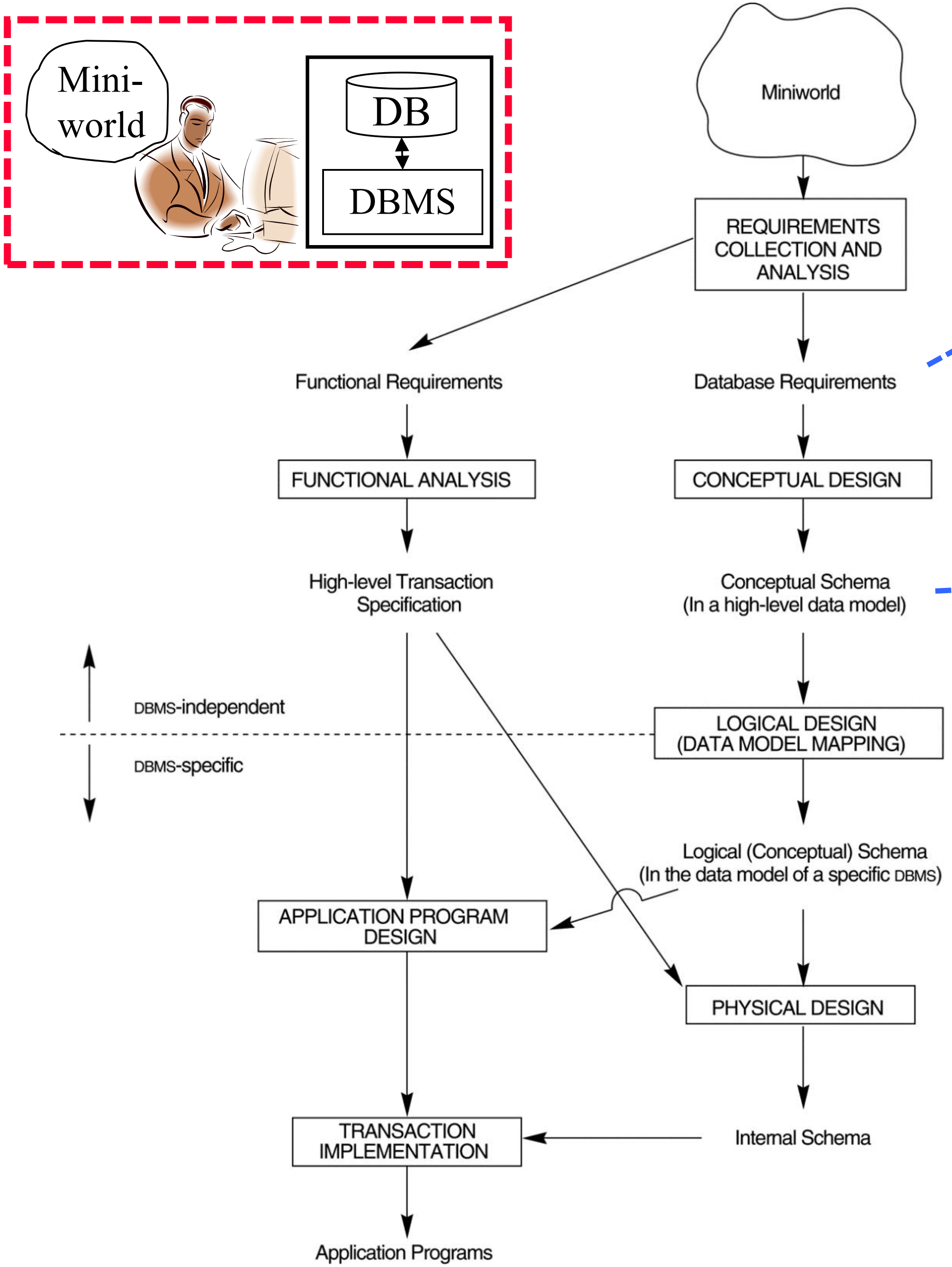
Chapter 14

Database Design Theory-Introduction to Normalization using Functional Dependencies and Multivalued Dependencies



Miniworld

REQUIREMENTS COLLECTION AND ANALYSIS

Functional Requirements

Database Requirements

FUNCTIONAL ANALYSIS

CONCEPTUAL DESIGN

High-level Transaction Specification

Conceptual Schema (In a high-level data model)

DBMS-independent

DBMS-specific

LOGICAL DESIGN (DATA MODEL MAPPING)

Logical (Conceptual) Schema (In the data model of a specific DBMS)

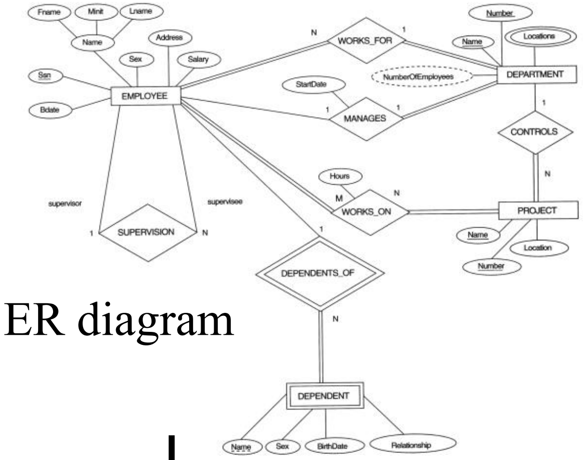
APPLICATION PROGRAM DESIGN

PHYSICAL DESIGN

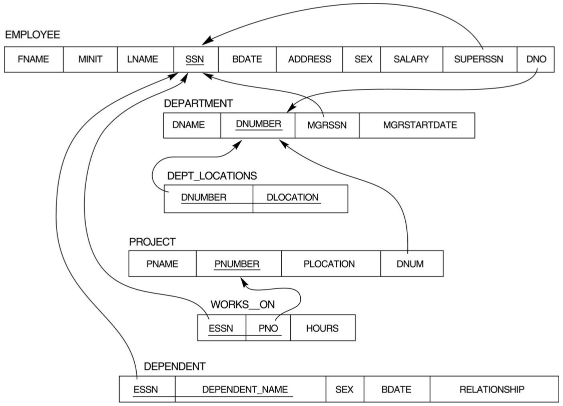
TRANSACTION IMPLEMENTATION

Internal Schema

Application Programs



ER diagram



EMPLOYEE

FNAME	MINIT	LNAME	SSN	BDATE	ADDRESS	SEX	SALARY	SUPERSSN	DNO

DEPARTMENT

DNAME	DNUMBER	MGRSSN	MGRSTARTDATE

DEPT_LOCATIONS

DNUMBER	DLOCATION

PROJECT

PNAME	PNUMBER	PLOCATION	DNUM

WORKS__ON

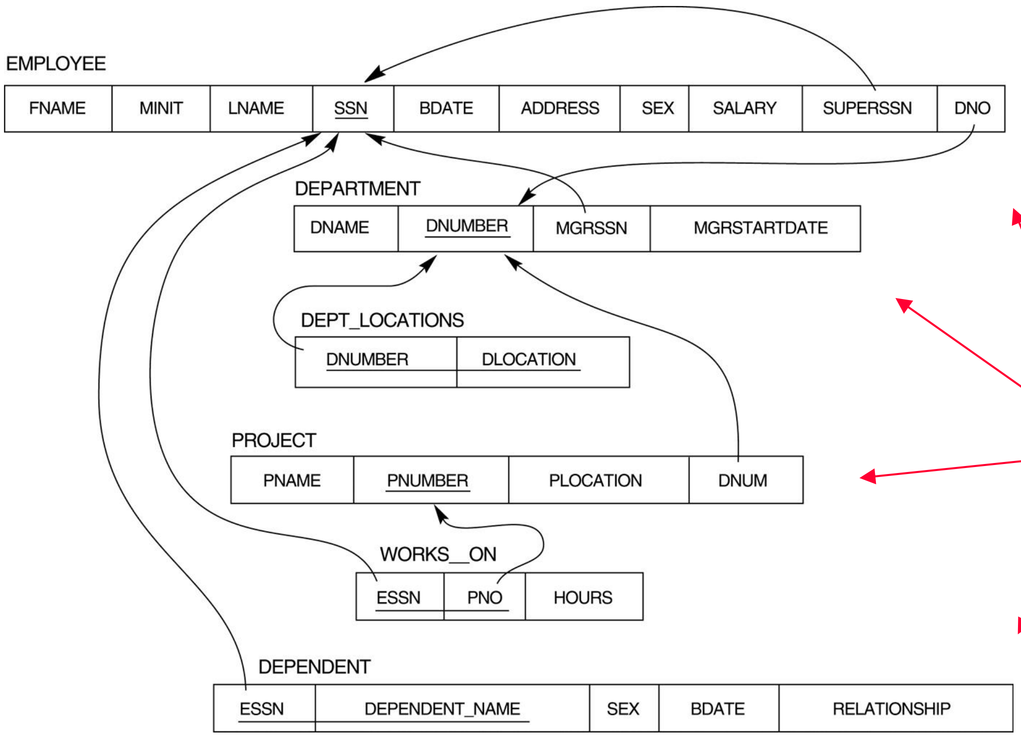
ESSN	PNO	HOURS

DEPENDENT

ESSN	DEPENDENT_NAME	SEX	BDATE	RELATIONSHIP

Quality? (ch. 14,15)

Relational Schema Quality



Schema Quality?

Which Normal Form?
1NF, 2NF, 3NF, BCNF?

Functional
Dependency

Chapter Outline

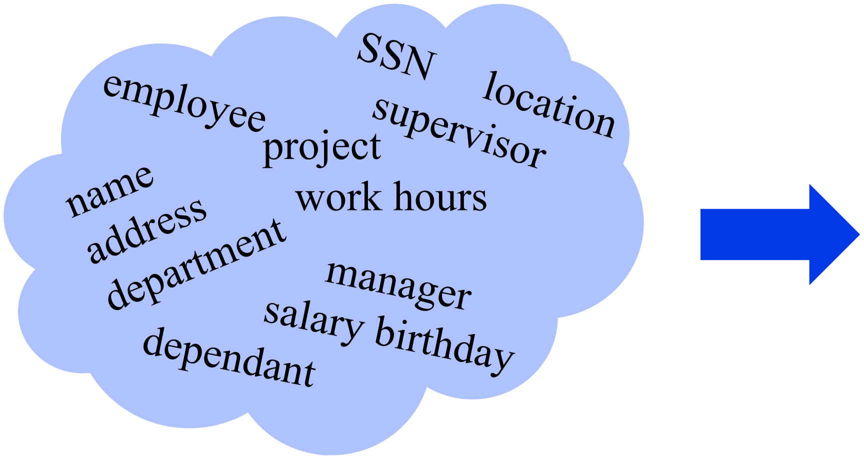
1. Informal Design Guidelines for Relational Databases
2. Functional Dependencies (FDs)
3. Normal Forms Based on Primary Keys
4. General Normal Form Definitions (For Multiple Keys)
5. BCNF (Boyce-Codd Normal Form)
6. Multivalued Dependency and Fourth Normal Form
7. Join Dependencies and Fifth Normal Form

Informal Design Guidelines

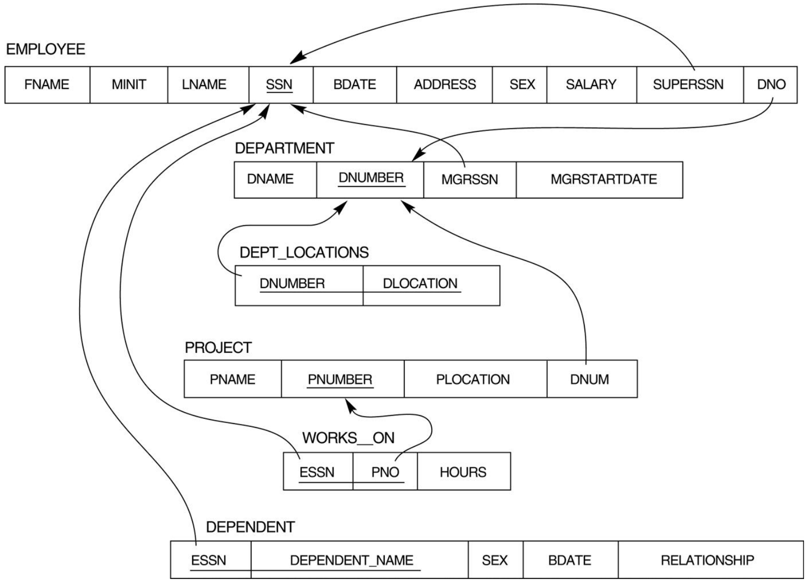
- What is relational database design?

  The grouping of attributes to form "good" relation schemas

- Two levels of relation schemas
  - The logical "user view" level
  - The storage "base relation" level
- Design is concerned mainly with base relations
- What are the criteria for "good" base relations?



Relevant Attributes of Mini-World



EMPLOYEE

FNAME	MINIT	LNAME	SSN	BDATE	ADDRESS	SEX	SALARY	SUPERSSN	DNO

DEPARTMENT

DNAME	DNUMBER	MGRSSN	MGRSTARTDATE

DEPT_LOCATIONS

DNUMBER	DLOCATION

PROJECT

PNAME	PNUMBER	PLOCATION	DNUM

WORKS__ON

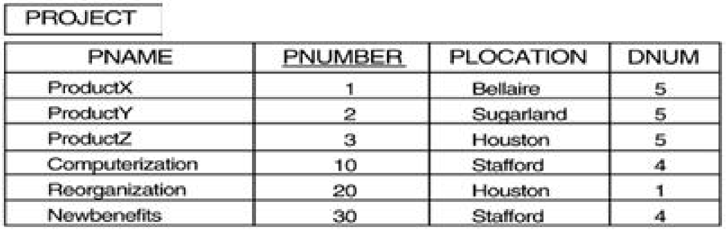
ESSN	PNO	HOURS

DEPENDENT

ESSN	DEPENDENT_NAME	SEX	BDATE	RELATIONSHIP

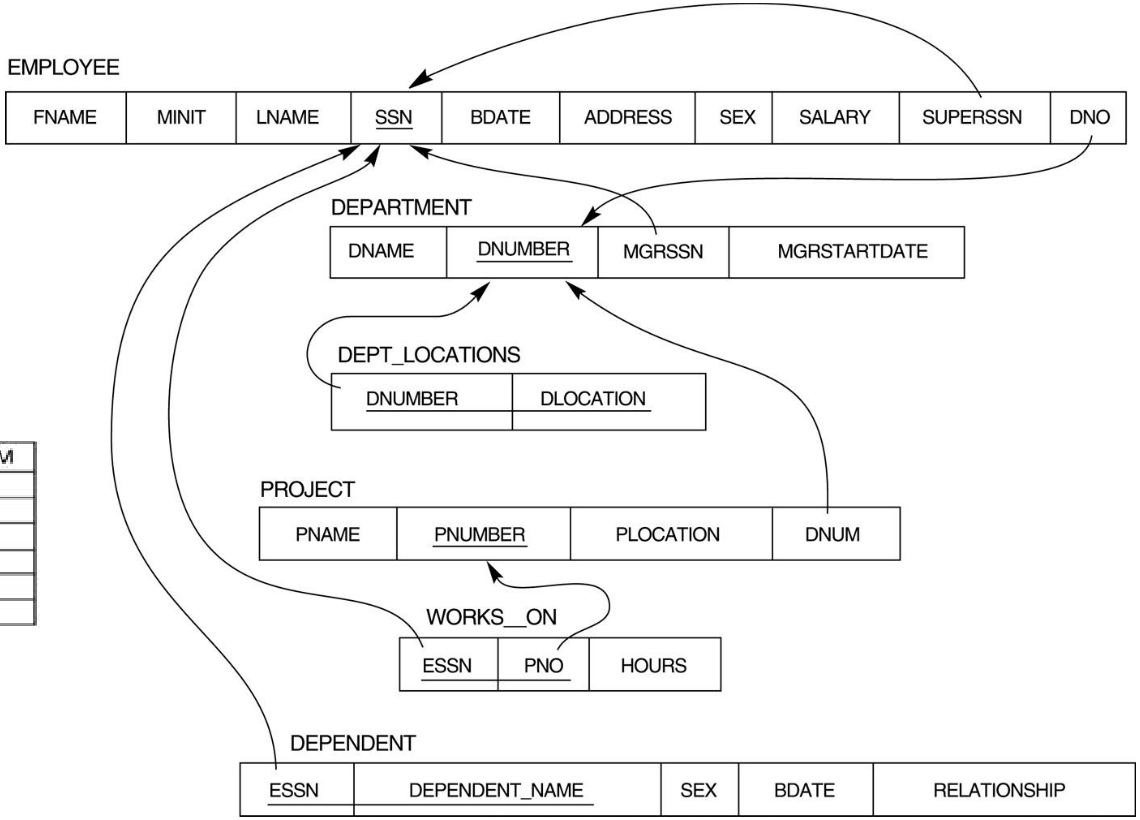
Informal Design Guidelines

1. Semantics of the Relation Attributes
2. Redundant Information in Tuples and Update Anomalies
3. Null Values in Tuples
4. Spurious Tuples



PROJECT

PNAME	PNUMBER	PLOCATION	DNUM
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5
Computerization	10	Stafford	4
Reorganization	20	Houston	1
Newbenefits	30	Stafford	4



EMPLOYEE

FNAME	MINIT	LNAME	SSN	BDATE	ADDRESS	SEX	SALARY	SUPERSSN	DNO

DEPARTMENT

DNAME	DNUMBER	MGRSSN	MGRSTARTDATE

DEPT_LOCATIONS

DNUMBER	DLOCATION

PROJECT

PNAME	PNUMBER	PLOCATION	DNUM

WORKS__ON

ESSN	PNO	HOURS

DEPENDENT

ESSN	DEPENDENT_NAME	SEX	BDATE	RELATIONSHIP

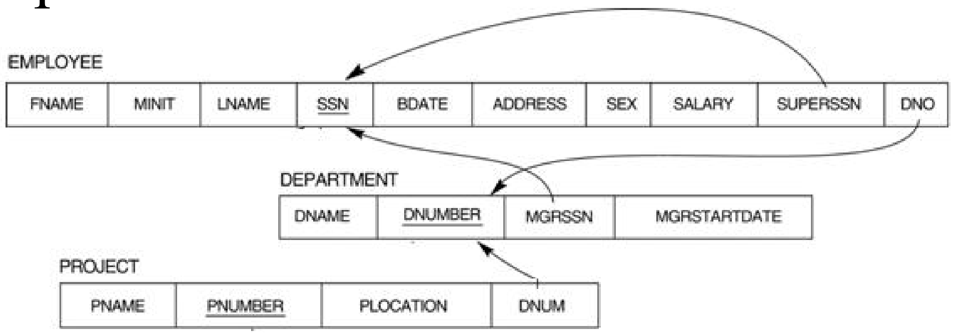
Semantics of the Relation Attributes

GUIDELINE 1

Informally, each tuple in a relation should **represent one entity or relationship instance**. (Applies to individual relations and their attributes).

