

# DIPLOMADO EN CIENCIA DE DATOS

Módulo: Minería de Datos  
Reglas de asociación

Universidad Nacional de Colombia

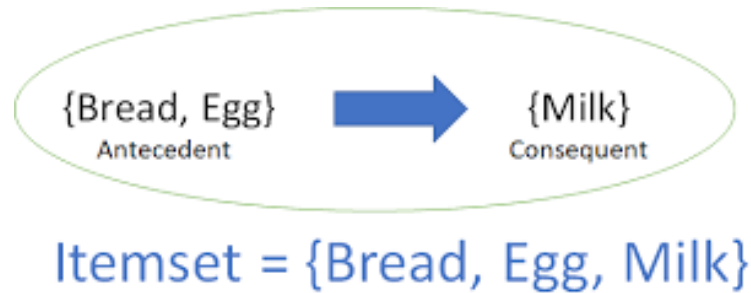
# Contenido

- Reglas de asociación
- Soporte, confianza, lift, convicción
- Algoritmo apriori
- Dificultades, filtros y ajustes
- Extensiones

# Reglas de asociación

Son un método para encontrar relaciones o patrones “interesantes” en transacciones

- Interesante, frecuente, raro, extraño?
- Identificar reglas fuertes en conjuntos de datos usando alguna medida de qué tan interesantes son

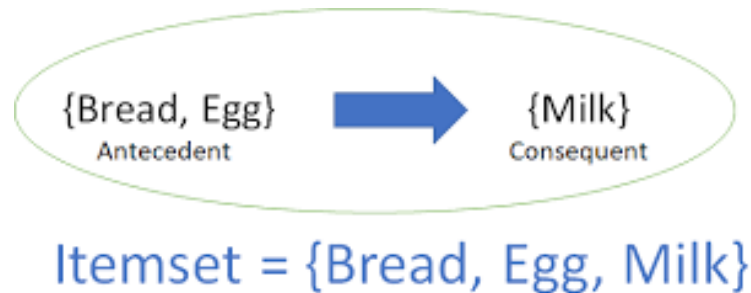


- Ejemplo **{cebolla, tomates, salsa de tomate}** → **{hamburguesa}**  
Cómo usar esta información para el beneficio del negocio?

# Reglas de asociación

Son un método para encontrar relaciones o patrones “interesantes” en transacciones

- Aplicaciones en market basket analysis, uso de páginas web, producción de productos.
- Generalmente no tiene en cuenta el orden de los items (sequence mining si)



**Lo que buscamos son co-ocurrencias, no causalidad!**

# {cerveza, pañales}?

<https://www.itbusiness.ca/news/behind-the-beer-and-diapers-data-mininglegend/>

136



Market basket analysis de 1.2 millones de transacciones de 25 Osco Drug stores  
Encontraron 30 patrones “interesantes” {cerveza, pañales} {jugo de fruta,  
jarabe para la tos}

# {cremas, calcio, zinc}?

<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#bf5973866686>



Análisis de compras de mujeres embarazadas

**{calcio, magnesio, zinc} → embarazo**

**{jabón sin fragancia, algodón, gel antibacterial} → ya casi nace el bebé**

Han identificado 25 productos que permiten asignar un puntaje de “predicción de embarazo”  
Target estima una fecha de Nacimiento para enviar promociones y recomendaciones

## ¿Y qué es una regla de asociación?

## Análisis de Reglas de asociación:

Es un método para descubrir patrones de relaciones entre ítems con alguna medida de interés

Ejemplo: {Leche, pañales}  $\rightarrow$  {Cerveza}

Entonces ¿cuándo un ítem es frecuente, cuándo un conjunto de ítems es frecuente?



- An algorithm behind “You may also like”

# ¿Y qué es una regla de asociación?

## Regla de asociación:

Una regla de asociación es una afirmación basada en evidencia de que con frecuencia cuando ocurre  $x$ , también ocurre  $y$ .

$x \rightarrow y$ , dónde  $x$  y  $y$  son conjuntos de ítems

Nota: Un **conjunto** puede estar conformado por uno o más ítems

Necesitamos los conceptos de:

- Base transaccional
- Ítems
- Conjunto de ítems (canastas)



-An algorithm behind  
"You may also like"



# Bases de datos transaccionales

Cada fila en la tabla representa una transacción  
Cada columna es un atributo (binario)

Tr. ID milk bread beer cheese wine spaghetti

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1



# Definamos que es un item y que es un basket o conjunto de items

Item

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

**Basket, canasta,  
transacción**

Conjunto de ítems

{leche}

{leche, pan}

{vino, leche, pasta}



## ¿Cómo extraer una regla “interesante”?

¿Qué significa tener una “buena” regla?

