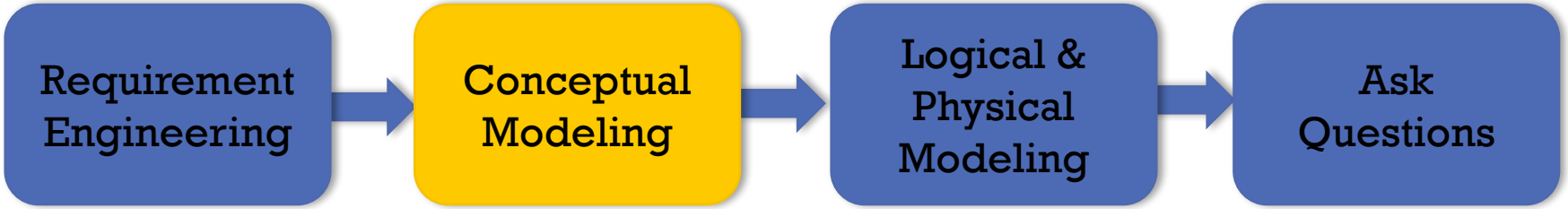
ER-MODEL

INTRODUCTION TO DATA SCIENCE

CARSTEN BINNIG

BROWN UNIVERSITY

DATABASES FOR DATA SCIENTIST



Book of duty

Conceptual Design (ER)

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

- Relational Algebra
- SQL

ENTITY/RELATIONSHIP MODEL

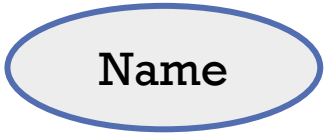
MODELING ELEMENTS

ENTITY/RELATIONSHIP (ER) MODEL

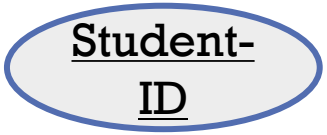
Entity

Relationship

Attribute



Key



Role

Other Notations exist!!!

ENTITY/RELATIONSHIP (ER) MODEL

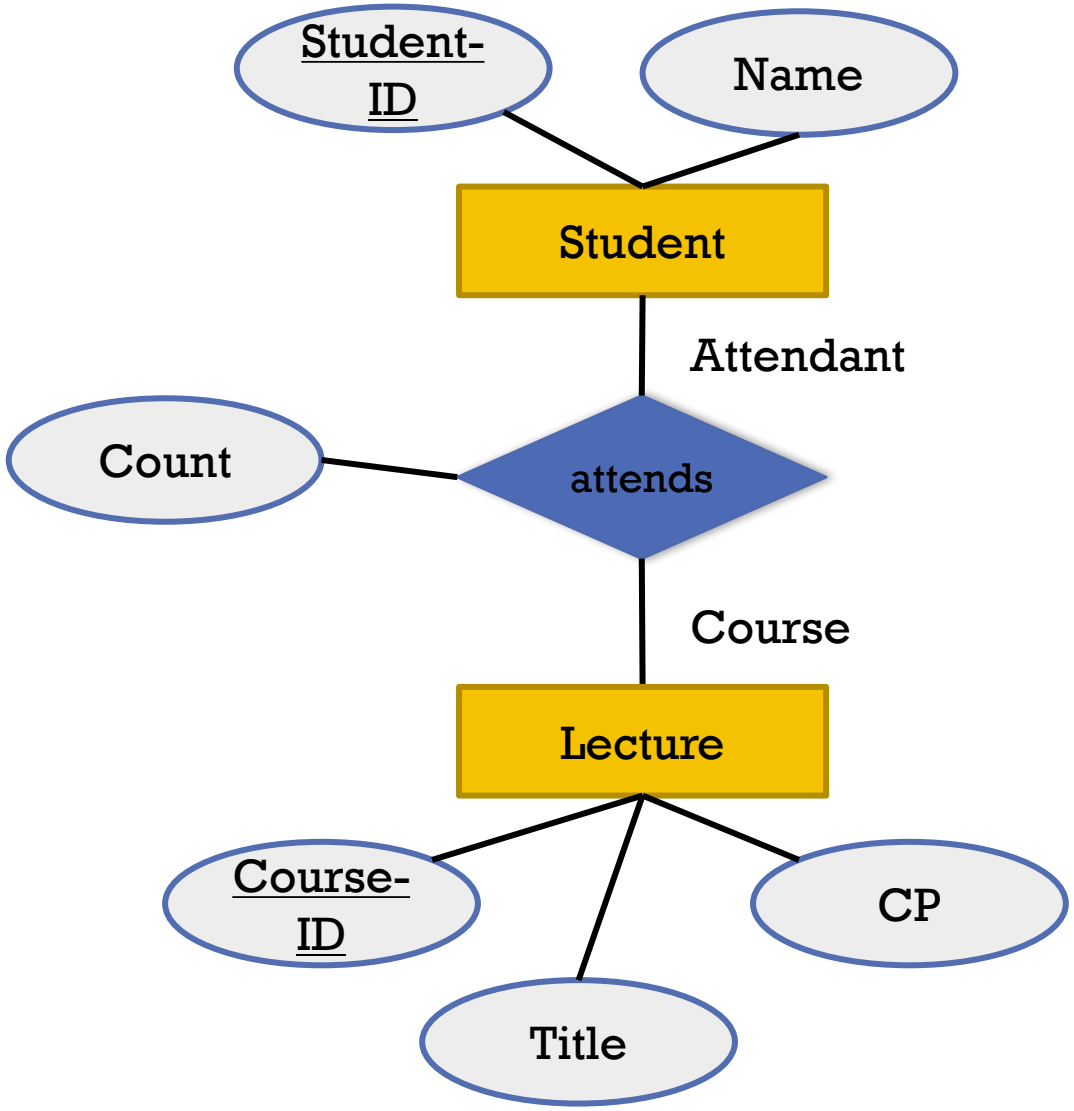
Entity

Relationship

Attribute

Key

Role



WHY ER?

Advantages

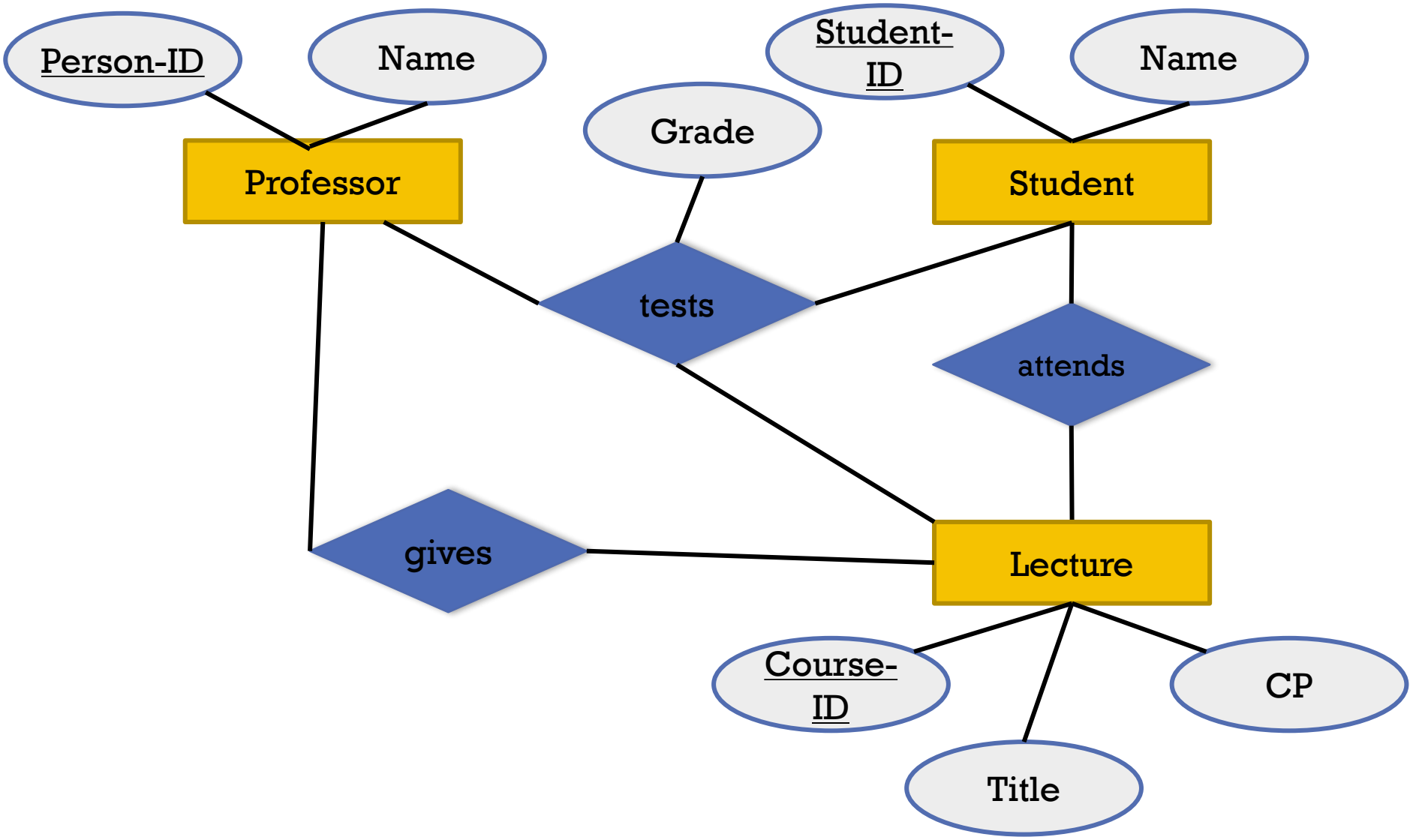
- ER diagrams are easy to create / edit/ read (from the layman)
- ER diagrams express all information requirements independent of concrete implementation (e.g., relational, XML, …)

