Do I Have the Data I Need?

Data Analysis

Does money make people happy?

money → happy

Source: freepik.com

Does money make people happy?

money → happy

What if we don't have money to give people?

Source: freepik.com

Does money make people happy?

money → happy

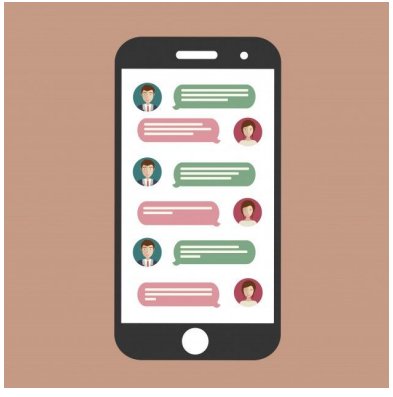
What if we don't have money to give people?

How will we measure happiness?

Source: freepik.com

Does texting while driving cause accidents?

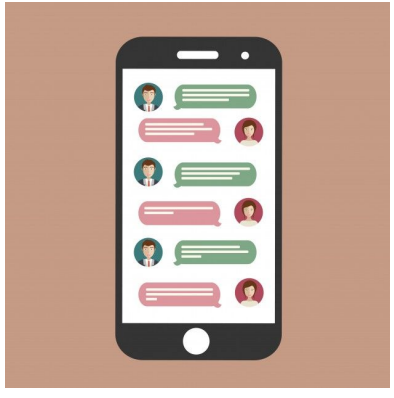
texting while driving → car accidents



Source: freepik.com

Does texting while driving cause accidents?

texting while driving → car accidents



Unethical to tell people to text while driving

Source: freepik.com

In this lesson...

- Determine if you have data you need
- Limitations of your data
- Considerations to make before analysis
- What to do to get the data

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

Email
Twitter
Github
Stack Overflow

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:

http://varianceexplained.org/r/trump-tweets/

What data do you need to answer your data science question and what limitations do these data have?

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

Email
Twitter
Github
Stack Overflow

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:

http://varianceexplained.org/r/trump-tweets/

The Perfect Dataset

What would the perfect dataset look like?

The Perfect Dataset

The data science question:
Do the tweets on the _@realDonaldTrump_ account differ between those posted using an iPhone vs. those posted using an Android? Are the Android tweets angrier and more negative?

The perfect dataset would contain **all** tweets and the variables:

- date
- time
- os
- author
- location
- tweet
- censored
  - censored_date
  - censored_time
- anger

The Data We Have

The data science question:
Do the tweets on the _@realDonaldTrump_ account differ between those posted using an iPhone vs. those posted using an Android? Are the Android tweets angrier and more negative?

The data we have includes ***some*** tweets and the variables:

- **date**
- **time**
- **os**
- ~~author~~
- ~~location~~
- **tweet**
- ~~censored~~
  - ~~censored_date~~
  - ~~censored_time~~
- ~~anger~~

The Data We Can Get (Easily)

Obtain data from available sources:

- Government Data
- APIs
- Open Datasets
- Your Company
- etc...

Data Collection

Different types of data by how they are collected:

- Observational data
  - Cross-sectional data
  - Longitudinal data
  - Panel data
- Experimental data

Data Collection

Different types of data by how they are collected:

- Observational data
  - Cross-sectional data
  - Longitudinal data
  - Panel data
- Experimental data

Data Collection

Different types of data by how they are collected:

- Observational data
  - Cross-sectional data
  - Longitudinal data
  - Panel data
- Experimental data

Data Collection

Different types of data by how they are collected:

- Observational data
  - Cross-sectional data
  - Longitudinal data
  - Panel data
- Experimental data

Data Collection

Different types of data by how they are collected:

- Observational data
  - Cross-sectional data
  - Longitudinal data
  - Panel data
- Experimental data

The Data We Can't Get

Limitations to Data Collection:

- Limited Resources
  - Money
  - Time
  - Access
- Ethical Limitations
  - Unethical Experiments
  - Invasive Data Collection
- Security

The Data We Can't Get

Limitations to Data Collection:

- Limited Resources
  - Money
  - Time
  - Access
- Ethical Limitations
  - Unethical Experiments
  - Invasive Data Collection
- Security

The Data We Can't Get

Limitations to Data Collection:

- Limited Resources
  - Money
  - Time
  - Access
- Ethical Limitations
  - Unethical Experiments
  - Invasive Data Collection
- Security

The Data We Can't Get

Limitations to Data Collection:

- Limited Resources
  - Money
  - Time
  - Access
- Ethical Limitations
  - Unethical Experiments
  - Invasive Data Collection
- Security

Questions to Ask Yourself

- Change the question?
- Still worth doing?
- Project feasible?
- If feasible: how rework the question & redesign data collection?

Are the data we have good data?

