Cơ sở dữ liệu

TS. Hồ Mạnh Tài

Khoa CNTT2

Học viện Công nghệ Bưu chính Viễn thông
2018

Chương 5: Chuẩn hóa cơ sở dữ liệu

1. Boyce-Codd Normal Form

Thiết kế khái niệm - Conceptual Design

Now that we know how to find FDs, it’s a straight-forward process:

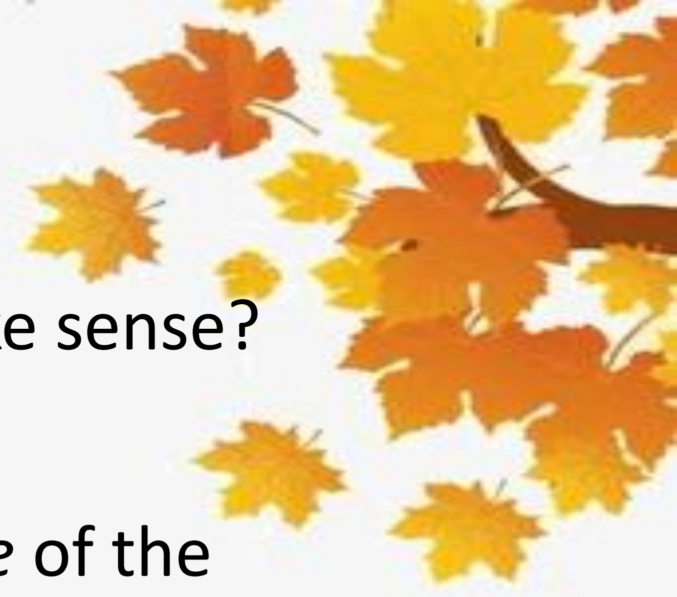
1. Search for “bad” FDs
2. If there are any, then *keep decomposing the table into sub-tables* until no more bad FDs
3. When done, the database schema is *normalized*

Recall: there are several normal forms…

Boyce-Codd Normal Form (BCNF)

- Main idea is that we define “good” and “bad” FDs as follows:
  - X → A is a *“good FD” if X is a (super)key*
    - In other words, if A is the set of all attributes
  - X → A is a *“bad FD”* otherwise
- We will try to eliminate the “bad” FDs!

Boyce-Codd Normal Form (BCNF)



- Why does this definition of “good” and “bad” FDs make sense?

- If X is *not* a (super)key, it functionally determines *some* of the attributes; therefore, those other attributes can be duplicated
  - Recall: this means there is <u>redundancy</u>
  - And redundancy like this can lead to data anomalies!

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

Boyce-Codd Normal Form (BCNF)

BCNF is a simple condition for removing anomalies from relations:

A relation R is **<u>in BCNF</u>** if:

if **$\{A_1, \ldots, A_n\}$** → **B** is a *non-trivial* FD in R

then **$\{A_1, \ldots, A_n\}$ is a superkey** for R

Equivalently: ∀ sets of attributes X, either ($X^+ = X$) or (X^+ = all attributes)

In other words: there are no “bad” FDs

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

{SSN} → {Name,City}

This FD is *bad* because it is **not** a superkey

⟹ **Not** in BCNF

What is the key?
{SSN, PhoneNumber}

Example

Name	SSN	City
Fred	123-45-6789	Seattle
Joe	987-65-4321	Madison

SSN	PhoneNumber
123-45-6789	206-555-1234
123-45-6789	206-555-6543
987-65-4321	908-555-2121
987-65-4321	908-555-1234

{SSN} → {Name,City}

This FD is now *good* because it is the key

Let's check anomalies:

- Redundancy ?
- Update ?
- Delete ?

Now in BCNF!

BCNF Decomposition Algorithm

```
BCNFDecomp(R):
```

BCNF Decomposition Algorithm

```
BCNFDecomp(R):
    Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]
```

Find a set of attributes X which has non-trivial "bad" FDs, i.e. is not a superkey, using closures

BCNF Decomposition Algorithm

```
BCNFDecomp(R):
  Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]

  if (not found) then Return R
```

If no “bad” FDs found, in BCNF!

