Introduction

What we're doing here

Why are we here?

- Understand what data scientists do
- Get some cool tools and skills
- Build visualizations - for decisions, M&E, funding
- Stop hand-waving and start making stuff

Who's helping?

- Prof:
  - Sara-Jayne Terp (bodacea on github)
- Teaching Assistants:
  - Nate Brennand
  - Henrique Gubert
  - Lin He

This Weekend

- Friday: basic concepts, set up tools, Python
- Saturday: data, science, visualizations
- Sunday: advanced concepts, continuing your journey

Some of you have to leave for an hour or two

- To go to church, lectures, etc (nb "hangover" doesn't count)
- That's okay… these things happen
- All slides are online, with notes
- And we have "activity sessions", designed to help you get further

Why are you here?

- …tell me what you want to get out of this weekend…
  - What are your favorite visualisations?
  - What's your favorite dataset?
  - What are your burning questions?
  - What do you want to build?

What you will learn this weekend

- Basics of a computing language
- Basics of data management
- Basics of creating a visualization
- Tools and places to help you

What you won't learn this weekend

- Statistics
- Specific algorithms like k-means clustering
- Specific application areas like machine learning

(Resources for these: Coursera, MITx, Stack overflow)

Process

- OSEMN: Obtain-Scrub-Explore-Model-Interpret
  - Obtain datasets
  - Clean, combine, transform data
  - Explore the data
  - Try models (classification, machine learning etc)
  - Interpret and communicate your results

http://machinelearningmastery.com/how-to-work-through-a-problem-like-a-data-scientist/

Data

- find data
- pull data (automatically)
- clean data
- reformat data

Responsibility: How bad data fed the Ebola epidemic, New York Times

Science

- explore data
- model data
- interpret
- predict
- test hypotheses

Visualisation

- Interpret data
- Results aren't useful if they don't *do* something
  - e.g. Persuade a decision-maker
- Good visualisation = insight, persuasion

Why not tool X?

- Lots of data science applications and tools, very few core concepts:
  - Data collection
  - Data cleaning
  - Visualisation
  - etc

Tools change: want you to focus on the concepts

Why Python, R, D3?

- Very flexible languages
- Lots of helpful libraries
- Huge communities

PS Ignore the holy wars - just use what works for you

Help after this weekend

- Local meetups, e.g. data driven NYC, NYC predictive analytics, DataKind NYC, Hacks/Hackers, NYC Pyladies, NYC datascience, data visualisation New York, Data skeptics, NYC data wranglers, …
- Data hackathons (e.g. DataKind)
- Websites:
  - Stack Exchange: http://datascience.stackexchange.com/
  - Datatau: http://www.datatau.com