- Attributes of different entities (EMPLOYEEs, DEPARTMENTs, PROJECTs) should not be mixed in the same relation
- Only foreign keys should be used to refer to other entities
- Entity and relationship attributes should be kept apart as much as possible.

Bottom Line: Design a schema that can be explained easily relation by relation. The semantics of attributes should be easy to interpret.



EMPLOYEE

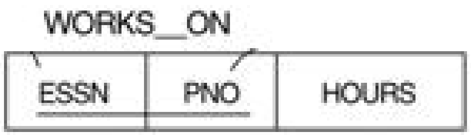
FNAME	MINIT	LNAME	SSN	BDATE	ADDRESS	SEX	SALARY	SUPERSSN	DNO

DEPARTMENT

DNAME	DNUMBER	MGRSSN	MGRSTARTDATE

PROJECT

PNAME	PNUMBER	PLOCATION	DNUM



WORKS_ON

ESSN	PNO	HOURS

A simplified COMPANY relational database schema

EMPLOYEE

				f.k.
ENAME	SSN	BDATE	ADDRESS	DNUMBER
	p.k.			

DEPARTMENT

		f.k.
DNAME	DNUMBER	DMGRSSN
	p.k.	

DEPT_LOCATIONS

f.k.	
DNUMBER	DLOCATION
p.k.	

PROJECT

			f.k.
PNAME	PNUMBER	PLOCATION	DNUM
	p.k.		

WORKS_ON

f.k.	f.k.	
SSN	PNUMBER	HOURS
p.k.		

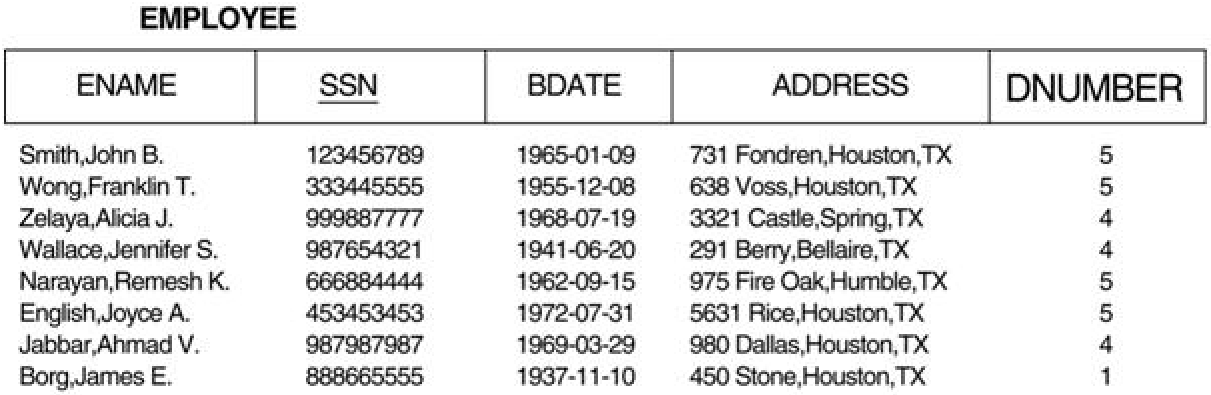
- Attributes of different entities (EMPLOYEEs, DEPARTMENTs, PROJECTs) should not be mixed in the same relation
- Only foreign keys should be used to refer to other entities.
- Entity and relationship attributes should be kept apart much as possible.

EMP_DEPT

ENAME	SSN	BDATE	ADDRESS	DNUMBER	DNAME	DMGRSSN

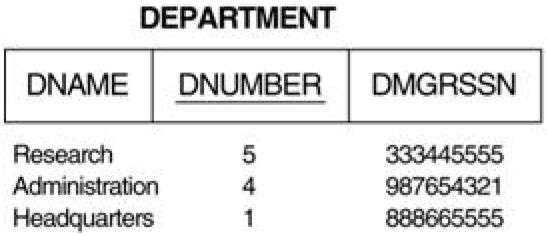
EMP_PROJ

SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION



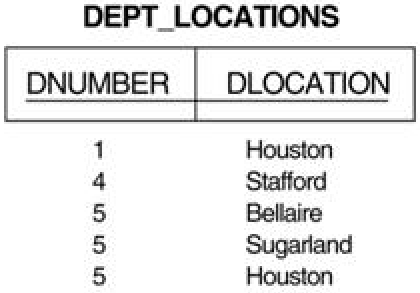
EMPLOYEE

ENAME	SSN	BDATE	ADDRESS	DNUMBER
Smith,John B.	123456789	1965-01-09	731 Fondren,Houston,TX	5
Wong,Franklin T.	333445555	1955-12-08	638 Voss,Houston,TX	5
Zelaya,Alicia J.	999887777	1968-07-19	3321 Castle,Spring,TX	4
Wallace,Jennifer S.	987654321	1941-06-20	291 Berry,Bellaire,TX	4
Narayan,Remesh K.	666884444	1962-09-15	975 Fire Oak,Humble,TX	5
English,Joyce A.	453453453	1972-07-31	5631 Rice,Houston,TX	5
Jabbar,Ahmad V.	987987987	1969-03-29	980 Dallas,Houston,TX	4
Borg,James E.	888665555	1937-11-10	450 Stone,Houston,TX	1



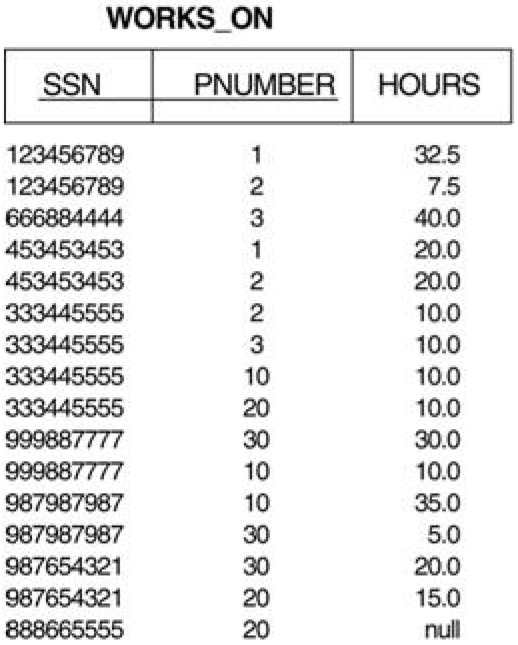
DEPARTMENT

DNAME	DNUMBER	DMGRSSN
Research	5	333445555
Administration	4	987654321
Headquarters	1	888665555



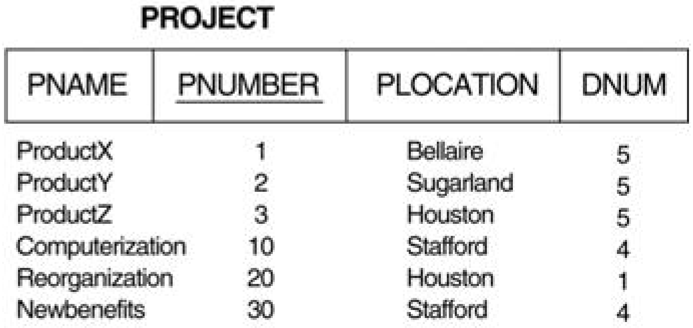
DEPT_LOCATIONS

DNUMBER	DLOCATION
1	Houston
4	Stafford
5	Bellaire
5	Sugarland
5	Houston



WORKS_ON

SSN	PNUMBER	HOURS
123456789	1	32.5
123456789	2	7.5
666884444	3	40.0
453453453	1	20.0
453453453	2	20.0
333445555	2	10.0
333445555	3	10.0
333445555	10	10.0
333445555	20	10.0
999887777	30	30.0
999887777	10	10.0
987987987	10	35.0
987987987	30	5.0
987654321	30	20.0
987654321	20	15.0
888665555	20	null



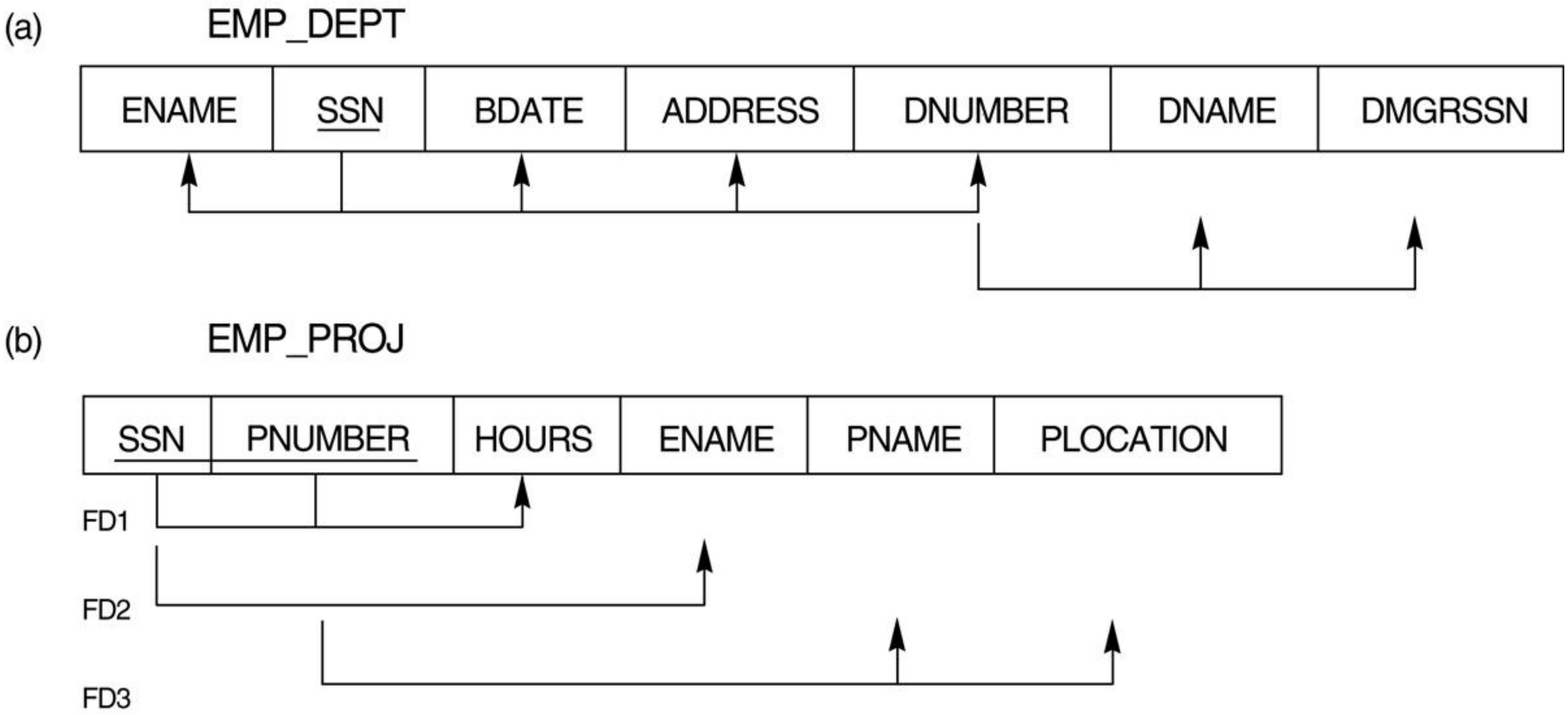
PROJECT

PNAME	PNUMBER	PLOCATION	DNUM
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5
Computerization	10	Stafford	4
Reorganization	20	Houston	1
Newbenefits	30	Stafford	4

Each tuple in a relation should represent **one entity** or **relationship instance**.

FIGURE 14.3

Two relation schemas suffering from update anomalies.



(a) EMP_DEPT

ENAME	SSN	BDATE	ADDRESS	DNUMBER	DNAME	DMGRSSN

(b) EMP_PROJ

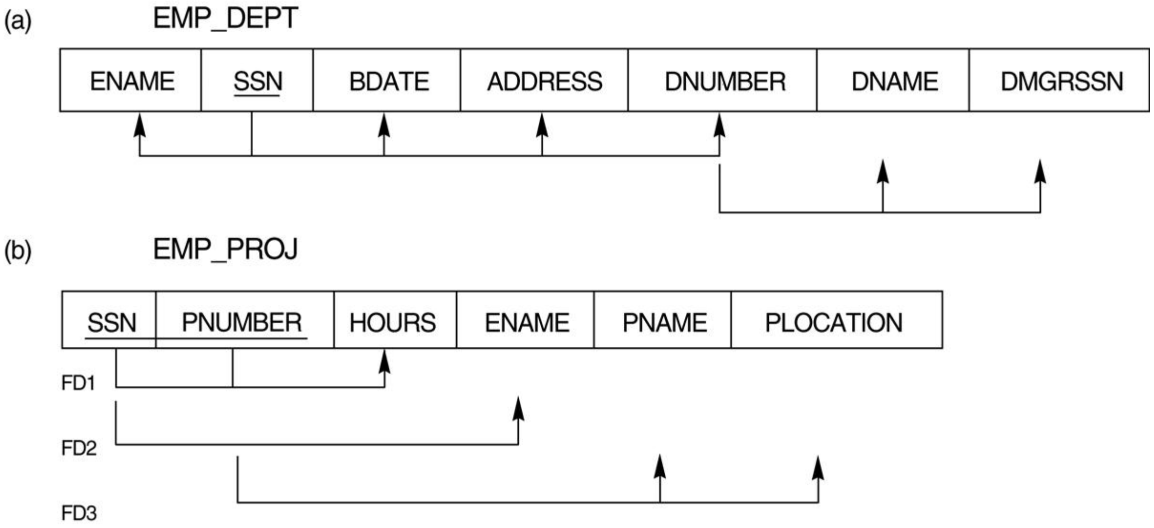
SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION

Bad design:

Violate Guideline 1 by mixing attributes from distinct real-world entities.

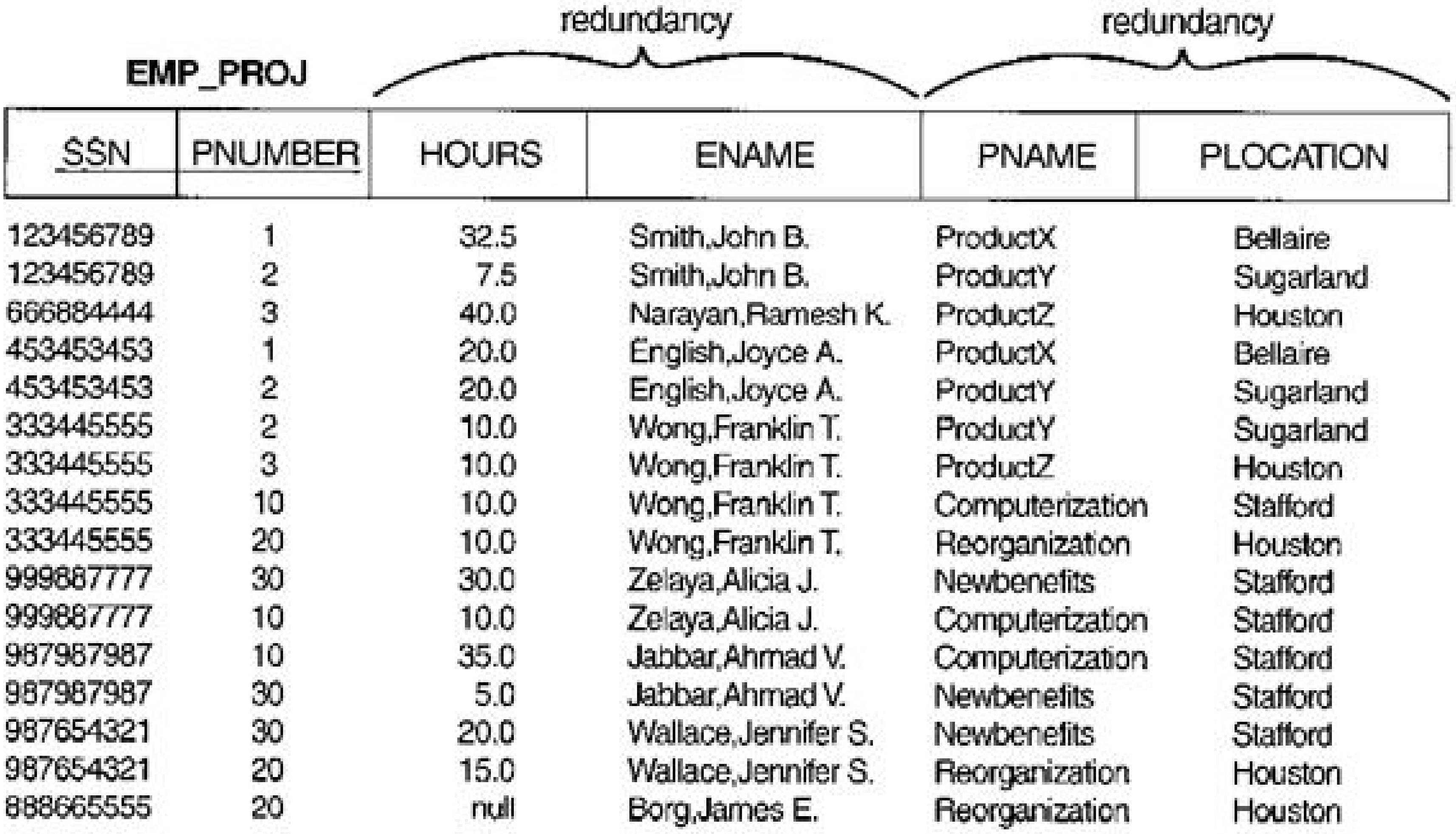
Redundant Information in Tuples and Update Anomalies

- Mixing attributes of multiple entities may cause problems
  - Information is stored redundantly wasting storage
  - Problems with update anomalies
    - Insertion anomalies
    - Deletion anomalies
    - Modification anomalies



Modification Anomaly:

Changing the name of project number P1 from “ProductX” to “Customer-Accounting” may cause this update to be made for all 100 employees working on project P1.

