# Database Design and Implementation

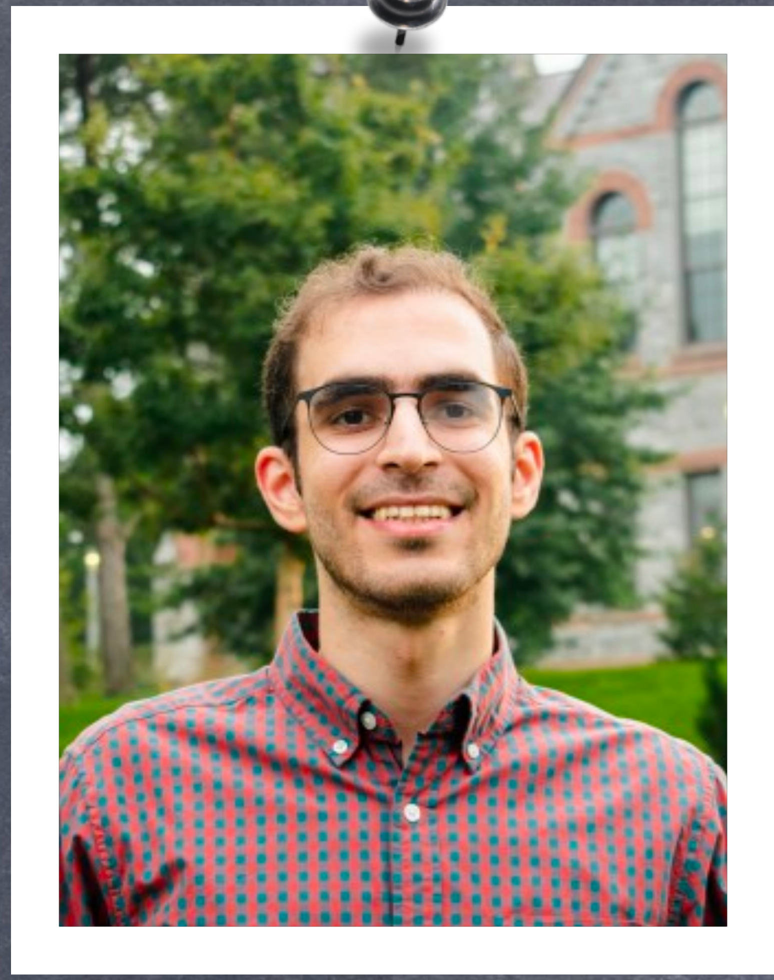
CS 645

Course Overview

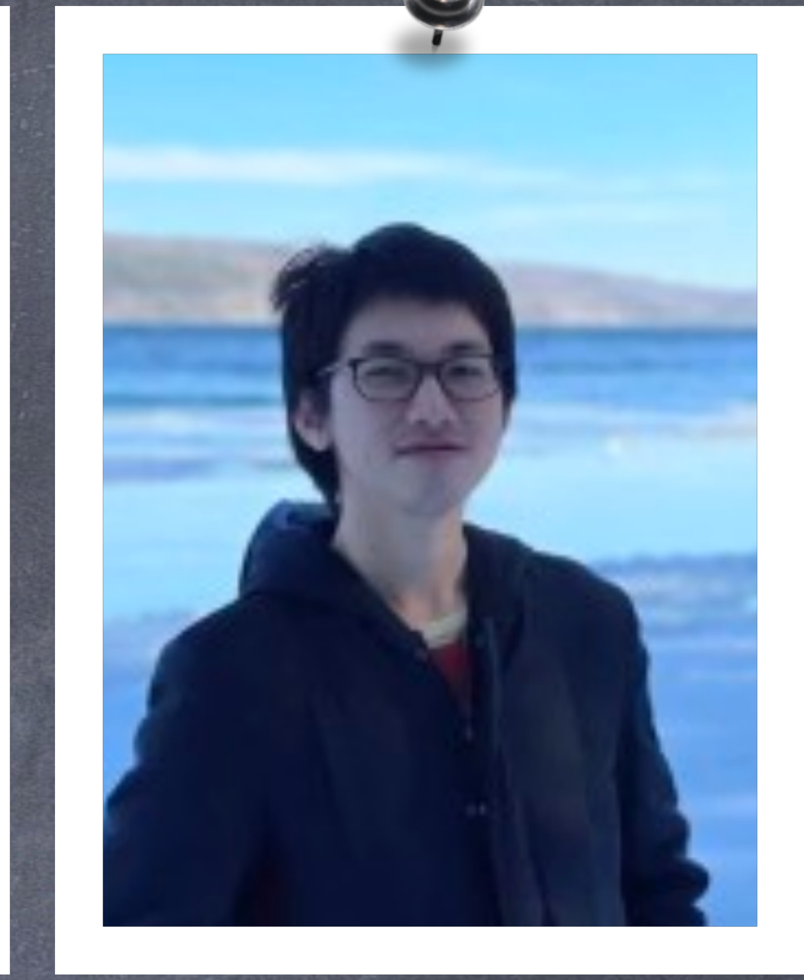
## Instructor

Alexandra Meliou

## Teaching assistants



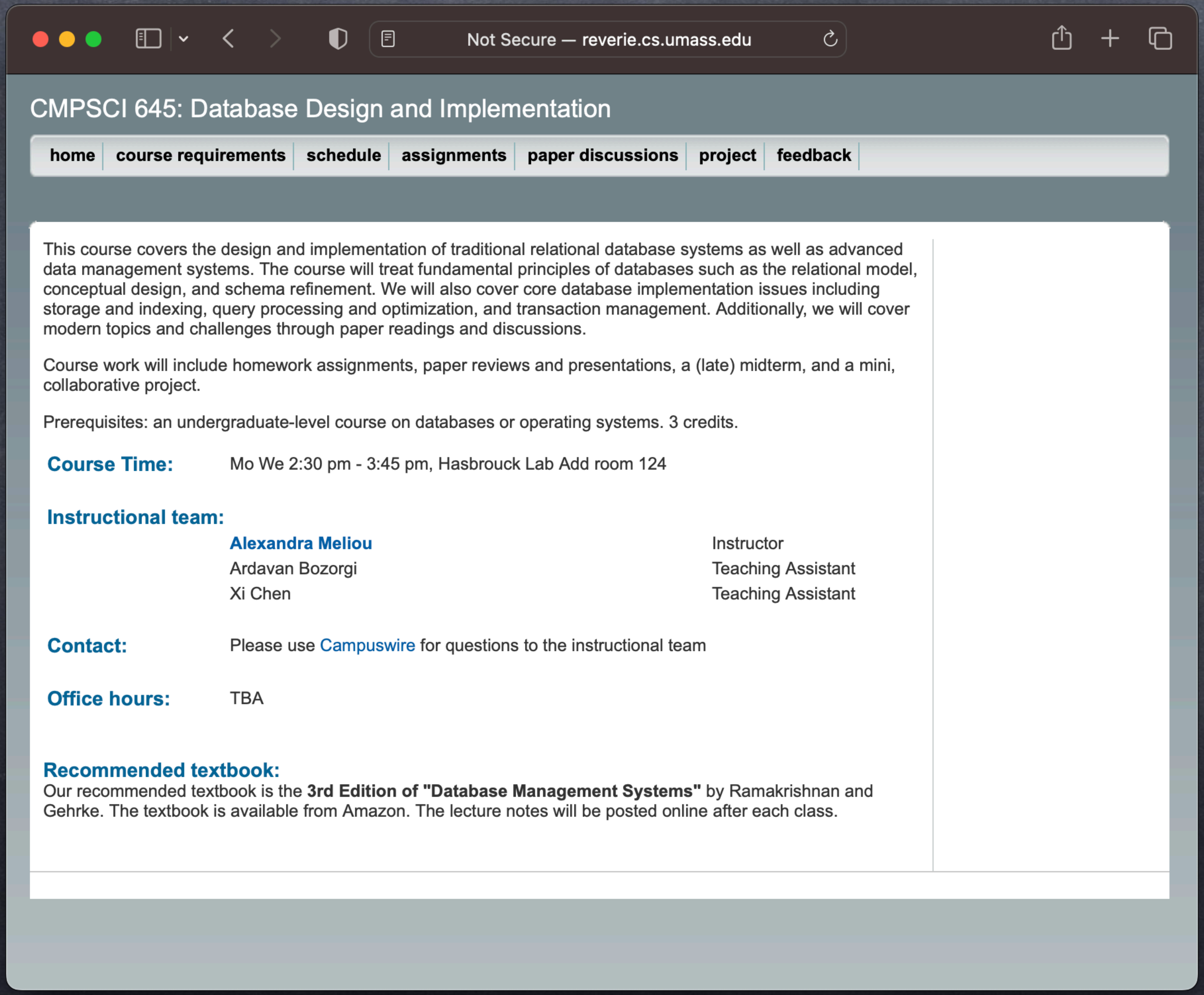
Ardavan Bozorgi



Xi Chen

# http://reverie.cs.umass.edu/courses/645

no https!


Not Secure — reverie.cs.umass.edu

CMPSCI 645: Database Design and Implementation

home | course requirements | schedule | assignments | paper discussions | project | feedback

This course covers the design and implementation of traditional relational database systems as well as advanced data management systems. The course will treat fundamental principles of databases such as the relational model, conceptual design, and schema refinement. We will also cover core database implementation issues including storage and indexing, query processing and optimization, and transaction management. Additionally, we will cover modern topics and challenges through paper readings and discussions.

Course work will include homework assignments, paper reviews and presentations, a (late) midterm, and a mini, collaborative project.

Prerequisites: an undergraduate-level course on databases or operating systems. 3 credits.

