

Association Rule - Market Basket Analysis

Presented by: Derek Kane

Overview of Topics

- ❖ Association Rules
 - ❖ Basic Terminology
 - ❖ Support
 - ❖ Confidence
 - ❖ Lift
- ❖ Apriori Algorithm
- ❖ Practical Examples
 - ❖ Grocery Shopping Basket Analysis
 - ❖ Voting Patterns in the House of Representatives



Association Rules

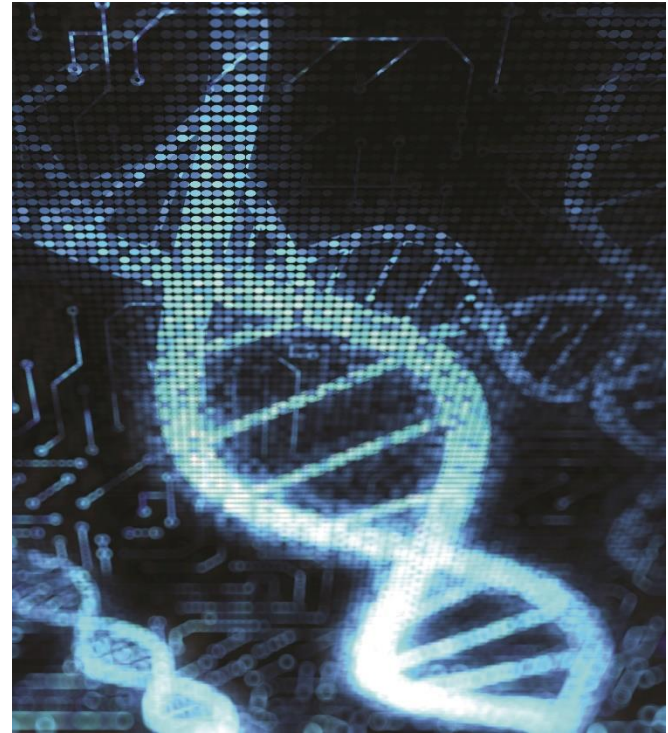


- ❖ A series of methodologies for discovering interesting relationships between variables in a database.
- ❖ The outcome of this technique, in simple terms, is a set of rules that can be understood as “if this, then that”.
- ❖ An example of an association rule would be:
- ❖ If a person buys Peanut Butter and Bread then they will also purchase Jelly.

Association Rules

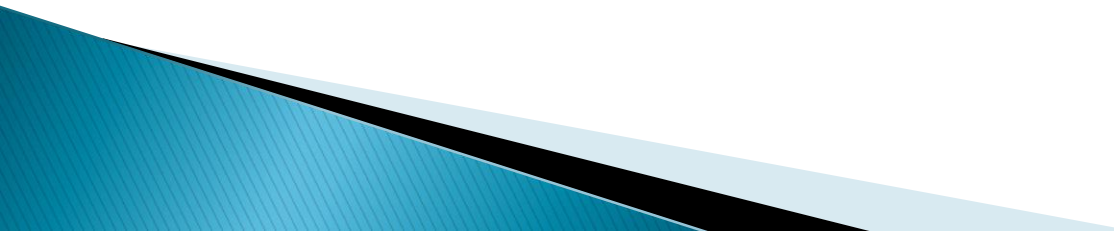
Applications

- ❖ Product recommendation
- ❖ Digital Media recommendations
- ❖ Politics
- ❖ Medical diagnosis
- ❖ Content optimisation
- ❖ Bioinformatics
- ❖ Web mining
- ❖ Scientific data analysis



- ❖ Example: The analysis of earth science data may reveal interesting connections among the ocean, land, and atmospheric processes. This may help scientists to better understand how these systems interact with one another.

Association Rules

- ❖ For example, maybe people who buy flour and casting sugar, also tend to buy eggs (because a high proportion of them are planning on baking a cake).
 - ❖ A retailer can use this information to inform:
 - ❖ Store layout (put products that co-occur together close to one another, to improve the customer shopping experience).
 - ❖ Marketing (e.g. target customers who buy flour with offers on eggs, to encourage them to spend more on their shopping basket).
 - ❖ Retailers can use these type of rules to identify new opportunities for cross selling/upselling their products to their customers.
- 

Company Profile - Netflix



- ❖ Asked engineers and scientists around the world to solve what might have seemed like a simple problem: improve Netflix's ability to predict what movies users would like by a modest 10%.
- ❖ From \$5 million revenue in 1999 reached \$3.2 billion revenue in 2011 as a result of becoming an analytics competitor.
- ❖ By analyzing customer behavior and buying patterns created a recommendation engine which optimizes both customer tastes and inventory condition.

