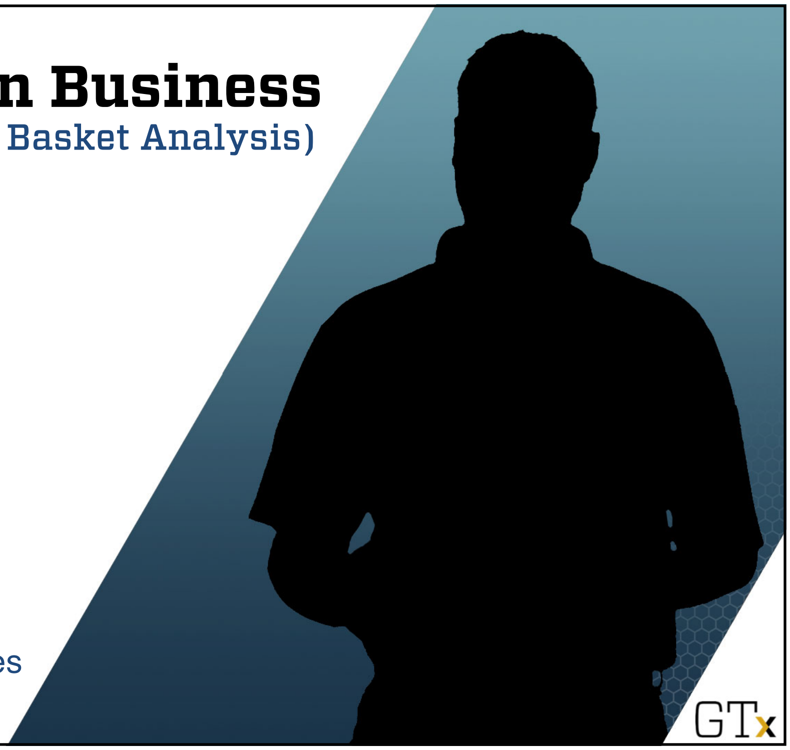
Data Analytics in Business

Association Rules (Market Basket Analysis)



Sridhar Narasimhan

Professor

Scheller College of Business

Data Mining and Association Rules

GTx

Lessons

A. Introduction to Association Rules (Market Basket Analysis)
B. Support and Confidence
C. Lift
D. Grocery Dataset Example Using R

GTx

Unsupervised Data Mining

- Analysts do not create the model before running the analysis
- Apply data-mining technique and observe the results
- Hypotheses are created after the analysis as explanation for the results
- Example:
    - Cluster analysis
    - Association Detection

GTx

Supervised Data Mining

- Model developed before the analysis
- Statistical techniques are used to estimate the parameters
- Examples:
    - Classification
    - Regression analysis

GTx

What Types of Data are Analyzed in Association Rules Mining?

- Records of customer relationships.
- Could be captured in Customer Relationship Management (CRM) systems
- Customer visits to websites
- Customers use of apps on mobile devices
- Call center records of customer support calls
- Sales receipts in stores
- These could be helpful to marketing executives who are trying to understand customer purchasing patterns
- Association rules mining can help recommend movies or music

Properties of Data Used in Associative Rules Mining

- Could be potentially huge data sets

Consider a typical supermarket

- It has several thousands of items for sale, millions of customers
- It has several thousands of daily transactions
- Each transaction (checkout by a customer) only involves a few items
- Each item is not purchased with nearly every other item!
- So the matrix of items and transactions is very sparse
- Correlations and regression techniques are not applicable since most of the matrix elements are empty!
- Hence, we need to use data mining techniques to tease out valuable insights that can be actionable by managers

Association, Transactions (Market Basket)

- Association rule mining – if some events (say items purchased together) occur together more often than is warranted by their individual rates of occurrence then we have a pattern worth exploring
- Assume beer occurs in 10% of sales transactions and peanuts in 5% of sales transactions in a supermarket.
- If in the transactions with beer sales, peanuts occurred 5% of the time, then we probably have no interesting relationship.
- If, however, in the transactions with beer sales, peanuts occurred 30% of the time, then we have an association!
- A **transaction** or **market basket** is the set of “things” that are purchased together or the set of “things” viewed by a user in a session.

