

# Mineração de Dados

## Aula 4 – parte 1

Especialização em Ciência de Dados e suas Aplicações

Dado um conjunto de transações, encontre regras para a predição da ocorrência de itens baseado na ocorrência de outros itens na transação

## Transações de supermercado

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## Exemplo de regras de associação

$\{\text{Diaper}\} \rightarrow \{\text{Beer}\},$   
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\},$   
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\},$

Implicação significa co-ocorrência, e não causalidade!

- **Itemset**
  - Coleção de um ou mais itens
    - ◆ Ex: {Milk, Bread, Diaper}
  - k-itemset
    - ◆ Itemset que contem k itens
- **Frequência de itemset ( $\sigma$ )**
  - E.g.  $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- **Suporte**
  - Fração de transações que contem um itemset
  - E.g.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- **Itemset frequente**
  - Itemset cujo suporte é maior ou igual a um limite *minsup*

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

# Regra de Associação

Ex:  $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

itemset                      itemset

Arrows point from the word "itemset" to the set {Milk, Diaper} and the set {Beer}.

Suporte (s) de  $A, B, C \longrightarrow D$

$$\frac{\text{número de clientes que compraram } A, B, C, D}{\text{Total de clientes}}$$

Confiança (c) de  $A, B, C \longrightarrow D$

$$\frac{\text{número de clientes que compraram } A, B, C, D}{\text{número de clientes que compraram } A, B, C}$$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

# Regra de Associação

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Exemplo

$$\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$$

$$\text{Suporte} = \{\text{Milk, Diaper, Beer}\} / \text{Total} = 2/5 = 0,4$$

$$\text{Confiança} = \{\text{Milk, Diaper, Beer}\} / \text{Frequência}\{\text{Milk, Diaper}\} = 2/3 = 0,67$$

Dado um conjunto de transações  $T$ , o objetivo da mineração de regras de associação é encontrar todas as regras com

Suporte  $\geq$  **minsup**

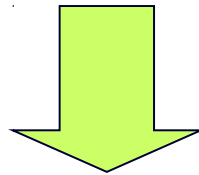
Confiança  $\geq$  **minconf**

## Força bruta:

- Listar todas as regras de associação possíveis
  - Computar o suporte e confiança para cada uma
  - Podar as regras que não atingirem **minsup** e **minconf**
- ⇒ Computacionalmente inviável!

- Princípio do Apriori:

Se um itemset é frequente



Todo subitemset é frequente !!

# Princípio do Apriori

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk



Itens (1-itemsets) -  $C_1$

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Suporte mínimo = 3

Se todo subconjunto é considerado:

$$C_1 + C_2 + C_3$$

$$6 + 15 + 20 = 41$$

Com poda baseada no suporte:

$$6 + 6 + 4 = 16$$



# Princípio do Apriori

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk



Itens (1-itemsets) -  $C_1$

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Suporte mínimo = 3

Se todo subconjunto é considerado:

$$C_1 + C_2 + C_3$$

$$6 + 15 + 20 = 41$$

Com poda baseada no suporte:

$$6 + 6 + 4 = 16$$

# Princípio do Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Itens (1-itemsets) -  $C_1$



Itemset
{Bread, Milk}
{Bread, Beer }
{Bread, Diaper}
{Beer, Milk}
{Diaper, Milk}
{Beer, Diaper}

Pares (2-itemsets) -  $C_2$

(Não é necessário gerar candidatos com coke ou eggs)

Suporte mínimo = 3

Se todo subconjunto é considerado:

$$C_1 + C_2 + C_3$$

$$6 + 15 + 20 = 41$$

Com poda baseada no suporte:

$$6 + 6 + 4 = 16$$

# Princípio do Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Itens (1-itemsets) -  $C_1$



Itemset	Count
{Bread,Milk}	3
{Beer, Bread}	2
{Bread,Diaper}	3
{Beer,Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Pares (2-itemsets) -  $C_2$

(Não é necessário gerar candidatos com coke ou eggs)

