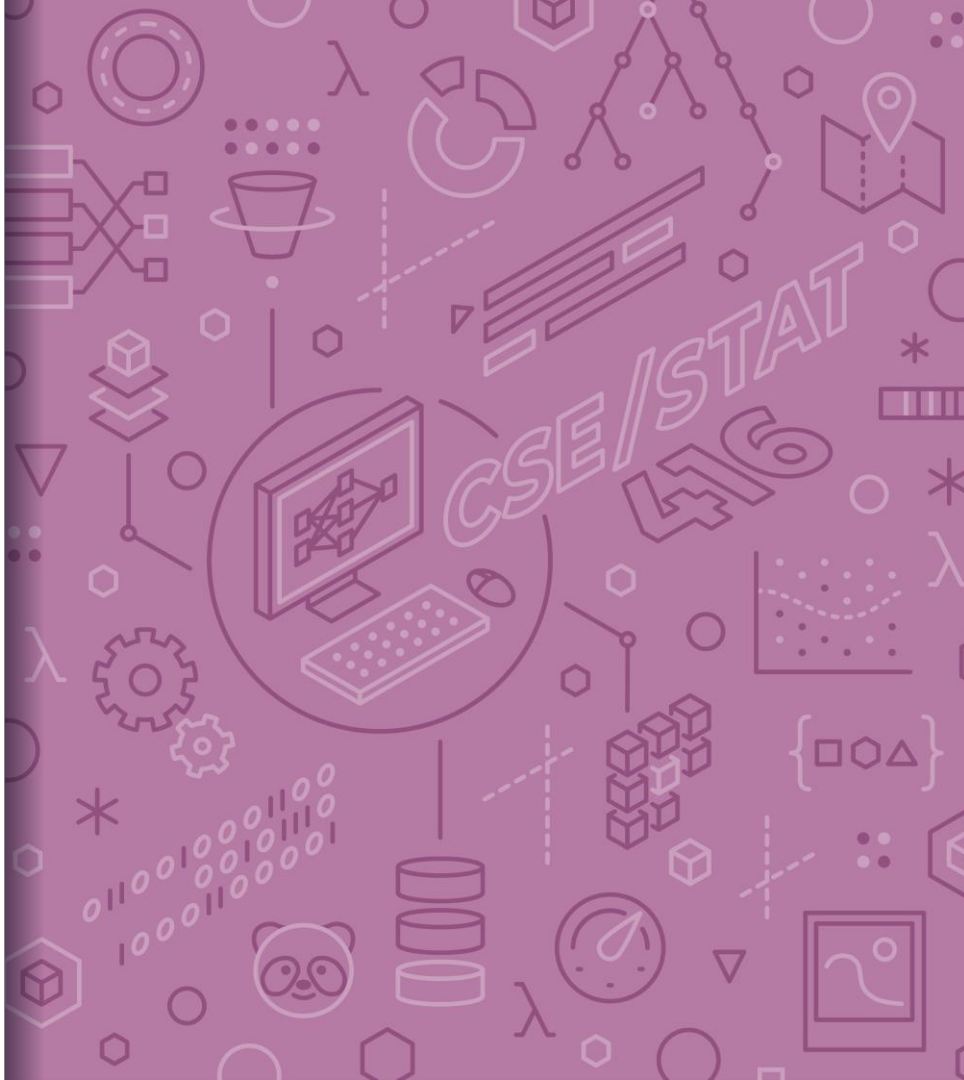
CSE/STAT 416

Fairness and Tradeoffs

Hunter Schafer
Paul G. Allen School of Computer Science & Engineering
University of Washington

