CSE/STAT 416

Classification

Amal Nanavati
University of Washington
July 11, 2022

Adapted from Hunter Schafer's Slides

Administrivia

We have now finished the "Regression" component of the course!

Next two weeks (4 lectures): Classification

HW2 due tomorrow 11:59PM

- Up to Thurs 7/14 11:59PM if you use late days

Reminder the NO HOMEWORK will be accepted more than 2 days late!

- Note that late days cause you to start the next assignment late.

HW3 released Wed

Lookahead: HW4 Programming is the **only** assignment that allows groups (up to 2)

- Ed post forthcoming with details about group formation.
- HW4 Concept is still individual

Exciting activity today on ethics, bias, and social impact in ML, led by our TA Karman!

- Last 30 mins of lecture.

Responding to Learning Reflection Questions

Standard Scaler (normalization)

normalization = standardization

Why do we fit the scaler only on the training set?

Q1: Why do we apply the transformation from the train set to the validation/test set?

- Say the “bedroom” feature of the **train set** had a mean of 3 and standard deviation of 1.5.
- Say the “bedroom” feature of the **validation set** had a mean of 5 and standard deviation of 2.
- A standardized value of “6” for the “bedroom” column of the **train set** would correspond to a house with 11 bedrooms.
- A standardized value of “6” for the “bedroom” column of the **validation set** would correspond to a house with 17 bedrooms.
- A model trained on the train set would perform poorly on the validation set because the input values represent different real-world quantities!

Takeaway: whatever transformations you do on the train set must be directly mimicked on the validation/test sets!

Why do we fit the scaler only on the training set?

Q2: Why do we not compute the mean and standard deviation of the whole dataset, as opposed to just the train set?

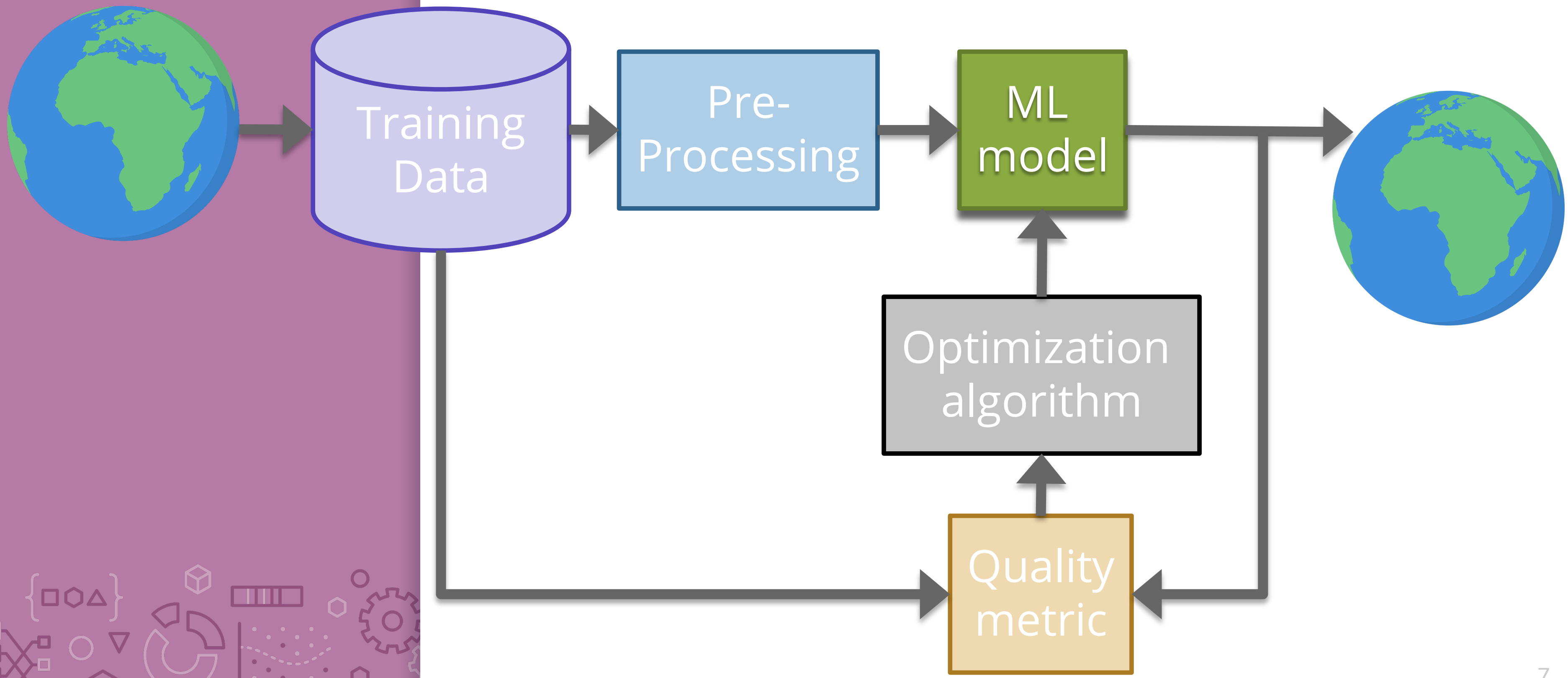
- If we used the mean and standard deviation on the whole dataset, train set values would be ”informed” or “ “influenced” by the test set.
- This violates the principle of only using the test set at the end.

Misc. LR Uncertainties

Does coefficient size indicate feature importance?

- Yes, but with two caveats:
    - **Caveat 1**: Only if features are normalized!
    - **Caveat 2**: It indicates feature importance in a model with those features. You can't eliminate a feature until its coefficient is 0!

