Data Science and Activism

Intro/Overview

A step back...

What is data?

What is data science?

A step back…

What is data?

- Information collected through observation

- Only becomes information suitable for making decisions once analyzed (quantitative/qualitative)
  - Our brains are doing this all the time (with sensory data)

A step back...

What is data science?

- Science of extracting knowledge and insights from data
- Unifies
    - statistics,
    - data analysis (transforming, cleaning, visualizing,...),
    - machine learning,
    - domain knowledge...

Example: In Police Scorecard, we saw stats, analysis, domain knowledge...

Data Science and Activism

How does an understanding of data science aid activism?

Data Science and Activism

How does an understanding of data science aid activism?

Our case studies fall into one of two categories:

- Criticize, subvert, draw attention to ways data is being (mis)used.
    - WMDs, CV Dazzle/Dazzle Club, limits of GDP
- Directly apply Data Science methods toward activist goals
    - Police Scorecard, VFRAME, Gringgo

Moreover, in both cases, a better understanding helps us figure out which questions to ask.

Data Science and Activism

Data science process flowchart from *Doing Data Science*, by Schutt & O'Neil (2013)
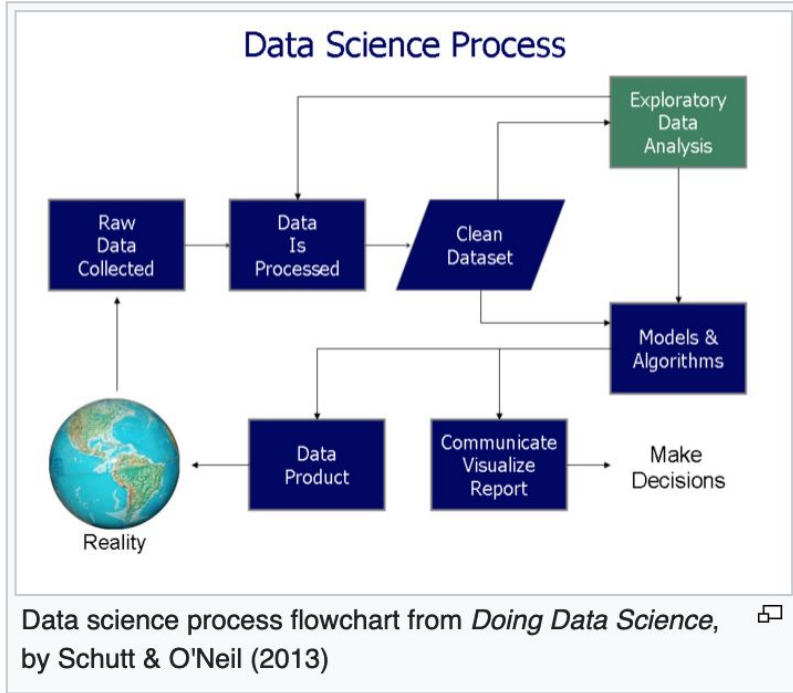
In reflecting on your potential role in data-driven activist project, it could help to look at this flowchart.

Which part or parts can you support/look at critically?

For example, can you help collect raw data? Can you help ensure that the collected data is reliable and representative?

Data Collection

What data is required? How should it be collected?

Examples from my projects:

- Labeled Images of waste
- Medical papers studying effectiveness of medical treatments

Qualitative data: numerical, measurable quantities. (ex. Number of drug arrests)

Quantitative data: descriptive, not measurable. (ex. images, text…)

Data Collection

Example tools/methods for data collection:

- Data scraping - computer program extracts data from human-readable input (e.g. websites)
  - https://scrapism.lav.io/ - guide to web scraping as artistic and critical practice (tool: scrapy)
  - "Regular expressions" - fancy 'ctrl+F'
- Data collection platforms
  - Mozilla Common Voice - build open-source database of voice data
  - Gringgo - App for photographing/manual labeling

Websites → Web Scraping → Structured Data

Speak

Donate your voice

Recording voice clips is an integral part of building our open dataset; some would say it's the fun part too.

