CMSC 473/673
Natural Language Processing

Instructor: Lara J. Martin (she/they)

TA: Duong Ta (he)

Slides modified from Dr. Frank Ferraro & Cynthia Matuszek

Learning Objectives

Recognize the ML data & training pipeline

Evaluate the effectiveness of a model regardless of what the model looks like

Review: ML/NLP Framework for Prediction

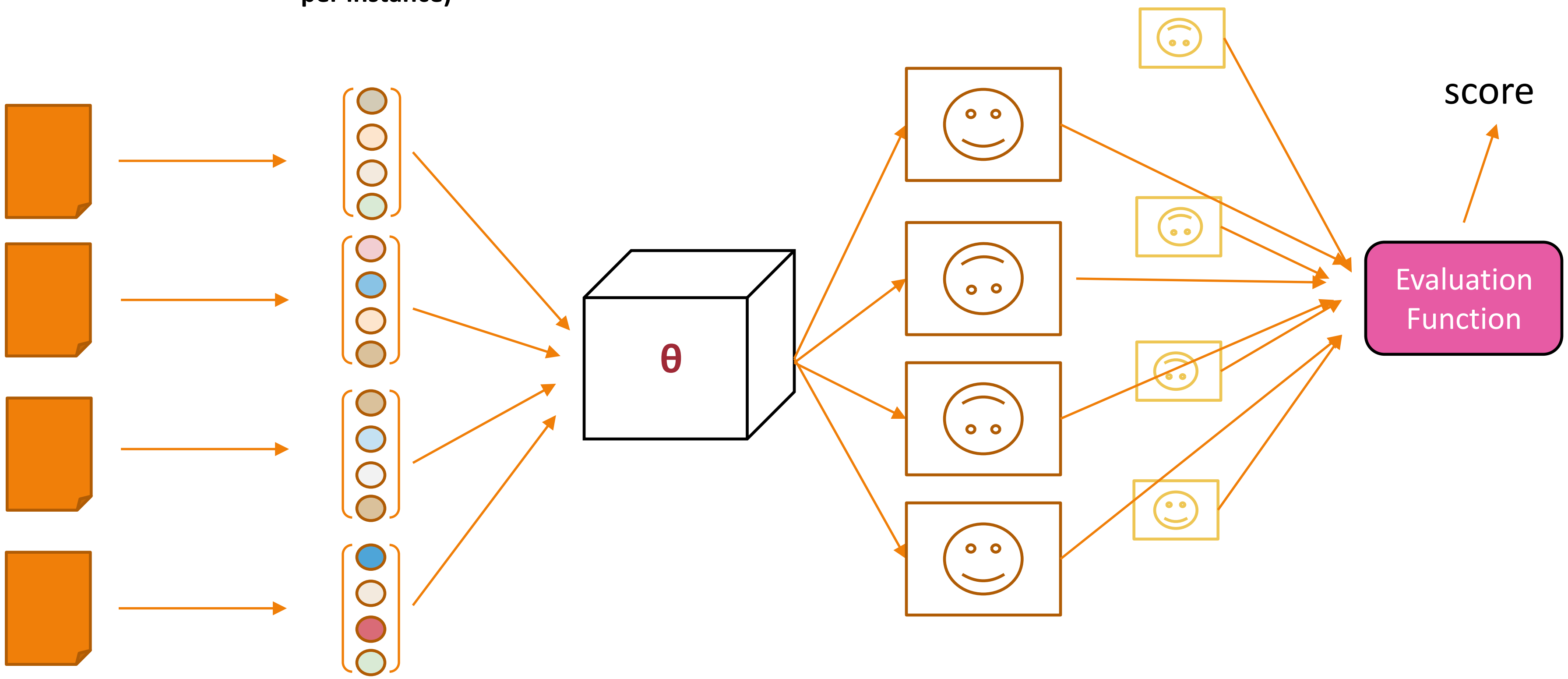
instances

features: **K-dimensional vector representations** (one per instance)

ML model:
- take in featurized input
- output scores/labels
- contains weights θ

"Gold" (correct) labels

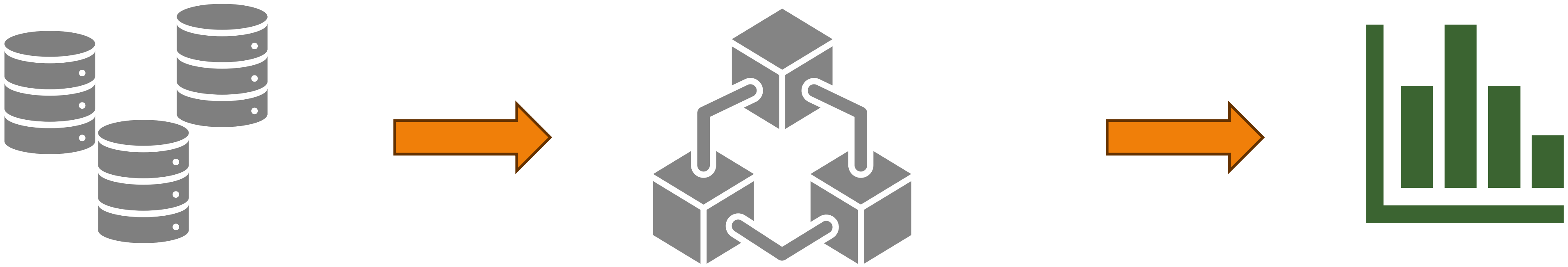
Evaluation Function



Review: Classification Types (Terminology)

Name	Number of Tasks (Domains) Labels are Associated with	# Label Types	Example
(Binary) Classification	1	2	Sentiment: Choose one of {positive or negative}
Multi-class Classification	1	> 2	Part-of-speech: Choose one of {Noun, Verb, Det, Prep, …}
Multi-label Classification	1	> 2	Sentiment: Choose multiple of {positive, angry, sad, excited, …}
Multi-task Classification	> 1	Per task: 2 or > 2 (can apply to binary or multi-class)	Task 1: part-of-speech Task 2: named entity tagging … ---------------------- Task 1: document labeling Task 2: sentiment

How do we learn models?



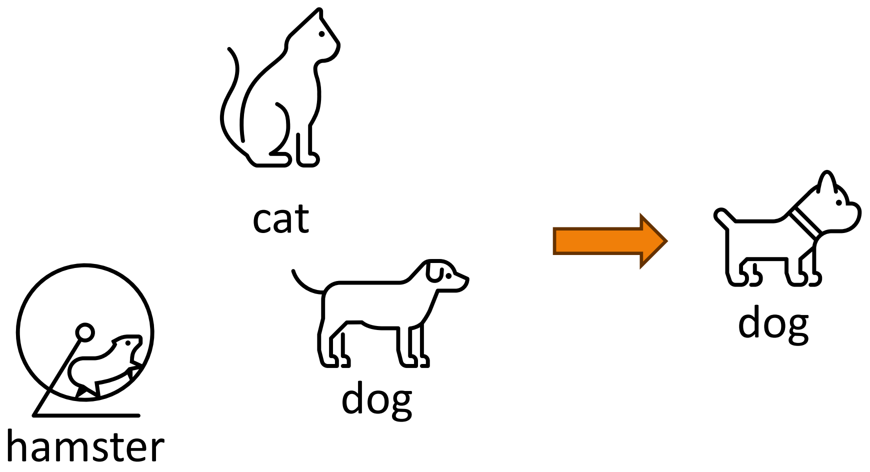
Take past experiences
(lots of data; corpus)

Find patterns
(the ML algorithm)