Para seleccionar reglas utilizamos medidas que cuantifican su nivel de interés/relevancia

Las más utilizadas son **soporte** (support) y **confianza** (confidence).

Hay más (lift, conviction)

# Soporte

$$\text{supp}(x) = \frac{\text{Número de transacciones en las que } x \text{ aparece}}{\text{Número total de transacciones}}$$

Usualmente se describe como un porcentaje

Rango: [0,1]

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

$$\text{supp}(\text{pan}) = \frac{4}{5} = 0.8$$

El ítem pan aparece en el 80% de las transacciones

=

El ítem pan tiene un soporte del 80%

# Confianza

$$conf(x \rightarrow y) = \frac{supp(x \cup y)}{supp(x)}$$

Describe que tan probable es que el ítem  $y$  sea comprado, cuando(dado que) el ítem  $x$  ha sido comprado

También se puede interpretar como una estimación de la probabilidad condicional  $P(y|x)$   
Rango:  $[0,1]$

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

$$conf(pan \rightarrow leche) = \frac{3/5}{4/5} = 0.75$$

La probabilidad de que se compre leche dado que se compró pan es de 0.75

# Lift

$$lift(x \rightarrow y) = \frac{supp(x \cup y)}{supp(x) * supp(y)}$$

Podría definirse como una métrica mejorada de la confianza: La confianza tiene en cuenta la frecuencia (o popularidad) de  $x$  pero no la de  $y$ . En cambio el lift tiene en cuenta las dos

Describe que tan probable es que el ítem  $y$  sea comprado cuando el ítem  $x$  ha sido comprador, teniendo en cuenta la popularidad de  $y$

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

# Lift

$$lift(x \rightarrow y) = \frac{supp(x \cup y)}{supp(x) * supp(y)}$$

- $Lift > 1$  Es probable que el ítem  $y$  sea comprado con el ítem  $x$   
 $Lift < 1$  Es poco probable que el ítem  $y$  sea comprado con el ítem  $x$   
 $Lift = 1$  La ocurrencia de los ítems es independiente  
Rango:  $[0,inf]$

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

$$lift(pan \rightarrow leche) = \frac{\frac{3}{5}}{\left(\frac{4}{5}\right) * \left(\frac{3}{5}\right)} = 1.25$$

$$lift(pan \rightarrow leche) = lift(leche \rightarrow pan)$$

# Convicción

$$conv(x \rightarrow y) = \frac{1 - supp(y)}{1 - conf(x \rightarrow y)}$$

Mide el independencia de los ítems en la regla. Compara la probabilidad de que aparezca  $x$  sin  $y$ , si fueran dependientes.

Rango: [0,inf]

1 significa independencia, un mayor valor de convicción significa mayor dependencia entre  $y$  y  $x$

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

$$conv(pan \rightarrow leche) = \frac{1 - (\frac{3}{5})}{1 - 0.75} = 1.6$$

Este valor indica que la regla sería errónea 60% veces mas si la asociación entre pan y leche fuera por pura casualidad.



# ¿Cómo extraer una regla “interesante”?

## Soporte

$$\textit{supp}(x) = \frac{\textit{Número de transacciones en las que } x \textit{ aparece}}{\textit{Número total de transacciones}}$$

## Confianza

$$\textit{conf}(x \rightarrow y) = \frac{\textit{supp}(x \cup y)}{\textit{supp}(x)}$$

## Lift

$$\textit{lift}(x \rightarrow y) = \frac{\textit{supp}(x \cup y)}{\textit{supp}(x) * \textit{supp}(y)}$$

# El algoritmo apriori– lo que hay detrás

Encontrar todos los conjuntos de ítems frecuentes en una base de datos puede ser difícil ya que hay que evaluar todas las posibles combinaciones de conjuntos

Se pueden formar hasta  $2^I - 1$  conjuntos, (I: número total de ítems)



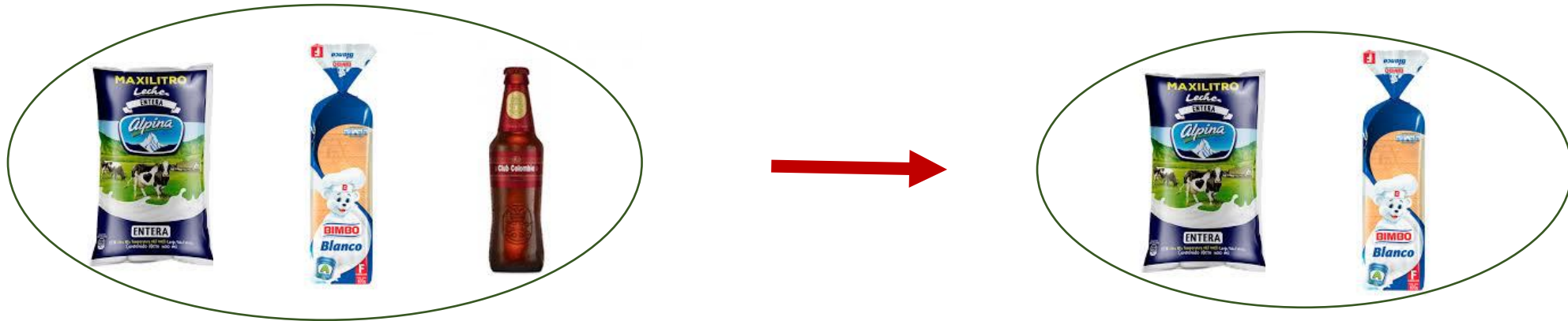
$$2^{6-1} = 32 \text{ conjuntos de ítems}$$

