

CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024



HVERFORD
COLLEGE

Admin

- **Lab 2** grades & feedback posted on Moodle

Outline for today

- Evaluation Metrics
 - Confusion matrices
 - Precision and recall
 - ROC curves
- Introduction to probability

Outline for today

- Evaluation Metrics
 - Confusion matrices
 - Precision and recall
 - ROC curves
- Introduction to probability

Goals of Evaluation

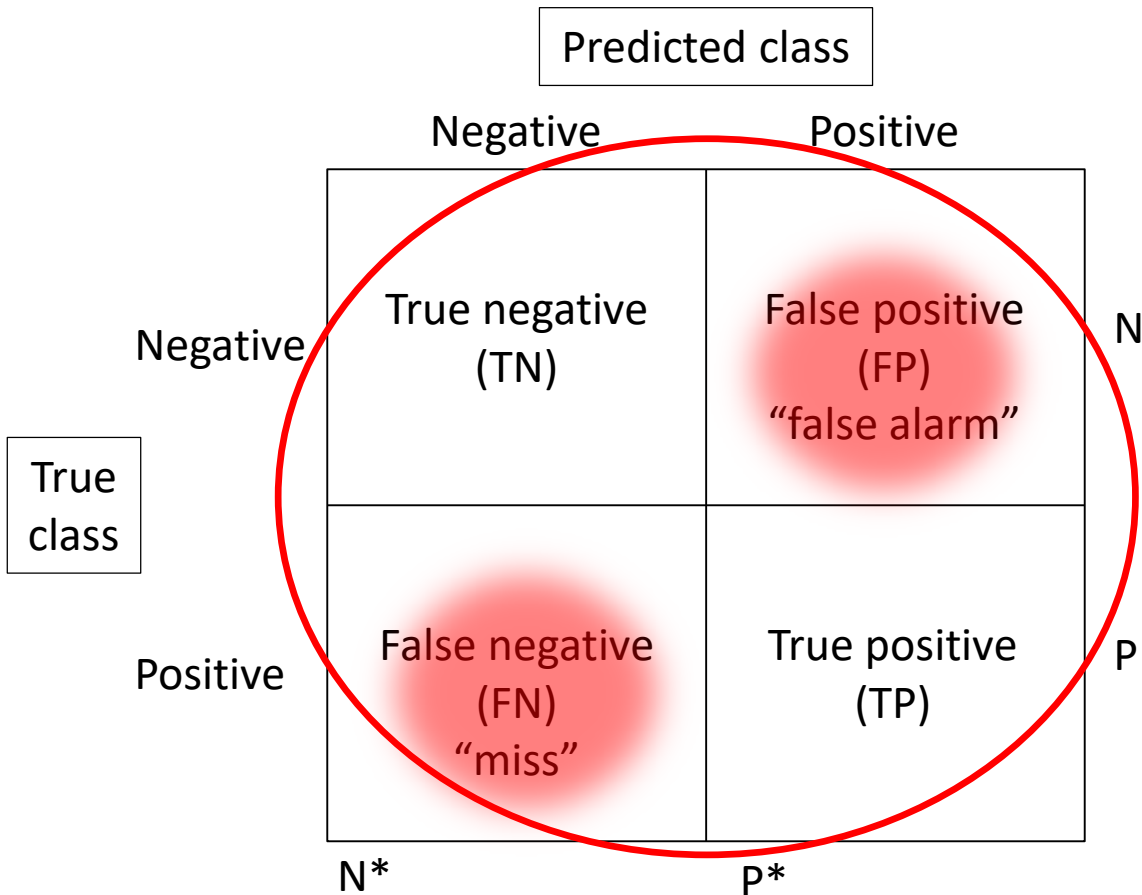
- Think about what metrics are important for the problem at hand
- Compare different methods or models on the same problem
- Common set of tools that other researchers/users can understand

Training and Testing

(high-level idea)

- **Separate** data into “**train**” and “**test**”
 - n = num training examples
 - m = num testing examples
- **Fit** (create) the model using **training data**
 - e.g. sea_ice_1979-2012.csv
- **Evaluate** the model using **testing data**
 - e.g. sea_ice_2013-2020.csv

