

# Introduction to Data Science

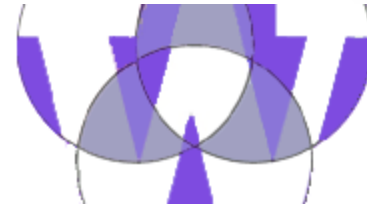
Lecture 09; May 21<sup>st</sup>, 2018

Ernst Henle

[ErnstHe@UW.edu](mailto:ErnstHe@UW.edu)

Skype: ernst-henle

# Agenda



- 18:00-18:05 Announcements
  - 4 Quizzes: 2 in class and 2 for homework
  - LinkedIn: start a discussion or comment on an existing discussion
  - Applebee's: 9:15 today; 13856 Bel Red Road ; (140<sup>th</sup> Ave)
- 18:05-18:25 Overfitting and Confusion Matrix
- 18:25-18:40 Lesson 09 Quiz 0a Confusion Matrix
- 18:40-19:05 ROC Chart Intro
- 18:55-19:05 Break
- 19:05-19:15 Lesson 09 Quiz 0b ROC Intro
- 19:25-19:45 How to make an ROC
- 19:45-19:55 Break
- 19:55-20:30 Lesson\_09
- 20:30-21:00 L09-AccuracyMeasures.py
- Predictive Analytics Iteration Trap (Time Permitting)

# Over-fitting and Confusion Matrix

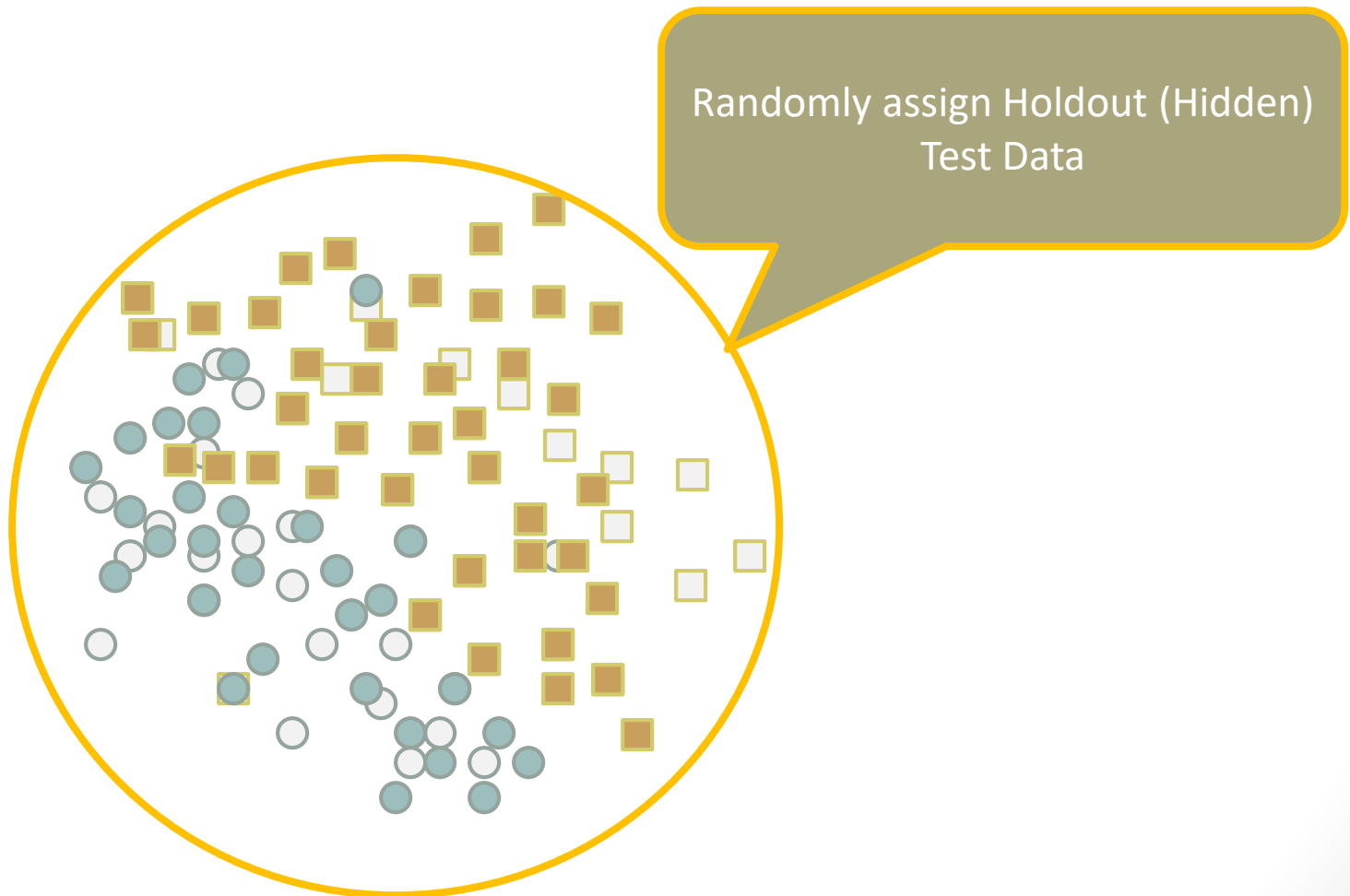
# Evaluate Model

- 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

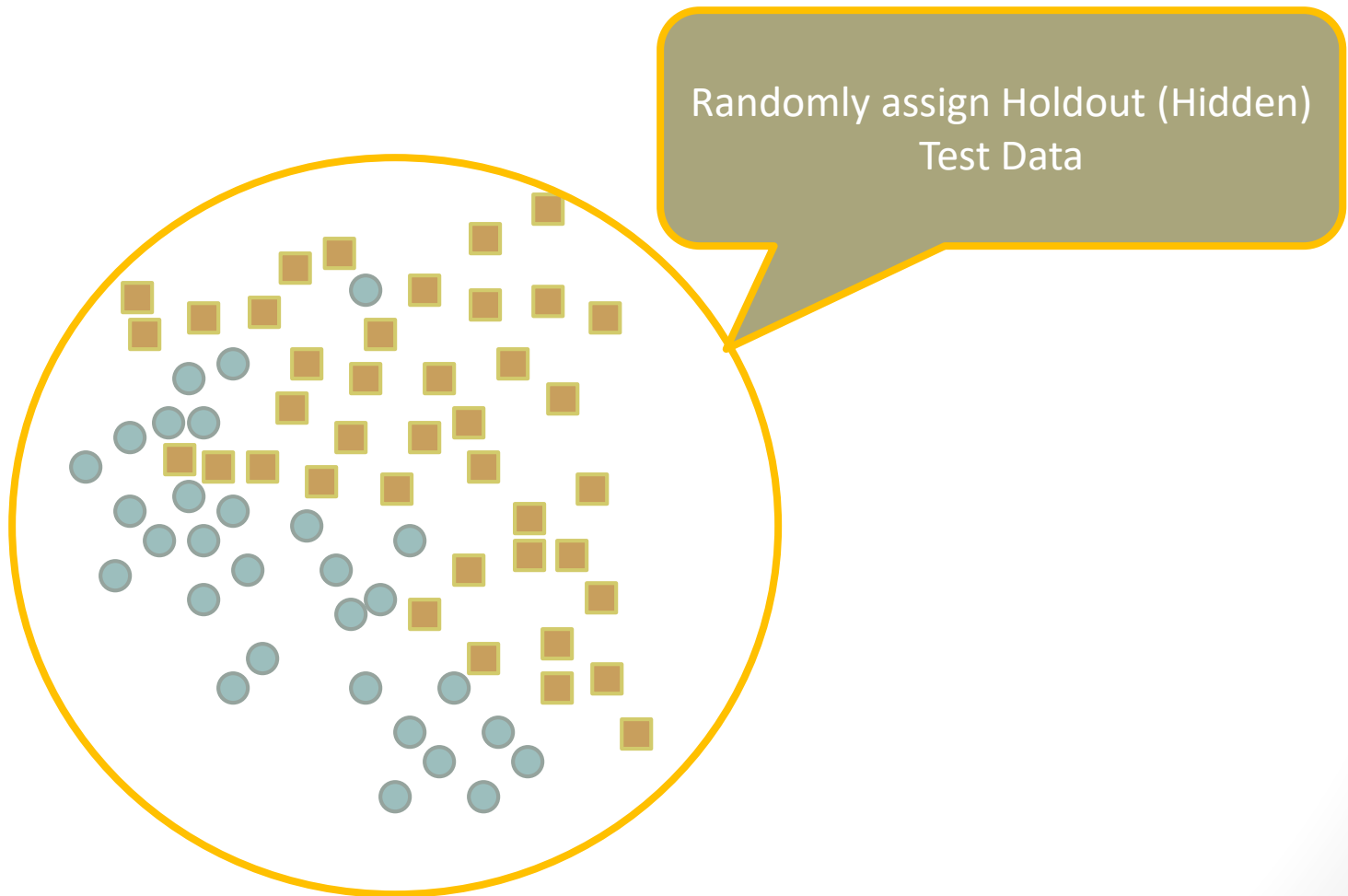
# Evaluate Model: All Data



# Evaluate Model: Test Data



# Evaluate Model: Test Data

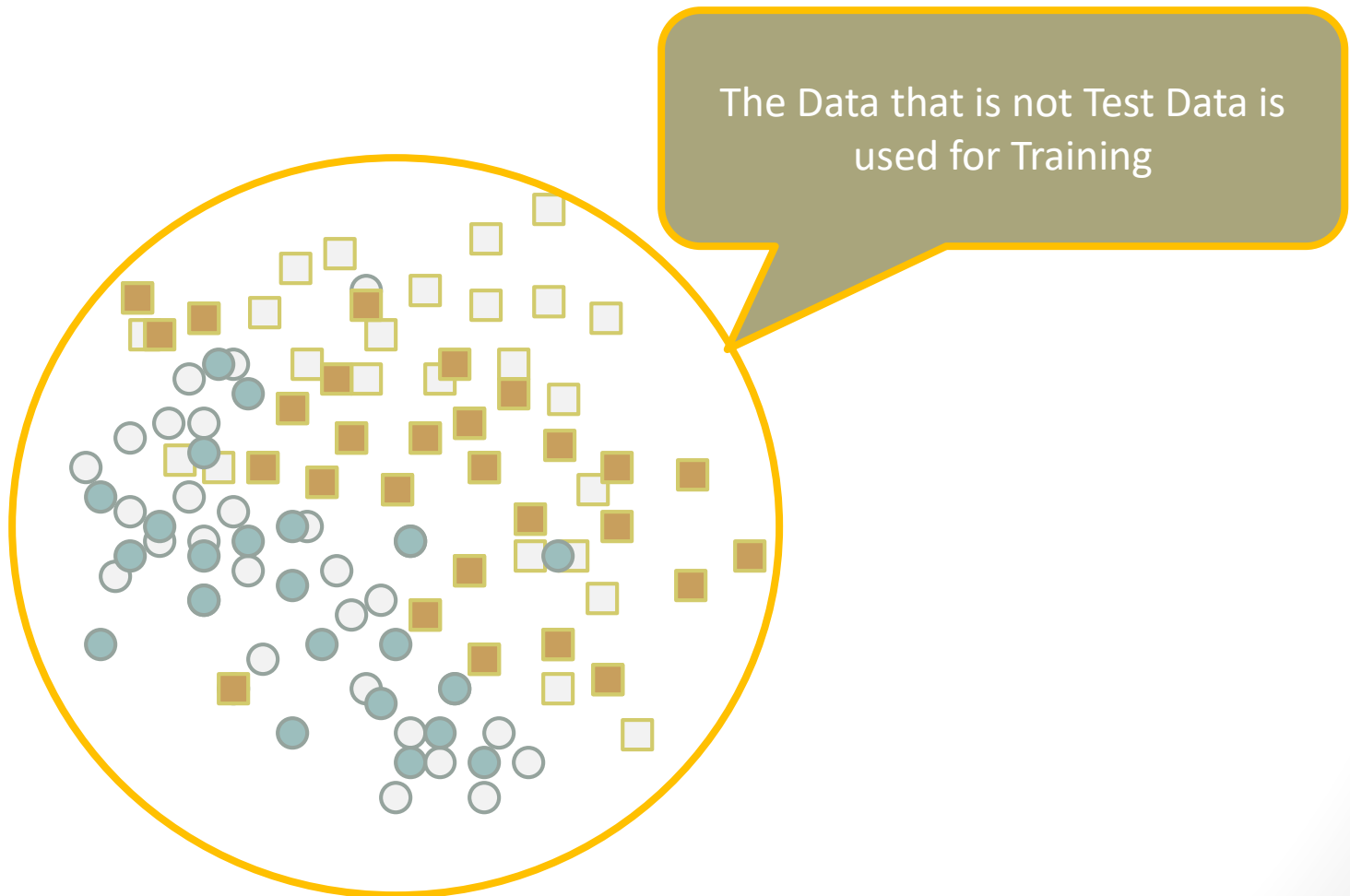


# Evaluate Model: All Data

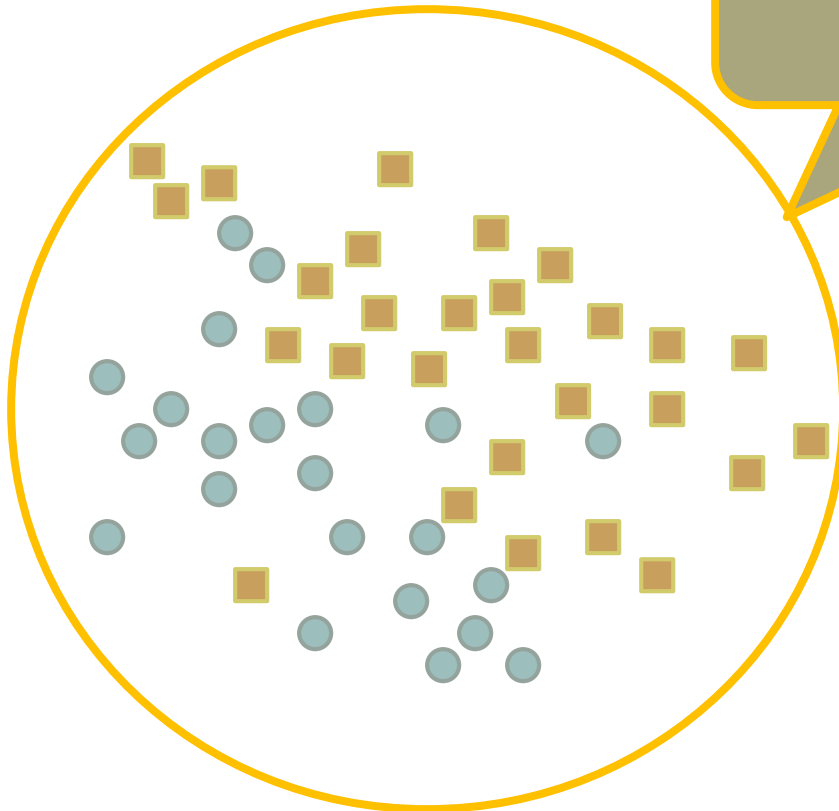




# Evaluate Model: Training Data

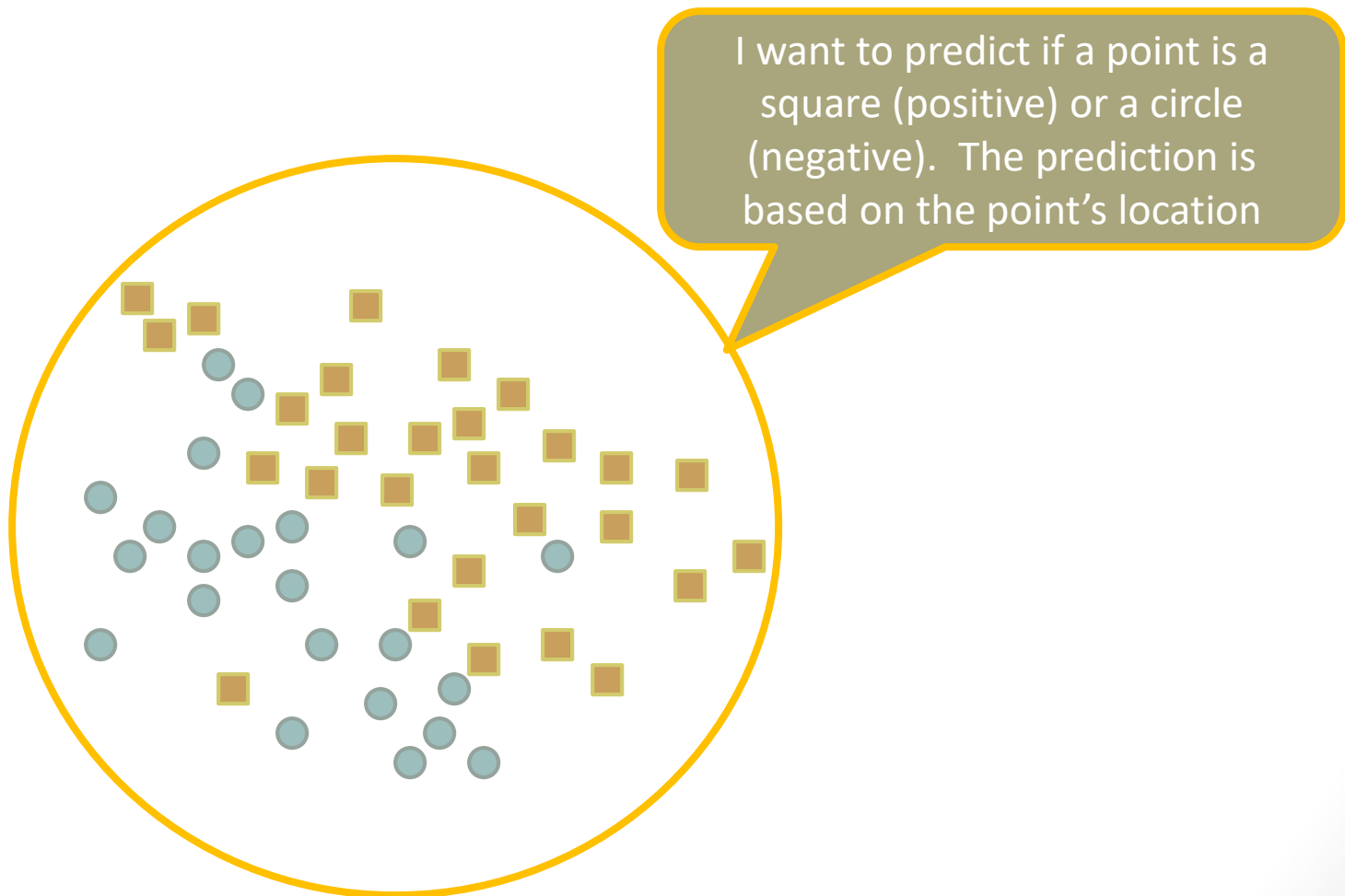


# Evaluate Model: Training Data

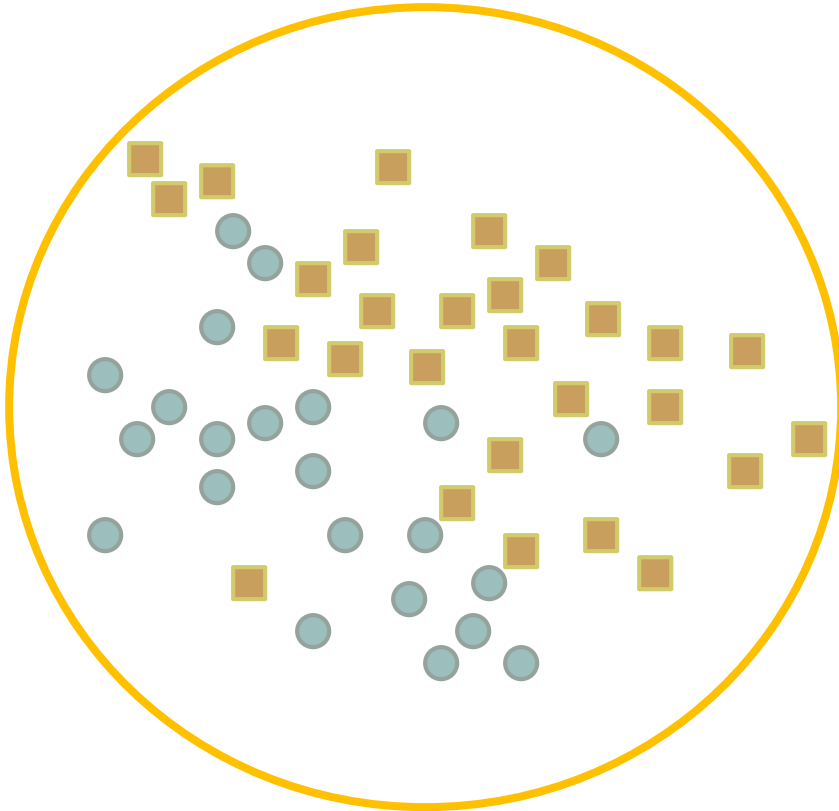


The Data that is not Test Data is  
used for Training

# Evaluate Model: Training



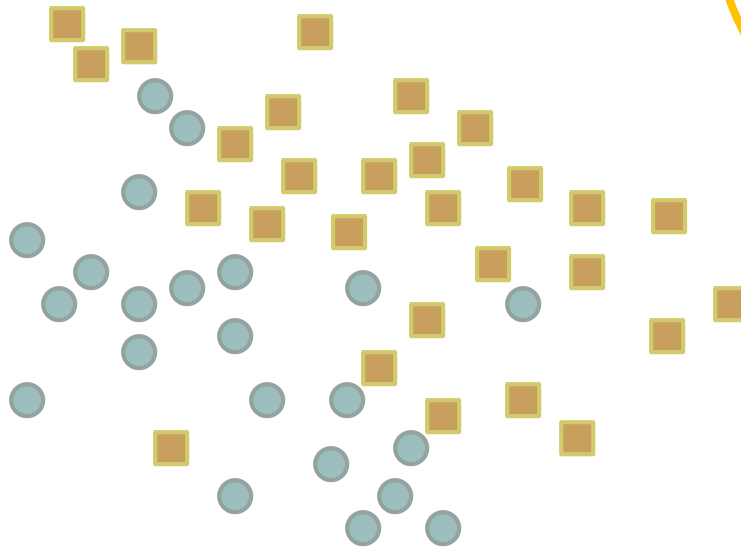
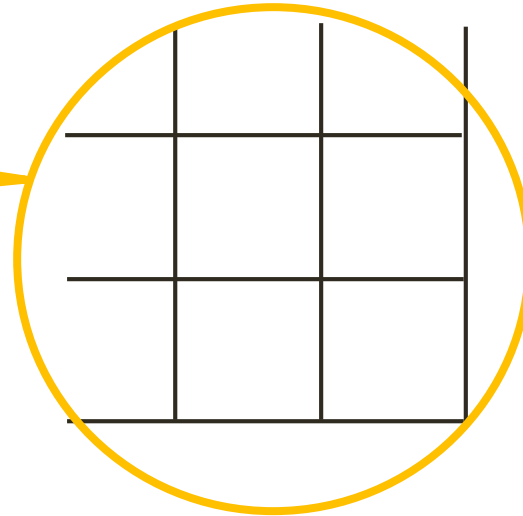
# Evaluate Model: Training



