UPPSALA UNIVERSITET

Normalization and Data quality (1)

lecturer: Mani Pelmo/Neena Thota

mpelmo@sherubtse.edu.bt

Neena.Thota@it.uu.se

Where are we?

Database design

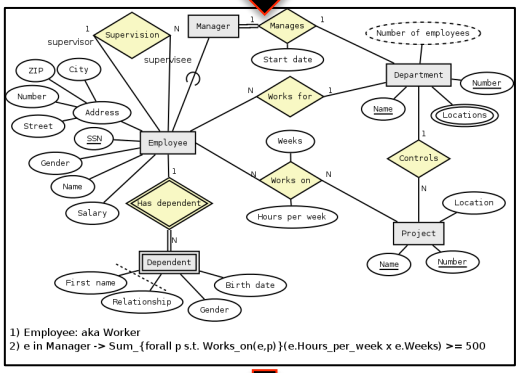
- Intro
- Database design 4 lectures
- LABS 1
- Data quality / Normalization 2 lectures
- LABS 2
- SQL

- Motivation & terminology
- The (E)ER conceptual model
- The relational model
- From ER to relational
- SQL and DBMSs (DDL)

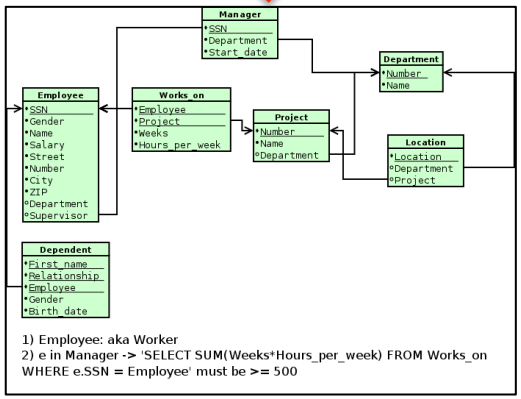
Natural language

An enterprise consists of a number of departments. Each department has a name, a number, a manager, and a number of employees. The starting date for every department manager should also be registered. A department can have several locations. Every department controls a number of projects. Each project has a unique name, a unique number (both unique only inside the project's department) and a location. For each employee, the following information is kept: name, social security number, address, salary and sex. An employee works for only one department but can work with several projects that can be related to different departments. An employee may also supervise one or more other workers. Information about the number of hours (per week) that an employee works with each project should be stored – to be a manager one must have worked at least 500 hours on projects. We also want to keep track of the dependents of each employee, for insurance purposes. We keep each dependent's first name, sex, birth date and relationship to the worker.