Other aspects

- Minimality
- Tools (e.g., Visio)
- Graphical representation

General

- Try to be concise, complete, comprehensible, and correct
- Controversy whether ER/UML is useful in practice
- No controversy that everybody needs to learn ER/UML


Person-ID
Name
Student-ID
Name
Grade
Professor
Student
tests
attends
gives
Lecture
Course-ID
CP
Title

RULES OF THUMB

Attribute vs. Entity

- Entity if the concept has more than one relationship
- Attribute if the concept has only one 1:1 relationship

Partitioning of ER Models

- Most realistic models are larger than a page
- Partition by domains (library, research, finances, …)

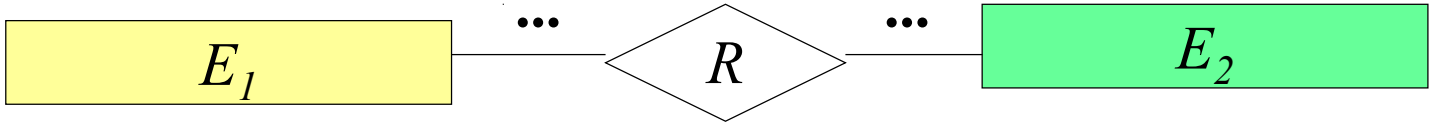
Good vs. Bad models

- Do not model redundancy or tricks to improve performance
- Less entities is better (the fewer, the better!)
- Follow the 5 C's (clear, concise, correct, complete, compliant)

