WHAT CAN WE LEARN FROM DATA MINERS?

Joel H. Steckel

Prepared for 2001 UC Berkeley Invitational Choice Symposium

June 1-5, 2001

My Thesis

- Scholars in the marketing and knowledge discovery (data mining) academic communities study similar problems, but study them very differently.
- As such, each group pays a price for their orientations relative to the other.
- What is the price marketers are paying?

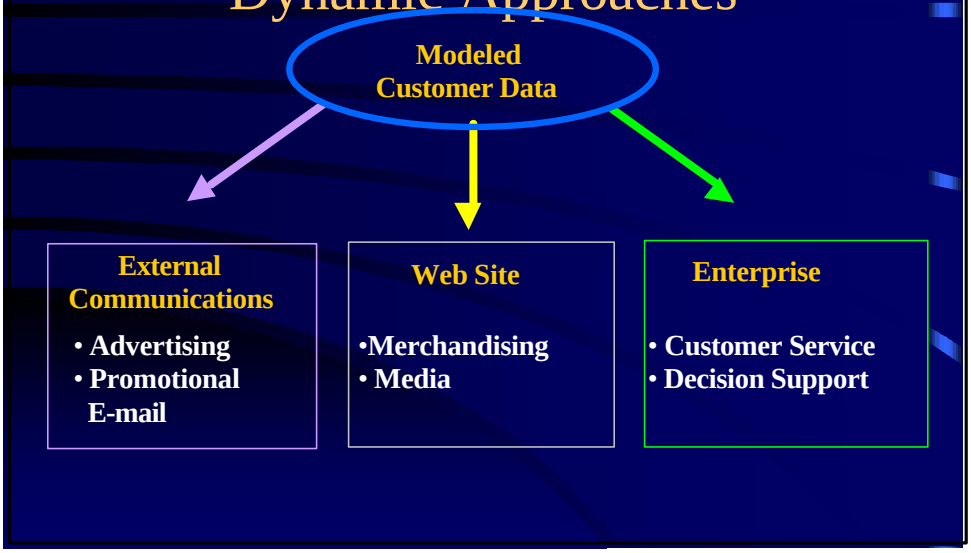
Agenda

- Background
  - Are we (marketers) missing something?
- What do Data Miners do?
  - How is it different from what marketers do?
    - Market Basket Analysis
- What Prices are Marketers Paying?
  - Amount of Data, Cleanliness of Data, Scope of Findings, Performance Metrics
- What can we Learn from Data Miners?

BACKGROUND

Are Marketers Missing Something?

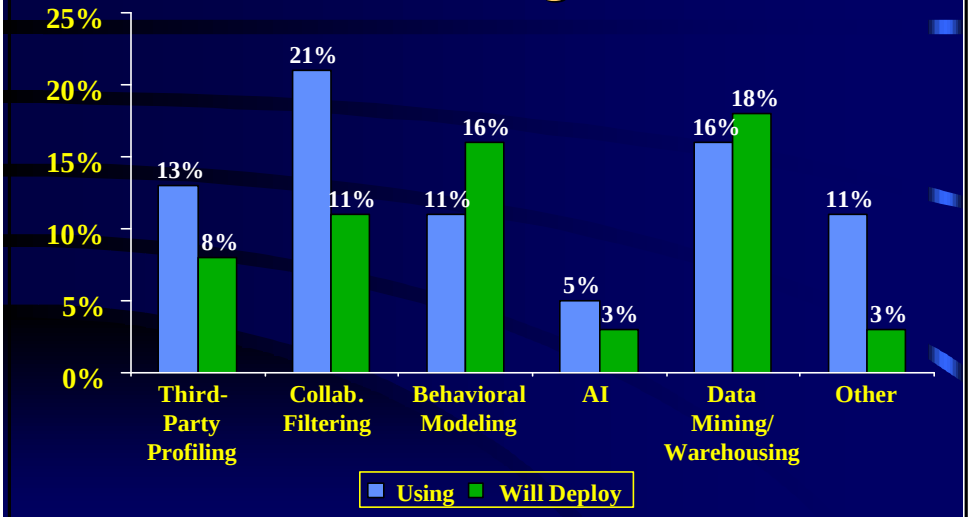
Consumer Data Feeds Myriad Dynamic Approaches



