Normalization

CIS 3730
Designing and Managing Data

J.G. Zheng
Fall 2010

Overview

- What is normalization?
- What are the normal forms?
- How to normalize relations?

Two Basic Ways To Design Tables

- Bottom-up:
  - Normalization: designing tables by splitting a big relation into multiple related tables to avoid anomalies
- Top-Down
  - Three-level data modeling approach: conceptual, logical, and physical design.

Table Structure Representation

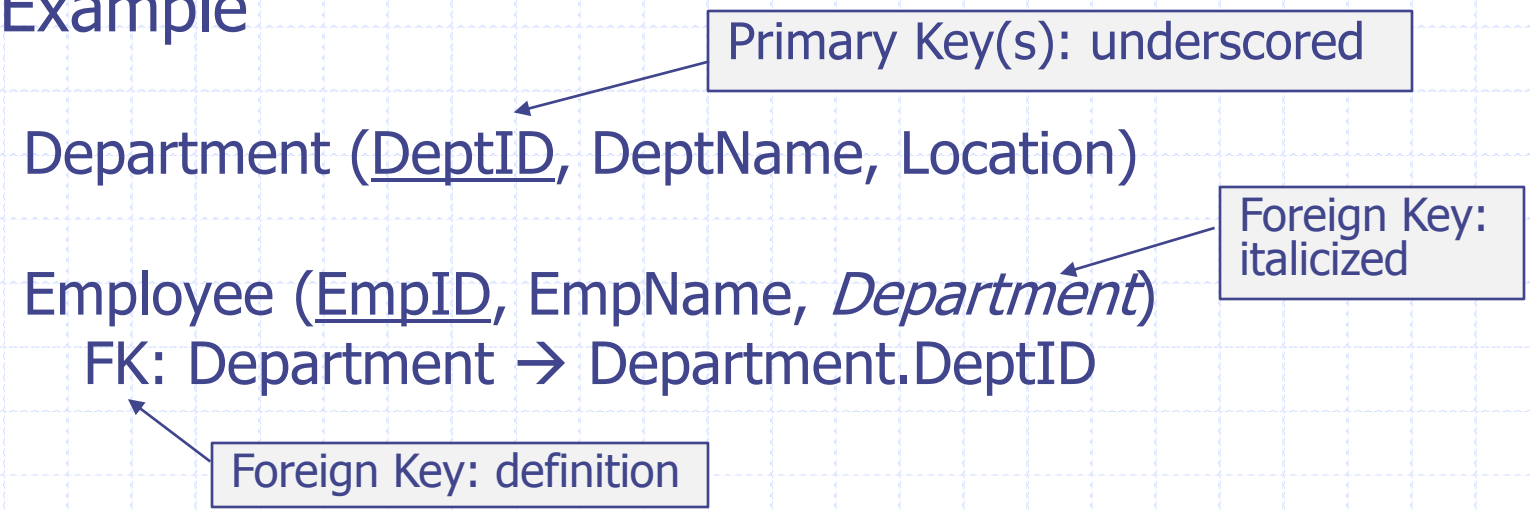
- Text style

  [Table Name]([Primary Key], attribute, [*Foreign Key*], attribute, …)

  FK: [Foreign Key] → [Reference Table].[Primary Key]

  FK: (if more than one foreign key)

- Example



Introduction

- How many, and what, relations (tables) should be used to store my data?
- Is this relation free of problems?

Stud_ID	Stud_Name	**Course_ID**	Course_Name	Instructor	Office	Room	Credit
224	Waters	CIS20	Intro CIS	Greene	CBA001	205G	5
224	Waters	CIS40	Database Mgt	Hong	CBA908	311S	5
224	Waters	CIS50	Sys.Analysis	Purao	CBA700	139S	5
351	Byron	CIS30	COBOL	Hong	CBA908	629G	3
351	Byron	CIS50	Sys.Analysis	Purao	CBA700	139S	5
421	Smith	CIS20	Intro CIS	Greene	CBA001	205G	5
421	Smith	CIS30	COBOL	Hong	CBA908	629G	3
421	Smith	CIS50	Sys.Analysis	Purao	CBA700	139S	5

Normalization

- Normalization is a process of producing a set of related relations (tables) with desirable attributes, given the data requirements of a domain
- The goal is to remove redundancy and data modification problems: insertion anomaly, update anomaly, and deletion anomaly
- Usually dividing a table into 2 or more tables
- Using Normal Forms as a formal guide

Anomaly Example

If Adviser **Baker** is changed to **Taing**, we need to change *AdviserEmail* as well. If changed to **Valdez**, we need to change *AdviserEmail*, *Department*, and *AdminLastName*.

	A	B	C	D	E	F	G
1	**LastName**	**FirstName**	**Email**	**AdviserLastName**	**AdviserEmail**	**Department**	**AdminLastName**
2	Andrews	Matthew	Matthew.Andrews@ourcampus.edu	Baker	Linda.Baker@ourcampus.edu	Accounting	Smith
3	Brisbon	Lisa	Lisa.Brisbon@ourcampus.edu	Valdez	Richard.Valdez@ourcampus.edu	Chemistry	Chaplin
4	Fischer	Douglas	Douglas.Fischer@ourcampus.edu	Baker	Linda.Baker@ourcampus.edu	Accounting	Smith
5	Hwang	Terry	Terry.Hwang@ourcampus.edu	Taing	Susan.Taing@ourcampus.edu	Accounting	Smith
6	Lai	Tzu	Tzu.Lai@ourcampus.edu	Valdez	Richard.Valdez@ourcampus.edu	Chemistry	Chaplin
7	~~Marino~~	~~Chip~~	~~Chip.Marino@ourcampus.edu~~	~~Tran~~	~~Ken.Tran@ourcampus.edu~~	~~InfoSystems~~	~~Rogers~~
8	Thompson	James	James.Thompson@ourcampus.edu	Taing	Susan.Taing@ourcampus.edu	Accounting	Smith
9	???	???	???	???	???	Biology	Kelly

