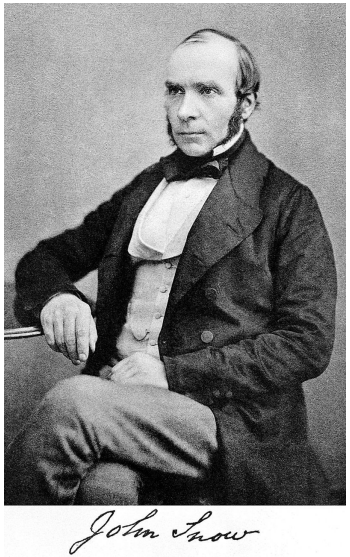
Quantitative Data Concepts

Data, models and uncertainty

Christine Choirat
HSPH Biostatistics & IQSS

“The cholera map that changed the world”

theguardian



John Snow (1813-1858) was a physician (anesthetics, cholera, maps)

John Snow's data journalism: the cholera map that changed the world

John Snow's map of cholera outbreaks from nineteenth century London changed how we saw a disease - and gave data journalists a model of how to work today

- Interactive map
- Download the data
- More data journalism and data visualisations from the Guardian

https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map

Next stop: Charles MGH

1846: first public (and successful) general anesthesia at MGH

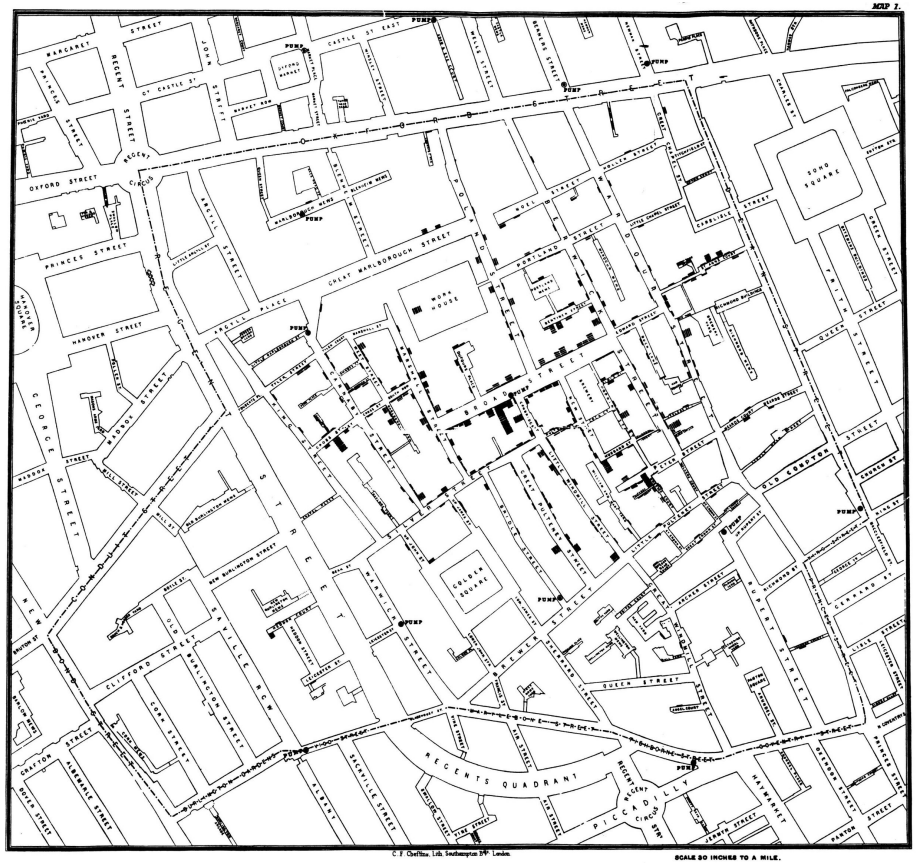
Snow performed obstetric anesthesia for Queen Victoria (1853, 1857)

Let’s take a closer look

Miasma theory: diseases (cholera, Black Death, ...) caused by *bad air* coming from rotting organic matter

