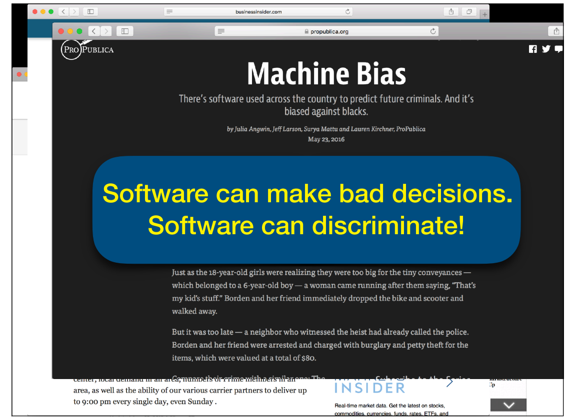
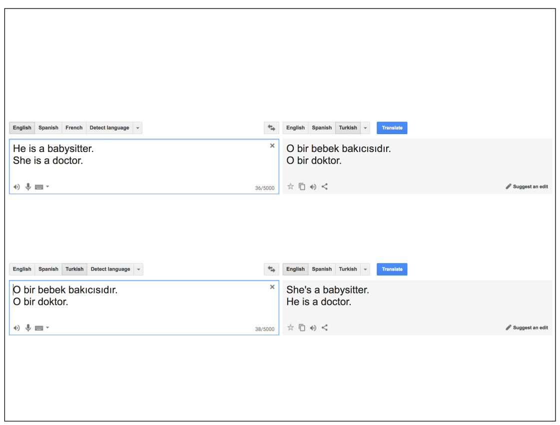
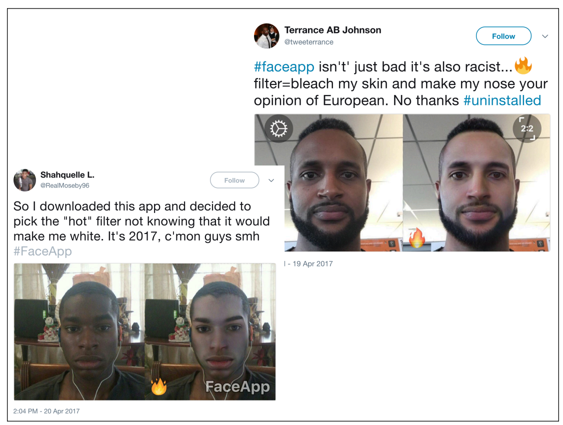
Fairness Testing

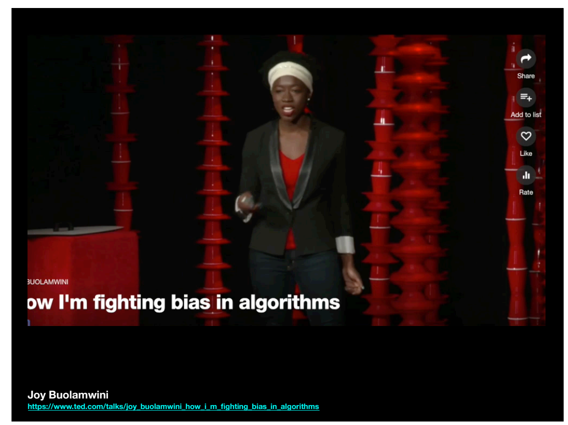
Fairness Testing: Testing Software for Discrimination
ESEC/FSE 2017

