Fairness, Accountability, and Transparency while Mining Data from the Web

Wagner Meira Jr.[1]

[1]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

October 24, 2017

An innacurate timeline of the Web

1991 digital library

1995 commercial web

1997 search engines

1999 e-commerce

1999 semantic web

2000 agents

2003 mobile

2005 web 2.0

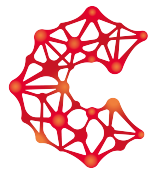
2007 social networks

2010 streaming dominates traffic

2011 IoT

2013 smart services

2016 fake news and misinformation



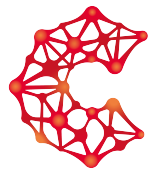
Mining Data from the Web and Social Networks

Mining web data used to deal with just:

- Dynamic behavior
- Complex relationships
- Heterogeneous data
- Incomplete information
- Noisy data
- Lack of scalability

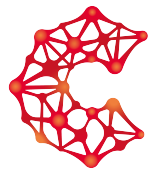
Mining Data from the Web and Social Networks

1. Data source
2. Data collection
3. Data processing
4. Data analysis, mining, and learning
5. Evaluation



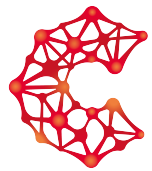
Mining Data from the Web and Social Networks

1. Data source
   - Functional: platform design and features shape user behavior
   - Normative: norms vary across platforms, communities and contexts
   - External: cultural elements and social contexts are reflected in data
   - Non-individuals: actions by organizations or bots
2. Data collection
3. Data processing
4. Data analysis, mining, and learning
5. Evaluation



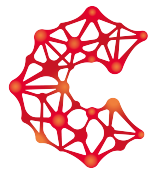
Mining Data from the Web and Social Networks

1. Data source
2. Data collection
   - Source selection: restricts the observations we make
   - Acquisition: API limits and opaque sampling strategies
   - Querying: limited expressiveness regarding information needs
   - Filtering: removal of apparently irrelevant data
3. Data processing
4. Data analysis, mining, and learning
5. Evaluation



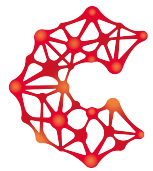
Mining Data from the Web and Social Networks

1. Data source
2. Data collection
3. Data processing
   - Cleaning: noisy and default values may introduce bias
   - Enrichment: manual or automated annotation are error-prone
   - Aggregation: information may be lost or amplified during aggregation
4. Data analysis, mining, and learning
5. Evaluation



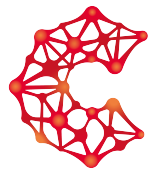
Mining Data from the Web and Social Networks

1. Data source
2. Data collection
3. Data processing
4. Data analysis, mining, and learning
   - Qualitative analysis: hard to generalize and sensitive to interpretation bias
   - Descriptive mining: sensitive to bias, confounders and randomness
   - Predictive mining: same data may generate different outcomes depending on target variable and representation
   - Observational studies: datasets are always incomplete and peer effects are hard to handle
5. Evaluation



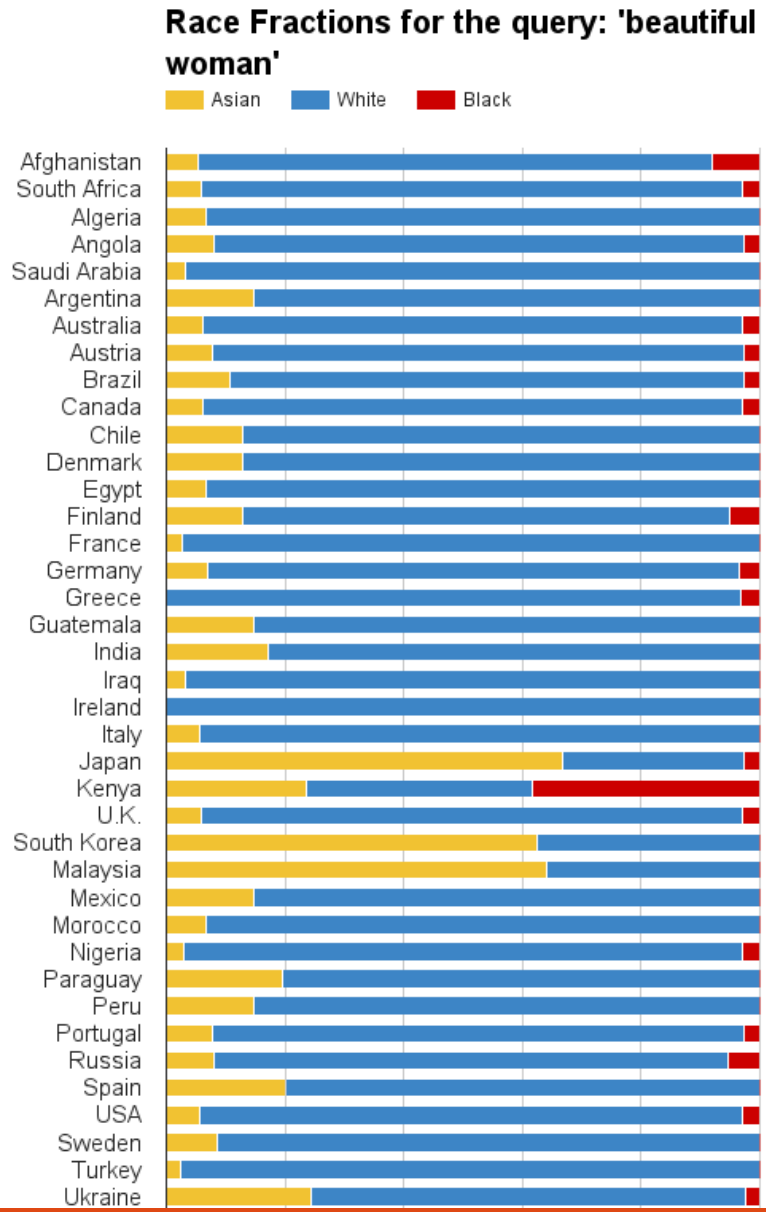
Mining Data from the Web and Social Networks

