Chapter 7

Normalization

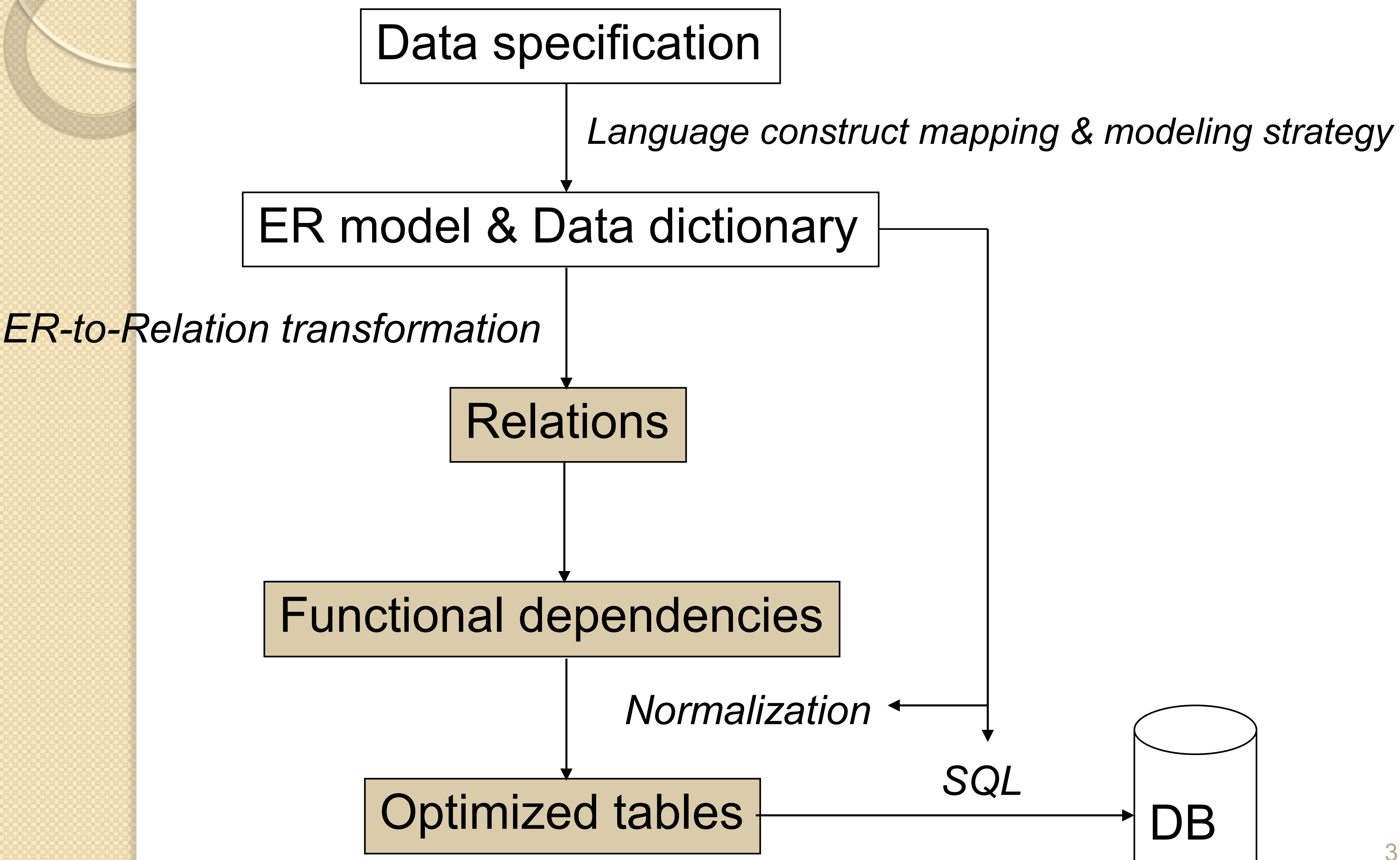
การทำให้เป็นบรรทัดฐาน

ผศ.ดร.เทพฤทธิ์ บัณฑิตวัฒนาวงศ์

Content

- 1NF
- 2NF
- 3NF
- BCNF
- 4NF
- 5NF
- Quick NF evaluation techniques

A Big Picture of RDB Development



Normalization

- An analysis of attribute interdependencies
  - To simplify relation representation and queries,
  - To reduce data redundancy so that insert/delete/update performance is improved and data inconsistency is reduced, and
  - To prevent information loss and update anomalies.

