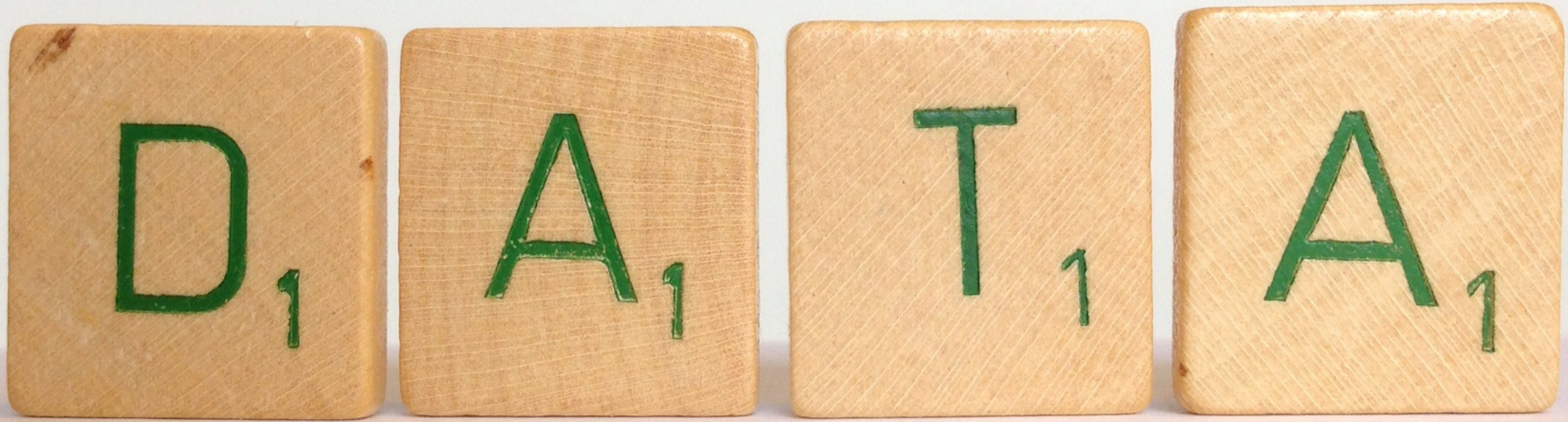
Data Science as a Superpower

Amelia McNamara @AmeliaMN
University of St Thomas St Paul, MN
Department of Computer & Information Sciences

flickr: Hannaford



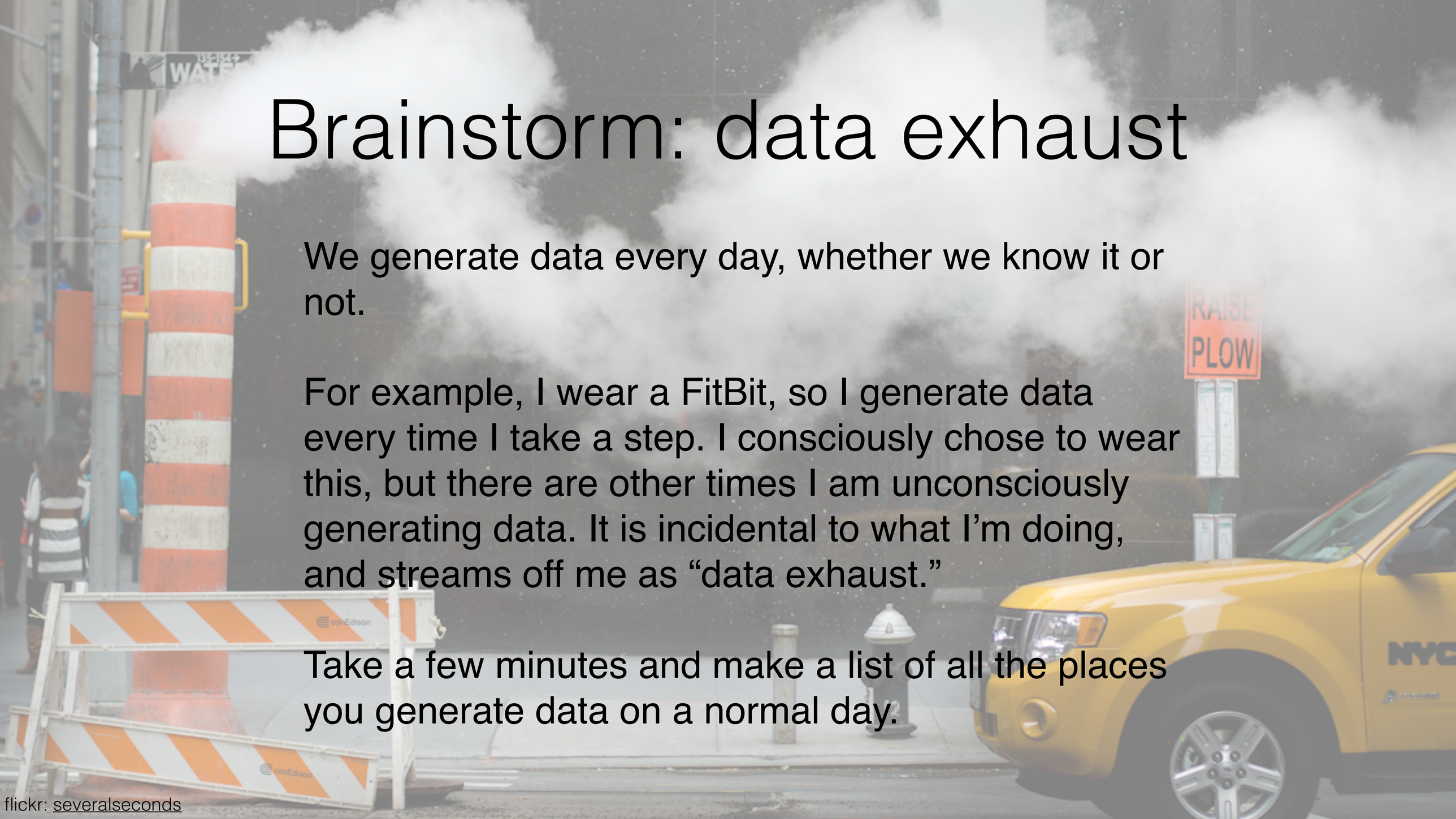
flickr: jannekestaaks

Brainstorm: data exhaust

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

For example, I wear a FitBit, 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.

