QTM 350: Data Science Computing
Spring 2024

Contact Information[1]

Instructor **Davi** Cordeiro **Moreira**, PhD
Class Schedule **TBD (TBD)** **TBD** on **TBD** and **TBD**
Office Hours ***TBD*** *Schedule posted in Canvas Page*
Appointment-only office hours (in person at PAIS 468 or via Zoom): **must set up appointments via email at least TWO days in advance**
E-Mail davi.moreira@emory.edu

****UNDER DEVELOPMENT****

Course Description

In 2022, over 60 percent of corporate data is stored in the cloud. This course equips students with advanced computing skills for data science applications focusing on cloud computing. It aims to provide to the new data scientists the job market demand of cloud computing skills. Prospective data scientists, statisticians, and other quantitative professionals will learn how to utilize foundational cloud services for data science and how to participate in this new computing paradigm by completing creative projects using Jupyter notebooks together with the latest cloud tools for data science.

Learning Outcomes

By the end of this course, students will be able to:
1. Utilize advanced data manipulation and querying.
2. Navigate and perform tasks in a cloud computing environment.
3. Implement high-level data analysis solutions in Python.
4. Demonstrate proficiency in version control systems like Git.
5. Understand the basics of distributed computing for data analysis.

Course Materials

- All attendees will be provided access to the cloud infrastructure used in the course.
- No software download is required, any web browser is sufficient.

Course References

[1] Thanks for Professor Jacobson for sharing the QTM350 syllabus, course material, and guidance.

- 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.
- 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.
- Computing Skills for Biologists: A toolbox with basic computational skills necessary for the course.
- Overview of Cloud Computing: An introductory textbook on cloud computing by Michael Wufka and Massimo Canonico.

Objectives

- Learn basic syntax and concepts necessary for utilizing Python for data science.
    - Learn important programming and computing concepts useful when working with computers: binary, ascii, unicode,
    - Learn Linux fundamentals necessary when working on cloud virtual machines.
    - Utilize fundamental shell commands for tasks such as cloning repos, installing and updating packages, and downloading data.
    - Practice markdown formatting and Python commands in Jupyter notebooks.
- Learn how to utilize a managed Jupyter notebook service running on cloud virtual machines (we will use AWS SageMaker) and take advantage of features of the cloud for notebook computing:
    - Use the interface to run different notebook kernels and virtual machines in SageMaker.
    - Explore AWS sample notebooks and new use cases of data science on the cloud.
    - Use the GitHub integration and Git via the graphical JupyterLab interface.
    - Get experience working with large datasets (GB and TB scale) in the cloud (we will use Open Datasets on AWS):
    - Use the command line interface (we will use the AWS CLI) to explore collections of files and buckets within cloud storage (Amazon S3). Copy, sync and move data to and from SageMaker for analysis.
    - Practice by implementing and building upon steps described in tutorial notebooks from the Registry of Open Data.
    - Write your own tutorial notebook explaining a use case you are interested in.
- Explore and test AWS Machine learning APIs:
    - Explore using Amazon Rekognition to compare with the state of the art in computer vision.
    - Explore using Amazon Comprehend to obtain valuable insights from text within documents.
    - Test and analyze the behavior of these machine learning services on your own data using AWS SageMaker.

- Write your own analysis notebook. Explain your unique insights into the performance of the ML services and demonstrate by testing on your own data.

Additional References

- 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.
- 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 realtime data systems: It describes a scalable, easy to understand approach to big data systems that can be built and run by a small team.
- 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

Grading

Details in the Canvas page. The breakdown is as follows:

- Three group homework assignments: 30%
- Weekly take-home quizzes: 20%
- Project proposal: 10%
- Final project: 30%
- In person final exam: 10%

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 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 Canvasmessages.
- **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 todiscuss 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 asparticipation 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 suspected 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. Therefore, the syllabus and course plan may change in the quarter. I always welcome feed- back from you about what is working and not working for your learning in the course.

Topics

- Notebook Computing
  - Project Jupyter
  - Data science environments
  - Managed notebook services
  - Amazon SageMaker Studio
- Cloud Concepts
  - Definition of a web service
  - Cloud providers
  - Six advantages of cloud computing
  - Different types of cloud computing models (e.g. IAAS, PAAS, SAAS)

    - 5 Principles of cloud computing
    - A new computing paradigm
- JupyterLab Interface
    - Jupyter notebook format
    - JupyterLab notebook model
    - Kernels
    - Instances
    - GitHub integration
    - Cloning repositories
- AWS Cloud Security and Billing
    - Shared responsibility model
    - AWS IAM
    - IAM users, groups, policies, and roles
    - AWS pricing model
    - Securing a new AWS account
    - AWS Console
    - AWS Billing and Cost Explorer
    - Setup Amazon CloudWatch Billing Alarms
    - AWS Cloud Shell
- Cloud Prerequisites
    - Common Linux distributions on AWS
    - YUM and APT
    - Basic commands such as ls, cp and chmod
    - JSON
    - RESTful APIs
- AWS Services
    - Main AWS service categories and core services
    - Regional and Zonal services
    - Services with no charge
    - AWS APIs
    - AWS CLI
    - AWS Python SDK
- Amazon Simple Storage Service (S3)
    - Block storage versus object storage
    - S3 overview
    - S3 storage classes
    - IAM policies
    - Bucket URLs (two styles)
    - Three common use cases
    - S3 pricing
    - AWS CLI commands for S3

    - Python boto3 for S3
    - Registry of Open Data on AWS
- AWS Machine Learning APIs
    - Amazon Rekognition (computer vision service)
    - Amazon Comprehend (NLP service)
    - Amazon Translate
    - Amazon Transcribe (speech to text service)
    - Amazon Polly (text to speech service)
- Amazon Elastic Compute Service (EC2)
    - Example use cases
    - EC2 overview
    - Amazon Machine Image
    - Instance types
    - User data scripts
    - Storage options
    - Tagging
    - Security group settings
    - EC2 pricing
    - Four pillars of cost optimization
- Amazon Elastic Container Registry (ECR)
    - Container basics
    - What is Docker
    - JupyterLab on EC2 via Docker
    - Amazon ECR overview
    - SageMaker Docker images for deep learning
- AWS Lambda
    - Serverless AWS services
    - Benefits of Lambda
    - Event sources
    - Lambda function configuration
    - AWS Lambda limits
    - Use Lambda to execute and schedule notebooks

Course Schedule - TBD