Confusion Matrices

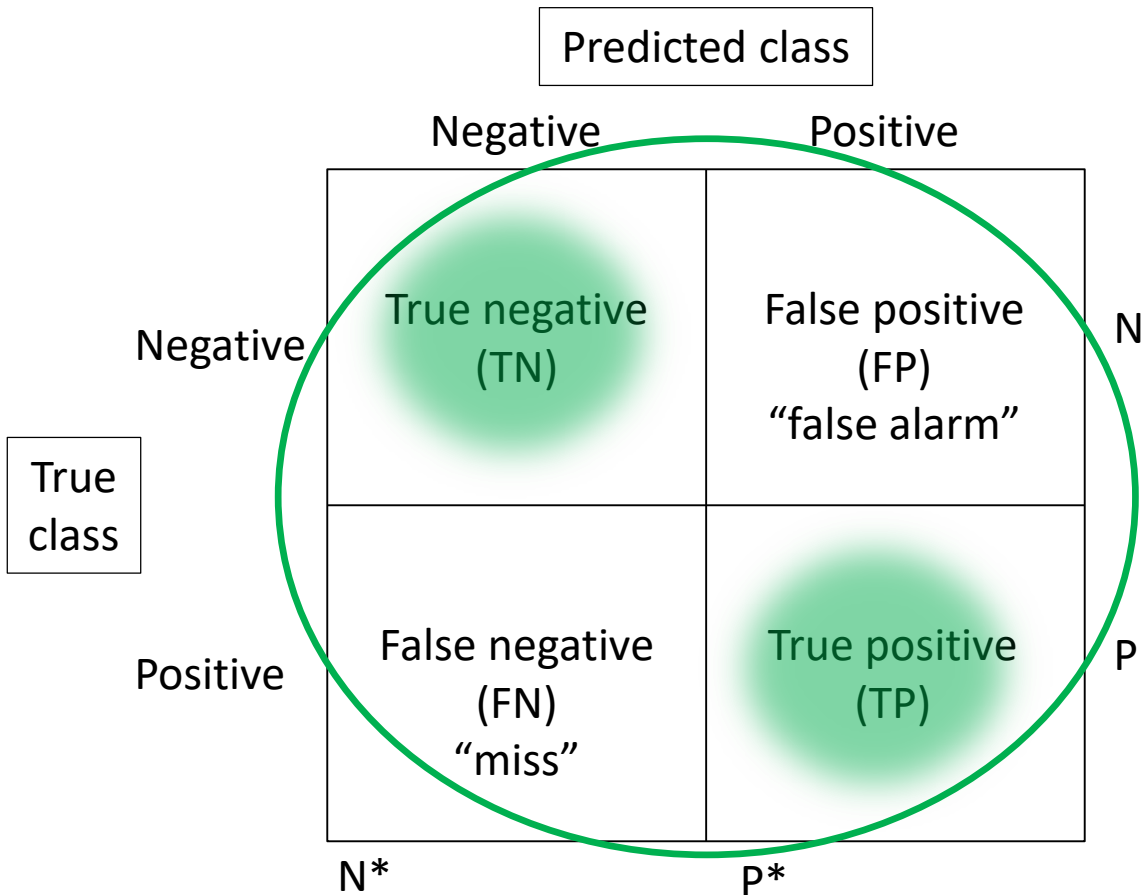


Error:

$$(FN+FP)/(TN+FP+FN+TP)$$

$$= (FN+FP)/(N+P)$$

Confusion Matrices



Accuracy = 1-Error:

$$(TN+TP)/(TN+FP+FN+TP)$$

$$= (TN+TP)/(N+P)$$

Confusion Matrices

		Predicted class		
		Negative	Positive	
True class	Negative	True negative (TN)	False positive (FP) "false alarm"	N
	Positive	False negative (FN) "miss"	True positive (TP)	P
		N*	p*	

Precision:

$$TP/(FP+TP) = TP/P^*$$

Confusion Matrices

		Predicted class		
		Negative	Positive	
True class	Negative	True negative (TN)	False positive (FP) "false alarm"	N
	Positive	False negative (FN) "miss"	True positive (TP)	P
		N*	p*	

Recall
(True Positive Rate):

$$TP/(FN+TP) = TP/P$$

Confusion Matrices

		Predicted class		
		Negative	Positive	
True class	Negative	True negative (TN)	False positive (FP) "false alarm"	N
	Positive	False negative (FN) "miss"	True positive (TP)	P
		N*	p*	

