DATA MINING ASSOCIATION ANALYSIS: BASIC CONCEPTS AND ALGORITHMS

https://www.comparitech.com/blog/vpn-privacy/netflix-statistics-facts-figures/

50+ Netflix statistics and facts stats that define the company's dominance [2020 version]

As the largest streaming service in the world, Netflix maintains some of the most impressive use and income stats, among other interesting tidbits. Here are just a few of the many stats and facts that highlight the streaming service's rapid success and continuing dominance.

NETFLIX

39. Netflix's personalized recommendation engine could be worth $1 billion per year (or more)

According to Business Insider Australia, Netflix believes its personalized recommendation engine is worth big bucks; roughly $1 billion *per year*, in fact.

40. Around 80% of Netflix users take the streaming service's title recommendations offered by its algorithm

While users quite often land on the most popular shows, the vast majority of Netflix users also click to watch the recommended shows from the Netflix recommendation algorithm. 80% of Netflix views were from the service's recommendations.

Each shopper has a distinctive list, depending on one's needs and preferences.

While we may know that certain items are frequently bought together, the question is, **how do we uncover these associations?**

WHY KNOW THESE ASSOCIATIONS?

If there is a pair of items, X and Y, that are frequently bought together:

- ✓ Both X and Y can be placed on the same shelf, so that buyers of one item would be prompted to buy the other.
- ✓ 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.

ASSOCIATION RULE LOOKS LIKE

Antecedent

Consequent

Note that implication here is co-occurrence and not causality

ASSOCIATION RULE LOOKS LIKE

Antecedent

Consequent

SAMPLE

{Bread, Egg} → {Milk}

Antecedent — Consequent

Itemset = {Bread, Egg, Milk}

ITEMSET

A collection of one or more items

Example: {Bread,Egg,Milk}

k-itemset

An itemset that contains k items

Ex. {X,Y} is a representation of the list of all items which form the association rule

SUPPORT

Mathematically, support is the fraction of the total number of transactions in which the itemset occurs.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

Example:

For itemsets which occur at least 100 times out of a total of 10,000 transactions

What is the support?

SUPPORT

Mathematically, support is the fraction of the total number of transactions in which the itemset occurs.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

If an itemset happens to have a very low support, we do not have enough information on the relationship between its items and hence no conclusions can be drawn from such a rule.

SUPPORT

This measure gives an idea of how frequent an itemset is in all the transactions.

Which has the higher support?	
itemset1 = {bread}	itemset2 = {shampoo}
itemset1 = {bread, butter} itemset2 = {bread, shampoo}	

Value of support helps us identify the rules worth considering for further analysis.

CONFIDENCE

This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents.

Example:

That is to answer the question — of all the transactions containing say, {Conditioner}, how many also had {Shampoo} on them?

CONFIDENCE

This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

Technically, confidence is the conditional probability of occurrence of consequent given the antecedent.