No germ theory yet (Pasteur, 1861)

Bars standing for number of cases



Big data in the 19th Century

- Several cholera outbreaks in London
- Many people, overrunning cesspools, no sewer system in Soho
- Major outbreak in Soho: **1854 Broad Street cholera outbreak** (617 deaths)

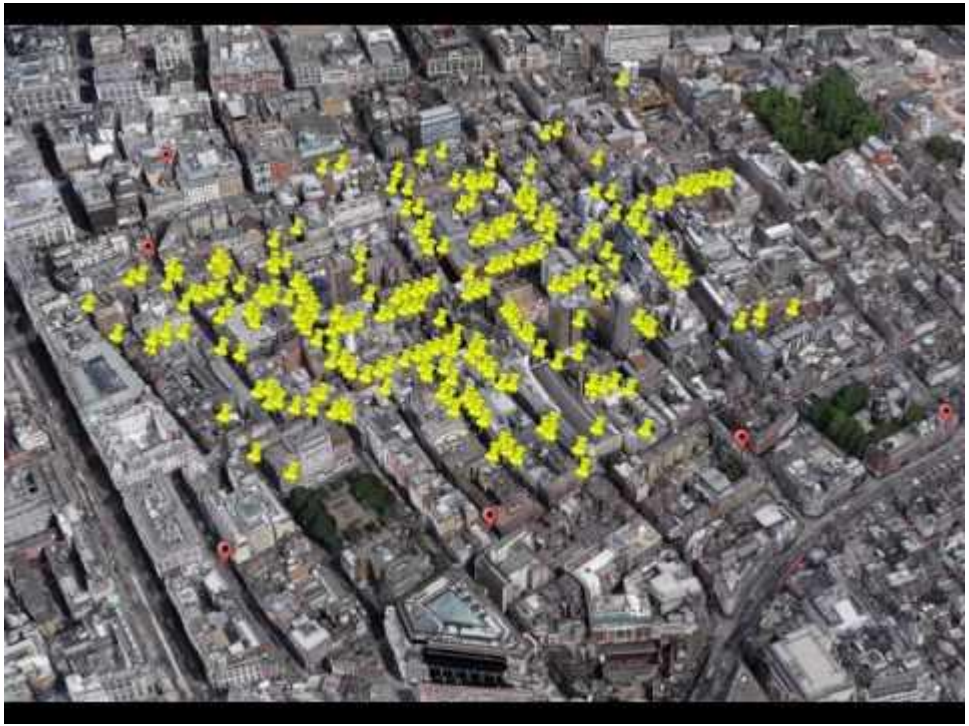
John Snow, who doubted miasma theory, studied the outbreak:

On the Mode of Communication of Cholera (1849, 1855): “NOT COMMUNICATED BY MEANS OF EFFLUVIA “

“There is, in our view, an entire failure of proof that the occurrence of any one case could be clearly and unambiguously assigned to water.” (London Medical Gazette, 1849)

http://www.johnsnowsociety.org/john-snow.html

Broad Street Outbreak: Exploring the data



- Which data?
  - Spatio-temporal data: Coordinates / addresses, count of cases
  - Interviews of the neighbors
- Which questions do we want to address?
  - Miasma theory or something else?
  - Can we identify the source of the outbreak?
  - Can we implement an intervention to stop the outbreak?

