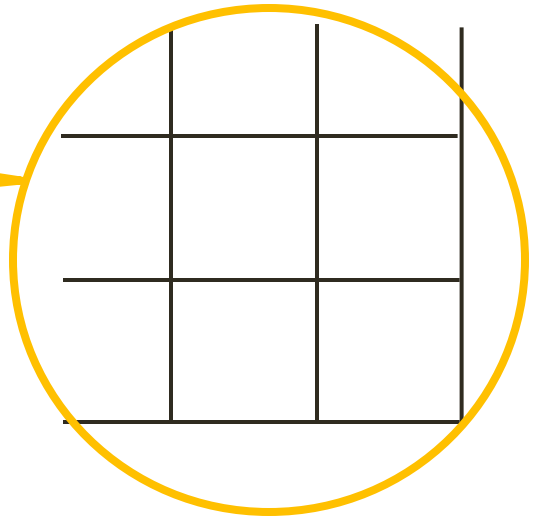
Evaluate Model: Training

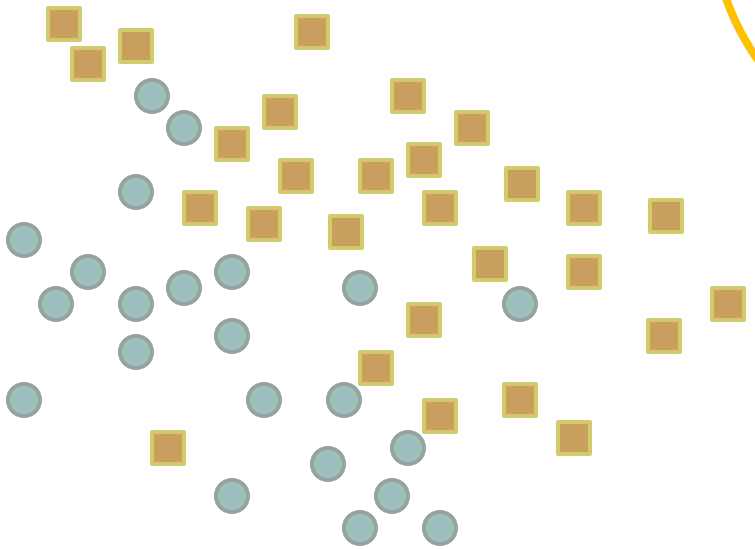


isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

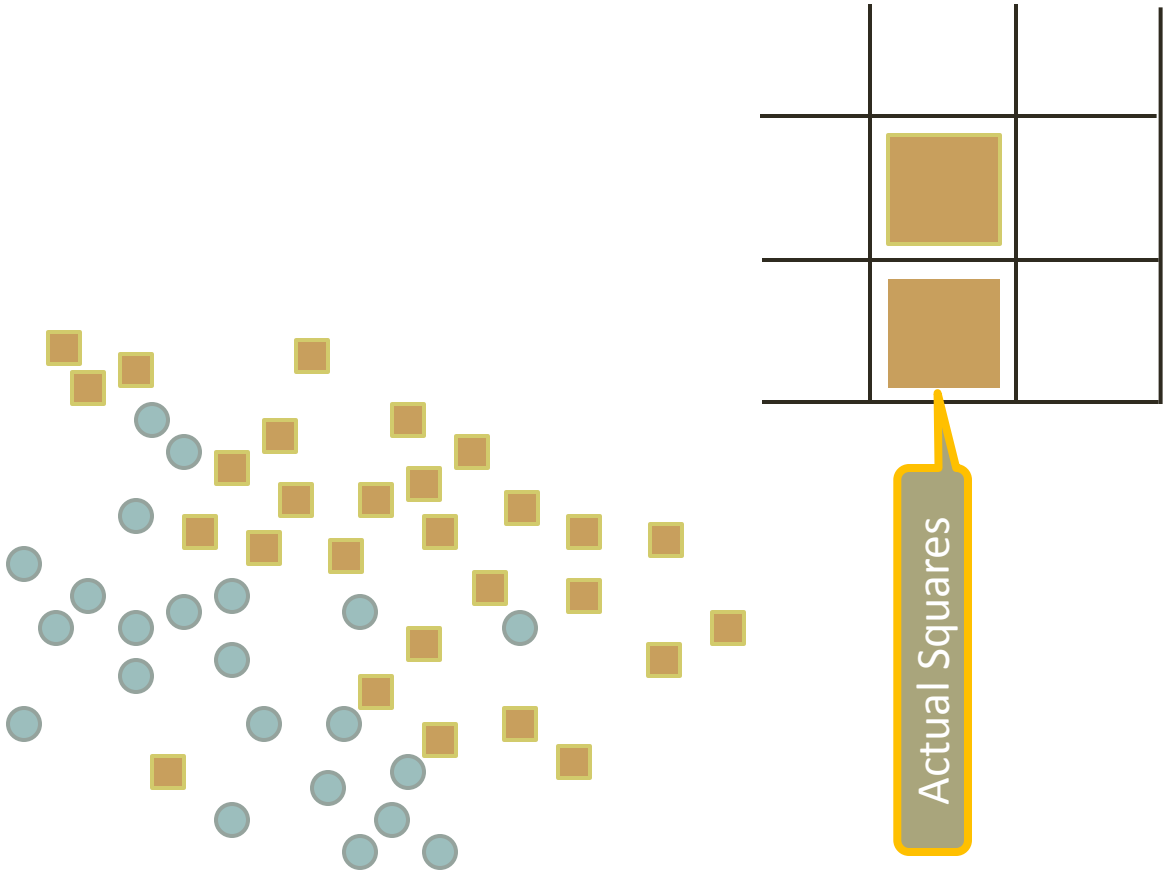
Confusion Matrix (Classification Matrix): Compare Squares and Circles with Predicted Squares and Circles





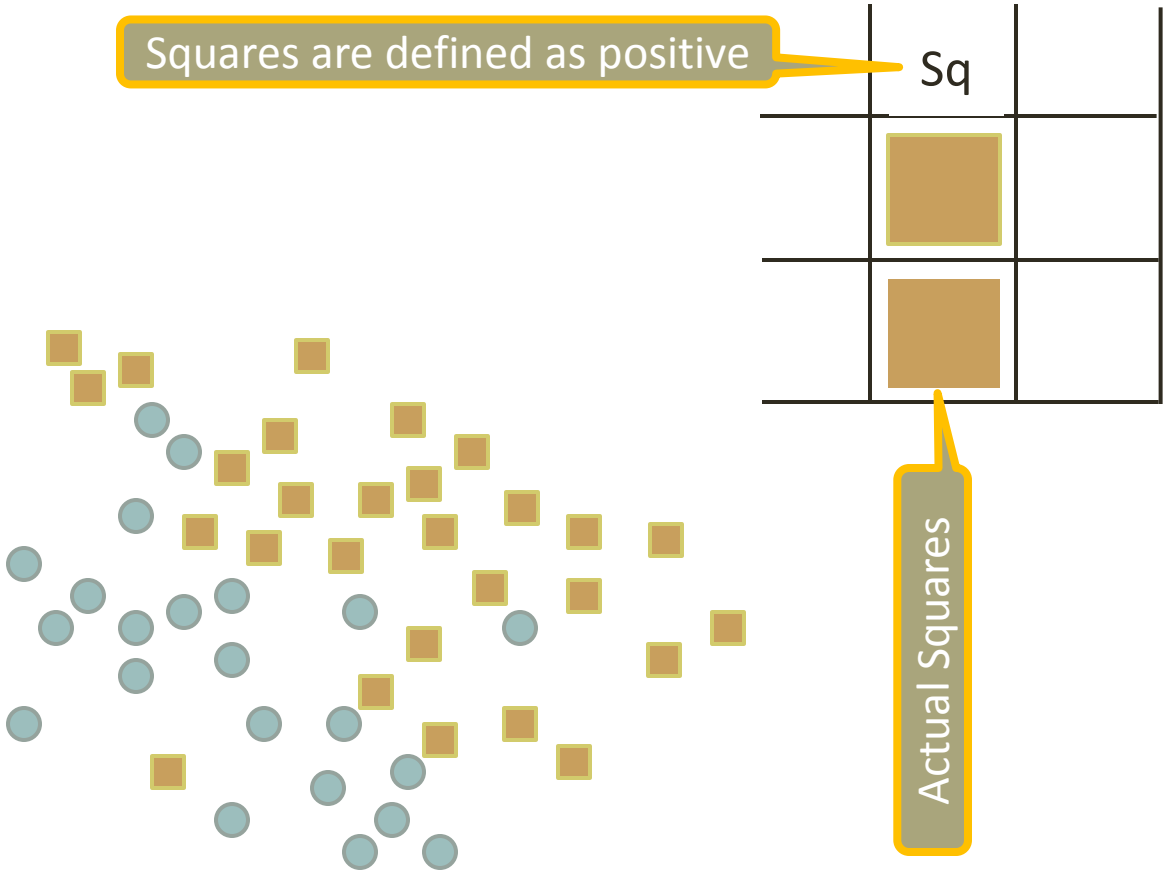
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



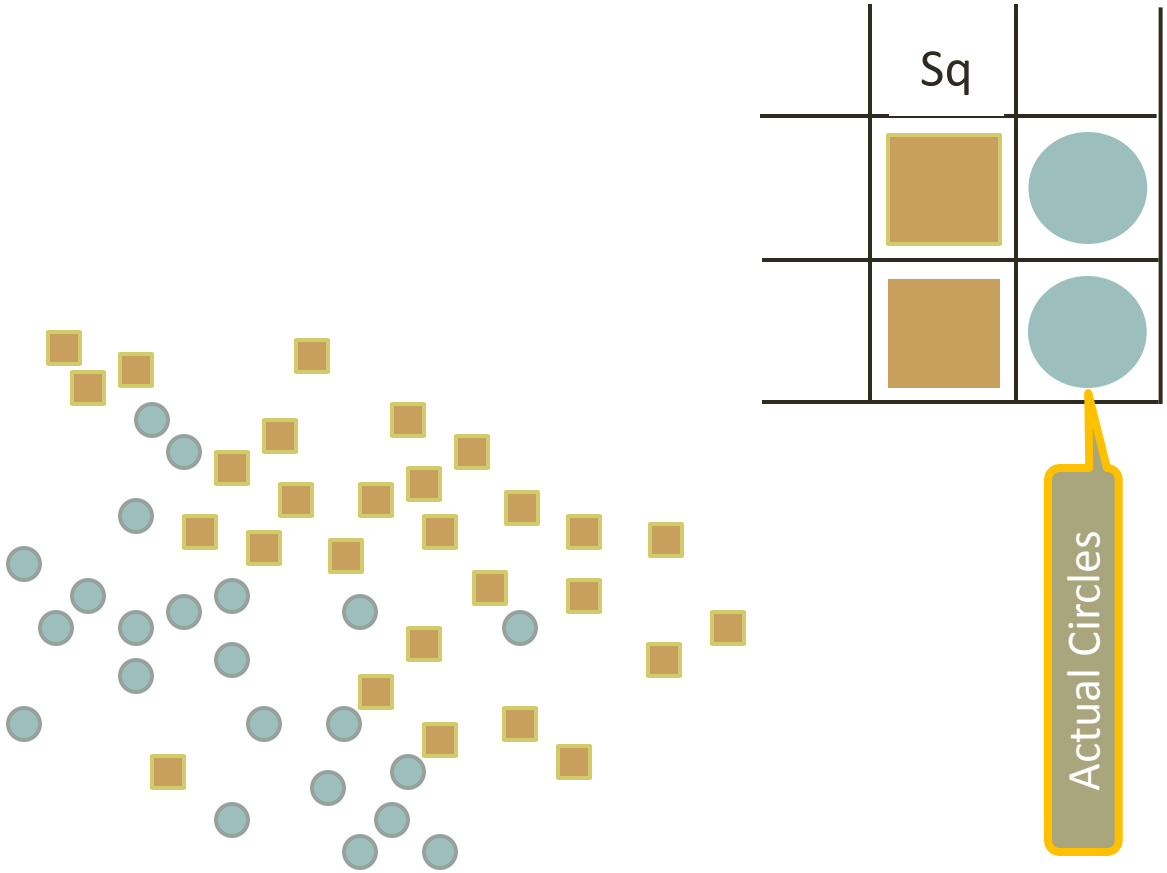
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



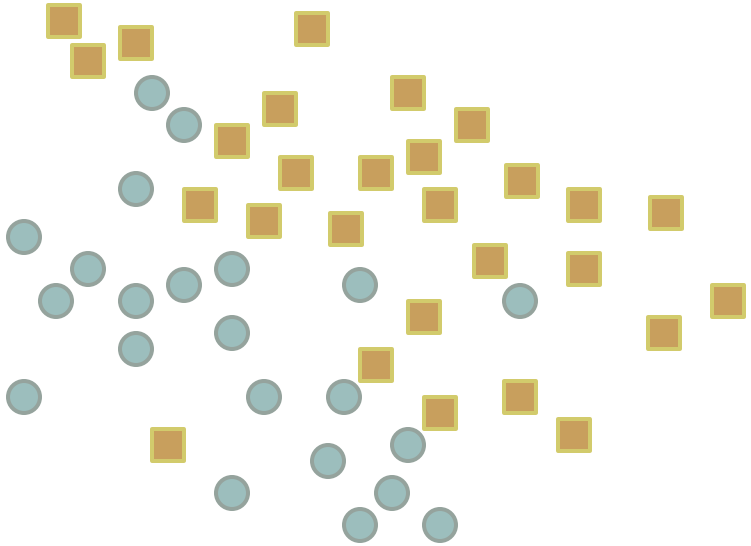
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



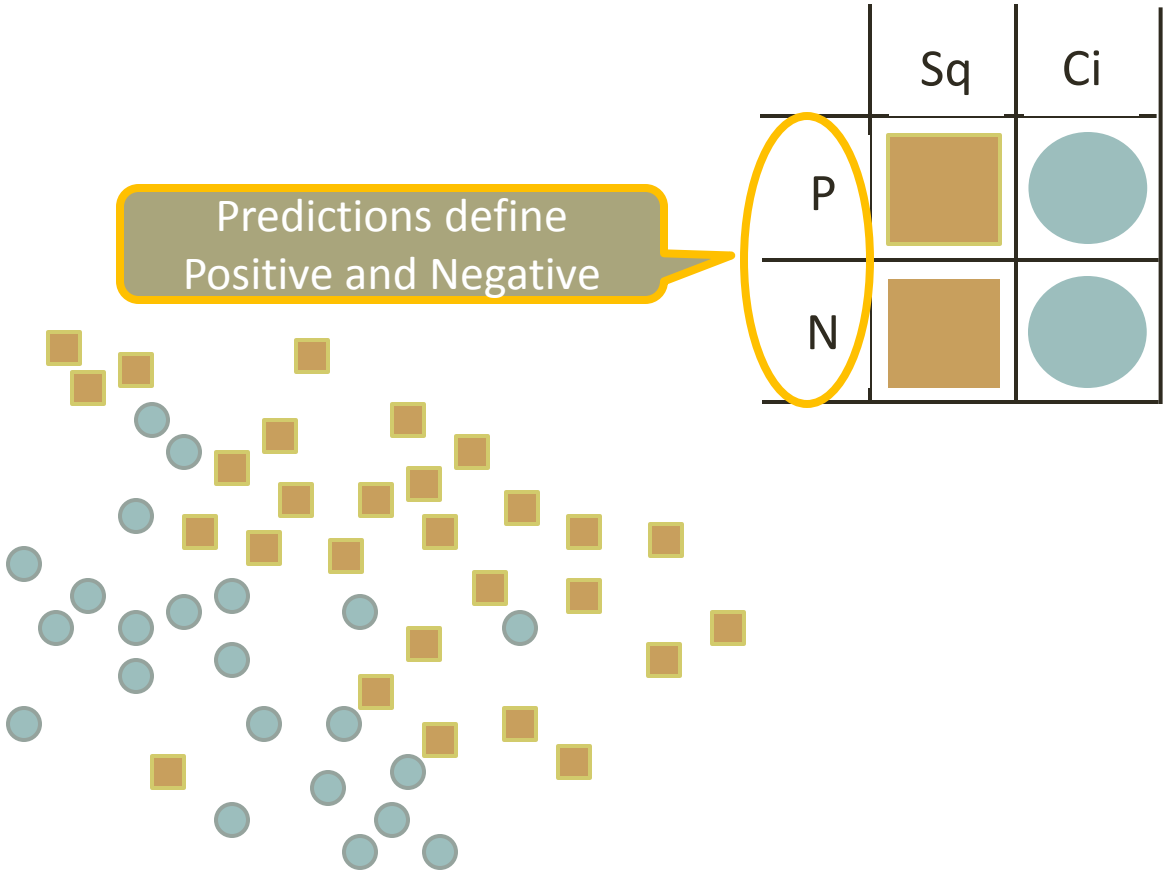
Circles are defined as negative

	Sq	Ci
		
		

Actual Circles

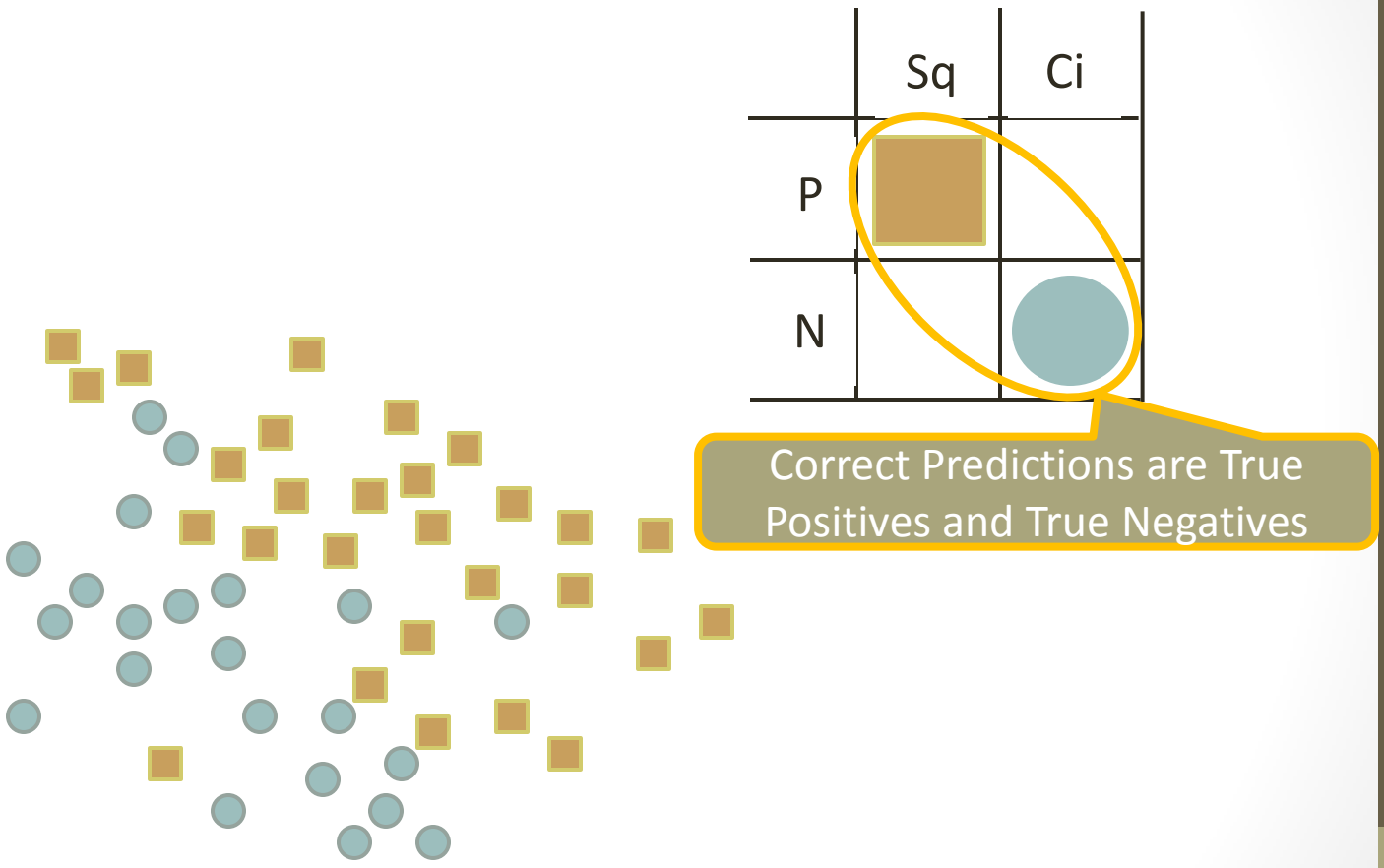
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



