

# Big Messy Data: Looking for the Signal in Noisy and Biased Data

*Alexandra Meliou*

*CRA-W Undergraduate Town Hall  
Oct 4<sup>th</sup>, 2018*



**CRA-W**

Computing Research Association  
Women

# Speaker & Moderator



*Alexandra Meliou*

Alexandra Meliou is an Assistant Professor in the College of Information and Computer Science, at the University of Massachusetts, Amherst. Prior to that, she was a Post-Doctoral Research Associate at the University of Washington. Alexandra received her PhD degree from the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley. She has received recognitions for research and teaching, including a CACM Research Highlight, an ACM SIGMOD Research Highlight Award, an ACM SIGSOFT Distinguished Paper Award, an NSF CAREER Award, a Google Faculty Research Award, and a Lilly Fellowship for Teaching Excellence. Her research focuses on data provenance, causality, explanations, data quality, and algorithmic fairness.



*Lori Pollock*

Dr. Lori Pollock is a Professor in Computer and Information Sciences at University of Delaware. Her current research focuses on program analysis for building better software maintenance tools, software testing, energy-efficient software and computer science education. Dr. Pollock is an ACM Distinguished Scientist and was awarded the University of Delaware's Excellence in Teaching Award and the E.A. Trabant Award for Women's Equity.



**CRA-W**  
Computing Research Association  
Women

EDITOR'S PICK

# Wisconsin Supreme Court allows state to continue using computer program to assist in sentencing

KATELYN FERRAL | The Capital Times | kferral@madison.com | @katelynferral Jul 13, 2016



is one with me," Stifelman says as his mechanized appendages pull tight another knot.

at 7,000 pieces of data primarily from credit create more for



# DATA



**CRA-W**  
Computing Research Association  
Women



A red, hand-drawn oval with a slightly textured, brush-like appearance, centered on the slide.

“What if data is dirty?”

“What if data is biased?”

Chicago Tribune

HOME NEWS BUSINESS SPORTS A&E LIFESTYLES OPINION REAL ESTATE CARS JOBS

## Data entry error wipes out life insurance coverage

Prudential says woman's policy on son had run out without warning because of wrong birth date

December 02, 2010 | By Jon Yates | V

Michael Mocny battled asthma for a long time, but he was far from his inhaler or his albuterol nebulizer. At 26 years old, he had grown adept at using his inhaler wrong when he woke up Oct. 5. After a terrible night of sleep during which he woke up, mother, Debbie McShane, he wanted to go to bed. He went upstairs to his room and attempted to check on her son. Mocny didn't respond, so he figured he was finally sleeping. McShane said she went back downstairs to check on her son's room and knocked a bit harder. She tried to shake him awake, but he didn't respond. She filled with albuterol. McShane had Mocny rushed to the hospital, but the doctor told McShane that her son's lungs were not working. Emergency crews couldn't revive him.

Related Articles

Collecting death benefit was not so simple  
April 23, 2009

Waiting for m.trb.com...

Oakland Unified makes \$7.6 million accounting error budget asking schools not to count on it

archived.oaklandlocal.com/article/oakland-unified-makes-76-million-accounting-error-budget-asking-schools-not-to-count-on-it

oakland LOCAL

This is the Oakland Local archive. To find more stories like this, you can search our main site.

Oakland Local is a news site. Questions? Contact us.

ABOUT STORY CATEGORIES COMMUNITY

Oakland Unified makes \$7.6 million accounting error budget asking schools not to count on it

Published on Thursday, February 07, 2013  
Last updated on 05:30PM, Monday, February 11, 2013

VERNON HAL Deputy Superintendent, Business & Operations

February 9, 2013

Recently, we informed you of a data entry error that inflated the budget figure for all schools sites by a cumulative total of \$7.6 million. We apologize for this mistake and also want to clarify what happened, the impact it had, and our response.

Prior to Winter Break, site administrators received a one-page summary of prior total allocations for each site. The summary indicated that funding for fiscal allocations, was flat, essentially the same as the current 2012-13 school's allocations were loaded in the tool used to plan expenses. This tool shows item present the full picture of expenses incurred by the site - it also shows utilizes the site level.

In this instance, an amount equal to the utility costs for a school site were error the tool associated with central office costs, and a second time in the section 9. The latter allocation was a mistake and made the site budget appear larger than it was.

amednews.com

AMERICAN MEDICAL NEWS

HOME PAST WEEKS ARCHIVES TOPICS COLUMNS MULTIMEDIA

Published by the American Medical Association

Search Articles Search tips

SECTIONS » Government Profession Business Opinion Health

TOPICS » Health Reform EHRs Medicare Liability AMA House » More

COLUMNS » Contract Language Ethics Forum In the Courts Practice Management Technically Speaking » More

LISTINGS » Issue dates Regions Columns Archives Writers

## Data entry is a top cause of medication errors

■ Training and design are seen as keys to reducing electronic prescribing errors.

By ANDIS ROBEZNIKES — Posted Jan. 24, 2005

PRINT | EMAIL | RESPOND | REPRINTS | LIKE | SHARE | TWEET

Computerized prescribing systems might cut the quantity and severity of medication mistakes, but they can't eliminate them entirely, said patient safety experts who reviewed the U.S. Pharmacopeia's 5th annual study of medication error reports.

The study of the more than 235,000 error reports submitted in 2003 by 570 health care facilities was the largest ever by USP. And as the number of reported errors goes up, the percentage that causes patient harm has gone down. But the findings that generated the most discussion are those indicating that electronic prescribing is creating new types of errors.

"Computer entry" was the fourth-leading cause of errors, accounting for 13% (27,711) of the medication errors reported in 2003. In contrast, illegible or unclear handwriting was the 15th-leading cause, and accounted for 2.9% (6,134) of reported errors.

