Data, Responsibly
fairness, neutrality and transparency in data analysis

Julia Stoyanovich
Drexel University, USA

Serge Abiteboul
INRIA, France

Gerome Miklau
UMass Amherst, USA

data RESPONSIBLY

EDBT 2016

Data for and about people

facebook

Gmail

R

Dropbox

freshdirect

foursquare

amazon

GitHub

Google Calendar

twitter

kindle

CHASE

skype

NETFLIX

UnitedHealthcare
A UnitedHealth Group Company

iTunes

Google+

Network Attached Storage

Google Docs

StateFarm

Picasa

tripadvisor

The promise of big data

power

enormous data sets: the 5Vs

enormous computational power

massively parallel processing

opportunity

improve people's lives, e.g., recommendation

accelerate scientific discovery, e.g., medicine

boost innovation, e.g., autonomous cars

transform society, e.g., open government

optimize business, e.g., advertisement targeting

goal - progress

Illustration: big data and health

Analysis of a person’s medical data, genome, social data

personalized medicine

personalized care and predictive measures

personalized insurance

expensive, or unaffordable, for those at risk

the same technology makes both possible!

Is data analysis impartial?

Big data is algorithmic, therefore it cannot be biased! And yet…

- All traditional evils of discrimination, and many new ones, exhibit themselves in the big data eco system
- We need novel technological solutions to identify and rectify **irresponsible data analysis practices**
- Technology alone won't do: also need **policy**, **user involvement** and **education** efforts, more on this later

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Data, responsibly

The problem is not only in the **technology**, but also in how its **used**

Because of its tremendous **power**, massive data analysis must be used **responsibly**

Roadmap

- ✓ Introduction
- Properties of responsible data analysis
  - ➡ Fairness
  - Diversity
  - Transparency
  - Neutrality
- Conclusion: towards a data responsible society

Staples online pricing

THE WALL STREET JOURNAL.

WHAT THEY KNOW

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.

WHAT PRICE WOULD YOU SEE?

lower prices offered to buyers who live in more affluent neighborhoods

Fairness is lack of bias

- Where does bias come from?
  - data collection
  - data analysis
- Analogy - scientific data analysis
  - collect a representative sample
  - do sound reproducible analysis
  - explain data collection and analysis

when data is about people, bias can lead to discrimination

The evils of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employer, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Disparate impact is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

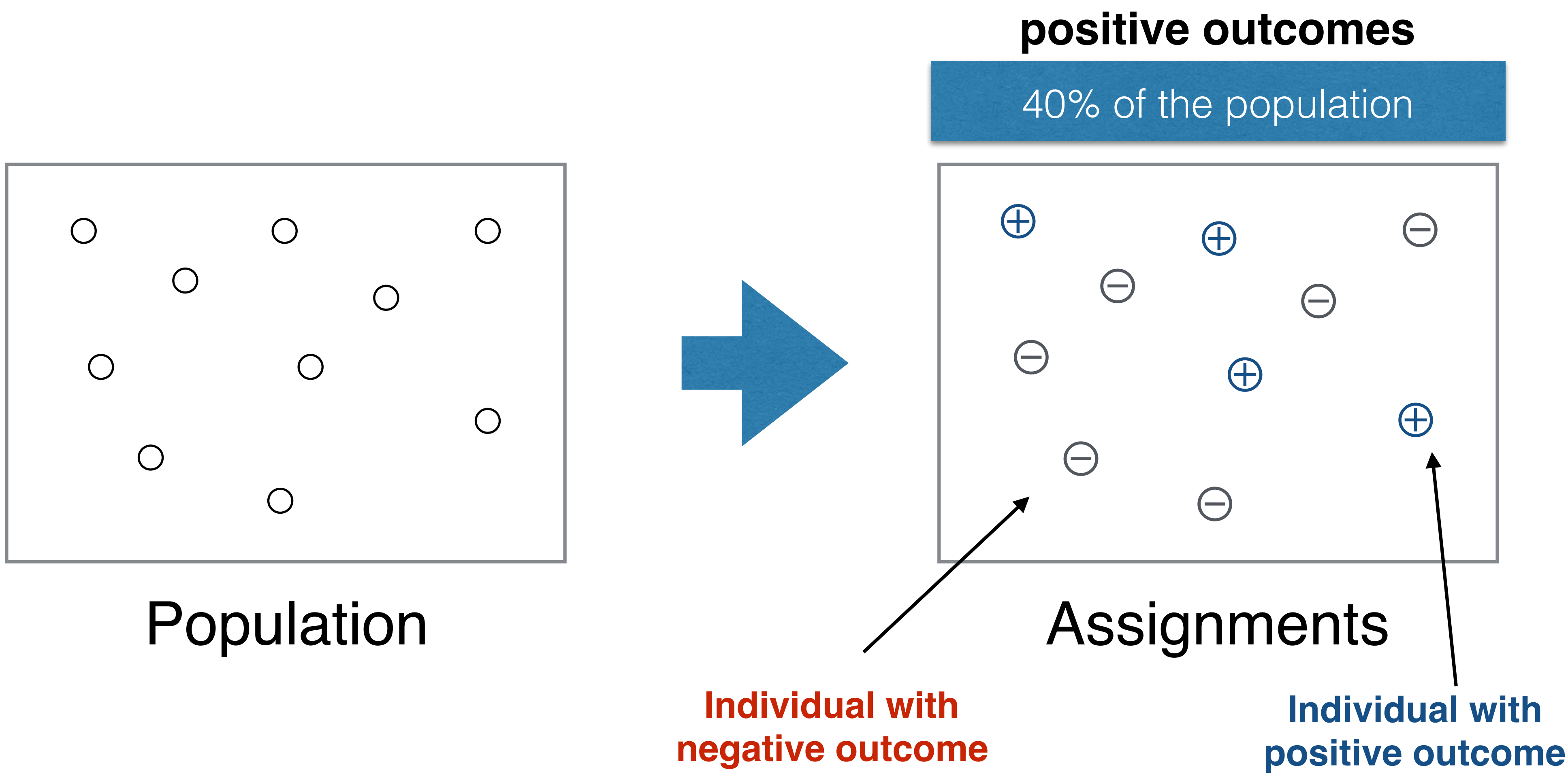
Outcomes

Consider a **vendor** assigning positive or negative **outcomes** to individuals.

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount

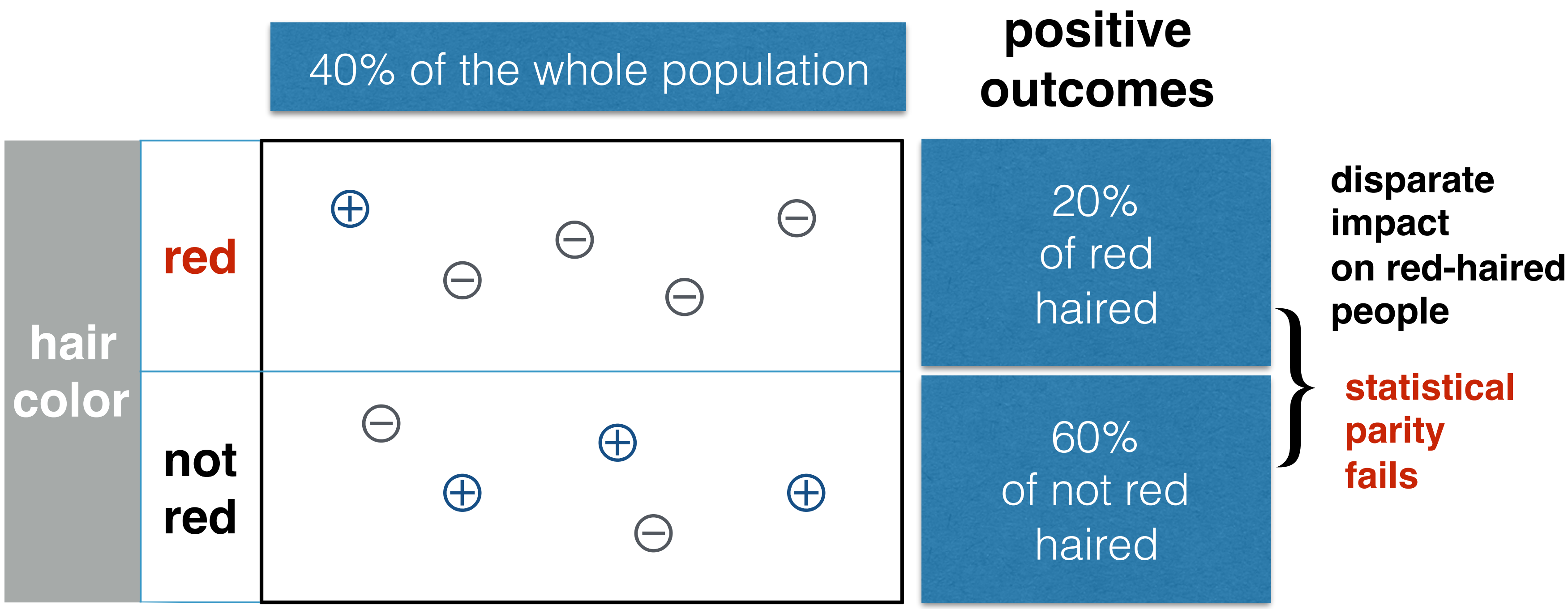
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



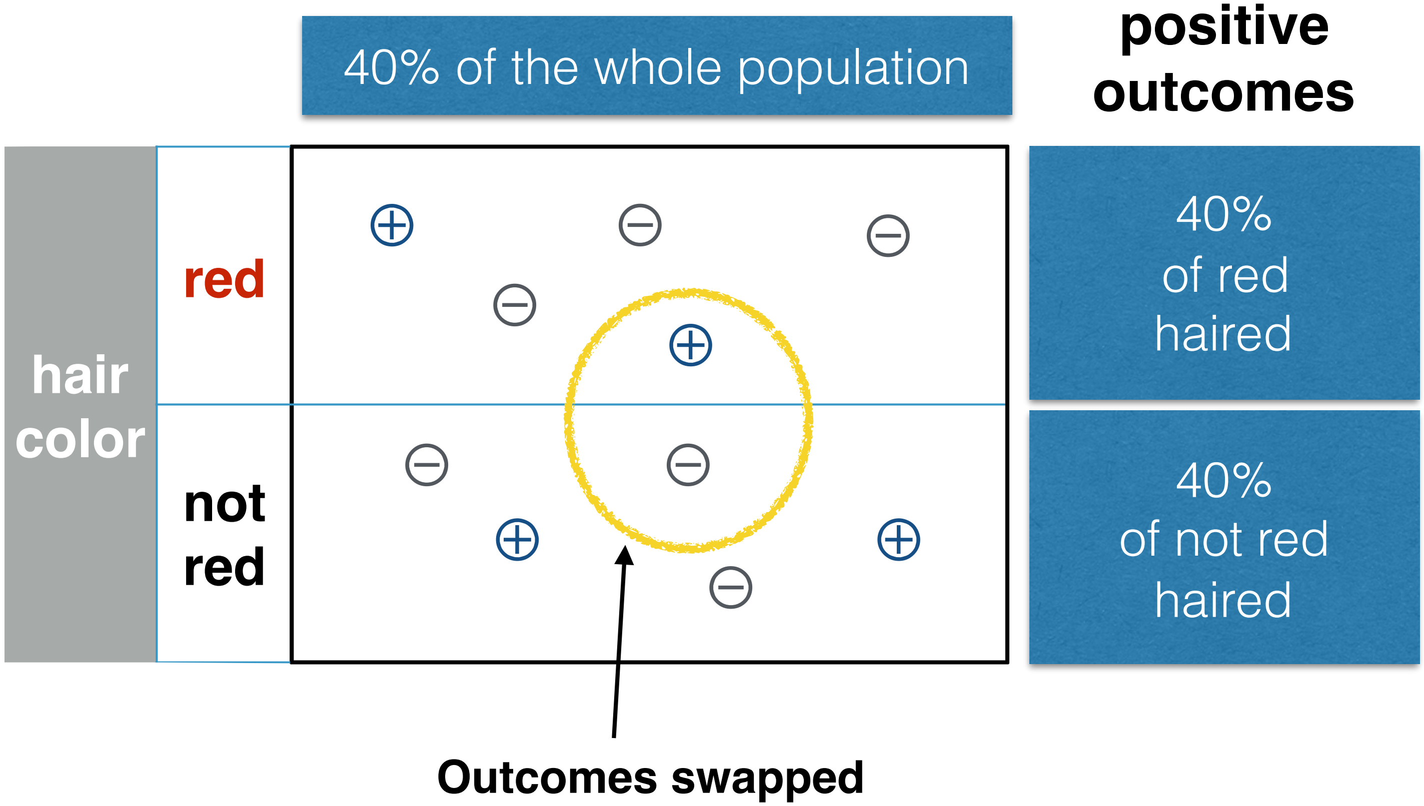
Sub-populations may be treated differently

Sub-population: those with red hair
(under the same assignment of outcomes)



Enforcing statistical parity

Statistical parity (aka **group fairness**)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Redundant encoding

Now consider the assignments under both **hair color** (protected) and **hair length** (innocuous)

		hair length		
		long	**not long**	**positive outcomes**
hair color	**red**	⊕	⊖ ⊖ ⊖ ⊖	20% of red haired
	not red	⊕ ⊕ ⊕	⊖ ⊖	60% of not red haired

Deniability
The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

Blinding does not imply fairness

Removing **hair color** from the vendor's assignment process does not prevent discrimination

		hair length	
		long	not long
hair color	red	⊕	⊖ ⊖ ⊖ ⊖
	not red	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20% of red haired

60% of not red haired

Assessing disparate impact
Discrimination is assessed by the effect on the protected sub-population, not by the input or by the process that lead to the effect.