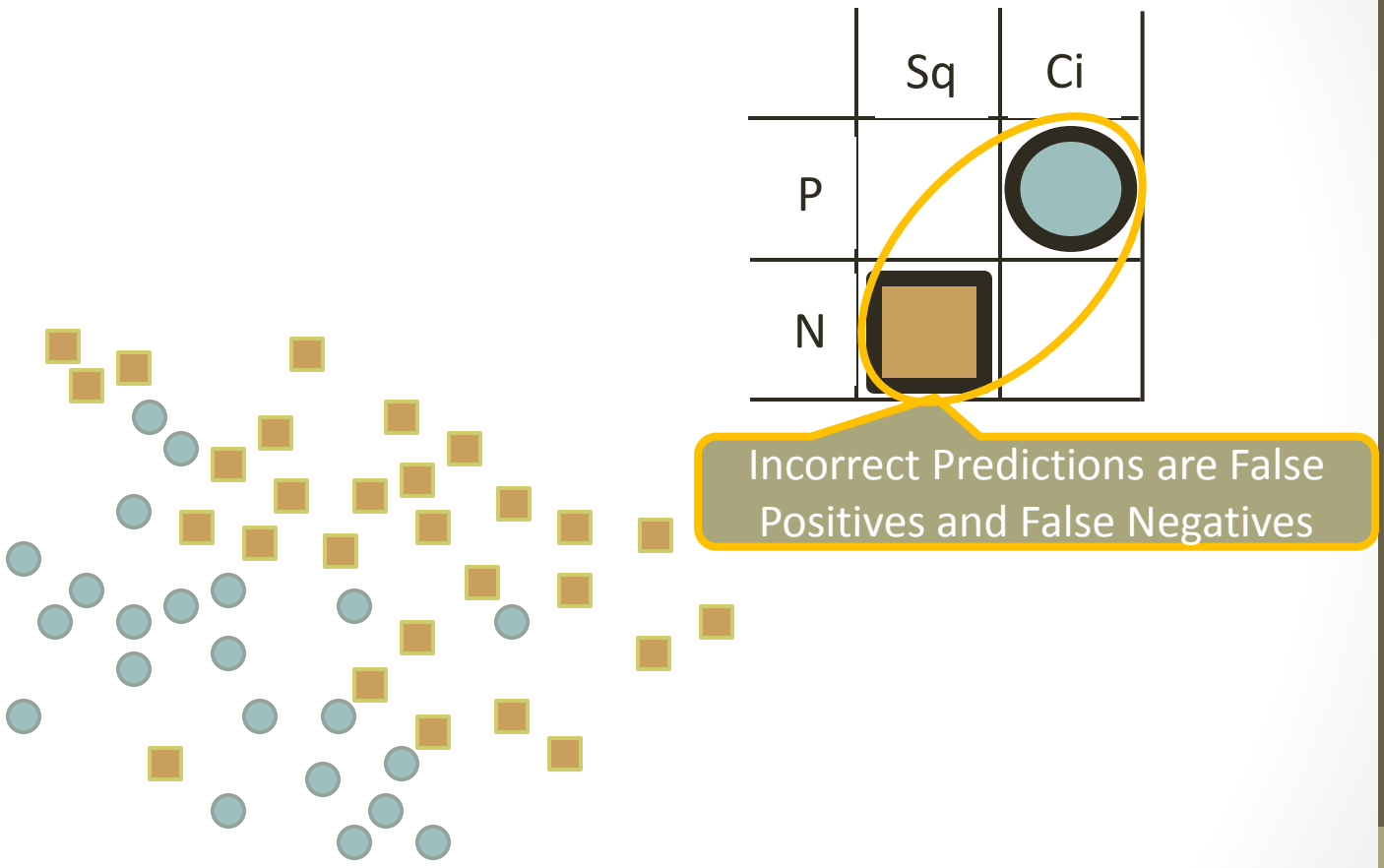
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

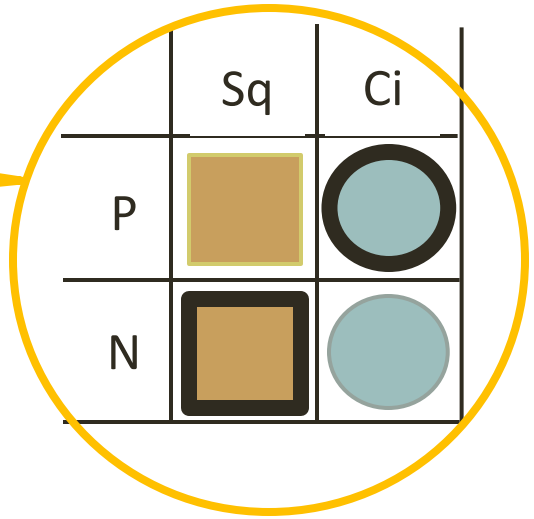
Evaluate Model: Confusion Matrix

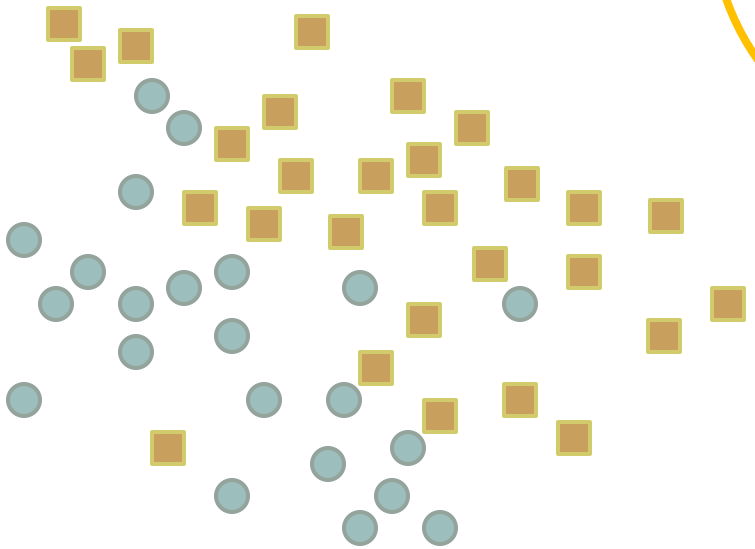


isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

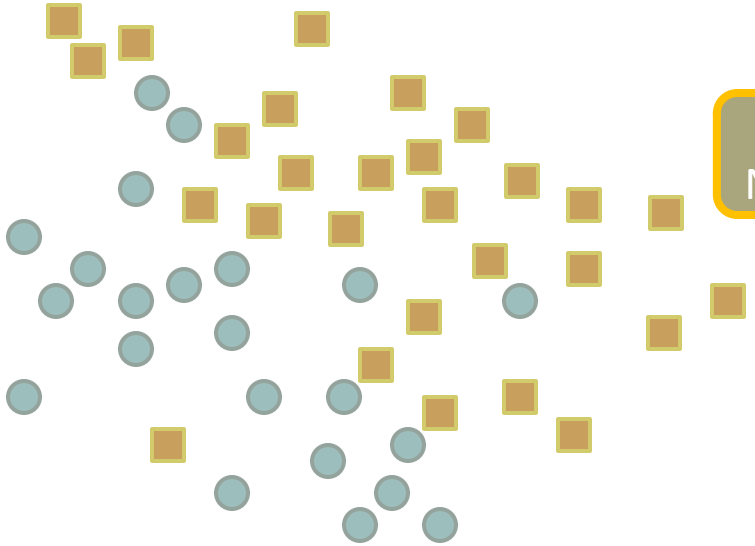
Confusion Matrix (Classification Matrix):
Vertical are actual classes
Horizontal are predicted classes