WITH THIS STORY:

- » The chief reasons
- » Most errors don't cause harm
- » External links
- » Related content

FEATURED

Confronting bias against obese patients

■ Medical educators are starting to raise awareness about how weight-related stigma can impair patient-physician communication and the treatment of obesity. [Read story](#)

Goodbye

■ American Medical News is ceasing publication after 55 years of serving physicians by keeping them informed of their rapidly changing profession. [Read story](#)

Policing medical practice employees after work

■ Doctors can try to regulate staff actions outside the office, but they must watch what

https://www.google.com/url?sa=t&ct=j&q=&esrc=s&source=web&cd=1&ved=0C88QFJAA&url=http%3A%2F%2Fwww.amednews.com%2Farticle%2F20050124%2Fprofession%2F301249959%2F4%...

## Poor-Quality Data Imposes Costs and Risks on Businesses, Says New Forbes Insights Report

Harvard  
Business  
Review

DATA

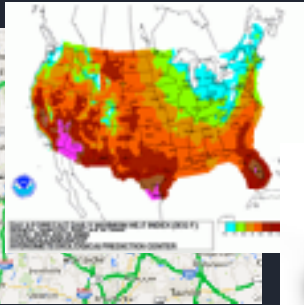
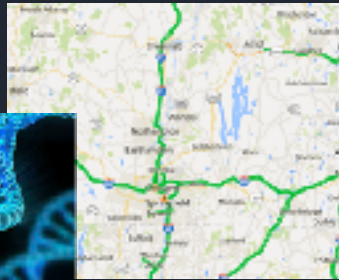
# Bad Data Costs the U.S. \$3 Trillion Per Year

by Thomas C. Redman

SEPTEMBER 22, 2016

SAVE SHARE COMMENT 10 TEXT SIZE PRINT \$8.95 BUY COPIES





data  
quality



DATA



The image features a hand in an orange rubber glove, covered in white foam, scrubbing a blue sponge against the word 'DATA'. The word is rendered in large, light blue, outlined capital letters. The background is white and filled with numerous small black dots, resembling dust or data points. The overall theme is data cleaning.

# DATA

**Data cleaning:**

Automated, semi-automated,  
manual, crowd-sourced...



**Barack Obama**  
44th U.S. President

Barack Hussein Obama II is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president, as well as the first born outside the contiguous United States.  
[Wikipedia](#)

**Born:** August 4, 1961 (age 55 years), Kapiolani Medical Center for Women and Children, Honolulu, HI  
**Height:** 6' 1"  
**Presidential term:** January 20, 2009 – January 20, 2017, 9:00 AM PST  
**Parents:** [Ann Dunham](#), [Barack Obama Sr.](#)  
**Education:** [Harvard Law School](#) (1988–1991), [More](#)






Quotes [View 7+ more](#)

*Change will not come if we wait for some other person or some other time. We are the ones we've been waiting for. We are the change that we seek.*

*If you're walking down the right path and you're willing to keep walking, eventually you'll make progress.*

*The future rewards those who press on. I don't have time to feel sorry for myself. I don't have time to complain. I'm going to press on.*

People also search for [View 15+ more](#)

 <a href="#">Donald Trump</a>	 <a href="#">Hillary Clinton</a>	 <a href="#">Michelle Obama</a> Spouse	 <a href="#">Ann Dunham</a> Mother	 <a href="#">Vladimir Putin</a>
---	--	---	---	---

