

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

Slide courtesy Jim Martin

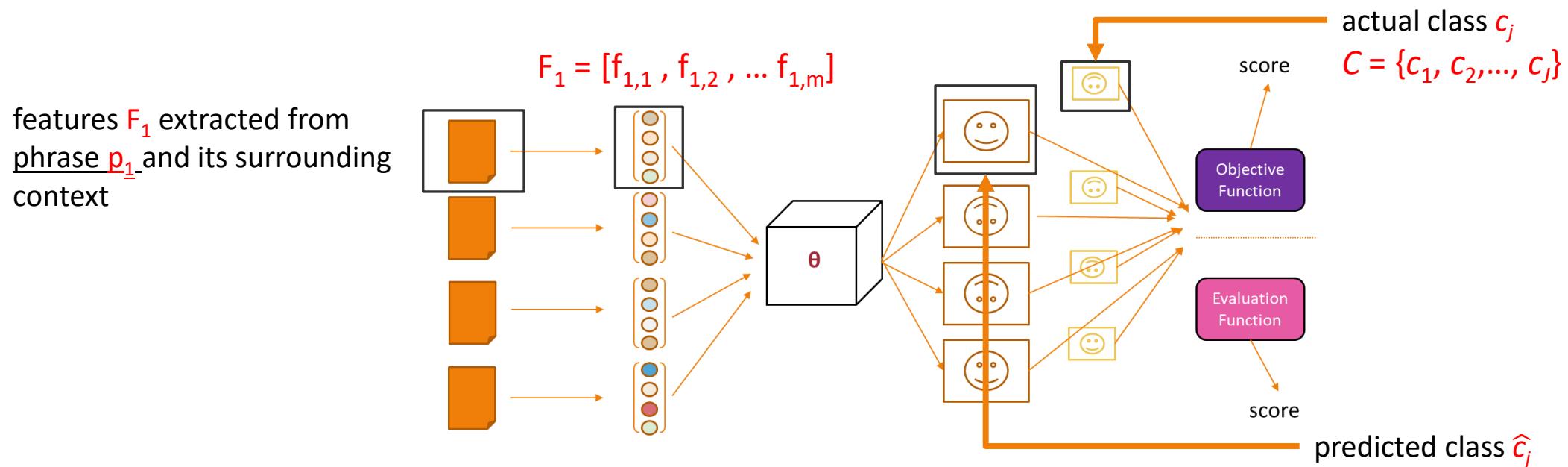
Named Entity Recognition

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



p(class | phrase in context)

Chunking Tasks

Named entity recognition (NER)

Information extraction (IE)

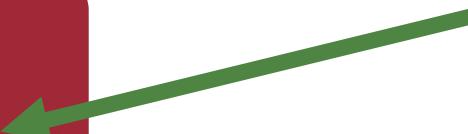
Identifying idioms

...

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.

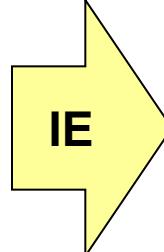
October 14, 2002, 4:00 a.m. PT

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...



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)

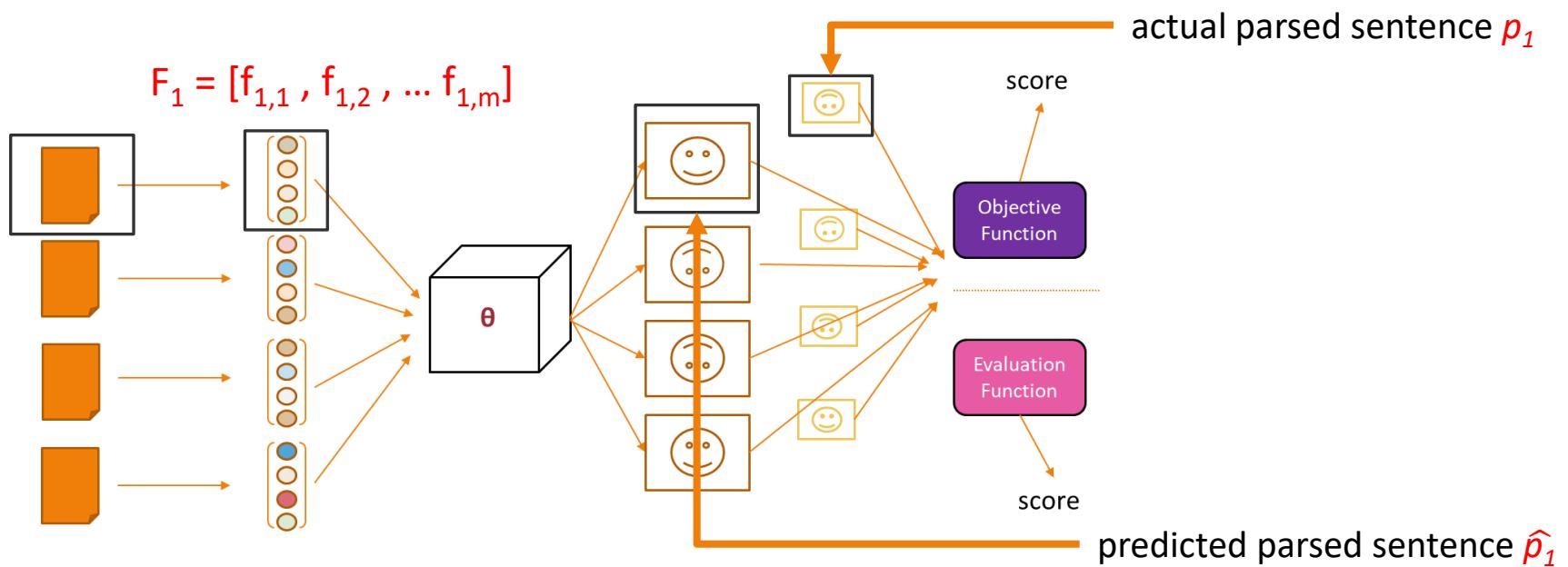
Text Annotation Tasks ("Classification" Tasks)

