

Introduction to Database Systems

CSE 414

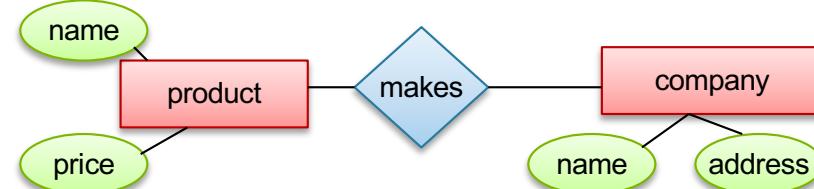
Lecture 22: Design Theory

Class Overview

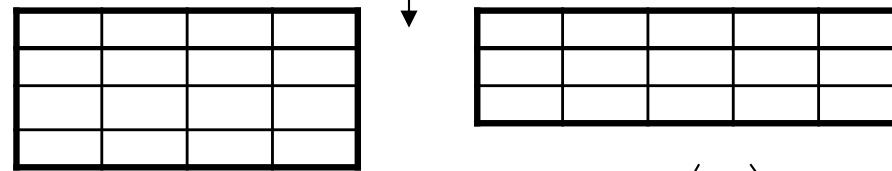
- Unit 1: Intro
- Unit 2: Relational Data Models and Query Languages
- Unit 3: Non-relational data
- Unit 4: RDMBS internals and query optimization
- Unit 5: Parallel query processing
- Unit 6: DBMS usability, conceptual design
 - E/R diagrams
 - Schema normalization
- Unit 7: Transactions

Database Design Process

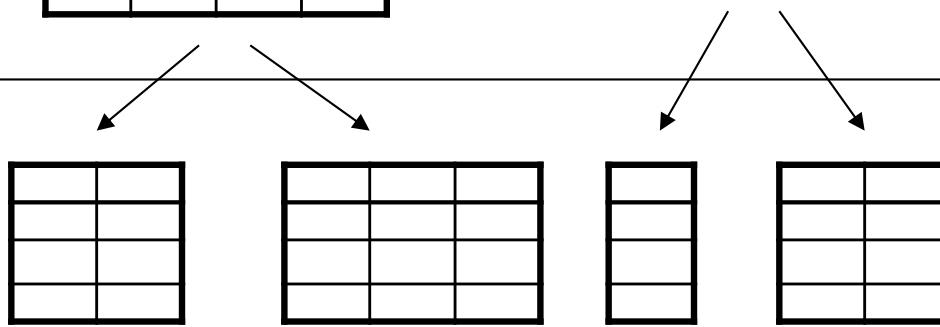
Conceptual Model:



Relational Model:
Tables + constraints
And also functional dep.



