Data Mining: An Introduction

Michael J. A. Berry and Gordon A. Linoff. Data Mining Techniques for Marketing, Sales and Customer Support, 2nd Edition, 2004

Data mining

- What promotions should be targeted to specific customers?
- What is the likelihood of this being a fraudulent claim ?
- Which customers will respond to this promotion?
  (Which customers would we like to respond to this promotion?)
- Which other products may be of interest to this customer?
- What is the likelihood of this customer defaulting on a loan?
- What is the risk of this policy generating a claim > $100K?
- What are the most profitable stocks to buy/sell during the next trading session?
- When should I buy/sell GOOG?
- Which customers are most likely to defect (churn)?
- What is the most cost-effective diagnosis for this problem?
- Why are defective goods higher in this plant?

Data Mining

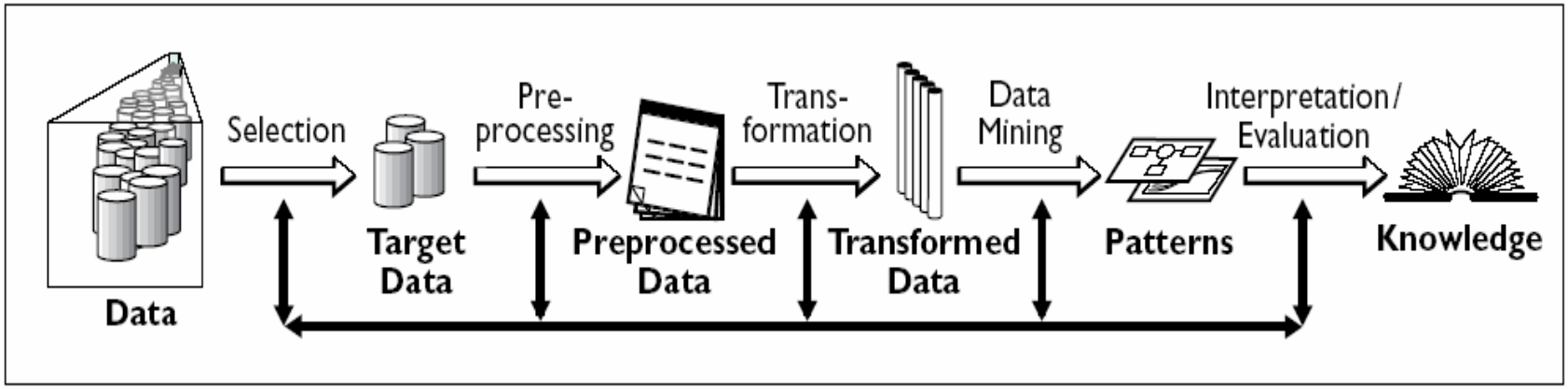
- "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"
- "the search for valuable information in large volume of data"
- "the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns or rules"

- Automated search for patterns, novel and interesting interactions, hidden relationships

