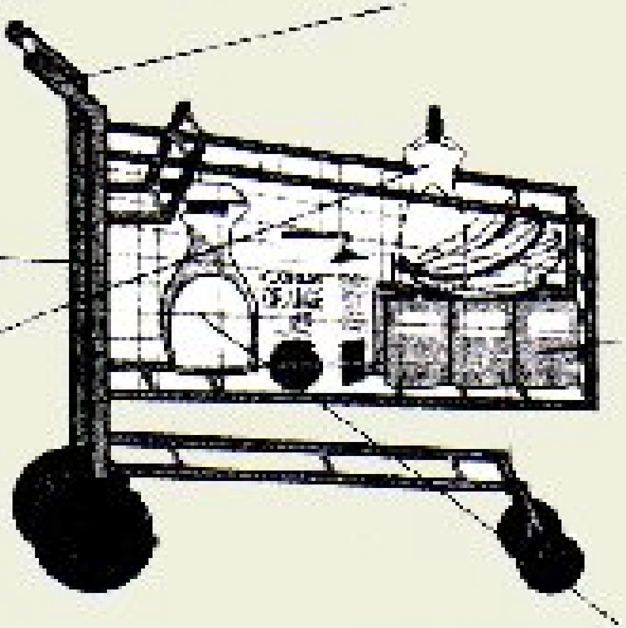
quart of orange juice, some bananas, dish detergent, window cleaner, and a six-pack of soda.



How are the demographics of the neighborhood affecting what customers are buying?

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

Where should detergents be placed in the store to maximize their sales?

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

On Analytical Tools for Market Basket (Associations) Analysis.

By Leonardo E. Auslender
SAS Institute, Research & Development
Leonardo dot Auslender “AT” sas dot com

Presented at NYC Informs,
New York City, NY, May 2004

Contents.

1. Definition and technicalities.
2. Business and other cases.
3. The practice of Market Basket Analysis.
   3.1 Barbie and Candy.
   3.2 Banking Application.
4. Adding Modeling to Association Effects.
   4.1 Giudici & Passerone.
   4.2 No Rose Garden.
   4.3 Other Methods.
5. Conclusion.
6. Bibliography.

1. Definitions and Technicalities.

Association Rules (Heuristic Definition and Application).

Identifies items/events that happen (or don’t) together:

retail: purchase of bread-butter, coffee-bagel.
clinical: fever-flu-cough.
Identifying patterns of telecommunications.

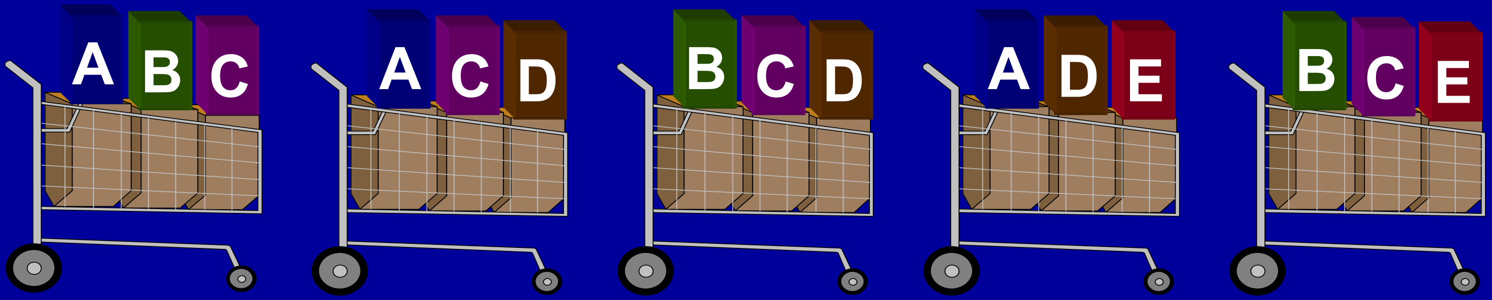
Attempts to provide interesting findings.

Identifies items that happen sequentially.
Household: home purchase- furniture.

Intends to categorize customer/criminal/patient…. behavior, by way of actionable profiles.

Intends to provide information on profitable strategies due to item/event identification, translated into catalog design, product placement and promotion, cross-selling, …

Association Rules (basket analysis, Agrawal et al 1993).



Rule	Support	Confidence
A ⇒ D	2/5	2/3 (P(A & D) /P(A)
C ⇒ A	2/5	2/4 (P(A & C) /P(C)
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3

A→B:

Support = Pr(A & B).

Confidence: if A, then B (A ‘causes’ B) = S (A →B) / S (A)

LIFT (A →B) = confidence / expected confidence if A and B are statistically independent: S(A→B) / S(A) S(B).

Association Rules: Implication?

Checking Account

		No	Yes	
Saving Account	No	500	3,500	4,000
	Yes	1,000	5,000	6,000
		1,500	8,500	10,000

S(CK) = 85%,
S(SVG) = 60%.
S(SVG →CK) = 50%
C(SVG → CK)= .5 / .6 = .83 (also 5000 / 6000)
Lift (SVG→CK) = .50 / (.85 * .6) = .83 / .85 = .98, CORR = -0.05
Lift(SVG→~CK) = .10 / (.15 * .6) = 1.11

Lift > 1 → "positively correlated" and vice-versa. **THESE ARE THE RULES TO LOOK AT.**

Association Rules. Present Practice.

Focus on transactions the lift of which is > 1, because it implies that items are positively correlated.

If lift (A→B) < 1 → lift (A→~B) (~B means 'not' B) > 1.

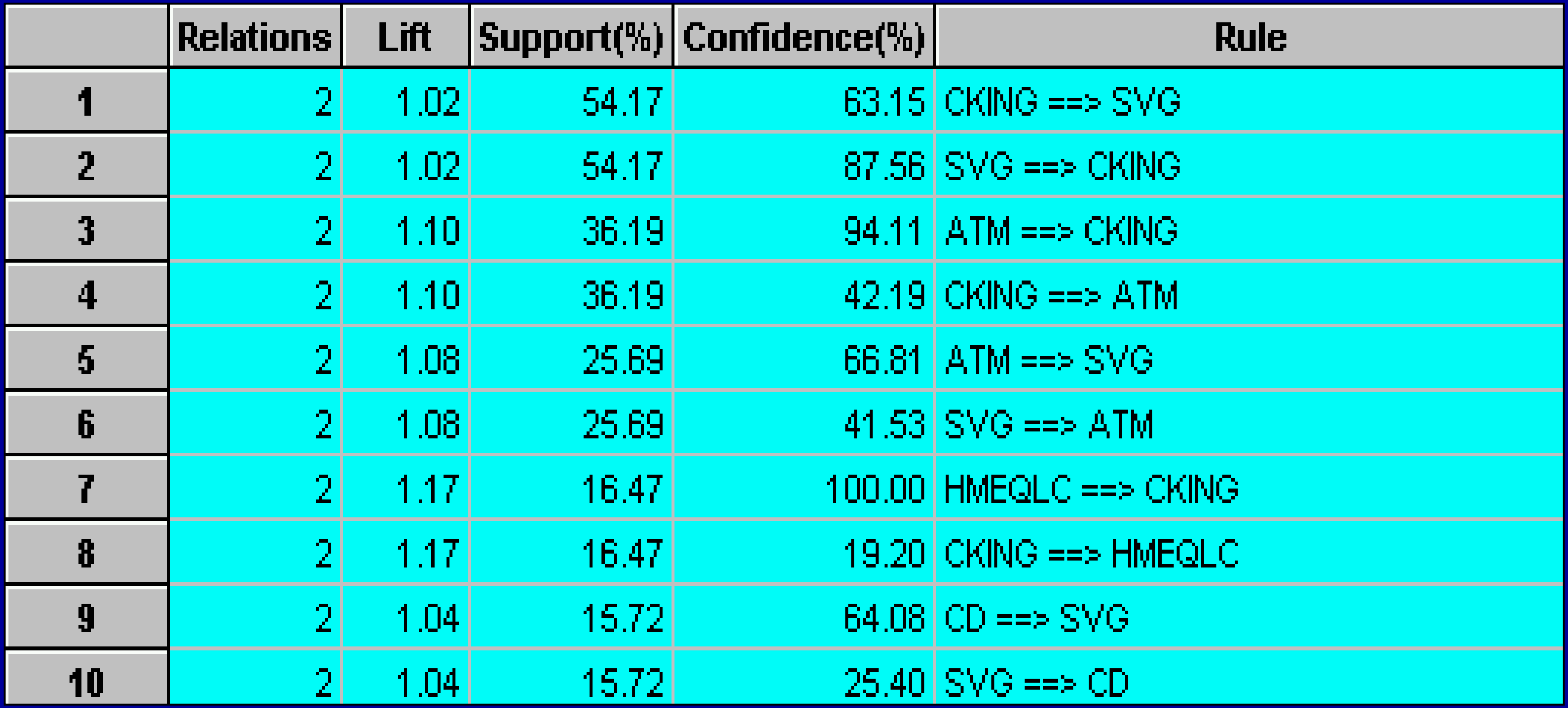
Market Basket Analysis typically "prunes" support at 5%. Tan et al (2002) show that this type of pruning tends to remove uncorrelated or negatively correlated rules. They also discuss extensively "measures of interestingness".

"Variable Selection" problem is not issue of bias/variance trade-off in AA, but of imputing measure of "importance-interestingness-relevance-substance…" to every rule to focus and lead the application.

Imputation might reduce number of items, but not necessarily.

Note: NO DISCUSSION ON COMPUTATIONAL ISSUES IN AA IN THIS PRESENTATION.

(SAS EM Association Node (clipped) set of rules below).



	Relations	Lift	Support(%)	Confidence(%)	Rule
1	2	1.02	54.17	63.15	CKING ==> SVG
2	2	1.02	54.17	87.56	SVG ==> CKING
3	2	1.10	36.19	94.11	ATM ==> CKING
4	2	1.10	36.19	42.19	CKING ==> ATM
5	2	1.08	25.69	66.81	ATM ==> SVG
6	2	1.08	25.69	41.53	SVG ==> ATM
7	2	1.17	16.47	100.00	HMEQLC ==> CKING
8	2	1.17	16.47	19.20	CKING ==> HMEQLC
9	2	1.04	15.72	64.08	CD ==> SVG
10	2	1.04	15.72	25.40	SVG ==> CD

Number of rules 'discovered' (reported in descending order of support) is usually large, what is relevant and interesting and what is merely anecdotal?

For (mostly SAS) programmers only.

How could I program this if I just wanted to?

Go Proc Summary and IML, with good macro doses. Make sure you understand the _type_ variable before you start.

2. Business and other Cases.

Association Rules (basket analysis by example).

Young-family high-income professional bank customers, two groups: high- and low-profit.

Mutual funds,
mortgages
credit cards

Mutual funds, credit cards. **SURE ACTION:**

Sell mortgages to this Group?

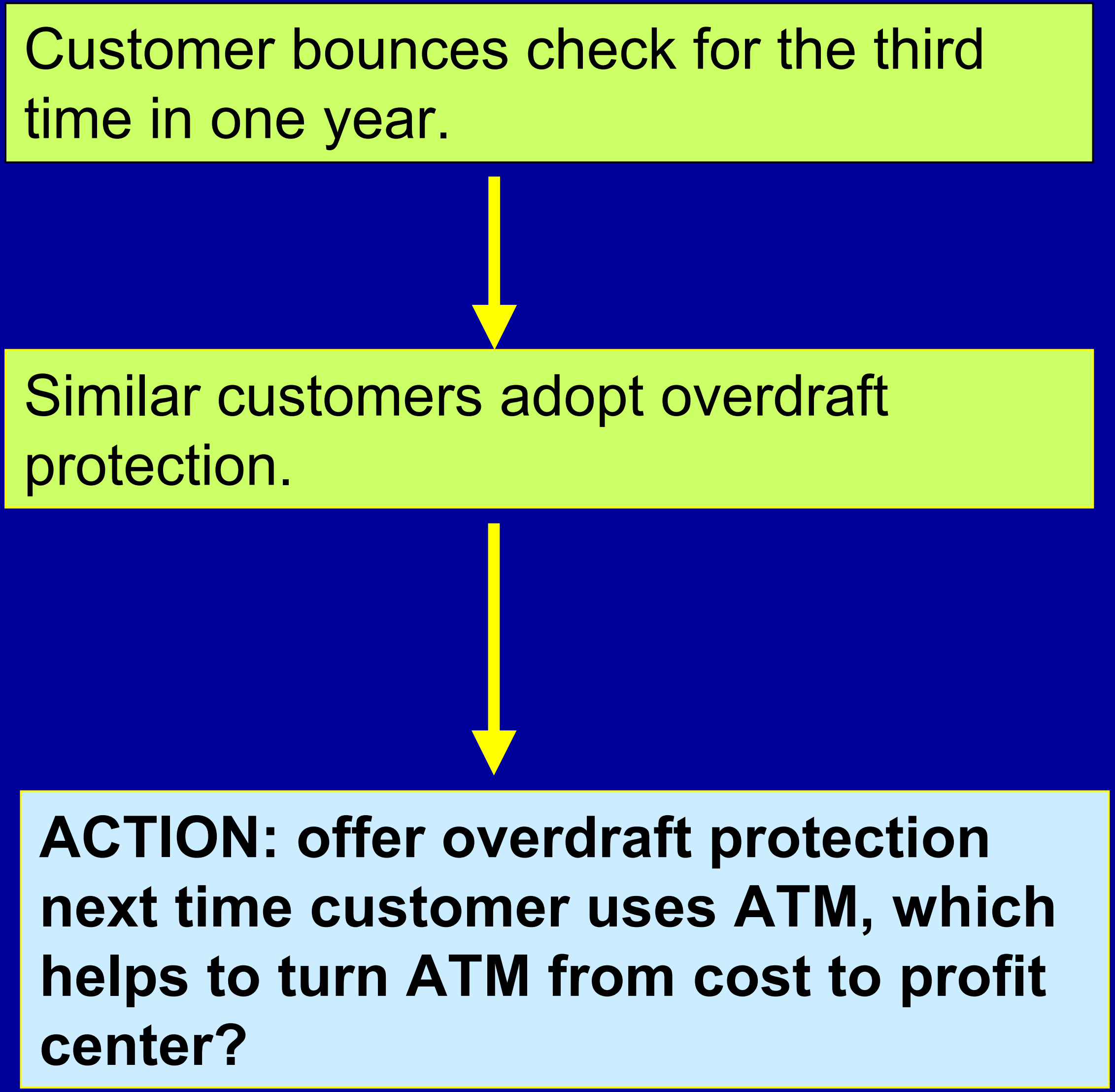
Association Rules (basket analysis by example).

Mutual funds,
credit cards. **SURE ACTION:**
Sell mortgages to this
Group.

Not necessarily successful. Banks have long profitably cross-sold checking and savings accounts, but successful cross-selling of other items requires **more** than 'opportunity' signal indicated by data mining:

1) Enhanced interdepartmental cooperation, and data sharing (pricing, short- versus long-view of the customer, etc).
2) Compatible customer programs.
3) Enhanced customer service.
4) Customer, and not product, focus.

Association Rules (basket analysis by example).



Customer bounces check for the third time in one year.

Similar customers adopt overdraft protection.

