



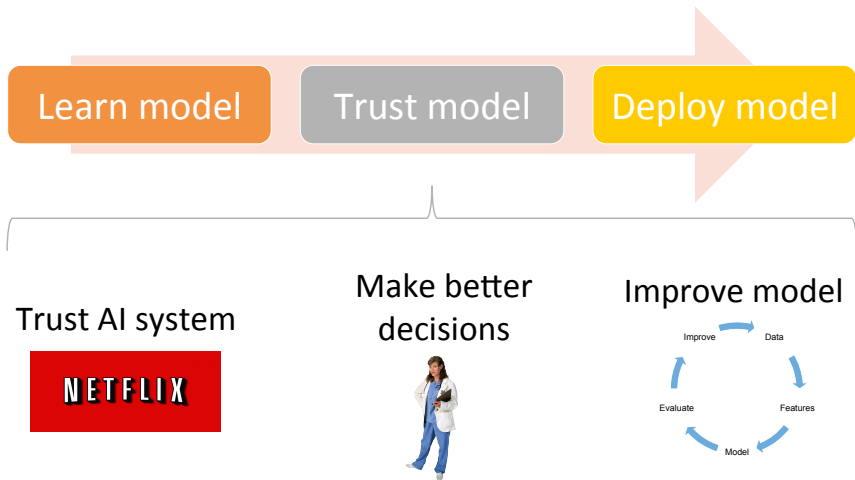
## Interpretability

### Advanced Machine Learning for NLP

Jordan Boyd-Graber

NEED FOR INTERPRETABILITY

## Trust Part of ML Pipeline



## 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

DYLAN FUGETT	BERNARD PARKER
Prior Offense 1 attempted burglary	Prior Offense 1 resisting arrest without violence
Subsequent Offenses 3 drug possessions	Subsequent Offenses None
LOW RISK 3	HIGH RISK 10

GREGORY LUGO	MALLORY WILLIAMS
Prior Offenses 3 DUIs, 1 battery	Prior Offenses 2 misdemeanors
Subsequent Offenses 1 domestic violence battery	Subsequent Offenses None
LOW RISK 1	MEDIUM RISK 6

JAMES RIVELLI	ROBERT CANNON
Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking	Prior Offense 1 petty theft
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	MEDIUM RISK 6

## ML is Everywhere

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

DYLAN FUGETT	BERNARD PARKER
Prior Offense 1 attempted burglary	Prior Offense 1 resisting arrest without violence
Subsequent Offenses 3 drug possessions	Subsequent Offenses None
LOW RISK 3	HIGH RISK 10

GREGORY LUGO	MALLORY WILLIAMS
Prior Offenses 3 DUIs, 1 battery	Prior Offenses 2 misdemeanors
Subsequent Offenses 1 domestic violence battery	Subsequent Offenses None
LOW RISK 1	MEDIUM RISK 6

JAMES RIVELLI	ROBERT CANNON
Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking	Prior Offense 1 petty theft
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	MEDIUM RISK 6

## Many Cars Tone Deaf To Women's Voices

Female voices pose a bigger challenge for voice-activated technology than men's voices



## To predict and serve?

Kristian Lum, William Isaac

First published: 7 October 2016 Full publication history



## Discrimination in Online Ad Delivery

Latanya Sweeney  
Harvard University  
latanya@fas.harvard.edu

January 28, 2013<sup>1</sup>

### Abstract

Uber seems to offer better service in areas with more white people. That raises some tough questions.

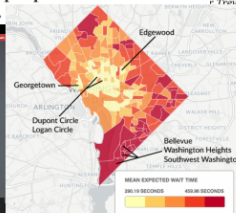
Search for a person's name, such as "Trevon Jones", may yield a ad for public records about Trevon that may be neutral, such as "Trevon Jones? ...", or may be suggestive of an arrest record, such as "Trevon Jones? ...". This writing investigates the delivery of these kinds of



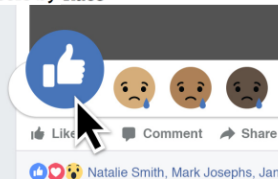
## Machine Bias

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



## Facebook Lets Advertisers Exclude Users by Race



David Haight/ProPublica

## Keep it Simple (Stupid)

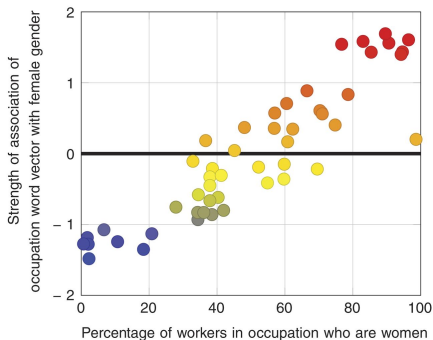
---

- 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