DATA MINING

Association Rules Mining

Ashish Pujari

Lecture Outline

- RFM Analysis
- Association Rules Mining

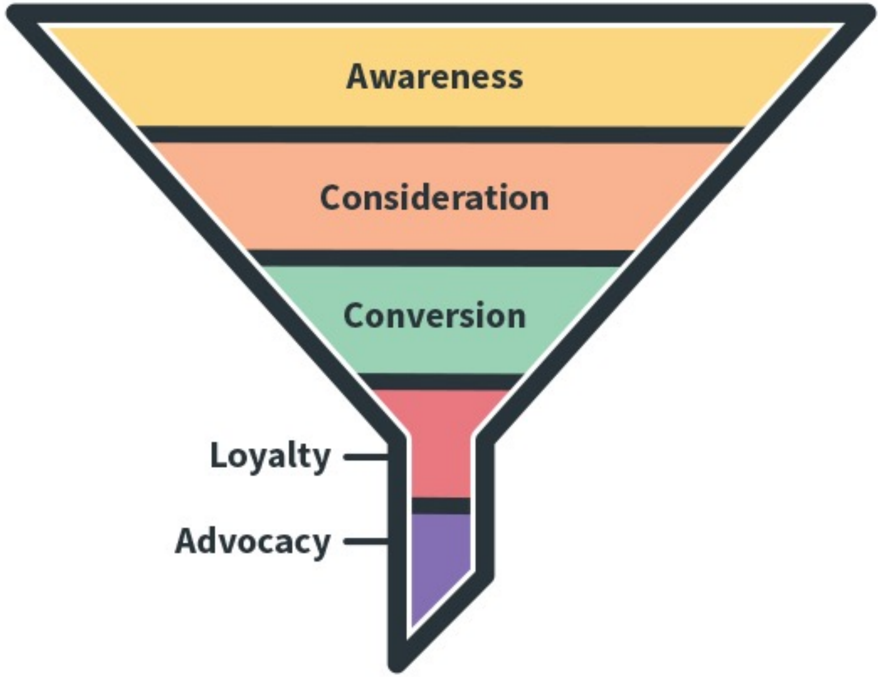
RFM ANALYSIS

Marketing Strategy

- A key element of a successful marketing strategy is identifying the target audience
- Tailor your marketing messages and tactics to increase engagement and conversion
- Understanding your customers:
  - Who are your best customers?
  - Which of your customers are at risk ?
  - Which of your customers can be retained?
  - Which of your customers are most likely to respond to engagement campaigns?
  - Who has the potential to become valuable customers?

Benefits of good marketing strategy

- Increased customer retention
- Increased customer loyalty
- Increased response rate
- Increased conversion rate
- Increased revenue
- Increasing sales
- Increased brand value



Marketing Funnel

Customer Retention

- Loyal customers boost sales
- It’s easier to sell to someone engaged with your brand
- It can cost 5x more to acquire a new customer than to retain an existing customer
- Retention and loyalty lead to word-of-mouth referrals

$$\boldsymbol{Customer\ Retention\ Rate} = \frac{(Total\ \#\ of\ Customers\ at\ the\ end\ of\ the\ Period\ -\ New\ Customers\ Acquired)}{Customers\ at\ the\ Start\ of\ the\ Period}$$

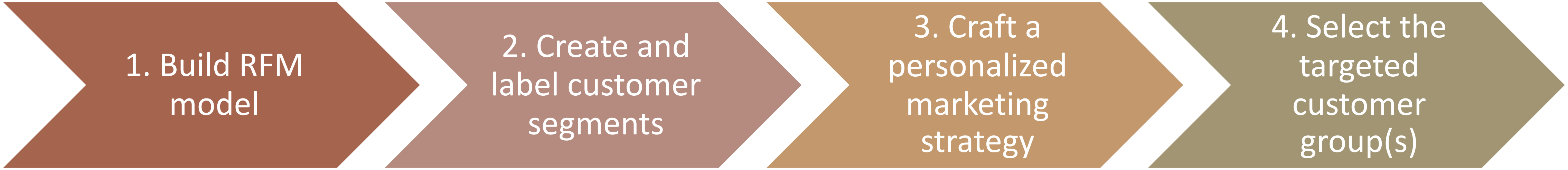
https://www.shopify.com/retail/customer-retention-retail

Recency, Frequency, Monetary (RFM) Analysis

- Marketing analysis tool to segment customers based on the spending habits
- Scores customers in three categories:
  - Recency (R): How recently a customer has made a purchase
  - Frequency (F): How often a customer makes a purchase
  - Monetary value (M): How much money a customer spends on purchases

Customer #	Recency	Frequency	Monetary
Cust1	5	3	2
Cust2	4	2	1
Cust3	4	4	5
Cust4	3	3	1
Cust5	4	5	4
Cust6	5	1	3

RFM Analysis: Steps



1. Build RFM model

2. Create and label customer segments

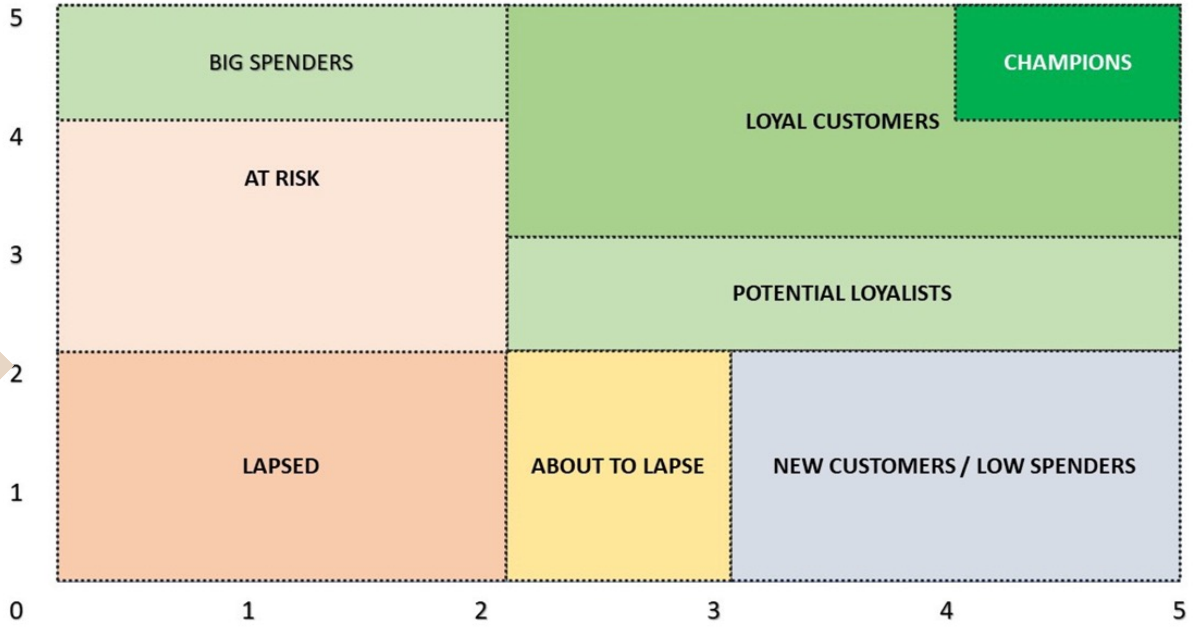
3. Craft a personalized marketing strategy