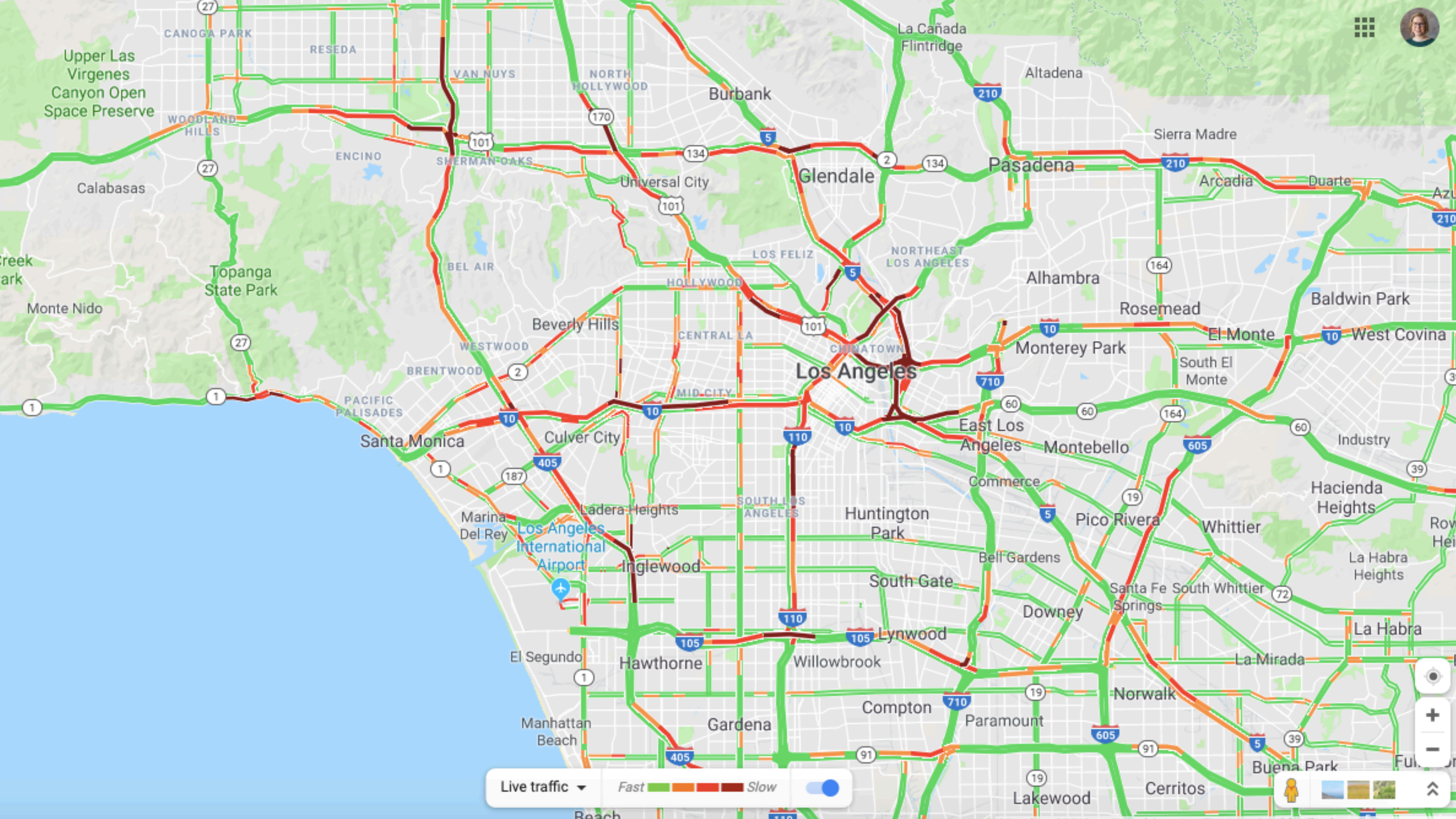


flickr: severalseconds

flickr: jannekestaaks

flickr: pchow98

flickr: jcolman


Upper Las Virgenes Canyon Open Space Preserve
CANOGA PARK
RESEDA
VAN NUYS
NORTH HOLLYWOOD
Burbank
La Cañada Flintridge
Altadena
WOODLAND HILLS
ENCINO
SHERMAN OAKS
Universal City
Glendale
Pasadena
Sierra Madre
Arcadia
Duarte
Calabasas
Topanga State Park
Monte Nido
BEL AIR
LOS FELIZ
NORTHEAST LOS ANGELES
HOLLYWOOD
Alhambra
Rosemead
Baldwin Park
Beverly Hills
CENTRAL LA
CHINATOWN
Monterey Park
El Monte
West Covina
WESTWOOD
BRENTWOOD
Los Angeles
South El Monte
PACIFIC PALISADES
MID-CITY
Santa Monica
Culver City
East Los Angeles
Montebello
Industry
Commerce
Hacienda Heights
Ladera Heights
SOUTH LOS ANGELES
Marina Del Rey
Huntington Park
Pico Rivera
Whittier
Los Angeles International Airport
Inglewood
Bell Gardens
La Habra Heights
South Gate
Santa Fe Springs
South Whittier
Downey
La Habra
Lynwood
El Segundo
Hawthorne
Willowbrook
La Mirada
Norwalk
Compton
Paramount
Manhattan Beach
Gardena
Buena Park
Lakewood
Cerritos
Live traffic
Fast
Slow

Select all images with a

bus

Click verify once there are none left.

VERIFY

flickr: sergiorojas

DataKind

Forecasting Water Demand in California When Every Drop Counts

December 2016

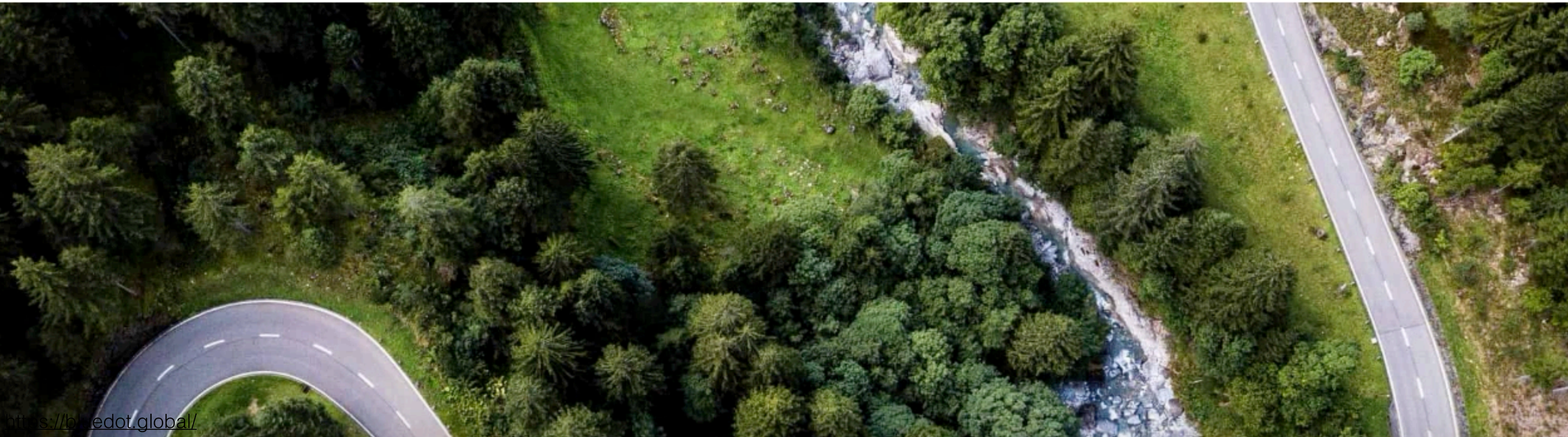
Objectives

- Develop a “proof of concept” water demand forecasting model using flow data at the microzone level for potential future scaling to other retailers in the California Data Collaborative, a unique water manager-led public private partnership that brings together utilities across the state to leverage data to help water managers ensure reliability.

http://www.datakind.org/projects/forecasting-water-demand-in-california-when-every-drop-counts

About Us

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



https://bluedot.global/

OCEAN HEALTH INDEX

GLOBAL SCORES OHI+ METHODOLOGY NEWS ABOUT

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.

http://www.oceanhealthindex.org/

GLOBAL ASSESSMENT

INDEPENDENT ASSESSMENTS

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

R for better science in less time. Julia Stewart Lowndes, useR 2019 (slides and video).

Algorithms Tour

How data science is woven into the fabric of Stitch Fix

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

...

$$\min_{a} \sum_{i} \sum_{j} a_{ij} q_{ij}$$

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

$$\sum_{j} a_{ij} = 1 \; \forall i$$

$$\sum_{i} a_{ij} < k_j \; \forall j$$

...

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

...

