Supported in part by NSF grant IIS 1250880

Moving Past the Wild West Era for Big Data

H. V. Jagadish
University of Michigan

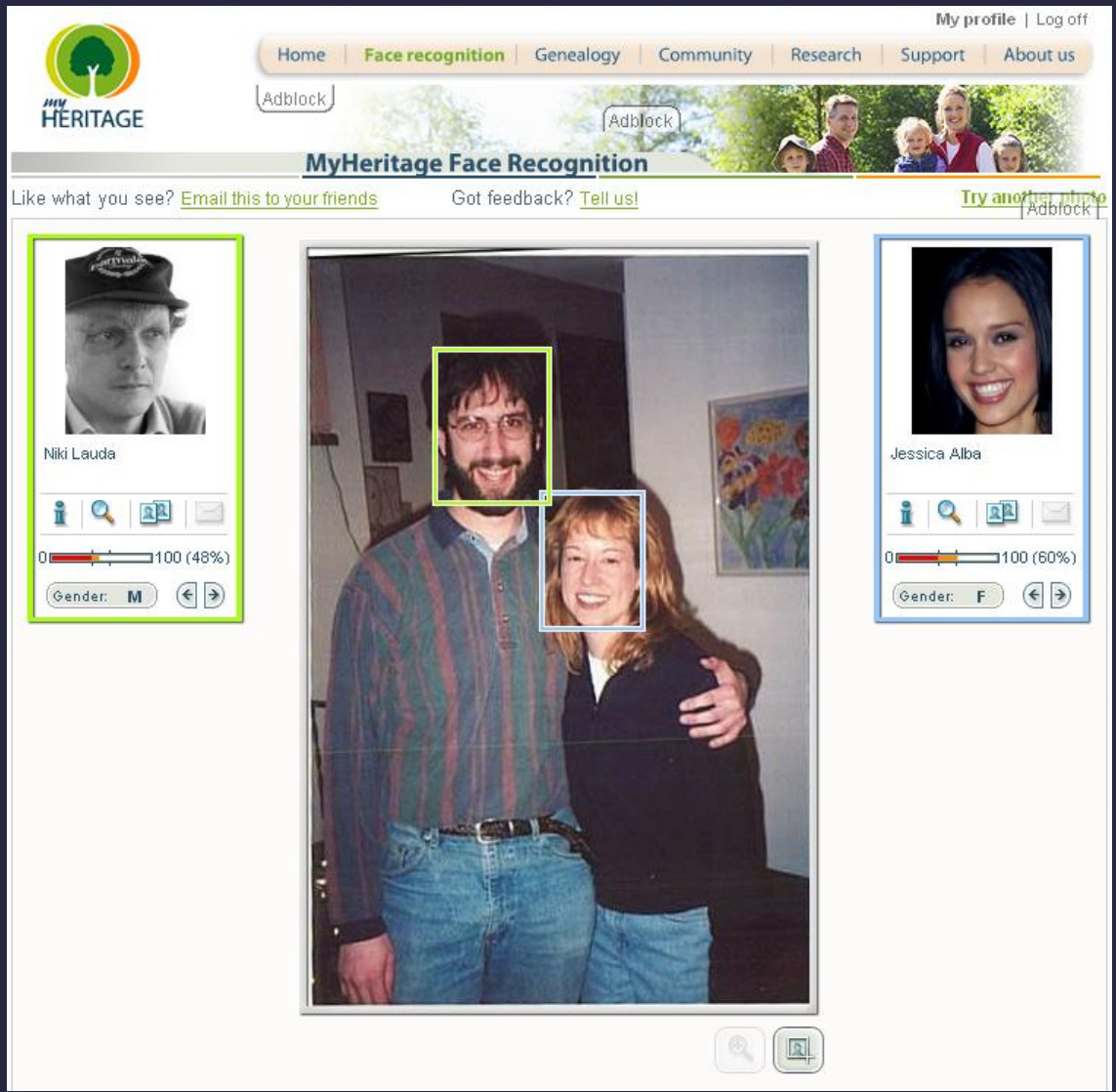
You take a photo at the beach

And Post it on Facebook

People who happened to be on the beach are in your photo. You do not know them, they are not tagged, and don’t even know about this photo you have made public

Face Recognition Software

With face recognition software, Don Juan and Lolita are identified in your photo.



Search Technology

- Lolita's spouse finds this photo and her affair with Don.
- Their marriage is on the rocks.

- Who is at fault?

Lolita and Don Juan?