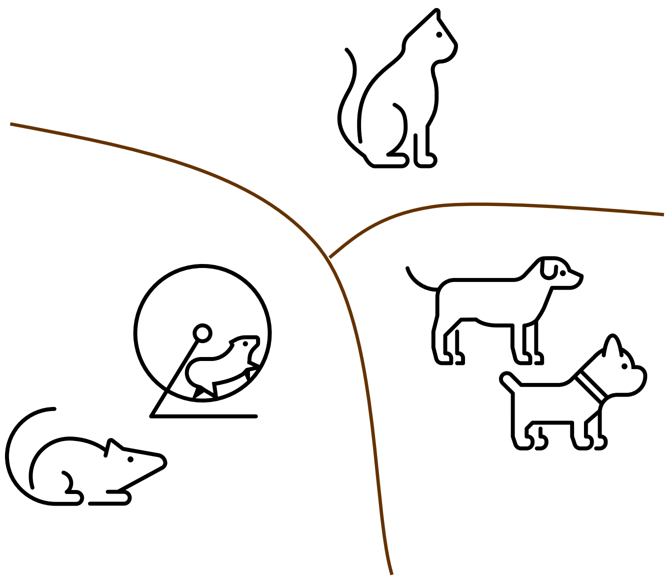
Use on new experiences
(save & test the model)

Types of Learning

SUPERVISED LEARNING



UNSUPERVISED LEARNING



Types of Learning

SUPERVISED LEARNING

Data has feedback (labels)

Data consists of input-output pairs

Learn mapping from input to output

Examples:

- Dataset classification
- How likely is it that this person will get into a car accident?

UNSUPERVISED LEARNING

No explicit feedback in data

Learn patterns directly from data

Examples:

- Clustering
- Do these people fall under multiple groups?

What are some other examples of these?

SUPERVISED LEARNING

- Machine translation
- Object segmentation (vision)
- Document classification

UNSUPERVISED LEARNING

- Clustering (e.g., topic modeling)
- Language modeling

The Machine Learning Framework

$$y = f(\mathbf{x})$$

output

prediction function

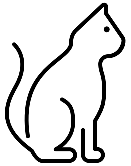
feature(s) of input

Training: given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

Testing: apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

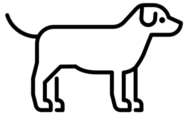
Slide credit: Svetlana Lazebnik

How do we learn models?



cat

hamster



dog

[position of lines]
[loops and dots]
[any other information relevant to our problem]

P(hamster|[line in this position],…)

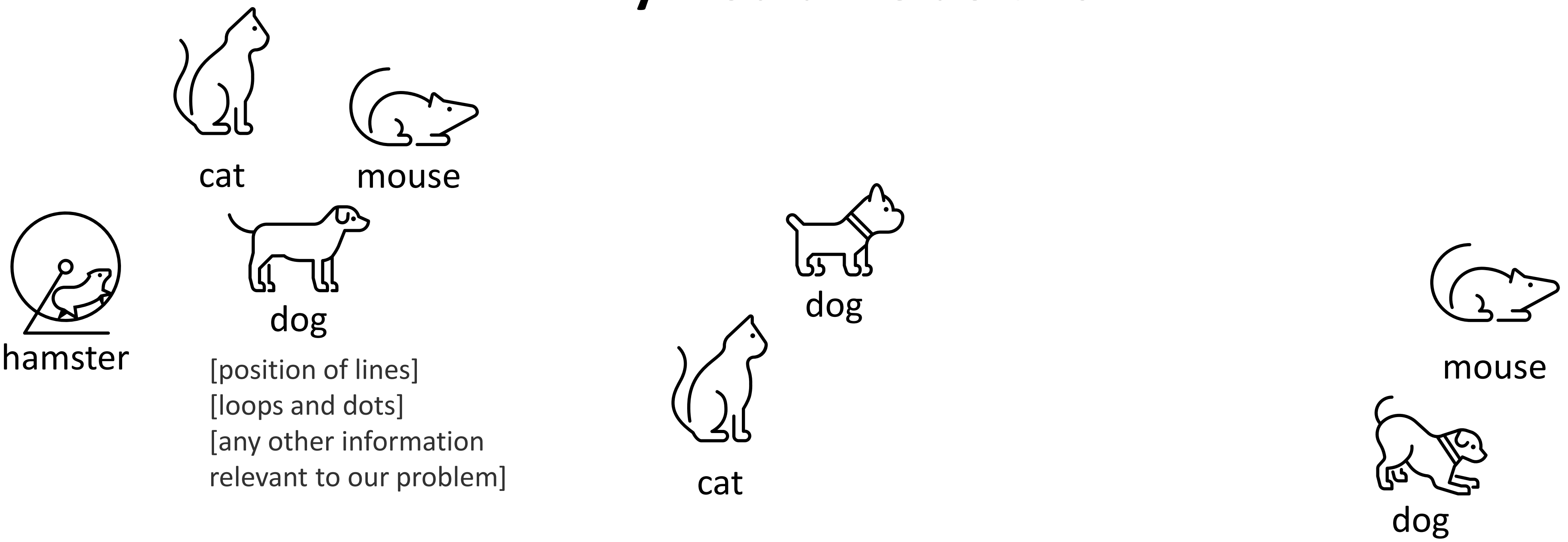
P(dog|[line in this other position],…)

Have data with features extracted (and possibly labels)

Learn associations between features and labels

Dividing up data for Training

Why would we do this?



