

Dr Renata Borovica-Gajic  
David Eccles



# INFO90002 Database Systems

Lecture 6  
Normalization



- By the end of this lecture, you should be able to:
  - Define normalization
  - Explain and identify database anomalies
  - Define and identify functional dependencies
  - Normalize relations to:
    - 1st Normal Form (1NF)
    - 2nd Normal Form (2NF)
    - 3rd Normal Form (3NF)
    - Boyce-Codd Normal Form (BCNF)



- What happens if we don't normalize?



# What's wrong with the *organization* of data in this table?

Student ID#	Student Name	Campus Address	Degree	Phone	Subject ID	Subject Title	Lecturer Name	Lecturer Office	Lecturer Phone	Sem.	Grade
A121	Joy Egbert	166 Grattan Street	B.Com.	555-7771	ACC101	Accounting	Davern	T240C	8344-1846	1-11	H1
A121	Joy Egbert	166 Grattan Street	B.Com.	555-7771	ECO101	Economics	Smyth	T240F	8344-1868	1-11	H2B
A121	Joy Egbert	166 Grattan Street	B.Com.	555-7771	ECO104	Quant. M.	Collier	T240D	8344-5716	1-11	H2B
A121	Joy Egbert	166 Grattan Street	B.Com.	555-7771	FIN101	Finance.	James	T240D	8344-5275	1-11	H2A
A121	Joy Egbert	166 Grattan Street	B.Com.	555-7771	ACC103	Processes	Wise	T240E	8344-5309	1-11	H3
A123	Larry Mueller	302 Royal Parade	B.Com.	555-1235	ACC101	Accounting	Davern	T240C	8344-1846	1-11	H1
A123	Larry Mueller	302 Royal Parade	B.Com.	555-1235	ECO101	Economics	Smyth	T240F	8344-1868	1-11	H2B
A123	Larry Mueller	302 Royal Parade	B.Com.	555-1235	ECO104	Quant. M.	Collier	T240D	8344-5716	1-11	H2A
A123	Larry Mueller	302 Royal Parade	B.Com.	555-1235	FIN101	Finance.	James	T240D	8344-5275	1-11	H3
A124	Mike Guon	224 Swanston St.	B.Eco.	555-2214	ACC101	Accounting	Davern	T240C	8344-1846	1-11	H2A
A124	Mike Guon	224 Swanston St.	B.Eco.	555-2214	ECO101	Economics	Smyth	T240F	8344-1868	1-11	H2A
A124	Mike Guon	224 Swanston St.	B.Eco.	555-2214	ECO104	Quant. M.	Collier	T240D	8344-5716	1-11	H2B
A124	Mike Guon	224 Swanston St.	B.Eco.	555-2214	ACC103	Processes	Wise	T240E	8344-5309	1-11	H2B
A126	Jackie Judson	85 Barry Street	B.Eco.	555-1245	ACC101	Accounting	Davern	T240C	8344-1846	1-11	H1
A126	Jackie Judson	85 Barry Street	B.Eco.	555-1245	ECO101	Economics	Smyth	T240F	8344-1868	1-11	H2B
A126	Jackie Judson	85 Barry Street	B.Eco.	555-1245	ECO104	Quant. M.	Collier	T240D	8344-5716	1-11	H2B
A126	Jackie Judson	85 Barry Street	B.Eco.	555-1245	ACC103	Processes	Wise	T240E	8344-5309	1-11	H2A
...	...	...	...	...	...	...	...	...	...	...	...



- Consider the following denormalized table (relation) :

Student-ID	Course-ID	Fee
130	C200	75
200	C300	100
250	C200	75
425	C400	150
500	C300	100
575	C500	50
...	...	...

- Insertion Anomaly:** A new course cannot be added until at least one student has enrolled (which comes first student or course?)
- Deletion Anomaly:** If student 425 withdraws, we lose all record of course C400 and its fee!
- Update Anomaly:** If the fee for course C200 changes, we have to change it in multiple records (rows), else the data will be inconsistent.

- A technique used to remove undesired redundancy from databases (Break one large table into several smaller tables).

