NLP Tasks 3 + ML Evaluation

Instructor: Lara J. Martin (she/they)

TA: Omkar Kulkarni (he)

https://laramartin.net/NLP-class/

Slides modified from Dr. Frank Ferraro & Dr. Jason Eisner

Learning Objectives

Formalize NLP Tasks at a high-level:

- What are the input/output for a particular task?
- What might the features be?
- What types of applications could the task be used for?

Enumerate different input scopes of tasks when thought of as classification

Fill out a contingency table and calculate accuracy, precision, and recall

Develop an intuition about precision & recall

Extend P/R to multi-class problems

Identify when you might want certain evaluation metrics over others

Text Annotation Tasks ("Classification" Tasks)

1. Classify the entire document ("text categorization")
2. Classify word tokens individually
3. Classify word tokens in a sequence
4. Identify phrases ("chunking")
5. Syntactic annotation (parsing)
6. Semantic annotation
7. Text generation

Slide courtesy Jason Eisner, with mild edits

Example: Finding Named Entities

Named entity recognition (NER)

Identify proper names in texts, and classification into a set of predefined categories of interest

- Person names
- Organizations (companies, government organisations, committees, etc.)
- Locations (cities, countries, rivers, etc.)
- Date and time expressions
- Measures (percent, money, weight, etc.),
- email addresses, web addresses, street addresses, etc.
- Domain-specific: names of drugs, medical conditions,
- names of ships, bibliographic references etc.

Cunningham and Bontcheva (2003, RANLP Tutorial)

NE Types

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Slide courtesy Jim Martin

Named Entity Recognition

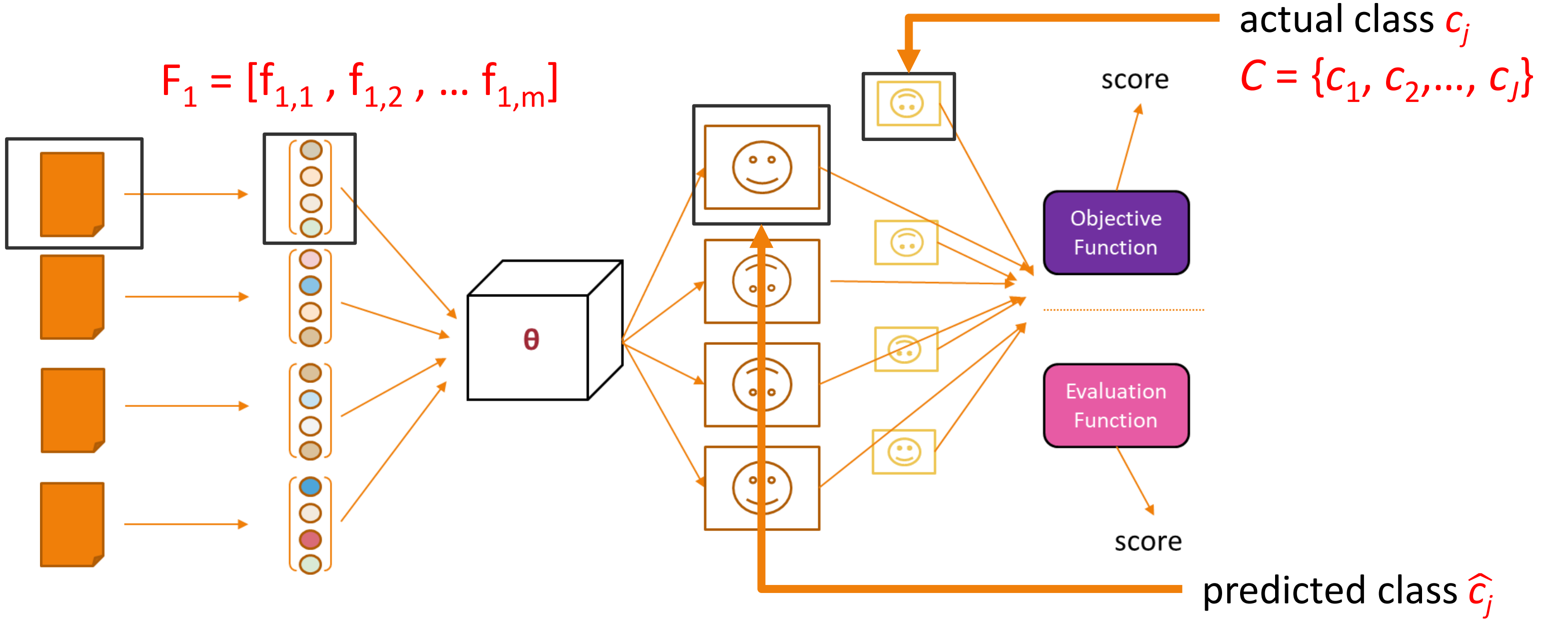
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

What are the input/output?
What are the features?
What types of applications?

Slide courtesy Jim Martin

Chunking Input/Output

features F_1 extracted from phrase p_1 and its surrounding context



p(class | phrase in context)

Chunking Tasks

Named entity recognition (NER)

Information extraction (IE)

Identifying idioms

…

Other examples?

What are the input/output?
What are the features?
What types of applications?

I've been referring to these as applications to get you thinking broadly, but sometimes they are tasks

Information Extraction as Task

As a task: **Filling slots in a database from sub-segments of text.**

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Slide from Chris Brew, adapted from slide by William Cohen

Example applications for IE

Classified ads

Restaurant reviews

Bibliographic citations

Appointment emails

Legal opinions

Papers describing clinical medical studies

What’s the difference between a task and an application?

Task vs Application

Task: A common problem that a community of people work on. They usually have their own benchmarks, usual ways of evaluating, etc.

- Tasks that aren’t as “popular” might not be as established

Application: Doesn’t really have a definition in NLP

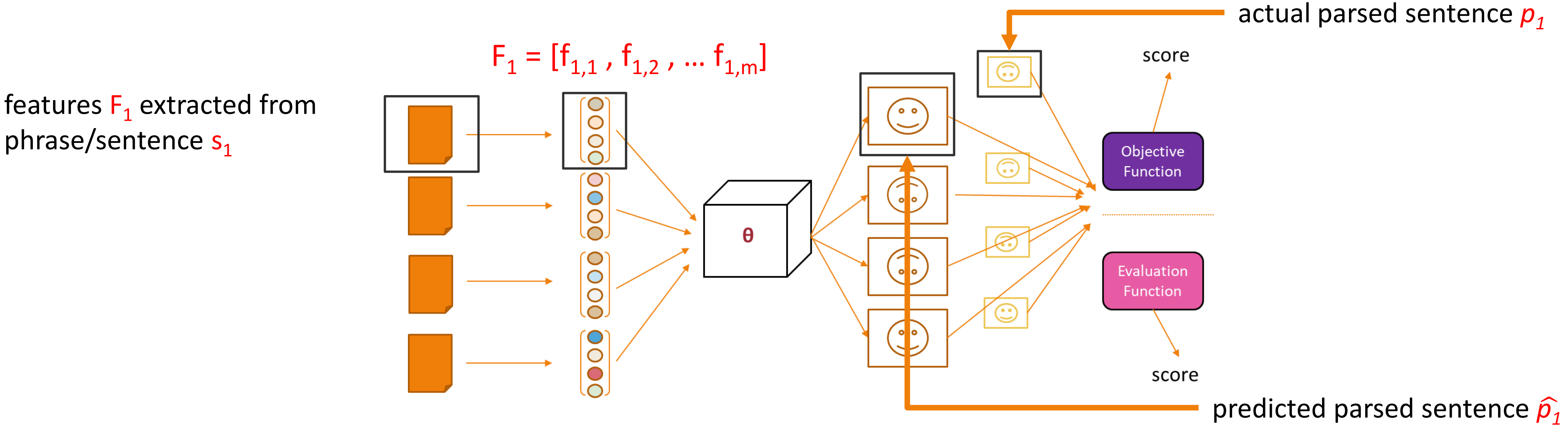
- It can refer to when a task is being used for a program/piece of software (like in the previous slide)

Text Annotation Tasks ("Classification" Tasks)

1. Classify the entire document ("text categorization")
2. Classify word tokens individually
3. Classify word tokens in a sequence
4. Identify phrases ("chunking")
5. Syntactic annotation (syntax parsing)
6. Semantic annotation
7. Text generation

Slide courtesy Jason Eisner, with mild edits

Syntax Parsing



Context Free Grammar

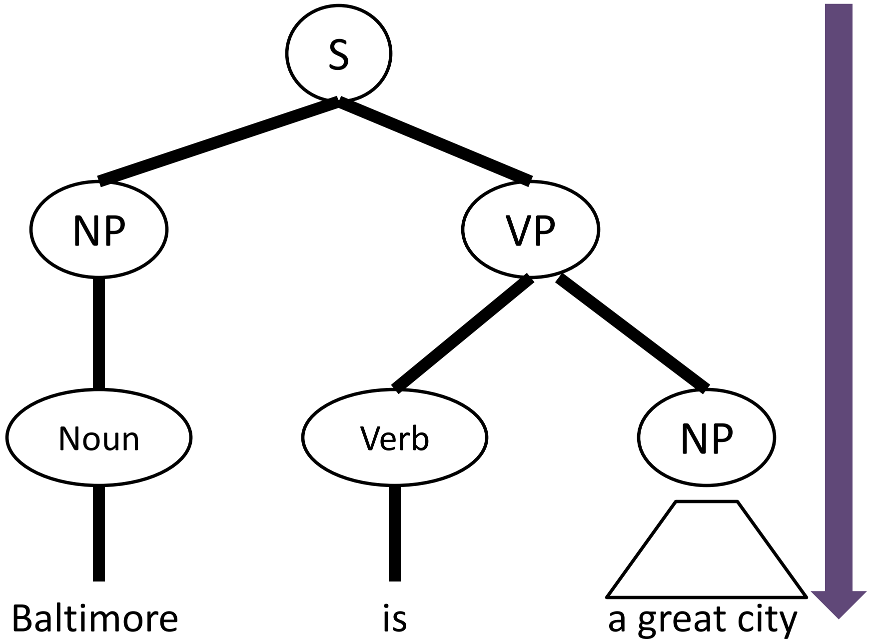
S → NP VP
NP → Det Noun
NP → Noun
NP → Det AdjP
NP → NP PP