$$2^{7-1} = 64$$

Este algoritmo nos ayuda a seleccionar las reglas eficientemente

# El algoritmo apriori – lo que hay detrás

Para hacerlo eficientemente se tiene que **cuando un conjunto de ítems es frecuente, todos sus subconjuntos de ítems son también frecuentes** (downward-closure)



El soporte de un conjunto de ítems nunca va a exceder el soporte de sus subconjuntos

Entonces si un **conjunto de ítems NO es frecuente, todo “super conjunto” (conjuntos que lo contengan) tampoco serán frecuentes**


# El algoritmo apriori – dos pasos

## ¿Cómo se define una regla?

1. Se define un valor **mínimo de soporte** para encontrar todos los **conjuntos de ítems frecuentes**
2. Se define un valor **mínimo de confianza** para estos conjuntos frecuentes para **formar reglas de asociación frecuentes**

# El algoritmo apriori – paso 1

## ¿Cómo evaluar qué conjuntos son frecuentes?

- Generar los conjuntos frecuentes de tamaño  $k=1$  basado en soporte
  - Generar los candidatos de conjuntos de tamaño  $k+1$
  - Descartar los candidatos que contienen subconjuntos de tamaño  $k$  que no son frecuentes
  - Calcular el soporte para cada candidato, considerar aquellos que superan el umbral
- 

Repetir hasta no encontrar conjuntos que superen el umbral

# El algoritmo apriori – paso 1

tr_id	leche	pan	cerveza	queso	vino	pasta
1	1	1	1	0	0	0
2	0	1	1	1	1	0
3	1	1	0	1	0	1
4	0	0	0	1	1	1
5	1	1	0	1	1	1

Definimos un soporte  $\geq 50\%$

Empezamos con los conjuntos de tamaño  $k=1$

Paso a paso, sólo necesitamos expandir los conjuntos que están por encima del umbral, con ítems que también están por encima del umbral.

{leche} (60%)  
{pan} (80%)  
{cerveza} (40%)  
{queso} (80%)  
{vino} (60%)  
{pasta} (60%)

k=1



{leche, pan} (60%)  
{leche, queso} (40%)  
{leche, vino} (20%)  
{leche, pasta} (40%)  
{pan, queso} (60%)  
{pan, vino} (40%)  
{pan, pasta} (40%)  
{queso, vino} (60%)  
{queso, pasta} (60%)  
{vino, pasta} (40%)

k=2



{queso, leche, pan} (40%)  
{vino, pan, queso} (20%)  
{pan, leche, pasta} (40%)  
{vino, pan, queso} (40%)  
{pasta, pan, queso} (40%)  
{leche, queso, vino} (20%)  
{pasta, queso, vino} (40%)  
{pasta, queso, vino} (40%)

k=3

## El algoritmo apriori – paso 2

Se define un umbral de confianza para los conjuntos de ítems frecuentes para formar las reglas de asociación

- Con cada conjunto frecuente, generar todos los subconjuntos no vacíos
- Para cada subconjunto calcular su confianza y validarla con el umbral definido

## El algoritmo a priori – paso 2

Ejemplo: Supongamos que el conjunto **{queso, vino, pasta}** es frecuente, validamos la confianza de