$$p(i \to j) = logit(\beta_0 + \beta_1 x_1 \dots)$$

http://algorithms-tour.stitchfix.com/

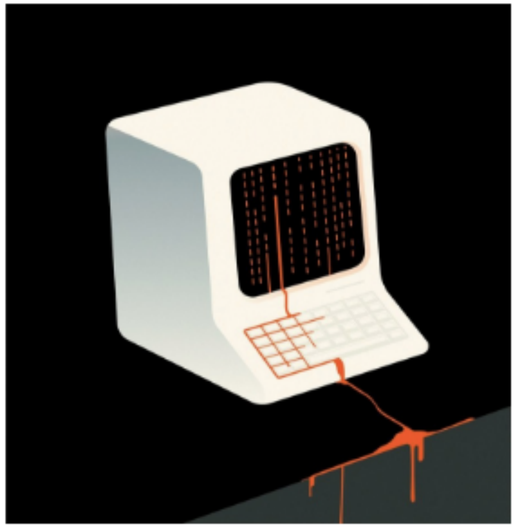
ANNALS OF CRIME NOVEMBER 27, 2017 ISSUE

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

https://www.newyorker.com/magazine/2017/11/27/the-serial-killer-detector

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

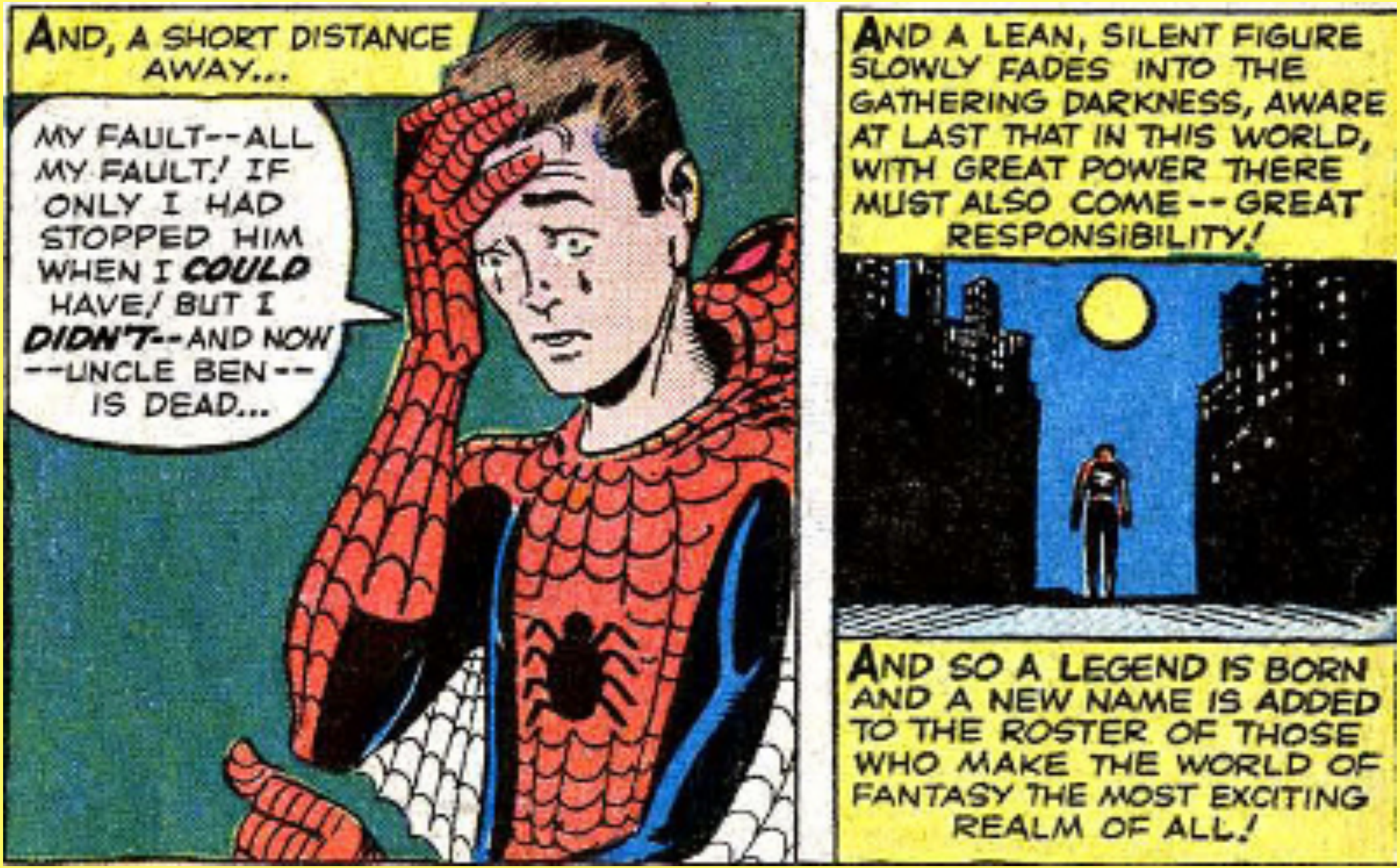
https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

ADAM ROGERS BUSINESS 04.02.18 07:00 AM

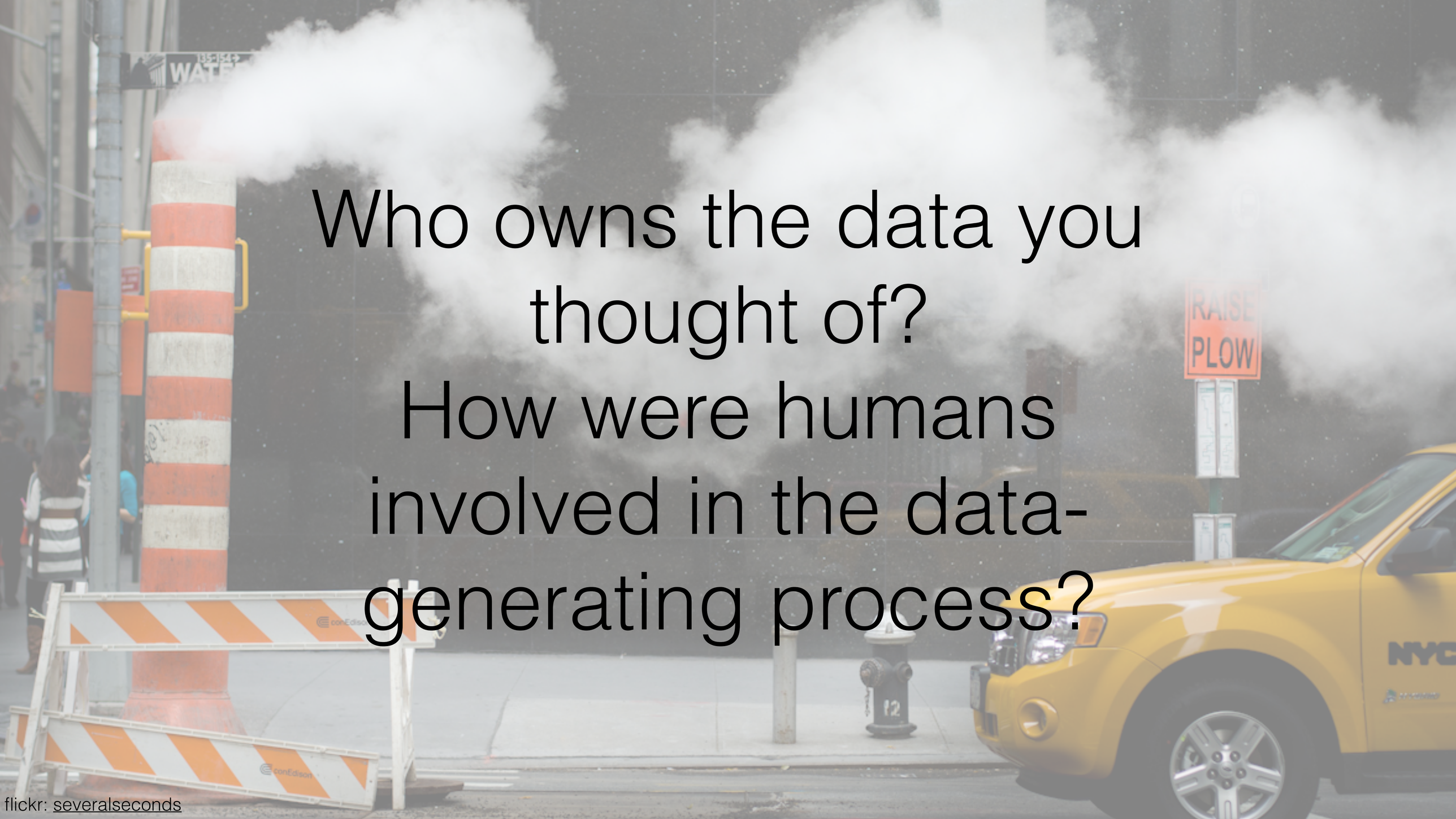
HOW GRUBHUB ANALYZED 4,000 DISHES TO PREDICT YOUR NEXT ORDER

JONATHAN KITCHEN/GETTY IMAGES

https://www.wired.com/story/how-grubhub-analyzed-4000-dishes-to-predict-your-next-order/



flickr: philwolff



Who owns the data you thought of?
How were humans involved in the data-generating process?

flickr: severalseconds

flickr: mermaid99

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.

https://www.forbes.com/sites/tomgroenfeldt/2013/10/28/kroger-knows-your-shopping-patterns-better-than-you-do/

