[639] Data Management for Data Science

Meenakshi Syamkumar

Why is Data Management crucial?

- Transforming raw data into knowledge and knowledge into business value!
- What does this course teach?
  - How to effectively use various data organization tools?
  - How to gain insights using predictive analysis?
  - How to tell stories using data?

Source: DALL·E 2

About myself

Dr. Meenakshi Syamkumar

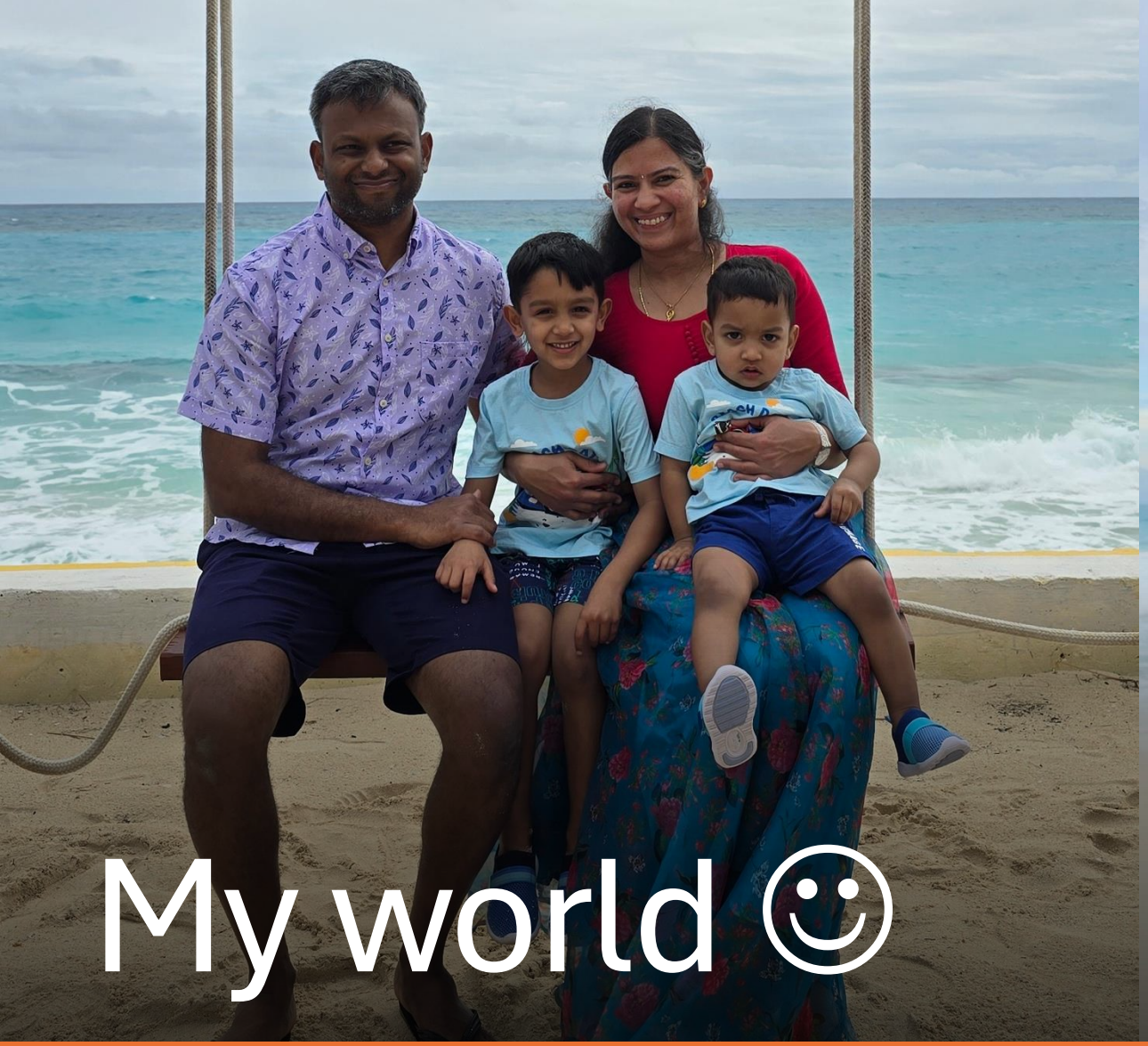
- Email: ms@cs.wisc.edu
- Please call me “Meena”
- Pronouns: she / her / hers

