Detecting Bias in Machine Learning Algorithms

Giulia Fanti

Based on slides by Aniko Hannak

Fall 2019

Administrative

- HW3 due today at 11:59 pm ET

- Mid-semester presentations
  - Wednesday, Oct. 30
  - Monday, Nov. 4

- This Friday (Oct. 25): Day for Community Engagement
  - No recitation
  - Sruti will hold her OH on Friday from 3-4pm ET
  - My OH will be by appointment this week

Mid-Semester Presentations

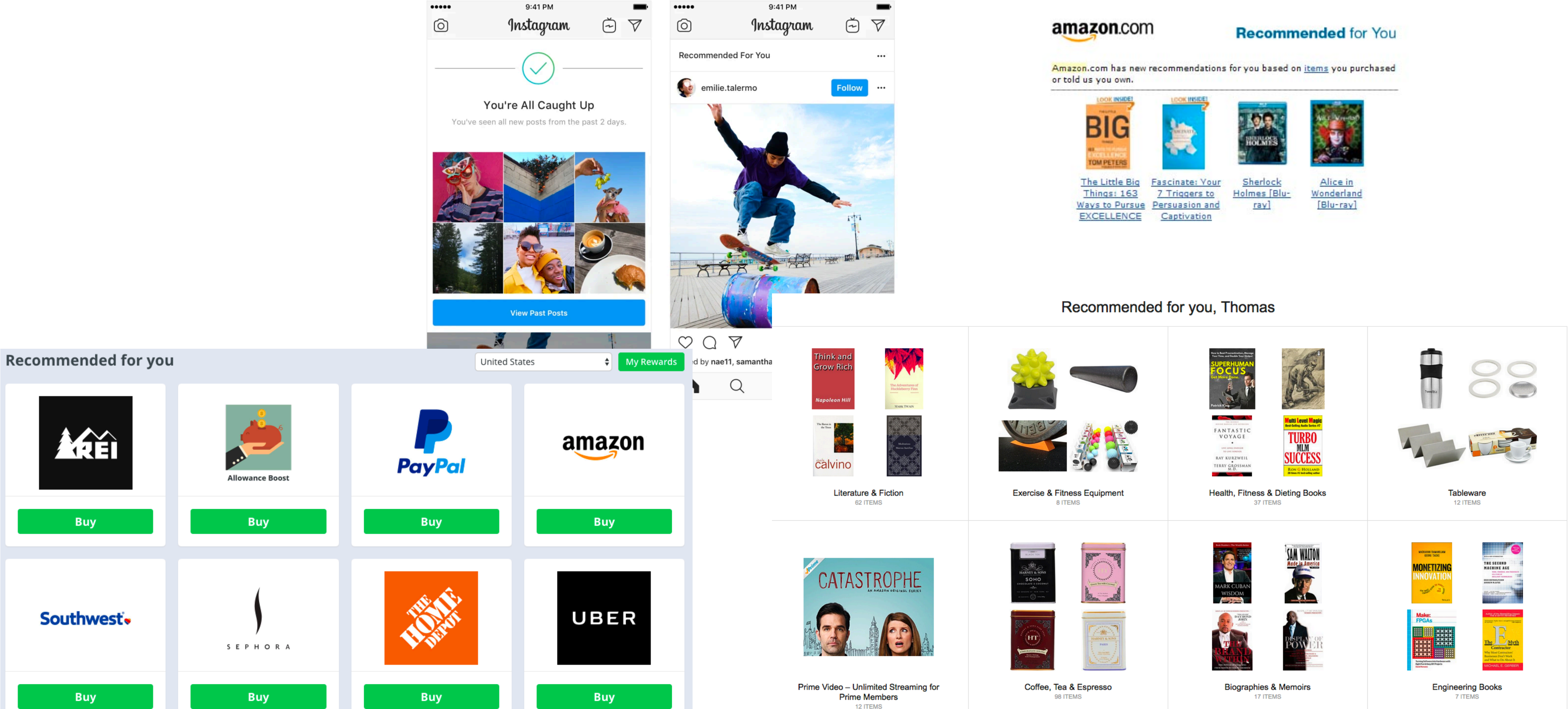
- Wednesday, Oct. 30 and Monday, Nov. 4
- Sign up for a slot here:
- Each presentation should be no more than 10 minutes (TIME YOURSELVES)
  - We WILL cut you off at 10 min; if you haven't covered all parts of the rubric, they will be counted as not present

- Rubric
  - Introduction and motivation
  - Background
  - Design/concept summary (i.e., what is the central idea underlying your project?)
  - Evaluation
    - Experiments if you are doing a practical project
      - Datasets
      - Plots
      - Interpretation
  - What you plan to do before the final report deadline?

In-class quiz

- Note on cheating
- On Canvas

Personalization on the Web



New Focus: Bias in ML models

- How does it manifest itself?
- How do we detect and measure it?

ProPublica

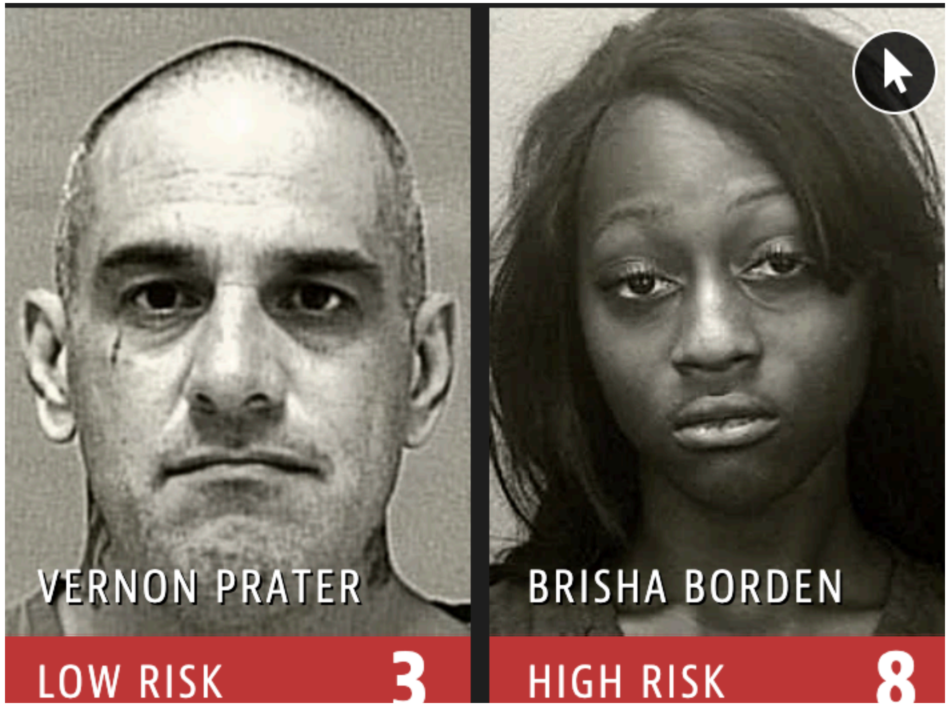
- 2015 Article by Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner
- Investigation of racial bias in software used in the criminal system

Two arrests

Stole $80 of tools from Home Depot.

Prior Offenses:
2 armed robberies
1 attempted armed robbery

Subsequent Offenses:
1 grand theft



Took a child's bicycle and scooter to go pick up her god-sister from school.

