Statistics: the big picture

Some reminders, take home messages, and most importantly, professional ethics

Joshua Loftus

5/7/2018

Check the data

```
head(df)
```

```
## # A tibble: 6 x 2
##       x     y
##   <dbl> <dbl>
## 1  55.4  97.2
## 2  51.5  96.0
## 3  46.2  94.5
## 4  42.8  91.4
## 5  40.8  88.3
## 6  38.7  84.9
```

Check summaries, fit models...

```
lm(y ~ x, df)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Coefficients:
## (Intercept)            x
##      53.453       -0.104
```

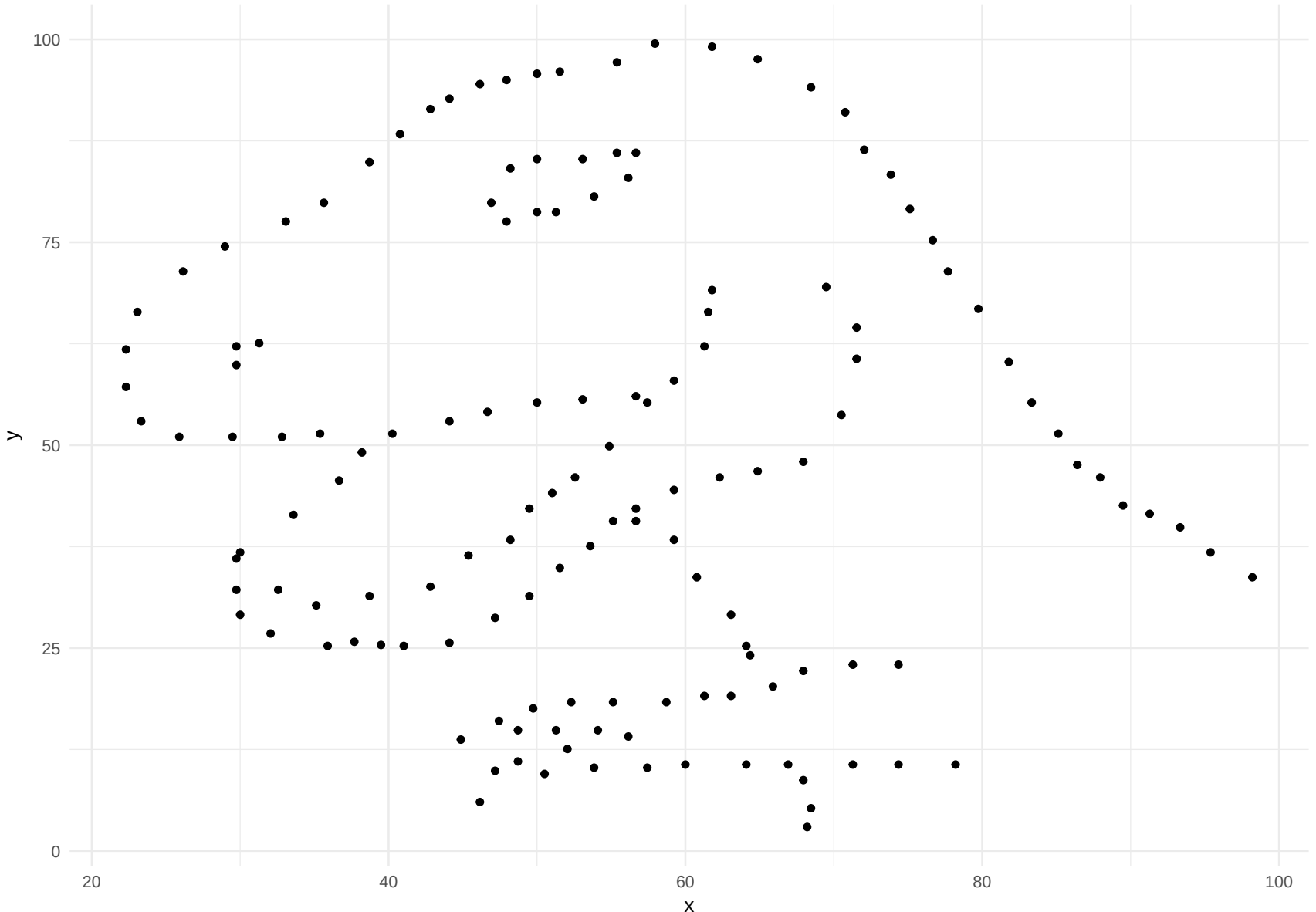
```
cor(df$x, df$y)
```

```
## [1] -0.06447
```

Did we forget something?

