SCIENCE
PASSION
TECHNOLOGY

Data Management
14 Q&A and Exam Preparation

Matthias Boehm

Graz University of Technology, Austria
Computer Science and Biomedical Engineering
Institute of Interactive Systems and Data Science
BMVIT endowed chair for Data Management

Last update: Jan 20, 2019

Exam Preparation

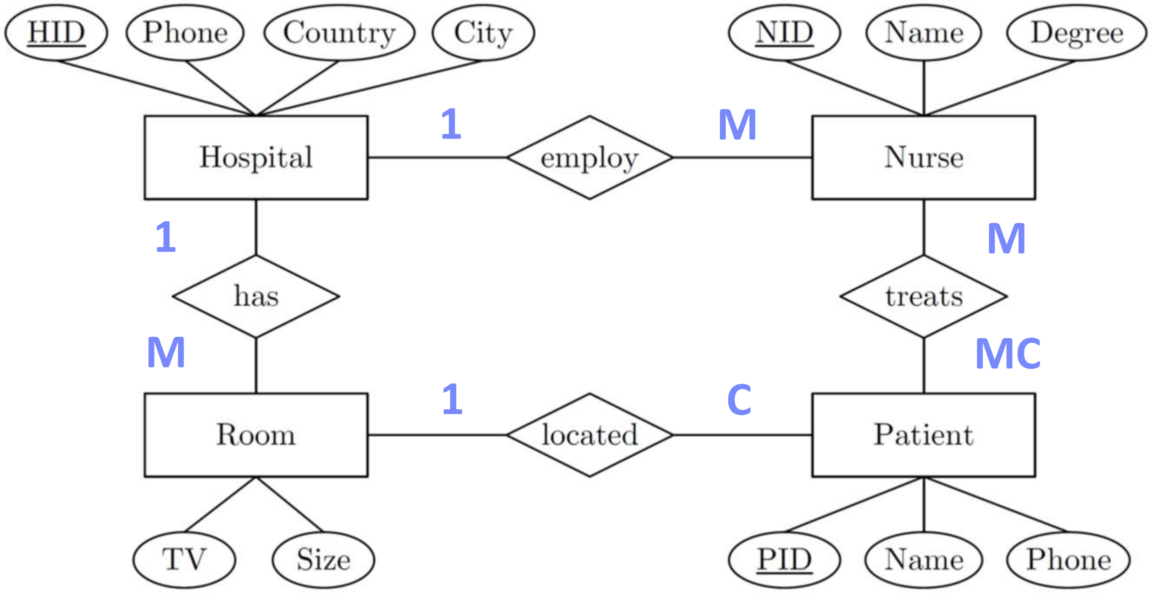
Basic focus: fundamental concepts and ability to apply learned techniques to given problems

Exam Logistics

- **Timing**
  - Exam starts 10min after official start
  - 90min working time (plenty of time to think about answers)
  - **Write into the worksheet if possible**, additional paper allowed
  - Grading will happen Feb 1 / Feb 2 → use exam Feb 6 as replacement
- **Covered Content**
  - **Must-have:** Data modeling/normalization, SQL query processing
  - Relational algebra, physical design, query and transaction processing
  - NoSQL, distributed storage and computation, streaming
- **Past Exams**
  - **3x Data Management** (previously known as Databases)
  - **3x Databases** (previously known as Databases 1)
  - https://mboehm7.github.io/teaching/ss19_dbs/index.htm

