Data Normalization

Data Normalization

- Primarily a tool to validate and improve a logical design so that it satisfies certain constraints that ***avoid unnecessary duplication of data***
- The process of decomposing relations with anomalies to produce smaller, ***well-structured*** relations

Well-Structured Relations

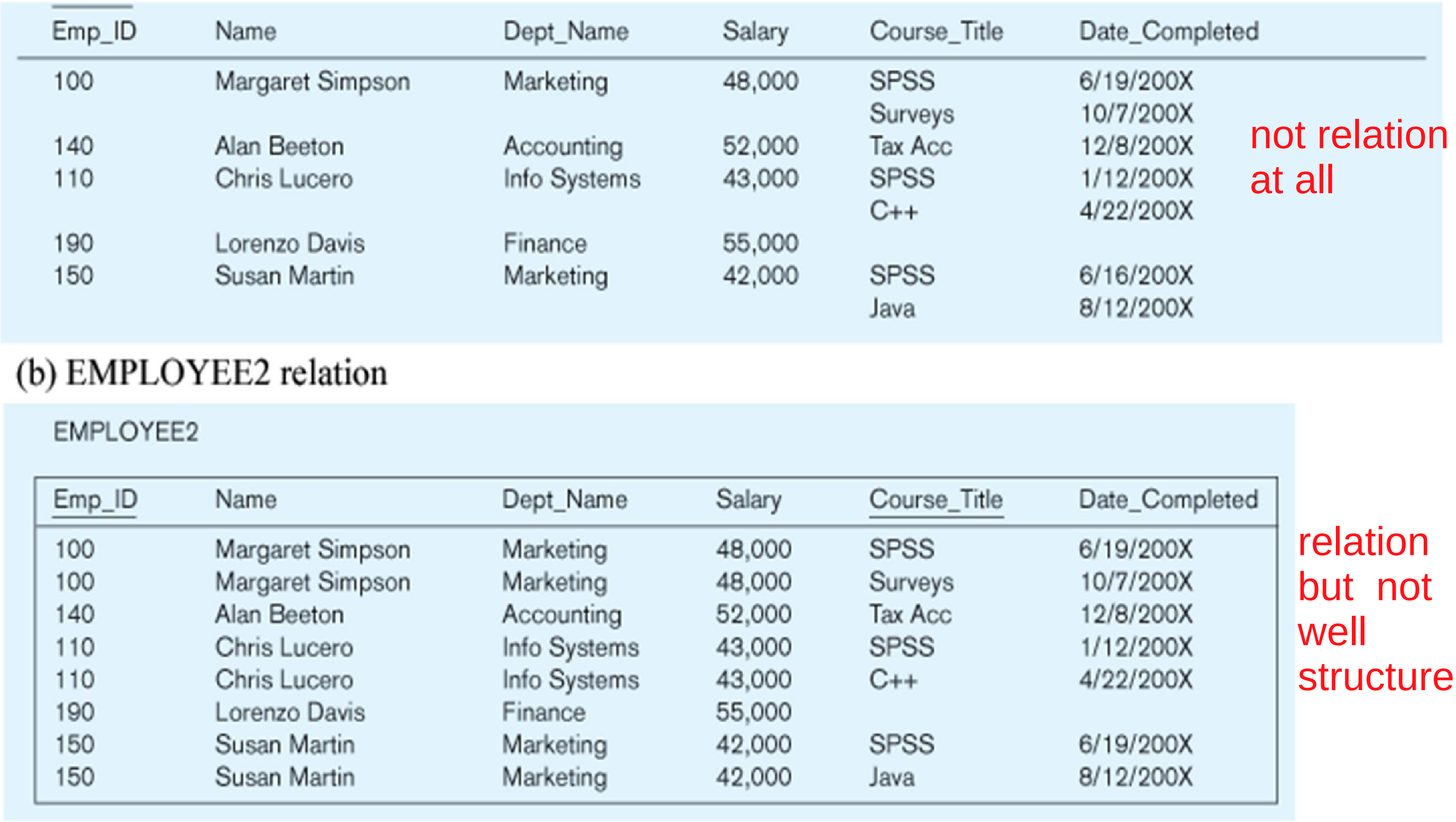
- A relation that contains minimal data redundancy and allows users to insert, delete, and update rows without causing data inconsistencies
- Goal is to avoid anomalies
  - **Insertion Anomaly** –adding new rows forces user to create duplicate data
  - **Deletion Anomaly** –deleting rows may cause a loss of data that would be needed for other future rows
  - **Modification Anomaly** –changing data in a row forces changes to other rows because of duplication

General rule of thumb: A table should not pertain to more than one entity type

Example

good example
try to solve its
with normalization

Figure 5-2 Eliminating multivalued attributes (a) Table with repeating groups



Emp_ID	Name	Dept_Name	Salary	Course_Title	Date_Completed
100	Margaret Simpson	Marketing	48,000	SPSS	6/19/200X
				Surveys	10/7/200X
140	Alan Beeton	Accounting	52,000	Tax Acc	12/8/200X
110	Chris Lucero	Info Systems	43,000	SPSS	1/12/200X
				C++	4/22/200X
190	Lorenzo Davis	Finance	55,000		
150	Susan Martin	Marketing	42,000	SPSS	6/16/200X
				Java	8/12/200X

