

# Data is Messy

Giorgio Quer  
(credits to Colin Jemmott)  
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
- Floating-point numbers
- Strings
- Booleans

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

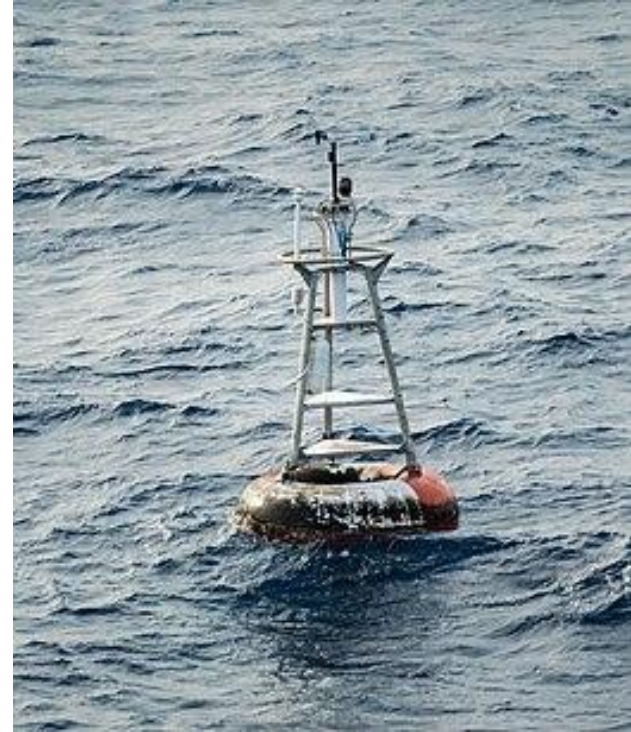
What might go wrong?

# Missing Values

- Null
- NaN
- 0, -1 or "" instead of null
- 1900 and 1970
- "Null Island" at  $0^{\circ}00'00.0''\text{N}+0^{\circ}00'00.0''\text{E}$

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](https://en.wikipedia.org/wiki/Null_Island)

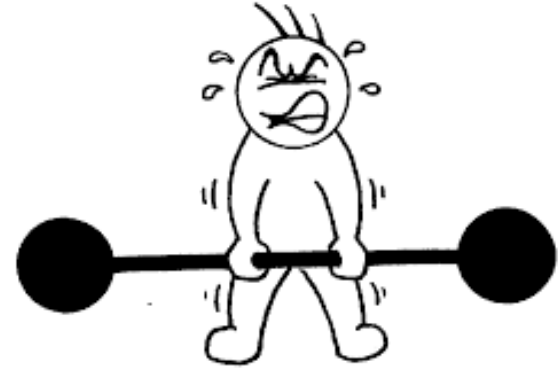
# Dates and Units

Which date is in September?

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

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

- 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

# Vehicle Stop Data

DSC 96

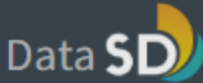
# Data Source




Secure | <https://data.sandiego.gov>

City of San Diego | Mayor Kevin L. Faulconer

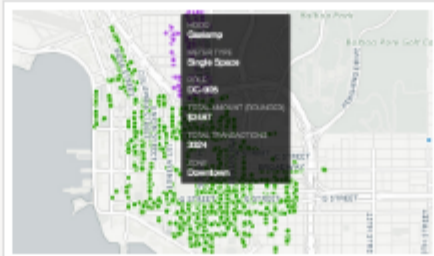
[Browse Data](#) [Stories](#)





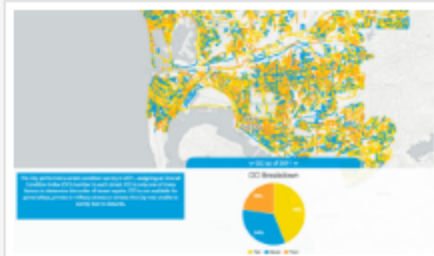
### SMART STREETLIGHTS

See where streetlights are getting upgraded



### PARKING METER REVENUE

Check how much each of the City's smart parking meters made last year



### STREET REPAIR WORK

Get a condition rating for each street in the City's network, plus view current and future street repair work

# Why Police Data?



# Police Vehicle Stops

Vehicle stops made by the San Diego Police Department. Vehicle Stops files contain all vehicle stops for a given year.

## Vehicle Stops (year-to-date)

*This is a preview. If you would like to view the full resource, please download it above.*

Show/Hide Column ▾

STOP_ID	STOP_CAUSE	SERVICE_AREA	SUBJECT_RACE	SUBJECT_SEX	SUBJECT_AGE	TIMESTAMP
Filter	Filter	Filter	Filter	Filter	Filter	Filter
1444799	Moving Violation	120	I	M	37	2017-0
1444821	Equipment Violation	520	W	M	22	2017-0
1447102	Moving Violation	520	W	M	29	2017-0
1444801	Equipment Violation	720	H	F	61	2017-0
1444802	Equipment Violation	120	H	M	24	2017-0
1444912	Equipment Violation	440	B	M	45	2017-0

# SDPD Vehicle Stop Data

1. Plot count of stops by age. Notice any issues? What should we do?
2. Make some time series plots! For example, stops by hour of day, day of week, month, etc. might be interesting.
3. Explore the “stop cause” variable. Notice any issues? What should we do?

Finally, explore and answer questions. When you find bad data, bring it up to the class.