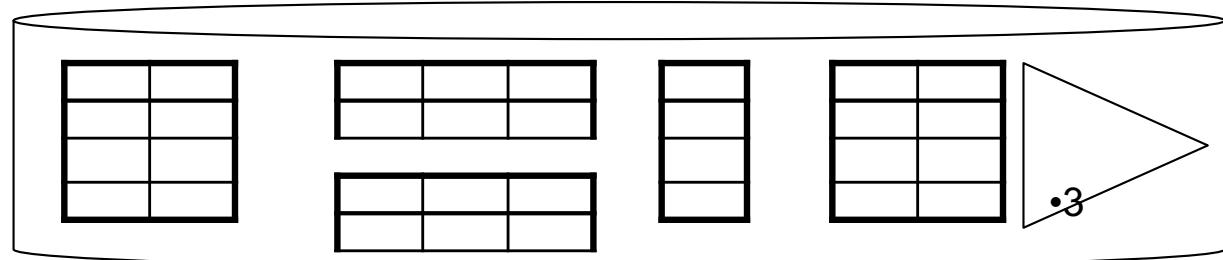
Normalization:
Eliminates anomalies



Conceptual Schema

Physical storage details

Physical Schema



Entity / Relationship Diagrams

- Entity set = a class
 - An entity = an object

Product

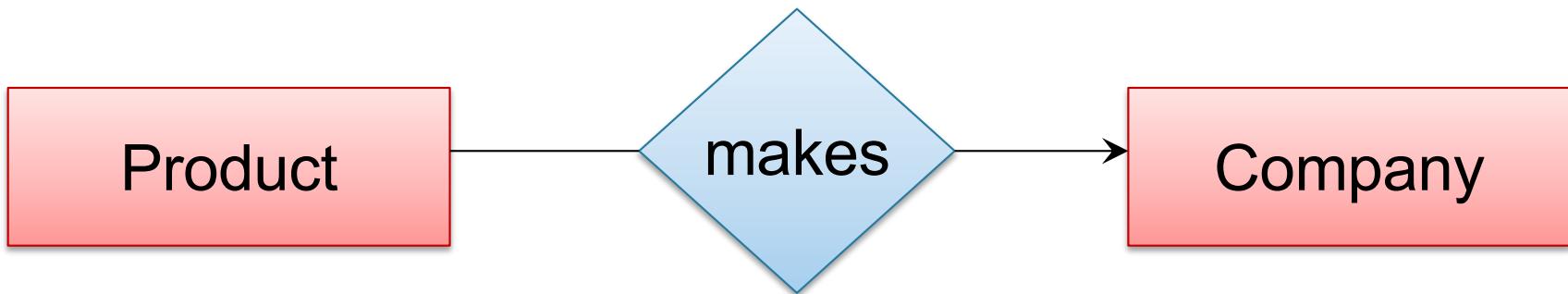
- Attribute

city

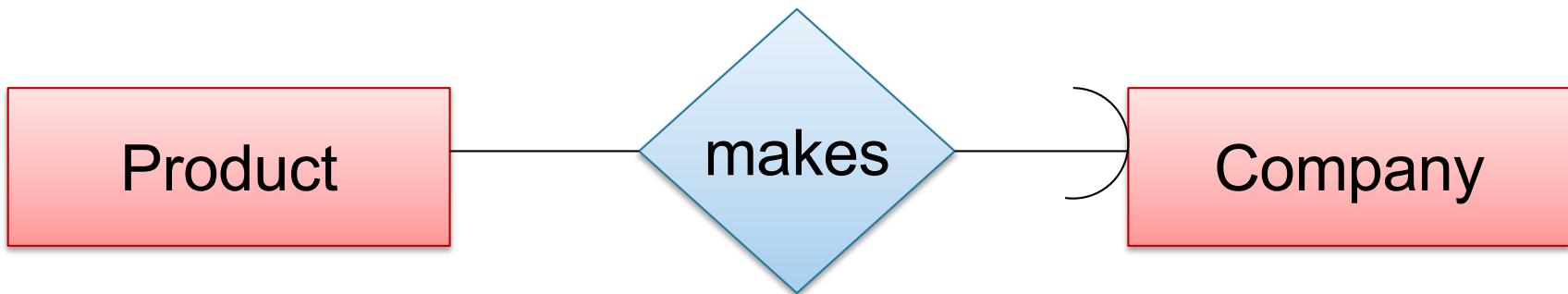
- Relationship

makes

Referential Integrity Constraints



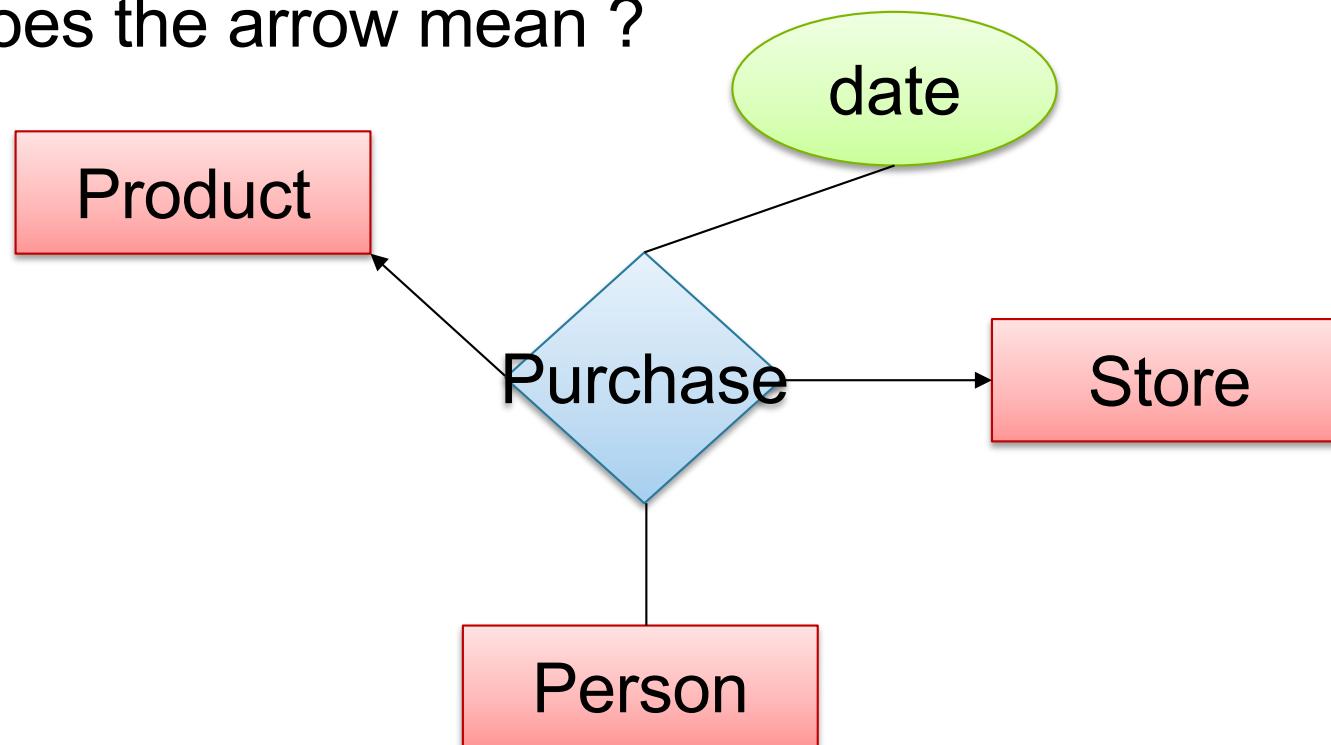
Each product made by at most one company.
Some products made by no company



Each product made by exactly one company.

