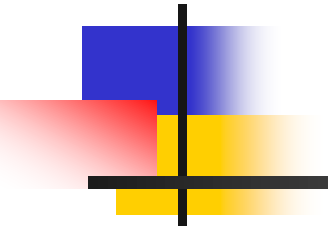


Association Analysis:
Pattern Discovery in Data

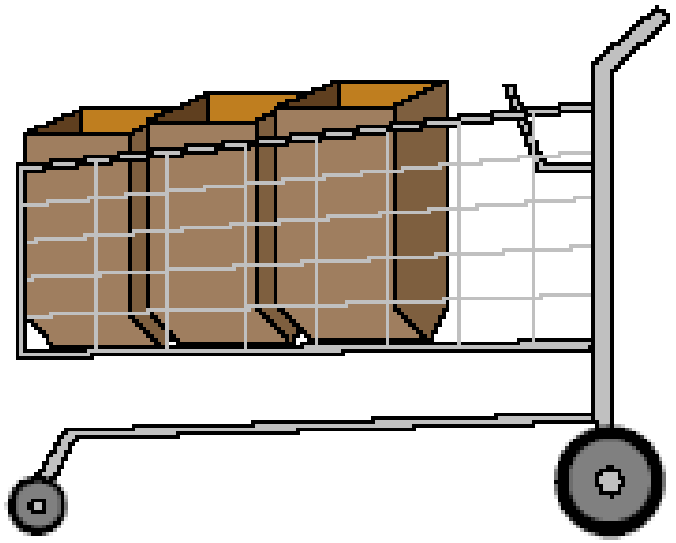
Dr. Goutam Chakraborty

Association Rules

Association rules specify which events are likely to occur together.

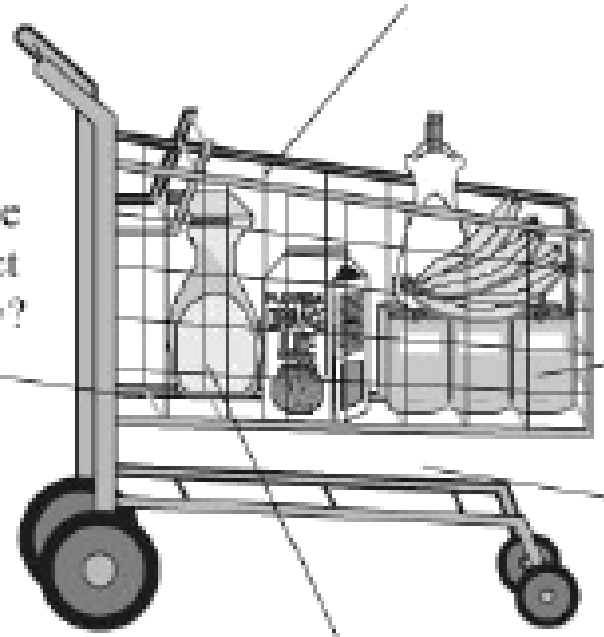
Market basket analysis includes association rules. However, market basket analysis is a more general term for retail analysis.

Sequential pattern analysis looks at the order that things occur in as well.



Questions in a Shopping Cart

In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, some window cleaner, and a six pack of soda.



How do the demographics of the neighborhood affect what customers buy?

Is soda typically purchased with bananas? Does the brand of soda make a difference?

What should be in the basket but is not?

Are window cleaning products purchased when detergent and orange juice are bought together?

4

Association Rules (Basics)

- Typical type of questions that are addressed by this type of analysis include (note that the questions go beyond retail environment):
  - When people buy milk do they also tend to buy chocolate or beer?
  - If people have high cholesterol and high blood sugar do they also tend to have high blood pressure?
  - If people claim rain damage and theft damage does that suggest possibility of insurance fraud?
- Although sounds simple, in practice with large data sets and many variables, it's very difficult to generate these types of rules unless we have efficient algorithms.