not relation
at all

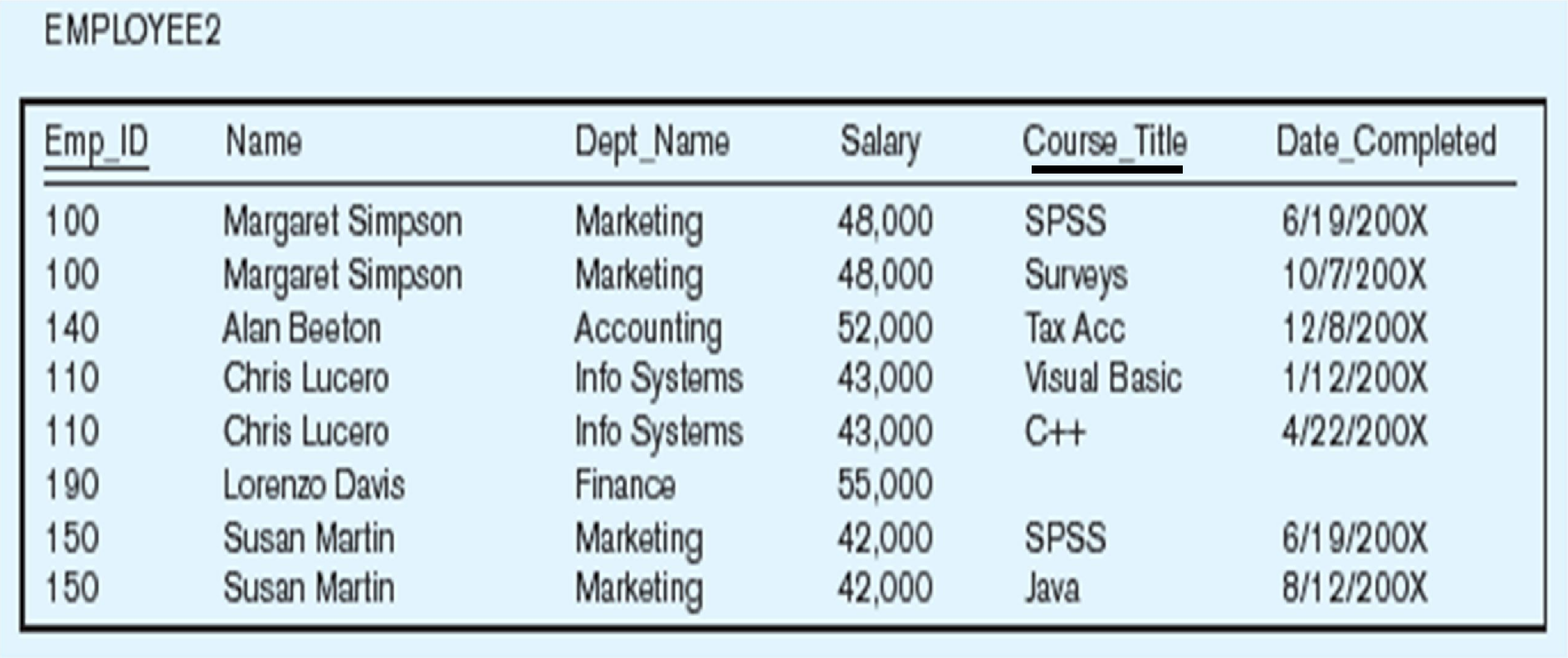
(b) EMPLOYEE2 relation

EMPLOYEE2

Emp_ID	Name	Dept_Name	Salary	Course_Title	Date_Completed
100	Margaret Simpson	Marketing	48,000	SPSS	6/19/200X
100	Margaret Simpson	Marketing	48,000	Surveys	10/7/200X
140	Alan Beeton	Accounting	52,000	Tax Acc	12/8/200X
110	Chris Lucero	Info Systems	43,000	SPSS	1/12/200X
110	Chris Lucero	Info Systems	43,000	C++	4/22/200X
190	Lorenzo Davis	Finance	55,000		
150	Susan Martin	Marketing	42,000	SPSS	6/19/200X
150	Susan Martin	Marketing	42,000	Java	8/12/200X

relation
but not
well
structured

Example –Figure 5-2b



EMPLOYEE2

Emp_ID	Name	Dept_Name	Salary	Course_Title	Date_Completed
100	Margaret Simpson	Marketing	48,000	SPSS	6/19/200X
100	Margaret Simpson	Marketing	48,000	Surveys	10/7/200X
140	Alan Beeton	Accounting	52,000	Tax Acc	12/8/200X
110	Chris Lucero	Info Systems	43,000	Visual Basic	1/12/200X
110	Chris Lucero	Info Systems	43,000	C++	4/22/200X
190	Lorenzo Davis	Finance	55,000		
150	Susan Martin	Marketing	42,000	SPSS	6/19/200X
150	Susan Martin	Marketing	42,000	Java	8/12/200X

Question–Is this a relation?

Answer–Yes: Unique rows and no multivalued attributes

Question–What's the primary key?

Answer–Composite: Emp_ID, Course_Title

Anomalies in this Table

- **Insertion**–can't enter a new employee without having the employee take a class
- **Deletion**–if we remove employee 140, we lose information about the existence of a Tax Acc class
- **Modification**–giving a salary increase to employee 100 forces us to update multiple records

Why do these anomalies exist?

Because there are two themes (entity types) in this one relation. This results in data duplication and an unnecessary dependency between the entities

Functional Dependencies and Keys

- Functional Dependency: The value of one attribute (the ***determinant***) determines the value of another attribute
- Candidate Key:
  - A unique identifier. One of the candidate keys will become the primary key
    - E.g. perhaps there is both credit card number and SS# in a table…in this case both are candidate keys
  - Each non-key field is functionally dependent on every candidate key