Major Classes of Personalization Applications

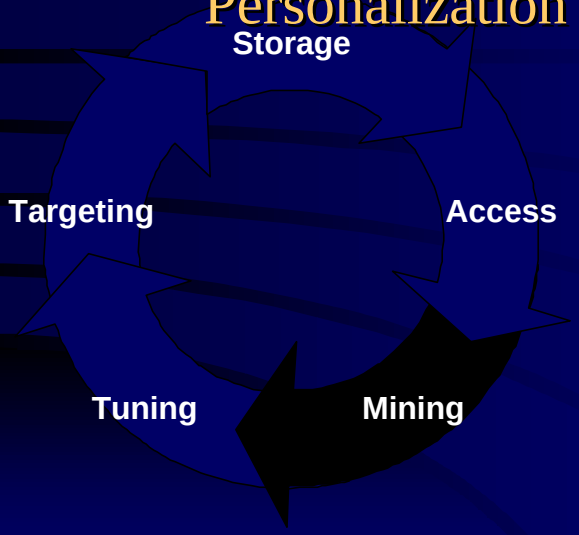
- Attraction
  - Advertising and Promotion
  - E-Mail Customization
- Retention
  - Leading Indicators of Attrition (hazard models)
- Cross Selling
  - Market Basket Analysis

Focus Is Shifting to Data Mining and Data Management



Source: Jupiter Executive Survey; N=44

Data Mining Part of Jupiter's Personalization Chain



- Behaviors, Explicit Data Matched With Business-Driven Segments

- SLP Inforware
- Broadbase
- E.Piphany
- Net Perceptions
- Etc.

WHAT DO DATA MINERS DO?

How is it Different from What Marketers Do?

Data Mining is at the Interface of

- Statistics
- Database Technology
- Pattern Recognition
- Machine Learning

Data Mining

… the process of ***inductive, computer*** analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners.

- Michael Berry & Gordon Linoff (1997), David Hand (1998)

… to enable and maximize the extraction of meaningful information from such a large database in an efficient and timely manner

- Fayad, Djorgovski, and Weir(1996)

Tasks and Tools of Data Miners

TASKS

- Classification
- Estimation
- Prediction
- Affinity Grouping
- Clustering
- Description

TOOLS

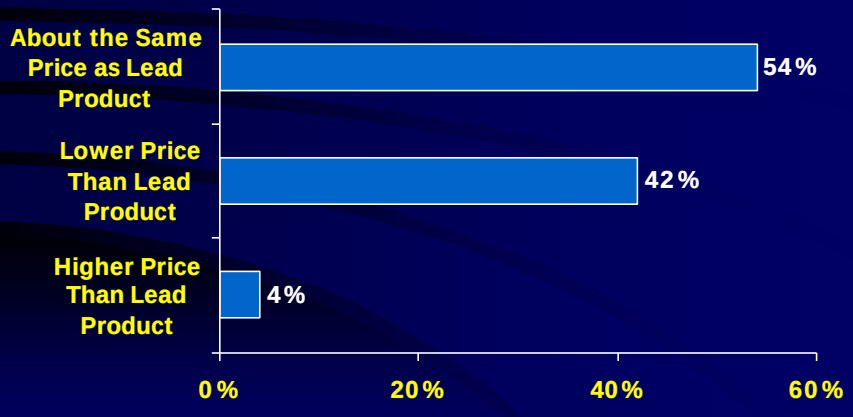
- Market Basket Analysis
- Memory Based Reasoning (MBR)
- Cluster Detection
- Link Analysis
- Decision Trees
- Neural Networks
- Genetic Algorithms

Market Basket Analysis

- Which products tend to be purchased together?
  - Items purchased on a credit card, such as rental cars and hotel rooms, give insight into the next product that customers are likely to purchase.
  - Optional services purchased by telecommunications customers (call waiting, call forwarding, etc.) help determine how to bundle them in order to maximize revenue.
  - Banking services used by retail customers (money markets, CDs, car loans) identify customers likely to want other services.
  - Medical patient histories can give indications of complications based on certain combinations of treatments.

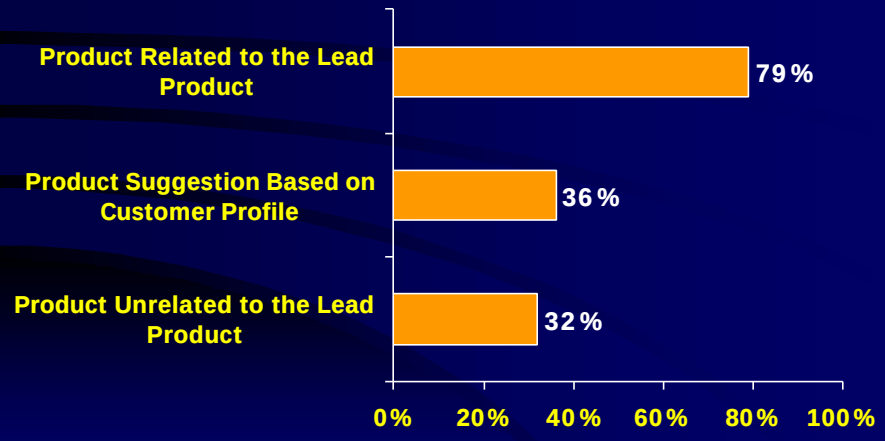
Current Focus: Increase Revenue Per Customer to Off-Set Acquisition Costs

