Introduction to Database Management Systems

Welcome and Course Logistics

Kaustubh Beedkar

Department of Computer Science and Engineering

Indian Institute of Technology Delhi

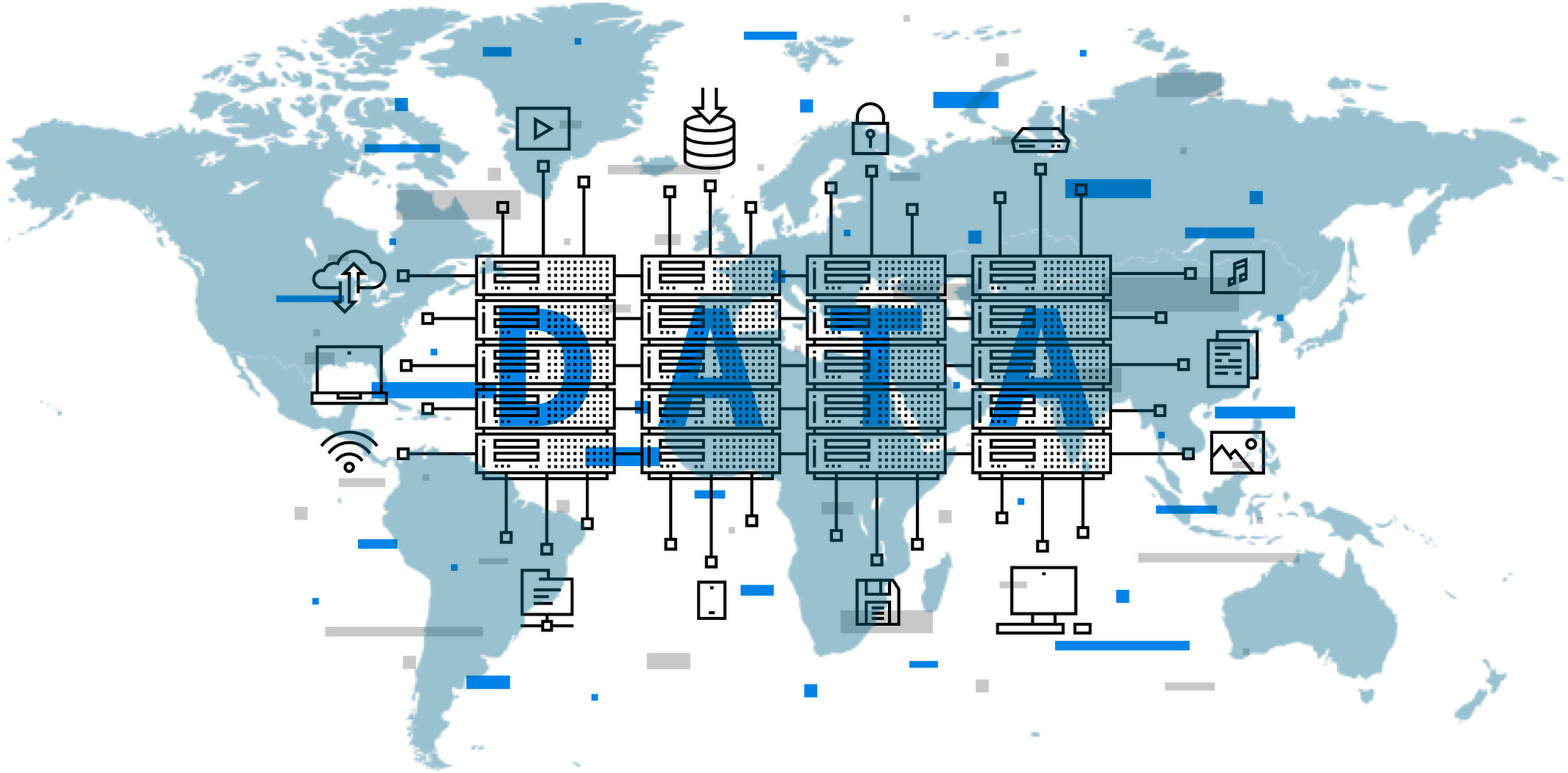
Today's Agenda

- Why DBMS?
- Course Introduction
- Course Logistics and Administrivia

Today's Agenda

- Why DBMS?
- Course Introduction
- Course Logistics and Administrivia

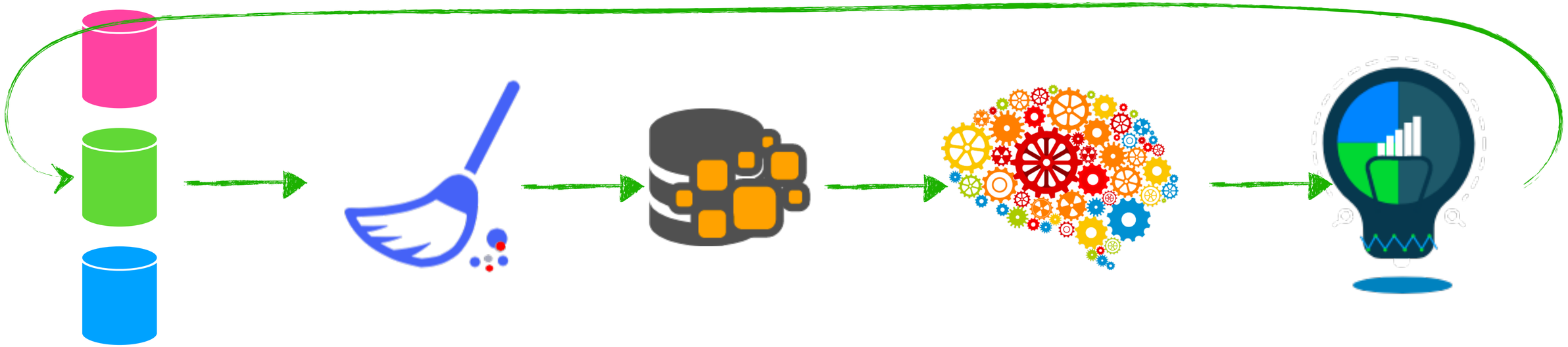
Data-driven World



Data is the New ~~Oil~~ Soil!