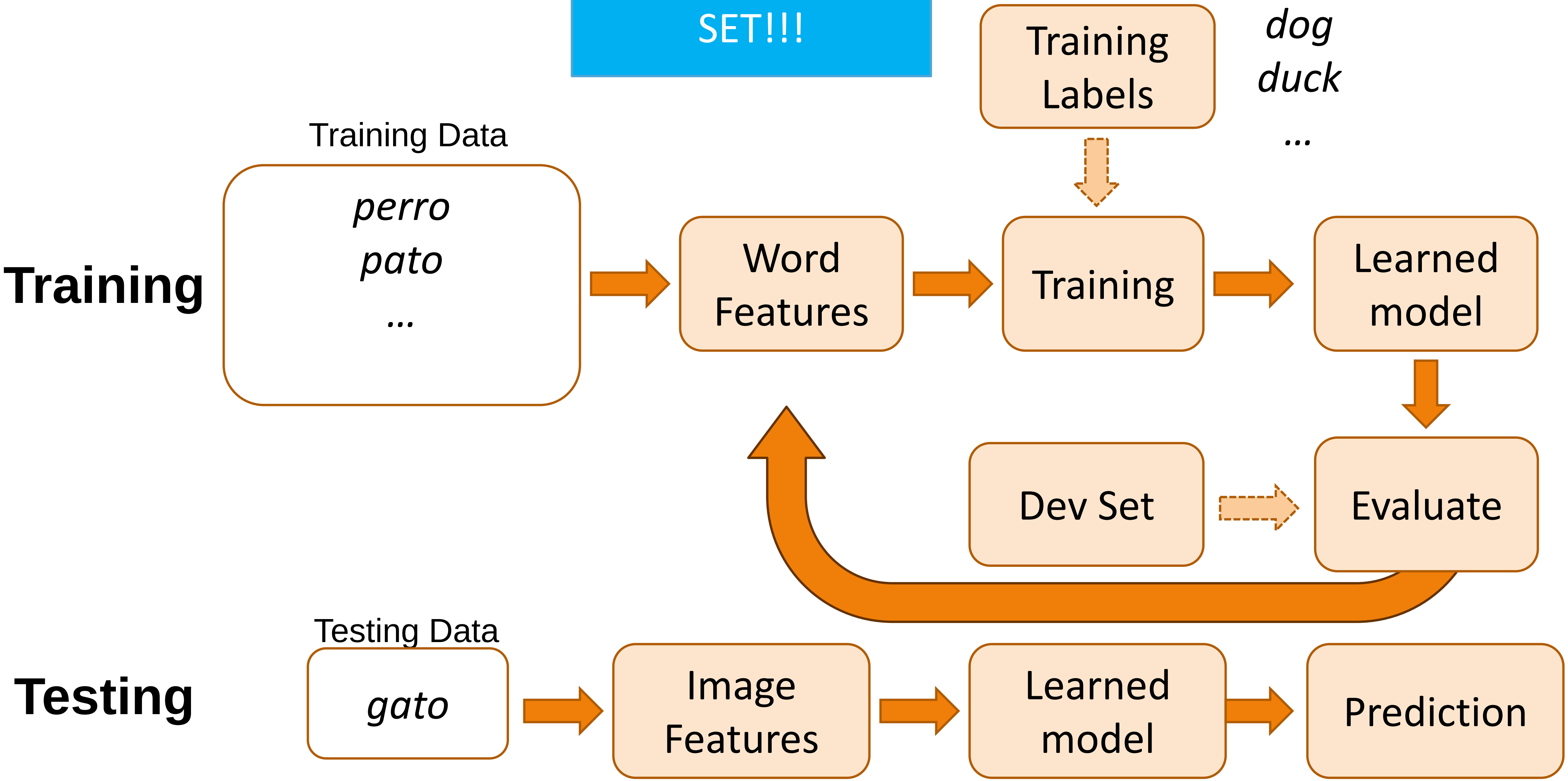
Training ~80%

Validation/Development ~10%

Testing ~10%

Steps

DO NOT ITERATE ON THE TESTING SET!!!



Types of models

CLASSIFICATION

Model outputs comes from a finite set of values

Discrete result

Examples:

- What type of animal is this a picture of?
- Predicting the weather (sunny, cloudy, or rainy?)
- Ranking: Is this result *better* than this result?

REGRESSION

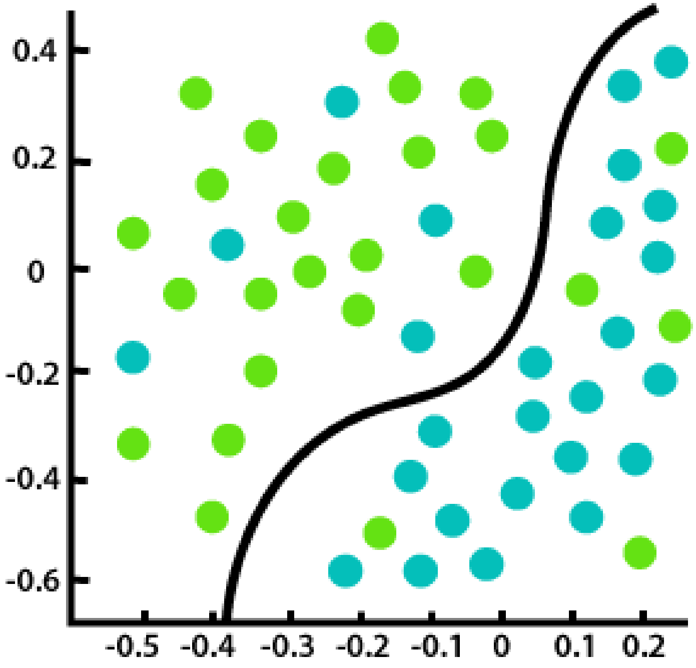
Model outputs are continuous values

Continuous result

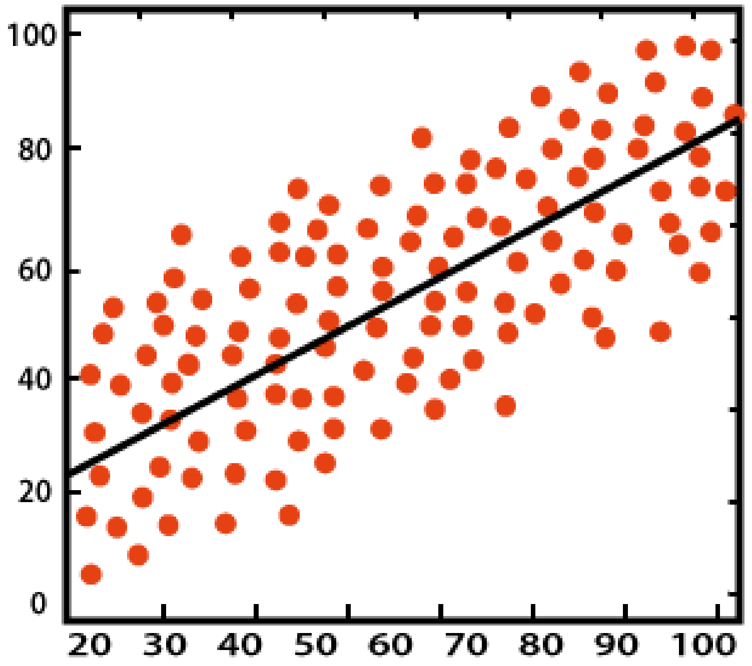
Examples:

- How far will I move if I drive my motors at this speed for 1 second?
- Predicting the weather (temperature)
- Ranking: *how good* is this result?

Types of models



Classification



Regression

https://medium.com/unpackai/classification-regression-in-machine-learning-7cf3b13b0b09

What are some other examples of these?

CLASSIFICATION

Tone tagging

Sentiment classification

Named entity recognition

REGRESSION

Quantity/scale of how much it sounds like a specific author

Numerical sentiment value

Political “score” from document

Likelihoods

Predicted Goodreads score

Types of Algorithms

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Source unknown

Central Question: How Well Are We Doing?

Classification

- Precision, Recall, F1
- Accuracy
- Log-loss
- ROC-AUC
- …

Regression

- (Root) Mean Square Error
- Mean Absolute Error
- …

Clustering

- Mutual Information
- V-score
- …

*the **task**: what kind of problem are you solving?*

This does not have to be the same thing as the loss function you optimize

Implementation: How To

1. scikit-learn: sklearn.metrics
   - very stable

2. huggingface evaluate module
   - community input
   - sometimes are based on sklearn

3. implement your own

Classification Evaluation:
the 2-by-2 contingency table

Assumption 1: There are two classes/labels

● ○

Assumption 2: ● is the "positive" label

Assumption 3: Given X, our classifier produces a score for each possible label

p(●|X) vs. p(○|X)

Examining Assumption 3

Given X, our classifier produces a score for each possible label

p(|X) vs. p(|X)

Normally (*but this can be adjusted!)

$$\text{best label} = \arg\max_{\text{label}} P(\text{label}|\text{example})$$

Example of `argmax`

Electronic alerts have been used to assist the authorities in moments of chaos and potential danger: after the Boston bombing in 2013, when the Boston suspects were still at large, and last month in Los Angeles, during an active shooter scare at the airport.

POLITICS	.05
TERRORISM	.48
SPORTS	.0001
TECH	.39
HEALTH	.0001
FINANCE	.0002
…	

Source: http://www.nytimes.com/2016/09/20/nyregion/cellphone-alerts-used-in-search-of-manhattan-bombing-suspect.html

Example of `argmax`

Electronic alerts have been used to assist the authorities in moments of chaos and potential danger: after the Boston bombing in 2013, when the Boston suspects were still at large, and last month in Los Angeles, during an active shooter scare at the airport.

