Chapter 7

Normalization of Relational Tables

From Whence Come Databases?

- Databases arise from three sources:
  - Existing data
  - Development of new information systems ("green field")
  - Redesign of existing information systems ("brown field")
- So far, we've approached database design from the "green field" standpoint
- Today, from the "existing data" standpoint

Scenario

- Someone sends you a spreadsheet of data and asks you to do some work with it
- You realize that putting the data in a database would facilitate the work
- Should you:
  - Keep the original structure?
  - Reorganize the data?

Outline

- Modification anomalies
- Functional dependencies
- Normal forms
- Practical concerns

Modification Anomalies

- Unexpected side effect
- Insert, modify, and delete more data than desired
- Caused by excessive redundancies in poorly designed databases

Big University Database Table

StdSSN	StdClass	OfferNo	OffYear	EnrGrade	CourseNo	CrsDesc
S1	JUN	O1	2006	3.5	C1	DB
S1	JUN	O2	2006	3.3	C2	VB
S2	SR	O3	2006	3.1	C3	OO
S2	SR	O2	2006	3.4	C2	VB

Modification Anomaly Examples

- Insertion
  - Insert more column data than desired
  - Must know student number and offering number to insert a new course
- Update
  - Change multiple rows to change one fact
  - Must change two rows to change student class of student S1
- Deletion
  - Deleting a row causes other facts to disappear
  - Deleting enrollment of student S2 in offering O3 causes loss of information about offering O3 and course C3

Identifying Bad Design

- The design of the “Big University Database Table” feels wrong
  - Undesirable data redundancy
- Is there a way to characterize the design problems precisely?
- Yes: Using
  - Functional Dependencies
  - Normal Forms

Consider Two Tables

- Course

CourseNo	CrsDesc	CrsUnits
FIN300	FUNDAMENTALS OF FINANCE	4
FIN450	PRINCIPLES OF INVESTMENTS	4
FIN480	CORPORATE FINANCE	4
…		

- Offering

OfferNo	CourseNo	OffTerm	OffYear	OffLocation	OffTime	FacSSN	OffDays
1111	IS320	SUMMER	2010	BLM302	10:30 AM		MW
1234	IS320	FALL	2009	BLM302	10:30 AM	098-76-5432	MW
2222	IS460	SUMMER	2009	BLM412	1:30 PM		TTH
3333	IS320	SPRING	2010	BLM214	8:30 AM	098-76-5432	MW

Suppose we combine them?

CourseOffering

OfferNo	CourseNo	OffTerm	OffYear	OffLocation	OffTime	FacSSN	OffDays	CrsDesc	CrsUnits
5555	FIN300	WINTER	2010	BLM207	8:30 AM	765-43-2109	MW	FUNDAMENTALS OF FINANCE	3
6666	FIN450	WINTER	2010	BLM212	10:30 AM	987-65-4321	TTH	PRINCIPLES OF INVESTMENTS	3
7777	FIN480	SPRING	2010	BLM305	1:30 PM	765-43-2109	MW	CORPORATE FINANCE	3
1111	IS320	SUMMER	2010	BLM302	10:30 AM		MW	FUNDAMENTALS OF BUS PROGRAMMING	4
1234	IS320	FALL	2009	BLM302	10:30 AM	098-76-5432	MW	FUNDAMENTALS OF BUS PROGRAMMING	4
3333	IS320	SPRING	2010	BLM214	8:30 AM	098-76-5432	MW	FUNDAMENTALS OF BUS PROGRAMMING	4

- IS320 is offered multiple times
- Need a way to express a rule that says “All offerings of course X must have the same CrsDesc and CrsUnits”
- We can do that by writing **CourseNo → CrsDesc** and **CourseNo → CrsUnits**

Functional Dependencies

- A functional dependency is a relationship between two or more columns in a table
- Notation: **Col1 → Col2**
- Requires that records in which Col1 have duplicate values must also have duplicate values in Col2

Functional Dependencies

- **Col1 → Col2**
- Col1 is the **determinant**
- Col2 is the **dependent**
- In English:
  - “Col1 functionally determines Col2”
  - “Col2 is functionally dependent on Col1”

FD Definition

- X $\rightarrow$ Y
- X and Y may be 1 or more columns
  - Example: **CourseNo $\rightarrow$ CrsDesc, CrsUnits**
- X "functionally determines" Y
  - For a given X value, there is a single Y value