Examples

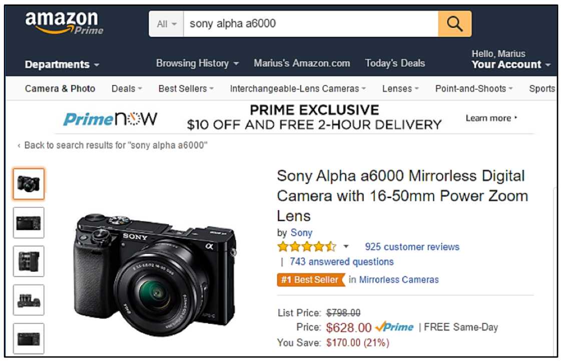
- Part of Amazon recommendations
  - What products are bought together? E.g., printers and paper, digital cameras and SD cards
  - Implication: what recommendations to make.
- Healthcare symptoms and disease
  - Symptoms and illnesses that manifest together, harmful drug interactions
  - E.g., “We know that a high percentage of patients with condition A also develop condition B, so we’re also going to start you on a drug regimen for condition B as a precaution.”

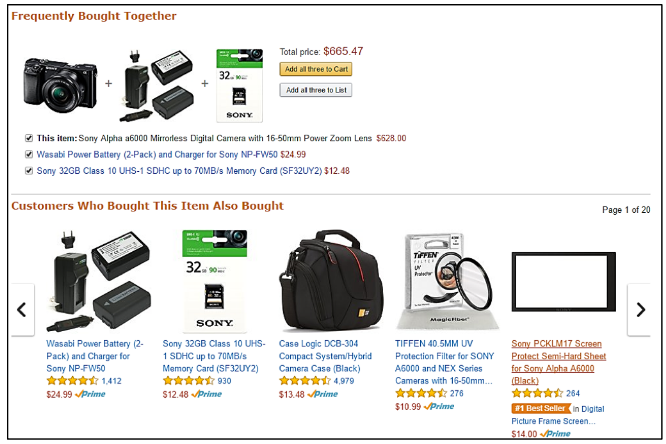
amazon.com

FDA Data Mining Uncovers Harmful Drug Interaction

Researchers mine FDA's Adverse Event Reporting System and corroborate findings using patient records to identify diabetes threat

Examples





Association Rules Mining: Market Basket Analysis

- **Market basket analysis:**
  - Type of association rules mining determines what products go together in a shopping cart at a retailer
  - Data-mining technique for determining sales patterns
  - Shows products that customers tend to buy **together**.
  - Example: On Thursday nights, people who buy beer may also buy peanuts.

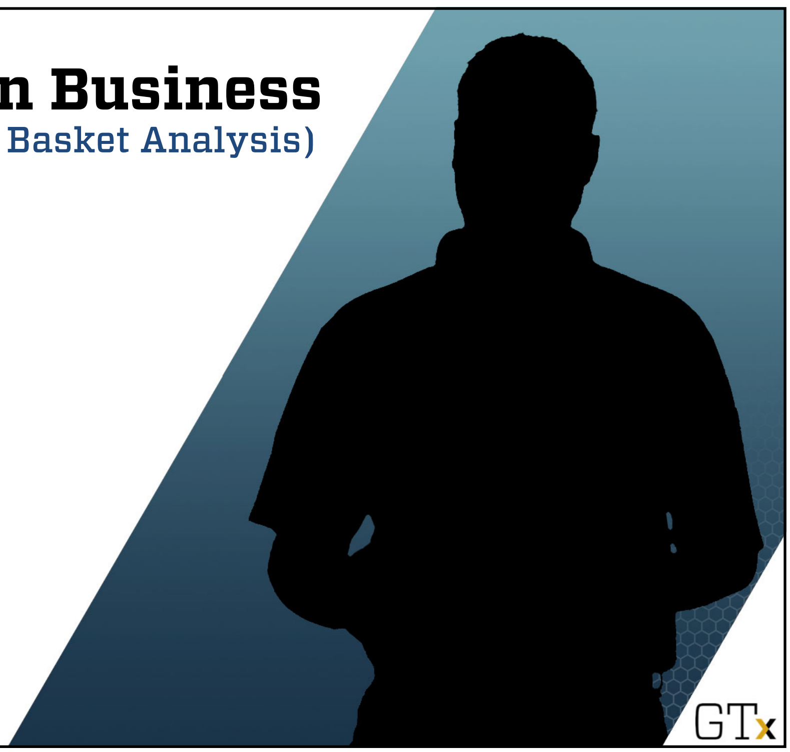
Market Basket Analysis

Product affinities (likelihood of two or more products being sold together) → cross-selling opportunities

