MS&E 226: “Small” Data
Lecture 1: Introduction

Ramesh Johari
ramesh.johari@stanford.edu

Fall 2015
What is this class about?
“Big” data

We are collecting data at unprecedented levels of granularity.
  ▶ Billions of: Facebook posts, tweets, medical tests, power meter readings...
  ▶ Often arriving faster than we can store and analyze it

Key feature of “big” data:
*Can’t be analyzed on a single machine.*

Requires new algorithms and tools to store, query, and analyze the data.
“Small” data

Data that *can* be analyzed, processed, etc., on a single machine. Keep in mind:

- Advances in technology means even “small” data is getting bigger
  (e.g., 32GB of RAM even on home PCs)
- Most analysis of “big” data starts by understanding “small” data

This class is a user’s manual for “small” data analysis.
Key features

- Conceptual rather than vocational: emphasis on how to reason about different approaches to data analysis

- Comparison and contrast between different approaches: machine learning, (frequentist and Bayesian) statistical inference

- Emphasis on articulating your objective carefully
1. **Summarization** (2 weeks).
   - Given a single data set, how do we summarize it?
   - Basic sample statistics; models; linear and logistic regression; in-sample fit ($R^2$ and residuals).

2. **Prediction** (2-3 weeks).
   - How do we generalize our understanding of a data set to new samples?
   - Binary classification; linear regression and logistic regression as approaches to prediction; model complexity and the bias-variance tradeoff; out-of-sample validation.
3. **Inference** (2-3 weeks).
   - How do we generalize our understanding of a data set to draw inferences about the population or system from which the data came?
   - Frequentist estimation and hypothesis testing; application to linear regression; bootstrap; multiple hypothesis testing. Comparison to Bayesian approaches.

4. **Causality** (2 weeks).
   - How do we determine the effect that changing a system will have?
   - The Rubin causal model, potential outcomes, and counterfactuals; randomized experiments; causal inference from observational data; data-driven decision making.
Course logistics
Basic info

- Website: http://web.stanford.edu/class/msande226
- Piazza: http://piazza.com/stanford/fall2015/mse226
- Details in syllabus on website
- Course assistants:
  Arpit Goel, Carlos Riquelme, Sven Schmit (as needed)
- Discussion sections: Fridays 1:30-2:50 PM
Important dates

No extensions or alternates!

▶ Problem sets
  ▶ PS1: Out 9/22, due 10/1
  ▶ PS2: Out 10/1, due 10/15
  ▶ PS3: Out 10/22, due 11/5
  ▶ PS4: Out 11/5, due 11/19
  ▶ PS5: Out 11/19, due 12/3

▶ Exams
  ▶ In-class midterm: 10/20
  ▶ Take-home midterm: Out 10/20, due 10/22
  ▶ Final exam: 12/7, 3:30-6:30 PM
Evaluation

- Each problem set: 10%
- In-class component of the midterm: 10%
- Take-home component of the midterm: 10%
- In-class final exam: 10%
- Mini-project: 20%
Notes

- Lecture notes will be posted to the site
- Suggested (but not required!) texts:
  - Wasserman, *All of Statistics* ([AoS]).
  - Freedman, *Statistical Models: Theory and Practice* ([SM]).
  - Gelman and Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* ([DAR]).