Of course …

Chain of Steps

- You took photo
- You posted on Facebook
- I used face recognition software with Big Data
- Recognized name tags were indexed by a search engine

- Each step looks fine in isolation
- Or does it?

The Role of Big Data

- Coincidences could have happened in the past
  - But would have been very rare
- With Big Data, such exposure becomes routine

Big Data => No Affairs

And No Privacy

Do we understand the consequences?

Privacy

- Ability to control sharing of information about self.
- Basic human need.
  - Even for people who have “nothing to hide”

Loss of Privacy

- Due to loss of control over personal data.
- I am OK with you having certain data about me that I have chosen to share with you or that is public, but I really do not want you to share my data in ways that I do not approve.

No Option to Exit

- In the past, one could get a fresh start by:
  - Moving to a new place
  - Waiting till the past fades
    - Reputations can be rebuilt over time.
- Big Data is universal and never forgets anything!!
- Can we develop techniques to forget?

An Algorithmic Vicious Cycle

- Company has only 10% women employees.
- Company has “boys’ club culture” that makes it difficult for women to succeed.
- Hiring algorithm trained on current data, based on current employee success, scores women candidates lower.
- Company ends up hiring even fewer women.

Predictive Policing

- It will deter reckless driving if patrol cars regularly follow known bad drivers.
- How long will you sustain your clean driving record if you were ticketed for every infraction?

Alex Smith (Washington DC)

Ethics Distinguish Right from Wrong

- Not just for privacy, but also for
  - Fairness
  - Validity
- Big Data has a big impact on society.
- Need to make smart choices about right and wrong to shape the world we live in.

A Framework for Discussion

- We need a shared sense of what is right and what is wrong.
- This shared sensibility can only arise through discussion.
- This talk attempts to establish a framework for such discussion.

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?

The data scientist's code of ethics

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Is it Your Data?

- OK, it is about you.
  - But is it yours?

- If I write your biography, I own copyright. If you dislike what I say, not much you can do, except sue for libel where I am inaccurate.

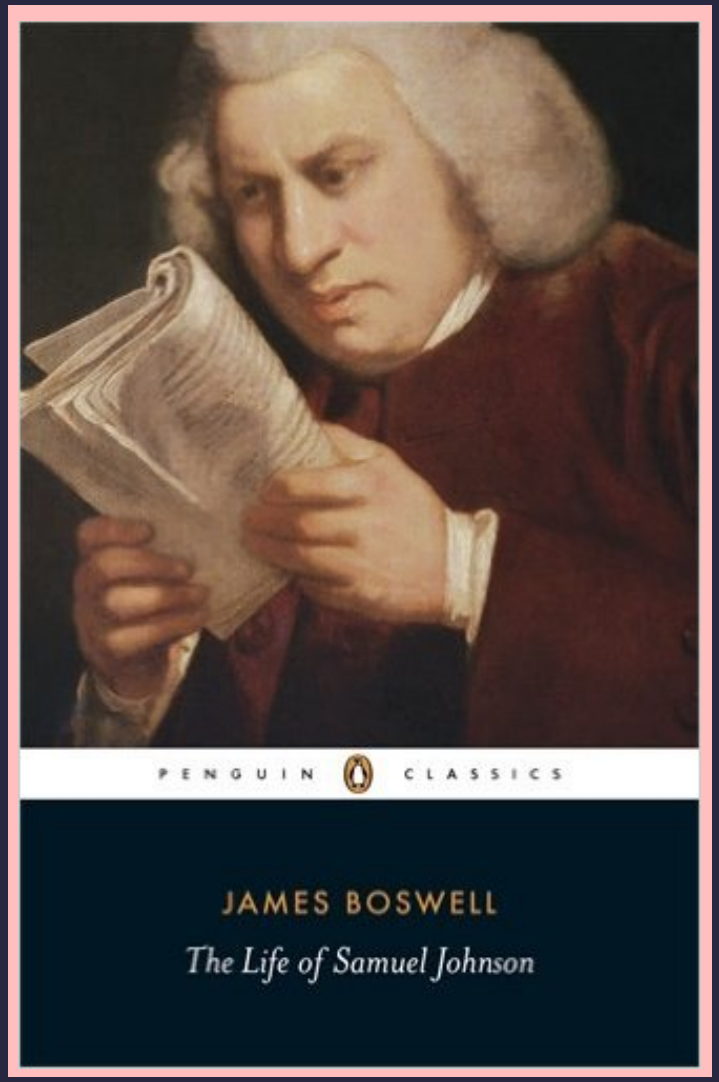


Photo Ownership

- If I photograph you, I own the photo.