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

features F_1 extracted from phrase/sentence s_1



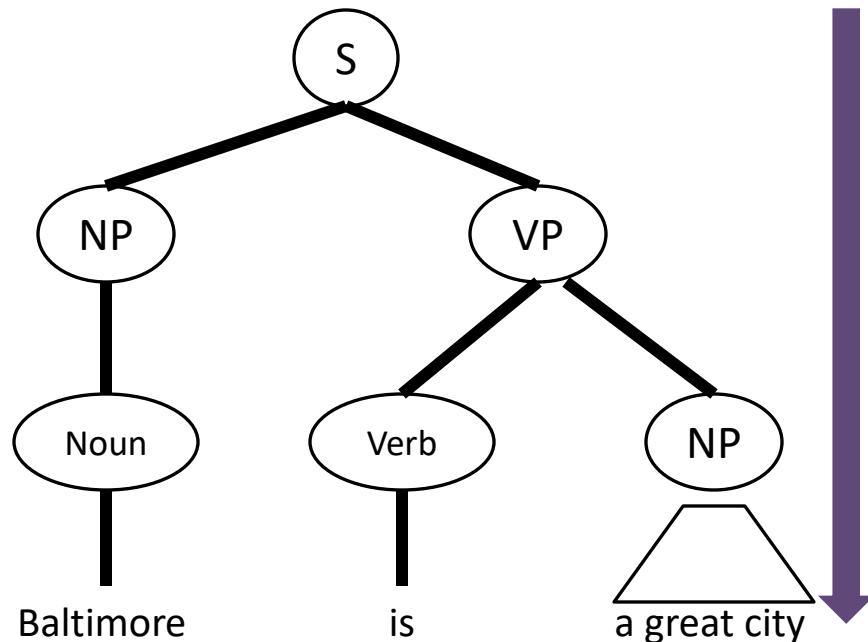
Context Free Grammar

```
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       ...
```

