Bias, Fairness, Accountability, and Transparency in Machine Learning

CS 115 Computing for the Socio-Techno Web

Instructor: Brian Brubach

Announcements

- Adjustment to deadline schedule
  - Assignment 5 due Tuesday
  - Project milestone 4 due Friday
- Elissa Redmiles remote lecture **Thursday 9:45-11:00am**
  - Reading posted
  - If you can’t make it, but have questions, email me by Wednesday night

Some questions

- How much data about each of us is collected online (and offline)?
- How are computers/websites/algorithms using that data to make decisions about us or that affect us?
- Can algorithms discriminate and how?
- Can we prevent algorithms from discriminating?
- Can algorithms combat discrimination and how?

Examples of computers making decisions

- Email spam filtering
  - Is an email spam or not?
- Advertising
  - Which ads should be shown to you?
- Social networks
  - What posts do you see? Who sees your posts?
- Web search
  - What results do you see when you search online?

Higher stakes examples of computer decisions

- Hiring and recruiting web sites
  - Who sees job ad? Which applications get filtered out?
- Banking
  - Which loans/credit cards do you qualify for? Amount? Interest rate?
- Criminal justice
  - Who is released on bail and how much? Which neighborhoods get patrolled?
- Self-driving cars
- Insurance
  - What should your insurance rate be? How risky are you?
- Healthcare
  - Who gets access to more urgent care?

Introduction to machine learning classification

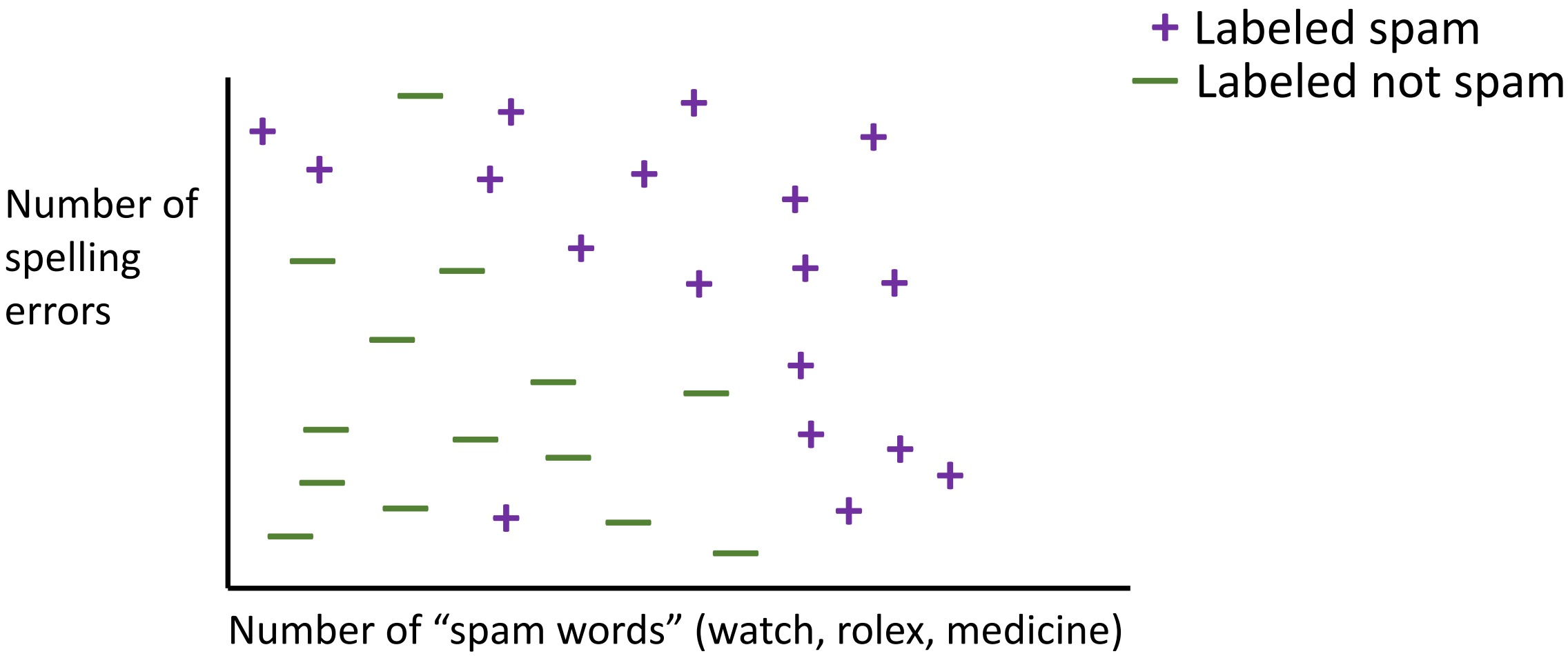
- Each data point has a set features and a label
  - Data point could be an email, job application, image, etc.
- Features → Information we have about the data point
  - Email → Length, spelling errors, common spam words (watch, Rolex, medicine, prince)?
  - Picture → Pixel colors, shapes
- Label → Something we want to know about the data
  - Email → Spam or not spam
  - Picture → This is a picture of a car, tree, horse, etc.
- Goal → Algorithm that can look at features for a data point and guess its label