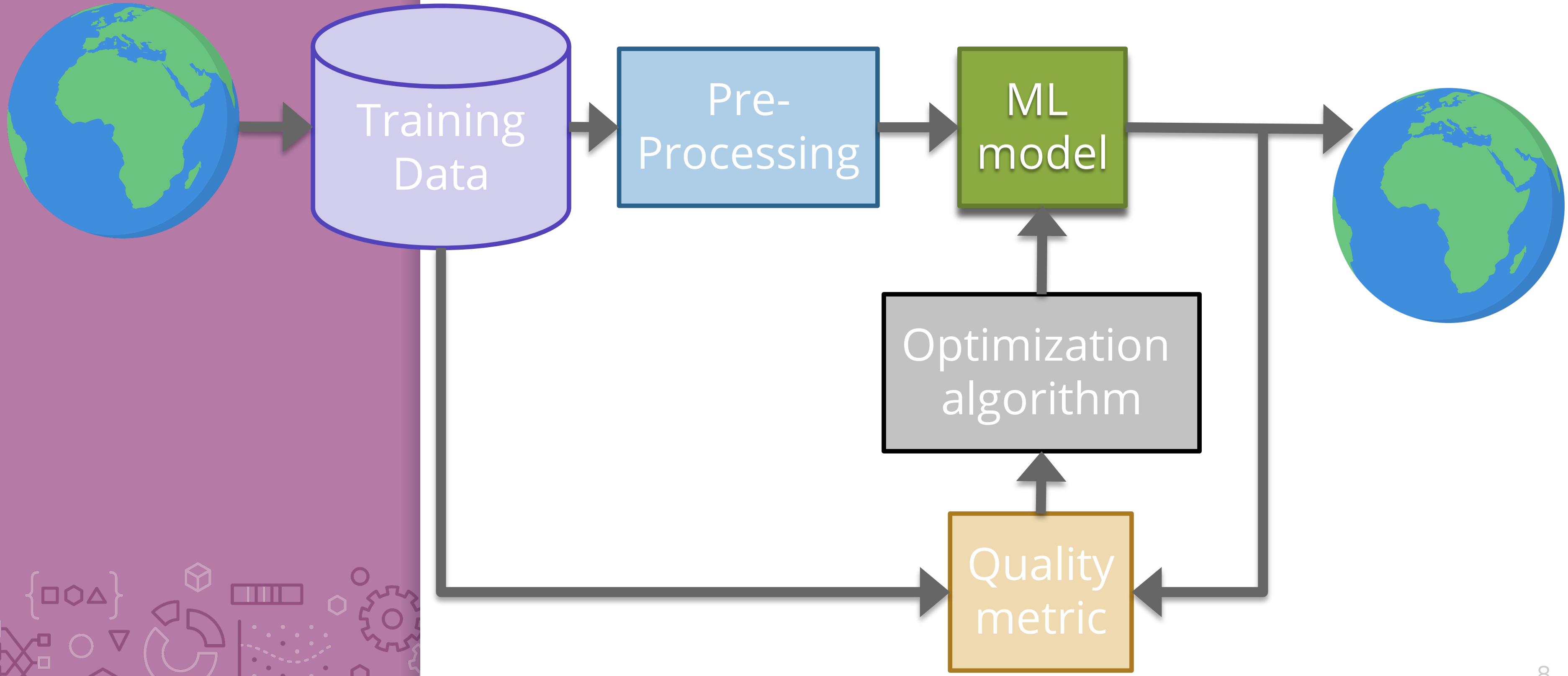
ML Pipeline



Training Data
Pre-Processing
ML model
Optimization algorithm
Quality metric

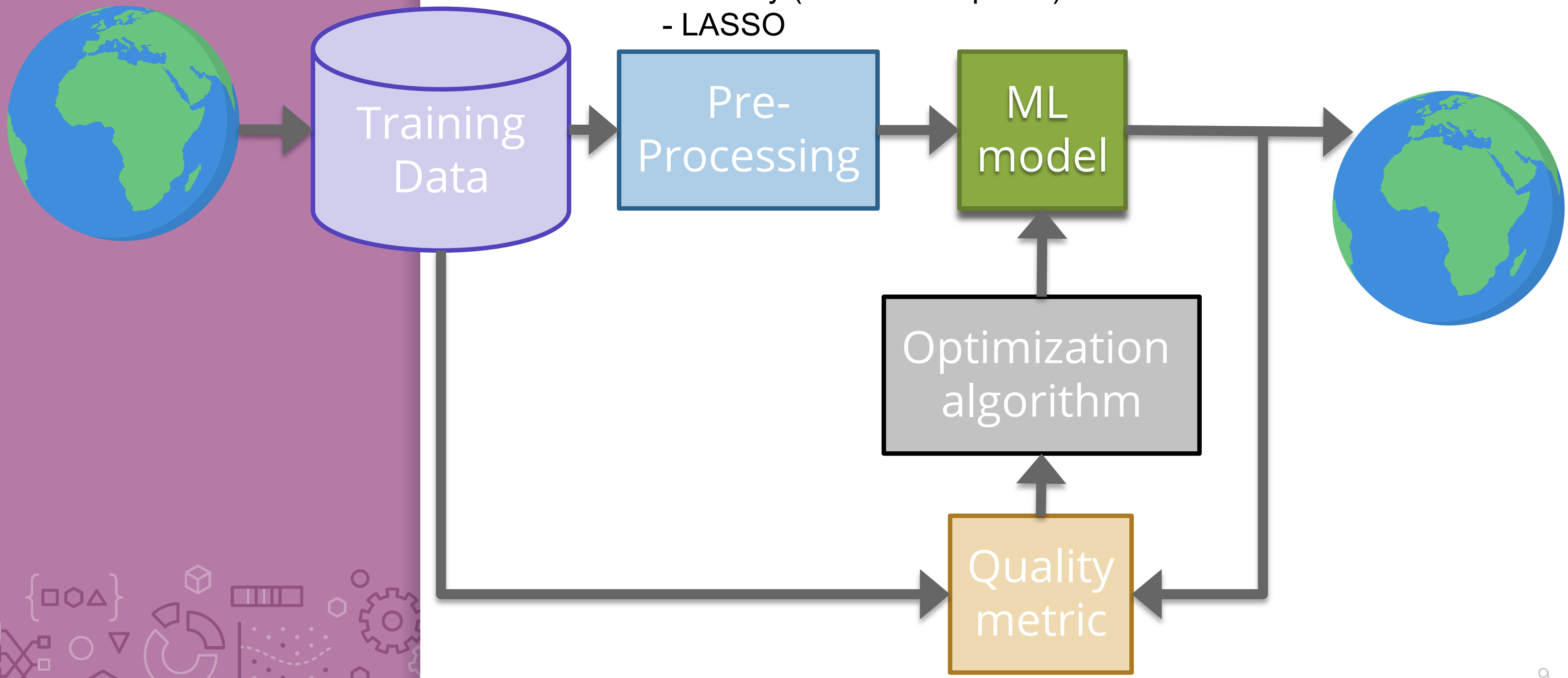
ML Pipeline

- Train-validation-test split



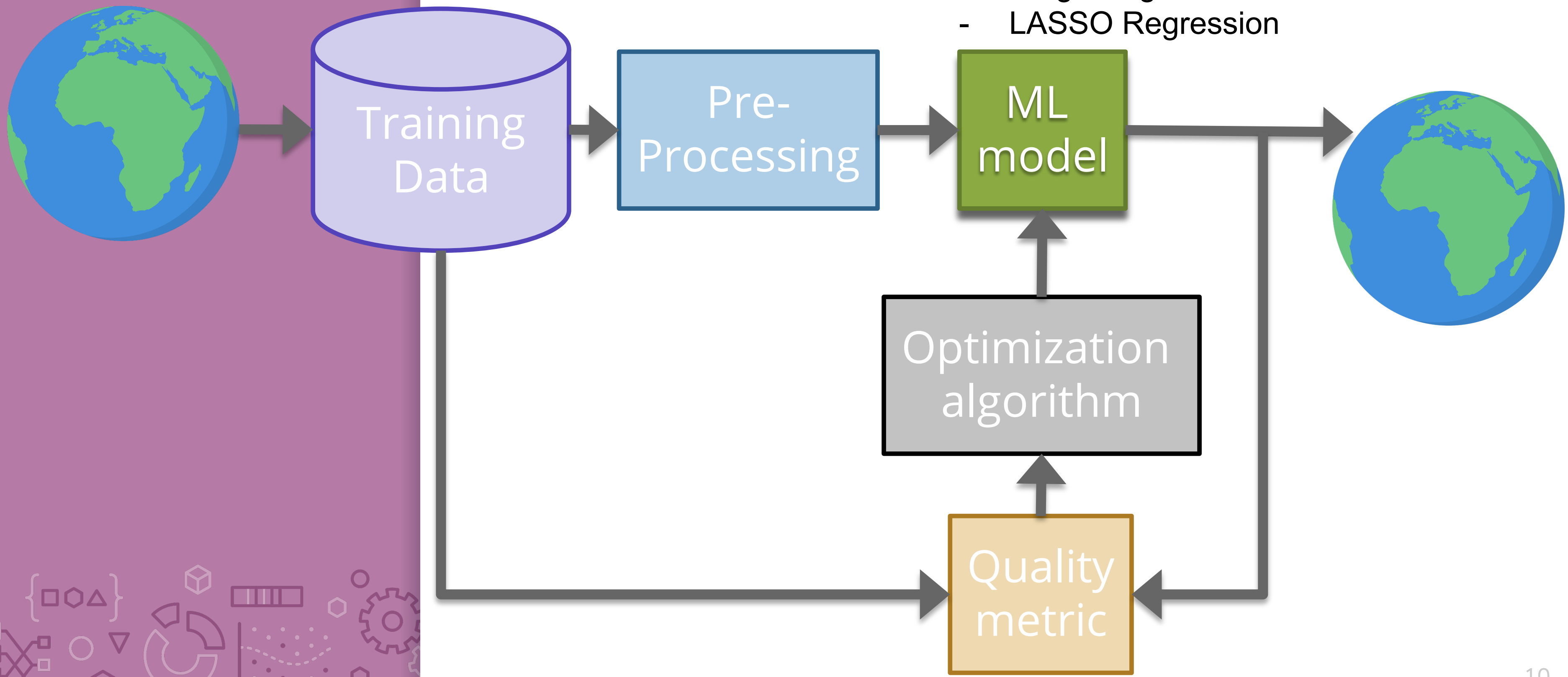
ML Pipeline

- Scaling / Normalization
- Feature Selection
  - All Subsets
  - Greedy (Forward Stepwise)
  - LASSO

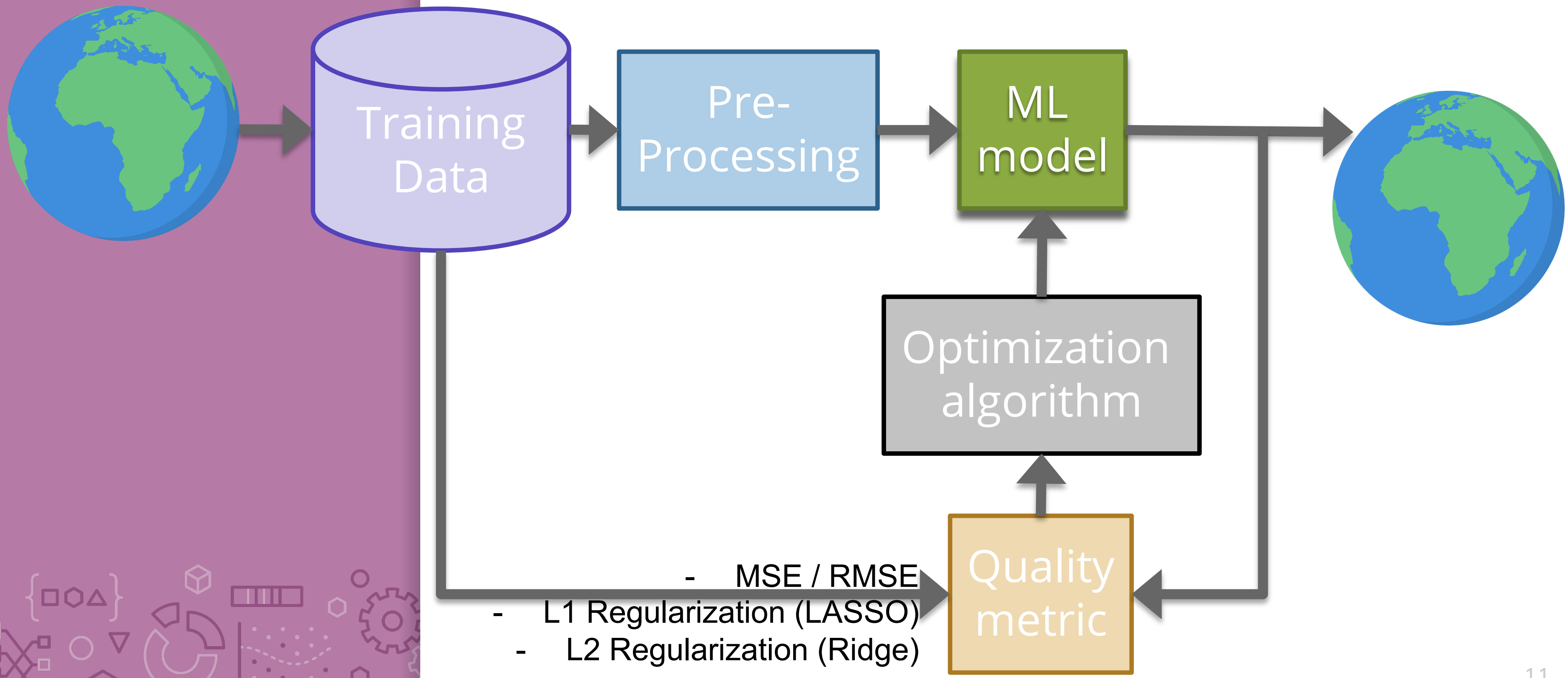


ML Pipeline

- Linear Regression
- Polynomial Regression
- Ridge Regression
- LASSO Regression

