CMPU 100 · Programming with Data

Relating Two Variables

Class 14

Remember that you have Assignment 5, from before break, due this week on the regular schedule.

Story time

When we introduced visualization at the end of Part 1 of the course, we talked about John Snow and the London cholera outbreak of 1854.

616 people died in the vicinity of Broad Street.

South side of Broad Street, c. 1875

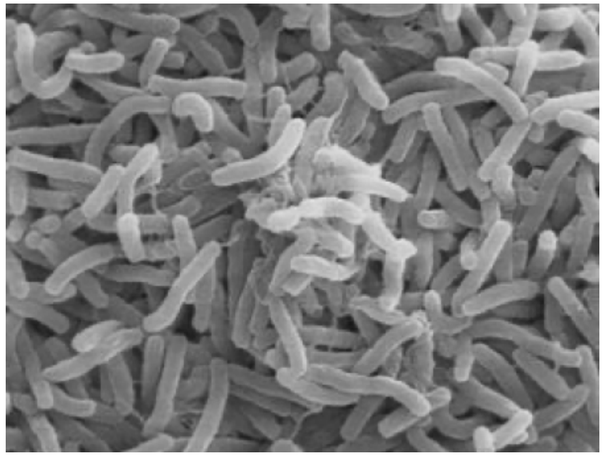
In the 1800s, there were two theories about cholera:

Miasma theory

Robert Seymour, 1831

Cholera is caused by *particles in the air* (*miasmata*), which arise from decomposing matter.