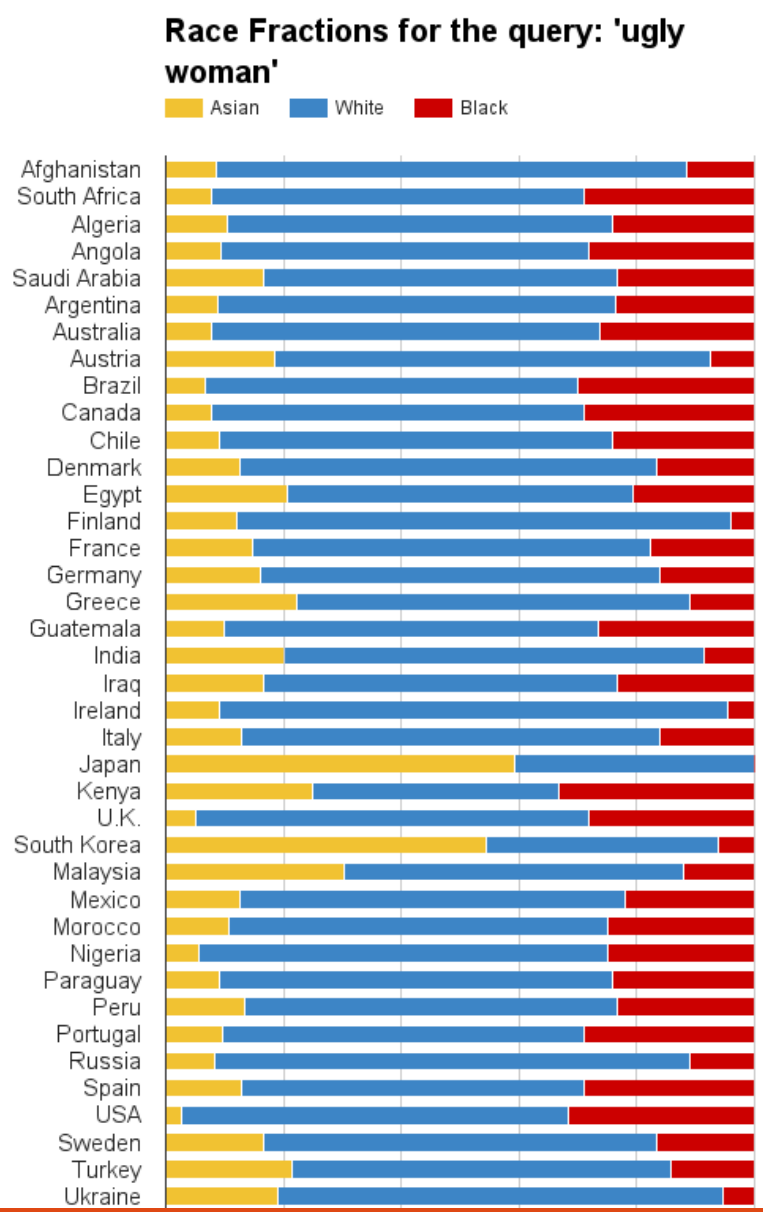
1. Data source
2. Data collection
3. Data processing
4. Data analysis, mining, and learning
5. Evaluation
   - Metrics: domain-specific indicators are rarely used.
   - Interpretation: data meaning changes with context and analyses should go beyond a single dataset of method.
   - Disclaimers: negative results are overlooked.