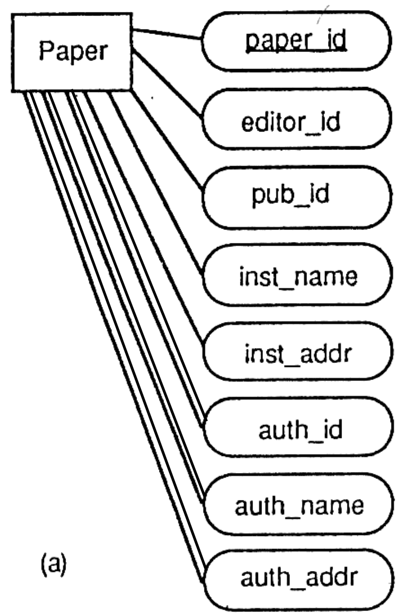
Six Levels of Normalization

- Each relation is normalized step-by-step from 1NF to the highest NF.

1. First normal form (1NF)
2. Second normal form (2NF)
3. Third normal form (3NF)
4. Boyce-Codd normal form (BCNF)
5. Fourth normal form (4NF)
6. Fifth normal form (5NF)

1NF

- Problems in UNF.



(a)

- Need table expansion & shrinking.
- Impose retrieval complexity.

Unnormalized relation(UNR) :

paper_id	inst_name	inst_addr	editor_id	pub_id	auth_id1	auth_name1	auth_addr1
4216	univ_mich	ann_arbor	woolf	14	7631	yang_d	peking_univ
5789	math_rev	providenc	bradlee	53	1126	umar_a	bellcore

auth_id2	auth_name2	auth_addr2	auth_id3	auth_name3 ....
4419	mantei_m	univ_toron	2692	koenig_j
7384	fry_j	mitre	3633	bolton_d

(b)

1NF (cont.)

- Partly solved by table restructuring.

paper_id	inst_name	inst_addr	editor_id	pub_id	auth_id	auth_name	auth_addr
4216	univ_mich	ann_arbor	woolf	14	7631	yang_d	peking_univ
4216	univ_mich	ann_arbor	woolf	14	4419	mantei_m	univ_toron
4216	univ_mich	ann_arbor	woolf	14	2692	koenig_j	math_rev
5789	math_rev	providenc	bradlee	53	1126	umar_a	bellcore
5789	math_rev	providenc	bradlee	53	7384	fry_j	mitre
5789	math_rev	providenc	bradlee	53	3633	bolton_d	math_rev

- Row indexing is still not possible.

1NF (cont.)

- Definition: A relation will be in 1NF if and only if there is no repeating groups.

1NF (cont.)

- 1NF of the previous UNF table.

Normalized relation (NR) :

paper_id	inst_name	inst_addr	editor_id	pub_id	auth_id	auth_name	auth_addr
4216	univ_mich	ann_arbor	woolf	14	7631	yang_d	peking_univ
4216	univ_mich	ann_arbor	woolf	14	4419	mantei_m	univ_toron
4216	univ_mich	ann_arbor	woolf	14	2692	koenig_j	math_rev
5789	math_rev	providenc	bradlee	53	1126	umar_a	bellcore
5789	math_rev	providenc	bradlee	53	7384	fry_j	mitre
5789	math_rev	providenc	bradlee	53	3633	bolton_d	math_rev

- Remark: this relation will be used as our running example for 1NF to 3NF.

1NF (cont.)

- Problems with 1NF:
  - **Insertion anomaly** occurs when inserting a new record causes many data items to be duplicated or violates entity integrity.
  - **Modification anomaly** occurs when modifying a record causes subsequent modifications in many other records.
  - **Deletion anomaly** occurs when remove record causes undesired information loss.
- Solved by transforming 1NF to 2NF.