ACTION: offer overdraft protection next time customer uses ATM, which helps to turn ATM from cost to profit center?

Association Rules (Basket analysis by example).

Firemen interested in types of wiring ‘not’ correlated with fires (negative associations).

Larry W. Mabra, Lancaster FD

Clinical data analysis: symptoms (presence and absence) in disease.

Association Rules (Basket analysis by example).

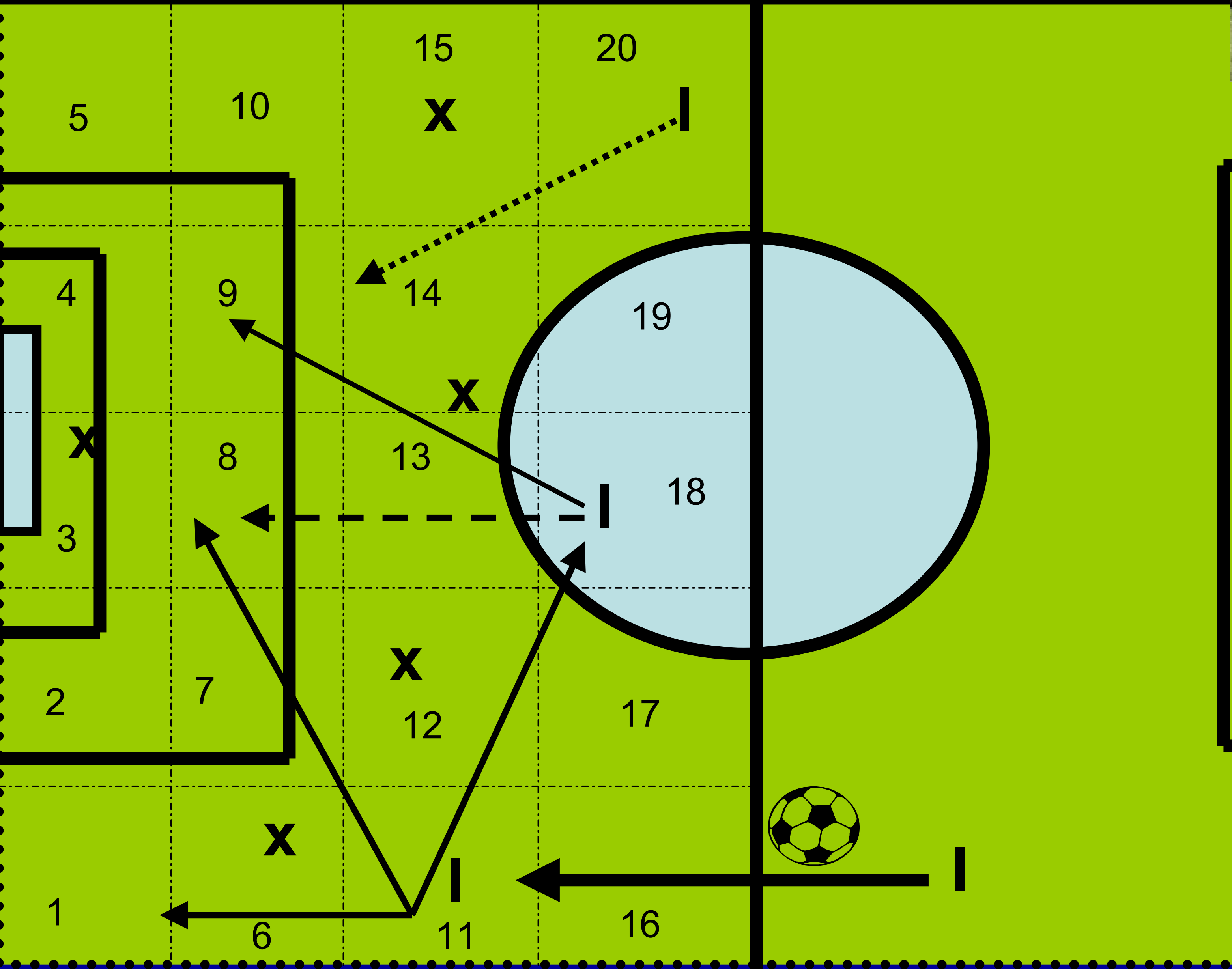
DuMouchel (1999): Empirical Bayes Analysis of Adverse Effects (AE) of Drugs, American Statistician.

Problem:

1) For given report, no certainty that Drug → reaction.
2) Sometimes, multiple reports per Patient, and no reports for many.
3) This implies that only reporting, And not occurrence, rates can be Calculated.

DuMouchel created very interesting Analytical method to find out Important Combinations of drug-AEs.

Association Rules (basket analysis by exa



Association Rules (basket analysis by example).

The science of deduction

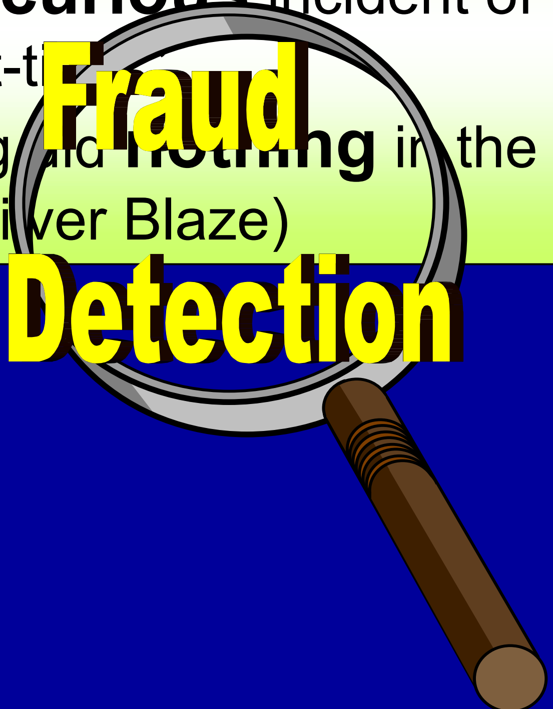
Screen role: Jeremy Brett was seen as the definitive Sherlock Holmes

"Once you eliminate the impossible, whatever remains, however improbable, must be the truth." (S. Holmes, Doyle 1981).

-"Is there any other point to which you would wish to draw my attention?"

-"To the **curious** incident of the dog in the night-time."

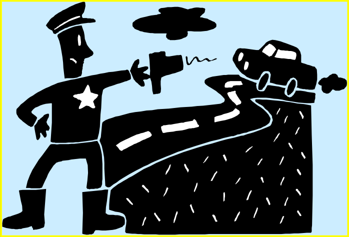
"The dog did **nothing** in the night-time". (Silver Blaze)



Association Rules (basket analysis by example).

Police Detection and Profiling

Crime #	Attr 1	Attr 2	Attr 3	Support	Confidence
1	#	#	#	%	%
2					
3					
4					
5					
6					
7					



Association Rules (basket analysis by example).

Analysis of Voter's preferences (MacDougall, 2003).

1) Analysts studied voters' survey of 2000 US election.

2) Did Clustering and Tree Modeling: un-interpretable and non-intuitive results.

3) Had 1800 variables plus voter preference.

4) By careful 'expert' perusal of rules generated, found 'profiles' to identify party preference. For instance, that 'female voters, with kids at home ….'.

5) Was also able to generalize by party preference. For instance, "Republicans may share similar backgrounds and a consensus of views within their party more than Democrats or Independents do."

Association Rules (basket analysis by example).

Recommender systems.

1) Amazon.com: "Readers who purchased this item, also purchased …"

2) Point of Sale Retail coupons: "Bought product of brand A, receive coupon for same product from brand B".

3) E-mail (spam) and snail solicitations due to web browsing, charity donation, credit card purchase ….

4) Web Search Engines (actually use "Link Analysis"). Some use Neural Networks in addition for clustering.

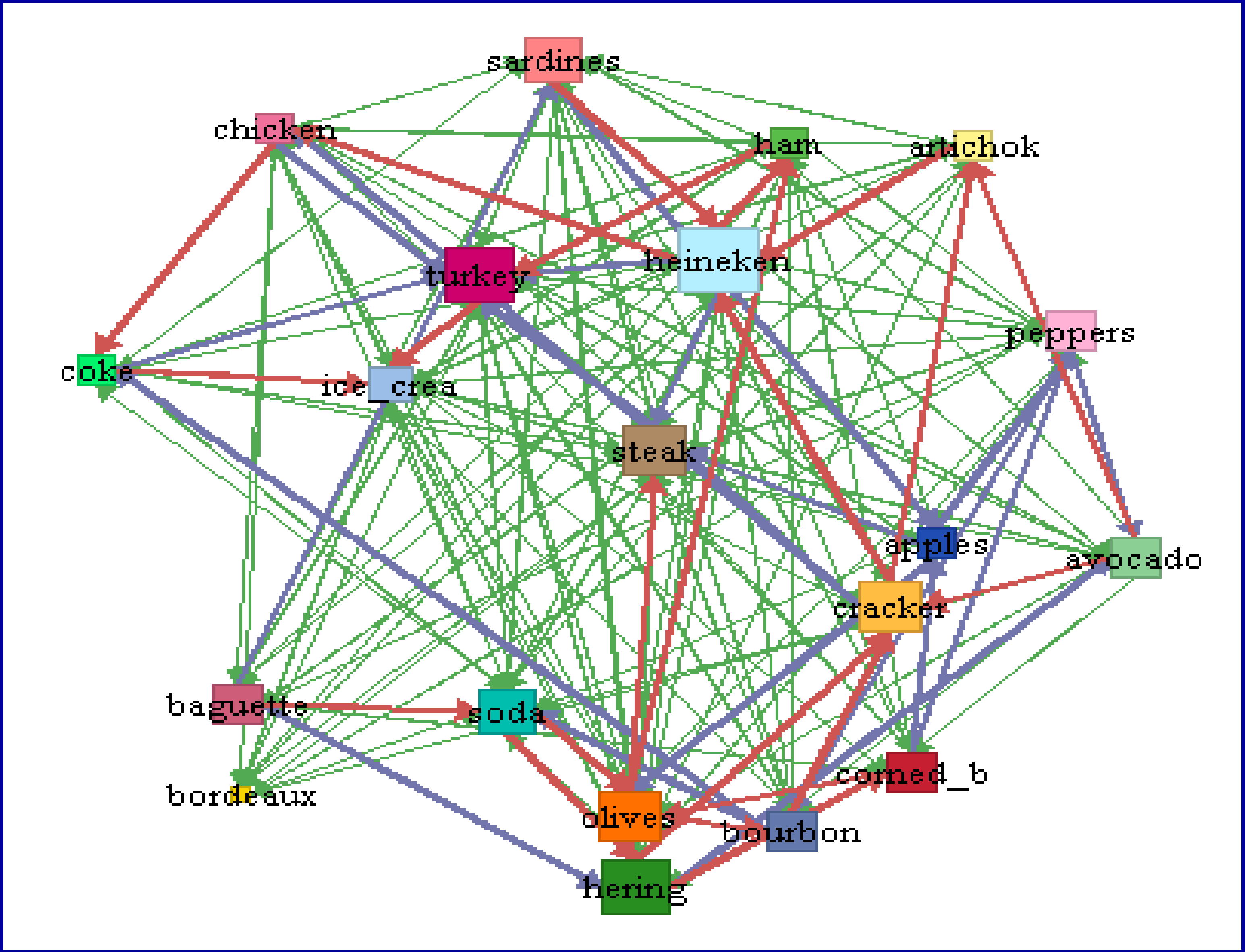
Slight Detour: Link Analysis.

Link Analysis **(LA) reveals and visually represents complex patterns of correlations between individual values of categorical and Boolean attributes.**

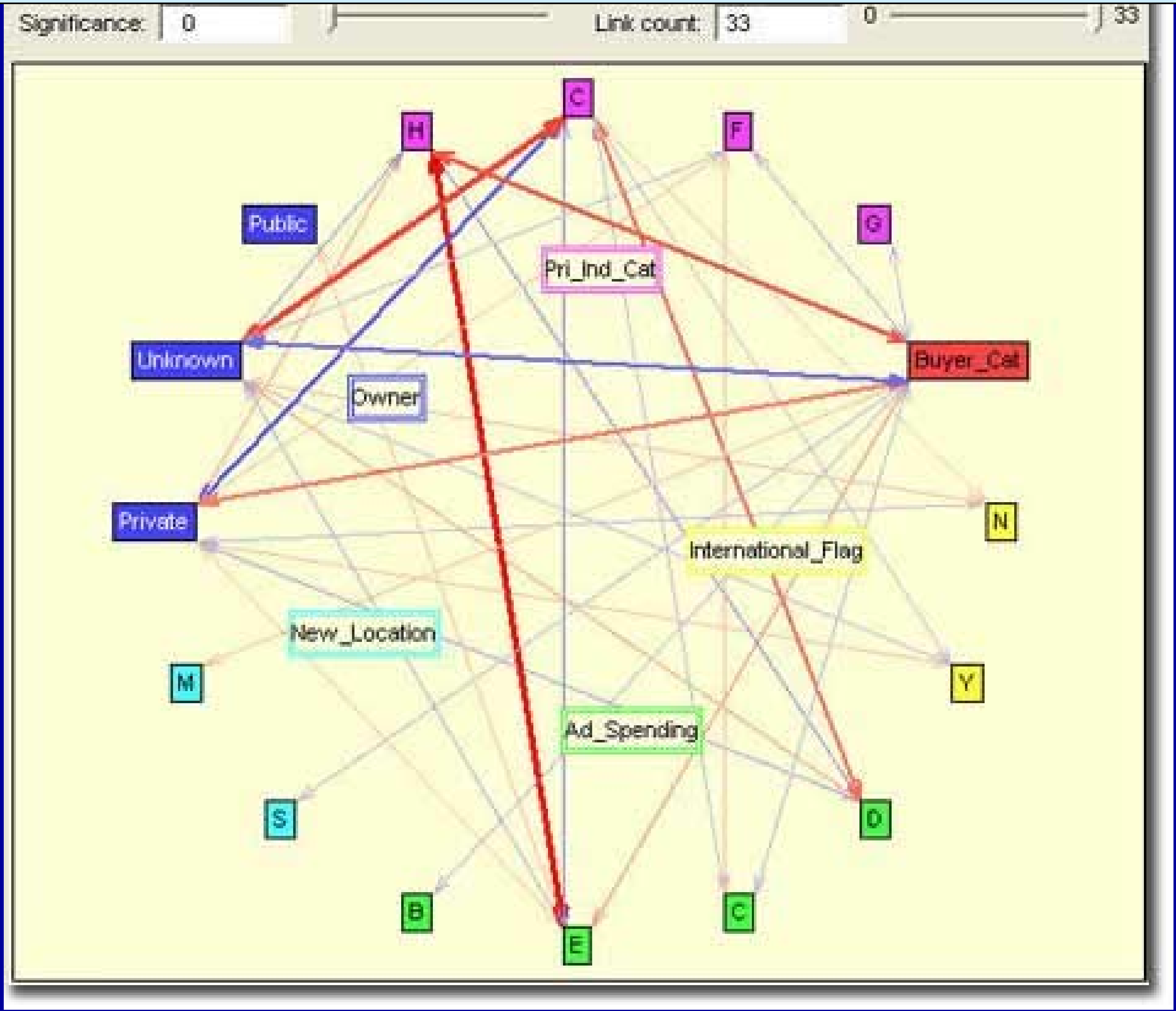
Results of analysis are displayed as graph of linked objects supporting various object manipulation and drill-down operations.