Figure 5.22 Steps in normalization

we take the first three steps only

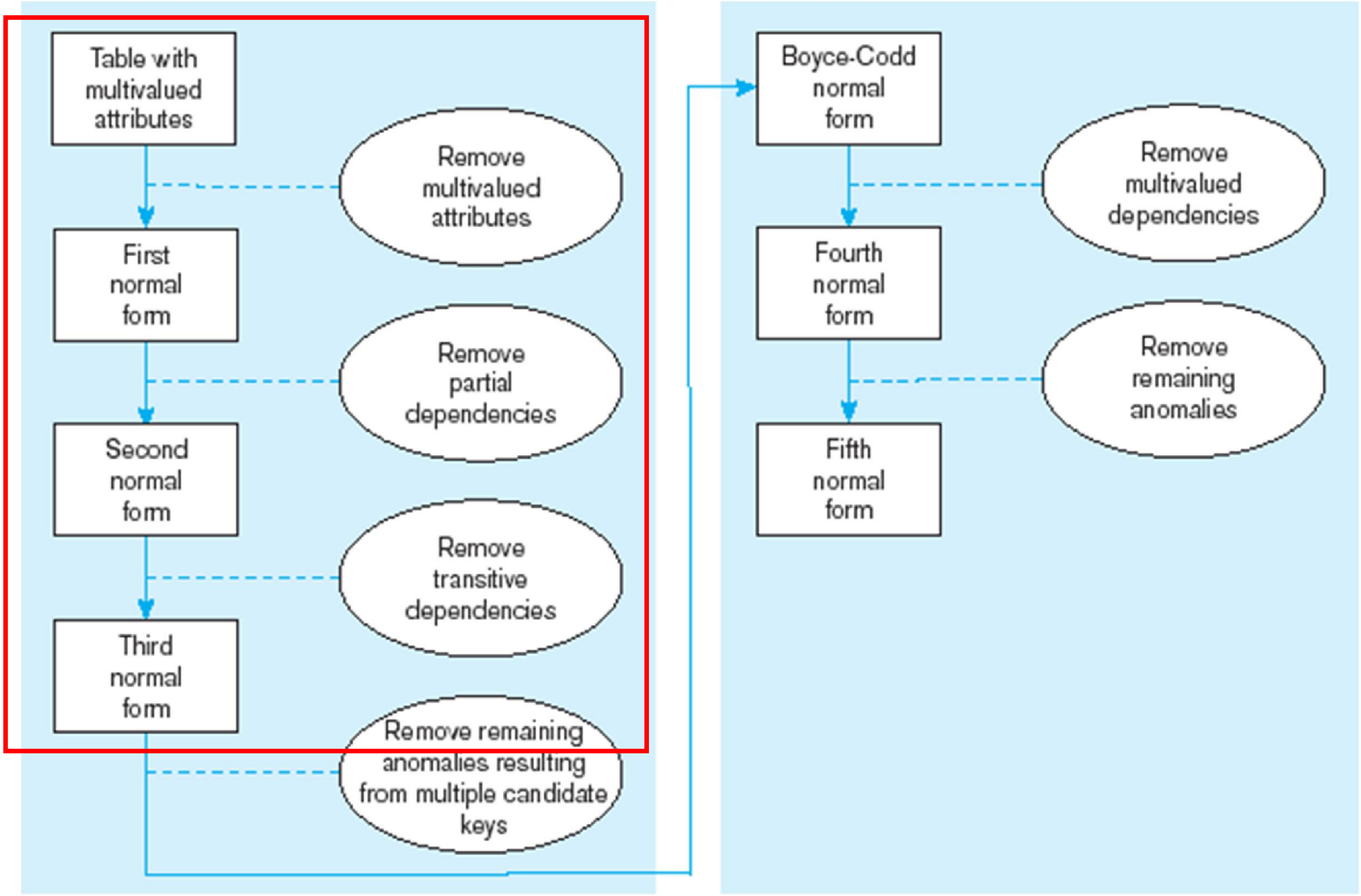


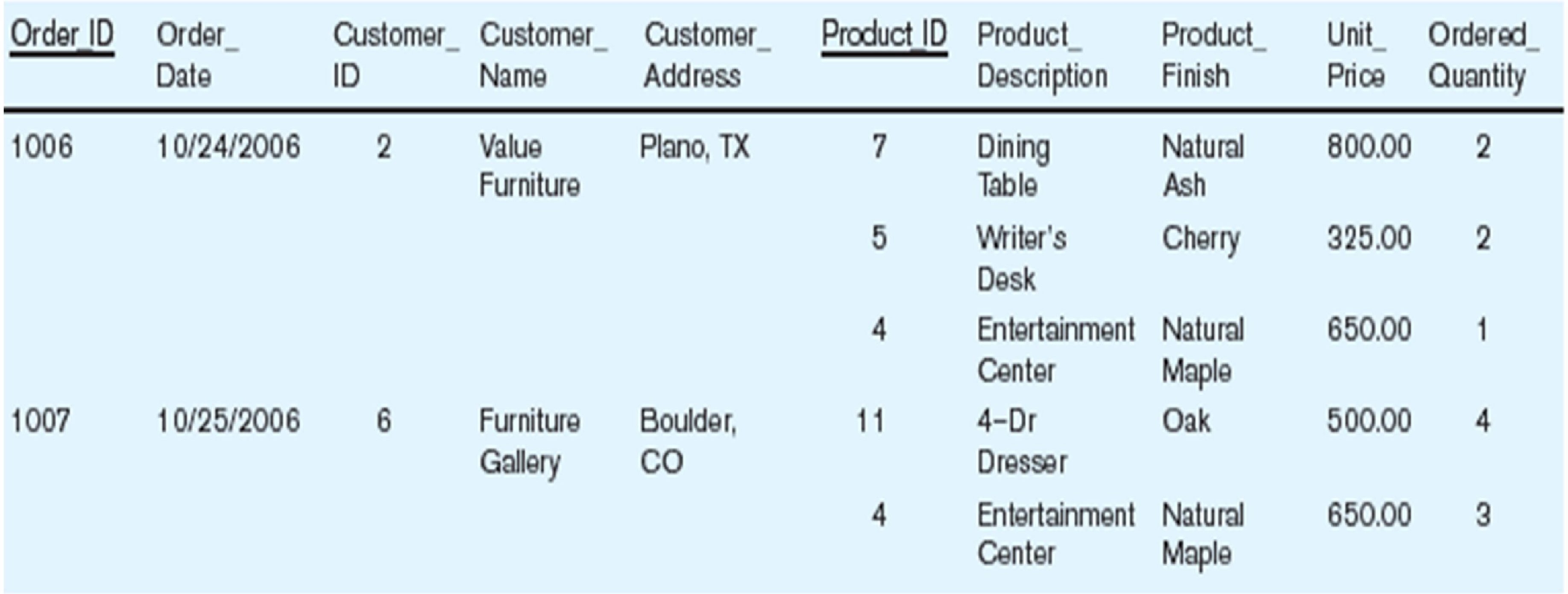
Table with multivalued attributes
Remove multivalued attributes
First normal form
Remove partial dependencies
Second normal form
Remove transitive dependencies
Third normal form
Remove remaining anomalies resulting from multiple candidate keys
Boyce-Codd normal form
Remove multivalued dependencies
Fourth normal form
Remove remaining anomalies
Fifth normal form

First Normal Form

- No multivalued attributes
- Every attribute value is atomic
- Fig. 5-25 *is not* in 1st Normal Form (multivalued attributes) ➔ it is not a relation
- Fig. 5-26 *is* in 1st Normal form
- ***All relations* are in 1st Normal Form**

Table with multivalued attributes, not in 1st normal form

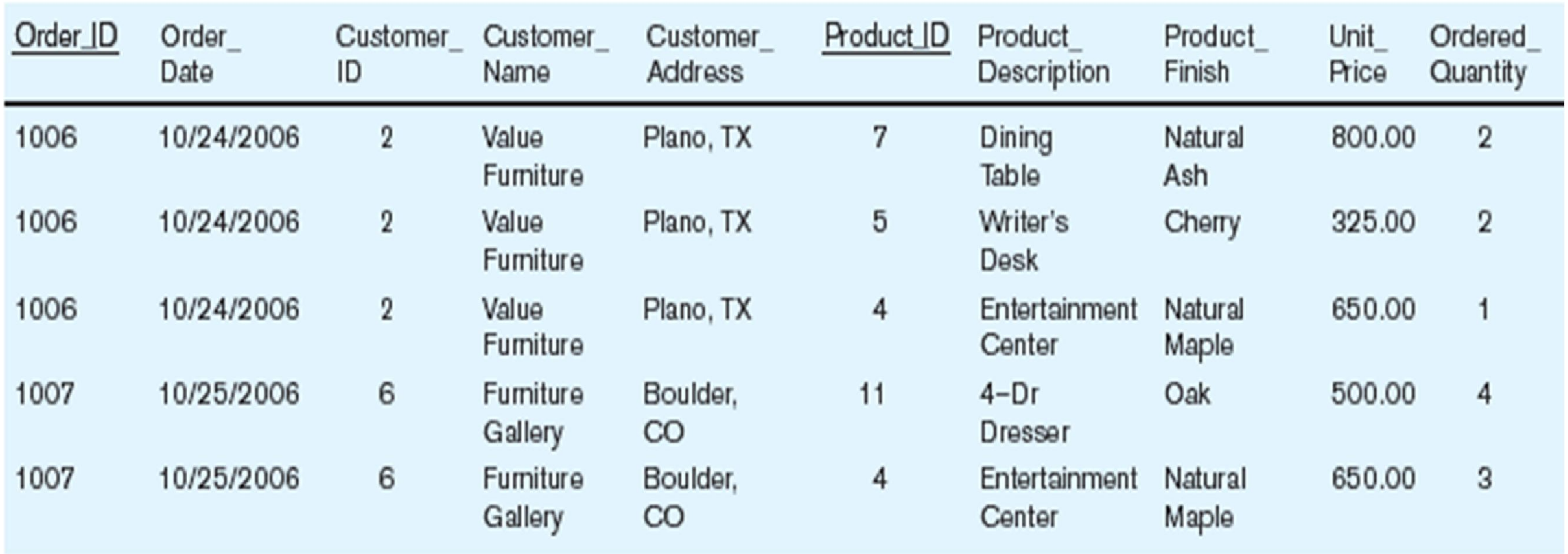
Figure 5-25
INVOICE data (Pine Valley Furniture Company)



Order_ID	Order_Date	Customer_ID	Customer_Name	Customer_Address	Product_ID	Product_Description	Product_Finish	Unit_Price	Ordered_Quantity
1006	10/24/2006	2	Value Furniture	Plano, TX	7	Dining Table	Natural Ash	800.00	2
					5	Writer's Desk	Cherry	325.00	2
					4	Entertainment Center	Natural Maple	650.00	1
1007	10/25/2006	6	Furniture Gallery	Boulder, CO	11	4–Dr Dresser	Oak	500.00	4
					4	Entertainment Center	Natural Maple	650.00	3

