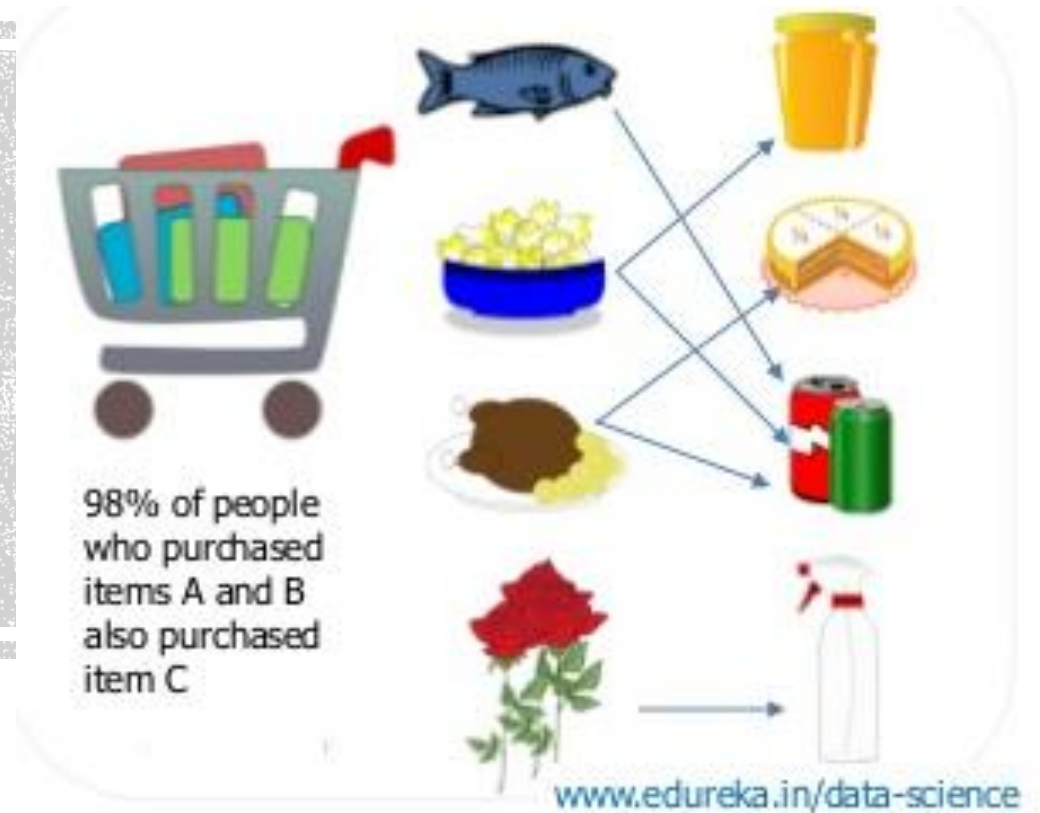


REGLAS DE ASOCIACIÓN — APRIORI



REGLAS DE ASOCIACIÓN

- **Market basket analysis (Análisis del carro de mercado):**
 - ¿Cuáles son los ítems mas propensos a ser comprados en conjunto? ...en los próximos 3 meses?
 - ¿Qué compran las personas con gustos similares?
 - Ofertas, amarres de productos, posición en el estante, venta cruzada
- **Predicción de navegación Web:**
 - Análisis de clickstream: ¿Cuál es el siguiente ítem más propenso a ser clickeado, o página a ser visitada?
- **Multimedia:**
 - Identificación de objetos en imágenes, videos o media social
 - Encontrar frases, entidades o atributos importantes en textos de gran volumen
- **Bioteología**
 - Encontrar secuencias de proteínas repetidas en secuencias genómicas del DNA
- **Social Networks**
 - Encontrar comunidades escondidas



REGLAS DE ASOCIACIÓN

- Aprendizaje no supervisado para descubrir **relaciones** significativas escondidas en el dataset
- **Transacción:** lista de productos comprados en conjunto en una misma visita a la tienda
- **Itemset:** Conjunto de uno o más productos
- **Itemset frecuente:** itemset cuyos ítems son frecuentemente comprados juntos (con respecto a un nivel mínimo de **soporte**)
- **Soporte:** Fracción de las transacciones que contienen un itemset dado → absoluta (conteo) o relativa (porcentaje)
- **Reglas:** ítem A → ítem B
- **Conocimiento de dominio:** Algunas reglas descubiertas pueden resultar inútiles por su obviedad (Papel → Lápiz), otras pueden resultar inesperadas, por tanto útiles (Pañal → Cerveza)

