Big Data, Society, and Data Science Education

Chinese University of Hong Kong, Shenzhen
Shenzhen, China, 29 December 2017

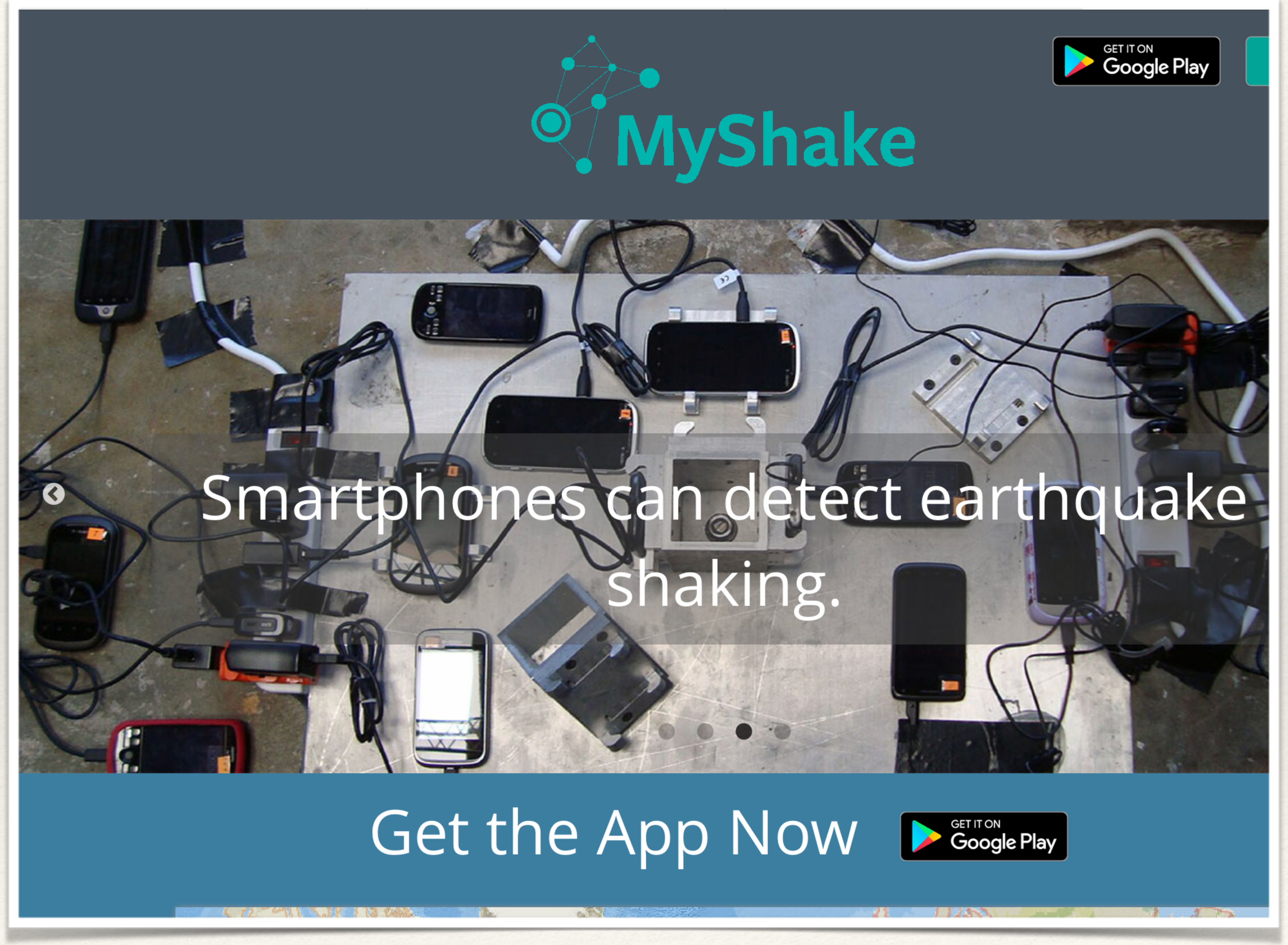
Philip B. Stark
University of California, Berkeley

Fallacies do not cease to be fallacies because they become fashions. —G.K. Chesterton

A Success Story: myShake

Join the MyShake community — help build the global seismic network.

Get the App Now



Smartphones can detect earthquake shaking.

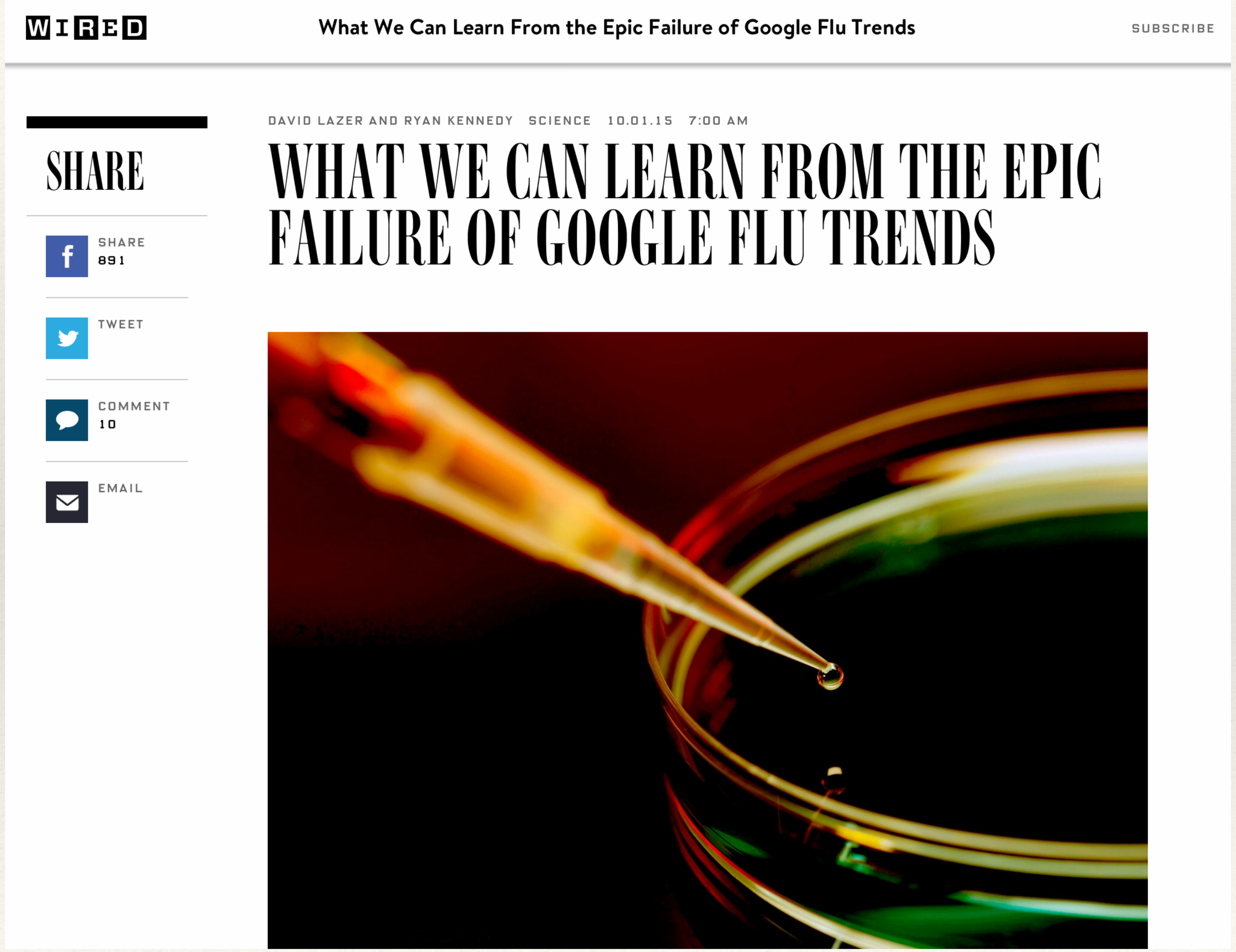
Get the App Now

Other successes

- Automation, robotics
- Supply chain, shipping
- Prediction of sepsis and other patient complications in hospitals
- Smart buildings, energy management
- Natural language processing
- Biometric identification

A Failure Story: Google Flu Trends

- Initial success "now casting" in 2008
- Serious failure in 2013; discontinued
- Blind modeling



WIRED What We Can Learn From the Epic Failure of Google Flu Trends SUBSCRIBE

SHARE

SHARE 891

TWEET

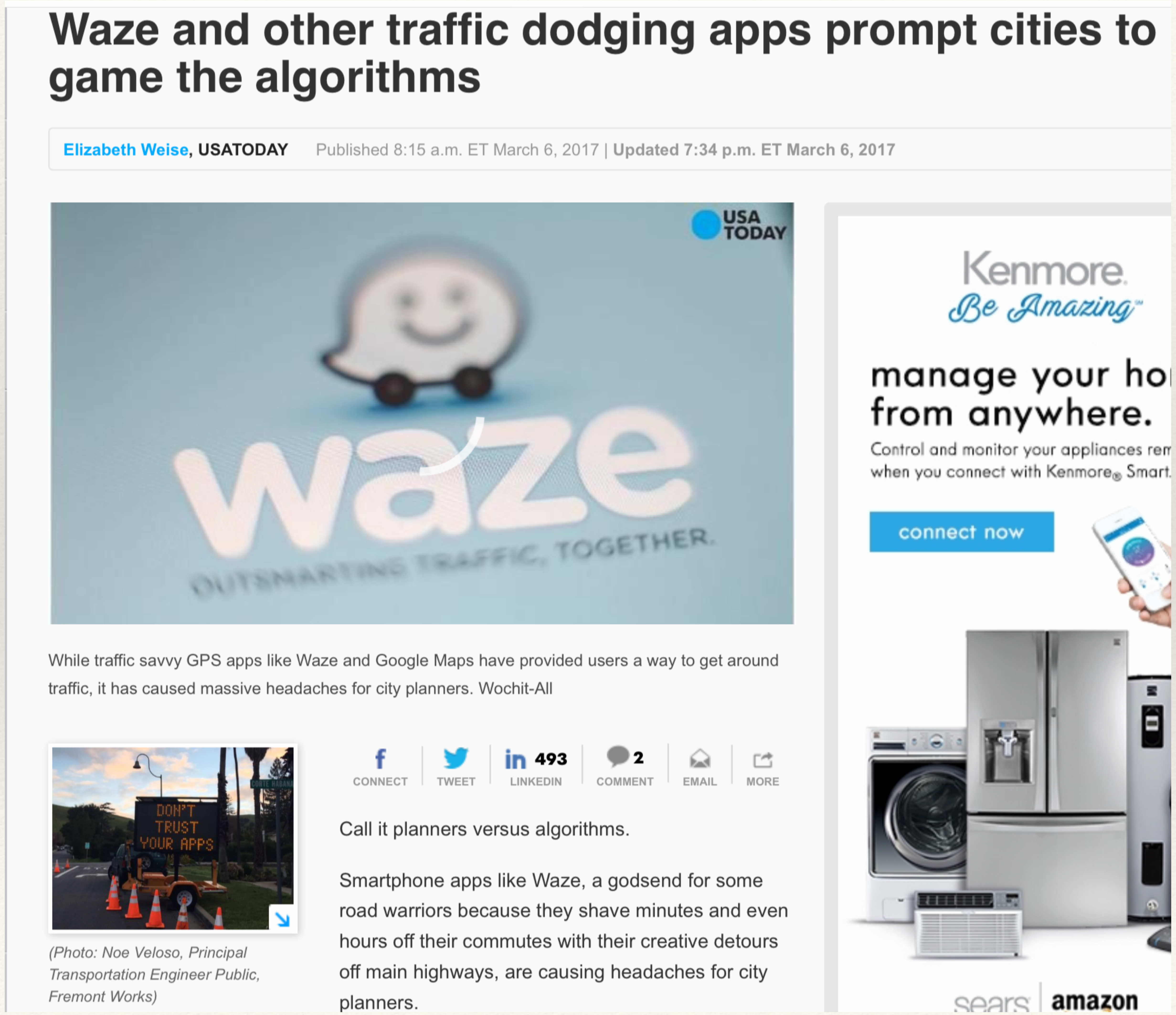
COMMENT 10

EMAIL

DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

Mixed: Traffic apps



Waze and other traffic dodging apps prompt cities to game the algorithms

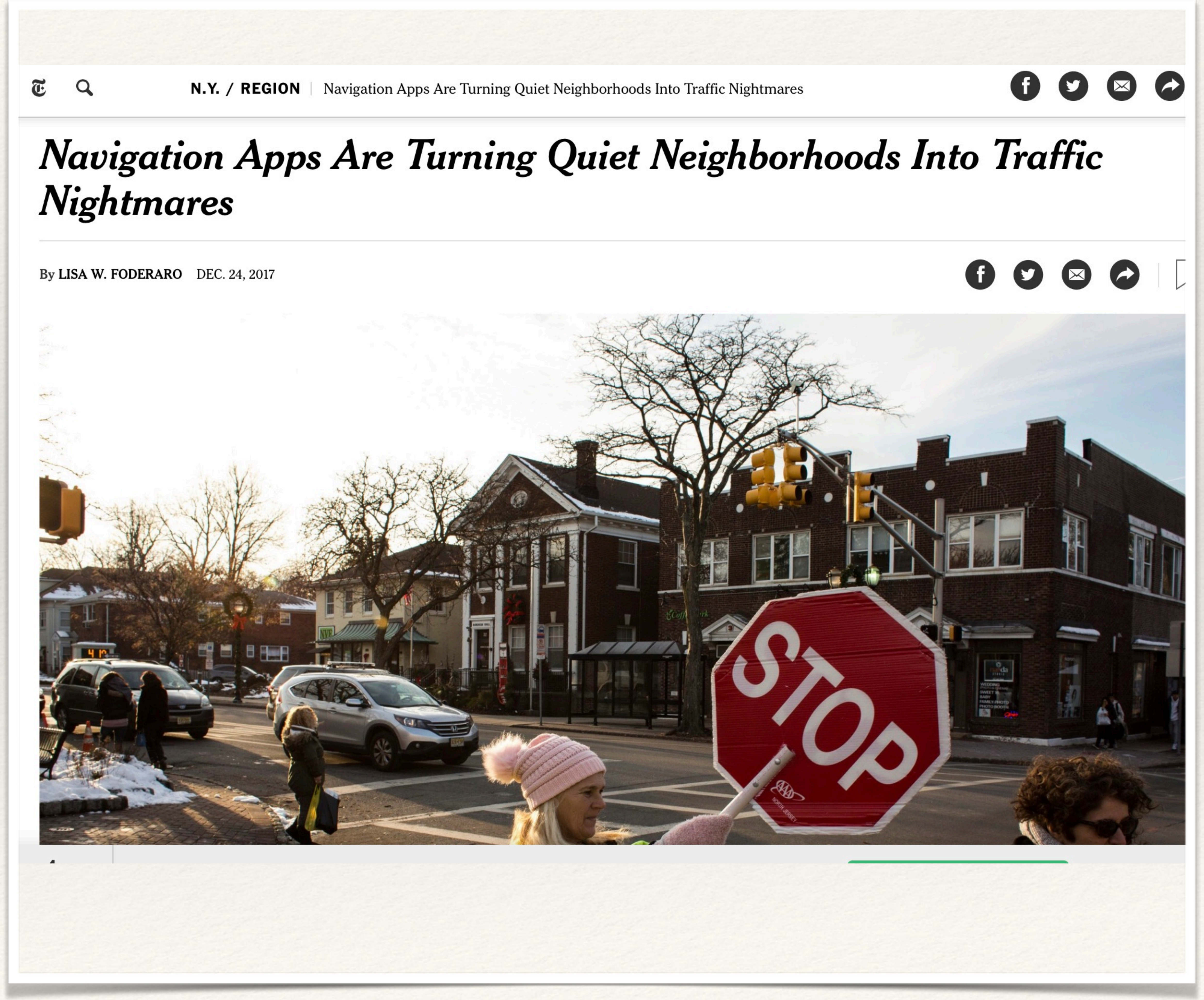
Elizabeth Weise, USATODAY Published 8:15 a.m. ET March 6, 2017 | Updated 7:34 p.m. ET March 6, 2017

While traffic savvy GPS apps like Waze and Google Maps have provided users a way to get around traffic, it has caused massive headaches for city planners. Wochit-All

(Photo: Noe Veloso, Principal Transportation Engineer Public, Fremont Works)

Call it planners versus algorithms.

Smartphone apps like Waze, a godsend for some road warriors because they shave minutes and even hours off their commutes with their creative detours off main highways, are causing headaches for city planners.

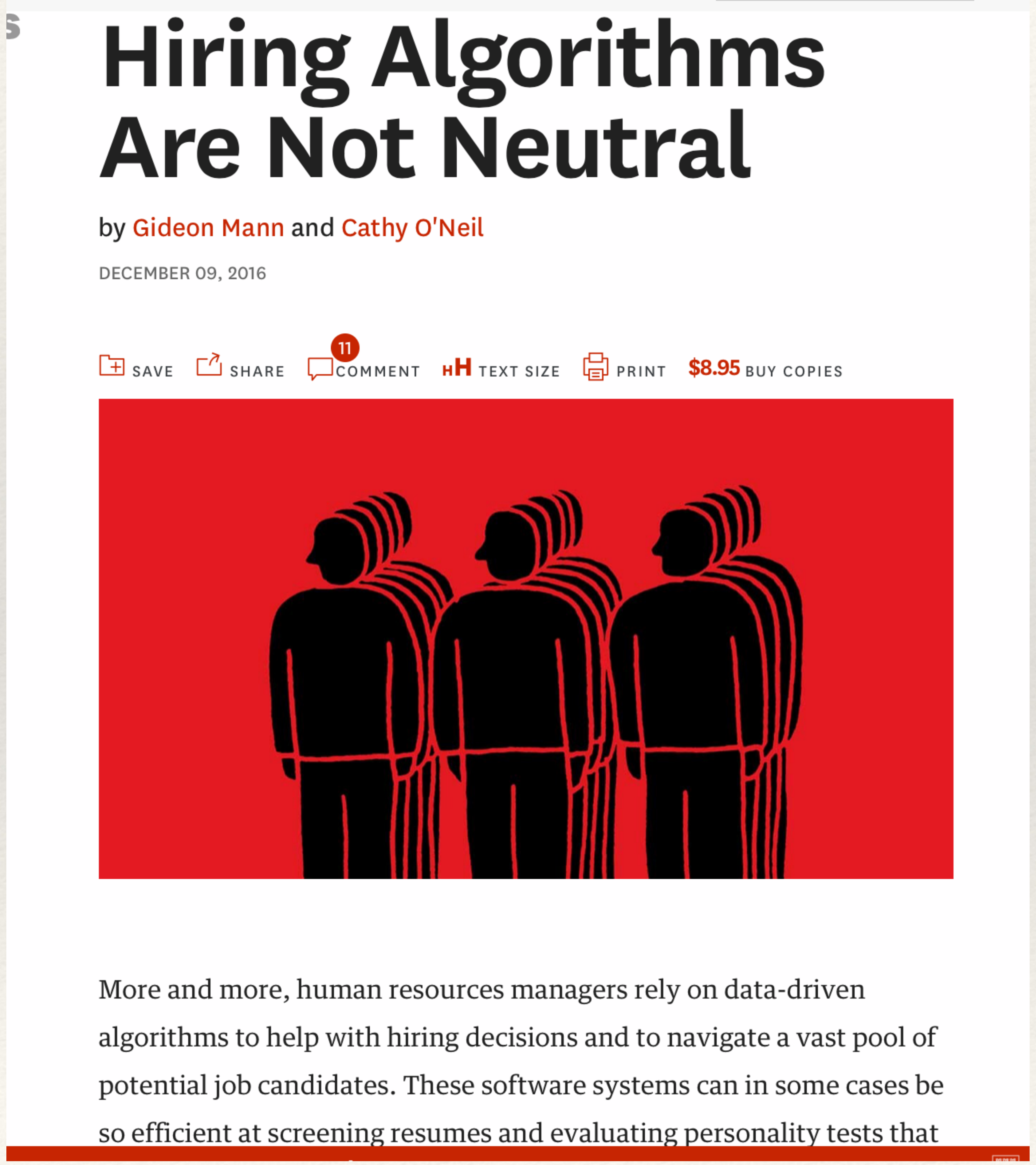


N.Y. / REGION | Navigation Apps Are Turning Quiet Neighborhoods Into Traffic Nightmares
STOP

Navigation Apps Are Turning Quiet Neighborhoods Into Traffic Nightmares

By LISA W. FODERARO DEC. 24, 2017

Reinforcing bias and inequality

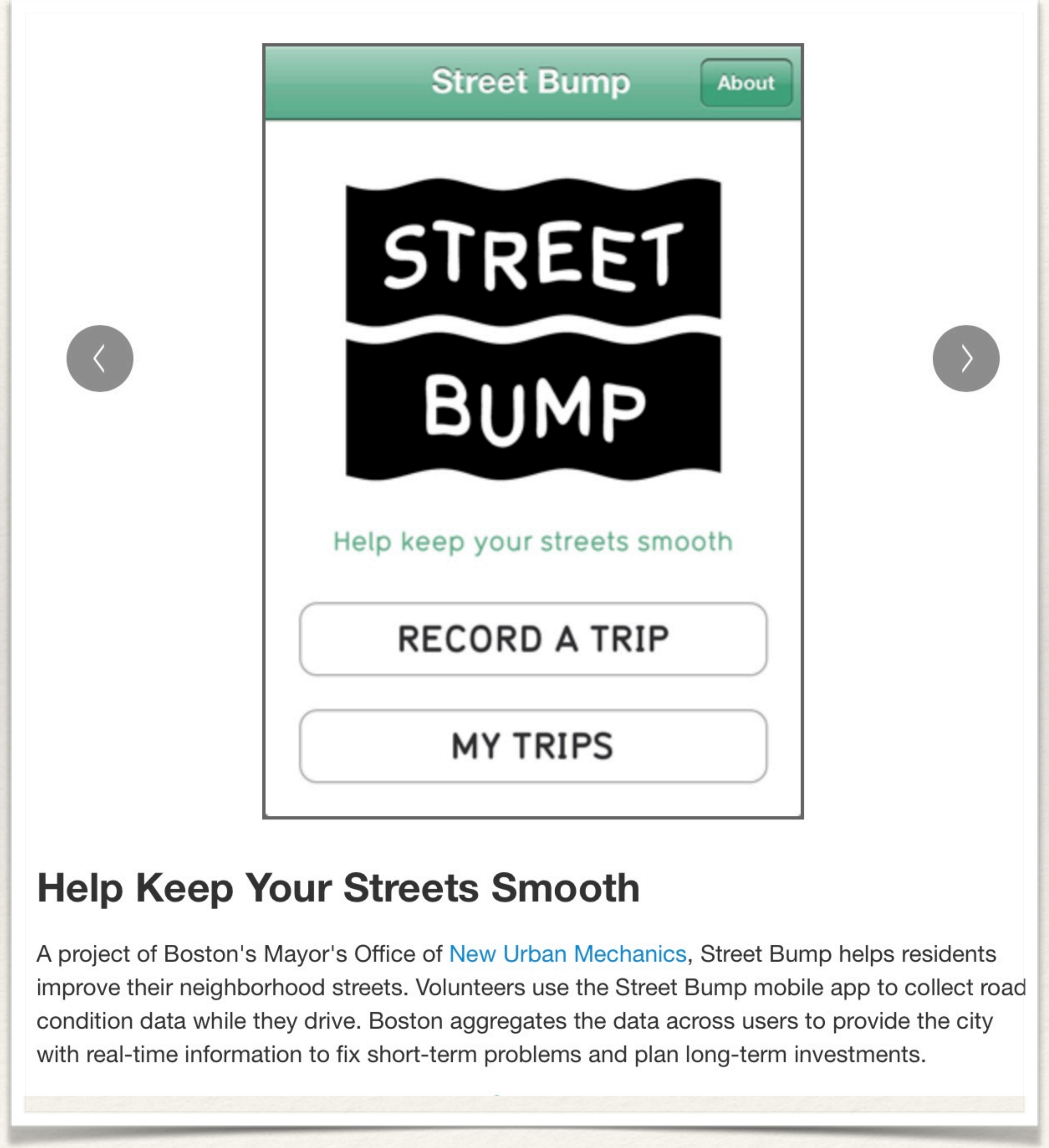


Hiring Algorithms Are Not Neutral

by Gideon Mann and Cathy O'Neil

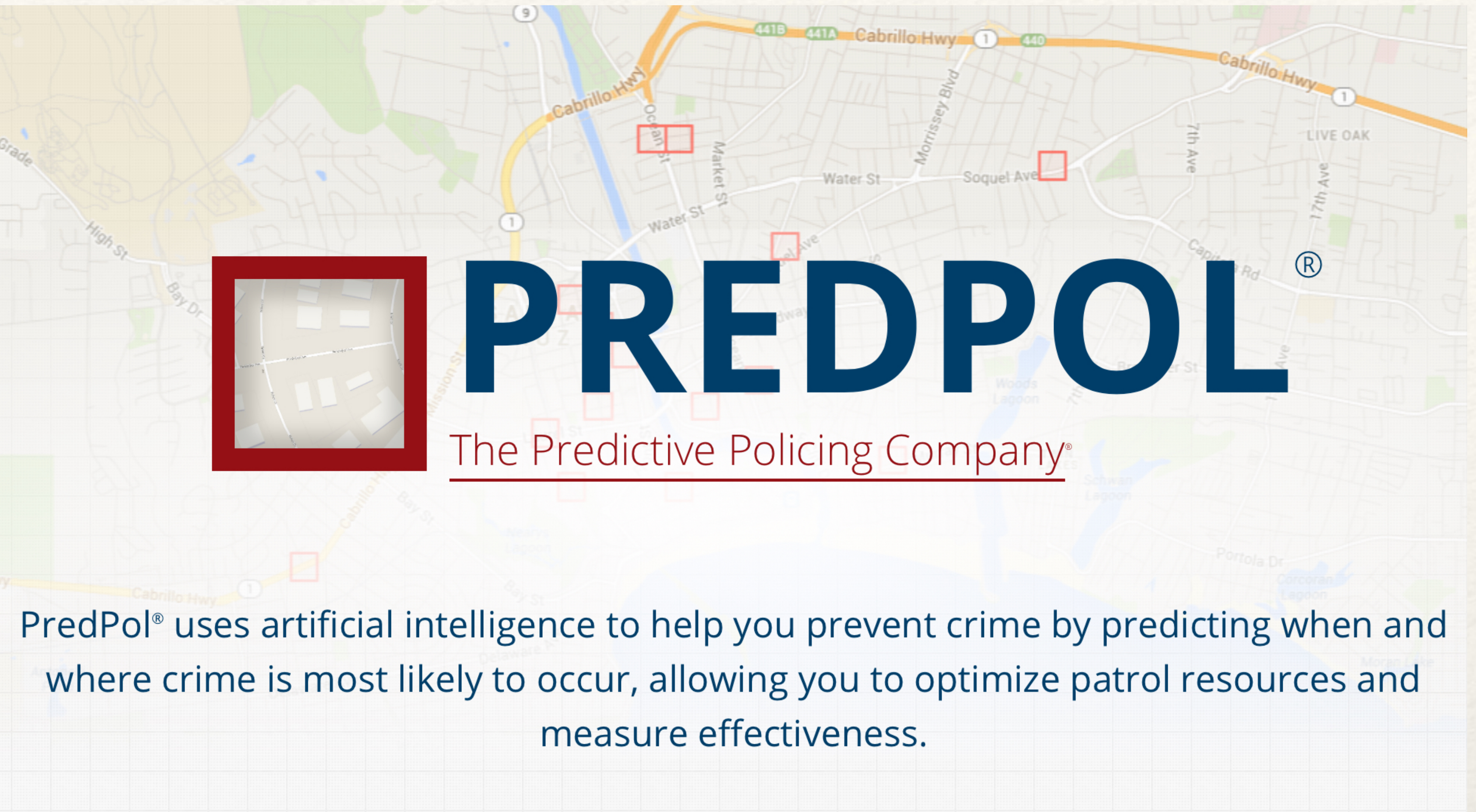
DECEMBER 09, 2016

More and more, human resources managers rely on data-driven algorithms to help with hiring decisions and to navigate a vast pool of potential job candidates. These software systems can in some cases be so efficient at screening resumes and evaluating personality tests that