Suporte mínimo = 3

Se todo subconjunto é considerado:

$$C_1 + C_2 + C_3$$

$$6 + 15 + 20 = 41$$

Com poda baseada no suporte:

$$6 + 6 + 4 = 16$$

# Princípio do Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Itens (1-itemsets) -  $C_1$



Itemset	Count
{Bread, Milk}	3
{Bread, Beer}	2
{Bread, Diaper}	3
{Milk, Beer}	2
{Milk, Diaper}	3
{Beer, Diaper}	3

Pares (2-itemsets) -  $C_2$

(Não é necessário gerar candidatos com coke ou eggs)



Trios (3-itemsets) -  $C_3$

Itemset
{ Beer, Diaper, Milk }
{ Beer, Bread, Diaper }
{ Bread, Diaper, Milk }
{ Beer, Bread, Milk }

Suporte mínimo = 3

Se todo subconjunto é considerado:

$$C_1 + C_2 + C_3$$

$$6 + 15 + 20 = 41$$

Com poda baseada no suporte:

$$6 + 6 + 4 = 16$$

# Princípio do Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Itens (1-itemsets) -  $C_1$



Itemset	Count
{Bread, Milk}	3
{Bread, Beer}	2
{Bread, Diaper}	3
{Milk, Beer}	2
{Milk, Diaper}	3
{Beer, Diaper}	3

Pares (2-itemsets) -  $C_2$

(Não é necessário gerar candidatos com coke ou eggs)

Suporte mínimo = 3

Se todo subconjunto é considerado:

$$C_1 + C_2 + C_3$$

$$6 + 15 + 20 = 41$$

Com poda baseada no suporte:

$$6 + 6 + 4 = 16$$



Trios (3-itemsets) -  $C_3$

Itemset	Count
{ Beer, Diaper, Milk }	2
{ Beer, Bread, Diaper }	2
{ Bread, Diaper, Milk }	2
{ Beer, Bread, Milk }	1