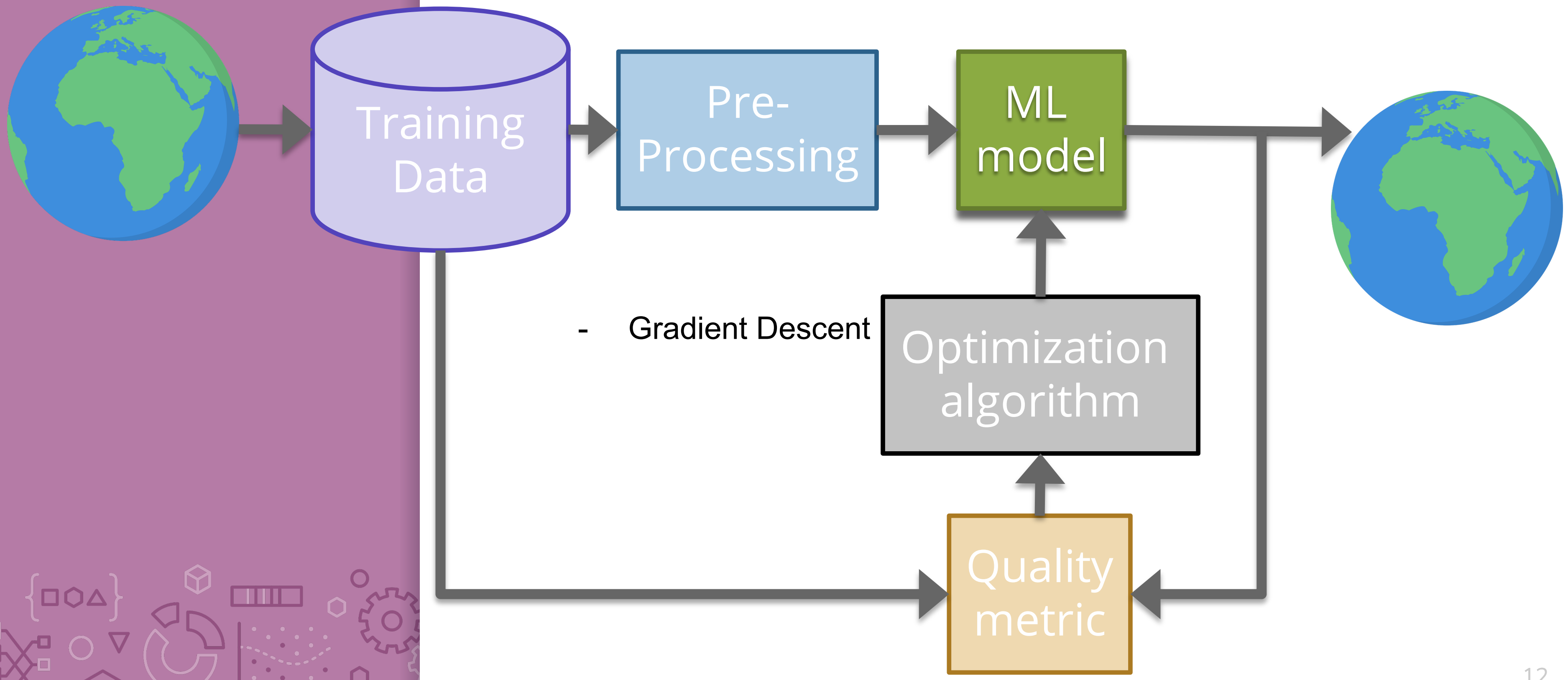


ML Pipeline



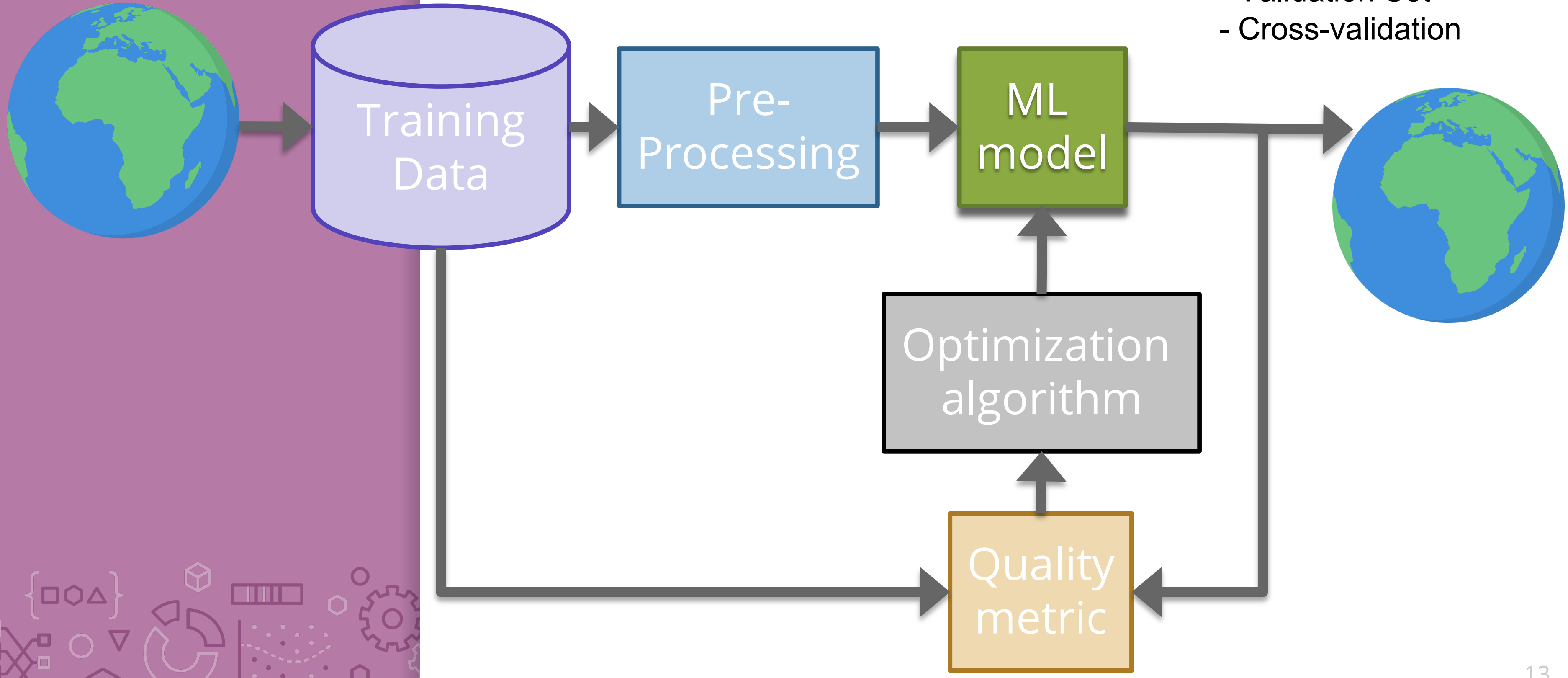
Training Data
Pre-Processing
ML model
Optimization algorithm
Quality metric
- MSE / RMSE
- L1 Regularization (LASSO)
- L2 Regularization (Ridge)

ML Pipeline


Training Data
Pre-Processing
ML model
- Gradient Descent
Optimization algorithm
Quality metric

ML Pipeline

Overarching Concepts:
- Overfitting
- Bias / Variance
- Hyperparameter Tuning:
  - Validation Set
  - Cross-validation



Classification

Roadmap So Far

1. Housing Prices - Regression
   - Regression Model
   - Assessing Performance
   - Ridge Regression
   - LASSO
2. Sentiment Analysis – Classification
   - Classification Overview
   - Logistic Regression

Regression vs. Classification

Regression problems involve predicting **continuous values**.

- E.g., house price, student grade, population growth, etc.

$y \in \mathbb{R}, \mathbb{Z}, [0,1]$

real numbers

integers

Classification problems involve predicting **discrete labels**

- e.g., spam detection, object detection, loan approval, etc.

$y \in \{+1, -1\}, \{cat, dog, horse\}$

Spam Filtering

Binary Classification

Osman Khan to Carlos — show details Jan 7 (6 days ago) — Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Miche — show details 3:15 PM (8 hours ago) — Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | x

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang — show details 9:52 PM (1 hour ago) — Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy

Output: y

Spam
+1

Not Spam (ham)
-1

Input: x

Text of email
Sender
Subject
...

Object Detection

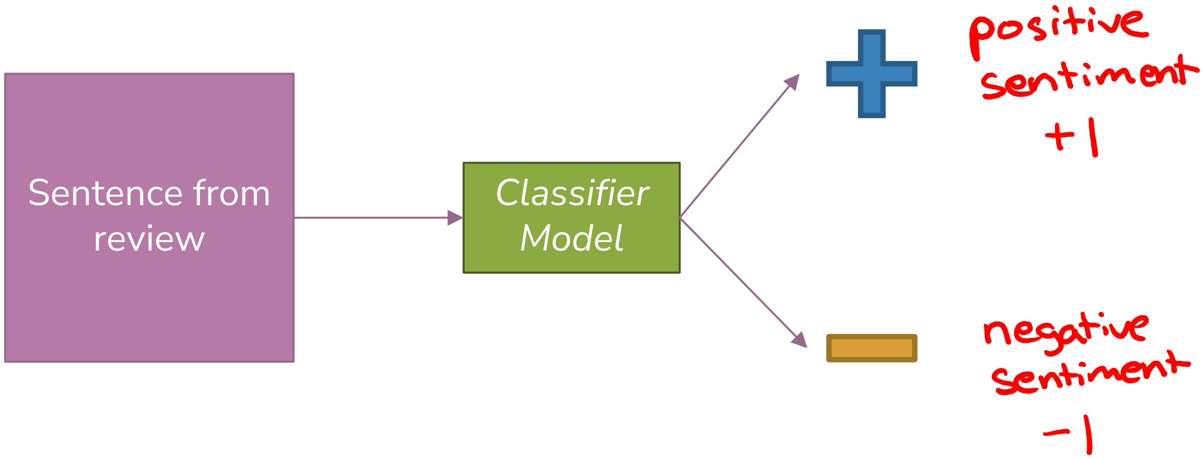
Multiclass Classification

Input: x
Pixels

Output: y
Class
(+ Probability)

Sentiment Classifier

In our example, we want to classify a restaurant review as positive or negative.



