

MINERÍA DE DATOS

Maximiliano Ojeda

muojeda@uc.cl



IIC-2433

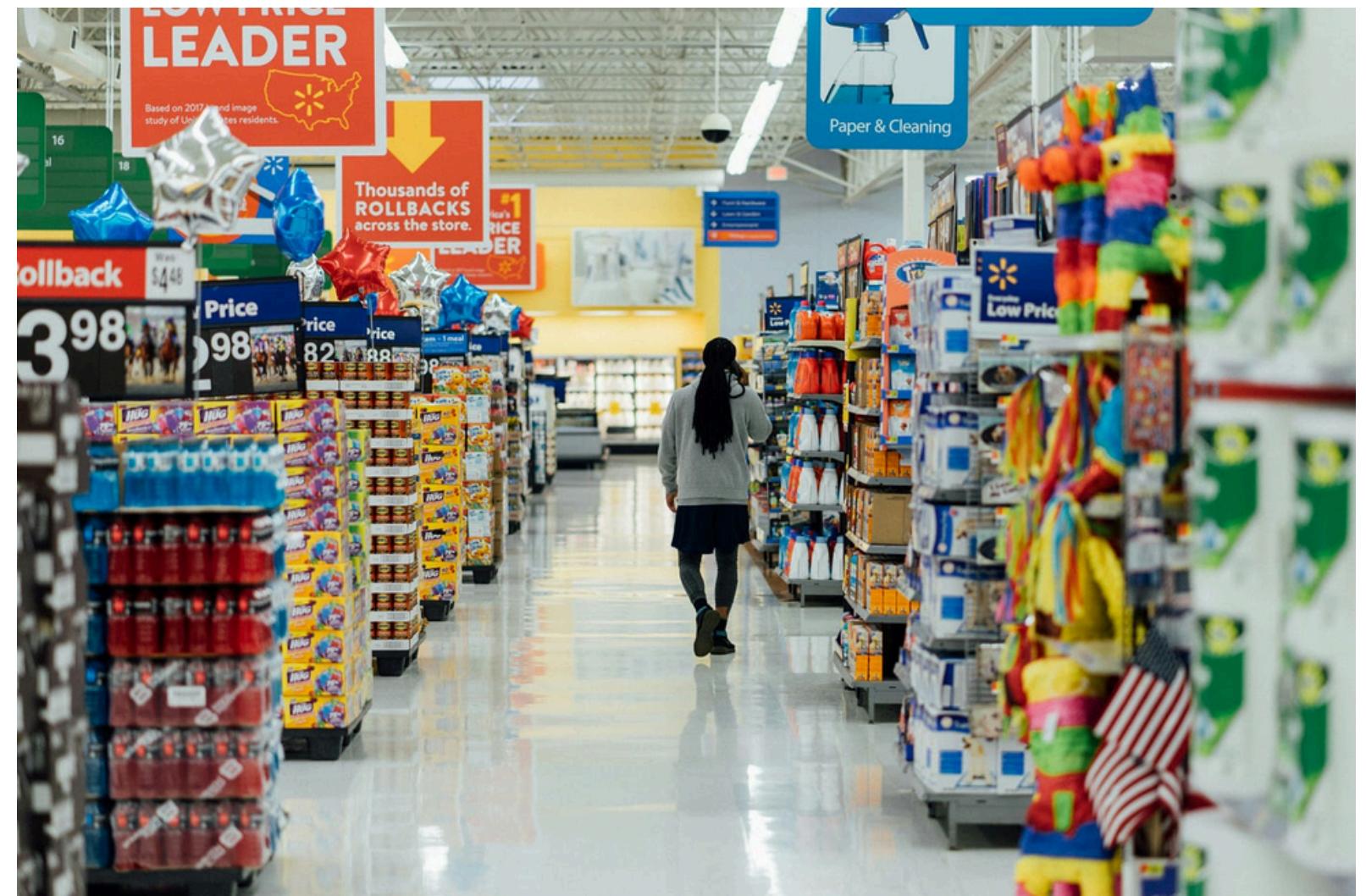


Reglas de Asociación

Reglas de Asociación

Compras en supermercados

- Todos los días se registran miles de compras en los supermercados.
- Personas compran: pan, queso, mantequilla, etc.
- ¿Existen patrones ocultos detrás de estas compras?