Soil → Worked → Treated → Exploitation → Result

Data Sources/Creation → Prepared → Enriched → Processed → Insights



Common Denominator

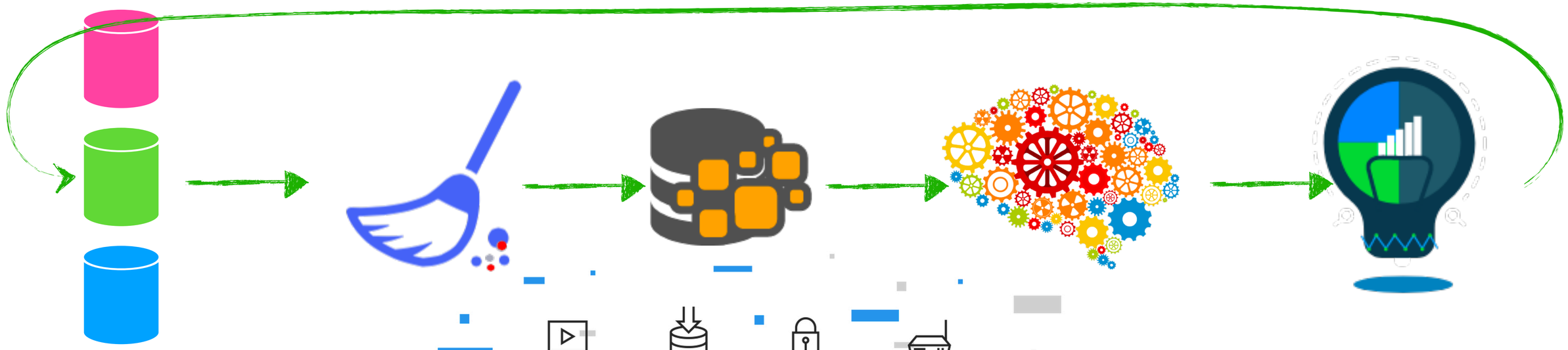
Data Sources/Creation

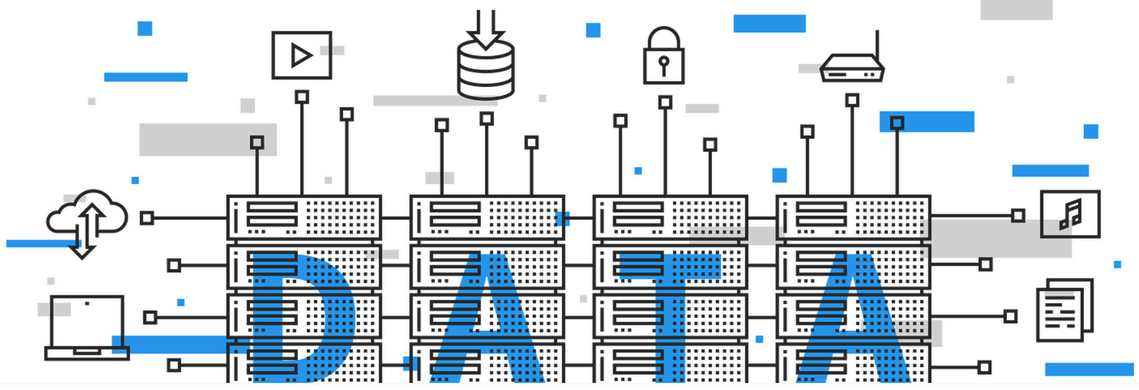
Prepared

Enriched

Processed

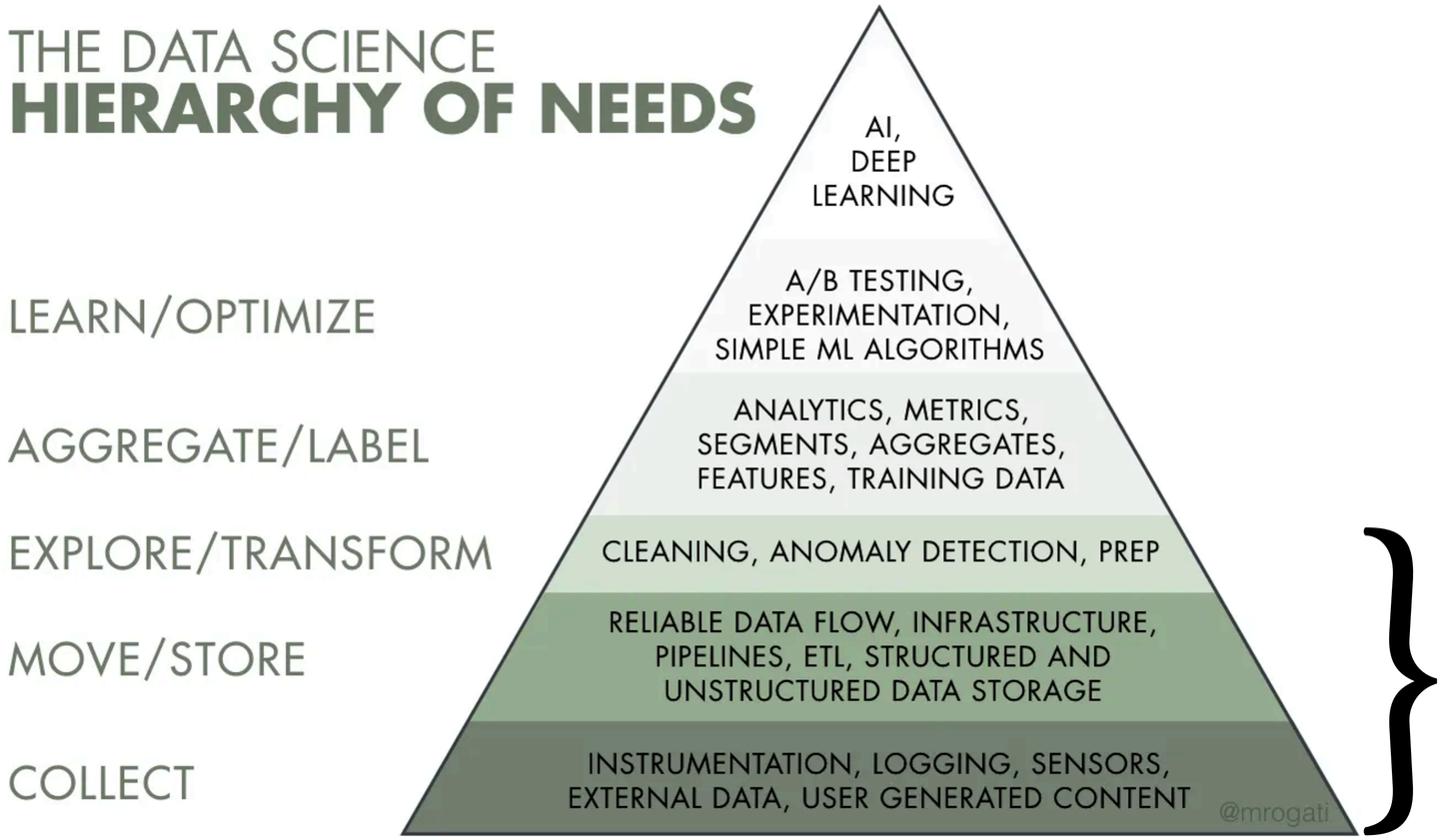
Insights





Mastering the data technology stack is crucial!

What About Data Science & AI?



You need
Data Systems
to do
Data Science & AI!

[Image Source]

