AT82.02

DATA MODELING AND MANAGEMENT

UNIT 1-6: RELATIONAL DB DESIGN AND NORMALIZATOIN

CHUTIPORN ANUTARIYA (CHUTI AT AIT DOT AC DOT TH)

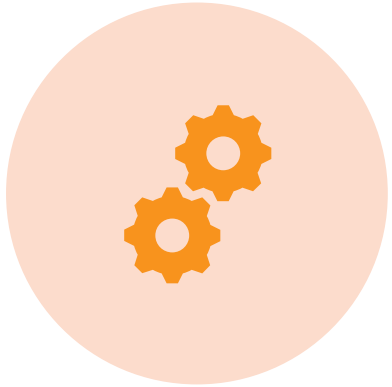
Co-funded by the Erasmus+ Programme of the European Union

DS&AI

Types of Database Design Process

FROM EXISTING DATA

NEW SYSTEMS DEVELOPMENT



DATABASE REDESIGN

Types of Database Design Process

From existing data

- Analyze spreadsheets and other data tables
- Extract data from other databases
- Design using normalization principles

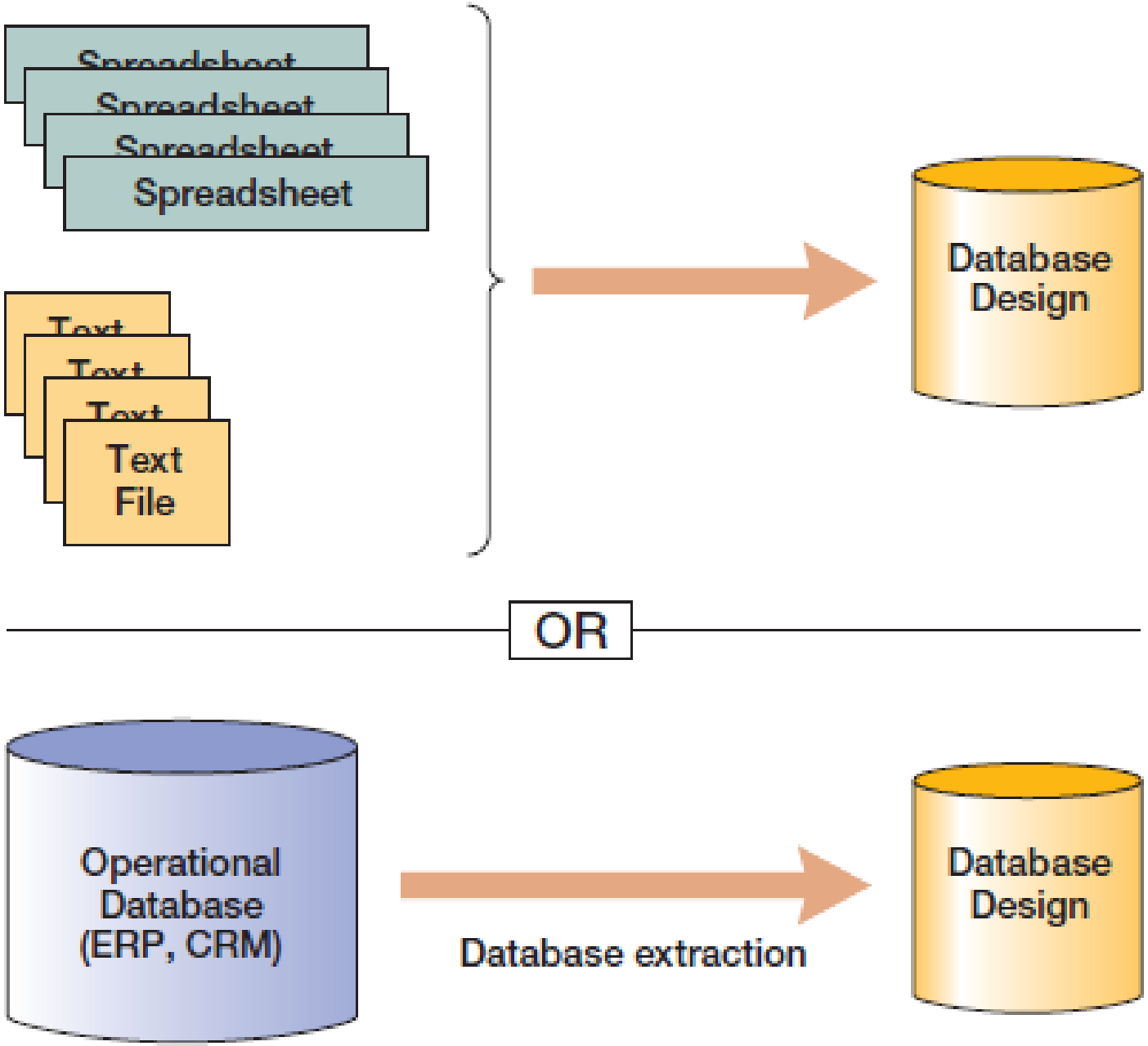
New systems development

- Create data model from application requirements
- Transform data model into database design