ENTITY/RELATIONSHIP MODEL

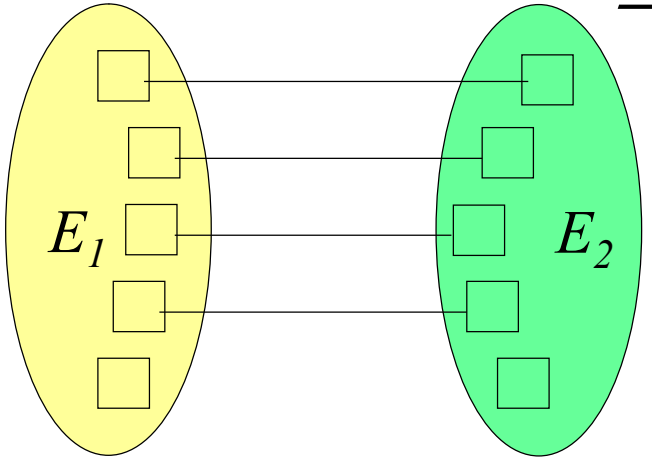
FUNCTIONALITIES

FUNCTIONALITIES

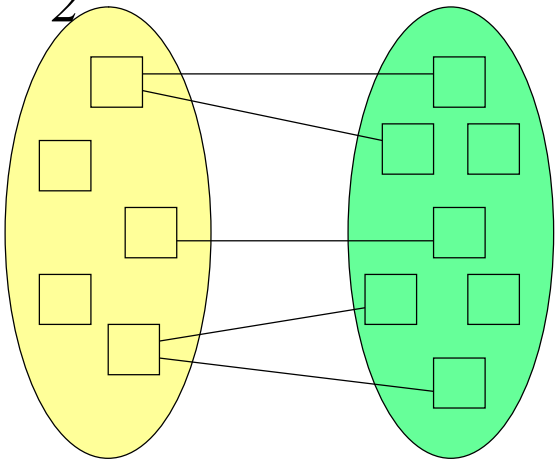


$$R \subseteq E_1 \times E_2$$

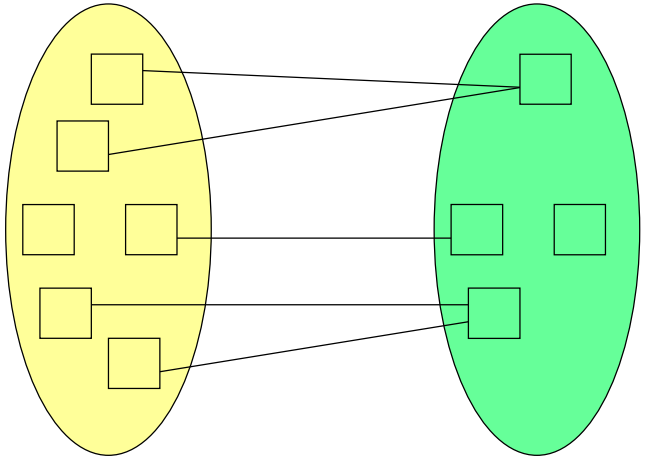
1:1



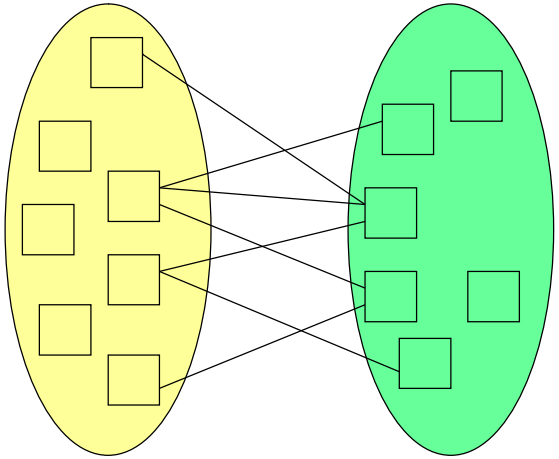
1:N



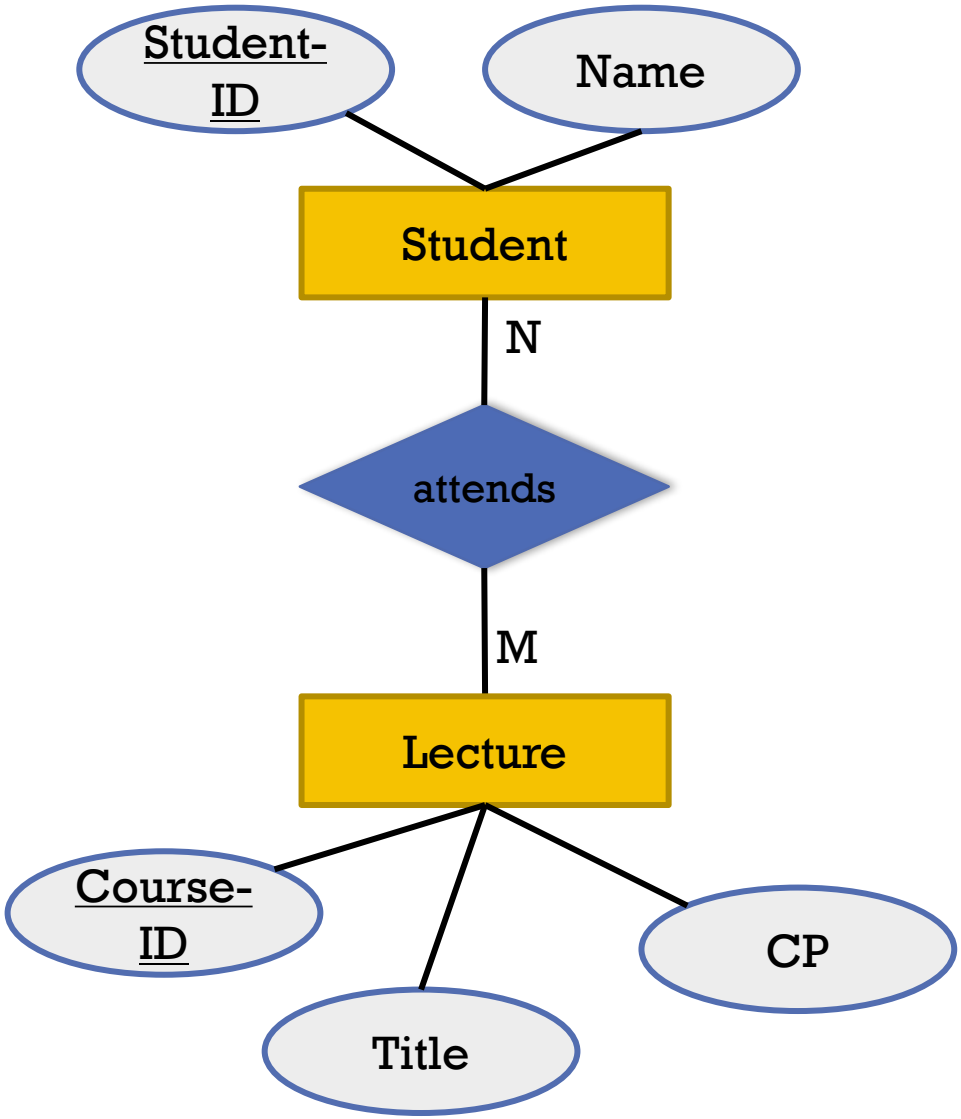
N:1



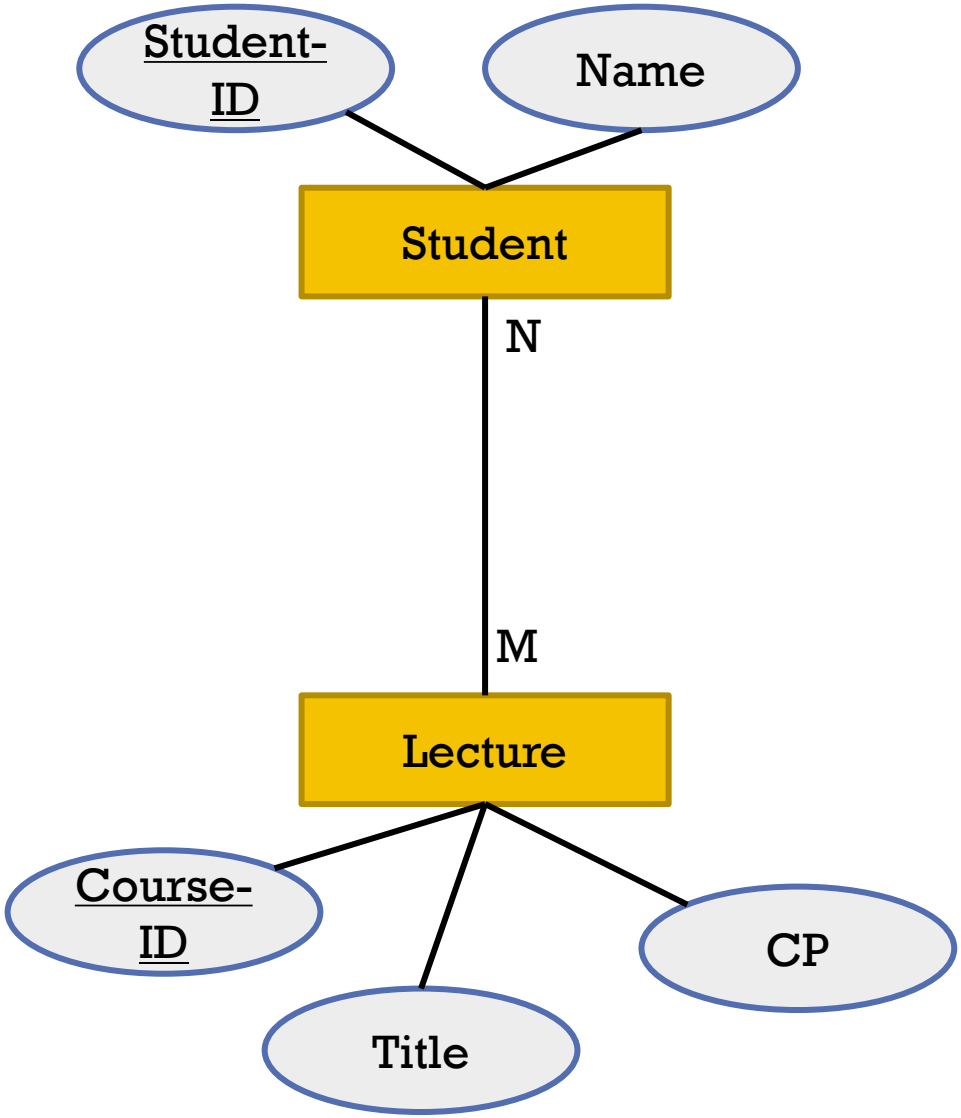
N:M



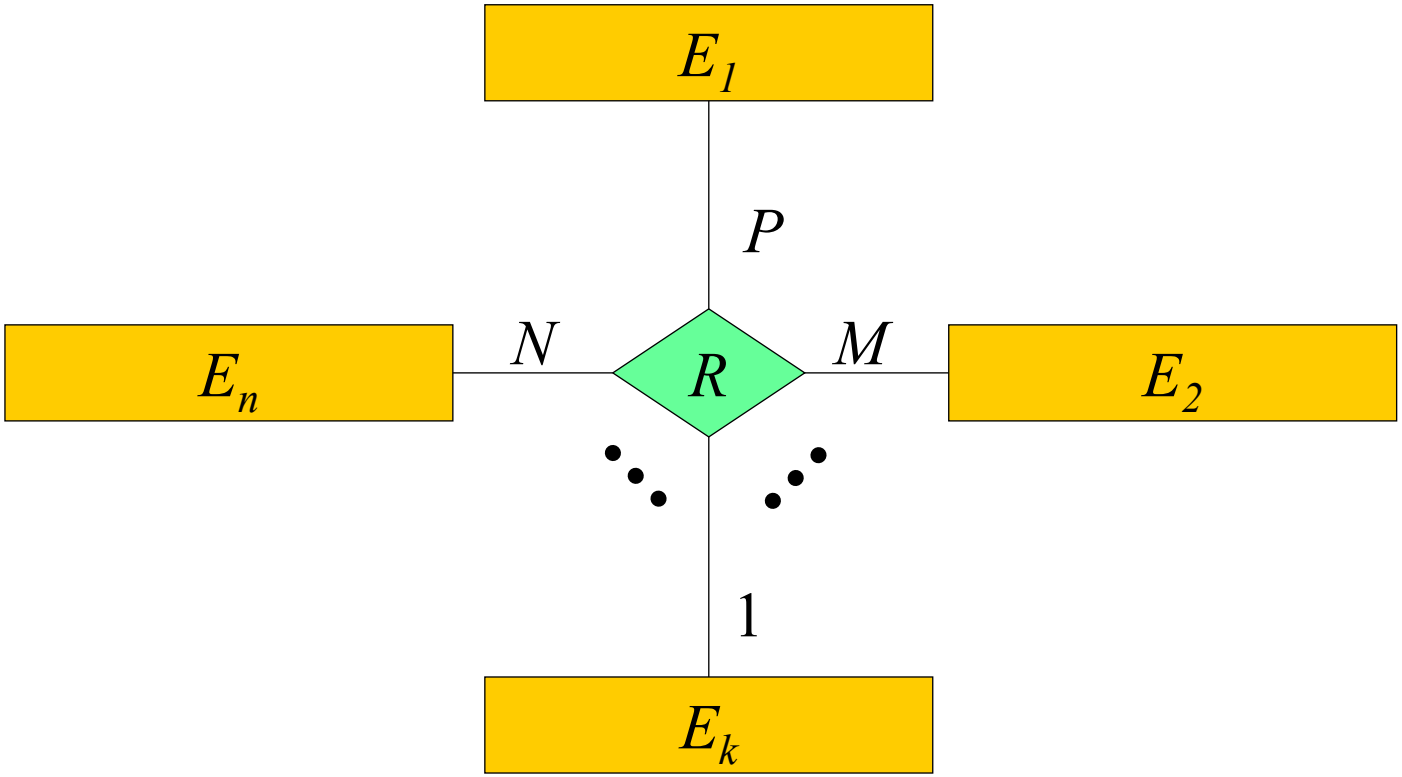
EXAMPLE: PROFESSOR <-> LECTURE



SOMETIMES ALSO SHOWN AS



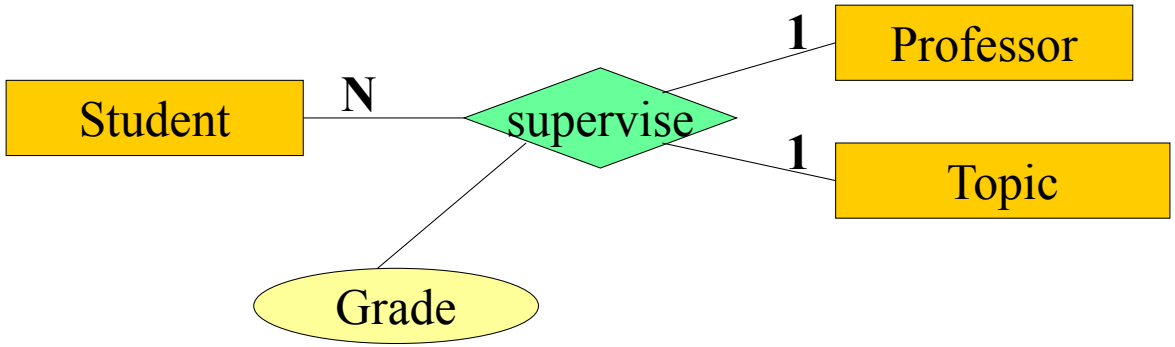
FUNCTIONALITIES OF N-ARY RELATIONSHIPS



$$R : E_1 \text{ x } \ldots \text{ x } E_{k-1} \text{ x } E_{k+1} \text{ x } \ldots \text{ x } E_n \rightarrow E_k$$