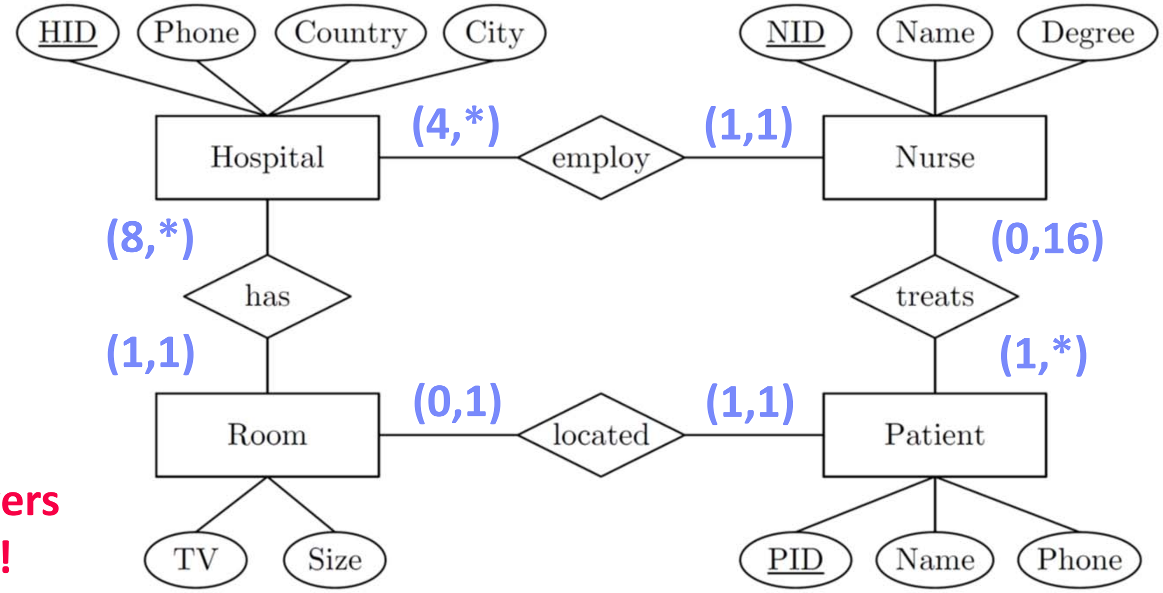
#1 Data Modeling

- **Task 1a: Specify the cardinalities in Modified Chen notation** (8 Points)
  - A hospital employs at least 4 nurses and has at least 8 patient rooms.
  - A nurse works in exactly one hospital and treats up to 16 patients.
  - A patient is treated by at least one but potentially many nurses.
  - Every patient has a room, a rooms belongs to exactly one hospital, and rooms are never shared by multiple patients.



#1 Data Modeling, cont.

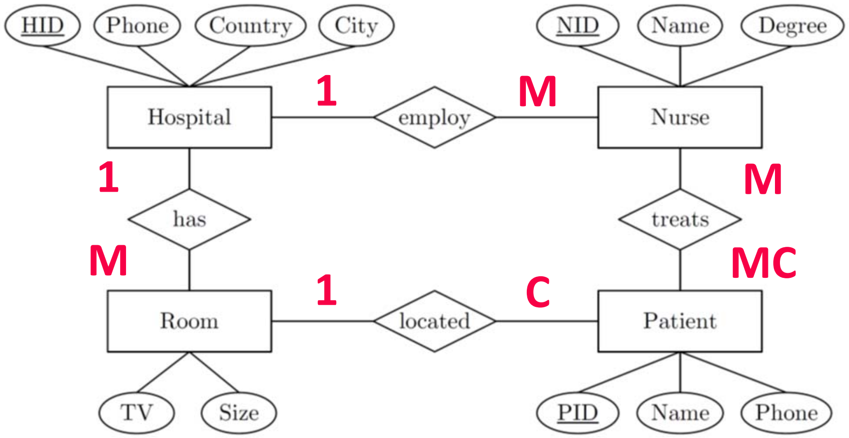
- **Task 1b: Specify the cardinalities in (min, max) notation** (4 Points)
  - A hospital employs at least 4 nurses and has at least 8 patient rooms.
  - A nurse works in exactly one hospital and treats up to 16 patients.
  - A patient is treated by at least one but potentially many nurses.
  - Every patient has a room, a rooms belongs to exactly one hospital, and rooms are never shared by multiple patients.



Only provide answers you're asked for!

#1 Data Modeling, cont.

