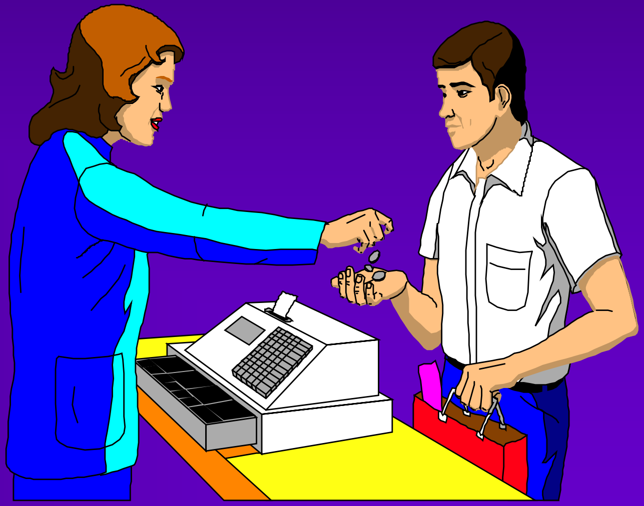
ICML Industrial Day 1999

Embedding Data Mining Technology in E-Commerce Applications

Ronny Kohavi
Director, Data Mining
Blue Martini Software

ronnyk@bluemartini.com
http://robotics.Stanford.EDU/~ronnyk/



Bodies in the Chasm

Geoffrey Moore (1995) wrote:

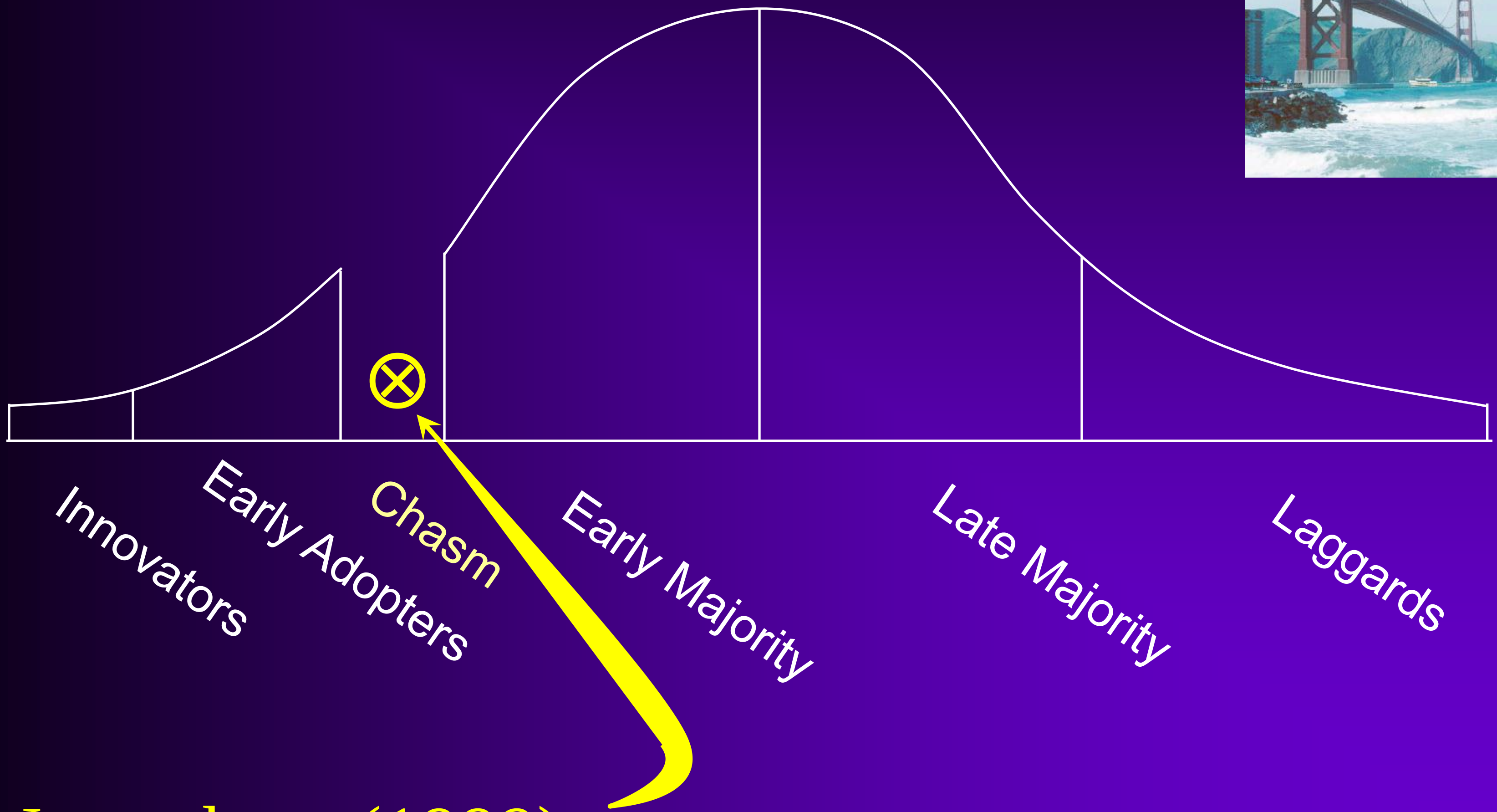
> *There were too many obstacles to its adoption…*
> *inability to integrate it easily into existing systems,*
> *no established design methodologies, and*
> *lack of people trained in how to implement it…*

