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#

#École Polytechnique Fédérale de Lausanne, *Columbia University, @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.
"Unfair" associations + consequences

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.
“Unfair” associations + consequences

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.*
Limits of preventative measures

What doesn’t work:
Limits of preventative measures

What doesn’t work:

• Hide protected attributes from data-driven application.
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.
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.
FairTest: a testing suite for data-driven apps

![Diagram showing User inputs to Data-driven application to Application outputs]
FairTest: a testing suite for data-driven apps
FairTest: a testing suite for data-driven apps
FairTest: a testing suite for data-driven apps

- Finds context-specific associations between protected variables and application outputs

![Diagram showing the flow of data from user inputs to protected variables, through a data-driven application, and finally to application outputs and association bug report]

User inputs ➔ Data-driven application ➔ Application outputs

- Protected vars.
- Context vars.
- Explanatory vars.

Association bug report for developer

location, click, ...

prices, tags, ...
FairTest: a testing suite for data-driven apps

- Finds **context-specific associations** between **protected variables** and **application outputs**
FairTest: a testing suite for data-driven apps

• Finds context-specific associations between protected variables and application outputs

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

Core of FairTest is based on statistical machine learning
A data-driven approach

Core of FairTest is based on statistical machine learning

Data

Ideally sampled from relevant user population
A data-driven approach

Core of FairTest is based on statistical machine learning

Ideally sampled from relevant user population
A data-driven approach

Core of FairTest is based on statistical machine learning

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

Find context-specific associations
Statistically validate associations

Training data
Test data

Ideally sampled from relevant user population
A data-driven approach

Core of FairTest is based on statistical machine learning

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

Report of associations of O=Income on Si=Race:
Global Population of size 24,421
p-value = 1.39e-53 ; NMI = [0.0063, 0.0139]

+----------------+--------+---------+----------+-----------+
| Income | Asian | Black | White   | Total     |
+----------------+--------+---------+----------+-----------+
| <=50K     | 556(73%)| 2061(88%)| 15647(75%)| 18640 (76%)|
| >50K      | 206(27%)| 287(12%) | 5238(25%) | 5781 (24%) |
| Total     | 762 (3%) | 2348(10%)| 20885(86%)| 24421(100%)|
+----------------+--------+---------+----------+-----------+

1. Subpopulation of size 341
Context = {Age <= 42, Job: Fed-gov, Hours <= 55}
p-value = 3.24e-03 ; NMI = [0.0085, 0.1310]

+----------------+--------+---------+----------+---------+
| Income | Asian | Black | White | Total |
+----------------+--------+---------+----------+---------+
| <=50K     | 10(71%) | 62(91%) | 153(63%) | 239 (70%) |
| >50K      | 4(29%)  | 6(9%)   | 91(37%)  | 102 (30%) |
| Total     | 14(4%)  | 68(20%) | 244(72%) | 341(100%) |
+----------------+--------+---------+----------+---------+

2. Subpopulation of size 14,477
Context = {Age <= 42, Hours <= 55}
p-value = 7.50e-31 ; NMI = [0.0070, 0.0187]

+----------------+--------+---------+----------+-----------+
| Income | Asian | Black | White | Total     |
+----------------+--------+---------+----------+-----------+
| <=50K     | 362(79%)| 1408(93%)| 10113(83%)| 12157 (84%)|
| >50K      | 97(21%) | 101 (7%) | 2098(17%) | 2320 (16%) |
| Total     | 459(3%) | 1509(10%)| 12213(84%)| 14477(100%)|
+----------------+--------+---------+----------+-----------+
Example: health care application

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

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

**FairTest’s finding:** significant contexts exhibiting strong association between age and prediction error rate.
Example: health care application

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

**FairTest’s finding:** significant contexts exhibiting strong association between age and prediction error rate.

- age, gender,
- # emergencies, ...

Hospital re-admission predictor

Will patient be re-admitted?

Association may translate to quantifiable harms (e.g., if app is used to adjust insurance premiums)!
Example: Berkeley graduate admissions

Admission into UC Berkeley graduate programs (Bickel, Hammel, and O’Connell, 1975)
Example: Berkeley graduate admissions

Admission into UC Berkeley graduate programs
(Bickel, Hammel, and O’Connell, 1975)

Bickel et al’s (and also FairTest’s) findings: gender bias in admissions at university level, but mostly gone after conditioning on department

age, gender, GPA, ...

Graduate admissions committees

Admit applicant?
Example: Berkeley graduate admissions

Admission into UC Berkeley graduate programs
(Bickel, Hammel, and O’Connell, 1975)

**Bickel et al’s (and also FairTest’s) findings**: gender bias in admissions at university level, but mostly gone after conditioning on department

---

FairTest helps developers understand & evaluate potential association bugs.
Closing remarks

• **Other applications studied using FairTest** ([http://arxiv.org/abs/1510.02377](http://arxiv.org/abs/1510.02377)):
  • Image tagger based on deep learning (on ImageNet data)
  • Simple movie recommender system (on MovieLens data)
  • Simulation of Staple’s pricing system
Closing remarks

• **Other applications studied using FairTest** ([http://arxiv.org/abs/1510.02377](http://arxiv.org/abs/1510.02377)):
  - Image tagger based on deep learning (on ImageNet data)
  - Simple movie recommender system (on MovieLens data)
  - Simulation of Staple’s pricing system

• **Other features in FairTest:**
  - Exploratory studies (e.g., find image tags with offensive associations)
  - Adaptive data analysis (preliminary) – i.e., statistical validity with data re-use
  - Integration with SciPy library
Closing remarks

- **Other applications studied using FairTest** ([http://arxiv.org/abs/1510.02377](http://arxiv.org/abs/1510.02377)):
  - Image tagger based on deep learning (on ImageNet data)
  - Simple movie recommender system (on MovieLens data)
  - Simulation of Staple’s pricing system

- **Other features in FairTest**:
  - Exploratory studies (e.g., find image tags with offensive associations)
  - Adaptive data analysis (preliminary) – i.e., statistical validity with data re-use
  - Integration with SciPy library

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

• Other applications studied using FairTest ([http://arxiv.org/abs/1510.02377](http://arxiv.org/abs/1510.02377)):
  • Image tagger based on deep learning (on ImageNet data)
  • Simple movie recommender system (on MovieLens data)
  • Simulation of Staple’s pricing system

• Other features in FairTest:
  • Exploratory studies (e.g., find image tags with offensive associations)
  • Adaptive data analysis (preliminary) – i.e., statistical validity with data re-use
  • Integration with SciPy library

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