Teaching the Basics of Data Journalism
Part 1: Thinking About Data

Mai Nguyen

October 2020

QTM and Emory Writing Center

Introduction

- Background:
  - PhD in Political Science from NYU
  - Worked as the Quantitative Editor for FiveThirtyEight
  - Currently work as a data scientist for Giant Oak

- Other fun facts:
  - Born and raised in Nebraska
  - Obsessesed with GBBO
  - Have a ridiculous collection of board games

Introduction

- Background:
  - PhD in Political Science from NYU
  - Worked as the Quantitative Editor for FiveThirtyEight
  - Currently work as a data scientist for Giant Oak

- Other fun facts:
  - Born and raised in Nebraska
  - Obsessesed with GBBO
  - Have a ridiculous collection of board games

What I want you to get out of this workshop

Introduction and resources on how to help you incorporate data journalism into the classroom:

1. Thinking About Data
2. Communicating Data

What I want you to get out of this workshop

Introduction and resources on how to help you incorporate data journalism into the classroom:

1. Thinking About Data
2. Communicating Data

Outline

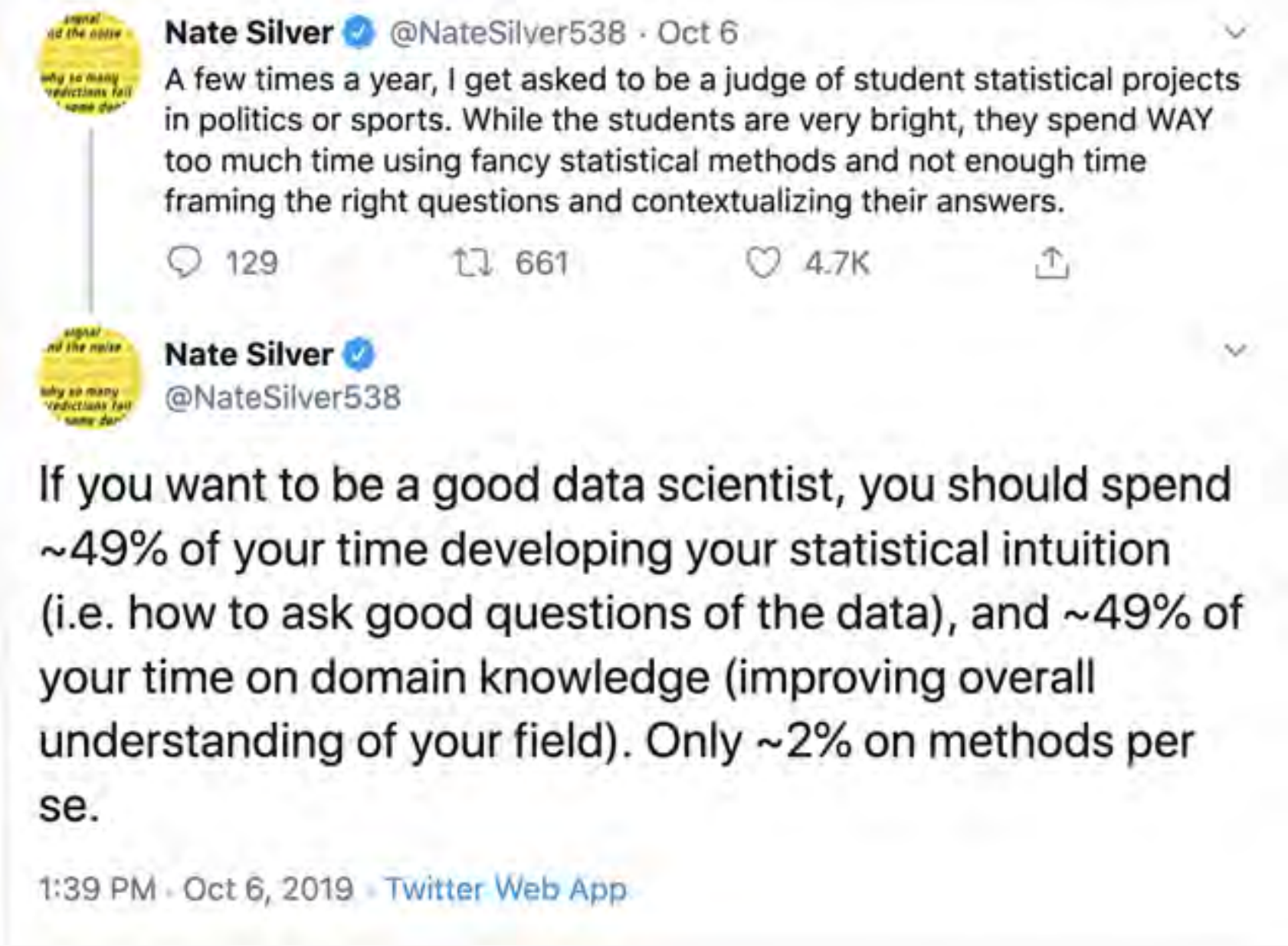
- What is data journalism?
- How to think about data
    - Data generating process
    - Measuring and conceptualizing data
    - Types of data stories
    - Potential pitfalls in data analysis

What is Data Journalism?

What is data journalism?

$\rightarrow$ Story-telling that involves data.

What skills are needed?



Nate Silver @NateSilver538 · Oct 6

A few times a year, I get asked to be a judge of student statistical projects in politics or sports. While the students are very bright, they spend WAY too much time using fancy statistical methods and not enough time framing the right questions and contextualizing their answers.

129 661 4.7K

Nate Silver
@NateSilver538

If you want to be a good data scientist, you should spend ~49% of your time developing your statistical intuition (i.e. how to ask good questions of the data), and ~49% of your time on domain knowledge (improving overall understanding of your field). Only ~2% on methods per se.

1:39 PM · Oct 6, 2019 · Twitter Web App

What skills are needed?

- In terms of methodology...
  - Is it nice to have some advanced stats skills? Yes
  - Don’t underestimate the power of simple summary statistics
  - In reality, most people already have the potential to create good stories using data

- The most important skill to have is the ability to **understand** data

What skills are needed?

- In terms of methodology...
  - Is it nice to have some advanced stats skills? Yes
  - Don’t underestimate the power of simple summary statistics
  - In reality, most people already have the potential to create good stories using data

- The most important skill to have is the ability to **understand** data

What skills are needed?

- In terms of methodology...
  - Is it nice to have some advanced stats skills? Yes
  - Don’t underestimate the power of simple summary statistics
  - In reality, most people already have the potential to create good stories using data

- The most important skill to have is the ability to **understand** data

What skills are needed?

- In terms of methodology...
  - Is it nice to have some advanced stats skills? Yes
  - Don’t underestimate the power of simple summary statistics
  - In reality, most people already have the potential to create good stories using data

- The most important skill to have is the ability to **understand** data

What skills are needed?

- In terms of methodology...
  - Is it nice to have some advanced stats skills? Yes
  - Don’t underestimate the power of simple summary statistics
  - In reality, most people already have the potential to create good stories using data

- The most important skill to have is the ability to **understand** data

How to think about data

Understanding how data is generated

Data = *information that is collected together for the purpose of examination, discussion or calculation*

- Data is contructed - by people!
- Different incentives. Data is in of itself political, social, etc.
- Healthy dose of skepticism is always required

Understanding how data is generated

Data = *information that is collected together for the purpose of examination, discussion or calculation*

- Data is contructed - by people!
- Different incentives. Data is in of itself political, social, etc.
- Healthy dose of skepticism is always required

Understanding how data is generated

Data = *information that is collected together for the purpose of examination, discussion or calculation*

- Data is contructed - by people!
- Different incentives. Data is in of itself political, social, etc.
- Healthy dose of skepticism is always required

Understanding how data is generated

Data = *information that is collected together for the purpose of examination, discussion or calculation*

- Data is contructed - by people!
- Different incentives. Data is in of itself political, social, etc.
- Healthy dose of skepticism is always required

