Introduction to Big Data and Data Science Syllabus

CSCE 5300: Introduction to Big Data and Data Science (grad)
CSCE 4930: Special Topics in Computer Science and Engineering (undergrad)
Spring 2020 at University of North Texas

Course Calendar: go here for details on content and assignments

Instructor Contact

Name: Ting Xiao
Office Location: Discovery Park F245
Office Hours: Tu/Th 3:30pm - 4:30pm
Email: ting.xiao@unt.edu

TA Name: Chengyuan Zhuang
TA Email: chengyuanzhuang@my.unt.edu
TA Office Hours: Tu/Th 2:30pm - 4:00pm
TA Office Location: CSE Help Lab, Room F232

Communication Expectations: You will be expected to regularly check university email and attend class regularly. When you miss a class, you are expected to check the course calendar shortly after class to be aware of assignments, quizzes, and other materials. Questions not answered in class are best asked before or after class. For quick questions, email is preferred and you can expect a response within 24 hours during the work week (M-F). For personal questions, office hours are preferred. It is also possible to arrange meetings in person as well as remotely through zoom, by phone, or skype by contacting the instructor over email.

Course Description

- Current course description: Big Data Flyer [PDF]
- Course description in the UNT system: Introduction to Big Data and Data Science includes an overview of the field, technical challenges, computational approaches, practical applications, structured and unstructured data processing, empirical methods in computer science, data analytics and learning, data visualization, privacy and ethics. Emphasis will be on Big Data and its effect on other topics within Data Science, its technical characteristics, and state-of-the-art Big Data analytics architectures and tools.

Course Structure

Time: Tu/Th 11:30am - 12:50pm
Location: Discovery Park B185
Credit hours: 3

Tentative topics

- Data visualization
- Basic data structures and data frames
- Databases (SQL)
- Machine learning concepts
- Deep learning concepts
- Cloud computing
- Big Data / parallel processing architectures (e.g. Spark/Hadoop)
- Virtualization and containers using Kubernetes

Course Prerequisites

- For CSCE 4930 specifically
    - Junior or senior standing in computer science, computer engineering or information technology, and consent of instructor.
- Experience with Python is beneficial as it is used extensively in the course, but significant prior programming experience with any language will be sufficient.

Course Objectives

By the end of the course, students will be able to:

1. Articulate key advances in contemporary data science and describe the skill sets needed to be successful in a data science career.
2. Use tools of data science and Big Data to obtain, assess, and prepare data for analysis.
3. Manage collections of data, create automated processes for analysis, use collaborative tools, and rapidly report quantitative findings.
4. Apply basic principles of predictive modeling for accurate, explainable predictions including proper hyperparameter tuning, model selection, and validation.
5. Administer cloud parallel computing for Big Data, and apply cloud-based tools to efficiently handle large data sets to solve problems.
6. Participate Kaggle competitions and learn tips to achieve high rankings.

Materials

All materials (readings, videos, tutorials, quizzes, and assignments) will be accessible online and posted on the course calendar on the respective class day at the latest. Readings will all be fairly dense, so please search for additional resources (e.g. wikipedia, coursera lectures) as needed. All attempts will be made to provide sufficient resources for everyone. There will be no required textbook as all reading materials will be linked from the course calendar and will be free and online accessible.

Course communication: Email will be the primary means of discussion outside of class.

- You will be expected to respond to emails from the instructor, TA, or students in the course in a timely way.
- When assignment or project groups form, it is important to share email and any other contact information at the earliest time to organize outside of class.
- You are free to organize alternative discussion boards with fellow students (e.g. via Slack, Piazza, Trello, etc…) but such boards are not official communication of course material. Also, be aware of the academic integrity policy in the course on such boards.
- Any emails to the class are for timely, *supplementary* communication; the course calendar will be the definitive source of requirements and course expectations and will be kept up to date to match any emails.

Technical Requirements and Skills

Minimum Technology Requirements

- Computers are **required for exams**
    - **In-class exams:** If computers are not available in the classroom, you will need to bring a laptop on exam days. *It is recommended to bring your own laptop.* Exams will be done individually on your computer in class. With a laptop, you will be expected to connect to the UNT wireless network.
    - Assignments and exams will use the Canvas system
        - Canvas Technical Requirements (https://clear.unt.edu/supported-technologies/canvas/requirements)
- Computers are **needed during class, outside of exam times, unless otherwise indicated on the course calendar**
    - You are expected to write programs during class.
    - There will be occasional in-class activities along with lecture that computers are not required, but may help in understanding and applying the material.