$C_1$  = Itemsets de tamanho 1

$F_1$  = Itemsets frequentes de  $C_1$

$k = 1$

Enquanto  $F_k$  não for vazio

$C_{k+1} = \text{Gerar}(F_k)$

$C_{k+1} = \text{Podar}(C_k, F_k)$

$F_{k+1} = \text{Validar}(BD, C_{k+1})$

$k = k + 1$

## □ Cálculo de $F(1)$

$T1 = \{ \text{Pao, Leite, Manteiga} \}$

$T2 = \{ \text{Pao, Leite, Acucar} \}$

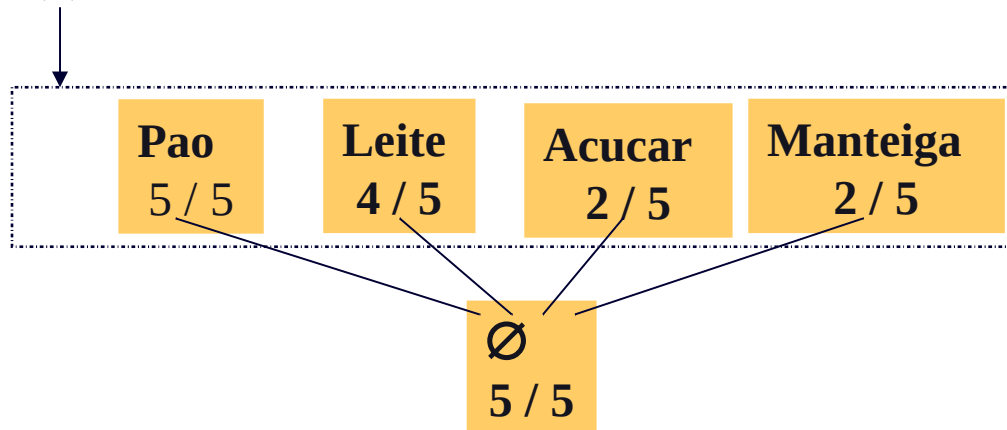
$T3 = \{ \text{Pao} \}$

$T4 = \{ \text{Pao, Leite} \}$

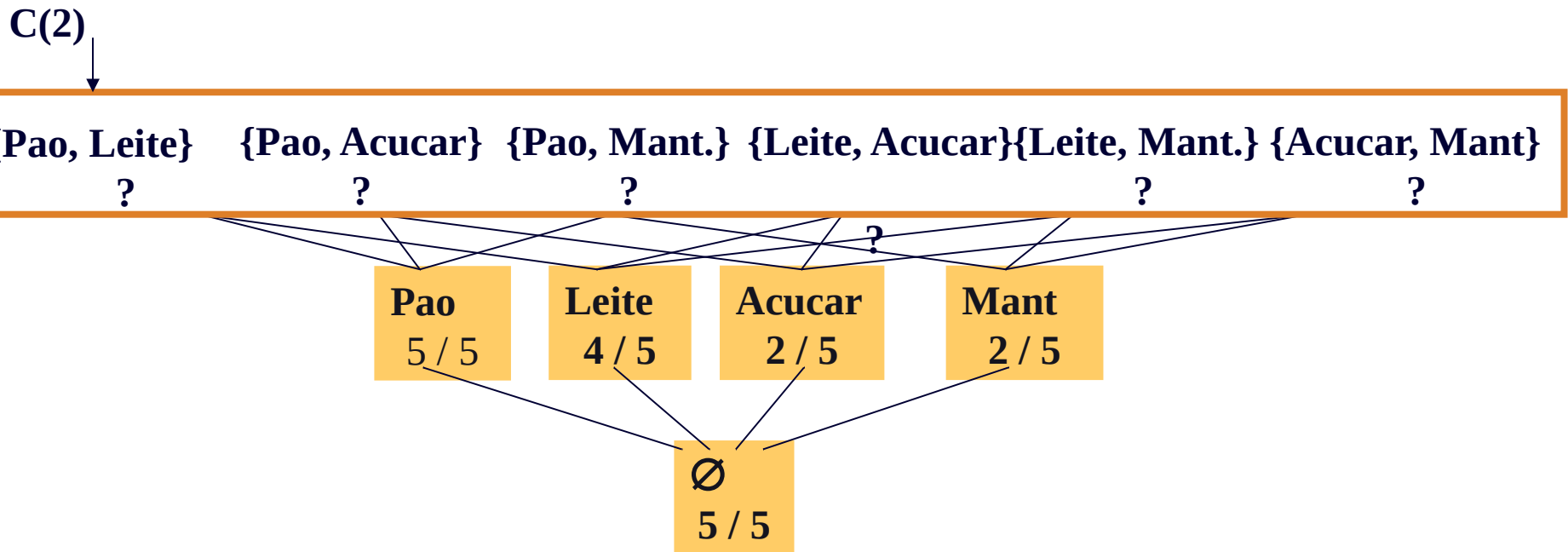
$T5 = \{ \text{Pao, Leite, Manteiga, Acucar} \}$

$\text{minsup} = 2 / 5$

$F(1)$



- Combinação dos elementos de F(1)
- Poda dos elementos de C(2) – nenhuma neste nível

