Introduction to Telling Stories with Data

Author: Nicholas G Reich

What are data?

da·tum

/ˈdātəm, ˈdatəm/

noun

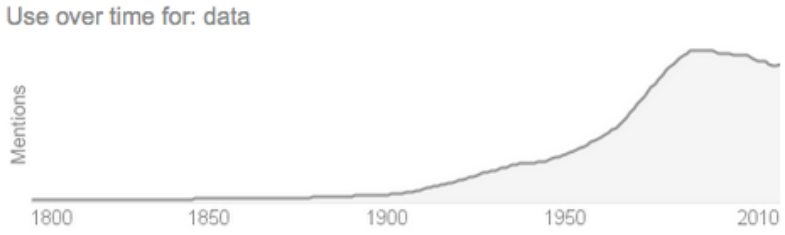
plural noun: **data**

1. a piece of information.
   - an assumption or premise from which inferences may be drawn.
2. a fixed starting point of a scale or operation.

Origin

LATIN	LATIN	
dare	datum	datum
give	something given	*mid 18th century*

mid 18th century: from Latin, literally 'something given,' neuter past participle of *dare* 'give.'



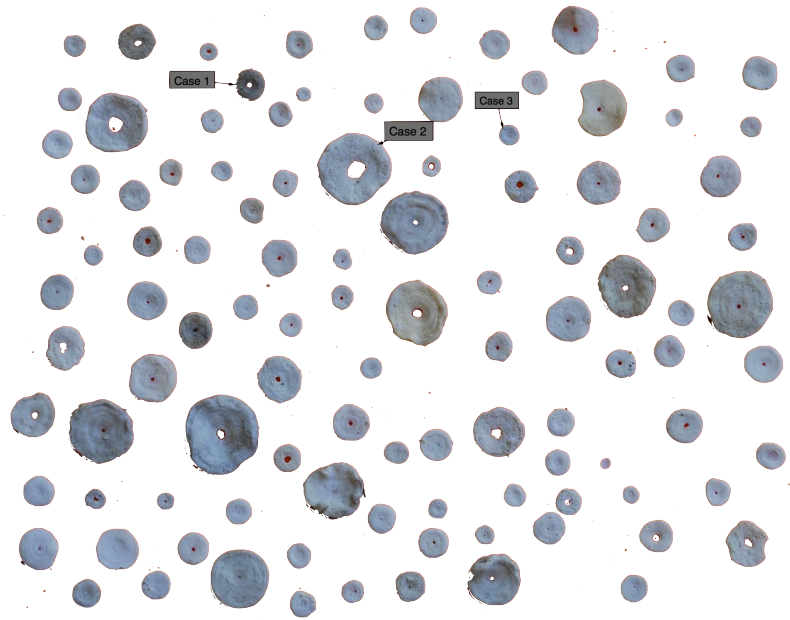
Data as building blocks

- Data are units of information, encoded for processing, e.g. "machine-readable"
- Data, like words, can be woven together to create new conceptual understanding
- Data are the raw material of the digital economy.

Data: cases, variables, and variation[1]

A data set consists of

- **cases**: objects in a collection, or sample
- **variables**: a measurable attribute or quantity of a case



[1] Image and definitions from Kaplan, Statistical Modeling

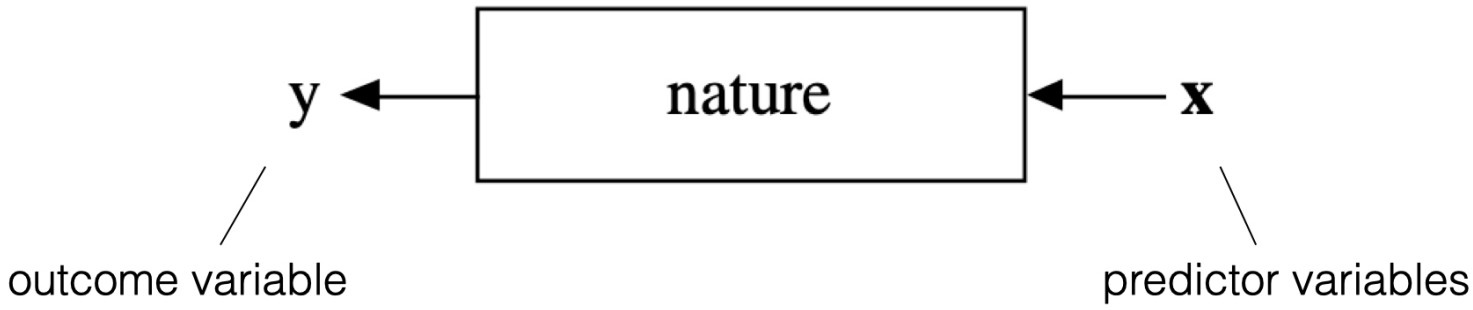
What are models?