Course Requirements

Readings, Competitions, and Exams

Readings/Tutorials: These will be available on the course calendar. Expectations of what is learned will be discussed in class and, as will be clear in the first few weeks, assignments will test your knowledge on the most important aspects of the readings only.

Competitions: In the course there will be Kaggle.com-style analytics competitions. Extra points will be distributed based on rank after the competition is finished, however the amount of points (or the variation in points across the class) will be minimal. These are primarily used as a barometer for your ability to apply these concepts quickly and efficiently.

Exams: There will be 3 midterm exams and a final exam. Exams will resemble assignments, but will be on a new data set with a slightly different request for tasks. Exams must be taken in the classroom. Time will be limited, and all work will be individual. You are strongly encouraged to attempt to solve the tasks iteratively and incrementally - write code that works first, but works poorly, and improve from there, rather than write perfect code top to bottom. Exam I, II, III cover the parts in the previous weeks. The final exam is cumulative, but the emphasis will be on the newer material. Exam days are already posted and are considered fixed. Prior arrangements can potentially be made without loss of points, but have to be discussed.
Missed exams: Exams cannot be missed without prior arrangements or later documented proof of extenuating circumstances.

Assignments, Presentations, and Projects

Assignments are designed to engage you in your learning, so you can begin to apply these principles in practice and tailor them to your needs.

Assignments: Unless otherwise specified, assignments are due at midnight exactly one week after they are assigned (allowing for questions in class on the day the assignment is due). Results should be presentable, with appropriate comments for someone to follow what you have done. Assignments can be done individually or in a group of no more than 3 people, but must be submitted to Canvas individually unless otherwise specified. Note, assignments will be given almost weekly, and the bulk of the assignments should be attempted in class with help from other students, the TA, and the instructor. To encourage in-class efforts and to provide precise feedback, **completed assignments must be shown to the instructor in class before submitting to Canvas to receive full points**. It is important to keep up given this pace of new assignments nearly weekly. Expected effort will generally match grade points and time given to assignments. Notably, low-point assignments may by small efforts more for learning and self-evaluation than assessment.

Short presentations: There are far more tools and techniques than we can possibly cover, many of which may be particularly relevant to your interests. Also, the ability to distill complex topics into a form useful for the audience is a critical skill to develop. **Students will be expected to present one concept, tool, or technique which goes beyond what is covered in the course**. Each presentation is to be 10-15 minutes with slides. These presentations will be spread over the first third of the course. Students may be quizzed on the material presented in these presentations, however, it will only be basic familiarity of the topic that can be learned from the slides or by paying attention to the presentation in class.

Research paper presentations: The research paper presentation is an opportunity for students to share their exposure to state of the art research efforts. Students will select among relevant recent papers from a provided list of conferences and journals. Each presentation will be 10-15 minutes with slides and will be spread over the second third of the course. Students may be quizzed on the material presented in these presentations, however, it will only be basic familiarity of the topic that can be learned from the slides or by paying attention to the presentation in class.

Projects: After a few weeks into the course you will select among a small number of collaborative projects. Project proposals, progress reports, and final reviews will be part of the process. You are required to work in groups, as this is part of a full and complete education. All people in the group are expected to contribute. This is your opportunity to demonstrate what you have learned in a way that reaches beyond the selection of tools, data sets, and approaches demonstrated in the course. Commonly students find a unique, complex data set and associated learning problem and apply the techniques presented in the class. The goal here is to create a coherent, completed project for presentation at the end of class. Essentially ask yourself what you would want to show an employer (or brag about to others) demonstrating what you have learned in the course.

Grading

Grades are determined by a simple points system, with a total of *at least* 100 pts given though more than 100 points are likely. The expected distribution of points is given below, with the exact scale determined by point values given for each assignment, presentation, competition, project, or exam - this is subject to minor modification based on actual points given. Note, due to the nature of the course, exams are a significant means of establishing your final grade, so please complete the assignments in a timely way and study appropriately prior to each exam.

