CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2025

HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

- **Sit somewhere new**
- **Lab 3** due tonight at midnight
- **Lab 4** posted (due next Tuesday)
  - pair-programming required, **different partner**

CAMPUS READ 2025

A special book talk and community discussion with marine biologist, policy expert, and writer, Dr. Ayana Elizabeth Johnson.

[RSVP here]

Tuesday, September 30

7:00 p.m.

Roberts Hall, Marshall Auditorium

A student-only book signing will be held in Lutnick 200 from 6:15–6:45!

Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Stochastic Gradient Descent for Linear Regression

Key Idea: take the derivative of **one datapoint** at a time and use that to update w

set **w = 0** vector

while **cost J(w)** still changing (or max iter reached):

    shuffle data points

    for i = 1…n:

        w <- **w – alpha(derivative of J(w) wrt xi)**

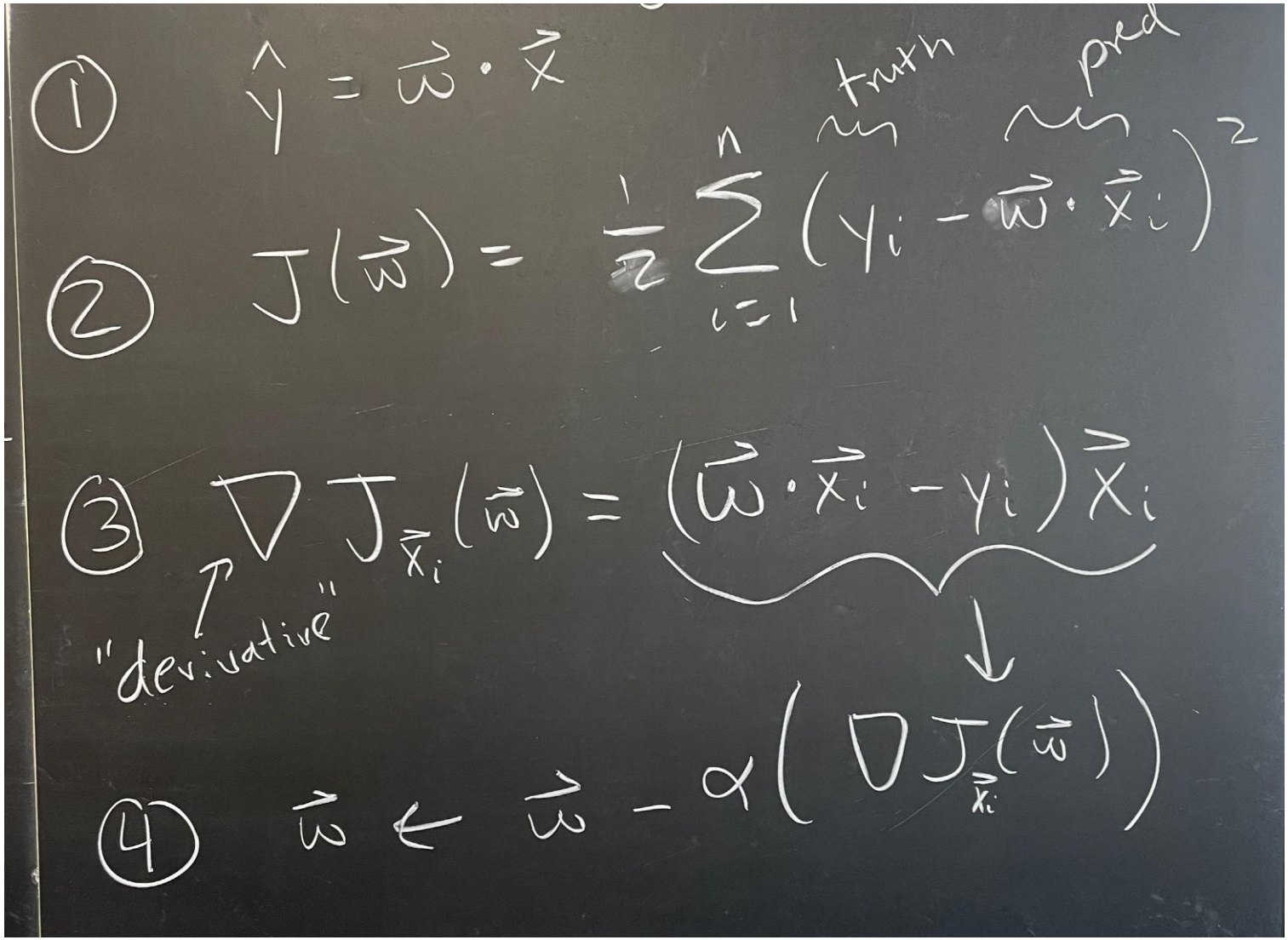
    store J(w)

Mini-quiz for linear regression

For each of the following terms/descriptions, write out the corresponding equation:

1) Linear regression **model**
2) Linear regression **cost function**
3) **Gradient** of cost function wrt one datapoint
4) **Gradient descent** weight vector update

Mini-quiz for linear regression



Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

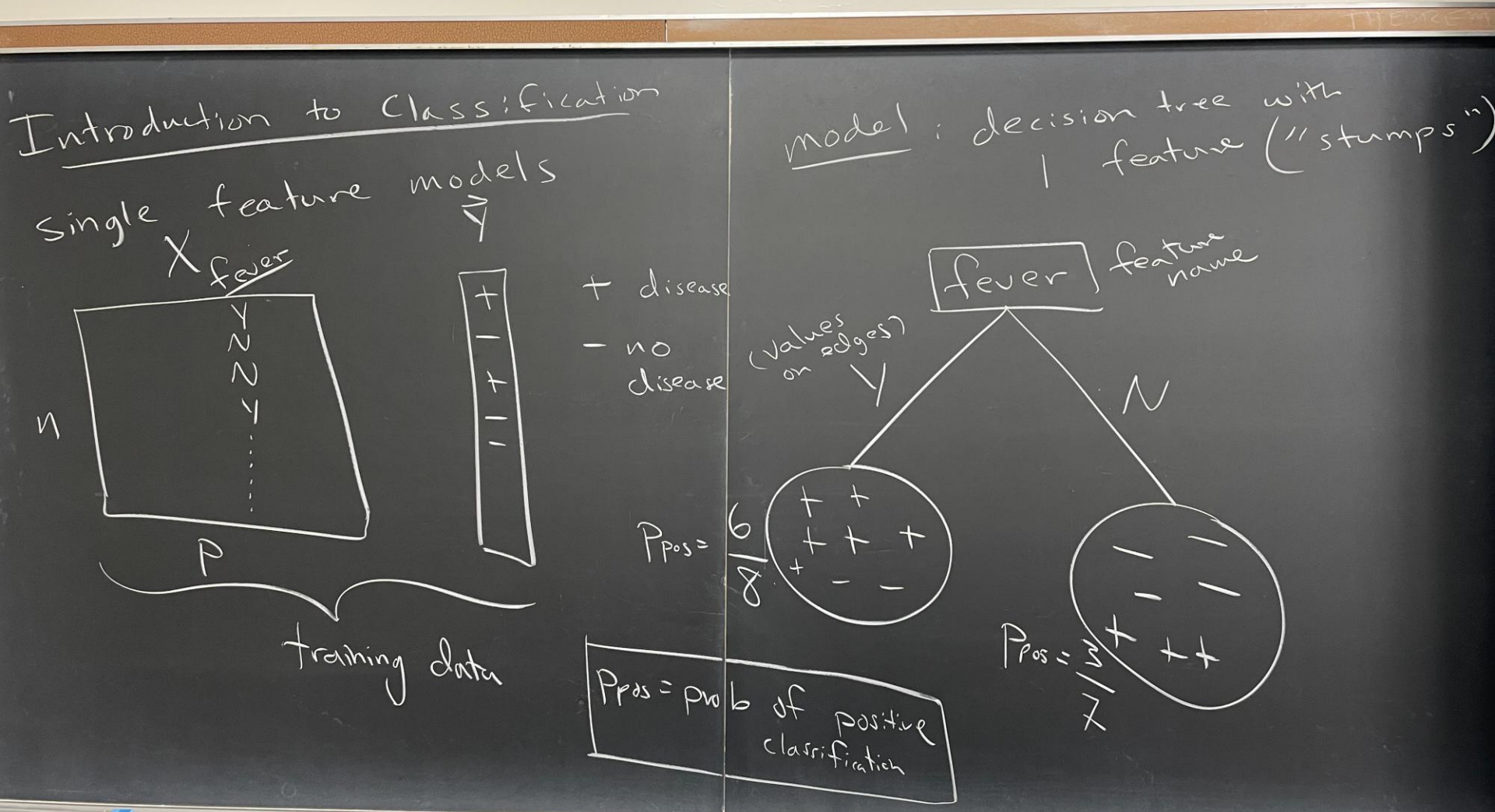
- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Binary classification examples