False Positive Rate:

$$FP/(TN+FP) = FP/N$$

Precision and Recall

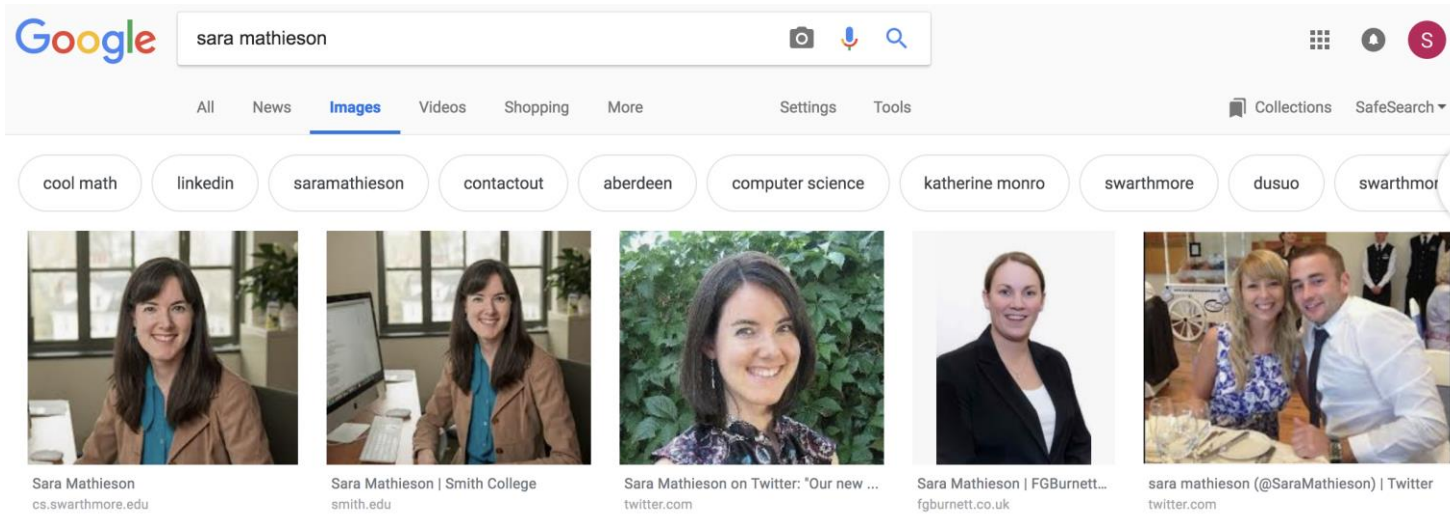
- Precision: of all the “flagged” examples, which ones are actually relevant (i.e. positive)?

(Purity)

- Recall: of all the relevant results, which ones did I actually return?

(Completeness)

