



Data Science as a Superpower

Amelia McNamara @amelia@vis.social

University of St Thomas St Paul, MN

Department of Computer & Information Sciences



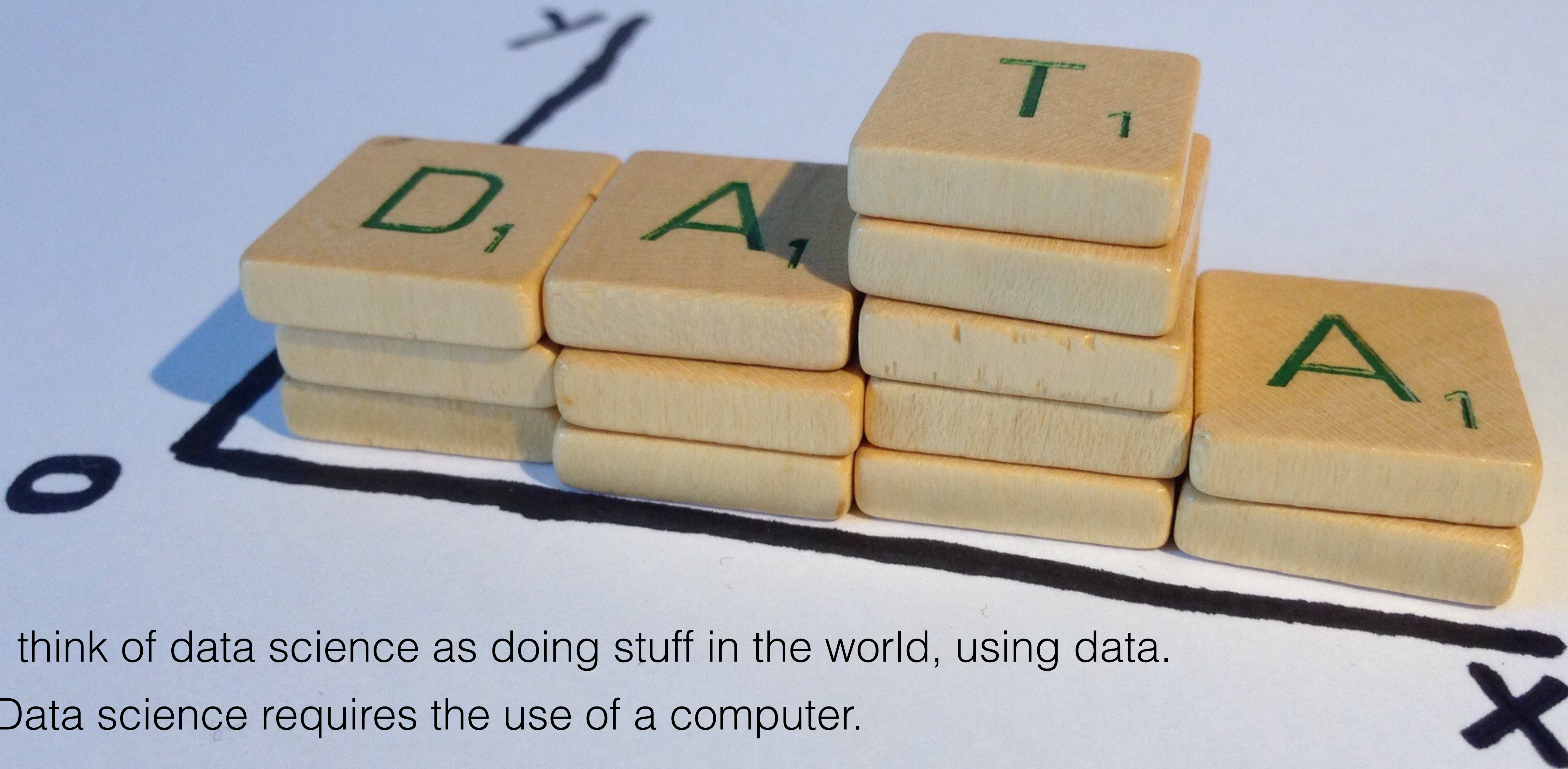
I think about data as any information that we can collect (write down or record on the computer) about the world.

Numbers are data, but images, text, and user behavior can be, too!

Brainstorm: data exhaust

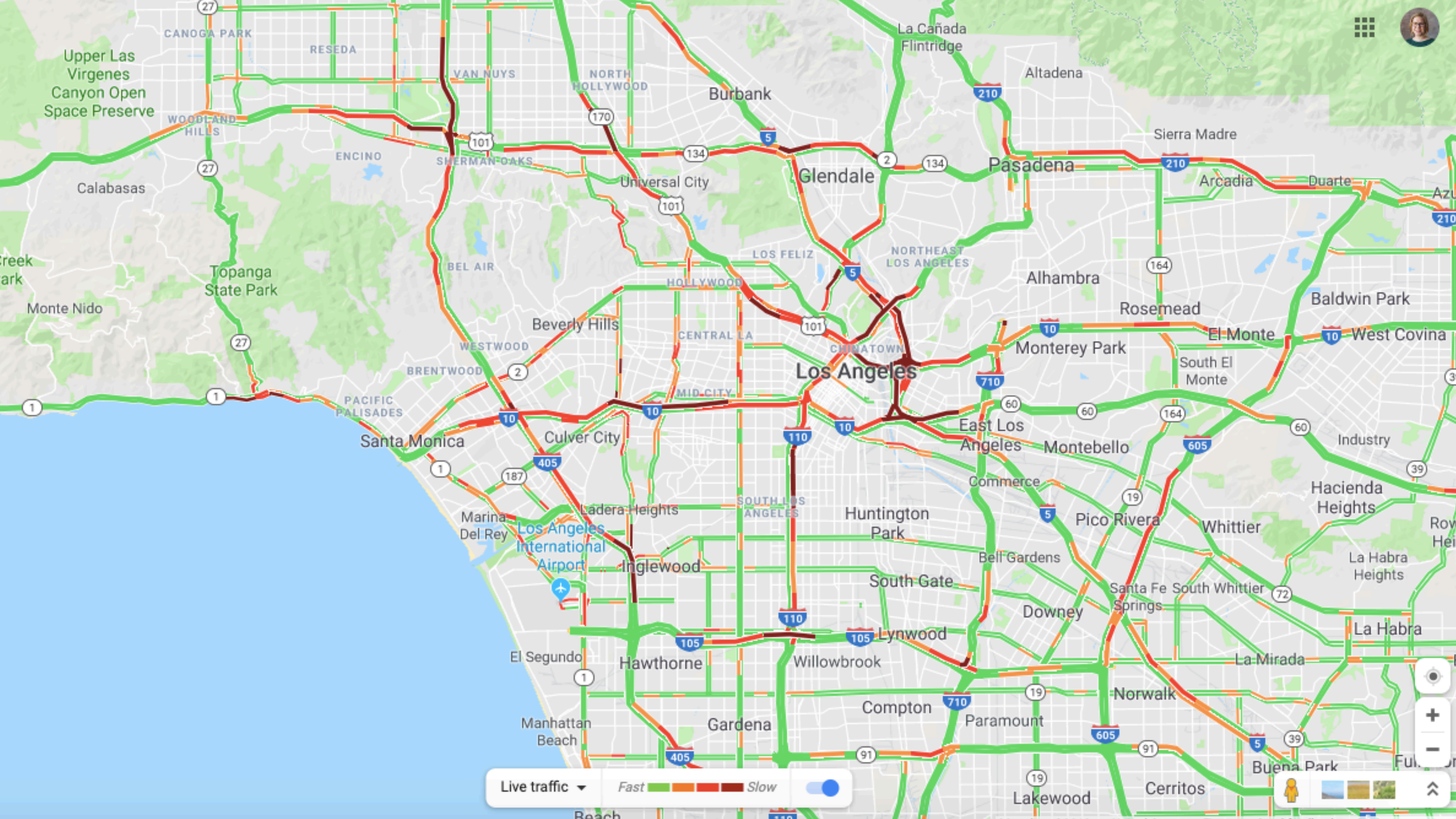
We generate data every day, whether we know it or not.

Take a few minutes and brainstorm some places you generate data exhaust on a normal day.



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

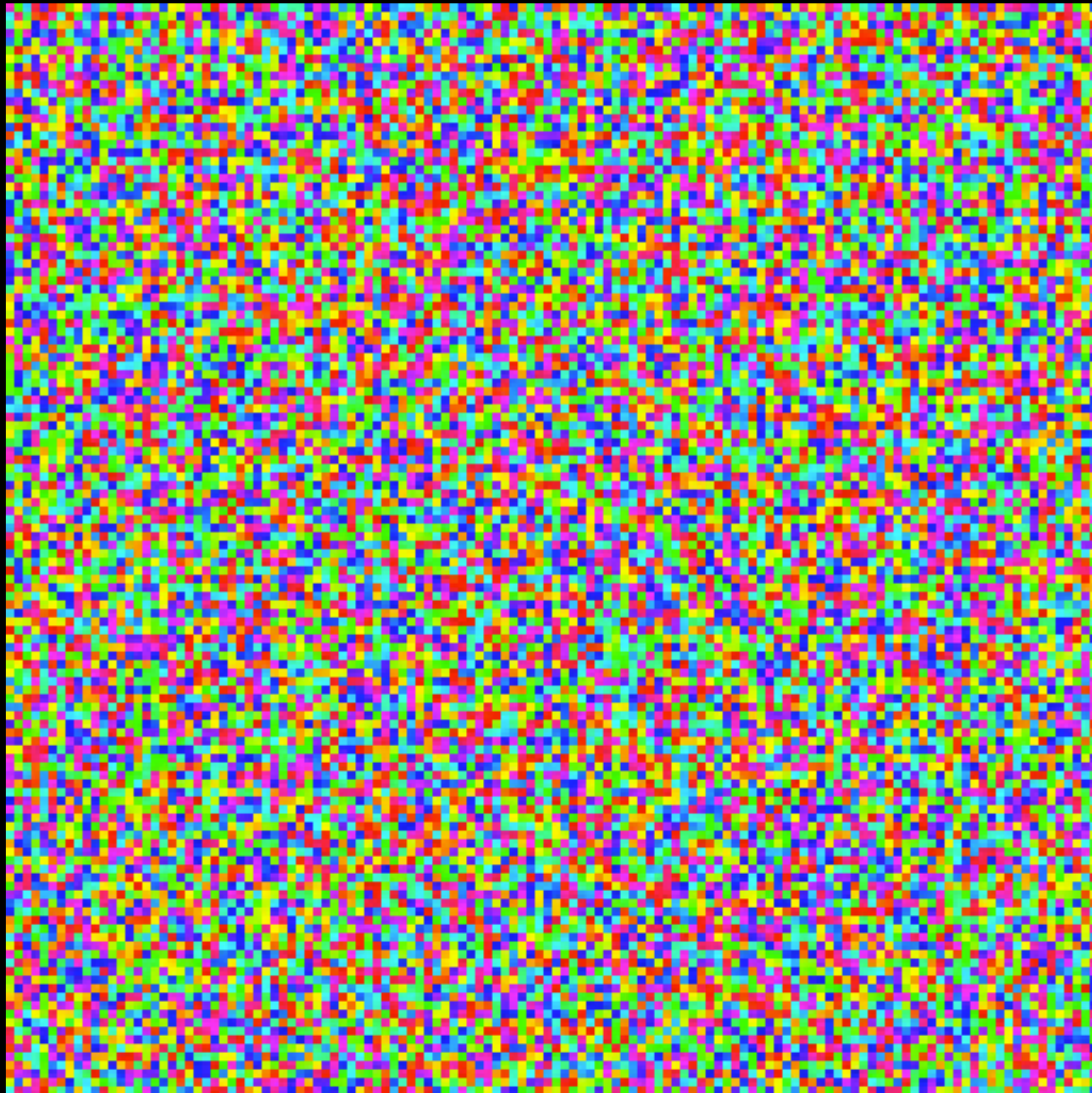
Data science requires the use of a computer.