Visual output of LA facilitates better understanding of hidden structure of investigated data, and helps to quickly isolate interesting patterns for further investigation.

Slight Detour: Link Analysis (cont. 1).



Slight Detour: Link Analysis (cont. 2).



Significance: 0

Link count: 33

Slight Detour: Link Analysis (cont. 3).

***Database Marketing*. Reveal typical characteristics of the best customers; identify frequent patterns in purchase behavior.**

***Fraud Detection*. Uncover and flag as suspicious those provider-patient or patient-procedure or customer-business pairs that demonstrate unusually high correlations.**

***Communications Analysis*. Visualize main communication patterns and potential network bottlenecks.**

***Criminal Investigations*. Display overall organizational structure, identify articulation points, and trace money laundering and illegal goods transfer patterns.**

***Text Analysis*. *A*bility to visually display clusters of terms extracted from textual notes to further investigate correlations.**

Association Rules. Lingo.

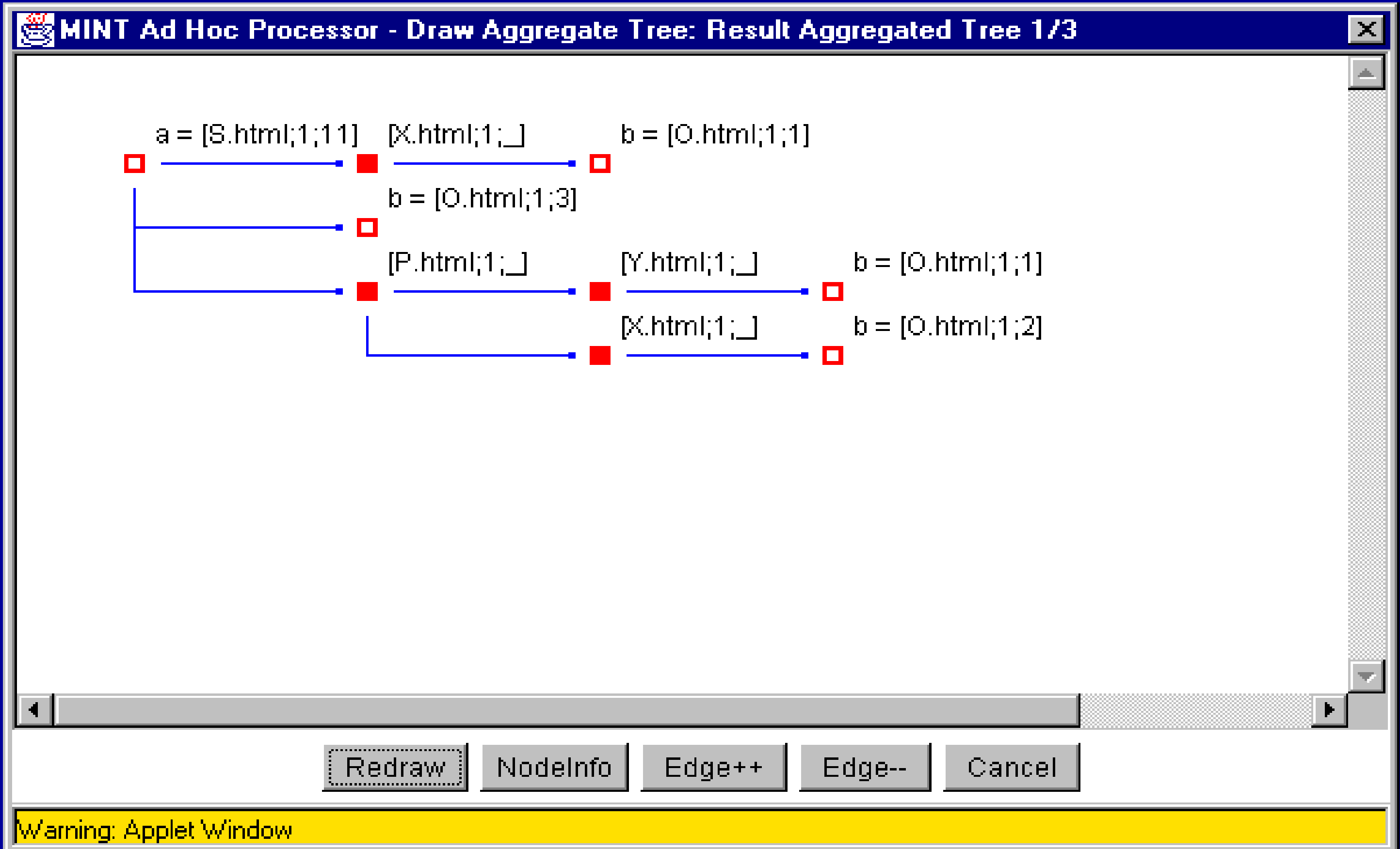
Association or Market Basket analysis was ‘the’ tool that brought in the term “Knowledge Discovery”.

Another piece of lingo is that data contains precious “Nuggets” of information that can be “mined”, and that brought in the label “Data Mining”.

Artificial Intelligence and Machine Learning were closer to robotic investigation and neural networks. Statisticians’ interest in the latter brought the fields closer.

Sequencing Rules.

Similar methodology, in which individual items have index of precedence associated with them. For instance, click-stream on web. Spiliopoulou (1999, WUM) provides sequencing tools based on support to identify underlying structure of click-streams in web-site.



MINT Ad Hoc Processor - Draw Aggregate Tree: Result Aggregated Tree 1/3

Warning: Applet Window

Analogy.

Since every item in every rule denotes implicitly presence or absence of the item in every rule, each rule can also be considered to be a representation of all available items as dummy variables, where a ‘0’ denotes absence of a specific item, and ‘1’ its presence.

Thus, if there are 5 UPCs in rule studied by retailer, rules could be considered to be as presented in following table.

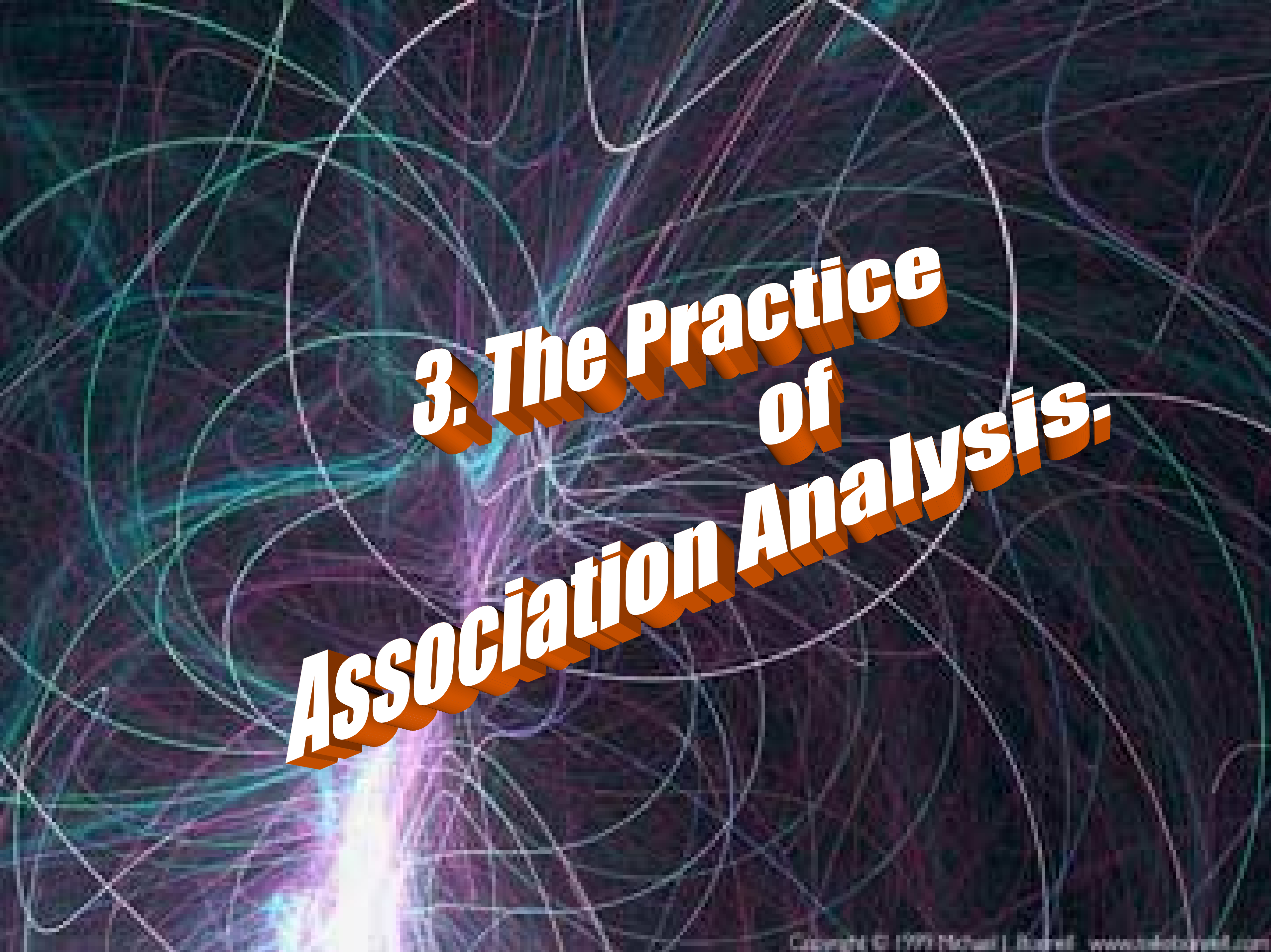
Rule	UPC 1	UPC 2	UPC 3	UPC 4	UPC 5
1	1	0	0	1	1
2	1	1	1	0	0
3	……	…	…	…	…
4		…			
5	…		…	…	…
6					
7					

Analogy (cont.1).

Therefore, if every rule (adding its negative rule as well as additional variables) can be linked to characteristic important to business, such as total expenditure (and remembering that the rules don’t incorporate quantities), a regression like environment to explain total expenditure is possible.

From it, insights into elasticity estimation are possible, for instance.

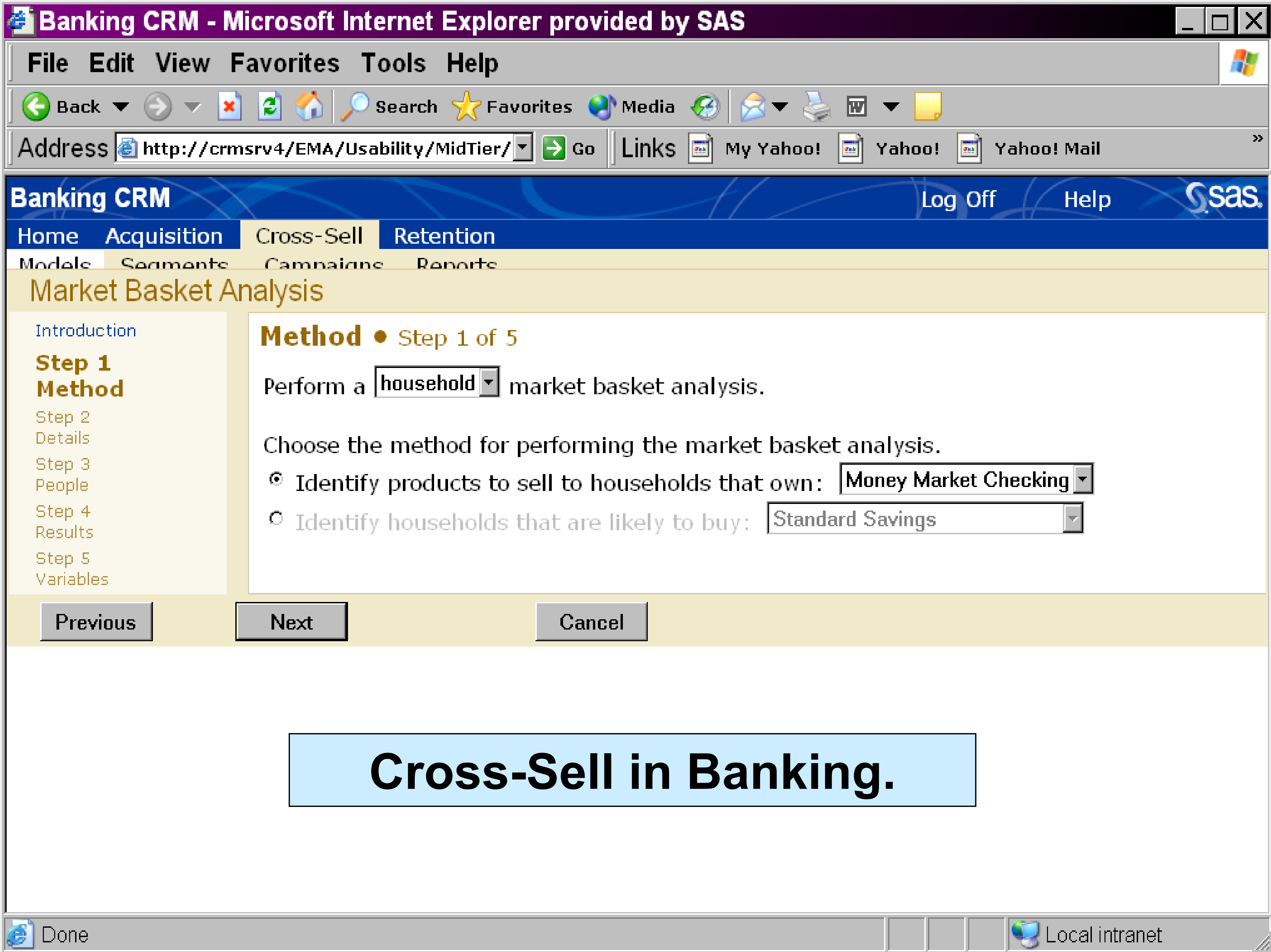
3. The Practice of Association Analysis.



Application: Barbie® ⇒ Candy (conf = 60%) Rules for action or inaction? Illuminating answers to On-line survey.

- 1. Put them closer together in the store.
- 2. Put them far apart in store (…. don’t hire this guy).
- 3. Package candy bars with dolls.
- 4. Package Barbie® + candy + poorly selling item.
- 5. Raise price on one, lower it on other.
- 6. Barbie® accessories for proofs of purchase.
- 7. Do not advertise or promote candy and Barbie® together (hire this guy).
- 8. Candies in the shape of a Barbie® Doll (send this guy back to marketing school).

Aside: Tobacco companies have been charged of strategically locating their products so that kids get ‘influenced’ at early age.