**Course Time:** Mo We 2:30 pm - 3:45 pm, Hasbrouck Lab Add room 124

**Instructional team:**

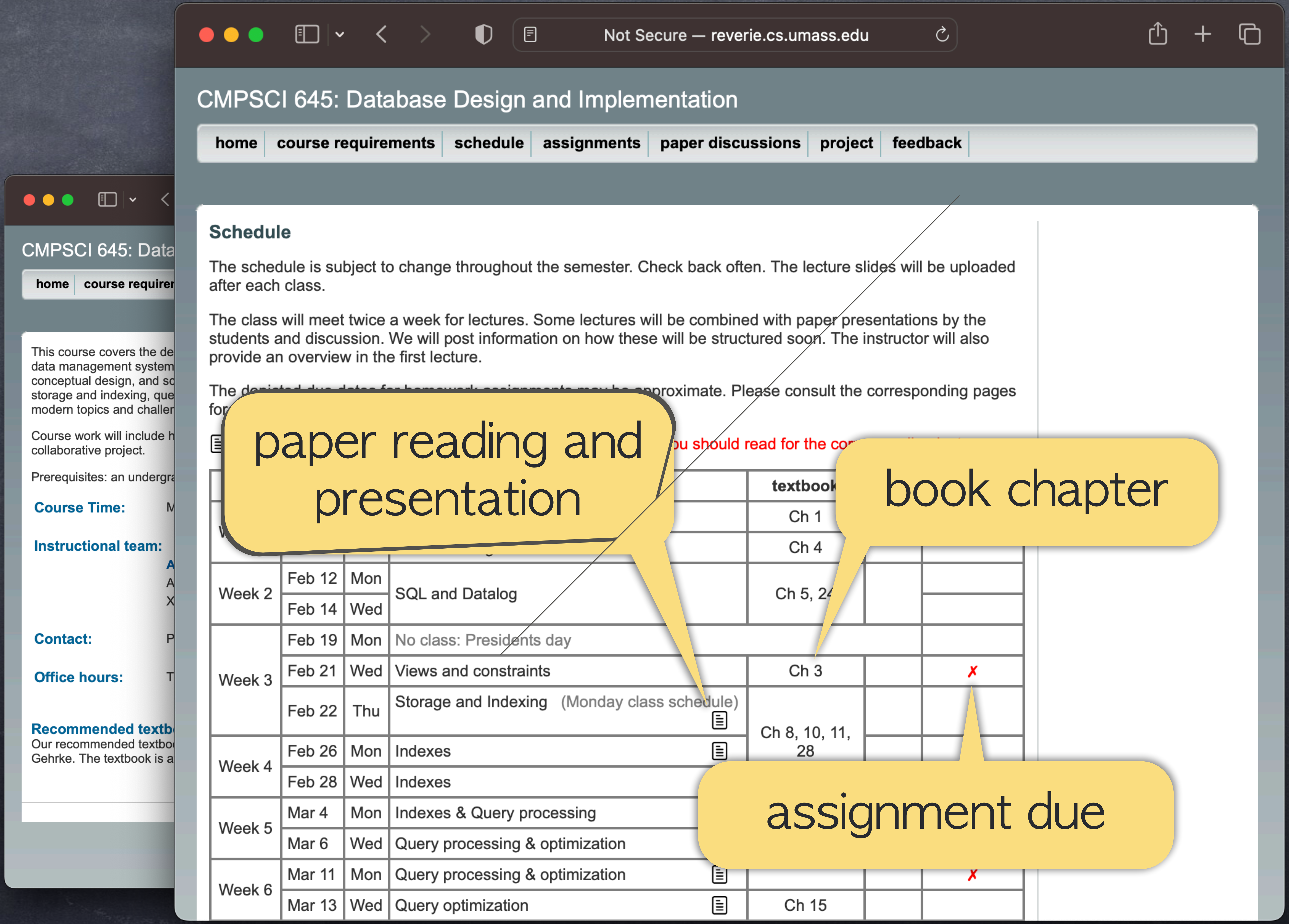
| | |
|---|---|
| **Alexandra Meliou** | Instructor |
| Ardavan Bozorgi | Teaching Assistant |
| Xi Chen | Teaching Assistant |

**Contact:** Please use Campuswire for questions to the instructional team

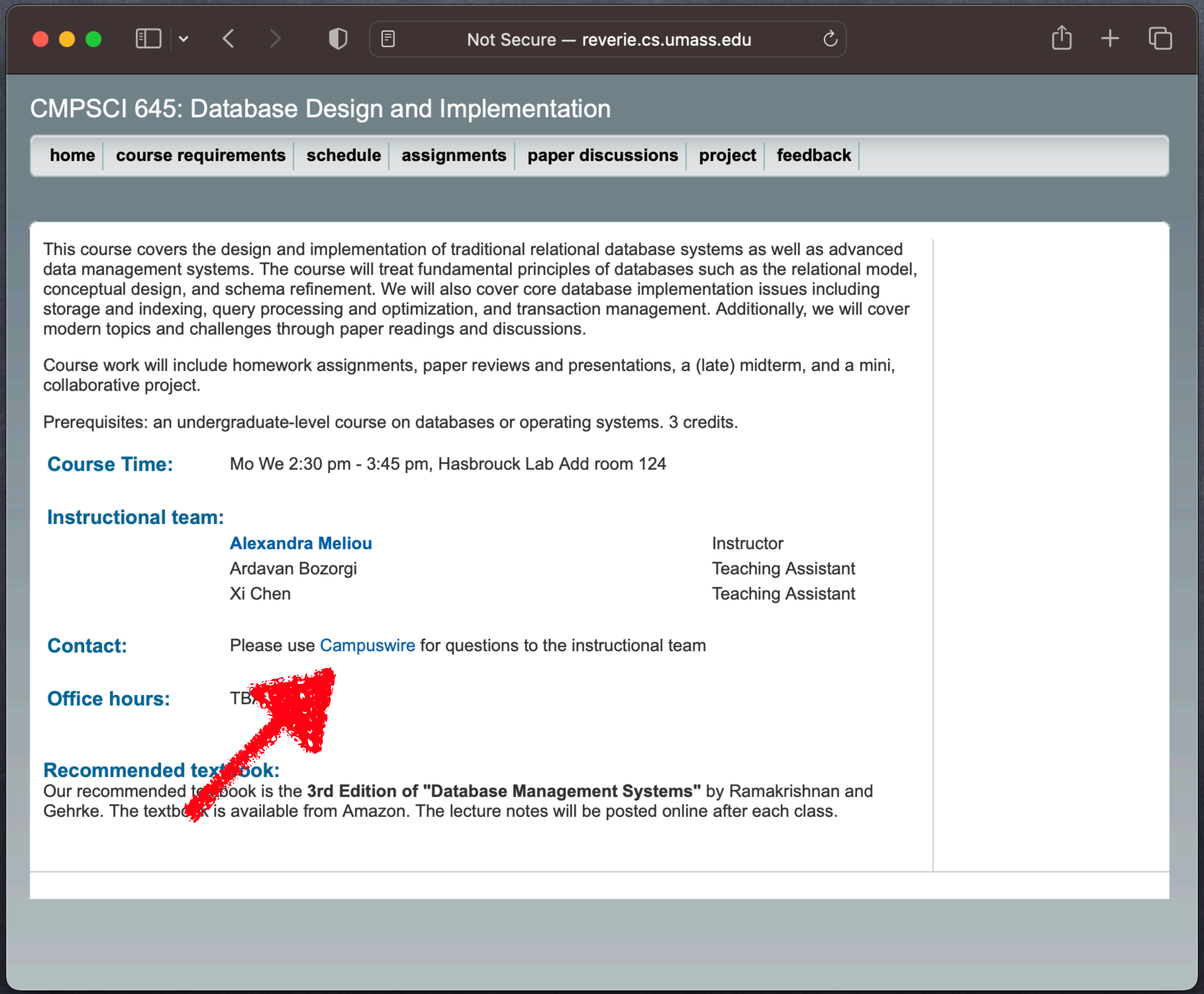
**Office hours:** TBA

**Recommended textbook:**
Our recommended textbook is the **3rd Edition of "Database Management Systems"** by Ramakrishnan and Gehrke. The textbook is available from Amazon. The lecture notes will be posted online after each class.

# http://reverie.cs.umass.edu/courses/645



# http://reverie.cs.umass.edu/courses/645



Not Secure — reverie.cs.umass.edu

CMPSCI 645: Database Design and Implementation

home | course requirements | schedule | assignments | paper discussions | project | feedback

This course covers the design and implementation of traditional relational database systems as well as advanced data management systems. The course will treat fundamental principles of databases such as the relational model, conceptual design, and schema refinement. We will also cover core database implementation issues including storage and indexing, query processing and optimization, and transaction management. Additionally, we will cover modern topics and challenges through paper readings and discussions.

Course work will include homework assignments, paper reviews and presentations, a (late) midterm, and a mini, collaborative project.

Prerequisites: an undergraduate-level course on databases or operating systems. 3 credits.