Introduction to machine learning classification

- One approach → “Train” a classifier
  - Classifier → An algorithm that performs the classification task
- Show the algorithm labeled data (training set)
- Have it develop rules for predicting labels on unlabeled data
- Supervised learning

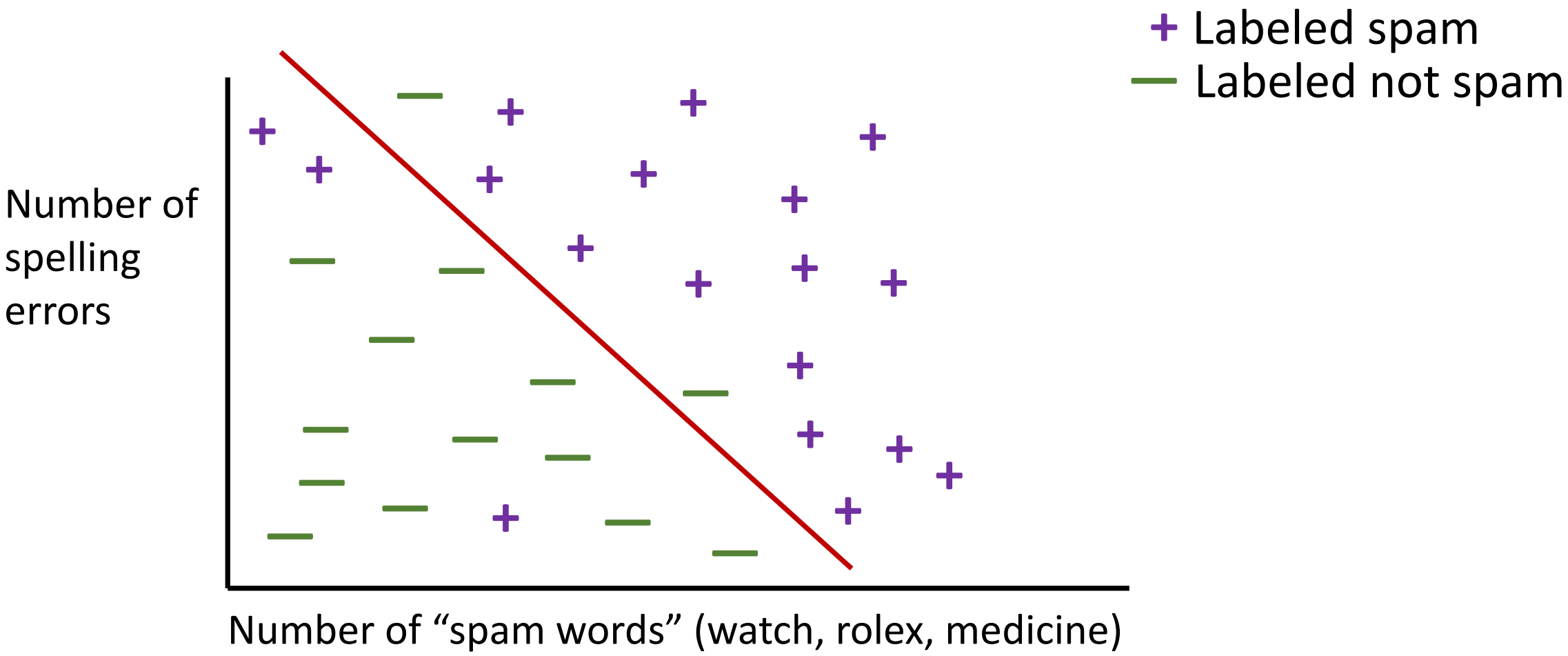
Spam filter example

- Linear classifier with two features



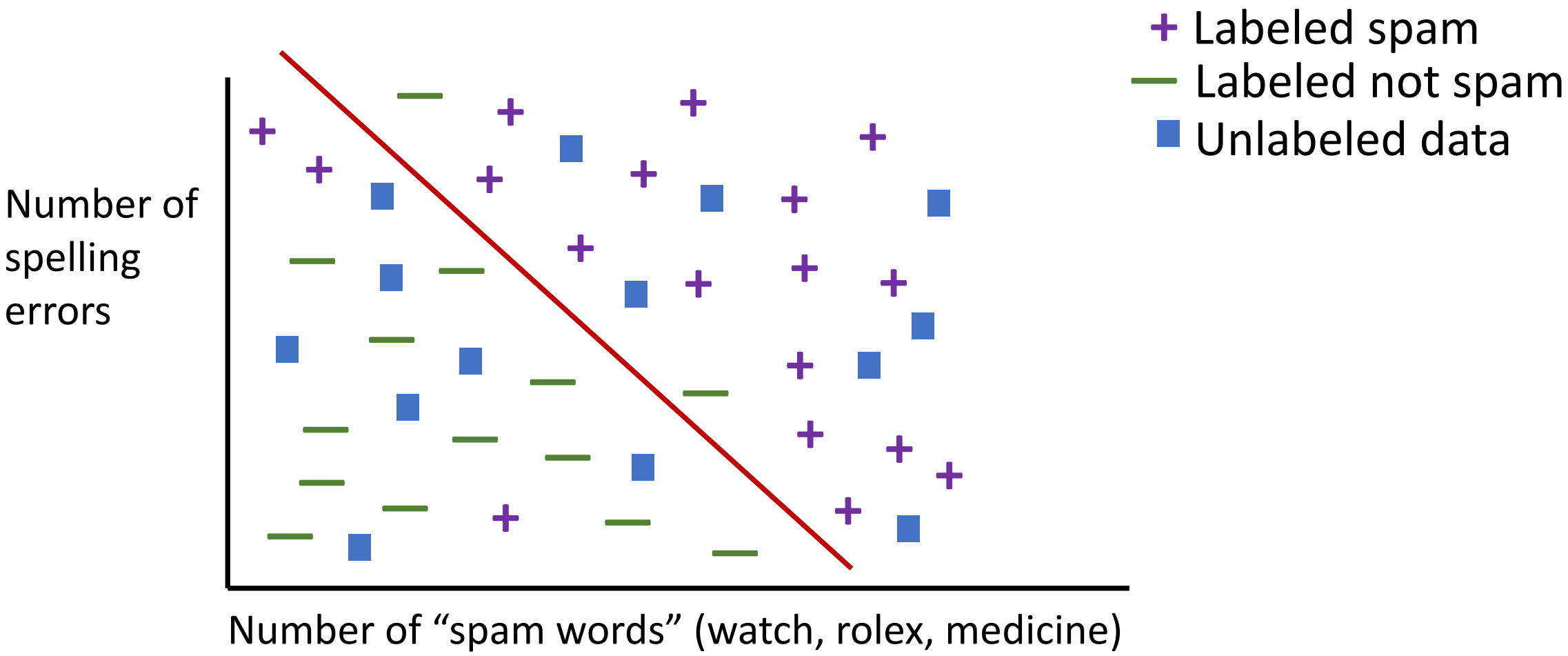
Spam filter example

- Linear classifier with two features



Spam filter example

- Linear classifier with two features



Real world classifiers

- May use thousands of features or more
  - Previous example was 2-dimensional
  - Imagine 3-dimensional, 4-dimensional, 1,000-dimensional
- Not limited to a linear classifier
  - Could be a curvy line, a list of conditional rules, or something else entirely
- Often not obvious why a classifier is making a decision
  - E.g., deep learning
- Obey the principle of garbage in, garbage out
  - But this is not the only problem!

