Lecture 2 Notes

Announcements

- Quiz 1 - available on Gradescope.
  - You should have an email (yesterday) with instructions for setting your password and logging in.
  - Take it between 2pm today and the start of class (1pm) Monday.
  - 15-minute time limit.
  - Quiz 1 only: dry run - full credit for participation.
- Start of quarter survey - on Canvas. Fill out by Sunday night.
- **Disclaimer**: My goal is to assume CSCI 141 and pre-algebra and nothing more.
  - There's a lot of variety in experience levels here - that's good!
  - Do not be intimidated by your peers, or by me.
  - I will slip up and assume you know things! I'm sorry. **Please** please **please** tell me if I do this so I can fix it. I will appreciate it, and the 20 other people in this class who also didn't know it will also silently appreciate it.
  - Conversely: I'm new to a lot of the tools we're using - we'll all be learning together! If you know of a better way to do something, let me know.

Questions on the syllabus?

What is data?

- (if time allows) What is data? Student suggestions
  - Properties of data
    - structured/unstructured
    - numerical/categorical
    - big/small
  - Structures of data
  - Types
    - `str`
    - integer
      - signed
        - `int32` (*roughly* -2b to +2 billion)
        - `int64` (*roughly* -9q to +9 quintillion)
      - unsigned
        - `uint8` (0 - 255)
        - `uint32` (*roughly* 0-4b)
        - `uint64` (*roughly* 0-18 quintillion)
    - floating-point

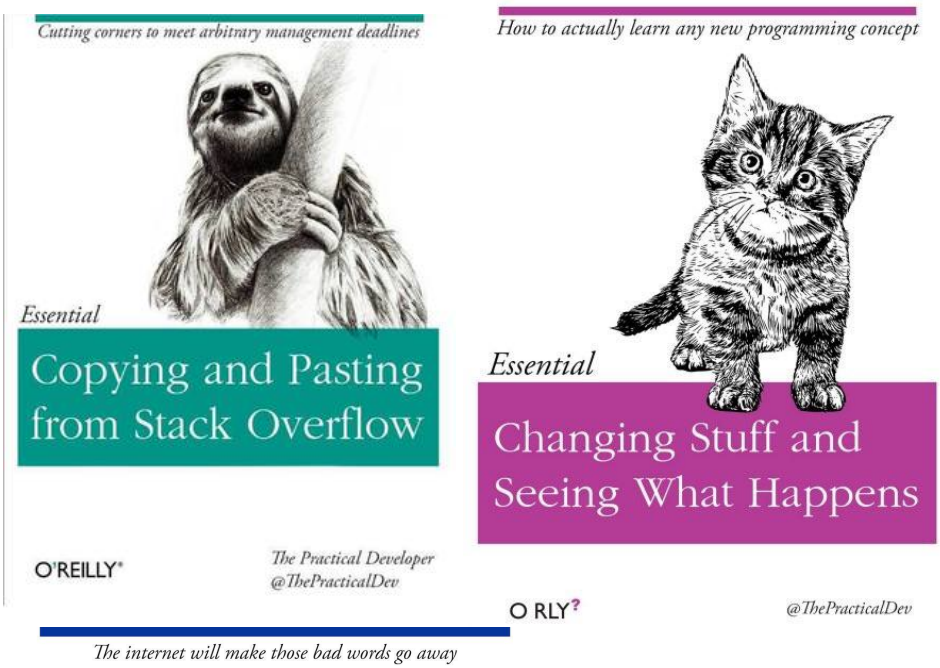
        - `float32` - *roughly* 7 decimal digits of precision
        - `float64` - *roughly* 15 decimal digits of precision
    - `object` - Pandas type that usually wraps columns of strings and other mixed types

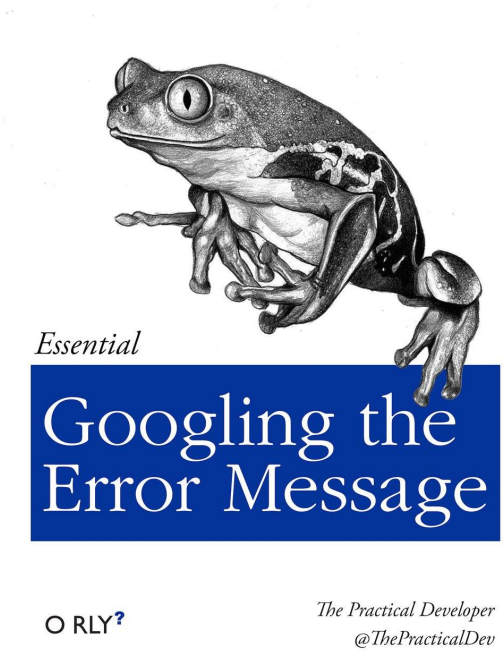
Lab 1 - Logistics

- Start in class today, finish and submit by Thursday at 10pm (this will be typical)
- Work in pairs in class (if you'd like to), and individually thereafter. Collaboration is still allowed, but be sure you're following the collaboration policy detailed on the syllabus.

Lab 1 - Demo

1. Environment Setup
2. Jupyter Concepts
    - Python cells
        - Contain Python code
        - You can run and re-run cells
        - State is maintained after running a cell
        - The value of the last line, if any, is displayed (not printed)
    - Markdown cells:
        - Allow you to intersperse formatted text with code.
        - Type your markdown syntax, then "run" the cell to see it rendered with formatting.
        - Basic markdown formatting
    - Why jupyter?
        - Interleaved display
        - Quick, interactive development cycle
        - Reproducibility. Cardinal rule of data science: **Always start with the raw data.**
3. Pandas Basics - How to Learn Pandas (and other tools we'll use in this class):





- But seriously: I won't teach you every little thing you need to use. I will expect you to be able to find and use functionality that gets the job done. I also won't quiz/test you on syntactic minutia.

Scavenger hunt demo:

```
data_url =
'https://fw.cs.wwu.edu/~wehrwes/courses/data311_21f/data/avengers/avengers.csv'
```

```
import pandas as pd
df = pd.read_csv(data_url, encoding='latin-1')
df
```

- Sample task:
    0. **Drop some columns**. Trim the table to drop the `"URL"`, `"Probationary Introl"`, `"Full/Reserve Avengers Intro"`, `"Honorary"`. Store the trimmed table to a variable called `avengers`.
    Useful function(s): `df.drop`