Entity-Relationship diagram



Relational model



SQL / DBMS


```
create table EMP
(SSN int ,
Name varchar,
…
```

UPPSALA UNIVERSITET

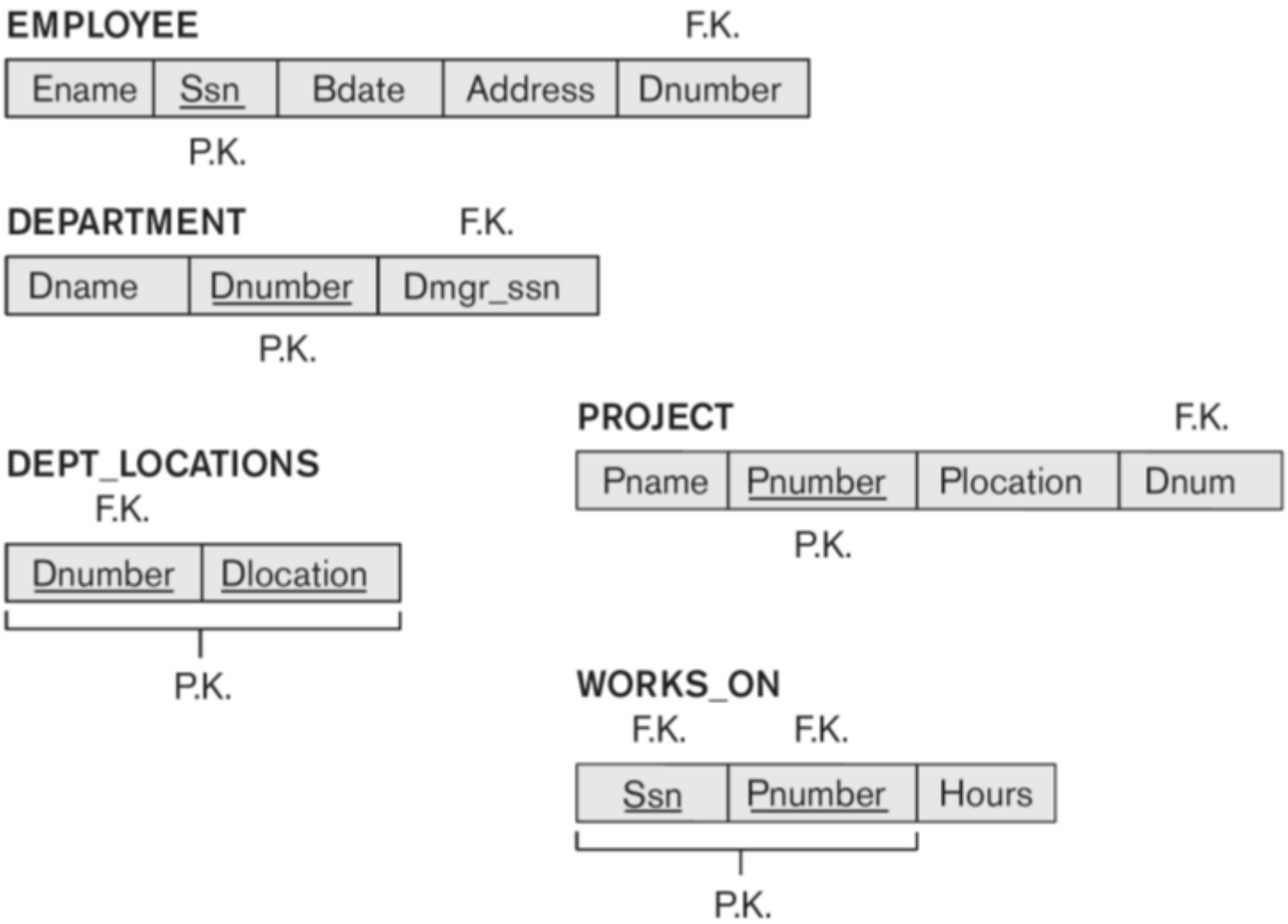
Intended Learning Outcomes

- Understand the main problems that may occur when a database schema is poorly designed.
  - Insertion, deletion and update problems, spurious tuples.
- Recognize poorly designed schemata.
- Explain the concepts of:
  - (Full) functional dependency.
  - Prime attributes.

UPPSALA UNIVERSITET

Overview

- What is relational database design?
  - The grouping of attributes to form "good" relation schemas
- What are the criteria for "good“ relations?
- We first discuss informal guidelines for good relational design
- Then we discuss formal concepts of functional dependencies and normal forms
  - 1NF (First Normal Form)
  - 2NF (Second Normal Form)
  - 3NF (Third Normal Form)
  - BCNF (Boyce-Codd Normal Form)


EMPLOYEE
F.K.
Ename
Ssn
Bdate
Address
Dnumber
P.K.
DEPARTMENT
F.K.
Dname
Dnumber
Dmgr_ssn
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.

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	291Berry, Bellaire, TX	4
Narayan, Ramesh 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	Dmgr_ssn
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

UPPSALA UNIVERSITET

Informal Design Guidelines for Relation Schemas

- These guidelines may be used as *measures to determine the quality of relation* schema design:
  1. Making sure that the semantics of the attributes is clear in the schema.
  2. Reducing the redundant information in tuples.
  3. Reducing NULL values in tuples.
  4. Disallowing the possibility of generating spurious tuples.