Be sure:

- each variable forms a column
- each observation forms a row
- each table/file stores data about one kind of observation
- if variables are collected from multiple sources, they are merged properly
- column names are easy to use and informative
- obvious mistakes in the data have been removed
- missing values are formatted uniformly and correctly
- variable values are internally consistent
- appropriate transformed variables have been added

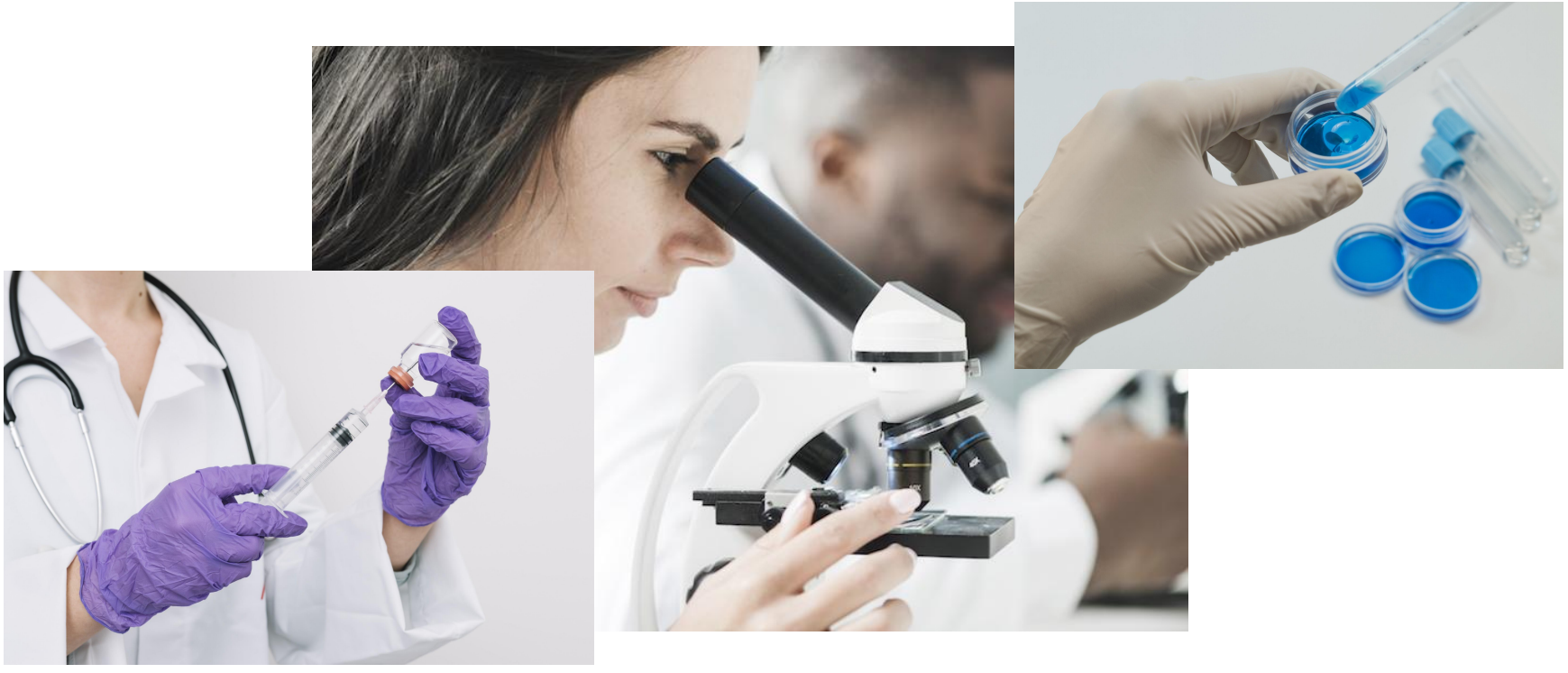
Reasons data aren’t good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Number of observations is too small



Source: Freepik.com

Reasons data aren't good

- Small number of observations

A large sample size is generally better than a small sample size

- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Dataset does not contain the exact variables you are looking for

Source: Freepik.com

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

proxy variables can be helpful!