Deleted row—Student, Adviser, and Department data lost

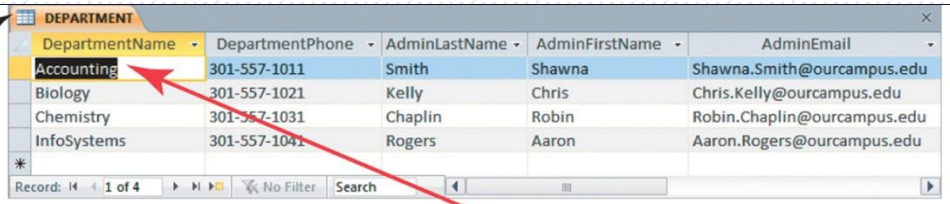
Inserted row—both Student and Adviser data missing

Normalized Tables

Can insert DEPARTMENT data as needed—no ADVISER or STUDENT data required

Can change STUDENT Adviser name as needed—new value is linked to its own data

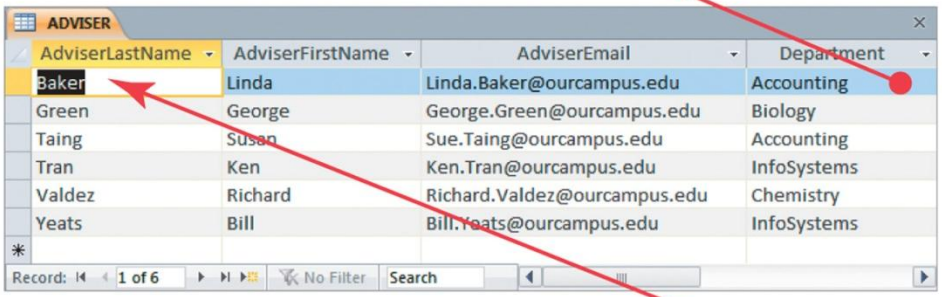
Can delete STUDENT data as needed—no DEPARTMENT or ADVISER data lost



DEPARTMENT

DepartmentName	DepartmentPhone	AdminLastName	AdminFirstName	AdminEmail
Accounting	301-557-1011	Smith	Shawna	Shawna.Smith@ourcampus.edu
Biology	301-557-1021	Kelly	Chris	Chris.Kelly@ourcampus.edu
Chemistry	301-557-1031	Chaplin	Robin	Robin.Chaplin@ourcampus.edu
InfoSystems	301-557-1041	Rogers	Aaron	Aaron.Rogers@ourcampus.edu

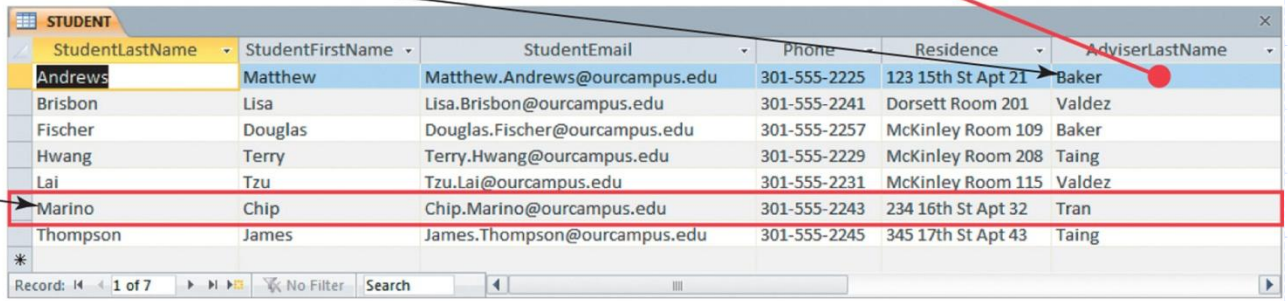
Record: 1 of 4 No Filter Search



ADVISER

AdviserLastName	AdviserFirstName	AdviserEmail	Department
Baker	Linda	Linda.Baker@ourcampus.edu	Accounting
Green	George	George.Green@ourcampus.edu	Biology
Taing	Susan	Sue.Taing@ourcampus.edu	Accounting
Tran	Ken	Ken.Tran@ourcampus.edu	InfoSystems
Valdez	Richard	Richard.Valdez@ourcampus.edu	Chemistry
Yeats	Bill	Bill.Yeats@ourcampus.edu	InfoSystems