1. Imparting Clear Semantics to Attributes in Relations

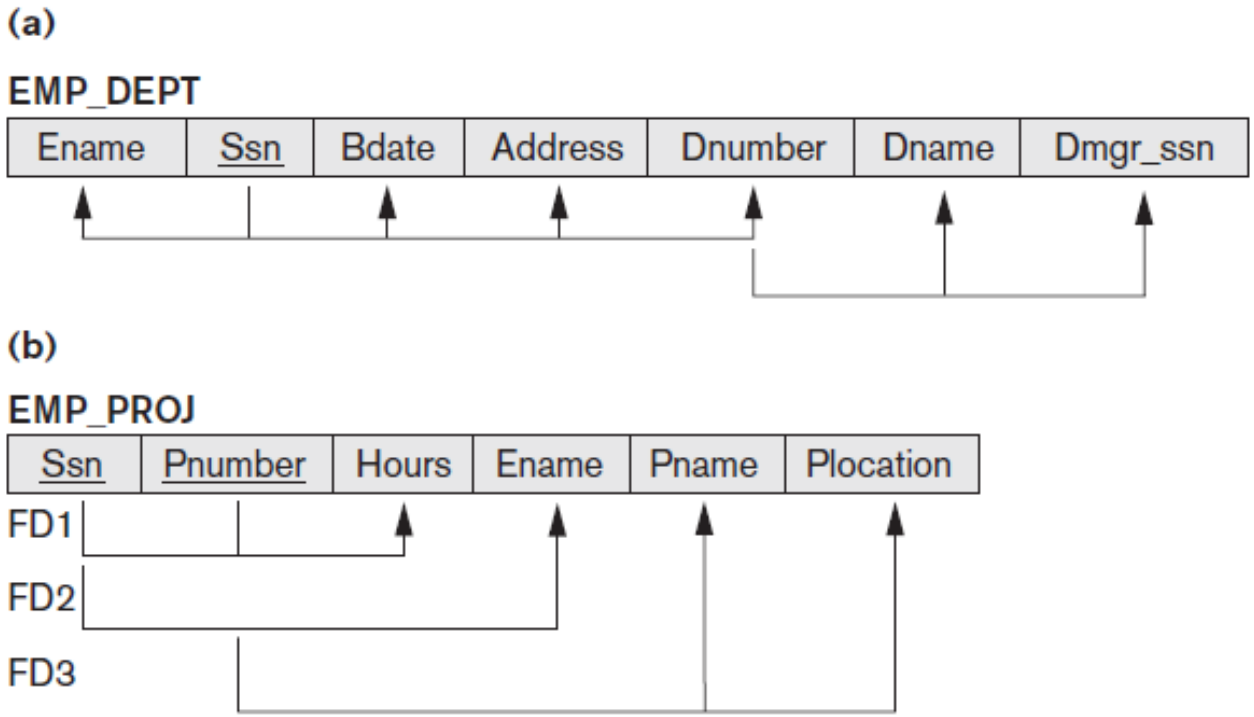
- **semantics of a relation r**efers to its meaning resulting from the interpretation of attribute values in a tuple.
- **Easier it is to explain the semantics of the relation**, the **better the relation schema design** will be.
- Ex. The meaning of the EMPLOYEE relation schema is quite simple:

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

Guideline 1 summary

- Design relation schema so that it is easy to explain its meaning
- Do not combine attributes from multiple entity types and relationship types into a single relation
    - Only foreign keys should be used to refer to other entities

Ex. Violating Guideline 1



Poor design!

- **EMP_DEPT** mixes attributes of **Employee** and **Department**
- **EMP_PROJ** mixes attributes of **Employee** and **Project** and the **Works_on** relationship.

2. Reducing the redundant information in tuples

- Minimize the storage space used by base relations
- Grouping attributes into relation schemas has a significant effect on storage space.

UPPSALA UNIVERSITET

Ex. Redundant information in tuples

Redundancy

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 FireOak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555

(Dnumber,Dname,Dmgrssn) are repeated.

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

UPPSALA UNIVERSITET

Reducing the redundant information in tuples (cont.)