From Broad Street to Baker Street

Cluster of cases around the Broad Street water pump

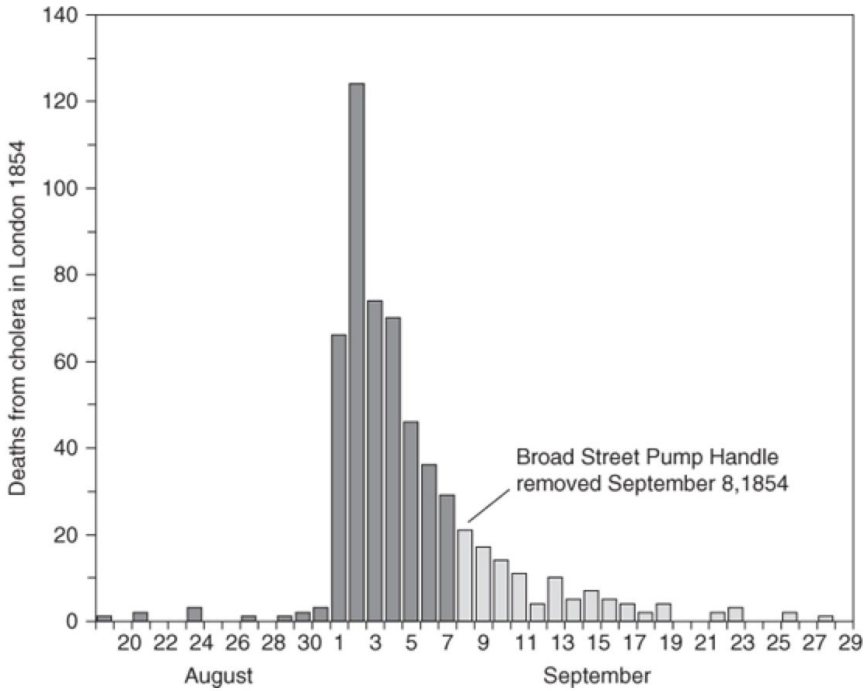
The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the ***persons who were in the habit of drinking the water*** *of the above-mentioned pump well.*

I had an interview with the Board of Guardians of St James's parish, on the evening of the 7th inst [7 September], and represented the above circumstances to them. In consequence of what I said, ***the handle of the pump was removed on the following day****.*

Inconclusive water analysis

Patient zero

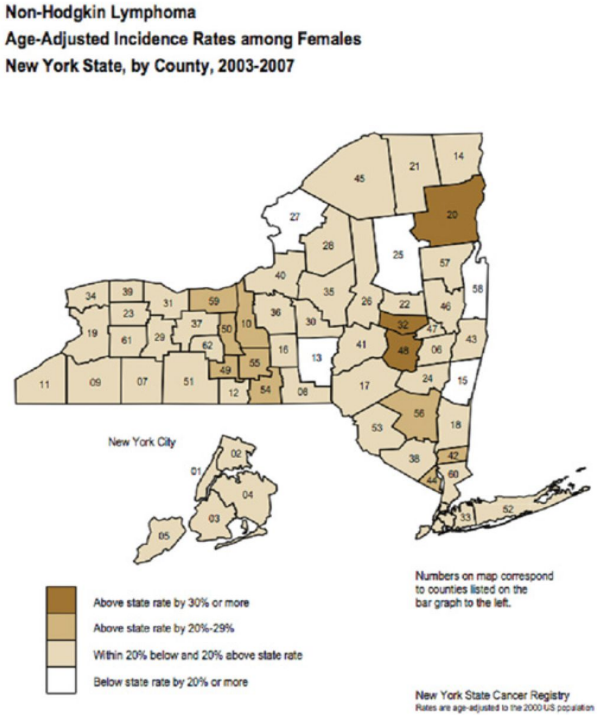
What was the impact of removing the pump?



What if... the pump hadn't been removed?

