FairTest: Discovering Unwarranted Associations in Data-Driven Applications

Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, Huang Lin

2017 IEEE European Symposium on Security and Privacy

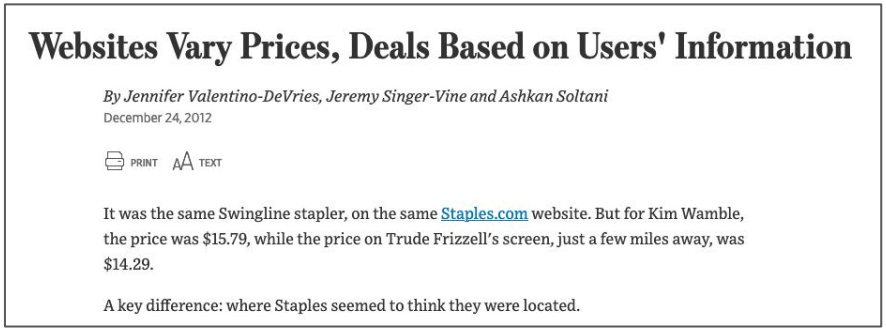
Presented by: Shubham Singh

Outline

- Problem Statement
- Contributions
- UA Framework
- FairTest Design
- Evaluation
- Discussion

Problem Statement

- User data collected by companies used to train algorithms.
- Such algorithms can lead to unwanted consequences -- causing harm to the users.
- The authors treat these biases as *bugs*.



Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani
December 24, 2012

PRINT TEXT

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

TECH

Google Photos labeled black people 'gorillas'

Jessica Guynn USA TODAY
Published 1:15 p.m. ET Jul. 1, 2015 | **Updated 2:10 p.m. ET Jul. 1, 2015**

How big data is unfair

Understanding unintended sources of unfairness in data driven decision making

Moritz Hardt Sep 26, 2014 · 8 min read

Association Bug

- Fairness in prior works -- Strong statistical dependency between algorithm output and protected user groups.
- But they lack
  - Wide-applicability
  - Scalable assessment
  - Inclusion of natural explanatory factors (Berkeley admissions)
- They define association bug in context of subpopulation and presence of explanatory factors, use it in a testing toolkit for a wide-variety of tasks and datasets.
- It’s a post-processing technique.

Contributions

- Unwarranted Association (UA) Framework
  - Define investigation primitives for widely applicable tasks.
- Association-guided tree construction algorithm
  - Find semantically meaningful user subpopulation.
- FairTest: Testing and Debugging Tool
  - A system and API to run tests.
- Evaluation on synthetic and real-world datasets

Outline

- Problem Statement
- Contributions
- UA Framework
- FairTest Design
- Evaluation
- Discussion

UA Framework

- Unwarranted Association is defined as

Any statistically significant association, in a semantically meaningful user subpopulation, between a protected attribute and an algorithmic output, where the association has no accompanying explanatory factor.

Methodology

- Data Collection and Pre-Processing
  - Output of algorithm: *O*
  - Protected attributes: *S*
  - Contextual attributes: *X*
  - Explanatory attributes: *E*
- Integration and Explanatory Factors
  - Some associations may be acceptable or necessary.
  - E.g. hiring based on qualifications, which can be proxy for age.

Methodology

- Selection of Appropriate Metric
  - Frequency Distribution Metric

$$Ratio = Pr(o_1|s_1)/Pr(o_2|s_2) - 1$$

$$Diff = Pr(o_1|s_1) - Pr(o_2|s_2)$$

  - Mutual Information (MI)

$$MI = \sum_{o,s} Pr(o,s)\ ln(\frac{Pr(o,s)}{Pr(o)Pr(s)})$$

  - Pearson's Correlation
  - Regression
  - Conditional Metric

$$\mathbb{E}_E(M(S;O)|E)$$

Metric	Description	When to Use
Binary Ratio & Difference [8]–[17]	compare probability of an output for groups	binary $\boldsymbol{S}, \boldsymbol{O}$; often for *Testing*
Mutual Information (MI) [18]	dependence measure for discrete variables	categorical $\boldsymbol{S}, \boldsymbol{O}$; often for *Testing*
Pearson Correlation (CORR)	linear dependence measure for scalar variables	scalar $\boldsymbol{S}, \boldsymbol{O}$; often for *Error Profiling*
Regression	for labeled outputs, measure each label's association	high dimensional $\boldsymbol{O}$; always for *Discovery*

TABLE 1. **Association Metrics for the UA Framework.**

Methodology

- Testing Across User Subpopulation
  - Testing on full user population is not enough.
  - Use contextual features X to successively split users into smaller subsets with stronger associations.
- Adaptive Debugging
  - Debugging needs subsequent investigations.
  - Statistical validity.