$\{\text{queso, vino}\} \rightarrow \{\text{pasta}\}$

$\{\text{queso, pasta}\} \rightarrow \{\text{vino}\}$

$\{\text{vino, pasta}\} \rightarrow \{\text{queso}\}$

$\{\text{queso}\} \rightarrow \{\text{vino, pasta}\}$

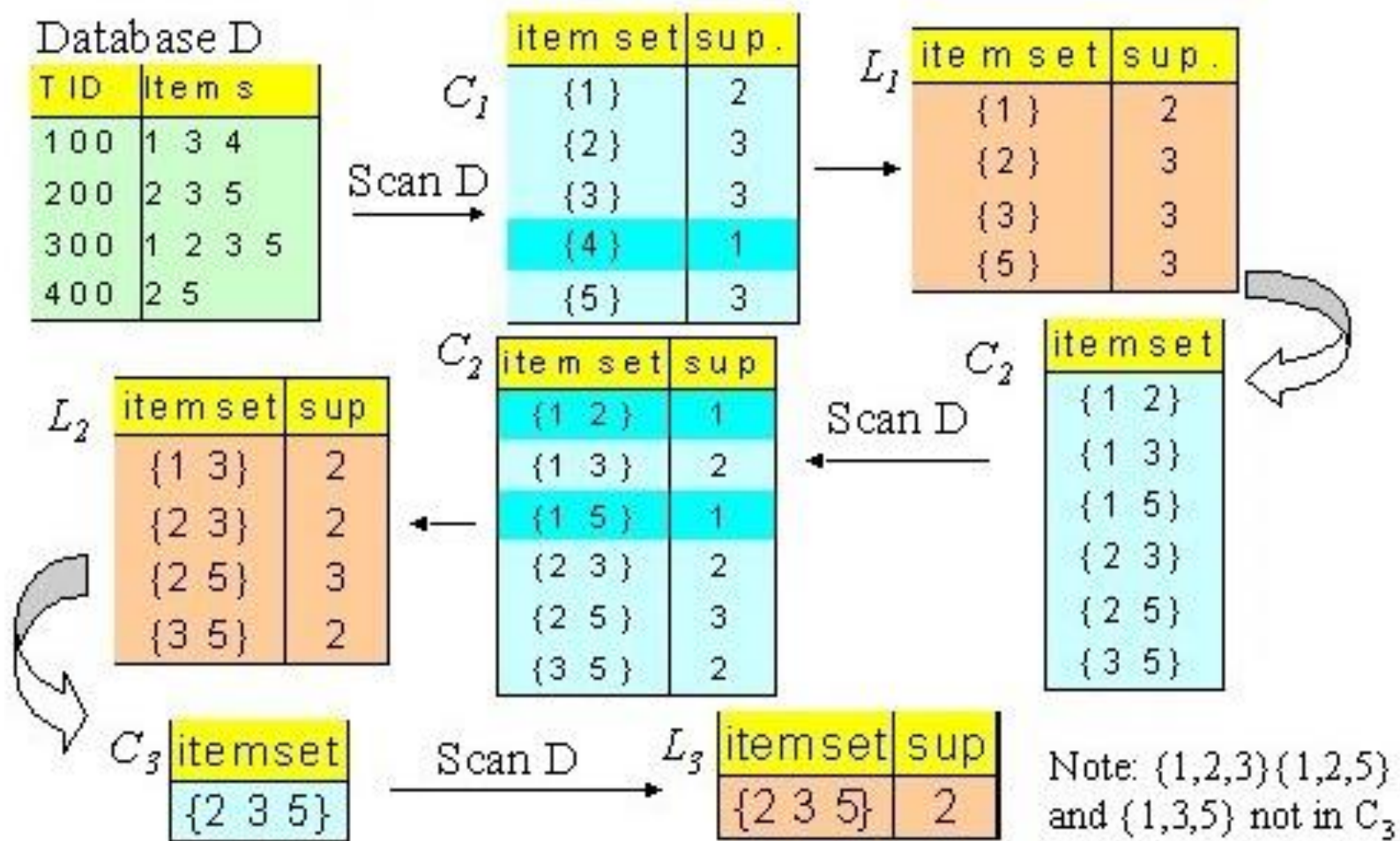
$\{\text{pasta}\} \rightarrow \{\text{queso, vino}\}$

$\{\text{vino}\} \rightarrow \{\text{queso, pasta}\}$

Y se generan reglas de aquellos que tengan suficiente confianza



## El algoritmo a priori



# Dificultades del algoritmo

- Definición del umbral de soporte y confianza
- Dimensiones (número de ítems). Si el número de conjuntos frecuentes incrementa los cálculos se vuelven computacionalmente más costosos
- Tamaño de la base de datos
- Tamaño de las transacciones

## Filtrar y Ajustar

Reglas de asociación es una metodología muy útil pero requiere de un gran trabajo con los resultados



Usualmente se pueden llegar a producir muchas reglas de asociación, no todas relevantes

Un post-procesamiento es necesario

# Filtrar y Ajustar

- Revisarlas con el conocimiento experto
- Hacer un análisis de sensibilidad probando distintos umbrales de soporte y confianza
- La confianza no necesariamente es la mejor métrica en todos los casos (lift, conviction)

Queremos llegar a identificar

Reglas no triviales (~~pasta~~ → ~~salsa para pasta~~)