Prior offenses:
4 Juvenile misdimeanors

Subsequent offenses:
None

Risk Scores

- Increasingly common in courtrooms
  - E.g., Northpointe software
- Used to inform who can be set free
  - Results are sometimes given to judges during sentencing
- Sentencing Reform and Corrections Act (2015)
  - Mandates use of these tools in federal prisons
- Eric Holder (US Attorney General, 2014) warned about bias in these tools
  - No formal action was taken

ProPublica Investigation of Risk Scores

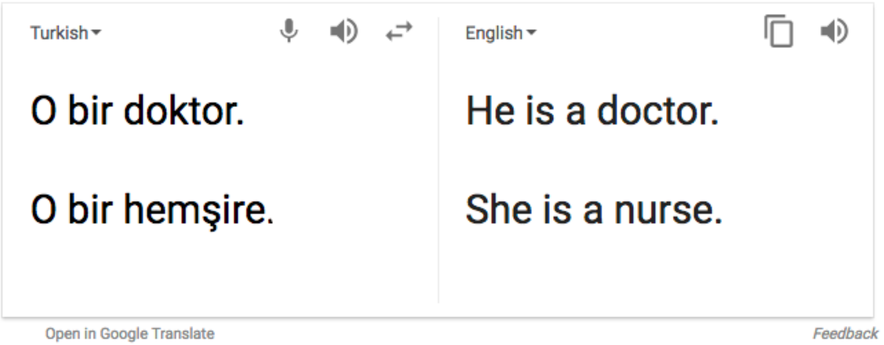
- Obtained risk scores for 7,000+ arrests in Broward County, FL
- Checked recidivism rate over next 2 years
- Result:
  - 20% of people predicted to commit violent crimes did so
  - 61% of people predicted to commit any crime did so

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

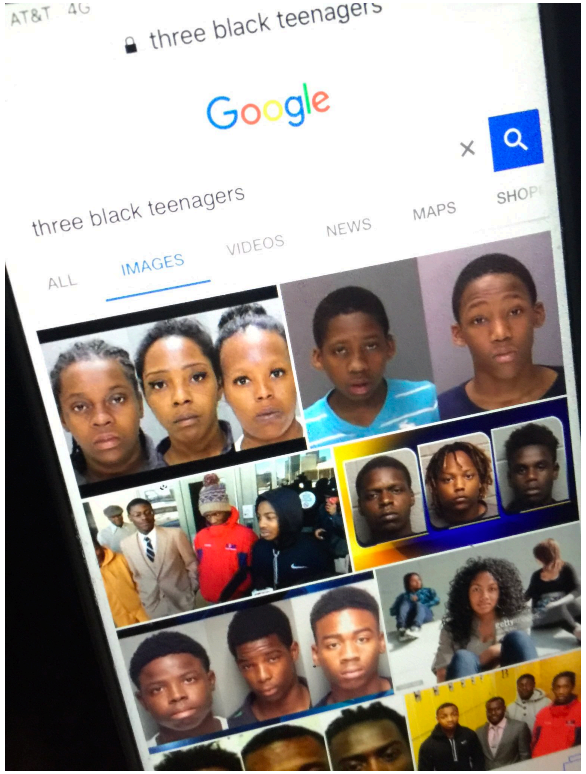
Why is this happening?

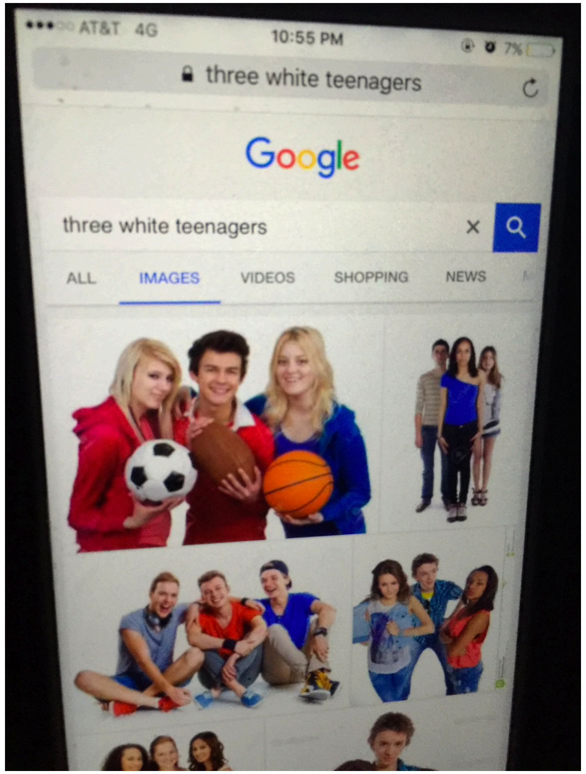
- Could it be statistical?
  - E.g. black defendants happen to recidivate at a higher rate?
  - No
  - Controlled for these effects, found software:
    - 77% more likely to classify black defendants as likely to commit future violent crimes
    - 45% more likely to classify black defendants as likely to commit any future crimes
- Northpointe's response
  - Disputes analysis
  - Scores are derived from 137 questions, none of them race
    - "Was one of your parents ever sent to jail or prison?"
    - "How many of your friends are taking drugs illegally?"
- How can this be?

Inadvertent bias is often present in ML models



It's all in the data!





How do we fix this?

How artificial intelligence learns to be racist

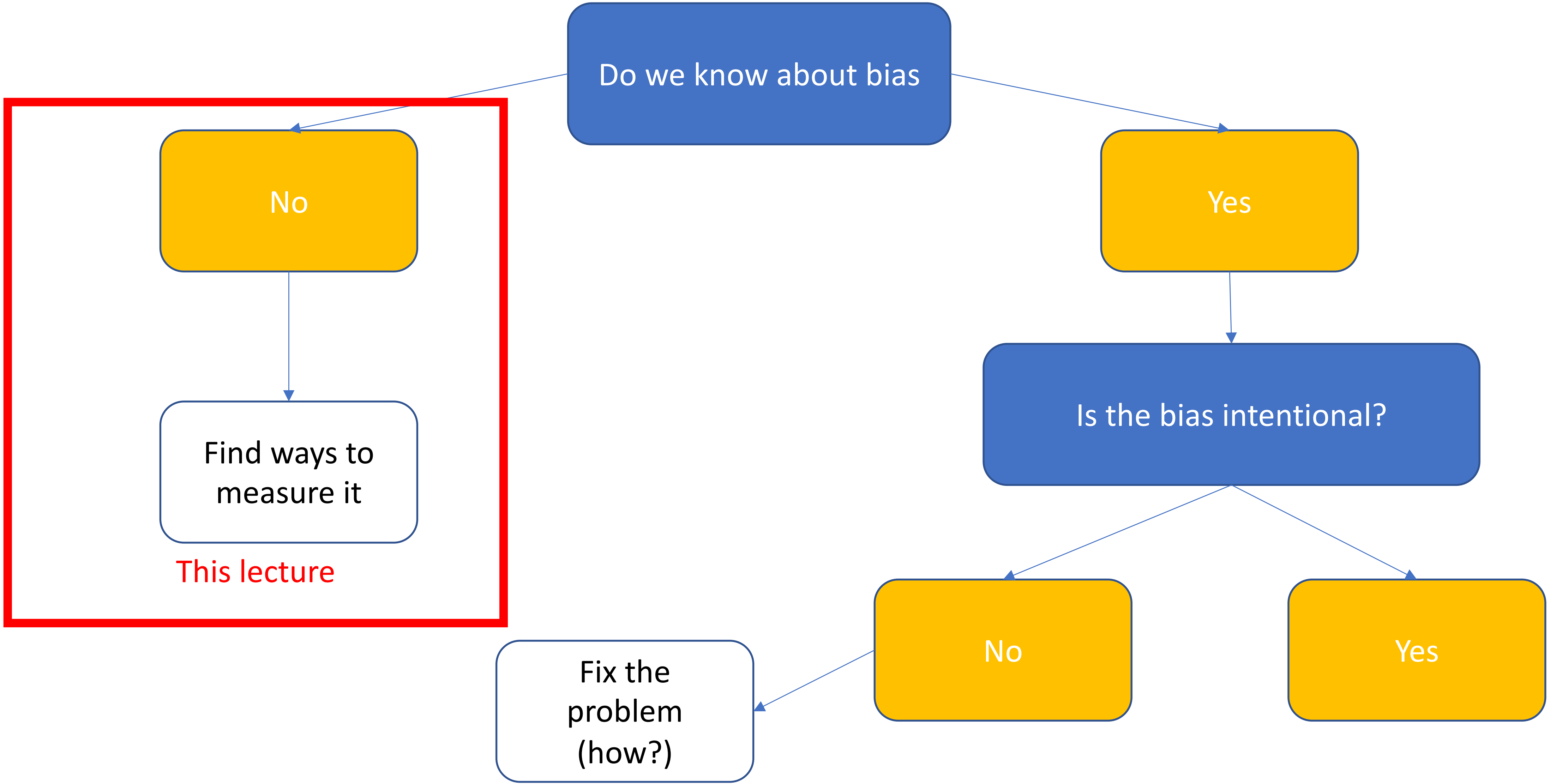
Simple: It's mimicking us.