A relation is normalized if  
all determinants are  
candidate keys

## How do we normalise?



## Invoice example

## Bill To

John  
synex Inc  
128 AA Juanita Ave  
Glendora  
CA 91740 US

## Ship To

John  
synex Inc  
128 AA Juanita Ave  
Glendora  
CA 91740 US

Date	14-Aug-2009	Order No		Sales Person	Charles Wooten
Shipping Date	13-Aug-2009	Shipping Terms		Terms	COD
ID	SKU / Description		Unit Price (USD)	Qty	Amount (USD)
PS.V860.005	AMD Athlon X2DC-7450, 2.4GHz/1GB/160GB/SMP-DVD/VB		580.00	6.00	3,480.00
PS.V880.037	PDC-E5300 - 2.6GHz/1GB/320GB/SMP-DVD/FDD/VB		645.00	4.00	2,580.00
LC.V890.002	LG 18.5" WLCD		230.00	10.00	2,300.00
HP.Q754.071	HP LaserJet 5200		1,103.00	1.00	1,103.00



Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Product ID	Product Name	Unit Price	Quantity	Amount	Sub Total
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	PSV880.006	AMD Athlon X2DC	580	6	3480	9463
					COD	PSV880.037	PDC E5300	645	4	2580	
					COD	LC.V890.002	LG 8.5" LCD	230	10	2300	
					COD	HPQ754.071	HP LaserJet 5200	1103	1	1103	
INV0013	15-Aug-09	Ann / ThisCo	123 Smith Street...	Mary Smith	COD	HP Q754.071	HP LaserJet 5200	1103	2	2206	3356
						LCV890.002	LG 8.5" LCD	230	5	1150	

This is not relational model



Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann / ThisCo	123 Smith Street...	Mary Smith	COD	3356	0	100	0

Product ID	Product Name	Unit Price	Quantity	Amount
PSV880.006	AMD Athlon X2DC	580	6	3480
PSV880.037	PDC E5300	645	4	2580
LC.V890.002	LG 8.5" LCD	230	10	2300
HPQ754.071	HP LaserJet 5200	1103	1	1103
HPQ754.071	HP LaserJet 5200	1103	2	2206
LCV890.002	LG 8.5" LCD	230	5	1150

Break into two  
But...  
How do we connect?



## Invoice example – Spreadsheet Format

Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann/This Co	123 Smith Street...	Mary Smith	COD	3356	0	100	0

Product ID	Product Name	Unit Price	Quantity	Amount	Invoice Number
PSV880.006	AMD Athlon X2DC	580	6	3480	INV0012
PSV880.037	PDC E5300	645	4	2580	INV0012
LC.V890.002	LG 8.5" LCD	230	10	2300	INV0012
HPQ754.071	HP LaserJet 5200	1103	1	1103	INV0012
HPQ754.071	HP LaserJet 5200	1103	2	2206	INV0013
LCV890.002	LG 8.5" LCD	230	5	1150	INV0013

Add FK



Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann / ThisCo	123 Smith Street...	Charles Wooten	COD	3356	0	100	0

Product ID	Product Name	Unit Price	Quantity	Amount	Invoice Number
PSV880.006	AMD Athlon X2DC	580	6	3480	INV0012
PSV880.037	PDC E5300	645	4	2580	INV0012
LC.V890.002	LG 8.5" LCD	230	10	2300	INV0012
HPQ754.071	HP LaserJet 5200	1103	1	1103	INV0012
HPQ754.071	HP LaserJet 5200	1103	2	2206	INV0013
LCV890.002	LG 8.5" LCD	230	5	1150	INV0013

This is about product

This is about Order (invoice)



Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann / ThisCo	123 Smith Street...	Charles Wooten	COD	3356	0	100	0

Product ID	Quantity	Amount	Invoice Number
PSV880.006	6	3480	INV0012
PSV880.037	4	2580	INV0012
LC.V890.002	10	2300	INV0012
HPQ754.071	1	1103	INV0012
HPQ754.071	2	2206	INV0013
LCV890.00	5	1150	INV0013