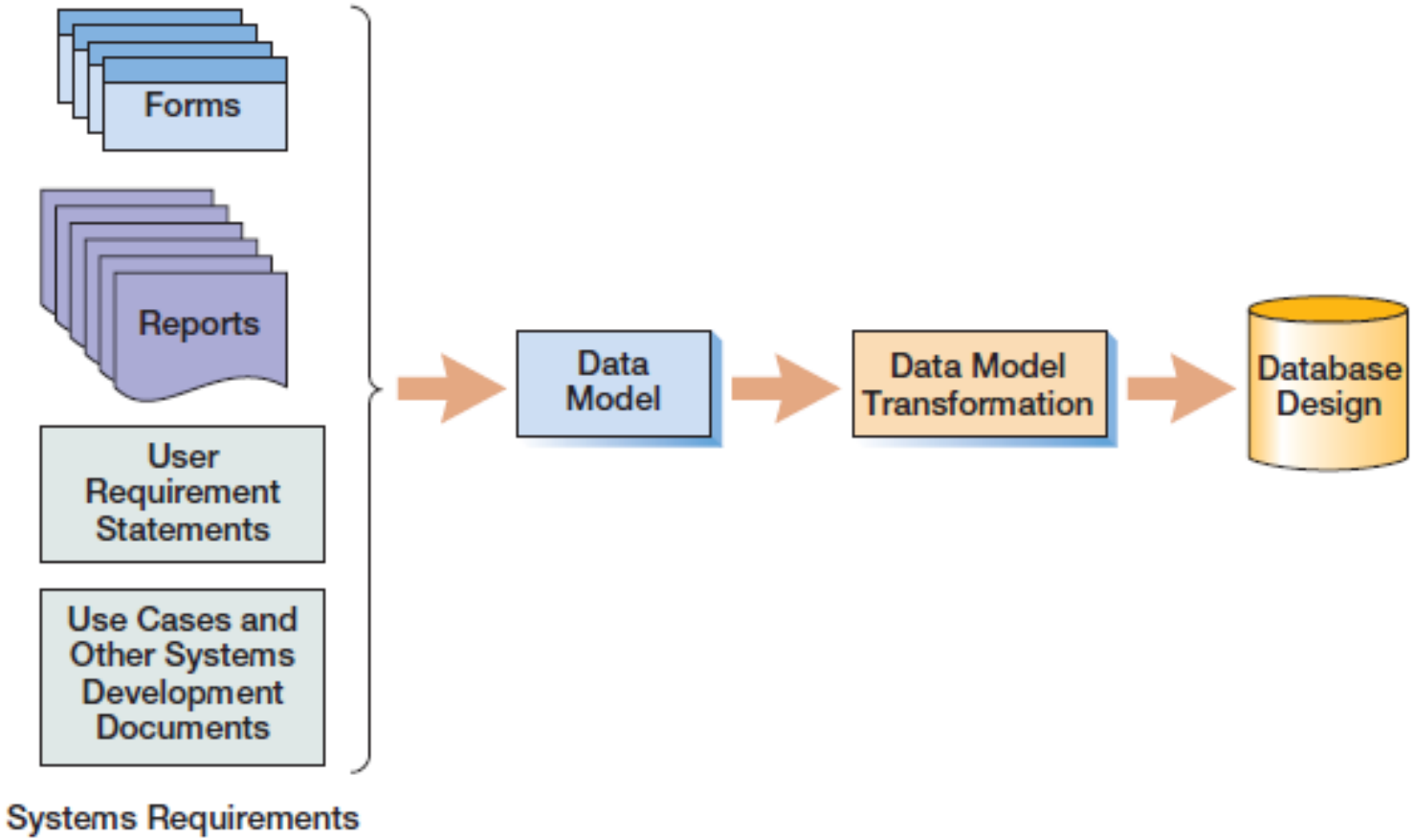
Database redesign

- Migrate databases to newer databases
- Integrate two or more databases
- Reverse engineer and design new databases using normalization principles and data model transformation