Core Investigation Primitives

- Testing
  - Using metrics to test for suspected association, conditioned on explanatory feature.
- Discovery
  - Applies to cases with large space outputs, where it's hard to know them *a priori*.
- Error Profiling
  - Measure the accuracy of the classifier w.r.t. to the *utility* to the user.

Outline

- Problem Statement
- Contributions
- UA Framework
- FairTest Design
- Evaluation
- Discussion

FairTest Design

- Association Report Example
  - Simulated Pricing Scheme, similar to Staples'.
  - Gives discounts to user located within 20 mi of competing OfficeDepot stores.
  - Protected attribute: 'income'
  - Output: 'price'
  - Shows contingency table

Report of associations of O=Price on S_i=Income:
Assoc. metric: norm. mutual information (NMI).

Global Population of size 494,436
p-value=3.34e-10 ; NMI=[0.0001, 0.0005]

Price	Income <$50K	Income >=$50K	Total
High	**15301 (6%)**	**13867 (6%)**	29168 (6%)
Low	234167(94%)	231101(94%)	465268 (94%)
Total	249468(50%)	244968(50%)	494436(100%)

1. Subpopulation of size 23,532
Context={State: CA, Race: White}
p-value=2.31e-24 ; NMI=[0.0051, 0.0203]

Price	Income <$50K	Income >=$50K	Total
High	**606 (8%)**	**691 (4%)**	1297 (6%)
Low	7116(92%)	15119(96%)	22235 (94%)
Total	7722(33%)	15810(67%)	23532(100%)

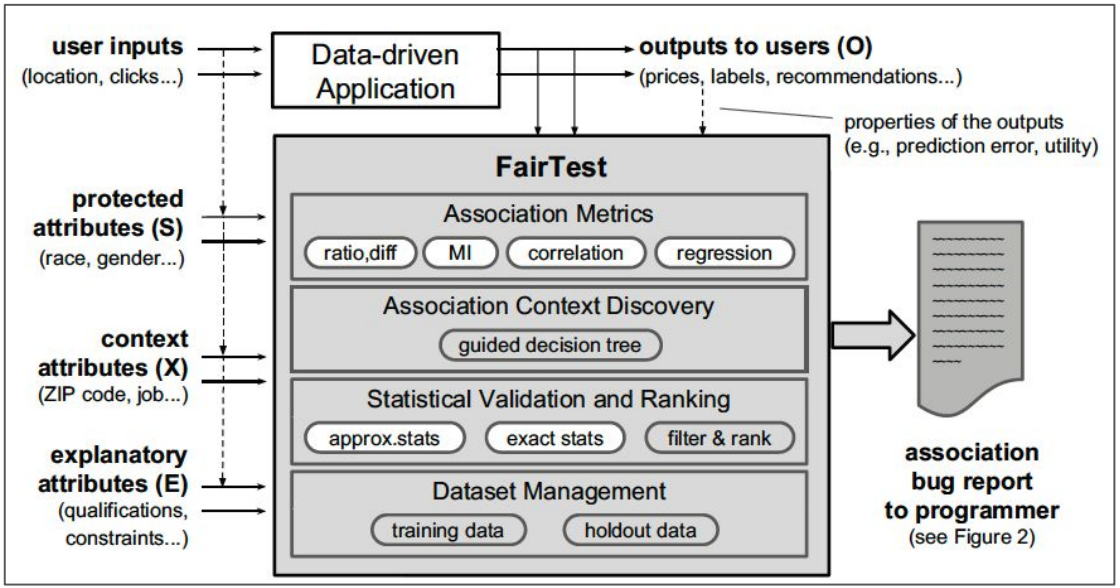
2. Subpopulation of size 2,198
Context={State: NY, Race: Black, Gender: Male}
p-value=7.72e-05 ; NMI=[0.0040, 0.0975]

Price	Income <$50K	Income >=$50K	Total
High	**52 (4%)**	**8 (1%)**	60 (3%)
Low	1201(96%)	937(99%)	2138 (97%)
Total	1253(57%)	945(43%)	2198(100%)

...more entries (sorted by decreasing NMI)...