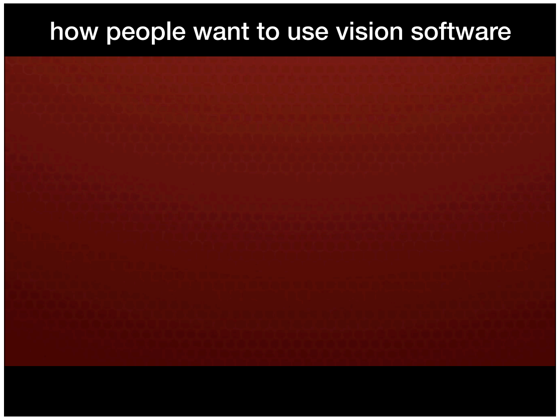
http://tinyurl.com/FairnessPaper



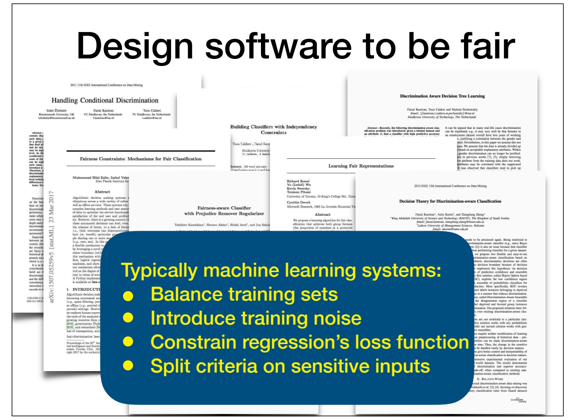


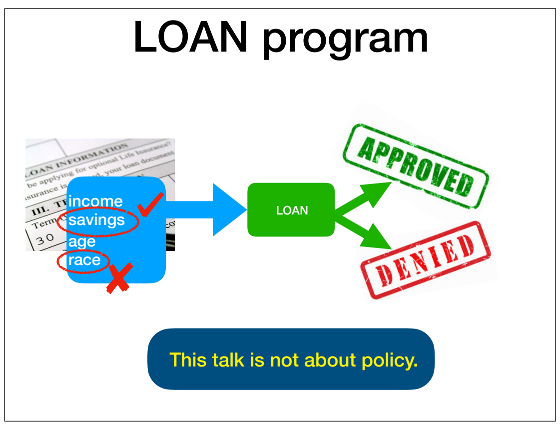


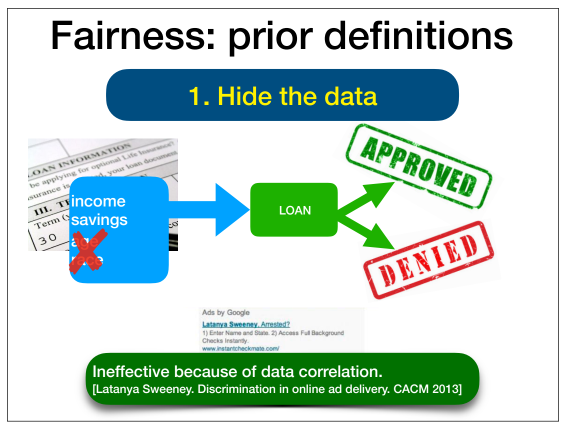


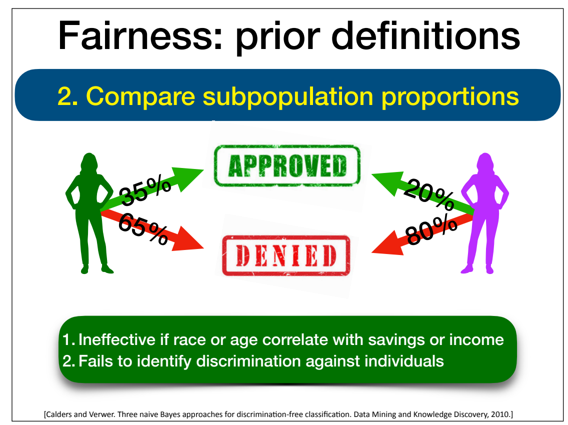

how people want to use vision software


today's goals
Define software discrimination.
Operationalize measuring discrimination through causal software testing.

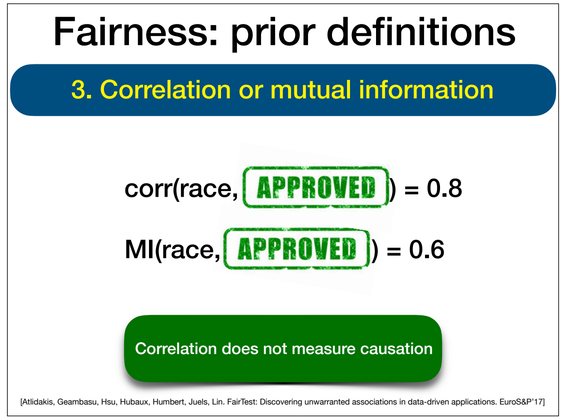

Design software to be fair
Handling Conditional Discrimination
Building Classifiers with Independency Constraints
Discrimination Aware Decision Tree Learning
Fairness Constraints: Mechanisms for Fair Classification
Learning Fair Representations
Fairness-aware Classifier with Prejudice Remover Regularizer
Decision Theory for Discrimination-aware Classification
Typically machine learning systems:
Balance training sets
Introduce training noise
Constrain regression's loss function
Split criteria on sensitive inputs


LOAN program
income
savings
age
race
LOAN
APPROVED
DENIED
This talk is not about policy.

