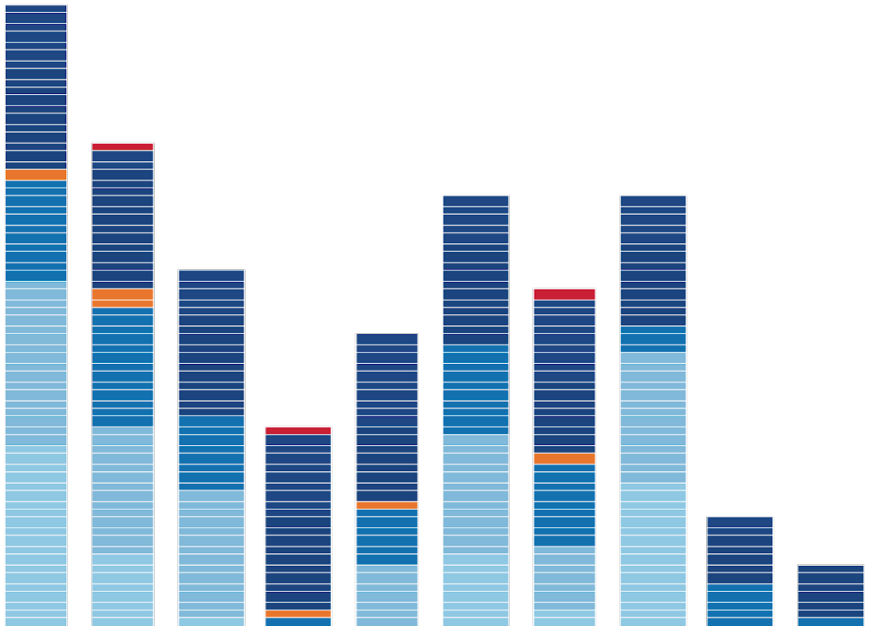
Announcements

- Office Hours:
    - Mondays, 6:30 PM - 7:30 PM, in HDSI Office E203 (SDSC).
- Readings (Week 1 due Friday 1/10)
- Email: gquer@ucsd.edu
- Subject line:
    - [DSC96 W20]: Week 01, Sec A/B, YourFirstName YourLastName
- Email content:
    - Your comments/ questions/ observations on the proposed lectures

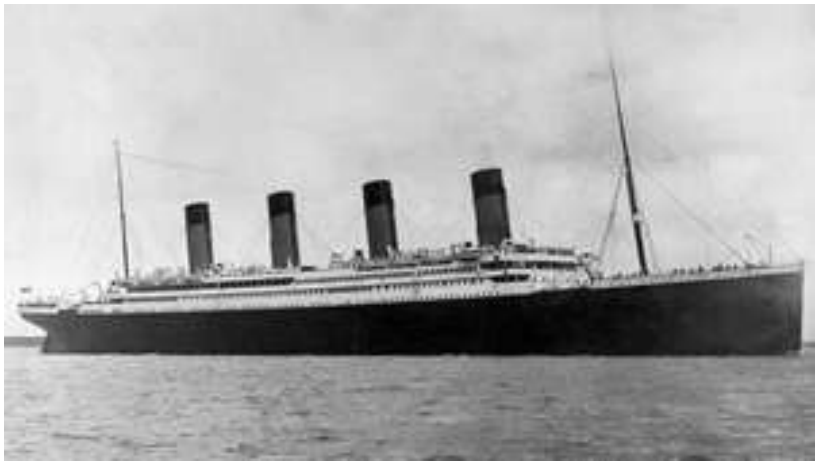
- **Tableau basics**



- Intro to interface
- Importing
- Dimensions and Measures
- Chart types
- Finding the story
- Practice on your own

Download a file from github

- Click on the file
  - You will see it (csv)
- Press Alt + click on “raw” (top right)



- Titanic dataset and cool things you can do with it: https://www.kaggle.com/c/titanic
- Tableau official training: https://www.tableau.com/learn
- Tableau examples with Titanic data: https://public.tableau.com/search/all/titanic

Titanic questions

- Sex, age, pclass
  - 1) Which one is affecting survival? Show it
  - 2) Are there confounding effects, or is each feature (sex, age, pclass) affecting survival?

3) Does group size have an effect on survival rate?