BCNF Decomposition Algorithm

```
BCNFDecomp(R):
  Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]

  if (not found) then Return R

  let $Y = X^+ - X$,  $Z = (X^+)^C$
```

Let Y be the attributes that ***X functionally determines*** (+ that are not in X)

And let Z be **the *complement*, the other attributes that it *doesn't***

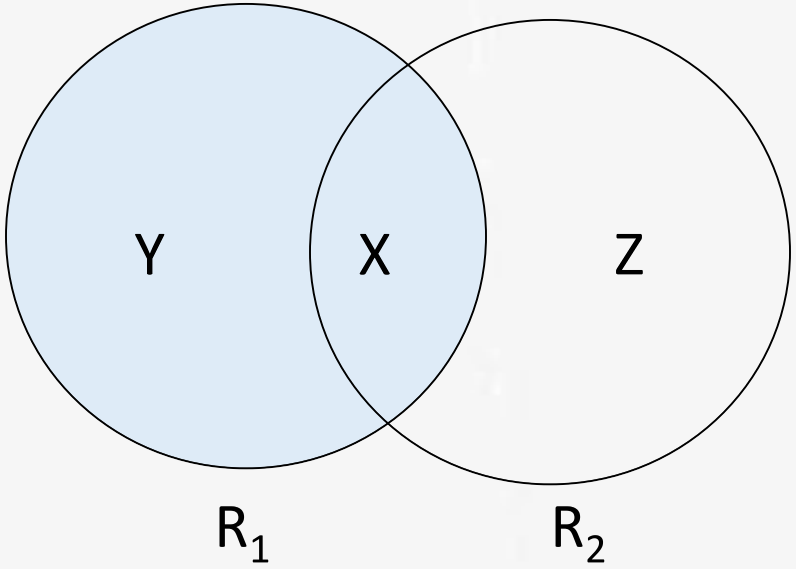
BCNF Decomposition Algorithm

```
BCNFDecomp(R):
   Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]

   if (not found) then Return R

   let $Y = X^+ - X$,  $Z = (X^+)^C$
   decompose R into $R_1(X \cup Y)$ and $R_2(X \cup Z)$
```

Split into one relation (table) with X plus the attributes that X determines (Y)…



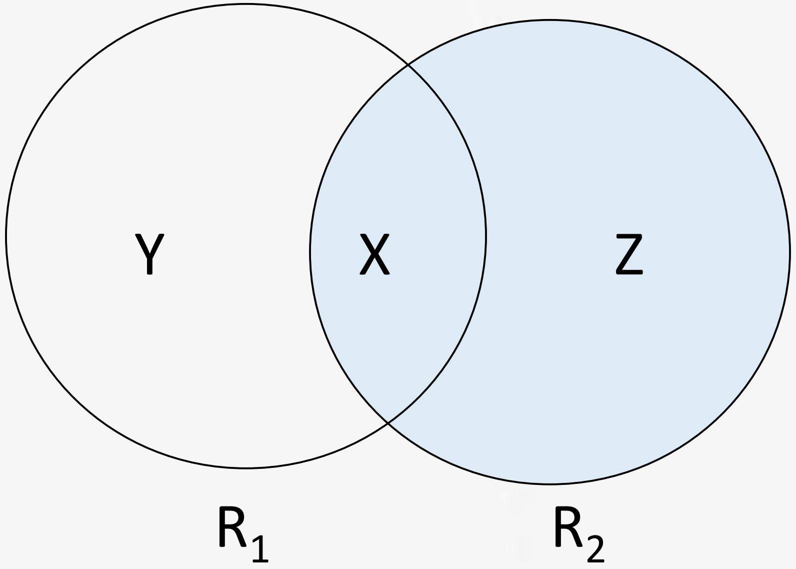
BCNF Decomposition Algorithm

```
BCNFDecomp(R):
  Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]

  if (not found) then Return R

  let $Y = X^+ - X$,  $Z = (X^+)^C$
  decompose R into $R_1(X \cup Y)$ and $R_2(X \cup Z)$
```

And one relation with X plus the attributes it *does not* determine (Z)



BCNF Decomposition Algorithm

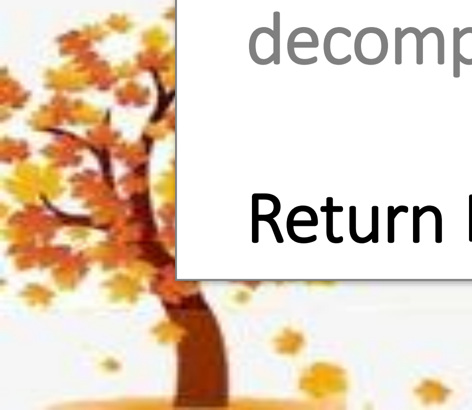
```
BCNFDecomp(R):
  Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]

  if (not found) then Return R

  let $Y = X^+ - X$,  $Z = (X^+)^C$
  decompose R into $R_1(X \cup Y)$ and $R_2(X \cup Z)$

  Return BCNFDecomp($R_1$), BCNFDecomp($R_2$)
```

Proceed recursively until no more “bad” FDs!



