



Data Science as a Superpower

Amelia McNamara [@AmeliaMN](https://twitter.com/AmeliaMN)

University of St Thomas St Paul, MN

Department of Computer & Information Sciences

#privilegealert

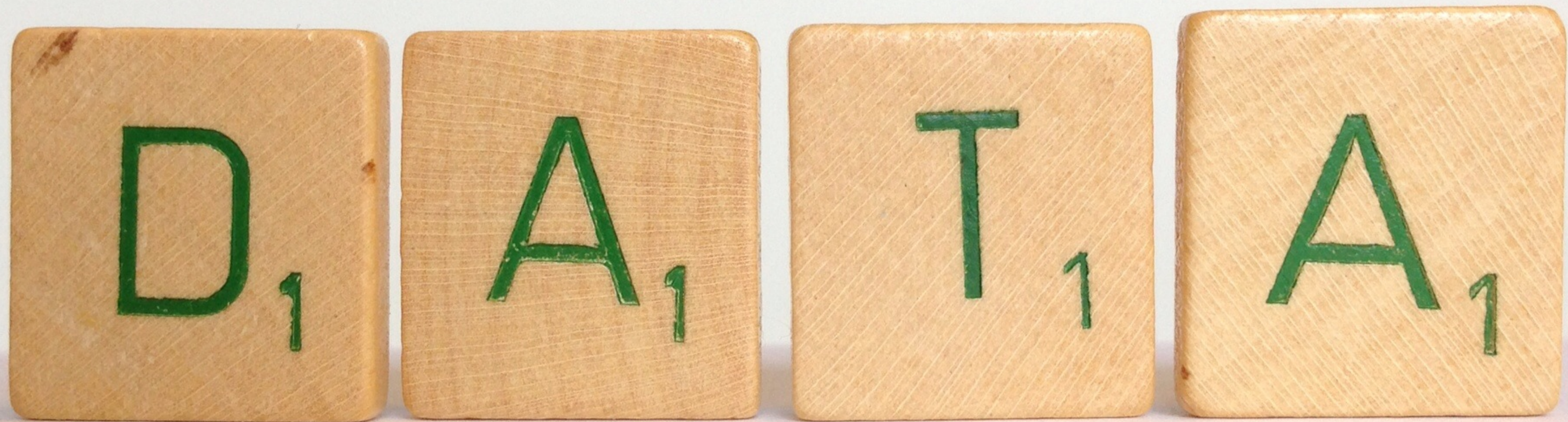
I am a

- white
- straight
- cisgender
- middle class
- highly educated
- American

lady

I'm doing the best I can when I talk about issues of race, class, gender, and other sensitive topics. But you should *always feel free* to call me out (publicly or privately).





I think about data as any information that we can collect (write down or record on the computer) about the world.

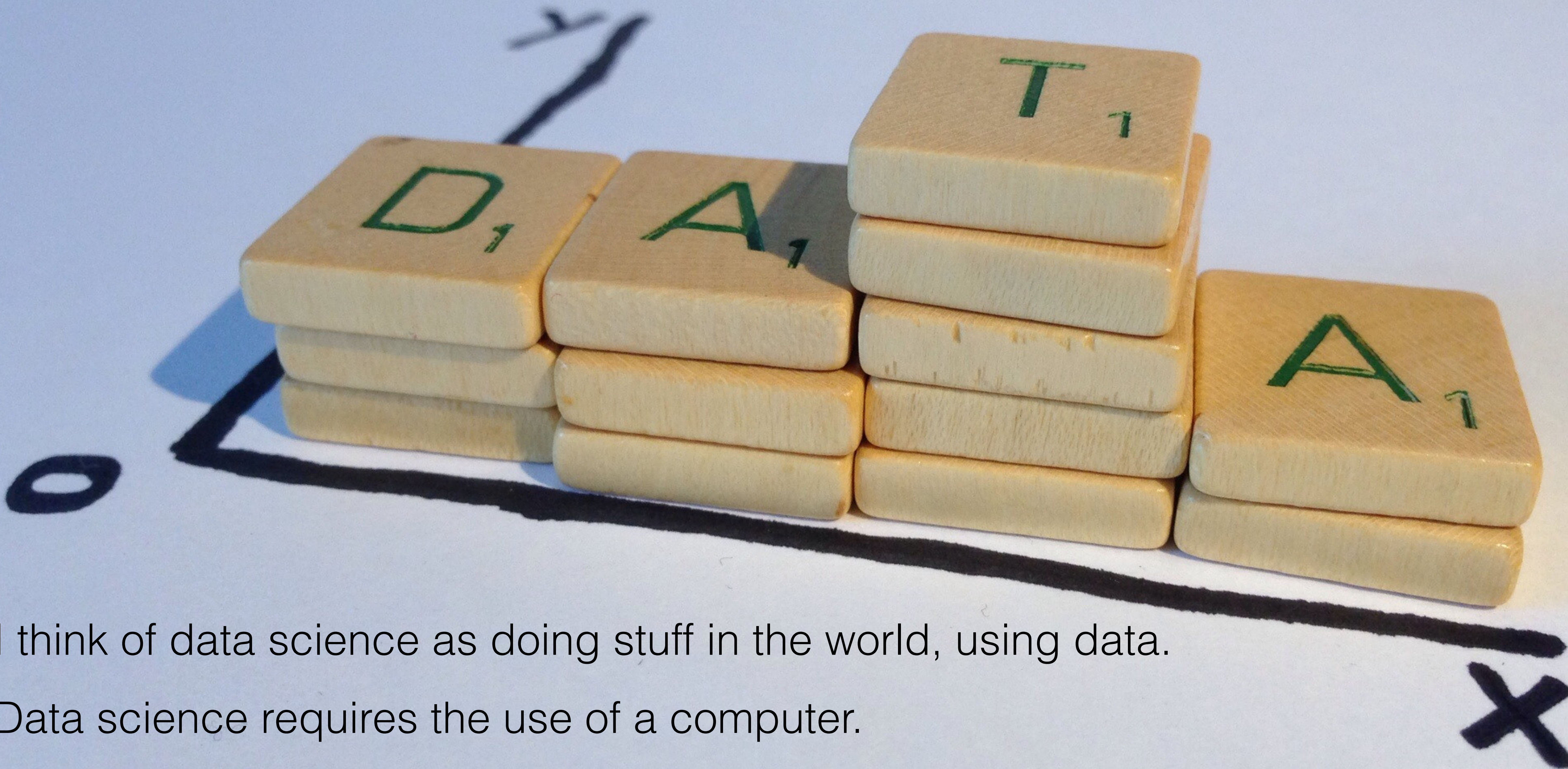
Numbers are data, but text can be, too!

Brainstorm: data exhaust

We generate data every day, whether we know it or not.

For example, I have a Withings watch, so I generate data every time I take a step. I consciously chose to wear this, but there are other times I am unconsciously generating data. It is incidental to what I'm doing, and streams off me as "data exhaust."

Take a few minutes and make a list of all the places you generate data on a normal day.



I think of data science as doing stuff in the world, using data.

Data science requires the use of a computer.

nest

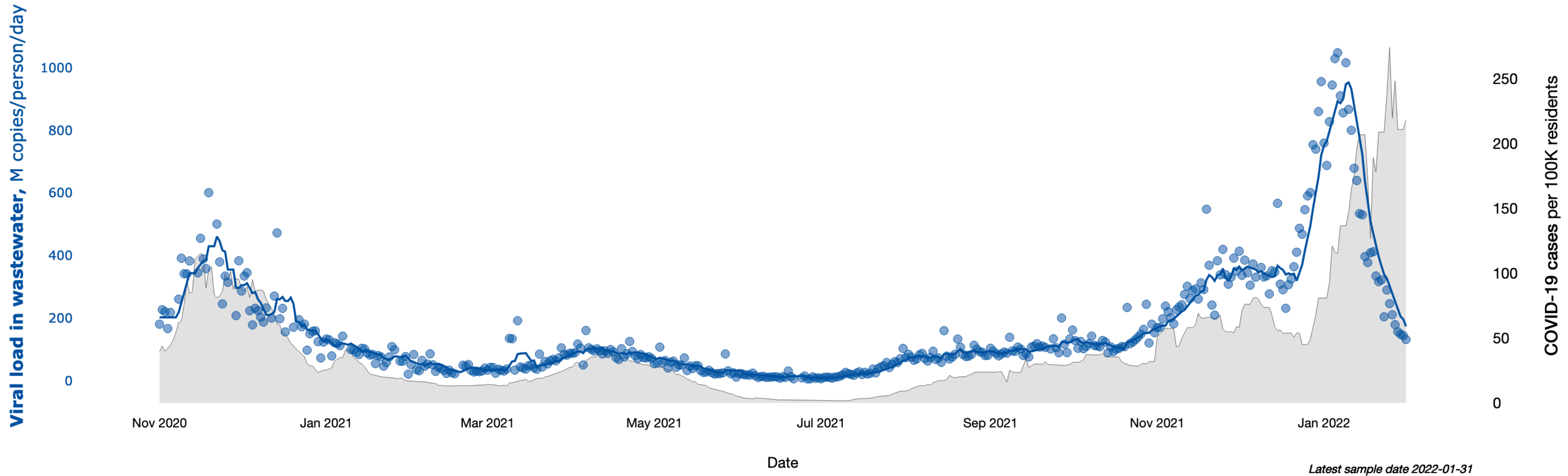
COOLING
73





Tracking COVID-19 Prevalence with Metro Plant Wastewater

The number of reported cases of COVID-19 infections in the seven-county metro area corresponds to the prevalence of the virus in wastewater samples at the Metro treatment plant in Saint Paul. The plant serves a large portion of the seven-county metro area.



The blue line and points show the total amount of SARS-CoV-2 viral RNA in wastewater flowing into the Metro Plant, in millions copies of the SARS-CoV-2 genome per person served by the wastewater area, per day. Blue points are daily values; the blue line is a running average of the previous 7 days. The gray line shows the average of the previous 7 days of new reported COVID-19 infections in the seven-county Metro area per 100,000 residents. Case data are provided by the Minnesota Department of Health and downloaded from USA Facts (<https://usafacts.org>). New cases tend to lag wastewater detection trends by about 6-8 days.

