INFO-UB 23: Introduction to Programming and Data Science

Katherine Hoffmann Pham

July 23, 2018

NYU Stern, Department of Information Systems

Agenda

1. Why databases?
2. Entity-relationship diagrams (ERD)
3. Business narratives to ERD
4. ERD to relational databases

Why Databases?

What is a database? Why do we care about database design?

Why Databases?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

Databases vs. Spreadsheets

When should you use a database instead of Excel?

1. Insertion anomalies
2. Deletion anomalies
3. Update anomalies

Anomalies in Un-normalized Data

(1) Insertion anomalies

- Example: Adding a new train line before knowing its stops?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

Anomalies in Un-normalized Data

(1) Insertion anomalies

- Want to insert information about an object *without* having to insert (fake) information about something else
- Example: Adding a new train line before knowing its stops?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

Anomalies in Un-normalized Data

(2) Deletion anomalies

- Example: Remove 6 train without losing Astor Place?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

Anomalies in Un-normalized Data

(2) Deletion anomalies

- Want to avoid losing information about one object when information about a different object is deleted
- Example: Remove 6 train without losing Astor Place?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

Anomalies in Un-normalized Data

(3) Update anomalies

- Example: Change the name of the 6th avenue line?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

Anomalies in Un-normalized Data

(3) Update anomalies

- Want to efficiently update and store information that appears multiple times
- Example: Change the name of the 6th avenue line?

stop_id	stop_name	borough	train_id	train_type	line_id	line_color	line_name
636	Astor Pl	Manhattan	6	Local	456	Green	Lexington Avenue
637	Bleecker St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	6	Local	456	Green	Lexington Avenue
D21	Broadway-Lafayette St	Manhattan	B	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	D	Express	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	F	Local	BDFM	Orange	6 Avenue
D21	Broadway-Lafayette St	Manhattan	M	Local	BDFM	Orange	6 Avenue
R21	8 St - NYU	Manhattan	R	Local	NQRW	Yellow	Broadway
R21	8 St - NYU	Manhattan	W	Local	NQRW	Yellow	Broadway

A Better Solution?

STOPS

stop_id	stop_name	borough
636	Astor Pl	Manhattan
637	Bleecker St	Manhattan
D21	Broadway-Lafayette St	Manhattan
R21	8 St - NYU	Manhattan

TRAINS

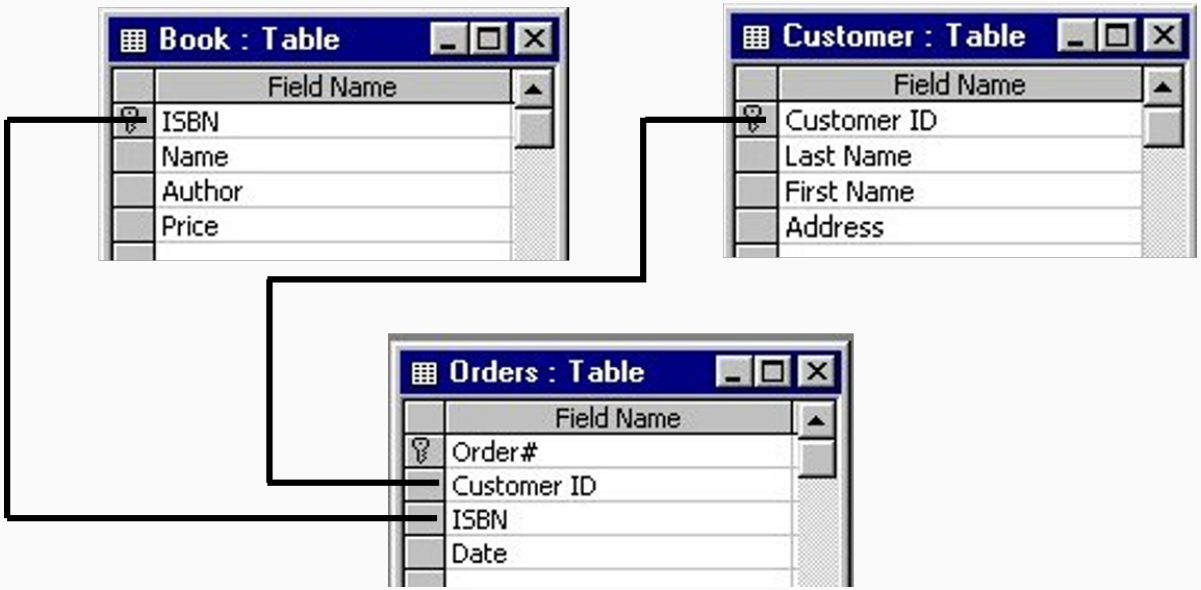
train_id	train_type
6	Local
B	Local
D	Express
F	Local
M	Local
R	Local
W	Local