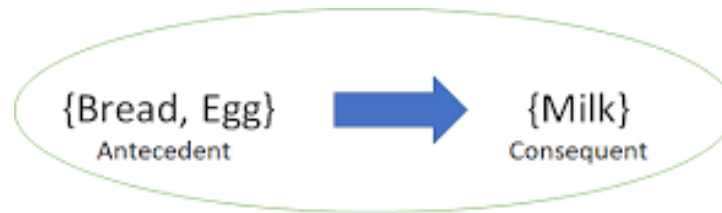
Reglas desconocidas/inesperadas

Patrones innovadores y accionables

# Filtrar y Ajustar

Importante tener conocimiento experto del negocio

Filtrar conjuntos de datos que siempre van a ser frecuentes sin importar los umbrales (canasta familiar: huevos, pan y leche)



Aceptar reglas interesantes de conjuntos que no alcanzan los umbrales

# Filtrar y Ajustar

Validar lo interesante de las reglas teniendo en cuenta el contexto de las transacciones

- Patrones raros, tienen bajo soporte pero son interesantes (relojes Rolex)
- Reglas segmentadas por grupos de interés  
Definir umbrales específicos por grupos de interés (edad, género, estrato)
- Patrones negativos  
Items con correlación negativa  
Hummer vs Tesla  
Cómida saludable vs Comida chatarra

A veces los patrones infrecuentes pueden ser más interesantes que los frecuentes

# Aplicaciones

## Market basket analysis

{comida para bebés, pañales} → {cerveza}

Ubicarlas mas cerca en la tienda

Ubicarlas más lejos

Poner productos de interés entre ellas

Subir el precio de un producto y bajar el del otro



## Sistemas de recomendación

Cientes que compran A, frecuentemente compran B

Recomendaciones sencillas pero con mucho potencial

## Detección de fraude

{Aseguradora A, Taller de reparación B, Oficial de policía C} con una alta frecuencia

# Extensiones

Minería en secuencias

Reglas de asociación multinivel

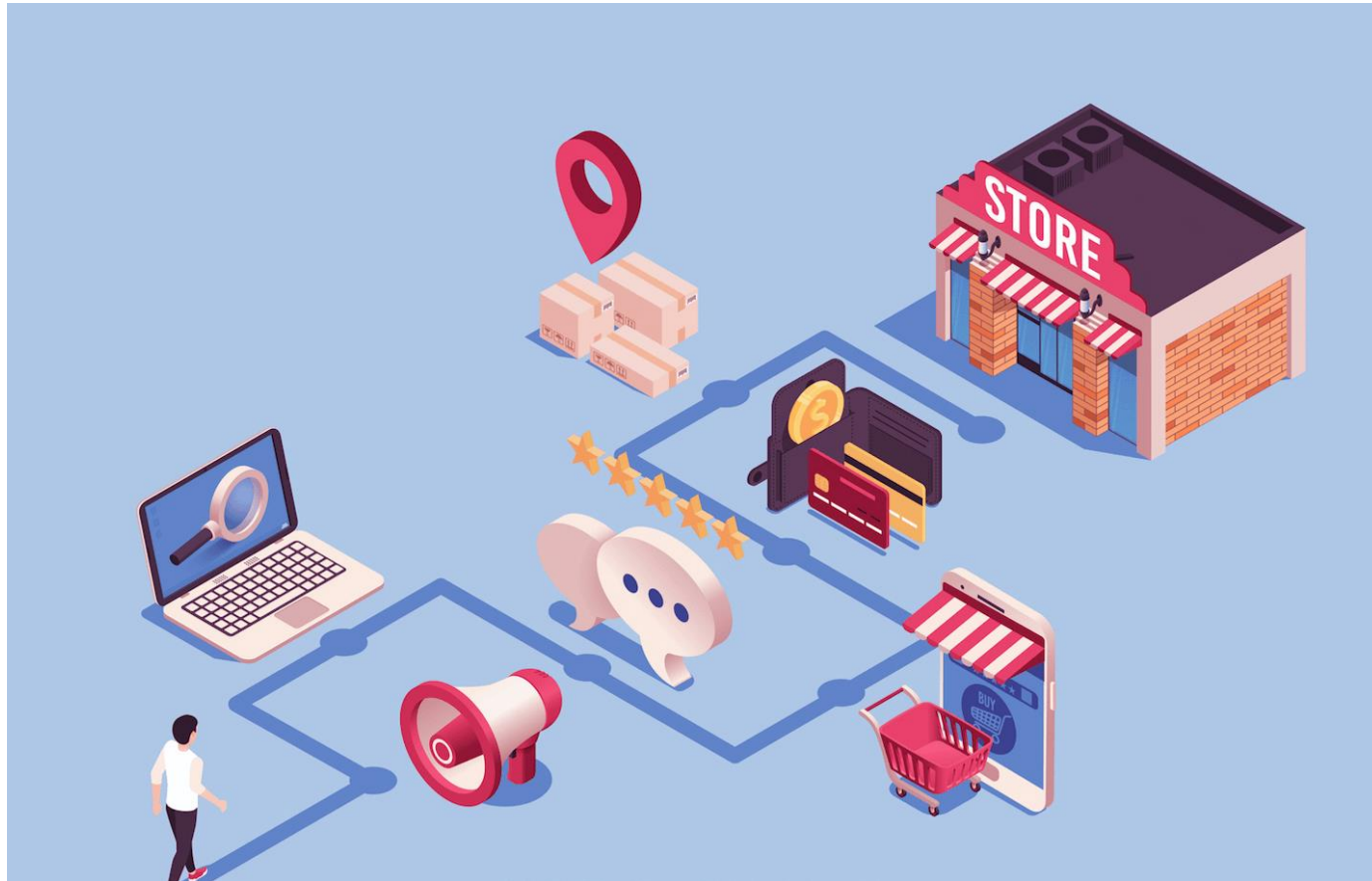
Reglas de asociación con variables mixtas