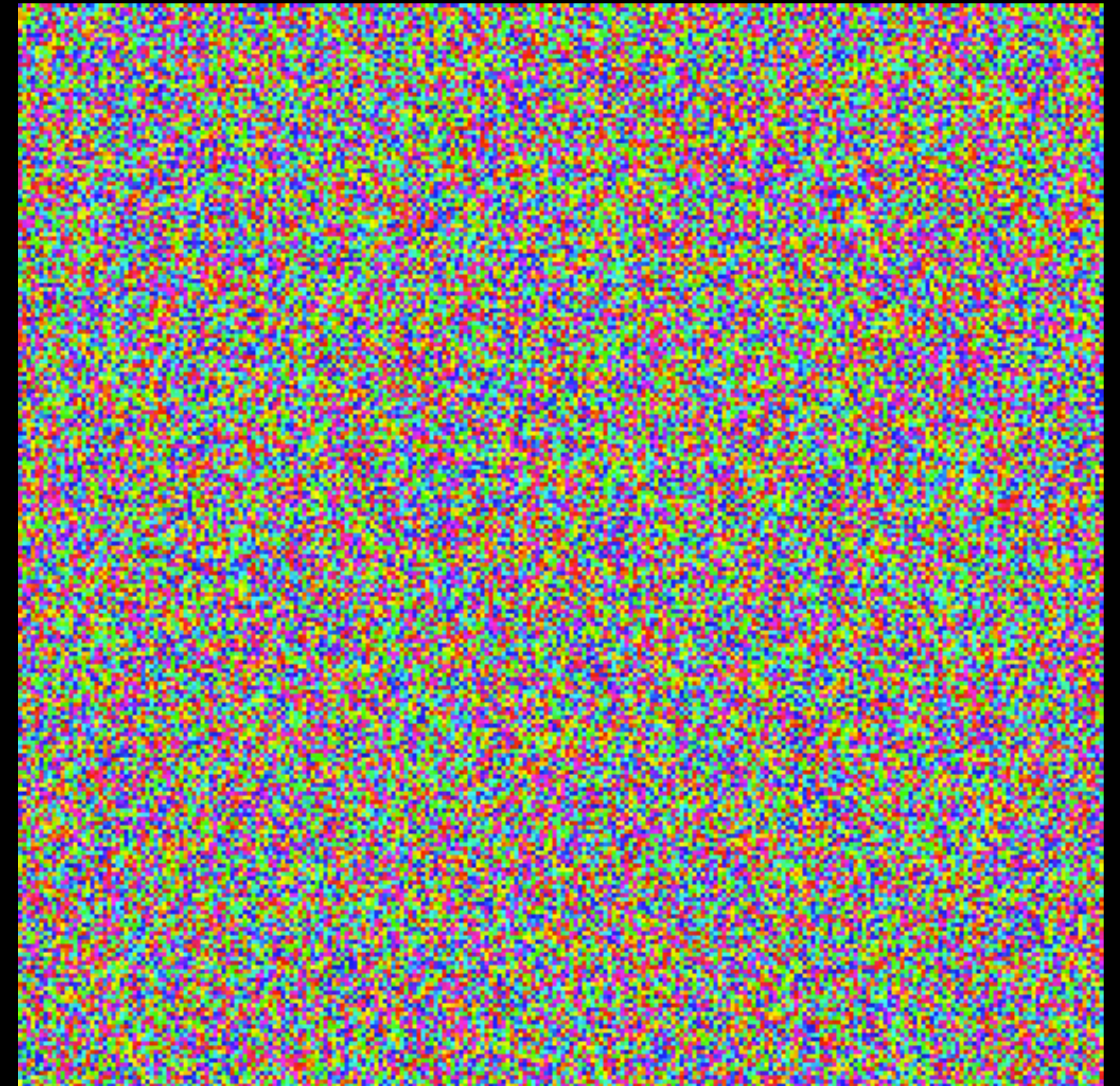
Live traffic ▼ Fast ■ ■ ■ Slow ■

Map navigation controls including a location pin, zoom in (+) and zoom out (-) buttons, a street view pegman icon, and a vertical arrow for map orientation.

An algorithm is a set of instructions
to achieve a specific goal



Merge sort, breadth first



Bubble sort







COVID-19 Surveillance at the Metro Plant

Total Viral Load

Viral Load by Variant

Variant Frequencies (%)

This Week's Summary

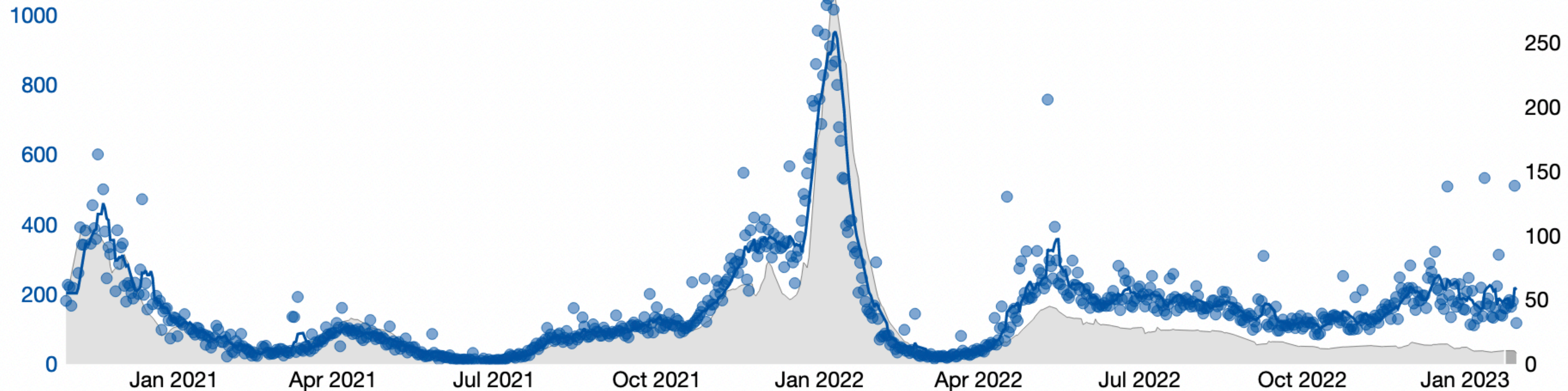
This graph shows the amount of SARS-CoV-2 viral RNA entering the Metro Plant each day (blue line) and the number of new daily COVID-19 cases in the Metro Plant's service area, by sample collection date (gray line; data from the Minnesota Department of Health). The most recent case data (darker gray) are incomplete and subject to change.

Last Sample Date: January 30, 2023.

*All data are preliminary and subject to revision

Viral load in wastewater
M copies/person/day

New Daily COVID-19 cases
per 100K residents 300



● Viral load, Metro Plant service area

— 7-day avg. viral load

— 7-day avg. cases per capita, Metro Service Area

— 7-day avg. cases per capita, Metro Service Area, Incomplete

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





MEET THE ANALYTICS TEAM!



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

ANNALS OF CRIME NOVEMBER 27, 2017 ISSUE

THE SERIAL-KILLER DETECTOR

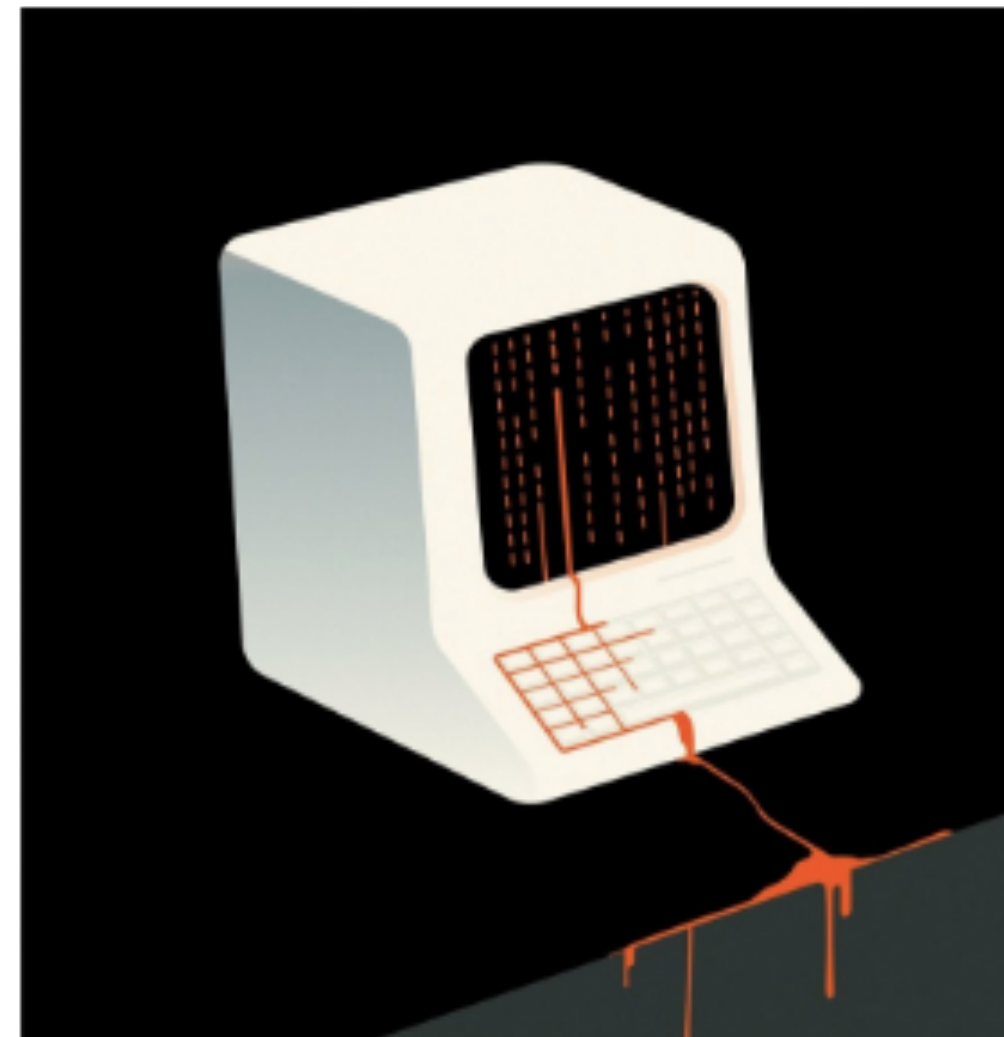
A former journalist, equipped with an algorithm and the largest collection of murder records in the country, finds patterns in crime.



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):

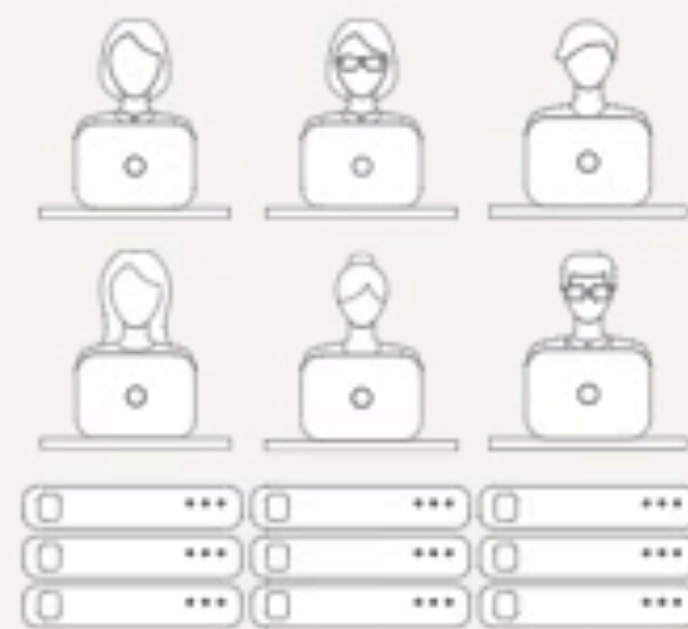
"I just took an overdose of Lithium and I'm letting it build up in my system for a few days."

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:



Algorithms Tour

How data science is woven into the fabric of Stitch Fix



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

...

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

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

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

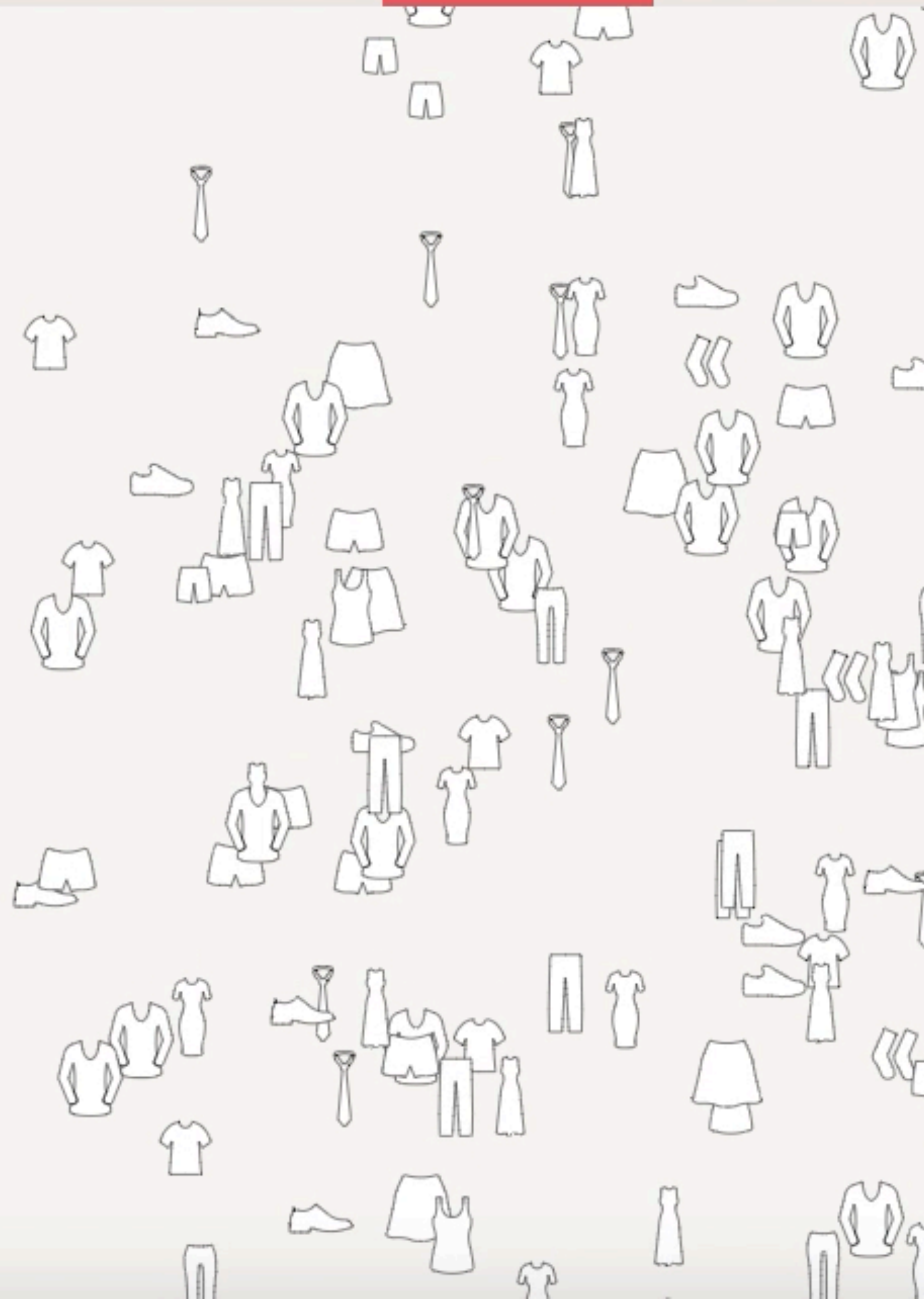
$$\sum_i a_{ij} < k, \forall j$$

...