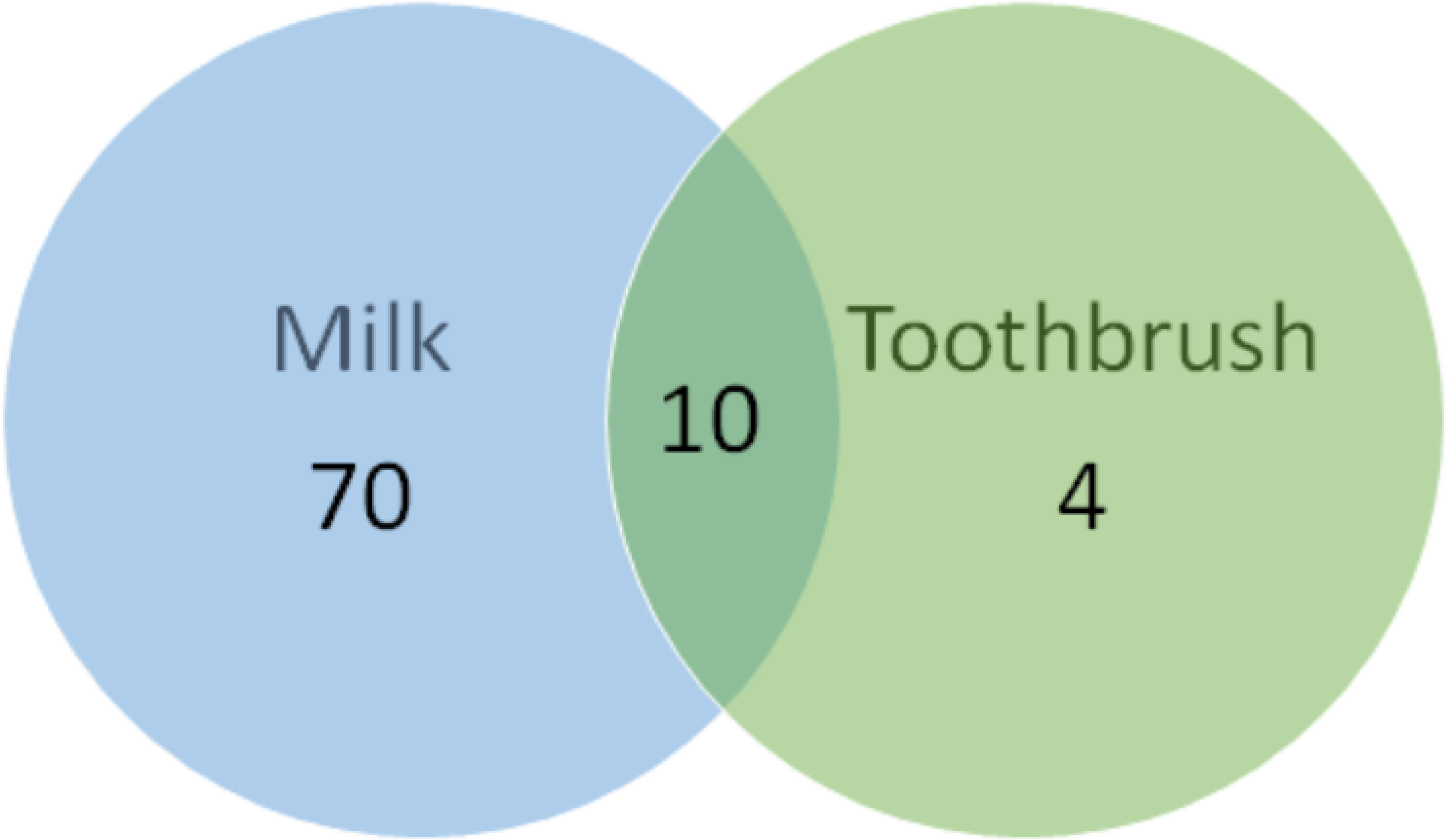
CONFIDENCE

This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents.

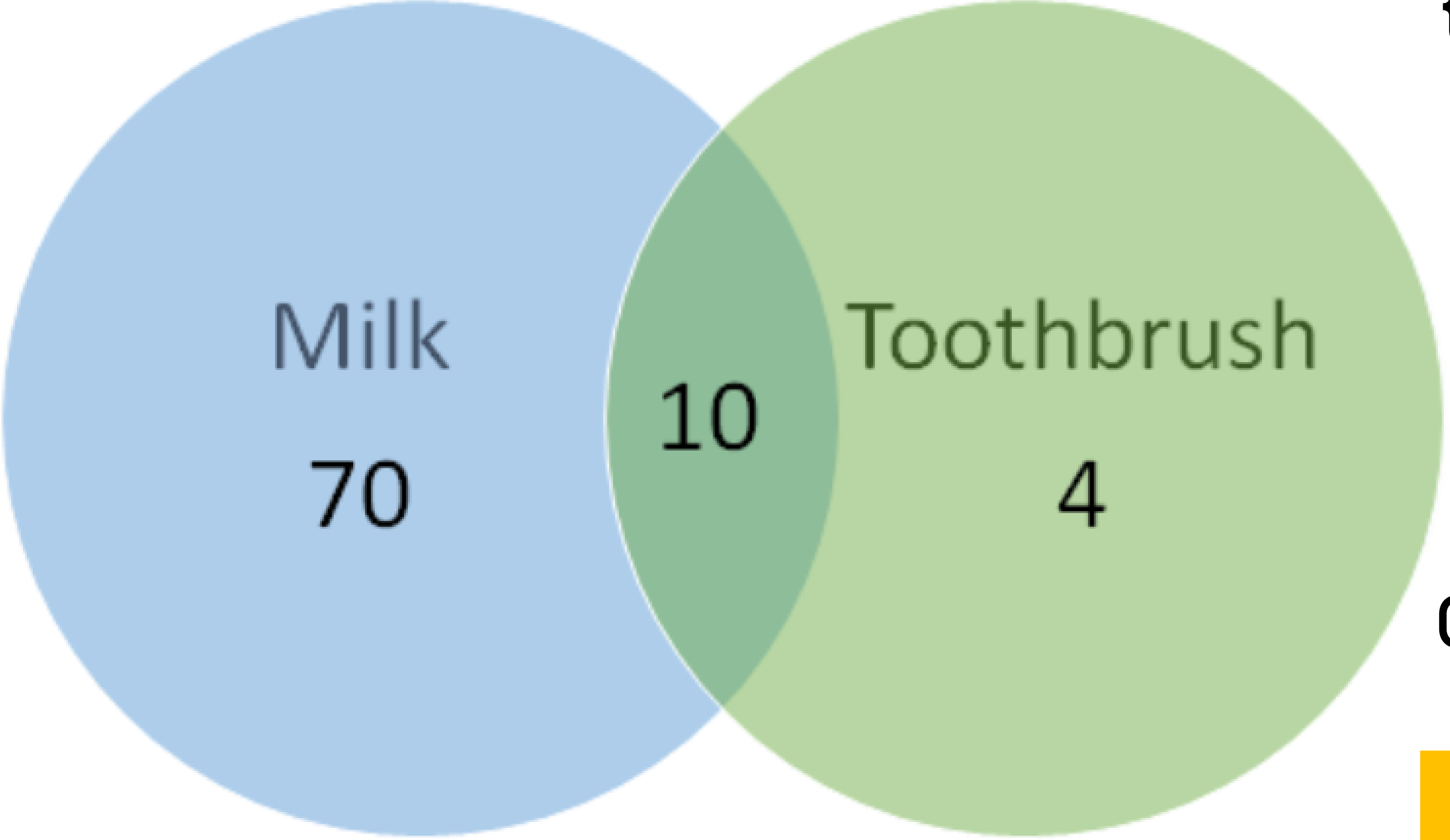
$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

