Data Analysis

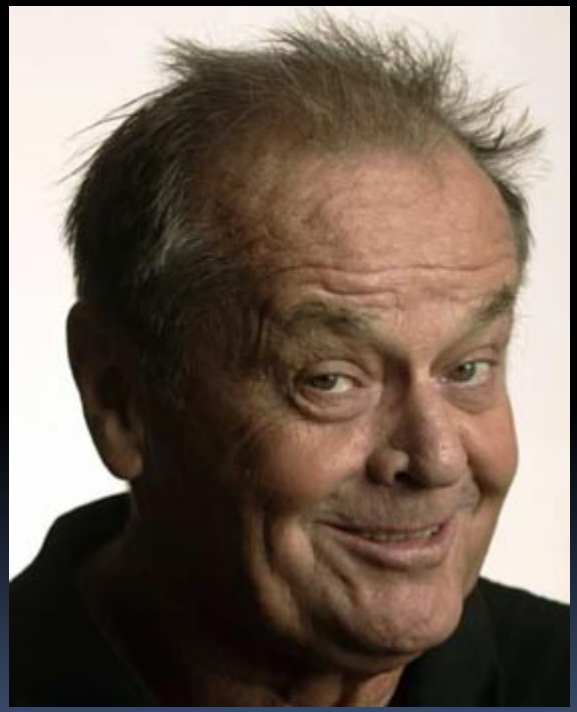
CSCI 447/547 MACHINE LEARNING

Outline

- Dataset Issues
  - Missing Data
  - Erroneous Data (Intentional or Not)
  - Data Format
- Multiple Source Issues
  - Different Fields
  - Different Answers
  - Population Sample

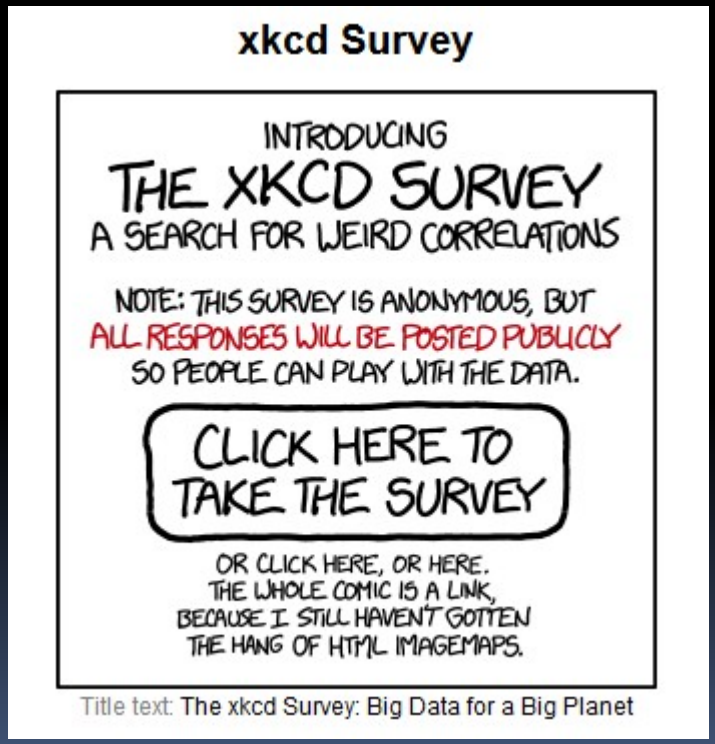
Data Analysis

I got you some data...