Input: x

Output: y
Predicted class

Poll Everywhere

Think

1 min

Think of 1-2 classification problem(s) not mentioned.

For each problem:

- What is the input into the model?
- What are the output classes?
- What are the social impacts of errors in the model?

Poll Everywhere

Group

2 min

Think of 1-2 classification problem(s) not mentioned.

For each problem:

- What is the input into the model?
- What are the output classes?
- What are the social impacts of errors in the model?

Stock prediction:
Input: historic data
Output: Up or Down

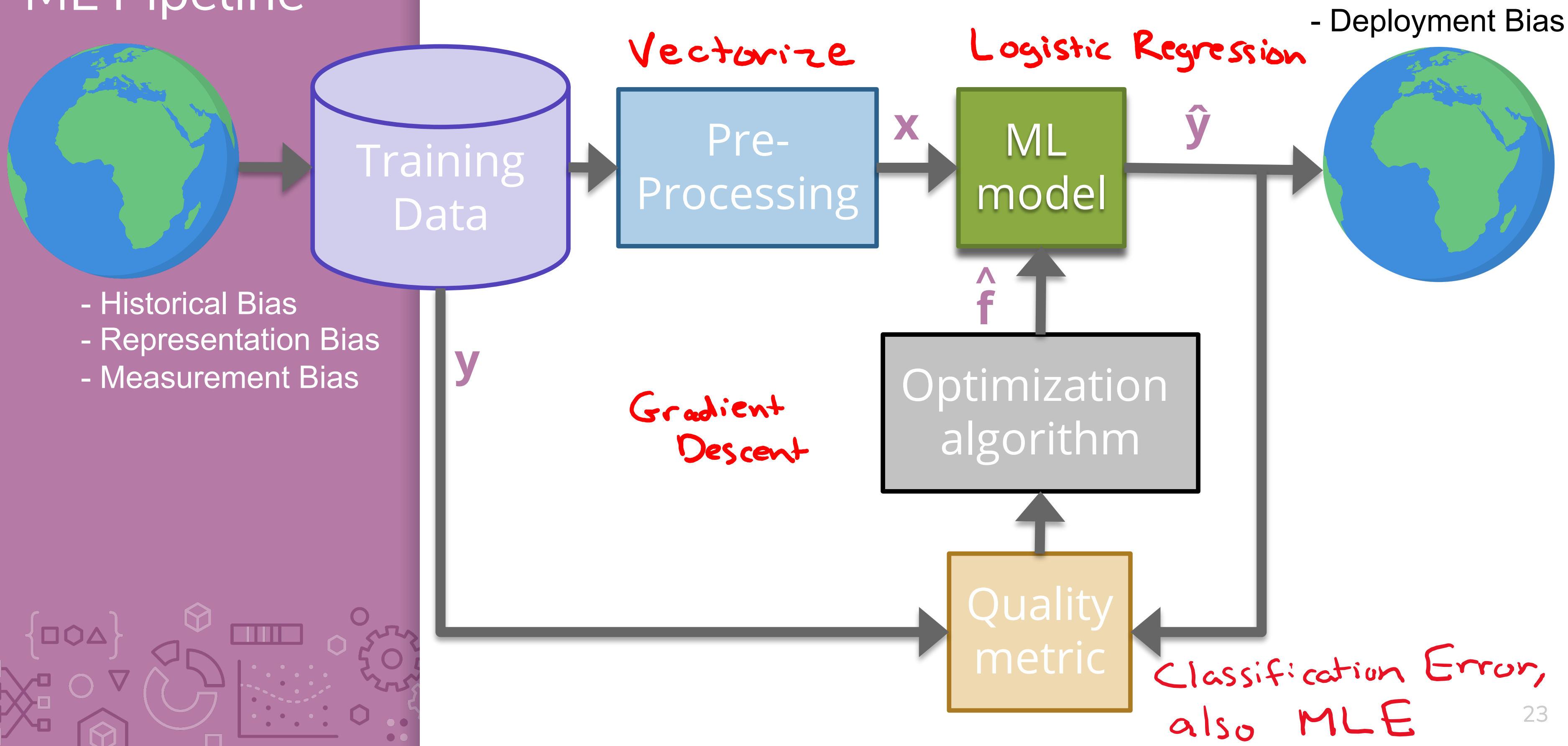
Farming:
Input: color/shape of fruit
Output: ripe/not ripe

Political Views:
In: FB posts
Output: political party

Sentiment Analysis

ML Pipeline

Input: text from reviews
Label: +1, -1 depending on sentiment



- Historical Bias
- Representation Bias
- Measurement Bias

Converting Text to Numbers (Vectorizing):

Bag of Words

Idea: One feature per word!

Example: "Sushi was great, the food was awesome, but the service was terrible"

sushi	was	great	the	food	awesome	but	service	terrible
1	3	1	2	1	1	1	1	1

This **has** to be too simple, right?

Stay tuned (today and Wed) for issues that arise and how to address them ☺

Pre-Processing: Sample Dataset

Review	Sentiment
"Sushi was great, the food was awesome, but the service was terrible"	+1
…	…
"Terrible food; the sushi was rancid."	-1

Vectorizer

Label

ALL words across all reviews

Sushi $h_1(x)$	was $h_2(x)$	great $h_3(x)$	the $h_4(x)$	food $h_5(x)$	awesome $h_6(x)$	but $h_7(x)$	service $h_8(x)$	terrible $h_9(x)$	rancid $h_{10}(x)$	Sentiment
1	3	1	2	1	1	1	1	1	0	+1
…	…	…	…	…	…	…	…	…	…	…
1	1	0	1	1	0	0	0	1	1	-1

GOAL: given a vectorized review, predict its sentiment

How to Implement Sentiment Analysis?

Attempt 1: Simple Threshold Analysis

Attempt 2: Linear Regression

Attempt 3: Linear Classifier

Attempt 4 (Wed): Logistic Regression

Attempt 1: Simple Threshold Classifier

Idea: Use a list of good words and bad words, classify review by the most frequent type of word

Word	Good?
sushi	None
was	None
great	Good
the	None
food	None
but	None
awesome	Good
service	None
terrible	Bad
rancid	Bad

Simple Threshold Classifier

Input x: Sentence from review

Count the number of positive and negative words, in x

If $\text{num\_positive} > \text{num\_negative}$:

- $\hat{y} = +1$

Else:

- $\hat{y} = -1$

Example: "Sushi was great, the food was awesome, but the service was terrible"

#pos : 2
neg : 1 $\Rightarrow \hat{y} = +1$

Limitations of Attempt 1 (Simple Threshold Classifier)

Words have different degrees of sentiment.

- Awesome > Great
- How can we weigh them differently?

Single words are not enough sometimes…

- “Good” → Positive
- “Not Good” → Negative

How do we get list of positive/negative words?

Words Have Different Degrees of Sentiments

What if we generalize good/bad to a numeric weighting per word?

Word	Good?
sushi	None
was	None
great	Good
the	None
food	None
but	None
awesome	Good
service	None
terrible	Bad
rancid	Bad

Word	Weight
sushi	0
was	0
great	1
the	0
food	0
but	0
awesome	2
service	0
terrible	-1
rancid	-2

Single Words Are Sometimes Not Enough!

What if instead of making each feature one word, we made it two?

- **Unigram**: a sequence of one word
- **Bigram**: a sequence of two words
- **N-gram**: a sequence of n-words

"Sushi was good, the food was good, the service was not good"

sushi	was	good	the	food	service	not
1	3	3	2	1	1	1

Unigrams

sushi was	was good	good the	the food	food was	the service	service was	was not	not good
1	2	2	1	1	1	1	1	1

Bigrams

Longer sequences of words results in more context, more features, and a greater chance of overfitting.

How do we get the word weights?

What if we learn them from the data?

$h_1(x)$	$h_2(x)$	$h_3(x)$	$h_4(x)$	$h_5(x)$	$h_6(x)$	$h_7(x)$	$h_8(x)$	$h_9(x)$
sushi	**was**	**great**	**the**	**food**	**awesome**	**but**	**service**	**terrible**
1	3	1	2	1	1	1	1	1

Word	Weight
sushi	w_1
was	w_2
great	w_3
the	w_4
food	w_5
awesome	w_6
but	w_7
service	w_8
terrible	w_9

In linear regression we learnt the weights for each feature. Can we do something similar here?

Poll Everywhere

Think

1 min

SKIPPED

Use a Simple Threshold Classifier to rate this review, by following the following steps:

- Decide whether you want to use unigrams or bigrams.
- Come up with lists of positive / negative words.
- Determine the predicted sentiment of the review.

(There is no one right answer)

“Their Good Old-Fashioned Burger was so good! I only wish service was faster; I did not enjoy waiting 1 hour for a burger.”

Poll Everywhere

Group

2 min

SKIPPED

Use a Simple Threshold Classifier to rate this review, by following the following steps:

- Decide whether you want to use unigrams or bigrams.
- Come up with lists of positive / negative words.
- Determine the predicted sentiment of the review.

(There is no one right answer)

“Their Good Old-Fashioned Burger was so good! I only wish service was faster; I did not enjoy waiting 1 hour for a burger.”

Poll Everywhere

Group

SKIPPED

"Their Good Old-Fashioned Burger was so good! I only wish service was faster; I did not enjoy waiting 1 hour for a burger."

"Their Good Old-Fashioned Burger was so good! I only wish service was faster; I did not enjoy waiting 1 hour for a burger."