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

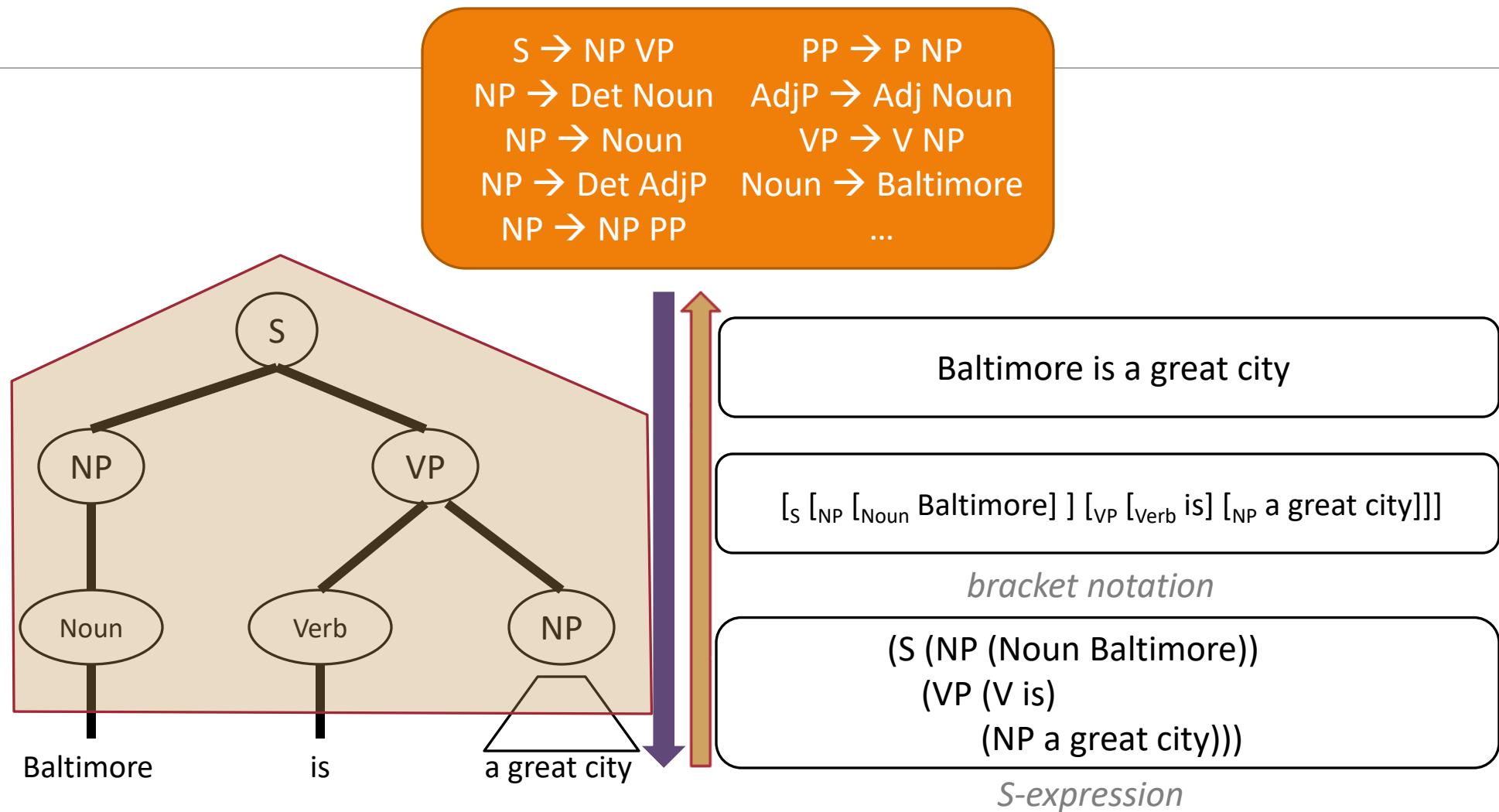
Generate from a Context Free Grammar

$S \rightarrow NP\ VP$ $PP \rightarrow P\ NP$
 $NP \rightarrow Det\ Noun$ $AdjP \rightarrow Adj\ Noun$
 $NP \rightarrow Noun$ $VP \rightarrow V\ NP$
 $NP \rightarrow Det\ AdjP$ $Noun \rightarrow Baltimore$
 $NP \rightarrow NP\ PP$...



Baltimore is a great city

Assign Structure (Parse) with a Context Free Grammar



Why is it useful?



Garden Path Sentences

The old man the boat .



Garden Path Sentences

The old man the boat .



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.



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

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

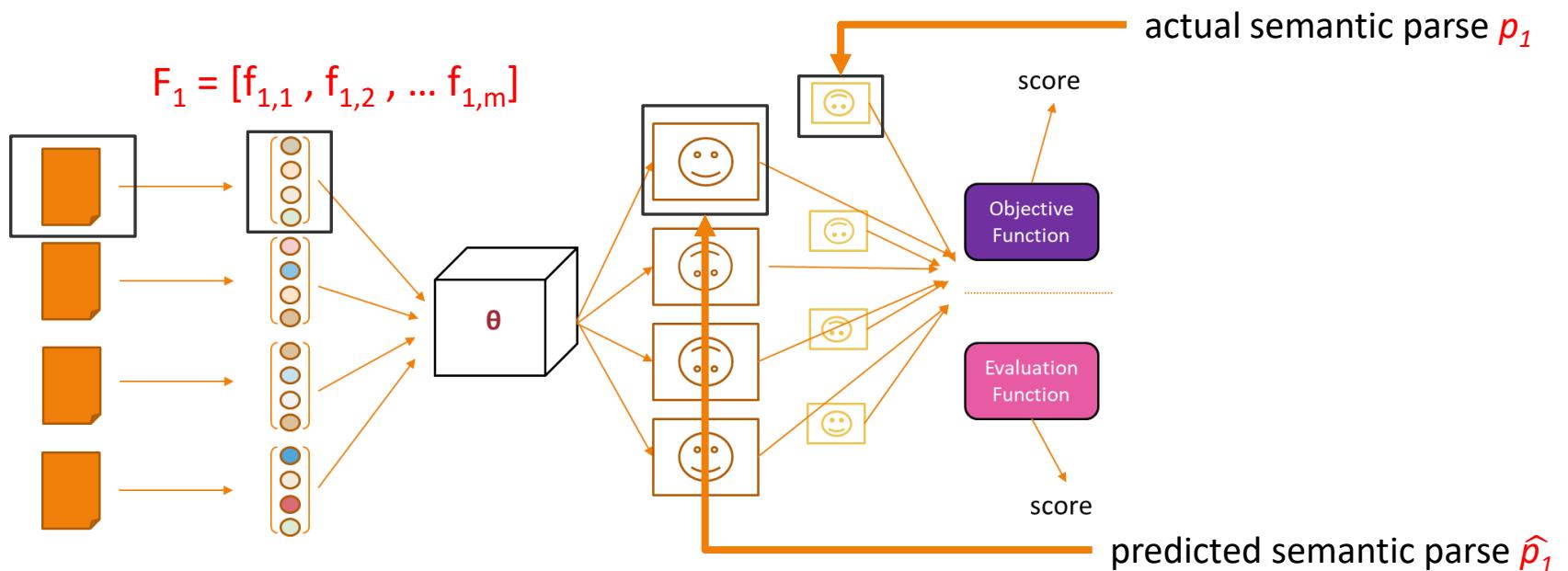
Slide courtesy Jason Eisner, with mild edits

