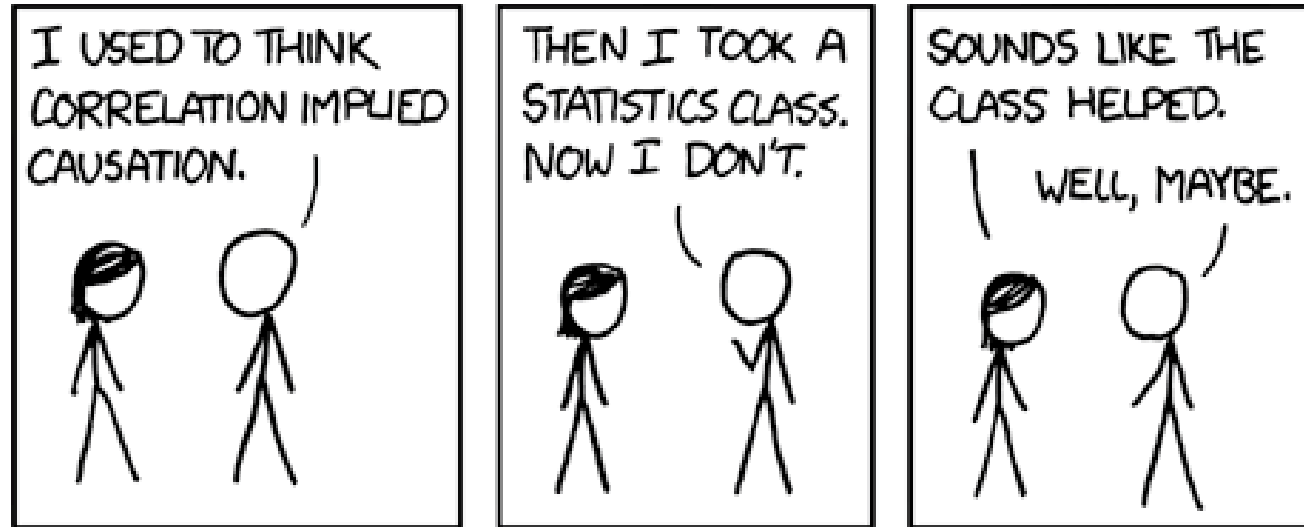


Patterns in data

Introduction to data analysis: Lecture 2

Ori Plonsky

Spring 2023



Source: xkcd

*Slides partially
based on data8
(<http://data8.org/>)*

Today

- Correlation vs. causation
- Establishing causality
- Confounders
- Experiments
- Biases in data

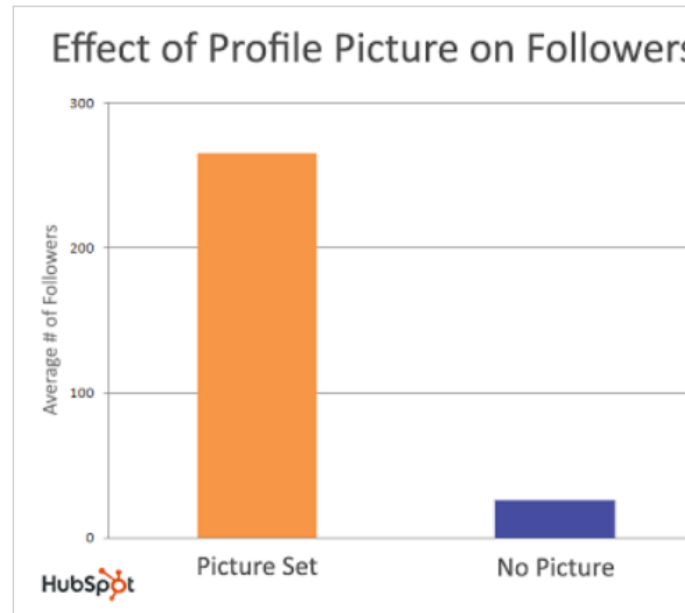
A pattern

WEDNESDAY, APRIL 7, 2010

Twitter Accounts with a Profile Picture Have 10 Times More Followers Than Those Without

When publishing Twitter statistics, one of the most common questions I'm asked is about the effect of setting a profile picture in your Twitter account.

I dug into Twitter Grader data and analyzed nearly 9 million Twitter profiles to produce the graph below.



Source: HubSpot

It doesn't take much, but if you want to get followers on Twitter, it's a good idea to upload a picture of yourself.

<http://anonymous1mill.blogspot.com/2010/04/twitter-accounts-with-profile-picture.html>

Observation

- Individuals (subjects, participants, units)
- Treatment
- Outcome

Observation

- Individuals (subjects, participants, units)
 - Twitter accounts
- Treatment
- Outcome

Observation

- Individuals (subjects, participants, units)
 - Twitter accounts
- Treatment
 - Having a profile picture
- Outcome

Observation

- Individuals (subjects, participants, units)
 - Twitter accounts
- Treatment
 - Having a profile picture
- Outcome
 - Number of followers

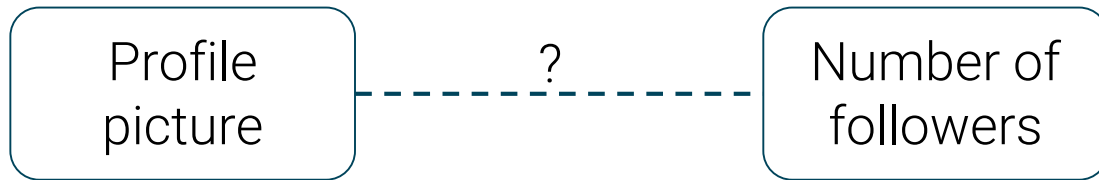
Observation

- Individuals (subjects, participants, units)
 - Twitter accounts
- Treatment
 - Having a profile picture
- Outcome
 - Number of followers

- Socrative student
 - Room name: PLONSKY

Question 1

Is there **any relation** between setting a profile picture and the number of followers?

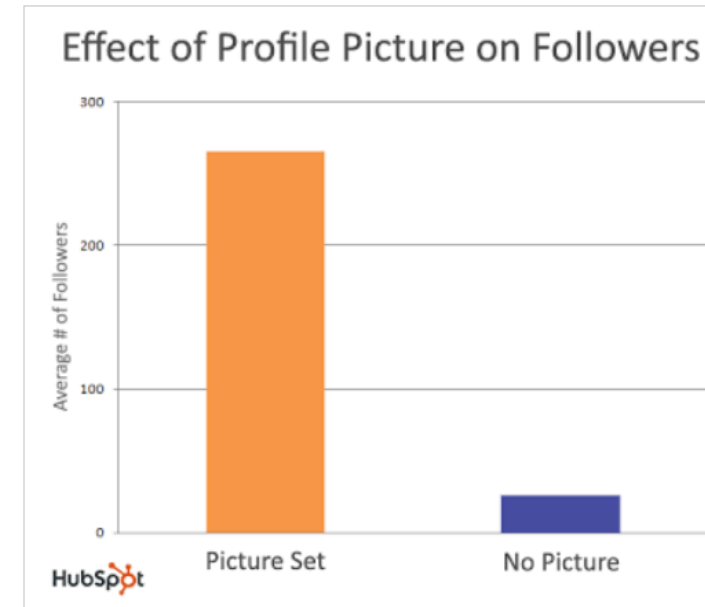


WEDNESDAY, APRIL 7, 2010

Twitter Accounts with a Profile Picture Have 10 Times More Followers Than Those Without

When publishing Twitter statistics, one of the most common questions I'm asked is about the effect of setting a profile picture in your Twitter account.

I dug into Twitter Grader data and analyzed nearly 9 million Twitter profiles to produce the graph below.

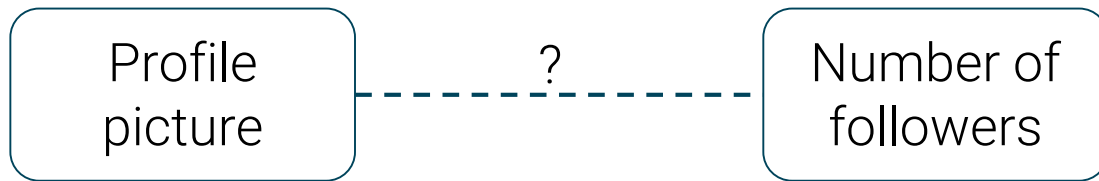


Source: [HubSpot](#)

It doesn't take much, but if you want to get followers on Twitter, it's a good idea to upload a picture of yourself.

Question 1

Is there **any relation** between setting a profile picture and the number of followers?



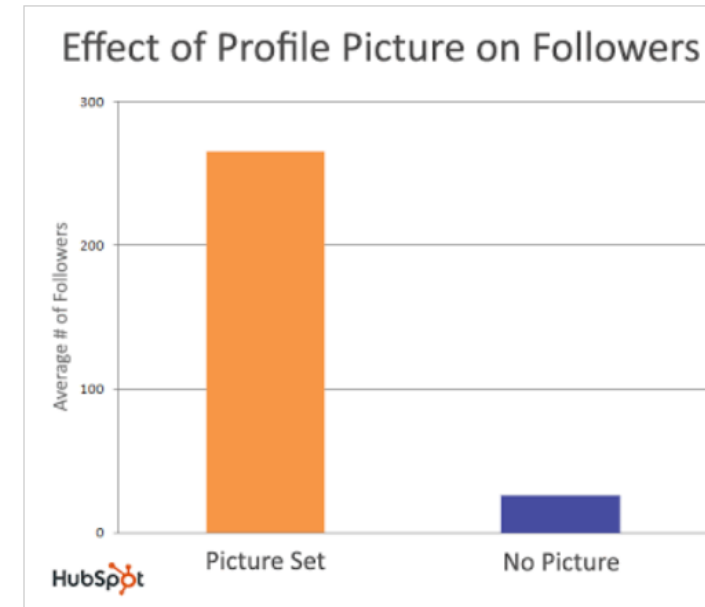
Correlation, association

WEDNESDAY, APRIL 7, 2010

Twitter Accounts with a Profile Picture Have 10 Times More Followers Than Those Without

When publishing Twitter statistics, one of the most common questions I'm asked is about the effect of setting a profile picture in your Twitter account.

I dug into Twitter Grader data and analyzed nearly 9 million Twitter profiles to produce the graph below.

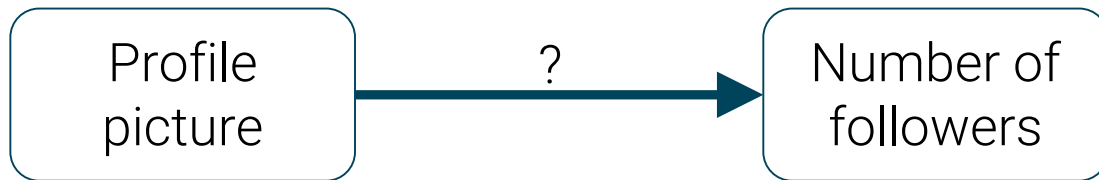


Source: [HubSpot](#)

It doesn't take much, but if you want to get followers on Twitter, it's a good idea to upload a picture of yourself.

Question 2

Does setting a profile picture **lead to** an increase in number of followers?

