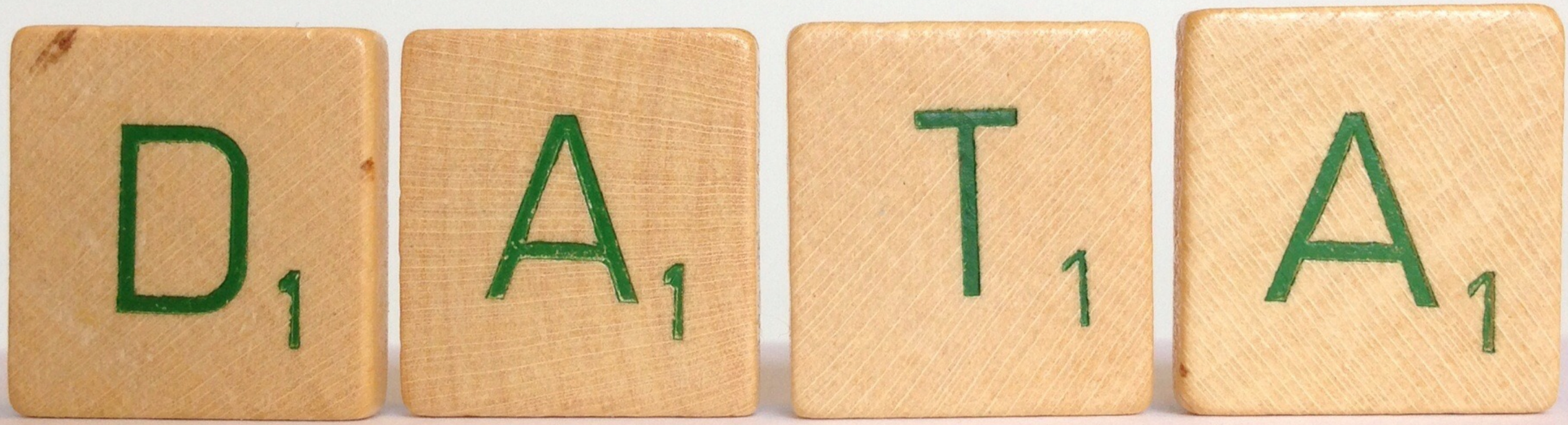
Data Science as a Superpower

Amelia McNamara
Smith College Program in Statistical and Data 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

Search Google Maps

ENCINO
SHERMAN OAKS
Universal City
Glendale
EAGLE ROCK
Pasadena
Sierra Madre
Arcadia
Duarte
BEL AIR
LOS FELIZ
HOLLYWOOD
Alhambra
Rosemead
Baldwin Pa
Beverly Hills
CENTRAL LA
WESTWOOD
CHINATOWN
Monterey Park
El Monte
West
BRENTWOOD
Los Angeles
South El Monte
PACIFIC PALISADES
DOWNTOWN
East Los Angeles
Montebello
City of Industry
Santa Monica
Culver City
Commerce
Hacienda Heights
SOUTH LOS ANGELES
Marina Del Rey
Ladera Heights
Los Angeles International Airport
Huntington Park
Pico Rivera
Whittier
Inglewood
Bell Gardens
South Gate
Santa Fe Springs
South Whittier
Downey
Lynwood
El Segundo
Hawthorne
Willowbrook
La Mirada
Norwalk
Compton
Paramount
Manhattan Beach
Gardena
Buena Park
Redondo Beach
Lakewood
Cerritos
Torrance
Carson

Satellite

Live traffic
Fast
Slow

Select all images with a

bus

Click verify once there are none left.

VERIFY

flickr: sergiorojas

BlueDot studies how infectious diseases disperse worldwide through analysis of big data

2003

SARS spreads from Hong Kong to Toronto, causing an outbreak that cripples the city with 44 deaths and $2 billion in losses

After completing his training as a physician specializing in infectious diseases and public health in New York City, Dr. Kamran Khan returns to Toronto just before the city's SARS outbreak begins.

2008

BioDiaspora is launched as a scientific research program at *St. Michael's Hospital* in Toronto

To understand how infectious diseases can spread worldwide, Dr. Khan launches BioDiaspora – a scientific research program to study how the world's population is connected through commercial air travel.

https://bluedot.global/

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

A measure of ocean health across countries and high seas regions.

