TECHNION | The Henry and Marilyn Taub Faculty of Computer Science

Structured ML

Course 02360605 – Advanced Topics in Computer Science 5

Introduction

Benny Kimelfeld
Technion – Israel Institute of Technology

Spring 2026

The Course

- Paradigms for ML over **relational databases**
  - To this aim, we learn ML over **tabular** and **graphs**
- Content:
  - 6 core modules
  - Tutorials: rehearsal, SQL, PyTorch, PyTorch Geo, …
    - In general, the *practical side of the modules*
  - Enrichment module on explainability
  - Invited lecturers
- Staff:
  - Lecturer: **Benny Kimelfeld**
  - TAs: **Boaz Berger**, **Daniel Elgarici**, **Ehud Gordon**
  - TA in charge: **Shunit Agmon**

Requirements

- 3 homework **assignments**
  - 10%, 15%, 25% of the grade
  - Each followed by:
    - A “class report” (assuming in-person, attendance required)
    - Or a personal interview (assuming online)
- Final **project**
  - 50% of the grade
  - Oral defense of the project
- In **pairs**
- Final project + all assignments **mandatory**

Assumed Background

- Basic course on **machine learning**
  - Not necessarily *deep learning*
- **Databases** – an advantage
  - Not mandatory, mainly simple concepts used here, can be learned throughout the course

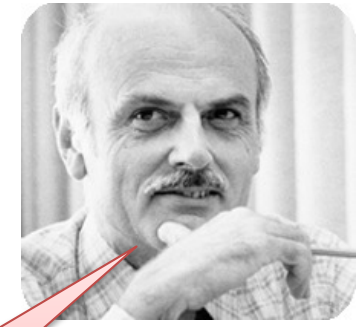
Structured ML

Context: Relational Databases

1970's Debate

General-purpose DB to be abstracted as a ***network***

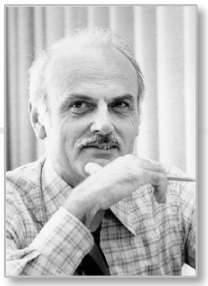
C. W. Bachman (1924-2017)



Edgar F. Codd (1923-2003)

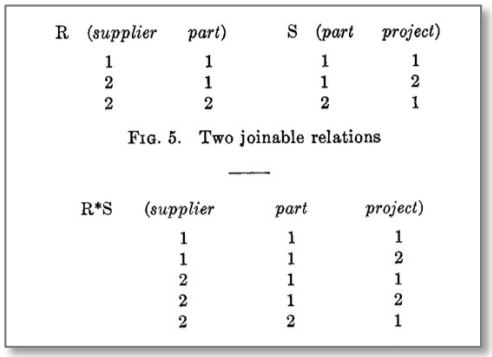
As logic!

Codd's Vision



Edgar F. Codd

- 1970: Codd invents the **relational database** model
  - Idea: interface via mathematical logic
    - Data = collection of relations, interconnected via keys
    - Questions via a query language over the schema
    - System translates queries into actual execution plans
  - Principle: separate logical from physical layers
  - Work done at IBM in San Jose, California
    - [E. F. Codd: *A Relational Model of Data for Large Shared Data Banks.* In Communications of the ACM 13(6): 377-387 (1970)]
    - [E. F. Codd: *Relational Completeness of Data Base Sublanguages.* IBM Research Report RJ987 (1972)]



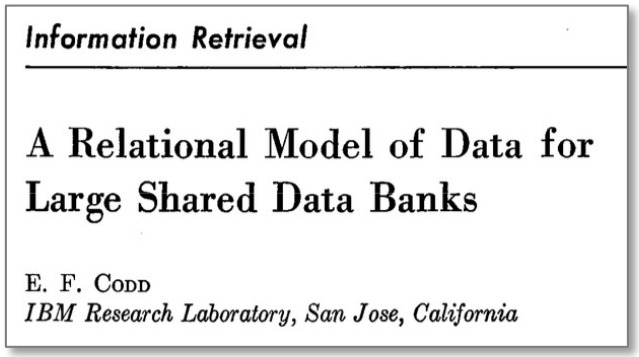
R (*supplier part*) S (*part project*)

1	1	1	1
2	1	1	2
2	2	2	1

FIG. 5. Two joinable relations

R*S (*supplier part project*)

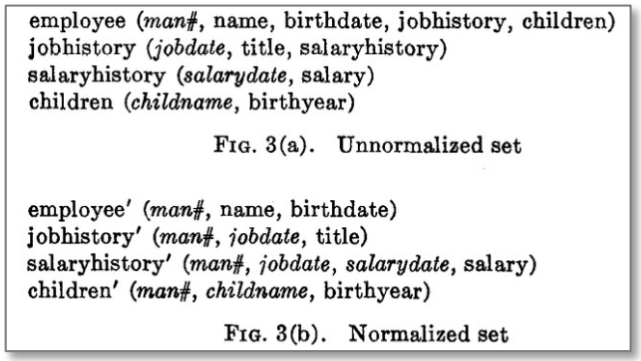
1	1	1
1	1	2
2	1	1
2	1	2
2	2	1



Information Retrieval

A Relational Model of Data for Large Shared Data Banks

E. F. CODD
IBM Research Laboratory, San Jose, California



employee (*man#*, name, birthdate, jobhistory, children)
jobhistory (*jobdate*, title, salaryhistory)
salaryhistory (*salarydate*, salary)
children (*childname*, birthyear)