By Brian Resnick | @B_resnick | brian@vox.com | Apr 17, 2017, 2:10pm EDT

Training AI robots to act 'human' makes them sexist and racist

By Mike Wehner, BGR April 17, 2017 | 1:00pm | Updated

How do we fix this?



Measuring Price Discrimination and Steering on E-commerce Web Sites

Aniko Hannak
Northeastern University
Boston, MA
ancsaaa@ccs.neu.edu

Gary Soeller
Northeastern University
Boston, MA
soelgary@ccs.neu.edu

David Lazer
Northeastern University
Boston, MA
d.lazer@neu.edu

Alan Mislove
Northeastern University
Boston, MA
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
Boston, MA
cbw@ccs.neu.edu

Peeking Beneath the Hood of Uber

Le Chen
Northeastern University
Boston, MA
leonchen@ccs.neu.edu

Alan Mislove
Northeastern University
Boston, MA
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
Boston, MA
cbw@ccs.neu.edu

E-commerce sites

- Online purchasing now extremely common
- Significant, comprehensive user tracking
  - Clear economic incentive to use data to increase sales
- These processes are hidden from users
  - What personal data is collected?
  - How is it used? Possibly to users' disadvantage

- Examine two trends: Price discrimination and steering

Price Discrimination

- Showing different users different prices
- Ex. Amazon in 2004
  - DVDs were sold for $3-4 more to some users
- Not illegal!
  - Anti-discrimination act doesn’t protect consumers

Price Steering

- Altering the rank order of products
  - E.g., high-priced items ranked higher for some people

- Ex: Orbitz in 2012
  - Users received hotels in different order when searching
  - Normal users: cheap hotels first
  - Mac users: expensive hotels first

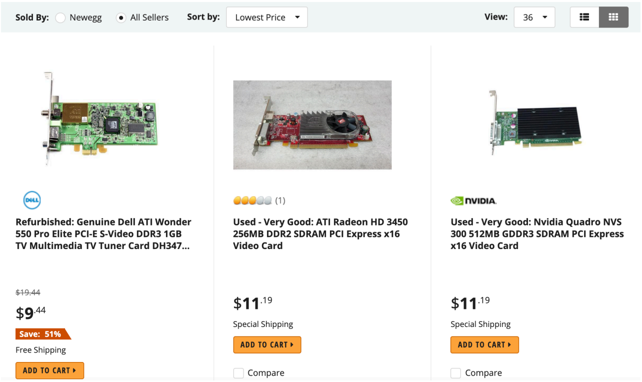
Goals of this paper

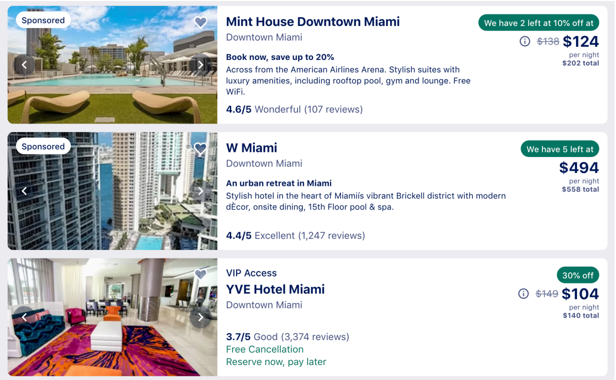
- Methodology to measure personalization of e-commerce
- Measure personalization on e-commerce sites
  - Price Discrimination
    - Are the same products offered at different prices to people?
  - Price Steering
    - Are products presented in a different order?
    - Do some people see more expensive products?
- Explore how online retailers personalize
  - What features do their algorithms personalize on?

Measurements

- 10 general retailers
  - BestBuy CDW HomeDepot JCPenney Macy's NewEgg OfficeDepot Sears Staples Walmart
- 6 travel sites
  - CheapTickets, Expedia, Hotels, Priceline, Orbitz, Travelocity

- Focus on products retuned by searches, 20 search terms / site





Are all differences personalization?

- No! Could be due to
  - Updates to inventory/prices
  - Tax/Shipping differences
  - Distributed infrastructure
  - Load-balancing

Only interested in personalization due to client-side state associated with request

How do we measure personalization?