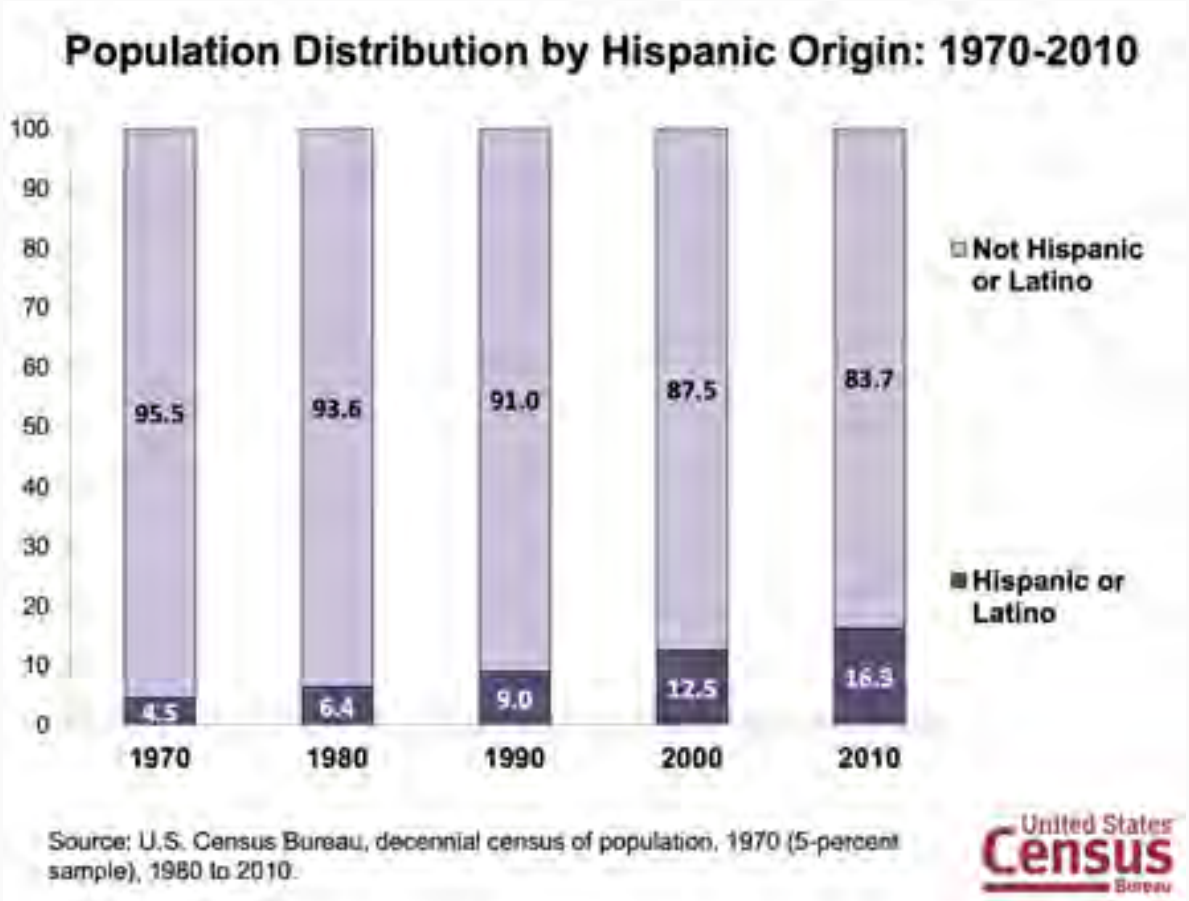
Understanding how data is generated

Fun fact: There were no Hispanic people in the United States prior to 1970

Understanding how data is generated

Fun fact: There were no Hispanic people in the United States prior to 1970, according to the U.S. Census Bureau...

Understanding how data is generated



$\rightarrow$ The U.S. Census didn't start consistently including a Hispanic or Latino option until 1970

Understanding how data is generated

- We can’t measure something we have not conceptualized
- Even when we have, issues abound when it comes to operationalizing and measuring concepts

Understanding how data is generated

- We can’t measure something we have not conceptualized
- Even when we have, issues abound when it comes to operationalizing and measuring concepts

Operationalizing and Measuring Data

Let's take a look at the example of hate crimes:

> FBI: *A hate crime is a traditional offense like murder, arson, or vandalism with an added element of bias... "criminal offense against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity."*

Operationalizing and Measuring Data

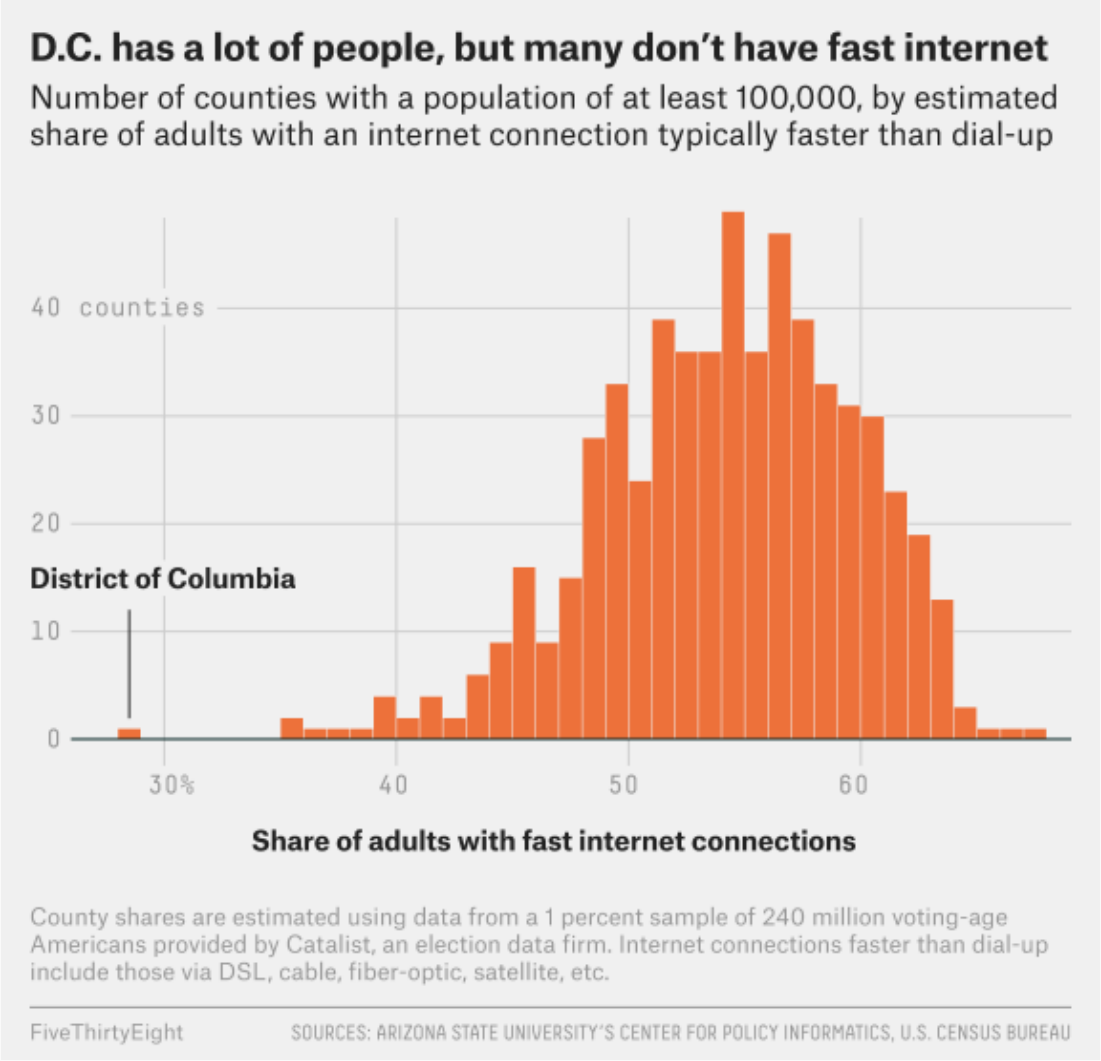
Let's take a look at the example of hate crimes:

> FBI: *A hate crime is a traditional offense like murder, arson, or vandalism with an added element of bias... "criminal offense against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity."*