FIG. 3(a). Unnormalized set

employee′ (*man#*, name, birthdate)
jobhistory′ (*man#*, *jobdate*, title)
salaryhistory′ (*man#*, *jobdate*, *salarydate*, salary)
children′ (*man#*, *childname*, birthyear)

FIG. 3(b). Normalized set

Codd Catches On (1)

- 1973: Michael Stonebraker and Eugene Wong implement Codd's vision in **INGRES**
  - Commercialized in 1983
  - Evolved to **Postgres** (now PostgreSQL) in 1989

M. Stonebraker

INGRES

E. Wong

PostgreSQL

Codd Catches On (2)

- 1974: A group from the IBM San Jose lab implements Codd's vision in **System R**, which evolved to **DB2** in 1983
  - SQL initially developed at IBM by Donald D. Chamberlin and Raymond F. Boyce
    - [Chamberlin, Boyce: SEQUEL: A Structured English Query Language. SIGMOD Workshop, Vol. 1 1974: 249-264]
- 1977: Influenced by Codd, Larry Ellison founded Software Development Labs
  - Becomes *Relational Software* in 1979
  - Becomes *Oracle* Systems Corp (1982), named after its **Oracle** database product


R. F. Boyce
(1947-1974)

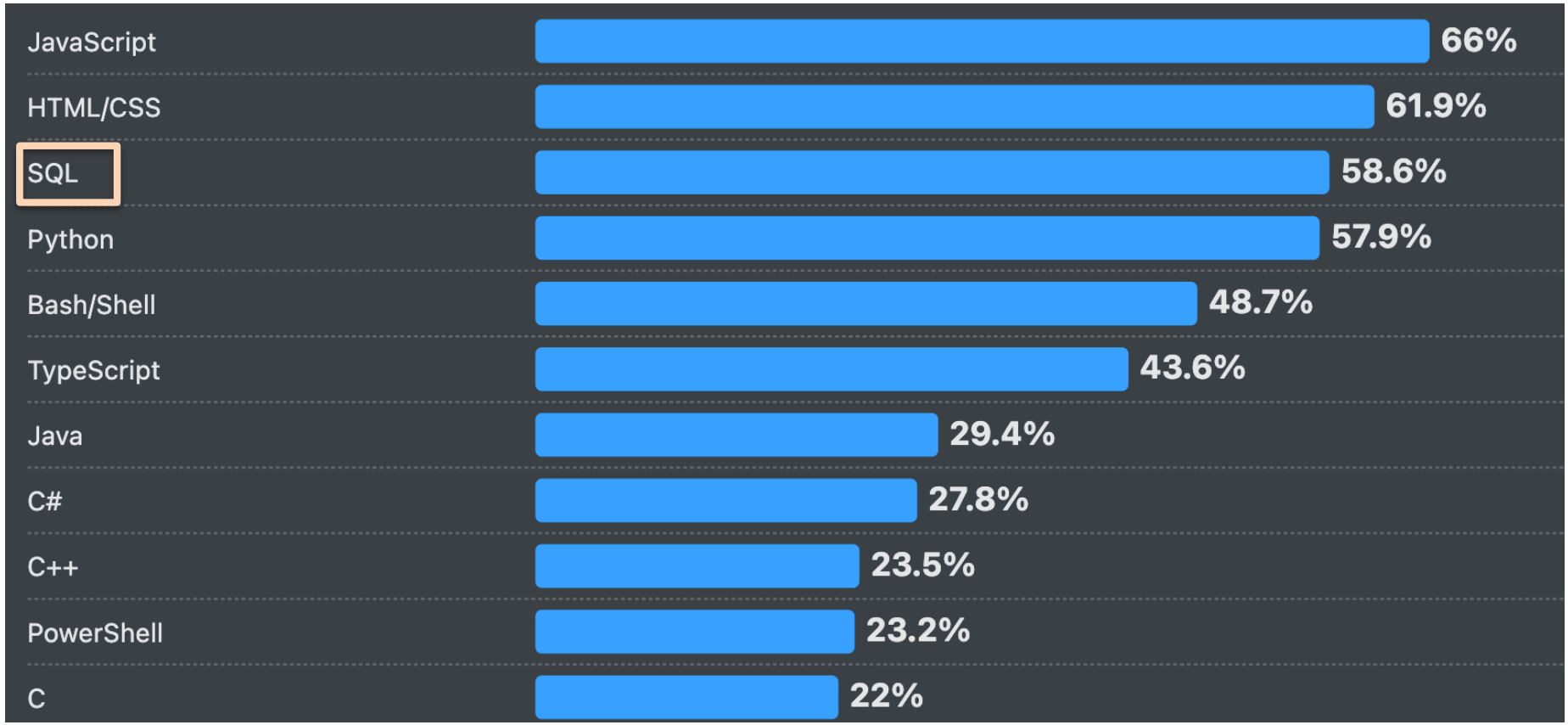

D. D.
Chamberlin


J. Grey


L. Ellison


P. G. Selinger

DB Popularity in Programming

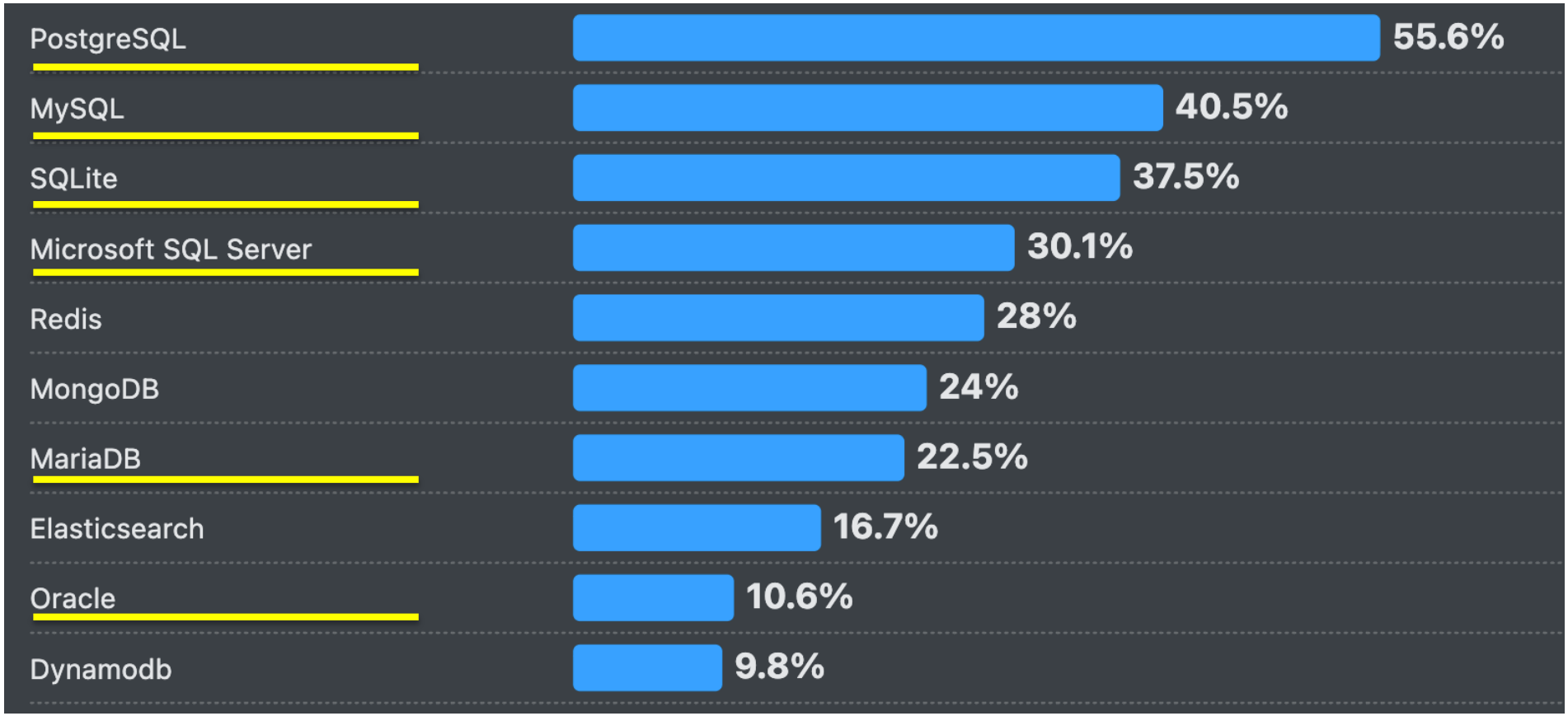


31,771 responses

Survey on Programming, Scripting, and Markup Languages

https://survey.stackoverflow.co/2025

Data is Stored in Relations



26,083 responses

Databases

https://survey.stackoverflow.co/2025

Structured ML

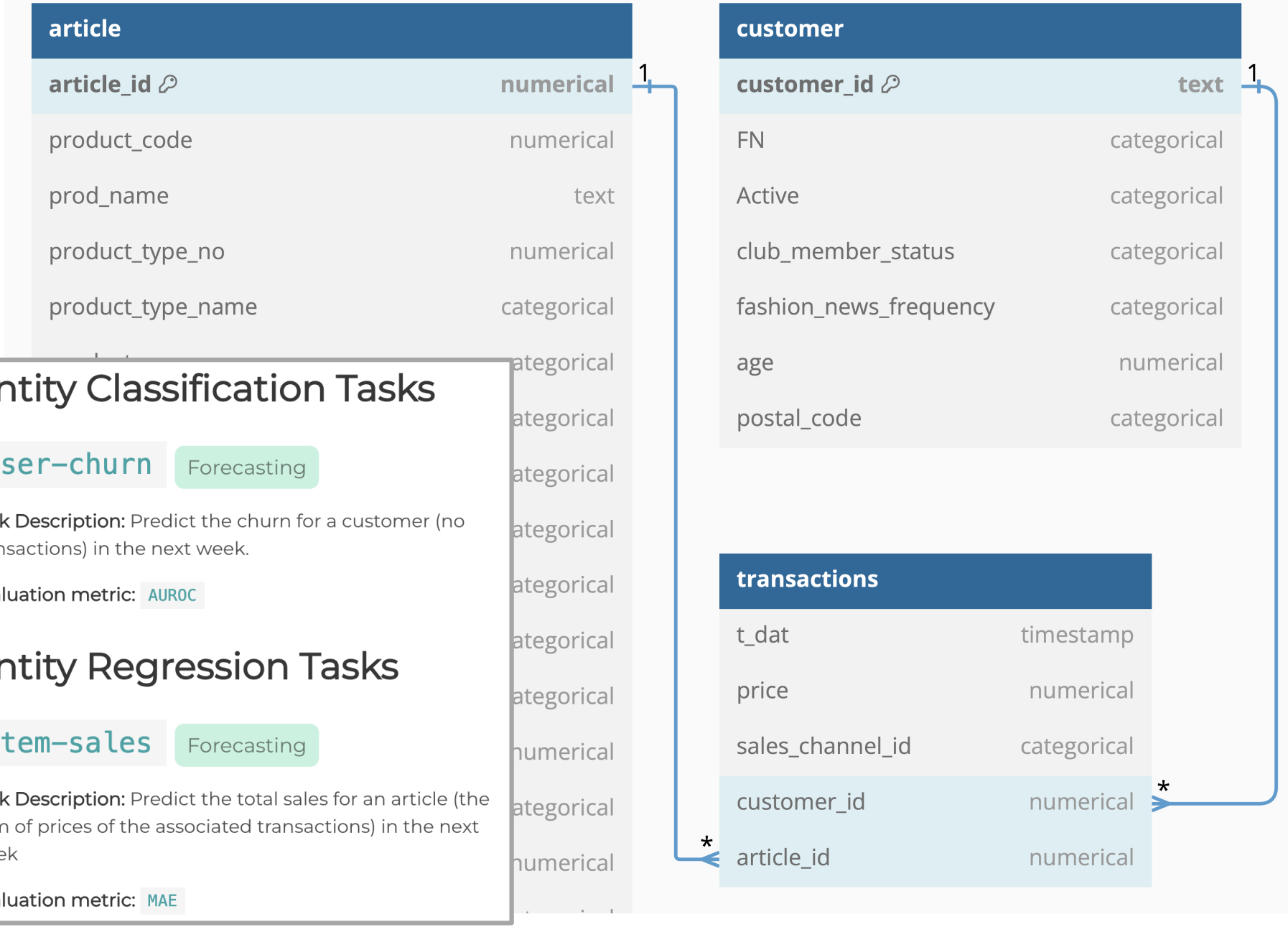
ML over Relational Databases

Databases: QL vs. ML

- Query languages answer questions that we can translate into **exact** algorithms over the database