Record: 1 of 6 No Filter Search



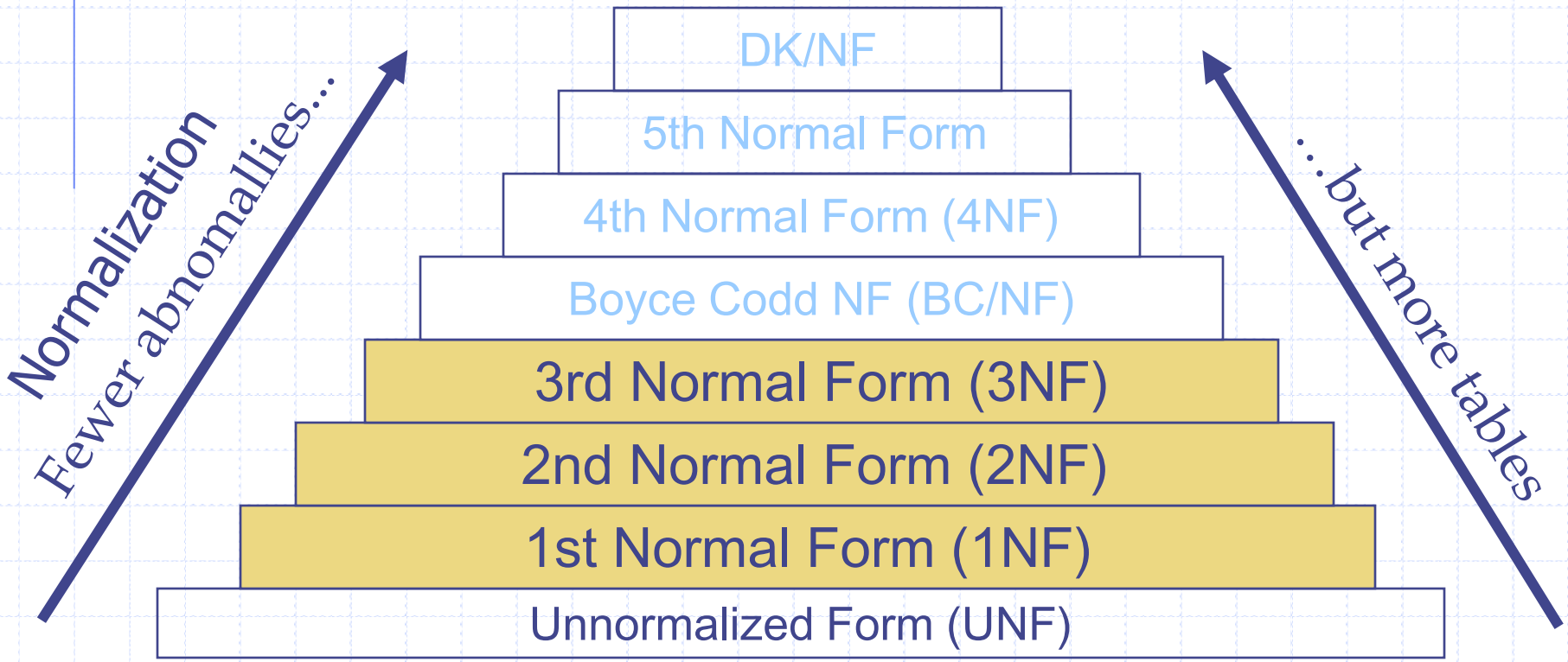
STUDENT

StudentLastName	StudentFirstName	StudentEmail	Phone	Residence	AdviserLastName
Andrews	Matthew	Matthew.Andrews@ourcampus.edu	301-555-2225	123 15th St Apt 21	Baker
Brisbon	Lisa	Lisa.Brisbon@ourcampus.edu	301-555-2241	Dorsett Room 201	Valdez
Fischer	Douglas	Douglas.Fischer@ourcampus.edu	301-555-2257	McKinley Room 109	Baker
Hwang	Terry	Terry.Hwang@ourcampus.edu	301-555-2229	McKinley Room 208	Taing
Lai	Tzu	Tzu.Lai@ourcampus.edu	301-555-2231	McKinley Room 115	Valdez
Marino	Chip	Chip.Marino@ourcampus.edu	301-555-2243	234 16th St Apt 32	Tran
Thompson	James	James.Thompson@ourcampus.edu	301-555-2245	345 17th St Apt 43	Taing

Record: 1 of 7 No Filter Search

Normal Forms

- Normal forms are formal guidelines (steps) for the normalization process



Normalization – 1NF

- A table is in 1NF if
  1. it satisfies the definition of a *relation*
     - Review: what are the features of a relation?
  2. no "repeating groups" (columns)

Repeating Groups

Customer ID	First Name	Surname	Telephone Number
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659 555-776-4100
789	Maria	Fernandez	555-808-9633

Customer ID	First Name	Surname	Tel. No. 1	Tel. No. 2	Tel. No. 3
123	Robert	Ingram	555-861-2025		
456	Jane	Wright	555-403-1659	555-776-4100	
789	Maria	Fernandez	555-808-9633		

Lots of Null values

Avoid Repeating Groups

- Transforming to additional rows, rather than additional columns

Customer ID	First Name	Surname	Telephone Number
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659
456	Jane	Wright	555-776-4100
789	Maria	Fernandez	555-808-9633

Transforming to 1NF: Example

- Another example

UNF

OrderNum	OrderDate	PartNum	NumOrdered
21608	10/20/2003	AT94	11
21610	10/20/2003	DR93 DW11	1 1
21613	10/21/2003	KL62	4
21614	10/21/2003	KT03	2
21617	10/23/2003	BV06 CD52	2 4
21619	10/23/2003	DR93	1
21623	10/23/2003	KV29	2

1NF

OrderNum	OrderDate	PartNum	NumOrdered
21608	10/20/2003	AT94	11
21610	10/20/2003	DR93	1
21610	10/20/2003	DW11	1
21613	10/21/2003	KL62	4
21614	10/21/2003	KT03	2
21617	10/23/2003	BV06	2
21617	10/23/2003	CD52	4
21619	10/23/2003	DR93	1
21623	10/23/2003	KV29	2

Problems in 1NF

- Basically it may have the same problem as spreadsheet tables
  - Redundancy, and anomalies
- What's the problem in this table?