- Assignments: 15 pts (1-2 pts each)
- Presentations: 10 pts
    - Short presentations: 5 pts
    - Research paper presentation: 5 pts

- Projects: 20 pts
    - Brainstorming: 1 pt
    - Proposal: 2 pts
    - Update: 2 pt
    - Project presentation: 5 pts
    - Report: 10 pts
- Competitions: 10 pts (~5 pts each)
- Exams: 45 pts
    - Exams I,II,III: 10 pts each
    - Final Exam: 15 points

Grading Scale: A=90, B=80-89.9, C=70-79.9, D=60-69.9, F=0-59.9 pts. No exceptions. If class grades are low (e.g. I expect the vast majority of students will end with A's and B's), extra quizzes or assignments will be given to add points to the class. (Note, **these grades are based on points and not percentages, so if 120 points are given, you only need 90/120 for an A!**)

CSCE 4000 vs 5000 level grading: In line with the added expectations for the graduate version of the course, **students enrolled in the undergraduate level course will not have to give a research paper presentation - they will be given full credit for that assignment.**

Course Evaluation

Student Perceptions of Teaching (SPOT) is the student evaluation system for UNT and allows students the ability to confidentially provide constructive feedback to their instructor and department to improve the quality of student experiences in the course. Spot evaluations will be available toward the end of the semester.

Course Policies

Examination Policy

Exams will be on the computer using the Canvas quiz system. You need to bring a laptop on the appropriate exam days. **Exams are open book, open note, and open internet unless otherwise specified in advance, however, no communication with others other than the instructor and the TA is allowed in any form (e.g., email, chatting, etc.).** Exams must be taken in the classroom unless special accommodations have been made through the Office of Disability Accommodation (ODA). Another other accommodations must be given by prior arrangement with the instructor, otherwise documentation proving an extenuating circumstance will have to be provided after the missed exam. Time will be limited, and all work will be individual. You are strongly encouraged to attempt to solve the tasks iteratively and incrementally - write code that works first, but works poorly, and improve from there, rather than write perfect code top to bottom. Exams will focus on the most recent material but are expected to be cumulative in scope.

Technical errors during exams

If during an online quiz or exam there is a technical error which affects your ability to complete the assignment, you are immediately to let the quiz or exam proctor know and the instructor will discuss ways to allow you to resume the test without giving an unfair advantage. In the event of any unexpected server outage or any unusual technical difficulty which prevents students from completing a time sensitive assessment activity, the instructor will extend the time windows and provide an appropriate accommodation based on the situation. Students should immediately report any problems to the instructor.

Late Policy

When assignments and project work are turned in after the due date, this places an undue burden on the instructor and TA, especially when this policy is abused. Acceptance of late assignments is at the discretion of the TA or instructor, but is not guaranteed, and a reduction of points will occur in a way that is consistent for the rest of the class.

Attendance Policy

You are expected to attend lectures and to complete all readings, however, this course does not use participation points and there is no penalty for missing days without exams, or group project efforts. There is no need to let the instructor know you have missed a class, however, you are responsible for keeping up with the material covered in the class if you are not present. If a class is missed, you are expected to proactively reach out to classmates, the TA, or the instructor if there are any questions.

Individual attendance is not required except on presentation days and exam days. There is no direct participation grading, but in the past there has been a strong correlation between engagement and accomplishment in courses - especially for those that are struggling with the material. Feel free to prioritize your time, but prioritize wisely.

Syllabus Change Policy

Any substantial changes to the syllabus after the first week will be highlighted in red on the online platform. Approximate point values are expected to vary but will be fixed when the assignment or exam is given.

UNT Policies

Academic Integrity Policy

Academic Integrity Standards and Consequences. According to UNT Policy 06.003, Student Academic Integrity, academic dishonesty occurs when students engage in behaviors including, but not limited to cheating, fabrication, facilitating academic dishonesty, forgery, plagiarism, and sabotage. A finding of academic dishonesty may result in a range of academic penalties or sanctions ranging from admonition to expulsion from the University.