Banking CRM - Microsoft Internet Explorer provided by SAS
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://crmsrv4/EMA/Usability/MidTier/ Go Links My Yahoo! Yahoo! Yahoo! Mail
Banking CRM
Log Off Help
sas
Home Acquisition Cross-Sell Retention
Models Segments Campaigns Reports
Introduction
Step 1 Method
Step 2 Details
Step 3 People
Step 4 Results
Step 5 Variables
Previous Next Cancel
Done
Local intranet

Market Basket Analysis

Method ● Step 1 of 5

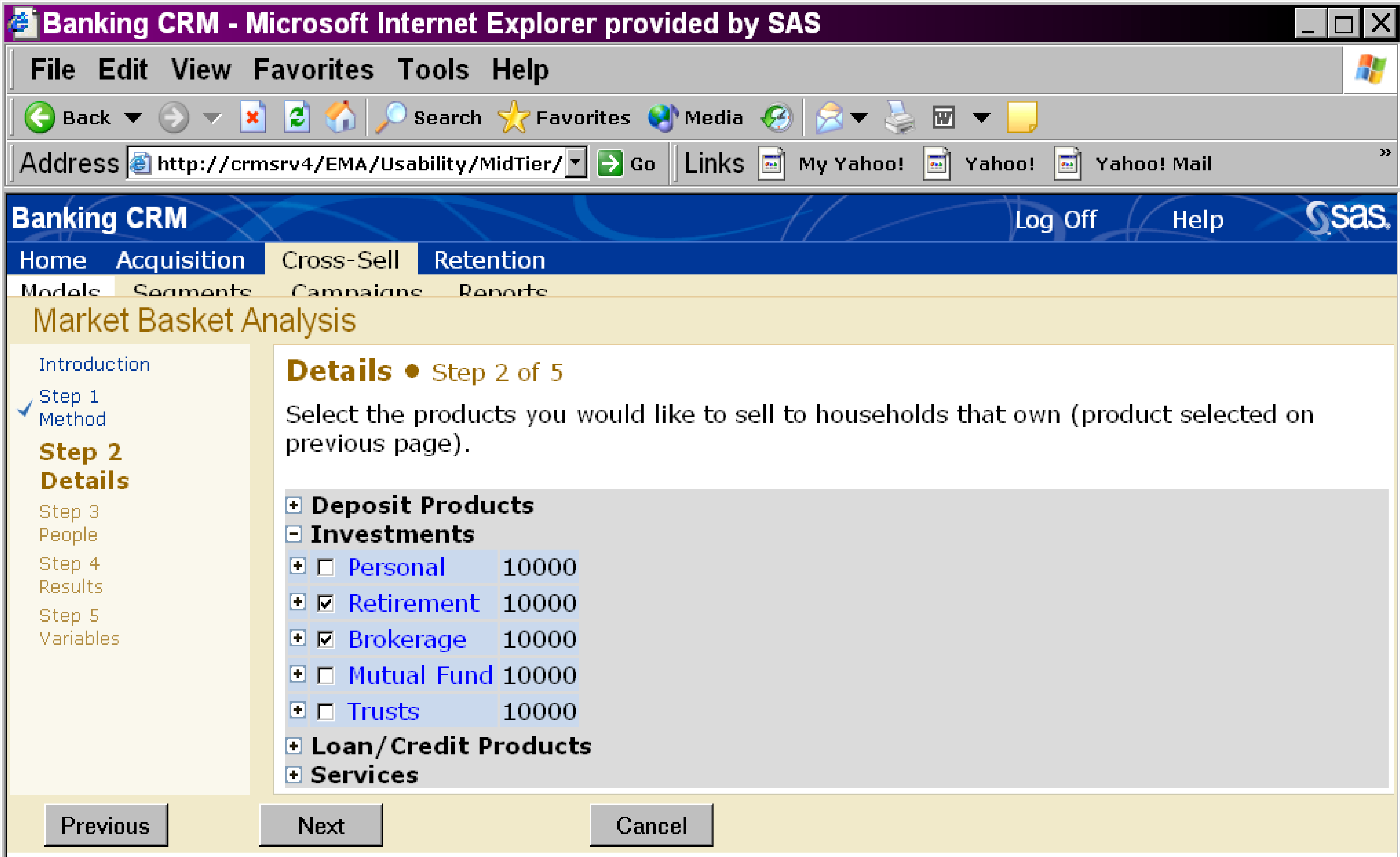
Perform a household market basket analysis.

Choose the method for performing the market basket analysis.

● Identify products to sell to households that own: Money Market Checking

○ Identify households that are likely to buy: Standard Savings

Cross-Sell in Banking.


Banking CRM - Microsoft Internet Explorer provided by SAS
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://crmsrv4/EMA/Usability/MidTier/ Go Links My Yahoo! Yahoo! Yahoo! Mail
Banking CRM Log Off Help sas
Home Acquisition Cross-Sell Retention
Models Segments Campaigns Reports
Introduction
✓ Step 1 Method
Step 2 Details
Step 3 People
Step 4 Results
Step 5 Variables
Previous Next Cancel

Market Basket Analysis

Details • Step 2 of 5

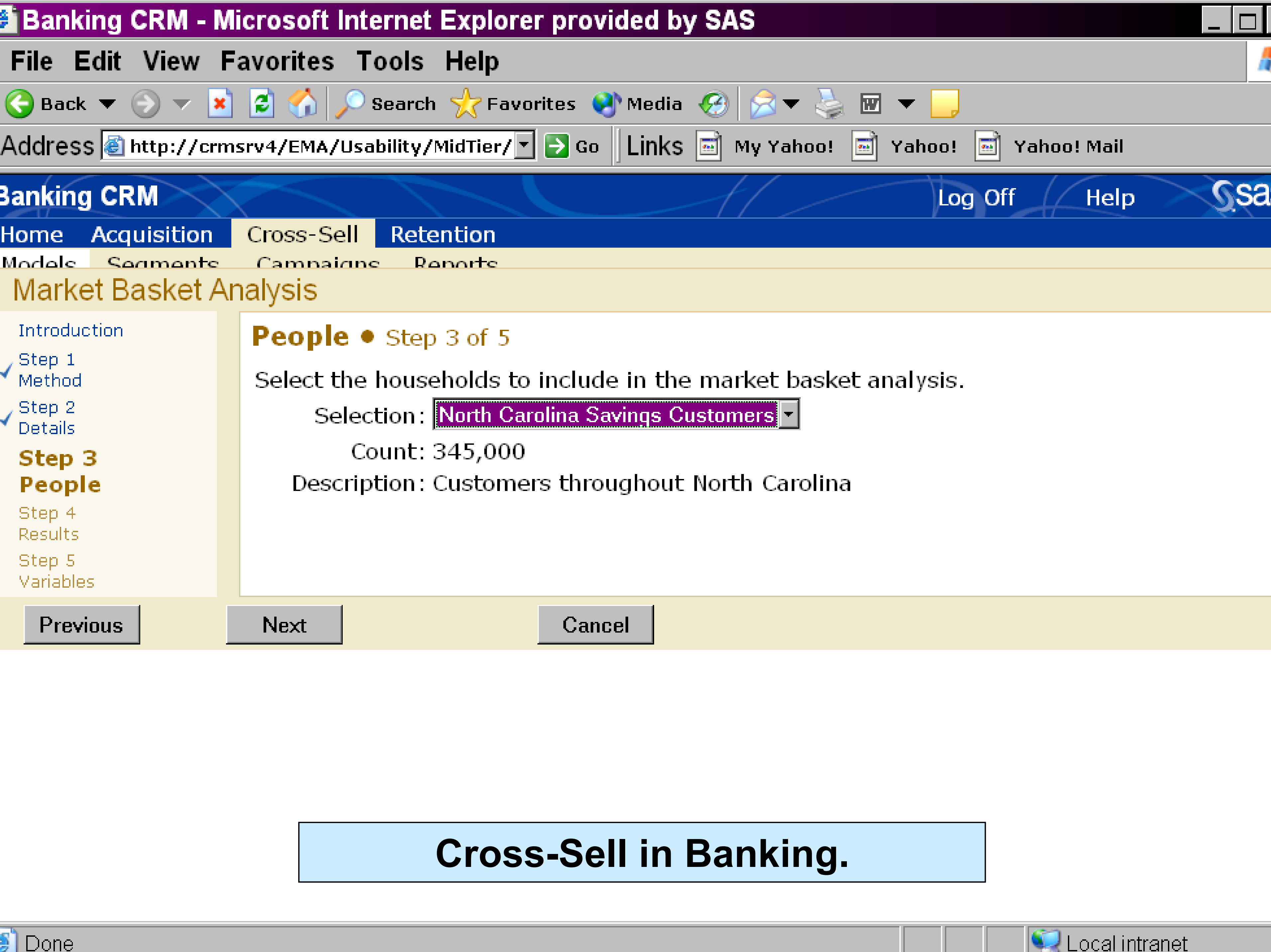
Select the products you would like to sell to households that own (product selected on previous page).

- ⊞ Deposit Products
- ⊟ Investments

⊞	□	Personal	10000
⊞	☑	Retirement	10000
⊞	☑	Brokerage	10000
⊞	□	Mutual Fund	10000
⊞	□	Trusts	10000

- ⊞ Loan/Credit Products
- ⊞ Services

Cross-Sell in Banking.



Market Basket Analysis

People ● Step 3 of 5

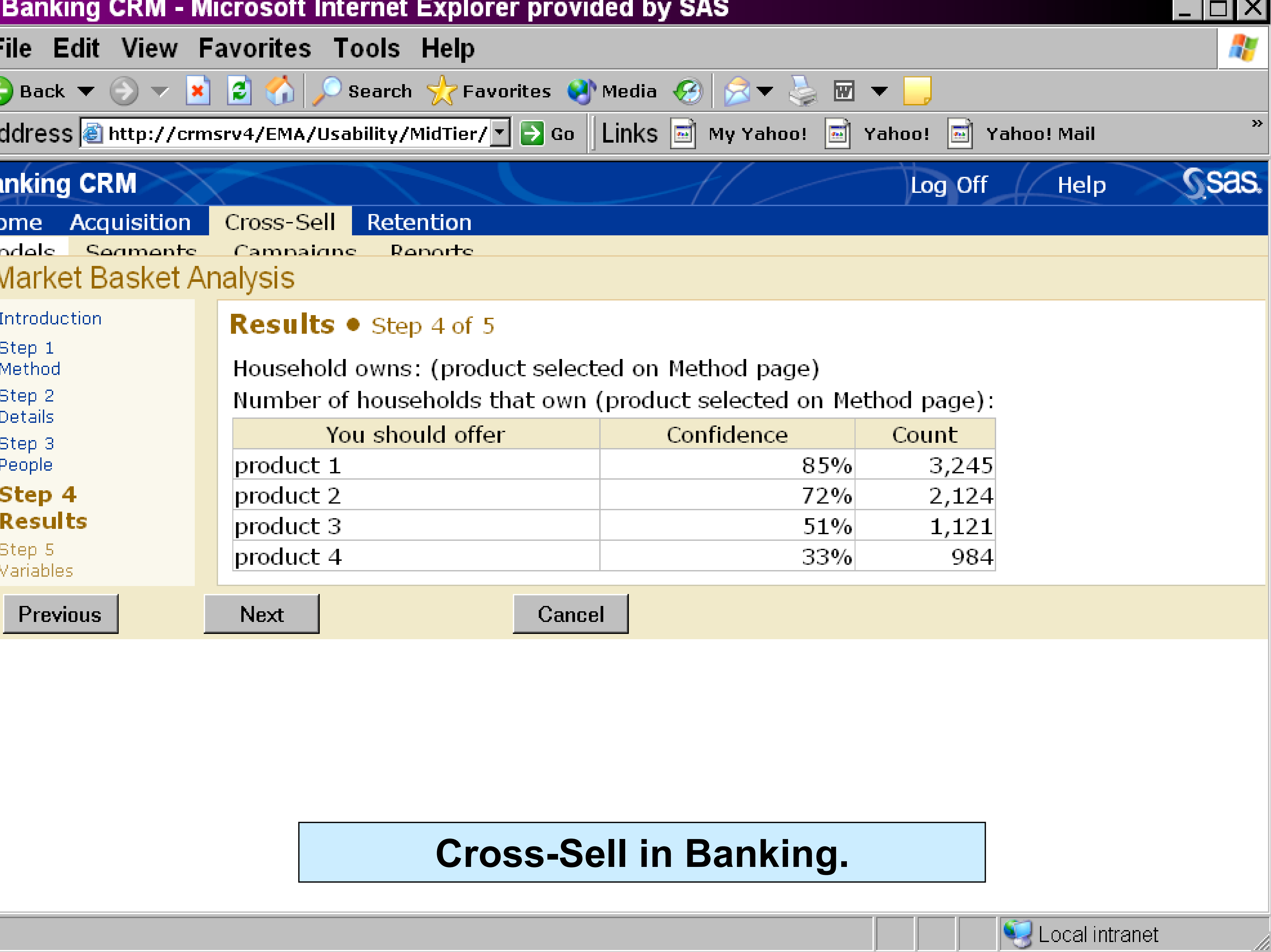
Select the households to include in the market basket analysis.

Selection: North Carolina Savings Customers

Count: 345,000

Description: Customers throughout North Carolina

Cross-Sell in Banking.


Banking CRM - Microsoft Internet Explorer provided by SAS
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress http://crmsrv4/EMA/Usability/MidTier/
Go
Links My Yahoo! Yahoo! Yahoo! Mail
nking CRM
Log Off
Help
sas
ome Acquisition Cross-Sell Retention
odels Segments Campaigns Reports
Introduction
Step 1 Method
Step 2 Details
Step 3 People
Step 4 Results
Step 5 Variables
Previous
Next
Cancel
Local intranet

Market Basket Analysis

Results ● Step 4 of 5

Household owns: (product selected on Method page)

Number of households that own (product selected on Method page):

You should offer	Confidence	Count
product 1	85%	3,245
product 2	72%	2,124
product 3	51%	1,121
product 4	33%	984

Cross-Sell in Banking.

General Observations.

1) Banking case seems to provide well defined and intelligible information of the form:

   account_1 and account_2,,, etc or activity_1 and activity_2, etc, possibly indexed by time.

   As such, rules found provide guide to action to ‘offer’ product or service (cross-sell).

   Note that political survey and police criminal profiles can also be used to produce ‘offers’ to the found profiles.

2) Sherlock Holmes example shows that lack of attribute creates interesting profile for solution.

General Observations (cont. 1).

3) In retailing case of items purchased together, ‘guidance’ is not so clear cut due to extensive number of rules. However, more extensive analysis and possible solution is presented below.

4) Soccer event exemplifies sequencing of events towards reaching goal. Basketball-applied software has been developed years ago. Web-mining shares the same principles, without passion usually associated with sports.

5) Unit of analysis is problem dependent. In retailing, the unit is each transaction, independent of the customer’s identity. Thus, small group of habitual shoppers may skew results. In clinical data analysis, the unit is patient, who contributes single observation.

4. Adding Modeling to Associations Effects.

Information issues, and Modeling?

1) What question/s does Associations Analysis answer, which one/s does not, and what information is not fully utilized?

2) Adding to 1), can we make sense of the information provided?

3) Modeling typically involves a target or dependent variable. Can modeling somehow be combined with Associations Analysis?