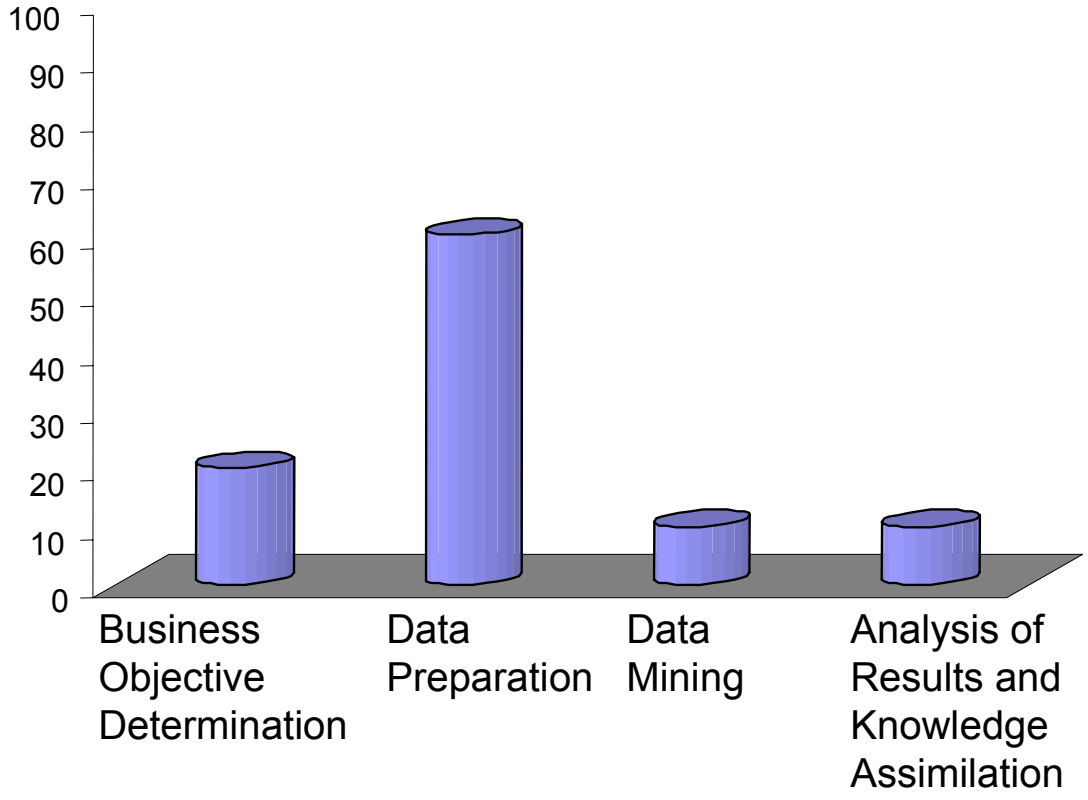
- Process of extracting *previously unknown*, *valid*, and *actionable* (understandable) information from large databases (knowledge)

- A step in the KDD process of applying data analysis and discovery algorithms
  - Obtained knowledge used for
    - Predicting new data
    - Describing existing data
    - Summarizing, visualizing a large database to facilitate decision-making

KDD Process



Business problems, Goals
Data selection, acquisition, integration
Data cleaning
noise, missing data, outliers, etc.
Exploratory data analysis
dimensionality reduction, transformations
Data mining
selecting appropriate method that match set goals (classification, regression, clustering, etc)
selecting algorithm
Testing and verification
Interpretation
Consolidation and use



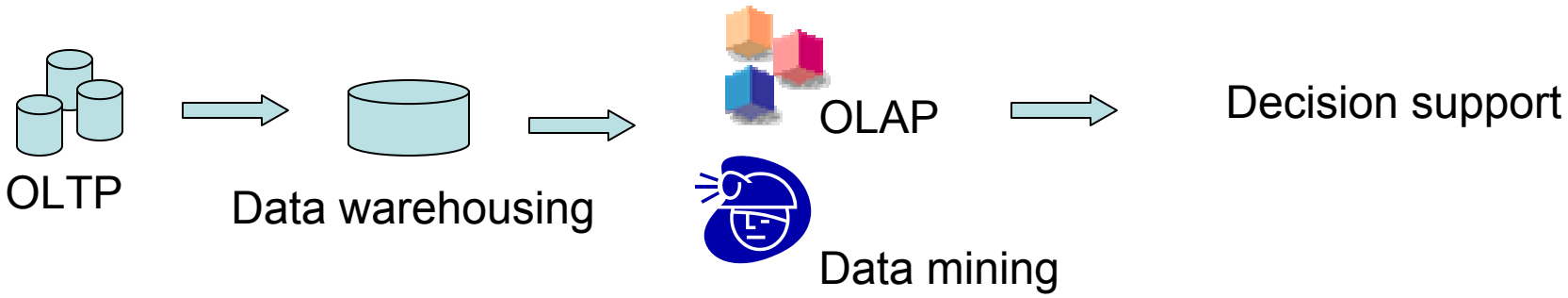
Effort for each data-mining process step

Source disciplines

- Statistics
- Artificial Intelligence, Machine Learning
- Databases
- Visualization

….

Role of Databases



OLAP (online analytical processing)

Query driven analysis, usually based on on multi-dimensional data

- Loss by region by month
- Number/value of claims by range of value by client

Data mining

Automated exploration and analysis

- Which cases constitute fraud?
- Estimating likelihood of fraud
- What represent high risk policies?