This project is open-source. See our GitHub repository here [🔗](#)

App last updated 2022-01-31



Data received so far:

- **93** Excel files containing
- **146** individual sheets with
- **66,294** rows and
- **345** unique columns from
- **3** agencies within
- **2** federal departments

Regarding

- **2672** children and
- **2876** adults from
- **16** unique countries speaking
- **19** languages

ACLU



RSTUDIO::CONF 2019

INDUSTRY

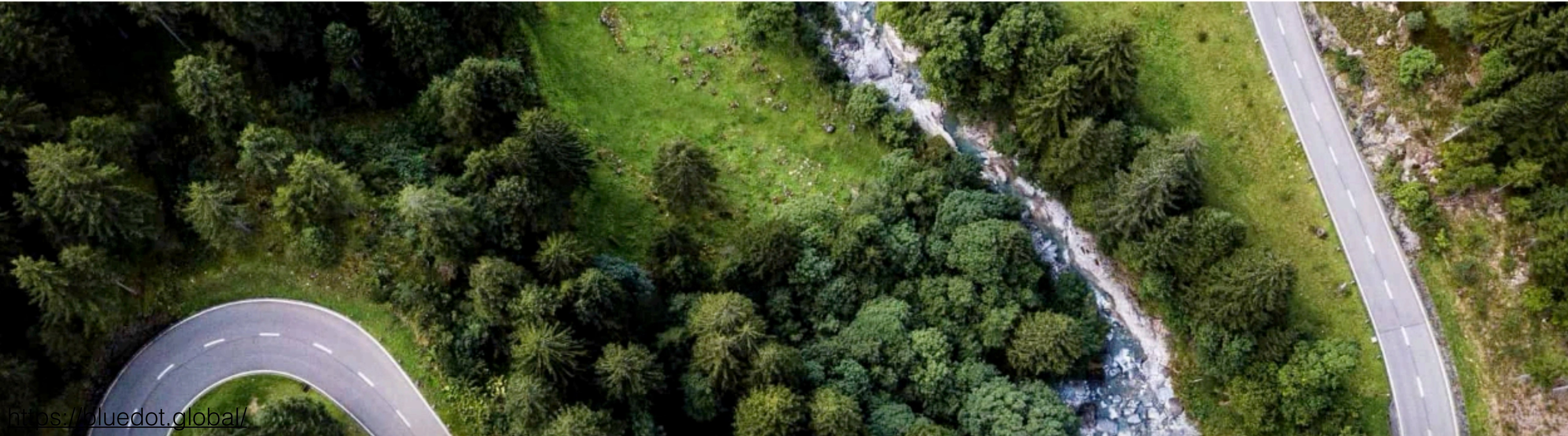
R at the ACLU: Joining tables to to reunite families

Brooke Watson | January 25, 2019

<https://www.rstudio.com/resources/rstudioconf-2019/r-at-the-aclu-joining-tables-to-to-reunite-families/>

About Us

BlueDot protects people around the world from infectious diseases with human and artificial intelligence.



OCEAN HEALTH INDEX

A healthy ocean sustainably delivers a range of benefits to people now and in the future. The Ocean Health Index is the comprehensive framework used to measure ocean health from global to local scales.

GLOBAL
ASSESSMENT

INDEPENDENT
ASSESSMENTS

<http://www.oceanhealthindex.org/>



R for better science in less time

Julia Stewart Lowndes, PhD
Marine Data Scientist & Mozilla Fellow
National Center for Ecological Analysis & Synthesis
University of California at Santa Barbara, USA

 @juliesquid

 lowndes@nceas.ucsb.edu

 jules32.github.io/useR-2019-keynote

How Netflix Reverse Engineered Hollywood

To understand how people look for movies, the video service created 76,897 micro-genres. We took the genre descriptions, broke them down to their key words, ... and built our own new-genre generator.

ALEXIS C. MADRIGAL | JAN 2, 2014 | TECHNOLOGY

Emotional Independent Sports Movies
Spy Action & Adventure from the 1930s
Cult Evil Kid Horror Movies
Cult Sports Movies
Sentimental set in Europe Dramas from the 1970s
Visually-striking Foreign Nostalgic Dramas
Japanese Sports Movies
Gritty Discovery Channel Reality TV
Romantic Chinese Crime Movies
Mind-bending Cult Horror Movies from the 1980s
Dark Suspenseful Sci-Fi Horror Movies
Gritty Suspenseful Revenge Westerns
Violent Suspenseful Action & Adventure from the 1980s
Time Travel Movies starring William Hartnell
Romantic Indian Crime Dramas
Evil Kid Horror Movies

ADAM ROGERS BUSINESS 04.02.18 07:00 AM

HOW GRUBHUB ANALYZED 4,000 DISHES TO PREDICT YOUR NEXT ORDER

SHARE

f SHARE 551

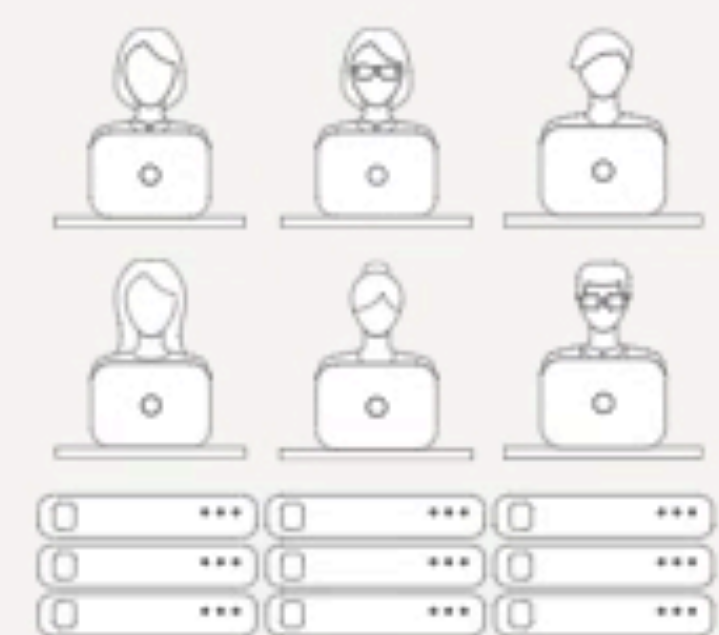
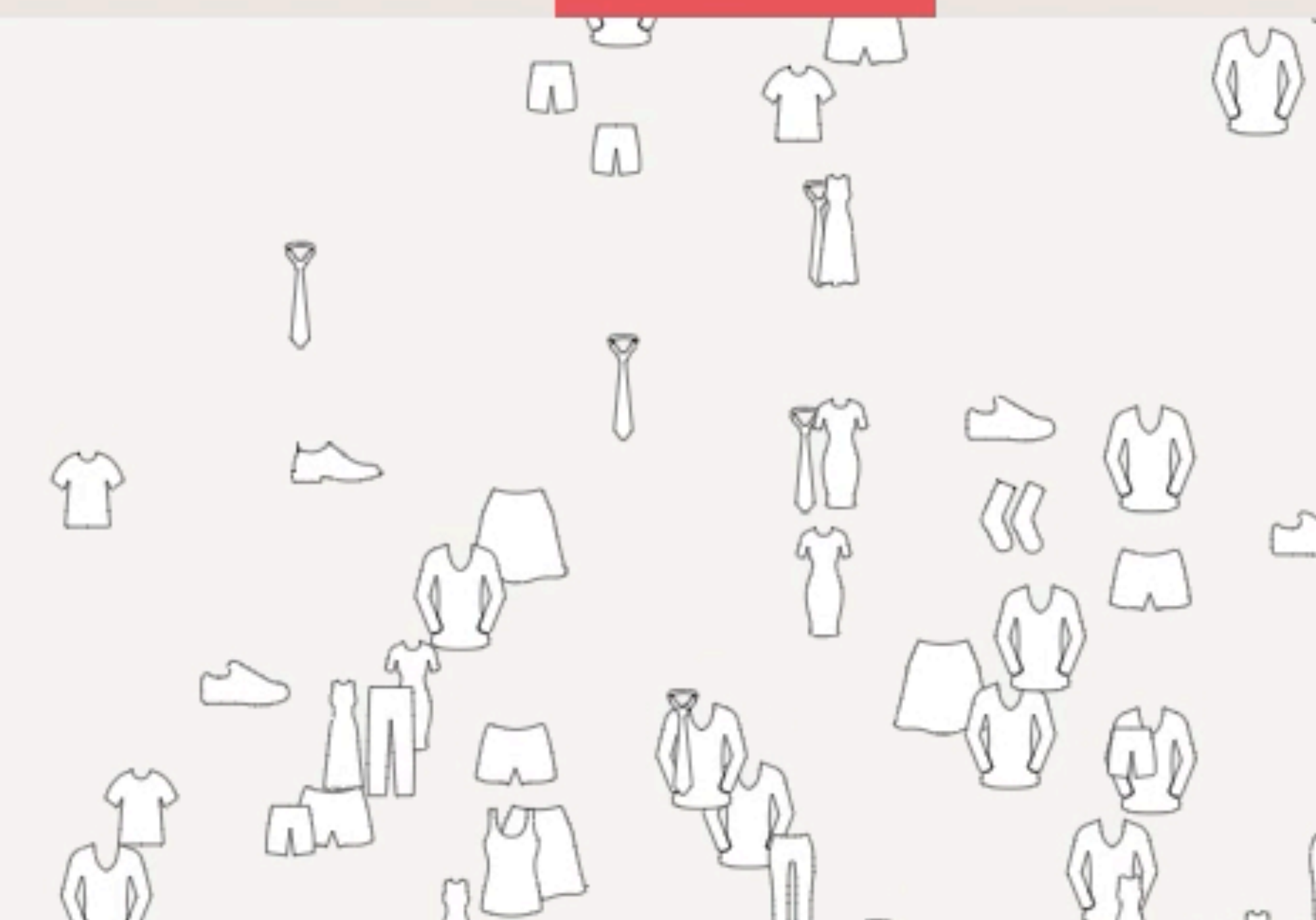
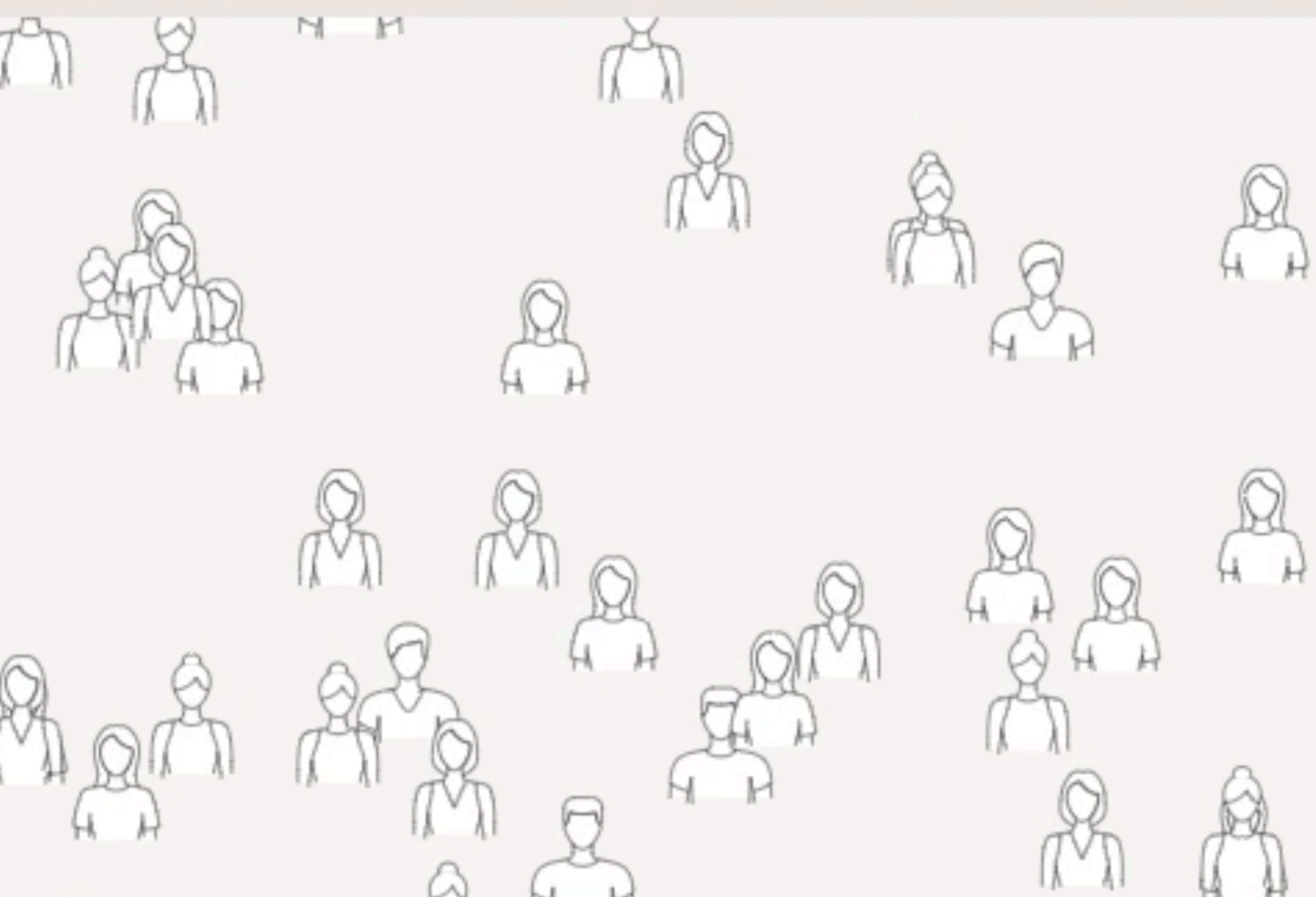
TWEET

COMMENT

EMAIL



JONATHAN KITCHEN/GETTY IMAGES



Algorithms Tour

How data science is woven into the fabric of Stitch Fix

$$\log \frac{p}{1-p} = a + X\beta + Zb$$

...

$$\min_a \sum_i \sum_j a_{ij} q_{ij}$$

$$s.t. a_{ij} \in \{0,1\}, \forall i,j$$

$$\sum_j a_{ij} = 1 \forall i$$

$$\sum_i a_{ij} < k, \forall j$$

...

$$\frac{\partial x}{\partial t} = f(x_t, u_t, w_t)$$

...

$$p(i \rightarrow j) = \text{logit}(\beta_0 + \beta_1 x_1 \dots)$$

ANNALS OF CRIME NOVEMBER 27, 2017 ISSUE

THE SERIAL-KILLER DETECTOR

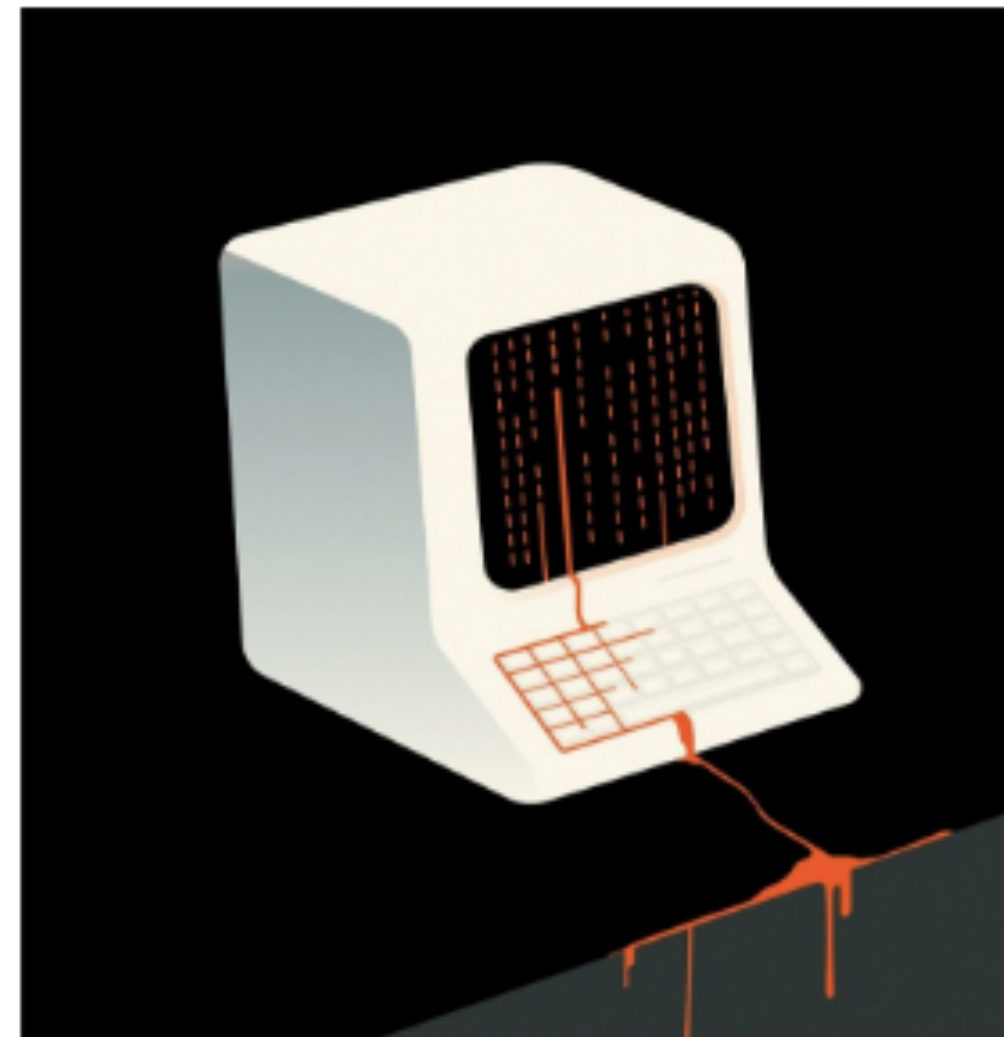
A former journalist, equipped with an algorithm and the largest collection of murder records in the country, finds patterns in crime.



By Alec Wilkinson



Thomas Hargrove is a homicide archivist. For the past seven years, he has been collecting municipal records of murders, and he now has the largest catalogue of killings in the country—751,785 murders carried out since 1976, which is roughly twenty-seven thousand more than appear in F.B.I. files. States are supposed to report murders to the Department of Justice, but some report inaccurately, or fail to report altogether, and Hargrove has sued some of these states to



Hargrove estimates that two thousand serial killers are at large in the U.S.

Illustration by Harry Campbell



PRODUCT MARCH 28, 2018

Detecting Crisis: An AI Solution

by Ankit Gupta, Senior Data Scientist

DATA SCIENCE

TECH

AI

MACHINE LEARNING

SUICIDE PREVENTION

Content warning: This post references words and phrases associated with suicide, in the context of how Crisis Text Line identifies texters at most imminent risk.

Editor's Note: In July 2017, The Cool Calm presented a post on how Crisis Text Line was using machine learning to triage texters by severity. This post follows up on the evolution of that product.

It's mid-evening on December 1, 2017. A post goes viral on Instagram, resulting in record texter volume at Crisis Text Line. Hundreds of volunteer Crisis Counselors pour into the system to respond. An opening message from one texter reads (paraphrased for confidentiality):

"I just took an overdose of Lithium and I'm letting it build up in my system for a few days."

The triaging algorithm flags this message as high-risk for a suicide attempt, and moves it straight to the top of the queue. A Crisis Counselor responds within 20 seconds. Within an hour, the texter has been located by emergency services, and is safe. This is the power of data at scale. Here's how we did it:



AND, A SHORT DISTANCE
AWAY...