Attempt 2: Linear Regression

Idea: Use the regression model we learnt! The output will be the sentiment!

weight of word

sushi

good

$$\text{Predicted Sentiment} = \hat{y} = \sum_{j=0}^{D} w_j h_j(x^{(i)})$$

$h_1(x)$	$h_2(x)$	$h_3(x)$	$h_4(x)$	$h_5(x)$	$h_6(x)$	$h_7(x)$	$h_8(x)$	$h_9(x)$
sushi	**was**	**great**	**the**	**food**	**awesome**	**but**	**service**	**terrible**
1	3	1	2	1	1	1	1	1

"Sushi was great, the food was awesome, but the service was terrible"

Word	**Weight**
sushi	0
was	0
great	1
the	0
food	0
awesome	2
but	0
service	0
terrible	-1

$\hat{y} = 1 \cdot 1 + 2 * 1 + -1 \cdot 1$

$= 2$

Issue: How do we measure the quality of a prediction?

Recall that the labels are binary: positive/negative sentiment.

Review	**Sentiment**
"Sushi was great, the food was awesome, but the service was terrible"	+1
…	…
"Terrible food; the sushi was rancid."	-1

However, regression models predict continuous values!

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Error for review 1: $(y_i - \hat{y}_i)^2 = (1-2)^2$

$= 1$

???? ?

Attempt 3: Linear Classifier

Idea: Only predict the sign of the output!

must already have weights

$$Score\left(x^{(i)}\right) = \hat{s} = w_0 h_0\left(x^{(i)}\right) + \cdots + w_D h_D\left(x^{(i)}\right)$$

$$= \sum_{j=0}^{D} w_j h_j\left(x^{(i)}\right)$$

$$= w^T h\left(x^{(i)}\right)$$

$$\text{Predicted } Sentiment = \hat{y} = sign(\, Score(x))$$

Attempt 3: Linear Classifier

(Another View)

Idea: Only predict the sign of the output!

$$Predicted\ Sentiment = \hat{y} = sign(\,Score(x))$$

Linear Classifier

Input x: Sentence from review

Compute $Score(x)$

If $Score(x) > 0$: ← Threshold

- $\hat{y} = +1$

Else:

- $\hat{y} = -1$

Earlier Example:

$Score(x) = 2$

$\hat{y} = +1$

Issue: How do we train this?

Say we were to use the MSE...

actual sentiment

predicted sentiment $\hat{y}$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \overbrace{sign(\,Score(x^{(i)}))}^{\text{predicted sentiment}}\,)^2$$

The derivative of the $sign$ function is 0!

Hence, Gradient Descent will no longer work ☹

Come back Wed for how to resolve this!

Brain Break

3:17

Decision Boundary

Decision Boundary

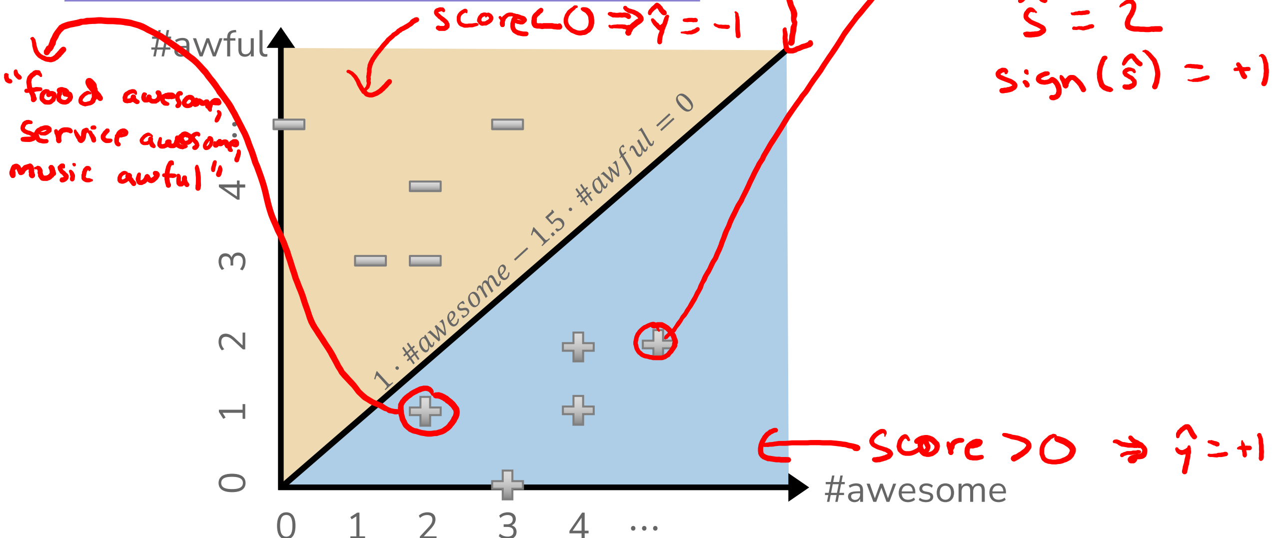
Consider if only two words had non-zero coefficients

Word	Coefficient	Weight
	w_0	0.0
awesome	w_1	1.0
awful	w_2	-1.5

$\hat{s} = 1 \cdot \#awesome - 1.5 \cdot \#awful$

defining this decision boundary

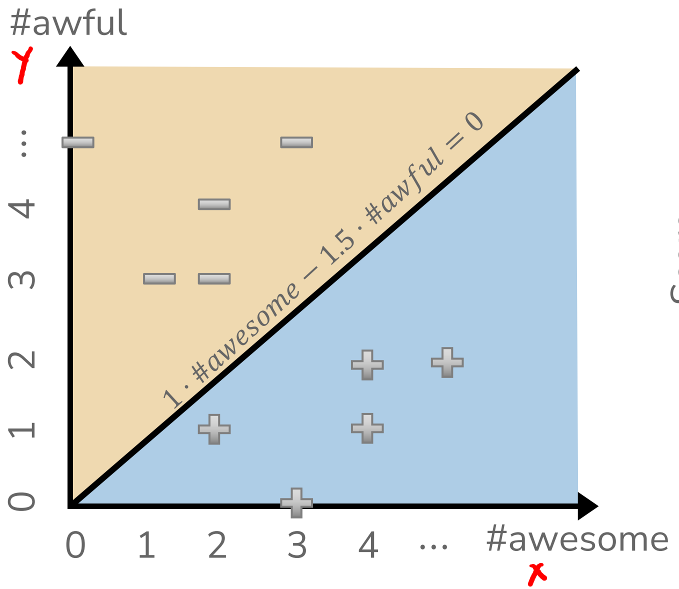
#awesome = 5
awful = 2
$\hat{s} = 2$
$\text{sign}(\hat{s}) = +1$



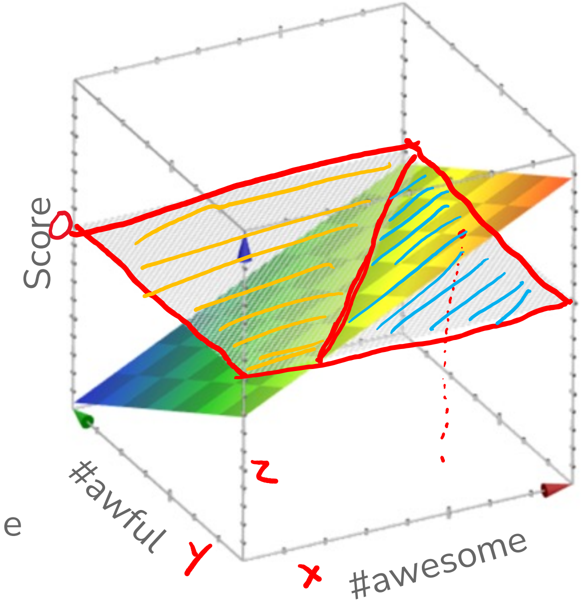
Decision Boundary

$$Score(x) = 1 \cdot \#awesome - 1.5 \cdot \#awful$$

2-dimensional view



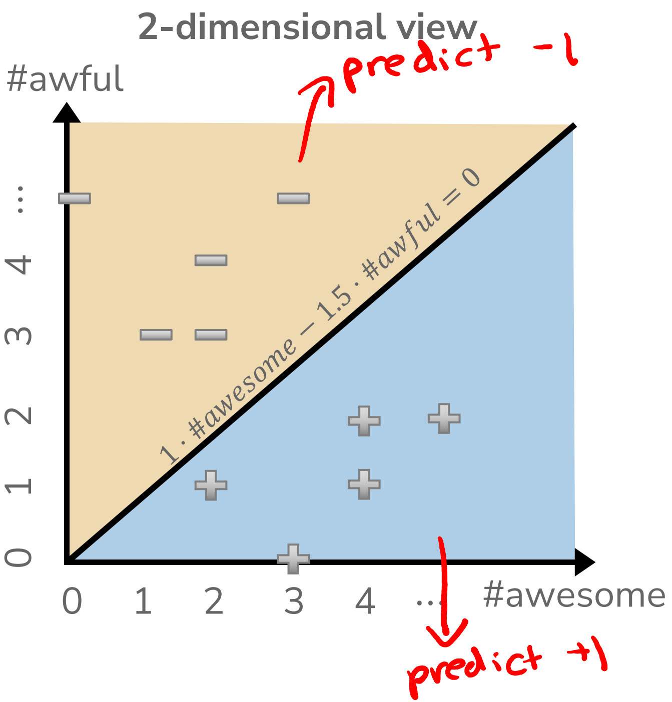
3-dimensional view

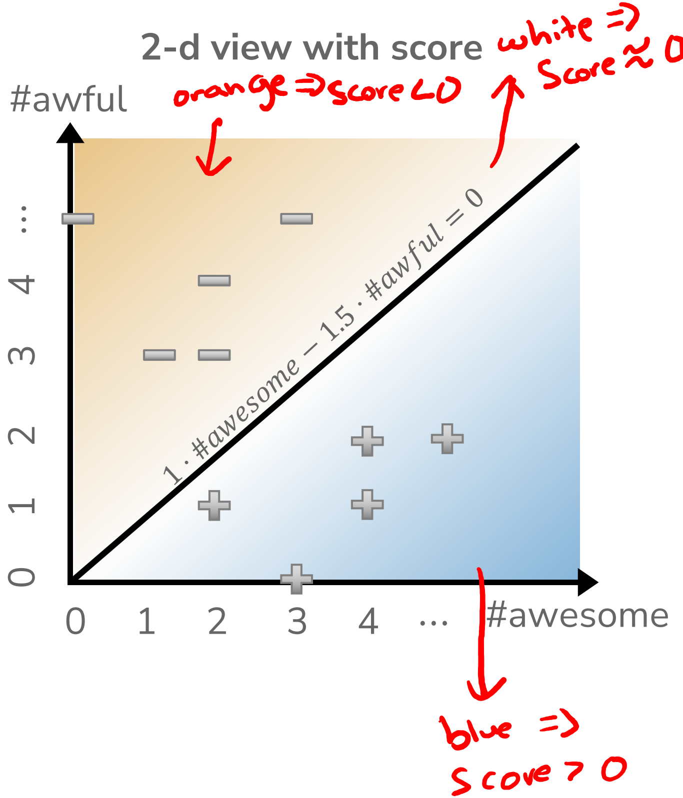


Generally, with classification we don't us a plot like the 3d view since it's hard to visualize, instead use 2d plot with decision boundary