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Data arise thanks to the black box of nature.



e.g. covid cases next week in Hampshire county

covid cases this week, behavior, vaccination, cases in all past weeks, ...

What are models?

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

"To extract some information about how nature is associating the response variables to the input variables."

One goal: **infer** something about nature from data.

We want to learn something about the "true" state of nature, but we will never be able to observe what the black box relationships are between all the $\mathbf{x}$ and y.

How do population structure, human behavior, biological features of a pathogen, etc... interact to cause an outbreak?

What are models?

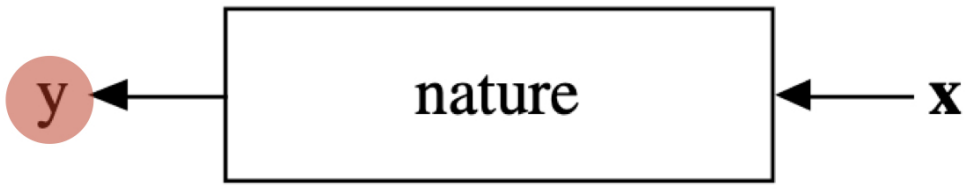
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

"To be able to predict what the responses are going to be to future input variables."

Another goal: **predict** new data.



In prediction, we might be less concerned learning about nature, and more with what the the outcome y will be. If we are careful, we can pick problems and settings where we can (eventually) know the truth about what y will be given some $\mathbf{x}$.

How many cases will be observed next week?

What are models?

A (simplified) representation of a system of inter-relationships.

- Statistical models are a mathematical way to describe that black box of nature.
- Classic quote: "all models are wrong, some are useful."

Mathematically, we can think of models like functions

$$y = f(x_1, x_2, ...) + error$$

Communicating ideas with evidence

What is a narrative? [From the OED]

An account of a series of events, facts, etc., given in order and with the establishing of connections between them; a narration, a story, an account.

- A data-driven narrative communicates the results of a model.
- Important to let the data lead our belief, not the other way around!!

Telling stories with data

- raw material: words, data
- tools: code, software, computers, statistical models
- finished product: visualization, explanations, inference