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

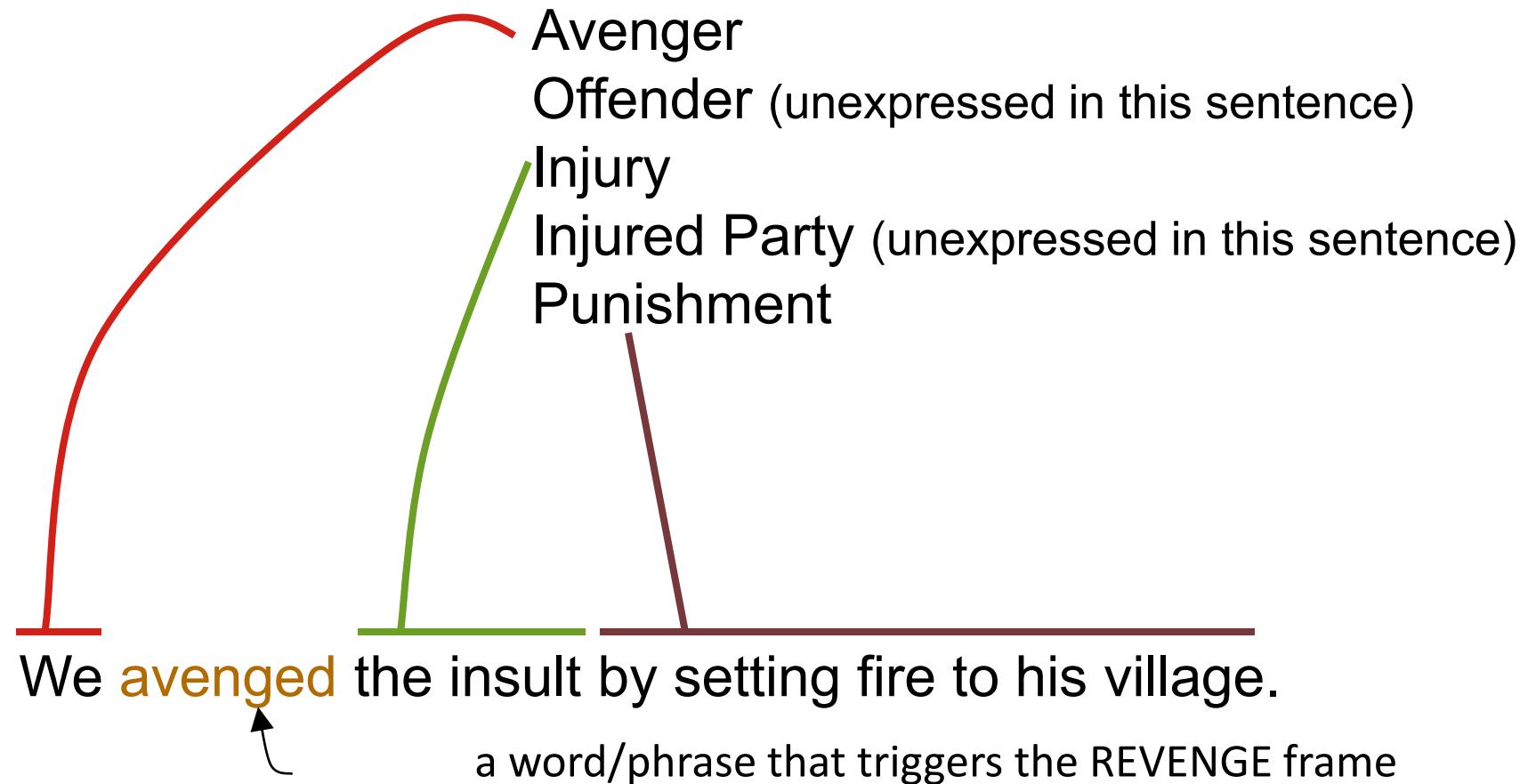
TimeBank – temporal expressions

Slide courtesy Jason Eisner, with mild edits

What type of applications might this have?

