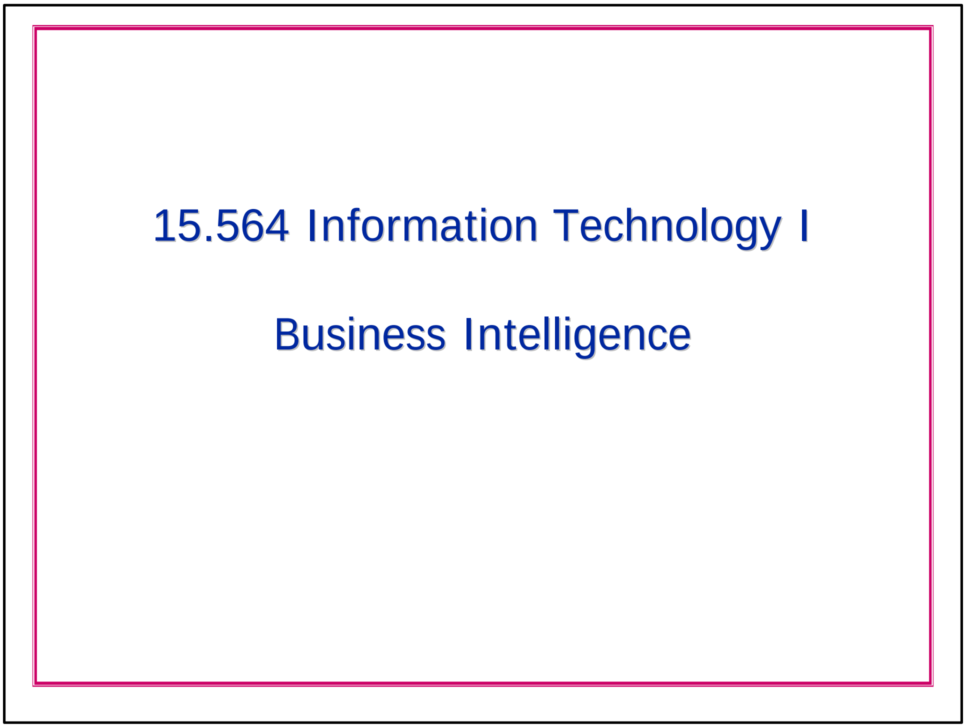
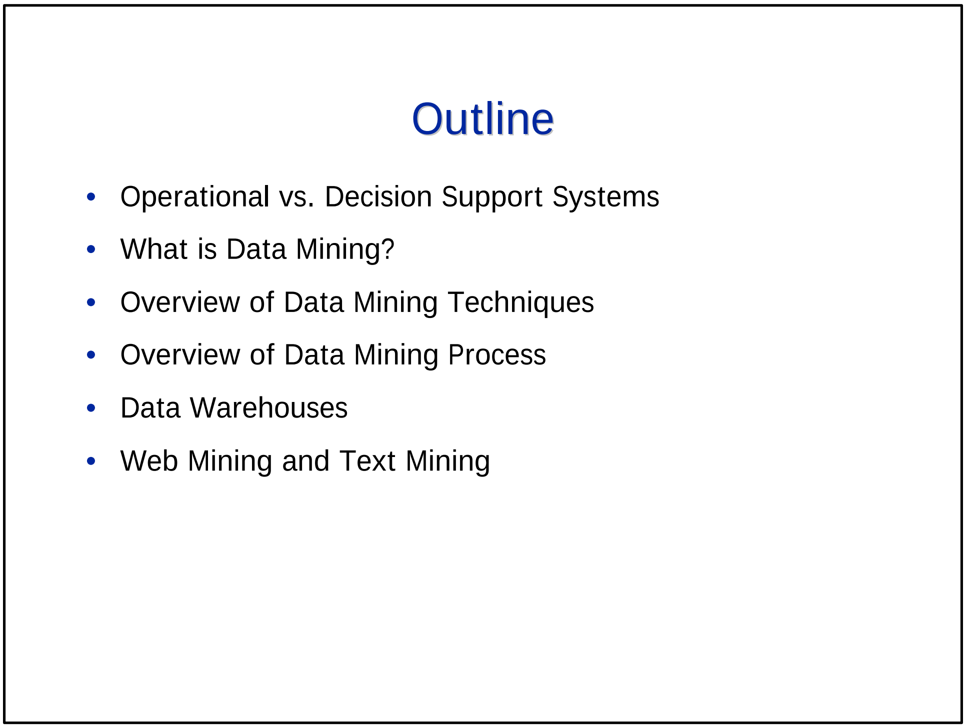


15.564 Information Technology I

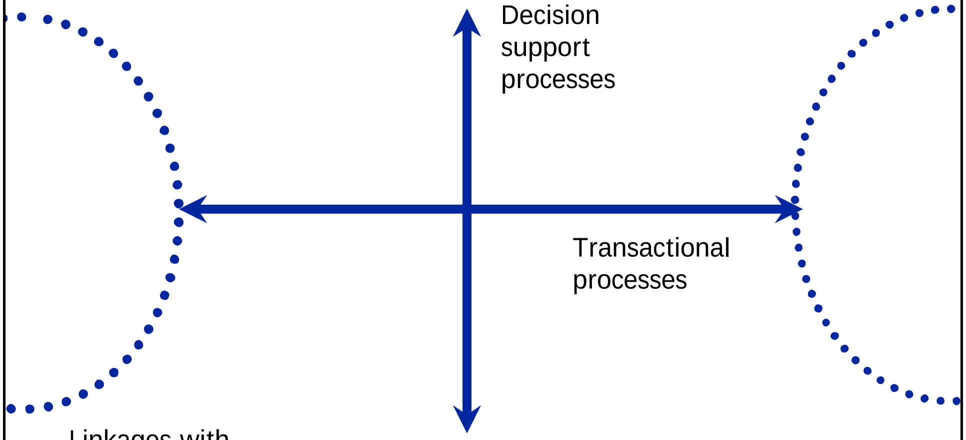
Business Intelligence



Outline

- Operational vs. Decision Support Systems
- What is Data Mining?
- Overview of Data Mining Techniques
- Overview of Data Mining Process
- Data Warehouses
- Web Mining and Text Mining