PUBLISHED JAN. 18, 2021, AT 6:00 AM

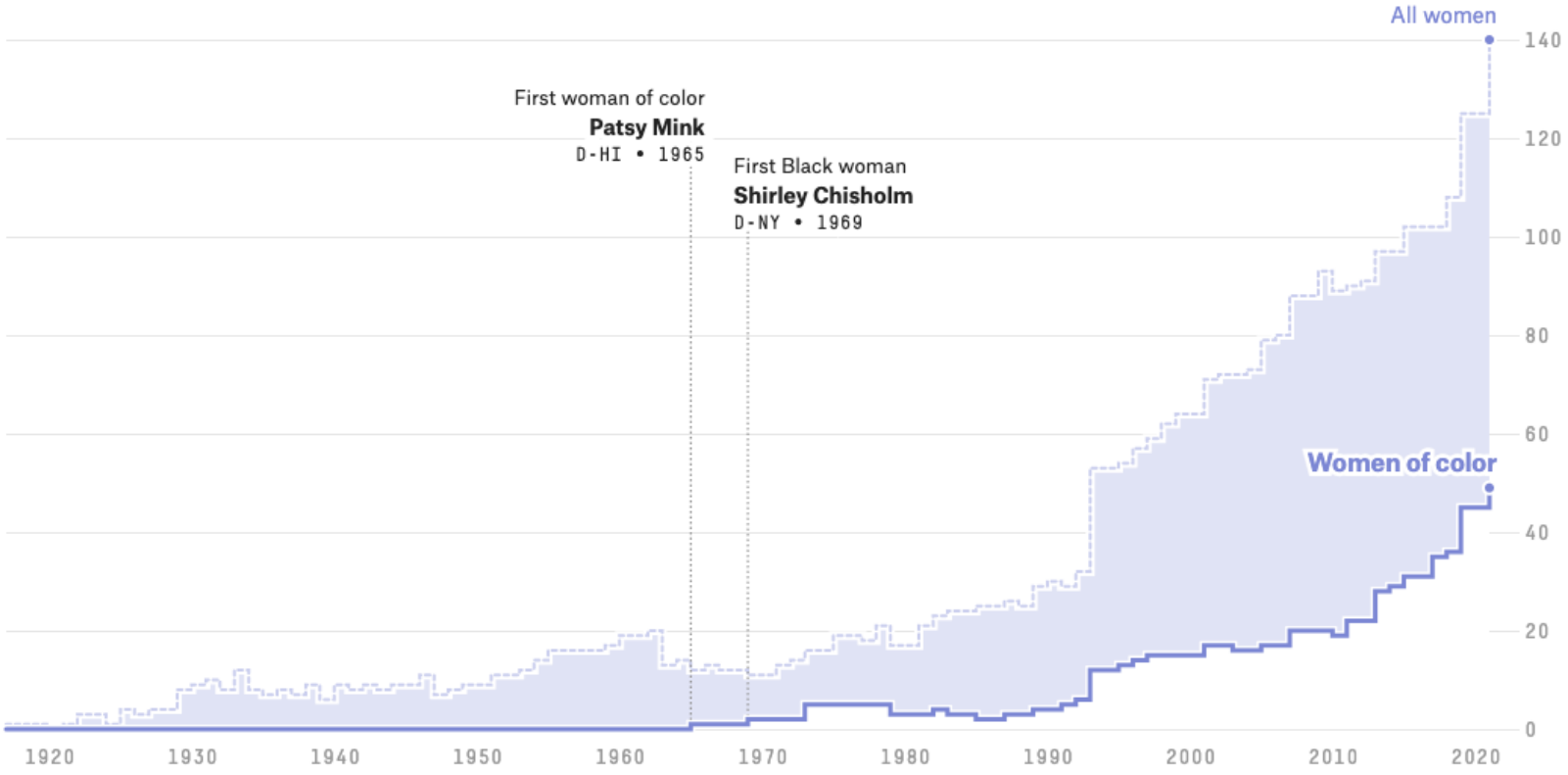
Women of Color Were Shut Out of Congress For Decades. Now They're Transforming It.