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)

https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

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

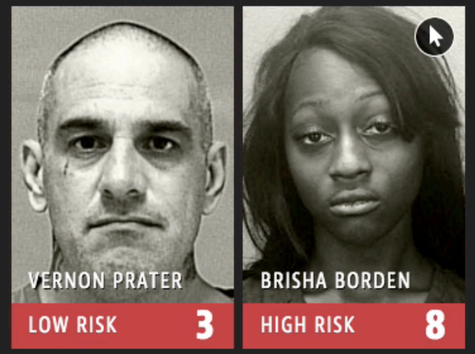
May 23, 2016

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

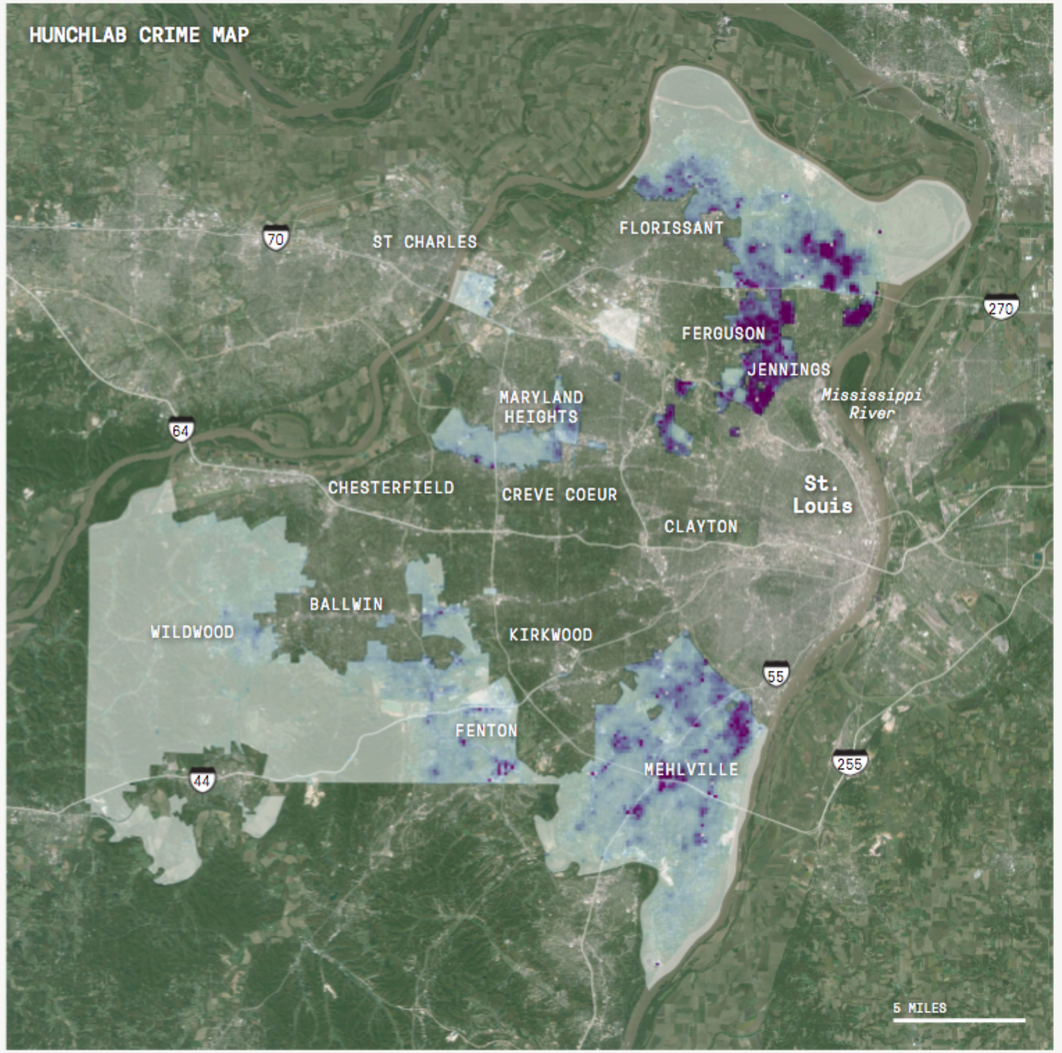
Policing the Future

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.
hallproject.org/2016/02/03/policing-the-future

Where the St. Louis County Police Patrol

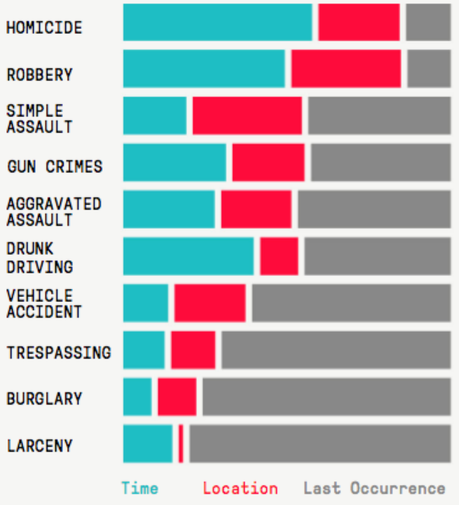
Dozens of small, local municipal agencies handle policing in parts of St. Louis County. The St. Louis County Police Department covers areas not policed by the "munis," including the city of Jennings, Mo. The ▪ DARKER AREAS in the map show the areas within their jurisdiction that HunchLab has identified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

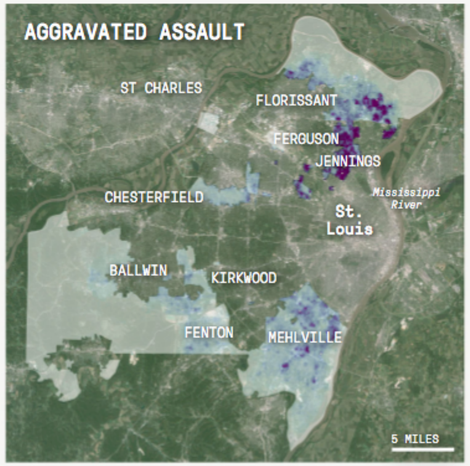


SOURCE: HUNCHLAB

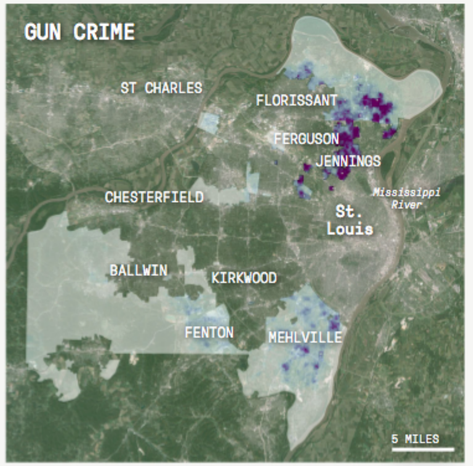
Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.
https://www.themarshallproject.org/2016/02/03/policing-the-future

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.

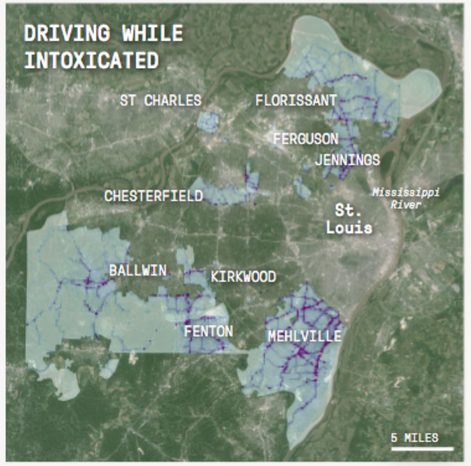




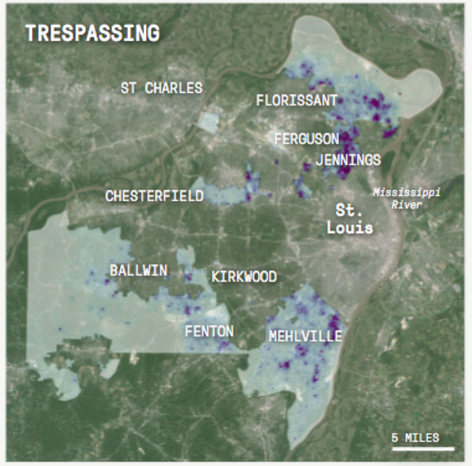
Aggravated assault (assault with a dangerous weapon) makes up 18.5 percent of the overall risk score assigned to a cell. The darkest regions on this map represent cells with a 1 in 320 chance of at least one aggravated assault taking place there during the shift.



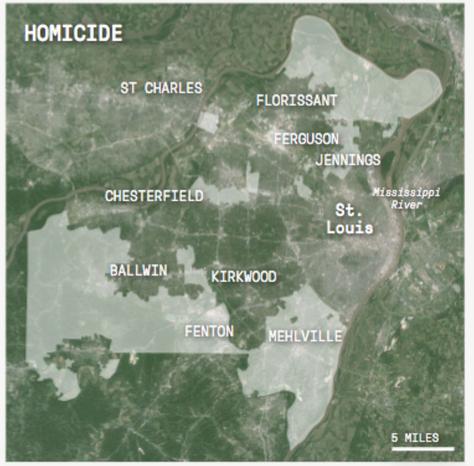
Gun crime (all homicides, robberies, and aggravated assaults with a firearm) makes up about 16.5 percent of the overall risk score. The darkest regions represent a 1 in 850 chance of at least one gun crime taking place.



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

DATA PRIVACY LAB

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 April 17 Year...