**Course Time:** Mo We 2:30 pm - 3:45 pm, Hasbrouck Lab Add room 124

**Instructional team:**

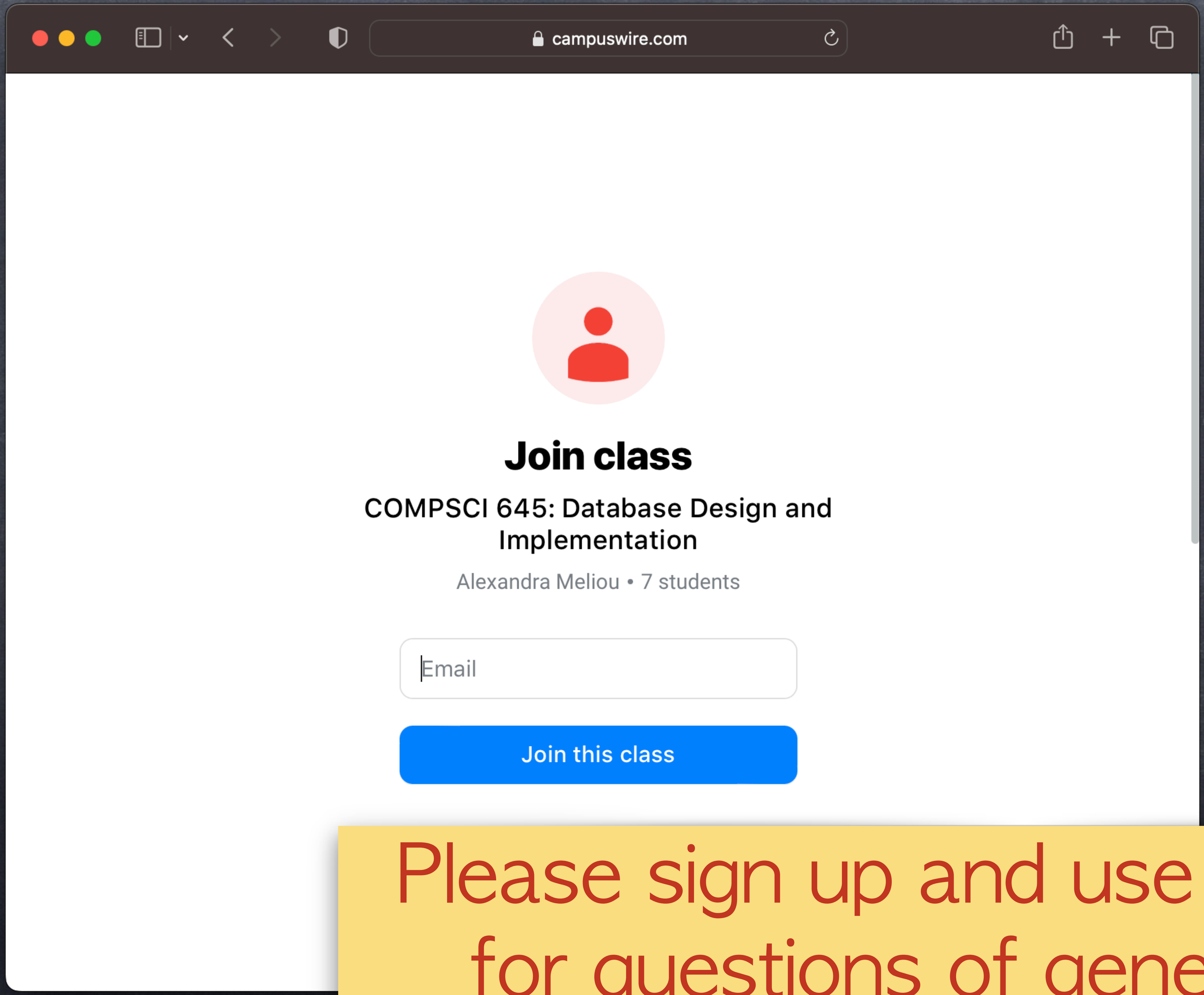
| | |
|---|---|
| Alexandra Meliou | Instructor |
| Ardavan Bozorgi | Teaching Assistant |
| Xi Chen | Teaching Assistant |

**Contact:** Please use Campuswire for questions to the instructional team

**Office hours:** TBA

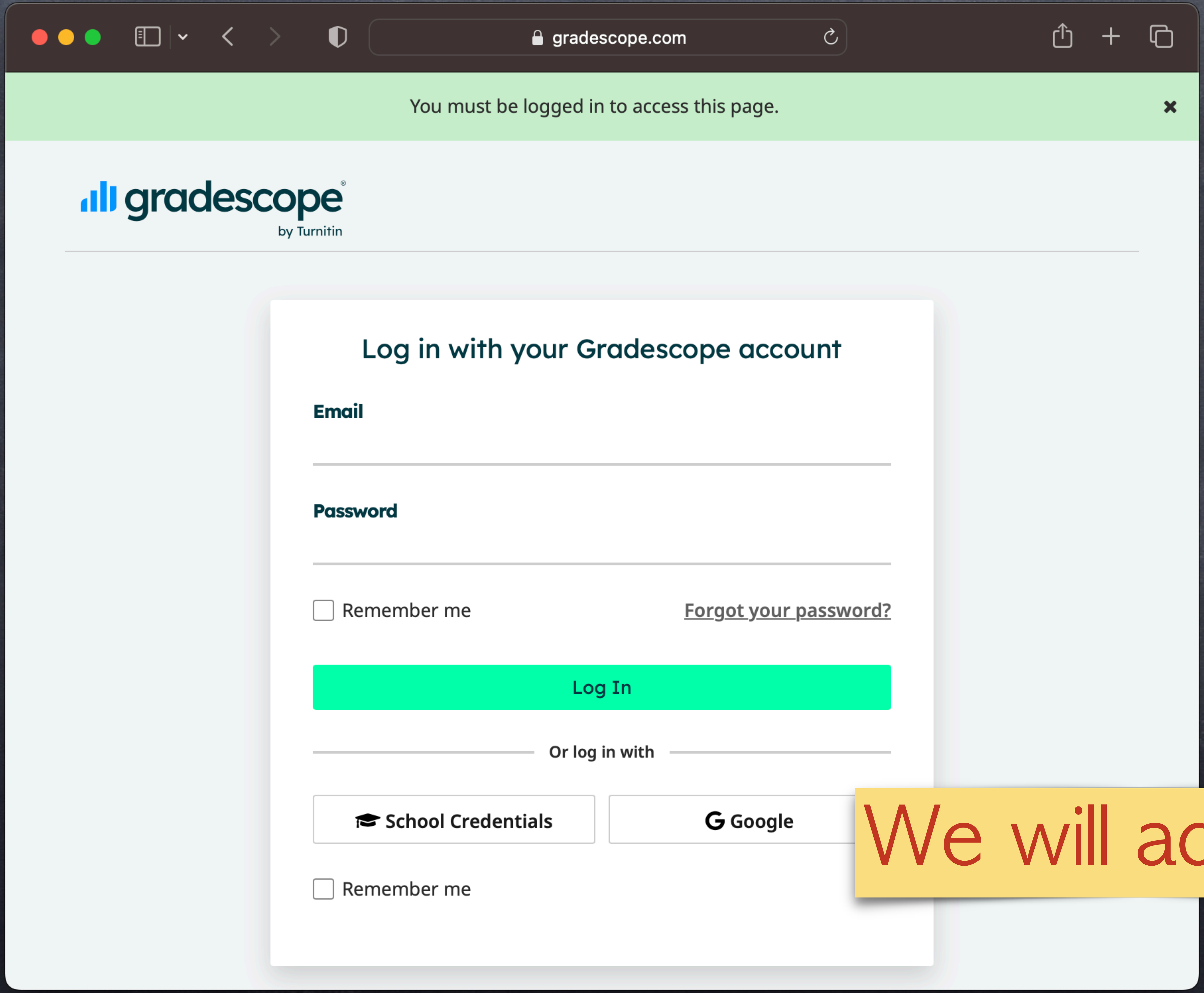
**Recommended textbook:**
Our recommended textbook is the **3rd Edition of "Database Management Systems"** by Ramakrishnan and Gehrke. The textbook is available from Amazon. The lecture notes will be posted online after each class.

https://campuswire.com/p/GB0492DFB

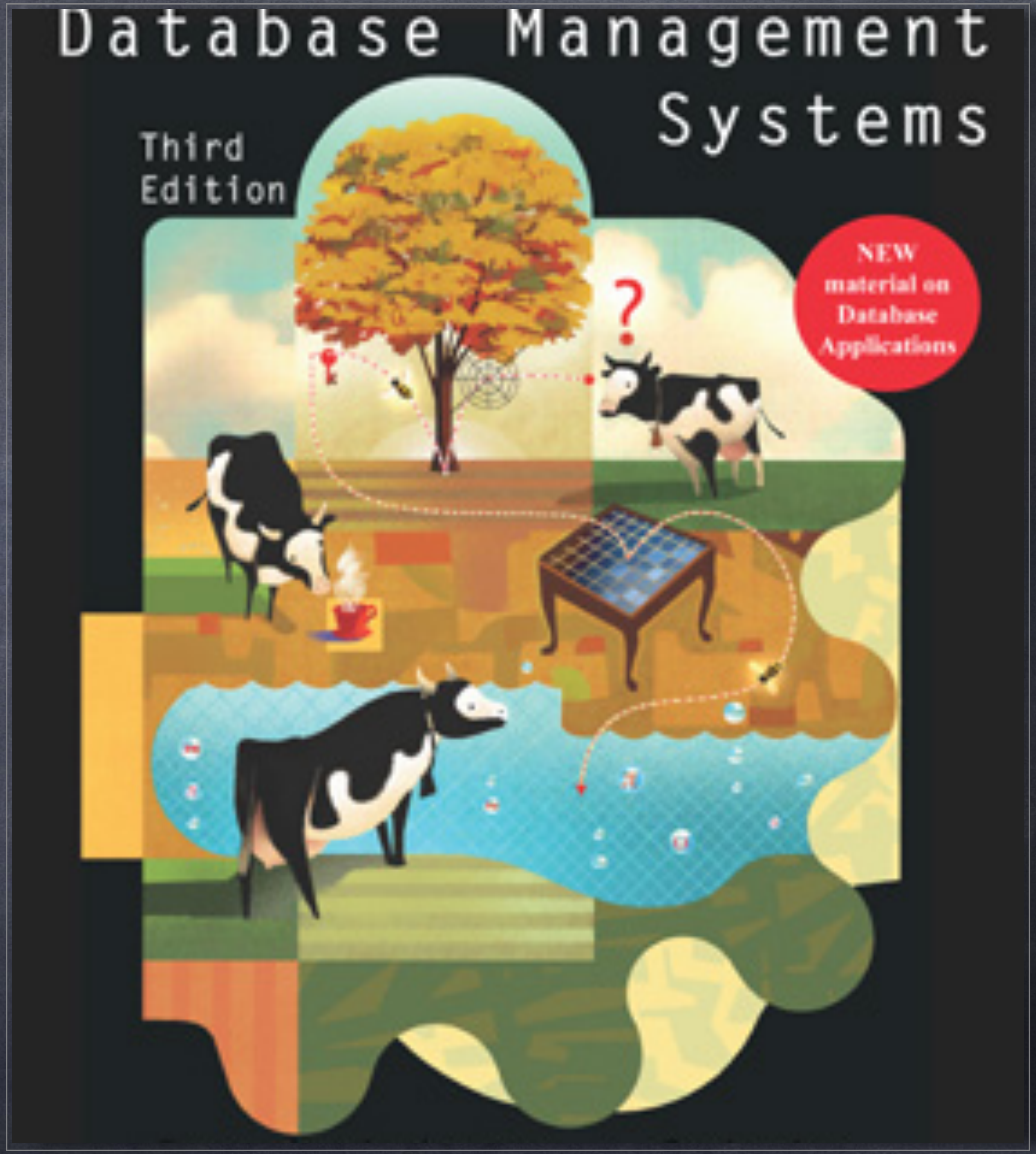



Please sign up and use Campuswire for questions of general interest

# submissions will be through Gradescope




We will add you




Database Management Systems
(3rd edition)

http://pages.cs.wisc.edu/~dbbook

# Course format

- Mo-Wed, 2:30-3:45pm, Hasbrouck Lab Add 124
- Student-led paper discussions
- Homework assignments
  - 5 individual assignments
  - group mini-project
- Late Midterm

# Disclaimer

- The class is actively designed, so there may be changes to the content, structure, and assignment types.

- You are a crucial part of this development
  - Be vocal about the things you like and the things you dont like
  - Feel free to make suggestions

# Grading

| | |
|---|---|
| Homework assignments | 40% |
| Paper reviews, presentations, and class participation | 20% |
| Midterm | 20% |
| Mini-project | 20% |

# Course work

- 5 assignments
  - Practical experience
  - Written problem sets
  - Late policy: 3 grace days, 10% penalty per day after that
- 20 paper presentations
  - role-playing format (more on this in a bit!)
- Reproducibility group project

# assignments

- to be done individually
- published on the webpage
- submissions through Gradescope
- first will be released by the end of this week

# paper reading and discussions

- Students participate in different roles:
  - paper author
  - academic researcher
  - academic historian
  - industry practitioner
  - peer reviewer

# paper reading and discussions

- Students participate in different roles:
  - paper author
  - academic researcher
  - academic historian
  - industry practitioner
  - peer reviewer

A team of 5-6 students prepares a 7-minute presentation highlighting the main contributions and results to deliver in class. Each student will take this role for one paper.

# paper reading and discussions

- Students participate in different roles:
  - paper author
  - academic researcher
  - academic historian
  - industry practitioner
  - peer reviewer

Two teams (5-6 students each) prepare 1-2 slides proposing imaginary follow-up work to this paper. Each student will take this role for two papers.

# paper reading and discussions

- Students participate in different roles:
  - paper author
  - academic researcher
  - academic historian
  - industry practitioner
  - peer reviewer

Two teams (5-6 students each) prepare 3-4 slides on the context of this paper with respect to prior and subsequent work. Each student will take this role for two papers.

# paper reading and discussions

- Students participate in different roles:
  - paper author
  - academic researcher
  - academic historian
  - industry practitioner
  - peer reviewer

Two teams (5-6 students each) prepare 1-2 slides from the perspective of a company or organization considering to use this work in an application or product. Each student will take this role for two papers.

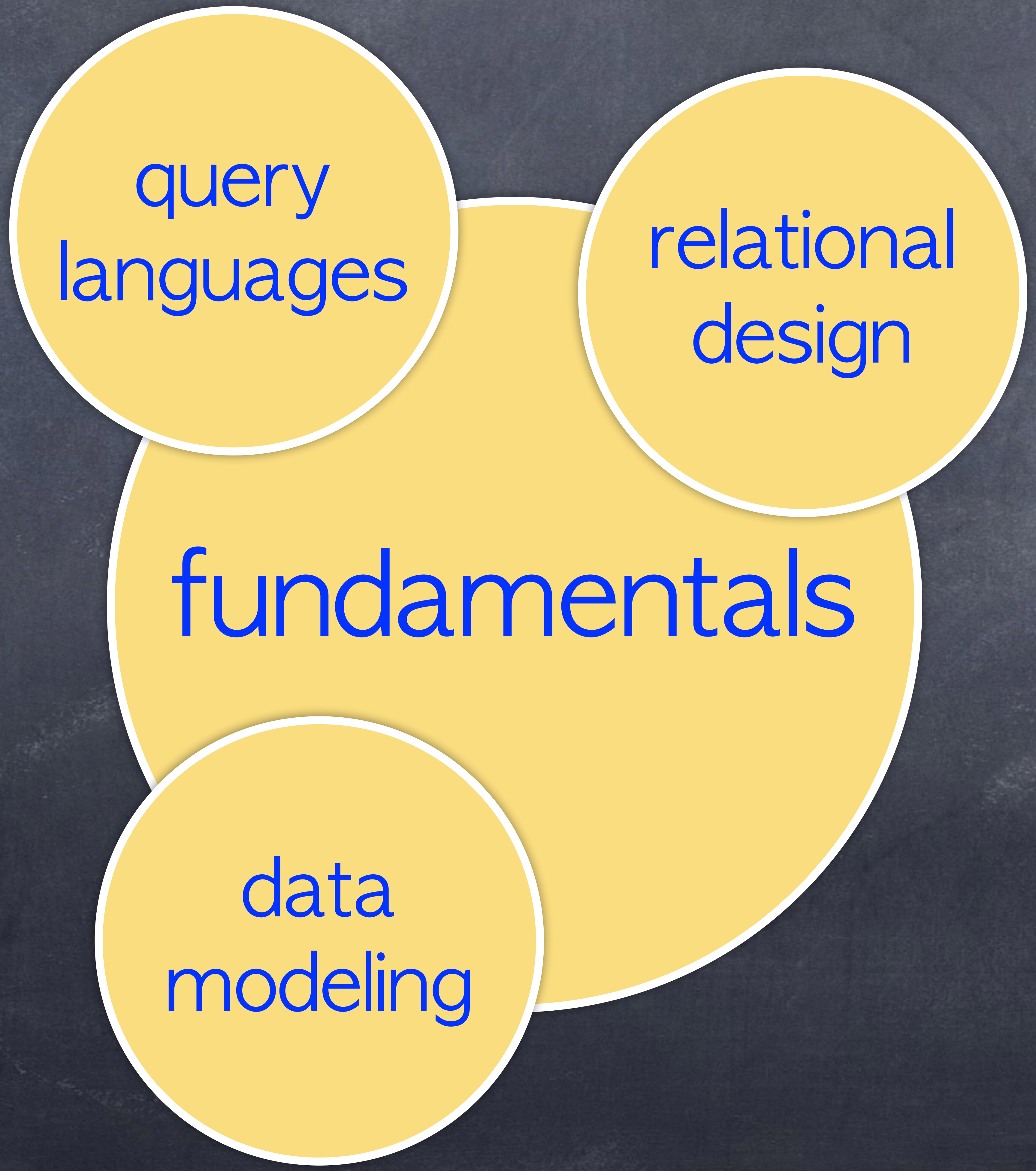
# paper reading and discussions

- Students participate in different roles:
  - paper author
  - academic researcher
  - academic historian
  - industry practitioner
  - peer reviewer

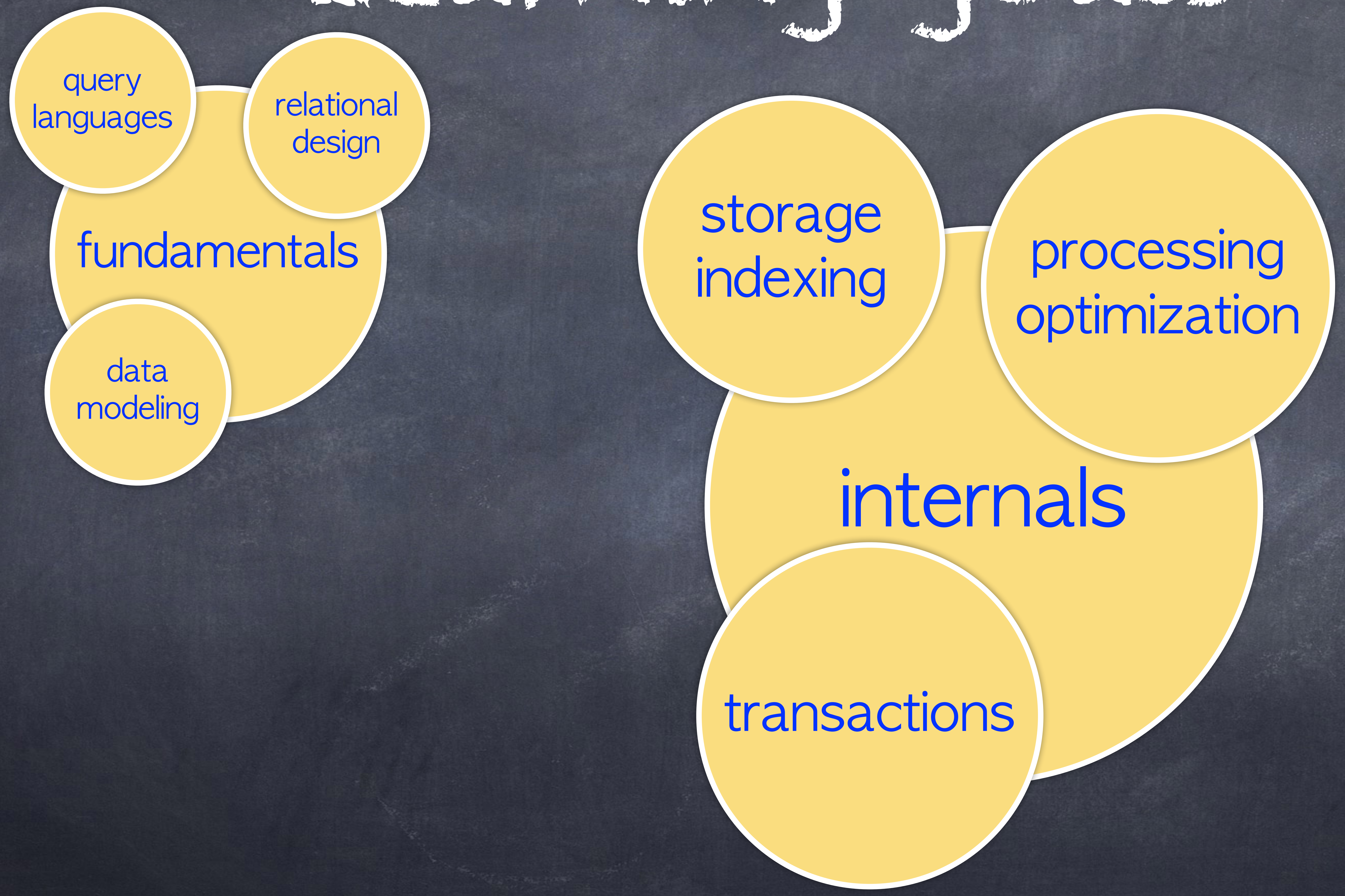
Students individually write reviews of the paper, highlighting strengths and weaknesses. Each student will take this role for three papers.