2NF

- The property that attribute(s) uniquely identifies other attribute(s) is called **functional dependency (FD)**.
  - FD definition: given a relation R, a set of attributes B (called **nondeterminant**) is functionally dependent on another set of attributes A (called **determinant**) if each A value is associated with only one B value. Such an FD is denoted by A → B.

2NF (cont.)

- R(<u>Paper_id</u>, Inst_name, Inst_addr, Editor_id, Pub_id, <u>Auth_id</u>, Auth_name, Auth_addr) has the following FDs:

1. Paper_id, Auth_id → Auth_name
2. Paper_id, Auth_id → Auth_addr
3. Paper_id, Auth_id → Editor_id
4. Paper_id, Auth_id → Pub_id
5. Paper_id, Auth_id → Inst_name
6. Paper_id, Auth_id → Inst_addr
7. Paper_id → Editor_id
8. Paper_id → Pub_id
9. Auth_id → Auth_name
10. Auth_id → Auth_addr
11. Inst_name → Inst_addr

2NF (cont.)

- Equivalent shorthand form:
  - Paper_id, Auth_id → Inst_name, Inst_addr, Editor_id, Pub_id, Auth_name, Auth_addr
  - Paper_id → Editor_id, Pub_id
  - Auth_id → Auth_name, Auth_addr
  - Inst_name → Inst_addr

2NF (cont.)

- **Fully FD**: an FD whose nondeterminant fully depends on its determinant.
- Based on the running example, fully FDs are:
  1. Paper_id → Editor_id
  2. Paper_id → Pub_id
  3. Auth_id → Auth_name
  4. Auth_id → Auth_addr
  5. Inst_name → Inst_addr
  6. Paper_id, Auth_id → Inst_name
  7. Paper_id, Auth_id → Inst_addr

2NF (cont.)

- **Partially FD**: an FD whose nondeterminant depends on not all attributes composing its determinant.
- Based on the running example, partially FDs are:

1. Paper_id, Auth_id → Editor_id
2. Paper_id, Auth_id → Pub_id
3. Paper_id, Auth_id → Auth_name
4. Paper_id, Auth_id → Auth_addr

2NF (cont.)

- Definition: A relation is 2NF if and only if every nonkey attribute is <u>fully dependent</u> on a primary key.

2NF (cont.)

- To solve the problems of 1NF, break down the running example relation as follows:
  - R1: Paper_id, Auth_id → Inst_name, Inst_addr
  - R2: Auth_id → Auth_name, Auth_addr
  - R3: Paper_id → Pub_id, Editor_id

  that are
  - R1(<u>Paper_id, Auth_id</u>, Inst_name, Inst_addr)
  - R2(<u>Auth_id</u>, Auth_name, Auth_addr)
  - R3(<u>Paper_id</u>, Pub_id, Editor_id)

2NF (cont.)

- Problems with 1NF may remain in 2NF but less severe.
  - Modification anomaly
  - Insertion anomaly
  - Deletion anomaly
- Solved by transforming 2NF to 3NF.

3NF

- Definition: A relation is in 3NF if and only if it is in 2NF and there is no **transitive dependency** with respect to a primary key.
  - Transitive dependency means if X $\rightarrow$ Y then there exists X $\rightarrow$ Z and Z $\rightarrow$ Y with X as a pk and Z is not a candidate key.
- Simplified definition: A relation is in 3 NF if it is 2NF and there is no functional dependency between nonkey attributes.

3NF (cont.)

