There are **multiple CSE6242 sections**. This is the course homepage for **campus CSE6242A/CX4242A**.

CSE6242A/CX4242A Fall 26

Data and Visual Analytics

Georgia Tech, College of Computing

Prof. Duen Horng (Polo) Chau

Professor, School of Computational Science & Engineering

Associate Director, Master of Science in Analytics

Linkedin Google Scholar YouTube Bluesky

Course Description

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 Objectives

- 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

Course Learning Outcomes

- Apply computational and visual techniques to analyze real-world datasets (e.g., PageRank, Random Forest, SQLite, OpenRefine).
- Understand core design principles and the strengths of common visualization techniques (e.g., bar charts, line charts, tables).
- Design and build interactive data visualizations (e.g., using D3.js).
- Gain hands-on experience with cloud platforms (e.g., AWS, GCP) for large-scale data analysis.
- Collaborate in small teams to develop a substantial project showcasing innovative visual and data analytics.
- Present the project's goals, problem statement, approach, novelty, impact, risks, and expected benefits.

Acknowledgement

We thank the generous support of **Amazon Web Services**, **Google Cloud Platform**, 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 Edstem for all announcements and discussion. Everyone must join this class's Ed Discussion through Canvas. Double check that you are joining the correct Edstem! 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.
Students must always use **Ed Discussion** to communicate with course staff or for any class-related questions. Ed Discussion will be used for general posts, including private and public posts, threads, mega threads, Q&A, and announcements.
If course staff needs to communicate with specific students (i.e. members of a project team), the **Ed Chat** feature of Ed Discussion will be used. Students can benefit from this feature to communicate with other students. e.g., to discuss forming a project.

IMPORTANT: Everyone must ensure that the notification setting is on for both Ed Discussion and its Ed Chat feature to stay up to date with the class requirements and prevent losing points because of missing updates and announcements on Ed Discussion.

The fastest way to get help with homework assignments is to post your questions on Ed Discussion. That way, you can get help from our TAs and instructor can help, as well as your peers.

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

For special cases, such as personal emergencies, you can contact the staff using via Ed Discussion's private post feature (i.e., check the "Individual Students(s) / Instructors(s)" radio box).

Canvas will be used for some submission of assignments and projects, but not for announcements or discussion. Homework assignments will in general be submitted via Gradescope.

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 Ed Discussion threads, where the TA running the office hour will be responding to comments within the thread live. We will share more information in our office hour announcement post on Ed Discussion.

Information about Polo's weekly office hour will be shared on Ed Discussion too. All questions are welcome, except homework assignment related questions, which are best addressed via Ed Discussion and TA office hours.

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

Course Schedule

For **all homework and project dates** used in this course, their times are **23:59 Anywhere on Earth (11:59 pm AoE)**, unless stated otherwise. 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. Lecture slides below will be updated as semseter progresses.

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. In the past, some students have waited until the last week to begin, and could not finish. It is critical to plan ahead and prepare for the significant time required to complete the homework assignments.

Almost all homework assignments involve **very large amounts of programming** (which naturally means that a lot of debugging will likely be needed. This can be time consuming, and students should be prepared for the time commitment required). **You should be proficient in at least one high-level programming language** (e.g., Python, C++, Java), and should be **efficient with debugging principles and practices**. For students not meeting these expectations, we recommend first taking introductory computing course(s) before taking this course (for exmaple, CSE 6040 for (OMS) Analytics students; CS 1301, CS 1331, CS 1332, CS 1371, etc. for on-campus students).

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 know all of the tools/skills needed in every programing task, so you are expected to learn many of them on the fly.**

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

Minimum Computer Requirements

- 8GB RAM (16GB recommended)
- 512GB disk (SSD recommended). Some assignments use data files that are more than a few GBs, and some uses virtual machines that can easily take up more than tens of GBs. It is typical for some project teams to use large datasets that are more than a few or tens of GBs.
- Dual-core Core i5 (8th generation or better recommended), or Mac with M1 processor or better

Accessing Course Materials Outside of US

You may need to use Georgia Tech's VPN. We also recommend checking out some solutions that seem to be working well for OMS students in different countries.