LINES

line_id	line_color	line_name
456	Green	Lexington Avenue
BDFM	Orange	6 Avenue
NQRW	Yellow	Broadway

A "Normalized" Version of the Spreadsheet

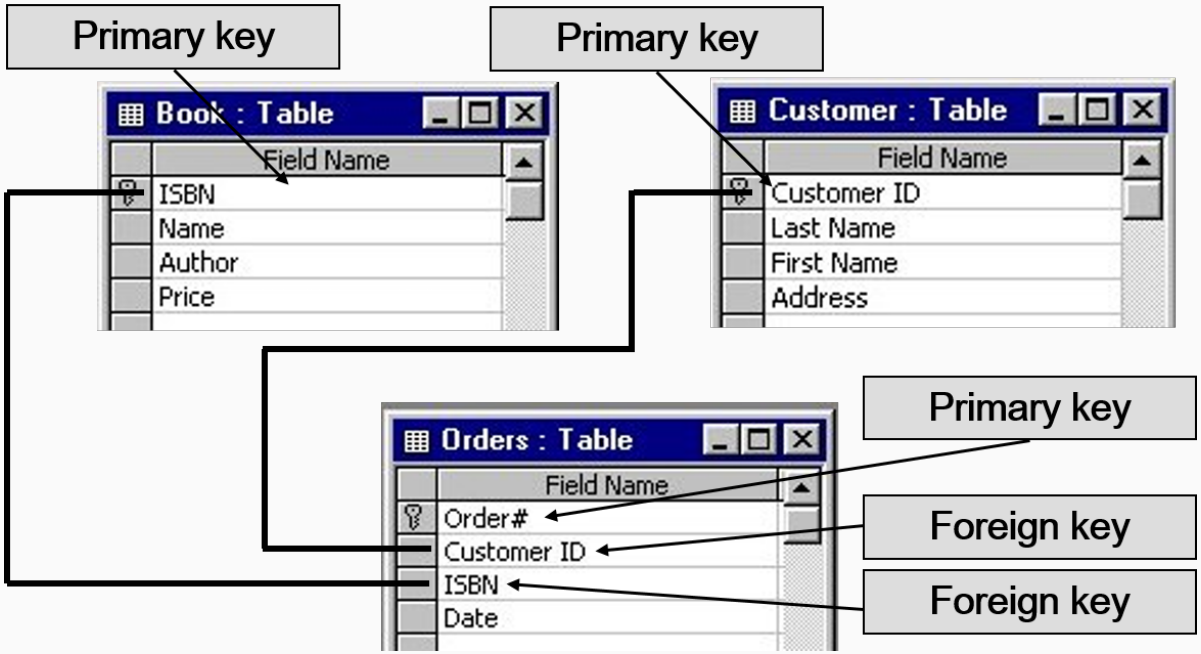
- Data stored in tables
- Cells contain single values
- Avoid insertion, deletion, and update anomalies



A Database Schema

Consists of:

- The tables, along with their fields and keys
- The relationships between the tables



Recap

Databases are helpful in:

1. Reducing redundancy, saving space
2. Updating information consistently
3. Controlling insertion of new information

How Do We Design Databases?