Stereotypes in the perception of physical attractiveness

Do search engines discriminate? (Socinfo'16)

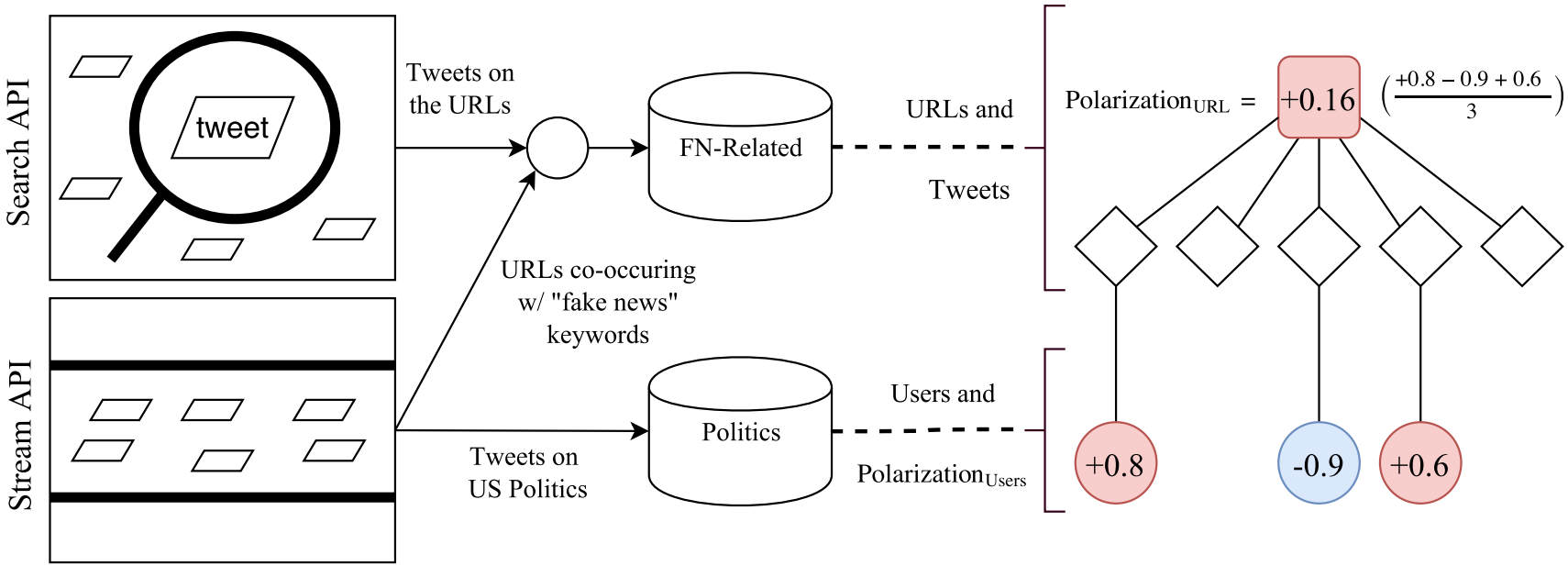




INCT CYBER

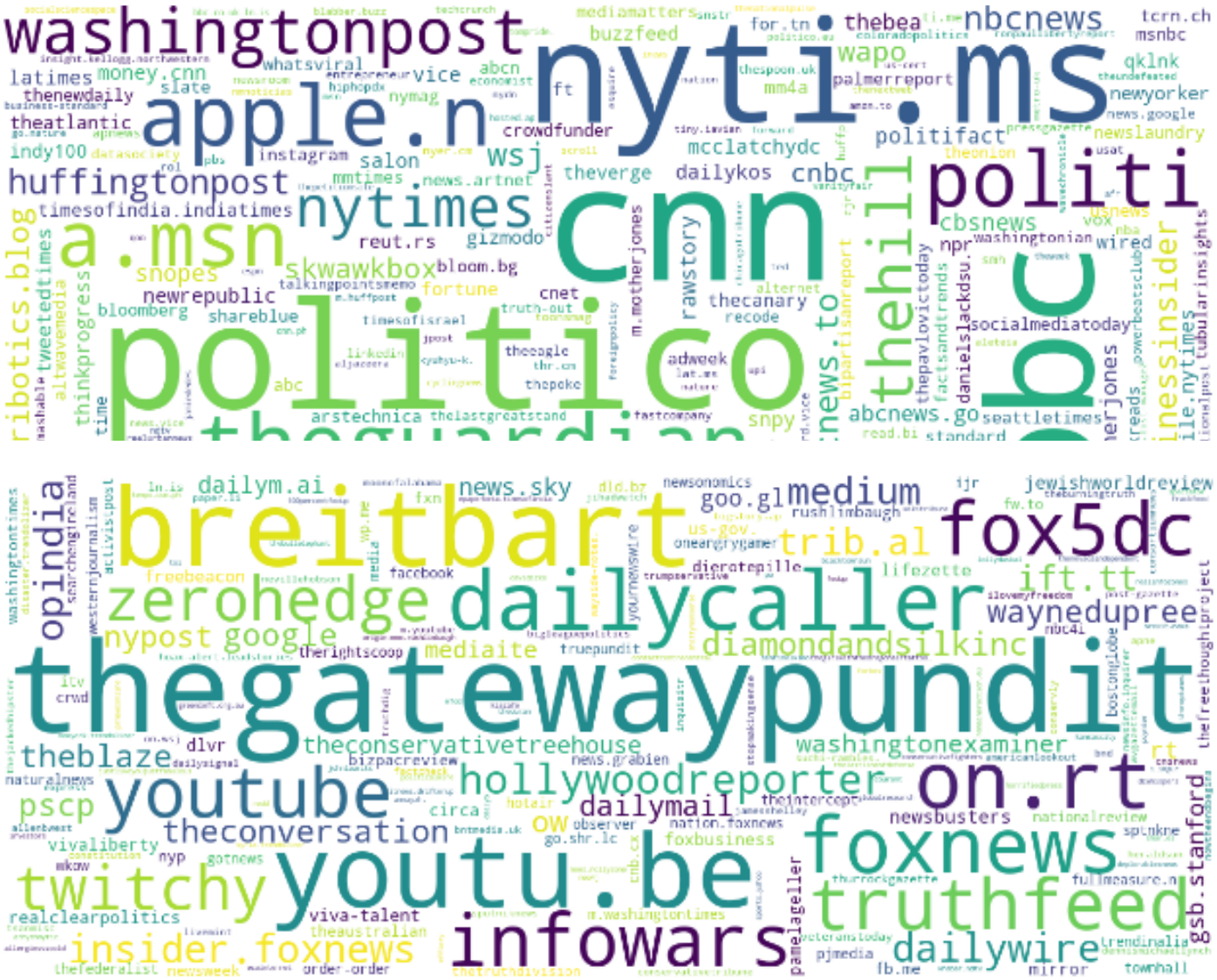
Everything I disagree with is #Fakenews

Does political polarization and spread of misinformation correlate? (DS+J, KDD'17)



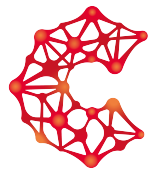
Everything I disagree with is #Fakenews

Does political polarization and spread of misinformation correlate? (DS+J, KDD'17)