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 (Canvas may show a weight greater than 15% because historically HW2 offers bonus points.)
- [15%] HW3: Spark, Docker, DataBricks, Cloud Services (AWS, Azure, GCP)
- [10%] HW4: PageRank, Random Forest, Scikit-Learn

We do not release solutions for homework. A solution is only one way of learning. This course provides multiple other ways, including immediate feedback from the autograder (with unlimited resubmissions), TA office hours, Ed Discussion, and we also welcome students to discuss with peers at the whiteboard level. All of these options offer a great variety of complementary ways to learn. If you have further questions about the assignment, please feel free to make a post on Ed or attend an office hour session. This course has been offered to 14,000+ students. While some students may think that seeing the solution is the "easy" way to identify where their code is not working, based on our experience, a lot of the time the error is so dependent on the student's implementation that seeing one "solution" does not help the student discover the specific, relevant reasons for

the error. We typically have 1100+ students each semester, across Atlanta and OMS campuses. When managing such a huge course, it is essential for the course staff to strike a balance between various factors when deciding on different aspects of the course. This includes designing and revising assignment questions, ensuring that we have robust autograders that can support the huge number of students and provide informative feedback, grading and offering feedback for team projects, and many more.

Can you release homework early? We understand that some students may prefer that homework assignments be released as soon as possible. Behind the scenes, our course staff work diligently to develop new questions, which means testing new datasets, new instructions, new auto graders, solution code, and more! Unfortunately, this means we cannot release assignments in advance.

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 course grade points. HW1: 10; HW2: 15; HW3: 15; HW4: 10.
2. There will be **one group project** worth 50 course grade points. The project components are:

   Proposal Deliverables

   1. Proposal Document (7 course grade points)
   2. Proposal Presentation (5)
   3. Proposal Peer Feedback (1)

   Midpoint Deliverables

   4. Progress Report (5)
   5. Midpoint Peer Feedback (2)

   Final Deliverables

   6. Final Poster Presentation (7)
   7. Final Report (21)
   8. Final Peer Feedback (2)
3. Bonus course grade points (up to 5.67 points)
   - HW2 has a maximum possible score of 100 points. Students have the option to complete any 90 points' worth of work to earn 15 course grade points. They may earn more course grade points by submitting additional work. For example, a student scoring 100 points will receive 16.67 course grade points (i.e., **1.67 bonus course grade points**, because 15 x 100/90 = 16.67).
   - There will be 4 bonus quizzes, each worth 1 bonus course grade point (answering, say, half of the questions correctly earns 0.5 points). Only the **3 highest-scoring quizzes** will count toward the course grade.
   - If more than 60% of the students in the class complete the CIOS survey, every student will receive **1 bonus course grade point**.
   - With bonus points, a student's total course grade points can reach a maximum of 105.67; this is why Canvas shows a total "weight" of 105.67%.
4. You must achieve at least 60 course grade points to pass the course.
5. Deliverables will be graded by TAs or autograders, except that the project poster presentation will be peer-graded.
6. When assigning course letter grades, we start with the standard grade thresholds (90, 80, etc.). We may lower (and never raise) the thresholds (i.e., to your benefit). For example, we may use 88 instead of 90.

Academic Integrity

- All learners are expected to know and abide by the **Georgia Tech Academic Honor Code** and the student **Code of Conduct**.
- 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. You must not put your code on public domain (e.g., public GitHub), because a (future) student could copy your code. That student obviously violates the honor code, and you may also be implicated.
4. 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

1. 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.
2. Every homework assignment deliverable and every project deliverable comes with a generous 48-hour "grace period".
    a. The grace period allows students to address **unexpected, minor** issues without facing penalties.
    b. Students can use the grace period without asking.
    c. Course staff support is **not** guaranteed during the grace period.
    d. Submissions during the grace period will display as "late" but will not incur a penalty.
    e. If a student decides to make a submission during the grace period, they are responsible for all issues associated with that submission (e.g., any Gradescope errors, including those triggered by student's syntax errors that crash Gradescope). Thus, a submisson that does not complete by the end of the grace period will receive **0** point.
3. For Canvas, a submission made during the grace period will be marked as "late", without point deduction. 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.**
4. For Gradescope, a submission made during the grace period will be marked as "late", without point deduction. Each submission and its score will be recorded and saved by Gradescope. By default, Gradescope uses **your last submission for grading**. To use a different submission, **you must "activate" it** prior to the end of the grace period (click "Submission History" button at bottom toolbar, then "Activate"). If Gradescope does not show up on Canvas's embedded panel, check your web browser's privacy or cookies settings.
5. 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!
6. No penalties for medical reasons or emergencies. And should they arise, you must contact the Dean of Students office. Any sensitive information, doctor's notes, medical documentation, explanation of emergencies, etc. should be submitted to the Dean's office. After their office receives the information, they will notify us on your behalf. Do not share any sensitive information with us. Accommodations are not retroactive.

