FairTest:
Discovering unwarranted associations in data-driven applications

IEEE EuroS&P
April 28th, 2017

Florian Tramèr\textsuperscript{1}, Vaggelis Atlidakis\textsuperscript{2}, Roxana Geambasu\textsuperscript{2}, Daniel Hsu\textsuperscript{2}, Jean-Pierre Hubaux\textsuperscript{3}, Mathias Humbert\textsuperscript{4}, Ari Juels\textsuperscript{5}, Huang Lin\textsuperscript{3}

\textsuperscript{1}Stanford University, \textsuperscript{2}Columbia University, \textsuperscript{3}École Polytechnique Fédérale de Lausanne, \textsuperscript{4}Saarland University, \textsuperscript{5}Cornell Tech
Websites Vary Prices, Deals Based on Users’ Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

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.

In what appears to be an unintended side effect of Staples’ pricing methods—likely a function of retail competition with its rivals—the Journal’s testing also showed that areas that tended to see the discounted prices had a higher average income than areas that tended to see higher prices.
Google Photos labeled black people 'gorillas'

Jessica Guynn, USA TODAY 2:10 p.m. EDT July 1, 2015

SAN FRANCISCO — Google has apologized after its new Photos application identified black people as "gorillas."

On Sunday Brooklyn programmer Jacky Alciné tweeted a screenshot of photos he had uploaded in which the app had labeled Alcine and a friend, both African American, "gorillas."

Yontan Zunger, an engineer and the company's chief architect of Google+, responded swiftly to Alciné on Twitter: "This is 100% Not OK." And he promised that Google's Photos team was working on a fix.

These are **software bugs**: need to *actively test for them* and *fix them (i.e., debug)* in data-driven applications... *just as with functionality, performance, and reliability bugs.*
Unwarranted Associations Model

User inputs → Data-driven application → Application outputs

Protected inputs
Limits of preventative measures

What doesn’t work:

- **Hide protected attributes** from data-driven application.
- Aim for **statistical parity** w.r.t. protected classes and service output.

Foremost challenge is to even detect these unwarranted associations.
A Framework for Unwarranted Associations

1. Specify **relevant data features**:
   - Protected variables (e.g., Gender, Race, …)
   - “Utility”: a function of the algorithm’s output (e.g., Price, Error rate, …)
   - Explanatory variables (e.g., Qualifications)
   - Contextual variables (e.g., Location, Job, …)

2. Find **statistically significant associations** between protected attributes and utility
   - *Condition on* explanatory variables
   - Not tied to any particular *statistical metric* (e.g., odds ratio)

3. Granular search in **semantically meaningful subpopulations**
   - Efficiently list *subgroups* with highest adverse effects
• Finds context-specific associations between protected variables and application outputs, conditioned on explanatory variables

• Bug report ranks findings by assoc. strength and affected pop. size
A data-driven approach

Core of FairTest is based on statistical machine learning

Find context-specific associations

Statistically validate associations

Statistical machine learning internals:
- top-down spatial partitioning algorithm
- confidence intervals for assoc. metrics
- ...

Data

Ideally sampled from relevant user population

Report of associations of $O=Price \text{ on } S=Income$:
Assoc. metric: norm. mutual information (NMI).
Global Population of size 494,436
p-value $= 3.34 \times 10^{-10}$; NMI $= (0.0001, 0.0005)$

<table>
<thead>
<tr>
<th></th>
<th>Income &lt;$50K</th>
<th>Income $\geq$50K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>2916 (6%)</td>
<td>13467 (6%)</td>
<td>16383 (59%)</td>
</tr>
<tr>
<td>Low</td>
<td>234167 (94%)</td>
<td>231303 (94%)</td>
<td>465468 (94%)</td>
</tr>
<tr>
<td>Total</td>
<td>249468 (50%)</td>
<td>244968 (50%)</td>
<td>494436 (100%)</td>
</tr>
</tbody>
</table>

1. Subpopulation of size 23,532
Context: [State: CA, Race: White]
p-value $= 2.31 \times 10^{-4}$; NMI $= (0.0031, 0.0039)$

<table>
<thead>
<tr>
<th></th>
<th>Income &lt;$50K</th>
<th>Income $\geq$50K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>606 (8%)</td>
<td>691 (4%)</td>
<td>1297 (6%)</td>
</tr>
<tr>
<td>Low</td>
<td>7116 (92%)</td>
<td>13119 (96%)</td>
<td>22235 (94%)</td>
</tr>
<tr>
<td>Total</td>
<td>7722 (13%)</td>
<td>15810 (100%)</td>
<td>23532 (100%)</td>
</tr>
</tbody>
</table>

2. Subpopulation of size 2,198
Context: [State: NY, Race: Black, Gender: Male]
p-value $= 7.73 \times 10^{-5}$; NMI $= (0.0040, 0.0075)$

<table>
<thead>
<tr>
<th></th>
<th>Income &lt;$50K</th>
<th>Income $\geq$50K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>52 (4%)</td>
<td>8 (1%)</td>
<td>60 (3%)</td>
</tr>
<tr>
<td>Low</td>
<td>120 (96%)</td>
<td>937 (99%)</td>
<td>2133 (97%)</td>
</tr>
<tr>
<td>Total</td>
<td>1253 (57%)</td>
<td>945 (43%)</td>
<td>2198 (100%)</td>
</tr>
</tbody>
</table>
• **Example:** simulation of location based pricing scheme

• Test for **disparate impact on low-income populations**
  
  • Low effect over whole US population
  
  • High effects in specific sub-populations
Association-Guided Decision Trees

Goal: find most strongly affected user sub-populations

Split into sub-populations with increasingly strong associations between protected variables and application outputs.
Association-Guided Decision Trees

- Efficient discovery of contexts with high associations
- Outperforms previous approaches based on frequent itemset mining
- Easily interpretable contexts by default
- Association-metric agnostic

<table>
<thead>
<tr>
<th>Metric</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary ratio/difference</td>
<td>Binary variables</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>Categorical variables</td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>Scalar variables</td>
</tr>
<tr>
<td>Regression</td>
<td>High dimensional outputs</td>
</tr>
<tr>
<td><strong>Plugin your own!</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Greedy strategy (some bugs could be missed)
Example: healthcare application

Predictor of whether patient will visit hospital again in next year (from winner of 2012 Heritage Health Prize Competition)

**FairTest findings:** strong association between age and prediction error rate

Association may translate to quantifiable harms (e.g., if model is used to adjust insurance premiums)
Debugging with FairTest

Are there **confounding factors**?
Do associations disappear **after conditioning**?
⇒ **Adaptive Data Analysis!**

Example: the healthcare application (again)
- Estimate **prediction confidence** (target variance)
- Does this **explain** the predictor’s behavior?
- Yes, partially

FairTest helps developers understand & evaluate potential association bugs.
Other applications studied using FairTest

• Image tagger based on ImageNet data
  ⇒ Large output space (~1000 labels)
  ⇒ FairTest automatically switches to regression metrics
  ⇒ Tagger has *higher error rate* for pictures of black people

• Simple movie recommender system
  ⇒ Men are assigned movies with *lower ratings* than women
  ⇒ Use personal preferences as *explanatory factor*
  ⇒ FairTest finds no significant bias anymore
The *Unwarranted Associations* Framework

- Captures a broader set of algorithmic biases than in prior work
- Principled approach for statistically valid investigations

**FairTest**

- The first end-to-end system for evaluating algorithmic fairness

---

Developers need better statistical training and tools to make better statistical decisions and applications.


Thanks!