The key questions of database design:

- Which tables to create?
- Which fields to put in each table?
- How to avoid duplication of data?
- How to ensure that database has no "multi valued" cells?
- How to select primary and foreign keys?

Entity-Relationship Diagrams (ERD)

Basic Concepts

- Entities
  - Primary keys
  - Attributes
- Relationships
  - Foreign keys
  - Cardinalities

Entities

Entities are collections of objects with the same properties, e.g.:

- Students
- Courses

Attributes are the properties of these entities, e.g.:

- Student name, student ID, age, gender
- Course ID, section ID, course description, location

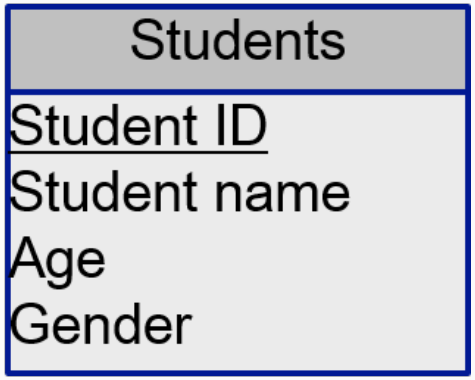
Instances are specific occurrences of an entity, e.g.:

- "Joe Doe", "N12897", "20", "M"
- "INFO.2346", "01","Dealing with Data", "Tisch-UC25"

Primary Keys (PK)

A primary key is an attribute whose value is unique for each instance, e.g.:

- Student ID in the students entity



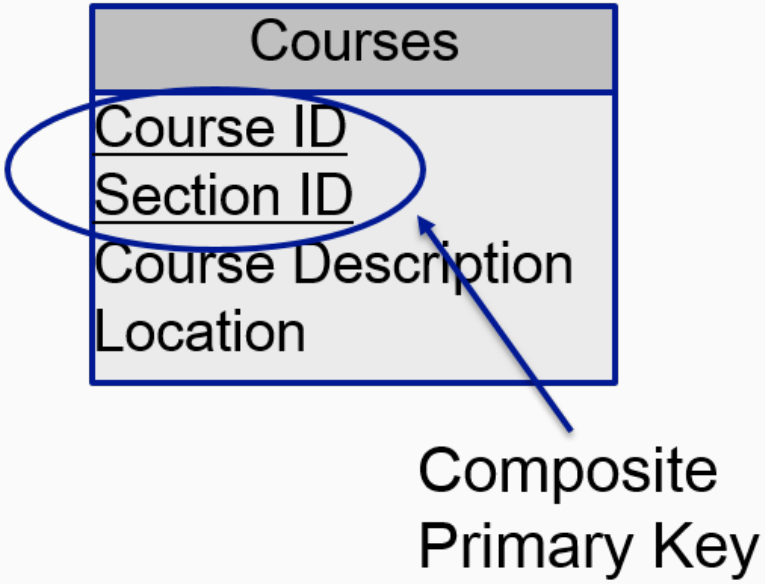
Note:

- Keys are typically denoted by underlining

Primary Keys (PK)

A composite primary key is a primary key that consists of two or more attributes, whose values together (but not separately) are unique for each instance in an entity, e.g.

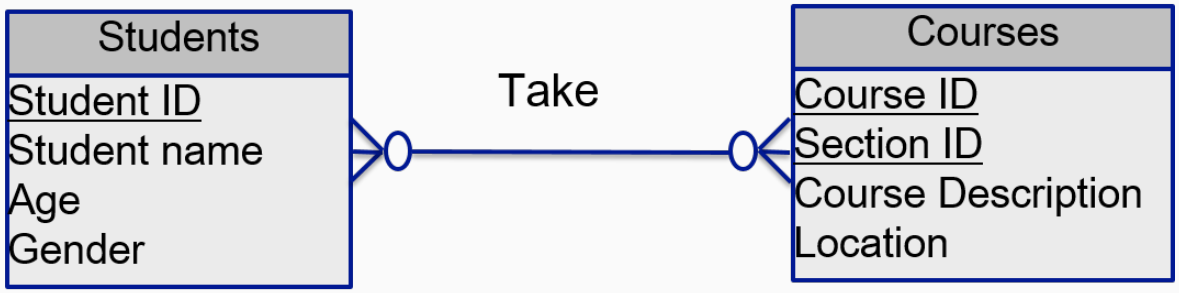
- Course ID and Section ID in the courses entity