Customer ID	First Name	Surname	Telephone Number
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659
456	Jane	Wright	555-776-4100
789	Maria	Fernandez	555-808-9633

Higher Normal Forms

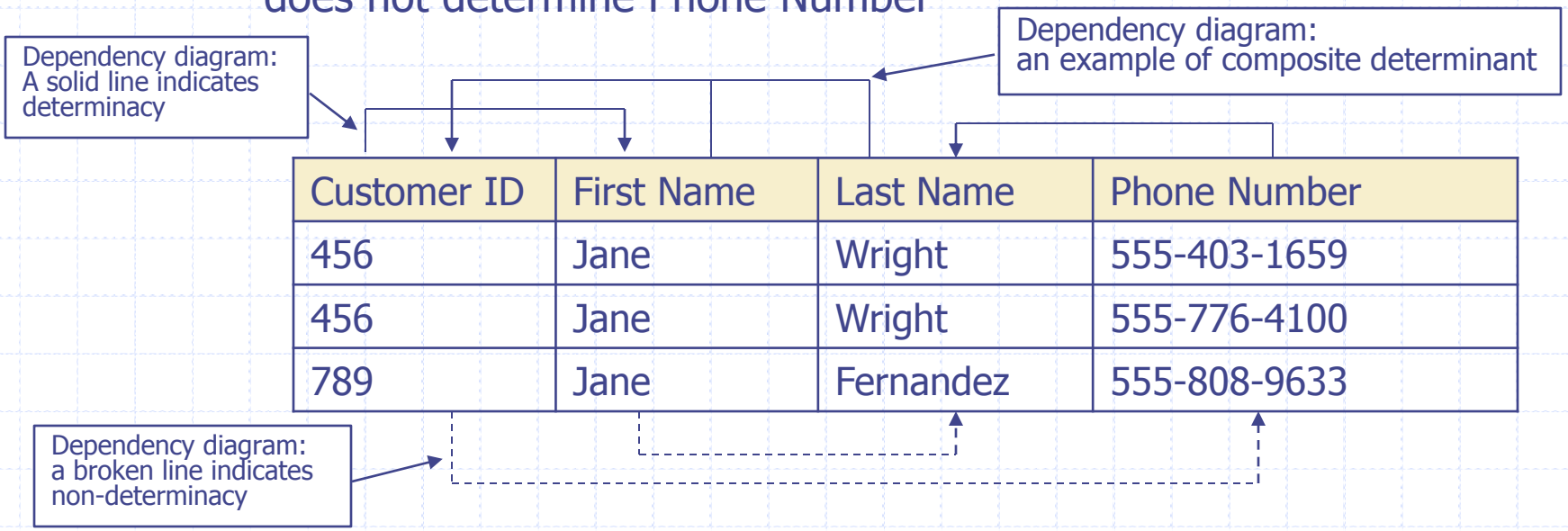
- Normal forms higher than 1NF deal with *functional dependency*
- Identifying the normal form level by analyzing the *functional dependency* between *attributes* (fields)

Functional Dependency

- If each value of attribute A is associated with only one value of attribute B, we say
  - A determines B
  - Or, B is dependent on A
  - Denoted as: A → B
- Functional dependence describes relationships between attributes (not relations)
- Composite determinant: A (and B) can be a set of fields
  - If A consists of column a and b, and a and b together determines c, then:
  - (a, b) → c

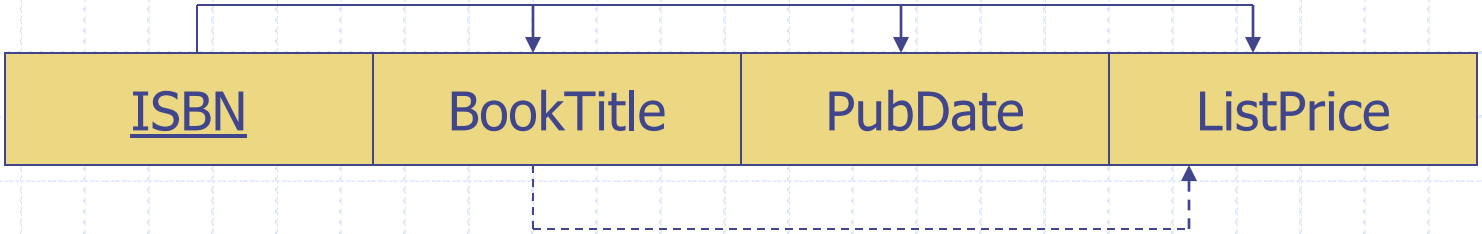
Functional Dependency Examples

- Dependency example
  - For each Customer ID, there is only one corresponding first name (or last name), so: Customer ID determines First Name, or Customer ID → First Name
  - Composite determinant: (First Name, Last Name) → Customer ID
- Non-dependency example
  - An Customer ID can have multiple phone numbers, so: Customer ID does not determine Phone Number