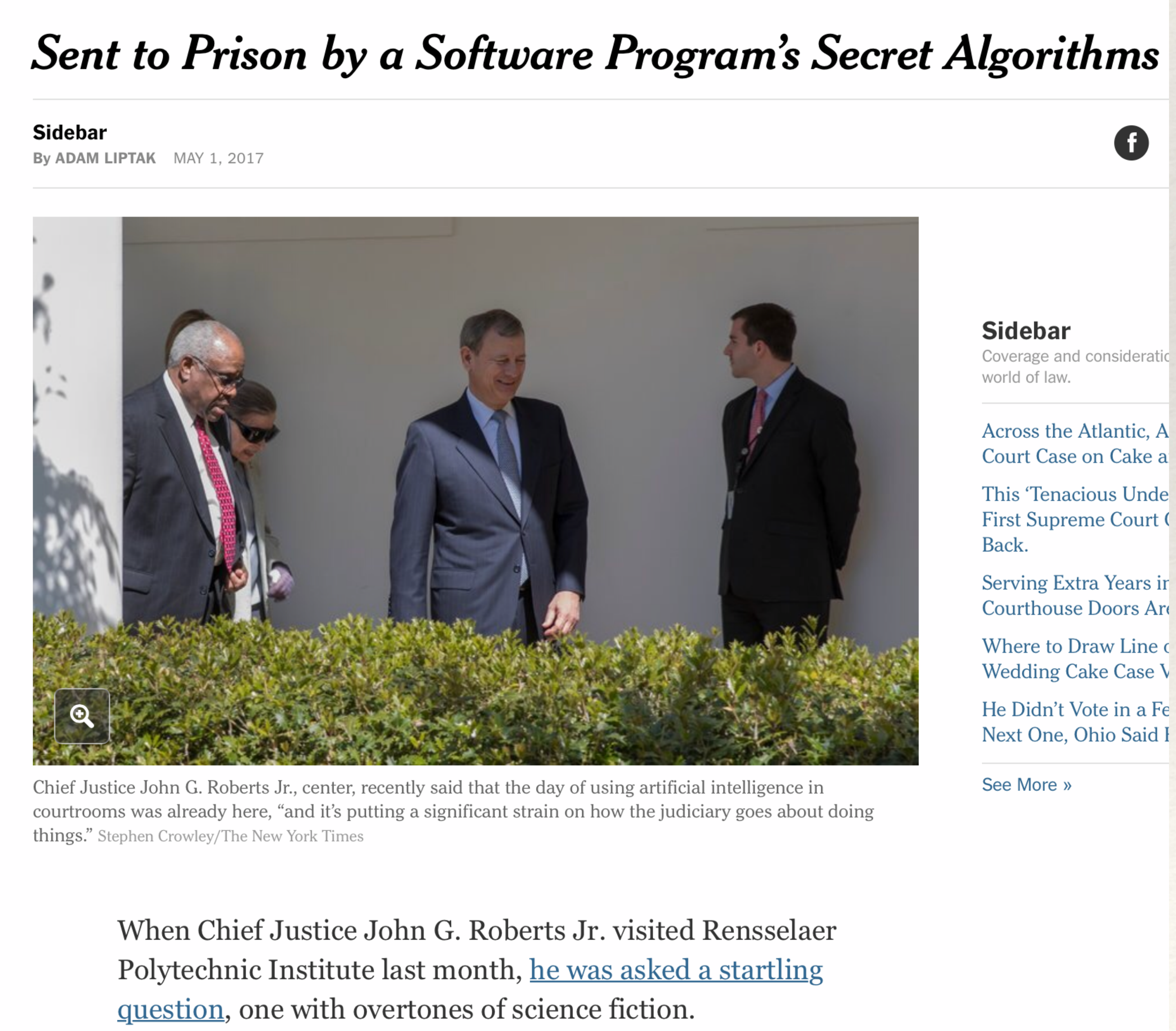


Help Keep Your Streets Smooth

A project of Boston's Mayor's Office of New Urban Mechanics, Street Bump helps residents improve their neighborhood streets. Volunteers use the Street Bump mobile app to collect road condition data while they drive. Boston aggregates the data across users to provide the city with real-time information to fix short-term problems and plan long-term investments.



PredPol® uses artificial intelligence to help you prevent crime by predicting when and where crime is most likely to occur, allowing you to optimize patrol resources and measure effectiveness.



Sent to Prison by a Software Program's Secret Algorithms

Sidebar

By ADAM LIPTAK MAY 1, 2017

Chief Justice John G. Roberts Jr., center, recently said that the day of using artificial intelligence in courtrooms was already here, "and it's putting a significant strain on how the judiciary goes about doing things." Stephen Crowley/The New York Times

Sidebar

Coverage and consideratic
world of law.

Across the Atlantic, A
Court Case on Cake a

This 'Tenacious Unde
First Supreme Court (
Back.

Serving Extra Years ir
Courthouse Doors Arc

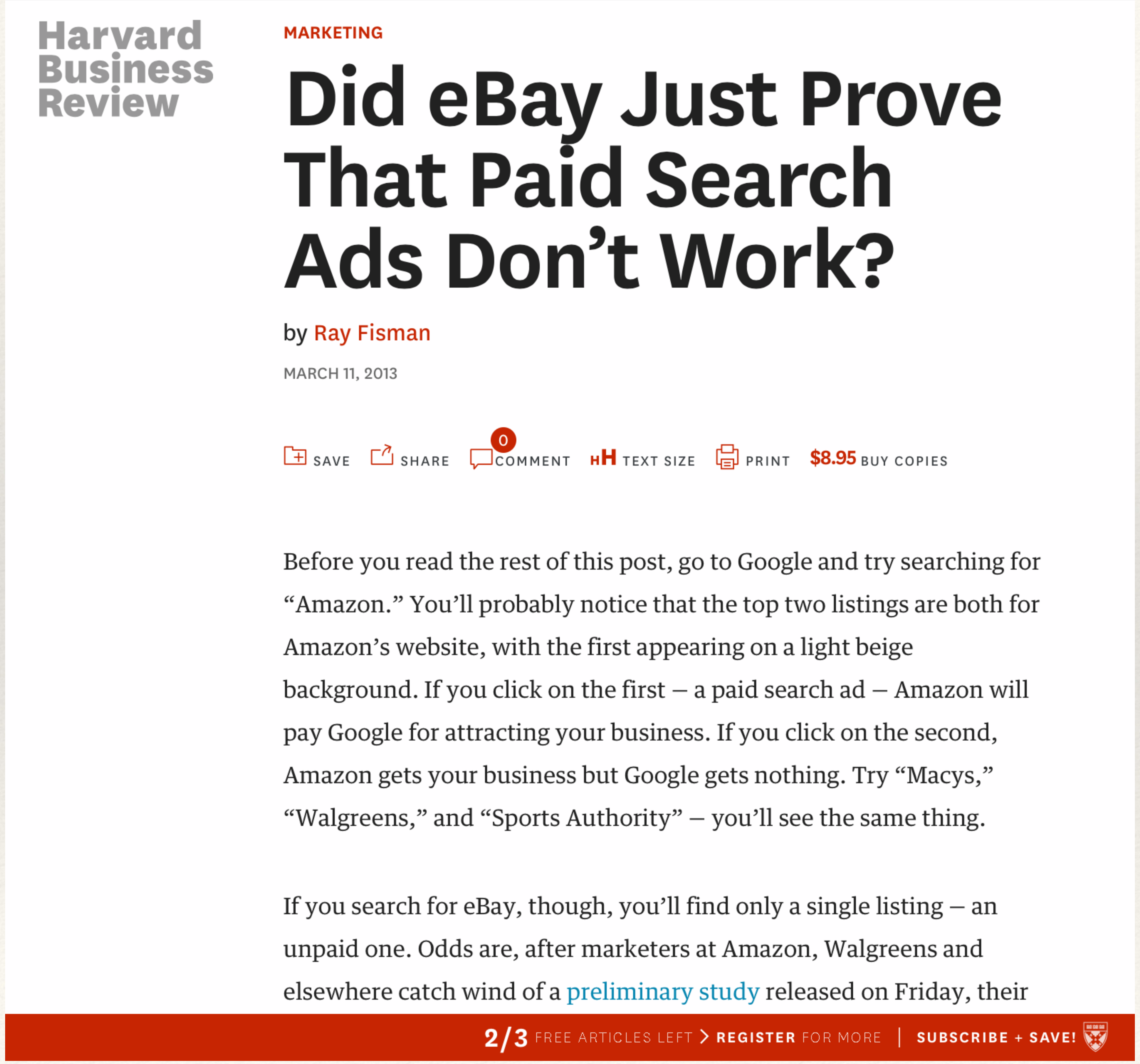
Where to Draw Line c
Wedding Cake Case V

He Didn't Vote in a Fe
Next One, Ohio Said I

See More »

When Chief Justice John G. Roberts Jr. visited Rensselaer Polytechnic Institute last month, he was asked a startling question, one with overtones of science fiction.

Bad Business



MARKETING

Did eBay Just Prove That Paid Search Ads Don't Work?

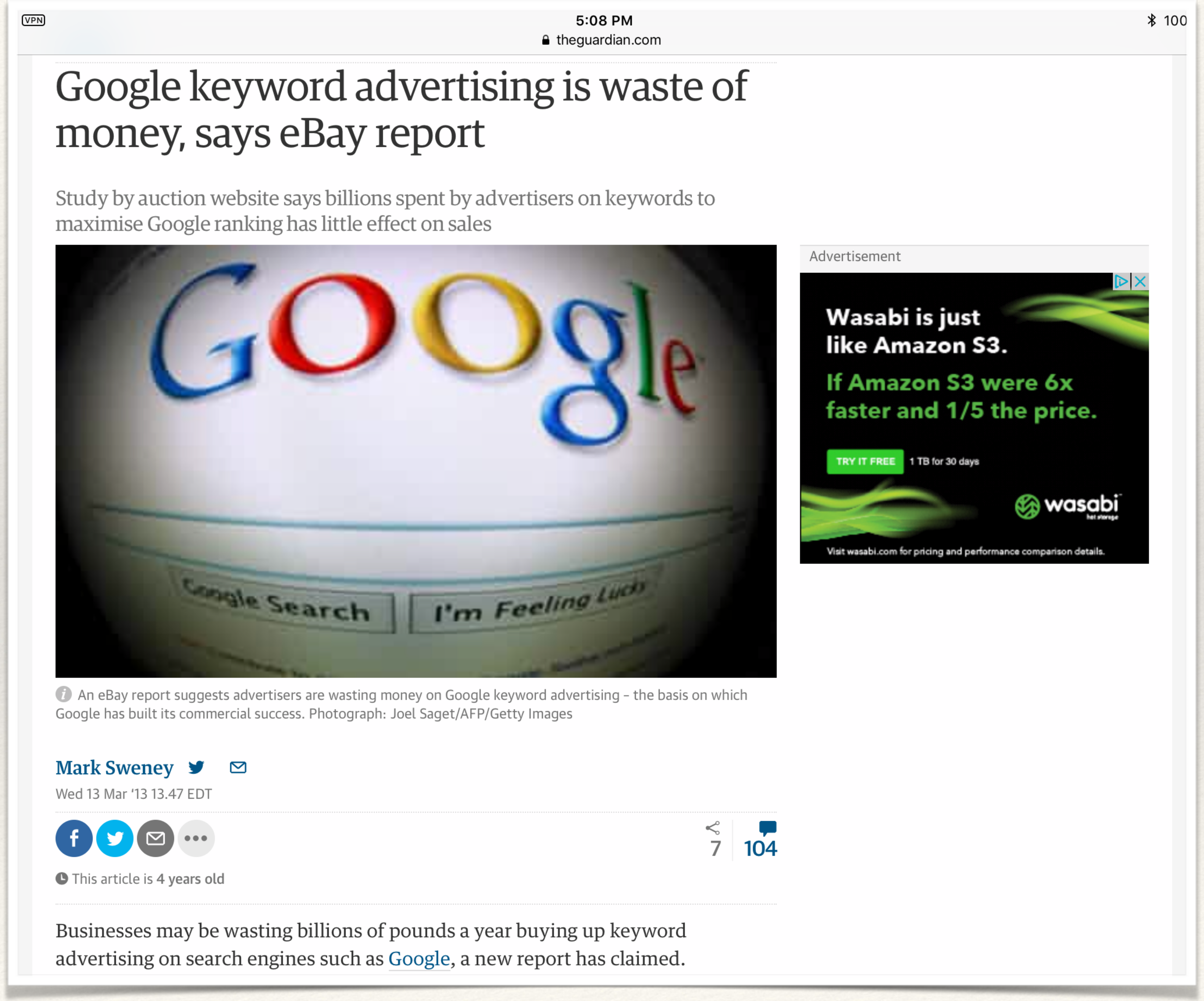
by Ray Fisman

MARCH 11, 2013

Before you read the rest of this post, go to Google and try searching for "Amazon." You'll probably notice that the top two listings are both for Amazon's website, with the first appearing on a light beige background. If you click on the first – a paid search ad – Amazon will pay Google for attracting your business. If you click on the second, Amazon gets your business but Google gets nothing. Try "Macys," "Walgreens," and "Sports Authority" – you'll see the same thing.

If you search for eBay, though, you'll find only a single listing – an unpaid one. Odds are, after marketers at Amazon, Walgreens and elsewhere catch wind of a preliminary study released on Friday, their

2/3 FREE ARTICLES LEFT > REGISTER FOR MORE | SUBSCRIBE + SAVE!



Google keyword advertising is waste of money, says eBay report

Study by auction website says billions spent by advertisers on keywords to maximise Google ranking has little effect on sales

An eBay report suggests advertisers are wasting money on Google keyword advertising - the basis on which Google has built its commercial success. Photograph: Joel Saget/AFP/Getty Images

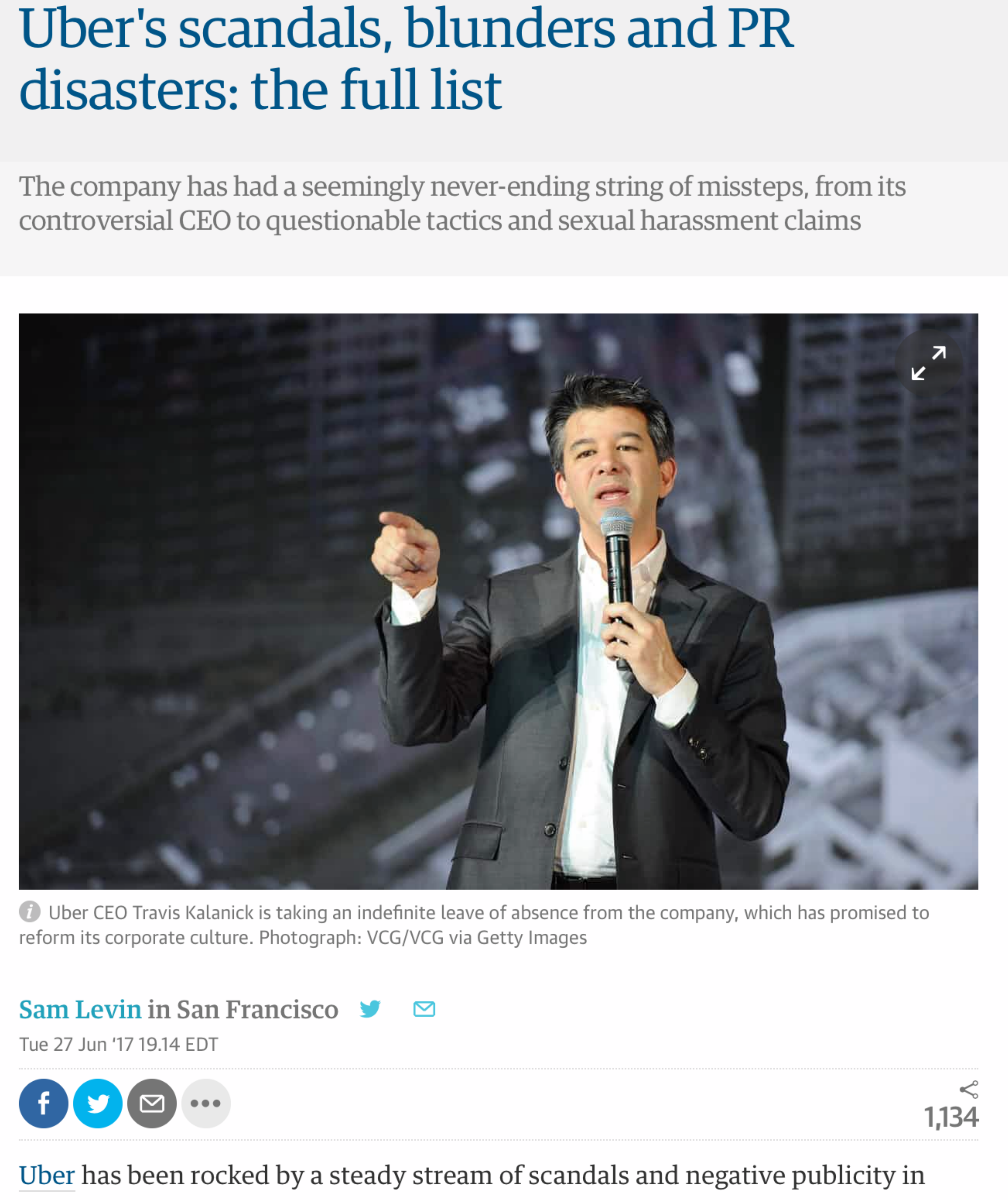
Mark Sweney

Wed 13 Mar '13 13.47 EDT

This article is 4 years old

Businesses may be wasting billions of pounds a year buying up keyword advertising on search engines such as Google, a new report has claimed.

Bad Businesses



Uber's scandals, blunders and PR disasters: the full list

The company has had a seemingly never-ending string of missteps, from its controversial CEO to questionable tactics and sexual harassment claims

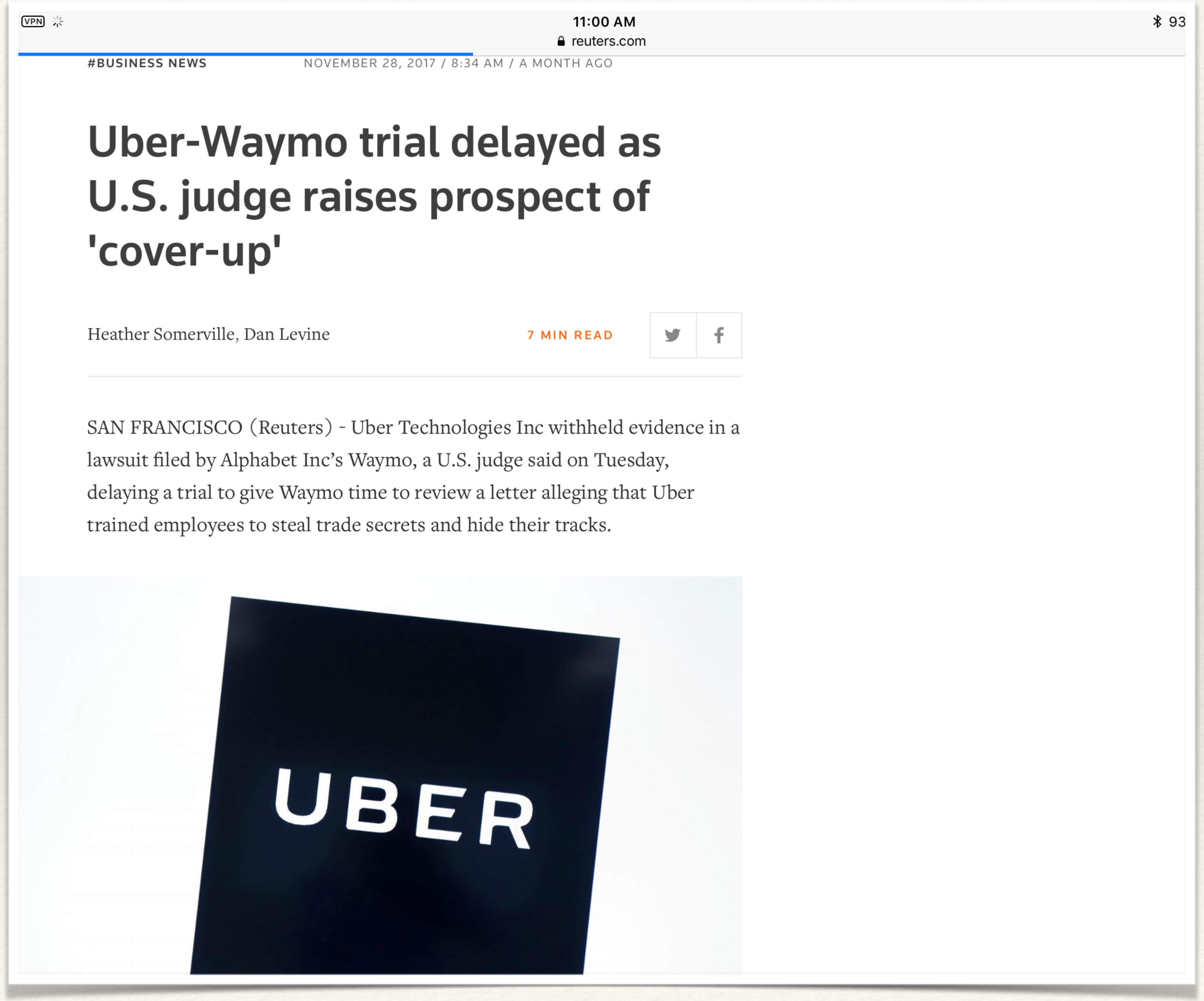
Uber CEO Travis Kalanick is taking an indefinite leave of absence from the company, which has promised to reform its corporate culture. Photograph: VCG/VCG via Getty Images

Sam Levin in San Francisco

Tue 27 Jun '17 19.14 EDT

1,134

Uber has been rocked by a steady stream of scandals and negative publicity in

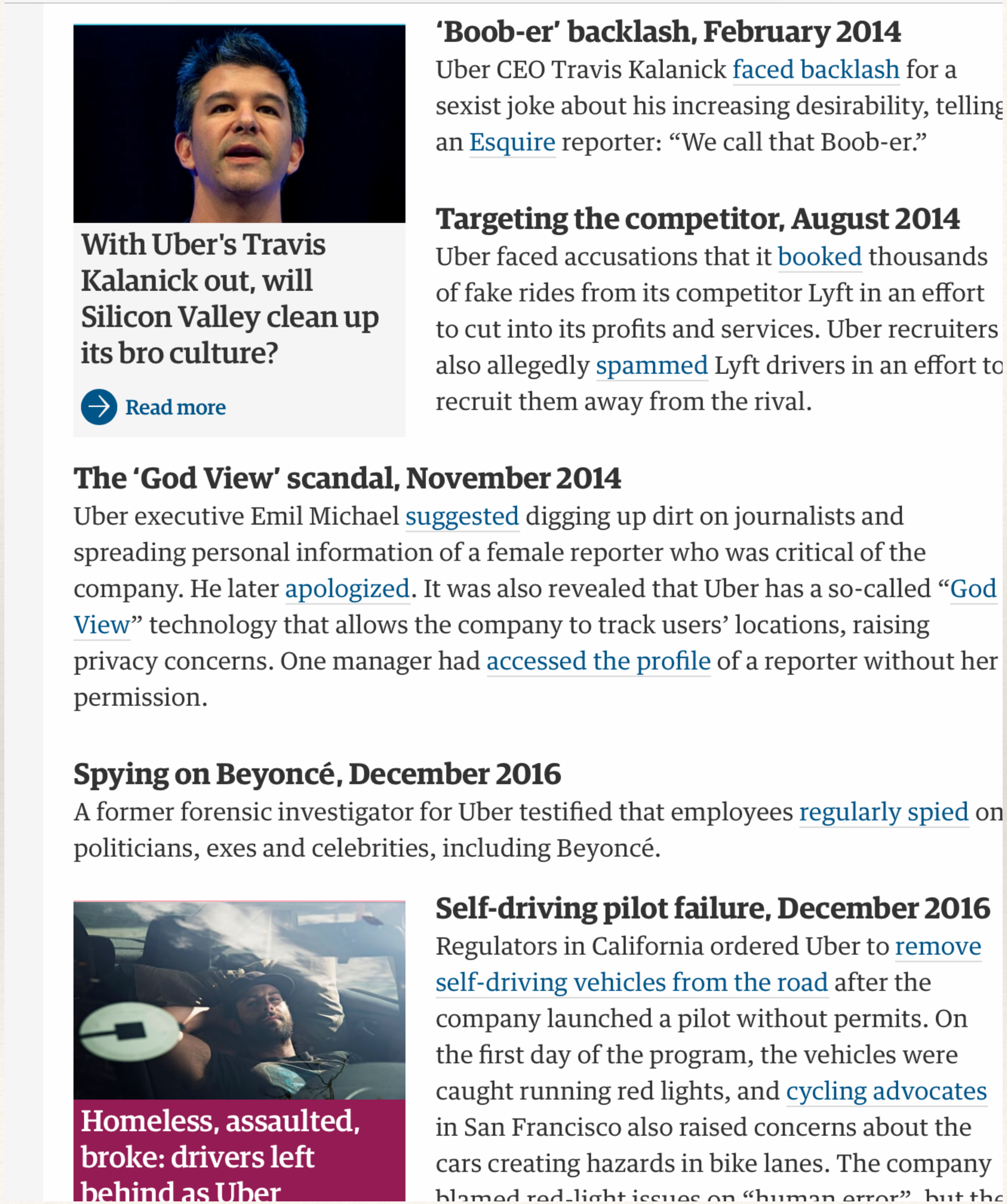


#BUSINESS NEWS NOVEMBER 28, 2017 / 8:34 AM / A MONTH AGO

Uber-Waymo trial delayed as U.S. judge raises prospect of 'cover-up'

Heather Somerville, Dan Levine 7 MIN READ

SAN FRANCISCO (Reuters) - Uber Technologies Inc withheld evidence in a lawsuit filed by Alphabet Inc's Waymo, a U.S. judge said on Tuesday, delaying a trial to give Waymo time to review a letter alleging that Uber trained employees to steal trade secrets and hide their tracks.



With Uber's Travis Kalanick out, will Silicon Valley clean up its bro culture?