Break into two

Product ID	Product Name	Unit Price
PSV880.006	AMD Athlon X2DC	580
PSV880.037	PDC E5300	645
LC.V890.002	LG 8.5" LCD	230
HPQ754.071	HP LaserJet 5200	1103



# Invoice example – Spreadsheet Format

Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann / ThisCo	123 Smith Street...	Charles Wooten	COD	3356	0	100	0

Product ID	Quantity	Amount	Invoice Number
PSV880.006	6	3480	INV0012
PSV880.037	4	2580	INV0012
LC.V890.002	10	2300	INV0012
HPQ754.071	1	1103	INV0012
HPQ754.071	2	2206	INV0013
LCV890.00	5	1150	INV0013

What about amount?

Product ID	Product Name	Unit Price
PSV880.006	AMD Athlon X2DC	580
PSV880.037	PDC E5300	645
LC.V890.002	LG 8.5" LCD	230
HPQ754.071	HP LaserJet 5200	1103



Invoice Number	Date	Customer Name	Customer Address	Sales Person	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	Charles Wooten	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann / ThisCo	123 Smith Street...	Charles Wooten	COD	3356	0	100	0

What about sales person?

Could be derived

Product ID	Quantity	Invoice Number
PSV880.006	6	INV0012
PSV880.037	4	INV0012
LC.V890.002	10	INV0012
HPQ754.071	1	INV0012
HPQ754.071	2	INV0013
LCV890.00	5	INV0013

Product ID	Product Name	Unit Price
PSV880.006	AMD Athlon X2DC	580
PSV880.037	PDC E5300	645
LC.V890.002	LG 8.5" LCD	230
HPQ754.071	HP LaserJet 5200	1103



# Invoice example – Spreadsheet Format

Invoice Number	Date	Customer Name	Customer Address	Sales Person ID	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	John / Synex	128 Juanita Ave...	1	COD	9463	0	780.70	0
INV0013	15-Aug-09	Ann / This co	123 Smith Street...	1	COD	3356	0	100	0

What about customer?

Product ID	Quantity	Invoice Number
PSV880.006	6	INV0012
PSV880.037	4	INV0012
LC.V890.002	10	INV0012
HPQ754.071	1	INV0012
HPQ754.071	2	INV0013
LCV890.00	5	INV0013

Sales Person ID	Sales Person
1	Charles Wooten
2	Mary James

Product ID	Product Name	Unit Price
PSV880.006	AMD Athlon X2DC	580
PSV880.037	PDC E5300	645
LC.V890.002	LG 8.5" LCD	230
HPQ754.071	HP LaserJet 5200	1103



Invoice Number	Date	Customer ID	Sales Person ID	Terms	Sub Total	Discount	Sales Tax	Shipping
INV0012	14-Aug-09	1	1	COD	9463	0	780.70	0
INV0013	15-Aug-09	2	1	COD	3356	0	100	0

