

Introduction to Data Mining

Scott Powers and Ryan Wang

January 4, 2014

About us

Scott is getting his PhD in statistics from Stanford.

Ryan is a Senior Consultant at The Greatest Good.

Course logistics

- ▶ The goal of the course is to introduce the basic concepts of data mining and teach you to work with structured data.
- ▶ You can do some data analysis in point-and-click applications like MS Excel or Tableau, but most work (e.g. data wrangling or predictive modeling) will require some basic programming.
- ▶ We will use the statistical programming language R, which is heavily used in both industry and academia. Other popular options are Python, SAS, and SQL.
- ▶ Each day we'll (roughly) spend:
 - ▶ Half the time teaching concepts by working through examples
 - ▶ 30 minutes on break, watching the movie Moneyball
 - ▶ Rest of the time exploring datasets that we've prepared

What is data mining?

- ▶ Data mining is the process of extracting insights and understanding from data
- ▶ Revolutions in computing have drastically lowered the cost of collecting, storing, and analyzing data, which has led to big changes in industry, academia, and government:
 - ▶ Facebook uses social data to predict "People you may know"
 - ▶ Retailers (e.g. Amazon and Target) use your purchasing history to advertise other relevant products
 - ▶ TO DO: Academia
 - ▶ City of Chicago uses crime data to target police patrols (link)
 - ▶ Federal government has released over 80,000 datasets since 2009 on <http://data.gov>

Course materials

- ▶ All course materials will be hosted on Dropbox
- ▶ This will include datasets, lecture notes, sample code, and links to external resources
- ▶ Link: `https://www.dropbox.com/sh/bp11oz5fkx3m11s/rJVU1jAoNC`
- ▶ TO DO: Put finalized materials on Dropbox? Or just give them the GitHub link?