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

...

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



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!

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

- Data Feminism, D’Ignazio & Klein



Kroger's 84.51° Labs: Pushing the Boundaries of Data Science



Mike O'Brien

October 11, 2022



In 2015, major grocer Kroger acquired the U.S. assets of its data science partner dunnhumby. Out of that grew [84.51°](#), a retail data and analytics unit that encompasses shopper insights, loyalty marketing and retail media advertising.

The unit is credited with helping [Kroger](#) become an omnichannel powerhouse in a sector that had been lagging others in retail, reaching \$10 billion in digital sales in 2020 and seeing 113% growth since. The output from the group benefits Kroger's CPG brand partners, as well.

ADAM ROGERS BUSINESS 04.02.18 07:00 AM

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

SHARE



JONATHAN KITCHEN/GETTY IMAGES



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



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

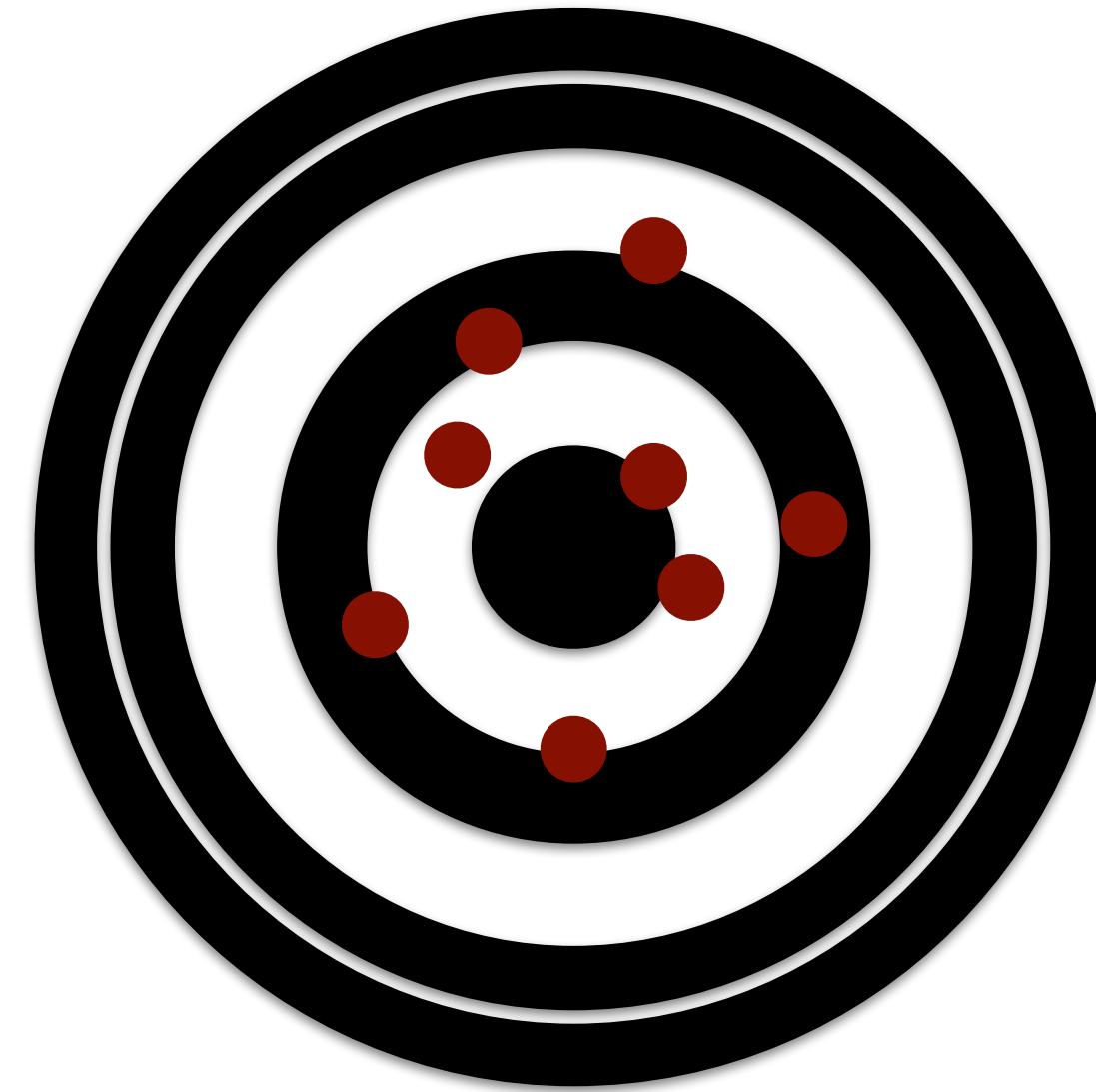
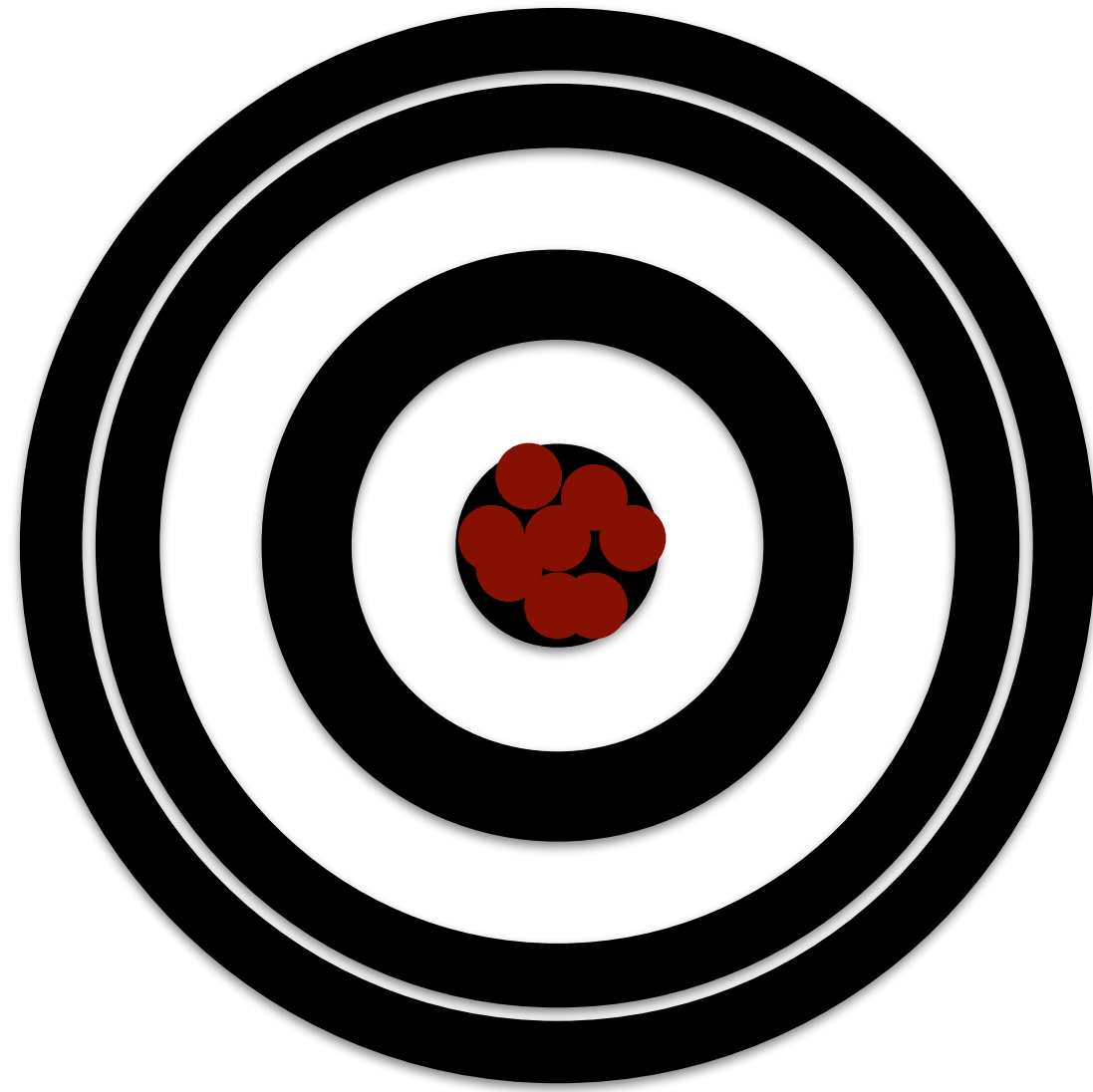
We want to ensure algorithms are fair