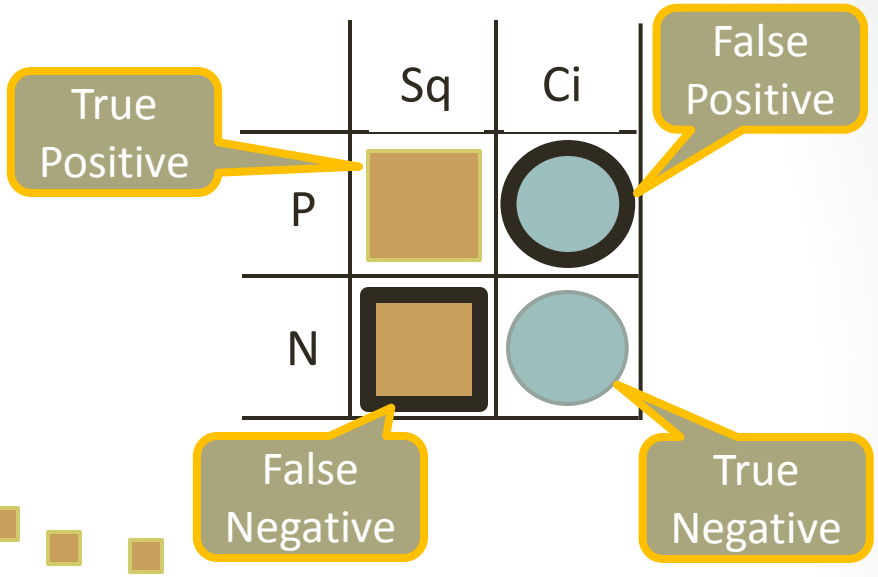




isSquare ~ xLocation + yLocation

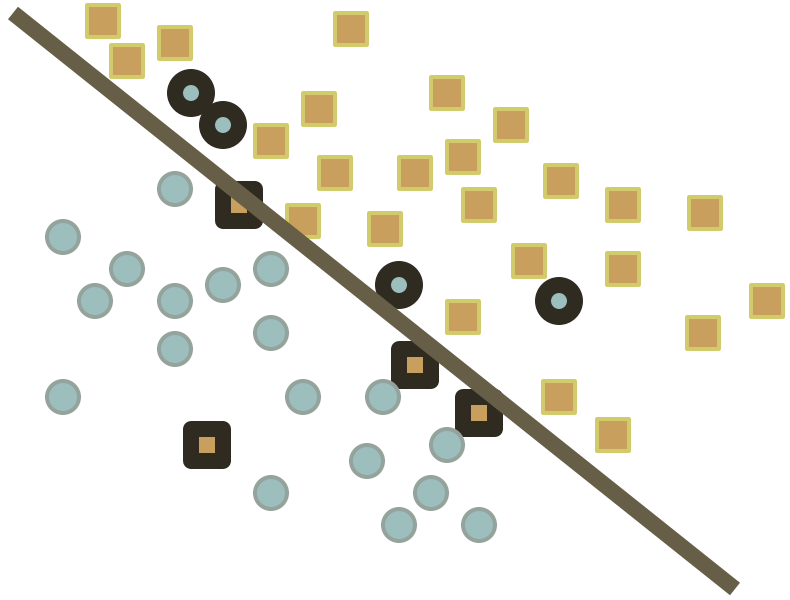
Evaluate Model: Confusion Matrix





isSquare ~ xLocation + yLocation

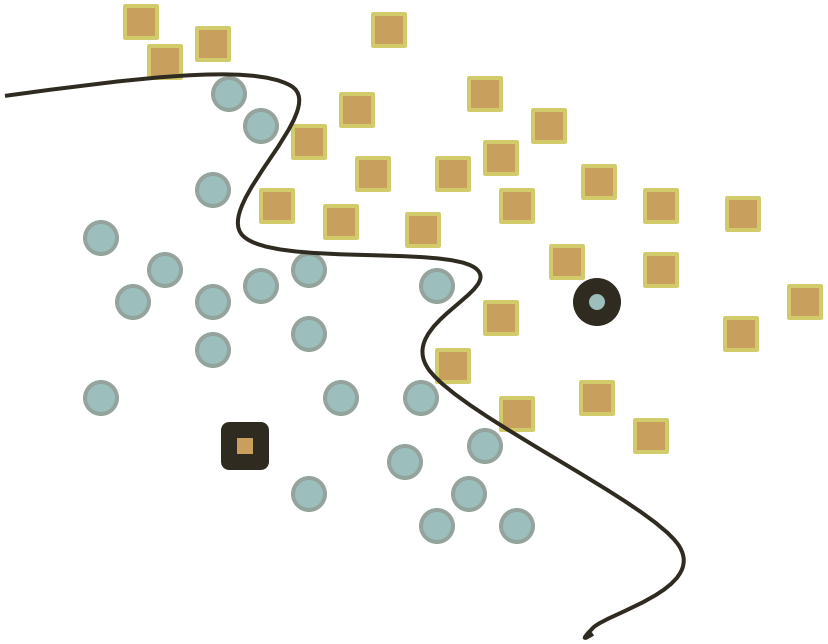
Evaluate Model : Train Model 1



	Sq	Ci
P	36	4
N	4	26

isSquare ~ xLocation + yLocation

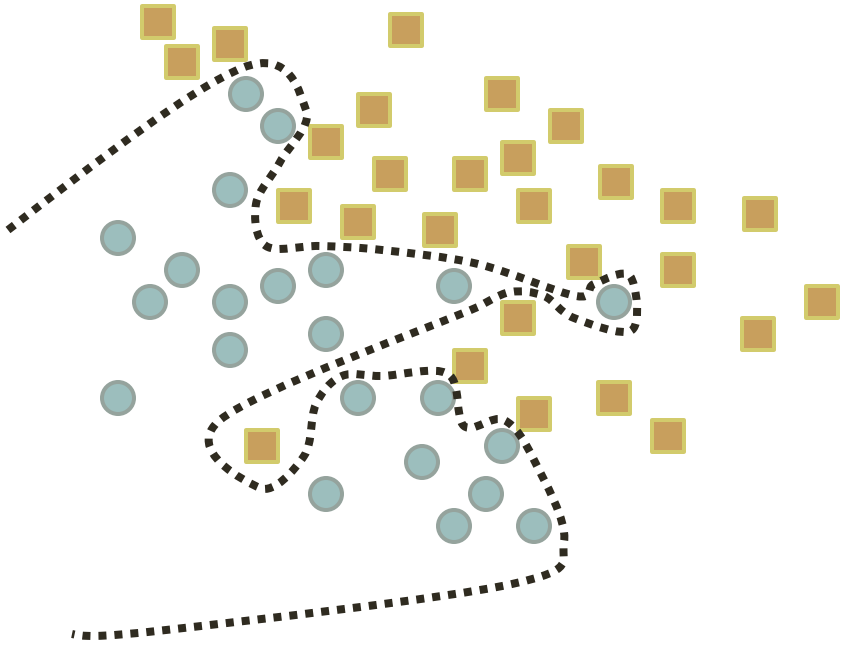
Evaluate Model : Train Model 2



	Sq	Ci
P	39	1
N	1	29

isSquare ~ xLocation + yLocation

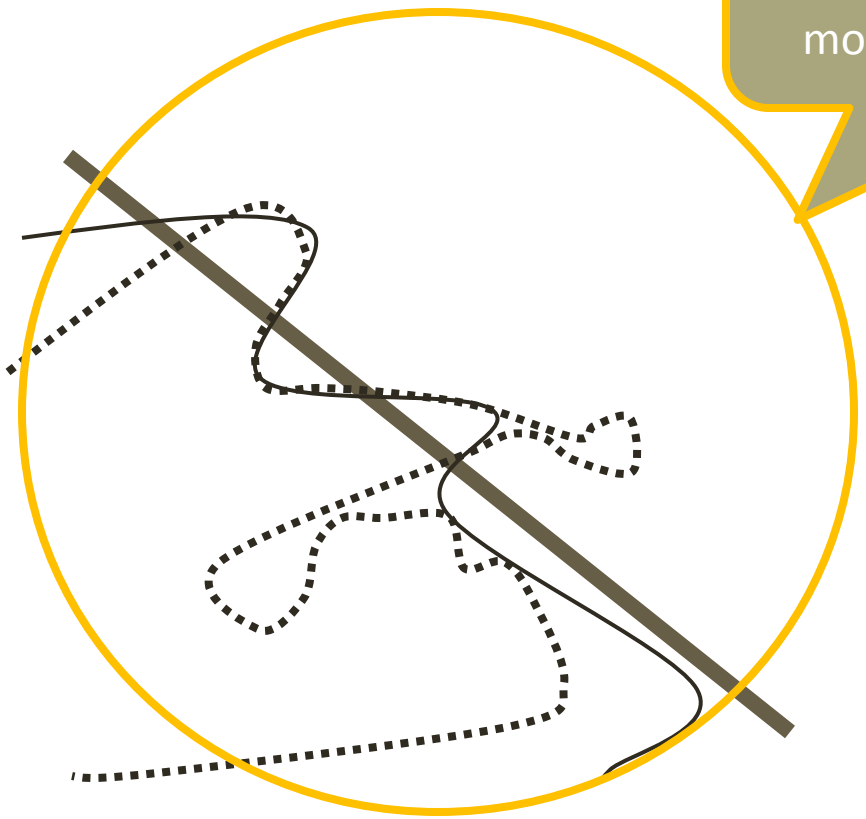
Evaluate Model : Train Model 3



	Sq	Ci
P	40	0
N	0	30

isSquare ~ xLocation + yLocation

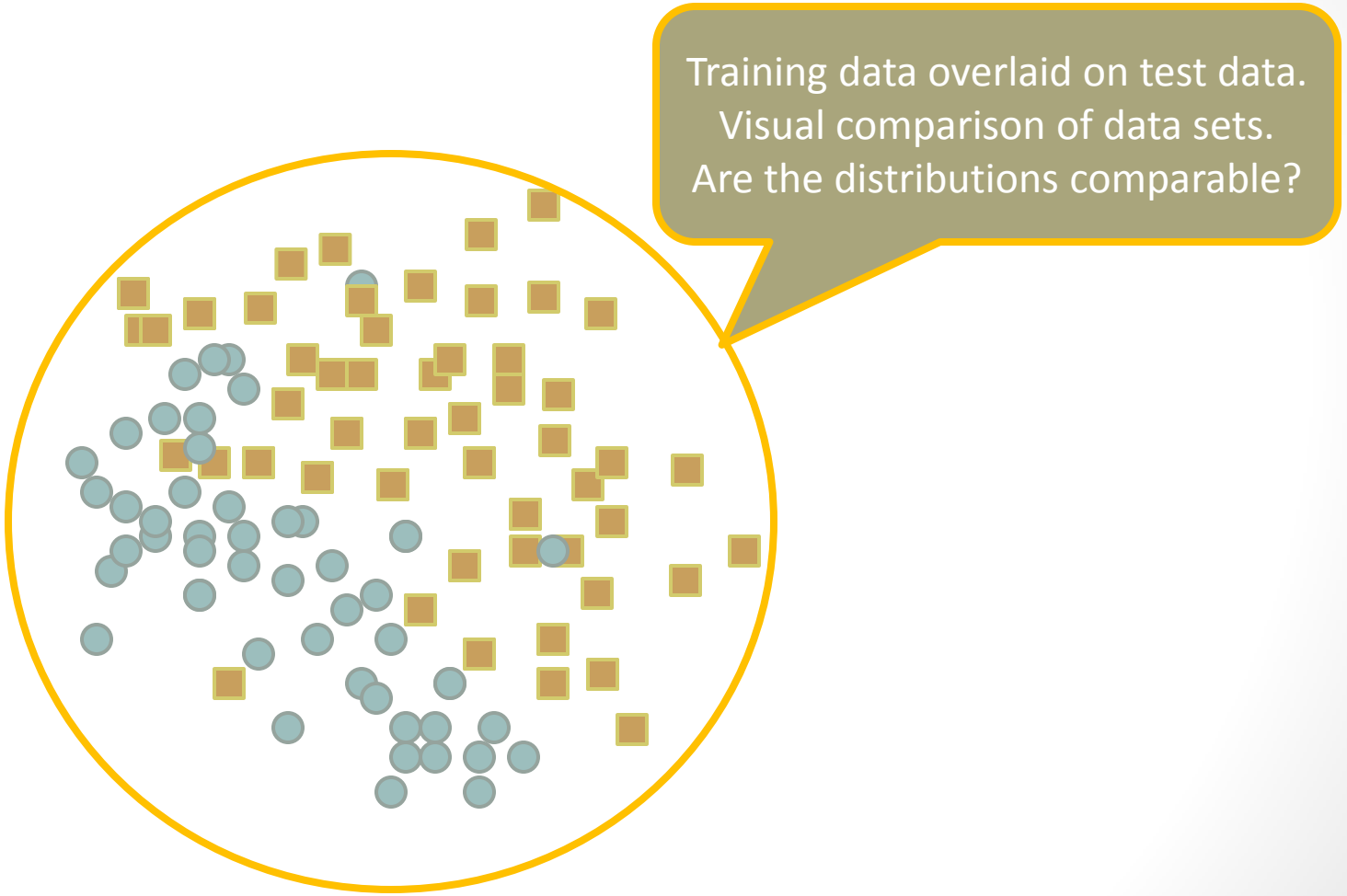
Evaluate Model : 3 Models



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

isSquare ~ xLocation + yLocation

Evaluate Model : All Data

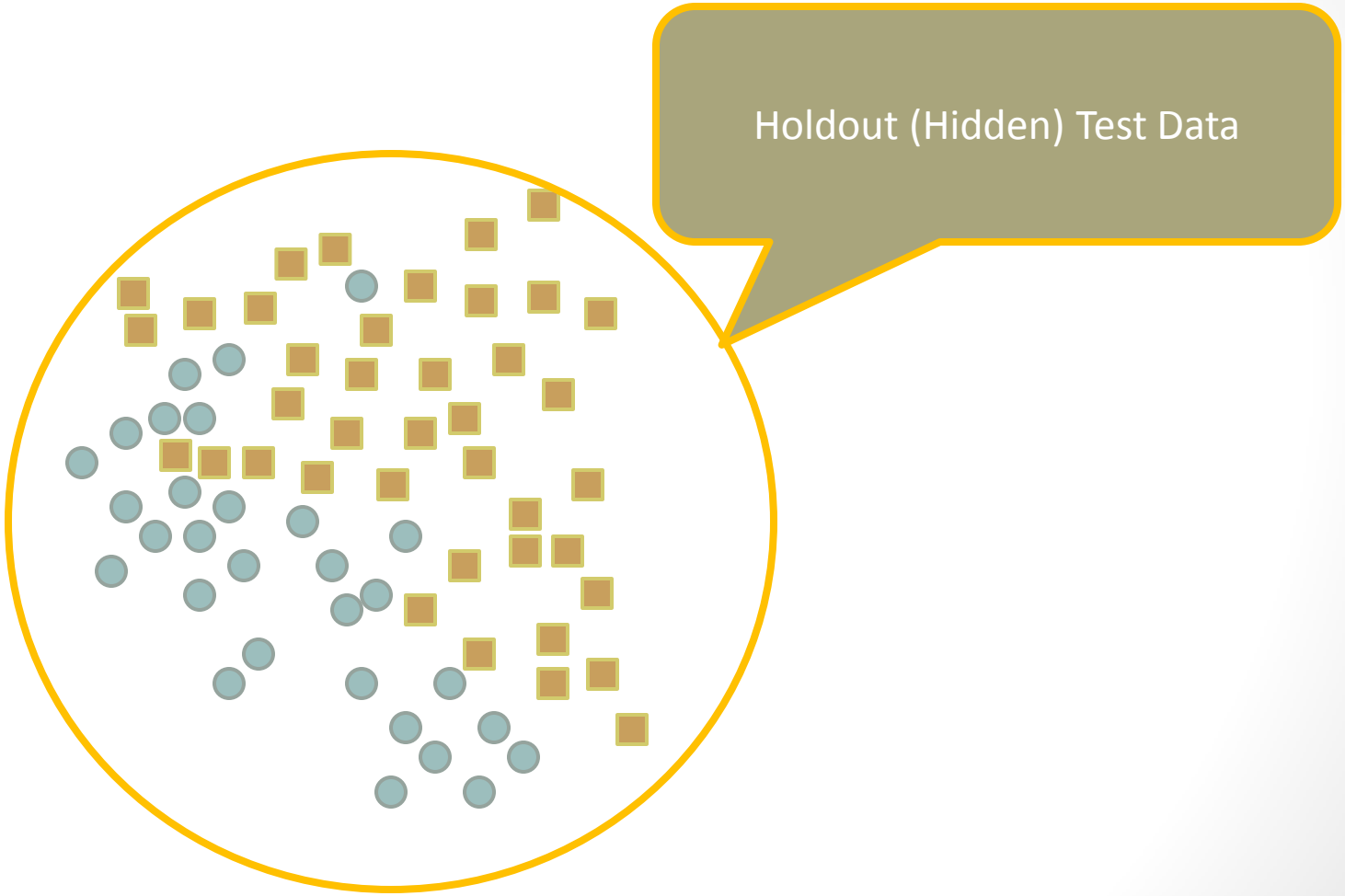


isSquare ~ xLocation + yLocation

Evaluate Model : Training Data

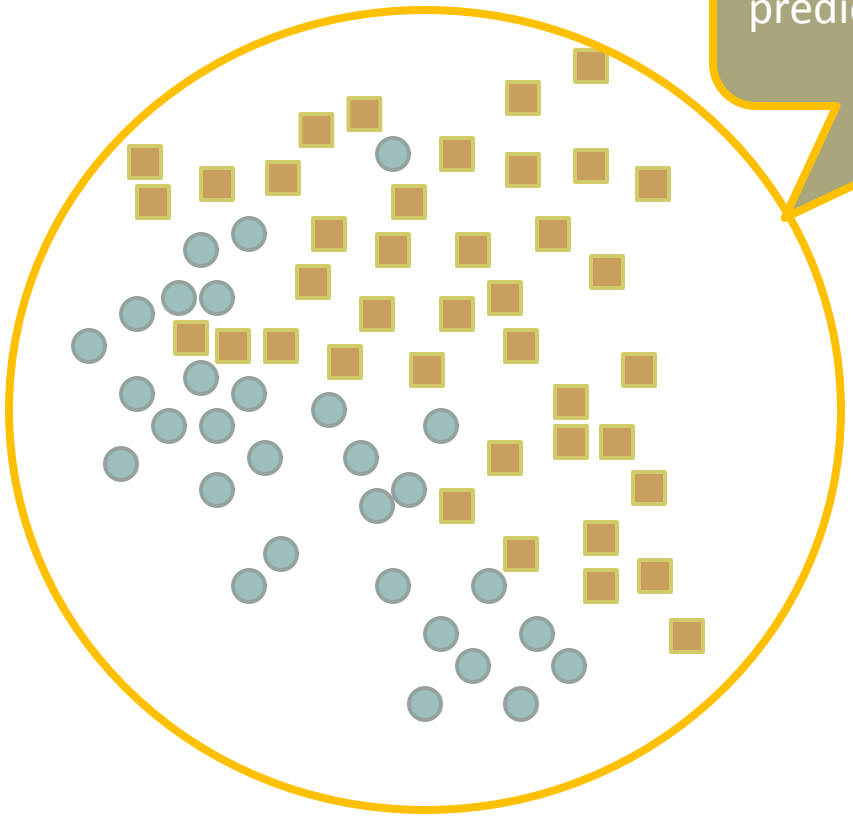
isSquare ~ xLocation + yLocation

Evaluate Model : Test Data



isSquare ~ xLocation + yLocation

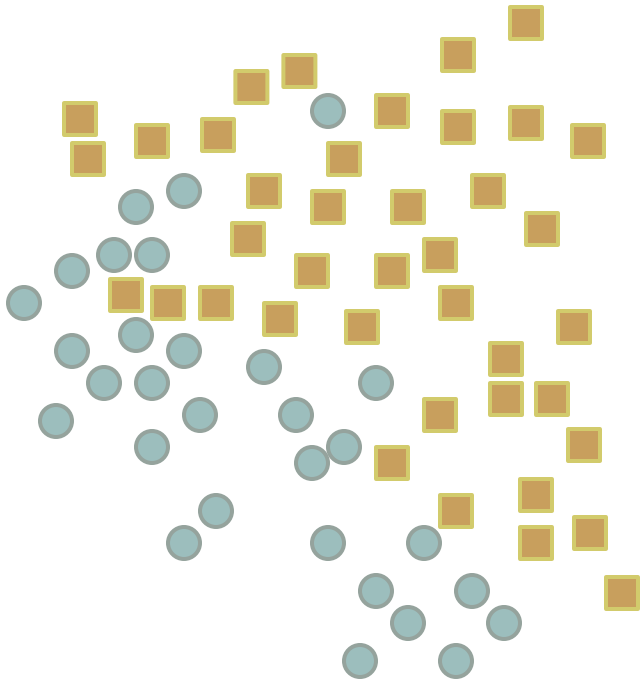
Evaluate Model : Test Data

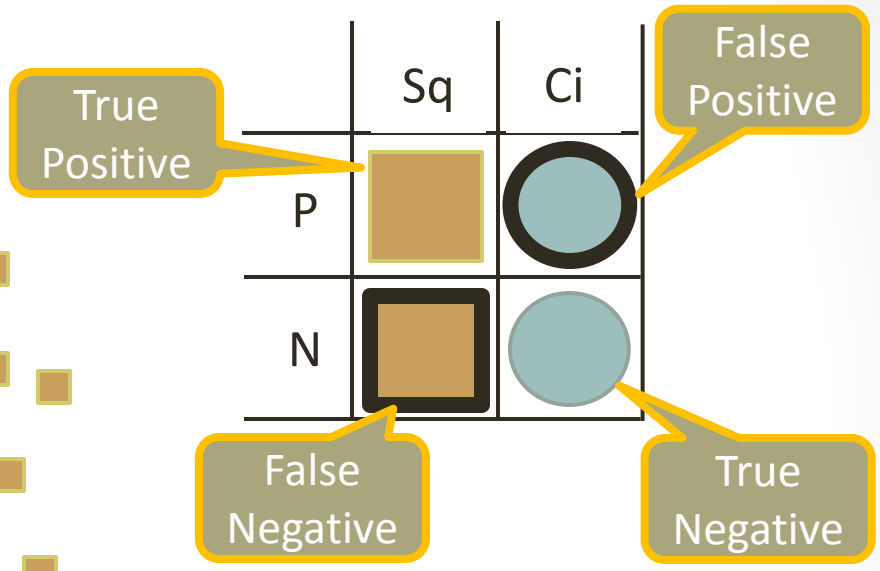


In the test data set:
I want to test if a square is predicted as positive and if a circle is predicted as negative

isSquare ~ xLocation + yLocation

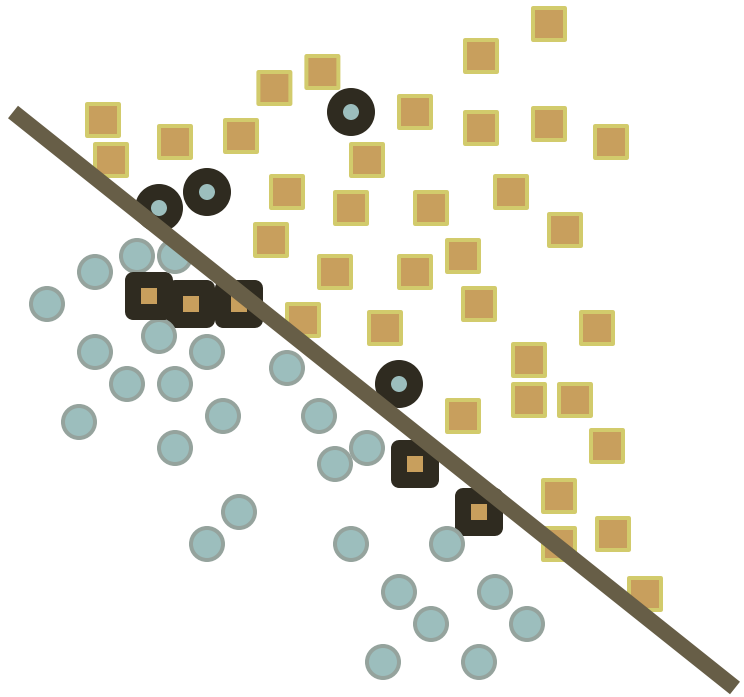
Evaluate Model : Test Data





isSquare ~ xLocation + yLocation

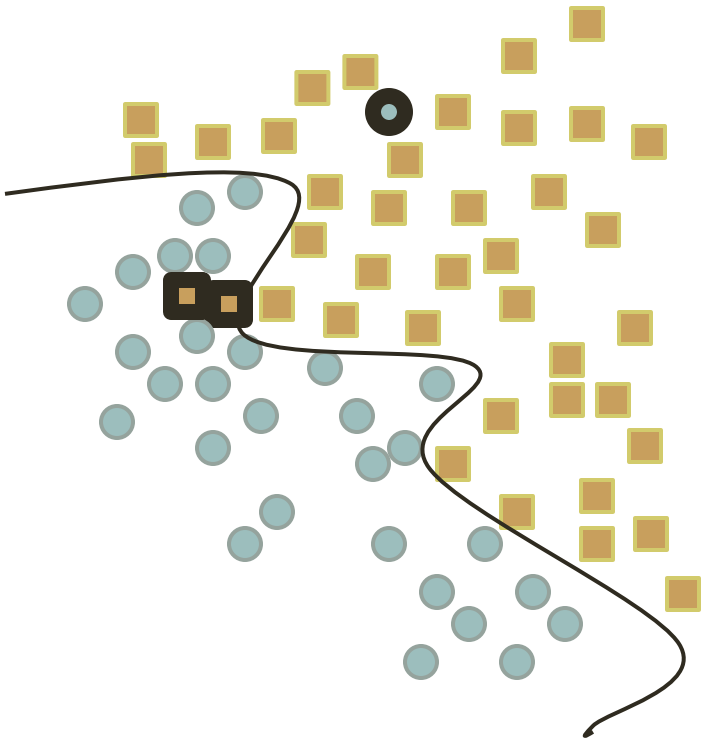
Evaluate Model : Test Model 1



	Sq	Ci
P	35	4
N	5	26

isSquare ~ xLocation + yLocation

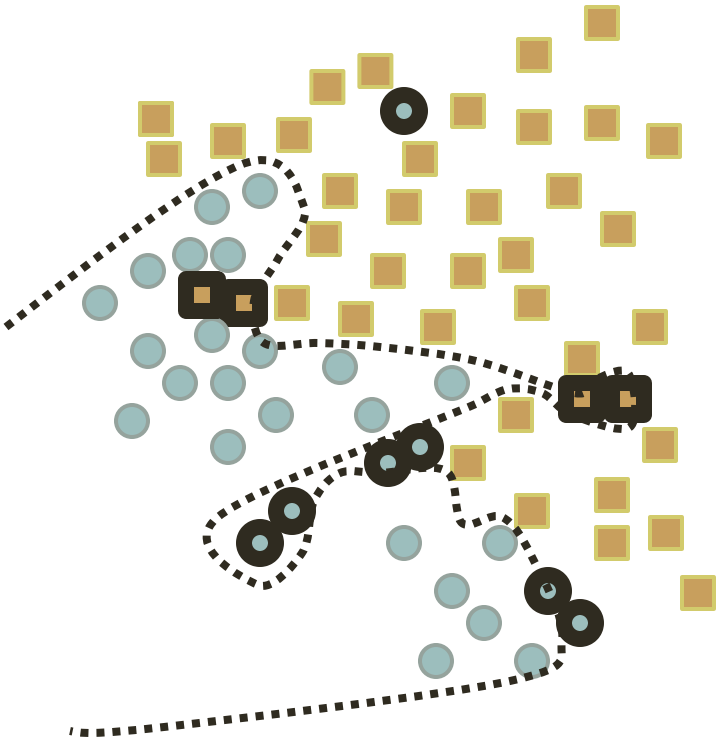
Evaluate Model : Test Model 2



	Sq	Ci
P	38	1
N	2	29

isSquare ~ xLocation + yLocation

Evaluate Model : Test Model 3



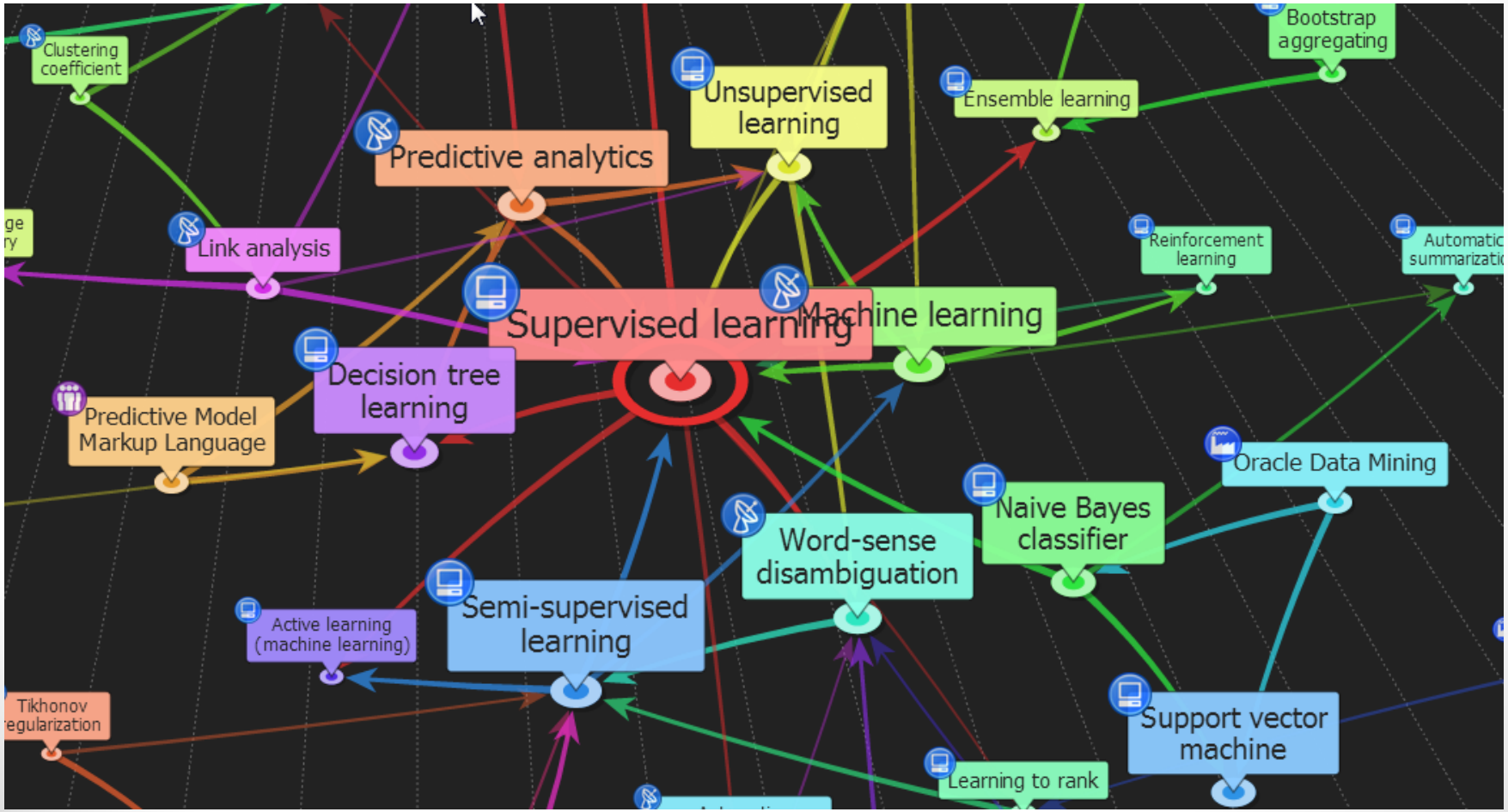
	Sq	Ci
P	36	7
N	4	23

isSquare ~ xLocation + yLocation

Over-fitting and Confusion Matrix

Video and Break

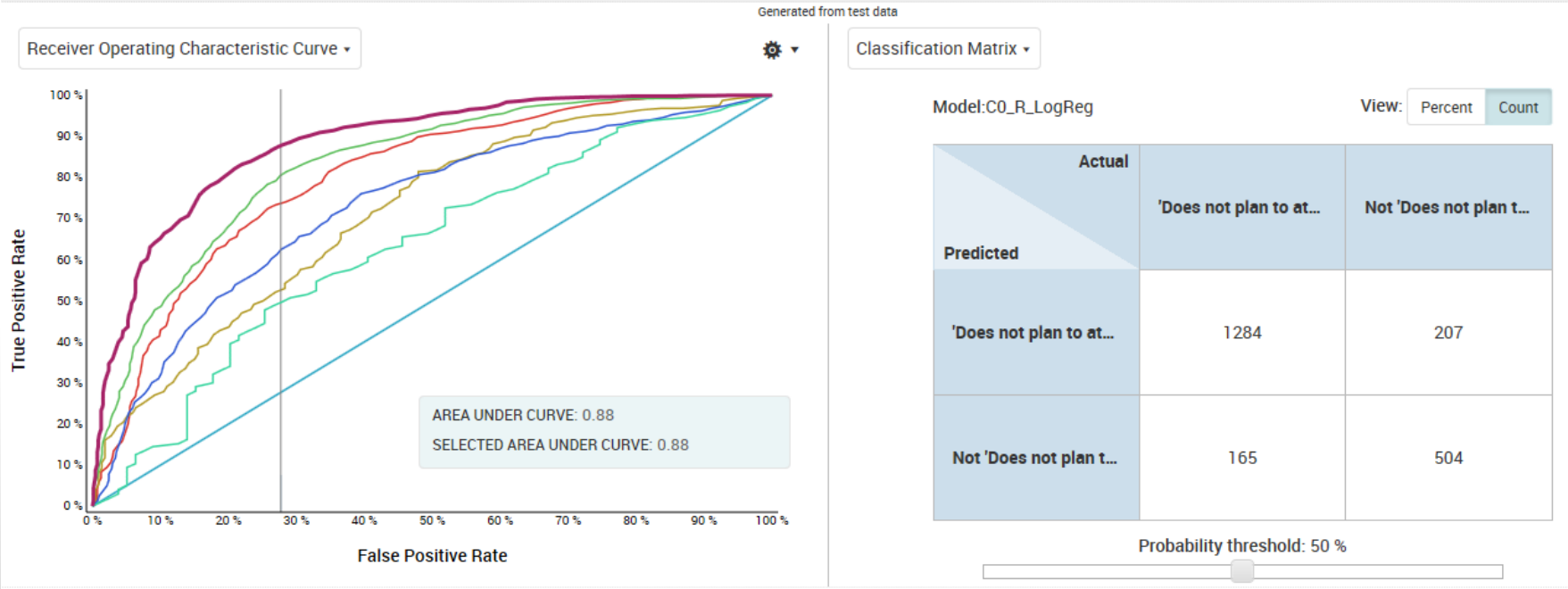
- Advertisement for IBM's predictive analytics: https://www.youtube.com/watch?v=iY3WRvXVogo



ROC Chart Intro

ROC Chart Intro (1)