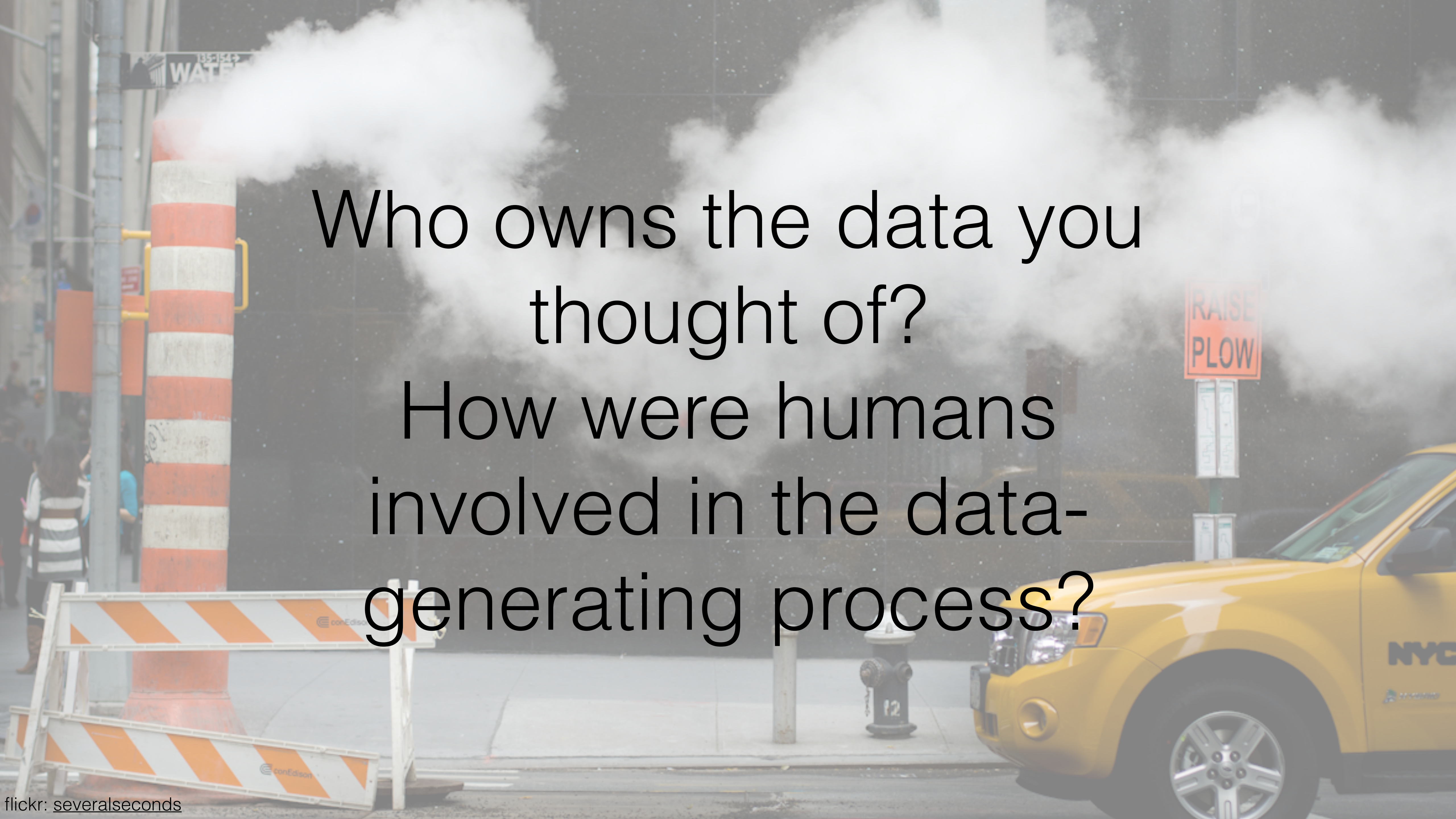
MY FAULT--ALL
MY FAULT! IF
ONLY I HAD
STOPPED HIM
WHEN I **COULD**
HAVE, BUT I
DIDN'T--AND NOW
--UNCLE BEN--
IS DEAD...



AND A LEAN, SILENT FIGURE
SLOWLY FADES INTO THE
GATHERING DARKNESS, AWARE
AT LAST THAT IN THIS WORLD,
WITH GREAT POWER THERE
MUST ALSO COME--GREAT
RESPONSIBILITY!




AND SO A LEGEND IS BORN
AND A NEW NAME IS ADDED
TO THE ROSTER OF THOSE
WHO MAKE THE WORLD OF
FANTASY THE MOST EXCITING
REALM OF ALL!



Who owns the data you
thought of?

How were humans
involved in the data-
generating process?





Data science often serves “the three Ss: science (universities), surveillance (governments), and selling (corporations).”

- Data Feminism, D’Ignazio & Klein



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

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

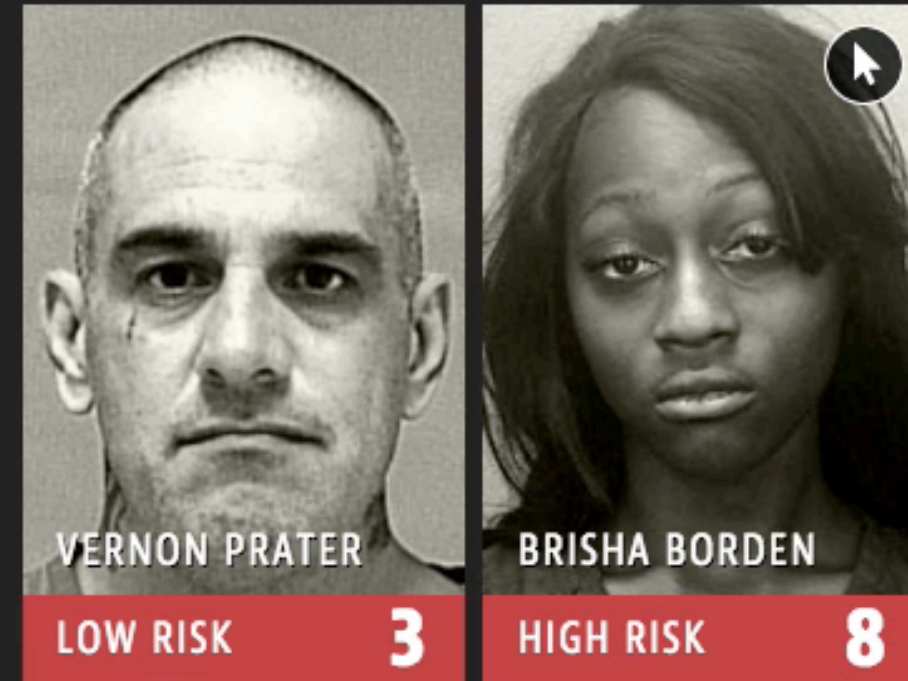
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing **reform bill** currently pending in Congress would mandate the use of such assessments in federal prisons.

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

The sentencing commission did not, however, launch a study of risk scores. So ProPublica did, as part of a larger examination of the powerful, largely

hidden effect of algorithms in American life.

We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the **same benchmark used** by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.



FEATURE

Policing the Future

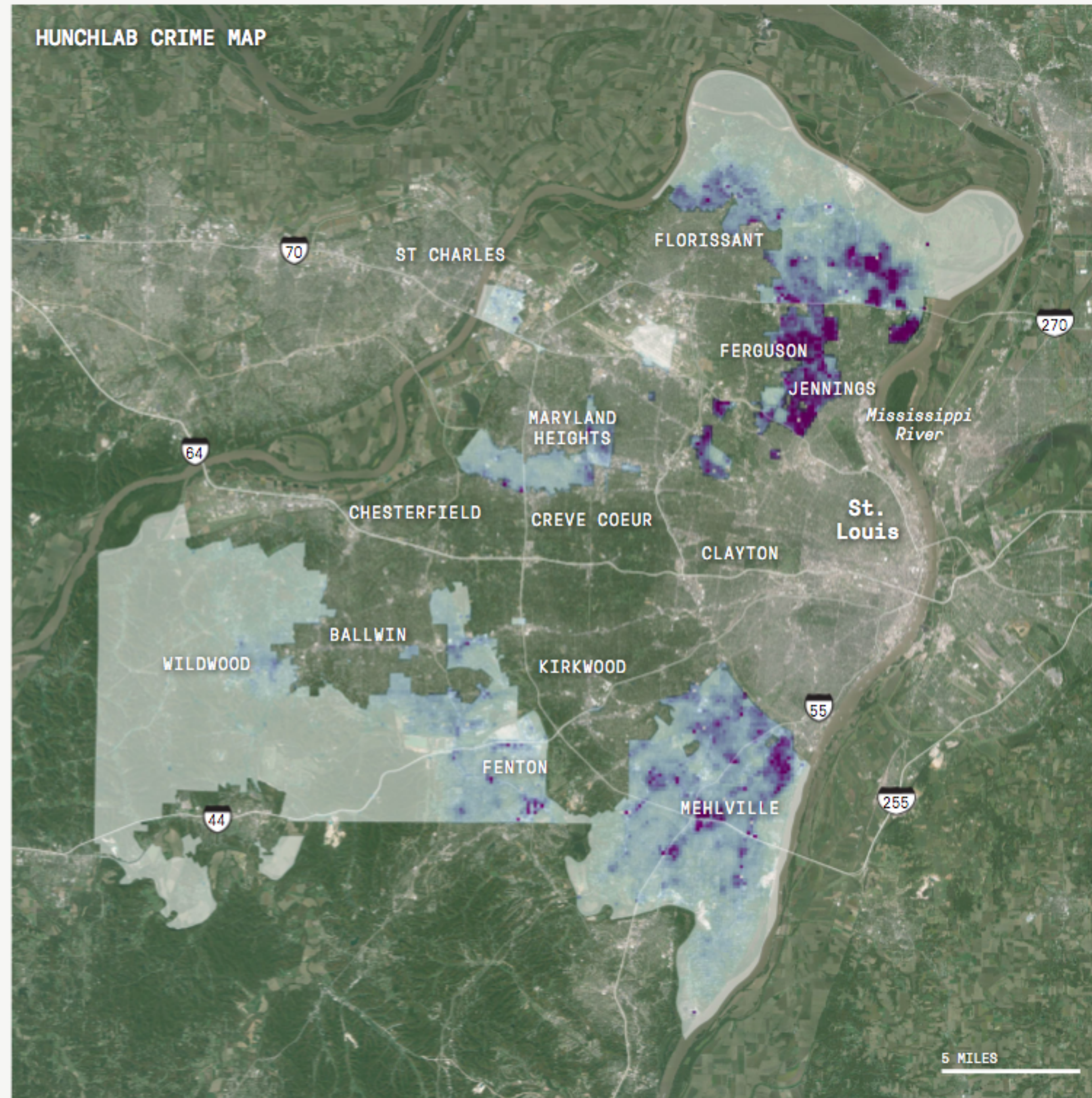
In the aftermath of Michael Brown's death, St. Louis cops embrace crime-predicting software.



Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.
<https://www.themarshallproject.org/2016/02/03/policing-the-future>

Where the St. Louis County Police Patrol

Dozens of small, local municipal agencies handle policing in parts of St. Louis County. The St. Louis County Police Department covers areas not policed by the "munis," including the city of Jennings, Mo. The **DARKER AREAS** in the map show the areas within their jurisdiction that HunchLab has identified as high risk.

