Georgia Tech

DATA ANALYTICS USING DEEP LEARNING

GT 8803 // FALL 2018 // JOY ARULRAJ

LECTURE #01: COURSE INTRODUCTION

CREATING THE NEXT®

TODAY'S AGENDA

- Course Objectives
- Course Logistics
- Course Overview

WHY SHOULD YOU TAKE THIS COURSE?

- There are many challenging problems in data analytics using machine learning (ML)
- Systems + ML developers are in demand
- If you are good enough to write code for a ML-driven data analytics system, then you can write code on almost anything else

COURSE DESCRIPTION

- This is a **research-oriented course**
  - Very much a “take what you want”
  - You will not be tested (exams, assignments) or taught (lectures) traditionally
- Instead, you will engage in research
  - Read, comment on, and discuss papers
  - I won’t be teaching: we will discuss together
  - Pursue a research project

COURSE DESCRIPTION

- That said: this is not an easy course
    - The research project requires dedication and ingenuity
    - Dealing with unpredictable research outcomes
    - If you have never done research, talk to me!

COURSE OBJECTIVES

- Learn about cutting-edge research topics in data analytics using machine learning
- Learn about modern practices in systems programming and machine learning
- We will cover state-of-the-art topics
- This is **not** a course on classical database systems

COURSE OBJECTIVES

- Students will become proficient in:
  - Critiquing and presenting technical papers
  - Identifying and tackling research problems
  - Writing correct and performant code
  - Reviewing, testing, and documenting code

BACKGROUND

- I assume that you have already taken an intro course on database systems & ML
- At a high level, you should be familiar with topics such as (or be willing to pick them up):
  - Query processing
  - Query optimization
  - Deep learning
  - Reinforcement learning

BACKGROUND

- You should be comfortable with programming in languages such as:
  - Python or C/C++
- For your project, you would be leveraging machine learning frameworks such as:
  - Tensorflow or PyTorch

BACKGROUND

- I am happy to have people from different backgrounds
  - But talk to me if you're not sure
  - Talk to me if you are pursuing MS/PhD in a different field

COURSE LOGISTICS

- Office: KACB 3324
- Email: jarulraj@cc.gatech.edu
  - Mention “CS 8803” in email title
- Course Policies + Schedule
  - Refer to course web page
  - If you are not sure, ask me
- Course email address
  - gt.8803.ddl.fall.2018@gmail.com

OFFICE HOURS

- Immediately before class
  - Mon/Wed 3:30 – 4:30 PM
- Things we can talk about:
  - Issues related to research projects
  - Paper clarifications/discussions
  - Relationship advice

WAITLIST

- Add your name to the sign-up sheet
  - I will add you to the class roster

CLASS STRUCTURE

- Seminar course
  - We read papers and talk about our feelings
- Since there are no textbooks or exams, I need to be convinced that you're learning
  - Everybody reads the assigned paper before class
  - One person presents the paper for an hour
  - Extra time for brainstorming sessions in which we will collectively discuss and develop new ideas related to the covered paper

READING REVIEWS

- One page per paper
- Standard conference review template
  - Overview
  - Three strong points
  - Three weak points
  - Technical questions or comments for the class
  - Looking for **innovative ideas** on new research directions related to the paper

READING REVIEWS

- If you are not presenting the paper, then you must turn in the review **by 11:59pm EST** on the night before the class
- Submit it via email to **the course email address and the presenter**
- Late submissions will not be accepted
- You can miss up to three submissions

PAPER PRESENTATIONS

- In depth description and analysis of the paper
- May need to incorporate information from supplemental sources
- Should be **60 minutes** long and then 20 minutes remaining for questions
- Send your presentation slides to the course email address **48 hrs** prior to your presentation

PAPER PRESENTATIONS

- If you are not sure what parts of the papers to present, ask me
- You are encouraged to reach out to the authors of the paper regarding the availability of presentation slides
  - If you borrow from other presentations, be sure to provide attribution

PAPER PRESENTATIONS

- You will be expected to lead a stimulating discussion of the questions & comments submitted by your peers in their reviews
  - You should engage the class by asking questions to carry the discussion forward
  - You are strongly encouraged to **propose new ideas related to the paper** and discuss with the class

PAPER PRESENTATION

- Lectures will be divided into two parts
  - Paper presentation (driven by a student/me)
  - Discussion (driven by me)
- For the discussion part, I will initiate an open-ended debate on the paper
  - What could the authors have done better?
  - What they did they do well?
  - Be prepared with your questions about the paper!

