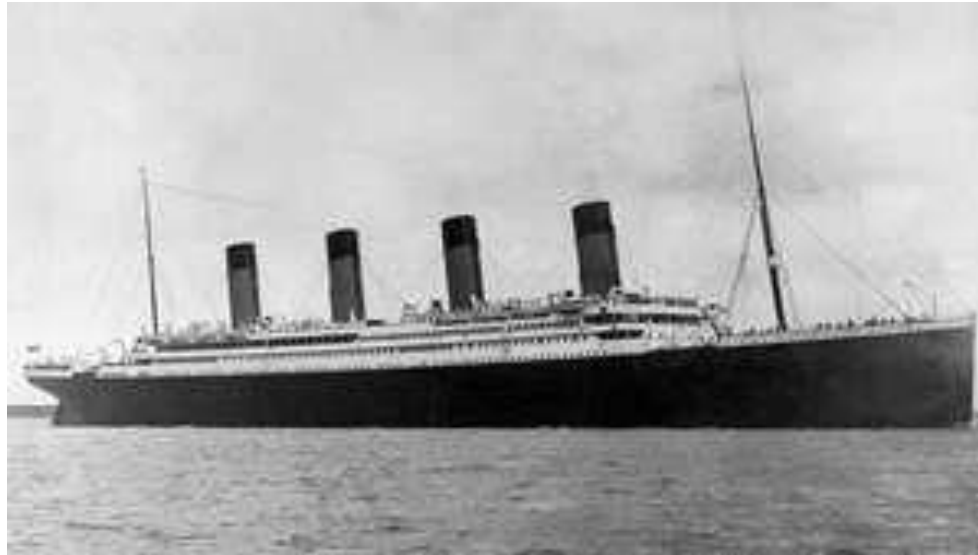**DSC 96**

Fall 2019

Class 2

Data Exploration

Announcements

- Readings (Week 1 due Friday, week 2 due Tuesday)
- Professor versus Lecturer

Titanic Data

- How many people were on the ship?
- What was the total of all fares paid?

Tableau

- Dimensions and Measures

- Chart Types

- Aggregations

- Filters

- Calculated Values

Titanic data

- Pclass:
  - It represent the class for each passenger (first, second, third)
  - 1,2,3
- Tableau
  - It guesses it is a number (like the amount paid): it is in **Measures**
  - It makes sense for Tableau to sum them up or to average them
  - What is the meaning of: average class is 1.85 ?
- Solution
  - Drag pclass it to **Dimensions** !
  - Change the data type to **string**
- What about age?
  - Age in bins (drag to Dimensions, Create, Bins)
 - Fix reasonable intervals (10 years?)

- **Titanic!**

  - 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

- Can you attribute survival to a single primary trait?
- In other words, can you attempt to tease apart the confounding effects of pclass/age/sex, when assessing one?

- Can you explain why the distribution of fares by pclass seems off? why are some 3rd class tickets more expensive than first class?

- Does group size have an effect on survival rate?
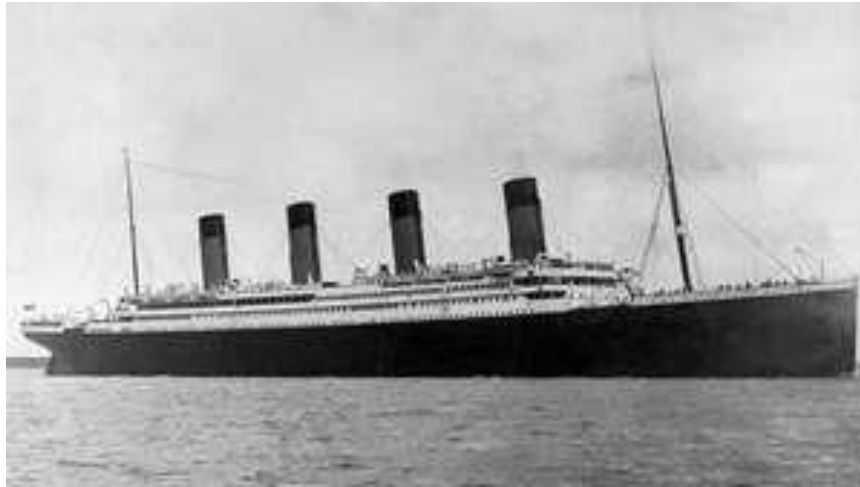
Data is Messy

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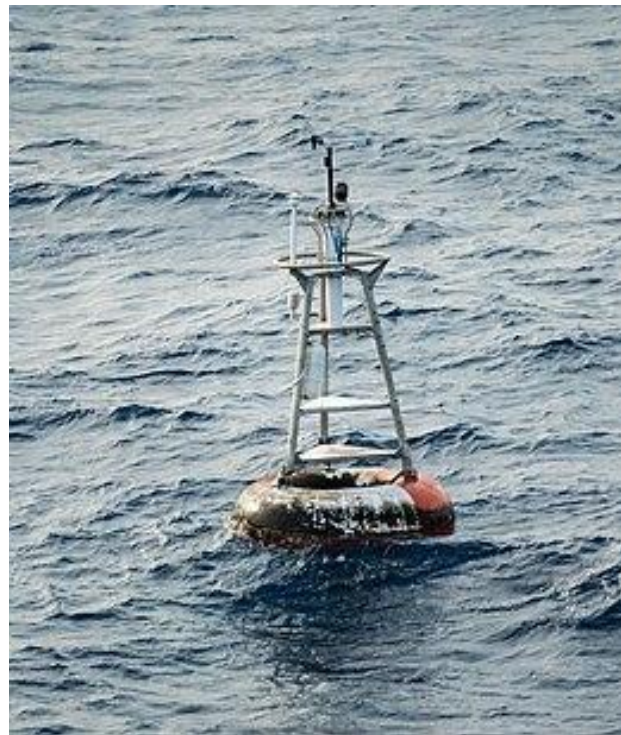
Even with these simple types, data can often be "messy" or bad".

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"?
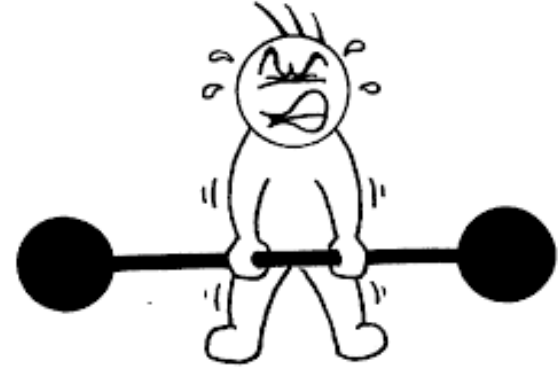
Numbers and "Numbers"

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
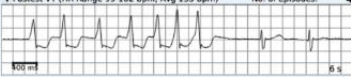
Other data types

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"