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

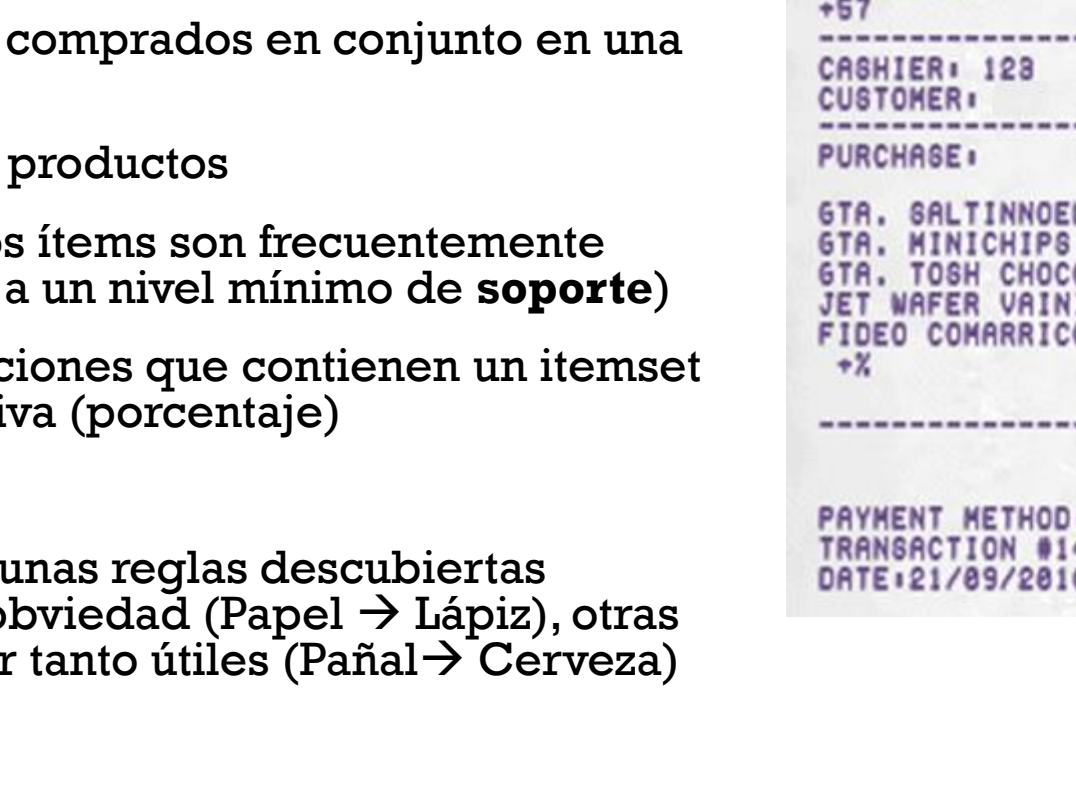
CASHIER: 123
CUSTOMER: **JUAN PEREZ**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE 86.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE: 21/09/2016 5:01:29 PM



REGLAS DE ASOCIACIÓN — MEDIDAS

- **Soporte** $(A \rightarrow B) = P(A \ \& \ B)$, **simétrica**
- **Confianza** $(A \rightarrow B) = P(B \mid A) = P(A \ \& \ B) / P(A)$, **asimétrica**
 - Probabilidad condicional
 - Indica qué tanto se puede confiar en la regla, pero no si se trata de una coincidencia
- **Lift** $(A \rightarrow B) = \text{Confianza}(A \rightarrow B) / P(B) = P(A \ \& \ B) / P(A) * P(B)$, **simétrica**
 - Cuántas veces más ocurren A y B juntas que lo que se esperaría si fueran independientes
 - $=1$: Regla inútil. A y B son **independientes** entre ellas (no hay relación significativa)
 - >1 : La regla es útil. Entre mayor el lift mejor. Se trata de productos **complementarios**
 - <1 : La regla es útil para identificar productos **sustitutos**
- **Leverage** $(A \rightarrow B) = P(A \ \& \ B) - P(A) * P(B)$, **simétrica**
 - Medida análoga al lift, pero aditiva, y utilizando 0 como el límite de decisión



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE



1-Itemset



4-Itemset



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Cuál es el soporte del itemset {Café Sello Rojo}?

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **XXX**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE:21/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **YYY**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 500Gx16PAST	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024266 -001
DATE:11/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **ZZZ**

PURCHASE:

GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475023934 -001
DATE:27/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **PPP**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL BP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024073 -001
DATE:24/09/2016 5:01:29 PM

Soporte de {Café Sello Rojo} = $3/4 = 75\%$



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Cuál es el soporte del itemset {Spaghetti Doria Clásica}?

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **XXX**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE:21/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **YYY**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 600Gx16PAG	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024266 -001
DATE:11/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **ZZZ**

PURCHASE:

GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475023934 -001
DATE:27/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **PPP**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL AP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024073 -001
DATE:24/09/2016 5:01:29 PM

Soporte de {Café Sello Rojo} = $3/4 = 75\%$

Soporte de {Spaghetti Doria Clásica} = $2/4 = 50\%$



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Cuál es el soporte del itemset {Café Sello Rojo, Spaghetti Doria Clásica}?