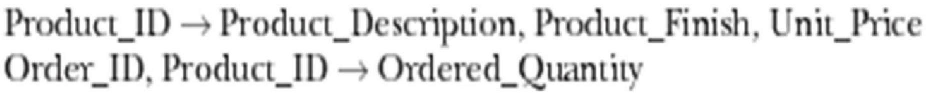
Note: this is NOT a relation

Table with no multivalued attributes and unique rows, in 1st normal form



Order_ID	Order_ Date	Customer_ ID	Customer_ Name	Customer_ Address	Product_ID	Product_ Description	Product_ Finish	Unit_ Price	Ordered_ Quantity
1006	10/24/2006	2	Value Furniture	Plano, TX	7	Dining Table	Natural Ash	800.00	2
1006	10/24/2006	2	Value Furniture	Plano, TX	5	Writer's Desk	Cherry	325.00	2
1006	10/24/2006	2	Value Furniture	Plano, TX	4	Entertainment Center	Natural Maple	650.00	1
1007	10/25/2006	6	Furniture Gallery	Boulder, CO	11	4–Dr Dresser	Oak	500.00	4
1007	10/25/2006	6	Furniture Gallery	Boulder, CO	4	Entertainment Center	Natural Maple	650.00	3

Figure 5-26
INVOICE relation (1NF) (Pine Valley Furniture Company)



Product_ID → Product_Description, Product_Finish, Unit_Price
Order_ID, Product_ID → Ordered_Quantity

Note: this is relation, but not a well-structured one

Anomalies in this Table

- **Insertion**–if new product is ordered for order 1007 of existing customer, customer data must be re-entered, causing duplication
- **Deletion**–if we delete the Dining Table from Order 1006, we lose information concerning this item's finish and price
- **Update**–changing the price of product ID 4 requires update in several records

Why do these anomalies exist?

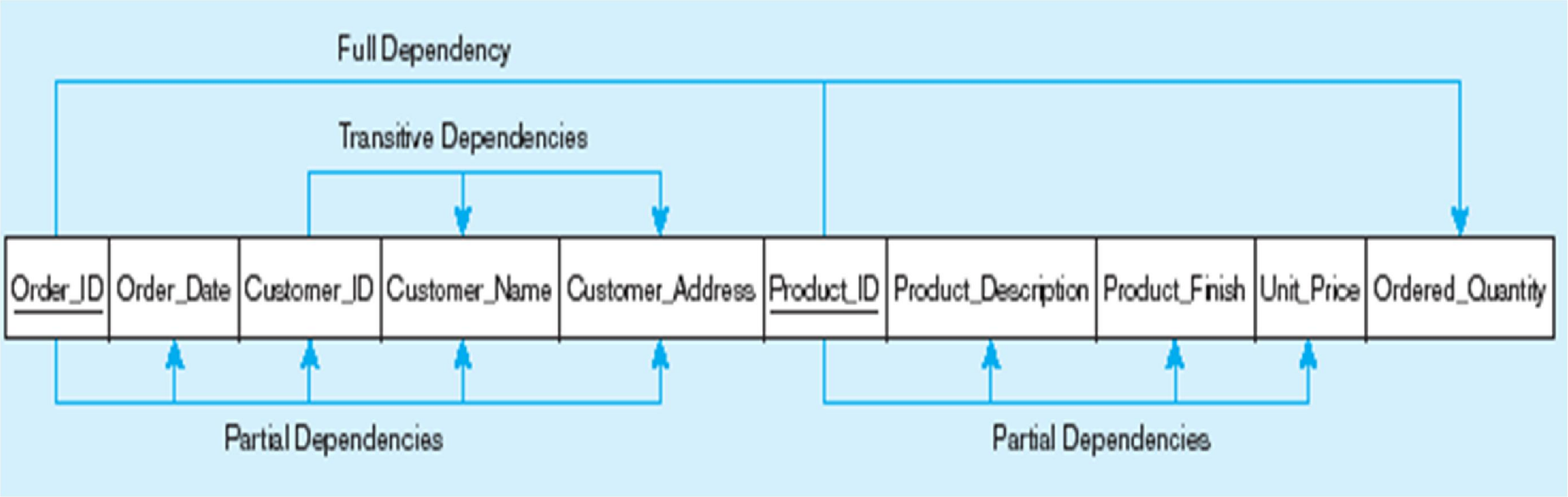
Because there are multiple themes (entity types) in one relation. This results in duplication and an unnecessary dependency between the entities

Second Normal Form

- Must be in 1st normal form and no partial dependency
- every ***non-key*** attribute is ***fully functionally dependent*** on the ***ENTIRE primary key***
  - Every non-key attribute must be defined by the entire key, not by only part of the key
  - No partial functional dependencies

we do this in the exam

Figure 5-27 Functional dependency diagram for INVOICE



Order_ID ➔ Order_Date, Customer_ID, Customer_Name, Customer_Address

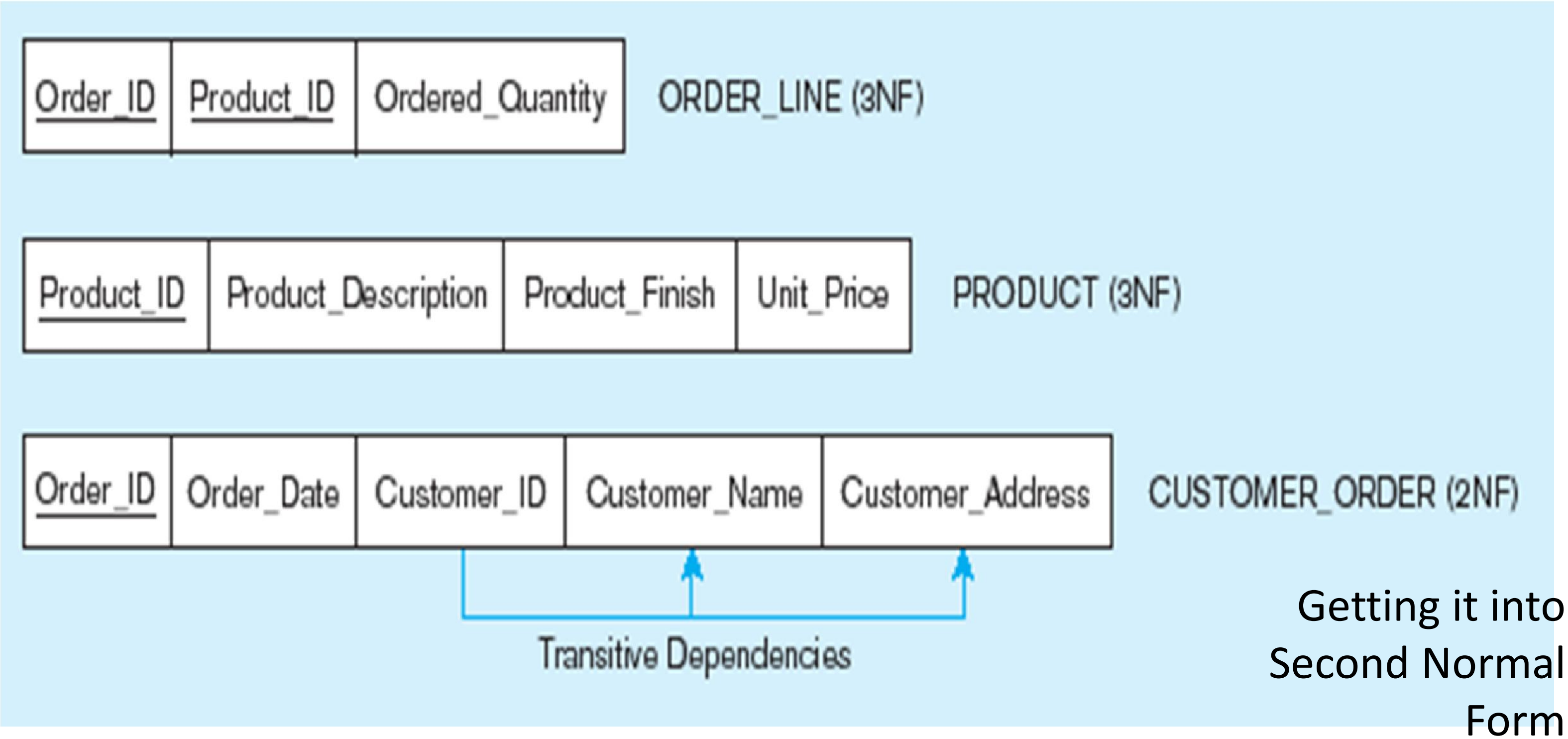
Customer_ID ➔ Customer_Name, Customer_Address