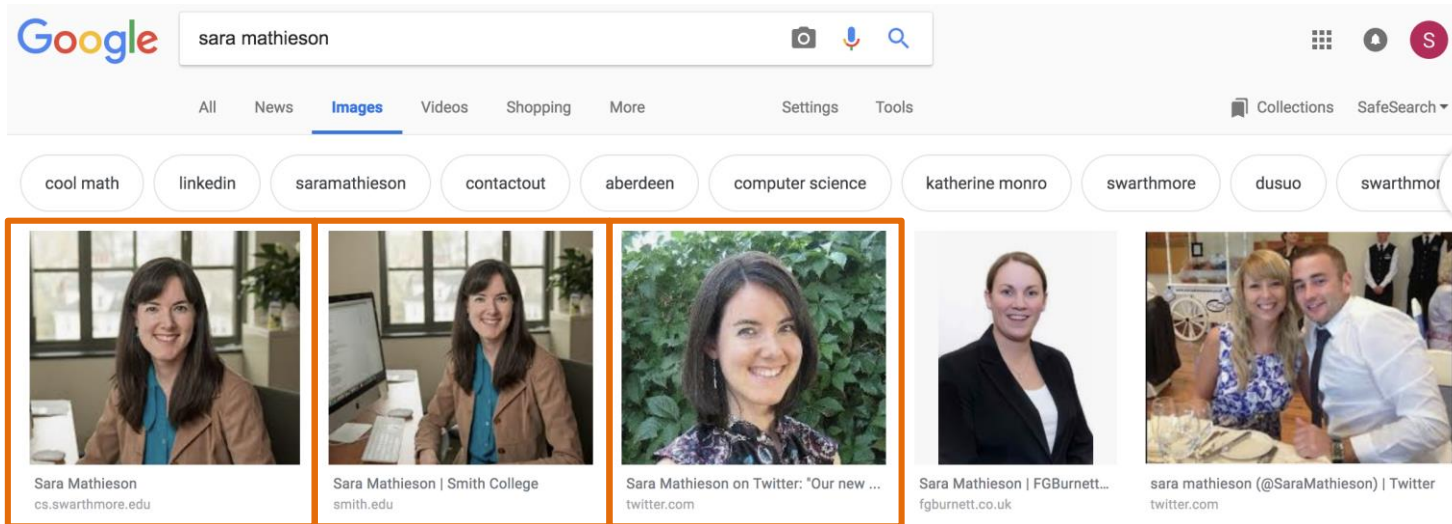
Precision and Recall



$P=6$ (number of images that are actually Sara)

- Precision?
- Recall?

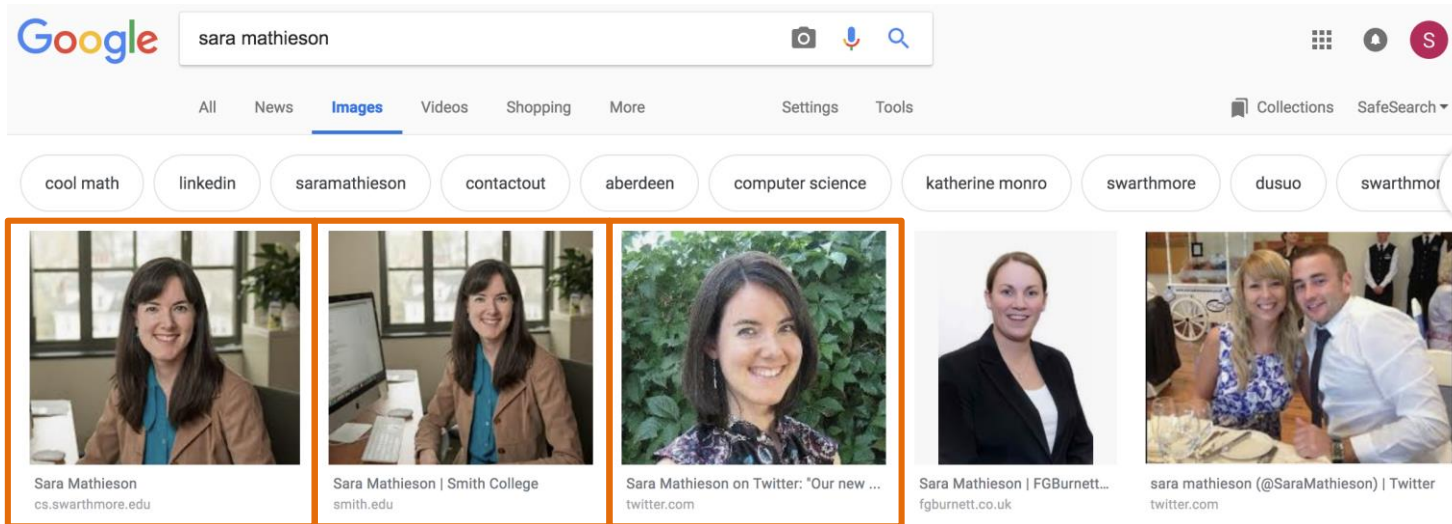
Precision and Recall



$P=6$ (number of images that are actually Sara)

- Precision = $TP/(FP+TP) = 3/5$
- Recall?

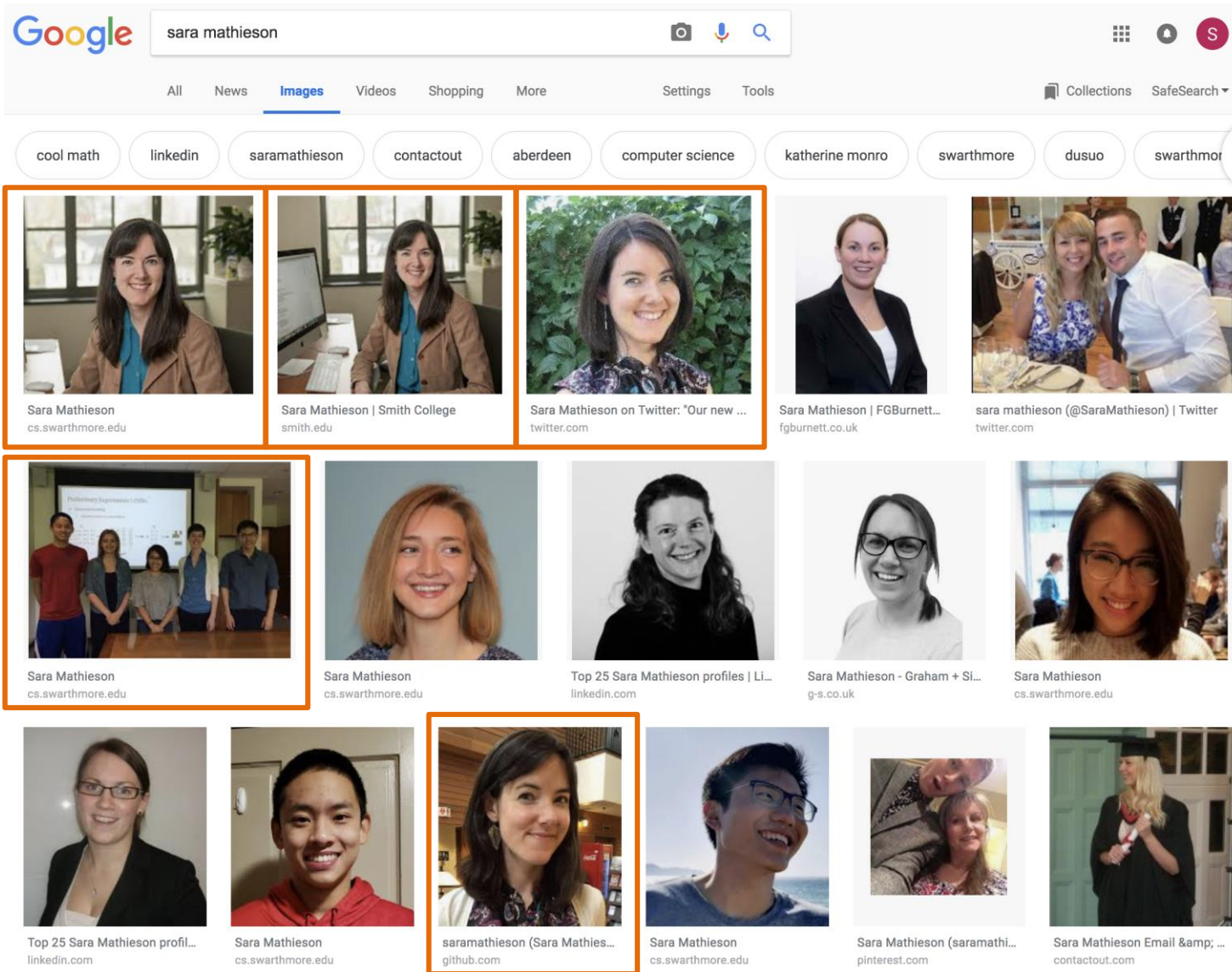
Precision and Recall



$P=6$ (number of images that are actually Sara)