Redundant encoding

Let's replace hair color with **race** (protected),
hair length with **zip code** (innocuous)

		zip code	
		10025	**10027**
race	**black**	⊕	⊖ ⊖ ⊖ ⊖
	white	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

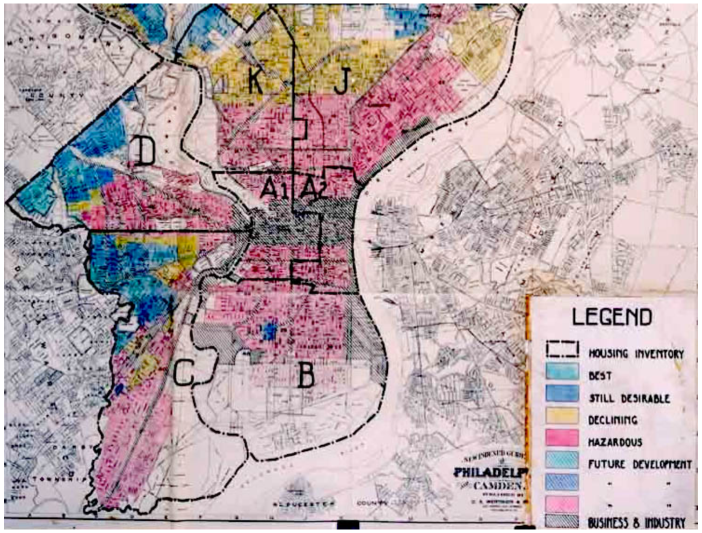
20%
of black

60%
of white

The evils of discrimination

Substituting hair color (protected) with hair length (innocuous) or race (protected) with zip code (innocuous) are examples of **redundant encoding**.

Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.



Discrimination may be unintended

Staples website estimated user’s location, **offering discounts** to those near rival stores, leading to discrimination w.r.t. to average income.

		rival store proximity	
		close	**far**
income	**low**	⊕	⊖ ⊖ ⊖ ⊖
	high	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20%
of low income

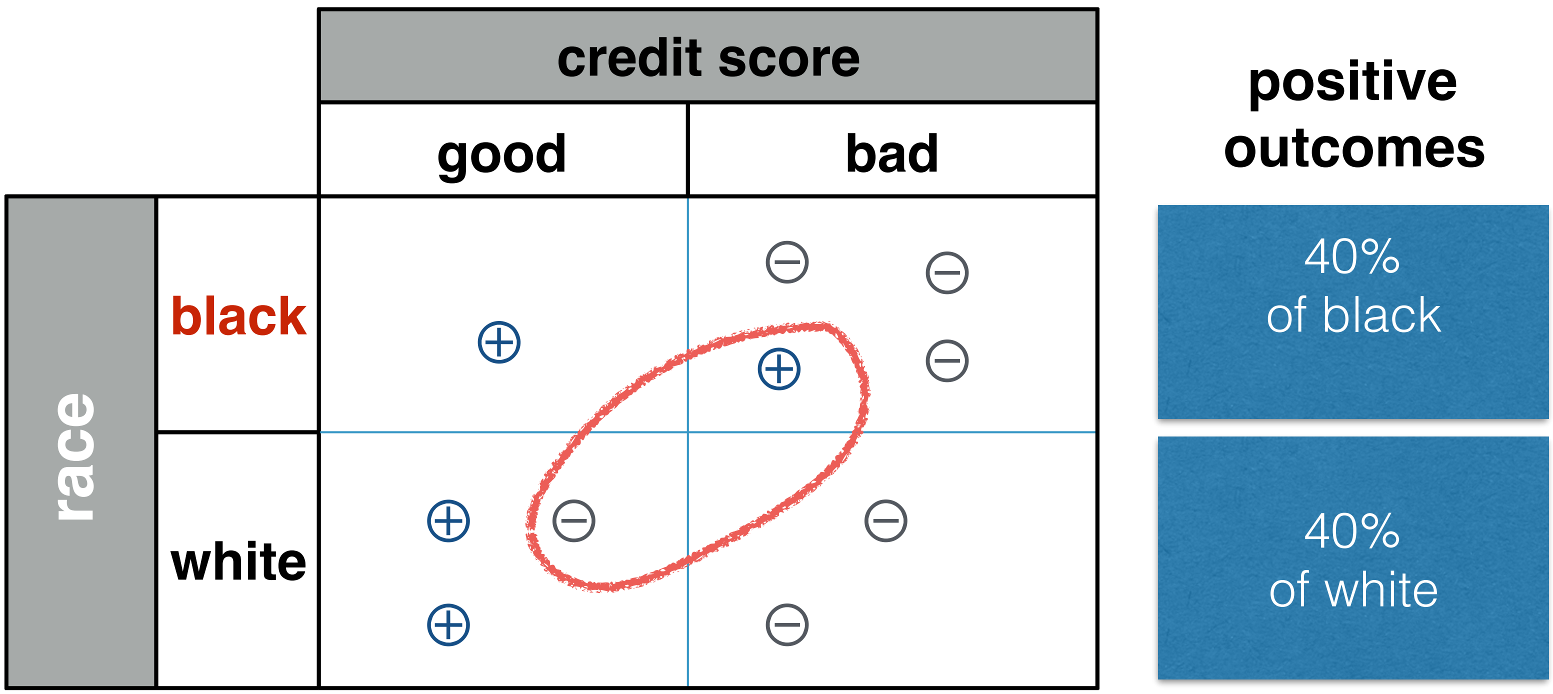
60%
of high income

Discrimination
Whether intentional or not, discrimination is unethical and, in many countries, illegal.

Imposing statistical parity

May be contrary to the goals of the vendor

positive outcome: offered a loan



Impossible to predict loan payback accurately.
Use past information, may itself be biased.

Justifying exclusion

Self-fulfilling prophecy
deliberately choosing the “wrong” (lesser qualified) members of the protected group to build bad track record

race	credit score: good	credit score: bad	positive outcomes
black	⊕	⊖ ⊕ ⊖ ⊖	40% of black
white	⊕ ⊖ ⊕	⊖ ⊖	40% of white

Justifying exclusion

Reverse tokenism
pointing to another "better" (more qualified) member of the protected group who also received a negative outcome

		credit score		
		good	**bad**	**positive outcomes**
race	**black**	⊕ ⊖ ⊖	⊖ ⊖	20% of black
	white	⊕ ⊕ ⊕	⊖ ⊖	60% of white

Effect on sub-populations

Simpson's paradox
disparate impact at the full population level disappears or reverses when looking at sub-populations!

		grad school admissions	
		admitted	**denied**
gender	**F**	1512	2809
	M	3715	4727

positive outcomes

35% of women

44% of men

UC Berkeley 1973: women applied to more competitive departments, with low rates of admission among qualified applicants.

Defeating statistical parity

If the vendor wants to avoid offering positive outcomes to red-hairs, they can try to find a disqualifying secondary attribute.

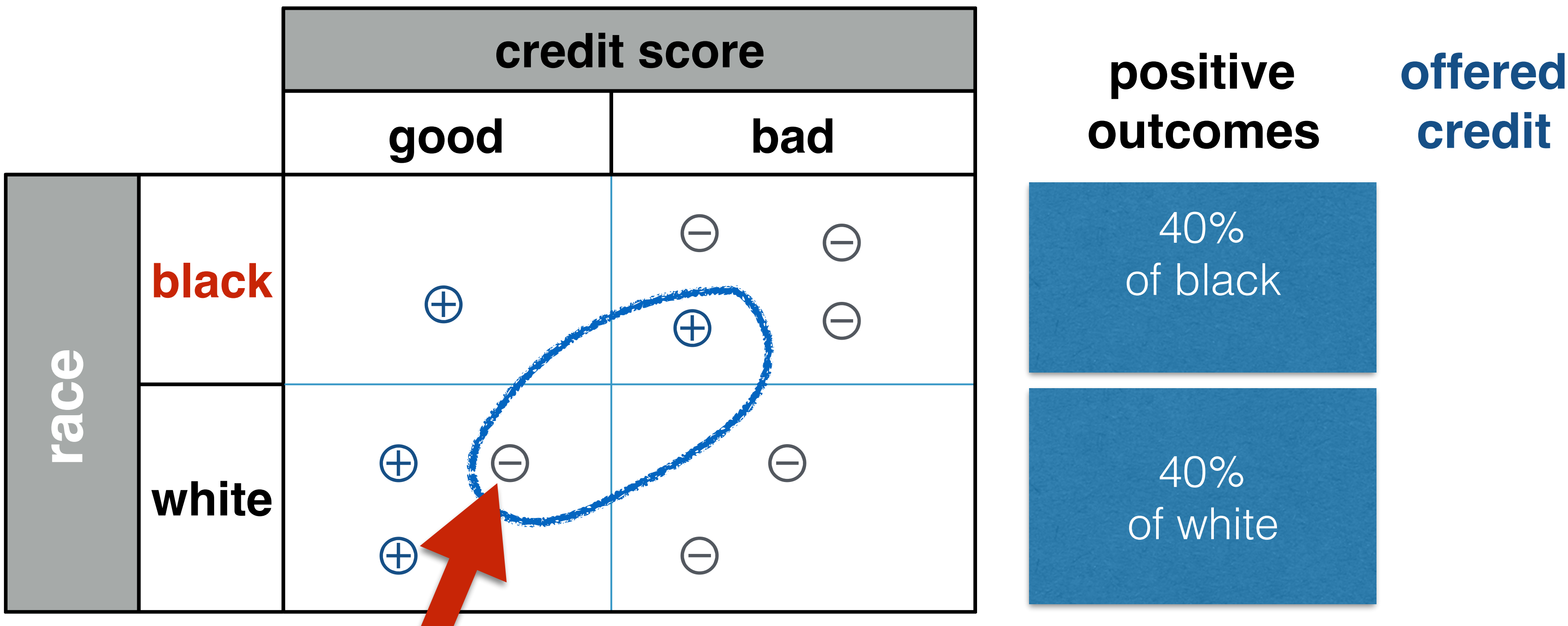
positive outcome: burger discount

		diet	
		vegetarian	**carnivore**
hair color	**red**	⊕ ⊕	⊖ ⊖ ⊖
	not red	⊖ ⊖	⊕ ⊖ ⊕

offered	**accepted**
40% of red haired	0% of red haired
40% of not red haired	40% of not red haired

Is statistical parity sufficient?