Gender Male Female

5-digit ZIP 01060

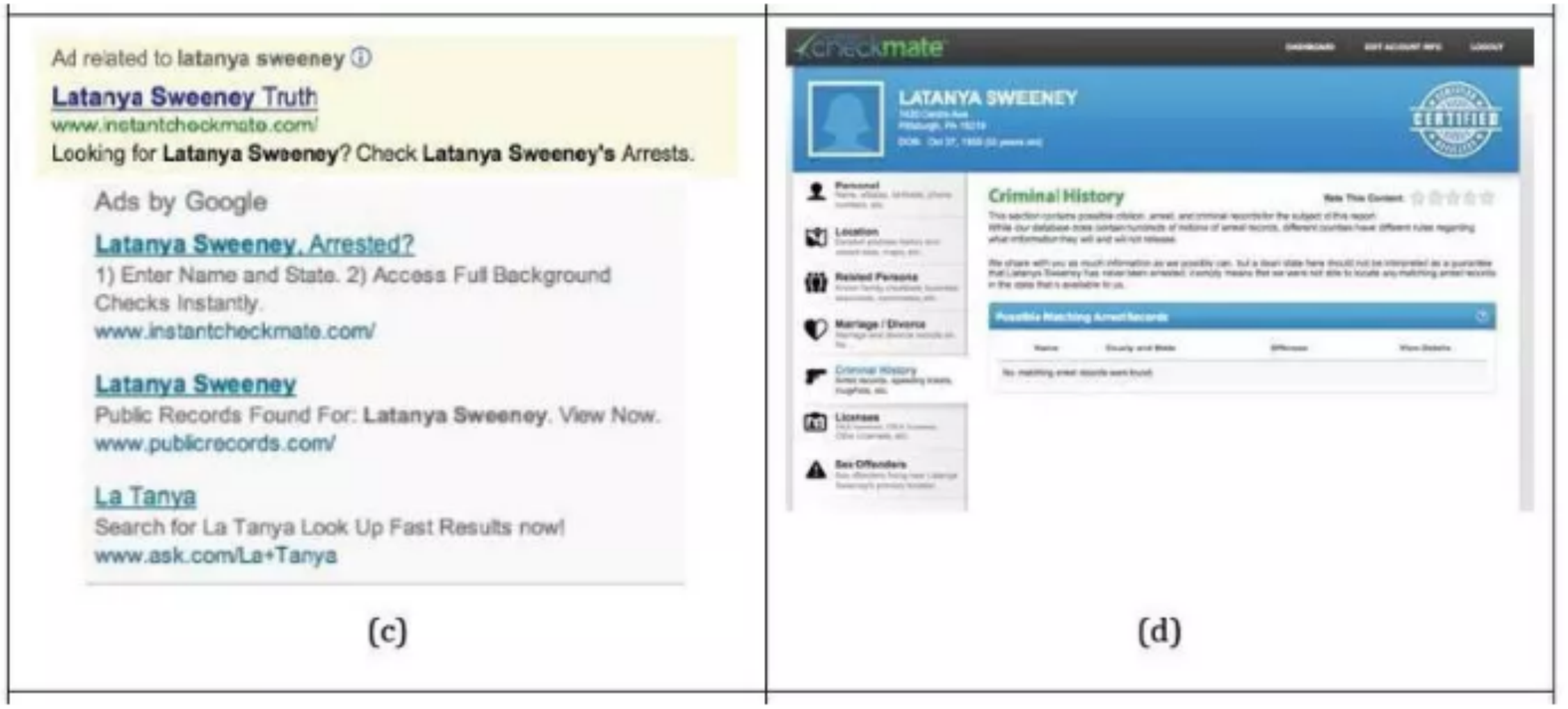
Submit

About | Samples | Harvard | Harvard Multi Years

https://aboutmyinfo.org/index.html

Racial bias alleged in Google's ad results

Names associated with blacks prompt link to arrest search



LATANYA SWEENEY

Web page results of ads that appeared on-screen when Harvard professor Latanya Sweeney typed her name in a google search. Ads featured services for arrest records. Sweeney conducted a study that concluded searches with "black sounding" names are more likely to get results with ads for arrests records and other negative information.

By Hiawatha Bray | GLOBE STAFF FEBRUARY 06, 2013

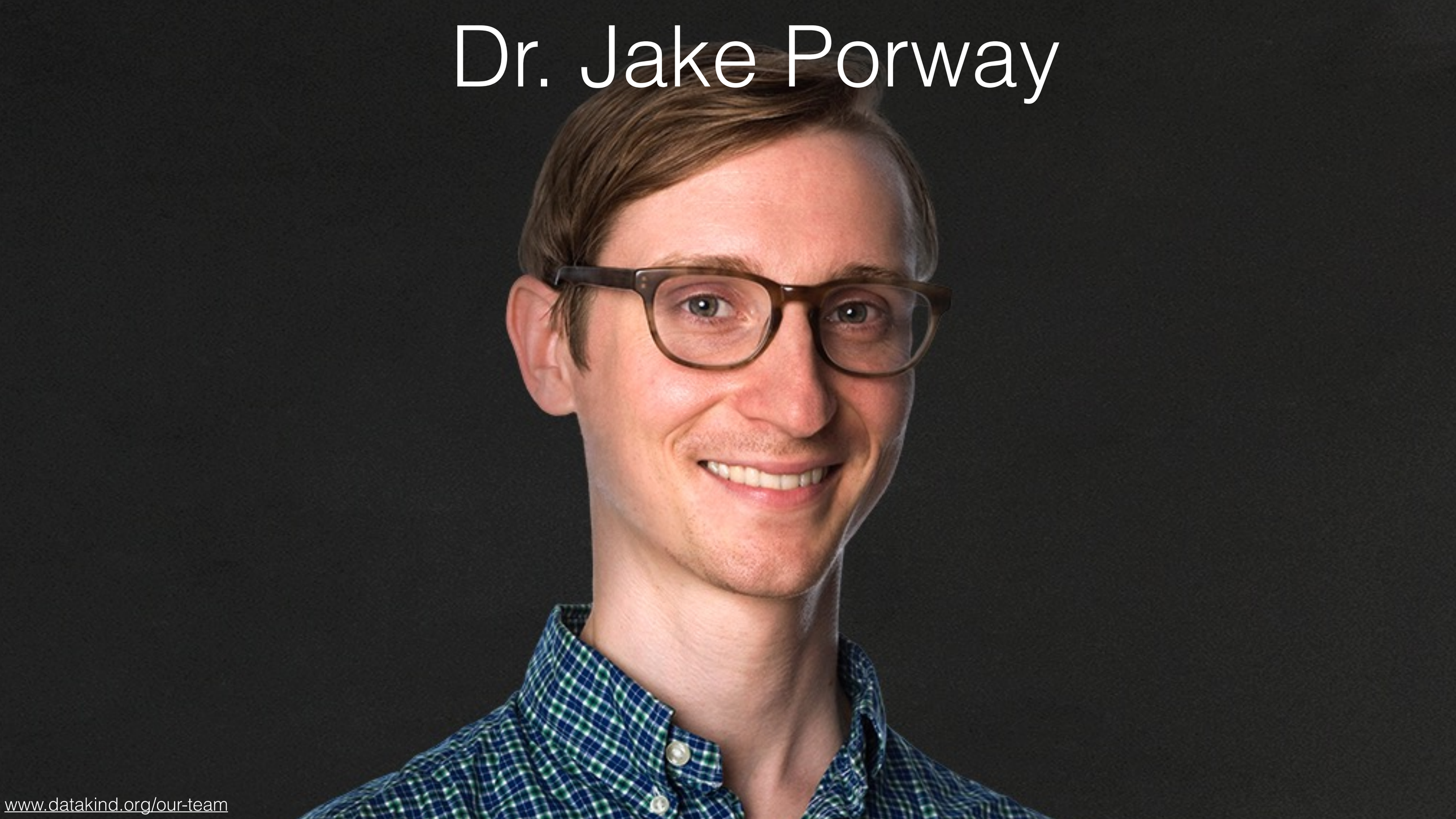
https://www.bostonglobe.com/business/2013/02/06/harvard-professor-spots-web-search-bias/PtOgSh1ivTZMfyEGj00X4I/story.html