- Precision = $TP/(FP+TP) = 3/5$
- Recall = $TP/(FN+TP) = 3/6$

Precision and Recall

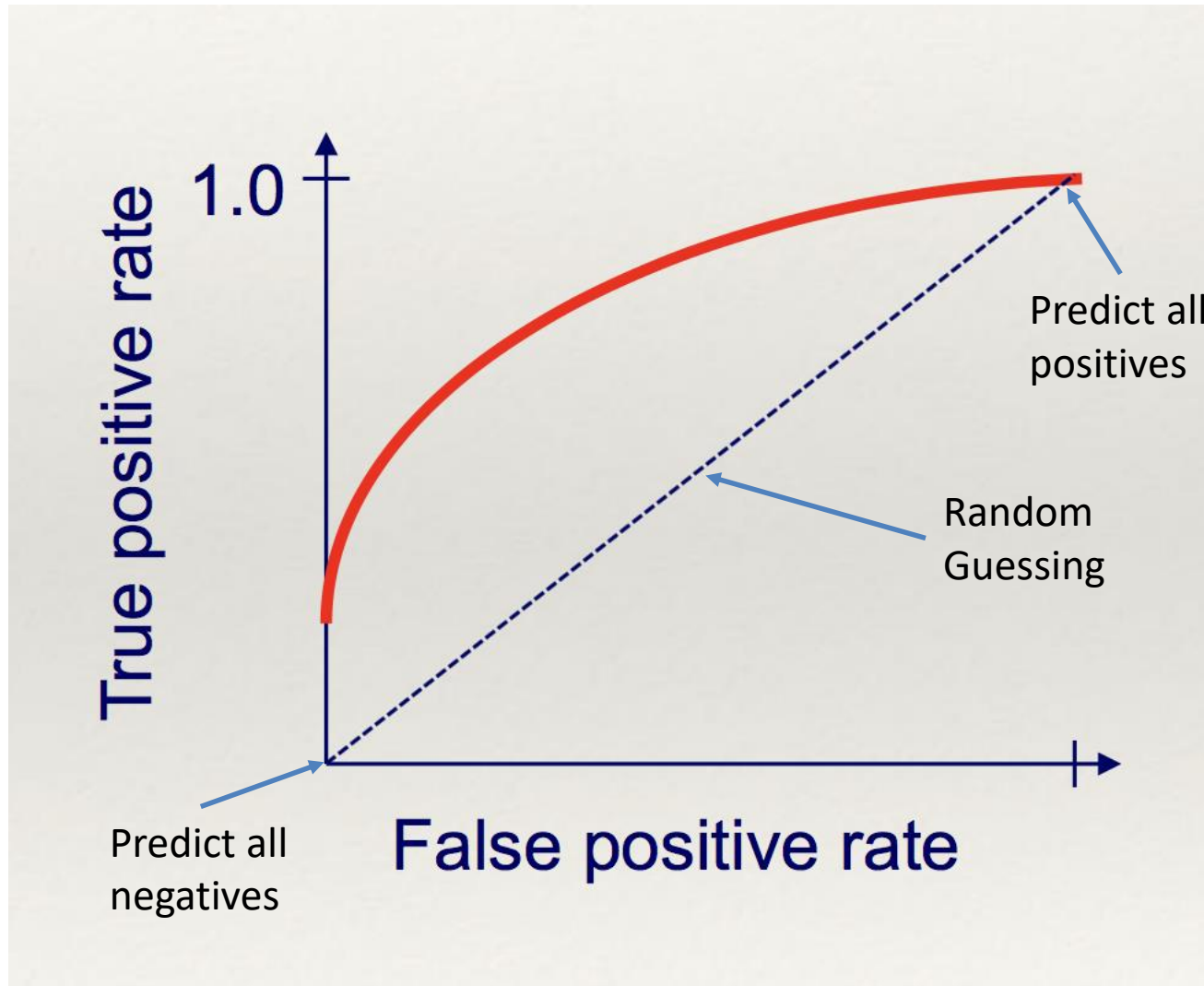


$P=6$ (number of images that are actually Sara)