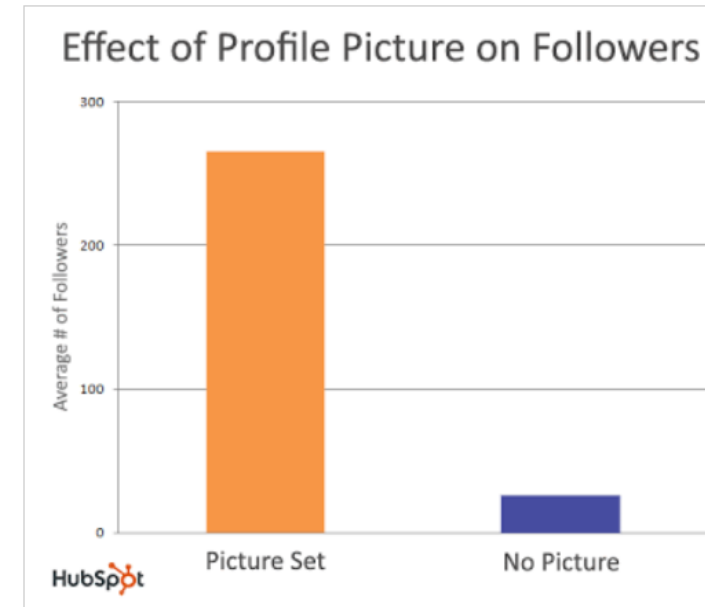


WEDNESDAY, APRIL 7, 2010

Twitter Accounts with a Profile Picture Have 10 Times More Followers Than Those Without

When publishing Twitter statistics, one of the most common questions I'm asked is about the effect of setting a profile picture in your Twitter account.

I dug into Twitter Grader data and analyzed nearly 9 million Twitter profiles to produce the graph below.

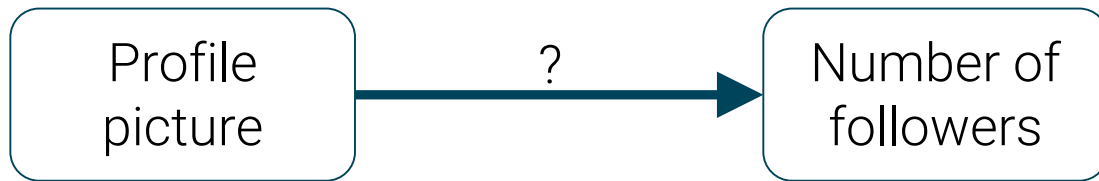


Source: [HubSpot](#)

It doesn't take much, but if you want to get followers on Twitter, it's a good idea to upload a picture of yourself.

Question 2

Does setting a profile picture **lead to** an increase in number of followers?



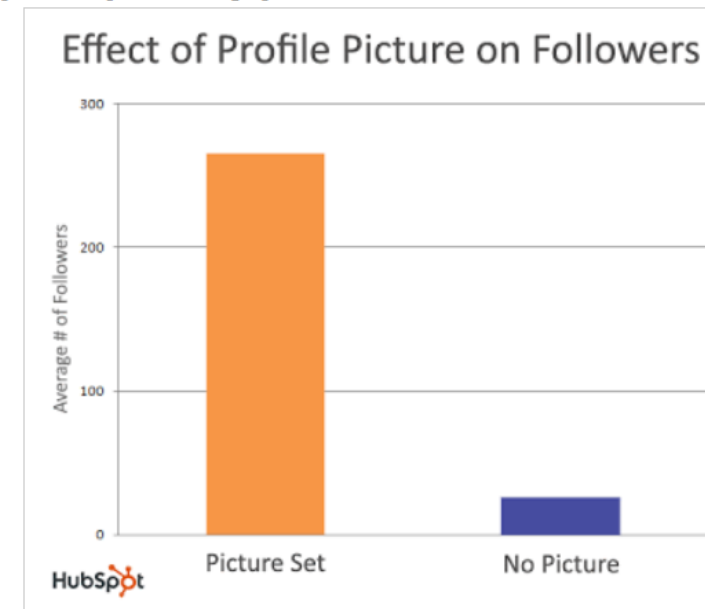
Causality
(Often much harder to answer)

WEDNESDAY, APRIL 7, 2010

Twitter Accounts with a Profile Picture Have 10 Times More Followers Than Those Without

When publishing Twitter statistics, one of the most common questions I'm asked is about the effect of setting a profile picture in your Twitter account.

I dug into Twitter Grader data and analyzed nearly 9 million Twitter profiles to produce the graph below.



Source: [HubSpot](#)

It doesn't take much, but if you want to get followers on Twitter, it's a good idea to upload a picture of yourself.

Correlations or causations?

מחקר: תרגול יוגה עשוי למנוע דמנציה

מחקר חוצה יבשות מצא כי עובי קליפת המוח של נשים מבוגרות המתרגלות יוגה עבה יותר מאצל נשים בנות אותו גיל שלא תרגלו כלל. "התוצאות", אומרים החוקרים "עשויות להצביע על שימור וחיזוק המוח במיוחד באזור האחראי על הזיכרון"



דנה פלג פורסם: 18.12.17, 11:00

בשיתוף מגדלי הים התיכון

רוצים לחזק את המוח - תתחילו כבר עכשיו לתרגל יוגה. חוקרים מברזיל וארצות הברית שיתפו פעולה במחקר חדש שמצא כי תרגול יוגה ממושר, ליתר דיוק במשך שמונה שנים, גורם לעיבוי וחזוק אזורי במוח המעורבים בזיכרון, תשומת לב וחלוקת זמן.

מחקר: למתרגלות יוגה מוח חזק

החוקרים בדקו כ-50 נשים בנות 60 ומעלה אשר תרגלו יוגה לפחות במשך עשור מול קבוצה של נשים בנות אותו הגיל אשר מקפידות על פעילות גופנית אחרת במשך תקופת זמן ממושכת. המשתתפות במחקר ביצעו סריקת MRI אשר בה ניתן היה לראות את השינוי בקליפת המוח.

"מצאנו כי בהשוואה של נשים מבוגרות, בריאות ופעילות אל מול נשים שתרגלו יוגה, בנות אותו הגיל, עובי קליפת המוח של מתרגלות היוגה היה גדול יותר", מספרת אליסה קוזאסה עורכת המחקר מברזיל. "הקליפה המעובה נמצאה בחלק המוח המכונה קדם-פרונטאלית מצד שמאל. אזור זה ידוע כקשור לביצוע משימות, היכולת לתשומת לב וזיכרון".

מחקר: ילדים קראו קומיקס ולמדו על מחלתם, מה שמעלה את סיכויים להבריא

חוקרים ישראלים מצאו כי ילדים שחולים בדלקת מפרקים כרונית אשר קראו עליה בספרון קומיקס שיפרו משמעותית את הידע שלהם אודותיה, באופן שעשוי להיטיב עם היענותם לטיפול

חדשות בריאות

מחקר: מזון מעובד מגביר את הסיכוי לסרטן

מחקר חדש של חוקרים מאוניברסיטת סורבון, מצא כי אנשים שצורכים אחוז גבוה יותר של מזון מעובד, מעלים את סיכוייהם ב-12% לחלות בסרטן • המסוכנים: חטיפים, מאכלי אינסטנט, נגיסי בשר ולחמניות

אסף גולן //

פורסם ב: 15.02.2018 15:47 / 1

מחקר של חוקרים צרפתיים מאוניברסיטת סורבון שהתבצע על 105 אלף אנשים, מצא קשר משמעותי בין מזונות שעברו עיבוד נרחב כגון עוגות מייצור המוני, נגיסי בשר שונים, ותערובות אינסטנט שונות כמו מרק, למחלת הסרטן.

לאחר שהתברר כי עודף משקל, עישון ואף אכילת בשר אדום מעלים את הסיכון לסרטן, רצו החוקרים

מחקר ענק: קיצור קיבה עשוי למנוע סרטן

לבדוק האם גם מזון מאוד העמידה, ובחן את התפר ששיעור המזון מסוג זה ע

חוקרים אמריקנים בדקו עשרות אלפי אנשים שסובלים מהשמנת יתר, חלקם עברו ניתוחים בריאטריים, וגילו כי הניתוח הפחית את סיכויי התחלואה בסרטן בכשליש. החוקרים: "הממצאים נכונים בעיקר עבור נשים - תוצאות משמעותיות מאוד נרשמו בסרטן השד ובסרטן הרחם"



ד"ר איתי גל פורסם: 08.10.17, 09:52

מנותחים בריאטריים לא רק מפחיתים במשקל, אלא גם בסיכוי לחלות בסרטן: מחקר רחב היקף, שכלל אלפי נבדקים, מצא כי ניתוחי קיצור קיבה הורידו כבשליש את סיכון התחלואה בסרטן אצל אנשים שסובלים ממשקל יתר משמעותי. הממצאים החדשים פורסמו בגליון האחרון של כתב העת Annals of Surgery.

Correlation vs. Causation

Two variables are **correlated** when knowing the value of one gives us some information about the (more likely) value of the other

Two states are **causally related** when one state influences the other via a cause-and-effect process

Correlation vs. Causation

Two variables are **correlated** when knowing the value of one gives us some information about the (more likely) value of the other

- Linear correlation: the degree to which the variables are linearly related

Two states are **causally related** when one state influences the other via a cause-and-effect process

Correlation vs. Causation

Two variables are **correlated** when knowing the value of one gives us some information about the (more likely) value of the other

- Linear correlation: the degree to which the variables are linearly related

Two states are **causally related** when one state influences the other via a cause-and-effect process

- When are we interested in each?

Correlation vs. Causation

Two variables are **correlated** when knowing the value of one gives us some information about the (more likely) value of the other

- Linear correlation: the degree to which the variables are linearly related

Two states are **causally related** when one state influences the other via a cause-and-effect process

- When are we interested in each?
- Correlation/association does not imply causation!

Correlation vs. Causation

Two variables are **correlated** when knowing the value of one gives us some information about the (more likely) value of the other

- Linear correlation: the degree to which the variables are linearly related

Two states are **causally related** when one state influences the other via a cause-and-effect process

- When are we interested in each?
- **Correlation/association does not imply causation!**
 - Causation also does not necessarily imply correlation, but this is beyond our scope
 - In most real world scenarios, when two things are causally related, they are also associated (not necessarily linearly!)



Hannah Fry ✓
@FryRsquared



Every single person who confuses correlation and causation ends up dying.



