DATA 140: Introduction to Data Structures and Management

T/Th 09:30 AM – 10:45 AM

Davie Hall Room 112

A little bit about me

- I am a third-year CS PhD student advised by Shashank Srivastava
  - Understanding **when and why deception arises** in large language models—and developing systematic methods to detect and mitigate it.
  - Why do LLMs produce deceptive outputs?
  - How can we systematically detect deception?
  - What interventions improve alignment and safety?
- Before starting my PhD I worked on Wall Street as a quantitative researcher
- Life Outside the Lab:
  - **Blues Guitar:** I don’t let myself practice for more than 2 hours on school days otherwise nothing gets done.
  - **Sports:** Lifelong Boston sports fan. Go Pats!
  - **Comedy:** I’m also a huge fan of stand-up comedy

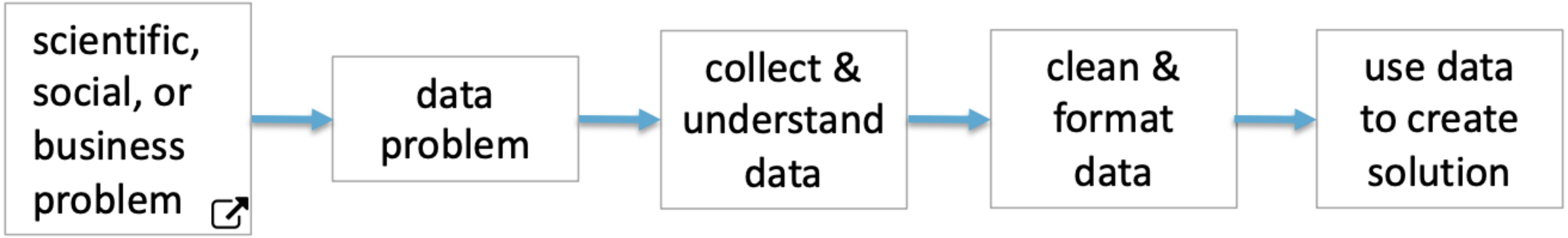
Course TAs

- TA: Zhaoyang Wang (zhaoyang@unc.edu)
- ULAs: Brad Royal (royalsbr@unc.edu), Tina Zou (tinazyx@unc.edu)
- Office hour times and locations can be found in syllabus

- **We get a lot of emails so please prefix the subject to all your emails “Data140 –”**
  - Ex. Data140 – Question about the Midterm
  - Data140 – Grading Issue

WHAT IS DATA SCIENCE?

…solving problems with data…



scientific, social, or business problem

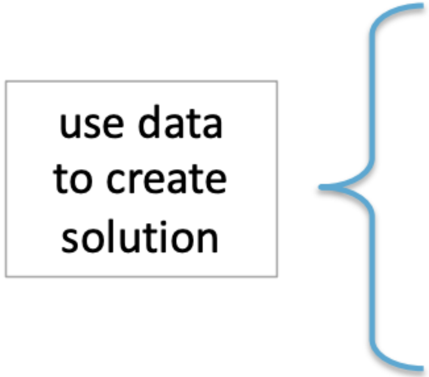
data problem

collect & understand data

clean & format data

use data to create solution

…which step is most challenging?