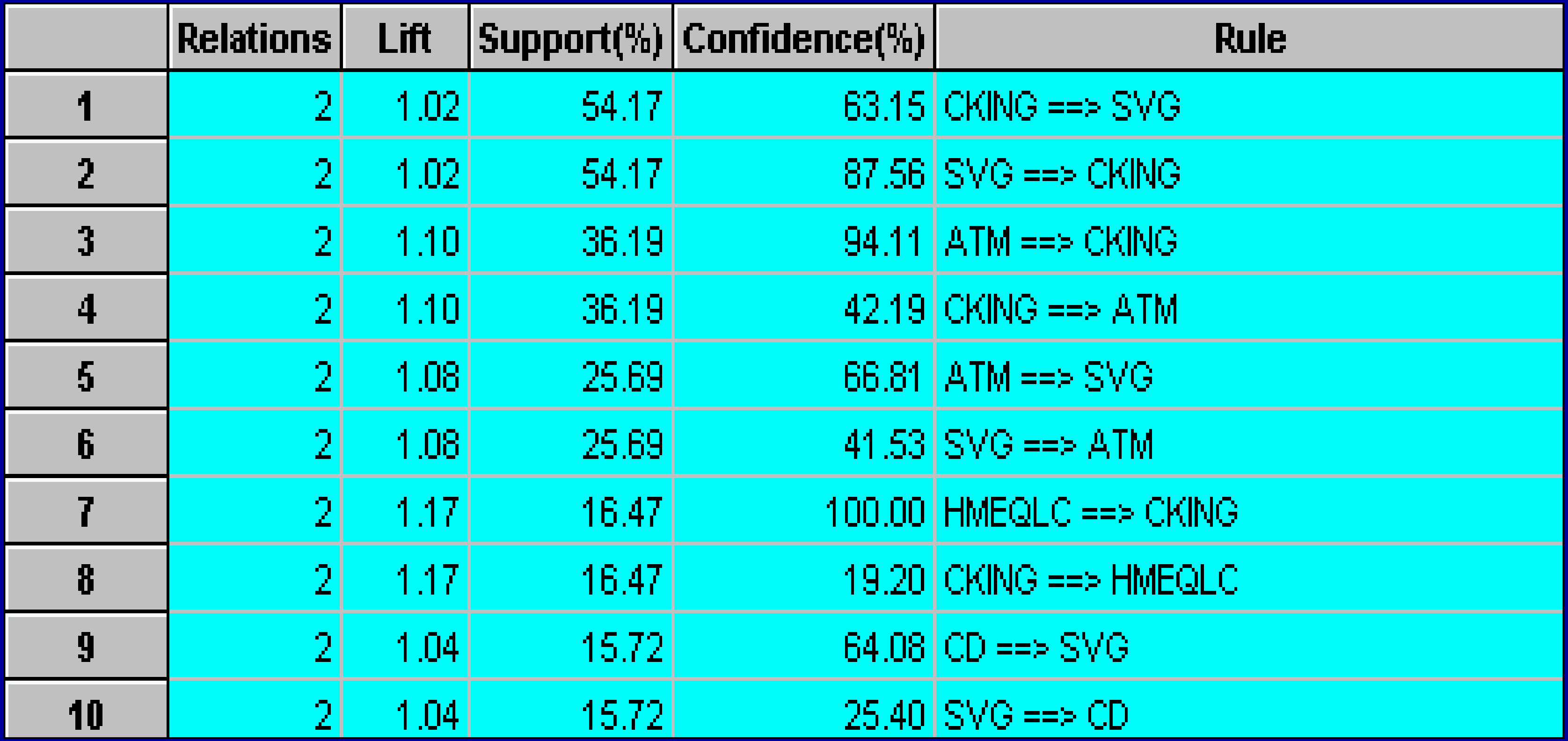
What questions are answered?

1) Generated rules are different combinations of K dummy variables. Rules over pre-specified support 5%) and only those with ‘present’ effects (‘1’) are typically reported.

2) This implies that ‘seemingly’ variable selection effect of association analysis is at best illusory. That is, if ‘correct’ set of variables is not analyzed, there is no rule to indicate that the findings ‘fit’ to maximize profits, to diagnose a disease correctly, etc.

Rule	UPC 1	UPC 2	UPC 3	UPC 4	UPC 5
1	1	0	0	1	1
2	1	1	1	0	0
3	……	…	…	…	…
4		…			
5	…		…	…	…
6					
7					

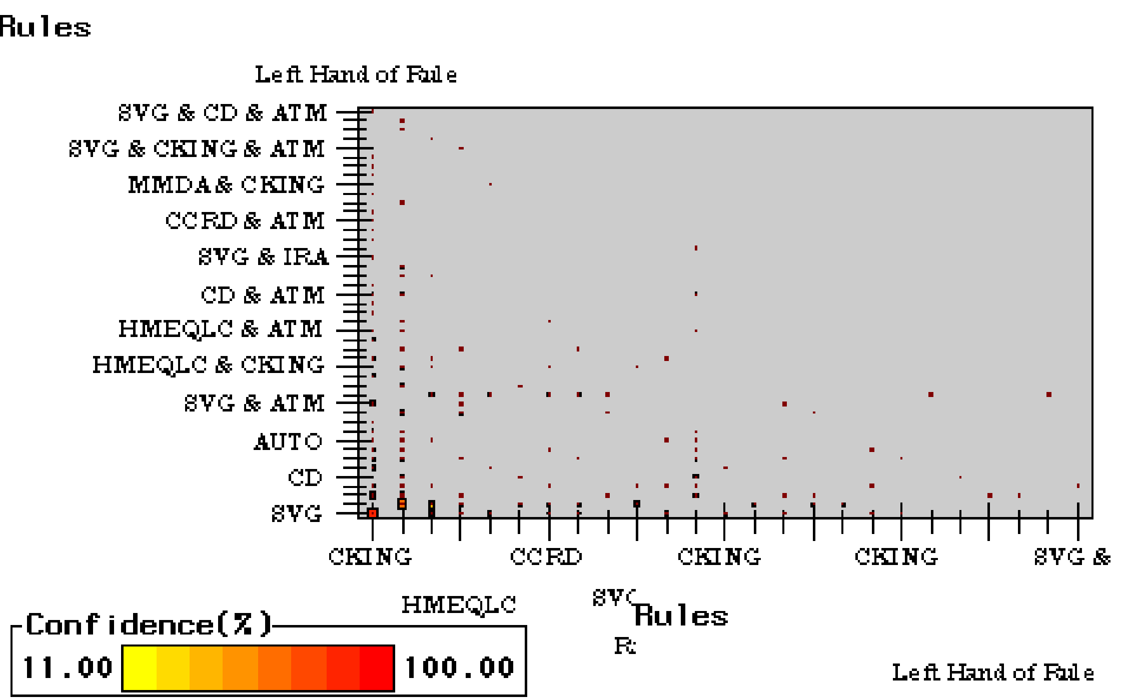
Biggest issue: is this informative?
Example: SAS Enterprise Miner Association Node (clipped) set of rules



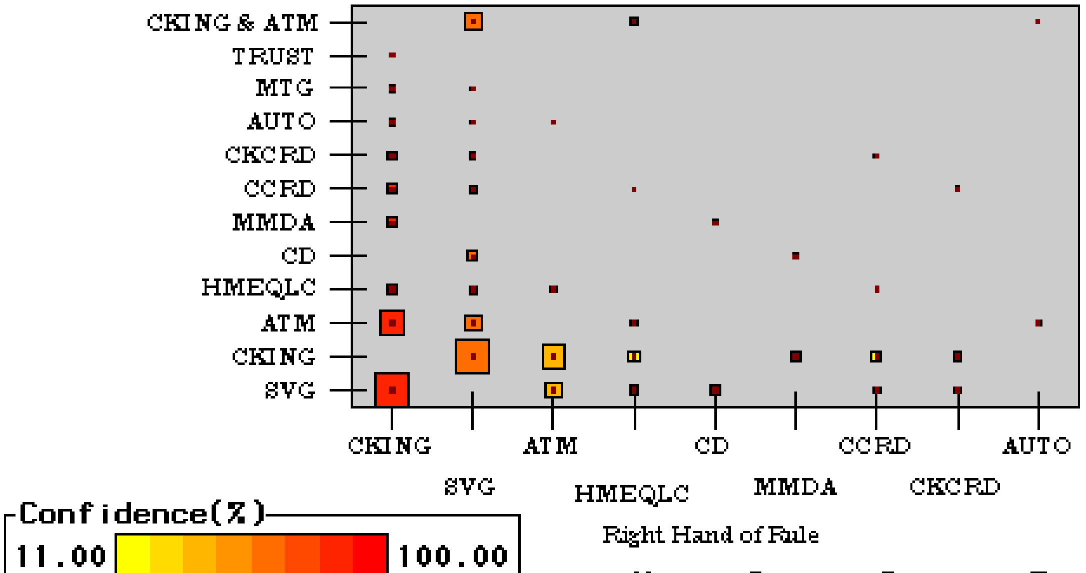
	Relations	Lift	Support(%)	Confidence(%)	Rule
1	2	1.02	54.17	63.15	CKING ==> SVG
2	2	1.02	54.17	87.56	SVG ==> CKING
3	2	1.10	36.19	94.11	ATM ==> CKING
4	2	1.10	36.19	42.19	CKING ==> ATM
5	2	1.08	25.69	66.81	ATM ==> SVG
6	2	1.08	25.69	41.53	SVG ==> ATM
7	2	1.17	16.47	100.00	HMEQLC ==> CKING
8	2	1.17	16.47	19.20	CKING ==> HMEQLC
9	2	1.04	15.72	64.08	CD ==> SVG
10	2	1.04	15.72	25.40	SVG ==> CD

If number of rules ‘discovered’ is, say, 500, how can we conceptualize hidden information, if any?
or do we merely consider this counting exercise (similar to accounting) with massive reporting printouts?

Difficulties of graphing and viewing Associations.



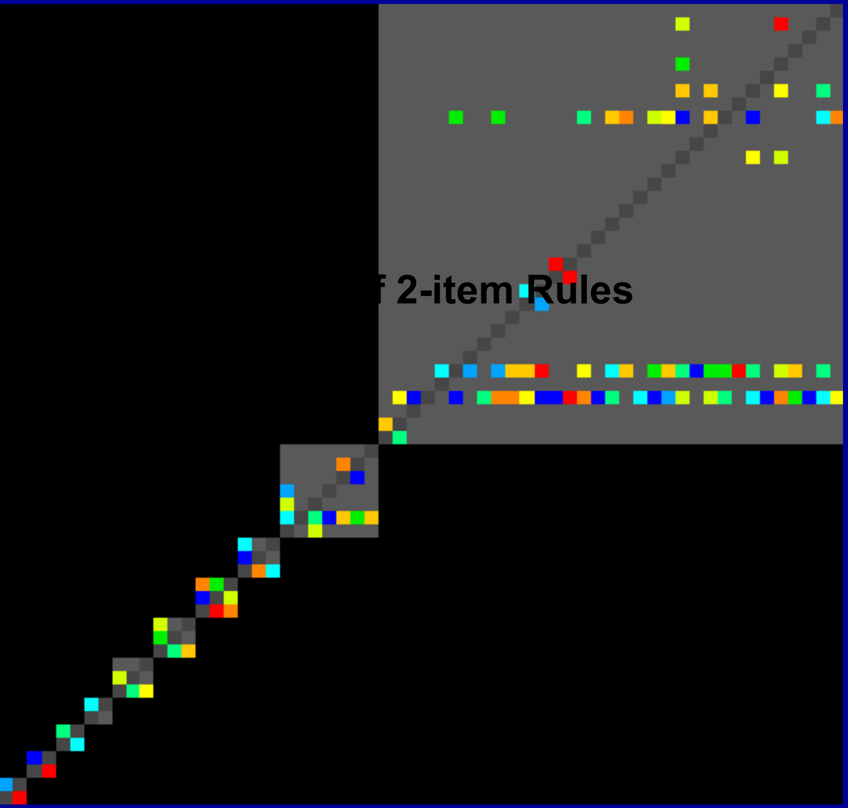
Enterprise Miner V3.01



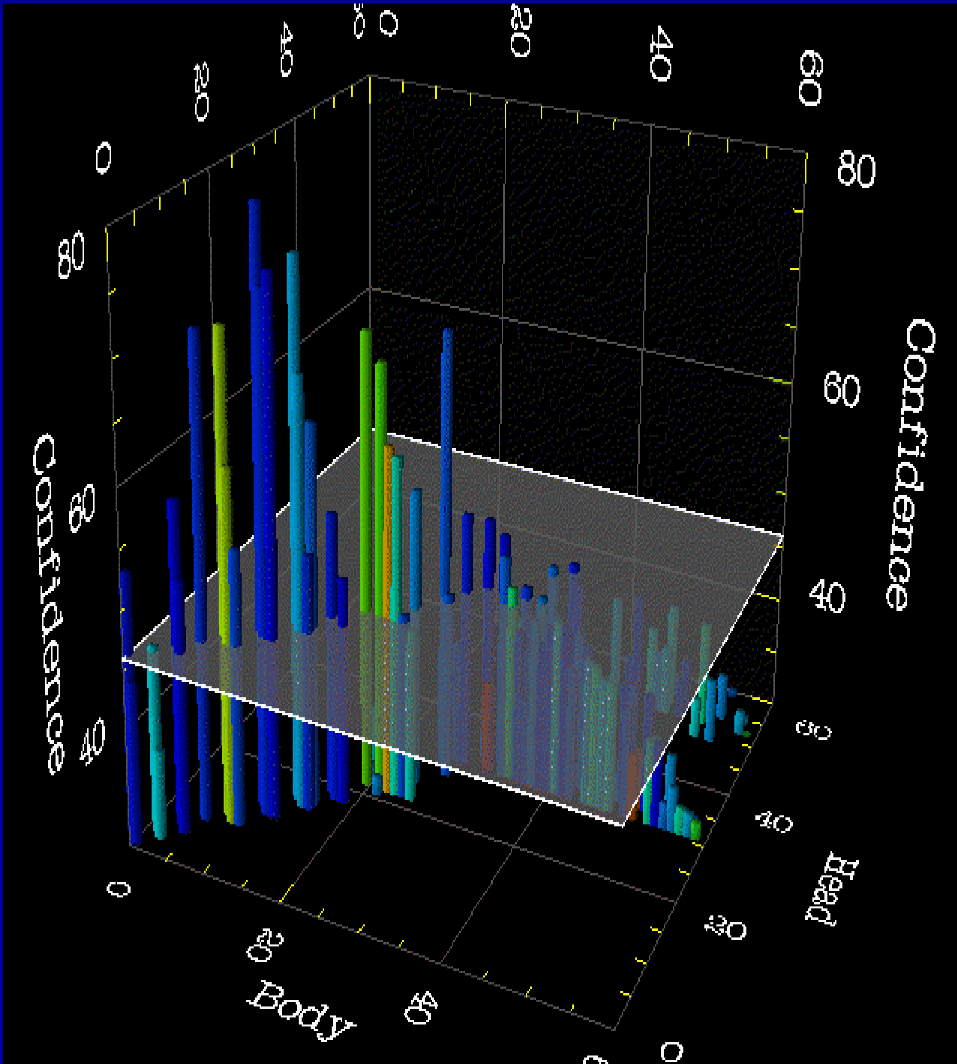
Note: Size = Support %

Difficulties of graphing and viewing Associations.