# Minería en secuencia (sequence mining)

En el algoritmo estándar apriori, el orden de los items no afecta el análisis. En sequence mining **el orden sí importa**.

Ya no se tienen conjuntos de ítems, sino secuencias de ítems



Ejemplos: Web mining, Customer journey analysis

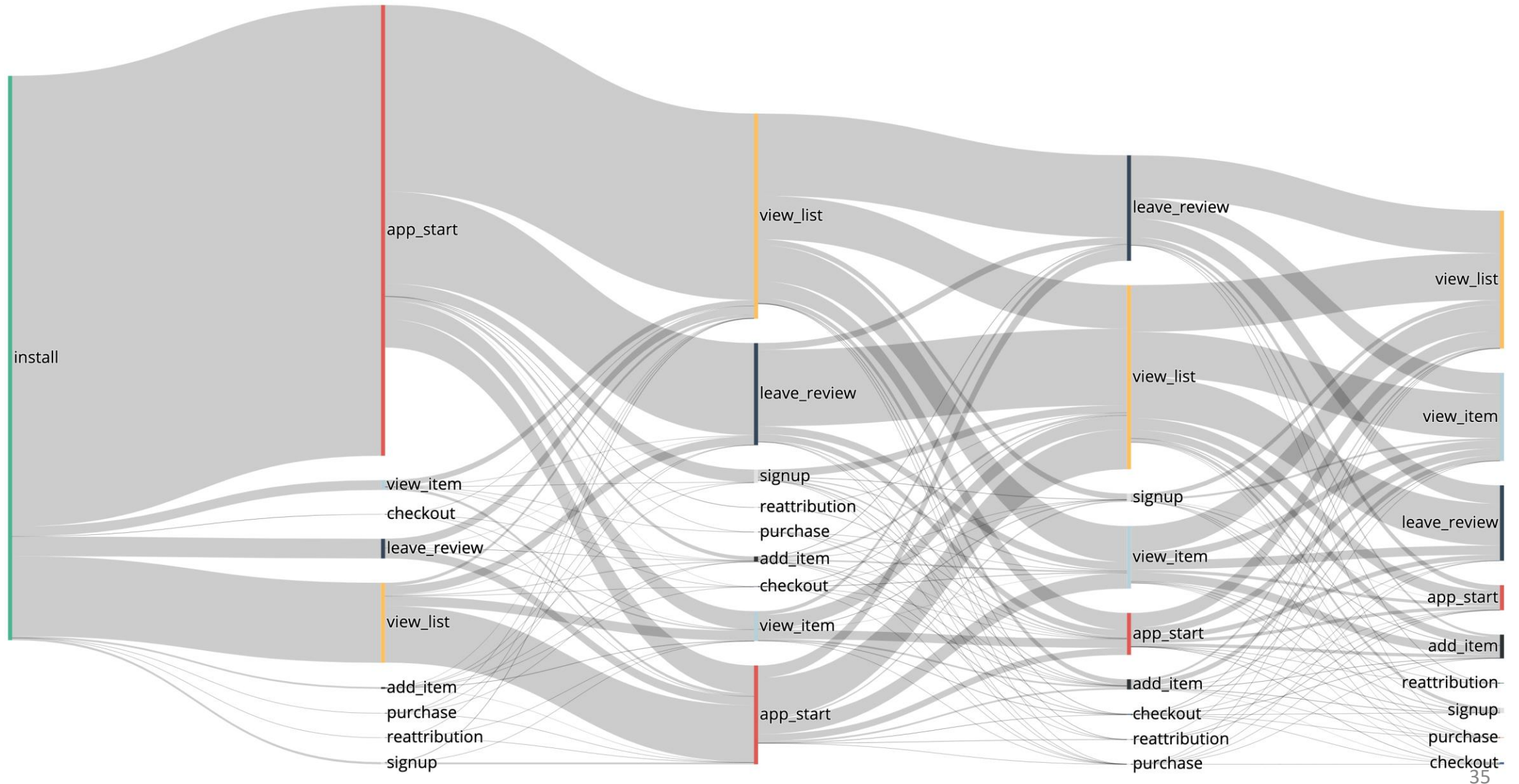
# Minería en secuencia (sequence mining)