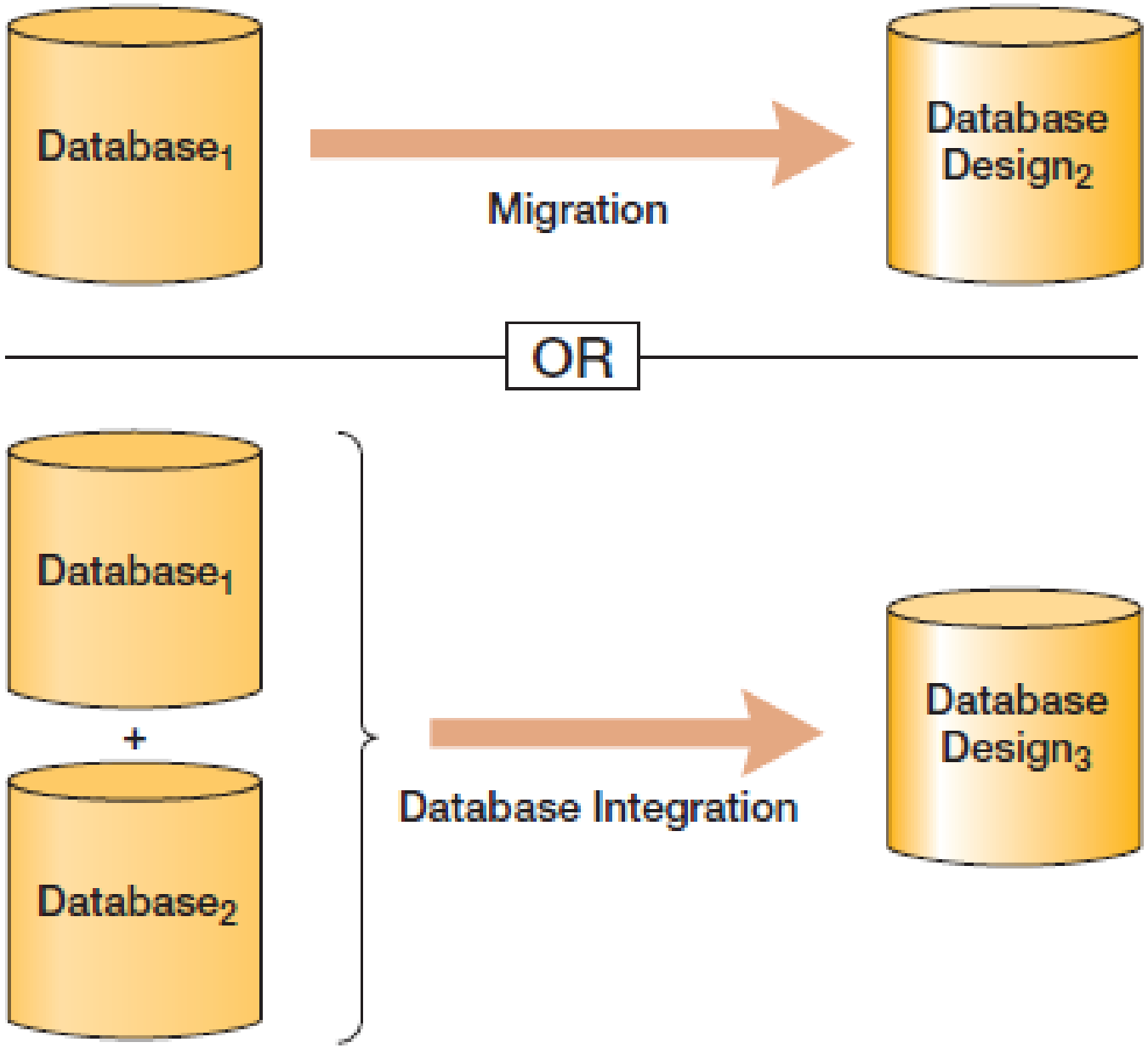
Database Originating from Existing Data



DB Originating from New System Development



Database Originating from DB Redesign



Informal Design Guidelines for Relation Schemas

Introduction

Levels at which we can discuss *goodness* of relation schemas

- Logical (or conceptual) level
- Implementation (or physical storage) level

Approaches to database design:

- Bottom-up or
- Top-down

Informal Design Guidelines for Relation Schemas

Measures of quality

- Making sure attribute semantics are clear
- Reducing redundant information in tuples
- Reducing NULL values in tuples
- Disallowing possibility of generating spurious tuples

Imparting Clear Semantics to Attributes in Relations

Semantics of a relation

- Meaning resulting from interpretation of attribute values in a tuple

Easier to explain semantics of relation

- Indicates better schema design

Figure 15.1
A simplified COMPANY relational database schema.