Lobby questions

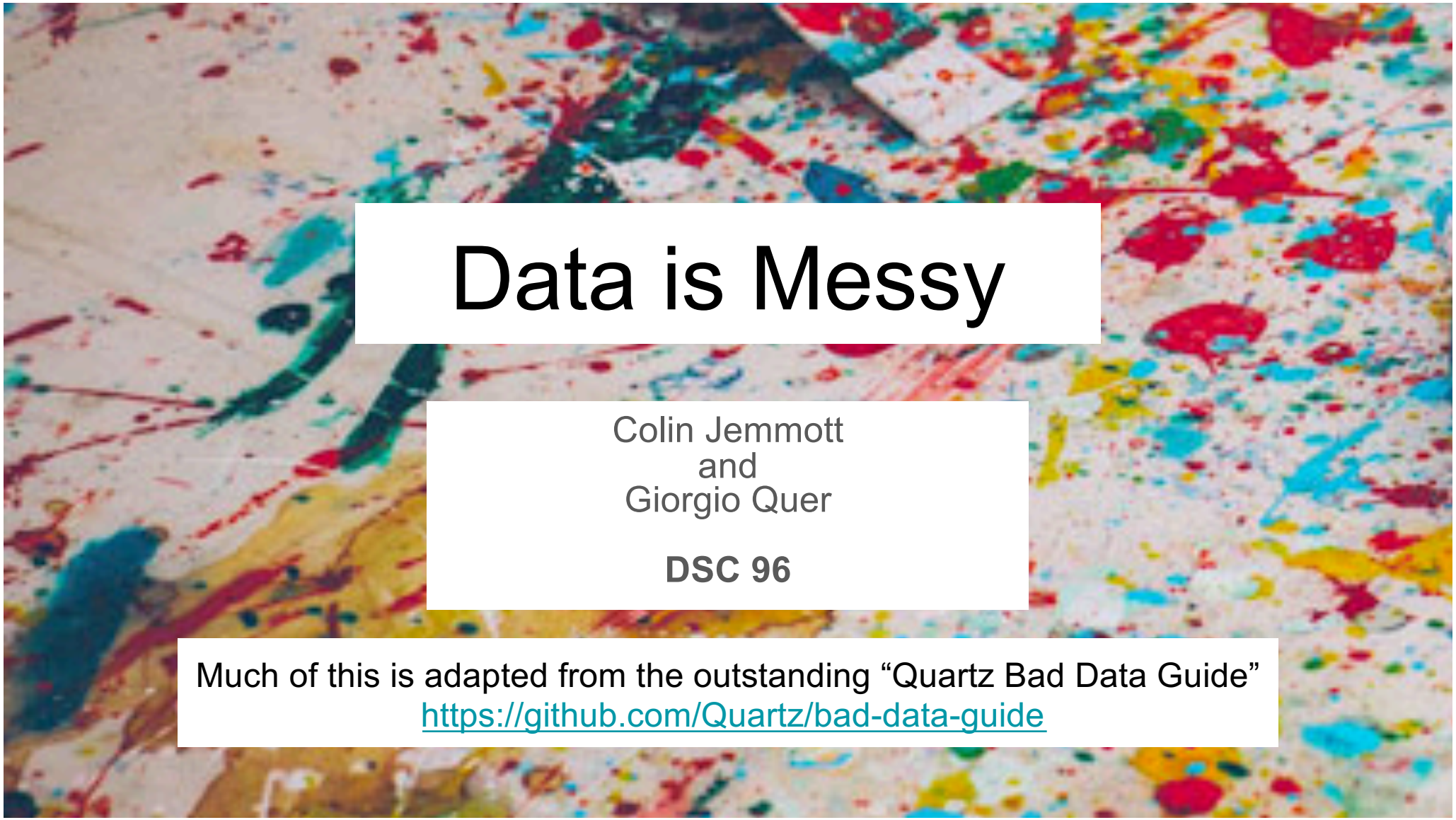
- Data on how much each local governmental agency has paid for lobbyists
- Compensation and expenses both contribute to total money for lobby
  - You should sum them up (right click, then Create -> Calculated field)

1) Which entity paid most?

2) Which entity type paid the most?

3) Tribes: are they paying lobbyists a lot?