What was it he was writing about?

Artificial Intelligence

In Crossing the Chasm, p. 23

Technology Adoption Life Cycle



We are here (1999)

Vertical Solutions: the Way Out of the Chasm

- Generic horizontal tools are hard to sell:
  - Mainstream users do not understand the technology
  - Integration effort is required but no-one to run it
  - Significant additional components required

- Vertical solutions are hard to build:
  - Need people with expertise in a vertical
  - Need to build multiple systems and glue them
  - Include integration with customer’s systems

Case Study: Blue Martini Software

- Vertical solution: E-Merchandising

  Allow retailers and manufacturers to effectively sell products on the Internet

- Solution includes
  - Web store module
  - Customer management module - manage attributes
  - Product management module - manage attributes
  - Micro-Marketing module (data mining, reporting, personalization)
  - Administration (e.g., Workflow)

Value Proposition

- Company’s brand is a strategic asset. Avoid diluting it with a mediocre web store. Leverage the internet to build your brand
- Collect data (both transactions and clickstreams) for improved personalization, yielding:
  - Higher conversion rates
  - Improved loyalty
  - Effective cross-sells
  - Larger baskets
  - Transfer insight back to bricks-and-mortar stores

Experiments in the Real World

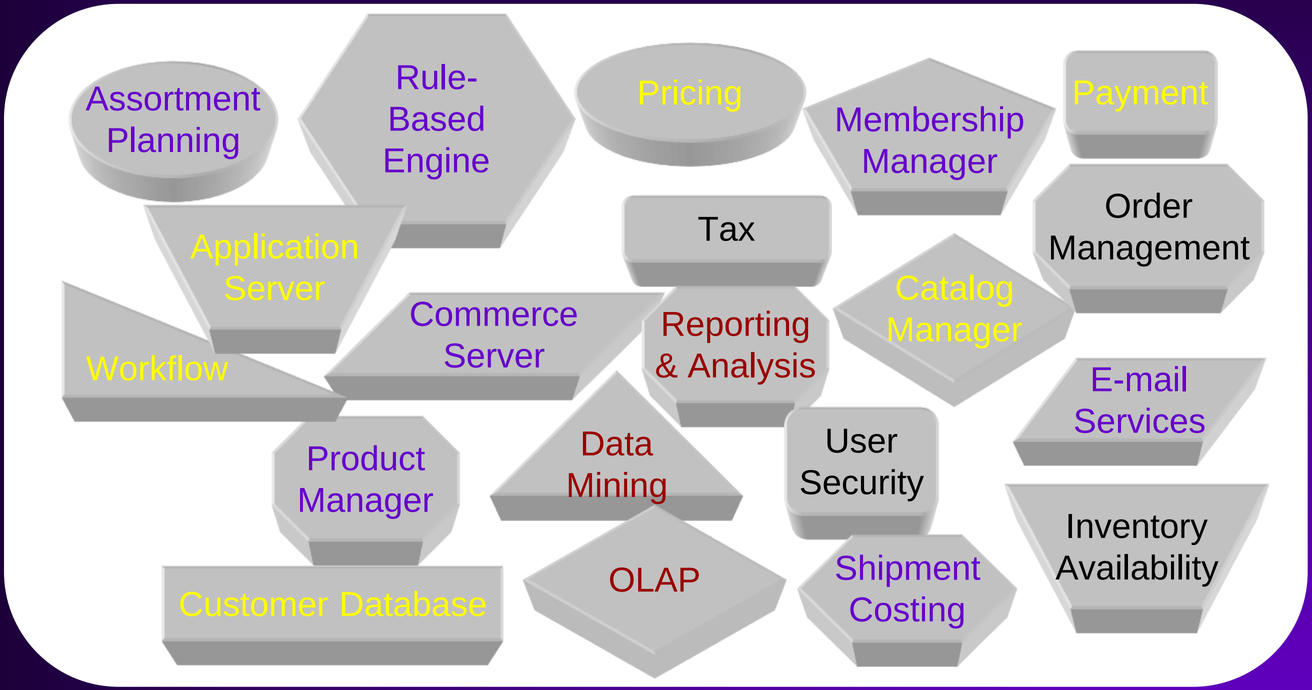
Experiments in bricks-and-mortar stores are hard. Here is a "log" from Why We Buy: the Science of Shopping:

> She's in the bath section. She's touching towels. Mark this down -- she's petted one, two, three, four of them so far. She just checked the price tag on one. Mark that down, too. Careful, her head's coming up -- blend into the aisle. She's picking up two towels from the tabletop display and is leaving the section with them. Get the time. Now, tail her into the aisle and on to her next stop.

EnviroSell Inc. goes through 14,000 hours of store videotapes a year to do behavioral research.

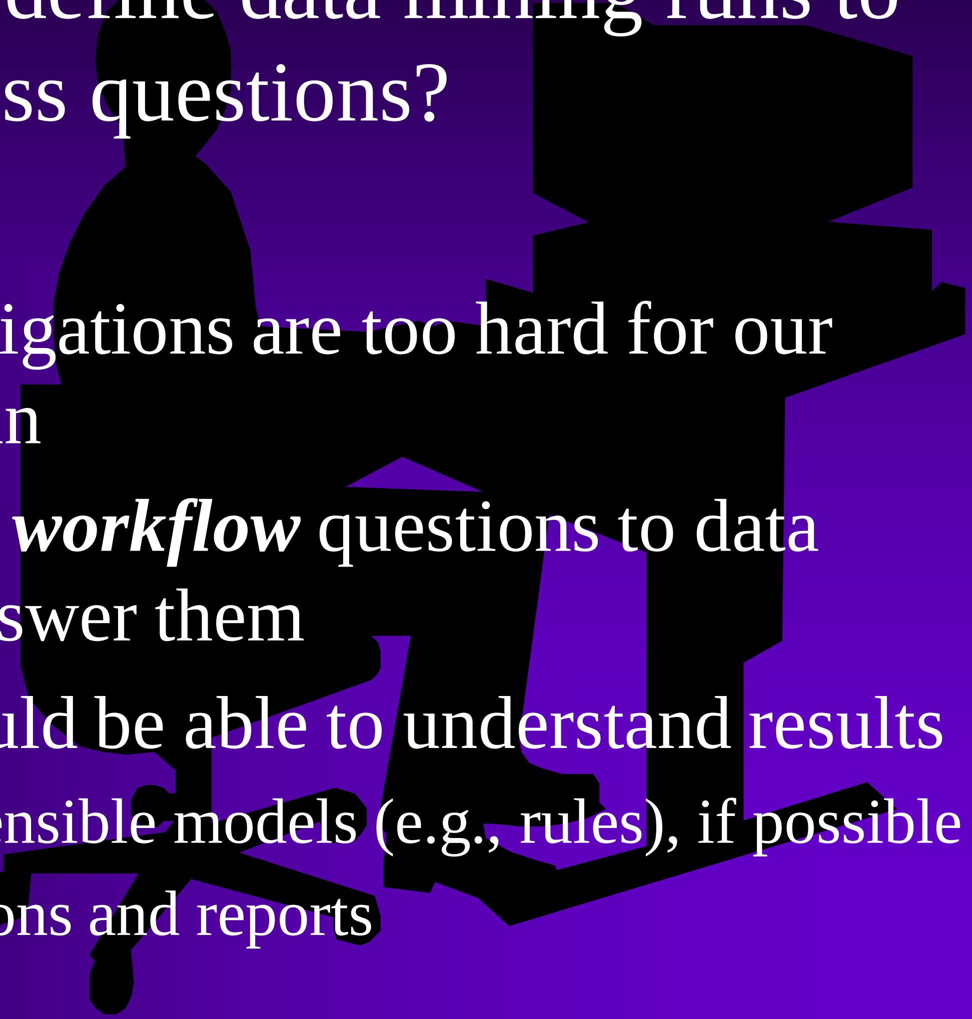
The web changes everything: clickstreams