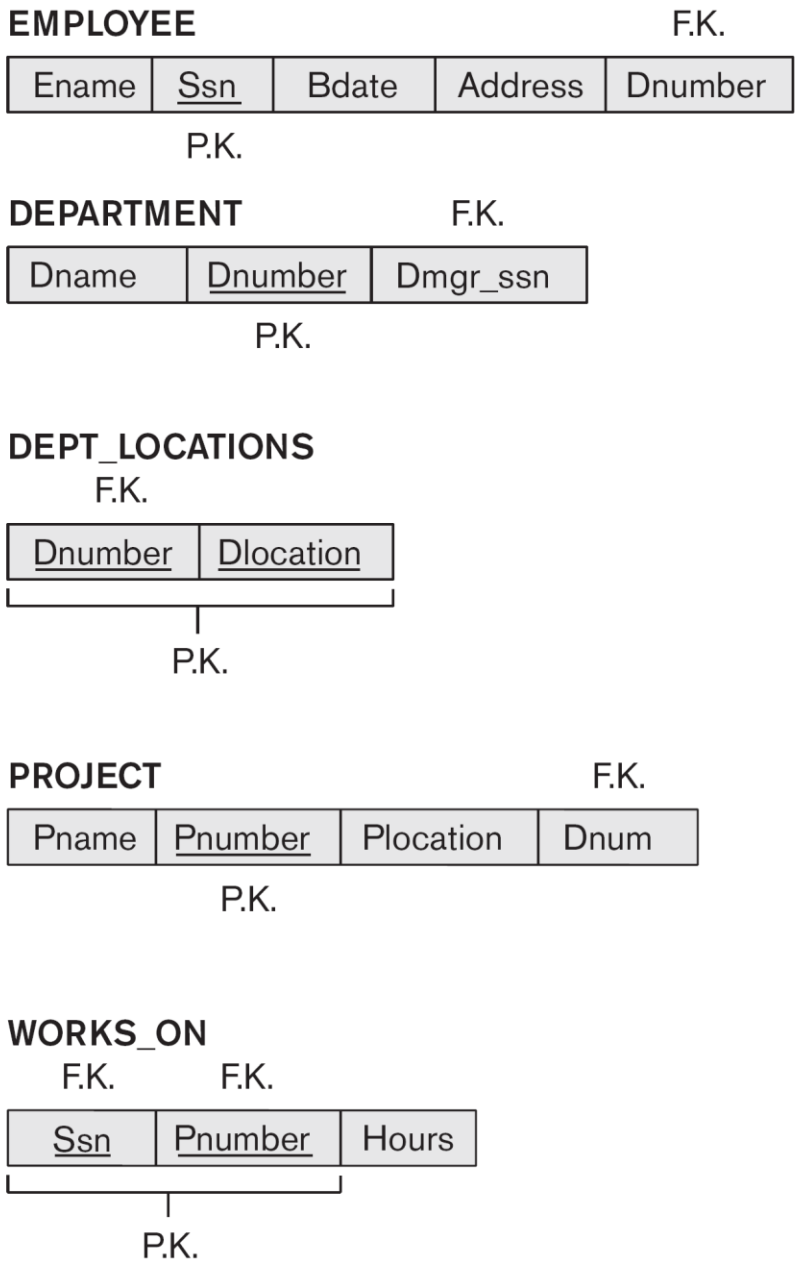


Figure 15.2
Sample database state for the relational database schema in Figure 15.1.

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	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

Continues...