- Information stored redundantly
  - Wastes storage
  - Causes problems with update anomalies
    - Insertion anomalies
    - Deletion anomalies
    - Modification anomalies

UPPSALA UNIVERSITET

Example of an insert anomaly

- Consider the relation

  EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321

Insert anomaly

- To insert new tuple for an employee who works in department number # must also enter the values for **Dname** and **Dmgrssn** correctly for *consistency*.
- To insert new department that has no employees yet results in placing NULL values in the attributes of employee.
  - causes a problem- **ssn** is the primary key of the relation.

UPPSALA UNIVERSITET

Example of a delete anomaly

- Consider the relation

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321

Delete anomaly

- Deleting any employee from the relation will delete all the information relating to that department.
- Deleting any department will result in deleting all the information of an employee who is working in that department.

UPPSALA UNIVERSITET

Example of an update anomaly

- Consider the relation

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321

Update anomaly

- Changing the manager of department number 5 may cause this update to be made for all employees working in department number 5.
- Failing to update some tuples would result in inconsistency.

Guideline 2 summary

- Design relation schemas that does not suffer from **insertion, deletion and update** anomalies.

UPPSALA UNIVERSITET

3. Reducing NULL values in tuples

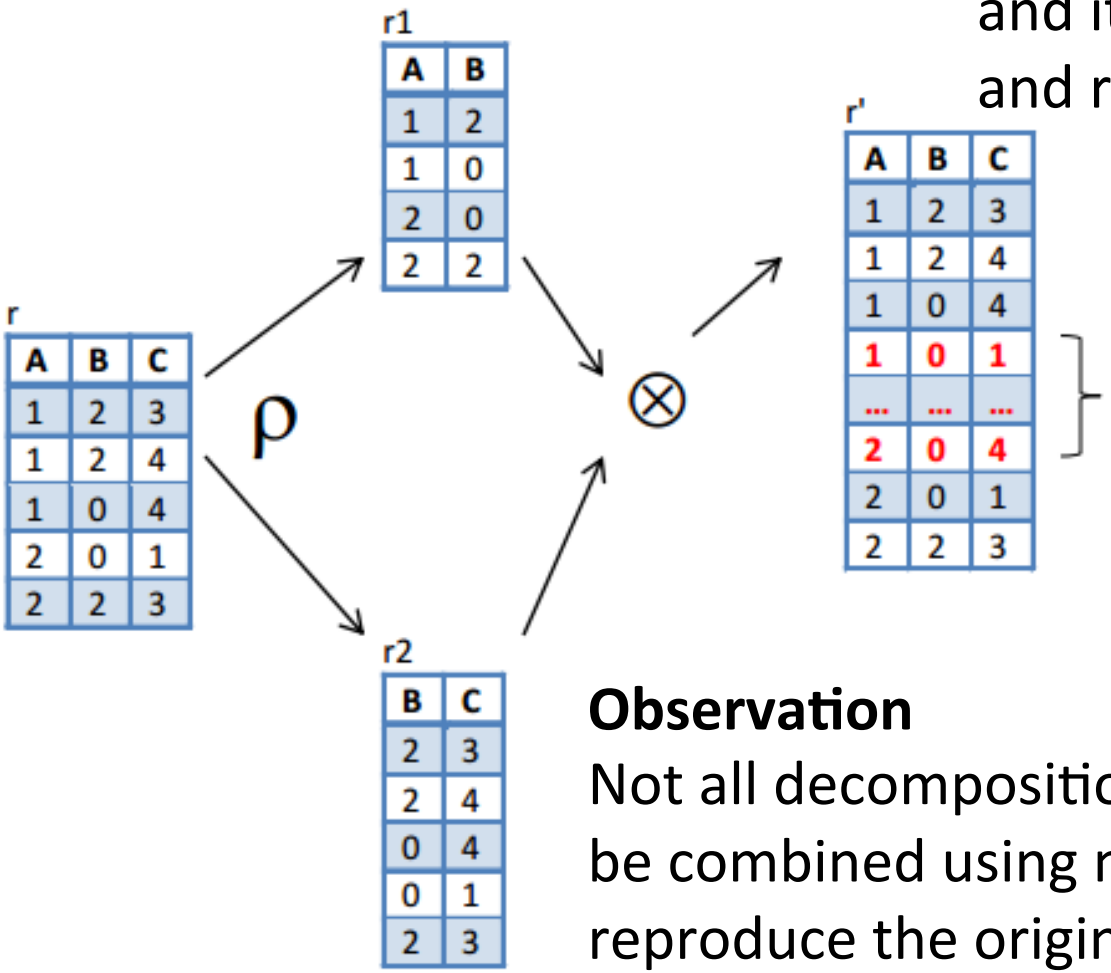
- May group many attributes together into a "fat" relation
  - If many of the attributes do not apply to all tuples in the relation, we end up with many NULLs
- Problems with NULLs
  - Wasted storage space
  - Problems understanding meaning
- Reasons for nulls:
  - Attribute not applicable or invalid
  - Attribute value unknown (may exist)
  - Value known to exist, but unavailable