Variables in the dataset are not collected in the same year

Source: Freepik.com

Reasons data aren't good

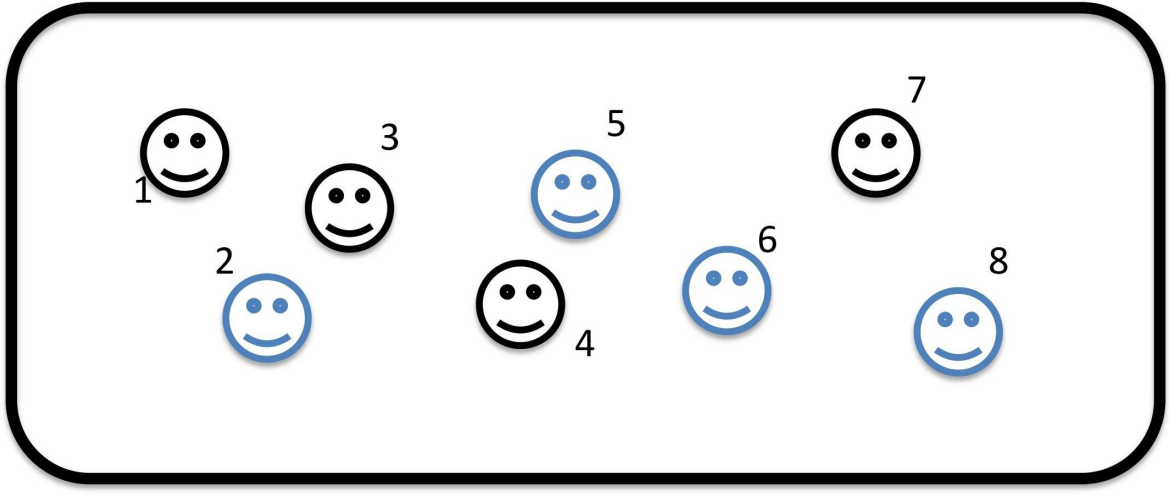
- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Go find data from the same time period!

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

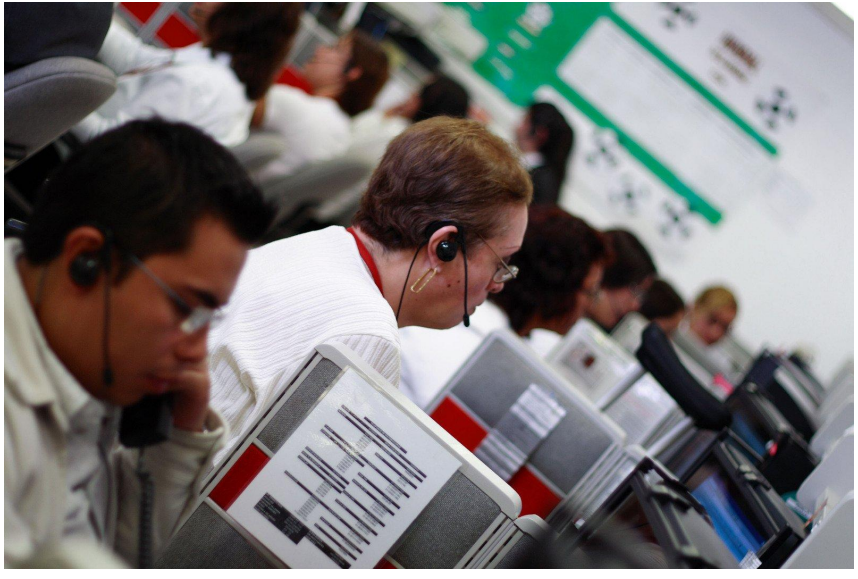
Dataset is not representative of the population that you are interested in



Dataset is not representative of the population that you are interested in

Source: Freepik.com

Dataset is not representative of the population that you are interested in



Source: freeimages.com

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Random sampling is important!

Examples of bad sampling:

- Surveying subscribers of a gun-related magazine for research on attitudes toward owning guns
- Surveying Facebook users for what TV shows Americans like

Best sampling practices:

- Always think about what your population is
- Collect data from a sample that is representative of your population
- If you have no choice but to work with a dataset that is not collected randomly and is biased, be careful not to generalize your results to the entire population

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

Some variables in the dataset are measured with error

Source: freepik.com

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error — Instruments have error & people's memories are not perfect!
- Confounding

Reasons data aren't good

- Small number of observations
- Variable of Interest not in dataset
- Variables of interest not from same year
- Dataset not representative of population
- Measurement error
- Confounding

popsicles → crime rate

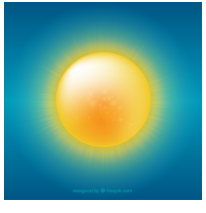
Source: freepik.com

Confounding variables

popsicles → crime rate

temperature → popsicles

temperature → crime rate



FiveThirtyEight

Politics Sports Science & Health Economics Culture

NOV. 9, 2016, AT 4:53 PM

The Polls Missed Trump. We Asked Pollsters Why.

By Carl Bialik and Harry Enten

Filed under 2016 Election

POLITICS

4 Possible Reasons The Polls Got It So Wrong This Year

November 14, 2016 · 2:28 PM ET

TheUpshot

POLITICAL CALCULUS

A 2016 Review: Why Key State Polls Were Wrong About Trump

By Nate Cohn

May 31, 2017

193

BUSINESS INSIDER

TECH FINANCE POLITICS STRATEGY LIFE INTELLIGENCE ALL

A group of major pollsters just released an autopsy report to explain why the polls were such a disaster in 2016

Allan Smith May 7, 2017, 12:23 PM

Image: freepik.com and Wikimedia Commons (Kyle Taylor)

After you have your data science question...

- Imagine the optimal dataset
- Determine data you have
- Identify data you can get
- Figure out limitations
- Do you need to re-work question?
- Explore, Wrangle, Analyze, Answer!

Garbage in, garbage out!