FrameNet Example

REVENGE FRAME



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

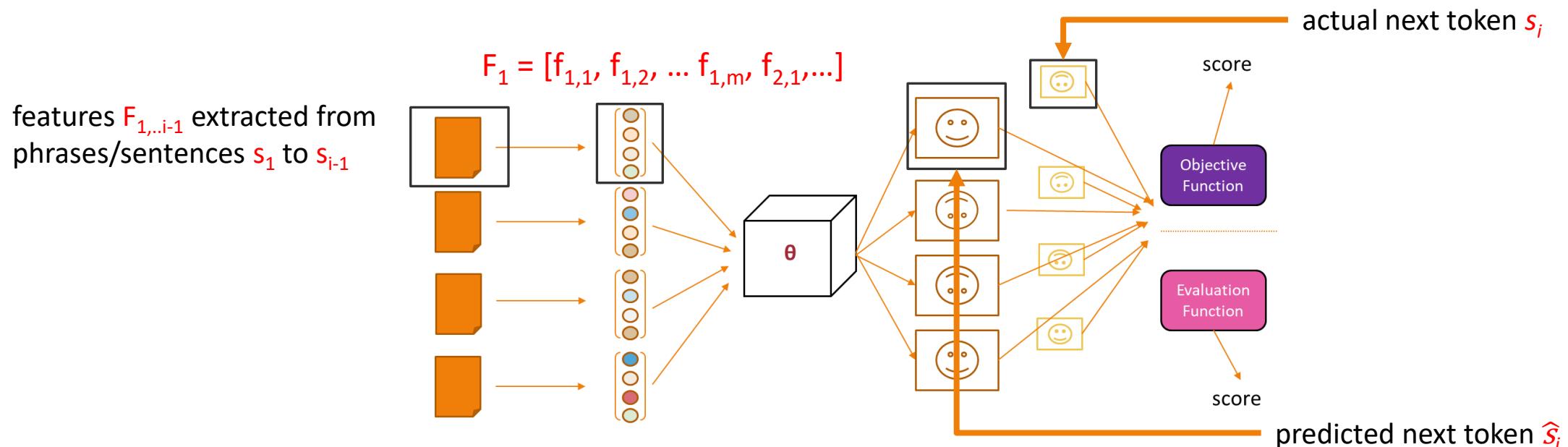
Slide courtesy Jason Eisner, with mild edits

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)

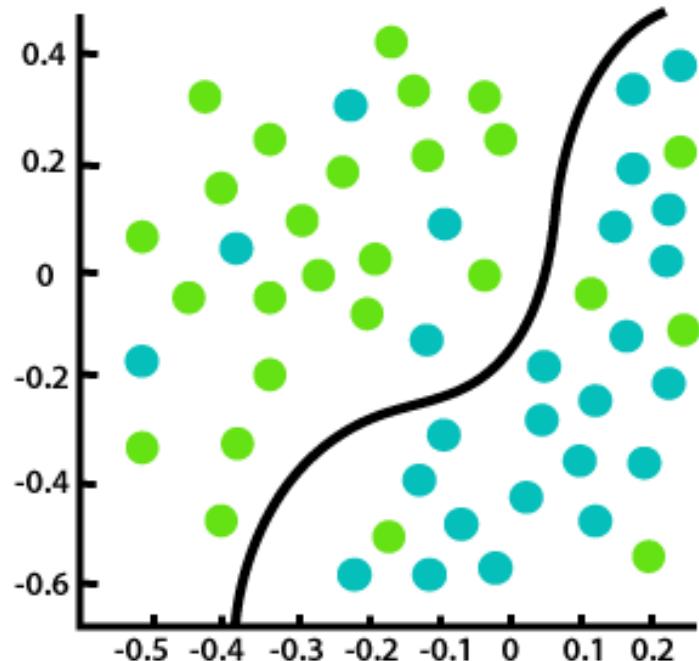
Summarization

Generating text from a structured representation

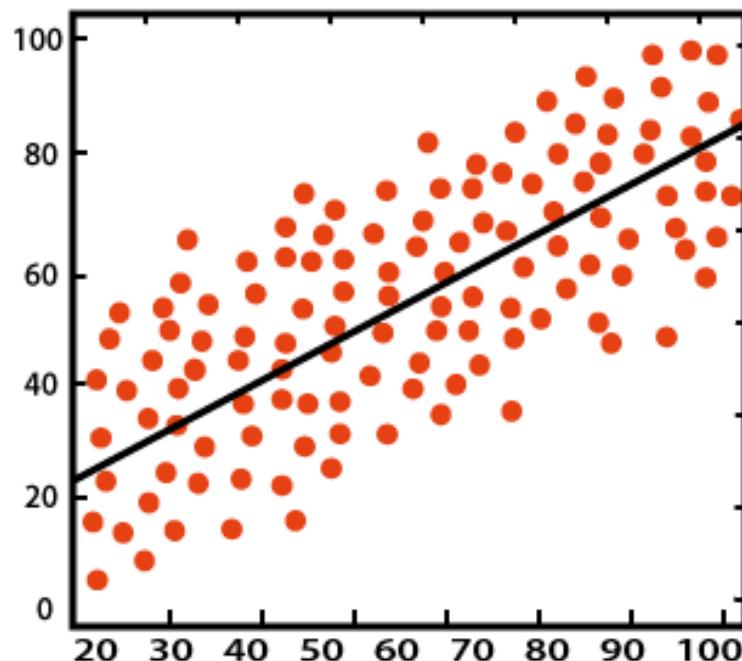
...