Product_ID ➔ Product_Description, Product_Finish, Unit_Price

Order_ID, Product_ID ➔ Order_Quantity

Therefore, NOT in 2nd Normal Form

Figure 5-28 Removing partial dependencies



Order_ID	Product_ID	Ordered_Quantity

ORDER_LINE (3NF)

Product_ID	Product_Description	Product_Finish	Unit_Price

PRODUCT (3NF)

Order_ID	Order_Date	Customer_ID	Customer_Name	Customer_Address

CUSTOMER_ORDER (2NF)

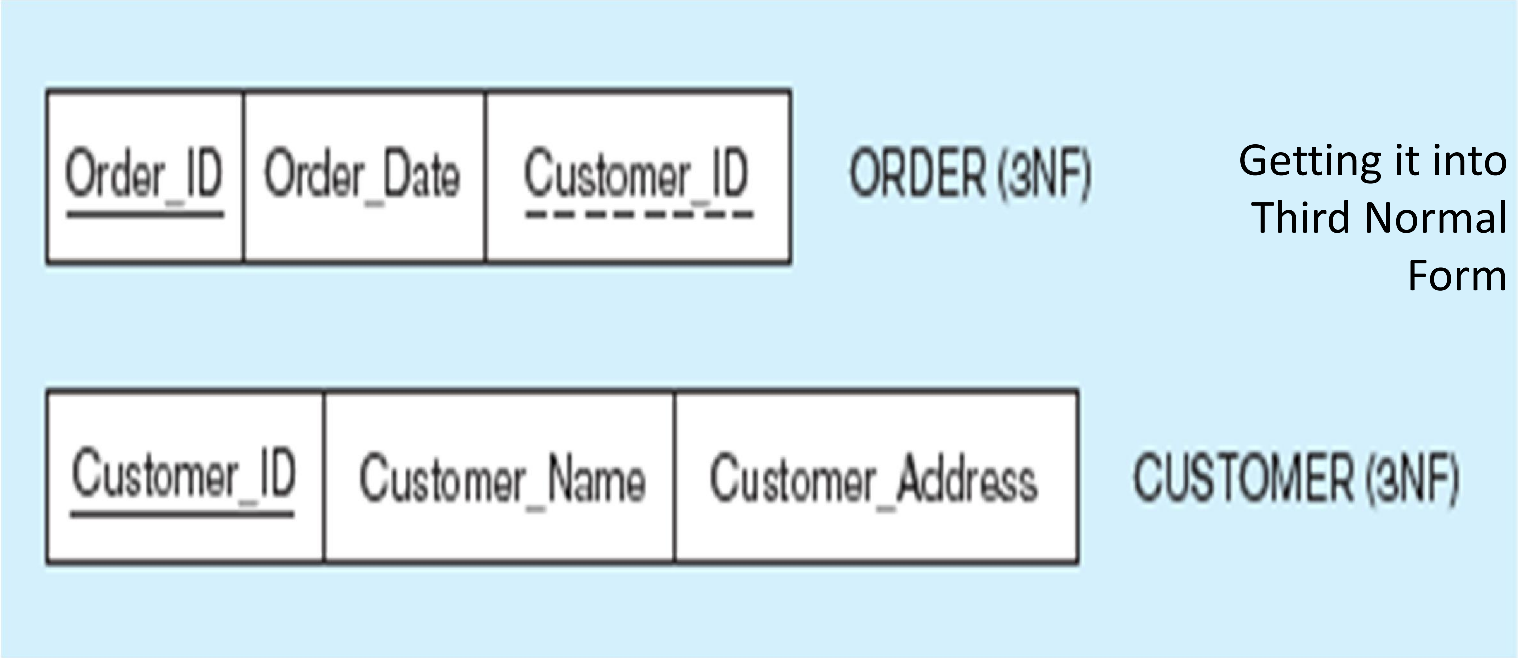
Getting it into Second Normal Form

Partial dependencies are removed, but there are still transitive dependencies

Third Normal Form

- 2NF PLUS ***no transitive dependencies*** (functional dependencies on non-primary-key attributes)
- Note: This is called transitive, because the primary key is a determinant for another attribute, which in turn is a determinant for a third
- Solution: Non-key determinant with transitive dependencies go into a new table; non-key determinant becomes primary key in the new table and stays as foreign key in the old table

Figure 5-28 Removing partial dependencies



Order_ID	Order_Date	Customer_ID

ORDER (3NF)

Customer_ID	Customer_Name	Customer_Address

CUSTOMER (3NF)

Getting it into Third Normal Form

Transitive dependencies are removed

not very important

Merging Relations

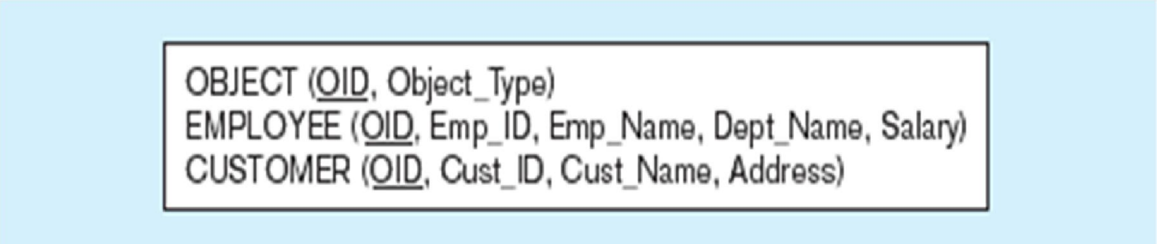
- **View Integration** – Combining entities from multiple ER models into common relations
- Issues to watch out for when merging entities from different ER models:
  - **Synonyms** –two or more attributes with different names but same meaning
  - **Homonyms** –attributes with same name but different meanings
  - **Transitive dependencies** –even if relations are in 3NF prior to merging, they may not be after merging
  - **Supertype/subtype** relationships –may be hidden prior to merging