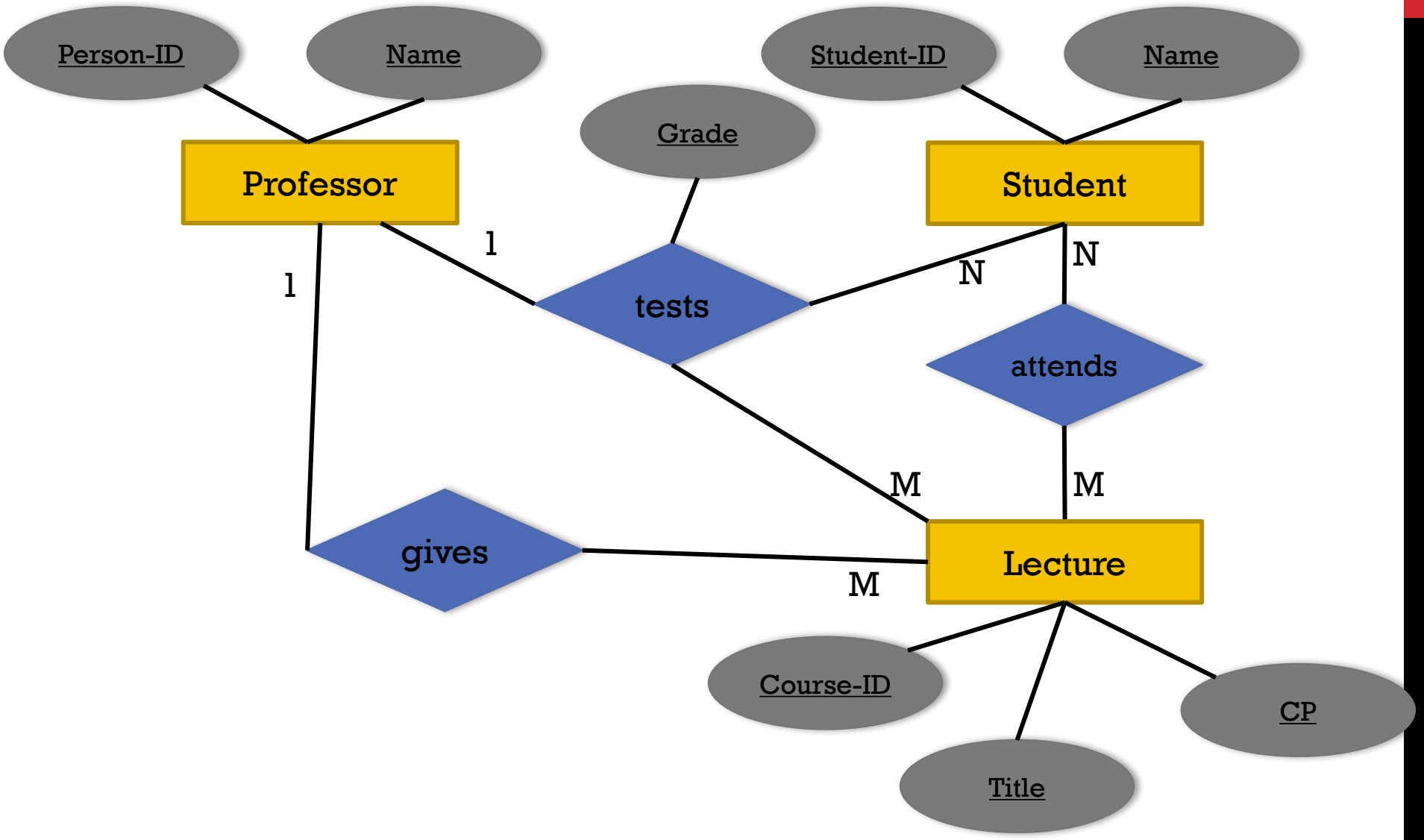
EXAMPLE: RESEARCH PROJECT SUPERVISION

- **Each student can be supervised by the same professor only on one research topic**
- **The student can not use the same research topic with different professors**
- **The same professor can give the same research topic to different students**



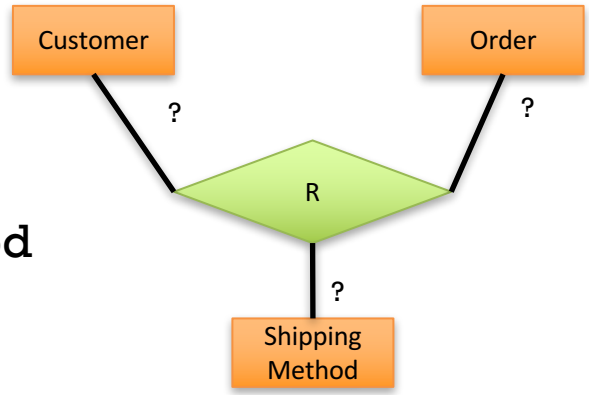
supervise : Professor x Student → Topic

supervise : Topic x Student → Professor

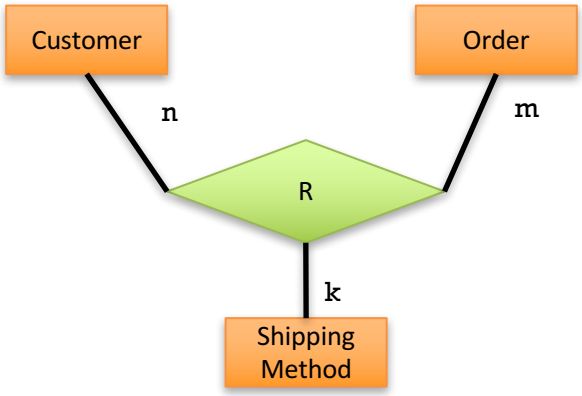

Person-ID
Name
Student-ID
Name
Grade
Professor
Student
1
1
tests
N
N
attends
M
M
gives
Lecture
M
Course-ID
CP
Title