Architectural Components

- Dataset, $D = (S, X, E, O)$
  - Split into train and test set, D_{train}, D_{test}



Association Context Discovery

- "Zoom into" as user population
- Use guided decision-tree construction
  - Similar to decision-tree learning
- Split D based on $X_i \in X$ into subsets $D = \{D_1, D_2, \ldots\}$
- If X_i is categorical:
  - Split into one subset per value
- If X_i is continuous:
  - Binary splits based on some threshold t
  - To choose t, we sort D with the values of X_i, and test unique value of X_i
- A valid split has higher association than the one over D

Params : MIN_SIZE // Minimum size of a context
MAX_DEPTH // Maximum tree depth
Metric // Association metric

Function findContexts($D = \{\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{E}, \boldsymbol{O}\}, \mathcal{P} = \emptyset$)
  Create a subpopulation defined by predicates $\mathcal{P}$
  if $|D| < MIN\_SIZE$ *or* $|\mathcal{P}| \geq MAX\_DEPTH$ **then return**
  for $\boldsymbol{X}_i \in \boldsymbol{X}$ **do**
    $\mathbb{D} = \{D_1, D_2, \ldots\} \leftarrow$ partition of D based on $\boldsymbol{X}_i$
    if $\exists\, D_k \in \mathbb{D} : \text{Metric}(D_k) > \text{Metric}(D)$ **then**
      // Avg. association for this split
      $\text{Score}_i \leftarrow \sum_{D_k \in \mathbb{D}} \text{Metric}(D_k) / |\mathbb{D}|$
    else $\text{Score}_i \leftarrow 0$
  if $\forall i : Score_i = 0$ **then**
    return // No split yields higher assoc.
  $\boldsymbol{X}_{\text{best}}, \mathbb{D}_{\text{best}} \leftarrow$ partition with highest score
  for $D_k \in \mathbb{D}_{best}$ **do**
    $V \leftarrow$ values taken by $\boldsymbol{X}_{\text{best}}$ in D_k
    findContexts($D_k, \mathcal{P} \cup \{\boldsymbol{X}_{best} \in V\}$)

Statistical Validation and Ranking

- Validation is needed cause Association Context Discovery maximizes association over a finite user sample (D_{train}).
- We need an independent test sample (D_{test}) for validation.
- Ideally, $|D_{train}| = |D_{test}|$
- Employ p-value under null hypothesis and association metrics are estimated with *confidence intervals (CI)*.
- Dataset Management
  - Test sets are independent of hypothesis in the first investigation, *not independent* of hypotheses formed over subsequent investigations.
  - Users specify a budget *B* upfront, so that FairTest can earmark *B* test sets for each investigation.

Explanatory Attributes

- After seeing unfair effects, the analyst re-runs the tests, using conditional association metrics.

Report of assoc. of O=Admitted on S_i=Gender, conditioned on attribute E=Department:

Global Population of size 2,213
p-value=7.98e-01 ; COND-DIFF=[-0.0382, 0.1055]

Admitted	Female	Male	Total
No	615(68%)	680(52%)	1295 (59%)
Yes	**295(32%)**	**623(48%)**	918 (41%)
Total	910(41%)	1303(59%)	2213(100%)

* **Department A: Population of size 490:**
p-value=4.34e-03 ; DIFF=[0.0649, 0.3464]

Admitted	Female	Male	Total
No	9(15%)	161(37%)	170 (35%)
Yes	51(85%)	269(63%)	320 (65%)
Total	60(12%)	430(88%)	490(100%)

* **Department B: Population of size 279:**
p-value=1.00e+00 ; DIFF=[-0.4172, 0.3704]

Admitted	Female	Male	Total
No	3(30%)	93(35%)	96 (34%)
Yes	7(70%)	176(65%)	183 (66%)
Total	10 (4%)	269(96%)	279(100%)

* **... Departments C-F, with high p-values ...**

Figure 4. **Disparate Admission Rates in the Berkeley Dataset.** Shows a *Testing* investigation with explanatory attribute $\boldsymbol{E}$ = Department. COND-DIFF is the binary difference metric (DIFF), conditioned on $\boldsymbol{E}$.

Summary

Input: Data D = (**S**, **X**, **E**, **O**);
Output: Association bug report.