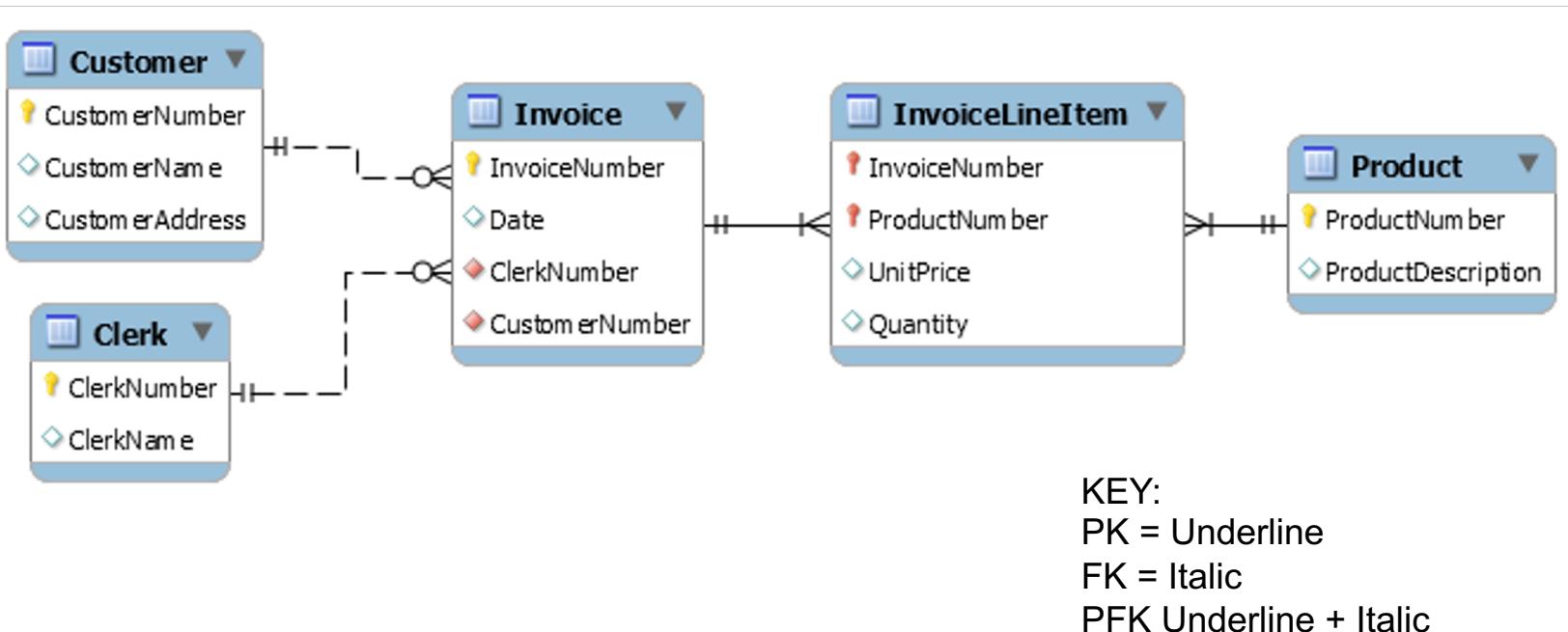
Product ID	Quantity	Invoice Number
PSV880.006	6	INV0012
PSV880.037	4	INV0012
LC.V890.002	10	INV0012
HPQ754.071	1	INV0012
HPQ754.071	2	INV0013
LCV890.00	5	INV0013

Sales Person ID	Sales Person
1	Charles Wooten
2	Mary James

Customer ID	Customer Name	Customer Address
1	John / Synex	128 Juanita Ave...
2	Ann / ThisCo	123 Smith Street...

Product ID	Product Name	Unit Price
PSV880.006	AMD Athlon X2DC	580
PSV880.037	PDC E5300	645
LC.V890.002	LG 8.5" LCD	230
HPQ754.071	HP LaserJet 5200	1103

- We can name the relations now
  - Customer (CustomerNumber, CustomerName, CustomerAddress)
  - Clerk (ClerkNumber, ClerkName)
  - Product (ProductNumber, ProductDescription)
  - Invoice (InvoiceNumber, Date, *CustomerNumber*, *ClerkNumber*)
  - InvoiceLineItem (InvoiceNumber, ProductNumber, UniPrice, Quantity)





- Now let's go back to theoretical concepts...



- A functional dependency concerns values of attributes in a relation
- A set of attributes **X determines** another set of attributes **Y** if each value of **X** is associated with only one value of **Y**
  - Written  $X \rightarrow Y$
  - **X determines Y** (If I know **X** then I also know **Y**)

- Emp#  Emp-name
- Emp#  Salary



- Determinants ( $X, Y \rightarrow Z$ )  $A(\underline{X}, \underline{Y}, Z, D)$ 
  - the attribute(s) on the left hand side of the arrow
- Key and Non-Key attributes
  - each attribute is either part of the primary key or it is not
- Partial functional dependency ( $Y \rightarrow Z$ )
  - a functional dependency of one or more non-key attributes upon part (but not all) of the primary key
- Transitive dependency ( $Z \rightarrow D$ )
  - a functional dependency between 2 (or more) non-key attributes