Data mining algorithm components

- **Model representation**
  - descriptions of discovered patterns
  - overly limited representation -- unable to capture data patterns
    too powerful -- potential for overfit

  (decision trees, rules, linear/non-linear regression & classification, nearest neighbor and case-based reasoning methods, graphical dependency models)

- **Model evaluation criteria**
  - how well a pattern (model) meets goals (objective function)
  - eg., accuracy, novelty, specific targeting goals etc.

- **Search method**
  - parameter search: optimization of parameters for a given model representation
  - model search: considers a family of models

Different methods suit different problems. Proper problem formulation crucial.

Issues and challenges

- Large data
  - number of variables (features), number of cases (examples)
  - multi gigabyte, terabyte databases
  - efficient algorithms, parallel processing

- High dimensionality
  - large number of features: exponential increase in search space
  - potential for spurious patterns
  - dimensionality reduction

- Over-fitting
  - models noise in training data, rather than just the *general* patterns

- Changing data, missing and noisy data

- Use of domain knowledge
  - utilizing knowledge on complex data relationships, known facts

- Understandability of patterns

Data Mining Methods

- Classification
  - Learning a function that maps a case to one of several classes
- Regression
  - Learning a function that maps a case to a real-value; functional relationships between variables
- Clustering
  - Identifying a set of categories or clusters to describe the data; natural groupings in data, based on similarity on multiple features
- Summarization
  - Compact descriptions of a subset of data. Descriptive statistics, summary rules, multivariate visualizations, functional relationships between vars.
- Database segmentation
  - groupings into sub-problems, normally by a single feature

Data mining methods

- Link analysis
  - Multi-field correlations, satisfying support and confidence thresholds (association rules)
- Dependency modeling
  - Dependencies between variables; belief networks
- Deviation detection
  - significant deviations from previous or expected values
- Sequential analysis
  - Time series patterns; to model the process generating the sequence of observations, or to report trends, deviations
- Visualization
- Text mining

Etc.

Problems, symptoms, assumptions

"Here's some data we have…can you find something interesting, useful for…"

- What is of interest and use for the business manager?
- Manager's interest in mining can arise from some dissatisfaction, sense of possible improvements – specifics?
- To alleviate certain symptoms

Establish a framework for the problem

- Data miner/modeler works with business manager (help identify the underlying problem)
- Exactly what one is looking to discover from the data, how it will be useful?
- Structuring the problem – data, assumptions, outputs, relationships, objectives, hidden assumption

Technically successful projects, but of little business use:

"Seemed reasonable that if we discovered the people who responded to our mailing that these would be the same people who would be willing to contribute money to our cause. Why didn't it work?"

"But if you can predict it, I thought one would be able to easily say what caused it, and how to stop it"