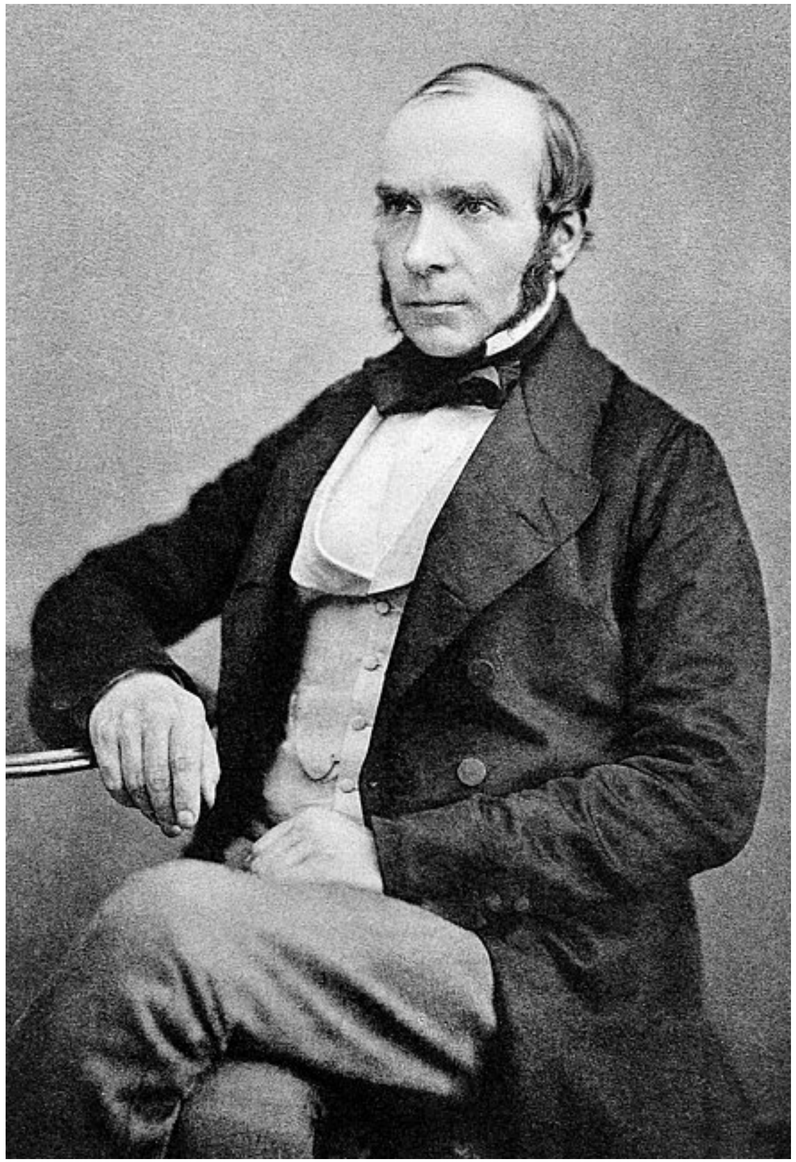
Germ theory

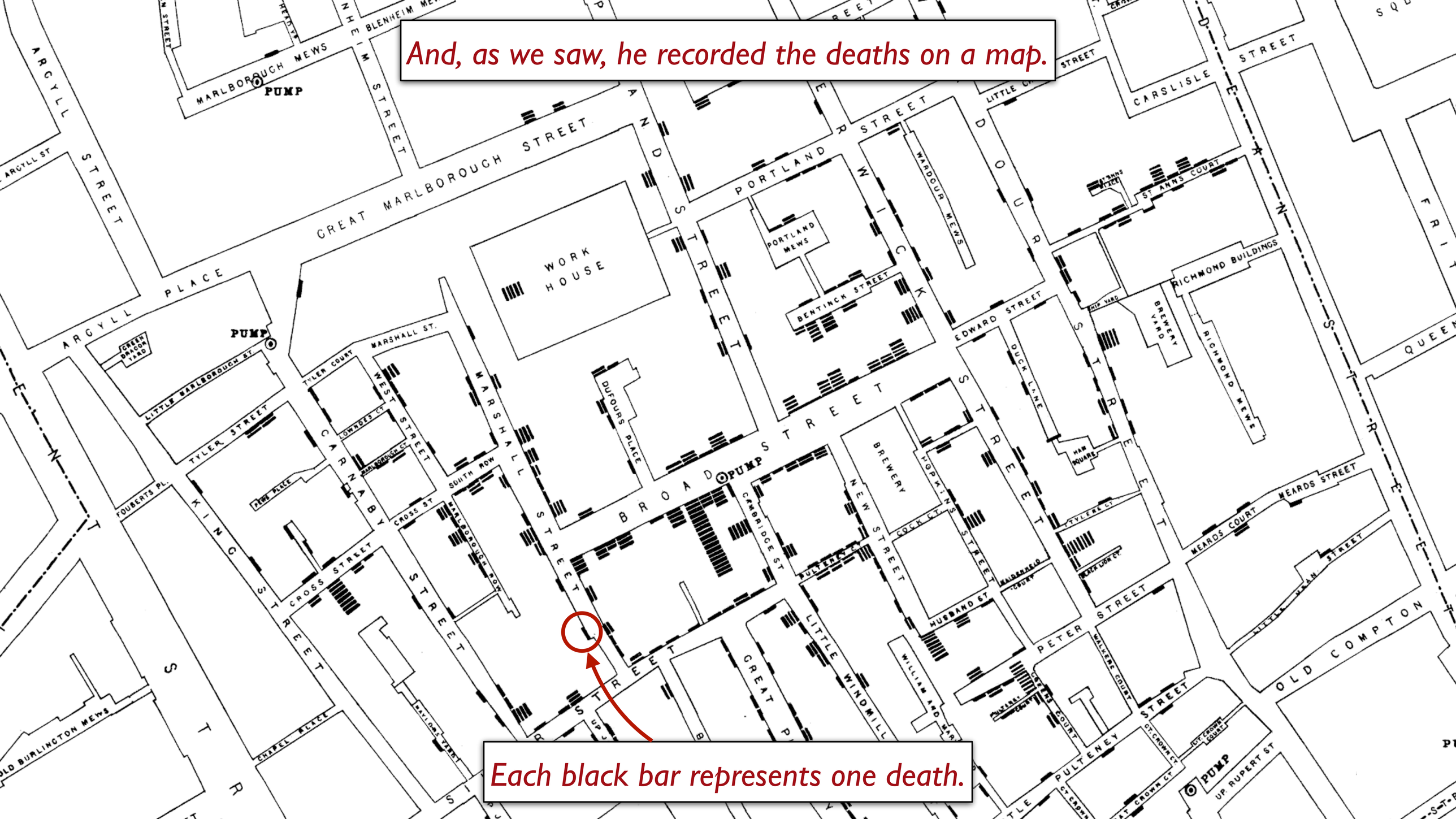


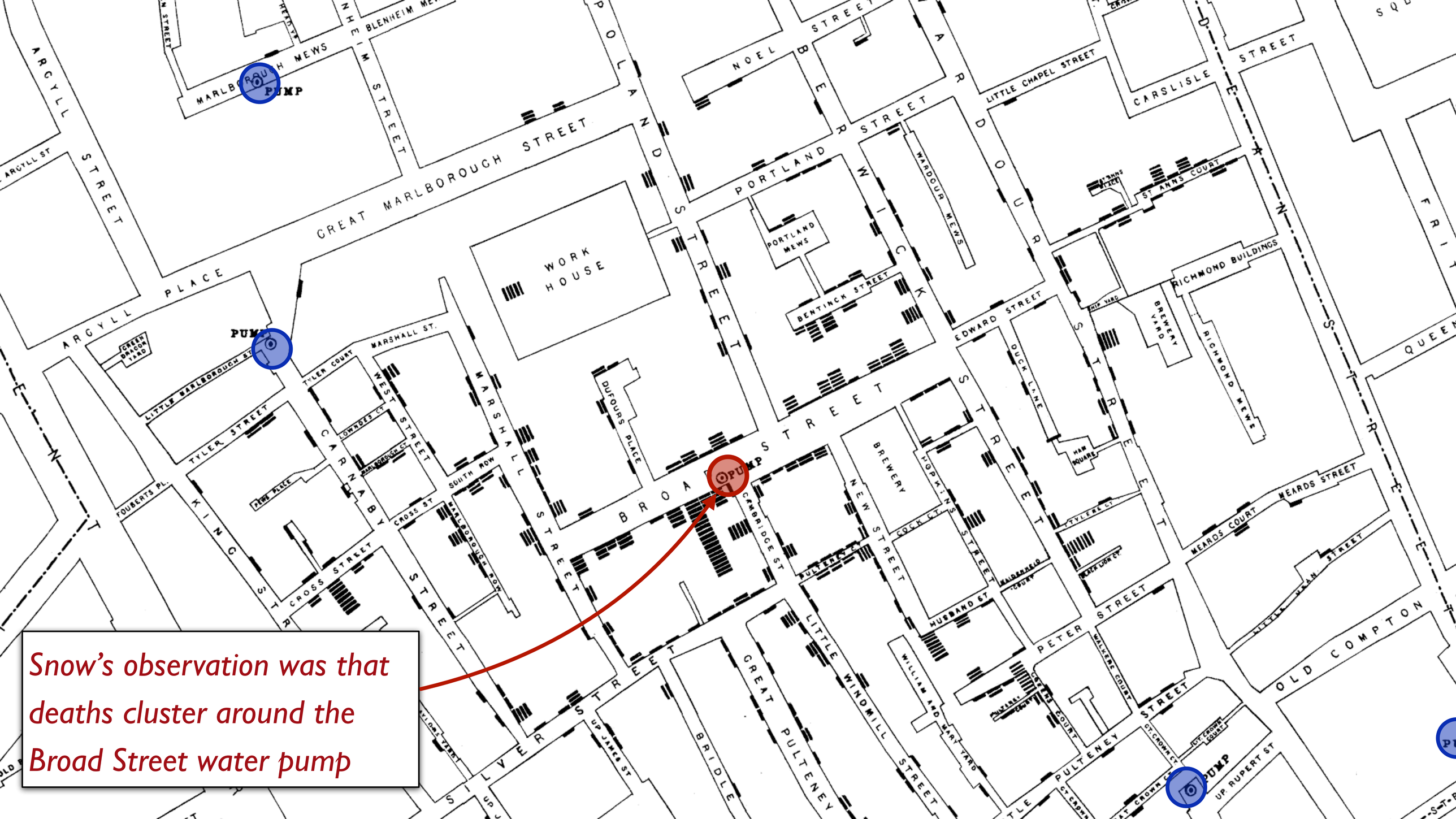
Dartmouth Electron Microscope Facility

Cholera is caused by germ cells (which had not yet been identified), *transmitted through food or drink*.

Physician John Snow (1813–1858) doubted miasma theory.




And, as we saw, he recorded the deaths on a map.
Each black bar represents one death.
PUMP
WORK HOUSE
BROAD STREET
GREAT MARLBOROUGH STREET
PORTLAND STREET
MARSHALL ST.
CARNABY STREET
KING STREET
CROSS STREET
TYLER STREET
LITTLE MARLBOROUGH ST
ARGYLL PLACE
ARGYLL STREET
REGENT STREET
DUFOURS PLACE
BENTINCK STREET
PORTLAND MEWS
WARDOUR MEWS
EDWARD STREET
RICHMOND BUILDINGS
RICHMOND MEWS
BREWERY YARD
DUKE STREET
CARSLISLE STREET
ST ANNS COURT
DUKE LANE
NEW STREET
BREWERY
CAMBRIDGE ST
GREAT PULTENEY
LITTLE WINDMILL
WILLIAM AND MARY
PETER STREET
MEARDS COURT
MEARDS STREET
OLD COMPTON
PULTENEY STREET
UP RUPERT ST
HUSBAND ST
COCK CT
HOPKINS STREET
TYLER CT
MARLBOROUGH ROW
SOUTH ROW
TYLER COURT
LOWNDES CT
MARLBOROUGH CT
CROSS ST
CHAPEL PLACE
OLD BURLINGTON MEWS
FOUBERTS PL
GREEN DRAGON YARD
MARLBOROUGH MEWS
BLENHEIM MEWS


Snow's observation was that deaths cluster around the Broad Street water pump

Snow used his map to convince local authorities to remove the handle of the Broad Street pump.

Snow used his map to convince local authorities to remove the handle of the Broad Street pump.

Later investigation found a cesspit, leaking a few feet from the well of the Broad Street pump; the pump's water was contaminated by sewage from houses of cholera victims.


Southwark and Vauxhall Co.
Lambeth Co.
CHELSEA
BATTERSEA
BATTERSEA PARK
FULHAM
PUTNEY
WIMBLEDON
WIMBLEDON
PARK.
CLAPHAM
Clapham Common
STOCKWELL
BRIXTON
BRIXTON DISTRICT
KENNINGTON
KENNINGTON 1ST DISTRICT
KENNINGTON 2ND DISTRICT
LAMBETH
CAMBERWELL
PECKHAM
PECKHAM NEW TOWN
HATCHAM
NEW CROSS DISTRICT
ST PAUL DEPTFORD
ST GEORGE DISTRICT
WALWORTH
ST PETER WALWORTH DISTRICT
NEWINGTON
SOUTHWARK
DULWICH
EAST DULWICH
DENMARK HILL
HERNE HILL
NUNHEAD
Nunhead Cemetery
CAMBERWELL DISTRICT
PECKHAM DISTRICT
LEWISHAM DISTRICT
SYDENHAM
SYDENHAM DISTRICT
FOREST HILL
BROCKLEY
STREATHAM
STREATHAM DISTRICT
Streatham Common
Tooting Common
UPPER TOOTING
TOOTING
LOWER NORWOOD
WANDSWORTH DISTRICT
CLAPHAM DISTRICT
BATTERSEA DISTRICT
RIVER WANDLE
Wandsworth Common
Radius of 4 Miles from Charing Cross
WEST END AND CRYSTAL PALACE RAILWAY

The River Thames


water flow
The River Thames
water flow

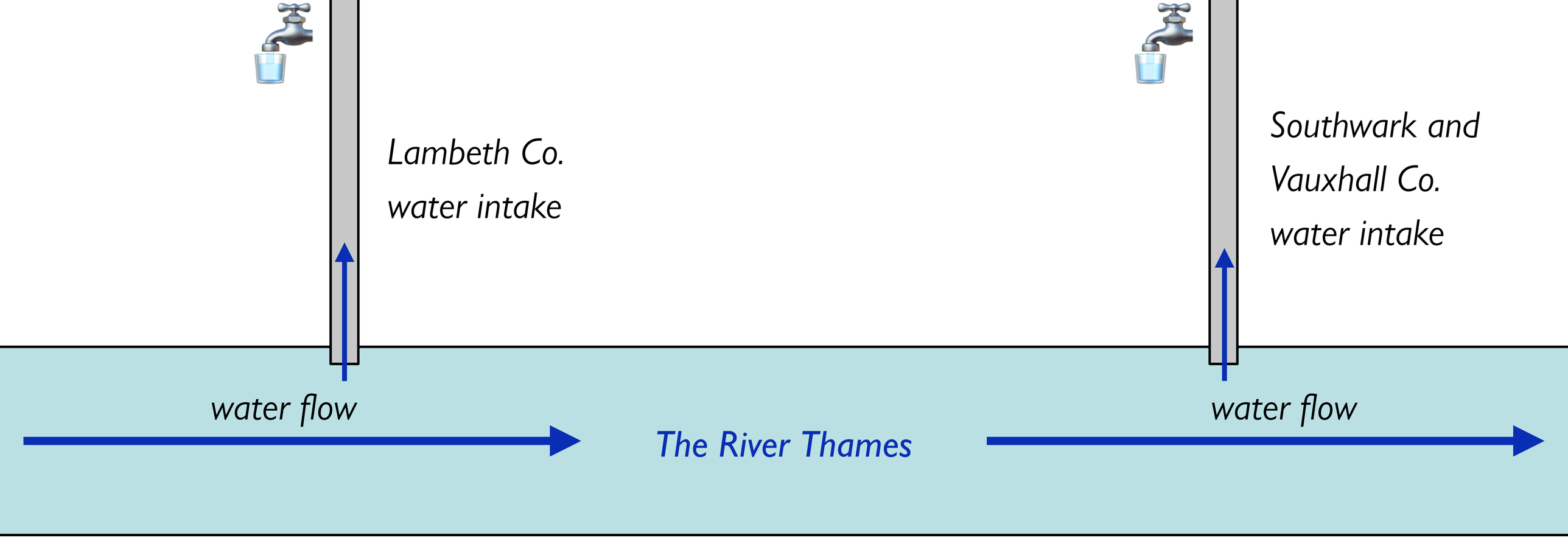
Lambeth Co.