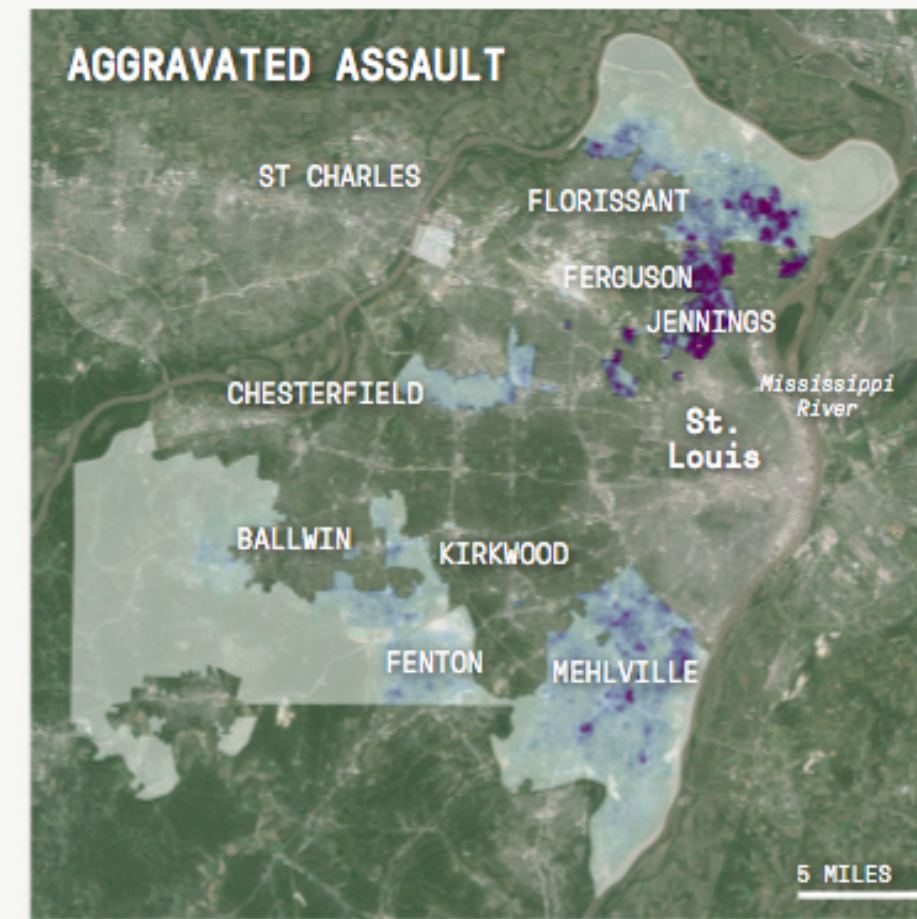
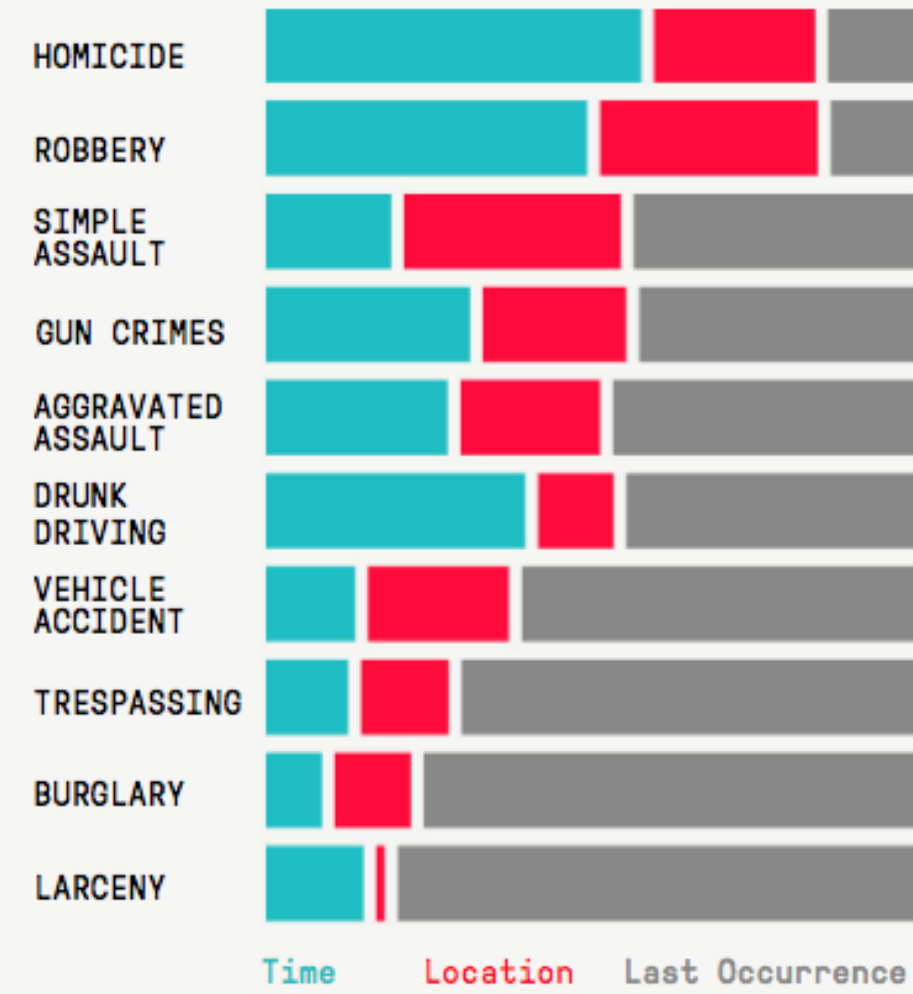


SOURCE: HUNCHLAB

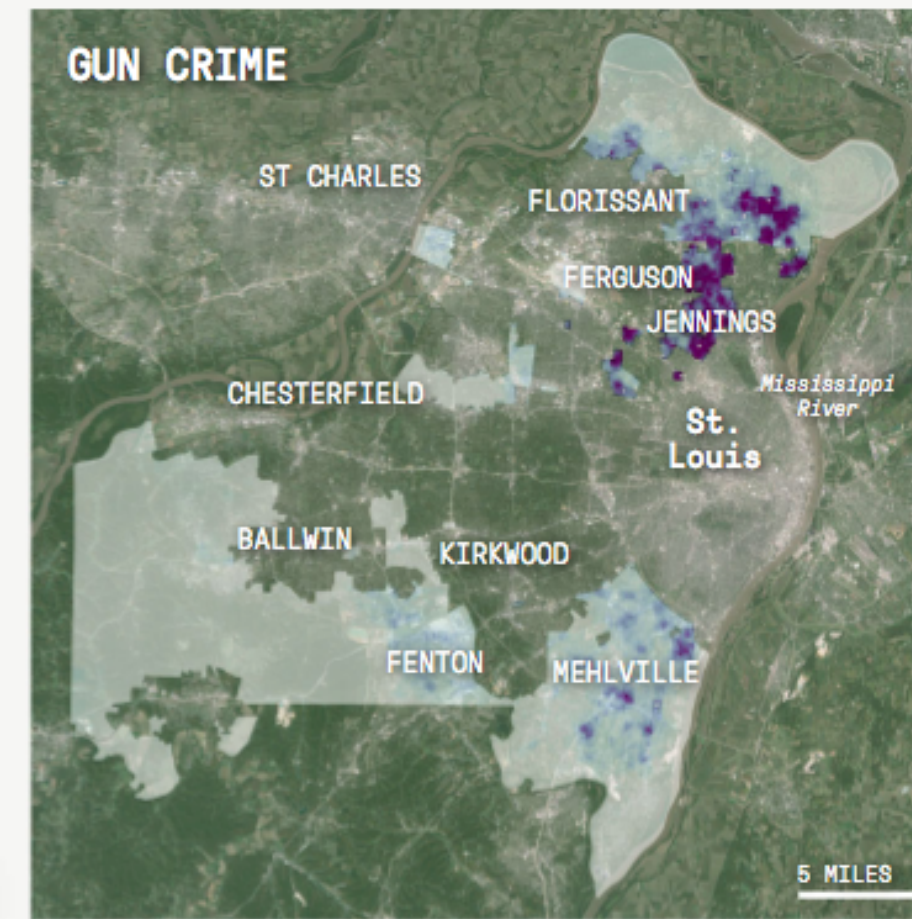
Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.

<https://www.themarshallproject.org/2016/02/03/policing-the-future>

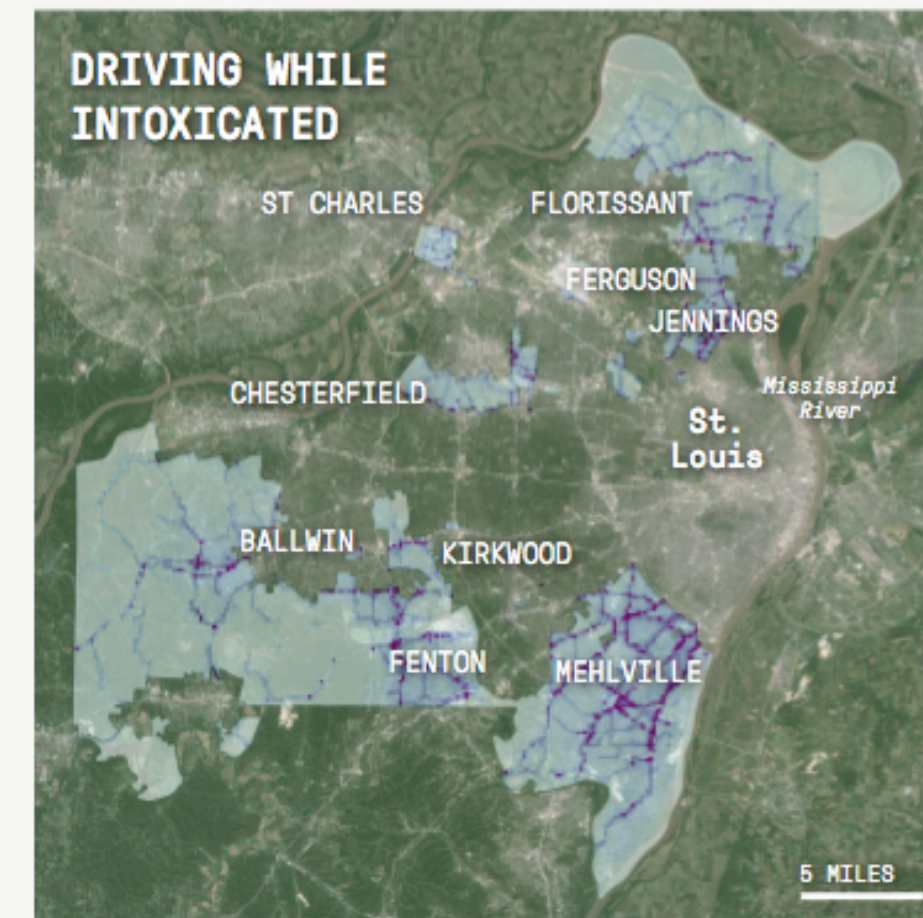
In St. Louis, the HunchLab algorithm took the 10 crimes that the police department had selected, calculated the risk-level for each, and combined them to determine where patrols would have the most impact.