Product	Association	Lift	Confidence
Orbit Sleeping Pad	Orbit Stuff Sack	222	37%
Bambini Tights Children's	Bambini Crewneck Sweater Children's	195	52%
Silk Crew Women's	Silk Long Johns Women's	304	73%
Cascade Entrant Overmitts	Polartec 300 Double Mitts	51	48%

Data Analytics in Business

Association Rules (Market Basket Analysis)



Sridhar Narasimhan
Professor
Scheller College of Business

Support, Confidence

What Does a Rule Look Like?

LHS → RHS

- The left hand side (LHS) contains a set of one or more 'things,' which are called **itemsets**.
- The right hand side contains a set of one or more 'things' or itemsets.

Some Examples

- {Cereal} → {Milk}
- {Pizza, FridayEvening} → {Beer}
- {Holiday, Turkey} → {CranberrySauce}

How Do We Evaluate Association Rules LHS → RHS?

- **Support of the rule:** how often do these things appear together?
- **Confidence**: given LHS, how often do we see RHS?
- **Lift**: how often does LHS appear with RHS, compared to what chance would predict?

Rule Strength: Support

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

- Consider the set of 5 transactions (market baskets) shown above
- **Count(n)** is the raw count of the number of transactions containing a particular {itemset}
  - n\{Milk, Bananas, Coke\} = 2 (baskets 3,5)
  - n\{Milk, Bananas\} = 3 (baskets 3,4,5)
  - n\{Coke\} = 2 (baskets 3,5)

GTx

Rule Strength: Support

- Support(s) is an empirical probability of the set of items (i.e., itemsets) in an association rule LHS → RHS
- *s(X)* is the fraction of transactions that contain all the items in an itemset. Note that the transaction could contain additional items!

$$s(X) = \frac{n(X)}{N}$$

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

n\{Milk, Bananas, Coke\} = 2 (baskets 3,5)
n\{Milk, Bananas\} = 3 (baskets 3,4,5)
n(\{Coke\}) = 2 (baskets 3,5)

N = Total # of baskets = 5

GTx

Rule Strength: Support

For example:

s(Coke) = $\frac{n(\{\text{Coke}\})}{5}$ = 2/5 = 0.4

s(Milk, Bananas) = $\frac{n(\{\text{Milk, Bananas}\})}{5}$ = 3/5 = 0.6

s(Milk, Bananas, Coke) = $\frac{n(\{\text{Milk, Bananas, Coke}\})}{5}$
= 2/5 = 0.4, which is the support for
Milk, Bananas → Coke or
Milk, Coke → Bananas, or
Coke, Bananas → Milk
because these three association rules have identical itemsets

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

n\{Milk, Bananas, Coke\} = 2 (baskets 3,5)
n\{Milk, Bananas\} = 3 (baskets 3,4,5)
n(\{Coke\}) = 2 (baskets 3,5)

N = Total # of baskets = 5

GTx

Rule Strength: Confidence

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

- **Confidence** is the strength of an association rule
- Confidence measures how often items in Y appear within those transactions that contain X – similar to an empirical conditional probability of $\boldsymbol{Y}$ given $\boldsymbol{X}$ - $P(Y|X)$
- $c(X \rightarrow Y) = \frac{s(X \rightarrow Y)}{s(X)} = \frac{s(X\&Y)}{s(X)} = \frac{n(X\&Y)}{n(X)}$
- Note that confidence is not **symmetric**. $c(X \rightarrow Y)$ is not necessarily equal to $c(Y \rightarrow X)$

GTx

Rule Strength: Confidence

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

- For example:

$$c(\{\text{Milk, Bananas}\} \rightarrow \{\text{Coke}\}) = \frac{n(\{\text{Milk, Bananas, Coke}\})}{n(\{\text{Milk, Bananas}\})} = \frac{2}{3} = 0.67$$

$$c(\{\text{Coke}\} \rightarrow \{\text{Milk, Bananas}\}) = \frac{n(\{\text{Milk, Bananas, Coke}\})}{n(\{\text{Coke}\})} = \frac{2}{2} = 1$$