Basic Ideas Behind Association Rules

- Works using counts and probabilities of things (purchases) happening together
  - Level of data needed is often simple and available at POS (transaction records at check out counters with possibly date/time stamp and often anonymous)
- Algorithms first creates a co-occurrence matrix that tells us number of times any pairs/triplets of products bought together (i.e., co-occur).
- Some examples follow

Grocery POS Transactions of 5 Customers

Customer	*Items*
1	orange juice, soda
2	milk, orange juice, window cleaner
3	orange juice, detergent
4	orange juice, detergent, soda
5	window cleaner, soda

Note that OJ (orange juice) was bought 4 times, Soda was bought 3 times and so on. Also, note that OJ and Soda were bought together 2 times.
A co-occurrence matrix will create a 5X5 table (there are 5 distinct products in all of these transactions) showing how each product co-occur with every other product

Co-occurrence Matrix for Pairs of Products

	OJ	Window Cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

Note in the co-occurrence matrix, the diagonal reflect the number of times a particular item was bought.

Some simple rules from this co-occurrence table are:

1) OJ & Soda are more likely to be purchased together than any other pair
2) Window cleaner is never purchased with Detergent
3) Milk is never purchased with Soda or Detergent

Probabilities Become Rules

Three important measures of rules:

- **Support** is the proportion of market baskets where the rule is true.
- **Confidence** measures how often the right side of the rule holds, given the left side.
- **Lift** (also called **improvement**) measures how much better the rule is for prediction than a random guess.
  - Values greater than one indicate a useful rule.

Basic Ideas (contd.)

- In the data, *two* out of the *five* transactions support the rule that **if a customer buys OJ, he/she also buys soda**
  - The support level is 40%
  - Support calculation is **Reflexive (or, Symmetric)**
  - If a customer buys Soda, the chance that he/she buys OJ is 40%

	OJ	*Window Cleaner*	*Milk*	*Soda*	*Detergent*
OJ	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

Basic Ideas (contd.)

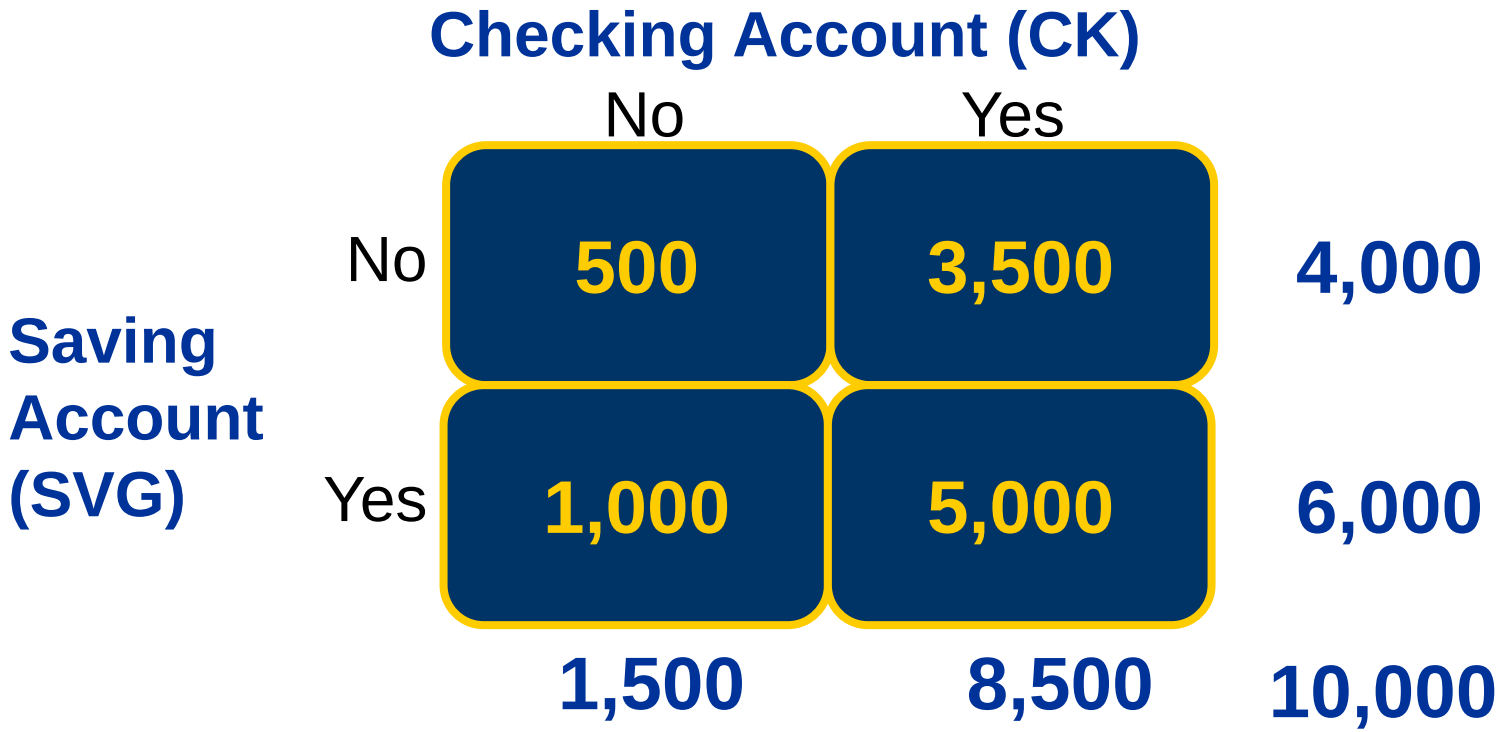
- Confidence is the ratio of number of transactions supporting the rule to the number of transactions where only the conditional part of the rule is true.
- Since two transactions involving Soda purchase (out of 3) also contain OJ, then confidence in the rule "**If Soda, then OJ**" is 66.7%
  - Confidence calculation is **Not Reflexive (asymmetric)**.
  - For e.g.,, "*If OJ, then Soda*" has only 50% confidence