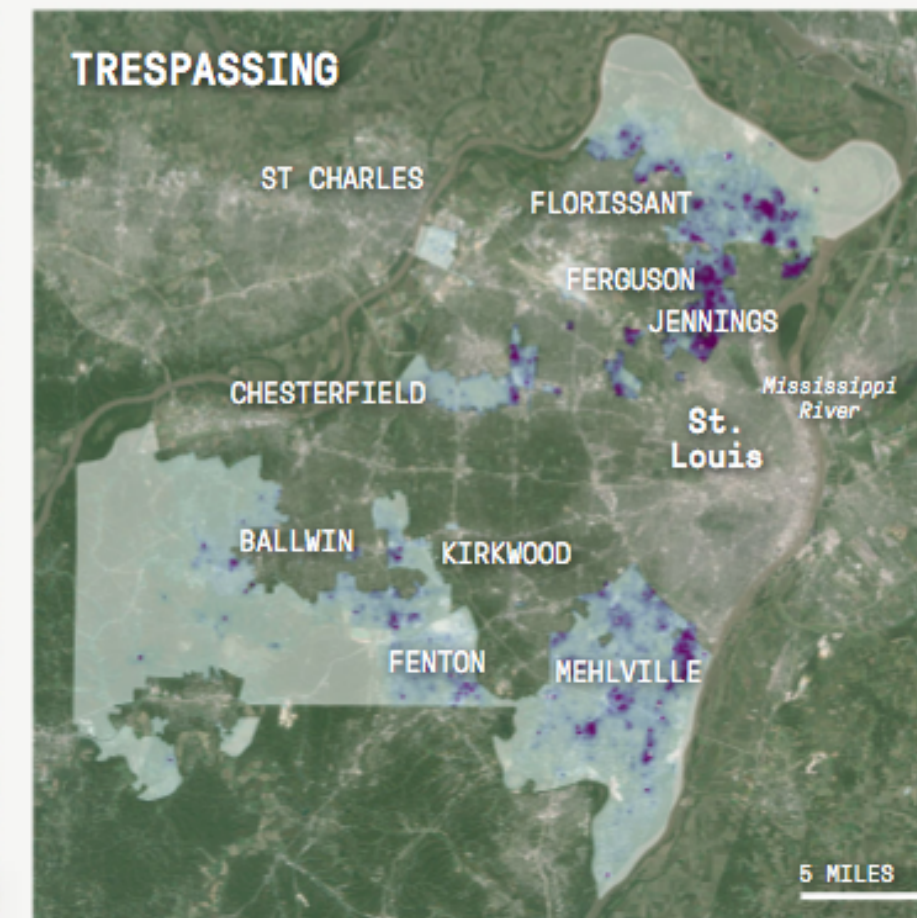
Aggravated assault (assault with a dangerous weapon) makes up 18.5 percent of the overall risk score assigned to a cell. The darkest regions on this map represent cells with a 1 in 320 chance of at least one aggravated assault taking place there during the shift.



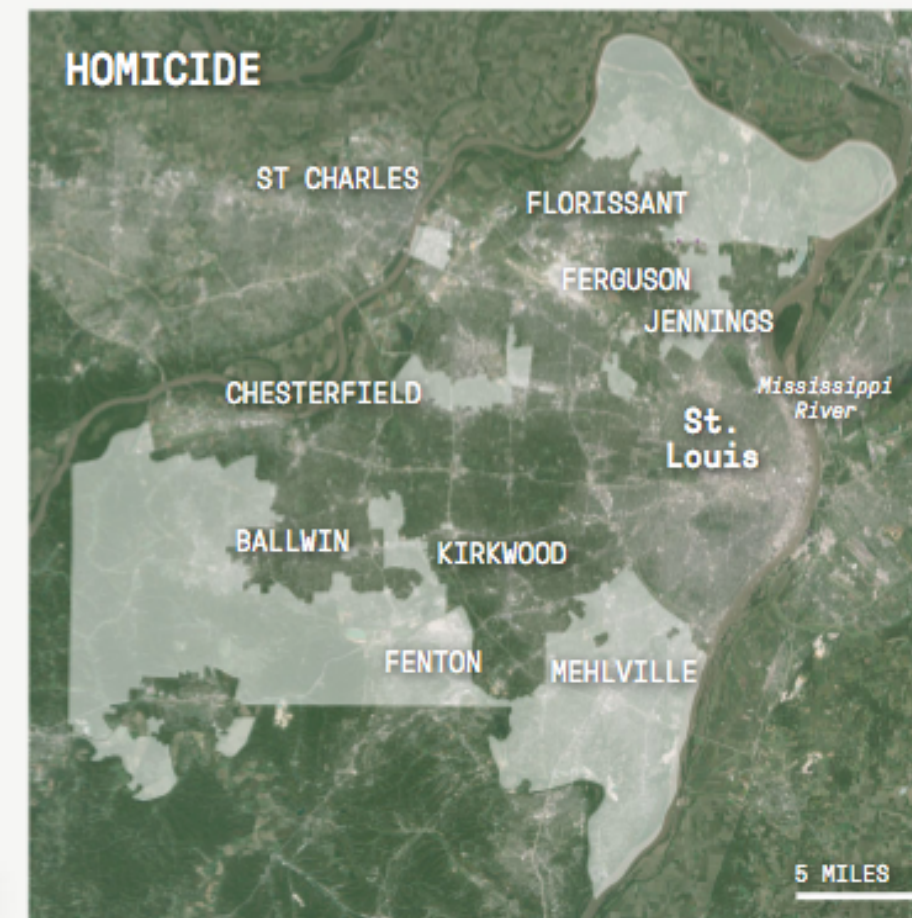
Gun crime (all homicides, robberies, and aggravated assaults with a firearm) makes up about 16.5 percent of the overall risk score. The darkest regions represent a 1 in 850 chance of at least one gun crime taking place.



Driving while intoxicated makes up 10 percent of the total risk score. The darkest regions represent a 1 in 1,300 chance of at least one DWI taking place.



Trespassing makes up about 10 percent of the total risk score. The darkest regions represent cells a 1.7 percent chance of at least one act of trespassing taking place.



Homicides make up 0.66 percent of the total risk score assigned to a cell. The two darkest cells on this map present a 3 percent chance of at least one homicide taking place.

SOURCE: HUNCHLAB

Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.

<https://www.themarshallproject.org/2016/02/03/policing-the-future>

OCT 28, 2013 @ 11:43 AM 42,089 👁

Kroger Knows Your Shopping Patterns Better Than You Do



Tom Groenfeldt, CONTRIBUTOR

I write about finance and technology. [FULL BIO](#) ✓

Opinions expressed by Forbes Contributors are their own.

Kroger KR -1.63%, the Cincinnati-based grocery store chain, calls the 11 million pieces of direct mail it sends to customers each quarter “snowflakes” -- because if any two are the same, it is a fluke. The redemption rate is over 70 percent within six weeks of the mailing.

Kroger is the nation’s largest traditional grocery chain with more than 2,400 stores and \$80.8 billion in sales last year, second only to Wal-Mart in grocery sales. It was named “Retailer of the Year” by **Progressive** PGR -0.08% Grocer magazine. “They have made significant investments in a best-in-class loyalty program, strong private label, and reinvested in their stores and technology,” Neil Stern at McMillanDoolittle said of the award, as reported by Progressive Grocer.



What do others think of you?

See a profile of what your browsing history suggests about you. Paste URLs from your web history into the box below. (Help)

Browsing history

Submit

[About](#) | [Samples](#) | [Comparison](#) | [Privacy](#)

This service is currently under development. Stay tuned...

data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

How unique are you?

Enter your ZIP code, date of birth, and gender to see how unique you are (and therefore how easy it is to identify you from these values).

Date of Birth

Gender Male Female

5-digit ZIP

Submit

[About](#) | [Samples](#) | [Harvard](#) | [Harvard Multi Years](#)



MACHINE BIAS



Facebook (Still) Letting Housing Advertisers Exclude Users by Race

After ProPublica revealed last year that Facebook advertisers could target housing ads to whites only, the company announced it had built a system to spot and reject discriminatory ads. We retested and found major omissions.

by Julia Angwin, Ariana Tobin and Madeleine Varner, Nov. 21, 2017, 1:23 p.m. EST



Facebook CEO Mark Zuckerberg speaks in San Jose, California, in October 2016. (David Paul Morris/Bloomberg via Getty Images)

FOLLOW PROPUBLICA

Twitter

Facebook

Podcast

RSS

Get our stories by email.

Email

Subscribe

MOST POPULAR STORIES

Most Read

Most Emailed

[Billion-Dollar Blessings](#)

[We're Hiring, a Lot. Here's What We're Looking For.](#)

[The Company Michael Cohen Kept — "Trump, Inc."](#)

POLICY / TECH / LABOR

Stitch Fix stylists reportedly quit in droves as the company leans on algorithms to serve customers

The new CEO said recently its stylists “play a very active role” in training machine learning models

By [Kim Lyons](#) | [@SocialKimLy](#) | Aug 20, 2021, 11:24am EDT

f 🐦 ↗ SHARE



verge deals

Subscribe to get the best Verge-approved tech deals of the week.

Email (required)

By signing up, you agree to our [Privacy Notice](#) and European users agree to the data transfer policy.

SUBSCRIBE



[Mental Health](#) [Mental Health](#)

Crisis Text Line tried to monetize its users. Can big data ever be ethical?

The crisis intervention service had concerns about its financial future, but made a huge mistake.

By [Rebecca Ruiz](#) on February 3, 2022



Crisis Text Line tried to make a business out of user data. It went terribly wrong. Credit: Vicky Leta / Mashable



Dr. Latanya Sweeney



Carnegie Mellon

DATA PRIVACY LAB

SSNs Social Security Numbers

Matching a Person to

Using publicly available information about SSNs and other data, the system identifies the issuing state, date issued, and other details.

Sample uses:

- Job Applications
- Apartment Rentals
- Insurance Claims
- Student Applications

Enter the SSN (9-digits) and select the state of the SSN. The system will automatically learn...

Results for SSN 078-000000000

Geography	New York
Date of issuance	Issued between 1980
Year of Birth (8-digit prefix)	1975 born 1980 to 1990 1975 born 1975 to 1980

If the person presenting the SSN fails to list or acknowledge New York as a prior residence, then it is extremely unlikely that the provided SSN was issued to that person.