GSP-Generalized Sequential Pattern algorithm (similar al apriori)

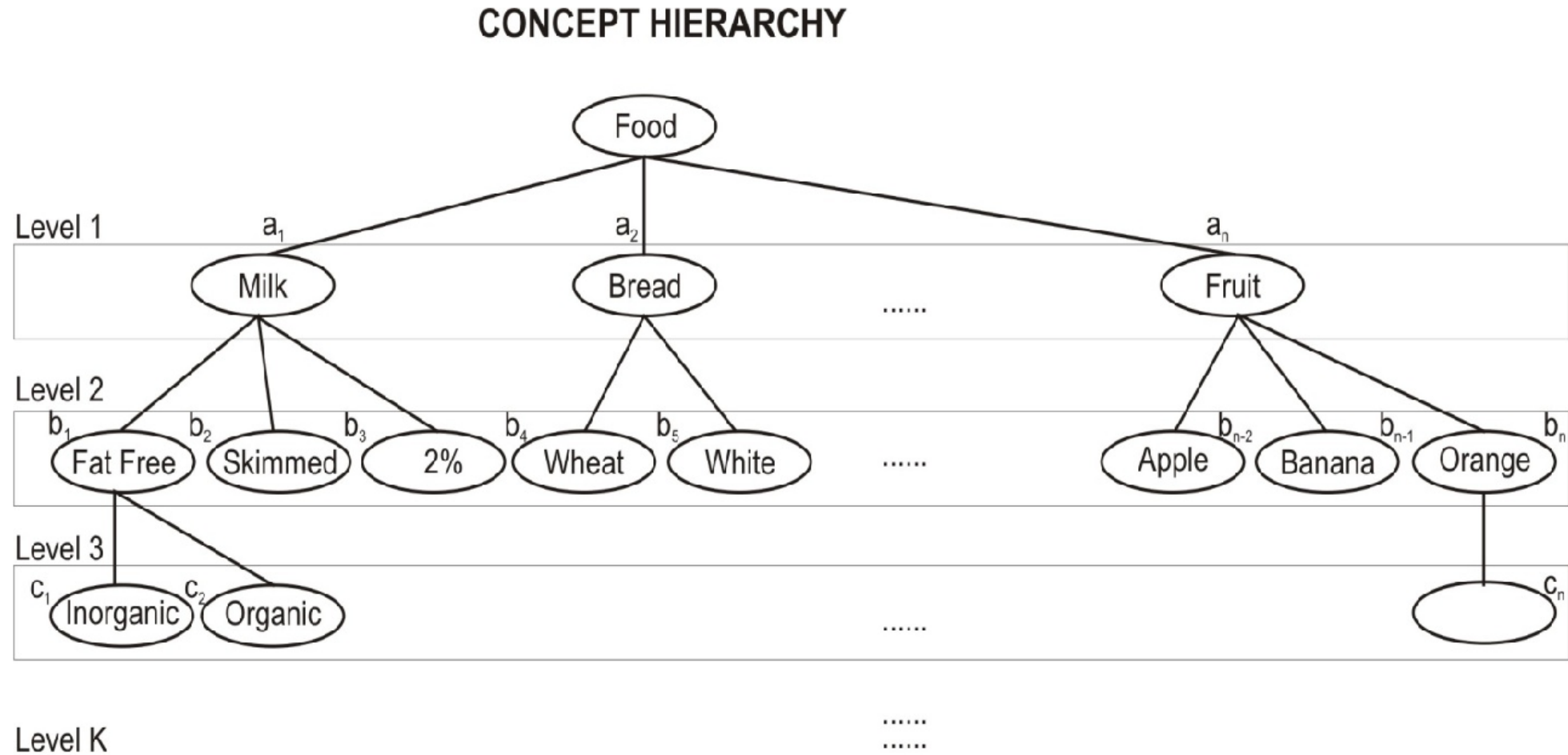
- Empezar con las secuencias de tamaño 1 más frecuentes
- La generación de candidatos es diferente
  - En apriori los conjuntos  $\{a,b\}$  y  $\{a,c\}$  llevan al candidato  $\{a,b,c\}$
  - Para secuencias se llega a  $\{a,b,c\}$  y  $\{a,c,b\}$