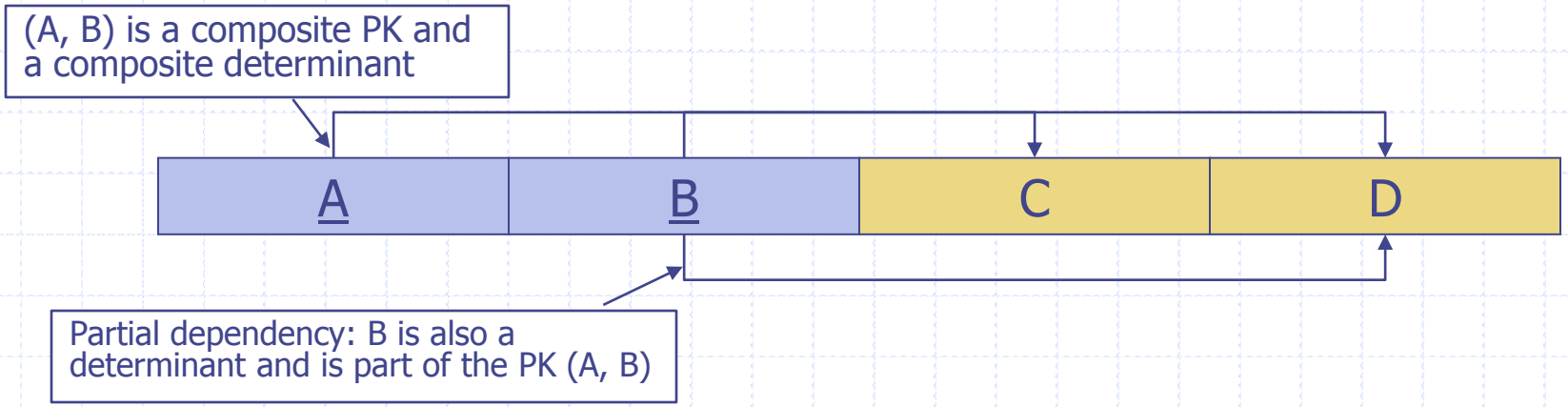
Functional Dependency and Keys

- By definition, a unique key functionally determines all other attributes
  - Primary key
  - Candidate key
  - Surrogate key
  - Composite primary key
- Example

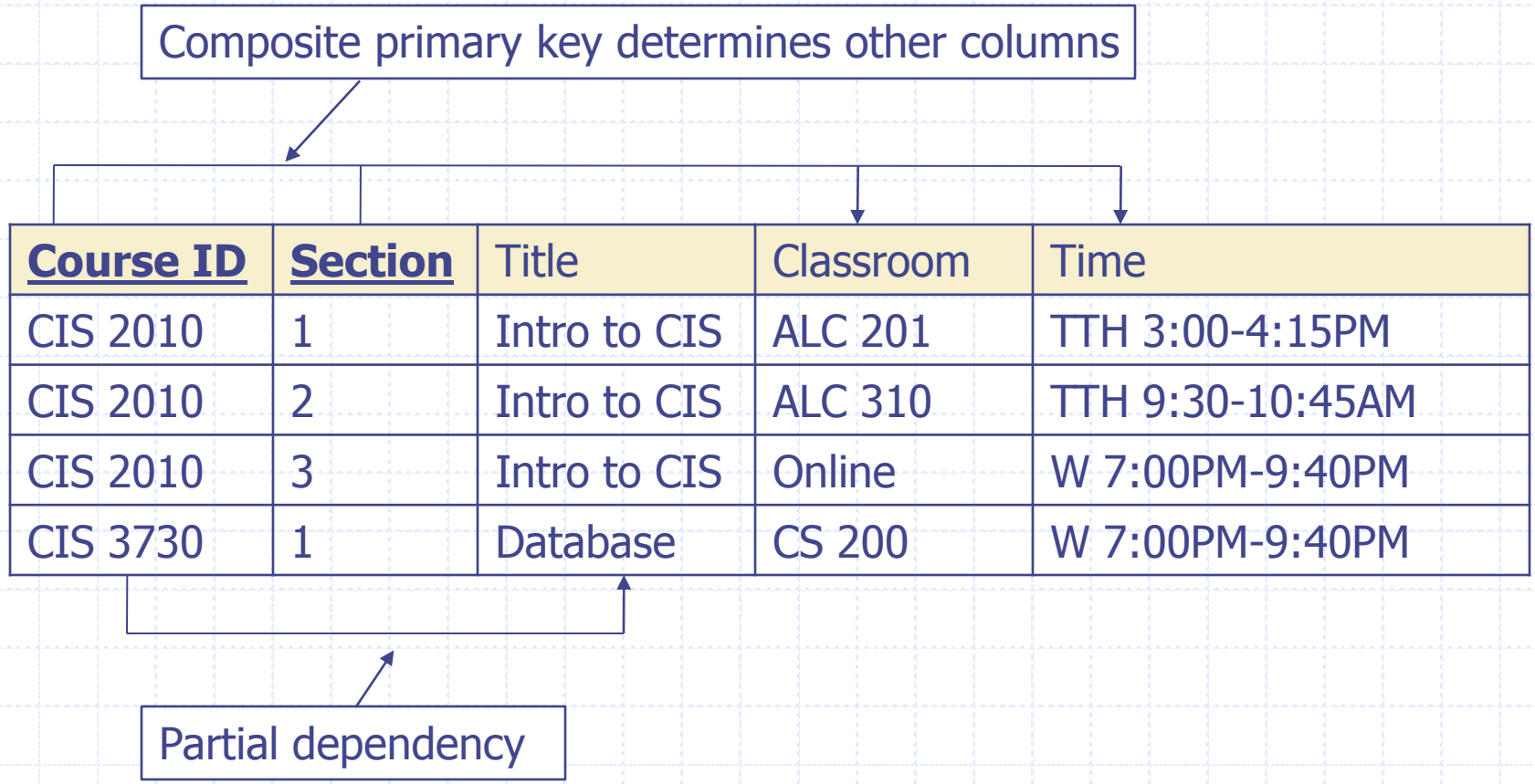


Normalization – 2NF

- A relation is in 2NF, if
  - It is in 1NF, and
  - All non-key attributes (attributes that are not part of any primary key or candidate key) must be functionally dependent on the whole primary (candidate) key
  - Or, NO *partial dependency*
- Partial dependency
  - A non-key attribute is dependent on part of a composite primary key
- Implication
  - A relation with only single-attribute primary key and candidate key does not have partial dependency problem; therefore, such a relation is in 2NF.