Problem: Complex System



- Multiple components from multiple vendors
  Need significant "glue" work in the white spaces
- Data Mining is just one piece of the puzzle

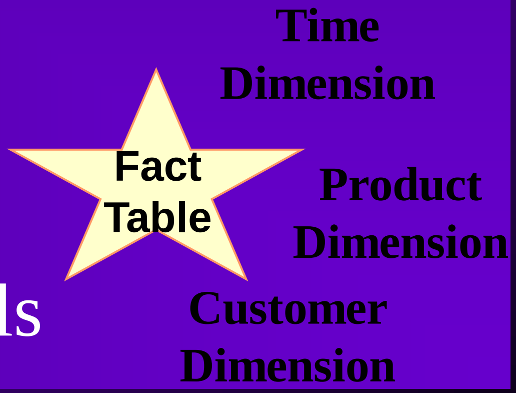
Problem: Who is the User?



- Can business users define data mining runs to answer their business questions?
- Answer:
  - Data Mining investigations are too hard for our business users to run
  - Business users will ***workflow*** questions to data miners who will answer them
  - Business users should be able to understand results
    - Generate comprehensible models (e.g., rules), if possible
    - Provide visualizations and reports

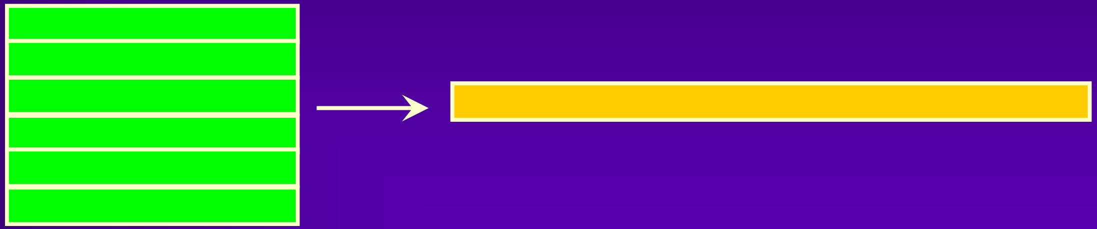
Issue: Web Store vs. Data Warehouse

- The Web Store is an On-Line Transaction Processing system (OLTP).
  Analysis should be done on a different system
- Solution:
  - Provide support for transferring the transactional data (normalized data) to a data warehouse (denormalized) using ***star schemas***
    - Bulk transfers with joins
    - Transfer meta data
  - Update store with scores from models



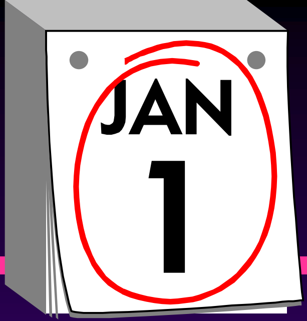
Problem: Customer Signature

- Data Mining algorithms assume records are independently and identically distributed (i.i.d)
- Need to summarize transactions/clickstreams into one record



- Solutions:
  - Provide aggregation/rollup operations.
    - Avg/min/max for numeric values (e.g., transaction price)
    - Count/percentages for values of discrete values (credit card brand)
  - Provide powerful expression language

Problem: Dates

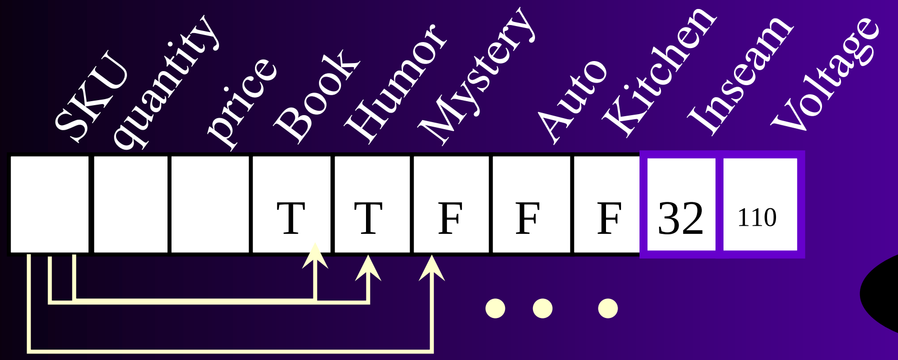


- Dates are ***very*** important, yet most data mining algorithms do not support them well