flickr: Hannaford

Dr. Latanya Sweeney

https://dataprivacylab.org

Dr. Jake Porway



www.datakind.org/our-team

Dr. Julia Stewart Lowndes

https://jules32.github.io/

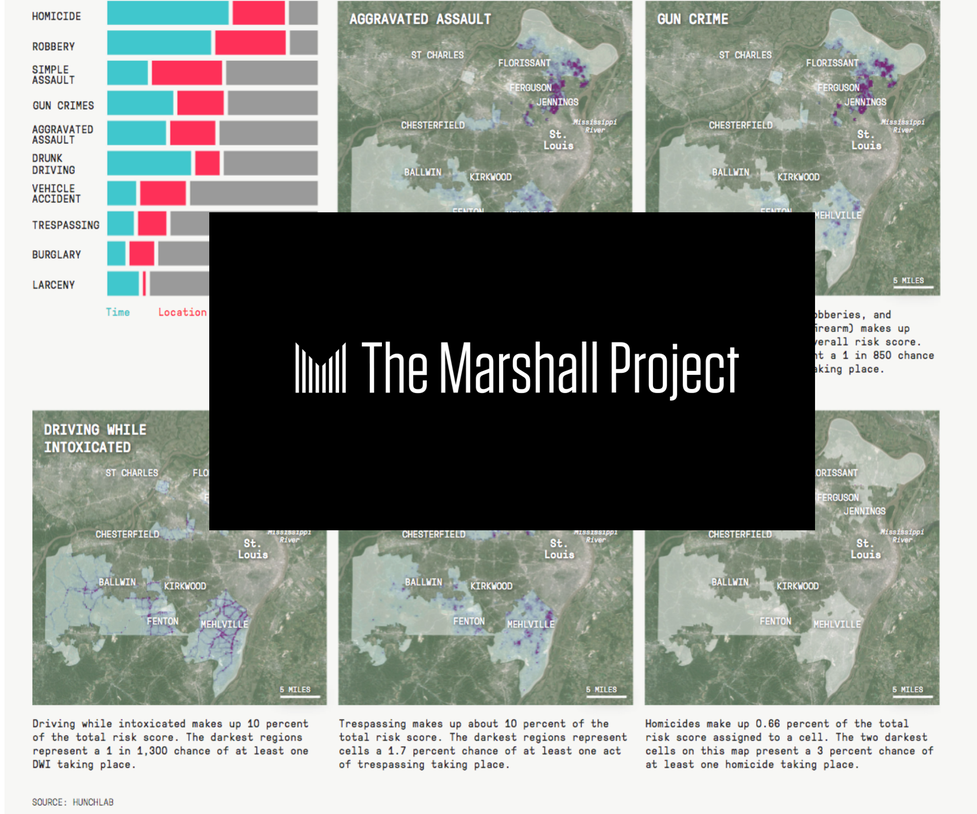
Joy Buolamwini

https://www.ajlunited.org/

Journalists can be data science superheroes

flickr: jdhancock

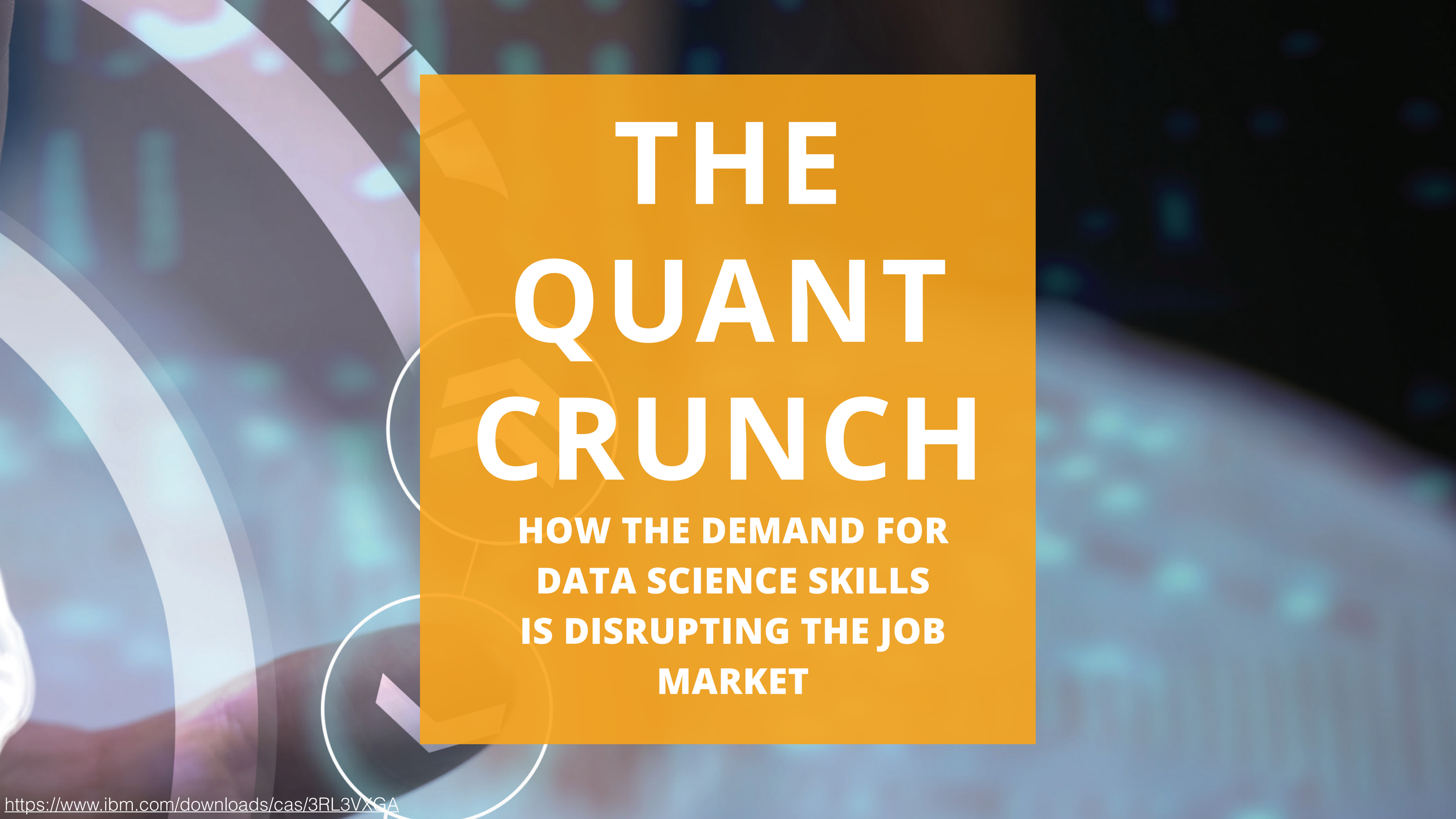

ProPublica
Donate
Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.
Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's 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
ProPublica
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.


HOMICIDE
ROBBERY
SIMPLE ASSAULT
GUN CRIMES
AGGRAVATED ASSAULT
DRUNK DRIVING
VEHICLE ACCIDENT
TRESPASSING
BURGLARY
LARCENY
Time
Location
AGGRAVATED ASSAULT
GUN CRIME
The Marshall Project
DRIVING WHILE INTOXICATED
ST CHARLES
FLORISSANT
FERGUSON
JENNINGS
CHESTERFIELD
St. Louis
BALLWIN
KIRKWOOD
FENTON
MEHLVILLE
5 MILES
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


SEARCH
The New York Times
POLITICS
How Trump Consultants Exploited the Facebook Data of Millions
Leer en español
By MATTHEW ROSENBERG, NICHOLAS CONFESSORE and CAROLE CADWALLADR MARCH 17, 2018
2086
The New York Times


Racial bias alleged in Google's ad results
Names associated with blacks prompt link to arrest search
19
The Boston Globe
(c)
(d)