FDs in Data

StdSSN	**StdClass**	**OfferNo**	**OffYear**	**EnrGrade**	**CourseNo**	**CrsDesc**
S1	JUN	O1	2006	3.5	C1	DB
S1	JUN	O2	2006	3.3	C2	VB
S2	JUN	O3	2006	3.1	C3	OO
S2	JUN	O2	2006	3.4	C2	VB

- Prove non existence (but not existence) by looking at data
- If two rows have the same X value but a different Y value, $X \not\rightarrow Y$

Class Exercise

- Find the Likely Functional Dependencies

StdSSN	StdName	OfferNo	CourseNc	CrsDesc	EnrGrade
123-45-6789	HOMER WELLS	5555	FIN300	FUNDAMENTALS OF FINANCE	3.20
123-45-6789	HOMER WELLS	4321	IS320	FUNDAMENTALS OF BUSINESS PROGRAMMING	3.50
123-45-6789	HOMER WELLS	1234	IS320	FUNDAMENTALS OF BUSINESS PROGRAMMING	3.30
123-45-6789	HOMER WELLS	5678	IS480	FUNDAMENTALS OF DATABASE MANAGEMENT	3.20
123-45-6789	HOMER WELLS	5679	IS480	FUNDAMENTALS OF DATABASE MANAGEMENT	2.00
124-56-7890	BOB NORBERT	5555	FIN300	FUNDAMENTALS OF FINANCE	2.70
124-56-7890	BOB NORBERT	4321	IS320	FUNDAMENTALS OF BUSINESS PROGRAMMING	3.20
124-56-7890	BOB NORBERT	9876	IS460	SYSTEMS ANALYSIS	3.50
124-56-7890	BOB NORBERT	5679	IS480	FUNDAMENTALS OF DATABASE MANAGEMENT	3.70
234-56-7890	CANDY KENDALL	6666	FIN450	PRINCIPLES OF INVESTMENTS	3.10
234-56-7890	CANDY KENDALL	1234	IS320	FUNDAMENTALS OF BUSINESS PROGRAMMING	3.50
234-56-7890	CANDY KENDALL	9876	IS460	SYSTEMS ANALYSIS	3.20
234-56-7890	CANDY KENDALL	5678	IS480	FUNDAMENTALS OF DATABASE MANAGEMENT	2.80

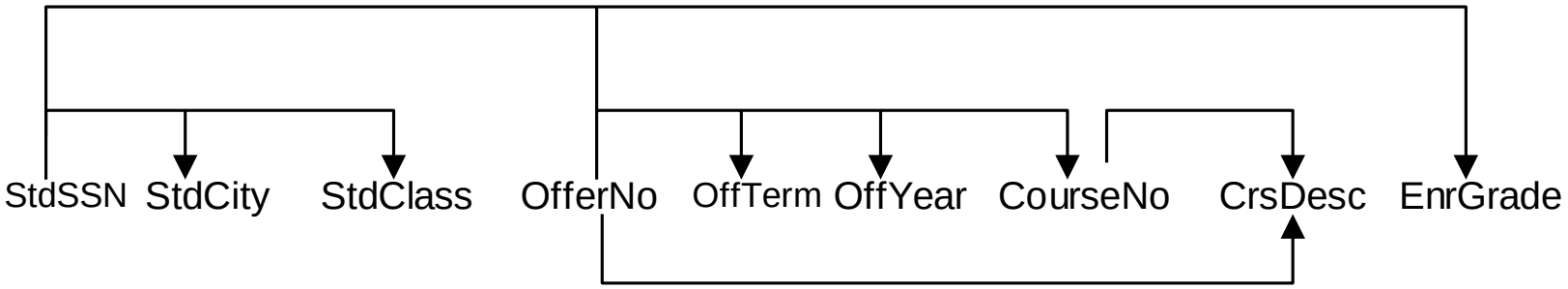
Answers

- StdSSN $\rightarrow$ StdName
- OfferNo $\rightarrow$ CourseNo
- CourseNo $\rightarrow$ CrsDesc
- StdSSN, OfferNo $\rightarrow$ EnrGrade

FD's and Keys

- A candidate or primary key is the determinant for all of the other columns in its table
- The dependency X $\rightarrow$ Y often implies that, in a properly designed database, Y should be a column in a table whose primary key is X