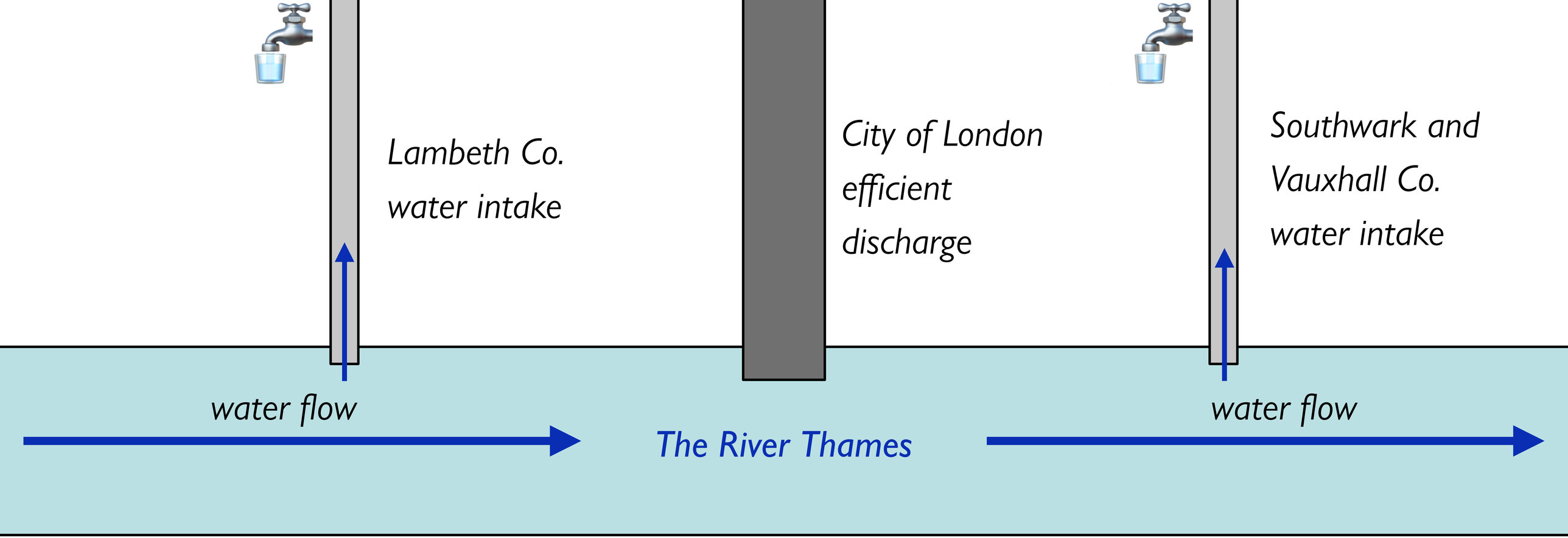
water intake

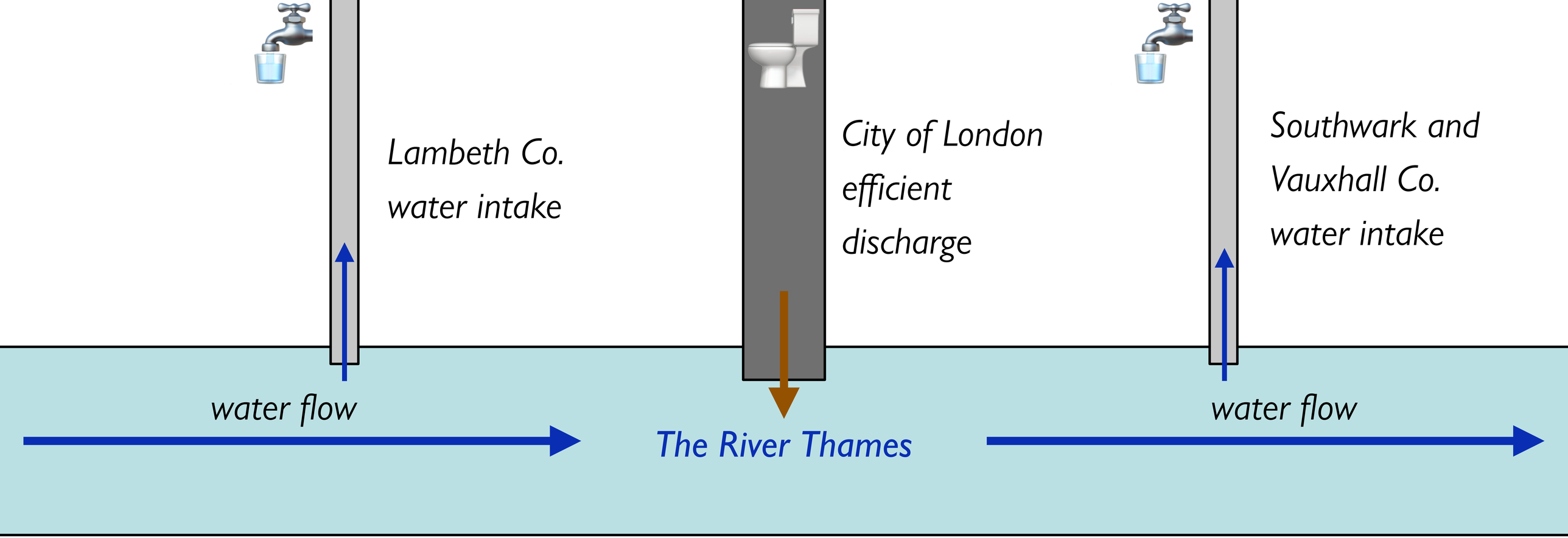
water flow

The River Thames

water flow


Lambeth Co.
water intake
Southwark and
Vauxhall Co.
water intake
water flow
The River Thames
water flow


Lambeth Co.
water intake
City of London
efficient
discharge
Southwark and
Vauxhall Co.
water intake
water flow
The River Thames
water flow


Lambeth Co.
water intake
City of London
efficient
discharge
Southwark and
Vauxhall Co.
water intake
water flow
The River Thames
water flow

"Each company supplies both rich and poor, both large houses and small; there is *no difference* either in the condition or occupation of the persons receiving the water of the different Companies… 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…"

John Snow, *On the Mode of Communication of Cholera*, 1855

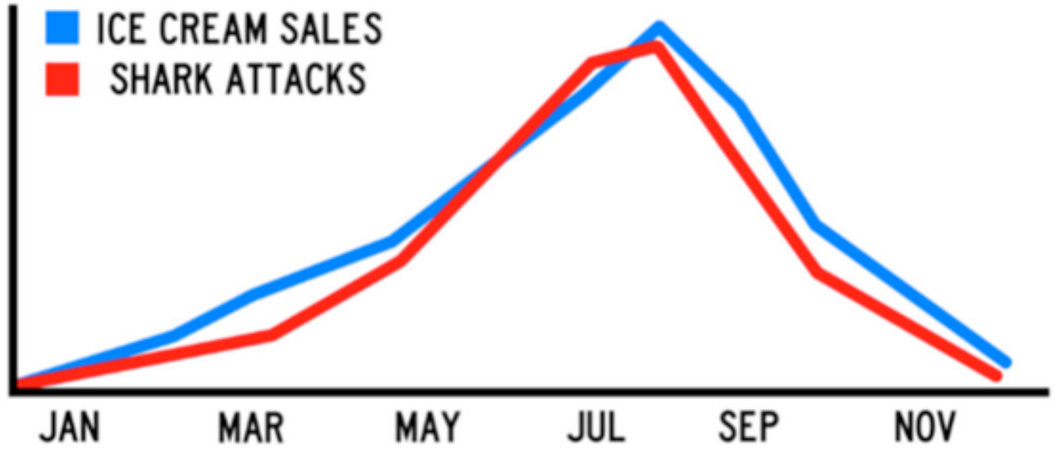
Water supply area	*Number of houses*	*Cholera deaths*
Southwark and Vauxhall Co.	40,046	1,263
Lambeth Co.	26,107	98
Rest of London	256,423	1,422

Water supply area	*Number of houses*	*Cholera deaths*	*Cholera deaths per 10,000 houses*
Southwark and Vauxhall Co.	40,046	1,263	315
Lambeth Co.	26,107	98	37
Rest of London	256,423	1,422	59

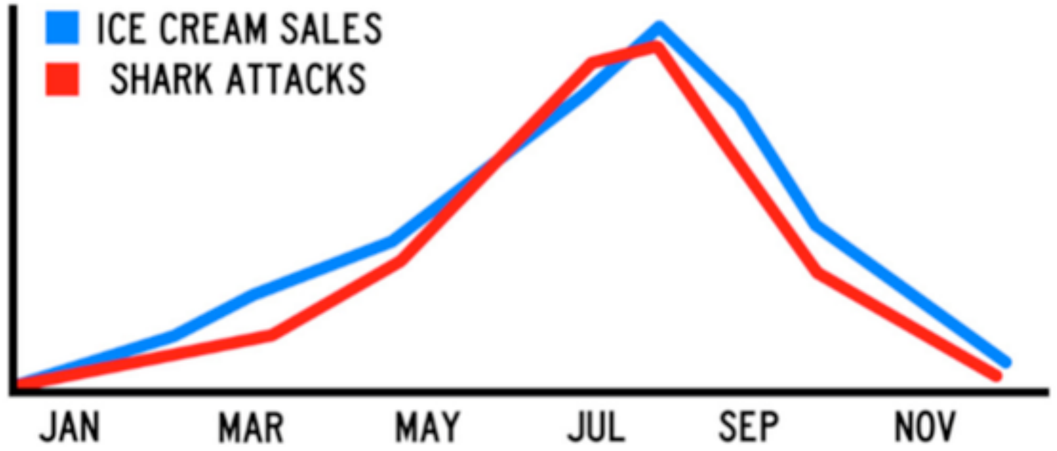
Cause and effect

Association: variables *A* and *B* change at the same time

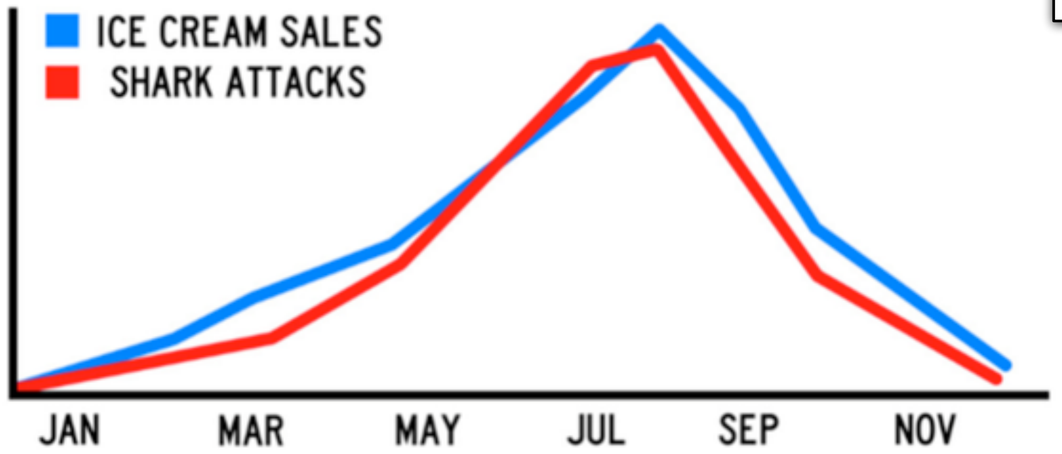
Causation: change in variable *A* makes variable *B* change

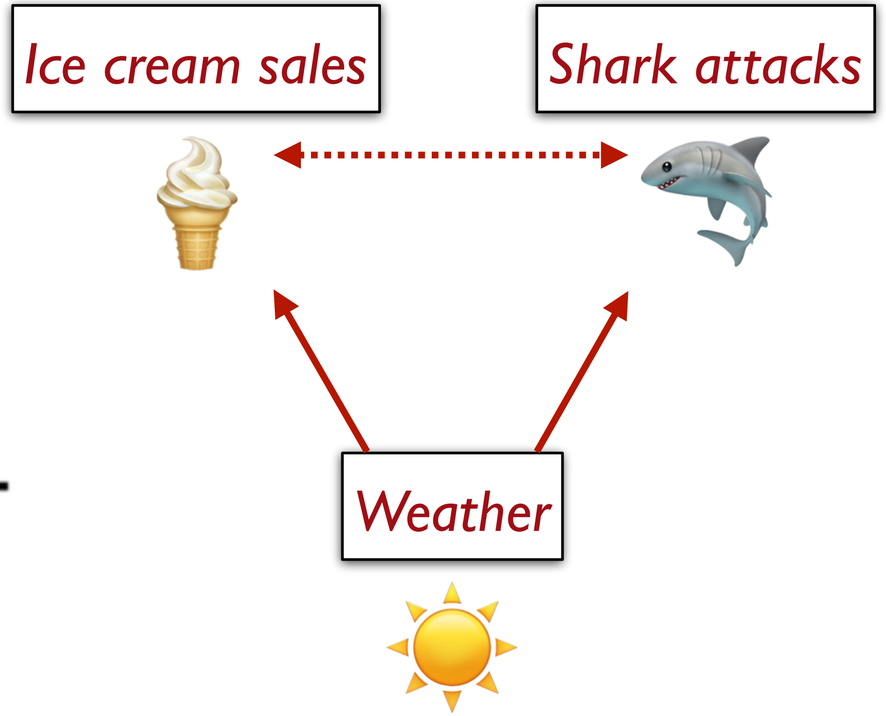


What type of conclusion – association or causation – can we draw from this graph?

