Visualizing and Learning the Structure in Data

Rebecca Nugent
Department of Statistics, Carnegie Mellon University

PCMI Undergraduate Summer School 2016

July 1, 2016
Who am I?

- Too many Math/Stat degrees via Rice, Stanford, University of Washington (some Spanish just for fun)
- Former Swimmer, Runner, Ultimate Frisbee player
- Lots of travel and big ideas that seemed good at the time
- Currently Professor at Carnegie Mellon Statistics
- Director of the Undergraduate Statistics program
- Academic Advisor concentrating on preparation and transition
- Coach for CMU Athletics Dept
- Research Interests: clustering, record linkage, extracting high-dimensional structure, variable selection
- Applications include: characterizing civilian casualties in Syrian Civil War, modeling semantic organization in young children, modeling student interaction with online tutoring systems, early interventions for patients with unknown depression trajectories
- Learn more at http://www.stat.cmu.edu/~rnugent
Who are you?

On a piece of paper, please jot down quick responses:

- What year are you?
- Major (current or former)
- What was your favorite/topic course in your major (so far)?
- What statistics classes have you had? (e.g. intro, regression, probability)
- What is/was your favorite topic from your statistics classes (if you took any)?
- How comfortable are you with the statistical programming language R? 
  *(1 - have no idea what you’re talking about; 5 - bring it on)*
- If you’ve graduated, what are you doing next year?
- If you haven’t graduated, what do you think you want to do after school?
- Any particular topic you would like to learn more about?
OkCupid Profiles

(created with OkCupid’s permission by Albert Kim)

- 59,946 OkCupid users
- within 25 miles of San Francisco
- active profiles on June 26, 2012 and had been active in the previous year
- at least one picture in their profile

- Collected variables include body type, diet, education, ethnicity, offspring, sexual orientation, pets, religion, and a set of essays
  - My self summary
  - I’m really good at
  - The six things I could never do without
  - The most private thing I am willing to admit

What kinds of questions might we ask about this population?
Distribution of Body Type Based on Gender

User Gender

Reported Body Type

- a little extra
- athletic
- average
- curvy
- fit
- full figured
- jacked
- overweight
- rather not say
- skinny
- thin
- used up

Standardized Residuals: