Data Science

Course Overview

Risa Myers
Christopher Jermaine
Marmar Orooji

Rice University

Welcome!

Please fill out the Introductions questionnaire

Welcome!

- Introductions
- Course objectives
- Syllabus / logistics
- Tools

Introductions!

I am

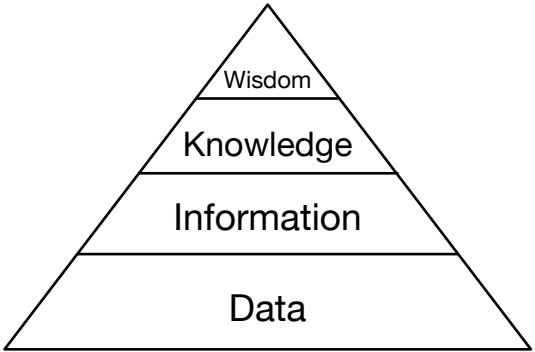
- Risa Myers
- She, Her, Hers
- Assistant Teaching Professor
- rbm2@rice.edu
- Duncan Hall 2062

Course objectives

- Understand the challenges and opportunities involved in using "big" data
- Become familiar with modern machine learning tools used with "big" data
- Be able to implement machine learning algorithms using these tools
- Develop basic skills in querying relational databases and processing data within a relational database
- Be familiar with the theoretical basis and underlying research that motivated the systems and models discussed in class
- Be familiar with storage infrastructure and programming models of large-scale computing

This class is about data science

- Extraction of actionable knowledge from large volumes of data
  - Encompasses methods from:
    - Computer science
    - Statistics
    - Optimization/Applied Math
  - Data Science also encompasses
    - Domain knowledge
    - Communication skills
    - Data management



What is "Big Data"

- Broad, general term
- Refers to tools & techniques for extracting knowledge from massive & complex datasets
- Term appeared in late 90s
- Typically considered data too large to fit in memory of an expensive server machine
  - 5GB in 2002, a couple of terabytes in 2018

"Big Data" historical example

- IBM IMS was a big data system decades years ago!
  - President Kennedy challenged the nation to send an American to Moon
  - Rockwell won the bid to build Saturn V rocket
  - Rockwell needed an automated system to keep track of millions of rocket parts and materials
  - IBM designed IMS in 1966
- Over time, IBM IMS expanded to adapt to exponential increase in data
- Now, IBM IMS can
  - Process more than 50 billion transactions a day
  - Manage 15 million GB of data

The V's of "big" data

- Primary characteristics – 3 Vs
  1. Volume
  2. Variety
  3. Velocity
- Additional characteristics – more Vs
  4. Veracity
  5. Variability
  6. Visualization
  7. Vulnerability
  8. Value
  9. ...

"Big" data – Volume

1. Quantity of data

- Scale varies over time
  - Couple of Gigabytes in 2002
  - Couple of Terabytes in 2018
  - Now, Petabytes/Exabytes
- Example
  - In 2018, global mobile data traffic was 19 EB/month (19 billion GB/month)

"Big" data – Variety

2 Type of data

- Beyond structured data
- Examples
  - Text
  - Image
  - Audio
  - Video
  - Social media

"Big" data – Velocity

3 Speed of data generation/processing

- High rate of data generation
- Real-time data processing
- Examples
  - Facebook ∼600 TB of data per day
  - Google ∼3.5 billion searches per day
    - Real-time processing: ad display for each search query
  - Credit card transactions in US ∼108 million transactions per day
    - Real-time processing: fraud detection

"Big" data – More Vs

4. Veracity – quality of data
   - Contains missing values, invalid entries, wrong formats, ...
5. Variability – changes in quality and / or content over time
   - Due to inconsistent sources
6. Visualization – difficult to create a meaningful visualization
   - Some approaches – data clustering, parallel coordinates, use of tree maps
7. Vulnerability – data breaches
   - May 2016, 167 million LinkedIn accounts & 360 million MySpace users were hacked
8. Value – utility of data
   - Deriving valuable, actionable knowledge

Course scope

- Volume – datasets that are too large to be stored in the memory of a single computer
- Variety – Text & numeric data

Examples of Data Science Tasks

- Given a huge set of per-customer sales data, build a model to predict customer "churn"
- Given a large graph of Medicare payout data, find suspicious (potentially fraudulent) referral patterns
- Given a set of EMR data, find previously unknown side effects (ex: Vioxx and heart disease)
- Given data from an online learning tool find markers that are an early sign of later academic achievement problems
- Many, many more!

What's involved

- You need advanced models to solve challenging prediction/analysis tasks
- You need computer systems that can scale those models to the largest data sets
- You need computer tools that make it easy to implement complicated models

How will we manage and use the Big Data?

- We need tools for manipulating large data sets
- Tools for scalable, distributed computation
- Specifically, we'll learn about:
  - SQL databases
  - Python programming (NumPy, pandas)
  - Distributed file systems
  - The MapReduce paradigm
  - Spark (distributed Big Data manipulation software)

As such, this class...

- Will introduce modern data management software...
  - Relational database systems and SQL
  - Distributed computing frameworks such as Hadoop and Spark
- Will look at approaches to analyzing big data sets...
  - Vectorized programming
  - Data preparation using Pandas
- Assignments will focus on implementing algorithms for analyzing big data and manipulating data with tools

Motivations

- Relational Databases
  - Ubiquitous
  - Scalable & secure
  - Well established storage and retrieval model
  - Foundational for big data systems
- Vectorized Programming
  - Efficient coding
  - Operating on volumes of data concurrently
- Distributed Computing
  - Necessary for data that can't fit in memory
  - Required to process big data in "reasonable" time
- Machine Learning
  - Inferring Information, Knowledge, and Wisdom from the data

Skills you need to succeed in this class

- Should be a reasonable programmer
  - Comfortable with Python
  - One analytical assignment
  - Two assignments use SQL (no knowledge assumed)
  - Remaining assignments use Python
- Attention to details: Submit your homework correctly and on time!!!
- Engage! There is a lot of active learning in this class. Come to class! Participate

Who are you?

Survey results

More skills you need to take this class

- Some background in probability/statistics
  - Common distributions (e.g. Gaussian)
  - Expected value
  - Variance, covariance
  - Norms (e.g. L_1, L_2)

Course norms

- If you don't understand something, say something... you're likely not the only one
- No stupid questions
- We may repeat lectures
- We may adapt assignments
- We may go over some basics that, depending on your background, might be review
- If an assignment is taking too long, speak up! Get help! There may be some knowledge gaps we need to fill

What about overlap with other classes?

- COMP 643 – Big Data
  - Online version of COMP 553
  - Has a database course prerequisite
  - Assumes declarative SQL experience
  - Covers a little more (e.g. Spark streaming)
- COMP 330/543 – Tools & Models - Data science
  - COMP 543 includes more models and theory and some different tools (no pandas, yes TensorFlow)
  - Both assume more familiarity with computing platforms
- COMP 430/533 – Introduction to Database Systems
  - Superset of the database material covered here

Tools

- Google Colab `colab.research.google.com`
- Colab / Jupyter Notebooks
- Amazon Web Services
- Relational DataBase Management System – PostrgreSQL
- pandas
- NumPy
- Hadoop Distributed File System
- Spark

Wrap up

- How can we use what we learned today?
- What do we know now that we didn't know before?