Structured information  
retrieved from  
unstructured web  
data

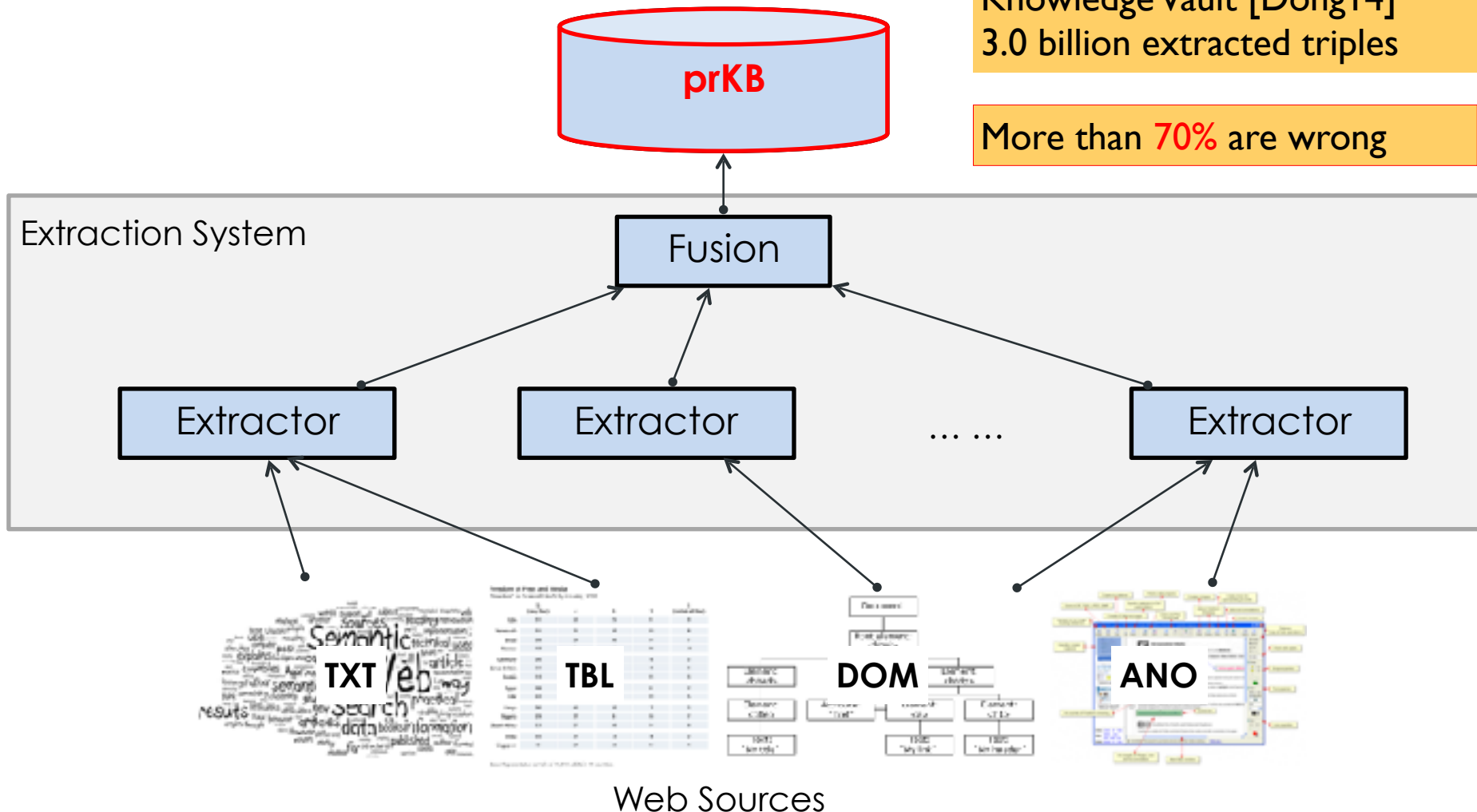


# a look at Knowledge Bases

Example:

Knowledge Vault [Dong14]  
3.0 billion extracted triples

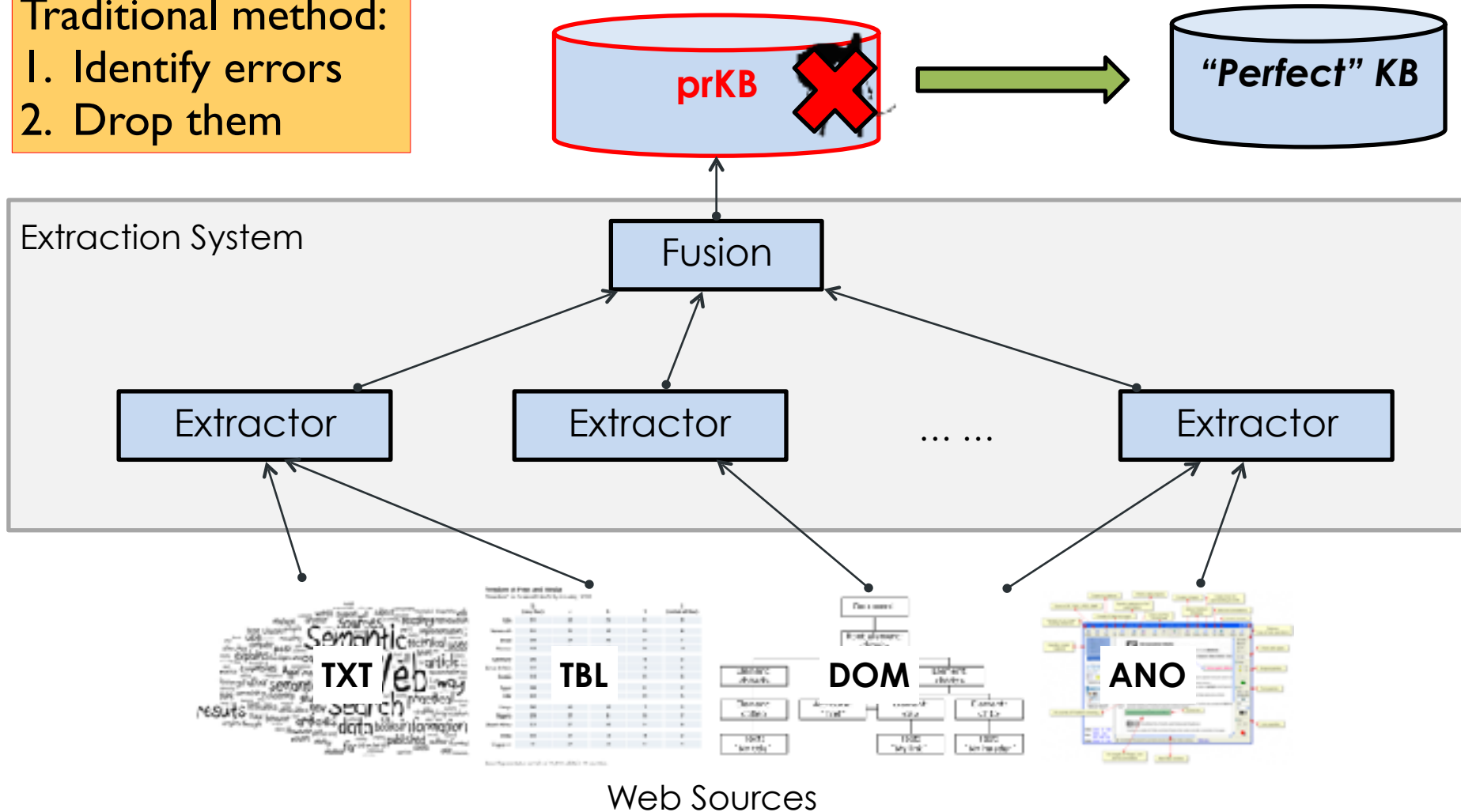
More than **70%** are wrong



# traditional data cleaning

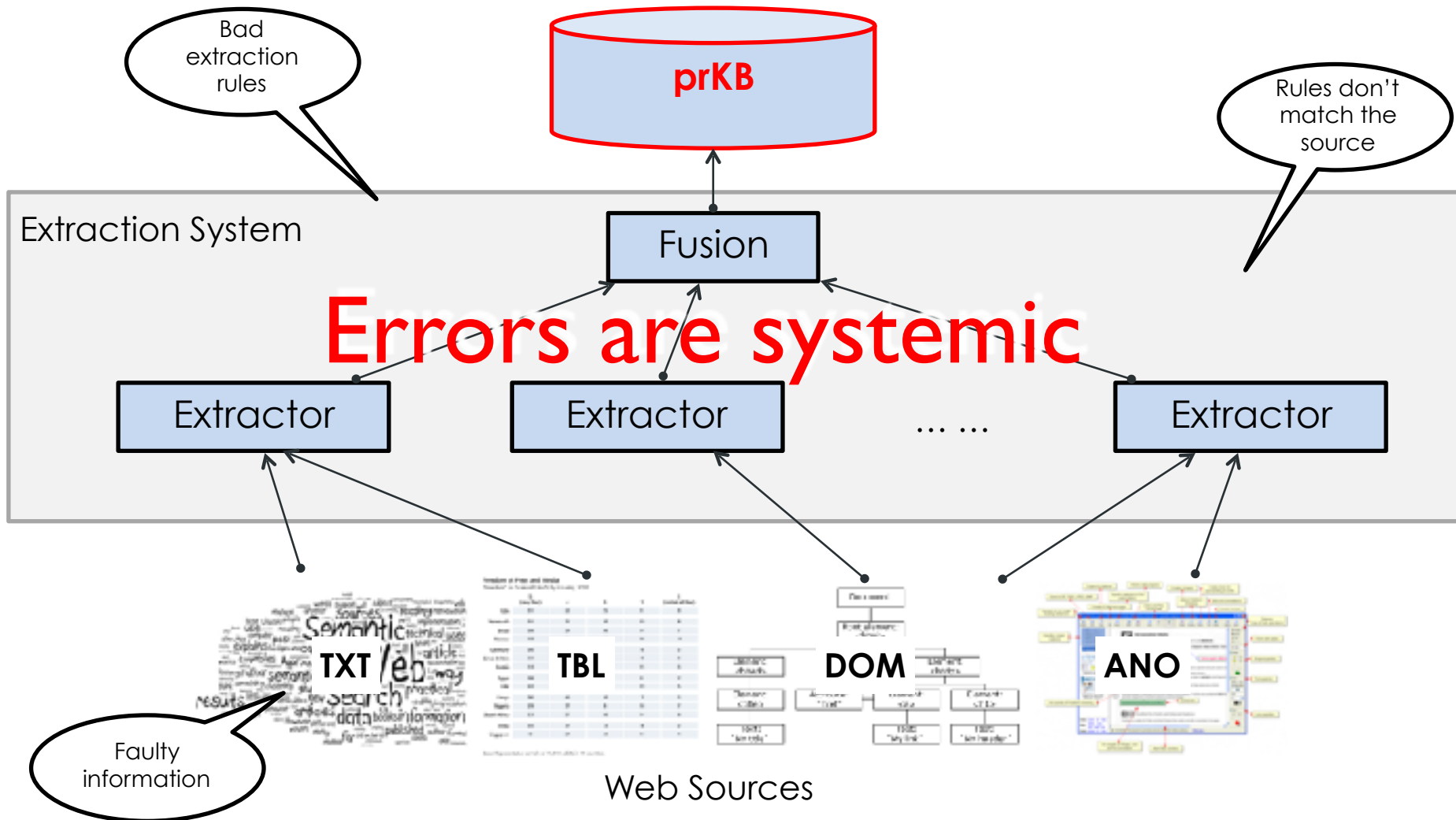
Traditional method:

1. Identify errors
2. Drop them

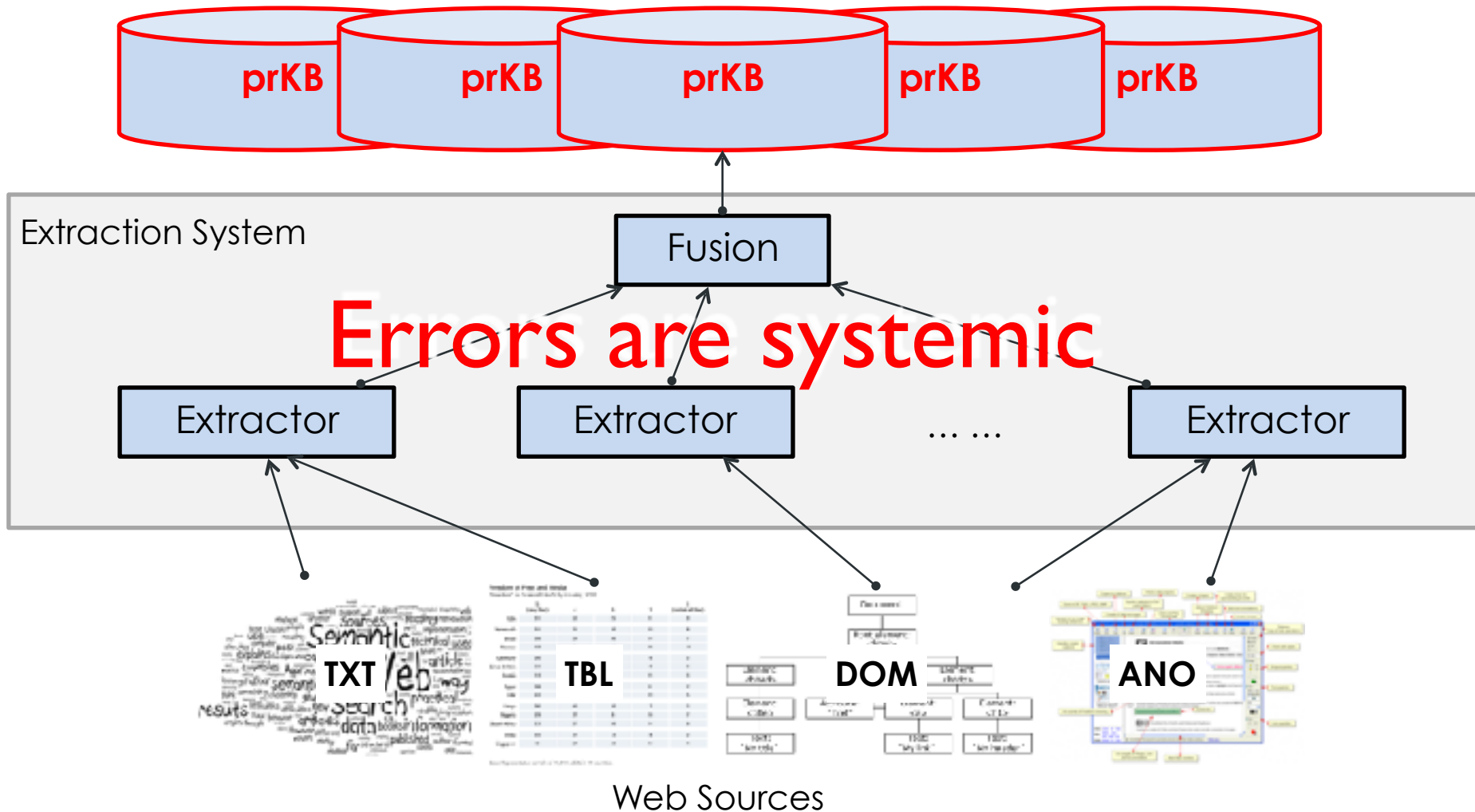




# but the errors are deeper...



...continue producing bad data



# CHALLENGES

- Massive scale

Sampling loses statistical strength and misses a lot of mistakes

- System complexity

Complex processes, thousands extraction patterns

- High error rates



**CRA-W**

Computing Research Association  
Women



# DATA X-RAY

Works on simple meta-data  
Parallelizable in MapReduce

## Example: Default value error

([besoccer.com](http://besoccer.com), `date_of_birth`, 1986-02-18)

# Triples 630

Error Rate 100%

Context: Date of birth of athletes extracted from [besoccer.com](http://besoccer.com) is set to default value 1986-02-18, due to copied html segments



**CRA-W**

Computing Research Association  
Women

“What if data is dirty?”

“What if data is biased?”



# Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica  
May 23, 2016

Software can make bad decisions.  
Software can discriminate!

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

...area, as well as the ability of our various carrier partners to deliver up to 9:00 pm every single day, even Sunday .

INSIDER

Real-time market data. Get the latest on stocks, commodities, currencies, funds, rates, ETFs, and

Up



# algorithms can exacerbate societal biases

The image displays three instances of Google Translate, illustrating how the algorithm perpetuates gender stereotypes in Turkish. In each instance, the source text is in English and the target text is in Turkish.

- Top instance:** The source text is "He is a nurse. She is a doctor." The Turkish translation is "O bir hemşire. O bir doktor." (He is a nurse. He is a doctor.). The word "hemşire" (nurse) is associated with "O" (He) and "doktor" (doctor) is associated with "O" (He). The interface shows "English", "Greek", and "Turkish" as source languages, and a "Translate" button.
- Middle instance:** The source text is "O bir hemşire. O bir doktor." (He is a nurse. He is a doctor.). The Turkish translation is "She is a nurse. He is a doctor." (She is a nurse. He is a doctor.). The word "hemşire" (nurse) is associated with "She" and "doktor" (doctor) is associated with "He". The interface shows "English", "Greek", and "Turkish" as source languages, and a "Translate" button.
- Bottom instance:** The source text is "O bir bilgisayar bilimcisidir" (He is a computer scientist). The Turkish translation is "He's a computer scientist". The interface shows "Turkish", "English", and "Spanish" as source languages, and a "Translate" button.

Each instance includes a "Suggest an edit" link and a "31/5000", "28/5000", and "29/5000" character count respectively.



TIME

SUBSCRIBE

IDEAS • TECHNOLOGY

## The Police Are Using Computer Algorithms to Tell If You're a Threat





Resilient cities Cities

## Predicting crime, LAPD-style

Cutting edge data-driven analysis directs Los Angeles patrol officers to likely future crime scenes - but critics worry that decision-making by machine will bring 'tyranny of the algorithm'

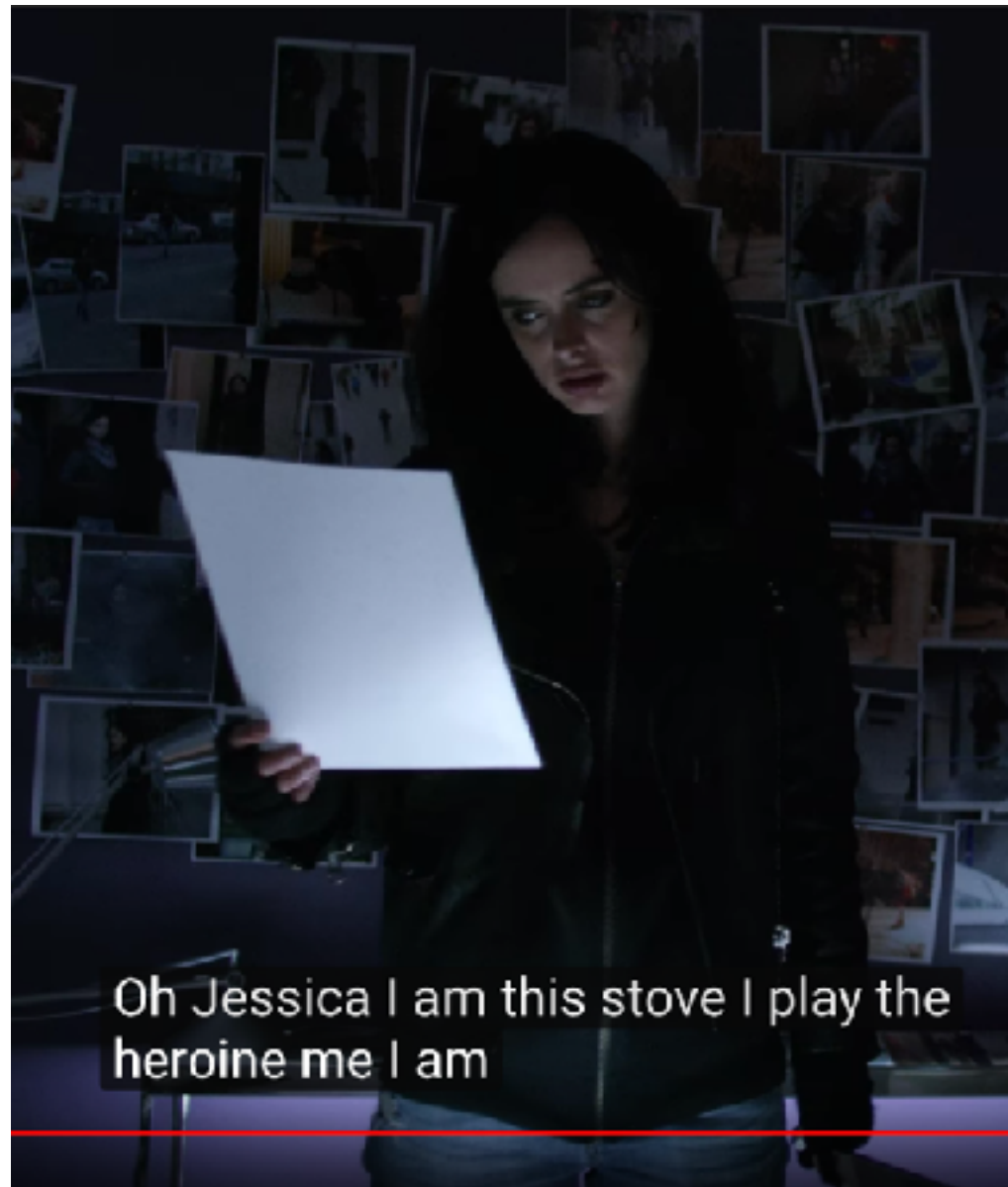
● Join our live Q&A with Homicide Watch this Friday