  - Designed for simple, high-level, declarative specification of information needs, without dealing with computation
- ML is needed whenever our question involves prediction/uncertainty
  - We cannot phrase an exact query to answer our need
  - ... and yet, we believe that the *database information can greatly reduce the uncertainty*

Example: H&M DB

RELBENCH
RELATIONAL DEEP LEARNING BENCHMARK



Entity Classification Tasks

`user-churn` Forecasting

Task Description: Predict the churn for a customer (no transactions) in the next week.

Evaluation metric: `AUROC`

Entity Regression Tasks

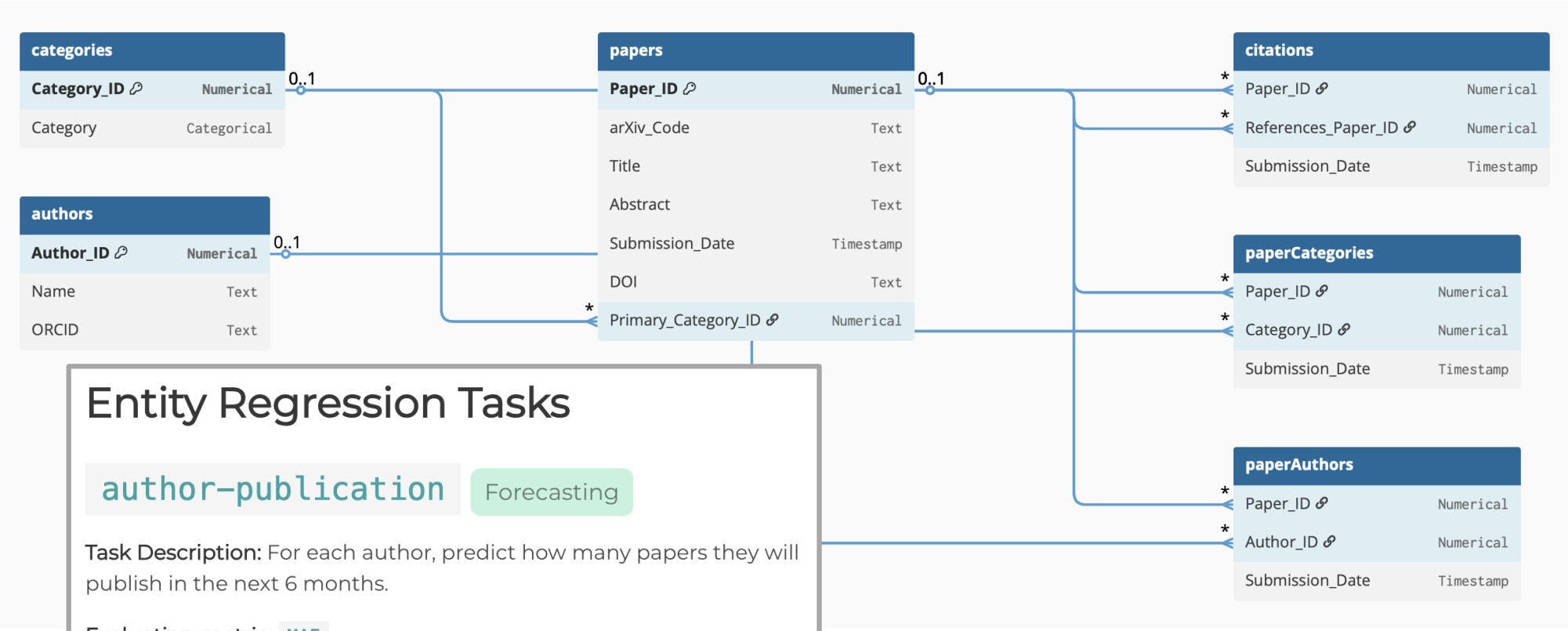
`item-sales` Forecasting

Task Description: Predict the total sales for an article (the sum of prices of the associated transactions) in the next week

