ANNOUNCEMENTS

1) **iClicker registration in Canvas**

Attendance is 5% of the course grade and is tracked via iClicker participation with answers being asked in class. Students must attend two-thirds of all classes to receive full attendance credit. Questions in class do not have to be answered correctly to get credit (but please don't just mash B). You must answer all questions during a certain class to receive credit for attending that day. iClickers may be set up on the course Canvas page.

2) **First assignment will be out today (due by 2/9)**

RELATIONAL DATABASES

INTRODUCTION TO DATA SCIENCE

CARSTEN BINNIG

BROWN UNIVERSITY

CLICKER QUESTION

How well do you know relational database systems

A. What are they?

B. I know SQL and tables

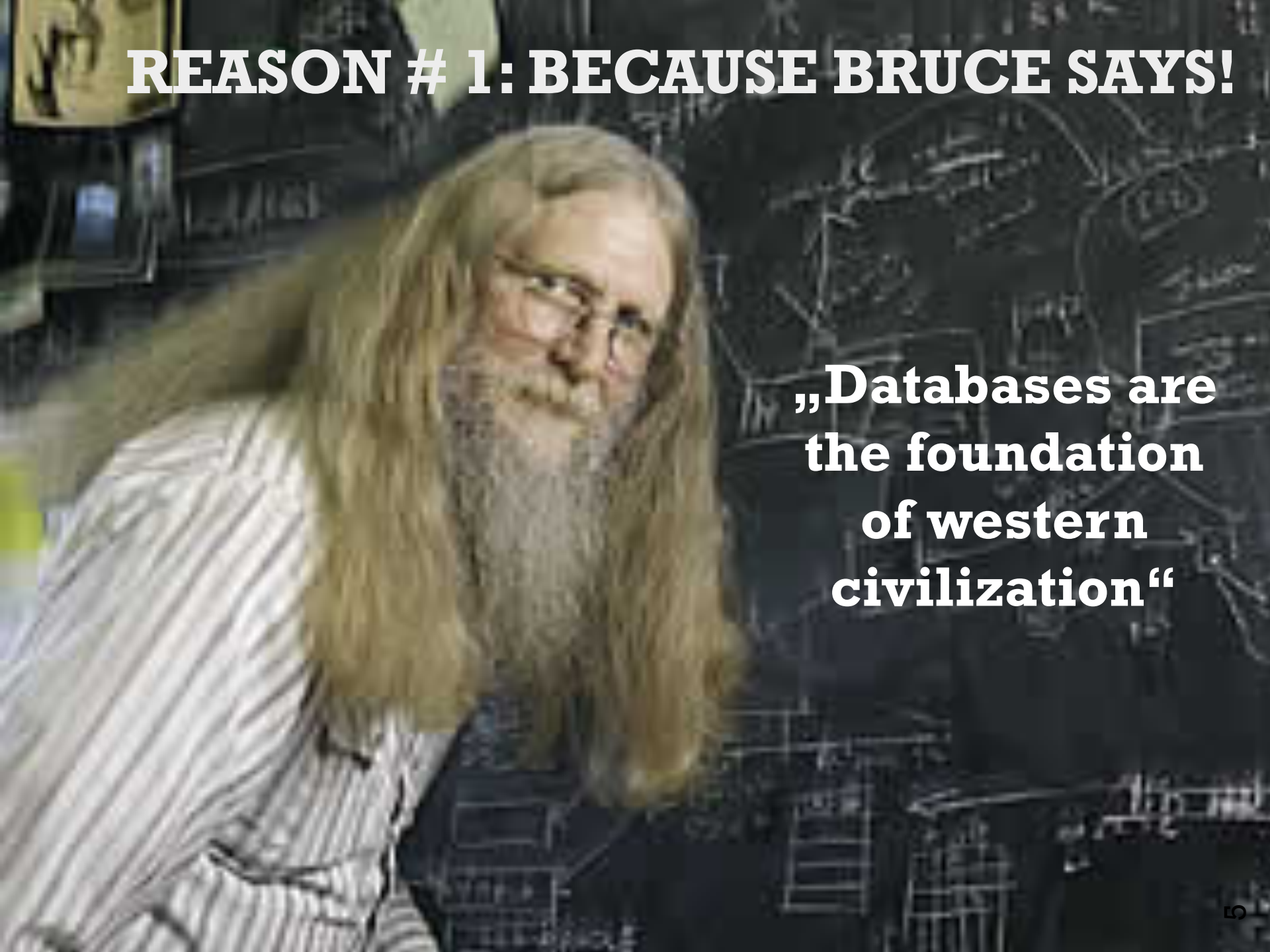
C. I know SQL, ER diagrams, and the relational algebra