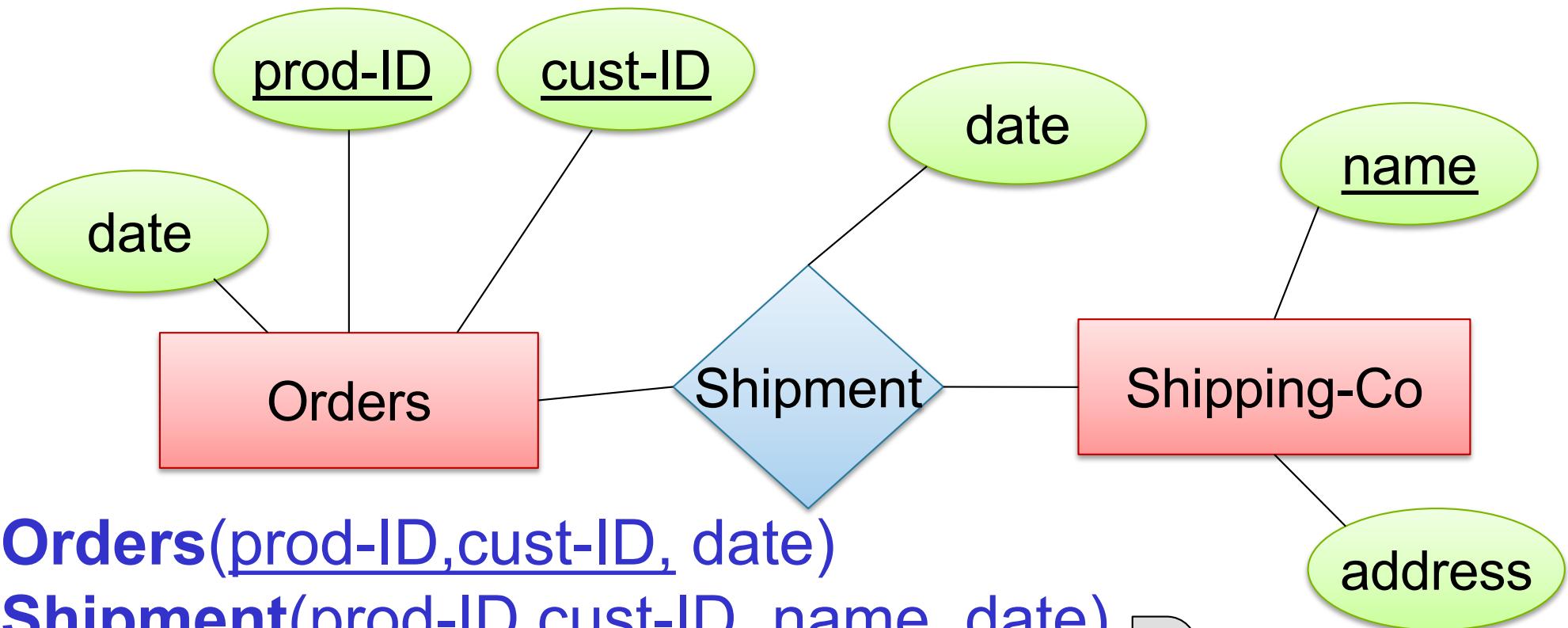
Arrows in Multiway Relationships

Q: What does the arrow mean ?



A: Any person buys a given product from at most one store
AND every store sells to every person at most one product

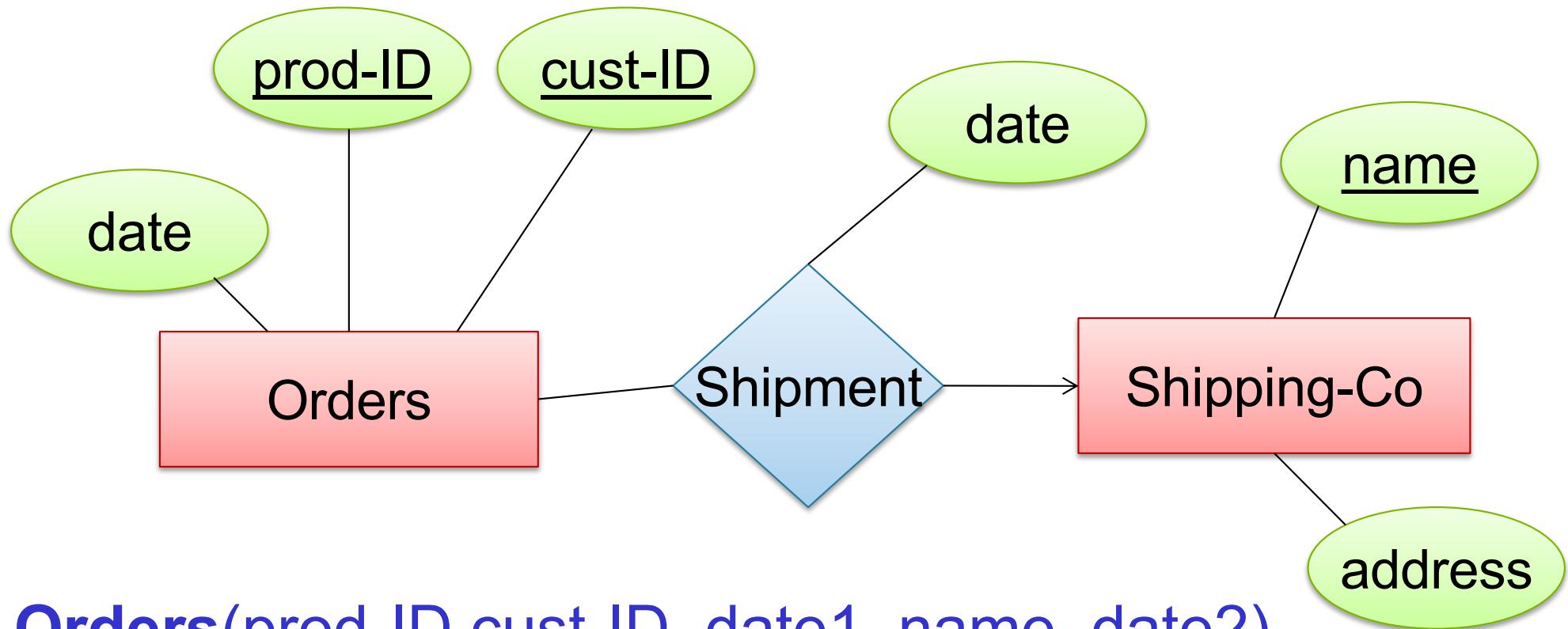
N-N Relationships to Relations



•7

<u>prod-ID</u>	<u>cust-ID</u>	<u>name</u>	date
Gizmo55	Joe12	UPS	4/10/2011
Gizmo55	Joe12	FEDEX	4/9/2011