- **Task 1c: Map the given ER diagram into a relational schema (10 points)**
  - Including data types, primary keys, and foreign keys



- **Solution**
  - Hospitals(
    HID:int, phone:char(16), Country:varchar(64), City:varchar(64))
  - Nurses(
    NID:int, Name:varchar(64), Degree:varchar(32), HID^{FK}:int)
  - Patient(
    PID:int, Name:varchar(64), Phone:char(16), RID^{FK}:int)
  - Room(
    RID:int, TV:boolean, Size:int, HID^{FK}:int)
  - Treated(
    NID^{FK}:int, PID^{FK}:int)

#1 Data Modeling, cont.

- **Task 1d: Bring your schema in 3rd normal form and explain why it is in 3NF (12 points)**
  - Let Hospital.Phone and Patient.Phone be multi-valued attributes
  - Assume the functional dependency City → Country

- **Solution**
  - Phones(Number:char(16), HID^{FK}:int, PID^{FK}:int)
  - Cities(City:varchar(64), Country:varchar(64))
  - Hospitals(HID:int, City^{FK}:varchar(64))

  - **1st Normal Form:** no multi-valued attributes
  - **2nd Normal Form:** 1NF + all non-key attributes fully functional dependent on PK
  - **3rd Normal Form:** 2NF + no dependencies among non-key attributes

#2 Structured Query Language

- **Task 2a: Compute the results for the following queries (15 points)**

Orders

OID	Customer	Date	Quantity	PID
1	A	‘2019-06-25’	3	2
2	B	‘2019-06-25’	1	3
3	A	‘2019-06-25’	1	4
4	C	‘2019-06-26’	2	2
5	D	‘2019-06-26’	1	4
6	C	‘2019-06-26’	1	1

Products

PID	Name	Price
1	X	100
2	Y	15
4	Z	75
3	W	120

```
Q1: SELECT DISTINCT Customer, Date
      FROM Orders O, Products P
      WHERE O.PID = P.PID AND Name IN('Y','Z')
```

Customer	Date
A	‘2019-06-25’
C	‘2019-06-26’
D	‘2019-06-26’

```
Q2: SELECT Customer, count(*) FROM Orders
      GROUP BY Customer
      ORDER BY count(*) DESC, Customer ASC
```

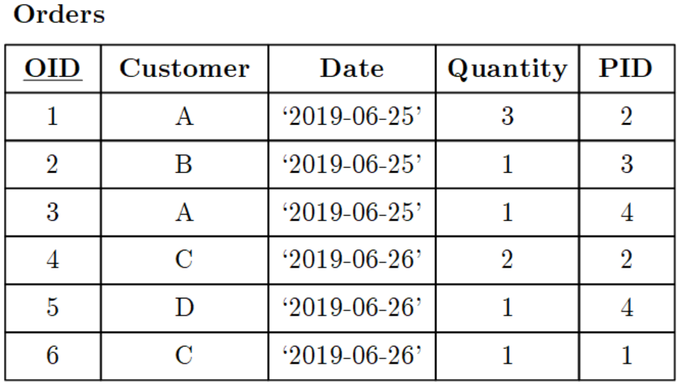
Customer	Sum
A	2
C	2
B	1
D	1

```
Q3: SELECT Customer, sum(O.Quantity * P.Price)
      FROM Orders O, Products P
      WHERE O.PID = P.PID
      GROUP BY Customer
```

Customer	Sum
A	120
B	120
C	130
D	75

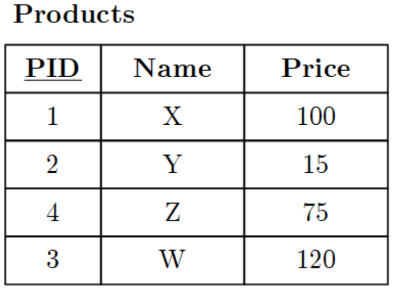
#2 Structured Query Language, cont.