Carny Portrait (Times Two)
George Norcus
cc-by-nc-2.0

Photo Ownership

- If I photograph you, I own the photo.
- Limits on me can be:
  - On taking the photo in certain private areas
    - In your home
    - In a bathroom in my store
  - On using the photo in certain ways
    - As implied endorsement
    - As implied libel

Data Ownership

- Similar limits on what data I can record about you and what I can do with it after
- Free to record and free to use otherwise.

- And we have done this forever …

Limits on Recording

- Recording is wrong when there is reasonable expectation of privacy
  - E.g. no cameras in clothing store fitting room

Walmart Fitting Room
Random Retail
cc-by-2.0

Limits on Recording

- Recording is wrong when there is reasonable expectation of privacy
  - E.g. no cameras in clothing store fitting room
- Similarly:
  - Phone company must not record the (content of) phone calls.
  - Email provider must not read emails.
    - Unless clearly agreed to.
  - Mobile apps must not record location except where they provide location-based service.

We Could Agree Otherwise

- You could pose for a photograph under an agreement that I will give you ownership of the photo.
- You could agree to participate in a research experiment under an agreement that has *informed consent*.

Informed Consent

- Human Subject must be
  - Informed about the experiment
  - Must consent to the experiment
    - Voluntarily, without coercion
  - Must have the right to withdraw consent at any time

Wrong Question??

- “Informed” is based on something hidden in multiple pages of fine print.
  - Unclear value in law
- “Voluntary” in spite of the requirement of consent to obtain a desired action at the time of the action
  - E.g. to use a software service
  - E.g. to buy product.

Prospective Data Collection

- Institutional Review Boards
- Driven by excesses of medical research
- Also applies to social science research involving human subjects
- Enforced by US federal law only for government funded experiments.
  - Some states go farther.
- Applies to user studies in CS!!

Human Subjects Research?

- Web Companies do A/B testing all the time. Human users are impacted.
- Companies try to “push our buttons” to buy more.
  - This is considered good business.
- Companies can have imperfect algorithms that give us poor results.

Human Subjects Research?

- But companies should not lie to us intentionally, even if effects are similar.
- Facebook/Cornell: PNAS paper on effects of altering user mood.
- OKCupid: Reported False Compatibility.

We Experiment On Human Beings!

July 28th, 2014 by Christian Rudder

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Video Cameras in Stores

- Can provide security
- Can even be used to improve placement
- But should not be published

Cell Phone Location Tracking

- Necessary to provide service.
- Required for many valuable applications.
- But can result in a huge loss of privacy.

Limits on Use

- Often there is a strong reason to record data, but also potential for misuse.
- Prefer to limit use rather than recording
- Allows desired legitimate use while disallowing other (undesirable) use.

Voluntary Limits on Use

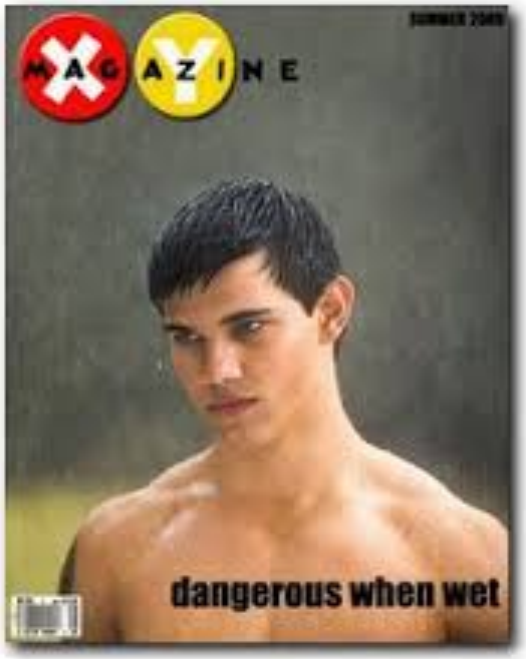
- Police reassure citizens that bodycam video will not be posted on the web.
- Businesses can reassure customers that data collected for one purpose will not be used for another.
- These assurances can
  - Have legal force
  - Remove barriers to many transactions.

Data Destruction

- Companies legitimately collect data as part of doing business.
- Companies need to retain goodwill of customers, and so will try not to do egregiously bad things with the data.
- Once the company ceases to do business (e.g. because of bankruptcy), this data is an asset that is likely sold to a third party who intends to misuse.
- Collected data must be destroyed, not sold.