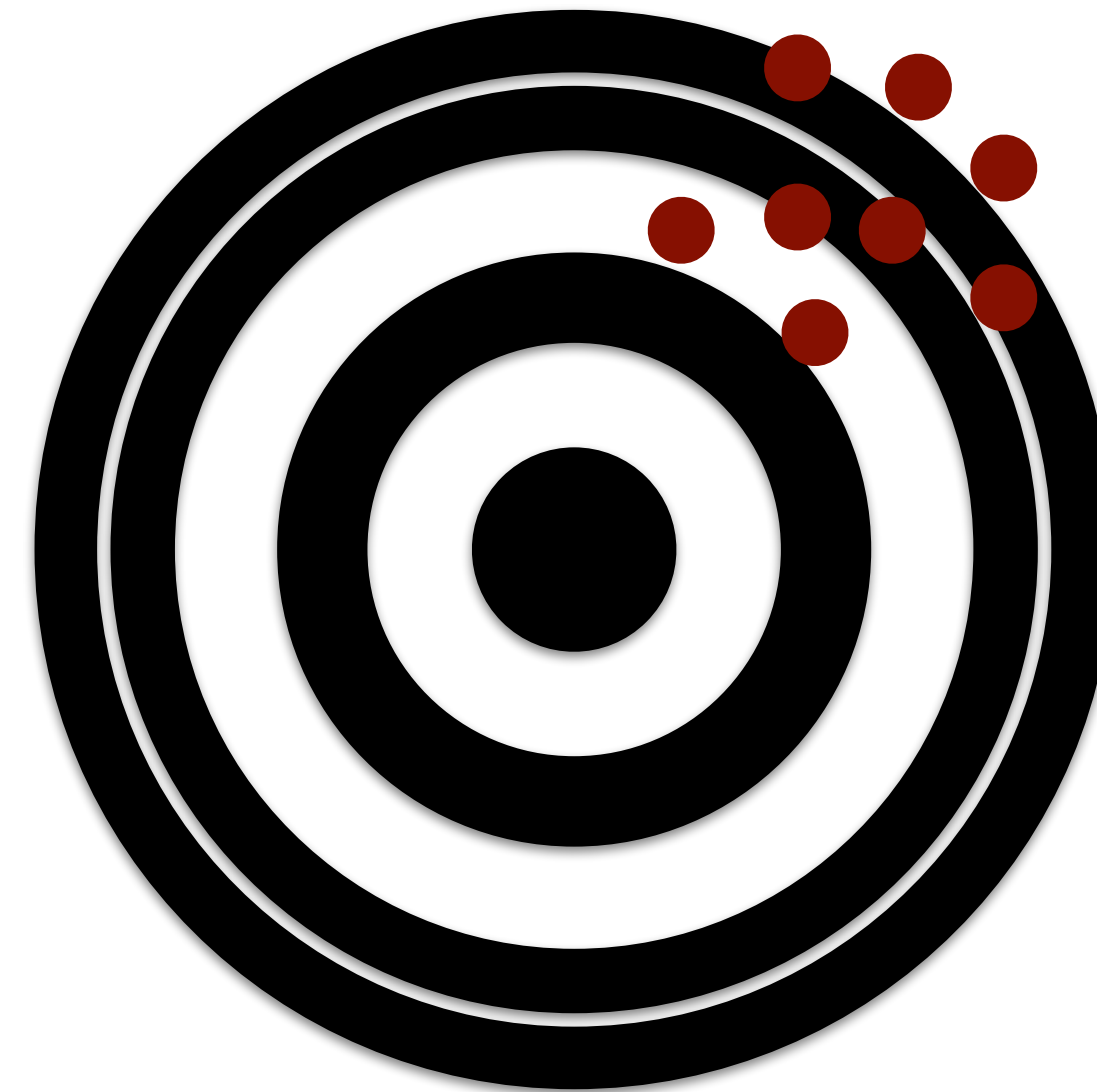
Low variance

High variance

Low bias



High bias



bias noun

Definition of *bias*

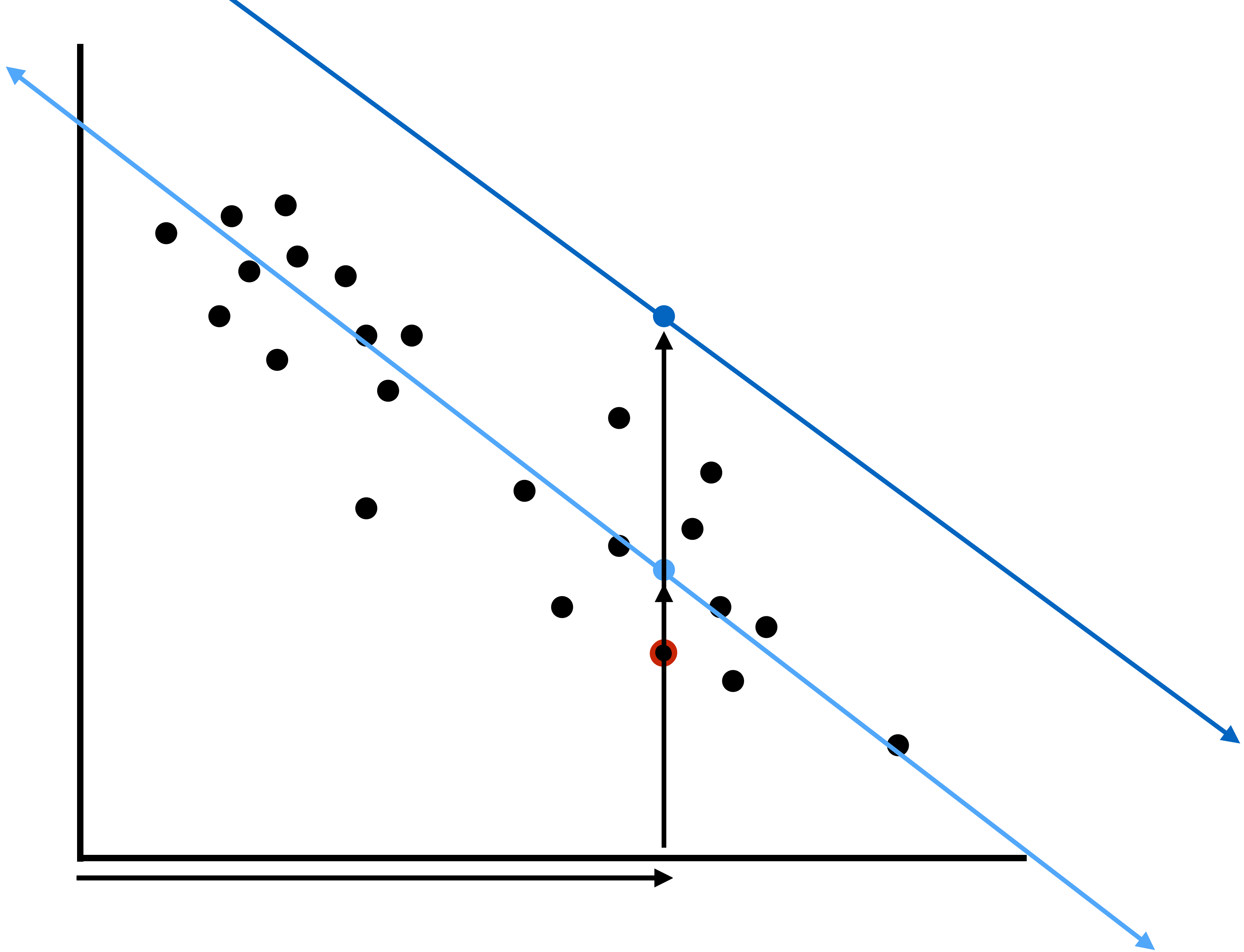
1.

a.

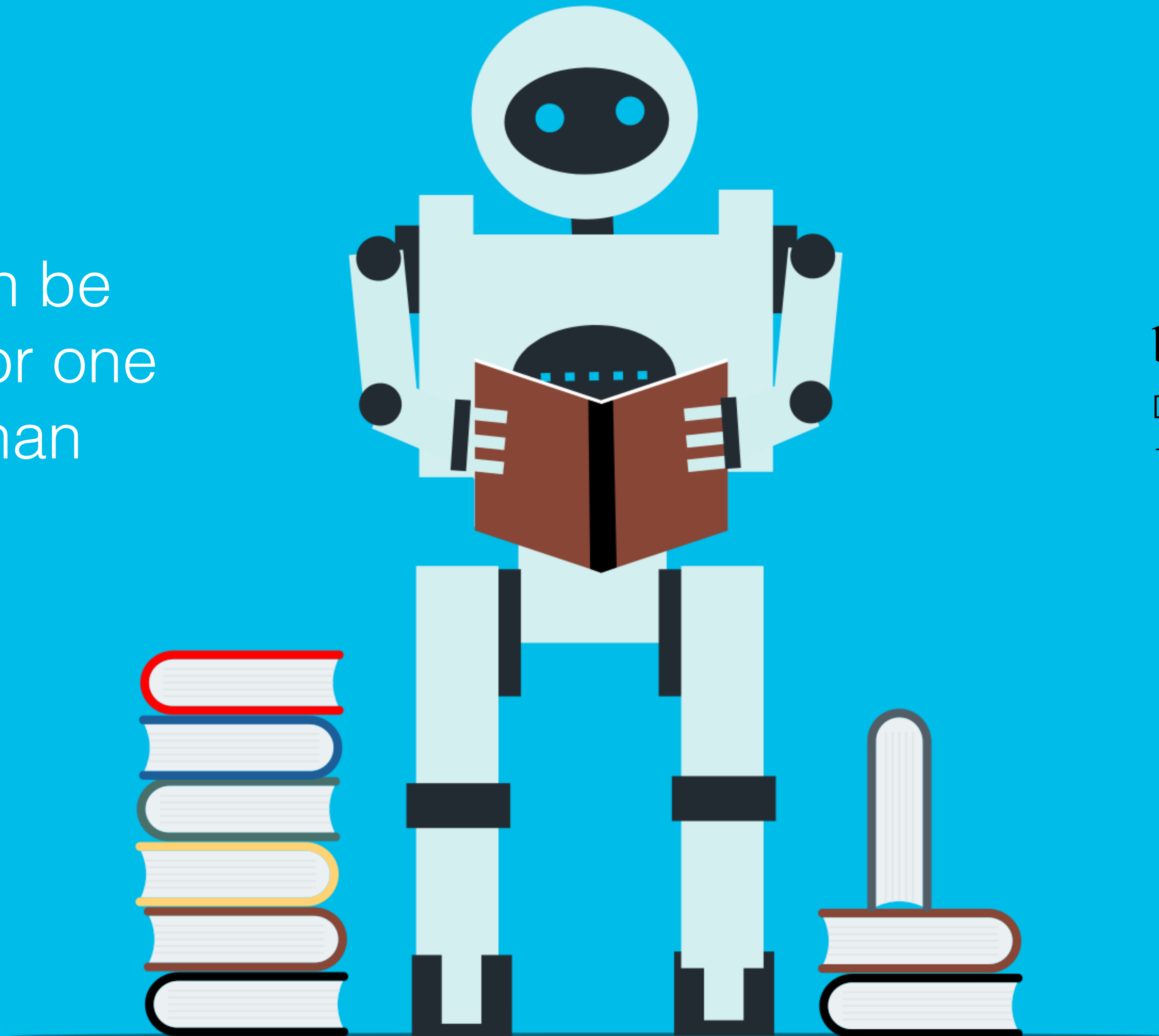
b.

c.

d. (1): deviation of the expected value of a statistical estimate from the quantity it estimates



Bias can be worse for one group than another



bias noun

Definition of *bias*

1.
 - a. an inclination of temperament or outlook
especially: a personal and sometimes unreasoned judgment: prejudice
 - b. an instance of such prejudice

MACHINE BIAS

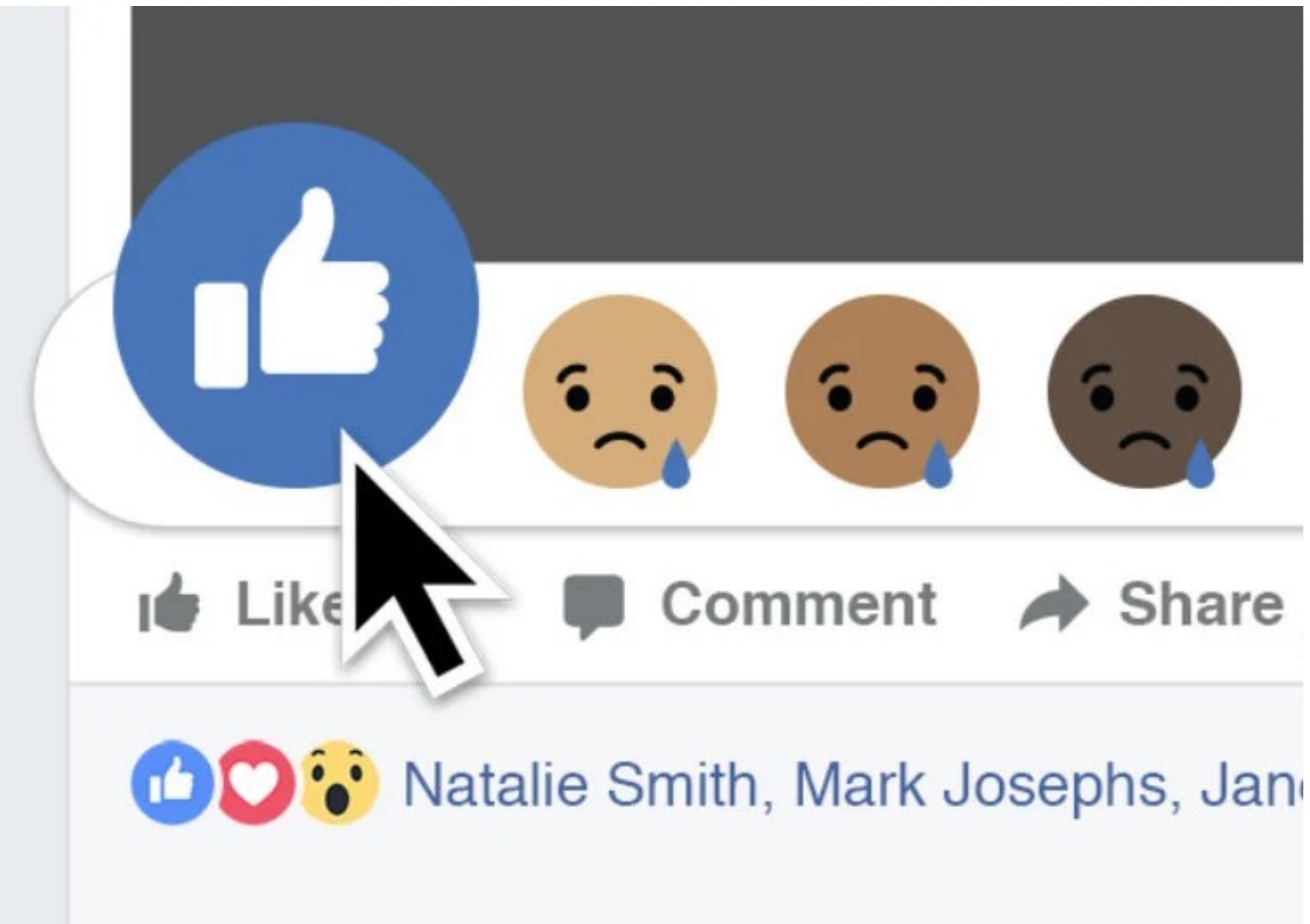


Facebook Lets Advertisers Exclude Users by Race

Facebook's system allows advertisers to exclude black, Hispanic, and other "ethnic affinities" from seeing ads.

by Julia Angwin and Terry Parris Jr., Oct. 28, 2016, 1 p.m. EDT

Oct. 28, 2016



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

Nov. 21, 2017



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



Technology

Facebook Finally Agrees to Eliminate Tool That Enabled Discriminatory Advertising

Six years after ProPublica revealed that Facebook allowed advertisers to exclude Black users and others, the company agreed to a settlement with the Justice Department to overhaul its ad algorithm system.