use data to create solution

data analysis

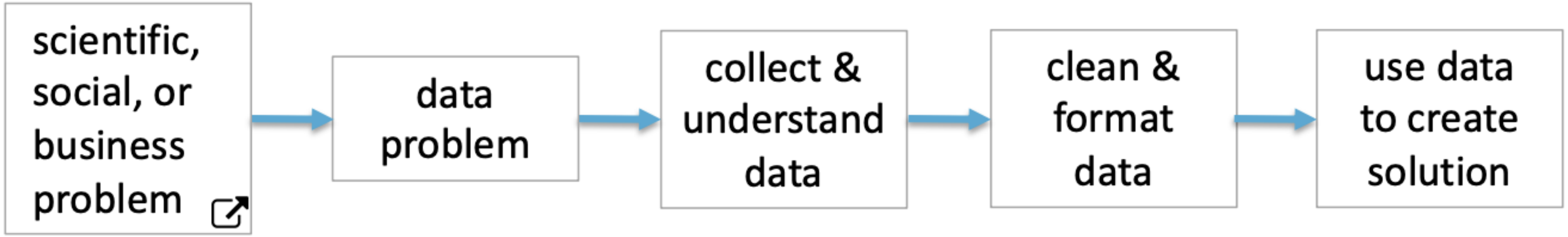
or

machine learning

(*or both*)

WHAT IS DATA SCIENCE?

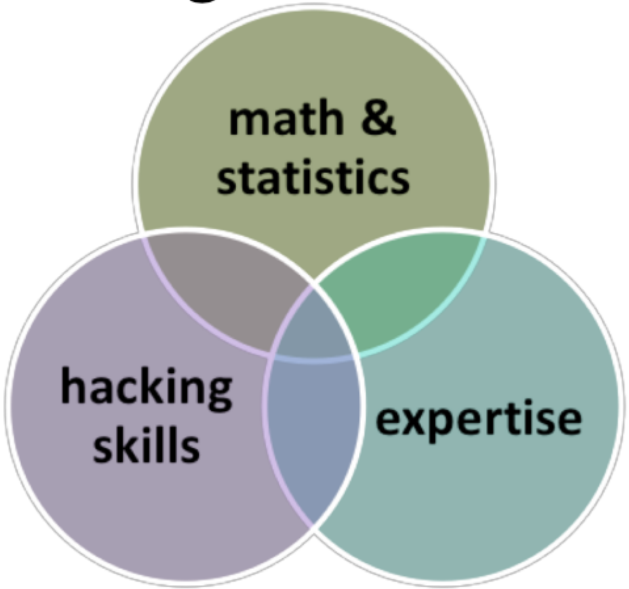
...solving problems with data...



scientific, social, or business problem → data problem → collect & understand data → clean & format data → use data to create solution

...sounds cool!

What makes a good data scientist?



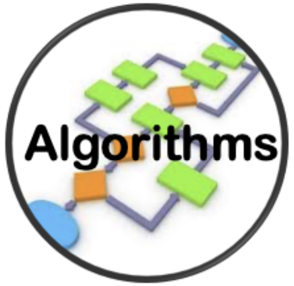
[PDSH] p xi
[DSFS] p 1-2

WHAT IS DATA ANALYSIS?

…using data to discover useful information…

- **data**: anything you can *measure* or *record*

- **statistics**: summarize (and visualize) *main characteristics* of the data



- **algorithms**: apply algorithms to find *patterns* in the data

WHAT IS MACHINE LEARNING?

…creating and using models that learn from data…

- **data**: anything you can *measure* or *record*

- **model**: specification of a (mathematical) *relationship* between different variables

- **evaluation**: how well does the model *work*?

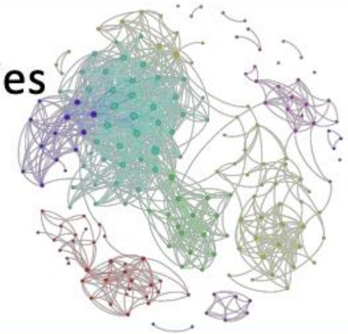
WHAT IS MACHINE LEARNING?