Read more

'Boob-er' backlash, February 2014

Uber CEO Travis Kalanick faced backlash for a sexist joke about his increasing desirability, telling an Esquire reporter: "We call that Boob-er."

Targeting the competitor, August 2014

Uber faced accusations that it booked thousands of fake rides from its competitor Lyft in an effort to cut into its profits and services. Uber recruiters also allegedly spammed Lyft drivers in an effort to recruit them away from the rival.

The 'God View' scandal, November 2014

Uber executive Emil Michael suggested digging up dirt on journalists and spreading personal information of a female reporter who was critical of the company. He later apologized. It was also revealed that Uber has a so-called "God View" technology that allows the company to track users' locations, raising privacy concerns. One manager had accessed the profile of a reporter without her permission.

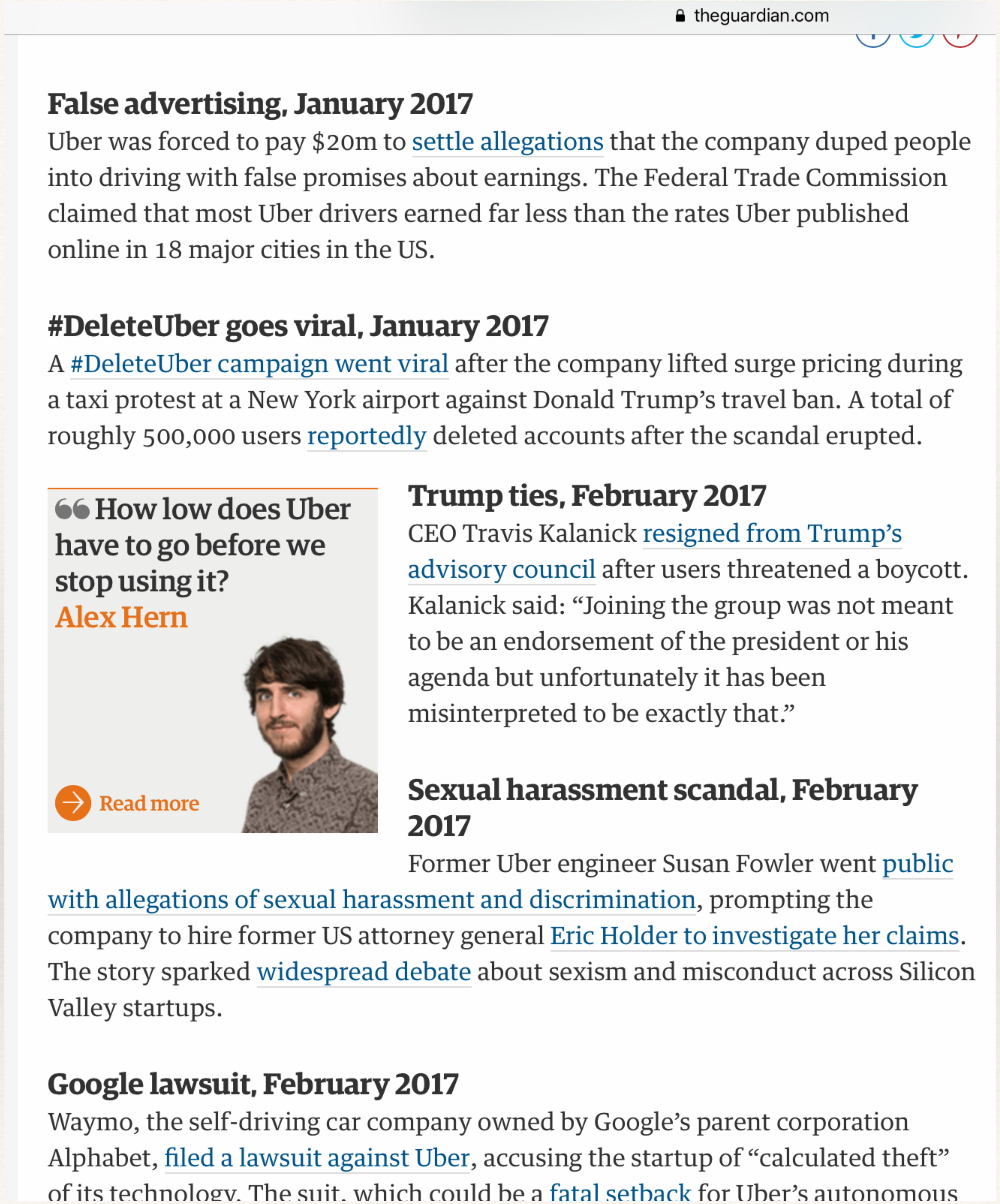
Spying on Beyoncé, December 2016

A former forensic investigator for Uber testified that employees regularly spied on politicians, exes and celebrities, including Beyoncé.

Homeless, assaulted, broke: drivers left behind as Uber

Self-driving pilot failure, December 2016

Regulators in California ordered Uber to remove self-driving vehicles from the road after the company launched a pilot without permits. On the first day of the program, the vehicles were caught running red lights, and cycling advocates in San Francisco also raised concerns about the cars creating hazards in bike lanes. The company blamed red-light issues on "human error", but the



theguardian.com

False advertising, January 2017

Uber was forced to pay $20m to settle allegations that the company duped people into driving with false promises about earnings. The Federal Trade Commission claimed that most Uber drivers earned far less than the rates Uber published online in 18 major cities in the US.

#DeleteUber goes viral, January 2017

A #DeleteUber campaign went viral after the company lifted surge pricing during a taxi protest at a New York airport against Donald Trump's travel ban. A total of roughly 500,000 users reportedly deleted accounts after the scandal erupted.

How low does Uber have to go before we stop using it?
Alex Hern

Read more

Trump ties, February 2017

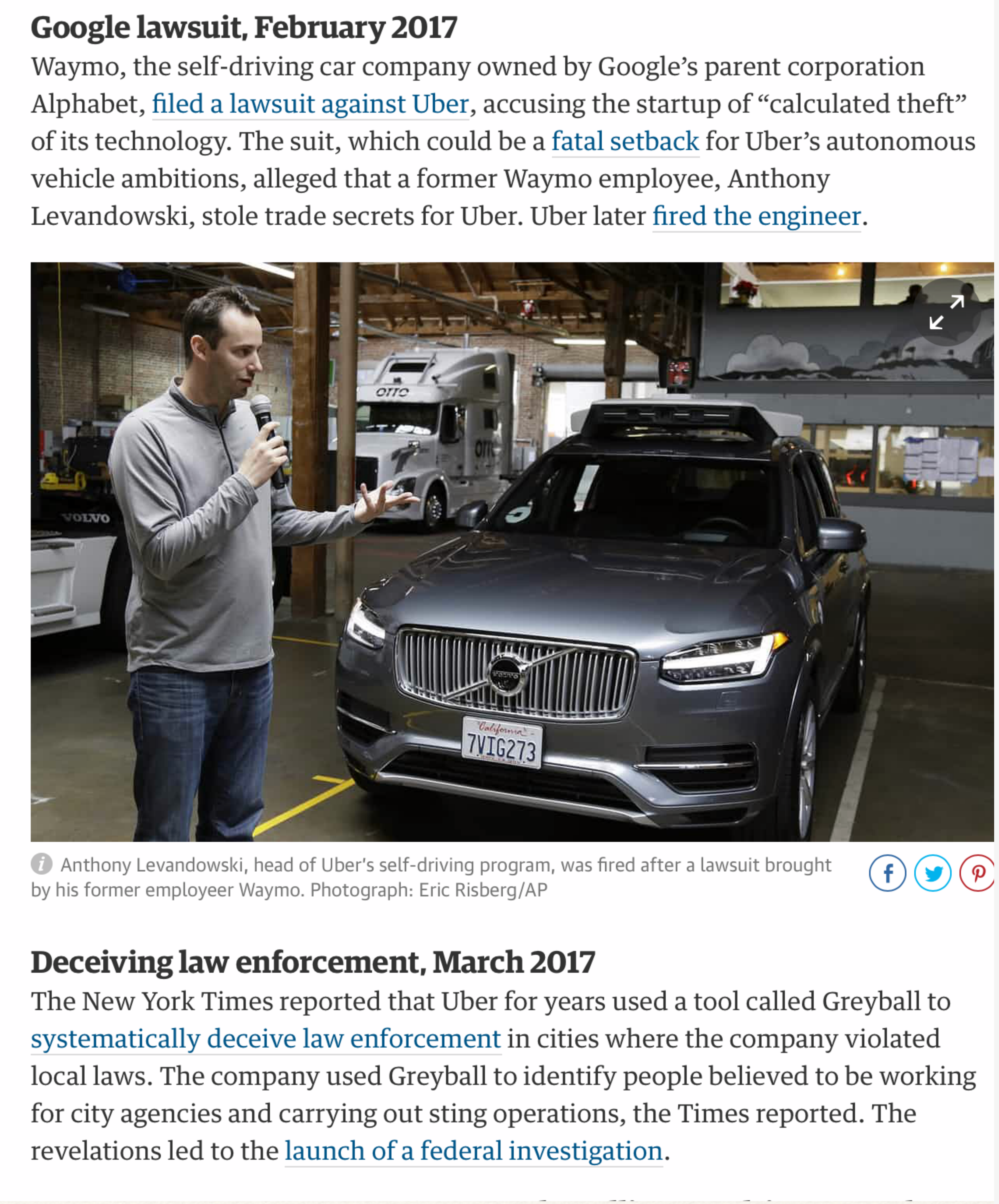
CEO Travis Kalanick resigned from Trump's advisory council after users threatened a boycott. Kalanick said: "Joining the group was not meant to be an endorsement of the president or his agenda but unfortunately it has been misinterpreted to be exactly that."

Sexual harassment scandal, February 2017

Former Uber engineer Susan Fowler went public with allegations of sexual harassment and discrimination, prompting the company to hire former US attorney general Eric Holder to investigate her claims. The story sparked widespread debate about sexism and misconduct across Silicon Valley startups.

Google lawsuit, February 2017

Waymo, the self-driving car company owned by Google's parent corporation Alphabet, filed a lawsuit against Uber, accusing the startup of "calculated theft" of its technology. The suit, which could be a fatal setback for Uber's autonomous



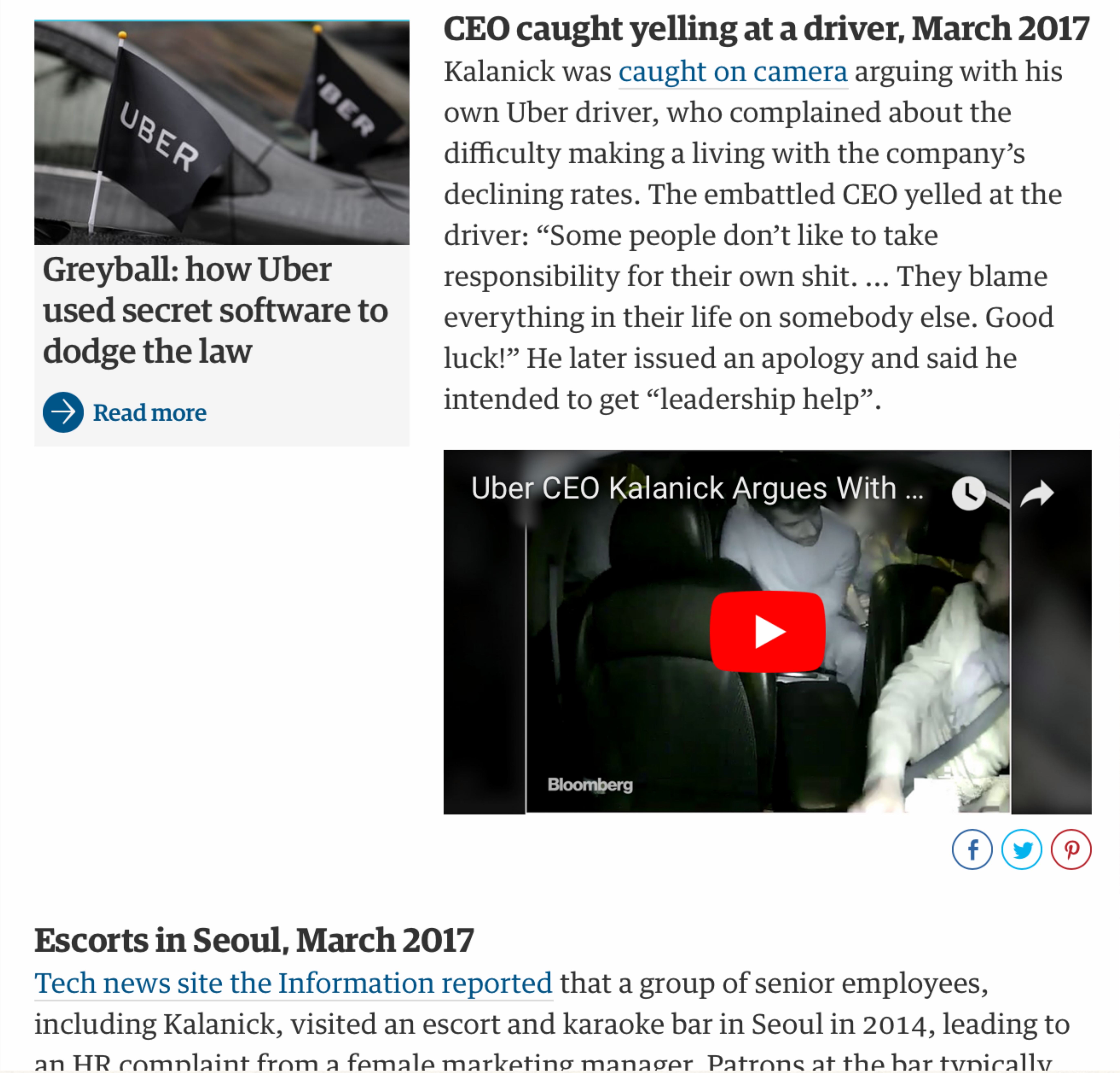
Google lawsuit, February 2017

Waymo, the self-driving car company owned by Google's parent corporation Alphabet, filed a lawsuit against Uber, accusing the startup of "calculated theft" of its technology. The suit, which could be a fatal setback for Uber's autonomous vehicle ambitions, alleged that a former Waymo employee, Anthony Levandowski, stole trade secrets for Uber. Uber later fired the engineer.

Anthony Levandowski, head of Uber's self-driving program, was fired after a lawsuit brought by his former employeer Waymo. Photograph: Eric Risberg/AP

Deceiving law enforcement, March 2017

The New York Times reported that Uber for years used a tool called Greyball to systematically deceive law enforcement in cities where the company violated local laws. The company used Greyball to identify people believed to be working for city agencies and carrying out sting operations, the Times reported. The revelations led to the launch of a federal investigation.



Greyball: how Uber used secret software to dodge the law

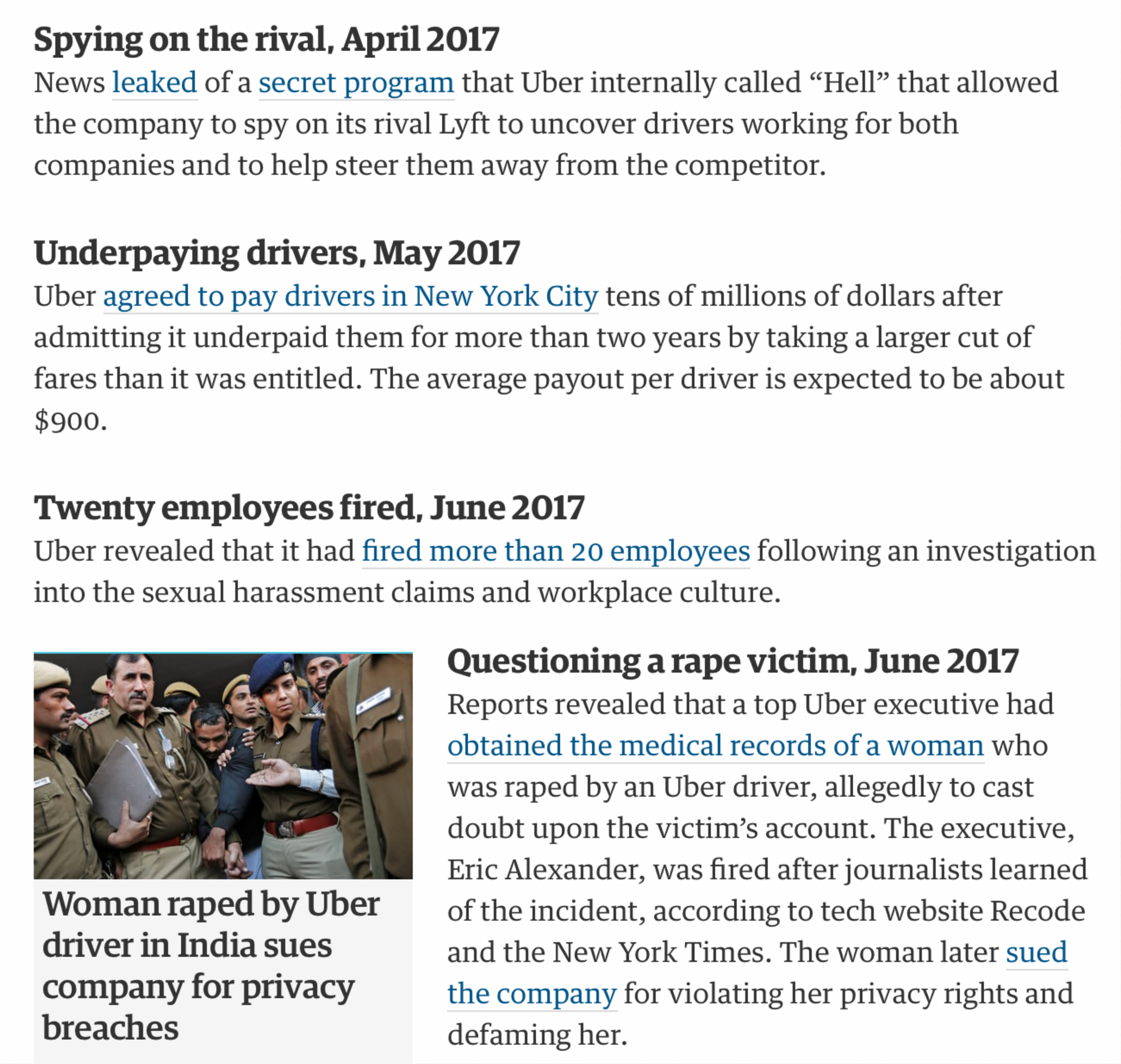
Read more

CEO caught yelling at a driver, March 2017

Kalanick was caught on camera arguing with his own Uber driver, who complained about the difficulty making a living with the company's declining rates. The embattled CEO yelled at the driver: "Some people don't like to take responsibility for their own shit. … They blame everything in their life on somebody else. Good luck!" He later issued an apology and said he intended to get "leadership help".

Escorts in Seoul, March 2017

Tech news site the Information reported that a group of senior employees, including Kalanick, visited an escort and karaoke bar in Seoul in 2014, leading to an HR complaint from a female marketing manager. Patrons at the bar typically



Spying on the rival, April 2017

News leaked of a secret program that Uber internally called “Hell” that allowed the company to spy on its rival Lyft to uncover drivers working for both companies and to help steer them away from the competitor.

Underpaying drivers, May 2017

Uber agreed to pay drivers in New York City tens of millions of dollars after admitting it underpaid them for more than two years by taking a larger cut of fares than it was entitled. The average payout per driver is expected to be about $900.

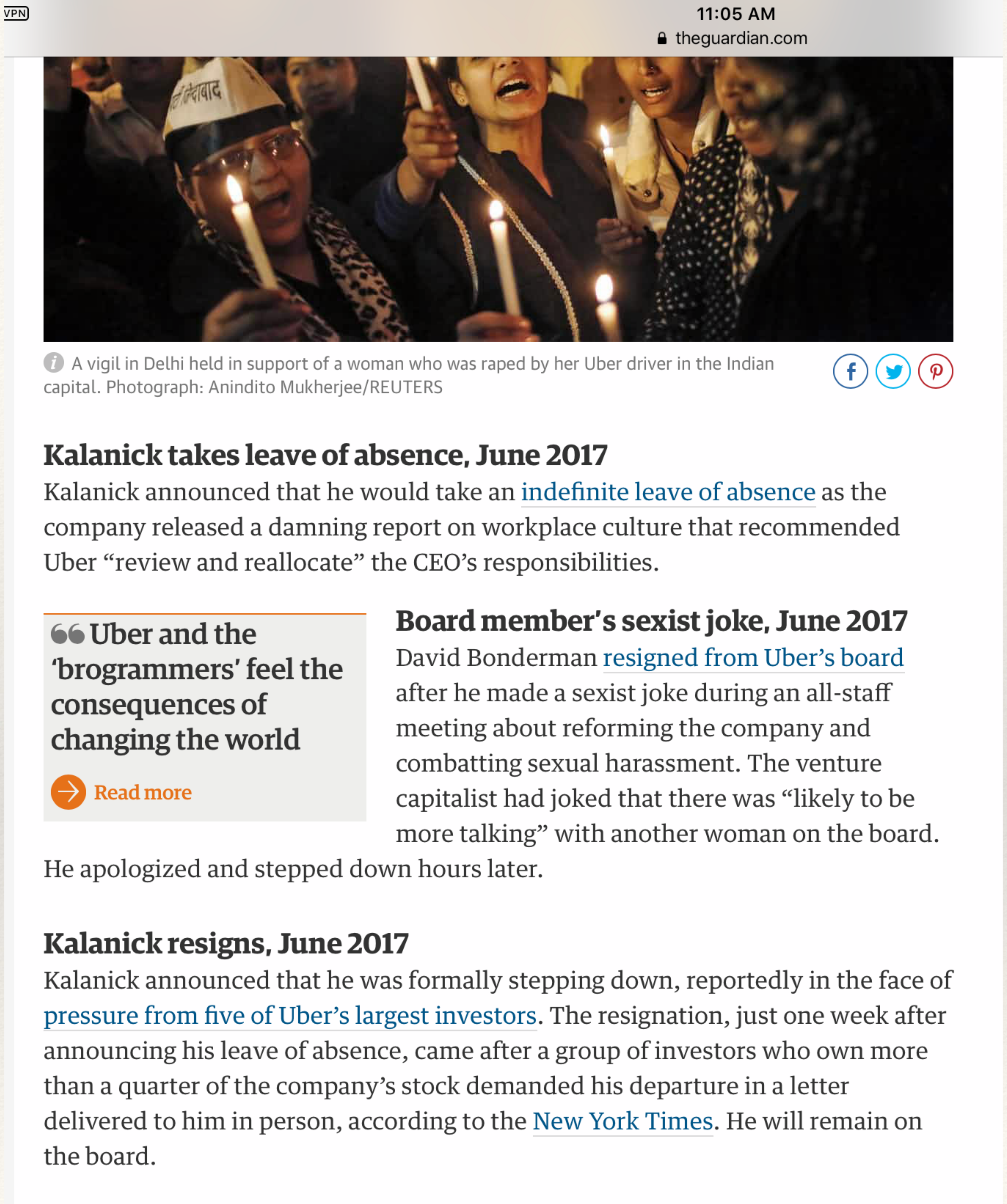
Twenty employees fired, June 2017

Uber revealed that it had fired more than 20 employees following an investigation into the sexual harassment claims and workplace culture.

Woman raped by Uber driver in India sues company for privacy breaches

Questioning a rape victim, June 2017

Reports revealed that a top Uber executive had obtained the medical records of a woman who was raped by an Uber driver, allegedly to cast doubt upon the victim’s account. The executive, Eric Alexander, was fired after journalists learned of the incident, according to tech website Recode and the New York Times. The woman later sued the company for violating her privacy rights and defaming her.



A vigil in Delhi held in support of a woman who was raped by her Uber driver in the Indian capital. Photograph: Anindito Mukherjee/REUTERS

Kalanick takes leave of absence, June 2017

Kalanick announced that he would take an indefinite leave of absence as the company released a damning report on workplace culture that recommended Uber “review and reallocate” the CEO’s responsibilities.

> Uber and the ‘brogrammers’ feel the consequences of changing the world
>
> Read more

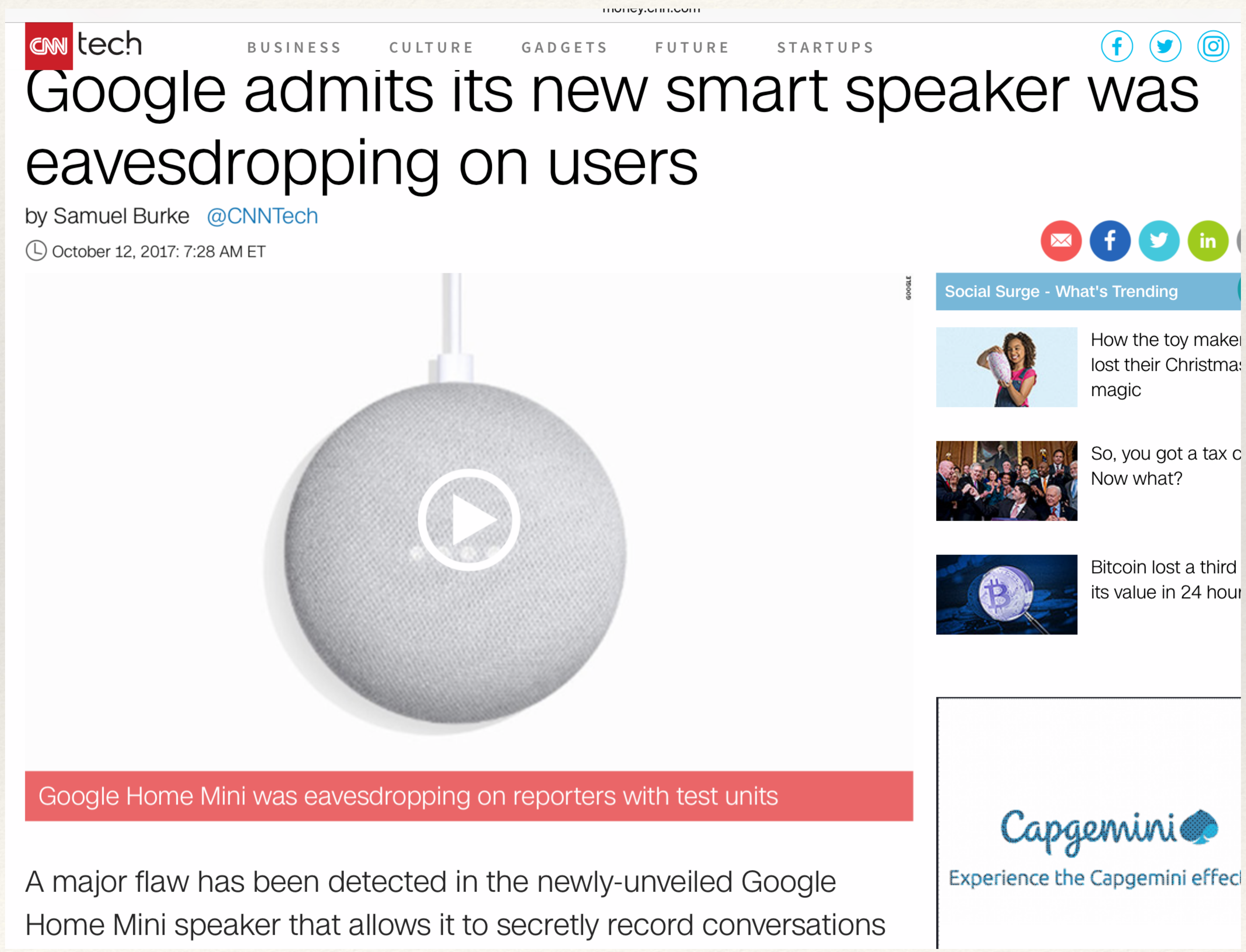
Board member’s sexist joke, June 2017

David Bonderman resigned from Uber’s board after he made a sexist joke during an all-staff meeting about reforming the company and combatting sexual harassment. The venture capitalist had joked that there was “likely to be more talking” with another woman on the board. He apologized and stepped down hours later.