- Confusion Matrix and ROC Chart

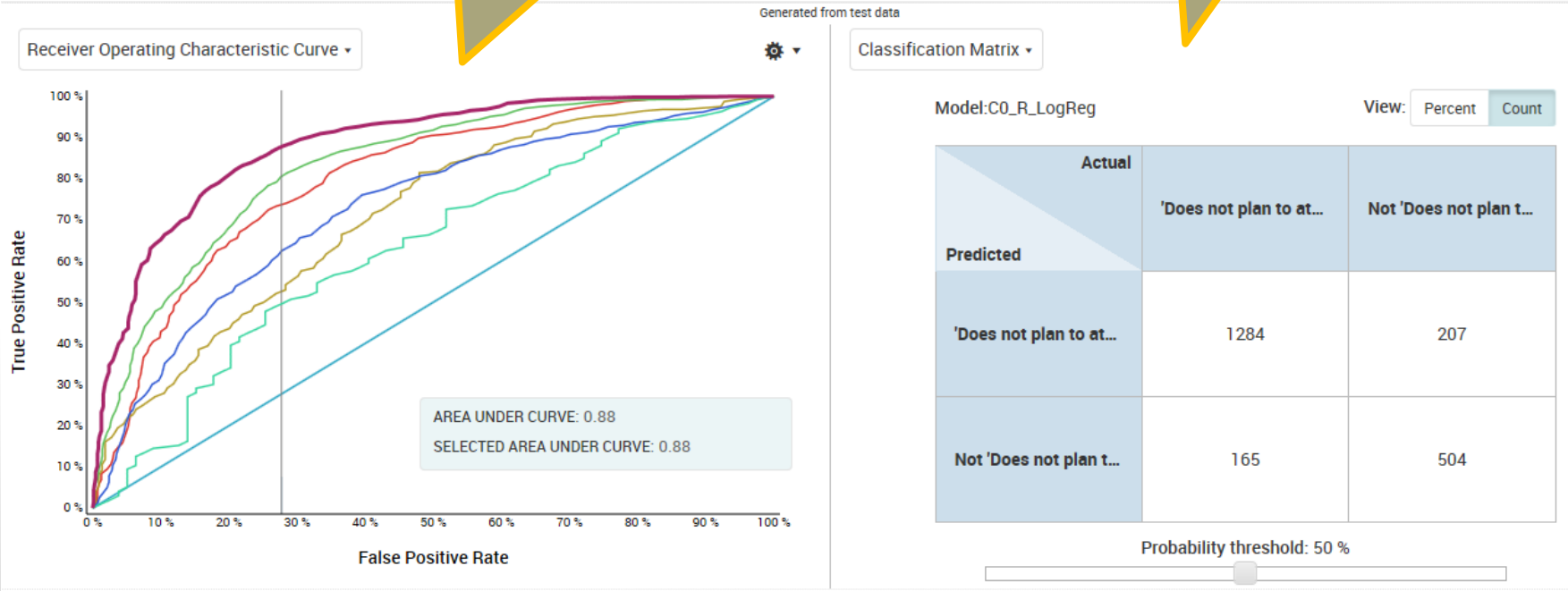


ROC Chart Intro (2)

- Confusion Matrix and ROC Chart

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

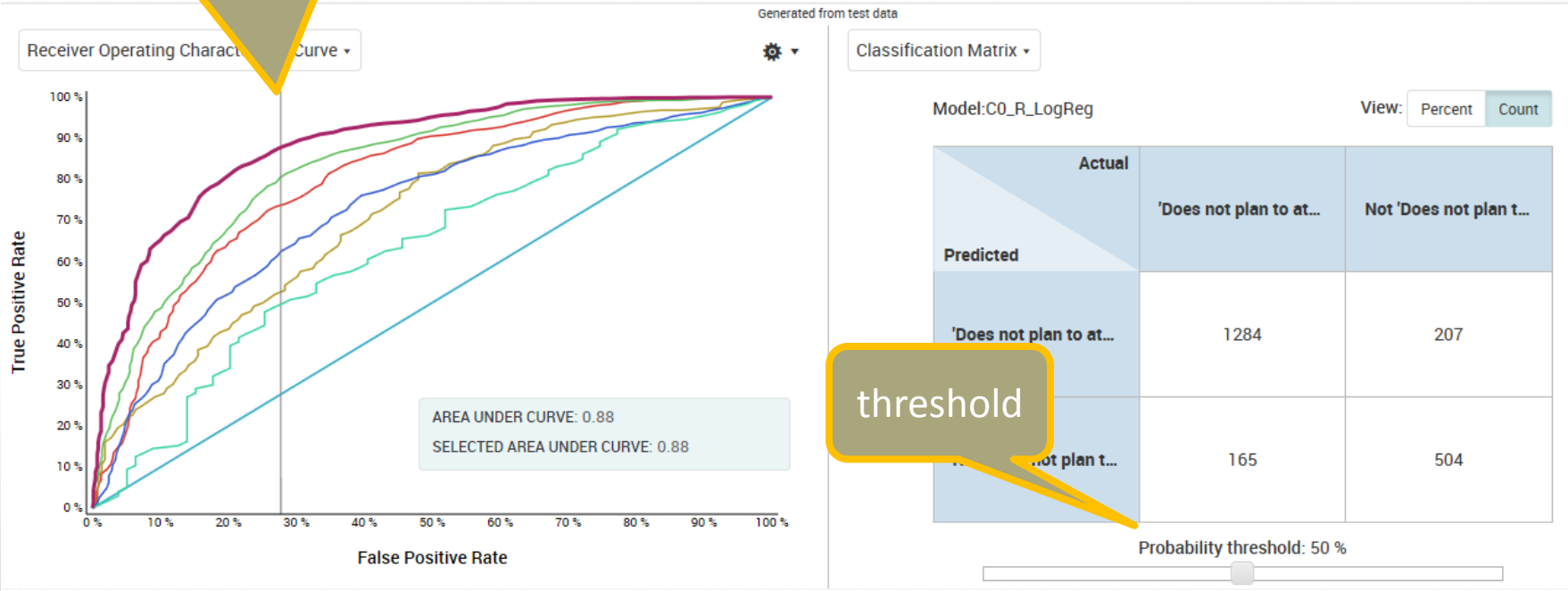
The confusion matrix for one model at one threshold



ROC Chart Intro (3)

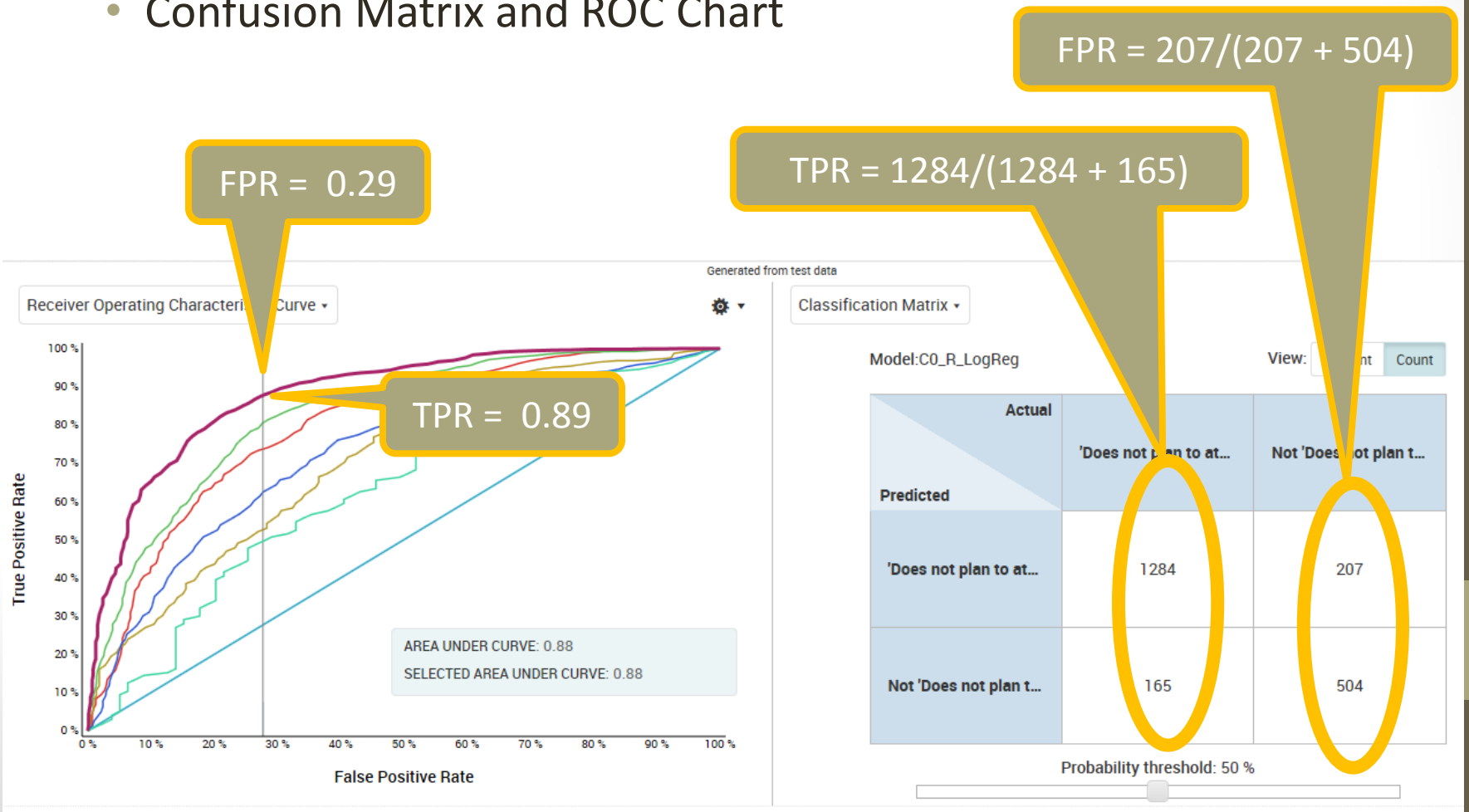
- Confusion Matrix and ROC Chart

This FPR (0.28) corresponds to the threshold (0.5) for the confusion matrix for the best model



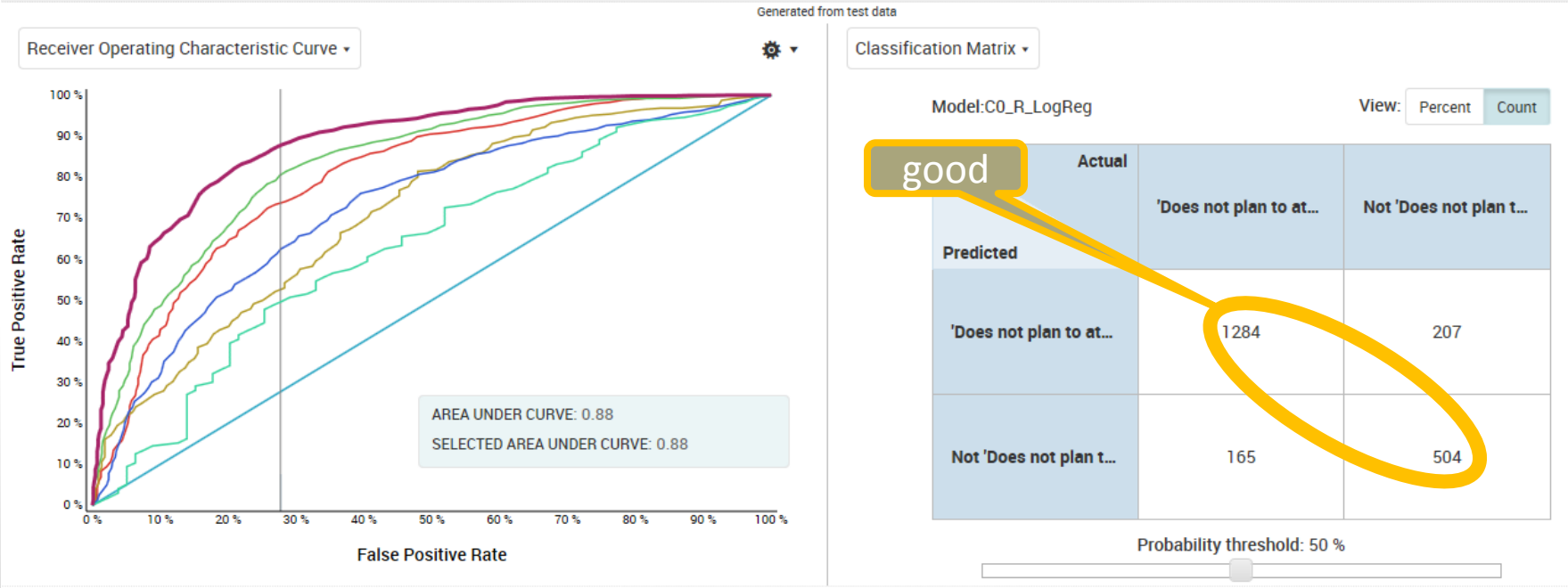
ROC Chart Intro (4)

- Confusion Matrix and ROC Chart



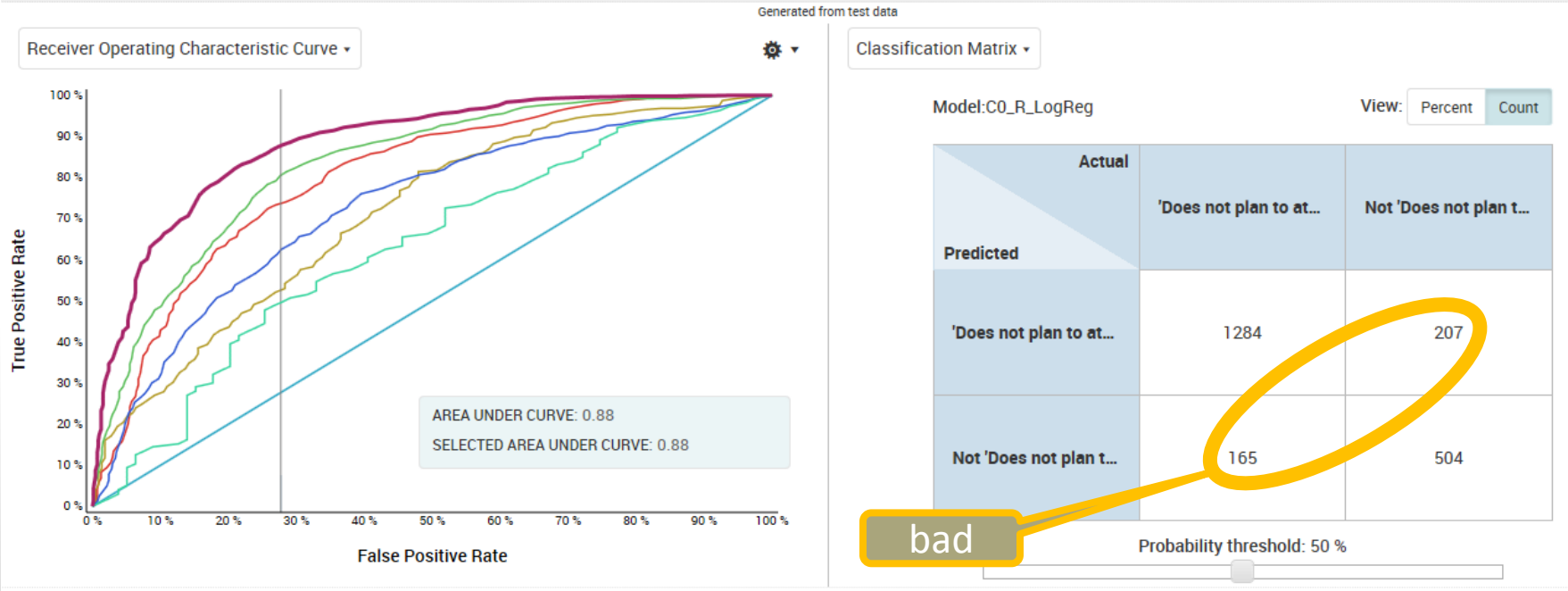
ROC Chart Intro (5)

- Confusion Matrix and ROC Chart



ROC Chart Intro (6)

- Confusion Matrix and ROC Chart

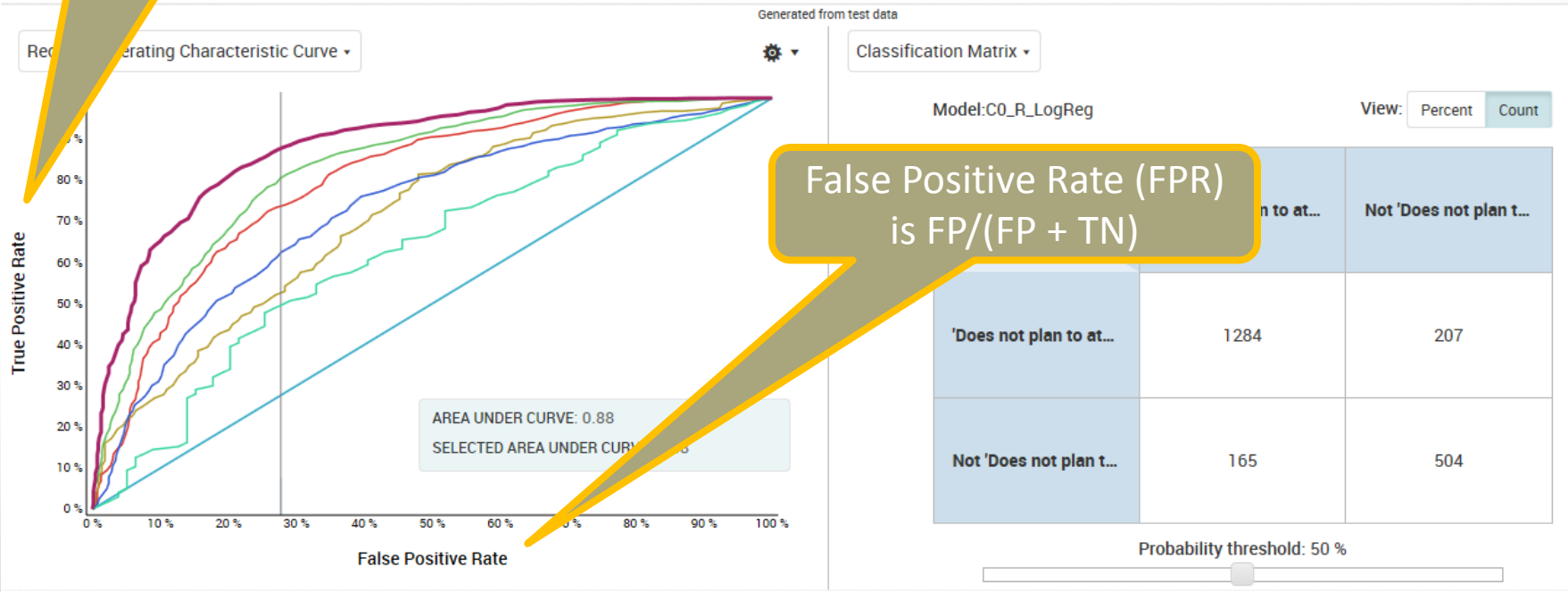


ROC Chart Intro (7)

- Confusion Matrix and ROC Chart

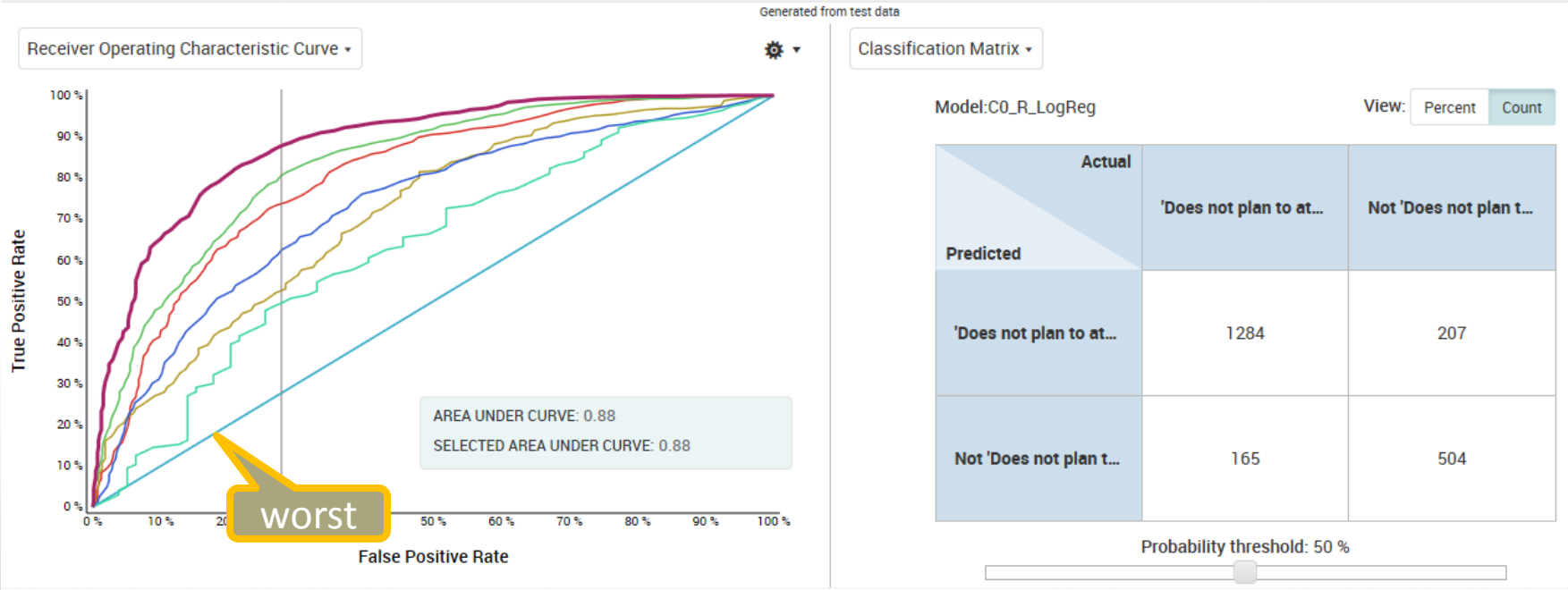
True Positive Rate (TPR) is TP/(TP + FN)

