CIS4930/6930 – Data Science: Large-scale Advanced Data Analysis
Fall 2011

Daisy Zhe Wang
CISE Department
University of Florida
Vital Information

• Instructor: Daisy Zhe Wang
• Office: E456
• Class time: Tuesdays 3-5pm, Thursdays 4-5pm
• Office hours: right after the class (one hour)
• TA: none
• Course page:

http://www.cise.ufl.edu/class/cis6930fa11lad/
(read announcements frequently!)
Overview

• **Trend**: Bigger Data and Deeper Analysis
• **Data Science**: Uses advanced analysis over large-scale data to **create data products**
Data Science – A Working Definition

**Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.
Course Goal

• In this course, we will have in-depth discussions of recent publications related to Data Science.

• I will put most emphasis on the systems, applications and algorithms for large-scale advanced data analysis.
This Course will

• Give you exposure to research topics and existing work in Data Science.
• Ask you to critique the papers we are going to read.
• Strongly encourage you to explore new research problems, come up with better solutions and make contribution.
This Course will NOT

• Teach you statistics, machine learning, database systems.
• Teach you programming.
• Teach you how to be an expert in map-reduce, statistical packages, parallel databases.
Expectations

• Require
  – Information and Database Systems I (CIS4301)
  – Data structures and algorithms, Coding (C, Java)
  – Good Maths and Statistics Background Knowledge

• Encourage
  – Actively participate in discussions in the classroom
  – Read Data Science literature in general
  – Experience in Machine Learning, NLP, Data Mining

• Academic honesty
Course Outline

• Data Analysis I – Systems and Frameworks
• Data Collection, Cleaning, and Integration
• Data Analysis II – Applications and Algorithms
• Interface Design and Data Visualization
Text Books

• Not required, but recommended.
• Class notes + papers.
Additional Reading Pointers

• Data Science Summit (Strata) (http://www.datascientistsummit.com/)
• Kaggle Competitions (http://www.kaggle.com/)
• Data Science course at Berkeley (http://datascienc.es/)
• Conferences and Journals
  – VLDB, ICDE, SIGMOD
  – CIDR, KDD, ICDM ...
Grading

• Homework (25 %)
• Project (55 %)
• Presentations (20 %)
• Participation (5% bonus)
• Novelty in Project (5% bonus)

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

How can I get an A?
Homework (25%)

• Literary reviews
(due before class in **hard copy with your name and ID – do NOT send email**)
  – Main contributions (goals, techniques, evaluations)
  – Positive Critiques
  – Negative Critiques

• This Thursday:

• Next Thursday:
Presentation (20%)

• Select 1-2 papers from the reading list (I will announce how to sign up)
• Prepare 1 hour presentation
• Send me the presentation (ppt) 1 week before the presentation in class
• Improve presentation according to feedback
• Deliver the presentation
• Lead the discussion
Project (55%)

• Work in groups of 2-3 people
• Project proposal (1-2 pages) (Sep 27)
  – Form the groups
• Project mid-term evaluation (Nov 3/8)
  – Novelties, Progress, and Deliverables
• Project final presentation and demo (Dec 6/8)
• Final Report (Dec 15)
Bonuses (5% each)

- Class participation
- Novelty in the Project
An Overview of Research Topics in Data Science
Goal of Data Science

Turn data into data products.
Data

- Application Databases
- Wireless Sensor Data, Seismic, Astronomy Data
- Text Data (Webpages, Wikipedia, Emails, Enterprise Documents)
- Social Media Data (Twitter, Blogs, Social Networks)
- Software Log Data (Server, API, Database Logs, Click Streams), Images, Videos, Music
- Scientific Data, Medical, Microarray, Genome Data

Data is getting Larger and more Diverse
Data Products – Twitter

- Text Analysis – Spam Filter/Similarity Search
- User Sentiment/Satisfaction/Feedback
- News Breakout
- Trend and Topics

200 million users as of 2011, generating over 200 million tweets and handling over 1.6 billion search queries per day
Data Products – Netflix

• Personalized Movie Ratings
• Movie Recommendations
• Similar Movies
• Movie Categories (e.g., 80’s movie with a strong female lead, Kung Fu movies)

BlockBuster is out of the business ...
Data Products – LinkedIn/Facebook

• People you may know
• Applications you may like
• Jobs/Events you might be interested
• Classifier for bad users and bad content
• With high accuracy, Facebook can guess whether you are single or married

Who does not have LinkedIn/Facebook Account?
Data Products – Splunk

• Degradation, Failure Detection
• Identify Security Breach
• Event Monitoring
• Troubleshoot Tools
• Cross-platform Event Correlation