GLOBAL SCORES

METHODOLOGY

INDEPENDENT ASSESSMENTS

OHI+ is an assessment tool customized by stakeholders to meet local management needs.

OHI+ ASSESSMENTS

CONDUCT AN ASSESSMENT

DataKind

California

Moulton Niguel Water District

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

THE NEW YORKER

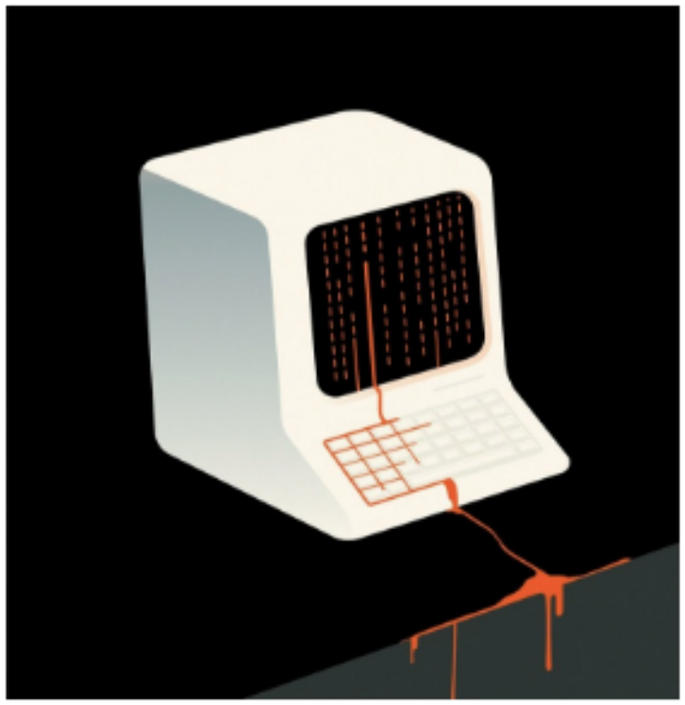
ANNALS OF CRIME NOVEMBER 27, 2017 ISSUE

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

Forbes / Tech

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/

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/

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/

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

...

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

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

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

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

...

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

...

$$p(i \rightarrow j) = logit(\beta_0 + \beta_i x_1 \ldots)$$

...

$$p(1) = \beta \, \text{ex\_cell} + \alpha_1 \, p(1 \mid \text{stylist}) + \alpha_2 \, p(1 \mid \text{client})$$

Algorithms Tour

How data science is woven into the fabric of Stitch Fix

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

facebook research

Research Areas | Publications | People | Academic Programs | Downloads & Projects | Careers | Blog

Our People

View All >

Moira Burke
Data Scientist

Eytan Bakshy
Empiricist

Lada Adamic
Data Scientist

Shaomei Wu
Data Scientist

Alex Dow
Data Scientist

Latest Publications

View All >

The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People with Visual Impairments