We now face three additional requirements:

- Fairness
- Accountability
- Transparency

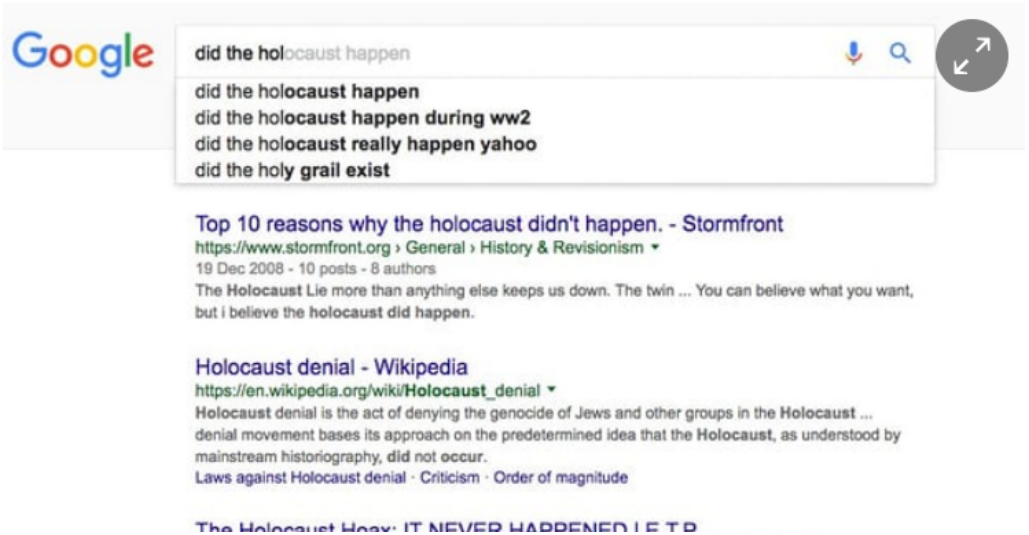


Fairness?

Google is not 'just' a platform. It frames, shapes and distorts how we see the world

Carole Cadwalladr

Last week, we reported how extremist sites 'game' the search engine, boosting their propaganda. In response, the web giant appears to have modified some results, but would like us not to notice



Fairness?