Functional dependencies can be identified using Armstrong's Axioms

$$A = (X_1, X_2, \dots, X_n) \text{ and } B = (Y_1, Y_2, \dots, Y_n)$$

1. Reflexivity:  $B \subseteq A \Rightarrow A \rightarrow B$

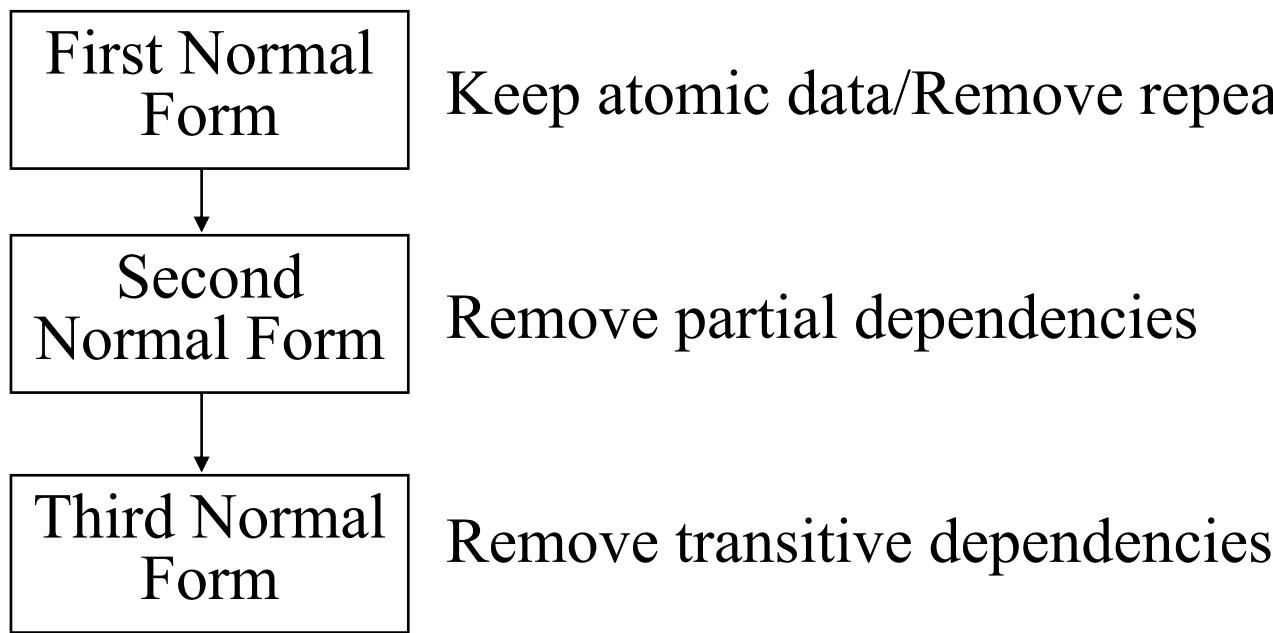
**Example:** Student\_ID, name  $\rightarrow$  name

2. Augmentation:  $A \rightarrow B \Rightarrow AC \rightarrow BC$

**Example:** Student\_ID  $\rightarrow$  name  $\Rightarrow$  Student\_ID, surname  $\rightarrow$  name, surname

3. Transitivity:  $A \rightarrow B$  and  $B \rightarrow C \Rightarrow A \rightarrow C$

**Example:** ID  $\rightarrow$  birthdate and birthdate  $\rightarrow$  age then ID  $\rightarrow$  age





## Remove Repeating Groups/Keep Atomic Data

- repeating groups of attributes cannot be represented in a flat, two dimensional table
- removing cells with multiple values (keep atomic data)  
Set of values

Example: Order-Item (Order#, Customer#, (Item#, Desc, Qty))

- Order-Item (Order#, Customer#, (Item#, Desc, Qty))



- Order-Item (Order#, Item#, Desc, Qty)
- Order (Order#, Customer#)