Operational vs. Decision Support Systems

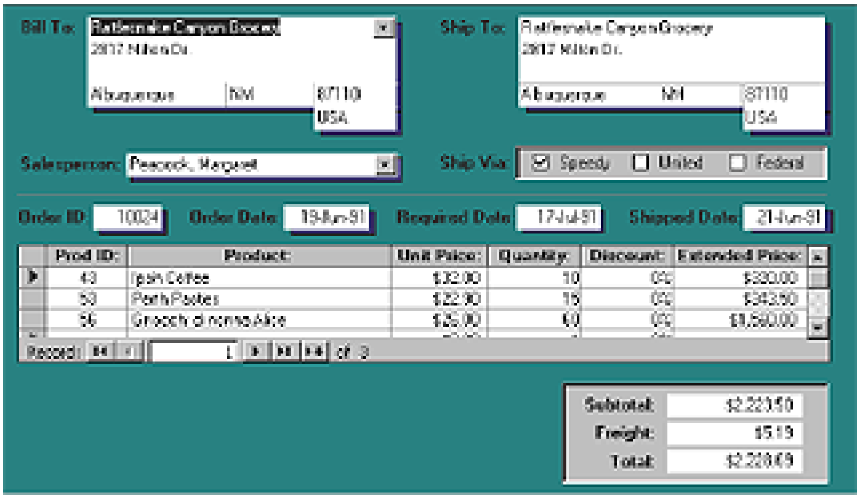


Linkages with
Suppliers, customers, partners

Operational vs. Decision Support Systems

- Operational Systems
  - Support day to day transactions
  - Contain current, “up to date” data
  - Examples: customer orders, inventory levels, bank account balances
- Decision Support Systems
  - Support strategic decision making
  - Contain historical, “summarized” data
  - Examples: performance summary, customer profitability, market segmentation

Example of an Op. Ap.: Order Entry



Bill To: Rattlesnake Canyon Grocery
2817 Milton Dr.
Albuquerque NM 87110
USA

Ship To: Rattlesnake Canyon Grocery
2817 Milton Dr.
Albuquerque NM 87110
USA

Salesperson: Peacock, Margaret

Ship Via: ☑ Speedy ☐ United ☐ Federal

Order ID: 10024 Order Date: 19-Jun-91 Required Date: 17-Jul-91 Shipped Date: 21-Jun-91

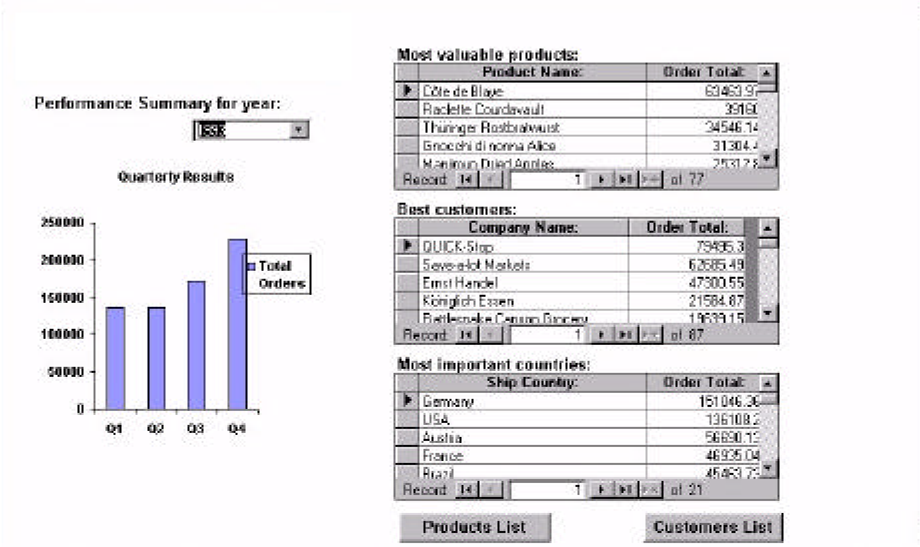
Prod ID:	Product:	Unit Price:	Quantity:	Discount:	Extended Price:
43	Ipoh Coffee	$32.00	10	0%	$320.00
53	Perth Pasties	$22.90	15	0%	$343.50
56	Gnocchi di nonna Alice	$26.00	60	0%	$1,560.00

Record: 1 of 3

Subtotal:	$2,223.50
Freight:	$5.19
Total:	$2,228.69

This screenshot is from the Microsoft® Access® software program.

Example of a DSS ap: Annual performance summary



Performance Summary for year: [illegible]

Most valuable products:

Product Name:	Order Total:
Côte de Blaye	63463.9
Raclette Courdavault	39160
Thüringer Rostbratwurst	34546.14
Gnocchi di nonna Alice	31304.4
Manjimup Dried Apples	25317.8

Record: 1 of 77

Best customers:

Company Name:	Order Total:
QUICK-Stop	79495.3
Save-a-lot Markets	62685.49
Ernst Handel	47300.55
Königlich Essen	21584.87
Rattlesnake Canyon Grocery	19639.15