http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html

New York’s cancer maps



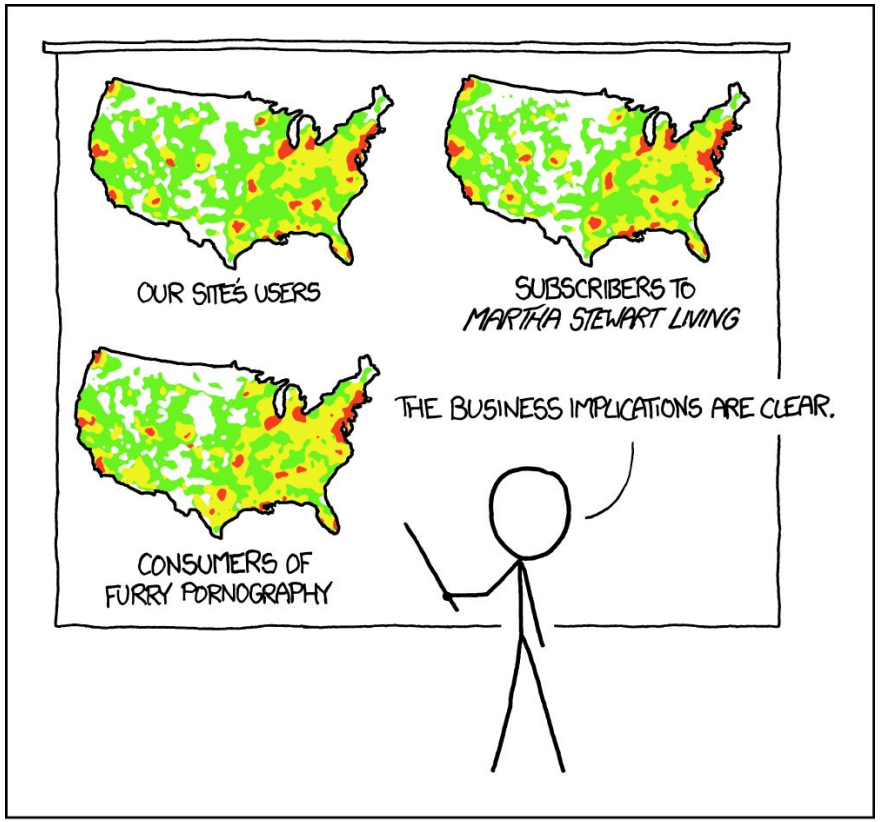
2010: Interactive maps with layers (hazardous sites, …)

Objections from the American Cancer Society

Warning on the website:

The map cannot explain why cancer may be higher or lower in certain areas. It does not show that an environmental facility causes cancer.

- *The map does not contain any information about important known individual risk factors for cancer.*
- *The environmental facility information only shows the locations of facilities.*
- *The cancer information reflects people's addresses at the time of their cancer diagnoses.*

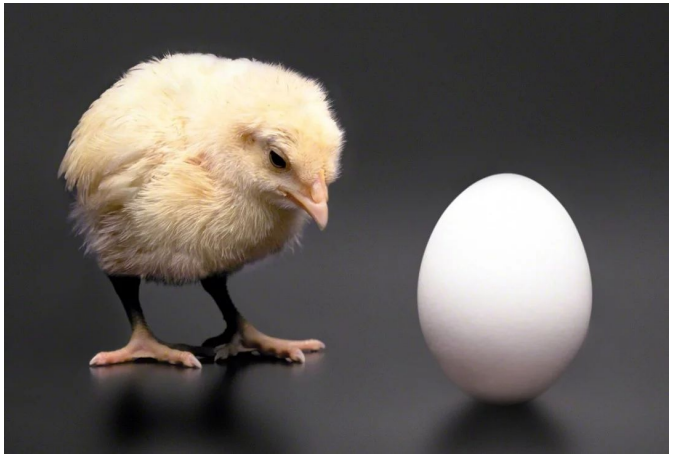


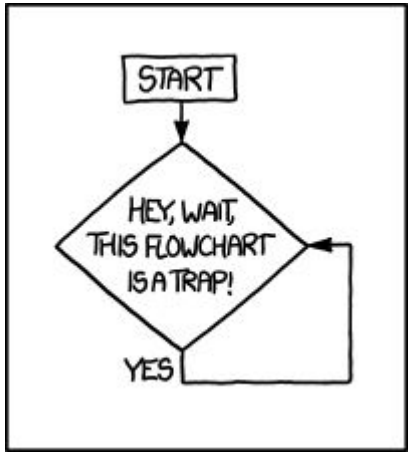
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

https://xkcd.com/1138/

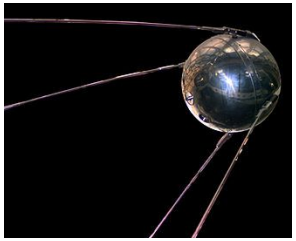
Why do we collect data? And which data?