- 2NF
  - R1: Paper_id, Auth_id → Inst_name, Inst_addr

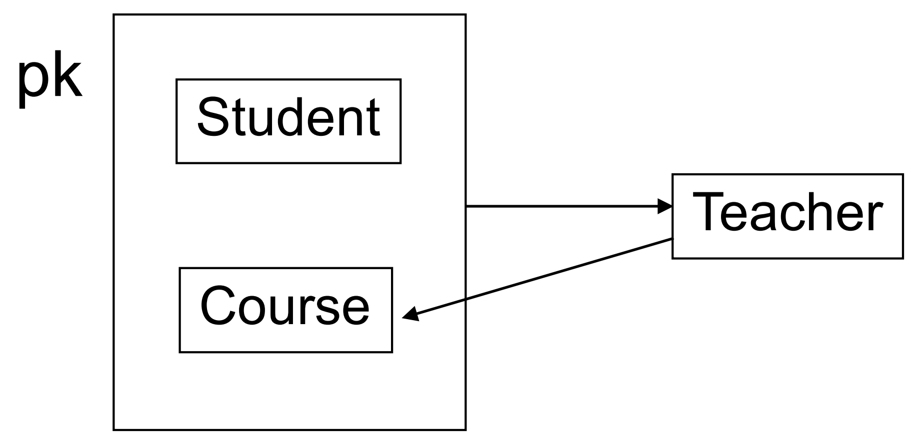
    Inst_name → Inst_addr
  - R2: Auth_id → Auth_name, Auth_addr
  - R3: Paper_id → Pub_id, Editor_id
- 3NF: splits R1 to R11 and R12
  - R11: Paper_id, Auth_id → Inst_name

    R12: Inst_name → Inst_addr
  - R2: Auth_id → Auth_name, Auth_addr
  - R3: Paper_id → Pub_id, Editor_id

3NF (cont.)

- A 3NF relation with deletion anomaly in remain.

<u>Student</u>	**<u>Course</u>**	**Teacher**
Smith	Math	Prof. White
Smith	Physics	Prof. Green
Jones	Math	Prof. White
Jones	Physics	Prof. Brown



Boyce-Codd NF (BCNF)

- Definition: A relation is in Boyce-Codd NF if and only if every determinant is super key.

BCNF (cont.)

- Pattern 1

  R(A,B,C): A, B → C

  C → B

  R is 3NF but not BCNF because C → B

- Solution

  Split R into 2 relations:

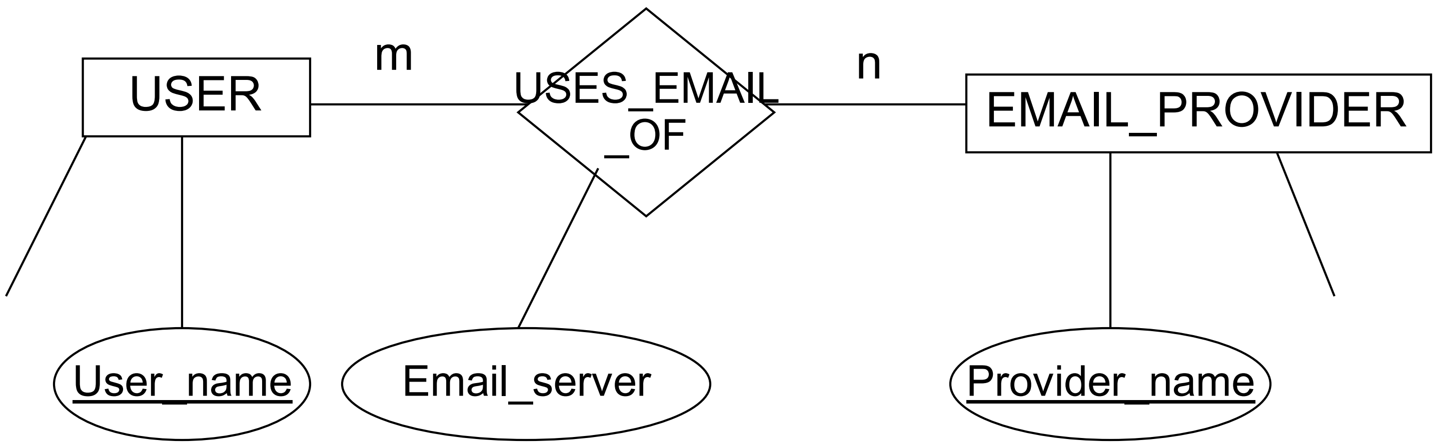
  R1: A, C that is R1(A,C)

  R2: C → B that is R2(C,B)

  R1 and R2 are in (3NF and) BCNF.

BCNF (cont.)

- Example of Pattern 1



User_name	Provider_name	Email_server
Aloha	Hotmail	mail1.hotmail.com
David	Hotmail	mail2.hotmail.com
David	Google_mail	mx.gmail.com
Peter	Google_mail	mx.gmail.com

BCNF (cont.)