Results for SSN

Geography

Date of

Dr. Jake Porway



Dr. Joy Buolamwini



Dr. Julia Stewart Lowndes



Data for Black Lives is a movement of activists, organizers, and mathematicians committed to the mission of using data science to create concrete and measurable change in the lives of Black people. Since the advent of computing, big data and algorithms have penetrated virtually every aspect of our social and economic lives. These new data systems have tremendous potential to empower communities of color. Tools like statistical modeling, data visualization, and crowd-sourcing, in the right hands, are powerful instruments for fighting bias, building progressive movements, and promoting civic engagement.

But history tells a different story, one in which data is too often wielded as an instrument of oppression, reinforcing inequality and perpetuating injustice. Redlining was a data-driven enterprise that resulted in the systematic exclusion of Black communities from key financial services. More recent trends like predictive policing, risk-based sentencing, and predatory lending are troubling variations on the same theme. Today, discrimination is a high-tech enterprise.

The Team



Founder &
Executive
Director

Yeshimabeit Milner



Co-Founder

Lucas Mason-Brown



Director of
Research

Jamelle Watson-
Daniels



National
Organizing
Director

Tawana Petty



Director of
Policy
Innovation

Akina (Aki) Young



Research
Associate

Paul Watkins



Research
Associate

Linda Denson

Principles

CARE PRINCIPLES

PRINCIPLES OF MĀORI
DATA SOVEREIGNTY



CARE Principles for Indigenous Data Governance

The CARE Principles for Indigenous Data Governance can be downloaded here in [summary](#) or [full](#)

LEARN MORE

#BeFAIRandCARE





DONATE

RACIAL JUSTICE REQUIRES ALGORITHMIC JUSTICE. SUPPORT THE MOVEMENT.





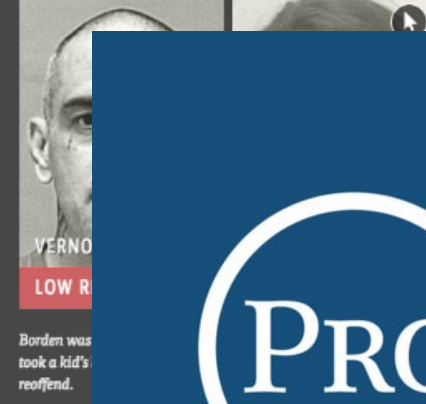
Journalists can be
data science
superheroes

Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing reform bill currently pending in Congress would mandate the use of such assessments in federal prisons.

Two Petty Theft Arrests

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing



algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

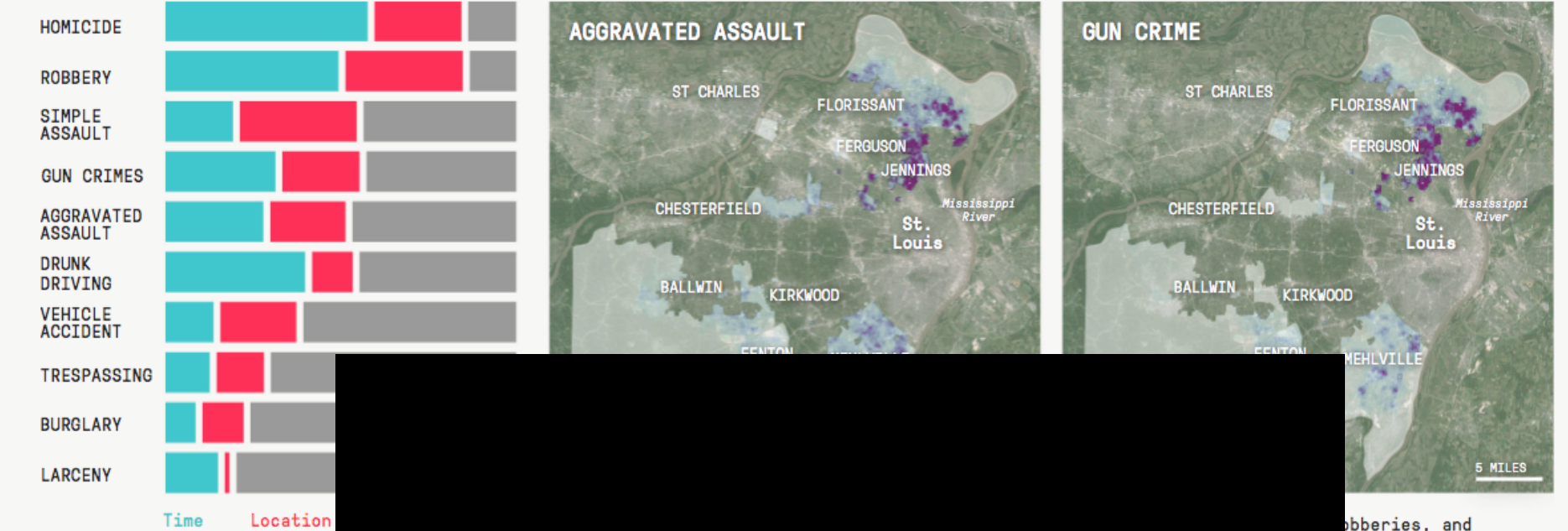
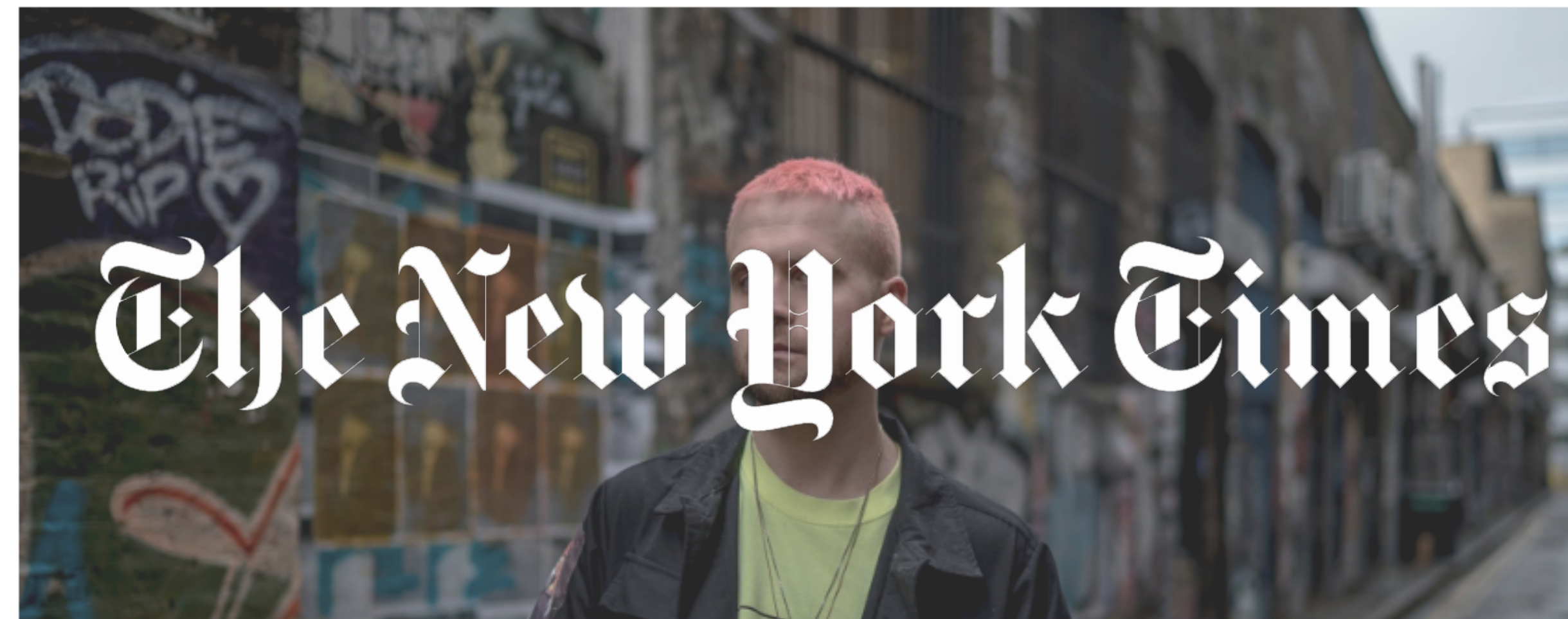
When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

How Trump Consultants Exploited the Facebook Data of Millions

Leer en español

By MATTHEW ROSENBERG, NICHOLAS CONFESSORE and CAROLE CADWALLADR MARCH 17, 2018



Driving while intoxicated makes up 10 percent of the total risk score. The darkest regions represent a 1 in 1,300 chance of at least one DWI taking place.

Trespassing makes up about 10 percent of the total risk score. The darkest regions represent a 1.7 percent chance of at least one act of trespassing taking place.

Homicides make up 0.86 percent of the total risk score assigned to a cell. The two darkest cells on this map present a 3 percent chance of at least one homicide taking place.

SOURCE: HUNCHLAB

Racial bias alleged in Google's ad results

Names associated with blacks prompt link to arrest search



(c)

(d)