4. Select the targeted customer group(s)

RFM Analysis: Example

Customer #	Recency	Frequency	Monetary
Cust1	5	3	2
Cust2	4	2	1
Cust3	4	4	5
Cust4	3	3	1
Cust5	4	5	4
Cust6	5	1	3

Customer Segments



Source: bettermarketing.pub

RFM Analysis: Sample Segments

Segment	RFM	Description	Marketing Strategy
Best Customers	444	Customers who bought most recently, most often and spend the most.	No price incentives, New products and loyalty programs
Loyal Customers	X4X	Customers who bought most recently	Use R and M to further segment.
Big Spenders	XX4	Customers who spend the most	Market your most expensive products.
Almost Lost	213	Haven't purchased for some time but purchased frequently and spend the most.	Aggressive price incentives
Lost Customers	122	Haven't purchased for some time but purchased frequently and spend more.	Agressive price incentives.
Lost Cheap Customers	111	Last purchase long ago, purchased few and spend little.	Don't spend too much trying to re-acquire.

ASSOCIATION RULES MINING

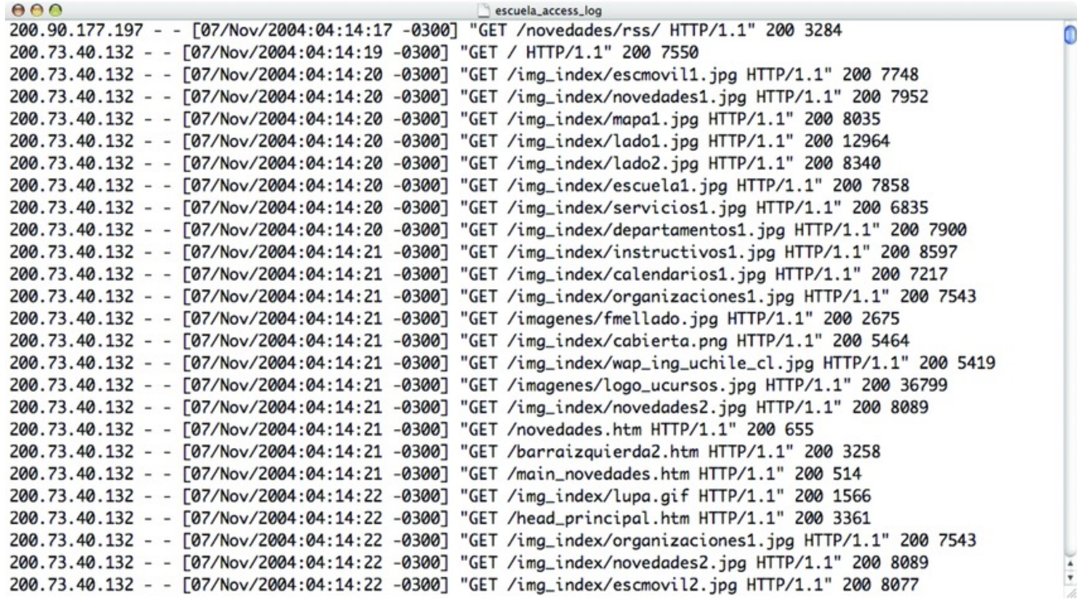
Item Association

- Association Analysis and Market Basket Analysis looks exclusively at content
  - The goal is to find items that are frequently consumed together

If a pair of items, X and Y, that are frequently bought together:
Both X and Y can be placed on the same shelf
Promotional discounts could be applied to just one out of the two items
Advertisements on X could be targeted at buyers who purchase Y
X and Y could be combined into a new product, such as having Y in flavors of X

Web Transaction Data

- Optimize the structure of the site
  - Discovery of rule A → B
  - special-offers/ , /products/software/ → shopping-cart/
  - Indication that promotional campaign on software products is positively affecting online sales



```
200.90.177.197 - - [07/Nov/2004:04:14:17 -0300] "GET /novedades/rss/ HTTP/1.1" 200 3284
200.73.40.132 - - [07/Nov/2004:04:14:19 -0300] "GET / HTTP/1.1" 200 7550
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/escmovil1.jpg HTTP/1.1" 200 7748
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/novedades1.jpg HTTP/1.1" 200 7952
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/mapa1.jpg HTTP/1.1" 200 8035
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/lado1.jpg HTTP/1.1" 200 12964
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/lado2.jpg HTTP/1.1" 200 8340
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/escuela1.jpg HTTP/1.1" 200 7858
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/servicios1.jpg HTTP/1.1" 200 6835
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/departamentos1.jpg HTTP/1.1" 200 7900
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/instructivos1.jpg HTTP/1.1" 200 8597
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/calendarios1.jpg HTTP/1.1" 200 7217
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/organizaciones1.jpg HTTP/1.1" 200 7543
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /imagenes/fmellado.jpg HTTP/1.1" 200 2675
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/cabierta.png HTTP/1.1" 200 5464
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/wap_ing_uchile_cl.jpg HTTP/1.1" 200 5419
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /imagenes/logo_ucursos.jpg HTTP/1.1" 200 36799
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/novedades2.jpg HTTP/1.1" 200 8089
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /novedades.htm HTTP/1.1" 200 655
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /barraizquierda2.htm HTTP/1.1" 200 3258
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /main_novedades.htm HTTP/1.1" 200 514
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/lupa.gif HTTP/1.1" 200 1566
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /head_principal.htm HTTP/1.1" 200 3361
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/organizaciones1.jpg HTTP/1.1" 200 7543
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/novedades2.jpg HTTP/1.1" 200 8089
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/escmovil2.jpg HTTP/1.1" 200 8077
```

Association Rules Mining

- Technique to uncover how items are associated to each other
- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Strong relationship between customers that purchased diapers, and beer in the same transaction.

$$\{\boldsymbol{Diapers}\}\ -\!>\ \{\boldsymbol{Beer}\}$$

(antecedent) (consequent)

$$\{\boldsymbol{Diaper}, \boldsymbol{Gum}\}\ -\!>\ \{\boldsymbol{Beer}, \boldsymbol{Chips}\}$$

(antecedent) (consequent)