MOTHERBOARD It's Our Fault That AI Thinks White Names Are More 'Pleasant' Than ... SHARE TWEET

It's Our Fault That AI Thinks White Names Are More 'Pleasant' Than Black Names

JORDAN PEARSON
Aug 26 2016, 10:00am

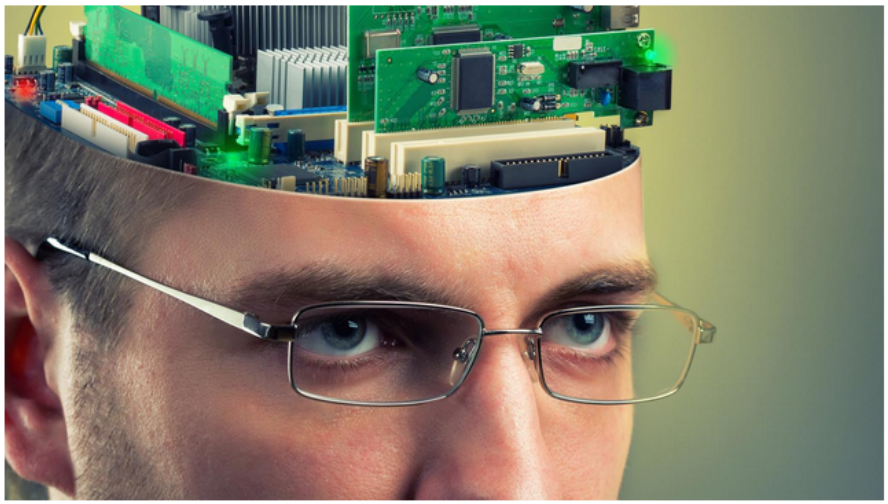


Image: Shutterstock

ADVERTISEMENT

LATAM AIRLINES

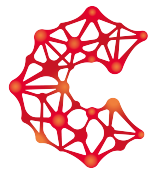
Ofertas imperdíveis para você descobrir o melhor do mundo.

Transferindo dados de dt.adsafeprotected.com...

e unsettling trends in our

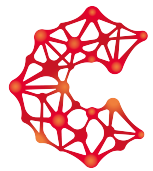
Discrimination

- To discriminate is to treat someone differently
  (Unfair) discrimination is based on group membership, not individual merit
- People's decisions include objective and subjective elements
  Hence, they can be discriminate
- Algorithmic inputs include only objective elements
  Hence, can they discriminate?



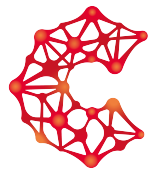
Data mining assumptions might not hold

- Data mining assumptions are not always observed in reality
  - Variables might not be independently identically distributed
  - Samples might be biased
  - Labels might be incorrect
- Errors might be concentrated in a particular class
- Sometimes, we might be seeking more simplicity than what is possible



Main concerns: data and algorithms

- Data inputs:
  - Poorly selected (e.g., observe only car trips, not bicycle trips)
  - Incomplete, incorrect, or outdated
  - Selected with bias (e.g., smartphone users)
  - Perpetuating and promoting historical biases (e.g., hiring people that "fit the culture")
- Algorithmic processing:
  - Poorly designed matching systems
  - Personalization and recommendation services that narrow instead of expand user options
  - Decision making systems that assume correlation implies causation
  - Algorithms that do not compensate for datasets that disproportionately represent populations
  - Output models that are hard to understand or explain hinder detection and mitigation of bias



Fairness-aware data mining

Goal: Develop a non-discriminatory decision-making process while preserving as much as possible the quality of the decision.
Steps:

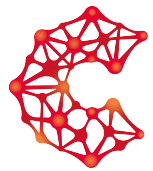
1. Defining anti-discrimination/fairness constraints
2. Transforming data/algorithm/model to satisfy the constraints
3. Measuring data/model utility

Fairness-aware data mining

Pre-processing: input data transformations to minimize discrimination while accuracy is maximized (e.g., supression, massaging, reweighing, sampling).

In-processing: novel algorithms that achieve the same goal (e.g., change split criterion and leaf relabeling in decision trees). A classifier is fair if it is not affected by the presence of sensitive data in the training set.

Post-processing: output models should not discriminate, how to clean the traces of discrimination (e.g., pattern sanitization, which is similar to anonymization).



Transparency?

SECTIONS HOME SEARCH The New York Times SUBSCRIBE NOW LOG IN

The Upshot

FOLLOW US:
GET THE UPSHOT IN YOUR INBOX

DANGER ZONE

Inside the Algorithm That Tries to Predict Gun Violence in Chicago

By JEFF ASHER and ROB ARTHUR JUNE 13, 2017

35

Gun violence in Chicago has surged since late 2015, and much of the news media attention on how the city plans to address this problem has focused on the Strategic Subject List, or S.S.L.

The list is made by an algorithm that tries to predict who is most likely to be involved in a shooting, either as perpetrator or victim. The algorithm is not public, but the city has now placed a version of the list — without names — online through its open data portal, making it possible for the first time to see how Chicago evaluates risk.

We analyzed that information and found that the assigned risk scores — and what characteristics go into them — are sometimes at odds with the Chicago Police Department's public statements and cut against some common perceptions.

■ Violence in the city is less concentrated at the top — among a group of about 1,400 people with the highest risk scores — than some public comments from the Chicago police have suggested.