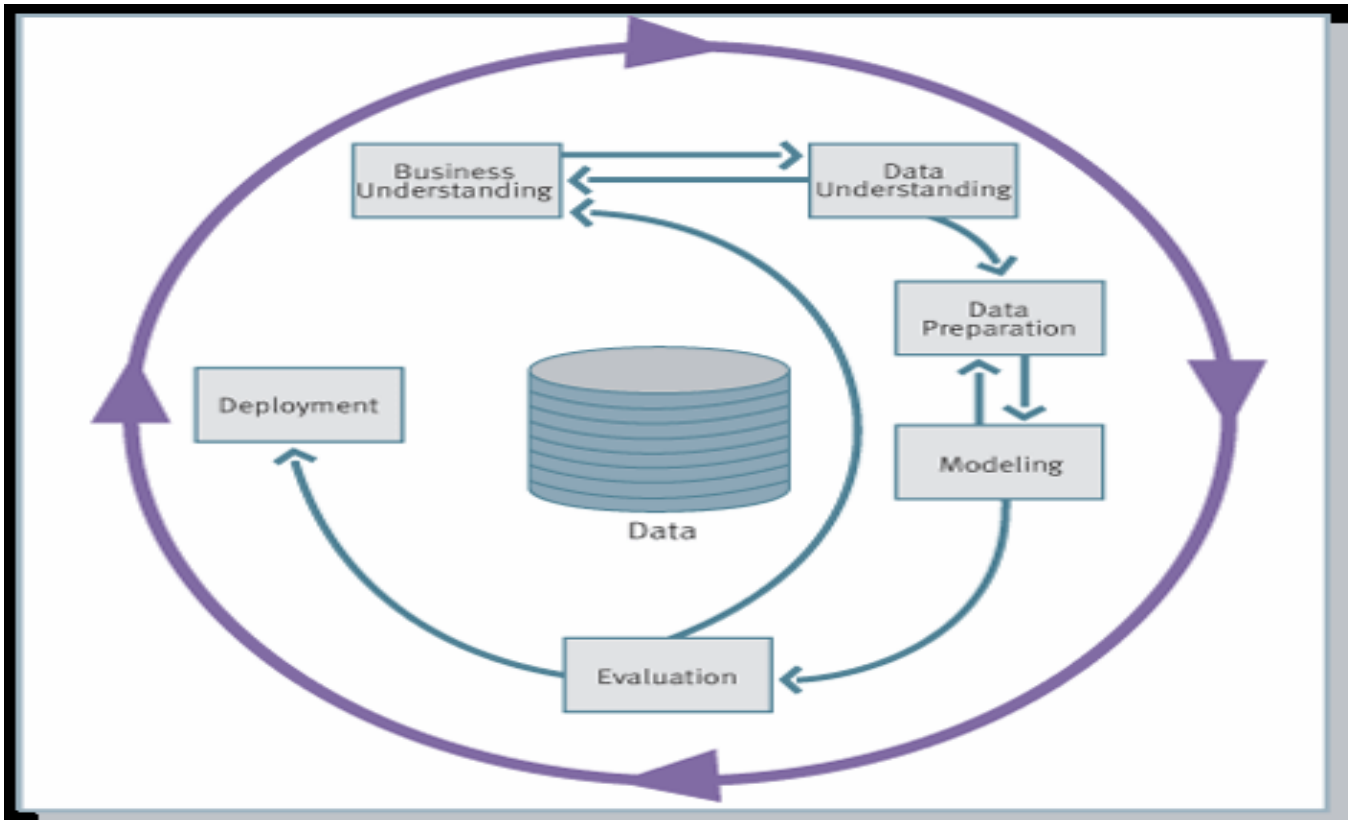
- Seeking the wrong answers! (Surely the business people can specify the question they want answered!)

Nature of data

- Ideal – measurement and data collection orchestrated with specific data mining intent
  - pertinent features
  - appropriate granularity of measurement
  - no missing data
  - consistently collected
  - validated to ensure that it represents phenomenon of interest

- Data used for mining typically collected without consideration of possible model or analysis
  - collected for some purpose - thereby imbibes certain assumptions, perspectives, interests (“objective” data?)
  - represents a simplified view of a complex reality, built-in sources of errors

Data mining process model



CRISP-DM (CRoss Industry Standard Process for Data Mining)
www.crisp-dm.org

Case Study – Marketing home equity loans

- HELOC failing to attract customers (BofA)
  - Lower the interest rate?
  - Hypothesis / Insights
    - People with college-age children borrow against home equity of pay college tuition
    - People with high but variable income use home equity to smooth out their income

