Interpretability

Advanced Machine Learning for NLP

Jordan Boyd-Graber

NEED FOR INTERPRETABILITY

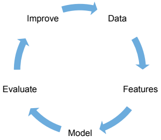
Trust Part of ML Pipeline

Learn model → Trust model → Deploy model

Trust AI system

Make better decisions

Improve model

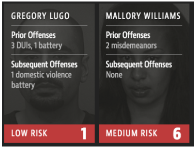


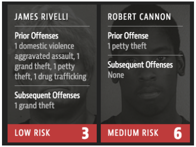
ML is Everywhere

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

ML is Everywhere

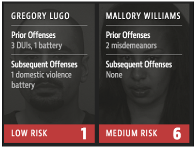
- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment

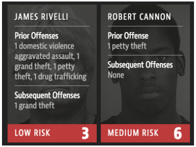


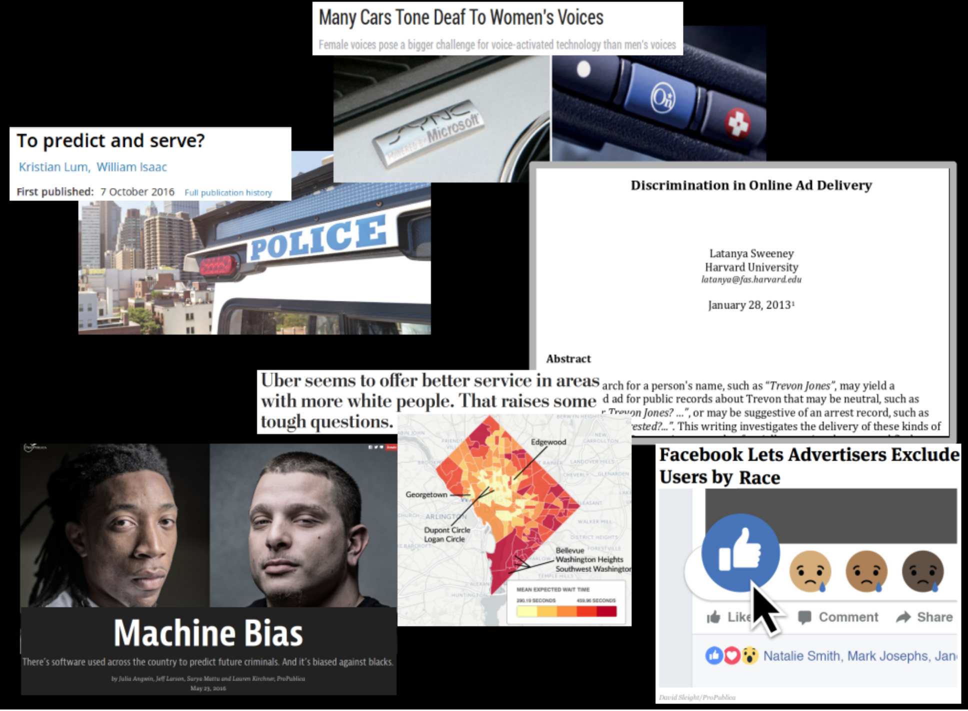


ML is Everywhere

- 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, 2013
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
", or may be suggestive of an arrest record, such as
". 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
ARLINGTON
Dupont Circle
Logan Circle
Bellevue
Washington Heights
Southwest Washington
MEAN EXPECTED WAIT TIME
Facebook Lets Advertisers Exclude Users by Race
Like
Comment
Share
Natalie Smith, Mark Josephs,
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?

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

- Intrinsic evaluation: topic models
- Intrinsic evaluation: embeddings
- Extrinsic evaluation: supervised ML
- Extrinsic evaluation: visualizations for supervised ML