Which of the following best describes the products that you cross-sell to the customer? The cross-sell product usually:



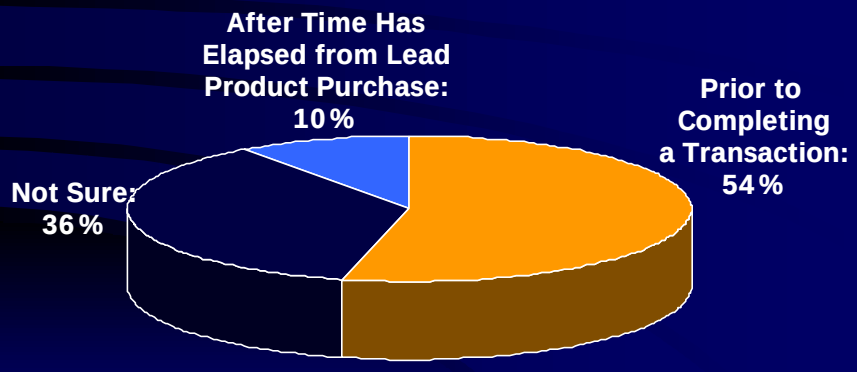
Online Cross-Selling Defined by Immediate Transaction...

Which of the following describes the types of products that you currently cross-sell online?



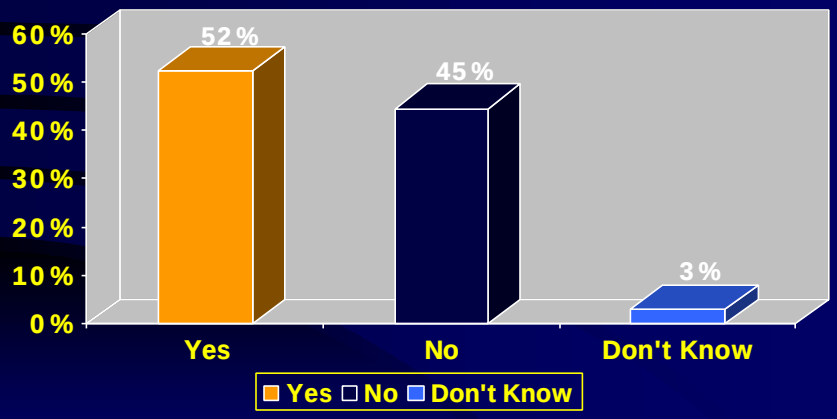
…Rather Than Developing a Long-Term Customer Relationship

In your experience, what is the most effective time to cross-sell a product to a customer?



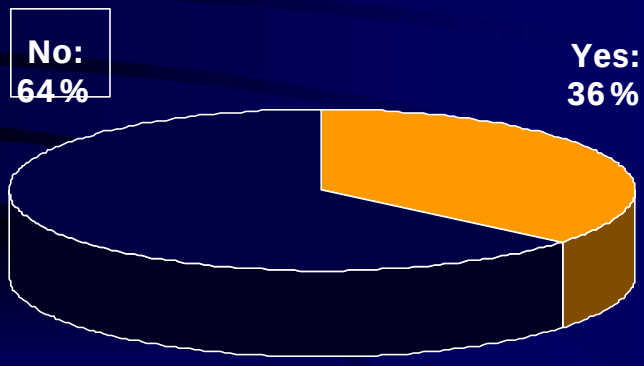
Consumers Are Primed for Cross-Sell Efforts...

Have you purchased anything online in the past 6 months that you would not have bought otherwise, had you not begun to shop over the Internet?