PAPER PRESENTATIONS

- Send me a PDF copy of your slides immediately after presenting in class
  - Be sure to include your name in the meta-data
  - I will publish the slide-deck on the course website

RESEARCH PROJECT

- Semester-long research project
  - Main component of the course
  - Everyone has to work in a team of **two people**
- Projects must:
  - Be relevant to the topics discussed in class
  - Require a significant programming effort from all team members
  - Be unique (i.e., two groups may not choose the same project topic)

RESEARCH PROJECT

- Build/design/test something new and cool!
  - Should be “original”, e.g., re-implementing an algorithm from a paper is not sufficient
  - Goal: Projects should eventually lead to a conference paper
  - Amaze us (of course, we will help!)

RESEARCH PROJECT

- Each team will present their proposals to the class to get feedback from their peers
  - Ask me if you are looking for ideas or a partner

PROJECT MILESTONES

- Project deliverables:
  - Week 6: Proposal Presentation + Report (3 pages)
  - Week 12: Project Status Update Presentation + Report (6 pages)
  - Week 18: Final Presentation + Report (10 pages)
  - Weeks 10 & 16: Code Reviews
  - Week 18: Code Drop

PROJECT PROPOSAL

- **Ten** minute presentation to the class that discusses the high-level topic
- Each proposal must discuss:
  - What is the problem being addressed?
  - Why is this problem important?
  - How will the team solve this problem?
  - How will you validate your implementation?
  - How will you evaluate its performance?

PROJECT STATUS UPDATE

- **Ten** minute presentation to update the class about the current status of your project
- Each presentation should include:
  - Current development status
  - Whether anything in your plan has changed
  - Any thing that surprised you

FINAL PRESENTATION

- **Ten** minute presentation on the final status of your project
- You’ll want to include any performance measurements or benchmarking numbers for your implementation

CODE REVIEWS

- Each group will be paired with another group and provide feedback on their code at least two times during the semester
- Grading will be based on participation

CODE DROP

- A project is **not** considered complete until:
  - All comments from code review are addressed
  - The group provides documentation in both the source code and in separate Markdown files
  - The project includes test cases that correctly verify that implementation is correct
  - The project includes benchmarks and data sets used for the empirical analysis

GOOD EXAMPLE

- Read 5+ state-of-the-art papers on video analytics using machine learning
- Develop a novel query optimization technique that improves performance
- Implement the technique in a ML framework and demonstrate its impact

BAD EXAMPLE

- Run a standard benchmark suite on a few systems and show a bunch of graphs

PROJECT TIPS

- Innovation will be highly appreciated!
- Try to present and read supplementary papers related to your project topic
- Start early so that you can learn the ML and systems programming techniques required for your project
  - Pitch your project ideas to me during Weeks 3 & 4

PROJECT RESOURCES

- During your project proposal, you should mention the resources will you need
  - Software
  - Hardware
  - Data sets or workloads
- Computing resources will be made available on a case-by-case basis

PROJECT RESOURCES

- You are encouraged to reach out to the authors of a paper regarding the availability of data sets and workloads in advance **before your proposal**

GRADE BREAKDOWN

- 30%: Reading Reviews + Class Participation
- 20%: Paper Presentations
- 10%: Project Intermediate Report
- 30%: Project Final Report
- 10%: Project Presentation and Poster

GRADING POLICY

- I will grade on an absolute scale
  - All of you could get A’s
  - Emphasis is on learning rather than testing you
  - If your project is truly amazing, you get an automatic A!