- USE_EMAIL_OF:
  - User_name, Provider_name → Email_server
  - Email_server → Provider_name
- This is 3NF but BCNF because of the second FD.
- Split the relation so that each of which is in BCNF.

User_name	**Email_server**
Aloha	mail1.hotmail.com
David	mail2.hotmail.com
David	mx.gmail.com
Peter	mx.gmail.com

User_name, Email_server (No FD)

Email_server	**Provider_name**
mail1.hotmail.com	Hotmail
mail2.hotmail.com	Hotmail
mx.gmail.com	Google_mail

FD: Email_server → Provider_name

BCNF (cont.)

- Exercise: Make the previous relation R(<u>Student, Course,</u> Teacher) BCNF.

BCNF (cont.)

- Pattern 2

  R(A,B,C,D): A, B → C, D

  C → B

  R is 3NF but not BCNF because C → B.

- Solution

  Split R into 2 relations:

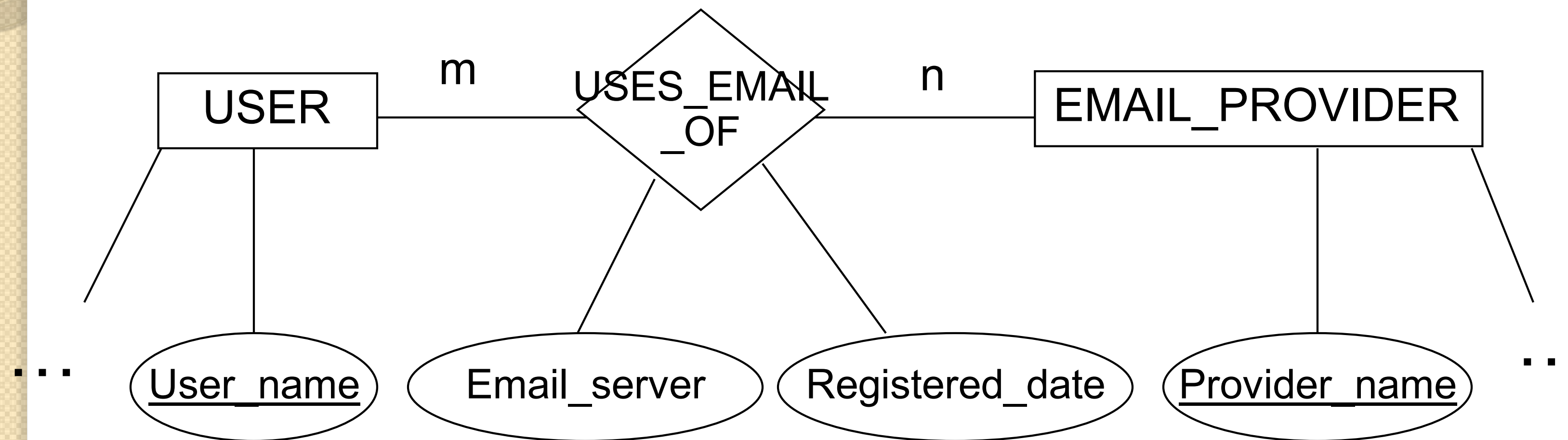
  R1: A, C → D

  R2: C → B

  R1 and R2 are in (3NF and) BCNF.

BCNF (cont.)

- Example of Pattern 2



User_name	Provider_name	Email_server	Registered_date
Aloha	Hotmail	mail1.hotmail.com	1/1/2000
David	Hotmail	mail2.hotmail.com	2/1/2000
David	Google_mail	mx.gmail.com	3/1/2000
Peter	Google_mail	mx.gmail.com	2/1/2000

BCNF (cont.)

- FD1&2: User_name, Provider_name → Email_server, Registered_date
- FD3: Email_server → Provider_name

  This is 3NF but BNCF because of the third FD.

- The relation is in 3NF but not BCNF, thus split the relation.