- **Task 2b: Write SQL queries to answer the following Qs** (15 points)



Orders

OID	Customer	Date	Quantity	PID
1	A	‘2019-06-25’	3	2
2	B	‘2019-06-25’	1	3
3	A	‘2019-06-25’	1	4
4	C	‘2019-06-26’	2	2
5	D	‘2019-06-26’	1	4
6	C	‘2019-06-26’	1	1



Products

PID	Name	Price
1	X	100
2	Y	15
4	Z	75
3	W	120

Q4: Which products where bought on 2019-06-25 (return the distinct product names)?

```
SELECT DISTINCT P.Name
  FROM Orders O, Products P
  WHERE O.PID = P.PID
    AND Date = ‘2019-06-25’
```

Q5: Which customers placed only one order?

```
SELECT Customer FROM Orders
  GROUP BY Customer HAVING count(*) = 1
```

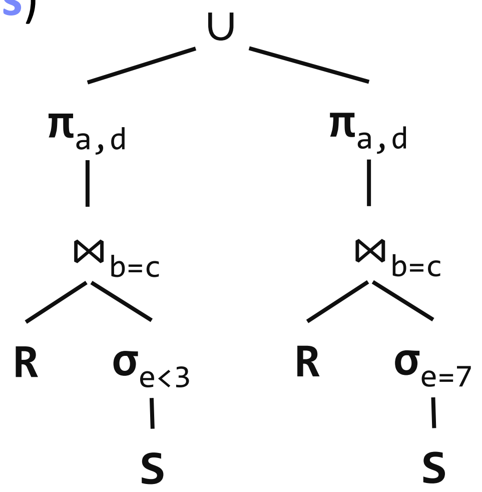
Q6: How much revenue (sum(O.Quantity * P.Price)) did products with a price less then 90 generate (return (product name, revenue))?

```
SELECT P.Name, sum(O.Quantity * P.Price)
  FROM Orders O, Products P
  WHERE O.PID = P.PID AND Price < 90
  GROUP BY P.Name
```

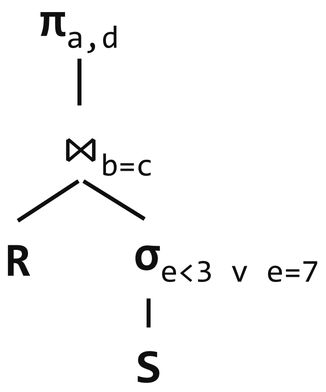
#3 Query Processing

- **Task 3a: Assume tables R(a,b), and S(c,d,e), draw a logical query tree in relational algebra for the following query: (5 points)**

```
Q7: SELECT R.a, S.d FROM R, S
      WHERE R.b = S.c AND S.e < 3
    UNION ALL
    SELECT R.a, S.d FROM R, S
      WHERE R.b = S.c AND S.e = 7
```

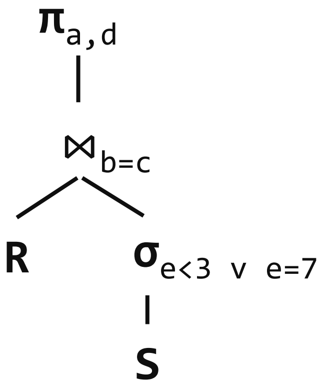


- **Task 3b: Draw an optimized logical query tree for the above query in relational algebra by eliminating the union operation (3 points)**

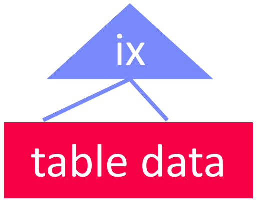


#3 Query Processing, cont.

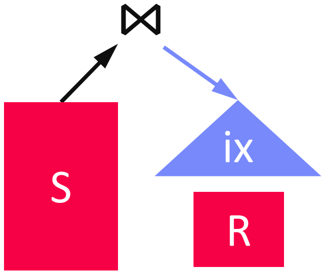
- **Task 3c: Given the schema and query above, which attribute or attributes are good candidates for secondary indexes and how could they be exploited during query processing?** (4 points)



- **Solution**
  - S.e → index scan
    (lookup e=7,
    lookup e=3 and scan DESC)



  - R.b (or S.c) → index nested loop join
    (for every S tuple s, loopup s.c in IX)



#3 Query Processing, cont.