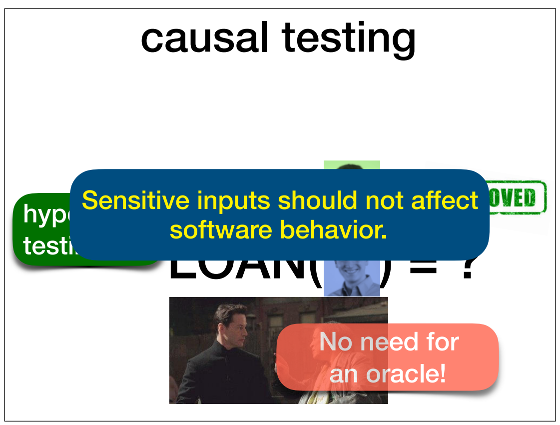

Fairness: prior definitions
1. Hide the data
income
savings
LOAN
APPROVED
DENIED
Ads by Google
Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/
Ineffective because of data correlation.
[Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]

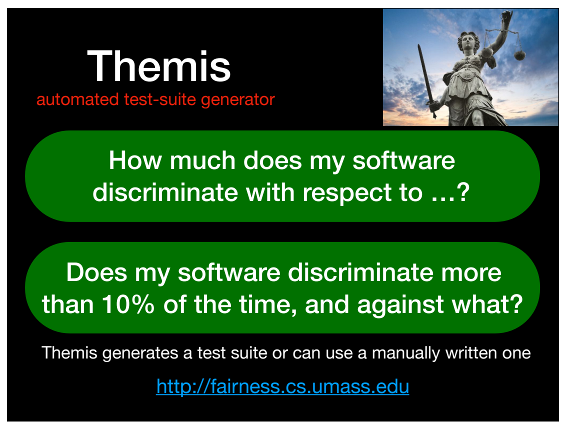

Fairness: prior definitions
2. Compare subpopulation proportions
35%
65%
APPROVED
DENIED
20%
80%
1. Ineffective if race or age correlate with savings or income
2. Fails to identify discrimination against individuals
[Calders and Verwer. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 2010.]

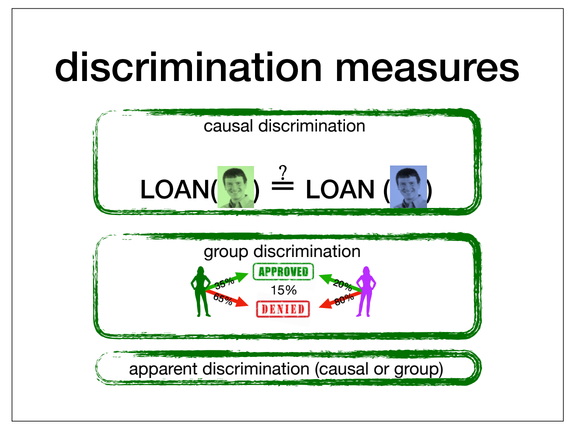

where group discrimination fails
Europe
Asia
recommend loans to all green and to no purple applicants
recommend loans to all purple and to no green applicants
Group discrimination can be 0.

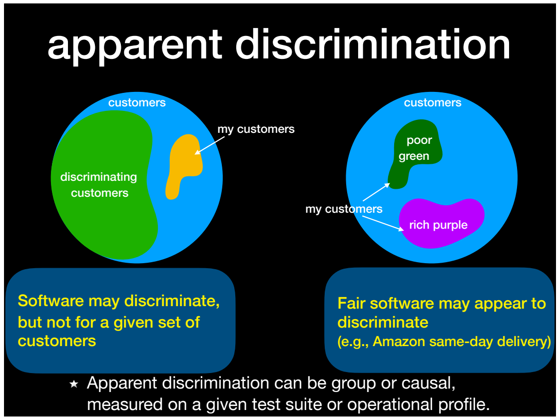

Fairness: prior definitions
3. Correlation or mutual information
corr(race, APPROVED) = 0.8
MI(race, APPROVED) = 0.6
Correlation does not measure causation
[Atlidakis, Geambasu, Hsu, Hubaux, Humbert, Juels, Lin. FairTest: Discovering unwarranted associations in data-driven applications. EuroS&P'17]

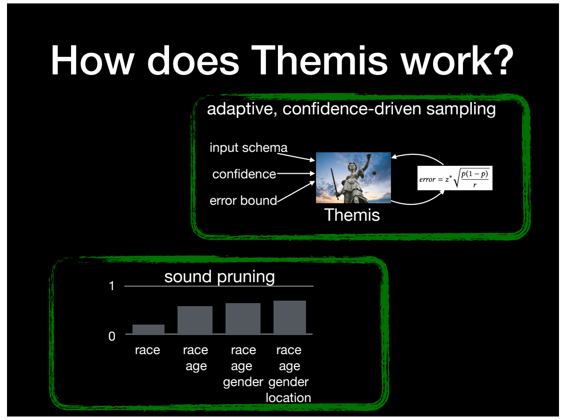

What is fairness?
Sensitive inputs should not affect software behavior.
We want to measure causality!
[Judea Pearl. Causal inference in statistics: An overview. Statistics Surveys 2009]