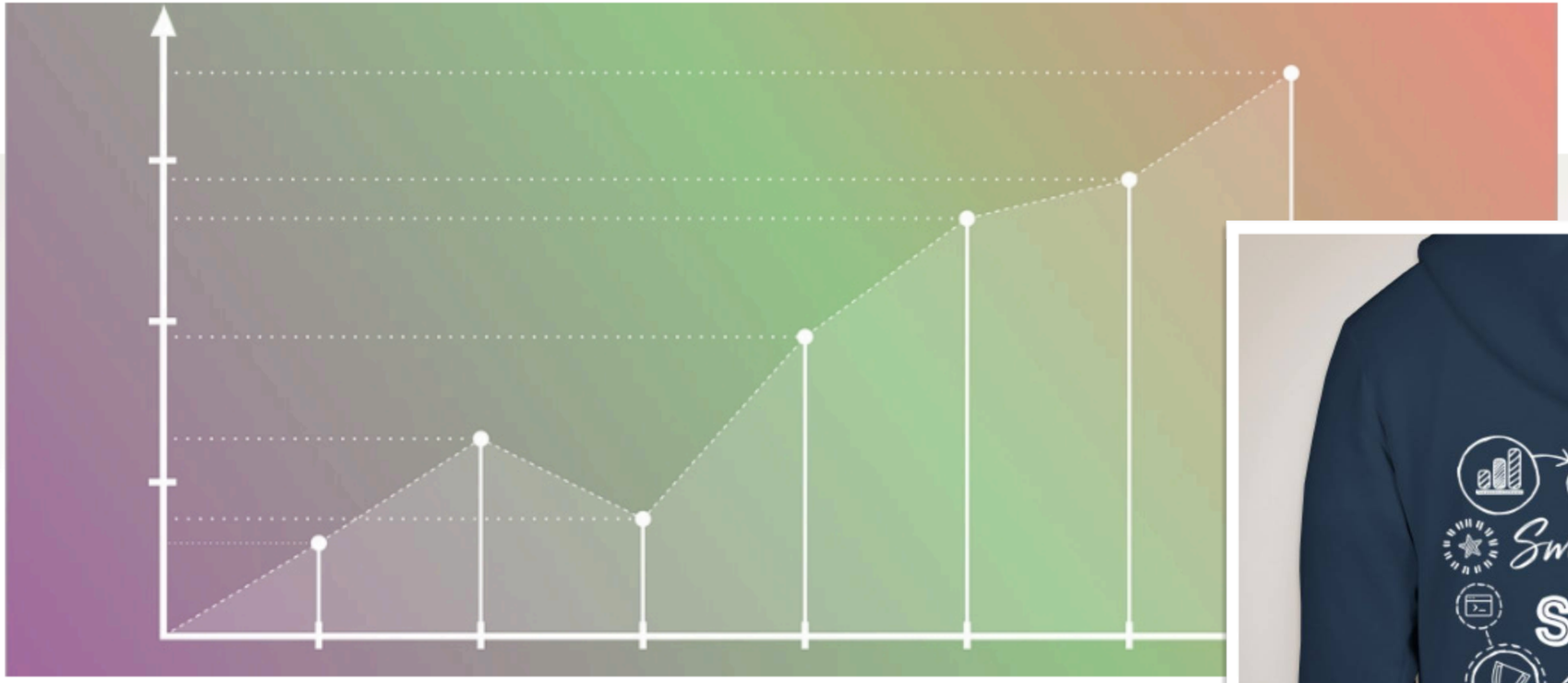
flickr: marcosit



https://www.ibm.com/downloads/cas/3RL3VXGA

https://www.introdatascience.org/

Statistical & Data Sciences



The Statistical & Data Sciences (SDS) Program links faculty and students from acros
interested in learning things from data. At Smith, students learn statistics by **doing**—

https://www.smith.edu/academics/statistics

Schedule

Please read the specified sections *after* the class posted, but *before* the next class.

Part 1: Descriptive Statistics

Mon, Sept 4

- NO CLASS

Wed, Sept 6

- NO CLASS

Thurs, Sept 7

- NO LAB

Fri, Sept 8

- Questionnaire
- Introduction
- Randomization: Is Yawning Contagious?
- Watch MythBusters on yawning
- Homework:
  - Read OI, Ch. 1.1-1.3

Mon, Sept 11

- Data & Data Collection (link to Google Spreadsheet)
- Homework:
  - Read OI, Ch. 1-4, 1-5

Wed, Sept 13

- **HW 1 due today**
- Spurious Correlations
- Sampling and Experiments
- Homework:
  - Read OI, Ch. 1.6

Thurs, Sept 14

- Intro to R and RStudio
- R Study Guide

Fri, Sept 15

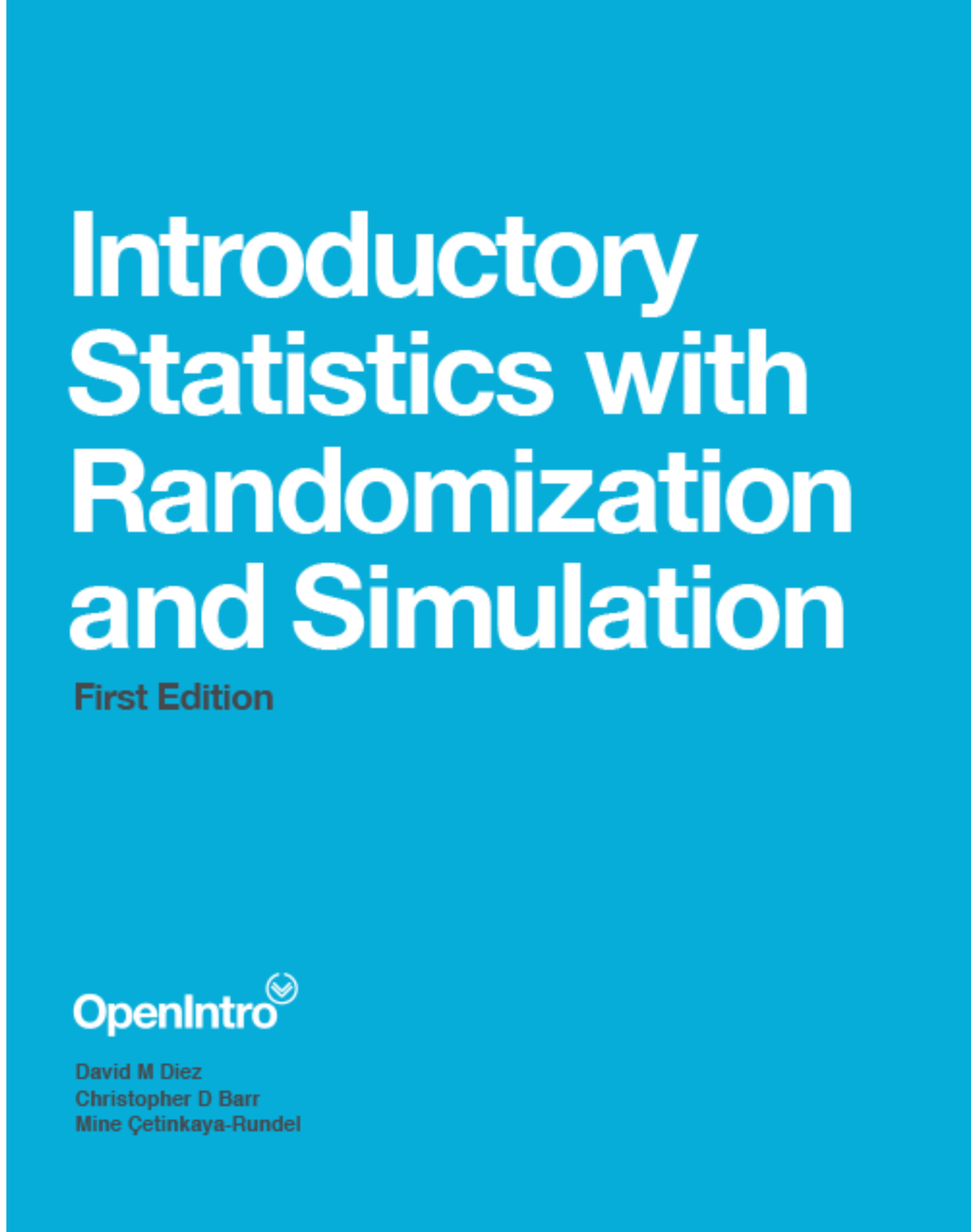
- Numerical data
- Homework:
  - Read OI, Ch. 1.7
  - Extra reading: Nature editorials on reproducible research
  - Watch videos from lab 1

http://www.amelia.mn/sds220/index.html

Introductory Statistics with Randomization and Simulation.

David M Diez,
Christopher D Barr, Mine Çetinkaya-Rundel

www.openintro.org



Schedule

Fri, Jan 25

- Introduction to Data Science
- Slack
- Questionnaire
- Reading: *Modern Data Science with R*, Ch. 1

Module 1: Data Visualization

Mon, Jan 28

- Introduction to R and RStudio
- Lab 1: Introduction to R and RStudio
- Sign Up for DataCamp

Wed, Jan 30

- Taxonomy of data graphics
- Lab 2: Deconstructing data graphics
- Reading: *MDSR*, Ch. 2