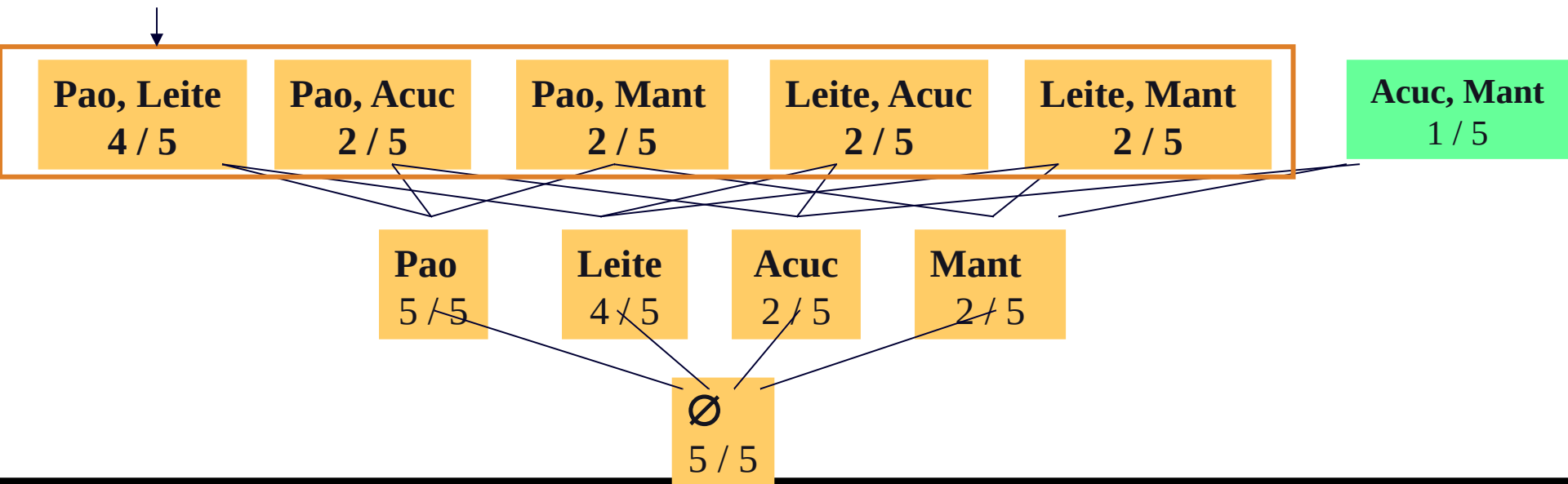




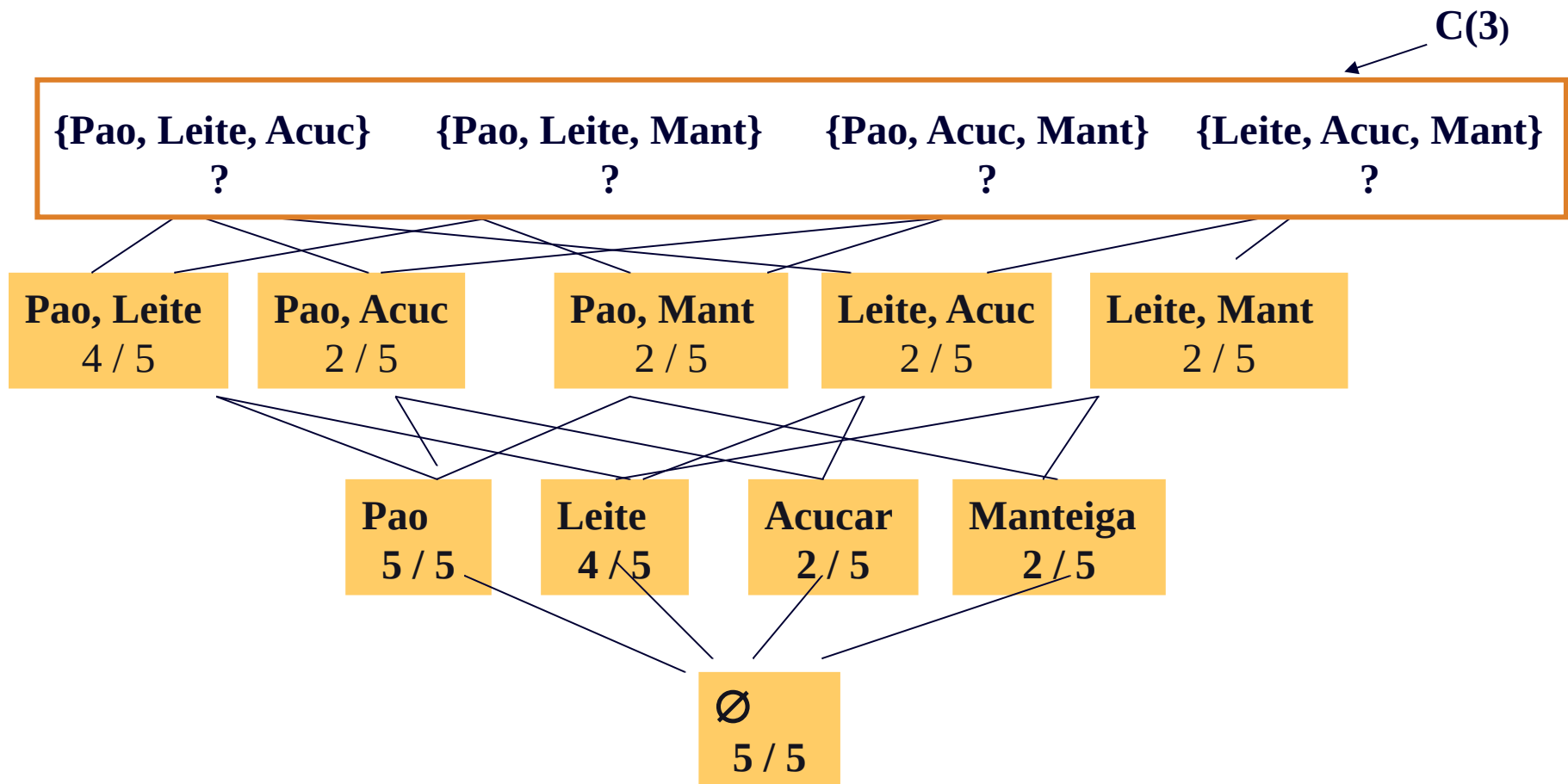
□ Cálculo de  $F(2)$   
Numa varrida dos dados