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

Hidden Technical Debt in Machine Learning Systems

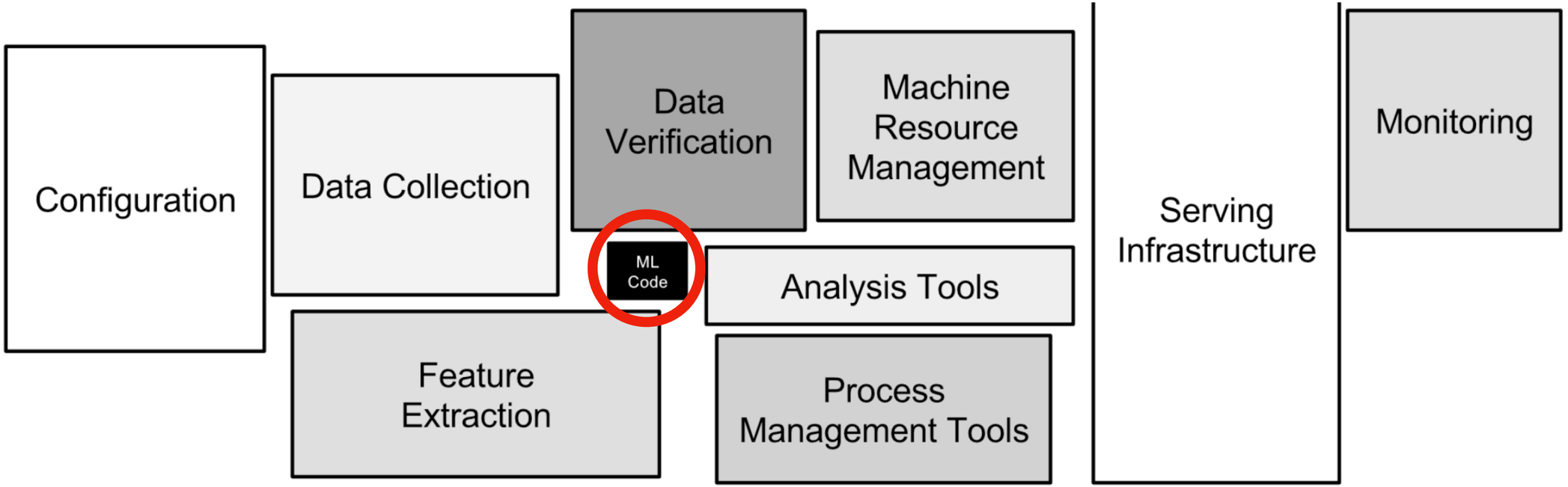


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source

Production ML

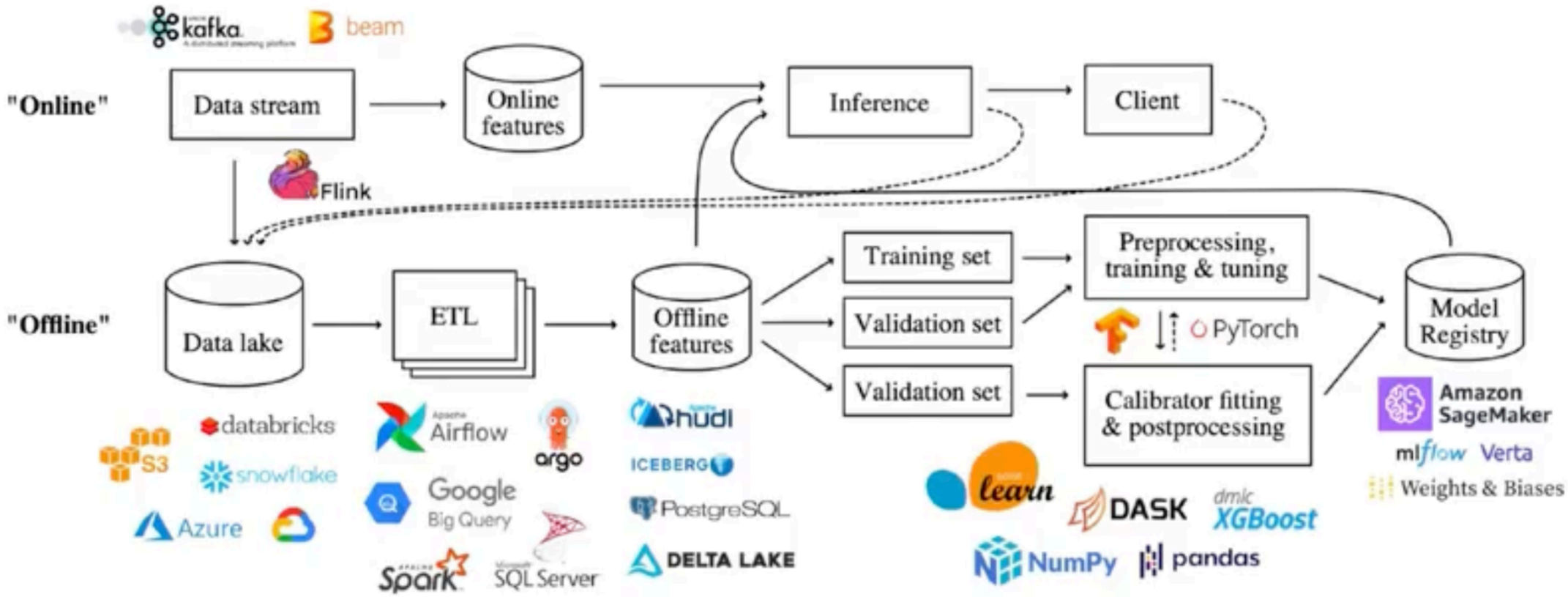
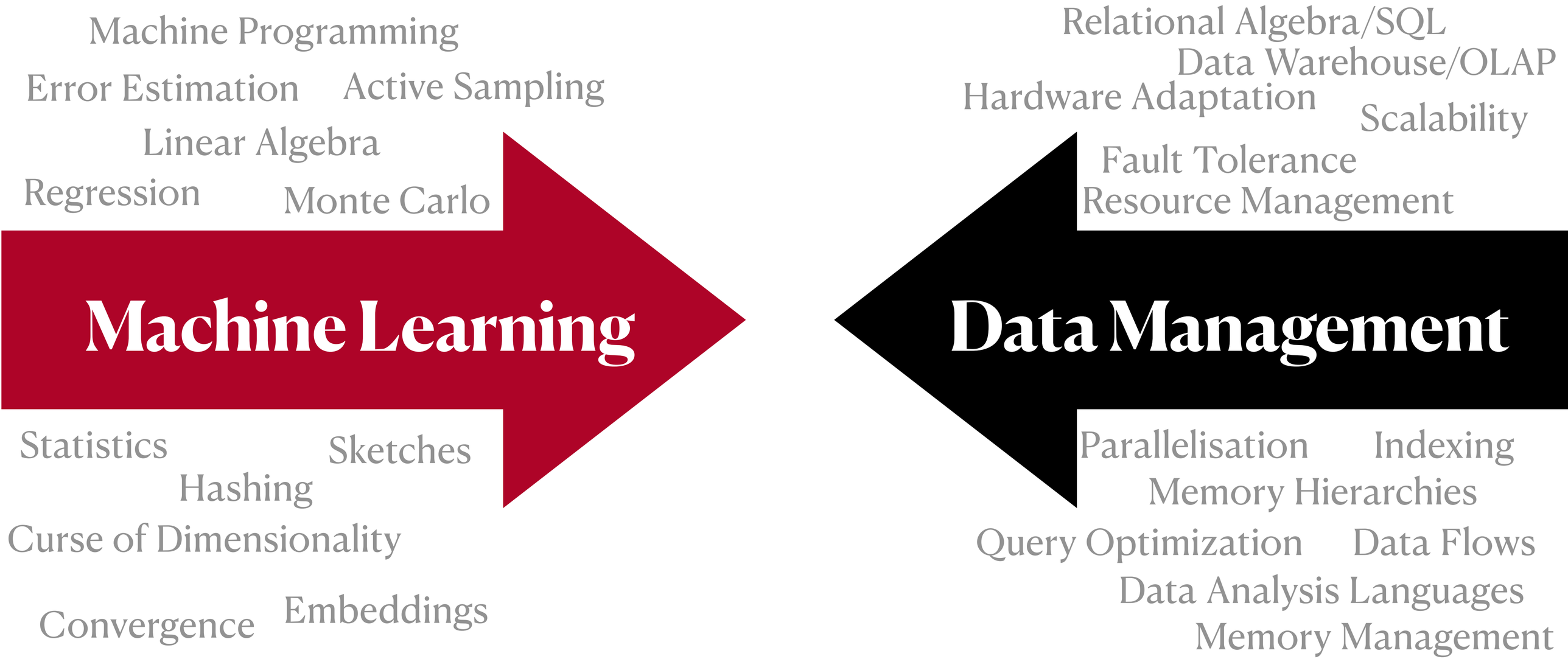


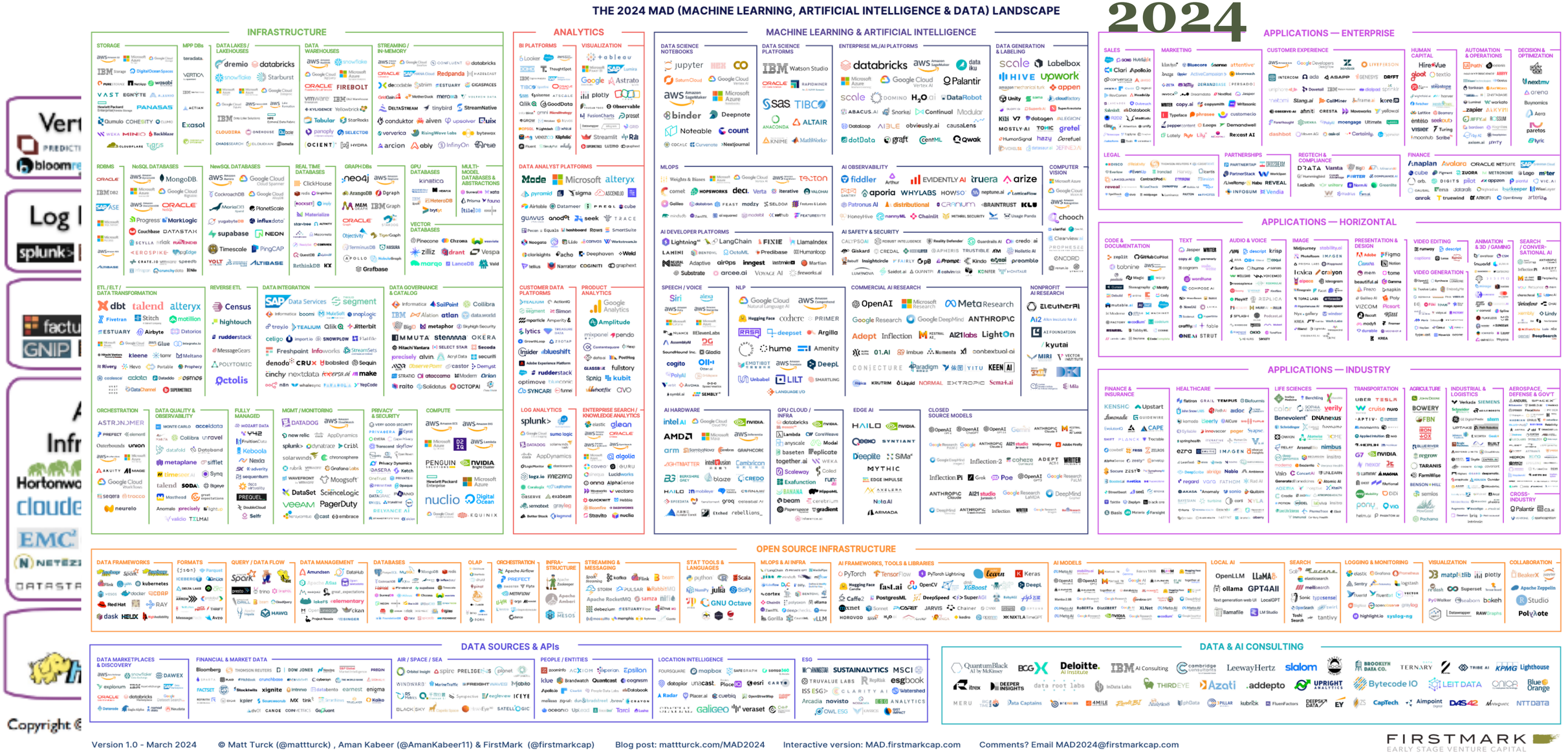
Figure 1: High-level architecture of a generic end-to-end machine learning pipeline. Logos represent a sample of tools used to construct components of the pipeline, illustrating heterogeneity in the tool stack. *Shankar et al. 2021*