June 22, 2022

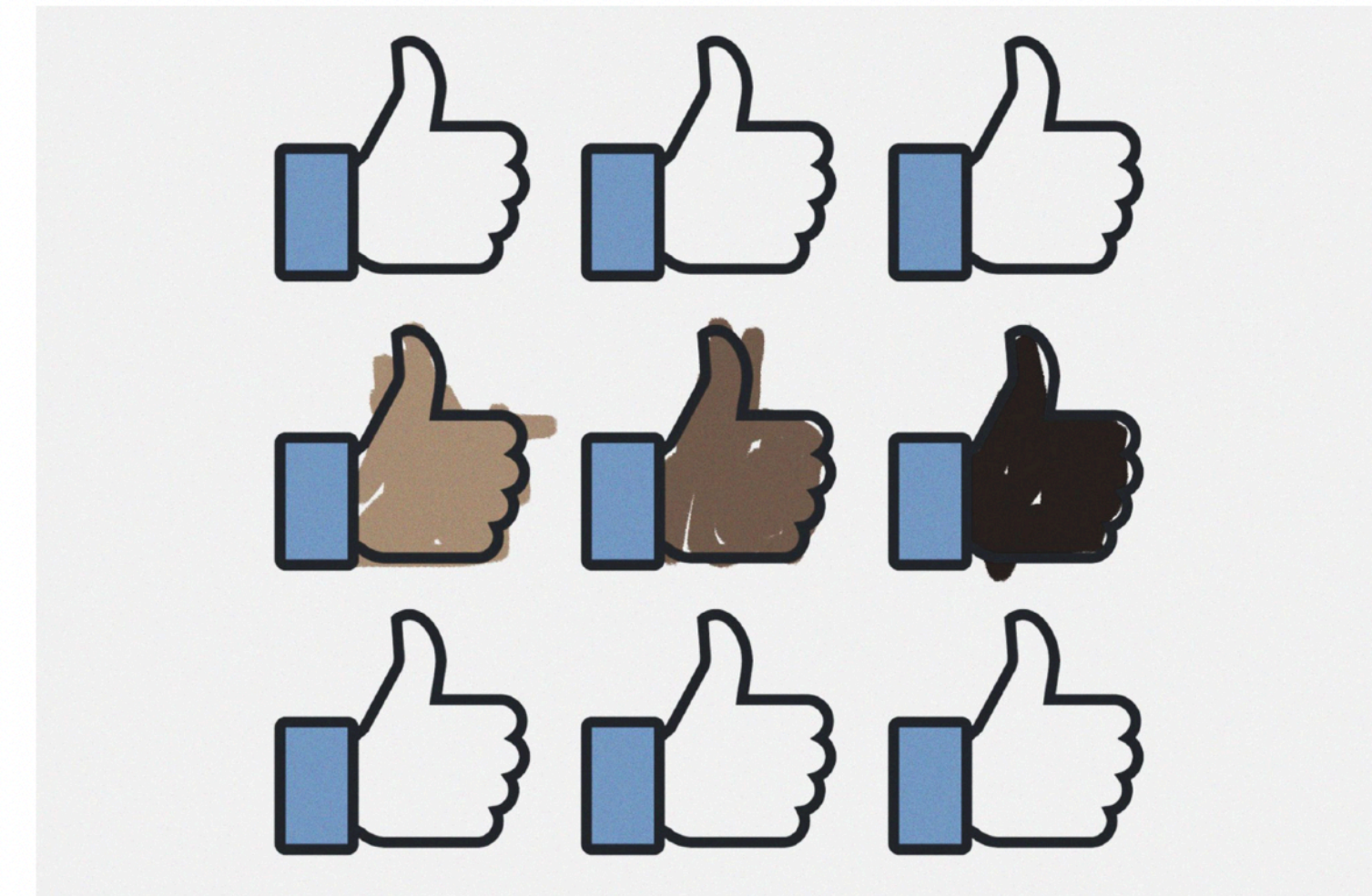


Illustration by ProPublica

by Ariana Tobin and Ava Kofman

June 22, 2022, 4:30 p.m. EDT

MACHINE BIAS



Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads

Among the companies we found doing it: Amazon, Verizon, UPS and Facebook itself. "It's blatantly unlawful," said one employment law expert.

by Julia Angwin, ProPublica, Noam Scheiber, The New York Times, and Ariana Tobin, ProPublica, Dec. 20, 2017, 5:45 p.m. EST



Mark Edelstein, a social media marketing strategist who is also legally blind, says he never had serious trouble finding a job until he turned 50. (Whitney Curtis for The New York Times)

TECH | AMAZON | ARTIFICIAL INTELLIGENCE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

21

The secret program penalized applications that contained the word “women’s”

By James Vincent | Oct 10, 2018, 7:09am EDT

f t SHARE



Illustration by Alex Castro / The Verge



Listen to this article



Bias in machine learning can be a problem even for companies with plenty of experience with AI, like Amazon. According to a [report from Reuters](#), the e-commerce giant scrapped an internal project that was trying to use AI to vet job applications after the software consistently downgraded female candidates.

Because AI systems learn to make decisions by looking at historical data they often perpetuate existing biases. In this case, that bias was the male-dominated working environment of the tech world. According to *Reuters*, Amazon's program penalized applicants who attended all-women's colleges, as well as any resumes that contained the word "women's" (as might appear in the phrase "women's chess club").



IMAGE: GETTY IMAGES / COMPOSITION: JASON KOEBLER

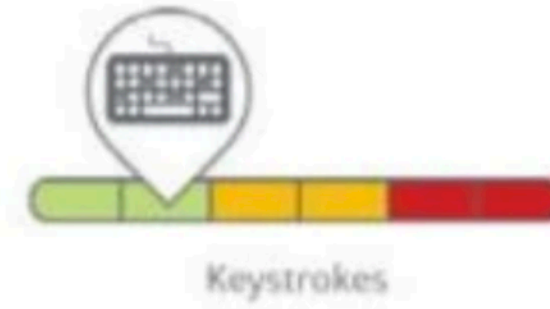
MOTHERBOARD
TECHBYVICE

Students Are Rebelling Against Eye-Tracking Exam Surveillance Tools

Invasive test-taking software has become mandatory in many places, and some companies are retaliating against those who speak out.

TF By [Todd Feathers](#)

By [Janus Rose](#)
NEW YORK, US



Suspicion Levels

Proctorio is intended to be used to uphold academic integrity by discouraging cheating. However, some incidents will still occur. When these do happen, Proctorio makes the identification of these cases as simple as possible. One way is through the suspicion level. The suspicion level is a percentage that represents low, medium, or high suspicion for an individual's exam attempt.

The suspicion level is a quick calculation based on the aggregation of frames (captured activity) during the exam that were deemed suspicious and the detection of abnormal (deviation from the class norm) behaviour. If the suspicion level shows a large percentage, then this is an exam attempt that should be considered for further review.

The suspicion level will increase or decrease depending on how heavily each **Behaviour Setting** is weighted and which abnormalities are enabled.

A SLIDE FROM PROCTORIO'S TRAINING MATERIALS, DETAILING HOW THE SYSTEM MEASURES "SUSPICION LEVELS" WHILE STUDENTS TAKE EXAMS.

Students' and educators' objections to exam proctoring software go beyond the privacy concerns around being watched and listened to in their bedrooms while they take a test. As more evidence emerges about how the programs work, and fail to work, critics say the tools are bound to hurt low-income students, students with disabilities, students with children or other



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.

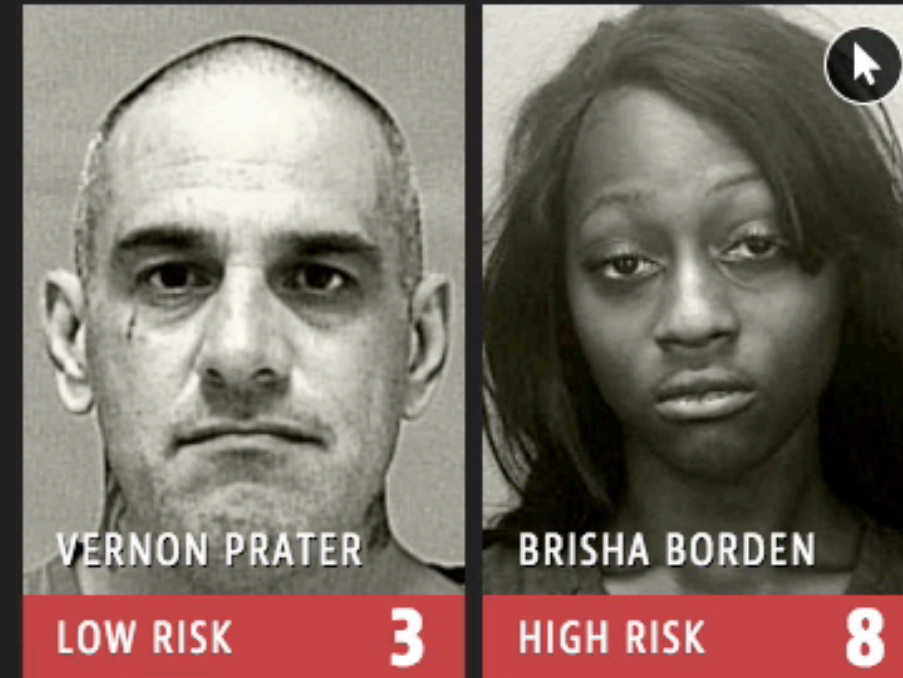
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

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.



- keep your eyes open to ways that data and algorithms are being used to perpetuate inequity
- talk to your representatives about legislation for algorithmic transparency and fairness
- look for the helpers

Dr. Latanya Sweeney



Carnegie Mellon

DATA PRIVACY LAB

SSNs Social Security Numbers

Matching a Person to

Using publicly available information about SSNs, a
Server identifies the issuing state, date issued, etc.

Check Validation

Verify status of an SSN.

Sample uses:

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

Enter the
"quick"
Enter the
whether

(SSN) and select
of the SSN
additionally learn

Results for SSN 078 - 08 -

Geography	New York
Date of issuance	Second Half of 1980
Year of Birth (8-digit prefix)	1975, from 1980 to 1990 1975, 1976, 1977, 1978

If the
allow
unlikely
issued to

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



Tawana Petty

oderator: *Patie Hear*



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

Dr. Joy Buolamwini





DONATE

RACIAL JUSTICE REQUIRES ALGORITHMIC JUSTICE. SUPPORT THE MOVEMENT.