User_name	Email_server	Registered_date
Aloha	mail1.hotmail.com	1/1/2000
David	mail2.hotmail.com	2/1/2000
David	mx.gmail.com	3/1/2000
Peter	mx.gmail.com	2/1/2000

Email_server	Provider_name
mail1.hotmail.com	Hotmail
mail2.hotmail.com	Hotmail
mx.gmail.com	Google_mail

FD: Email_server → Provider_name

FD: User_name, Email_server → Registered_date

BCNF (cont.)

Course	Teacher	Textbook
Physics	Green	Mechanics
Physics	Green	Optics
Physics	Brown	Mechanics
Physics	Brown	Optics
Math	Green	Mechanics
Math	Green	Vector
Math	Green	Trigonometry

- This relation is BCNF but there remain the problems of insertion, deletion and modification anomalies.

4 NF

- **Multi-valued dependence (MVD)** holds in a relation R(A,B,C) if a set of values B and a set of values C are determined by A value, and that A determines B must be independent of that A determines C. This is denoted by:
  - A →→ B and A →→ C, or
  - A →→ B|C

4NF (cont.)

- Example of MVD

determines

<u>Course</u>	<u>Teacher</u>	<u>Textbook</u>
Physics	Green	Mechanics
		Optics
	Brown	Mechanics
		Optics
Math	Green	Mechanics
		Vector
		Trigonometry

4NF (cont.)

- FDs are actually a special form of MVD in that, for instance, a value A in R(<u>A</u>, B, C) <u>determines a single value</u> of B and C.
- Example: R2(<u>Auth_id</u>, Auth_name, Auth_addr)
  - MVD: Auth_id →→ Auth_name|Auth_addr
  - FDs: Auth_id → Auth_name, Auth_addr

4NF (cont.)

- Definition: A relation is 4NF if and only if it contains no MVD or MVD is FDs.
- Relation R(<u>A,B,C</u>) can be lossless-decomposed into two relations R1(<u>A,B</u>) and R2(<u>A,C</u>) if and only if the MVD A→→B|C holds in R.

4NF (cont.)

Example of lossless decomposition.

Course	Teacher	Textbook
Physics	Green	Mechanics
Physics	Green	Optics
Physics	Brown	Mechanics
Physics	Brown	Optics
Math	Green	Mechanics
Math	Green	Vector
Math	Green	Trigonometry

Course →→ Teacher

Course	Teacher
Physics	Green
Physics	Green
Physics	Brown
Physics	Brown
Math	Green
Math	Green
Math	Green

Course→→Textbook

Course	Textbook
Physics	Mechanics
Physics	Optics
Physics	Mechanics
Physics	Optics
Math	Mechanics
Math	Vector
Math	Trigonometry

4NF (cont.)

- The decomposed relations.

Course	Teacher
Physics	Green
Physics	Brown
Math	Green

Course	Textbook
Physics	Mechanics
Physics	Optics
Math	Mechanics
Math	Vector
Math	Trigonometry

4NF (cont.)

- This relation is 4NF due to no MVD but there remain the problems of insertion, deletion and update anomalies, which are caused by a **cyclic constraint**.

Supplier_no	Product_no	Project_no
S1	P1	J2
S1	P2	J1
S2	P1	J1
S1	P1	J1

5NF

- **Join dependency (JD)** *(X,Y,...,Z) holds in a relation R if and only if the relation can be derived by joining the projections on X, Y, ..., Z, where X, Y, ..., Z are subsets of the attributes of R.
- JD is **trivial** if one of the projections is R itself.

- Examining JD by two projections.
- No JD found.

Supplier_no	Product_no	Project_no
S1	P1	J2
S1	P2	J1
S2	P1	J1
S1	P1	J1

Supplier_no	Product_no
S1	P1
S1	P2
S2	P1
S1	P1

Product_no	Project_no
P1	J2
P2	J1
P1	J1
P1	J1

Supplier_no	Product_no	Project_no
S1	P1	J2
S1	P1	J1
S1	P2	J1
S2	P1	J2
S2	P1	J1

Spurious row

- Examining JD by three projections.
- A JD is found.