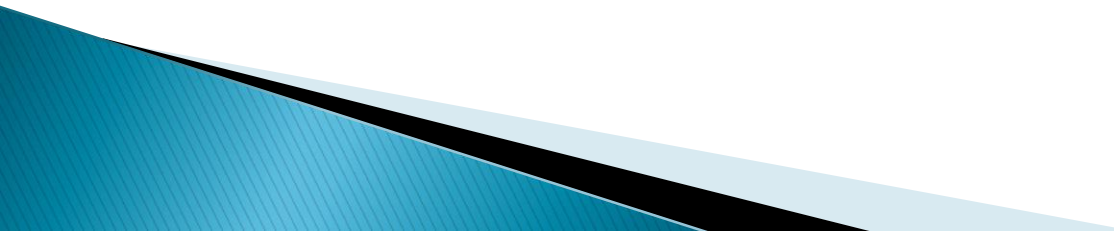
Basic Terminology & Mathematics

- ❖ A rule is typically written in the following format: $\{ i_1, i_2 \} \Rightarrow \{ i_k \}$
- ❖ The $\{ i_1, i_2 \}$ represents the left hand side, LHS, of the rule and the $\{ i_k \}$ represents the right hand side, RHS.
- ❖ This statement can be read as “if a user buys an item in the item set on the left hand side, then the user will likely buy the item on the right hand side too”.
- ❖ A more human readable example is:

$$\{\text{coffee, sugar}\} \Rightarrow \{\text{milk}\}$$

- ❖ If a customer buys coffee and sugar, then they are also likely to buy milk.

Basic Terminology & Mathematics

- ❖ Before we can begin to employ association rules, we must first understand three important ratios; the support, confidence and lift.
 - ❖ **Support:** The fraction of which our item set occurs in our dataset.
 - ❖ **Confidence:** Probability that a rule is correct for a new transaction with items on the left.
 - ❖ **Lift:** The ratio by which the confidence of a rule exceeds the expected confidence.
- 

Support

- ❖ The support of an item or item set is the fraction of transactions in our data set that contain that item or item set.
- ❖ Ex. A grocer has 15 transactions in total. Of which, Peanut Butter => Jelly appears 6 times. The support for this rule is $6 / 15$ or 0.40.
- ❖ In general, it is nice to identify rules that have a high support, as these will be applicable to a large number of transactions.
- ❖ Support is an important measure because a rule that has a low support may occur simply by chance. A low support rule may also be uninteresting from a business perspective because it may not be profitable to promote items that are seldom bought together. For these reasons, support is often used to eliminate uninteresting rules.
- ❖ For super market retailers, this is likely to involve basic products that are popular across an entire user base (e.g. bread, milk). A printer cartridge retailer, for example, may not have products with a high support, because each customer only buys cartridges that are specific to his / her own printer.

Confidence

- ❖ The confidence of a rule is the likelihood that it is true for a new transaction that contains the items on the LHS of the rule. (I.e. it is the probability that the transaction also contains the item(s) on the RHS.) Formally:
- ❖ $\text{Confidence}(X \rightarrow Y) = \text{Support}(X \cup Y) / \text{Support}(X)$
- ❖ $\text{Confidence}(\text{Peanut Butter} \Rightarrow \text{Jelly}) = 0.26 / 0.40 = 0.65$
- ❖ This means that for 65% of the transactions that contain Peanut Butter and Jelly, the rule is correct.
- ❖ Confidence measures the reliability of the inference made by a given rule. For a given rule $X \rightarrow Y$, the higher the confidence the more likely it is for Y to be present in transactions that contain X.
- ❖ Association analysis results should be interpreted with caution. The inference made by an association rule does not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between the items in the antecedent and consequent of the rule.