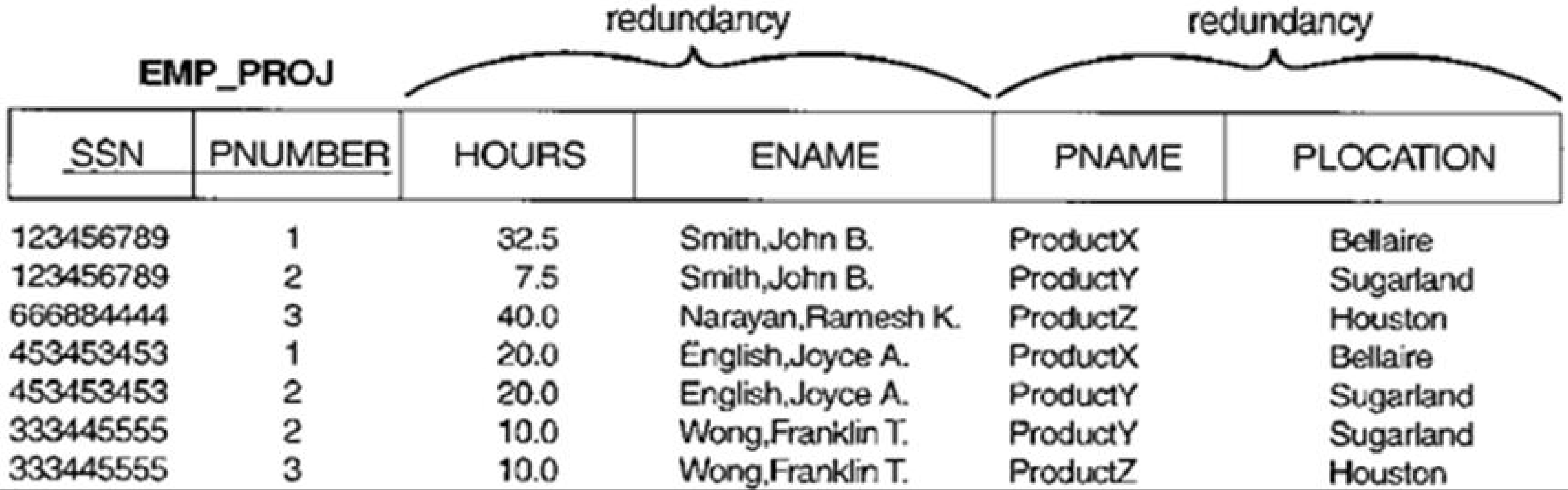


redundancy
redundancy

EMP_PROJ

SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION
123456789	1	32.5	Smith,John B.	ProductX	Bellaire
123456789	2	7.5	Smith,John B.	ProductY	Sugarland
666884444	3	40.0	Narayan,Ramesh K.	ProductZ	Houston
453453453	1	20.0	English,Joyce A.	ProductX	Bellaire
453453453	2	20.0	English,Joyce A.	ProductY	Sugarland
333445555	2	10.0	Wong,Franklin T.	ProductY	Sugarland
333445555	3	10.0	Wong,Franklin T.	ProductZ	Houston
333445555	10	10.0	Wong,Franklin T.	Computerization	Stafford
333445555	20	10.0	Wong,Franklin T.	Reorganization	Houston
999887777	30	30.0	Zelaya,Alicia J.	Newbenefits	Stafford
999887777	10	10.0	Zelaya,Alicia J.	Computerization	Stafford
987987987	10	35.0	Jabbar,Ahmad V.	Computerization	Stafford
987987987	30	5.0	Jabbar,Ahmad V.	Newbenefits	Stafford
987654321	30	20.0	Wallace,Jennifer S.	Newbenefits	Stafford
987654321	20	15.0	Wallace,Jennifer S.	Reorganization	Houston
888665555	20	null	Borg,James E.	Reorganization	Houston

- **Insert Anomaly:**
  - Cannot insert a project unless an employee is assigned to.
  - ***Inversely*** cannot insert an employee unless an he/she is assigned to a project.
- **Delete Anomaly:**
  - When a project is deleted, it will result in deleting all the employees who work on that project.
  - Alternately, if an employee is the sole employee on a project, deleting that employee would result in deleting the corresponding project.


redundancy
redundancy

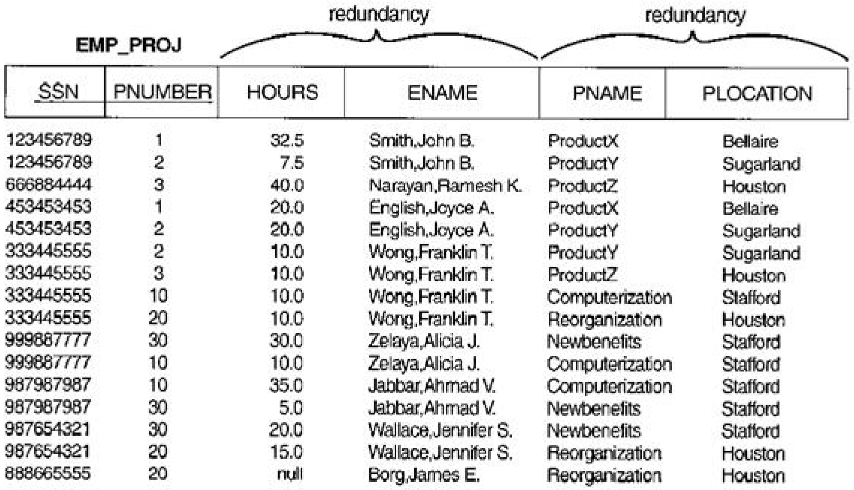
EMP_PROJ

SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION
123456789	1	32.5	Smith,John B.	ProductX	Bellaire
123456789	2	7.5	Smith,John B.	ProductY	Sugarland
666884444	3	40.0	Narayan,Ramesh K.	ProductZ	Houston
453453453	1	20.0	English,Joyce A.	ProductX	Bellaire
453453453	2	20.0	English,Joyce A.	ProductY	Sugarland
333445555	2	10.0	Wong,Franklin T.	ProductY	Sugarland
333445555	3	10.0	Wong,Franklin T.	ProductZ	Houston

Guideline to Redundant Information

- **GUIDELINE 2:**

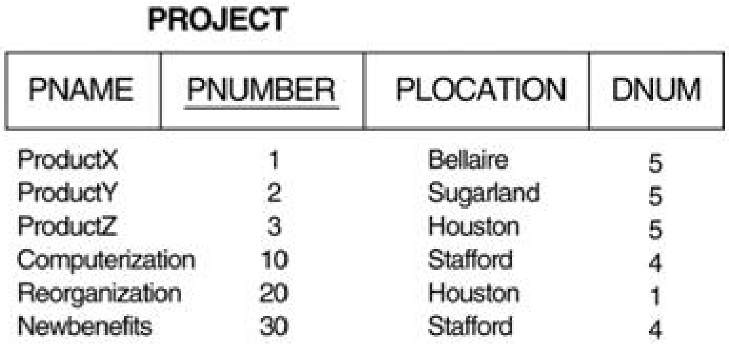
  Design a schema that does not suffer from the insertion, deletion and update anomalies. If there are any present, then note them so that applications can be made to take them into account



redundancy
redundancy

EMP_PROJ

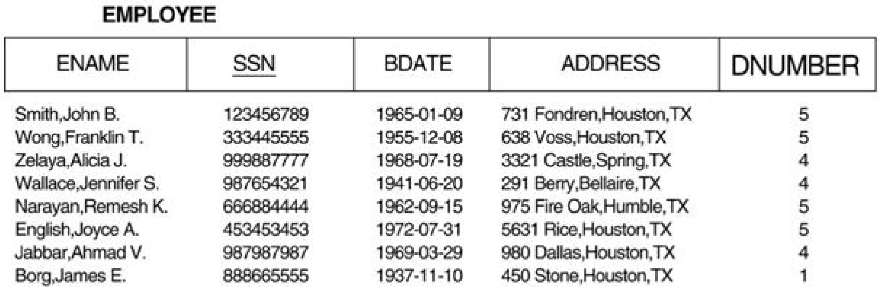
SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION
123456789	1	32.5	Smith,John B.	ProductX	Bellaire
123456789	2	7.5	Smith,John B.	ProductY	Sugarland
666884444	3	40.0	Narayan,Ramesh K.	ProductZ	Houston
453453453	1	20.0	English,Joyce A.	ProductX	Bellaire
453453453	2	20.0	English,Joyce A.	ProductY	Sugarland
333445555	2	10.0	Wong,Franklin T.	ProductY	Sugarland
333445555	3	10.0	Wong,Franklin T.	ProductZ	Houston
333445555	10	10.0	Wong,Franklin T.	Computerization	Stafford
333445555	20	10.0	Wong,Franklin T.	Reorganization	Houston
999887777	30	30.0	Zelaya,Alicia J.	Newbenefits	Stafford
999887777	10	10.0	Zelaya,Alicia J.	Computerization	Stafford
987987987	10	35.0	Jabbar,Ahmad V.	Computerization	Stafford
987987987	30	5.0	Jabbar,Ahmad V.	Newbenefits	Stafford
987654321	30	20.0	Wallace,Jennifer S.	Newbenefits	Stafford
987654321	20	15.0	Wallace,Jennifer S.	Reorganization	Houston
888665555	20	null	Borg,James E.	Reorganization	Houston



PROJECT

PNAME	PNUMBER	PLOCATION	DNUM
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5
Computerization	10	Stafford	4
Reorganization	20	Houston	1
Newbenefits	30	Stafford	4

- Change P1’s name
- Insert a new project without assigning any worker yet
- Delete a project or an employee

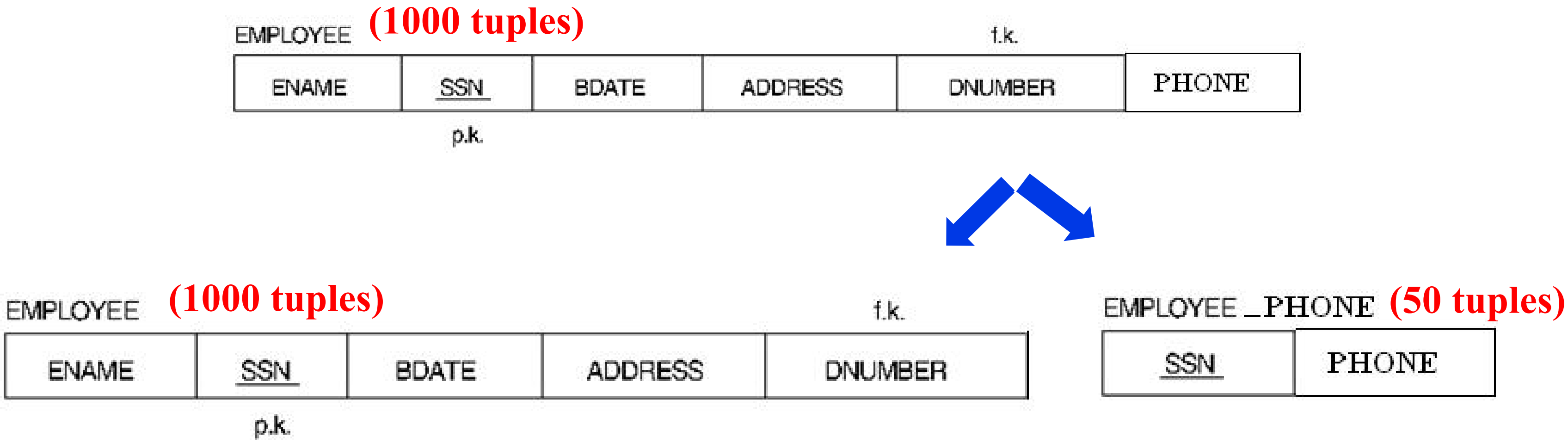


EMPLOYEE

ENAME	SSN	BDATE	ADDRESS	DNUMBER
Smith,John B.	123456789	1965-01-09	731 Fondren,Houston,TX	5
Wong,Franklin T.	333445555	1955-12-08	638 Voss,Houston,TX	5
Zelaya,Alicia J.	999887777	1968-07-19	3321 Castle,Spring,TX	4
Wallace,Jennifer S.	987654321	1941-06-20	291 Berry,Bellaire,TX	4
Narayan,Remesh K.	666884444	1962-09-15	975 Fire Oak,Humble,TX	5
English,Joyce A.	453453453	1972-07-31	5631 Rice,Houston,TX	5
Jabbar,Ahmad V.	987987987	1969-03-29	980 Dallas,Houston,TX	4
Borg,James E.	888665555	1937-11-10	450 Stone,Houston,TX	1

Guideline to Redundant Information

- **GUIDELINE 3:**
  - Relations should be designed such that their tuples will have as few NULL values as possible
  - Attributes that are NULL frequently could be placed in separate relations (with the primary key)
  - Reasons for nulls:
    - attribute not applicable or invalid (e.g. office phone no. of a student)
    - attribute value unknown (may exist) (e.g. name of spouse)
    - value known to exist, but unavailable (e.g. weight of a female)



EMPLOYEE **(1000 tuples)**

ENAME	<u>SSN</u>	BDATE	ADDRESS	DNUMBER	PHONE

EMPLOYEE **(1000 tuples)**

ENAME	<u>SSN</u>	BDATE	ADDRESS	DNUMBER

EMPLOYEE_PHONE **(50 tuples)**

<u>SSN</u>	PHONE

Spurious Tuples

- Bad designs for a relational database may result in erroneous results for certain JOIN operations
- The "lossless join" property is used to guarantee meaningful results for join operations

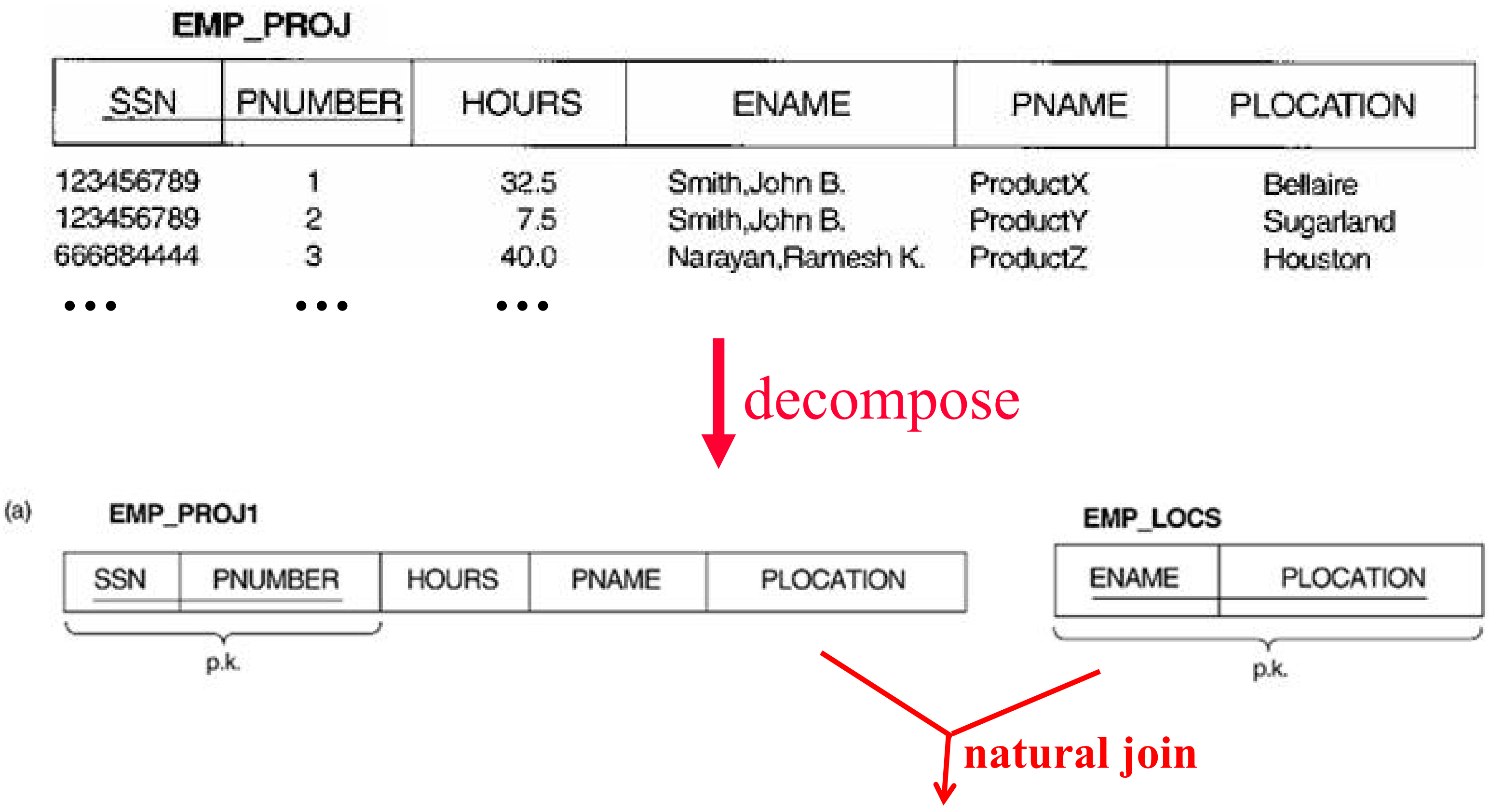
GUIDELINE 4

The relations should be designed to satisfy ***the lossless join condition***:

No spurious tuples should be generated by doing a natural-join of any relations.

FIGURE 14.5 Particularly poor design for the EMP_PROJ relation of Figure 14.3b.

(a) The two relation schemas EMP_LOCS and EMP_PROJ1.
(b) The result of projecting the extension of EMP_PROJ from Figure 14.4 onto the relations EMP_LOCS and EMP_PROJ1.



EMP_PROJ

SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION
123456789	1	32.5	Smith,John B.	ProductX	Bellaire
123456789	2	7.5	Smith,John B.	ProductY	Sugarland
666884444	3	40.0	Narayan,Ramesh K.	ProductZ	Houston
• • •	• • •	• • •			

decompose

(a) EMP_PROJ1

SSN	PNUMBER	HOURS	PNAME	PLOCATION

p.k.