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



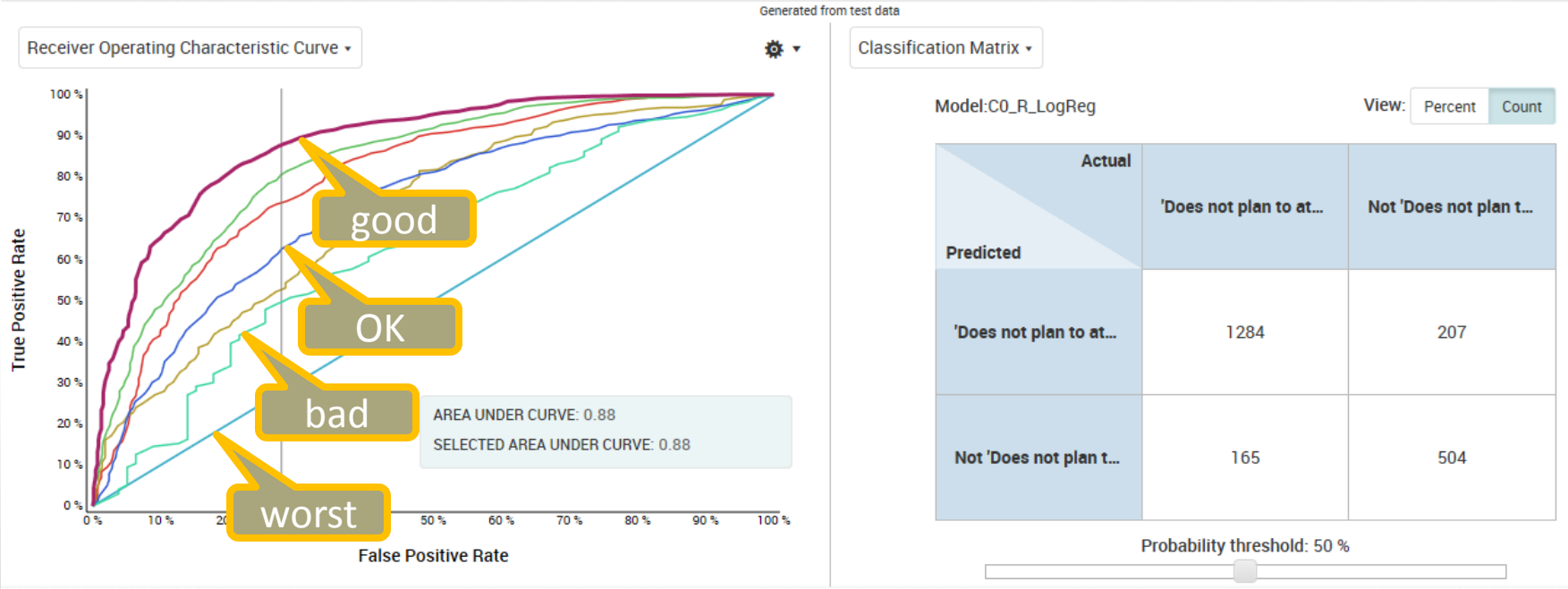
ROC Chart Intro (8)

- Confusion Matrix and ROC Chart



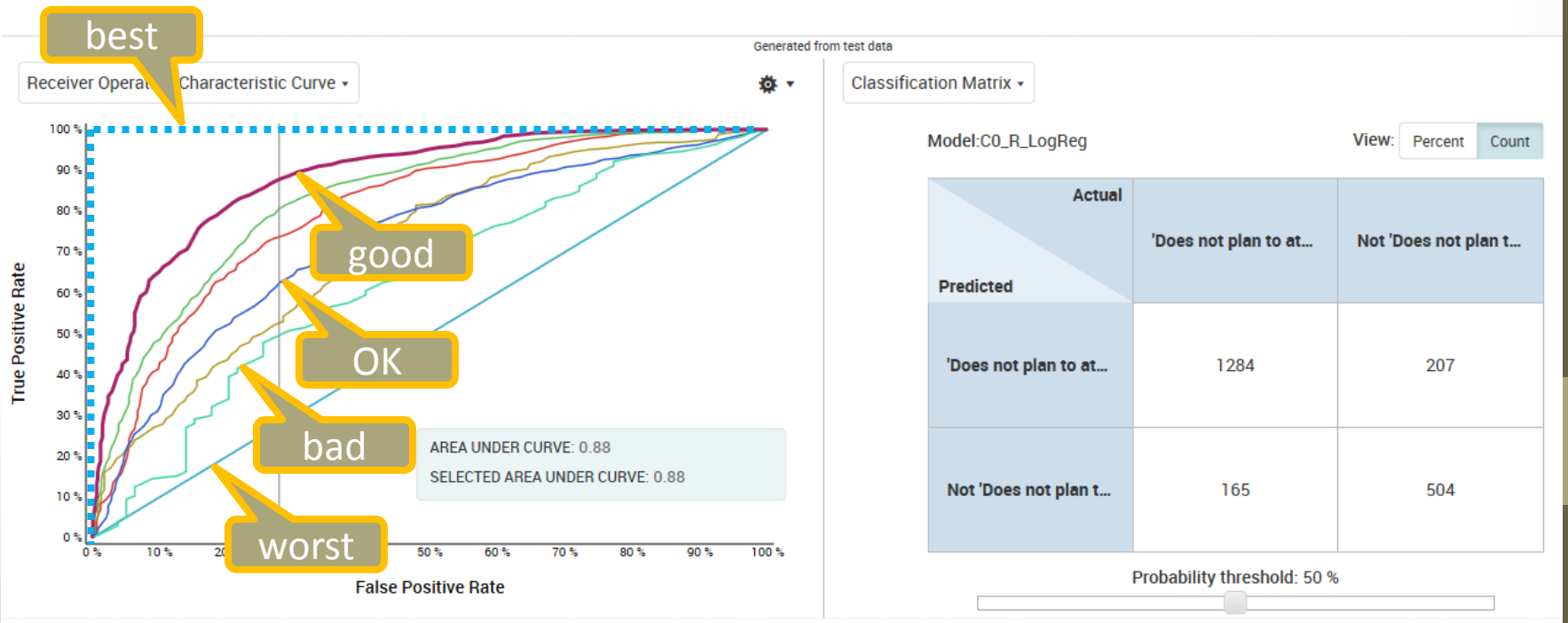
ROC Chart Intro (9)

- Confusion Matrix and ROC Chart



ROC Chart Intro (10)

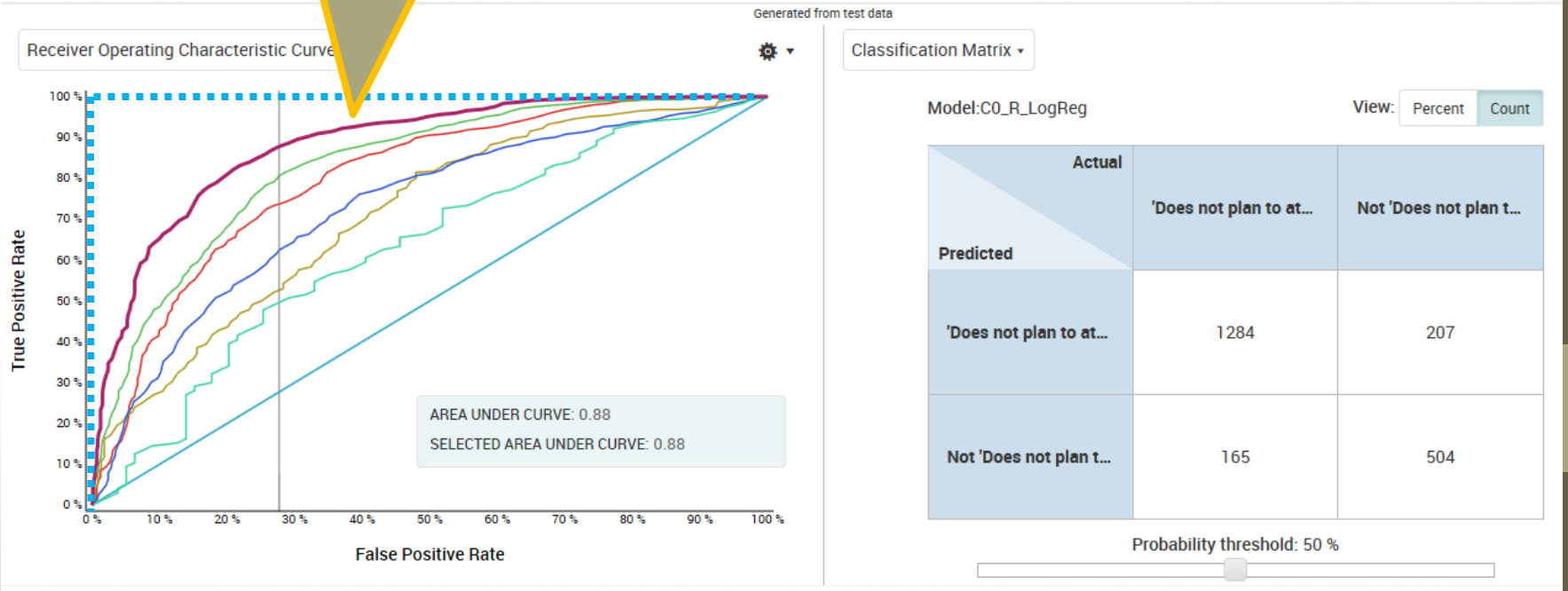
- Confusion Matrix and ROC Chart



ROC Chart Intro (11)

- Confusion Matrix and ROC Chart

ROC charts are non-decreasing functions



ROC Chart Intro

Quiz Confusion Matrix ROC

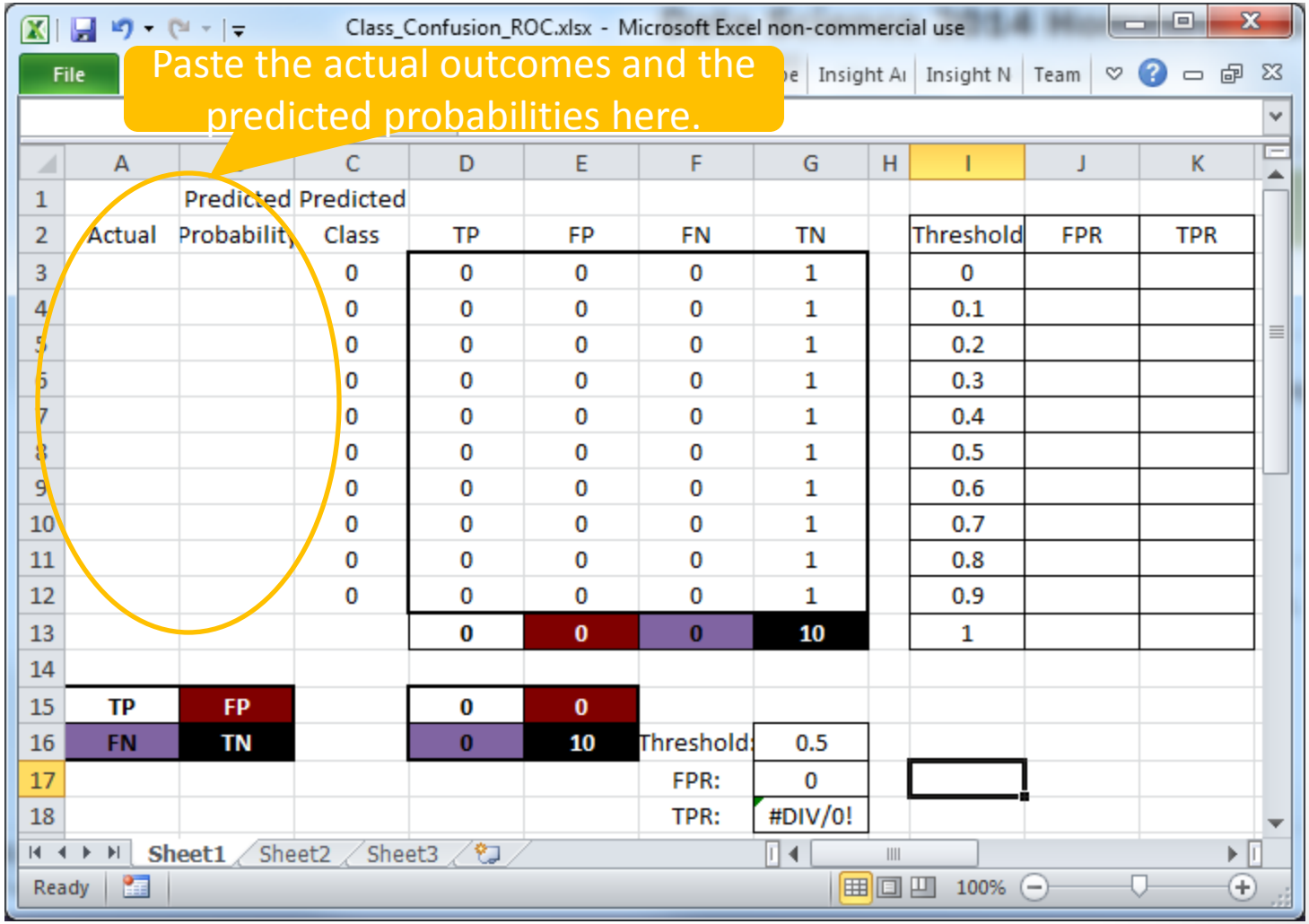
- Quiz Confusion Matrix ROC Intro
  - Test and Accuracy Measures

How to make an ROC

How to make an ROC (0)

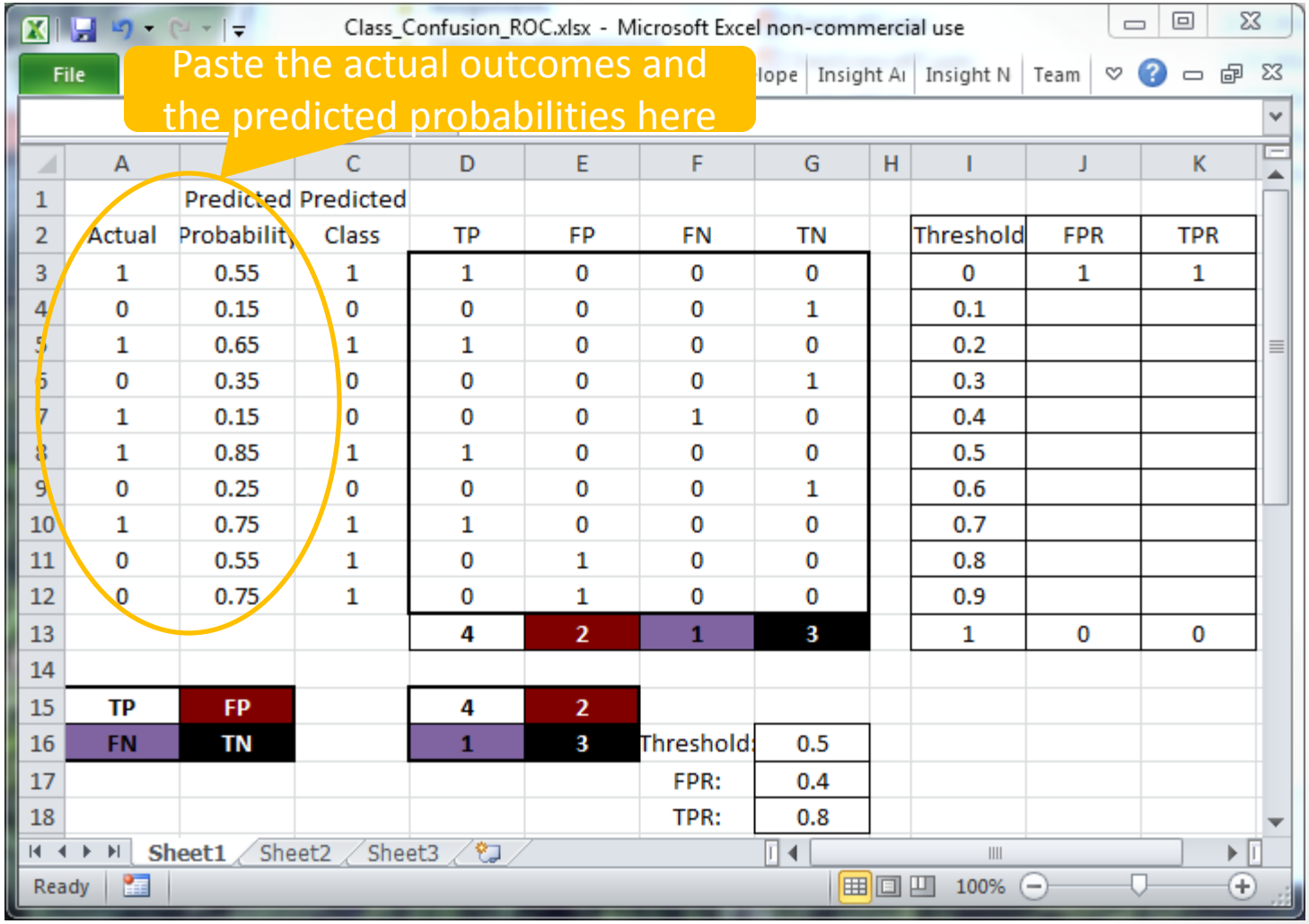
- From Probabilities to ROC:
- Probabilities -> Threshold -> Predictions -> Confusion Matrix -> ROC
- Get Excel workbook: HowToMakeAnROC.xls
- Note that at the bottom of the worksheet are the actual outcomes and the predicted probabilities.

How to make an ROC (1)



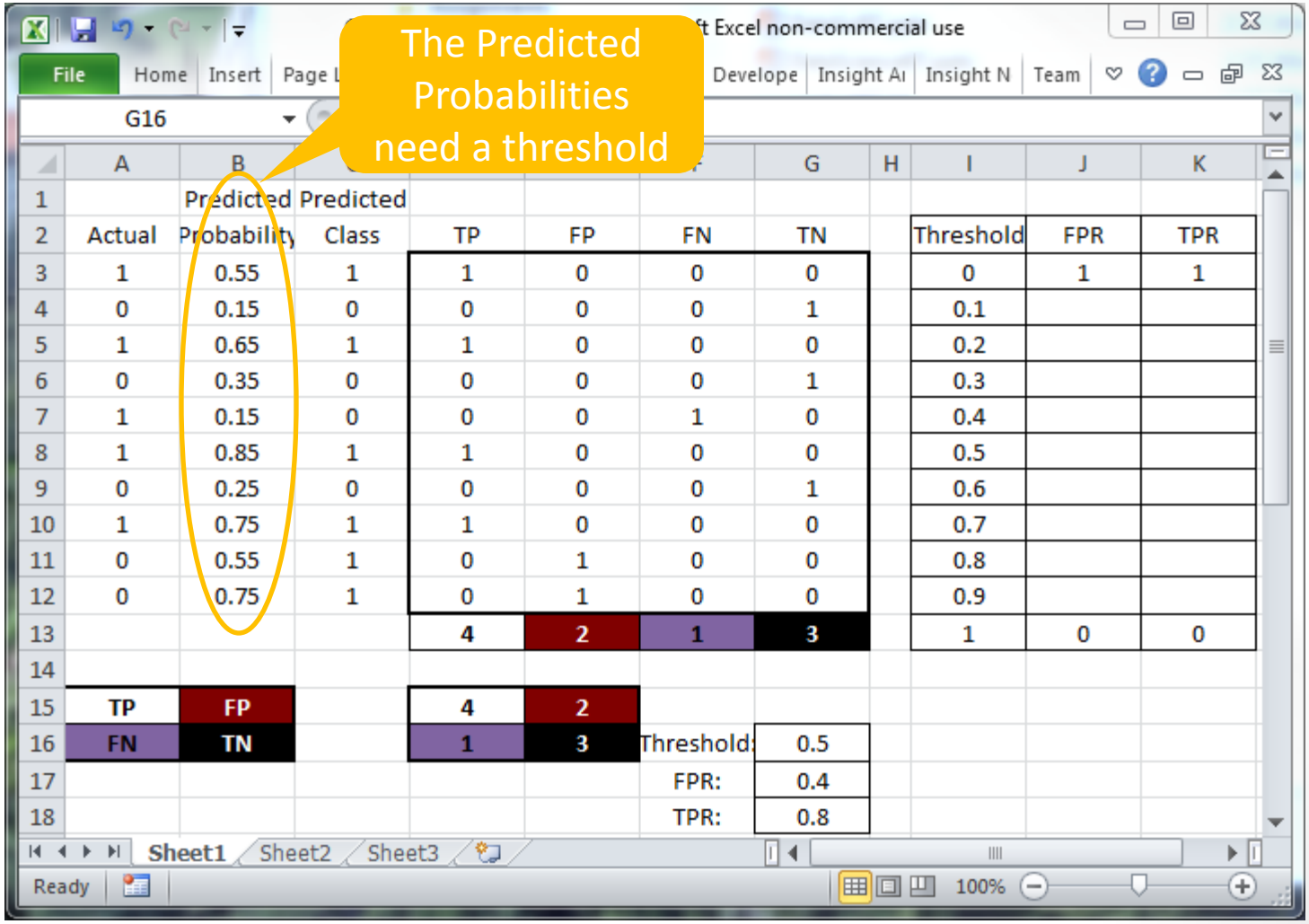
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3			0	0	0	0	1		0		
4			0	0	0	0	1		0.1		
5			0	0	0	0	1		0.2		
6			0	0	0	0	1		0.3		
7			0	0	0	0	1		0.4		
8			0	0	0	0	1		0.5		
9			0	0	0	0	1		0.6		
10			0	0	0	0	1		0.7		
11			0	0	0	0	1		0.8		
12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