Lift

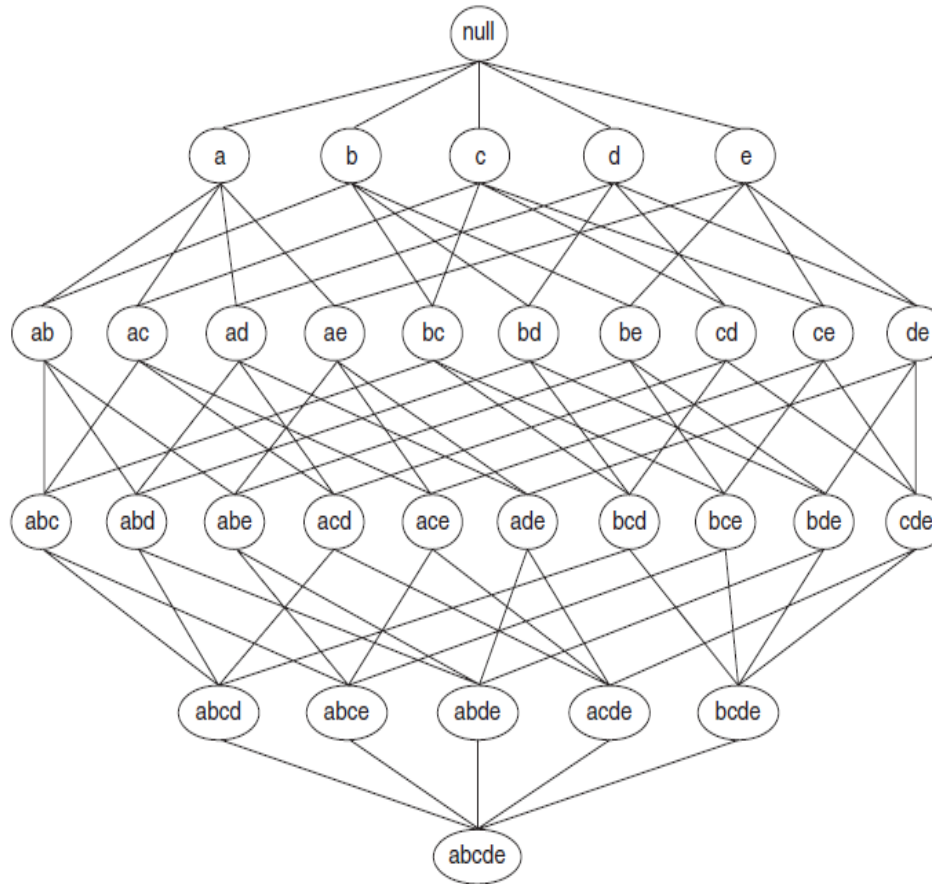
- ❖ The lift of a rule is the ratio of the support of the items on the LHS of the rule co-occurring with items on the RHS divided by probability that the LHS and RHS co-occur if the two are independent.
- ❖ $\text{Lift}(X \rightarrow Y) = \text{Support}(X \cup Y) / (\text{Support}(Y) * \text{Support}(X))$
- ❖ $\text{Lift}(\text{Peanut Butter} \Rightarrow \text{Jelly}) = 0.26 / (0.46 * 0.40) = 1.4$
- ❖ If lift is greater than 1, it suggests that the presence of the items on the LHS has increased the probability that the items on the right hand side will occur on this transaction.
- ❖ If the lift is below 1, it suggests that the presence of the items on the LHS make the probability that the items on the RHS will be part of the transaction lower.
- ❖ If the lift, is 1 it indicates that the items on the left and right are independent.
- ❖ When we perform market basket analysis, then, we are looking for rules with a lift of more than one and preferably with a higher level of support.

Apriori Algorithm

- ❖ The Apriori algorithm is perhaps the best known algorithm to mine association rules.
- ❖ Apriori Theorem: *"If an itemset is frequent, then all of its subsets must be also be frequent."*
- ❖ It uses a breadth first strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support (anti-monotonicity).
- ❖ The approach follows a 2 step process:
 - ❖ First, minimum support is applied to find all frequent itemsets in the database.
 - ❖ Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