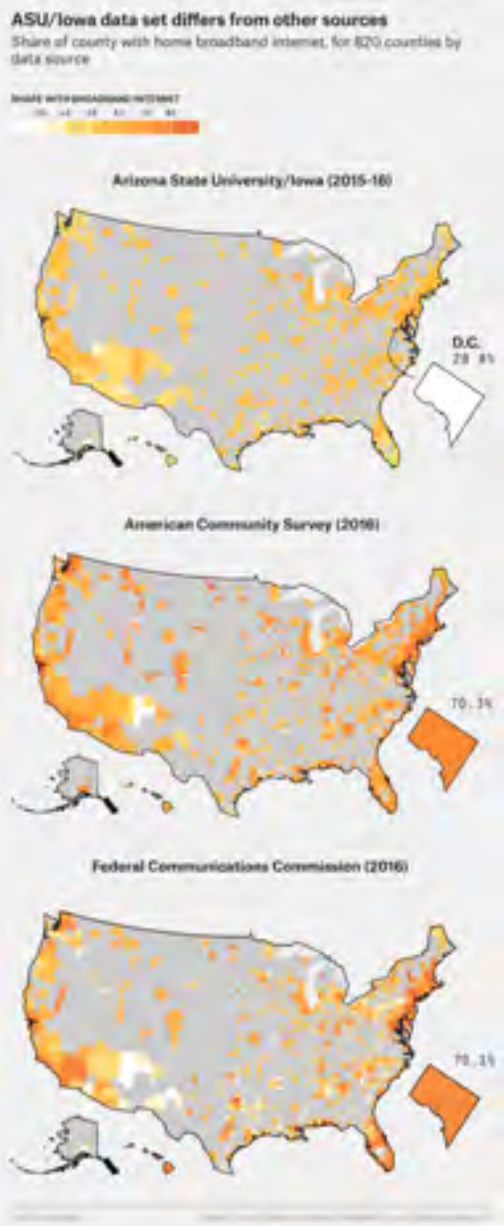
Operationalizing and Measuring Data

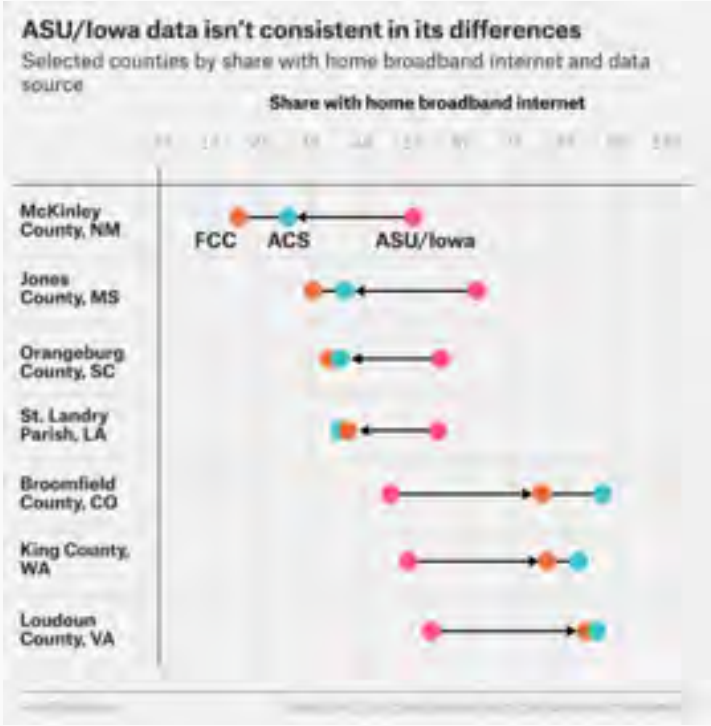
Hard Lessons Learned

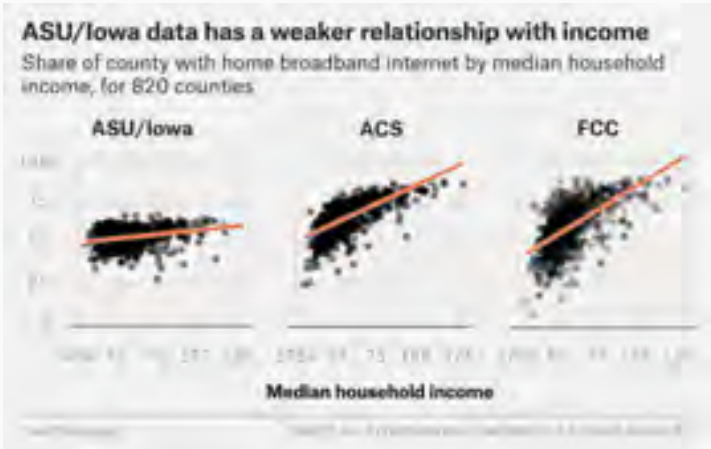
→ Broadband Series at FiveThirtyEight:



Hard Lessons Learned







Hard Lessons Learned

From the original codebook:

"Denotes interest in 'high tech' products and/or services as reported via Share Force. This would include personal computers and internet service providers. Blended with modeled data."

$\rightarrow$ Quite different than what was expected

Hard Lessons Learned

From the original codebook:

"Denotes interest in 'high tech' products and/or services as reported via Share Force. This would include personal computers and internet service providers. Blended with modeled data."

$\rightarrow$ Quite different than what was expected

Hard Lessons Learned

- Understanding how the data were generated is vital
  - Know how things are defined and measured
  - Understand how the data were collected
- Diligence and a healthy dose of skepticism is invaluable
  - Perform sanity checks
  - Cross-reference against other data sources

Hard Lessons Learned

- Understanding how the data were generated is vital
  - Know how things are defined and measured
  - Understand how the data were collected
- Diligence and a healthy dose of skepticism is invaluable
  - Perform sanity checks
  - Cross-reference against other data sources

Hard Lessons Learned

- Understanding how the data were generated is vital
  - Know how things are defined and measured
  - Understand how the data were collected
- Diligence and a healthy dose of skepticism is invaluable
  - Perform sanity checks
  - Cross-reference against other data sources

Hard Lessons Learned

- Understanding how the data were generated is vital
  - Know how things are defined and measured
  - Understand how the data were collected
- Diligence and a healthy dose of skepticism is invaluable
  - Perform sanity checks
  - Cross-reference against other data sources

Different types of data stories

- ***Measurement***
- ***Trends***
- ***Comparisons***
- ***Factors/Composition***
- ***Relationships***

Types of Data Stories

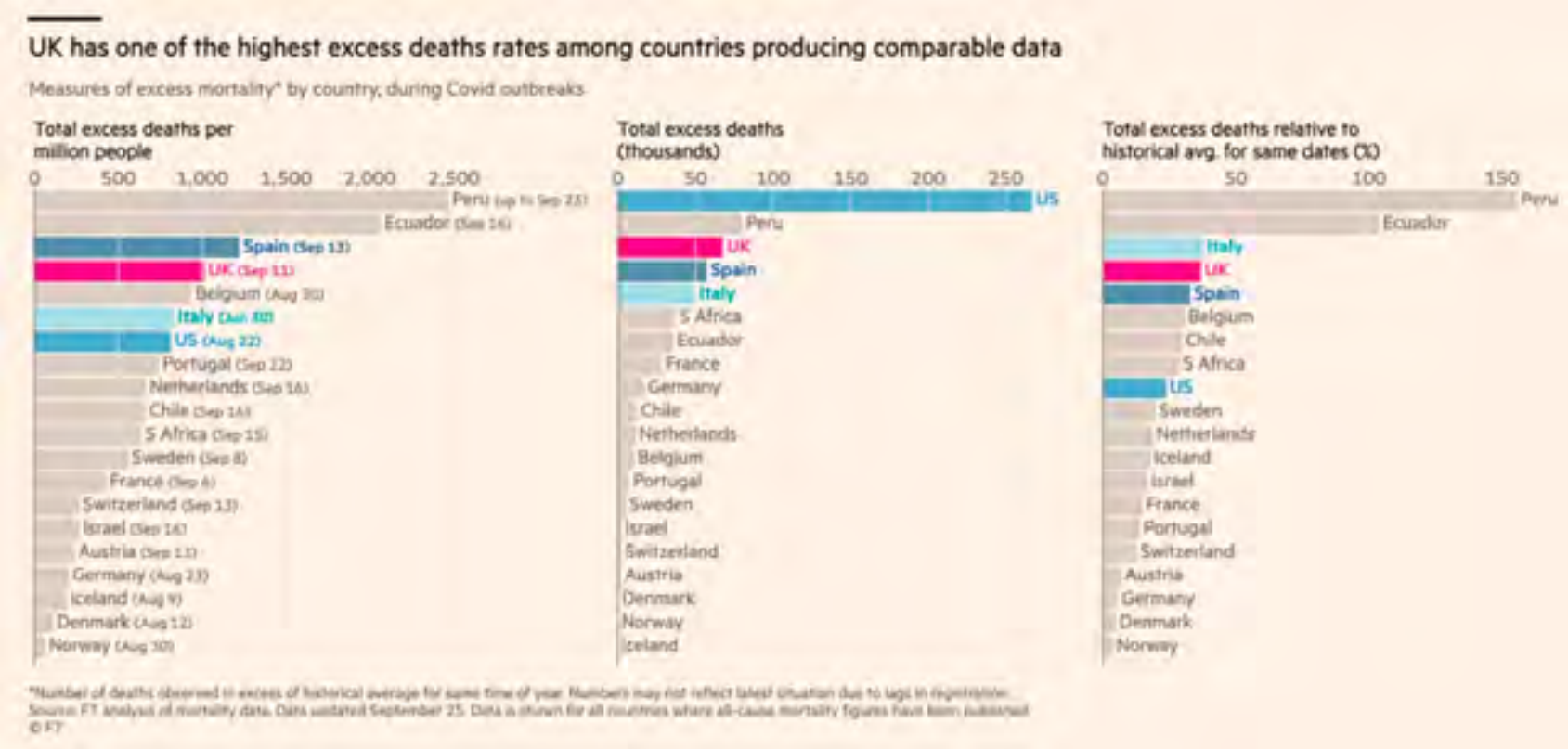
- ***Measurement***: How many people have been affected by coronavirus and how prevalent is it?