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

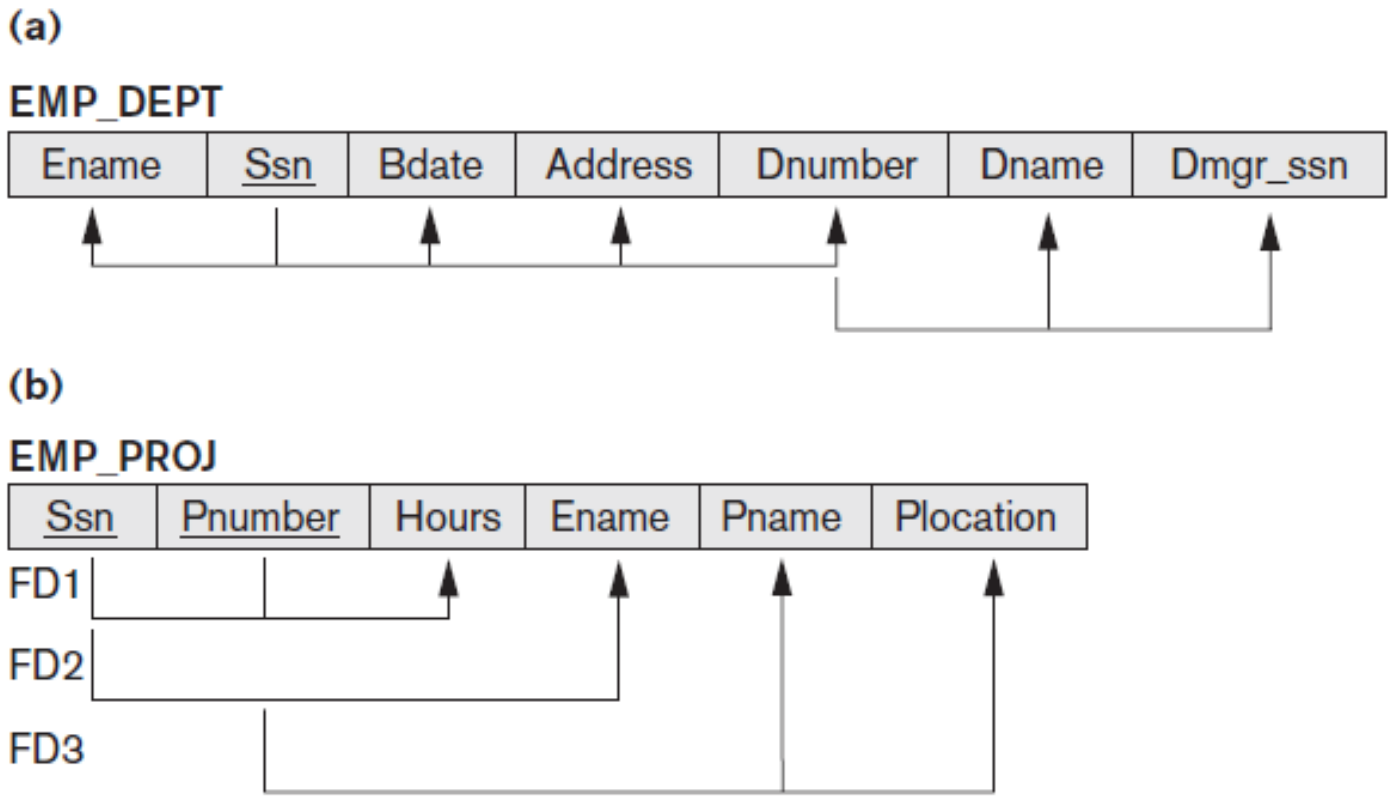
PROJECT

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

Guideline 1

Design	Design relation schema so that it is easy to explain its meaning
Do not combine	Do not combine attributes from multiple entity types and relationship types into a single relation

Figure 15.3
Two relation schemas suffering from update anomalies. (a) EMP_DEPT and (b) EMP_PROJ.



Example of violating Guideline 1

Informal Design Guidelines for Relation Schemas

Measures of quality

- Making sure attribute semantics are clear
- Reducing redundant information in tuples
- Reducing NULL values in tuples
- Disallowing possibility of generating spurious tuples

Redundant Information in Tuples and Anomalies

Grouping attributes into relation schemas

- Significant effect on storage space

Storing natural joins of base relations leads to data anomalies

Types of data anomalies:

- Insertion
- Deletion
- Update

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

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

Figure 15.4
Sample states for EMP_DEPT and EMP_PROJ resulting from applying NATURAL JOIN to the relations in Figure 15.2. These may be stored as base relations for performance reasons.

Data Anomalies

Update anomaly

- Can we modify the Dname of the 1st tuple of EMP_DEPT?

Insertion anomaly

- What if we want to insert a new department but does not have an employee in the department yet.

Deletion anomaly

- If we delete the employee Borg, James E., we lose information about the department 5!

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

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

Figure 15.4
Sample states for EMP_DEPT and EMP_PROJ resulting from applying NATURAL JOIN to the relations in Figure 15.2. These may be stored as base relations for performance reasons.

Guideline 2

Design base relation schemas so that no data anomalies are present in the relations

If any anomalies are present:

- Note them clearly
- Make sure that the programs that update the database will operate correctly

Informal Design Guidelines for Relation Schemas

Measures of quality

- Making sure attribute semantics are clear
- Reducing redundant information in tuples
- Reducing NULL values in tuples
- Disallowing possibility of generating spurious tuples

NULL Values in Tuples

May group many attributes together into a “fat” relation

- Can end up with many NULLs

Problems with NULLs

- Wasted storage space
- Problems understanding meaning

Guideline 3

Avoid placing attributes in a base relation whose values may frequently be NULL

If NULLs are unavoidable:

Make sure that they apply in exceptional cases only, not to a majority of tuples

Informal Design Guidelines for Relation Schemas

Measures of quality

- Making sure attribute semantics are clear
- Reducing redundant information in tuples
- Reducing NULL values in tuples
- Disallowing possibility of generating spurious tuples

Generation of Spurious Tuples

Figure 15.5(a)

- Relation schemas EMP_LOCS and EMP_PROJ1

NATURAL JOIN

- Result produces many more tuples than the original set of tuples in EMP_PROJ
- Called **spurious tuples**
- Represent spurious information that is not valid