Supplier_no	Product_no	Project_no
S1	P1	J2
S1	P2	J1
S2	P1	J1
S1	P1	J1

Supplier_no	Product_no
S1	P1
S1	P2
S2	P1
S1	P1

Product_no	Project_no
P1	J2
P2	J1
P1	J1
P1	J1

Supplier_no	Project_no
S1	J2
S1	J1
S2	J1
S1	J1

Supplier_no	Product_no	Project_no
S1	P1	J2
S1	P1	J1
S1	P2	J1
S2	P1	J2
S2	P1	J1

Join over (Supplier_no, Project_no)

The relation before decomposition

JD: *((Supplier_no,Product_no), (Product_no,Project_no), (Supplier_no, Project_no))

5NF (cont.)

- AKA. **Project-join normal form (PJNF)**
- Definition: A relation is 5NF if and only if either <u>there is only trivial JD</u> or <u>if there is nontrivial JD(s), every projection of every nontrivial JD is implied by candidate key(s)</u>.
- To make a relation 5NF, splitting it according to JD projections.

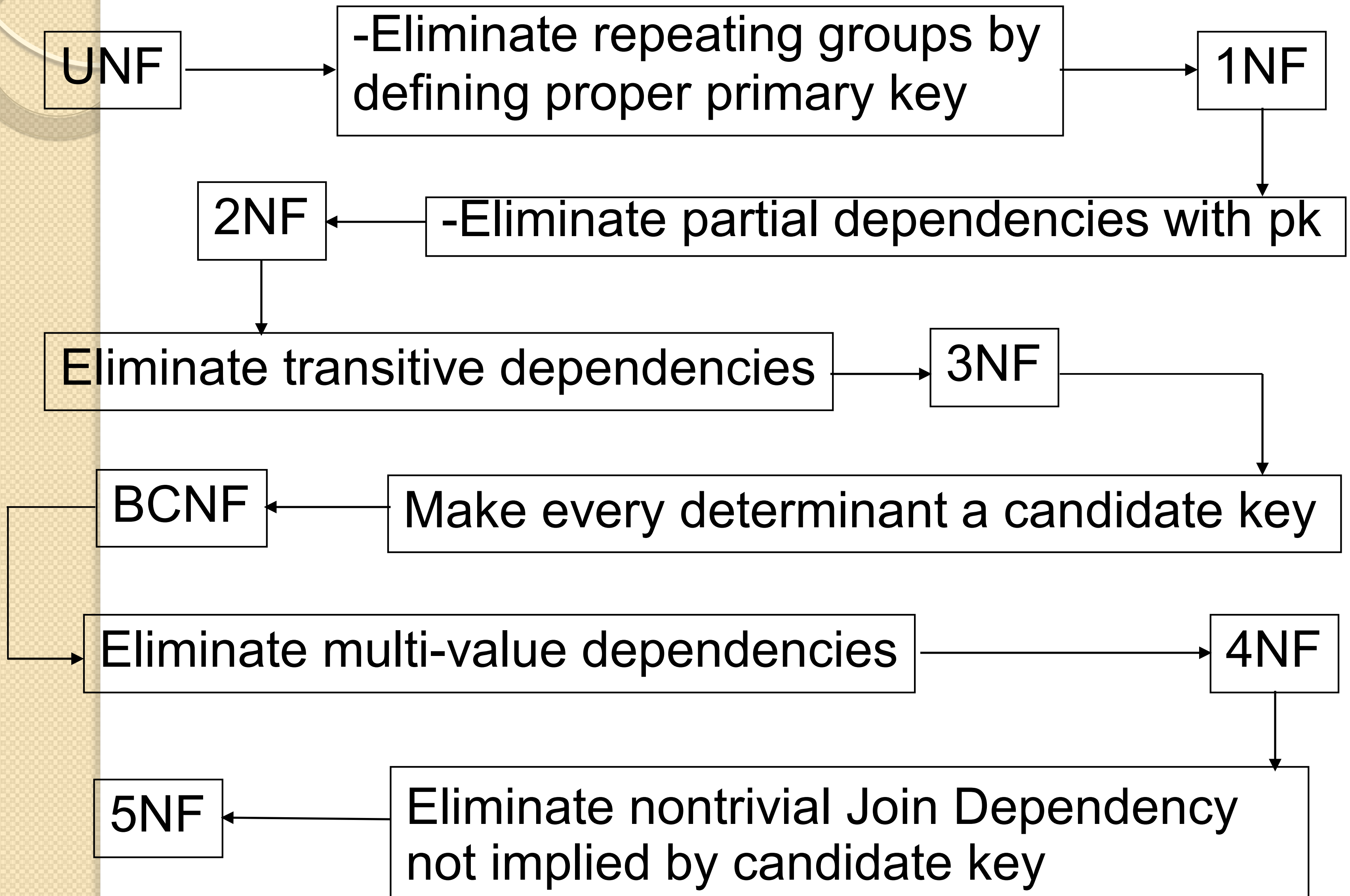
5NF (cont.)