EMP_LOCS

ENAME	PLOCATION

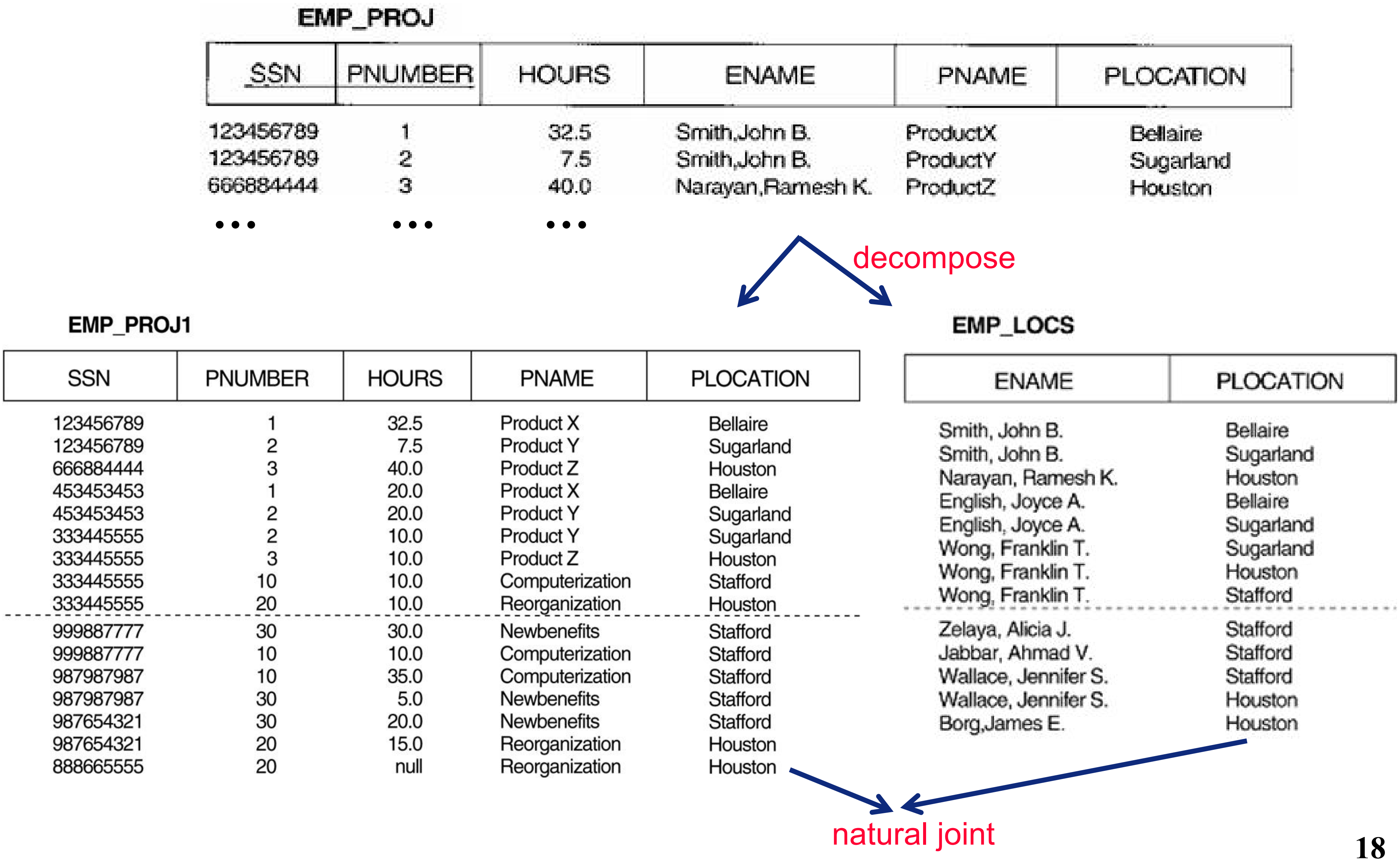
p.k.
natural join

EMP_PROJ

Shall not generate spurious tuples.

FIGURE 14.5 (continued)

The result of projecting the extension of EMP_PROJ from Figure 14.4 onto the relations EMP_LOCS and EMP_PROJ1.



EMP_PROJ

SSN	PNUMBER	HOURS	ENAME	PNAME	PLOCATION
123456789	1	32.5	Smith,John B.	ProductX	Bellaire
123456789	2	7.5	Smith,John B.	ProductY	Sugarland
666884444	3	40.0	Narayan,Ramesh K.	ProductZ	Houston
• • •	• • •	• • •			

EMP_PROJ1

SSN	PNUMBER	HOURS	PNAME	PLOCATION
123456789	1	32.5	Product X	Bellaire
123456789	2	7.5	Product Y	Sugarland
666884444	3	40.0	Product Z	Houston
453453453	1	20.0	Product X	Bellaire
453453453	2	20.0	Product Y	Sugarland
333445555	2	10.0	Product Y	Sugarland
333445555	3	10.0	Product Z	Houston
333445555	10	10.0	Computerization	Stafford
333445555	20	10.0	Reorganization	Houston
999887777	30	30.0	Newbenefits	Stafford
999887777	10	10.0	Computerization	Stafford
987987987	10	35.0	Computerization	Stafford
987987987	30	5.0	Newbenefits	Stafford
987654321	30	20.0	Newbenefits	Stafford
987654321	20	15.0	Reorganization	Houston
888665555	20	null	Reorganization	Houston

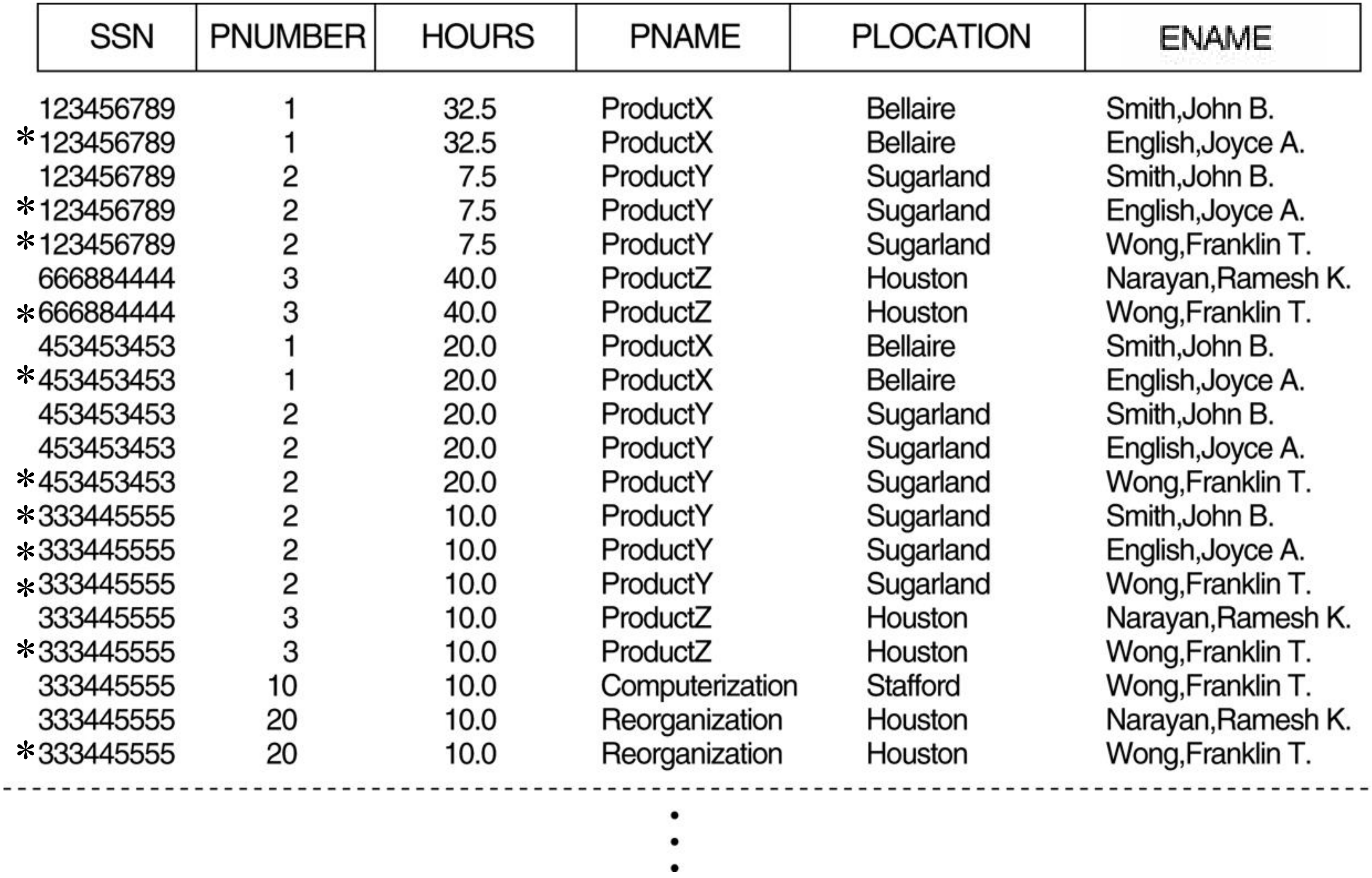
EMP_LOCS

ENAME	PLOCATION
Smith, John B.	Bellaire
Smith, John B.	Sugarland
Narayan, Ramesh K.	Houston
English, Joyce A.	Bellaire
English, Joyce A.	Sugarland
Wong, Franklin T.	Sugarland
Wong, Franklin T.	Houston
Wong, Franklin T.	Stafford
Zelaya, Alicia J.	Stafford
Jabbar, Ahmad V.	Stafford
Wallace, Jennifer S.	Stafford
Wallace, Jennifer S.	Houston
Borg,James E.	Houston

FIGURE 14.6

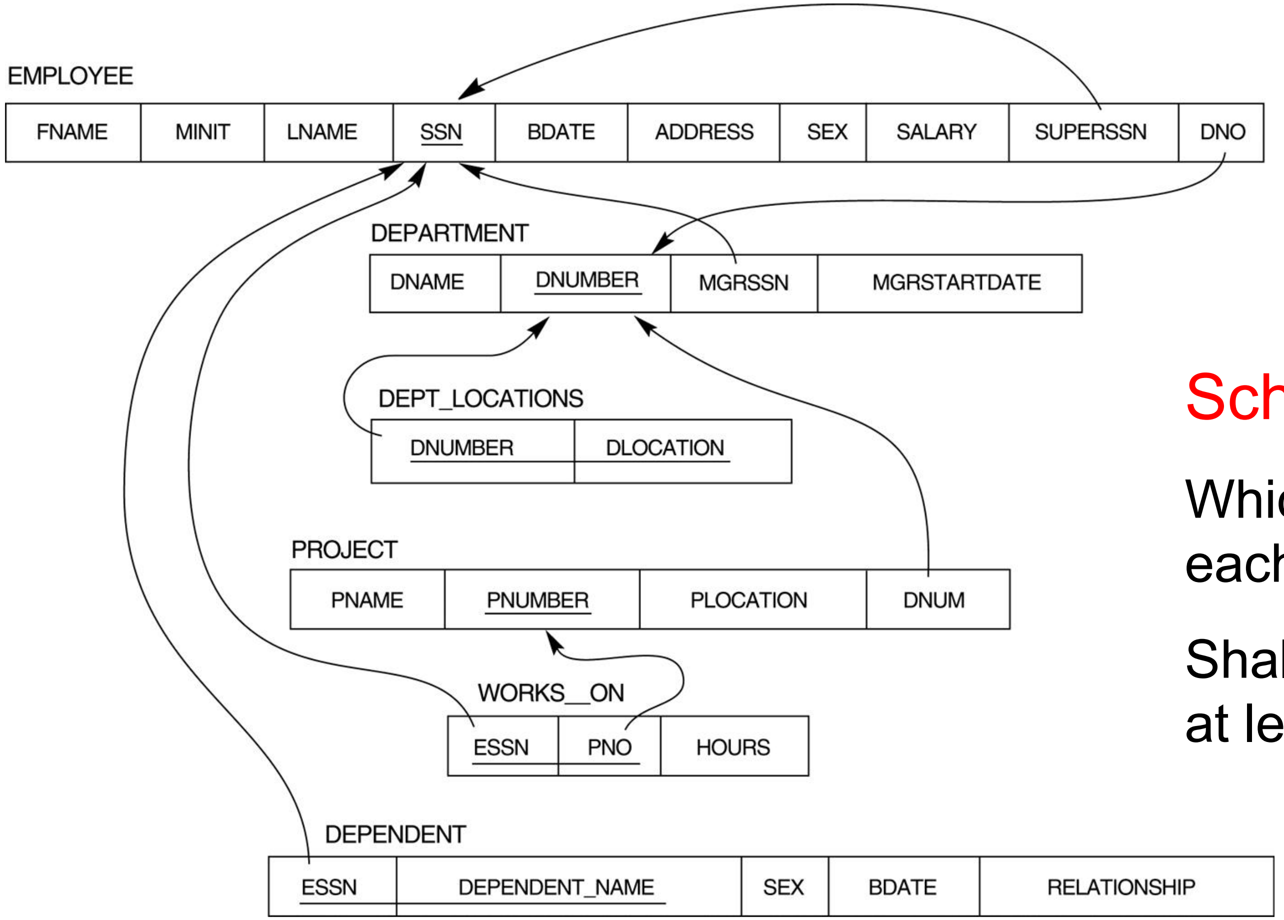
Result of applying NATURAL JOIN to the tuples above the dotted lines in EMP_PROJ1 and EMP_LOCS of Figure 14.5.

Generated spurious tuples are marked by asterisks (*).



SSN	PNUMBER	HOURS	PNAME	PLOCATION	ENAME
123456789	1	32.5	ProductX	Bellaire	Smith,John B.
*123456789	1	32.5	ProductX	Bellaire	English,Joyce A.
123456789	2	7.5	ProductY	Sugarland	Smith,John B.
*123456789	2	7.5	ProductY	Sugarland	English,Joyce A.
*123456789	2	7.5	ProductY	Sugarland	Wong,Franklin T.
666884444	3	40.0	ProductZ	Houston	Narayan,Ramesh K.
*666884444	3	40.0	ProductZ	Houston	Wong,Franklin T.
453453453	1	20.0	ProductX	Bellaire	Smith,John B.
*453453453	1	20.0	ProductX	Bellaire	English,Joyce A.
453453453	2	20.0	ProductY	Sugarland	Smith,John B.
453453453	2	20.0	ProductY	Sugarland	English,Joyce A.
*453453453	2	20.0	ProductY	Sugarland	Wong,Franklin T.
*333445555	2	10.0	ProductY	Sugarland	Smith,John B.
*333445555	2	10.0	ProductY	Sugarland	English,Joyce A.
*333445555	2	10.0	ProductY	Sugarland	Wong,Franklin T.
333445555	3	10.0	ProductZ	Houston	Narayan,Ramesh K.
*333445555	3	10.0	ProductZ	Houston	Wong,Franklin T.
333445555	10	10.0	Computerization	Stafford	Wong,Franklin T.
333445555	20	10.0	Reorganization	Houston	Narayan,Ramesh K.
*333445555	20	10.0	Reorganization	Houston	Wong,Franklin T.

Quality of Database Schema



EMPLOYEE

FNAME	MINIT	LNAME	SSN	BDATE	ADDRESS	SEX	SALARY	SUPERSSN	DNO

DEPARTMENT

DNAME	DNUMBER	MGRSSN	MGRSTARTDATE

DEPT_LOCATIONS

DNUMBER	DLOCATION

PROJECT

PNAME	PNUMBER	PLOCATION	DNUM

WORKS__ON

ESSN	PNO	HOURS

DEPENDENT

ESSN	DEPENDENT_NAME	SEX	BDATE	RELATIONSHIP

Schema Quality?

Which **Normal Form** is each relation in?

Shall be in 3NF or BCNF at least .

Functional Dependencies and Normal Forms

2. Functional Dependencies (FDs)

3. Normal Forms Based on Primary Keys

   3.1 Normalization of Relations

   3.2 Practical Use of Normal Forms

   3.3 Definitions of Keys and Attributes Participating in Keys

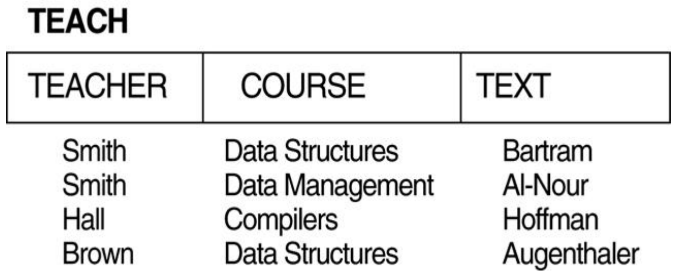
   3.4 First Normal Form

   3.5 Second Normal Form

   3.6 Third Normal Form

4. General Normal Form Definitions (For Multiple Keys)

5. BCNF (Boyce-Codd Normal Form)



TEACH

TEACHER	COURSE	TEXT
Smith	Data Structures	Bartram
Smith	Data Management	Al-Nour
Hall	Compilers	Hoffman
Brown	Data Structures	Augenthaler

Which NF?

Functional Dependency

Examples of Functional Dependency

- Social security number determines employee name

  SSN → ENAME

  ENAME → SSN (?)

- Project number determines project name and location

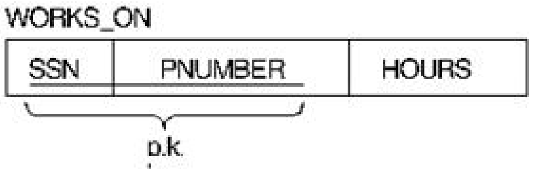
  PNUMBER → {PNAME, PLOCATION}

  PNUMBER → PNAME (?)

- Employee ssn and project number determines the hours per week that the employee works on the project

  {SSN, PNUMBER} → HOURS

  SSN → HOURS (?)