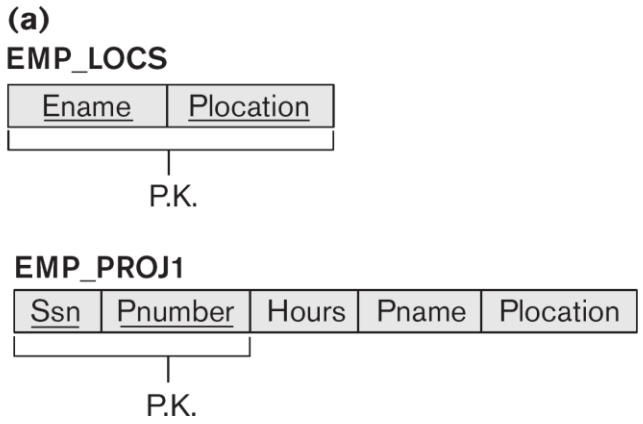


Figure 15.5
Particularly poor design for the EMP_PROJ relation in Figure 15.3(b). (a) The two relation schemas EMP_LOCS and EMP_PROJ1. (b) The result of projecting the extension of EMP_PROJ from Figure 15.4 onto the relations EMP_LOCS and EMP_PROJ1.

(b)

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_PROJ1

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

Example: Spurious Tuples

- Result of applying NATURAL JOIN to EMP_LOCS and EMP_PROJ1
- 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.