Lindsay Reynolds, Shaomei Wu, Shiri Azenkot, Yuhang Zhao

CSCW 2018 - November 3, 2018

A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab

Lindsay Reynolds, Shaomei Wu, Shiri Azenkot, Yuhang Zhao

CHI 2018 - April 21, 2018

Social Structure and Trust in Massive Digital Markets

David Holtz, Diana Lynn MacLean, Sinan Aral

ICIS 2017 - December 10, 2017

https://research.fb.com/category/data-science/

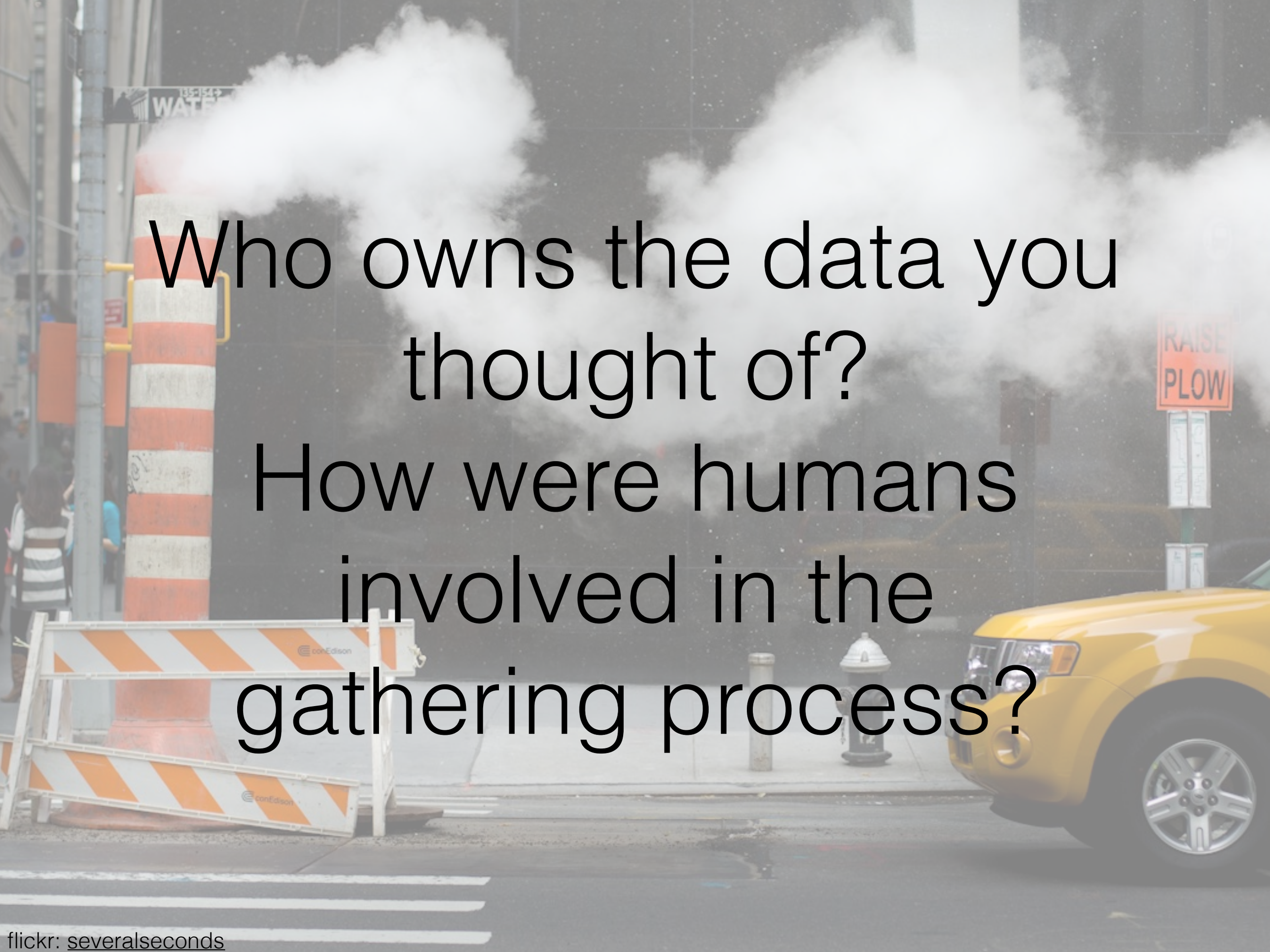
AND, A SHORT DISTANCE AWAY...

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!

flickr: philwolff



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

flickr: severalseconds

flickr: mermaid99

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

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

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.

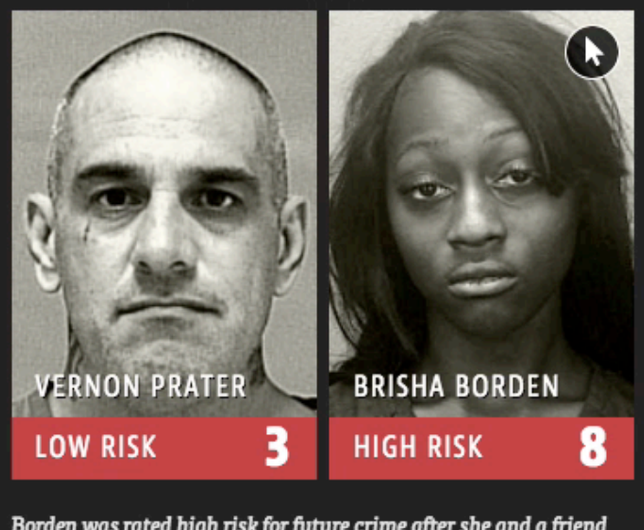
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