2-D Visualization of 2-item Rules



3-D Visualization of 2-item Rules



IBM Intelligent Miner (from WEB, 2001)

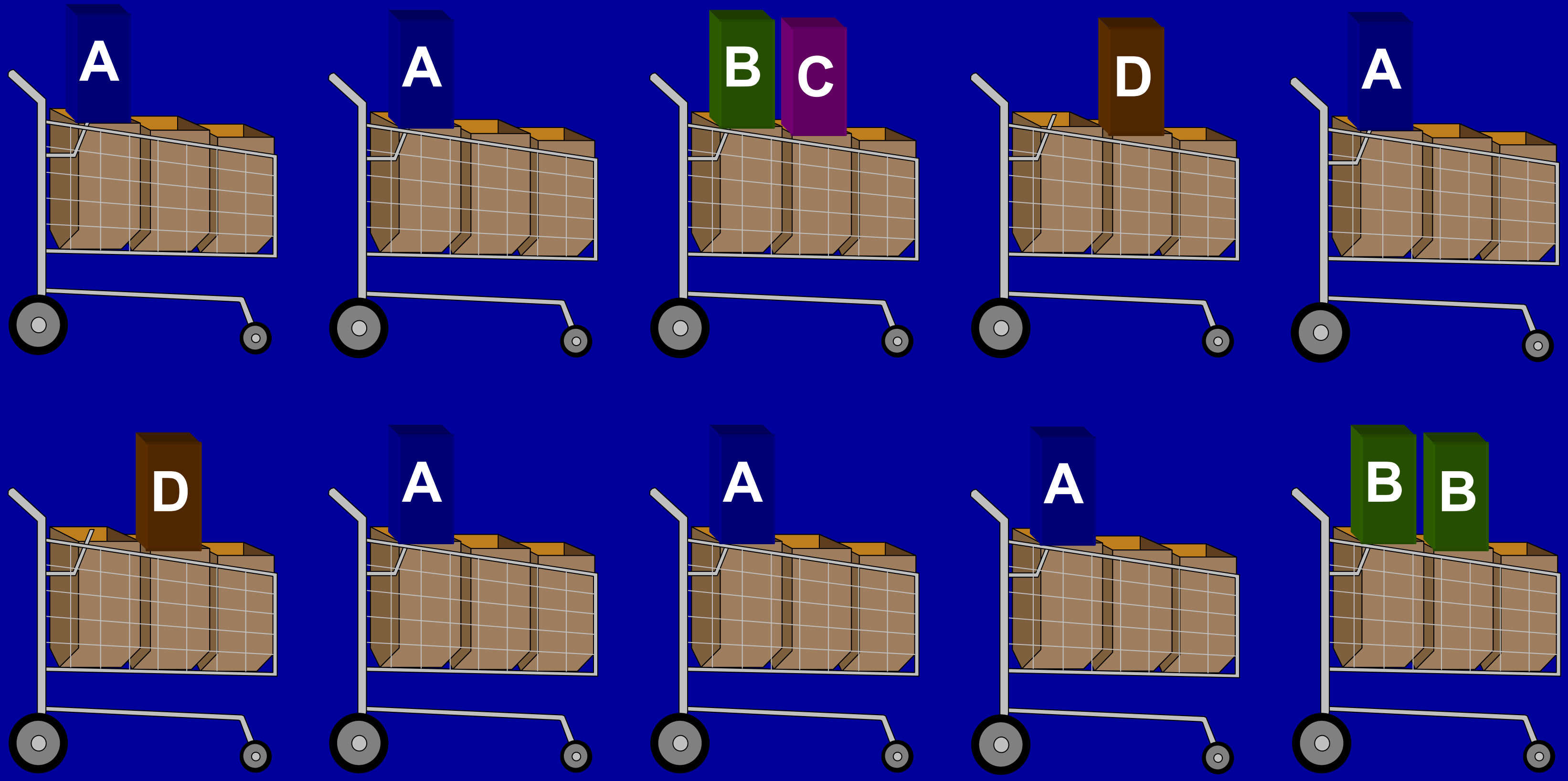
Additional Issues.

Present methodology does not allow for different amounts of items. I.e., there is no way to relate rule “3 pairs of socks → 3 pairs of shoes” with the rule “1 pair of socks → 1 pair of shoes”, nor with any “1 watermelon → 1 pair of socks”. Quantities alter the meaning of items.

There is no “interestingness” measure. Interestingness is provided by human brains, if there are any, of course (see Holmes’ Silver Blaze above).

Similar to “interestingness”, “propensity to purchase”, “probability of contracting a disease”, etc, are interpretations that must be provided by a business case analysis, a scientific model, or sheer common sense, if any.

Data Capacity: is the data informative for the problem at hand?



Data become sparse in a multidimensional setting.

Data Capacity: Curse of Dimensionality.

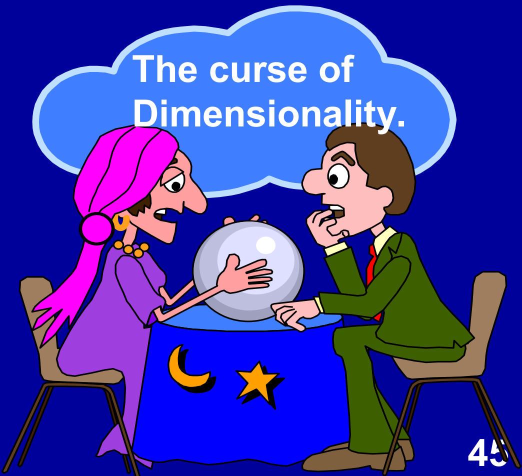
Let k indicate number of binary attributes, indicating presence/absence, purchase/no-purchase, sick/healthy, etc. Let 0/1 denote the binary nature of each attribute.

For k = 2, maximal number of cells is 4, { (0, 0), (0, 1), (1, 0), (1, 1)}. For k = 10, maximal number of cells is 1024, and for k = 20 it is 1,048,576.

That is, maximal number of cells is 2 ** k, which grows very quickly. Associations Analysis, a counting technique, does not take this into consideration.

Modeling, on the other hand, cannot avoid it.

Note: extremely unlikely that customer/
Patient would score in 2 ** k items/adverse
Effects, for k "large" ➔ need for heavy
Pruning of non-core rules.



4.1 Giudici

Passerone_ and

Odds and Odds Ratios.

Example: Location Preference of Soldiers training in WWII (North/South) (Rudas, 1998)

	Personal Preference.	
Origin	N	S
N	3,092	958
S	959	3,027
	4,051(50.4%)	3985(49.6%)

Total = 8,036

Odds (N vs. S) = 4,051/3,985 = 1.02

Odds (N vs. S / N origin) = 3.23
Odds (N vs. S/ S origin) = 0.32

➔Suspect association between origin and preference.

Similarly, there is association between present location and preference. Which one is stronger, and how are they related?

Strength: odd ratios; Relation: conditional Odd ratios.

Odds and Odds Ratios (cont. 2).

2-way Contingency Table – Conditional Probabilities.

		x	
		1	0
Y	1	π_1 (π_{11})	$(1 - \pi_1)$ (π_{12})
	0	π_2	$(1 - \pi_2)$

Let π = Probability of success = Pr (Y = 1 / X = 1); and (1 – π) = Pr (Y = 1 / X = 0).

Odds (Y = 1) = Ώ = π / (1 – π) >= 0.

When Ώ > 1 ➔ success more likely than failure. Odds equally defined for Y = 0 ➔

Odds and Odds Ratios (cont. 3).

Sample Odds ratio (2 – way contingency table).

Odds ratio = θ = Ώ1 / Ώ2.

For joint distributions with cell probabilities $\{\pi_{ij}\}$, odds ratio is

$$\theta = \pi_{11}\ \pi_{22}\ /\ \pi_{12}\ \pi_{21}.$$

Interpretation and Properties of Odd Ratios.

$0 \leq \theta \leq \infty$. If $\theta = 1$ ➔ independence of X and Y. When $\theta > 1$, subjects in row 1 are more likely to have 'success' than subjects in row 2, i.e., $\pi_1 > \pi_{2.}$

E.g., $\theta = 5$ ➔ odds(Y = 1) is 5 * odds (Y = 0), not that $\pi_1 = 5\ \pi_2$.

Another measure is relative risk: $\pi_1 / \pi_{2.}$

Odds and Odds Ratios (cont. 4).

Estimated odd ratio: θ(hat) = n_{11} n_{22} / n_{12} n_{21}., which equals 0 or ∞ if any element is 0, and undefined if any two entries in numerator and denominator are 0.

The distribution of θ(hat) is highly skewed because, unless n is very large, the lower bound is 0 and the upper one is ∞ with non-negligible probability.

➔ log-transform that converges to normality more quickly.

Estimated SE for log(θ(hat)) = σ(hat) (θ(hat)) =

$$sqrt(\sum_{i}^{2}\sum_{j}^{2} 1/n_{ij})$$

By asymptotic normality, the CI is:

Log θ(hat) + | z $_{\alpha/2}$ σ(hat) (θ(hat))|

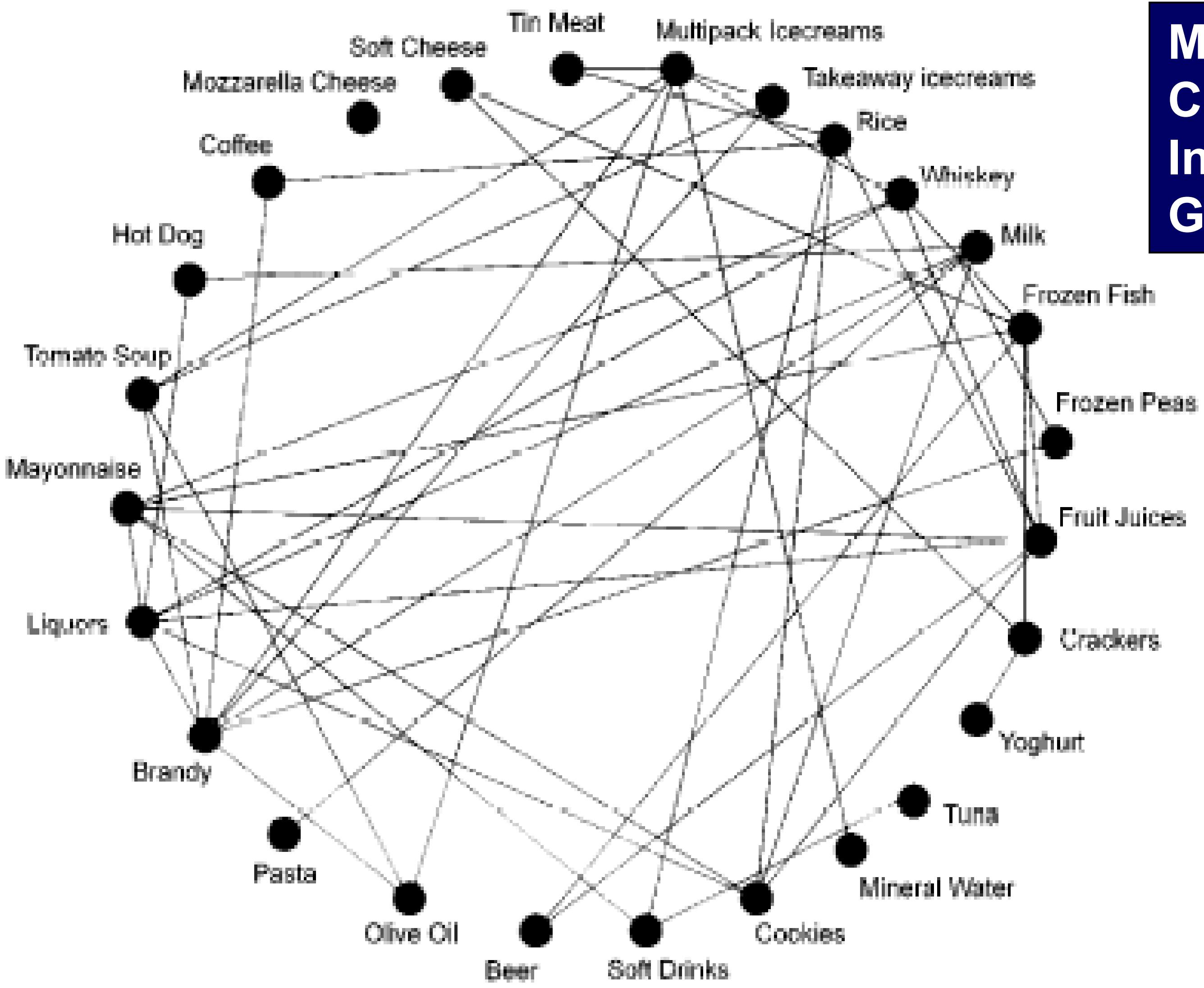
Giudici and Passerone (2002).

Giudici's and Passerone's (2002) approach focuses on pairs of grocery items. Given K items, form all K (K – 1) / 2 contingency tables, and conclude that pair of items is associated whenever odds ratio, corresponding to marginal independence, lies outside confidence interval of odds = 1.

They also provide graphical description of relations, borrowed from link analysis, which shows 36 dependency relations, given by edges of the graph. They focus analysis then on absolute odd ratios significantly larger than 5, which leads them to 5 'clusters', completely disconnected from each other.

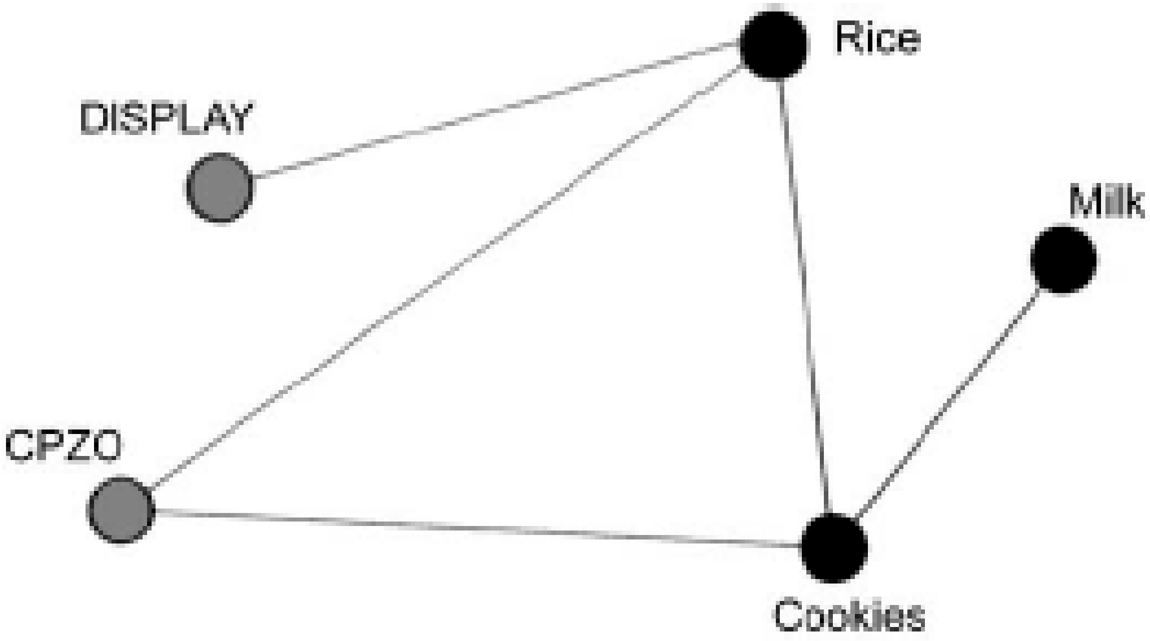
Promotional variables are then added to each cluster, and check whether promotions affect sales. They analyze one cluster in the paper.