Evaluation metric: `MAE`

Example: Arxiv (pub) DB

RELBENCH
RELATIONAL DEEP LEARNING BENCHMARK



Entity Regression Tasks

`author-publication` Forecasting

Task Description: For each author, predict how many papers they will publish in the next 6 months.

Evaluation metric: `MAE`

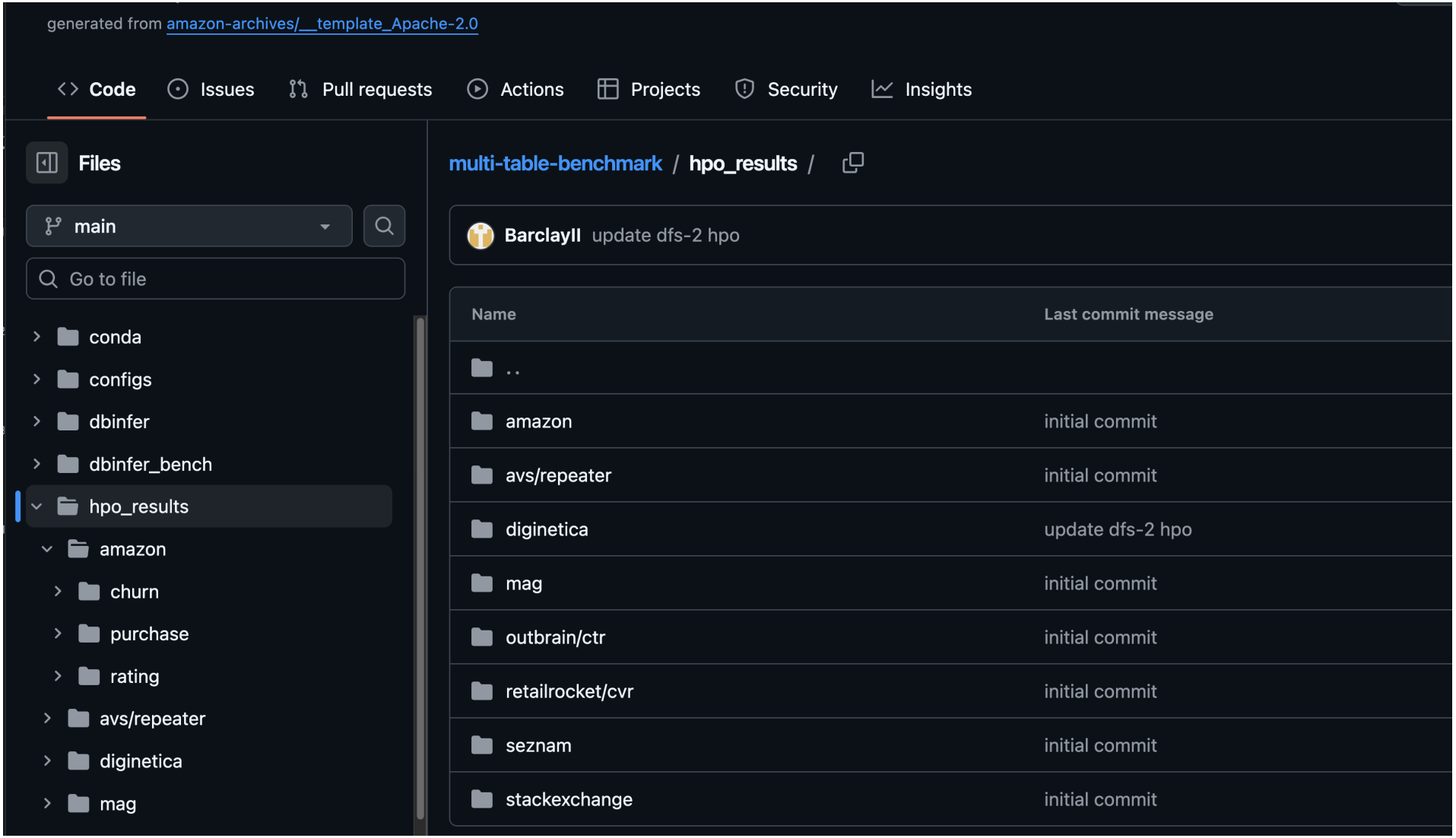
Link Prediction Tasks

`co-citation`

Task Description: For each paper, predict which other papers will be co-cited with it (i.e., cited in the same reference list) in the next 6 months.