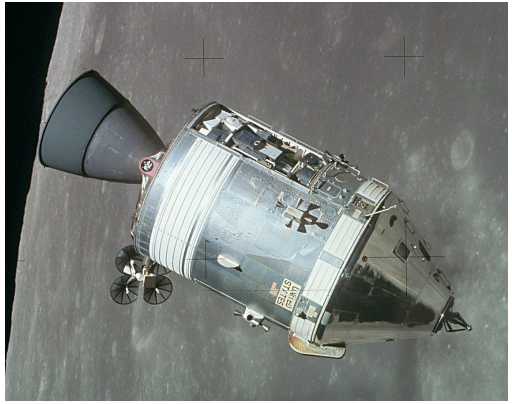
- Because we can
- Because we think it will be “useful” at some point
- To make pretty (interactive) graphs / maps / widgets
- To make better decisions
- Better?





Keeping all the data?





Sufficient statistics

$$X_1, X_2, \ldots, X_t$$

$$M_t = \frac{X_1 + X + \ldots + X_t}{t}$$

$$X_1, X_2, \ldots, X_t, X_{t+1}$$

$$M_{t+1} = \frac{X_1 + X_2 + \ldots + X_t + X_{t+1}}{t+1}$$

$$M_{t+1} = \frac{tM_t + X_{t+1}}{t+1}$$

$$(t, M_t, X_{t+1})$$

Data, models and uncertainty

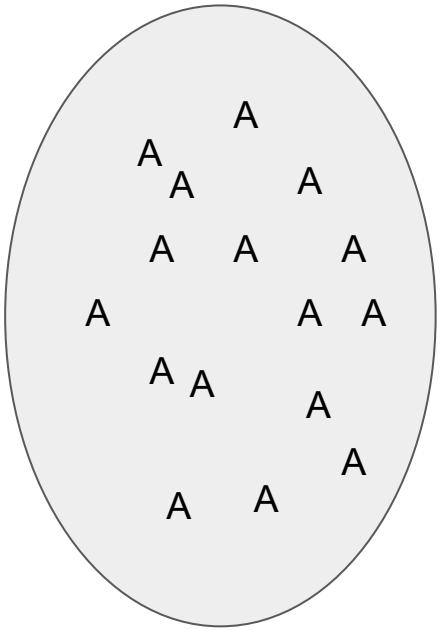
Collect data

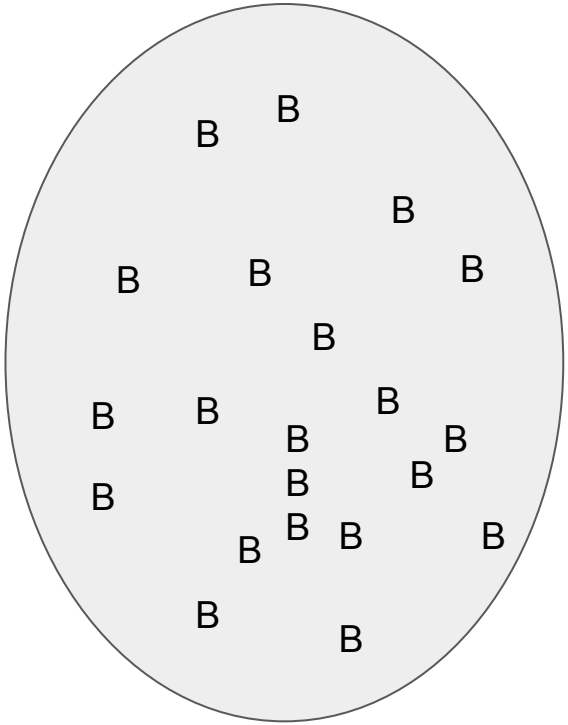
```
2 3 4 6 2 6 6 4 4 1 2 2 5 3 5 3 5 6 3 5 6 2 4 1 2 3
1 3 6 3 3 4 3 2 5 5 5 1 5 3 5 4 5 4 4 5 1 3 5 5 3 6
3 2 1 1 2 4 4 3 6 2 3 2 4 2 3 5 1 6 3 6 3 3 3 6 6 3
5 6 3 5 3 2 5 2 5 1 2 1 2 1 4 6 5 5 3 3 5 4
```

