There are multiple CSE6242 sections. This is the course homepage for **online CSE6242OAN,O01**.

CSE6242OAN,O01, Spring 2019

Data and Visual Analytics

Georgia Tech, College of Computing

[Prof. Duen Horng (Polo) Chau]
Associate Professor, School of Computational Science & Engineering
Associate Director, Master of Science in Analytics
Machine Learning Area Leader, College of Computing
Director, Polo Club of Data Science

Linkedin Twitter Google Scholar YouTube

This course will introduce you to broad classes of techniques and tools for analyzing and visualizing data at scale. It emphasizes on how to *complement* computation and visualization to perform effective analysis. We will cover methods from each side, and hybrid ones that combine the best of both worlds. Students will work in small teams to complete a significant project exploring novel approaches for interactive data & visual analytics.

Course Goals

- Learn **visual** and **computation** techniques and tools, for typical data types
  - Learn how to **complement** each kind of methods
  - Gain a **breadth** of knowledge
- Work on **real datasets and problems**
- Learn **practical** know-how (useful for jobs, research) through significant hands-on programming assignments

Acknowledgement

We thank the generous support of **Amazon Web Services** and **Microsoft Azure** for free cloud credits, **Intel** for curriculum development of the memory mapping module (scaling up algorithms with virtual memory), and **Tableau** for data visualization software.

Announcements and Discussion

> **We use Piazza for all announcements and discussion. Everyone must join this class's Piazza (link available on Canvas). Double check that you are joining the correct Piazza!** There are multiple concurrent course sections with the same name and course number taking place, e.g., online for OMSA and OMSCS, and campus for Atlanta-based students.

The fastest way to get help with homework assignments is to post your questions on Piazza. That way, only our TAs and instructor can help, your peers can too.

If you prefer that your question addresses to only our TAs and the instructor, you can use the private post feature (i.e., check the "Individual Students(s) / Instructors(s)" radio box).

While we welcome everyone to share their experiences in tackling issues and helping each other out, but please do not post your answers, as that may affect the learning experience of your fellow classmates.

For special cases such as failed submissions due to system errors, missing grades, failed file uploads, emergencies that prevent you from submitting, personal issues, you can contact the staff using a private Piazza post.

> Canvas will be used for submission of assignments and projects, but not for announcements or discussion.

Course Staff & Office Hours

TAs plan to hold office hours starting week 2, except on Georgia Tech holidays (e.g., thanksgiving, MLK day, spring break). Each office hour session will be run by at least one TA, and is 1 hour long. See GT's academic calendar for the full list of holidays (https://registrar.gatech.edu/calendar). We will spread the office hours across weekdays, and across time of the day. We will announce the office hour times.

We will hold office hours via Slack, where the TA running the office hour will be responsive. We will share information about how to join the appropriate Slack group.

Please note that you are always welcome to ask questions on Piazza. Office hours supplement Piazza, and do not replace it.

Course Schedule

For **all dates** used in this course, their times are **23:59 Anywhere on Earth (11:59 pm AoE)**. For example, a due date of "January 8" is the same as "January 8, 23:59pm AoE". Convert the times to your local times using a Time Zone Converter.

Wk	Dates		Topics	Homework (HW)	Project
1	Jan	7-11	* Course Introduction * Analytics Building Blocks * Data Science Buzzwords	**HW1 out** Fri, Jan 11	
2		14-18	* Data Collection * SQLite * Data Cleaning * Class Project Overview * Code Back-up & Version Control		
3		21-25	* Data Integration * Data Analytics, Concepts and Tasks		
4		28-1	* Visualization 101 * Fixing Common Visualization Issues	**HW1 due** Fri, Feb 1 (Sat, 06:59 ET) (Sat, 11:59 UTC) **HW2 out** Fri, Feb 1	
5	Feb	4-8	* Data Visualization for Web (D3)		**Form project teams by** Fri, Feb 8
6		11-15	* Scalable Computing: Hadoop * Scalable Computing: Pig * Scalable Computing: Hive		
7		18-22	* Scalable Computing: Spark * Scalable Computing: HBase	**HW2 due** Fri, Feb 22 **HW3 out** Fri, Feb 22	
8		25-1	* Classification * Visualization for Classification		**Proposal Document due** Fri, Mar 1 **Proposal Presentation Slides and Video due** Fri, Mar 1
9	Mar	4-8	* Introduction to Clustering		