WORKS_ON

SSN	PNUMBER	HOURS

- An FD is a property of the attributes in the schema R
- The constraint must hold on ***every relation instance*** r(R)

Functional Dependencies 功能相依

- X → Y holds
  - If whenever two tuples have the same value for X, they ***must have*** the same value for Y

    e.g. {StudentID} → {Name}

    {SSN} → {Address}

    {Name, Birthday} → {Address, Dept., Sex}
  - For any two tuples t1 and t2 in any relation instance r(R)

    If $t1[X] = t2[X]$, **then** $t1[Y] = t2[Y]$
- X → Y in R specifies a *constraint* on all relation instances r(R)
- FDs are derived from the real-world constraints on the attributes

Definition of Functional Dependency

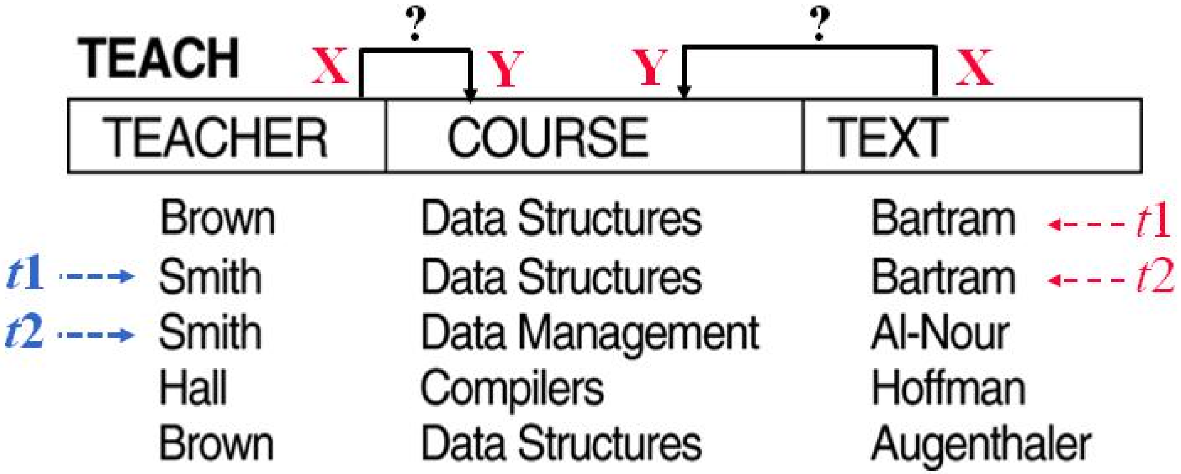
X → Y

- A set of attributes X ***functionally determines*** a set of attributes Y if the value of X determines a unique value for Y.
- **If** $t1[X] = t2[X]$, **then** $t1[Y] = t2[Y]$

A relation state of TEACH with

A *possible* functional dependency **TEXT → COURSE**.

However, **TEACHER → COURSE** is ruled out. (Smith違反)

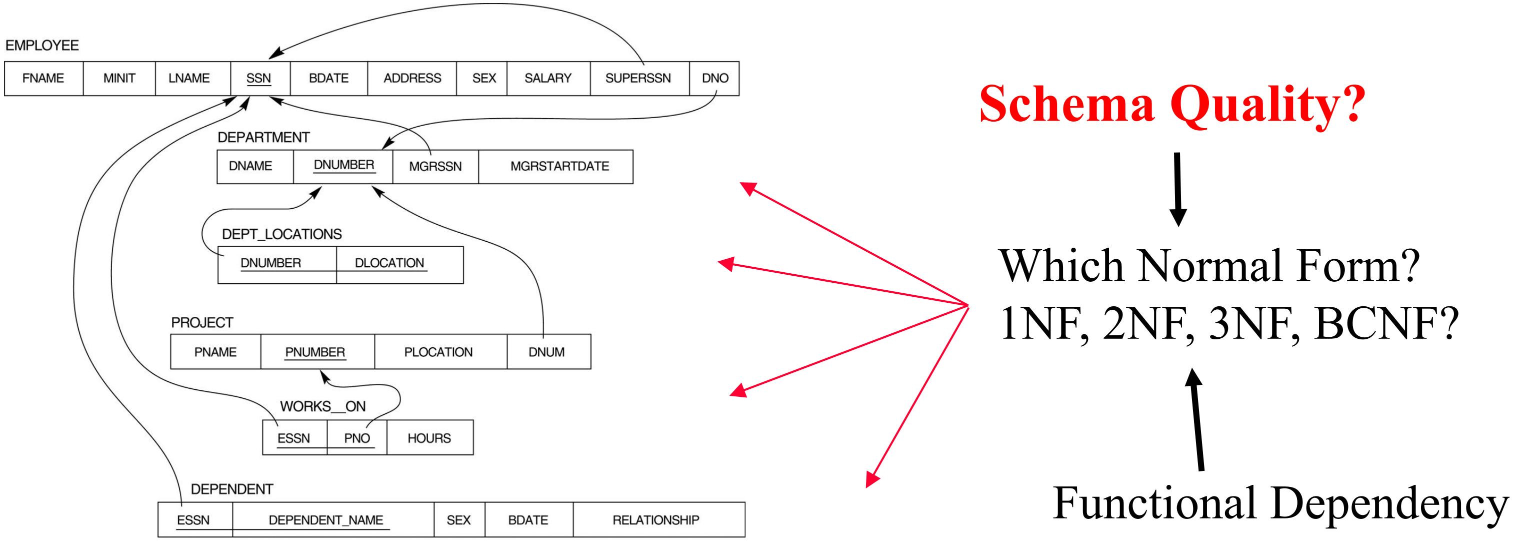


TEACH
X
?
Y
Y
?
X
t1
t2
t1
t2

TEACHER	COURSE	TEXT
Brown	Data Structures	Bartram
Smith	Data Structures	Bartram
Smith	Data Management	Al-Nour
Hall	Compilers	Hoffman
Brown	Data Structures	Augenthaler

Functional Dependencies

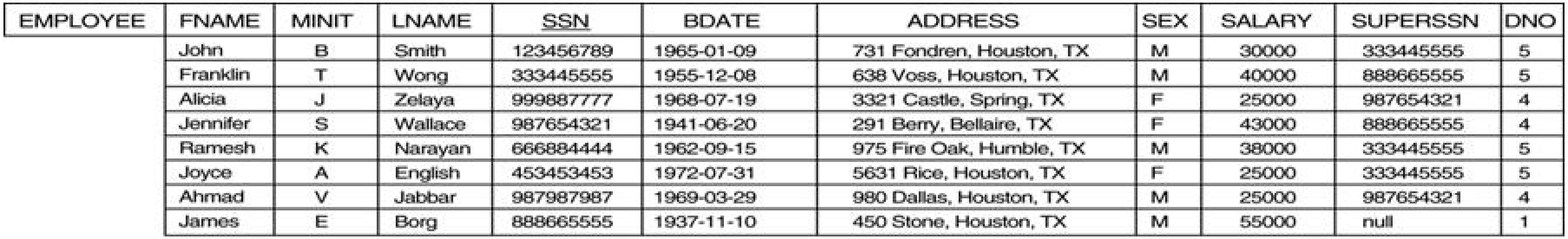
- Functional dependencies (FDs) are used to specify *formal measures* of the "goodness" of relational designs
- FDs and keys are used to define **normal forms** for relations
- FDs are **constraints** that are derived from the *meaning* and *interrelationships* of the data attributes



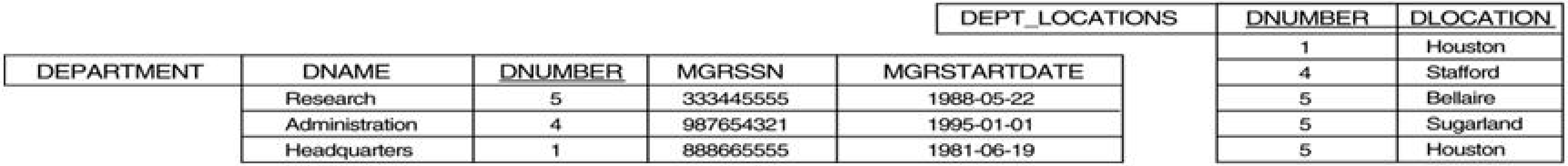
Schema Quality?

Which Normal Form?
1NF, 2NF, 3NF, BCNF?

Functional Dependency

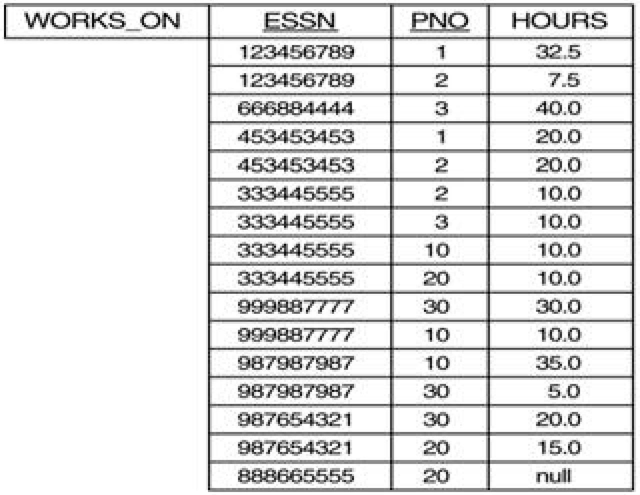


EMPLOYEE	FNAME	MINIT	LNAME	SSN	BDATE	ADDRESS	SEX	SALARY	SUPERSSN	DNO
	John	B	Smith	123456789	1965-01-09	731 Fondren, Houston, TX	M	30000	333445555	5
	Franklin	T	Wong	333445555	1955-12-08	638 Voss, Houston, TX	M	40000	888665555	5
	Alicia	J	Zelaya	999887777	1968-07-19	3321 Castle, Spring, TX	F	25000	987654321	4
	Jennifer	S	Wallace	987654321	1941-06-20	291 Berry, Bellaire, TX	F	43000	888665555	4
	Ramesh	K	Narayan	666884444	1962-09-15	975 Fire Oak, Humble, TX	M	38000	333445555	5
	Joyce	A	English	453453453	1972-07-31	5631 Rice, Houston, TX	F	25000	333445555	5
	Ahmad	V	Jabbar	987987987	1969-03-29	980 Dallas, Houston, TX	M	25000	987654321	4
	James	E	Borg	888665555	1937-11-10	450 Stone, Houston, TX	M	55000	null	1

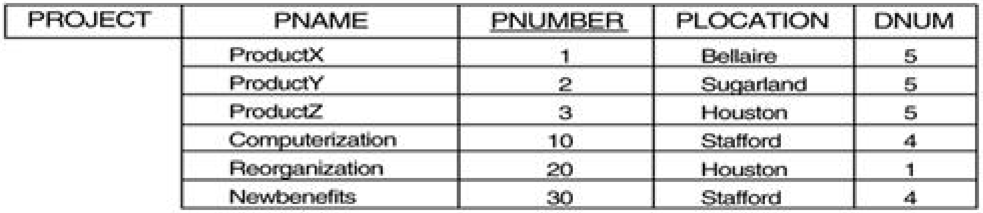


DEPT_LOCATIONS	DNUMBER	DLOCATION
	1	Houston
	4	Stafford
	5	Bellaire
	5	Sugarland
	5	Houston

DEPARTMENT	DNAME	DNUMBER	MGRSSN	MGRSTARTDATE
	Research	5	333445555	1988-05-22
	Administration	4	987654321	1995-01-01
	Headquarters	1	888665555	1981-06-19



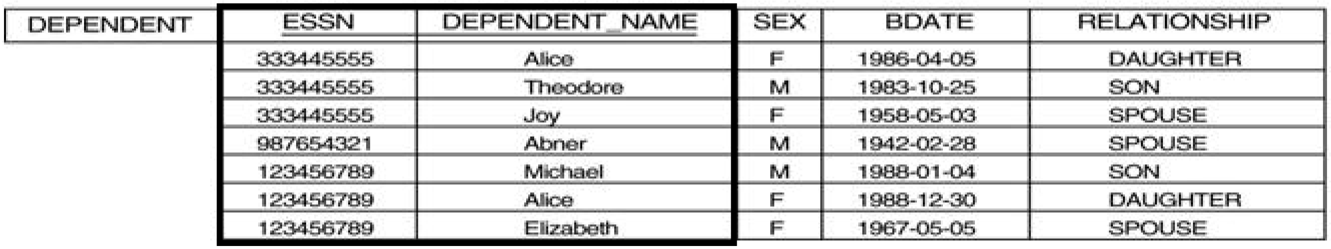
WORKS_ON	ESSN	PNO	HOURS
	123456789	1	32.5
	123456789	2	7.5
	666884444	3	40.0
	453453453	1	20.0
	453453453	2	20.0
	333445555	2	10.0
	333445555	3	10.0
	333445555	10	10.0
	333445555	20	10.0
	999887777	30	30.0
	999887777	10	10.0
	987987987	10	35.0
	987987987	30	5.0
	987654321	30	20.0
	987654321	20	15.0
	888665555	20	null



PROJECT	PNAME	PNUMBER	PLOCATION	DNUM
	ProductX	1	Bellaire	5
	ProductY	2	Sugarland	5
	ProductZ	3	Houston	5
	Computerization	10	Stafford	4
	Reorganization	20	Houston	1
	Newbenefits	30	Stafford	4

- If *K* is a key of R, then *K* functionally determines all attributes in R (since we never have two distinct tuples with t1[*K*] = t2[*K*])
- The FD constraint must hold on *every relation instance* r(R)

{ESSN, DEPENDENT_NAME} → {SEX, BDATE, RELATIONSHIP}



DEPENDENT	ESSN	DEPENDENT_NAME	SEX	BDATE	RELATIONSHIP
	333445555	Alice	F	1986-04-05	DAUGHTER
	333445555	Theodore	M	1983-10-25	SON
	333445555	Joy	F	1958-05-03	SPOUSE
	987654321	Abner	M	1942-02-28	SPOUSE
	123456789	Michael	M	1988-01-04	SON
	123456789	Alice	F	1988-12-30	DAUGHTER
	123456789	Elizabeth	F	1967-05-05	SPOUSE

key

3 Normal Forms Based on Primary Keys

3.1 Normalization of Relations

3.2 Practical Use of Normal Forms

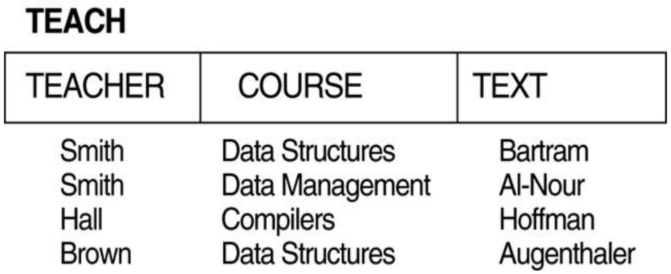
3.3 Definitions of Keys and Attributes Participating in Keys

3.4 First Normal Form

3.5 Second Normal Form

3.6 Third Normal Form

Which NF is TEACH in?

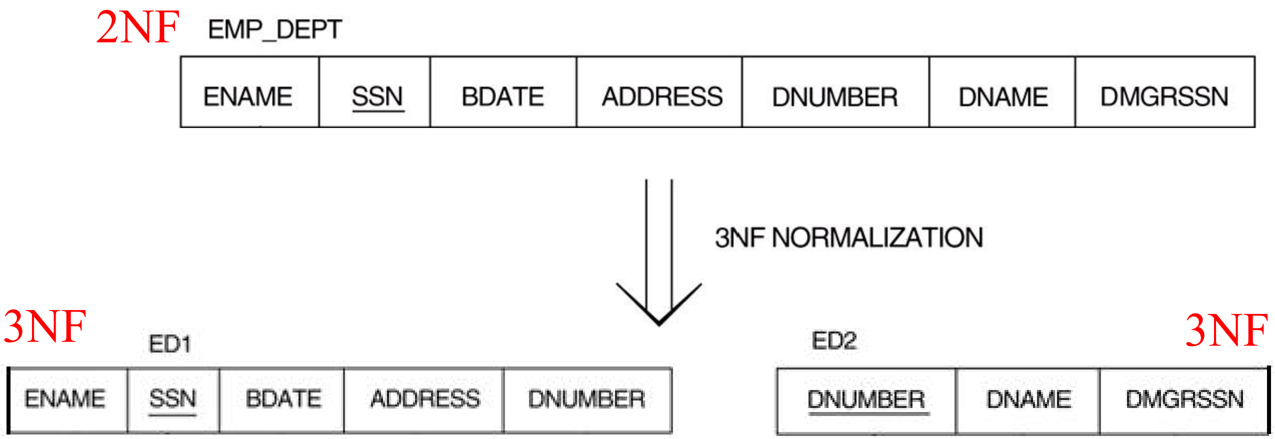


TEACH

TEACHER	COURSE	TEXT
Smith	Data Structures	Bartram
Smith	Data Management	Al-Nour
Hall	Compilers	Hoffman
Brown	Data Structures	Augenthaler

14.3.1 Normalization of Relations

- **Normalization**:
  - The process of decomposing unsatisfactory "bad" relations by breaking up their attributes into smaller relations
- **Normal form**:
  - Condition using keys and FDs of a relation to certify whether a relation schema is in a particular normal form
- 2NF, 3NF, BCNF based on keys and FDs of a relation schema
- 4NF based on keys and multi-valued dependencies (MVD)
- 5NF based on keys and join dependencies (JD)



2NF EMP_DEPT

ENAME	SSN	BDATE	ADDRESS	DNUMBER	DNAME	DMGRSSN

3NF NORMALIZATION

3NF ED1

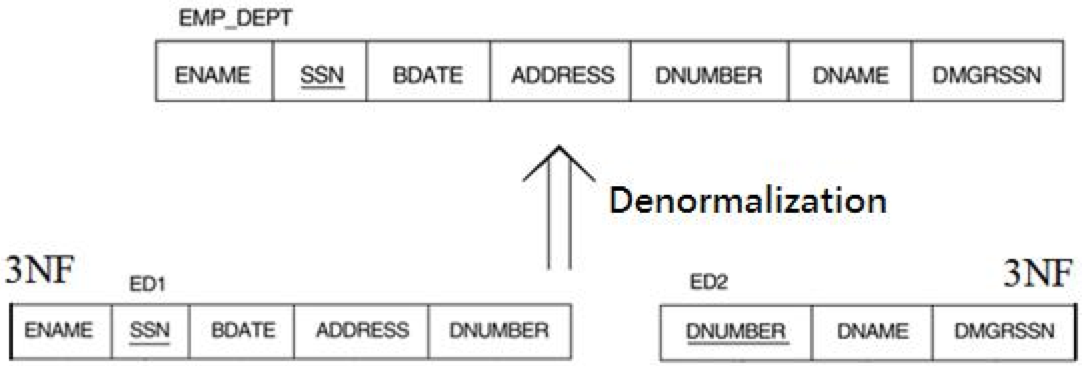
ENAME	SSN	BDATE	ADDRESS	DNUMBER

ED2 3NF

DNUMBER	DNAME	DMGRSSN

14.3.2 Practical Use of Normal Forms

- Normalization is carried out in practice so that the resulting designs are of high quality and meet the desirable properties
- The practical utility of these normal forms becomes questionable when the constraints on which they are based are **hard to understand** or to **detect**
- The database designers ***need not*** normalize to the highest possible normal form. (usually up to 3NF, BCNF or 4NF)
- **Denormalization:**
  - the process of storing the join of higher normal form relations as a base relation—which is in a lower normal form



14.3.3 Definitions of Keys and Attributes Participating in Keys (1)

- A **superkey** of a relation schema $R = \{A_1, A_2, ..., A_n\}$
  - a set of attributes S *subset-of* R with the property that no two tuples t_1 and t_2 in any legal relation state r of R will have $t_1[S] = t_2[S]$
- A **key** K is a superkey with the *additional property* that removal of any attribute from K will cause K not to be a superkey any more.
- If a relation schema has more than one key, each is called a **candidate key.** One of the candidate keys is *arbitrarily* designated to be the **primary key,** and the others are called *secondary keys*.
- A **Prime attribute** must be a member of ***some candidate key.***
- A **Nonprime attribute** is not a prime attribute.

  that is, it is not a member of any candidate key.