$\text{isSquare} \sim x\text{Location} + y\text{Location}$

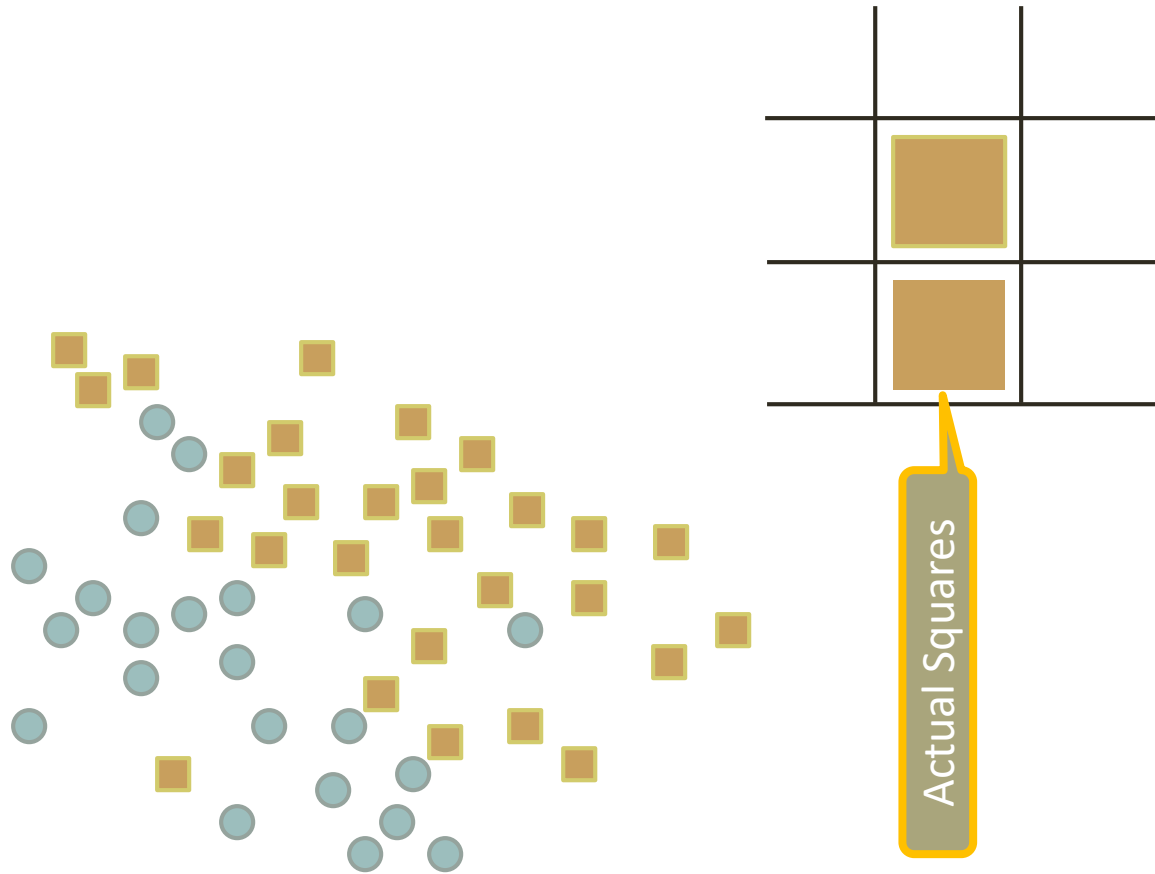
# Evaluate Model: Confusion Matrix

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



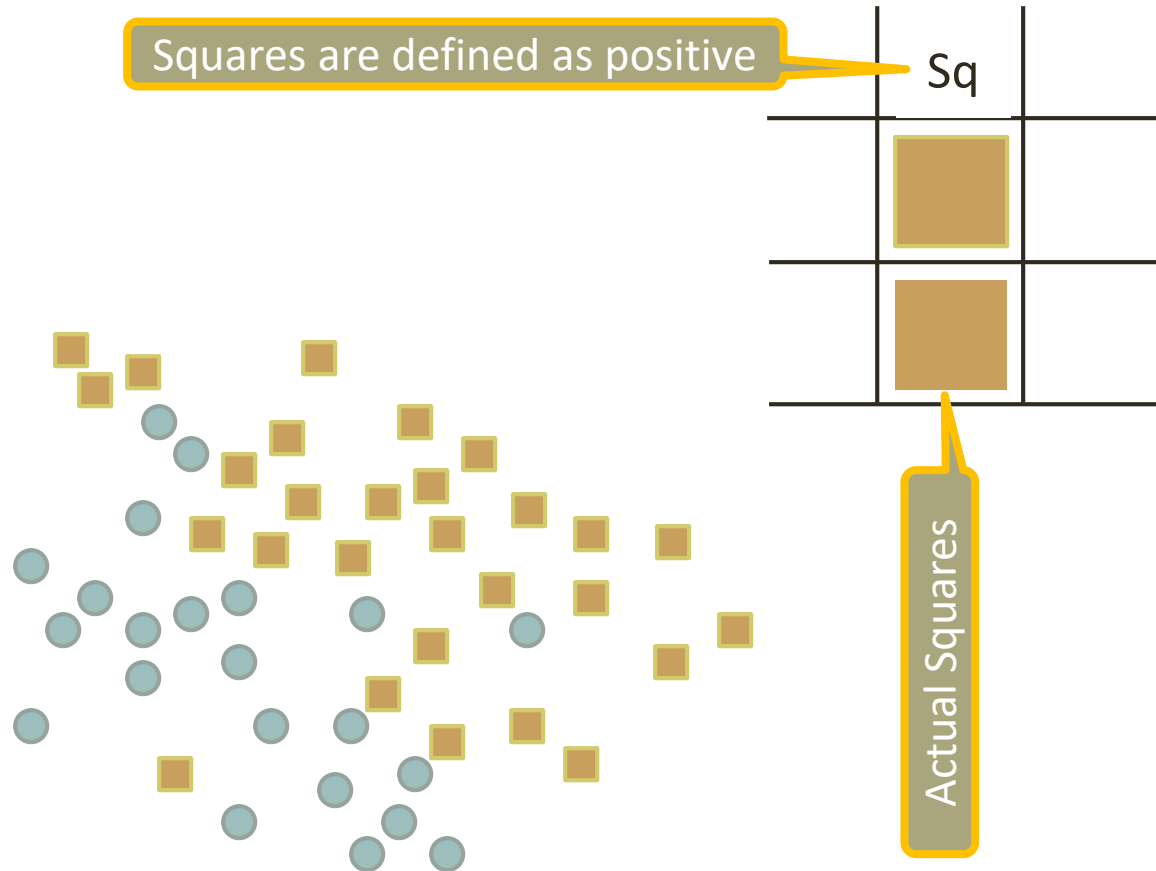
$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

# Evaluate Model: Confusion Matrix



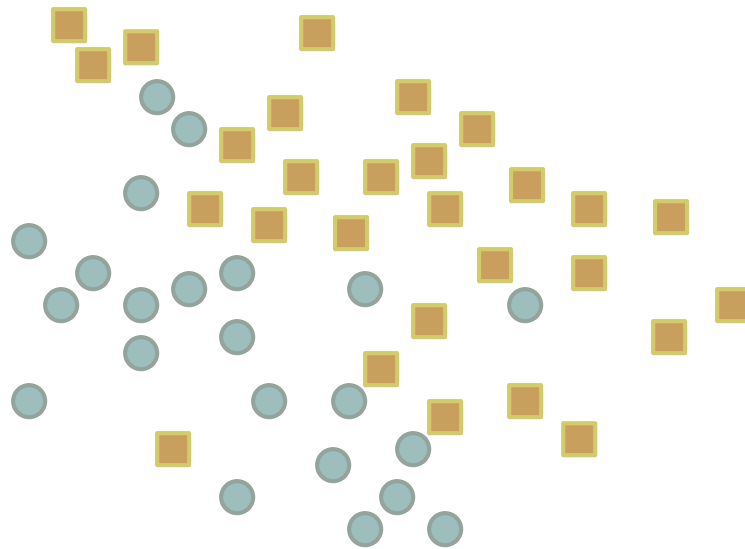
$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$





# Evaluate Model: Confusion Matrix



$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

# Evaluate Model: Confusion Matrix



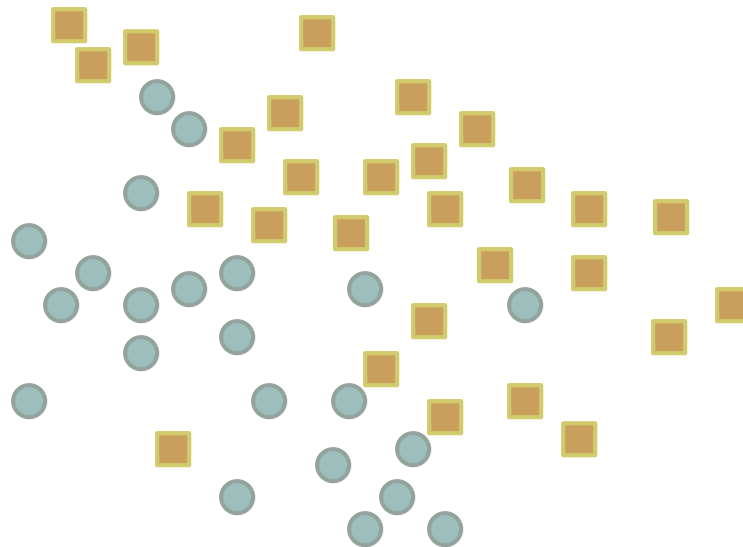
	Sq	
		
		

Actual Circles





$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$



# Evaluate Model: Confusion Matrix



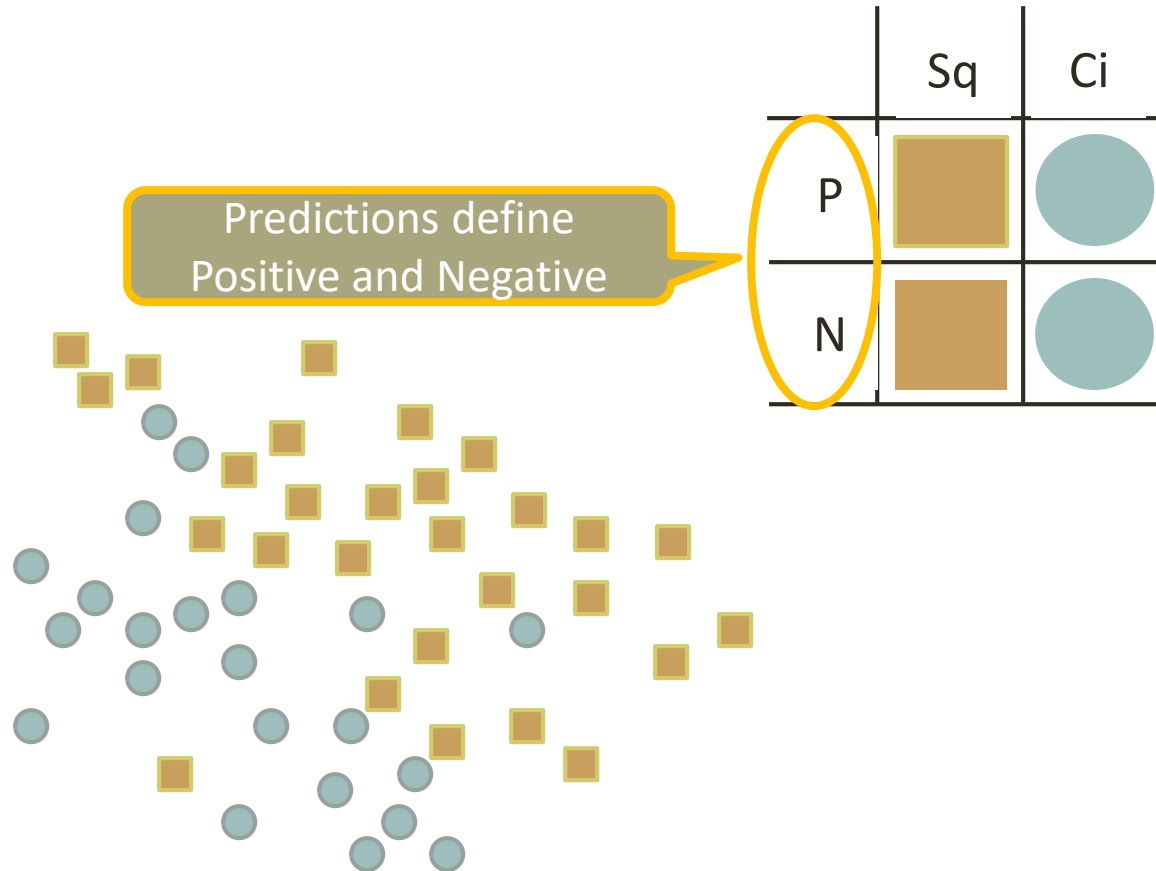
Circles are defined as negative

	Sq	Ci
		
		

Actual Circles

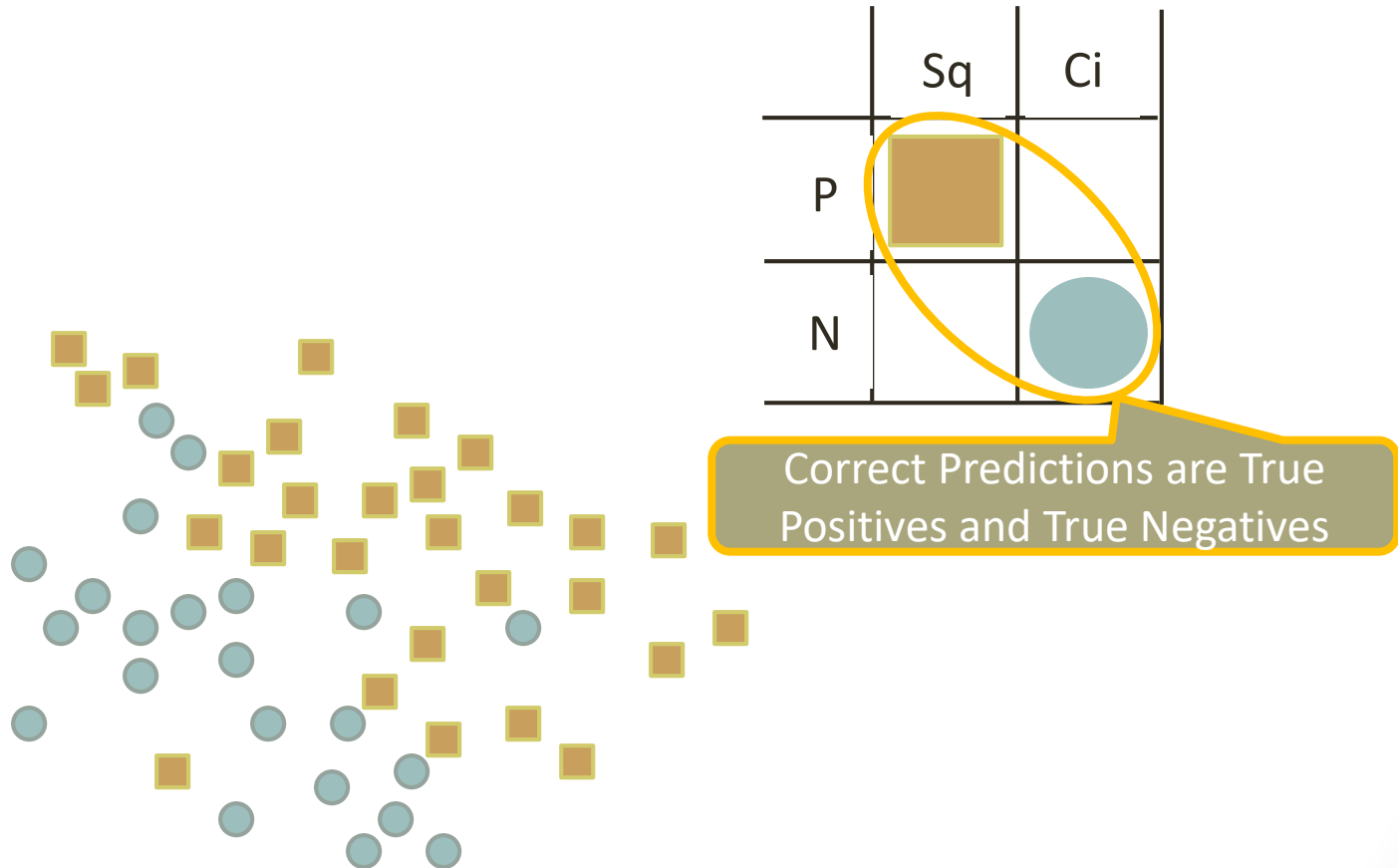
$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

# Evaluate Model: Confusion Matrix



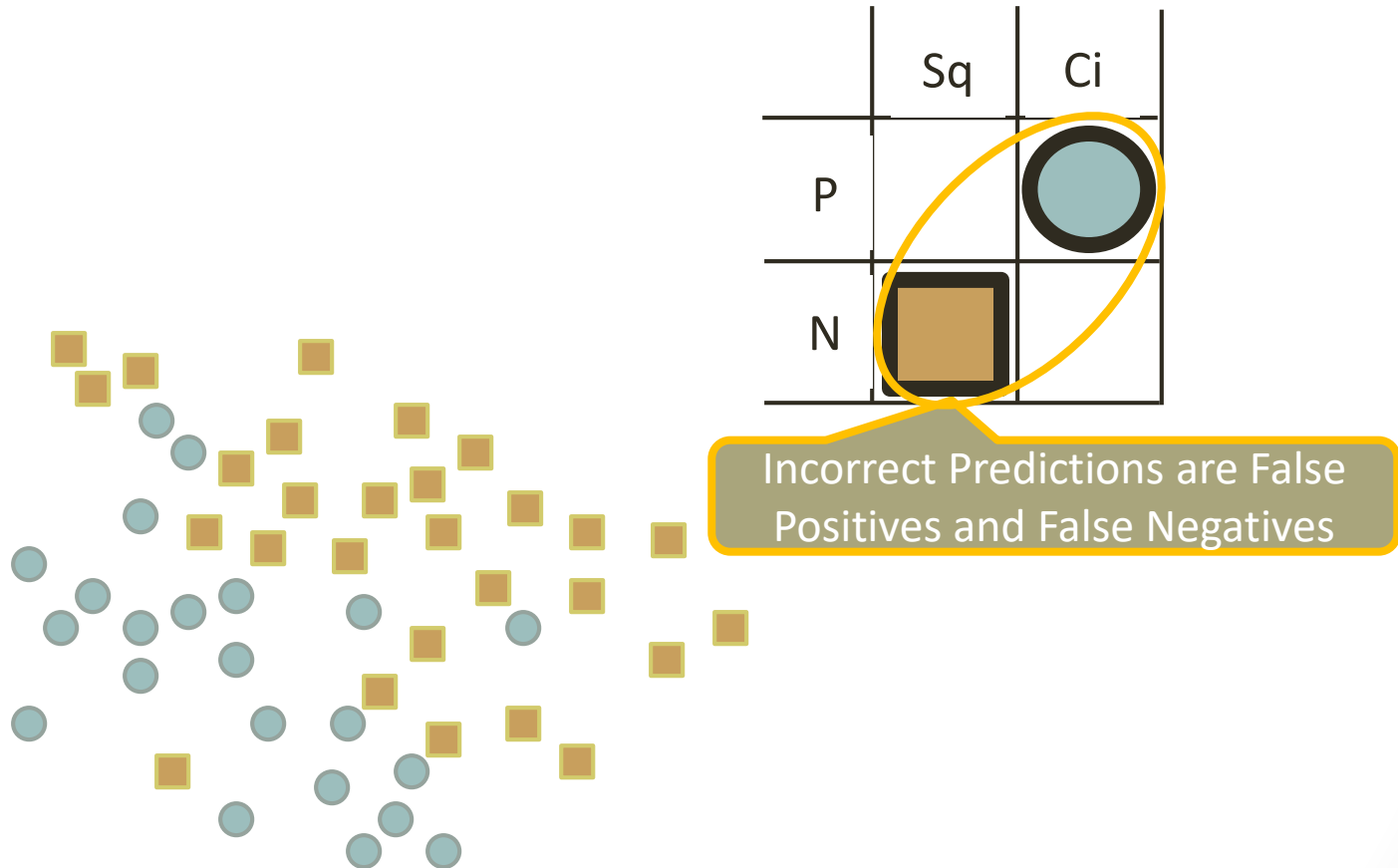
$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