Founded 2004, Rumor has it – Close to IPO
Data Products – Google

• Web Search
• News Recommendation Engine
• Google Map
• Google Ads
• Google Analytics

Still the hottest IT company to work for now -- Microsoft of 90’s, IBM of the 70’s
Techniques used in Data Science

• Statistics
• Machine Learning
• Data Management
• Visualization
• HCI ...
Related Research Areas

• Databases, Systems
• Data warehouse, ETL, Data Cubes
• Statistics and Machine Learning
• Data Mining
• Data Visualizations
• Human Computer Interactions
• Privacy
Challenges in Data Science

• Preparing Data (Noisy, Incomplete, Diverse, Streaming ...)

• Analyze Data (Scalable, Accurate, Real-time, Advanced Methods, Probabilities and Uncertainties ...)

• Represent Analysis Results (i.e. data product) (Story-telling, Interactive, explainable...)
“Sexy Job” in the next 10 years

“The sexy job in the next ten years will be statisticians... The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill.”

-- Hal Varian, Google Chief Economist, 2009
“Sexy Job” in the next 10 years

“The sexy job in the next ten years will be data scientists... The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill.”

-- Hal Varian, Google Chief Economist, 2009
Skill Set of a Data Scientist

• Data Management
  – Data collection, storage, cleaning, filtering, integration ...

• Statistics and Machine Learning
  – Data modeling, inference, prediction, pattern recognition ...

• Interface and Data Visualization
  – HCI design, visualization, story-telling ...
The Life of Data (state-of-the-art)

Collect → Clean → Integrate → Analysis → Visualization

Data Sources

Users

Interface
Data Collection – Data in the First Mile

• Collect data effectively
  – Sensors: Acquisitional data collection
  – Surveys: Model-based Re-ordering of questions (Usher)

• Collect Structured Data from the Web
  – WebTables
  – Scuba
  – Extraction from Text (more later...)
Data Cleaning

• Deal with dirty, noisy, incomplete data
  – E.g., Census, Sensor Networks, Text Extractions

• Interactive Data Transformation and Cleaning
  – Data Wranger, 2011
  – Potter’s Wheel, 2001

• Pay-as-you-go user feedback

• More automatic (avoid boring tasks) – still a research challenge
Data Integration

• Merging multiple data sources
• Schema Mapping
• Entity Resolution (SERF)
• Querying over Multiple data sources
• Pay-as-you-go data integration
• Probabilistic data integration
• Fusion Table: web-centered data integration
Data Analysis (I)
Systems and Frameworks

- MADLib
- Mahout
- Spark
- Map-Reduce Online
- SciDB
- RIOT
- MauveDB
- DataPath
- Dremel: Google Analytics
Data Analysis (II)
Applications and Algorithms

• Text Analysis (OpenIE, Computational Journalism, DBLife, BayesStore)
  – Classification
  – Information extraction
  – Relation extraction
  – Reference reconciliation (Co-reference)
• On-line advertisement (MADLib)
• Fraud Detection (Splash)
Data Analysis (II)
Applications and Algorithms (cont.)

• Risk Management (MCDB)
• Probabilistic Databases (MystiQ, Trio, BayesStore..)
• Mechanical Turks (CrowdDB, CrowdFlow, CrowdSearch...) hot topic now!
• Recommendation Systems, Log Analysis ...
Interface Design and Data Visualization

• Polaris – Tableau (Stanford)
• Spreadsheet (e.g., Excel) (MIT)
• Interactive Querying interface (Berkeley)
  – CONTROL
  – Interactive Cleaning ($D^p$)
• Query-by-Example (IBM)
  – ManyEyes
Course Project (I) – An Application of Data Science

• Using a Framework We discussed in class
  – Relational Database, Parallel Database
  – Hadoop, Map-Reduce, Mechanical Turk

• Using a statistical/machine learning algorithm
  – Text Analysis (Classification, Information Extraction, Entity Resolution)
  – A/B Testing, MCMC simulation ...
Course Project (II) – Improving Data Science Framework

• New Interface Design for data cleaning/integration/querying/feedback
• New Technology to improve Crowdsourcing Service
• New Framework supporting Data Science applications
Research Directions

• Interactive Query-Driven Text Analysis
  – Information/Relation Extraction
  – Reference Reconciliation
  – Classification

• Pay-as-you-go Machine Learning
  – On-line Learning
  – Quality Control, Lineage

• Probabilistic Knowledge Base

• Probabilistic Database + Crowd-sourcing
Homework Today

• Think about Project, form groups early (project proposal due Sep 27)
• Reviews due