- Sign-up opens Feb 9; link will be posted on the website.
- You cannot sign up for 2 different roles for the same paper.
- You should not erase or move anyone else's name in the spreadsheet.
- You need to sign up in total for:
    - 1 paper as a paper author
    - 2 papers as an academic researcher
    - 2 papers as an academic historian
    - 2 papers as an industry practitioner
    - 3 papers as a scientific reviewer

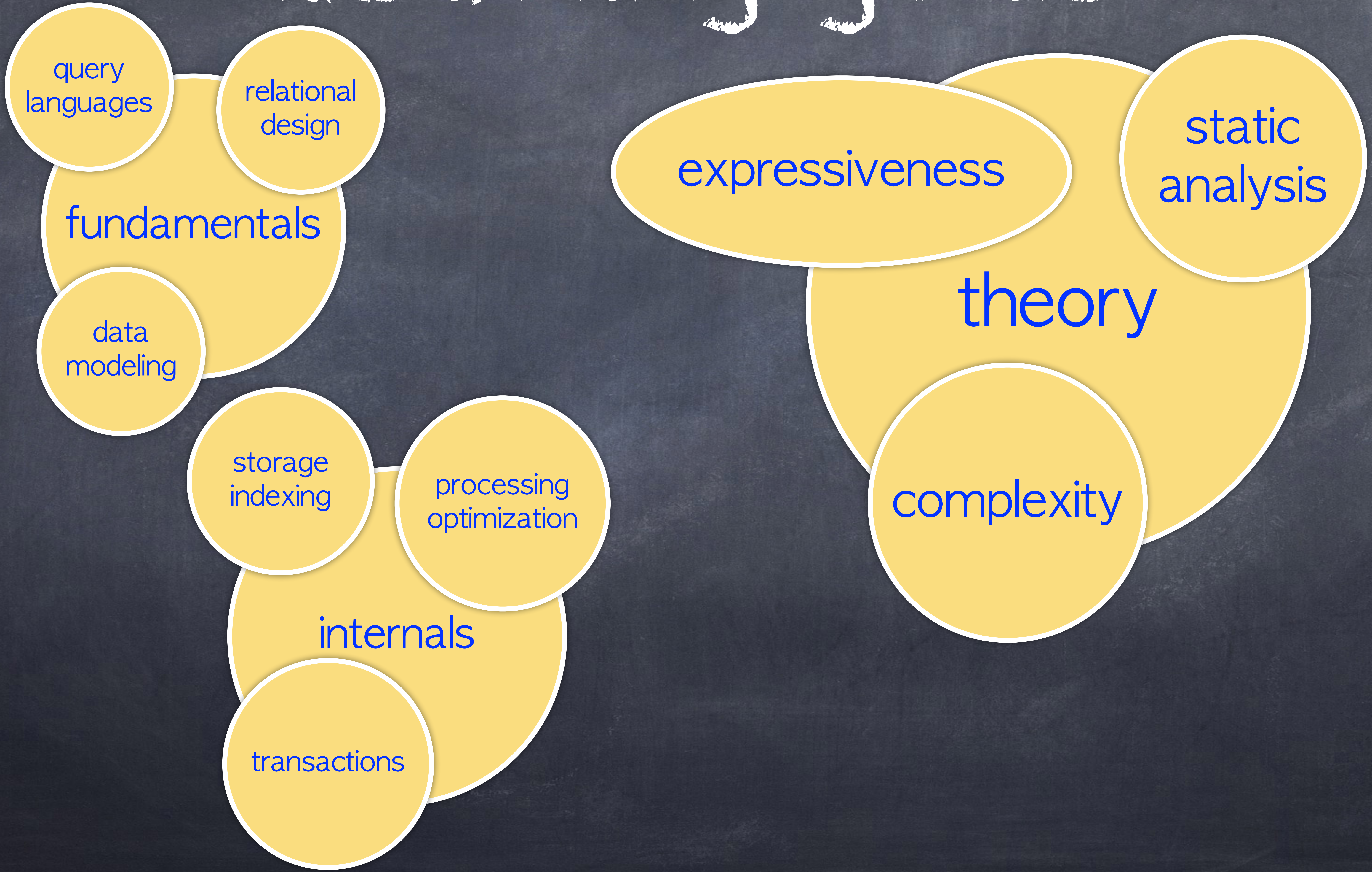
# Learning goals



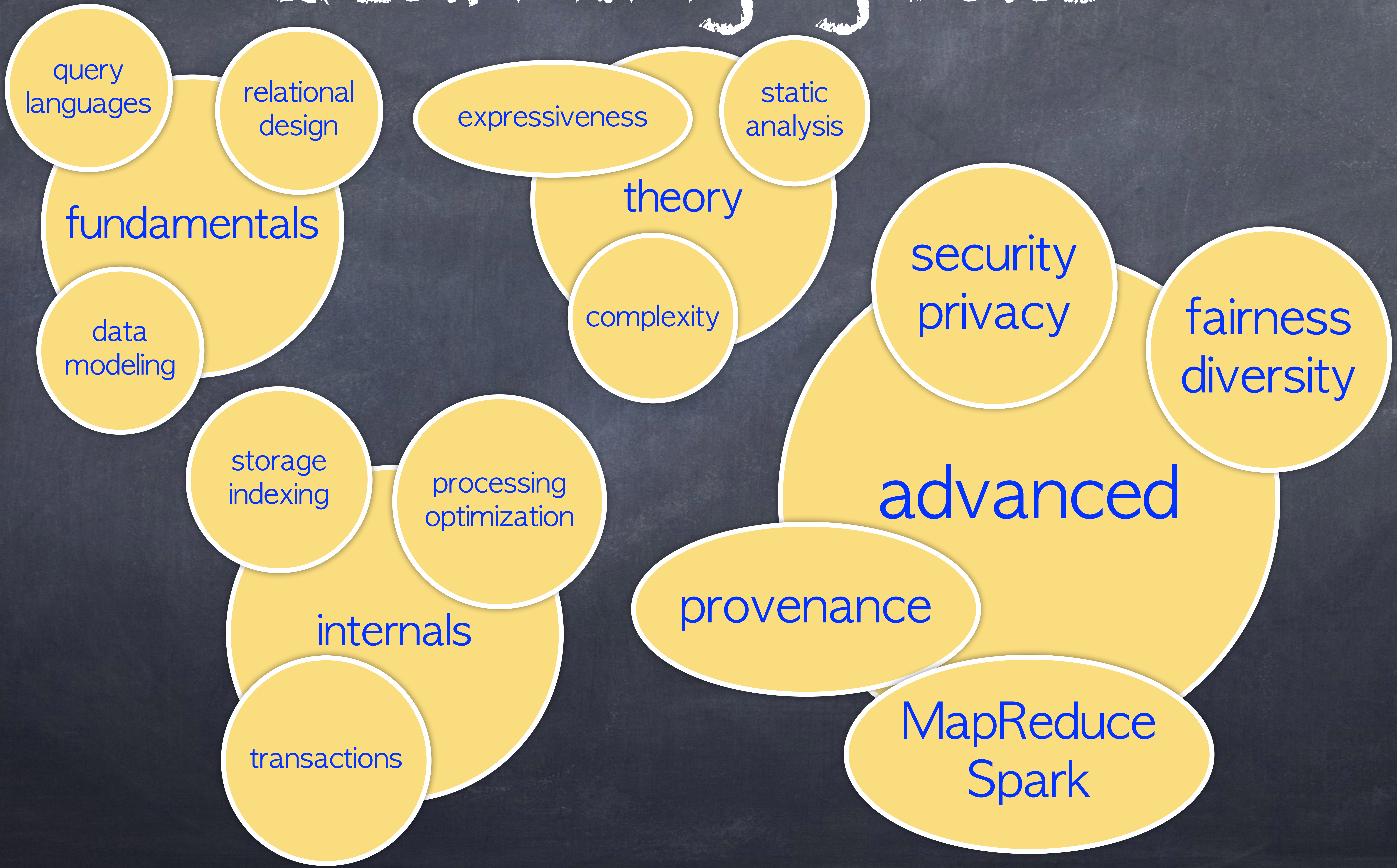
# Learning goals



# Learning goals



# Learning goals



# Why database research is exciting

- One of the broadest areas
  - Well integrated theory and systems
- A microcosm of CS:
  - Languages, operating systems, data structures, theory, algorithms, distributed systems, statistics

# What is a DBMS?

large integrated collection of data

- declarative
- efficient querying
- concurrent users
- reliable storage
- access control

# what about file systems?

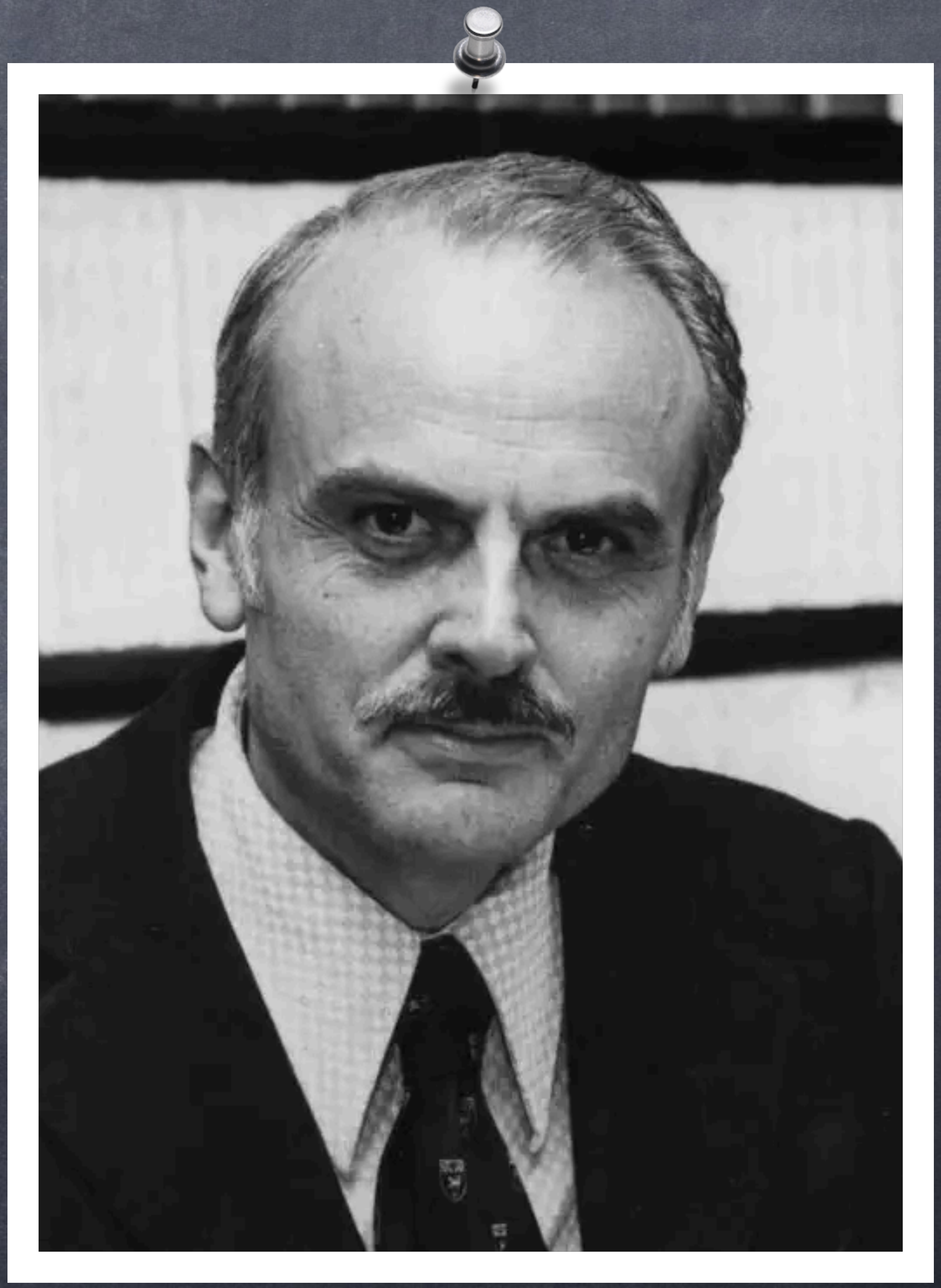
no efficient access
no query language
no specialized buffering
no recovery from failure
no safe concurrent access