Rules Having the Same Itemset

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

All of these rules below are binary partitions of the same itemset: {Milk, Bananas, Beer}

n{Milk, Bananas, Beer} = 2 (baskets 4,5)
n{Milk, Bananas} = 3 (baskets 3,4,5)
n{Milk, Beer} = 2 (baskets 3,4)
n{Bananas, Beer} = 3 (baskets 2,3,4)
n{Beer} = 3 (baskets 2,3,4)
n{Bananas} = 4 (baskets 2,3,4,5)
n{Milk} = 4 (baskets 1,3,4,5)

Association Rule	Support (s)	Confidence (c)
{Milk, Bananas} → {Beer}	2/5 = 0.4	2/3 = 0.67
{Milk, Beer} → {Bananas}	2/5 = 0.4	2/2 = 1.0
{Bananas, Beer} → {Milk}	2/5 = 0.4	2/3 = 0.67
{Beer} → {Milk, Bananas}	2/5 = 0.4	2/3 = 0.67
{Bananas} → {Milk, Beer}	2/5 = 0.4	2/4 = 0.5
{Milk} → {Bananas, Beer}	2/5 = 0.4	2/4 = 0.5

Venn Diagram

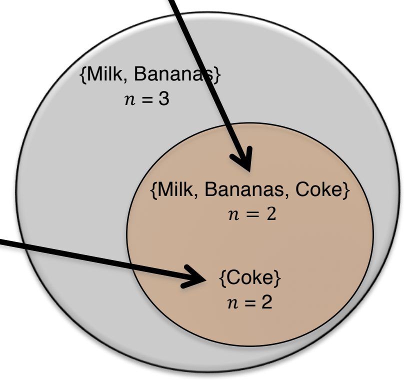
Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

$$c(\{\text{Milk, Bananas}\} \rightarrow \{\text{Coke}\}) = \frac{n(\{\text{Milk, Bananas, Coke}\})}{n(\{\text{Milk, Bananas}\})} = \frac{2}{3} = 0.67$$

67% of total that buy Milk and Bananas also buy Coke

$$c(\{\text{Coke}\} \rightarrow \{\text{Milk, Bananas}\}) = \frac{n(\{\text{Milk, Bananas, Coke}\})}{n(\{\text{Coke}\})} = \frac{2}{2} = 1$$

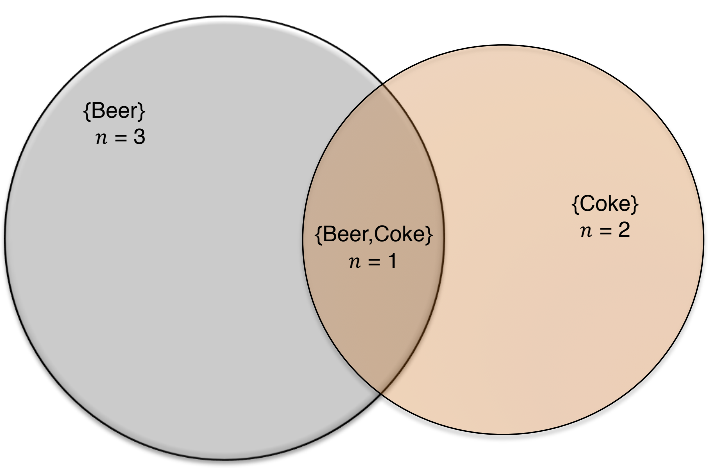
100% of total that buy Coke also buy Milk and Bananas