Guideline 3 summary

- Design relations such that their tuples will have as few NULL values as possible
- Place attributes that are NULL frequently in separate relations (with the primary key)
- Ex. if only 15 percent of employees have individual offices,
  - there is little justification for including an attribute Office_number in the EMPLOYEE relation;
  - Create relation **EMP_OFFICES(Essn, Office_number)** to include tuples for only the employees with individual offices.

4. Generation of spurious tuples

Consider relation r(ABC) and its projections r1(AB) and r2(BC).

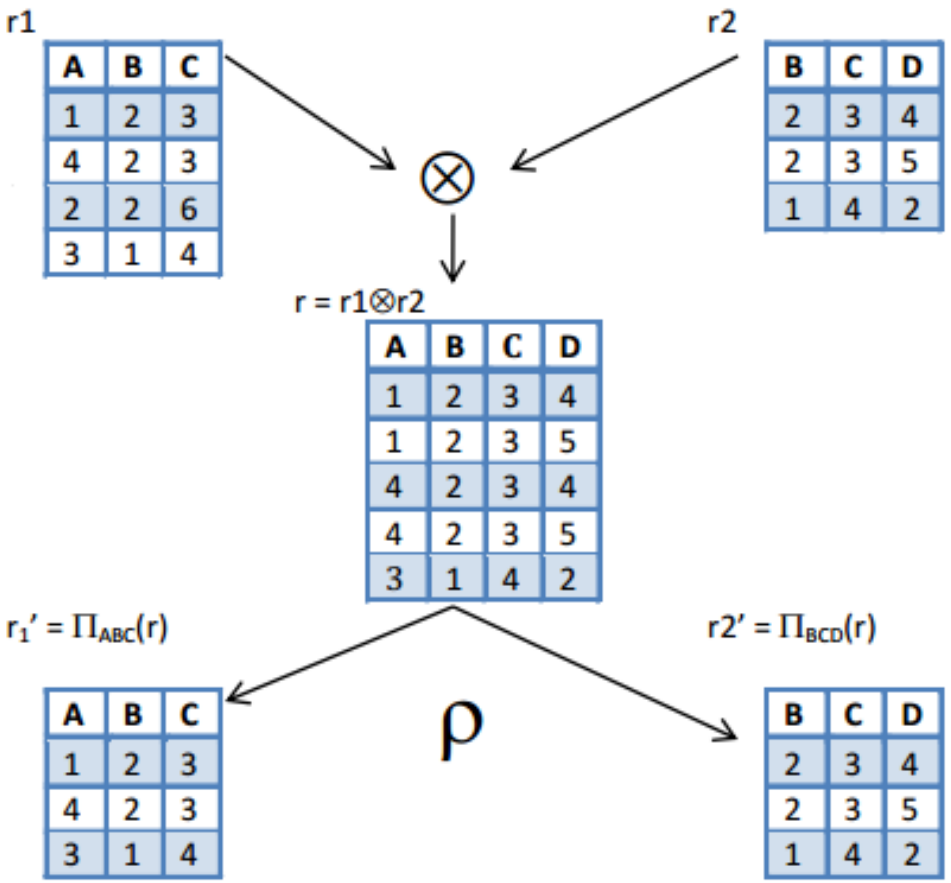


Observation

Not all decompositions of a table can be combined using natural join to reproduce the original table.