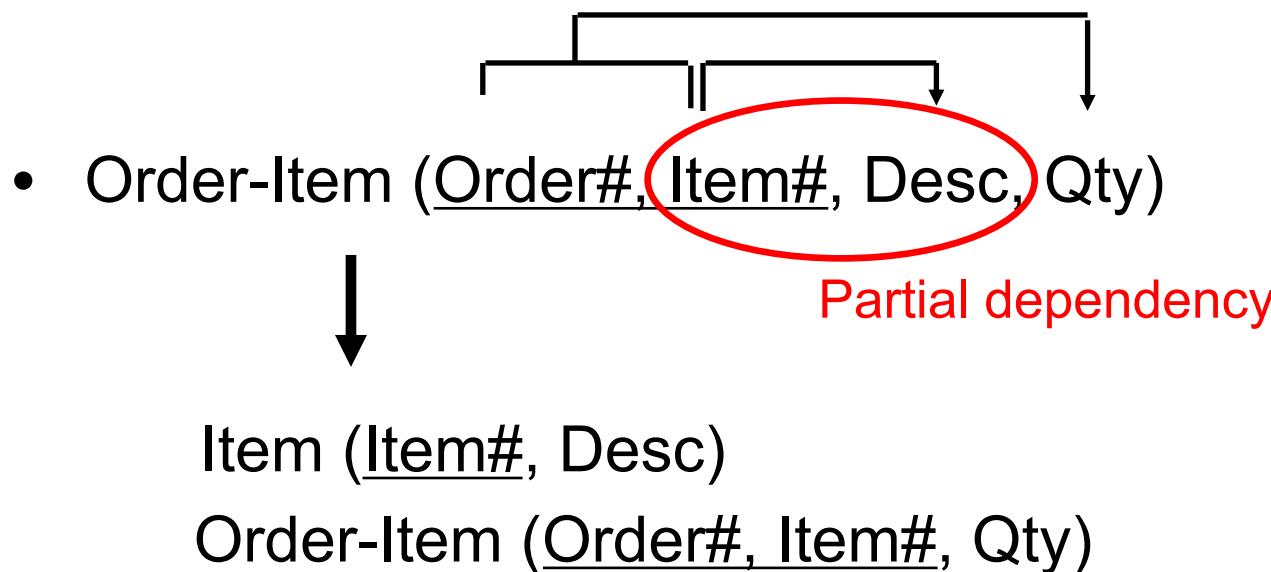
Break them into two  
Use PK/FK to connect



## Remove Partial Dependencies

a non-key attribute cannot be identified  
by part of a composite key

Example: Order-Item (Order#, Item#, Desc, Qty)





## Order-Item (Order#, Item#, Desc, Qty)

Order#	Item#	Desc	Qty
27	873	nut	2
28	402	bolt	1
28	873	nut	10
30	495	washer	50

- *UPDATE* change item desc in many places
- *DELETE* data for last item lost when last order for that item is deleted
- *INSERT* cannot add new item until it is ordered