Beers & Diapers

Esta historia es clásica en DM, basada en realidades y mitos

- Se analizaron millones de tickets de compras en WalMart
- Los hombres jóvenes de entre 25 y 35 años, que compraban pañales los jueves o viernes, también solían comprar cerveza.
- Al investigar el patrón, se propuso que muchos eran padres jóvenes que hacían las compras del hogar al salir del trabajo, antes del fin de semana.



Reglas de Asociación

{Diapers} → {Beers}

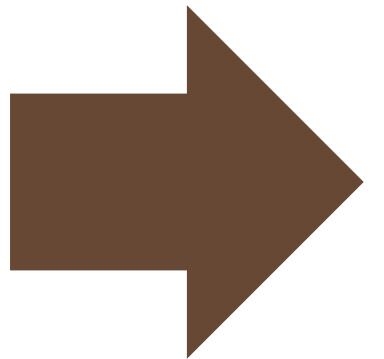
- Indica que si ocurre el antecedente, es probable que ocurra el consecuente
- Usualmente son variables categóricas nominales.
- Se usan {} para indicar asociaciones entre conjuntos denominados itemsets
- La cardinalidad de los conjuntos puede ser mayor a uno:

{Pan, Mantequilla} → {Queso}

{Queso, Jamón} → {Pan}

Reglas de Asociación

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}



ID	Bread	Milk	Daipers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Reglas de Asociación

ID	Bread	Milk	Daipers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Items

$$I = \{i_1, i_2, i_3, \dots, i_d\}$$

Para la tabla de ejemplo, la cantidad **d** corresponde a la cantidad de columnas

Transacciones

$$T = \{t_1, t_2, t_3, \dots, t_N\}$$

N corresponde a la cantidad de transacciones o filas

Soporte de un itemset

$$\sigma(X) = |\{t_i \mid X \subseteq t_i, t_i \in T\}|$$

Reglas de Asociación

Una regla de asociación es una implicancia $X \rightarrow Y$ tal que $X \cap Y = \emptyset$

Support

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

Frecuencia con que aparece la combinación $X \cup Y$

Confidence

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Probabilidad de comprar Y dado que se compró X

Reglas de Asociación

Estrategia para generación de reglas

1. Búsqueda de itemsets frecuentes
2. Generación de reglas candidatas a partir de itemsets frecuentes

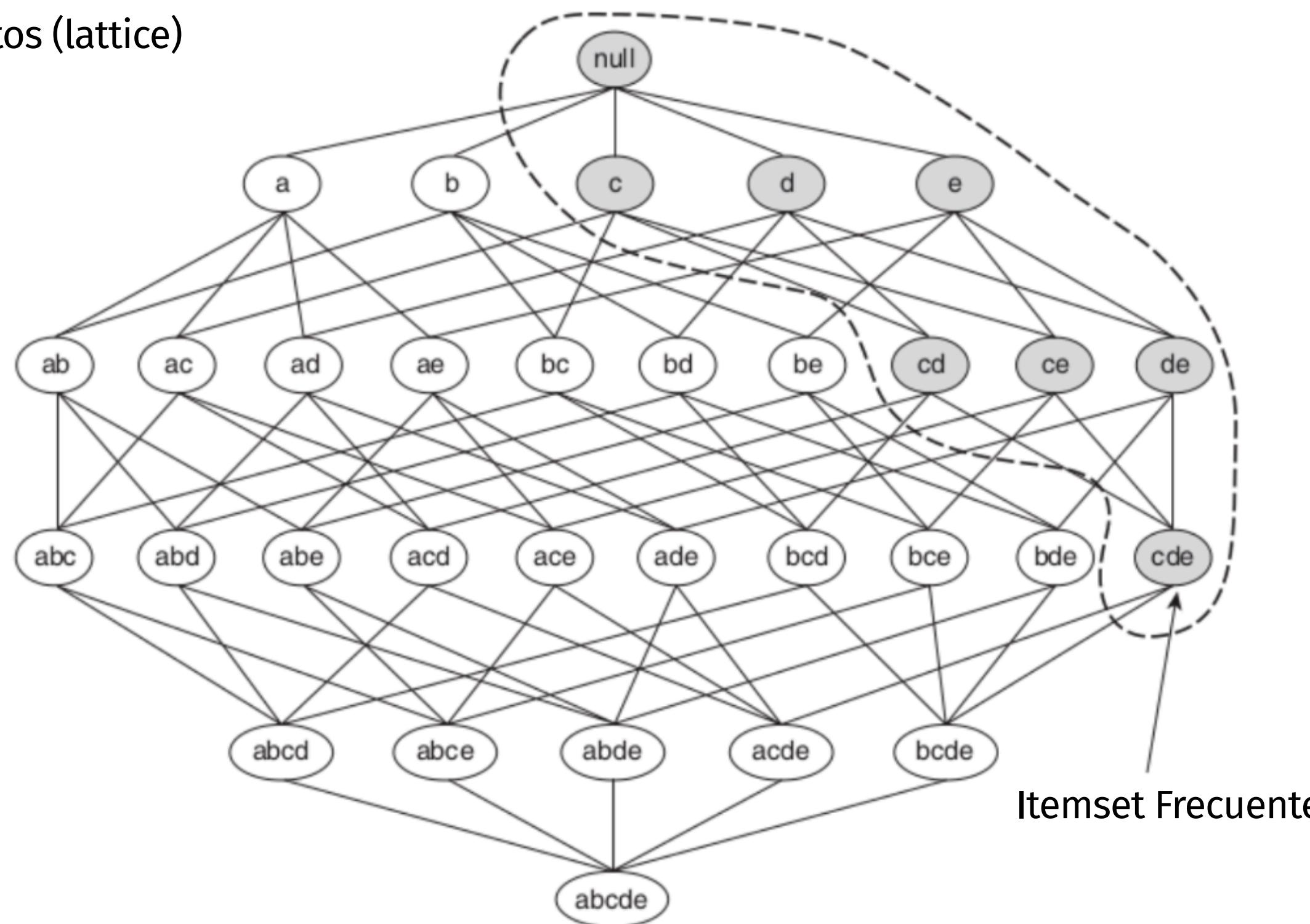
Buscar asociaciones puede ser muy ineficiente. La eficiencia de la búsqueda es un tema crítico en reglas de asociación