Example

```
BCNFDecomp(R):
  Find a set of attributes X s.t.: $X^+ \neq X$ and $X^+ \neq$ [all attributes]

  if (not found) then Return R

  let $Y = X^+ - X$,  $Z = (X^+)^C$
  decompose R into $R_1(X \cup Y)$ and $R_2(X \cup Z)$

  Return BCNFDecomp($R_1$), BCNFDecomp($R_2$)
```

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

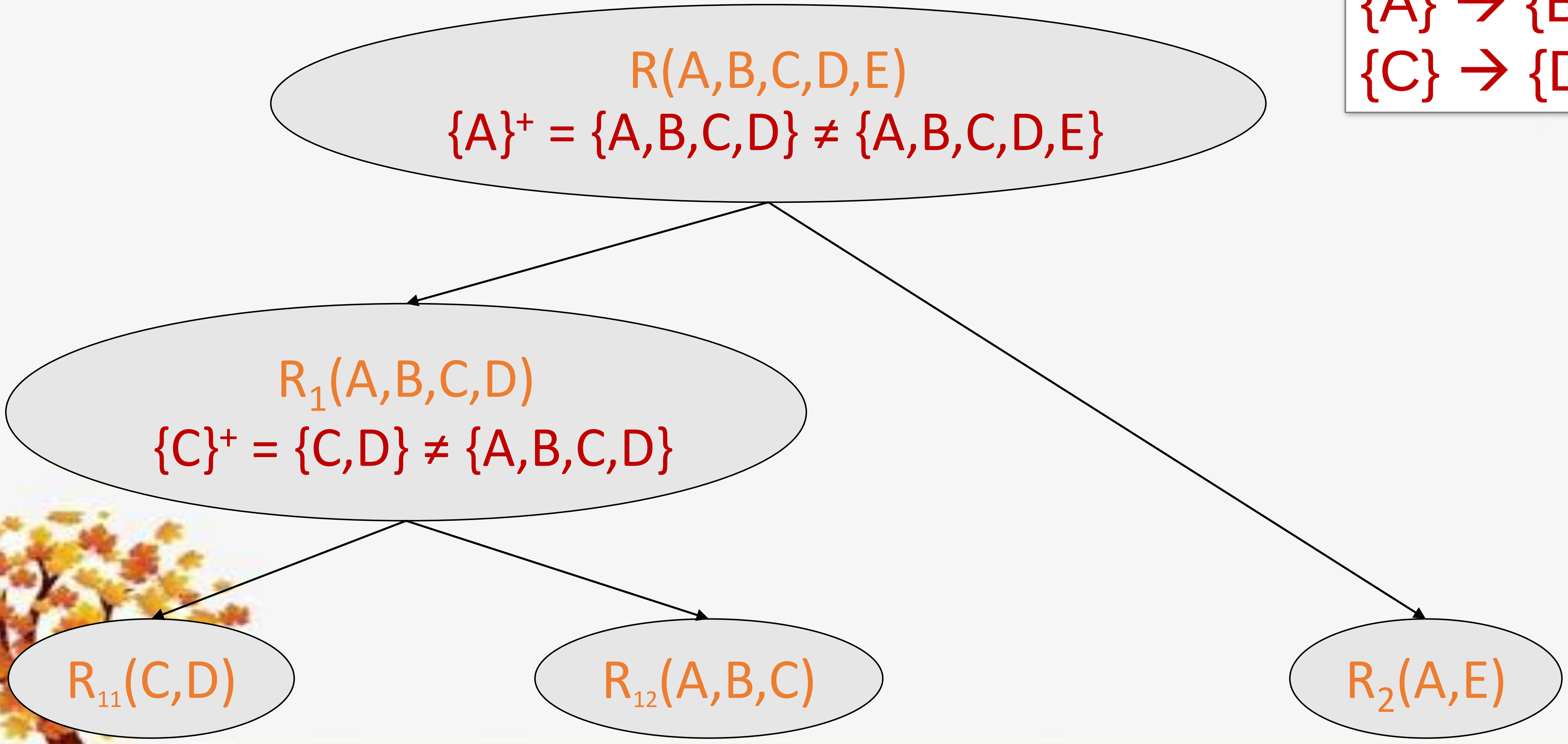
$\{A\} \rightarrow \{B,C\}$
$\{C\} \rightarrow \{D\}$