Greg Hands ✓ @GregHands · Feb 12

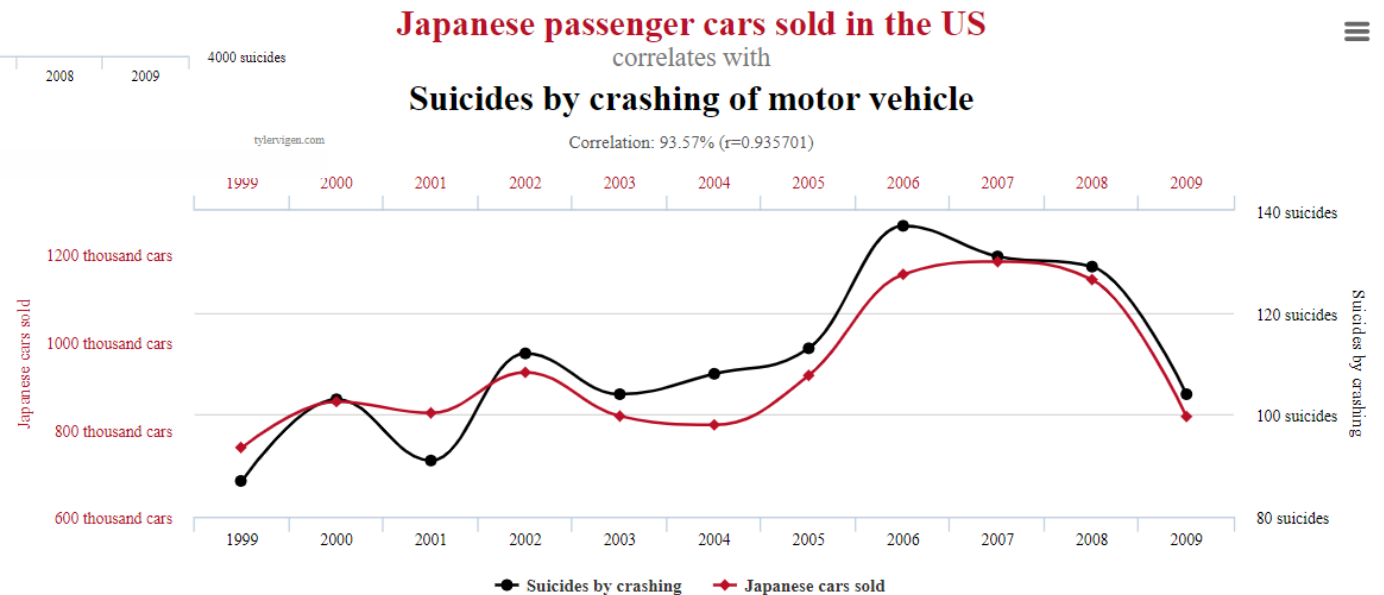
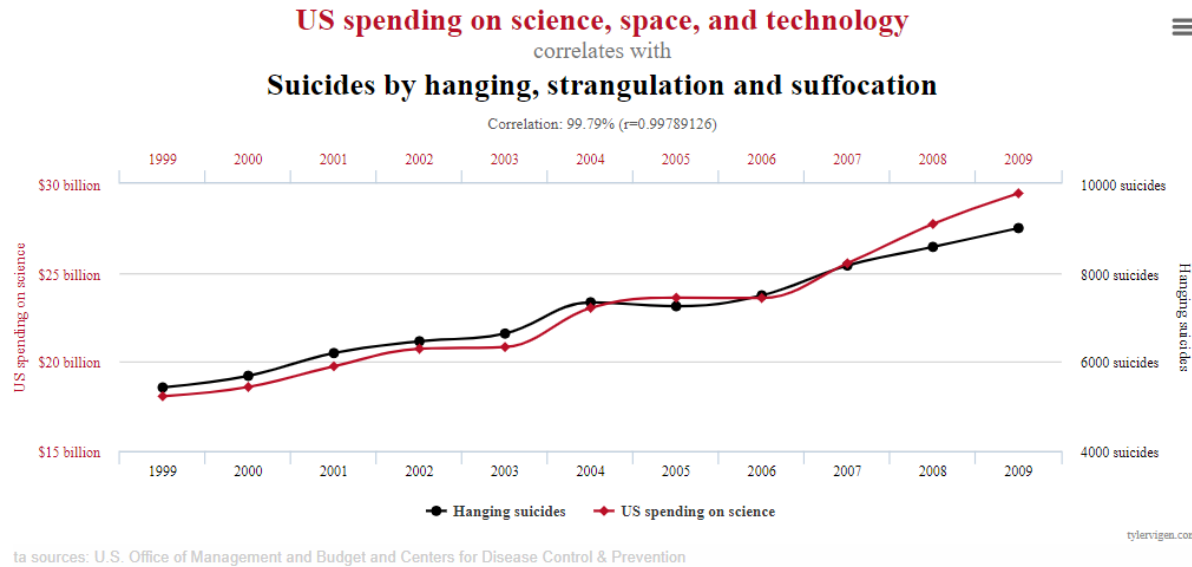
People more likely to be victims of crime in Labour-run areas, new study shows.

telegraph.co.uk/news/2023/02/1...

1:15 AM · Feb 14, 2023 · **3.9M** Views

5,087 Retweets **256** Quotes **43.3K** Likes **656** Bookmarks

Beware of spurious correlations



<http://>ata sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

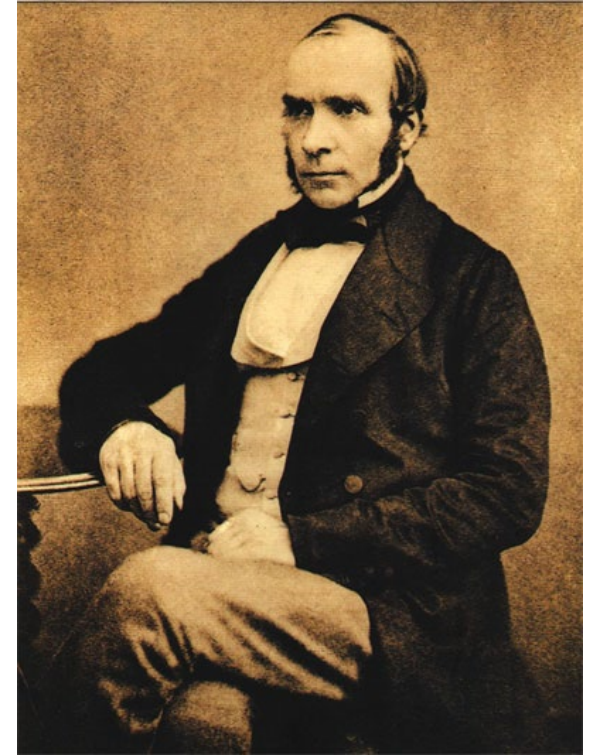
Association

Example: London, 19th century



Jon Snow

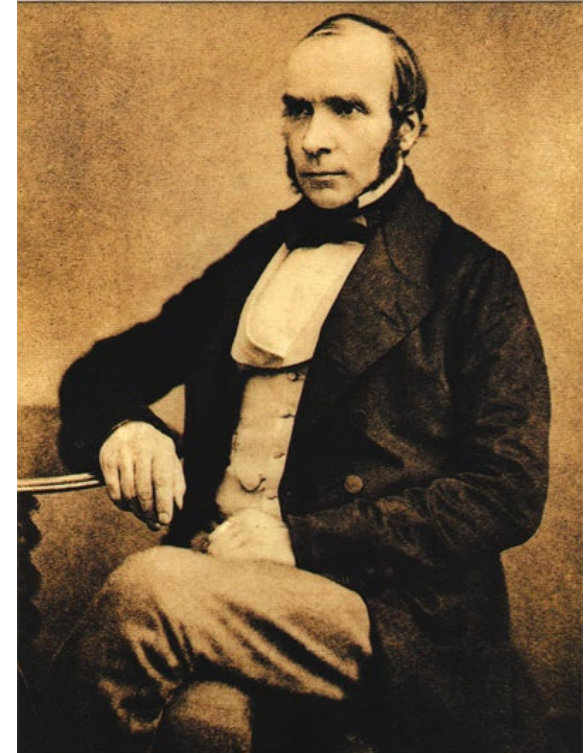
Example: London, 19th century



John Snow,
1813-1858

Example: London, 19th century

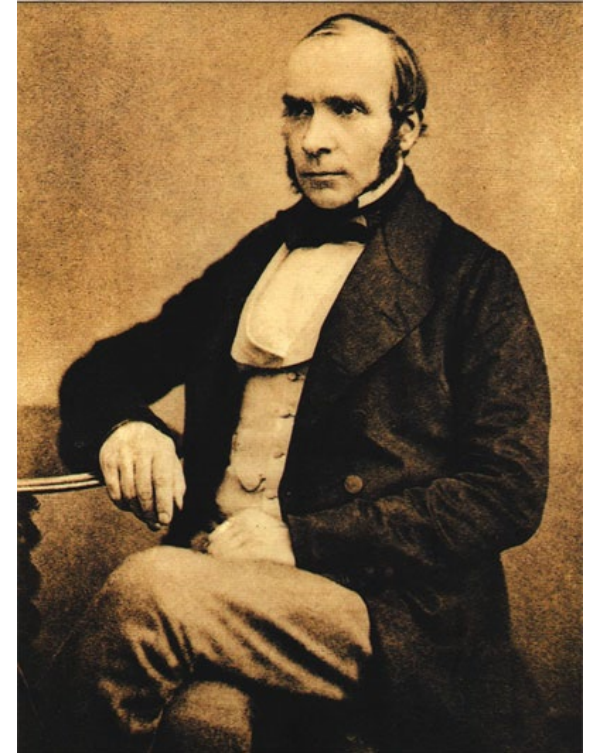
- Cholera outbreaks
 - *"These problems are, and will probably ever remain, among the inscrutable secrets of nature. They belong to a class of questions radically inaccessible to the human intelligence."*
—The Times of London, September 1849, on how cholera is contracted and spread
- "Miasmas" = invisible particles in the air
 - Cause bad smells
 - Believed to cause cholera



John Snow,
1813-1858

Example: London, 19th century

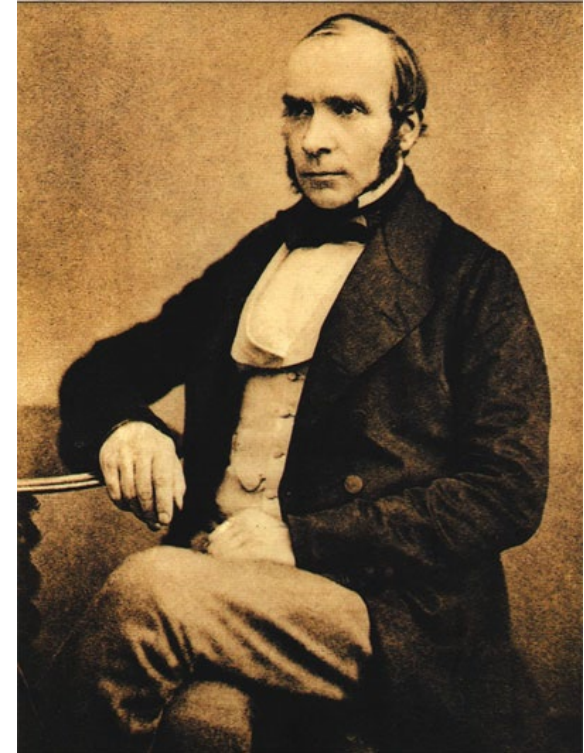
- Cholera outbreaks
 - *"These problems are, and will probably ever remain, among the inscrutable secrets of nature. They belong to a class of questions radically inaccessible to the human intelligence."*
—The Times of London, September 1849, on how cholera is contracted and spread
- "Miasmas" = invisible particles in the air
 - Cause bad smells
 - Believed to cause cholera
- Snow thought otherwise
 - Next door neighbors: same treatment (air), different outcomes
 - Contaminated food/water