Dr. Stephanie Russo Carroll





GIDA

Global Indigenous
Data Alliance





Journalists can be
data science
superheroes

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

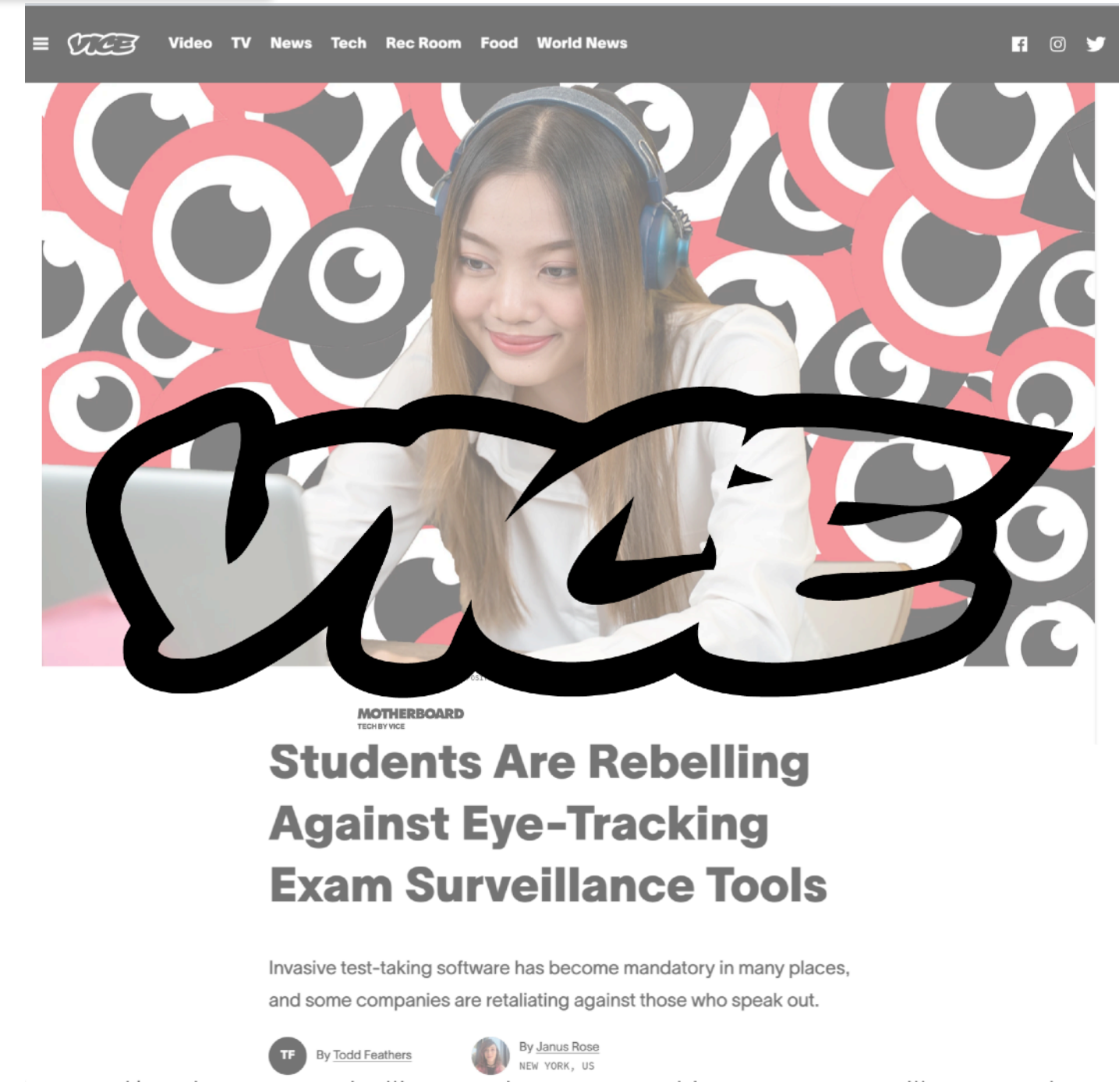
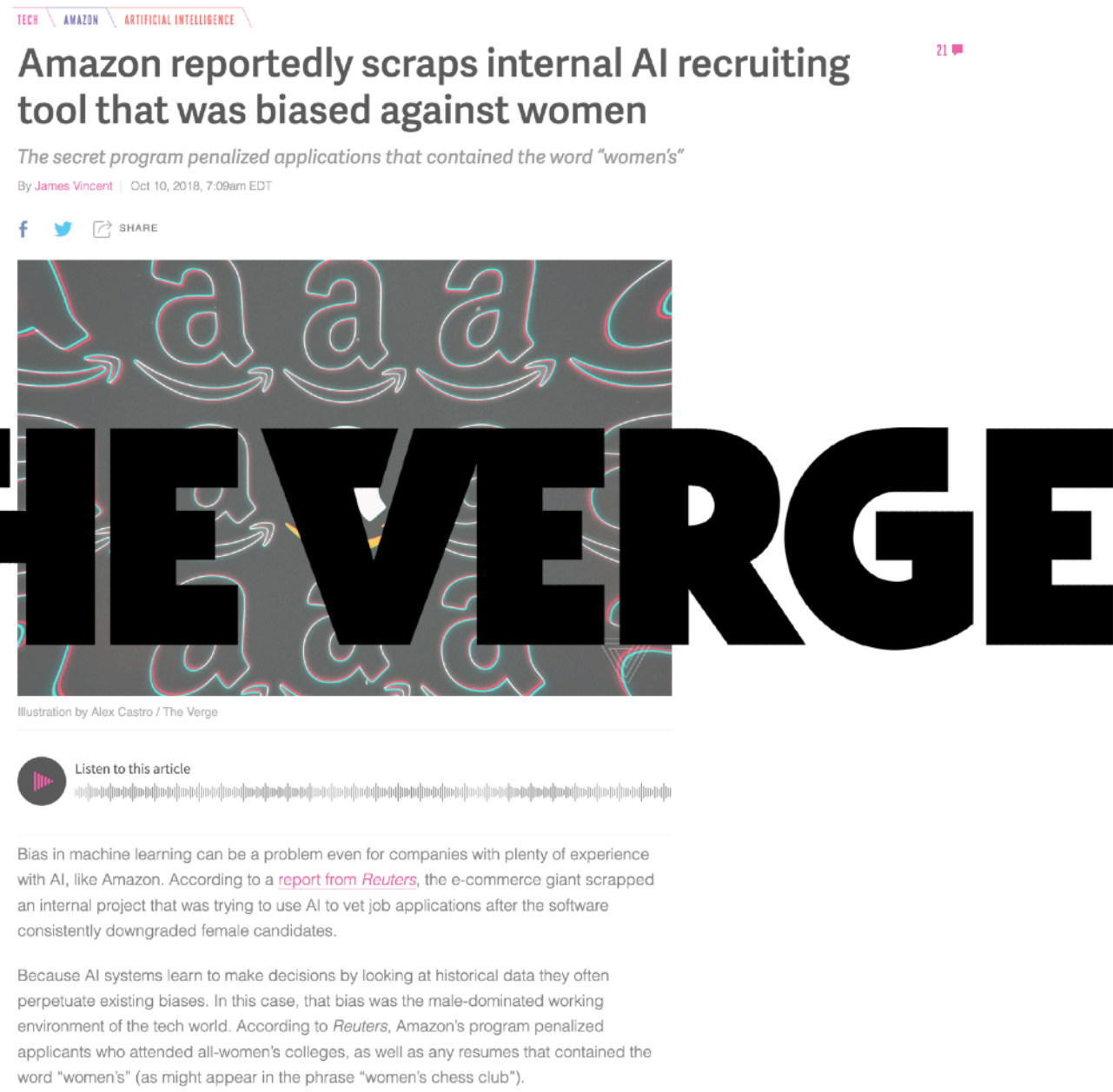


Onning
ed an
orden
m
.
ces —
"That's
and

unl
and
dov

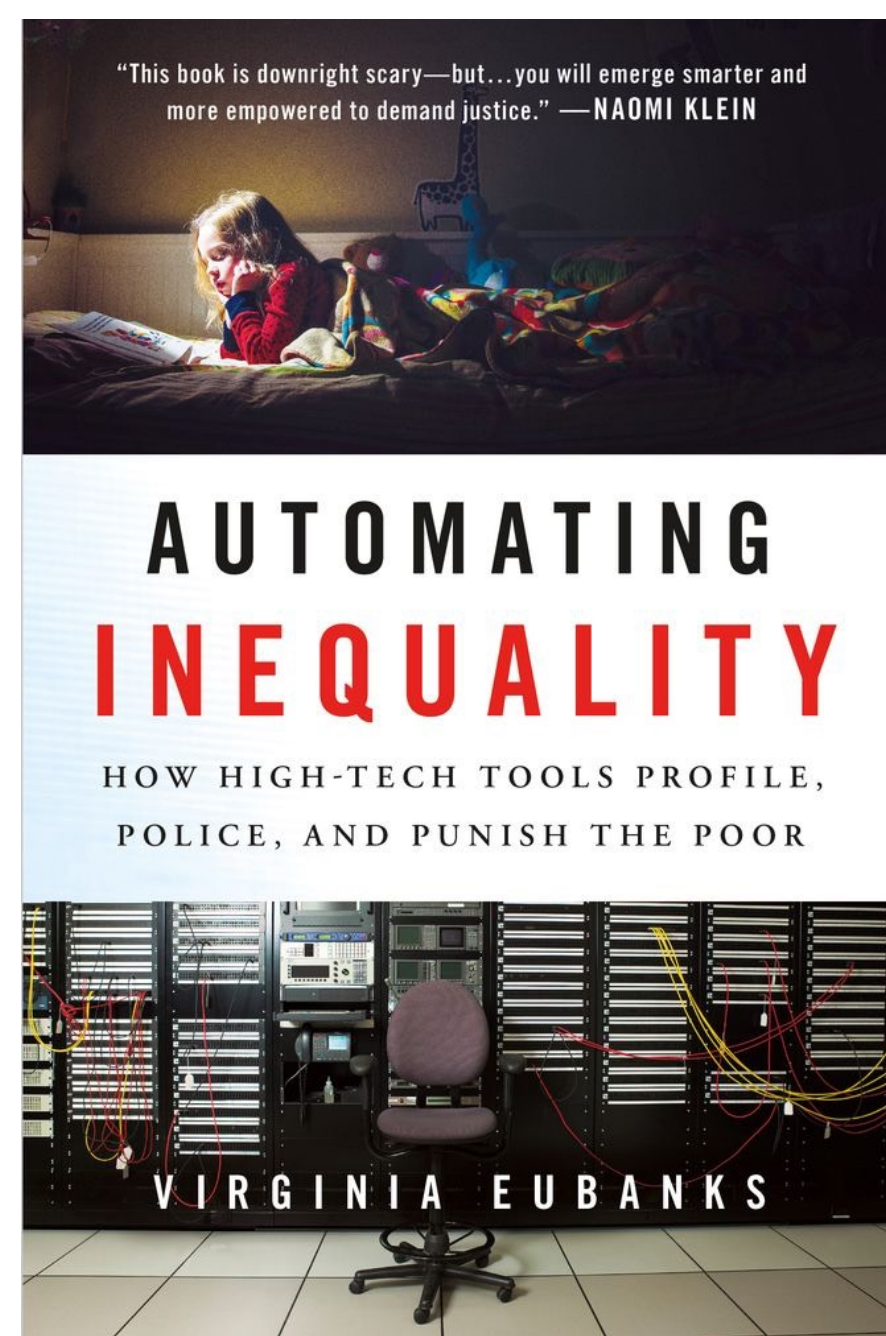
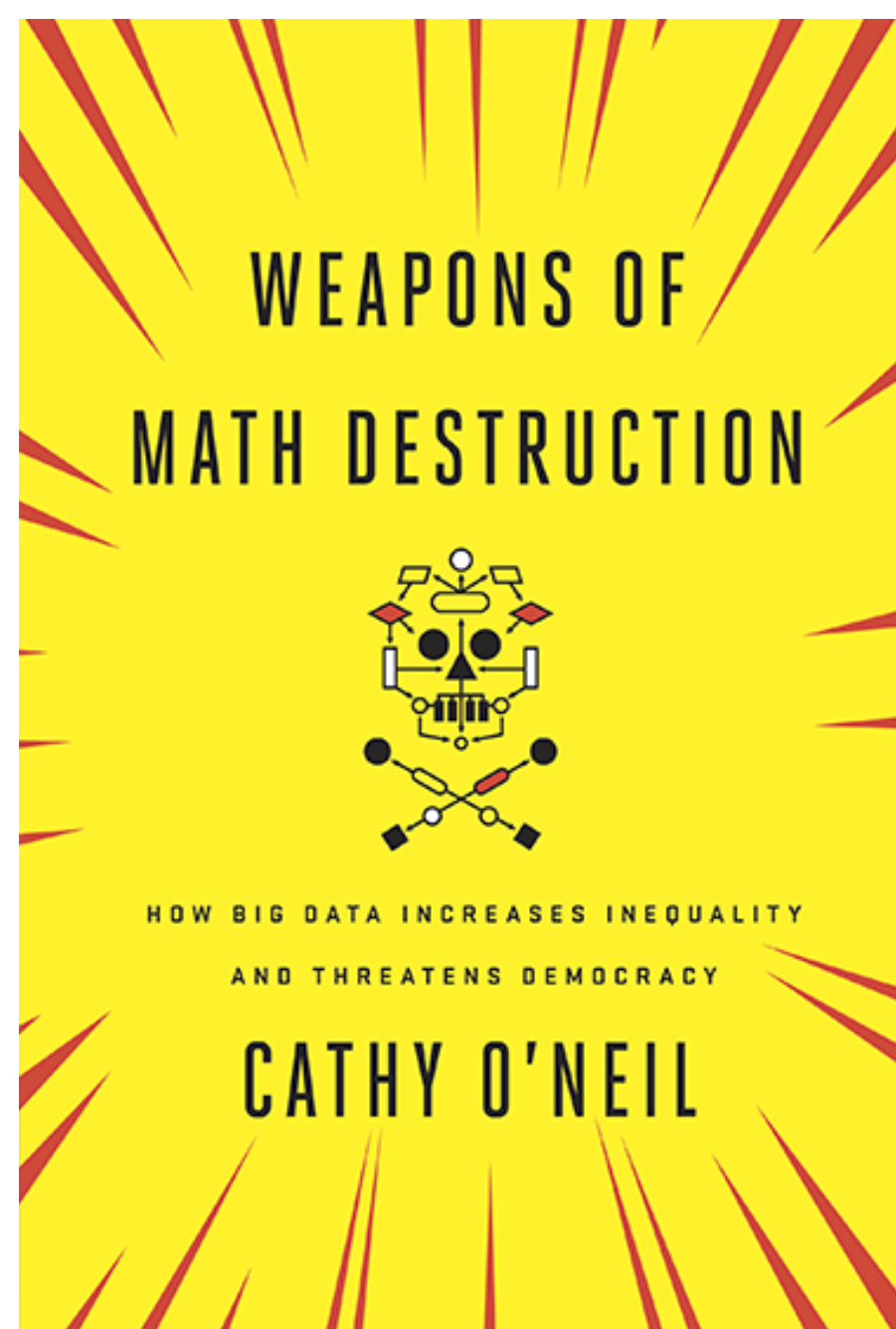
Just
whic
my k
walked away.