FD Diagrams and Lists



StdSSN → StdCity, StdClass

OfferNo → OffTerm, OffYear, CourseNo, CrsDesc

CourseNo → CrsDesc

StdSSN, OfferNo → EnrGrade

Identifying FDs

- Look for:
  - Statements about uniqueness
  - PKs and CKs resulting from ERD conversion
  - 1-M relationship: FD from child to parent
- Problematic FD's:
  - LHS is not a PK or CK in a converted table
  - LHS is subset of a compound primary or candidate key
- Ensure minimality of LHS

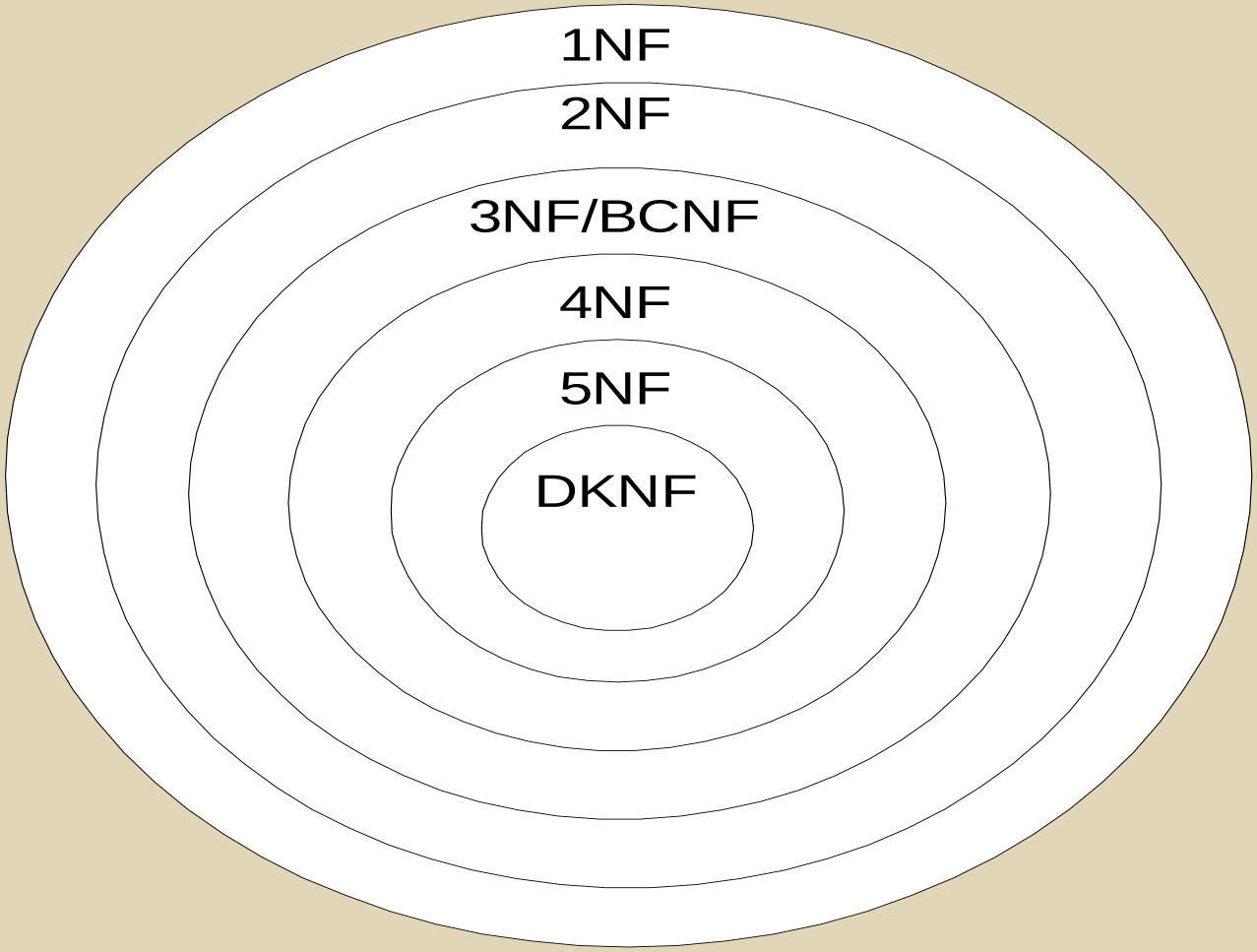
Normalization

- Process of refining database design by removing unwanted redundancies
- Apply normal forms
  - Identify FDs
  - Determine whether FDs meet normal form
  - Split the table to meet the normal form if there is a violation