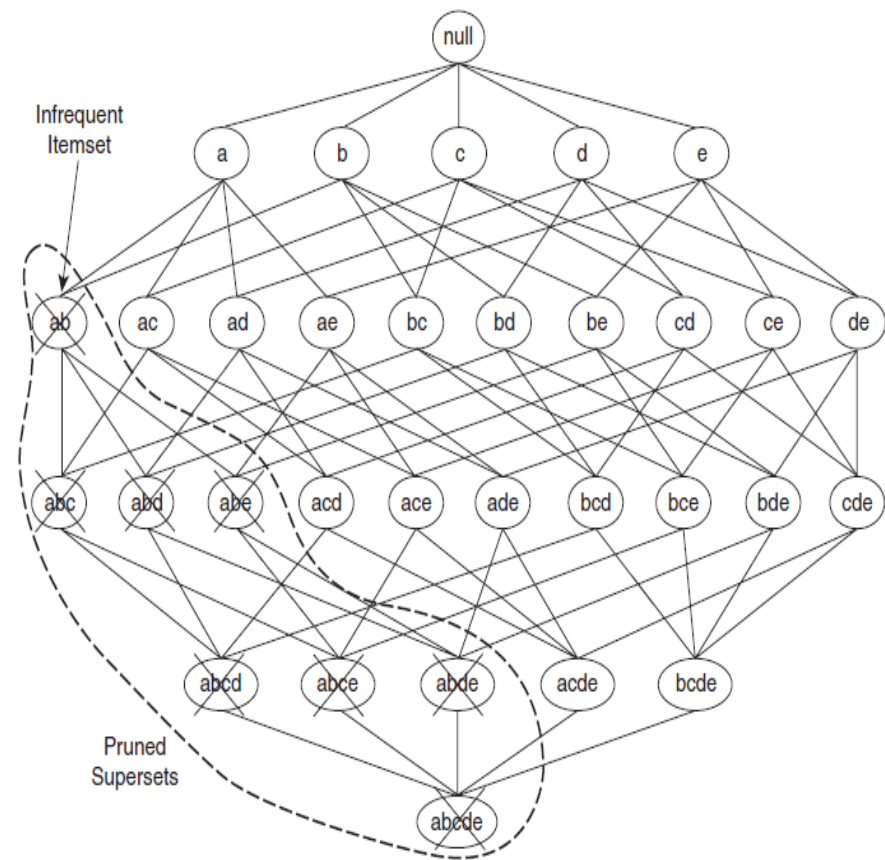
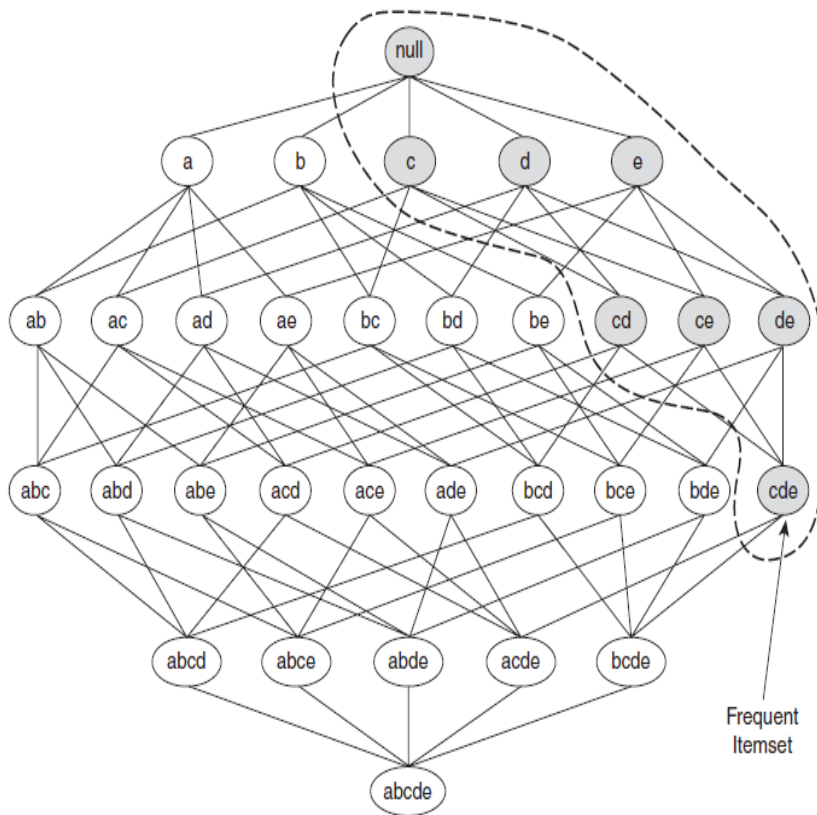
Frequent Itemset

- ❖ Finding all frequent itemsets in a database is difficult since it involves searching for all item combinations. The set of possible item combinations is the power set over I and has the size of $2^n - 1$.