▲ PredPol co-developer P. Jeffrey Brantingham at the Unified Command Post in Los Angeles. 'This is not Minority Report,' he said. Photograph: Damian Dowarganes/AP

www.dhammadownload.com

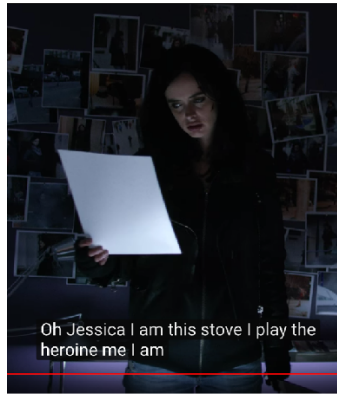
algorithms don't provide the same service to all



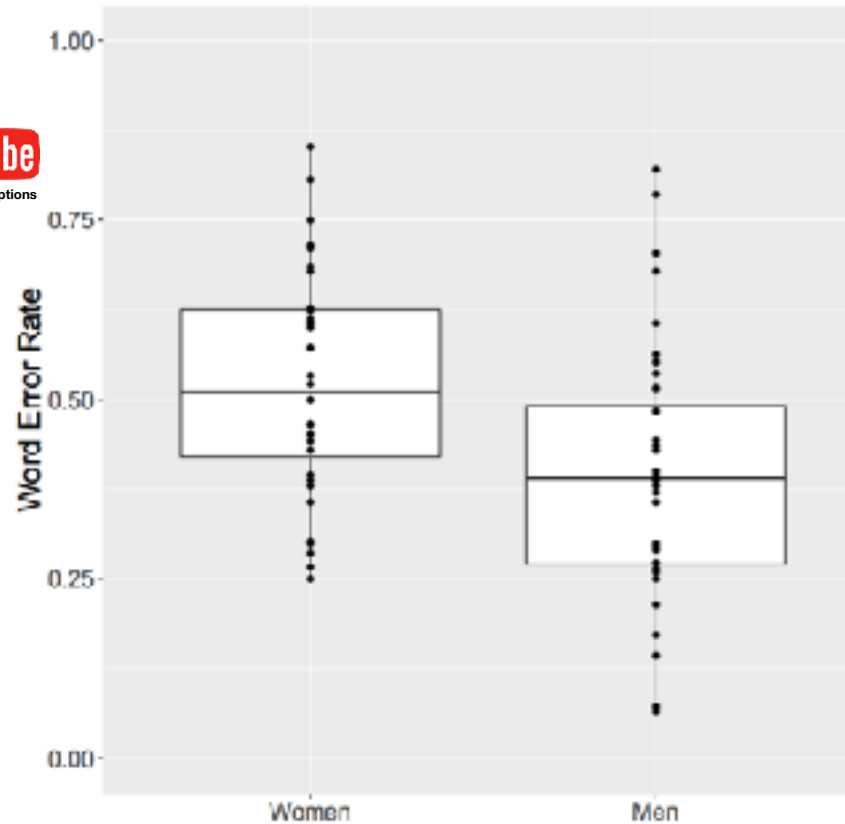
Oh Jessica I am this stove I play the  
heroine me I am