Types of Data Stories

- ***Trends***: How does the number of coronavirus cases change over time?

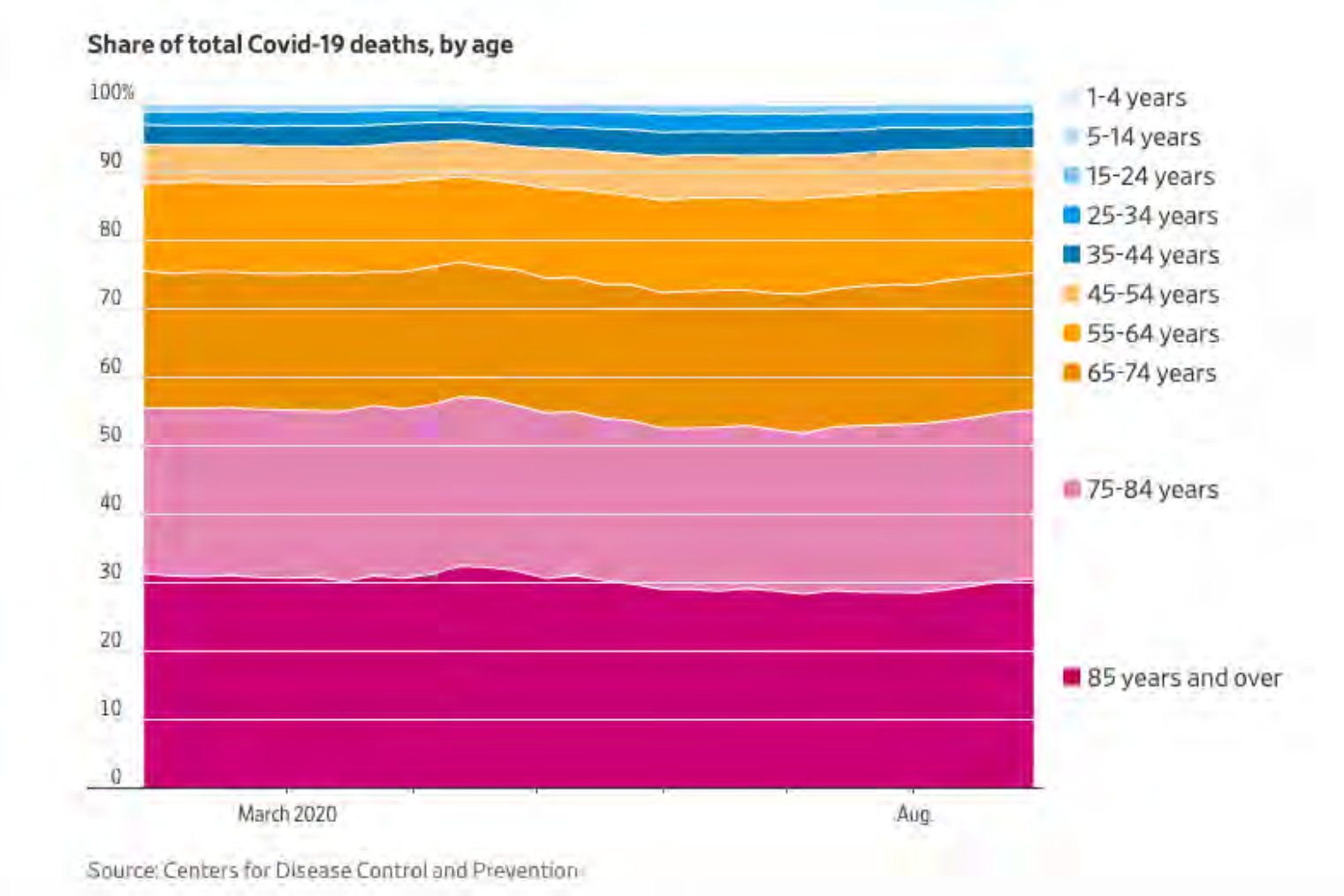
Types of Data Stories

- ***Comparisons***: How does coronavirus fatality differ between countries?



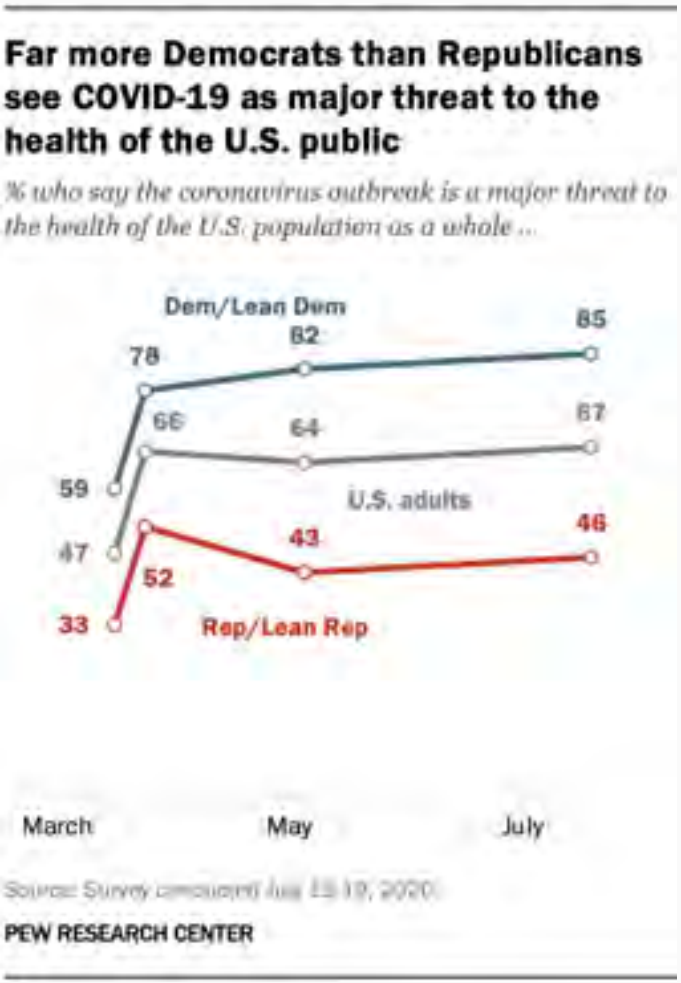
Types of Data Stories

- ***Factors/Composition***: What is the breakdown of people affected by coronavirus?



Types of Data Stories

- ***Relationships***: What affects people's attitudes towards coronavirus?



Understanding relationships within data

What the data are really telling us...

Let’s go over some potential pitfulls

Understanding relationships within data

What the data are really telling us...

Let’s go over some potential pitfulls

Everyone knows this...

Correlation does not imply causation

But why?