April 21, 2021

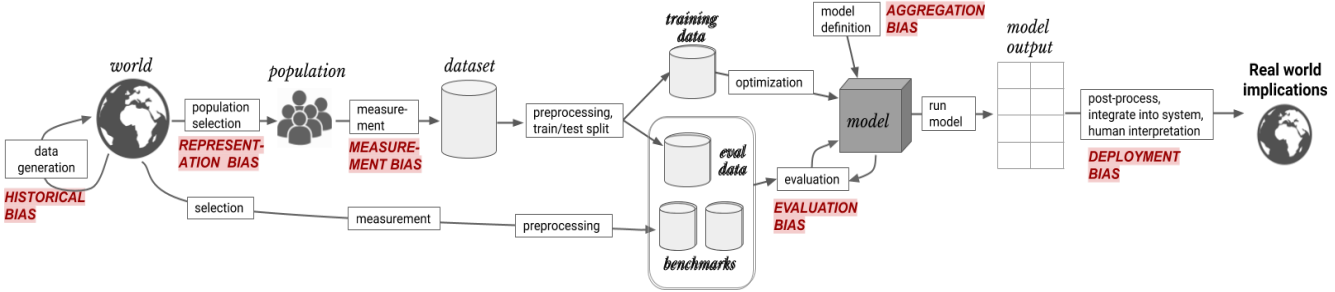


Sources of Bias

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

Fairness

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

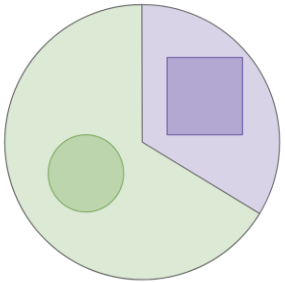
Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren't allowed.
- Different definitions of fairness can be contradictory!

Example: College Admissions

Will use a very simplified example of college admissions. This is **not** an endorsement of such a system or a statement of how we think the world does/should work. Will make MANY simplifying assumptions (which are unrealistic).

- There is a single definition of "success" for college applicants, and the goal of an admissions decision is to predict "success"
- The only thing we will use as part of our decision is SAT Score
- To talk about group fairness, will assume everyone belongs to exactly one of two races: Circles (66%) or Squares (33%).



Group Fairness

- Fairness through Unawareness
- Statistical Parity
  - Require admissions match demographics in data
- Equal Opportunity
  - Require false-negative rate to be equal across groups
- Predictive Equality
  - Require false-positive rate to be equal across groups

(Im)possibility of Fairness

Four reasonable conditions we want in a real world ML Model:

1. Statistical Parity
2. Equal Opportunity (Equality across false negative rates)
3. Predictive Equality (Equality across false positive rates)
4. Good accuracy of the model across subgroups

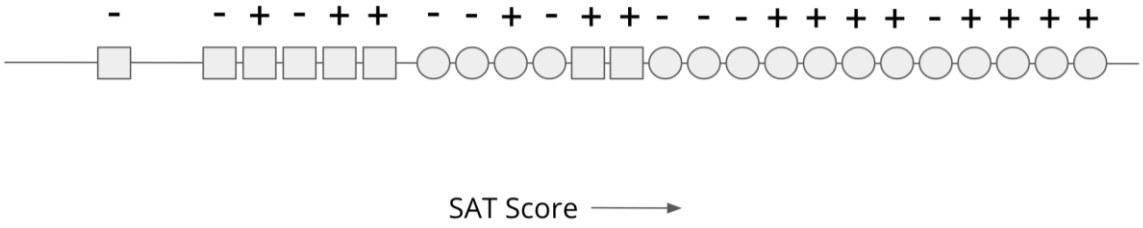
In general, can't satisfy all 4 simultaneously *unless* groups have the exact same underlying distribution.

- This condition is rarely met in practice as we mentioned earlier when there are so many places for bias to enter our data collection.

College Admissions - Continued

Continuing overly simplistic college admissions example, with a fake dataset.

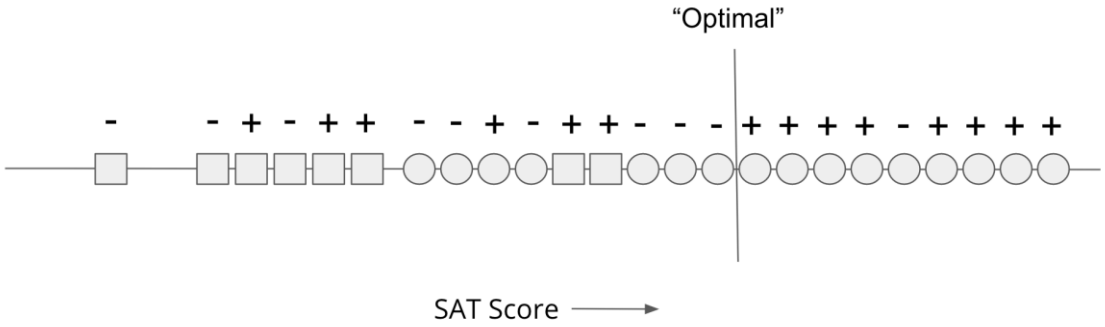
- Majority (2/3) are Circle, the remaining 1/3 are Square
- SAT score for Circles tends to be inflated when compared to Squares. Possibility: Systematic barriers and access to SAT Prep
- Even though we see statistical differences between groups in our data, the rate in which they are actually successful is the same.