Lawyers are Resolving This!!

radioshack



connectedu
EDUCATION & CAREER MANAGEMENT

BORDERS®

Repurposing is not all bad

- I share medical data with my hospital to get better care, but I may gladly support its repurposed use for medical research.
- I understand that my credit card company collects data about my purchases and payments, and I am willing to accept that they also share some of this with a credit reporting agency, even if I don’t particularly like it.

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Anonymity

NETFLIX PRIZE

Closeted Lesbian Sues Netflix For Potential Outing

By Laura Northrup on December 19, 2009 3:00 PM

Here's the problem with anonymized data: if it were truly anonymized, it wouldn't be useful to anyone for anything. With enough data about a person—say, their age, gender, and zip code—it's not hard to narrow down who someone is. That's the idea behind a class-action lawsuit against Netflix regarding the customer data they released to the public as part of the Netflix Prize project, a contest to help create better movie recommendations. A closeted lesbian alleges that the data available about her could reveal her identity.

Consumerist.com

Settled for $9 million after >2 years of litigation

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Anonymity is Impossible

- Anonymity is virtually impossible, with enough other data.
  - Diversity of entity sets can be eliminated through joining external data
  - Random perturbation works only if we can guarantee a one-time perturbation
  - Aggregation works only if there is no known structure among entities aggregated
- Faces can be recognized in image data.
  - Progressively, even under challenging conditions, such as partial occlusion

Limit Publication of Datasets

- If anonymity is not possible, the simplest way to prevent misuse is not to publish the dataset.
  - E.g. government agencies should not make public potentially sensitive data
- Yet access to data is crucial for many desirable purposes, including:
  - Medical research
  - Public watchdogs

License Data to Trusted Parties

- Need simple licensing regime for access to potentially sensitive data, including de-identified data.
- Enforce through contracts in the business world.
- Enforce through professional standards in the research world.
  - Investigator promises not to re-identify
  - Else, loses reputation and future access
  - Similar to double-blind review

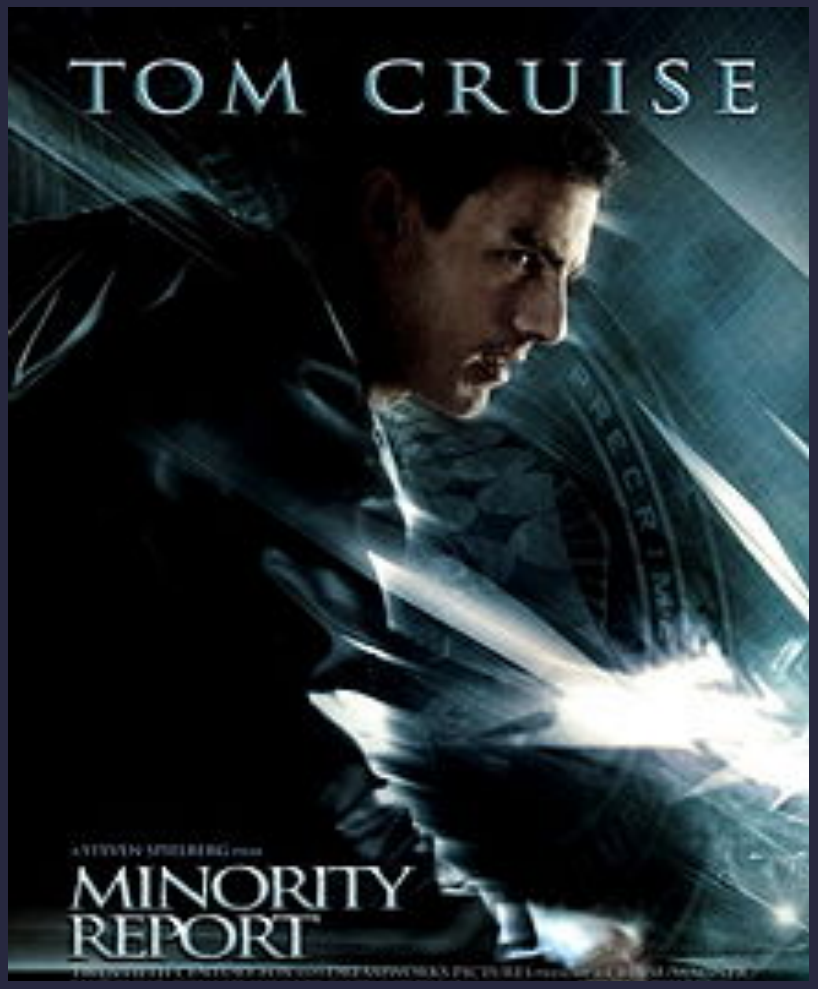
Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Predict a Crime??