Industry and Teaching experience

- Citrix, Cisco, and Microsoft
- CS220, CS300, CS320, CS367, CS544
- Guest lectures CS640, CS740

Research

- CS / DS education
- Past: network measurements



My world ☺

Passions

- Running / working out
- Gardening
- Women in CS / DS
- Promoting breastmilk donorship

Jin Pan

Ricky Wang

Avi Trost

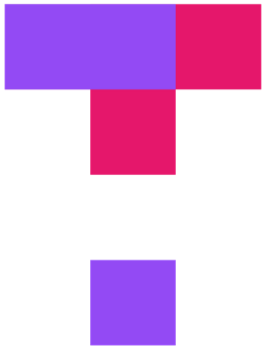
CS639 team

Yuna Hwang

(Head TA)

Lisheng Ren

Who are you?



TOP HAT

Current standing? Major?

CS course experience

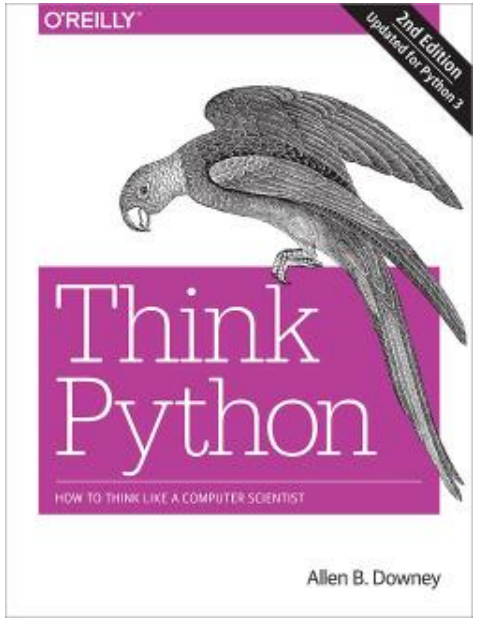
- 320
- 400
- 537/544/564/640?