Normal Forms

- Normal Forms – rules for database design based on FD concept
- Invented by Codd, father of relational database

Relationships of Normal Forms



First Normal Form (1NF)

- Basic rules for valid tables:
  - Must have primary key
  - No repeating groups or multivalued attributes
- The table below is not in 1NF
  - Telephone Number is a multivalued attribute

Customer

Customer ID	First Name	Surname	Telephone Number
123	Pooja	Singh	555-861-2025, 192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53; 182-929-2929
789	John	Doe	555-808-9633

From: https://en.wikipedia.org/wiki/First_normal_form

1NF Continued

- The table below is not in 1NF
  - Telephone1 and Telephone2 form a repeating group

Customer

Customer ID	First Name	Surname	Telephone Number1	Telephone Number2
123	Pooja	Singh	555-861-2025	192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53	182-929-2929
789	John	Doe	555-808-9633	

Violating 1NF

An order is placed for multiple products

- Design #1

OrderID	CustID	OrdDate	Products
1001	1	2017-02-15	1,25,16,32
1002	2	2017-02-18	35,16

- Design #2

OrderID	CustID	OrdDate	Prod1	Qty1	Prod2	Qty2
1001	1	2017-02-15	1	3	25	3
1002	2	2017-02-18	35	1	16	20

Revising to 1NF

OrderID	CustID	OrdDate
101	1	2017-02-15
102	2	2017-02-18

OrderItemID	OrderID	ProdID	Qty
1001	101	1	10
1002	101	25	3
1003	102	35	5
1004	102	16	10

Repeating Groups

- Why avoid repeating groups?
- Consider writing queries to solve the following:
  - Which customer has phone number X?
  - Display a list of all phone numbers with duplicates removed

Eliminating Repeating Groups / Multivalued Attributes

- How would you rework the telephone number example?

2NF and 3NF

- Outdated – superseded by BCNF
- We will not discuss

Boyce-Codd Normal Form (BCNF)

- Rule: Every determinant in a table X must be a candidate key in table X.

Violating BCNF

- ProdID → ProdDescription
- Why is this table design undesirable?

OrderItemID	OrderID	ProdID	ProdDescription	Qty
1001	101	1	Toothpaste	10
1002	101	25	Hot dogs	3
1003	102	1	Toothpaste	5
1004	102	16	Ground Beef	10

Revising to BCNF

- Move ProdDescription to separate table, together with ProdID
- Leave ProdID in original table

OrderItemID	OrderID	ProdID	Qty
1001	101	1	10
1002	101	25	3
1003	102	1	5
1004	102	16	10

ProdID	ProdDescription
1	Toothpaste
16	Ground Beef
25	Hot dogs

How to put a table into BCNF

- Identify every Functional Dependency in the table
- Identify every candidate key
- If there is a F.D. that has a determinant that is not a candidate key:
  - Move columns of F.D. to a new table whose primary key is the determinant of the F.D.
  - Leave a copy of the determinant in the original table as a foreign key

BCNF Example

- See Big University Database Table
- Primary key: (OfferNo, StdSSN)
- Many violations of BCNF:
  - StdSSN → StdCity, StdClass
  - OfferNo → OffTerm, OffYear, CourseNo
  - CourseNo → CrsDesc

Normalization Practice

- Big Patient Table:
  - VisitNo, VisitDate, PatNo, PatAge, PatCity, PatZip, ProviderNo, ProviderSpecialty, Diagnosis
- FD's:
  - PatNo $\rightarrow$ PatAge, PatCity, PatZip
  - ProviderNo $\rightarrow$ ProviderSpecialty
  - VisitNo, ProviderNo $\rightarrow$ Diagnosis
  - PatZip $\rightarrow$ PatCity

Review

- Taking Stock

Multivalued Dependencies

- $X \rightarrow\rightarrow Y$
- Read "X multidetermines Y"
- Values of X uniquely determine a set of values Y (Y is "multivalued")
- Example: StdSSN $\rightarrow\rightarrow$ OfferNo

