

# QTM 350: Data Science Computing

## Spring 2024

### Contact Information

<b>Instructor</b>	<b>Davi Cordeiro Moreira</b> , PhD	
<b>Class Schedule</b>	TBD (TBD)	TBDAM on TBD and TBD
<b>Office Hours</b>	<b>TBD</b>	Schedule posted in <a href="#">Lecture Canvas Page</a> <i>Appointment-only</i> office hours (in person at PAIS 468 or via Zoom): <b>must set up appointments via email at least TWO days in advance</b>
<b>E-Mail</b>	<a href="mailto:davi.moreira@emory.edu">davi.moreira@emory.edu</a>	

**\*\* Under Development \*\***

### Course Description

This course aims to equip students with advanced computing skills for data science applications, focusing on SQL, cloud computing, and programming for data analysis and manipulation. The course is designed for students who have already taken an introductory data science or programming course.

### Learning Outcomes

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

1. Utilize SQL for advanced data manipulation and querying.
2. Navigate and perform tasks in a cloud computing environment.
3. Apply best practices in data science computing.
4. Implement high-level data analysis solutions in R/Python
5. Demonstrate proficiency in version control systems like Git.
6. Understand the basics of distributed computing.
7. Implement APIs for data collection and manipulation.

### Honor code

All student work is governed by the [Emory Honor Code](#). Honor code violations (e.g., copies from any source including your colleagues and internet sites without reference and delineation of your contributions) will be referred to the [Emory Honor Council](#).

### Accommodations

All information shared with me regarding your identity, disability and accommodations is confidential. Please contact me as soon as possible so that I can provide the approved accommodations outlined on the accommodation notification letter. I am happy to consult with ADSR in providing additional accommodations.

### References

- [Computing Skills for Biologists](#): TBD

- [Elements of Data Science](#): TBD
- [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 QTM350.
- [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](https://careers.google.com/students/) in a cloud-first world.
- Big Data: Principles and best practices of scalable realtime data systems: TBD
- Data Science from Scratch: TBD
- SQL Performance Explained: TBD
- Certainly, Dr. Moreira. The following references are highly regarded in their respective areas and can serve as excellent supplementary materials to help students meet the learning outcomes you have specified:
- [Pro Git Book](#): A comprehensive resource for learning Git, covering everything from the basics to advanced topics by Scott Chacon and Ben Straub.
- [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.
- [Data-Intensive Text Processing with MapReduce](#): A useful resource for understanding MapReduce, a key technology for distributed computing. By Jimmy Lin and Chris Dyer
- Hadoop: The Definitive Guide: Covers the Hadoop ecosystem, a popular distributed computing framework, in great detail. By Tom White\*\*
- Web Scraping with Python: A Comprehensive Guide" by Ryan Mitchell
- APIs: A Strategy Guide" by Daniel Jacobson, Greg Brail, and Dan Woods
- RESTful API Design: Best Practices in API Design with REST" by Matthias Biehl
- [Postman Learning Center](#): Offers tutorials and articles on API testing and development.

## Grading

- \* Attendance: 5%
- \* 4 Homeworks: 40%
- \* 6 Quizzes: 30%
- \* Midterm: 10%
- \* Final exam: 15%

## Attendance

In each class, one of the groups gives a five-minute lightning talk presentation on one of the elements that the class is working on mastery of in the homeworks. Every student must submit their presentation feedback to the instructor in order to satisfy the attendance requirement.

## Homeworks

There are four homeworks, one on each of the core topics: spreadsheets; SQL; data APIs; ML APIs. Students are split into groups at the start of the semester and all homeworks are completed as groups.

### **Quizzes**

There are six take-home style quizzes that students complete individually on Canvas.

### **Midterm and Final Exam**

The midterm and final exam are cumulative. Questions from the assigned quizzes are pooled and a selection of these, or similar, make up the midterm and final exam. They are administered on Canvas.

### **Course Schedule - TBD**