Example

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

{A} → {B,C}

{C} → {D}



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

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

$R_1(A,B,C,D)$

$\{C\}^+ = \{C,D\} \neq \{A,B,C,D\}$

$R_{11}(C,D)$

$R_{12}(A,B,C)$

$R_2(A,E)$

2. Decompositions

Recap: Decompose to remove redundancies

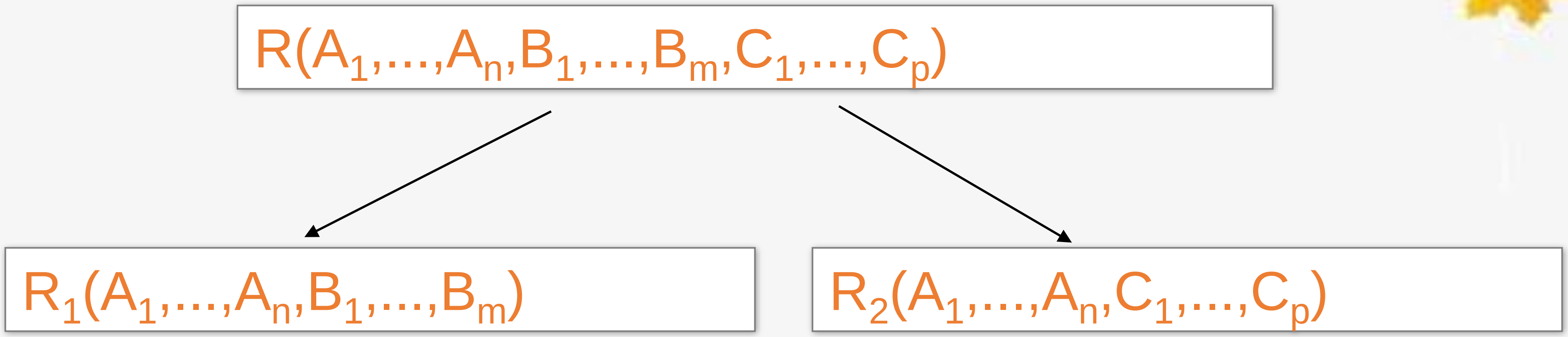
1. We saw that **redundancies** in the data (“bad FDs”) can lead to data anomalies

2. We developed mechanisms to **detect and remove redundancies by decomposing tables into BCNF**
    1. BCNF decomposition is *standard practice*- very powerful & widely used!

3. However, sometimes decompositions can lead to **more subtle unwanted effects…**

When does this happen?

Decompositions in General



$R(A_1,\ldots,A_n,B_1,\ldots,B_m,C_1,\ldots,C_p)$

$R_1(A_1,\ldots,A_n,B_1,\ldots,B_m)$

$R_2(A_1,\ldots,A_n,C_1,\ldots,C_p)$

R_1 = the *projection* of R on $A_1, \ldots, A_n, B_1, \ldots, B_m$

R_2 = the *projection* of R on $A_1, \ldots, A_n, C_1, \ldots, C_p$

Theory of Decomposition

Name	Price	Category
Gizmo	19.99	Gadget
OneClick	24.99	Camera
Gizmo	19.99	Camera

Name	Price
Gizmo	19.99
OneClick	24.99
~~Gizmo~~	~~19.99~~

Name	Category
Gizmo	Gadget
OneClick	Camera
Gizmo	Camera

Sometimes a decomposition is "correct"