COURSE MAILING LIST

- On-line Discussion through Piazza:
  - https://piazza.com/class/jkt7fvdtqzh64t
- If you have a technical question about the projects, please use Piazza
  - Don't email me directly
  - All non-project questions should be sent to me

WHY SHOULD YOU TAKE THIS COURSE

- There are many challenging problems in database systems & machine learning
- Systems + ML developers are in demand
- If you are good enough to write code for a ML-driven data analytics system, then you can write code on almost anything else

BIG DATA ERA

- We have more data now than ever before
  - 2.5 million terabytes of data created each day
  - Accelerating with growth of the Internet of Things
- Every minute:
  - YouTube: 400 hours of video uploaded
  - Instagram: 50 thousand photos uploaded
  - Twitter: 500 thousand tweets posted

Source: How much data do we create, Forbes, August 2018

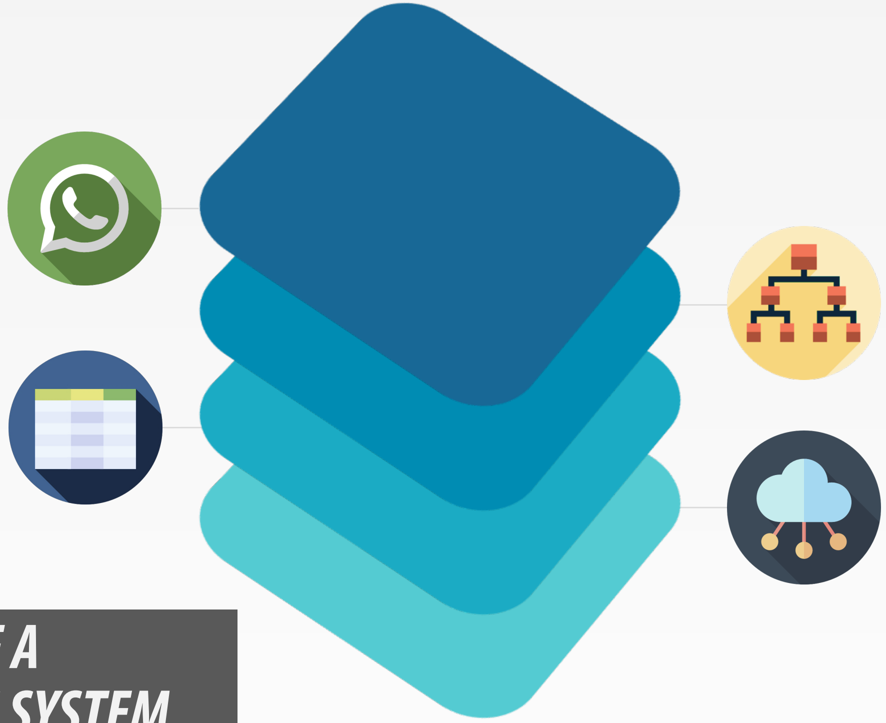
UNSTRUCTURED DATA & QUERIES

- Traditional DB research focuses on structured data and queries
  - Unstructured Data: Images, videos, and speeches make up the bulk of the generated data
  - Unstructured Queries: Novice data analysts can’t construct sophisticated database queries
  - Need to integrate ML techniques to handle unstructured data & queries

WHY IS THIS IMPORTANT NOW?

- This will enable lots of important applications
  - Personal memex
    - Store and retrieve everything a person sees and hears
  - Developmental psychology
    - Psychologists can quickly distill behavioral data in videos
  - Data science
    - Data analysts can ask queries in natural languages
  - Public transportation
    - Intelligent dash cams can help drivers avoid accidents

THEMES OF THE COURSE



MACHINE TRANSLATION

STORAGE MANAGEMENT

DATA ANALYTICS

HARDWARE ACCELERATION

LAYERS OF A DATA ANALYTICS SYSTEM

THEMES OF THE COURSE

- Machine Translation
  - Natural language query processing
- Data Analytics
  - Video analytics, Speech analytics, Data exploration
- Storage Management
  - Non-volatile Memory
- Hardware acceleration
  - FPGAs, GPUs

NEXT CLASS

- First paper review is due on Tuesday night
- Sign up for top 5 papers you’d like to present
- Links will be sent out on Piazza

ALL ABOUT YOU

- Introduce yourself
  - Which department/program you are in?
  - What are your goals for this course?
  - What research topics are you excited about?