Venn Diagram

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

$$c(\{\text{Beer}\} \rightarrow \{\text{Coke}\}) = \frac{1}{3} = 0.33$$

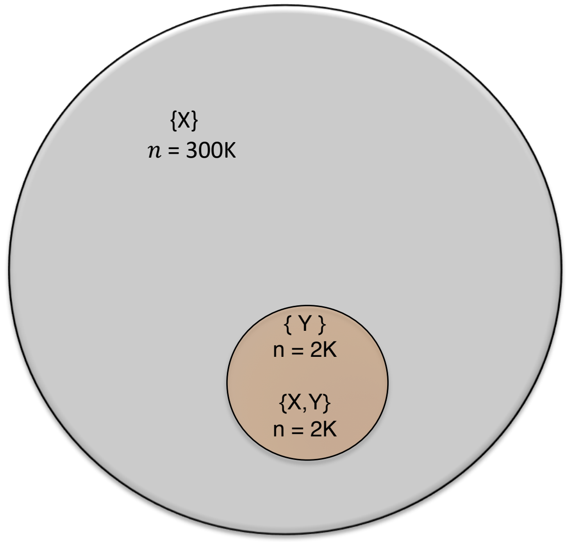


$$c(\{\text{Coke}\} \rightarrow \{\text{Beer}\}) = \frac{1}{2} = 0.5$$

Scenario 1: Product Recommendations After a Revealed Preference

<u>Confidence – direction matters</u>
Suppose you have the following scenario with repeat customers:

- 300K customers buy X
- 2K customers buy Y (and also happen to buy X)



GTx

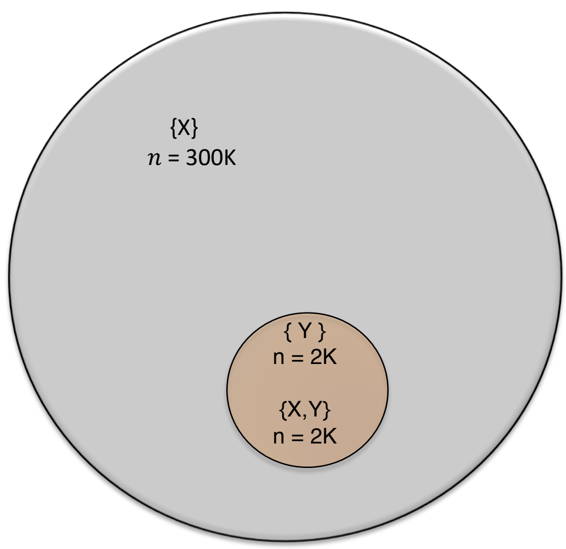
Scenario 1: Product Recommendations After a Revealed Preference

Would you rather

… **promote Y** to all customers interested in X (300K marketing effort units - when you know it might not matter in 298K cases)?

or

… **promote X** to all customers interested in Y (2K marketing effort units - it might get us 2K sales of X, for the other 298K it might not matter anyway)?

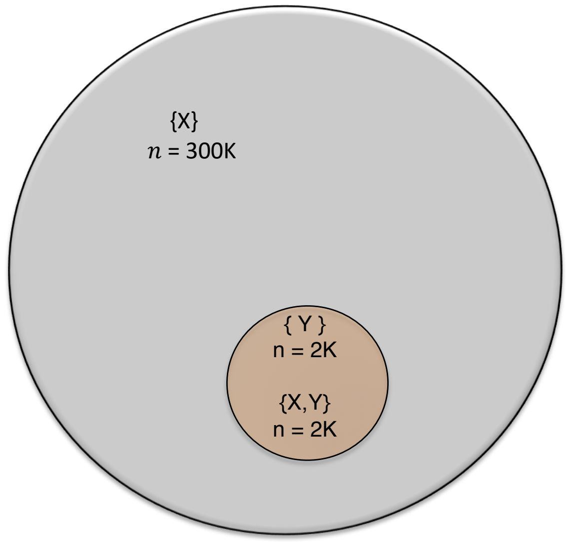


GTx

Scenario 2: Product Relocation in Stores

Confidence – direction matters
Suppose you have the following scenario with repeat customers:

- 300K customers buy X
- 2K customers buy Y (and also happen to buy X)



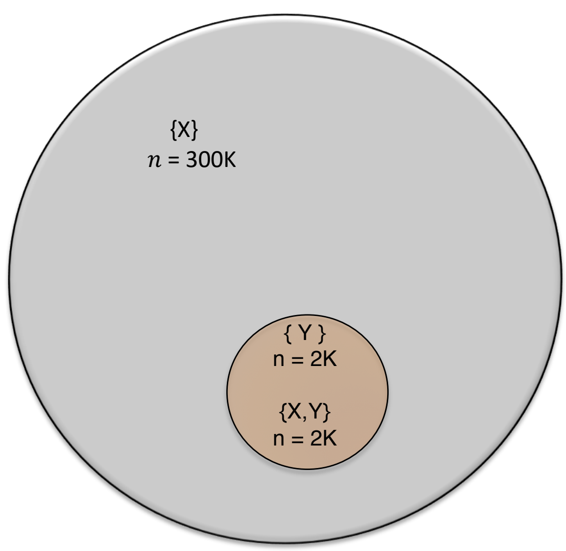
Scenario 2: Product Relocation in Stores