…creating and using models that learn from data…

<u>Examples</u>

Identifying zip code from handwritten digits

Detecting communities in social networks



Predicting the traffic volume at rush hour

Detecting fraudulent credit card transactions

Determining the location of distribution centers based on customers' residence

[DSFS] p 3-13

Learning Path Overview

Module 1: Foundations of Data & Data Quality (Jan 8 - Feb 5)

- **Data Types & Structures:** Primitive types, complex types, vectors
- **Data Organization:** Tables, keys, indexing, normalization
- **Data Quality:** Cleaning, validation, handling missing or inconsistent data

Module 2: Data Structures and Algorithms (Feb 10 – Mar 10)

- **Searching Algorithms:** Linear, binary, DFS, BFS
- **Sorting Algorithms:** Insertion, Merge, Quick Sort
- **Hashing:** Hash tables, collisions, efficient lookups
- **Applications:** Coding interview practice, algorithm review sessions

Module 3: Big Data, Visualization & Machine Learning (Mar 12 – Apr 27)

- **Big Data Concepts:** Distributed computing, MapReduce
- **Data Visualization:** Basic visualization techniques and best practices
- **Machine Learning Foundations:** Linear regression, classification, clustering, anomaly detection, NLP

Course Deliverables

- Participation (10%)
- Perusall Assignments (20%)
  - 4 reading assignments with community discussion
- Colab Assignments (10%)
  - 4 hands-on coding projects in Google Colab
  - No coding experience required or expected, we will be “vibe coding”
- 2 Midterms (15% each)
  - Non-cumulative exams testing material in each module
- Final Exam (30%)
  - Cumulative, covering entire course

Participation (10%)

- Participation will be tracked via a **short 2-question multiple-choice quiz** at the end of each lecture.
- Quizzes are **password-protected**; the password will be shared at the end of each lecture.
- Quizzes are **open-note and open-book**.
- Students who score **50% or higher** will receive **full participation credit** for that lecture.

Perusall Assignments (20%)

- 4 Readings graded automatically by student engagement
  - Comments left, replies to other students, likes, etc.
  - No required textbook! (All readings available through Perusall)
  - Additional Recommended books provided in Syllabus

1. Classics on Tidy Data and Normalization
2. Avoiding Data Pitfalls
3. Netflix Recommender System
4. Technical Debt in Machine Learning

Colab Assignments (10%)

- Notebooks will be provided that walk you through how to perform Data Analysis in Practice
- You will not be required to code but only “Vibe Code” or prompt Gemini

1. **SQL Commands** – You will gain hands on experience using SQL to query anonymized UNC student grades data
2. **Sorting Algorithms** – You will gain hands on experience programming fundamental sorting algorithms
3. **Linear Regression** – You will gain hands on experience using linear regression models to predict the NCAA March Madness Tournament
4. **PCA Exercises** - You will gain hands on experience applying PCA to image data using images of Pokémon Sprite’s