	OJ	*Window Cleaner*	*Milk*	*Soda*	*Detergent*
OJ	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

Basic Ideas (contd.)

- The lift of a rule is the confidence of a rule divided by the expected confidence, assuming items sets are independent.
- Calculation of lift requires understanding of independence in a co-occurrence matrix
- Lift values greater than 1 for lift indicates an useful rule
- A simple example with two products (Checking and Savings account in a bank) follows
- Note that calculation of lift **is Reflexive** (that is, lift of rule A=>B is the same as the lift of rule B=>A

Implication?



Support(SVG ⇒ CK) = 50%

Confidence(SVG ⇒ CK) = 83%

Expected Confidence(SVG ⇒ CK) = 85%

Lift(SVG ⇒ CK) = 0.83/0.85 < 1

Basic Ideas (contd.)

- These ideas can be generalized to any number of items, not *just pairs*
- For example, we can have co-occurrence among triplets, quadruplets etc.
  - Clearly, the number of combinations can get very large very quickly
  - Imagine a typical retail store containing 20,000+ SKUs!
  - Thus, the need for having efficient algorithms to sort through these data becomes important

How can businesses use these rules?

- *Forbes* (Palmeri 1997) reported that a major retailers has determined that customers who buy Barbie dolls have a 60% likelihood of buying one of the three types of candies.
- Is rule (Barbie => Candy) support or rule confidence 60%?
- Does this rule make sense?
  - Can we think of typical situations where families/customers may buy these items together?
- What can a retailer do now that this rule (Barbie => Candy) has been discovered?

Barbie® ⇒ Candy Rule Potential Usage

1. Put them closer together in the store.
2. Put them far apart in the store.
3. Package candy bars with the dolls.
4. Package Barbie + candy + poorly selling item.
5. Raise the price on one, lower it on the other.
6. Offer Barbie accessories for proofs of purchase of candies
7. Do not advertise candy and Barbie together
8. And many others…

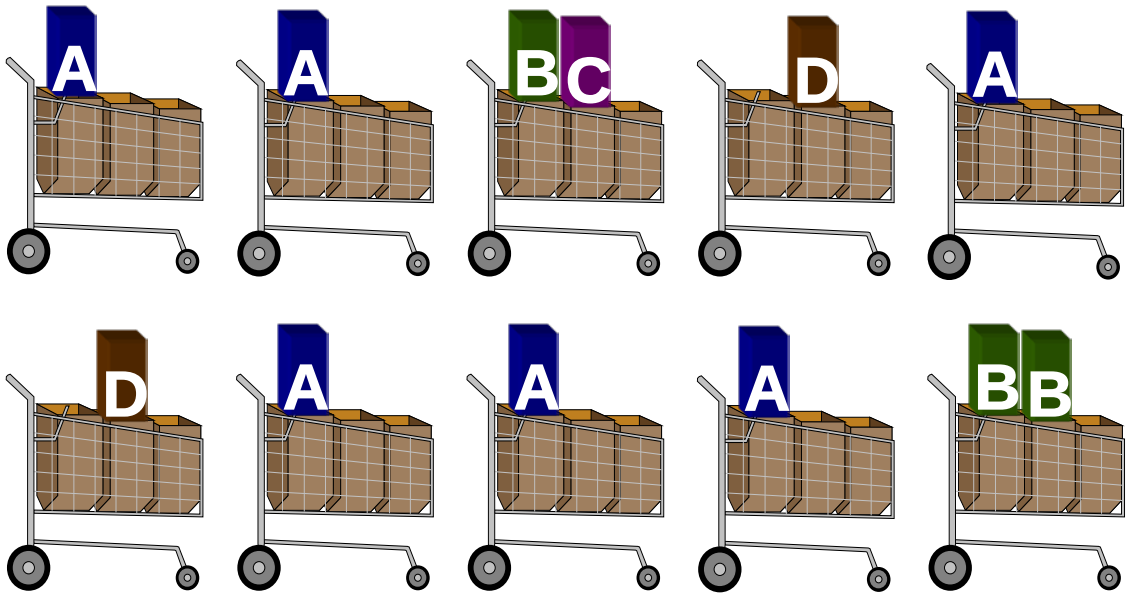
MBA may not always identify useful rules...