Source

Big Data Analytics = ML + DM



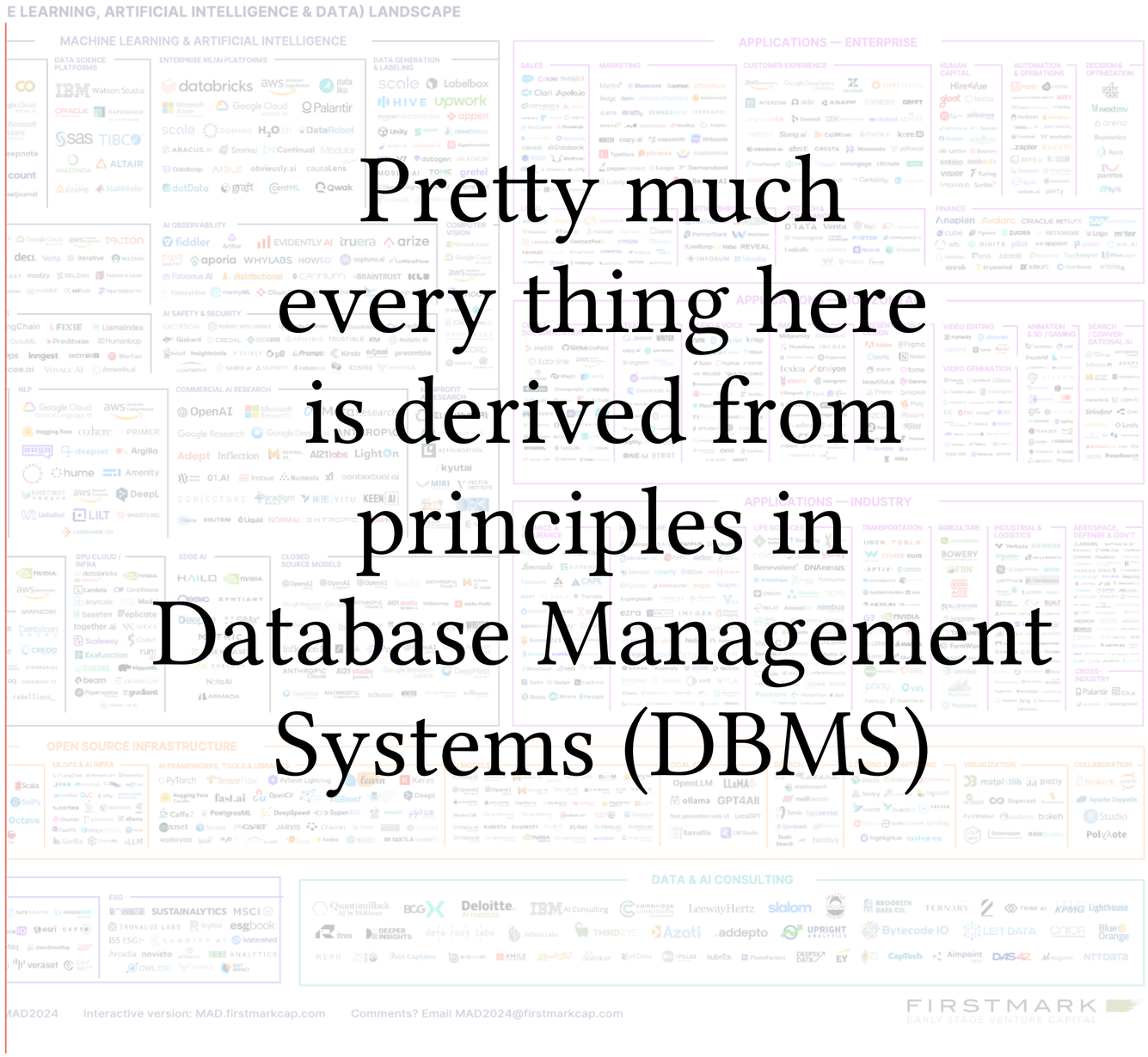
Big Data and AI Landscape Growth



Version 1.0 - March 2024 © Matt Turck (@mattturck) , Aman Kabeer (@AmanKabeer11) & FirstMark (@firstmarkcap) Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK EARLY STAGE VENTURE CAPITAL

DBMS roots



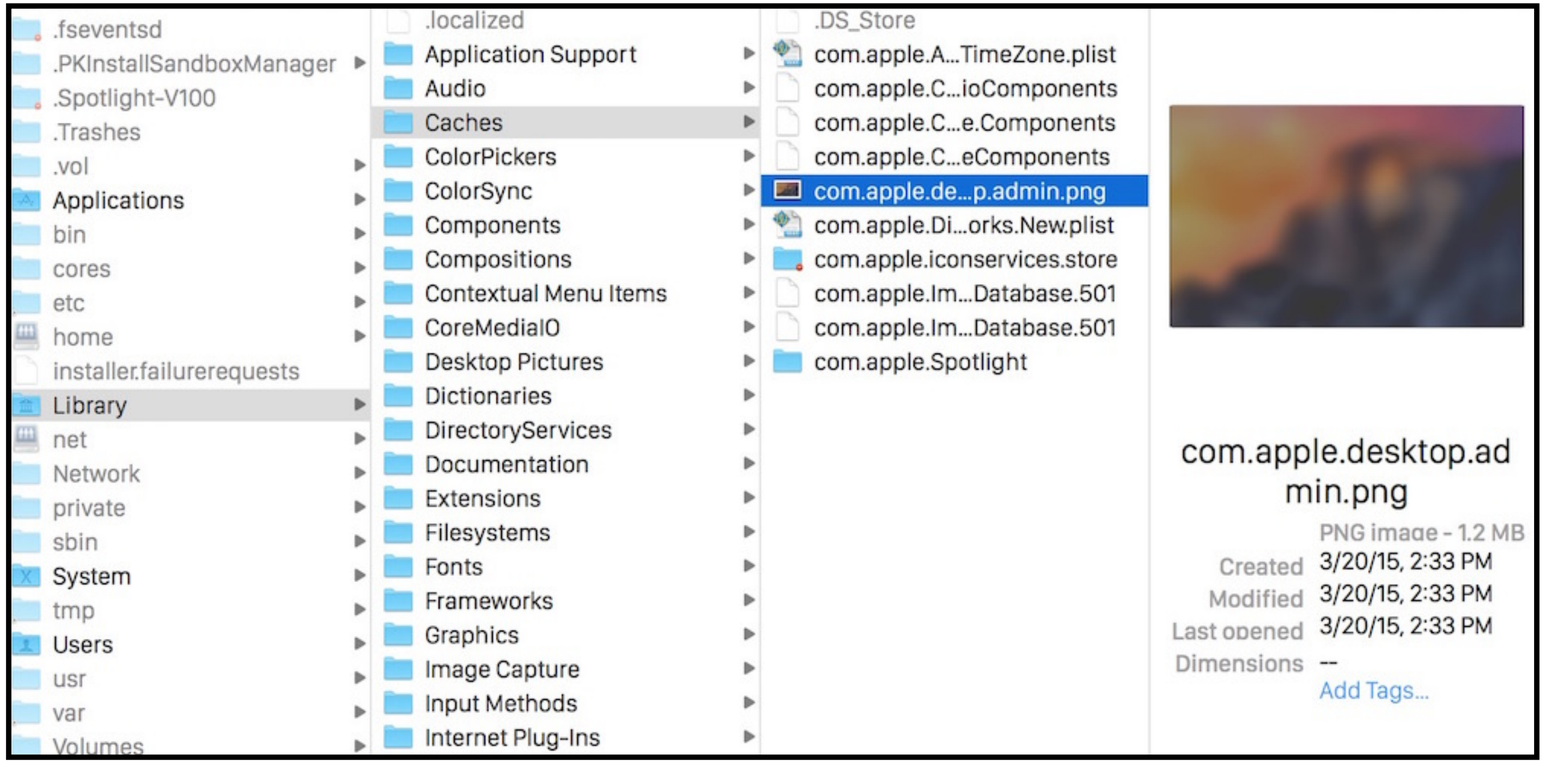
Pretty much every thing here is derived from principles in Database Management Systems (DBMS)

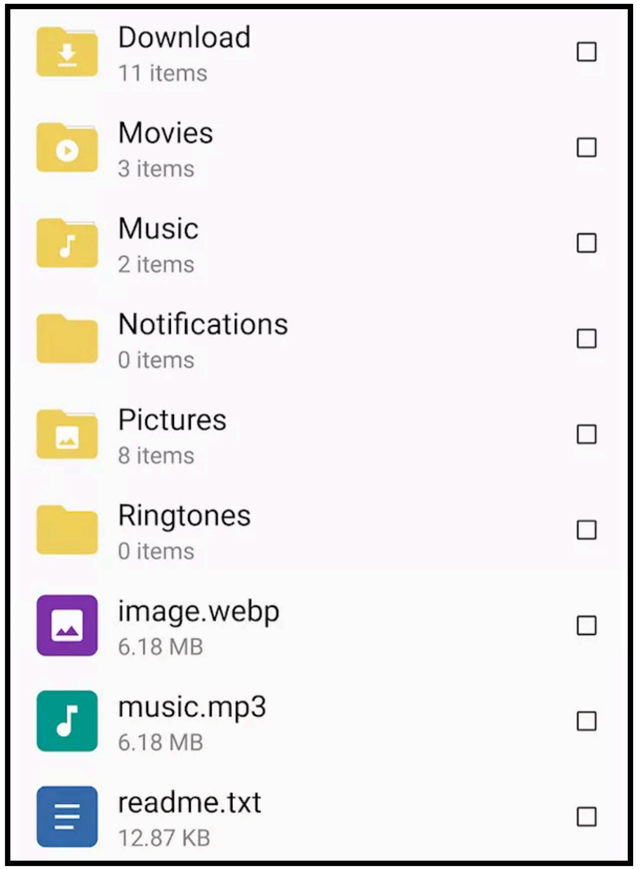
Today's Agenda

- Why DBMS?
- Course Introduction
- Course Logistics and Administrivia

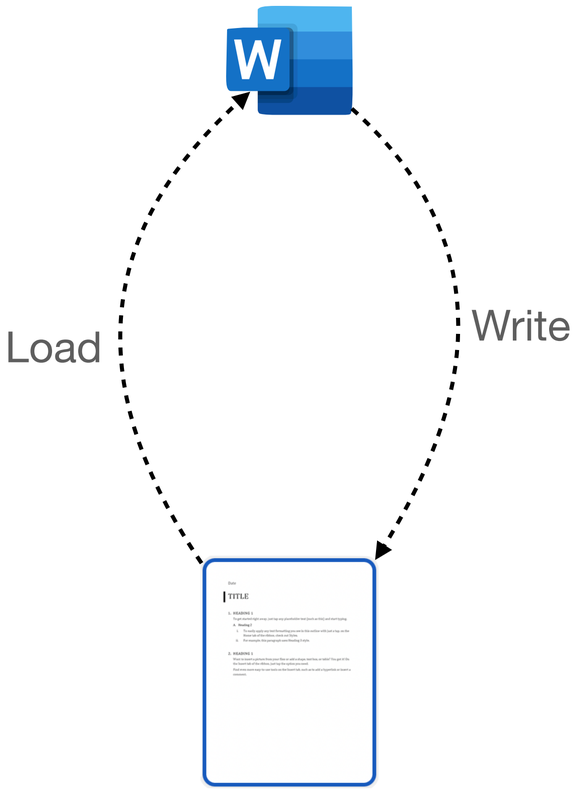
Why Database Management Systems?