# Evolution

- Early DBMSs evolved from file systems
- Many small items, many queries and updates
  - e.g., banking, reservations
- Hierarchical / network model
  - users had to think about how data was stored

# the relational model

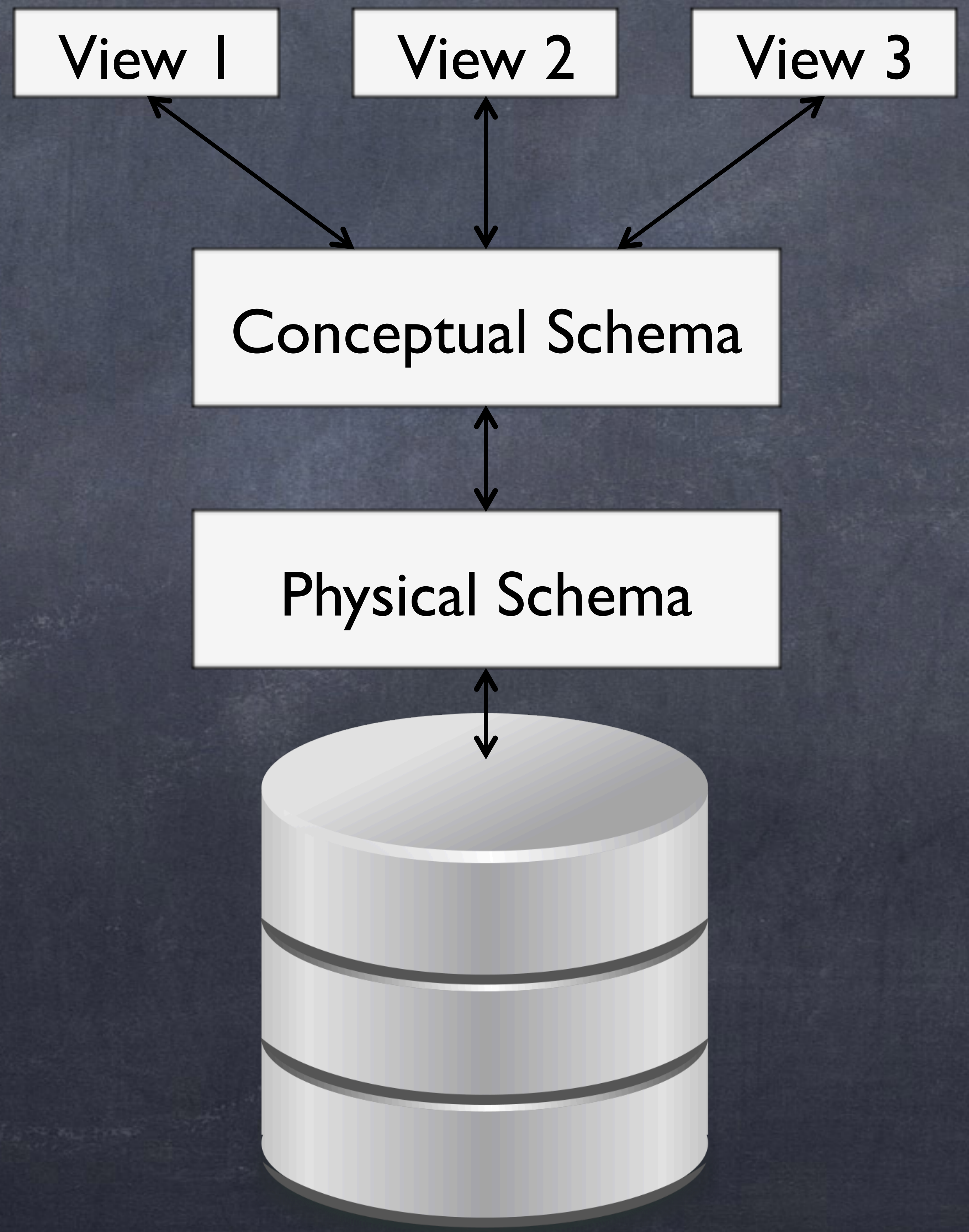
- E. F. Codd, 1970
  - data independence
  - declarative language
  - mathematical foundation



# generality & declarativity

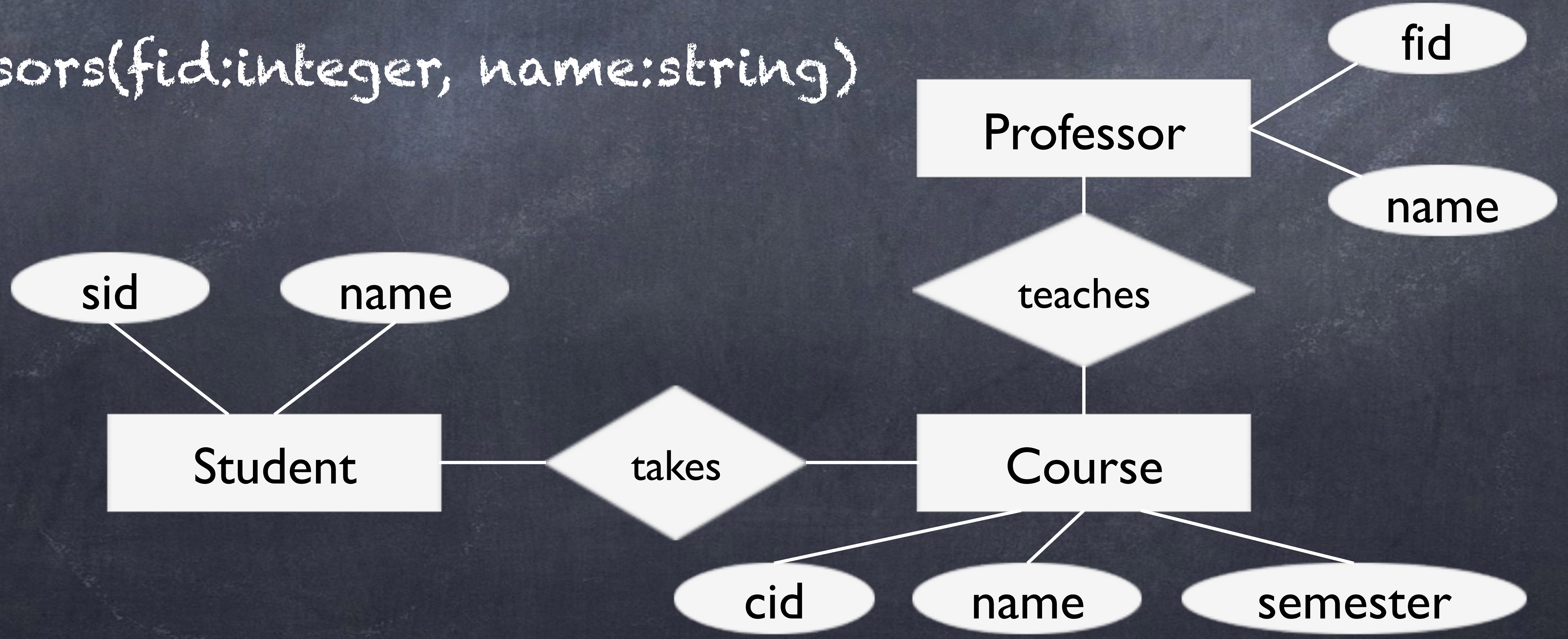
- Programmers and users do not need to know about storage, indexes, sort orders, concurrent users, etc.
- Use logical model, high-level schema
- The DBMS determines how to retrieve the data

# levels of abstraction



# Example: university DB

- Conceptual schema:
  - Students(sid:integer, name:string)
  - Courses(cid:integer, name:string, semester:string)
  - Professors(fid:integer, name:string)