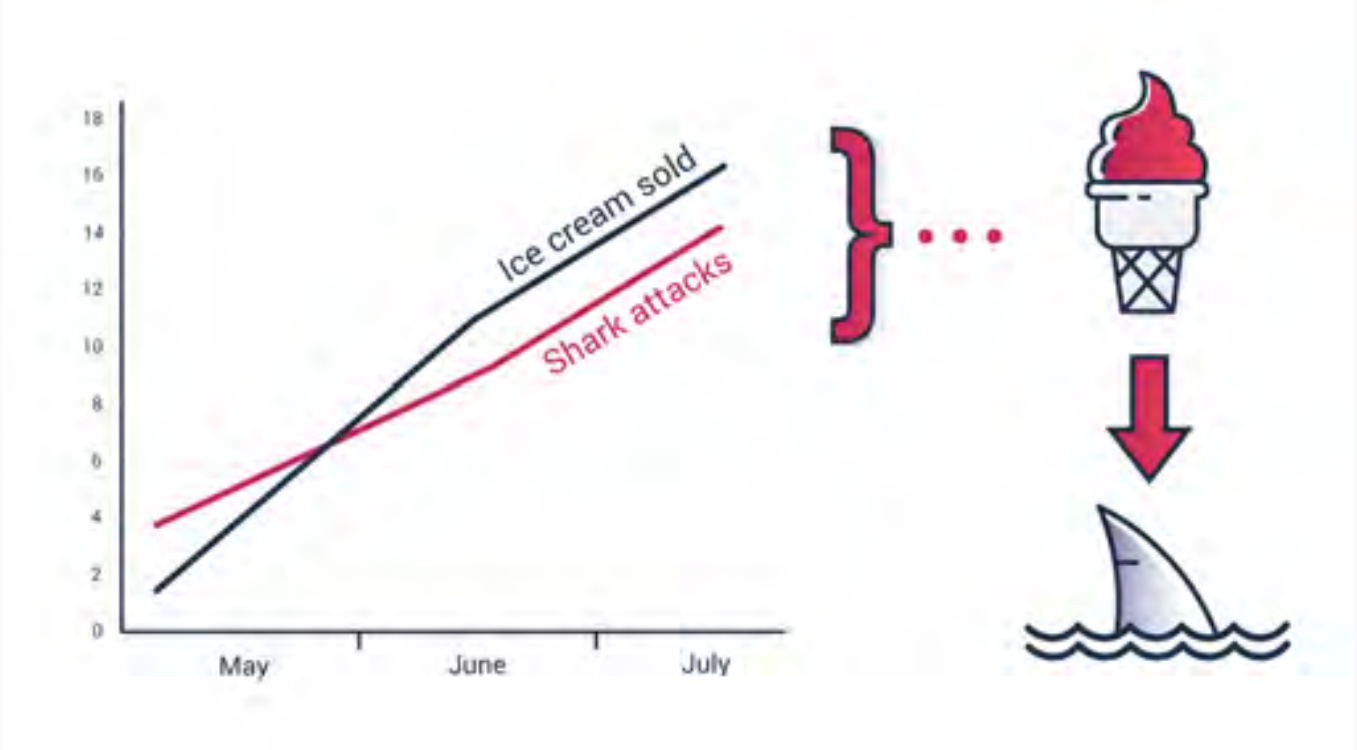
Everyone knows this...

Correlation does not imply causation

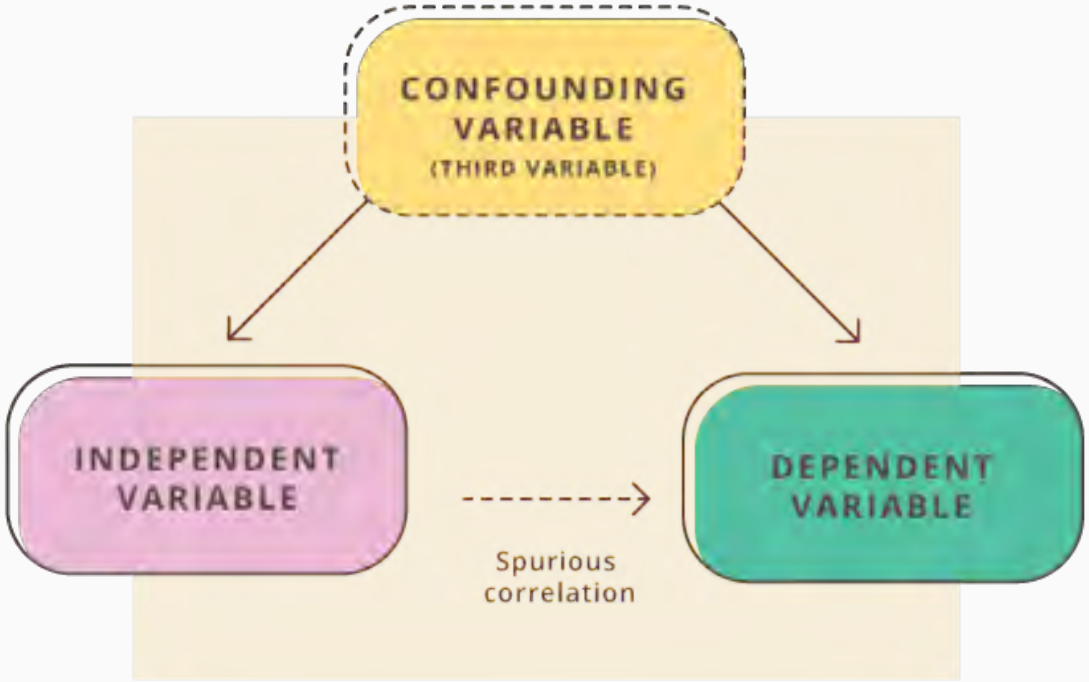
But why?

