CSCE 590B: Big Data Analytics

John R. Rose

Computer Science and Engineering

University of South Carolina

Overview

- Big Data Analytics: What is it?
- First, what do we mean by data?
- Second what do we mean by analytics?
- Third what do we mean by big?

What do we mean by “data”?

- In principal, any kind of data:
  - Corporate sales
  - Email
  - Tweets
  - Sensor output
  - Video
  - Photos
  - Omics
  - Website click streams

What do we mean by "data"?

- What is the structure of this data:
  - Corporate sales – structured (tables in a DB)
  - Email – unstructured (free text)
  - Tweets – unstructured (free text)
  - Sensor output – structured (DB or stream)
  - Video – unstructured
  - Photos – unstructured
  - Omics – semi-structured (XML-like DB)
  - Website click streams – quasi-structured

What do we mean by “analytics”?

- Broadly refers to the method of analysis
- Depends on what we want to learn from the data.
- → method/model used to make sense of the data.
- Depends on the nature of the data.

What do we mean by "analytics"?

- Example: When will social security go broke?
  - Data: Historical data over 50 years
    - Yearly balance
    - Payments in
    - Payments out
    - Size of working population
    - Size of retired population
    - Life expectancy
  - Analytical method: ?

What do we mean by “big”?

– Any thoughts on what we might mean by “big”?

What do we mean by "big"?

- Examples:
  - Genomics data: human genome is 3 billion basepairs
    - "mapping" of human genome exceeds 8 petabytes.
  - New York Times public archive consists of millions of pdf files.
  - Chemical reaction databases containing millions of reactions.
  - Library of Congress collection of tweets: 170 billion tweets (as of January 8, 2013)
  - Others?

What are we actually going to do?

- Depends on your background.
- What do you know about SQL?
- What about NoSQL?
- What do you know about statistical analysis?
  - Regression
  - Clustering
  - Association rules
  - Decision trees
  - Neural networks
  - Support vector machines
  - Hidden Markov models

Plan 0

- Introduce SQL or refresh your SQL memory
- Introduce R in the context of RStudio
- Review basic statistical methods
- Investigate advanced data mining techniques
- Investigate "Big Data" techniques

Plan 0

- Introduce SQL or refresh your SQL memory
  - Next lecture or two will cover SQL
  - Will do more if needed.
- Introduce R in the context of RStudio
  - Instructions for downloading R & RStudio on class webpage
  - Install on your own machine to work at home
- Review basic statistical methods
  - Will use RStudio for hands-on in class

Plan 0

- Investigate advanced data mining techniques
  - Will use methods implemented in R packages
  - No need to rewrite existing tools
  - More important to understand use and limitation of tools
- Investigate "Big Data" techniques
  - Hadoop
  - HDFS
  - PIG
  - HIVE