# designing a schema

- Convert to tables and constraints
- Physical design: disk layout, indices

students

| sid | name |
|---|---|
| 1 | Jill |
| 2 | Bo |
| 3 | Maya |

takes

| fid | cid |
|---|---|
| 1 | 621 |
| 1 | 645 |
| 3 | 390 |

courses

| cid | name | sem. |
|---|---|---|
| 645 | DB | S'18 |
| 621 | Soft. Eng. | S'18 |
| 345 | DB | F'17 |

professors

| sid | name |
|---|---|
| 1 | Brun |
| 2 | Meliou |
| 3 | Miklau |

teaches

| fid | cid |
|---|---|
| 1 | 621 |
| 2 | 645 |
| 2 | 345 |

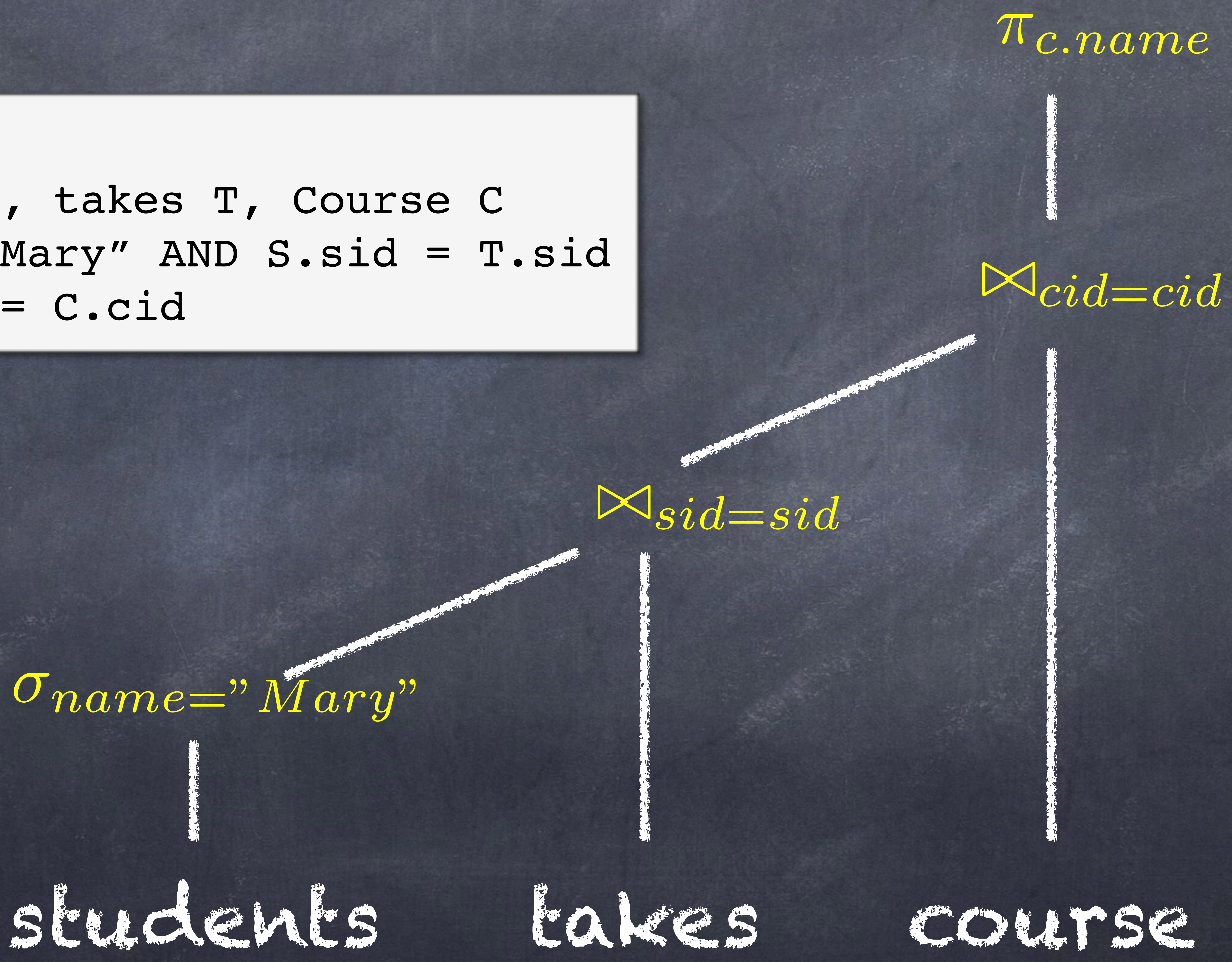
# queries

```
SELECT C.name
FROM   Students S, takes T, Course C
WHERE  S.name = "Mary" AND S.sid = T.sid
       AND T.cid = C.cid
```

find all courses that Mary takes

# behind the scenes

```
SELECT C.name
FROM   Students S, takes T, Course C
WHERE  S.name = “Mary” AND S.sid = T.sid
       AND T.cid = C.cid
```



# DBMSs and DB research

- Huge industry
  - Large data warehouses
  - Distributed databases
  - Integration
- But: not all data is in a DBMS
  - Scientific data
  - Personal data
  - www
- Data management research has expanded

# DB research is broad

- core topics (DB internals, processing, optimization, transactions)
- scientific data
- streaming data
- provenance, security, privacy
- cleaning, matching, integration
- distributed data / querying
- usability, visualization
- crowdsourcing
- ...

6:22pm

6:15pm

ORBITZ

**American Airlines # 119**

**Leg 1: In Transit**

Departs: Newark (EWR) View real-time airport conditions at I

Gate: 32

| **Scheduled** | **Estimated** | **Actual** |
|---|---|---|
| **6:22p** Dec 8 | - | **6:32p** Dec 8 |

Arrives: Los Angeles (LAX) View real-time airport conditions

Gate: 42B

| **Scheduled** | **Estimated** | **Actual** |
|---|---|---|
| **9:54p** Dec 8 | **9:47p** Dec 8 | |

American Airlines Flight Number 119 (AA119)

FlightView

Departure

| | |
|---|---|
| Airport: | |
| Scheduled Time: | 6:15 PM, Dec 08 |
| Takeoff Time: | 6:53 PM, Dec 08 |
| Terminal - Gate: | Terminal A - 32 |

ArrivalStatus: In Air

| | |
|---|---|
| Airport: | |
| Scheduled Time: | 9:40 PM, Dec 08 |
| Estimated Time: | 9:42 PM, Dec 08 |
| | Track This Flight Live! |
| Time Remaining: | 25 min |
| Terminal - Gate: | Terminal 4 - 42B |
| Baggage Claim: | 4 |

FlightAware

| | |
|---|---|
| **Aircraft** | Boeing 737-800 (twin-jet) (B738/Q - track or photos) |
| **Origin** | Terminal A / Gate 32 / Newark Liberty Intl (KEWR - track or info) |
| **Destination** | Terminal 4 / Gate 42B / Los Angeles Intl (KLAX - track or info) |
| | *Other flights between these airports* |
| **Route** | ZIMMZ Q42 BTRIX Q480 AIR J80 VHP J80 MCI J24 SLN J102 ALS J44 RSK J64 PGS RIIVR2 (Decode) |
| **Date** | 2011年 12月 08日 (Thursday) |
| **Duration** | 5 hours 43 minutes |
| **Progress** | 20 minutes left<br>5 hours 23 minutes |
| **Status** | En Route (2,284 sm down; 168 sm to go) |
| **Distance** | Direct: 2,451 sm Planned: 2,458 |
| **Fare** | $51.99 to $3,561.11; average: $241.96 (airline insight) |
| **Cabin** | First: Dinner / Economy: Food for sale |

| | Scheduled | 7-day Average | *Actual/Estimated* |
|---|---|---|---|
| **Departure** | 06:15PM EST | 07:08PM EST | 06:53PM EST |
| **Arrival** | 08:33PM PST | 09:17PM PST | *09:36PM PST* |

9:54pm

9:40pm

8:33pm

# Bogdan Alexe

-- show all -- | by year | Trier 1

## [–] 2010 – today

**2014**

- [j10] Bogdan Alexe, Mary Roth, Wang-Chiew Tan: **Preference-aware Integration of Temporal Data.** PVLDB 8(4): 365-376 (2014)

**2013**

- [c18] Tyler Baldwin, Yunyao Li, Bogdan Alexe, Ioana Roxana Stanoi: **Automatic Term Ambiguity Detection.** ACL (2) 2013: 804-809
- [c17] Mauricio A. Hernández, Kirsten Hildrum, Prateek Jain, Rohit Wagle, Bogdan Alexe, Rajasekar Krishnamurthy, Ioana Roxana Stanoi, Chitra Venkatramani: **Constructing consumer profiles from social media data.** BigData Conference 2013: 710-716
- [c16] Bogdan Alexe, Douglas Burdick, Mauricio A. Hernández, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana Stanoi, Ryan Wisnesky: **High-Level Rules for Integration and Analysis of Data: New Challenges.** In Search of Elegance in the Theory and Practice of Computation 2013: 36-55
- [c15] Bogdan Alexe, Wang-Chiew Tan: **A New Framework for Designing Schema Mappings.** In Search of Elegance in the Theory and Practice of Computation 2013: 56-88

**2012**

- [j9] Thomas Deselaers, Bogdan Alexe, Vittorio Ferrari: **Weakly Supervised Localization and Learning with Generic Knowledge.** International Journal of Computer Vision 100(3): 275-293 (2012)
- [j8] Bogdan Alexe, Thomas Deselaers, Vittorio Ferrari: **Measuring the Objectness of Image Windows.** IEEE Trans. Pattern Anal. Mach. Intell. 34(11): 2189-2202 (2012)
- [j7] Bogdan Alexe, Mauricio A. Hernández, Lucian Popa, Wang Chiew Tan: **MapMerge: correlating independent schema mappings.** VLDB J. 21(2): 191-211 (2012)
- [c14] Bogdan Alexe, Nicolas Heess, Yee Whye Teh, Vittorio Ferrari: **Searching for objects driven by context.** NIPS 2012: 890-898
- [c13] Bogdan Alexe, Mauricio A. Hernández, Kirsten Hildrum, Rajasekar Krishnamurthy, Georgia Koutrika, Meenakshi Nagarajan, Haggai Roitman,

# MapMerge: Correlating Independent Schema Mappings

Bogdan Alexe
UC Santa Cruz

Mauricio Hernández
IBM Almaden

Lucian Popa
IBM Almaden

Wang-Chiew Tan
IBM Almaden & UC Santa Cruz

**ABSTRACT**

One of the main s
to design the mapp
ships between the s

# Preference-aware Integration of Temporal Data

Bogdan Alexe
IBM Almaden
balexe@us.ibm.com

Mary Roth
IBM Almaden and UCSC
torkroth@us.ibm.com

Wang-Chiew Tan
UCSC
tan@cs.ucsc.edu

**ABSTRACT**

A complete description of an
data source, but rather, it is of
sources. Applications based o