RELATED COVERAGE

Opinion | Op-Ed Contributor
When a Computer Program Keeps You in Jail JUNE 13, 2017

Drug Deaths in America Are Rising Faster Than Ever JUNE 5, 2017

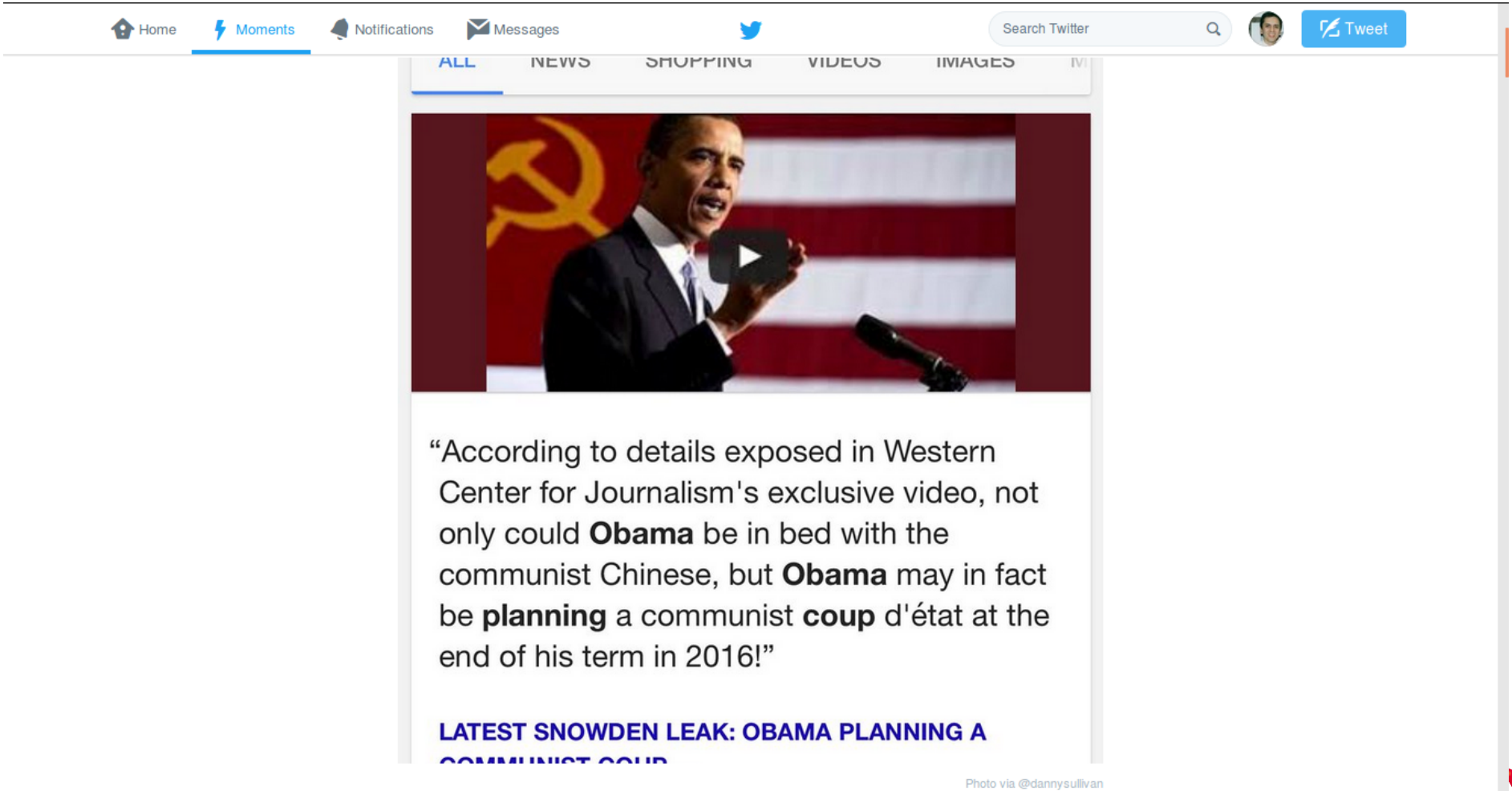
UNEQUAL JUSTICE
This small Indiana county sends more people to prison than San Francisco and Durham, N.C., combined. Why? SEPT. 2, 2016

Your Rabbi? Probably a Democrat. Your Baptist Pastor? Probably a Republican. Your Priest? Who Knows. JUNE 12, 2017

TAKING A TOLL
How Prejudice Can Harm Your Health

INCT CYBER

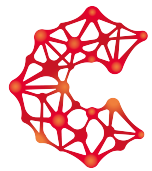
Transparency?



Transparency

Transparency may imply, in a broader sense, model interpretability:

- **Trust:** Confidence that a model will perform well. More specifically, not only how often it performs well, but also for which cases.
- **Causality:** To what extent may we generalize associations to infer properties?
- **Transferability:** Capacity of transferring learned skills to unfamiliar situations.
- **Informativeness:** How actionable is the pattern or model?
- **Fair and Ethical-Decision Making:** Are the models fair? Do they follow ethical patterns?



Properties

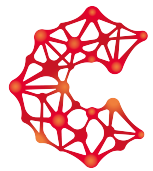
- **Transparency:** How does the model work?
- **Post-hoc explanations:** What else can the model tell me?

Transparency

- **Simulatability:** A human should be able to take the input data together with the parameters and, in reasonable time, *compute* the model.
- **Decomposability:** Each part of the model (input, parameter, calculation) admits an intuitive explanation.
- **Algorithmic transparency:** We should be able to understand how the model was built, i.e., its principles, capabilities and limitations.

Post-hoc interpretability

- **Text explanations:** Build an additional model that explains textually the outputs of a primal model.
- **Visualizations:** Render visualizations of the model and its outputs to ease understanding and usage.
- **Local explanations:** Zoom in the search space associated with input data and build a *local* model.
- **Explanation by example:** Report which training samples resemble the input data.



Accountability?

Tim Berners-Lee calls for tighter regulation of online political advertising

Inventor of the worldwide web described in an open letter how it has become a sophisticated and targeted industry, drawing on huge pools of personal data

Tim Berners-Lee: 'Targeted advertising allows a campaign to say completely different, possibly conflicting things to different groups. Is that democratic?' Photograph: Bloomberg/Bloomberg via Getty Images

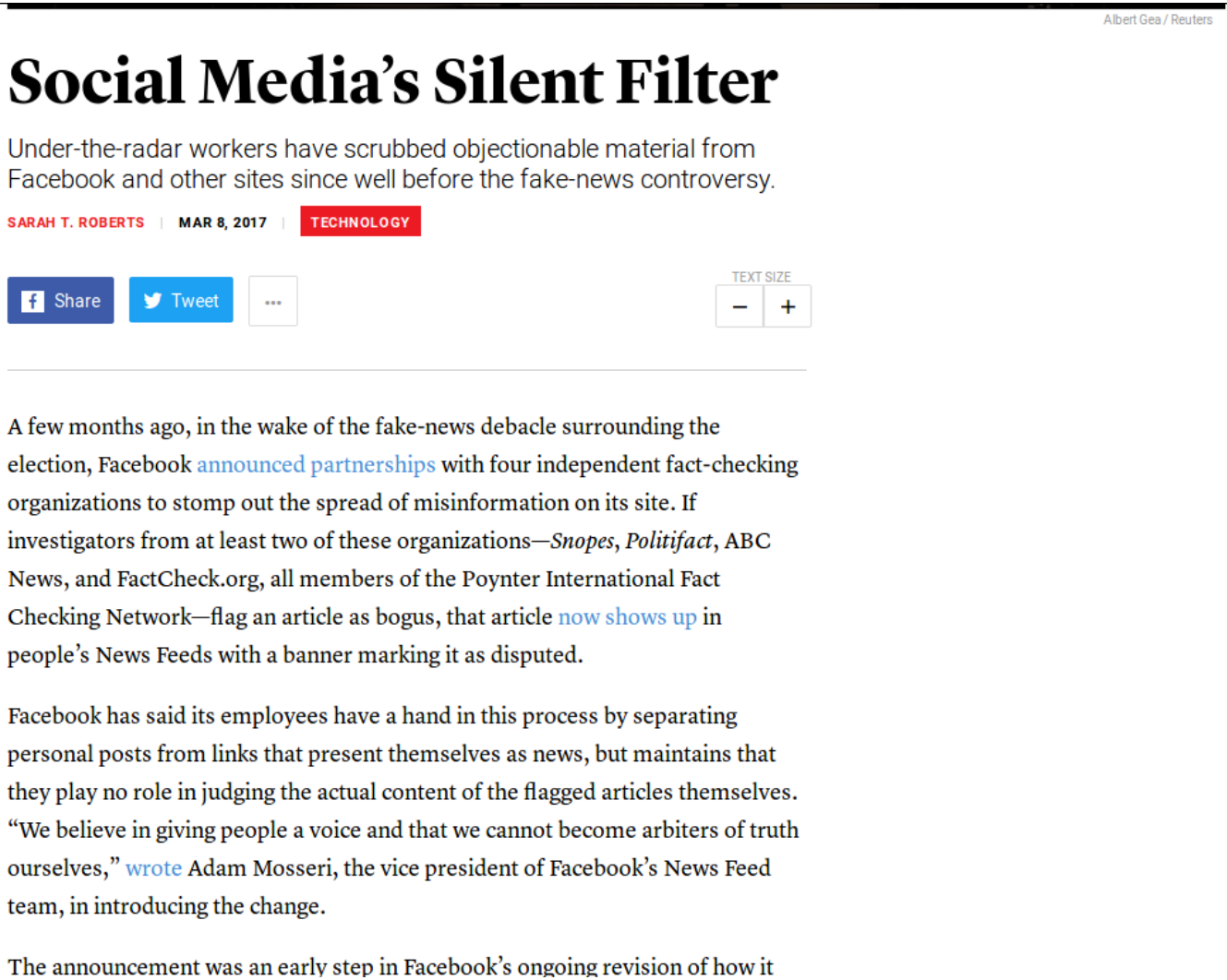
Esperando por ophan.theguardian.com...