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

FEATURE

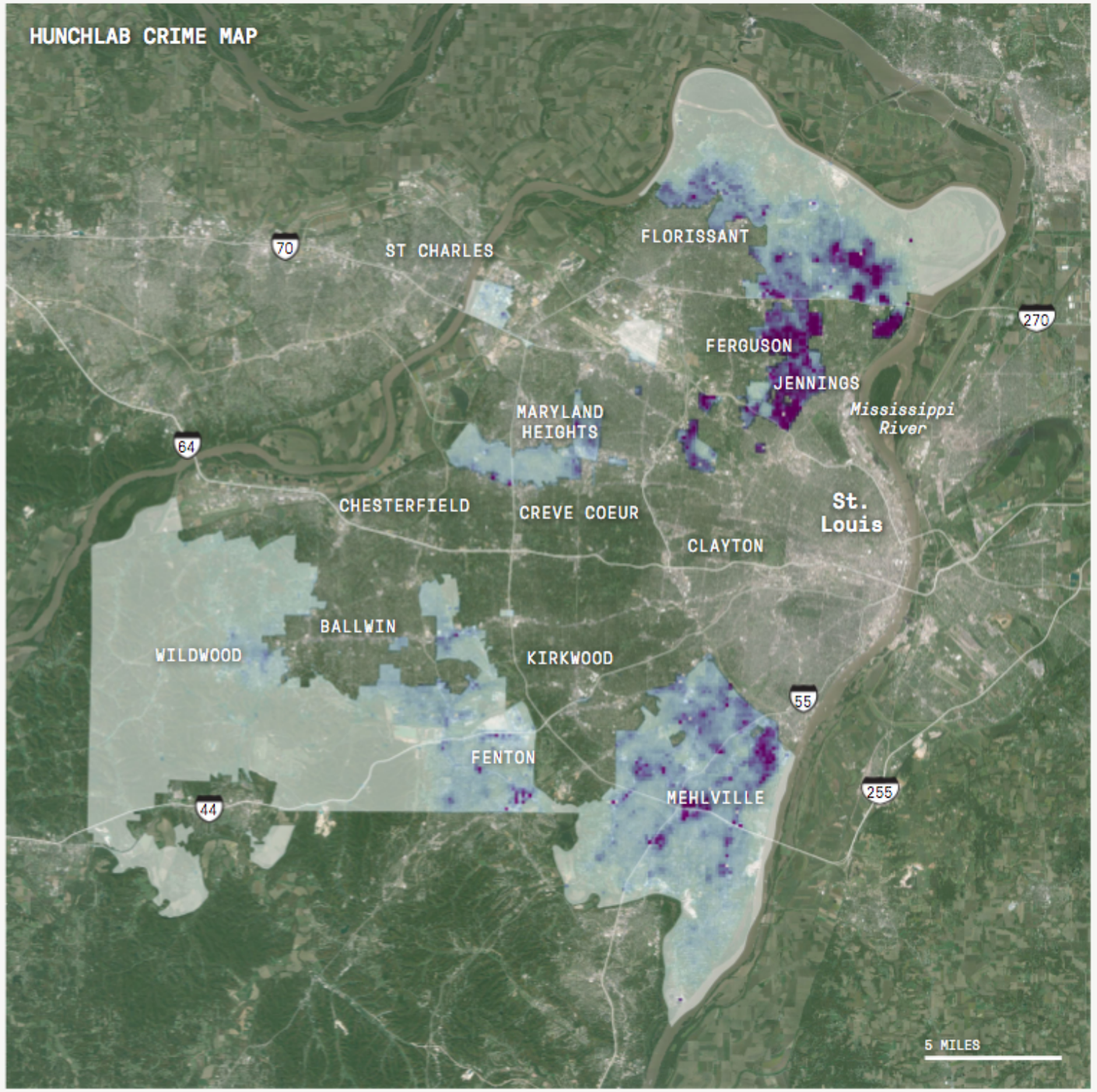
Policing the Future

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

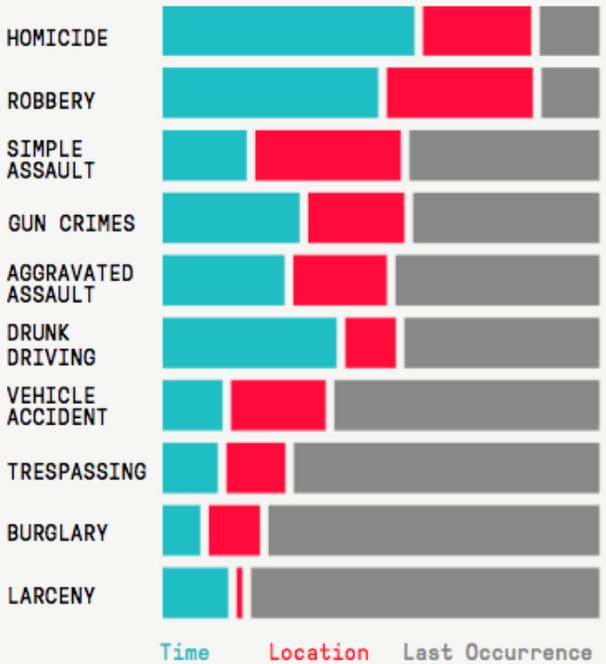
Dozens of small, local municipal agencies handle policing in parts of St. Louis County. The St. Louis County Police Department covers areas not policed by the "munis," including the city of Jennings, Mo. The ▪ DARKER AREAS