Technically, confidence is the conditional probability of occurrence of consequent given the antecedent.

CONFIDENCE



CONFIDENCE



Confidence for {Toothbrush} → {Milk}?

10/(10+4)
= 0.7

Oops! Looks like a high confidence value.

But we know intuitively that these two products have a weak association.

BUT

Considering just the value of confidence limits our capability to make any business inference.

SUPPORT

Mathematically, support is the fraction of the total number of transactions in which the itemset occurs.

LIFT

Lift controls for the support (frequency) of consequent while calculating the conditional probability of occurrence of {Y} given {X}.

LIFT is the rise in probability of having {Y} on the cart with the knowledge of {X} being present over the probability of having {Y} on the cart without any knowledge about presence of {X}

LIFT

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{\boxed{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}}{Fraction\ of\ transactions\ containing\ Y}$$

LIFT is the rise in probability of having {Y} on the cart with the knowledge of {X} being present over the probability of having {Y} on the cart without any knowledge about presence of {X}

LIFT

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

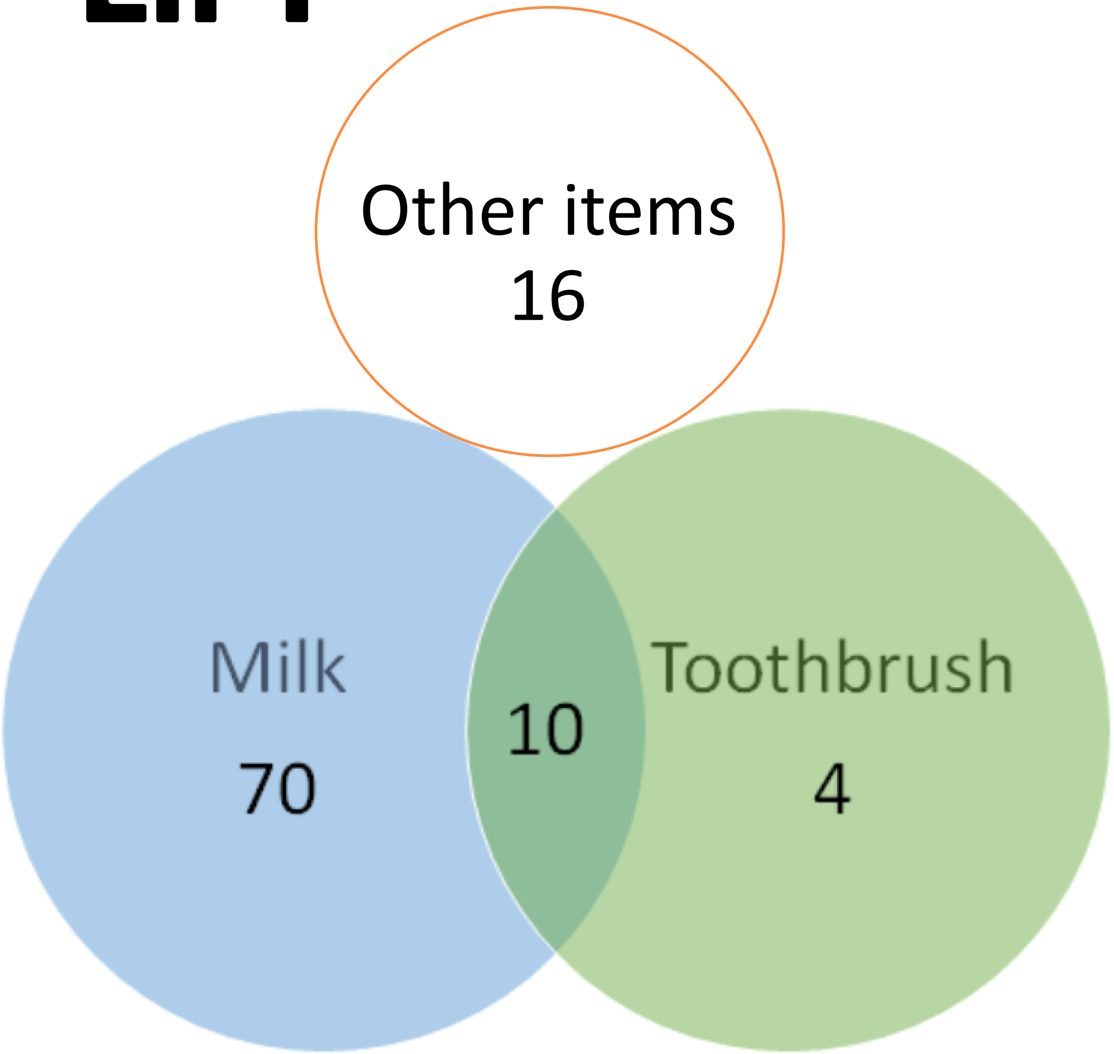
$$Lift(\{X\} \rightarrow \{Y\}) = \frac{\boxed{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}}{Fraction\ of\ transactions\ containing\ Y}$$