- **Useful rule** (actionable information with plausible explanation)
  - On Thursdays, consumers often purchase diapers and beers together
- **Trivial rule** (known to everyone or artifact of business practices)
  - Consumers who buy extended warranties are likely to buy large appliances
  - If a customer purchases three-way calling, he/she will also buy call waiting
- **Inexplicable rule**
  - When a new hardware store opens, one of the most commonly sold items is “toilet bowl cleaners”

Issues in Using Market Basket Analysis

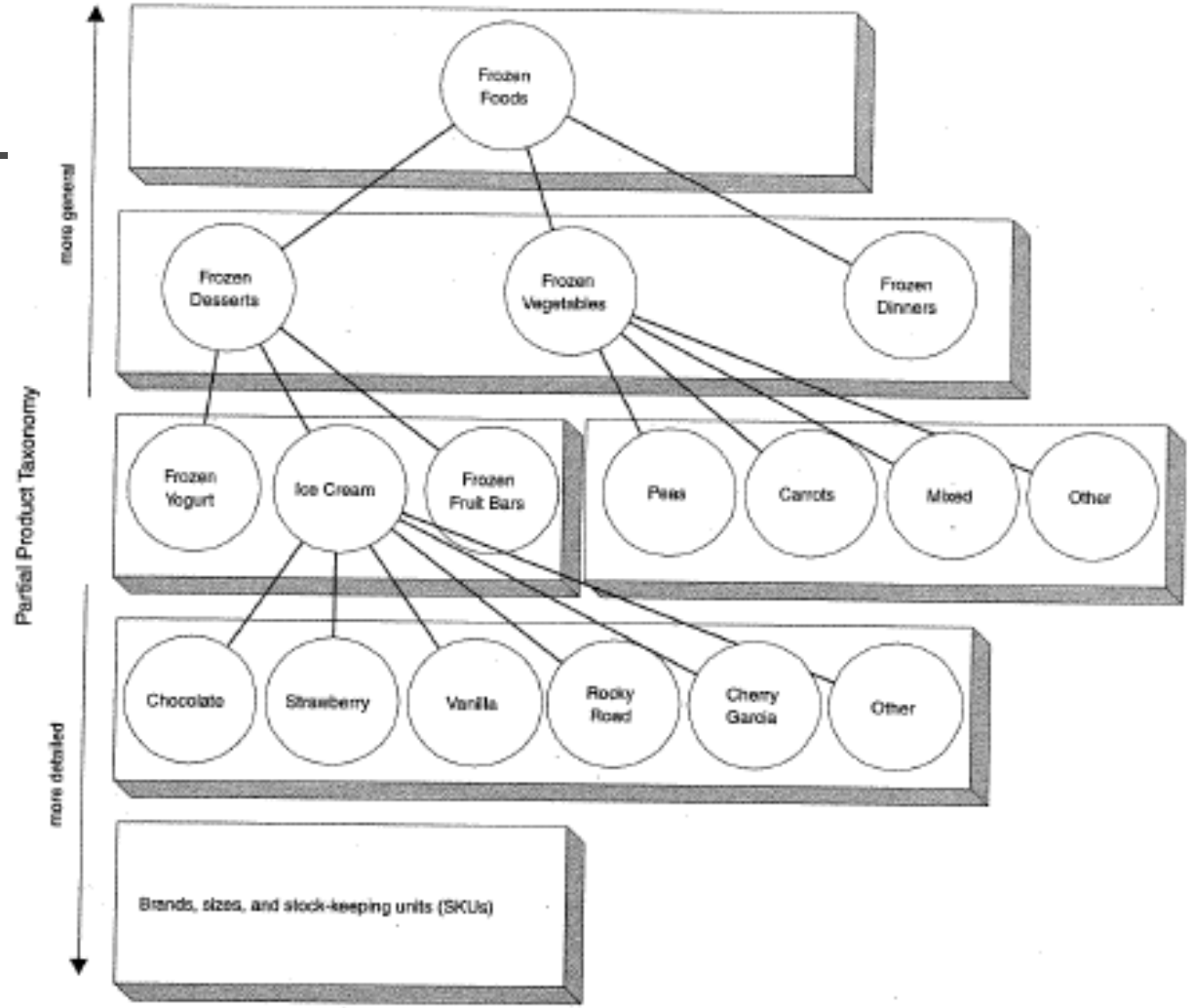
- Choosing the right set of items
- Generating rules by using algorithms
- Overcoming practical limits imposed by large number of items appearing in combinations large enough to seem interesting