*
*
*

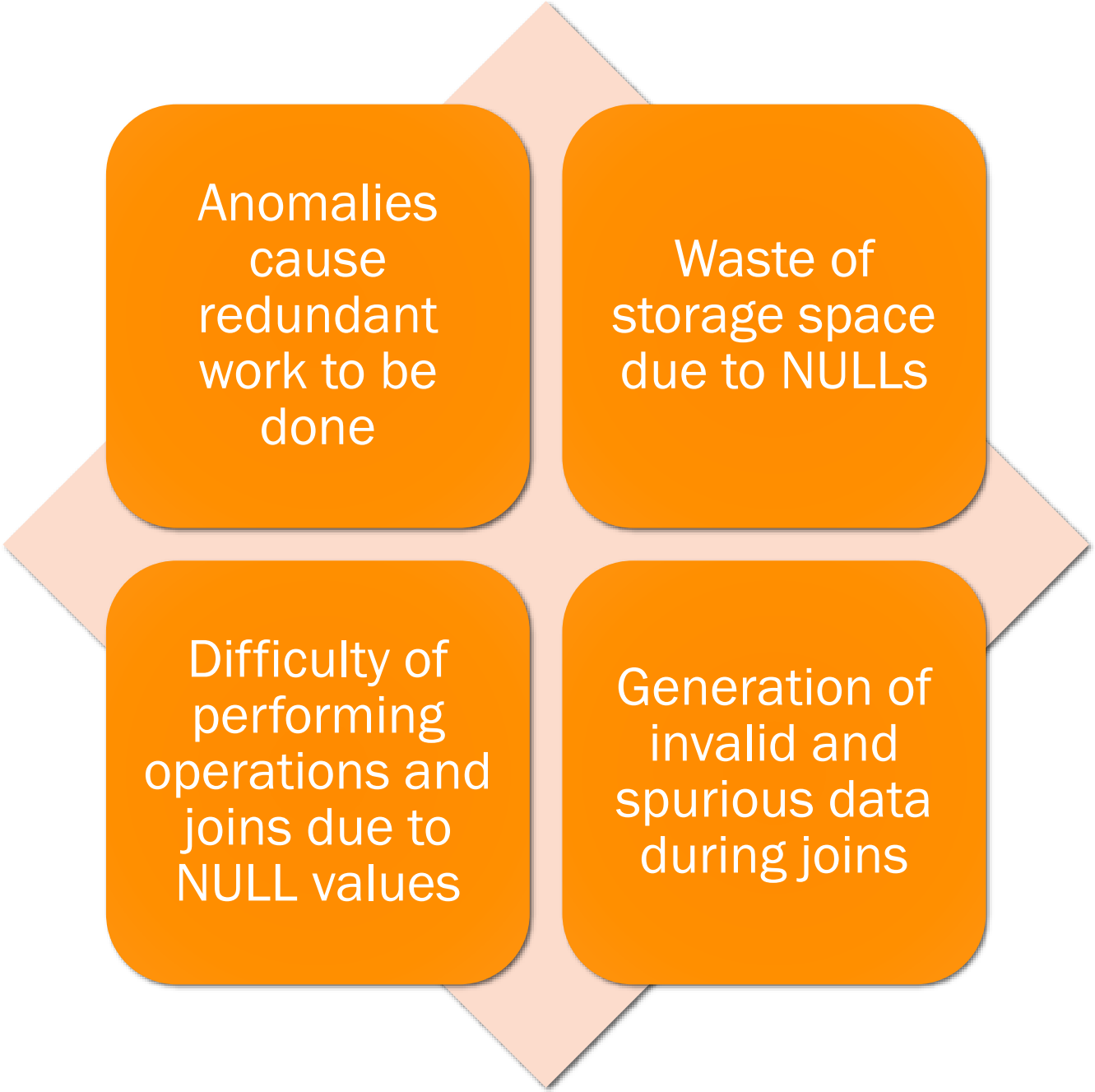
Guideline 4

Design relation schemas to be joined with equality conditions on attributes that are appropriately related

- Guarantees that no spurious tuples are generated

Avoid relations that contain matching attributes that are not (foreign key, primary key) combinations

Summary and Discussion of Design Guidelines



Basic Functional Dependency and Normal Forms

Review Questions

Discuss insertion, deletion and modification anomalies. Why are they considered bad? Illustrate with examples.

State the informal guidelines for relation schema design that we discussed. Illustrate how violation of these guidelines may be harmful.

Normalization

Process for evaluating and correcting table structures to minimize data redundancies

- helps eliminate data anomalies

Works through a series of stages called normal forms:

- Normal form (1NF)
- Second normal form (2NF)
- Third normal form (3NF)
- Boyce-Codd Normal Form

2NF is better than 1NF; 3NF is better than 2NF

For most business database design purposes, BCNF is highest we need to go in the normalization process

Highest level of normalization is not always most desirable

The Evils of Redundancy

Redundancy is at the root of several problems associated with relational schemas:

- redundant storage, insert/delete/update anomalies

Integrity constraints, in particular *functional dependencies*, can be used to identify schemas with such problems and to suggest refinements.

Main refinement technique: *decomposition*

- replacing ABCD with, say, AB and BCD, or ACD and ABD.

Decomposition should be used judiciously

Functional Dependencies (FDs)

$X \rightarrow Y$ means

- Given any two tuples in *r*, if the X values are the same, then the Y values must also be the same. (but not vice versa)
- That is, we cannot have two tuples with the same X value but different Y values.

Can read "$\rightarrow$" as "determines"

- note that this symbol (and its meaning) is *different* than the one for logical implication!!!

Constraints on the set of legal relations.

Require that the value for a certain set of attributes determines uniquely the value for another set of attributes.

A functional dependency is a generalization of the notion of a *key.*

ER Model and Normalization

When an E-R diagram is carefully designed, identifying all entities correctly, the tables generated from the E-R diagram should not need further normalization.

However, in a real (imperfect) design, there can be functional dependencies from non-key attributes of an entity to other attributes of the entity

- Example: an *employee* entity with attributes *department_name* and *building*, and a functional dependency *department_name* $\rightarrow$ *building*
- Good design would have made department an entity

Denormalization for Performance

May want to use non-normalized schema for performance.

Pros and Cons?

- Faster lookup
- Data anamalies
- Extra space and extra execution time for updates
- Extra coding work for programmer and possibility of error in extra code

Thank you.

Exit Slip:
Discuss 3 important things / concepts we have learned today.