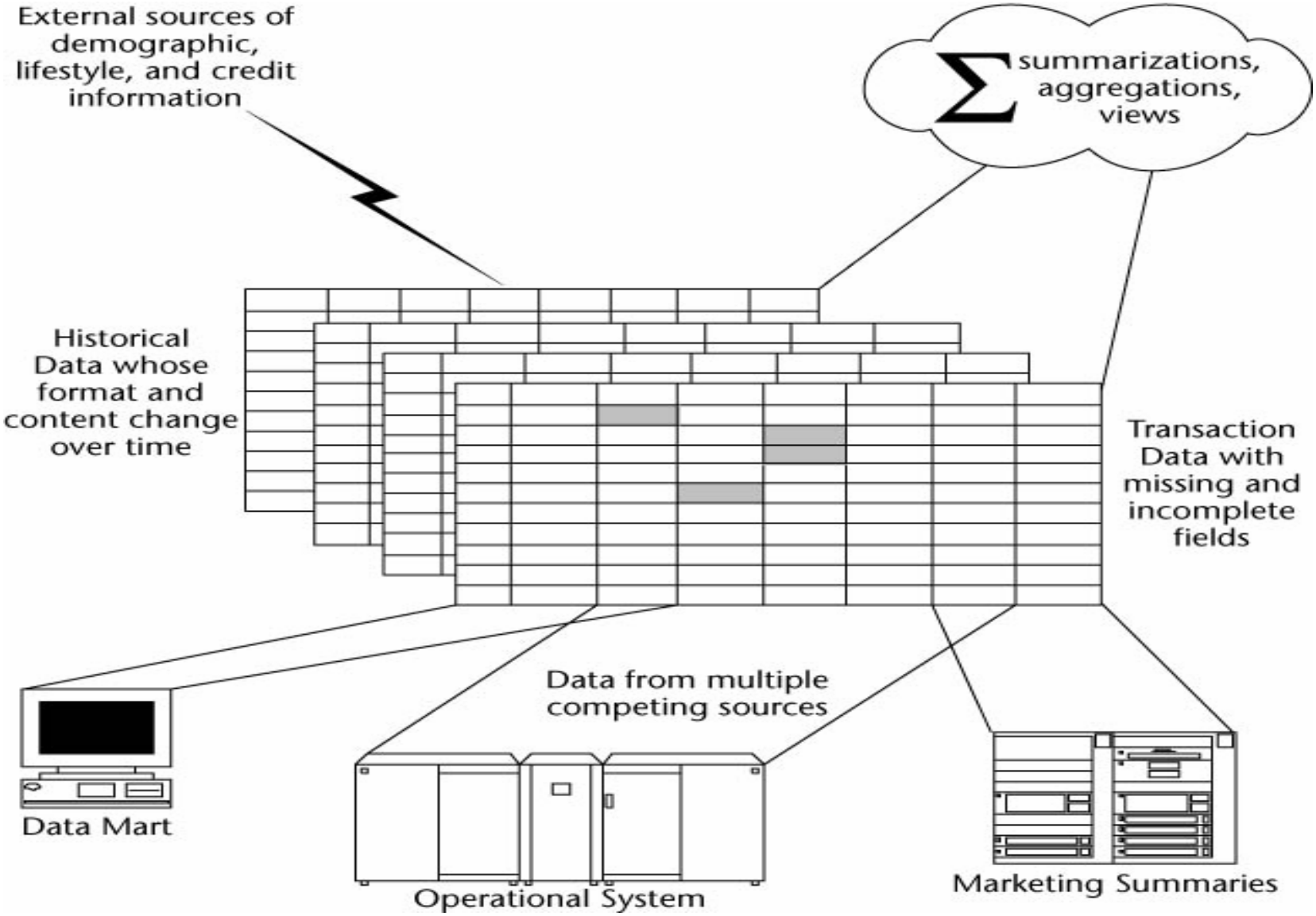
    Disappointing results!
- Data mining
  - Data from multiple systems cleansed, transformed, and integrated into data warehouse
  - Data on past cases of those who had obtained HELOC and those who had not used to build a decision tree model - used to label cases:" good prospect " or "bad prospect"
  - Sequential pattern tool used to determine when customers are most likely to want HELOC – sequence of events that preceded past successful solicitations
  - Clustering – 14 clusters, mostly uninteresting
    1 cluster: with over 1/4th of customers classified as "good prospect";
    39% had both business and personal accounts
  - Market research, with additional question " Will loan be used to start a business?"
  - Response for home equity campaign: from 0.7% to 7%
  - Transformation of the retail side of the bank from mass marketing to a leaning institution

Virtuous Cycle of Data Mining

- Identify business opportunities
- Data mining to transform data into actionable information
- Act on the information
- Measure results to complete learning cycle

Closed loop learning process

Operational data ⇨ data mining ⇨ execute programs and test ⇨ new data



Various sources of data

Case study

Once upon a time there were two companies ….
Airhead Industries, Jetstream Inc.