1. Split D into D_{train} and D_{test}.
2. for each protected attribute $\mathbf{S}_i$ in **S**:
  2.1 Choose an association metric M, given **O**, $\mathbf{S}_i$, **E**
  2.2. Using D_{train}, derive association contexts by building a decision tree on **X** guided by the value of the metric M between $\mathbf{S}_i$ and **O**.
  2.3. for each context:
    Using D_{test}, compute confidence interval (CI) and statistical significance (p-value) for M
3. Correct CIs, p-values for multiple testing across all protected attributes and contexts.
4. for each protected attribute $\mathbf{S}_i$ in **S**:
  4.1. Filter results on p-value.
  4.2. Rank results on CIs.
5. return association bugs for each $\mathbf{S}_i$.

Outline

- Problem Statement
- Contributions
- UA Framework
- FairTest Design
- Evaluation
- Discussion

Evaluation

The evaluation is designed at address three questions:

1. Is FairTest effective at detecting association bugs?
2. Is it fast enough to be practical?
3. Is it used to identify and debug association bugs in a variety of applications?

Detection Effectiveness (Q1)

- Microbenchmark
  - Generate ~1M synthetic users from U.S. Census data.
  - Introduce disparity to algorithm outcome.
    - Output “1” to 60% of high-income users and 40% of low-income users, for White users in CA.
    - Δ implies difference in output proportions.
  - Inject 10 randomly chosen discrimination contexts, for various subpopulation sizes.

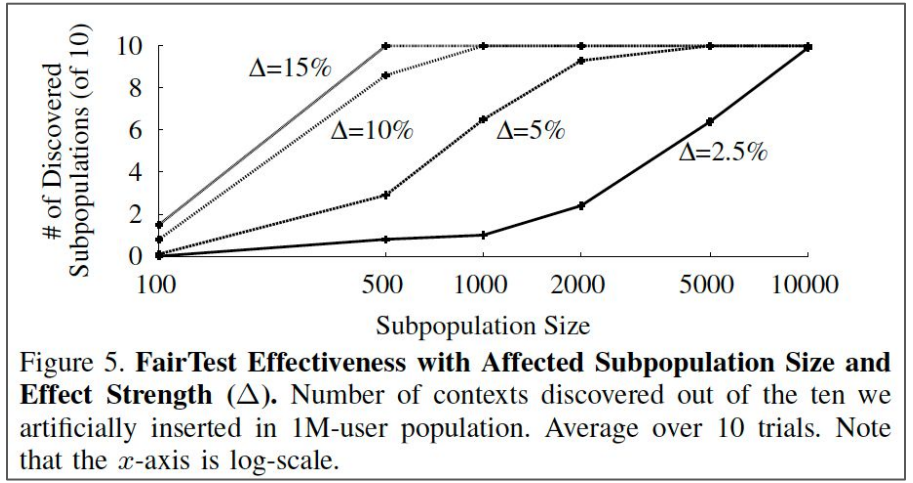


Figure 5. **FairTest Effectiveness with Affected Subpopulation Size and Effect Strength (Δ).** Number of contexts discovered out of the ten we artificially inserted in 1M-user population. Average over 10 trials. Note that the x-axis is log-scale.

Detection Effectiveness (Q1)

- Real-World Apps and Datasets
  - Staples' Pricing Scheme
  - Predictive Healthcare app
  - Image Tagger based on Caffe, trained on ImageNet
  - Movie Recommender, trained on MovieLens data
- No ground truth available for these datasets
  - Detections on discrimination contexts of different sizes appear accurate, and revelatory for an investigator

					Association Contexts			
Application	**Invest.**	**Users**	**Attr.**	**Metric(s)**	**Discovered**	**Validated**	**Reported**	**Smallest Reported**
Microbenchmark	T	988871	4	NMI	n/a	n/a	n/a	n/a
Staples Pricing	T	988871	4	NMI	224	100	21	211
Predictive Healthcare	EP	86359	128	CORR	33	33	2	91
Image Tagger	D,T	2648	1	REG,DIFF	1	1	1	1324
Movie Recommender	T	6040	3	CORR	15	10	7	511
Adult Census	T	48842	13	NMI	108	57	10	104
Berkeley Admission	T	4425	2	DIFF	1	0	1	2213

TABLE 2. **Workloads.** Investigations: *Discovery* (D), *Testing* (T), *ErrorProfiling* (EP). Metrics: normalized mutual information (NMI), correlation (CORR), binary difference (DIFF), regression (REG). For each application, we report the number of potential association contexts found by FairTest's guided-tree construction, the number that were found to be statistically significant (p-value $< 5\%$), and the number of reported bugs.

