

Introduction to Telling Stories with Data

Author: Nicholas G Reich

*Slides available under the Creative Commons Attribution-ShareAlike 3.0 Unported
License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US*

What are data?

da·tum

/ˈdætəm, ˈdætəm/ ⓘ

noun

plural noun: data

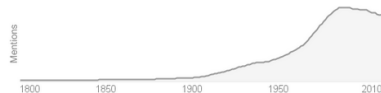
1. a piece of information.
 - an assumption or premise from which inferences may be drawn.
2. a fixed starting point of a scale or operation.

Origin



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

Use over time for: data



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.

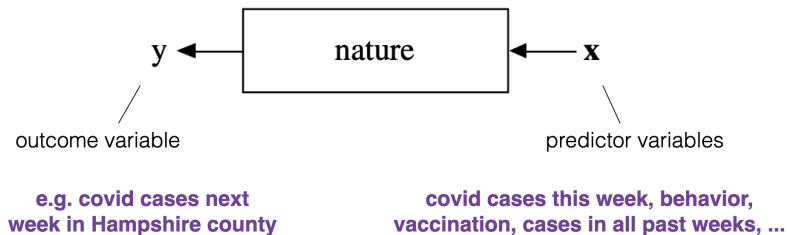
What are models?

Statistical Science
2001, Vol. 16, No. 3, 189-201

Statistical Modeling: The Two Cultures

Leo Breiman

Data arise thanks to the black box of nature.



What are models?

Statistical Science
2001, Vol. 16, No. 3, 189-201

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. 18, No. 3, 119-131

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

$$y = f(x_1, x_2, \dots) + \textit{error}$$

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 8:00 AM

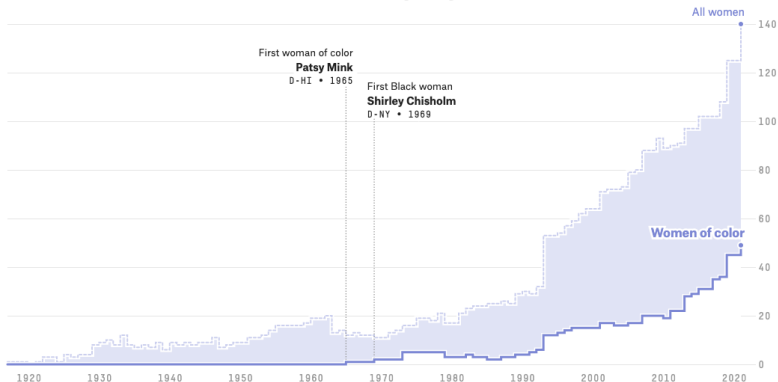
Women of Color Were Shut Out of Congress For Decades. Now They're Transforming It.

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

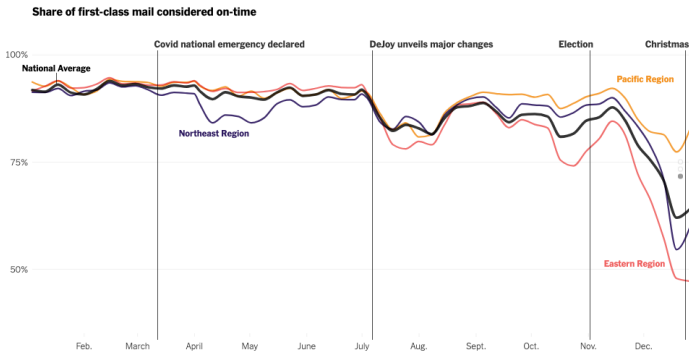
The number of women serving in Congress since 1917



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

The Postal Service Survived the Election. But It Was Crushed by Holiday Packages.

By Emily Badger, Quoctrung Bui and Margot Sanger-Katz Jan. 19, 2021

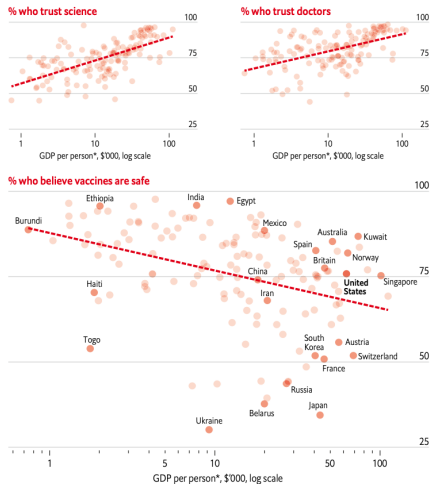


Note: Includes three of the seven Postal Service regions. Region definitions can be found [here](#). Source: United States Postal Service

<https://www.nytimes.com/interactive/2021/01/19/upshot/postal-service-survived-election-but-crushed-by-holidays.html>

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

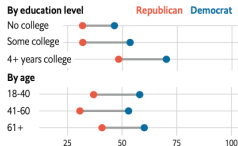
Trust v GDP per person*, 2018, by country



United States, % responding "yes"

By political alignment, July-August 2020

"If and when a coronavirus vaccine becomes available, will you get vaccinated?"



"In general, do you think vaccines given to children for diseases like measles are safe?"



"Do you think parents should be required to have their children vaccinated?"



"If and when a coronavirus vaccine is available, will you get your school-aged children vaccinated?"



"Do you believe it would be safe for the country to fully reopen, before a vaccine became available?"



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

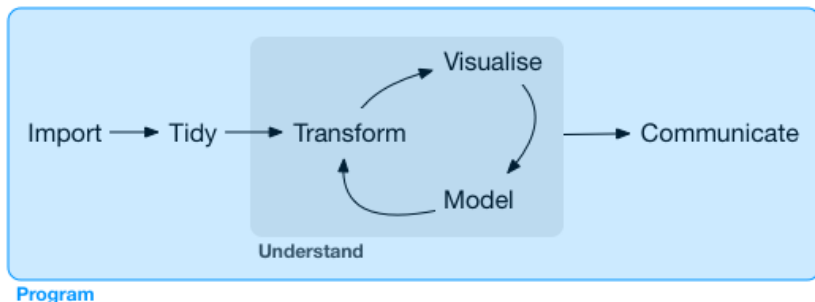
Telling stories with data requires

- ▶ 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 annotations)

A process for data analysis



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

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