T1 = { Pao, Leite, Mant }  
T2 = { Pao, Leite, Acuc }  
T3 = { Pao }  
T4 = { Pao, Leite }  
T5 = { Pao, Leite, Mant, Acuc }  
minsup = 2 / 5

$F(2)$



Combinar somente os itemsets cujos primeiros elementos são idênticos



# Poda de C(3)

C(3)



{Pao, Leite, Acuc}

?

{Pao, Leite, Mant}

?

{Pao, ~~Acuc~~, Mant}

?

{Leite, ~~Acuc~~, Mant}

?

Pao, Leite

4 / 5

Pao, Acuc

2 / 5

Pao, Mant

2 / 5

Leite, Acuc

2 / 5

Leite, Mant

2 / 5

Pao

5 / 5

Leite

4 / 5

Acuc

2 / 5

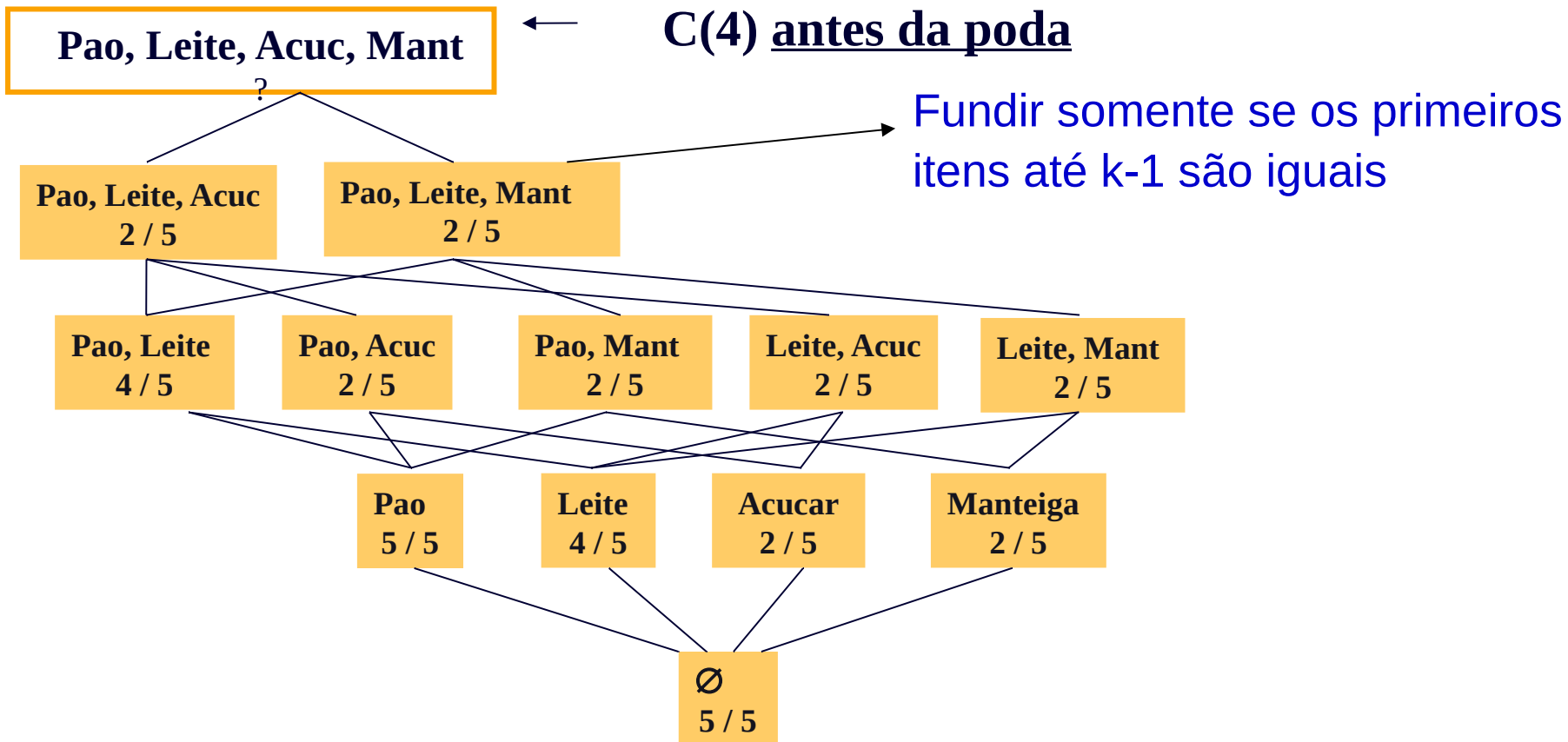
Mant

2 / 5

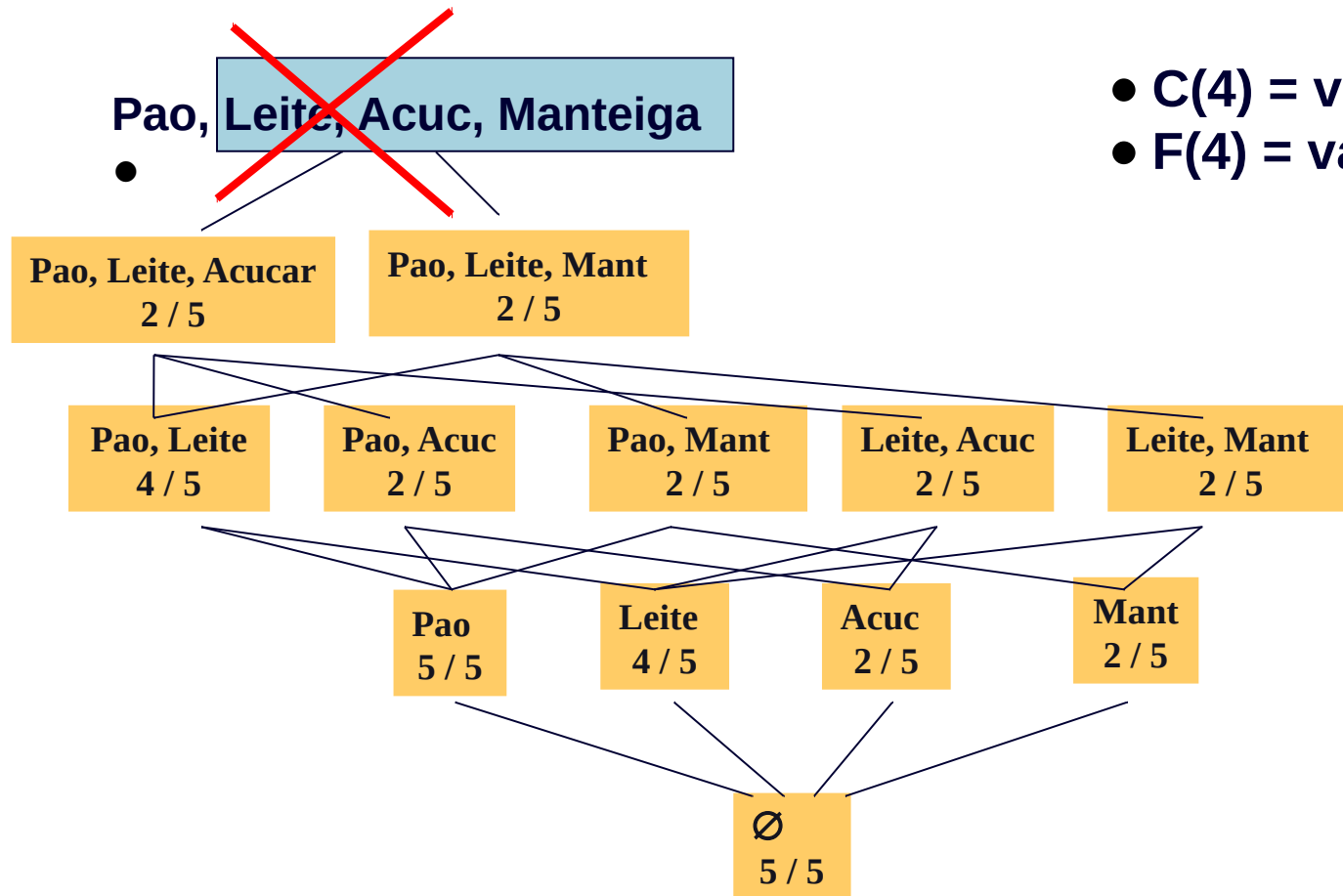
∅

5 / 5

## □ Combinação dos elementos de F(3)



# Poda de C(4)



- C(4) = vazio
- F(4) = vazio

- Para todo k-itemset **X** frequente  
(com  $k > 1$ ) e  $Y \subset X$ 
  - Calcular a *confiança* da regra de associação  
 $X - Y \Rightarrow Y$