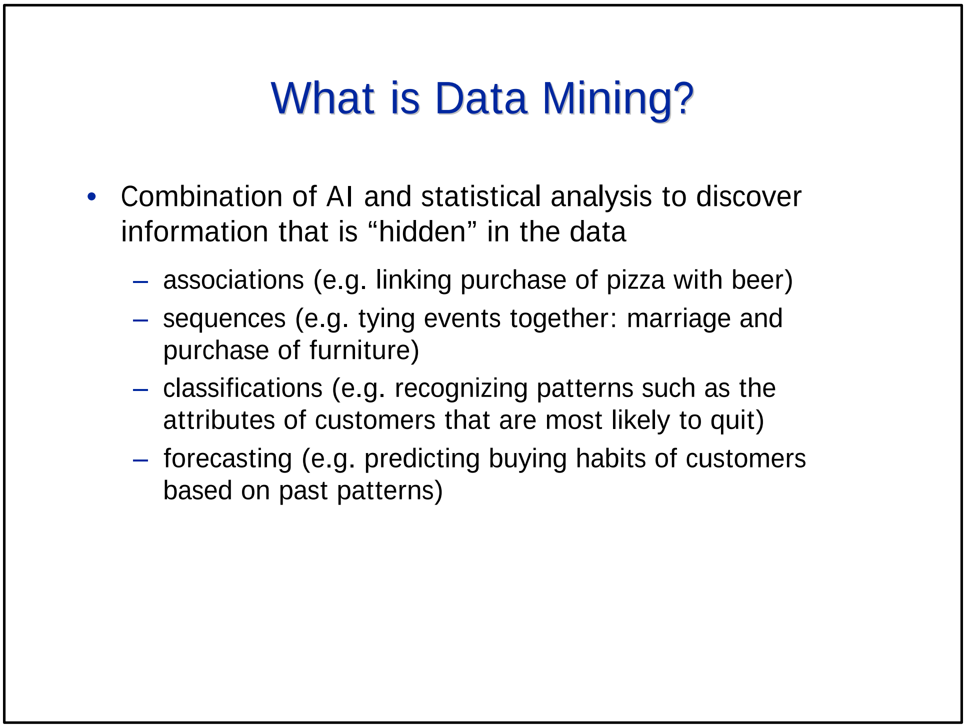
Record: 1 of 87

Most important countries:

Ship Country:	Order Total:
Germany	151046.36
USA	136108.2
Austria	56690.13
France	46935.04
Brazil	45463.73

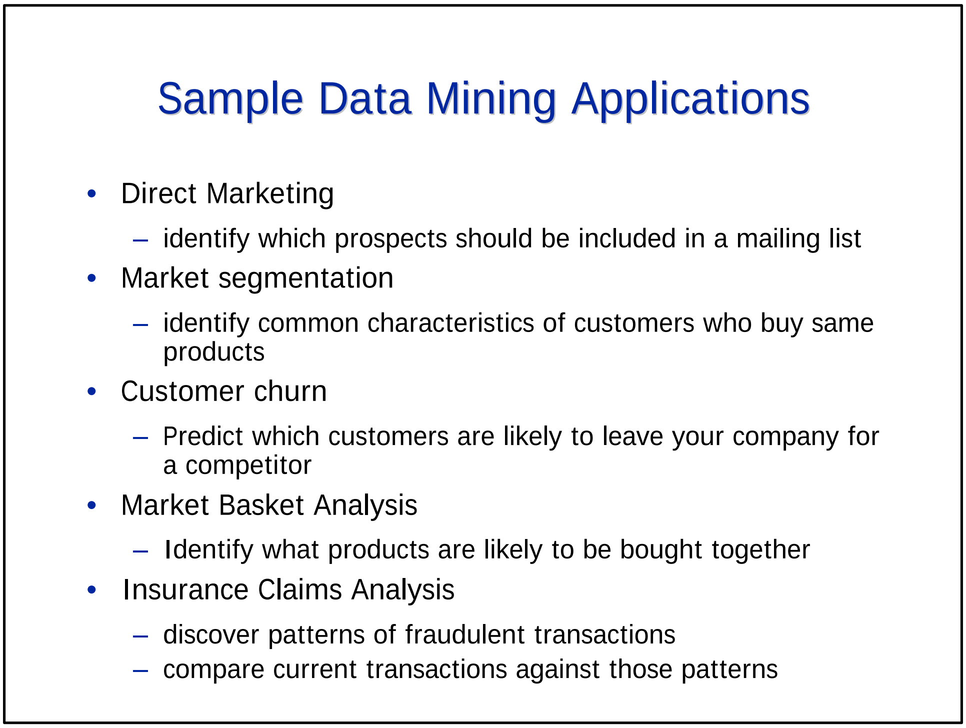
Record: 1 of 21

This screenshot is from the Microsoft® Access® software program.



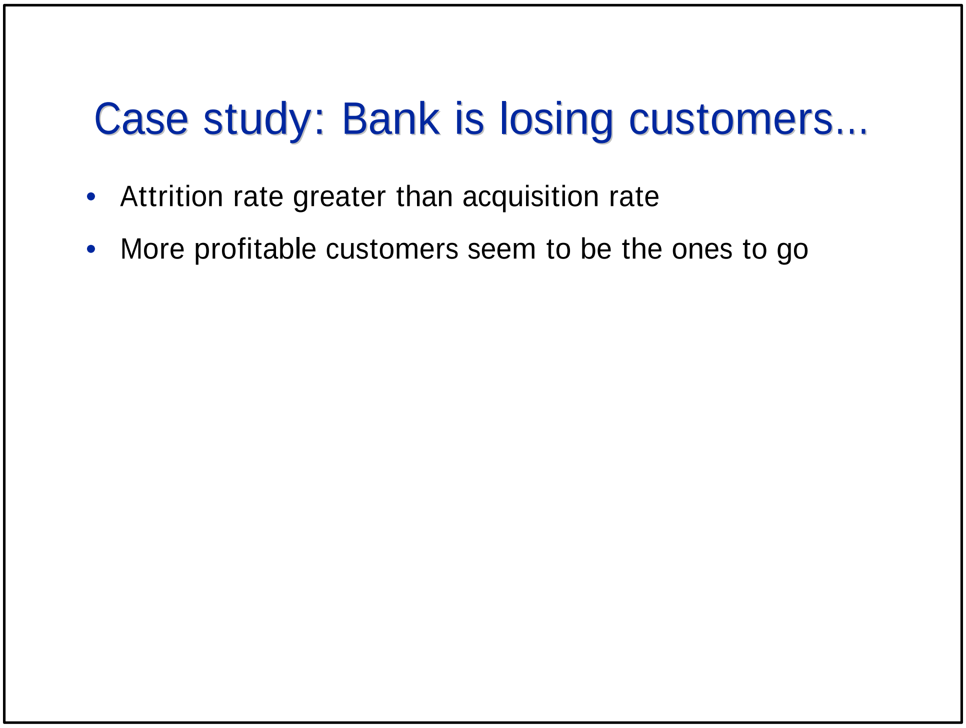
What is Data Mining?