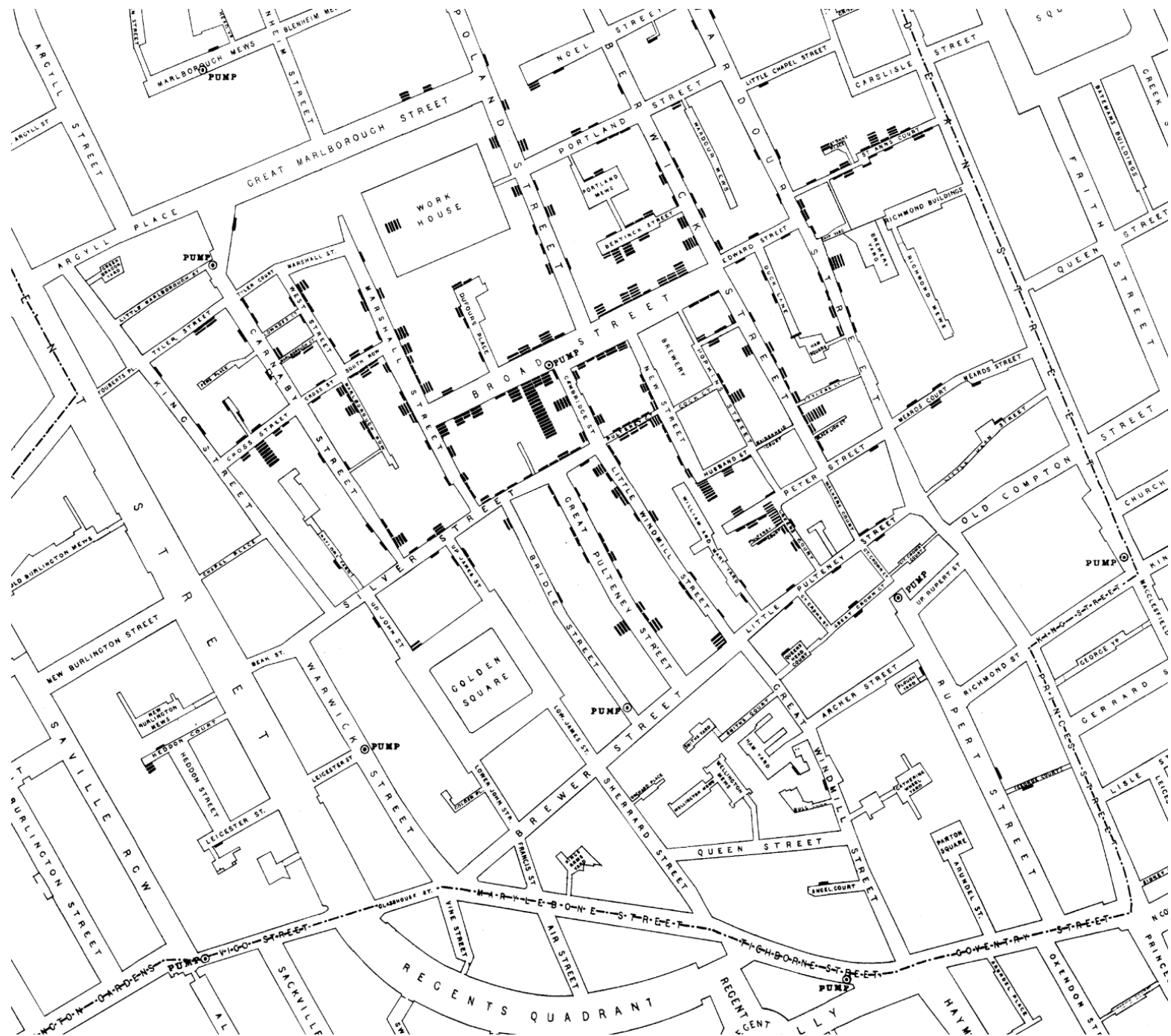
John Snow,
1813-1858

Example: London, 19th century

- Cholera outbreaks
 - *"These problems are, and will probably ever remain, among the inscrutable secrets of nature. They belong to a class of questions radically inaccessible to the human intelligence."*
—The Times of London, September 1849, on how cholera is contracted and spread
- "Miasmas" = invisible particles in the air
 - Cause bad smells
 - Believed to cause cholera
- Snow thought otherwise
 - Next door neighbors: same treatment (air), different outcomes
 - Contaminated food/water
- 1854: Soho district outbreak



John Snow,
1813-1858








Two questions

- Did Snow find a causal relation between contaminated water and the spread of cholera?
- Did Snow find a causal relation between the deaths from cholera in the Soho district 1854 outbreak and the Broad St. pump?

John Snow



John Snow

3.7 ★★★★★ 193 reviews

Pub

Directions


SAVE


NEARBY


SEND TO YOUR PHONE


SHARE


Dark-wood saloon bar serving Yorkshire ales, named after doctor who traced London cholera outbreak. - Google


 39 Broadwick St, Carnaby, London W1F 9QJ, UK

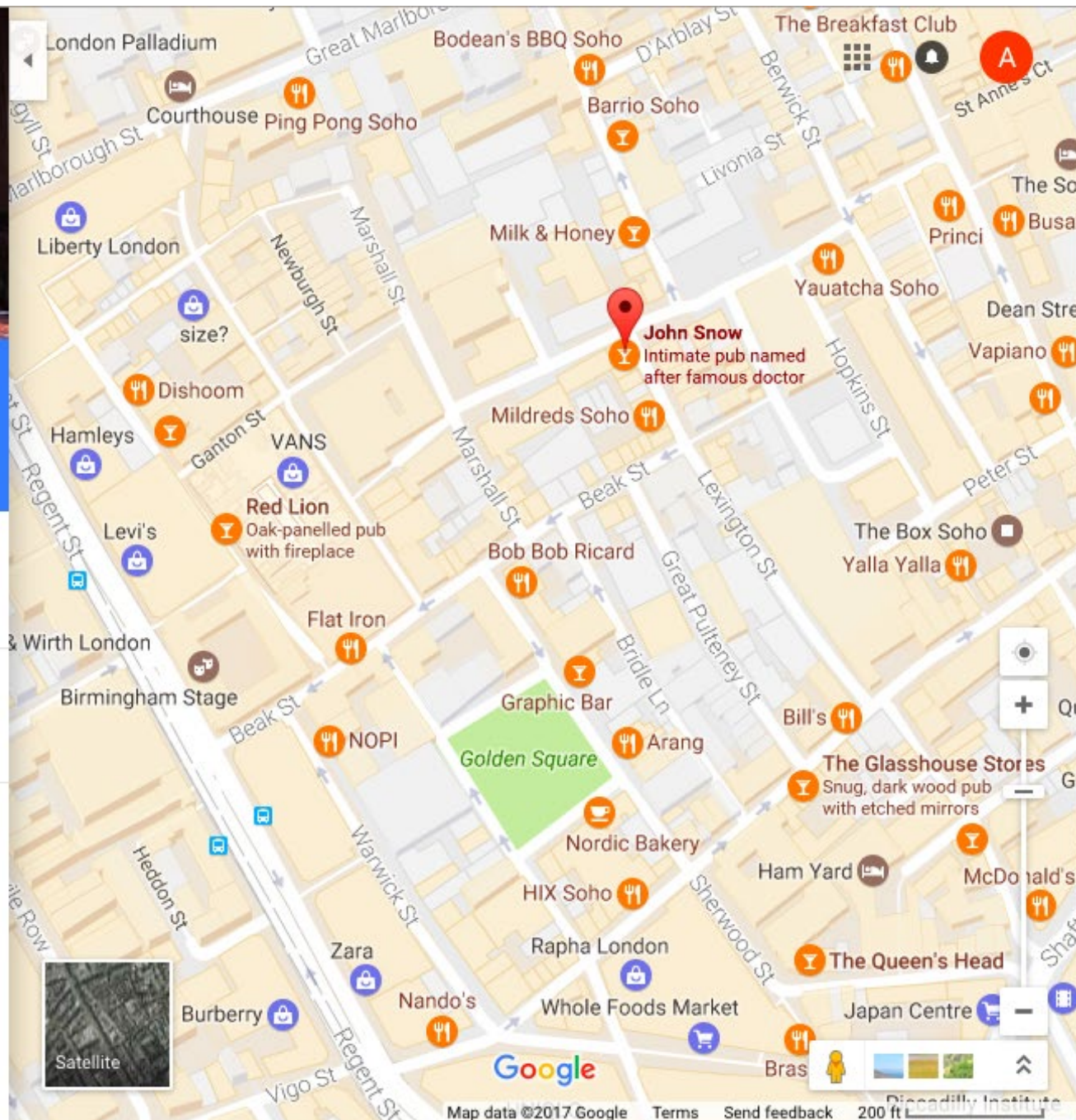
 +44 20 7437 1344

 Closed. Opens at 12:00 PM ▾

 Claim this business

 Suggest an edit

 Add a label



A Google Map of Soho, London, centered on the John Snow pub. The map shows a grid of streets including Great Marlborough St, Newburgh St, Marshall St, Beak St, Lexington St, and Regent St. Numerous other businesses and landmarks are labeled, such as the London Palladium, Courthouse Ping Pong Soho, Barrio Soho, Milk & Honey, Yauatcha Soho, The Breakfast Club, The Soho Bus, Princi, Vapiano, The Box Soho, Yalla Yalla, The Glasshouse Stores, McDonald's, The Queen's Head, Japan Centre, Whole Foods Market, Nando's, Zara, Burberry, HIX Soho, Nordic Bakery, Arang, Bill's, Flat Iron, Red Lion, Dishoom, Hamleys, Levi's, and The Soho. A red pin marks the location of John Snow, with a text box stating 'Intimate pub named after famous doctor'. A green square highlights Golden Square. The map includes a satellite view inset in the bottom left corner and a scale bar at the bottom right.

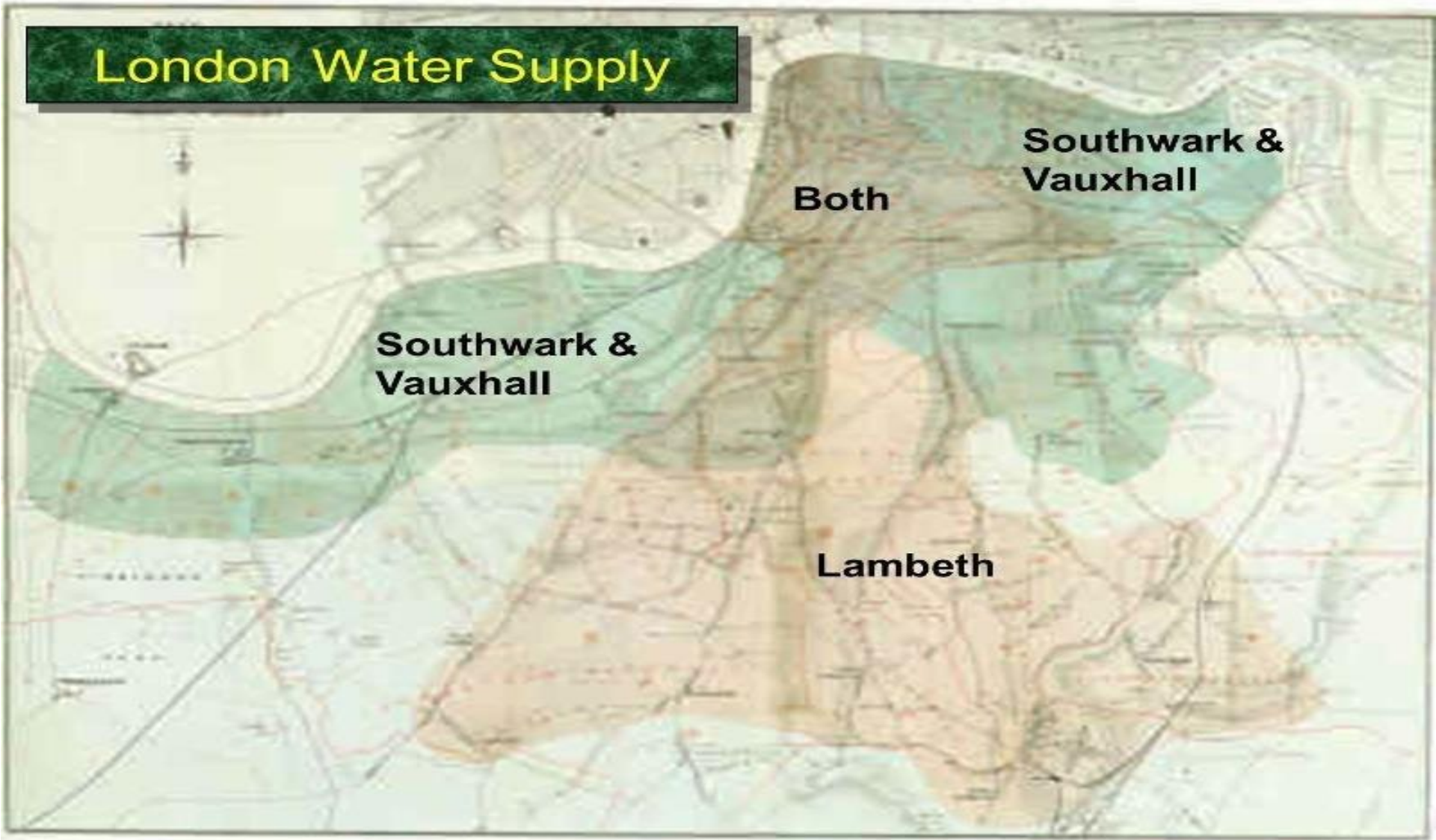


Causality

Comparison

- Treatment group
 - Gets treatment
 - Control group
 - Does not get treatment
-
- Outcomes differ -> association between treatment and outcome

London Water Supply



Snow's Grand Experiment

"... there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded ..."