Principio Apriori

Si un itemset es frecuente, entonces todos sus subconjuntos son frecuentes

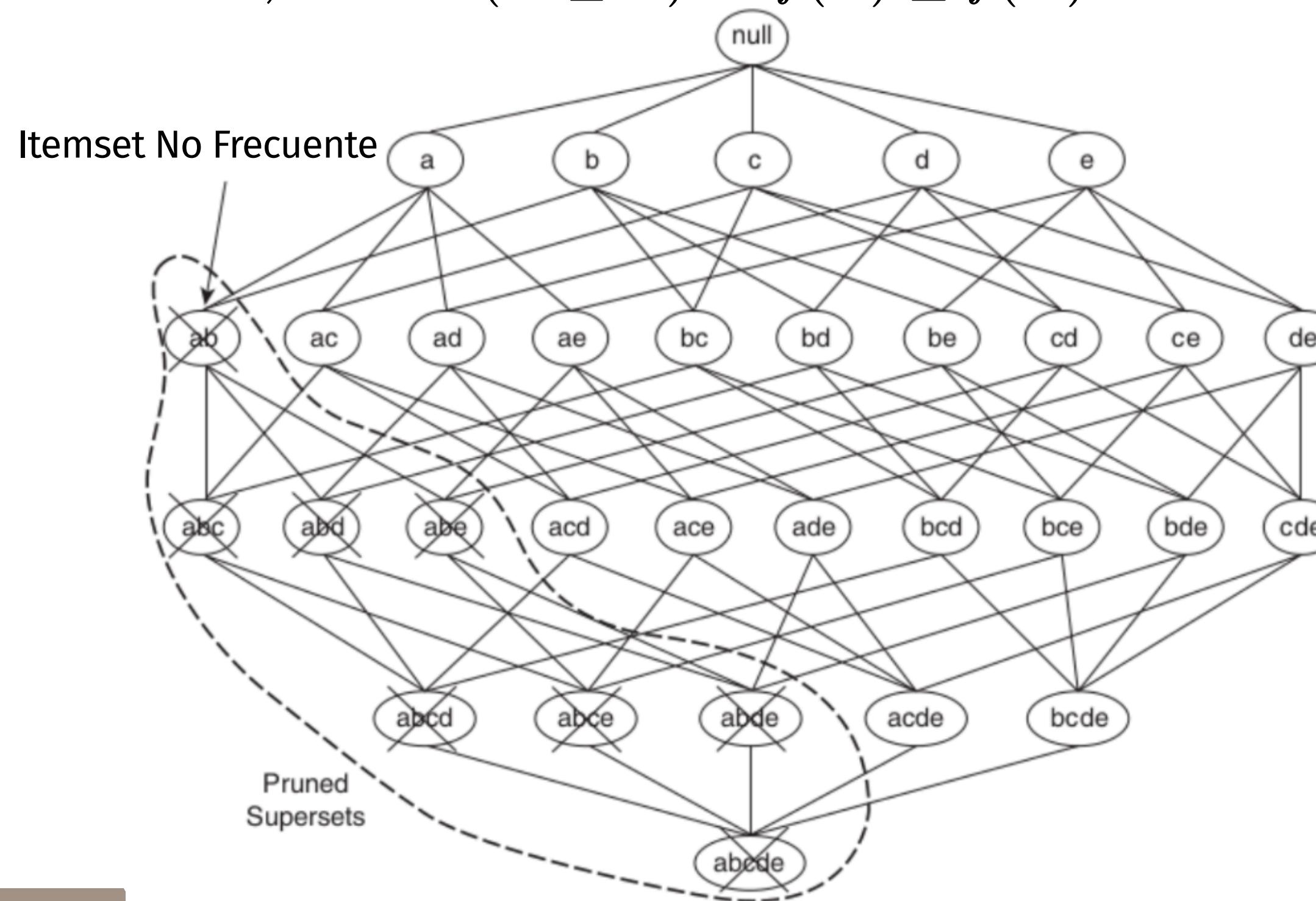
Reglas de Asociación

Estructura de subconjuntos (lattice)