How to make an ROC (2)



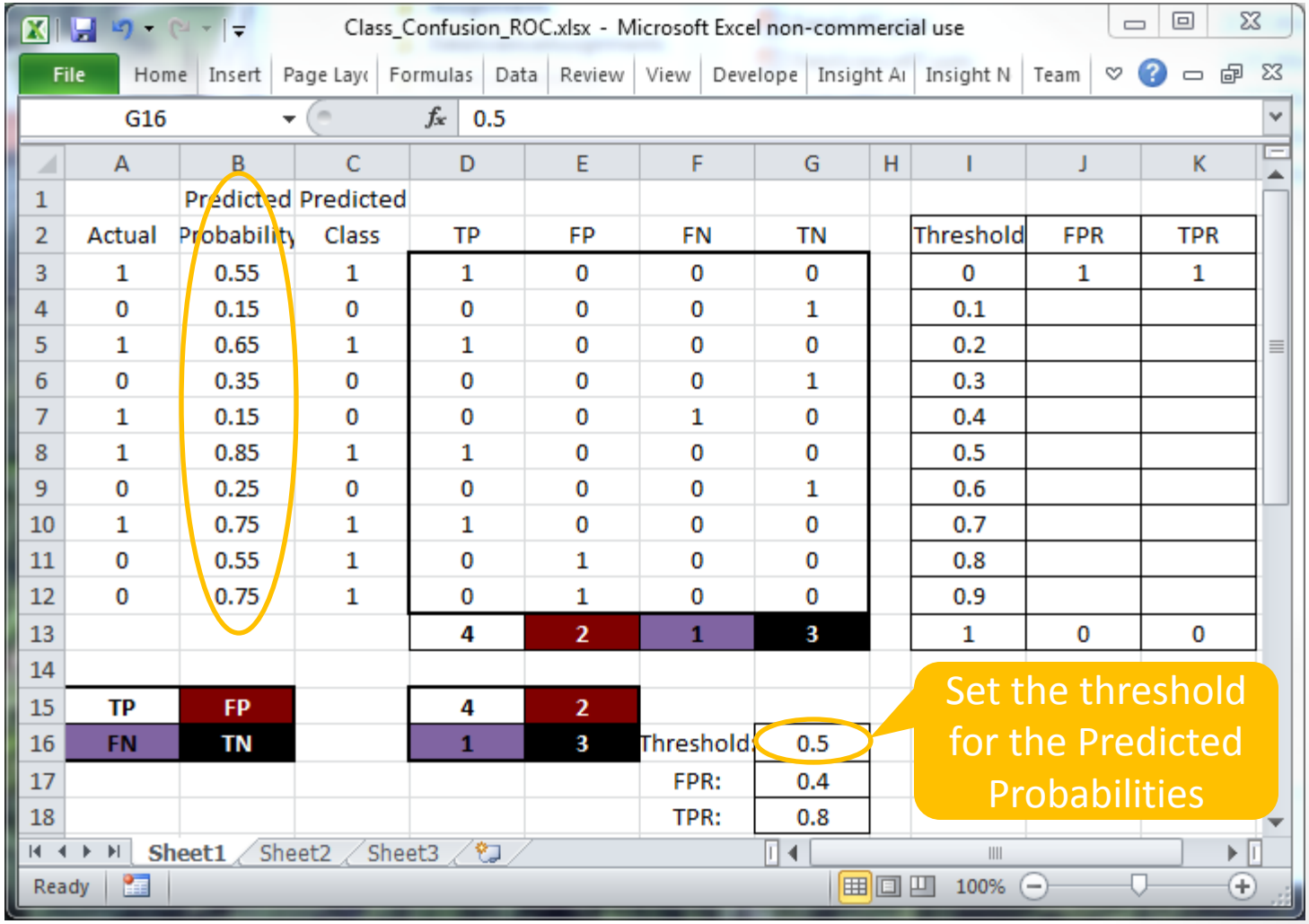
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

How to make an ROC (3)

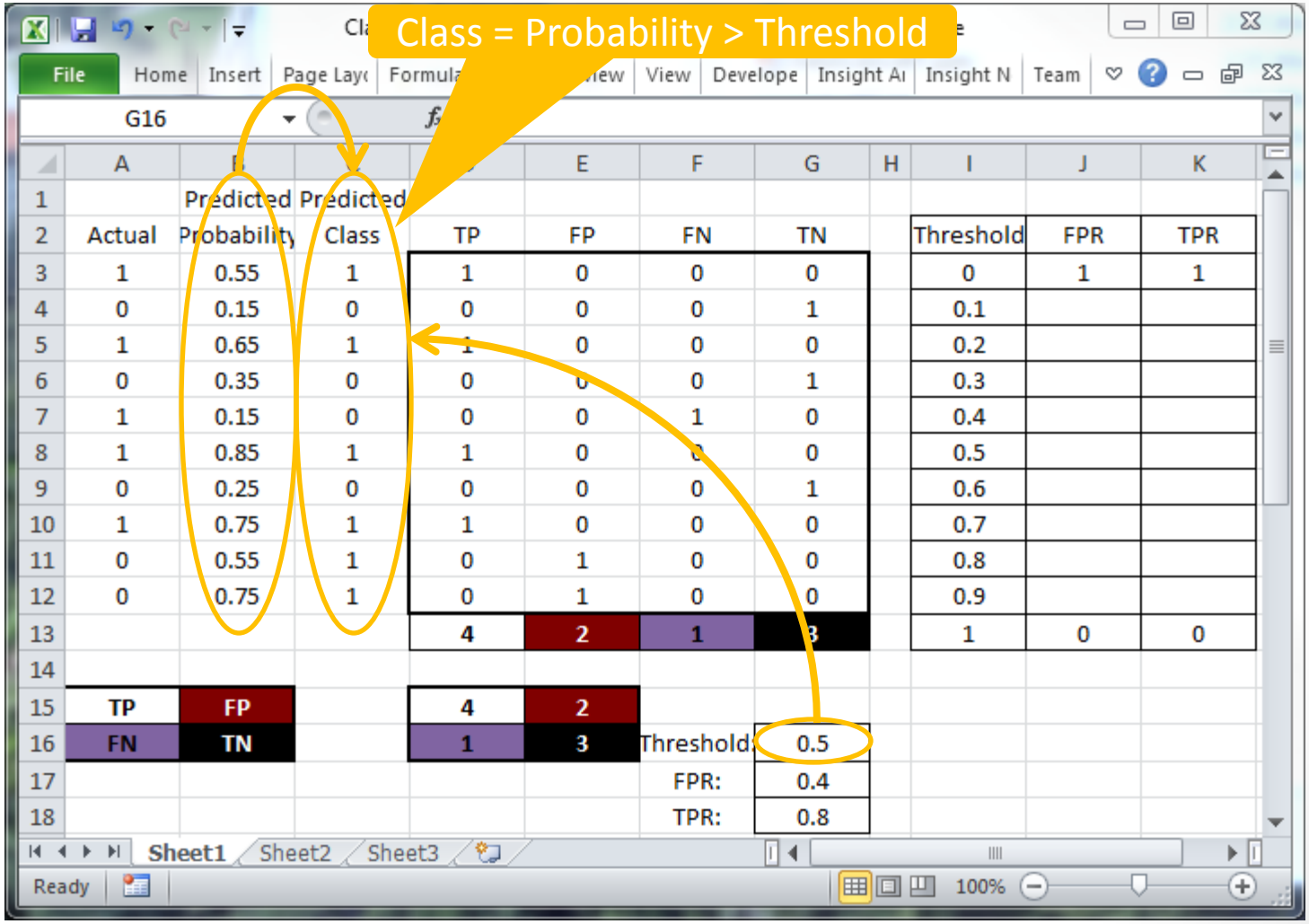


	A	B	C	TP	FP	FN	TN		Threshold	FPR	TPR
1		Predicted	Predicted								
2	Actual	Probability	Class								
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	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				

How to make an ROC (4)

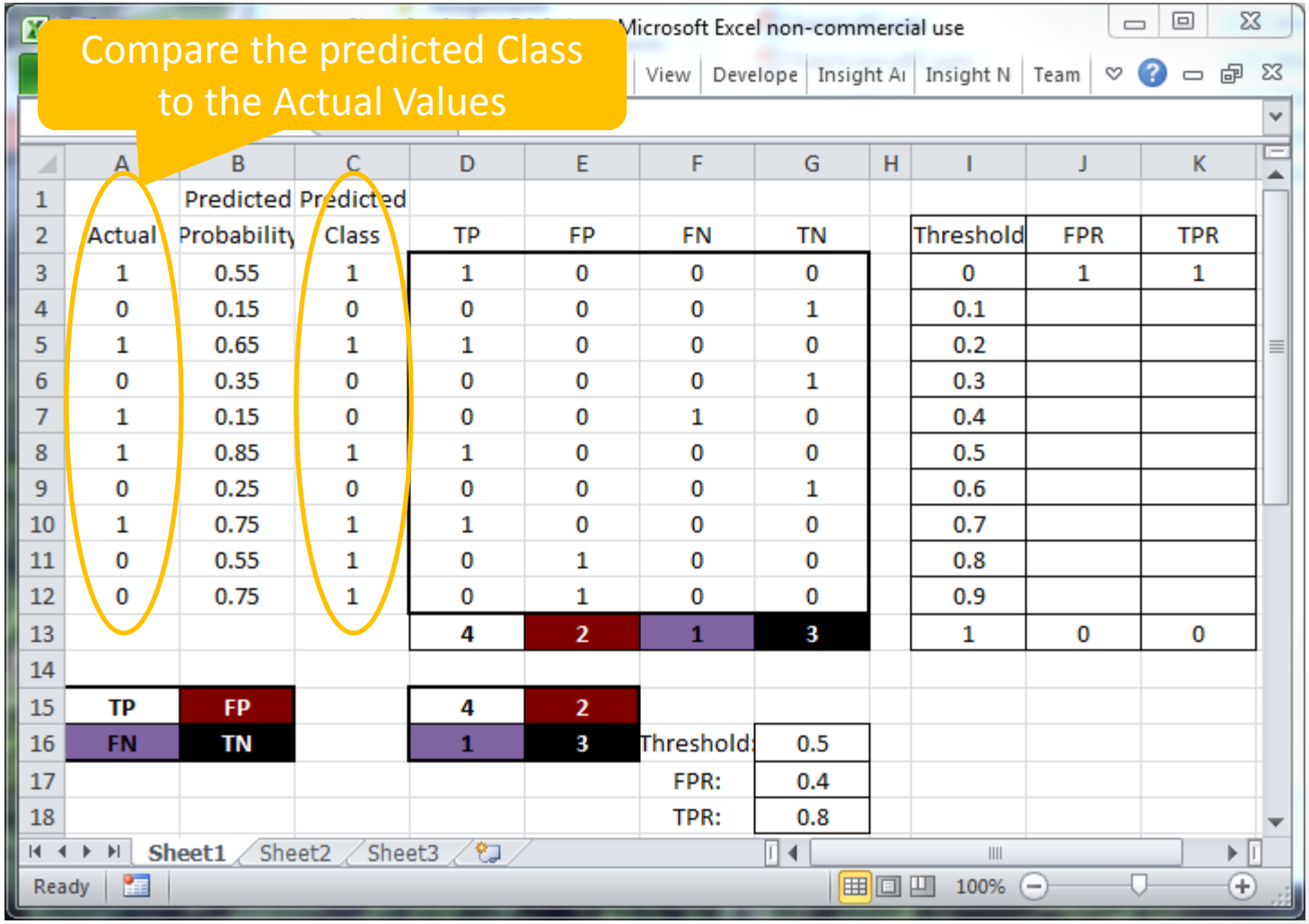


How to make an ROC (5)



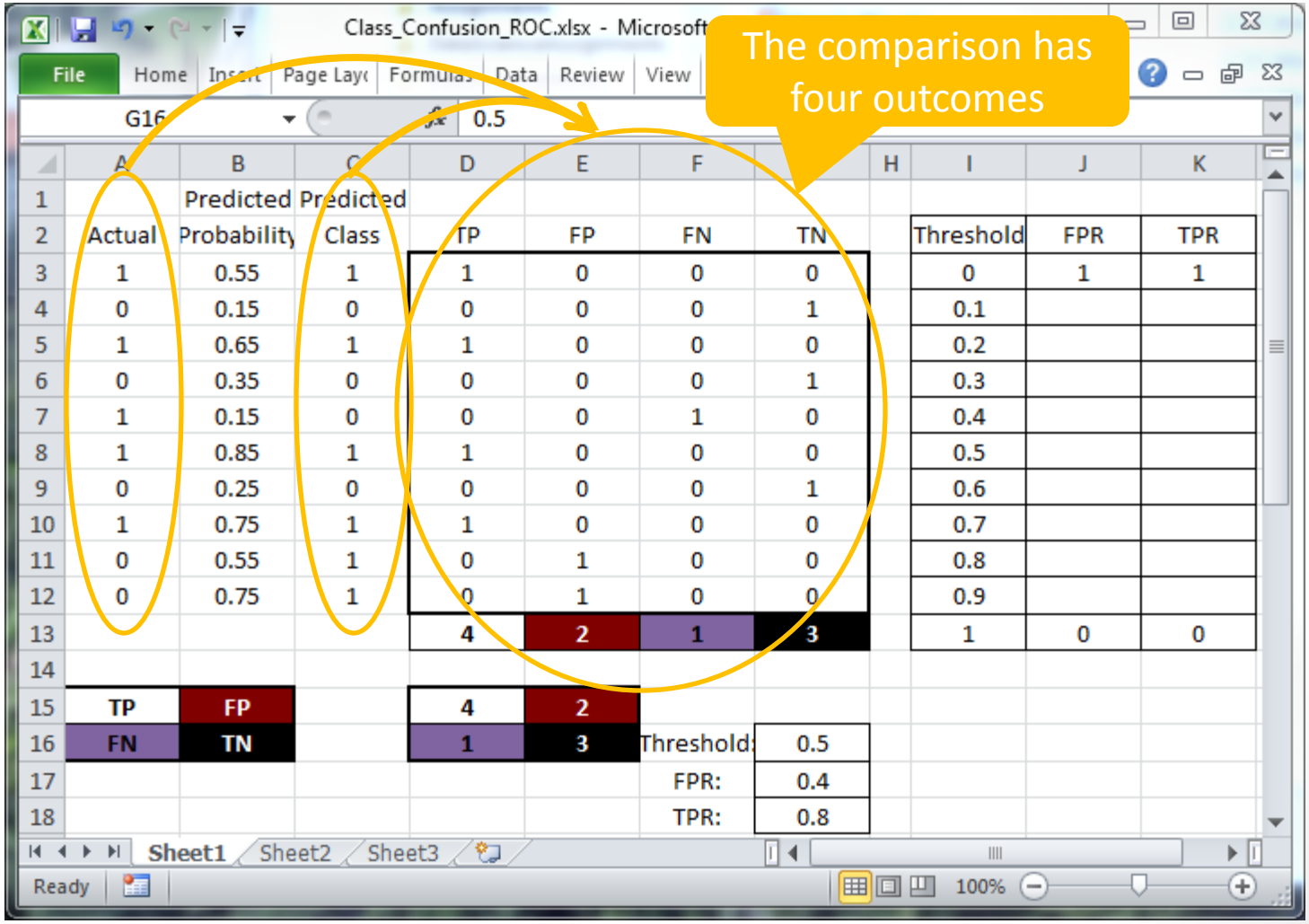
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

How to make an ROC (6)



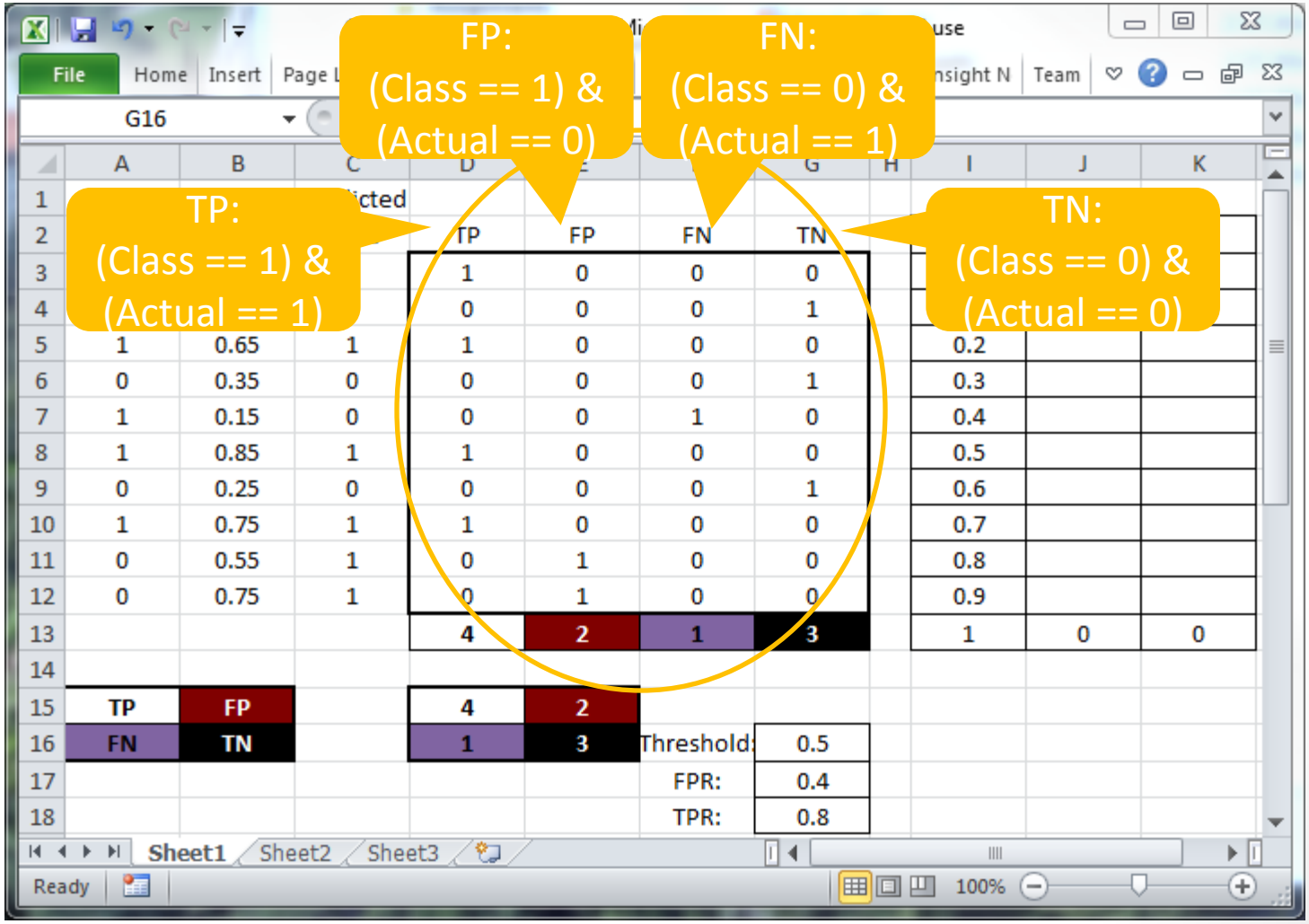
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

How to make an ROC (7)

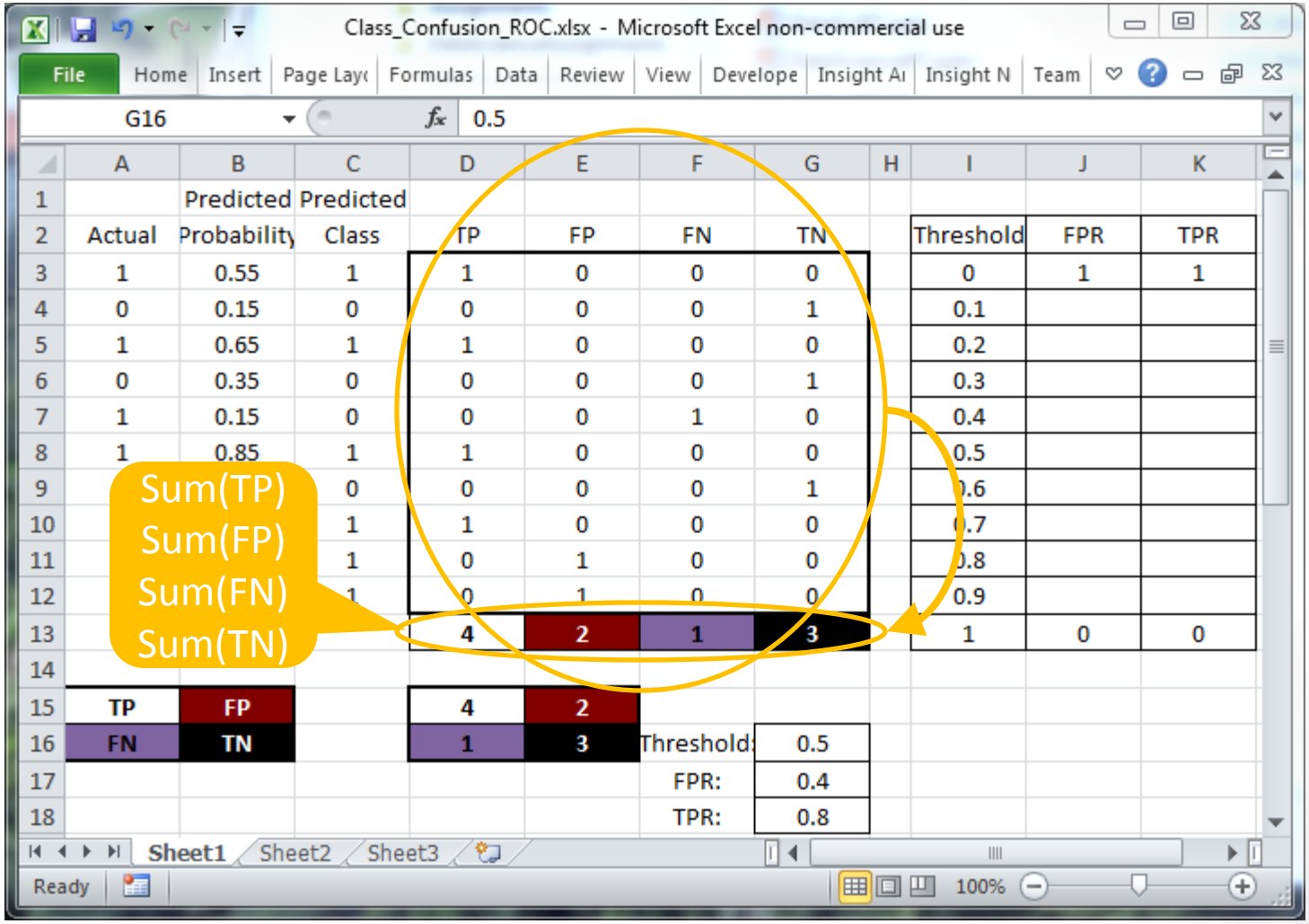


	A	B	C	D	E	F	G	H	I	J	K
1		Predicted Probability	Predicted Class								
2	Actual			TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