Performance (Q2)

Timing

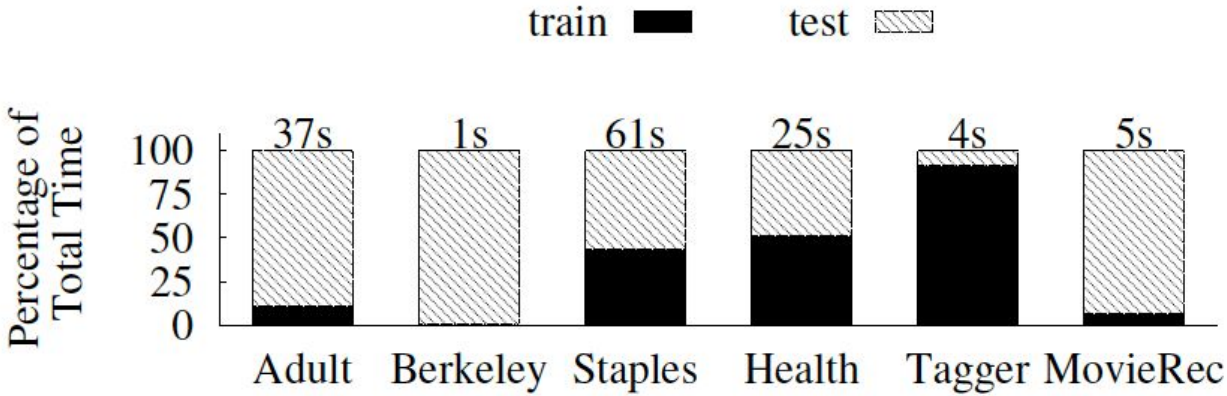


Figure 6. **FairTest Performance.** Total analysis time (labels above bars) broken down into training and testing times (bars).

Performance (Q2)

Subpopulation Discovery

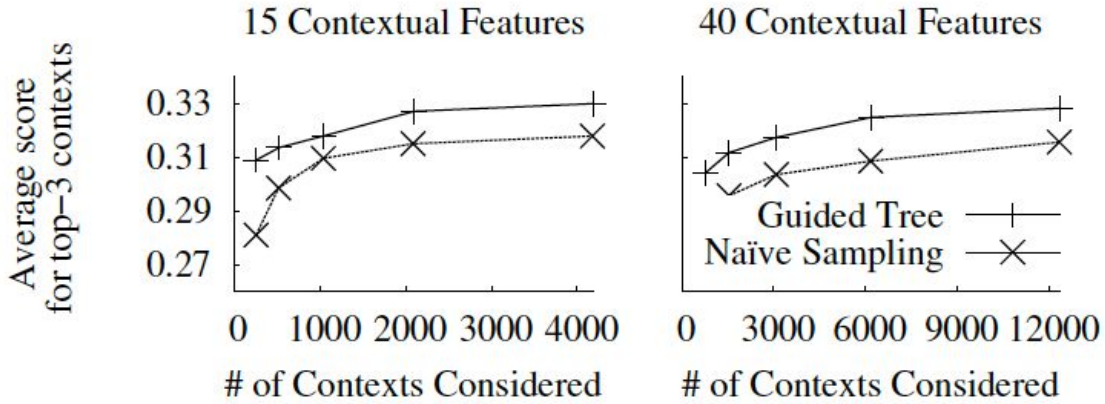


Figure 7. **Comparison between FairTest's guided tree mechanism and naïve subpopulation sampling.** Shows the average of the 3 highest associations scores uncovered in the testing phase, for an increasing number of candidates considered in the training phase.