By Meredith Conroy, Amelia Thomson-DeVeaux and Anna Wiederkehr

Illustration by Chelsea Alexander

It wasn't until 1965 that a woman of color arrived in Congress

The number of women serving in Congress since 1917

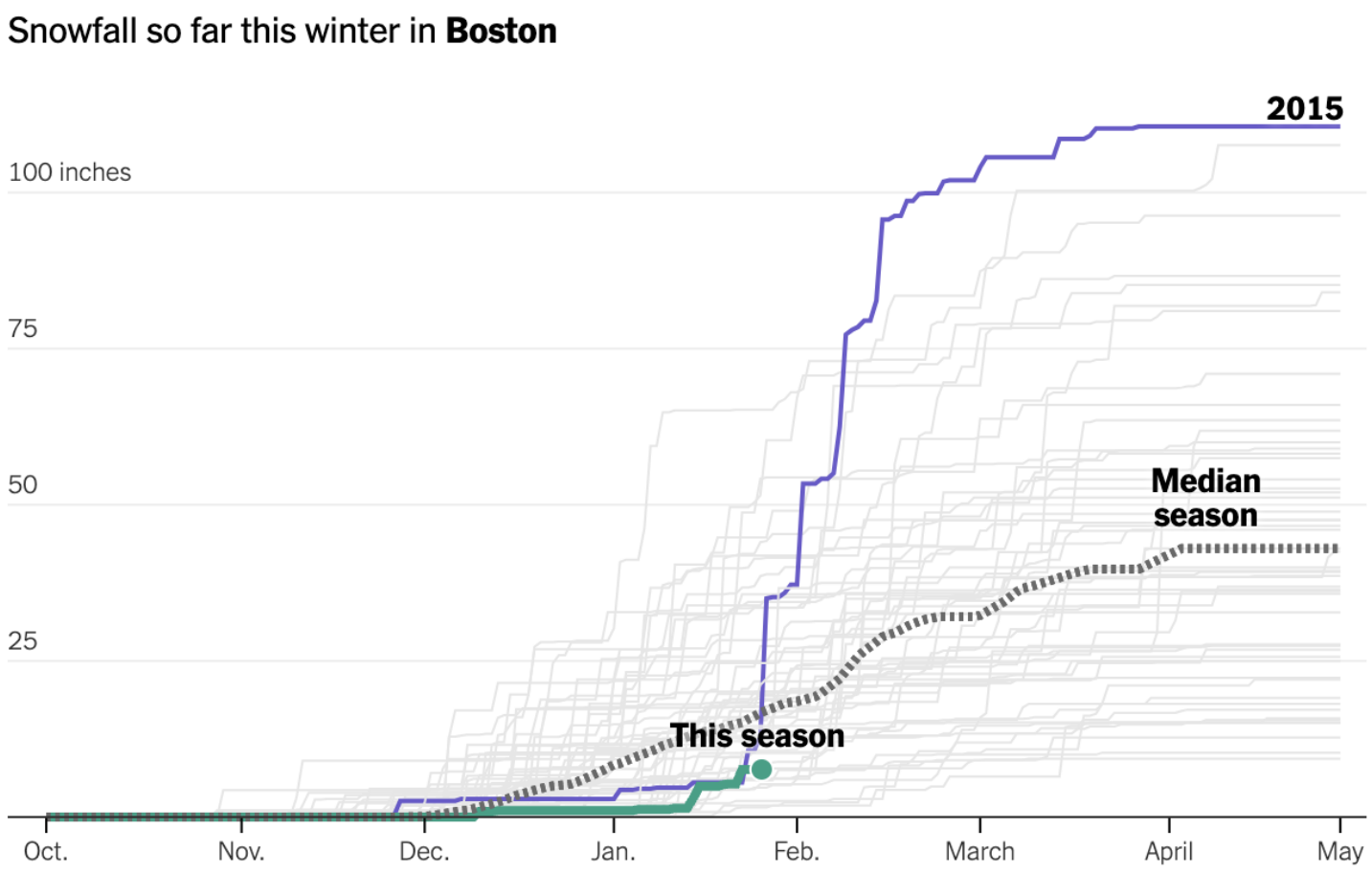


https://projects.fivethirtyeight.com/women-of-color-congress-2020/

TheUpshot

You Call That Snow?! See How This Winter Stacks Up.

By Francesca Paris Jan. 26, 2023

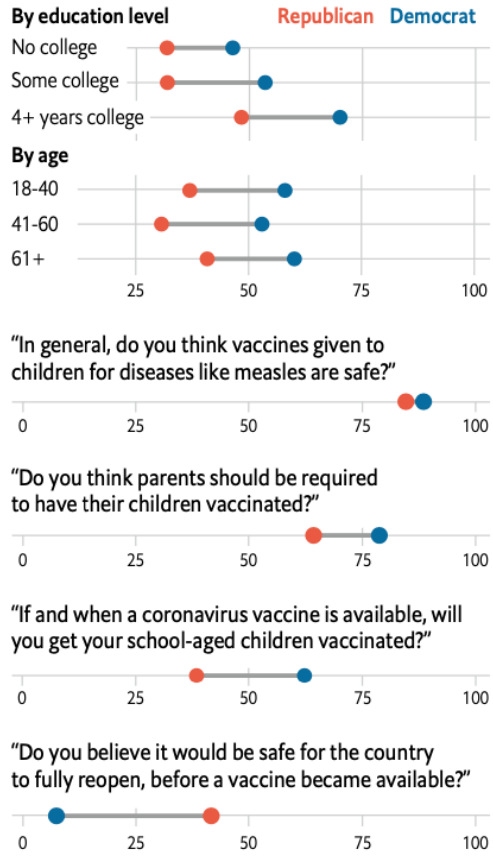


https://www.nytimes.com/interactive/2023/01/26/upshot/city-snowfall-totals.html

→ Rich countries have more trust in science and doctors, but less in vaccines

Trust v GDP per person*, 2018, by country

United States, % responding "yes"
By political alignment, July-August 2020



Sources: Wellcome Trust; World Bank; YouGov/*The Economist*
*At purchasing-power parity