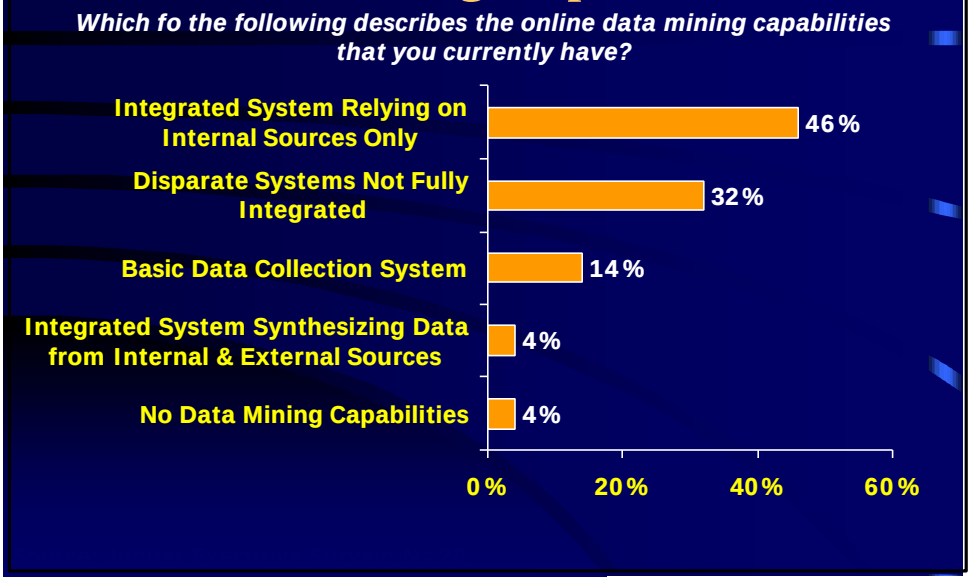


…But Under-Utilized Cross-Selling Leaves Money on the Table

In the past six months, have you purchased a product or service that a merchant suggested to you, either while or after you purchased another product or service from that same merchant?



Effective Cross-Selling Will Depend on Data Mining Capabilities



How Market Basket Analysis Might Be Performed

- Marketers/Statisticians
  - A logit model where one of the predictor variables for choice probability of one product is the prior purchase of another product. (e.g. Harlam and Lodish 1995).
- Knowledge Discovery
  - Logic based rules such as "If A and not B, then C" obtained from highly complex analyses of multidimensional co-occurence matrices.

Contrasting Approaches to Market Basket Analysis

MARKETING	KDD
Concerned with conditional probability	Concerned with conditional probability
Focused on estimating conditional probability	Focused on finding the highest ones
Uses a “random” sample of data	Use “all” the data
Talk about models	Talk about patterns
Use parametric statistical models designed for specific problems	Use non-parametric data processing algorithms that don’t have a mechanism for recognizing uncertainty
Focus on theory	Focus on prediction
Assessed by statistical criteria	Assessed by “business” criteria

WHAT PRICES DO MARKETERS PAY?

Amount of Data, Cleanliness of Data, Scope of Findings, Performance Metrics

Statistics Texts

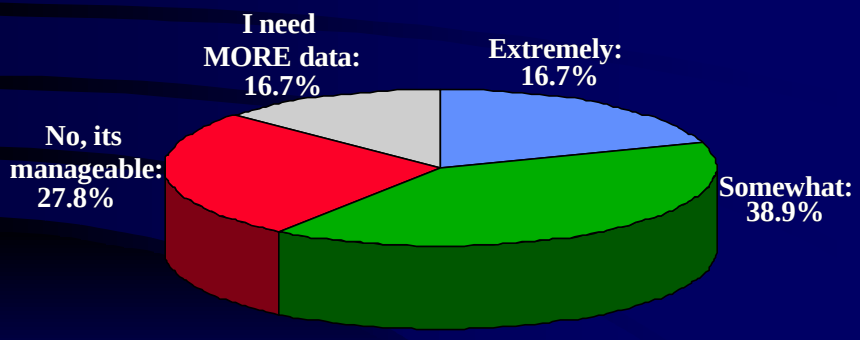
- Are characterized by data sets which are
  - Small
  - Clean
  - Sampled in an IID manner
- Permit straightforward answers to specific problems via intensive analysis of single data sets

Data Set Size

- Wal Mart makes over 30 million transactions daily
- AT&T has 100 million customers with 200 million calls daily.
- Statistical methods were developed because of data scarcity, not data abundance.