POLITICS	.05
TERRORISM	**.48**
SPORTS	.0001
TECH	.39
HEALTH	.0001
FINANCE	.0002
…	

Source: http://www.nytimes.com/2016/09/20/nyregion/cellphone-alerts-used-in-search-of-manhattan-bombing-suspect.html

Classification Evaluation: the 2-by-2 contingency table

	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")		
Not selected/ not guessed ("○")		

Classification Evaluation: the 2-by-2 contingency table

	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP) ● *Actual* ● *Guessed*	
Not selected/ not guessed ("○")		

Classification Evaluation: the 2-by-2 contingency table

	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive ● *Actual* (TP) ● *Guessed*	False Positive ○ *Actual* (FP) ● *Guessed*
Not selected/ not guessed ("○")		

Classification Evaluation: the 2-by-2 contingency table

	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP) ● *Actual* ● *Guessed*	False Positive (FP) ○ *Actual* ● *Guessed*
Not selected/ not guessed ("○")	False Negative (FN) ● *Actual* ○ *Guessed*	

Classification Evaluation: the 2-by-2 contingency table

	What is the actual label?	
What label does our system predict? (↓)	**Actual Target Class ("●")**	**Not Target Class ("○")**
Selected/ Guessed ("●")	True Positive (TP) ● Actual ● Guessed	False Positive (FP) ○ Actual ● Guessed
Not selected/ not guessed ("○")	False Negative (FN) ● Actual ○ Guessed	True Negative (TN) ○ Actual ○ Guessed

Classification Evaluation: the 2-by-2 contingency table

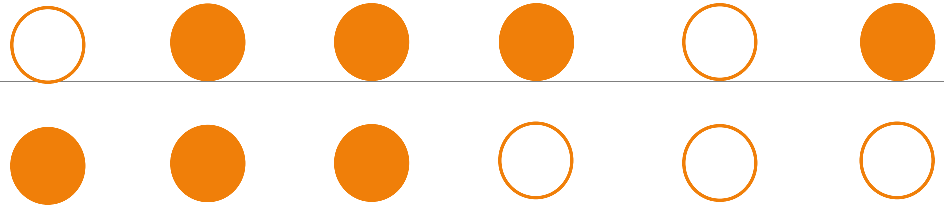
	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP) ● *Actual* ● *Guessed*	False Positive (FP) ○ *Actual* ● *Guessed*
Not selected/ not guessed ("○")	False Negative (FN) ● *Actual* ○ *Guessed*	True Negative (TN) ○ *Actual* ○ *Guessed*

Construct this table by *counting* the number of TPs, FPs, FNs, TNs

Contingency Table Example

Predicted:

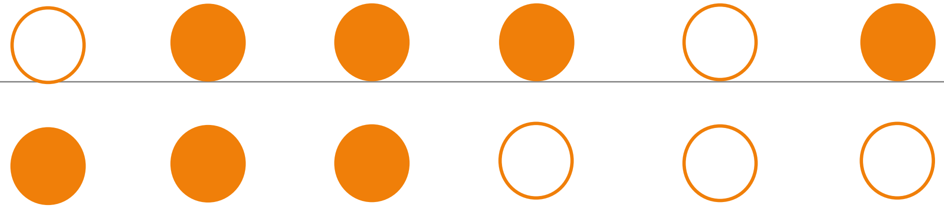
Actual:



Contingency Table Example

Predicted:

Actual:

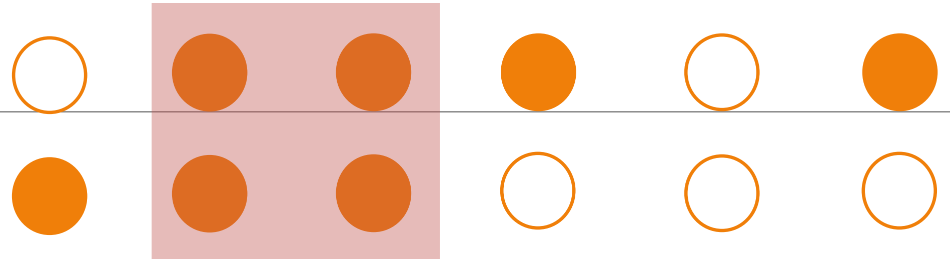


	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP)	False Positive (FP)
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)

Contingency Table Example

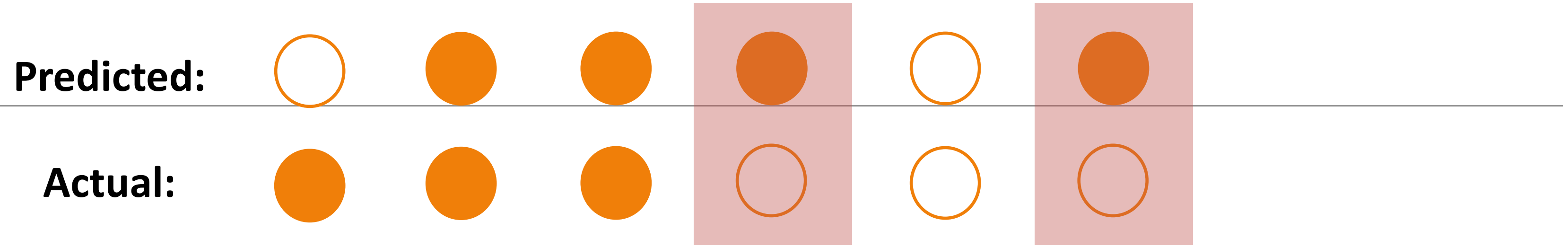
Predicted:

Actual:



	What is the actual label?	
What label does our system predict? (↓)	**Actual Target Class ("●")**	**Not Target Class ("○")**
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP)
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)

Contingency Table Example

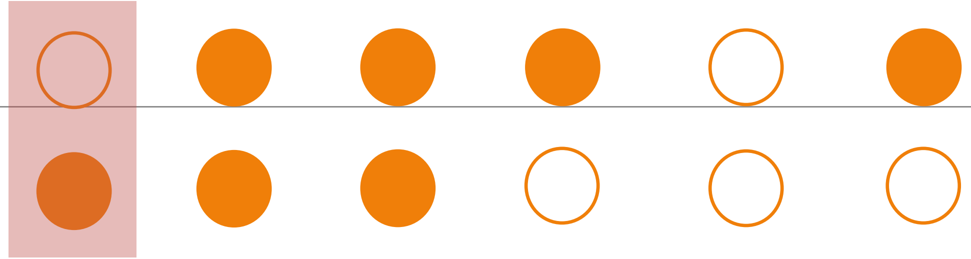


	What is the actual label?	
What label does our system predict? (↓)	**Actual Target Class ("●")**	**Not Target Class ("○")**
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)

Contingency Table Example

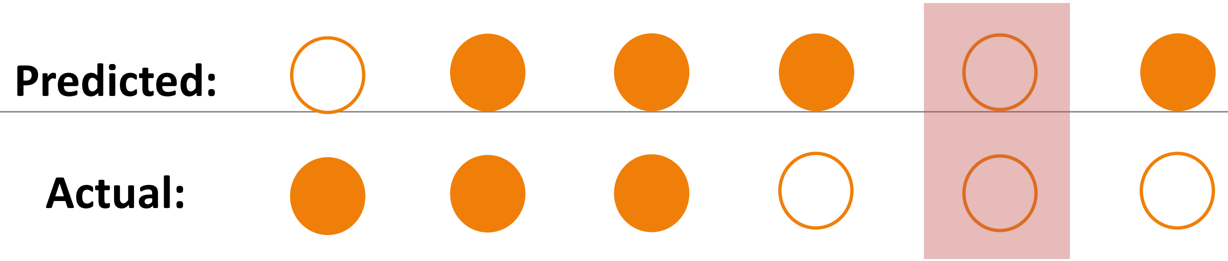
Predicted:

Actual:



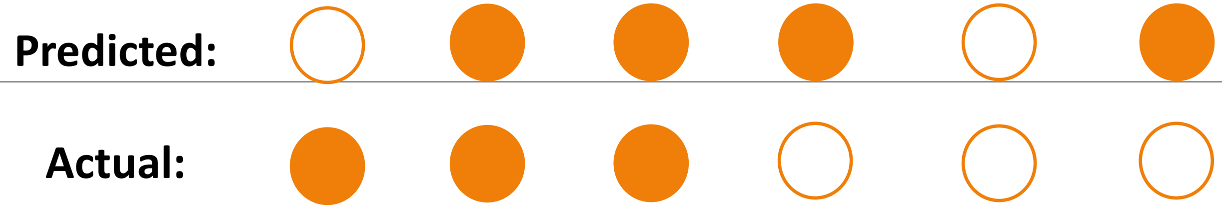
	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN) = 1	True Negative (TN)

Contingency Table Example



	What is the actual label?	
What label does our system predict? (↓)	**Actual Target Class ("●")**	**Not Target Class ("○")**
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN) = 1	True Negative (TN) = 1

Contingency Table Example



	What is the actual label?	
What label does our system predict? (↓)	**Actual Target Class ("●")**	**Not Target Class ("○")**
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN) = 1	True Negative (TN) = 1

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

Precision: % of selected items that are correct

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

Precision: % of selected items that are correct

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall: % of correct items that are selected

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

	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

Precision: % of selected items that are correct

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall: % of correct items that are selected

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

Min: 0 ☹
Max: 1 😲

	Actually Target	**Actually Not Target**
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

The Importance of “Polarity” in Binary Classification

Fundamentally: what are you trying to “identify” in your classification?

Are you trying to find ● or ○?

The Importance of “Polarity” in Binary Classification

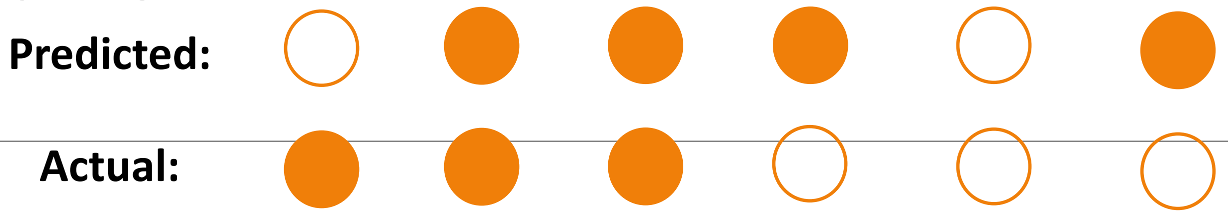
		Correct Value	
		●	○
Guessed Value	●	#	#
	○	#	#

Try to find ● : Where do the TP / FP / FN / FN values go?

The Importance of “Polarity” in Binary Classification

		Correct Value	
Guessed Value		TP	FP
		FN	TN

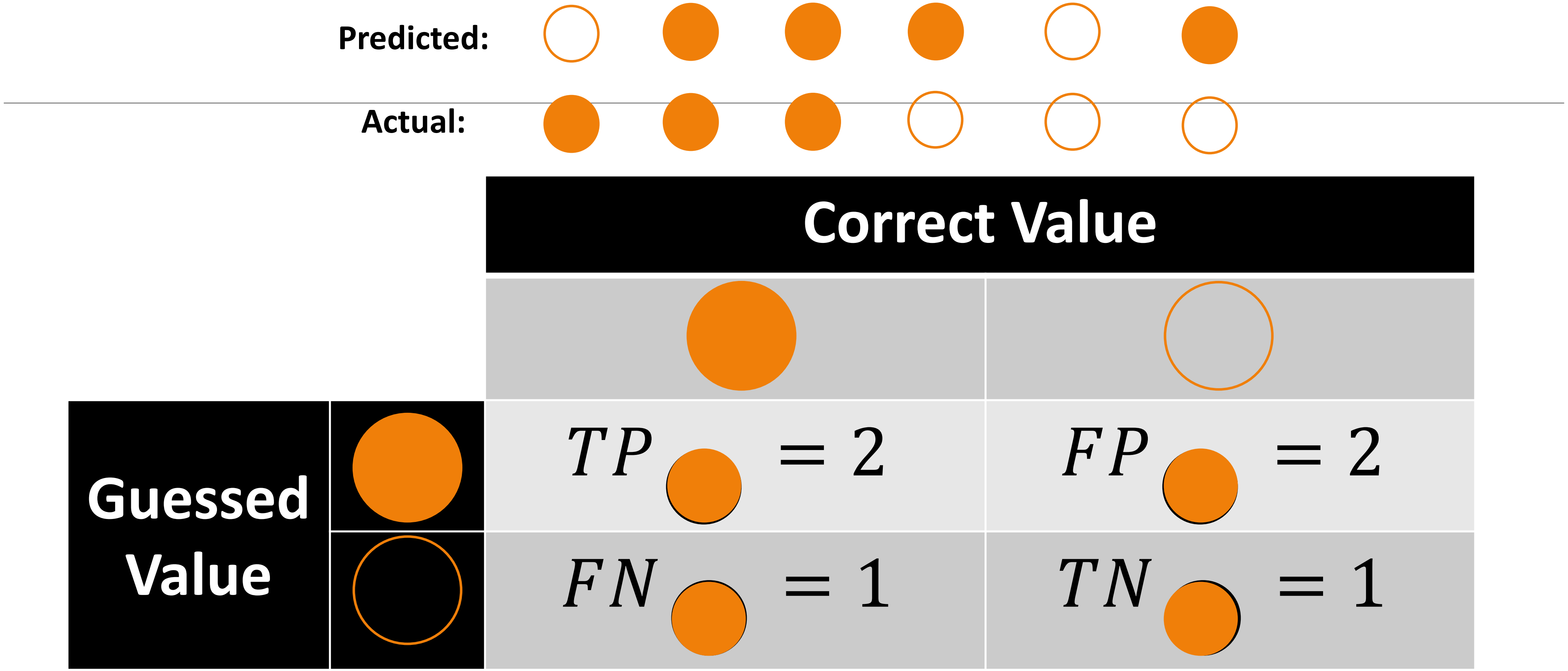
The Importance of “Polarity” in Binary Classification



		Correct Value	
Guessed Value		$TP = 2$	$FP = 2$
		$FN = 1$	$TN = 1$

What are the accuracy, recall, and precision values?

The Importance of "Polarity" in Binary Classification



What are the accuracy, recall, and precision values?