Data Analysis

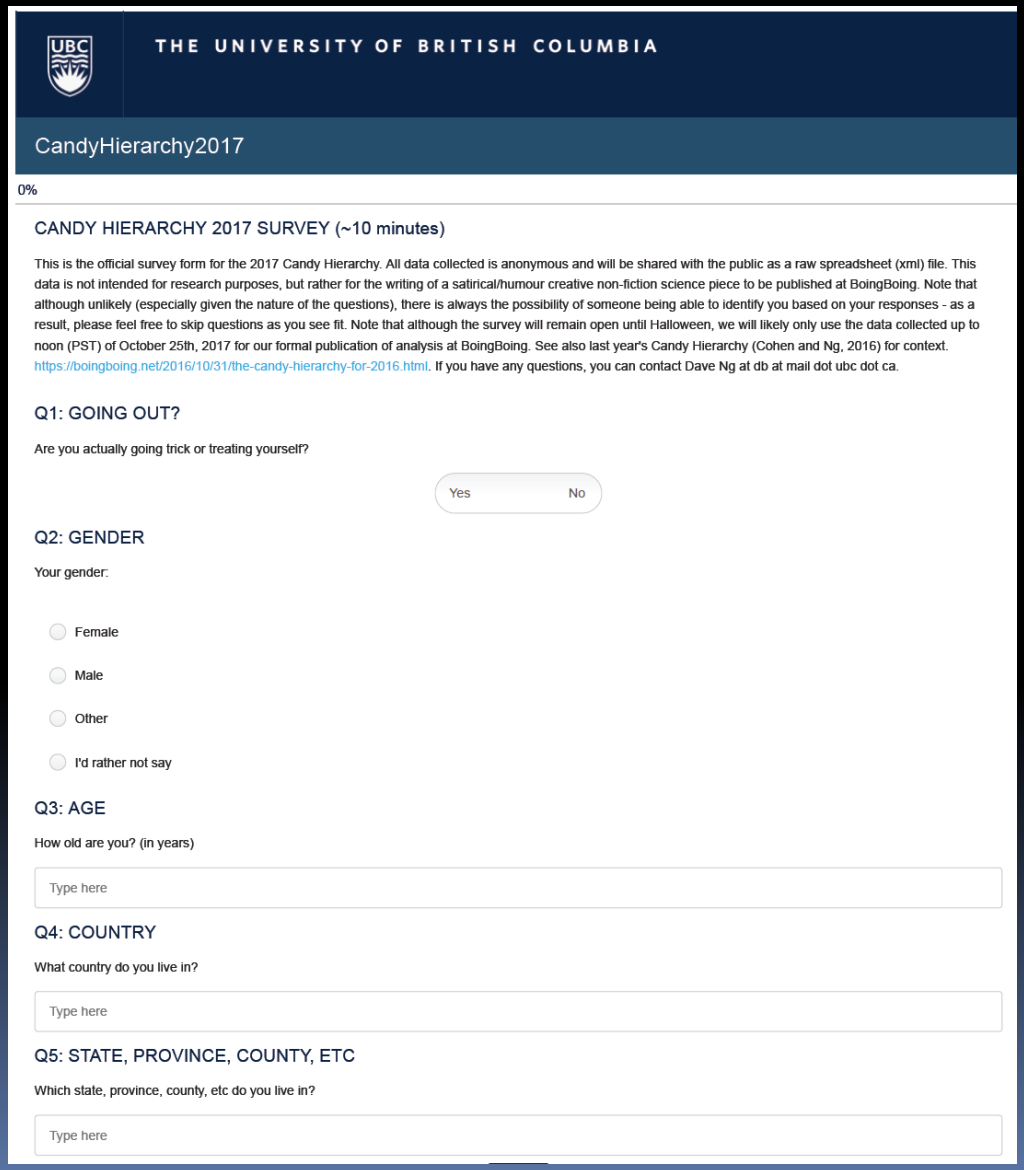
- xkcd Survey



Data Analysis: Case Study

- Data Source:
  - The Science Creative Quarterly
    - Candy Data (Halloween Survey)
      - 2015 (n=5459)
      - 2016 (n=1232)
      - 2017 (n=2460)

Survey Questions 2017



THE UNIVERSITY OF BRITISH COLUMBIA

CandyHierarchy2017

0%

CANDY HIERARCHY 2017 SURVEY (~10 minutes)

This is the official survey form for the 2017 Candy Hierarchy. All data collected is anonymous and will be shared with the public as a raw spreadsheet (xml) file. This data is not intended for research purposes, but rather for the writing of a satirical/humour creative non-fiction science piece to be published at BoingBoing. Note that although unlikely (especially given the nature of the questions), there is always the possibility of someone being able to identify you based on your responses - as a result, please feel free to skip questions as you see fit. Note that although the survey will remain open until Halloween, we will likely only use the data collected up to noon (PST) of October 25th, 2017 for our formal publication of analysis at BoingBoing. See also last year's Candy Hierarchy (Cohen and Ng, 2016) for context. https://boingboing.net/2016/10/31/the-candy-hierarchy-for-2016.html. If you have any questions, you can contact Dave Ng at db at mail dot ubc dot ca.

Q1: GOING OUT?

Are you actually going trick or treating yourself?

Yes No

Q2: GENDER

Your gender:

- Female
- Male
- Other
- I'd rather not say

Q3: AGE

How old are you? (in years)

Type here

Q4: COUNTRY

What country do you live in?