D. I know normalization (e.g., 4th normal form) and, star and snowflake schemas

E. I know database internals (e.g., query optimization)

WHY RELATIONAL DATABASES?

REASON # 1: BECAUSE BRUCE SAYS!



„Databases are the foundation of western civilization“

Reason # 2:
A Major Skill

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Reasons 3: Database Systems got many things right

WHY NOT FLAT FILES?

… they are so simple!

REASON 1: DATA CONSISTENCY

[Name, Course, Grade]
Stan Zdonik, CS22, A;
Mike Stonebraker, CS123, B;
Tim Kraska, CS22, A;
…
Binnig, C., CS113, 1;

File system does not even know this

Correct Format?

Correct Reference?

Correct Datatype?

REASON 2: SCALABILITY

Modern Database Systems (DBMSs) are designed to **scale to large amounts of data (i.e., >> a terabyte)**

Must be **well optimized on a single node**

Must be able to **execute on 100's or 1000's of nodes.**

REASON 3: DATA RETRIEVAL

[Name, Course, Grade]
Stan Zdonik, CS22, A;
Mike Stonebraker, CS123, B;
Tim Kraska, CS22, A;
…

Querying

- Find all courses of Tim Kraska
- Count students with an A in CS22

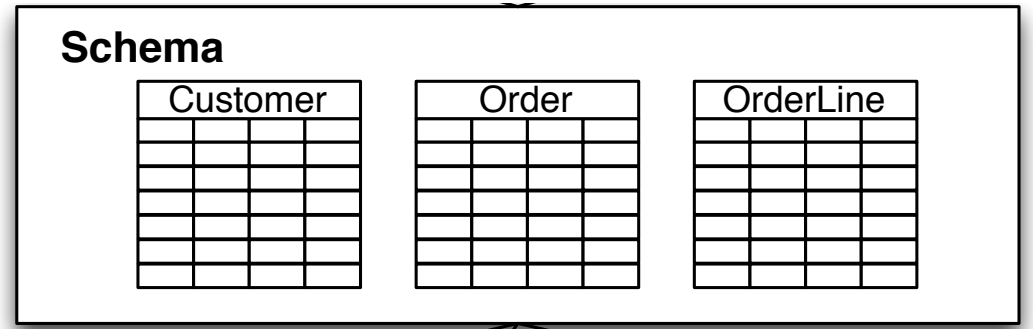
For every query we would need to **write a program!**

Each program needs to **read the whole file?**

SQL is easy to write (Querying)