Wk	Dates		Topics	Homework (HW)	Project
10		11-15	* Graph Analytics * Ensemble Method * Scaling up Algorithms with Virtual Memory	**HW3 due** Fri, Mar 15 (Sat, 07:59 ET) (Sat, 11:59 UTC) **HW4 out** Fri, Mar 15	
11		18-22	[Work on Project]		
12		25-29	[Work on Project]		**Progress Report due** Fri, Mar 29
13	Apr	1-5	* Text Analytics		
14		8-12	[Work on Project]	**HW4 due** Fri, Apr 12	
15		15-19	[Work on Project]		**Poster Presentation Video due** Fri, Apr 19 **Final Report due** Fri, Apr 19
16		22-26	Wrap up peer assessment		**Poster Presentation Video grading starts** Mon, Apr 22 **Poster Presentation Video grading due** Fri, Apr 26

This course can be very tough for many!

> **WARNING!** You are expected to quickly learn many things simultaneously, and for some materials you will need to learn them on your own (e.g., Linux commands, for working with MS Azure/Amazon AWS). This can be very intimidating for many students.

The amounts of time students spend on this class **greatly vary**, based on their backgrounds, and what they may already know. Some former students told us they spent about **40-60 hours** on each homework assignment (we have 4 big assignments, and no exams), and some reported much less. For example, for the homework assignment about D3 visualization programming, students who are completely new to javascript, css, and html likely will spend significantly more time than their peers who have already tried them before. Some former students who do not have a computer science background found the homework assignments challenging, would take significant time and effort, but were rewarding, fun, and "do-able."

Students have at least 3 weeks to complete each homework assignment. Some students waited until the last week, and could not finish. It is critical to plan ahead and prepare for the significant time needed.

> Almost all homework assignments involve **very large amount of programming tasks** (which naturally means likely a lot of debugging will be needed, thus can be time consuming). **You should be proficient in at least one high-level programming language** (e.g., Python, C++, Java), and is **efficient with debugging principles and practices**. If not, you should **NOT** take this course. Instead, **you should first take CSE 6040 (for OMS Analytics students)** and, if needed, CS 1301 and CS 1371 as well.

Some programming assignments involve high-level languages or scripting (e.g., Python, Java, SQL etc.). Some assignments involve web programming and D3 (e.g., Javascript, CSS, HTML). For example, an assignment on Hadoop and Spark may require you to learn some basic Java and Scala quickly, which should not be too challenging if you already know another high-level language like Python or C++. **It is unlikely that you all know tools/skills needed in the programing tasks, so you are expected to learn many of them on the fly.**

Basic linear algebra, probability and statistics knowledge is also expected.

Homework

We have 4 big assignments in total (subject to change). Visit this course's Canvas site for the assignment documents. See the schedule table above for deliverable due dates.

- [10%] HW1: Collecting & visualizing data, SQLite, D3 warmup, OpenRefine
- [15%] HW2: D3 Graphs and Visualization
- [15%] HW3: Hadoop, Spark, Pig and Azure
- [10%] HW4: Scalable PageRank via Virtual Memory (MMap), Random Forest, Scikit-Learn

Project

See project description. See the schedule table above for deliverable due dates.

Grading Policy

1. There will be **4 homework assignments**. Together, they are worth 50% (10%, 15%, 15%, 10%) of the course grade.
2. There will be one course **group project** worth 50% of the course grade. The project components are:
    1. Proposal (7.5% of course grade)
    2. Proposal presentation (5%) (video recording)
    3. Progress report (5%)
    4. Final poster presentation (7.5%) (video recording)
    5. Final report (25%)
3. You must achieve an overall weighted average of 60% to pass the course.
4. All deliverables will be graded by our TAs, except the project poster presentation, which will be peer-graded.
5. When assigning course grades, I will start with the standard grade thresholds (90, 80, etc.). I may lower (and never raise) the thresholds (i.e., to your benefits). For example, I may use 88 instead of 90.

Plagiarism, Collaboration Policy, and Student Honor Code

- All course participants (myself, teaching assistants, and learners) are expected to know and abide by the Georgia Tech Academic Honor Code.
- Ethical behavior is extremely important in **all facets of life**.

1. Plagiarism is a **serious offense**. You are responsible for completing your own work. You are not allowed to copy and paste, or paraphrase, or submit materials created or published by others, as if you created the materials. All materials submitted must be your own.
2. You may discuss high-level ideas with other students at the "whiteboard" level (e.g., how cross validation works, use hashmap instead of array) and review any relevant materials online. However, each student must write up and submit his or her own answers.
3. All incidents of suspected dishonesty, plagiarism, or violations of the Georgia Tech Honor Code will be subject to the institute's Academic Integrity procedures (e.g., reported to and directly handled by the Office of Student Integrity (OSI)). **Consequences can be severe, e.g., academic probation or dismissal, grade penalties, a 0 grade for assignments concerned, and prohibition from withdrawing from the class.**

Late Policy and Due Dates

- All homework and project deliverables are due at the times shown in the Course Schedule. These times are subject to change so please check back often. Convert the times to your local times using a Time Zone Converter.
- Every homework assignment deliverable and every project deliverable comes with a 48-hour "grace period". You do **not** need to ask before using this grace period.