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.
- You are responsible for staying engaged throughout the semester, recognizing that updates to the course schedule or coursework may be necessary (e.g., due to system issues or unforeseen events). For example, during peer grading periods, you are expected to remain attentive and plan accordingly to ensure all tasks are completed by the deadline, as task assignments may be updated during the grading window.

Attendance Policy

- Class attendance is mandatory for the project proposal presentation days (see course schedule for the exact dates).
- To provide flexibility for students who may not able to come to class due to illness, pre-recorded lecture videos from the OMS section of this course are available in the course schedule table above, and can be downloaded in the Media Gallery on Canvas.

Student-Faculty Expectations Agreement

At Georgia Tech, we believe that it is important to strive for an atmosphere of mutual respect, acknowledgement, and responsibility between faculty members and the student body. The Student-Faculty Expectations articulate some basic expectations that you can have of me and that I have of you. In the end, simple respect for knowledge, hard work, and cordial interactions will help build the environment we seek. Therefore, I encourage you to remain committed to the ideals of Georgia Tech while in this class.

Dataset Ideas (may need API, or scraping)

- "List of datasets for machine-learning research" Wikipedia page (which links to some of the datasets below)
- "Machine Learning Datasets" on Papers with Code. Thanks Justin!
- DiffusionDB: A large-scale text-to-image prompt gallery dataset based on Stable Diffusion
- FiveThirtyEight data
- MalNet graphs (300GB) and MalNet images (80GB): 1262024 unique function call graphs across 696 families and 47 types of malware.
- Google Dataset Search
- Google public datasets. Thanks Revant!
- Kaggle public datasets
- Data science competitions for Africa. Thanks Krupa!
- 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). TLC Trip Record Data. "Yellow and green taxi trip records include fields capturing pickup and drop-off dates/times, pickup and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts."
- Large datasets publicly available. Thanks Gopi!
- 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 pickups in NYU
- 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
- Movies data: IMDB
- List of lists of datasets for recommendations.
  Thanks Jon!
- Million Song Dataset.
  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
- 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: Open Academic Graph, DBLP
- Retrosheet: MLB statistics (Game/Play logs)
- Classification datasets
  Thanks Amish!
- Climate Data Online (CDO) provides free access to NCDC's archive of global historical weather and climate data in addition to station history information.
- 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

Required Course Materials

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

Resources

Accommodations for Students with Disabilities

If you are a student with learning needs that require special accommodation, contact the Office of Disability Services (404-894-2563) as soon as possible to make an appointment to discuss your special needs and to obtain an accommodations letter.

Support Services

Graduate Student Resources and academic and personal support services : Office of the Dean of Students, Counseling Center, Health Serivces, Women's Resource Center, LGBTQIA Resource Center, Veteran's Resource Center, Georgia Tech Police.

Recommended Reading

All Georgia Tech students have FREE access to **https://www.oreilly.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

- Debugging
- Clean Code
- Refactoring
- Design Patterns: Elements of Reusable Object-Oriented Software
- 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

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

How to manage multiple versions of Python packages?

To get started, we recommend the excellent article on Which Python package manager should you use?

If you've decided to go with pyenv, I recommend Managing Multiple Python Versions With pyenv.

If you use Mac, we recommend to also check out The right and wrong way to set Python 3 as default on a Mac.

Students in my reserach group said that Poetry seems to be fast replacing conda envs, and may even replace setuptools for pypi packages in the future.

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)

Course offerings and Registration

Auditing & Pass/Fail

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

Major restrictions and when they will be lifted; phase 1 and phase 2 registration info

CX 4242: https://www.cc.gatech.edu/undergraduate-registration
CSE 6242: https://www.cc.gatech.edu/student-registration-dates-information-graduate

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 an AWS partner university and part of AWS Educate's private beta.

We thank Google Cloud Platformfor providing free cloud credit for GCP. We are excited to be a GCP partner education.

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