Problem: 60% of technical support calls passed on to engineering!

Airhead: add 8 more tech support personnel

Jetstream: Analyze tech support database, interview staff

Categorized problem types, and frequency

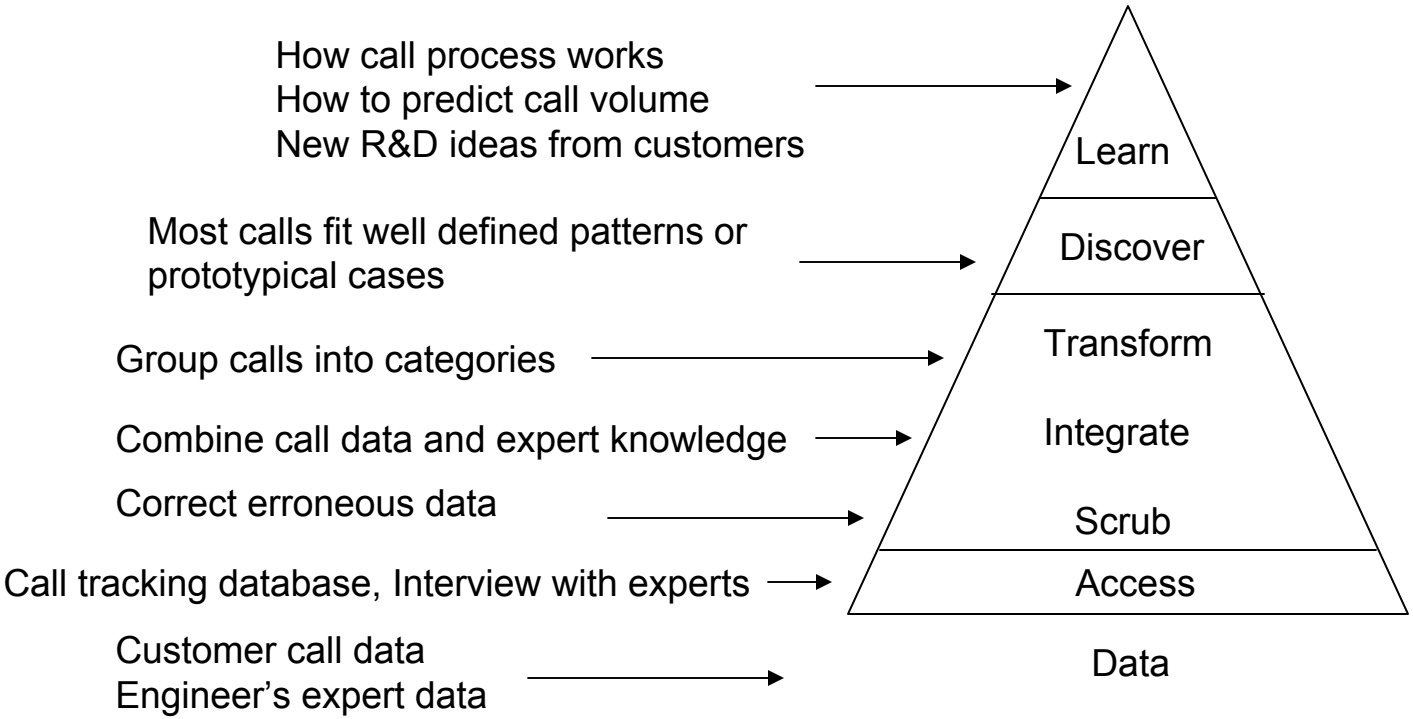
Interviewed engineering experts – most calls did not require an engineer once problem is correctly identified

Formalized expertise for addressing different types of problems, categorized into prototypical cases, available in online system for tech support

Transferred calls: 5%

Dhar, V. and R. Stein. Intelligent decision Support Methods. 1997.Prentice Hall

Case study


How call process works
How to predict call volume
New R&D ideas from customers
Most calls fit well defined patterns or prototypical cases
Group calls into categories
Combine call data and expert knowledge
Correct erroneous data
Call tracking database, Interview with experts
Customer call data
Engineer's expert data
Learn
Discover
Transform
Integrate
Scrub
Access
Data

Data mining methodology

- Helps avoid
  - patterns that aren't true
  - patterns that aren't useful
- Models that are stable
  - performance as expected when applied to new data

Are the patterns ‘for real’

- The party that does not hold the White House picks up seats in Congress in off-year election
- When the American League wins the World Series, Republicans take the White House
- When the Redskins win their last home game, the incumbent party keeps n the White House
- In U.S. presidential contests, the taller man usually wins.