Fri, Feb 1

- Catch-up
- Reading: *MDSR*, Ch. 3.1

Mon, Feb 4

- `ggplot2`
- Lab 3: `ggplot2`
- Data Viz cheat sheet
- Reading: *MDSR*, Ch. 3.2

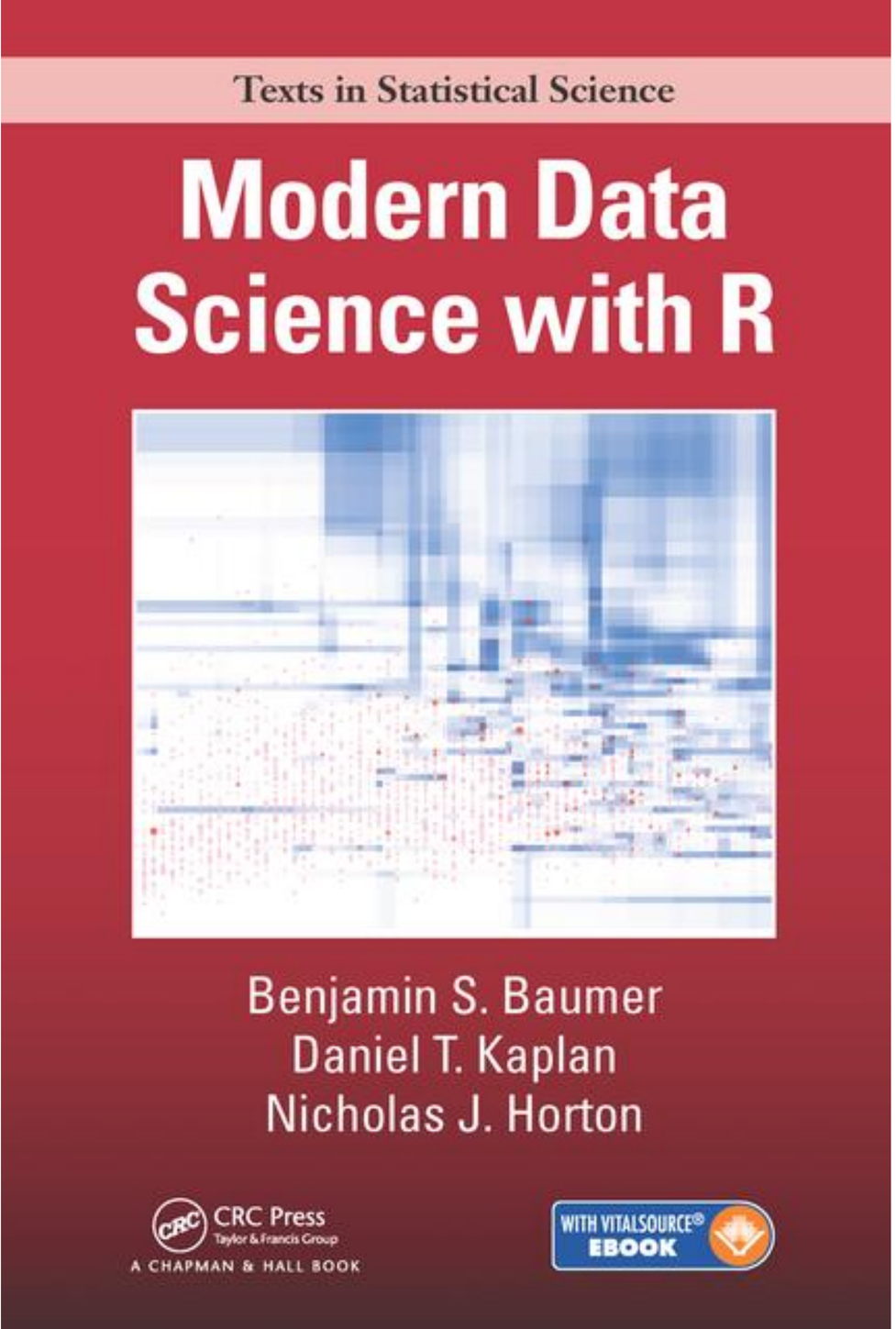
Wed, Feb 6

- Color
- Lab 4: `colorbrewer` and `ggthemes`
- Reading: *MDSR*, Ch. 3.3–3.4
- DataFest info session
  7–8pm, Sabin-Reed 301
- MULTIPLE IMPUTATION METHODS IN CLUSTER RANDOMIZED TRIALS
  Brittney Bailey, Amherst College
  Thursday, Feb. 7 12:10 pm,
  McConnell B15

Fri, Feb 8

- Common mistakes
- Lab 5: respond to prompt on `#lab5` Slack channel

https://beanumber.github.io/sds192/index.html



Modern Data Science with R.

Ben Baumer, Danny Kaplan, Nick Horton.

https://mdsr-book.github.io/

Search...

Built with ♥ and blogdown, logo by muuuuge.

Hello #dsbox

Hello! And welcome!

There is a little bit of something for everyone (who wants to teach/learn) data science in this box.

If you are an educator we recommend consuming the content in the order presented here: first familiarize yourself with the design principles, course syllabus, and the tech stack, then browse the course content, and then review the details of the computing infrastructure of the course.

If you are a learner, you might consider jumping straight into the course content.

https://datasciencebox.org/

Design principles

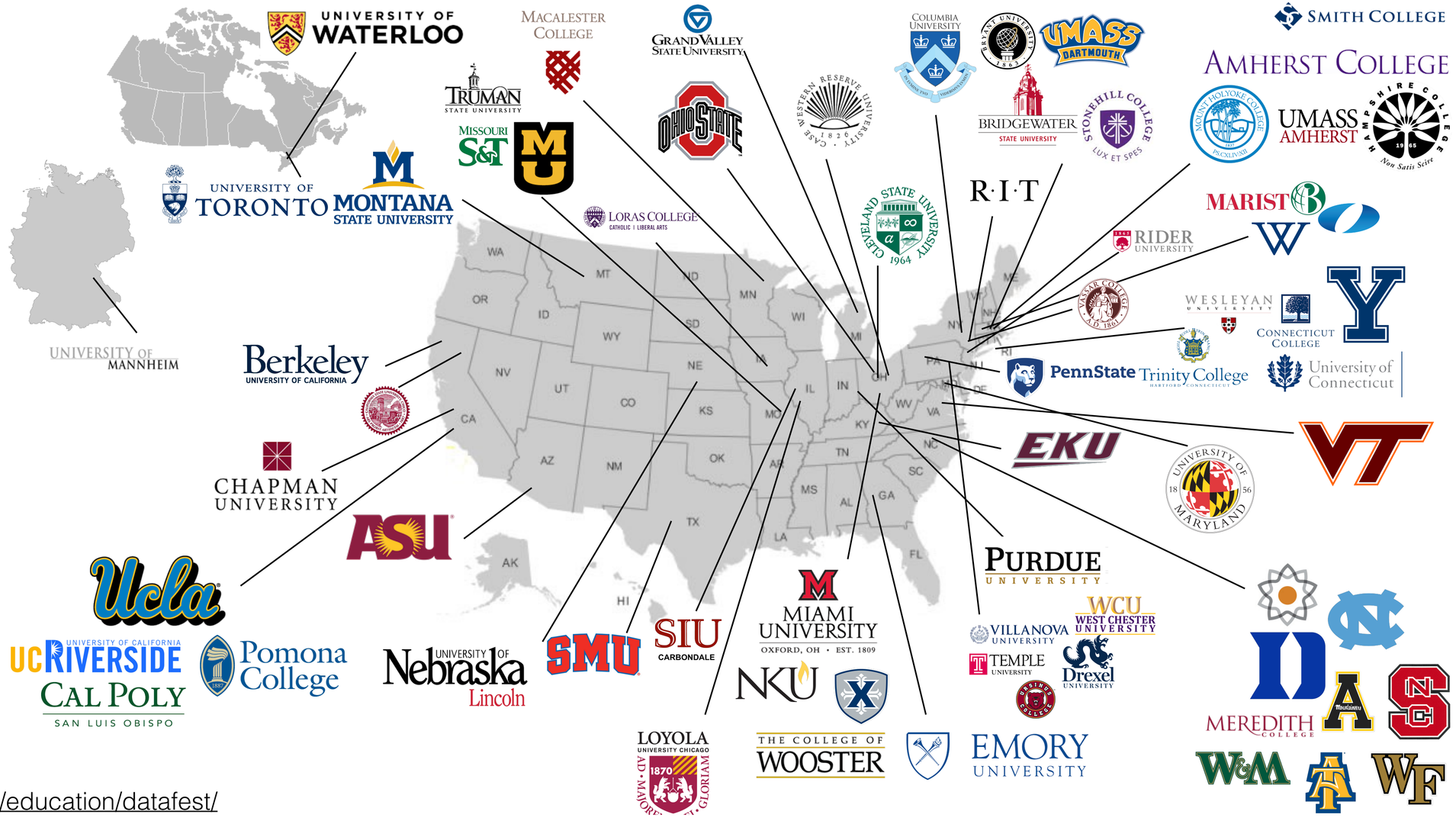
This course is designed with five principles in mind:

1. Start with cake
2. Cherish day one
3. Skip baby steps
4. Hide the veggies
5. Leverage the ecosystem

Start with cake

Assuming you like chocolate and strawberries, which of the following images is more likely to make you want to learn to bake a cake? I'm guessing the answer is the image on the left: the cake.

ASA DataFest™



https://ww2.amstat.org/education/datafest/

flickr: sheila_dee

flickr: Hannaford