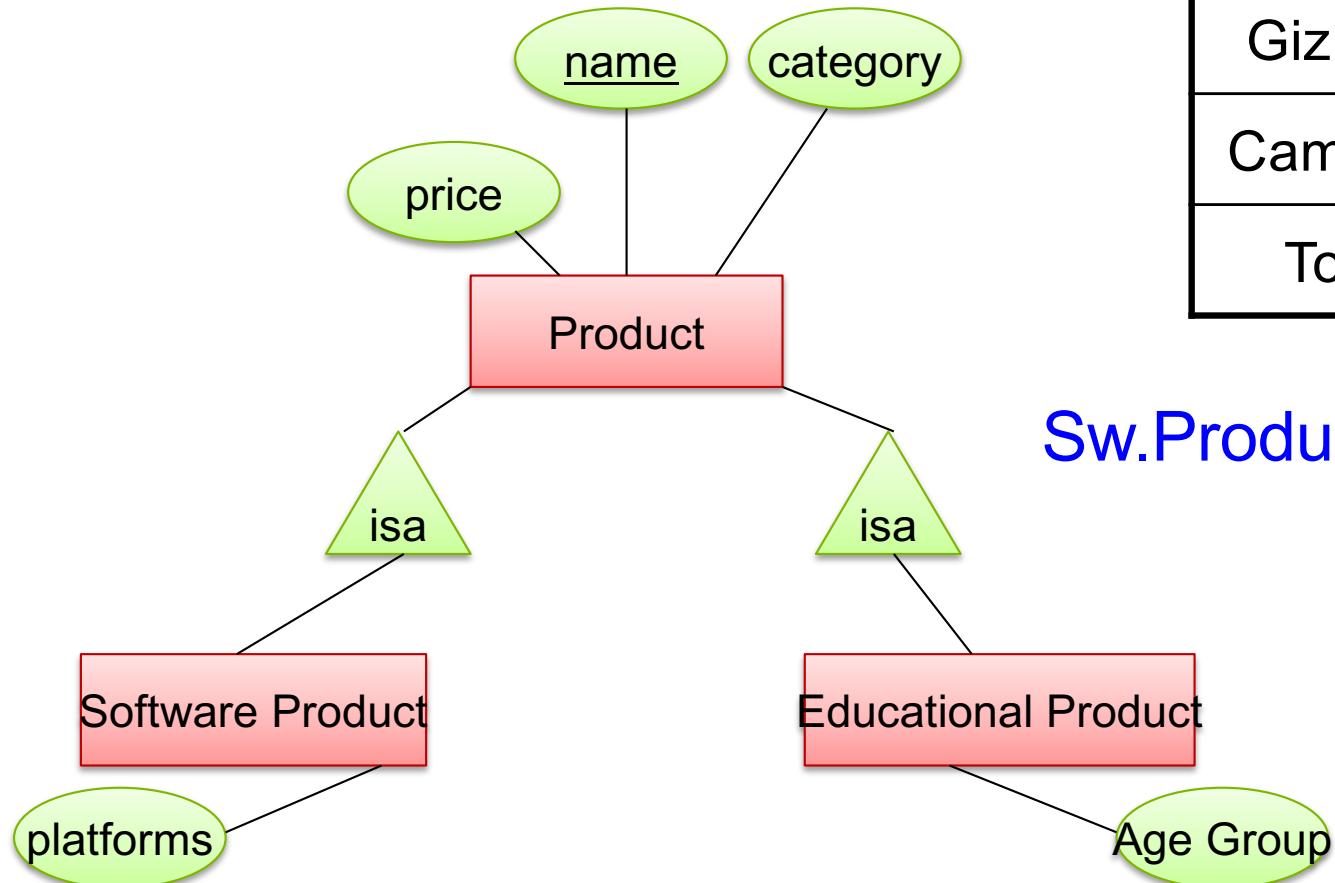
N-1 Relationships to Relations



Orders(prod-ID,cust-ID, date1, name, date2)
Shipping-Co(name, address)