La reglas se definen igualmente, a partir de umbrales de soporte y confianza

# Minería en secuencia – Sankey Diagram

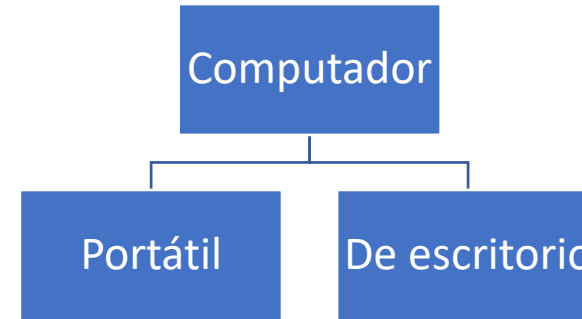


# Reglas de asociación multinivel



# Reglas de asociación multinivel

Incorpora conceptos de jerarquía en los ítems



Reglas en niveles inferiores pueden no tener suficiente soporte para considerarlas como frecuentes

Reglas en niveles inferiores tienden a ser muy específicas

$\{\text{leche descremada}\} \rightarrow \{\text{pan blanco}\}$

$\{\text{leche 2\%,}\} \rightarrow \{\text{pan integral}\}$

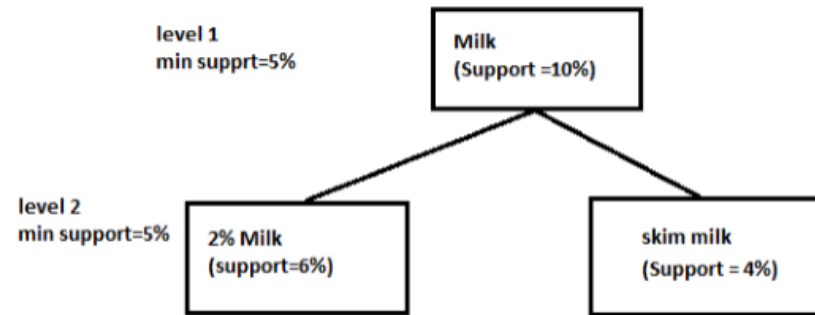
Reglas en niveles superiores pueden ser muy generales

$\{\text{leche}\} \rightarrow \{\text{pan}\}$

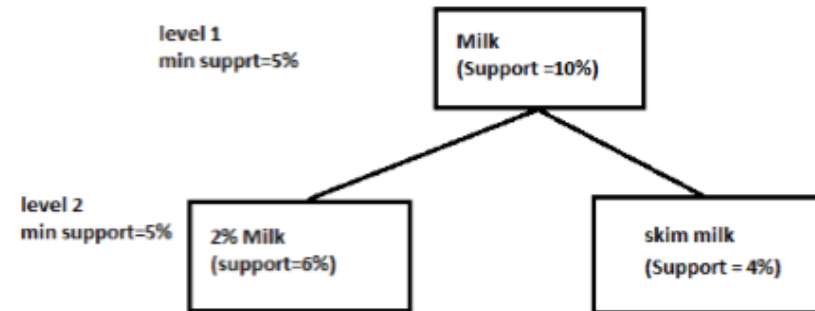
# Reglas de asociación multinivel

Dos opciones para definir los conjuntos frecuentes

- Mesmo soporte para todos los niveles - general



- Umbral de soporte definido para cada nivel - específico



# Reglas de asociación multinivel

Extender las reglas de asociación aumentando el conjunto con los ítems de niveles superiores

**Asociación encontrada: {leche descremada, pan blanco}**

**Asociación aumentada: {leche descremada, pan blanco, leche, pan}**

Dificultades:

- Ítems en niveles superiores tienen soporte mucho mas alto
- Si el umbral es general, con un soporte alto perdemos asociaciones de niveles inferiores. Con un soporte bajo, generamos demasiadas reglas
- La dimensionalidad aumenta
- Si el umbral es específico, se pueden perder asociaciones entre niveles

# Reglas de asociación con variables mixtas

Ejemplo: {buscador:'Chrome', páginas visitadas: < 15} → {compra: 'no'}

Transformar cada variable a ítems binarios (dummy): categorizar variables continuas y luego hacer dummies

- Recategorizar (agrupar) variables categóricas con muchas categorías a menos
- Eliminar niveles frecuentes que no son interesantes. Si el 90% usan Chrome, no es información de utilidad
- El agrupamiento de variables continuas en muchas categorías puede llevar a soportes o confianzas muy bajas



# Conclusiones

- Manejo de datos transaccionales
- Definición de métricas para entender que tan relevante/importante es una regla de asociación
- Algoritmo a priori
- Aplicaciones de reglas de asociación
- Extensiones de las reglas de asociación