Would you rather

… **relocate X** close to Y in the store (and risk perhaps 298K customers - who knew where X was before and wanted X but not Y - not finding X)?

or

… **relocate Y** close to X in the store (the 2K customers interested in Y were interested in X anyway so they know where X is and they will find Y in the new location)?



Quiz (True/False)

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

- The support for {Beer} → {Milk, Bananas} = 2/5 = 0.4
  Answer: **TRUE**.

- The confidence for {Bananas} → {Milk, Beer} = 2/5 = 0.4
  Answer: **FALSE**. The confidence = 2/4 = 0.5

Data Analytics in Business

Association Rules (Market Basket Analysis)

Sridhar Narasimhan
Professor
Scheller College of Business

Lift

Why Do We Need Lift?

Confidence and support can be misleading

Example: {Bread} → {Bananas}

- Support = 3/5 = 60%
- Confidence = 3/4 = 75%
- However, Bread and Bananas both occur frequently in general – high confidence may still just be due to random co-occurrence chance

Basket	Items
1	Bread, Milk
2	Bread, Bananas, Beer, Eggs
3	Milk, Bananas, Beer, Coke
4	Bread, Milk, Bananas, Beer
5	Bread, Milk, Bananas, Coke

We need a measure that can account for **chance**.

GTx

The Lift Metric

- **Lift** tries to factor in the expected probability of co-occurrence based only on chance. It tries to see if events are truly independent from one another.
- If A and B are independent, *P(A&B) = P(A) * P(B)*
- Lift of an Association Rule is the ratio of observed support to the expected support under independence

$$L(X \to Y) = \frac{s(X \to Y)}{s(X) \times s(Y)} = \frac{c(X \to Y)}{s(Y)}$$

- Lift is symmetric. That is $L(X \to Y) = L(Y \to X)$

GTx

Lift

- $L(X \rightarrow Y) = \frac{s(X \rightarrow Y)}{s(X) \times s(Y)} = \frac{s(X \& Y)}{s(X) \times s(Y)} = \frac{c(X \rightarrow Y)}{s(Y)} \approx \frac{P(Y|X)}{P(Y)}$
- This metric is called lift because it calculates how much the presence of X "lifts" the probability of Y.
- For example, suppose the rule {Cereal} → {Milk} has lift of 4. This can be interpreted as: "shoppers who buy cereal are 4x (or 400%) more likely to buy milk than the average shopper."

GTx

Lift Interpretation

Lift > 1 indicates positive correlation – co-occurrence is more likely than chance.

Lift ≈1 indicates almost no correlation – the events seem independent.

Lift < 1 indicates negative correlation – co-occurrence is less likely than chance. In other words, actions occur in opposite directions – customers who buy one item (or a given itemset) are less likely (than average) to purchase the other.

GTx

Lift

Example:

- One would expect
  - **Lift > 1** for complementary products that often times are bought together (e.g. toothbrush and toothpaste, toothpaste and floss)
  - **Lift < 1** for substitutes (e.g., wines of the same kind – Pinot brand X and Pinot brand Y – customers who bought one brand might be less likely to buy the other brand of the same kind of wine)
- However, there may be more subtle connections (affinities) between item sets which may lead to high lift values (unexpected opportunities for cross-selling). It is important to identify those.

Example for Calculating Lift

Consider the following data describing what financial instruments consumers have (1000 consumers).

		Credit Card (CC)		
		No	**Yes**	**Total**
Savings Account (SA)	**No**	50	350	**400**
	Yes	100	500	**600**
	Total	**150**	**850**	**1000**

Consider the association rule $SA \rightarrow CC$

Compute:

$s(SA) =$

$s(CC) =$

$s(SA \rightarrow CC) =$

$c(SA \rightarrow CC) =$

$L(SA \rightarrow CC) =$

Example for Calculating Lift

Consider the following data describing what financial instruments consumers have (1000 consumers).