How to make an ROC (8)

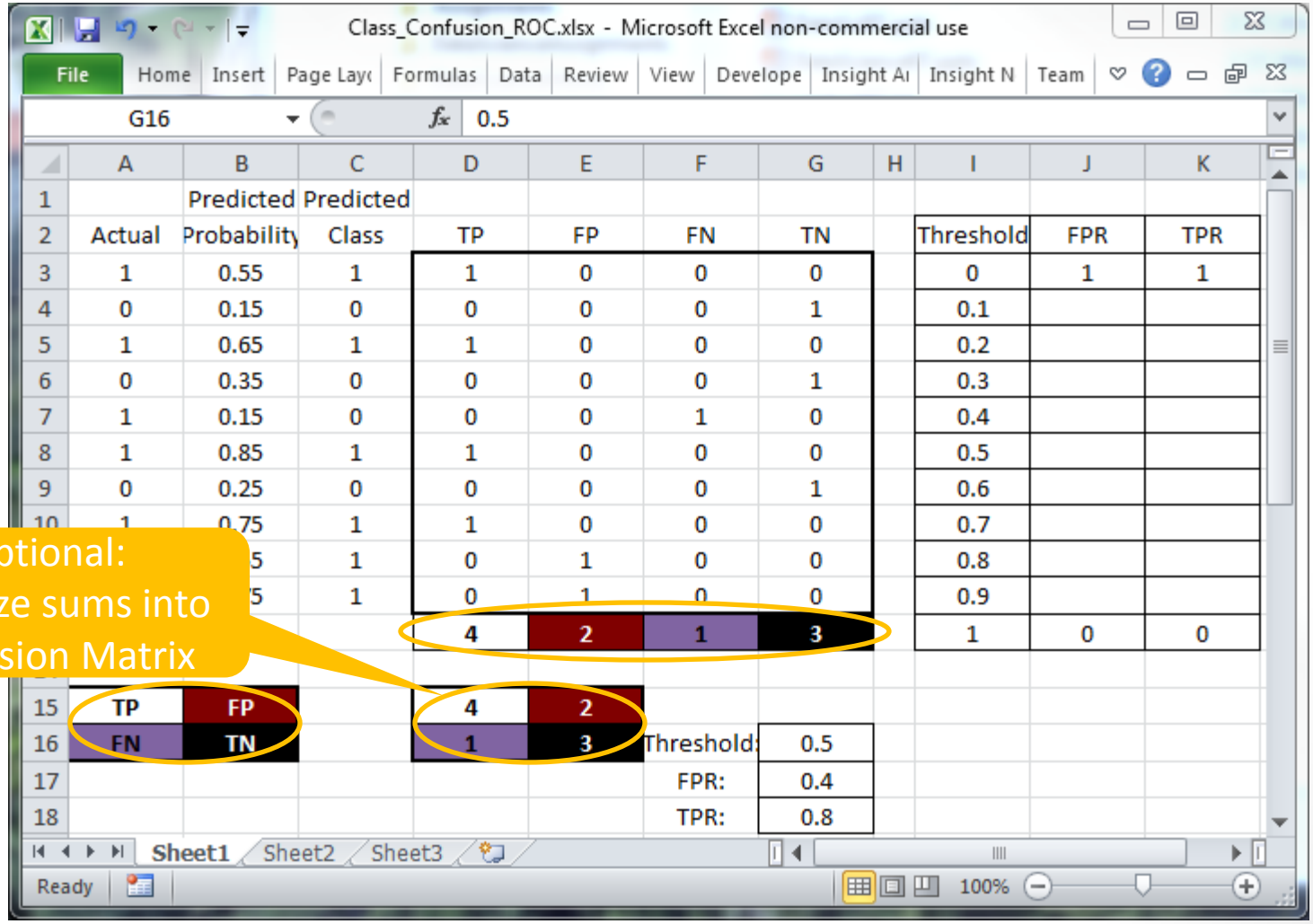


How to make an ROC (9)



	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						
17											
18											

How to make an ROC (10)

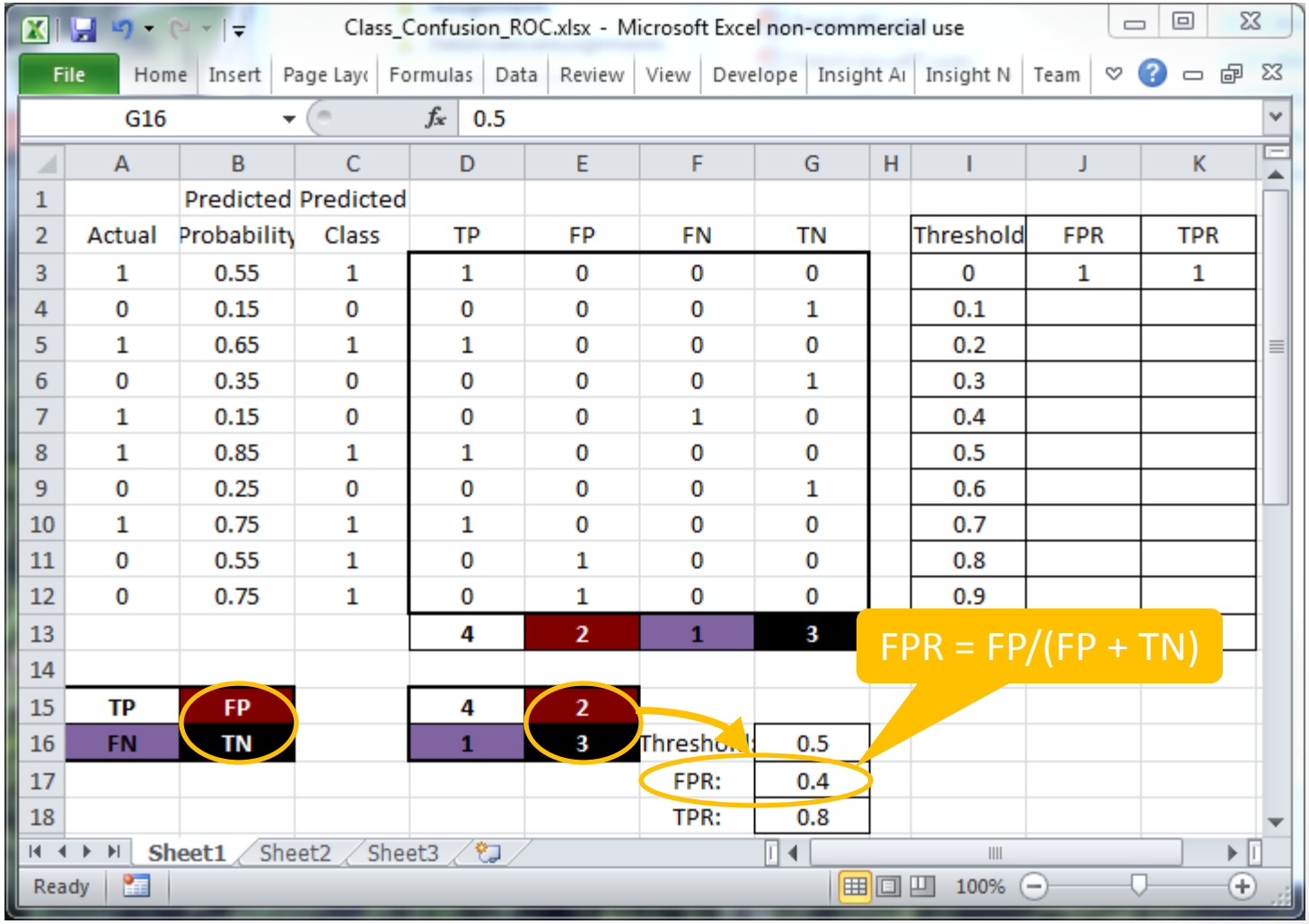


	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

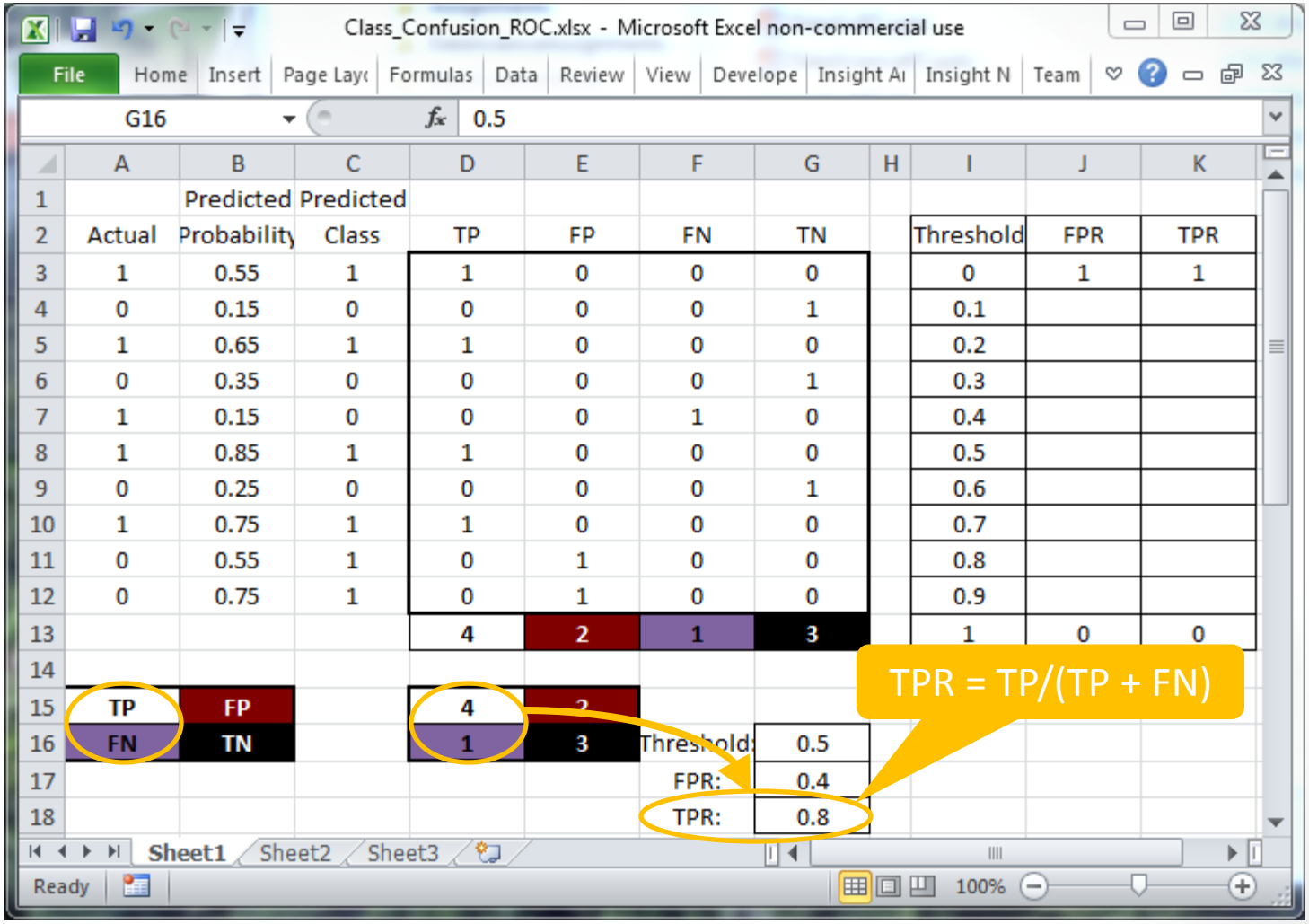
TP	FP
FN	TN

4	2
1	3

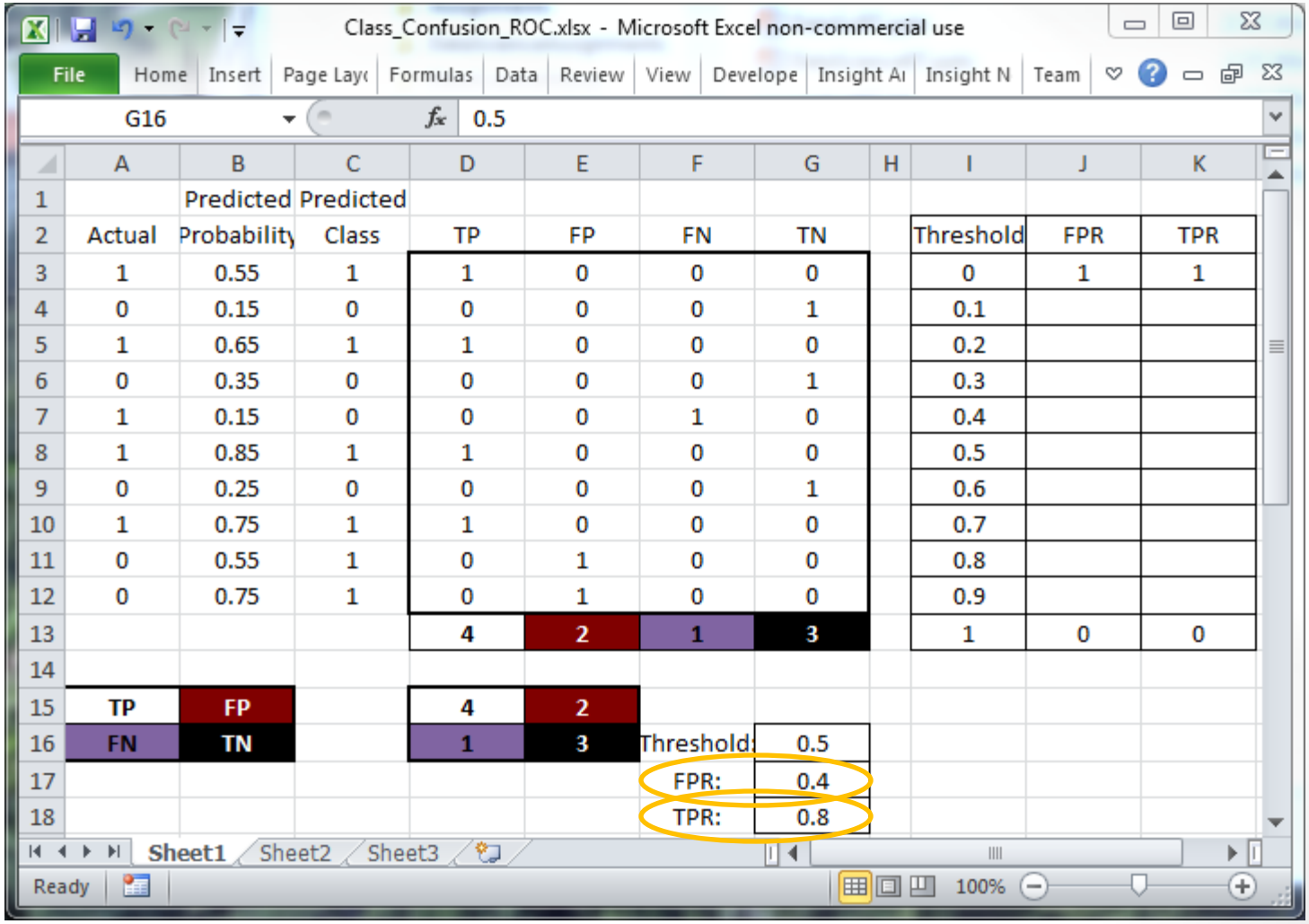
How to make an ROC (11)



How to make an ROC (12)

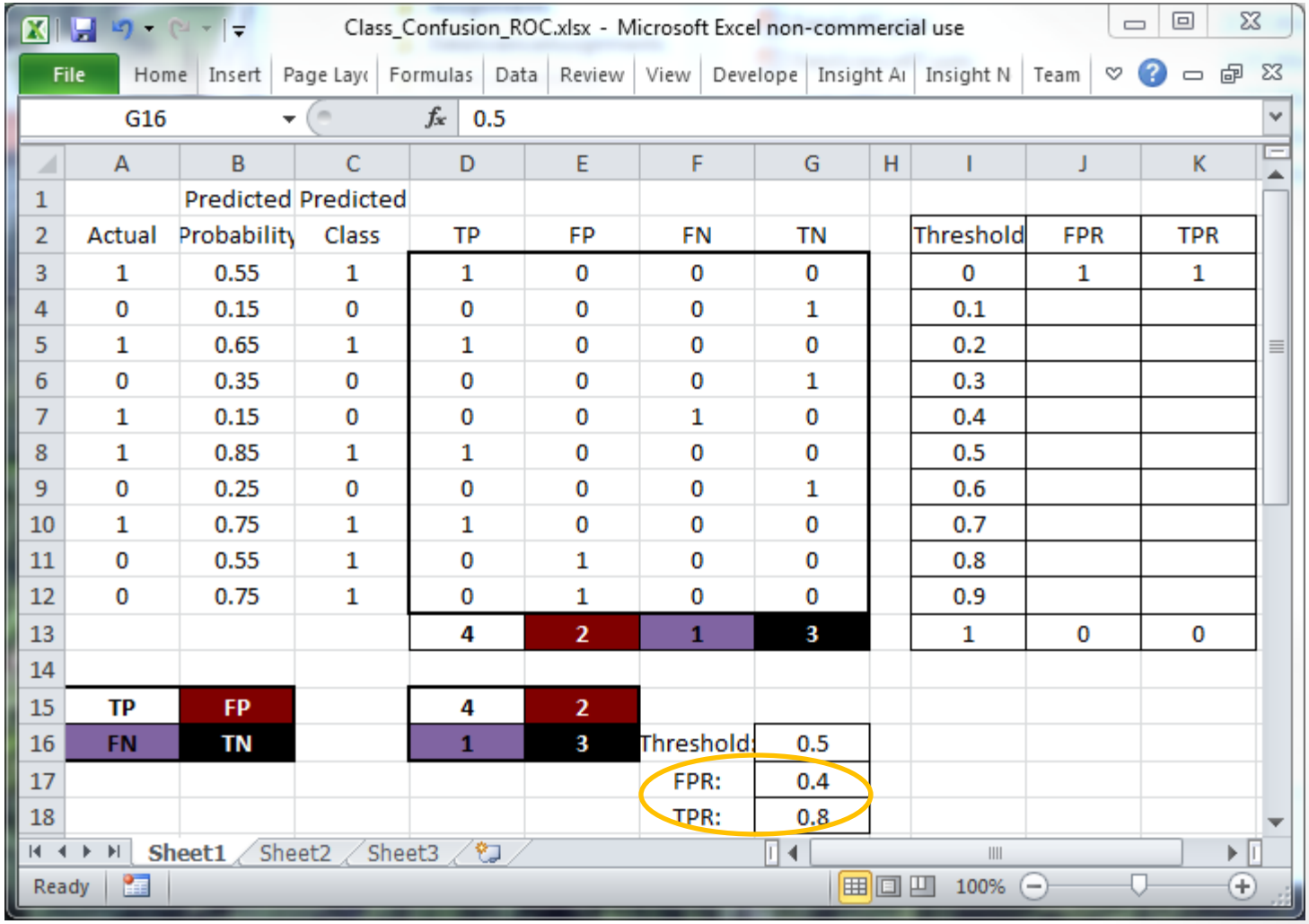


How to make an ROC (13)