- The two groups were **similar except for the treatment**.

Snow's table

Supply Area	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Key to establishing causality

If the treatment and control groups are **similar apart from the treatment**, then differences between the outcomes in the two groups can be ascribed to the treatment.

Trouble

If the treatment and control groups have **systematic differences other than the treatment**, then it might be difficult to identify causality.

When they lead researchers astray, they are called **confounding factors**.

Trouble

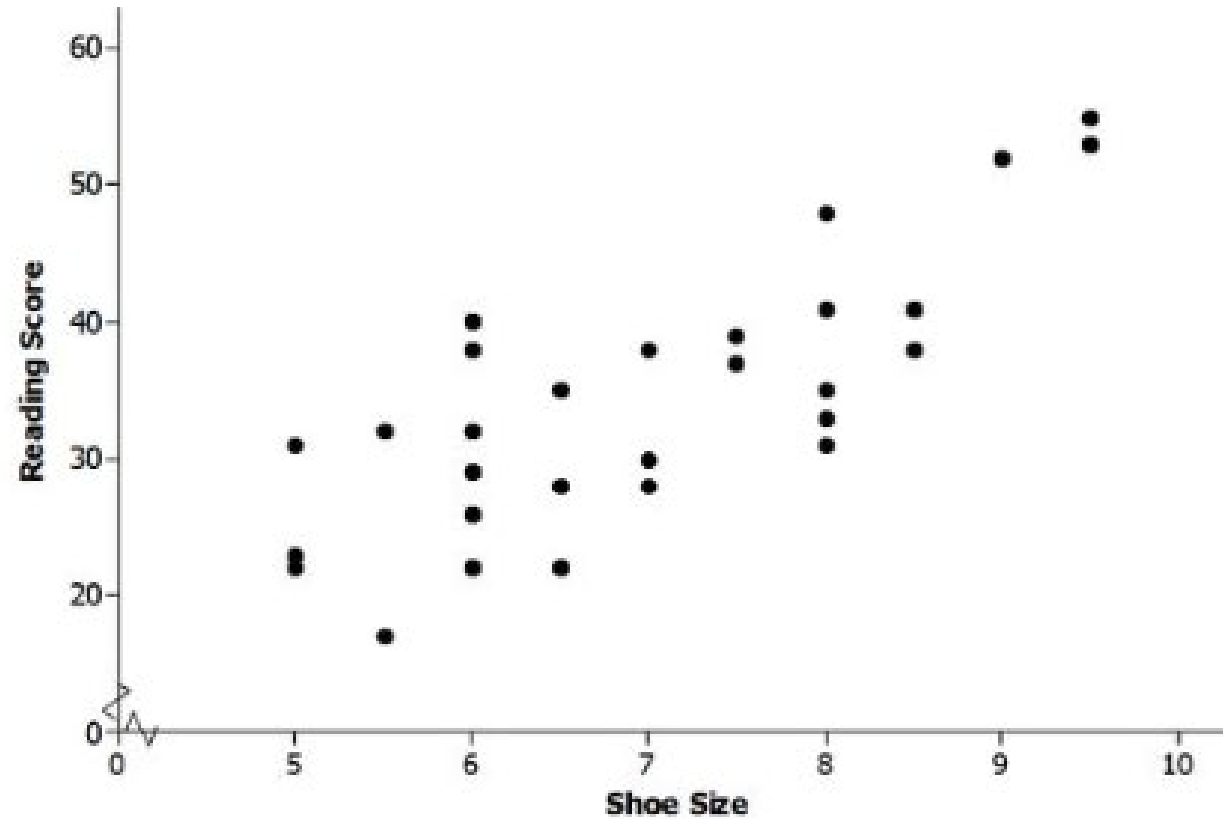
If the treatment and control groups have **systematic differences other than the treatment**, then it might be difficult to identify causality.

When they lead researchers astray, they are called **confounding factors**.

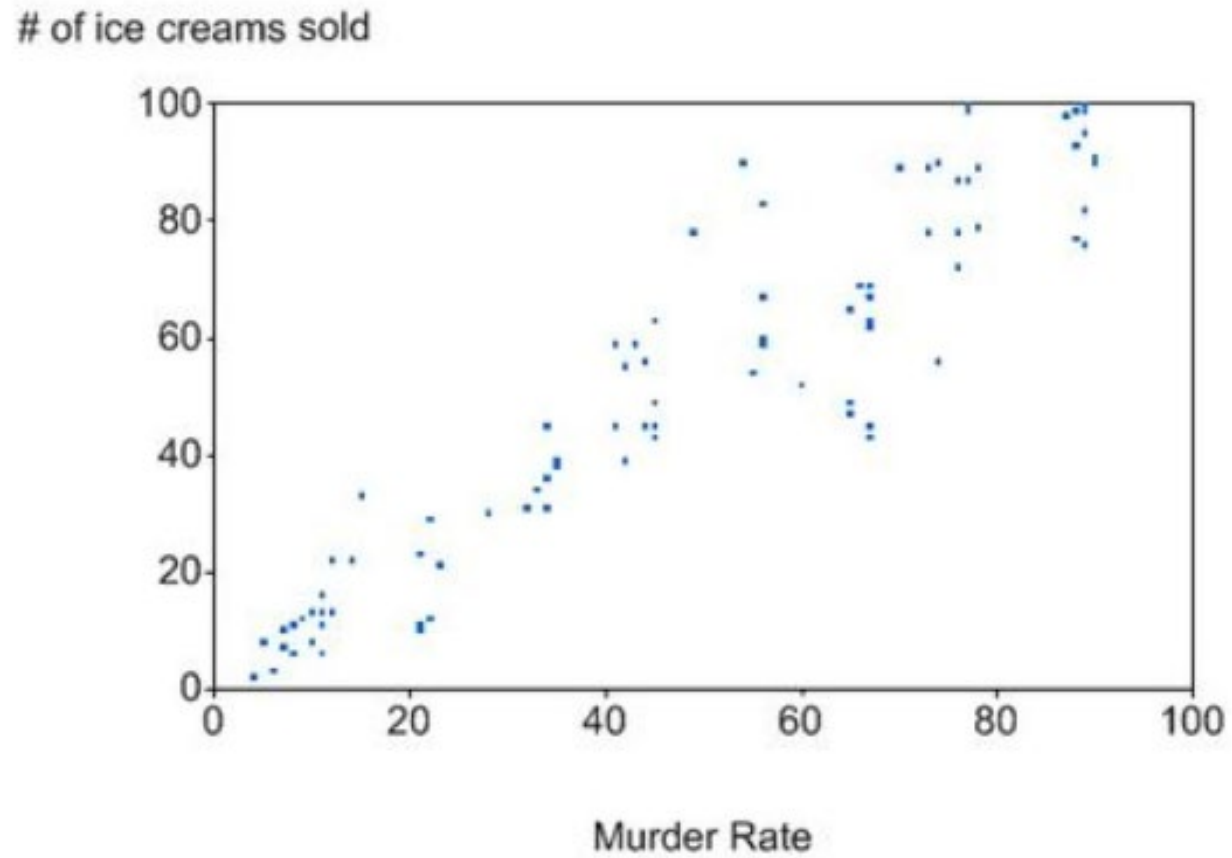
Such differences are often present in **observational studies**.

- Observational study: The researcher does not choose who gets the treatment and who doesn't
- Controlled experiment: The researcher determines the assignment to treatment and control