# Evaluate Model: Confusion Matrix



$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

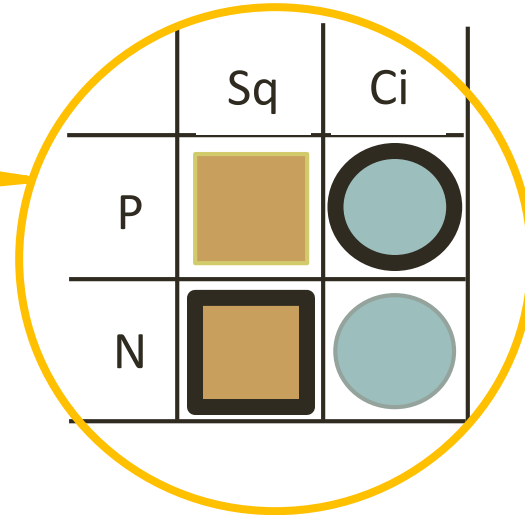
# Evaluate Model: Confusion Matrix




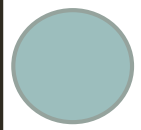


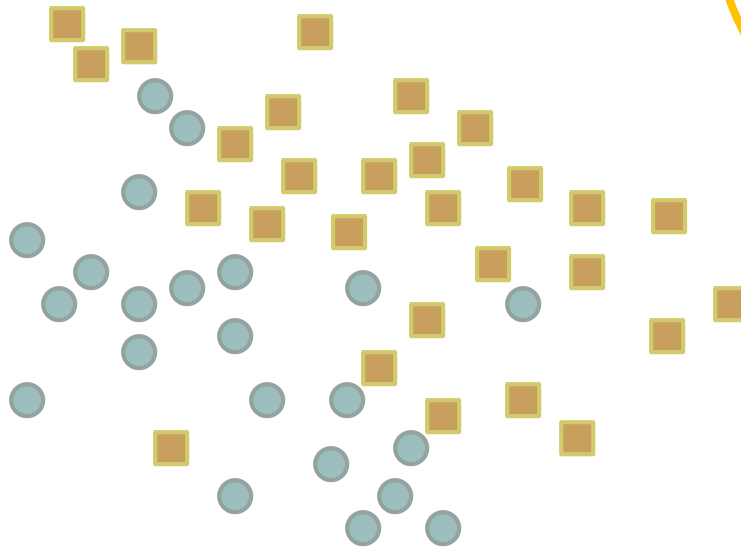
$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

# Evaluate Model: Confusion Matrix

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

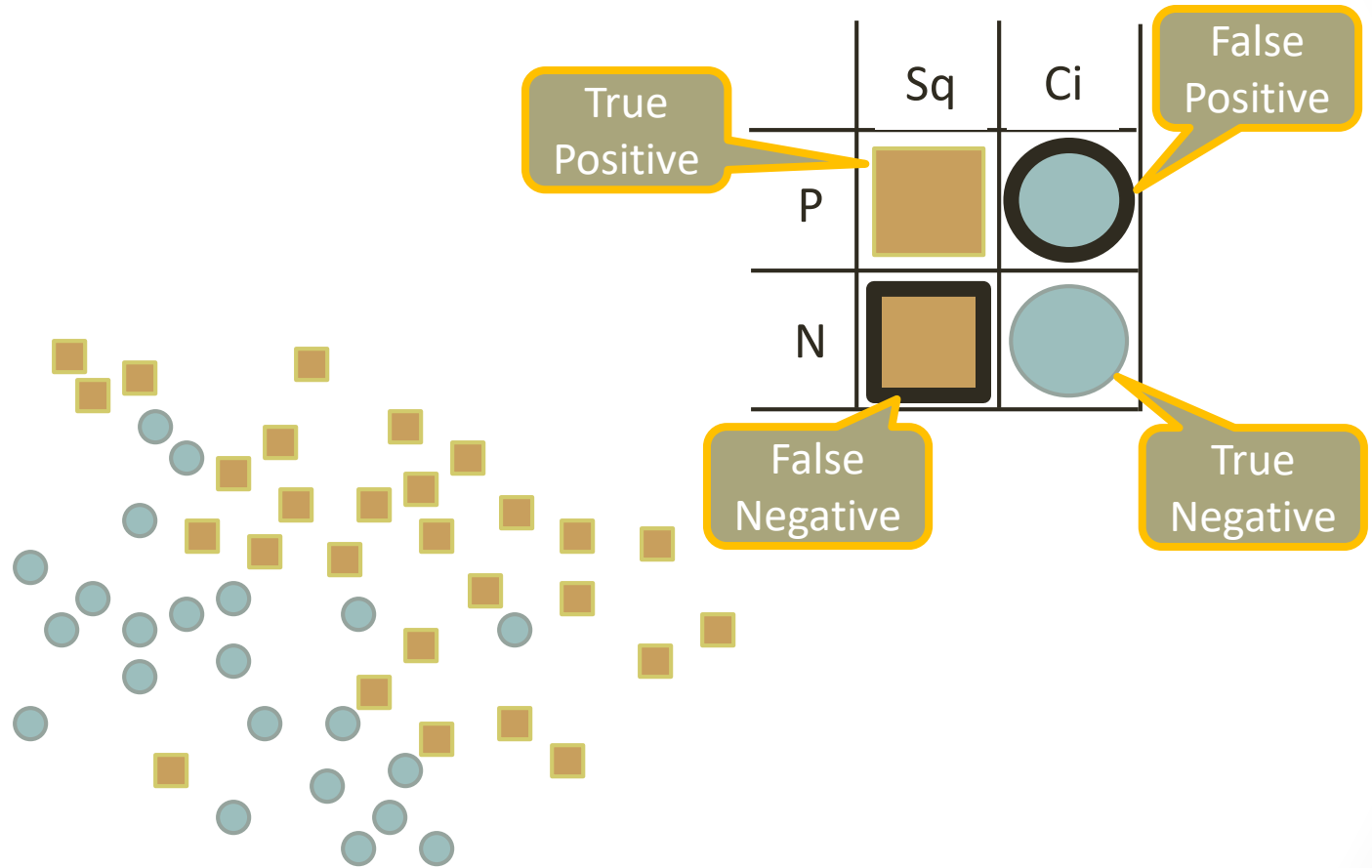


	Sq	Ci
P		
N		



$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

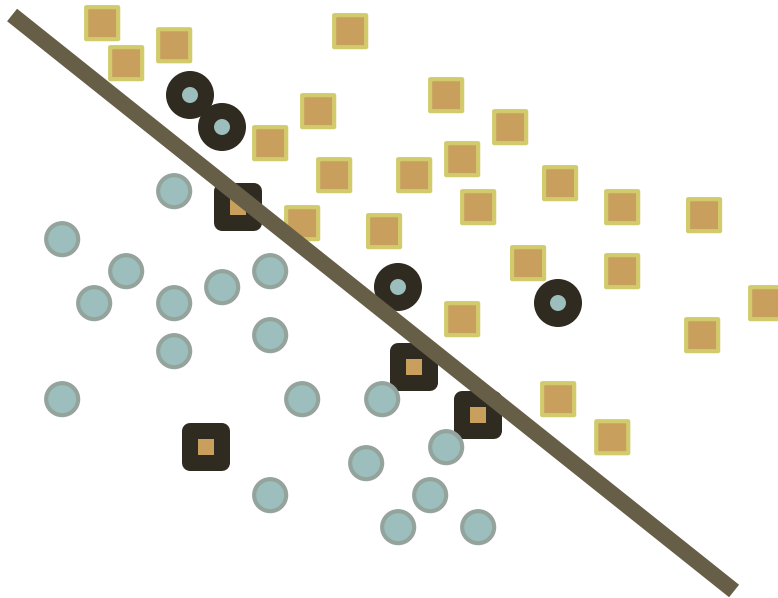
# Evaluate Model: Confusion Matrix



$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

# Evaluate Model : Train Model 1

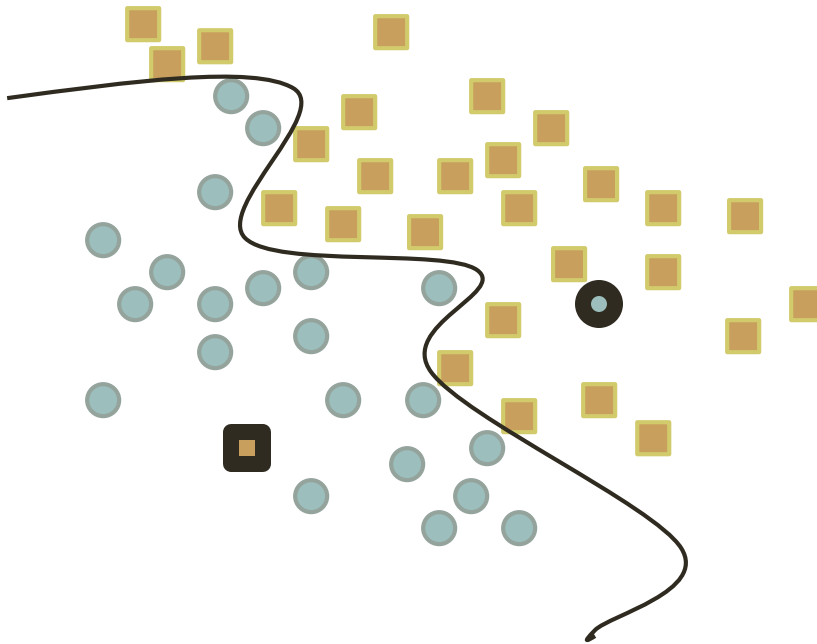
	Sq	Ci
P	36	4
N	4	26



$\text{isSquare} \sim x\text{Location} + y\text{Location}$

# Evaluate Model : Train Model 2

	Sq	Ci
P	39	1
N	1	29



$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$



# Evaluate Model : Train Model 3

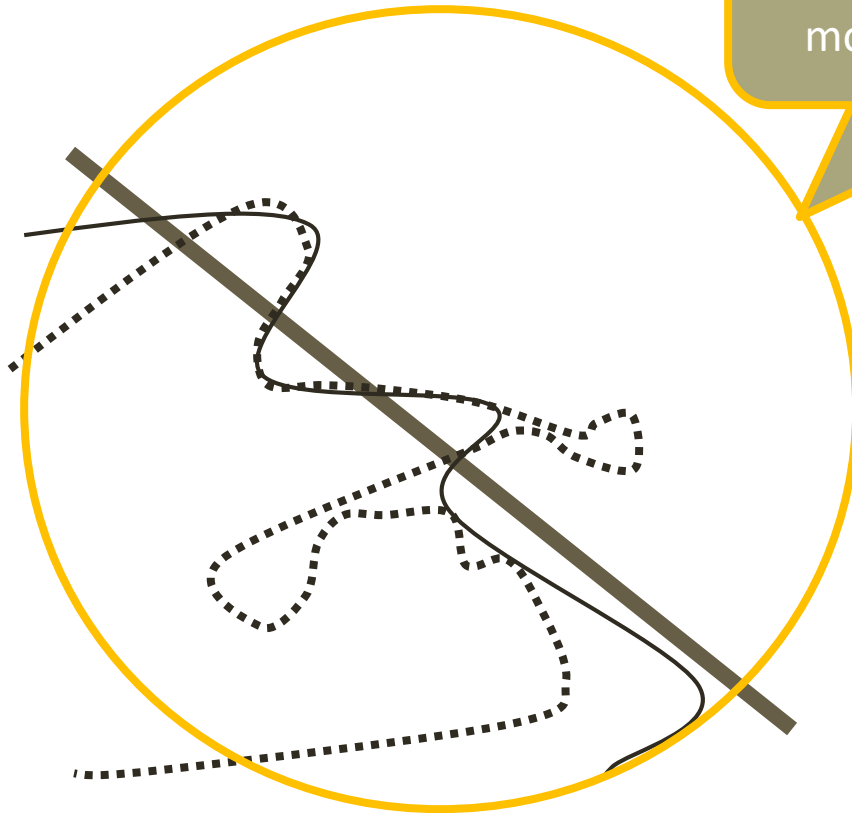
	Sq	Ci
P	40	0
N	0	30



$\text{isSquare} \sim x\text{Location} + y\text{Location}$

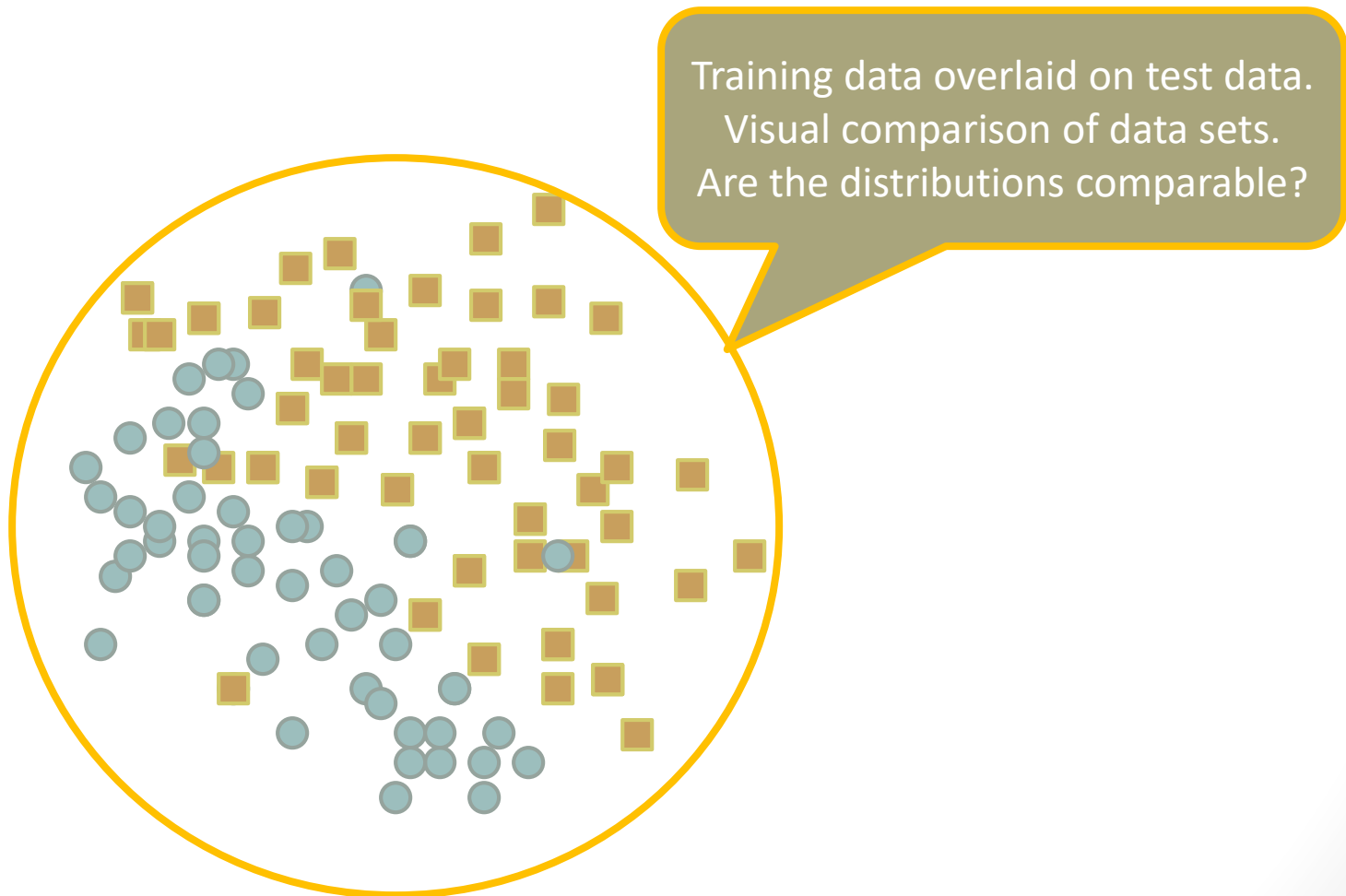
# Evaluate Model : 3 Models

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



$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

# Evaluate Model : All Data



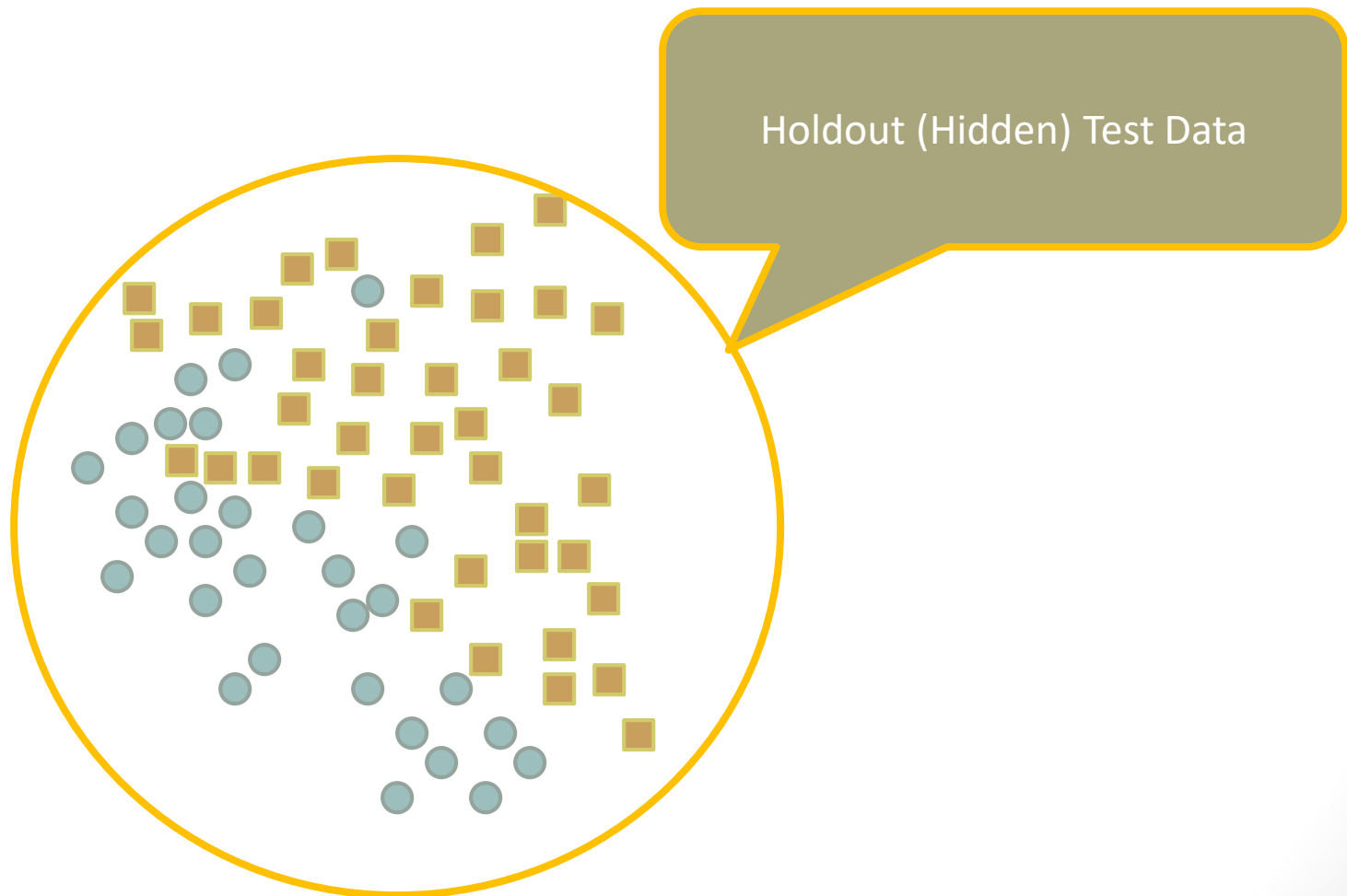
$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