PP → P NP
AdjP → Adj Noun
VP → V NP
Noun → Baltimore
…

Set of rewrite rules, comprised of terminals and non-terminals

Generate from a Context Free Grammar

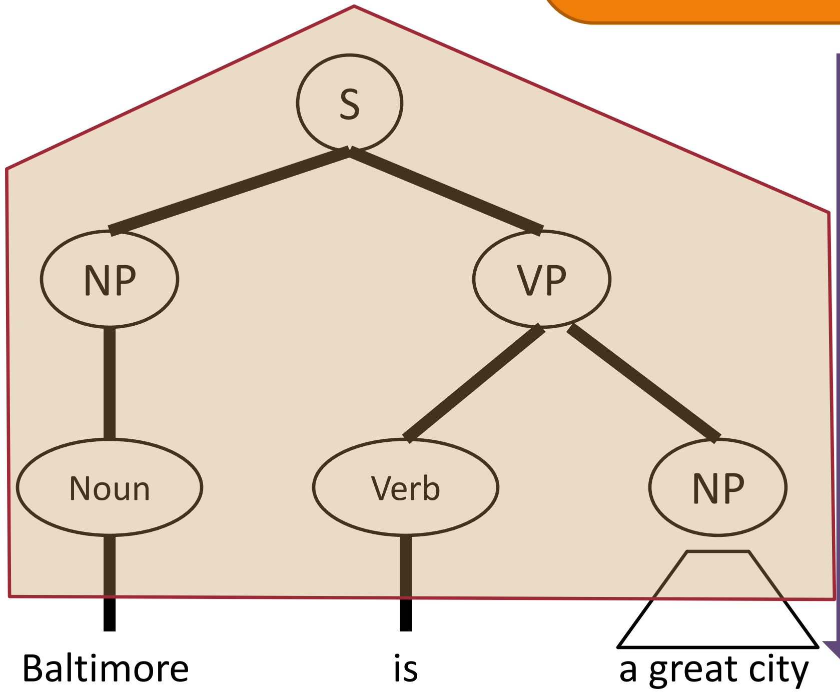
S → NP VP	PP → P NP
NP → Det Noun	AdjP → Adj Noun
NP → Noun	VP → V NP
NP → Det AdjP	Noun → Baltimore
NP → NP PP	…



Baltimore is a great city

Assign Structure (Parse) with a Context Free Grammar

S → NP VP
NP → Det Noun
NP → Noun
NP → Det AdjP
NP → NP PP
PP → P NP
AdjP → Adj Noun
VP → V NP
Noun → Baltimore
…



Baltimore is a great city

$[_{S}\ [_{NP}\ [_{Noun}$ Baltimore]] $[_{VP}\ [_{Verb}$ is] $[_{NP}$ a great city]]]

bracket notation

(S (NP (Noun Baltimore))
(VP (V is)
(NP a great city)))

S-expression

Why is it useful?

https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

The old man the boat .

Garden Path Sentences

https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

The old man the boat .

https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

The rat the cat the dog chased killed ate the malt.

https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

The rat *that* the cat the dog chased killed ate the malt.

https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

The rat *that* the cat *that* the dog chased killed ate the malt.


https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

The rat *that* the cat *that* the dog chased killed ate the malt.

Garden Path Sentences


https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

The rat *that* the cat *that* the dog chased killed ate the malt.


https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

The rat *that* the cat *that* the dog chased killed ate the malt.


https://www.housebeautiful.com/uk/garden/g45582875/garden-path-ideas/

Garden Path Sentences

[The rat [the cat [the dog chased] killed] ate the malt].

Language can have recursive patterns

Syntactic parsing can help identify those

Text Annotation Tasks (“Classification” Tasks)

1. Classify the entire document (“text categorization”)
2. Classify word tokens individually
3. Classify word tokens in a sequence
4. Identify phrases (“chunking”)
5. Syntactic annotation (syntax parsing)
6. Semantic annotation
7. Text generation

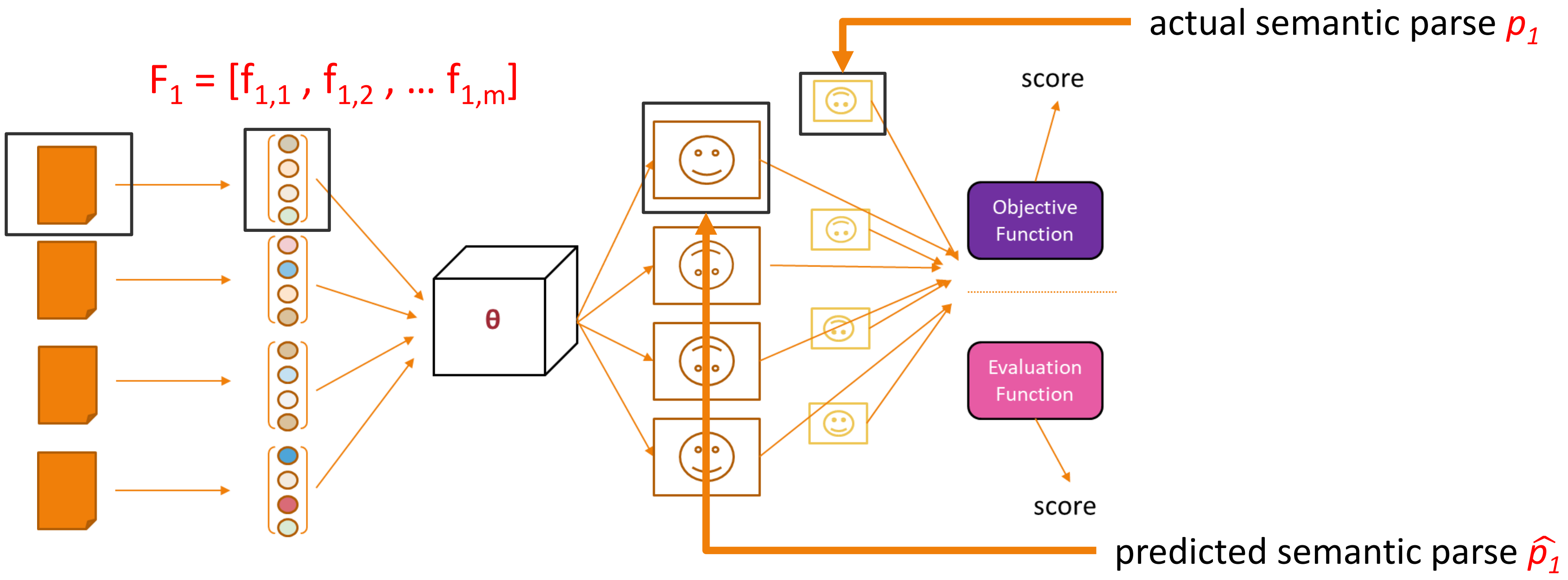
Slide courtesy Jason Eisner, with mild edits

Semantic Parsing

Task: Semantic role labeling (SRL)

What is a semantic parse?

features F_1 extracted from phrase/sentence s_1 and its surrounding context



Semantic Role Labeling (SRL)

For each predicate (e.g., verb)
1. find its arguments (e.g., NPs)
2. determine their **semantic roles**

John drove Mary from Austin to Dallas in his Toyota Prius.

The hammer broke the window.

- agent: Actor of an action
- patient: Entity affected by the action
- source: Origin of the affected entity
- destination: Destination of the affected entity
- instrument: Tool used in performing action.
- beneficiary: Entity for whom action is performed

Slide thanks to Ray Mooney (modified)

Other Current Semantic Annotation Tasks (similar to SRL)

PropBank – coarse-grained roles of verbs

NomBank – similar, but for nouns

FrameNet – fine-grained roles of any word

TimeBank – temporal expressions

Slide courtesy Jason Eisner, with mild edits

What type of applications might this have?

FrameNet Example

REVENGE FRAME

Avenger
Offender (unexpressed in this sentence)
Injury
Injured Party (unexpressed in this sentence)
Punishment

We avenged the insult by setting fire to his village.

a word/phrase that triggers the REVENGE frame

Slide thanks to CJ Fillmore (modified)

Text Annotation Tasks (“Classification” Tasks)

1. Classify the entire document (“text categorization”)
2. Classify word tokens individually
3. Classify word tokens in a sequence
4. Identify phrases (“chunking”)
5. Syntactic annotation (syntax parsing)
6. Semantic annotation
7. Text generation

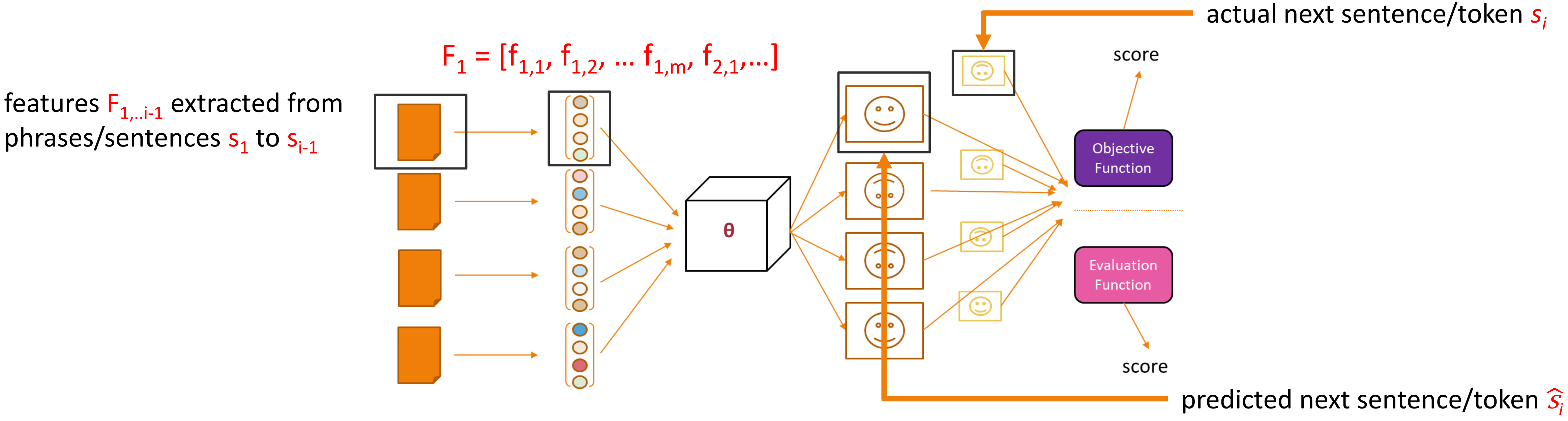
Slide courtesy Jason Eisner, with mild edits

Generation as a Classification Problem

Treating the word we want to generate as a label

What are the input/output?
What are the features?
What types of applications or tasks?