CLICKER QUESTION I

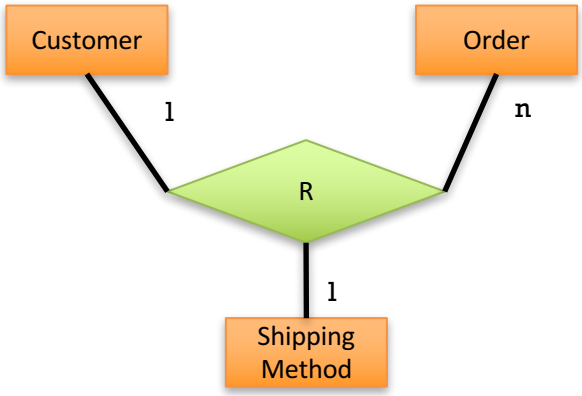
- A customer can have several orders
- An order belongs to a single customer
- Every order has exactly one shipping method (e.g., Post, Fexed, UPS,…)



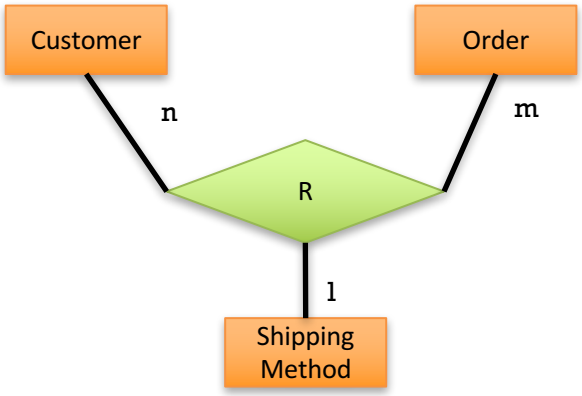
(A)



(B)

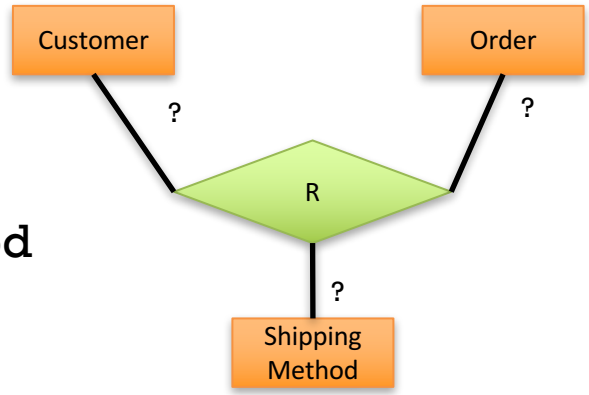


(C)

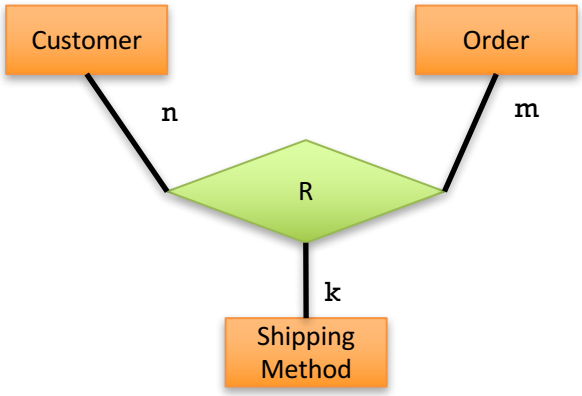


CLICKER QUESTION I

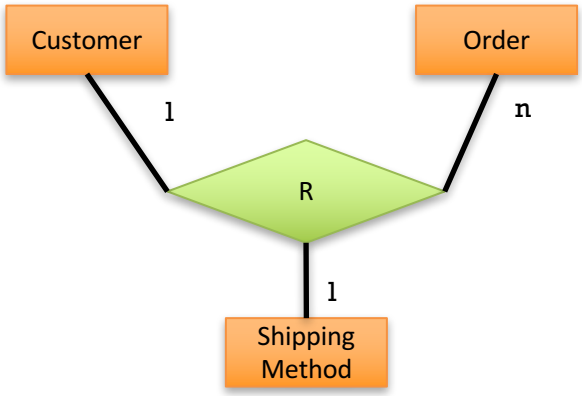
- A customer can have several orders
- An order belongs to a single customer
- Every order has exactly one shipping method (e.g., Post, Fexed, UPS,…)



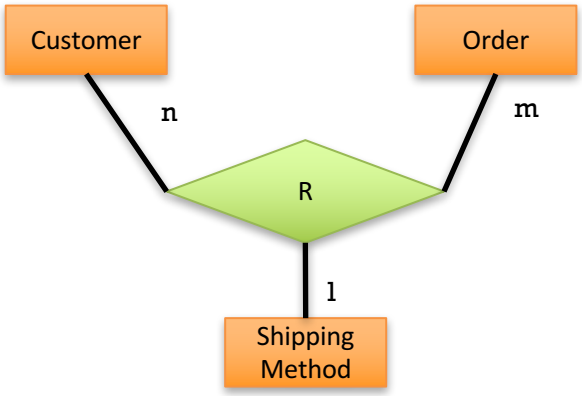
(A)



(B)

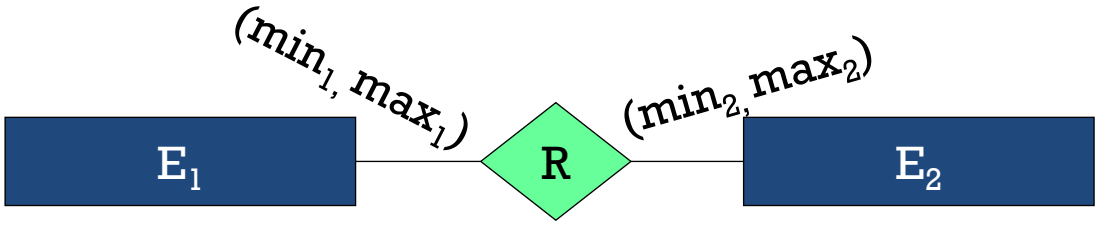


(C)



(MIN,MAX) NOTATION

(MIN, MAX)-NOTATION



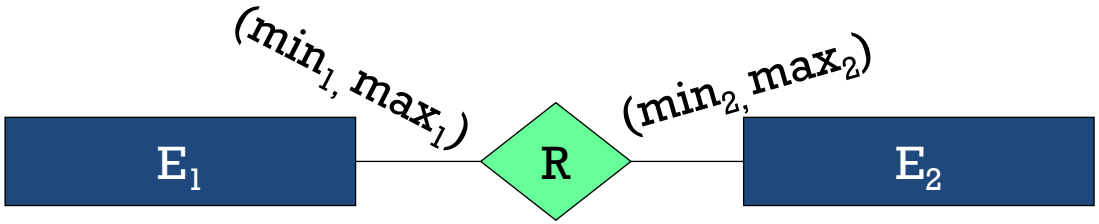
$$R \subseteq E_1 \times E_2$$

(min, max)-Notation is more precise than functionalities:

- **Lower-bound (min)**
- **Upper-bound (max)**

BUT: different definition!

(MIN, MAX)-NOTATION



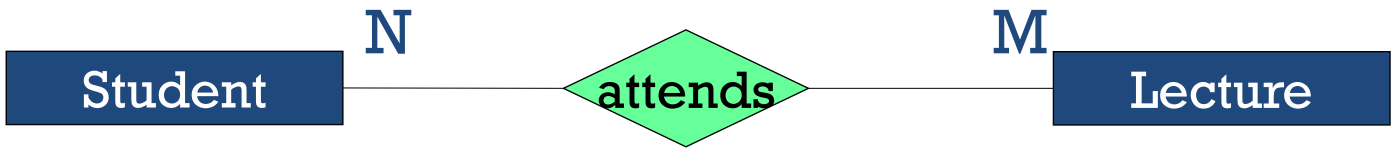
$R \subseteq E_1 \times E_2$

Definition **(min_1, max_1) means that for each $e_1 \in E_1$ there exist**

- At least min_1 relationships with $(e_1, *) \in R$
- Maximally max_1 relationships with $(e_1, *) \in R$

* means **any $e_2 \in E_2$**

CLICKER QUESTION



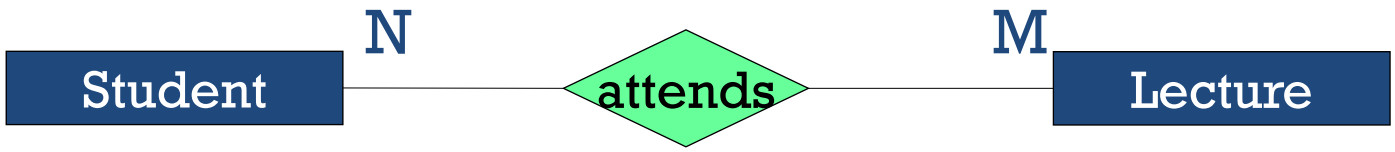
A lecture can be attended by maximally 100 students and a student can not attend more than 20 lectures. What is the correct (min, max) notation to represent these two cases?