- Combination of AI and statistical analysis to discover information that is “hidden” in the data
  - associations (e.g. linking purchase of pizza with beer)
  - sequences (e.g. tying events together: marriage and purchase of furniture)
  - classifications (e.g. recognizing patterns such as the attributes of customers that are most likely to quit)
  - forecasting (e.g. predicting buying habits of customers based on past patterns)



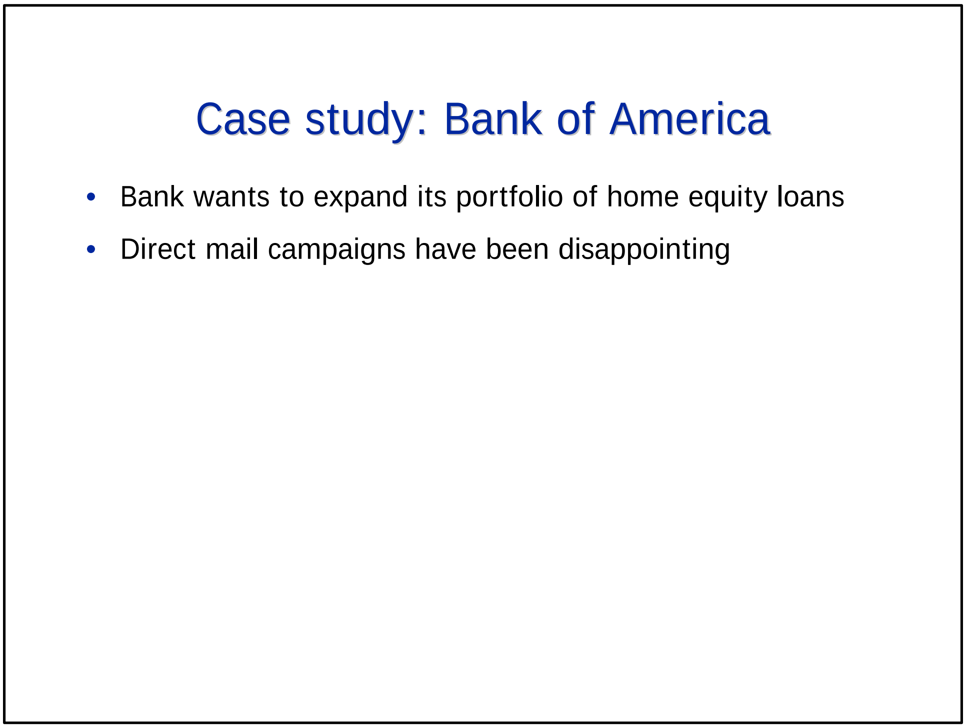
Sample Data Mining Applications

- Direct Marketing
  - identify which prospects should be included in a mailing list
- Market segmentation
  - identify common characteristics of customers who buy same products
- Customer churn
  - Predict which customers are likely to leave your company for a competitor
- Market Basket Analysis
  - Identify what products are likely to be bought together
- Insurance Claims Analysis
  - discover patterns of fraudulent transactions
  - compare current transactions against those patterns



Case study: Bank is losing customers...

- Attrition rate greater than acquisition rate
- More profitable customers seem to be the ones to go



Case study: Bank of America

- Bank wants to expand its portfolio of home equity loans
- Direct mail campaigns have been disappointing

The “Virtuous Circle” of Data Mining

- Identify the business problem
- Use data mining to transform “the data” into actionable information
    - What is the “right” data and where do we get it from?
    - What are the “right” techniques?
- Act on the information
- Measure the results

Business uses of data mining

Essentially six tasks...

- Classification
    - Classify credit applicants as low, medium, high risk
    - Classify insurance claims as normal, suspicious
- Estimation
    - Estimate the probability of a direct mailing response
    - Estimate the lifetime value of a customer
- Prediction
    - Predict which customers will leave within six months
    - Predict the size of the balance that will be transferred by a credit card prospect

Business uses of data mining

- Affinity Grouping
  - Find out items customers are likely to buy together
  - Find out what books to recommend to Amazon.com users
- Clustering
  - Difference from classification: classes are unknown!
- Description
  - Help understand large volumes of data by uncovering interesting patterns

Overview of Data Mining Techniques

- Market Basket Analysis
- Memory-Based Processing (Collaborative Filtering)
- Automatic Clustering
- Decision Trees and Rule Induction
- Neural Networks