Measuring personalization

129.10.115.14

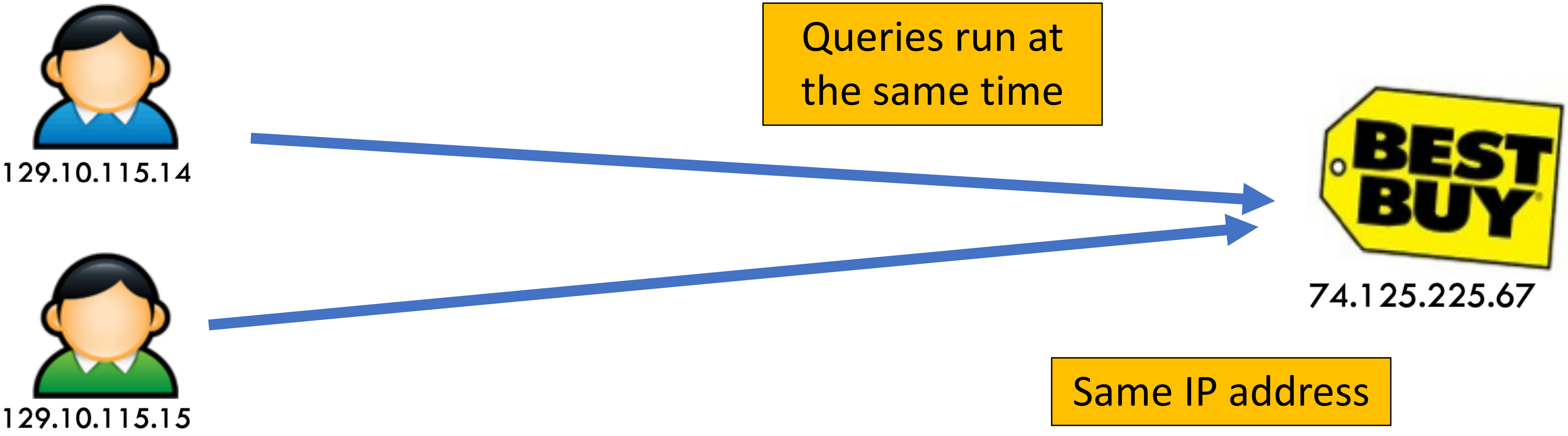
IP addresses in the same /24

129.10.115.15

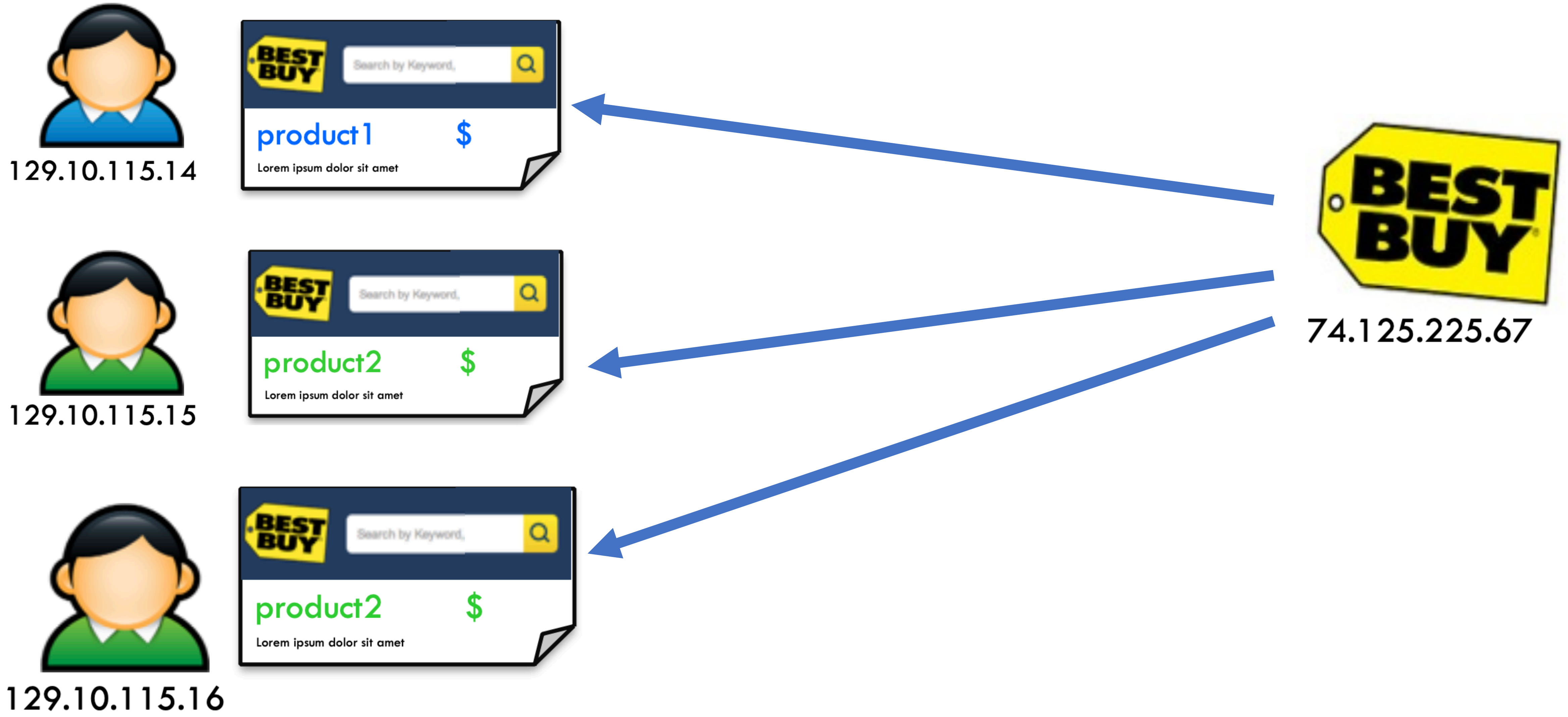
BEST BUY

74.125.225.67

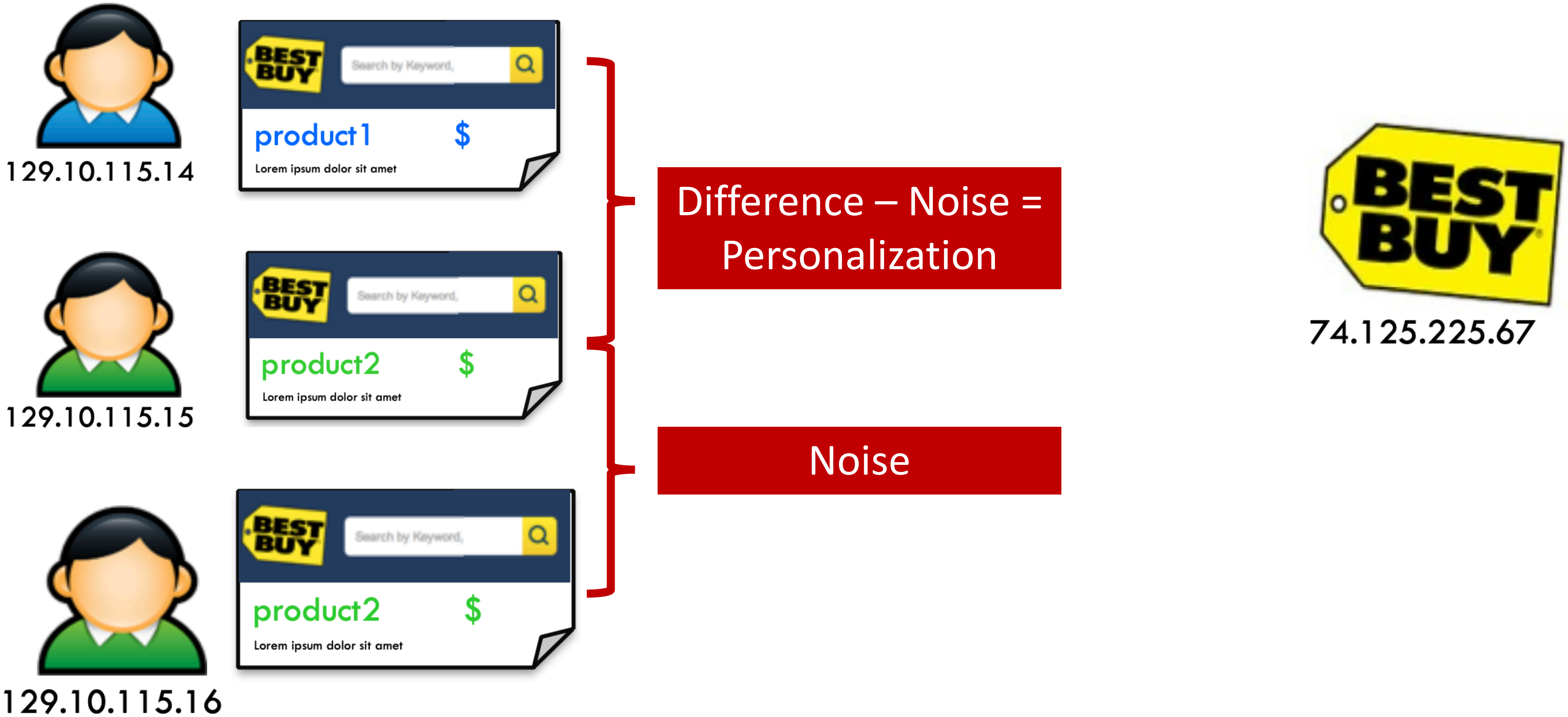
Measuring personalization



Measuring personalization



Measuring personalization



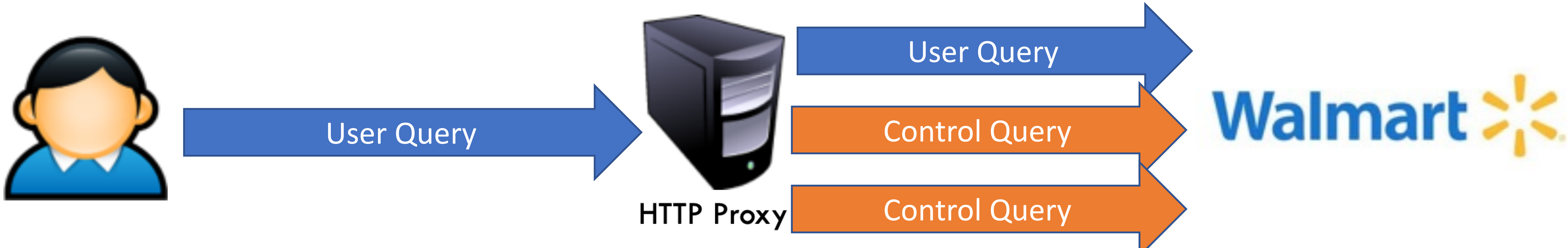
Measuring Price Discrimination

- Real user accounts
- Synthetic user accounts
- Key questions:
  - To what extent are products personalized?
  - What user features drive personalization?