LIFT is the rise in probability of having {Y} on the cart with the knowledge of {X} being present over the probability of having {Y} on the cart without any knowledge about presence of {X}

LIFT

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{\boxed{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}}{Fraction\ of\ transactions\ containing\ Y}$$



{Toothbrush} → {Milk}

Probability of having milk on the cart with the knowledge that toothbrush is present

Confidence : 10/(10+4) = 0.7

Probability of having milk on the cart without any knowledge about toothbrush

80/100 = 0.8

These numbers show that having toothbrush on the cart actually reduces the probability of having milk on the cart from 0.8 to 0.7

LIFT

$$Confidence(\{X\} \to \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

$$Lift(\{X\} \to \{Y\}) = \frac{\boxed{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}}{Fraction\ of\ transactions\ containing\ Y}$$

{Toothbrush} → {Milk}

Probability of having milk on the cart with the knowledge that toothbrush is present

Probability of having milk on the cart without any knowledge about toothbrush

$$\frac{0.7}{0.8} = 0.87$$

Now this is the lift!

These numbers show that having toothbrush on the cart actually reduces the probability of having milk on the cart from 0.8 to 0.7

LIFT

{Toothbrush} → {Milk}

$$\frac{0.7}{0.8} = 0.87$$

Now this is the lift!

A value of lift less than 1 shows that having toothbrush on the cart does not increase the chances of occurrence of milk on the cart in spite of the rule showing a high confidence value.

A value of lift greater than 1 vouches for high association between {Y} and {X}. More the value of lift, greater are the chances of preference to buy {Y} if the customer has already bought {X}.

LIFT

{Toothbrush} → {Milk}

Lift is the measure that will help store managers to decide product placements on aisle.

A value of lift less than 1 shows that having toothbrush on the cart does not increase the chances of occurrence of milk on the cart in spite of the rule showing a high confidence value.

A value of lift greater than 1 vouches for high association between {Y} and {X}. More the value of lift, greater are the chances of preference to buy {Y} if the customer has already bought {X}.

ASSOCIATION RULE MINING

We now understand how to quantify the importance of association of products within an itemset

The next step is to
(1) generate rules from the entire list of items and
(2) identify the most important ones

The next step is to
(1) generate rules from the entire list of items
(2) identify the most important ones

First step in generation of association rules is to **get all the frequent itemsets on which binary partitions can be performed** to get the antecedent and the consequent.