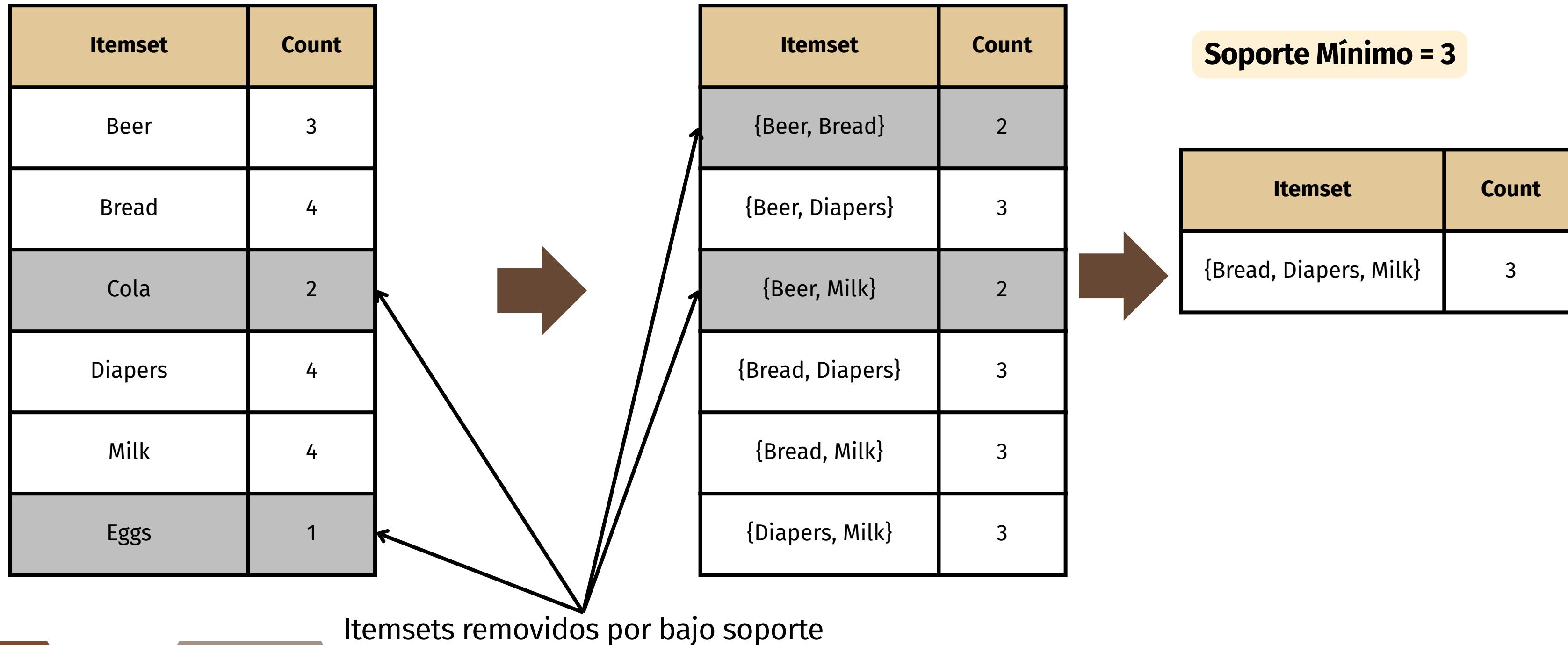
Reglas de Asociación

Principio de Monotonidad $\forall X, Y \in J : (X \subseteq Y) \rightarrow f(Y) \leq f(X)$



Reglas de Asociación

Principio de Monotocidad: Buscamos por orden creciente de cardinalidad



Reglas de Asociación

Ahora vamos a generar reglas a partir de itemsets frecuentes

Support

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