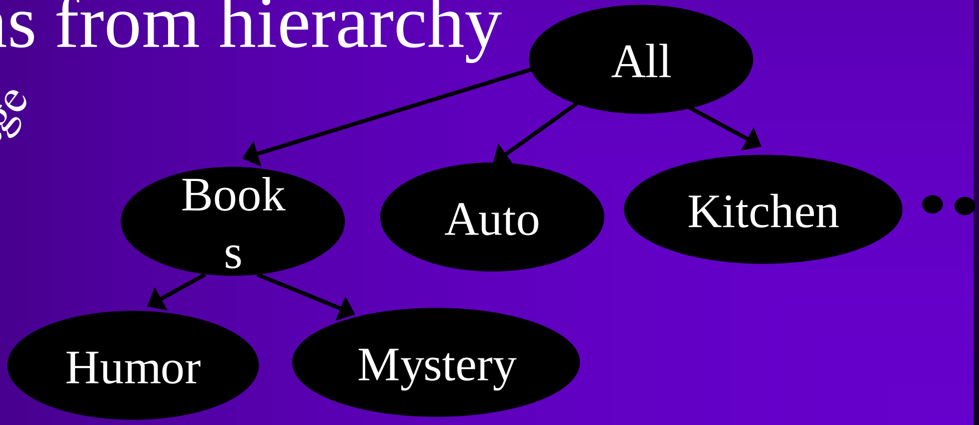
- Solution:
  - Provide well-used measurements in industry, such as Recency and Frequency (of RFM).
  - Provide strong support for date operations (days between dates, day-of-week, etc).

Product Hierarchies

- Products are typically arranged in a hierarchy. Most algorithms expect same-size records
- Solution:
  - Flatten product attributes (lots of nulls).
  - Allow users to choose parts of hierarchy for pivots based on product id (SKU).
    Add Boolean columns from hierarchy



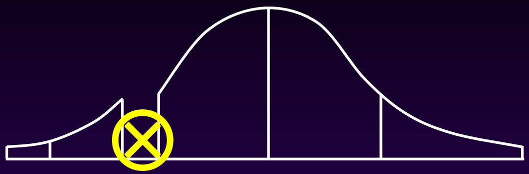
SKU	quantity	price	Book	Humor	Mystery	Auto	Kitchen	Inseam	Voltage
			T	T	F	F	F	32	110



Machine Learning Algorithms

- Problem: data mining vendors are shrinking
  - Nov 98: DataMind changes to a vertical solution provider (1-1 marketing) as RightPoint.
  - Nov 98: Gentia acquired Compression Sciences' K.wiz
  - Dec 98: Yahoo acquired HyperParallel
  - Jan 99: SPSS acquired ISL Clementine
  - June 99: Oracle acquired Thinking Machines’ Darwin
  - June 99: Unica announced move to marketing automation
- Few vendors are setup for OEM relationships
- Solution: mix of build (e.g. transformations) and buy (e.g., C5.0)

Summary (1 of 2)



- Data Mining/Machine Learning is a technology
- Data mining needs to be used by business people, who care about their vertical application
- To make it simpler and usable, it needs to be integrated into solutions, requiring people with diverse backgrounds in different areas
- E-commerce is a great source of reliable data, so the combination with DM makes great sense

Summary (2 of 2)

Important areas for research include:

- Generating insight through comprehensible models, visualization, and filtering techniques.
- Better transactional data handling, not necessarily forcing transformations into customer signatures
- Better support for data types: dates, nulls, multimedia
- Support for large hierarchical attributes
- Post mining integration (scoring, acting, validating)
- The usual: scalable anytime algorithms, use meta data, use of star schemas, and non-propositional models.

Some images used herein where obtained from IMSI's MasterClips/Master Photo(C) Collection,
1895 Francisco Blvd East, San Rafael 94901-5506, USA