Remember: no separate relations for many-one relationship

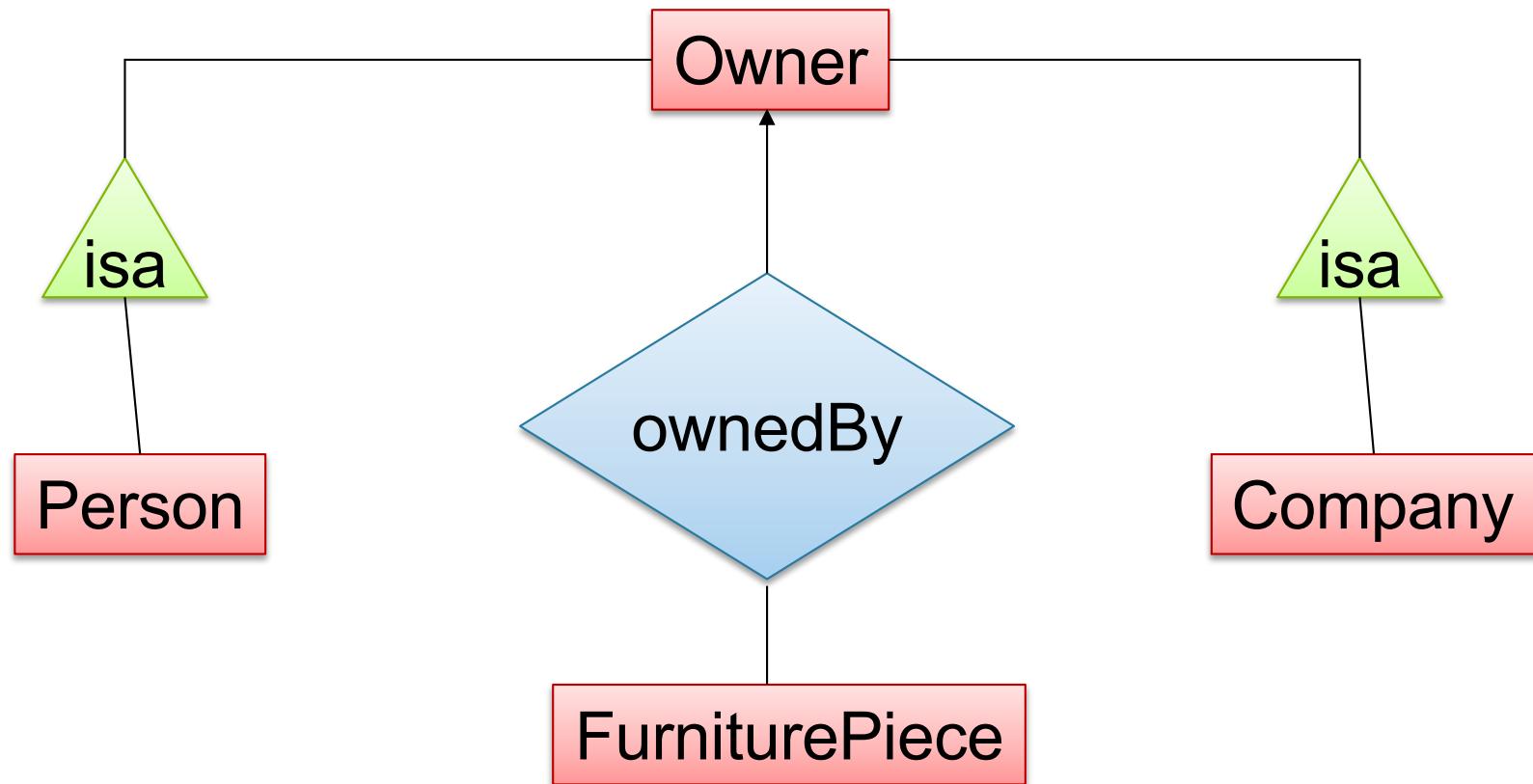
Subclasses to Relations



Other ways to convert are possible

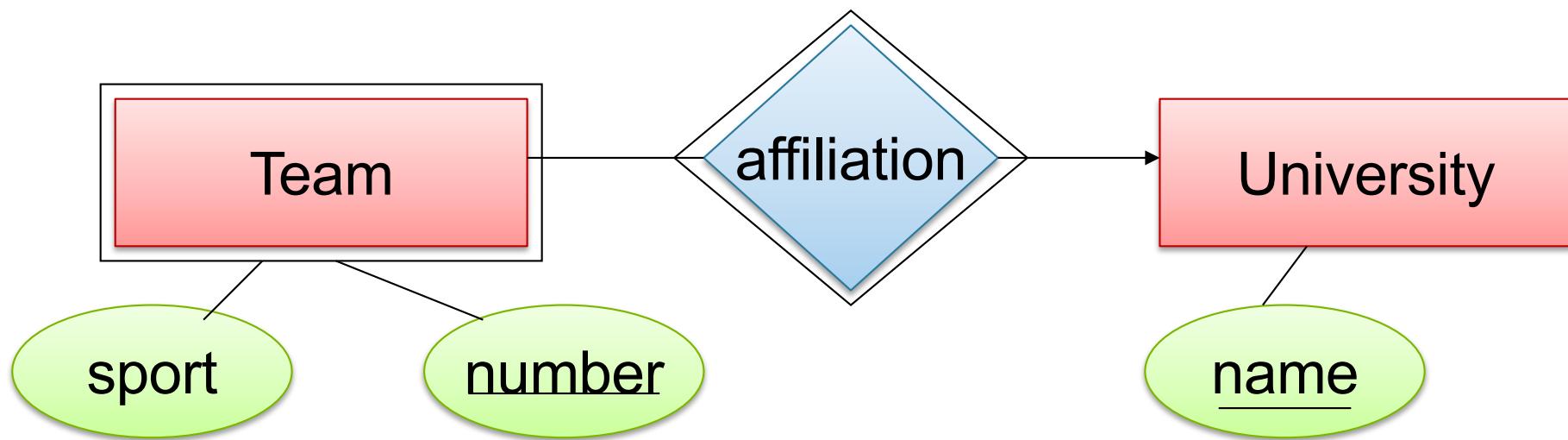
Modeling Union Types with Subclasses

Solution 2: better, more laborious



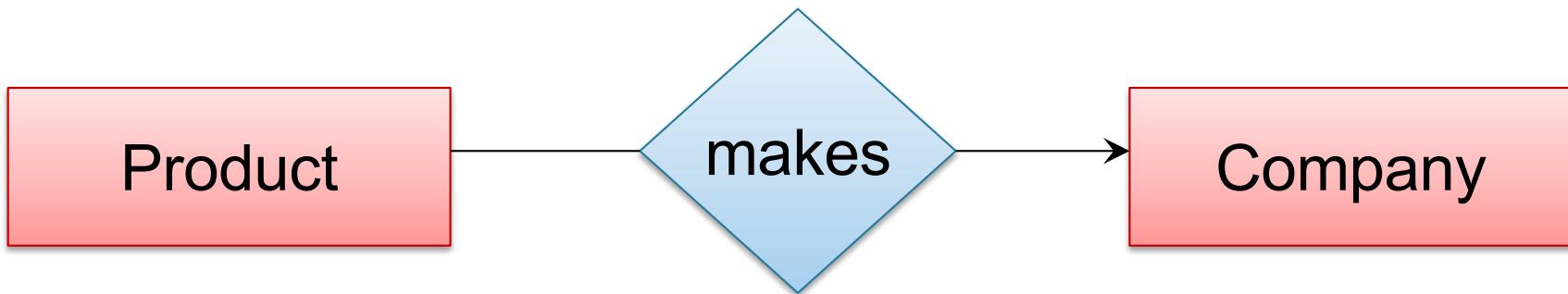
Weak Entity Sets

Entity sets are weak as their key comes from other classes to which they are related.

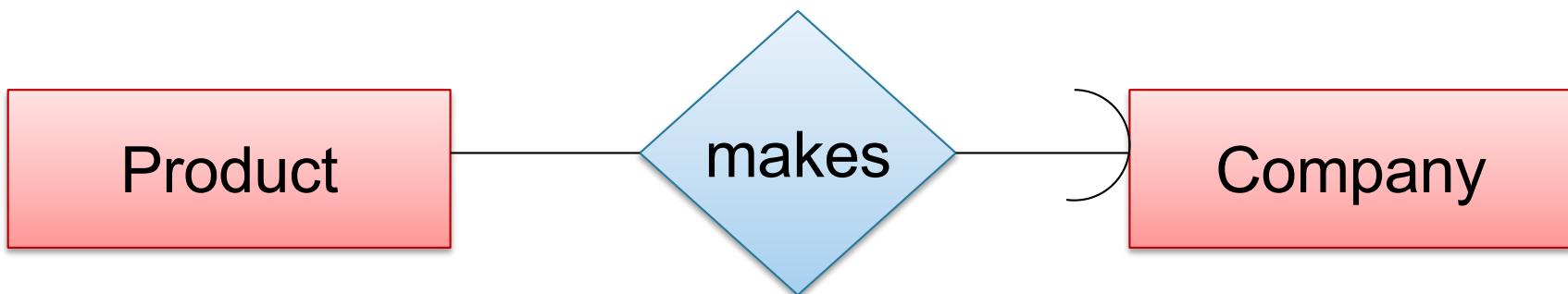


Team(sport, number, universityName)
University(name)