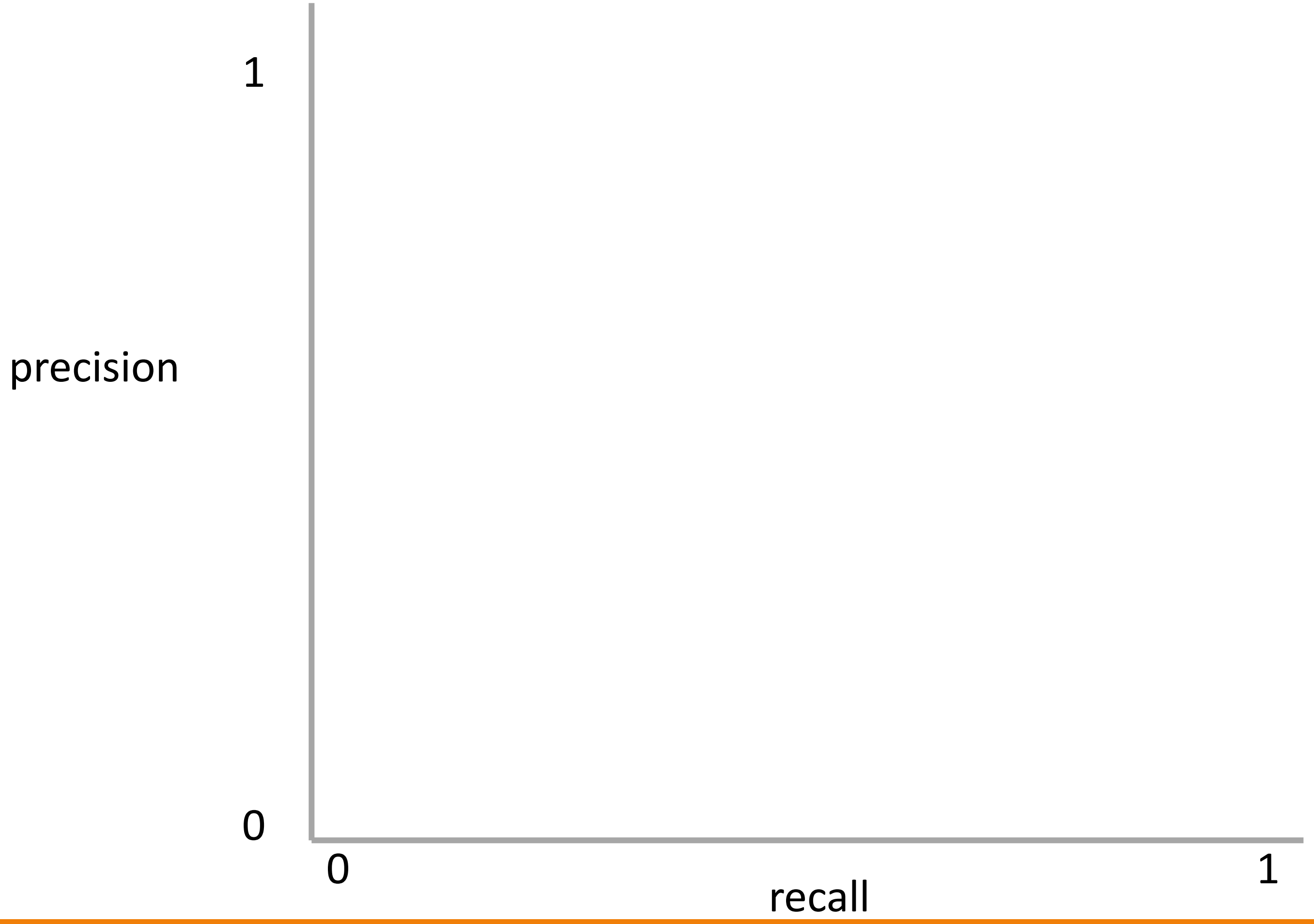
Accuracy: 50%
Recall: 66.67%
Precision: 50%

The Importance of “Polarity” in Binary Classification

		Correct Value	
		●	○
Guessed Value	●	$TP_{\bullet} = TN_{\circ}$	$FP_{\bullet} = FN_{\circ}$
	○	$FN_{\bullet} = FP_{\circ}$	$TN_{\bullet} = TP_{\circ}$

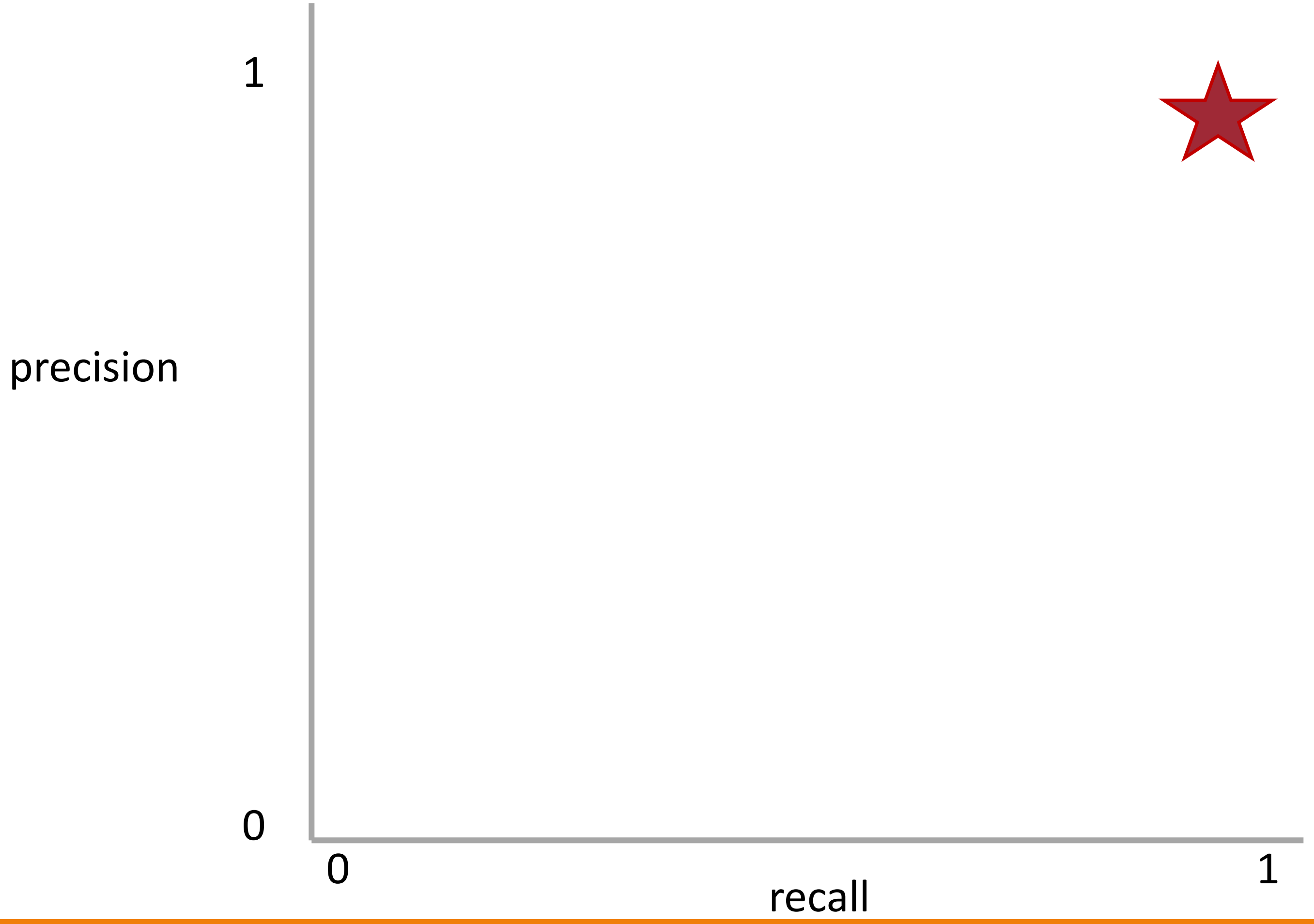
Remember: what are you trying to “identify” in your classification?

Precision and Recall Present a Tradeoff



Q: Where do you want your ideal model ?