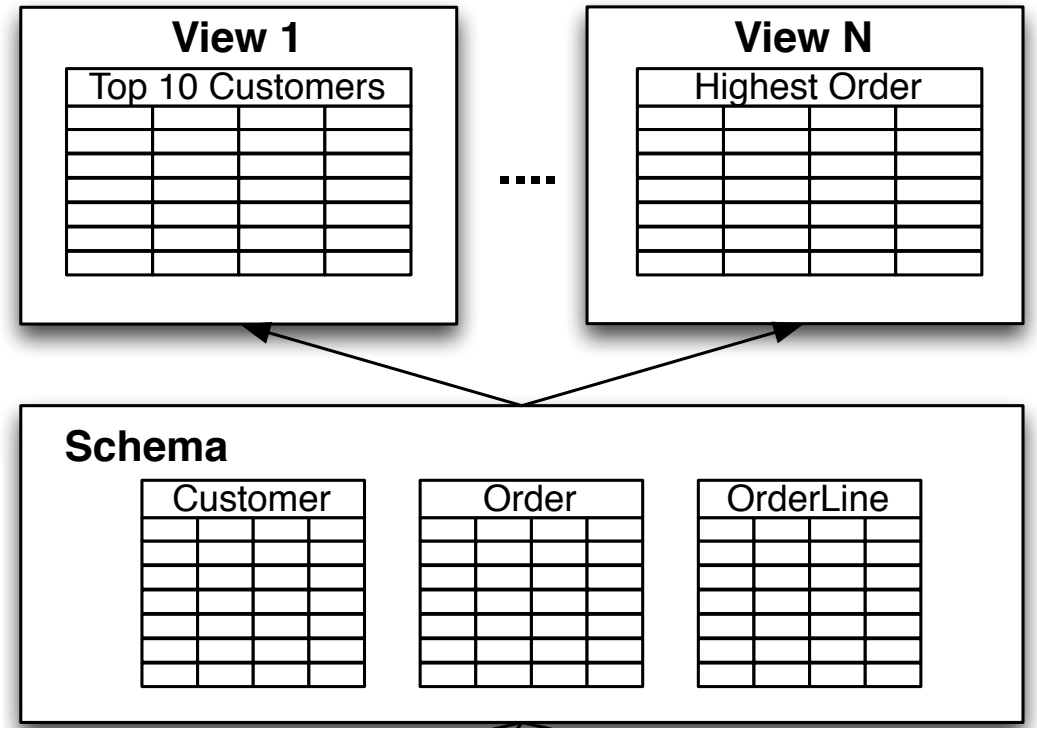
SQL is efficiently executable (Retrieval)