- Preemptively arrest person with criminal intent?
- In general, NO.
- But possibly yes for terrorism?



Prediction is Probabilistic

- Probabilistic – only indicates likelihood.
- Suggests greater surveillance
  - More audits of tax returns for people more likely to have cheated.

401K-2012 (1040 Tax Return)
cc-by-sa-2.0

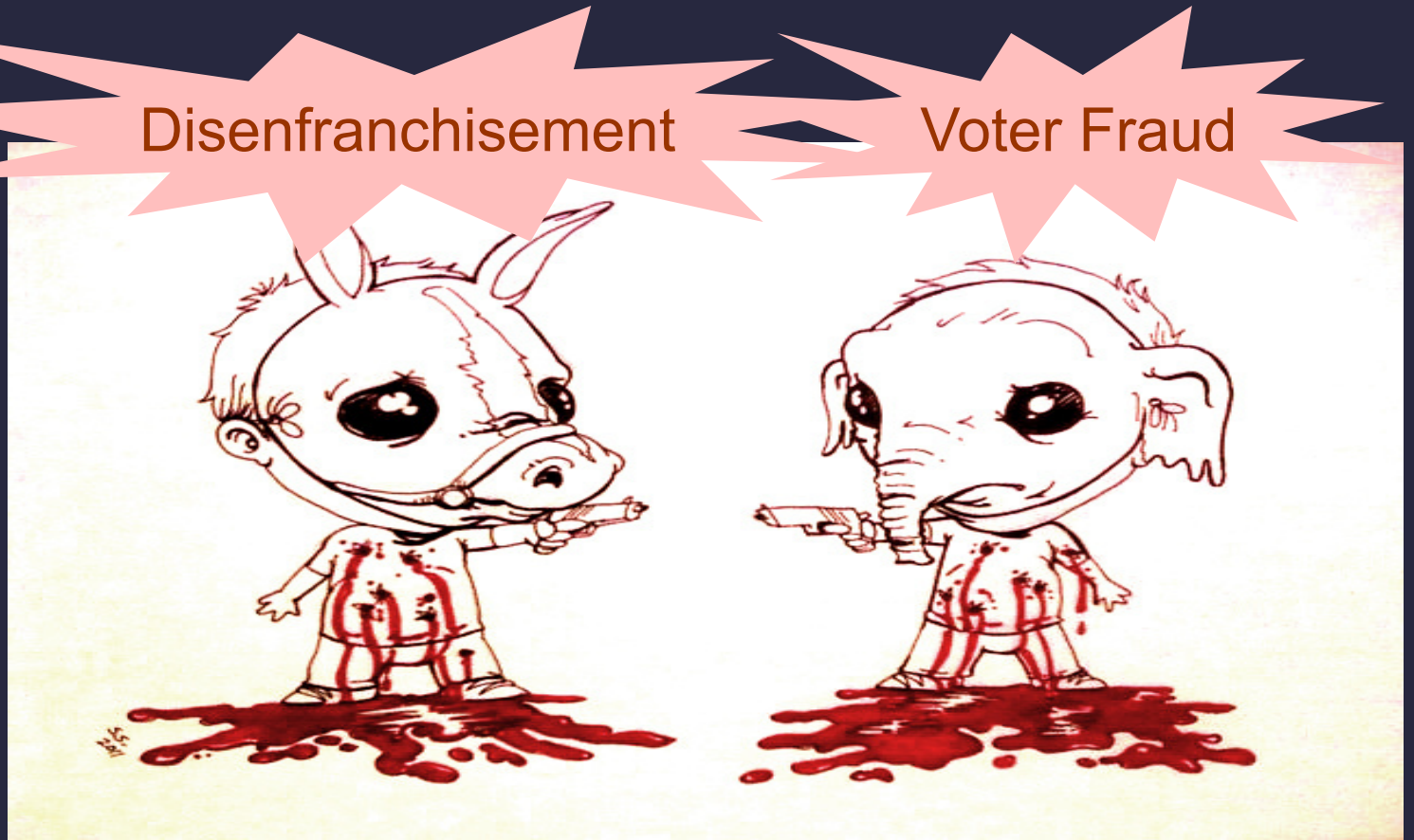
Predictive Policing

- Deployment of police forces to higher crime areas in greater numbers.
- Vicious cycle?
  - More surveillance can lead to more detected crime.
- Police car follows known bad drivers.