- Come up with a colorful and clear way to present your data
- New story (bottom of the workbook)
- Give a title
- Import you figures (double click from left tab)
- Write the story (Add a caption) on your data



Data is Messy

Colin Jemmott
and
Giorgio Quer

DSC 96

Much of this is adapted from the outstanding “Quartz Bad Data Guide”
https://github.com/Quartz/bad-data-guide

Data Types

Many different data types exist. Common types include:

- Integers : 5, 2790, 342, 1200124
- Floating-point numbers: 13.540394542 , 3.14159… , 22.7421341321514
- Strings: ‘Hello’ , ‘This data is a mess!’, ‘92122’
- Booleans: True, False

Even with these simple types, data can often be “messy” or bad”.

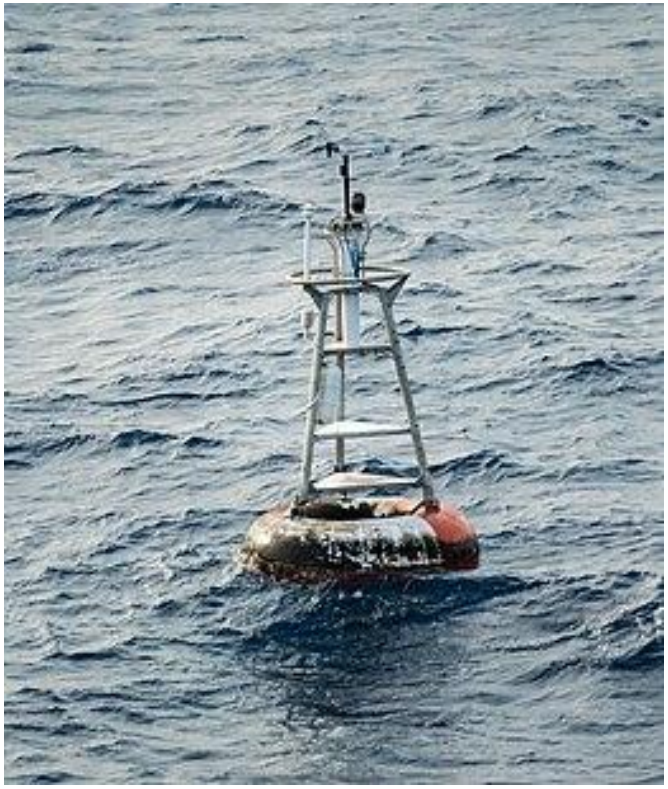
What might go wrong?

Missing Values

- Null
- NaN
- 0, -1 or “” instead of null
- 1900 and 1970
- “Null Island” at 0°00'00.0"N+0°00'00.0"E

Related: missing data that you know should be there

- how many states should be listed in national data?

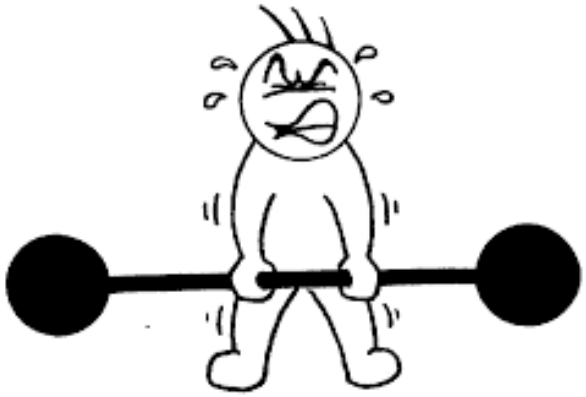


Null Island is one of the most popular jogging locations according to the Strava fitness tracking app.
https://en.wikipedia.org/wiki/Null_Island

Dates and Units

Which date is in September?

- 9/10/18
- 10/9/18



Object A is listed as “weight=87”. Can you lift it?

Does “Los Angelos” == “Los Angeles”?

Numbers and “Numbers”

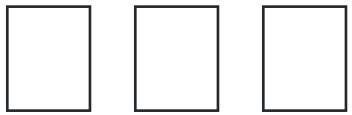
1537660383 looks like a number, but is probably a date (Unix timestamp)

“**USD 1,000,000**” looks like a string, but is actually a number and a unit.