Generation of spurious tuples (cont.)

Consider the following two relations r1(ABC) and r2 (BCD).

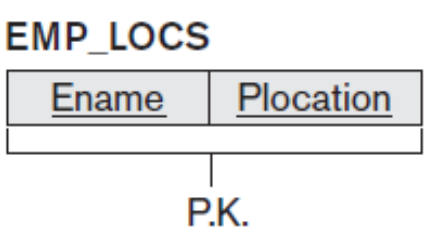


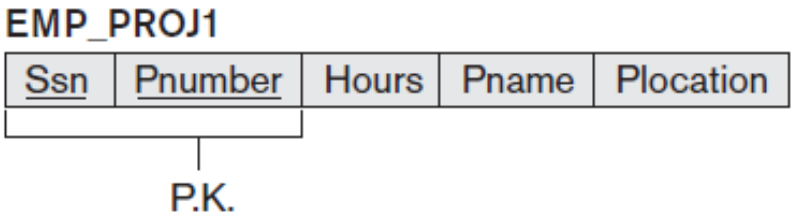
Compute natural join $r = r_1 * r_2$

Evaluate projections
$r1' = \Pi_{ABC}(r)$ and
$r2' = \Pi_{BCD}(r)$

Observation
Tables r2 and r2' are the same however tuple <2,2,6> in r1 is not present in r1'.

Generation of spurious tuples (cont.)





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

EMP_P...

Ssn	Pnumber	Hours	Pname	Plocation
123456789	1	32.5	ProductX	Bellaire
123456789	2	7.5	ProductY	Sugarland
666884444	3	40.0	ProductZ	Houston
453453453	1	20.0	ProductX	Bellaire
453453453	2	20.0	ProductY	Sugarland
333445555	2	10.0	ProductY	Sugarland
333445555	3	10.0	ProductZ	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

Generation of spurious tuples (cont.)

Natural join of EMP_LOCS and EMP_PROJ1

	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.

*= spurious tuples

Result produces many more tuples than the original set of tuples in EMP_PROJ

- Called **spurious tuples**
- Represent spurious information that is not valid

Guideline 4 summary

- Design relation schemas that can be joined with equality conditions using only the primary key and foreign keys.
- Avoid relations that contain matching attributes that are not foreign and primary keys.
- The relations should be designed to satisfy the lossless join condition.

Let’s have Break!

Normalization

- A set of principles to be followed systematically to prevent the aforementioned problems.
- In 1972, Codd defined a set of such principles.
- To fulfill them, the relation schema is divided into smaller schemas in several steps.
- This process is called **normalization**.
- We need to study the following concepts for performing the normalization:
  - Normal forms for relations.
  - Functional dependencies.

UPPSALA UNIVERSITET

Motivation

Why Normalization?

- Formal way to analyze why one grouping of attributes into a relational schema is better than another.
- Provides algorithms to improve the design.
- If you design the database as we have done so far, you will probably NOT need normalization.
- Unfortunately, typically many people put their hands on a database during its lifecycle.

First Normal Form (INF)

- A relation is in first normal form (1NF) if
  - There are no repeating groups in the relation, i.e. **all column values must be atomic.**
  - A **primary key has been defined**, which uniquely identifies each row in the relation.
- INF disallows:
  - multivalued attributes
  - **nested relations** (combination of composite and multivalued)

EMP_PROJ

		Projs	
Ssn	Ename	Pnumber	Hours

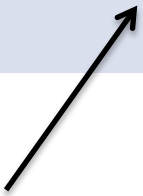
Ssn	Ename	Pnumber	Hours
123456789	Smith, John B.	1	32.5
		2	7.5
666884444	Narayan, Ramesh K.	3	40.0

UPPSALA UNIVERSITET

Is this relation in first normal form - 1NF?