# Evaluate Model : Training Data



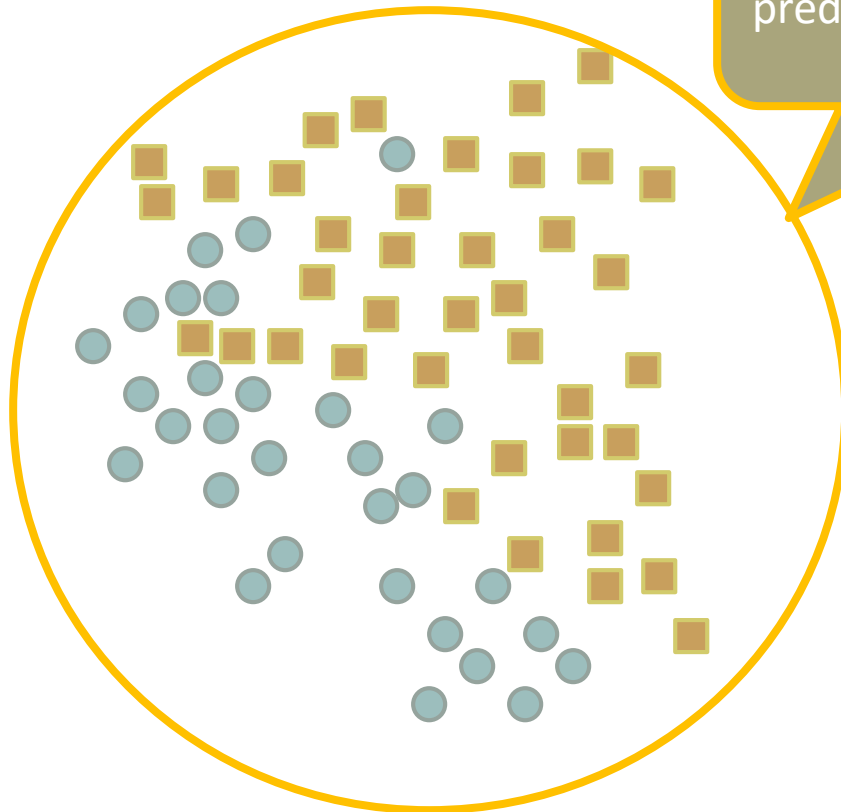
$$\text{isSquare} \sim x\text{Location} + y\text{Location}$$

# Evaluate Model : Test Data



$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

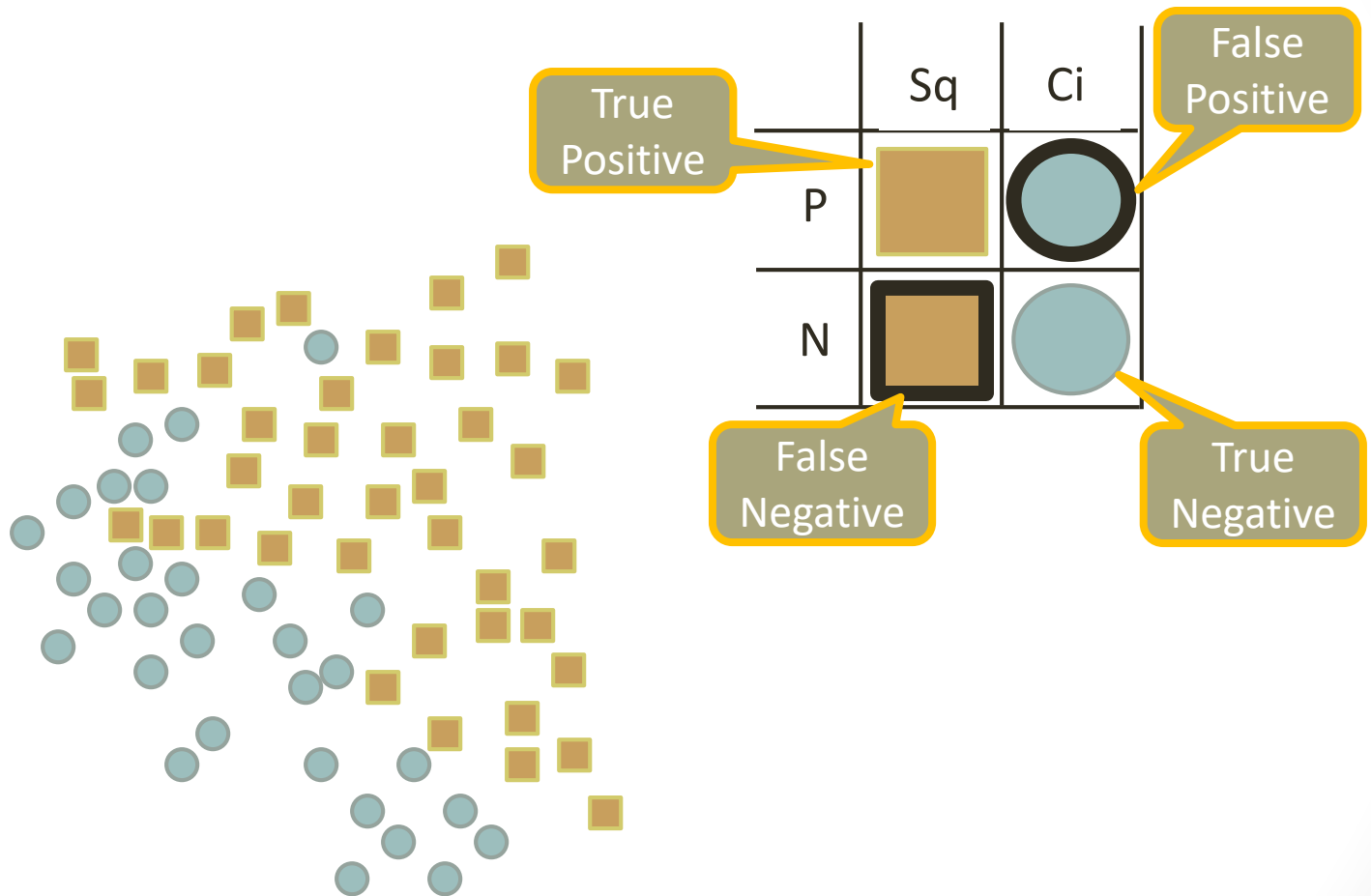
# Evaluate Model : Test Data



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

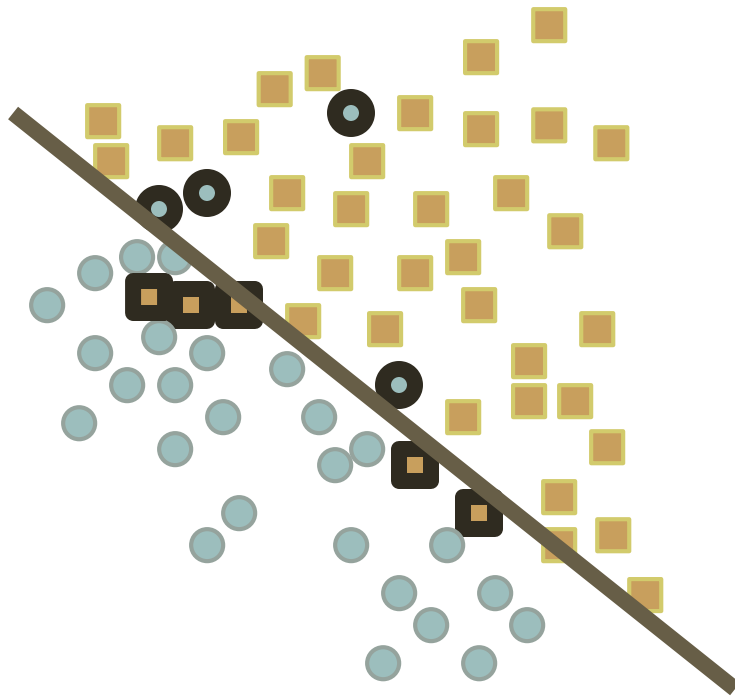
$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

# Evaluate Model : Test Data



$$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$$

# Evaluate Model : Test Model 1

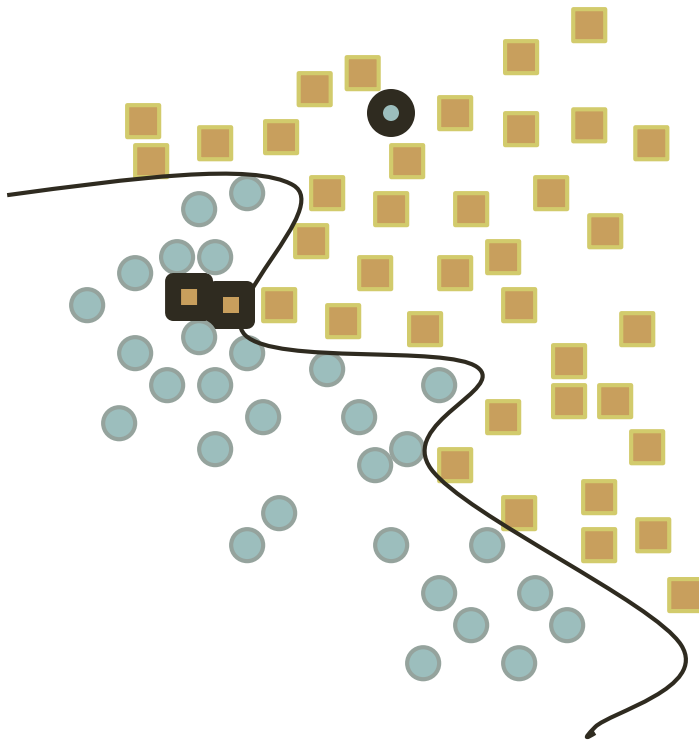


	Sq	Ci
P	35	4
N	5	26

$\text{isSquare} \sim x\text{Location} + y\text{Location}$



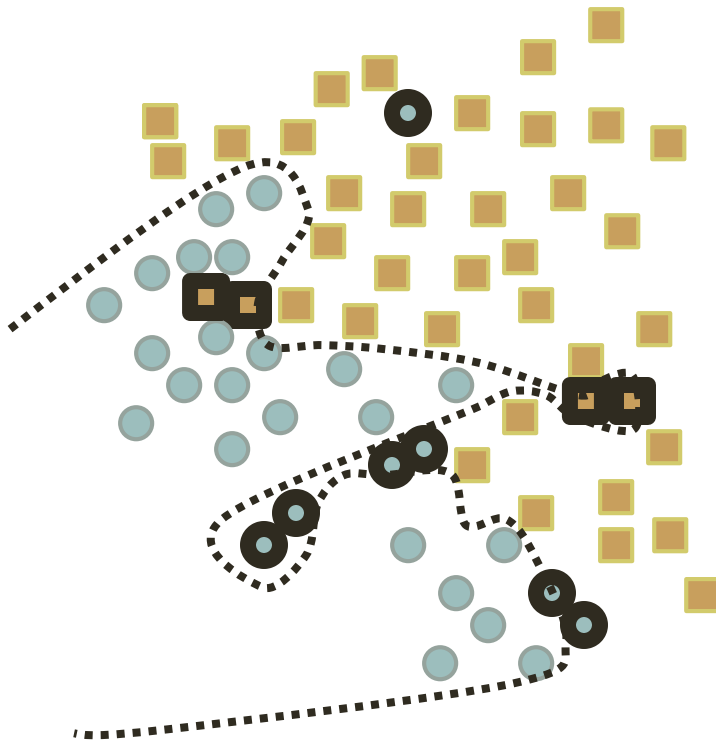
# Evaluate Model : Test Model 2



	Sq	Ci
P	38	1
N	2	29

$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

# Evaluate Model : Test Model 3



	Sq	Ci
P	36	7
N	4	23

$\text{isSquare} \sim \text{xLocation} + \text{yLocation}$

# Over-fitting and Confusion Matrix

# Lesson 09 Quiz 0a Confusion Matrix

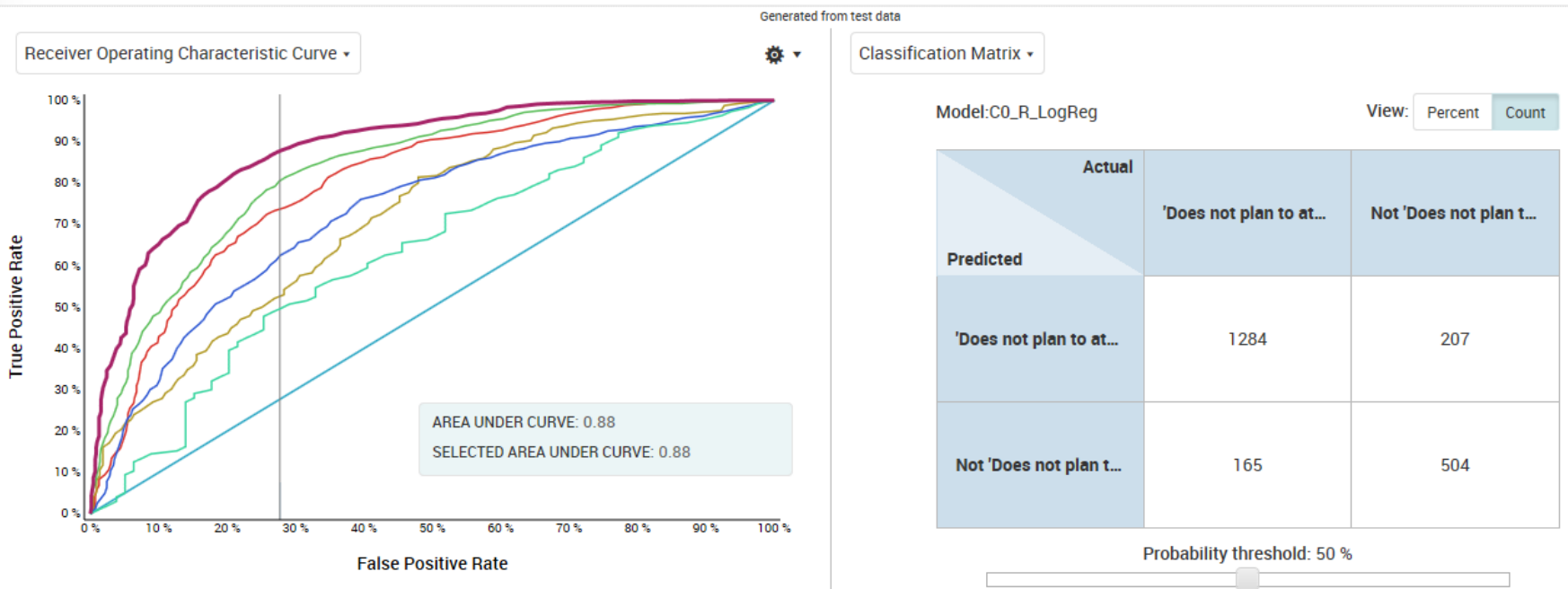
- Quiz Confusion Matrix
  - Test and Accuracy Measures



# ROC Chart Intro

# ROC Chart Intro (1)

- Confusion Matrix and ROC Chart

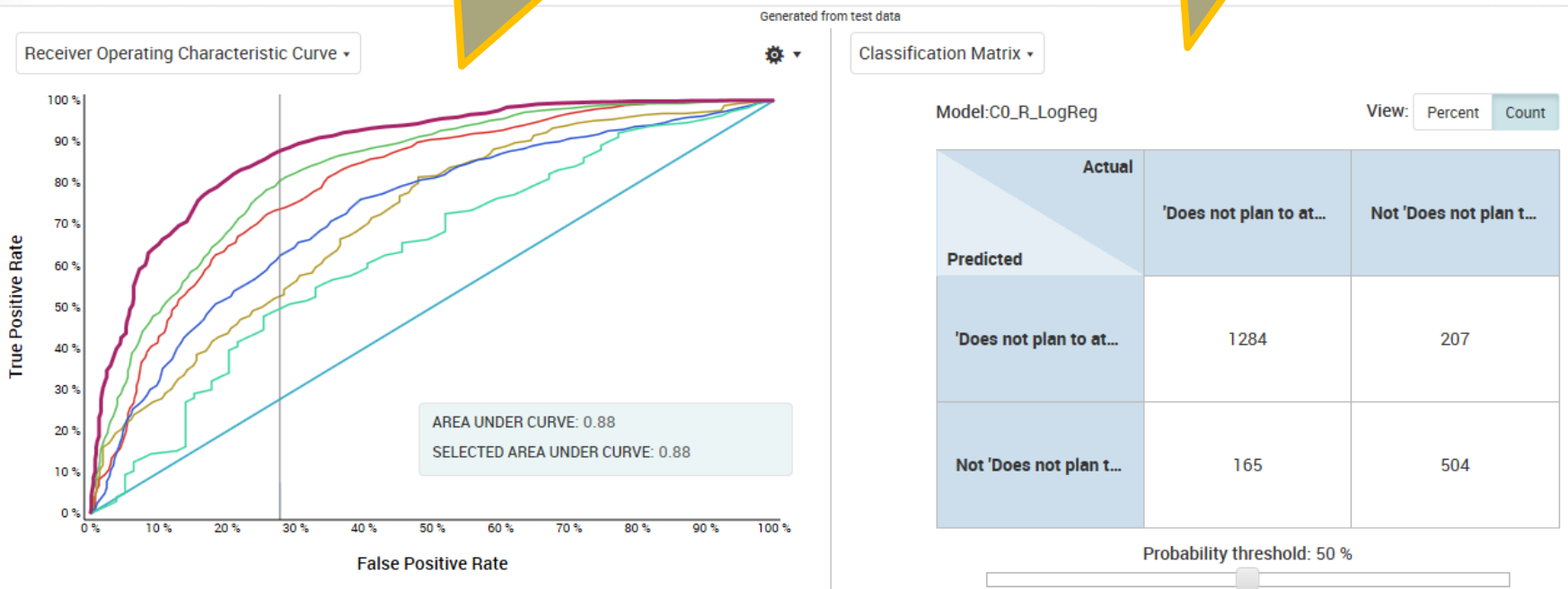


# ROC Chart Intro (2)

- Confusion Matrix and ROC Chart

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

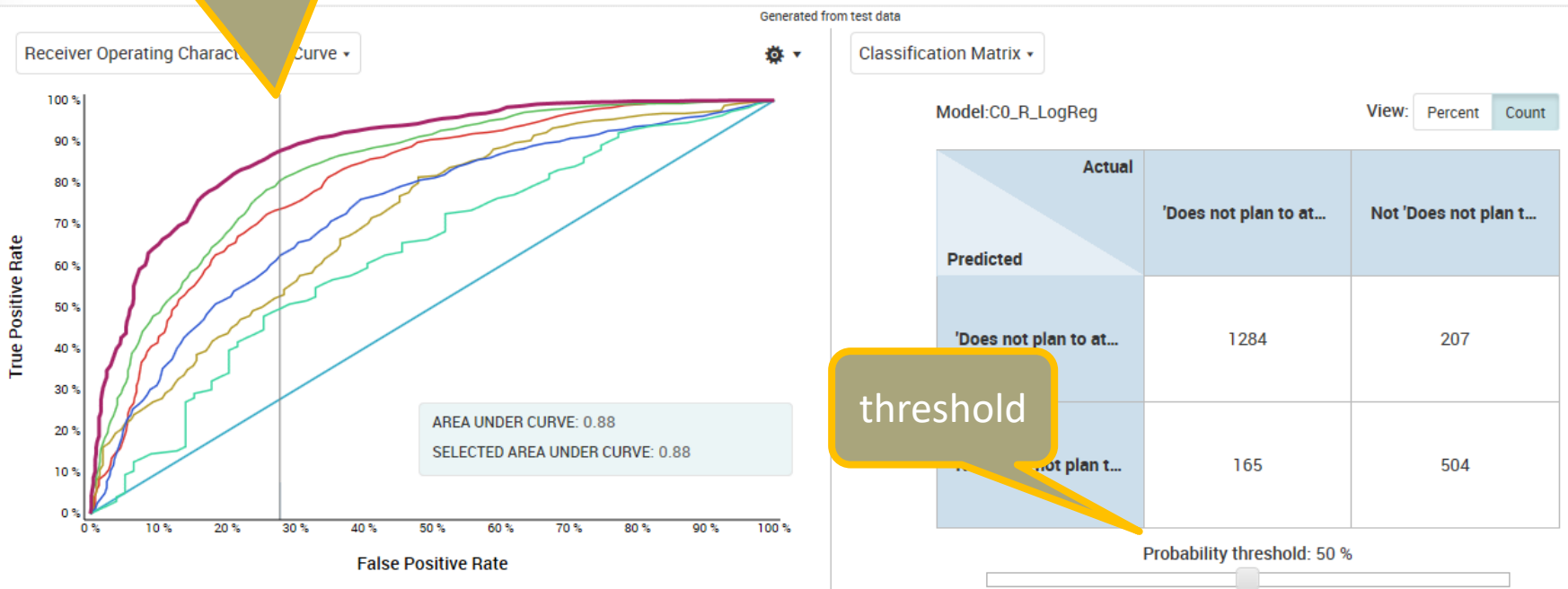
The confusion matrix for  
one model at one threshold