Data Capacity (Aggregation) Issues in Association Analysis



Ask if the data, as it exists, has the capacity to meet the objectives. For example, quantifying affinities among related items would be pointless if very few transactions involved multiple items.

Taxonomy of Products in one Department of a Grocery Store

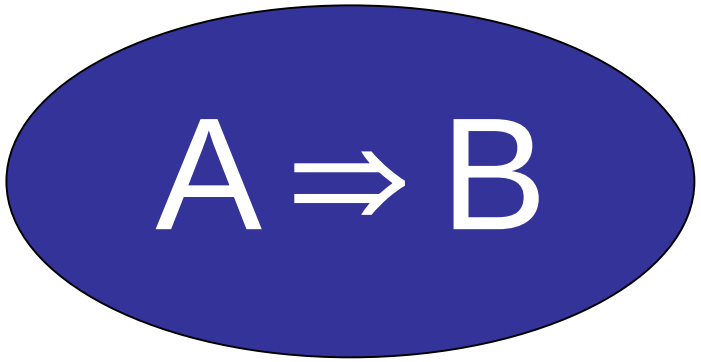


What level of taxonomy to use for market basket analysis often depends on business objectives, data availability etc.

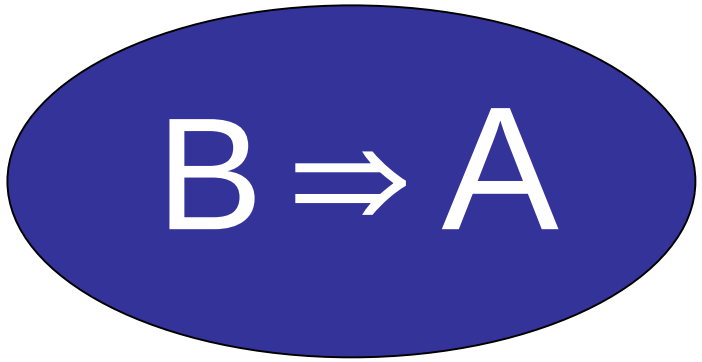
Other issues in Using Association Rules

- How do we use demographics in Association rules?
  - We can use demographics (after appropriate binning if these are numeric fields) as antecedents along with the rules
  - Or, we could investigate profile of demographics of customers of a particular rule to get more insights
- Using other “virtual variables”
  - such as new/old store if we have data pulled over multiple stores
  - Date of purchase or season or geographic locations (multiple store data)
- Use training/validation data to explore stability of rules

Association or Sequence?



versus



Is there a difference?
Is the order meaningful?
Do we have data about the order of events?

Sequence Analysis

- Typical Goal: To determine common sequences in time-ordered data and sequences associated with an event
- Business questions:
  - What are the most common sequences of web-clicks in a retailer site that lead to a purchase?
  - In a service repair situation, what are the common sequences that lead to a success or failure?
  - If a customer purchases wine this week, how likely is he/she to purchase beer next week?

Sequence Analysis

- The meaning of support and Confidence is same as in Association analysis. Except, **the sequence of events** make a difference in this analysis.
  - Transaction count is the total number of customers who purchased the products **in the order shown in Rule**
  - The percent *support* is the transaction count divided by the total number of customers
  - The percent *confidence* is the transaction count divided by the transaction count for the left side of the sequence