in the map show the areas within their jurisdiction that HunchLab has identified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

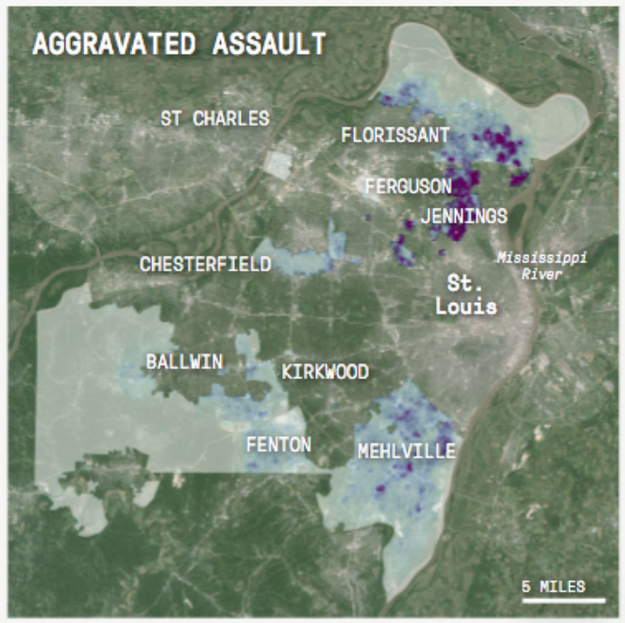


SOURCE: HUNCHLAB

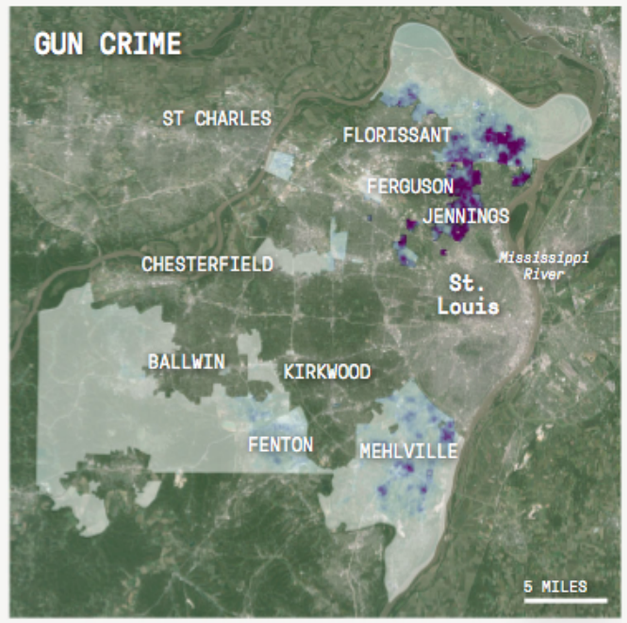
Maurice Chammah, with additional reporting by Mark Hansen. Policing the Future.
https://www.themarshallproject.org/2016/02/03/policing-the-future

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.