Frequent Itemset

- ❖ The downward-closure property of support allows for efficient search and guarantees that for a frequent itemset, all of its subsets are also frequent. Additionally, for an infrequent itemset, all of its supersets must also be infrequent.



Practical Example - Groceries



Market Basket Analysis

- ❖ Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer's basket – and therefore 'Market Basket Analysis'.
- ❖ That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item.
- ❖ For each transaction, there can be only distinct Item(s) without repeating entries. This allows for us to create a binary (0,1) representation whether a particular item was purchased under a specific transaction.

Transaction	Items
A	X
A	Y
B	X
B	Z
C	Y
C	Z
etc...	etc...

Transaction	Items
A0001	citrus fruit
A0001	margarine
A0001	ready soups
A0001	semi-finished bread
A0002	coffee
A0002	tropical fruit

Association Rules – Binary Representation

- ❖ The dataset will need to first be flipped across the horizontal axis in to a cross tabulation. Notice that the “Items” are now the column headings. This preparation ensures that the dataset can be read properly into the apriori market basket algorithm.

Transaction	Items
A	X
A	Y
B	X
B	Z
C	Y
C	Z
etc...	etc...



Transactions	<u>X</u>	<u>Y</u>	<u>Z</u>	<u>etc...</u>
<u>A</u>	1	1	0	etc...
<u>B</u>	1	0	1	etc...
<u>C</u>	0	1	1	etc...
<u>etc...</u>	etc...	etc...	etc...	etc...

Association Rules – Binary Representation

Transaction	Items
A0001	citrus fruit
A0001	margarine
A0001	ready soups
A0001	semi-finished bread
A0002	coffee
A0002	tropical fruit



Transaction	citrus fruit	margarine	ready soups	semi finished bread	coffee	tropical fruit
A0001	1	1	1	1	0	0
A0002	0	0	0	0	1	1
A0003	0	0	0	0	0	0
A0004	0	0	0	0	0	0

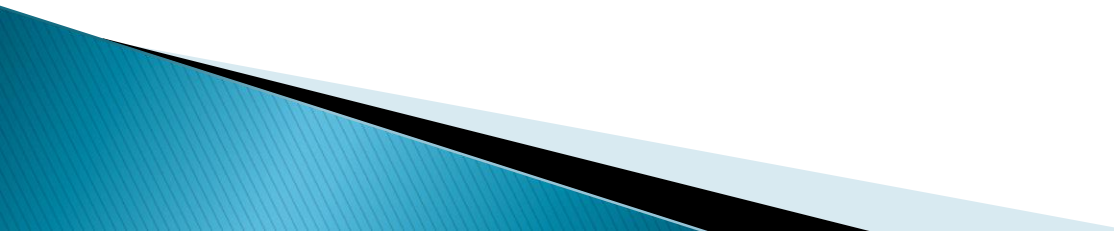
Practical Example - Groceries

- ❖ The market basket analysis algorithm requires setting a threshold for detecting patterns in the dataset.

Rules	Support	Confidence	Lift
{liquor, red/blush wine} => {bottled beer}	0.002	0.90	11.24
{cereals, yogurt} => {whole milk}	0.002	0.81	3.17
{butter, jam} => {whole milk}	0.001	0.83	3.26
{chocolate, pickled vegetables} => {whole milk}	0.001	0.86	3.35
{grapes, onions} => {other vegetables}	0.001	0.92	4.74
{hard cheese, oil} => {other vegetables}	0.001	0.92	4.74

- ❖ For this example, we specified a support value of 0.001 (due to the high volume of receipts and large product offering) and a confidence level of 0.70.
- ❖ Additionally, we set the length of the rule not to exceed three elements. This ensures that we will have a maximum of 2 items on the LHS and that the assessment will produce more meaningful insights at a tertiary glance.

Tuning the Algorithm Parameters

- ❖ Different businesses will have distribution of items by transaction that look very different, and so very different support and confidence parameters may be applicable.
 - ❖ To determine what works best, organizations need to experiment with different parameters: as you reduce them, the number of rules generated will increase, which will give us more to work with.
 - ❖ However, we will need to sift through the rules more carefully to identify those that will be more impactful for your business.
 - ❖ There is no steady fast rule on where to begin so experiment with loose parameters and go from there.
- 

Practical Example - Groceries

- ❖ To showcase how we can leverage this insight further let's focus in on 3 specific items of interest:
 - ❖ Yogurt
 - ❖ Tropical Fruit
 - ❖ Bottled Beer.
- ❖ We can run the algorithm (with the same thresholds) and specify that these terms are used as criterion for the RHS of the ruleset generation.