## Order-Item

Order#	Item#	Qty
27	873	2
28	402	1
28	873	10
30	495	50

*delete last  
order for item,  
but item  
remains*

*add new item at any time*

Item#	Desc
873	nut
402	bolt
495	washer

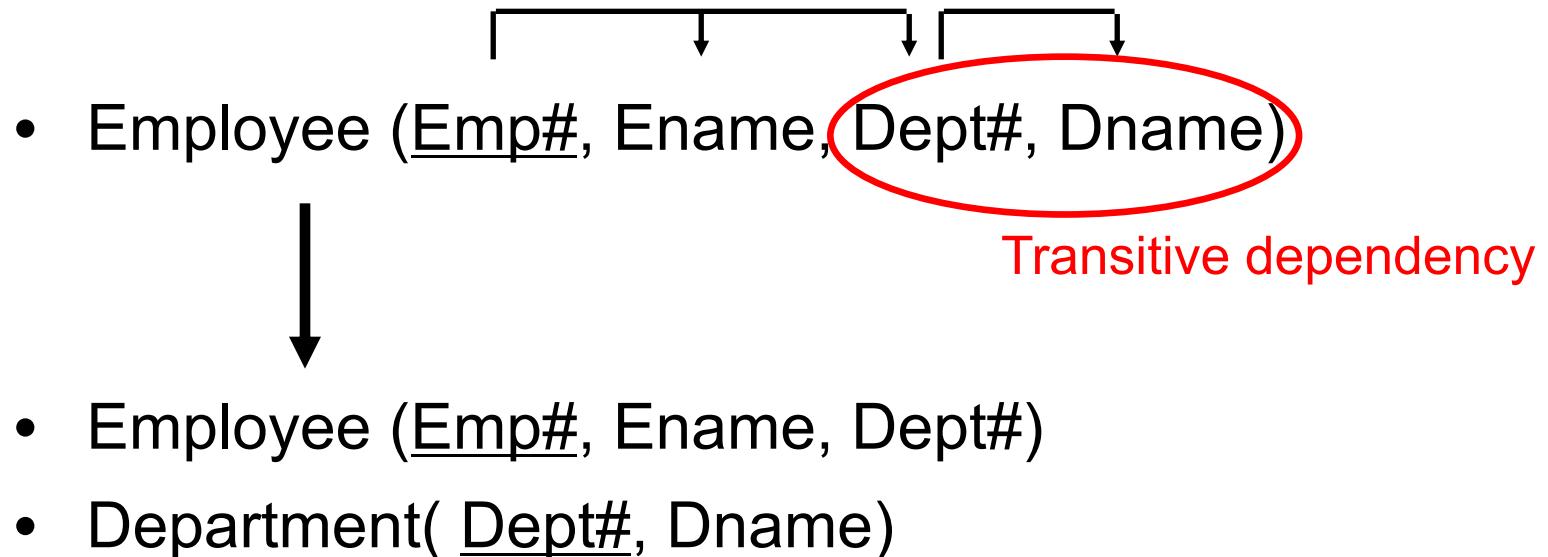
*change item  
description in one  
place*



## Remove Transitive Dependencies

a non-key attribute **cannot be identified**  
by another non-key attribute

Example: Employee (Emp#, Ename, Dept#, Dname)





## Example: Employee (Emp#, Ename, Dept#, Dname)

Emp#	Ename	Dept#	Dname
10	Smith	D5	MIS
20	Jones	D7	Finance
25	Smith	D7	Finance
30	Black	D8	Sales

- *UPDATE* change dept name in many places
- *DELETE* data for dept lost when last employee for that dept is deleted
- *INSERT* cannot add new dept until an employee is allocated to it



# Solution to these Anomalies

Employee

Emp#	Ename	Dept#
10	Smith	D5
20	Jones	D7
25	Smith	D7
30	Black	D8

*delete last emp in dept, but dept remains*

*add new dept at any time*

*change dept name in one place*

Dept#	Dname
D5	MIS
D7	Finance
D8	Sales

Every determinant must be a candidate key

"Every non-key attribute must provide a fact about the key, the whole key, and nothing but the key.  
(So help me Codd)"

Example: Allocation (Student#, Subject, Teacher)

- Allocation (Student#, Subject, Teacher, GPA)
  - ↓
    - Allocation (Student#, Teacher, Subject)
    - Assignment (Student#, Teacher, GPA)
    - Department( Teacher, Subject)



## Allocation

Stud#	Subject	Teacher
123	Physics	Einstein
123	Music	Mozart
456	Biology	Darwin
789	Physics	Bohr
999	Physics	Einstein

- *UPDATE* student changing subject may lose teacher for that subject
- *DELETE* deleting student may lose teacher for that subject
- *CREATE* cannot assign a teacher to a subject until a student takes it