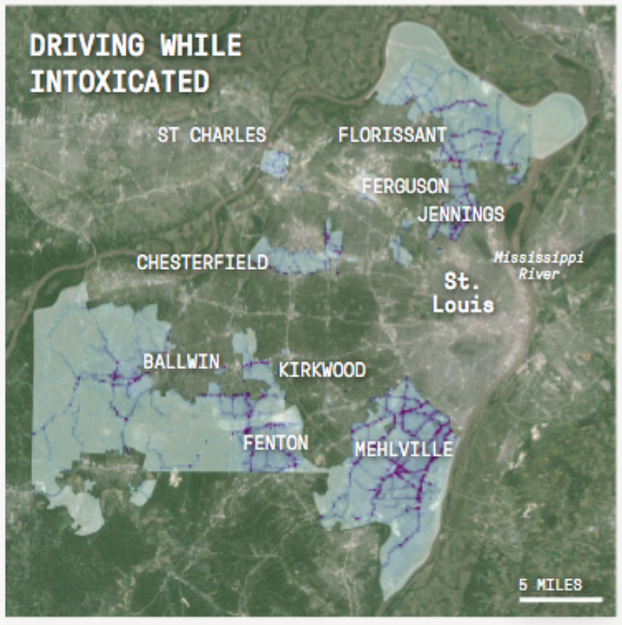




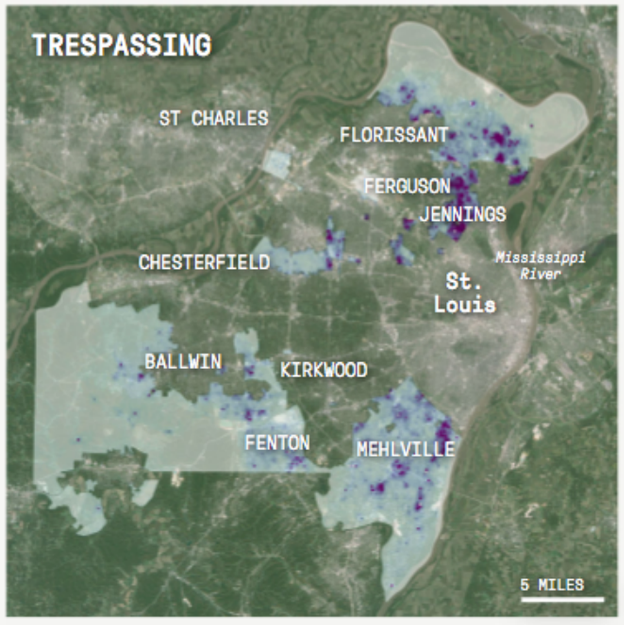
Aggravated assault (assault with a dangerous weapon) makes up 18.5 percent of the overall risk score assigned to a cell. The darkest regions on this map represent cells with a 1 in 320 chance of at least one aggravated assault taking place there during the shift.



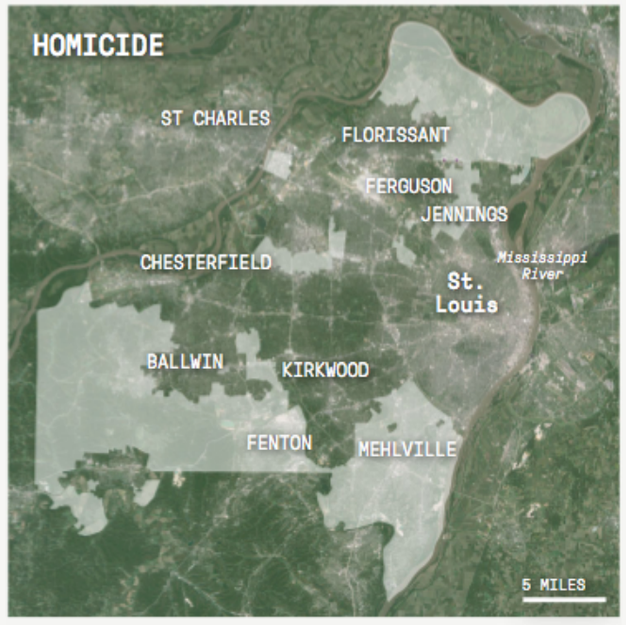
Gun crime (all homicides, robberies, and aggravated assaults with a firearm) makes up about 16.5 percent of the overall risk score. The darkest regions represent a 1 in 850 chance of at least one gun crime taking place.



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.