Relationships

A relationship is an association among entities, e.g.:

- Students take courses



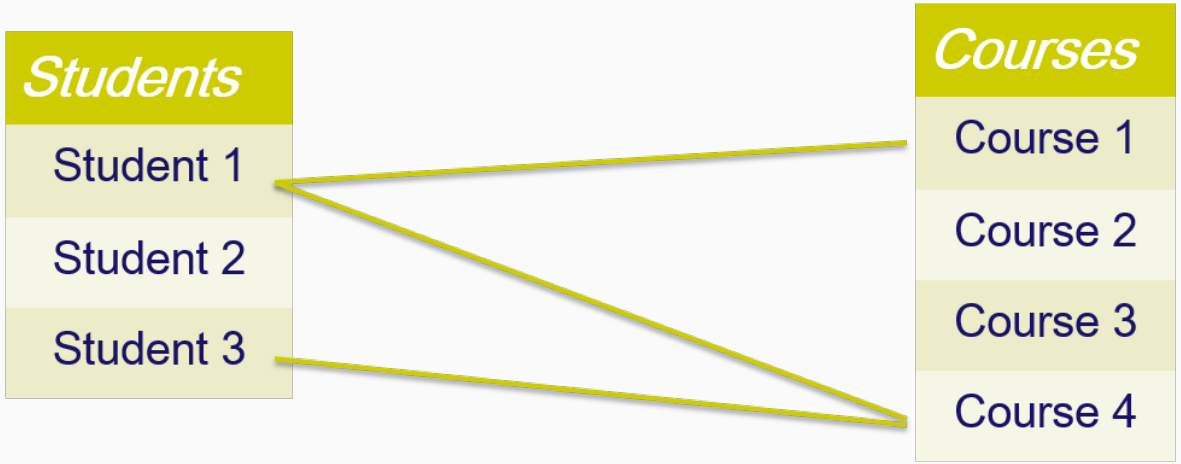
Note:

- A relationship is shown as a line connecting the associated entities, labeled with the name of the relationship
- A relationship name is usually a verb (e.g., takes)