https://www.economist.com/graphic-detail/2020/08/29/conspiracy-theories-about-covid-19-vaccines-may-prevent-herd-immunity

How to tell a story using data

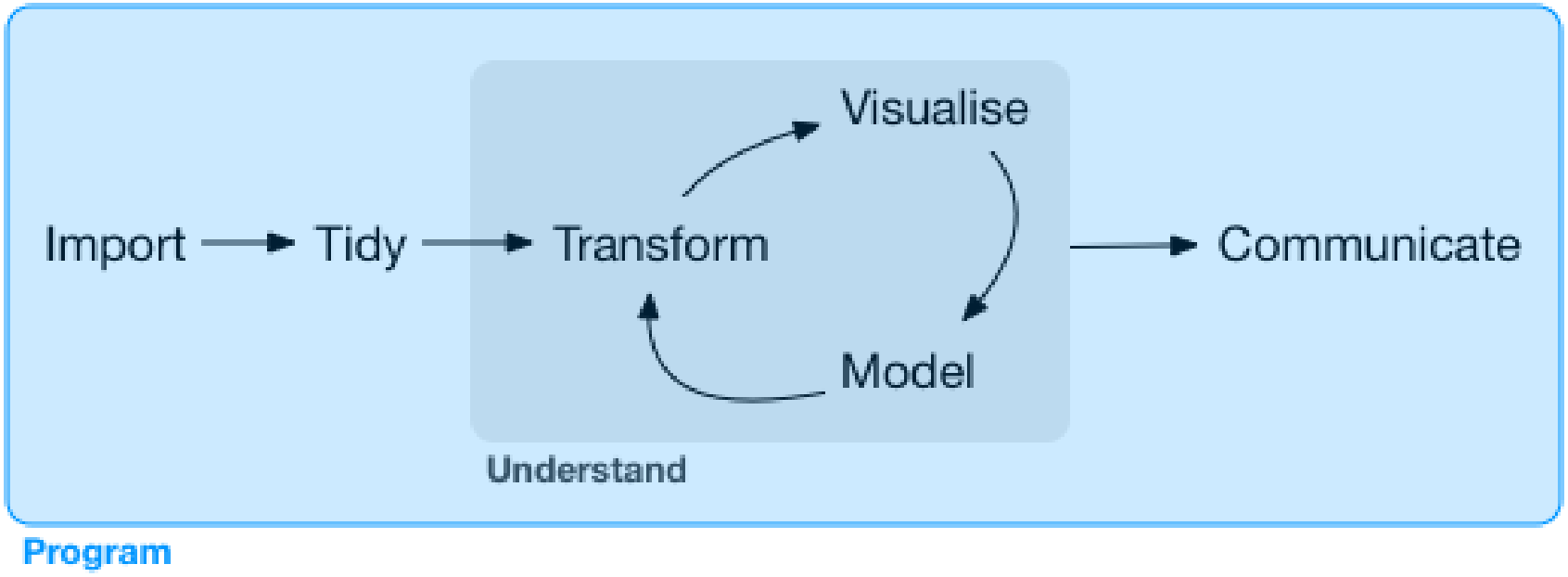
Telling stories with data requires

- a topic for the story you want to tell
- detective work
- creativity, both scientific and artistic
- experimentation with different storylines
- statistical literacy
- good data (good data does not necessarily equal "big data")

Common mistakes in data storytelling

- not knowing the audience
- making the story too complex too quickly
- trying to make it too complicated
- encouraging correlation to be seen as causation
- missing the little things (clear axis labels, plot annnotations)

A process for data analysis



This is where we are going to spend most of our time this semester. With a focus on public health applications.

Figure credits: https://r4ds.had.co.nz/introduction.html