Sensitive features

- Some common features associated with people
  - Browsing and shopping history
  - Location
  - Ratings (how they rate movies, recipes, books, etc.)
  - Content of emails and social media posts
  - Pictures of the person or pictures they share
  - Medical history
- Common “sensitive” features
  - Race, gender, age, disability status, etc.
  - Often things you can’t legally discriminate based on
- How can we avoid bias and discrimination based on sensitive features?
- Simple idea → What if we just remove sensitive features from our data?

Redundant encoding: the invisible red line

- Redundant encoding → Information about one feature can be inferred from other features
- Well-known examples → Redlining and congressional districting
- Redlining → Discrimination based on residential location that masks discrimination based on a sensitive feature, often race
  - Historic practice of color-coding a map based partly on racial and ethnic demographics and designating certain neighborhoods as risky to loan to
  - Modern equivalent → Using a person's address as a feature to determine their insurance rate or whether they qualify for a loan
- Sensitive features are redundantly encoded in the location feature

Redundant encoding: the invisible red line

- Why not also remove features that redundantly encode sensitive features?
- Might throw away too much useful information
  - Location information can be useful
- Might be hard to identify which features to remove
  - Which shopping data?
- Deeply engrained in image classification and facial recognition

Other issues (not an exhaustive list)

- Feature selection
- Biased training set
- Perpetuating existing biases
- Proxy labels
- Lack of diversity in tech

Feature selection

- Recall features are information about a data point
- Millions of features we could use
- Need to choose a smaller number of features for most classifiers
- Programmers get to choose which features to use
- Including or excluding certain features may lead to bias
  - Redundant encoding of sensitive features
  - Favoring features which measure one group better than another
- Intersects with lack of diversity in tech
- Can you think of examples?

Biased training set

- Ideal training set → Random sample of data points with accurate labels
- Reality → Nope!
- Biased labeling → How is the training set labeled? How will the bias of a human labeler affect the outcome?
- Biased sampling → Do the data points in the training set represent a random sample of the data points in the real word?
- Example → Bail recommendation software
  - Predict likelihood someone will jump bail to decide whether to release a person on bail and what to set the bail at
  - Also used in sentencing in some places

Perpetuating existing bias

- Algorithms can perpetuate biases existing in society even if humans are trying not to
- Wage gap problem → Different groups of people paid differently
- Human perpetuation → Employers ask about previous salary
- Possible legislative solution → Ban employers from asking about previous salary
- Algorithmic problem → Previously salary can be inferred from other data
  - How do we even know if this is happening?
- Capable of magnifying bias
- Can you think of other examples?

Proxy labels

- Proxy label → Different from true label you want predict
  - Used in classier training when true labels are hard to get
  - Hopefully correlated with true label
- Triage problem → Predict which patients need extra care and attention
- True label to predict → Future healthcare needs
  - Give those patients more attention and preventative care
- Proxy label used → Future healthcare expenses
- Problem → Racial disparities influence healthcare expenses
- Result → Healthier white patients prioritized over sicker black patients
- Good news → Computer science researchers contacted software company and they made improvements

Some solutions

- Fairness
  - Can we make algorithms more fair than human decision-makers?
  - Efforts to define “fair”
  - Actually using sensitive features in the training step
- Accountability
  - Testing algorithms for bias/discrimination
  - Requiring companies to justify their decisions
- Transparency
  - Translating the computer classifiers into something humans can read and interpret
  - Interactive machine learning → Lets us ask an algorithm why it made a decision
  - Huge efforts to understand deep learning

Testing for bias/discrimination

“We set the agents’ gender to female or male on Google’s Ad Settings page. We then had both the female and male groups of agents visit webpages associated with employment. We established that Google used this gender information to select ads, as one might expect. The interesting result was how the ads differed between the groups: during this experiment, Google showed the simulated males ads from a certain career coaching agency that promised large salaries more frequently than the simulated females, a finding suggestive of discrimination.”

-Automated Experiments on Ad Privacy Settings (Datta, Tschantz, and Datta, 2015)

Some questions

- How much data about each of us is collected online (and offline)?
- How are computers/websites/algorithms using that data to make decisions about us or that affect us?
- Can algorithms discriminate and how?
- Can we prevent algorithms from discriminating?
- Can algorithms combat discrimination and how?