Confidence

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Si una regla $X \rightarrow Y - X$ no satisface el umbral de confianza, entonces toda regla de la forma $\tilde{X} \rightarrow Y - \tilde{X}$, donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza

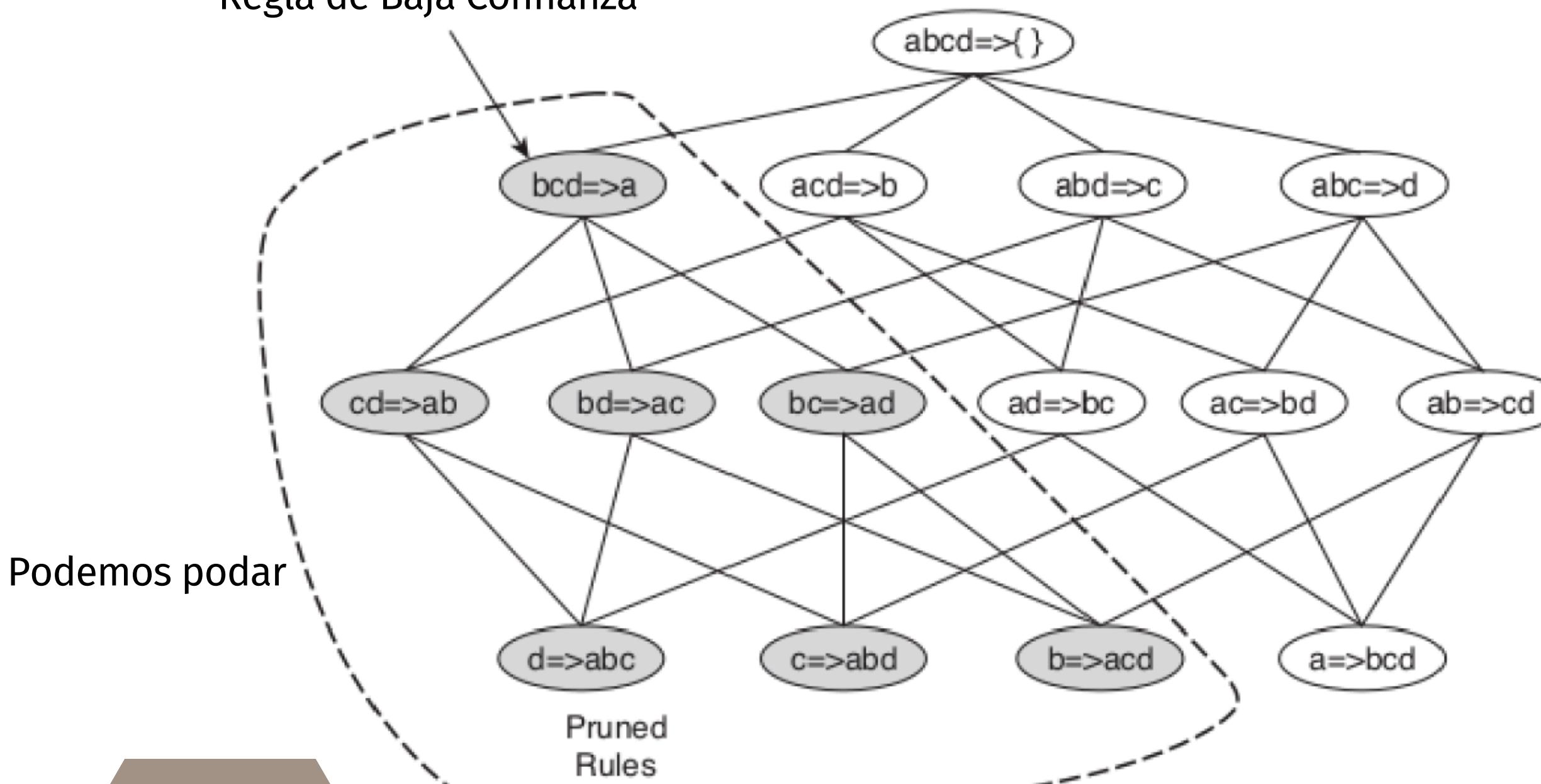
Principio de Monotocidad $\forall X, Y \in J : (X \subseteq Y) \rightarrow f(Y) \leq f(X)$