Alex Smith (Washington DC)

Discriminatory Intent

- Voter ID laws – vast majority of those impacted are likely to vote Democratic.



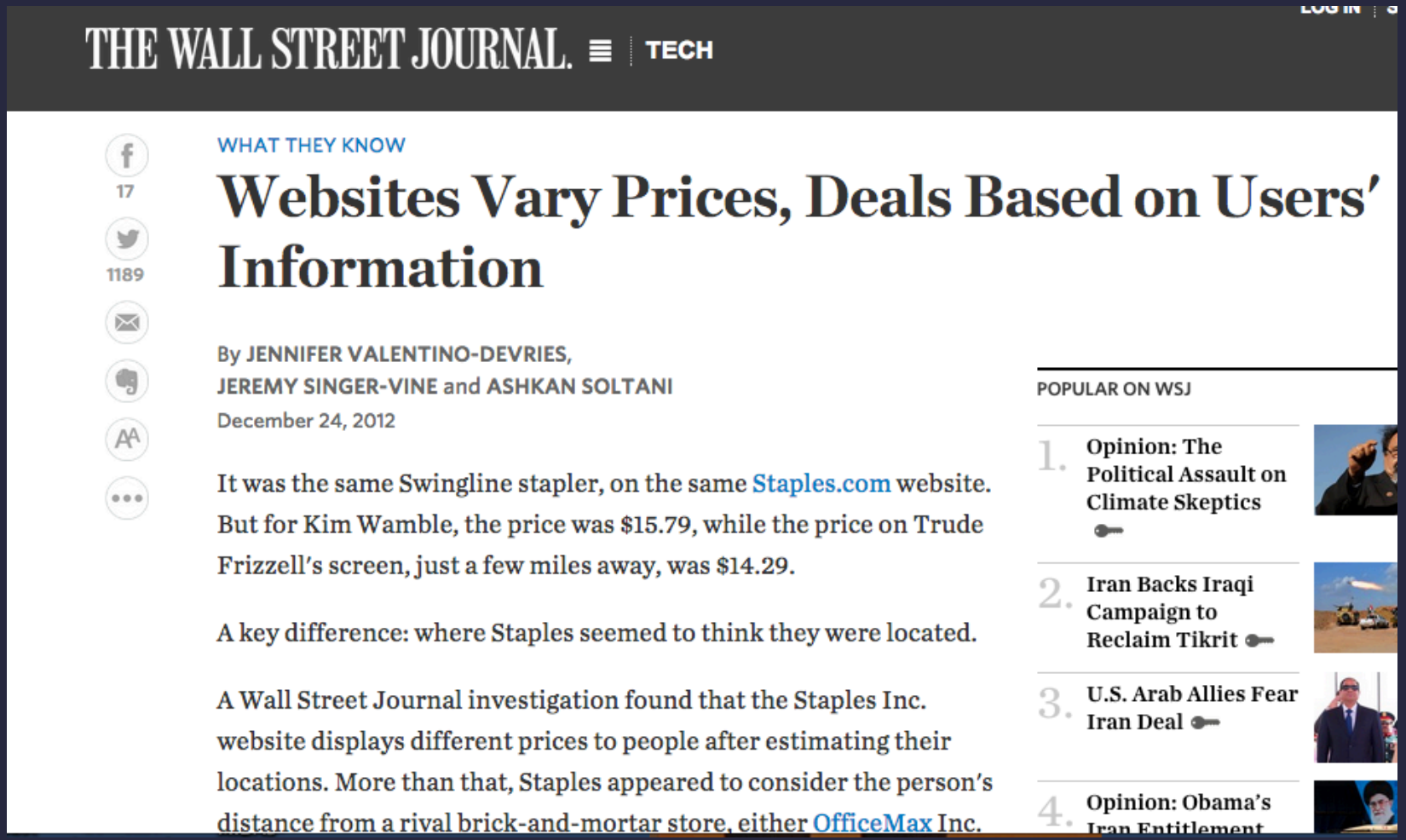
Racial Discrimination

- Universities prohibited by law from considering race in admissions
  - can find surrogate features that get them close, without violating the letter of the law
- Lender prohibited by law from redlining borrowers on account of race
  - can find surrogate features that get them close
- In general, proxy attributes can be found.
- Concept of “indirect discrimination”

Discriminatory Intent (contd.)

- Big Data provides the technology to facilitate such proxy discrimination.
- Whether this technology is used this way becomes a matter of intent.
- Big Data also provides the technology to detect and address discrimination.