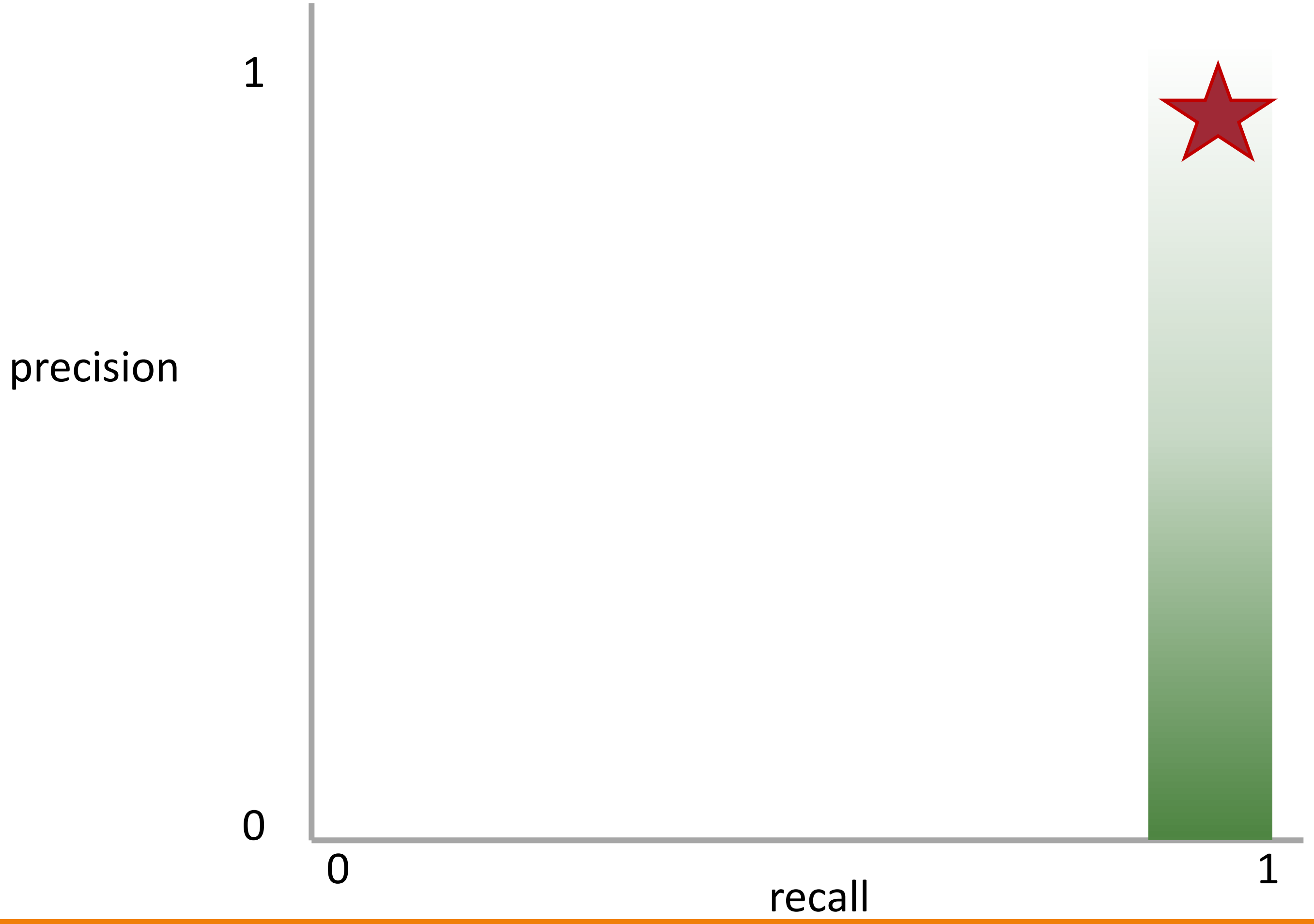
Precision and Recall Present a Tradeoff



Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Precision and Recall Present a Tradeoff

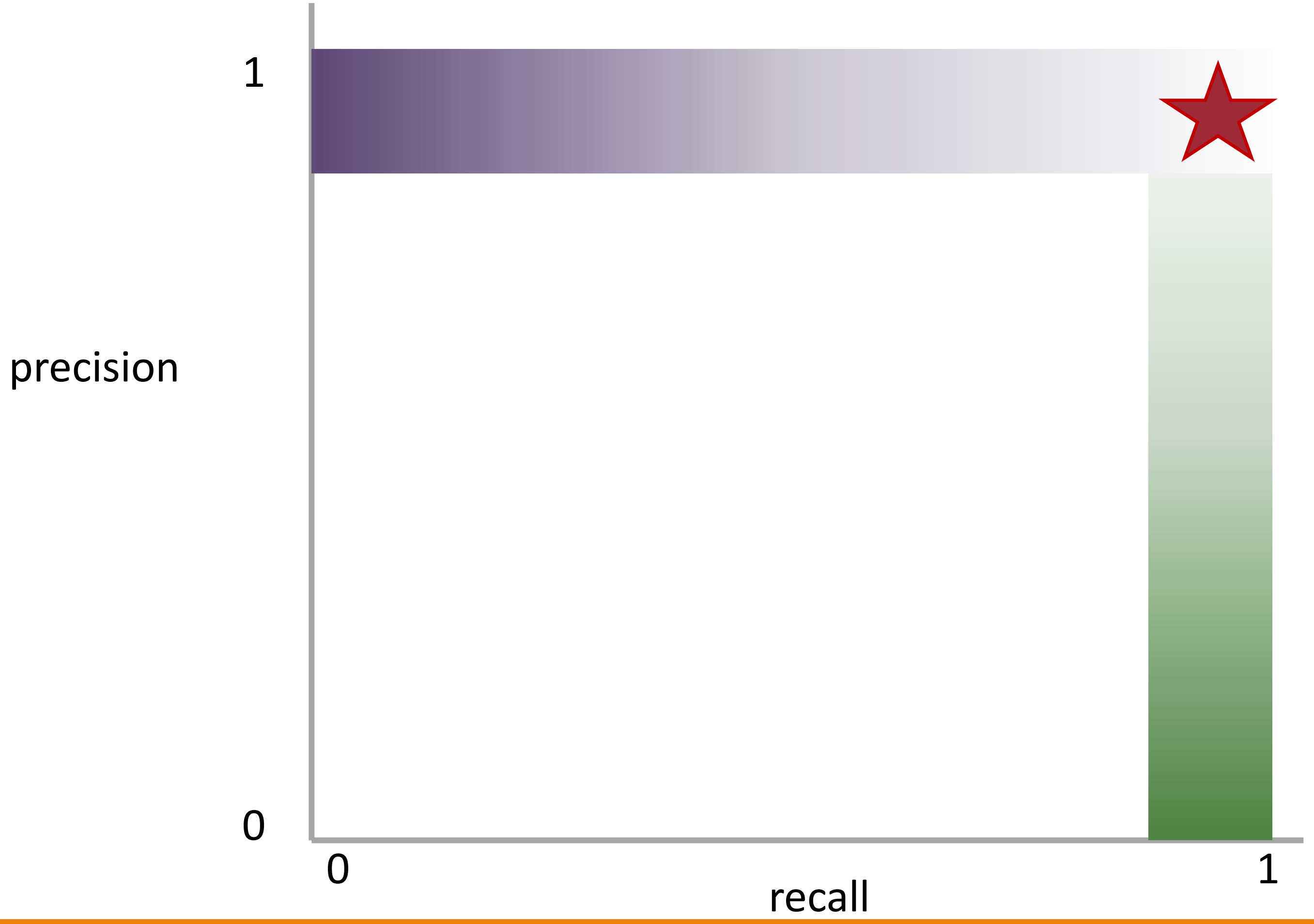


Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Precision and Recall Present a Tradeoff