Decision Boundary with Score

$$Score(x) = 1 \cdot \#awesome - 1.5 \cdot \#awful$$

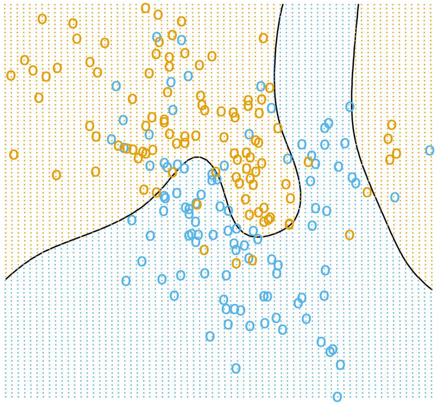


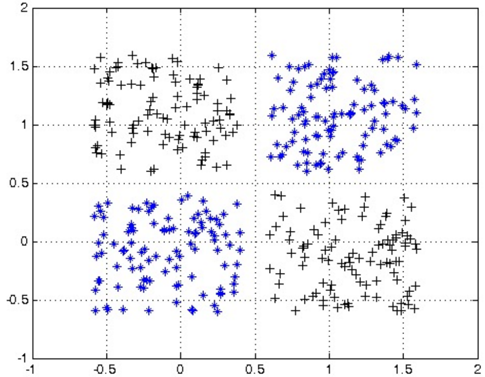


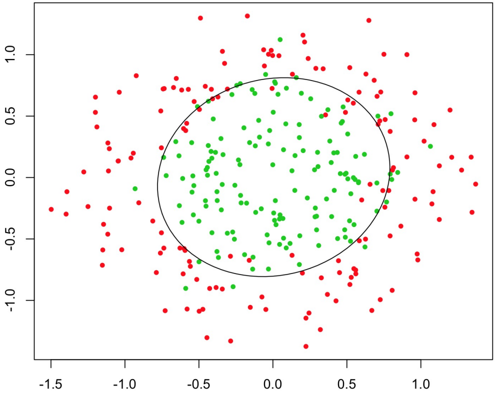
Complex Decision Boundaries?

What if we want to use a more complex decision boundary?

- Need more complex model/features! (Come back Wed)



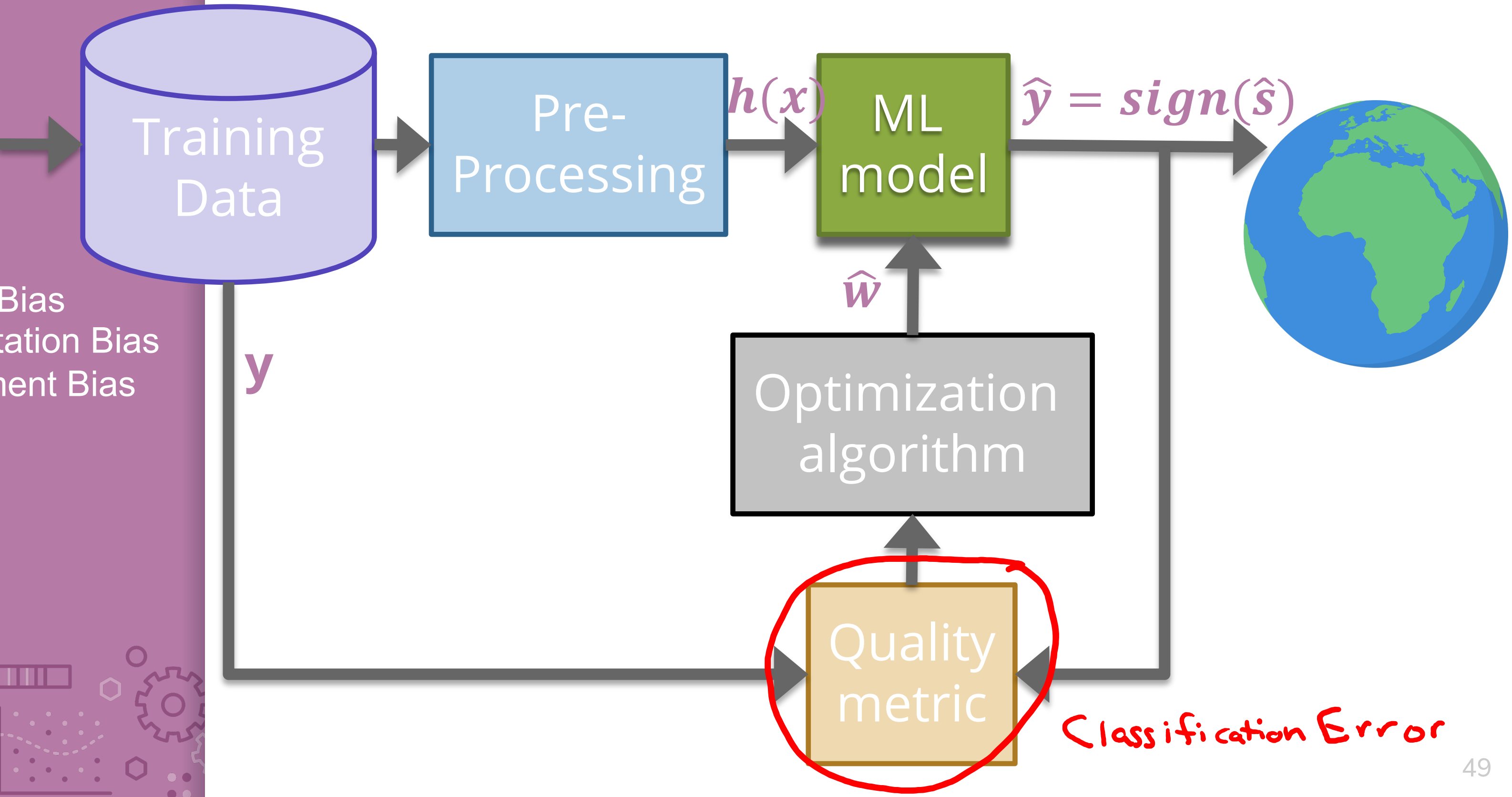




Evaluating Classifiers

ML Pipeline

- Historical Bias
- Representation Bias
- Measurement Bias



Classification Error

Ratio of examples where there was a mistaken prediction

What's a mistake?

If the true label was positive ($y = +1$), but we predicted negative ($\hat{y} = -1$) → False Negative

If the true label was negative ($y = -1$), but we predicted positive ($\hat{y} = +1$) → False Positive

Classification Error

$$\frac{\#\text{mistakes}}{\#\text{examples}} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i \neq \hat{y}_i\}}{n}$$

Classification Accuracy

$$\frac{\#\text{correct}}{\#\text{examples}} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = \hat{y}_i\}}{n} = 1 - \text{error}$$

What's a good accuracy?

For binary classification:

Should at least beat random guessing...

Accuracy should be at least 0.5

For multi-class classification (k classes):

Should still beat random guessing

Accuracy should be at least: 1 / k

- 3-class: 0.33
- 4-class: 0.25
- ...

Digit Classification:
Accuracy > 10%

Besides that, higher accuracy means better, right?

Detecting Spam

Imagine I made a “Dummy Classifier” for detecting spam

The classifier ignores the input, and always predicts spam.

This actually results in 90% accuracy! Why?

- Most emails are spam…

This is called the **majority class classifier.**

A classifier as simple as the majority class classifier can have a high accuracy if there is a **class imbalance**.

A class imbalance is when one class appears much more frequently than another in the dataset

This might suggest that accuracy isn't enough to tell us if a model is a good model.

Assessing Accuracy

Always digging in and ask critical questions of your accuracy.

Is there a **class imbalance**?

How does it compare to a baseline approach?

- Random guessing
- Majority class
- …

Most important: **What does my application need?**

- What's good enough for user experience?
- What is the impact of a mistake we make?

Confusion Matrix

For binary classification, there are only two types of mistakes

$$\hat{y} = +1, \quad y = -1$$

$$\hat{y} = -1, \quad y = +1$$

Generally we make a **confusion matrix** to understand mistakes.

Complete the sentence: "my prediction was a ..."

		Predicted Label	
		+	−
True Label	+	True Positive (TP)	False Negative (FN)
	−	False Positive (FP)	True Negative (TN)

Tip on remembering: complete the sentence "My prediction was a …"

Confusion Matrix Example

		Predicted Label		
		+	−	
True Label	+	True Positive (TP) 50	False Negative (FN) 10	60
	−	False Positive (FP) 15	True Negative (TN) 35	50
		65	45	

Total # examples: $50 + 10 + 15 + 35 = 110$

$$\text{Accuracy} = \frac{\#\text{ correct}}{\#\text{ examples}} = \frac{50+35}{110} = \frac{85}{110}$$

Which is Worse?