Referential Integrity Constraints

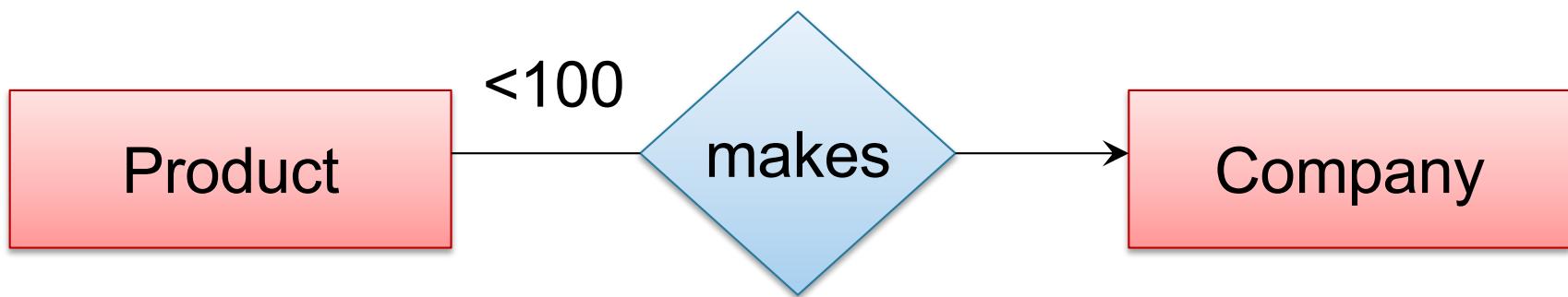


Each product made by at most one company.
Some products made by no company



Each product made by exactly one company.

Other Constraints



Q: What does this mean ?

A: A Company entity cannot be connected by relationship to more than 99 Product entities

Constraints in SQL

Constraints in SQL:

- Keys, foreign keys
- Attribute-level constraints
- Tuple-level constraints
- Global constraints: assertions
- The more complex the constraint, the harder it is to check and to enforce

simplest

Most
complex

What happens when data changes?

- SQL has three policies for maintaining referential integrity:
- NO ACTION reject violating modifications (default)
- CASCADE after delete/update do delete/update
- SET NULL set foreign-key field to NULL
- SET DEFAULT set foreign-key field to default value
 - need to be declared with column, e.g.,
CREATE TABLE Product (pid INT DEFAULT 42)

Relational Schema Design

Name	<u>SSN</u>	<u>PhoneNumber</u>	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield

One person may have multiple phones, but lives in only one city

Primary key is thus (SSN, PhoneNumber)

What is the problem with this schema?

Relational Schema Design

Name	<u>SSN</u>	<u>PhoneNumber</u>	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield

Anomalies:

- Redundancy = repeat data
- Update anomalies = what if Fred moves to “Bellevue”?
- Deletion anomalies = what if Joe deletes his phone number?

Relation Decomposition

Break the relation into two:

Name	SSN	PhoneNumber	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield

Name	<u>SSN</u>	City
Fred	123-45-6789	Seattle
Joe	987-65-4321	Westfield

<u>SSN</u>	<u>PhoneNumber</u>
123-45-6789	206-555-1234
123-45-6789	206-555-6543
987-65-4321	908-555-2121

Anomalies have gone:

- No more repeated data
- Easy to move Fred to “Bellevue” (how ?)
- Easy to delete all Joe’s phone numbers (how ?)

Relational Schema Design (or Logical Design)

How do we do this systematically?

- Start with some relational schema
- Find out its *functional dependencies* (FDs)
- Use FDs to *normalize* the relational schema

Functional Dependencies (FDs)

Definition

If two tuples agree on the attributes

A_1, A_2, \dots, A_n

then they must also agree on the attributes

B_1, B_2, \dots, B_m

Formally:

$A_1 \dots A_n$ determines $B_1 \dots B_m$

$A_1, A_2, \dots, A_n \rightarrow B_1, B_2, \dots, B_m$

$A_1 \rightarrow B_1$ and $A_2 \rightarrow B_2$ then $A_1, A_2 \rightarrow B_1, B_2$

Functional Dependencies (FDs)

Definition $A_1, \dots, A_m \rightarrow B_1, \dots, B_n$ holds in R if:

for all $\forall t, t' \in R,$ *and* $(t.A_1 = t'.A_1 \wedge \dots \wedge t.A_m = t'.A_m \rightarrow t.B_1 = t'.B_1 \wedge \dots \wedge t.B_n = t'.B_n)$

R	A_1	\dots	A_m		B_1	\dots	B_n	
t								
t'								

if t, t' agree here then t, t' agree here

Example

An FD holds, or does not hold on an instance:

EmplID	Name	Phone	Position
E0045	Smith	1234	Clerk
E3542	Mike	9876	Salesrep
E1111	Smith	9876	Salesrep
E9999	Mary	1234	Lawyer

EmplID → Name, Phone, Position

Position → Phone

but not Phone → Position

Example

EmplID	Name	Phone	Position
E0045	Smith	1234	Clerk
E3542	Mike	9876 ←	Salesrep
E1111	Smith	9876 ←	Salesrep
E9999	Mary	1234	Lawyer

Position → Phone

Example

EmplID	Name	Phone	Position
E0045	Smith	1234 →	Clerk
E3542	Mike	9876	Salesrep
E1111	Smith	9876	Salesrep
E9999	Mary	1234 →	Lawyer

But not Phone → Position

Example

name → color
category → department
color, category → price

name	category	color	department	price
Gizmo	Gadget	Green	Toys	49
Tweaker	Gadget	Green	Toys	99

Do all the FDs hold on this instance?

Example

name → color
category → department
color, category → price

name	category	color	department	price
Gizmo	Gadget	Green	Toys	49
Tweaker	Gadget	Green	Toys	49
Gizmo	Stationary	Green	Office-supp.	59

What about this one ?

Buzzwords

- FD **holds** or **does not hold** on an instance
- If we can be sure that *every instance of R* will be one in which a given FD is true, then we say that **R satisfies the FD**
- If we say that R satisfies an FD, we are **stating a constraint on R**

Why bother with FDs?