02111 looks like a number, but is really a zip code (and isn’t equal to 2,111)

Strings

- **Encoding problems**
  - Presence of weird characters in the middle of a word
- **Solution**
  - Ask the source
  - Best guess



Data definition

- Data is too coarse:
  - You needs months, but you only have years
- Data is too granular:
  - You have daily “number of steps”, but you need monthly steps for your statistical analysis

Data collection problems

- We have a great dataset:
  - Physical activity for 1 year from 10M people in US with an activity tracker!
  - We want to describe the physical activity of US citizens !
  - **Can we?**

Data collection problems

- We have a great dataset:
  - Physical activity for 1 year from 10M people in US who bought an activity tracker!
  - We want to describe the physical activity of US citizens !
  - Can we?

- Ok, let's collect the data properly:
  - 1000 people randomly selected (any age or physical status or income) in San Diego county
  - 3 months of data (May, June, July)
  - Are we ok now?

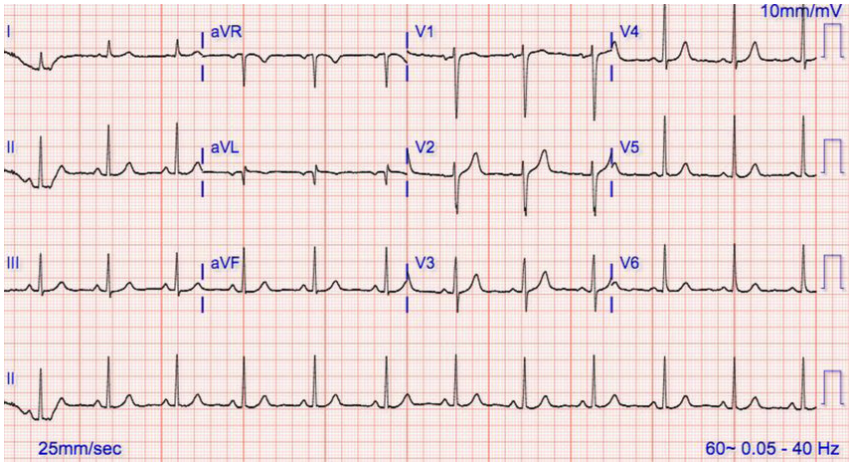
Data collection problems

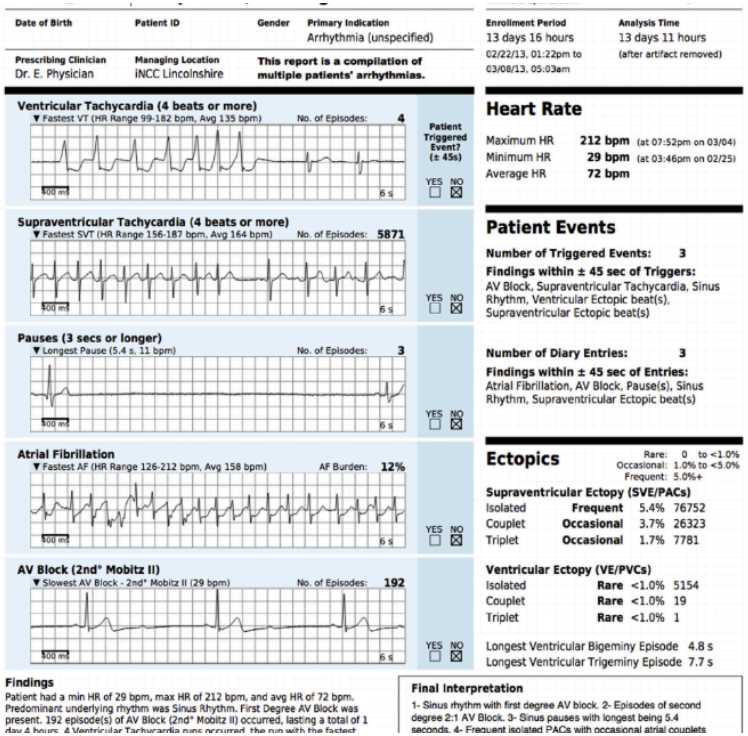
- Sample is not random
  - You have the number of steps, but the population is composed of very active people
- Seasonal variation
  - You have number of steps from a good population, but only in summer time
- Results are p-hacked
  - The data collection stopped once a significant result was observed

Other data types

Data doesn't always come in in nicely formatted packages

- CSV, escaping, and the lack of standards
- Data are in a PDF - what now?
- Images and sound recordings as data





from: Barrett et al, "Comparison of 24-hour Holter Monitoring with 14-dayNovel Adhesive Patch Electrocardiographic Monitoring"