Accuracy and Fairness

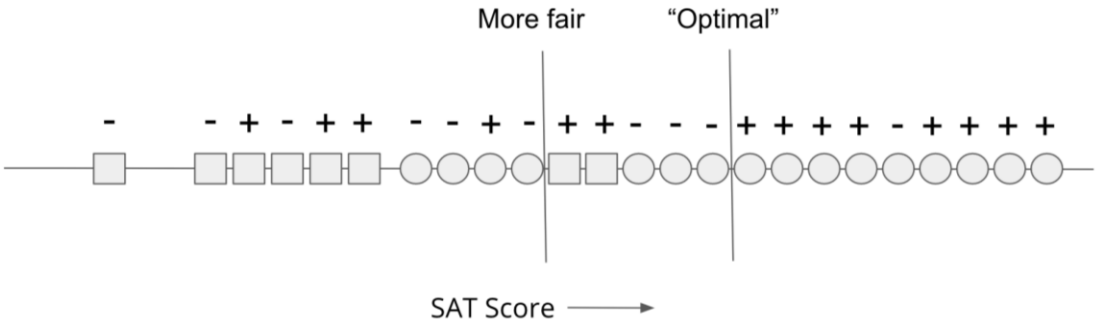
With only one feature, we will consider a simple threshold classifier (a linear classifier with 1 input!).

The most accurate model is not necessarily the most fair.



Fairness-Accuracy Tradeoff

In general, we find there is a tradeoff between accurate models and fair models. Making a model more fair tends to decrease accuracy by some amount.



Notes on Tradeoff

Might argue that my example is overly simplistic (it is!), but I'll claim this is a proof of concept. We saw lots of examples of "accurate" models that were unfair.

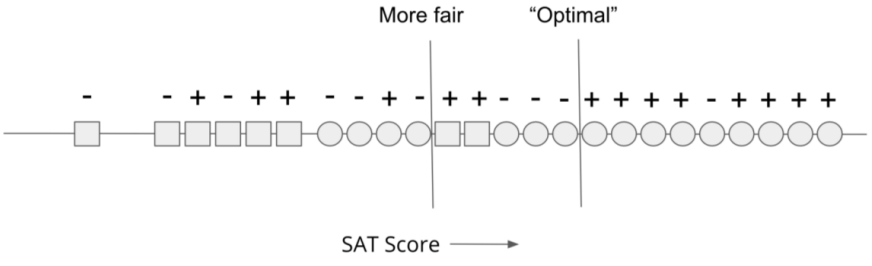
This is not a statement that a tradeoff necessarily must exist, it just generally happens in real-world datasets.

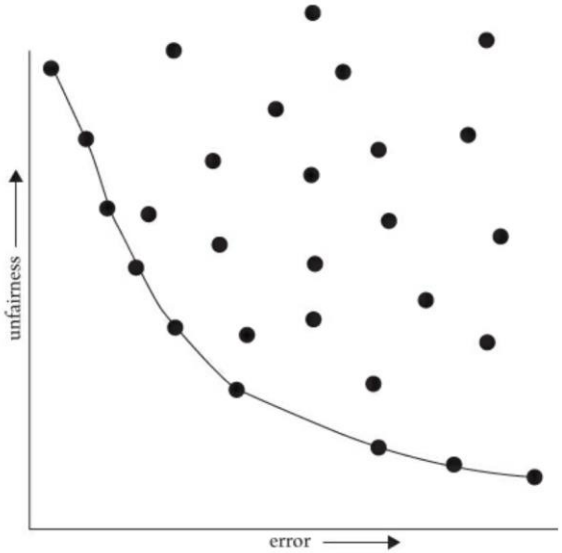
- Originally just cared about finding the most accurate model, saw unfairness as a byproduct. Controlling for fairness will yield a different model than you found before.
- If we recognize data can encode biases and accuracy is determined in terms of that data, trying to achieve fairness will likely hurt accuracy.
  - In the example before, the artificial difference in SAT scores caused the problem.

Pareto Frontier

Visualizing the tradeoff between fairness and accuracy

- **Does not** tell you which tradeoff is appropriate!