Name	<u>SSN</u>	<u>PhoneNumber</u>	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield

Anomalies:

- Redundancy = repeat data
- Update anomalies = what if Fred moves to “Bellevue”?
- Deletion anomalies = what if Joe deletes his phone number?

An Interesting Observation

If all these FDs are true:

$\text{name} \rightarrow \text{color}$
 $\text{category} \rightarrow \text{department}$
 $\text{color, category} \rightarrow \text{price}$

Then this FD also holds:

$\text{name, category} \rightarrow \text{price}$

If we find out from application domain that a relation satisfies some FDs, it doesn't mean that we found all the FDs that it satisfies! There could be more FDs implied by the ones we have.

Closure of a set of Attributes

Given a set of attributes A_1, \dots, A_n

The **closure** is the set of attributes B, notated $\{A_1, \dots, A_n\}^+$,
s.t. $A_1, \dots, A_n \rightarrow B$

Example:

1. name \rightarrow color
2. category \rightarrow department
3. color, category \rightarrow price

Closures:

$$\text{name}^+ = \{\text{name}, \text{color}\}$$

$$\{\text{name}, \text{category}\}^+ = \{\text{name}, \text{category}, \text{color}, \text{department}, \text{price}\}$$

$$\text{color}^+ = \{\text{color}\}$$

Closure Algorithm

$X = \{A_1, \dots, A_n\}$.

Repeat until X doesn't change **do:**
if $B_1, \dots, B_n \rightarrow C$ is a FD **and**
 B_1, \dots, B_n are all in X
then add C to X .

Example:

1. name \rightarrow color
2. category \rightarrow department
3. color, category \rightarrow price

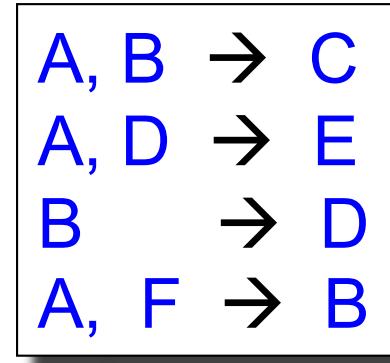
$\{name, category\}^+ =$
 { name, category, color, department, price }

Hence: name, category \rightarrow color, department, price

Example

In class:

R(A,B,C,D,E,F)



Compute $\{A, B\}^+$ $X = \{A, B,$

Compute $\{A, F\}^+ \quad X = \{A, F,$

Example

In class:

$R(A,B,C,D,E,F)$

$A, B \rightarrow C$
$A, D \rightarrow E$
$B \rightarrow D$
$A, F \rightarrow B$

Compute $\{A, B\}^+$ $X = \{A, B, C, D, E\}$

Compute $\{A, F\}^+$ $X = \{A, F, \}$

Example

In class:

$R(A,B,C,D,E,F)$

$A, B \rightarrow C$
$A, D \rightarrow E$
$B \rightarrow D$
$A, F \rightarrow B$

Compute $\{A, B\}^+$ $X = \{A, B, C, D, E\}$

Compute $\{A, F\}^+$ $X = \{A, F, B, C, D, E\}$

Example

In class:

$R(A,B,C,D,E,F)$

A, B → C
A, D → E
B → D
A, F → B

Compute $\{A, B\}^+$ $X = \{A, B, C, D, E\}$

Compute $\{A, F\}^+$ $X = \{A, F, B, C, D, E\}$

Practice at Home

Find all FD's implied by:

$$\begin{array}{l} A, B \rightarrow C \\ A, D \rightarrow B \\ B \rightarrow D \end{array}$$

Step 1: Compute X^+ , for every X :

$$A^+ = A, \quad B^+ = BD, \quad C^+ = C, \quad D^+ = D$$

$$\begin{aligned} AB^+ &= ABCD, \quad AC^+ = AC, \quad AD^+ = ABCD, \\ &\quad BC^+ = BCD, \quad BD^+ = BD, \quad CD^+ = CD \end{aligned}$$

$$ABC^+ = ABD^+ = ACD^+ = ABCD \text{ (no need to compute— why ?)}$$

$$BCD^+ = BCD, \quad ABCD^+ = ABCD$$

Step 2: Enumerate all FD's $X \rightarrow Y$, s.t. $Y \subseteq X^+$ and $X \cap Y = \emptyset$:

$$AB \rightarrow CD, \quad AD \rightarrow BC, \quad ABC \rightarrow D, \quad ABD \rightarrow C, \quad ACD \rightarrow B$$

Keys

- A **superkey** is a set of attributes A_1, \dots, A_n s.t. for any other attribute B , we have $A_1, \dots, A_n \rightarrow B$
i.e. the closure set $\{A_1, A_2, \dots, A_n\}^+ = \{\text{all attributes}\}$
- A **key** is a minimal superkey
 - A superkey and for which no subset is a superkey

Computing (Super)Keys

- For all sets X , compute X^+
- If $X^+ = [\text{all attributes}]$, then X is a superkey
- Try reducing to the minimal X 's to get the key

Example

Product(name, price, category, color)

name, category → price
category → color

What is the key ?

Example

Product(name, price, category, color)

name, category → price
category → color

What is the key ?

(name, category) + = { name, category, price, color }

Hence (name, category) is a key

Key or Keys ?

Can we have more than one key ?

Given $R(A,B,C)$ define FD's s.t. there are two or more distinct keys

Key or Keys ?

Can we have more than one key ?

Given $R(A,B,C)$ define FD's s.t. there are two or more distinct keys

$$\begin{array}{l} A \rightarrow B \\ B \rightarrow C \\ C \rightarrow A \end{array}$$

or

$$\begin{array}{l} AB \rightarrow C \\ BC \rightarrow A \end{array}$$

or

$$\begin{array}{l} A \rightarrow BC \\ B \rightarrow AC \end{array}$$

what are the keys here ?

Eliminating Anomalies

Main idea:

- $X \rightarrow A$ is OK if X is a (super)key
- $X \rightarrow A$ is not OK otherwise
 - Need to decompose the table, but how?