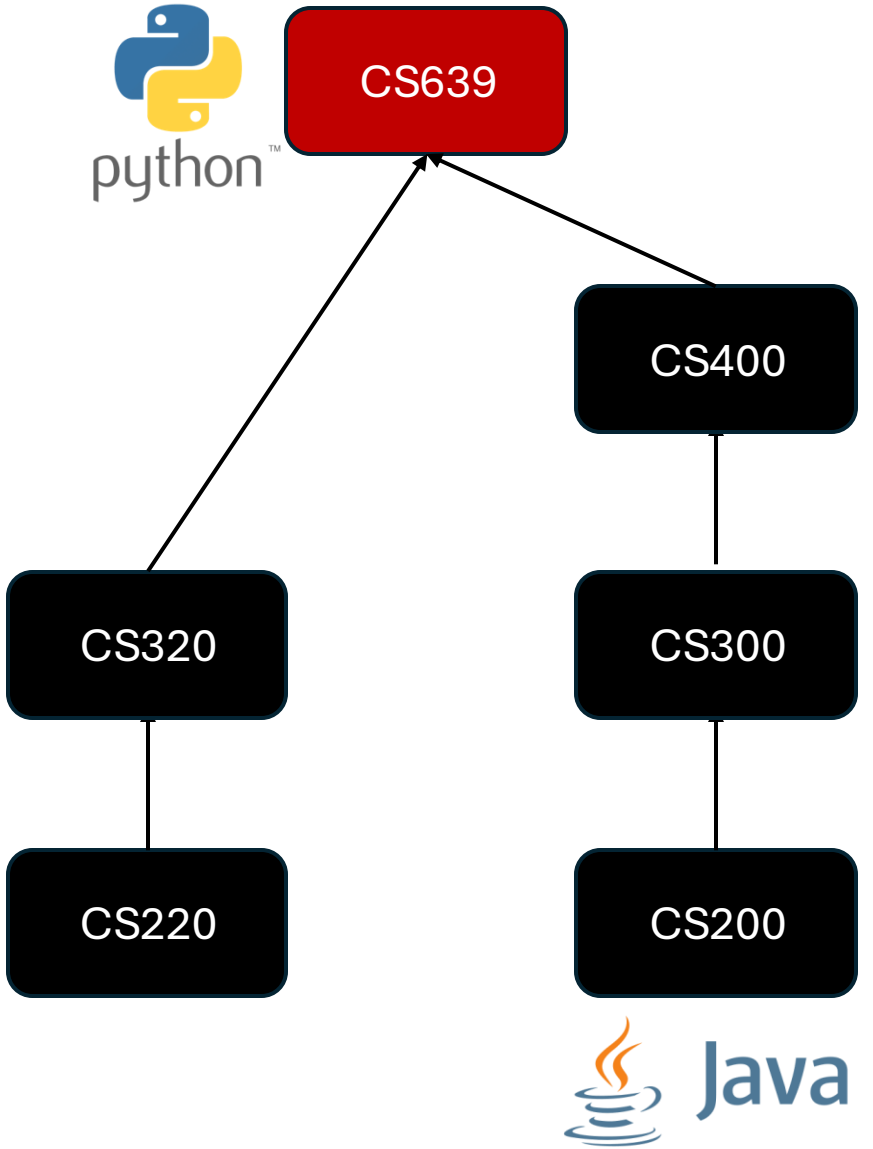
Participation credit

- Please fill this form: https://forms.gle/p9hz8dt8R9GCd1gu9
- **Due:** *Monday, February 3rd*

Where does CS639 fit?

- “Advanced Computing” course for Data Science Major
- Numbered course: Spring 2026
- Programming language: Python