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **XXX**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE:21/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **YYY**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 500Gx16PAGT	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024266 -001
DATE:11/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **ZZZ**

PURCHASE:

GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475023934 -001
DATE:27/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **PPP**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL AP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024073 -001
DATE:24/09/2016 5:01:29 PM

Soporte de {Café Sello Rojo} = $3/4 = 75\%$

Soporte de {Spaghetti Doria Clásica} = $2/4 = 50\%$

Soporte de {Café Sello Rojo, Spaghetti Doria Clásica} = $2/4 = 50\%$



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Confianza, lift y leverage de {Café Sello Rojo → Spaghetti Doria Clásica}?

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: XXX	
PURCHASE:	
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00
TOTAL: \$16.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475024200 -001 DATE:21/09/2016 5:01:29 PM	

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: YYY	
PURCHASE:	
CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 500Gx16PAGT	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00
TOTAL: \$15.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475024266 -001 DATE:11/09/2016 5:01:29 PM	

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: ZZZ	
PURCHASE:	
GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00
TOTAL: \$15.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475023934 -001 DATE:27/09/2016 5:01:29 PM	

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: PPP	
PURCHASE:	
CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL AP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00
TOTAL: \$16.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475024073 -001 DATE:24/09/2016 5:01:29 PM	

Soporte de {Café Sello Rojo} = $3/4 = 75\%$

Soporte de {Spaghetti Doria Clásica} = $2/4 = 50\%$

Soporte de {Café Sello Rojo, Spaghetti Doria Clásica} = $2/4 = 50\%$

Confianza de {Café Sello Rojo → Spaghetti Doria Clásica} = $(2/4)/(3/4) = 66,6\%$

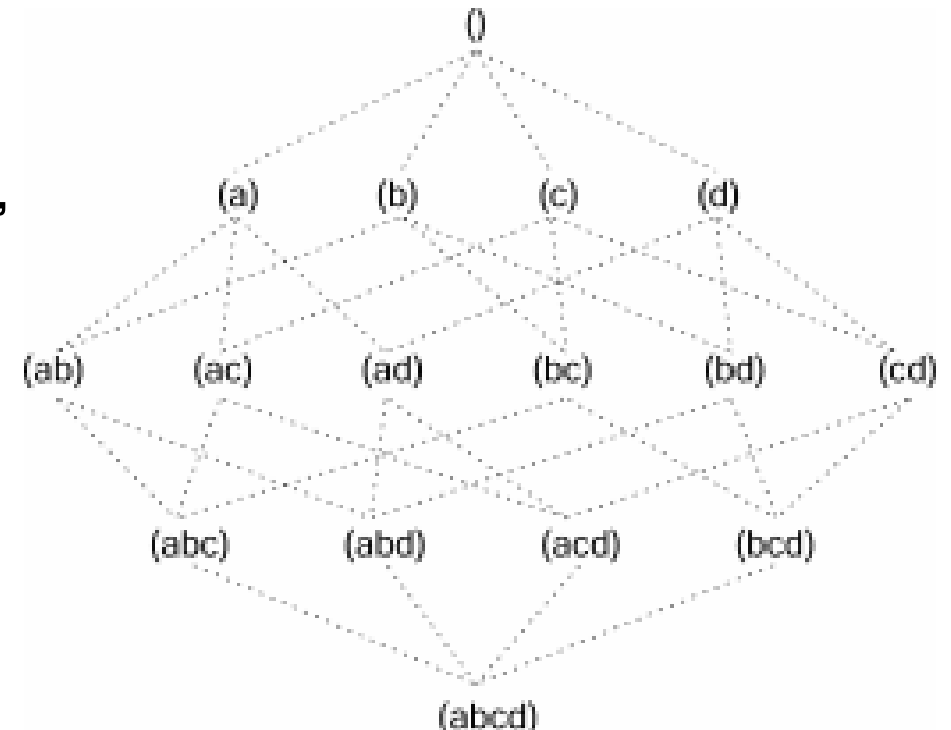
Lift de {Café Sello Rojo → Spaghetti Doria Clásica} = $(2/4) / ((3/4)*(2/4)) = 4/3 = 1,33$

Leverage de {Café Sello Rojo → Spaghetti Doria Clásica} = $(2/4) - ((3/4)*(2/4)) = 4/3 = 0,125$