Confounders

Relationship between Ice Cream and Shark Attacks

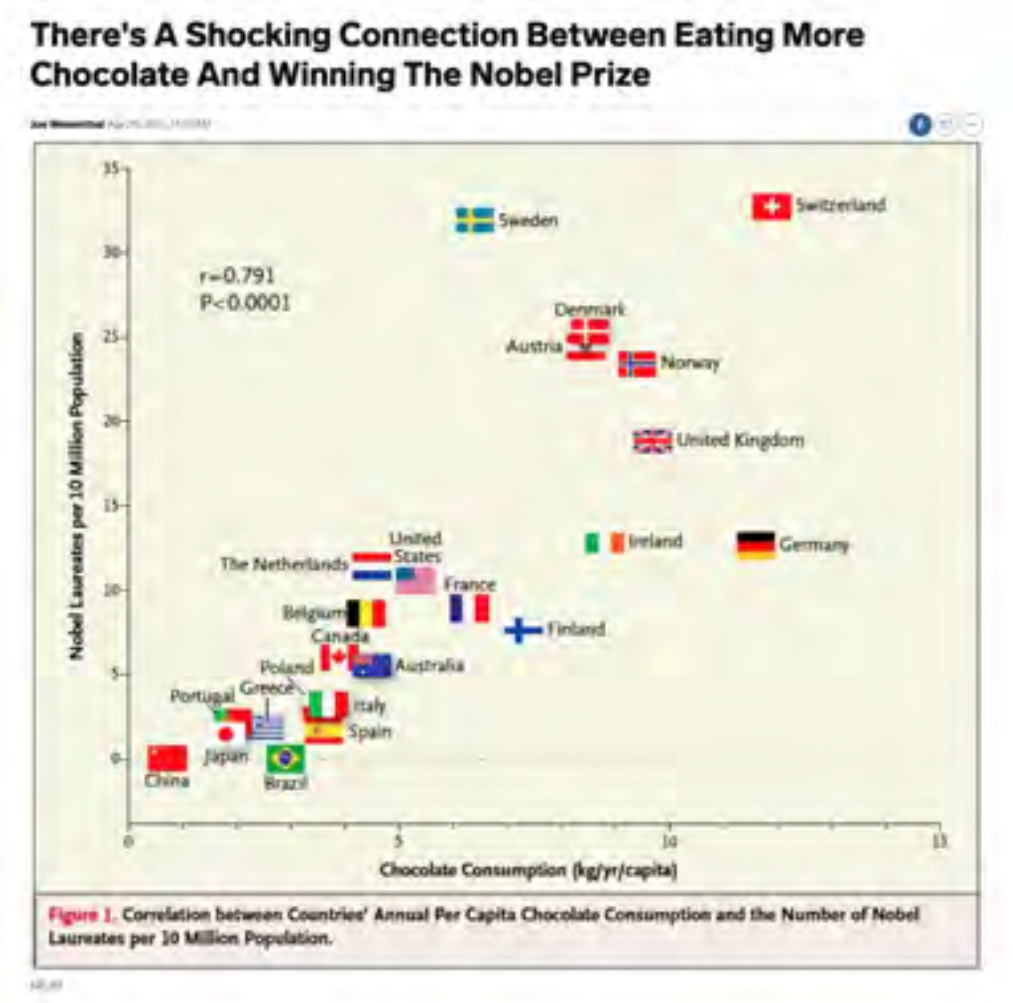


Confounders



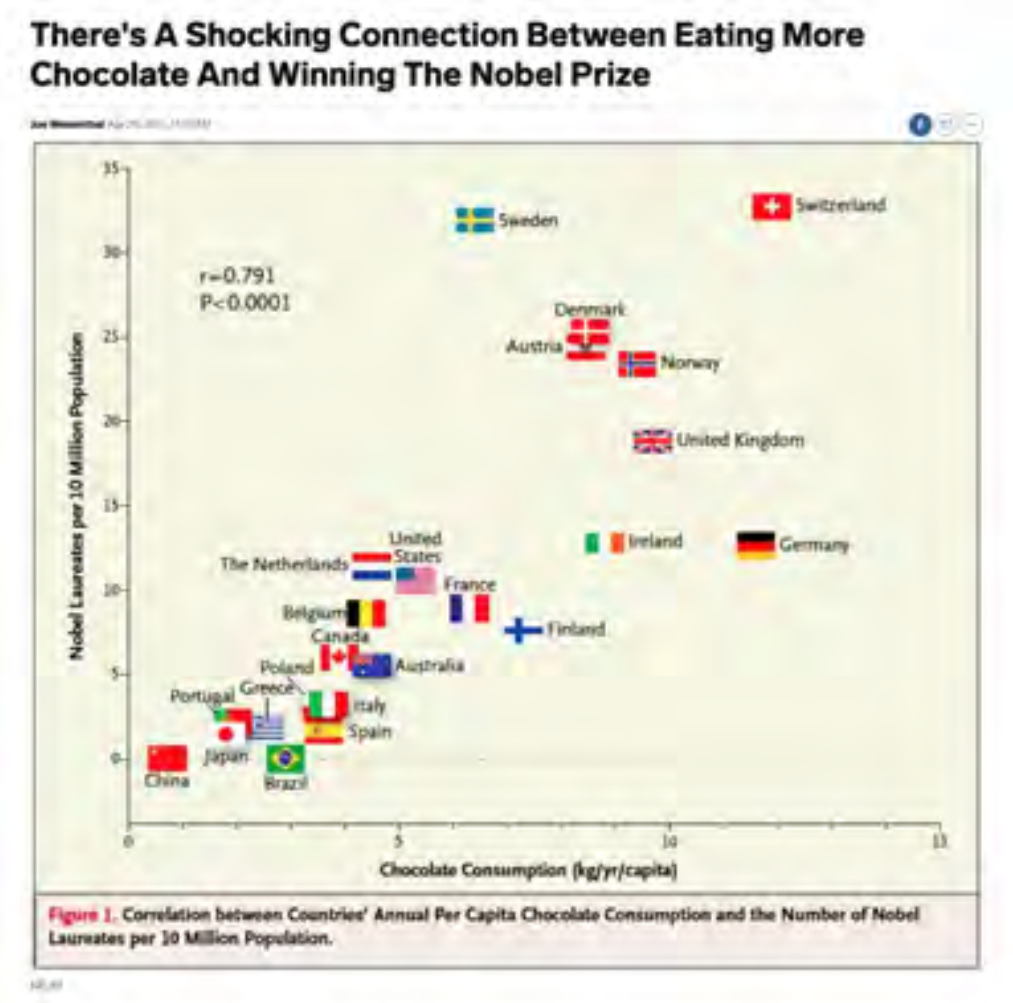
→ A confounder is a variable that influences both the independent variable and dependent variable, that leads us to see a *spurious correlation*

Confounders



→ What could be a potential confounder here?

Confounders



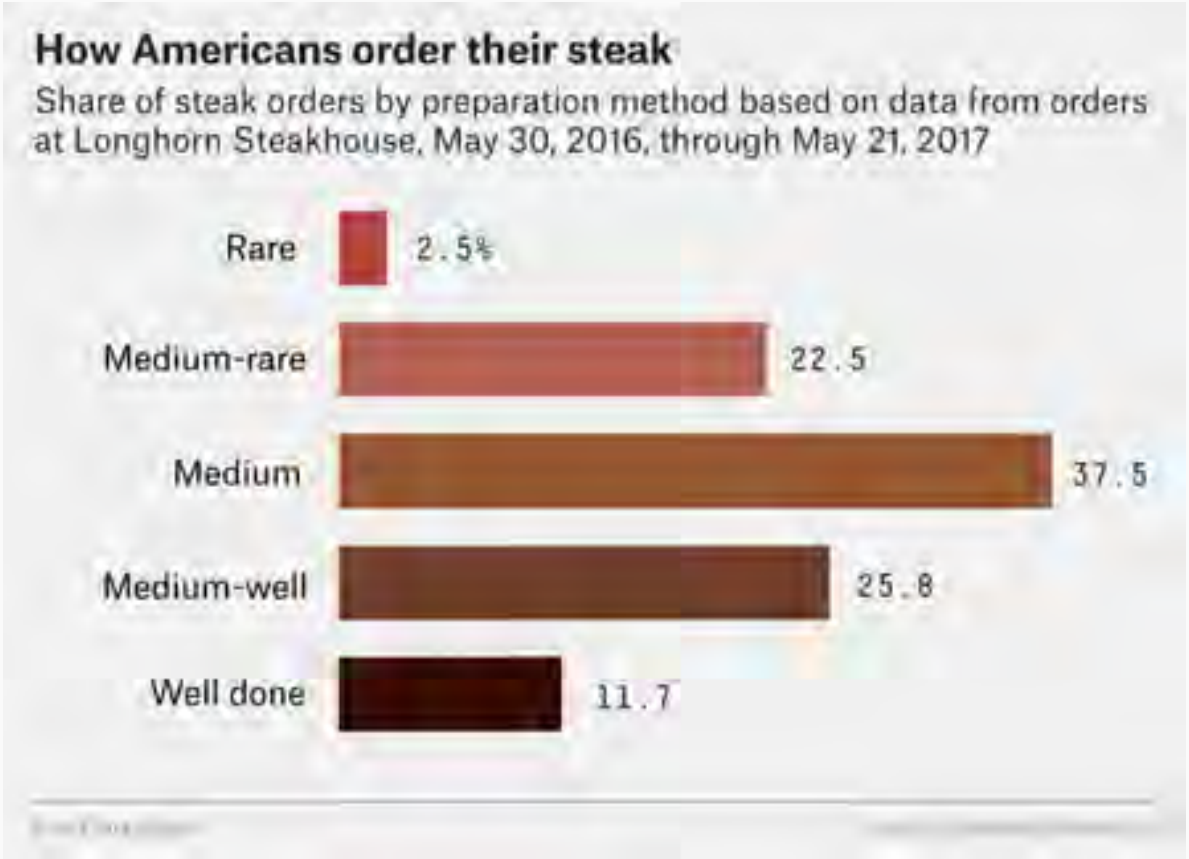
→ What could be a potential confounder here?

Issues with Selection

I've Interviewed 300 High Achievers About Their Morning Routines. Here's What I've Learned.

Your morning routine should suit your needs, but there are some habits everyone should try.