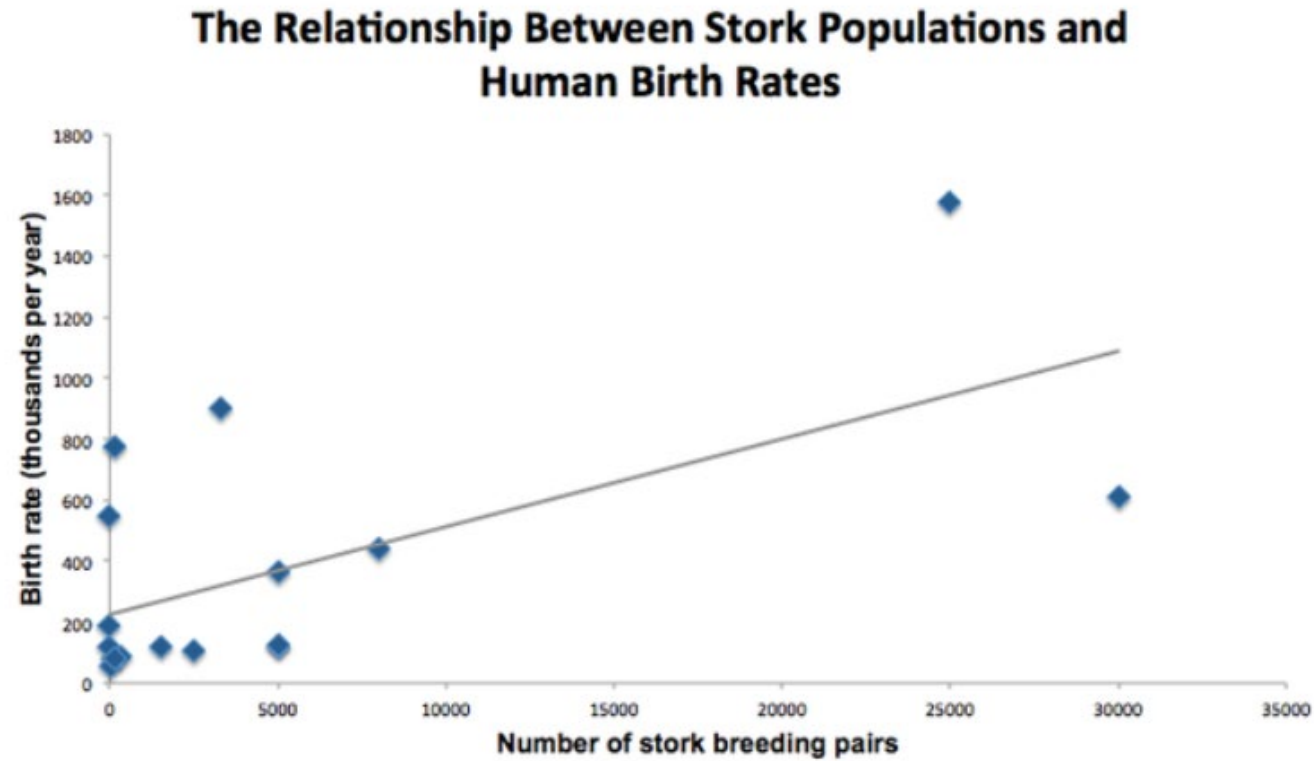
Confounders



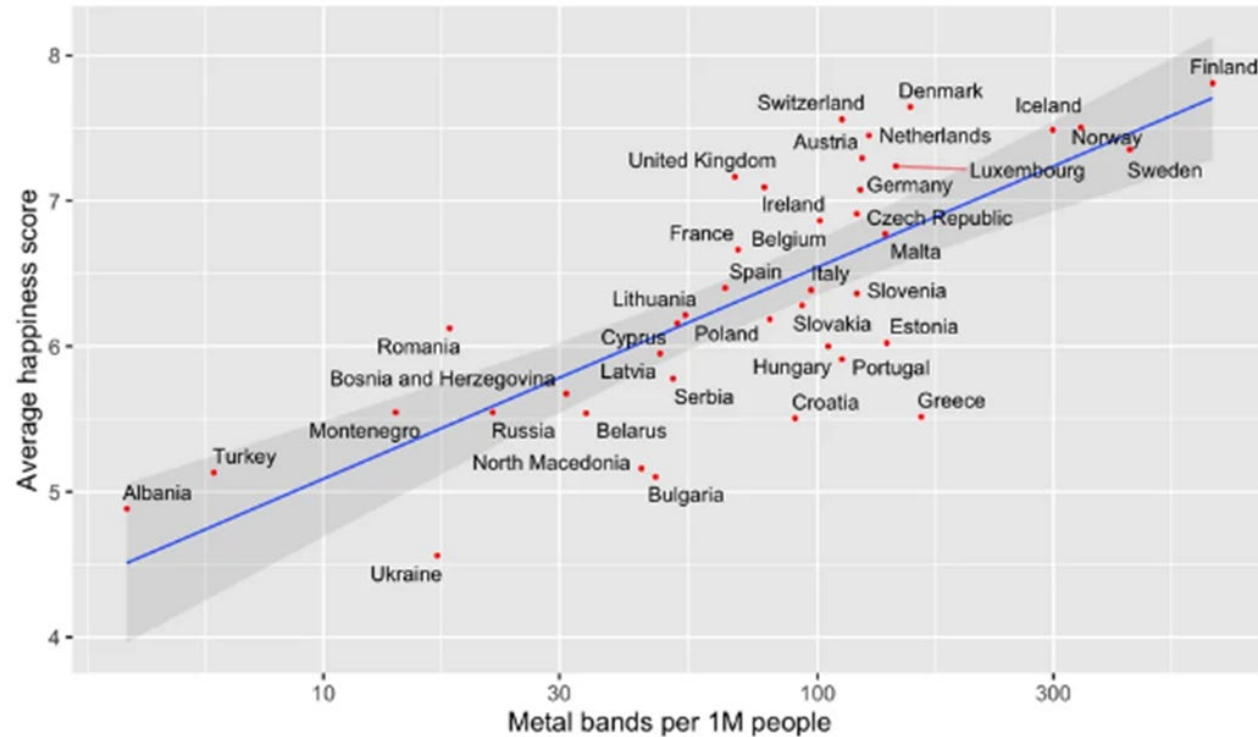
Confounders



Confounders



Confounders



Data sources: Enc. Metallum (2016), after Jakub Marian; World Happiness Report (2022). Chart by Piotr Migdal, p.migdal.pl, CC-BY.

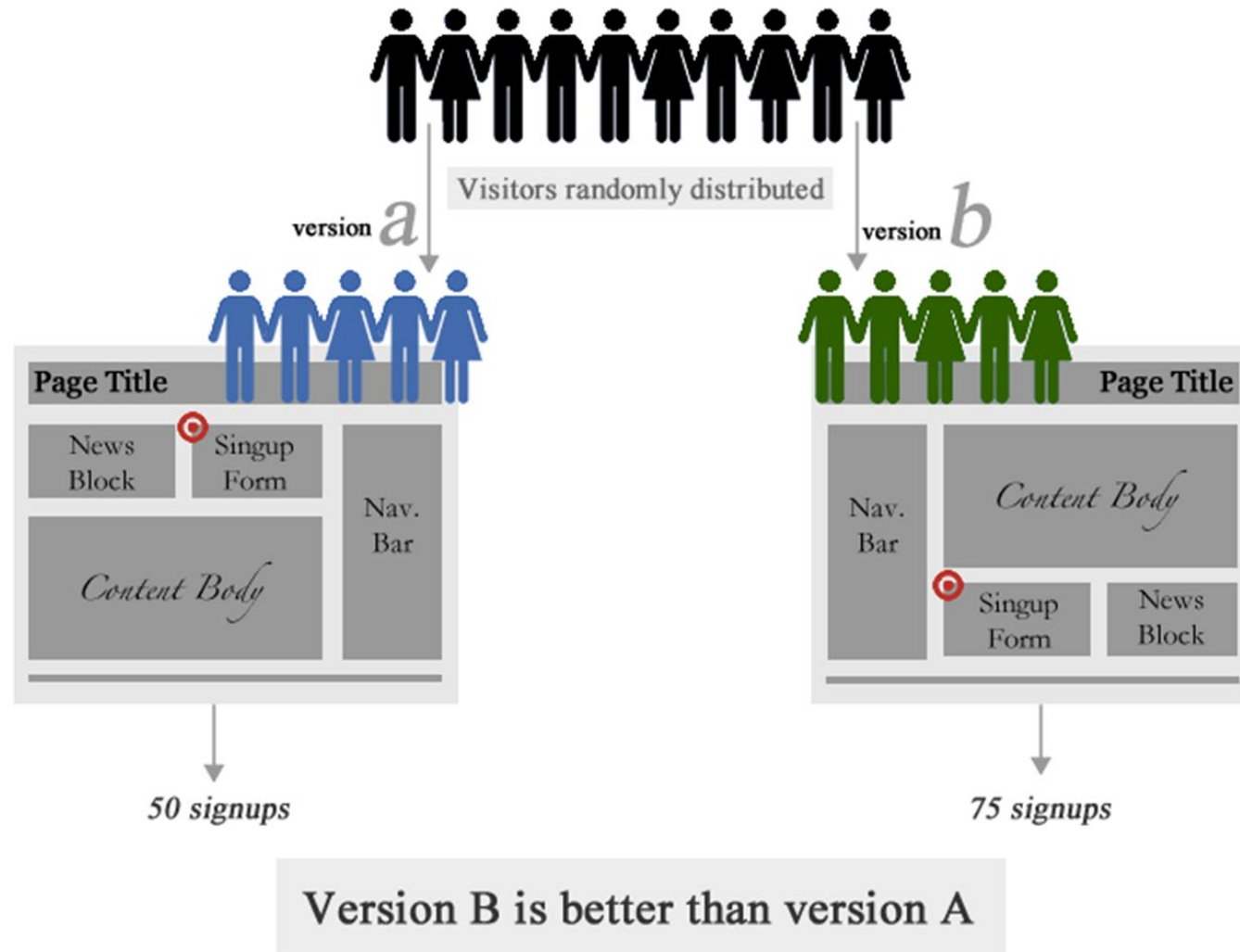
Solution: Randomize!

- If assignment of individuals to treatment and control is done **at random**, then the two groups are *likely* to be similar apart from the treatment
- And we can account – mathematically – for the chance of a systematic difference
 - We know the likelihood that the groups are dissimilar by chance

Solution: Randomize!

- If assignment of individuals to treatment and control is done **at random**, then the two groups are *likely* to be similar apart from the treatment
- And we can account – mathematically – for the chance of a systematic difference
 - We know the likelihood that the groups are dissimilar by chance
- Randomized Controlled Experiment
 - “Double blind”

Example: A/B Testing



Is temporal order enough?

- Say 2 variables:
 1. have an association (correlation)
 2. one variable always precedes the other

Does this mean the first causes the second?

Is temporal order enough?

- Say 2 variables:
 1. have an association (correlation)
 2. one variable always precedes the other

Does this mean the first causes the second?

- No!
- Example: Car baby seats and human births...



Donald J. Trump ✓

@realDonaldTrump

 Follow



Healthy young child goes to doctor, gets pumped with massive shot of many vaccines, doesn't feel good and changes - AUTISM. Many such cases!

RETWEETS

11,555

LIKES

9,293



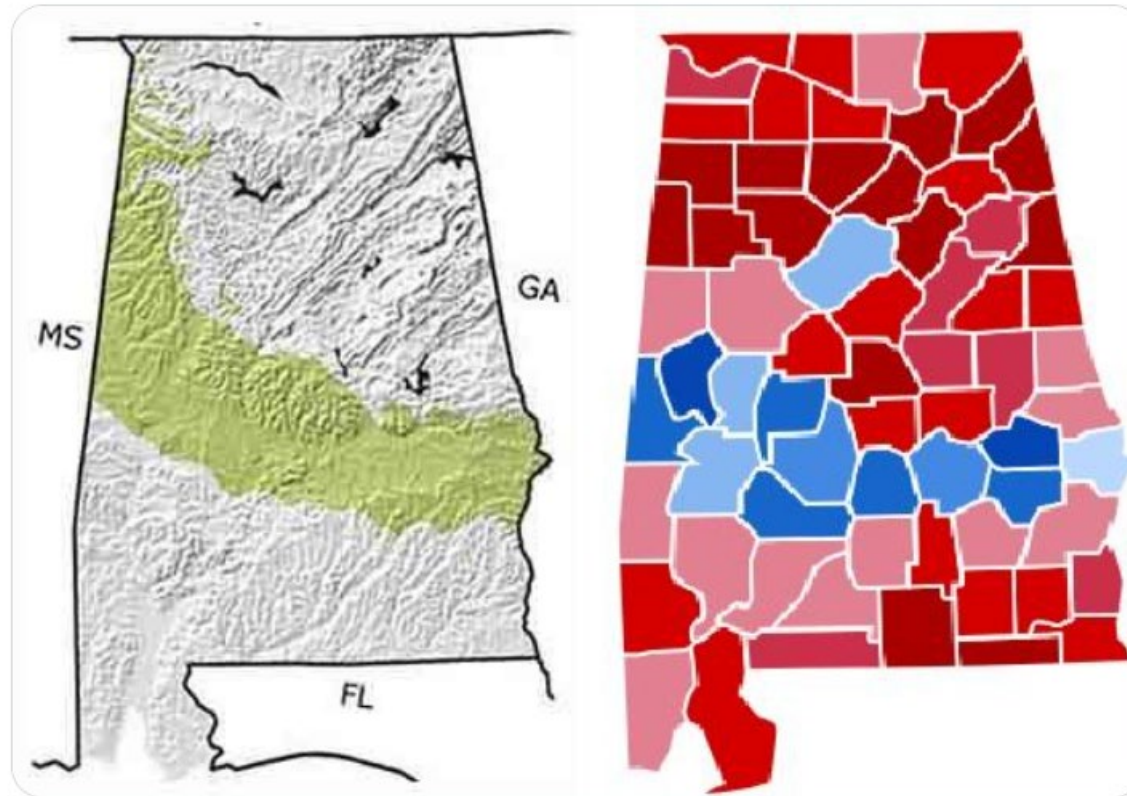
5:35 AM - 28 Mar 2014



Terrible Maps ✓
@TerribleMaps

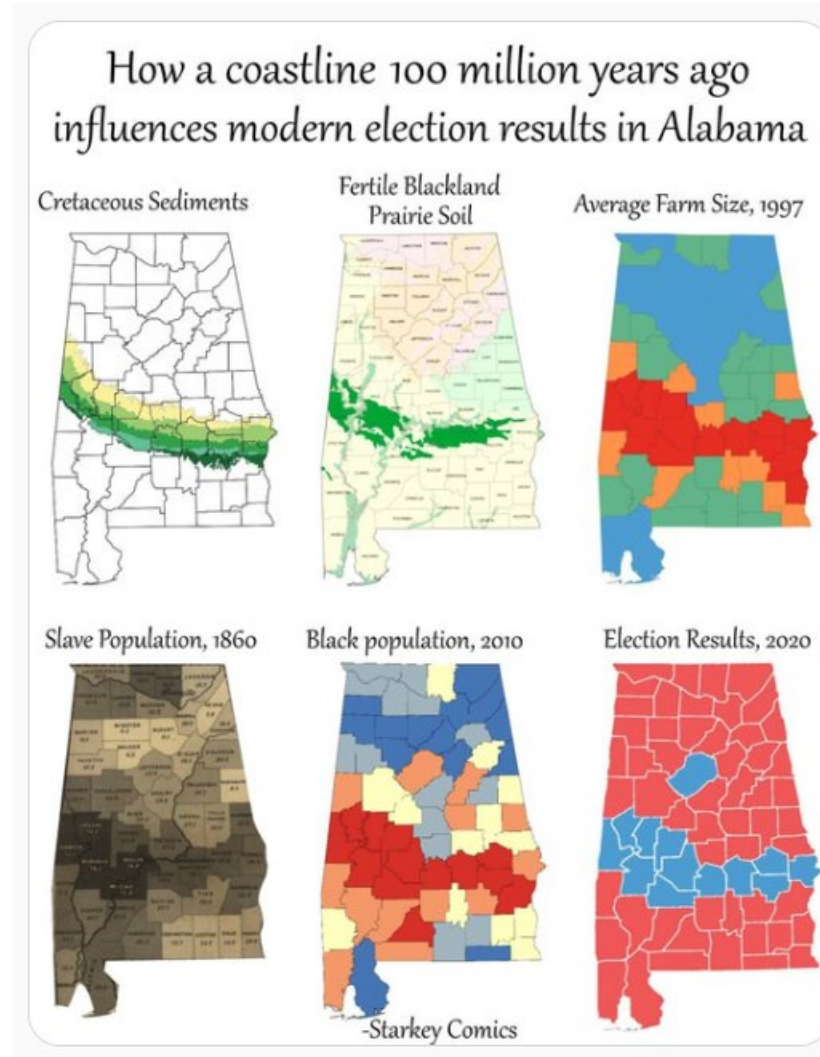


The cretaceous period (145 to 66 million years ago)
seashore in Alabama vs how Alabama voted in the
2016 election



8:51 PM · Dec 14, 2022

Sometimes, correlation may actually reflect causation...