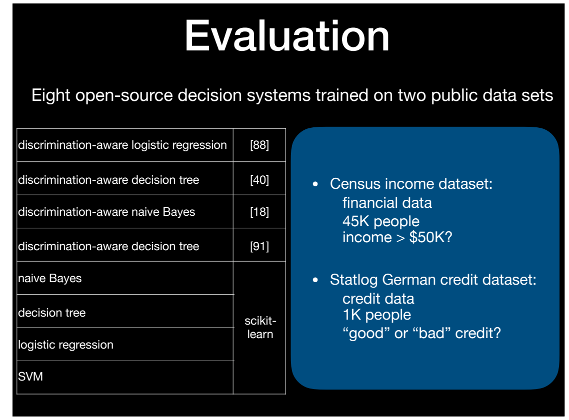

causal testing
Sensitive inputs should not affect software behavior.
No need for an oracle!


Themis
automated test-suite generator
How much does my software discriminate with respect to …?
Does my software discriminate more than 10% of the time, and against what?
Themis generates a test suite or can use a manually written one
http://fairness.cs.umass.edu

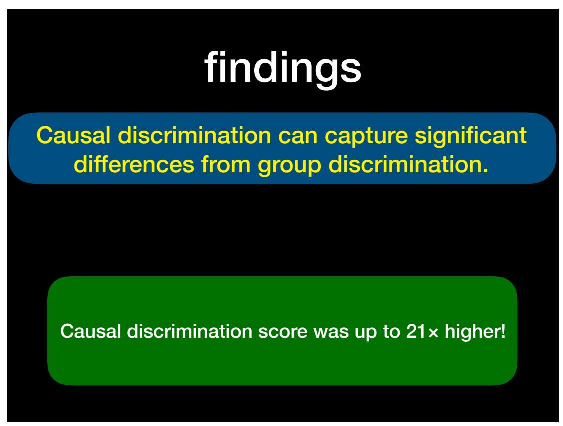

discrimination measures
causal discrimination
LOAN( ) ≟ LOAN ( )
group discrimination
APPROVED
15%
DENIED
apparent discrimination (causal or group)


apparent discrimination
customers
discriminating customers
my customers
customers
poor green
my customers
rich purple
Software may discriminate, but not for a given set of customers
Fair software may appear to discriminate
(e.g., Amazon same-day delivery)
★ Apparent discrimination can be group or causal, measured on a given test suite or operational profile.


How does Themis work?
adaptive, confidence-driven sampling
input schema
confidence
error bound
Themis
sound pruning
1
0
race
race age
race age gender
race age gender location


Evaluation
Eight open-source decision systems trained on two public data sets
discrimination-aware logistic regression [88]
discrimination-aware decision tree [40]
discrimination-aware naive Bayes [18]
discrimination-aware decision tree [91]
naive Bayes
decision tree
logistic regression
SVM
scikit-learn
• Census income dataset:
financial data
45K people
income > $50K?
• Statlog German credit dataset:
credit data
1K people
“good” or “bad” credit?


findings
Group discrimination is not enough.
More than 11% of the individuals had the output flipped just by altering the individual’s gender.
Decision tree trained not to group discriminate against gender causal discriminated against gender: 0.11.


findings
Causal discrimination can capture significant differences from group discrimination.
Causal discrimination score was up to 21× higher!


findings
Trying to avoid group discrimination may introduce other discrimination.
Training a decision tree not to discriminate against gender made it discriminate against race 38.4% of the time.

findings

Pruning is highly effective.

- The more a system discriminates, the more efficient Themis is.
- On average, pruning reduced test suites by 148× for causal and 2,849× for group discrimination. Best improvement was 13,000×.

related work

Ways of measuring discrimination

- CV score [19]
- correlation, mutual information [79]
- Output probability distributions [51]

Discrimination-aware algorithms
[18, 40, 88, 91]

Measuring discrimination with manually-written tests [79]

Causal model inference
[Maier et al., UAI'13]

Fairness verification
[Albarghouthi et al., OOPSLA'17]

Contributions

http://fairness.cs.umass.edu

- Causality-based definition and method for measuring software fairness
- Themis, an automated test-suite generator for fairness testing
- Provably-sound pruning test-suite reductions
- Evaluation on real-world software, demonstrating Themis' effectiveness