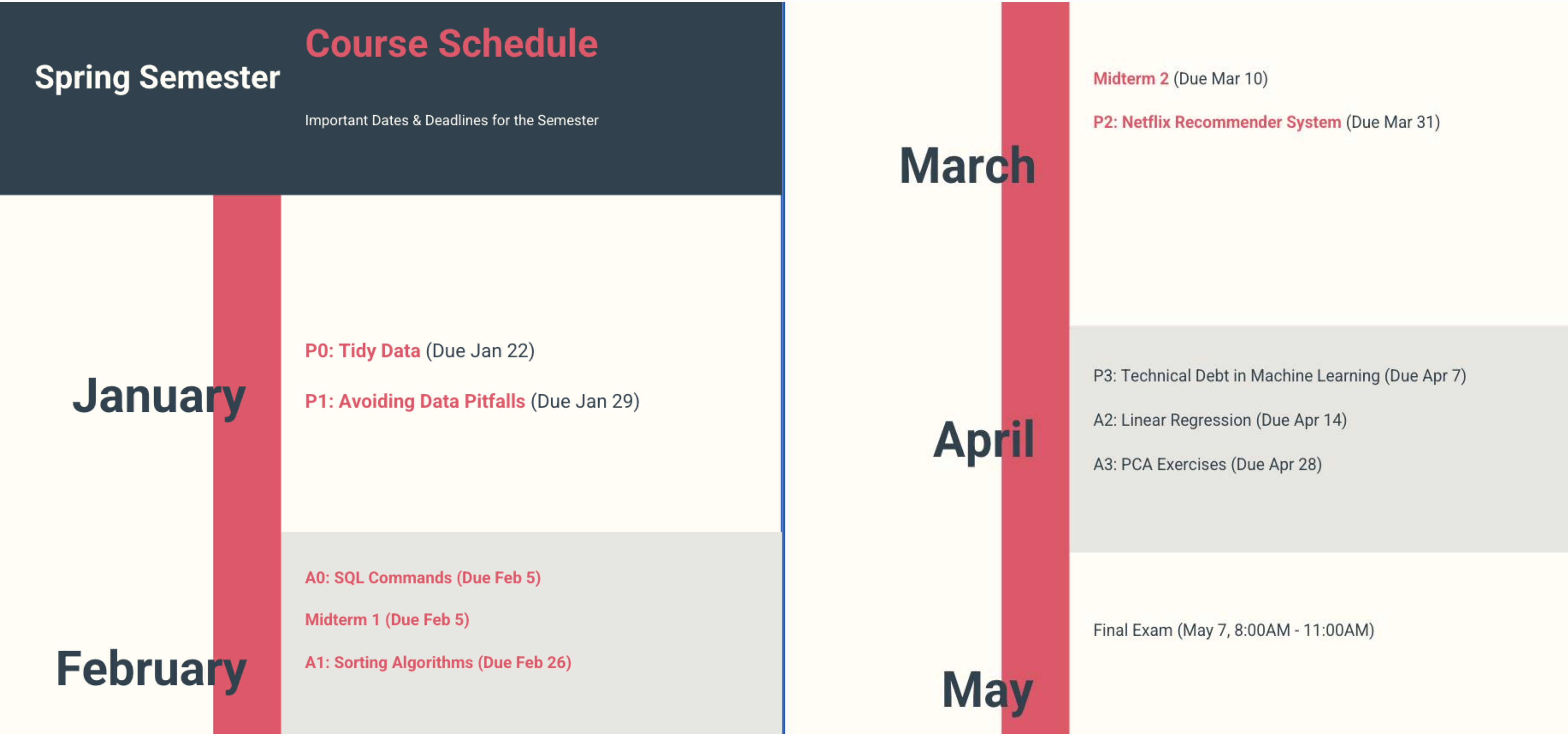
We encourage you to use AI Tools

1. ChatGPT/Gemini/Perplexity
   - Coding assignments (debugging, walkthroughs, code review suggestions)
   - Reading assistant for papers and docs (explanations, paraphrasing, analogies)
   - Studying (practice quizzes, notecards, exam-style questions from your notes)

2. NotebookLLM
   1. Course-pack / notes hub (upload slides, papers, homework specs into one notebook)
   2. Auto study materials (flashcards, targeted quizzes, concept maps grounded in your sources)
   3. Research assistant (summaries, syntheses, and “explain like I’m new to this topic” views)

- Perplexity / Claude / Poe
  - Research starting point (high-level overviews plus linked sources for further reading)
  - Comparing viewpoints or methods (e.g., different ML models or papers side-by-side)

Important Dates



Spring Semester

Course Schedule

Important Dates & Deadlines for the Semester

January

P0: Tidy Data (Due Jan 22)

P1: Avoiding Data Pitfalls (Due Jan 29)

February

A0: SQL Commands (Due Feb 5)

Midterm 1 (Due Feb 5)

A1: Sorting Algorithms (Due Feb 26)

March

Midterm 2 (Due Mar 10)

P2: Netflix Recommender System (Due Mar 31)

April

P3: Technical Debt in Machine Learning (Due Apr 7)

A2: Linear Regression (Due Apr 14)

A3: PCA Exercises (Due Apr 28)

May

Final Exam (May 7, 8:00AM - 11:00AM)

Welcome to Data140

For Next Time:

1. Read the Syllabus on Canvas
2. Join the Course on Perusall and Gradescope
3. Have a great weekend!