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.

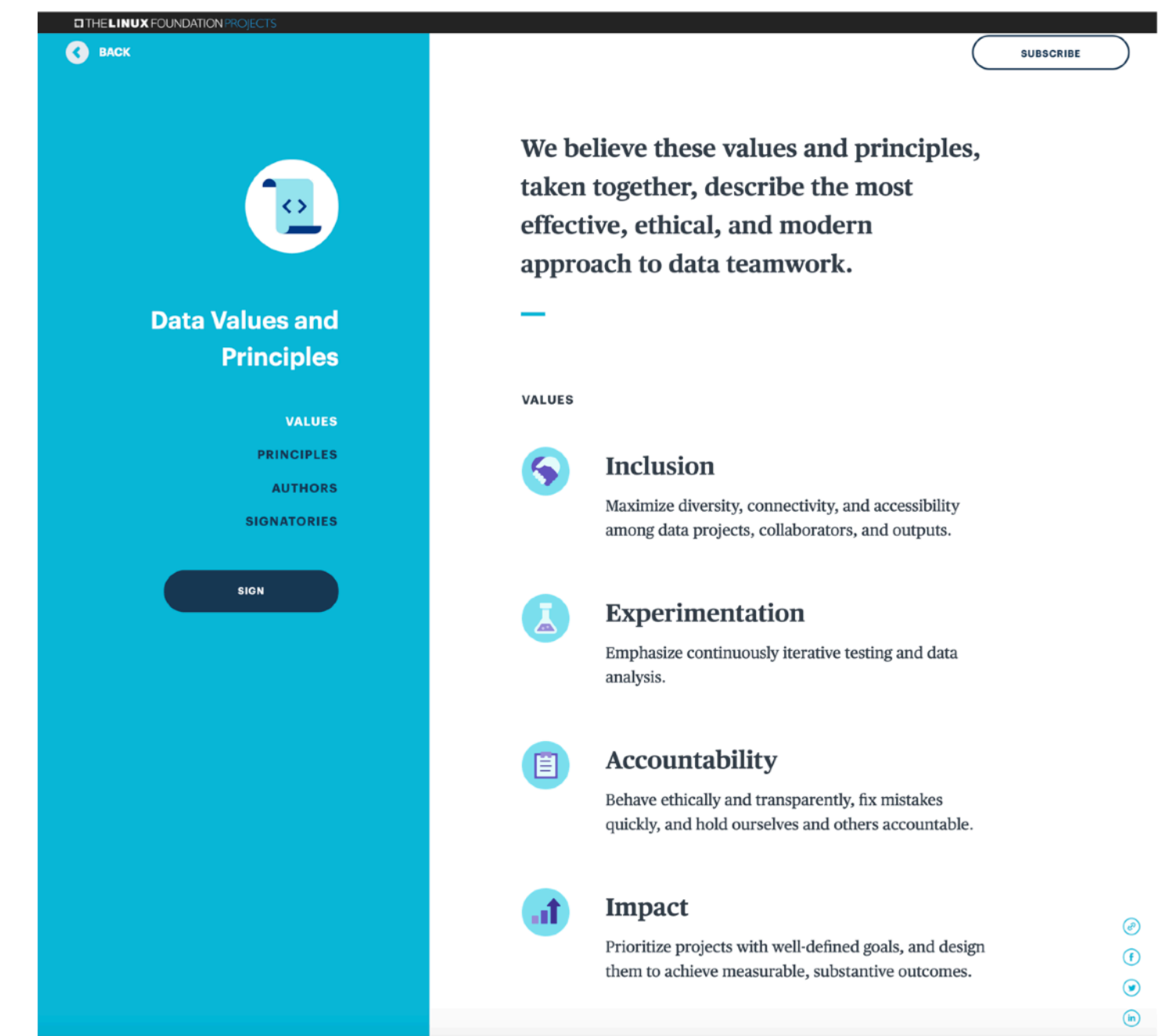
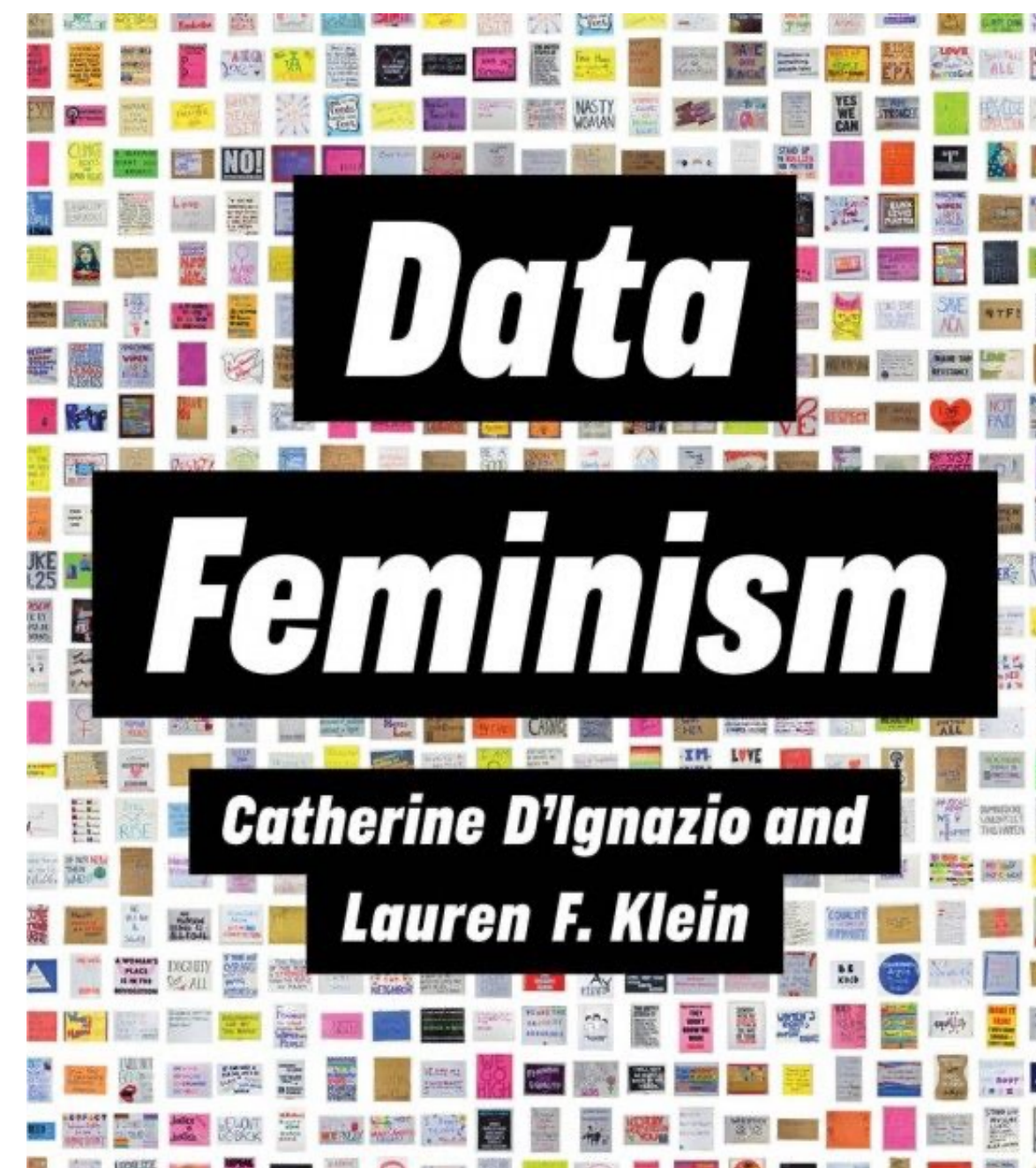
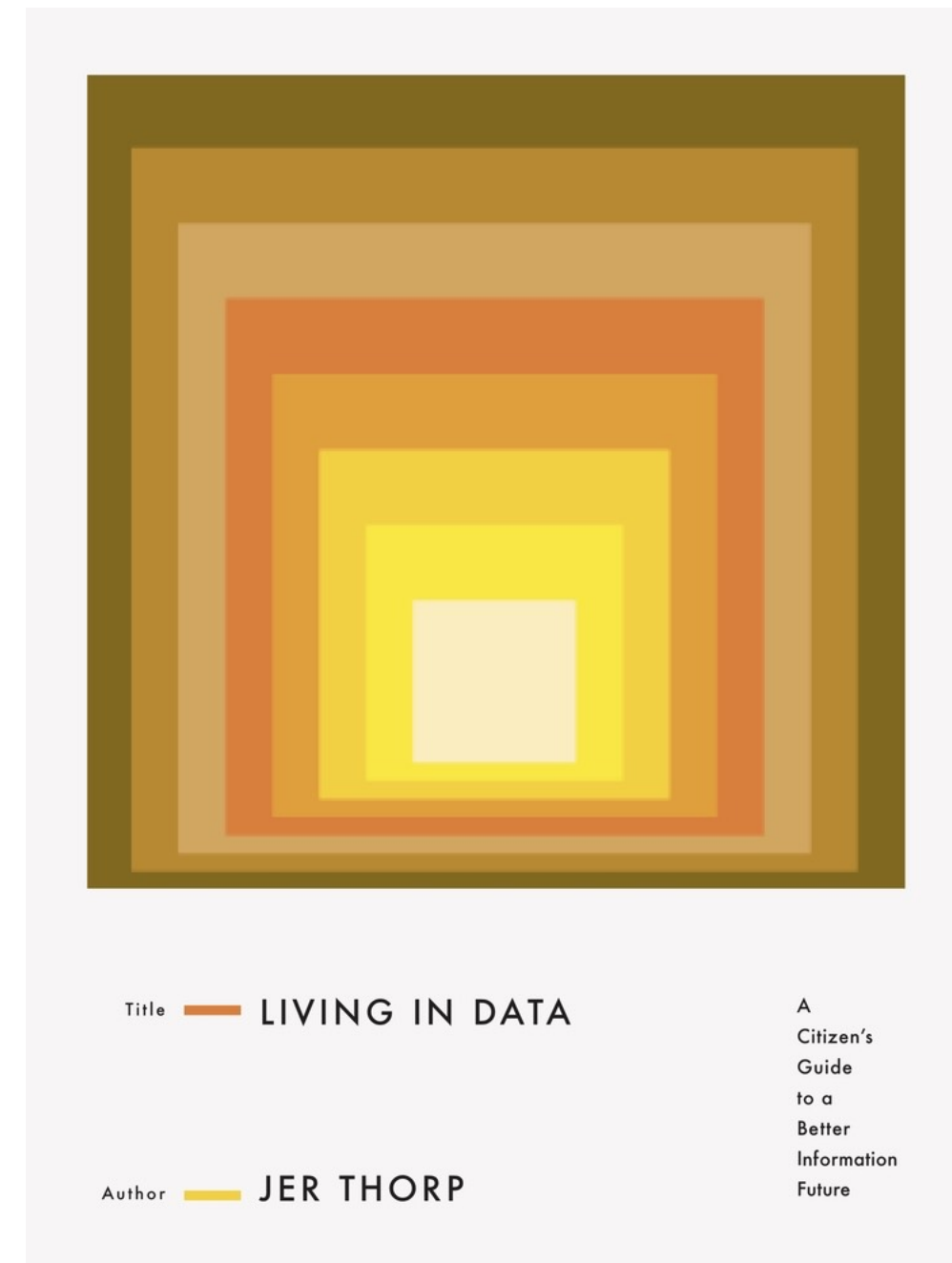
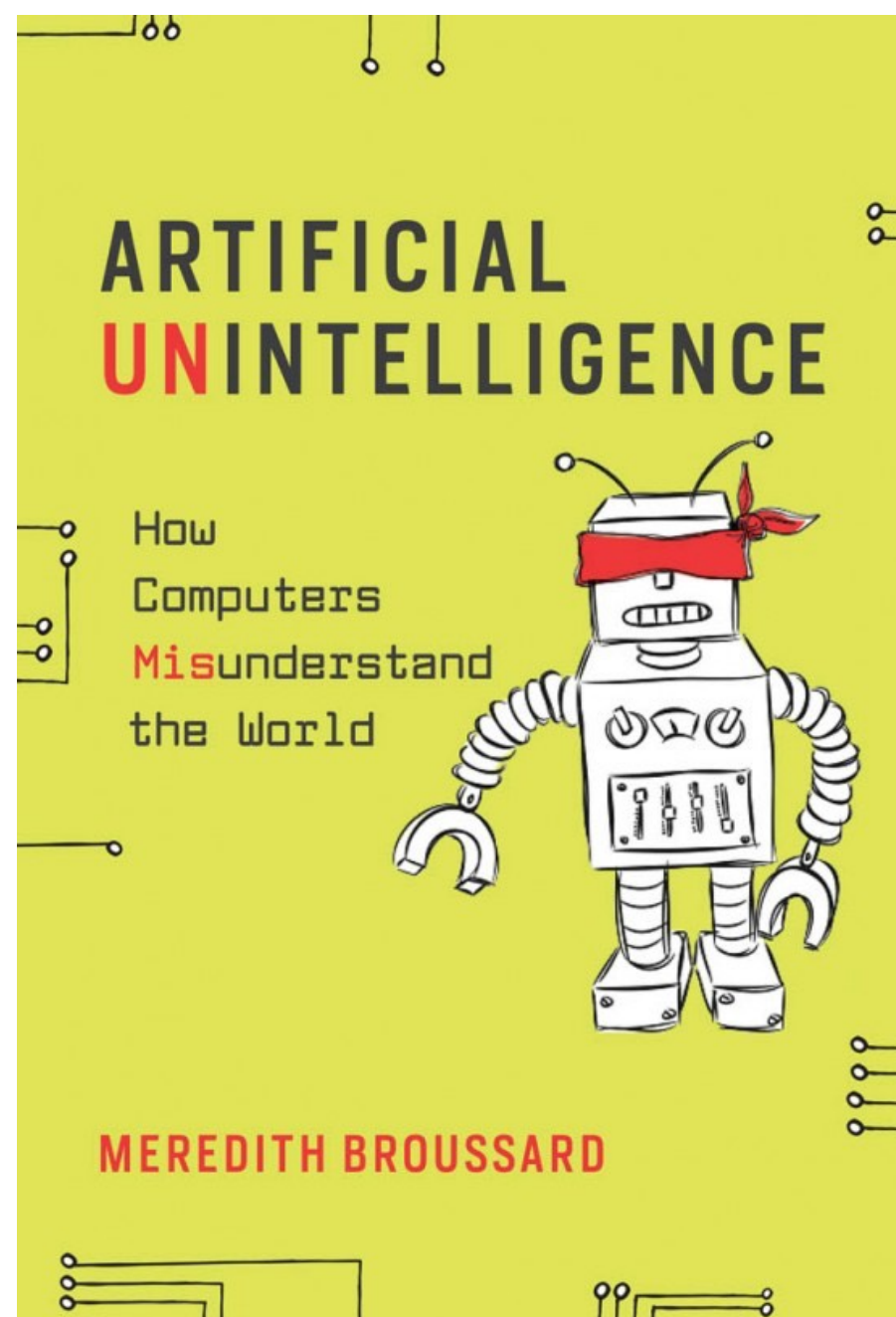
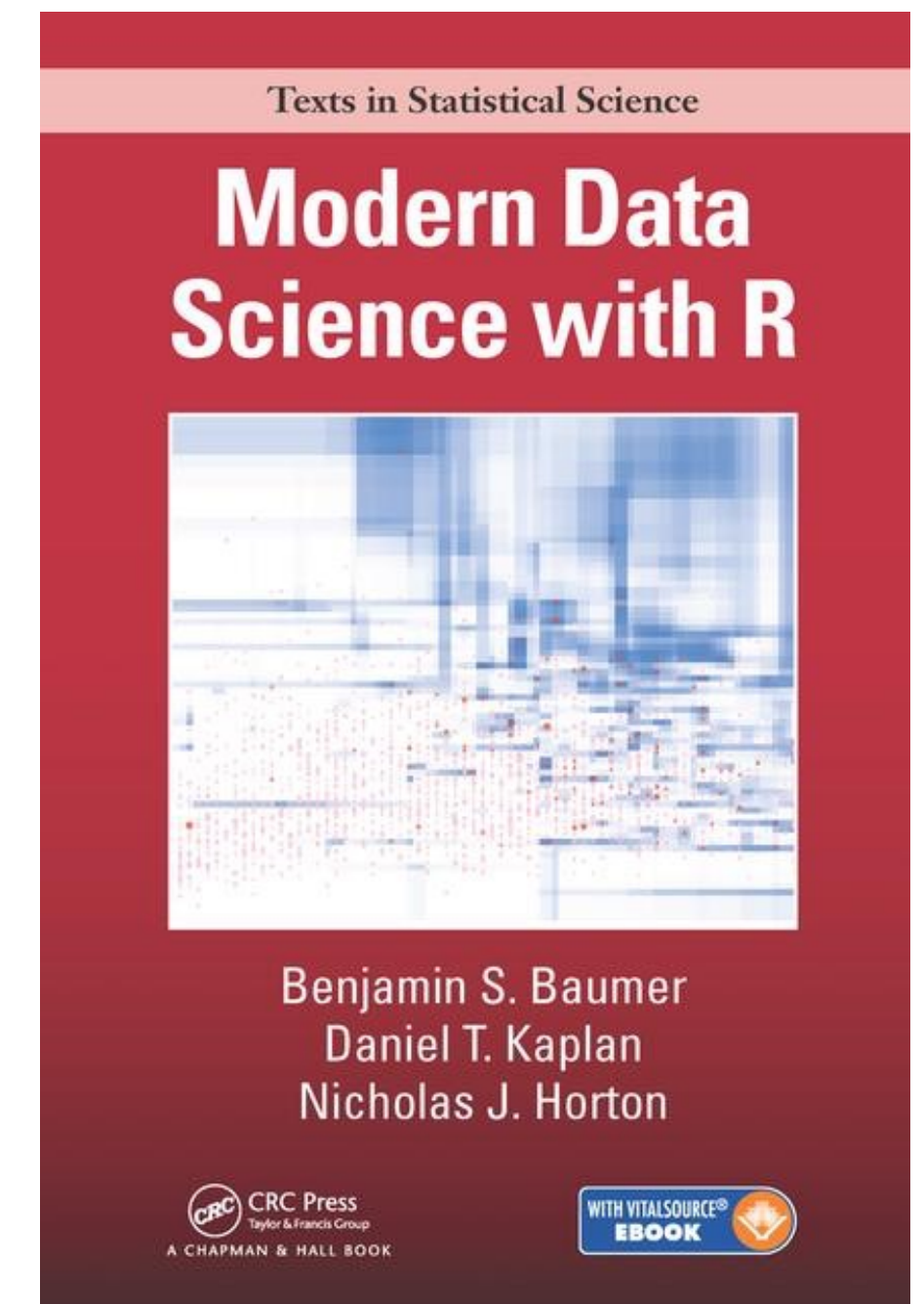