A. Student (0,100) -----<attends>----- (0,20) Lecture

B. Student(20,20) -----<attends>----- (100,100) Lecture

C. Student(0,20) -----<attends>----- (0,100) Lecture

CLICKER QUESTION



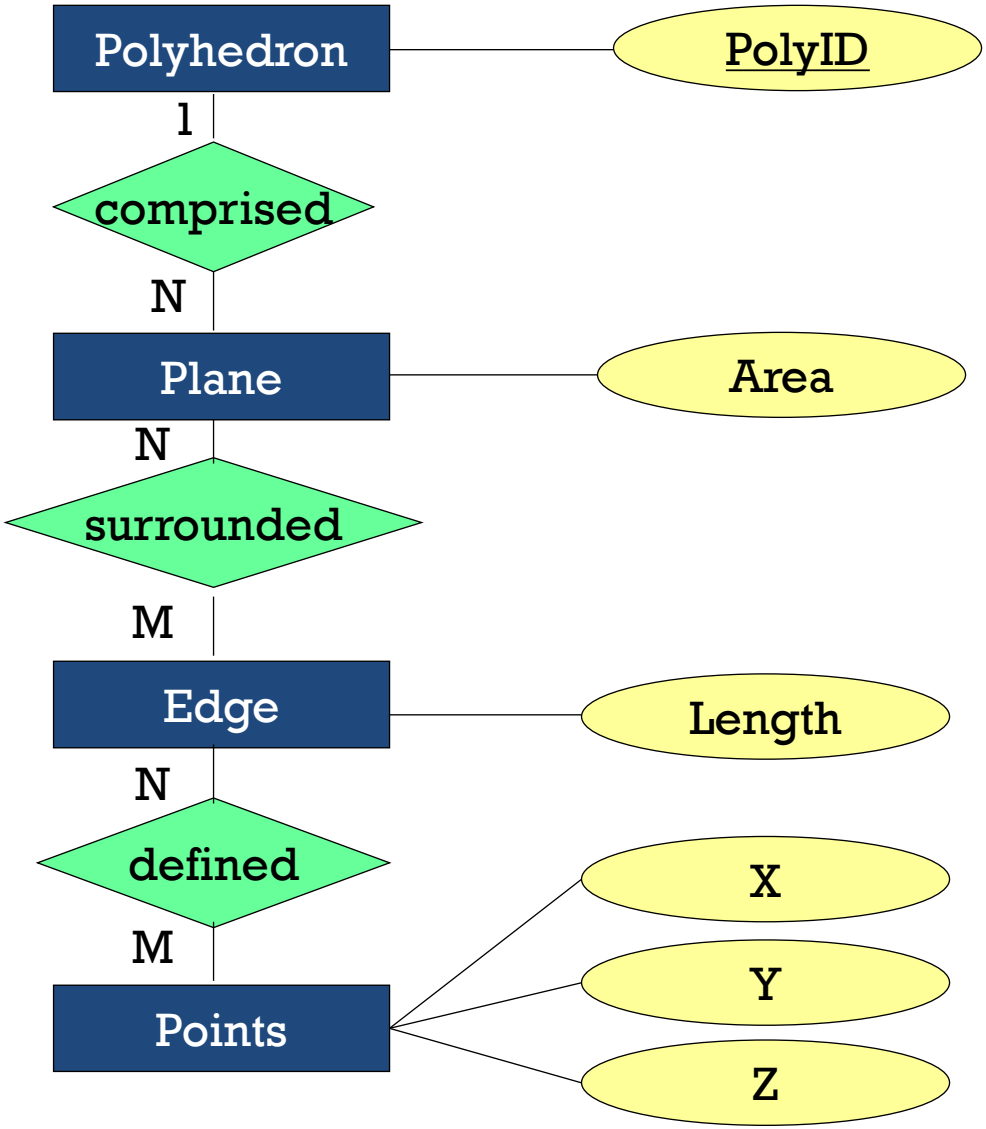
A lecture can be attended by maximally 100 students and a student can not attend more than 20 lectures. What is the correct (min, max) notation to represent these two cases?

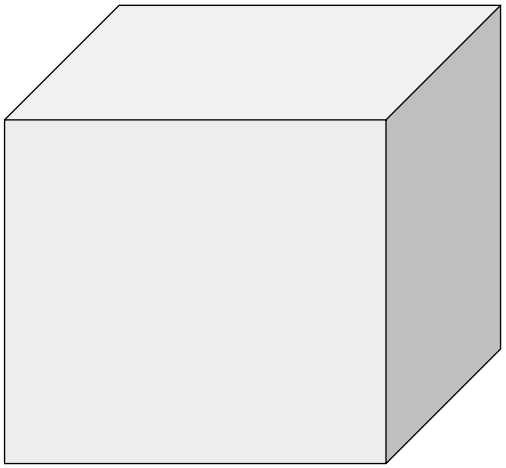
A. Student (0,100) -----<attends>----- (0,20) Lecture

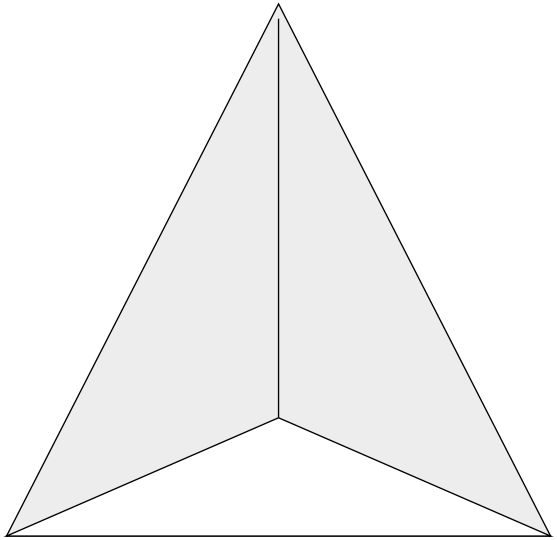
B. Student(20,20) -----<attends>----- (100,100) Lecture