  Your deliverable may be submitted (and resubmitted) up to 48 hours after the official deadline **without penalty**, but Canvas will mark your submission as "late".

  Canvas automatically appends a "version number" to files that you re-submit. You do not need to worry about these version numbers, and there is no need to delete old submissions. **We will only grade the most recent submission.**

- Any deliverable submitted after the grace period will get **0** credit. We recommend that you submit your work before the grace period begins.

- We will **not** consider late submission of any missing parts of a deliverable. To make sure you have submitted everything, download your submitted files to double check. If your submitting large files, you are responsible for making sure they get uploaded to the system in time. You have 48 hours to verify your submissions!

- No penalties for medical reasons or emergencies. And should they arise, you must contact the Dean of Students office. Doctor's notes, medical documentation, explanation of emergencies, etc. should be submitted to the Dean's office. After their office receives the information, we will notify me on your behalf.

Timing Policy

- The course videos follow a logical sequence that includes knowledge-building and experience-building (assignments).
- Assignments should be completed by their due dates, in order for timely peer assessment. Peer assessments should also be completed by their due dates, to give timely feedback.
- You will have access to the course content for the scheduled duration of the course.

Attendance Policy

- This is a fully online course.
- Login on a regular basis to complete your work, so that you do not have to spend a lot of time reviewing and refreshing yourself regarding the content.

Netiquette

- Netiquette refers to etiquette that is used when communicating on the Internet. Review the Core Rules of Netiquette. When you are communicating via email, discussion forums or synchronously (real-time), please use correct spelling, punctuation and grammar consistent with the academic environment and scholarship[1].
- We expect all participants (learners, faculty, teaching assistants, staff) to interact respectfully. Learners who do not adhere to this guideline may be removed from the course.

1. Conner, P. (2006-2014). Ground Rules for Online Discussions, Retrieved 4/21/2014 from http://teaching.colostate.edu/tips/tip.cfm?tipid=128

Dataset Ideas (may need API, or scraping)

- Google Dataset Search
- Google public datasets. Thanks Revant!
- Awesome Public Datasets. Thanks Marcel Gwerder!
- NYC Taxi data for 2013 (suggested by Chris Wong). 2013 Trip Data (11.0GB). 2013 Fare Data (7.7GB). Visualization for a days trip. Thanks Jitesh.
- Large datasets publicly available. Thanks Gopi!
- Georgia Tech's campus data (has APIs): bus info, directory, building, T-square, room reservation, building facilities usage (e.g., electricity, lights, A/C, etc.), Oscar/course info/registration, etc.
- Yahoo WebScope
- Data.gov: U.S. Government's open data
- IPEDS data: Postsecondary education data from National Centre for Education Statistics
- Bureau of Labor Statistics data
- Uber data: Anonymized data from over 2 billion trips
- Freebase
- Yelp
- Microsoft Academic Graph
- Numerous APIs from Google (e.g., Maps, Freebase, YouTube, etc.)
- Zillow: real estate listing site
- Numerous graph datasets (large and small): SNAP, Konect
- Movies data: IMDB
- List of lists of datasets for recommendations.
  Thanks Jon!
- Million song dataset by Echo Nest.
  It contains not only the basic information of songs (artist, genre, year, length etc), but also some musical features(like tempo, pitch, key, brightness).
  Thanks Minwei!
- Dataset about soccer games, players, clubs.
  No API, but easy to scrape.
  For a soccer player: transfer history, performance, nationality, birth date, etc.
  For a soccer club: performance, squad, etc.
  Thanks Ding!
- The Free 'Big Data' Sources Everyone Should Know
- Quandl - a dataset search engine for time-series data.
  Thanks Henry!
- UCI also has a collection of links to various datasets sorted for various tasks (Classification, Regression, etc)
  Thanks Vinodh!
- Amazon AWS Public Data Sets (Thanks Jonathan!)