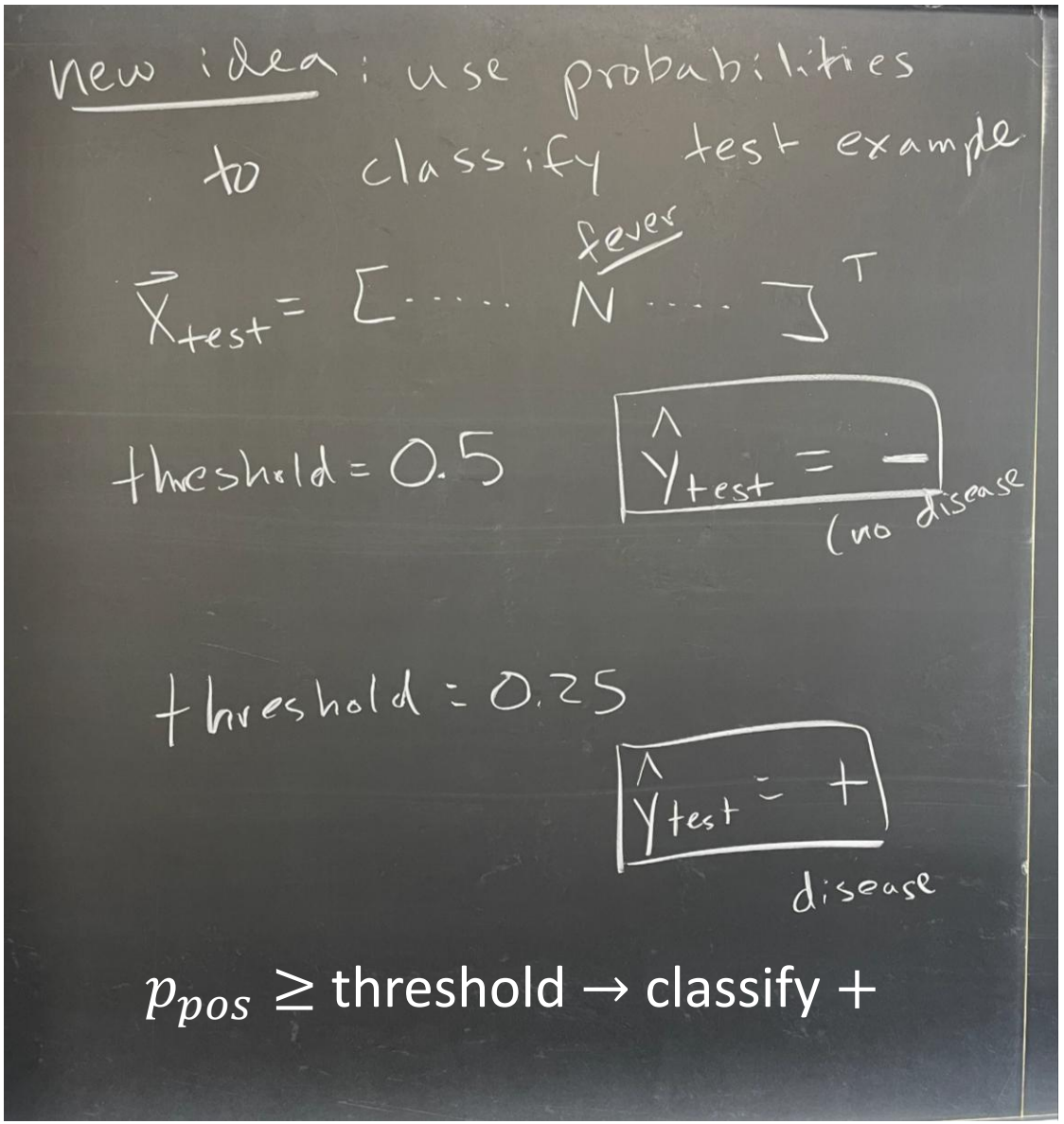
- Transactions that indicate credit card fraud
- Accounts that are bots
- Detecting which scans show tumors
- Prenatal test for Down's Syndrome
- Finding genes under natural selection
- Regions of the environment that contains the object the robot is searching for

In all these examples, we are trying to find unusual items ("needle in a haystack") -- we call these *positives*

