QTM 350: Data Science Computing
Spring 2024

Contact Information

Instructor	**Davi** Cordeiro **Moreira**, PhD
E-mail:	davi.moreira@emory.edu
Class Schedule	**T/Th 4-5:15 p.m.** on **PAIS 290**
Office Hours	*Instructor: Tue 9-10:15 a.m. ([Zoom](#)), or by appointment*
	TA: Tue 9-10:15 a.m. (Link TBD)

Course Description

This course equips students with computing skills and knowledge for data science applications. Students will gain knowledge foundations and hands-on experience with technologies such as Version Control, Project Collaboration, Data Structures and Algorithms, Database Designs, Database Management, and Cloud Computing. Prospective data scientists, statisticians, and other quantitative professionals will learn how to efficiently utilize database structures and foundational cloud services for data science by completing creative projects.

Learning Outcomes
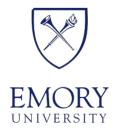
By the end of this course, students will be able to:
1. Demonstrate proficiency in data science project collaboration and version control.
2. Utilize advanced data storage, manipulation, and querying.
3. Perform tasks in a cloud computing environment.
4. High-level understanding of data structures and algorithms, and cloud computing concepts.
5. Critically navigate in the emergent trends in data science computing.

Course Materials

- No software download is required, any web browser is sufficient.

Course References

- [Computing Skills for Biologists:](#) A toolbox with basic computational skills necessary for the course.
- [Elements of Data Science](#): a digital textbook by Allen Downey written in the form of Jupyter notebooks. It provides an introduction to data science in Python for students with limited programming experience.
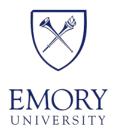
- [Think Python](#): An introduction to programming using Python.
- [Applied Computing](#): Applied Computing is an [online textbook](#). It provides an introduction to spreadsheets and SQL. To view the book, students need to [register](#) using the course name.
- [SQL & NoSQL Databases: Models, Languages, Consistency Options, and Architectures for Big Data Management:](#) Explores relational (SQL) and non-relational (NoSQL) databases. Covers database management, modeling, languages, consistency, architecture, and more.
- [Pro Git Book:](#) A comprehensive resource for learning Git, covering everything from the basics to advanced topics by Scott Chacon and Ben Straub.
- [Overview of Cloud Computing](#): An introductory textbook on cloud computing by Michael Wufka and Massimo Canonico.

Objectives

- **Conceptual Understanding:** To provide students with a foundational grasp of data structures, algorithms, and data modeling techniques pertinent to SQL and NoSQL databases and data analytics.
- **Technical Proficiency:** To equip students with practical skills in version control, Python programming, and database management using SQL and MySQL, enabling them to execute data manipulation and analysis tasks proficiently.
- **Cloud Computing Literacy:** To introduce students to the fundamentals of cloud computing, focusing on models, components, and security considerations, and offer hands-on experience.
- **Integrated Learning:** To offer a holistic educational experience that combines theoretical learning with practical exercises, ensuring students can apply their knowledge to real-world data science computing projects.
- **Critical Analysis:** To cultivate the ability to compare and evaluate different data structure algorithms, cloud platforms and technologies, fostering an awareness of emerging trends in the data science computing landscape.

Additional References

- [Google Cloud Computing Foundations:](#) Google Cloud experts develop and maintain the Computing Foundations reference material specifically for university courses such as this to ensure it keeps pace with cloud innovation and to prepare students seeking to launch or pivot to [careers](#) in a cloud-first world.
- [Big Data: Principles and best practices of scalable real-time data systems:](#) It describes a scalable, easy to understand approach to big data systems that can be built and run by a small team.
- [Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development](#): This book explains how Git works and how to use it effectively. By Jon Loeliger and Matthew McCullough

- [GitHub Learning Lab](#): Offers a variety of exercises to get hands-on experience using Git and GitHub.
- [AWS Academy Cloud Foundations (login required):](#) AWS experts develop and maintain the Cloud Foundations reference material to ensure it keeps pace with cloud innovation and prepares students for real-world, industry challenges. This digital textbook with accompanying labs comes in the form of a Canvas page, separate from the class Canvas page.