Calculating Evaluation Metrics for Classification

Review: Types of models



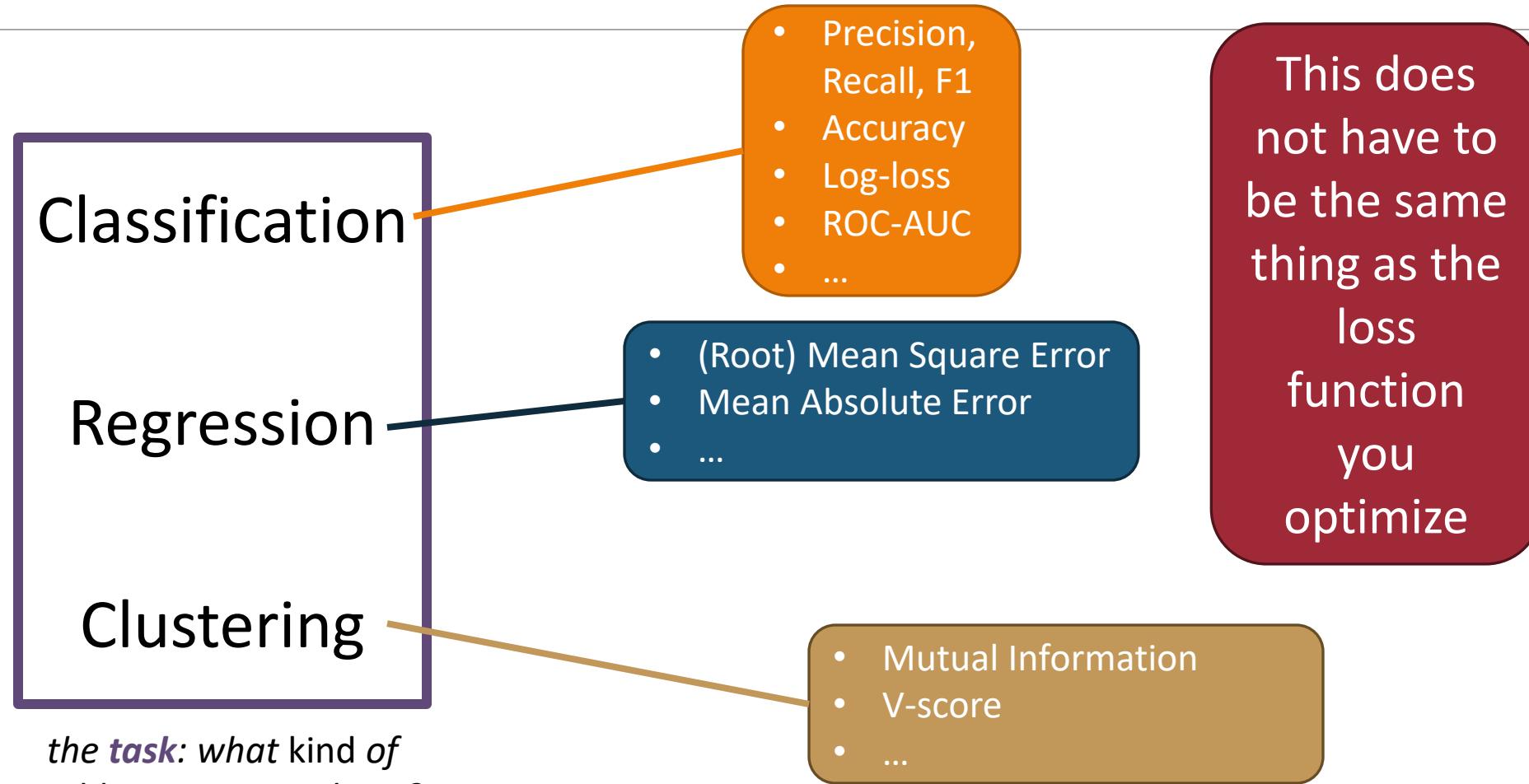
Classification



Regression

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

Central Question: How Well Are We Doing?



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. nltk
4. 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

$$p(\text{solid circle} | X) \text{ vs. } p(\text{hollow circle} | X)$$

Examining Assumption 3

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

$$p(\bullet|X) \text{ vs. } p(\circ|X)$$

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

...

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

...