A relation is in 1NF if all attributes contain only atomic values.

Ssn	Ename	Pnumber	Hours
1234	Smith	1	12
		2	7
4534	Wong	3	40
		2	26



NOT in 1NF

UPPSALA UNIVERSITET

Functional Dependencies (FD)

- Formal tool for analysis of relational schemas
- Enables us to detect and describe some of the above-mentioned problems in precise terms
- Are used to specify *formal measures of the* "goodness" of relational designs
- Are **constraints** between two attributes or two sets of attributes.

UPPSALA UNIVERSITET

Definition of Functional Dependency

- A set of attributes X *functionally determines a set* of attributes Y (denoted by X→Y) if the value of X determines a unique value for Y
- X → Y holds if whenever two tuples have the same value for X, they *must have the same value for Y*
  - 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)

UPPSALA UNIVERSITET

Examples of FD constraints

- SSN number determines employee name
  - SSN → ENAME
- Project number determines project name and location
  - PNUMBER → {PNAME, PLOCATION}
- Employee SSN number and project number determines the hours per week that the employee works on the project
  - {SSN, PNUMBER} -> HOURS

UPPSALA UNIVERSITET

Definition of Functional Dependency (cont.)

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

Ex. Determine all the FDs that hold and that does not hold.

A	B	C	D
a1	b1	c1	d1
a1	b2	c2	d2
a2	b2	c2	d3
a3	b3	c4	d3

The following FDs *holds:*

$B \rightarrow C$

$C \rightarrow B$

$\{A, B\} \rightarrow C$

$\{A, B\} \rightarrow D$

$\{C, D\} \rightarrow B$

The following do not:

$A \rightarrow B$ (tuples 1 and 2 violate this constraint);

$B \rightarrow A$ (tuples 2 and 3 violate this constraint);

$D \rightarrow C$ (tuples 3 and 4 violate it).

UPPSALA UNIVERSITET

Second Normal Form(2NF)

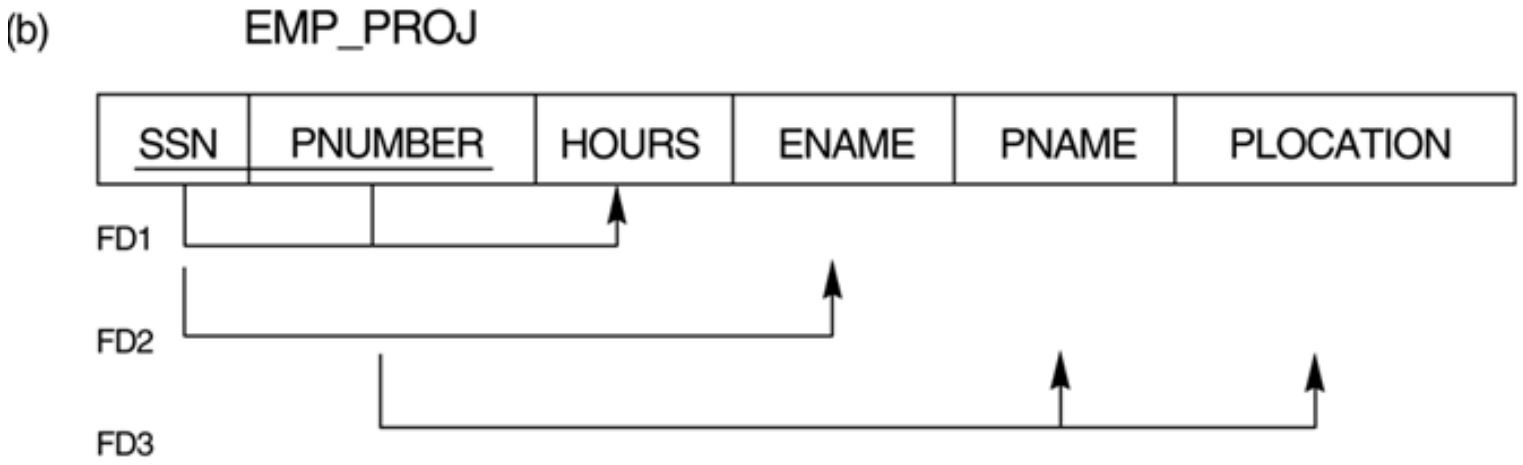
- **Prime attribute**: An attribute that is member of the primary key K.
- **Full functional dependency:** a functional dependency X→Y is a full functional dependency if removal of any attribute from X means that the dependency does not hold any more.
- **Partial functional dependency:** a functional dependency X→Y is a partial functional dependency if some attribute can be removed from X and the dependency still holds.