For example, if there are 3 items on all transactions

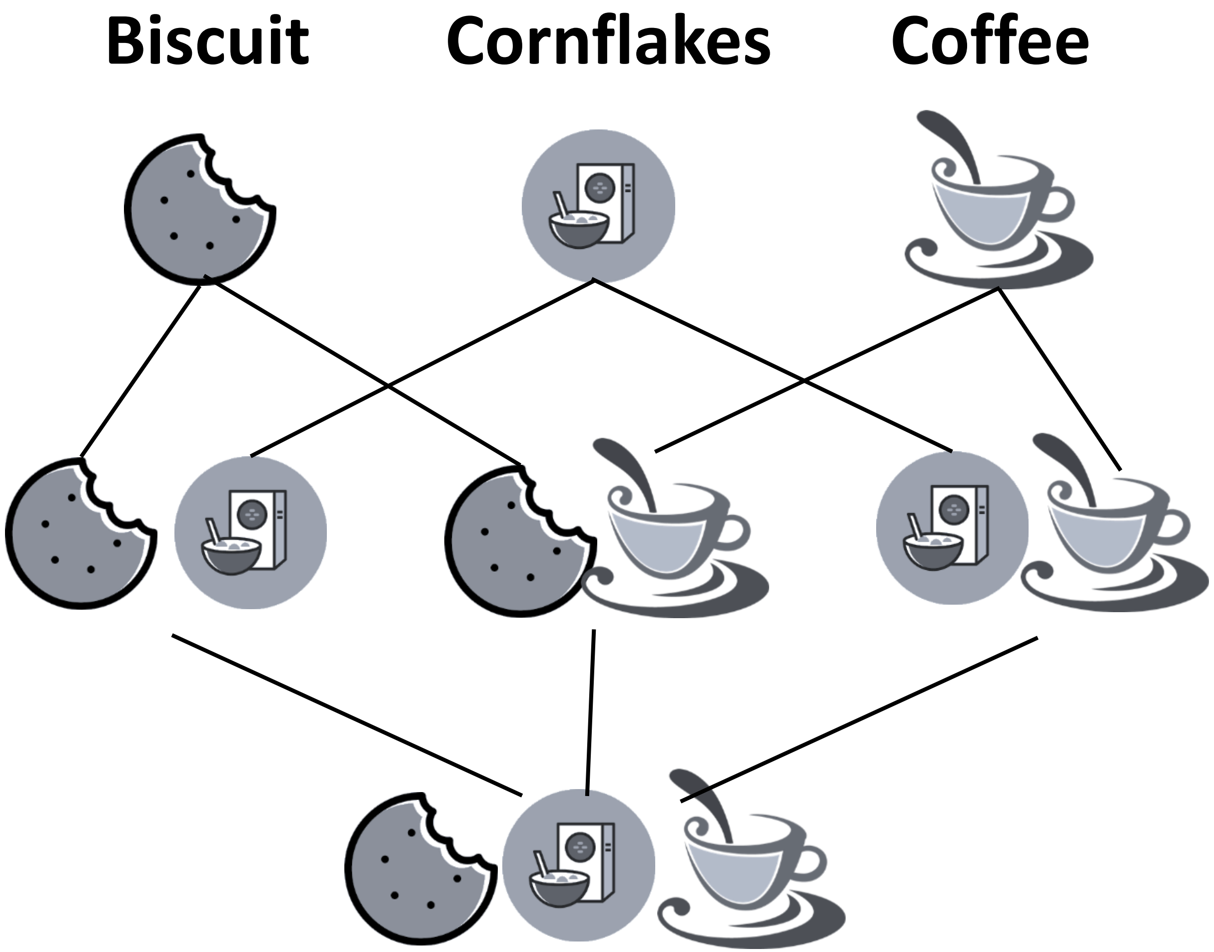
{Biscuit, Cornflakes, Coffee}

How many itemsets can be generated?

ASSOCIATION RULE MINING

The next step is to
(1) generate rules from the entire list of items
(2) identify the most important ones

HOW MANY ITEMSETS CAN BE GENERATED?



For example, if there are 3 items on all transactions

{Biscuit, Cornflakes, Coffee}

1 {Biscuit}
2 {Cornflakes}
3 {Coffee}
4 {Biscuit, Cornflakes}
5 {Biscuit, Coffee}
6 {Cornflakes, Coffee}
7 {Biscuit, Cornflakes, Coffee}

ASSOCIATION RULE MINING

The next step is to
(1) generate rules from the entire list of items
(2) identify the most important ones