$$\begin{aligned} \tilde{X} \subseteq X &\quad \Rightarrow \quad \sigma(X) \leq \sigma(\tilde{X}) \\ &\quad \Rightarrow \quad \frac{\sigma(X \cup Y)}{\sigma(X)} \geq \frac{\sigma(X \cup Y)}{\sigma(\tilde{X})} \end{aligned}$$

Reglas de Asociación

Si una regla $X \rightarrow Y - X$ no satisface el umbral de confianza, entonces toda regla de la forma $\tilde{X} \rightarrow Y - \tilde{X}$, donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza

Regla de Baja Confianza



Mlxtend

```

from mlxtend.frequent_patterns import apriori, association_rules
import pandas as pd

# Dataset con 6 ítems: Pan, Mantequilla, Leche, Huevos, Cerveza, Cola
data = {
    'Pan': [1,1,0,1,1,0,1,1],
    'Mantequilla': [1,1,1,0,1,0,1,0],
    'Leche': [0,1,1,1,1,1,0,1],
    'Huevos': [0,0,1,0,0,1,1,1],
    'Cerveza': [0,1,1,1,0,0,0,1],
    'Cola': [0,0,1,1,1,1,0,0],
}
df = pd.DataFrame(data).astype(bool)

# 1) Ítems frecuentes
frequent_items = apriori(df, min_support=0.35, use_colnames=True)
print(frequent_items)

# 2) Reglas de asociación
rules = association_rules(frequent_items, metric="confidence", min_threshold=0.6)
print(rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']])

```

	support	itemsets
0	0.750	(Pan)
1	0.625	(Mantequilla)
2	0.750	(Leche)
3	0.500	(Huevos)
4	0.500	(Cerveza)
5	0.500	(Cola)
6	0.500	(Mantequilla, Pan)
7	0.500	(Pan, Leche)
8	0.375	(Pan, Cerveza)
9	0.375	(Mantequilla, Leche)
10	0.375	(Huevos, Leche)
11	0.500	(Cerveza, Leche)
12	0.500	(Cola, Leche)
13	0.375	(Pan, Cerveza, Leche)

	antecedents	consequents	support	confidence	lift
14	(Cerveza)	(Pan, Leche)	0.375	0.750000	1.500000
12	(Pan, Leche)	(Cerveza)	0.375	0.750000	1.500000
11	(Pan, Cerveza)	(Leche)	0.375	1.000000	1.333333
8	(Leche)	(Cerveza)	0.500	0.666667	1.333333
7	(Cerveza)	(Leche)	0.500	1.000000	1.333333
9	(Cola)	(Leche)	0.500	1.000000	1.333333
10	(Leche)	(Cola)	0.500	0.666667	1.333333
0	(Mantequilla)	(Pan)	0.500	0.800000	1.066667
1	(Pan)	(Mantequilla)	0.500	0.666667	1.066667
4	(Cerveza)	(Pan)	0.375	0.750000	1.000000
6	(Huevos)	(Leche)	0.375	0.750000	1.000000
13	(Cerveza, Leche)	(Pan)	0.375	0.750000	1.000000
2	(Pan)	(Leche)	0.500	0.666667	0.888889
3	(Leche)	(Pan)	0.500	0.666667	0.888889
5	(Mantequilla)	(Leche)	0.375	0.600000	0.800000

MINERÍA DE DATOS

Maximiliano Ojeda

muojeda@uc.cl



IIC-2433