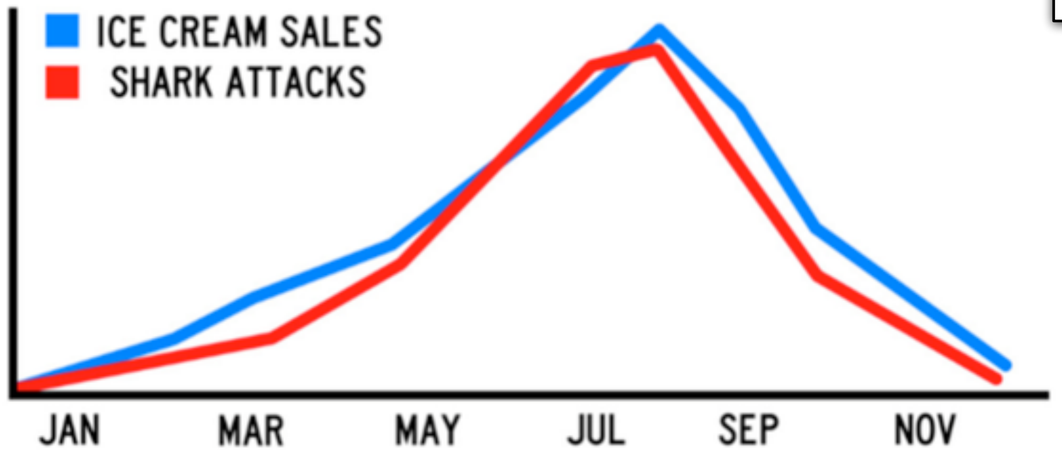


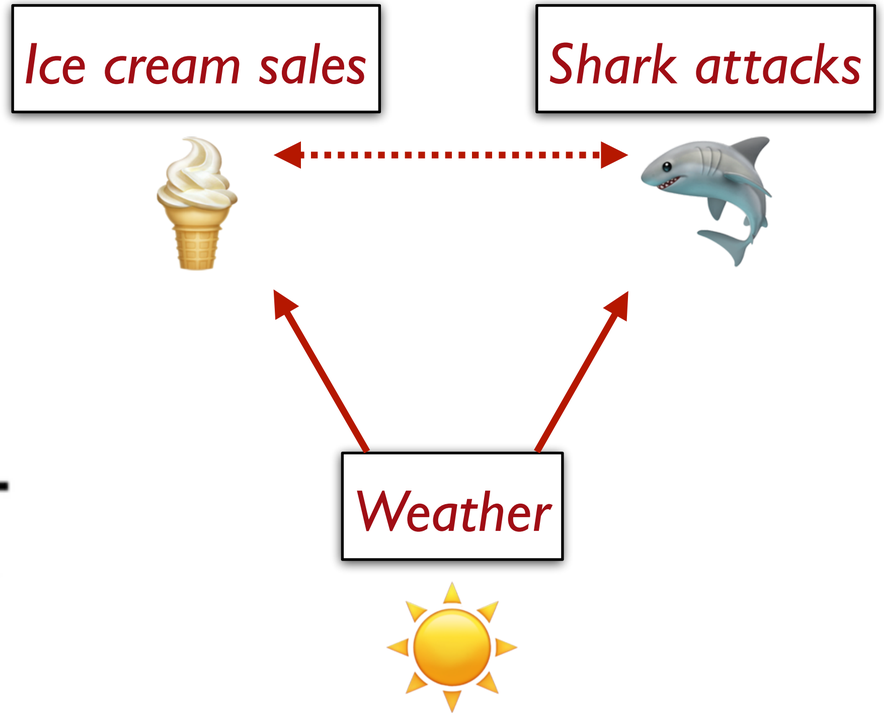
Ice cream sales don't *cause* shark attacks (or vice versa); there's a third thing that causes an increase in both.





Ice cream sales don't *cause* shark attacks (or vice versa); there's a third thing that causes an increase in both.





Ice cream sales don't *cause* shark attacks (or vice versa); there's a third thing that causes an increase in both.

This is called a *confounding variable* (because it can "confound" researchers and lead them astray).

If we can *control for any confounding variables* – that is, remove the possibility that they have an effect on the data – then we can make *causal claims*.

One way to do that is with a *randomized controlled trial*.

Treatment: the variable a researcher would (possibly) manipulate

Outcome: the variable (possibly) affected by treatment

Treatment group: those who receive the treatment

Control group: those who do not receive the treatment

Population: the group for which the conclusions apply

Treatment: the variable a researcher would (possibly) manipulate

Outcome: the variable (possibly) affected by treatment

Treatment group: those who receive the treatment

Control group: those who do not receive the treatment

Population: the group for which the conclusions apply

COVID vaccines

Treatment:

Outcome:

Treatment group:

Control group:

Population:

Treatment: the variable a researcher would (possibly) manipulate

Outcome: the variable (possibly) affected by treatment

Treatment group: those who receive the treatment

Control group: those who do not receive the treatment

Population: the group for which the conclusions apply

COVID vaccines

Treatment: COVID vaccine

Outcome: severity of COVID symptoms

Treatment group: people receiving the COVID vaccine

Control group: people receiving a placebo

Population: the US population

Treatment: the variable a researcher would (possibly) manipulate

Outcome: the variable (possibly) affected by treatment

Treatment group: those who receive the treatment

Control group: those who do not receive the treatment

Population: the group for which the conclusions apply

Cholera and germs

Treatment:

Outcome:

Treatment group:

Control group:

Population:

Treatment: the variable a researcher would (possibly) manipulate

Outcome: the variable (possibly) affected by treatment

Treatment group: those who receive the treatment

Control group: those who do not receive the treatment

Population: the group for which the conclusions apply

Cholera and germs

Treatment: Sewage in drinking water

Outcome: Death from cholera

Treatment group: Households getting Southwark & Vauxhall Co. water

Control group: Households getting Lambeth Co. water

Population: The areas of London served by these companies

While much of this class is focused on the details of working with data, we also want you to think critically about other people's data analyses, which you read or hear about.

The Salt WHAT'S ON YOUR PLATE

EATING AND HEALTH

Chocolate, Chocolate, It's Good For Your Heart, Study Finds

JUNE 19, 2015 · 5:03 AM ET

HEARD ON MORNING EDITION

Allison Aubrey

Treatment: chocolate consumption

Outcome: heart disease

Population: 20,000 European adults followed for 12 years

“Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn’t eat chocolate.”

"Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn't eat chocolate."

What's the relationship between chocolate consumption and heart disease?

"Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn't eat chocolate."

What's the relationship between chocolate consumption and heart disease?

This is just asking about association.

“Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn’t eat chocolate.”

"Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn't eat chocolate."

Does chocolate consumption lead to a reduction in heart disease?

"Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn't eat chocolate."

Does chocolate consumption lead to a reduction in heart disease?

This is asking about causation.

It's often harder to answer questions about causality.

> "[The study] ***doesn't*** prove a cause-and-effect relationship between chocolate and reduced risk of heart disease and stroke."

JoAnn Manson, chief of Preventative Medicine at Brigham and Women's Hospital, Boston

Is the fact that people ate more chocolate the only possible cause for the observed effect of decreased heart disease risk?

For example, suppose the people who ate more chocolate tended to live in European countries with better health care?

What if wealthier people eat more chocolate – and can also afford better health care?

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

If you assign individuals to treatment and control *at random*, then the two groups are likely to be similar apart from the treatment.

Randomized controlled experiments are the *gold standard* for establishing cause and effect.

Observational study: Treatment is non-random.

Research must statistically adjust for confounding variables.

Randomized control experiment: Research manipulates treatment. Controls for confounding variables by experimental design

Natural experiment: Treatment occurs naturally and randomly (no researcher manipulation). Control for confounding naturally.

If randomized controlled experiments can establish causality while observational studies are subject to confounding, why are so many studies in the real world observational?

Review: Scatter plots

Scatter plots show the relationship between two numerical variables.

From 1973 to 2013, Peter and Rosemary Grant visited the Galápagos island of Daphne Major and collected data on finches, to study Darwin's theory of evolution.

The Grants spent years observing, tagging, and measuring finches and their environment.