# Searching for objects driven by context

**Bogdan Alexe**
BIWI
ETH Zurich

**Nicolas Heess**
Gatsby Unit
UCL

**Yee Whye Teh**
Department of Statistics
University of Oxford

**Vittorio Ferrari**
School of Informatics
University of Edinburgh

## Abstract

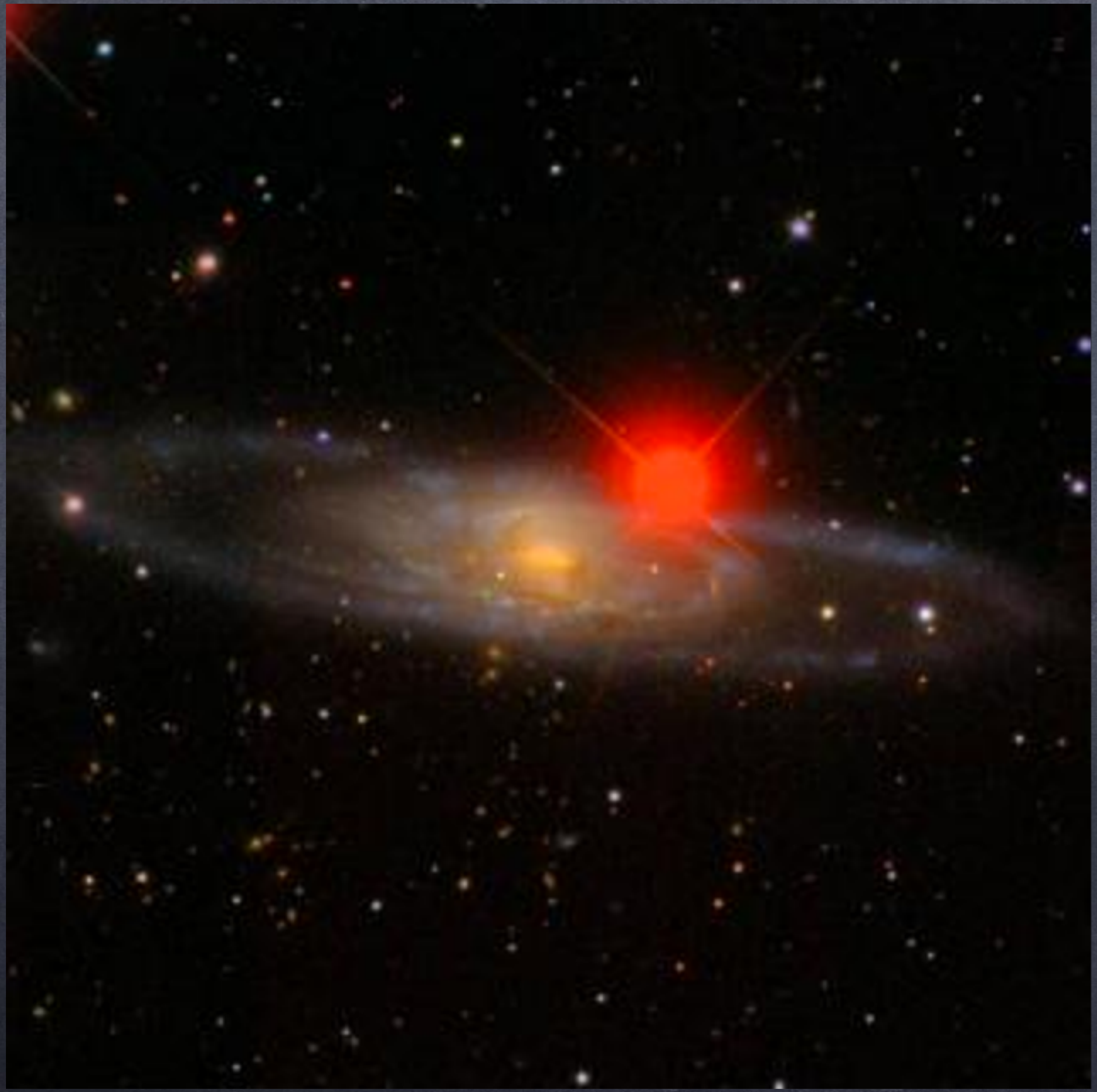
The dominant visual search paradigm for object class detection is sliding windows. Although simple and effective, it is also wasteful, unnatural and rigidly hardwired. We propose strategies to search for objects which intelligently explore the space of windows by making sequential observations at locations decided based on previous observations. Our strategies adapt to the class being searched and to the content of a particular test image, exploiting context as the statistical

how to debug?

```
SELECT  a3.fname, a3.lname
FROM  Actor a0, Casts c0, Casts c1,
   Casts c2, Casts c3, Actor a3
WHERE a0.fname = 'Kevin' AND a0.lname = 'Bacon' AND
   c0.pid = a0.id AND c0.mid = c1.mid AND
   c1.pid = c2.pid AND c2.mid = c3.mid AND
   c3.pid = a3.id AND
   NOT (a3.fname = 'Kevin' and a3.lname = 'Bacon') AND
         NOT EXISTS (SELECT xc1.pid
                FROM    Actor xa0, Casts xc0, Casts xc1
                      WHERE  xa0.fname = 'Kevin' AND
           xa0.lname = 'Bacon' AND
                             xa0.id = xc0.pid AND
           xc0.mid = xc1.mid AND xc1.pid = a3.id)
GROUP BY        a3.id, a3.fname, a3.lname;
```

alternatives to writing queries?

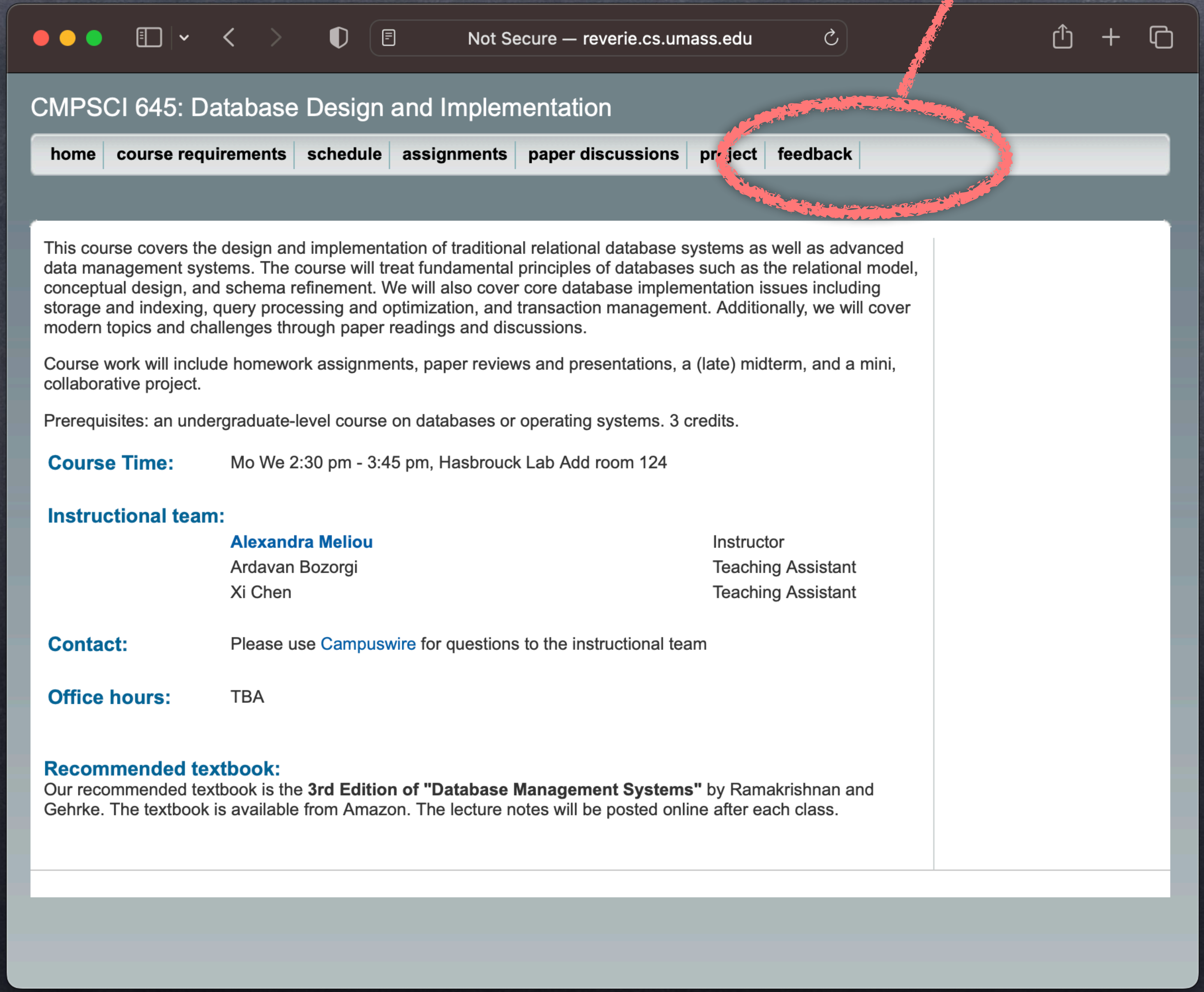
# databases for applications



sloan digital sky survey

# Questions?

please give us feedback!




CMPSCI 645: Database Design and Implementation

home | course requirements | schedule | assignments | paper discussions | project | feedback

This course covers the design and implementation of traditional relational database systems as well as advanced data management systems. The course will treat fundamental principles of databases such as the relational model, conceptual design, and schema refinement. We will also cover core database implementation issues including storage and indexing, query processing and optimization, and transaction management. Additionally, we will cover modern topics and challenges through paper readings and discussions.

Course work will include homework assignments, paper reviews and presentations, a (late) midterm, and a mini, collaborative project.

Prerequisites: an undergraduate-level course on databases or operating systems. 3 credits.

**Course Time:** Mo We 2:30 pm - 3:45 pm, Hasbrouck Lab Add room 124

**Instructional team:**

| | |
|---|---|
| **Alexandra Meliou** | Instructor |
| Ardavan Bozorgi | Teaching Assistant |
| Xi Chen | Teaching Assistant |

**Contact:** Please use Campuswire for questions to the instructional team

**Office hours:** TBA

**Recommended textbook:**
Our recommended textbook is the **3rd Edition of "Database Management Systems"** by Ramakrishnan and Gehrke. The textbook is available from Amazon. The lecture notes will be posted online after each class.