Learning Objectives

01

Course logistics, and policies

02

Key aspects of Data Management

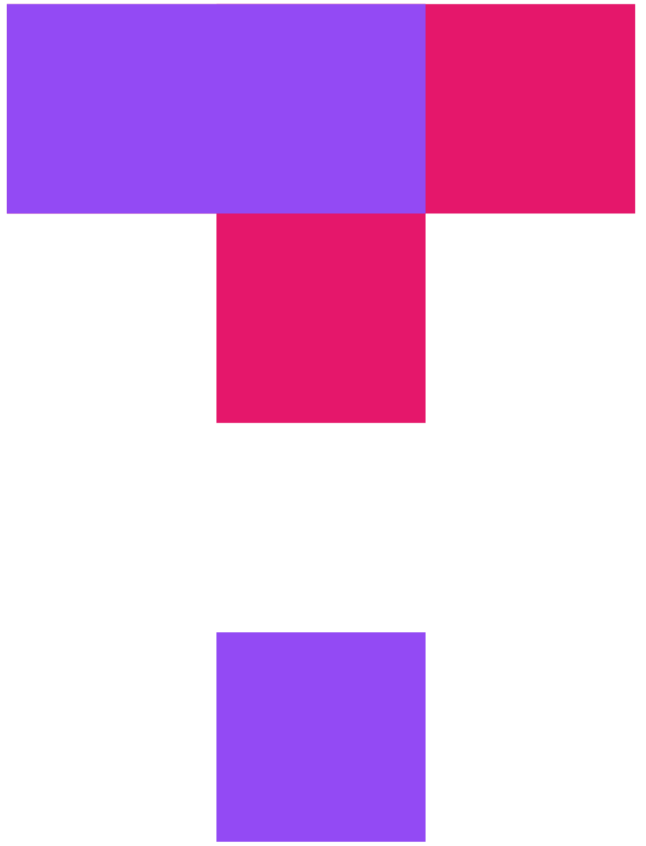
03

Overall course plan

Course logistics, and policies

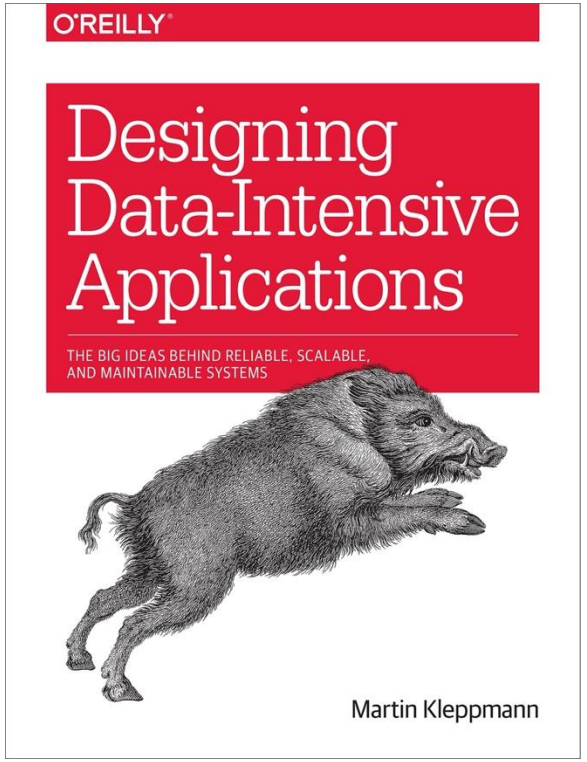
Lectures

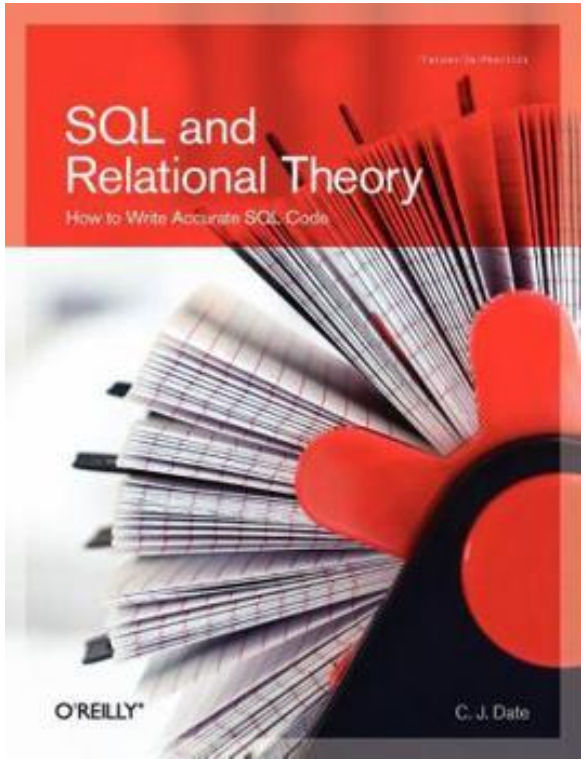
- In-person lectures
- TopHat:
  - 0.5 points for attendance and 0.5 for correctness
  - 2% extra credit
- Recordings:
  - Lecture recordings will not be posted for in-person sessions