Rules	Support	Confidence	Lift
{pip fruit, sausage, sliced cheese} => {yogurt}	0.001	0.86	6.14
{butter, cream cheese , root vegetables} => {yogurt}	0.001	0.91	6.52
{butter, margarine, tropical fruit} => {yogurt}	0.001	0.85	6.07
{butter, curd, other vegetables, tropical fruit} => {yogurt}	0.001	0.83	5.97
{liquor, red/blush wine} => {bottled beer}	0.002	0.90	11.24
{citrus fruit, fruit/vegetable juice, grapes} => {tropical fruit}	0.001	0.85	8.06
{ham, other vegetables, pip fruit, yogurt} => {tropical fruit}	0.001	0.83	7.94

Practical Example - Groceries

❖ Now as criterion for the LHS of the ruleset generation.

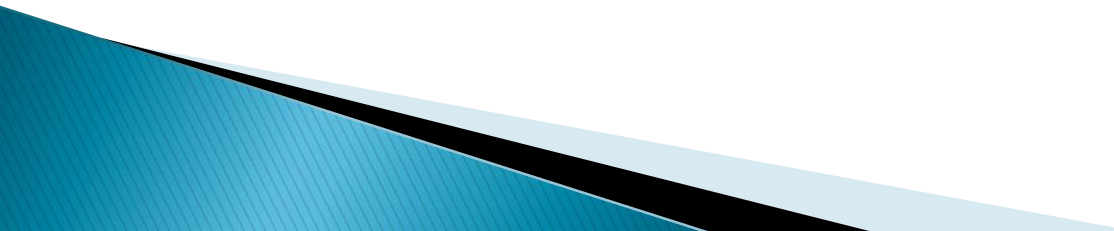
Rules	Support	Confidence	Lift
{yogurt} => {whole milk}	0.056	0.40	1.57
{tropical fruit} => {other vegetables}	0.036	0.34	1.77
{yogurt} => {rolls/buns}	0.034	0.25	1.34
{tropical fruit} => {rolls/buns}	0.025	0.23	1.27
{bottled beer} => {soda}	0.017	0.21	1.21
{bottled beer} => {bottled water}	0.016	0.20	1.77
{tropical fruit} => {pip fruit}	0.020	0.19	2.57
{tropical fruit} => {citrus fruit}	0.020	0.19	2.29

Using the Analysis for Business Decisions



- ❖ Before we use the data to make any kind of business decision, it is important that we take a step back and remember something important:
- ❖ The output of the analysis reflects how frequently items co-occur in transactions. This is a function both of the strength of association between the items, and the way the business has presented them to the customer.
- ❖ To say that in a different way: items might co-occur not because they are “naturally” connected, but because we, the people in charge of the organization, have presented them together.

Targeted Marketing

- ❖ The market basket results can be used to drive targeted marketing campaigns.
 - ❖ For each patron, we pick a handful of products based on products they have bought to date which have both a high uplift and a high margin, and send them a e.g. personalized email or display ads etc.
 - ❖ How we use the analysis has significant implications for the analysis itself: if we are feeding the analysis into a machine-driven process for delivering recommendations, we are much more interested in generating an expansive set of rules.
 - ❖ If, however, we are experimenting with targeted marketing for the first time, it makes much more sense to pick a handful of particularly high value rules, and action just them, before working out whether to invest in the effort of building out that capability to manage a much wider and more complicated rule set.
- 

Web Based Marketing

- ❖ There are a number of ways we can use the data to drive site organization:
 - ❖ Large clusters of co-occurring items should probably be placed in their own category / theme.
 - ❖ Item pairs that commonly co-occur should be placed close together within broader categories on the website. This is especially important where one item in a pair is very popular, and the other item is very high margin.
 - ❖ Long lists of rules (including ones with low support and confidence) can be used to put recommendations at the bottom of product pages and on product cart pages. The only thing that matters for these rules is that the lift is greater than one. (And that we pick those rules that are applicable for each product with the high lift where the product recommended has a high margin.)
 - ❖ In the event that doing the above (3) drives significant uplift in profit, it would strengthen the case to invest in a recommendation system, that uses a similar algorithm in an operational context to power automatic recommendation engine on your website.

