Fairness, Accountability, and Transparency

Machine Learning: Jordan Boyd-Graber
University of Maryland
NEED FOR INTERPRETABILITY

Trust Part of ML Pipeline

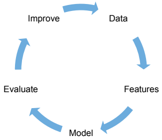
Learn model → Trust model → Deploy model

Trust AI system

NETFLIX

Make better decisions

Improve model

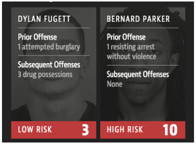


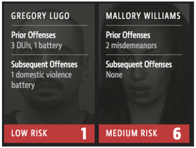
ML is Everywhere

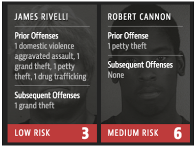
- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment

ML is Everywhere

- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment

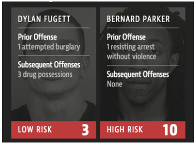


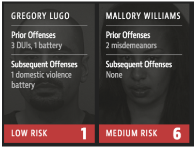


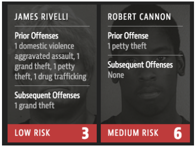


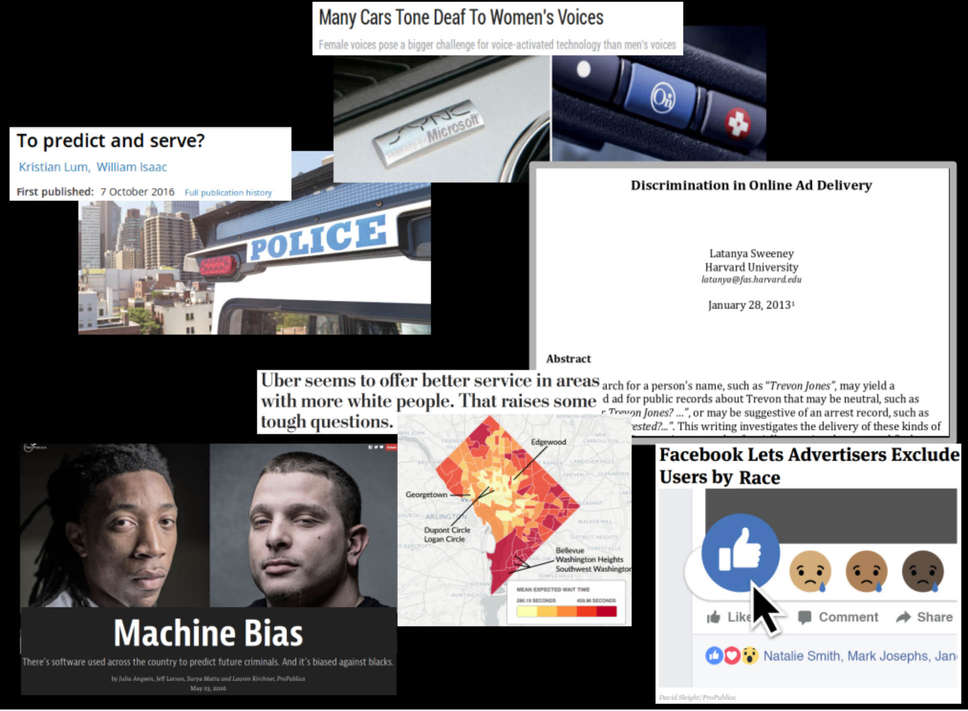
ML is Everywhere

- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment
- How do we know it isn't being incompetent/evil?









Many Cars Tone Deaf To Women's Voices
Female voices pose a bigger challenge for voice-activated technology than men's voices
Microsoft
To predict and serve?
Kristian Lum, William Isaac
First published: 7 October 2016 Full publication history
POLICE
Discrimination in Online Ad Delivery
Latanya Sweeney
Harvard University
latanya@fas.harvard.edu
January 28, 2013¹
Abstract
arch for a person's name, such as "Trevon Jones", may yield a
d ad for public records about Trevon that may be neutral, such as
r Trevon Jones? ...", or may be suggestive of an arrest record, such as
rested?...". This writing investigates the delivery of these kinds of
Uber seems to offer better service in areas with more white people. That raises some tough questions.
Edgewood
Georgetown
Dupont Circle
Logan Circle
Bellevue
Washington Heights
Southwest Washington
MEAN EXPECTED WAIT TIME
290.19 SECONDS
459.96 SECONDS
Facebook Lets Advertisers Exclude Users by Race
Like
Comment
Share
Natalie Smith, Mark Josephs, Jan
David Sleight/ProPublica
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Keep it Simple (Stupid)

- Clear preference for interpretability
- Even at the cost of performance: decision trees still popular
- But what about all of the great machine learning we've talked about?