Kalanick resigns, June 2017

Kalanick announced that he was formally stepping down, reportedly in the face of pressure from five of Uber’s largest investors. The resignation, just one week after announcing his leave of absence, came after a group of investors who own more than a quarter of the company’s stock demanded his departure in a letter delivered to him in person, according to the New York Times. He will remain on the board.

Vulnerabilities: home automation



money.cnn.com

CNN tech BUSINESS CULTURE GADGETS FUTURE STARTUPS

Google admits its new smart speaker was eavesdropping on users

by Samuel Burke @CNNTech

October 12, 2017: 7:28 AM ET

Google Home Mini was eavesdropping on reporters with test units

Social Surge - What's Trending

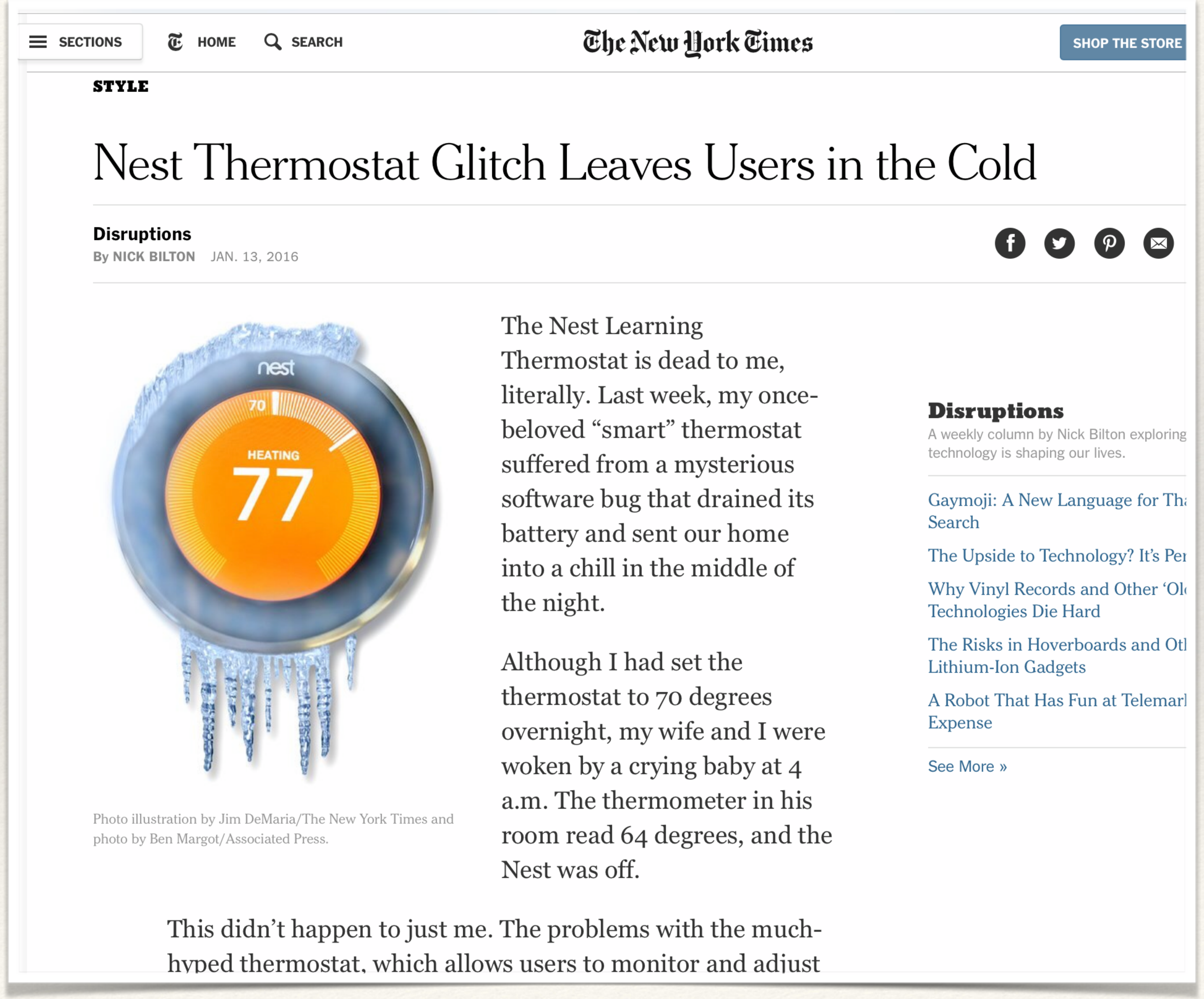
How the toy maker lost their Christmas magic

So, you got a tax c Now what?

Bitcoin lost a third its value in 24 hour

Capgemini Experience the Capgemini effec

A major flaw has been detected in the newly-unveiled Google Home Mini speaker that allows it to secretly record conversations



SECTIONS HOME SEARCH The New York Times SHOP THE STORE

STYLE

Nest Thermostat Glitch Leaves Users in the Cold

Disruptions

By NICK BILTON JAN. 13, 2016

Photo illustration by Jim DeMaria/The New York Times and photo by Ben Margot/Associated Press.

The Nest Learning Thermostat is dead to me, literally. Last week, my once-beloved “smart” thermostat suffered from a mysterious software bug that drained its battery and sent our home into a chill in the middle of the night.

Although I had set the thermostat to 70 degrees overnight, my wife and I were woken by a crying baby at 4 a.m. The thermometer in his room read 64 degrees, and the Nest was off.

This didn’t happen to just me. The problems with the much-hyped thermostat, which allows users to monitor and adjust

Disruptions

A weekly column by Nick Bilton exploring technology is shaping our lives.

Gaymoji: A New Language for Tha Search

The Upside to Technology? It’s Per

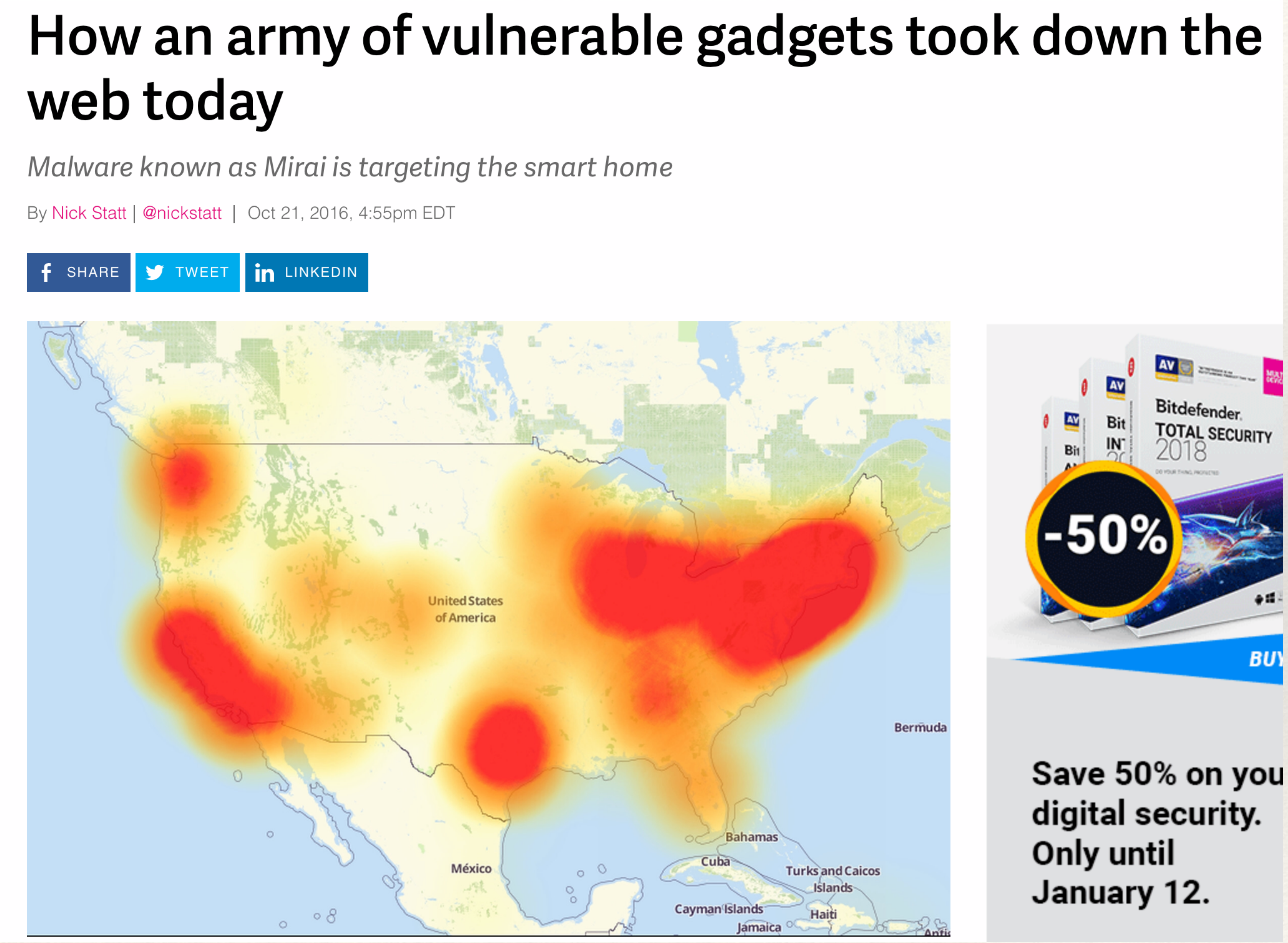
Why Vinyl Records and Other ‘Ol Technologies Die Hard

The Risks in Hoverboards and Ot Lithium-Ion Gadgets

A Robot That Has Fun at Telemar Expense

See More »

Vulnerabilities: IoT

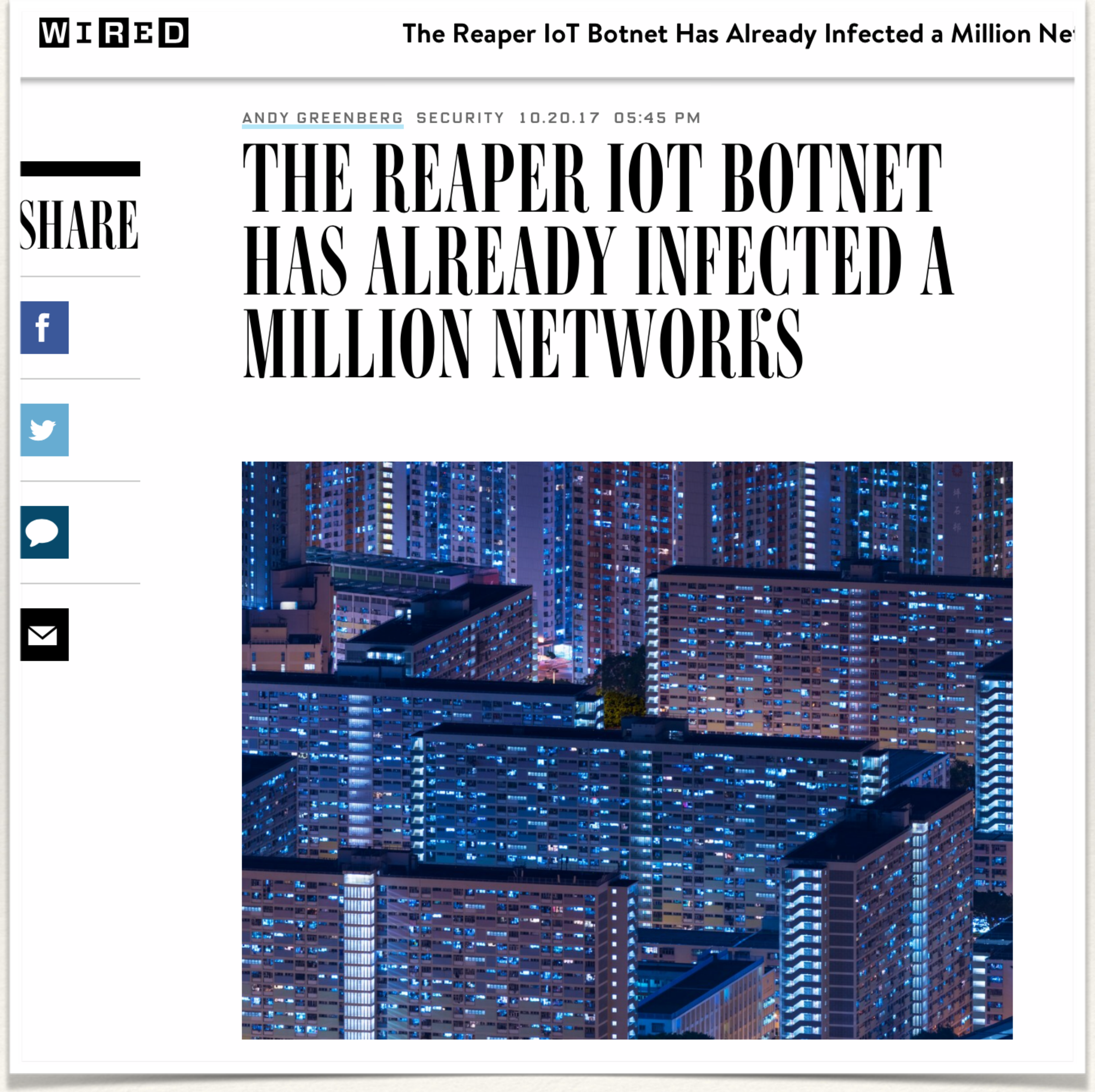


How an army of vulnerable gadgets took down the web today

Malware known as Mirai is targeting the smart home

By Nick Statt | @nickstatt | Oct 21, 2016, 4:55pm EDT

SHARE TWEET LINKEDIN



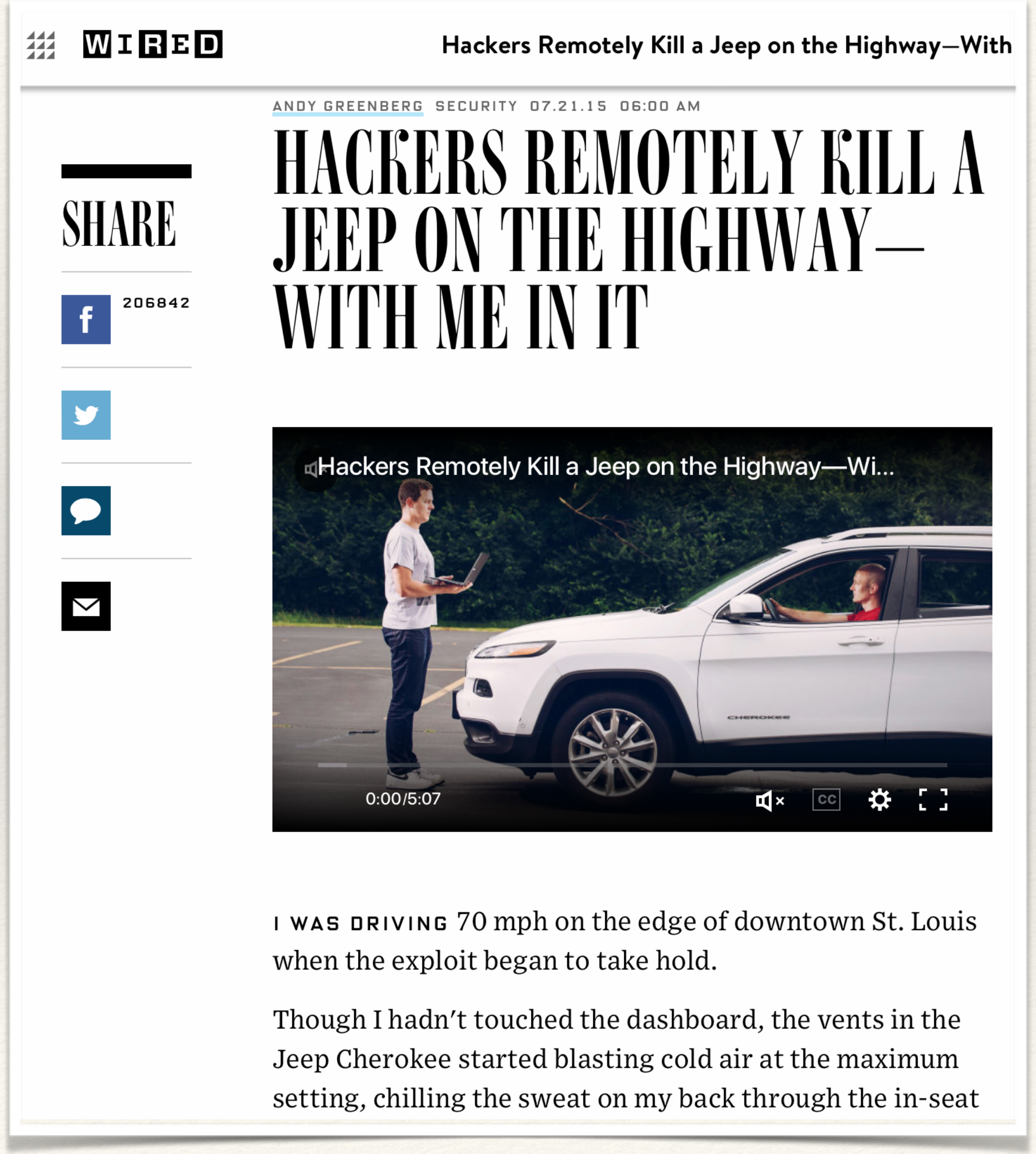
WIRED The Reaper IoT Botnet Has Already Infected a Million Ne

SHARE

ANDY GREENBERG SECURITY 10.20.17 05:45 PM

THE REAPER IOT BOTNET HAS ALREADY INFECTED A MILLION NETWORKS

Vulnerabilities: cars


WIRED Hackers Remotely Kill a Jeep on the Highway—With

ANDY GREENBERG SECURITY 07.21.15 06:00 AM

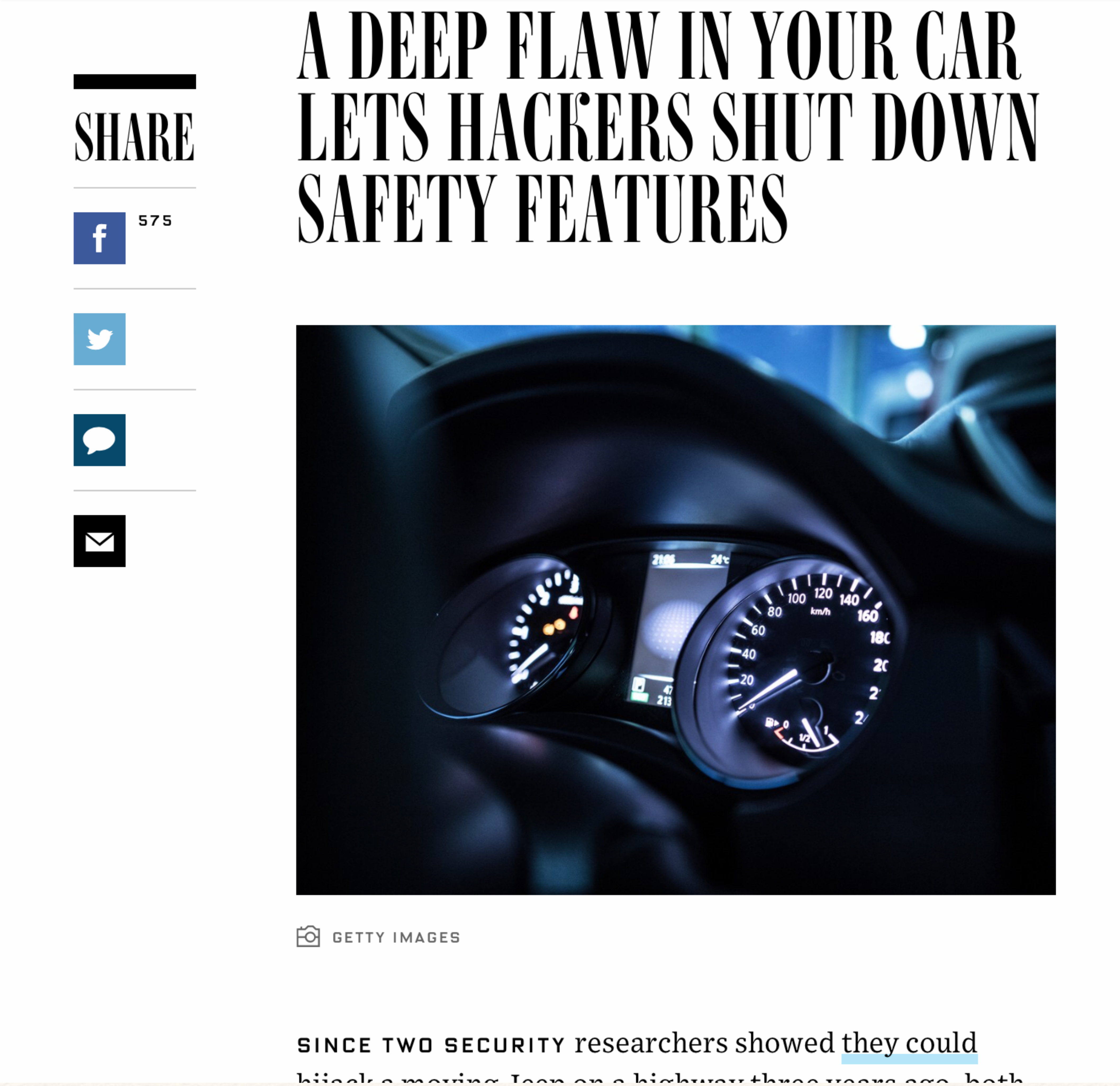
SHARE

206842

HACKERS REMOTELY KILL A JEEP ON THE HIGHWAY—WITH ME IN IT

I WAS DRIVING 70 mph on the edge of downtown St. Louis when the exploit began to take hold.

Though I hadn't touched the dashboard, the vents in the Jeep Cherokee started blasting cold air at the maximum setting, chilling the sweat on my back through the in-seat


SHARE

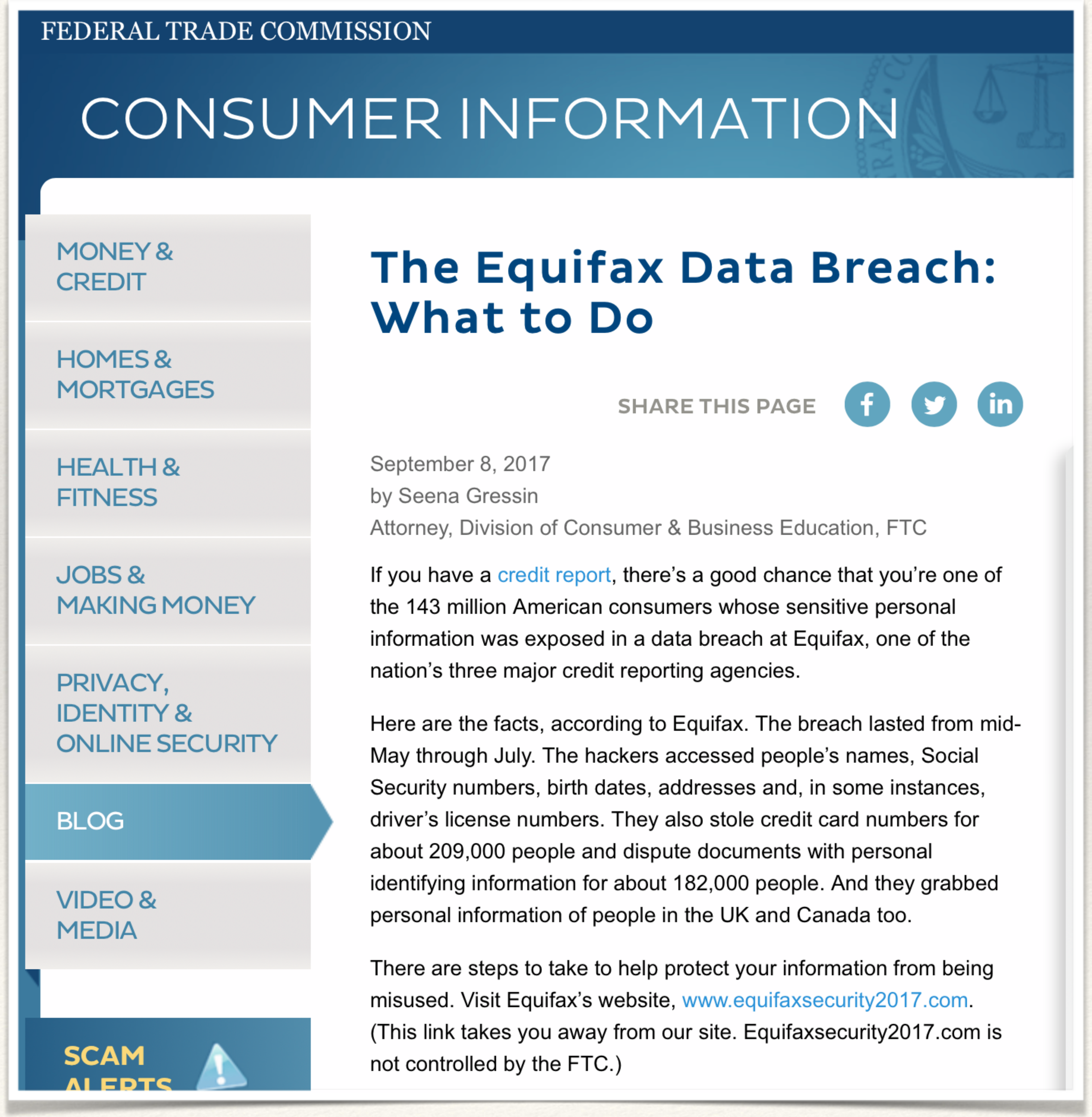
575

A DEEP FLAW IN YOUR CAR LETS HACKERS SHUT DOWN SAFETY FEATURES

GETTY IMAGES

SINCE TWO SECURITY researchers showed they could hijack a moving Jeep on a highway three years ago, both

Vulnerabilities: financial, social



FEDERAL TRADE COMMISSION

CONSUMER INFORMATION

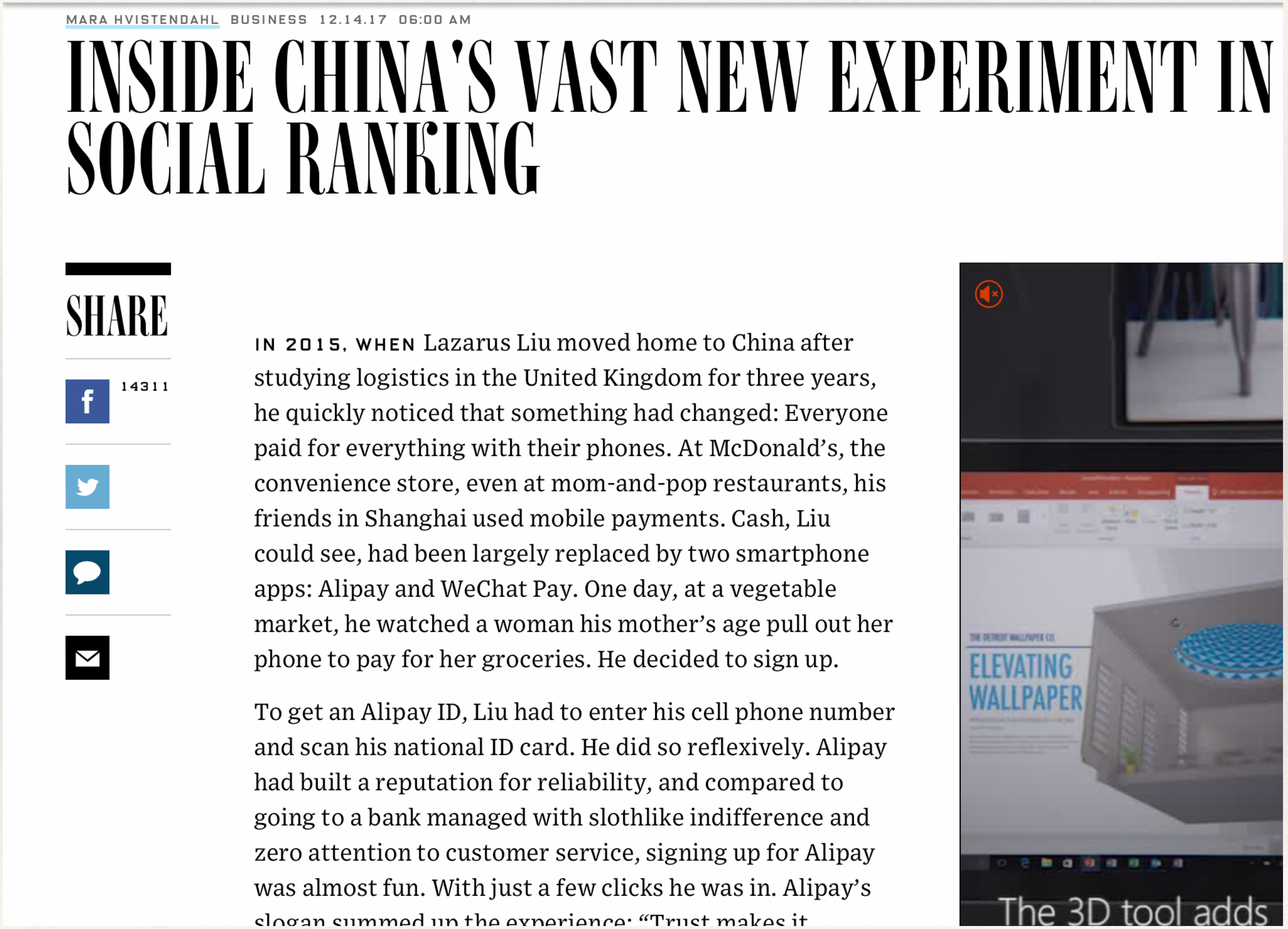
The Equifax Data Breach: What to Do

September 8, 2017
by Seena Gressin
Attorney, Division of Consumer & Business Education, FTC

If you have a credit report, there's a good chance that you're one of the 143 million American consumers whose sensitive personal information was exposed in a data breach at Equifax, one of the nation's three major credit reporting agencies.