		Credit Card (CC)		
		No	Yes	Total
Savings Account (SA)	No	50	350	**400**
	Yes	100	500	**600**
	Total	**150**	**850**	**1000**

Consider the association rule $SA \rightarrow CC$

Compute:

$s(SA) =$ **600/1000 = 0.60**

$s(CC) =$ **850/1000 = 0.85**

$s(SA \rightarrow CC) =$ **500/1000 = 0.50**

$c(SA \rightarrow CC) =$ **500/600 = 0.83**

$L(SA \rightarrow CC) =$ **0.83/0.85 = 0.98**

Another Example for Lift Calculation

$$L(Beer \rightarrow Hot\ Dog) = \frac{s(Beer \rightarrow Hot\ Dog)}{s(Beer) \times s(Hot\ Dog)}$$

$$= \frac{400/1000}{{}^{500}/_{1000} \times {}^{500}/_{1000}} = 0.4/.25 = 1.6$$

$$L(Wine \rightarrow Hot\ Dog) = \frac{s(Wine \rightarrow Hot\ Dog)}{s(Wine) \times s(Hot\ Dog)}$$

$$= \frac{100/1000}{{}^{500}/_{1000} \times {}^{500}/_{1000}} = 0.1/.25 = 0.4$$

		Food		
		Fruit	Hot Dog	Total
Wine/Beer	Beer	100	400	**500**
	Wine	400	100	**500**
	Total	**500**	**500**	**1000**

Association Rules

How many different association rules can we have in total?

- With ***k*** items, there are $3^k - 2^{k+1} + 1$ different association rules
- k = 5 → total 180; k = 20 → total 3,484,687,250
- The total number increases exponentially

We want to find meaningful and strong association rules X→Y

- Minimum support requirement
  - Supports need to be large enough to be statistically significant
- Minimum confidence requirement
  - Y must occur at a reasonable frequency among transactions that contain X
- Lift must be high
  - We want co-occurrence of events that happens significantly more often than chance

Quiz (True/False)

- Lift is asymmetric; that is $L(X \rightarrow Y) \neq L(Y \rightarrow X)$.
  Answer: **FALSE.** Lift is symmetric.

- The definition of lift is $L(X \rightarrow Y) = \frac{s(X \rightarrow Y)}{s(X) \times s(Y)} = \frac{c(X \rightarrow Y)}{s(Y)}$.
  Answer: **TRUE.**

Data Analytics in Business

Association Rules (Market Basket Analysis)

Sridhar Narasimhan
Professor
Scheller College of Business

Grocery Dataset Example Using R

Using the *arules* and *arulesviz* Packages in R

#load the packages for association rules and the groceries dataset

library(arules)
library(arulesViz)
data("Groceries") # built in dataset with the arules package. The Groceries data set contains 1 month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions and the items are aggregated to 169 categories.

summary(Groceries)

Counts for the top 5 most frequent items

```
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda
            2513             1903             1809             1715
          yogurt          (other)
            1372            34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
  17   18   19   20   21   22   23   24   26   27   28   29   32
  29   14   14    9   11    4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
            labels
1 abrasive cleaner
2 artif. sweetener
3   baby cosmetics
> |
```

Frequency of transactions of different size (e.g. 1005 transactions that contained four products)

GTx

inspect(head(Groceries,10))

items
[1] {citrus fruit,semi-finished bread,margarine,ready soups}
[2] {tropical fruit,yogurt,coffee}
[3] {whole milk}
[4] {pip fruit,yogurt,cream cheese ,meat spreads}
[5] {other vegetables,whole milk,condensed milk,long life bakery product}
[6] {whole milk,butter,yogurt,rice,abrasive cleaner}
[7] {rolls/buns}
[8] {other vegetables,UHT-milk,rolls/buns,bottled beer,liquor (appetizer)}
[9] {pot plants}
[10] {whole milk,cereals}

GTx

rules <- apriori(Groceries,parameter=list(supp = 0.001, conf=0.8))