not very important

Enterprise Keys

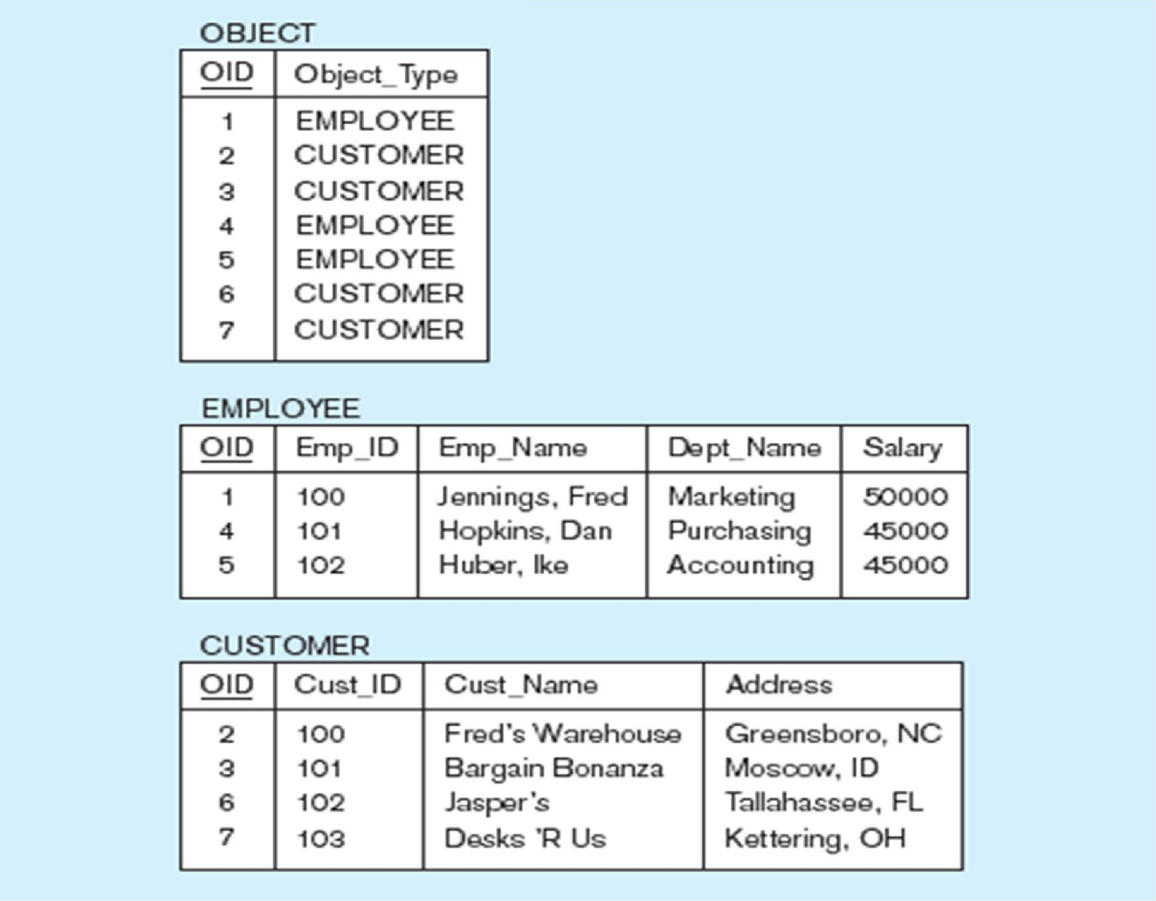
- Primary keys that are unique in the whole database, not just within a single relation
- Corresponds with the concept of an object ID in object-oriented systems

Figure 5-31 Enterprise keys



OBJECT (OID, Object_Type)
EMPLOYEE (OID, Emp_ID, Emp_Name, Dept_Name, Salary)
CUSTOMER (OID, Cust_ID, Cust_Name, Address)

a) Relations with enterprise key



OBJECT

OID	Object_Type
1	EMPLOYEE
2	CUSTOMER
3	CUSTOMER
4	EMPLOYEE
5	EMPLOYEE
6	CUSTOMER
7	CUSTOMER

EMPLOYEE

OID	Emp_ID	Emp_Name	Dept_Name	Salary
1	100	Jennings, Fred	Marketing	50000
4	101	Hopkins, Dan	Purchasing	45000
5	102	Huber, Ike	Accounting	45000

CUSTOMER

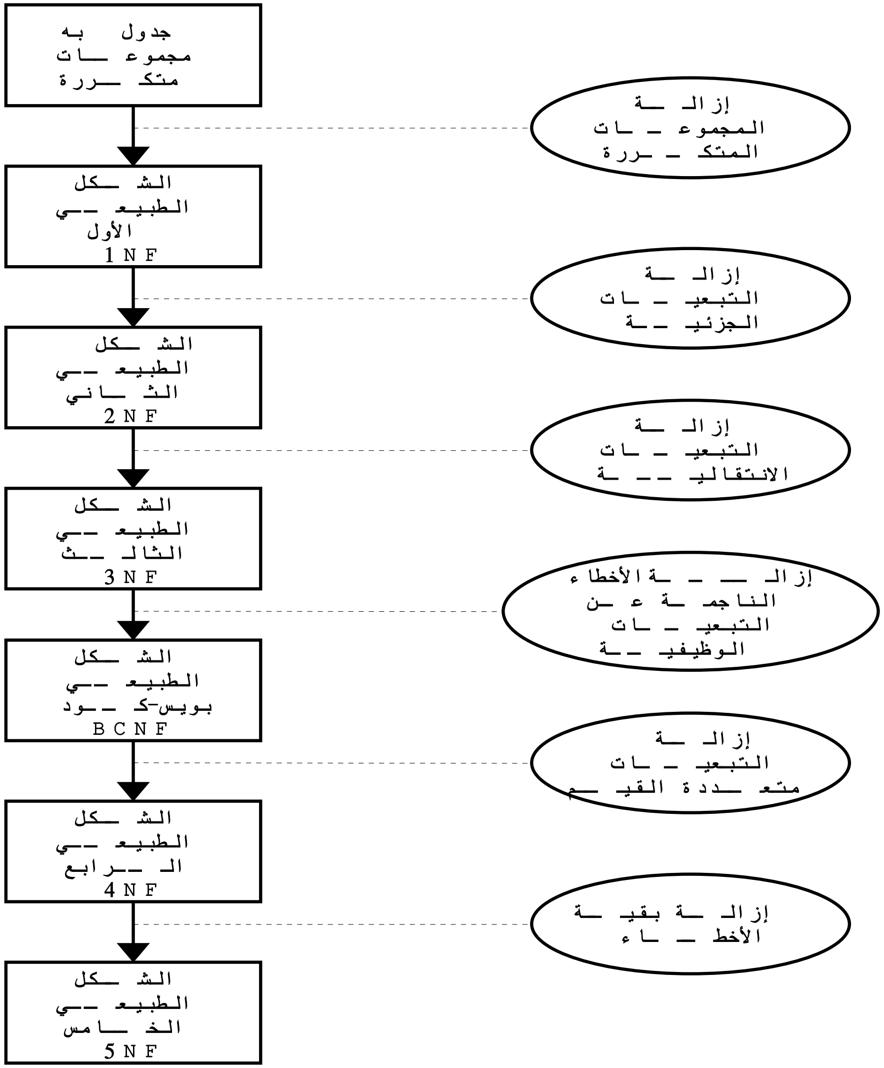
OID	Cust_ID	Cust_Name	Address
2	100	Fred's Warehouse	Greensboro, NC
3	101	Bargain Bonanza	Moscow, ID
6	102	Jasper's	Tallahassee, FL
7	103	Desks 'R Us	Kettering, OH

b) Sample data with enterprise key

التطبيع
(Normalization)

التطبيع هو تحويل هياكل البيانات المركبة إلى هياكل بيانات بسيطة ومُستقرة.