Statistical parity (aka **group fairness**)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Individual fairness
any two individuals who are similar w.r.t. a particular task should receive similar outcomes

Discrimination-aware data analysis

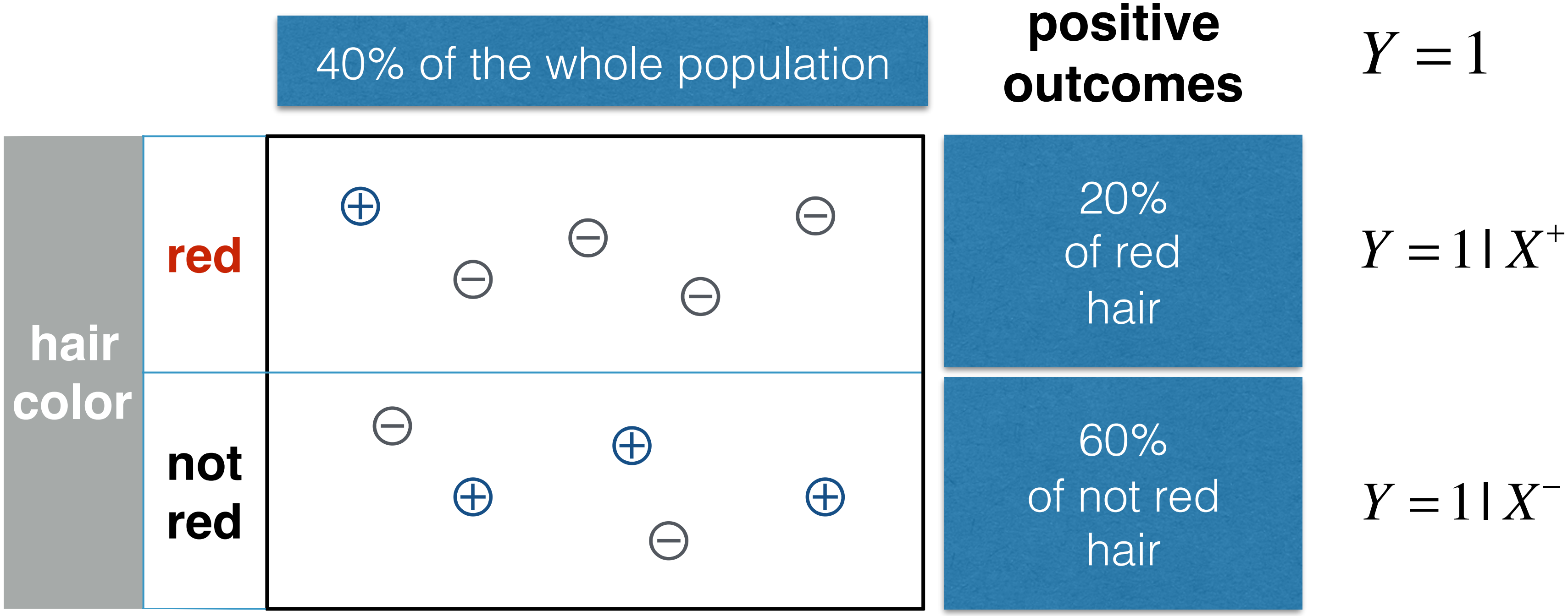
- **Identifying discrimination**
  - mining for discriminatory patterns in (input) data
  - verifying data-driven applications
- **Preventing discrimination**
  - data pre-processing
  - model post-processing
  - model regularization

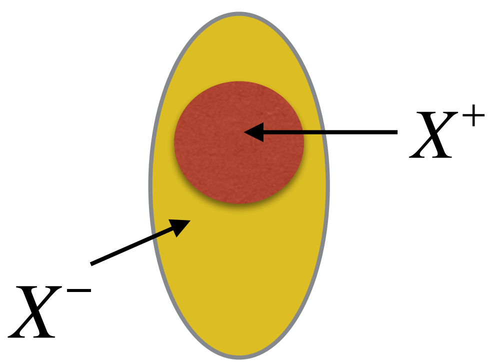
[Ruggieri *et al.*; 2010]
[Luong *et al.*; 2011]
[Pedresci *et al.*; 2012]
[Romei *et al.*; 2012]
[Hajian & Domingo-Ferrer; 2013]
[Mancuhan & Clifton; 2014]
[Kamiran & Calders; 2009]
[Kamishima *et al.*; 2011]
[Mancuhan & Clifton; 2014]
[Feldman *et al.*; 2015]
[Dwork *et al.*; 2012]
[Zemel *et al.*; 2013]

many more….

both rely on discrimination criteria

How do we quantify discrimination?





discrete (binary) protected feature *S*

X^+ are members of X with $S=1$
X^- are members of X with $S=0$

Discrimination criteria

[Indre Zliobaite, CoRR abs/1511.00148 (2015)]

- **Statistical tests** check how likely the difference between groups is due to chance - is there discrimination?
- **Absolute measures** express the absolute difference between groups, quantifying the magnitude of discrimination
- **Conditional measures** express how much of the difference between groups cannot be explained by other attributes, also quantifying the magnitude of discrimination
- **Structural measures** how wide-spread is discrimination? Think Simpson's paradox, individual fairness.

Discrimination criteria

[Indre Zliobaite, CoRR abs/1511.00148 (2015)]

Table III. Summary of absolute measures. Checkmark (✓) indicates that it is directly applicable in a given machine learning setting. Tilde (∼) indicates that a straightforward extension exists (for instance, measuring pairwise).

	Protected variable			Target variable		
Measure	Binary	Categoric	Numeric	Binary	Ordinal	Numeric
Mean difference	✓	∼		✓		✓
Normalized difference	✓	∼		✓		
Area under curve	✓	∼		✓	✓	✓
Impact ratio	✓	∼		✓		
Elift ratio	✓	∼		✓		
Odds ratio	✓	∼		✓		
Mutual information	✓	✓	✓	✓	✓	✓
Balanced residuals	✓	∼		∼	✓	✓
Correlation	✓		✓	✓		✓

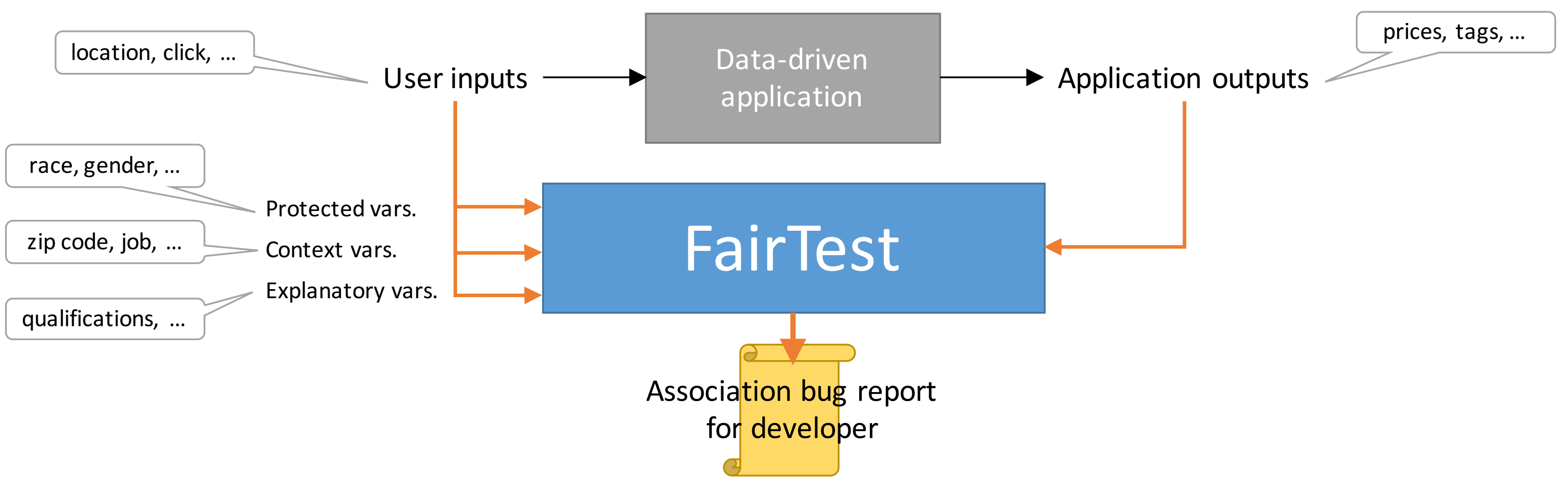
a proliferation of task-specific measures

FairTest: identifying discrimination

[F. Tramèr *et al.*, arXiv:1510.02377 (2015)]

A test suite for data analysis applications

- Tests for **unintentional discrimination** according to several representative discrimination measures
- Automates search for **context-specific associations** (recall Simpson's paradox) between protected variables and application outputs
- Report findings, ranked by association **strength** and affected **population size**

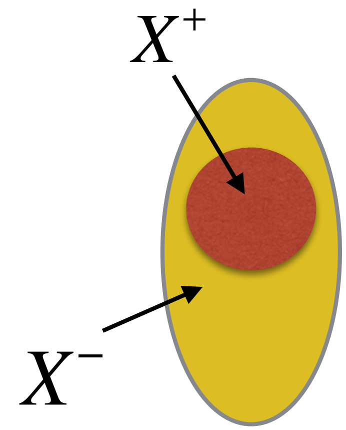


http://www.cs.columbia.edu/~djhsu/papers/fairtest-privacycon.pdf

dataRESPONSIBLY

FairTest: discrimination measures

[F. Tramèr *et al.*, arXiv:1510.02377 (2015)]



Binary ratio / difference compares probabilities of a single output for two groups

$$\Pr(Y=1 \mid X^+) - \Pr(Y=1 \mid X^-)$$

Easy to extend to non-binary outputs, not easy to overcome binary protected class membership

$$\frac{\Pr(Y=1 \mid X^+)}{\Pr(Y=1 \mid X^-)} - 1$$

Mutual information measures statistical dependence between outcomes and protected group membership

Works for non-binary outputs, class membership, can be normalized; bad for continuous values, does not incorporate of order among values

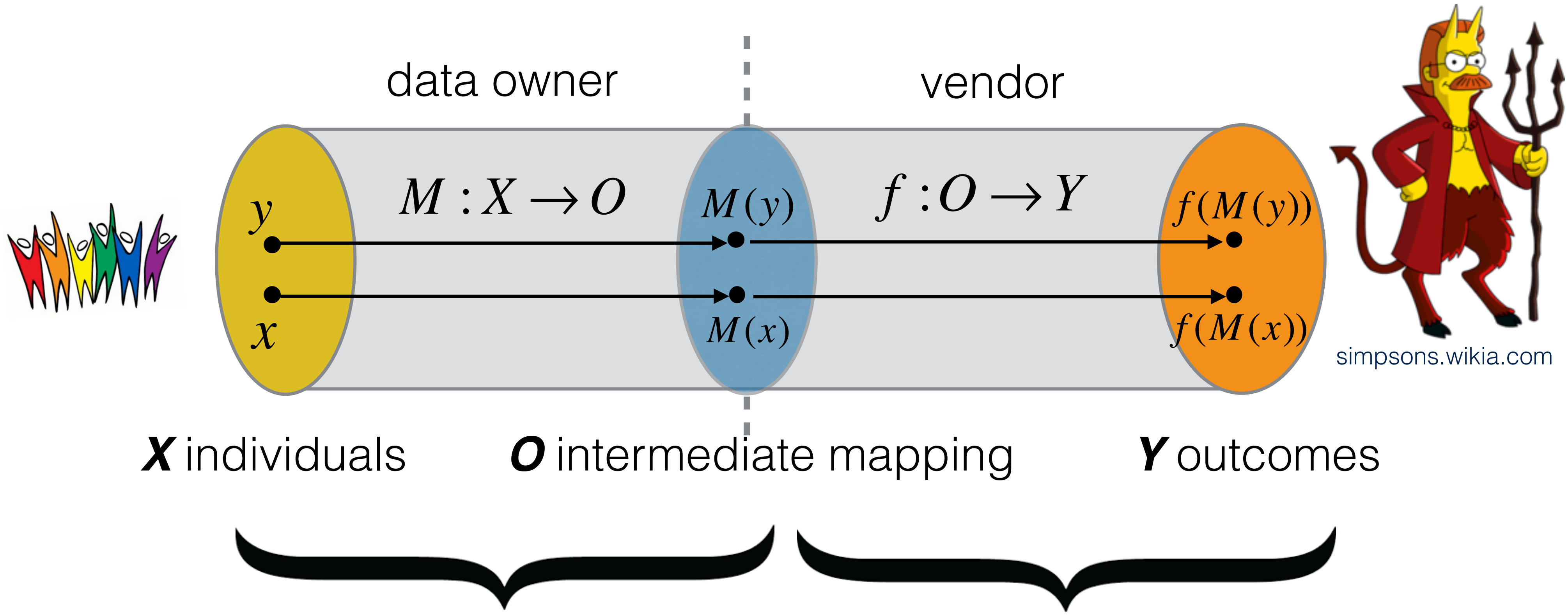
$$\sum \Pr(y,s) \ln \frac{\Pr(y,s)}{\Pr(y)\Pr(s)}$$

Pearson's correlation measures strength of linear relationship between outcomes and protected group membership

Works well for ordinal and continuous values, may detect non-linear correlations, is easy to interpret; finding a 0 correlation does not imply that S and Y are independent

Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

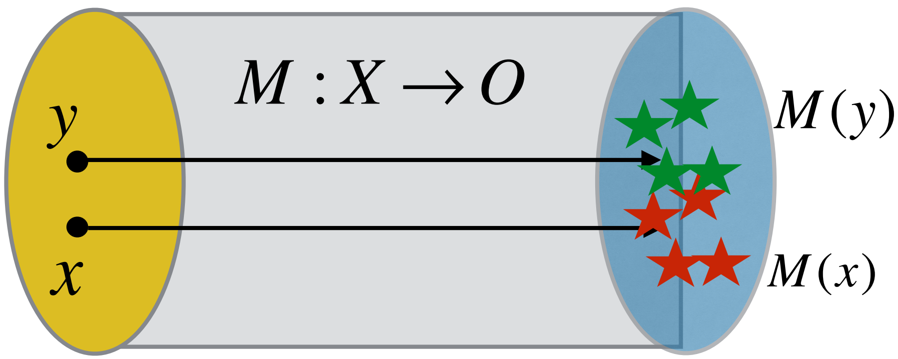


fairness enforced at this step **vendor cannot introduce bias**

Task-specific fairness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



X individuals ***O*** intermediate mapping

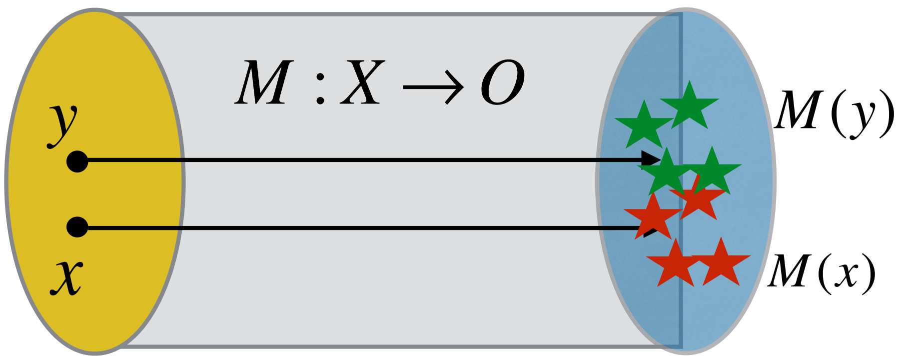
A task-specific similarity metric is given $d(x,y)$

$M : X \rightarrow O$ is a **randomized mapping**: an individual is mapped to a distribution over outcomes

Fairness through a Lipschitz mapping

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



A task-specific similarity metric is given $d(x,y)$

X individuals ***O*** intermediate mapping

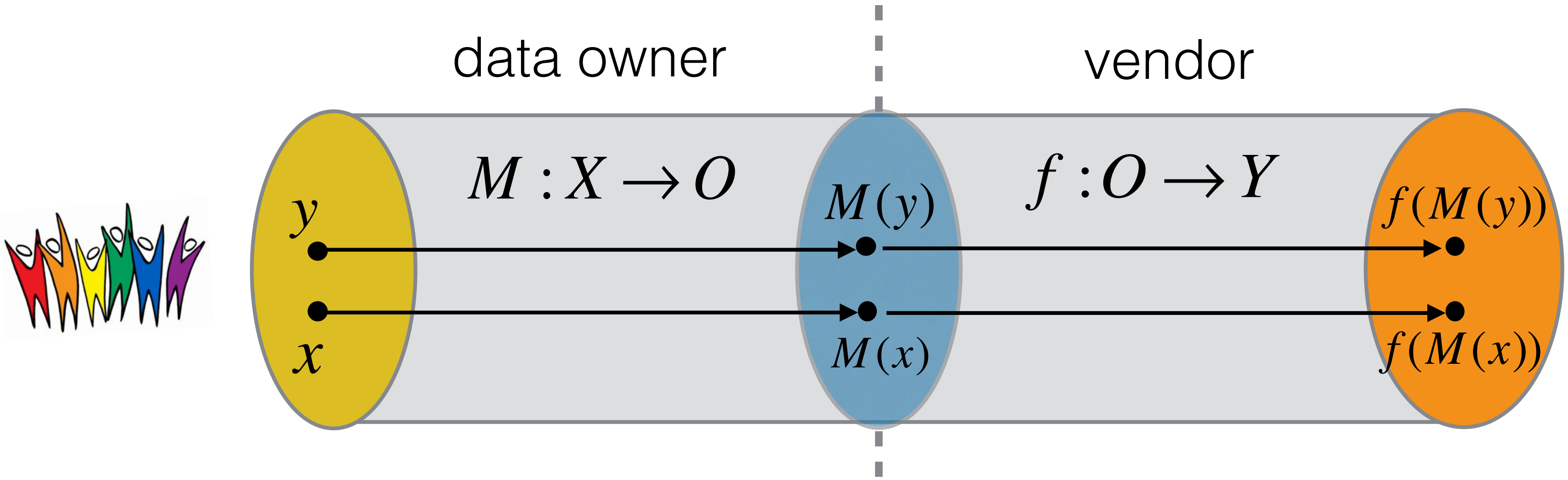
M is a Lipschitz mapping if $\forall x,y \in X \quad \|M(x),M(y)\| \leq d(x,y)$

close individuals map to close distributions

What about the vendor?

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Vendors can efficiently maximize expected utility, subject to the Lipschitz condition



X individuals   ***O*** intermediate mapping   ***Y*** outcomes

Computed with a linear program of size $poly(|X|,|Y|)$

the same mapping can be used by multiple vendors

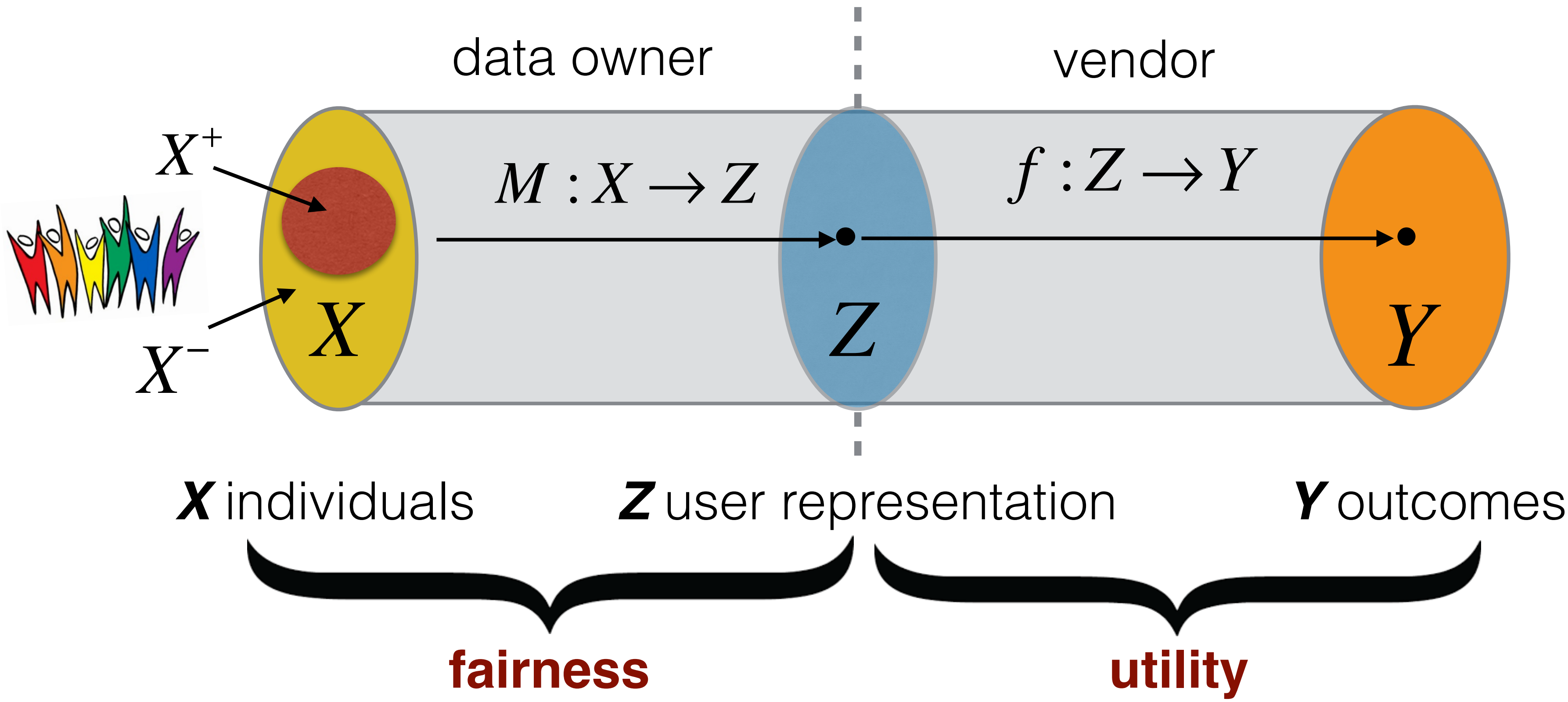
Fairness through awareness: summary

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

- An early work in this space, proposes a principled data pre-processing approach
- Stated as an **individual fairness** condition but also leads to **group fairness**
- Relies on an externally-supplied task-specific similarity metric - magic!
- Is not formulated as a learning problem, does not generalize to unseen data

Learning fair representations

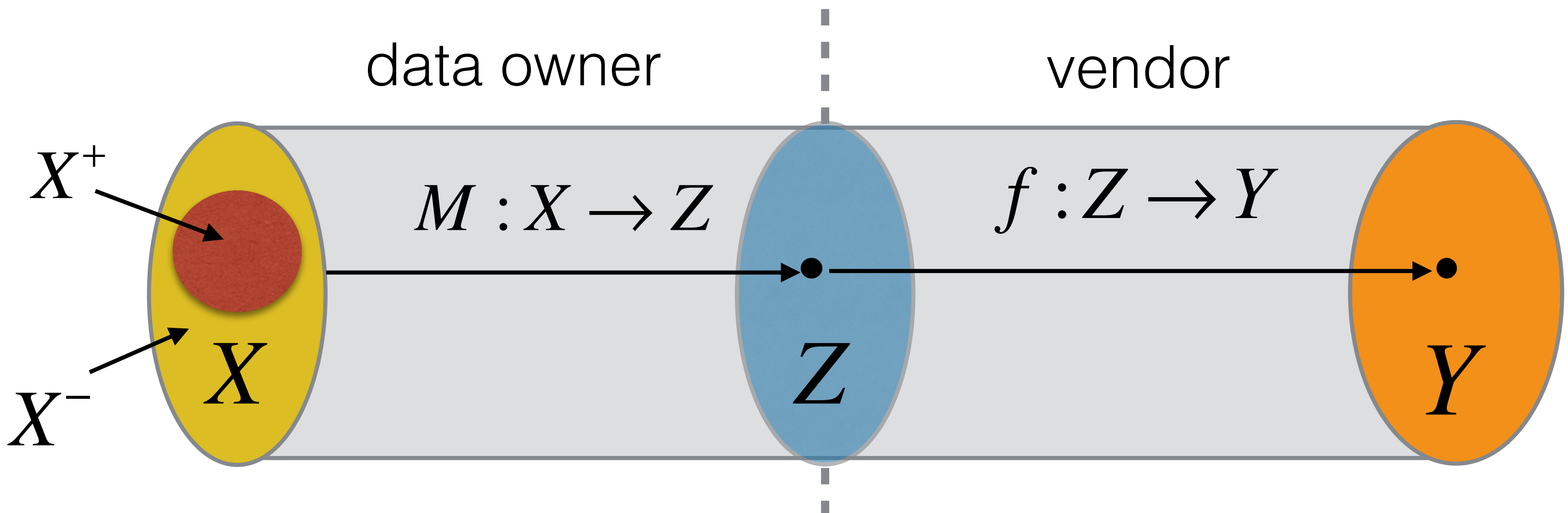
[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



Idea: remove reliance on a "fair" similarity measure, instead **learn** representations of individuals, distances

Fairness and utility

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



Learn a **randomized mapping** M(X) to a set of K prototypes Z

M(X) should lose information about membership in S $\quad P(Z \mid S=0) = P(Z \mid S=1)$