Thoughts on Pareto Frontier

> This feels a bit cold-hearted, it's okay to like this is weird. Michael Kearns and Aaron Roth write in *The Ethical Algorithm*

While the idea of considering cold, quantitative trade-offs between accuracy and fairness might make you uncomfortable, the point is that there is simply no escaping the Pareto frontier. Machine learning engineers and policymakers alike can be ignorant of it or refuse to look at it. But once we pick a decision-making model (which might in fact be a human decision-maker), there are only two possibilities. Either that model is not on the Pareto frontier, in which case it's a "bad" model (since it could be improved in at least one measure without harm in the other), or it is on the frontier, in which case it implicitly commits to a numerical weighting of the relative importance of error and unfairness. Thinking about fairness in less quantitative ways does nothing to change these realities—it only obscures them.

Making the trade-off between accuracy and fairness quantitative does **not** remove the importance of human judgment, policy, and ethics—it simply focuses them where they are most crucial and useful, which is in deciding exactly which model on the Pareto frontier is best (in addition to choosing the notion of fairness in the first place, and which group or groups merit protection under it, [...]). Such decisions should be informed by many factors that cannot be made quantitative, including what the societal goal of protecting a particular group is and what is at stake. Most of us would agree that while both racial bias in the ads users are shown online and racial bias in lending decisions are undesirable, the potential harms to individuals in the latter far exceed those in the former. So in choosing a point on the Pareto frontier for a lending algorithm, we might prefer to err strongly on the side of fairness—for example, insisting that the false rejection rate across different racial groups be very nearly equal, even at the cost of reducing bank profits. We'll make more mistakes this way—both false rejections of creditworthy applicants and loans granted to parties who will default—but those mistakes will not be disproportionately concentrated in any one racial group.

Brain Break

Fairness as Worldview

Context

So far have discussed notions of **group fairness**, but other notions of fairness exist. Provide a framework for how to approach learning tasks and what assumptions we make. Based on Friedler et al. (2016).

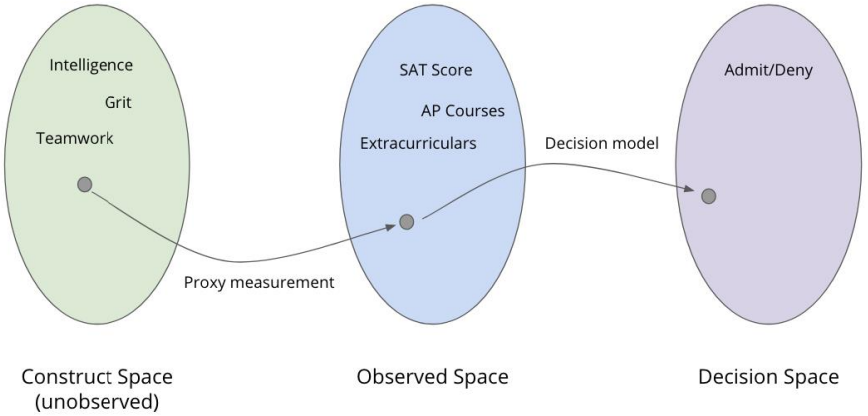
High level ideas:

- Data gathering and modeling
- Individual fairness vs. group fairness
- Common world-views that dictate which fairness is appropriate
- How these worldviews can contradict each other

ML and Spaces

Defined modeling as transformation through three spaces

- **Construct space:** True quantities of interest (unobserved)
- **Observed space:** Data gathered to (hopefully) represent constructs. Achieved through measurement of proxies.
- **Decision space:** The decisions of the model. Models take observed data and make decisions.

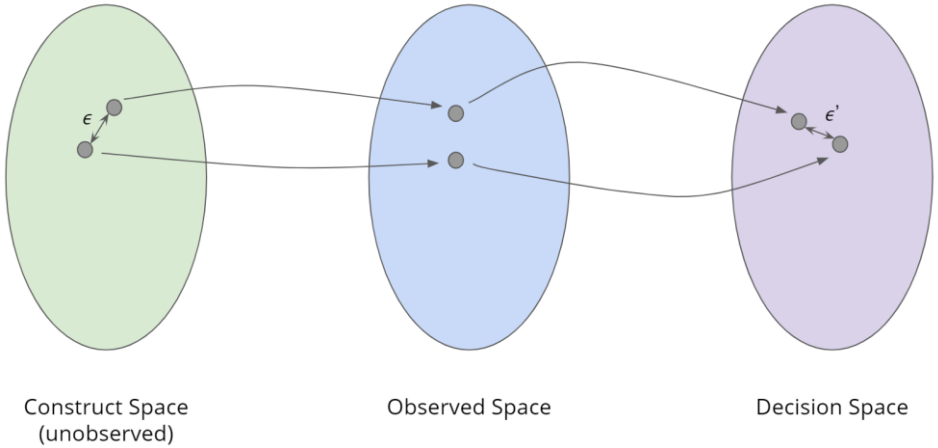


Individual Fairness

Idea: If two people are close in construct space, they should receive similar decisions.