Evaluation metric: `MAP`

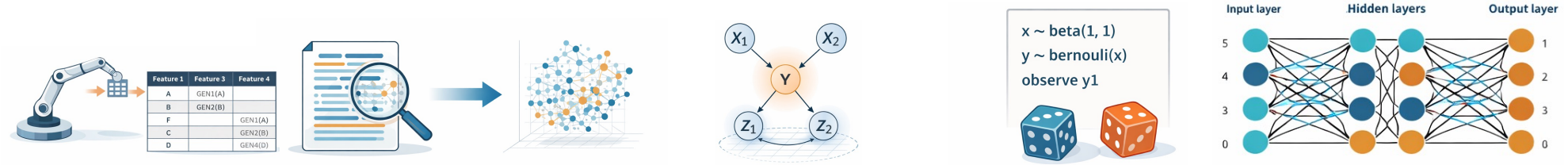
More from Amazon's Git



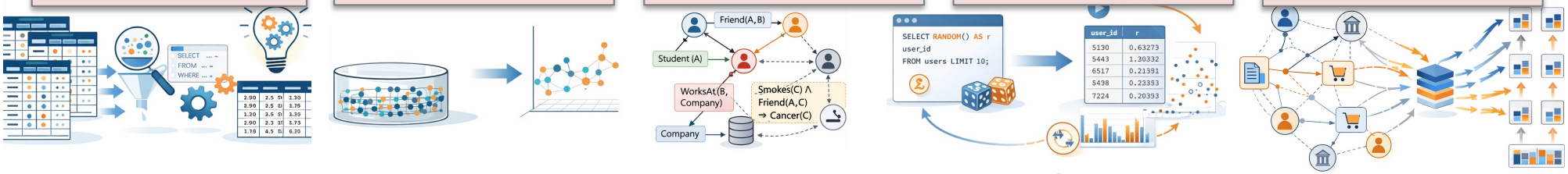
https://github.com/awslabs/multi-table-benchmark/tree/main/hpo_results

ML Paradigms: General & DB-Case

General ML World



Feature generation	Low-dim embedding	Prob graphical models	Probabilistic programming	Deep neural networks
DB feature extraction	Tuple embedding	Markov Logic Networks	Generative queries	Hetero GNNs; rel. transform
Ex: ACORA	Ex: EmbDI	Ex: DeepDive	Ex: StarfishDB	Ex: KUMO



ML in Relational DB World

Databases: Tables + Graph of Records

Student

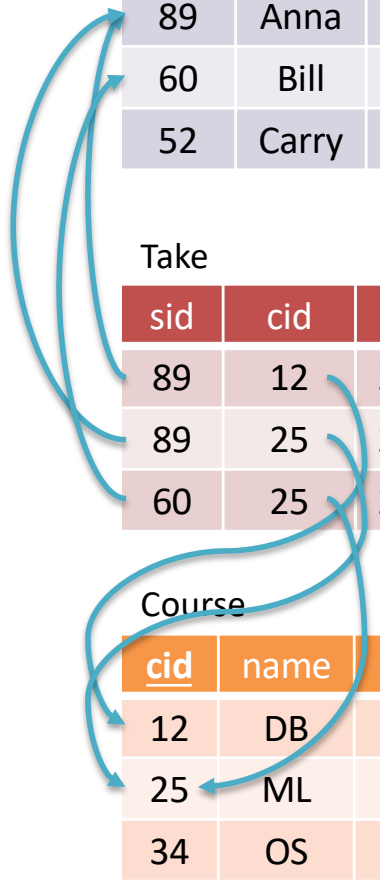
sid	name	started	address	gender	track
89	Anna	2022	Haifa	F	CS3
60	Bill	2024	Haifa	M	CS4
52	Carry	2023	Tel-Aviv	F	SE

Take

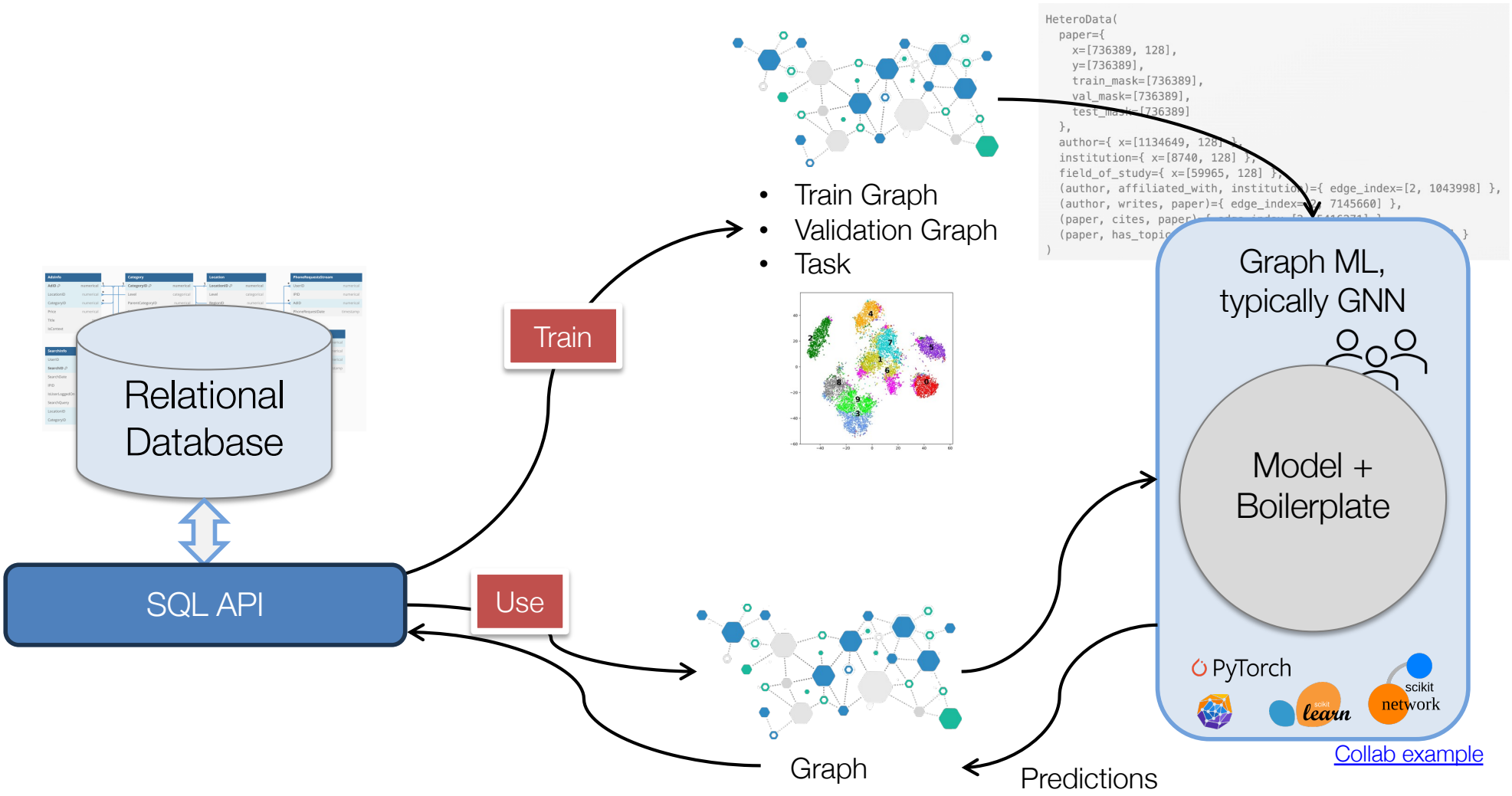
sid	cid	year	semester
89	12	2025	1
89	25	2024	2
60	25	2024	2