# ROC Chart Intro (3)

- Confusion Matrix and ROC Chart

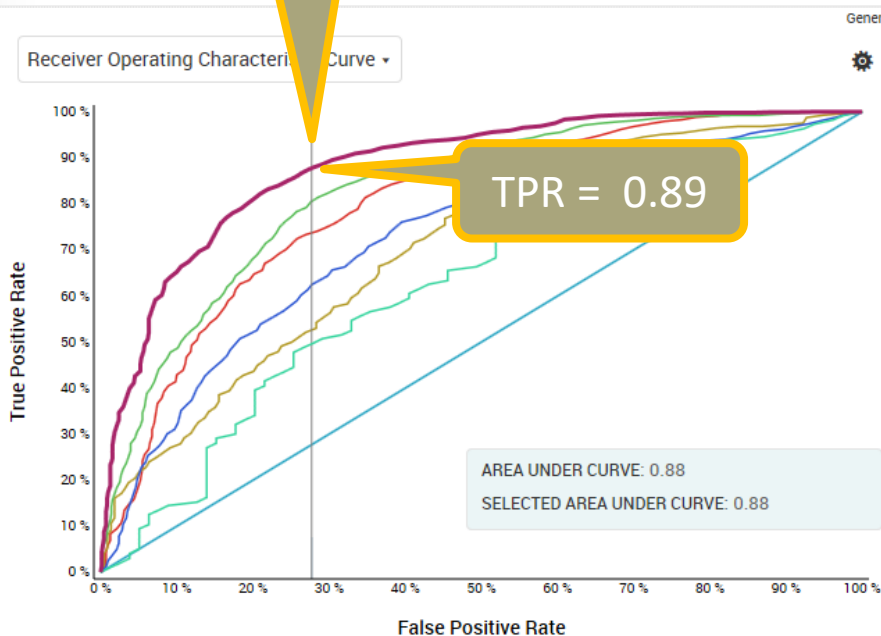
This FPR (0.28) corresponds to the threshold (0.5) for the confusion matrix for the best model





# ROC Chart Intro (4)

- Confusion Matrix and ROC Chart



Generated from test data

Classification Matrix

Model: C0\_R\_LogReg

View: Count

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

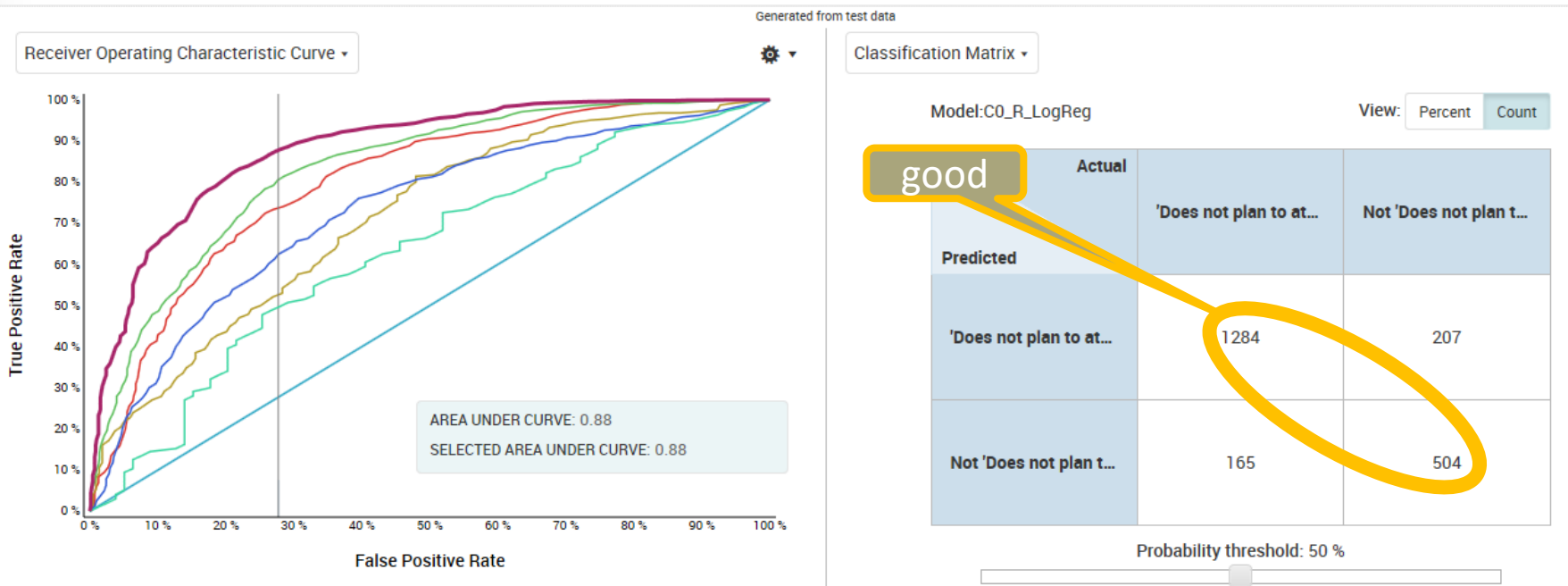
Probability threshold: 50 %

$$\text{FPR} = 207 / (207 + 504)$$

$$\text{TPR} = 1284 / (1284 + 165)$$

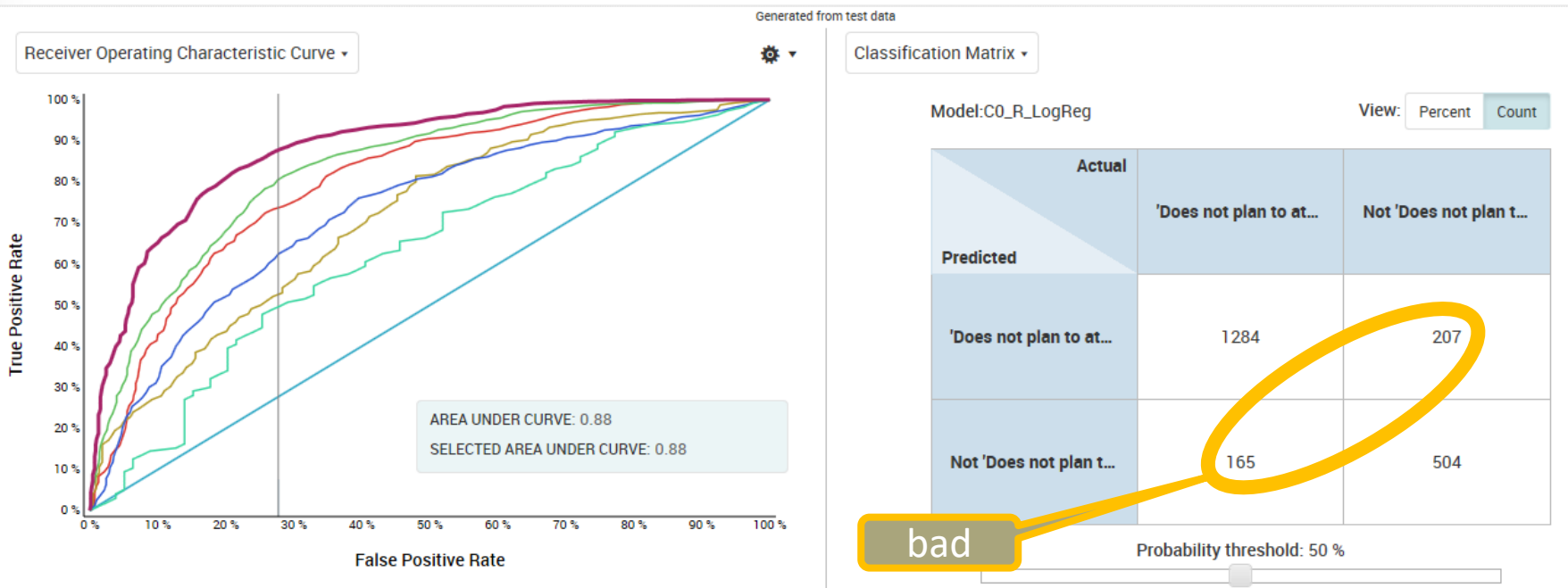
# ROC Chart Intro (5)

- Confusion Matrix and ROC Chart



# ROC Chart Intro (6)

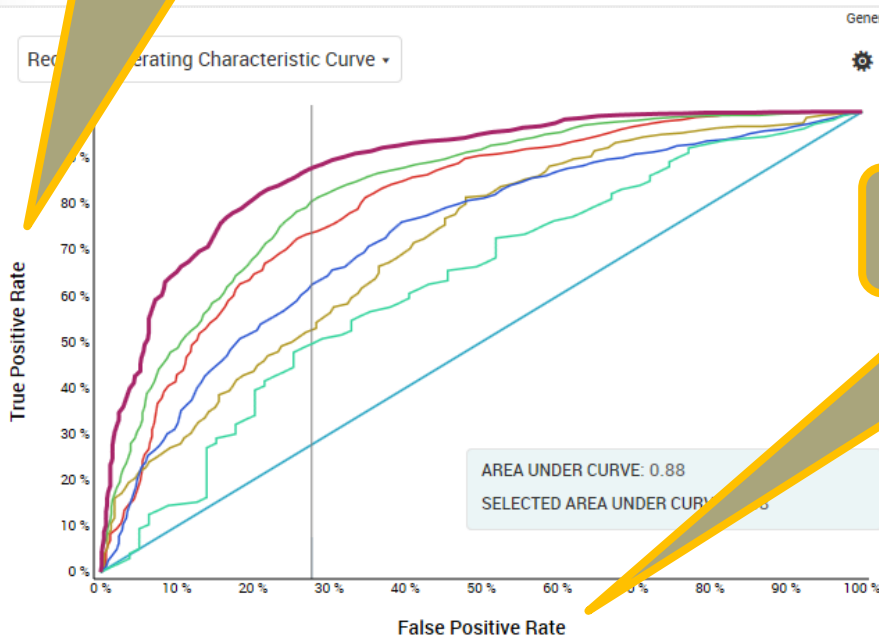
- Confusion Matrix and ROC Chart



# ROC Chart Intro (7)

- Confusion Matrix and ROC Chart

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



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

Generated from test data

Classification Matrix

Model: C0\_R\_LogReg

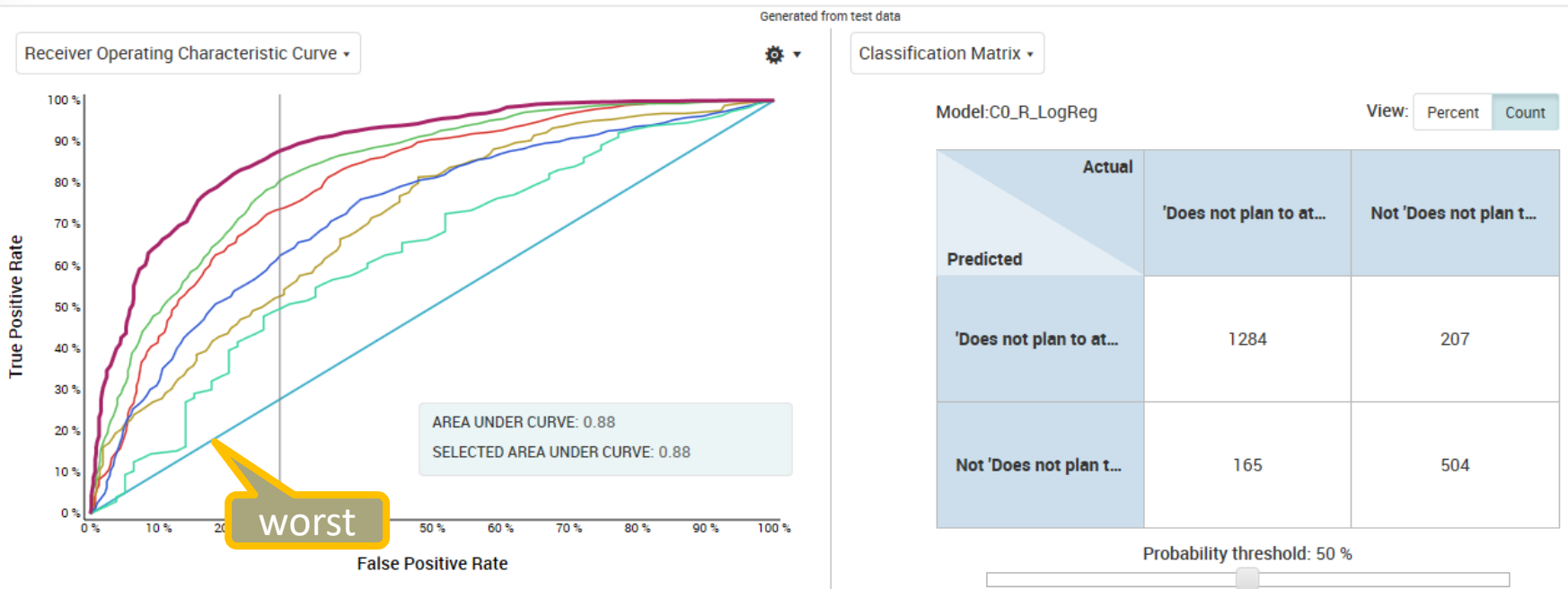
View: Percent Count

	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 %

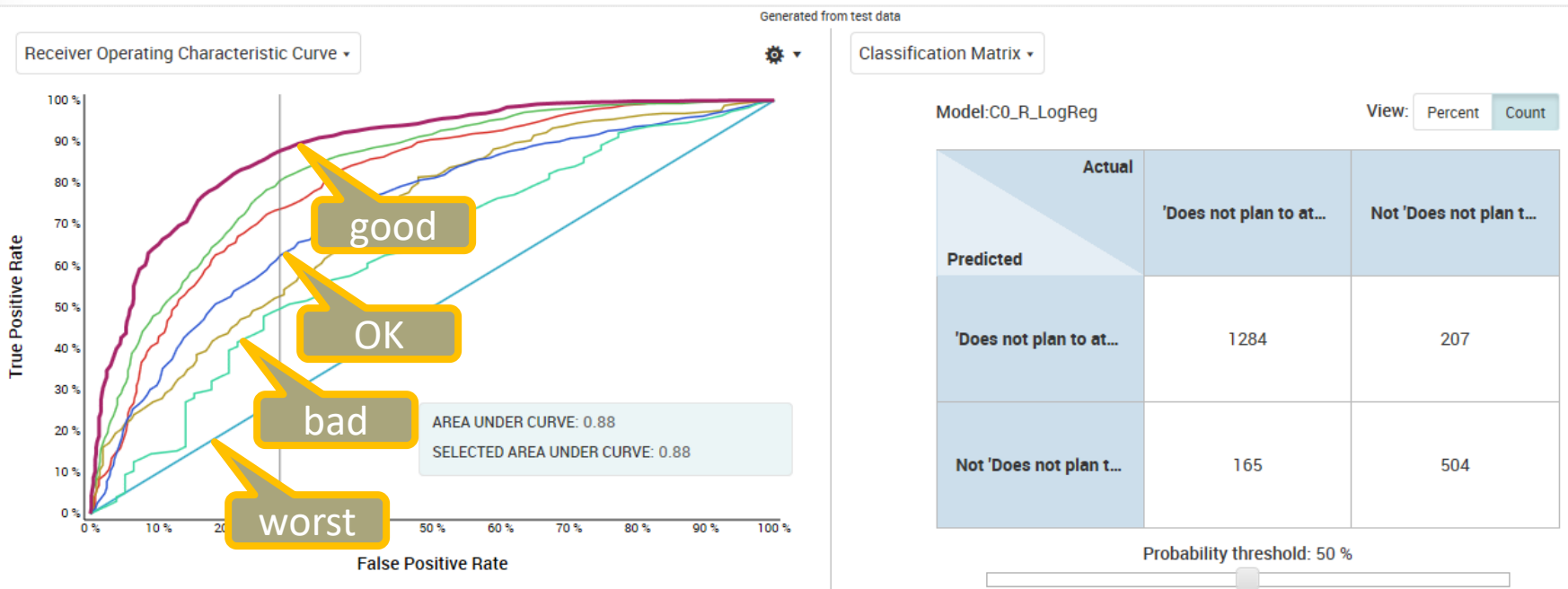
# ROC Chart Intro (8)

- Confusion Matrix and ROC Chart



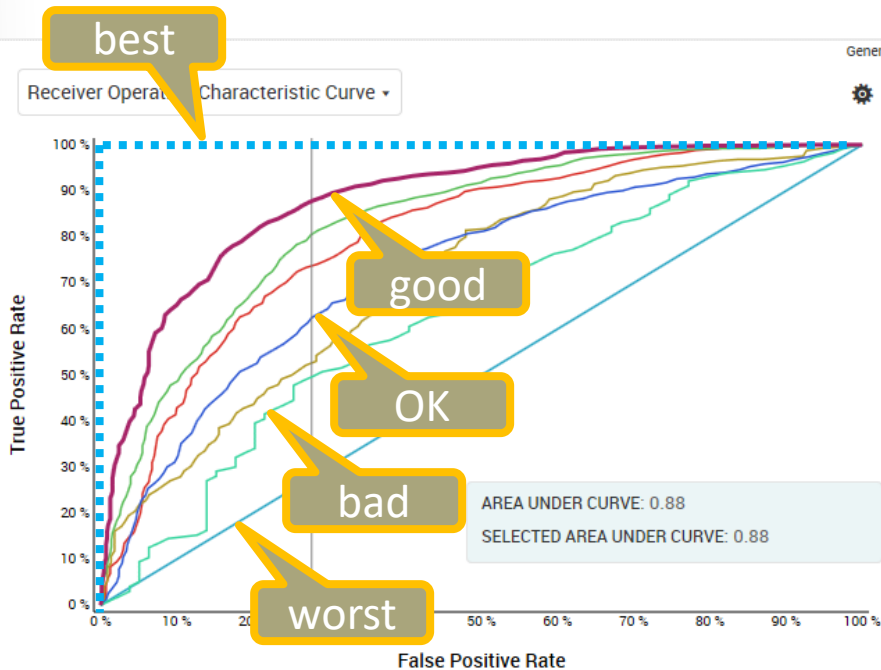
# ROC Chart Intro (9)

- Confusion Matrix and ROC Chart



# ROC Chart Intro (10)

- Confusion Matrix and ROC Chart



Classification Matrix

Model: C0\_R\_LogReg

View: Percent Count

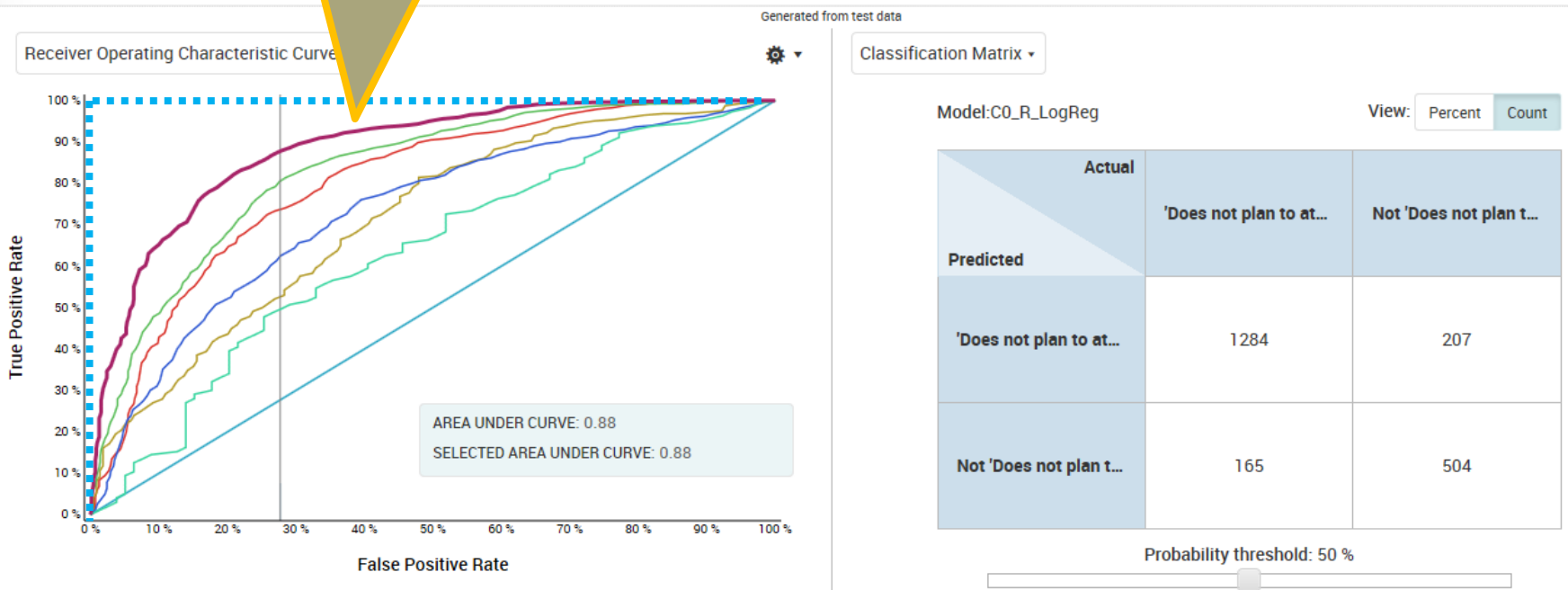
Actual \ Predicted	'Does not plan to at...	Not 'Does not plan t...
'Does not plan to at...	1284	207
Not 'Does not plan t...	165	504

Probability threshold: 50 %

# ROC Chart Intro (11)

- Confusion Matrix and ROC Chart

ROC charts are non-decreasing functions





# ROC Chart Intro

# Lesson 09 Quiz 0b ROC Intro

- Quiz ROC Intro



# 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)

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			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					
17						Threshold:	0.5			
18						FPR:	0			
						TPR:	#DIV/0!			