# Solution to these Anomalies

Stud#	Teacher
123	Einstein
123	Mozart
456	Darwin
789	Bohr
999	Einstein

*Change enrolment  
without losing  
teacher/subject*



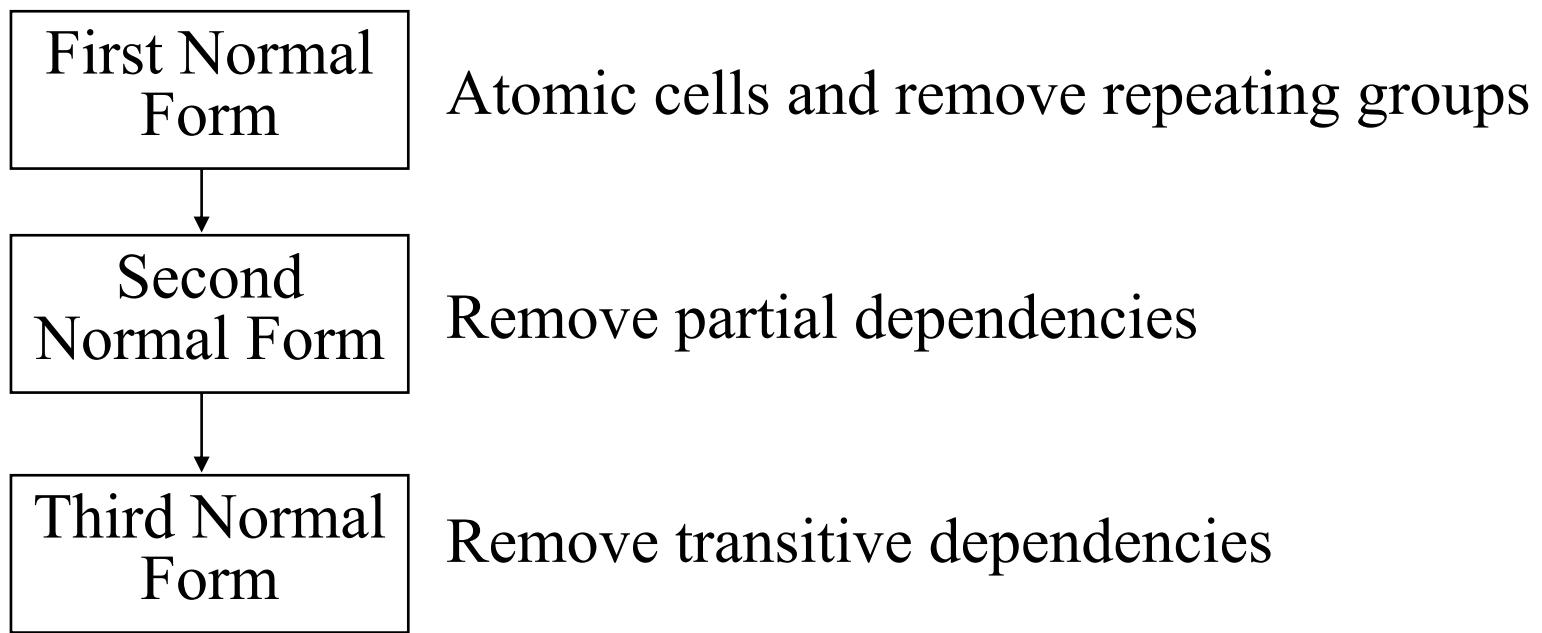
*delete student but  
subject remains*

Teacher	Subject
Einstein	Physics
Mozart	Music
Darwin	Biology
Bohr	Physics

*assign teacher to  
subject*



Boyce-Codd is taught for completeness but is **not examinable**





## Normalisation:

- Normalised relations contains a minimum amount of redundancy and allow users to **insert**, **modify**, and **delete** rows in tables without errors or inconsistencies (anomalies)

## Denormalization:

- The pay-off: query speed.
- The price: extra work on updates to keep redundant data consistent.
- Denormalization may be used to improve performance of time-critical operations.
  - Essential for Data Analytics! (Data Warehouse lecture)



- Normalisation Process (1NF -> 2NF -> 3NF)
- Anomalies
- Functional dependencies
- Denormalisation



- ER Modelling LIVE!
- Make sure you have read the Medicare Case Study
- Please attempt a conceptual model