Practical Example - Politics



Example: Congressional Voting Records

- ❖ We will apply the results of association analysis to the voting records of members of the United States House of Representatives. The data is obtained from the 1984 Congressional Voting Records Database, which is available at the UCI machine learning data repository.
- ❖ Each transaction contains information about party affiliation along with his or her voting record on 16 key issues.

Attribute Information:	
1. Class Name:	2 (democrat, republican)
2. handicapped-infants:	2 (y,n)
3. water-project-cost-sharing:	2 (y,n)
4. adoption-of-the-budget-resolution:	2 (y,n)
5. physician-fee-freeze:	2 (y,n)
6. el-salvador-aid:	2 (y,n)
7. religious-groups-in-schools:	2 (y,n)
8. anti-satellite-test-ban:	2 (y,n)
9. aid-to-nicaraguan-contras:	2 (y,n)
10. mx-missile:	2 (y,n)
11. immigration:	2 (y,n)
12. synfuels-corporation-cutback:	2 (y,n)
13. education-spending:	2 (y,n)
14. superfund-right-to-sue:	2 (y,n)
15. crime:	2 (y,n)
16. duty-free-exports:	2 (y,n)
17. export-administration-act-south-africa:	2 (y,n)

Practical Example - Politics

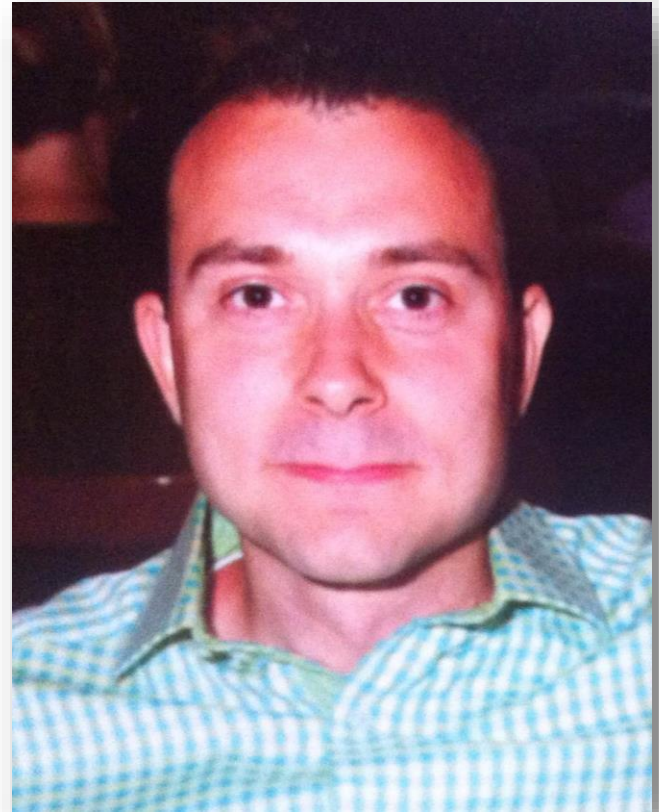
Rules	Support	Confidence	Lift
{physician fee freeze = No} => {Democrat}	0.51	0.99	1.86
{physician fee freeze = Yes, mx missile = No} => {Republican}	0.40	0.96	2.06
{physician fee freeze = Yes} => {Republican}	0.46	0.95	2.03
{adoption of the budget resolution = Yes, education spending = No} => {Democrat}	0.43	0.94	1.76
{education spending = No} => {Democrat}	0.47	0.87	1.63
{mx missile = Yes} => {Democrat}	0.42	0.87	1.62
{crime = Yes, duty free exports = No} => {Republican}	0.41	0.80	1.71
{mx missile = No} => {Republican}	0.40	0.78	1.68

Observations:

- ❖ A vote in favor of the Physician Pay Freeze indicates a Republican (95% confidence), a vote against indicates a Democrat (99% confidence).
- ❖ The voting patterns are relatively clear in this example with a high degree of support and confidence. We can see which party favors a specific law without knowing the contents of the legislation itself.

About Me

- ❖ Reside in Wayne, Illinois
- ❖ Active Semi-Professional Classical Musician (Bassoon).
- ❖ Married my wife on 10/10/10 and been together for 10 years.
- ❖ Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- ❖ Pet Maine Coons' named Maximus Power and Nemesis Gul du Cat.
- ❖ Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- ❖ Self proclaimed Data Nerd and Technology Lover.



Fine