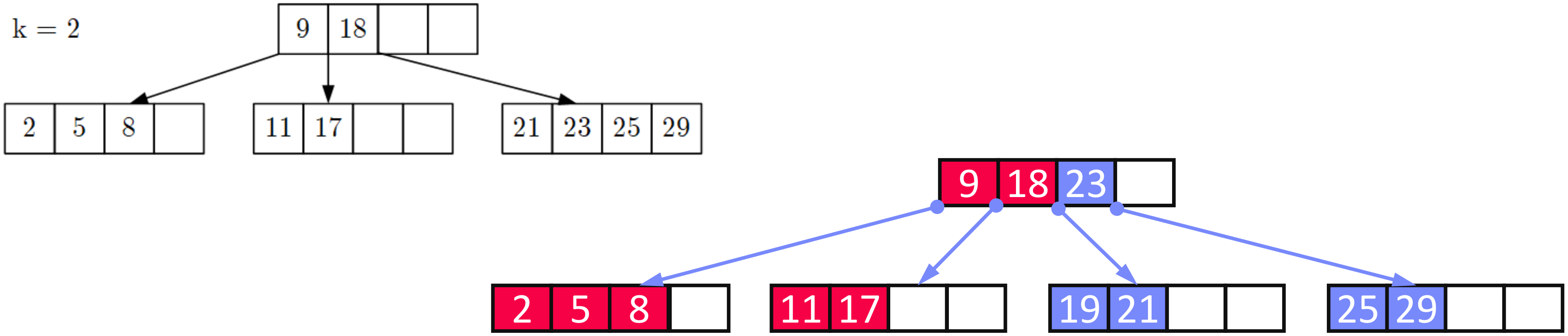
- **Task 3d: Describe the volcano (open-next-close) iterator model by example of a selection operator and discuss the space complexity of this selection operator.** (6 points)

- **Solution**
  - Open, next, close calls propagate from root to leafs
  - Open: operator initialization
  - Next: compute next tuple (selection: call next of input until next qualifying tuple found)
  - Close: cleanup resources
  - **Space complexity:** O(1)

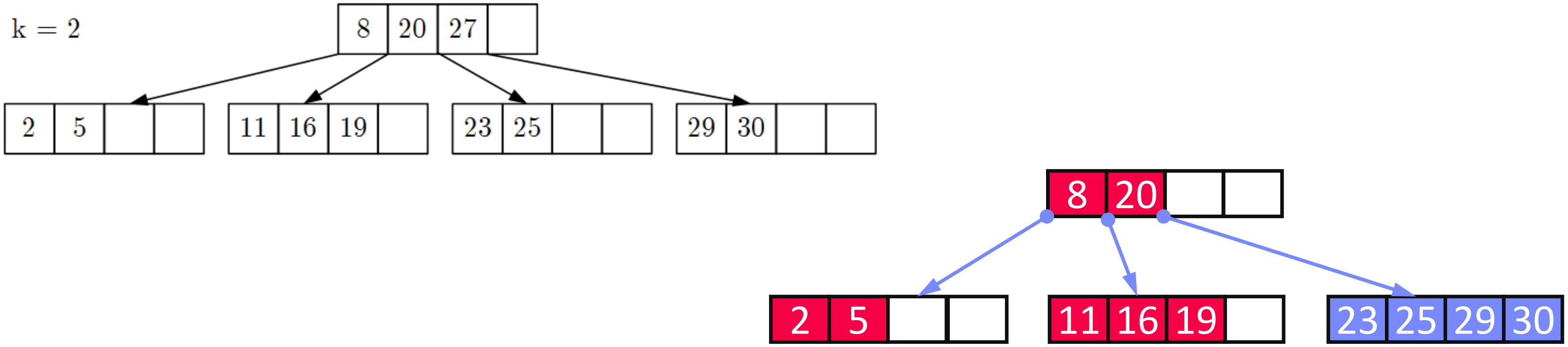
```
void open() { R.open(); }

void close() { R.close(); }

Record next() {
  while( (r = R.next()) != EOF )
    if( p(r) ) //A==7
      return r;
  return EOF;
}
```