What's worse, a false negative or a false positive?

It entirely depends on your application!

Detecting Spam

email was spam / predicted not

False Negative: Annoying

False Positive: Email lost

Medical Diagnosis

False Negative: Disease not treated

False Positive: Wasteful treatment

In almost every case, how treat errors depends on your context.

Will pick up from here on Wed

Errors and Fairness

We mentioned on the first day how ML is being used in many contexts that impact crucial aspects of our lives.

Models making errors is a given, what we do about that is a choice:

- Are the errors consequential enough that we shouldn't use a model in the first place?
- Do different demographic groups experience errors at different rates?
  - If so, we would hopefully want to avoid that model!

Will talk more about how to define whether or a not a model is fair / discriminatory later in the course. Will use these notions of error as a starting point!

Binary Classification Measures

Notation

$$C_{TP} = \#\mathrm{TP}, \quad \mathrm{C_{FP}} = \#\mathrm{FP}, \quad \mathrm{C_{TN}} = \#\mathrm{TN}, \quad \mathrm{C_{FN}} = \#\mathrm{FN}$$

$$N = C_{TP} + C_{FP} + C_{TN} + C_{FN}$$

$$N_P = C_{TP} + C_{FN}, \quad N_N = C_{FP} + C_{TN}$$

Error Rate

$$\frac{C_{FP} + C_{FN}}{N}$$

Accuracy Rate

$$\frac{C_{TP} + C_{TN}}{N}$$

False Positive rate (FPR)

$$\frac{C_{FP}}{N_N}$$

False Negative Rate (FNR)

$$\frac{C_{FN}}{N_P}$$

True Positive Rate or Recall

$$\frac{C_{TP}}{N_P}$$

Precision

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

F1-Score

$$2\,\frac{Precision \cdot Recall}{Precison + Recall}$$

[See more!]

Multiclass Confusion Matrix

Consider predicting (*Healthy, Cold, Flu*)

Predicted Label

True Label	Healthy	Cold	Flu
Healthy	60	8	2
Cold	4	12	4
Flu	0	2	8

Poll Everywhere

Think

1 min

pollev.com/cs416

Suppose we trained a classifier and computed its confusion matrix on the training dataset. **Is there a class imbalance in the dataset and if so, which class has the highest representation?**

Predicted Label

True Label	Pupper	Doggo	Woofer
Pupper	2	27	4
Doggo	4	25	4
Woofer	1	30	2

Poll Everywhere

Group

2 min

pollev.com/cs416

Suppose we trained a classifier and computed its confusion matrix on the training dataset. **Is there a class imbalance in the dataset and if so, which class has the highest representation?**

Predicted Label

True Label	Pupper	Doggo	Woofer
Pupper	2	27	4
Doggo	4	25	4
Woofer	1	30	2

Next Time

We will address the issues highlighted with the Linear Classifier approach from today by predicting the probability of a sentiment, rather than the sentiment itself.

$$P(y|x)$$

Normally assume some structure on the probability (e.g., linear)

$$P(y|x,w) \approx w^T x$$

Use machine learning algorithm to learn approximate $\hat{w}$ such that $\hat{P}(y|x)$ is close to $P(y|x)$, where:

$$\hat{P}(y|x) = P(y|x,\hat{w})$$

Activity: Bias Framework in Machine Learning

ML and Society

ML Systems Gone Wrong

INDEPENDENT

GOOGLE'S ALGORITHM SHOWS PRESTIGIOUS JOB ADS TO MEN, BUT NOT TO WOMEN

REUTERS — Business · Markets · World · Politics · TV · More

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin · 8 MIN READ

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019 · 168

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

COMPAS

An ML model created by NorthPointe used to predict likelihood of inmates to “recidivate”. Eventually started use in Florida in judges’ decision for parole

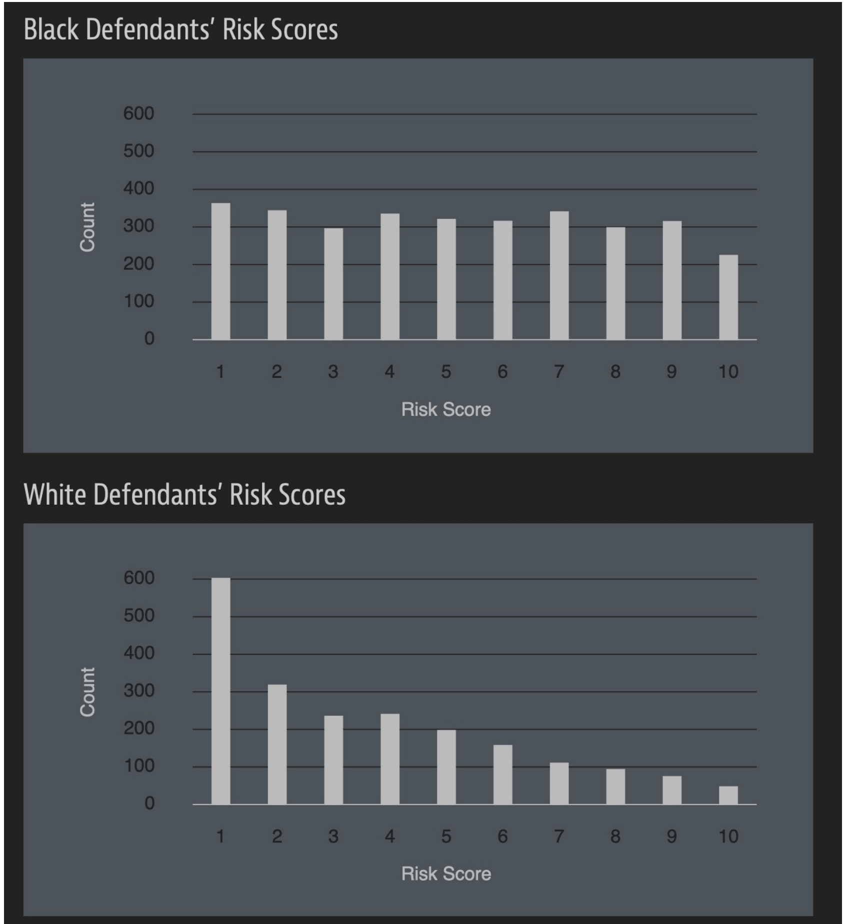
ProPublica (a news org) investigated the model and wrote that the model exhibited biased behavior against people of color. Particularly, they found that the model would predict higher risk scores for black people.

Northpointe countered and claimed that their scores were well **calibrated** (e.g., when the predict score of 9/10 that person recidivates about 90% of the time).

- Interesting follow up from ProPublica

So the question is: Who is right? Is it right to use this model?

COMPAS



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Why Biased Outcomes?

Probably not the case that someone explicitly coded the model to be biased against a particular race. In fact, race was not even a question that was on the survey inmates took!

More often than not, biased outcomes from a model come from **the data it learns from** rather than some explicit choice from the modeler.

> “Garbage in → Garbage out”
>
> “Bias in → Bias out”

Sources of Bias

Sources of Bias

Discussion heavily based on Suresh and Guttag (2020)

Six common sources of bias:

Historical bias

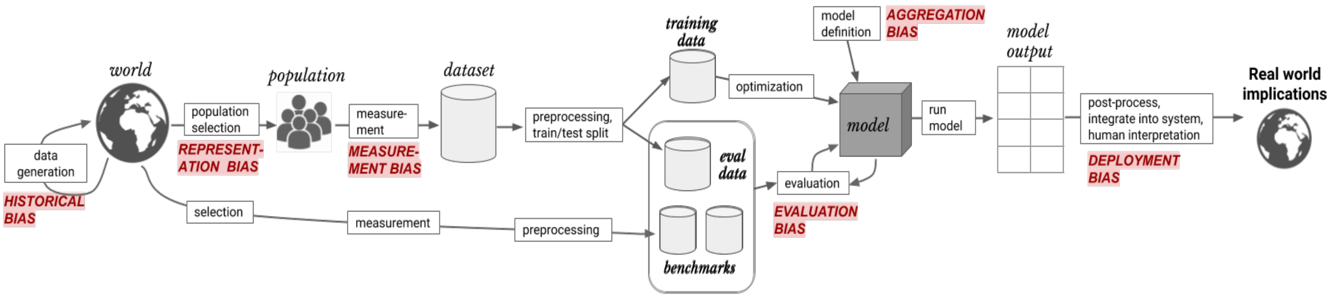
Representation Bias

Measurement Bias

Aggregation Bias

Evaluation Bias

Deployment Bias



[A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING](), BY HARINI SURESH AND JOHN V. GUTTAG, 2020

Historical Bias

The world we lived in is one that contains biases for/against certain demographics. Even ‘accurate’ data could still be harmful.

- Historical bias exists even with perfect sampling or feature measurement (other sources of bias are possible)!

Examples:

- In 2018, 5% of Fortune 500 CEOs were women. Should search results for “CEO” match this statistic? Could reflecting the world (even if accurately) perpetuate more harm?

Representation Bias

When the *training data* we collect does not contain representative samples of the true distribution.

Examples:

- If we use data gathered from smart phones, we would likely be underestimating poorer and older populations.
- ImageNet (a very popular image dataset) with 1.2 million images. About 45% of these images were taken in the US and the majority of the rest in North America and Western Europe. Only about 1% and 2.1% of the images come from China and India respectively.

Measurement Bias

Often we are gathering data that contains (noisy) proxies of characteristics of interest. Some examples:

Financial responsibility → Credit Score

Crime Rate → Arrest Rate

Intelligence → SAT Score

If these measurements are not measured equally across groups or places (or aren't relevant to the task at hand), this can be another source of bias.

Measurement Bias (cont.)

Examples:

If factory workers are monitored more often, more errors are spotted. This can result in a **feedback loop** to encourage more monitoring in the future.

- Same principles at play with predictive policing. Minoritized communities were more heavily policed in the past, which causes more instances of documented crime, which then leads to more policing in the future.

Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.

The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

Aggregation Bias

When we use a “one-sized fits all” model that does not accurately serve every group equally.

Examples:

> HbA1c levels (used to monitor and diagnose diabetes) differ in very complex ways across ethnicities and sexes. One model for everyone might not be the right choice, even if everyone is represented well in the training data.

Evaluation Bias

Similar to representation bias, but focused more on the data we evaluate or test ourselves against. If the evaluation dataset or benchmark doesn't represent the world well, we have evaluation bias.

> **Benchmarks** are common datasets used to evaluate models from different researchers.

Examples:

> If it is common to report accuracy on a benchmark, this might hide disparate performance on subgroups.
>
> Drastically worse performance for facial recognition software when used on faces of darker-skinned females. Common evaluation datasets for facial recognition only had 5-7% had faces of darker-skinned women.

Deployment Bias

When there is a difference in how a model was intended to be used and how it is actually used when deployed in the real-world.

Examples:

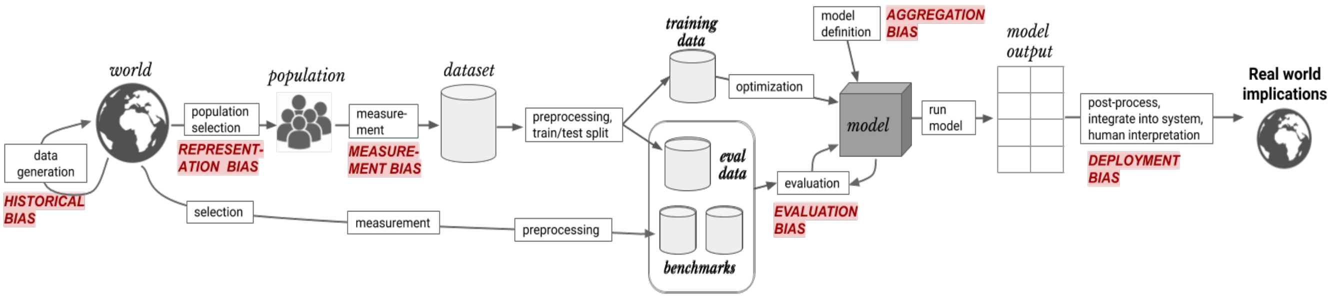
- Crime risk prediction models might be evaluated to achieve good calibration, but the model designers might not have evaluated the model's use in the context of determining prison sentence lengths.
- People are complex and when using models to aid their decisions, might make incorrect assumptions about what a model says.

Sources of Bias

Discussion heavily based on Suresh and Guttag (2020)

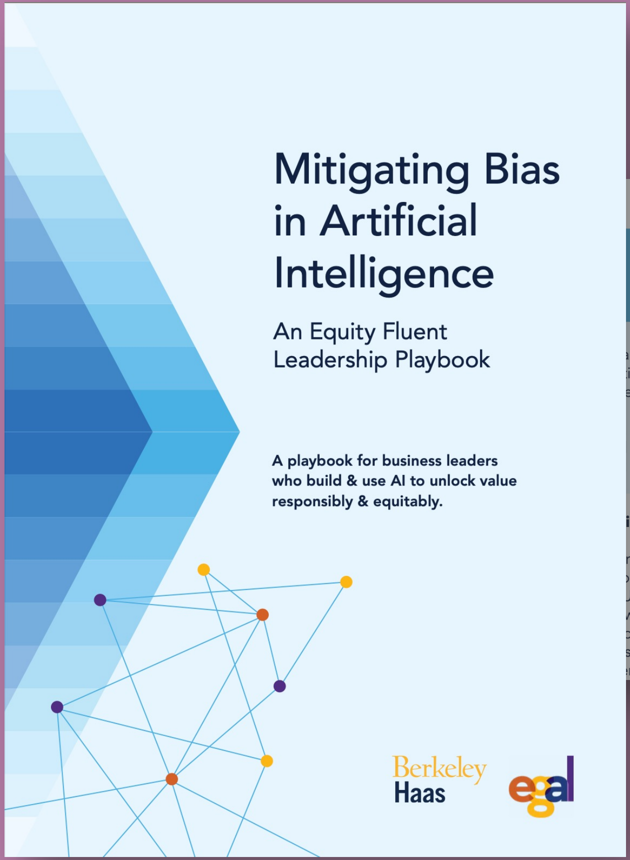
Six common sources of bias:

- **Historical bias**
- **Representation Bias**
- **Measurement Bias**
- **Aggregation Bias**
- **Evaluation Bias**
- **Deployment Bias**

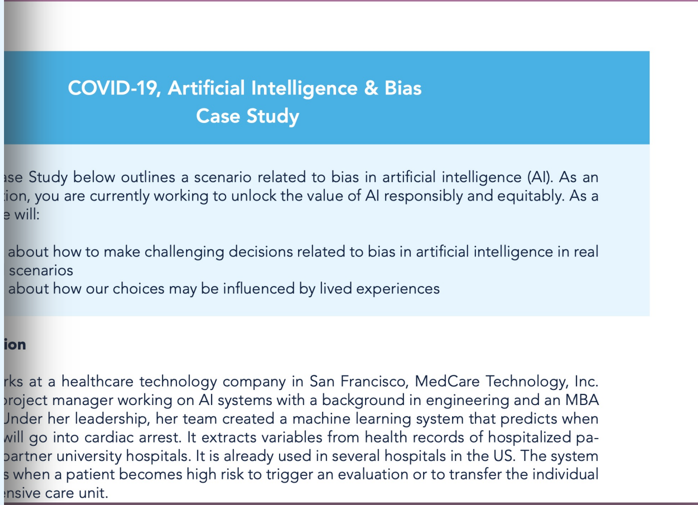


[A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING](), BY HARINI SURESH AND JOHN V. GUTTAG, 2020

ML Bias Case Study



https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf



COVID-19, Artificial Intelligence & Bias
Case Study

ase Study below outlines a scenario related to bias in artificial intelligence (AI). As an
ion, you are currently working to unlock the value of AI responsibly and equitably. As a
e will:

about how to make challenging decisions related to bias in artificial intelligence in real
scenarios
about how our choices may be influenced by lived experiences

ion

rks at a healthcare technology company in San Francisco, MedCare Technology, Inc.
project manager working on AI systems with a background in engineering and an MBA
Under her leadership, her team created a machine learning system that predicts when
will go into cardiac arrest. It extracts variables from health records of hospitalized pa-
partner university hospitals. It is already used in several hospitals in the US. The system
s when a patient becomes high risk to trigger an evaluation or to transfer the individual
nsive care unit.

https://haas.berkeley.edu/wp-content/uploads/Quick-win_Bias-in-AI-Case-Study.pdf

COVID-19, Artificial Intelligence & Bias Case Study

The Case Study on the next few slides outline a scenario related to bias in artificial intelligence (AI).

As an organization, you are currently working to unlock the value of AI responsibly and equitably. As a group, we will:

- Think about how to make challenging decisions related to bias in artificial intelligence in real world scenarios
- Think about how our choices may be influenced by lived experiences

Scenario

Anita works at a healthcare technology company in San Francisco, MedCare Technology, Inc. She is a project manager working on AI systems with a background in engineering and an MBA degree. Under her leadership, her team created a machine learning system that predicts when patients will go into cardiac arrest. It extracts variables from health records of hospitalized patients at partner university hospitals. It is already used in several hospitals in the US. The system highlights when a patient becomes high risk to trigger an evaluation or to transfer the individual to an intensive care unit.

In March 2020, the novel coronavirus SARS-COV-2 (COVID-19) was declared a pandemic by the World Health Organization and shortly after, a national emergency was declared in the United States regarding the outbreak. Given the scale of the pandemic, it was anticipated that hospitals in locations globally would be overrun and doctors overwhelmed, straining doctors' capacity to assess patient risk and make critical decisions timely and effectively. Anita's company immediately kicked into gear wondering how it could adapt their cardiac arrest tool to help doctors and COVID-19 patients. They asked themselves, "How might we use AI to predict which patients will be high risk to COVID-19 complications? How might an early warning system help inform deployment and allocation of life-saving resources like ventilators?" The team was excited – many hospitals, particularly in New York, were already tearing at the seams with doctors attempting to support as many patients as possible and volunteers looking for direction. Her team could do something.

Email

During the team's exploratory phase, Dr. Martin, a lung specialist doctor working at a large Bay Area hospital and advisor to the team, shared the following email:

Hi Team,
Anita asked me to share some information that might be relevant as you develop your AI model. Hope this helps and let me know if you have any follow up questions.

- *Patients that are higher risk tend to include: older patients and those with underlying medical conditions (e.g., obesity, type 2 diabetes, asthma). More from the Center for Disease Control and the underlying medical conditions is here.*
- *We have and can share data from chest CT scans of patients that received them at our hospital. CT scans use x-rays to identify COVID-19 signs and the extent of the virus in the lungs.*
- *We have and can share extensive health data from other COVID-19 patients we've had to date. Of course, all information shared will go through rigorous privacy and licensing procedures.*

Under Anita's guidance, your team gets to work.

Task

1. Individually read the case study. (Suggested time: 5 minutes)
2. In groups of 2-3, answer the following questions: (Suggested time: 15 minutes)
    1. What concerns do you have about the data referenced by Dr. Martin? How might this data be biased?
    2. Do you have any follow up questions for Dr. Martin? What other information or data would you like?
    3. What types of features would you like to include in the algorithmic model?
    4. Are there any ways that these features could embed bias? If so, how? (Hint: The medical system has a history of discrimination, an example of that is here)
    5. Are there other ways that bias could creep into the algorithm?
    6. If the biases discussed so far were to lead to an inaccurate prediction, what impact(s) would that have?

Recap

Theme: Describe high level idea and metrics for classification

Ideas:

- Applications of classification
- Linear classifier
- Decision boundaries
- Classification error / Classification accuracy
- Class imbalance
- Confusion matrix
- Sources of bias in machine learning