Notebook: *Finch data*

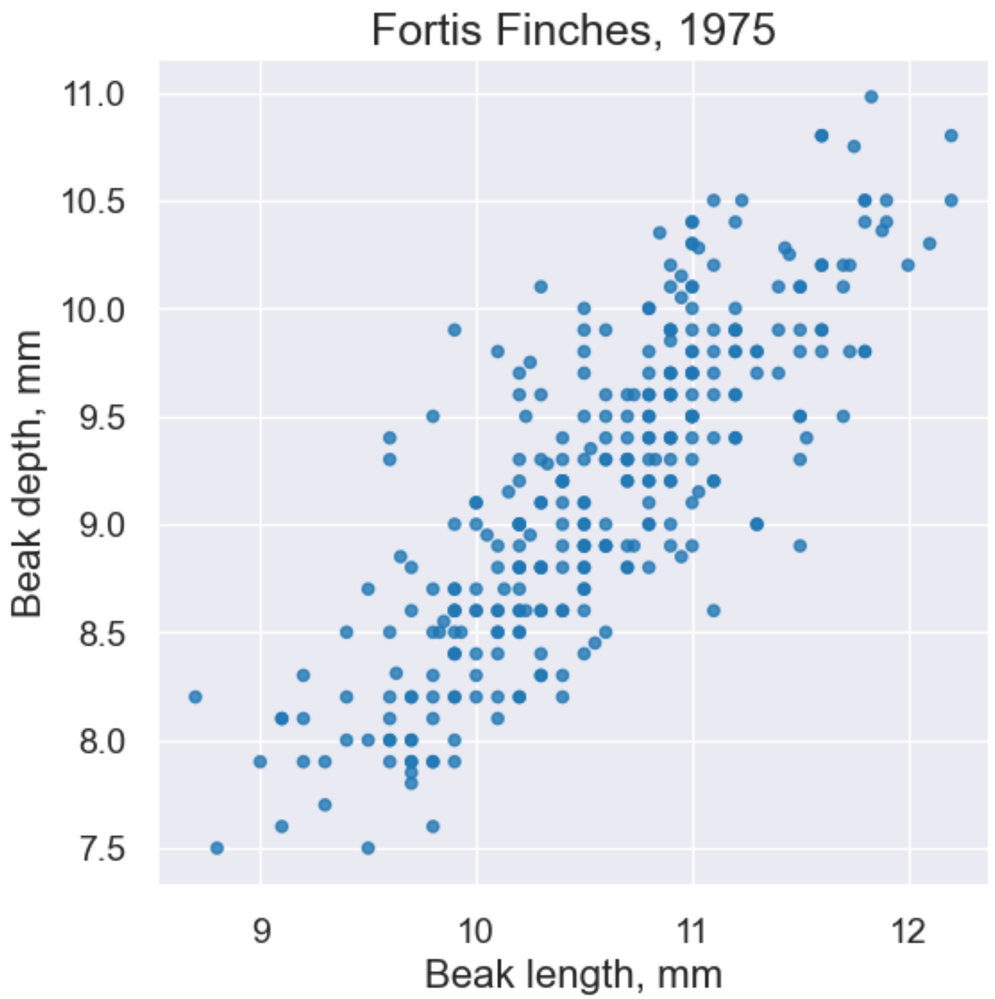
fortis_1975

Beak length, mm	Beak depth, mm
9.4	8
9.2	8.3
9.5	7.5

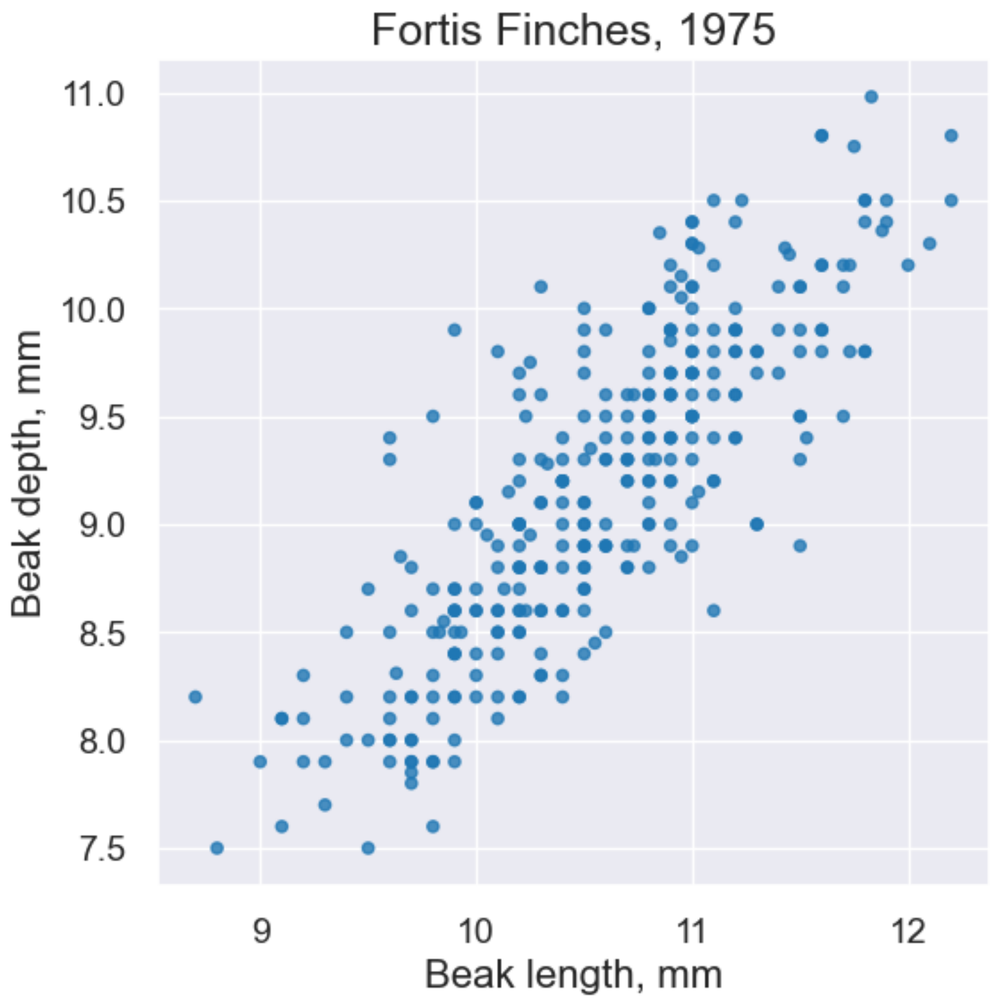
```
fortis_1975
```

Beak length, mm	Beak depth, mm
9.4	8
9.2	8.3
9.5	7.5

```
fortis_1975.scatter(
    "Beak length, mm",
    "Beak depth, mm"
)
```



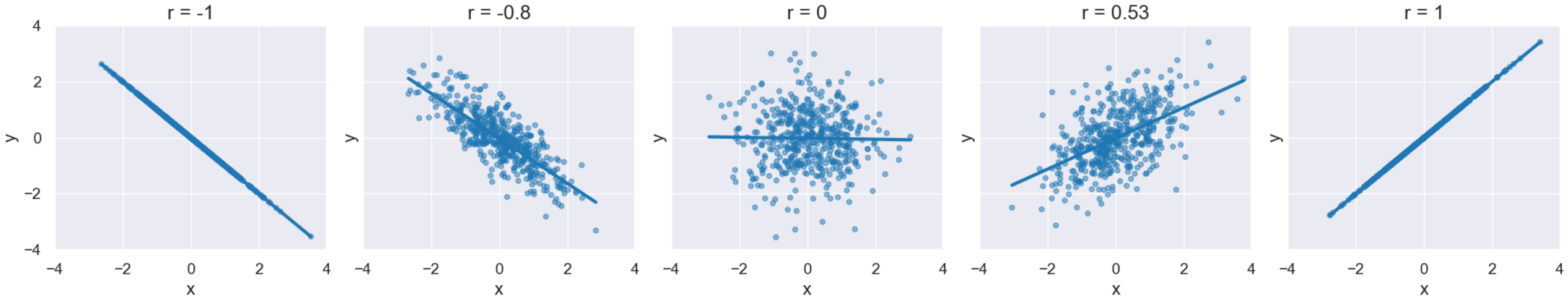
"On average, as beak length increases, beak depth increases."



Correlation

Quantifying linear association

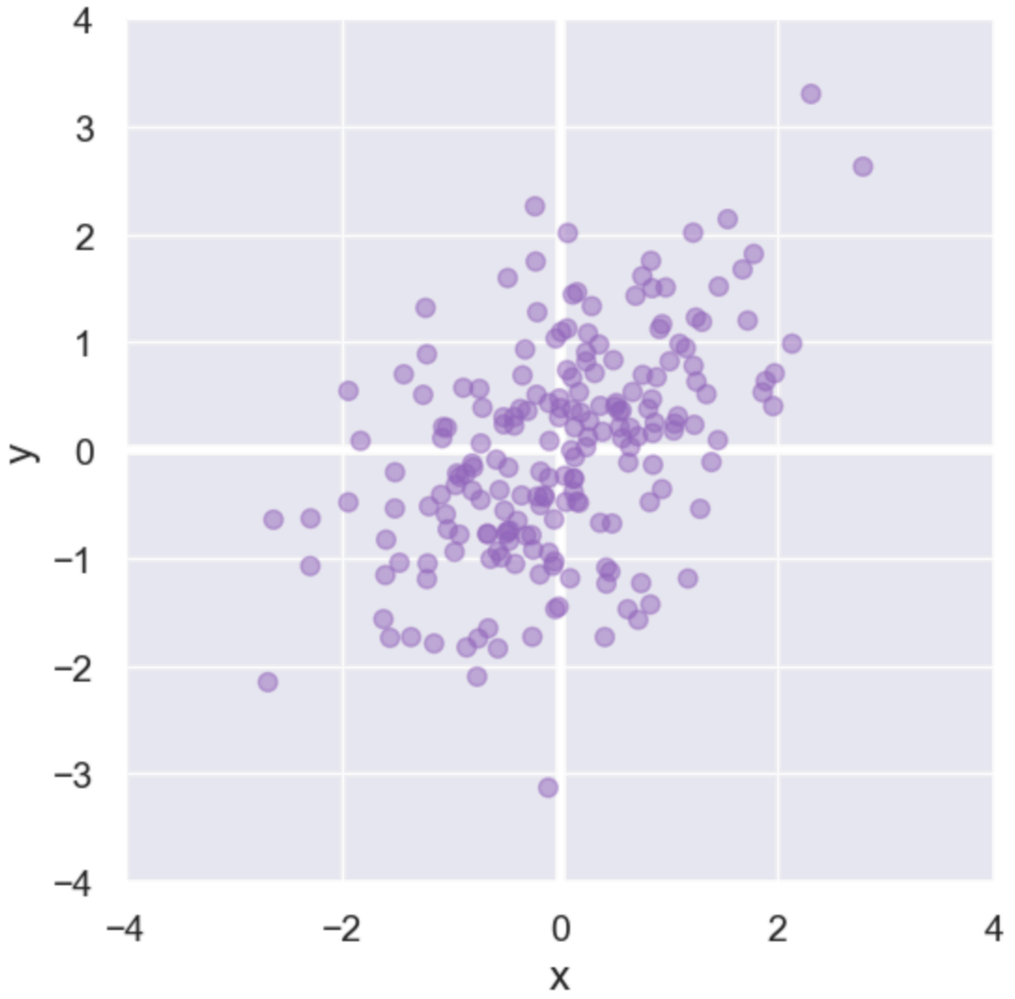
Stronger association means the points are more tightly clustered on a straight line.