REASON 4: DATA INDEPENDENCE

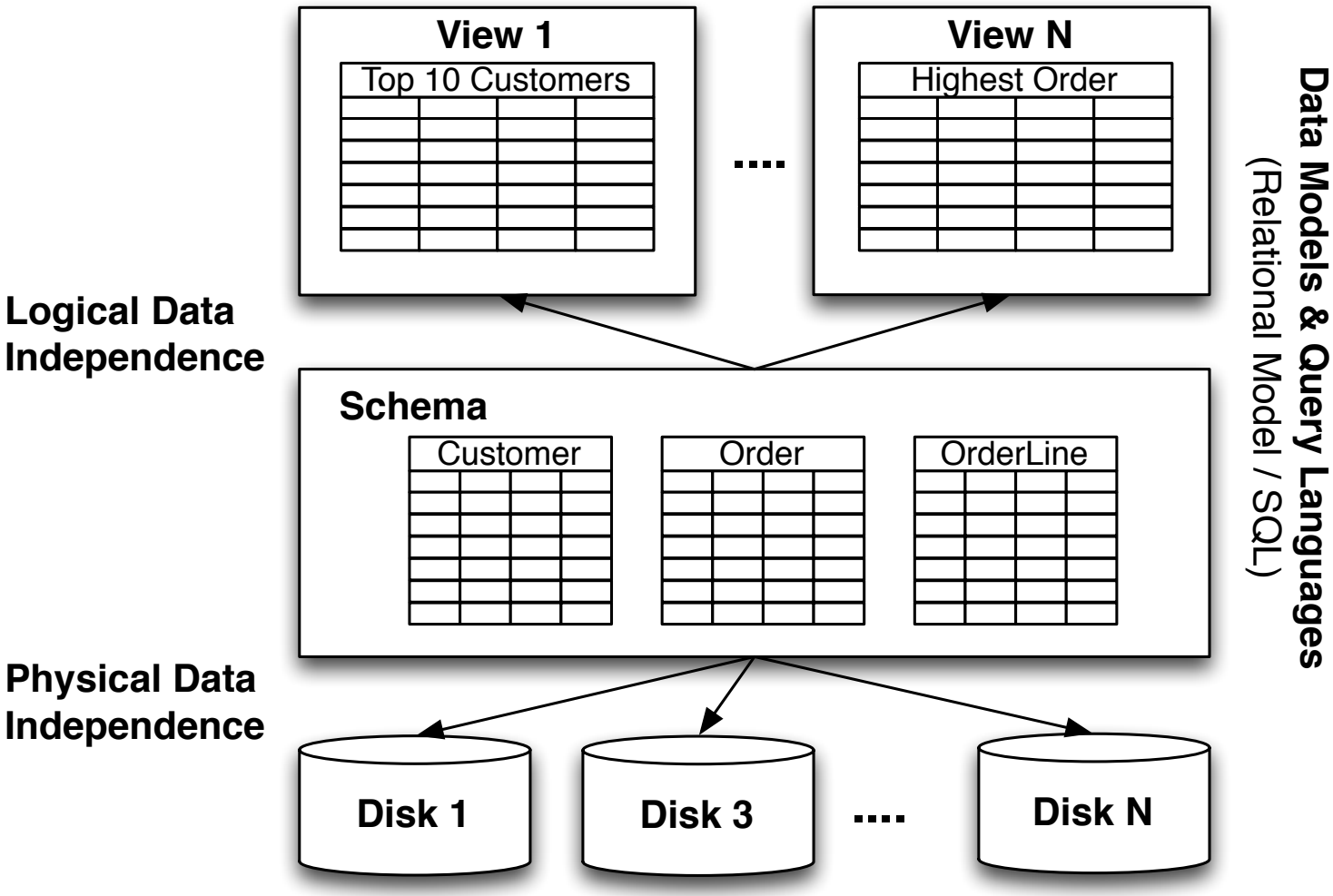


REASON 4: DATA INDEPENDENCE

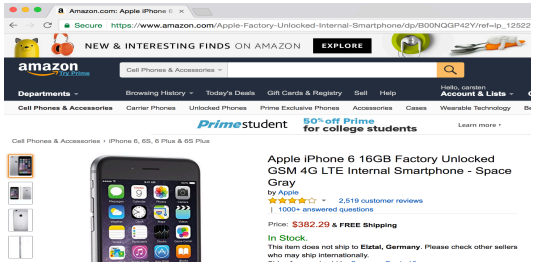
Logical Data Independence

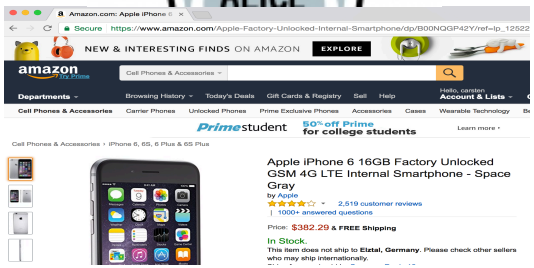


REASON 4: DATA INDEPENDENCE



REASON 5: CONCURRENT ACCESS





Stock = 100

Stock = 99

Stock = 99

Stock = 100

ProductID	ProductName	Stock
1	iPhone 7	100
…	…	…

WHY NOT FLAT FILES?