Text Generation Input/Output



p(word | history of words)

Text Generation Applications

Question answering (QA)

Speech recognition (ASR)

Machine translation (MT)

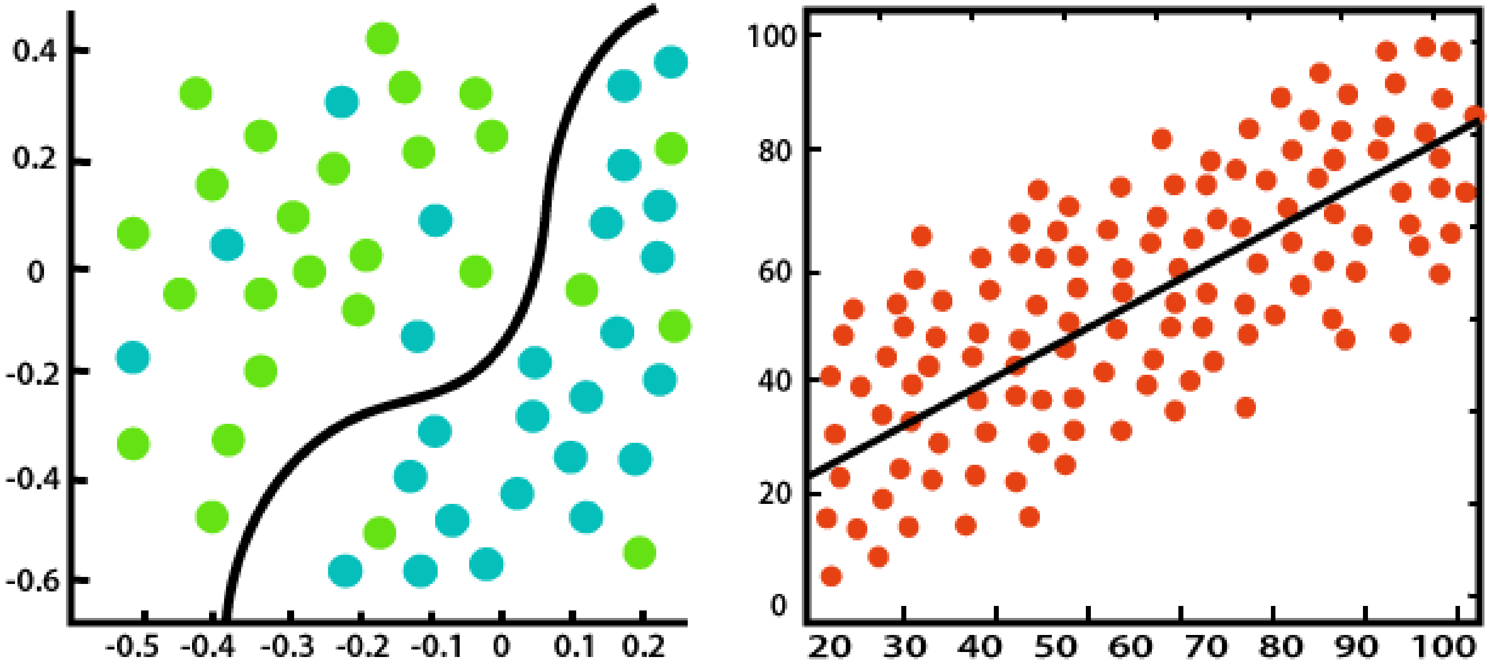
Summarization

Generating text from a structured representation

…

Calculating Evaluation Metrics for Classification

Review: Types of models



Classification Regression

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

Central Question: How Well Are We Doing?

Classification

Regression

Clustering

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

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

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

- Mutual Information
- V-score
- …

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

Some Classification Metrics

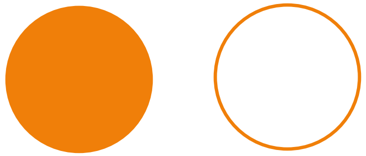
- Accuracy
- Precision
- Recall
- AUC (Area Under Curve)
- F1
- Confusion Matrix

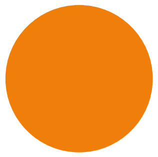
Implementation: How To

1. scikit-learn: sklearn.metrics
2. huggingface evaluate module
3. implement your own

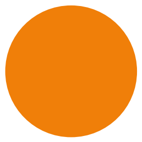
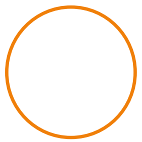
Classification Evaluation: the 2-by-2 contingency table

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`

Amazon acquired MGM in 2022, taking over a sprawling library that includes more than 4,000 feature films and 17,000 television shows. The tech behemoth also earned the rights to distribute all the Bond movies, but the new deal solidifies the company's oversight of Bond's big-screen future.

POLITICS	.002
MOVIES	.48
SPORTS	.0001
TECH	.39
HEALTH	.0001
FINANCE	.05
…	

Adapted from https://www.nbcnews.com/pop-culture/movies/amazon-taking-control-james-bond-movie-franchise-rcna192967

Example of `argmax`

Amazon acquired MGM in 2022, taking over a sprawling library that includes more than 4,000 feature films and 17,000 television shows. The tech behemoth also earned the rights to distribute all the Bond movies, but the new deal solidifies the company's oversight of Bond's big-screen future.

Politics	.002
Movies	**.48**
Sports	.0001
Tech	.39
Health	.0001
Finance	.05
…	

Adapted from https://www.nbcnews.com/pop-culture/movies/amazon-taking-control-james-bond-movie-franchise-rcna192967

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 (TP) ● *Actual* ● *Guessed*	False Positive (FP) ○ *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 (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*

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

Contingency Table (out of table form)

Query:
Articles about dogs

Simple model classifies based on presence of “dog” or “dogs”

FN

Adopting an Animal at the Pound

Why Huskies Howl

Who’s a good boy?

TP

10 Tips for Grooming your Dog

What Foods to Hide from Your Dog During the Holidays

FP

It’s raining cats and dogs

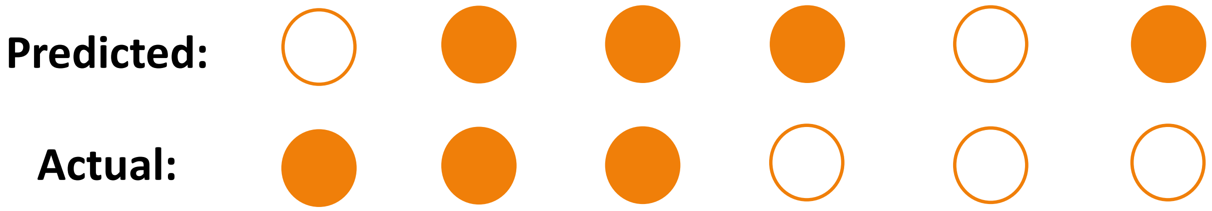
TN

15 Weird Animals You Can Find in North America

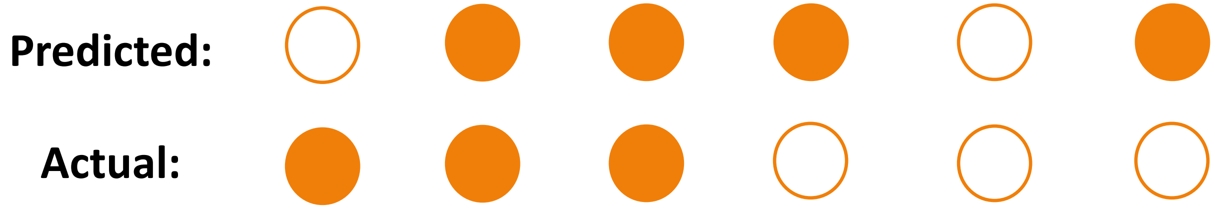
When Cats Rule the World

Meme from: *https://www.reddit.com/r/AdviceAnimals/comments/ck8xh0/yo_dawg_i_heard_you_like_old_memes/*

Contingency Table Example



Contingency Table Example

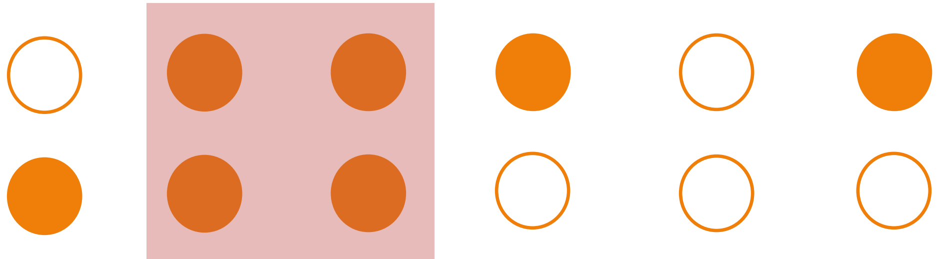


	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP)	False Positive (FP)
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)

Contingency Table Example

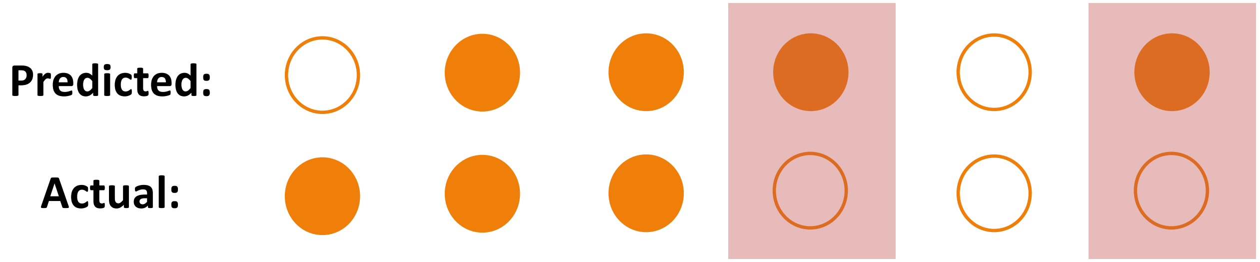
Predicted:

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

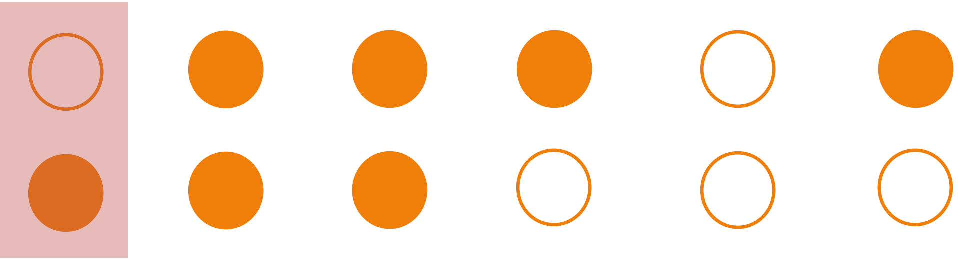


	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)

Contingency Table Example

Predicted:

Actual:

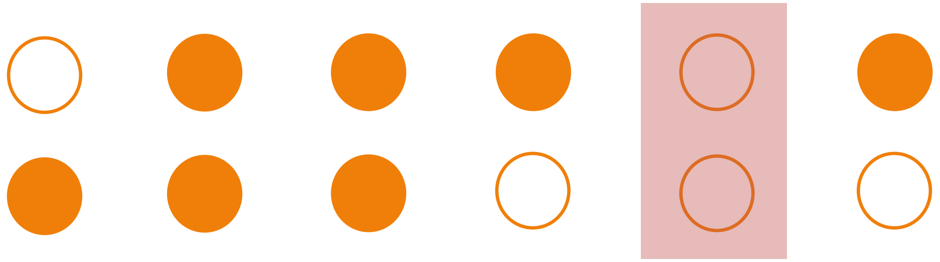


	What is the actual label?	
What label does our system predict? (↓)	**Actual Target Class ("●")**	**Not Target Class ("○")**
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN) = 1	True Negative (TN)

Contingency Table Example

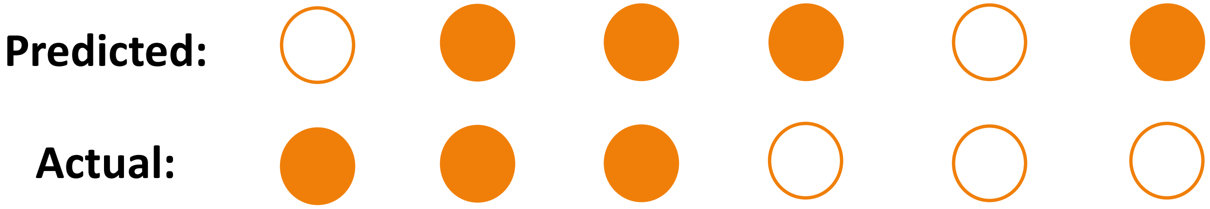
Predicted:

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) = 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

Min: 0 😟
Max: 1 😀

Accuracy: % of items correct

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

Precision: % of selected items that are correct

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

"**Precision** measures the percentage of the items that precision the system detected (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels"
SLP, ch. 4

	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

Min: 0 😟
Max: 1 😀

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

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

"**Recall** measures the percentage of items actually present in the input that were correctly identified by the system."
SLP, ch. 4

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

We're going to evaluate a Dogs vs Muffins classifier

https://petcentral.chewy.com/are-blueberries-safe-for-dogs-and-everything-else-you-could-possibly-want-to-know-about-dogs-and-blueberries/

Knowledge Check

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

1. Fill out a contingency table for this example. Your target class is **Dog**.

2. Then calculate precision, recall, and accuracy.

Actual:
Blueberry Blueberry Dog Dog Blueberry

Predicted:
Blueberry Dog Dog Blueberry Blueberry

	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)

The Importance of “Polarity” in Binary Classification

What are you trying to “identify” in your classification?