C. Student(0,20) -----<attends>----- (0,100) Lecture

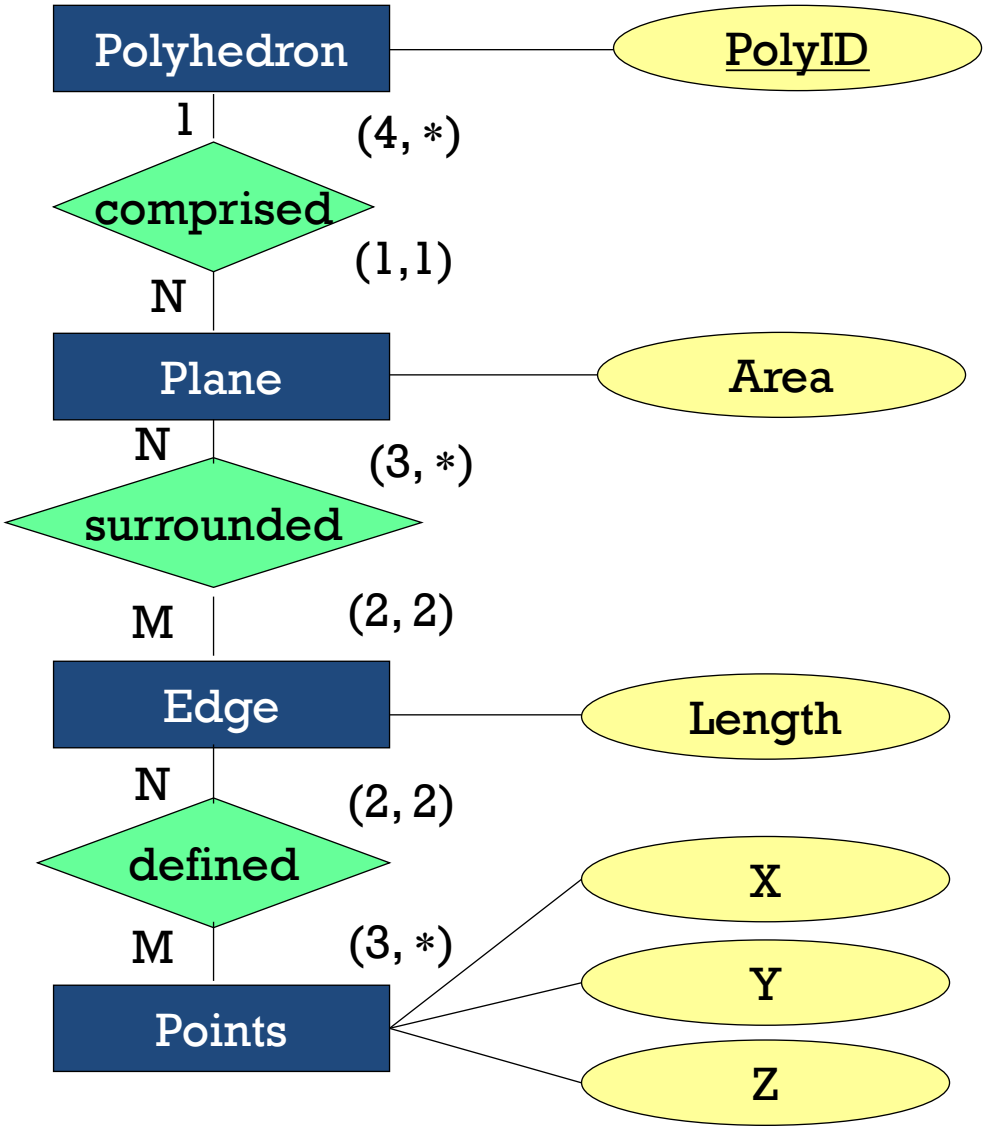
EXAMPLE: POLYHEDRON

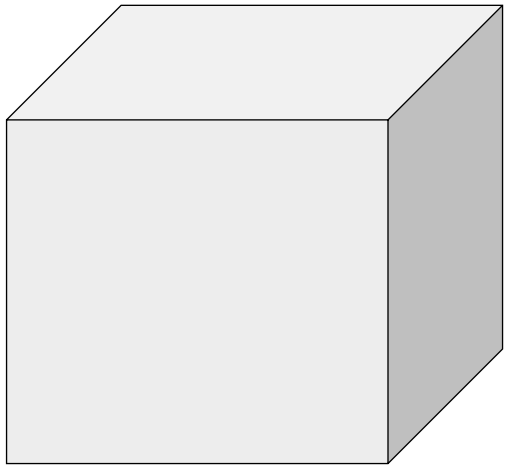


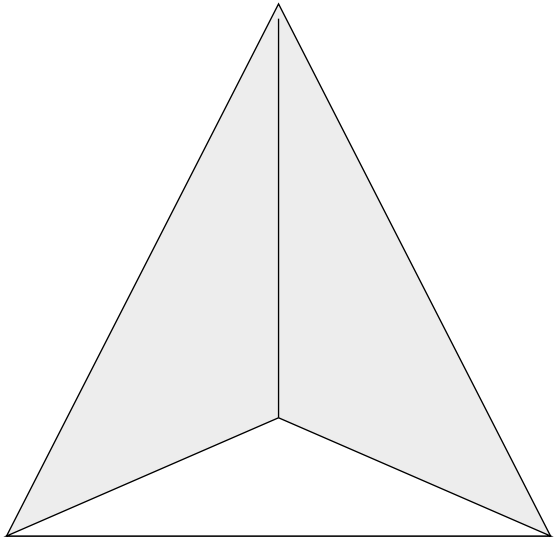




EXAMPLE: POLYHEDRON







NOT COVERED

Weak Entities

Compound attributes

Enhanced ERM

- **Generalization (i.e., inheritance)**
- **Aggregation**

…

CLICKER QUESTION

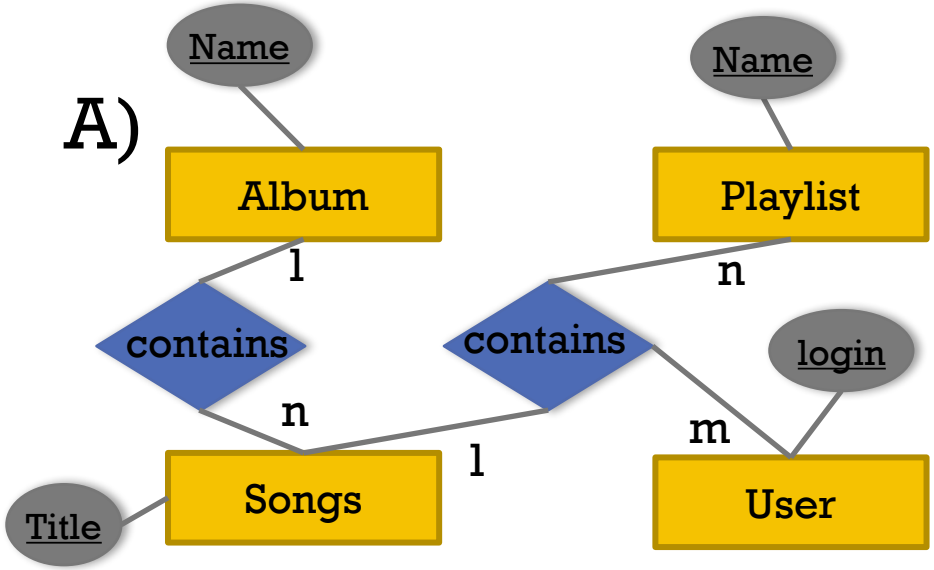
Model a music record database

- An album has a unique name and songs have unique titles
- An album contains several songs
- A playlist has a unique name and is created by one user with a unique login
- A playlist contains several songs from potential different albums

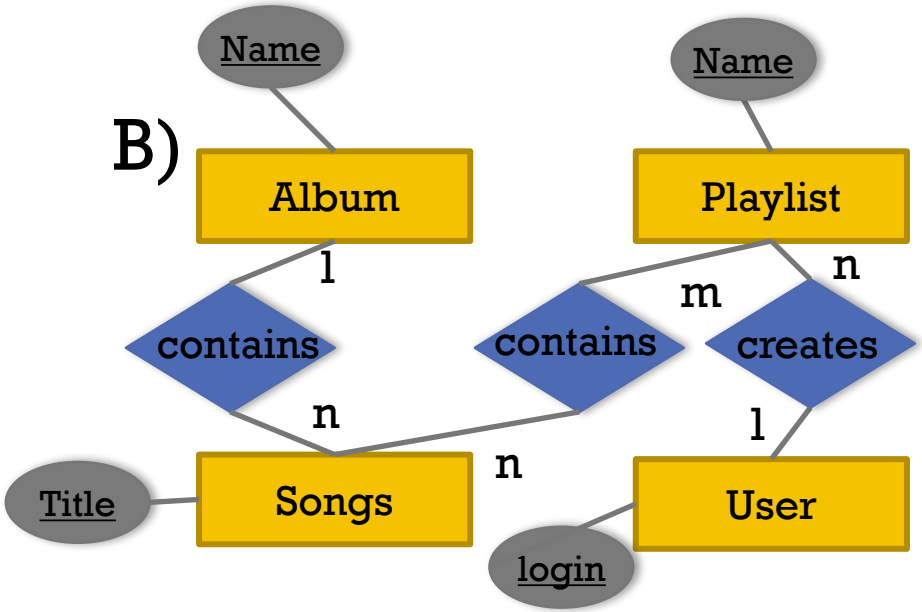
CLICKER QUESTION

- An album has a unique name and songs have unique titles
- An album contains several songs
- A playlist has a unique name and is created by one user with a unique login
- A playlist contains several songs from potential different albums