Here are the facts, according to Equifax. The breach lasted from mid-May through July. The hackers accessed people's names, Social Security numbers, birth dates, addresses and, in some instances, driver's license numbers. They also stole credit card numbers for about 209,000 people and dispute documents with personal identifying information for about 182,000 people. And they grabbed personal information of people in the UK and Canada too.

There are steps to take to help protect your information from being misused. Visit Equifax's website, www.equifaxsecurity2017.com. (This link takes you away from our site. Equifaxsecurity2017.com is not controlled by the FTC.)



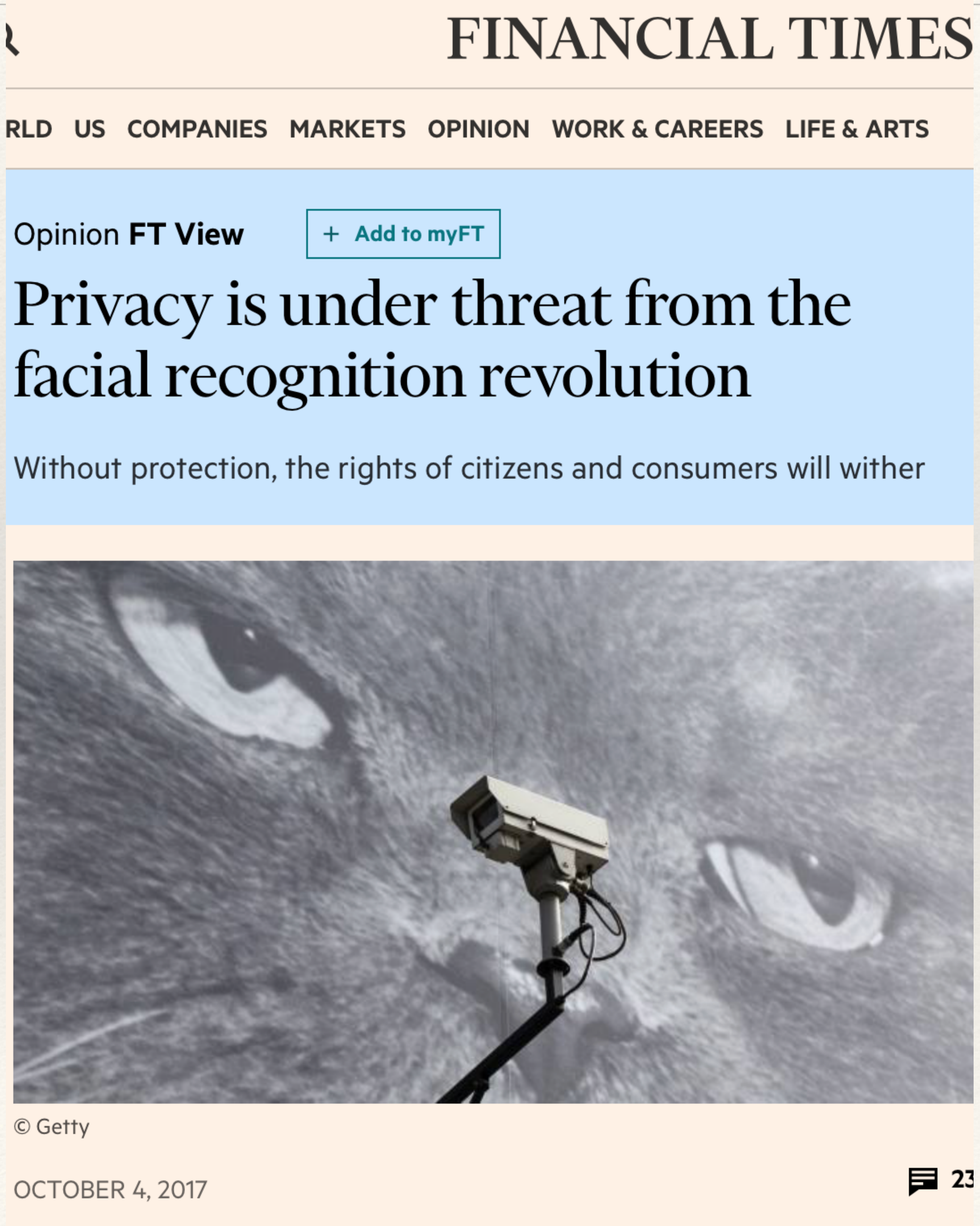
MARA HVISTENDAHL BUSINESS 12.14.17 06:00 AM

INSIDE CHINA'S VAST NEW EXPERIMENT IN SOCIAL RANKING

IN 2015, WHEN Lazarus Liu moved home to China after studying logistics in the United Kingdom for three years, he quickly noticed that something had changed: Everyone paid for everything with their phones. At McDonald's, the convenience store, even at mom-and-pop restaurants, his friends in Shanghai used mobile payments. Cash, Liu could see, had been largely replaced by two smartphone apps: Alipay and WeChat Pay. One day, at a vegetable market, he watched a woman his mother's age pull out her phone to pay for her groceries. He decided to sign up.

To get an Alipay ID, Liu had to enter his cell phone number and scan his national ID card. He did so reflexively. Alipay had built a reputation for reliability, and compared to going to a bank managed with slothlike indifference and zero attention to customer service, signing up for Alipay was almost fun. With just a few clicks he was in. Alipay's slogan summed up the experience: "Trust makes it

Vulnerabilities: privacy



FINANCIAL TIMES

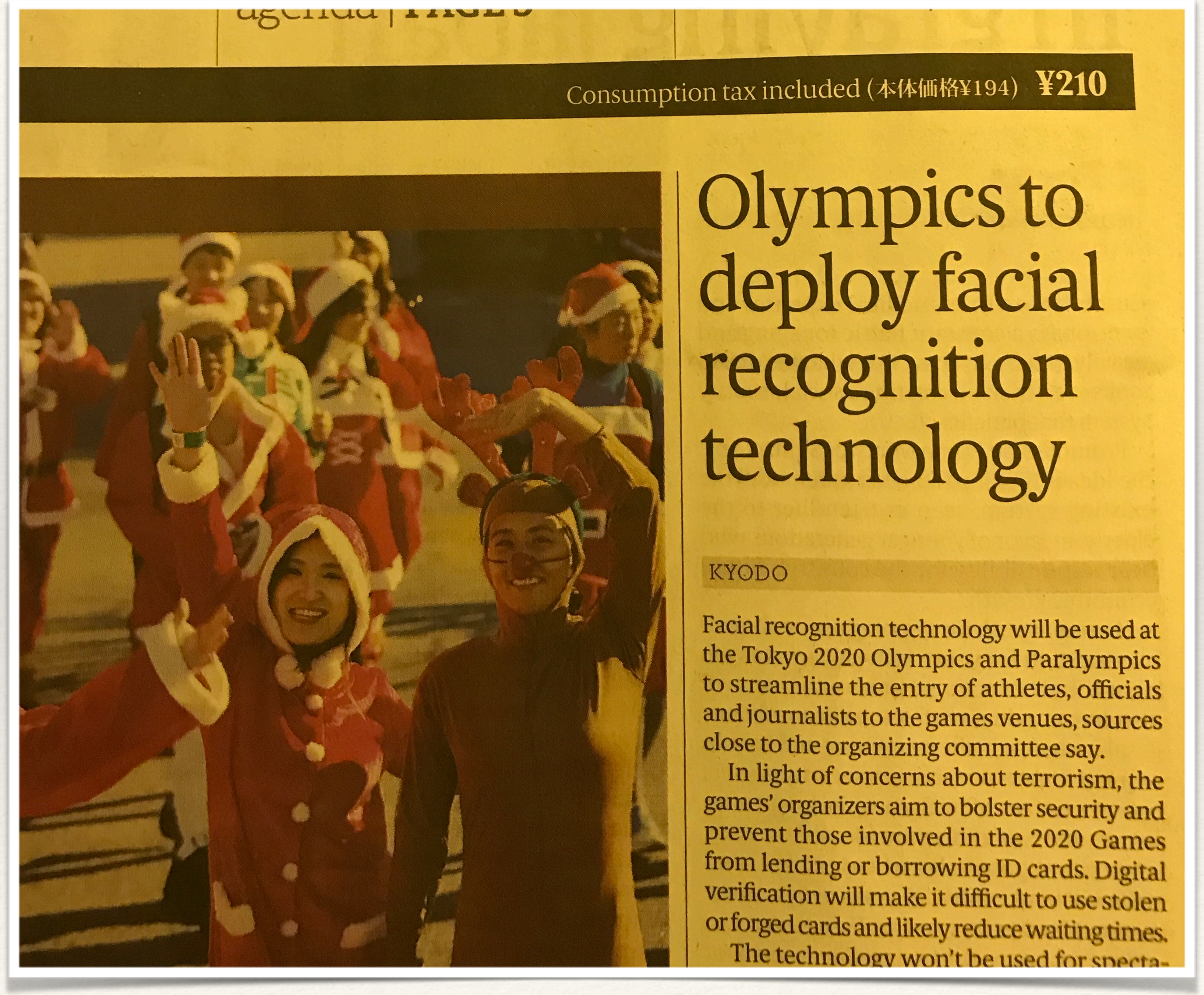
Opinion **FT View**

Privacy is under threat from the facial recognition revolution

Without protection, the rights of citizens and consumers will wither

© Getty

OCTOBER 4, 2017



agenda |

Consumption tax included (本体価格¥194) ¥210

Olympics to deploy facial recognition technology

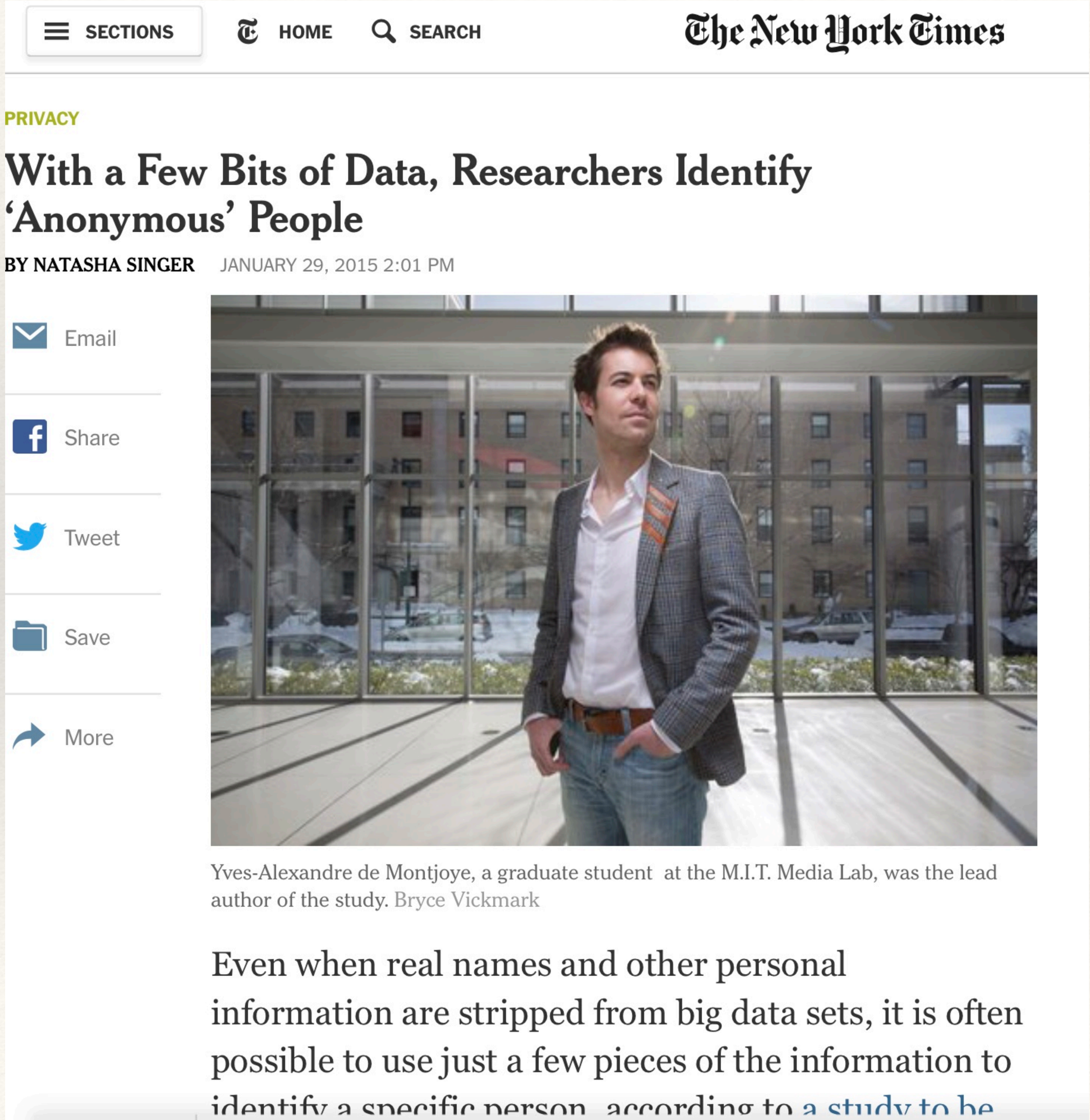
KYODO

Facial recognition technology will be used at the Tokyo 2020 Olympics and Paralympics to streamline the entry of athletes, officials and journalists to the games venues, sources close to the organizing committee say.

In light of concerns about terrorism, the games' organizers aim to bolster security and prevent those involved in the 2020 Games from lending or borrowing ID cards. Digital verification will make it difficult to use stolen or forged cards and likely reduce waiting times.

The technology won't be used for specta-

Vulnerabilities: privacy



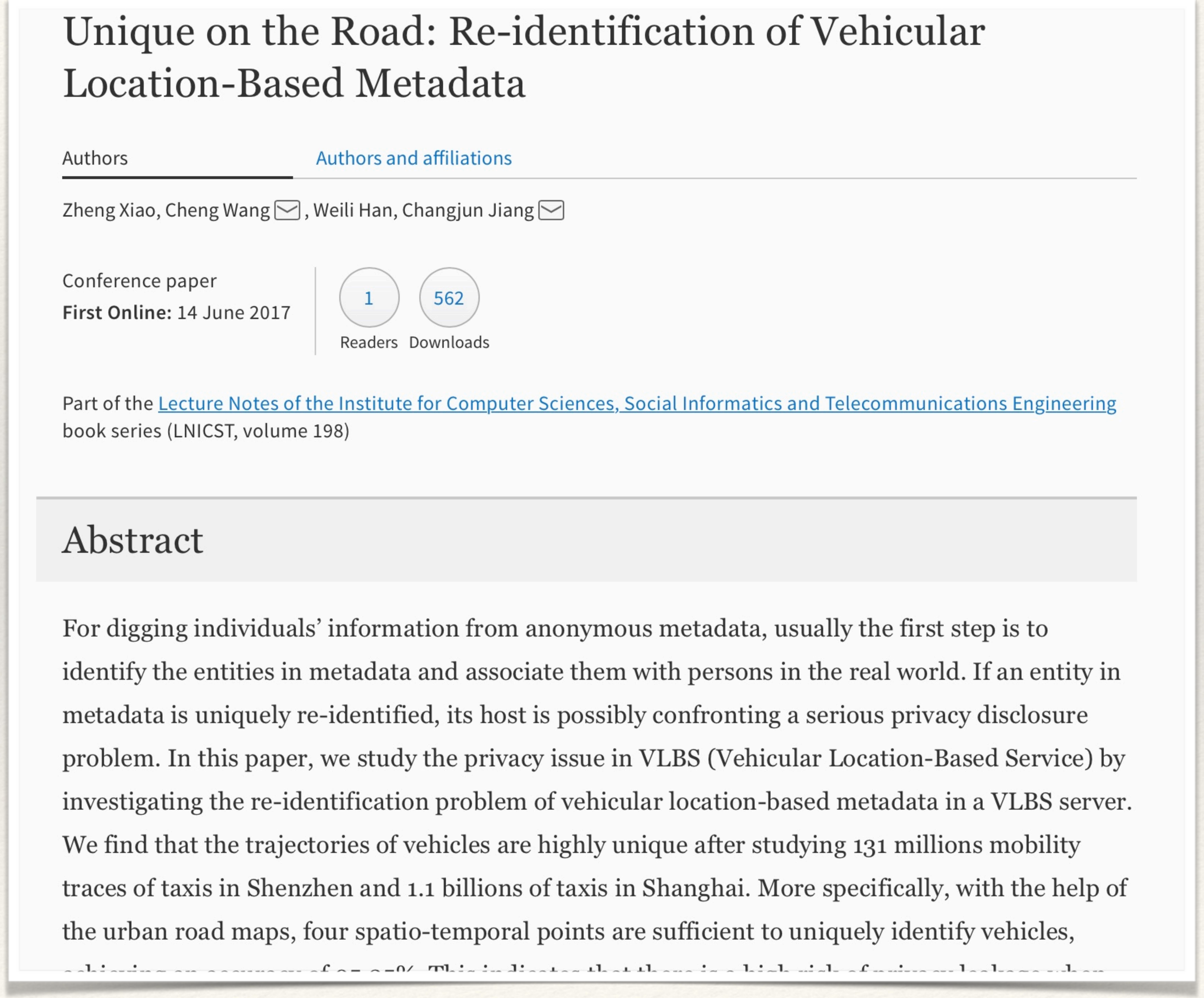
PRIVACY

With a Few Bits of Data, Researchers Identify 'Anonymous' People

BY NATASHA SINGER JANUARY 29, 2015 2:01 PM

Yves-Alexandre de Montjoye, a graduate student at the M.I.T. Media Lab, was the lead author of the study. Bryce Vickmark

Even when real names and other personal information are stripped from big data sets, it is often possible to use just a few pieces of the information to identify a specific person, according to a study to be



Unique on the Road: Re-identification of Vehicular Location-Based Metadata

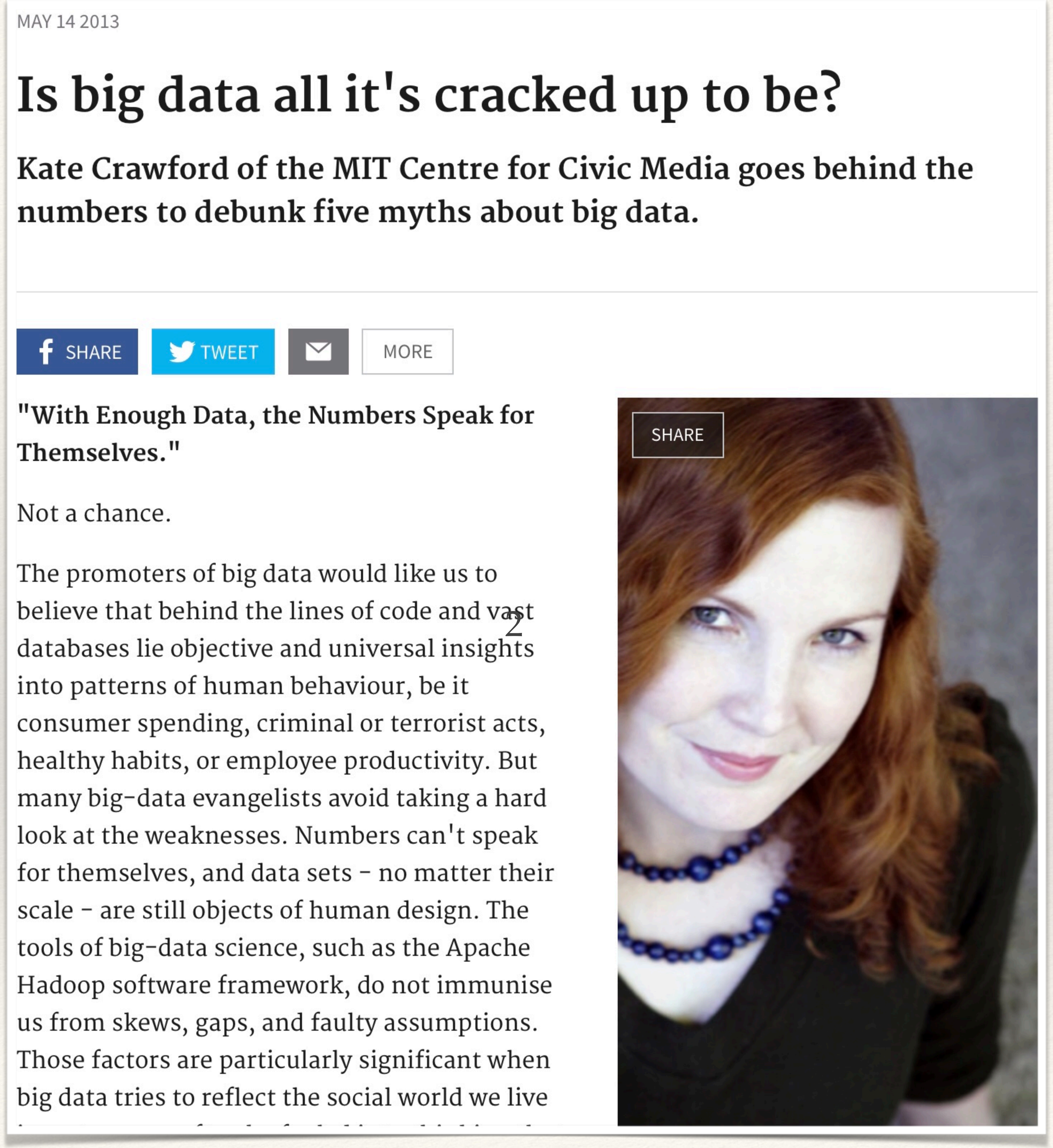
Zheng Xiao, Cheng Wang, Weili Han, Changjun Jiang

Conference paper
First Online: 14 June 2017

Part of the Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering book series (LNICST, volume 198)

Abstract

For digging individuals' information from anonymous metadata, usually the first step is to identify the entities in metadata and associate them with persons in the real world. If an entity in metadata is uniquely re-identified, its host is possibly confronting a serious privacy disclosure problem. In this paper, we study the privacy issue in VLBS (Vehicular Location-Based Service) by investigating the re-identification problem of vehicular location-based metadata in a VLBS server. We find that the trajectories of vehicles are highly unique after studying 131 millions mobility traces of taxis in Shenzhen and 1.1 billions of taxis in Shanghai. More specifically, with the help of the urban road maps, four spatio-temporal points are sufficient to uniquely identify vehicles,



MAY 14 2013

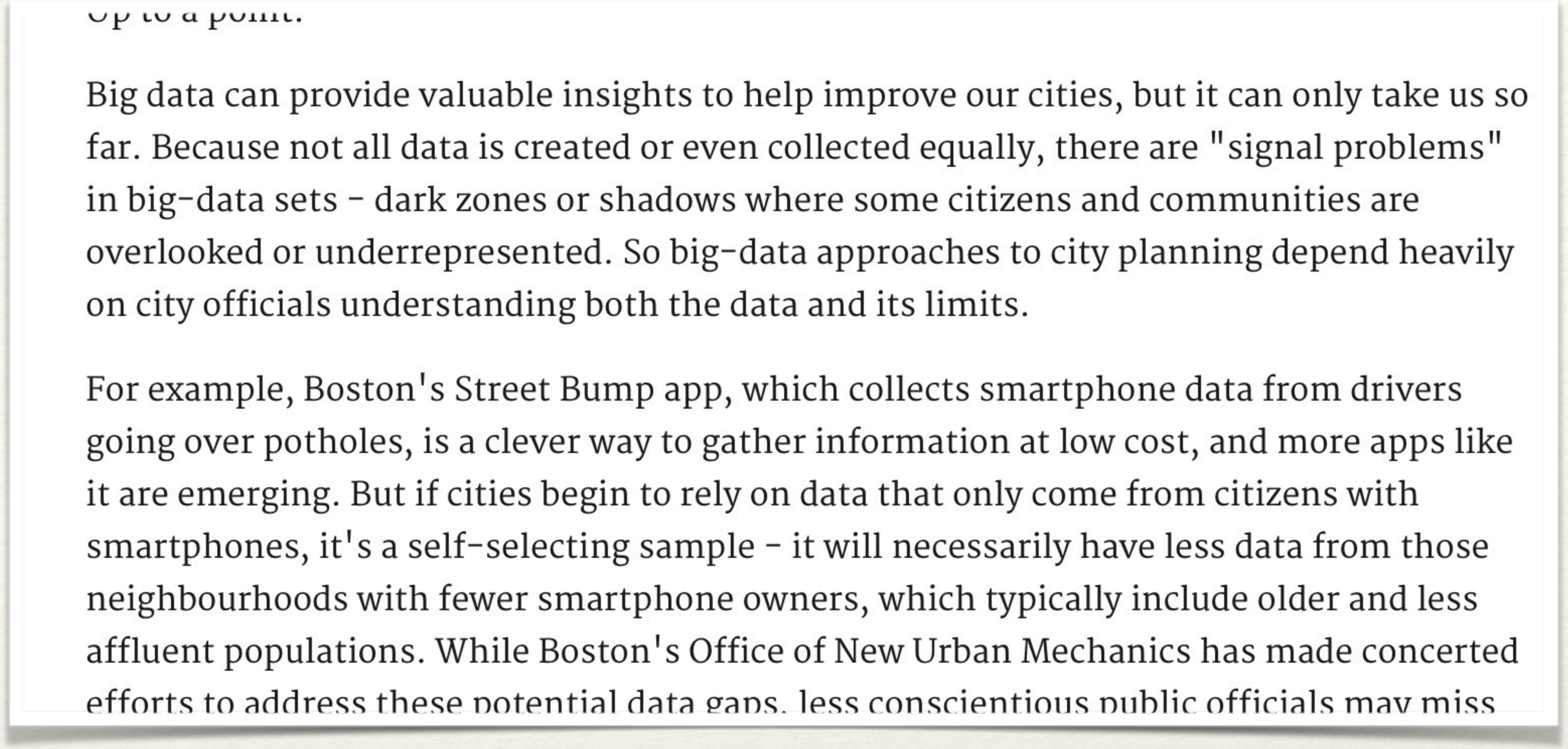
Is big data all it's cracked up to be?