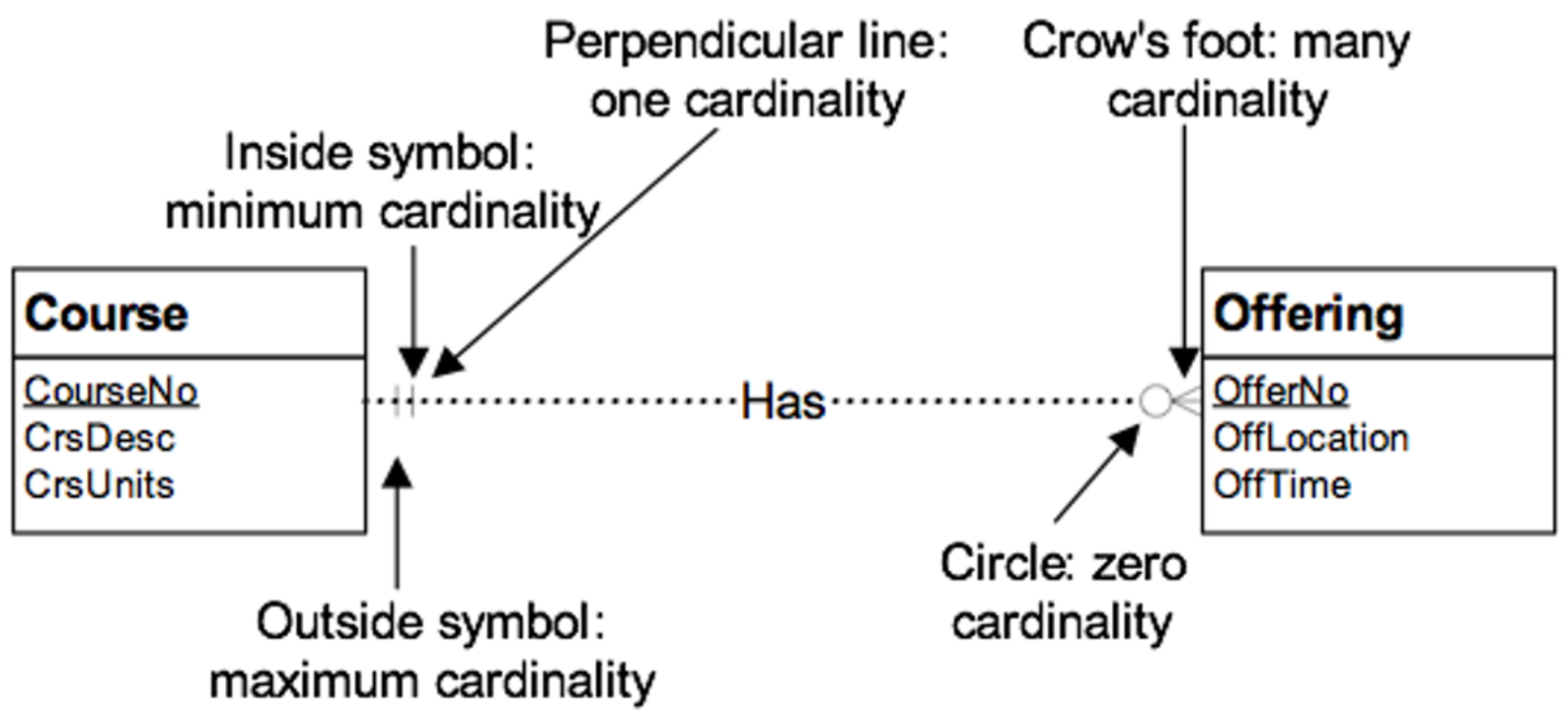
Relationship Cardinalities

Cardinalities describe the number of instances that participate in a relationship, e.g.:

- A student can take 0, 1, or more courses (many).
- A course can be taken by 0, 1, or more students (many).



Cardinality Notation Example



Cardinality Notation

Symbol	Meaning
	One - mandatory
	Many - mandatory
	One - optional
	Many - optional

Cardinality Notation Examples

- Minimum-one, Maximum-one: $\|$

  e.g. a professor must have one and only one office
- Minimum-one, Maximum-many: $|\!<$

  e.g. a department must have at least one instructor, but may have many
- Minimum-zero, Maximum-many: $0\!<$

  e.g. a person can have no phones or many phones
- Minimum-zero, Maximum-one: $0|$

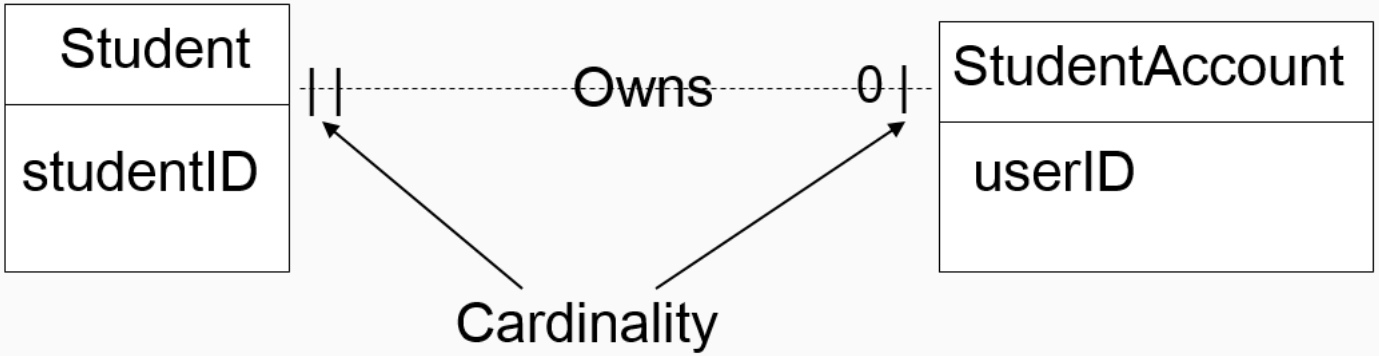
  e.g. a student may have 0 or 1 university accounts