Individual Fairness: A model $f: CS \rightarrow DS$ is said to be fair if objects close in CS are close in DS. Specifically, it is $(\varepsilon, \varepsilon')$-fair if for any $x, y \in CS$

$$d_{CS}(x, y) \leq \varepsilon \quad \Rightarrow \quad d_{DS}\big(f(x), f(y)\big) \leq \varepsilon'$$



Worldview 1: WYSIWYG

Problem: We can't tell if two objects are close in CS. So if we want to use individual fairness, we must make an assumption about how the world workds

What You See is What You Get (WYSIWYG): The Observed Space is a good representation of the Construct Space.

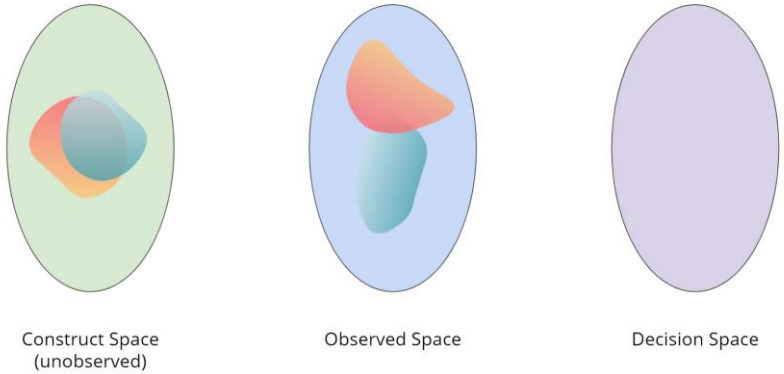
- Example: For college admissions, things like SAT correlate well with intelligence.

With WYSIWYG, you can ensure fairness by comparing objects in the Observed Space as a good proxy for the Construct Space

Worldview 2: Structural Bias + WAE

What if we don't believe the Observed Space represents the Construct Space well? What if there is some **structural bias** that make people close in the construct space look different in the observed space?

- Example: SAT doesn't just measure intelligence, but also measures ability to afford SAT prep. People who are just as intelligent as someone else, can end up with different observations.



Worldview 2: Structural Bias + WAE

When considering Structural Bias, commonly will also assume **We're All Equal (WAE).**

We're All Equal (WAE): Membership in some protected group (e.g., race) *should not* be the cause of a meaningful difference for the task at hand (e.g., academic preparation). Not saying every group is exactly equal in all ways, but for the task at hand we are equal enough that it shouldn't be the cause of difference.

- Differences seen in groups in Observed Space are the result of structural bias!

Notions of group fairness make sense with Structural Bias + WAE

Which One?

So which is right? WYSIWYG or Structural Bias + WAE?

- No way to know! They are statements of belief!
- Which worldview you use determines what you think is fair

If you assume WYSIWYG

- Individual fairness is right and easy to achieve
- Non-discrimination may violate individual fairness

If you assume Structural Bias + WAE

- Non-discrimination is right and is possible (saw group fairness mechanisms)
- Attempts to achieve individual fairness may result in discrimination.

Takeaways

- Models can have a huge impact on society, both positive and negative.
  - If we are not careful, our models will at best, perpetuate and at worst, amplify injustice in our society.
- Historically, people thought defining things like accuracy was easy but defining what is/isn't fair was not. Only recently (~10 years) have ML researchers tried to define what fairness might mean and how to enforce it in our models.
- It's clear that defining and enforcing fairness, but what fairness and how is a crucial problem we need humans (and not just ML engineers) in the loop to determine. These are questions of values, and we need humans to make informed decisions of what is right.

Recap

Theme: Thinking about fairness and the limitations of learning as a worldview.