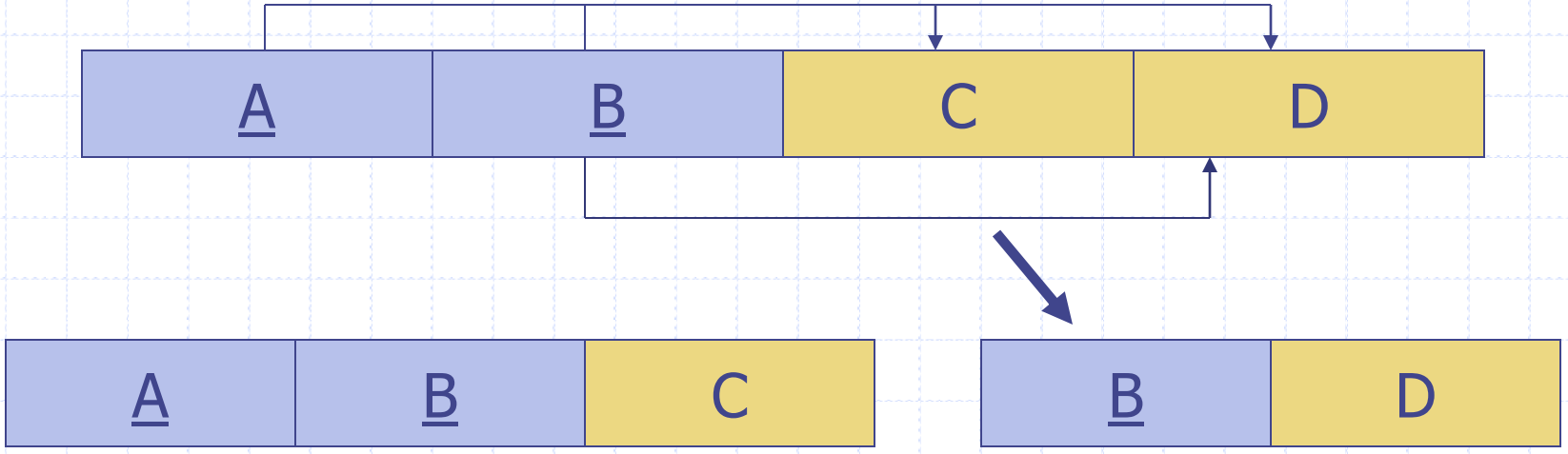
A Relation in 1NF but Not in 2NF



Course ID	**Section**	Title	Classroom	Time
CIS 2010	1	Intro to CIS	ALC 201	TTH 3:00-4:15PM
CIS 2010	2	Intro to CIS	ALC 310	TTH 9:30-10:45AM
CIS 2010	3	Intro to CIS	Online	W 7:00PM-9:40PM
CIS 3730	1	Database	CS 200	W 7:00PM-9:40PM

Transforming to 2NF

- Steps
  - Identify the primary key (PK).
  - If PK consists of only one field, then it is in 2NF.
  - If PK is a composite key, then look for partial dependency.
  - If there is partial dependency, move the partial dependency involved attributes to another relation.



Transforming to 2NF: Example

Course ID	**Section**	Title	Classroom	Time
CIS 2010	1	Intro to CIS	ALC 201	TTH 3:00-4:15PM
CIS 2010	2	Intro to CIS	ALC 310	TTH 9:30-10:45AM
CIS 2010	3	Intro to CIS	Online	W 7:00PM-9:40PM
CIS 3730	1	Database	CS 200	W 7:00PM-9:40PM

Course ID	Title
CIS 2010	Intro to CIS
CIS 3730	Database

Redundancy is avoided

Course ID	**Section**	Classroom	Time
CIS 2010	1	ALC 201	TTH 3:00-4:15PM
CIS 2010	2	ALC 310	TTH 9:30-10:45AM
CIS 2010	3	Online	W 7:00PM-9:40PM
CIS 3730	1	CS 200	W 7:00PM-9:40PM

Course (CourseID, Title)

Schedule (*CouseID*, Section, Classroom, Time)
FK: CourseID→Course.CourseID

Problems in 2NF

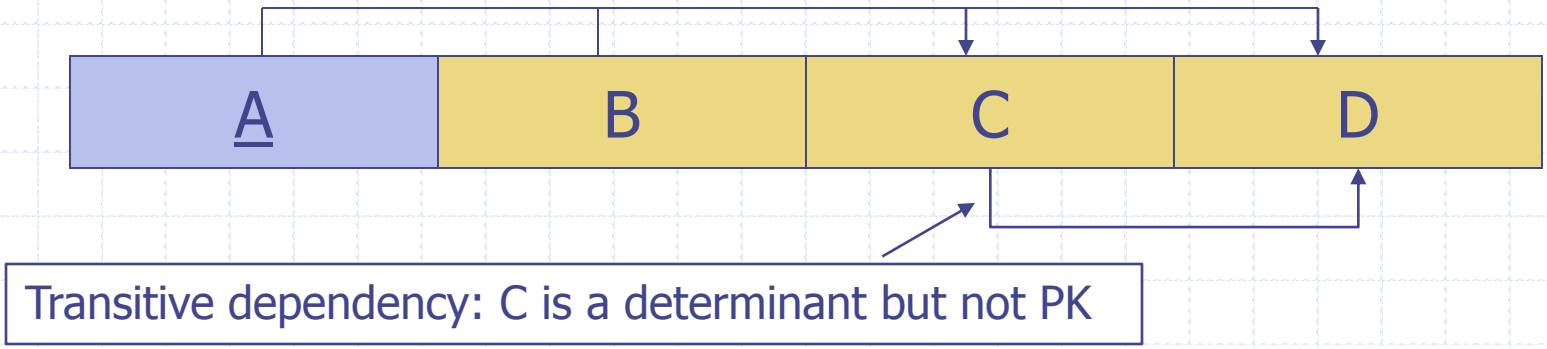
- Again, there could be redundancy and potential inconsistency

Order_ID	Order_Date	Cust_ID	Cust_Name	Cust_Address
1006	10/24/2004	2	Value Furniture	Plano, TX
1007	10/25/2004	6	Furniture Gallery	Boulder, CO
1008	11/1/2004	2	Value Furniture	Plano, TX