Students and
Teachers can be
data science
superheroes





Read



Watch

Bias Remediation



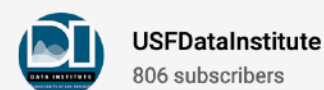
- 🔍 📊 Evaluate/diagnose the bias in a set of models (based on metrics like AIR, SMD, etc.)
- If there are no acceptable models, then: 🔧
 - Fix data: row weights or up/down sampling
 - Fix model: regularize impact of demographic features; adversarial debiasing.
 - Fix predictions: re-weight preds to improve fairness

19:40 / 1:30:12

Erin LeDell, Chief Learning Machine Scientist at H2O.ai: Admissible Machine Learning

65 views • Feb 22, 2022

👍 4 🗨 DISLIKE ➦ SHARE ⚙ SAVE ...



USFDataInstitute
806 subscribers

SUBSCRIBE

Erin Ledell is Chief Machine Learning Scientist at H2O.ai, the company that produces the open source, distributed machine learning platform, H2O. At H2O.ai, she leads the development of the H2O AutoML algorithm. She is also the founder of WiMLDS and co-founder of R-Ladies Global.

CODED BIAS

AVAILABLE TO STREAM APRIL 5TH ON

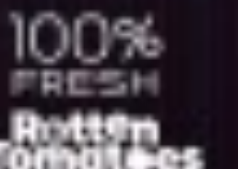
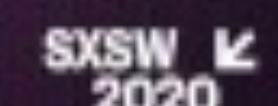
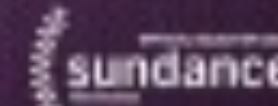
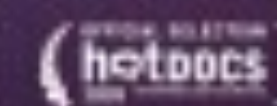
NETFLIX



"THE BEST OF SUNDANCE"
Denver Post

"'CODED BIAS' SERVES AS BOTH A WAKE-UP CALL AND A CALL TO ACTION"
Variety

"THE MOST CLEAREYED OF SEVERAL RECENT DOCUMENTARIES ABOUT THE PERILS OF BIG TECH"
The New York Times



codedbias.com/screen

Listen



A SPOTIFY ORIGINAL

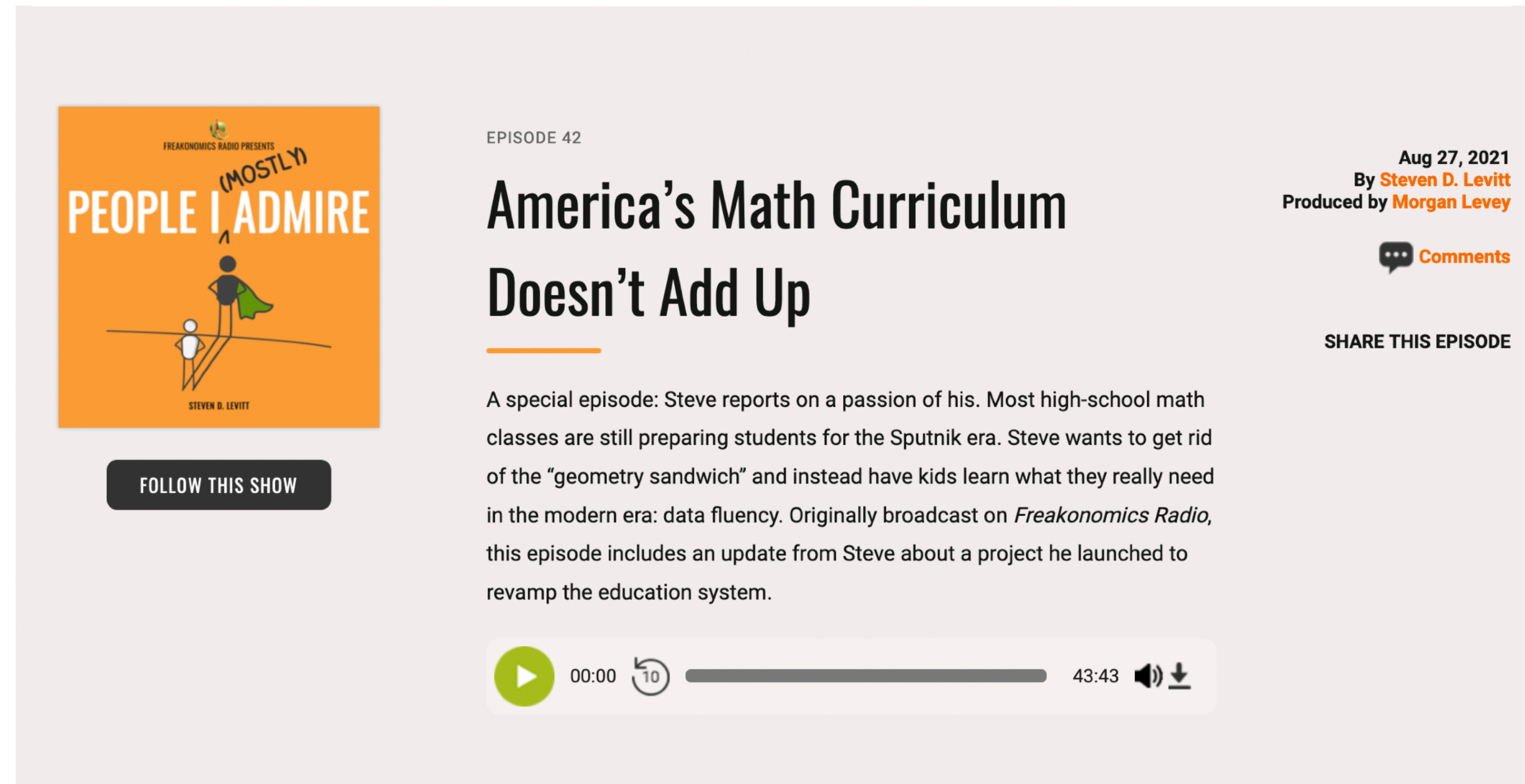
OCTOBER 12, 2018

#127 The Crime Machine, Part I

by REPLY ALL

LISTEN NOW

LISTEN ON 



FREAKONOMICS RADIO PRESENTS

PEOPLE I ADMIRE (MOSTLY)

STEVEN D. LEVITT

FOLLOW THIS SHOW

EPISODE 42

America's Math Curriculum Doesn't Add Up

A special episode: Steve reports on a passion of his. Most high-school math classes are still preparing students for the Sputnik era. Steve wants to get rid of the "geometry sandwich" and instead have kids learn what they really need in the modern era: data fluency. Originally broadcast on *Freakonomics Radio*, this episode includes an update from Steve about a project he launched to revamp the education system.

00:00 10 43:43

Aug 27, 2021
By **Steven D. Levitt**
Produced by **Morgan Levey**

[Comments](#)

SHARE THIS EPISODE



NOT SO STANDARD DEVIATIONS

with
Roger Peng
and
Hilary Parker



- keep your eyes open to ways that data and algorithms are being used to perpetuate inequity
- talk to your representatives about legislation for algorithmic transparency and fairness
- look for the helpers



Thank you