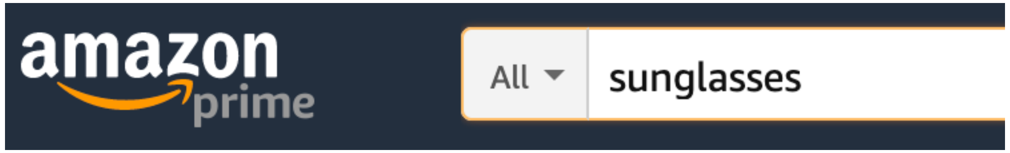
Real User Data	Synthetic Data
Make real-life measurements	Control account characteristics
Only valid for users with historic data	Measure the impact of specific features

Personalization for Real Users

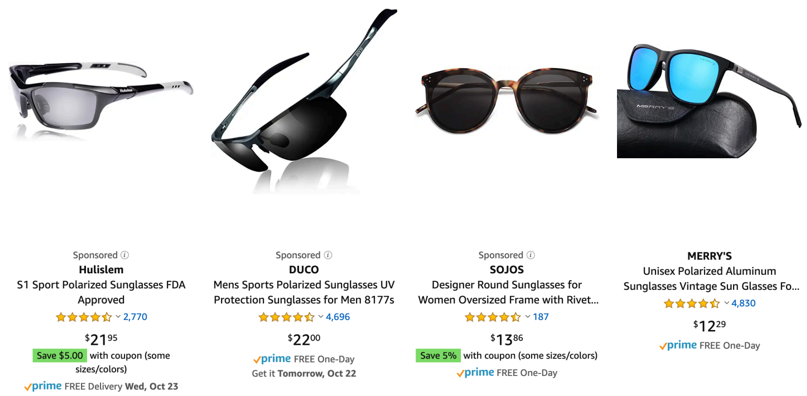
- Gather data from Mechanical Turk
- 300 users
  - 100 users for each category: e-commerce, hotels, rental cars
- 20 searches for each site
- Use web server + proxy to launch, intercept searches



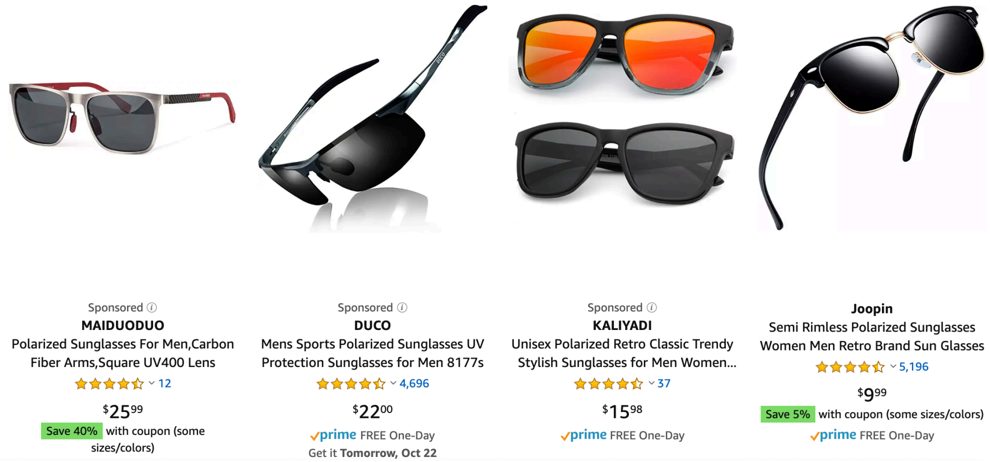
Comparing Results: Jaccard Similarity



Control Results



User results



$$\mathrm{Jacc}(\mathrm{ctrl}, \mathrm{usr}) = \frac{|\mathrm{ctrl} \cap \mathrm{usr}|}{|\mathrm{ctrl} \cup \mathrm{usr}|}$$

In this example $= \frac{1}{7}$

Comparing Ordering: Kendall's Tau

Control Results

1) A

2) B

3) C

User results

1) B

2) C

3) A

$$\tau(\mathrm{ctrl}, \mathrm{usr}) = \frac{(\#\text{concordant pairs}) - (\#\text{discordant pairs})}{\binom{n}{2}}$$