Steps in Normalization خطوات التطبيع



جدول به مجموعات متكررة

إزالة المجموعات المتكررة

الشكل الطبيعي الأول
1NF

إزالة التبعيات الجزئية

الشكل الطبيعي الثاني
2NF

إزالة التبعيات الانتقالية

الشكل الطبيعي الثالث
3NF

إزالة الأخطاء الناجمة عن التبعيات الوظيفية

الشكل الطبيعي بويس-كود
BCNF

إزالة التبعيات متعددة القيم

الشكل الطبيعي الرابع
4NF

إزالة بقية الأخطاء

الشكل الطبيعي الخامس
5NF

التبعيات الوظيفية و المفاتيح
Functional Dependence & Keys

التبعية الوظيفية هي علاقة معينة بين خاصتين. في العلاقة “R”، تعتبر الخاصية “B” تابعة
وظيفيًا للخاصية “A” إذا كانت كل قيمة “A” تحدد قيمة واحدة “B”. وتُمَثل هكذا A → B.

EMPCRS (EMP#, CRS#, DATE_COMPLETED)

EMP#, CRS# → DATE_COMPLETED

المحدد (Determinant) هو خصائص الجانب الأيسر للتبعية الوظيفية.

قواعد التبعيات الوظيفية
Rules of Functional Dependency

If X, Y, Z, and W are attributes in a relation, then:

1. X → X (reflexivity)الارتداد

2. If X → Y then XZ → Y (augmentation)الازدياد

3. If X → Y and X → Z then X → YZ (union)الاتحاد

4. If X → Y then X → Z where Z is a subset of Y (decomposition)التفكيك

5. If X → Y and Y → Z then X → Z (transitivity)الانتقالية

6. If X → Y and YZ → W then XZ → W (pseudotransitivity)الانتقالية الزائفة

أمثلة لقواعد التبعيات الوظيفية

1.الازدياد

STD# → STD_NAME then STD#, CRS# → STD_NAME

2.الانتقالية

STD# → MAJOR and MAJOR → ADVISOR

then STD# → ADVISOR

3.الانتقالية الزائفة

STD# → MAJOR and MAJOR, CLASS → ADVISOR

then STD#, CLASS → ADVISOR

الأشكال الطبيعية الأساسية
The Basic Normal Forms

GRADE REPORT

FALL SEMESTER

NAME : Saad Aldousary STUDENT#: 2773777
ADDRESS : P.O. Box 777 Riyadh 11147
MAJOR : Information Systems

COURSE#	TITLE	INST. NAME	INST. LOC.	GRADE
IS 350	Database Mgt	Saleh	1024	A
IS 465	System Analysis	Ahmad	1030	B

عينة بيانات تقرير الدرجات

GRADE_REPORT

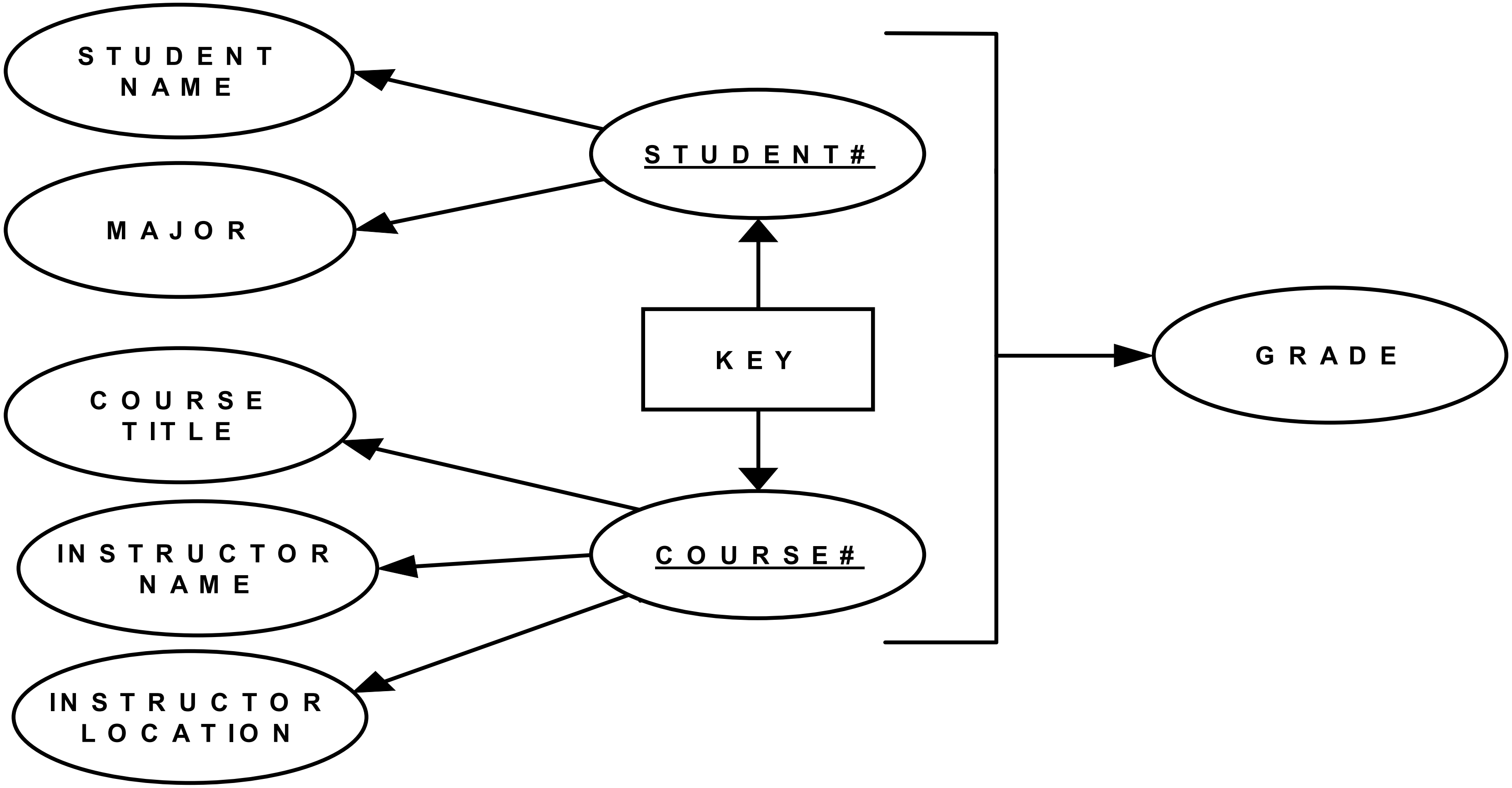
STUDENT#	STUDENT NAME	MAJOR	COURSE#	COURSE TITLE	INSTRUCTOR NAME	INSTRUCTOR LOCATION	GRADE
2773777	Saad	IS	IS 350 IS 465	Database Mgt System Analysis	Saleh Ahmad	1024 1030	A B
6917773	Ali	PM	IS 465 PM 300 QM 440	System Analysis Production Mgt Operations Res	Ahmad Soud Ahmad	1030 1025 1030	C A B
…							