Association Rules: Measures

- Suppose that we want to learn an association rule $X \rightarrow Y$
  - e.g., customers who buy X often buy Y as well

$$\text{Support } (X \rightarrow Y) \equiv P(X, Y)$$

- $Rule\ (X \rightarrow Y)$

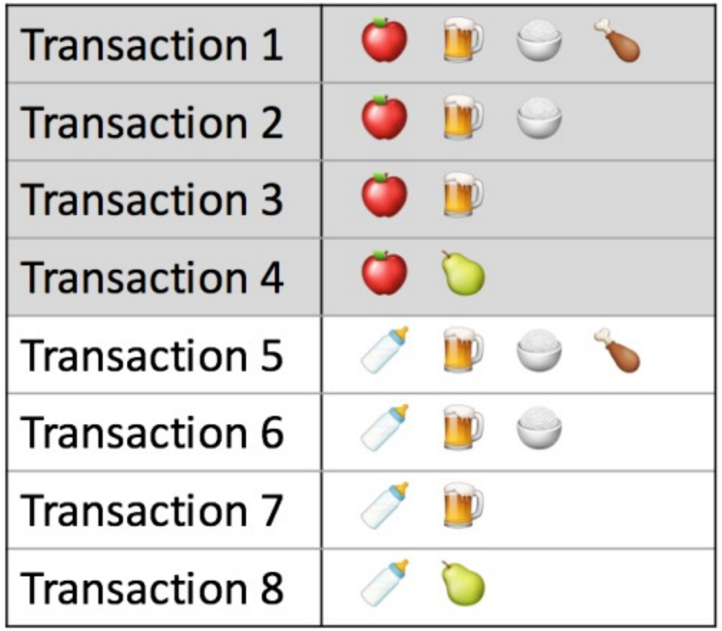
$$\text{Confidence}(X \rightarrow Y) \equiv P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$

$$\text{Lift } (X \rightarrow Y) \equiv \frac{P(X, Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{\mathrm{P}(Y)}$$

- $Lift > 1$: Y is more likely to be bought if X is bought
- $Lift = 1$: no association between items
- $Lift < 1$: Y is less likely to be bought if item X is bought

Support

- How popular is an itemset, as measured by the proportion of transactions in which an itemset appears.



Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Support {🍎} = $\frac{4}{8}$

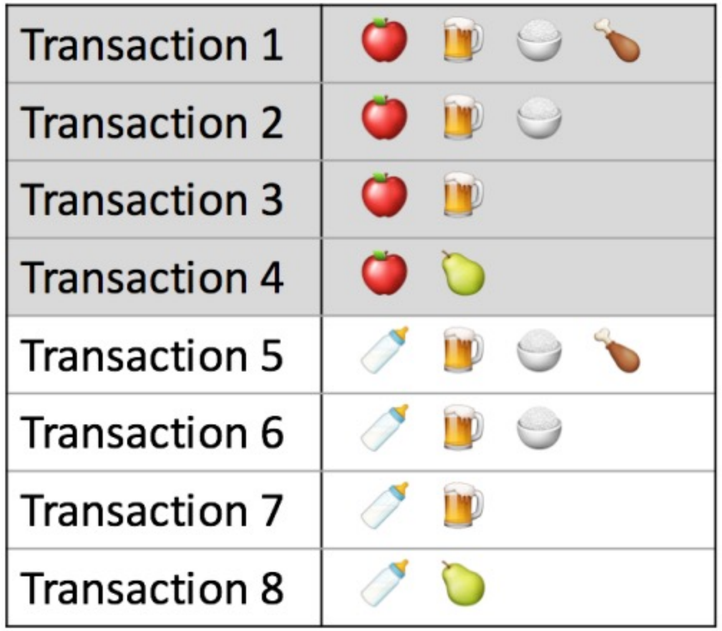
$$Support\ \{apple\} = \frac{4}{8} = \ 50\%$$

$$Support\ \{apple,\ beer,\ rice\} = \frac{2}{8} = 25\%$$

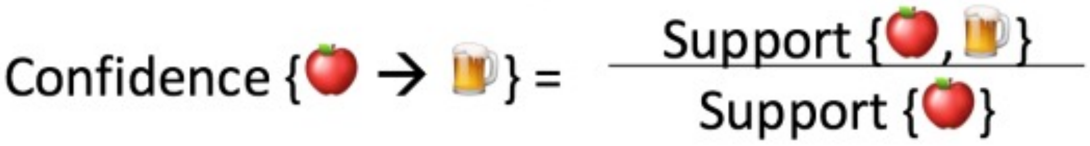
Source: kdnuggets.com

Confidence

- How likely is Y is to be purchased when X is purchased; Expressed as $\{X\ -> \ Y\}$
- Proportion of transactions with item X, in which item Y also appears.



Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐



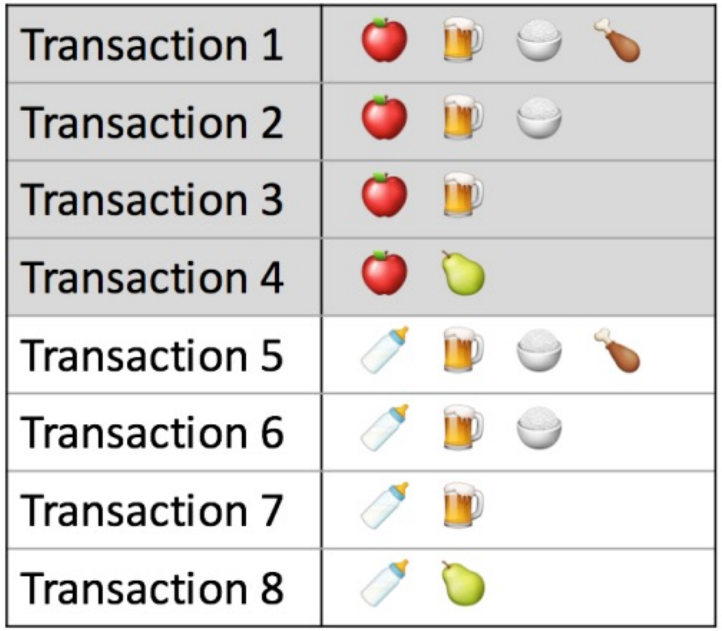
Confidence {🍎 → 🍺} = Support {🍎, 🍺} / Support {🍎}

$$Confidence\,\{apple\ -> \ beer\} = \frac{3}{4} = \ 75\%$$

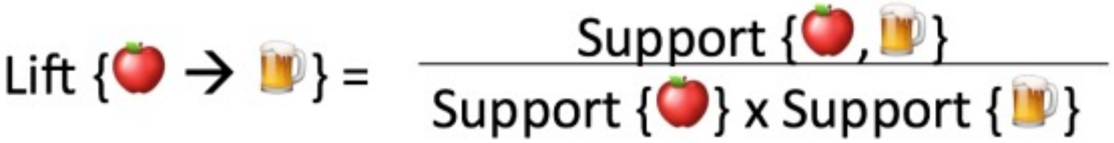
Source: kdnuggets.com

Lift

- Ratio of *confidence* to baseline probability of occurrence of $\{Y\}$
- How likely is it that item Y is purchased when item X is purchased, while controlling for how popular item Y is



Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐



Lift {🍎 → 🍺} = Support {🍎, 🍺} / Support {🍎} x Support {🍺}

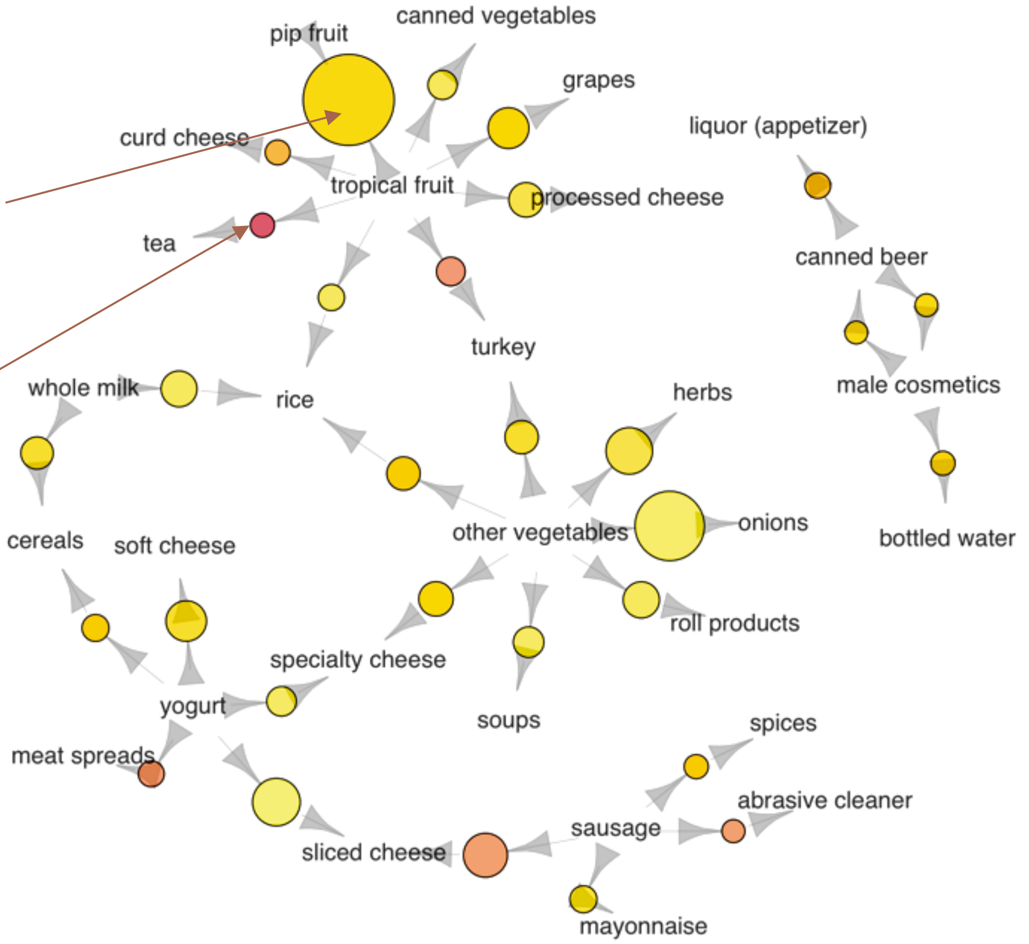
$$Lift\ \{apple\ \ -{>}\ \ beer\} = 1$$

Source: kdnuggets.com

Frequent Itemset Visualization

Larger circles imply higher support

Red circles imply higher lift



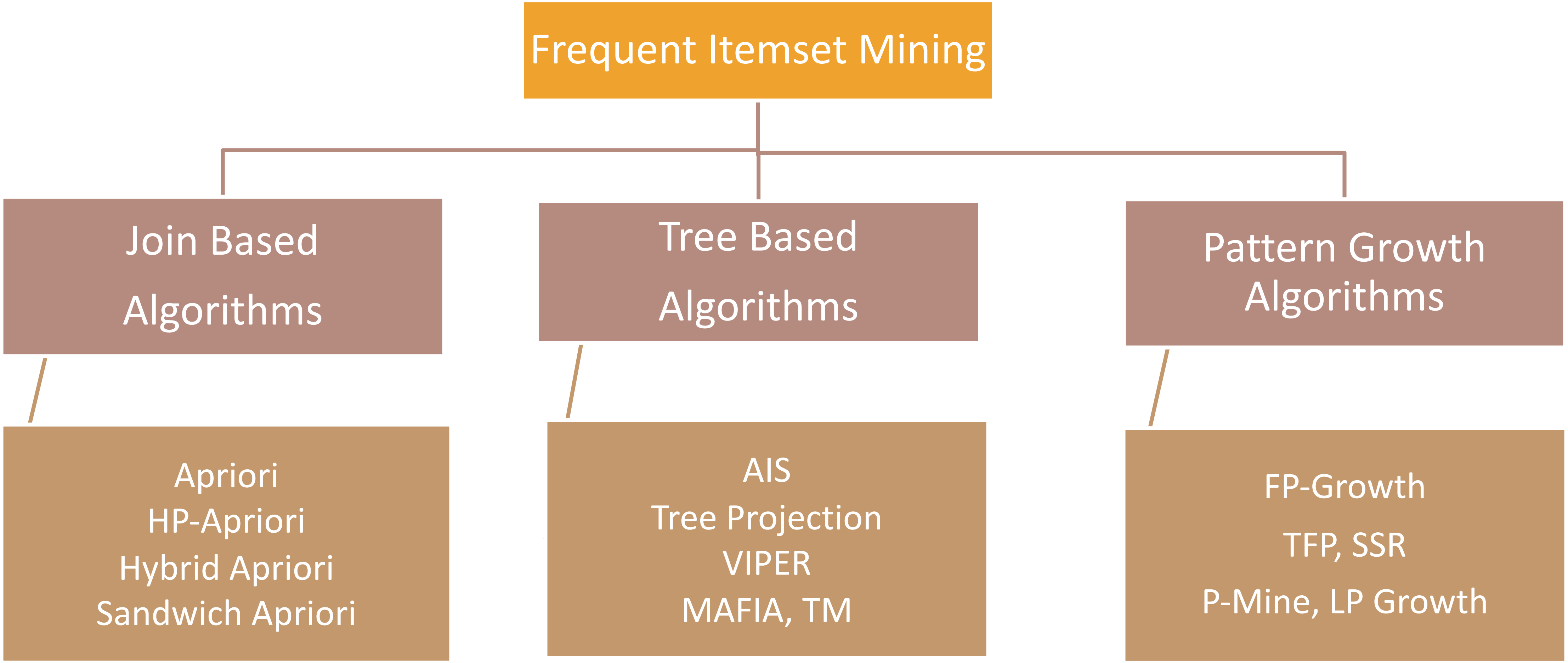
The network graph shows associations between selected items