# How to make an 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
		Predicted	Predicted								
	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%

# How to make an ROC (3)

The Predicted Probabilities need a threshold

	A	B		G	H	I	J	K
		Predicted	Predicted					
2	Actual	Probability	Class	TP	FP	FN	TN	Threshold
3	1	0.55	1	1	0	0	0	0
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
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)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

G16 fx 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						
17											
18											

Set the threshold for the Predicted Probabilities

Threshold: 0.5

FPR: 0.4

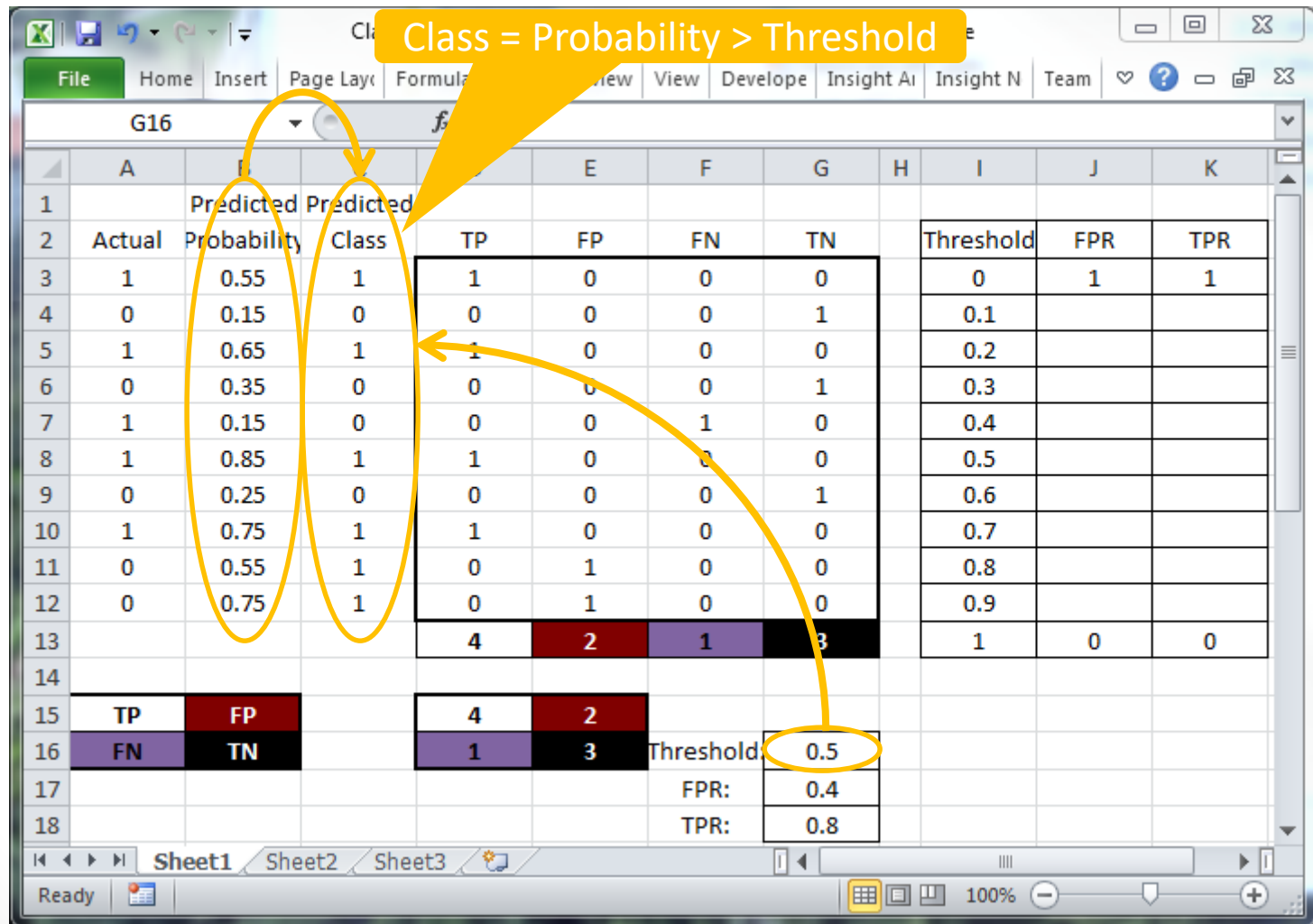
TPR: 0.8

Sheet1 Sheet2 Sheet3

Ready 100%



# How to make an ROC (5)

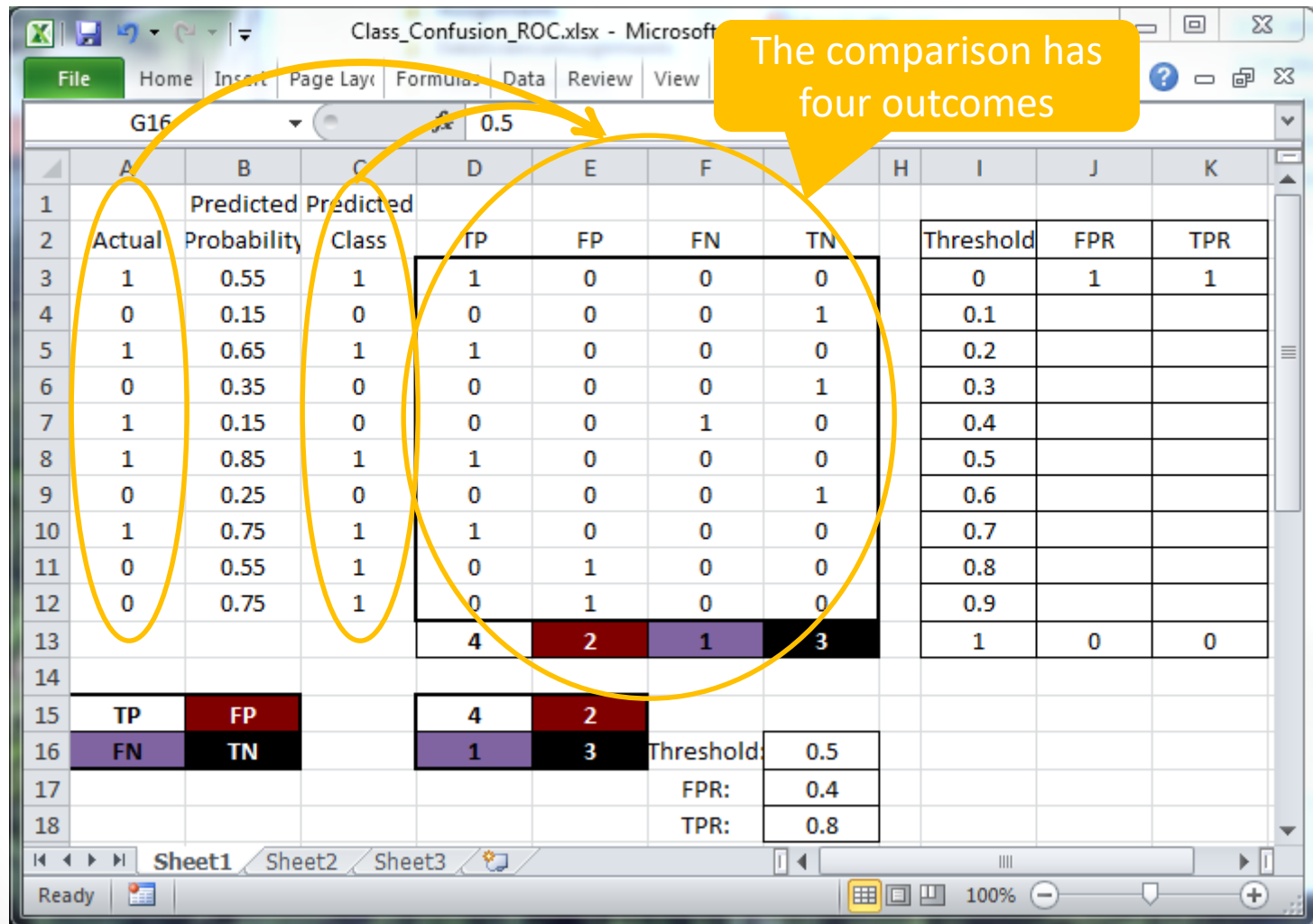


# How to make an ROC (6)

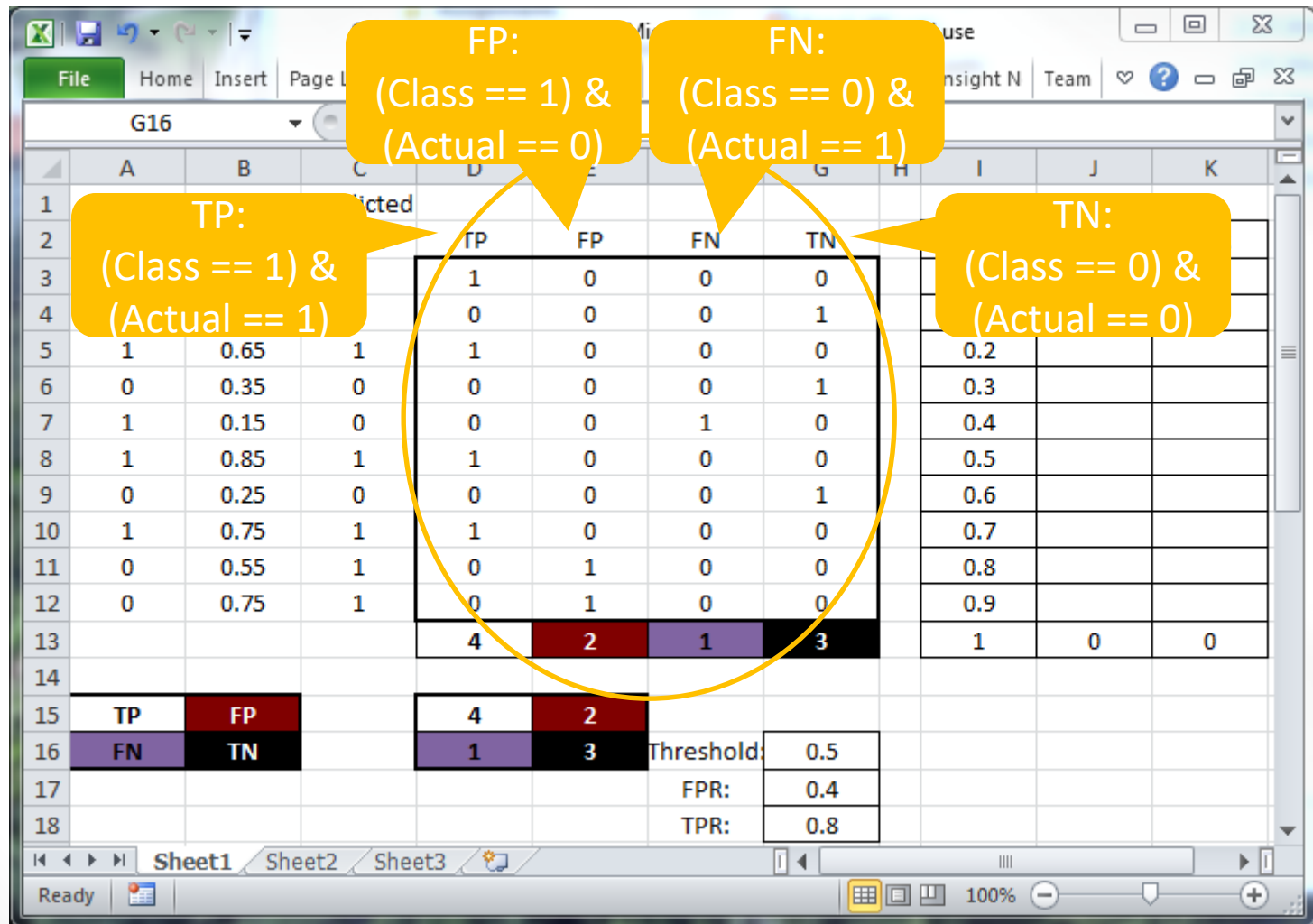
Compare the predicted Class to the Actual Values

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

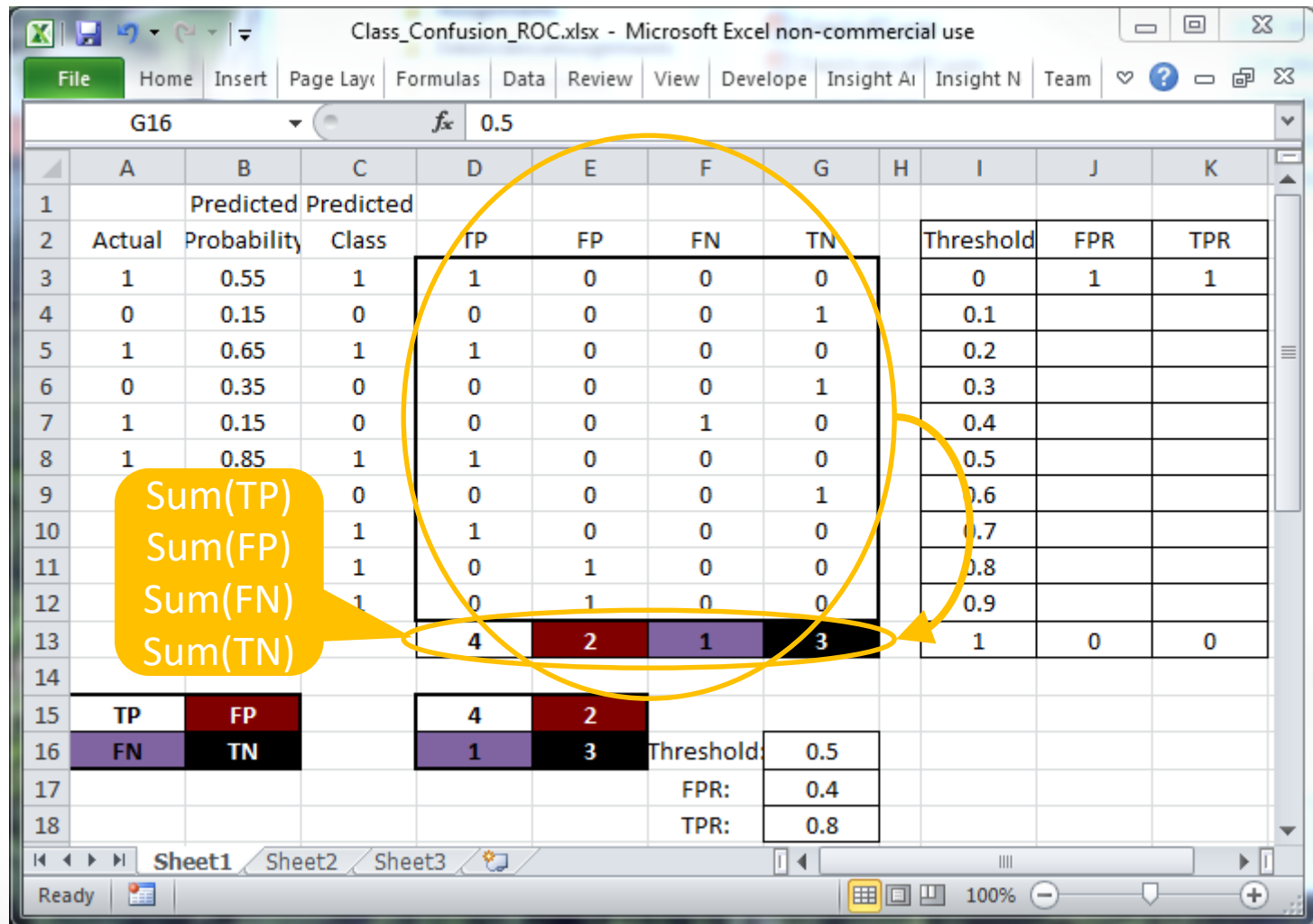
# How to make an ROC (7)



# How to make an ROC (8)



# How to make an ROC (9)



# How to make an ROC (10)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

G16 fx 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	0.5	1	0	1	0	0	0.8			
12	0	0.5	1	0	1	0	0	0.9			
13				4	2	1	3	1	0	0	
14											
15	TP	FP		4	2			Threshold:	0.5		
16	FN	TN		1	3			FPR:	0.4		
17								TPR:	0.8		
18											

Optional:  
Organize sums into  
Confusion Matrix

# How to make an ROC (11)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

G16 fx 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						
17											
18											

FPR = FP/(FP + TN)

Threshold: 0.5

FPR: 0.4

TPR: 0.8

Sheet1 Sheet2 Sheet3

Ready 100%

# How to make an ROC (12)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

G16 fx 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						
17											
18											

TPR = TP / (TP + FN)

Threshold: 0.5  
FPR: 0.4  
TPR: 0.8

Sheet1 Sheet2 Sheet3

Ready 100%



# How to make an ROC (13)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

G16 fx 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%

# How to make an ROC (14)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

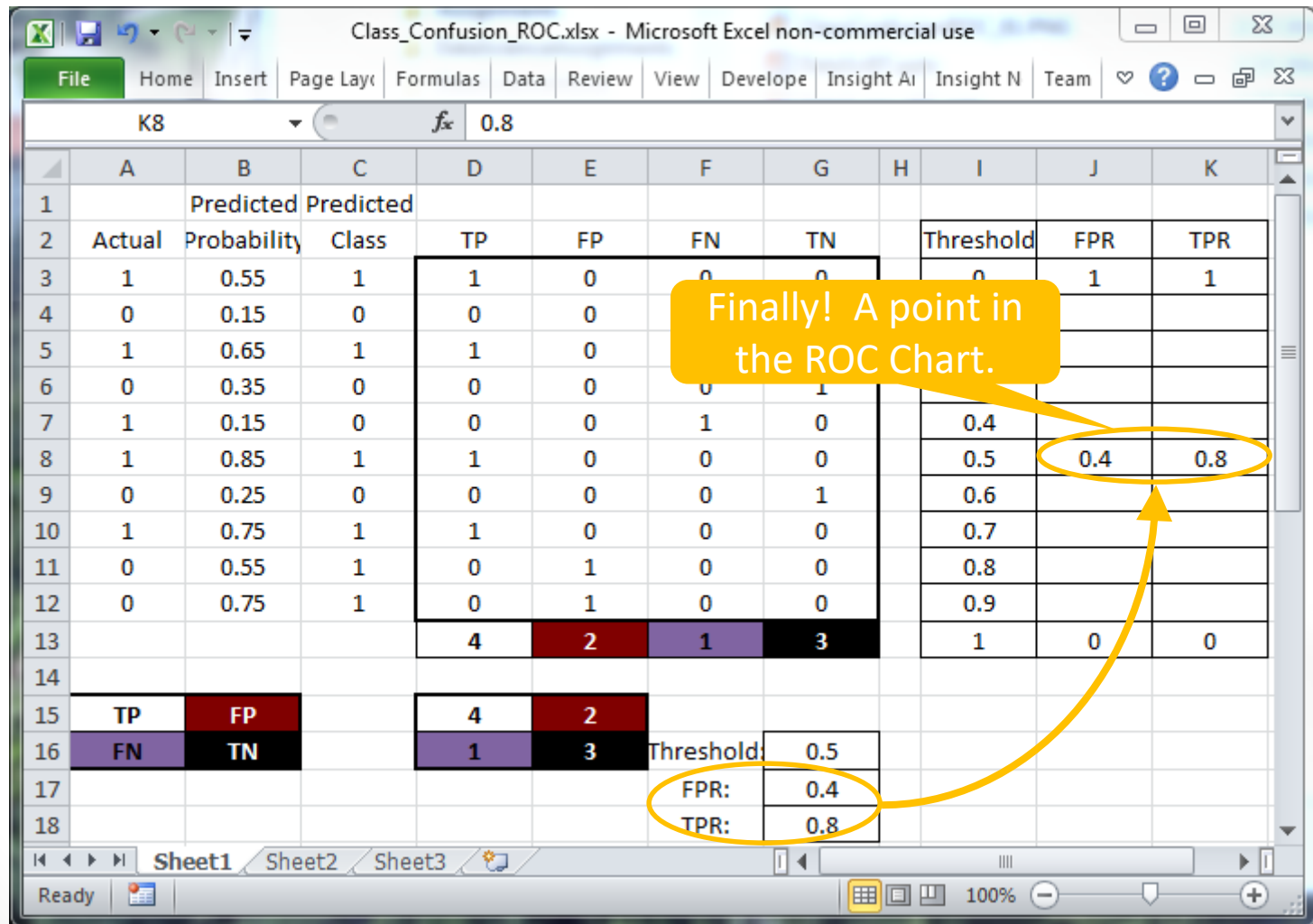
G16 fx 0.5

	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%

# How to make an ROC (15)



# How to make an ROC (16)

Class\_Confusion\_ROC.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Developer Insight AI Insight N Team