I.e. it is a **Lossless decomposition**

Lossy Decomposition

Name	Price	Category
Gizmo	19.99	Gadget
OneClick	24.99	Camera
Gizmo	19.99	Camera

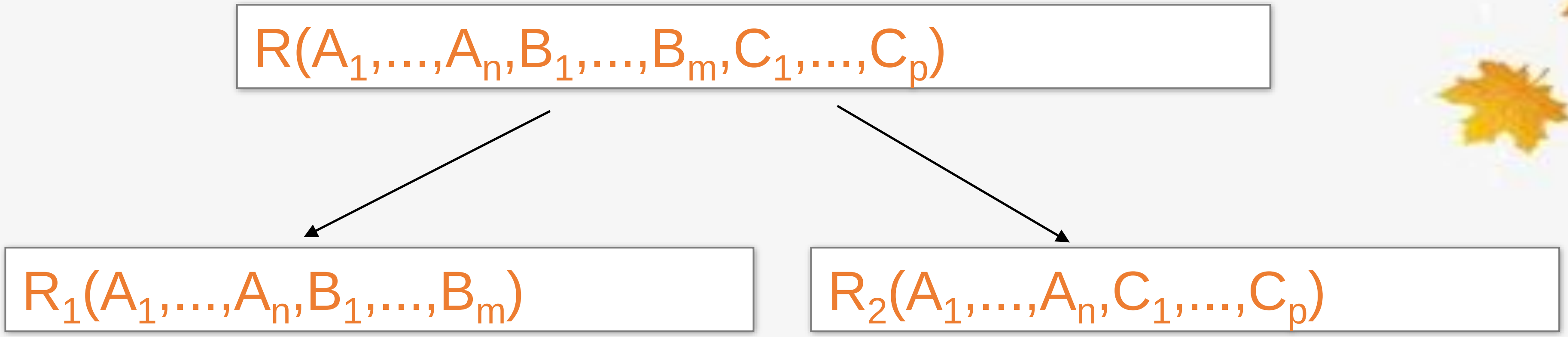
Name	Category
Gizmo	Gadget
OneClick	Camera
Gizmo	Camera

Price	Category
19.99	Gadget
24.99	Camera
19.99	Camera

However sometimes it isn’t

What’s wrong here?

Lossless Decompositions



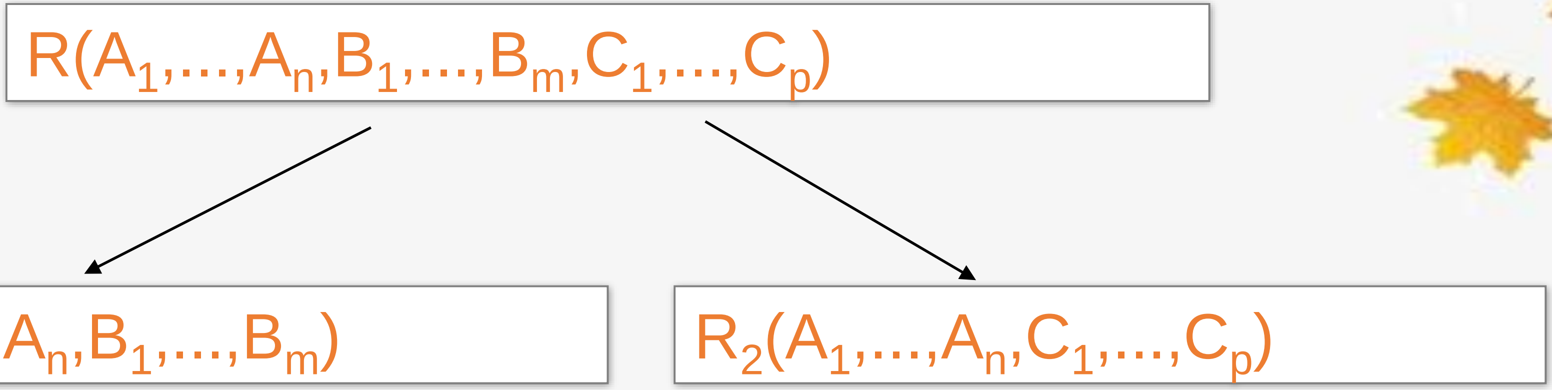
$R(A_1,\ldots,A_n,B_1,\ldots,B_m,C_1,\ldots,C_p)$

$R_1(A_1,\ldots,A_n,B_1,\ldots,B_m)$

$R_2(A_1,\ldots,A_n,C_1,\ldots,C_p)$

A decomposition R to (R1, R2) is **lossless** if R = R1 Join R2

Lossless Decompositions



$R(A_1,\ldots,A_n,B_1,\ldots,B_m,C_1,\ldots,C_p)$

$R_1(A_1,\ldots,A_n,B_1,\ldots,B_m)$

$R_2(A_1,\ldots,A_n,C_1,\ldots,C_p)$

If $\{A_1, \ldots, A_n\} \rightarrow \{B_1, \ldots, B_m\}$
Then the decomposition is lossless