Introduction to Classification

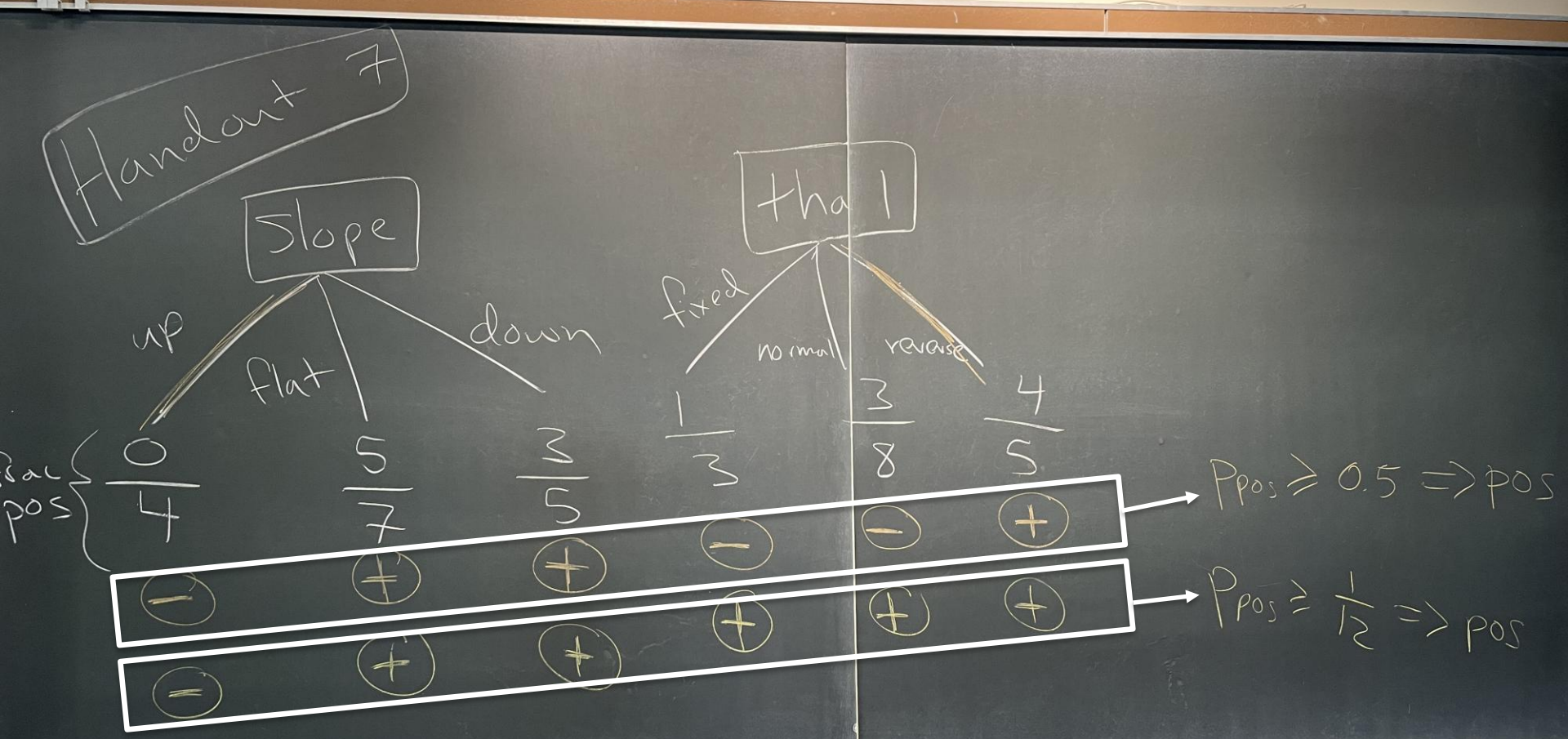


Introduction to Classification



Handout 7

Handout 7


Handout 7
Slope
up
flat
down
thal
fixed
normal
reverse
0/4
5/7
3/5
1/3
3/8
4/5
pos
$P_{pos} \geq 0.5 \Rightarrow$ pos
$P_{pos} \geq \frac{1}{2} \Rightarrow$ pos

Outline for today

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Goals of Evaluation

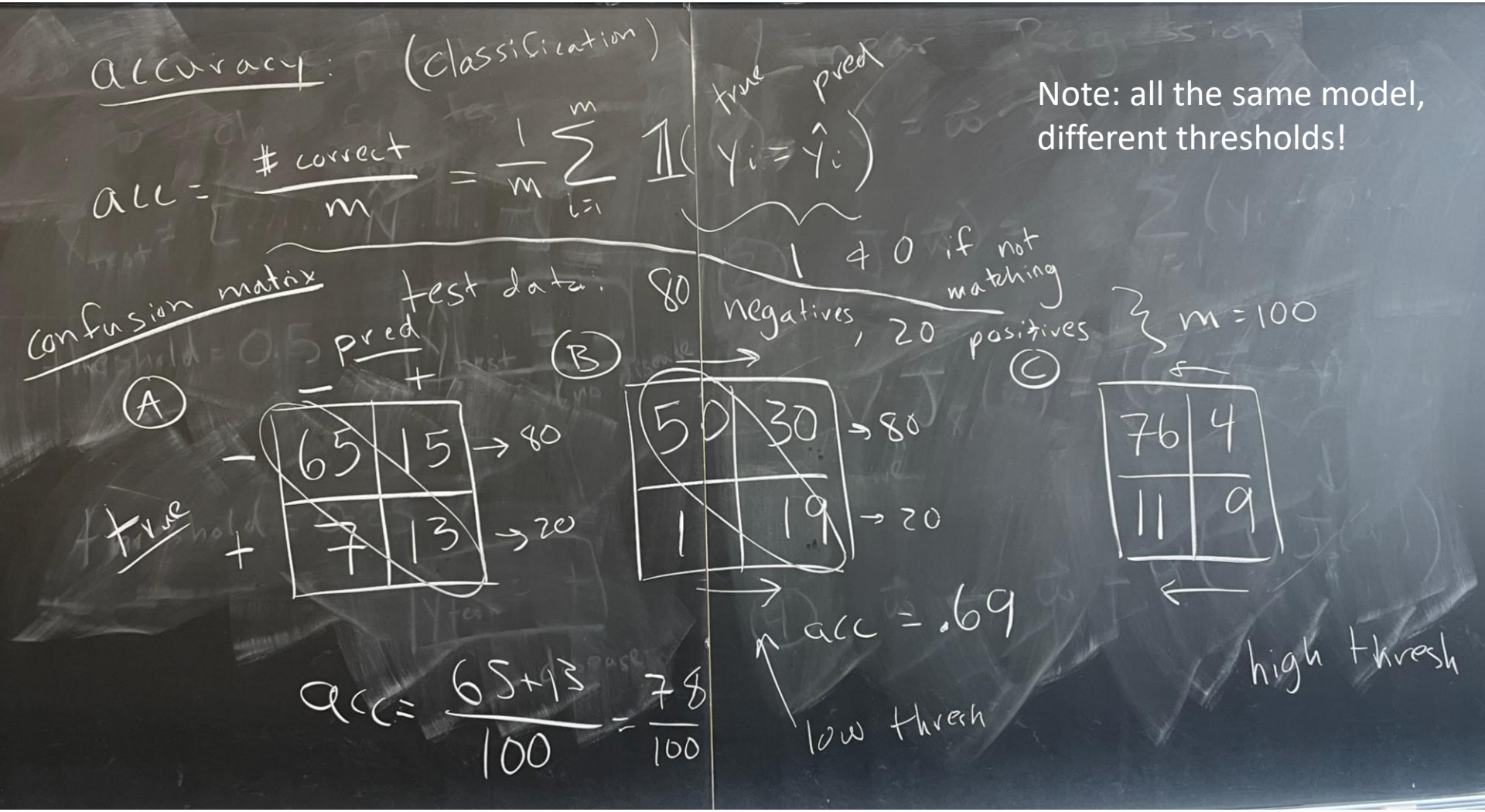
- Think about what metrics are important for the problem at hand
- Compare different methods or models on the same problem
- Common set of tools that other researchers/users can understand

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

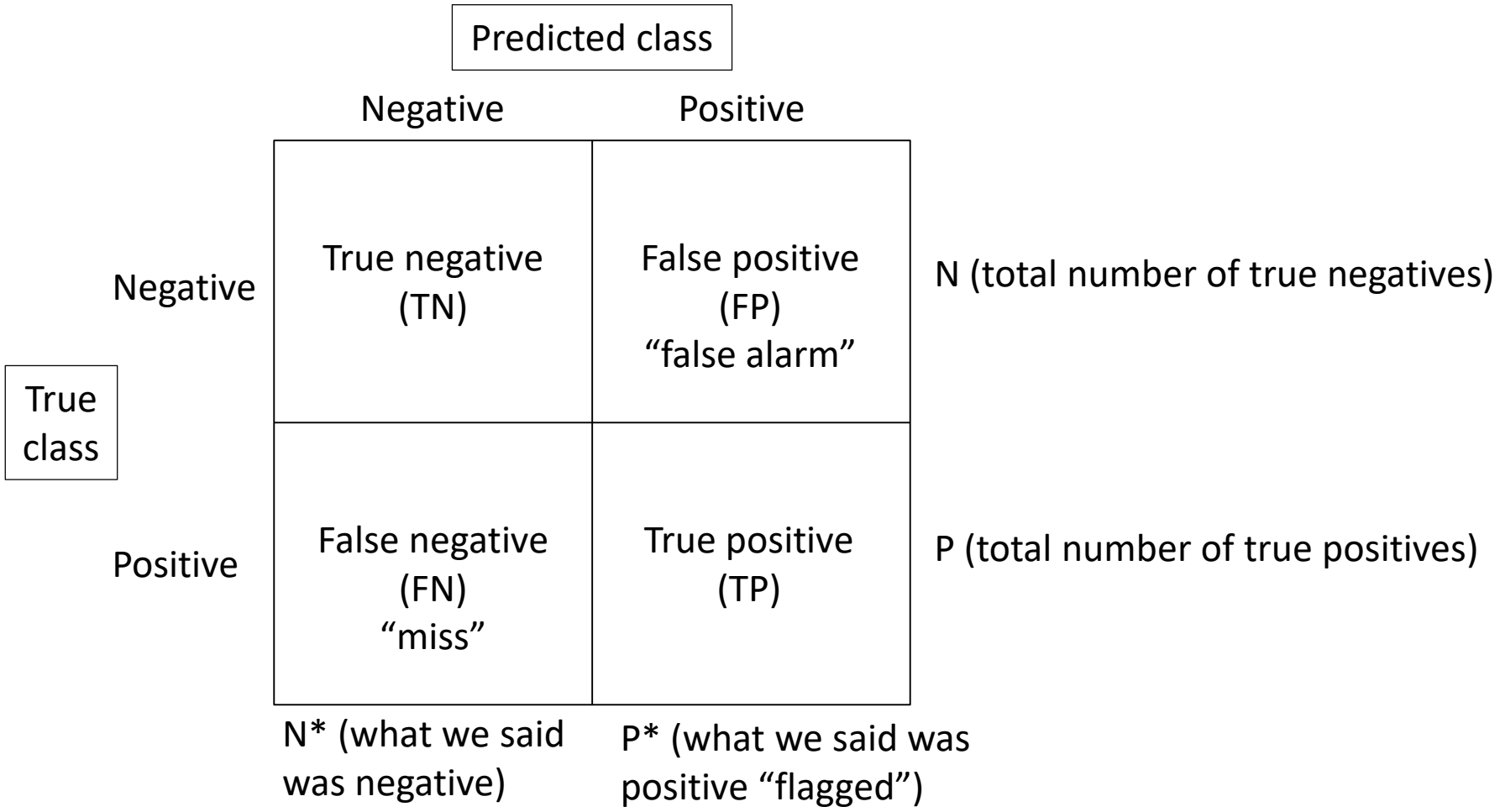


Note: all the same model, different thresholds!

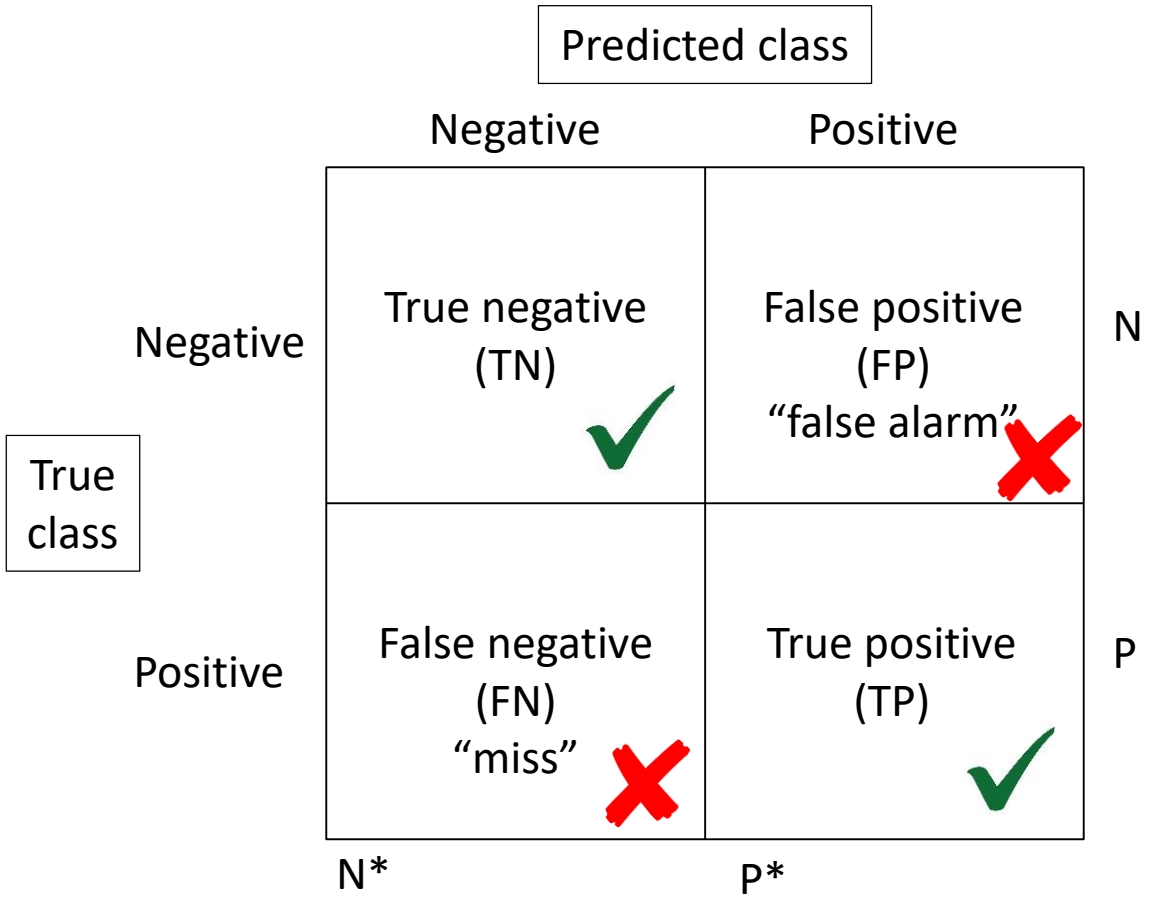
Confusion Matrices

True class \ Predicted class	Negative	Positive
Negative	True negative (TN)	False positive (FP)
Positive	False negative (FN)	True positive (TP)