Have you read our Terms?

Today's Progress
396 / 1200
Clips recorded

Help us get to 1,200

Data Processing

Not a well-defined term…

Essentially: "Making the data useable."

Ex. placing data in single, structured, easily decipherable, table format.

Data cleaning - checking for incompleteness, duplicates, errors

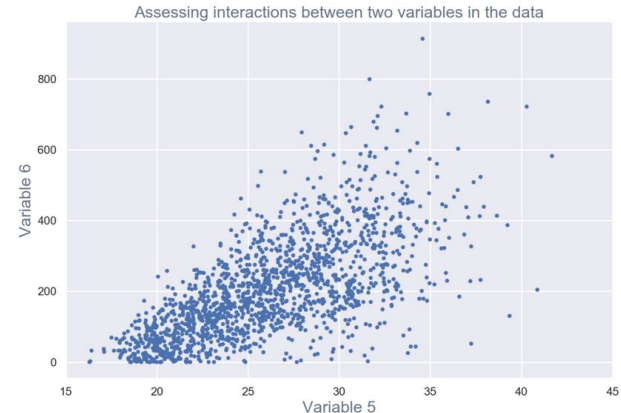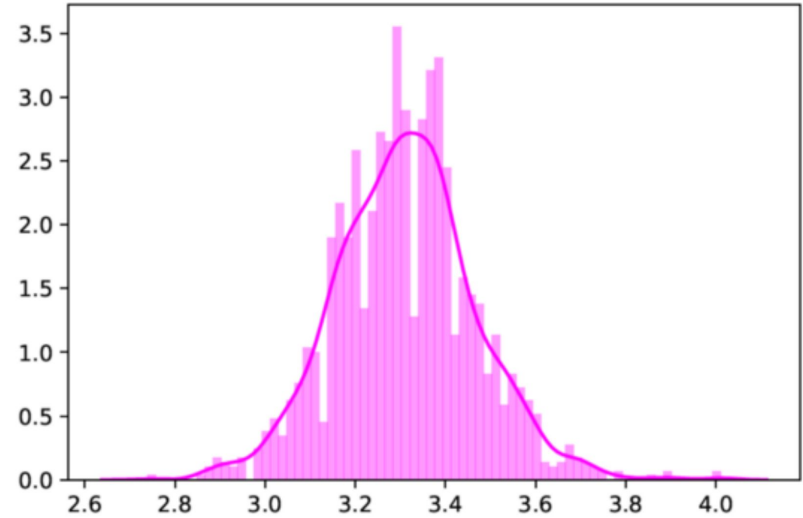- Tools: spell checker, outlier removal

Exploratory Data Analysis

Examples:

Descriptive statistics: average, median,...

Visualization: plot data points, histograms,... (tool: Matplotlib)

Objectives: make hypotheses based on data, suggest statistical tools to apply or further data collection


Assessing interactions between two variables in the data

Models and Algorithms

"Model" has different meanings in different contexts. Sometimes used interchangeably with "algorithm."

Algorithm: sequence of instructions taking given input and producing some output

Ex. A machine learning algorithm produces a trained model, itself an algorithm (input: images of faces, output: identity of person)