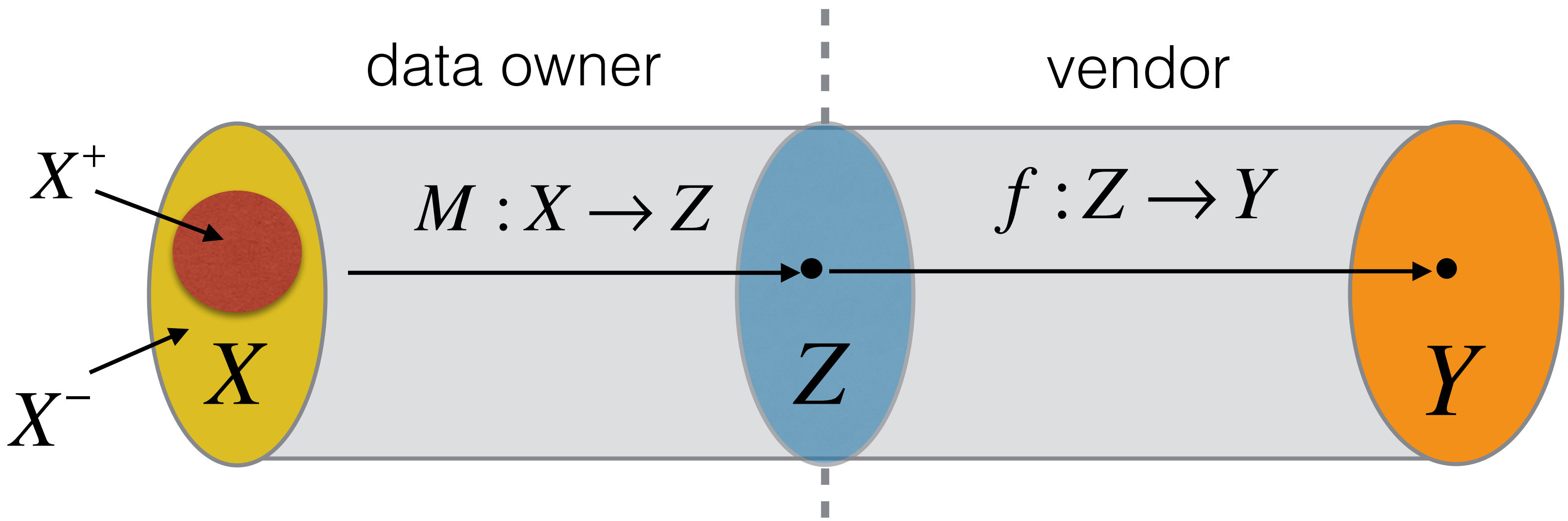
M(X) should preserve other information so that vendor can maximize utility

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

The objective function

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

$$P_k^+ = P(Z = k \mid x \in X^+)$$

$$P_k^- = P(Z = k \mid x \in X^-)$$

$$L_z = \sum_k \left| P_k^+ - P_k^- \right| \qquad L_x = \sum_n (x_n - \widehat{x}_n)^2$$

$$L_y = \sum_n -y_n \log \widehat{y}_n - (1 - y_n) \log(1 - \widehat{y}_n)$$

data RESPONSIBLY

Learning fair representations: summary

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]

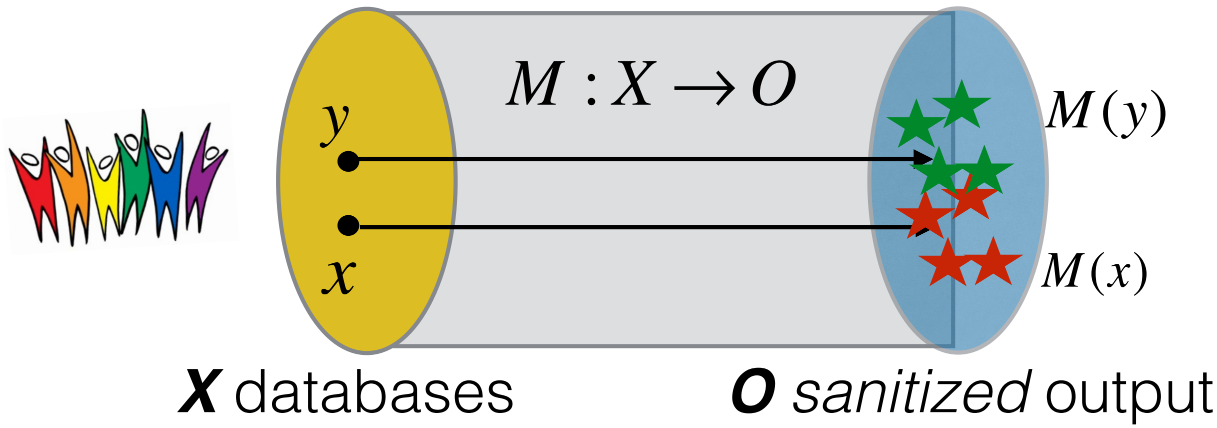
- A principled learning framework in the data pre-processing / classifier regularization category
- **Evaluation** of accuracy, discrimination (group fairness) and consistency (individual fairness), promising results on real datasets
- Not clear how to set *K*, so as to trade off accuracy / fairness
- The mapping is **task-specific**

Gaps and directions

- Handling a broader range of tasks, beyond task-specific measures
- Fairness in multi-step data processing pipelines
- Connection between fairness and privacy

Connection to privacy

Fairness through awareness generalizes differential privacy



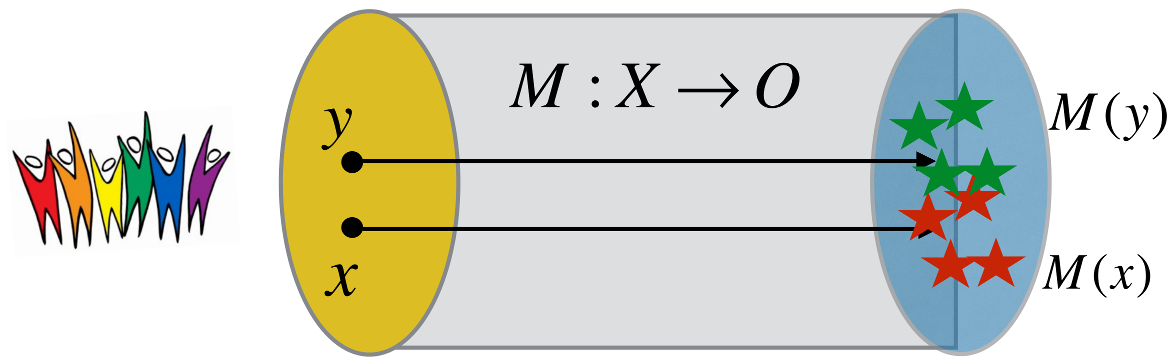
close databases map to close output distributions

Databases that differ in one record.

Connection to privacy

Does the fairness mapping provide privacy?

Similar individuals (according to $d(x,y)$) are hard to distinguish in the intermediate mapping. This provides a form of protection similar to anonymity based privacy.



X individuals ***O*** intermediate mapping

It depends on the metric *d* and on whether individual similarity is based on sensitive properties.

Roadmap

- ✓ Introduction
- Properties of responsible data analysis
  - ✓ Fairness
  - ➡ Diversity
  - Transparency
  - Neutrality
- Conclusion: towards a data responsible society

Illustration: online dating

Dating query: female, 40 or younger, at least some college, in order of decreasing income

Results are homogeneous at top ranks

Both the seeker (asking the query) and the matches (results) are dissatisfied

Crowdsourcing, crowdfunding, ranking of Web search results, … - all subject to this problem

the rich get richer, the poor get poorer


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K

… 999 matches


PhD, 36 years old
makes $100K

… 9999 matches


BS, 27 years old
makes $80K

What do we mean by diversity?

- For a given user consuming information in search and recommendation, relevance is important, but so are:
  - **diversity** - avoid returning similar items
  - **novelty** - avoid returning known items
  - **serendipity** - surprise the user with unexpected items
- For a set of users
  - uncommon information needs must be met: **less popular** "in the tail" queries constitute the overwhelming majority
  - lack of diversity can lead to **exclusion**

> **Jonas Lerman:** "… the nonrandom, systematic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle…"

Diversity when data is about people

- Data must be **representative** - bias in data collection may be amplified in data analysis, perpetuating the original bias

- In this sense diversity is related to **coverage**

Result diversification

From the pool of relevant items, identify a subset with items that are dissimilar and maintain a high cumulative relevance.

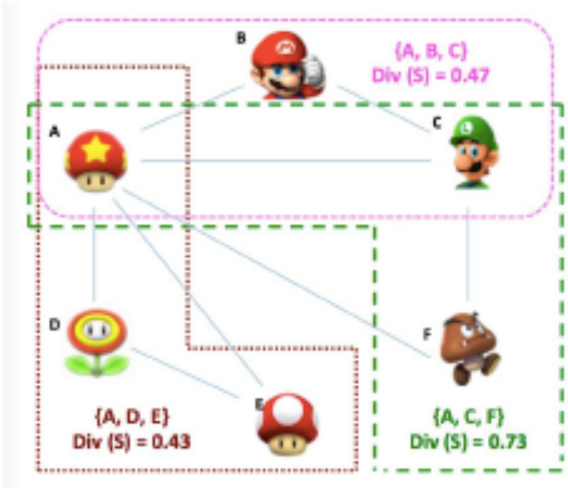
- Web search, product search, recommendation
  - diversity is defined for **pairs of items** (a distance)
  - pair-wise diversity aggregated into **set-wise** diversity (avg, min, max)
  - NP-hard, clever heuristics / approximations
- Diversity in composite items (bundles), e.g., travel package
- Building teams - items are people, based on **complementarity**, not explicitly on diversity

[C.Yu, L. Lakshmanan, S. Amer-Yahia; EDBT 2009]
[T.Deng, *et al.*; PVLDB 2013]
[S. Abbar, S. Amer-Yahia, P. Indyk, S. Mahabadi; WWW 2013]
[D.C. Thang, N.T. Tam, N.Q. Viet Hung, K. Aberer; DEXA 2015]
many more….

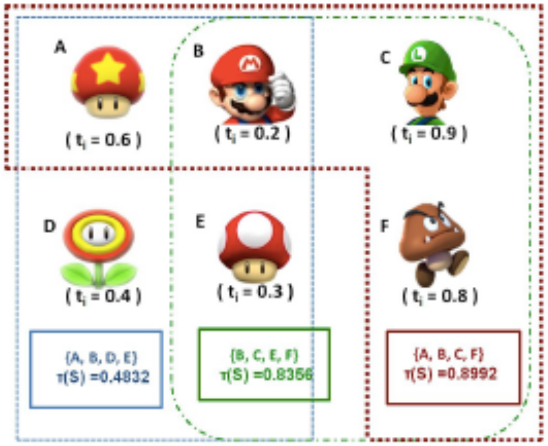
Diversity of opinion in crowdsourcing

[T. Wu, L. Chen, P. Hui, C.J. Zhang, W. Li; PVLDB 2015]

- Importance of diversity of opinion for **accuracy** is well-understood in the social sciences
  - Diversity is crucial in crowdsourcing, see Surowiecki "*The Wisdom of the Crowds*" 2005
  - The "Diversity trumps ability theorem"



- Crowd diversity: an aggregate of pair-wise diversity
- **S-Model:** similarity-driven / task-independent
- **T-Model**: task-driven, opinions are probabilistic



Rank-aware clustering

[J. Stoyanovich, S. Amer-Yahia, T. Milo; EDBT 2011]


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K

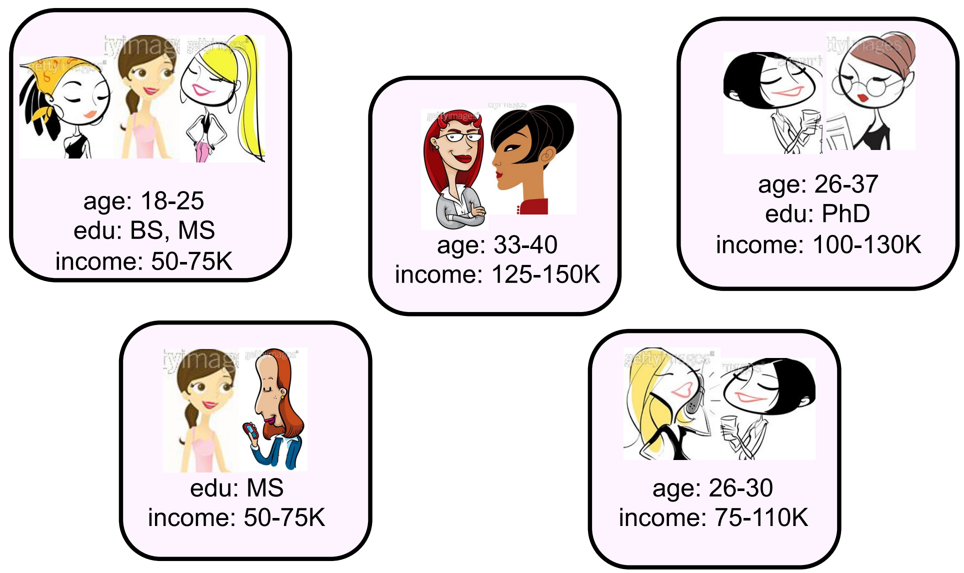
… 999 matches


PhD, 36 years old
makes $100K

… 9999 matches


BS, 27 years old
makes $80K

Return clusters that expose **best from among comparable** items (profiles) w.r.t. user preferences



More diverse items seen, and liked, by users

Users are more engaged with the system

Gaps and directions

- **An extremely important topic**: we are witnessing lack of diversity in a wide variety of domains, with serious consequences
- Technically, a variety of application-specific formulations and heuristic solutions
- Not explicitly related to coverage / fairness
- Data specifically about people is rarely considered