Normalization – 3NF

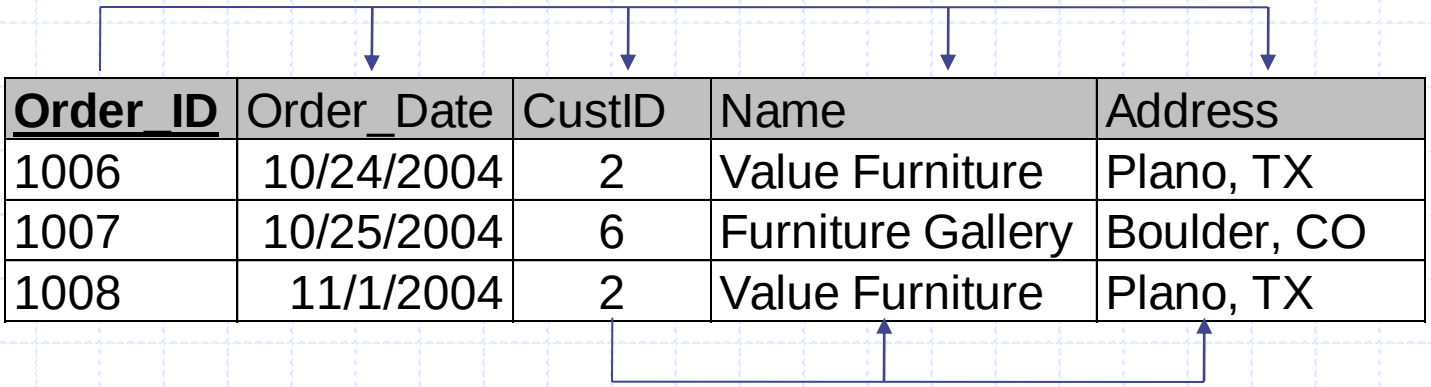
- A relation is in 3NF, if
  - It is in 2NF, and
  - All (non-key) attributes must, and only, be functionally dependent on the primary key
  - Or, NO *transitive dependency*

- Transitive dependency
  - A → B and B → C, then A → C



A Relation in 2NF but Not in 3NF

- Identify primary key (PK) and look for transitive dependency



Order_ID	Order_Date	CustID	Name	Address
1006	10/24/2004	2	Value Furniture	Plano, TX
1007	10/25/2004	6	Furniture Gallery	Boulder, CO
1008	11/1/2004	2	Value Furniture	Plano, TX

Transitive dependency

Transforming to 3NF

- Move the attributes involved in a transitive dependency to another relation

Order

Order_ID	Order_Date	CustID	Name	Address
1006	10/24/2004	2	Value Furniture	Plano, TX
1007	10/25/2004	6	Furniture Gallery	Boulder, CO
1008	11/1/2004	2	Value Furniture	Plano, TX

Order

Order_ID	Order_Date	Customer
1006	10/24/2004	2
1007	10/25/2004	6
1008	11/1/2004	2

Customer

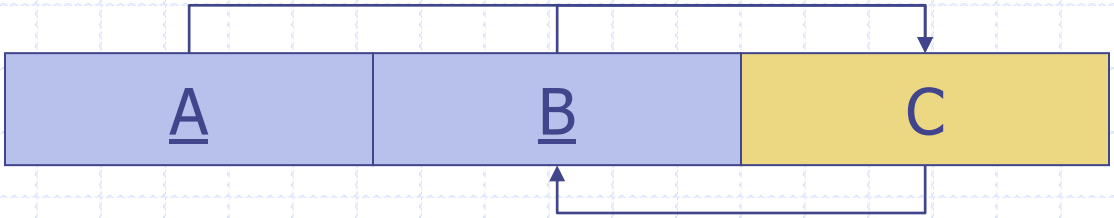
CustID	Name	Address
2	Value Furniture	Plano, TX
6	Furniture Gallery	Boulder, CO

Customer (CustID, Name, Address)

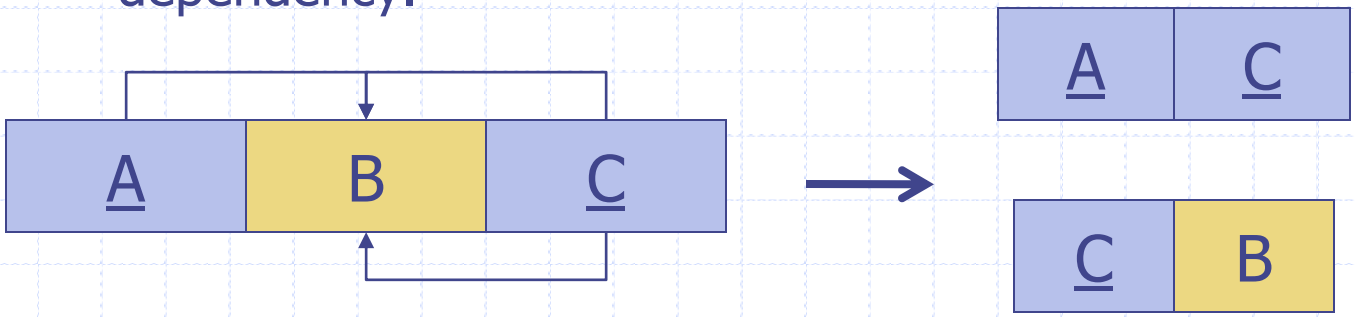
Order (Corder_ID, Order_Date, *Customer*)
FK: Customer→Customer.CustID

BC/NF

- BC/NF is a stricter form of 2NF and 3NF



  - (A, B) is a candidate key
  - (A, C) is also a candidate key
  - A, B, C are all key attributes; there are no non-key attributes
  - Can be viewed as special case of transitive dependency; or can be transformed into a similar pattern as partial dependency.