الشكل الطبيعي الأول (1NF)

GRADE_REPORT

STUDENT#	STUDENT NAME	MAJOR	COURSE#	COURSE TITLE	INSTRUCTOR NAME	INSTRUCTOR LOCATION	GRADE
2773777	Saad	IS	IS 350	Database Mgt	Saleh	1024	A
2773777	Saad	IS	IS 465	System Analysis	Ahmad	1030	B
6917773	Ali	PM	IS 465	System Analysis	Ahmad	1030	C
6917773	Ali	PM	PM 300	Production Mgt	Soud	1025	A
6917773	Ali	PM	QM 440	Operations Res	Ahmad	1030	B
…							

الشكل الطبيعي الثاني (2NF)



الشكل الطبيعي الثاني (2NF)

تحليل التبعيات الوظيفية

1. STUDENT(STUDENT#, STUDENT_NAME,MAJOR)
2. COURSE_INSTRUCTOR(COURSE#, COURSE_TITLE, INSTRUCTOR_NAME, INSTRUCTOR_LOCATION)
3. REGISTRATION(STUDENT#, COURSE#, GRADE)

العلاقة في الشكل الطبيعي الثاني إذا كانت في
الشكل الطبيعي الأول ولا تحتوي على تبعيات جزئية.

STUDENT

STUDENT#	STUDENT NAME	MAJOR
2773777	Saad	IS
6917773	Ali	PM
…		

RGISTRATION

STUDENT#	COURSE#	GRADE
2773777	IS 350	A
2773777	IS 465	B
6917773	IS 465	C
6917773	PM 300	A
6917773	QM 440	B
…		

COURSE_INSTRUCTOR

COURSE#	COURSE TITLE	INSTRUCTOR NAME	INSTRUCTOR LOCATION
IS 350	Database Mgt	Saleh	1024
IS 465	System Analysis	Ahmad	1030
PM 300	Production Mgt	Soud	1025
QM 440	Operations Res	Ahmad	1030

الشكل الطبيعي الثالث (3NF)

- التبعيات الوظيفية في ”COURSE_INSTRUCTOR”

1. COURSE# → COURSE_TITLE, INSTRUCTOR_NAME, INSTRUCTOR_LOCATION

2. INSTRUCTOR_NAME → INSTRUCTOR_LOCATION

COURSE

COURSE #	COURSE TITLE	INSTRUCTOR NAME
IS 350	Database Mgt	Saleh
IS 465	System Analysis	Ahmad
PM 300	Production Mgt	Soud
QM 440	Operations Res	Ahmad

INSTRUCTOR

INSTRUCTOR NAME	INSTRUCTOR LOCATION
Saleh	1024
Ahmad	1030
Soud	1025
...	

الشكل الطبيعي الثالث (3NF)

العلاقة في الشكل الطبيعي الثالث إذا كانت في الشكل الطبيعي الثاني ولا تحتوي علي تبعيات انتقالية.

1. STUDENT(STUDENT#, STUDENT_NAME,MAJOR)
2. COURSE_INSTRUCTOR(COURSE#, COURSE_TITLE, INSTRUCTOR_NAME)
3. INSTRUCTOR(INSTRUCTOR_NAME, INSTRUCTOR_LOCATION)
4. REGISTRATION(STUDENT#, COURSE#, GRADE)

أشكال طبيعية إضافية
(Additional Normal Forms)

- الشكل الطبيعي بويس- كود
(Boyce-Codd Normal Form “BCNF”)

STUDENT_MAJOR_ADVISOR

STUDENT#	MAJOR	ADVISOR
123	PHYSICS	EINSTEIN
123	MUSIC	MOZART
456	BIOL	DARWIN
789	PHYSICS	BOHR
999	PHYSICS	EINSTEIN

STUDENT#, MAJOR → ADVISOR

ADVISOR → MAJOR

تابع الشكل الطبيعي بويس-كود

العلاقة في الشكل الطبيعي بويس- كود “BCNF” إذا كان كل محدد فيها مفتاح مرشح.

ST_ADV

STUDENT	ADVISOR
123	EINSTEIN
123	MOZART
456	DARWIN
789	BOHR
999	EINSTEIN

ADV_MAJ

ADVISOR	MAJOR
EINSTEIN	PHYSICS
MOZART	MUSIC
DARWIN	BIOL
BOHR	PHYSICS

الشكل الطبيعي الرابع
(4th Normal Form “4NF”)

OFFERING

COURSE	INSTRUCTOR	TEXTBOOK
Management	Ali	Drucker
Management	Ahmad	Drucker
Management	Saad	Drucker
Management	Ali	Peters
Management	Ahmad	Peters
Management	Saad	Peters
Finance	Gamil	Weston
Finance	Gamil	Gulford

COURSE →→ INSTRUCTOR

COURSE →→ TEXTBOOK

الشكل الطبيعي الرابع
(4th Normal Form“4NF”)

العلاقة في الشكل الطبيعي الرابع اذا كانت في الشكل الطبيعي بويس-كود ”BCNF“ ولا تحتوي على تبعيات متعددة القيم.

TEACHER

<u>COURSE</u>	<u>INSTRUCTOR</u>
Management	Ali
Management	Ahmad
Management	Saad
Finance	Gamil

TEXT

<u>COURSE</u>	<u>TEXTBOOK</u>
Management	Drucker
Management	Peters
Finance	Weston
Finance	Gulford

(Merging Relations) دمج العلاقات

EMPLOYEE1(EMP#, NAME, ADDRESS, PHONE)

EMPLOYEE2(EMP#, NAME, ADDRESS, JOBCODE, #YEARS)

تمثلان نفس الكينونة ويمكن دمجهما لتُكَوِّنا العلاقة:

EMPLOYEE(EMP#, NAME, ADDRESS, PHONE, JOBCODE, #YEARS

مشكلات دمج العلاقات

- المترادفات: (Synonyms)

يجب إعطاء أسماء قياسية للخصائص المترادفة عند الدمج وحذف المترادفات الأخرى.

- تماثل الأسماء واختلاف المعنى (Homonyms)

STUDENT1(STD#, NAME, ADDRESS)

STUDENT2(STD#, NAME, PHON#, ADDRESS)

عنوان الطالب في العلاقة الأولي هو عنوانه في الجامعة فى حين أن عنوانه في العلاقة الثانية هو عنوانه المنزلي.

HOME_ADD)STUDENT(STD#, NAME, PHON#, CAMPUS_ADD,

مشكلات دمج العلاقات

- التبعيات الانتقالية (Transitive Dependencies)

STUDENT1(STUDENT_ID, MAJOR)

STUDENT2(STUDENT_ID, ADVISOR)

بدمج هاتين العلاقتين تنتج العلاقة:

STUDENT(STUDENT_ID, MAJOR, ADVISOR)

بفرض الحالة MAJOR → ADVISOR، تصبح العلاقة STUDEN في الشكل الطبيعي الثاني ويتم تحويلها للشكل الطبيعي الثالث كالتالي:

STUDENT(STUDENT_ID, MAJOR)

MAJ_ADV (MAJOR, ADVISOR)