K9 fx 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		

Sheet1 Sheet2 Sheet3

Ready 100%

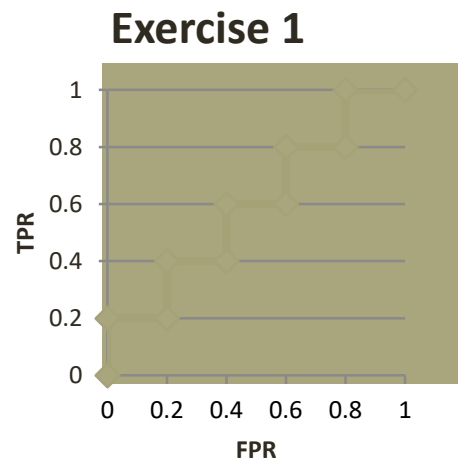
Repeat the process for all thresholds

# 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



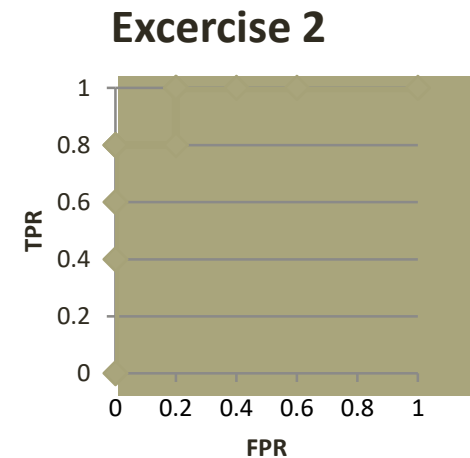
FPR	TPR
1	1
0	0



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



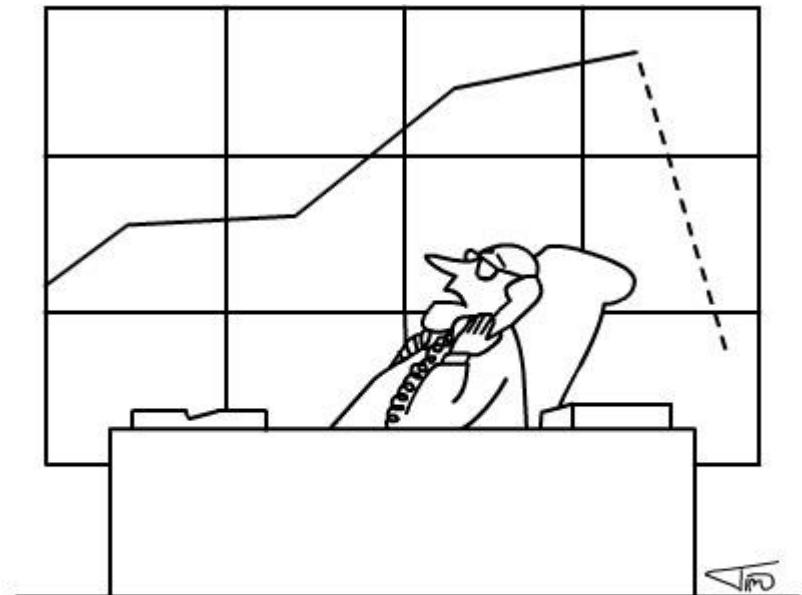
FPR	TPR
1	1
0	0



# How to make an ROC

# Break

- Documentaries on Predictive Analytics:
  - <https://www.analyticsvidhya.com/blog/2015/11/7-watch-documentaries-statistics-machine-learning/>



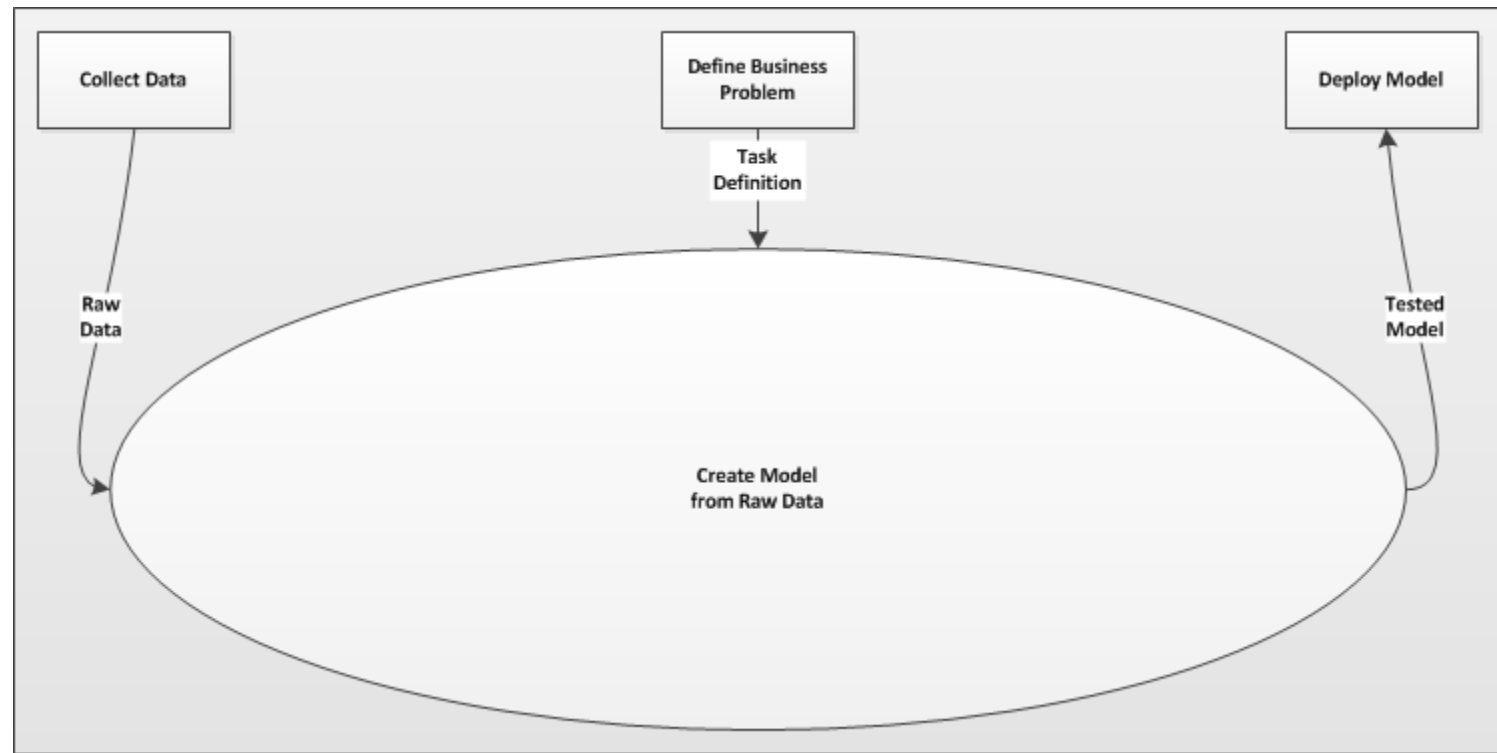
*"BI tech support? The predictive analysis system is giving the wrong answer again—can you please fix it?... "*

# Modeling Iteration Trap

The problem with iterative data preparation and training refinements



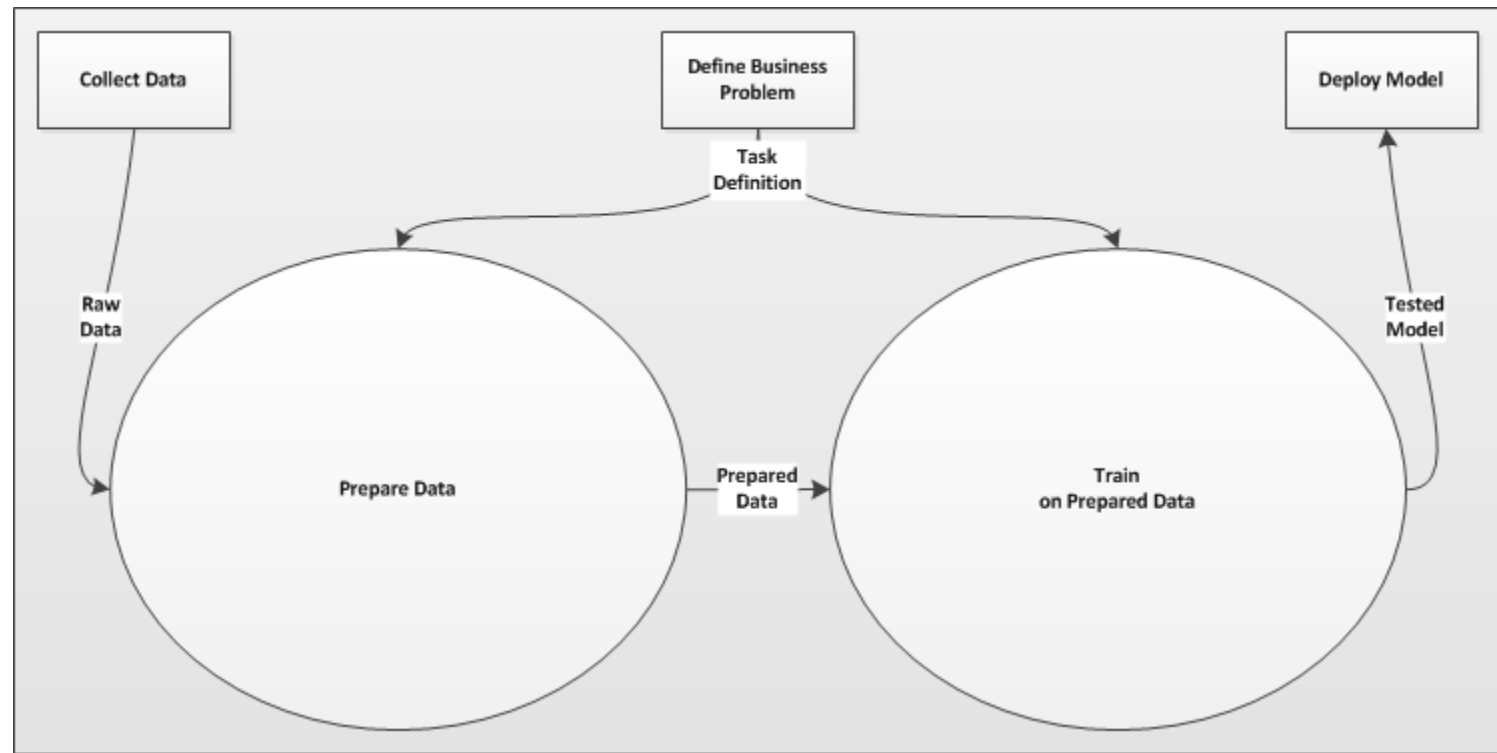
# Modeling Iteration Trap (0)



Basic Predictive Analytics

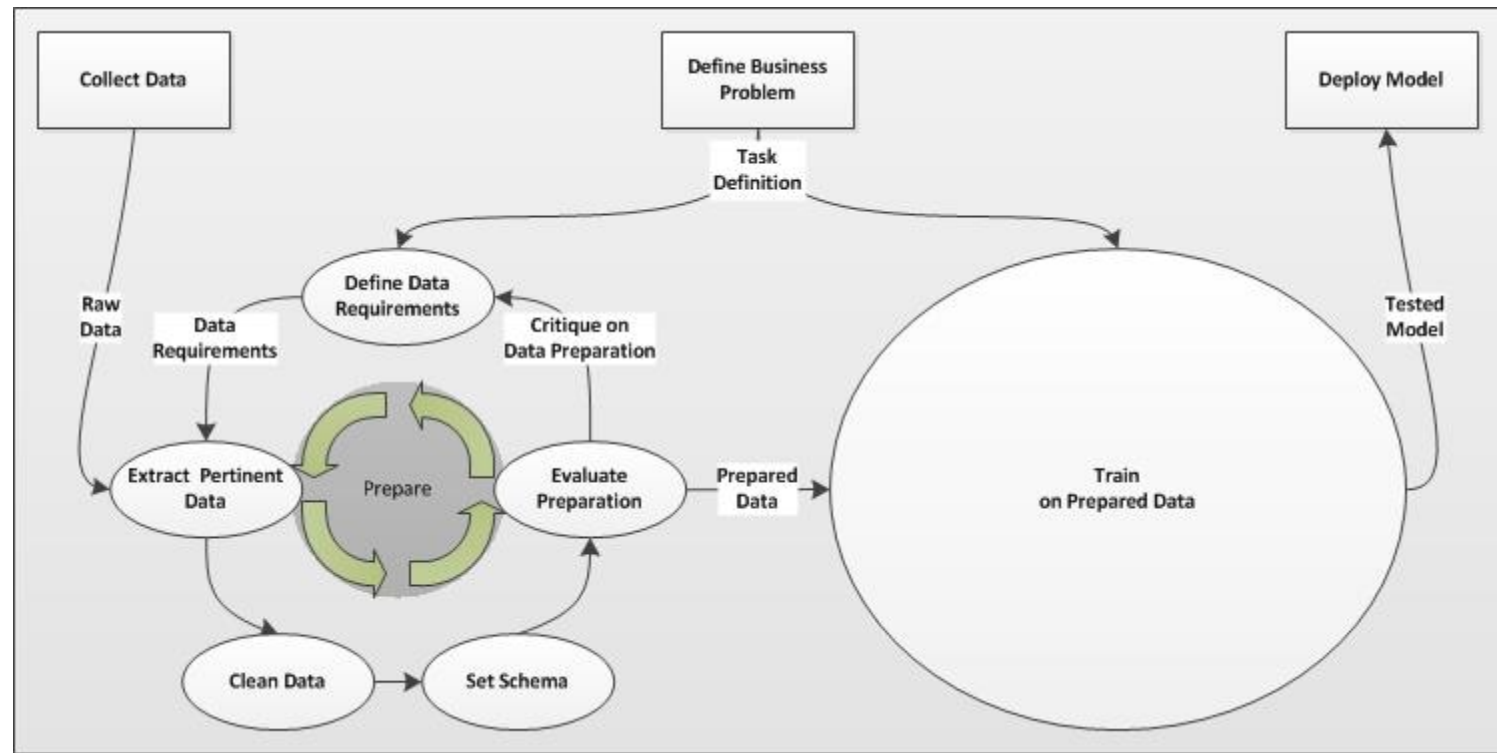
LO-DFD

# Modeling Iteration Trap (1)



Predictive Analytics consists of Data  
Preparation and Modeling  
L1-DFD

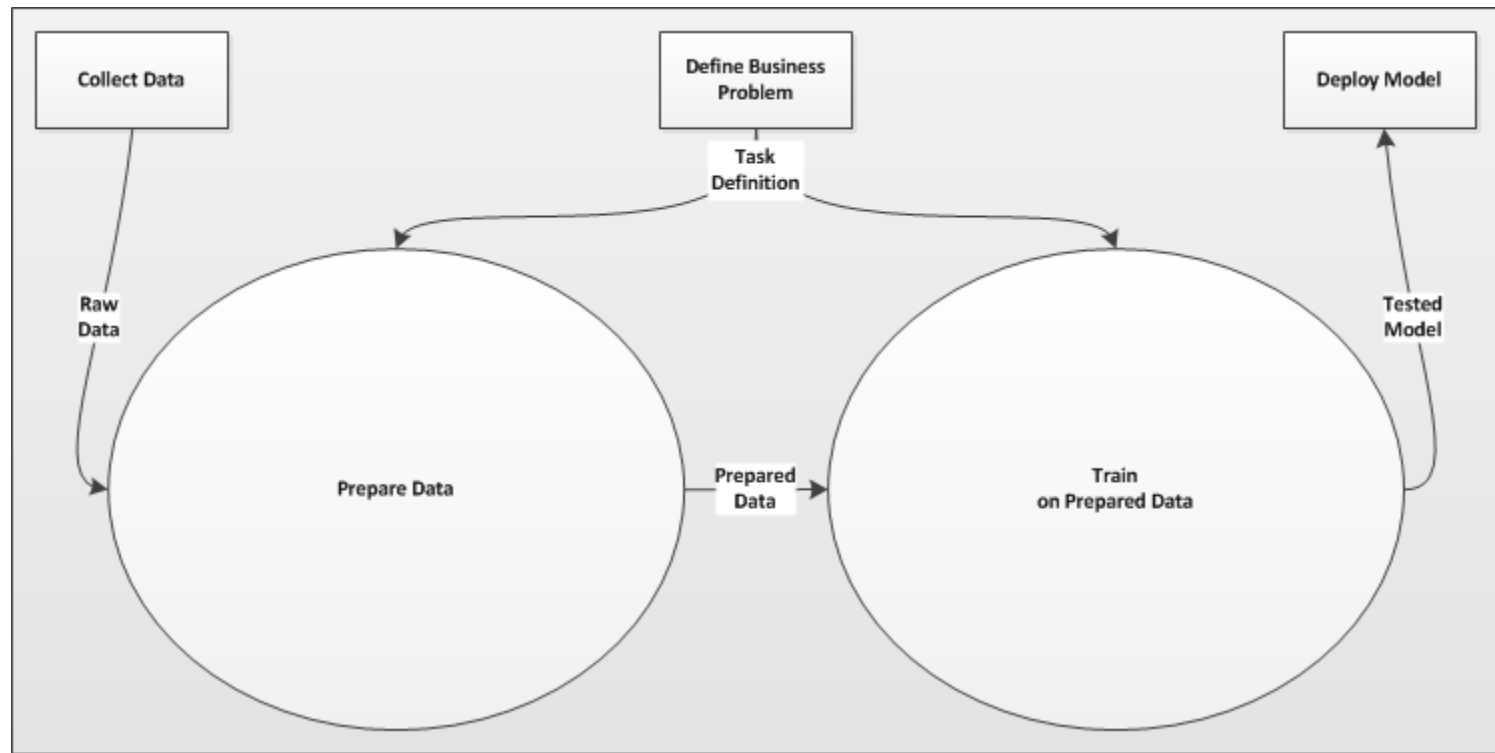
# Modeling Iteration Trap (2)