Giudici and Passerone (2002) (cont. 1).



Marginal, not a Conditional, Independence Graph.

Giudici and Passerone (2002) (cont. 2).

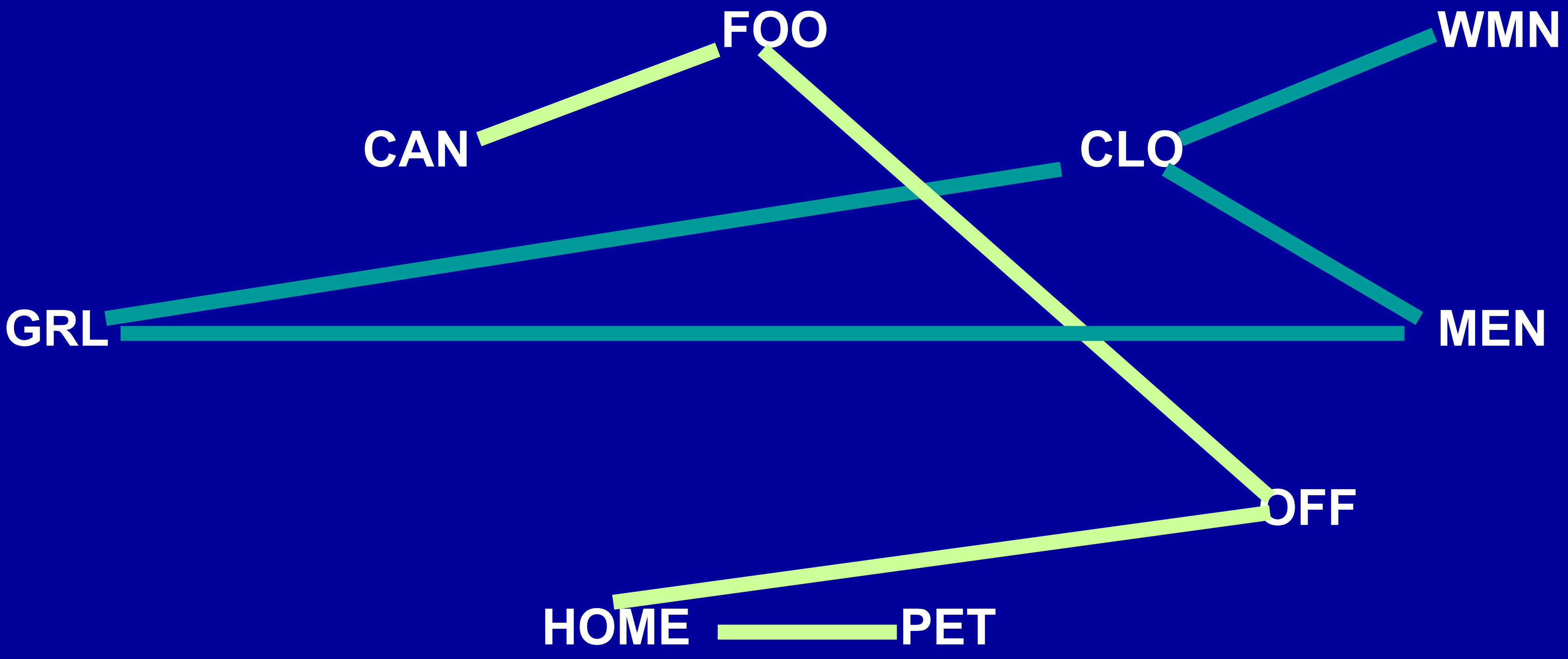


In this cluster, CPZO denotes a price reduction promotion, only indirectly affecting milk. Optimally, only one of rice and cookies should be discounted, since all three products are positively associated → sales increase of the three, estimated by the odds ratios. Rice should be on display, to further enhance sales.

Giudici and Passerone (2002) (application).

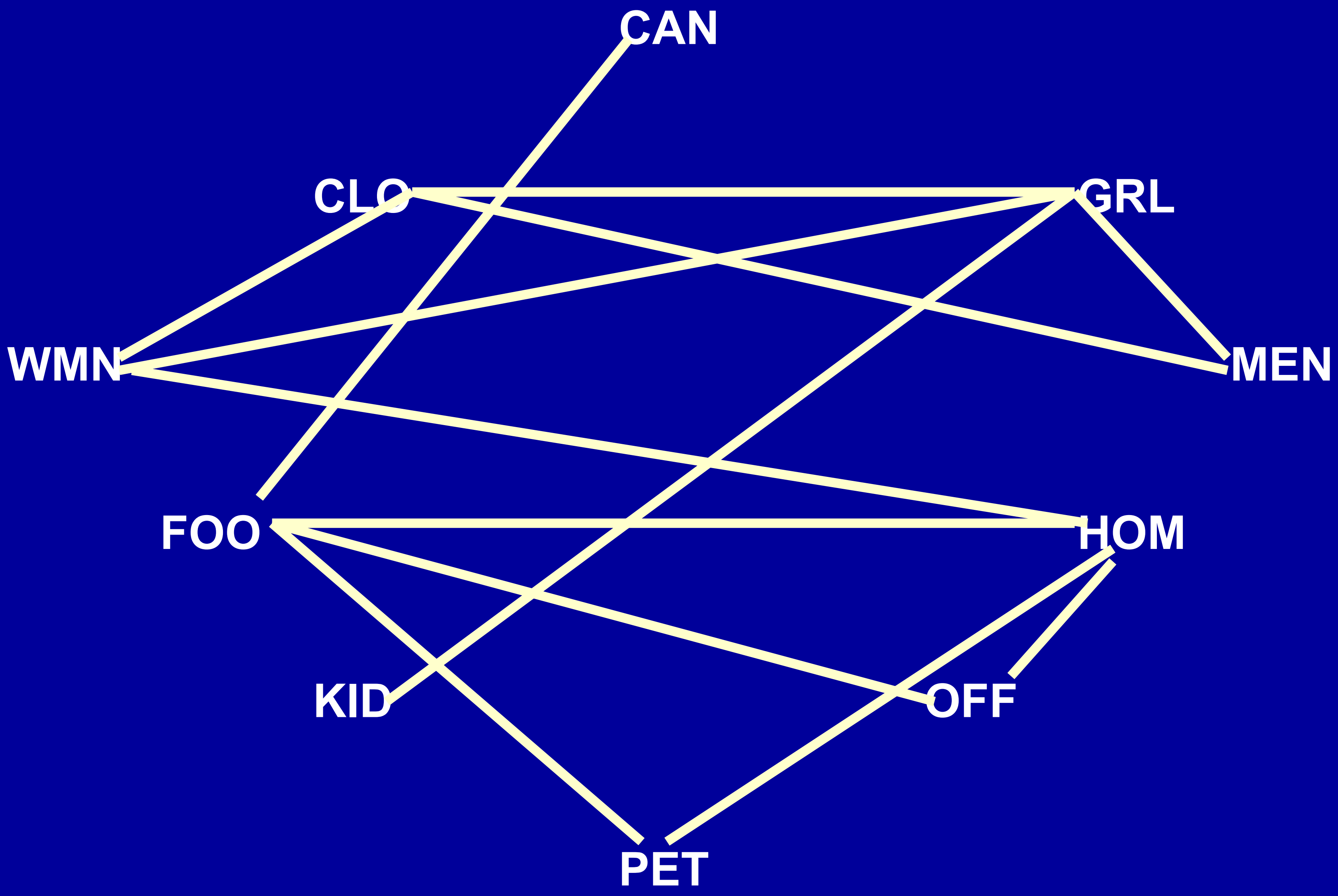
Retailer has some transaction data for 15 categories of products: CAN(dy) CLO(thing) CRA(fts) GRL WMN MEN FOO(d) FUN HLT(health) HOM(e) HRD(hardware) JWL(jewelry) KID OFF(ice) PET.

Threshold for Odds_ratio = 3 ➔ 2 disjoint clusters (different colors).

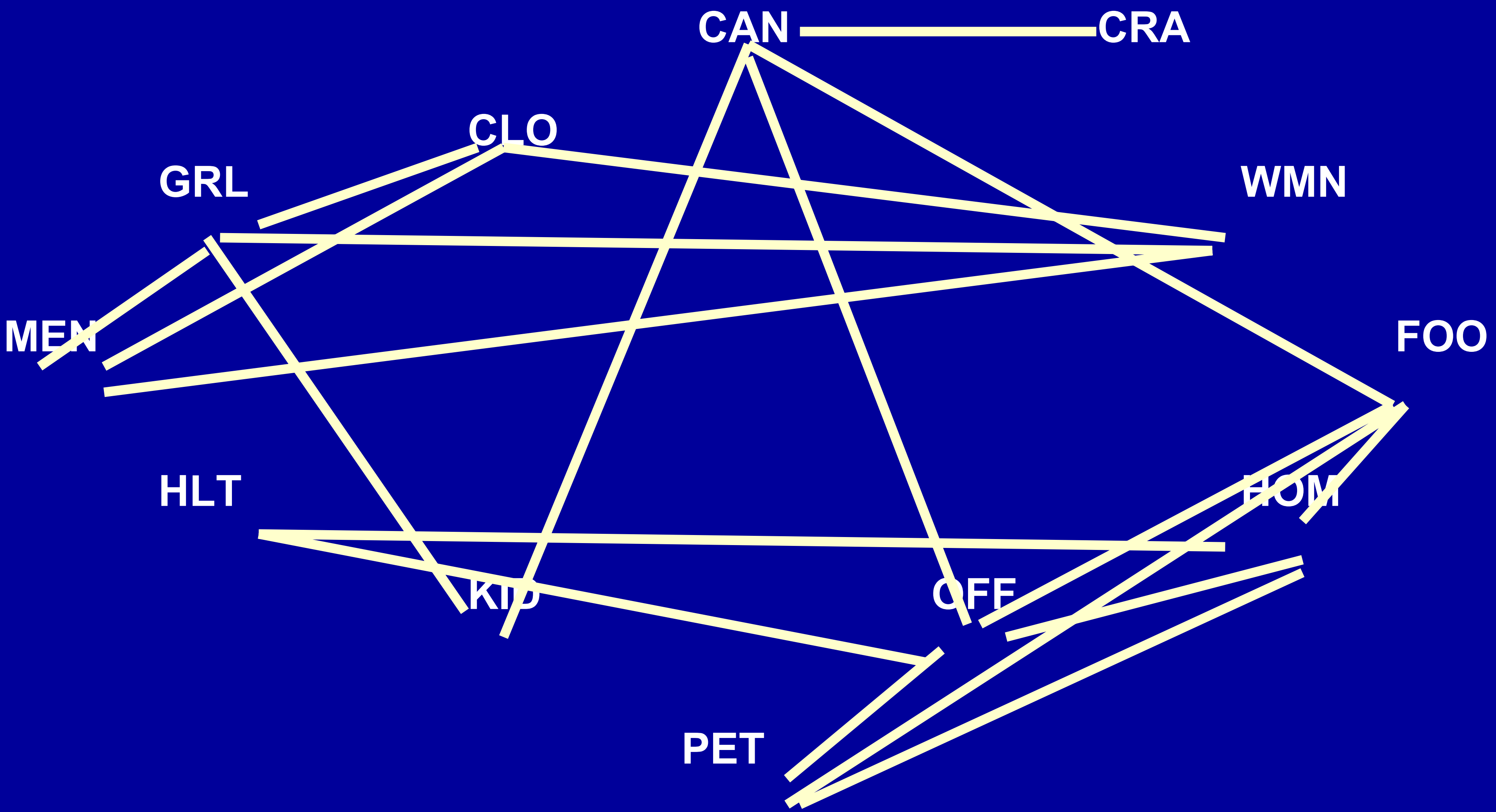


Giudici and Passerone (2002) (application cont. 1).

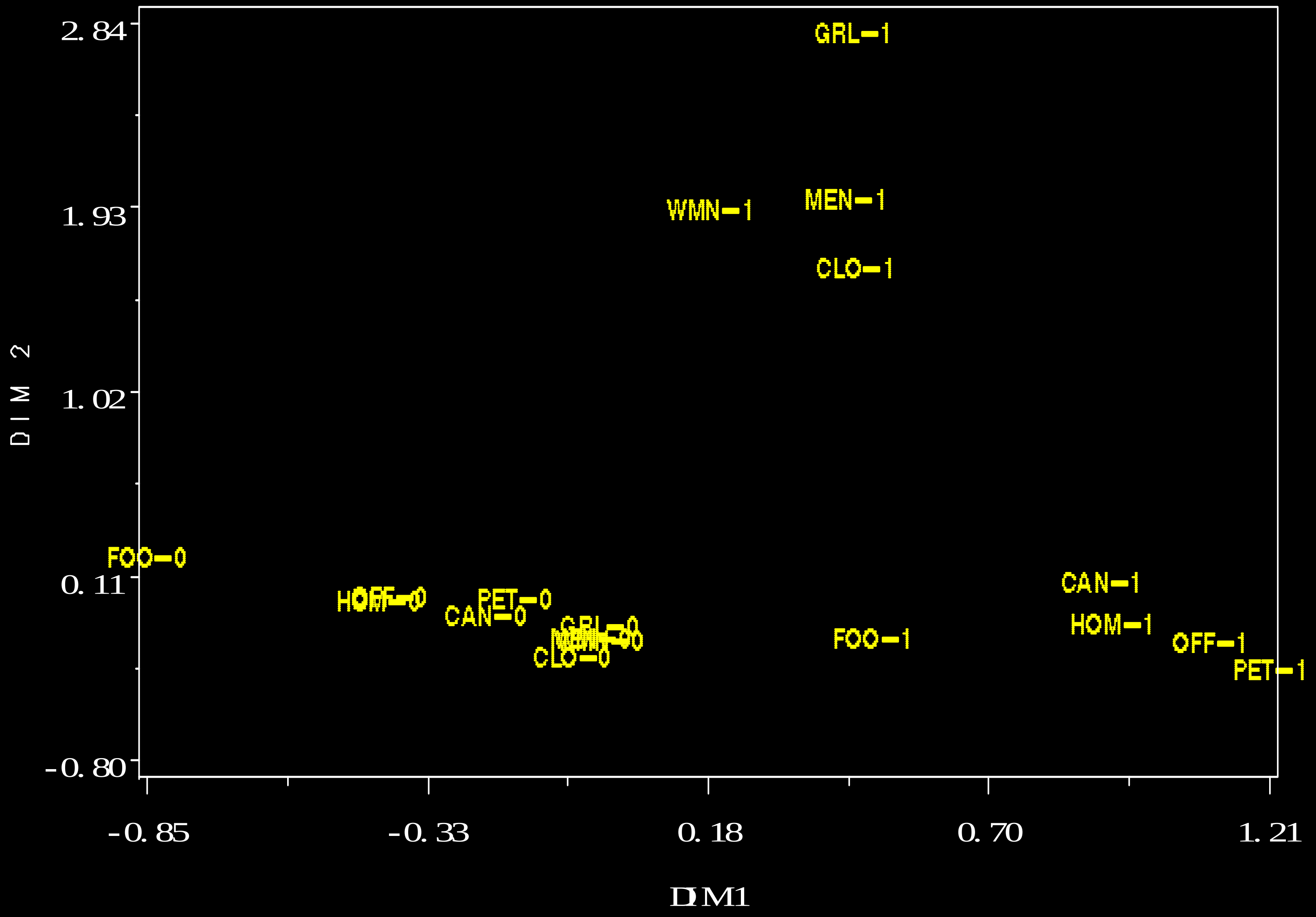
Threshold = 2.5 (13 Relations, 10 items).