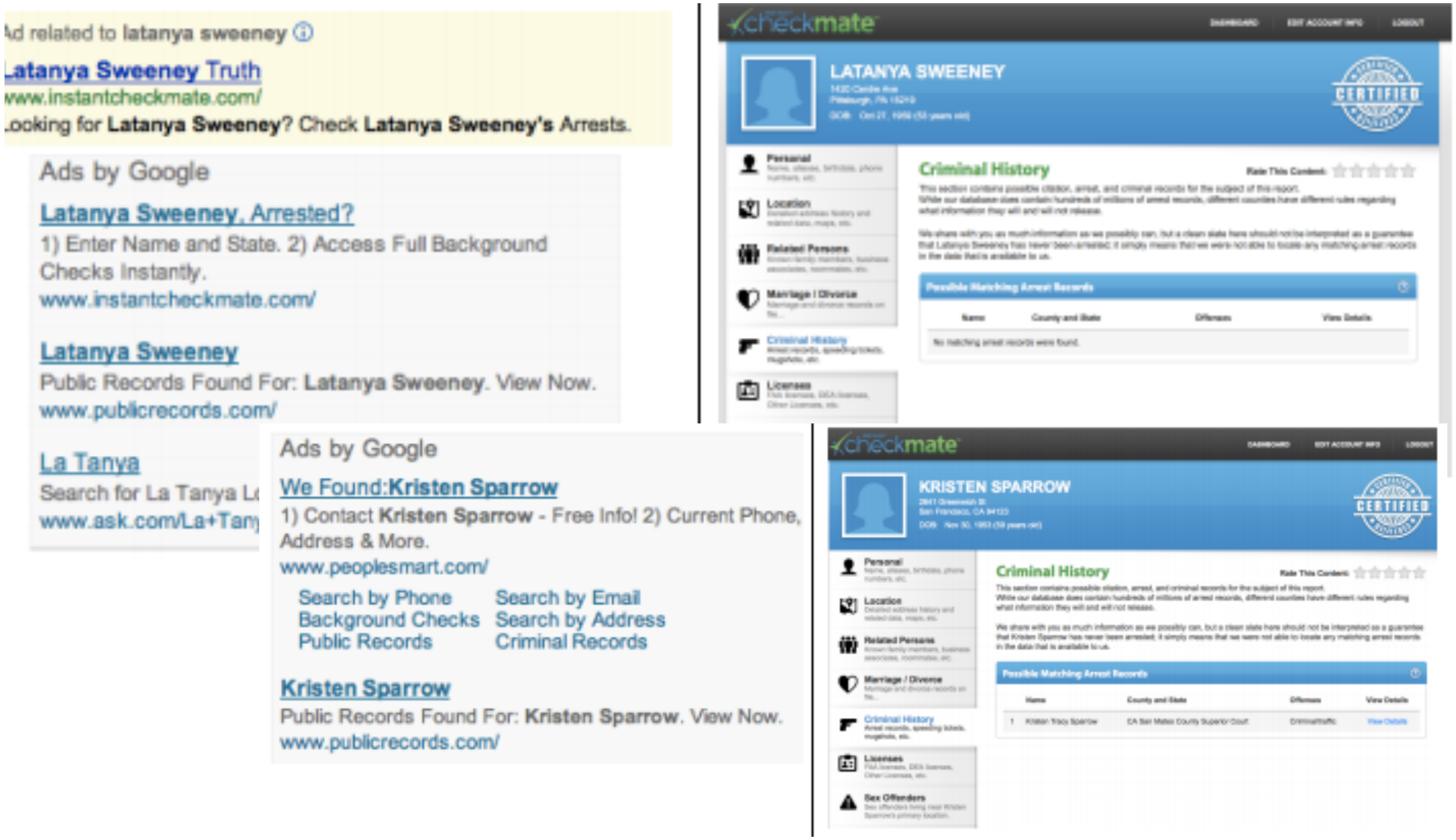
Roadmap

- ✓ Introduction
- • Properties of responsible data analysis
  - ✓ Fairness
  - ✓ Diversity
  - ➡ Transparency
  - • Neutrality
- • Conclusion: towards a data responsible society

Racially identifying names

[Latanya Sweeney; *CACM 2013*]

Google AdSense



racially identifying names trigger ads suggestive of an arrest record

Transparency and accountability

- Users and regulators must be able to **understand** how raw data was selected, and what operations were performed during analysis
- Users want to **control** what is recorded about them and how that information is used
- Users must be able to **access** their own information and correct any errors (US Fair Credit Reporting Act)
- **Transparency** facilitates **accountability** - verifying that a services performs as it should, and that data is used according to contract
- Related to **neutrality**, more on this later

the problem is broad, we focus on a specific case

Specifically: Ad targeting online

- **Users** browse the Web, consume content, consume ads (see / click / purchase)
- **Content providers** outsource advertising to third-party ad networks, e.g., Google's DoubleClick
- **Ad networks** track users across sites, to get a global view of users' behaviors
- **Google Ad Settings** aims to provide **transparency** / give **control to users** over the ads that they see

do users truly have transparency / choice or is this a placebo button?

Google Ads Settings

Your Google profile

Gender

35–44

Age

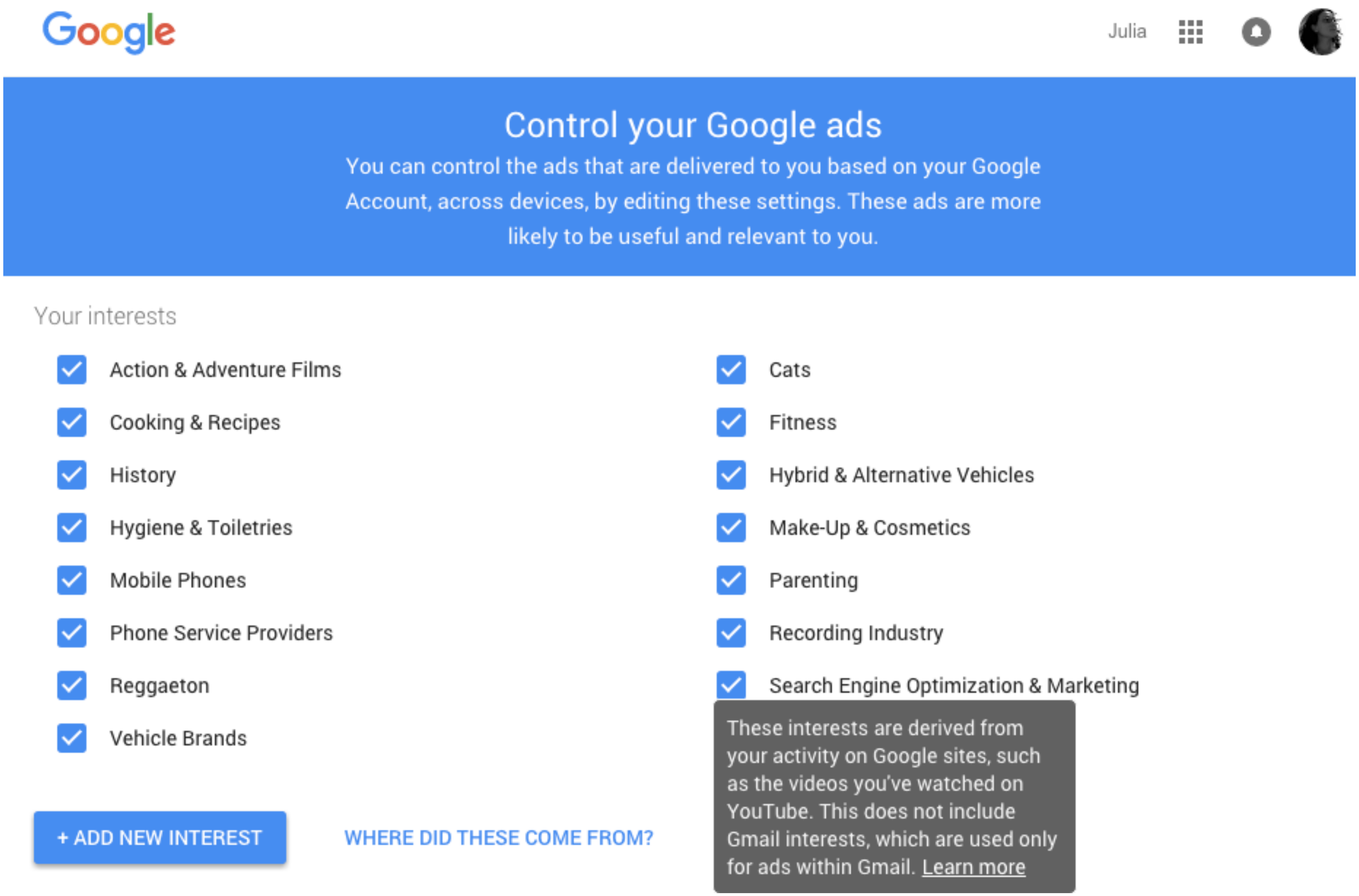
Ads based on your interests — ON

Improve your ad experience when you are signed in to Google sites

With Ads based on your interests ON	With Ads based on your interests OFF
• The ads you see will be delivered based on your prior search queries, the videos you've watched on YouTube, as well as other information associated with your account, such as your age range or gender • On some Google sites like YouTube, you will see ads related to your interests, which you can edit at any time by visiting this page • You can block some ads that you don't want to see	• You will still see ads and they may be based on your general location (such as city or state) • Ads will not be based on data Google has associated with your Google Account, and so may be less relevant • You will no longer be able to edit your interests • All the advertising interests associated with your Google Account will be deleted

http://www.google.com/settings/ads

Google Ads Settings



http://www.google.com/settings/ads

AdFisher

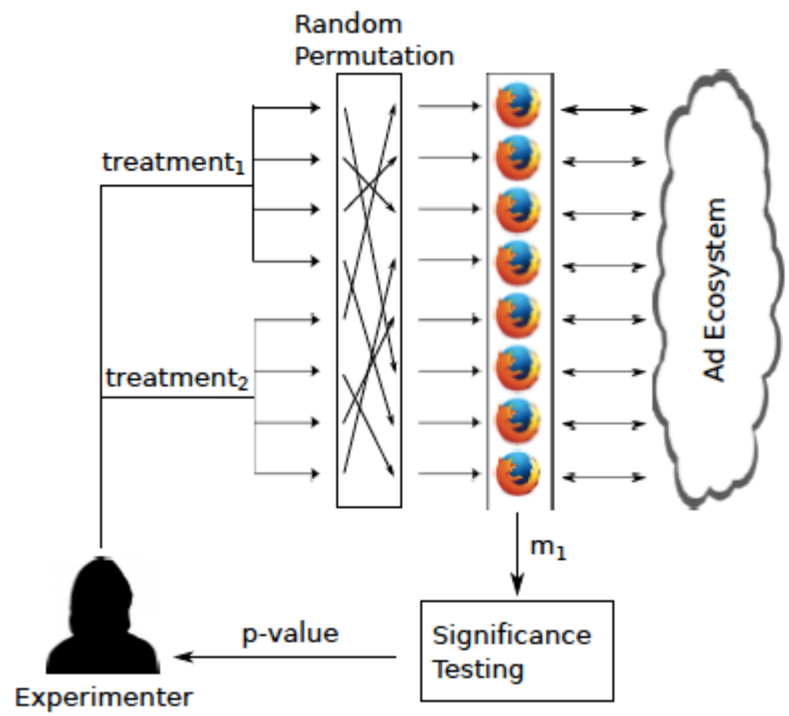
[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

From anecdotal evidence to statistical insight:

How do user behaviors, ads and ad settings interact?

Automated randomized controlled experiments for studying online tracking

Individual data use transparency: ad network must share the information it uses about the user to select which ads to serve to him



data RESPONSIBLY

AdFisher: discrimination

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

Non-discrimination: Users differing only in protected attributes are treated similarly

Causal test: Find that a protected attribute changes ads

Experiment 1: **gender and jobs**

Specify gender (male/female) in Ad Settings, simulate interest in jobs by visiting employment sites, collect ads from Times of India or the Guardian

Result: males were shown ads for higher-paying jobs significantly more often than females (1852 vs. 318)

violation

AdFisher: transparency

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

Transparency: User can view data about him used for ad selection

Causal test: Find attribute that changes ads but not settings

Experiment 2: **substance abuse**

Simulate interest in substance abuse in the experimental group but not in the control group, check for differences in Ad Settings, collect ads from Times of India

Result: no difference in Ad Settings between the groups, yet significant differences in what ads are served: rehab vs. stocks + driving jobs

violation

AdFisher: accountability

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

Ad choice: Removing an interest decreases the number of ads related to that interest.