# Data Preparation is Iterative

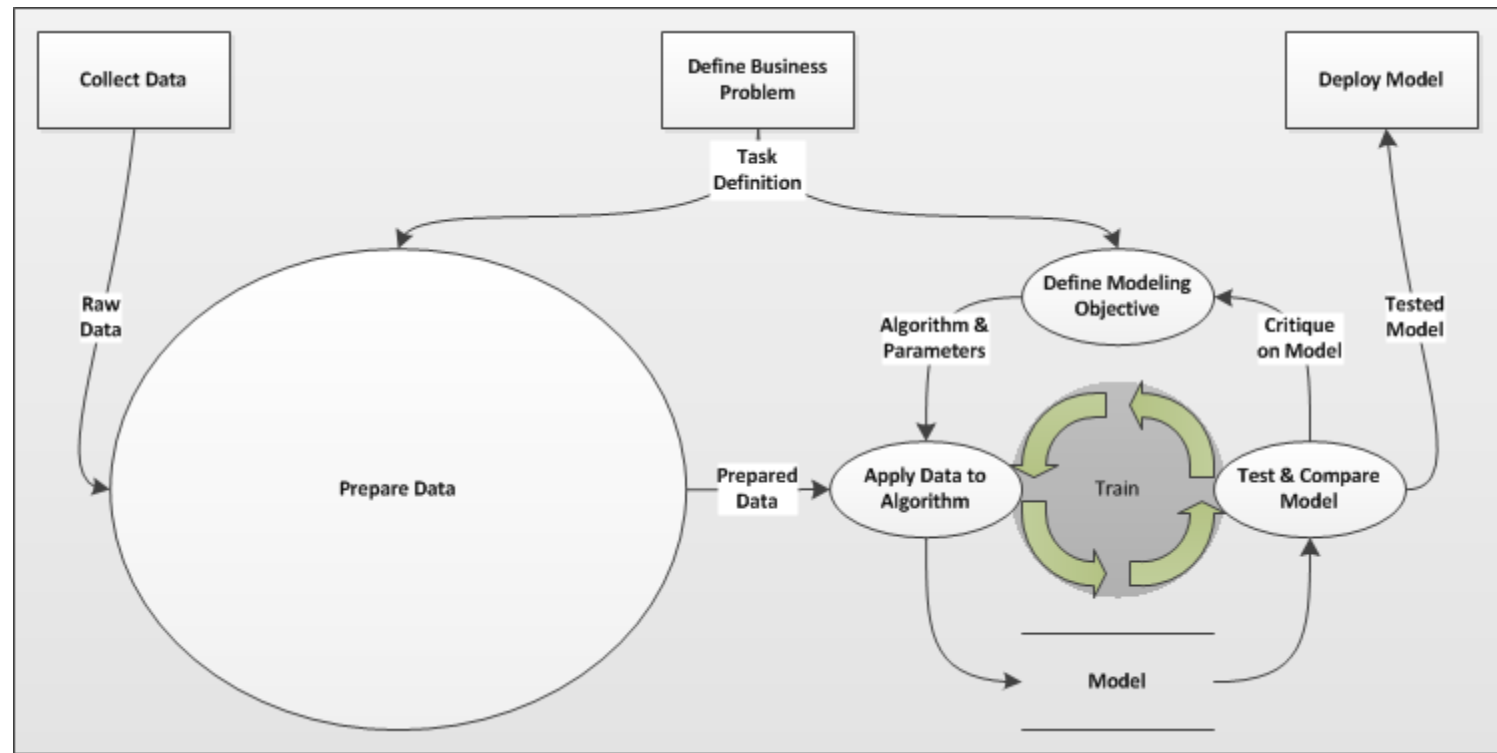
## L2-DFD

# Modeling Iteration Trap (3)



Predictive Analytics consists of Data Preparation and Modeling  
L1-DFD

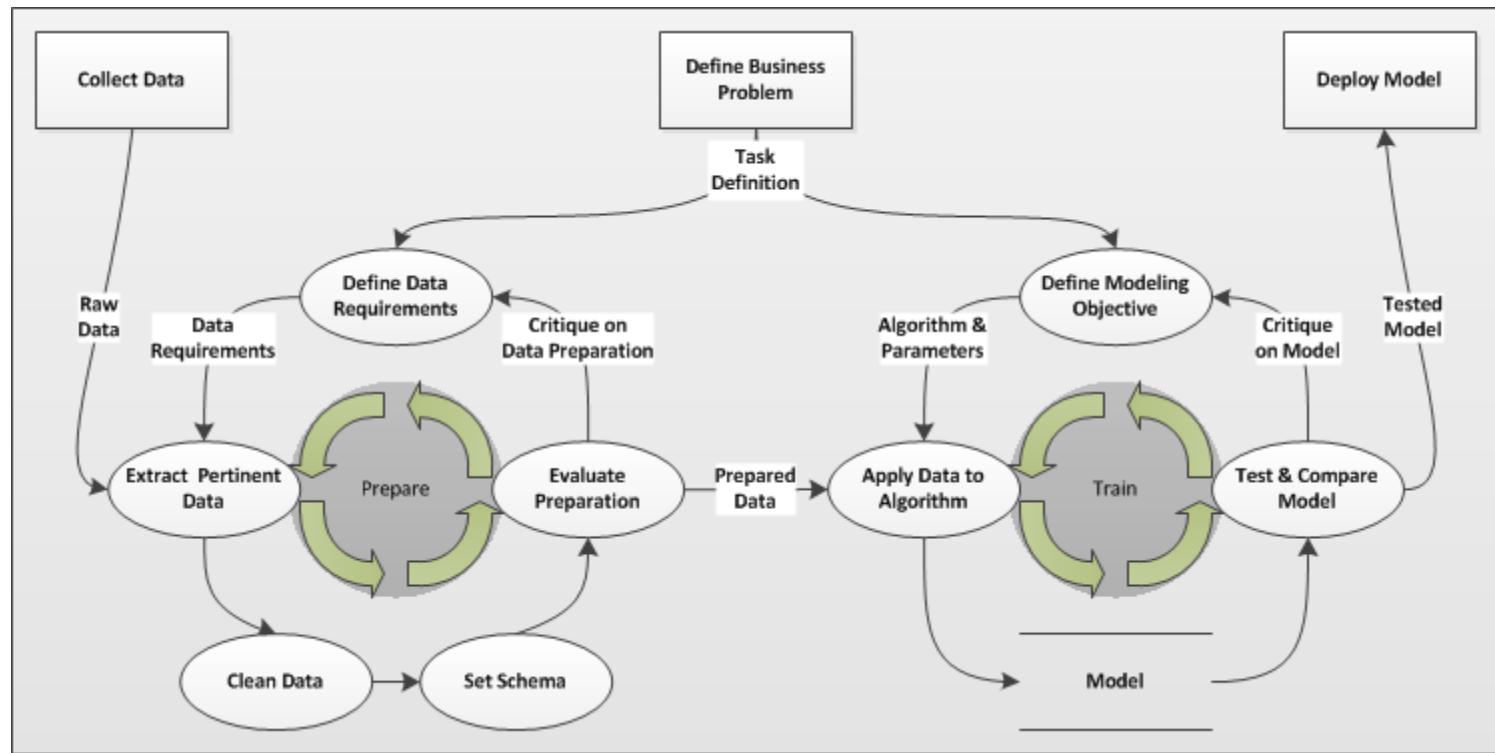
# Modeling Iteration Trap (4)



Training is Iterative

L2-DFD

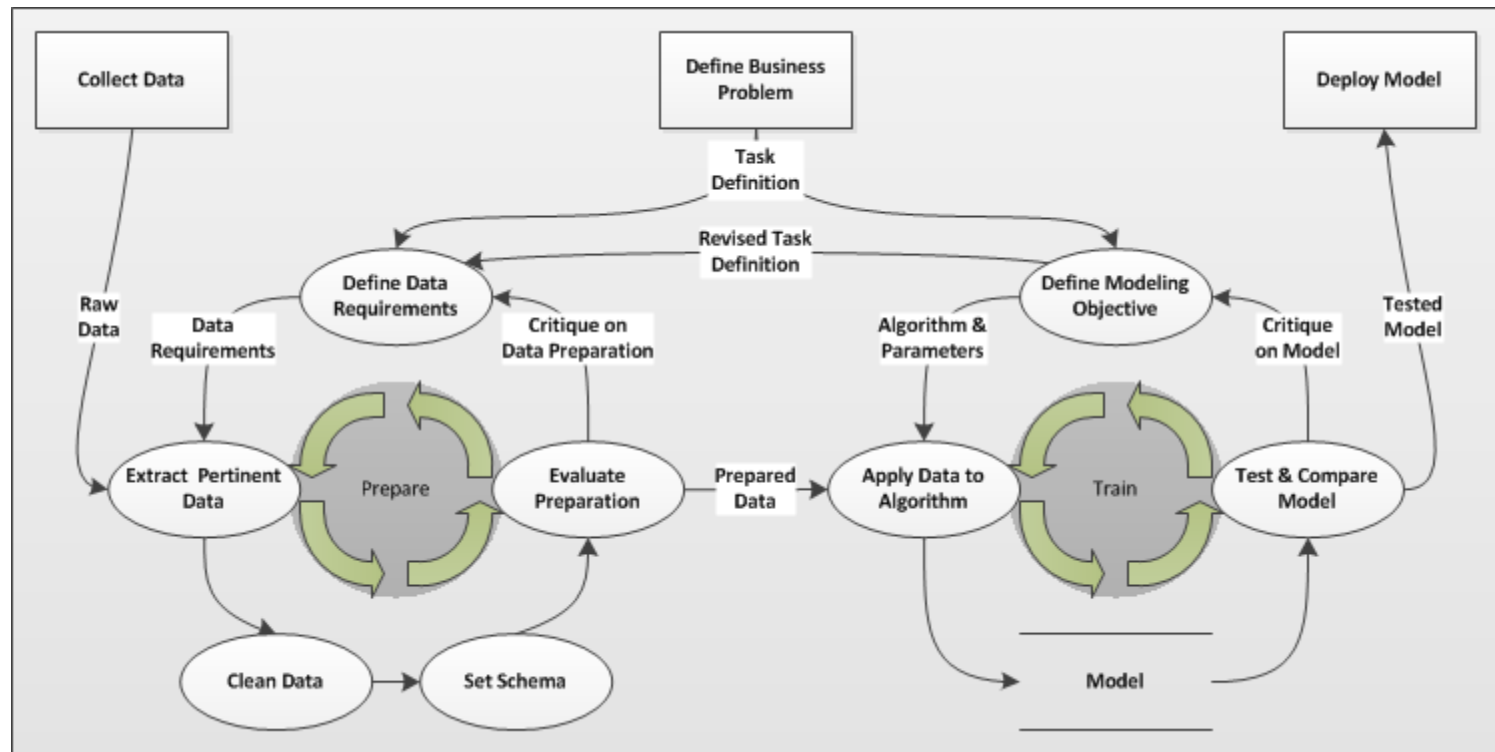
# Modeling Iteration Trap (5)



Data Preparation and Training

L2-DFD

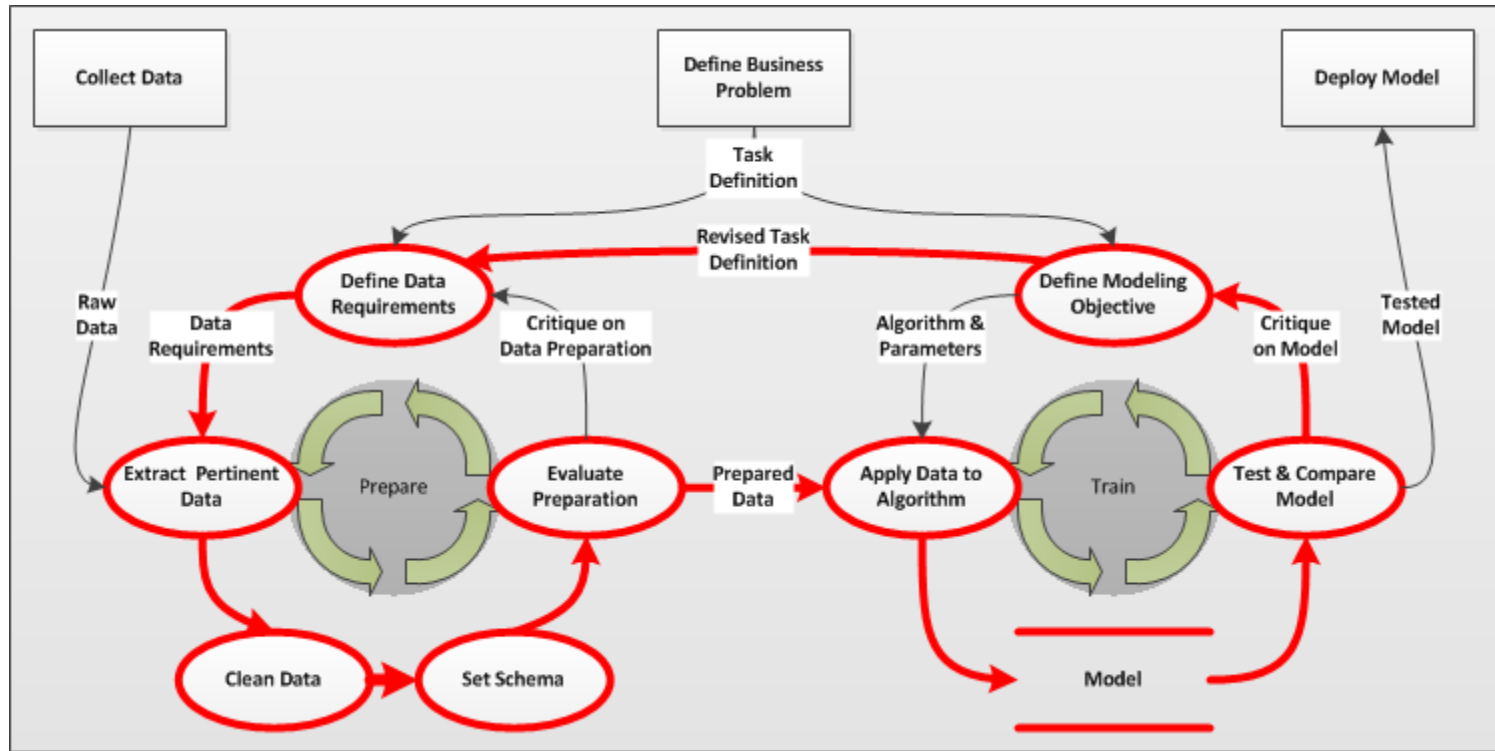
# Modeling Iteration Trap (6)



Modeling Revises the Task Definition

L2-DFD

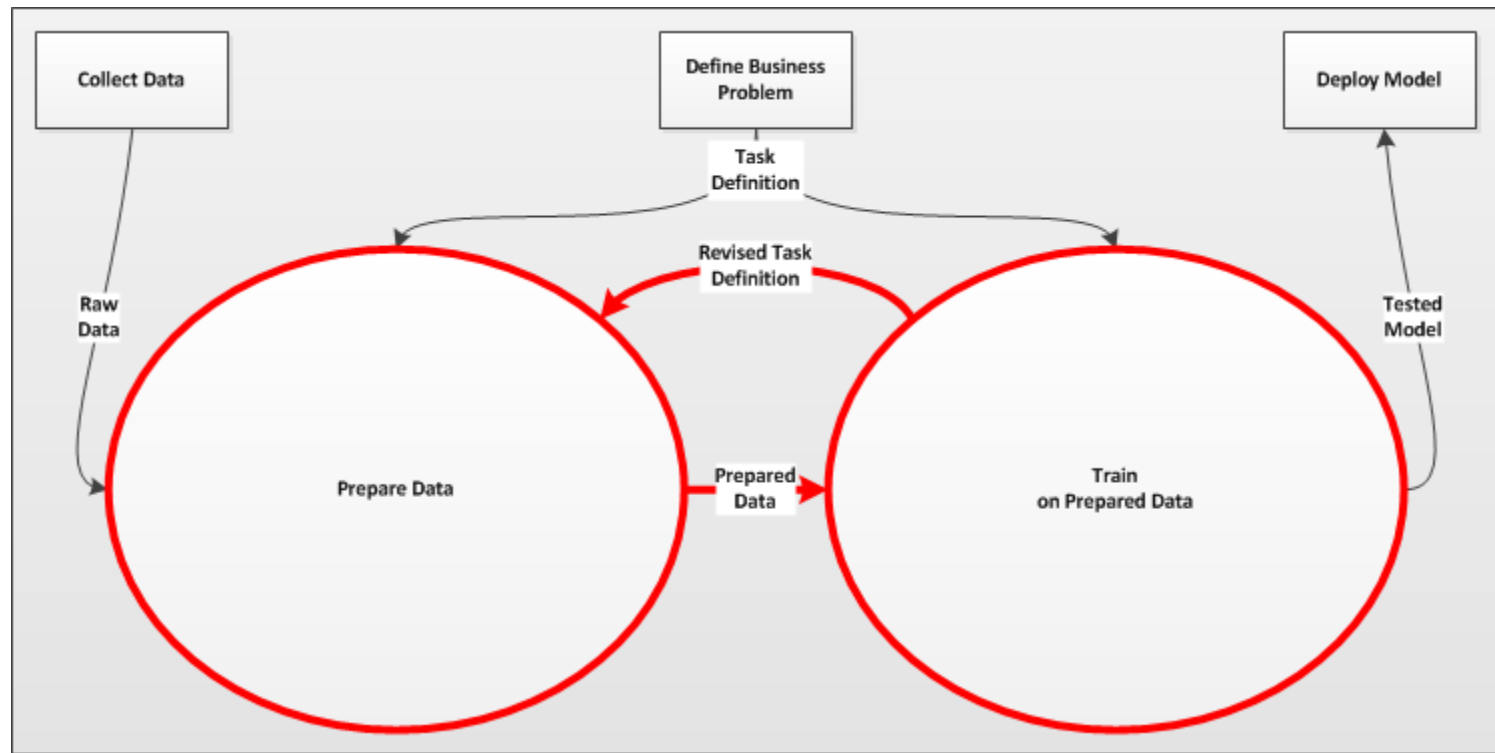
# Modeling Iteration Trap (7)



Revised Task Definition Completes a Cycle that is hard to optimize. Engineers get caught in a loop.

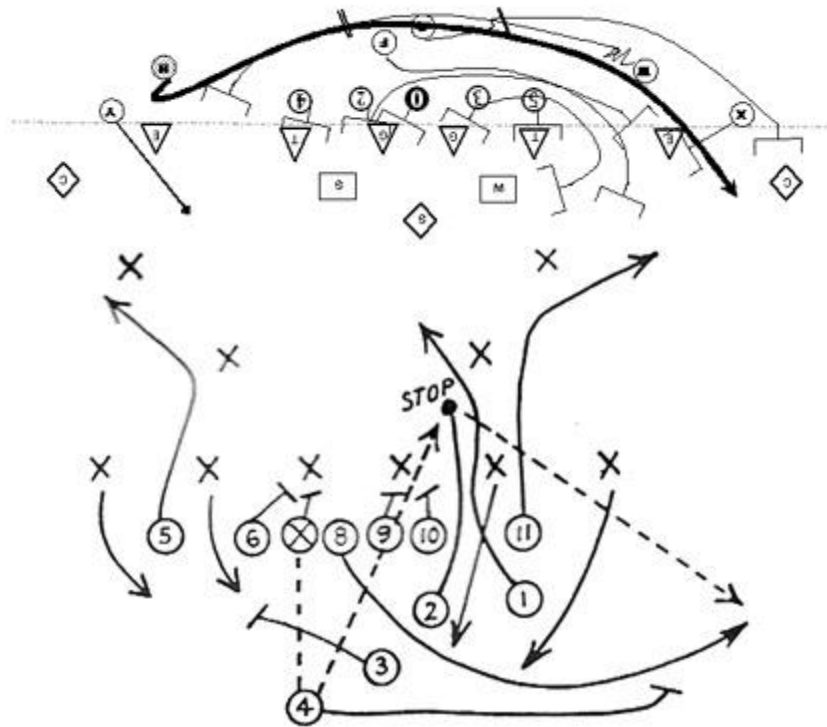


# Modeling Iteration Trap (8)



Revised Task Definition Completes a Cycle that is hard to optimize. Engineers get caught in a loop.

# Modeling Iteration Trap (9)



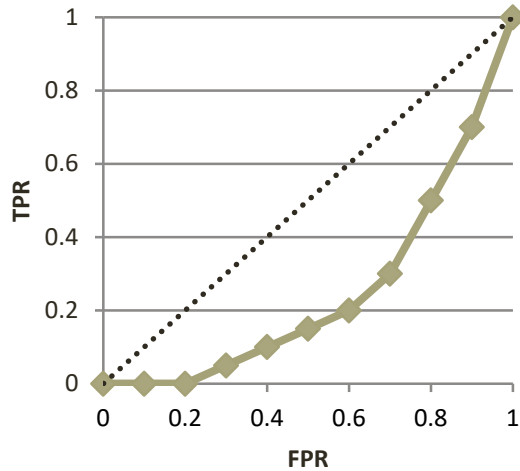
# Modeling Iteration Trap

The problem with iterative data preparation and training refinements

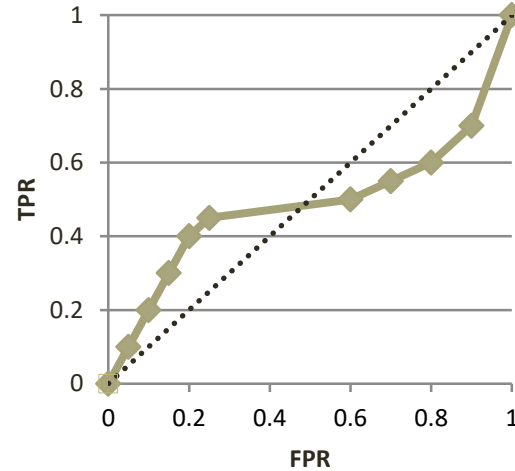
# Lesson 09 Quiz 1 Question 9:

## Which ROC curves are incorrect?

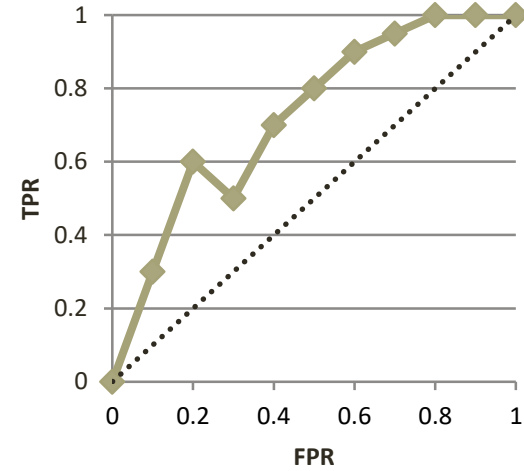
**ROC 1**



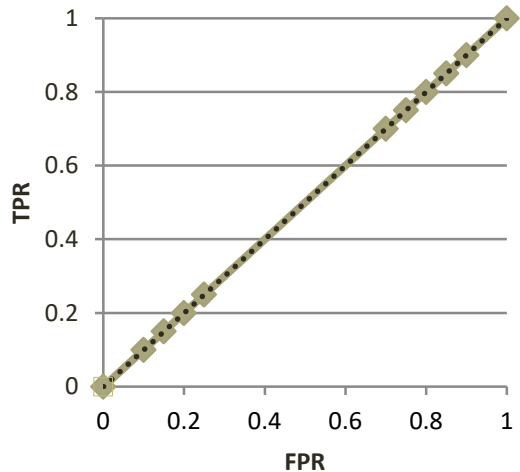
**ROC 2**



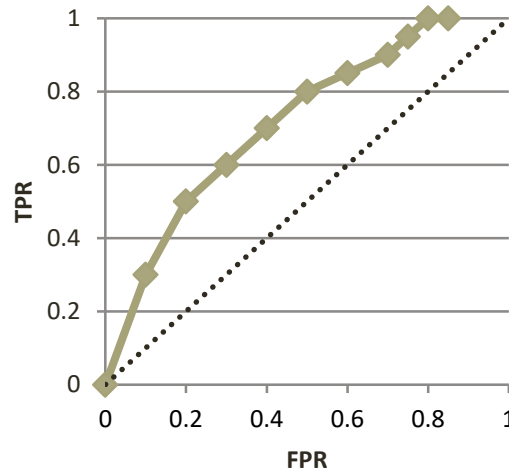
**ROC 3**



**ROC 4**



**ROC 5**



# Review Terminology

- Algorithm
- Anomaly detection
- Association
- Attribute
- Binarize Categories
- Binary Column
- Case
- Category Column
- Character Column
- Classification
- Clustering
- Coercion
- Column
- Column Header
- Confusion Matrix
- Data
- Data Dimensionality
- Data Frame
- Data Type
- DFD
- Dummy Variable
- Estimation
- False Positive
- False Negative
- Feature Scaling
- Field
- Hypothesis
- Key Column
- Machine Learning
- Market-basket analysis
- Matrix
- Missing Data
- Model
- Multinomial Column
- Normalization
- Numeric Column
- Observation
- One-hot encoding
- Outcome
- Outlier Removal
- Predictive Analytics
- Rectangular Data
- Relabeling
- ROC curve
- 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

# Introduction to Data Science