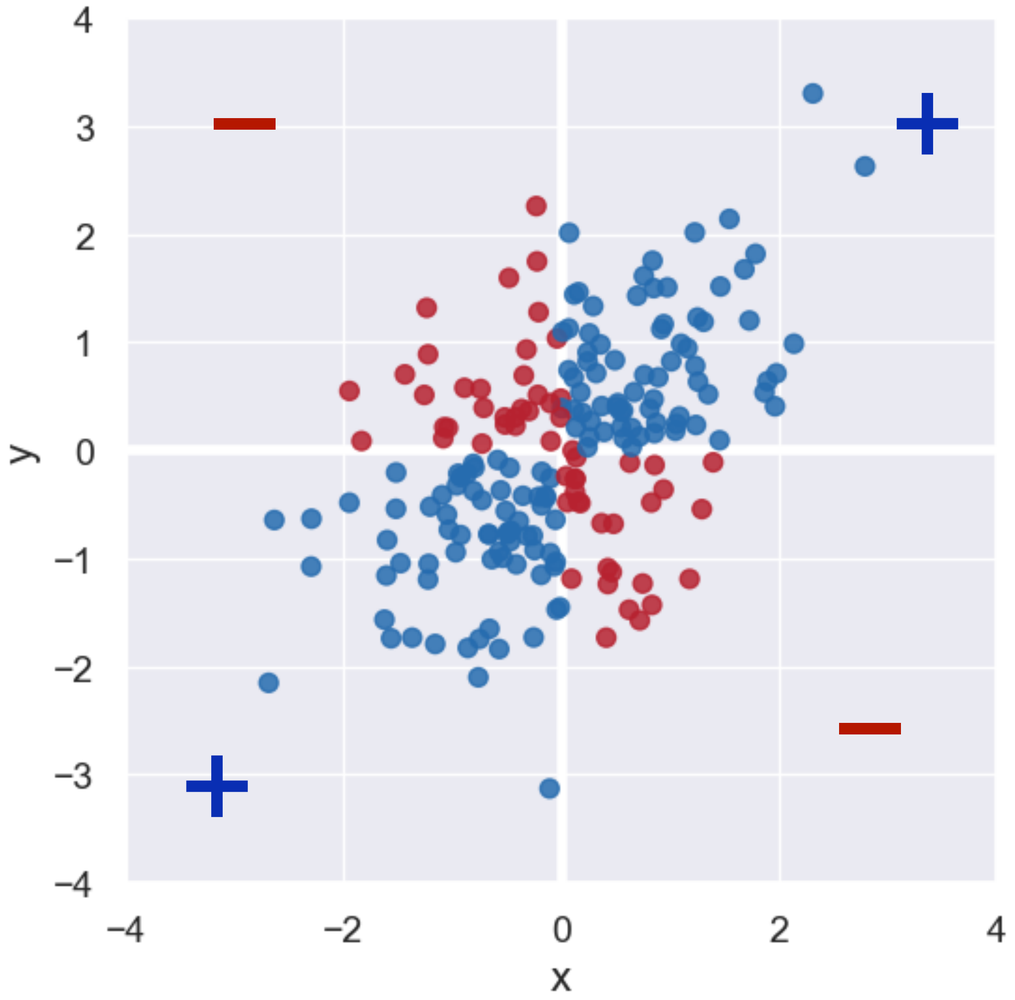


Most negative association

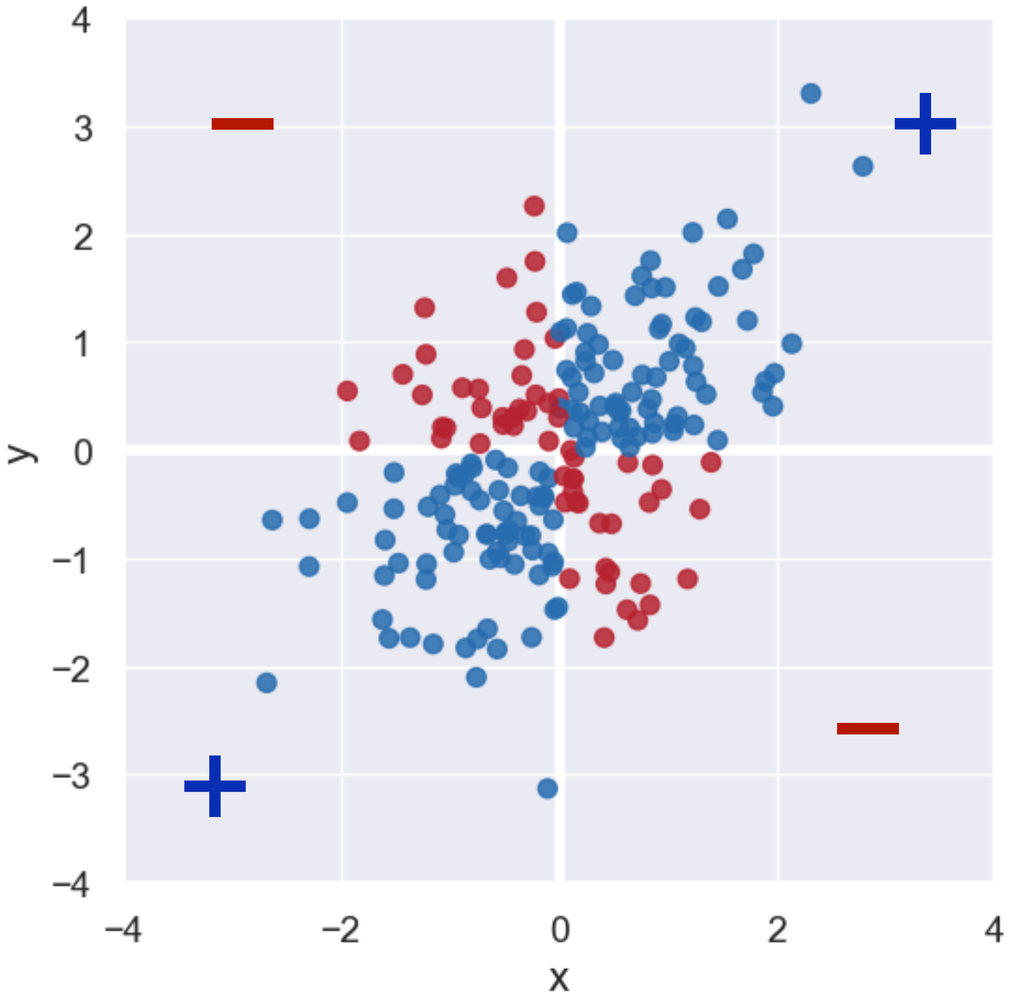
No association

Most positive association

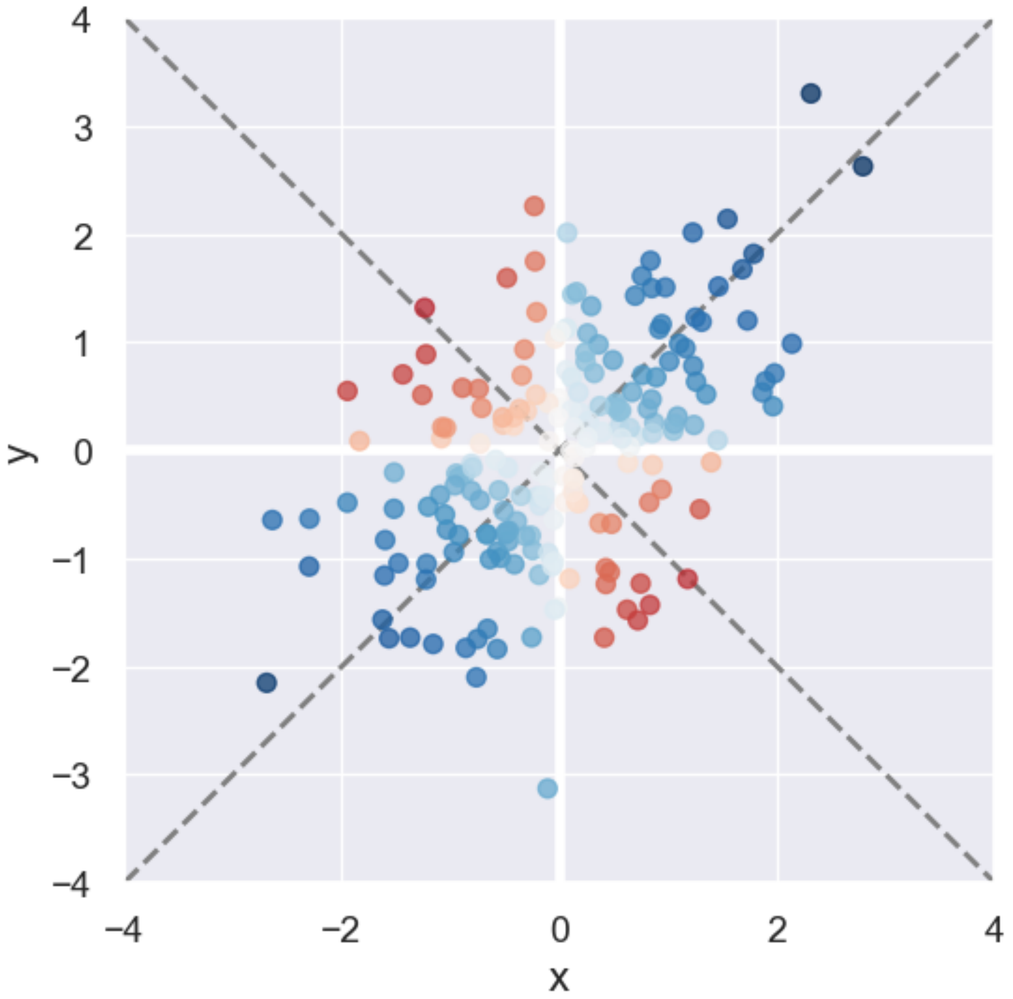

4
3
2
1
0
−1
−2
−3
−4
y
−4
−2
0
2
4
x


4
3
2
1
0
−1
−2
−3
−4
y
−4
−2
0
2
4
x

For each point (x, y), the *sign* of the product xy indicates whether the point suggests a positive or negative correlation.