For example, if there are 3 items on all transactions

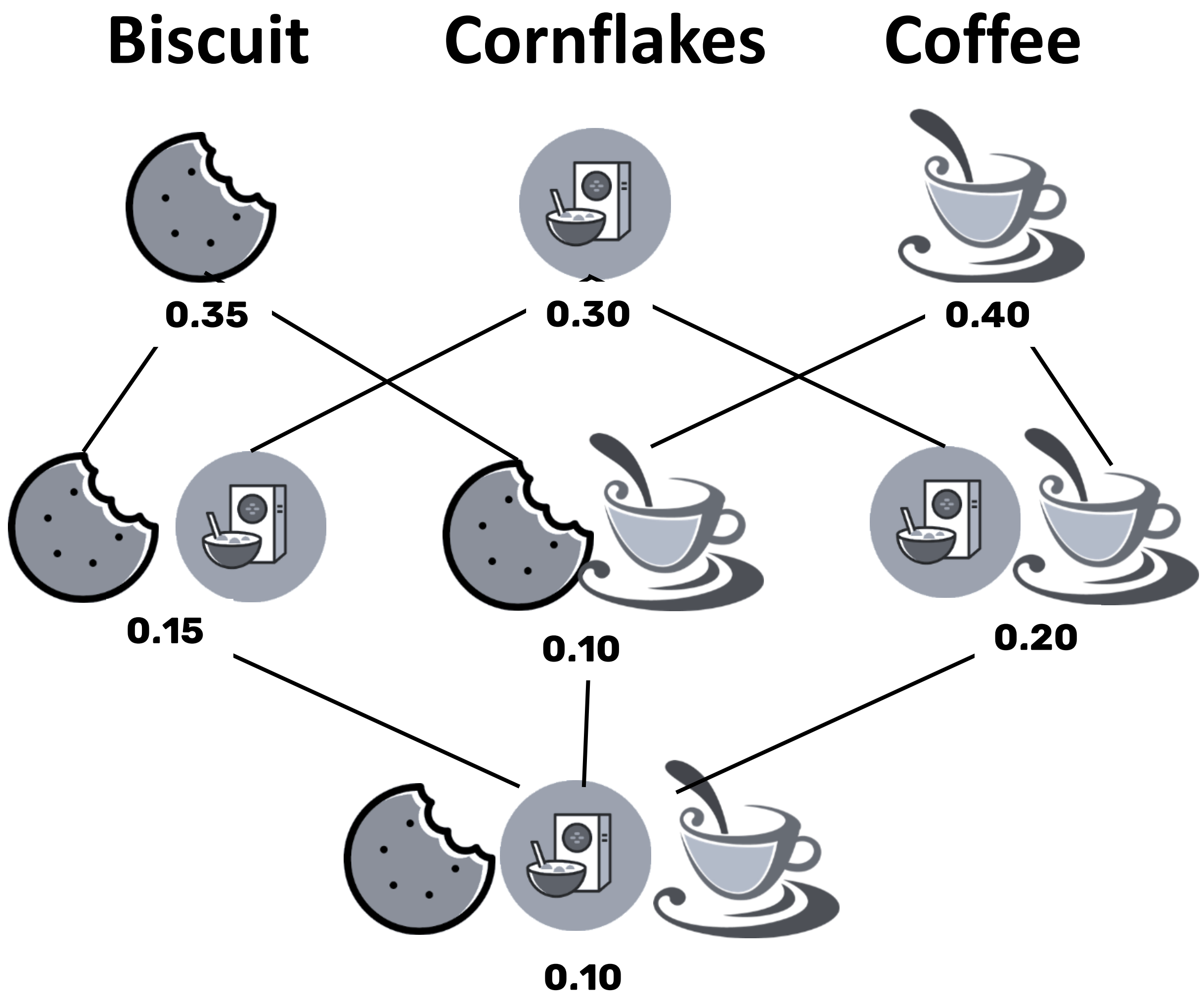
{Biscuit, Cornflakes, Coffee}

1 {Biscuit}
2 {Cornflakes}
3 {Coffee}
4 {Biscuit, Cornflakes}
5 {Biscuit, Coffee}
6 {Cornflakes, Coffee}
7 {Biscuit, Cornflakes, Coffee}

ASSOCIATION RULE MINING

The next step is to
(1) generate rules from the entire list of items
(2) identify the most important ones

HOW MANY ITEMSETS CAN BE GENERATED?



This what happens if we use
BRUTE FORCE

Every itemset will be generated

	support	itemsets
0	0.35	(BISCUIT)
1	0.30	(CORNFLAKES)
2	0.40	(COFFEE)
3	0.15	(BISCUIT, CORNFLAKES)
4	0.10	(COFFEE, BISCUIT)
5	0.20	(COFFEE, CORNFLAKES)
6	0.10	(COFFEE, BISCUIT, CORNFLAKES)

