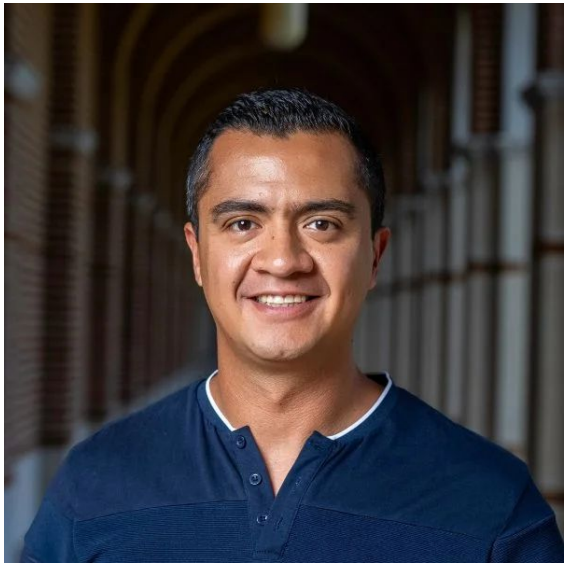
COMP 543:
Tools & Models for Data Science



Drs. Luis Guzman & Sinan Kockara

Meeting Time and Location

- MWF 10:00 am
- Herzstein Hall 210

Staff

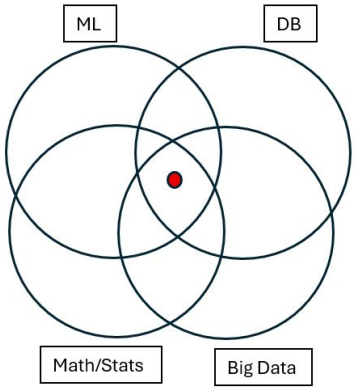
- Siddharth Balakrishnan (sb233@rice.edu)
- Janet Jiang (jj65@rice.edu)
- Chandana Mohan (cm203@rice.edu)
- Shivam Pathak (sp151@rice.edu)
- Yiyi Tang (yt61@rice.edu)
- Samhita Vinay (sv57@rice.edu)
- Ziyou Zhao (zz122@rice.edu)

So, what is COMP 543 about?

Course Description

- Introduction to Data Science
  - Extracting actionable, non-trivial knowledge from data
  - Focus on the software tools used by practitioners
  - Mathematical, statistical, and ML models used in conjunction with such tools

- Relation to other classes
  - First 3~4 weeks are on DB
    - Assignments 1 and 2 will be similar to COMP 430/530
  - Some overlap with ML and math/statistics classes
    - Linear regression, logistic regression, KNNs, Neural Networks, clustering
    - **NOT** a machine learning class - **Big-data focus**



So, what is actually covered?

Course Contents (Tentative)

- Intro to Data Science
- Intro to Relational Databases
- The Relational Calculus & Relational Algebra
- Declarative SQL
- Imperative SQL
- Intro to Modeling
- Optimization basics: Gradient descent
- Optimization basics: Newton's method
- Intro to Big Data: Map Reduce Paradigm
- Hadoop Programming (Java)
- Optimization basics: The EM Algorithm (Math derivations)
- Spark Programming (Python)
- Intro to Supervised Learning
- Linear Regression
- Generalized Linear Models (Math derivations)
- Overfitting and Regularization
- Outliers
- Intro to Neural Networks
- Learning in Neural Networks: Backpropagation (Math derivations)
- Recurrent Neural Networks
- Deep Learning with LSTM
- Intro to Unsupervised Learning
- Dimensionality Reduction
- Mixture Models

That seems like a lot! How do you evaluate?

Course Evaluation

- Final grade is computed as follows:
  - Assignments 60% (6 total, 10% each)
  - Homeworks 25% (5 total, 5% each)
  - Quizzes 10% (~35, ~0.28..% each)
  - Labs 5% (5 total, 1% each)

- Usual letter grade ranges (i.e., A- >= 90%, B- >= 80%, etc.)
  - A+ >= 97%, no curve

Deliverable Features

	Duration	Late Submissions	Extensions	Due Time
Assignments	2 weeks	Yes	Yes	11:55 pm
Homeworks	~7-10 days	Yes	Yes	11:55 pm
Labs	~5 days	No	No	11:55 pm
Quizzes	~12 hrs	No	No	11:59 pm