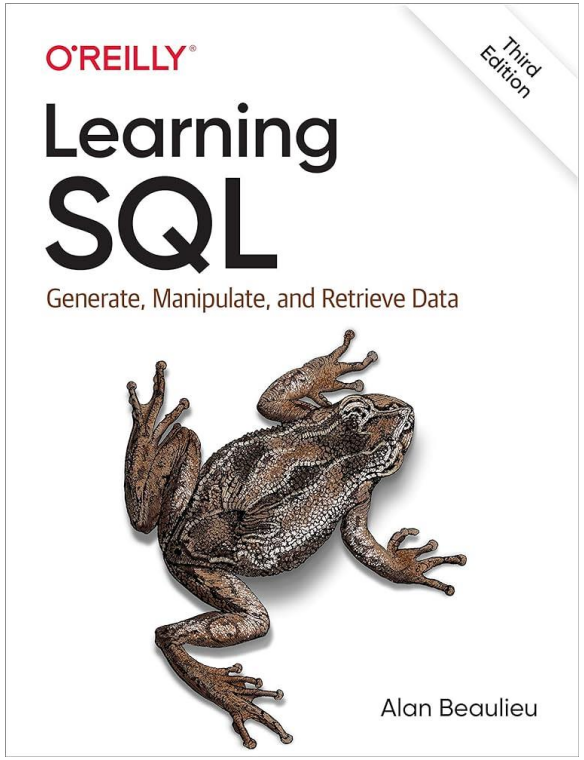


TOP HAT

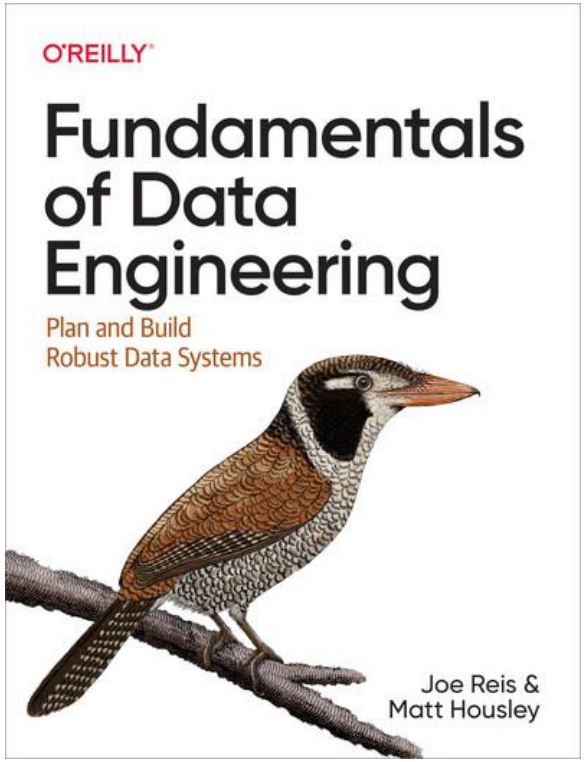
Readings (optional)

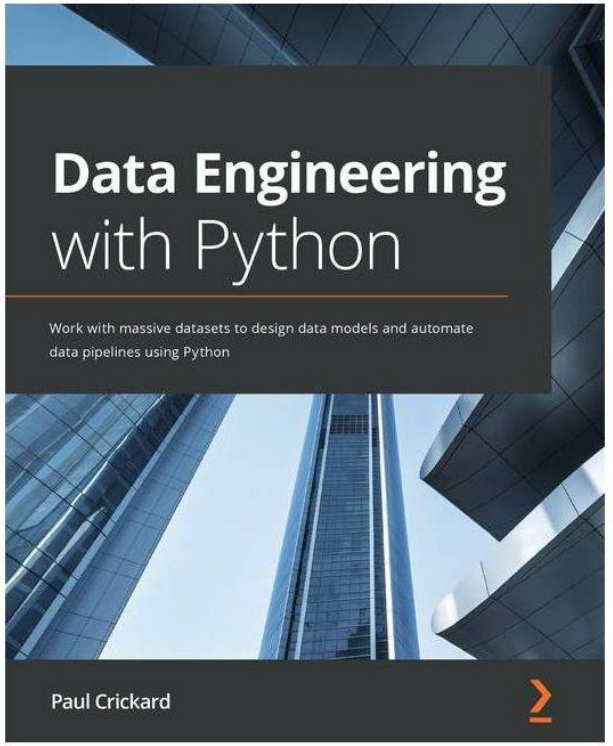






Readings (optional)





Websites

Course website

- https://ms.sites.cs.wisc.edu/cs639/s25/schedule.html
- schedule, course content, syllabus
- Deadlines:
  - Quizzes:
    - Assigned on Mondays, due on Wednesdays
  - Projects:
    - Approximately bi-weekly
  - Midterm:
    - March 3rd

Github repository

- https://github.com/CS639-Data-Management-for-Data-Science/s25
- Project specification and lecture code

Canvas

- announcements (class-wide & personal)
- quizzes
- grade summaries

Other resources

- Piazza (asking questions of general interest)
  - don't post project code
  - answer other student’s questions (you’ll get credit)

- Office hours (asking questions of individual interest)
  - Email: least-preferred (try to utilize office hours)
  - our goal: responses <2 business days
  - feel free to escalate by CC'ing instructor on same thread after 2 days
  - if contacting multiple staff members about same issue, please keep all on the same thread (don't start multiple threads)

- GitHub classroom
  - you'll be given a private repo for each project
  - we'll provide feedback on GitHub

Grading

Programming projects – 48%	6 projects – 8% each (no drops)
	Max group size: 2
	Hard deadline: 3 days after soft deadline (10% late penalty per day)
Quizzes – 15%	12 quizzes (2 drops) – 1.5% each
	Open materials
Exams – 35%	Midterm – 15%
	Final – 20%
Participation – 2%	Surveys
	Endorsed Piazza contribution
TopHat – 2%	Extra credit