Causal test: Find that removing an interest causes a decrease in related ads

Experiment 3: **online dating**

Simulate interest in online dating in both groups, remove "Dating & Personals" from the interests on Ad Settings for experimental group, collect ads

Result: members of experimental group do not get ads related to dating, while members of the control group do

compliance

Other work

- XRay [Lecuyer et al.; *USENIX Security 2014*], Sunlight [Lecuyer et al., CCS 2015]: statistical testing of lack of transparency, discrimination in online advertising
- **Privacy**: awareness of privacy leaks, usability of tools
- **Tracking**: awareness of tracking, reverse-engineering
- Pricing transparency, e.g., Uber surge pricing [L. Chen, A. Mislove, C. Wilson; *IMC 2015*]
- Data Transparency Lab: technology + policy, see DTL 2015 for pointers (datatransparencylab.org)

DATA TRANSPARENCY LAB

Is this down to privacy?

A shift from privacy and consent to responsible use!
[E. Kenneally; *SIGCAS 2015*]

Gaps and directions

- There is more to transparency than on-line behavioral marketing
- Promising approaches to help support transparency
  - personal information management
  - provenance & distributed access control
  - program verification

Personal data

- Lots of personal data, raising many problems
  - loss of functionality because of fragmentation
  - loss of control over data
  - loss of freedom: vendor lock-in
- A few companies concentrate most of the world's data and analytical power
- A few companies control all your personal data

enter personal information management systems (PIMS)

The PIMS: a paradigm shift

many Web services, each running

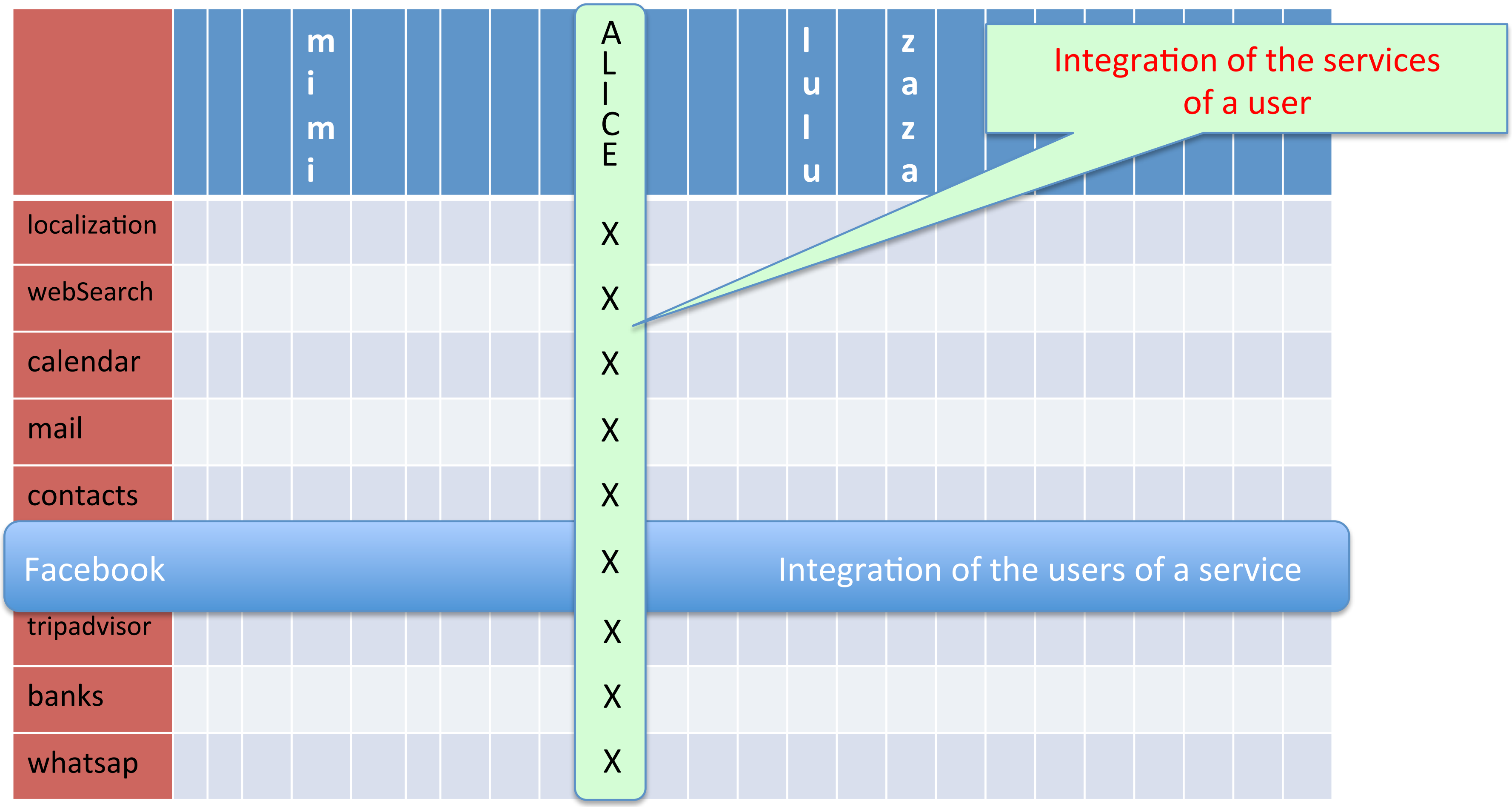
- on some unknown machine
- with your data
- with some unknown software

your PIMS

- on your machine
- with your data, possibly a replica of the data from systems you like
- with your software, or with wrappers to external services

[S. Abiteboul, B. Andre, D. Kaplan; CACM 2015]
[S. Abiteboul, A. Marian, EDBT 2015]
[H. Haddadi *et al.*, CoRR abs/1501.04737 (2015)]

Horizontal vs. vertical data integration



Code verification

- Possible if open-source - otherwise auditing
- Specify properties that should be verified
- Verification based on static analysis, in the spirit of theorem proving
- Lots of work in different areas
  - security, safety, optimization, privacy
- Little on responsibility

Provenance & distributed access control

Provenance specifies the origin of the data and the processing that has been performed

- Provenance [Green *et al.*, PODS 2007], [Green *et al.*, SIGMOD 2007]
- Common for scientific data, essential for verifying that data collection and analysis were performed responsibly
- Provenance and privacy [Davidson *et al.*, ICDT 2011]
- Managing access in the distributed setting, e.g., Webdamlog [Moffitt *et al.*, SIGMOD 2015; Abiteboul *et al.*, ICDT 2016], social networks: [Cheng *et al.*, PASSAT 2012; Hu *et al.*, TKDE 2013]

Roadmap

- ✓ Introduction
- Properties of responsible data analysis
  - ✓ Fairness
  - ✓ Diversity
  - ✓ Transparency
  - ➡ Neutrality
- Conclusion: towards a data responsible society

Google antitrust case

theguardian

European commission announces antitrust charges against Google

Inquiry will focus on accusations that internet search and tech multinational has unfairly used its products to oust competitors

Sam Thielman in New York
@samthielman
Wednesday 15 April 2015 07.27 EDT

Ruth Porat replaces Patrick Pichette as Google's chief finance officer. Photograph: Georges Gobet/AFP/Getty Images

The European Union accused Google on Wednesday of cheating competitors by distorting Internet search results in favour of its Google Shopping service and also launched an antitrust probe into its Android mobile operating system.

Facebook “like” button

REUTERS EDITION: U.S.

Technology | Wed Mar 9, 2016 1:22pm EST

Related: TECH, FACEBOOK, REGULATORY NEWS, BREAKINGVIEWS

German court rules against use of Facebook "like" button

FRANKFURT

A German court has ruled against an online shopping site's use of Facebook's "like" button on Wednesday, dealing a further legal blow to the world's biggest social network in Germany.

The Duesseldorf district court said that retailer Peek & Cloppenburg failed to obtain proper consent before transmitting its users' computer identities to Facebook, violating Germany's data protection law and giving the retailer a commercial advantage.

The court found in favor of the North Rhine-Westphalia Consumer Association, which had complained that Peek & Cloppenburg's Fashion ID website had grabbed user data and sent it to Facebook before shoppers had decided whether to click on the "like" button or not.

Neutrality

- Net and platform neutrality (CNNum report)
  - **net neutrality** - the network is transporting data with no bias based on source, destination, content …
  - **platform neutrality** - big internet platforms should not discriminate in favor of their own services
- Related to fairness and diversity, verified with transparency tools

the rich get richer, the poor get poorer

Power comes with responsibility

power

A handful of big players command most of the world's computational resources and most of the data, including all of your personal data - an **oligopoly**

danger

can destroy business competition

control what information you receive

can guide your decisions

can infringe on your privacy and freedom

Roadmap

- ✓ Introduction
- ✓ Properties of responsible data analysis
  - ✓ Fairness
  - ✓ Diversity
  - ✓ Transparency
  - ✓ Neutrality
- ➡ Conclusion: towards a data responsible society

Technology is not the whole answer

Technology enables responsible data analysis: specify and verify

- But will companies simply feel compelled to act responsibly?
- Who sets the standards for what is ethical and legal?

Users and regulators!

- But they have to be educated

User organization

- Users are data, users are consumers of data, users have **tremendous power**!
- Example: Instagram 2012, gave FB (new owner) broad access to user data and photos for commercial use. Forced to change back under pressure from users.
- Limitations: user education, lack of proper tools

Public policy

- Should the government regulate the big data industry?
  - regulate
  - define good practices
  - evaluate responsibility
- Issues:
  - which government?
  - lack of competence, agility

US legal mechanisms

[Big Data: A tool for inclusion or exclusion? FTC Report; 2016]

https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

- Fair Credit Reporting Act - applies to consumer reporting agencies, must ensure correctness, access and ability to correct information
- Equal opportunity laws - prohibit discrimination based on race, color, religion, … - plaintiff must show disparate treatment / disparate impact
- FTC Act - prohibits unfair or deceptive acts or practices to companies engaged in data analytics

lots of gray areas, much work remains, enforcement is problematics since few auditing tools exist

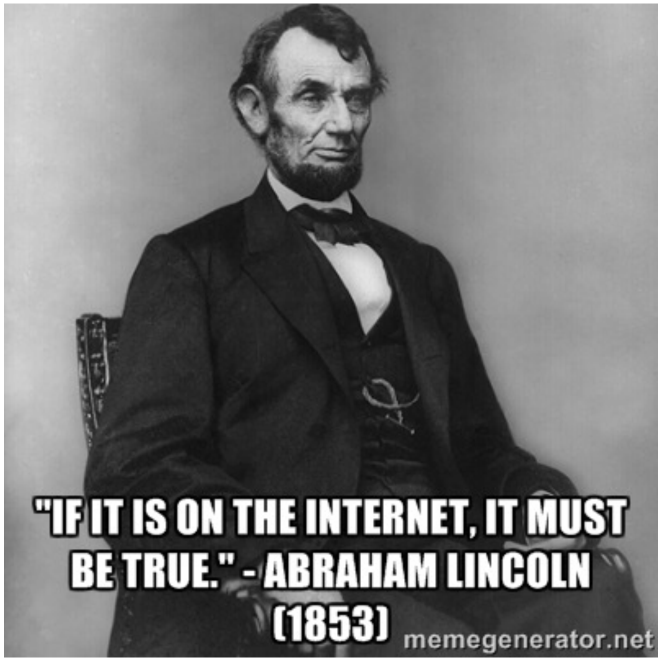
EU legal mechanisms

- **Transparency**
  - Open data policy: legislation on re-use of public sector information
  - Open access to research publications and data
- **Neutrality**
  - Net neutrality: a new law, but with some limitations
  - Platform neutrality: the first case against Google search
- Different countries are developing specific laws, e.g., portability agains user lock-in (France)

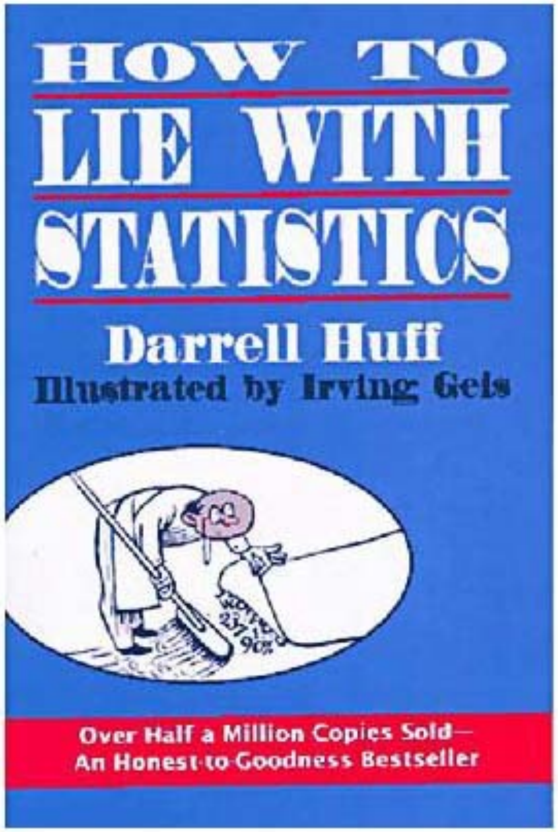
Education

- Concepts
  - **understanding** data acquisition methods and data analysis processes
  - **verifying** the data and the process: provenance, credit attribution, trust
  - **interpreting** results
- Tools: computer science, probability and statistics, what people need to know about **data science**!

learn to question!