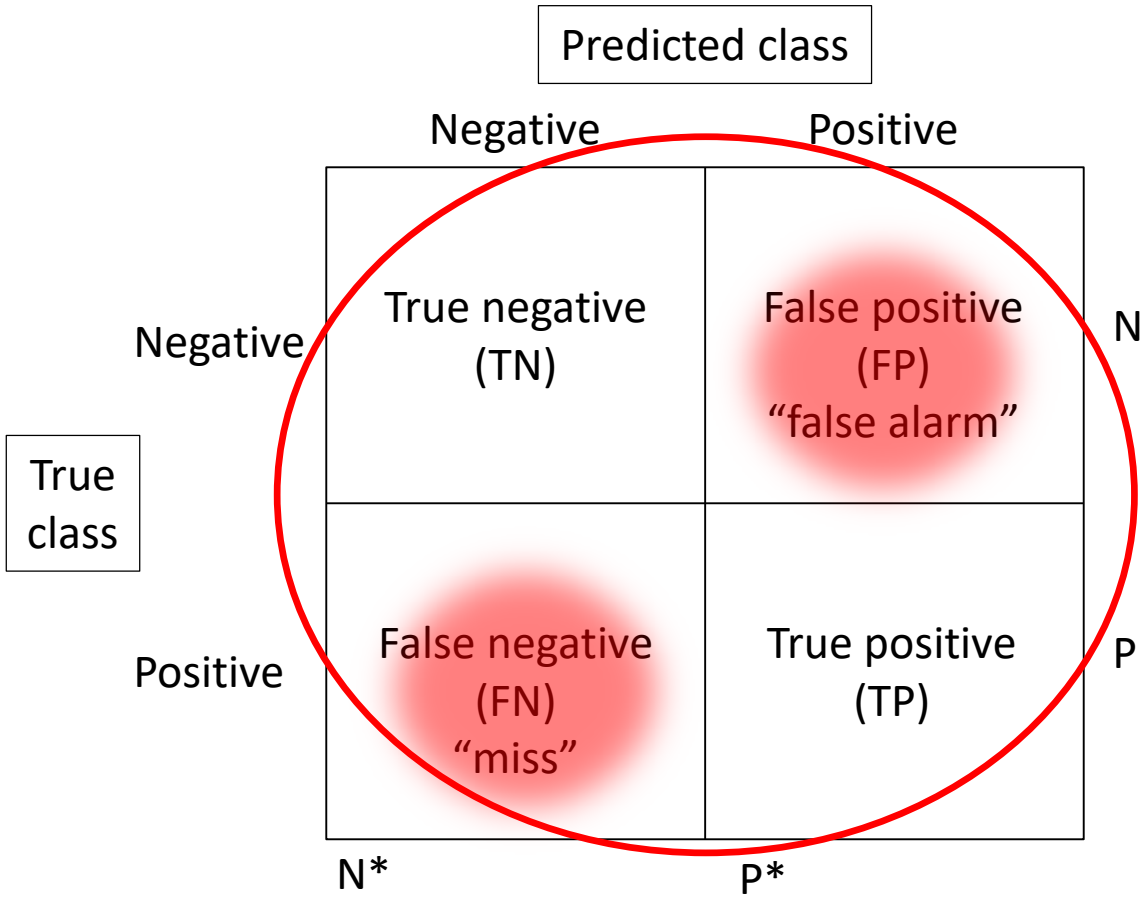
Confusion Matrices



Confusion Matrices



Confusion Matrices

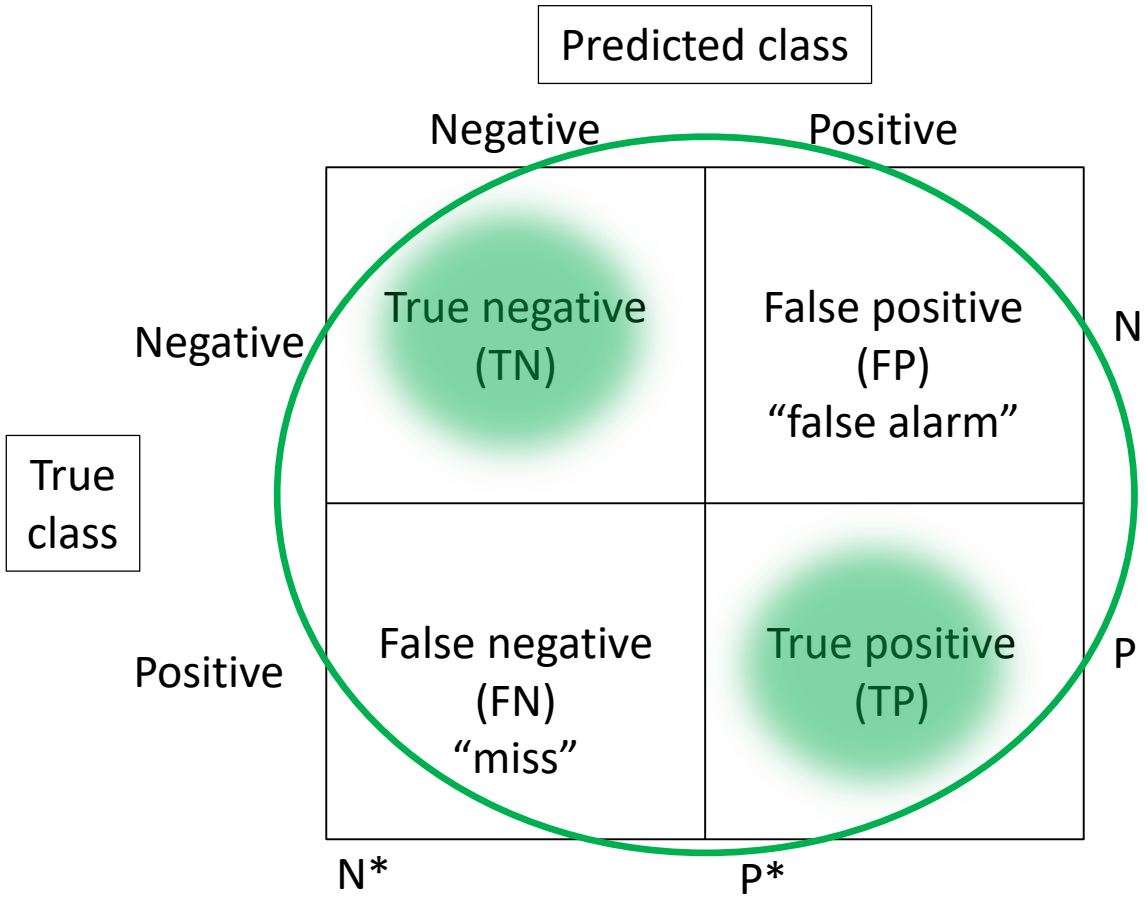


Error:

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

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

Confusion Matrices

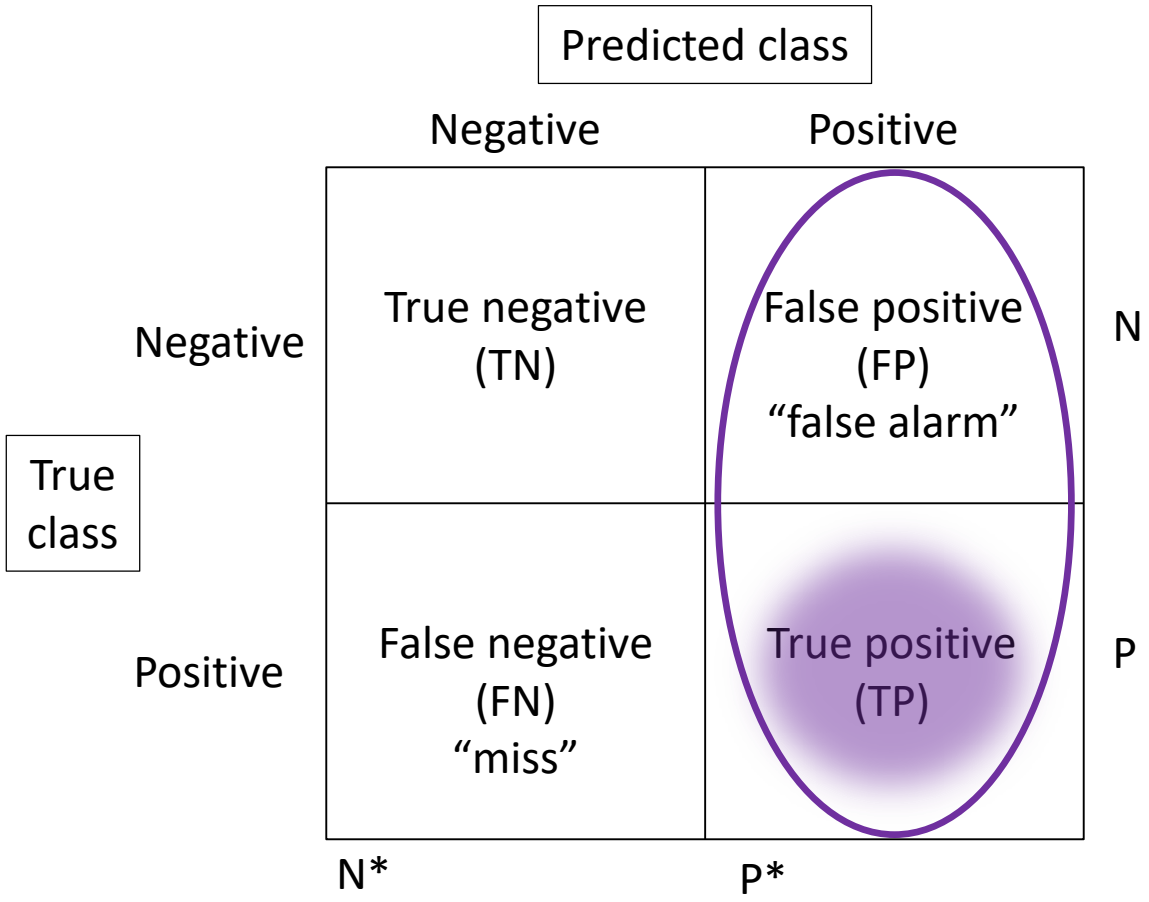


Accuracy = 1-Error:

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

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

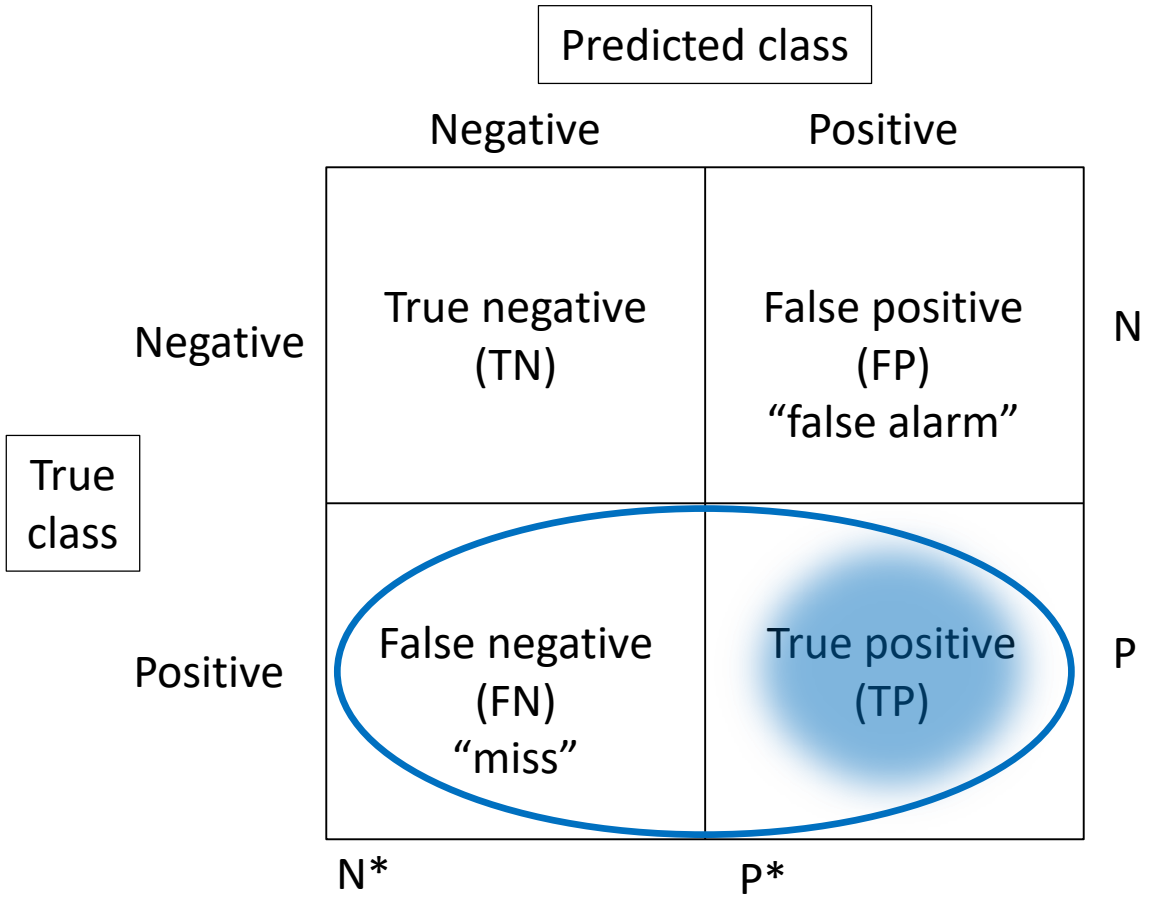
Confusion Matrices



Precision:

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

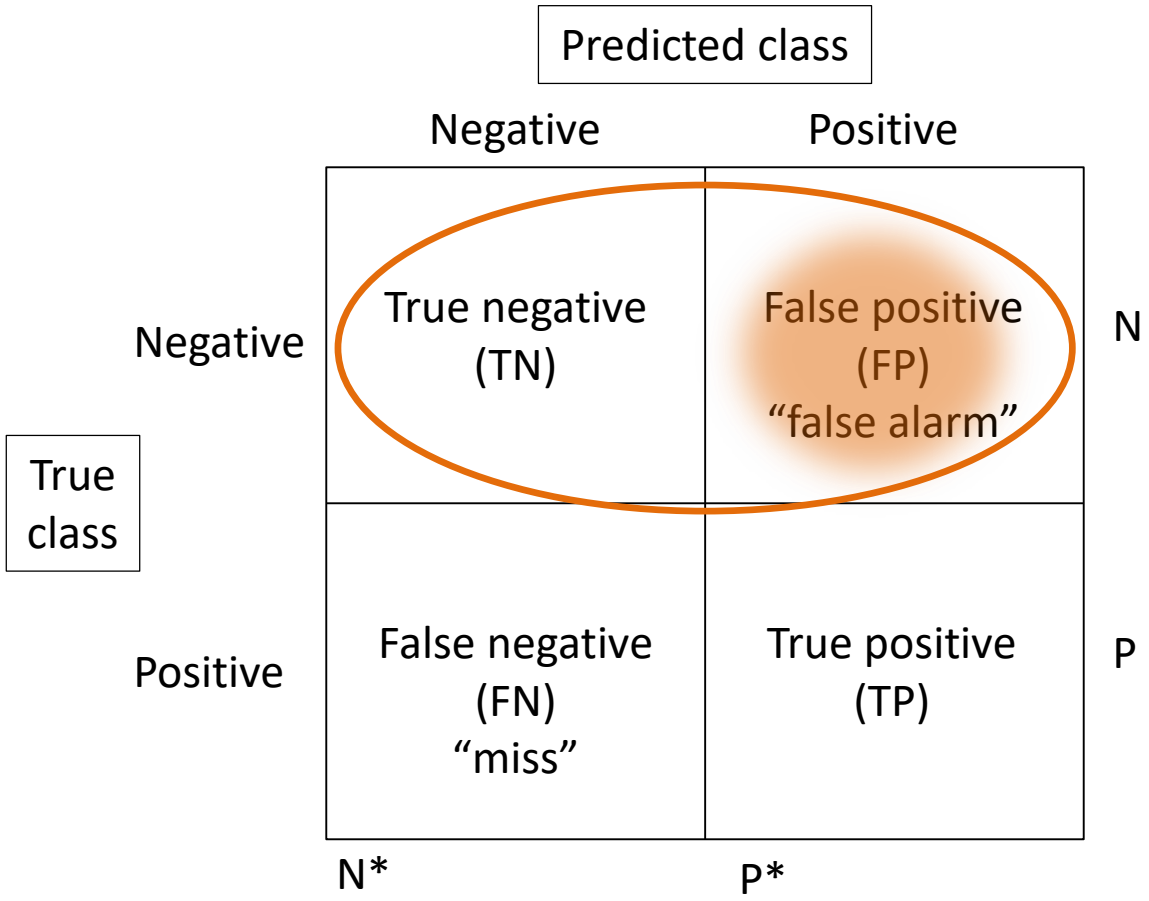
Confusion Matrices



Recall
(True Positive Rate):

TP/(FN+TP) = TP/P

Confusion Matrices



False Positive Rate:

FP/(TN+FP) = FP/N

Precision and Recall

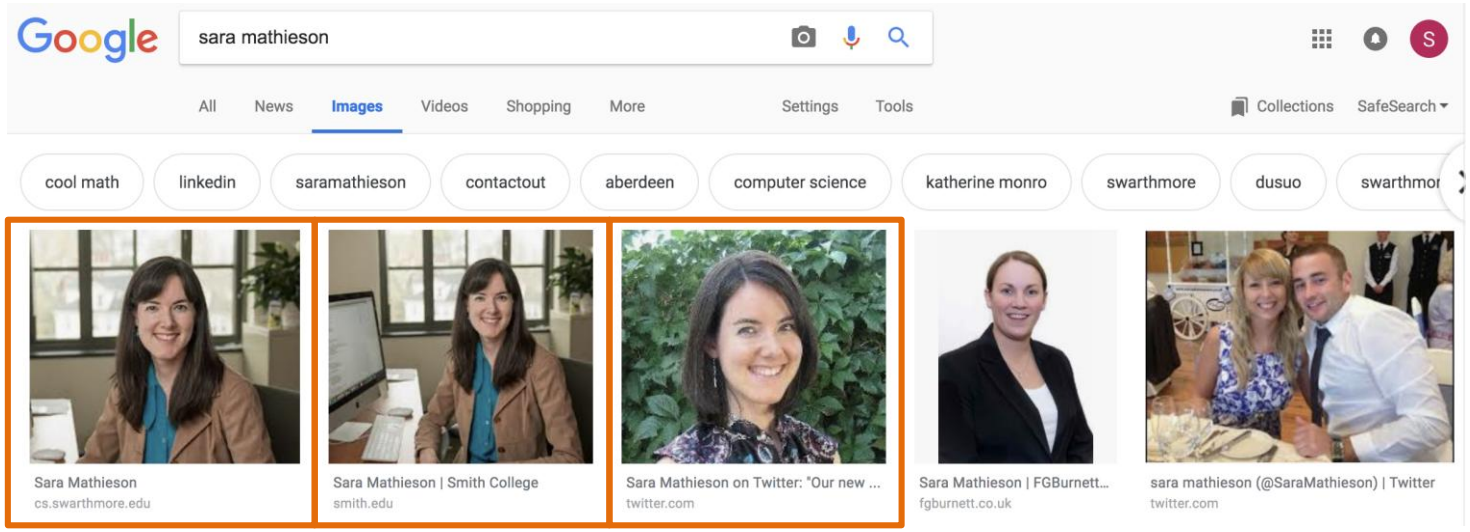
- Precision: of all the "flagged" examples, which ones are actually relevant (i.e., positive)?

  (Purity)

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

  (Completeness)

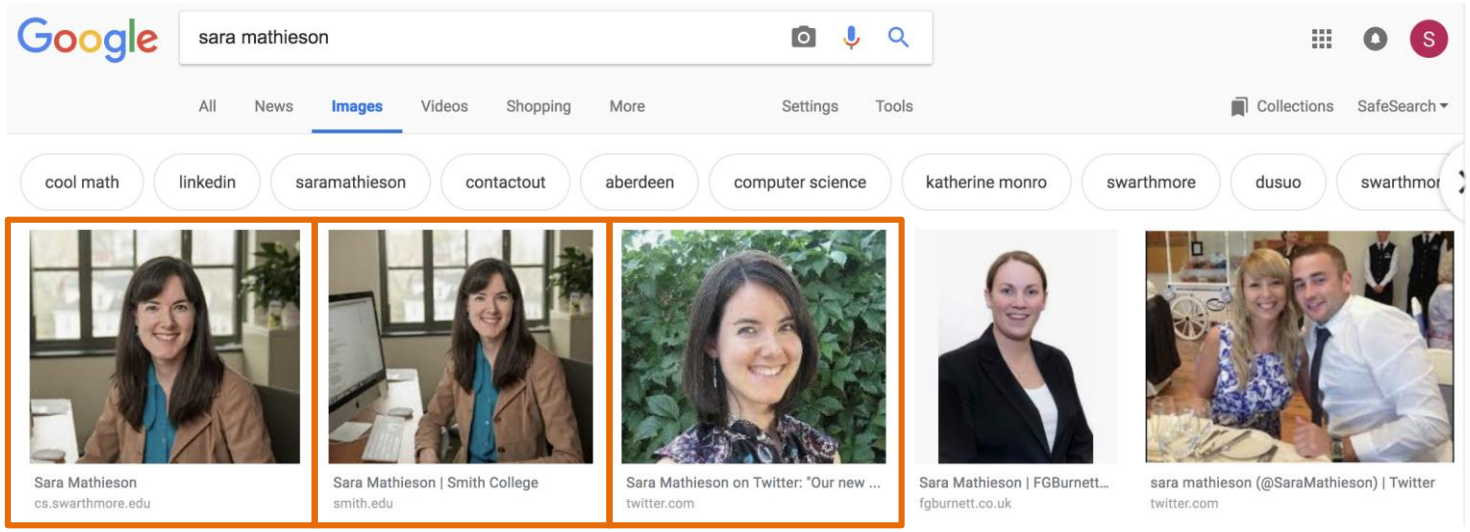
Precision and Recall



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

- Precision?
- Recall?