Giudici and Passerone (2002) (application cont. 2). Threshold = 2 (19 relations, 12 items).



MULTIPLE CORRESPONDENCE ANAL. 1/0: PURCHASE/NO PURCH.



4.2 Sorry, No Rose Garden.

Danger: Simpson's Paradox (1951).

In general, we expect that aggregates should evince same relationships as the categories, levels or individuals over which aggregate was formed.

Simpson's paradox: Direction of association is reversed when third variable is introduced. It happens rarely in actual practice but may happen more often the more you search.

Bickel et al (1975) example: apparent gender bias in acceptances to graduate school at Berkeley in 1973. As a whole, males were accepted proportionately more than females. However, at the individual department level, reverse was happening.

Females had overall lower rate of admission because they applied disproportionately more often to departments with lowest rates of acceptance.

Danger: Simpson’s Paradox (1951) (cont. 1).

Example from Agresti (1996, p. 54).

Victim’s Race	Defendants’ Race	Death Penalty			% Yes
Z	X	YES	NO		
White	White	53	414	467	11.3
	Black	11	37	48	22.9
Black	White	0	16	16	0
	Black	4	139	143	2.8
Total	White	53	430		11.0
	Black	15	176		7.9
Grand total		68 +	606 =	674	10.0

Danger: Simpson's Paradox (1951) (cont. 2).

Let Y = death penalty X = Defendant's race Z = Victim's race (control).

Let us look at conditional probabilities, controlling for Z = W.

P(Y / Z = W & X = B) = 22.9%.
P(Y / Z = W & X = W) = 11.3%
and similarly for Z = B → controlling for Z,

P(Y = yes / X = B) = 2.8% > P (Y = yes / X = W) = 0%

Marginal probabilities instead (ignoring Z, bottom part of previous table).

P(Y = yes / X = B) = 7.9% < P (Y = yes / X = W) = 11% →

ASSOCIATION HAS REVERSED DIRECTION when data from several groups is looked at as a single one.

Why?

Danger: Simpson's Paradox (1951) (cont. 3).

1) Association between X and Y is extremely strong. Marginal odds ratio of X and Y is (467 * 143) / (48 * 16) = 87 → odds that X = W had Z = W are 87 times the odds that X = B had Z = W.

2) P (Y = yes / Z = W) = 11% > P (Y = yes / Z = B) = 7.9%

From 1) and 2), whites tend to kill whites, and white victims lead more often to death penalty → marginal association shows greater tendency for X = W to receive the death penalty than conditional association would show.

There is similar Simpson's paradox effect in context of correlations. See Hassler and Thadewald (2003/10).

NOTE: Y is "acting" as a "dependent" variable.

Danger: Simpson's Paradox (1951) (cont. 4).

Another Example. "*Ask Marilyn*"

(Parade Magazine, 28 April 1996, p 6 Marilyn vos Savant).

Company offers 70 White collar and 385 blue collar jobs.

Gender	Job Type	%Hired
Male	W	15
	B	75
Females	W	20
	B	85

200 M and 200 F applied for Job Type = W. 400 M applied for job type = B, many more than W.

Marginals:

P(Gender = F / Hired) = 42%. P(Gender = Male / Hired) = 55%.

Danger: Simpson's Paradox (1951) (cont. 5).

Another Example (derived from EdStat-L, 2/25/04).

Fast food manager restaurant argues that 75% or more of visits are drive-through. Analyst collects 50 days of data to test hypothesis and his point estimate is proportion of drive-through visits. However, his estimate may be Simpsonially affected:

	Drive-through(Y, N, %)		
Tue	30	10	.75
Wed	30	10	.75
Thu	30	10	.75
Fri	30	90	.25
Sat	30	90	.25
Sun	30	90	.25
			=======
Total	180	300	0.375

To Simpson or not to Simpson.

Be extremely CAREFUL when generalizing over aggregated data. Associations analysis provides huge amounts of support, confidence and lift measures, which are marginal and conditional probabilities.

Since practice is lift > 1 ➔ look at “interesting” confidence, ensure that Simpson is not lurking by looking at rules with previous conditioners omitted.

Thus, if A & B → C is “interesting”, look also at A ➔ C, B ➔ C., etc.

4.3

Other Methods.

Other Methods.

1) Association Tree Tool (Liu et al. 1998): creates a tree structure based on descending Confidence and prunes upwards based on misclassification rate.

2) Association Chi-Square (Auslender, 2000a and 2000b): Items deemed “dependent” per Chi-Square test become “composites” and search continues until no more composites are possible.

3) Terse Representation of AA (Auslender, 2002): Obtains regression type representation of items, thus simplifying conceptualization.

4) Naïve-Bayes: One item becomes dependent variable. NB “learns’ classification rules transactions and classifies future transactions where true nature of dependent variable is unknown. Main assumption is that explanatory items are independent.

Other Methods (cont. 1).

5) Bayesian Networks: Directed A-Cyclical Graphical Models. Items are variables that are linked by probability statements. Masuda et al (1999) performed a study linking up AA and BNs.

6) Wu et al (2003) propose log-linear model to summarize results from AA.

5. CONCLUSIONS

AND

FOOD FOR THOUGHT.

Conclusions and future development.

1) Despite ‘discovery’ claims by data miners, ridiculously extended reports of groups of items do not lend themselves easily to conceptual knowledge. Human brains are irreplaceable in this and other situations as well. Columbus was a discoverer, human software is not (see Boorstin, 1985, and very forcefully, Goodman, 2002).

Conclusions and future development (cont. 1).

2) There are many instances of items with very small support that are bypassed by traditional tools. It could well be that it is their rarity that makes these items valuable. Therefore, weighted support (and confidence), or completely different index might be necessary to identify these nuggets of information (Liu et al, 1999).

3) Alternative to searching for ‘interesting’ items is to eliminate ‘non-interesting’ ones. E.g., husband → married, orange → fruit. Sahar (1999) provides recursive algorithm for this. Of course, if we find “husband → unmarried”, either data base information needs cleaning, or else civil laws of nation have changed, or …

Conclusions and future development (cont. 2).

4) Padmanabhan and Tuzhilin (1998) created an algorithm where 'interestingness' is provided as contradiction to established beliefs. The beliefs are elicited from experts or derived from perusing the data.

5) Associations Analysis can also be understood within framework of Bayesian Networks. With on-going developments of algorithms to identify structure from data, it is possible to venture much further in simplifying the information provided. Still, present applied work has not revealed full potential of the method.

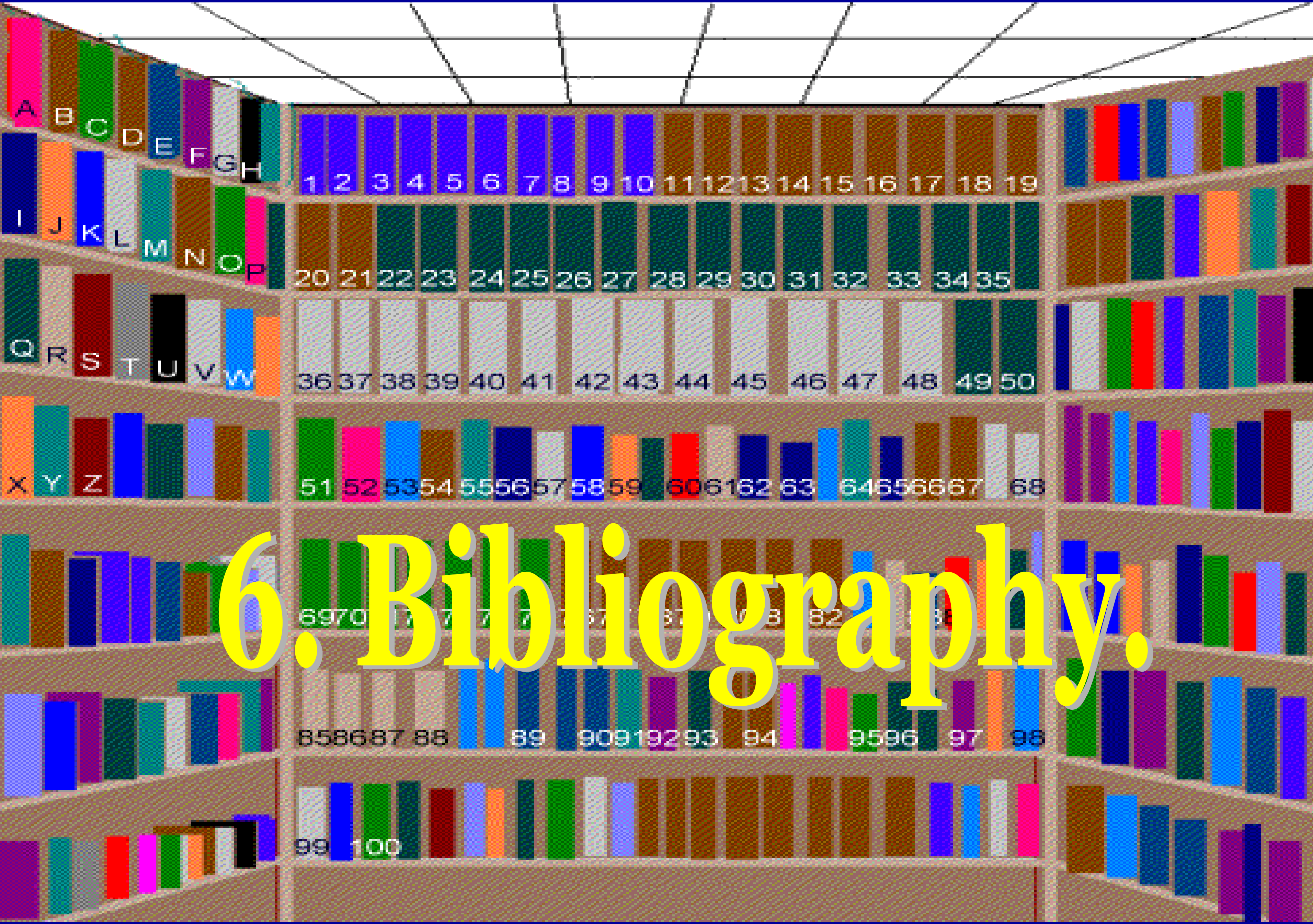
6) Giudici's and Passerone's approach is easy to investigate. Choice of magical "5" is easily modifiable.

10 Commandments of Associations Analysis.

10 Commandments of Associations Analysis.

1. Once thou hath eliminated the impossible, whatever remains, however improbable, is thy truth.
2. Thou shalt knoweth thy data relationships visually.
3. Thou shalt not gather thy networks in vain.
4. Behold thy Prior information with respect.
5. Thou shalt hanker for answers to your questions in thy data.
6. Thou shalt not covet relationships against the facts in thy data.
7. Fear not the unexpected, lest thou shalt kill unexpected and Simpson’s results in haste.
8. Thou shalt not bear false witness from thy Associations.
9. Thou shalt not stealeth stories nor maketh them up in thy data for your failure to seek the truth to the third generation of thy data set.
10. Thou shalt rejoice in thy data and rest when the truth be unveiled.

6. Bibliography.



Agrawal R., Imielinski T., and Swami A. (1993), *Mining association rules between sets of items in very large databases*, Proceedings of the ACM SIGMOD Conference on Management of data, pages 207-216, Washington D. C.

Agresti A. (1996): *An Introduction to Categorical Data Analysis*, Wiley.

Agresti A. (1998): *Applied Categorical Data Analysis*, Wiley.

Auslender, Leonardo (2000a): *On Analytical Tools for Market Basket Analysis*, Proceedings of the Diamond SAS Users Group Conference.

Auslender L. (2000b), *"On visualizing Direct and Partial Correlations – ELI plots"*, Proceedings of the Diamond-SUG Meeting, San Francisco,

Auslender L. (2002): *On terse representations of Associations Analysis*, ART Forum.

Bickel P., Hammel E., O'Connell J. (1975): *Sex bias in graduate admissions: Data from Berkeley,* Science, 187, 398-404.

Boorstin D., *The Discoverers*, Vintage Books, 1985.

Doyle A. C. (1981): *The Celebrated Cases of Sherlock Holmes,* Octopus Books Limited.

DuMouchel W. (1999): *Bayesian Data Mining in Large Frequency Tables, with an application to the FDA Spontaneous Reporting System,* The American Statistician.

Giudici P., Passerone G. (2002): *Data Mining of Association Structures to model consumer behavior,* Computational Statistics and Data Analysis, 38, 4, 533-541.

Goodman A. *Challenges, Checklists and Scorecards for Data Miners, Statisticians and Clients who fund them,* presented at the SAS 2002 Data Mining Conference, 2002.

Hassler U. and Thadewald T. (2003, October): *Nonsensical and Biased Correlation due to pooling heterogeneous samples,* The Statistician, 52, Part 3, pp. 367-379.

Liu Bing, Hsu Wynne and Ma Yiming (1998), *Integrating Classification and Association Rule Mining,* Proceedings of The Fourth International Conference on Knowledge Discovery & Data Mining, New York.

Liu B., Hsu W., Ma Y. (1999): *Mining Association Rules with multiple Minimum Supports,* Proceedings of the 1999 Knowledge Discovery & Data Mining.

MacDougall M. (2003): *Shopping for Voters: Using Association Rules to discover relationships in Election Survey Data*, Proceedings of the SAS Users Group International.

Masuda G., Yano R., Sakamoto N., Ushijima K. (1999): *Discovering and Visualizing Attribute Associations Using Bayesian networks and their use in KDD,* Proceedings of the 3rd European Conference on Principles of DMKD (PKDD-99), LNAI, vol. 1704, pp. 61-70, Springer-Verlag.

Megiddo Nimrod and Srikant Ramakrishnan (1998), *Discovering Predictive Association Rules*, Proceedings of The Fourth International Conference on Knowledge Discovery & Data Mining, New York.

Padmanabhan B., Tuzhilin A. (1998): *A belief-driven method for discovering unexpected patterns,* Proceedings of the 1998 Conference on Knowledge Discovery & Data Mining.

Rudas T. (1998): *Odds Ratios in the Analysis of Contingency Tables,* Sage.

Sahar S. (1999): *Interestingness via what is not interesting,* Proceedings of the 1999 conference on Knowledge Discovery & Data Mining.

Simpson, E. H. (1951), *The Interpretation of Interaction in Contingency Tables*, Journal of the Royal Statistical Society, Ser. B, 13, 238-241.

Spiliopoulou, M.: (1999/03): *Laborious way from data mining to web mining*, Computer Systems, Science & Engineering.

Tan P., Kumar V., Srivastava J. (2002): *Selecting the Right Interestingness Measure for Associations Patterns*, Proceedings of the 8th ACM SIGKDD International Conference on KDD, pp. 32-41, 2002.

Wu X., Barbara D., Ye Y. (2003): *Screening and Interpreting Multi-Item Associations Based on Log-Linear Modeling*, Proceedings of the 2003 KDD Conference on DMKD.

The ==> End.