Data Consistency

Scalability

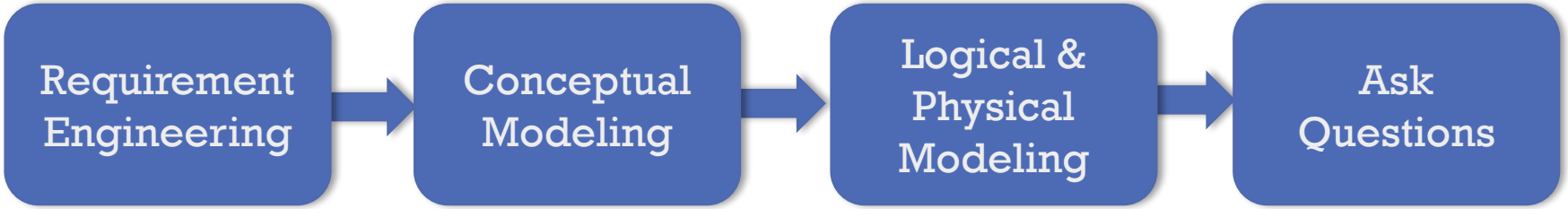
Data Retrieval

Data Independence

Concurrent Access

…

DATABASES FOR DATA SCIENTIST



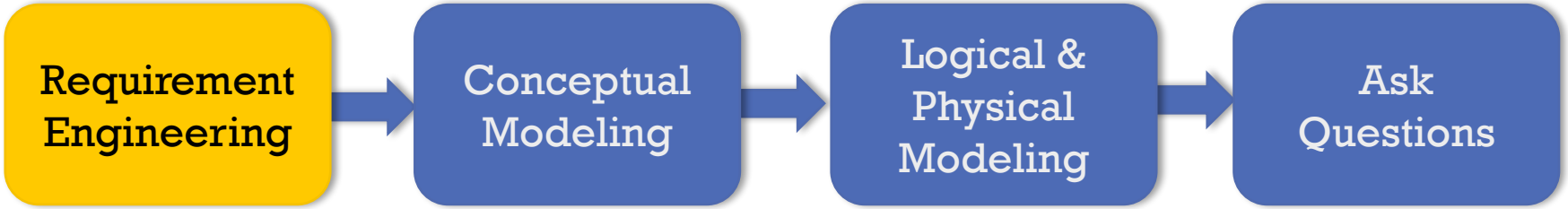
Book of duty

Conceptual Design (ER)

- Logical design (relational schema)
- Physical design (index, hints)

- Relational Algebra
- SQL

DATABASES FOR DATA SCIENTIST



Book of duty

Conceptual Design (ER)

- Logical design (relational schema)
- Physical design (index, hints)

- Relational Algebra
- SQL

BOOK OF DUTY

Informal description (often using text)

Describe information requirements

- Objects used (e.g., student, professor, lecture)
- Domains of attributes of objects
- Identifiers, references / relationships

Describe processing requirements

- Cardinalities: how many students?
- Distributions: skew of lecture attendance
- Workload: read-write/write-heady
- Priorities and service level agreements