  - Se é superior ou igual a *minconf*, **então** a regra gerada é **interessante**


$$\text{Minconf} = 3 / 4$$

$X = \{ \text{Pao, Leite, Mant} \}$  é frequente:  $\text{suporte}(X) = 2/5$

- ☐ Pao, Leite  $\Rightarrow$  Mant **não interessante**  
 $\text{sup}(\text{Pao, Leite}) = 4/5$  ,  $\text{Conf} = (2/5) / (4/5) = 2/4$
- ☐ Pao, Mant  $\Rightarrow$  Leite **interessante** pois  $\text{Conf} = 2 / 2$
- ☐ Leite, Mant  $\Rightarrow$  Pain **interessante** pois  $\text{Conf} = 2 / 2$
- ☐ Pao  $\Rightarrow$  Lait, Mant **não interessante** pois  $\text{Conf} = 2 / 5$
- ☐ Leite  $\Rightarrow$  Pao , Mant **não interessante** pois  $\text{Conf} = 2 / 4$
- ☐ Mant  $\Rightarrow$  Pao , Leite **não interessante** pois  $\text{Conf} = 2 / 4$

$$\text{lift} = \frac{\text{confiança da regra}}{\text{suporte do consequente}}$$

Parte “direita”  
da regra



Regra com **lift=1**: a probabilidade de ocorrer o antecessor e o consequente é independente um do outro

**Regra não interessante**

Se A -> leite

**Confiança** = 90%

**Suporte de leite** = 90% (aparece 90% das transações)

**Lift = 1: não interessante.** A probabilidade de A e leite são independentes. Isto significa que leite pode ser comprado muito frequentemente, independente de A



$$\text{lift} = \frac{\text{confiança da regra}}{\text{suporte do consequente}}$$

Regra com **lift>1**

**Se C -> D**

**Confiança = 70%**

**Suporte de D = 10%**

**Lift = 7**

Criar regras de associação com os personagens do filme Titanic.

Instalar o pacote: arules

Carregar os dados: "titanic.raw.rdata"

Executar o script em R fornecido.

Parte deste material é derivado do livro:  
Introduction to Data Mining - Tan, Steinbach,  
Kumar