Enrollment

StdSSN	OfferNo
Fred	1001
Fred	1002
Rita	1001
Joe	1002
Joe	1003

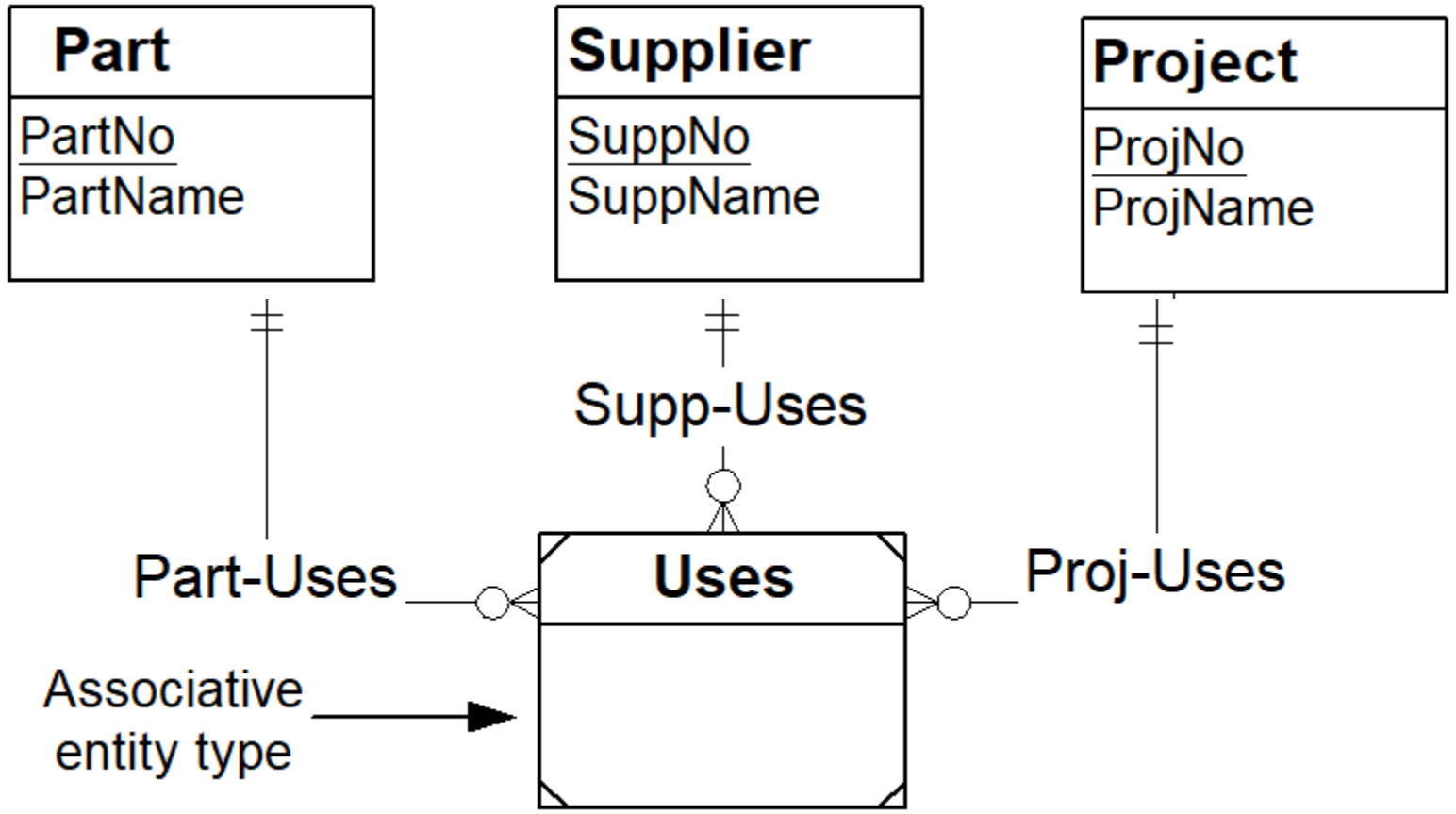
Fourth Normal Form (4NF)

- In a table containing a multivalued dependency, the MVD must be trivial
- A non-trivial MVD exists when a table contains a MVD, together with another multivalued attribute