EXAMPLE: BOOK OF DUTY

Students have a student ID, first and last name, DOB, …

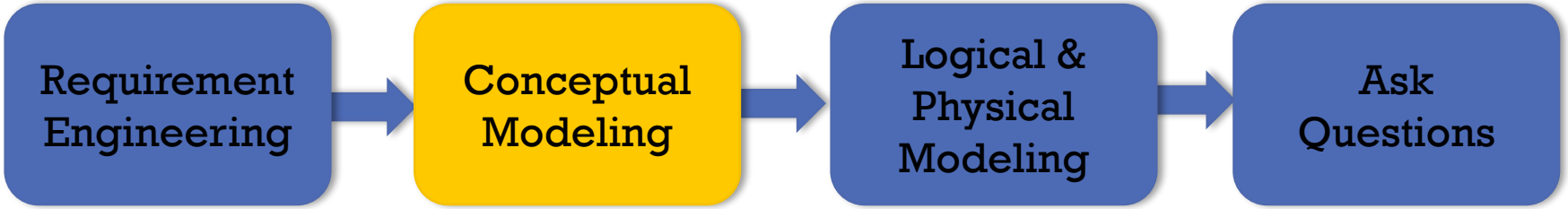
Student IDs are unique

Students can attend multiple courses

Courses can be attended by maximally 100 students

…

DATABASES FOR DATA SCIENTIST



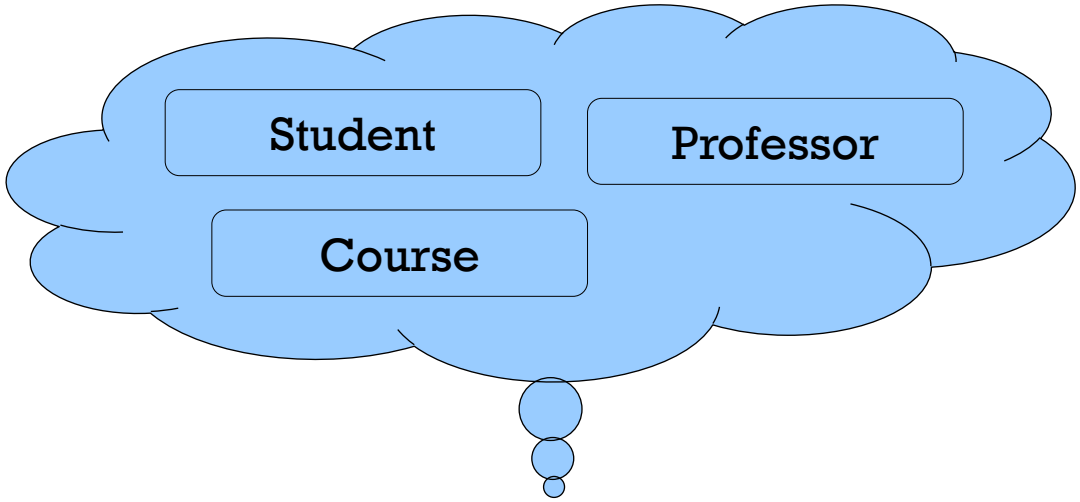
Book of duty

Conceptual Design (ER)

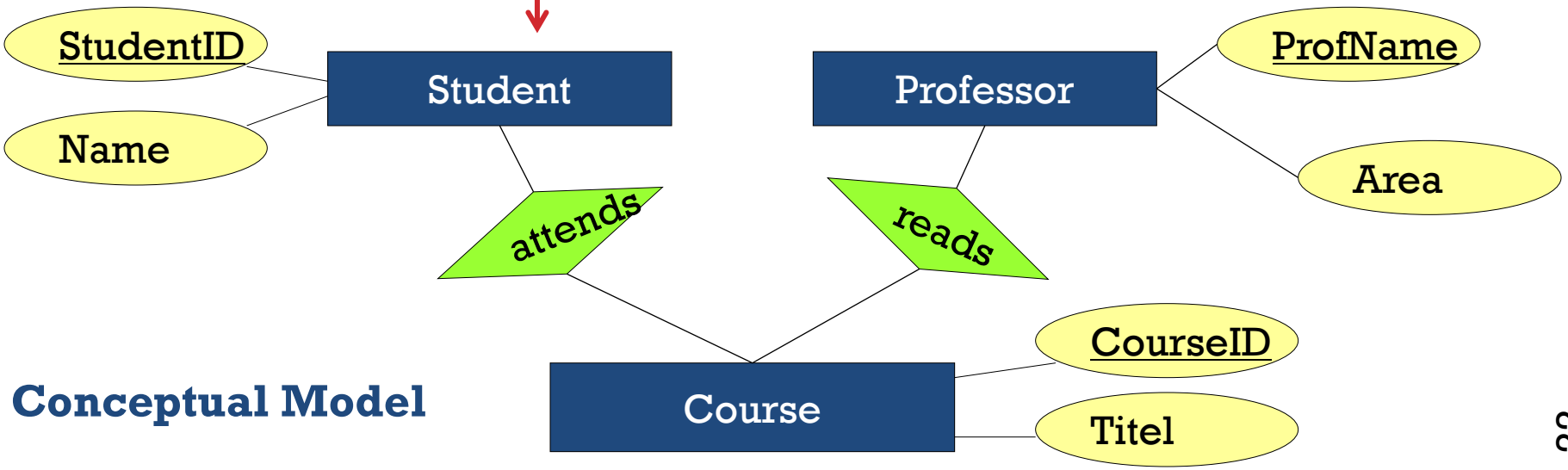
- Logical design (schema)
- Physical design (index, layout)