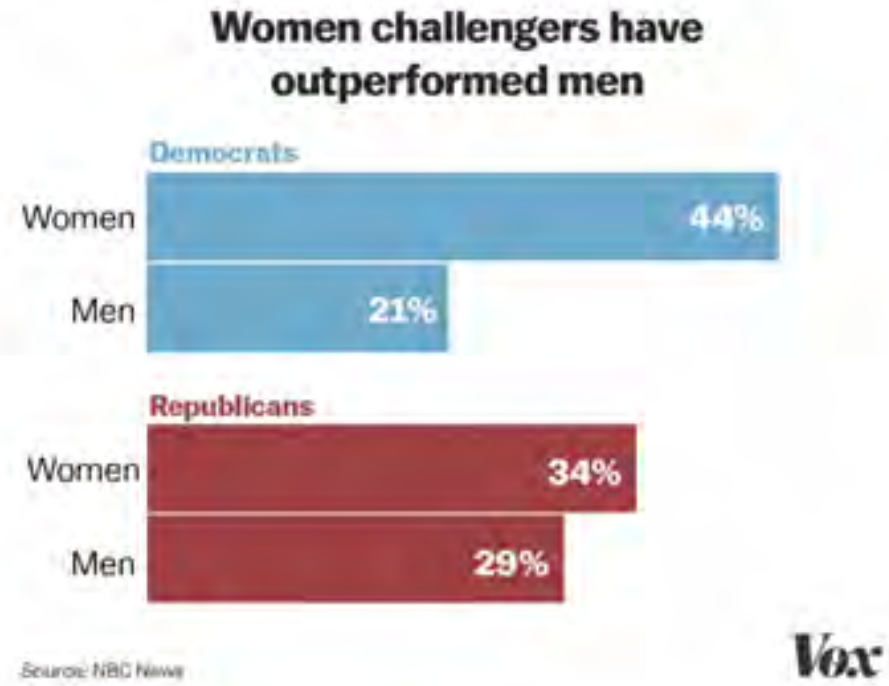
Issues with Selection



Issues with Selection

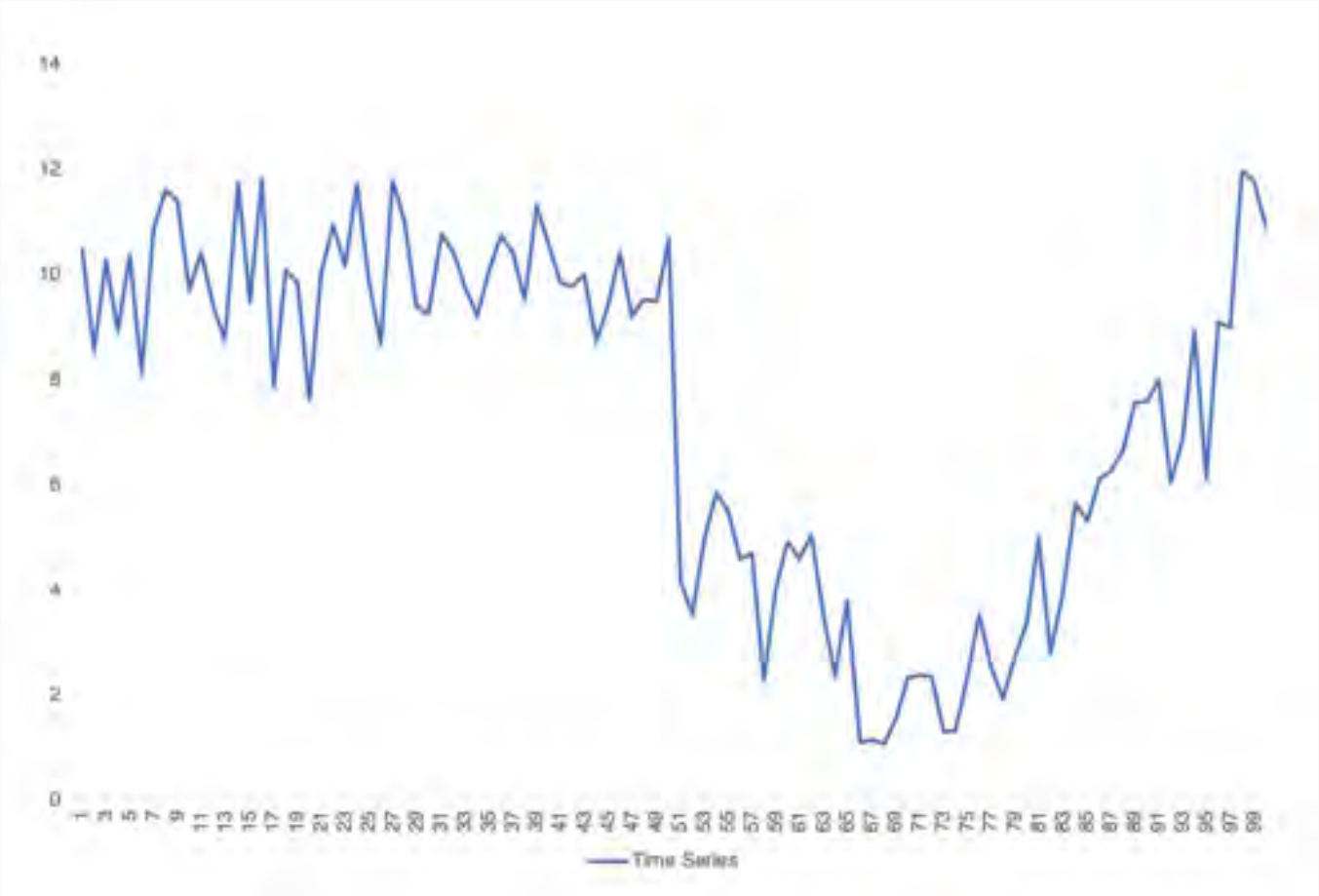
Women are outperforming men

One of the most interesting dynamics to emerge in this year's primaries is that women candidates consistently did better than their male counterparts. **According to an NBC News analysis**, 44 percent of non-incumbent Democratic women won their primaries, compared to 21 percent of Democratic men. Similarly, non-incumbent Republican women have also outperformed their male counterparts, with 34 percent of non-incumbent Republican women winning versus 29 percent of Republican men.

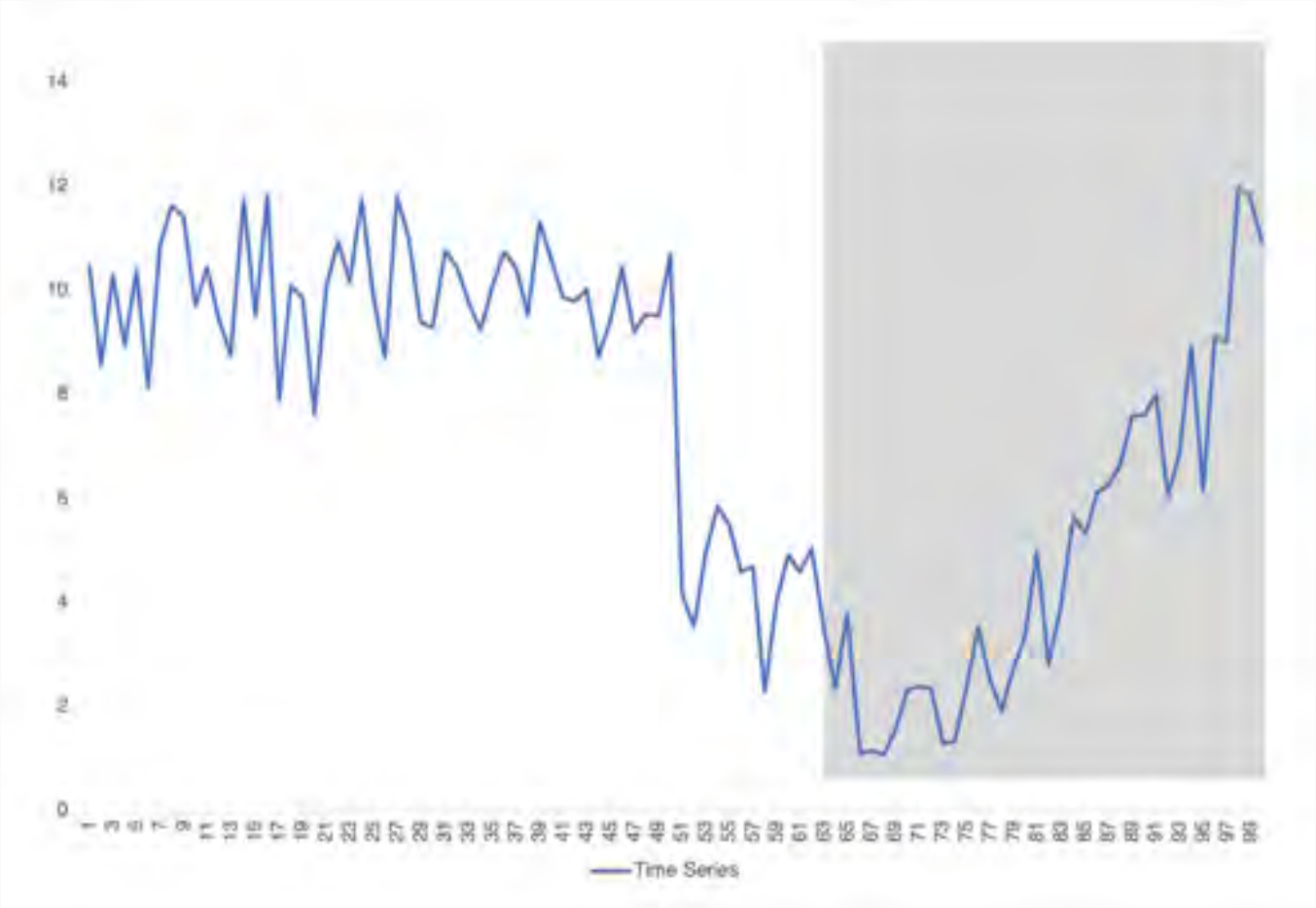


As Vox's Ella Nilsen reported, "[Democratic] female candidates in 2018 are **more likely to defeat male candidates** than the other way around."