For each point (x, y), the *sign* of the product xy indicates whether the point suggests a positive or negative correlation.



For each point (x, y), the *magnitude* of the product xy indicates the strength of correlation.

We can sum up the contributions of each point:

$$r = \sum_i x_i y_i$$

We can sum up the contributions of each point:

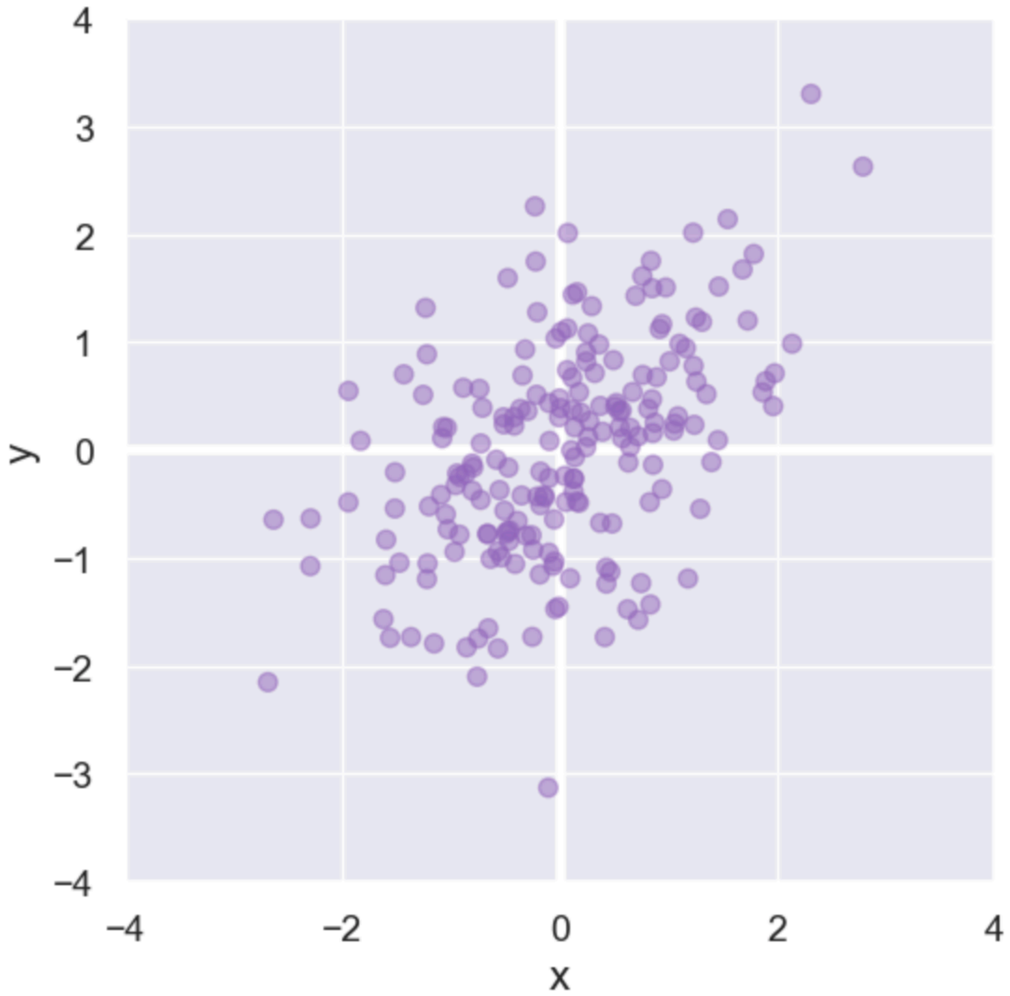
$$r = \sum_i x_i y_i = x_0 y_0 + x_1 y_1 + \cdots + x_n y_n$$

We can sum up the contributions of each point:

$$r = \sum_i x_i y_i = x_0 y_0 + x_1 y_1 + \cdots + x_n y_n$$

And then normalize it to be in the range $[-1, 1]$ to make it interpretable:

$$r = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$



$r = 0.48$

There's one more issue we need to deal with to have a good measure of how associated two variables are.

r = 0.24

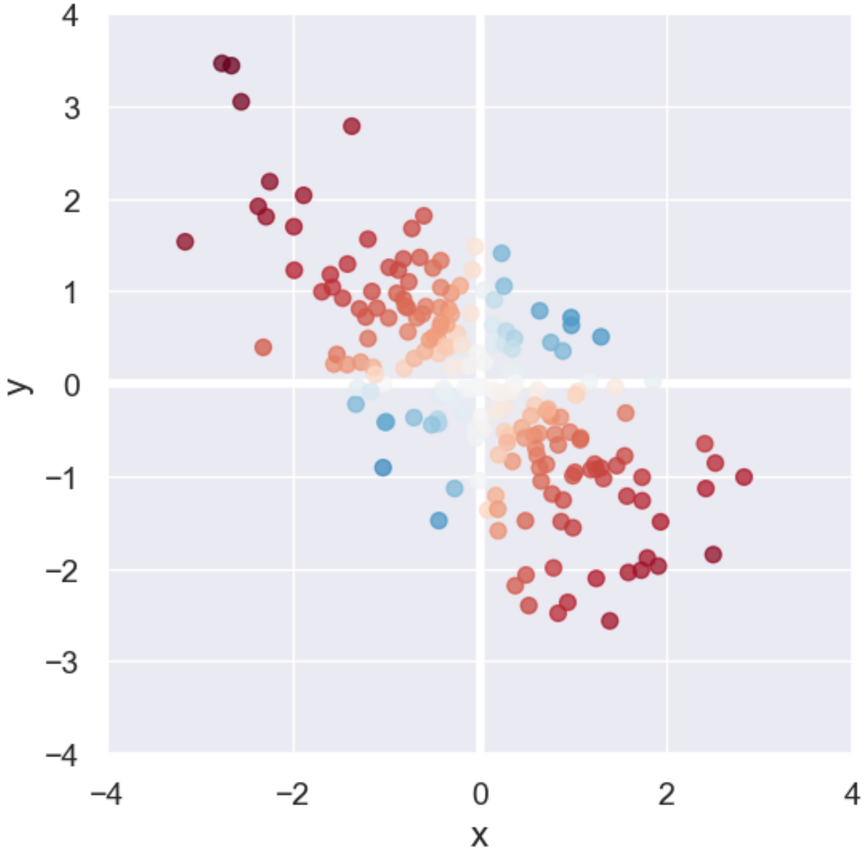
But this should have a negative correlation!

If the points aren't centered at the origin, it messes with the *r* we compute:

r = 0.24

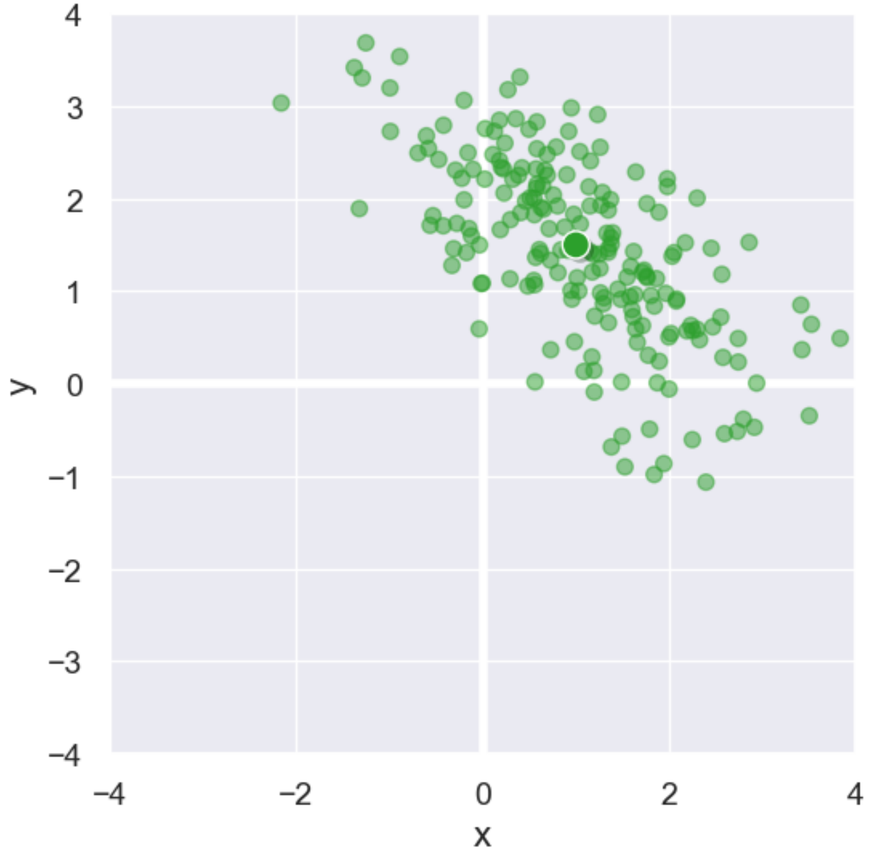
If the points aren't centered at the origin, it messes with the *r* we compute:

r = 0.24

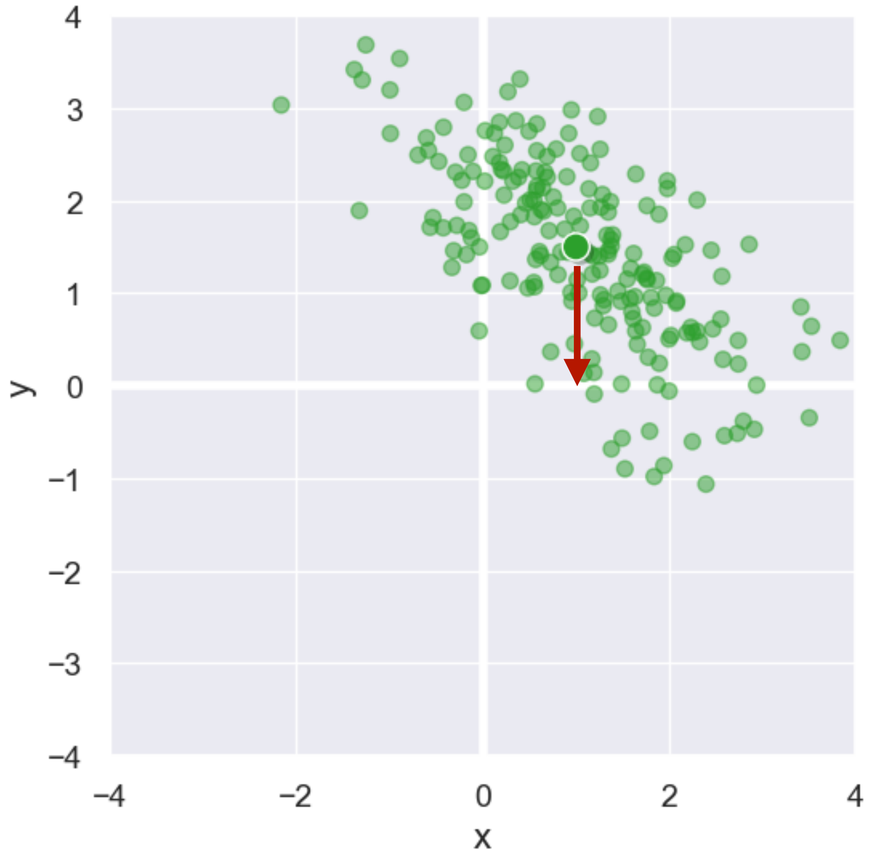


r = −0.70

So we center the data at (0, 0) by subtracting the average *x* and *y* from each point:

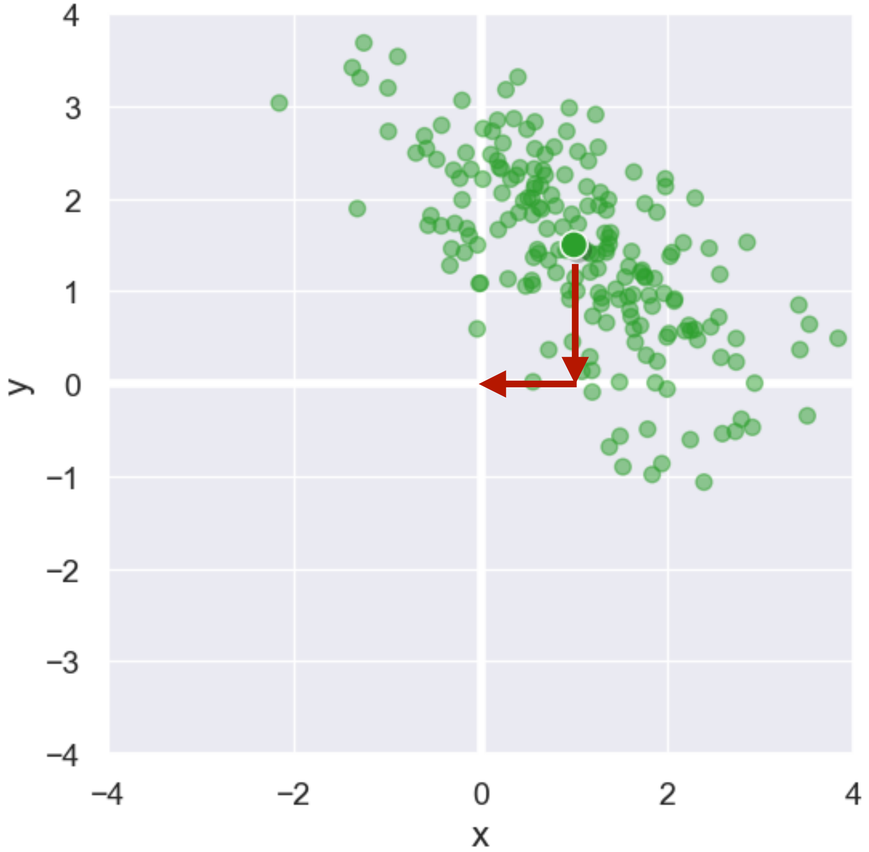


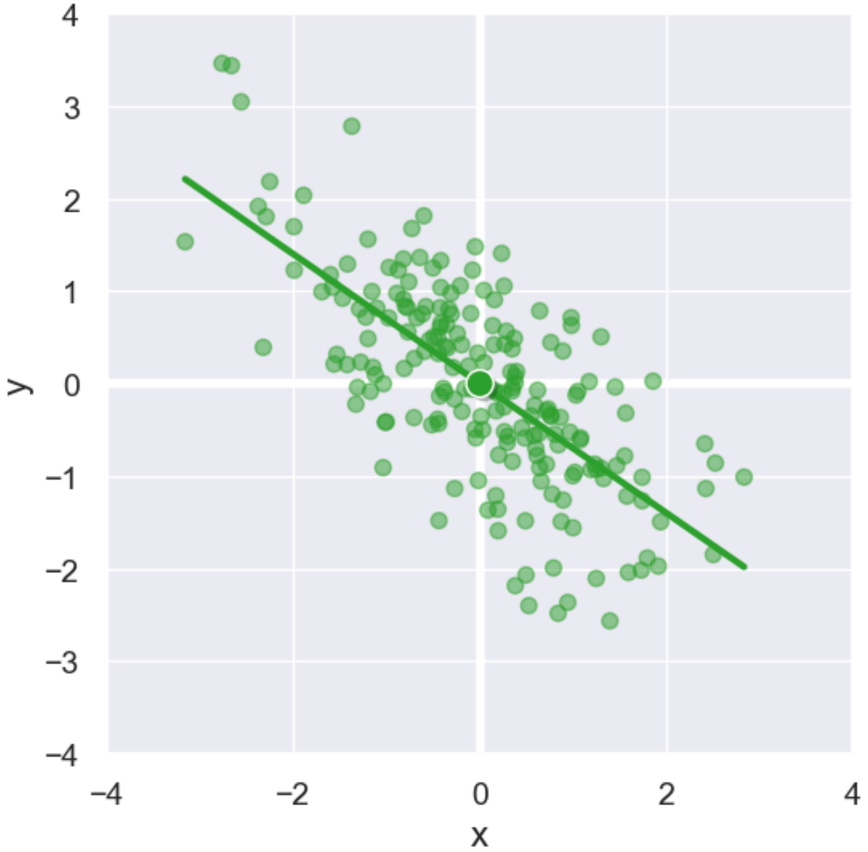
So we center the data at (0, 0) by subtracting the average *x* and *y* from each point:



So we center the data at (0, 0) by subtracting the average *x* and *y* from each point:

So we center the data at (0, 0) by subtracting the average *x* and *y* from each point:



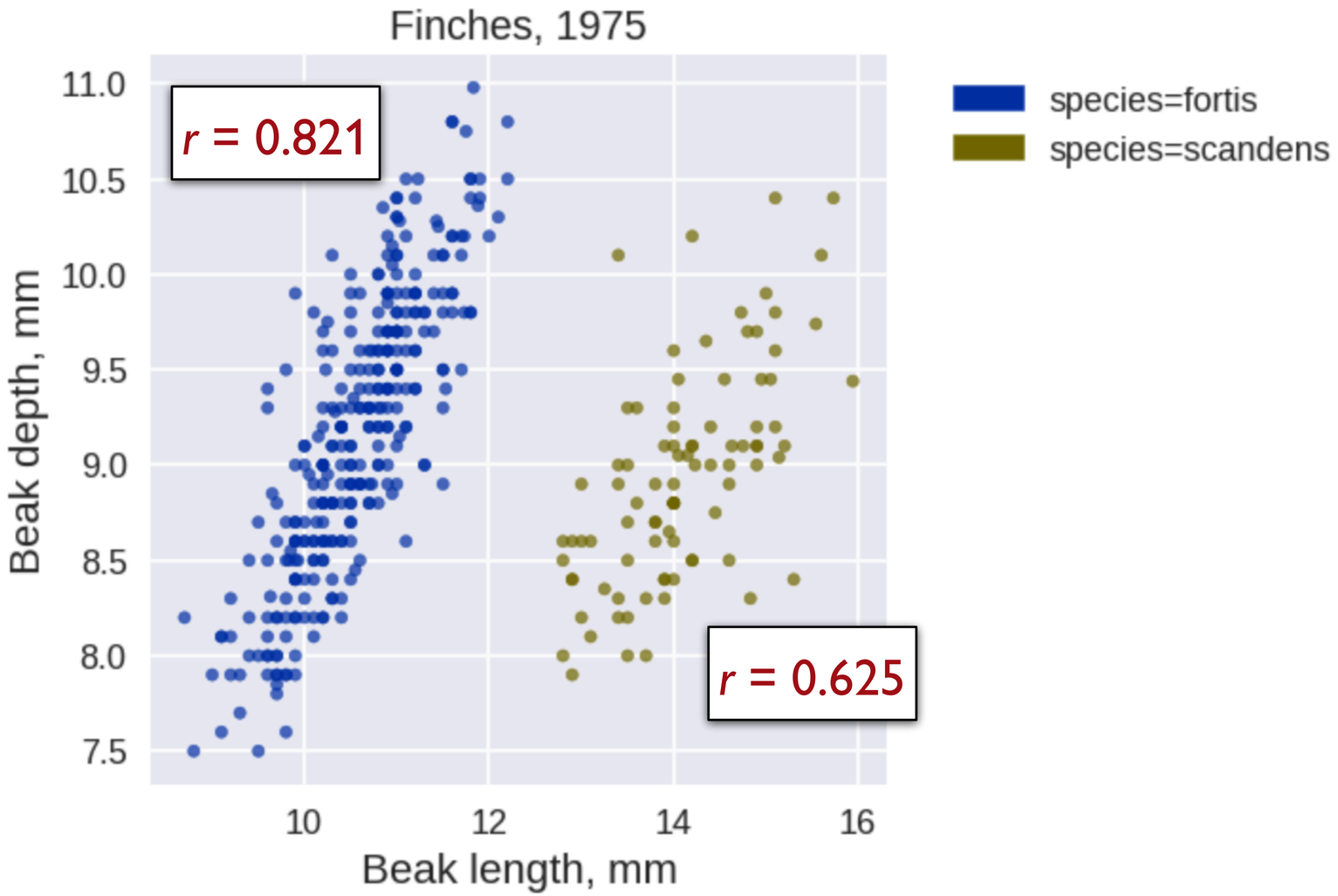


Pearson's correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Notebook: *Correlation*

Associations between beak length and depth



Acknowledgments

This class incorporates material from:

Stephen Freund and Katie Keith, Williams College

Data 8, University of California, Berkeley