Always plot the data!

```
qplot(x, y, data = df)
```



Appreciate variability

- Samples of data are fallible, incomplete, often noisey / low quality, and sometimes only indirectly related to the question you are really interested in. Even the census, which is supposed to be the whole population and not a sample, is imperfect
- Estimates calculated from data: means, regression coefficients, predicted outcomes, classifications, etc, all have some amount of inherent variation in them due to being calculated from samples
- Always insist on information about that variability, like confidence intervals

Appreciate bias

- Think about how the data is collected...
- WW2, Abraham Wald, Statistical Research Group, planes returning with bullet holes in them. Military analyzed the data, found the parts of the plane with the most bullet holes, and decided that was where to put extra armor. Wald said to do the opposite! Why?
- Survey our class–how many ways would that not be representative of (1) all students at NYU, (2) everyone in NYC, (3) everyone in the US, (4) everyone in the world right now, (5) all people throughout history
- Measurement bias: do the variables actually measure what we think they're measuring? Is there systematic error?
- The "woman without a face" - traces of her DNA found at 40 different crime scenes. Police spent 15 years investigating, even offered 3 million euro reward... It was just contaminated cotton swabs

Use your imagination

- Try to think of **interesting counterfactuals** – "what if" questions
- What if there are confounding variables?
- What if the associations in the data are due to ecological correlations / Simpson's paradox? What other data would you like to see to check for this?
- What if we analyzed a different dataset, with possibly different predictor or outcome variables? What if we used a different set of predictors?
- What if the analysis had begun with a different question?

Be vigilant

- Check the details. What are the units? Is the plot misleading? Does the more detailed content of the article contradict the title or abstract/summary?
- Don't assume other people are either honest or dishonest without evidence of either first
- Consider your own biases, those of your audience, and the motives of whoever collected and analyzed the data (publish or perish in science, prove value in business)
- If and when a sufficient standard of evidence has been reached, let it change your mind!
- Seek more evidence, question your assumptions, etc
- Ask for advanced methods (e.g. deep learning) to be compared to simple baselines like linear/logistic regression

Ethics

- Link: Guidelines from the American Statistical Association

The ethical statistician:

- Identifies and mitigates any preferences on the part of the investigators or data providers that might predetermine or influence the analyses/results.
- Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms.
- Reports the limitations of statistical inference and possible sources of error.

Ethics, continued

The ethical statistician:

- Where appropriate, addresses potential confounding variables not included in the study.
- Keeps informed about and adheres to applicable rules, approvals, and guidelines for the protection and welfare of human and animal subjects.
- Recognizes any statistical descriptions of groups may carry risks of stereotypes and stigmatization. Statisticians should contemplate, and be sensitive to, the manner in which information is framed to avoid disproportionate harm to vulnerable groups.

