# CSCE 587 / STAT 587

## Big Data Analytics

# Data

- What do we mean by data?
- Structured
  - Known types and semantics. e.g. records in a database table
- Unstructured
  - e.g. text, email, tweets, photos, text

# Analytics

- What do we mean by analytics?
- Broadly refers to the method of analysis
- Depends on the nature of the data
- Depends on what we want to learn from the data…
- … which helps determine the method / model used

# Analytics

- Example: When will social security run out?
- Data:
  - Annual balance
  - Payments in
  - Payments out
  - Size of working population
  - Size of retired population
  - Life expectancy
- Analytical method?

# Big

- What do we mean by big?
- Examples:
  - Genomics data: human genome is 3.2 billion base pairs
  - New York Times public archive of millions of PDFs
  - Tweets: 500+ million tweets per day
  - Instagram: ~1.3 billion photos shared per day

# What we'll do in this class..

- Introduce R using RStudio
- Review basic statistical methods
- Investigate advanced data mining techniques
- Introduce (or review) SQL (needed for Hive)
- Investigate big data techniques

# What we'll do in this class..

- Introduce R using RStudio
  - Install on your computer, if you'd like: https://posit.co/download/rstudio-desktop/
- Review basic statistical methods
  - We'll use RStudio in class
- Introduce Hadoop
  - HDFS – Hadoop Distributed File System
  - MapReduce
  - Pig
  - Hive
- SQL
  - SQL syntax needed for Hive

# What we'll do in this class..

- Investigate advanced data mining techniques
  - We'll use methods from existing R packages
  - Understand use and limitations of tools
- Investigate big data techniques
  - Hadoop
  - HDFS
  - Pig
  - Hive

# Virtual Machines

Each student will have a unique VM in the form

vm-hadoop-XX.cse.sc.edu

where XX is the student's unique 2-digit number

RStudio: http://vm-hadoop-XX.cse.sc.edu:8787

RStudio account:

- VM account: student
- VM initial password: qwertyCSCE587

Change the VM password to something you'll remember!

# Virtual Machines

On campus computer labs: https://cse.sc.edu/resources/labs

To access your assigned virtual machine from off campus, you'll need to be connected to the VPN. https://cse.sc.edu/resources/vpn