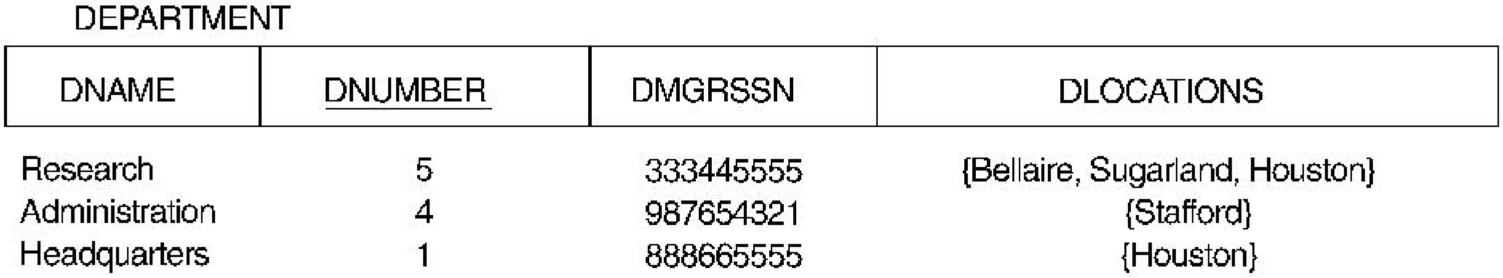
ED1

ENAME	SSN	EID	BDATE	ADDRESS	DNUMBER

- {SSN, EID, ADDRESS}
- {SSN}
- {EID}
- {DNUMBER}

14.3.4 First Normal Form

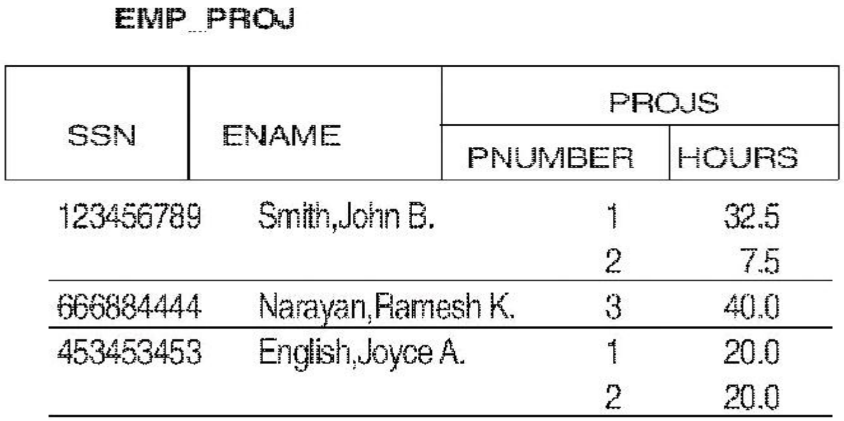
- 1NF disallows **composite attributes**, **multivalued attributes**, and **nested relations**; attributes whose values *for an individual tuple* are non-atomic



DEPARTMENT

DNAME	DNUMBER	DMGRSSN	DLOCATIONS
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

Multivalued attribute



EMP_PROJ

SSN	ENAME	PROJS	
		PNUMBER	HOURS
123456789	Smith,John B.	1	32.5
		2	7.5
666884444	Narayan,Ramesh K.	3	40.0
453453453	English,Joyce A.	1	20.0
		2	20.0

Nested relation

Normalization into 1NF

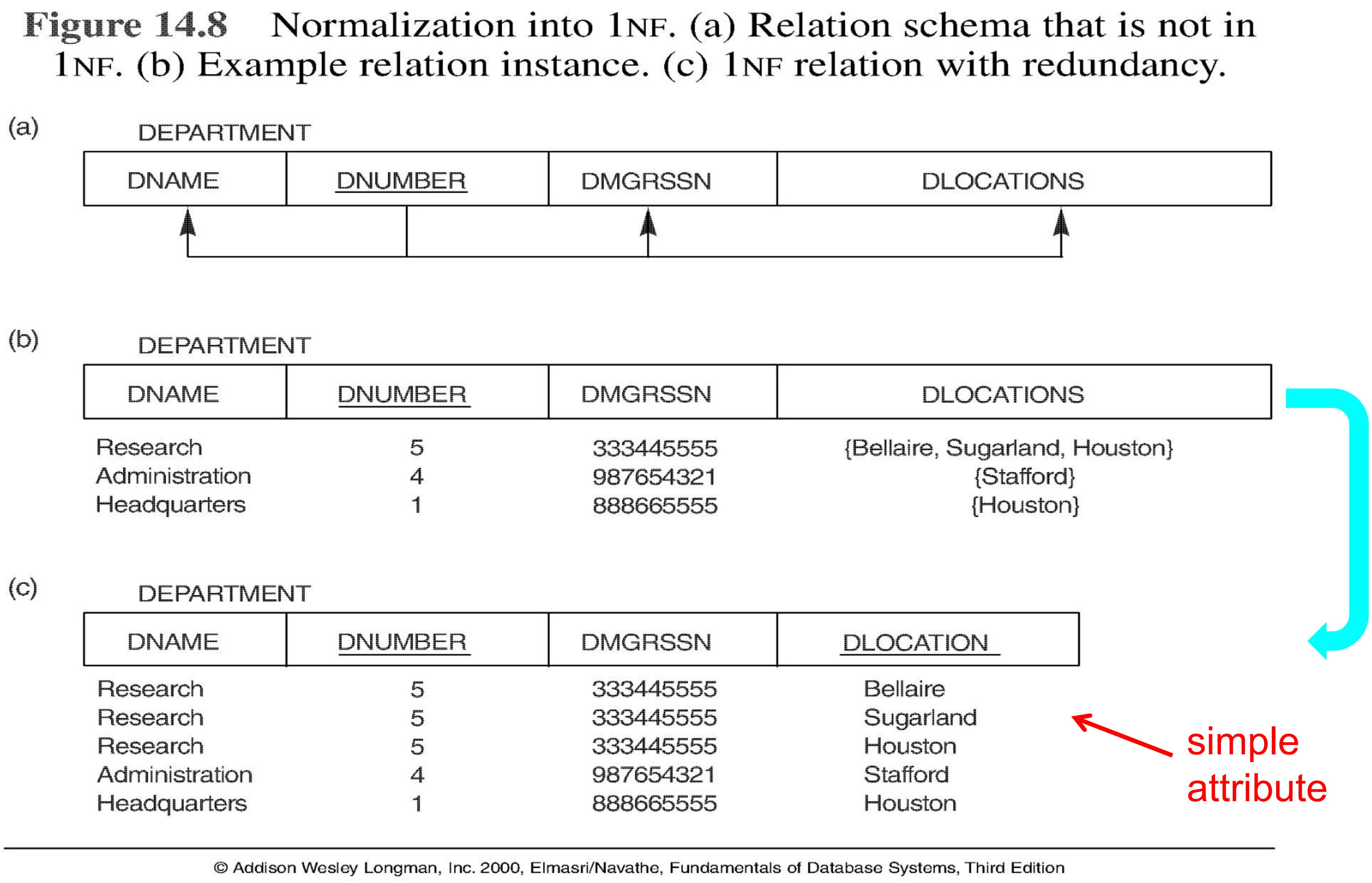


Figure 14.8 Normalization into 1NF. (a) Relation schema that is not in 1NF. (b) Example relation instance. (c) 1NF relation with redundancy.

(a) DEPARTMENT

DNAME	DNUMBER	DMGRSSN	DLOCATIONS

(b) DEPARTMENT

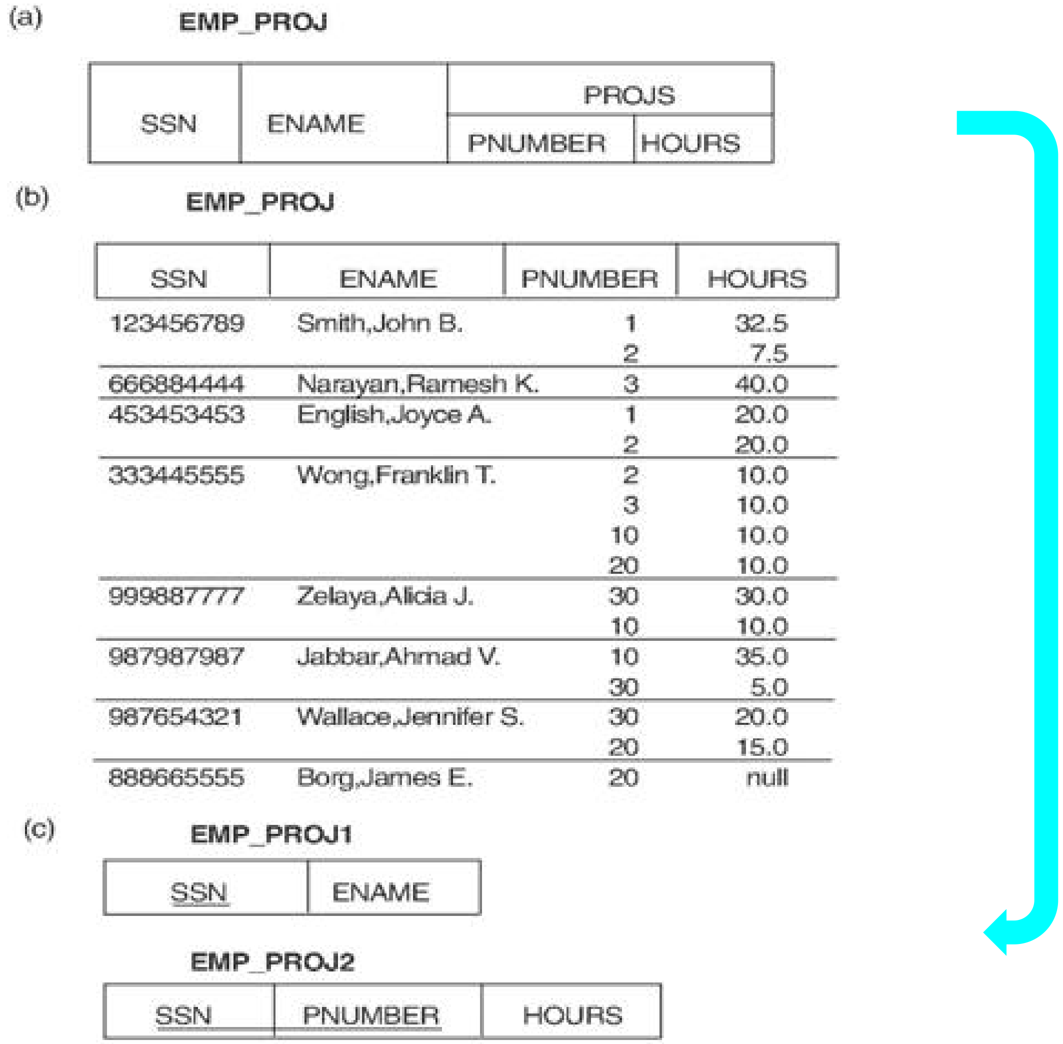
DNAME	DNUMBER	DMGRSSN	DLOCATIONS
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

(c) DEPARTMENT

DNAME	DNUMBER	DMGRSSN	DLOCATION
Research	5	333445555	Bellaire
Research	5	333445555	Sugarland
Research	5	333445555	Houston
Administration	4	987654321	Stafford
Headquarters	1	888665555	Houston

simple attribute

Normalization nested relations into 1NF



(a) EMP_PROJ

SSN	ENAME	PROJS	
		PNUMBER	HOURS

(b) EMP_PROJ

SSN	ENAME	PNUMBER	HOURS
123456789	Smith,John B.	1	32.5
		2	7.5
666884444	Narayan,Ramesh K.	3	40.0
453453453	English,Joyce A.	1	20.0
		2	20.0
333445555	Wong,Franklin T.	2	10.0
		3	10.0
		10	10.0
		20	10.0
999887777	Zelaya,Alicia J.	30	30.0
		10	10.0
987987987	Jabbar,Ahmad V.	10	35.0
		30	5.0
987654321	Wallace,Jennifer S.	30	20.0
		20	15.0
888665555	Borg,James E.	20	null

(c) EMP_PROJ1

SSN	ENAME

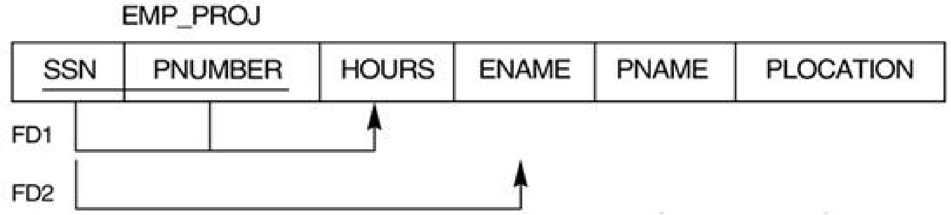
EMP_PROJ2

SSN	PNUMBER	HOURS

Figure 14.9 Normalizing nested relations into 1NF. (a) Schema of the EMP_PROJ relation with a "nested relation" PROJS. (b) Example extension of the EMP_PROJ relation showing nested relations within each tuple. (c) Decomposing EMP_PROJ into 1NF relations EMP_PROJ1 and EMP_PROJ2 by propagating the primary key.

Full Functional Dependency

- Uses the concepts of **primary key** and **FDs**
- **Prime attribute**
  - attribute that is member of the primary key K
- **Full functional dependency**
  - a FD Y → Z where removal of any attribute from Y means the FD does not hold any more



Examples:

- **{SSN, PNUMBER} → ENAME is *not*** a full FD

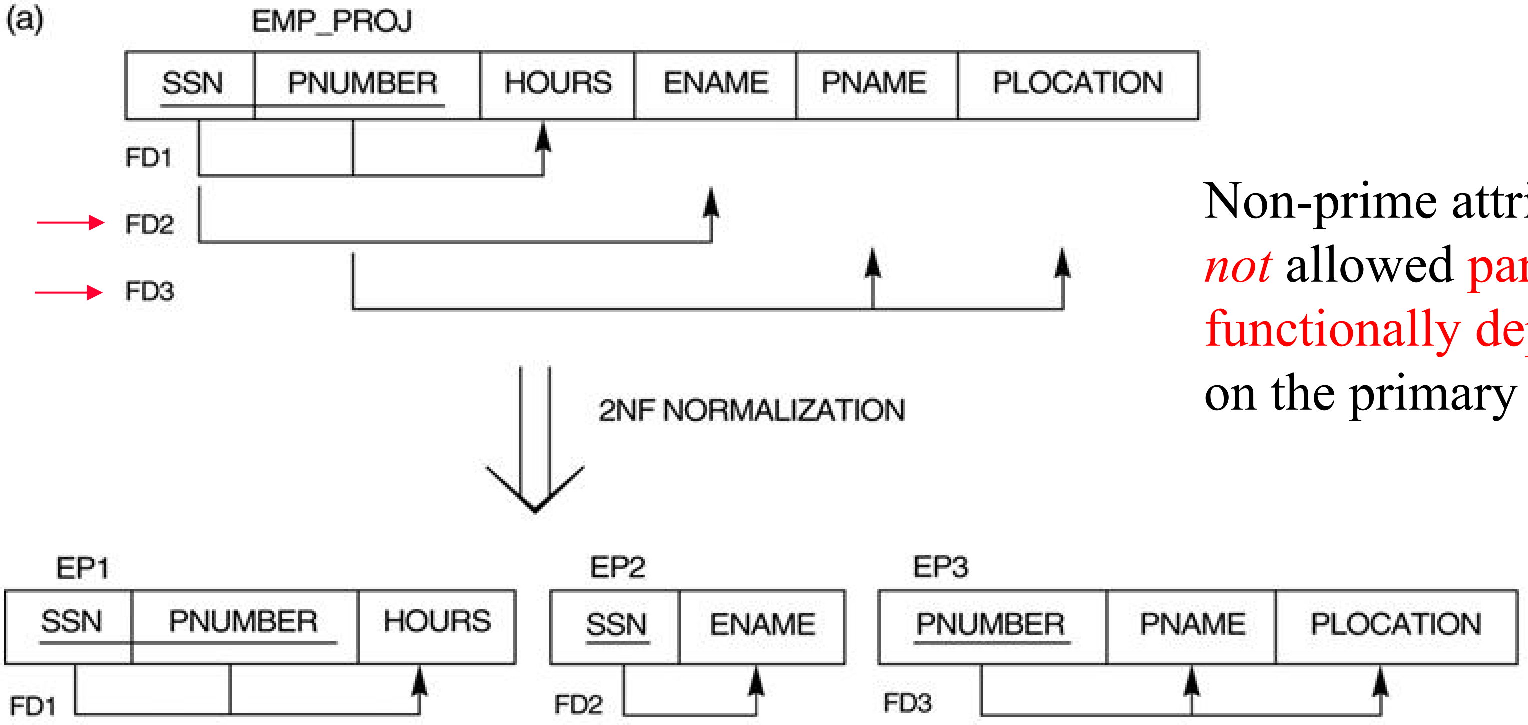
  since SSN → ENAME also holds (it is called a ***partial dependency***)

- **{SSN, PNUMBER} → HOURS is** a full FD

  since neither SSN →HOURS nor PNUMBER → HOURS hold

Second Normal Form

- A relation schema R is in **second normal form (2NF)**
  - if every non-prime attribute A in R is fully functionally dependent on the **primary** key
- R can be decomposed into 2NF relations via the process of 2NF normalization



Non-prime attribute is *not* allowed partially functionally dependent on the primary key.