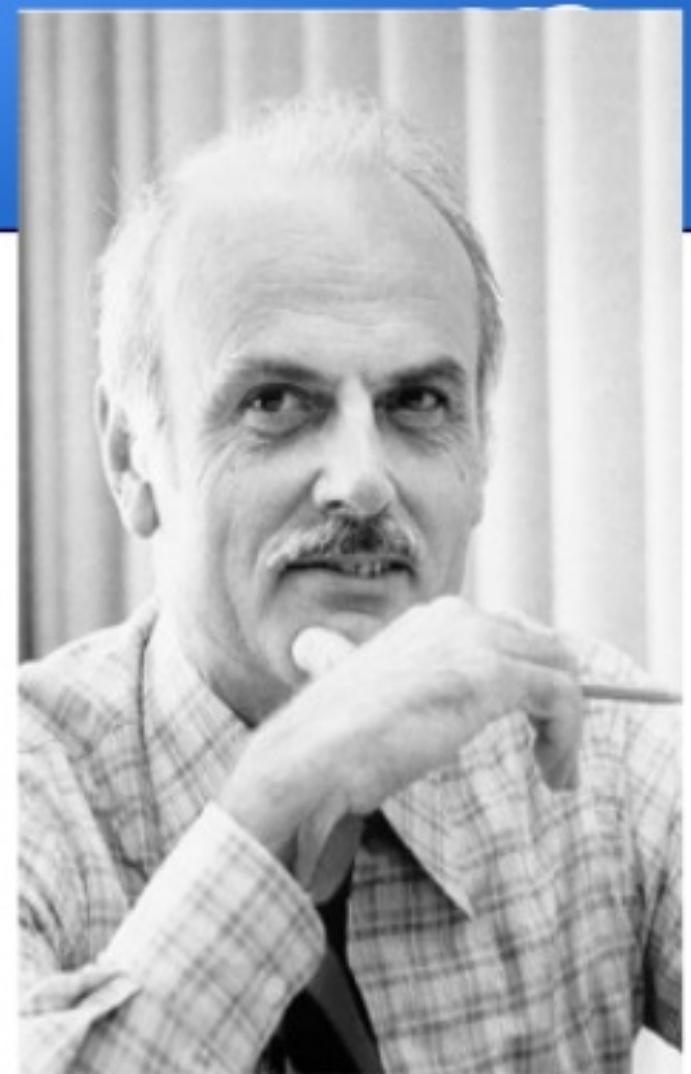
Boyce-Codd Normal Form

Boyce-Codd Normal Form

Dr. Raymond F. Boyce

Edgar Frank “Ted” Codd

"A Relational Model of Data for Large Shared Data Banks"



Boyce-Codd Normal Form

There are no
“bad” FDs:

Definition. A relation R is in BCNF if:

Whenever $X \rightarrow B$ is a non-trivial dependency,
then X is a superkey.

Equivalently:

Definition. A relation R is in BCNF if:

$\forall X$, either $X^+ = X$ or $X^+ = [\text{all attributes}]$

BCNF Decomposition Algorithm

Normalize(R)

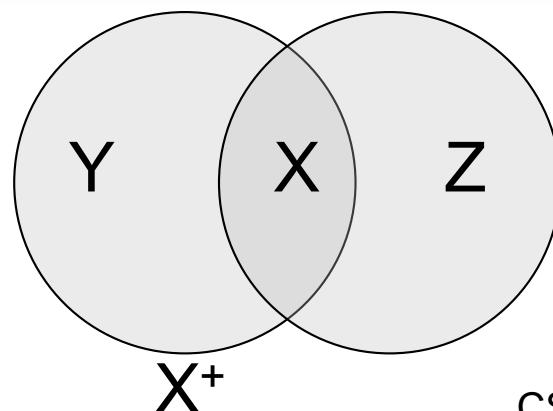
find X s.t.: $X \neq X^+$ and $X^+ \neq [\text{all attributes}]$

if (not found) then “R is in BCNF”

let $Y = X^+ - X$; $Z = [\text{all attributes}] - X^+$

decompose R into $R_1(X \cup Y)$ and $R_2(X \cup Z)$

Normalize(R_1); Normalize(R_2);



Example

Name	SSN	PhoneNumber	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield
Joe	987-65-4321	908-555-1234	Westfield

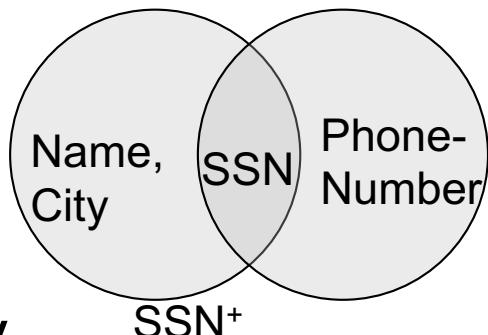
$\text{SSN} \rightarrow \text{Name, City}$

The only key is: {SSN, PhoneNumber}

Hence $\text{SSN} \rightarrow \text{Name, City}$ is a “bad” dependency

In other words:

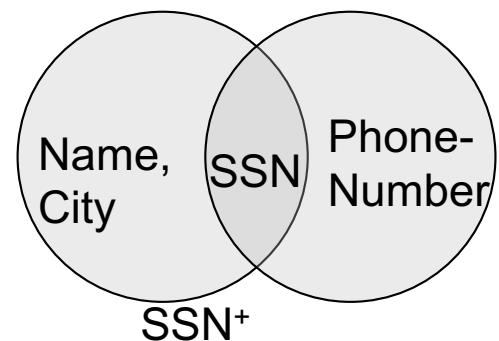
$\text{SSN}^+ = \text{SSN, Name, City}$ and is neither SSN nor All Attributes



Example BCNF Decomposition

Name	<u>SSN</u>	City
Fred	123-45-6789	Seattle
Joe	987-65-4321	Westfield

$\text{SSN} \rightarrow \text{Name, City}$



<u>SSN</u>	<u>PhoneNumber</u>
123-45-6789	206-555-1234
123-45-6789	206-555-6543
987-65-4321	908-555-2121
987-65-4321	908-555-1234

Let's check anomalies:

- Redundancy ?
- Update ?
- Delete ?