Association Rule Mining: Problem

- Problem:
  - Given a set of transactions T, the goal of association rule mining is to find all rules having
  - $support \ \geq \ minsup$ threshold
  - $confidence \ \geq \ minconf$ threshold

- Two step Process:
  - Find all frequent item-sets
  - Generate strong association rules from frequent item-sets

Association Rule Mining: Algorithms



Frequent Itemset Mining

Join Based Algorithms

Apriori
HP-Apriori
Hybrid Apriori
Sandwich Apriori

Tree Based Algorithms

AIS
Tree Projection
VIPER
MAFIA, TM

Pattern Growth Algorithms

FP-Growth
TFP, SSR
P-Mine, LP Growth

Frequent Itemsets: Example

$\{Beer, Diapers, Milk\}$:

$$\{Beer\}, \{Diapers\}, \{Milk\},$$

$$\{Beer, Diapers\}\ , \{Beer, Milk\}, \{Diapers, Milk\}$$

$$\{Beer, Diapers, Milk\},$$

Association Rules: Example

$\{Beer, Diapers, Milk\}$:

$$\{Beer, Diapers\} \rightarrow \{Milk\},$$

$$\{Beer, Milk\} \rightarrow \{Diapers\},$$

$$\{Diapers, Milk\} \rightarrow \{Beer\},$$

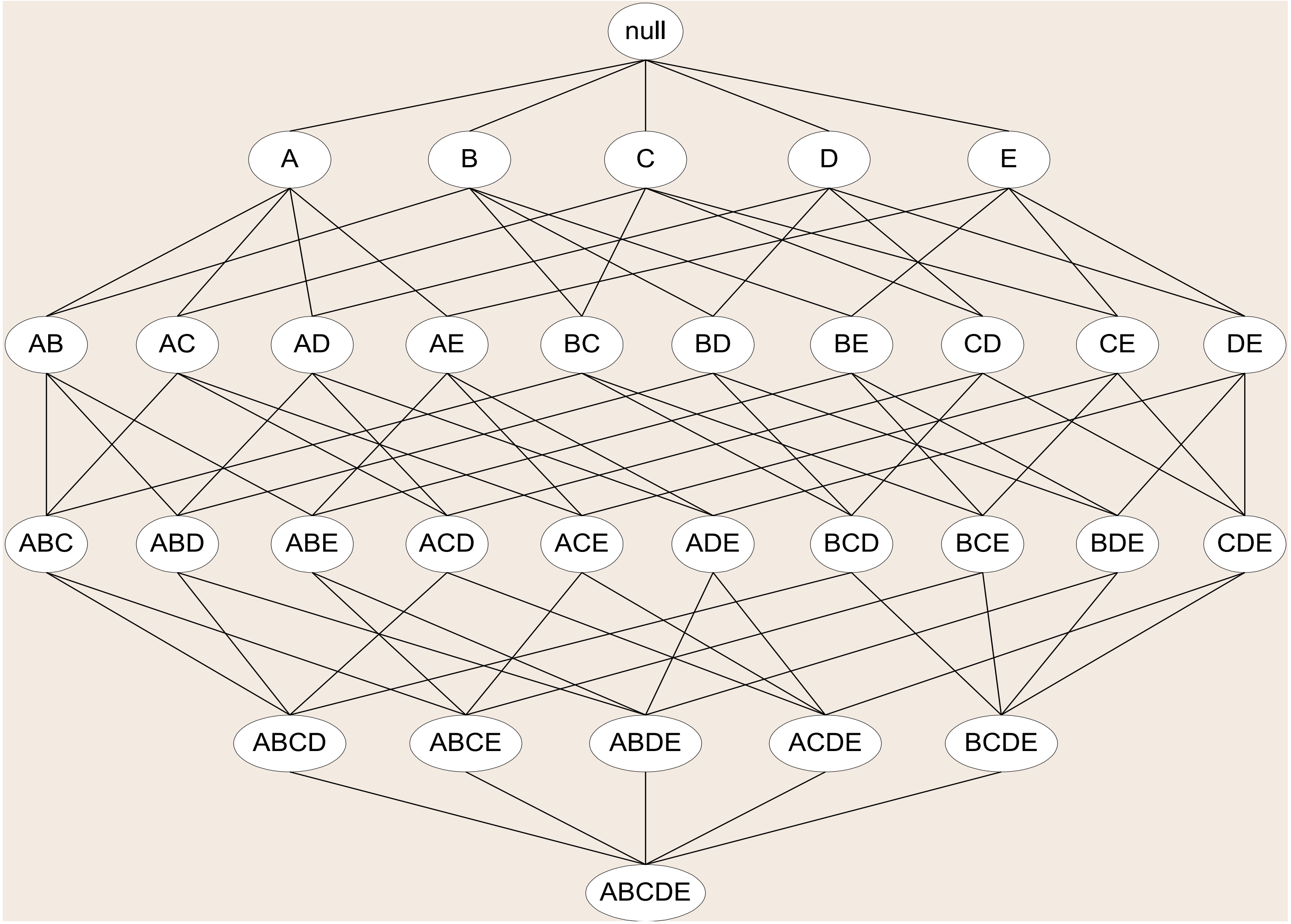
$$\{Beer\} \rightarrow \{Diapers, Milk\},$$