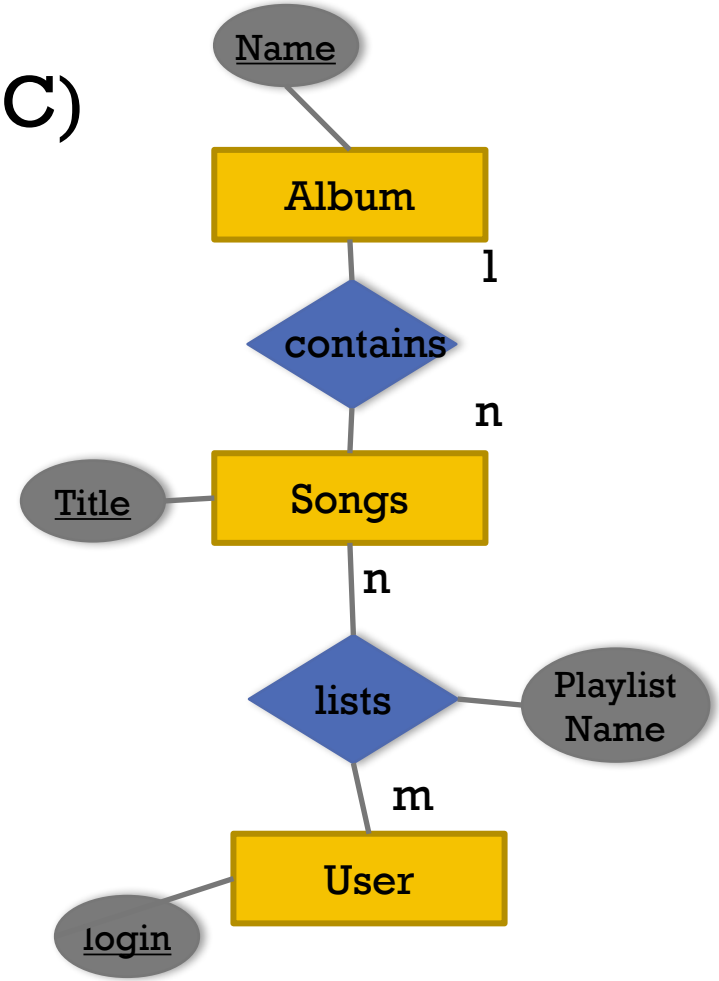
A)



B)



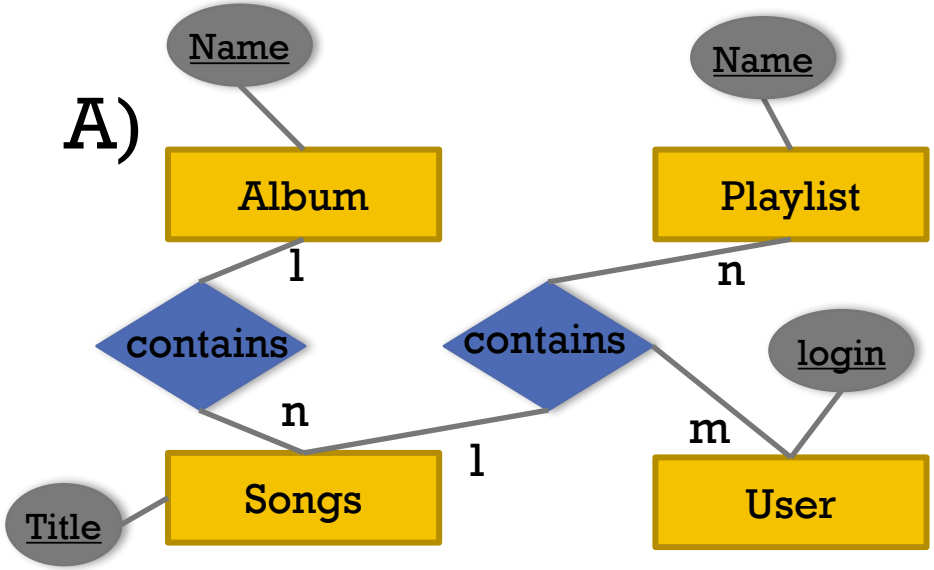
C)



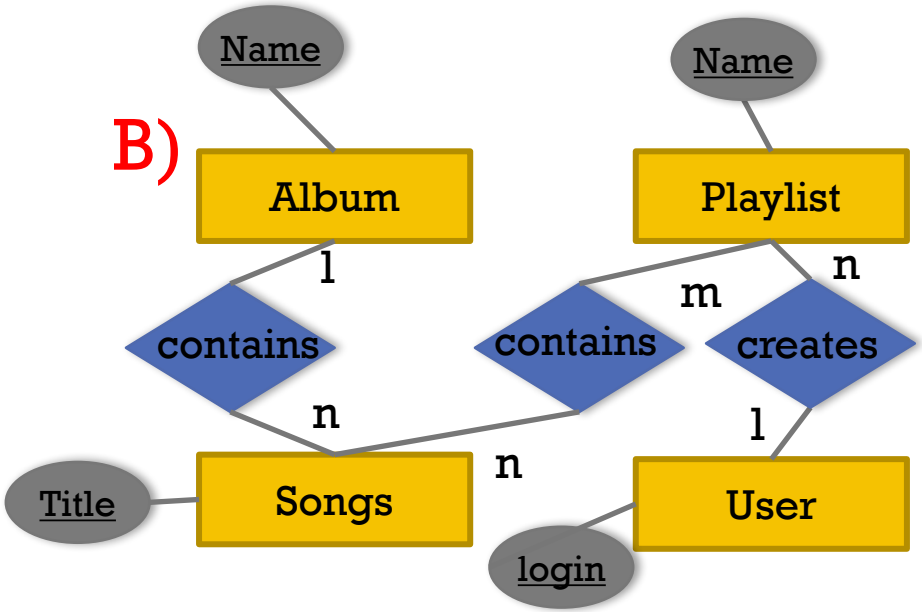
CLICKER QUESTION

- An album has a unique name and songs have unique titles
- An album contains several songs
- A playlist has a unique name and is created by one user with a unique login
- A playlist contains several songs from potential different albums

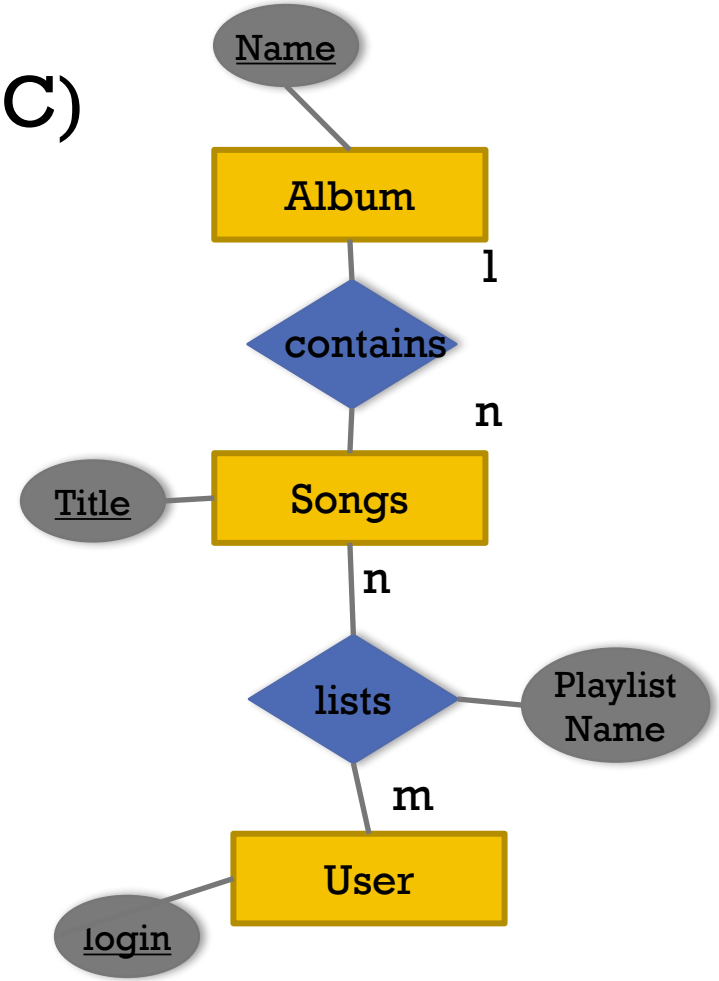
A)



B)



C)



SUMMARY

Elements of ER-Models

Functionalities

- **1:1**
- **1:N**
- **N:M**

(min,max)-Notation