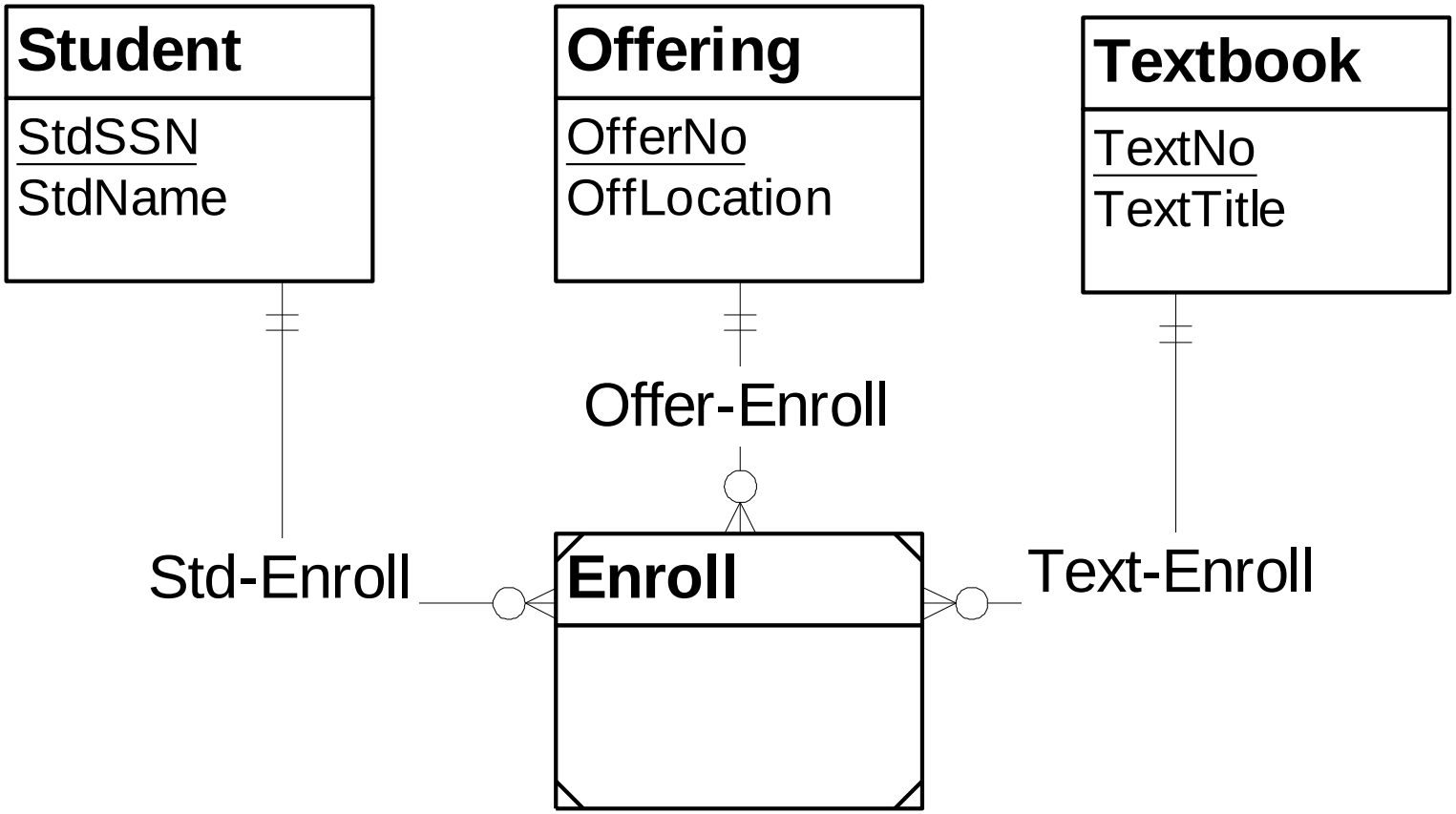
4NF Continued

- Common violations of 4NF: Inappropriate M-way relationships
- An M-way relationship that can be derived from binary relationships should be split into binary relationships

Review: Valid M-way Relationship



Relationship Independence Problem



Analysis

- StdSSN $\rightarrow\rightarrow$ OfferNo
- OfferNo $\rightarrow\rightarrow$ BookNo