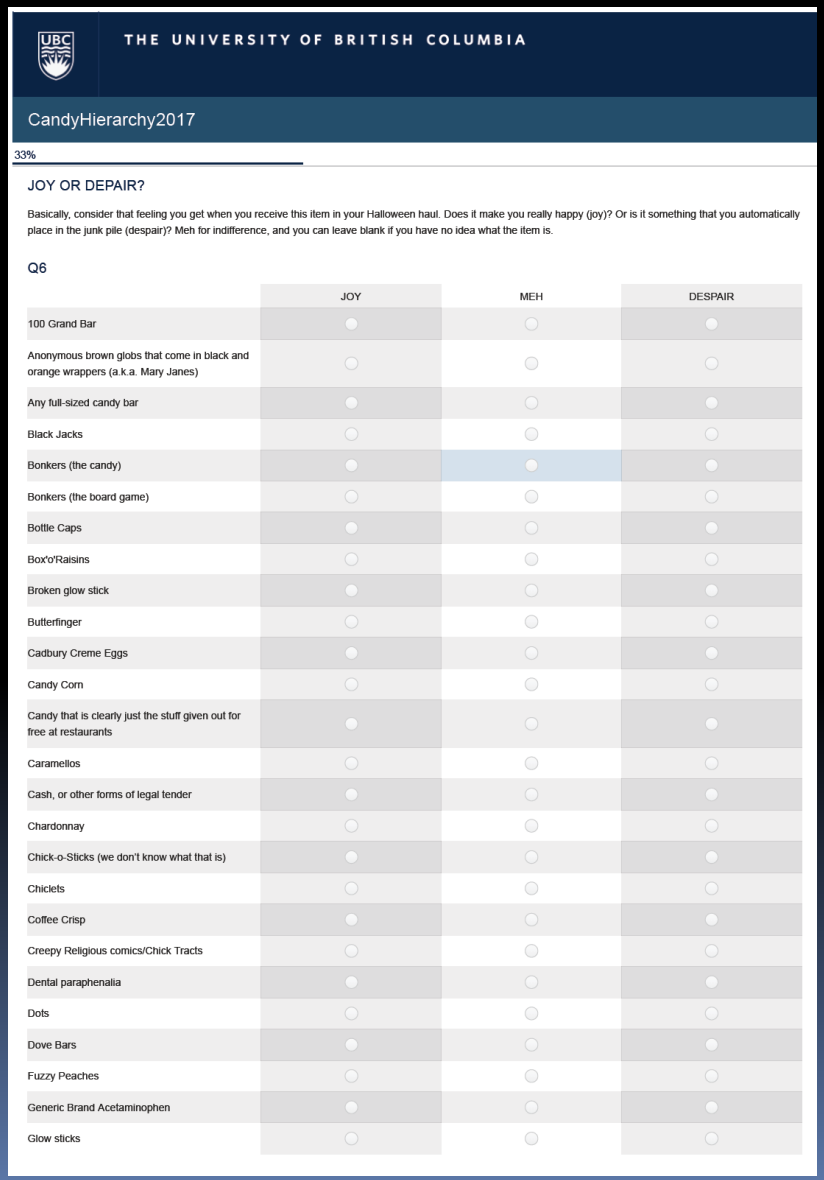
Type here

Q5: STATE, PROVINCE, COUNTY, ETC

Which state, province, county, etc do you live in?

Type here

Survey Questions 2017



THE UNIVERSITY OF BRITISH COLUMBIA

CandyHierarchy2017

33%

JOY OR DEPAIR?

Basically, consider that feeling you get when you receive this item in your Halloween haul. Does it make you really happy (joy)? Or is it something that you automatically place in the junk pile (despair)? Meh for indifference, and you can leave blank if you have no idea what the item is.

Q6

	JOY	MEH	DESPAIR
100 Grand Bar			
Anonymous brown globs that come in black and orange wrappers (a.k.a. Mary Janes)			
Any full-sized candy bar			
Black Jacks			
Bonkers (the candy)			
Bonkers (the board game)			
Bottle Caps			
Box'o'Raisins			
Broken glow stick			
Butterfinger			
Cadbury Creme Eggs			
Candy Corn			
Candy that is clearly just the stuff given out for free at restaurants			
Caramellos			
Cash, or other forms of legal tender			
Chardonnay			
Chick-o-Sticks (we don't know what that is)			
Chiclets			
Coffee Crisp			
Creepy Religious comics/Chick Tracts			
Dental paraphenalia			
Dots			
Dove Bars			
Fuzzy Peaches			
Generic Brand Acetaminophen			
Glow sticks			

Survey Questions 2017

Q7: JOY OTHER

Please list any items not included above that give you JOY.

Type here

Q8: DESPAIR OTHER

Please list any items not included above that give you DESPAIR.

Type here

Q9: OTHER COMMENTS

Please leave any witty, snarky or thoughtful remarks or comments regarding your choices. This is your chance to leave some qualitative data folks! Candy anthropology if you will.

Type here

Survey Questions 2017

This next section is for scientific purposes. Seriously*.

* as in satirically (is that even a word?)