That is, are you trying to find ● or ○ ?

If ● is our target

		Correct Value	
		●	○
Guessed Value	●	?	?
	○	?	?

Where do
TP / FP / FN / FN go?

If is our target

		Correct Value	
Guessed Value		TP	FP
		FN	TN

If ○ is our target

		Correct Value	
		●	○
Guessed Value	●	?	?
	○	?	?

Where do
TP / FP / FN / FN go?

If ○ is our target

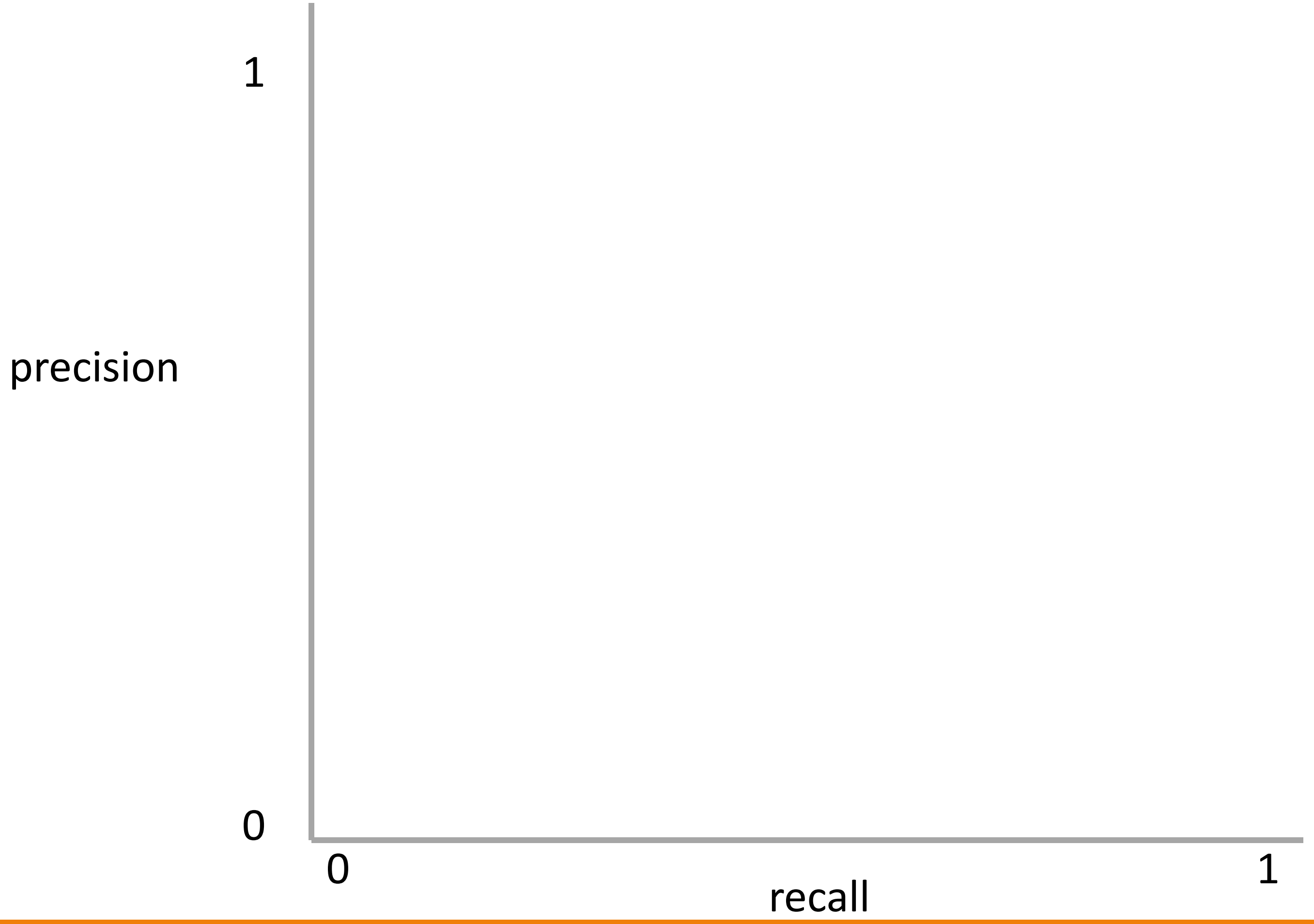
		Correct Value	
		●	○
Guessed Value	●	TN ○	FN ○
	○	FP ○	TP ○

When there are two classes, TP/TN & FP/FN are symmetrical

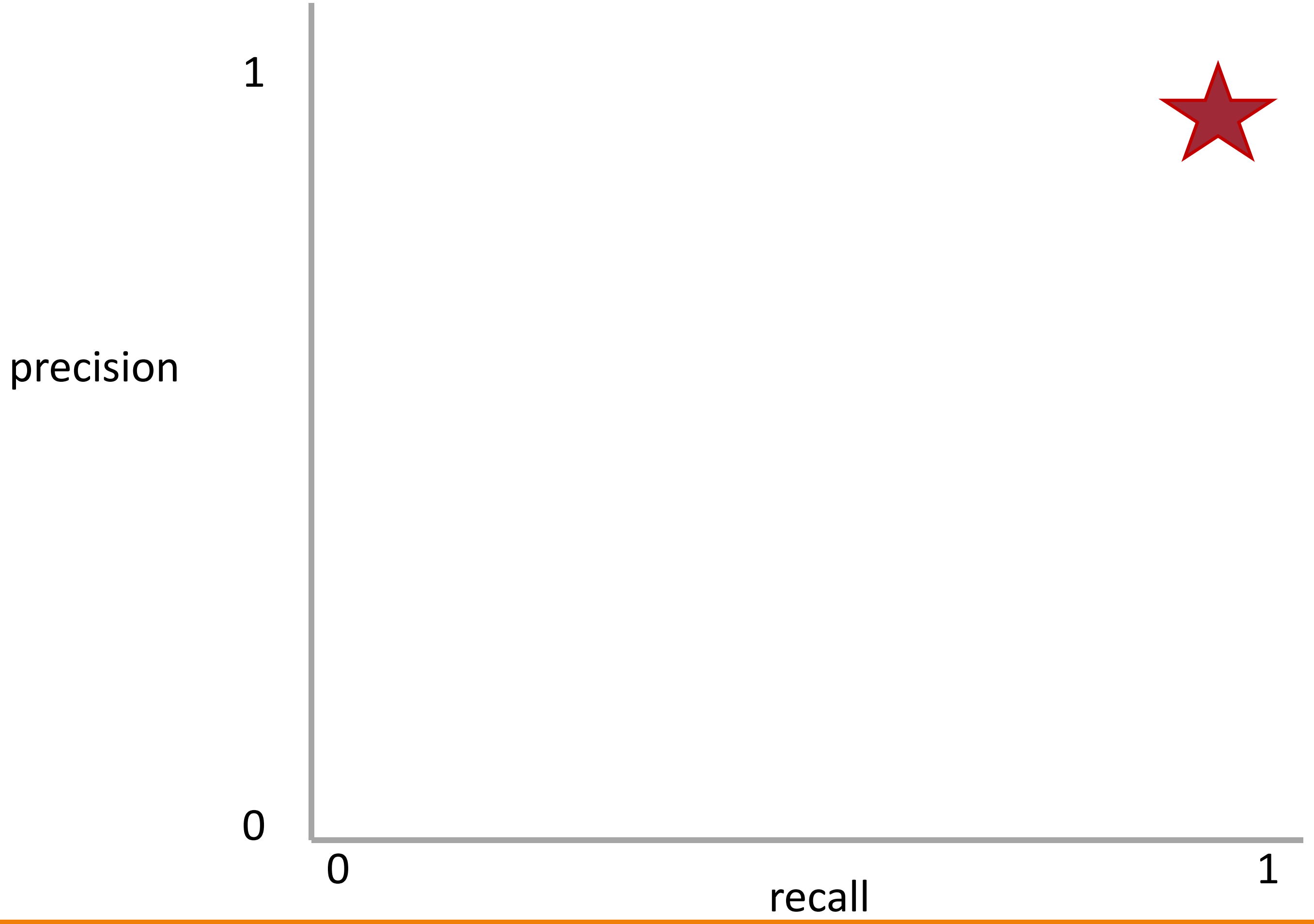
		Correct Value	
		●	○
Guessed Value	●	TP ● $= TN$ ○	FP ● $= FN$ ○
	○	FN ● $= FP$ ○	TN ● $= TP$ ○

Precision and Recall Present a Tradeoff

Q: Where do you want your ideal model?