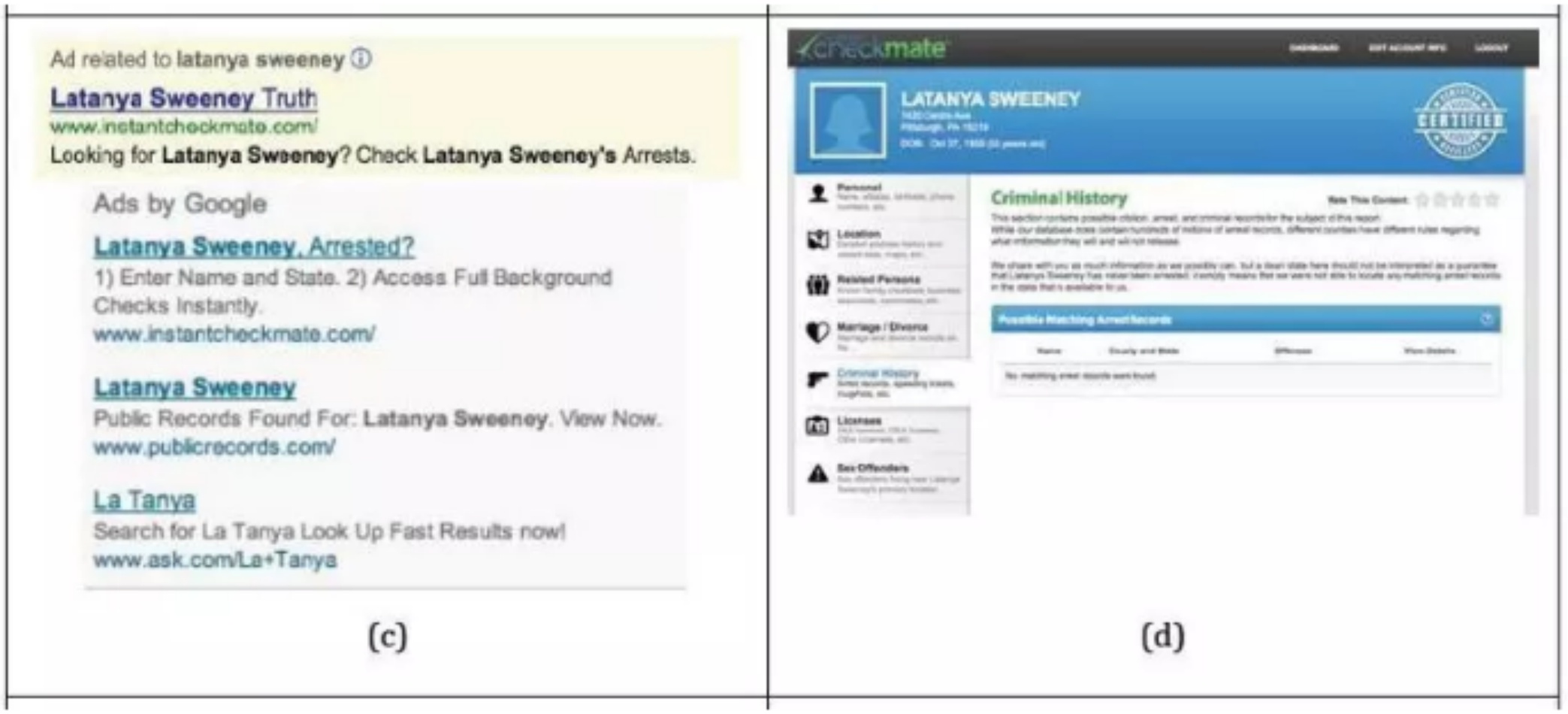
Homicides make up 0.66 percent of the total risk score assigned to a cell. The two darkest cells on this map present a 3 percent chance of at least one homicide taking place.

SOURCE: HUNCHLAB

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

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.h

flickr: Hannaford

Dr. Latanya Sweeney

https://dataprivacylab.org

Dr. Jake Porway

www.datakind.org/our-team

Journalists can be data science superheroes

flickr: jdhancock

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 Commission to study their use. "Although these measures were crafted with the best of intentions, I

Borden was rated high risk f took a kid's bike and scooter reoffend.

hidde

We ob
Count
crime
algori

The s

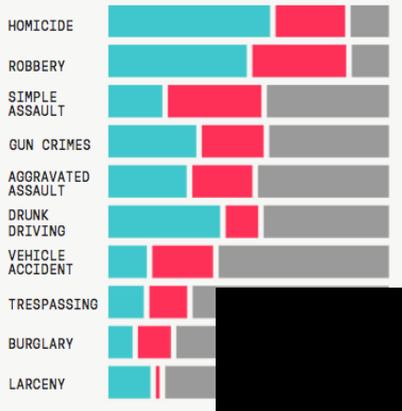
the people predicted to commit violent crimes actually went on to do so.