Education: data literacy



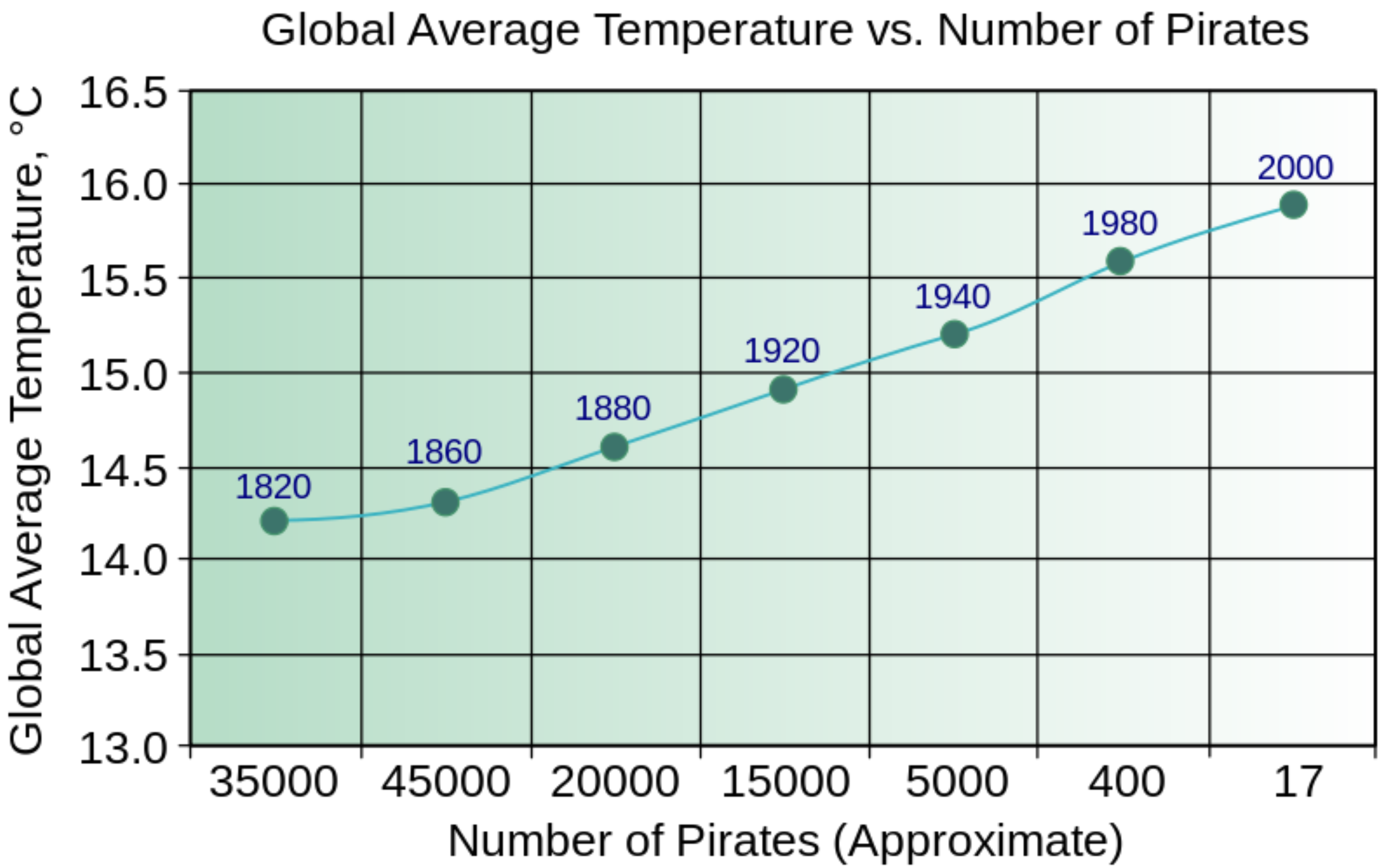
statistics

BIG DATA

Statistics scares people, big data REALLY scares people!

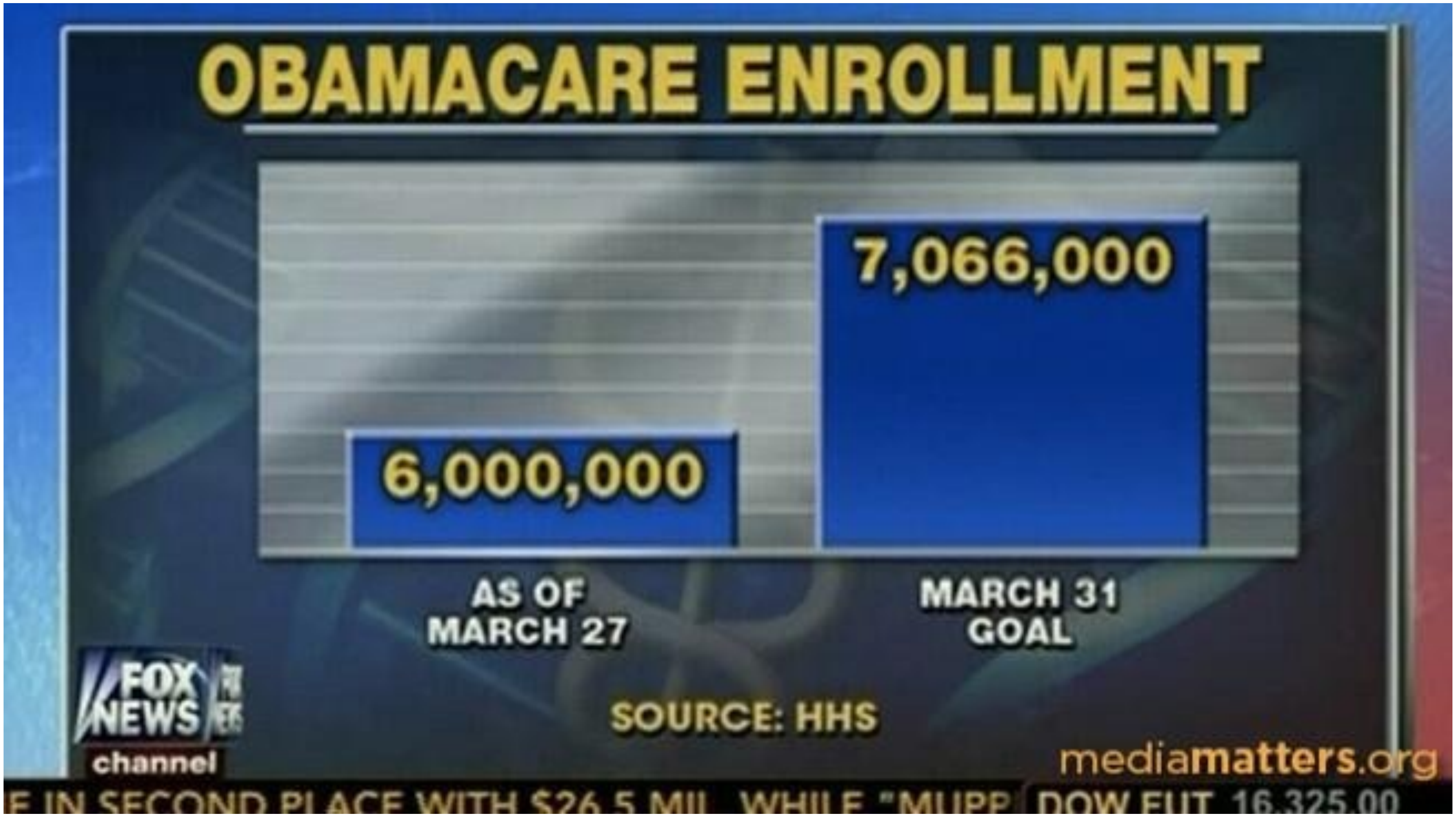
data RESPONSIBLY

Education: correlation, causation



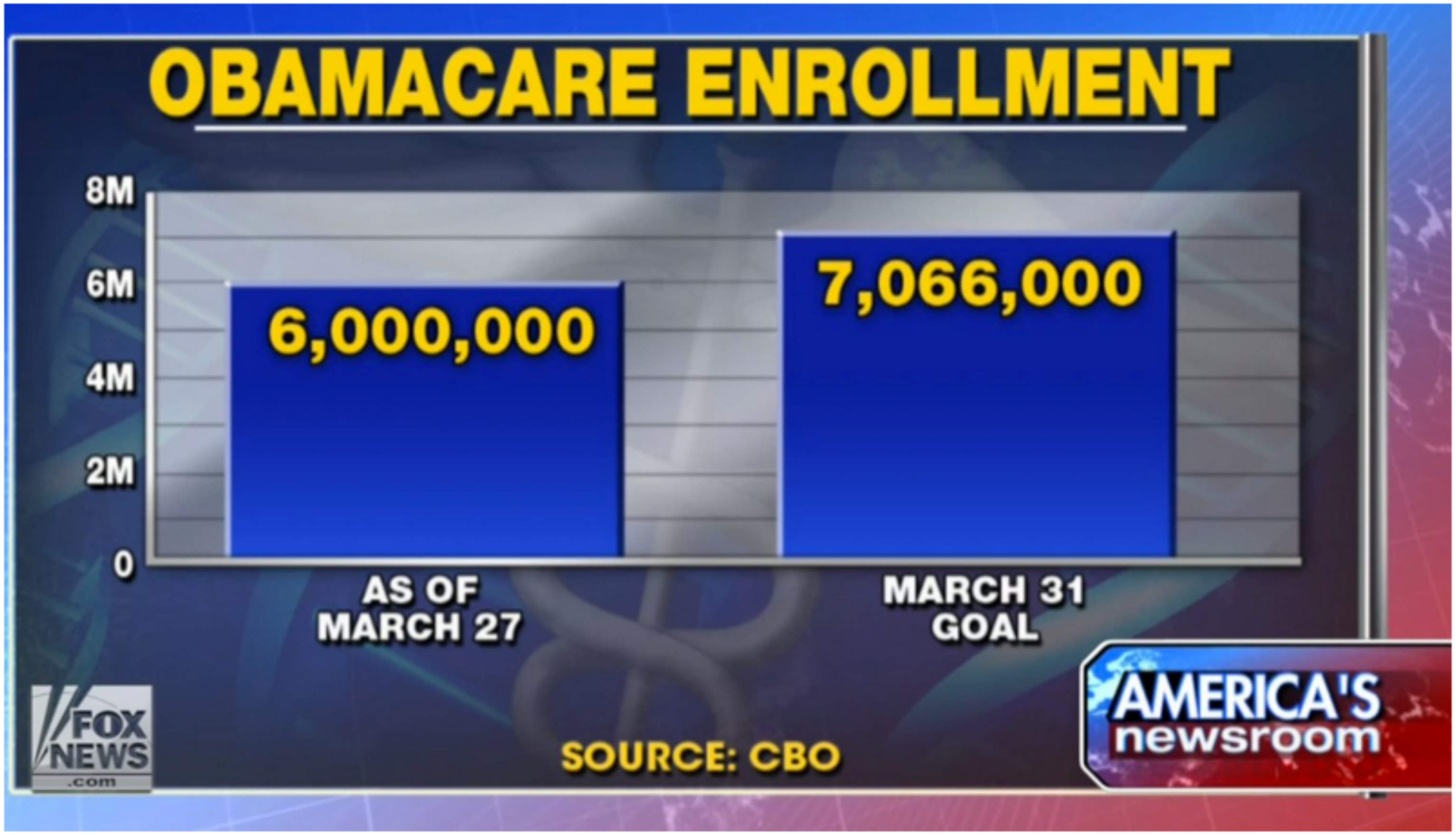
https://en.wikipedia.org/wiki/File%3aPiratesVsTemp%28en%29.svg

Education: data visualization



http://www.businessinsider.com/fox-news-obamacare-chart-2014-3

Education: data visualization



Fox News

http://www.businessinsider.com/fox-news-obamacare-chart-2014-3

Roadmap

- ✓ Introduction
- ✓ Properties of responsible data analysis
  - ✓ Fairness
  - ✓ Diversity
  - ✓ Transparency
  - ✓ Neutrality
- ✓ Conclusion: towards a data responsible society

Thank you!