- Why can't we store all data in "simple files"?





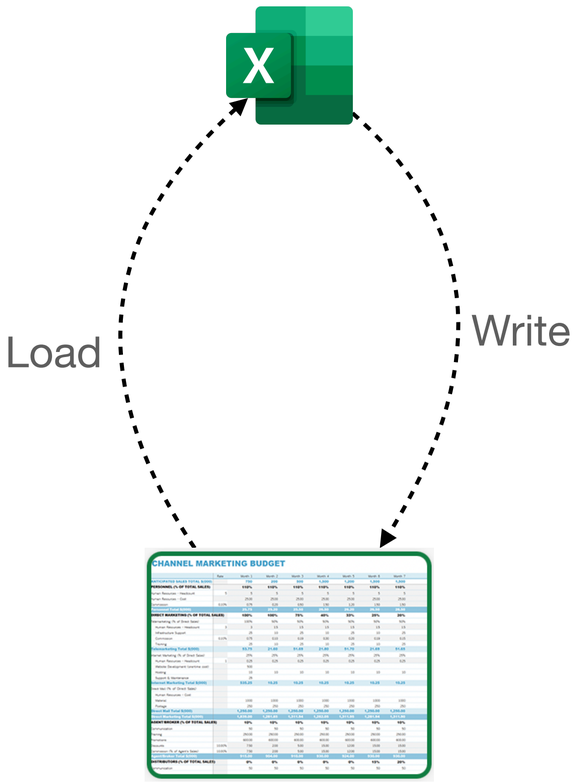
Why Database Management Systems?



Application

File System

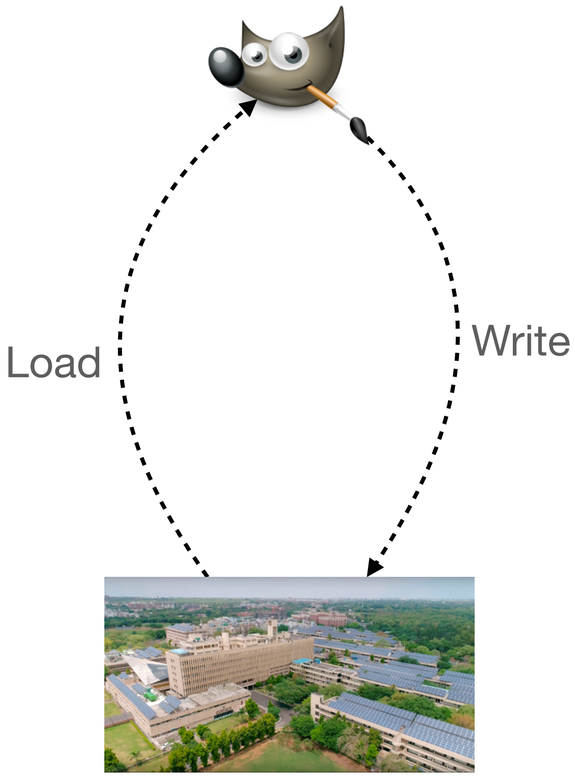
Why Database Management Systems?



Application

File System

Why Database Management Systems?



Application

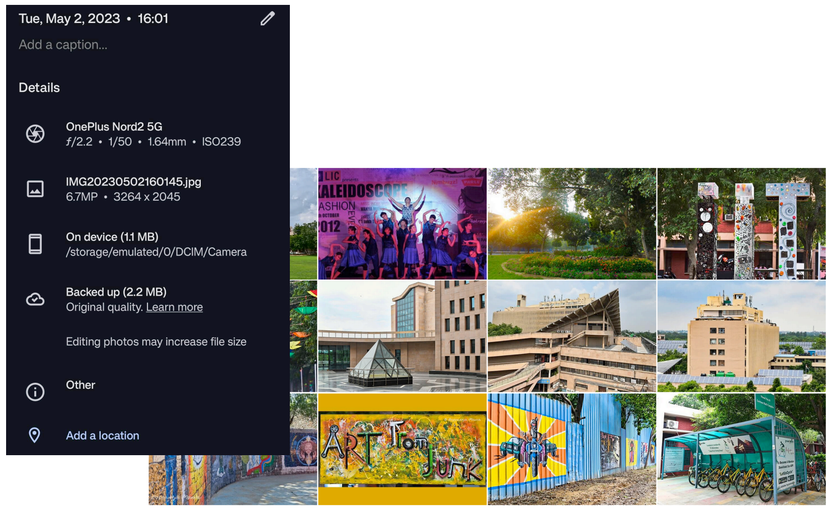
File System

What about finding data in one or more files?

Why Database Management Systems?

IITD Photo Manager

Application



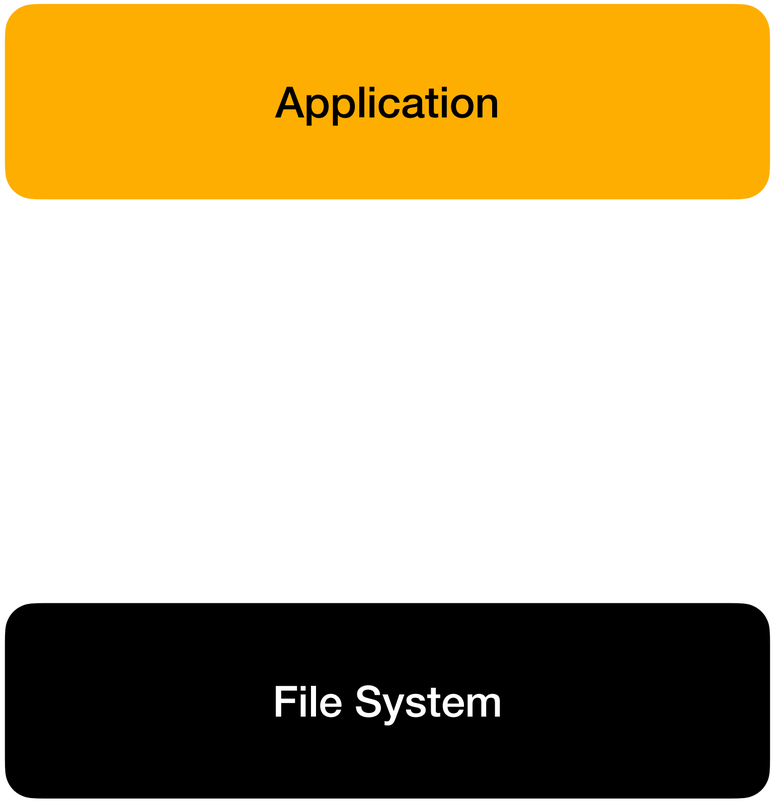
File System

Why Database Management Systems?

- Find all images with f/2.2 and ISO > 250

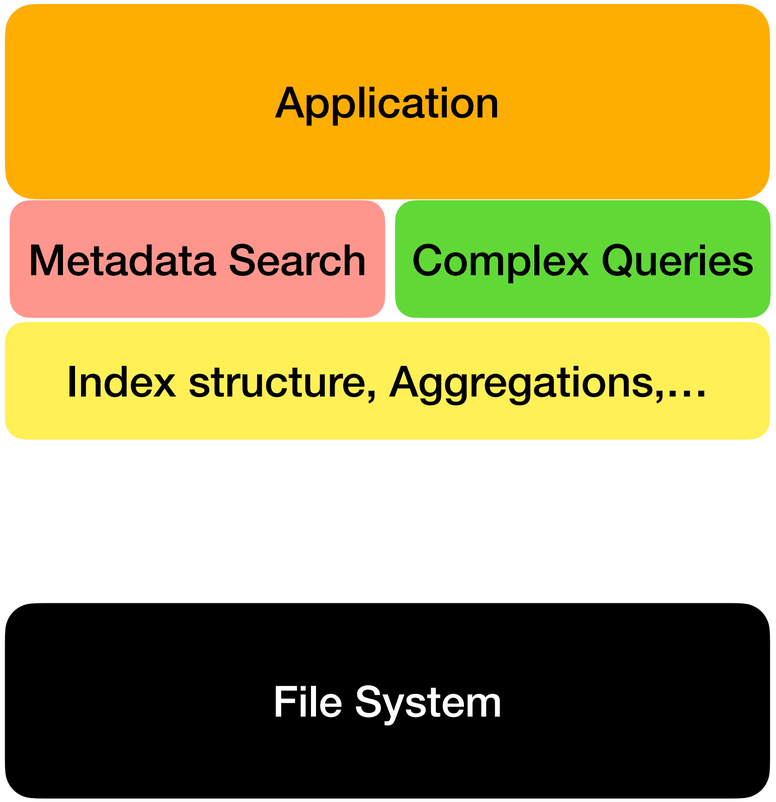
- Find all images of the main building taken by Bappaditya Manna