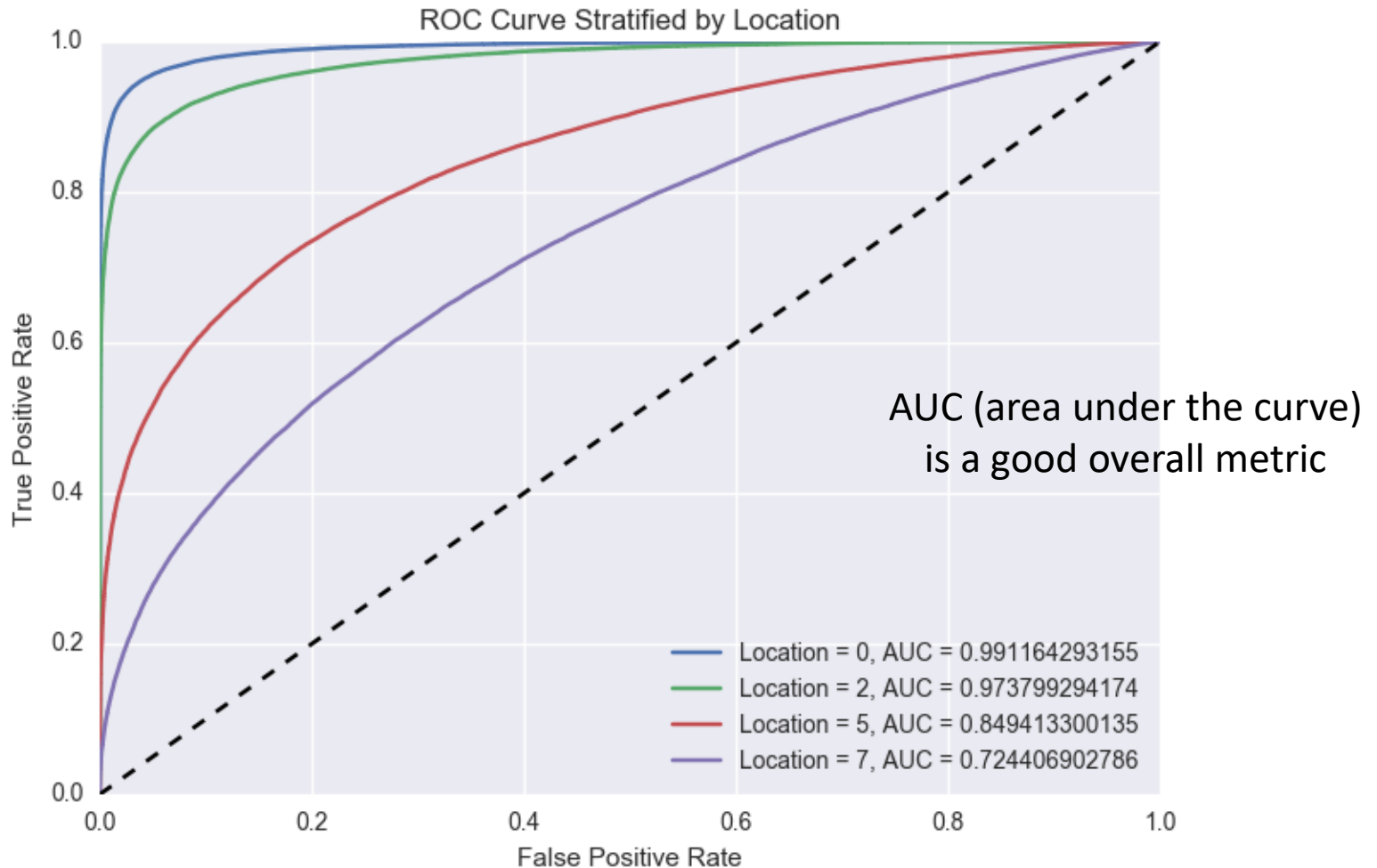
- Precision = $5/16$
- Recall = $5/6$

ROC curve (Receiver Operating Characteristic)

More history here! https://en.wikipedia.org/wiki/Receiver_operating_characteristic



ROC curve example: comparing methods



Example of a ROC curve

Chan, Perrone, Spence, Jenkins, Mathieson, Song

How to get a ROC curve for probabilistic methods?

- Usually we use 0.5 as a threshold for binary classification
- Vary the threshold! (i.e. choose 0, 0.1, 0.2,...)
 - $P(y=1 \mid x) \geq 0.2$ \Rightarrow classify as 1 (positive)
 - $P(y=1 \mid x) < 0.2$ \Rightarrow classify as 0 (negative)

Handout 8

Handout 8

	-	+	
-	77	3	$N = 80$
+	13	7	$P = 20$

$$N^* = 90 \quad P^* = 10$$

$$\text{precision} = \frac{7}{10}$$

$$\text{recall} = \frac{7}{20} = 0.35$$

$$\text{FPR} = \frac{3}{80}$$

"X"

"Y"