Unintentional Discrimination



THE WALL STREET JOURNAL. ≡ | TECH

17

1189

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.

A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, Staples appeared to consider the person's distance from a rival brick-and-mortar store, either OfficeMax Inc.

POPULAR ON WSJ

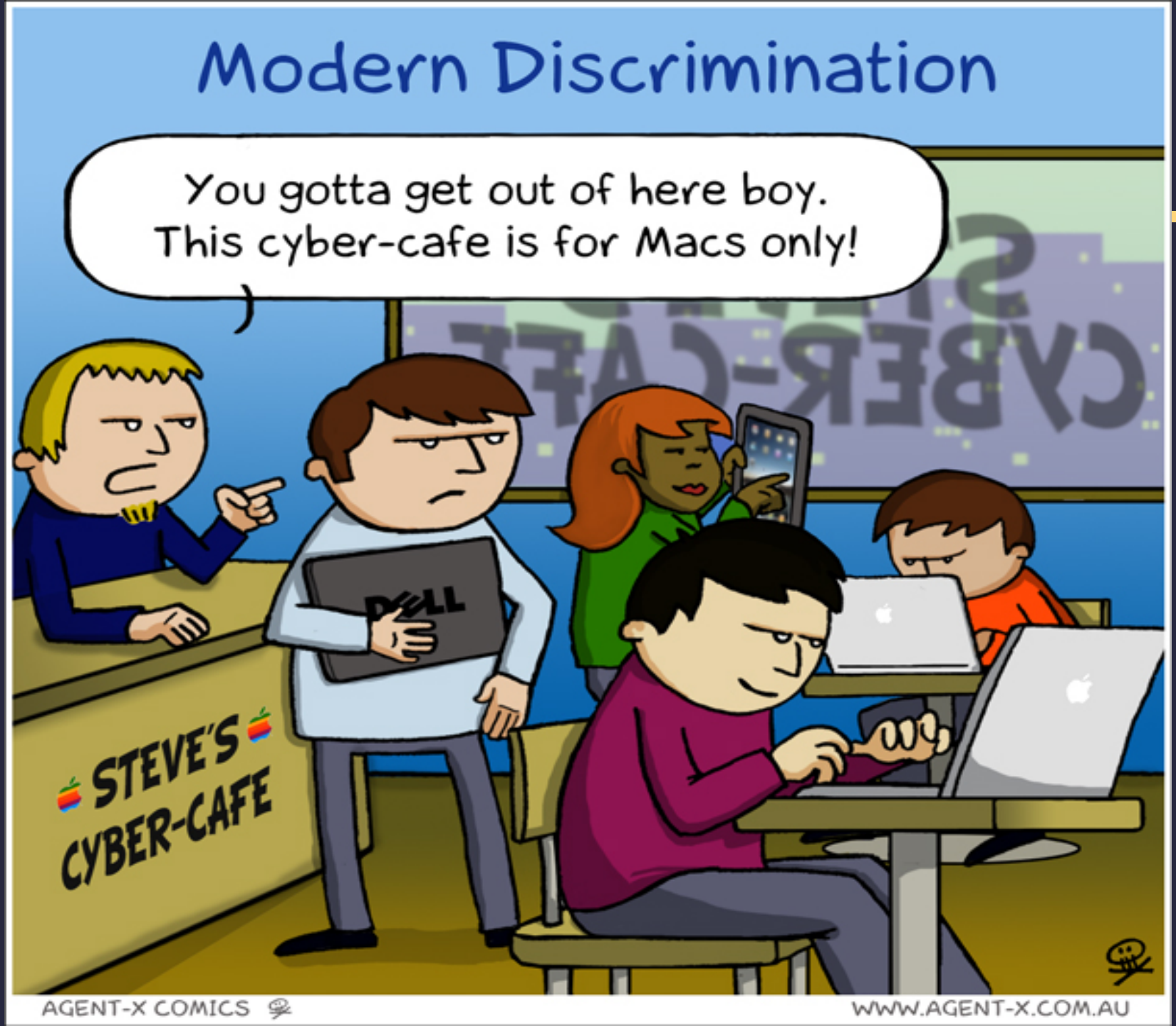
1. Opinion: The Political Assault on Climate Skeptics
2. Iran Backs Iraqi Campaign to Reclaim Tikrit
3. U.S. Arab Allies Fear Iran Deal
4. Opinion: Obama's Iran Entitlement

Unintentional Discrimination

- Staples offered lower prices online to customers who lived near a competitor.
- Staples charged higher prices online to poorer customers (who live in inner cities or rural areas).