- Total number of images with main building



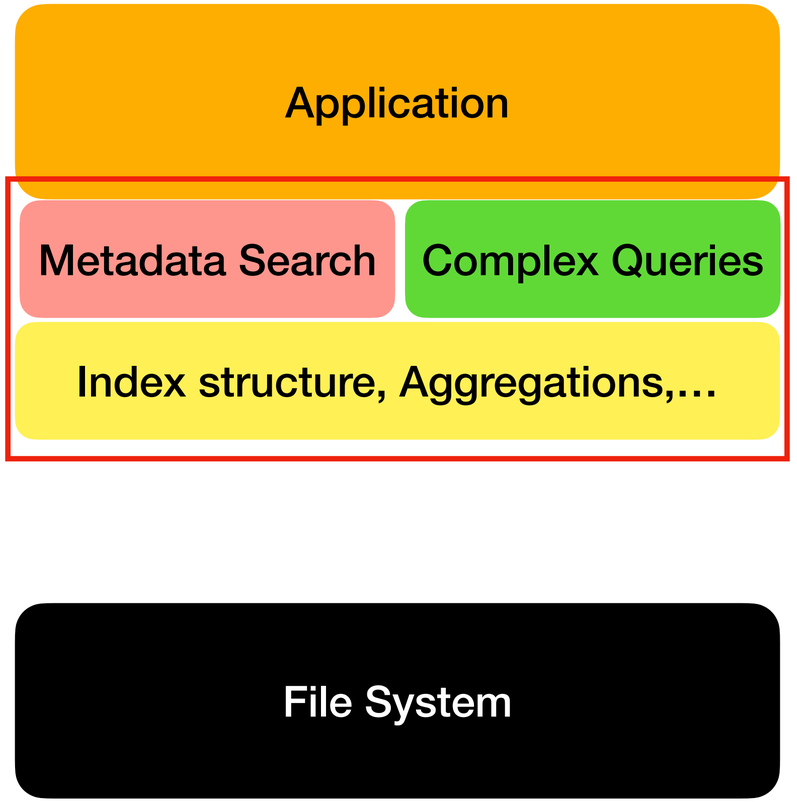
Why Database Management Systems?

- Find all images with f/2.2 and ISO > 250
- Find all images of the main building taken by Bappaditya Manna
- Total number of images with main building



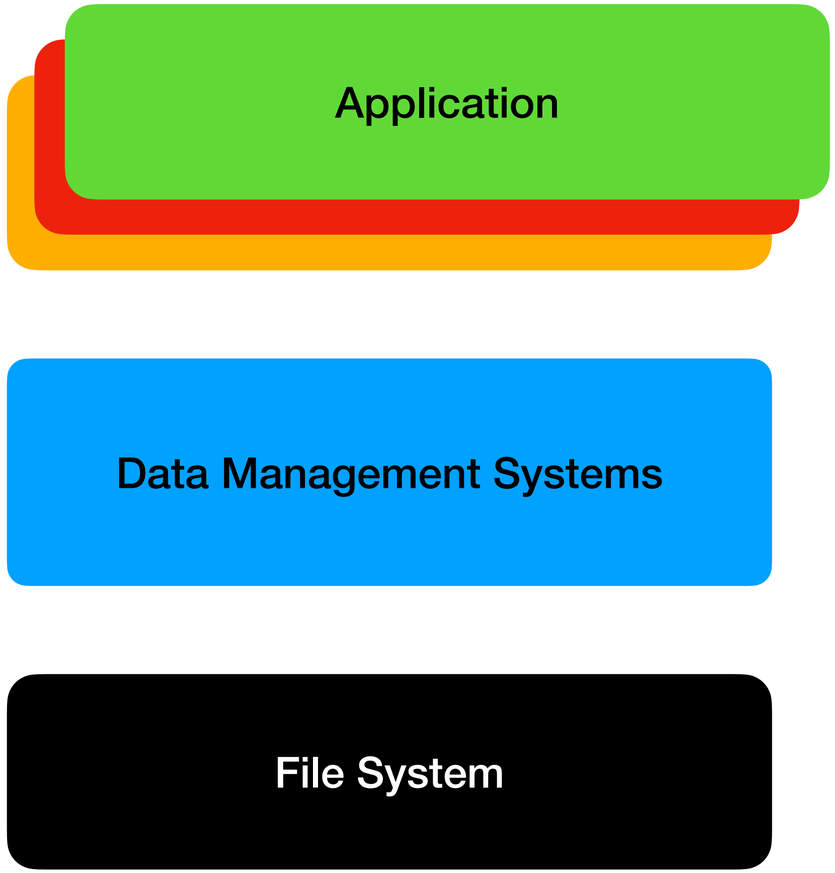
Why Database Management Systems?

This functionality is general to lot of applications!