APRIORI

- Encontrar los itemsets frecuentes es un problema Np-Hard.
- El algoritmo **Apriori** poda el espacio de búsqueda, para luego definir las reglas resultantes
- Búsqueda bottom-up de los itemsets frecuentes:
 - se debe especificar un umbral de **sopORTE** mínimo
- Las reglas son extraídas de los itemsets frecuentes encontrados:
 - se pueden especificar condiciones adicionales para las reglas encontradas con respecto a métricas de **confianza**, **lift** y/o **leverage**.
 - Itemset antecedente → Itemset consecuente

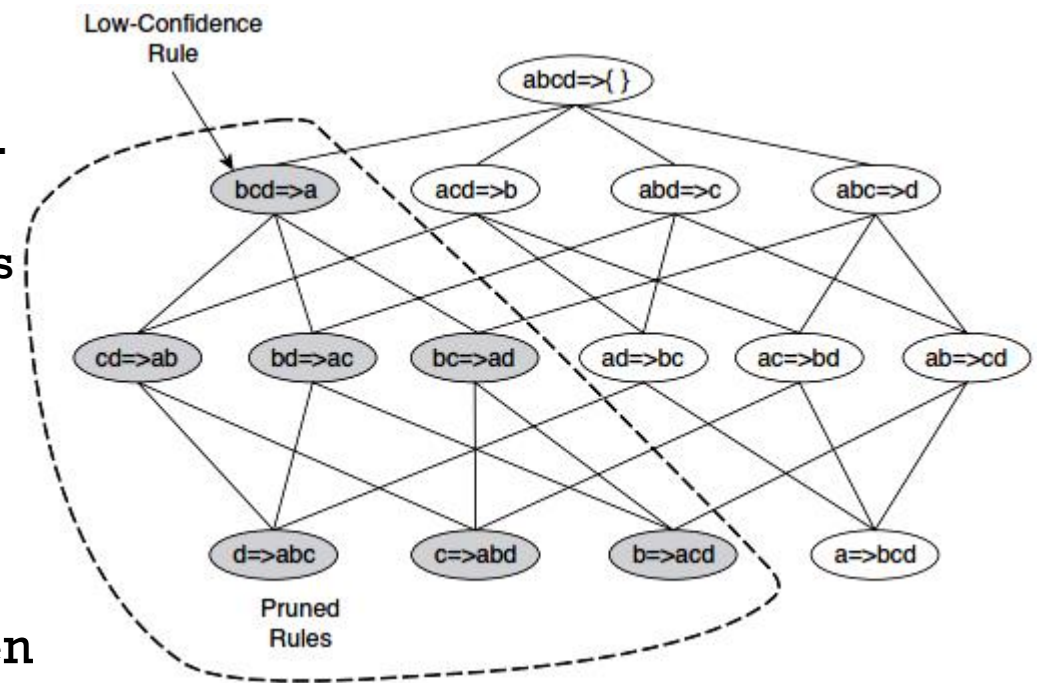


Espacio de búsqueda de 4 ítems



APRIORI - ALGORITMO

- Encontrar los itemsets candidatos (para un **soporte** definido):
 - Retener los 1-itemsets frecuentes como candidatos.
 - Descartar los 1-itemsets no frecuentes
 - Para (n en 2:N): Encontrar los (n)-itemset frecuentes combinando los (n-1)-itemsets candidatos. Guardarlos como candidatos
 - Repetir hasta el final, hasta que los itemsets se queden por debajo del soporte, o hasta llegar a un máximo de cardinalidad especificada
- Combinar los miembros de los itemsets candidatos, encontrando las reglas que satisfacen un mínimo de otra medida definida (**confidence**, **lift**, **leverage**)



Ejemplo de resultados encontrados

<http://www.paulallen.ca/apriori-algorithm-rule-generation/>



APRIORI

- Consideraciones:
 - La búsqueda en anchura genera una **complejidad** computacional temporal y espacial alta: cuando el número de productos y/o transacciones es muy grande, es necesario adoptar estrategias adicionales para reducir el espacio de búsqueda
 - En grandes datasets, la mayoría de los eventos van a ser raros (soportes y confianzas bajas)
 - La minería de reglas de asociación debe hacerse iterativamente, teniendo en cuenta la opinión de expertos del dominio en el equipo de analítica
- Alternativas
 - Eclat: algoritmo de búsqueda en profundidad
 - FP Growth



REFERENCIAS

- *Introduction to recommender Systems*, Joseph Konstan, 2015
- EMC2, “Data science and big data analytics”, 2015, John Wiley & Sons
- *Data Science for Business*, Foster Provost & Tom Fawcett, O’Reilly, 2013
- *Practical Data Science with R*, Nina Zumel & John Mount, 2014
- *Mining association rules between sets of items in large databases*, R. Agrawal, T. Imielinski, and A. Swami, en Proc. of SIGMOD'93, 2013
- *Discovering frequent closed itemsets for association rules*, N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, en Proc. of ICDT'99, 1999
- http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