Estimate the shape of the dice

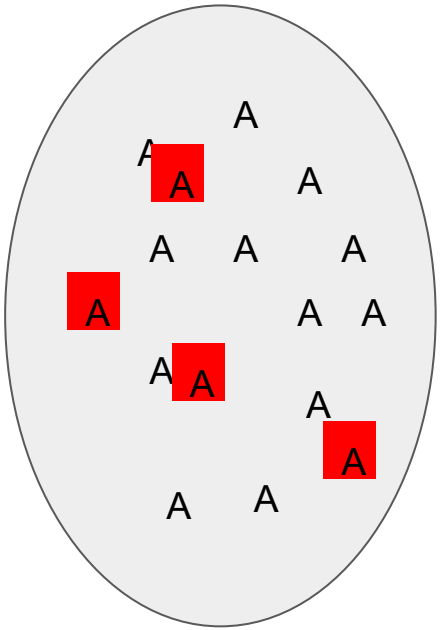
Throw the dice!

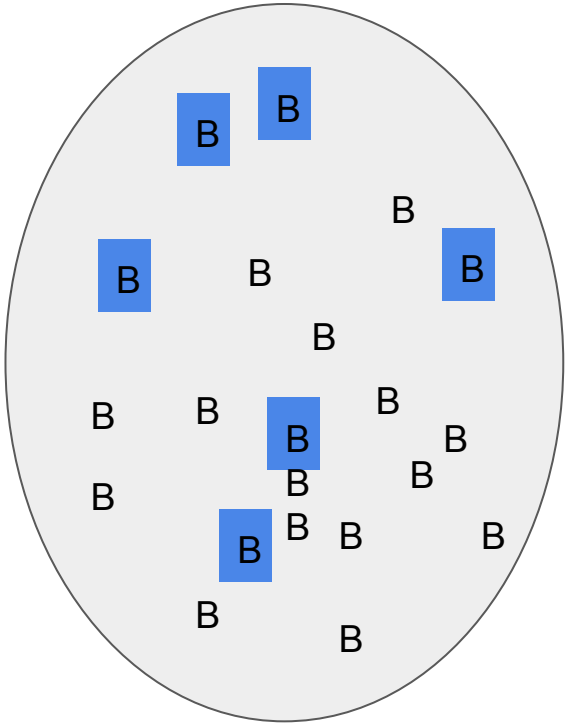
Candidate A or candidate B?





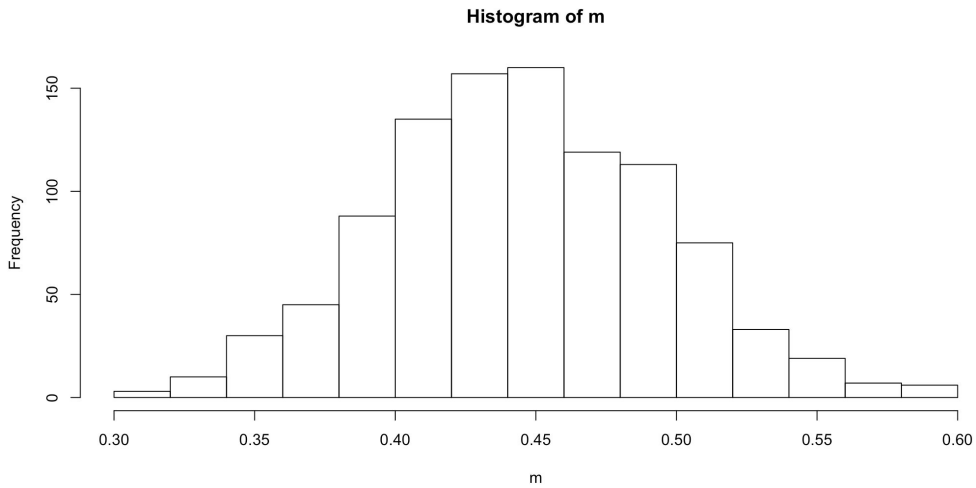
Candidate A or candidate B?