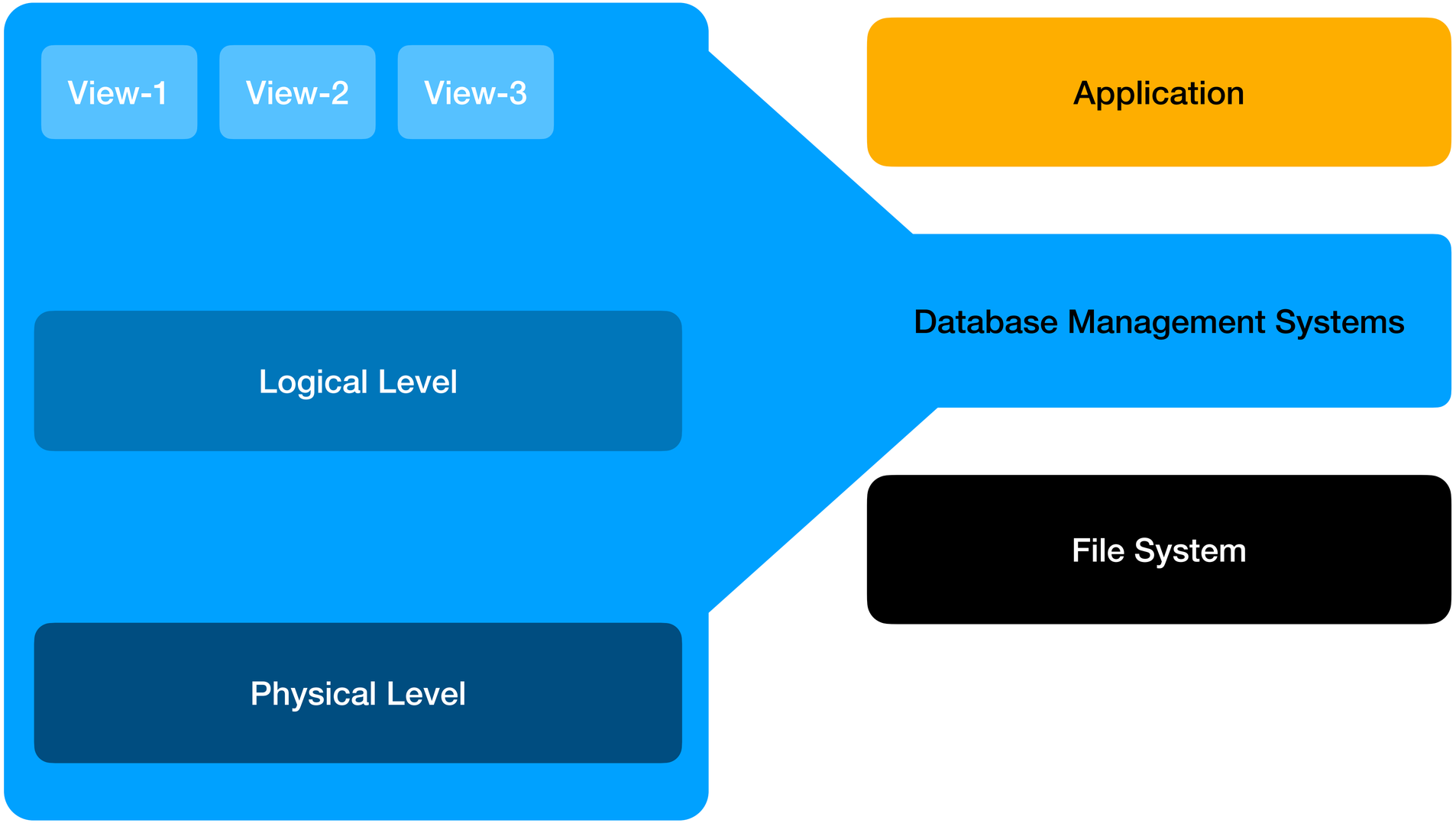


Database Management Systems

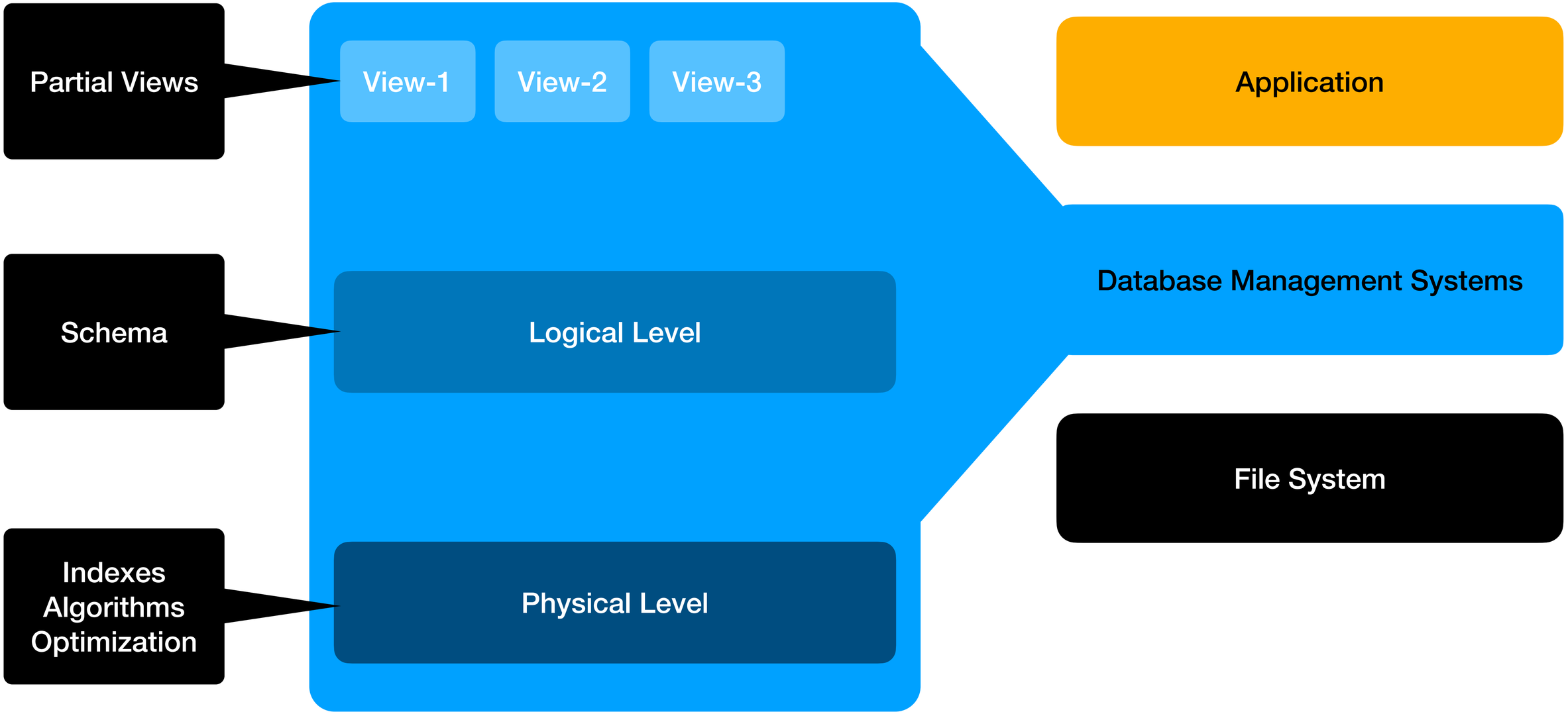
- Large amounts of data, persisting “forever”
  - Think database, think disk
- Physical and logical independence
  - Declarative languages for data manipulation
- Operations on data
  - Creating a database
  - Insert, delete, modify, retrieve data
- Guarantees



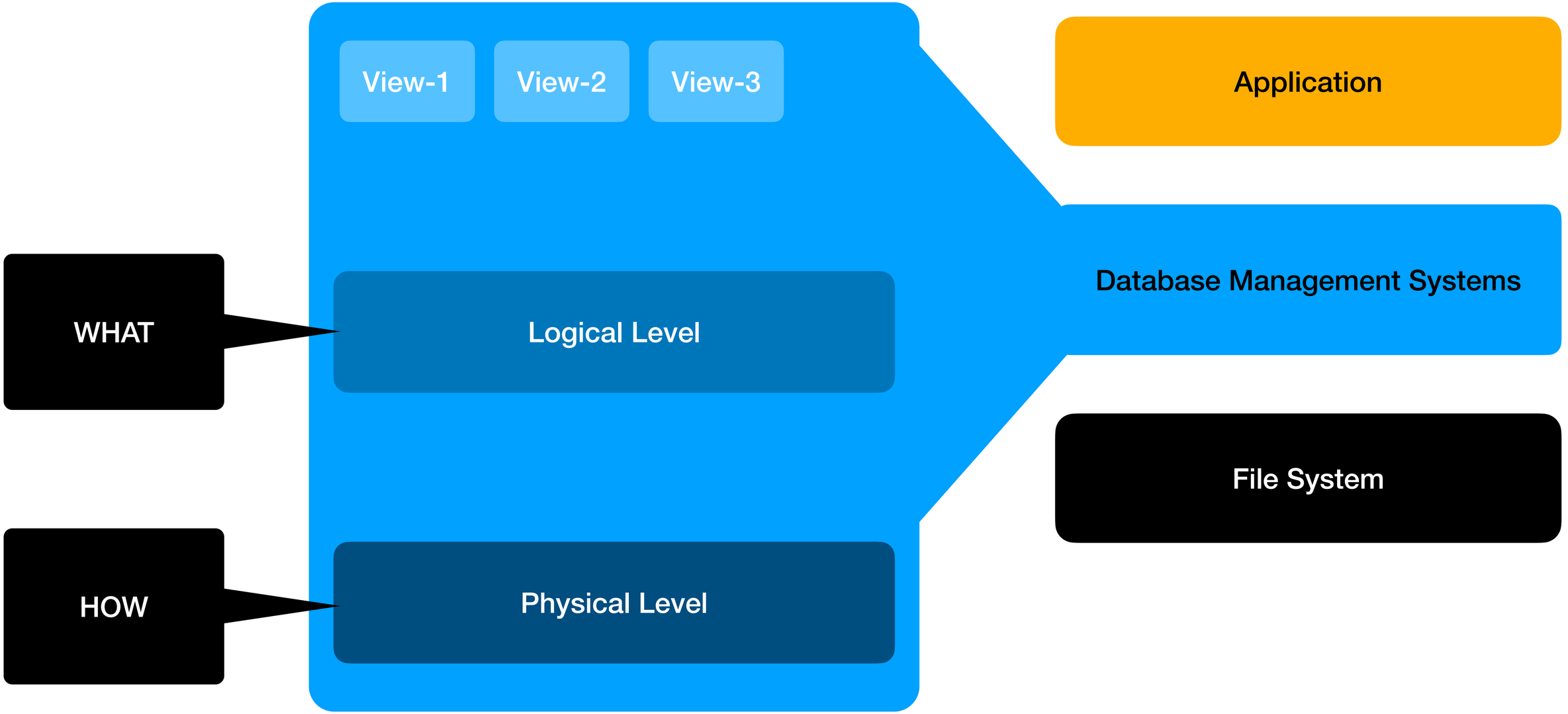
Physical and Logical Independence



Physical and Logical Independence



Physical and Logical Independence



Syllabus - I

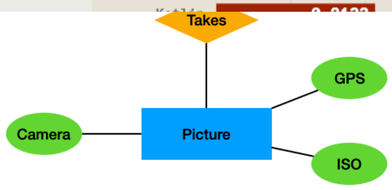
Database design and programming

- Data Modelling
  - ER Model
  - Relational Model
  - Schema design
- Relational Algebra
  - Relations
  - Basic/Adv. Operators
- SQL
  - Basic
  - Advanced

From RealWorld to Database Schema

ase Schema

Picture			
Camera	GPS	ISO	Apperture

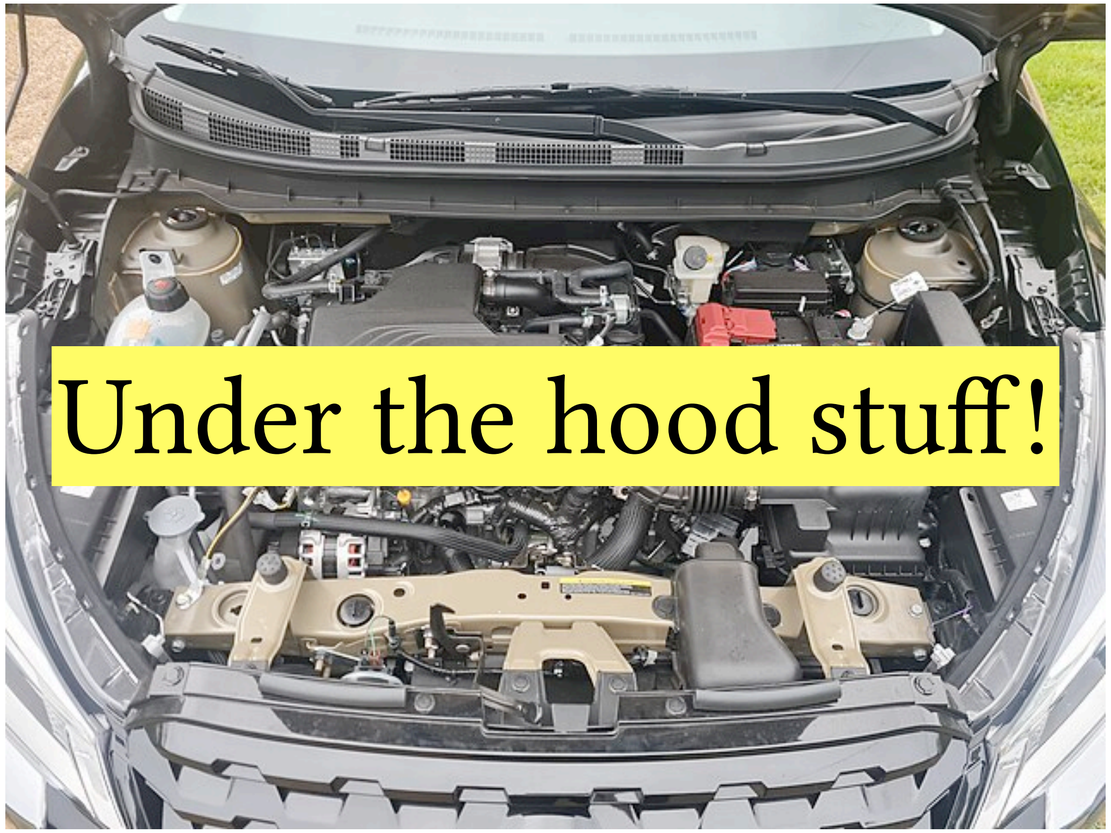


Photographer = {[Name: String, Age: Integer, ID: Integer]}
Picture = {[Camera: String, GPS: Point]}