APRIORI PRINCIPLES

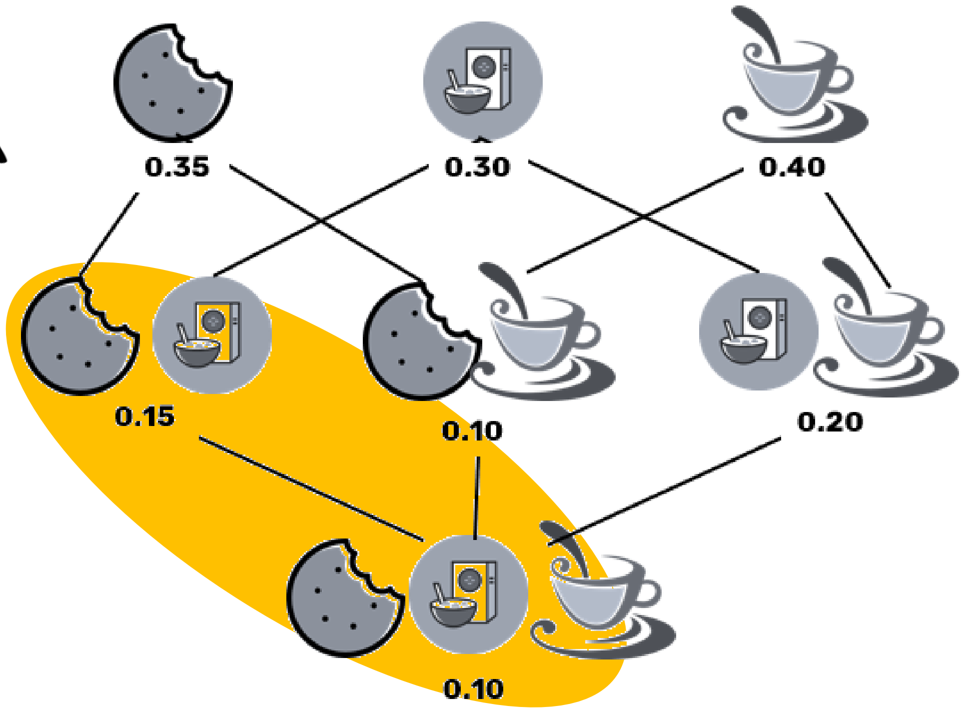
All subsets of a frequent itemset must also be frequent.

Remember this?

As you can see as you move up, each subset has greater than or equal support compared to the super set.

Example:

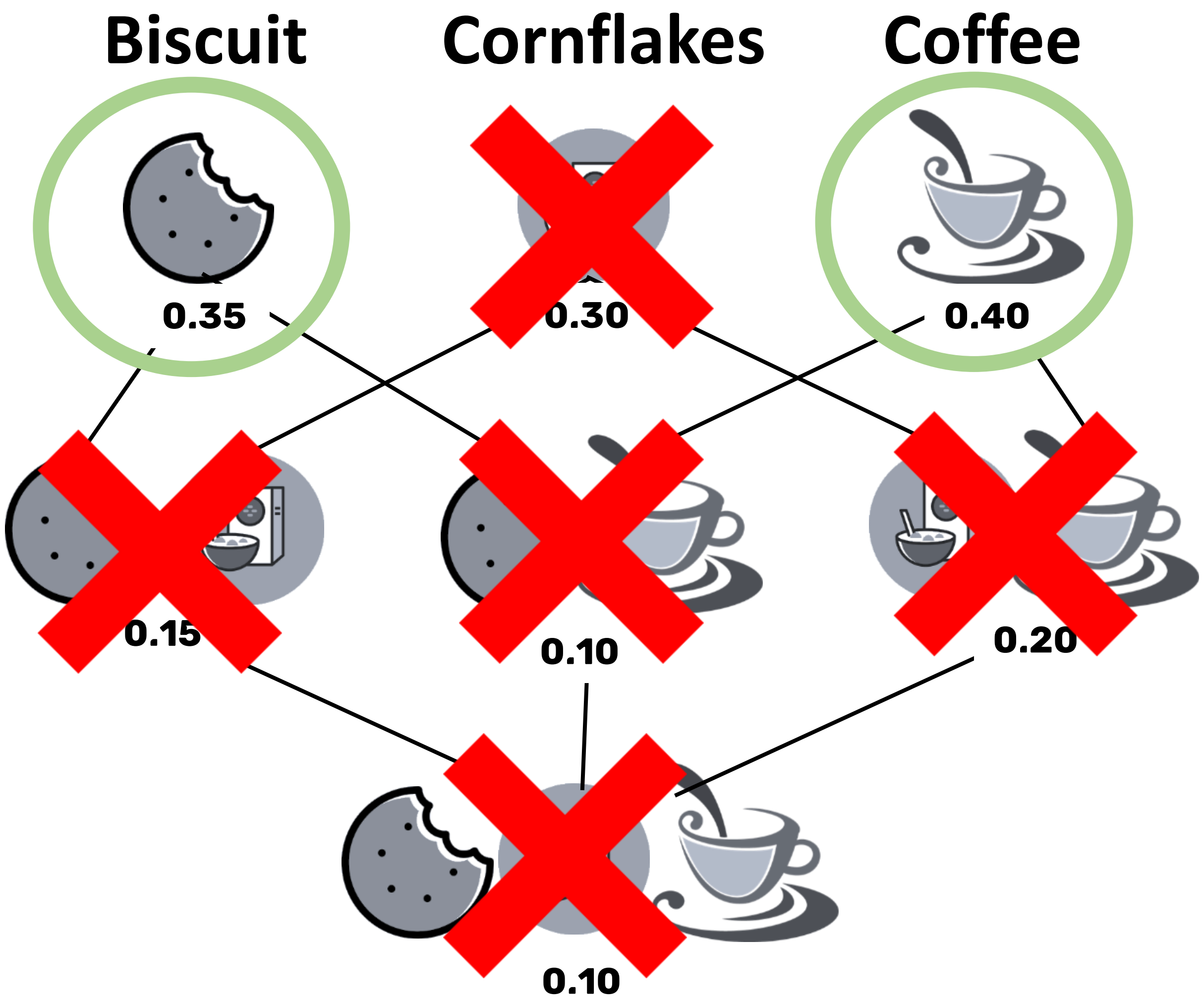
Support of {Biscuit, Cornflakes} > Support of {Biscuit, Cornflakes, Coffee}