Classification Evaluation: the 2-by-2 contingency table

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

Classification Evaluation: the 2-by-2 contingency table

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

Classification Evaluation: the 2-by-2 contingency table

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

Classification Evaluation: the 2-by-2 contingency table

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

Classification Evaluation: the 2-by-2 contingency table

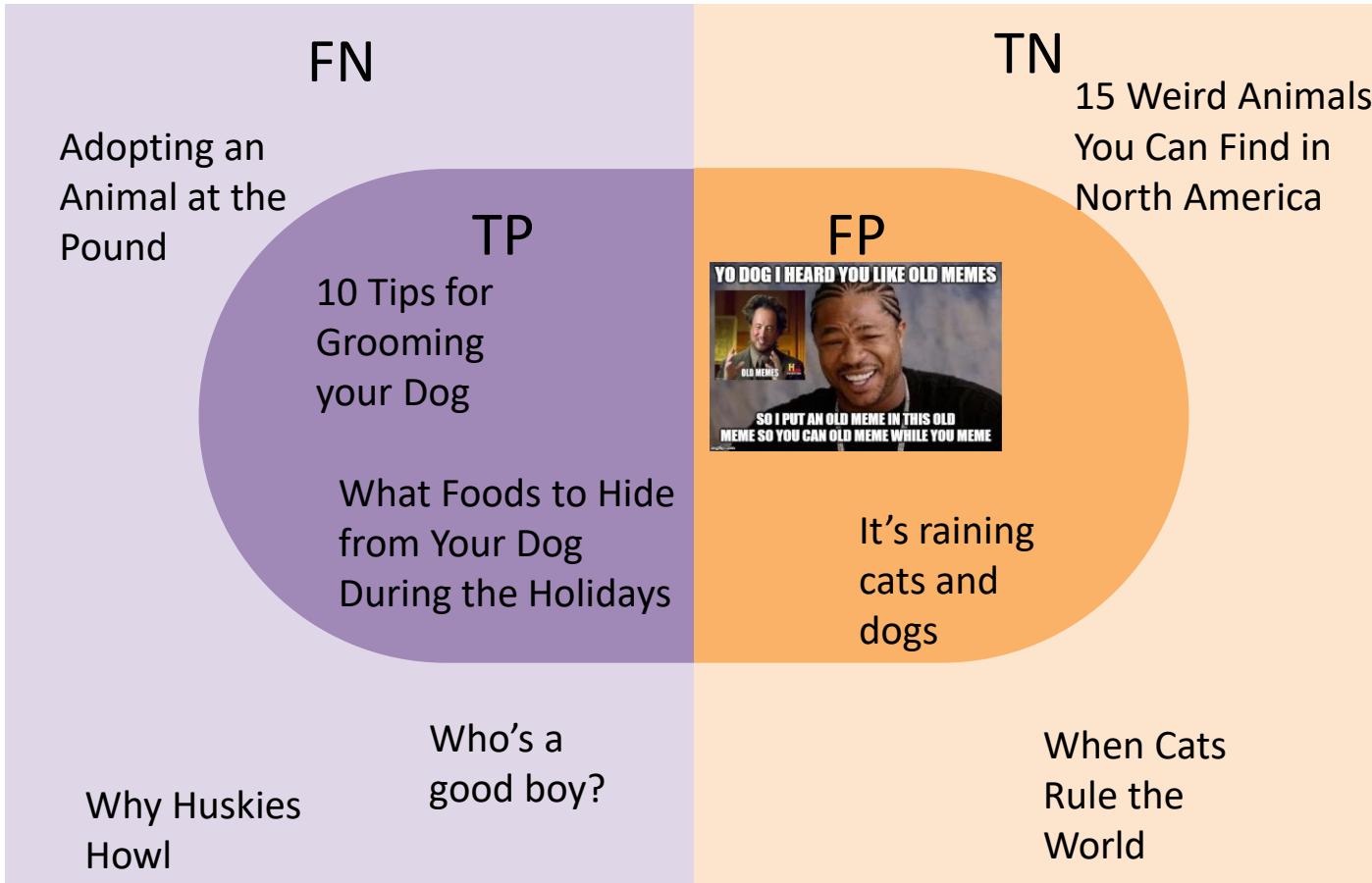
		<i>What is the actual label?</i>	
<i>What label does our system predict? (↓)</i>	Actual Target Class (“●”)		Not Target Class (“○”)
Selected/ Guessed (“●”)	True Positive <small>Actual</small> (TP)	False Positive <small>Actual</small> (FP)	False Positive <small>Guessed</small> (FP)
Not selected/ not guessed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

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”



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

Contingency Table Example

Predicted:	○	●	●	●	○	●
Actual:	●	●	●	○	○	○

Contingency Table Example

Predicted: ○ ● ● ● ○ ●
Actual: ● ● ● ○ ○ ○

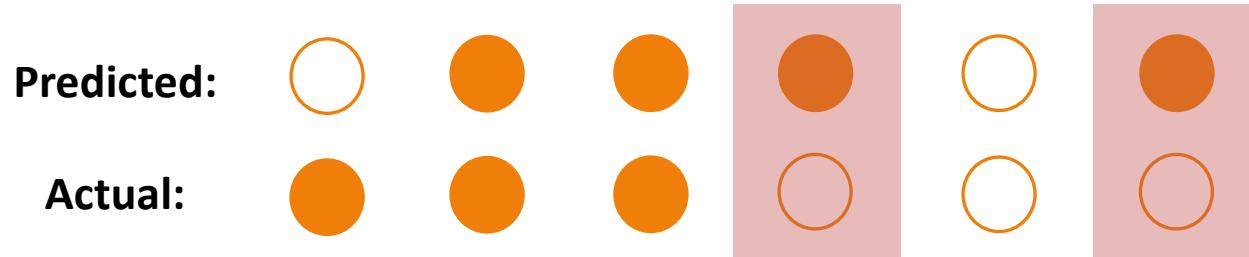
		<i>What is the actual label?</i>	
<i>What label does our system predict? (↓)</i>		Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP)	False Positive (FP)	
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)	