- Example1
  - R(Supplier_no, Product_no, Project_no) has nontrivial JD, which is not implied by the candidate key.
  - To make R 5NF, split it into:
    - R1(Supplier_no,Product_no)
    - R2(Product_no,Project_no)
    - R3(Supplier_no, Project_no)

5NF (cont.)

- Example 2
  - SUPPLIER(<u>Supplier_no</u>, Supplier_name, Status, City)
    - Suppose there are two candidate keys, Supplier_no and Supplier_name.
  - SUPPLIER holds multiple JDs:
    - *((<u>Supplier_no</u>, Supplier_name, Status), (<u>Supplier_no</u>, City))
    - *((<u>Supplier_no</u>, Supplier_name), (<u>Supplier_no</u>, Status , City))
    - *((<u>Supplier_no</u>, Status), (<u>Supplier_no</u>, Supplier_name, City))
    - *((<u>Supplier_no</u>, Supplier_name), (<u>Supplier_no</u>, Status), (<u>Supplier_name</u>, City))
    - …
  - All of the JDs are nontrivial but implied by candidate keys, thus already in 5NF.

A Big Picture of Normalization



Quick NF Evaluation

- A relation derived by ER transformation is 1NF.
- A relation is 2NF if its pk is not composite.
- A relation is 3NF if it has fewer than two nonkey attributes.
- A relation is BCNF if no nonkey is determinant.
- A relation is 4NF if it has fewer than three fields.
- A relation is 5NF if it has fewer than three fields.

Case Study 1 (revisited)

- You are given the following relations, normalize them.
  - EMPLOYEE(Ssn, Fname, Minit, Lname, Bdate, Address, Sex, Salary, Super_ssn, Dnumber)
  - DEPARTMENT(Dnumber, Dname, Mgr_ssn, Mgr_start_date)

Case Study 1 (cont.)

- DEPT_LOCATIONS(<u>Dlocation</u>, Dnumber)

- PROJECT(<u>Pnumber</u>, Pname, Plocation, Dnumber)

Case Study 1 (cont.)

- WORKS_ON(<u>Ssn, Pnumber</u>, Hours)

- DEPENDENT(<u>Ssn, Dependent_name</u>, Sex, Bdate)

Exercises

1. อธิบายข้อดีและข้อเสียของการทำให้เป็นบรรทัดฐาน
2. อธิบายปัญหาและการแก้ไขซึ่งแสดงให้เห็นถึงความสำคัญของรูปแบบบรรทัดฐานที่หนึ่ง
3. อธิบายปัญหาและการแก้ไขซึ่งแสดงให้เห็นถึงความสำคัญของรูปแบบบรรทัดฐานที่สอง
4. อธิบายปัญหาและการแก้ไขซึ่งแสดงให้เห็นถึงความสำคัญของรูปแบบบรรทัดฐานที่สาม
5. อธิบายปัญหาและการแก้ไขซึ่งแสดงให้เห็นถึงความสำคัญของรูปแบบบรรทัดฐานที่บีซีเอ็นเอฟ
6. การพึ่งพิงที่ใช้ในการทำให้เป็นบรรทัดฐานมีทั้งหมดกี่ประเภท อธิบายแต่ละประเภทโดยสังเขป
7. หารูปแบบบรรทัดฐานของตารางความสัมพันธ์ กิจกรรมนักศึกษา(<u>รหัสนักศึกษา, ชมรม, งานอดิเรก</u>) อธิบายและระบุสมมติฐานที่ใช้ถ้ามี