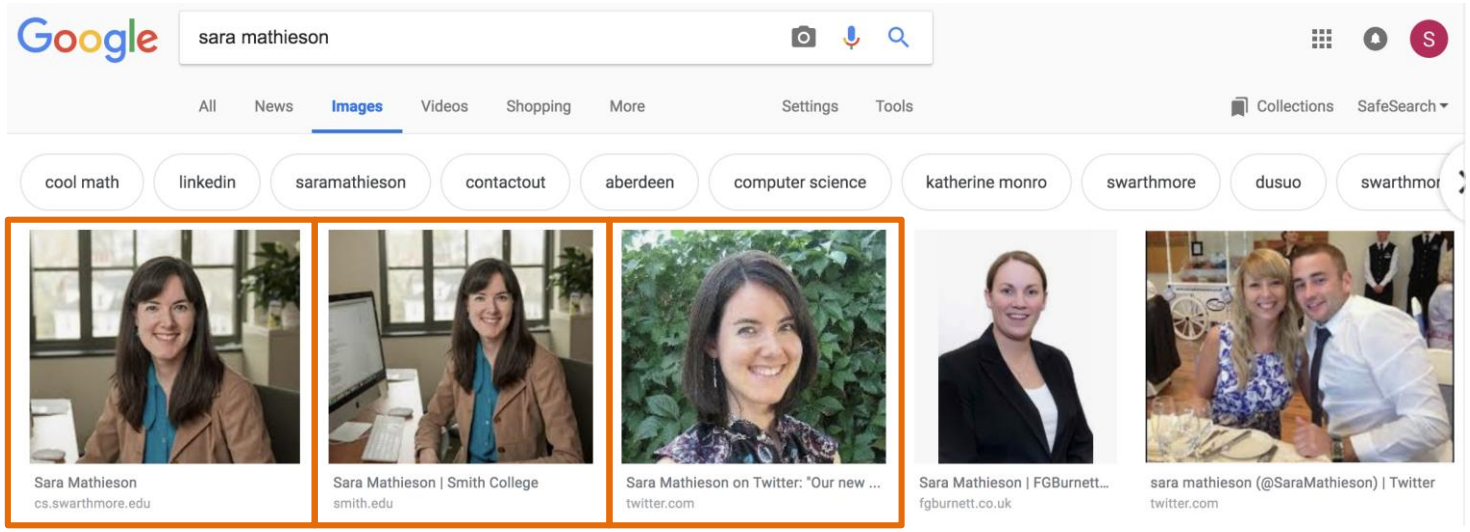
Precision and Recall



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

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

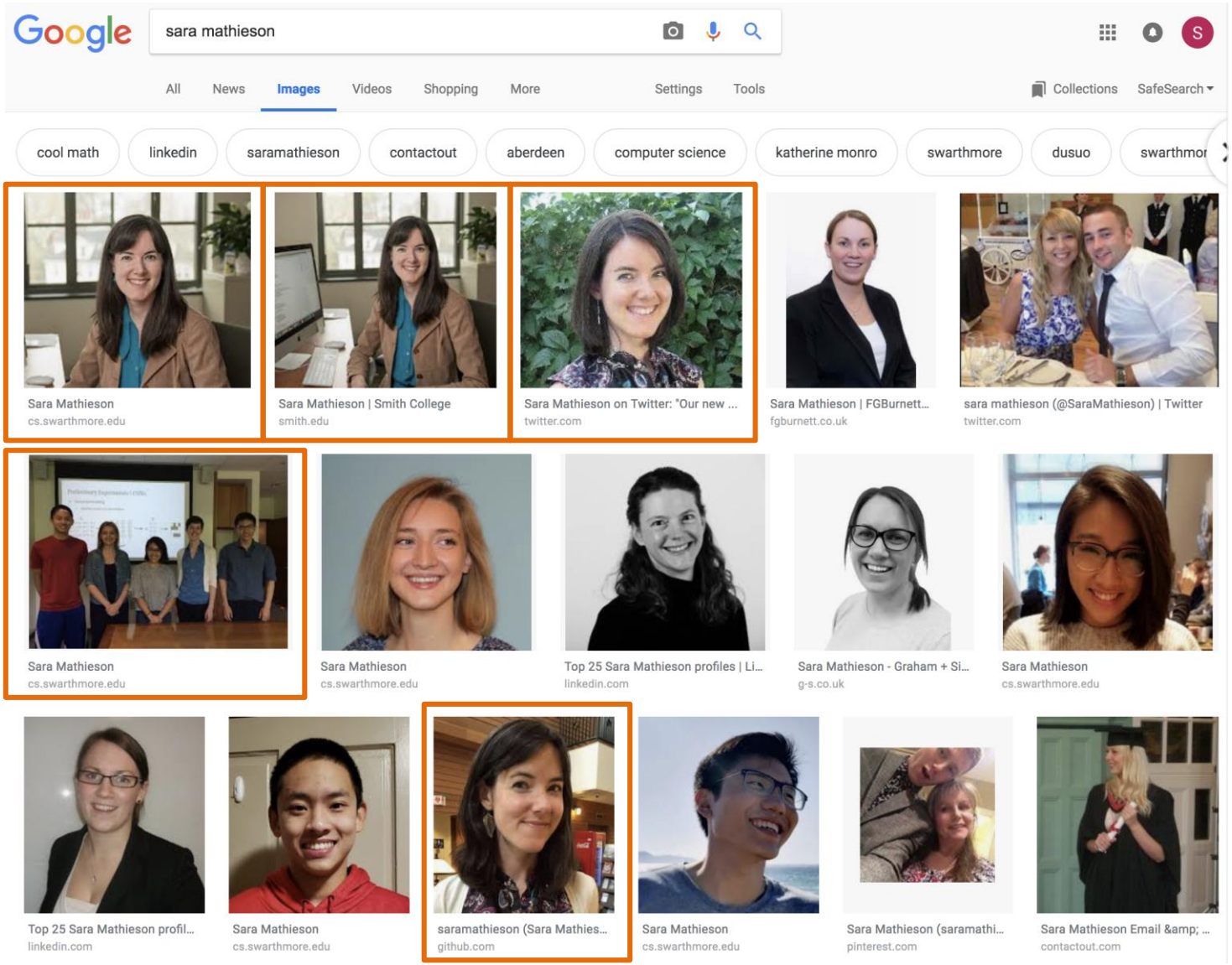
Precision and Recall



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

Modified from Ameet Soni

Outline for today

- Recap SGD (stochastic gradient descent)
- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation
- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Next time!