1,599 780

Advertisement

INCT CYBER

Accountability?



Albert Gea / Reuters

Social Media's Silent Filter

Under-the-radar workers have scrubbed objectionable material from Facebook and other sites since well before the fake-news controversy.

SARAH T. ROBERTS | MAR 8, 2017 | TECHNOLOGY

Share Tweet ...

TEXT SIZE – +

A few months ago, in the wake of the fake-news debacle surrounding the election, Facebook announced partnerships with four independent fact-checking organizations to stomp out the spread of misinformation on its site. If investigators from at least two of these organizations—*Snopes*, *Politifact*, ABC News, and FactCheck.org, all members of the Poynter International Fact Checking Network—flag an article as bogus, that article now shows up in people's News Feeds with a banner marking it as disputed.

Facebook has said its employees have a hand in this process by separating personal posts from links that present themselves as news, but maintains that they play no role in judging the actual content of the flagged articles themselves. "We believe in giving people a voice and that we cannot become arbiters of truth ourselves," wrote Adam Mosseri, the vice president of Facebook's News Feed team, in introducing the change.

The announcement was an early step in Facebook's ongoing revision of how it

INCT CYBER

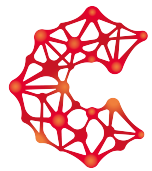
Accountability

- What are the personal rights regarding his/her collected data?
- What are the acceptable uses of data?
- Who is liable when something goes wrong?
- How can we report on algorithmical *abuse*?

GDPR

General Data Protecion Regulation

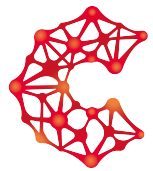
- Approved in April, 2016.
- Effective in 2018
- Three basic rights:
  - Right to access
  - Right to be forgotten
  - Right to explanation



GDPR

Article 22: Automated individual decision making, including profiling

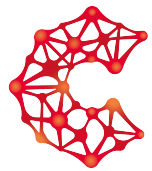
1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similiarity significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
    1. is necessary for entering into, or performance of, a contract between the data subject and the data controller.
    2. is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's right and freedoms and legitimate interests.
    3. is based on data subject's explicit consent.



Right to explanation

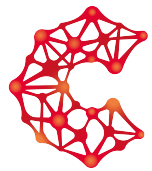
Basically, is the right for some interpretability and fairness assurance, but it is challenging:

- intentional concealment on the part of the institutions;
- gaps in technical literacy which mean that having access to technical details is not enough;
- a mismatch between the computational models and the demands of human-scale reasoning and styles of interpretation.



Principles for Algorithmic Transparency and Accountability

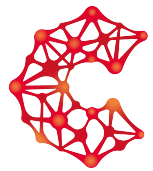
1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data provenance
6. Auditability
7. Validation and testing



Principles for Algorithmic Transparency and Accountability

1. Awareness

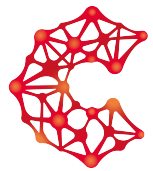
Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.



Principles for Algorithmic Transparency and Accountability

2. Access and redress

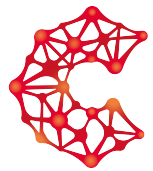
Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.



Principles for Algorithmic Transparency and Accountability

3. Accountability

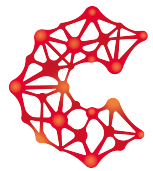
Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.



Principles for Algorithmic Transparency and Accountability

4. Explanation

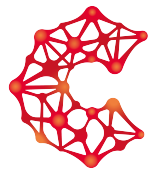
Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.



Principles for Algorithmic Transparency and Accountability

5. Data Provenance

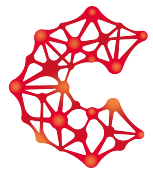
A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.



Principles for Algorithmic Transparency and Accountability

6. Auditability

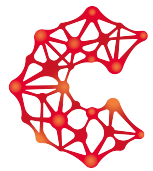
Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.



Principles for Algorithmic Transparency and Accountability

7. Validation and Testing

Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.



Conclusions

- The web was just a technology.
- Are we ready for understanding and exploiting this "new" web?
- The increasing usage of algorithms in the Web apps also comes with responsibilities.
- Making algorithms compatible with ethics and legal requirements may be hard.
- Research and development opportunities in all levels.
- Optimistic view about CS and its impact on society. Another opportunity!