$n = 3$ items

$\binom{n}{2} = 3$ pairs

(A, B)=dis

(B, C)=con

(A, C)=dis

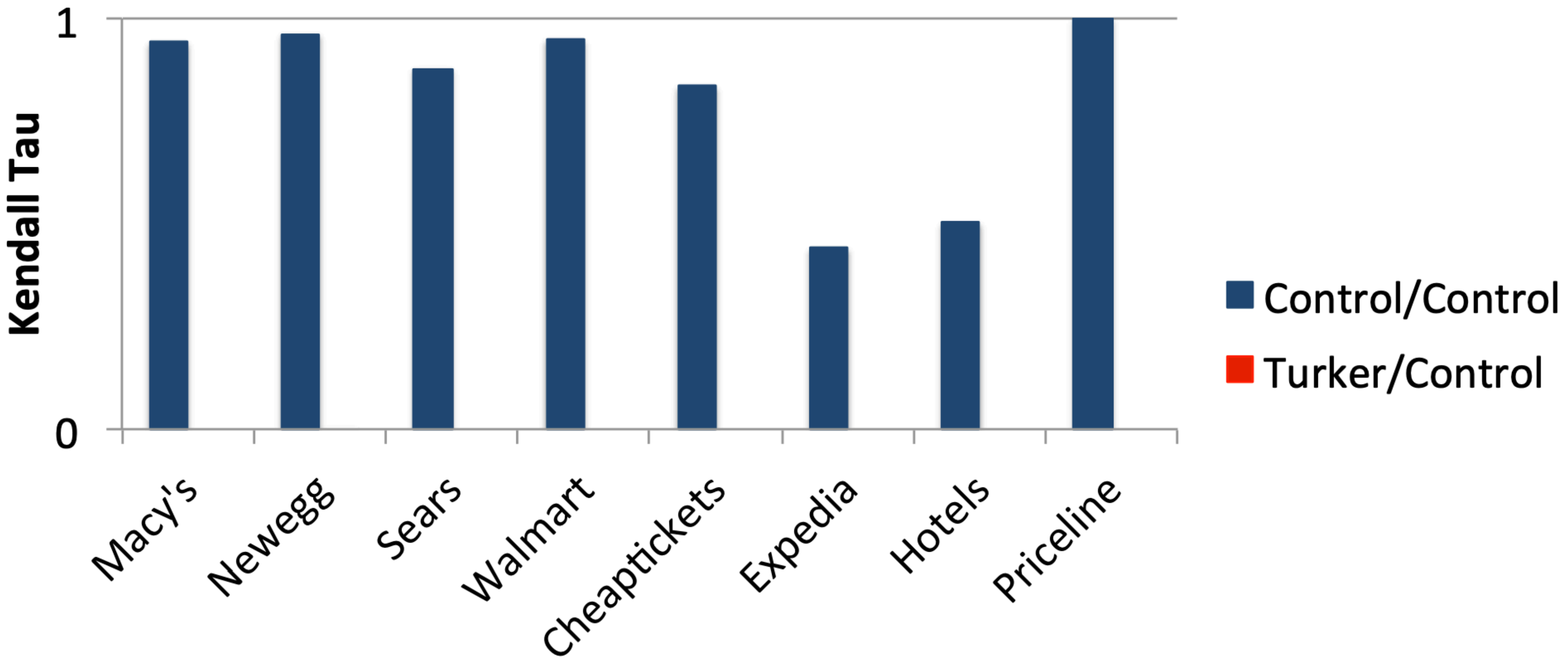
<u>Concordant</u> <u>Discordant</u>

I II

$$\tau(\mathrm{ctrl}, \mathrm{usr}) = \frac{1-2}{3} = -\frac{1}{3}$$

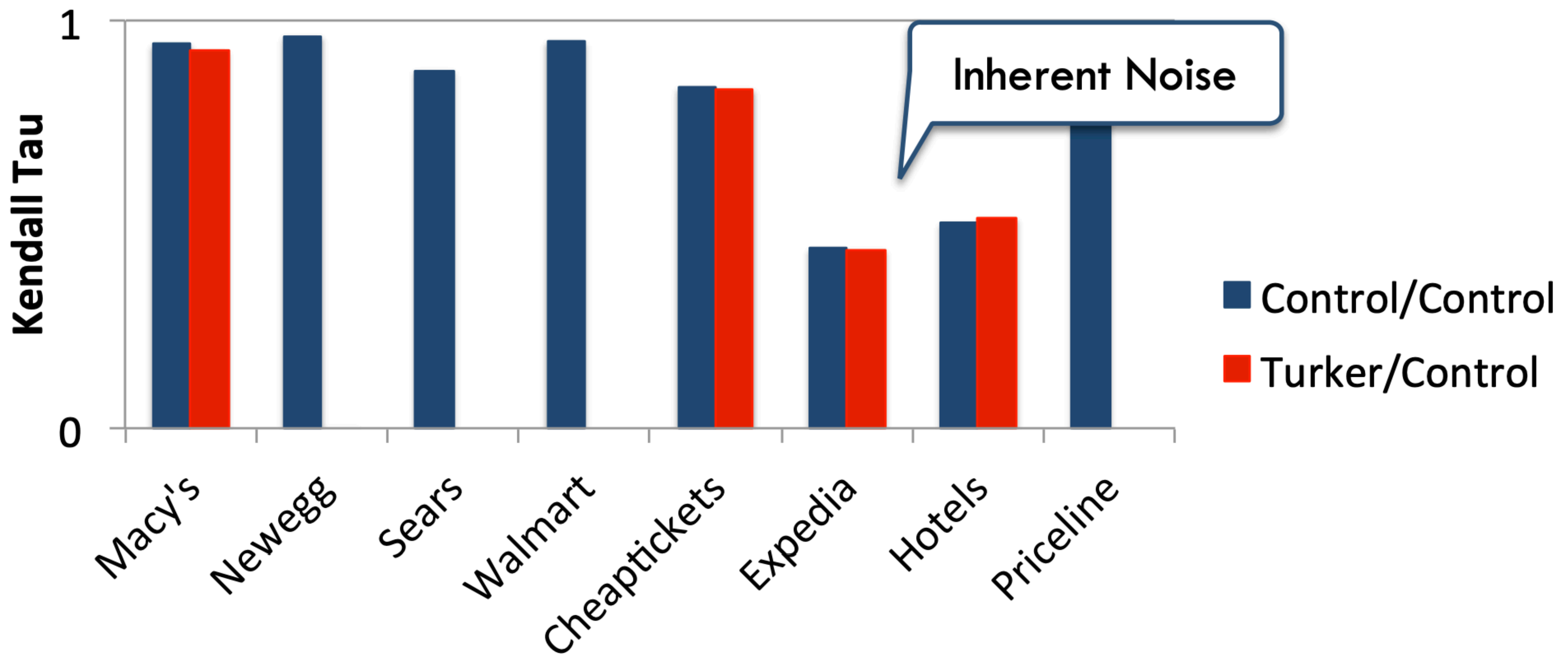
Price Steering for Real Users

- Are products presented in the same order?
  - Kendall's Tau Correlation



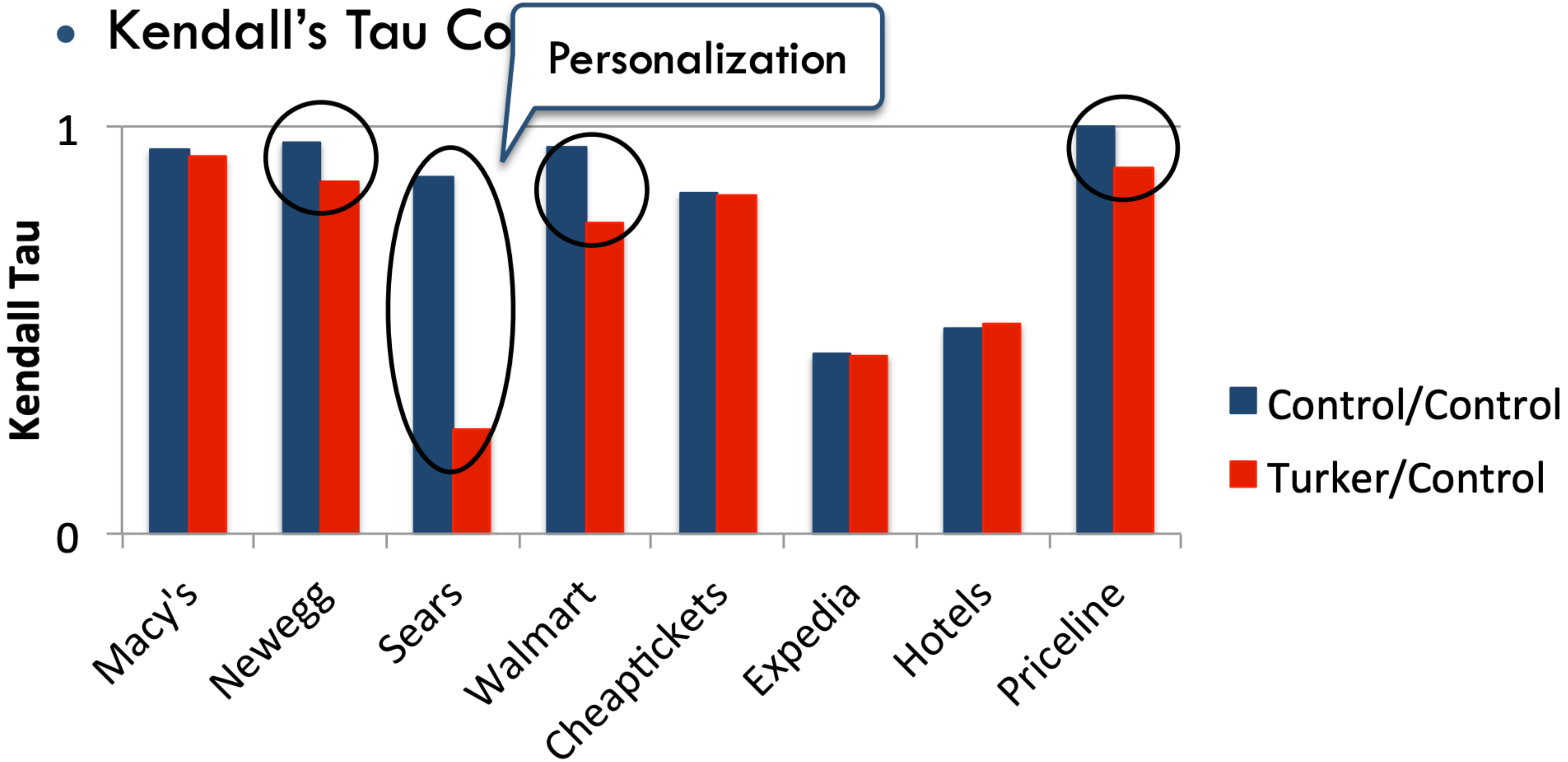
Price Steering for Real Users

- Are products presented in the same order?