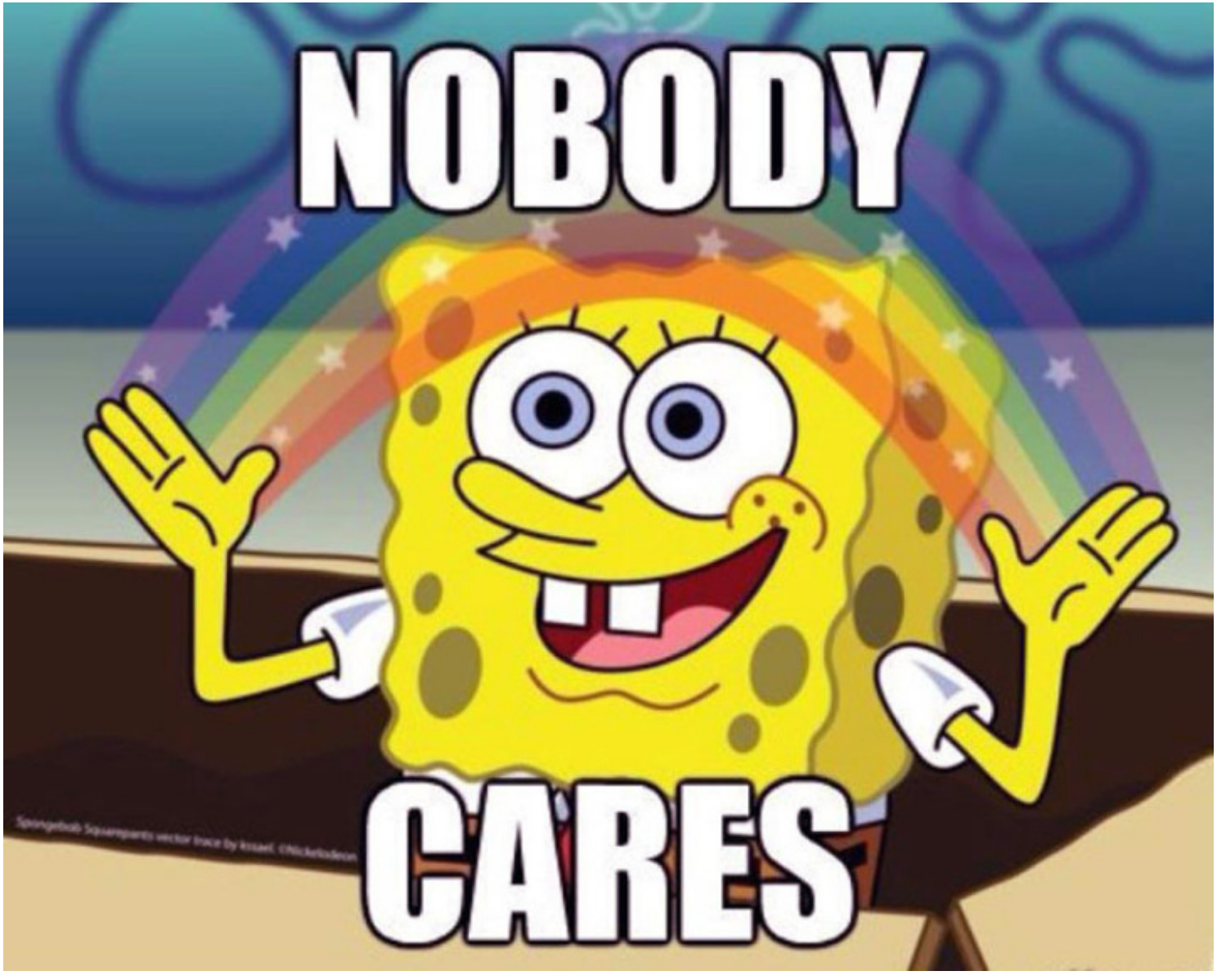
Ethics for employers of statisticians

Those employing statisticians are expected to:

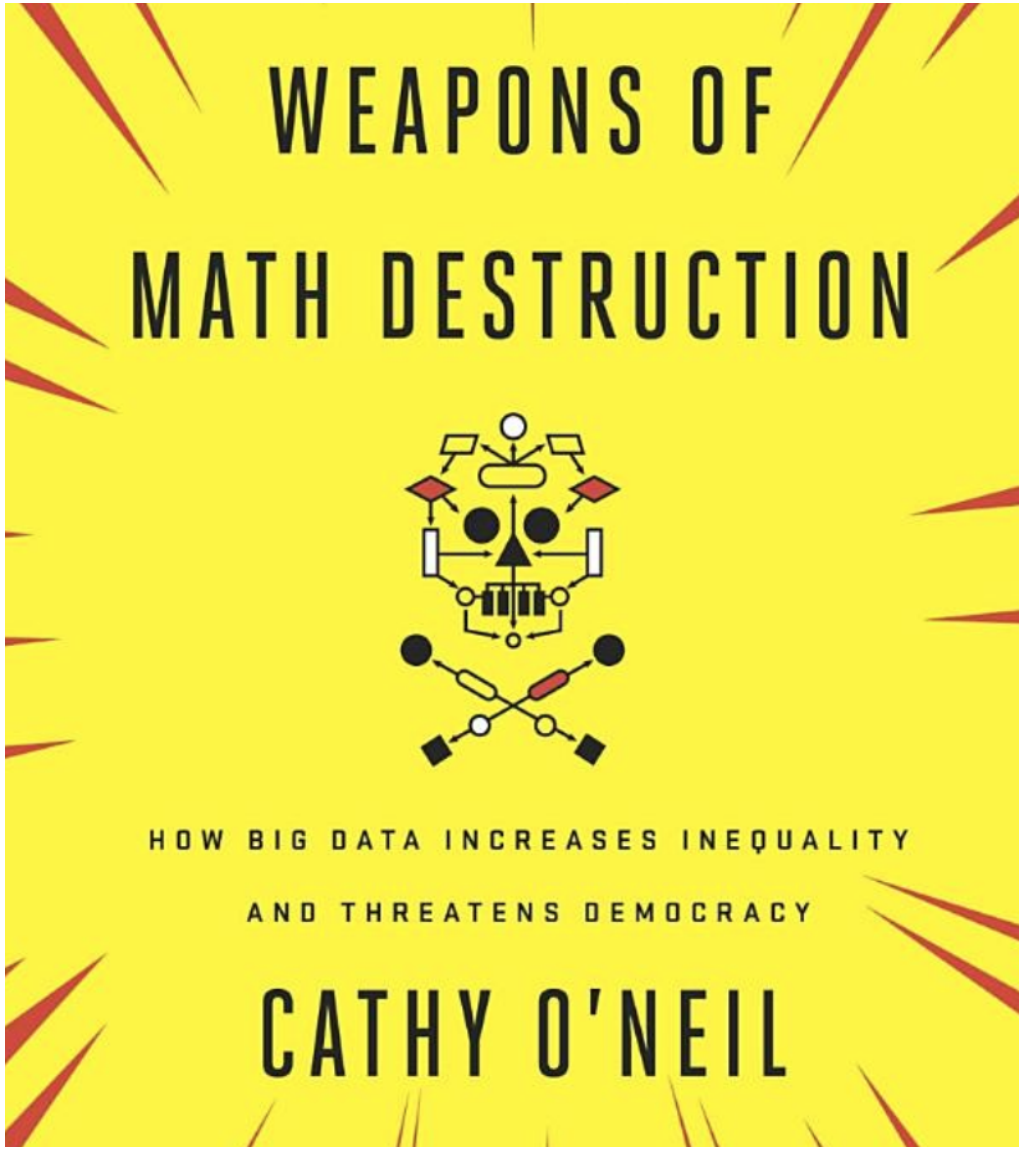
- Recognize that valid findings result from competent work in a moral environment.
- Recognize it is contrary to these guidelines to report or follow only those results that conform to expectations without explicitly acknowledging competing findings

How long is he gonna go on about ethics?

Whew, it's over!



Why you should care...



- Admissions
- Credit
- Employment
- Insurance
- Healthcare

Unique dangers: scale, opacity, biased data, faux objectivity

What if you were falsely convicted for a crime because of “science”?

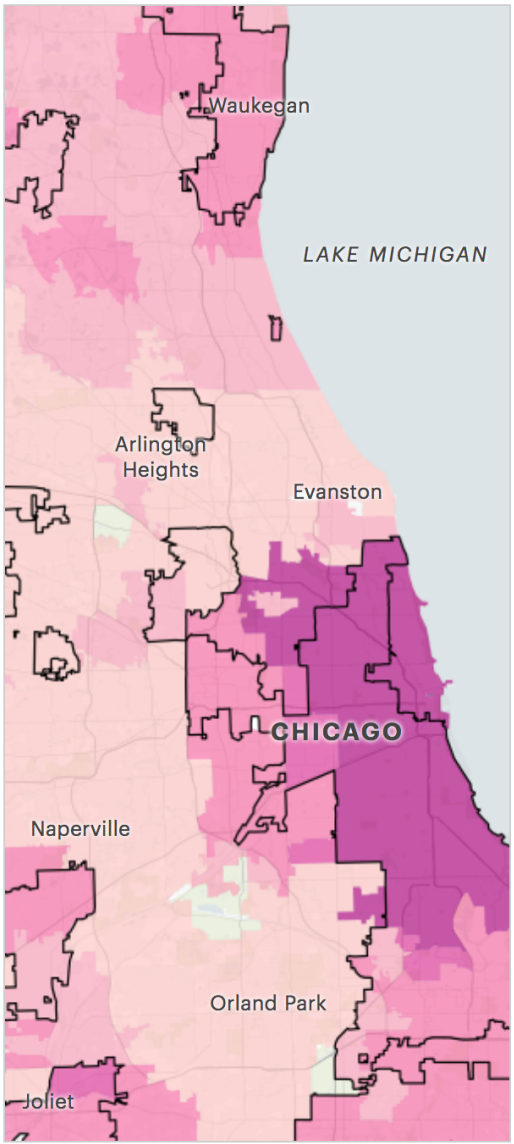
Traces of Crime: How New York’s DNA Techniques Became Tainted