#4 Physical Design – B-Trees

- **Task 4a: Given B-tree, insert key 19 and draw resulting B-tree (7 points)**



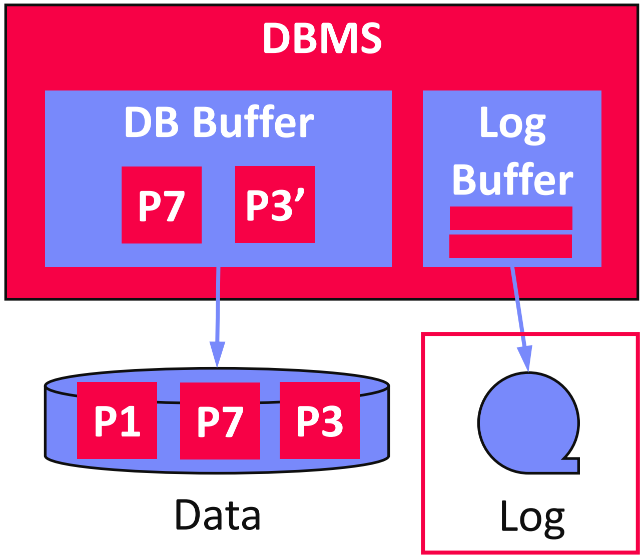
- **Task 4b: Given B-tree, delete key 27, and draw resulting B-tree (8 points)**



#5 Transaction Processing

- **Task 5a: Describe the concept of a database transaction log, and explain how it relates to the ACID properties Atomicity and Durability** (7 points)

- **Solution**
  - Log: append-only TX changes, often on separate devices
  - Write-ahead logging (log written before DB, forced-log on commit)
  - Recovery: forward (REDO) and backward (UNDO) processing

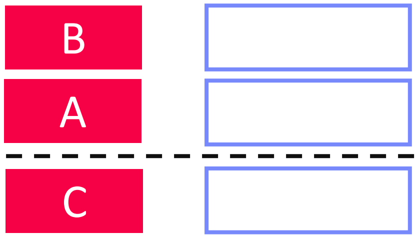


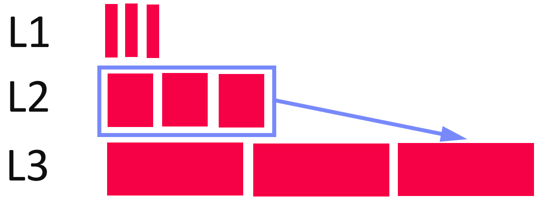
  - **#1 Atomicity:** A TX is executed atomically (**completely or not at all**); on failure/aborts no changes in DB (**UNDO**)
  - **#2 Durability: Guaranteed persistence** of changes of successful TXs; in case of system failures, the database is recoverable (**REDO**)

#6 NoSQL

- **Task 6a:** Describe the concept and system architecture of a **key-value store**, including techniques for achieving **high write throughput**, and **scale-out** in distributed environments. Please focus specifically on aspects of physical design such as **index structures**, and **distributed data storage**. (10 points)

- **Solution**
  - KV store: simple map of key-value pairs, w/ get/put interface, often distributed
  - Index structure for high write throughput: Log-structured merge trees (LSM)
  - Distributed data storage for scale-out: horizontal partitioning (sharding) via hash or range partitioning, partitioning via selection, reconstruction via union eventual consistency for high availability and partition tolerance





Remaining Questions & Answers

Course Content

Data Management in general

Conclusions and Q&A

- **Summary**
  - **13 Data Stream Processing Systems**
  - **14 Q&A and Exam Preparation**

- **Next Week: NO lecture** (use time for exam prep)
  - Office hours Mo 3pm as usual

- **Exams**
  - **Jan 30, 5.30pm Exam** DM VO / DB VU, HS i13
  - **Jan 31, 5.30pm Exam** DM VO / DB VU, HS i13
  - **Feb 6, 4pm Exam** DM VO / DB VU, HS i13 (also as replacement exam)