Concepts:

- Impossibility to achieve all fairness and accuracy
- Fairness-accuracy tradeoff
- Pareto Frontier
- Modeling Spaces
  - Construct space
  - Observed space
  - Decision space
- Individual fairness
- What You See is What You Get (WYSIWYG)
- Structural Bias + We're All Equal (WAE)
- Conflicting Worldviews

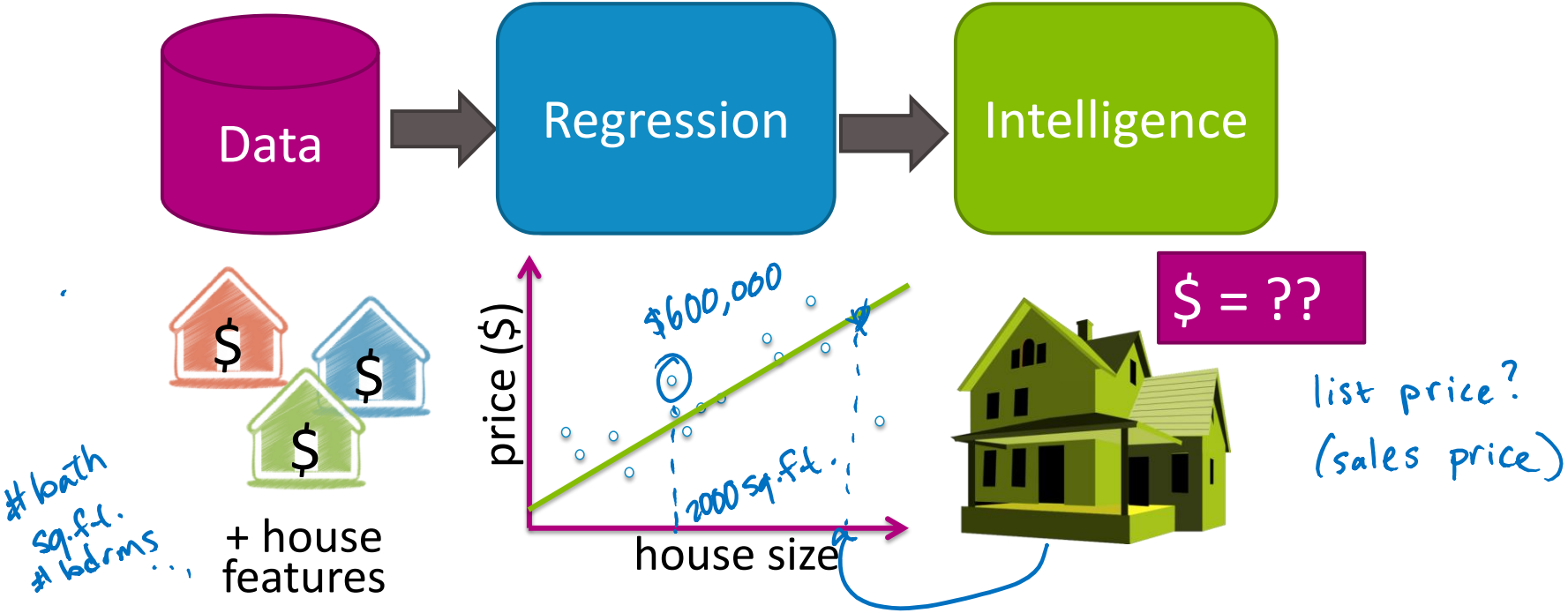
Brain Break

One Slide

- Regression
- Overfitting
- Training, test, and generalization error
- Bias-Variance tradeoff
- Ridge, LASSO
- Cross validation
- Gradient descent
- Classification
- Logistic regression
- Bias and Fairness

Case Study 1:

Predicting house prices



Regression

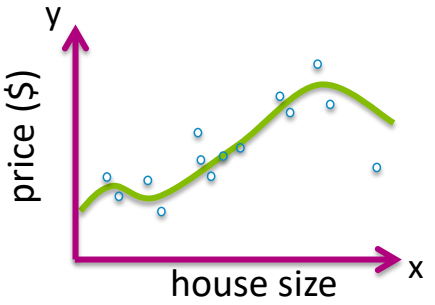
Case study: Predicting house prices

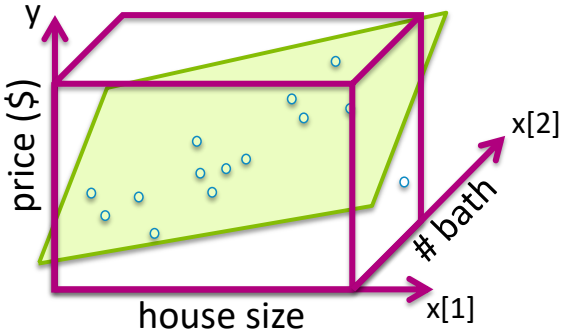
Models

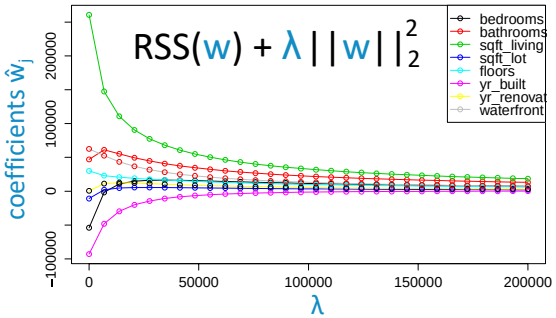
- Linear regression
- Regularization:
  Ridge (L2), Lasso (L1)

Including many features:

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- …







Regression

Case study: Predicting house prices

Algorithms

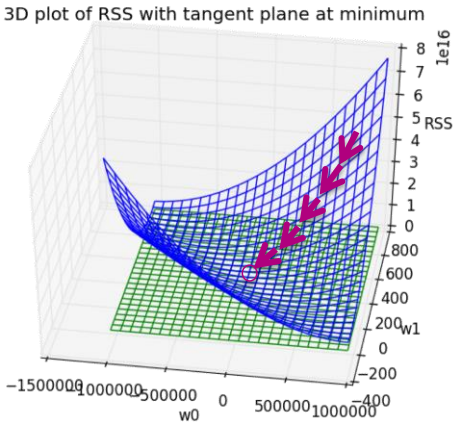
- Gradient descent

$$RSS(w_0,w_1) = (\$_{house\ 1}-[w_0+w_1 sq.ft._{house\ 1}])^2 + (\$_{house\ 2}-[w_0+w_1 sq.ft._{house\ 2}])^2 + (\$_{house\ 3}-[w_0+w_1 sq.ft._{house\ 3}])^2 + \ldots$$

[include all houses]