Note: don't need
$\{A_1, \ldots, A_n\} \rightarrow \{C_1, \ldots, C_p\}$

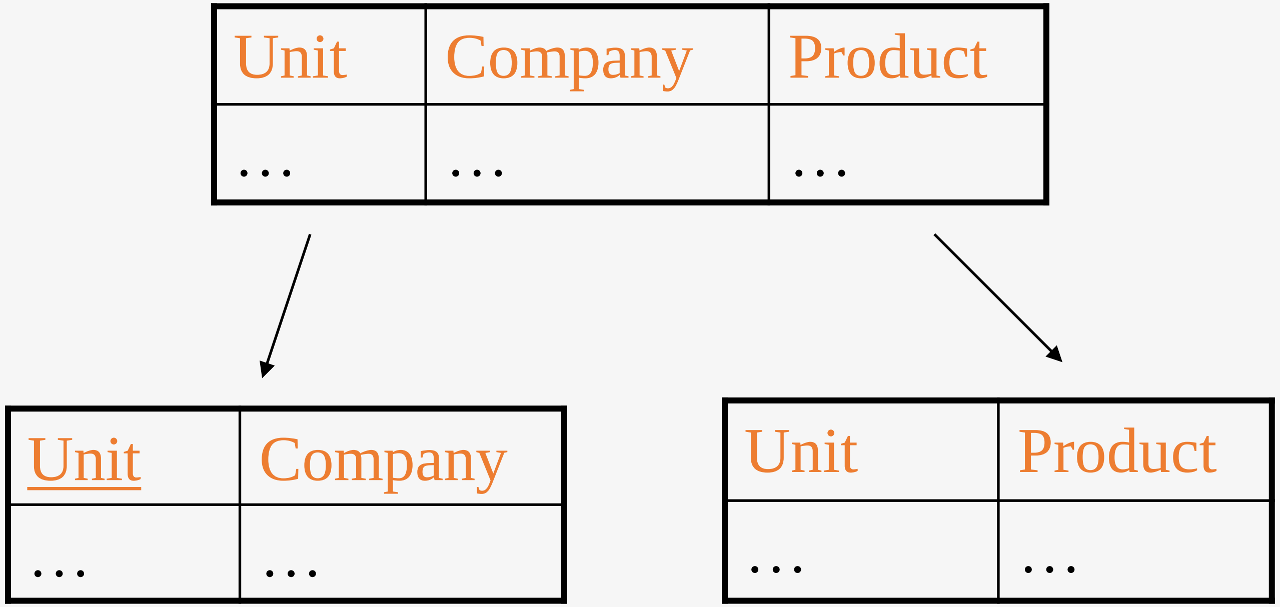
BCNF decomposition is always lossless. Why?

A problem with BCNF

Problem: To enforce a FD, must reconstruct original relation—*on each insert!*

Note: This is historically inaccurate, but it makes it easier to explain

A Problem with BCNF



Unit	Company	Product
…	…	…

Unit	Company
…	…

Unit	Product
…	…

{Unit} → {Company}
{Company,Product} → {Unit}

We do a BCNF decomposition on a “bad” FD: $\{\text{Unit}\}^{+} = \{\text{Unit, Company}\}$

{Unit} → {Company}

We lose the FD {Company,Product} → {Unit}!!

So Why is that a Problem?

Unit	Company
Galaga99	UW
Bingo	UW

Unit	Product
Galaga99	Databases
Bingo	Databases

No problem so far. All *local* FD's are satisfied.

{Unit} → {Company}

Unit	Company	Product
Galaga99	UW	Databases
Bingo	UW	Databases

Let's put all the data back into a single table again:

Violates the FD {Company,Product} → {Unit}!!

The Problem

- We started with a table R and FDs F
- We decomposed R into BCNF tables R_1, R_2, … with their own FDs F_1, F_2, …
- We insert some tuples into each of the relations—which satisfy their local FDs but when reconstruct it violates some FD **across** tables!

Practical Problem: To enforce FD, must reconstruct R—*on each insert!*

Possible Solutions

- Various ways to handle so that decompositions are all lossless / no FDs lost
  - For example 3NF- stop short of full BCNF decompositions. See Bonus Activity!

- Usually a tradeoff between redundancy / data anomalies and FD preservation…

BCNF still most common- with additional steps to keep track of lost FDs…