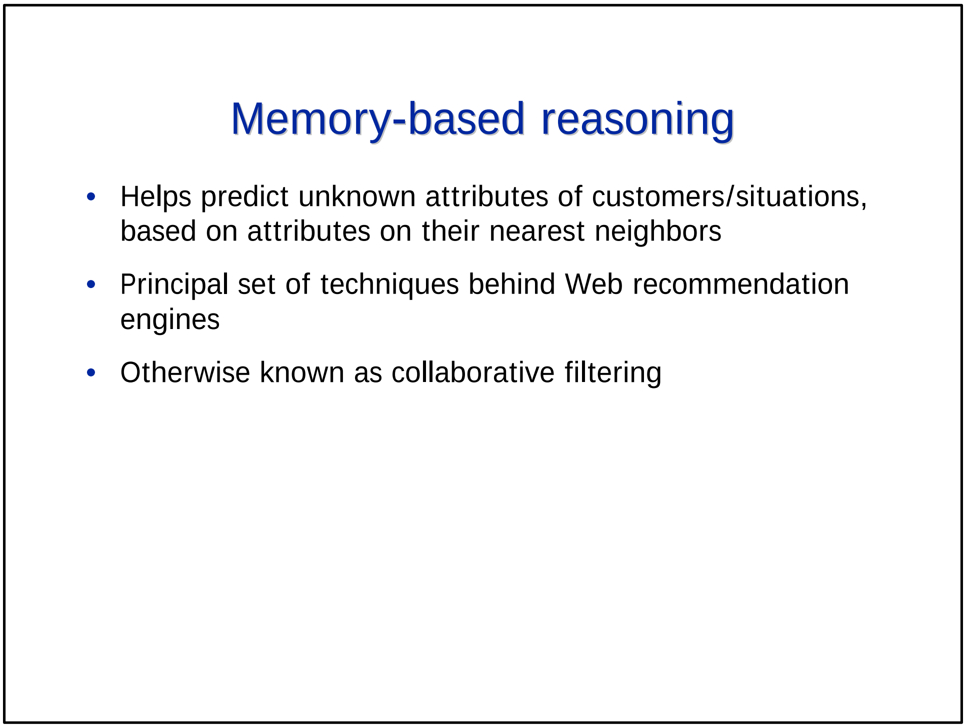
Market Basket Analysis

- Association and sequence discovery
- Principal concepts
  - Support or Prevalence: frequency that a particular association appears in the database
  - Confidence: conditional predictability of B, given A
- Example:
  - Total daily transactions: 1,000
  - Number which include “soda”: 500
  - Number which include “orange juice”: 800
  - Number which include “soda” and “orange juice”: 450
  - SUPPORT for “soda and orange juice” = 45% (450/1,000)
  - CONFIDENCE of “soda → orange juice” = 90% (450/500)
  - CONFIDENCE of “orange juice → soda” = 56% (450/800)

Applying Market Basket Analysis

- Create co-occurrence matrix
  - What is the right set of items???
- Generate useful rules
  - Weed out the trivial and the inexplicable from the useful
- Figure out how to act on them

- Similar techniques can be applied to time series for mining useful *sequences* of actions



Memory-based reasoning

- Helps predict unknown attributes of customers/situations, based on attributes on their nearest neighbors
- Principal set of techniques behind Web recommendation engines
- Otherwise known as collaborative filtering

Example: Amazon.com book recommendations

- Example: Identify books to recommend to customers
- Company keeps log of past customer purchases
- Represent each customer as a vector whose components are the past purchases
- Define a “distance” function for comparing customers
- Based on this distance function, identify the customer’s nearest neighbor set (NNS)
- Identify books that have been purchased by a large percentage of the nearest neighbor set but not by the customer
- Recommend these books to the customer as possible next purchases

(Screenshot from Amazon.com
showing book recommendations.)

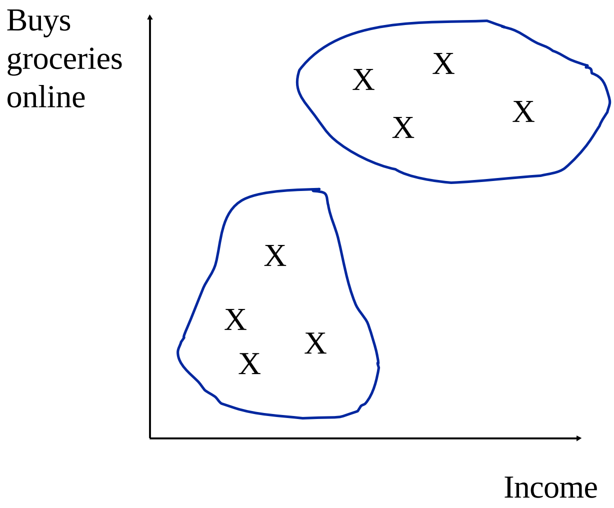
Another example: Personalized restaurant recommendations

- Alice is asking Zagat.com for a personalized rating of “Border Cafe”
- Alice has already submitted ratings for 20 other restaurants in the past 12 months
- Zagat.com finds other members whose ratings for those 20 restaurants are similar to Alice’s
- Zagat.com calculates the average (or weighted average) rating that these nearest neighbors have given to Border Cafe
- This is Alice’s personalized rating of Border Cafe
- Note that this may be quite different from the average rating of Border Cafe based on the entire population of raters!

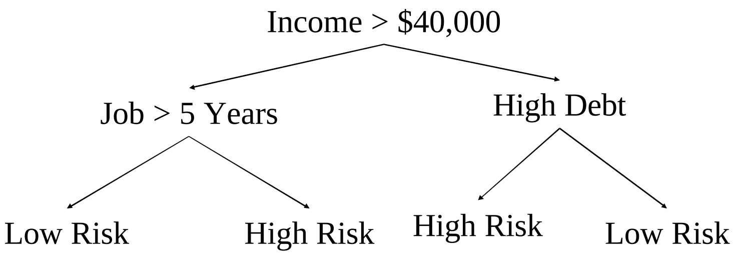
Clustering

- Divide (segment) a database into groups
- Goal: Find groups that are very different from each other, and whose members are similar to each other
- Number and attributes of these groups are *not known in advance*

Clustering (example)



Decision Trees


Income > $40,000

Job > 5 Years

High Debt

Low Risk

High Risk

High Risk

Low Risk

- Data mining is used to construct the tree
- Example algorithm: CART (Classification and Regression Trees)

Decision tree construction algorithms

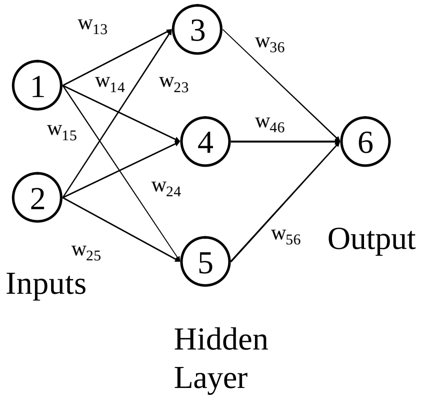
- Start with a training set (i.e. preclassified records of loan customers)
  - Each customer record contains
    - Independent variables: income, time with employer, debt
    - Dependent variable: outcome of past loan
- Find the independent variable that best splits the records into groups where one single class (low risk, high risk) predominates
  - Measure used: entropy of information (diversity)
  - Objective:
    - max[diversity before – (diversity left + diversity right)]
- Repeat recursively to generate lower levels of tree

Decision Tree pros and cons

- Pros
  - One of the most intuitive techniques, people really like decision trees
  - Really helps get some intuition as to what is going on
  - Can lead to direct actions/decision procedures
- Cons
  - Independent variables are not always the best separators
  - Maybe some of them are correlated/redundant
  - Maybe the best splitter is a linear combination of those variables (remember factor analysis)