Late submissions

- Up to 2 days (Quizzes excluded)
- 10% penalty per day

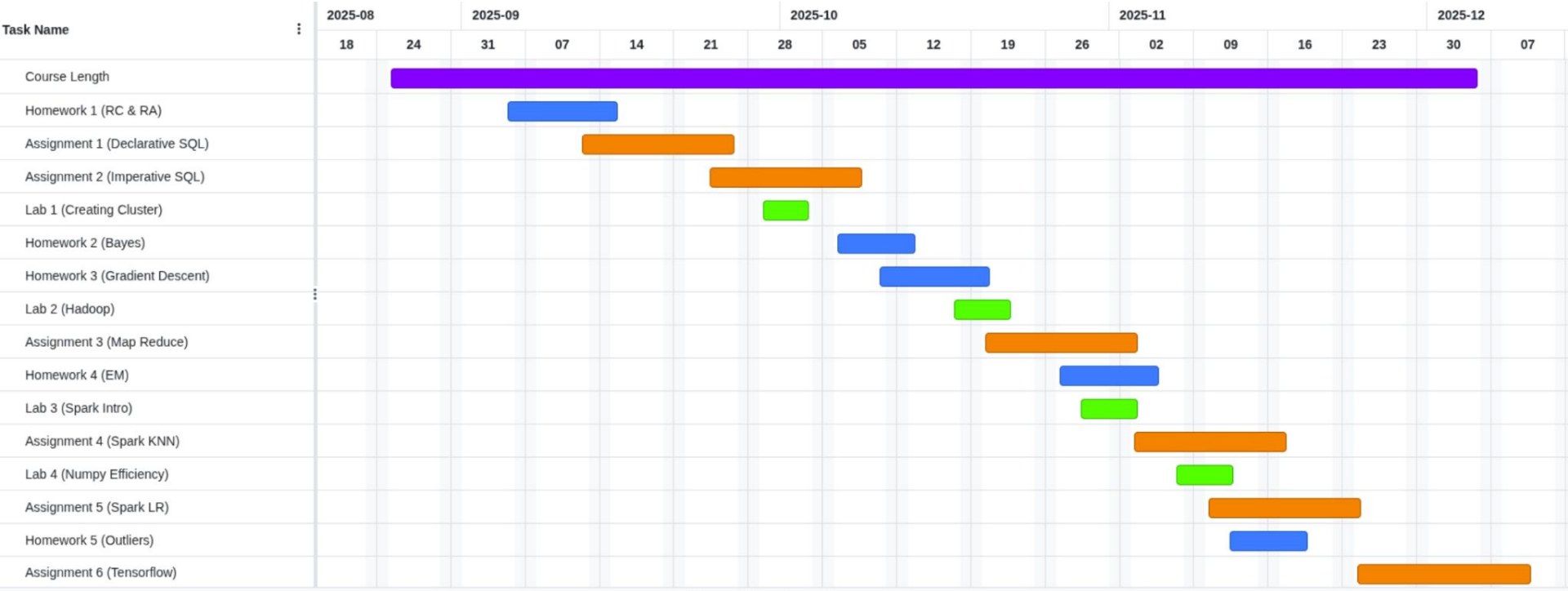
Extensions

- Super chill
  - Just ask **via email** (either instructor) with >= 7 days before the **due date**
    - Explain the reason for the extension request to instructors (e.g., illness, conference, or other valid circumstances).
  - Extension will be reflected in Canvas
  - Warning: Extensions **do NOT** affect other deliverable due dates
- Why are there **no extensions on Labs or Quizzes**?
  - Labs are warm ups to assignments, it would defeat their purpose
  - Quizzes aim to promote attendance, 4 lowest scores are dropped
  - Auditors must complete at least 60% of the quizzes (~21 Quizzes).