- KDD Cup: annual competition in data mining, like Kaggle
- Academic domain: Microsoft Academic Search, DBLP
- Retrosheet: MLB statistics (Game/Play logs)
- Classification datasets
  Thanks Amish!
- Various geophysical datasets for the oceans (magnetism, gravity, seismology, etc).
  Thanks Ryan!
- Social trends (Thanks Jonathan!)
- Beer data (Thanks Jonathan!). Website offline :(. Older version at web.archive.org
- Academic torrents (terabytes) (Thanks Vaibhav!)
- Article Search API from the New York Times (all the way back to 1851!) (Thanks Guido!)
- (Kayak: flight, hotel, car, etc.)
- Data Science Initiative - Microsoft Research has various datasets and access to tools that can aid in data science research

Resources

All content and course materials can be accessed online. There is no textbook for this course.

All Georgia Tech students have FREE access to https://www.safaribooksonline.com, where you can find a huge number of highly rated and classic books (e.g., the "animal" books) from O'Reilly and Pearson covering a wide variety of computer science topics, including some of those listed below. Just log in with your official GT email address, e.g., jdoe3@gatech.edu.

Software engineering; become a better programmer and developer

- Design Patterns: Elements of Reusable Object-Oriented Software
- Clean Code
- The Pragmatic Programmer: From Journeyman to Master

D3 Visualization; Javascript

- Interactive Data Visualization for the Web, 2nd Edition
- JavaScript: The Good Parts

Big Data

- Hadoop: The Definitive Guide, 2nd Edition
- HBase: The Definitive Guide, 2nd Edition

We also recommend the following books and resources.

Python

- Python Bootcamp, for campus MS Analytics students. By Chris Simpkins

Data science, machine learning, data mining

- Data Science for Business by Foster Provost and Tom Fawcett
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, and Jerome Friedman

Visualization

- A nice D3 tutorial
- The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures by Dona Wong
- Information Dashboard Design: Displaying Data for At-a-Glance Monitoring by Stephen Few
- Show Me the Numbers: Designing Tables and Graphs to Enlighten by Stephen Few

SQL

- Video courses and lectures: (1) Coursera, (2) lynda.gatech.edu and search for SQL courses
- Interactive SQL tutorials
- Books: (1) SQL Cookbook (recipes to solve specific problems), Visual QuickStart Guide (succinct topic-by-topic), SQL Pocket Guide (covers syntax variations of MySQL, Oracle, etc.)
- Introductory tutorial

Probability

- FREE probability book, by Prof. Guy Lebanon. (From Amazon.)

Human Computation

- Human Computation book by Edith Law and Luis von Ahn

Office of Disability Services

The Office of Disability Services offers accommodations for students with disabilities. Please contact the office should you need help.

Prerequisites

Review Polo's "warnings" before taking this course.

Additional formal prerequisites for CSE 6242

None, but you should have taken courses similar to those listed in the next section, at Georgia Tech or at another school.

If you are an Analytics (OMS or campus) degree student, you should first take CSE 6040 and do very well in it; if necessary, please also first take CS 1301.

Additional formal prerequisites for CX 4242

(Undergraduate Semester level MATH 2605 Minimum Grade of D or
Undergraduate Semester level MATH 2401 Minimum Grade of D or
Undergraduate Semester level MATH 24X1 Minimum Grade of D) or
and
(Undergraduate Semester level MATH 3215 Minimum Grade of D or
Undergraduate Semester level MATH 3225 Minimum Grade of D or
Undergraduate Semester level ECE 3077 Minimum Grade of D or
Undergraduate Semester level ISYE 2027 Minimum Grade of D)
and
(Undergraduate Semester level CS 1371 Minimum Grade of C or
Undergraduate Semester level CS 1372 Minimum Grade of C or
Undergraduate Semester level CX 4010 Minimum Grade of C or
Undergraduate Semester level CX 4240 Minimum Grade of C)

Auditing & Pass/Fail

Due to the class size, we are not offering audit and pass/fail option.

Previous offerings

See https://poloclub.github.io/#cse6242 for all past course offerings.

Acknowledgment & Related Classes

We thank Intel's support in curriculum development for the memory mapping module (scaling up algorithms with virtual memory).

We thank Amazon Educate for providing free cloud credit for Amazon Web Services. We are excited to be am AWS partner university and part of AWS Educate's private beta.

We thank Microsoft Azure's special grant for providing free cloud credit.

We thank Tableau for Teaching program's data visualization software.

Many thanks to my colleagues for sharing their course materials:

- Prof. John Stasko - Information Visualization - Fall 2012
- Prof. Jeff Heer - Research Topics in Interactive Data Analysis - Spring 2011
- Prof. Christos Faloutsos - Multimedia Databases and Data Mining - Fall 2012

© Georgia Tech