Today we will use an algorithm (linear regression) to model the relationship between two variables.
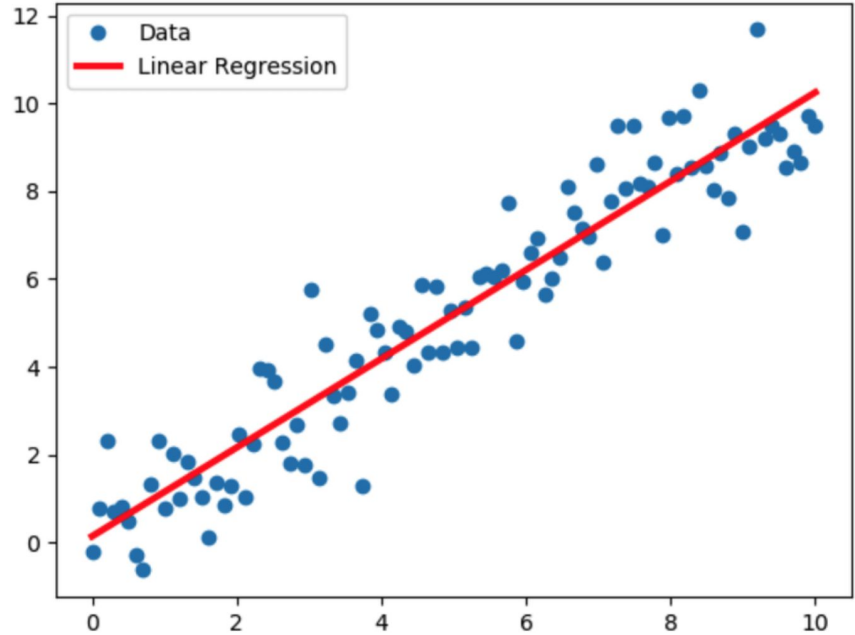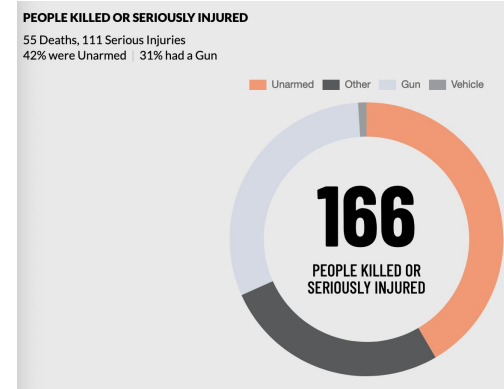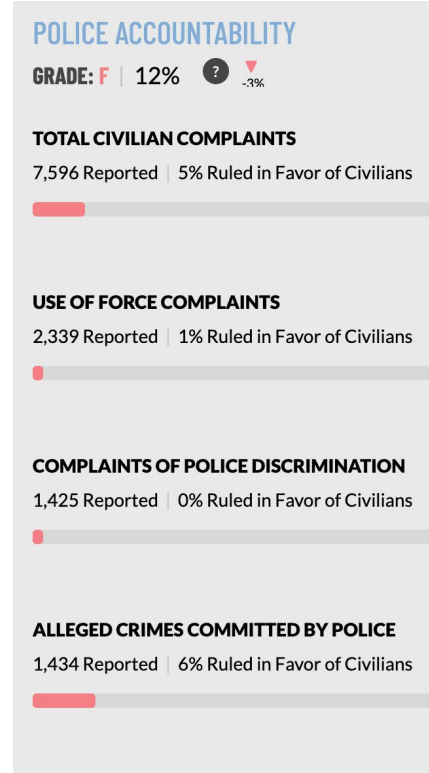
Report/Visualize/Decisions

Communicate results of analysis, model accuracy, tables, more visualization...

Ex. Police Scorecard website reports findings of their data analysis and score computation algorithm. Aids decisions regarding policing policy

Ex. After using Google AutoML, we decided to use a different model

POLICE ACCOUNTABILITY

GRADE: F | 12% ? ▼ -3%

TOTAL CIVILIAN COMPLAINTS
7,596 Reported | 5% Ruled in Favor of Civilians

USE OF FORCE COMPLAINTS
2,339 Reported | 1% Ruled in Favor of Civilians

COMPLAINTS OF POLICE DISCRIMINATION
1,425 Reported | 0% Ruled in Favor of Civilians

ALLEGED CRIMES COMMITTED BY POLICE
1,434 Reported | 6% Ruled in Favor of Civilians

PEOPLE KILLED OR SERIOUSLY INJURED
55 Deaths, 111 Serious Injuries
42% were Unarmed | 31% had a Gun

Unarmed   Other   Gun   Vehicle

166
PEOPLE KILLED OR
SERIOUSLY INJURED

Data Product

Refers to computer application taking data input and generating output.

Example: Packaging an image recognition model into an app with an intuitive UI.