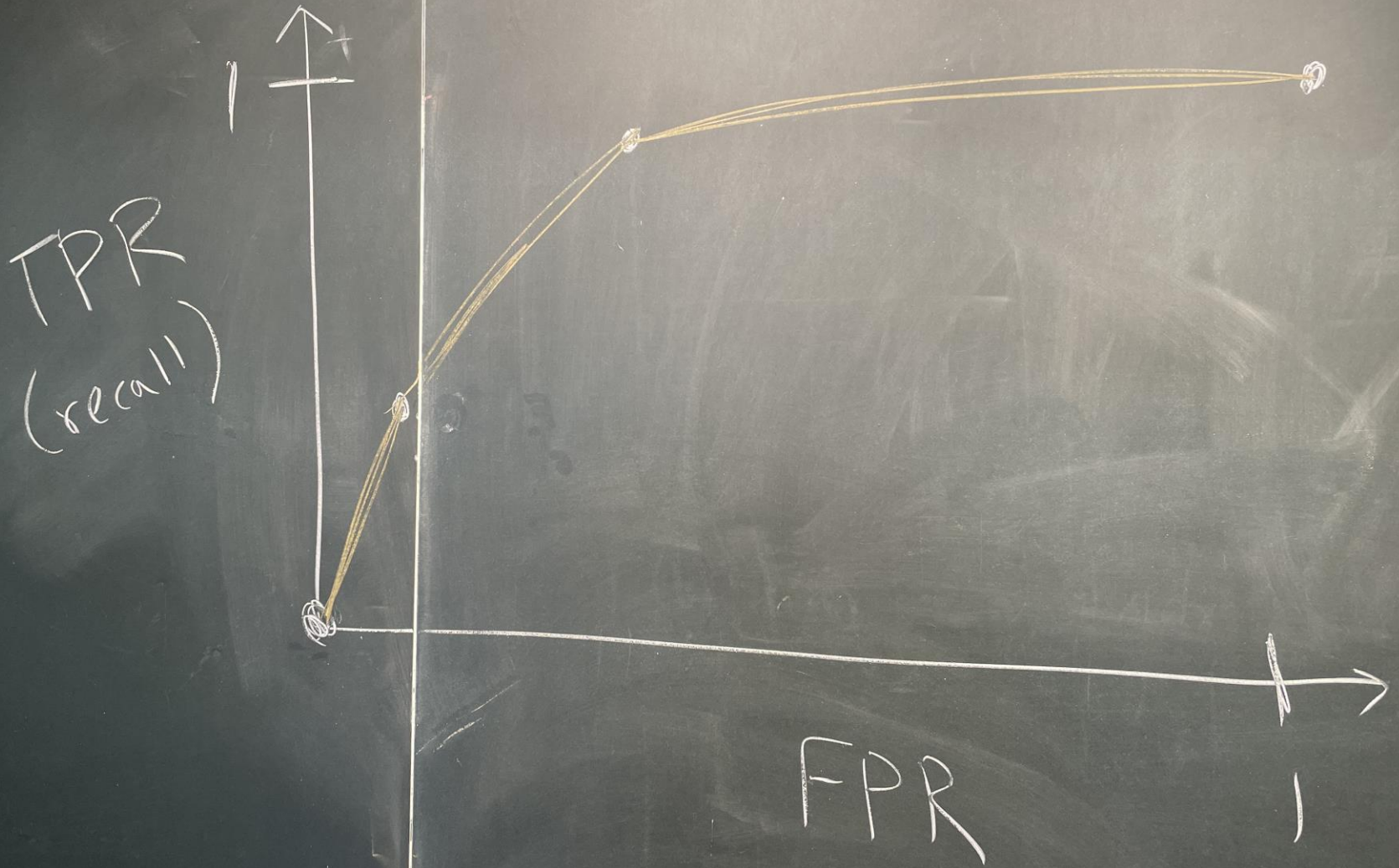
68	12	
2	18	$P = 20$

$$P^* = 30$$

$$\text{TPR} = 18/20 = 0.9$$

$$\text{FPR} = 12/80 = 0.15$$

Handout 8



Outline for today

- Evaluation Metrics
 - Confusion matrices
 - Precision and recall
 - ROC curves
- Introduction to probability

Intro to Probability

- The **probability** of an **event** e has a number of epistemological interpretations
- Assuming we have **data**, we can count the number of times e occurs in the dataset to estimate the probability of e , $P(e)$.

$$P(e) = \frac{\text{count}(e)}{\text{count}(\text{all events})}.$$

- If we put all events in a bag, shake it up, and choose one at random (called **sampling**), how likely are we to get e ?

Intro to Probability



- Suppose we flip a fair coin
- What is the probability of heads, $P(e = H)$?

Intro to Probability



- Suppose we flip a fair coin
- What is the probability of heads, $P(e = H)$?
- We have "all" of two possibilities, $e \in \{H, T\}$.
- $$P(e = H) = \frac{\text{count}(H)}{\text{count}(H) + \text{count}(T)}$$

Intro to Probability



- Suppose we have a fair 6-sided die.
- What's the probability of getting "1"?

Intro to Probability



- Suppose we have a fair 6-sided die.
- What's the probability of getting "1"?

$$\frac{\textit{count}(s)}{\textit{count}(1) + \textit{count}(2) + \textit{count}(3) + \cdots + \textit{count}(6)} = \frac{1}{1 + 1 + 1 + 1 + 1 + 1} = \frac{1}{6}$$