Pneumonia Example (Caruana)

- Prediction task:
  - LOW Risk: outpatient: antibiotics, call if not feeling better
  - HIGH Risk: admit to hospital (10% of pneumonia patients die)
- Most accurate ML method: multitask neural nets

Pneumonia Example (Caruana)

- Prediction task:
  - LOW Risk: outpatient: antibiotics, call if not feeling better
  - HIGH Risk: admit to hospital (10% of pneumonia patients die)
- Most accurate ML method: multitask neural nets
- Used logistic regression

Pneumonia Example (Caruana)

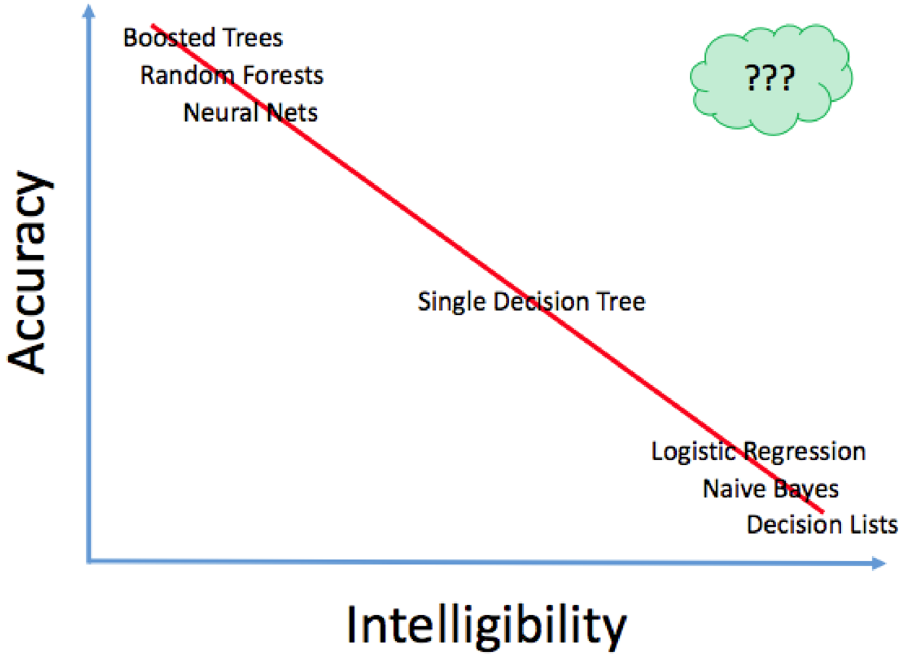
- Prediction task:
  - LOW Risk: outpatient: antibiotics, call if not feeling better
  - HIGH Risk: admit to hospital (10% of pneumonia patients die)
- Most accurate ML method: multitask neural nets
- Used logistic regression
- Learned rule: HasAsthma(x) $\longrightarrow$ LessRisk(x)

Why?

- asthmatics presenting with pneumonia considered very high risk
- receive agressive treatment and often admitted to ICU
- history of asthma also means they often go to healthcare sooner
- treatment lowers risk of death compared to general population

Lessons Learned (Caruana)

- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you donâĂŹt understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop


Boosted Trees
Random Forests
Neural Nets
???
Accuracy
Single Decision Tree
Logistic Regression
Naive Bayes
Decision Lists
Intelligibility

We've already seen problems

- Gender/racial bias
- Generalization failures
- Malicious Input

We've already seen problems

- Gender/racial bias
- Generalization failures
- Malicious Input

Can we just remove problematic variables?

- Not obvious *a priori*
- Can find correlated features
- More of a problem in deep learning

Subject for Today

- How to measure interpretability
- How to fix biased data
- How to unbias supervised algorithms