A relation schema R is in 2NF if every nonprime attribute in R is fully functionally dependent on the primary key of R.

The test for 2NF involves testing for FDs whose LHS attributes are part of the PK. If the PK contains a single attribute, the test does not need to be done.

UPPSALA UNIVERSITET

2NF - Example



FD1: ssn,pnumber →Hours

FD2: ssn→ename

FD3: pnumber→pname,plocation

Third Normal Form(3NF)

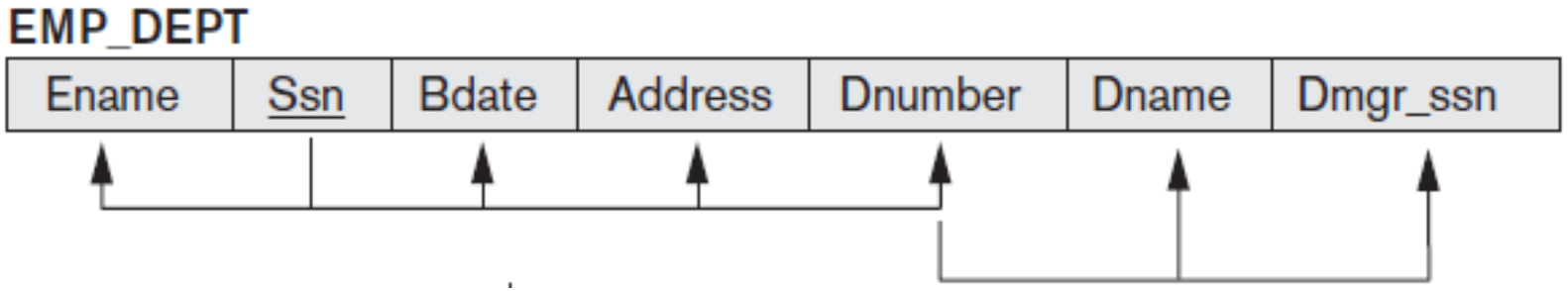
- **Transitive dependency:** a functional dependency $X \rightarrow Y$ in a relation schema R is a transitive dependency if there exists a set of attributes Z in R that is neither a candidate key nor a subset of an key of R, and both $X \rightarrow Z$ and $Z \rightarrow Y$ hold.

A relation schema R is in third normal form (3NF) if

- it is in 2NF
- no non-prime attribute A in R is transitively dependent on the primary key

UPPSALA UNIVERSITET

3NF - Example



- The dependency Ssn→Dmgr_ssn is transitive through Dnumber in EMP_DEPT as

  Ssn → Dnumber

  Dnumber → Dmgr_ssn

  Note that Dnumber is neither a key itself nor a subset of the key of EMP_DEPT.

UPPSALA UNIVERSITET

Note

- In X → Z and Z → Y, with X as the primary key, we consider this a problem only if Z is not a candidate key.
- If Z is a candidate key, there is no problem with the transitive dependency .
- Example:
  - Consider EMP (SSN, Emp#, Salary).

Here, SSN → Emp# → Salary

Emp# is a candidate key.

UPPSALA UNIVERSITET

Ex: Find FDs

Consider the following relation for published books:

BOOK (Book_title, Author_name, Book_type, List_price, Author_affil, Publisher)

Book_title → Publisher, Book_type

Book_type → List_price

Author_name → Author_affil