Better

Enrollment

StdSSN	OfferNo
Fred	1001
Fred	1002
Rita	1001
Joe	1002
Joe	1003

CourseTextbooks

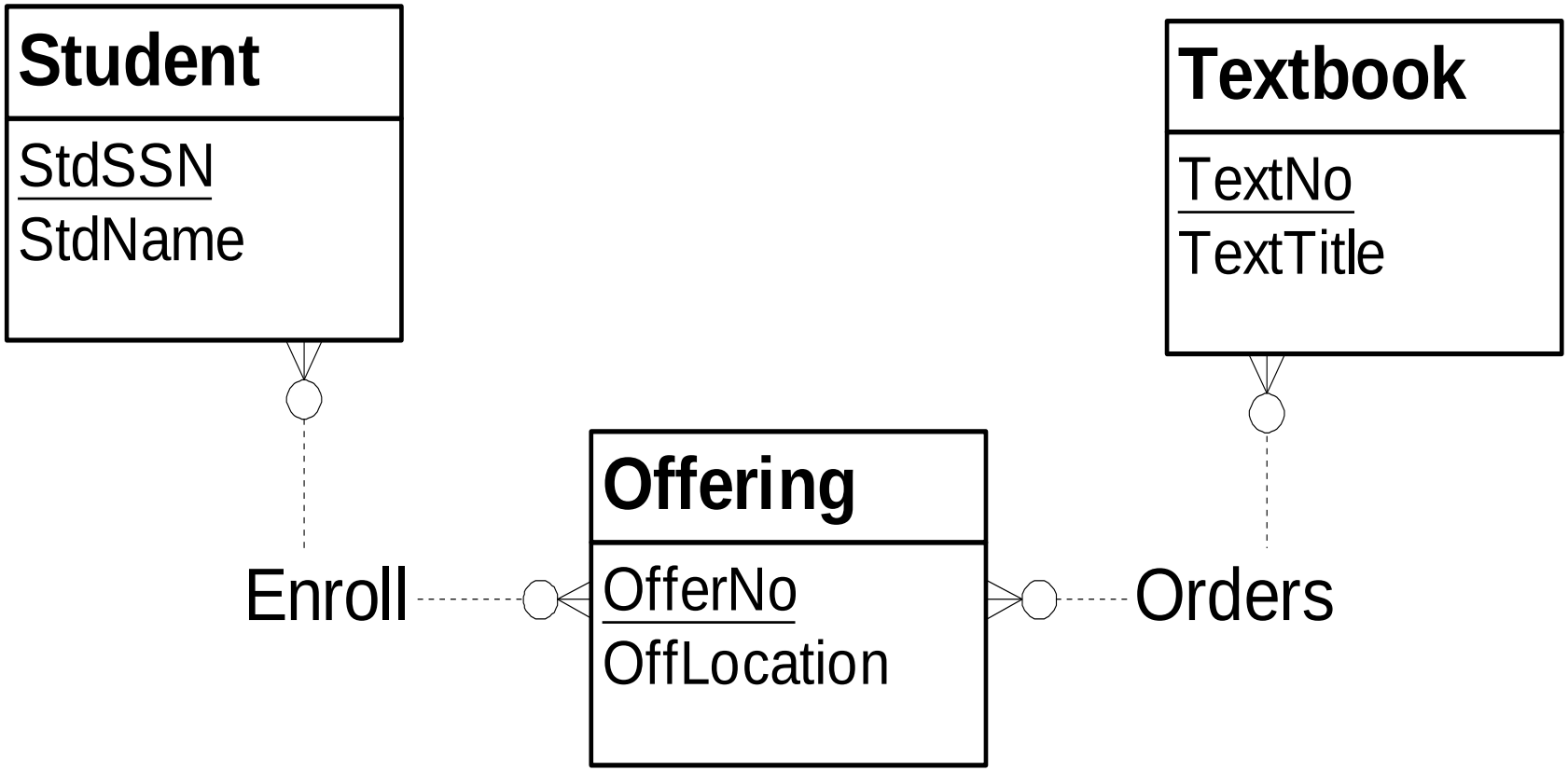
OfferNo	BookNo
1001	Comp1
1001	Writing1
1002	History1
1003	Piano1

Bad

BadEnroll

StdSSN	OfferNo	BookNo
Fred	1001	Comp1
Fred	1001	Writing1
Fred	1002	History1
Rita	1001	Comp1
Rita	1001	Writing1
Joe	1002	History1
Joe	1003	Piano1

Relationship Independence Solution



Higher Level Normal Forms

- 5NF and DKNF
  - We will not consider these
  - Deal with problems that rarely crop up in practice

Role of Normalization

- Refinement of New Database Design
  - Use after ERD
  - Apply to table design or ERD
- Creating Database from External Data
  - Excel file
  - CSV file

Summary

- Beware of unwanted redundancies
- FDs are important constraints
- Strive for BCNF
- Important tool of database development
- Focus on the normalization objective