  - Kendall's Tau Correlation



Price Steering for Real Users

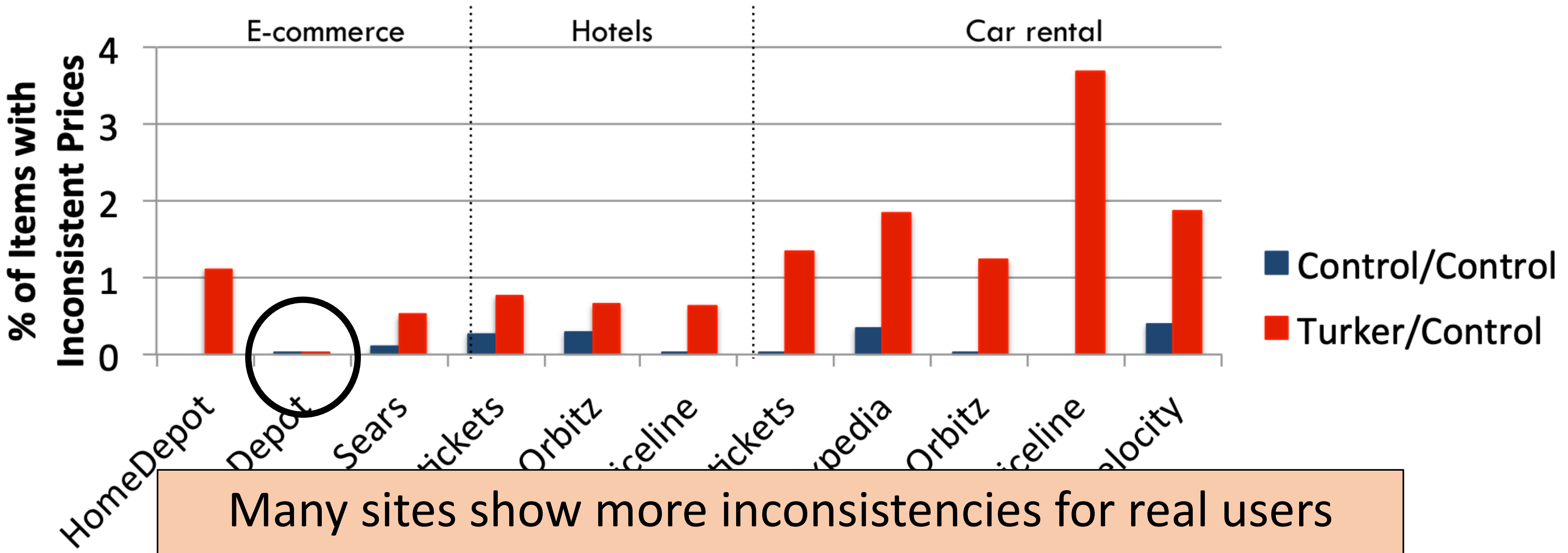
- Are products presented in the same order?
  - Kendall's Tau Co



Price Discrimination for real users

- Do users see the same prices for the same products?

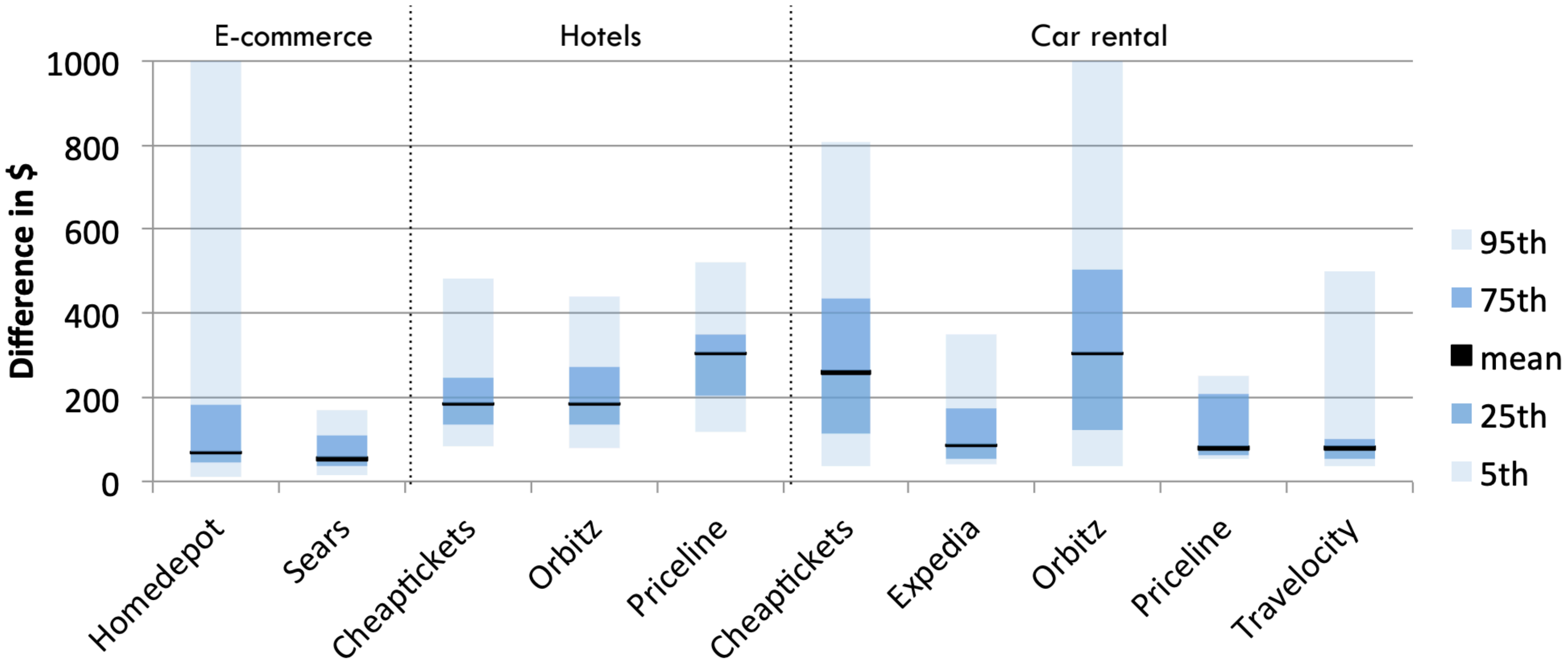
Percentage of products with inconsistent pricing



Many sites show more inconsistencies for real users
Up to 3.6% of all products!

Price Discrimination for real users

- How much money are we talking about?