$$\{Milk\} \rightarrow \{Beer, Diapers\},$$

$$\{Diapers\} \rightarrow \{Beer, Milk\}$$

Candidate Itemsets

Given d items, there are 2^d possible candidate itemsets



Itemset lattice

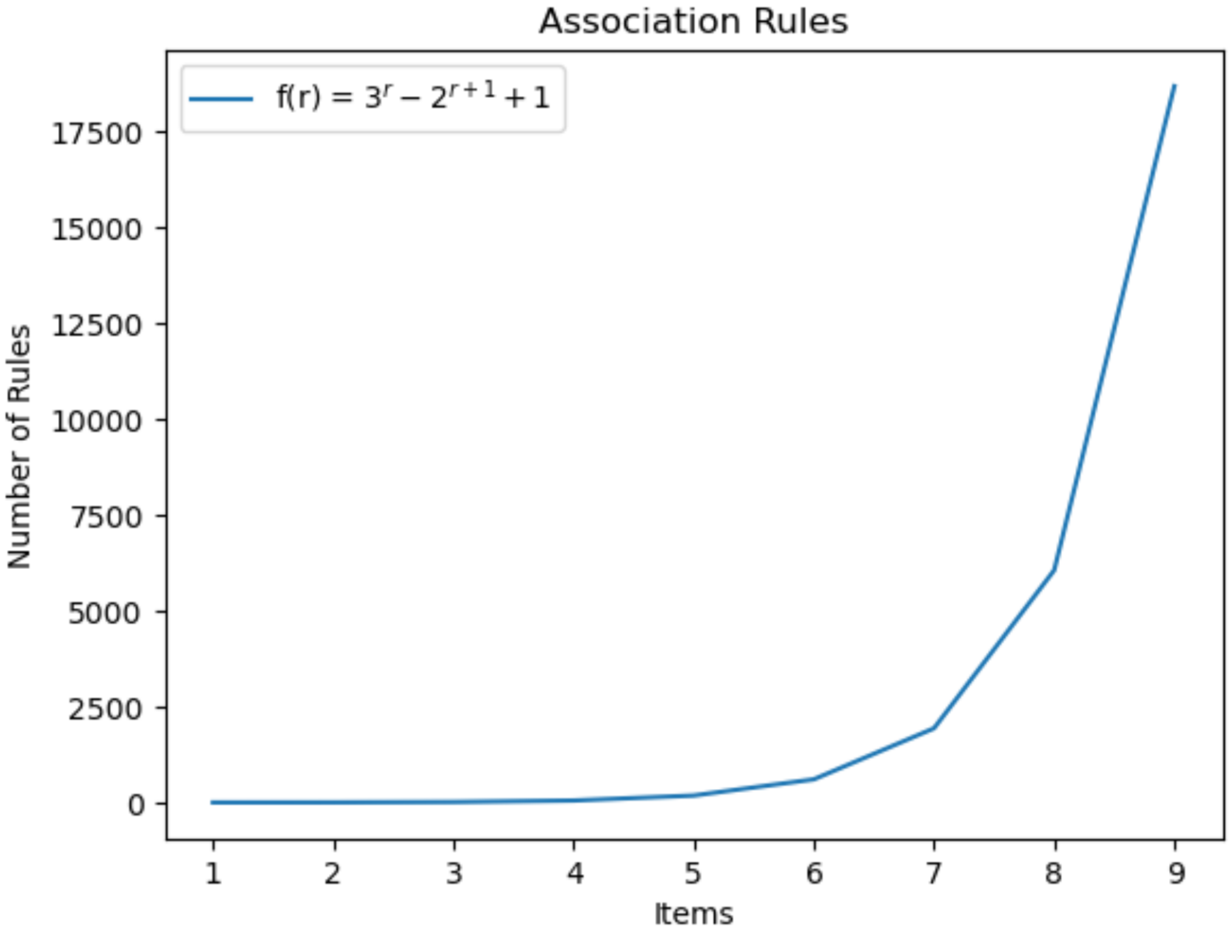
https://www-users.cse.umn.edu/~kumar001/dmbook/ch6.pdf

Computational Complexity

- Brute Force Approach
  - List all possible association rules
- Given r unique items:
  - Total number of itemsets = $2^r - 1$
  - Total number of possible association rules:

$$R = \sum_{k=1}^{r-1}\left[\binom{r}{k} \times \sum_{j=1}^{r-k}\binom{r-k}{j}\right]$$
$$= 3^r - 2^{r+1} + 1$$

If r=6, R = 602 rules



Computationally prohibitive!

Frequent Itemset Generation Strategies

- Reduce the number of candidates (M)
  - Complete search: M=2^d
  - Use pruning techniques to reduce M
- Reduce the number of transactions (N)
  - Reduce size of N as the size of itemset increases
  - Used by *DHP* and vertical-based mining algorithms
- Reduce the number of comparisons (NM)
  - Use efficient data structures to store the candidates or transactions
  - No need to match every candidate against every transaction

Reducing Number of Candidates

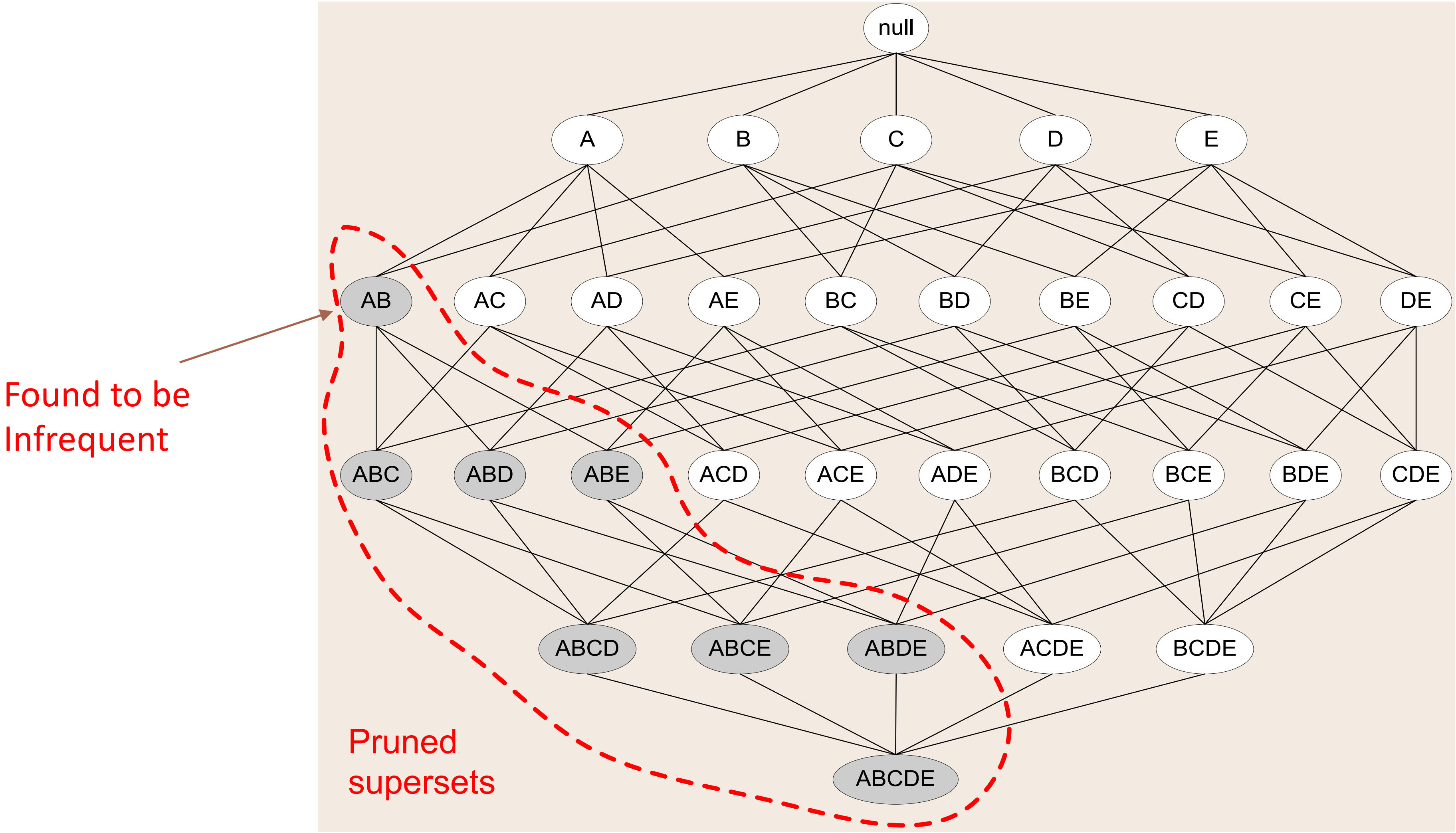
- Apriori principle:
  - If an itemset is frequent, then all of its subsets must also be frequent
  - Example: if {beer} was found to be infrequent, we can expect {beer, pizza} to be equally or even more infrequent.

- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - Support of an itemset never exceeds the support of its subsets
  - This is known as the anti-monotone property of support

Apriori Principle: Example

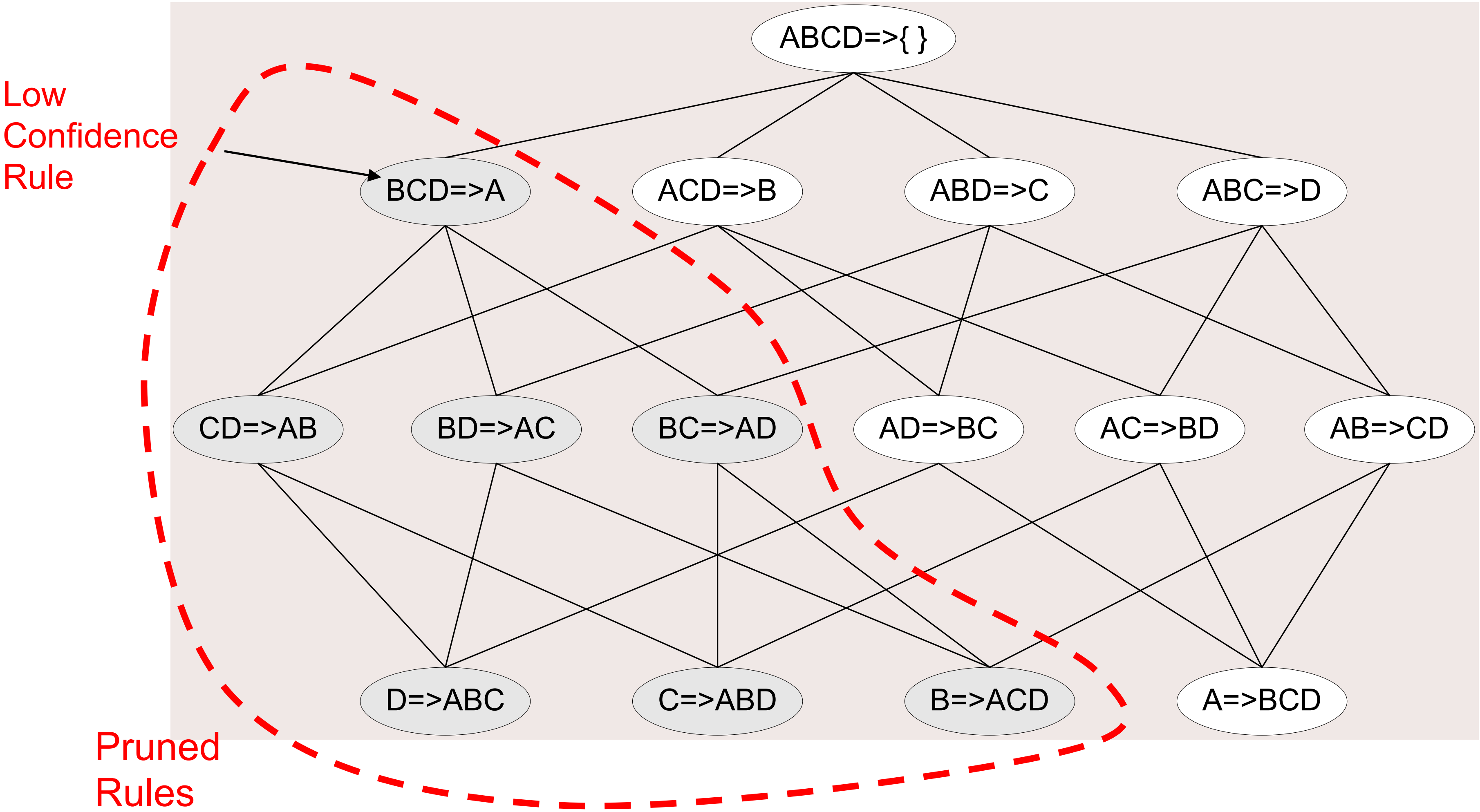


Pruning Hyperparameters

- Support Threshold ($minsup$)
  - Minimum lower bound for the Support measure of resulting association rules
  - E.g., If sales of items beyond a certain proportion tend to have a significant impact on profits, you might consider using that proportion as your support threshold.

- Confidence Threshold ($minconf$)
  - Identify rules that fall above a minimum confidence level ($minconf$).
  - Confidence does not show the *anti-monotone* property as the support measure
    - Knowing that $c(X \ -> \ Y) \ < \ minconf$, we cannot tell whether $c(X' \ -> \ Y') \ < minconf$ or $c(X' \ -> \ Y') > minconf$, for $X' \subseteq X$ and $Y' \subseteq Y$

Rule Generation: Example



Apriori Algorithm

- Using the apriori principle, the number of itemsets that have to be examined can be pruned, and the list of popular itemsets can be obtained in these steps:

- Start with itemsets containing just a single item, such as {apple} and {pear}.
- **Repeat**
  1. Determine the support for itemsets. Keep the itemsets that meet your minimum support threshold and remove itemsets that do not.
  2. Using the itemsets you have kept from Step 1, generate all the possible itemset configurations.
- **Until** there are no new itemsets.