**You Tube**  
automatic captions

# algorithms don't provide the same service to all



**YouTube**  
automatic captions





algorithms don't provide the same service to all



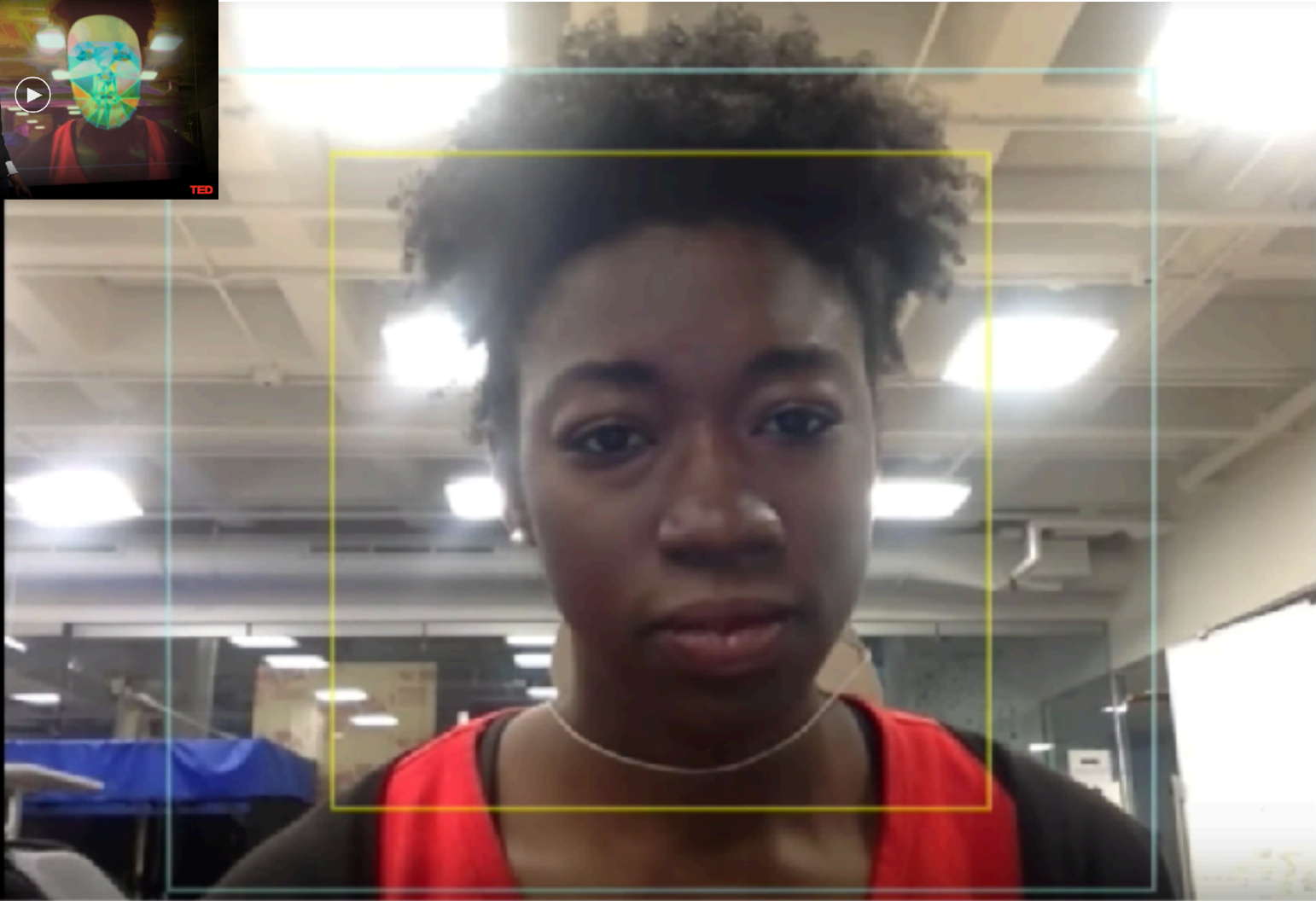
**JOY BUOLAMWINI**  
HOW I'M FIGHTING BIAS IN ALGORITHMS



Joy Buolamwini

[https://www.ted.com/talks/joy\\_buolamwini\\_how\\_i\\_m\\_fighting\\_bias\\_in\\_algorithms](https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms)

# algorithms don't provide the same service to all

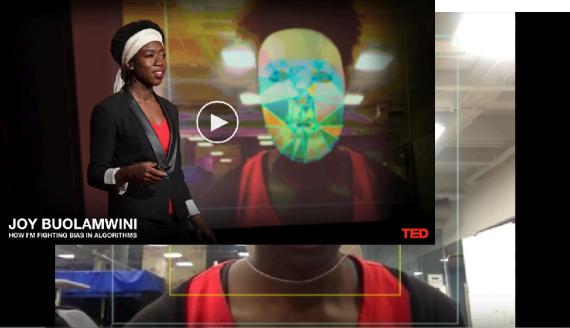


**Joy Buolamwini**

[https://www.ted.com/talks/joy\\_buolamwini\\_how\\_i\\_m\\_fighting\\_bias\\_in\\_algorithms](https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms)



# algorithms don't provide the same service to all



**Joy Buolamwini**

[https://www.ted.com/talks/joy\\_buolamwini\\_how\\_i\\_m\\_fighting\\_bias\\_in\\_algorithms](https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms)

# Research in Algorithmic Bias

- ♦ How to detect it
- ♦ How to measure it
- ♦ How to repair the algorithms
- ♦ How to repair the data

# **How to Fail!**

## **A Guide to Anticipating and Overcoming Failures**

our successes are public

our failures are not

what about awards and scholarships I got rejected for?

what about schools I didn't get into?

## Alexandra Meliou

College of Information and Computer Sciences 40-413-545-5788  
University of Massachusetts ameli@cs.umass.edu  
80 Governors Dr., Amherst, MA 01003-0264 <http://www.cs.umass.edu/~ameli>

### Research Interest

My research interests data management with user-facing functionality that helps people make sense of their data and use it effectively at a time when data becomes increasingly unpredictable, unsteady, and unmanageable. I focus on issues of provenance, causality, explanations, data quality, usability, and data and algorithmic bias. In the past, I have worked on a variety of topics spanning data management, machine learning, and sensor networks.

### Education

UNIVERSITY OF CALIFORNIA, BERKELEY ..... Berkeley, CA, USA  
2008 Doctor of Philosophy in Computer Science  
Dissertation: Querying uncertain data in resource-constrained settings  
Advisors: Joseph M. Hellerstein, Carlos Guestrin  
2005 Master of Science in Computer Science  
Thesis: Data gathering from sensor networks  
Advisors: Joseph Hellerstein, Carlos Guestrin  
NATIONAL TECHNICAL UNIVERSITY OF ATHENS ..... Athens, Greece  
2003 Diploma in Electrical and Computer Engineering (5-year degree)  
Thesis: Modeling and exploring the algebraic properties of hierarchical structures  
Advisor: Tamas Szell

### Employment History

UNIVERSITY OF MASSACHUSETTS ..... Amherst, MA, USA  
8/2002 - present Assistant Professor  
UNIVERSITY OF WASHINGTON ..... Seattle, WA, USA  
8/2009 - 8/2011 Postdoctoral Research Associate  
Advisor: Doug Gribble  
UNIVERSITY OF CALIFORNIA, BERKELEY ..... Berkeley, CA, USA  
8/2003 - 12/2003 Teaching Assistant  
1/2004 - 6/2005 Research Assistant  
IBM RESEARCH ..... Almaden, CA, USA  
3/2004 - 8/2004 Extension Researcher  
GO-ONLINE PROGRAM ..... Athens, Greece  
6/2002 - 12/2002 Networking Coordinator  
INTRACOM CORP ..... Athens, Greece  
6/2000 - 8/2000 Software Developer