$$\downarrow$$

$$\hat{w}$$

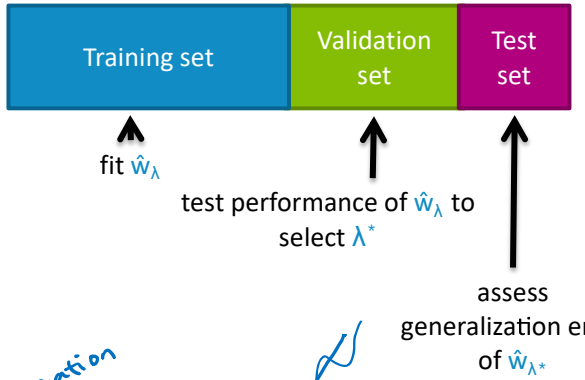


Regression

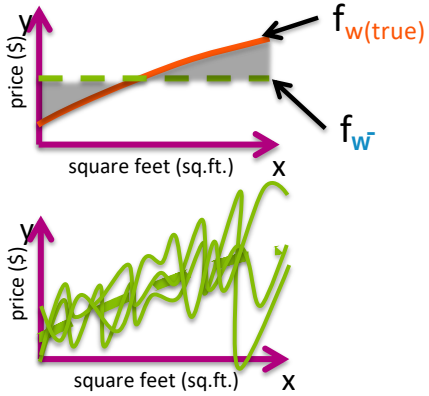
Case study: Predicting house prices

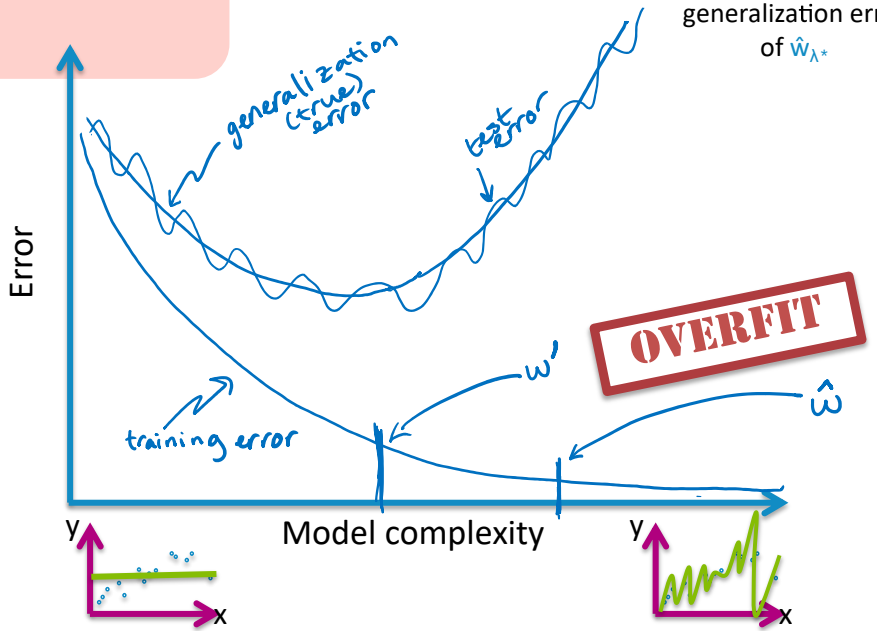
Concepts

- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection



1. Noise
2. Bias
3. Variance





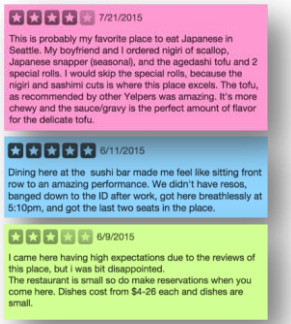
Case Study 2:
Sentiment analysis

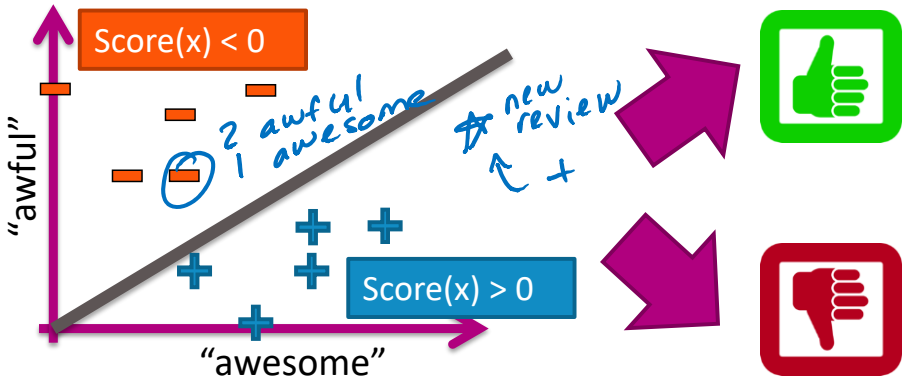
Sushi was awesome,
the food was awesome,
but the service was awful.

stars / +/-

text

All reviews:



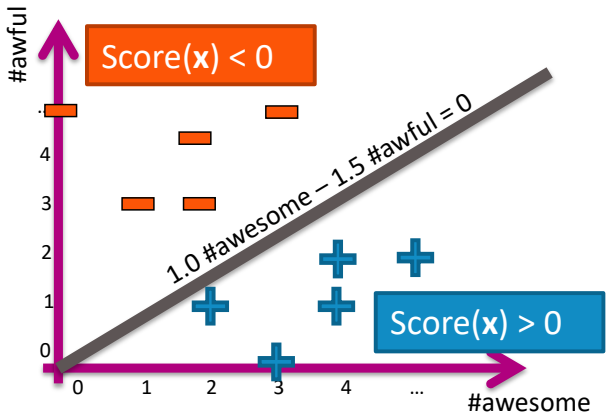


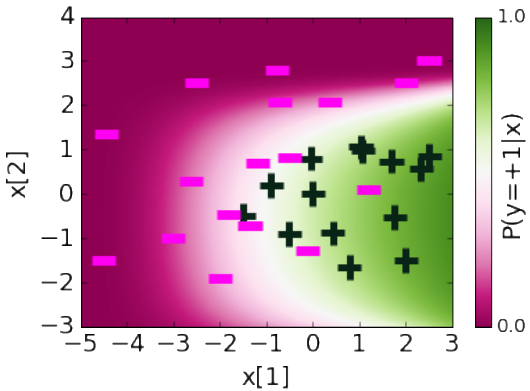
Classification

Case study: Analyzing sentiment

Models

- Linear classifiers (logistic regression)





Classification

Case study: Analyzing sentiment

Concepts

- Decision boundaries, maximum likelihood estimation