Relationships

We often categorize relationships using maximum cardinality in both directions, e.g.:

- One-to-one
- One-to-many
- Many-to-many

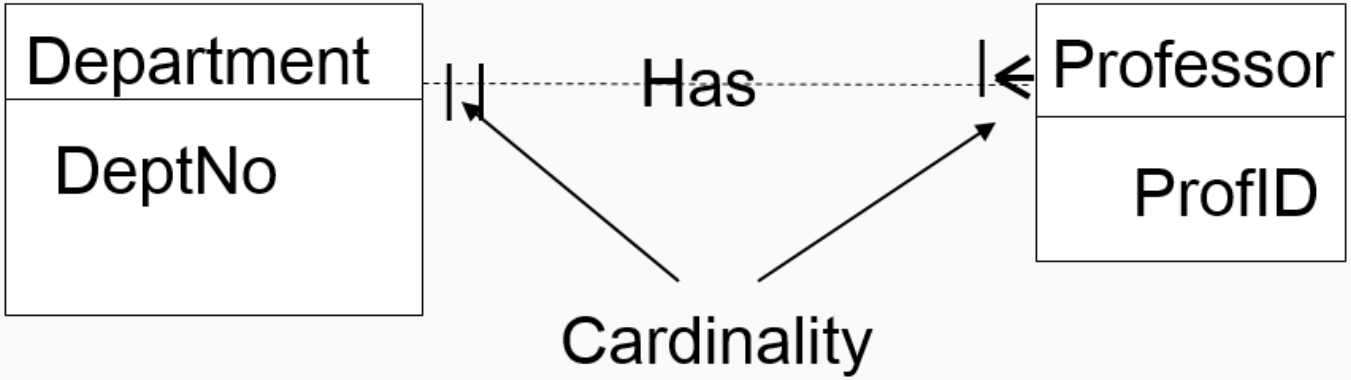
Relationships



One to one, e.g.:

- Each account is owned by exactly one (||) student
- Each student can own zero or one (0|) account

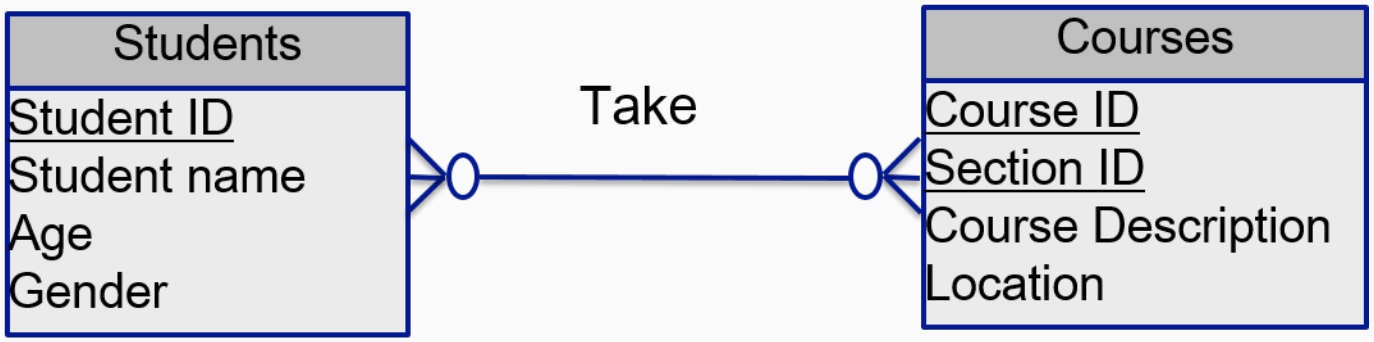
Relationships



One to many, e.g.:

- Each department can have one or more ($\vdash\!\!<$) professors
- Each professor is affiliated with one and only one ($\vdash\!\!\vdash$) department

Relationships

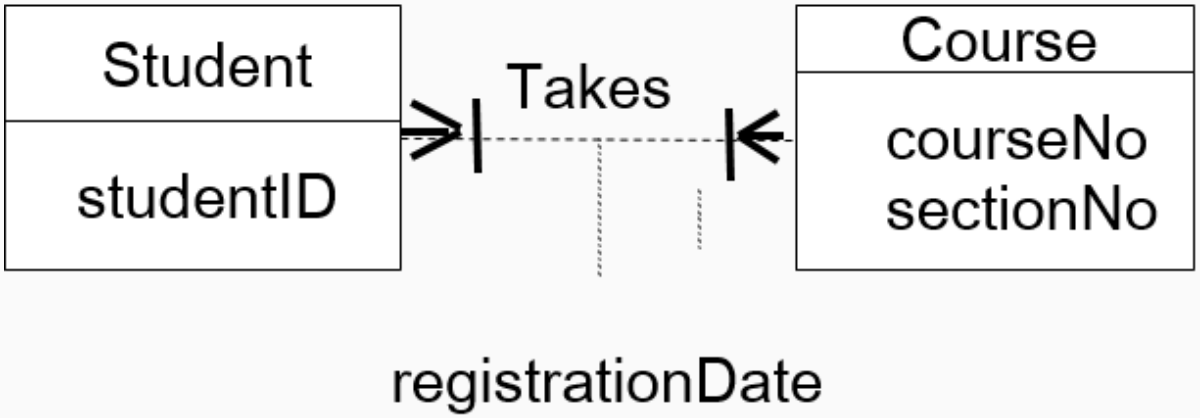


Many to many, e.g.:

- Each student can take 0 or more (0 <) courses
- Each course can enroll 0 to many (0 <) students.

Relationships

Note that relationships can also have attributes . . .



Recap

- Entities are collections of objects with the same attributes
- Primary keys are attributes whose values uniquely identify each instance
- Relationships are associations among entities
- Cardinalities describe the number of instances that participate in a relationship

Challenge: Design an ER model for Uber

Joe (646-889-4539) rides from Harlem to Chelsea for $11.25 and pays with credit card *3945. He is picked up by Jill at 3:37 pm in a red Toyota Corolla, and driven for a distance of 4.2 miles.

Design decisions:

- What tables to create?
- What fields to put in each table?
- How to avoid duplication of data?
- How to ensure that database has no "multi valued" cells?
- How to select primary and foreign keys?

Business Narratives to ERD

From Narratives to ER Diagrams

Typcially, want to convert:

Business narrative

$\rightarrow$ Entity-Relationship Model

$\longrightarrow$ Relational Database

- How can we create an ER diagram from scratch?
- How can we go from an ER diagram to a design for a database?

From Narratives to ER Diagrams

The procedure for analysis:

1. Identify entities and attributes
2. Determine primary keys
3. Identify relationships
4. Determine relationship cardinalities
5. Refine the ERD

Defining entities and primary keys

- What entities/tables should we create?
- What primary keys for the report below?
- Are there fields that are redundant once you create the tables?

Employee ID	Name	Department Num	Department Name	Num of Employees	Job Number	Job Name	Hours
1234	Jones	43	Residential	3	14 23	Acct Sales	4 4
2345	Smith	15	Commercial	1	14	Acct	8
6548	Joslin	43	Residential	3	23 46	Sales Admin	6 2
9087	Mills	43	Residential	3	23 14	Sales Acct	5 3
8797	Jones	69	Non-profit	1	39	Maint	8