CIS4930/6930 – Introduction to Data Science/Data Intensive Computing
Spring 2014

Dr. Sanjay Ranka and Dr. Daisy Zhe Wang
CISE Department
University of Florida
Jan 6th 2014
Vital Information

• Instructor: Sanjay Ranka, Daisy Zhe Wang
• Office: E532, E456
• Class time: Mon/Wed/Fri 12:50-1:40pm
• Office hours: Wed/Fri 3:00-4:00pm, Mon 1:40-3pm/Wed 1:40-2:30pm or by appointment
• TA: Kun Li, Yupeng Yan (Office hour: TBA)
• Course page will be up later this week:
  http://www.cise.ufl.edu/class/cis6930sp14IDS/
  (read announcements frequently!)
Data science is the study of the generalizable extraction of knowledge from data. It incorporates varying elements and builds on techniques and theories from many fields with the goal of extracting meaning from data and creating data products.

- Data -> Knowledge -> Action
- My Research *probabilistic Knowledge Base construction (SML + DB)*
- Overview of Data Science ➔ next class
Course Goals

• Teach state-of-the-art tools used for Big Data Science application development

  • Data processing: SQL, Map-Reduce, Parallel Computing

  • Data Modeling: Machine Learning and Data Mining, Unstructured Data Analysis

  • Data Representation: Visualization Tools
This Course will

• Teach you how to process data at scale using SQL and map-reduce over cloud services.
• Teach you parallel computing library.
• Teach you basic statistical modeling, machine learning and unstructured data analysis.
• Teach you common programming languages for analytics in R and Python.
• Teach you how to create visualizations.
This Course will NOT

• Touch the frontier of Data Science Research
• Attempt to improve existing Data Science systems and algorithms
• Read research papers or perform research-oriented projects

• First of the three-course series in the Data Science curriculum, followed by
  • Advanced Topics in Data Science
  • Projects in Data Science
Pre-requisites

• Require
  – Data Structures and Algorithms (COP3530)
  – Or equivalent

• Prefer
  – Information and Database Systems I (CIS4301)
  – Statistics and Probabilities (STA 5325/5328)

• Academic honesty
Course Outline

- Part 0: Introduction
- Part 1: Data Manipulation, at Scale
- Part 2: Statistical Analytics
- Part 3: Graph/Text Data Analysis & Communicating Results
- Part 4: Parallel Computing
Part 1: Data Manipulation, at Scale

• MapReduce, Hadoop, AWS, EMR, algorithms, extensions, languages
• Databases, SQL and the relational algebra
• Parallel databases, parallel query processing, in-database analytics
• Key-value stores and NoSQL; tradeoffs of SQL and NoSQL
Part 3: Graph/Text Data Analysis & Communicating Results

- Graph Analytics: PageRank, community detection, recursive queries, iterative processing
- Text Analytics: TF/IDF, conditional random fields, HMM
- Visualization: Communicate data products, visual data analytics
Part 2: Statistical Analytics

• Programming in Python and R
• Basic Data Mining
  – Basic statistical modeling, introduction to machine learning, over-fitting
  – Supervised learning: Linear and Logistic Regression, Classification
  – Unsupervised learning: Clustering, Association Rule mining
Part 4: Parallel Computing

- Concurrency and Data Decomposition
- Message Based Parallelism – MPI
- Thread Based Parallelism – OpenMP
Suggested Readings


Course Evaluation

• Project 1 (25 %) – Map-Reduce on AWS
• Project 2 (25 %) – SQL and Tableau on AWS
• Project 3 (25 %)
• Project 4 (25 %)

• Late submission: 20% per day for up to 5 days.
Computing Resources

- Amazon Web Services
  - need a credit card to create an AWS account
  - $100 AWS credits per student will be provided
  - Will be used for Project 1 & 2
  - Should be enough to complete the projects
  - Beyond the credit limit is at your own cost
Project 1 (25%)

- Work in groups of ~2 people
- Get AWS started (Week 1)
- Finish AWS tutorials on AWS account and S3 setup, create and run a job flow, command line tools, AWS instance types and pricing, EMR, debugging, etc. (Week 2)
- Implementation of a well-defined algorithm over a given dataset using Map-Reduce on AWS (Week 3-4)
- Evaluation: correctness, performance and selected code review
Project 2 (25%)

• Work in groups of ~2 people
• Given dataset with open-ended novel analysis and visualization using SQL and Tableau on AWS (Week 5-8)

• Evaluation: Presentation slides on analysis goal, Selected presentation and demo

Best projects may be presented at the 2nd Data Science Exposition in late April
Grading

Roughly the boundaries will be:

• 90 -- 100 A
• 85 -- 89 B+
• 80 -- 84 B
• 75 -- 79 C+
• 70 -- 74 C
• 60 -- 69 D
• 0 -- 59 F