ADA Policy

UNT makes reasonable academic accommodations for students with disabilities. Students seeking accommodations must first register with the Office of Disability Accommodation (ODA) to verify their eligibility. If a disability is verified, the ODA will provide a student with an accommodation letter to be delivered to faculty to begin a private discussion regarding one's specific course needs. Students may request accommodations at any time, however, ODA notices of accommodation should be provided as early as possible in the semester to avoid any delay in implementation. Note that students must obtain a new letter of accommodation for every semester and must meet with each faculty member prior to implementation in each class. For additional information see the ODA website (https://disability.unt.edu/).

Emergency Notification & Procedures

UNT uses a system called Eagle Alert to quickly notify students with critical information in the event of an emergency (i.e., severe weather, campus closing, and health and public safety emergencies like chemical spills, fires, or violence). In the event of a university closure, please refer to Blackboard for contingency plans for covering course materials.

Retention of Student Records

Student records pertaining to this course are maintained in a secure location by the instructor of record. All records such as exams, answer sheets (with keys), and written papers submitted during the duration of the course are kept for at least one calendar year after course completion. Course work completed via the Blackboard online system, including grading information and comments, is also stored in a safe electronic environment for one year. Students have the right to view their individual record; however, information about student's records will not be divulged to other individuals without proper written consent. Students are encouraged to review the Public Information Policy and the Family Educational Rights and Privacy Act (FERPA) laws and the University's policy. See UNT Policy 10.10, Records Management and Retention for additional information.

Acceptable Student Behavior

Student behavior that interferes with an instructor's ability to conduct a class or other students' opportunity to learn is unacceptable and disruptive and will not be tolerated in any instructional forum at UNT. Students engaging in unacceptable behavior will be directed to leave the classroom and the instructor may refer the student to the Dean of Students to consider whether the student's conduct violated the Code of Student Conduct. The University's expectations for student conduct apply to all instructional forums, including University and electronic classroom, labs, discussion groups, field trips, etc. Visit UNT's Code of Student Conduct (https://deanofstudents.unt.edu/conduct) to learn more.

Access to Information - Eagle Connect

Students' access point for business and academic services at UNT is located at: my.unt.edu. All official communication from the University will be delivered to a student's Eagle Connect account. For more information, please visit the website that explains Eagle Connect and how to forward e-mail Eagle Connect (https://it.unt.edu/eagleconnect).

Student Evaluation Administration Dates

Student feedback is important and an essential part of participation in this course. The student evaluation of instruction is a requirement for all organized classes at UNT. The survey will be made available during weeks 13, 14 and 15 of the long semesters to provide students with an opportunity to evaluate how this course is taught. Students will receive an email from "UNT SPOT Course Evaluations via IASystem Notification" (no-reply@iasystem.org) with the survey link. Students should look for the email in their UNT email inbox. Simply click on the link and complete the survey. Once students complete the survey they will receive a confirmation email that the survey has been submitted. For additional information, please visit the SPOT website (http://spot.unt.edu/) or email spot@unt.edu.

Getting Help

Technical Assistance

UIT Help Desk (http://www.unt.edu/helpdesk/index.htm)
Email: helpdesk@unt.edu
Phone: 940-565-2324
In Person: Sage Hall, Room 130
Walk-In Availability: 8am-9pm
Telephone Availability:

- Sunday: noon-midnight
- Monday-Thursday: 8am-midnight
- Friday: 8am-8pm
- Saturday: 9am-5pm

Laptop Checkout: 8am-7pm

Student Support Services

- Registrar (https://registrar.unt.edu/registration)
- Financial Aid (https://financialaid.unt.edu/)
- Student Legal Services (https://studentaffairs.unt.edu/student-legal-services)
- Career Center (https://studentaffairs.unt.edu/career-center)
- Multicultural Center (https://edo.unt.edu/multicultural-center)
- Counseling and Testing Services (https://studentaffairs.unt.edu/counseling-and-testing-services)
- Student Affairs Care Team (https://studentaffairs.unt.edu/care)
- Student Health and Wellness Center (https://studentaffairs.unt.edu/student-health-and-wellness-center)
- Pride Alliance (https://edo.unt.edu/pridealliance)

Academic Support Services

- Academic Resource Center (https://clear.unt.edu/canvas/student-resources)
- Academic Success Center (https://success.unt.edu/asc)
- UNT Libraries (https://library.unt.edu/)
- Writing Lab (http://writingcenter.unt.edu/)
- MathLab (https://math.unt.edu/mathlab)