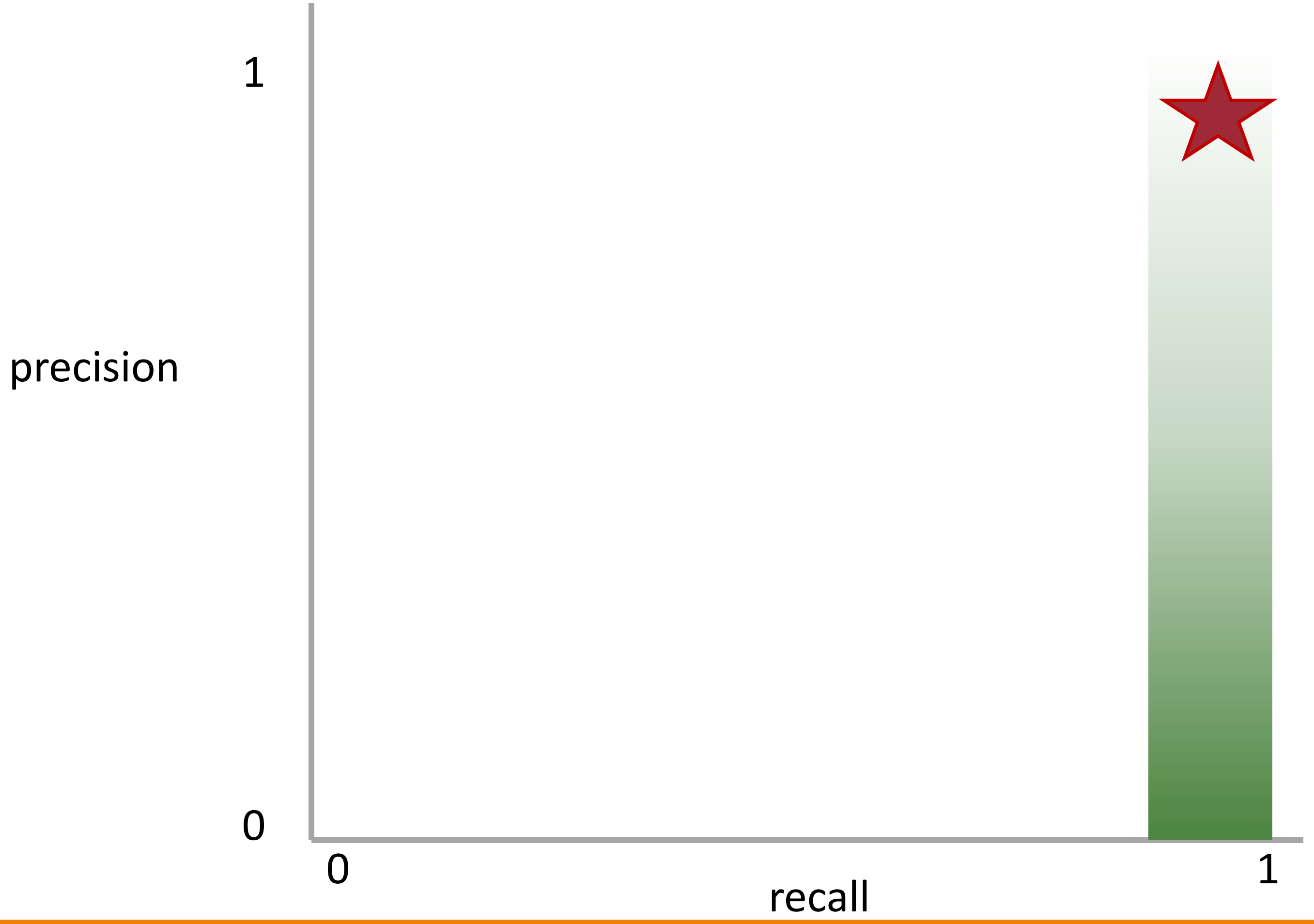
Precision and Recall Present a Tradeoff



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

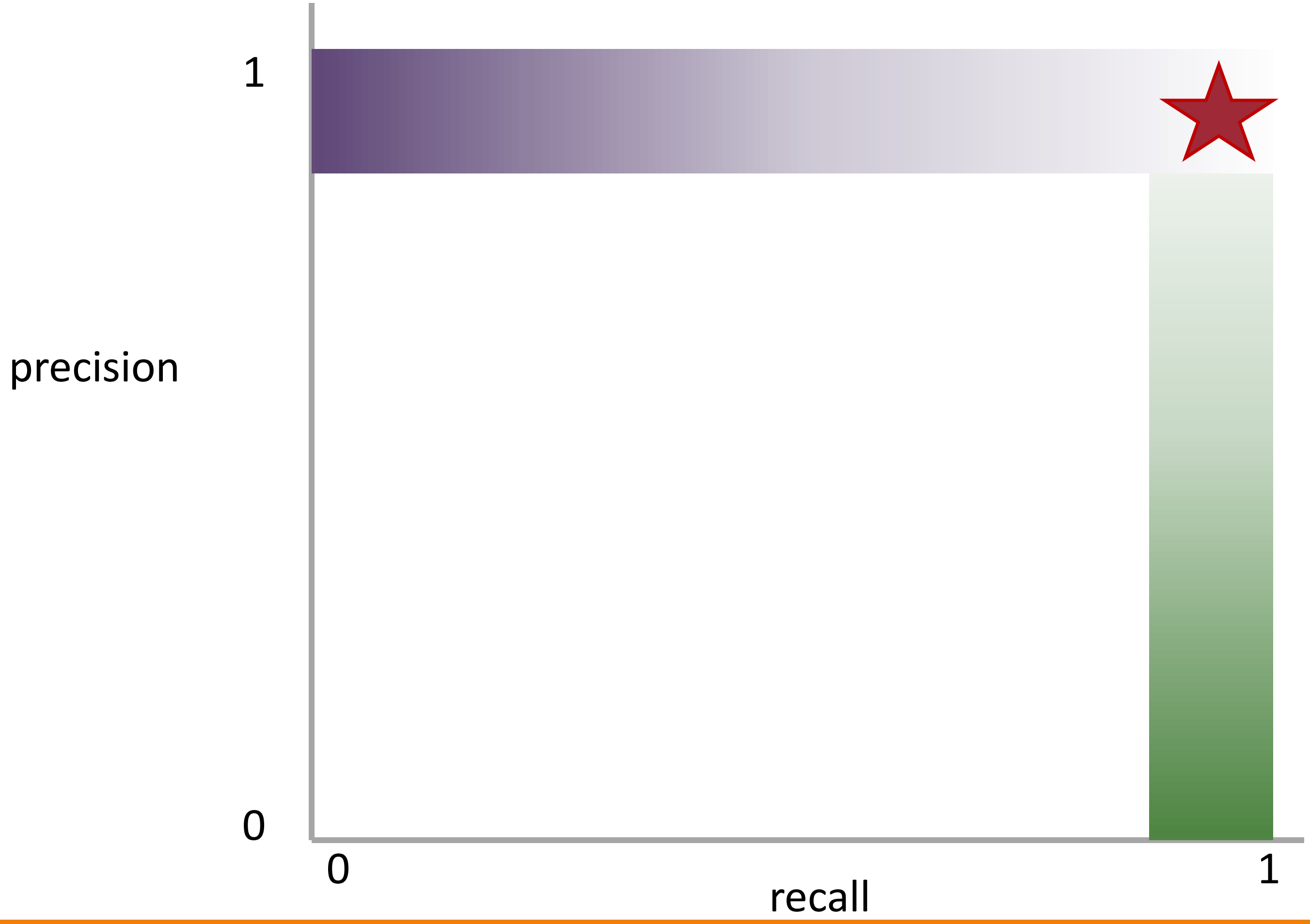


Q: Where do you want your ideal model ?

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

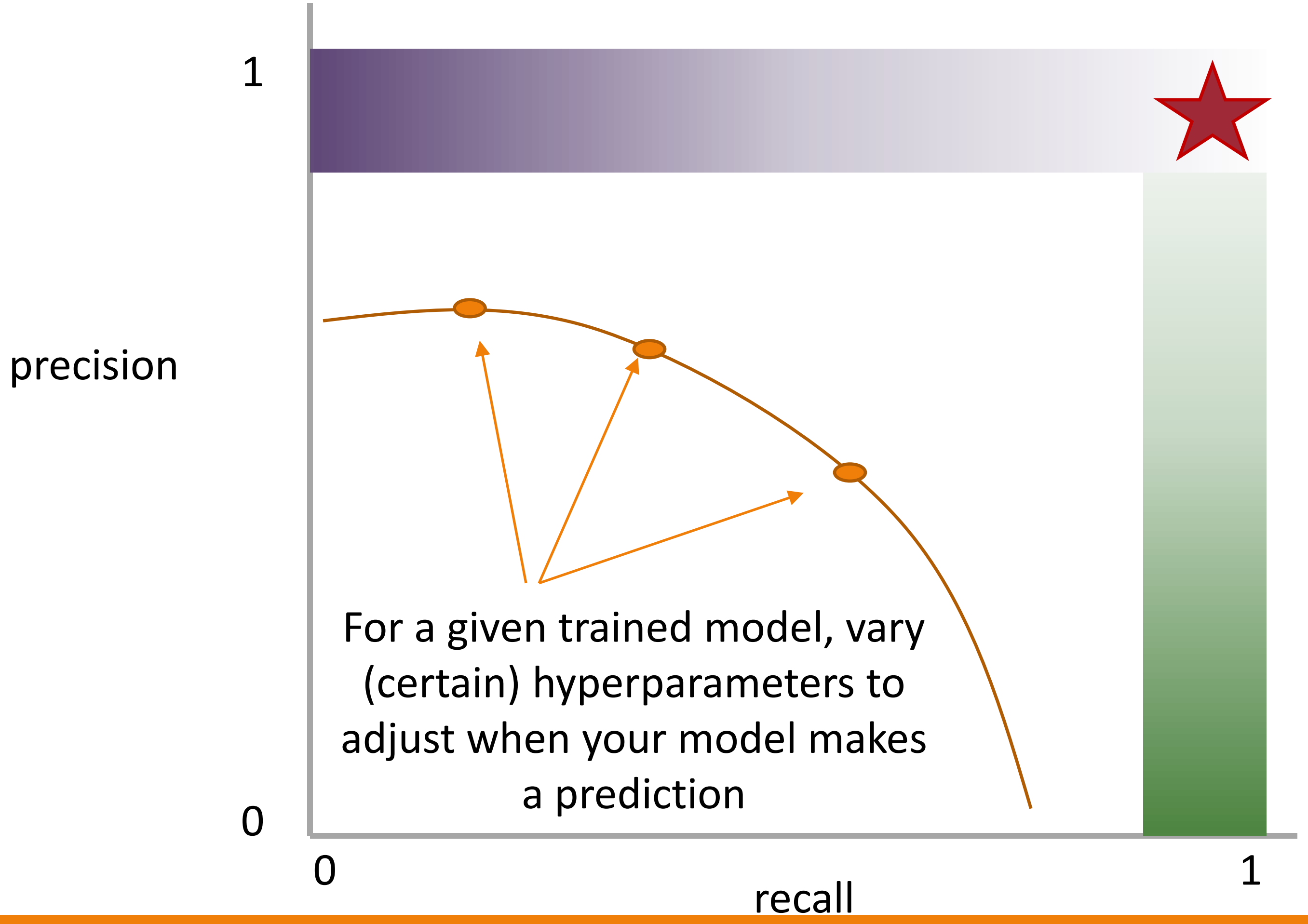


Q: Where do you want your ideal model?