Key Aspects of Data Management

Data-driven world

- Data Collection & Ingestion
- Data Cleaning & Transformation
- Data Organization / Storage
- Data Integration / Fusion
- Data Security / Compliance
- Data Orchestration
- Data Cataloging

Source: DALL·E 2

Data Organization / Storage

- Databases
  - Relational data – SQL
  - Non-relational data – NoSQL
- Data Warehousing
  - Centralized repositories of structured / semi-structured data
  - Optimized for querying and reporting
- Data Lakes
  - Storage of raw, unstructured, semi-structured, and structured data
  - Ideal for big data analytics, and machine learning

amazon REDSHIFT

Google Cloud Storage

Azure Data Lake Storage

Data Integration / Fusion

- Extract, Transform, Load (ETL)
  - Extract data from multiple sources
  - Transform data: cleaning, filtering, aggregation, normalization
  - Load: data warehouse / database
  - Highly structured and consistent data
- Extract, Load, Transform (ELT)
  - Load: data lake
  - Raw data
- Fusion
  - Integration
  - Reduction

Data Governance

- Data quality
  - Pre-written rules
  - Accuracy, completeness, and reliability
- Data security
  - Preventing unauthorized access
  - Encryption
  - Access control
- Data compliance
  - Laws governing how to collect, store, and process data
  - The General Data Protection Regulation (GDPR)
  - The Health Insurance Portability and Accountability Act (HIPAA)
  - The California Consumer Privacy Act (CCPA)

Data Orchestration

- Data Integration
  - Integrates SILOs
- Data mapping
- Data modeling
- Data governance
- Ideal for real-time data analysis

Data Cataloging

- Metadata management
  - Descriptions
  - Classifications
  - Lineage
- Data documentation
  - Sources
  - Structures
  - Pre-processing
  - Ensures transparency and reproducibility

Apache Atlas

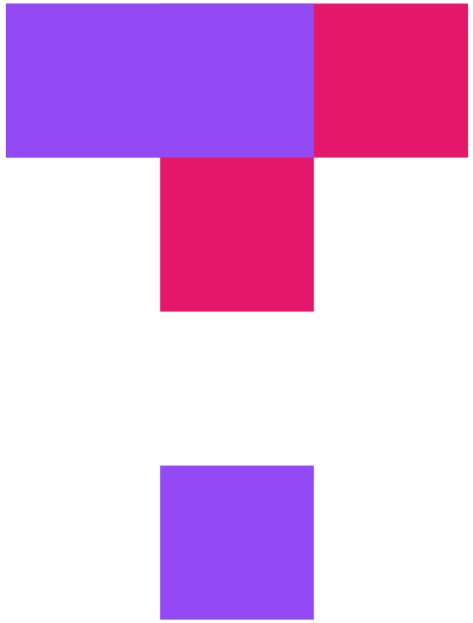
Overall course plan

Tentative topic list

- *Linux + Docker*
- *RDBMS + SQL*
- *NoSQL: MongoDB*
- *NoSQL: Elasticsearch*
- *Data generation and source systems*
- *Data ingestion*
- *Data lakes + Data warehouses (Snowflake)*
- *ETL, ELT, & Data pipelines*
- *Predictive analytics: time series analysis*
- *Predictive analytics: boosting - XGBoost, LightGBM, CatBoost*
- *Vector databases: RAG, prompt engineering*
- *Visualization: dashboards, data storytelling*

Course learning objectives

- *Deploy SQL and NoSQL databases to manage structured, semi-structured, and unstructured data and write programs for analysis*
- *Gain proficiency in data integration techniques such as ETL and ELT, and demonstrate competencies with each stage of the ETL process*
- *Apply and compare prebuilt implementations of popular gradient boosting algorithms for time series prediction*
- *Develop foundational skills in fine-tuning Large Language Model (LLM) implementations with custom unstructured data*
- *Understand the fundamentals of Retrieval-Augmented Generation (RAG) by learning to work with vector databases*
- Demonstrate visualization competencies for creating interactive plots, dashboards and data stories



TOP HAT

Action items

P1 will be released by mid of next week	We'll keep you posted as soon as we get approval for GCP credits
Fill out class participation survey:	https://forms.gle/p9hz8dt8R9GCd1gu9
Review course syllabus	https://ms.sites.cs.wisc.edu/cs639/s25/syllabus.html