Buyers Feel Overwhelmed by the Amount of Data They Analyze

Do you feel overwhelmed by the amount of data you analyze:



Source: Jupiter Executive Survey;N=19

Contaminated Data

- Outlier detection is a major activity
- **But they may be what we want!**
  - Wealthy individuals: .1 of 1% of adults in US is 200,000 people.
  - A study of 10,000 people would only have 10
    - **ARE THESE OUTLIERS?**

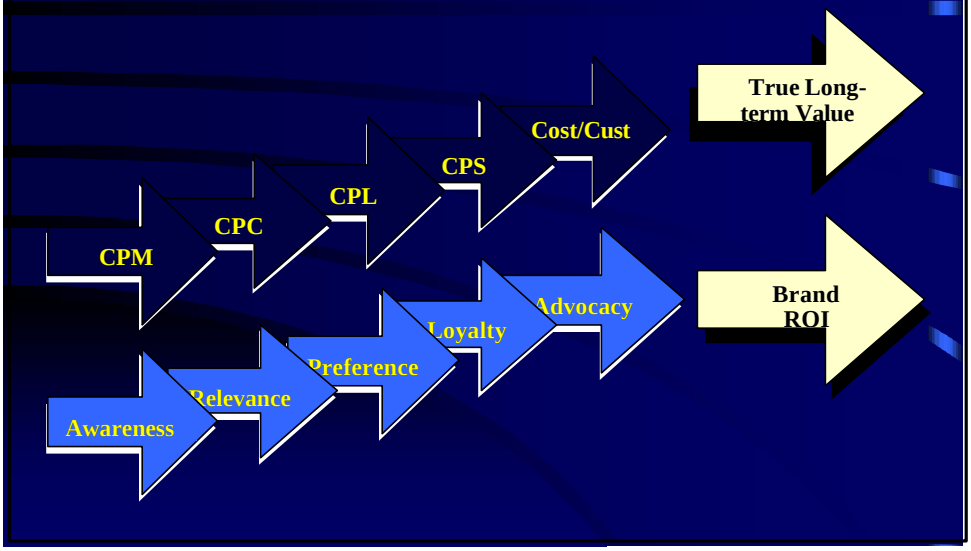
Finding Interesting Relationships

- Instructions cannot be “find interesting relationships” or “extract meaningful patterns”. Must define “INTERESTING”
  - Evidence (defined by some statistical criterion)
  - Redundancy (based on incremental knowledge)
  - Usefulness (goals of the user)
  - Novelty (deviation from prior knowledge)
  - Simplicity (MDL)
  - Generality (fraction of the population)
    - » Klosgen (1996)

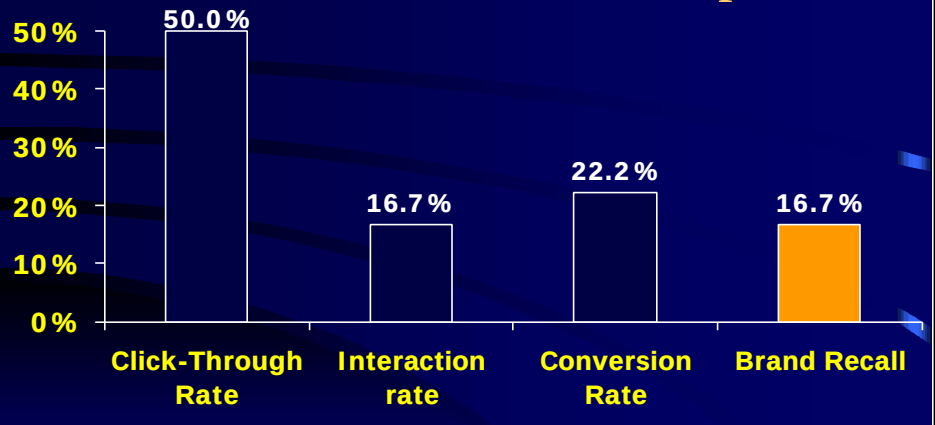
Interesting Market Basket Analysis

- Objective not to simply characterize database as conditional probabilities would.
- Make inferences to future likely co-occurences of items in a basket and ideally make (potentially) causal statements about purchase patterns.
  - ***If someone can be persuaded to buy item A then they are also likely to buy item B.***

Optimization Is Important Regardless Of Performance Metric



Click-Through Rate Still Primary Metric to Assess Brand Impact



Source: Jupiter Executive Survey,N=13

Click Rate Inaccurate Predictor of Conversion

	$ Spent	Click-through Rate	Conversion Rate	# of Conversions	Cost per Conversion
Campaign A	$50,000	14.52%	0.108%	157	$318.85
Campaign B	$50,000	1.45%	2.10%	305	$164.20

Evaluation Criteria

- Statistical
  - Goodness of Fit
  - Objective of "model" (rule?) determination
- KDD Community Measures Results
  - Lift = P ($class_t$|sample)/ P ($class_t$|population)
- ***MEASURE RESULTS, NOT MODELS***

What Can We Learn from Data Miners?

- Statistical methods may not be as crucial for large data sets. Furthermore, adherence to standard statistical methods may be counterproductive.
- Samples taken from large databases can obscure important classes
- Define “interesting”.
- Measure Results, Not Models