Transitive Functional Dependency

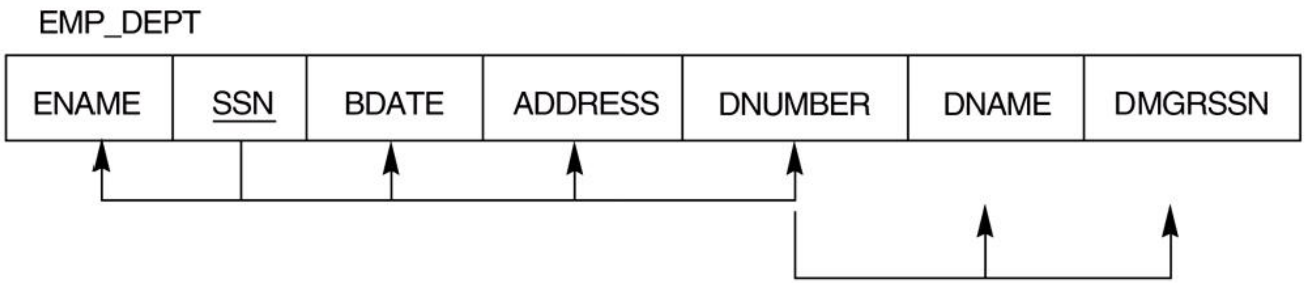
Definition:

- **Transitive functional dependency**
  - a FD X→Z that can be derived from two FDs X→Y and Y→Z

Examples:

- **SSN→DMGRSSN** is a ***transitive*** FD
  since SSN→DNUMBER and DNUMBER→DMGRSSN hold

- **SSN→ENAME** is ***non-transitive***
  since there is no set of attributes X where SSN→X and X→ENAME



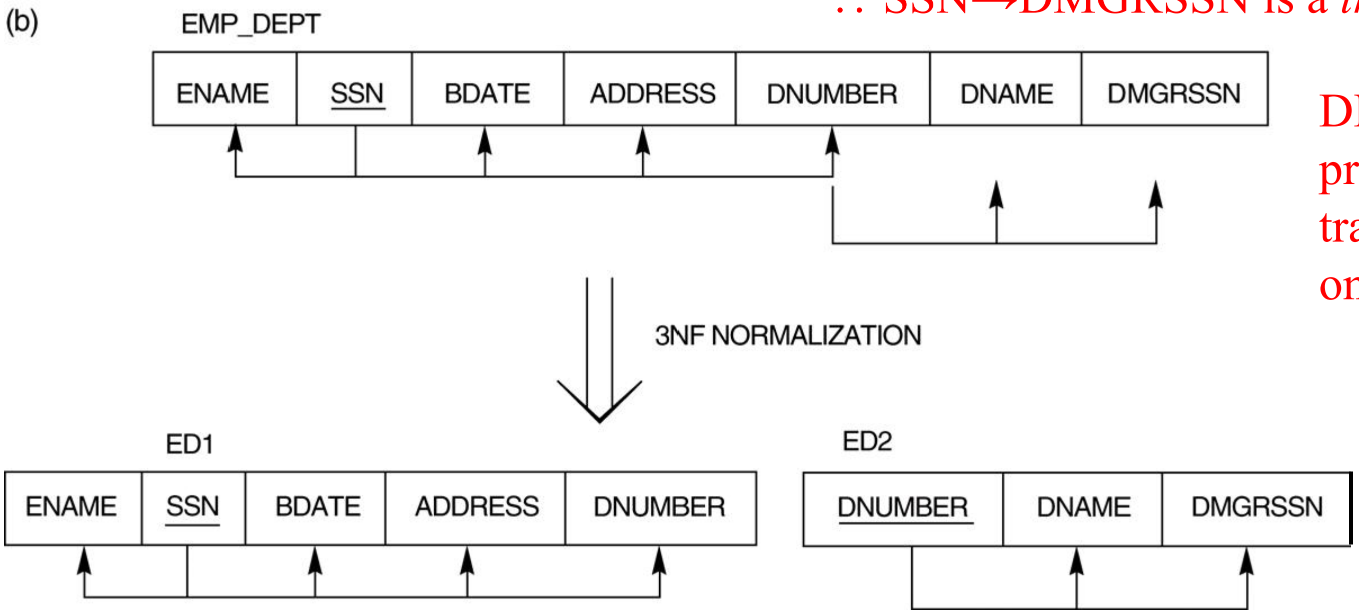
EMP_DEPT

ENAME	SSN	BDATE	ADDRESS	DNUMBER	DNAME	DMGRSSN

Third Normal Form

- A relation schema R is in **third normal form (3NF)**
  - if it is in 2NF and ***no*** **non-prime attribute** A in R is transitively dependent on the primary key
- R can be decomposed into 3NF relations via the process of 3NF normalization

SSN→DNUMBER
DNUMBER→DMGRSSN
∴ SSN→DMGRSSN is a *transitive* FD



(b)

EMP_DEPT

ENAME	SSN	BDATE	ADDRESS	DNUMBER	DNAME	DMGRSSN

3NF NORMALIZATION

ED1

ENAME	SSN	BDATE	ADDRESS	DNUMBER

ED2

DNUMBER	DNAME	DMGRSSN

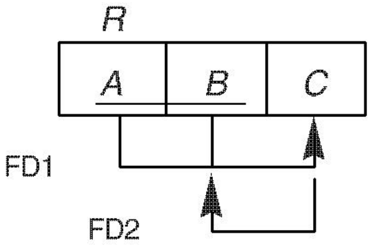
DMGRSSN is a non-prime attribute and transitively dependent on the primary key.

4. General Normal Form Definitions (For Multiple Keys)

- The previous definitions consider the primary key only
- The following more general definitions take into account relations with multiple candidate keys
- **Superkey** of relation schema R
  - a set of attributes S of R that contains a key of R
- **Second normal form (2NF) R:**
  - if every non-prime attribute A in R is fully functionally dependent on ***every key*** of R
- **Third normal form (3NF) R:**
  - if whenever a FD X→Y holds in R, then either:

    (a) X is a superkey of R, or

    (b) Y is a prime attribute of R



$AB \rightarrow C$

$C \rightarrow B$

Example of General Third Normal Form

- **Example**

  In X → Y and Y → Z, with X as the primary key,
  we consider this a problem only if **Y** is ***not*** a **superkey**.

  When **Y** is a **superkey**, there is no problem with the transitive dependency .

  E.g., Consider EMP(<u>SSN</u>, Emp#, Salary).

  SSN → Emp# ; Emp# → Salary

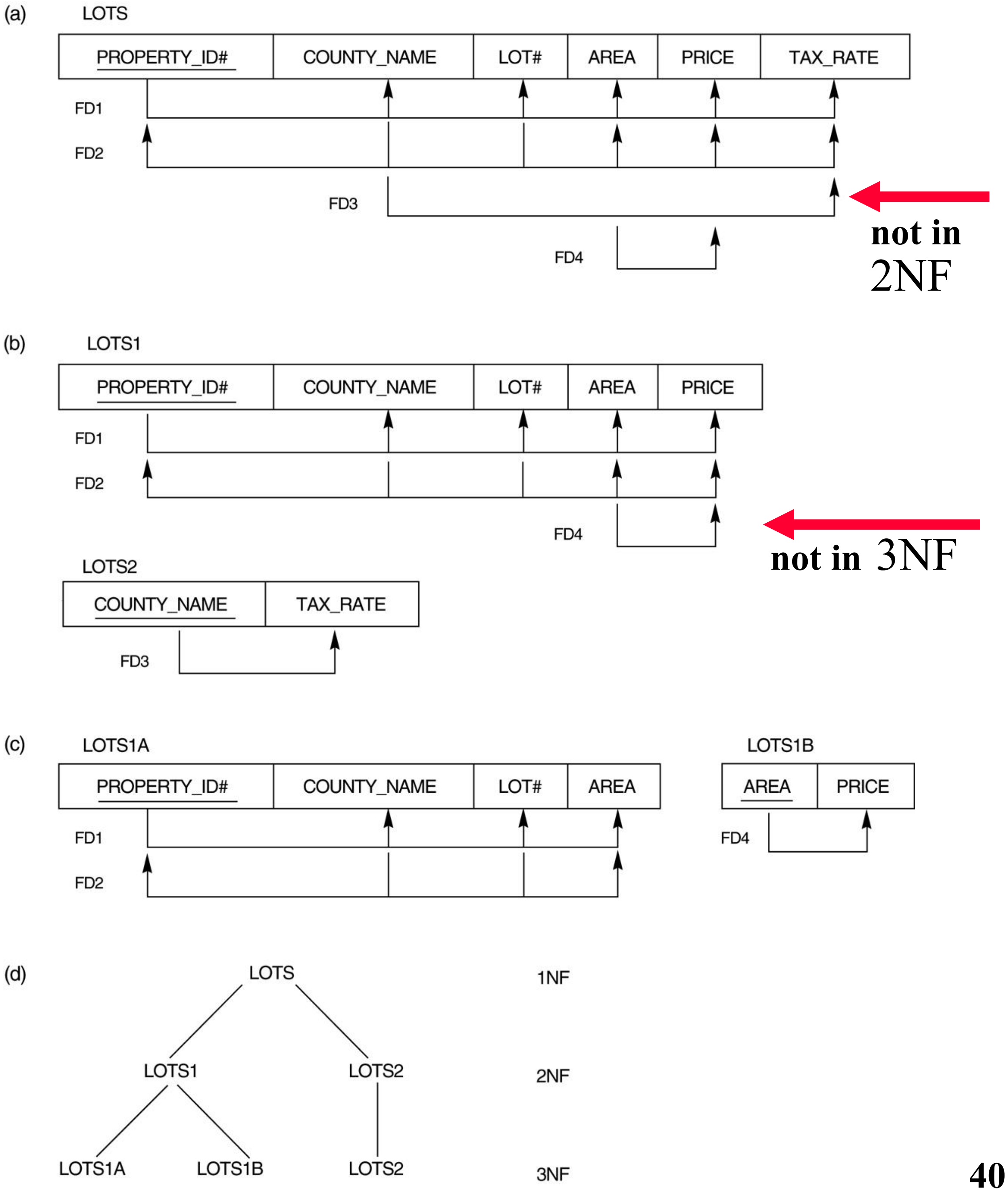
  Here, SSN → Salary (no problem, since **Emp#** is a **superkey**)

  The relation is in 3rd normal form.

FIGURE

Normalization into 2NF and 3NF.

(a) the LOTS relation with its functional dependencies FD1 though FD4.
(b) Decomposing into the 2NF relations LOTS1 and LOTS2.
(c) Decomposing LOTS1 into the 3NF relations LOTS1A and LOTS1B.
(d) Summary of the progressive normalization of LOTS.



(a) LOTS

PROPERTY_ID# | COUNTY_NAME | LOT# | AREA | PRICE | TAX_RATE

FD1

FD2

FD3

FD4

not in 2NF

(b) LOTS1

PROPERTY_ID# | COUNTY_NAME | LOT# | AREA | PRICE

FD1

FD2

FD4

not in 3NF

LOTS2

COUNTY_NAME | TAX_RATE

FD3

(c) LOTS1A

PROPERTY_ID# | COUNTY_NAME | LOT# | AREA

FD1

FD2

LOTS1B

AREA | PRICE

FD4

(d)

LOTS

LOTS1

LOTS2

LOTS1A

LOTS1B

LOTS2

1NF

2NF

3NF

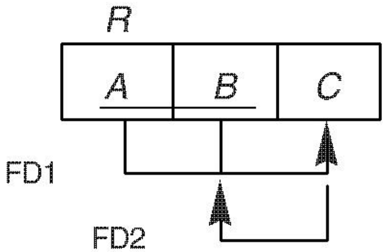
In (a)
PROPERTY_ID#
is a primary key.

{COUNTY_NAME, LOT#}
is a candidate key.

5 BCNF (Boyce-Codd Normal Form)

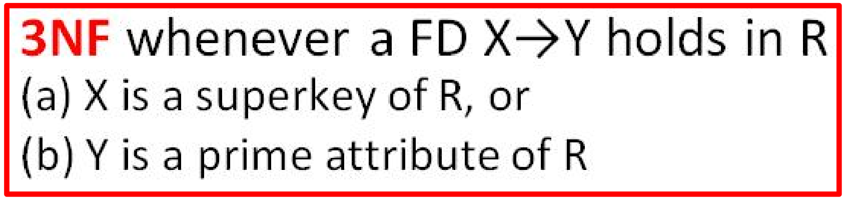
- A relation schema R is in **Boyce-Codd Normal Form (BCNF)**
  - if whenever an FD X→Y holds in R, then X is a superkey of R

Example: R in 3NF but not in BCNF



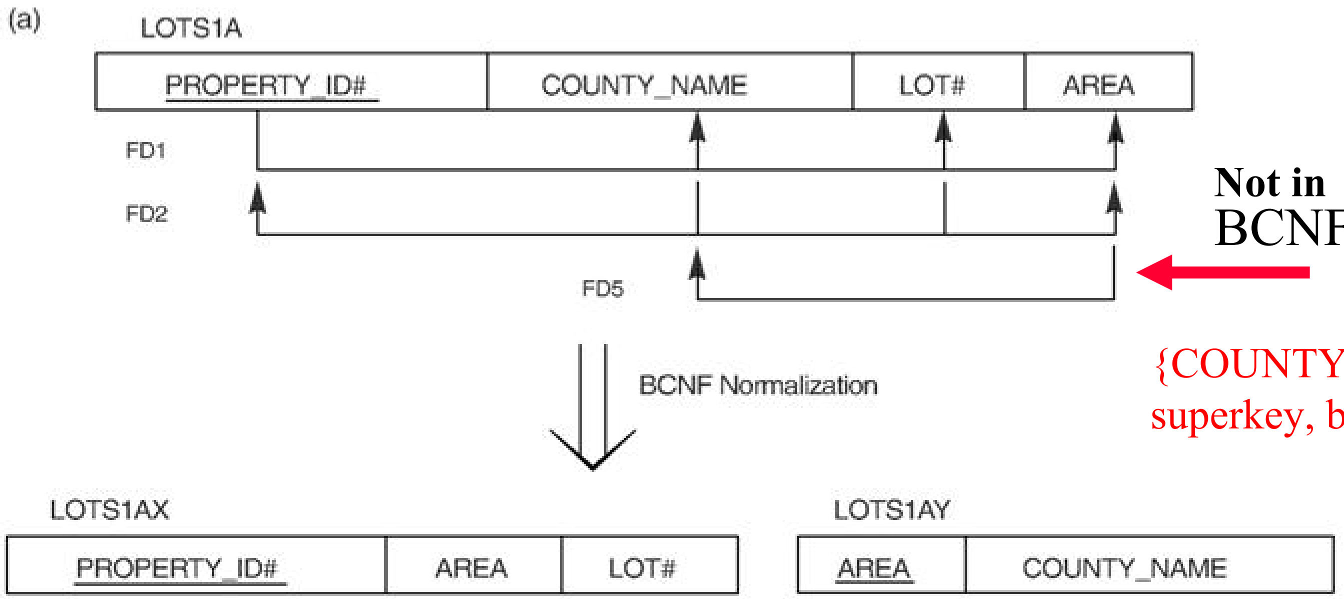
AB→C
C→B

- Each normal form is strictly stronger than the previous one
  - Every 2NF relation is in 1NF
  - Every 3NF relation is in 2NF
  - Every BCNF relation is in 3NF
- There exist relations that are in 3NF but not in BCNF
- DB design goal is to have each relation in BCNF (or 3NF)


3NF whenever a FD X→Y holds in R
(a) X is a superkey of R, or
(b) Y is a prime attribute of R

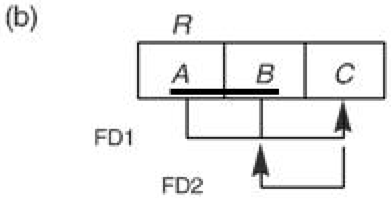
Boyce-Codd Normal Form

Figure 14.12 Boyce-Codd normal form. (a) BCNF normalization with the dependency of FD2 being "lost" in the decomposition. (b) A relation *R* in 3NF but not in BCNF.



Not in BCNF

{COUNTY_NAME, LOT#} is a superkey, but {AREA} is not.



AB→C
C→B

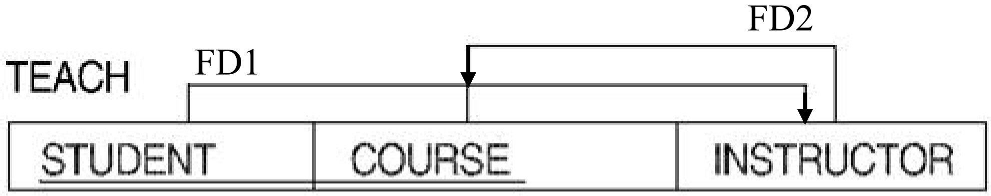
3NF whenever a FD X→Y holds in R
(a) X is a superkey of R, or
(b) Y is a prime attribute of R

BCNF whenever a FD X→Y holds in R then X is a superkey of R

Figure
a relation TEACH that is in 3NF but not in BCNF

FD1: STUDENT COURSE → INSTRUCTOR

FD2: INSTRUCTOR→COURSE



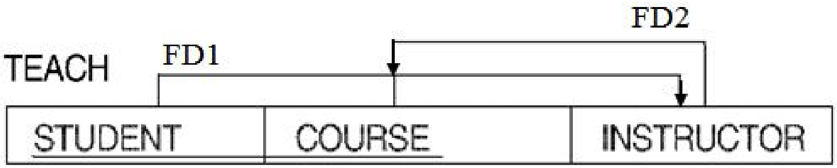
STUDENT	COURSE	INSTRUCTOR
Narayan	Database	Mark
Smith	Database	Navathe
Smith	Operating Systems	Ammar
Smith	Theory	Schulman
Wallace	Database	Mark
Wallace	Operating Systems	Ahamad
Wong	Database	Omiecinski
Zelaya	Database	Navathe

3NF whenever a FD X→Y holds in R
(a) X is a superkey of R, or
(b) Y is a prime attribute of R

BCNF whenever a FD X→Y holds in R then X is a superkey of R

Achieving the BCNF by Decomposition (2)

- Three possible decompositions for relation TEACH
  1. {student, instructor} and {student, course}
  2. {course, instructor } and {course, student}
  3. **{instructor, course } and {instructor, student}**
- All three decompositions will lose FD1. We have to settle for sacrificing the functional dependency preservation. But we **cannot** sacrifice the non-additivity property (i.e., lossless joint property) after decomposition.
- Out of the above three, **only the 3rd decomposition will not generate spurious tuples after join**.(and hence has the non-additivity property).
- A test to determine whether a binary decomposition (decomposition into two relations) is nonadditive (lossless) is discussed in section 15.2.4 under Property LJ1. Verify that the third decomposition above meets the property.