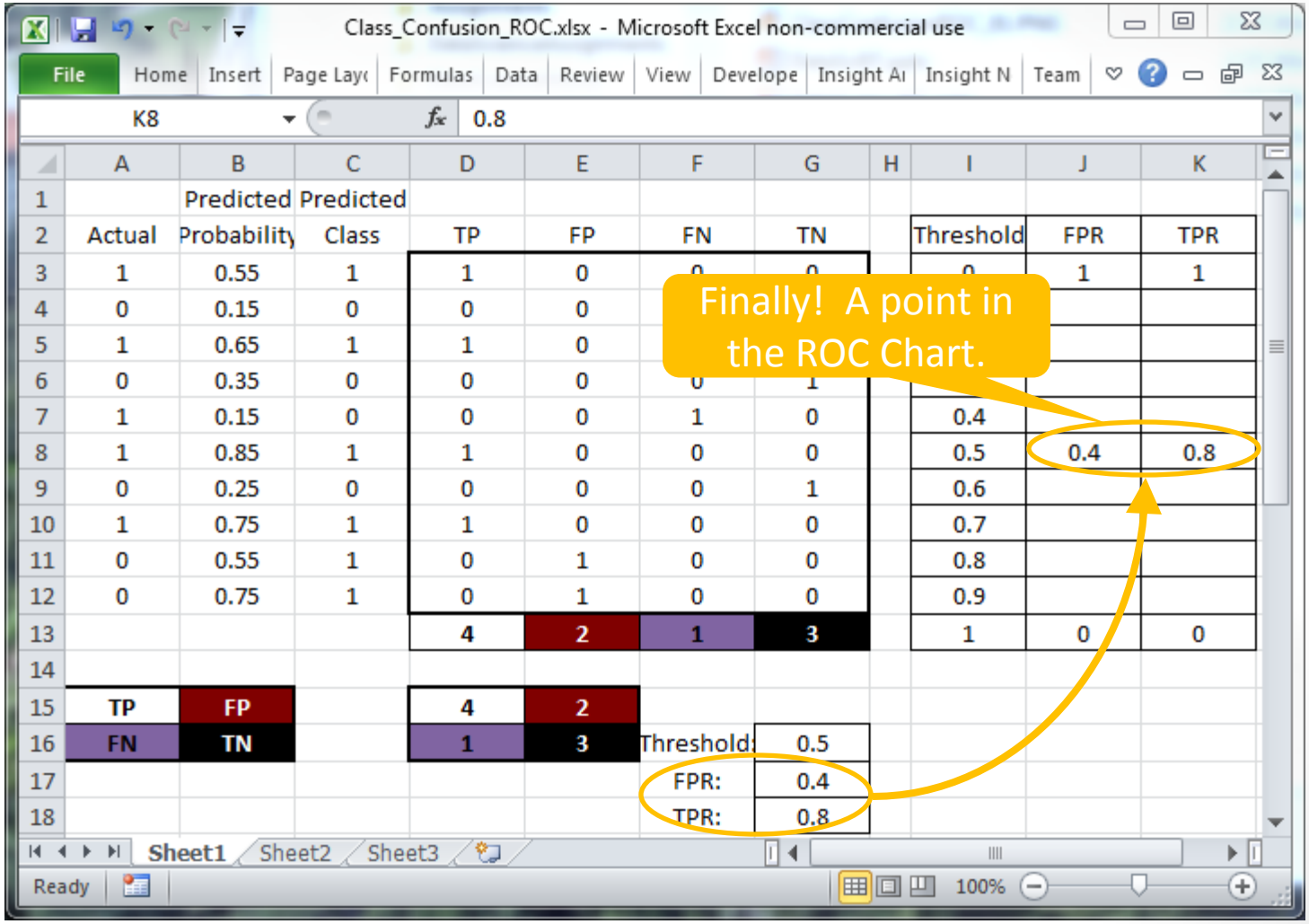
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

How to make an ROC (14)



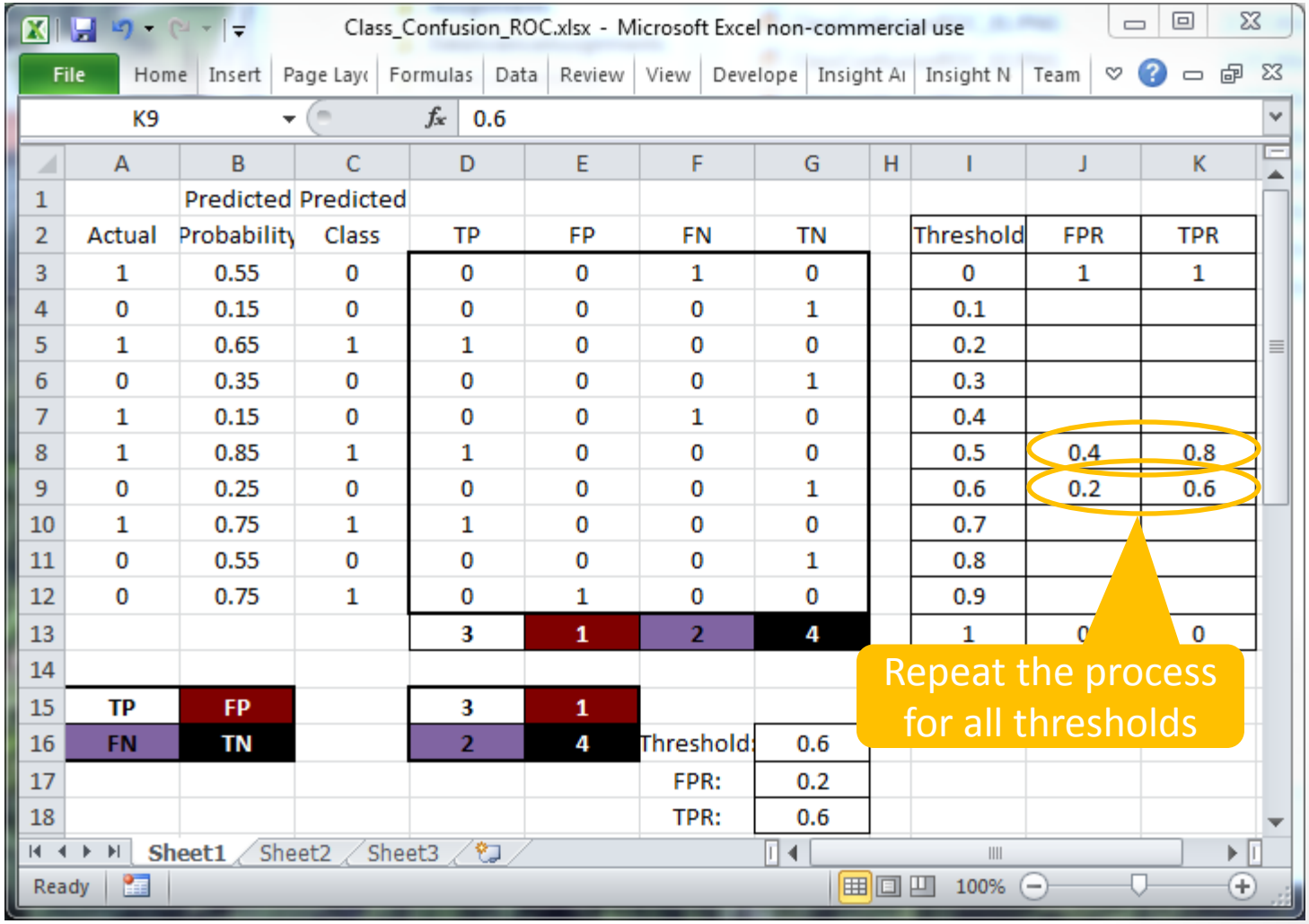
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75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

How to make an ROC (15)



	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				

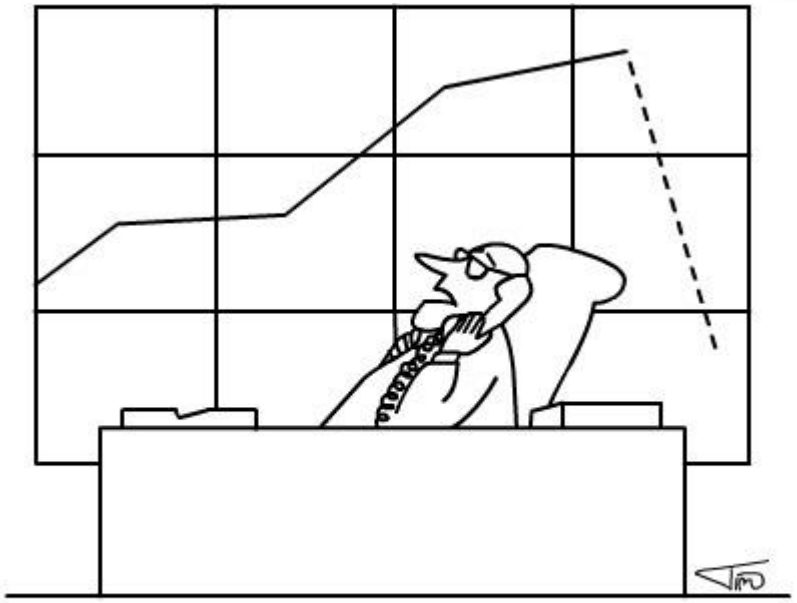
How to make an ROC (16)



	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				

Video and Break

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

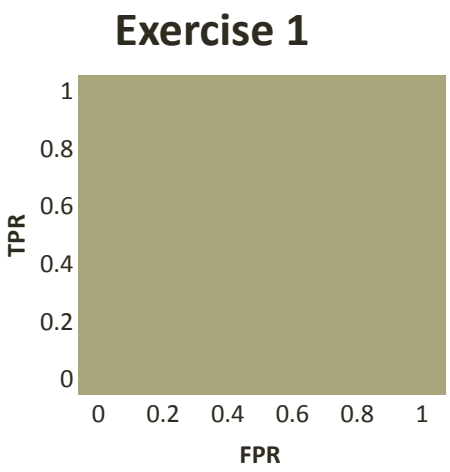


"BI tech support? The predictive analysis system is giving the wrong answer again—can you please fix it?... "

How to make an ROC (17)

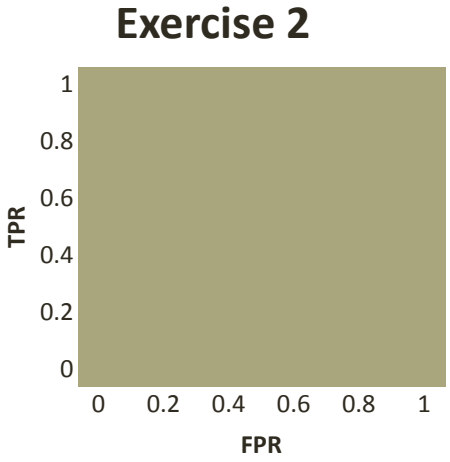
Actual	Predicted Probability
1	0.55
0	0.15
1	0.65
0	0.35
1	0.15
1	0.85
0	0.25
1	0.75
0	0.55
0	0.75

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

Data Structures

Data Structures (1)

Terminology and Concepts

- Data
  - Dataset is a set of Data. A set implies a commonality. The commonality is expressed as a type or a relation.
  - A data type provides structure and meaning to the data. Just like there is no such thing as un-structured data, there is no such thing as un-typed data. Data can be insufficiently typed and structured.
- Rectangular Data
  - Datasets are often 2D matrices, which are organized into rows and columns. The column and row order is not important .
  - Columns are named with a header; A columns may be also referred to as an attribute or field. The number of columns is often called the dimensionality of the data.
  - Rows are not named. A row is often referred to as a case or observation. Number of rows in a category is called support.
- Data dimensionality
  - A data frame or a table can be considered a sparse multi-dimensional matrix
  - The dimensionality for un-supervised learning is #columns
  - The dimensionality for supervised learning is #columns - 1 because one column represents the value and not the dimension. This structure is very similar to a star schema

Data Structures (2)

Terminology and Concepts

- Predictive Analytics (Machine Learning , Artificial Intelligence)
  - Algorithms (often called Methods)
    - Supervised Learning
      - Classification
      - Estimation
    - Unsupervised Learning
      - Clustering
      - Association (Market-basket analysis)
      - Anomaly detection
    - Time Series
      - Forecasting (Arima)
      - Regression with time lags
      - Survival analysis

Data Structures (3)

Terminology and Concepts

- Supervised Learning Algorithms
  - Classification Algorithms predict classes or categories
    - Logistic Regression (Deterministic)
    - Decision Trees (Deterministic)
    - Naïve Bayes (Deterministic)
    - Neural Net (Non-Deterministic)
    - Random Forest (Non-Deterministic)
  - Estimation Algorithms predict continuous (numeric) values
    - Generalized Linear Modeling abbreviated: GLM (Deterministic)
      - Linear Regression
      - Logistic Regression
    - Regression Trees (Deterministic)
    - Neural Net (Non-Deterministic)

Data Structures (4)

Terminology and Concepts

- Un-Supervised Learning Algorithms
  - Segmentation Algorithms, also called Clustering, create clusters or segments. These clusters can be thought of as categories.
    - Mixture of Gaussians aka Probabilistic (Deterministic)
    - Hierarchical (Deterministic)
    - K-Means (Non-Deterministic)
  - Association Algorithms associate or link items by a common attribute called the transaction ID.
    - Market Basket Analysis (Deterministic)
    - Affinity Analysis (Deterministic)
  - Anomaly Detection is used to find unusual or anomalous data like outliers

Data Structures (5)

Terminology and Concepts

- Forecasting (Time Series) is used to estimate future values based on past behaviors.
  - ARIMA / Auto ARIMA
  - Survival Analysis

Data Structures (6)

Major types of Data Sets

- Univariate
- Rectangular
- Time Series
- Nested
- Graphs (later in the course)

Data Structures (7)

Univariate

- A collection of data. The data do not have a particular order. Example: Students' age. This type of data is often (mistakenly) called unstructured data, especially when the values are strings of indeterminate length. (Ragged Array)
- Example usage: anomaly detection.

Data Structures (8)

Univariate

Parent Income
40,000
53,000
60,000

Data Structures (9)

Rectangular Data

- The data set has columns and rows. Each cell has a value or is null.
- A Rectangular dataset is often called a matrix, data frame, or table.
- Example usage: classifications and estimations

Data Structures (10)

Rectangular Data

- Columns have descriptive headers like: Name, Age, Height, Weight of each student.
- Columns are also called attributes and fields.
- All values within a column have the same data type

Data Structures (11)

Rectangular Data

- Rows generally do not have names. If a row has a name, then the names could be considered another column.
- Rows are also called observations or cases
- The number of rows in a category is called support.

Data Structures (12)

Rectangular Data

ID	IQ	Parent Income	Moral Support	Gender	College Plans
835	107	40,000	Yes	Female	Applied
016	99	53,000	Yes	Male	Applied
490	105	60,000	No	Male	Did not apply

Data Structures (13)

Time Series

- A rectangular data set where the independent variable is time. The observations are sorted by time.
- Example usage: forecasting.

Data Structures (14)

Time Series

Date	Red Wine Sales	White Wine Sales	Rose Sales
1/22/13	$103.00	$300.50	$19.00
1/23/13	$35.50	$204.00	$44.00
1/24/13	$217.50	$74.50	$80.00

Data Structures (15)

Nested

- A rectangular data set where a cell contains a table. The nested structure can have a flat representation that is not nested.
- Example usage: associations (shopping basket analyses).

Data Structures (16)

Nested

Transaction ID	Item
1	Milk
	Sugar
2	Lumber
3	Milk
	Sugar
	Flour

Data Structures (17)

Transaction ID	Item
1	Milk
1	Sugar
2	Lumber
3	Milk
3	Sugar
3	Flour

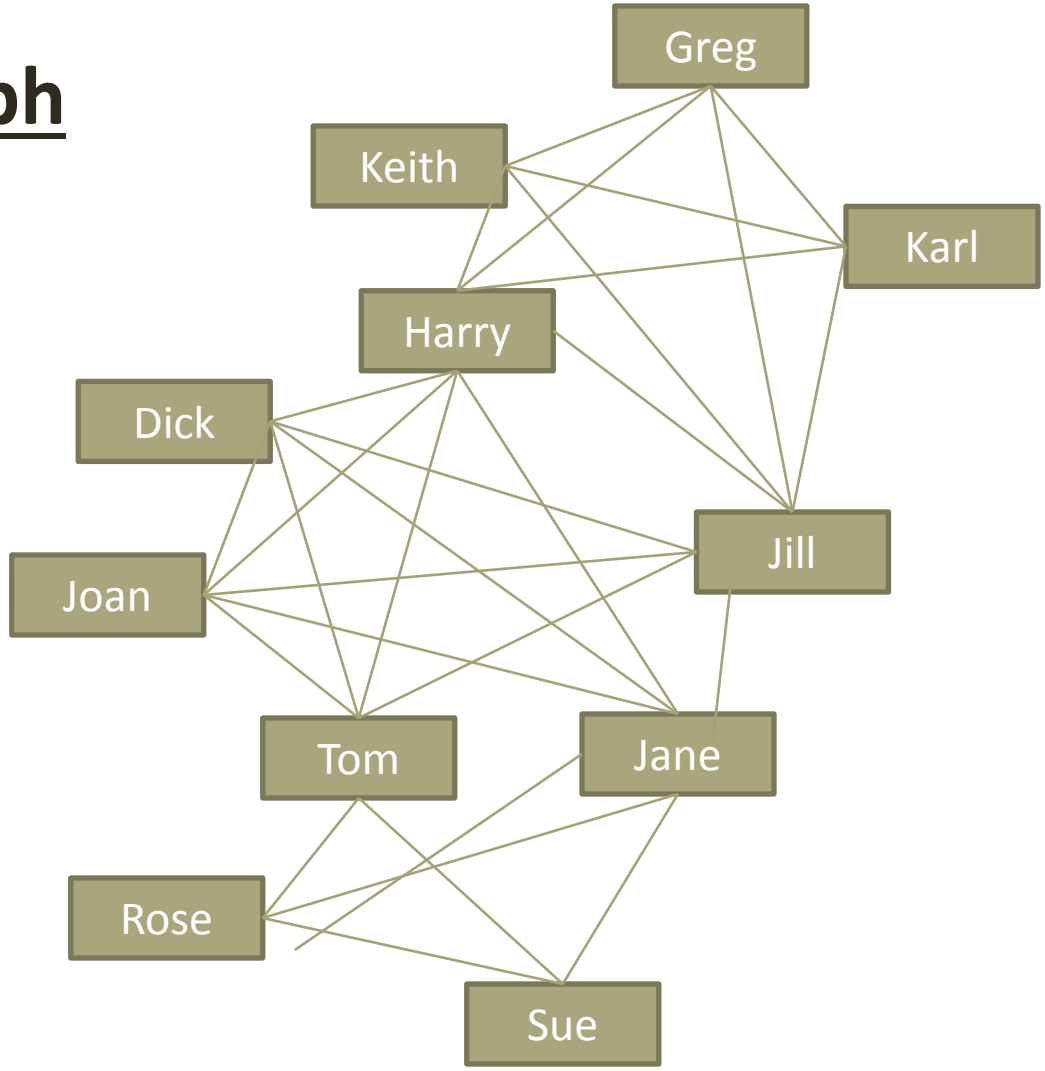
Data Structures (18)

Nested

Transaction ID	Item
1	Milk
1	Sugar
2	Lumber
3	Milk
3	Sugar
3	Flour

Data Structures (19)

Graph



Data Structures

Review: Terminology

- Algorithm
- Anomaly detection
- Association
- Attribute
- Binarize Categories
- Binary Column
- Case
- Category Column
- Character Column
- Classification
- Clustering
- Coercion
- Column
- Column Header
- Data
- Data Dimensionality
- Data Frame
- Data Type
- DFD
- Dummy Variable
- Estimation
- Feature Scaling
- Field
- Hypothesis
- Key Column
- Machine Learning
- Market-basket analysis
- MATLAB
- Matrix
- Missing Data
- Model
- Multinomial Column
- Normalization
- Numeric Column
- Observation
- Outcome
- Outlier Removal
- Predictive Analytics
- R
- Rectangular Data
- Relabeling
- Row
- Schema
- Shaping Data
- Sparse Multi-Dimensional Matrix
- Standard Deviation
- States
- String
- Supervised Learning
- Support
- Table
- Target Column
- Text Column
- Theory
- Un-structured Data
- Unsupervised Learning
- Z-score

Assignment (1)

1. Training vs Test Data
   a) In general, for any modeling data, why are accuracy measures 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) You have two datasets. You used one to train the model and the other to test the model. You lost the test results and forgot which one you used for training or testing. How can you determine which of these datasets is the testing 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. Present the confusion matrix and 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. Verify that your graph is montonic non-decreasing. 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/zmkrb58b3uqmic7/Cloudera-Training-VM-4.2.1.p-vmware_prist2.zip?dl=0
   - Install and setup the Hadoop VM according to SetupVirtualMachine.pdf".
   - Create Screenshot as directed on last slide of SetupVirtualMachine.pdf
7. Submit answers to items 1 through 4 in a text file with a "txt" suffix. 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 11:57 PM.
8. Complete Quizzes
   - Read Lecture_03.pdf Classification Schema and Lecture_04.pdf Data Structures
   - Quiz Predictive Analytics Overview
   - Quiz Schema and Attributes for Supervised Learning

Assignment (4)

9. On LinkedIn, start a discussion, make a comment on an existing discussion, or ask questions about homework.
10. Reading Assignments
    - 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
    - Review terminology at the end of this slide deck
    - Read Quiz Previews. They might not be posted until tomorrow.
    - Relational Model, Relational Algebra, and Relational Calculus
        - http://en.wikipedia.org/wiki/Relational_algebra
        - http://sentences.com/docs/amd.pdf (Pages 35 to 48 only)
        - http://en.wikipedia.org/wiki/Relational_model
        - http://www.youtube.com/watch?v=NvrpuBAMddw

Introduction to Data Science