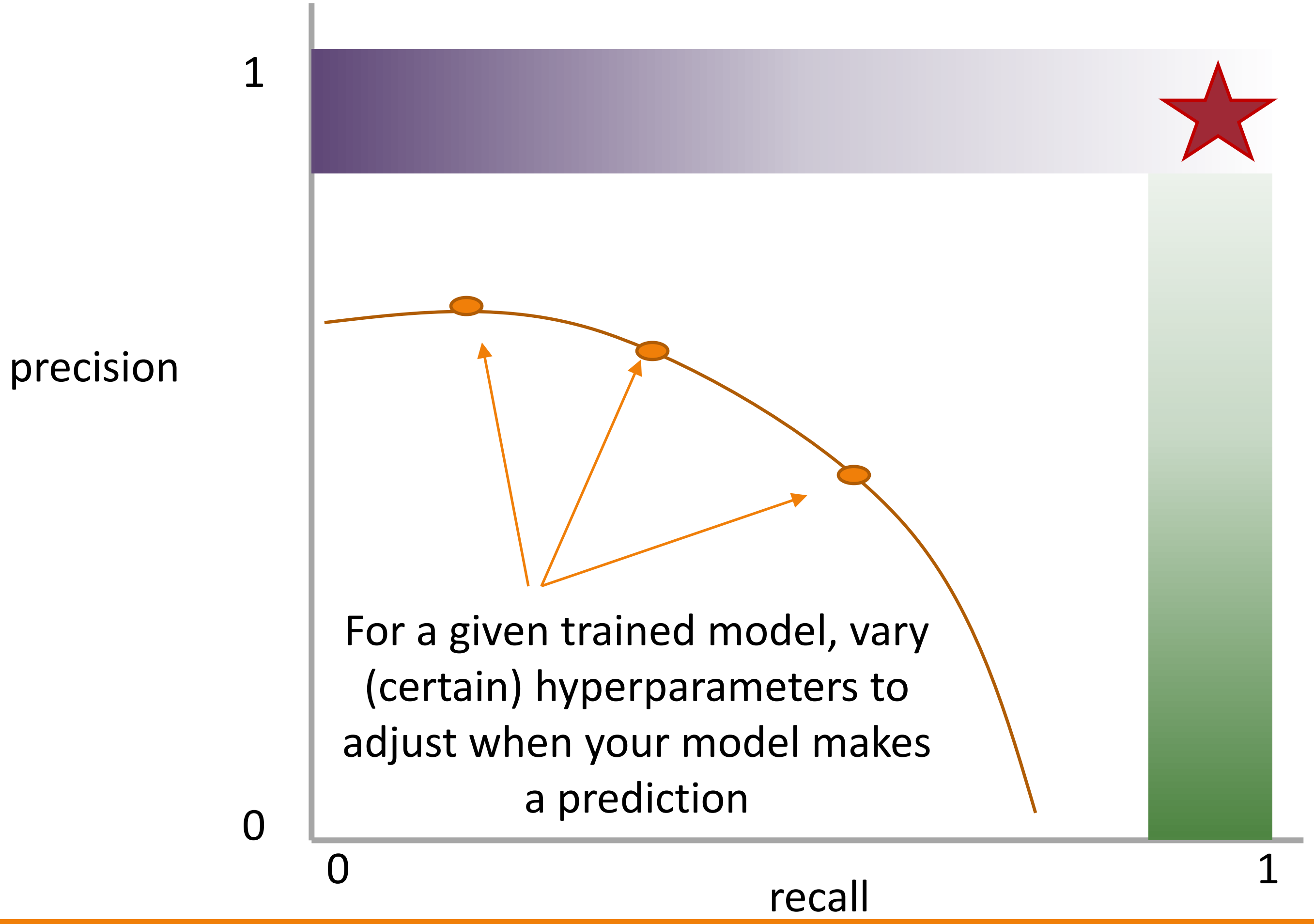


Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Precision and Recall Present a Tradeoff



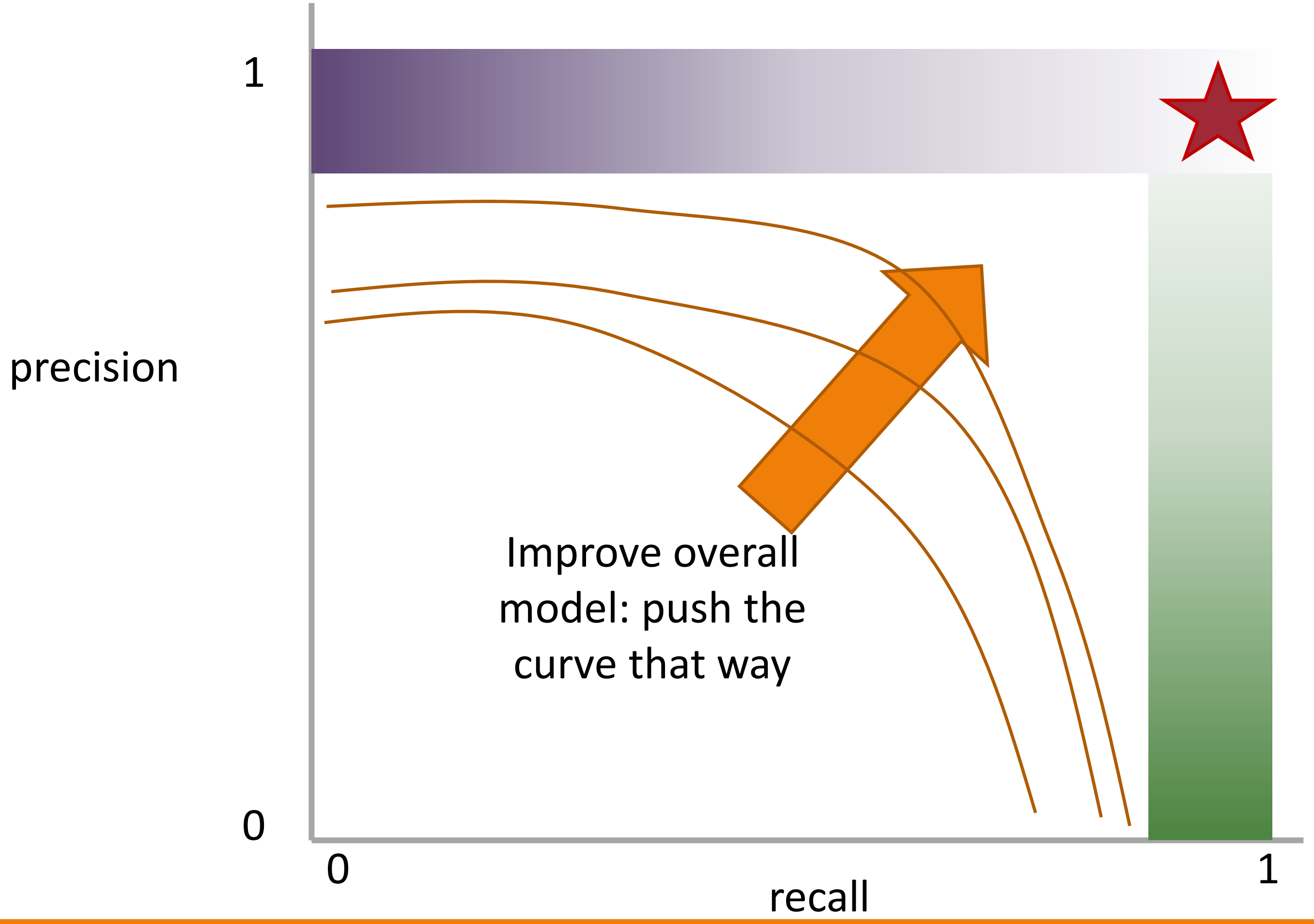
Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Idea: measure the tradeoff between precision and recall

Precision and Recall Present a Tradeoff



Q: Where do you want your ideal model?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Idea: measure the tradeoff between precision and recall