An Algorithmic Vicious Cycle

- Company has only 10% women employees.
- Company has “boys’ club culture” that makes it difficult for women to succeed.
- Hiring algorithm trained on current data, based on current employee success, scores women candidates lower.
- Company ends up hiring even fewer women.



Algorithmic Fairness

- Do the data “speak for themselves”?
- Can algorithms be biased?
- Can we make algorithms unbiased?
  - Is training data set representative of the population?
  - Is past population representative of future population?
  - Are observed correlations due to confounding processes?

Algorithmic Fairness

- Humans have many biases.
  - No human is perfectly fair, even with the best of intentions.
- Biases in algorithms usually easier to measure, even if outcome is no fairer.
- Mathematical definitions of fairness can be applied, proving fairness, at least within the scope of the assumptions.

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Validity

- Bad data leads to bad decisions.
- But most data are dirty.
- If decision-making is opaque, results can be bad in the aggregate, and catastrophic for an individual.
- What if someone has a loan denied because of an error in the data analyzed?

Third Party Data

- Material decisions can often be made on the basis of public data or data provided by third parties.
- There often are errors in these data.
- Does the affected subject have a mechanism to correct errors?
  - Credit rating data on steroids.
- Does the affected subject even know what data were used?

p-Hacking

- Multiple hypothesis testing
- For a given **single** hypothesis, a p-value of 0.05 says that there is only a 5% probability of observing values by chance, without the hypothesis being true.
- What if you test 100 independent hypotheses?
- What if you devised your hypothesis to fit the observed data?

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Information Is Power

- Technology can be the great leveler
- Or a force for greater inequity

- We need to think through societal implications of each new algorithm or analysis or system.
- The engaged data scientist will discuss these issues with decision-makers.
- Cannot abdicate responsibility.

Known Solutions?

- Many of these problems are known
- Many even have proposed solutions
  - Separate test data from training data to avoid p-hacking
  - Use privacy-preserving data mining algorithms
  - Use homomorphic encryption to avoid sharing data.
  - New theory on fairness of algorithms

When the West Was Wild

There were few rules

Buffalo Bill and his Wild West Show

You Took What You Could

Civilization is Based on Rules

We can no longer do everything we can.

We have to follow rules

Need Data Ethics

- Ethics are principles that help us distinguish right from wrong.
- Ethics are the cornerstone of civilization.
- Ethics are the basis for the rules we all voluntarily choose to follow because that makes the world a better place for all of us.

Need Data Ethics

- Ethical principles stop me from stealing your wallet,
  - Even if you are smaller and weaker
  - Even if there is no chance I will be caught.

Ethics are Not Laws

- Suppose you tell me a secret and I promise to tell no one. If I then break my promise, I may not have broken any law, but I have certainly been unethical.

Ethics => Laws

- Ethics guide the creation of laws, so the two are often in consonance.
- Not everyone will be ethical.
  - That is why people go to jail for theft
  - That is why we still have to fight spam
    - But today, in the US, no upstanding business will own up to spamming intentionally.

Professional Codes

- Some professions have established codes of conduct.
  - E.g. Hippocratic oath in medicine
    - “First, do no harm”
  - E.g. Lawyers’ oath on admission to the bar.
- Need a similar code for Data Scientists.

We Should Own Our Destiny

- I am proud of being a data scientist and excited about the good that Big Data can do.
- If we do not self-regulate, there will be a public backlash that sets us back in ways that would hurt us and keep us from realizing the potential of Big Data.
- We need to act ethically so that we can continue to be proud and successful.

Framework: Issues to Consider

- Who owns the data?
- What can the data be used for?
- What can you hide in exposed data?
- Is the data analysis fair?
- Is the data analysis valid?
- What are the societal consequences?

The data scientist's code of ethics

Jagadish's Code of Ethics

1. Do not surprise
   - Do not surprise the data subject with what data (about the subject) you collected, shared, or used.
   - It is fine to surprise others, not subjects.
2. Own the outcomes
   - If the process leads to undesirable outcomes, work to modify the process even if there is nothing "wrong" with it.

Jagadish's Code of Ethics

1. Do not surprise
   - Who owns the data?
   - What can the data be used for?
   - What can you hide in exposed data?
2. Own the outcomes
   - Is the data analysis fair?
   - Is the data analysis valid?
   - What are the societal consequences?

Discussion

- I would love to talk more.

- Please visit

http://www.bigdatadialog.com