Biases in data

Types of biases in data

- Measurement bias/error
 - Available data is wrong
 - Random noise vs. systematic noise
- Selection bias
 - Available data is correct, but does not represent well what we really want
 - *Sample distribution* different from *population distribution* (we'll get there)
- Data missing not at random
 - Kind of selection bias
 - Right-censored data
- ... (more later in this course)

CALENDAR OF MEANINGFUL DATES

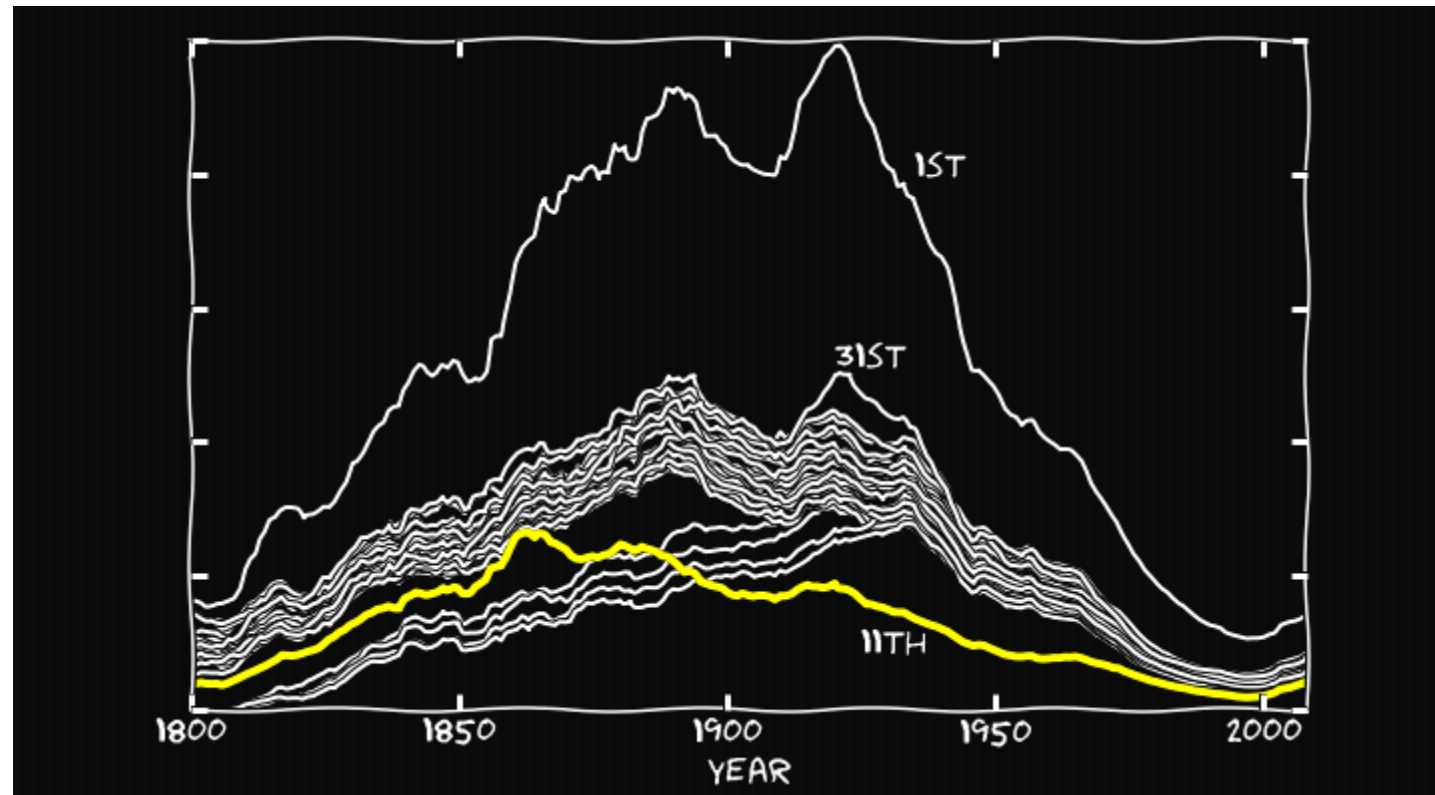
EACH DATE'S SIZE REPRESENTS HOW OFTEN IT IS REFERRED TO BY NAME
(E.G. "OCTOBER 17TH") IN ENGLISH-LANGUAGE BOOKS SINCE 2000
(SOURCE: GOOGLE NGRAMS CORPUS)



Source: xkcd

Google books data

- Ngrams
- <https://books.google.com/ngrams/>

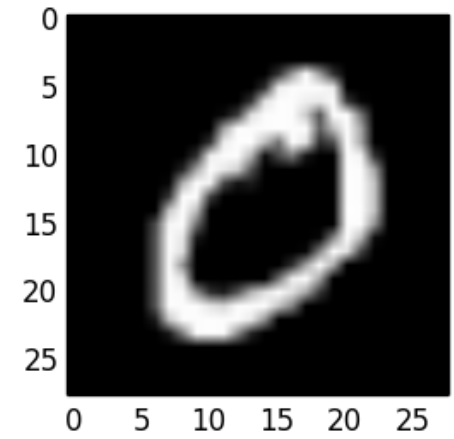
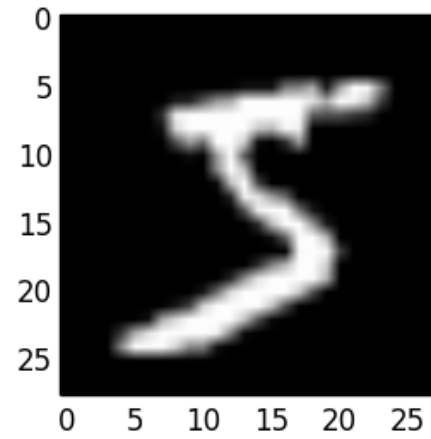


Systematic measurement errors

11th → llth Nth 1lth l1th ...

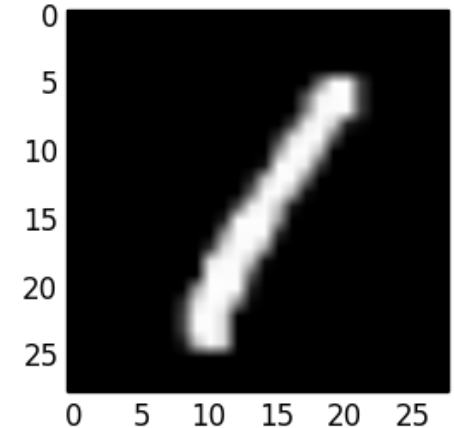
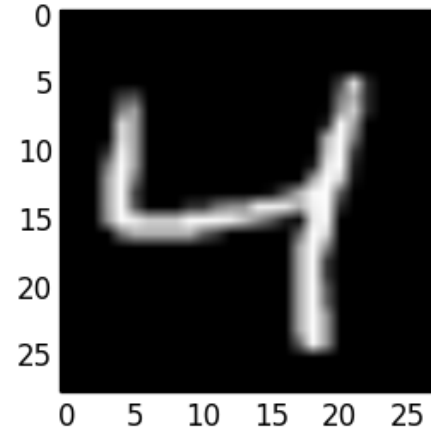
1850:

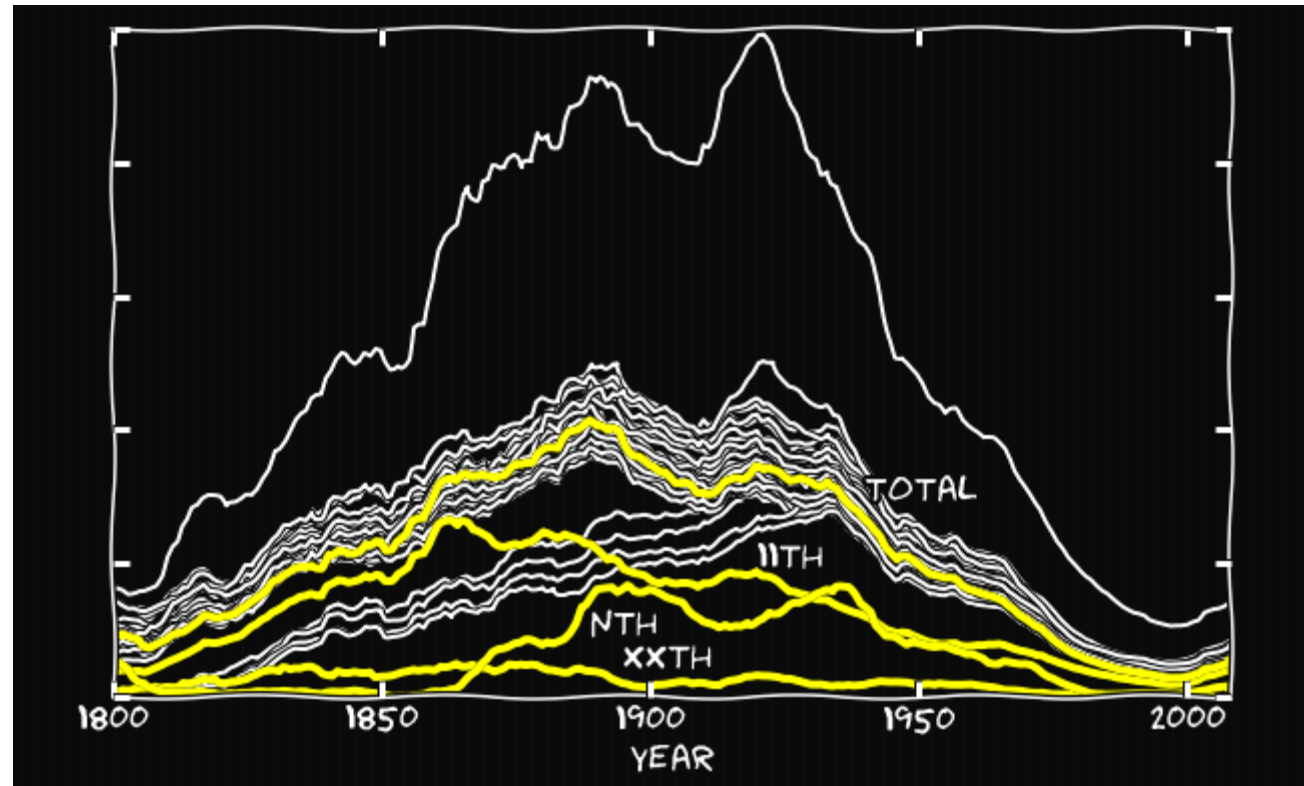
Council meeting January 11th, 1850.