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



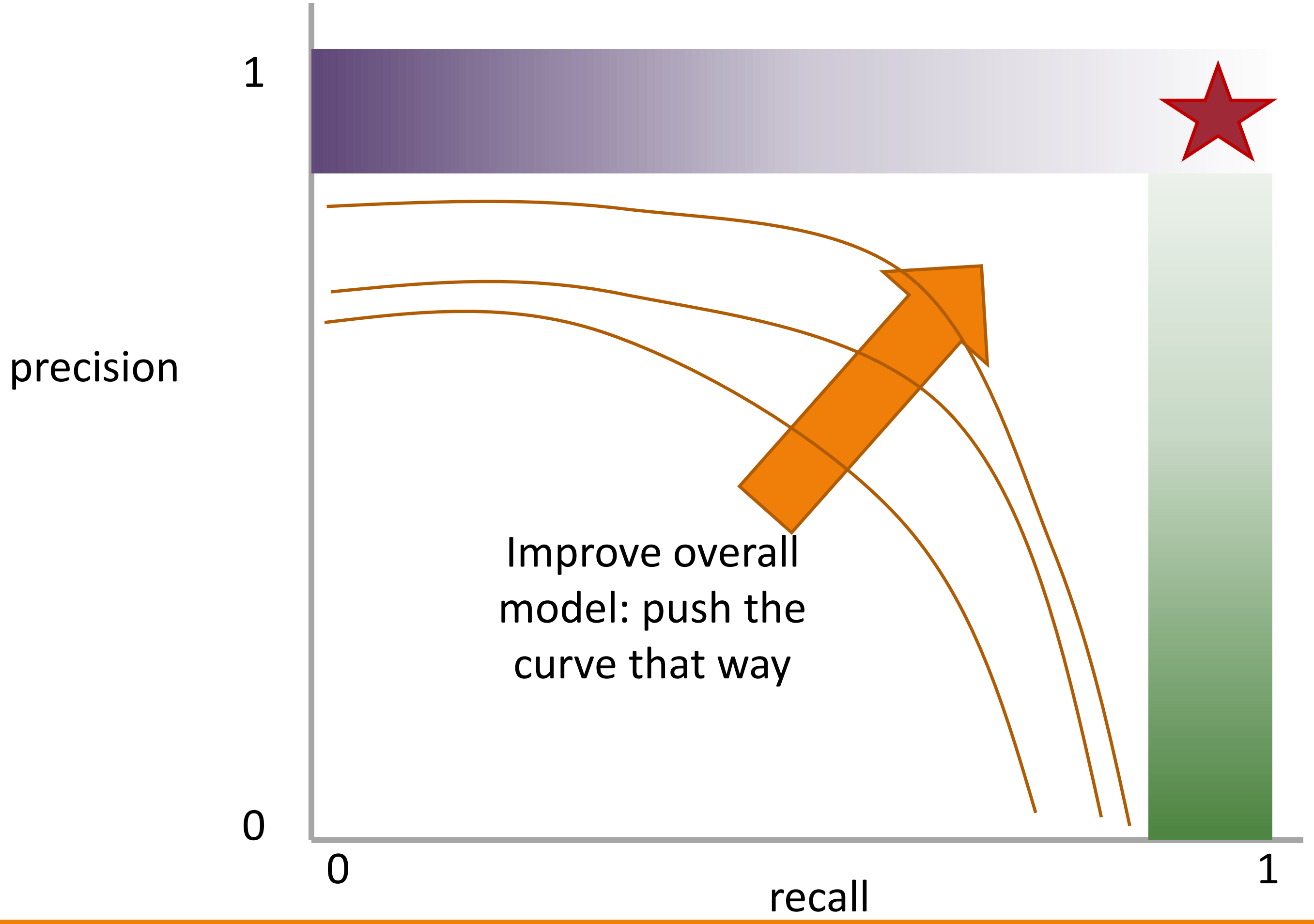
Q: Where do you want your ideal model ?

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



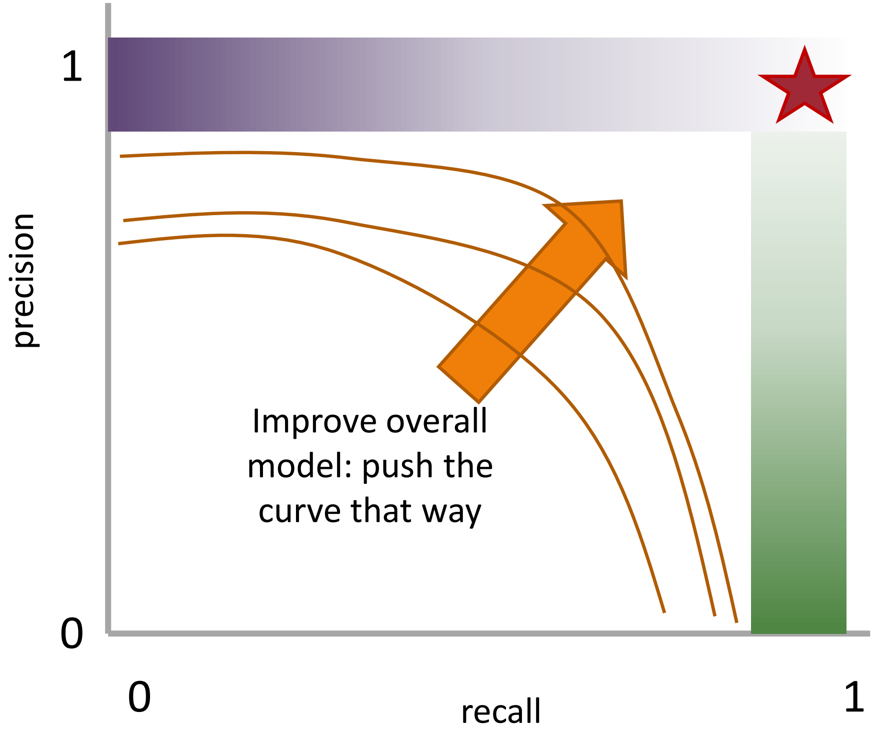
Q: Where do you want your ideal model ?

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?

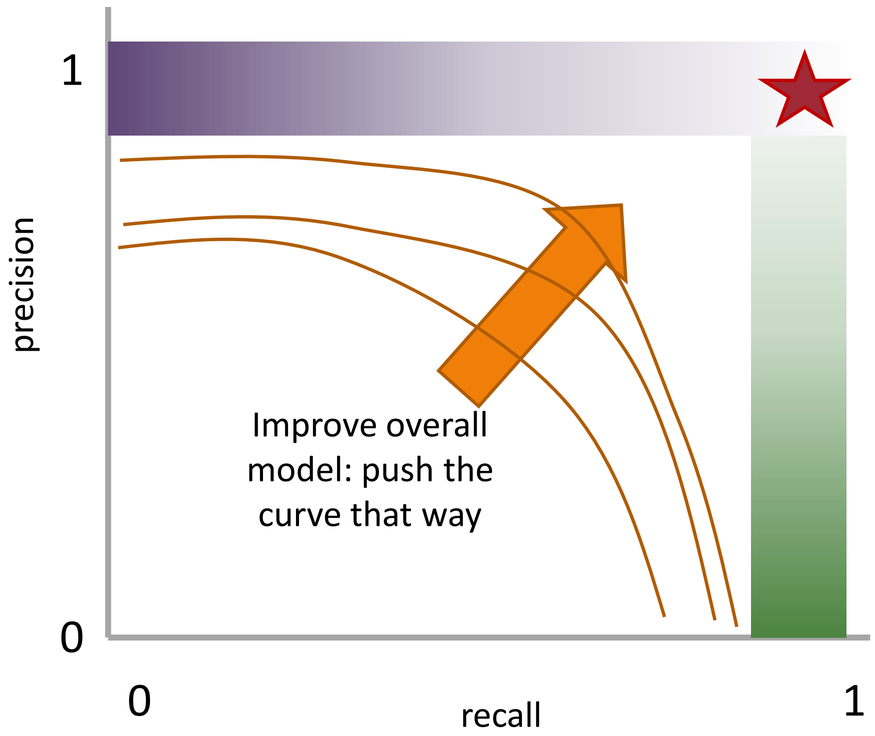
Idea: measure the tradeoff between precision and recall

Measure this Tradeoff: Area Under the Curve (AUC)



AUC measures the area under this tradeoff curve

Measure this Tradeoff: Area Under the Curve (AUC)



AUC measures the area under this tradeoff curve

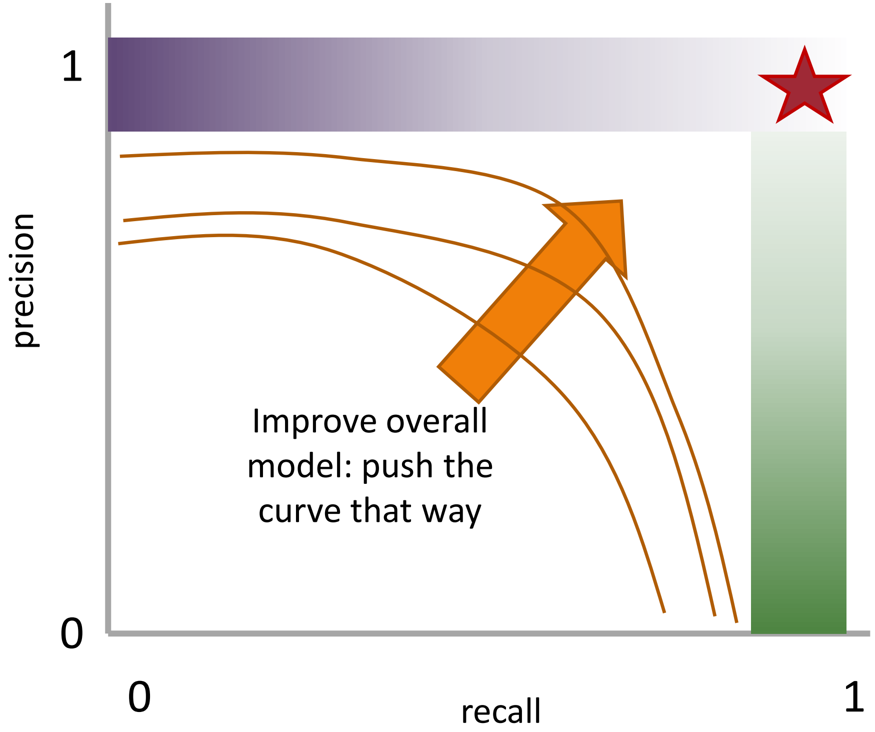
1. Computing the curve

   You need true labels & predicted labels with some score/confidence estimate

   Threshold the scores and for each threshold compute precision and recall

Measure this Tradeoff: Area Under the Curve (AUC)

Min AUC: 0 🙁
Max AUC: 1 😀



AUC measures the area under this tradeoff curve

1. Computing the curve

   You need true labels & predicted labels with some score/confidence estimate

   Threshold the scores and for each threshold compute precision and recall

2. Finding the area

   How to implement: trapezoidal rule (& others)

In practice: external library like the sklearn.metrics module

A combined measure: F1 (or F-score)

Weighted (harmonic) average of **P**recision & **R**ecall

F1 measure: equal weighting between precision and recall

$$F_1 = \frac{2 \ * P \ * R}{P + R}$$

A combined measure: F1 (or F-score)

Weighted (harmonic) average of **P**recision & **R**ecall

F1 measure: equal weighting between precision and recall

$$F_1 = \frac{2 * P * R}{P + R} = \frac{2 * TP}{2 * TP + FP + FN}$$

(useful when $P = R = 0$)

Comparing Accuracy & F1

Accuracy: % of items correct

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

$$F_1 = \frac{2 * P * R}{P+R} = \frac{2*TP}{2*TP+FP+FN}$$

Accuracy takes everything in consideration

F-Score is focused on TP

When would you want to use accuracy vs F1?