Contingency Table Example

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

Predicted: Actual:

Contingency Table Example

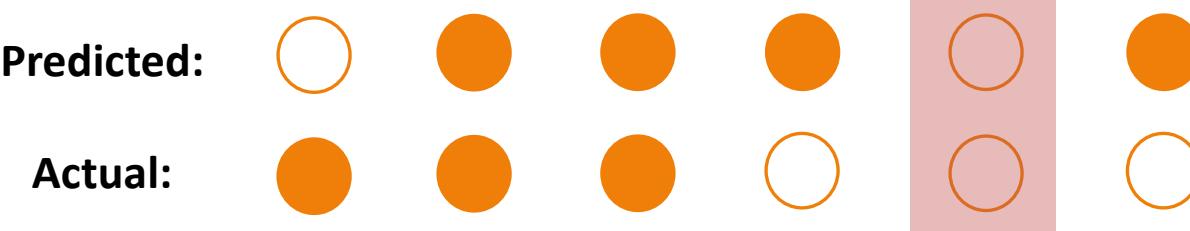


What is the actual label?		
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive (TP) = 2	False Positive (FP) = 2
Not selected/ not guessed ("○")	False Negative (FN)	True Negative (TN)

Contingency Table Example

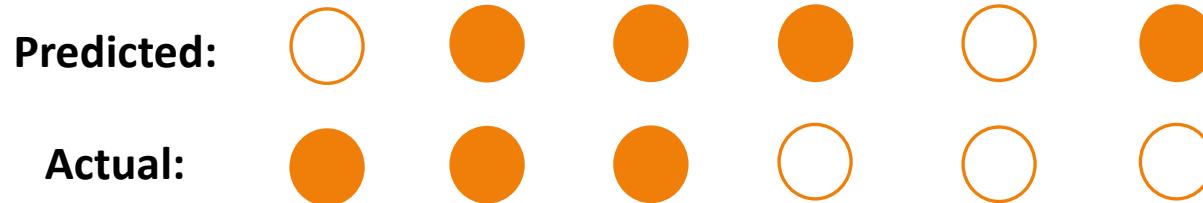
		What is the actual label?					
		Actual Target Class ("●")			Not Target Class ("○")		
What label does our system predict? (↓)		True Positive (TP) = 2		False Positive (FP) = 2			
Selected/ Guessed ("●")		True Positive (TP) = 2		False Positive (FP) = 2			
Not selected/ not guessed ("○")		False Negative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

Contingency Table Example



		What is the actual label?					
		Actual Target Class ("●")			Not Target Class ("○")		
What label does our system predict? (↓)		True Positive (TP) = 2		False Positive (FP) = 2			
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



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

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

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

Min: 0 😞
Max: 1 😃

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

Precision: % of selected items that are correct

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

“Precision measures the percentage of the items that 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)

Min: 0 😞
Max: 1 😃

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

Precision: % of selected items that are correct

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

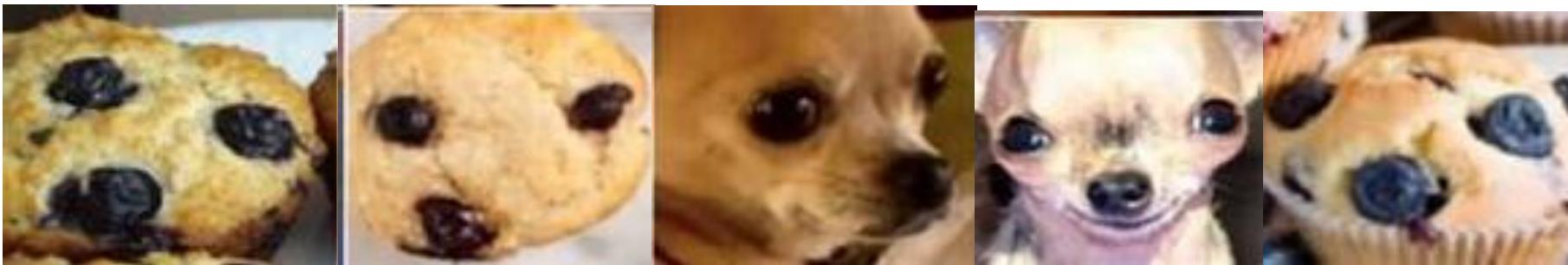
Recall: % of correct items that are selected

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

“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/>

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

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

Knowledge Check

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

		<i>What is the actual label?</i>	
<i>What label does our system predict? (↓)</i>		Actual Target Class	Not Target Class
Selected/ Guessed	True Positive (TP)	True Positive (TP)	False Positive (FP)
	False Negative (FN)	False Negative (FN)	True Negative (TN)