1920:

Hotel on Sunday January 11th





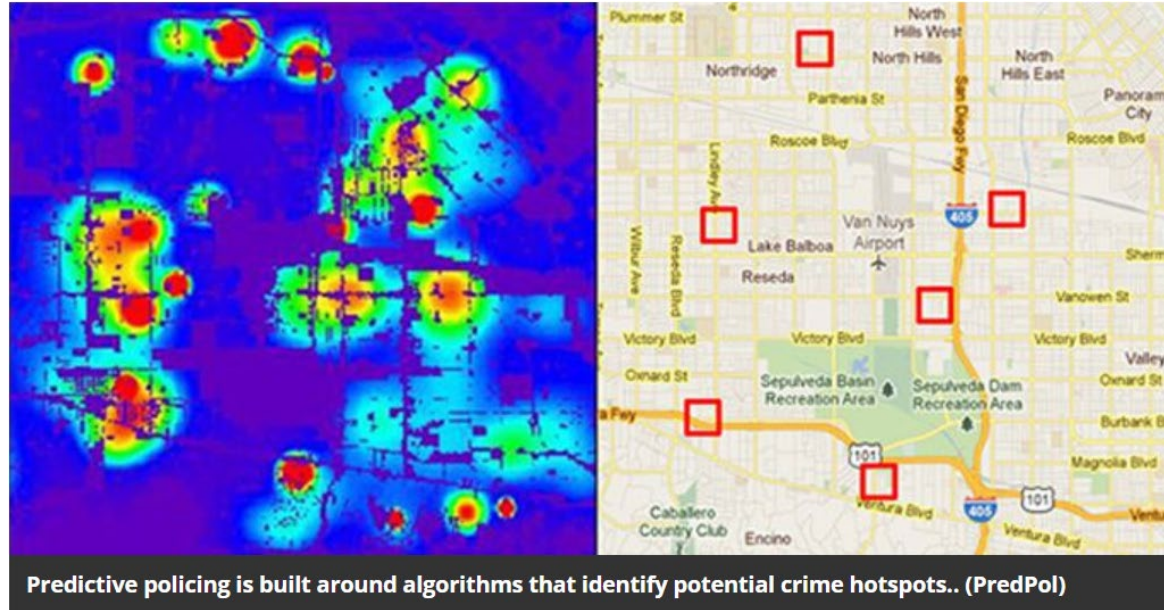
Boston potholes

Smartphones, Big Data Help Fix Boston's Potholes

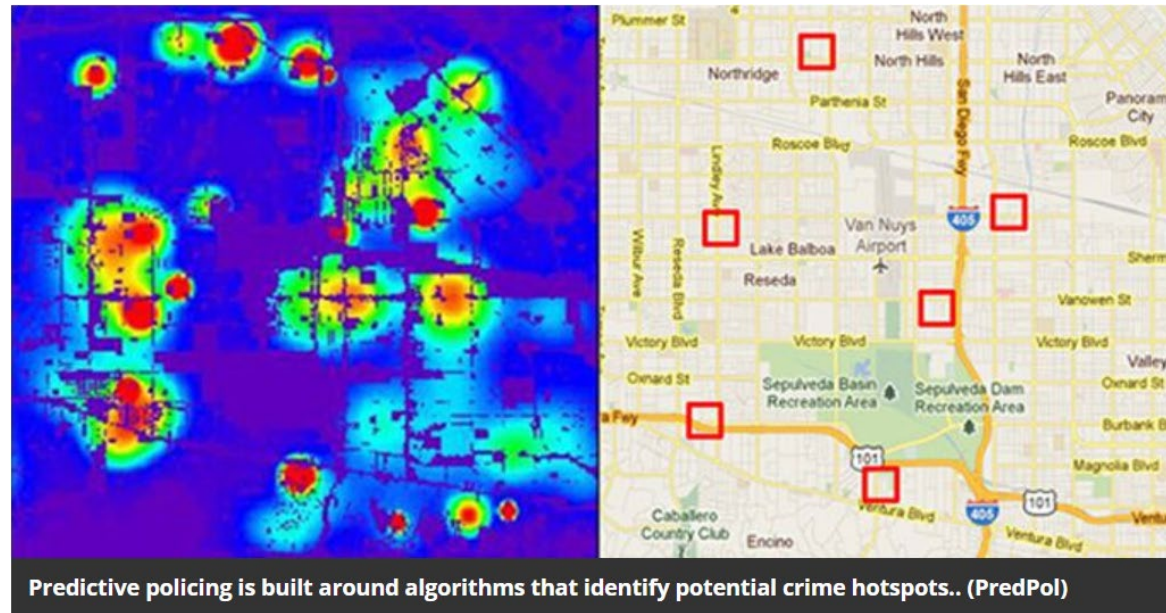
Boston has found an innovative way to find and fix street potholes: a free smartphone app, a crowdsourced competition, and lots of data from motorists.



Predictive policing



Predictive policing



- “Pile loads of resources into a certain area and it becomes a self-fulfilling prophecy, purely because there’s more policing going into that area, not necessarily because of discrimination on the part of officers”
- Feedback loops: predicting policing, not crime

Data missing not at random

- Parole decisions
- Fraud detection
- Navigation apps
- The “hot stove effect”



Bar Shem-Ur בר שם-אור
@Bar_ShemUr

...

קטע הזוי. שוק השוכרים מורכב לפי ההערכות משני מיליון איש בשנה שזה ציבור ענק של מצביעים. הלשכה המרכזית לסטטיסטיקה, כלומר המדינה, פרסמה בתקשורת שחלה עליה ממוצעת בשנה האחרונה של רק בין שלושה לשבעה אחוז (בדירות גדולות) בשכר הדירה.



בר שם-אור
@Bar_Shem

אלא שמניתוח עשרות אלפי מודעות מאתרי דירות שנעשתה עבור

דה מרקר יצא ששכר הדירה עלה מתחילת השנה בין 13 ל-22 אחוז! וממש לא רק בתל אביב, אלא בראש העין, אשדוד, רמלה, כולל עליה בפתח תקווה לבדה של שלושה עשר אחוזים!

ככה יוצא שהמדינה מורידה בסטטיסטיקה את גודל המחדל כנראה כדי להראות שהשד לא נורא כל כך בזמן שבשטח מרגישים היטב שהמצב על סף פיצוץ



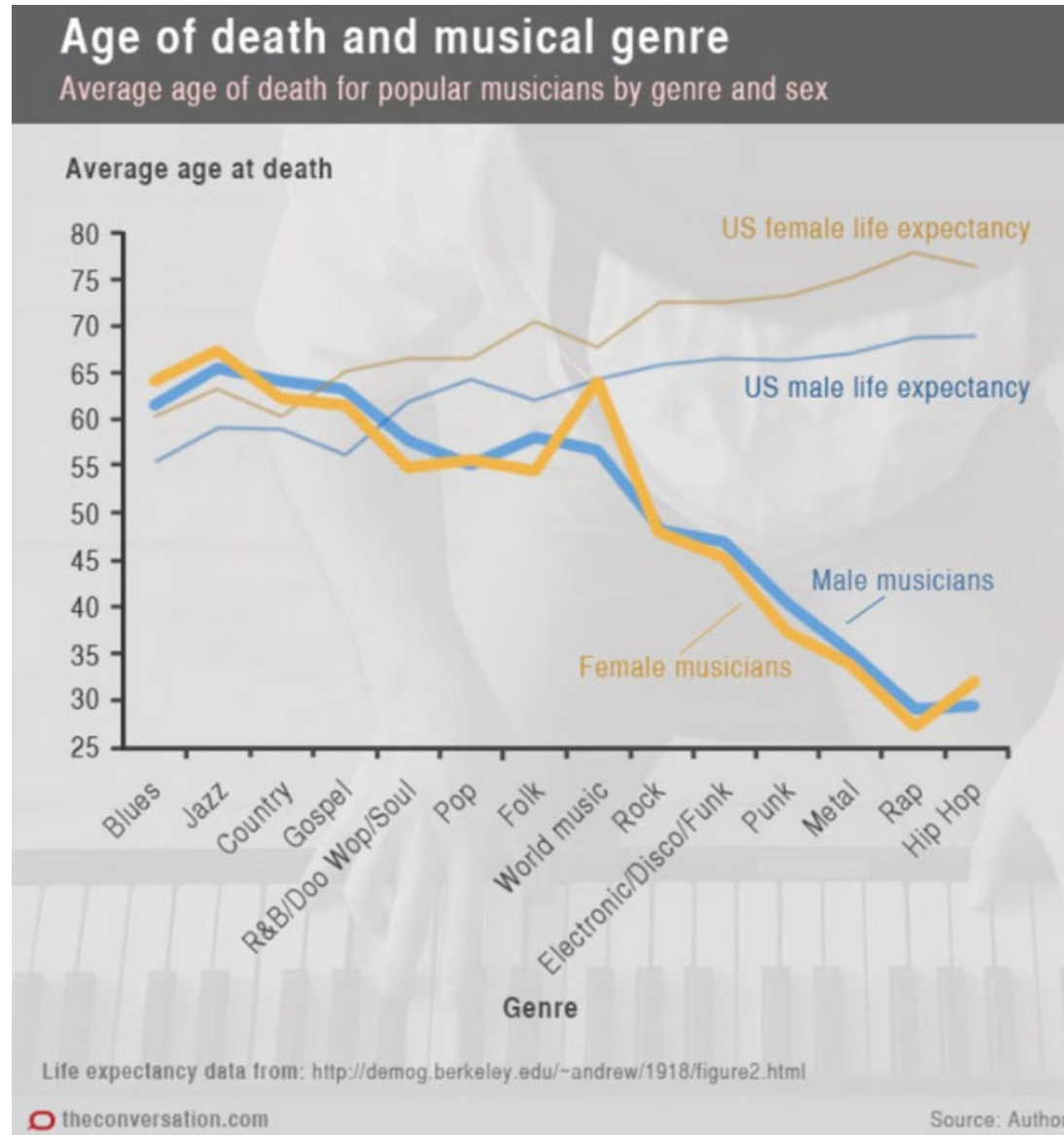
בר שם-אור
@Bar_ShemUr

בכל הארץ. הטיעון "אז אל תגורו בת"א" לא ממש רלוונטי אם שכר הדירה באזורי ביקוש בפ"ת וראשל"צ זינק לשמים, כשבאשדוד בשכונות מסוימות שכר הדירה זינק ב-18 אחוז לפי המחקר. עבור מי שעובד בפריפריה ומרוויח פחות מאשר במרכז זו עליה מורגשת ובלתי נתפסת. בקיצור, הלשכה המרכזית לעבודה בעיניים

[Translate Tweet](#)

9:45 PM · Jun 7, 2022

Musicians mortality



Right-censored data

Right-censoring: a data point is above a certain value, but it is unknown by how much

Right-censored data

Right-censoring: a data point is above a certain value, but it is unknown by how much

- Not all rap stars die young, but rap stars who already died, died young
 - What's the difference?

Right-censored data

Right-censoring: a data point is above a certain value, but it is unknown by how much

- Not all rap stars die young, but rap stars who already died, died young
 - What's the difference?
 - [P.s. apparently, they actually do]
- Look at the lines for life expectancy in the US...



Wordle No. 454 from Friday, Sept. 16

A bot studied almost 40,000 Wordles.

6.4 out of 6

The average number of turns
players took to solve the puzzle

What can we do?

- Collect more/better data
- Use more sophisticated methods
 - Future courses
- Highlight limitations