Course

cid	name	credits	faculty
12	DB	3	CS
25	ML	3	CS
34	OS	4	ECS



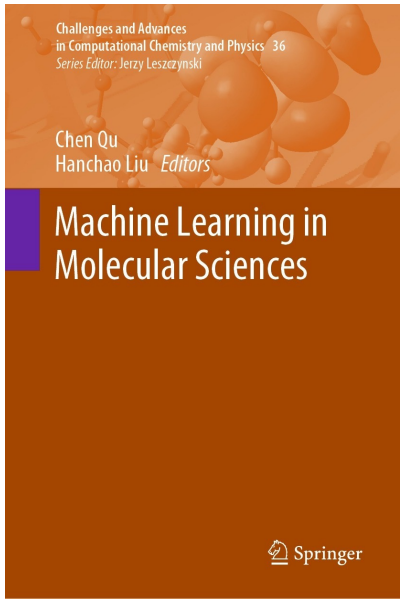
Common Workflow Nowadays

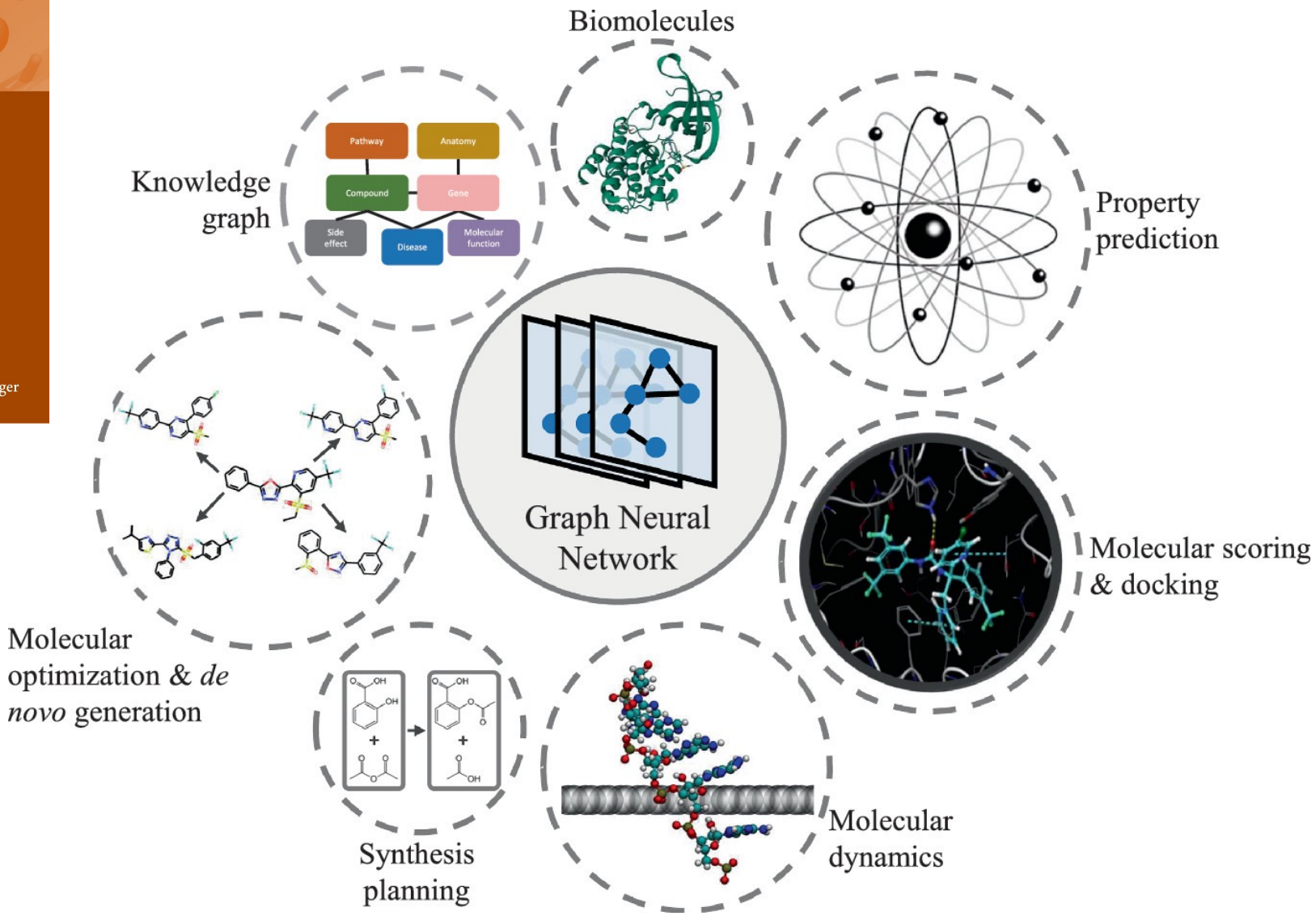


Graphs, Everywhere

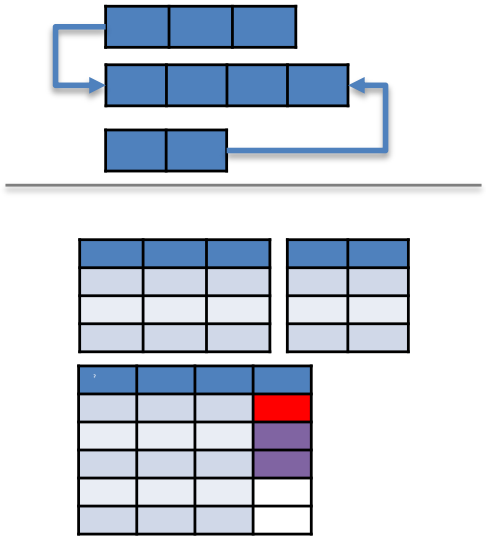
- Social networks
- Communication networks
- Molecules
- Internet pages
- Financial interactions
- Navigation and supply chains
- …

Graph Learning for Molecular Prediction

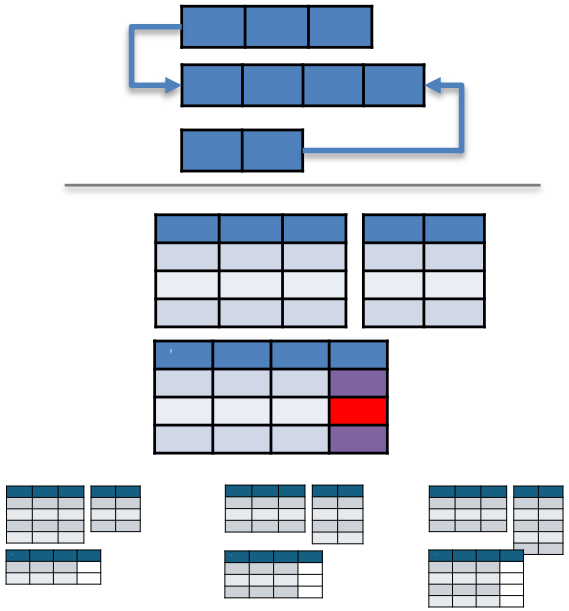




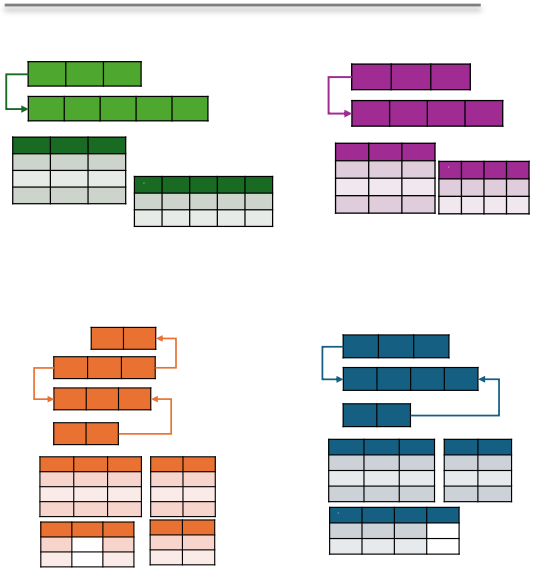
Transductive model:
Fixed DB



Inductive model:
Fixed schema & task, unknown DB



Foundation model:
Nothing known



Structured ML

Course Content

Core Modules in the Course

1. Basic **concepts**
   - Tables, Relations, Graphs, ML, Struct. ML
2. DB ML via **feature generation** & tabular ML
3. Low-dimensional **embedding** of structures
4. Overview of **deep learning**
   - Feedforward, perceptron, attention, transformers, back-propagation, GPU
5. DB ML via graph neural networks (**GNNs**)
6. Relational **foundation models**

Additional Treats (as time permits)

- Module on explainability of ML models over structured data
  - TA **Boaz Berger**
- Invited lectures
  - **Avi Mendelson** on foundation models
  - **Haggai Maron** on graph ML

Good Luck!