Assessment and Grading

Details in the Canvas page. The breakdown is as follows:
- Quizzes: 20%
- Homework Assignments: 30%
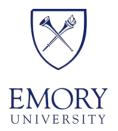- Project proposal: 5%
- Final Project: 35%
- Class Participation and Peer Review: 10%

Each student's final grade will be based on the following after rounding up to the nearest point:

	Lower	Upper
A	91	100
A-	86	90
B+	81	85
B	76	80
B-	71	75
C	66	70
D	60	65
F	0	59

Communication

- Check our **[Canvas course page](#)** regularly to keep yourself informed with up-to-date information about the course. Also, be sure **to check the course syllabus before asking any questions about the course schedule/policies.**
- If you cannot attend the office hours due to conflicts with other course schedule or attending the university-sanctioned events (proof required), **email the instructor at least two days in advance** to set up an appointment. Note that **each appointment will be 15-minutes long**, and it may be done in a small group or individually. **No appointments will be allowed nearing the exam dates. Instead, the instructor will hold extra office hours to help you prepare for**

the exam.
- **When attending virtual office hours, make sure you are in a private setting with a little to no background noise. The use of headphones is strongly encouraged.** This is especially true when you are discussing private matters with the instructor.
- Do NOT use email for asking content-related questions, and do NOT use Canvas messages.
- **Do NOT email me your private stories**. **Keep it brief**, and you will receive a response from me within 48 hours, except for the weekends. Similarly, if you receive an email from me, you are also expected to respond within 48 hours. Set up an individual appointment to discuss such things.
- **Finally,** if you are experiencing situations that negatively impact your overall student life, you should immediately contact the Office of Undergraduate Education.
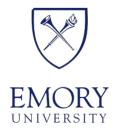
Regarding absences

- If you miss a lecture for any reasons, understand that you are still responsible for the missed course materials. First, review the missed materials, then you may attend the instructor office hours to ask specific questions.
- Attendance is **not** monitored in lecture except on the exam dates.
- Emory College of Arts and Sciences policy states, "**A student who fails to take any required midterm or final examination at the scheduled time may not make up the examination without written permission from a dean in the Office for Undergraduate Education**. Permission will be granted only for illness or other compelling reasons, such as participation in scheduled events off-campus as an official representative of the University.

Access and Disability Resources

Students with medical/health conditions that might impact academic success should visit the Department of Accessibility Services (DAS) to determine eligibility for appropriate accommodations. Students who receive accommodations must contact the instructor with an Accommodation Letter from the DAS at the beginning of the semester, or as soon as the accommodation is granted.

If you have DAS accommodations, you must inform the instructor by no later than the schedule change deadline (TBD) after confirming that your accommodation letter is available in the DAS web portal. The instructor will respond to your email confirming which accommodations you will receive for this class. If you wish to do so, you may request an individual meeting to further discuss the specific accommodations.

AI policy

I encourage you to use AI tools you believe will enhance your individual or group performance. In fact, it is possible that some course assignments will require it. Learning to use AI is a valuable and emerging skill, and I am available to provide support and assistance with these tools during office hours or by appointment.

Be aware of the following guidelines:

- Providing low-effort prompts will result in low-quality outputs. You must refine your prompts to achieve desirable outcomes. Use the course knowledge for that!
- Do not blindly trust the information provided by the output. If the output contains a number, index, analysis, conclusion, or fact, assume it is incorrect and check its veracity. Any errors or omissions resulting from using the AI tool will be your responsibility. Remember, the AI tool works better for topics that you already understand.
- While AI is a tool, you must acknowledge its use. Always cite! Include a paragraph or note at the end of any document to mention that you used AI on its development.