ASSOCIATION RULE MINING

The next step is to
(1) generate rules from the entire list of items
(2) identify the most important ones

HOW MANY ITEMSETS ARE FREQUENT?



Given minimum support (minsup) of 0.33

USING BRUTE FORCE

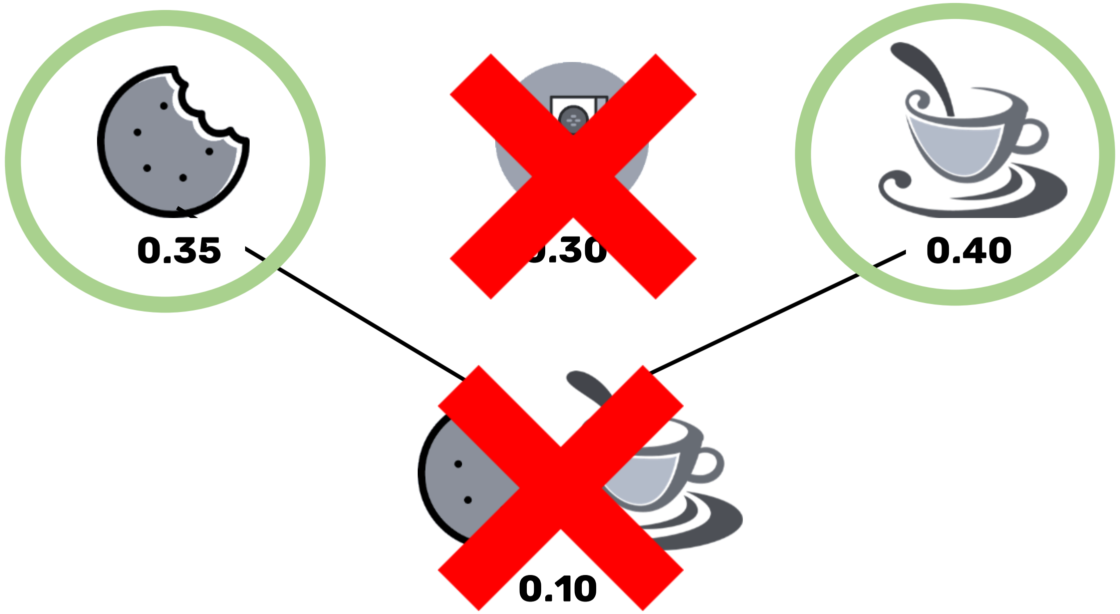
1. Brute force algorithm will first generate the list of all itemsets and compute their support
2. It will scan the whole database to check which itemsets have greater support than the minsup

ASSOCIATION RULE MINING

The next step is to
(1) generate rules from the entire list of items
(2) identify the most important ones

HOW MANY ITEMSETS ARE FREQUENT?

Biscuit Cornflakes Coffee



GIVEN NO MORE TO TEST FOR MINSUP, THE ALGORITHM STOPS AND REPORTS THE IDENTIFIED FREQUENT ITEMSETS

Given minimum support (minsup) of 0.33

USING APRIORI

1. Take each basic itemset
2. Compute support
3. Evaluate if greater than minsup, if yes then continue, if not don't consider the itemset as frequent
4. Go the identified frequent, go to the next level itemset and go back to step (2)