- Relational Algebra
- SQL

CONCEPTUAL MODEL

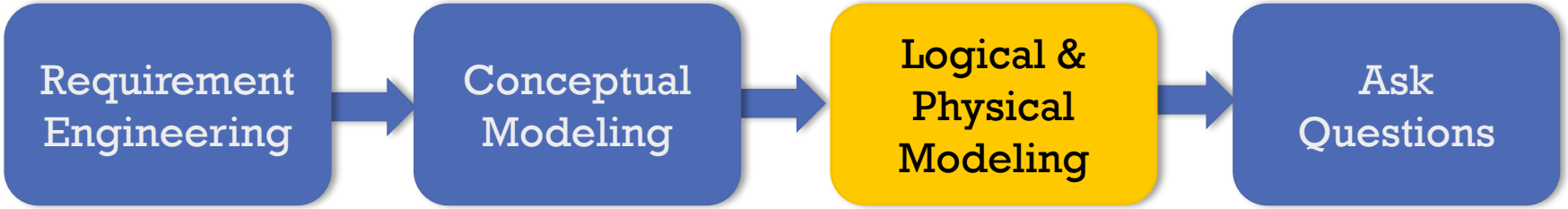


Book of Duty



Conceptual Model

DATABASES FOR DATA SCIENTIST



Book of duty

Conceptual Design (ER)

- Logical design (schema)
- Physical design (index, hints)

- Relational Algebra
- SQL

LOGICAL AND PHYSICAL DESIGN

Student

StudentID	Name
1	Thran
2	Lyons
…	…

attends

StudentID	CourseID
1	CS22
2	CS123
…	…

Course

CourseID	Title
CS22	DB
CS123	ML
…	…

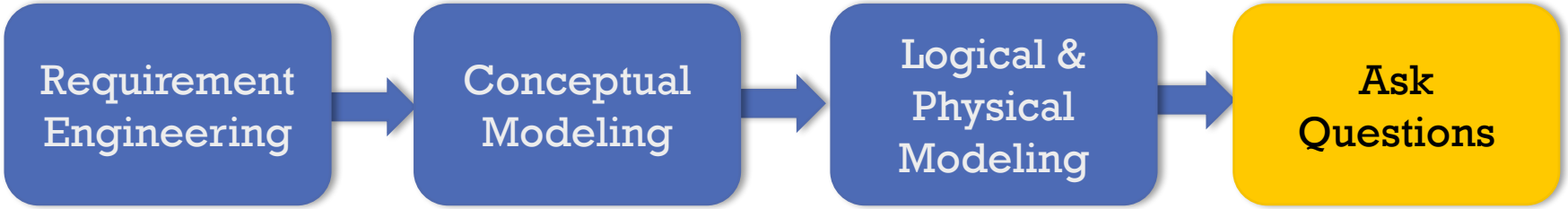
Logical Design

- Table / column names
- Data types
- Constraints
- …

Physical Design

- Indexes to speed up retrieval
- Memory layout
- Compression
- …

DATABASES FOR DATA SCIENTIST



Book of duty

Conceptual Design (ER)

- Logical design (relational schema)
- Physical design (index, hints)

- Relational Algebra
- SQL

ASK QUESTIONS

How many students take the course CS123?

Student

StudentID	Name
1	Thran
2	Lyons
3	Bertsch
...	...

attends

StudentID	CourseID
1	CS22
2	CS123
3	CS123
...	...

SQL Query:

```
SELECT COUNT(*)
FROM Student s, attends a
WHERE s.StudentID=a.StudentID
AND a.CourseID='CS123'
```

OUTLINE FOR NEXT LECTURES

ER-Models

Relational Model

Relational Algebra

SQL