Investigation Experience (Q3)

Predictive Healthcare

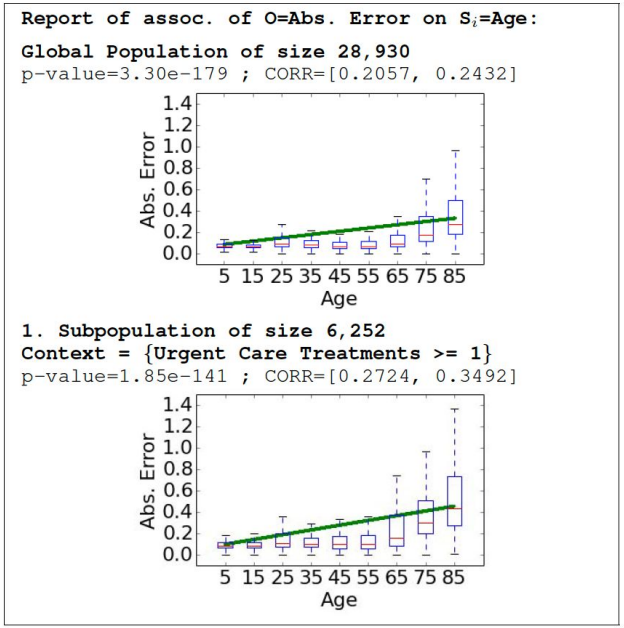


Figure 8. **Error Profile for Health Predictions**. Shows the global population and the subpopulation with highest effect (correlation). Plots display correlation between age and prediction error, for predictions of $\log(1 + \text{number of visits})$. For each age-decade, we display standard box plots (box from the 1st to 3rd quantile, line at median, whiskers at 1.5 IQRs). The straight green line depicts the best linear fit over the data.

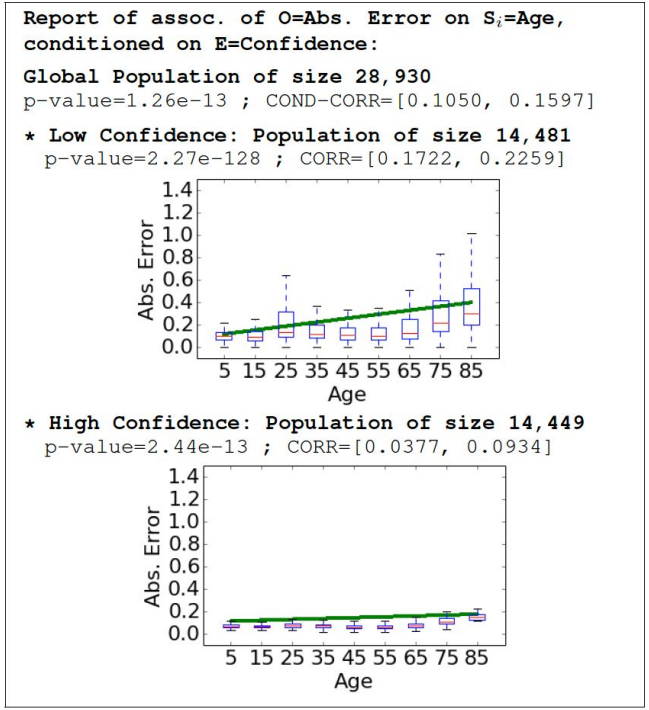


Figure 9. **Error Profile for Health Predictions using prediction confidence as an explanatory attribute**. Shows correlations between prediction error and user age, broken down by prediction confidence.

Investigation Experience (Q3)

Image Tagger

Report of associations of O=Labels on S_i=Race:

Global Population of size 1,324

*** Labels associated with Race=Black:**

Label	Black	White	DIFF	p-value
Cart	4%	0%	[0.014,0.065]	3.31e-05
Drum	4%	0%	[0.010,0.060]	3.83e-04
Helmet	8%	3%	[0.010,0.089]	2.34e-03
Cattle	**2%**	**0%**	[0.0037,0.0432]	4.73e-03

*** Labels associated with Race=White:**

Label	Black	White	DIFF	p-value
Face Powder	1%	10%	[-0.134,-0.053]	5.60e-12
Maillot	4%	15%	[-0.159,-0.058]	3.46e-10
Person	**96%**	**99%**	[-0.056,-0.004]	6.06e-03
Lipstick	1%	4%	[-0.062,-0.003]	1.03e-02

Figure 10. **Racial Label Associations in the Image Tagger.** Shows partial report of a *Discovery* (top_k=35); the four most strongly associated labels (for the binary difference metric DIFF) are shown for each race.

Discussion

- Use of regression to find bugs with large output space.
- Defining fairness as a measure of utility for the user subgroup.
  - Error Rate
- Extensive Testing Suite
  - Open sourced code
- Investigate Explanatory Factors for association bugs
- Limited to statistical tests
- Remediation: Batteries not included

Thanks!

Questions?