Uncertainty I

```
set.seed(1)

A <- rep(1, 45000)

B <- rep(0, 55000)

population <- c(A, B)

m <- rep(NA, 1000)

for (i in 1:1000)

 m[i] <- mean(sample(population, size =
100))

hist(m)
```



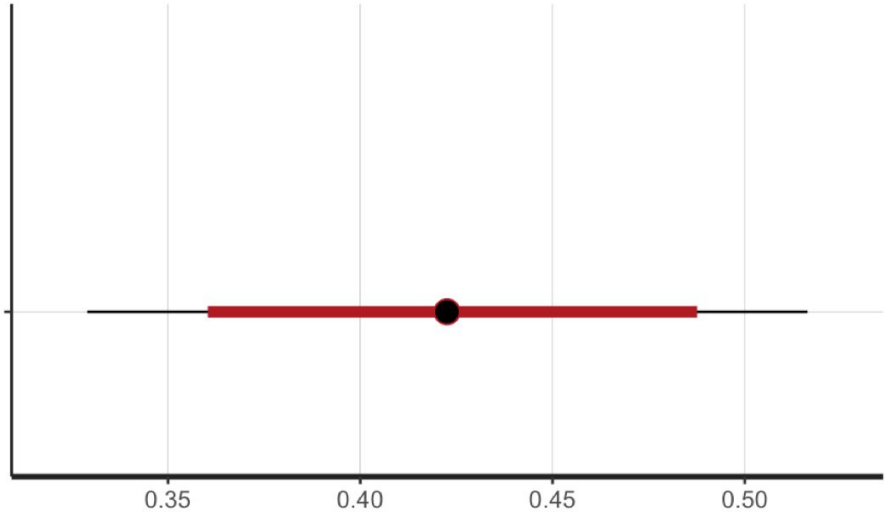
Uncertainty II: Throw the dice!

```
set.seed(1)

rbinom(n = 1, prob = 0.45, size = 1)                    # 0
```

Eventually...

```
set.seed(1)

mean(rbinom(n = 1, prob = 0.45, size = 1))          # 0

mean(rbinom(n = 100, prob = 0.45, size = 1))        # 0.47

mean(rbinom(n = 10000, prob = 0.45, size = 1))      # 0.445

mean(rbinom(n = 1000000, prob = 0.45, size = 1))    # 0.4512
```