Alexandra Meliou Curriculum Vitae August 29, 2018 Page 1 of 11

### Honors, Awards, Fellowships

2019 Communications of the ACM (C-ACM) Research Highlight  
2018 VLDB Distinguished Reviewer  
2018 SIGMOD Distinguished IC Member  
2017 ACM SIGMOD Distinguished Paper Award  
2017 ACM SIGMOD Research Highlight Award  
2016 Best paper of VLDB 2016  
2016 Lilly Fellowship for Teaching Excellence  
2015 NSF CAREER Award  
2013 Google Faculty Research Award  
2012 SIGMOD Best Demonstration Award  
2008 Gehlert Scholarship awarded by the Gehlert Foundation

### Research Grants

- SHF: Medicine: Inference in Software Systems  
NSF: The National Science Foundation  
Duration: Sep 15, 2018 - Aug 31, 2020  
PIs: Yury Brun, Alexandra Meliou  
\$1,050,000
- UAGER: Exploring the Feasibility of Software Testing Techniques to Evaluate Privacy Algorithms in Software Systems  
NSF: The National Science Foundation  
Duration: Sep 1, 2017 - Aug 3, 2018  
PIs: Yury Brun, Alexandra Meliou  
\$133,130
- CAREER: Reverse Data Management: Reverse Engineering Data Transformations to Understand, Diagnose, and Manipulate Data  
NSF: The National Science Foundation  
Duration: Aug 15, 2015 - July 31, 2020  
PI: Alexandra Meliou  
\$576,002
- DB: Small: Collaborative research: The package query problem  
NSF: The National Science Foundation  
Duration: Sep 1, 2011 - Aug 31, 2017  
PIs: Alexandra Meliou, Arava Aboussaid (New York University, Abu Dhabi)  
\$488,538 / Merit grant \$940,937
- UAGER: Data debugging  
NSF: The National Science Foundation  
Duration: Sep 1, 2011 - Feb 28, 2012  
PIs: Emory Berger, Alexandra Meliou  
\$139,000

Alexandra Meliou Curriculum Vitae August 29, 2018 Page 2 of 11

### • Bidirectional data cleaning

Google: Google Faculty Research Award  
Duration: Aug 16, 2013 - Aug 16, 2019  
PIs: Alexandra Meliou  
\$54,400

### Publications

#### Journal and Conference Publications

- [1] M. Brancato, A. Aboussaid, and A. Meliou. Scalable computation of high-order optimization queries. *Communications of the ACM*, 2018. [\[Research Highlight\]](#) (in production)
- [2] X. Wang, L. Han, and A. Meliou. Explaining Data Integration. *IEEE Data Engineering Bulletin*, 36(7):473-50, June 2018.
- [3] Y. Wang, A. Meliou, and G. Mikias. BI-Index: Diversifying Answers to Range Queries. *PVLDB*, 11(7):773-786, Sept. 2018.
- [4] M. Brancato, A. Aboussaid, and A. Meliou. Package queries: efficient and scalable computation of high-order constraints. *The VLDB Journal: Special Issue on Best Papers of VLDB 2016*.
- [5] I. Giallouras, Y. Irmak, and A. Meliou. Fairness testing: Testing software for discrimination. In *Proceedings of 2017 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering*, pages 488-510, 2017. [\[ACM SIGSOFT Distinguished Paper Award\]](#)
- [6] M. Brancato, A. Aboussaid, and A. Meliou. A Scalable Execution Engine for Package Queries. *SIGMOD Record*, 46(1):24-31, 2017. [\[ACM SIGMOD Research Highlight Award\]](#)
- [7] X. Wang, A. Meliou, and E. Wu. QFix: Diagnosing errors through query histories. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 1369-1386, 2017.
- [8] H. Zhang, Y. Xiao, and A. Meliou. Exstream: Explaining anomalies in event stream monitoring. In *36th International Conference on Extending Database Technology (EDBT)*, pages 161-183, 2017.
- [9] Y. Wang, A. Meliou, and G. Mikias. Lifting the Rate off the Cloud: A Consumer Centric Market for Database Computation in the Cloud. *PVLDB*, 11(1):373-384, 2018.
- [10] M. Brancato, I. I. Bekturov, A. Aboussaid, and A. Meliou. Scalable Package Queries in Relational Database Systems. *PVLDB*, 9(7):679-695, 2016. [\[Best Paper of VLDB 2016\]](#)
- [11] C. Frein, W. Gatterbauer, N. Immerman, and A. Meliou. Characterization of the complexity of inference and responsibility for self join free conjunctive queries. *PVLDB*, 9(3):183-194, 2016.
- [12] X. Wang, X. L. Dong, and A. Meliou. Data X-Ray: A Diagnostic Tool for Data Errors. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2011.
- [13] X. Wang, Y. Irmak, and A. Meliou. Preventing Data Errors with Continuous Testing. In *Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*, Baltimore, MD, USA, July 2015.

Alexandra Meliou Curriculum Vitae August 29, 2018 Page 3 of 11

what about my rejected grants?

what about jobs I didn't get?

what about all my paper rejections

**Let's demystify failure!**

# Let's demystify failure!

some lessons learned from my  
personal experiences

**an exam failure**



# an exam failure

Lesson: helped me shape a  
stronger background that I am  
now proud of

**research failure**

# research failure

Lesson: the slow down was temporary, and the core of what I learned guides my current work

**failure to network**

# failure to network

Lesson: seek collaborators and mentors, don't hesitate to ask and share

**job search failure**

# job search failure

Lesson: the slow down was temporary, and it allowed me to pursue better opportunities



# Good practices in handling failures

- ◆ *Do not isolate yourself*
- ◆ *Seek mentors (also aside from failures!)*
- ◆ *Think about what pieces of a failed attempt are reusable*
- ◆ *Do not let it reflect on your abilities and prospects*
  - ◆ *easier said than done, but it is true!*
- ◆ *Confront the reasons*

## **Remember:**

Even the most accomplished have failed numerous times



**CRA-W**

Computing Research Association  
Women