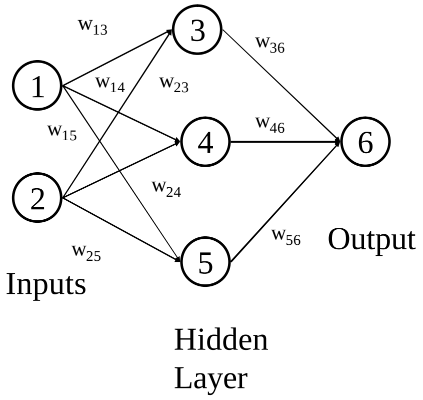
Neural Networks

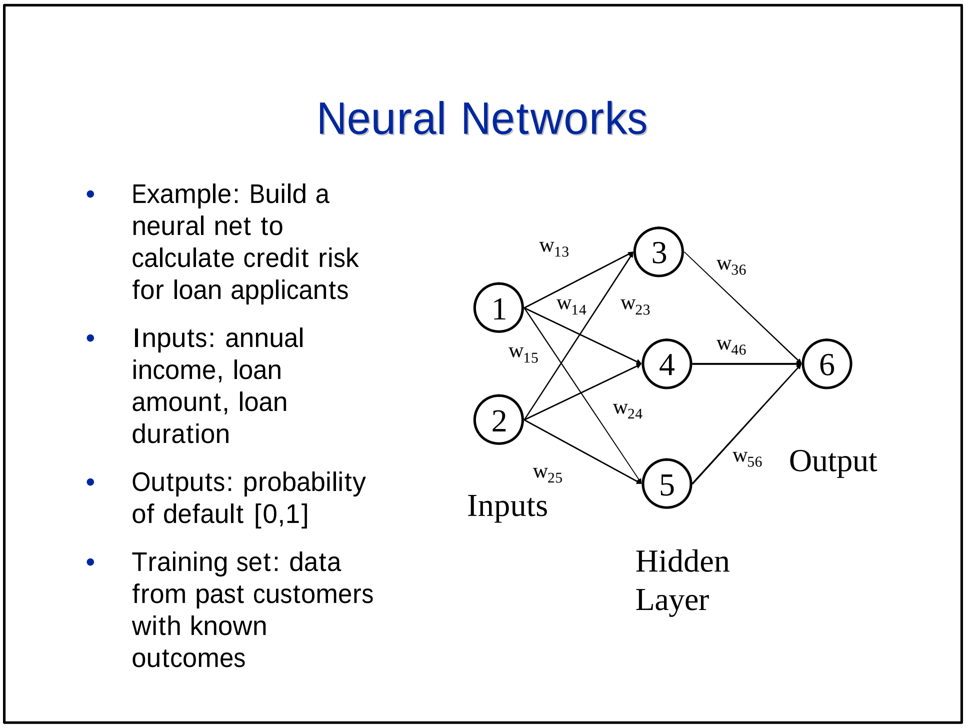
- Powerful method for constructing predictive models
- Each node applies an activation function to its input
- Activation function results are multiplied by w_{ij} and passed on to output



Neural Networks

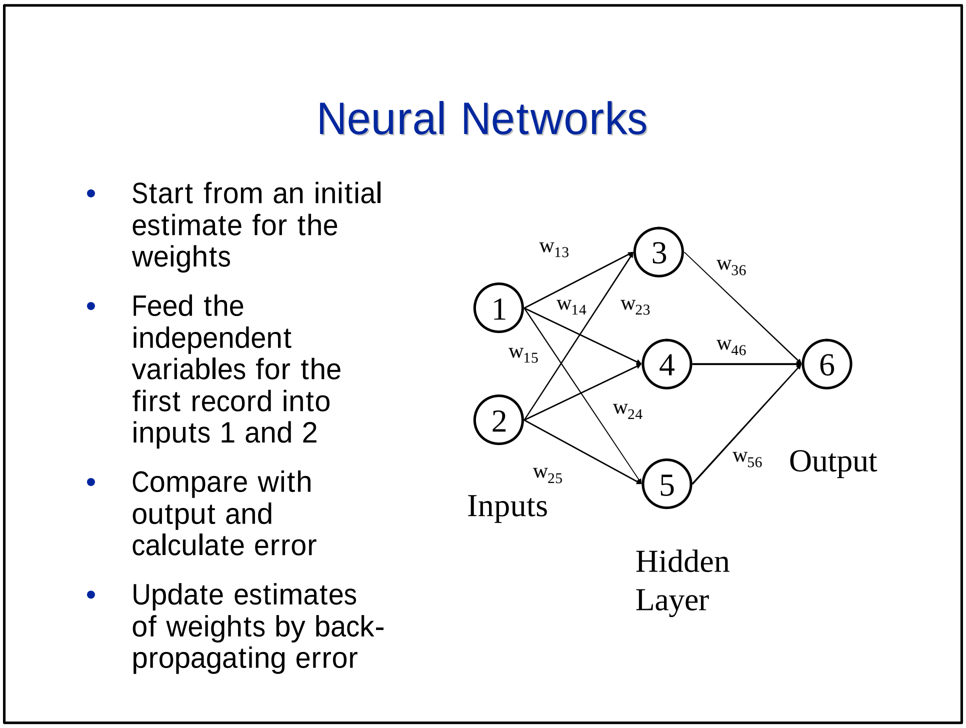
- Weights are determined using a “training set”, I.e. a number of test cases where both the inputs and the outputs are known





Neural Networks

- Example: Build a neural net to calculate credit risk for loan applicants
- Inputs: annual income, loan amount, loan duration
- Outputs: probability of default [0,1]
- Training set: data from past customers with known outcomes

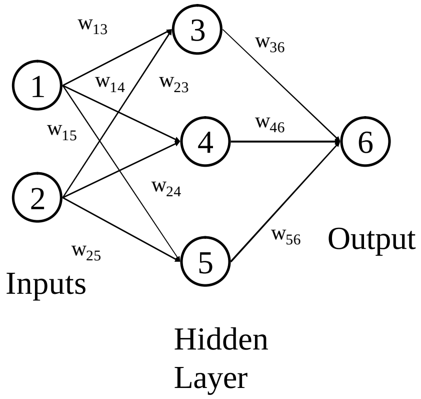


Neural Networks

- Start from an initial estimate for the weights
- Feed the independent variables for the first record into inputs 1 and 2
- Compare with output and calculate error
- Update estimates of weights by back-propagating error