Q10: DRESS

"That dress* (see image below) that went viral a few years back - when I first saw it, it was _____"

- ○ Blue and black
- ○ White and gold

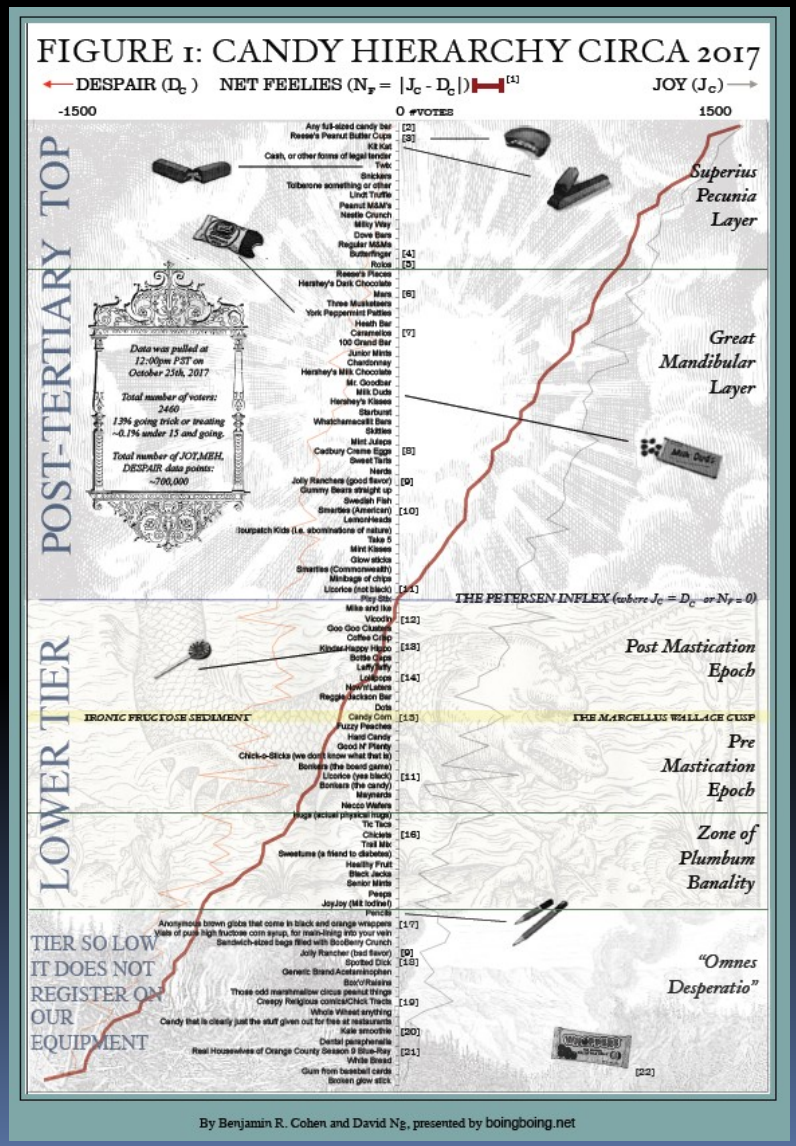
*Also a nod to xkcd's survey (http://tinyurl.com/ndyf5gw)

Q11: DAY

Which day do you prefer, Friday or Sunday?

- ○ Friday
- ○ Sunday

Descriptive Statistics 2017



Dataset Issues: Missing Data

- In class policy:
  - With records that have “some” missing values, impute missing values with nearest neighbor approach
    - Could do average of values
    - Could set to value “unknown”
    - Could remove record
  - If “few” answers, remove record
  - If class value missing, delete record

Dataset Issues: Erroneous Data

- In class policy:
  - Assign out of range value – make sure to exclude from statistics
  - Depends on:
    - Knowledge of data
    - Rest of data in record
  - Analyze data and see if this record is outlier on many answers
    - If so, delete record
  - Correct city + state to state

Dataset Issues: Erroneous Data

- Before gathering data, can add “lie scale” questions to assess truthfulness of individual responders

Dataset Issues: Data Format

- Correct format where correction is obvious

Multiple Source Issues: Different Fields

- Each of the longitudinal data sets have different questions
  - Betty or Veronica in 2015, 2016 sets
  - Friday or Sunday in 2017 dataset
- Need to decide how to handle this

Multiple Source Issues: Different Answers

- 2015 survey only gave choices of “Joy” or “Despair” and didn’t include “Meh”
  - Do we throw out all 5,000+ records?
  - Do we assume “Meh” for all blank answers?
    - Could be bad assumption – missing answer could mean they were not familiar with that particular candy item

Multiple Source Issues: Population Sample

- Can we assume that the population sampled in each of the three datasets is the same?
  - If not, can we still do longitudinal study of the data?

Summary

- Dataset Issues
  - Missing Data
  - Erroneous Data (Intentional or Not)
  - Data Format
- Multiple Source Issues
  - Different Fields
  - Different Answers
  - Population Sample