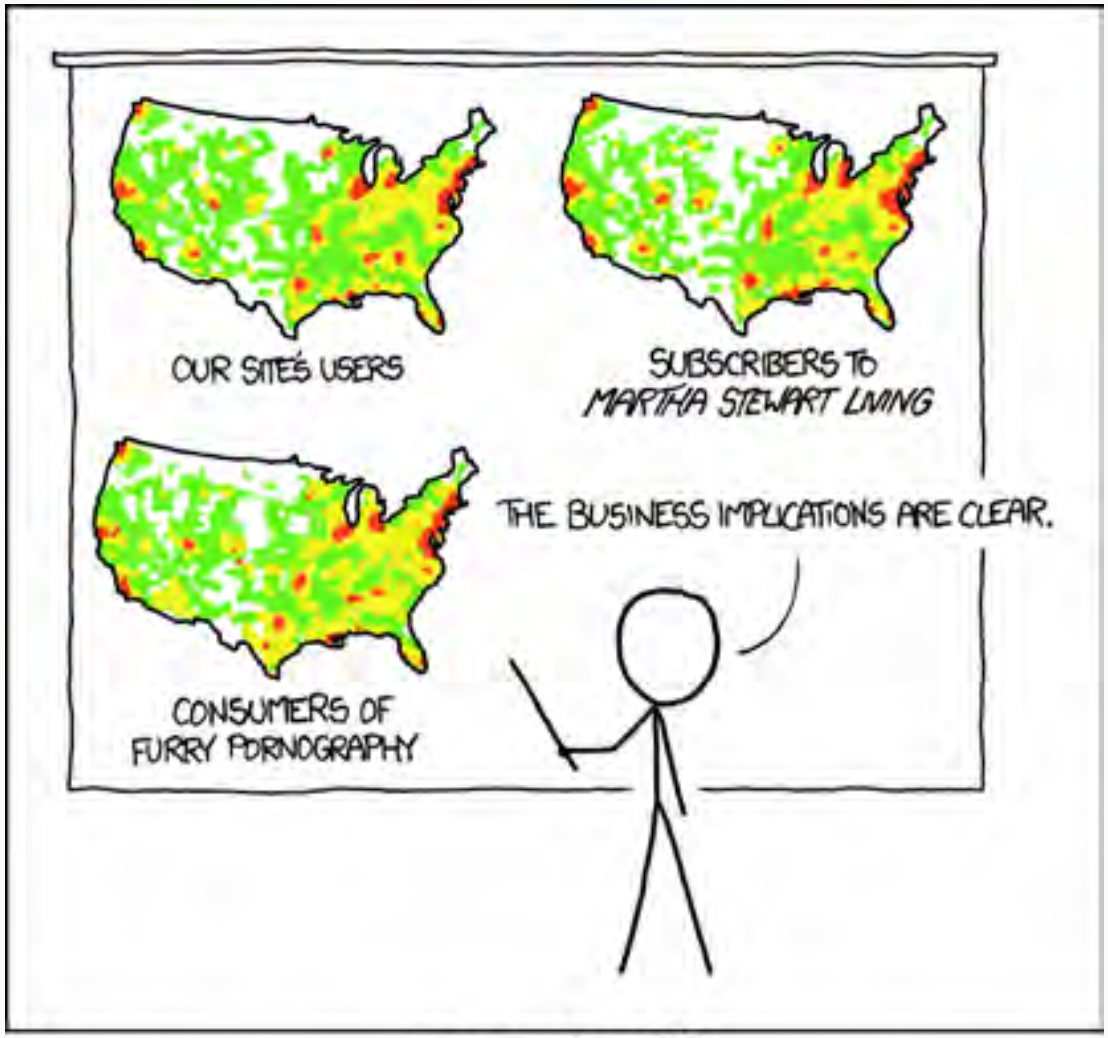
Beware of Cherry-picking



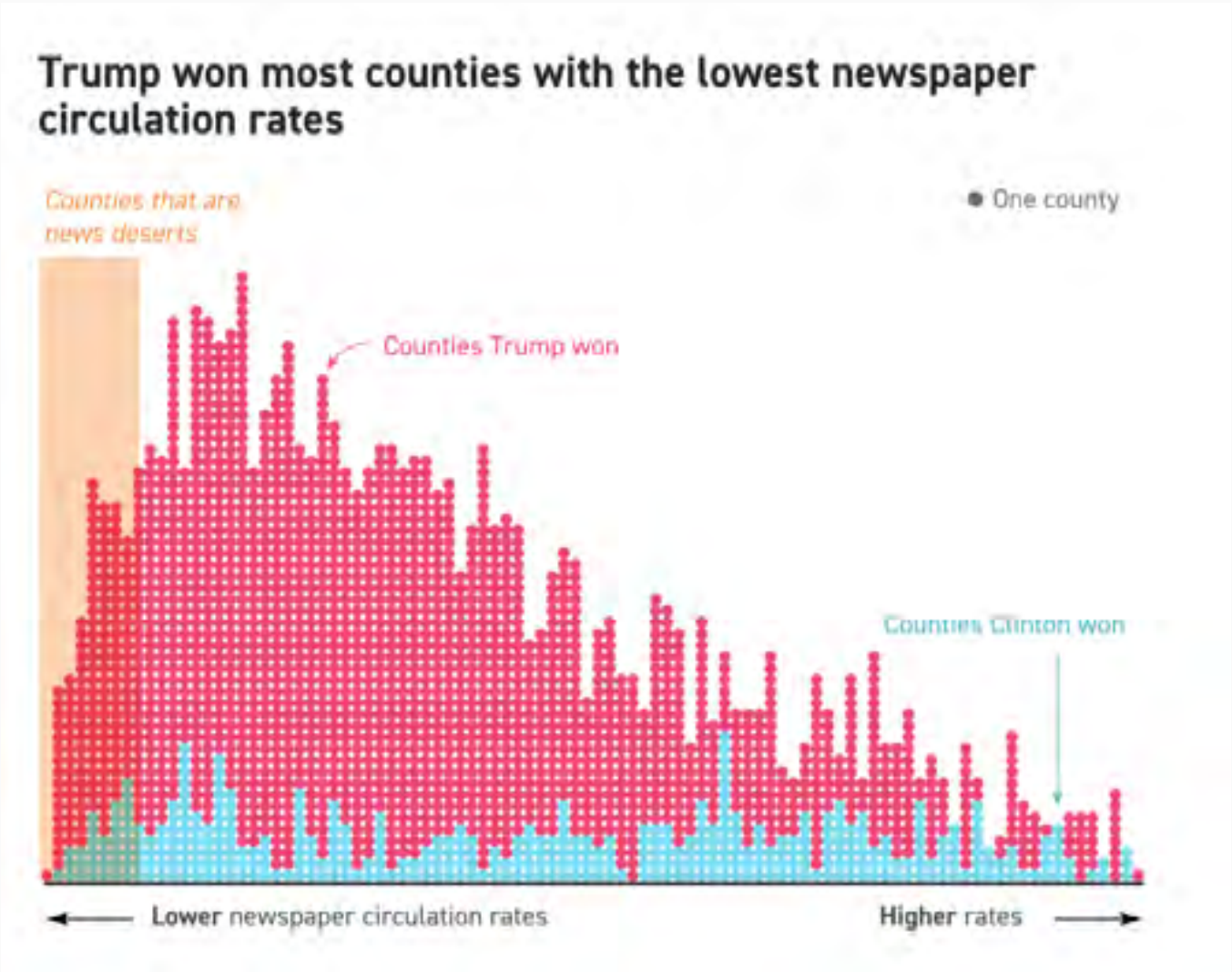
Beware of Cherry-picking



What am I really Identifying?



What am I really Identifying?



Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not

- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not

- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not
- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not
- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not
- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not

- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Final Thoughts on Thinking about Data

- If I run into any of these problems, is my data story doomed and I should quit?
  - No, of course not

- Data and data analysis is **never, ever** perfect
  - The point is to understand the limits of what we can and can't say
  - Being honest about the potential flaws of what we do provides credibility
  - And sometimes, finding and investigating flaws leads to good data projects on their own

Thanks!