Kate Crawford of the MIT Centre for Civic Media goes behind the numbers to debunk five myths about big data.

"With Enough Data, the Numbers Speak for Themselves."

Not a chance.

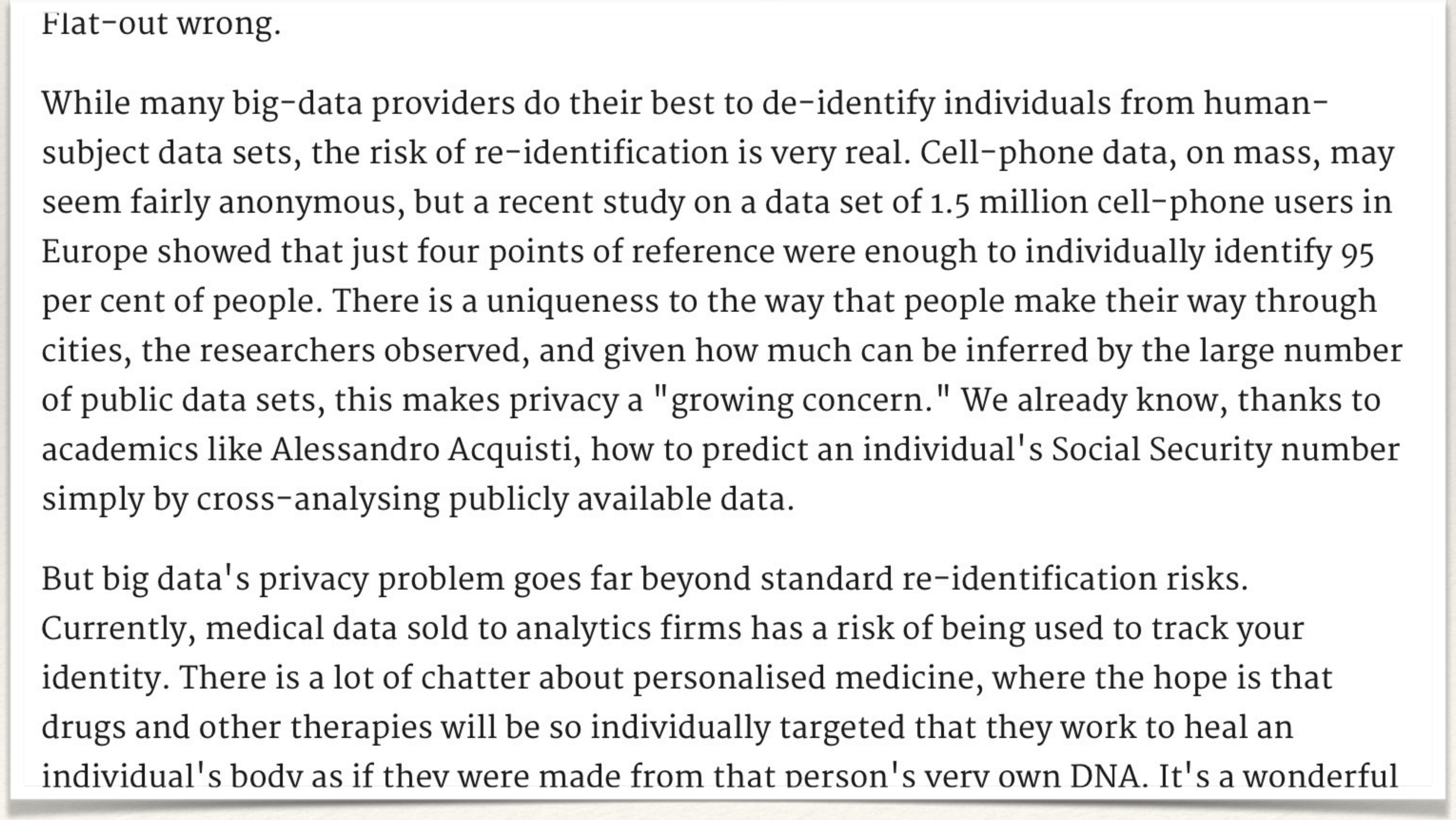
The promoters of big data would like us to believe that behind the lines of code and vast databases lie objective and universal insights into patterns of human behaviour, be it consumer spending, criminal or terrorist acts, healthy habits, or employee productivity. But many big-data evangelists avoid taking a hard look at the weaknesses. Numbers can't speak for themselves, and data sets - no matter their scale - are still objects of human design. The tools of big-data science, such as the Apache Hadoop software framework, do not immunise us from skews, gaps, and faulty assumptions. Those factors are particularly significant when big data tries to reflect the social world we live



Up to a point.

Big data can provide valuable insights to help improve our cities, but it can only take us so far. Because not all data is created or even collected equally, there are "signal problems" in big-data sets - dark zones or shadows where some citizens and communities are overlooked or underrepresented. So big-data approaches to city planning depend heavily on city officials understanding both the data and its limits.

For example, Boston's Street Bump app, which collects smartphone data from drivers going over potholes, is a clever way to gather information at low cost, and more apps like it are emerging. But if cities begin to rely on data that only come from citizens with smartphones, it's a self-selecting sample - it will necessarily have less data from those neighbourhoods with fewer smartphone owners, which typically include older and less affluent populations. While Boston's Office of New Urban Mechanics has made concerted efforts to address these potential data gaps, less conscientious public officials may miss



Flat-out wrong.

While many big-data providers do their best to de-identify individuals from human-subject data sets, the risk of re-identification is very real. Cell-phone data, on mass, may seem fairly anonymous, but a recent study on a data set of 1.5 million cell-phone users in Europe showed that just four points of reference were enough to individually identify 95 per cent of people. There is a uniqueness to the way that people make their way through cities, the researchers observed, and given how much can be inferred by the large number of public data sets, this makes privacy a "growing concern." We already know, thanks to academics like Alessandro Acquisti, how to predict an individual's Social Security number simply by cross-analysing publicly available data.

But big data's privacy problem goes far beyond standard re-identification risks. Currently, medical data sold to analytics firms has a risk of being used to track your identity. There is a lot of chatter about personalised medicine, where the hope is that drugs and other therapies will be so individually targeted that they work to heal an individual's body as if they were made from that person's very own DNA. It's a wonderful

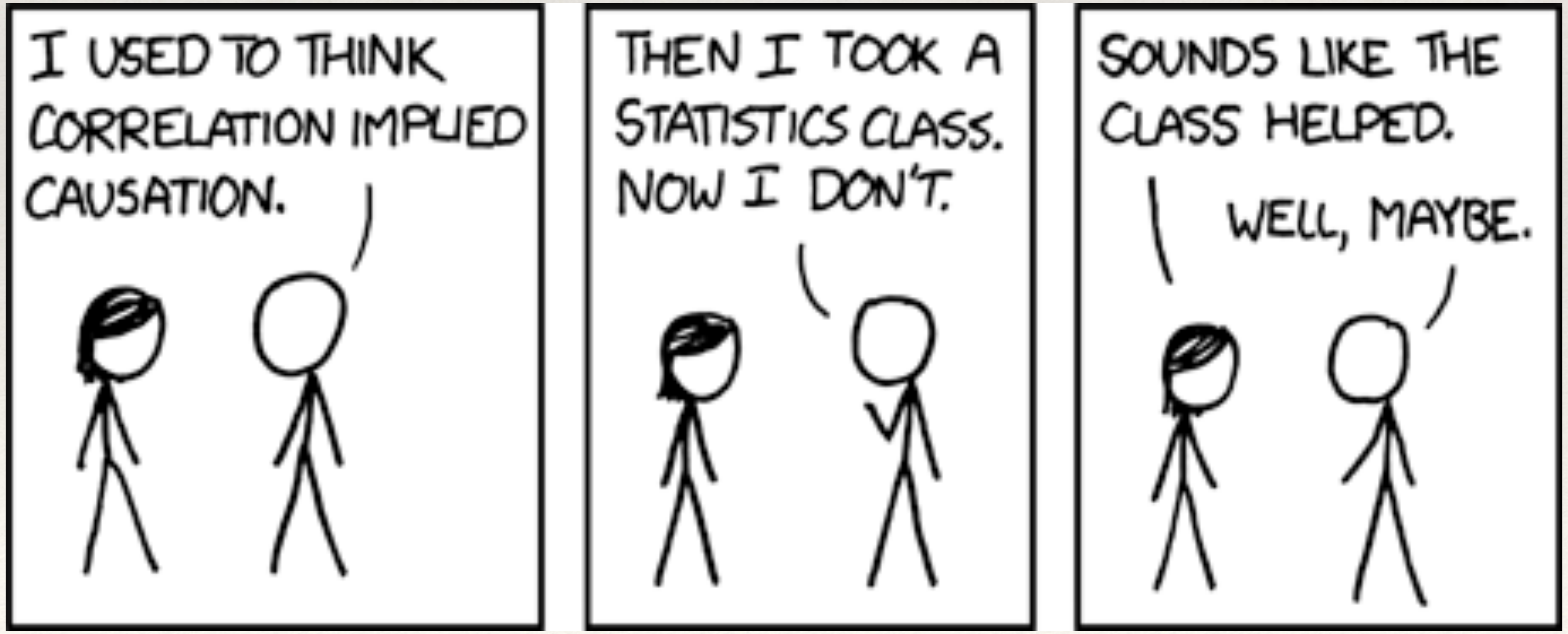
What kinds of things go wrong?

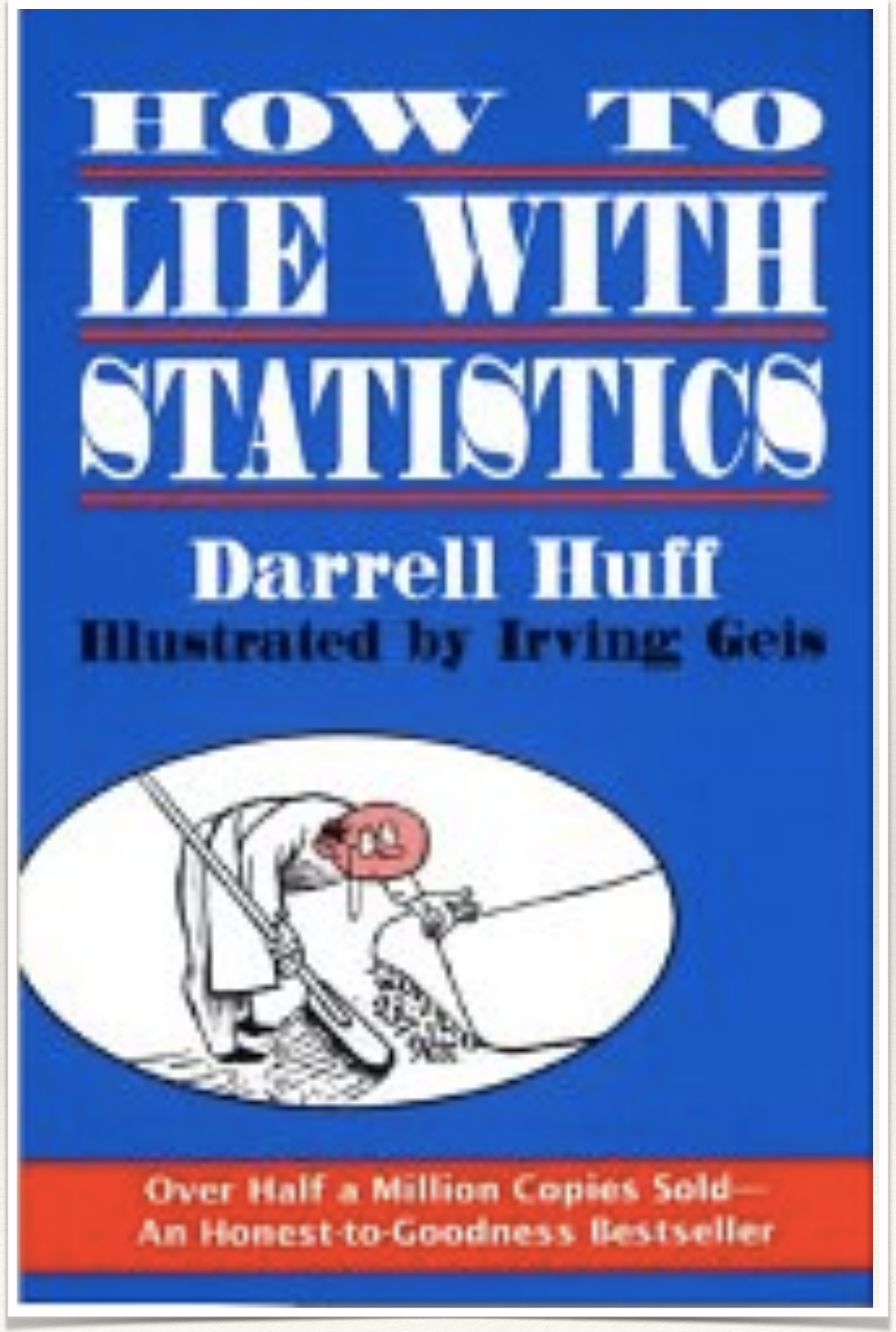
- Training using unrepresentative samples
- Assuming stationary / ergodicity
- Confusing more with better
- Confusing correlation with causation, fit with prediction, & prediction with response schedules
- Confusing models/simulations with reality
- Assuming algorithms are objective and unbiased
- Not considering privacy, bias, social consequences
- Trying to "bolt on" security instead of building it in
- Cargo-cult statistics, quantifauxcation

Ceteris are never paribus.

—Andrea Saltelli

- The sample with built-in bias
- The little figure that isn’t there
- The gee-whiz graph
- Post hoc ergo propter hoc
- …


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.


HOW TO LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

Quantifauxcation

Assign a meaningless number, then conclude it must be meaningful because it is quantitative.

Can be hard to identify when it involves models, complex calculations, etc.

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

—J.W. Tukey

The core of [the scientific method] is remembering your own level of ignorance.

—Jaron Lanier

Role of Models

- All models are wrong, but some are useful. —*George Box*
    - For what?
    - How can you tell?
- It is inappropriate to be concerned about mice when there are tigers abroad. —*George Box*

Cargo-cult Statistics

Go through the motions of computing estimates, uncertainties, etc., without concern for whether the assumptions hold or the models bear any relation to reality.

Use the language of statistics and the calculations of statistics, but not the thinking.

There’s Bayesian cargo-cult statistics, too.

Examples:

- IPCC uncertainties for climate predictions
- Earthquake probabilities
- Political polls

Where does probability come from?

- Phenomenon is random (quantum, thermo)
- Randomness created (random sample, randomized experiment)
- Randomness “invented” in the model itself
- Randomness as metaphor

Altamont Wind Farm

Wind turbines kill raptors.

How many?

What siting and design characteristics matter?

- Data quality: shrinkage, scavenging, background mortality, pieces, attribution
- Models:0-inflated Poisson, etc.

Wind-power company to replace bird-killing Altamont turbines - SFGate

Consultant's solution

- Collisions random, Poisson distributed
- Same process for all birds
- Birds are independent
- Rates follow hierarchical Bayes model with covariates for site, turbine design

Explanation:

- When bird approaches turbine, tosses coin to decide whether to collide
- Chance of heads depends on site & turbine design characteristics
- All birds use the same coin for a given site/design
- Birds toss their coins independently

Is the model reasonable?

Why random?

Why Poisson?

Why particular dependence on site / design?

Why no dependence on size, coloration, ground over, etc?

Why independent across birds, sites, etc.?

What about background mortality?

Complications

Random not same as unpredictable

Do we want to know how many birds are killed? Or the value of some parameter in a model?

Nonstationarity from season, scavenging, time between surveys, etc.

The model changes the subject!

Freedman's Rabbit Theorem

Axioms:

- For the number of rabbits in a closed system to increase, the system must contain at least two rabbits.
- No negative rabbits.

Theorem: *To pull a rabbit from a hat, at least one rabbit must first be placed in the hat.*

Corollary: *You cannot pull a rabbit out of an empty hat, even with a binding promise to return it later.*

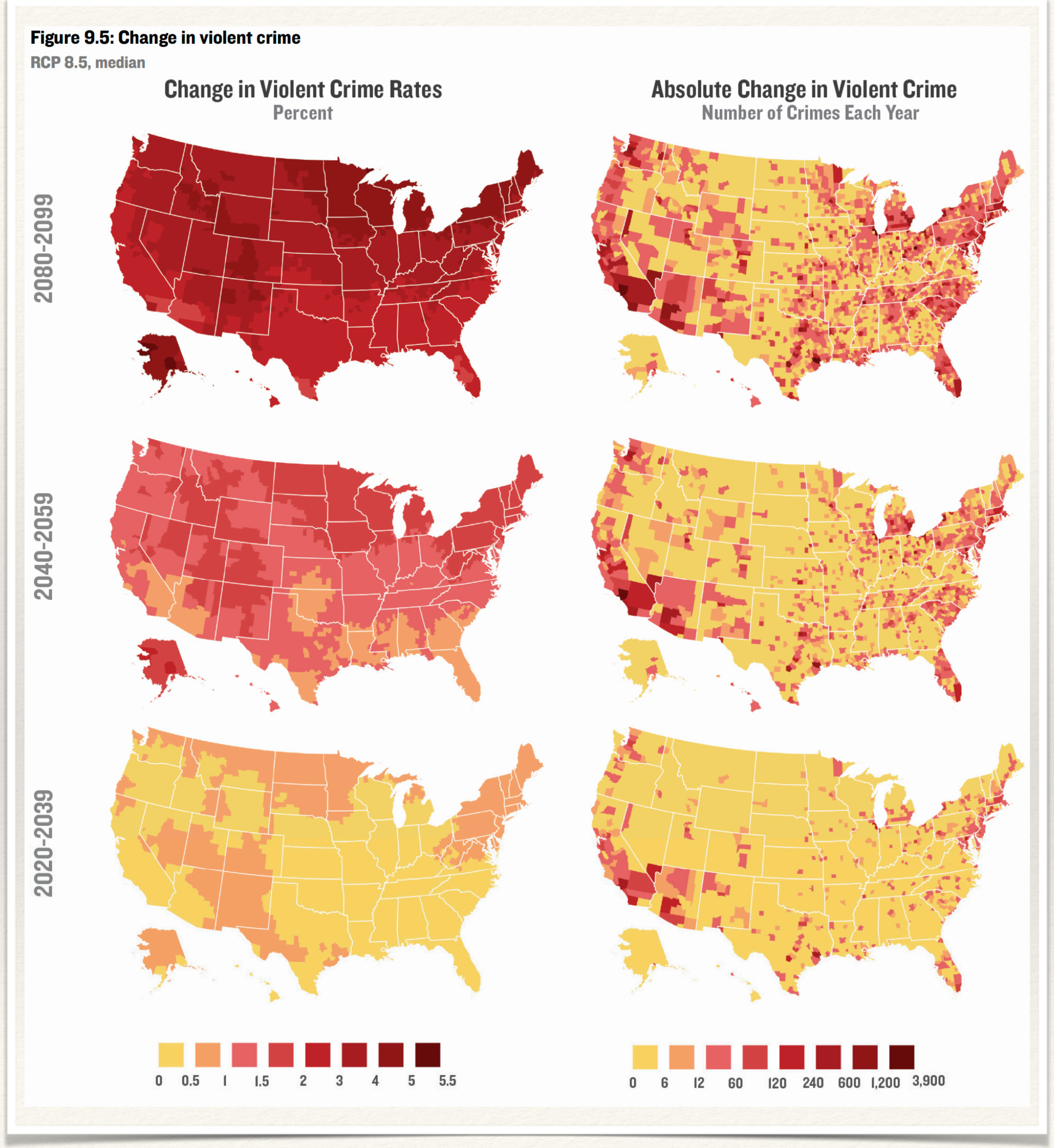


Figure 9.5: Change in violent crime

RCP 8.5, median

Change in Violent Crime Rates
Percent

Absolute Change in Violent Crime
Number of Crimes Each Year

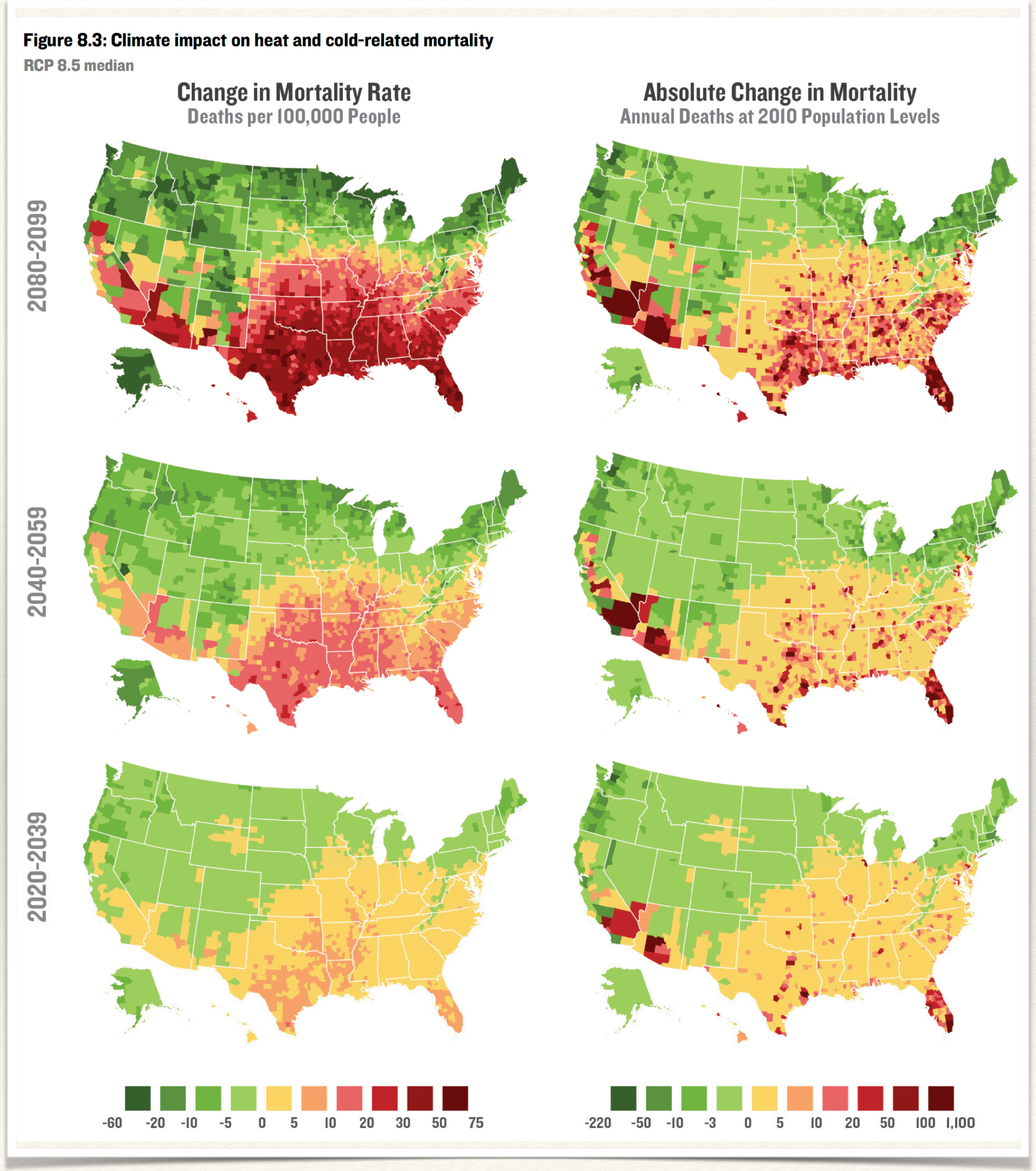


Figure 8.3: Climate impact on heat and cold-related mortality

RCP 8.5 median

Change in Mortality Rate
Deaths per 100,000 People

Absolute Change in Mortality
Annual Deaths at 2010 Population Levels

Data Science Education

- Reproducible research & work practices
- Statistical thinking
- Computational thinking
- Algorithmic thinking
- Evidence
- Ethics, societal impact
- Computational hygiene & software engineering
- MSDSE, Stat 157, Data 8, Division of Data Science

Research

Data science is advanced across Berkeley and integral to a wide range of domains. Faculty and students are pushing forward frontiers from fundamental research to advanced applications. Here are some of the institutes and centers advancing data science across UC Berkeley.