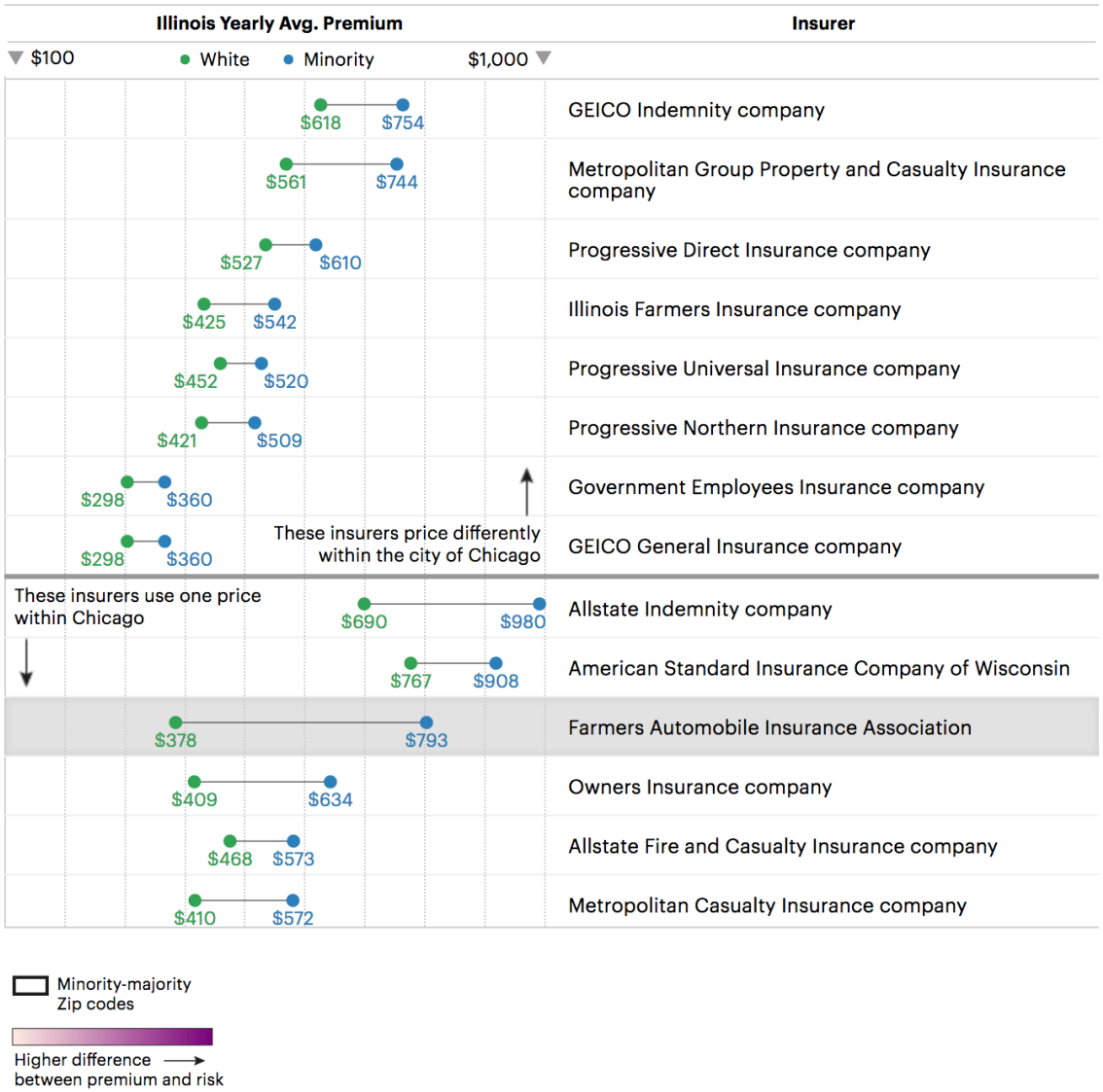
The city’s medical examiner has been a pioneer in analyzing complex DNA samples. But two methods were recently discontinued, raising questions about thousands of cases.