Accuracy works better if the dataset is balanced

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

P/R/F in a Multi-class Setting: Micro- vs. Macro-Averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

P/R/F in a Multi-class Setting: Micro- vs. Macro-Averaging

Macroaveraging: Compute performance for each class, then average.

$$\text{macroprecision} = \frac{1}{C}\sum_{c}\frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FP_c}} = \frac{1}{C}\sum_{c}\text{precision}_c$$

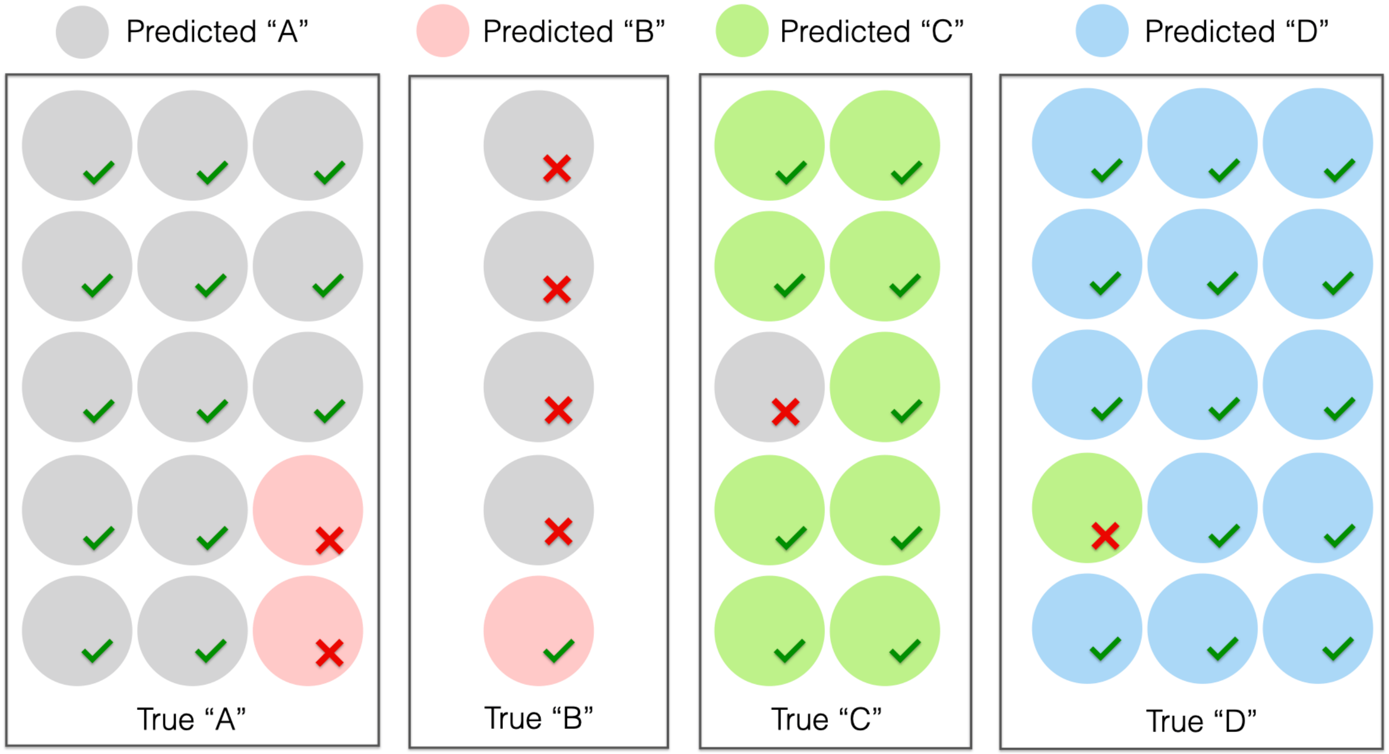
$$\text{macrorecall} = \frac{1}{C}\sum_{c}\frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FN_c}} = \frac{1}{C}\sum_{c}\text{recall}_\mathrm{c}$$

Microaveraging: Collect decisions for all classes, compute contingency table, evaluate.

$$\text{microprecision} = \frac{\sum_\mathrm{c} \mathrm{TP_c}}{\sum_\mathrm{c} \mathrm{TP_c} + \sum_\mathrm{c} \mathrm{FP_c}}$$

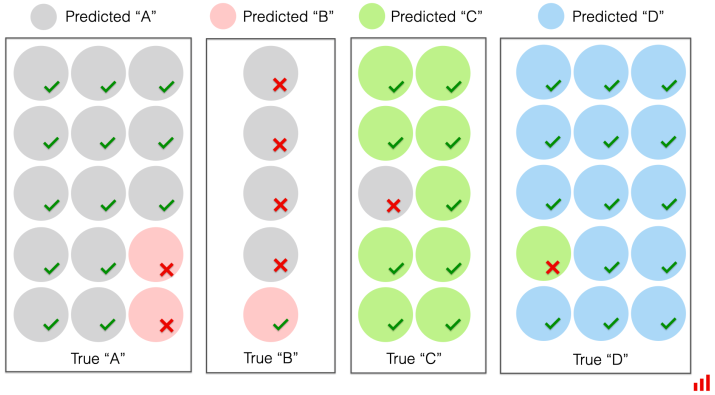
$$\text{microrecall} = \frac{\sum_\mathrm{c} \mathrm{TP_c}}{\sum_\mathrm{c} \mathrm{TP_c} + \sum_\mathrm{c} \mathrm{FN_c}}$$

Macro/Micro Example

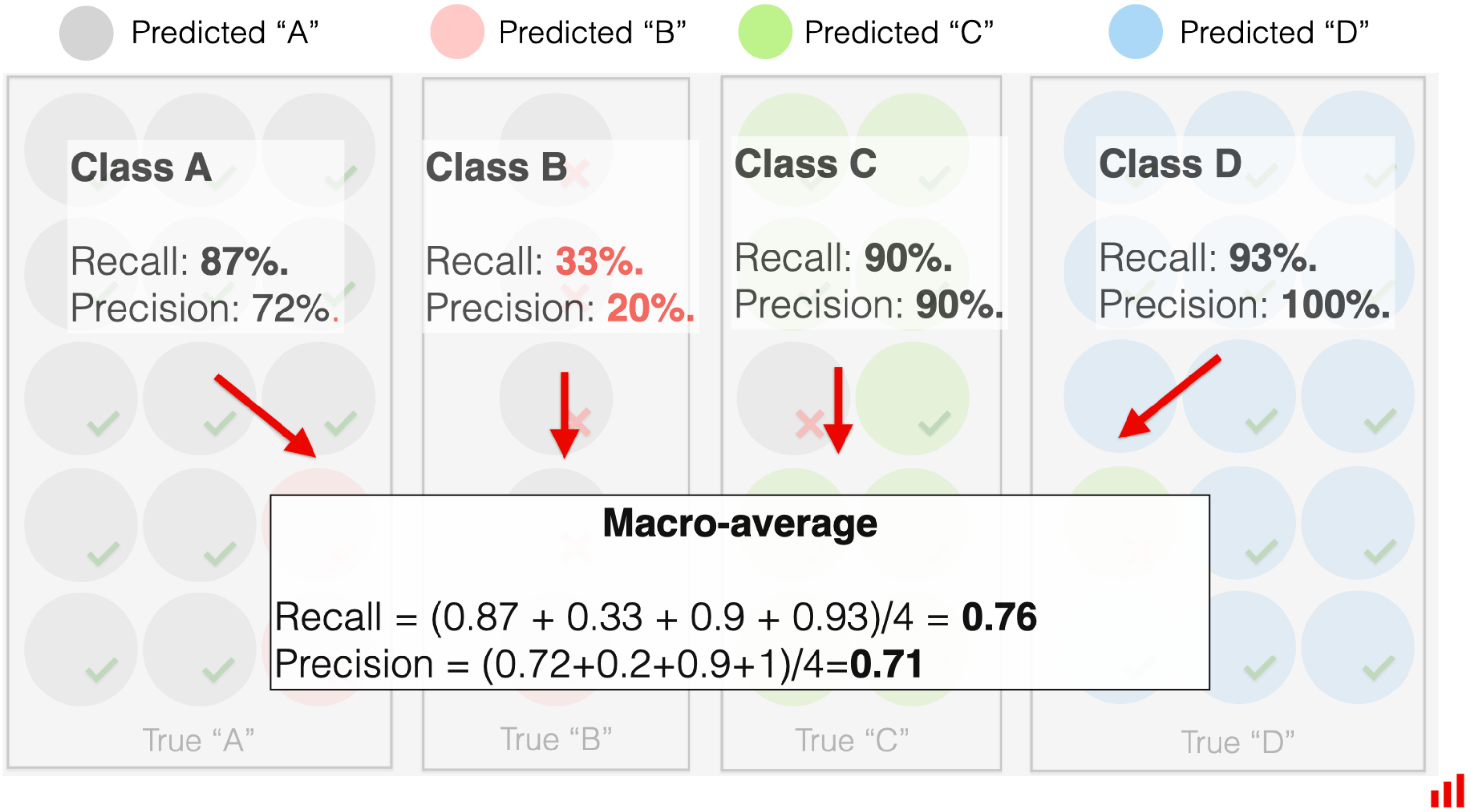


https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Each *class* has equal weight



Macro-Average

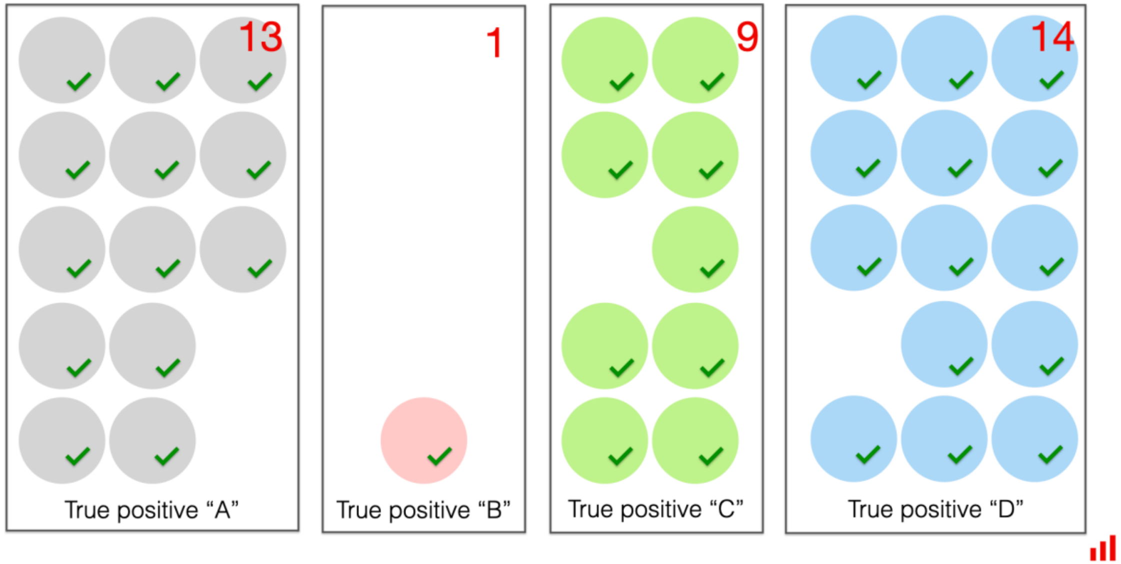


https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Each *instance* has equal weight