Association Rules: Example

- All the below rules are binary partitions of the same itemset:

$$\{Milk, Diapers, Beer\}$$

- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

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

Rules:

$$\{Milk, Diapers\} \rightarrow \{Beer\}\ (s = 0.4, c = 0.67)$$
$$\{Milk, Beer\} \rightarrow \{Diapers\}\ (s = 0.4, c = 1.0)$$
$$\{Diapers, Beer\} \rightarrow \{Milk\}\ (s = 0.4, c = 0.67)$$
$$\{Beer\} \rightarrow \{Milk, Diapers\}\ (s = 0.4, c = 0.67)$$
$$\{Diapers\} \rightarrow \{Milk, Beer\}\ (s = 0.4, c = 0.5)$$
$$\{Milk\} \rightarrow \{Diapers, Beer\}\ (s = 0.4, c = 0.5)$$

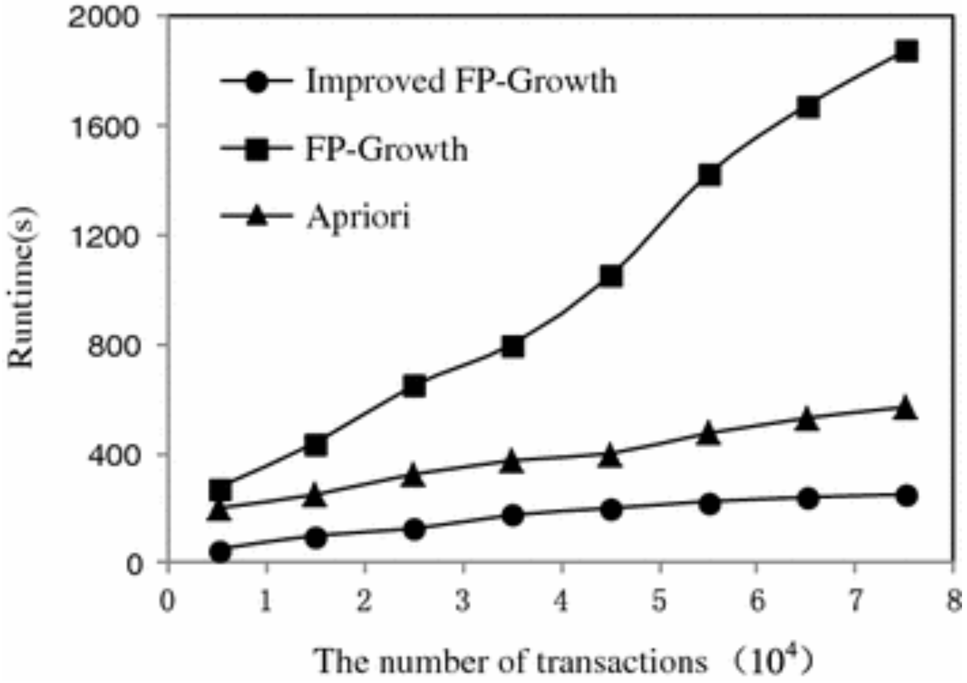
Apriori: Pros/Cons

- Pros
  - It is an easy-to-implement and easy-to-understand algorithm.
  - It can be used on large itemsets.
- Cons
  - Computationally Expensive
  - Choice of minimum support threshold
  - Spurious Associations

https://arxiv.org/ftp/arxiv/papers/1403/1403.3948.pdf

Frequent Pattern (FP) Growth

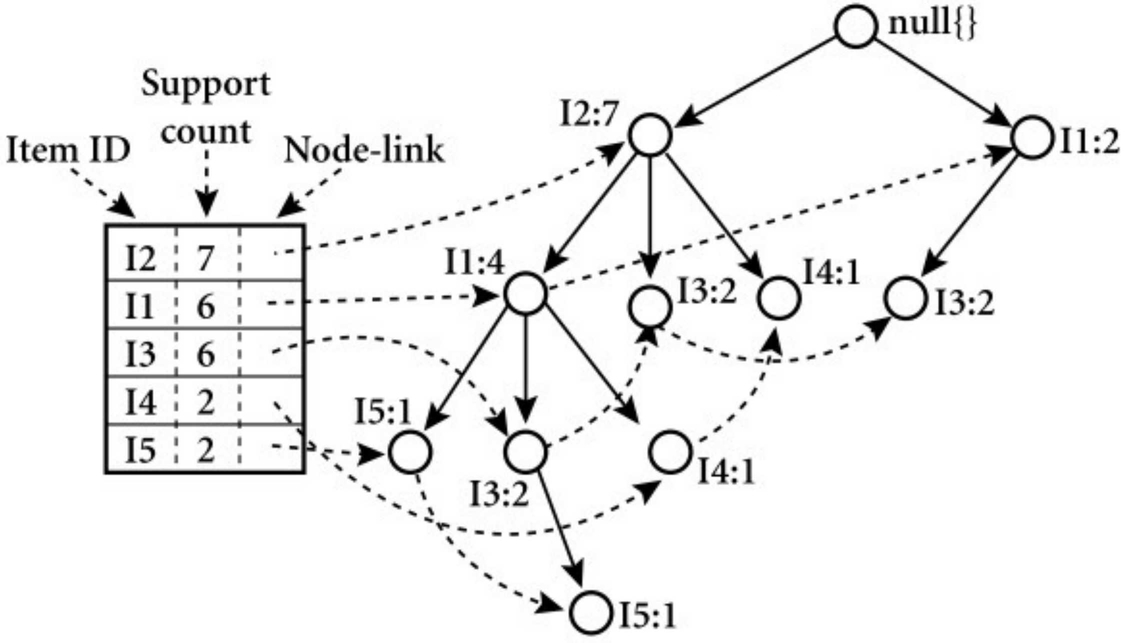
- Efficient and scalable method for using an FP-tree data structure
- FP-tree
  - Compact data structure that represents the data set in the tree form.
  - Each transaction is read and then mapped onto a path in the FP-tree. This is done until all transactions have been read.
  - Different transactions that have common subsets allow the tree to remain compact because their paths overlap.



https://link.springer.com/chapter/10.1007/978-3-662-49370-0_14

FP-Growth: Algorithm

- Scan the data set to determine the support count of each item, discard the infrequent items and sort the frequent items in decreasing order.
- Scan the data set one transaction at a time to create the FP-tree. For each transaction:
    - If it is a unique transaction form a new path and set the counter for each node to 1.
    - If it shares a common prefix itemset then increment the common itemset node counters and create new nodes if needed.
- Continue this until each transaction has been mapped unto the tree.



Item ID	Support count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	

FP-tree

Apriori vs FP-Growth

Algorithm	Technique	Runtime	Memory Usage	Parallelizability
Apriori	Generate singletons, pairs, triplets, etc.	Candidate generation is extremely slow. Runtime increases exponentially depending on the number of different items.	Saves singletons, pairs, triplets, etc.	Candidate generation is very parallelizable
FP-Growth	Insert sorted items by frequency into a pattern tree	Runtime increases linearly, depending on the number of transactions and items	Stores a compact version of the database	Data are very inter-dependent, each node needs the root

https://www.singularities.com/blog/our-blog-1/post/apriori-vs-fp-growth-for-frequent-item-set-mining-11