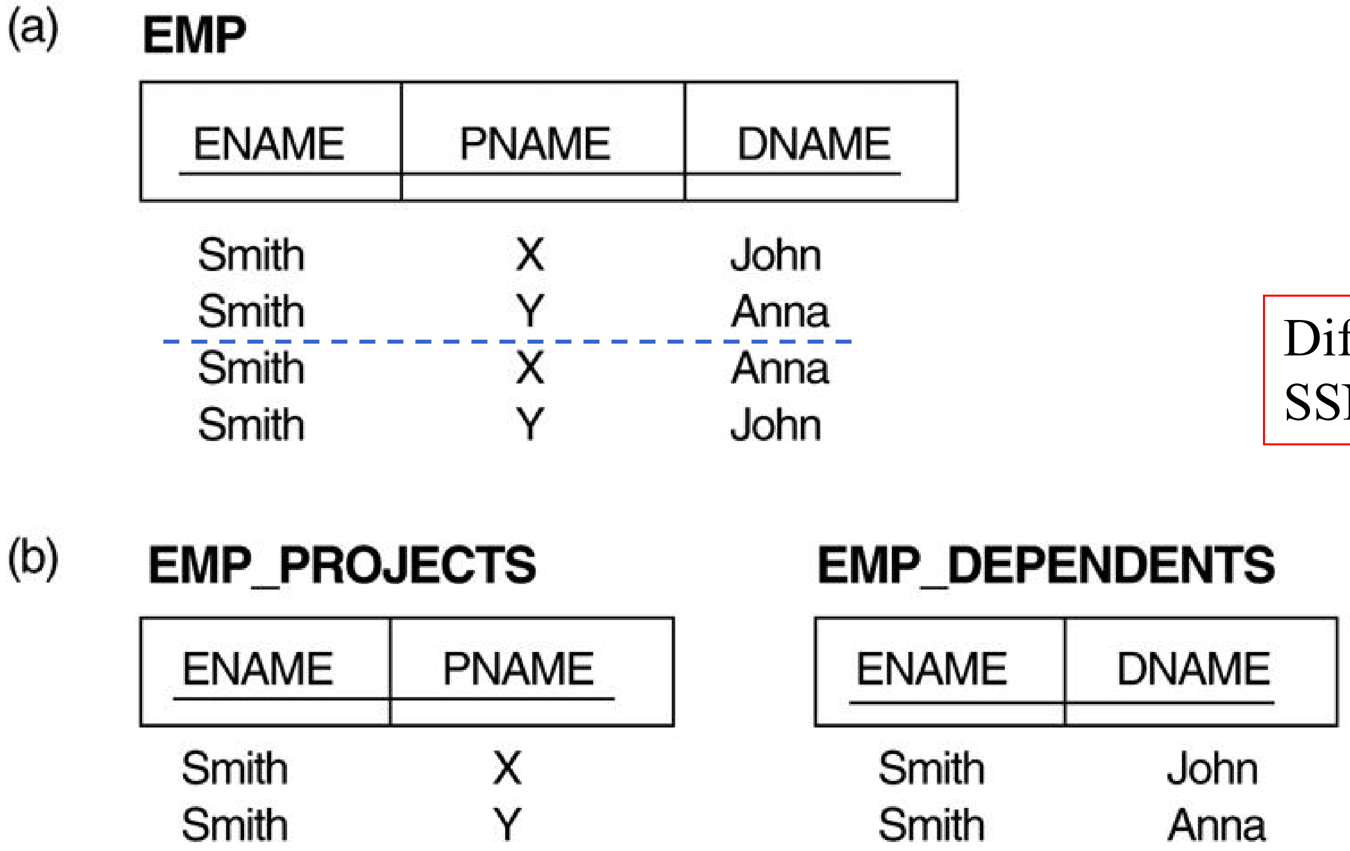


FD1: STUDENT COURSE → INSTRUCTOR
FD2: INSTRUCTOR → COURSE

6. Multivalued Dependencies and Fourth Normal Form

(a) The EMP relation with two MVDs: ENAME —>> PNAME and ENAME —>> DNAME.

(b) Decomposing the EMP relation into two 4NF relations EMP_PROJECTS and EMP_DEPENDENTS.



(a) **EMP**

ENAME	PNAME	DNAME
Smith	X	John
Smith	Y	Anna
Smith	X	Anna
Smith	Y	John

(b) **EMP_PROJECTS**

ENAME	PNAME
Smith	X
Smith	Y

EMP_DEPENDENTS

ENAME	DNAME
Smith	John
Smith	Anna

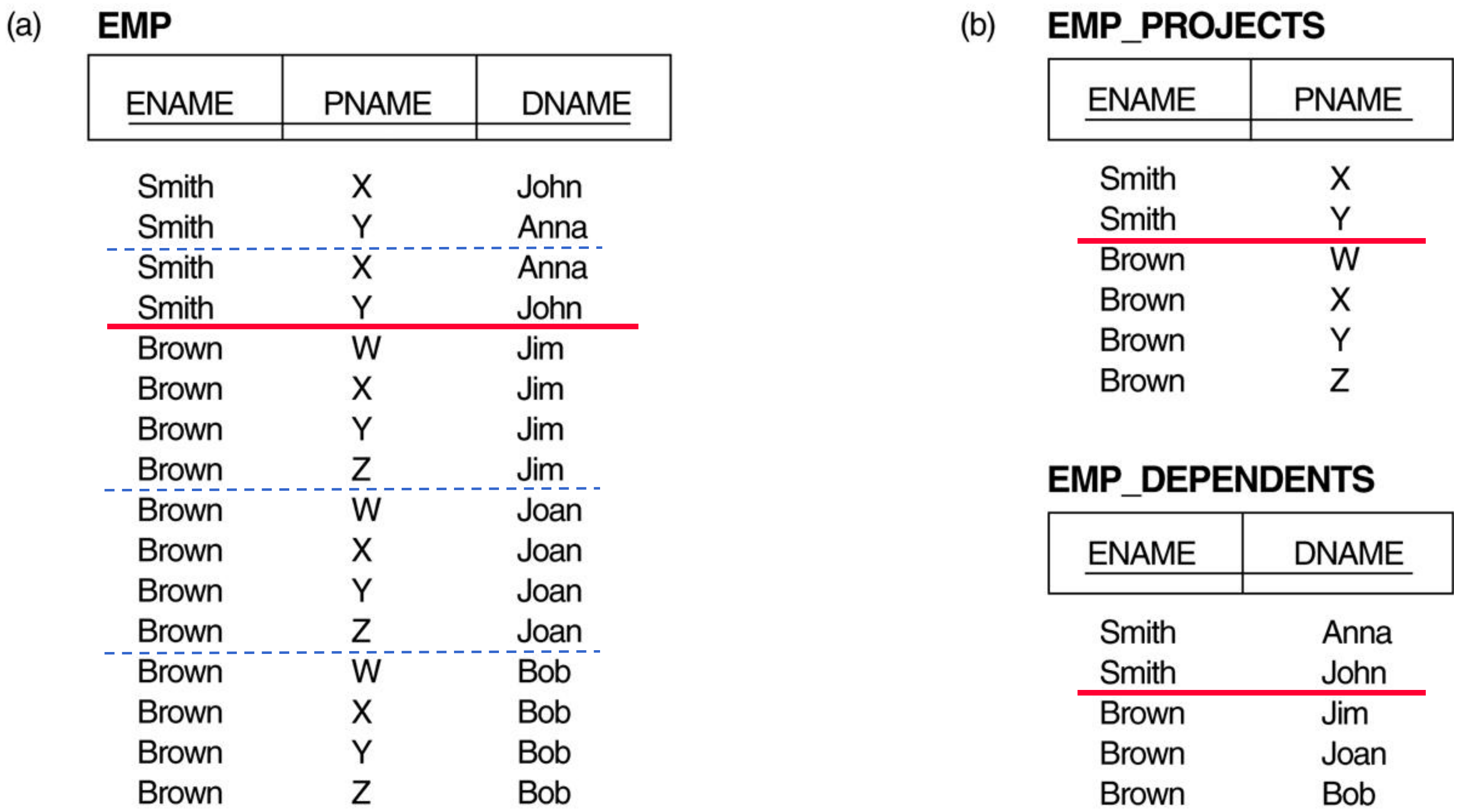
Different from F.D.:
SSN→ENAME

Multivalued Dependencies and Fourth Normal Form

Decomposing a relation state of EMP that is not in 4NF.

(a) EMP relation with additional tuples.

(b) Two corresponding 4NF relations EMP_PROJECTS and EMP_DEPENDENTS.



(a) **EMP**

ENAME	PNAME	DNAME
Smith	X	John
Smith	Y	Anna
Smith	X	Anna
Smith	Y	John
Brown	W	Jim
Brown	X	Jim
Brown	Y	Jim
Brown	Z	Jim
Brown	W	Joan
Brown	X	Joan
Brown	Y	Joan
Brown	Z	Joan
Brown	W	Bob
Brown	X	Bob
Brown	Y	Bob
Brown	Z	Bob

(b) **EMP_PROJECTS**

ENAME	PNAME
Smith	X
Smith	Y
Brown	W
Brown	X
Brown	Y
Brown	Z

EMP_DEPENDENTS

ENAME	DNAME
Smith	Anna
Smith	John
Brown	Jim
Brown	Joan
Brown	Bob

Multivalued Dependencies

- A **multivalued dependency (MVD)** X —>> Y specified on relation schema R, where X and Y are both subsets of R, specifies the following constraint on any relation state r of R:

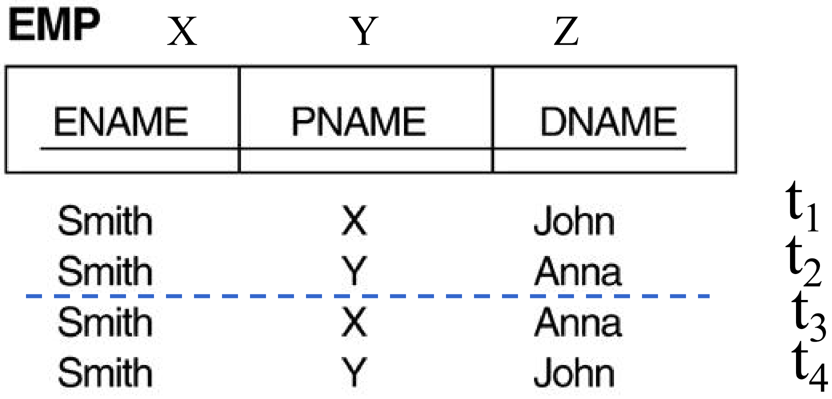
  If two tuples t_1 and t_2 exist in r such that $t_1[X] = t_2[X]$, then two tuples t_3 and t_4 should also exist in r with the following properties, where we use Z to denote $(R - (X \cup Y))$:

  $t_3[X] = t_4[X] = t_1[X] = t_2[X]$.

  $t_3[Y] = t_1[Y]$ and $t_4[Y] = t_2[Y]$.

  $t_3[Z] = t_2[Z]$ and $t_4[Z] = t_1[Z]$.

ENAME —>> PNAME



EMP

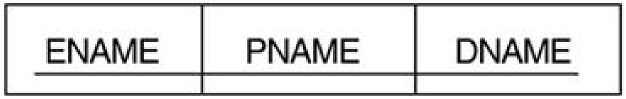
X	Y	Z	
ENAME	PNAME	DNAME	
Smith	X	John	t_1
Smith	Y	Anna	t_2
Smith	X	Anna	t_3
Smith	Y	John	t_4

Fourth Normal Form

- A relation schema *R* is in 4NF with respect to a set of dependencies *F* (including functional dependencies and multivalued dependencies)

  If, for every ***nontrivial*** multivalued dependency *X* —>> *Y* in F^+, *X* is a superkey for R.

EMP



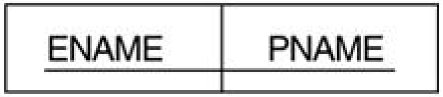
ENAME	PNAME	DNAME

ENAME —>> PNAME

- Note:
  - An MVD *X* —>> *Y* in *R* is called a **trivial MVD**

    if (a) *Y* is a subset of *X*, or (b) $X \cup Y = R$.
  - F^+ is the (complete) set of all dependencies (functional or multivalued) that will hold in every relation state *r* of *R* that satisfies *F*. It is also called the **closure** of *F*.

EMP_PROJECTS



ENAME	PNAME

ENAME —>> PNAME

4. Join Dependencies and Fifth Normal Form

- A **join dependency (JD)**, denoted by JD(R_1, R_2, ..., R_n), specified on relation schema *R*,
  - Every legal state *r* of *R* should have a non-additive join decomposition into R_1, R_2, ..., R_n;
  - that is, for every such *r* we have

    $* (\pi_{R1}(r), \pi_{R2}(r), ..., \pi_{Rn}(r)) = r$

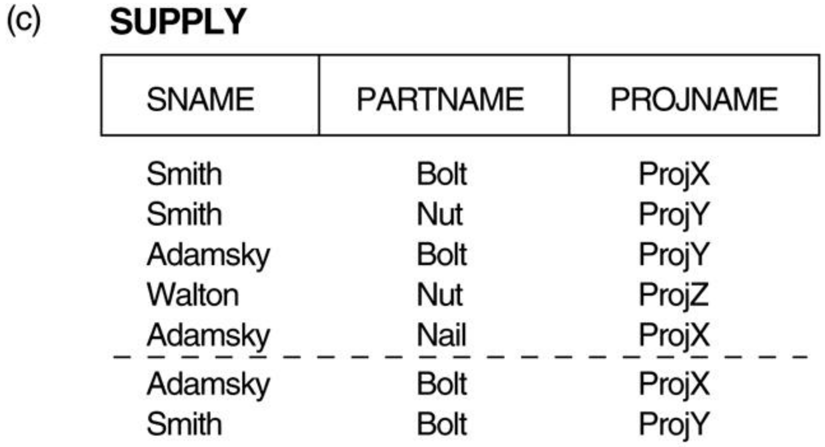
  ***Note**: an MVD is a special case of a JD where n = 2.*

- A join dependency JD(R_1, R_2, ..., R_n), specified on relation schema *R*, is a **trivial JD** if one of the relation schemas R_i in JD(R_1, R_2, ..., R_n) is equal to *R*.

Join Dependencies and Fifth Normal Form

- A relation schema *R* is in **fifth normal form (5NF)** (or **Project-Join Normal Form (PJNF)**) with respect to a set *F* of functional, multivalued, and join dependencies

  if, for every nontrivial join dependency JD(R_1, R_2, ..., R_n) in F^+ (that is, implied by *F*), every R_i is a superkey of *R*.

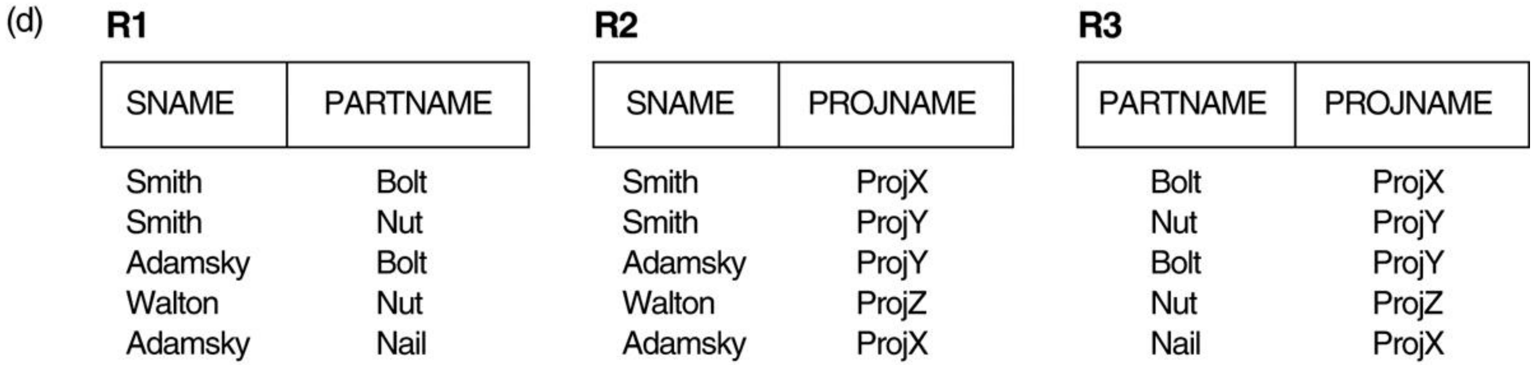


(c) **SUPPLY**

SNAME	PARTNAME	PROJNAME
Smith	Bolt	ProjX
Smith	Nut	ProjY
Adamsky	Bolt	ProjY
Walton	Nut	ProjZ
Adamsky	Nail	ProjX
Adamsky	Bolt	ProjX
Smith	Bolt	ProjY

(c) The relation SUPPLY with no MVDs is in 4NF but not in 5NF if it has the JD(R1, R2, R3).

(d) Decomposing the relation SUPPLY into the 5NF relations R1, R2, and R3.



(d) **R1**

SNAME	PARTNAME
Smith	Bolt
Smith	Nut
Adamsky	Bolt
Walton	Nut
Adamsky	Nail

R2

SNAME	PROJNAME
Smith	ProjX
Smith	ProjY
Adamsky	ProjY
Walton	ProjZ
Adamsky	ProjX

R3

PARTNAME	PROJNAME
Bolt	ProjX
Nut	ProjY
Bolt	ProjY
Nut	ProjZ
Nail	ProjX