Micro-Average

All true positives

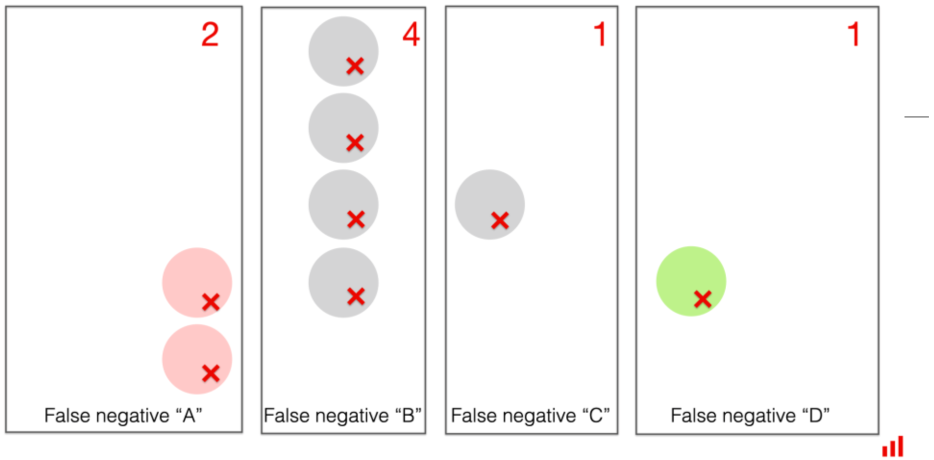


Total TP | Total FP | Total FN

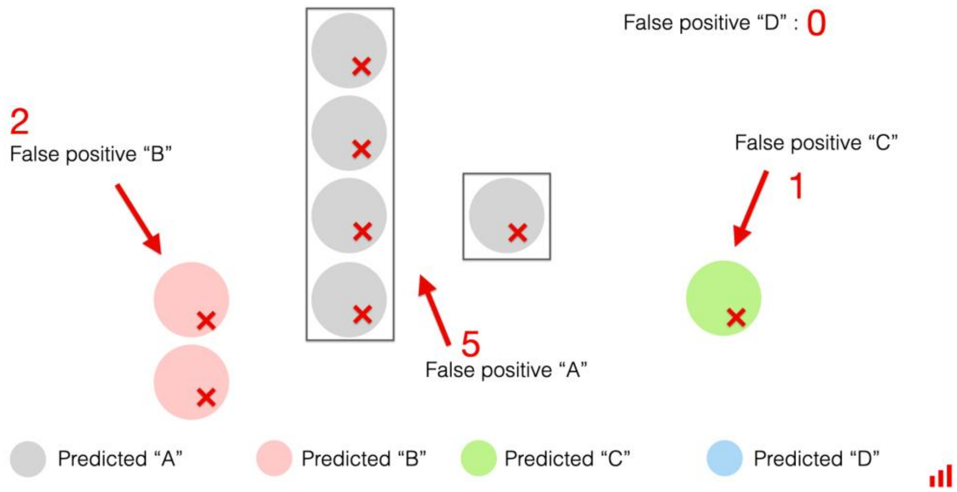
$$\text{Precision}_{\text{Micro-average}} = \frac{13+1+9+14}{13+1+9+14 + \boxed{2+5+1+0}} = 0.82$$

$$\text{Recall}_{\text{Micro-average}} = \frac{13+1+9+14}{13+1+9+14 + 2+4+1+1} = 0.82$$

All false negatives



All false positives



https://www.evidentlyai.com/classification-metrics/multi-class-metrics

P/R/F in a Multi-class Setting: Micro- vs. Macro-Averaging

Macroaveraging: Compute performance for each class, then average.

$$\text{macroprecision} = \frac{1}{C}\sum_{c} \frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FP_c}} = \frac{1}{C}\sum_{c} \text{precision}_c$$

$$\text{macrorecall} = \frac{1}{C}\sum_{c} \frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FN_c}} = \frac{1}{C}\sum_{c} \text{recall}_\mathrm{c}$$

Microaveraging: Collect decisions for all classes, compute contingency table, evaluate.

$$\text{microprecision} = \frac{\sum_\mathrm{c} \mathrm{TP_c}}{\sum_\mathrm{c} \mathrm{TP_c} + \sum_\mathrm{c} \mathrm{FP_c}} \qquad \text{microrecall} = \frac{\sum_\mathrm{c} \mathrm{TP_c}}{\sum_\mathrm{c} \mathrm{TP_c} + \sum_\mathrm{c} \mathrm{FN_c}}$$

When would we want to prefer micro-averaging vs macro-averaging?

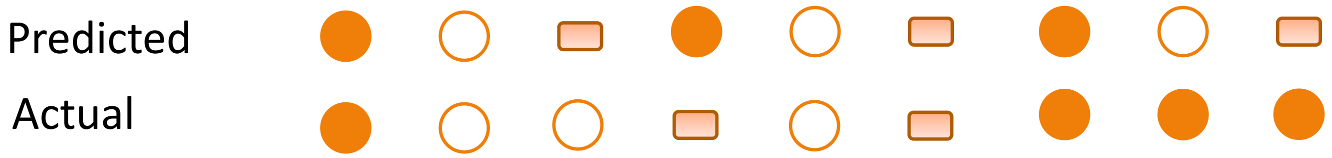
But how do we compute stats for multiple classes?

We already saw how the “polarity” affects the stats we compute…

Two main approaches. Either:

1. Compute “one-vs-all” 2x2 tables. OR
2. Generalize the 2x2 tables and compute per-class TP / FP / FN based on the diagonals and off-diagonals

1. Compute “one-vs-all” 2x2 tables



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

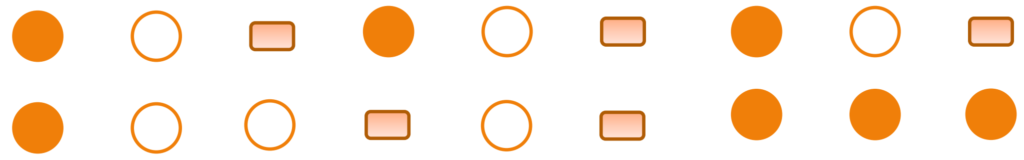
Look for	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

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

1. Compute “one-vs-all” 2x2 tables

Predicted

Actual



Look for	Actually Target	Actually Not Target
Selected/Guessed	2	1
Not select/not guessed	2	4

Look for	Actually Target	Actually Not Target
Selected/Guessed	2	1
Not select/not guessed	1	5

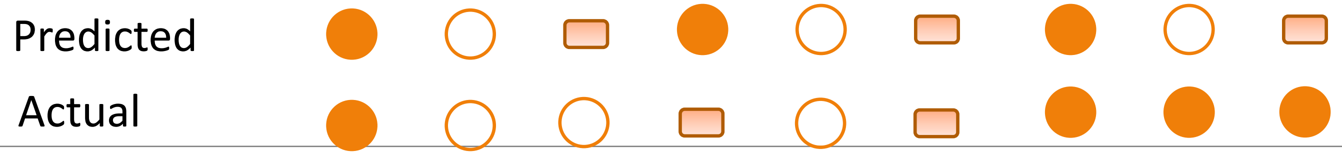
Look for	Actually Target	Actually Not Target
Selected/Guessed	1	2
Not select/not guessed	1	5

2. Generalizing the 2-by-2 contingency table

		Correct Value		
Guessed Value		#	#	#
		#	#	#
		#	#	#

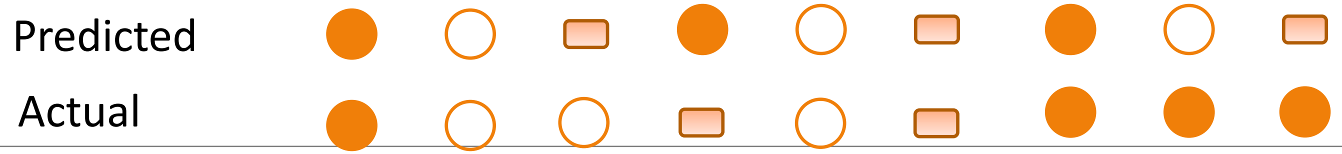
This is also called a **Confusion Matrix**

2. Generalizing the 2-by-2 contingency table



		Correct Value		
Guessed Value		a #	b #	c #
		d #	e #	f #
		g #	h #	i #

2. Generalizing the 2-by-2 contingency table



		Correct Value		
Guessed Value				
		2	0	1
		1	2	0
		1	1	1

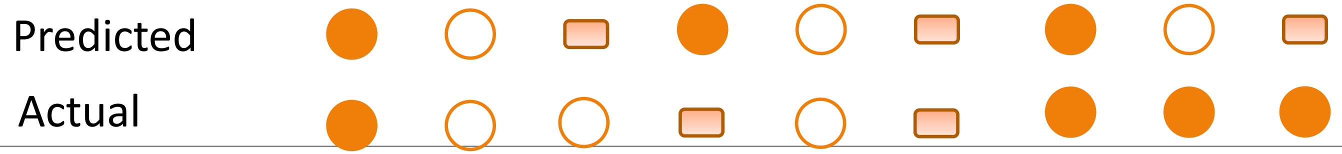
2. Generalizing the 2-by-2 contingency table

Predicted	●	○	▭	●	○	▭	●	○	▭
Actual	●	○	○	▭	○	▭	●	●	●

		Correct Value		
		●	○	▭
Guessed Value	●	a 2	b 0	c 1
	○	d 1	e 2	f 0
	▭	g 1	h 1	i 1

How do you compute $TP_{\bullet}$?

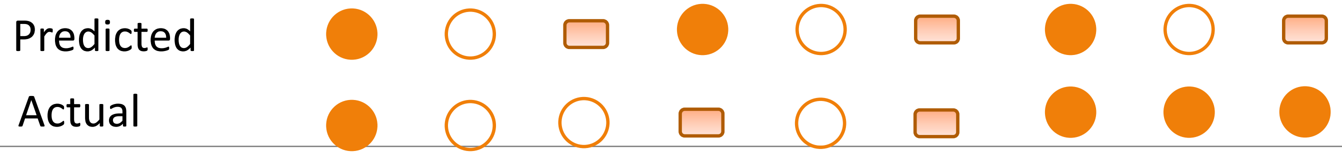
2. Generalizing the 2-by-2 contingency table



		Correct Value		
Guessed Value		a 2	b 0	c 1
		d 1	e 2	f 0
		g 1	h 1	i 1

How do you compute FN?

2. Generalizing the 2-by-2 contingency table

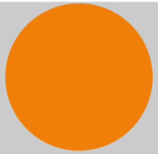
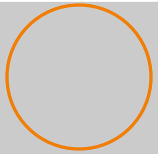
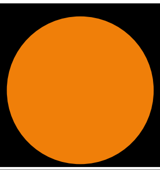
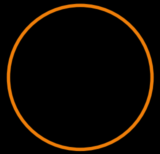


		Correct Value		
		●	○	▭
Guessed Value	●	a 2	b 0	c 1
	○	d 1	e 2	f 0
	▭	g 1	h 1	i 1

How do you compute $FP_{\bigcirc}$?

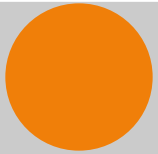
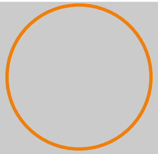
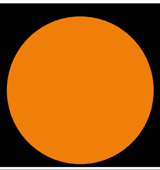
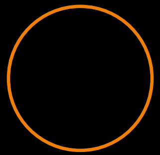
Performance of a Classifier using a Confusion Matrix

Q: Is this a good result?

		Correct Value		
				
Guessed Value		80	9	11
		7	86	7
		2	8	9

Performance of a Classifier using a Confusion Matrix

Q: Is this a good result?

		Correct Value		
				
Guessed Value		30	40	30
		25	30	50
		30	35	35

Performance of a Classifier using a Confusion Matrix

Q: Is this a good result?

		Correct Value		
		●	○	▭
Guessed Value	●	7	3	90
	○	4	8	88
	▭	3	7	90