Academic Integrity

Upon every individual who is a part of Emory University falls the responsibility for maintaining in the life of Emory a standard of unimpeachable honor in all academic work. The [Honor Code of](#) [Emory College](#) is based on the fundamental assumption that every loyal person of the University not only will conduct his or her own life according to the dictates of the highest honor, but will also refuse to tolerate in others action which would sully the good name of the institution.Academic misconduct is an offense generally defined as any action or inaction which is offensive to the integrity and honesty of the members of the academic community. **The typical sanction for a violation of the Emory Honor Code is an F in the course. Any <u>suspected</u> case of academic misconduct will be referred to the Emory Honor Council.**

Subject to Change Policy

While I will try to adhere to the course schedule as much as possible, I also want to adapt to your learning pace and style. The syllabus and course plan may change in the semester.
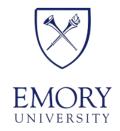
Topics

Getting started: Syllabus and Set up.
- Course Introduction
- Syllabus Overview
- Software Installation
- Setting up Development Environment

Version Control
- Introduction to Git
- Git Basics: Clone, Commit, Push, Pull

- Branching and Merging
- Git Workflows

Python essentials I
- Python Syntax and Variables
- Data Types
- Control Structures: Loops and Conditionals
- Functions and Modules
- Python Libraries: pandas, numpy, and matplotlib

Python essentials II
- File I/O
- Exception Handling
- List Comprehensions
- Lambda Functions and Map/Reduce

API Interactions and Web Scraping
- Introduction to APIs and Interactions
- Ethical Considerations in Web Scraping
- Introduction to Web Scraping: Understanding HTML and CSS
- Using BeautifulSoup and Requests
- Dynamic Scraping with Selenium

Overview of Data Structures and Algorithms
- Arrays and Linked Lists
- Stacks and Queues
- Hash Tables
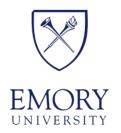- Trees: Binary Search Trees and their Relevance to Decision Algorithms

SQL and NoSQL
- Introduction to NoSQL
- Introduction to NoSQL: Advantages and Use-cases
- Overview of Document-based (e.g., MongoDB), Key-Value (e.g., Redis), and Column-family Stores (e.g., Cassandra)

SQL Operations and Operators
- SQL Basics: SELECT, INSERT, UPDATE, DELETE
- WHERE Clause and Operators
- JOIN Operations
- Aggregate Functions

Database Design, Structures and Management

- Database Normalization
- ER Diagrams
- ACID Properties
- Database Management Systems (DBMS)

Create, Populate and Manipulate databases
- Creating Tables and Schemas
- Data Import and Export
- CRUD Operations
- Transactions and Concurrency

Advanced Data Modeling
- Data Warehousing
- OLAP and OLTP
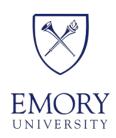- Star and Snowflake Schemas

Cloud Computing
- Introduction to Cloud Computing
- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Cloud Security and Compliance
- Cloud Providers: AWS, Azure, Google Cloud

Cloud Computing: Hands-on
- Setting up a Cloud Environment
- Deploying Applications
- Cloud Storage and Databases
- Monitoring and Scaling

Course Schedule

Week	Date	Topic
1	18-Jan	Getting started: Syllabus and Set up
2	23-Jan	Version Control
2	25-Jan	Version Control
3	30-Jan	Python essentials I
3	1-Feb	Python essentials I
4	6-Feb	Python essentials II
4	8-Feb	Python essentials II
5	13-Feb	API Interactions and Web Scraping

5	15-Feb	API Interactions and Web Scraping
6	20-Feb	Overview of Data Structure and Algorithms
6	22-Feb	Overview of Data Structure and Algorithms
7	27-Feb	Overview of Data Structure and Algorithms
7	1-Mar	SQL and NoSQL
8	6-Mar	SQL and NoSQL
8	8-Mar	SQL and NoSQL
9	**13-Mar**	**Spring Break (no classes, no office hours)**
9	**15-Mar**	**Spring Break (no classes, no office hours)**
10	20-Mar	SQL Operations and Operators
10	22-Mar	SQL Operations and Operators
11	27-Mar	SQL Operations and Operators
11	29-Mar	Database Design, Structures and Management
12	3-Apr	Database Design, Structures and Management
12	5-Apr	Create, Populate and Manipulate databases
13	10-Apr	Create, Populate and Manipulate databases
13	12-Apr	Advanced Data Modeling
14	17-Apr	Cloud Computing
14	19-Apr	Cloud Computing
15	24-Apr	Cloud Computing: Hands-on
15	26-Apr	Cloud Computing: Hands-on