Syllabus-II

Database Implementation and Internals

- Storage
  - Memory hierarchy
  - Data layouts
- Indexing
  - Tree-based indexes
  - Hash-based indexes
  - Bitmaps, bloom filters, etc.
- Query Processing Algorithms
  - Scans, Joins, Grouping/aggregation, Sorting
- Query Planning and Optimization
  - Join orders, plan enumeration, equivalences
  - Cost estimation, cardinalities
  - DP, interesting orders
  - Execution models (iterator/pipelining, materialised, vectorized)
- Transactions

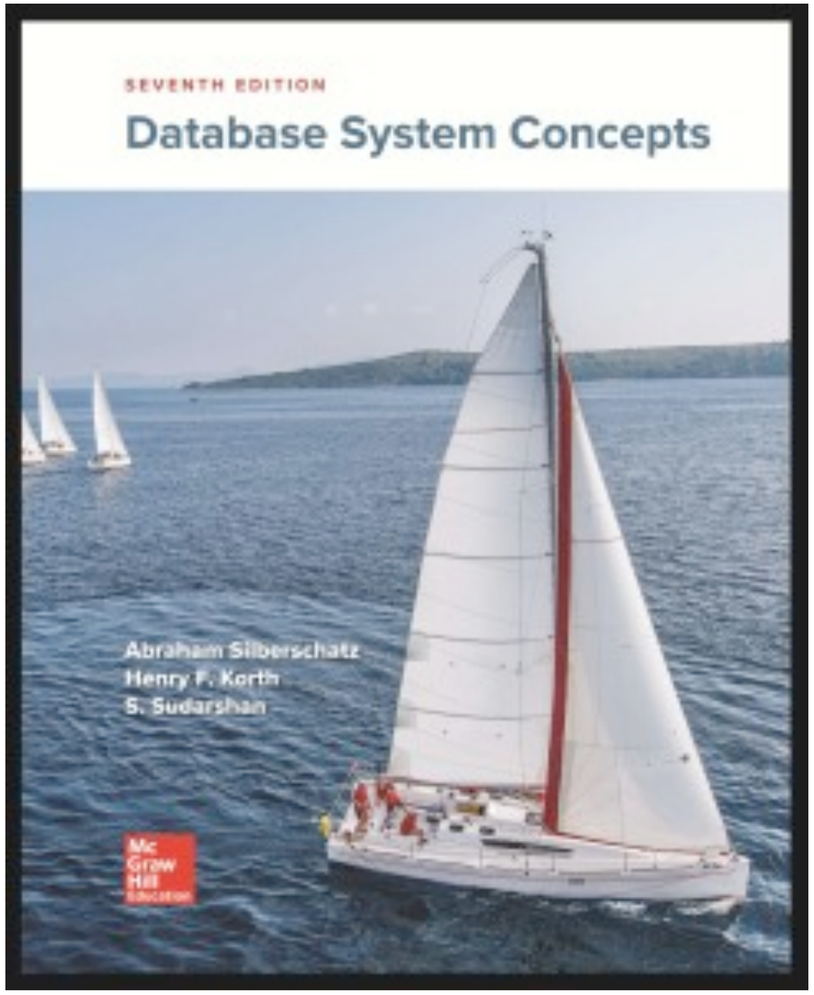


Syllabus-III

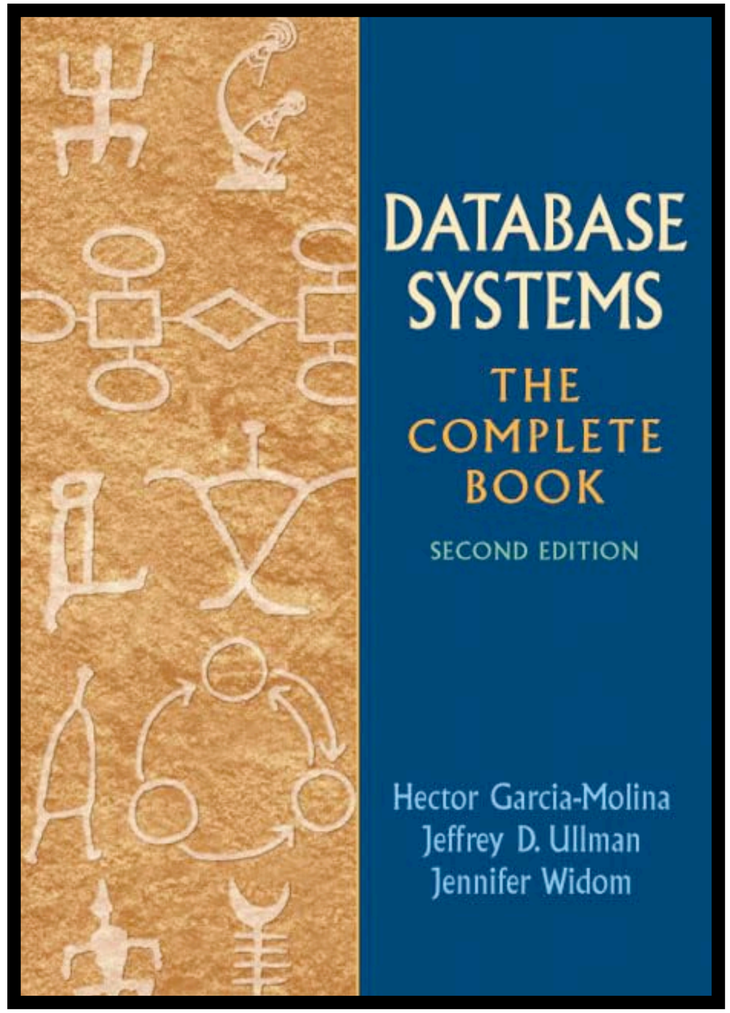
Misc. Topics (depending on time)

- Parallel and distributed databases
- Big Data technologies
- Graph Databases
- Vector Databases

Recommended Textbooks



Database System Concepts (7 ed.) by Silberschatz, Korth and Sudarshan, McGraw-Hill



Database Systems: The Complete Book by Garcia-Molina, Ullman, Widom, PrenticeHall

Today's Agenda

- Why should you take this course
- Introduction to DBMS
- Course Logistics and Administrivia

Teaching Assistants

- Abhinav Barnawal
- Rajat Bhardwaj
- Sadiya Ajaz Churoo
- Nilanjan Ghosh
- Rahul Kumar
- Manshi Sagar

Course Logistics and Administrivia

Prerequisites

- COL 106
- Programming in Java/C++

Lectures

- Tuesdays, Wednesdays, Fridays
- 10:00 — 11:00
- LH 308

Information

- Course Website (https://web.iitd.ac.in/~kbeedkar/teaching/col362-h-25/)
- Piazza (https://piazza.com/iitd.ac.in/spring2025/col362632/home)
- Moodle

Course Evaluation

Audit Policy

- Score 50% overall
- Attempt and submit all assignments and score at least 50%.

Pass Criteria

- Score at least 30% to pass the course (D grade)

Grading

- Mid-term 25%
- End-term 25%
- Assignments and Quizzes 50%
- Class participation 4%

Course Evaluation

Late Submission Policy

- You have a total of 72 late hours (across all assignments).
- Late hours will be counted in granularity of hours. For example, if you submit an assignment 20 mins past its deadline, 1 hour will be subtracted from your late hour.
- You will get a penalty of (-20)% for each late day after you have exhausted all your late hours.

Course Evaluation

Attendance Policy

- If attendance is less than 75%, the student will be awarded one grade less than the actual grade that he/she has earned. For example, a student who has got an A grade but has attendance less than 75% will be awarded an A(-) grade.

- A student cannot get NP for an audit course if his/her attendance is less than 75%.

Attendance

- You will be assigned a dedicated seat that you must sit on!
- We'll collect your preferences in the next days and try out best to accommodate them
- We'll follow photo-based attendance

Contact

- Use piazza!
- Emails will be ignored by TAs and me
- No post without me as one of the receiver

TA harassment will be severely punished

- Sending emails
- Calling over phone
- Social pressure/intimidation tactics

Cheating & Plagiarism

- Zero tolerance towards anything even remotely resembling dishonesty regarding your assignments, classwork, exams, and attendance

- Will lead to -100% of the course component. Repeated offenders will be referred to the department and institute disciplinary committee