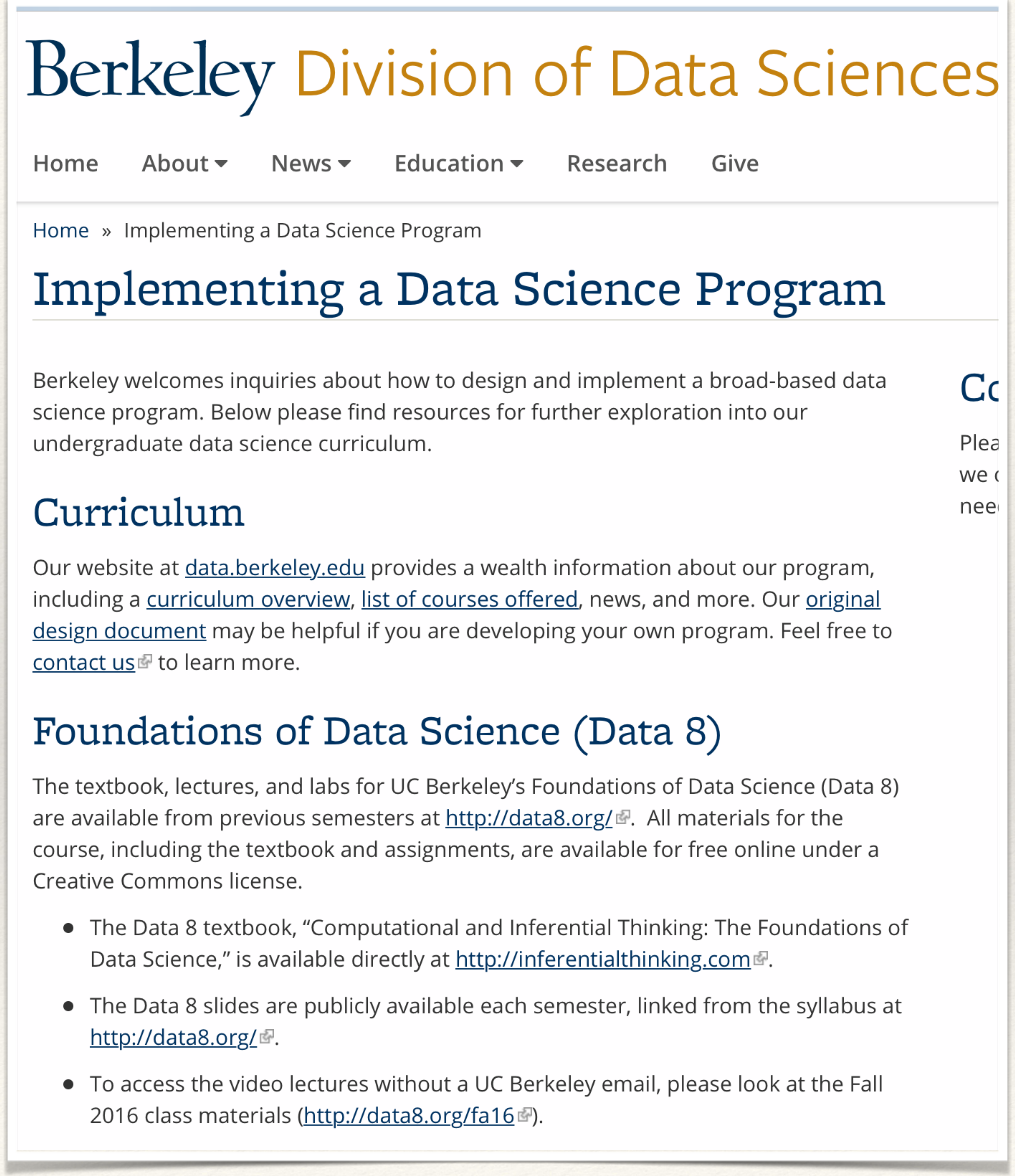
Affiliated Research Centers

BIDS (Berkeley Institute for Data Science)

BIDS is a central hub of research and education at UC Berkeley designed to facilitate and nurture data-intensive science.

D-Lab

D-Lab serves data intensive social science and humanities with in-depth consulting and advising, access to staff support, and training and provisioning for software and other infrastructure needs.



Home » Implementing a Data Science Program

Implementing a Data Science Program

Berkeley welcomes inquiries about how to design and implement a broad-based data science program. Below please find resources for further exploration into our undergraduate data science curriculum.

Curriculum

Our website at data.berkeley.edu provides a wealth information about our program, including a curriculum overview, list of courses offered, news, and more. Our original design document may be helpful if you are developing your own program. Feel free to contact us to learn more.

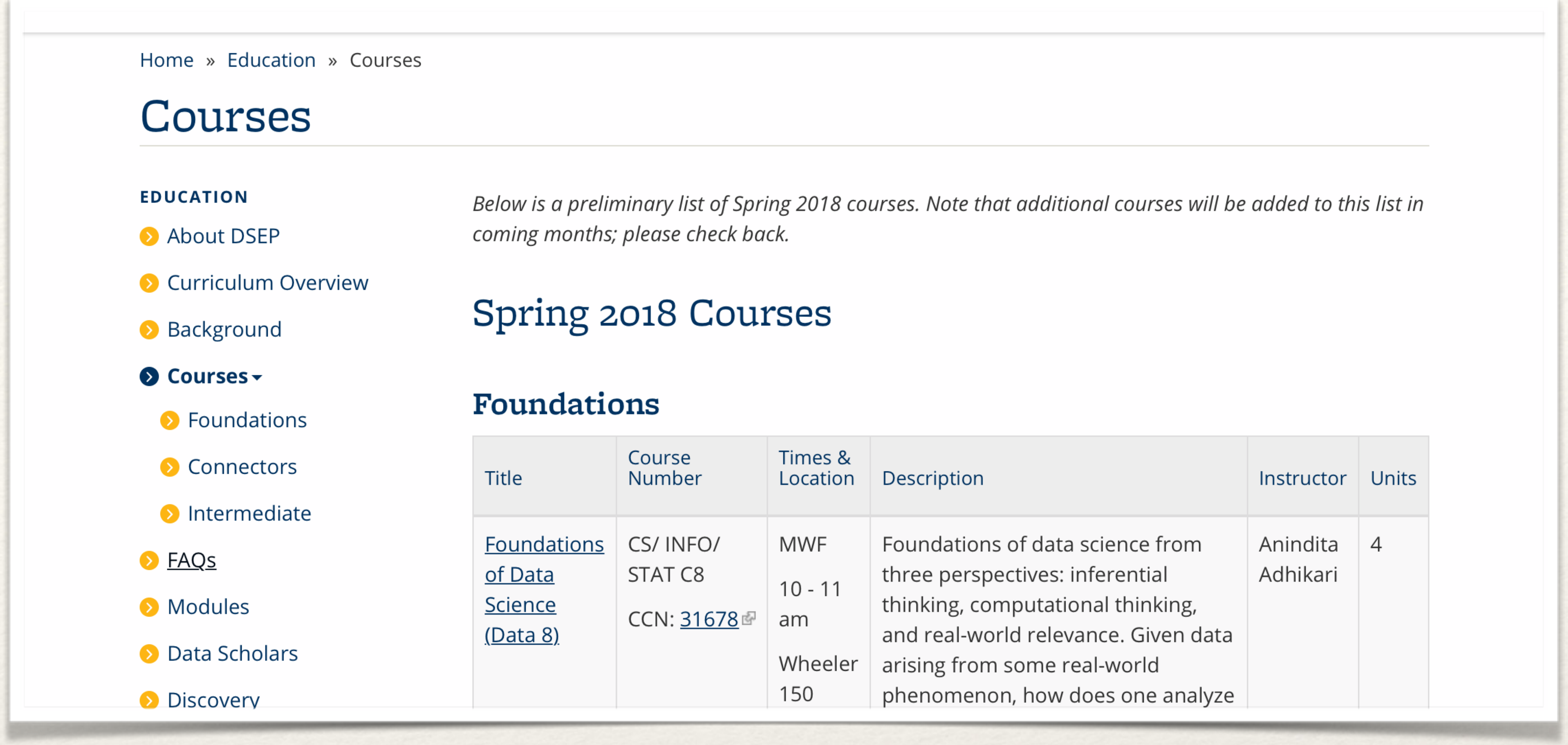
Foundations of Data Science (Data 8)

The textbook, lectures, and labs for UC Berkeley's Foundations of Data Science (Data 8) are available from previous semesters at http://data8.org/. All materials for the course, including the textbook and assignments, are available for free online under a Creative Commons license.

- The Data 8 textbook, "Computational and Inferential Thinking: The Foundations of Data Science," is available directly at http://inferentialthinking.com.
- The Data 8 slides are publicly available each semester, linked from the syllabus at http://data8.org/.
- To access the video lectures without a UC Berkeley email, please look at the Fall 2016 class materials (http://data8.org/fa16).

Cc

Plea
we c
nee


Home » Education » Courses

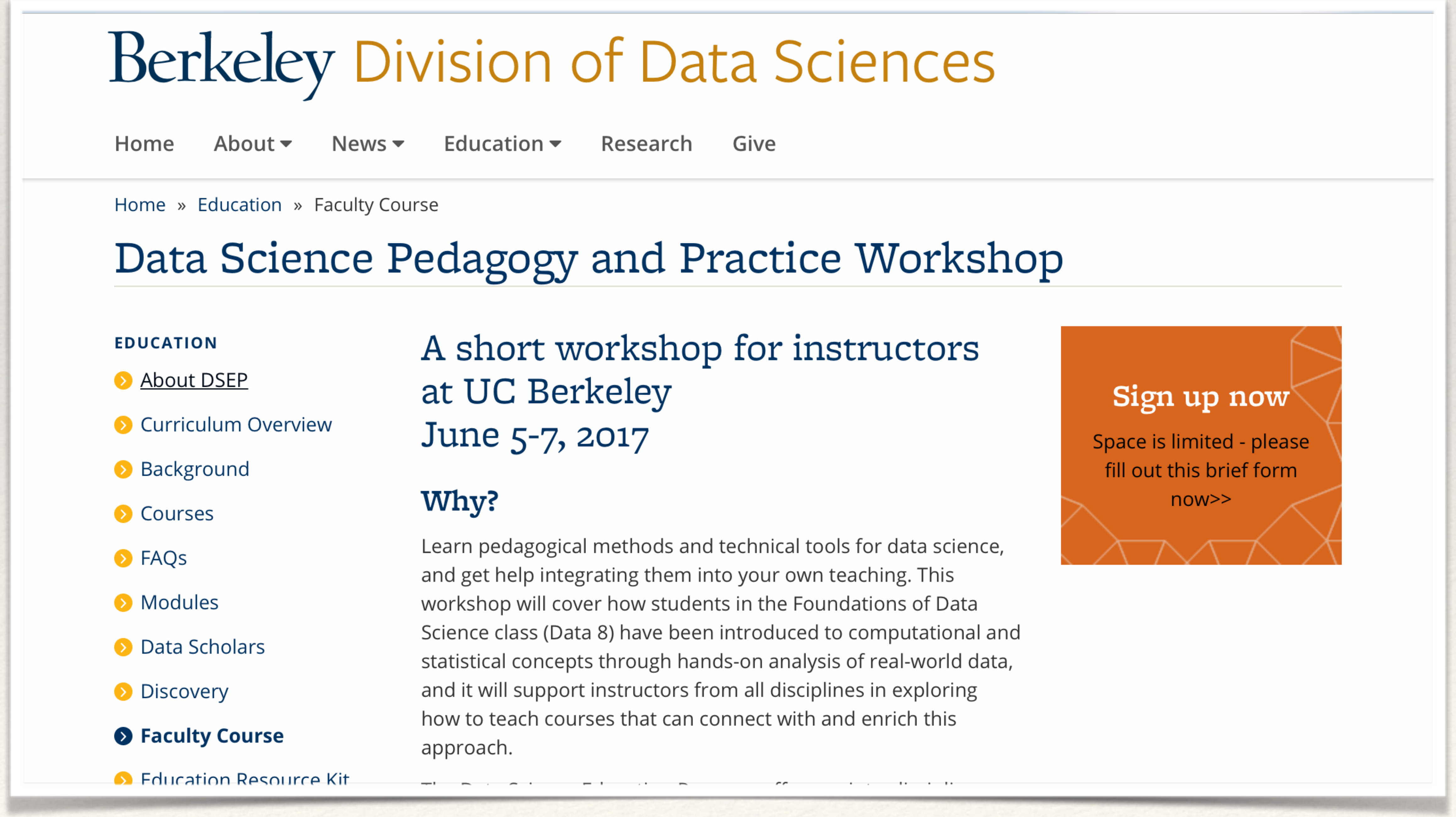
Courses

Below is a preliminary list of Spring 2018 courses. Note that additional courses will be added to this list in coming months; please check back.

Spring 2018 Courses

Foundations

Title	Course Number	Times & Location	Description	Instructor	Units
Foundations of Data Science (Data 8)	CS/ INFO/ STAT C8 CCN: 31678	MWF 10 - 11 am Wheeler 150	Foundations of data science from three perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze	Anindita Adhikari	4



Home » Education » Faculty Course

Data Science Pedagogy and Practice Workshop

A short workshop for instructors at UC Berkeley June 5-7, 2017

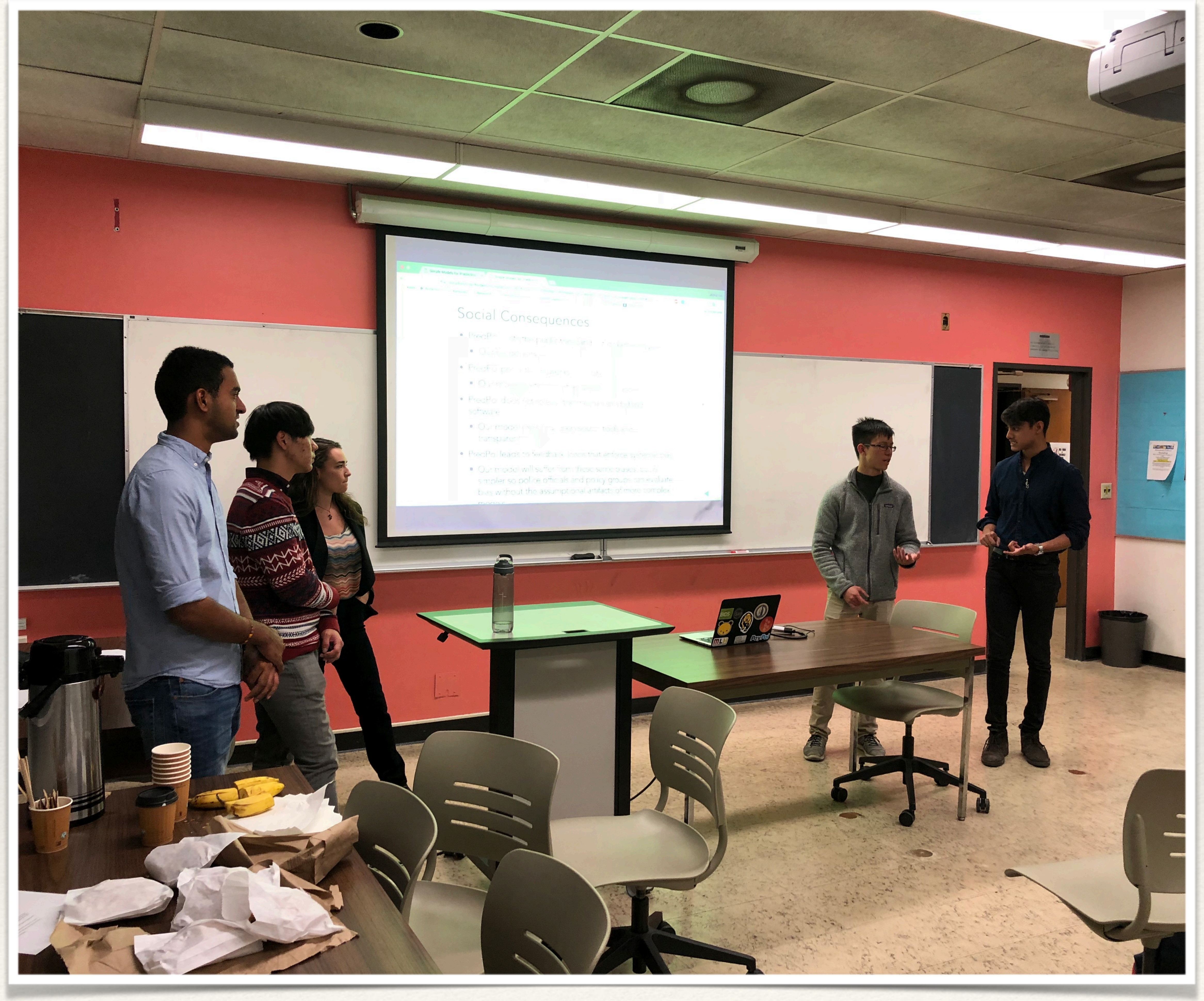
Why?

Learn pedagogical methods and technical tools for data science, and get help integrating them into your own teaching. This workshop will cover how students in the Foundations of Data Science class (Data 8) have been introduced to computational and statistical concepts through hands-on analysis of real-world data, and it will support instructors from all disciplines in exploring how to teach courses that can connect with and enrich this approach.

Data Science with Applications to Social Good

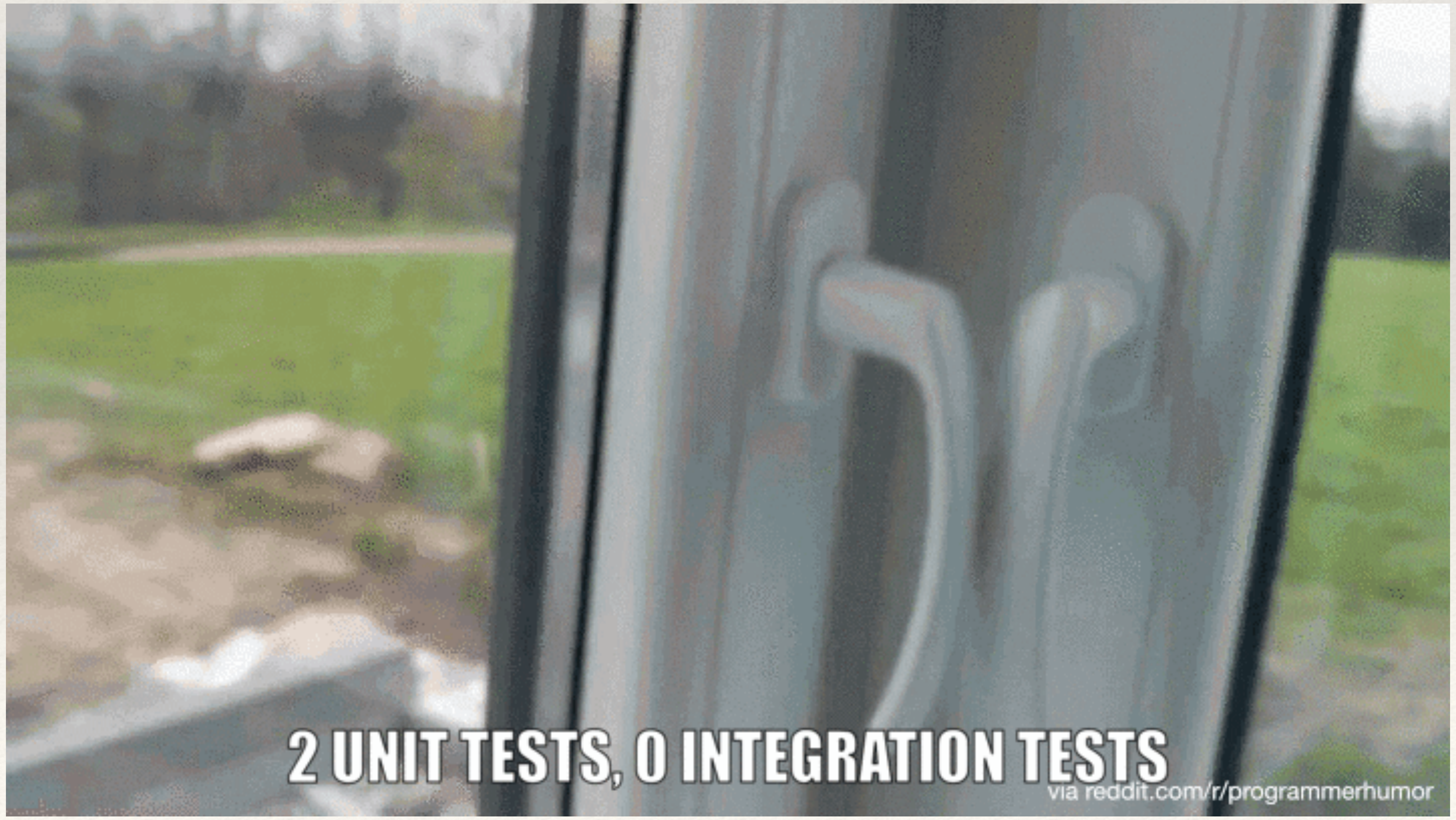
Term projects:

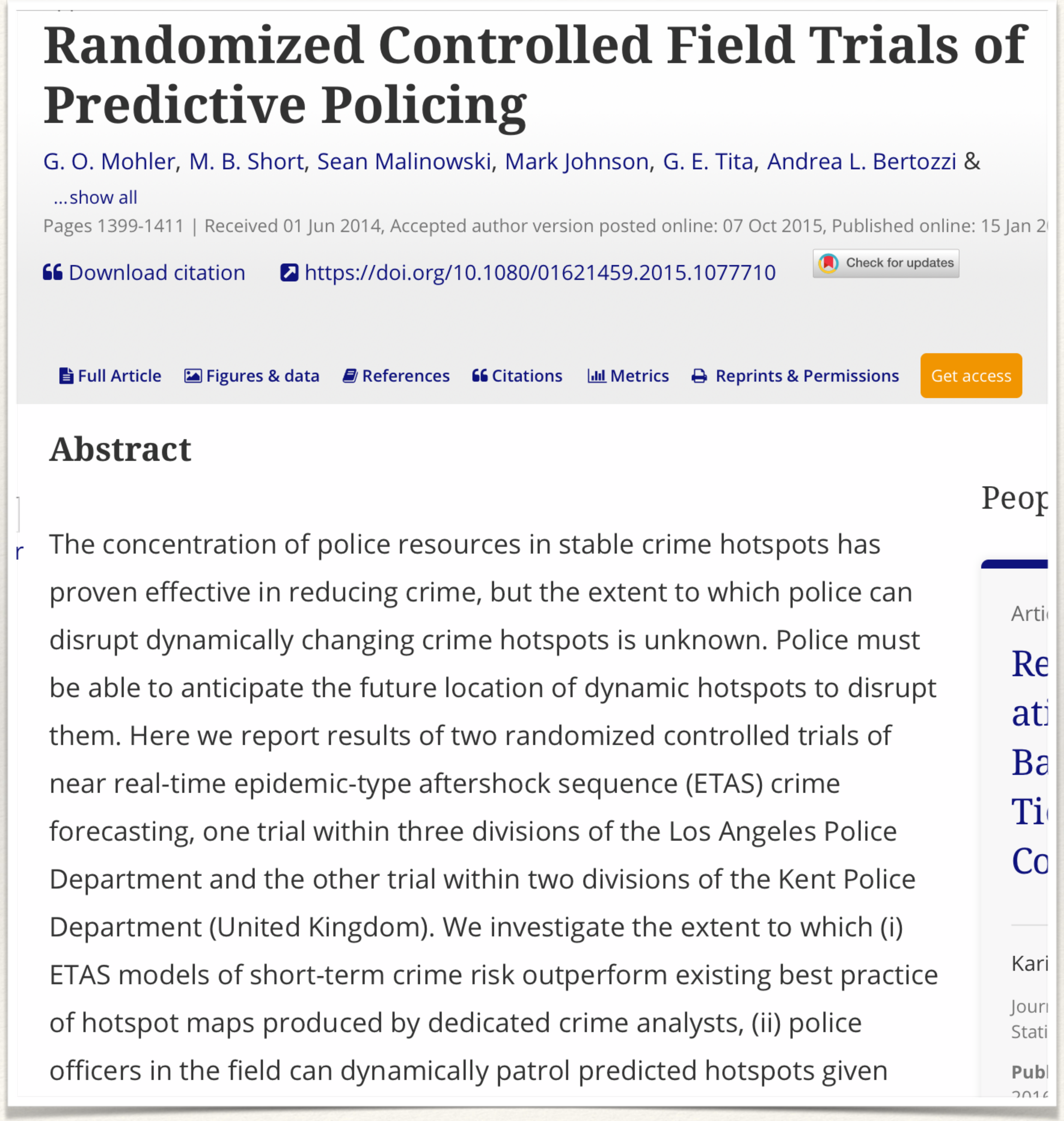
- Predictive policing
- Water injection and induced seismicity
- Open-source Python library of nonparametric permutation methods



Work Habits

- Revision control systems (e.g., git)
- Unit tests, integration tests, regression tests, coverage tests, automated
- Scripted analyses
- Pair programming
- Code review
- Documentation


2 UNIT TESTS, 0 INTEGRATION TESTS
via reddit.com/r/programmerhumor



Randomized Controlled Field Trials of Predictive Policing

G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi &

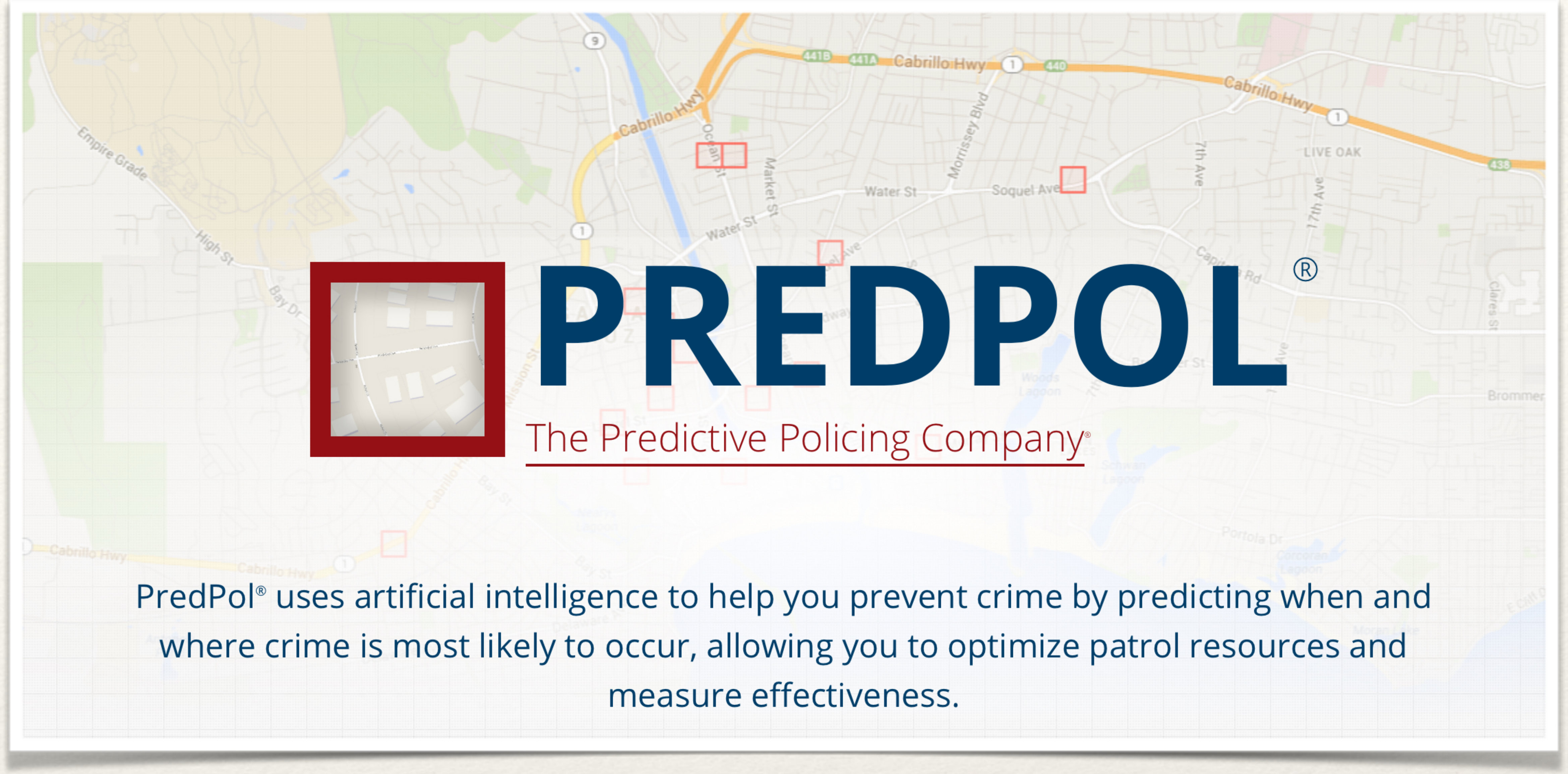
...show all

Pages 1399-1411 | Received 01 Jun 2014, Accepted author version posted online: 07 Oct 2015, Published online: 15 Jan 2