By LAUREN KIRCHNER SEPT. 4, 2017

Source and some good news

What if your car insurance was more expensive because you lived in a minority neighborhood?





Source

Facebook ads. . . don't even get me started

AARP and the senators were reacting to a Dec. 20 report by ProPublica and The New York Times that dozens of the nation's leading employers, including Facebook itself, narrow their audience for job ads on Facebook and other platforms by age. The ability of advertisers to direct their messages at specific groups is a cornerstone of Facebook's business. But such micro-targeting becomes controversial when it fosters discrimination in legally protected categories such as race and age. ProPublica has reported that Facebook also accepted ads aimed at "Jew-haters" as well as housing ads that discriminated by race, gender, disability and other factors.

Read More

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.

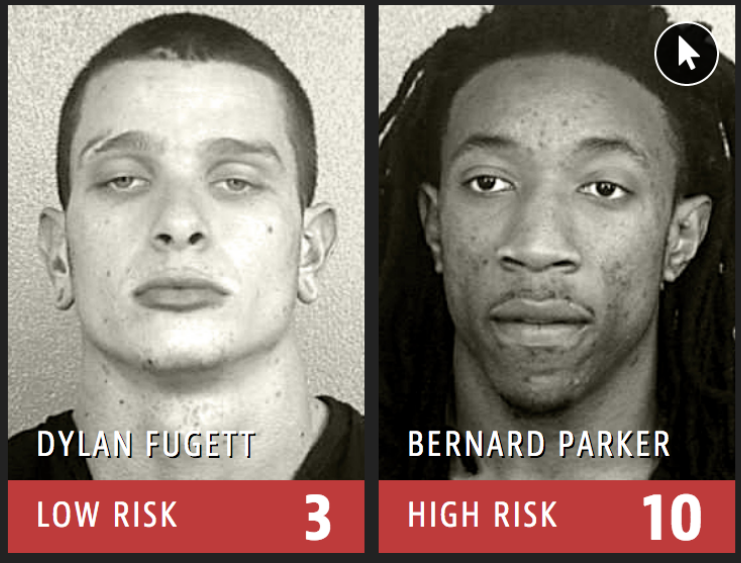
Source

What if your freedom depended on a proprietary algorithm that might be biased against you?

Dozens of risk assessments are being used across the nation — some created by for-profit companies such as Northpointe and others by nonprofit organizations. (One tool being used in states including Kentucky and Arizona, called the Public Safety Assessment, was developed by the Laura and John Arnold Foundation, which also is a funder of ProPublica.)

There have been few independent studies of these criminal risk assessments. In 2013, researchers Sarah Desmarais and Jay Singh examined 19 different risk methodologies used in the United States and found that "in most cases, validity had only been examined in one or two studies" and that "frequently, those investigations were completed by the same people who developed the instrument."

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Source

What consequences may follow AI learning discriminatory word associations?

Here is a poem written by Google Translate on the topic of gender. It is the result of translating Turkish sentences using the gender-neutral "o" to English (and inspired by this Facebook post).

Gender

by Google Translate

he is a soldier
she's a teacher
he is a doctor
she is a nurse

he is a writer
he is a dog
she is a nanny
it is a cat

Source

Ethical failures

Discrimination

- Failure to apply relevant laws/regulations
- Poor design/incentives

"Unintentional" disparate impact

- Non-diverse workplaces
- Bias introduced by data collection/use
- Unfair world: many predictor variables correlated with both outcome and protected attributes (race, sex, etc)
- Not an excuse, ethical statistical practice requires us to anticipate problems like these

More people are starting to take these things seriously

- The EU has a General Data Protection Regulation, passed in 2016, which will take effect on May 25. It's a big deal
- Researchers now developing methods for preventing algorithms from learning to be racist/sexist/etc. First Conference on Fairness, Accountability, and Transparency occurred here at NYU in February this year
- This is an area that I work in, so if these topics interest you and you want to talk more about it after this semester stay in touch!

Complete your CFE

- Course Faculty Evaluation (CFE) – Stern takes these seriously! Your input is valued
- Constructive feedback is appreciated
- Please take the time to complete it now
- You should have received an email from the school with a link, and instructions

I hope that both your knowledge and interest in statistics has increased *significantly*