BC/NF Example

Physician	Patient	Bill Number
A	1	101
B	1	101
A	2	102
B	2	103

Physician	Bill Number
A	101
B	101
A	102
B	103

Bill Number	Patient
101	1
102	2
103	2

4NF Brief

- Multi-value dependency
  - Employee →→ Skill (determines multiple skills)
  - Employee →→ Degree (determines multiple degrees)

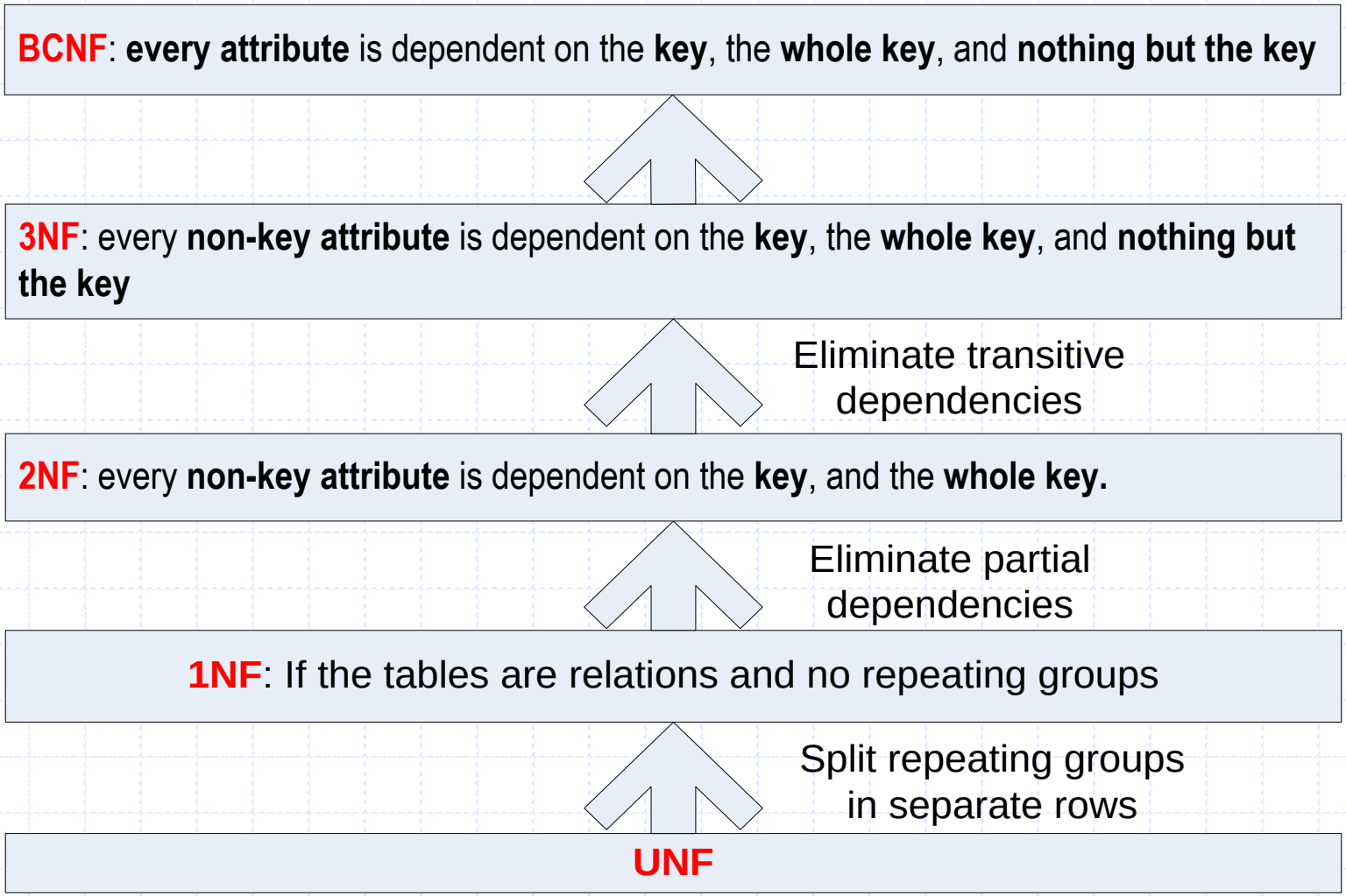
Employee	Skill	Degree
Jack	SQL, Teaching	BA, MS, PhD
Michael	SQL, C#, Java, Network	BA, MBA

Employee	Skill	Degree
Jack	SQL	BA
Jack	Teaching	MS
Jack		PhD
Michael	SQL	BA
…		

Practical Tips

- To identify the normalization level, determine the primary key and candidate keys first; then look for partial dependency (check if there is a composite PK or candidate key) and transitive dependency
- Design a relation that is easy to explain its meaning
  - If there are attributes of different things in one table, there are usually problems; for example, students and courses are in one table, or customer and products are in one table; etc.
- Attributes that potentially generate many Null values might be moved into another table
- Generally relations in the 3NF are considered to be well formed; going higher may introduce unnecessary structural complexity which is inefficient for data queries
- Very often tables can go for lower normal forms (de-normalization) depending on design requirements

Normal Forms Summary



Summary

- Key concepts
  - Anomalies
  - Normalization and de-normalization
  - Normal forms: 1NF to 3NF
  - Functional dependency
    - Partial dependency
    - Transitive dependency

- Key skills: identify and normalize tables from 1NF to 3NF
  - Be able to identify the normal form of a given relation
  - Be able to identify functional dependency among attributes
  - Be able to apply normalization principles to normalize a relation up to the 3rd normalization form