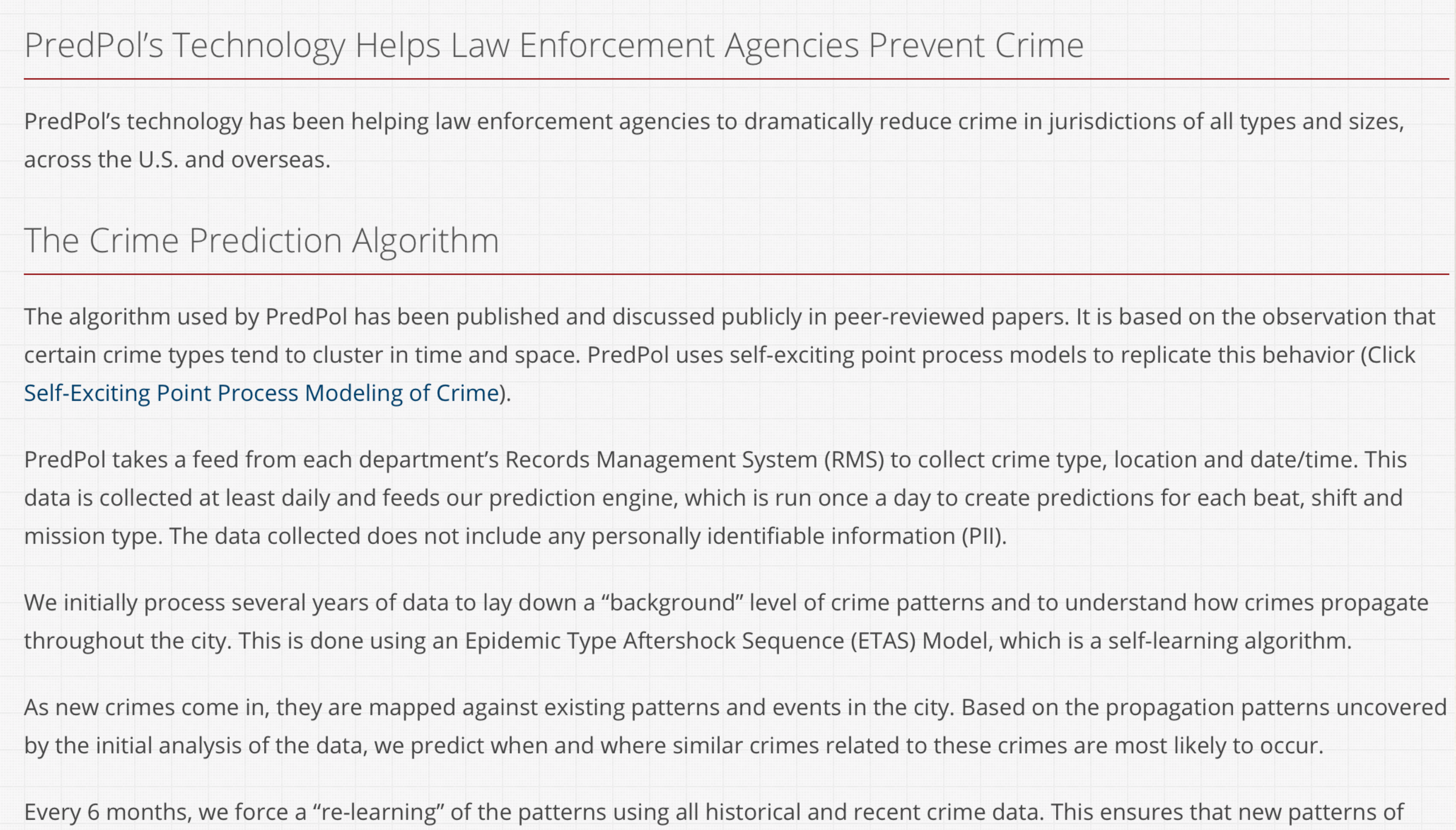
Download citation https://doi.org/10.1080/01621459.2015.1077710 Check for updates

Abstract

The concentration of police resources in stable crime hotspots has proven effective in reducing crime, but the extent to which police can disrupt dynamically changing crime hotspots is unknown. Police must be able to anticipate the future location of dynamic hotspots to disrupt them. Here we report results of two randomized controlled trials of near real-time epidemic-type aftershock sequence (ETAS) crime forecasting, one trial within three divisions of the Los Angeles Police Department and the other trial within two divisions of the Kent Police Department (United Kingdom). We investigate the extent to which (i) ETAS models of short-term crime risk outperform existing best practice of hotspot maps produced by dedicated crime analysts, (ii) police officers in the field can dynamically patrol predicted hotspots given



PredPol® uses artificial intelligence to help you prevent crime by predicting when and where crime is most likely to occur, allowing you to optimize patrol resources and measure effectiveness.



PredPol's Technology Helps Law Enforcement Agencies Prevent Crime

PredPol's technology has been helping law enforcement agencies to dramatically reduce crime in jurisdictions of all types and sizes, across the U.S. and overseas.

The Crime Prediction Algorithm

The algorithm used by PredPol has been published and discussed publicly in peer-reviewed papers. It is based on the observation that certain crime types tend to cluster in time and space. PredPol uses self-exciting point process models to replicate this behavior (Click Self-Exciting Point Process Modeling of Crime).

PredPol takes a feed from each department's Records Management System (RMS) to collect crime type, location and date/time. This data is collected at least daily and feeds our prediction engine, which is run once a day to create predictions for each beat, shift and mission type. The data collected does not include any personally identifiable information (PII).

We initially process several years of data to lay down a "background" level of crime patterns and to understand how crimes propagate throughout the city. This is done using an Epidemic Type Aftershock Sequence (ETAS) Model, which is a self-learning algorithm.

As new crimes come in, they are mapped against existing patterns and events in the city. Based on the propagation patterns uncovered by the initial analysis of the data, we predict when and where similar crimes related to these crimes are most likely to occur.

Every 6 months, we force a "re-learning" of the patterns using all historical and recent crime data. This ensures that new patterns of

IN DETAIL

To predict and serve?

Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data? **Kristian Lum** and **William Isaac** consider the evidence – and the social consequences

population?") with survey data from the 2011 National Survey on Drug Use and Health (NSDUH). This approach allowed us to obtain high-resolution *estimates* of illicit drug use from a non-criminal justice, population-based data source (see "How do we estimate the number of drug users?") which we could then compare with police records. In doing so, we find that drug crimes known to police are not a representative sample of all drug crimes.

While it is likely that estimates derived from national-level data do not *perfectly* represent drug use at the local level, we still believe these estimates paint a more accurate picture of drug use in Oakland than the arrest data for several reasons. First, the US Bureau of Justice Statistics – the government body responsible for compiling and analysing criminal justice data – has used data from the NSDUH as a more representative measure of drug use than police reports.[2] Second, while arrest data is collected as a by-product of police activity, the NSDUH is a well-funded survey designed using best practices for obtaining a statistically representative sample. And finally, although there is evidence that some drug users do conceal illegal drug use from public health surveys, we believe that any incentives for such concealment apply much more strongly to police records of drug use than to public health surveys, as public health officials are not empowered (nor inclined) to arrest those who admit to illicit drug use. For these reasons, our analysis continues under the assumption that our public health-derived estimates of drug crimes represent a ground truth for the purpose of comparison.

Figure 1(a) shows the number of drug arrests in 2010 based on data obtained from the Oakland Police Department; Figure 1(b) shows the estimated number of drug users by grid square. From comparing these figures, it is clear that police databases and public health-derived estimates tell dramatically different stories about the pattern of drug use in Oakland. In Figure 1(a), we see that drug arrests in the police database appear concentrated in neighbourhoods around West Oakland (1) and International Boulevard (2), two areas with largely non-white and low-income populations. These neighbourhoods experience about 200 times more drug-related arrests than areas outside of these clusters. In contrast, our estimates (in Figure 1(b)) suggest that drug crimes are much more evenly distributed across the city. Variations in our estimated number of drug users are driven primarily by differences in population density, as the estimated rate of drug use is relatively uniform across the city. This suggests that while drug crimes exist everywhere, drug arrests tend to only occur in very specific locations – the police data appear to disproportionately represent crimes committed in areas with higher populations of non-white and low-income residents.

To investigate the effect of police-recorded data on predictive policing models, we apply a recently published predictive policing algorithm to the drug crime records in Oakland.[9] This algorithm was developed by PredPol, one of the largest vendors of predictive policing systems in the USA and one of the few companies to publicly release its algorithm in a peer-reviewed journal. It has been described by its founders ▶

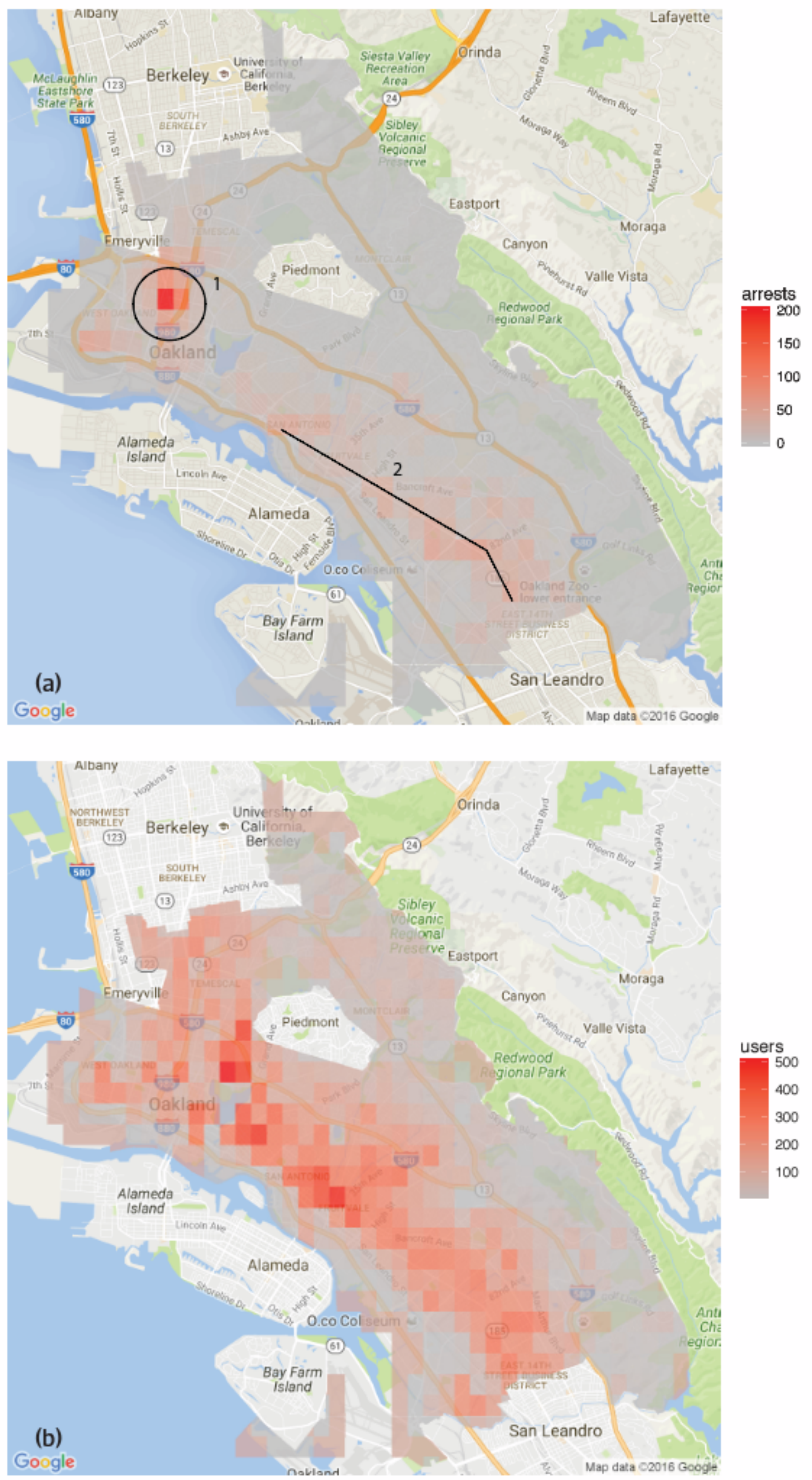
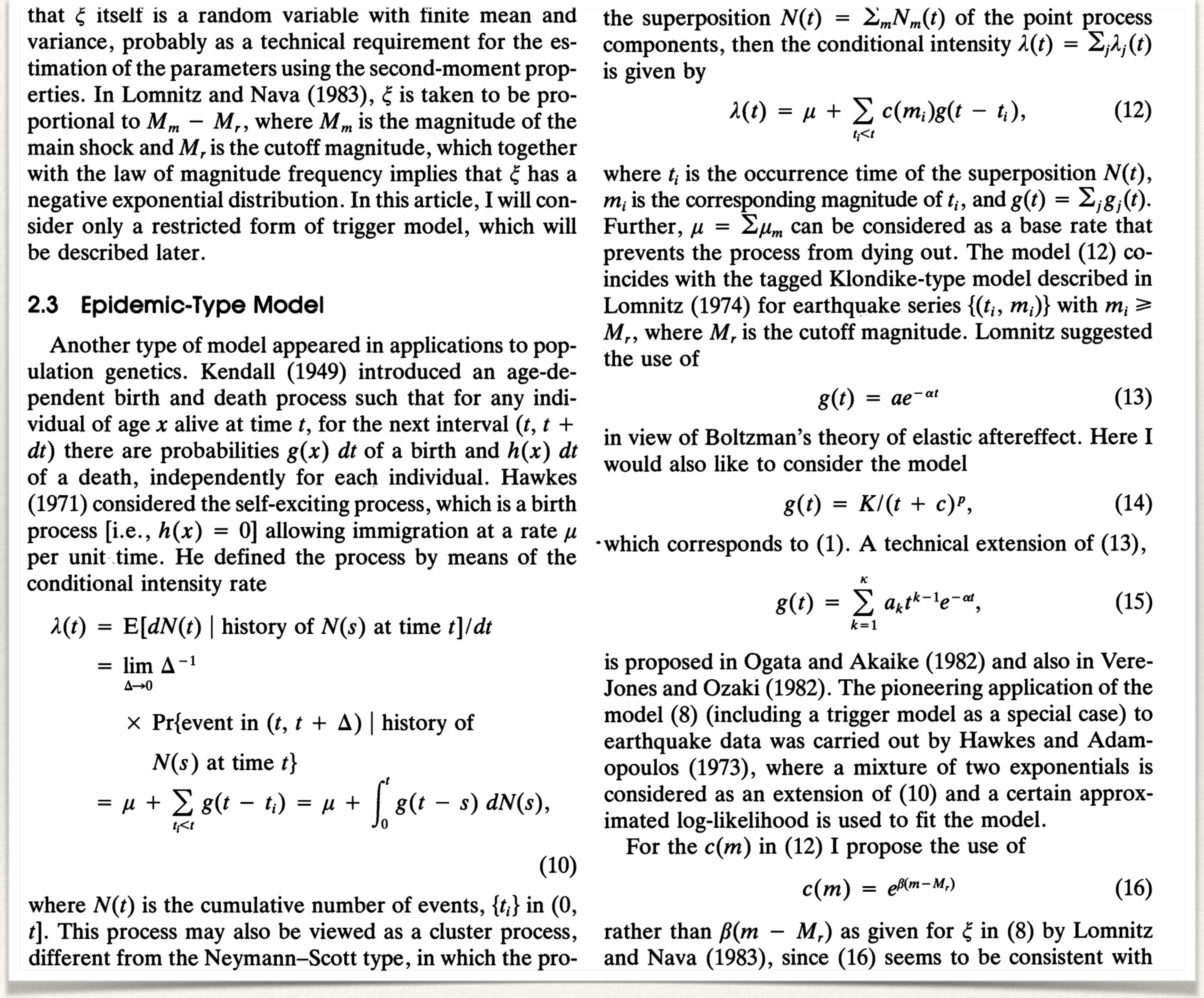


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

ETAS: self-exciting linear Hawkes process



that ξ itself is a random variable with finite mean and variance, probably as a technical requirement for the estimation of the parameters using the second-moment properties. In Lomnitz and Nava (1983), ξ is taken to be proportional to $M_m - M_r$, where M_m is the magnitude of the main shock and M_r is the cutoff magnitude, which together with the law of magnitude frequency implies that ξ has a negative exponential distribution. In this article, I will consider only a restricted form of trigger model, which will be described later.

2.3 Epidemic-Type Model

Another type of model appeared in applications to population genetics. Kendall (1949) introduced an age-dependent birth and death process such that for any individual of age x alive at time t, for the next interval $(t, t + dt)$ there are probabilities $g(x)\,dt$ of a birth and $h(x)\,dt$ of a death, independently for each individual. Hawkes (1971) considered the self-exciting process, which is a birth process [i.e., $h(x) = 0$] allowing immigration at a rate μ per unit time. He defined the process by means of the conditional intensity rate

$$\begin{aligned}\lambda(t) &= \mathrm{E}[dN(t) \mid \text{history of } N(s) \text{ at time } t]/dt \\ &= \lim_{\Delta\to0} \Delta^{-1} \\ &\quad \times \Pr\{\text{event in } (t, t+\Delta) \mid \text{history of} \\ &\quad N(s) \text{ at time } t\} \\ &= \mu + \sum_{t_i<t} g(t - t_i) = \mu + \int_0^t g(t-s)\,dN(s), \qquad (10)\end{aligned}$$

where $N(t)$ is the cumulative number of events, $\{t_i\}$ in $(0, t]$. This process may also be viewed as a cluster process, different from the Neymann–Scott type, in which the pro- the superposition $N(t) = \Sigma_m N_m(t)$ of the point process components, then the conditional intensity $\lambda(t) = \Sigma_j \lambda_j(t)$ is given by

$$\lambda(t) = \mu + \sum_{t_i<t} c(m_i) g(t - t_i), \qquad (12)$$

where t_i is the occurrence time of the superposition $N(t)$, m_i is the corresponding magnitude of t_i, and $g(t) = \Sigma_j g_j(t)$. Further, $\mu = \Sigma \mu_m$ can be considered as a base rate that prevents the process from dying out. The model (12) coincides with the tagged Klondike-type model described in Lomnitz (1974) for earthquake series $\{(t_i, m_i)\}$ with $m_i \geqslant M_r$, where M_r is the cutoff magnitude. Lomnitz suggested the use of

$$g(t) = ae^{-\alpha t} \qquad (13)$$

in view of Boltzman's theory of elastic aftereffect. Here I would also like to consider the model

$$g(t) = K/(t + c)^p, \qquad (14)$$

which corresponds to (1). A technical extension of (13),

$$g(t) = \sum_{k=1}^{K} a_k t^{k-1} e^{-\alpha t}, \qquad (15)$$

is proposed in Ogata and Akaike (1982) and also in Vere-Jones and Ozaki (1982). The pioneering application of the model (8) (including a trigger model as a special case) to earthquake data was carried out by Hawkes and Adamopoulos (1973), where a mixture of two exponentials is considered as an extension of (10) and a certain approximated log-likelihood is used to fit the model.

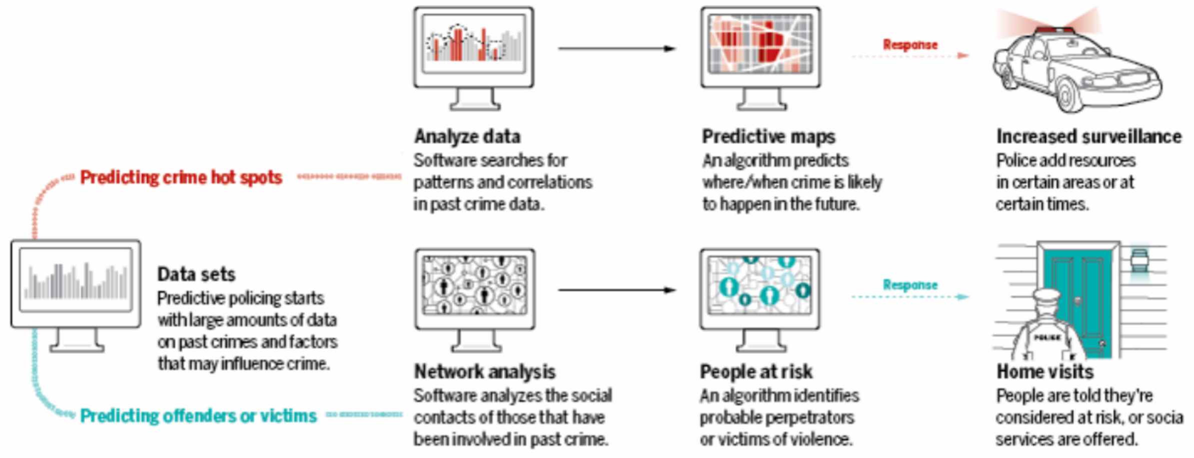
For the $c(m)$ in (12) I propose the use of

$$c(m) = e^{\beta(m - M_r)} \qquad (16)$$

rather than $\beta(m - M_r)$ as given for ξ in (8) by Lomnitz and Nava (1983), since (16) seems to be consistent with

Crimes of the future

One commonly used approach in predictive policing seeks to forecast where and when crime will happen; another focuses on who will commit crime or become a victim.



Predicting crime hot spots

Predicting offenders or victims

Data sets
Predictive policing starts with large amounts of data on past crimes and factors that may influence crime.

Analyze data
Software searches for patterns and correlations in past crime data.

Predictive maps
An algorithm predicts where/when crime is likely to happen in the future.

Response

Increased surveillance
Police add resources in certain areas or at certain times.

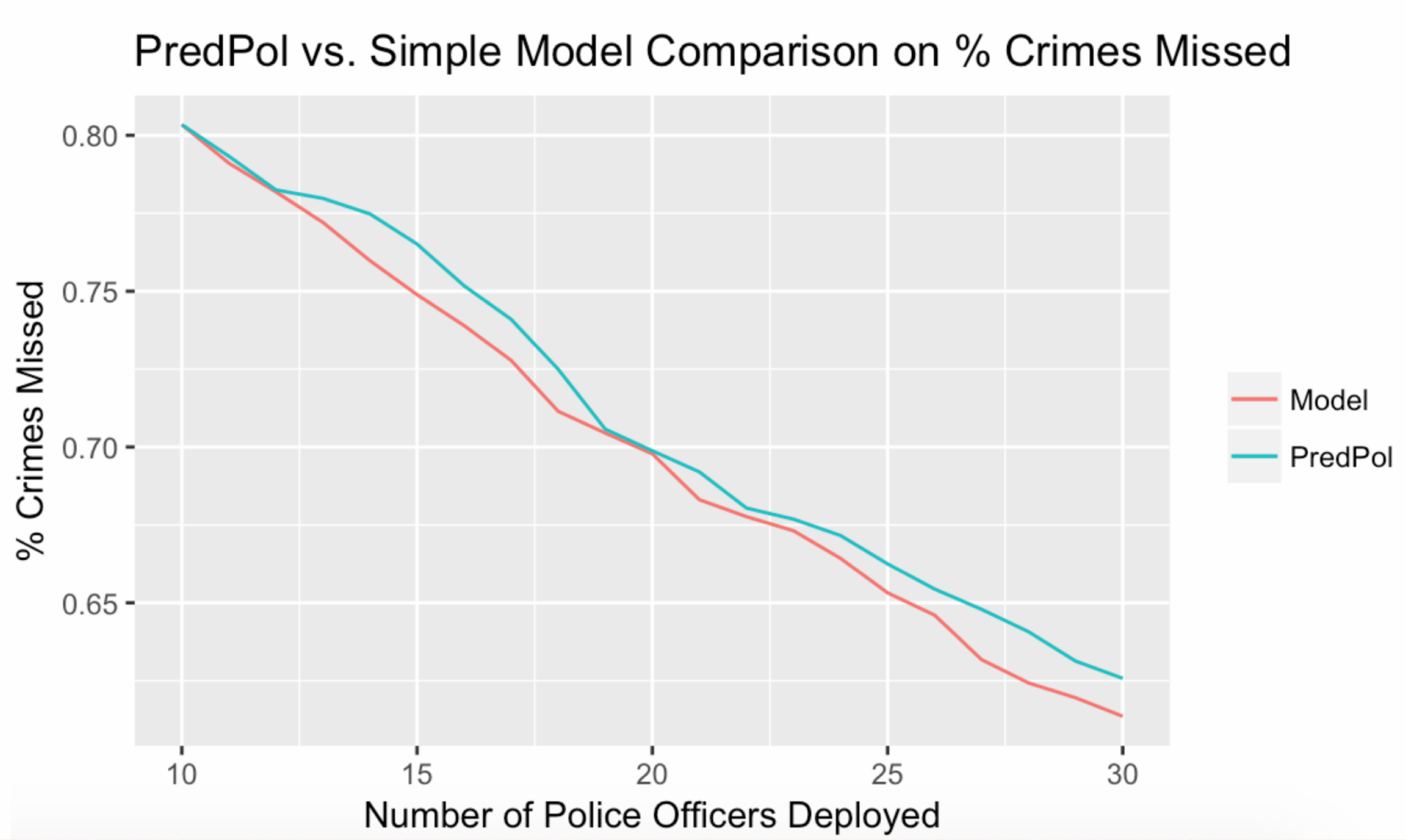
Network analysis
Software analyzes the social contacts of those that have been involved in past crime.

People at risk
An algorithm identifies probable perpetrators or victims of violence.

Response

Home visits
People are told they're considered at risk, or socia services are offered.

Diagram: G. Grullón/*Science*



Data Science Questions

- What's the underlying experiment?
- How were data collected/selected/processed to get the “data”?
- What analysis was reported to have been done on the “data”?
- Was it the right analysis to do? Was it done correctly? Was the implementation stable/sound?
- If the results involve probability, where did the probability come from?
- If there’s a model, where did it come from? Is it based on the “physics” of the situation?
- Were the results reported correctly?
- How many analyses were tried? What were they? What were the results? How was multiplicity treated?
- Were there ad hoc aspects to the analysis? What if different choices were made?
- Can someone else re-use/re-purpose the tools?