Take-Aways

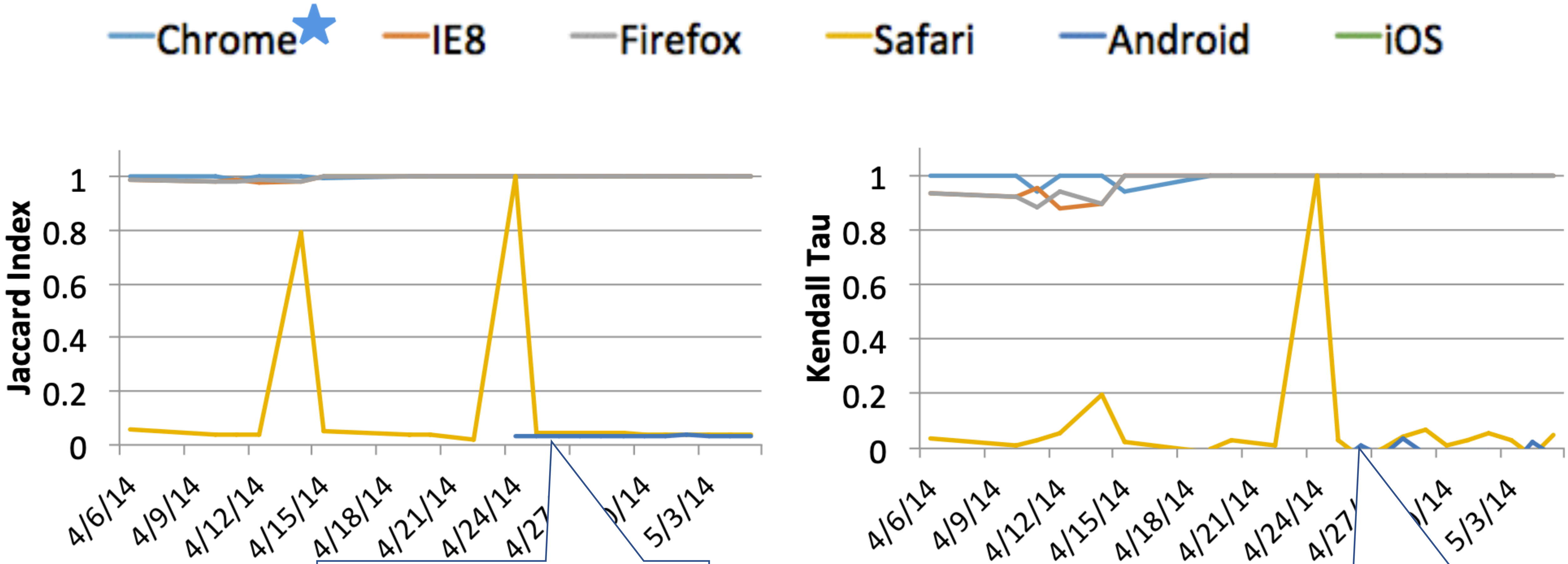
- Methodology is able to identify personalization
  - Manually verified incidents in HTML source
- Significant levels of price steering and discrimination
  - Not random — a small group of users are often personalized
- But, cannot say how or why these users get different prices
  - Could be due to browsers, purchase history, etc

What features enable personalization?

- Methodology: use synthetic (fake) accounts
  - Give them different features, look for personalization
  - Each day for 1 month, run standard set of searches
  - Add controls

Category	Feature	Tested Features
Account	Cookie	No Account, Logged In, No Cookies
User-Agent	OS	Win XP, Win 7, OS X, Linux
	Browser	Chrome 33, Android Chrome 34, IE 8, Firefox 25, Safari 7, iOS Safari 6
History	Click	Big Spender, Low Spender
	Purchase	Big Spender, Low Spender

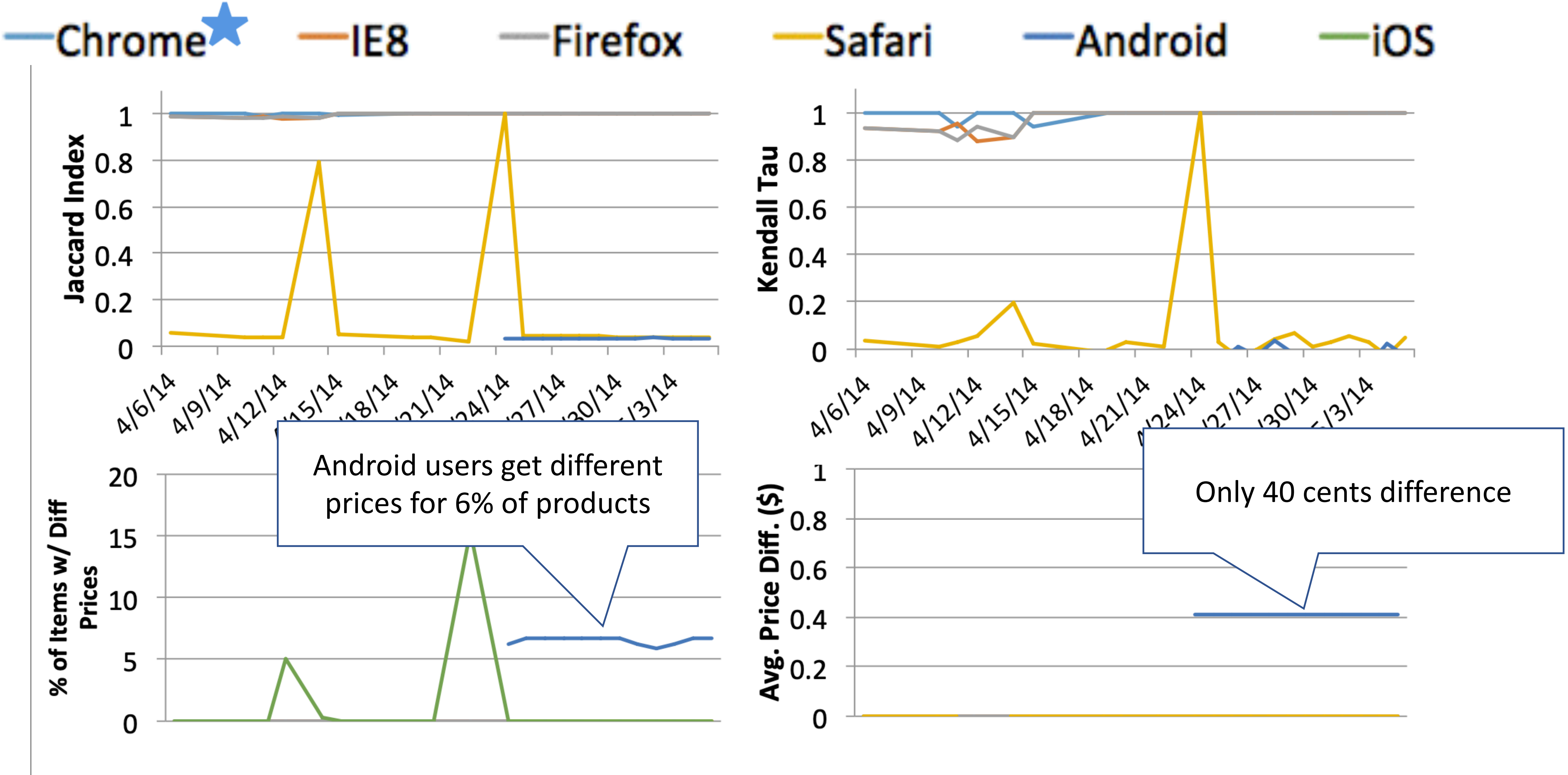
Example Result: Home Depot



Mobile users see completely different products!

… in a completely different order

Example Result: Home Depot



Results for different sites

- Orbitz & Cheaptickets
  - Logged in users get cheaper prices ($12 on average)
- Expedia & Hotels
  - A/B testing: assigns users to random bucket upon first visit
  - Some buckets are steered towards higher prices
  - $17 difference between buckets
- Travelocity: discriminates in favor of mobile users
  - $15 cheaper for mobile on average
- Priceline: recognizes cheapskates
  - They get different products in different order

Recap

- Developed methodology, measurement infrastructure to study price discrimination and steering
- Collected real-world data from 300 users
  - Evidence of personalization on 9 of the measured sites Conducted controlled experiments to identify features
  - Observed sites altering results based on based on: Account, Browser/OS, Purchase History

Discussion

- Part of a larger project
  - Understanding how web services collect data
  - How it affects the information users see
- Transparency
  - People don’t know when and how they are discriminated
  - Raising awareness is important
- Continuous Monitoring
  - Observe if, when, and how algorithms are changing
  - Develop active defense mechanisms