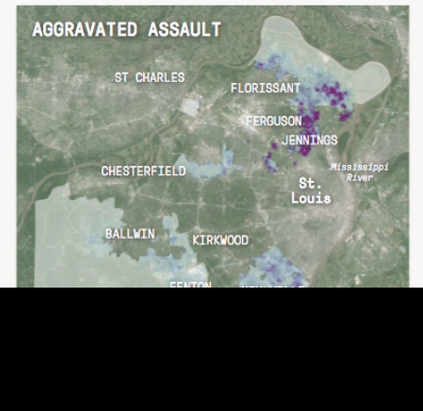
When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

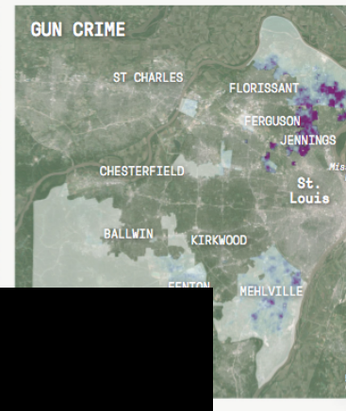
We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

PRO PUBLICA

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

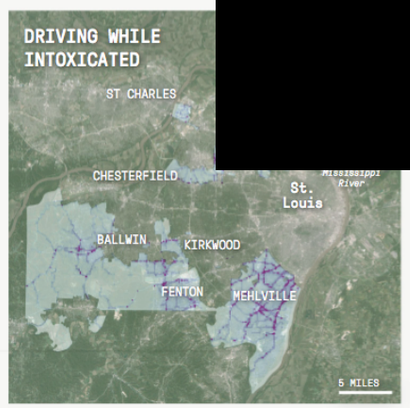


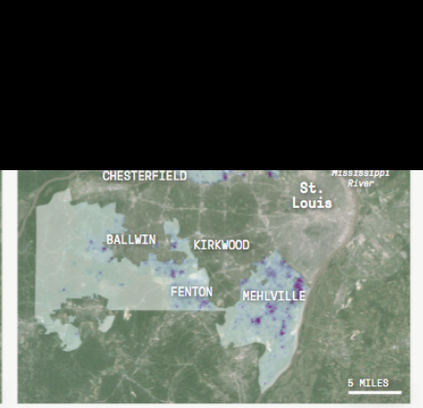


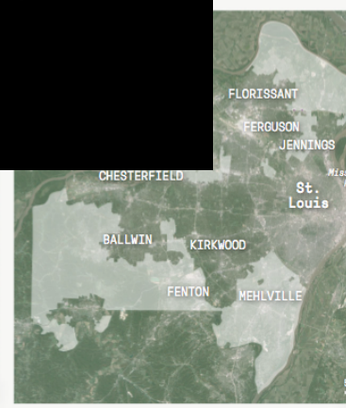


, robberies, and
a firearm) makes
e overall risk s
esent a 1 in 850
e taking place.

The Marshall Project







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 tot risk score assigned to a cell. The two da cells on this map present a 3 percent cha at least one homicide taking place.

SOURCE: HUNCHLAB

How Trump Consultants Exploited the Facebook Data of Millions

Leer en español

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

2086

Racial bias alleged in Google's ad resul

Names associated with blacks prompt link to arrest searc

The Boston Globe


flickr: sheila_dee

THE QUANT CRUNCH

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

flickr: jannekestaaks

MassMutual

Development | Engineering | Product | Science | Visualization | Community Data Consulting | Careers

Development

Our Data Science Development Program offers a unique industry leading program that combines high-impact industry experience with hands-on training, mentorship, and coursework.

Class of 2019

Lizzie Kumar

Jordan Menter

Martha Miller

Edward Pantridge

Karla Villalta

https://datascience.massmutual.com/development

ASA DataFest™

flickr:mujitra

Thank you

flickr: Hannaford