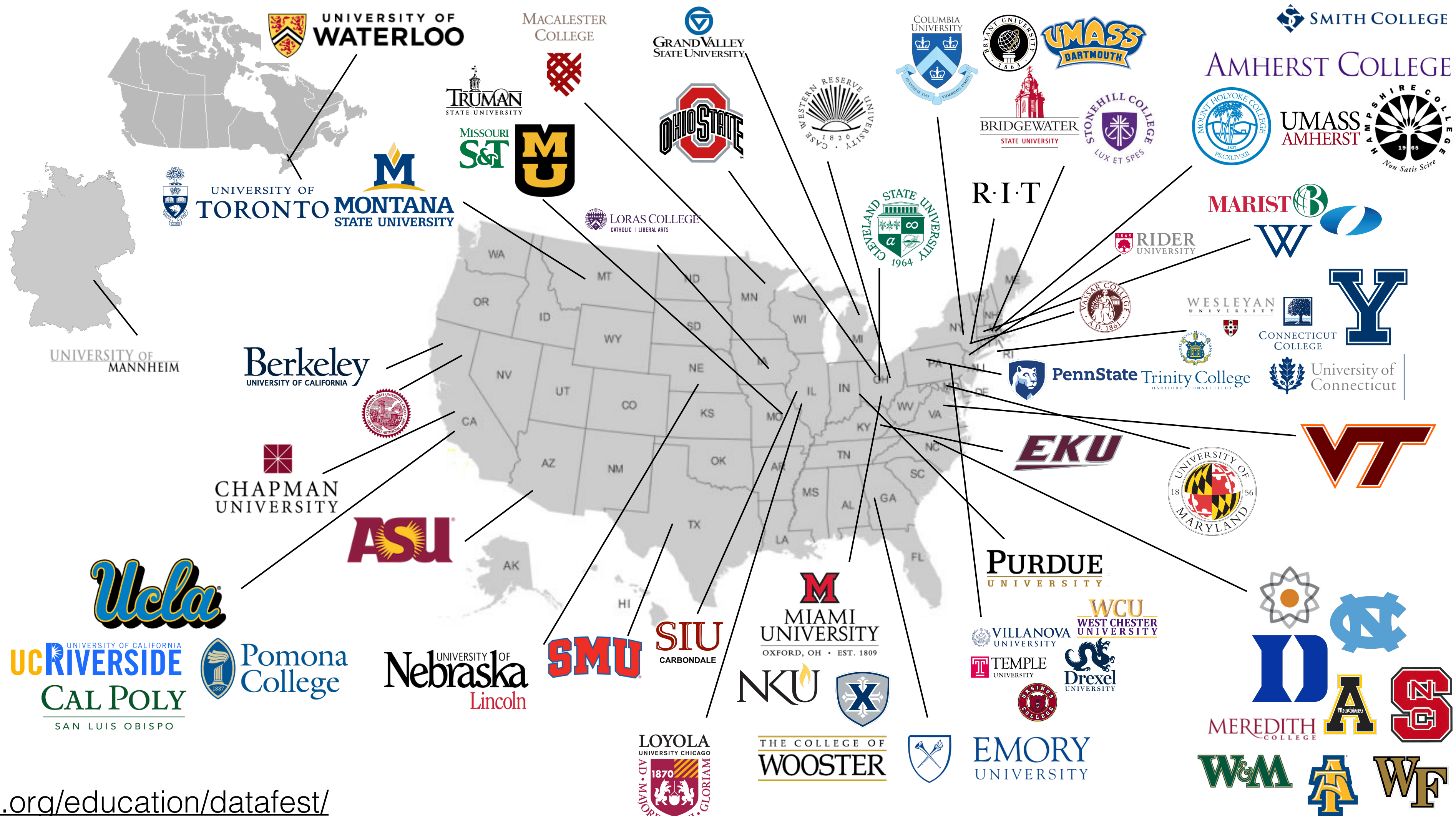
Students and
Teachers can be
data science
superheroes



THE QUANT CRUNCH

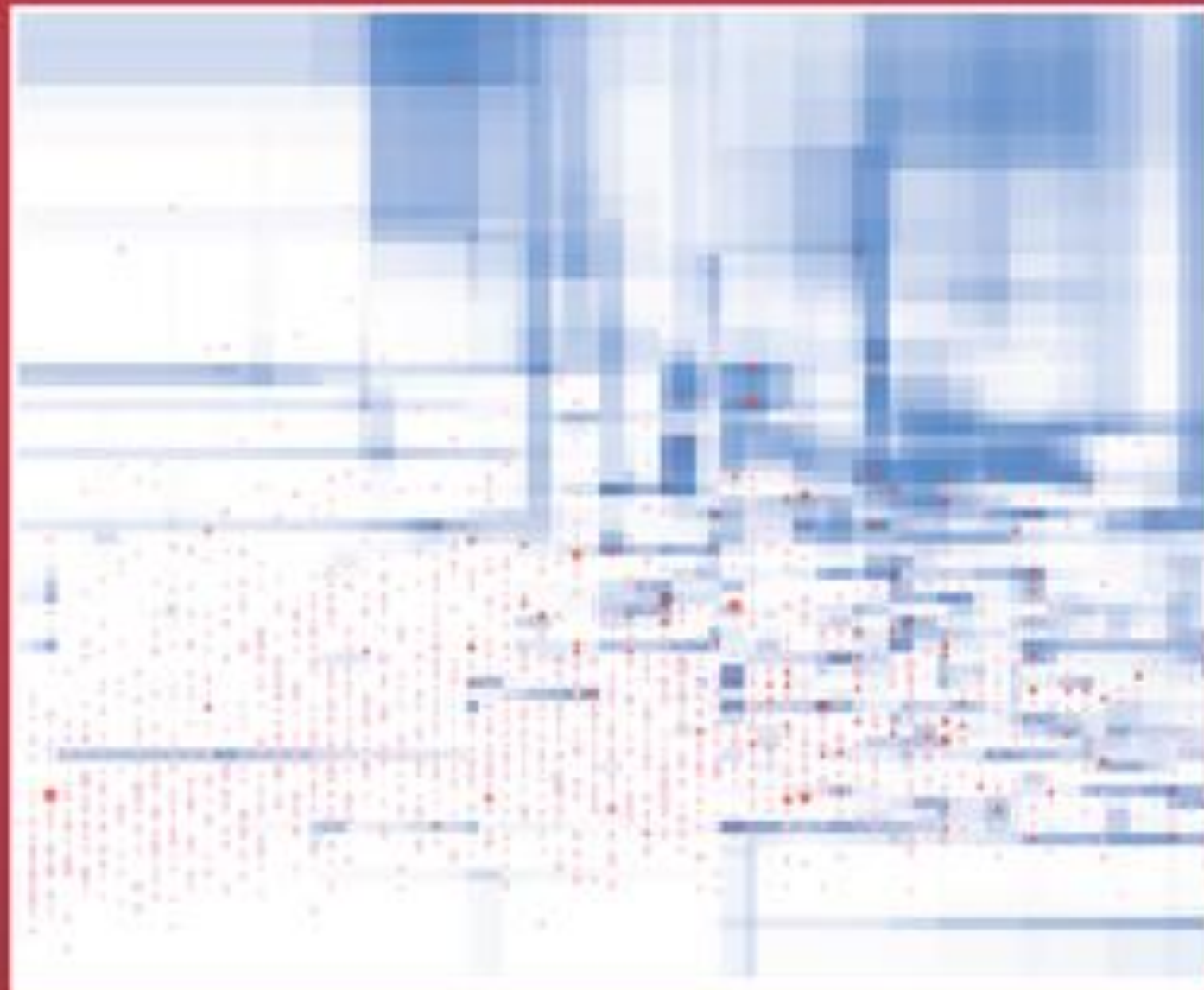
**HOW THE DEMAND FOR
DATA SCIENCE SKILLS
IS DISRUPTING THE JOB
MARKET**

ASA DataFest™



Texts in Statistical Science

Modern Data Science with R



Benjamin S. Baumer
Daniel T. Kaplan
Nicholas J. Horton

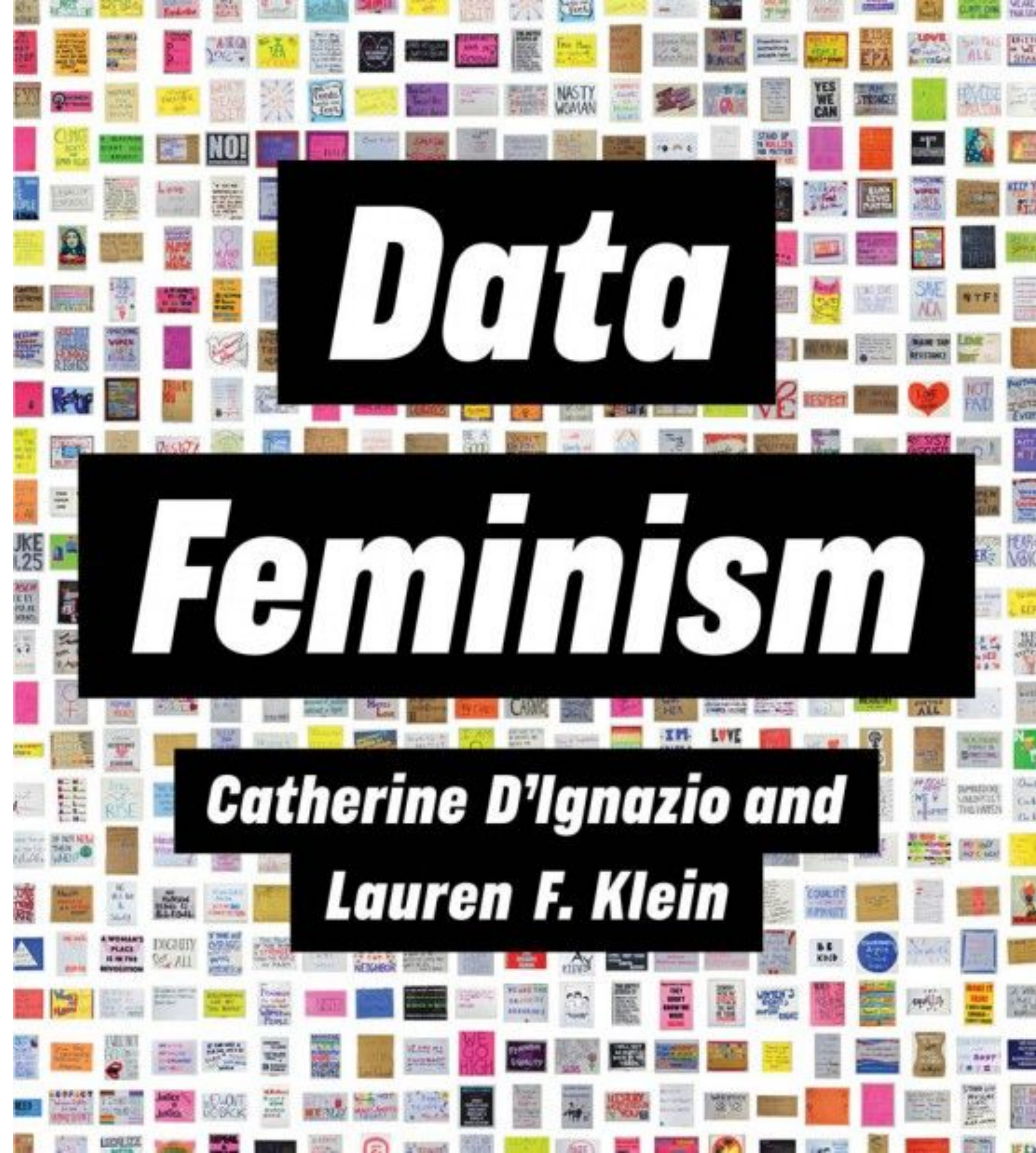
Modern Data Science with R.

Ben Baumer, Danny Kaplan, Nick Horton.

<https://mdsr-book.github.io/mdsr2e/>

 CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

WITH VITALSOURCE®
EBOOK 



Data

Feminism

Catherine D'Ignazio and

Lauren F. Klein

Data Feminism

Catherine D'Ignazio and Lauren F Klein

<https://data-feminism.mitpress.mit.edu/>



Thank you