Going Bayesian



Prior

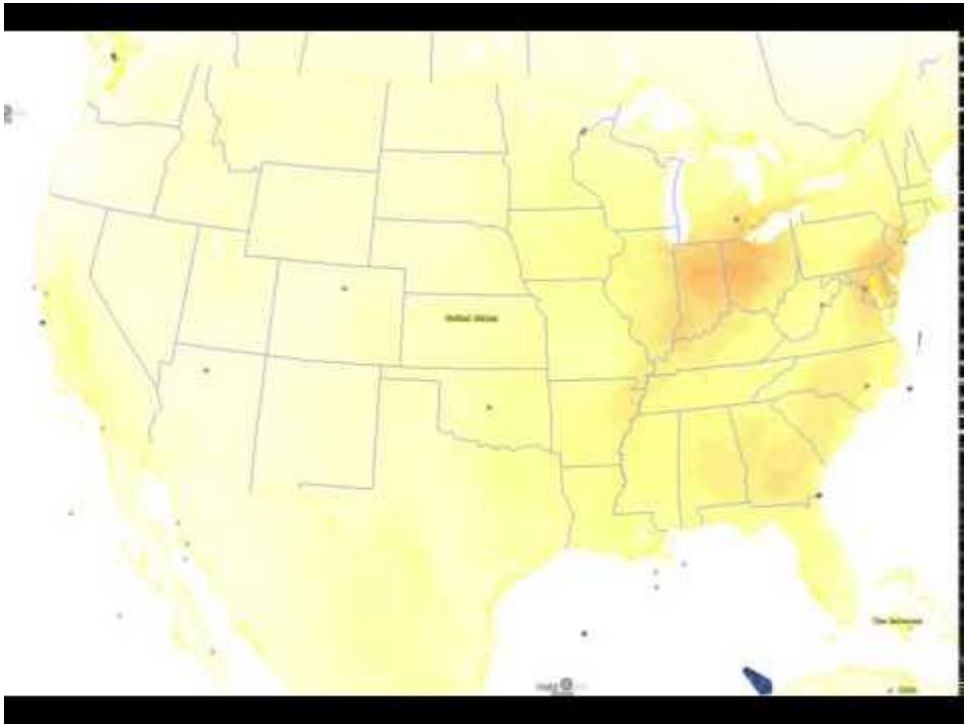
- Uniform: proportion could be anything
- Unimodal centered at 0.25, 0.5, 0.75, …
- ...

Posterior

Credibility intervals

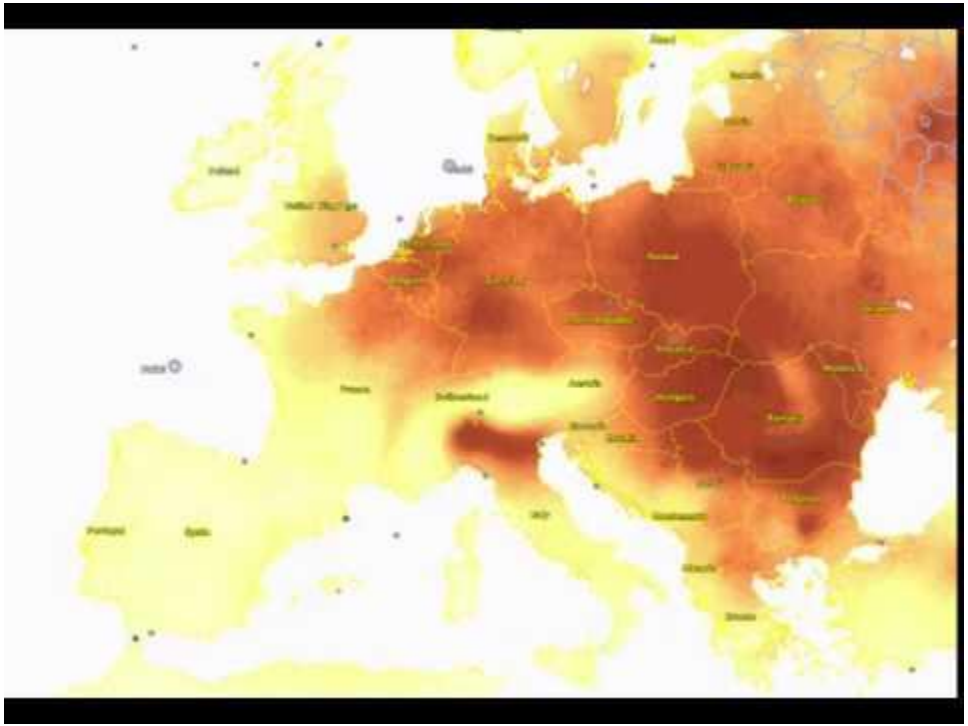
https://goo.gl/aL6MbU

Particulate matter (PM 2.5): 1998-2015



Data from
http://fizz.phys.dal.ca/~atmos/martin/

Particulate matter (PM 2.5): 1998-2015



Data from
http://fizz.phys.dal.ca/~atmos/martin/

What was the impact of…? What if…?