Regrade Requests

- Again, very chill
  - Just email your grader: should resolve most issues
  - Otherwise, **submit in writing (hard copy)** to instructor(s)

Deliverables Schedule



Communications and Resources

- **Piazza**
  - Most technical and logistical issues
  - Short announcements: changes to OHs, quiz releases, assignment tips
  - Post publicly as much as possible, post privately only if needed
- **Extensions and special requests**
  - Email instructors directly
  - **MUST include COMP 543** in subject line
- **No textbook**
  - Lecture slides are posted on Piazza
- **Canvas**
  - Mostly just for assignment submissions and grading

Office Hour Schedule

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
			Wednesday, 9:00 AM *Sinan Kockara* DCH 3051			
	Monday, 10:00 AM *Lecture* Herzstein Hall 210		**Wednesday, 10:00 AM** *Lecture* Herzstein Hall 210		**Friday, 10:00 AM** *Lecture* Herzstein Hall 210	**Saturday, 10:00 AM** *Siddharth Balakrishnan* Zoom
	Monday, 11:00 AM *Janet Jiang* Zoom					
	Monday, 1:00 PM *Yiyi Tang* TBD	**Tuesday, 1:00 PM** *Yiyi Tang* Zoom		**Thursday, 1:00 PM** *Chandana Mohan* TBD		
	Monday, 2:00 PM *Shivam Pathak* Sid Rich Commons		**Wednesday, 2:00 PM** *Shivam Pathak* Sid Rich Commons		**Friday, 2:00 PM** *Luis Guzman* DCH 3092	
		Tuesday, 3:00 PM *Ziyou Zhao* DCH 3rd floor		**Thursday, 3:00 PM** *Chandana Mohan* TBD		
		Tuesday, 4:00 PM *Siddharth Balakrishnan* TBD				
			Wednesday, 5:00 PM *Ziyou Zhao* Zoom			
	Monday, 7:00 PM *Samhita Vinay* Zoom		**Wednesday, 7:00 PM** *Samhita Vinay* Wiess Commons			

Student Reviews

Absolutely rockstar of a course

- *This is a very good course that **gives you the basics fundamentals of Big Data, SQL Database, and also for data science**.*
- *As a survey course of data analysis techniques, there was a lot of ground to cover. I feel it **did a good job hitting the highlights of each area, giving me sufficient information to serve starting points for my own further exploration**.*
- *It's a **great course**. The content is **rich and interesting**, and the **workload and difficulty are moderate**.*
- ***Great course** that offers a lot of opportunities for help.*
- ***Great course** for big data analysis and NLP.*
- *This is a helpful course for finding a SDE job.*
- *It was an **amazing course**.*

Student Reviews

With super clear and detailed assignment instructions

- *Good course materials with **detailed instructions for each assignment** experience.*
- *This course is a mixture of lots of concepts, models, and tools. The **instructions on assignments are clear** while some assignments would be challenging and time consuming.*

Or, are they?

- *Course introduced many concepts at a low level. **Some assignment details felt unclear** but overall made sense.*
- *Overall a very organized course with comprehensive contents. But I think **the specs for homework are somewhat vague**. If those could be more clear, the course will be better.*

Student Reviews

Somewhat superficial and workload-heavy

- *A well-organized course.* ***My only regret is that, in my opinion, there is so much to learn, that we cannot delve deeply into these fields****.*
- *The course was very reliant upon assignments, though the lecture material, while feeling brief due to the short class times, did cover the basics of what was necessary to learn. However,* ***it did feel overwhelming at points personally****.*
- *The* ***workload is a bit heavy****, but there is no group work or exam, which is good.*
- *We were* ***trying to cover way too many things*** *to really learn any one thing.*
- *Lots of* ***overlapping assignments***

- **Accommodations**
  - Send the instructors your DRC letter as soon as possible
  - We will try to accommodate your needs as best as we can

- **Disclaimer for AWS credits**
  - Students will receive a set amount of credits (TBD)
  - During the account creation process, students will need to provide credit card information. This credit card will be used to cover any costs that exceed the provided credits
  - Instructors are not responsible for any costs incurred by students beyond the allotted credits
  - Once the credits have been depleted, they cannot be replenished
  - Students are responsible of monitoring their AWS usage and to terminate any unused compute nodes to avoid additional charges.

- **Academic Misconduct**
  - Standard honor code policies
  - Cannot transmit or receive code from or to anyone in class in any way.
  - Collaboration with anyone outside of class is not allowed either:
    - Looking for solutions from prior semesters
    - Posting on StackOverflow
  - GenAI models are allowed for learning purposes only
    - Ask general questions, clarify in-clas concepts, debugging code
    - AI-generated solutions are considered honor-code violations
    - Do not make yourself replaceable
  - Cite outside sources if using small snippets of code
    - Include URL or prompt/model

Looking forward to working with all of you!

Wish you a great beginning of the semester!