Neural Networks

- Repeat with next training set record until model converges



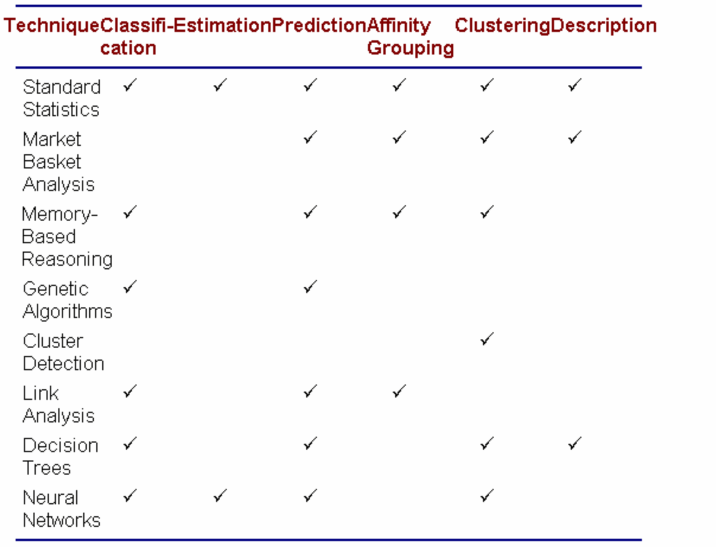
Neural networks pros and cons

- Pros
  - Versatile, give good results in complicated domains
- Cons
  - NN cannot explain the data
  - All inputs and outputs must be massaged to [0,1]

Data Mining Process

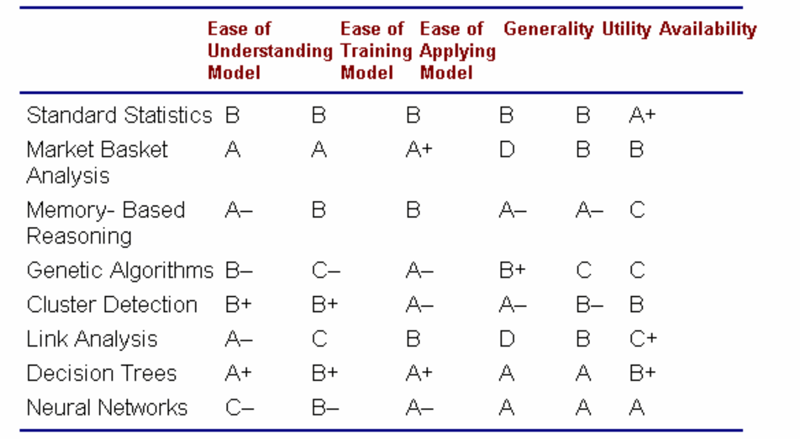
- Define business problem
- Build data mining database
- Explore data
- Prepare data for modeling
- Build model
- Evaluate model
- Deploy model and results

Selecting the right data mining technique



Technique	Classification	Estimation	Prediction	Affinity Grouping	Clustering	Description
Standard Statistics	✓	✓	✓	✓	✓	✓
Market Basket Analysis			✓	✓	✓	✓
Memory-Based Reasoning	✓		✓	✓	✓	
Genetic Algorithms	✓		✓			
Cluster Detection					✓	
Link Analysis	✓		✓	✓		
Decision Trees	✓		✓		✓	✓
Neural Networks	✓	✓	✓		✓	

Evaluating the various techniques

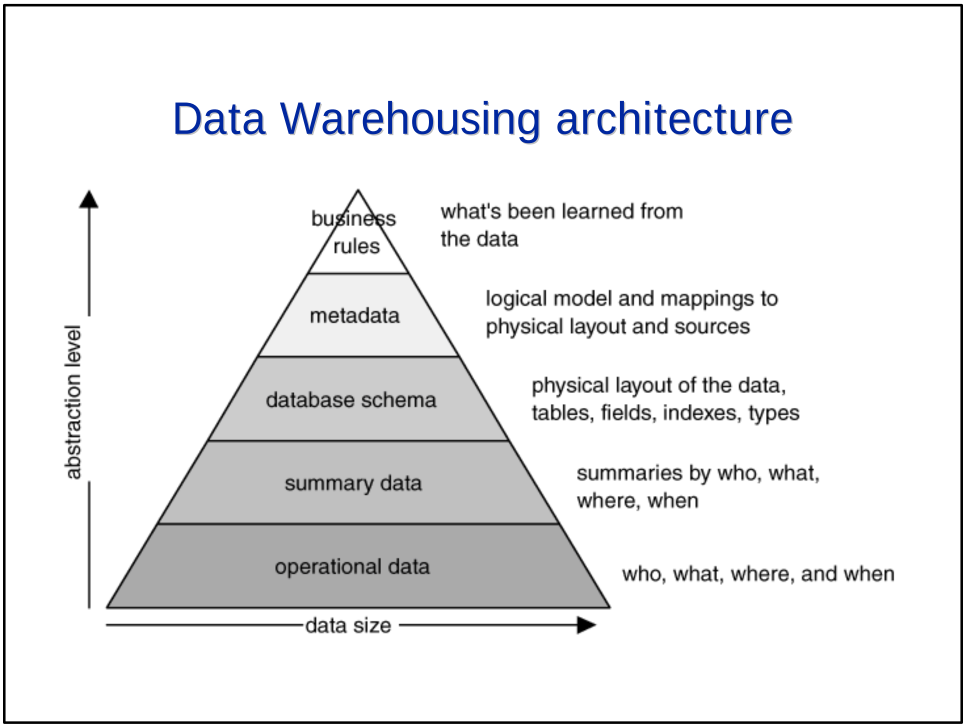


	Ease of Understanding Model	Ease of Training Model	Ease of Applying Model	Generality	Utility	Availability
Standard Statistics	B	B	B	B	B	A+
Market Basket Analysis	A	A	A+	D	B	B
Memory- Based Reasoning	A–	B	B	A–	A–	C
Genetic Algorithms	B–	C–	A–	B+	C	C
Cluster Detection	B+	B+	A–	A–	B–	B
Link Analysis	A–	C	B	D	B	C+
Decision Trees	A+	B+	A+	A	A	B+
Neural Networks	C–	B–	A–	A	A	A

What is a data warehouse?