- We use the **Apriori algorithm** to compute the association rules from the Groceries dataset of transactions
- So **rules** is an object that we have created using the Apriori algorithm
- This algorithm allows us to specify the **support** (**supp**) and **confidence** (**conf**) level of rules that we are interested in
- For example, specifications: **supp** = 0.001 and **conf**=0.8 in the function below mean that the rules that have **support** less than 0.001 and **confidence** less than 0.8 will be dropped
- We can also observe (in the next slide) that the absolute minimum support count was 9 items (the minimum number of times any product appeared in any transaction was 9) and also that 410 association rules were found

GTx

rules <- apriori(Groceries,parameter=list(supp = 0.001, conf=0.8))

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
        0.8    0.1    1 none FALSE            TRUE       5   0.001      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> |
```

GTx

Apriori Algorithm

Advantages of the Apriori Algorithm	Disadvantages of the Apriori Algorithm
• Computationally more efficient than brute force association algorithms. • Allows us to specify rules with predefined characteristics of interest.	• Eliminates product sets with low support at the early stage of computation. Thus, we might potentially lose some interesting observations that were not frequent. • Also produces association rules with a single item on the RHS.

GTx

Inspect the generated association rules

> options(digits=2)
> inspect(rules[1:5])

```
    lhs                         rhs            support confidence lift
[1] {liquor,red/blush wine}  => {bottled beer} 0.0019  0.90       11.2
[2] {cereals,curd}           => {whole milk}   0.0010  0.91        3.6
[3] {cereals,yogurt}         => {whole milk}   0.0017  0.81        3.2
[4] {butter,jam}             => {whole milk}   0.0010  0.83        3.3
[5] {bottled beer,soups}     => {whole milk}   0.0011  0.92        3.6
```

GTx

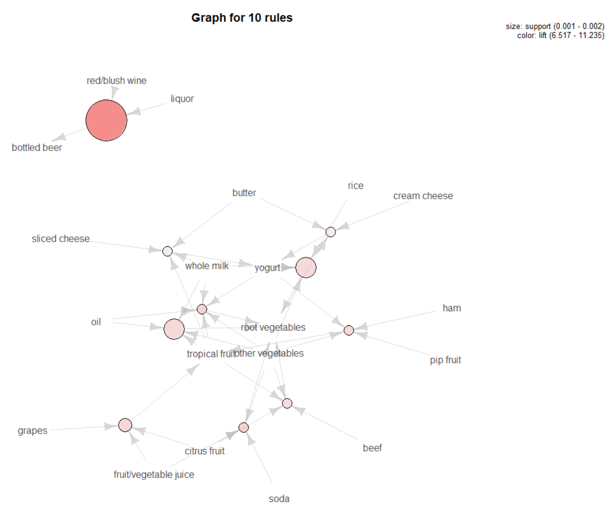
Summary(rules)

```
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 29 229 140  12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0     4.0     4.0     4.3     5.0     6.0

summary of quality measures:
    support           confidence        lift           count
 Min.   :0.00102   Min.   :0.80   Min.   : 3.1   Min.   :10.0
 1st Qu.:0.00102   1st Qu.:0.83   1st Qu.: 3.3   1st Qu.:10.0
 Median :0.00122   Median :0.85   Median : 3.6   Median :12.0
 Mean   :0.00125   Mean   :0.87   Mean   : 4.0   Mean   :12.3
 3rd Qu.:0.00132   3rd Qu.:0.91   3rd Qu.: 4.3   3rd Qu.:13.0
 Max.   :0.00315   Max.   :1.00   Max.   :11.2   Max.   :31.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.001        0.8
```

inspect(head(sort(rules, by ="lift"),3))

```
 lhs                                                           rhs                support confidence lift count
[1] {liquor,red/blush wine}                                    => {bottled beer}    0.0019  0.90       11.2 19
[2] {citrus fruit,other vegetables,soda,fruit/vegetable juice} => {root vegetables} 0.0010  0.91        8.3 10
[3] {tropical fruit,other vegetables,whole milk,yogurt,oil}    => {root vegetables} 0.0010  0.91        8.3 10
```

```
subrules2<-head(sort(rules,by="lift"),10)
plot(subrules2,method="graph")
```



Recap of the Lessons

A. Introduction to Association Rules (Market Basket Analysis)
B. Support and Confidence
C. Lift
D. Grocery Dataset Example Using R

Data Analytics in Business

Association Rules (Market Basket Analysis)

Sridhar Narasimhan
Professor
Scheller College of Business

End

GTx