CAP4773/CIS6930
Projects in Data Science
Fall 2014

Dr. Sanjay Ranka and Dr. Daisy Zhe Wang
CISE Department
University of Florida
August 25th 2014
Vital Information

• Instructor: Dr. Sanjay Ranka, Dr. Daisy Zhe Wang
• Office: E532, E456
• Class time: Mon/Wed/Fri 3:00-3:50pm (8th periods)
• Office hours: Mon/Wed 3:50-4:50pm (Dr. Wang); Mon 2-3pm Fri 4-5pm (Dr. Ranka) or by appointment
• TA: Yang Chen (Office hour: TBA)
• Course page will be up later this week: TBA (read announcements frequently!)
(Big) Data Science

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 (e.g., decision support)
• My Research (*SML + DB*)
  – *In-Database Statistical ML*
  – *Information Extraction from Text, Image, Crowds*
  – *Probabilistic Knowledge Bases*
Course Goals

• Review and apply state-of-the-art tools for Big Data Science application development
  • Data processing: SQL, Map-Reduce, Parallel Computing
  • Data Modeling: Unstructured Data Analysis, Graph Analysis, Probabilistic Graphical Models, Machine Learning and Data Mining,
  • Data Representation: Visualization Tools
  • Applications: Knowledge Extraction, Pattern Finding, Image/Text Extraction/Retrieval, etc.
This Course will

- Review tools and systems to process data at scale using SQL, map-reduce, parallel computing frameworks, etc..
- Review basic statistical modeling, data mining, unstructured data analysis, machine learning..
- Survey literature, state-of-the-art tools, and applications of big data science
- Mentor groups of students on the design, development and presentation of a big data science application based on the parallel frameworks and statistical methods..
- Lead to possible research short papers..
This Course will NOT

• Teach the basics and details of parallel data processing systems (e.g., DBMS, map-reduce, AWS)
• Teach the basics and details of statistical machine learning models and visualization tools
• Improve the state-of-the-art Data Science systems, algorithms and models
• Teach advanced machine learning models (e.g., Bayesian Networks, Markov Logic Networks, sum-product algorithm, MCMC inference..)

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

• Require
  – Data Structures and Algorithms (COP3530) or equivalent
  – Introduction to Data Science or equivalent

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

• Academic honesty
Review Data Science Tools

• **Data Processing using Map-Reduce and SQL, at Scale**
• **Graph/Text Data Analysis & Communicating Results via Visualization**
• **Statistical Analytics and Data Mining**
• **Parallel Computing**
• **Data Science Applications and State-of-the-art Solutions**
Data Processing, 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
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
Statistical Analytics

• Statistical ML library (e.g., Rapid Miner)
• 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
Parallel Computing

- Data Preprocessing
- Message Based Parallelism – MPI
- Thread Based Parallelism – OpenMP
Suggested Readings


• Various papers in data science applications, data science and data mining (e.g., KDD, ICDE, ICDM) – individual project specific
How can I get an A?

Course Evaluation

- Project Proposal (20 %)
- Midterm Evaluation (20 %)
- Final Presentation (25 %)
- Final Report (25 %)
- Class participation (10%)
- Project Novelty (bonus 5%)

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

One of the following:

• Amazon Web Services
• Department Servers
• Department Computing Resources dedicated to Data Science Education
Project (90%)

• Work in groups of ~3 people
• Project proposal (1-2 pages) (mid Sept.)
  – Form the groups
• Project mid-term evaluation (mid to late Oct.)
  – Application definition & Data, Input/output, Model/System/Algorithm Design, Progress, Deliverables, Evaluation Metrics, Result Viz, etc.
• Project final presentation and demo (late Nov. to early Dec.)
• Final Report (Mid Dec.)
Others (10%)

• Class participation (10%)
  – class attendance
    • Learn from each other (e.g., style of presentation, project management, system & algorithms design)
    • Different Data Science Applications and Tools
  – in-class discussions
    • Observations, insight, questions, and critique
    • Recommendations and suggestions
    • Experience and advise
• Project Novelty (bonus 5%): techniques, problem, data, result...
Project Proposal (20%)

• Form groups of ~3 people
• Problem Statement
• Datasets, Data Sources and Data Collection
• Possible Systems, Algorithms and Models
• Expected Results and Visualizations
• Evaluation Metrics
• Deadline: Mid September
• Evaluation Metrics: reports and/or presentations
Grading

Roughly the boundaries will be:

• 90 -- 100 A
• 85 -- 89 B+
• 80 -- 84 B
• 75 -- 79 C+
• 70 -- 74 C
• 65 -- 69 D+
• 60 -- 64 D
• 0 – 59 E