- Data Mining has a hard requirement for clean and consistent data
- Decision Support Data
  - Are found in many different databases
    - within the company
    - outside the company
  - Are often inconsistent and “unclean”
  - In practical terms, locating and integrating all this information in real time is very difficult
- Solution:
  - Create separate repositories of data for decision support

  ⇒ data warehouses

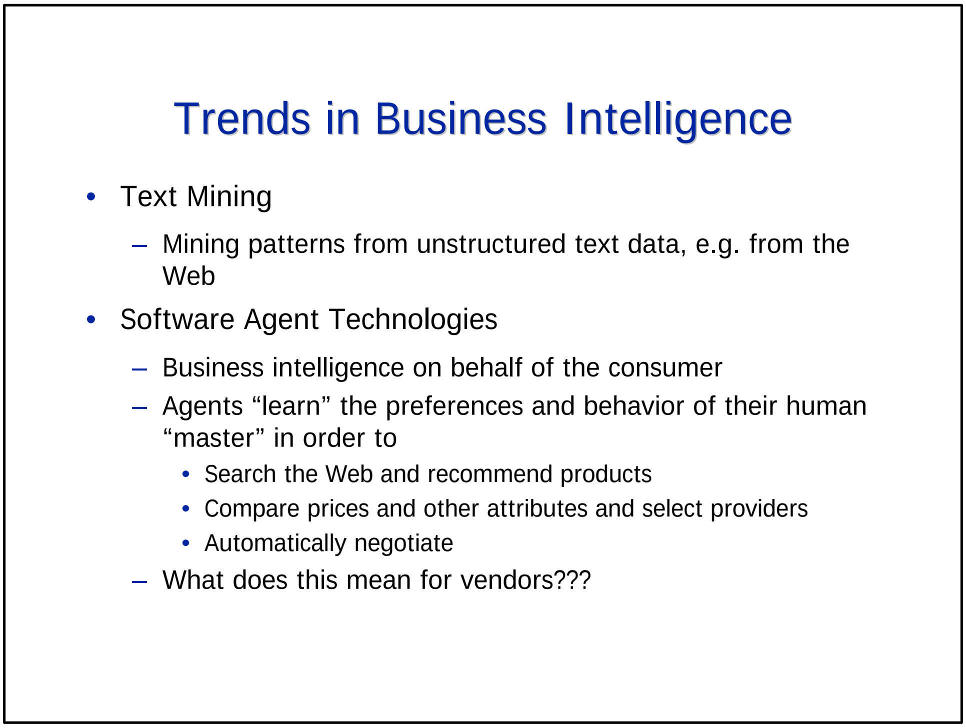


Data Warehousing architecture

business rules
what's been learned from the data
metadata
logical model and mappings to physical layout and sources
database schema
physical layout of the data, tables, fields, indexes, types
summary data
summaries by who, what, where, when
operational data
who, what, where, and when
abstraction level
data size

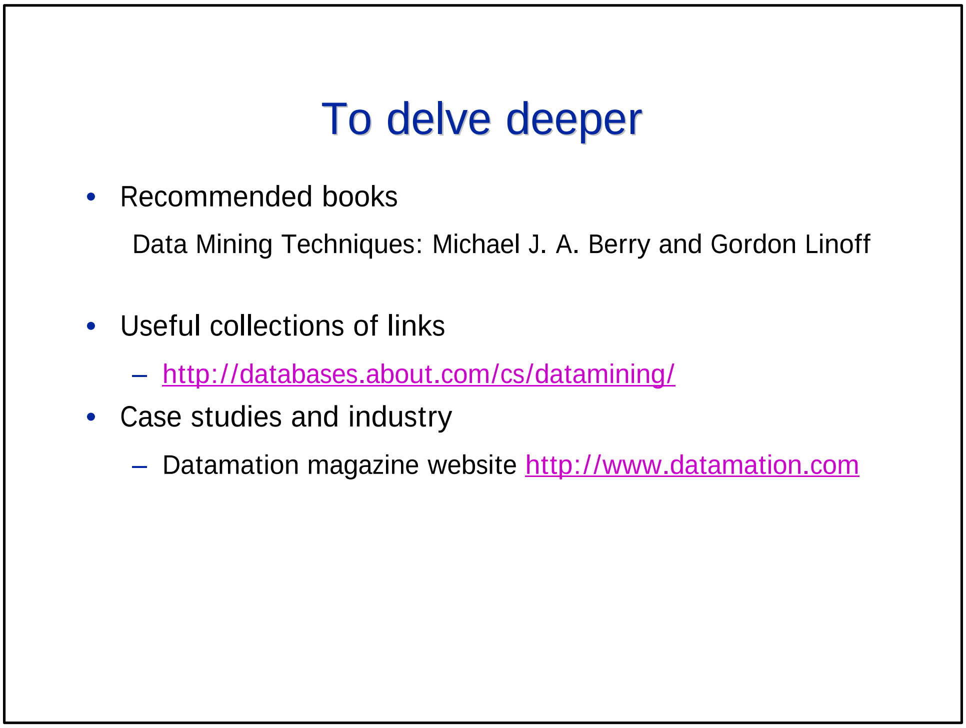
Data Warehousing considerations

- What data to include?
- How to reconcile inconsistencies?
- How often to update?



Trends in Business Intelligence

- Text Mining
  - Mining patterns from unstructured text data, e.g. from the Web
- Software Agent Technologies
  - Business intelligence on behalf of the consumer
  - Agents “learn” the preferences and behavior of their human “master” in order to
    - Search the Web and recommend products
    - Compare prices and other attributes and select providers
    - Automatically negotiate
  - What does this mean for vendors???



To delve deeper

- Recommended books

  Data Mining Techniques: Michael J. A. Berry and Gordon Linoff

- Useful collections of links
  - http://databases.about.com/cs/datamining/
- Case studies and industry
  - Datamation magazine website http://www.datamation.com