Height correlates positively with success?

Correlation does not mean causation!

Data mining challenge – which patterns are predictive?

Unstable models – overfitting

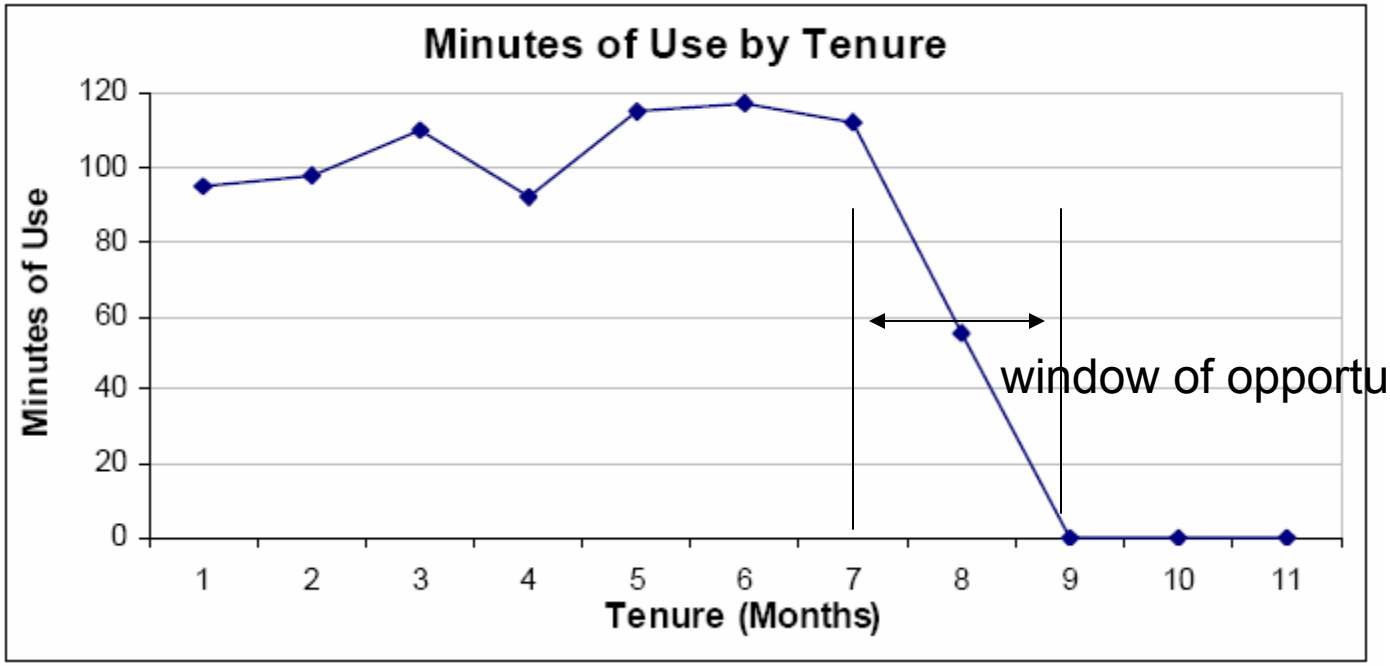
Does the model set reflect the population

- **Model Set**
  - **Training Set** – used to build a set of DM models
  - **Validation Set** – used to choose best DM model
  - **Test Set** – used to determine how the model performs

- **Sample bias**

  Valid inference from model requires data that is representative of the population that the model is intended to describe
  - Customers are different from prospects
  - Customers who web-register are different from those who do not
  - Customers of acquired company are not the same as existing customers
  - Records with certain missing values may be different

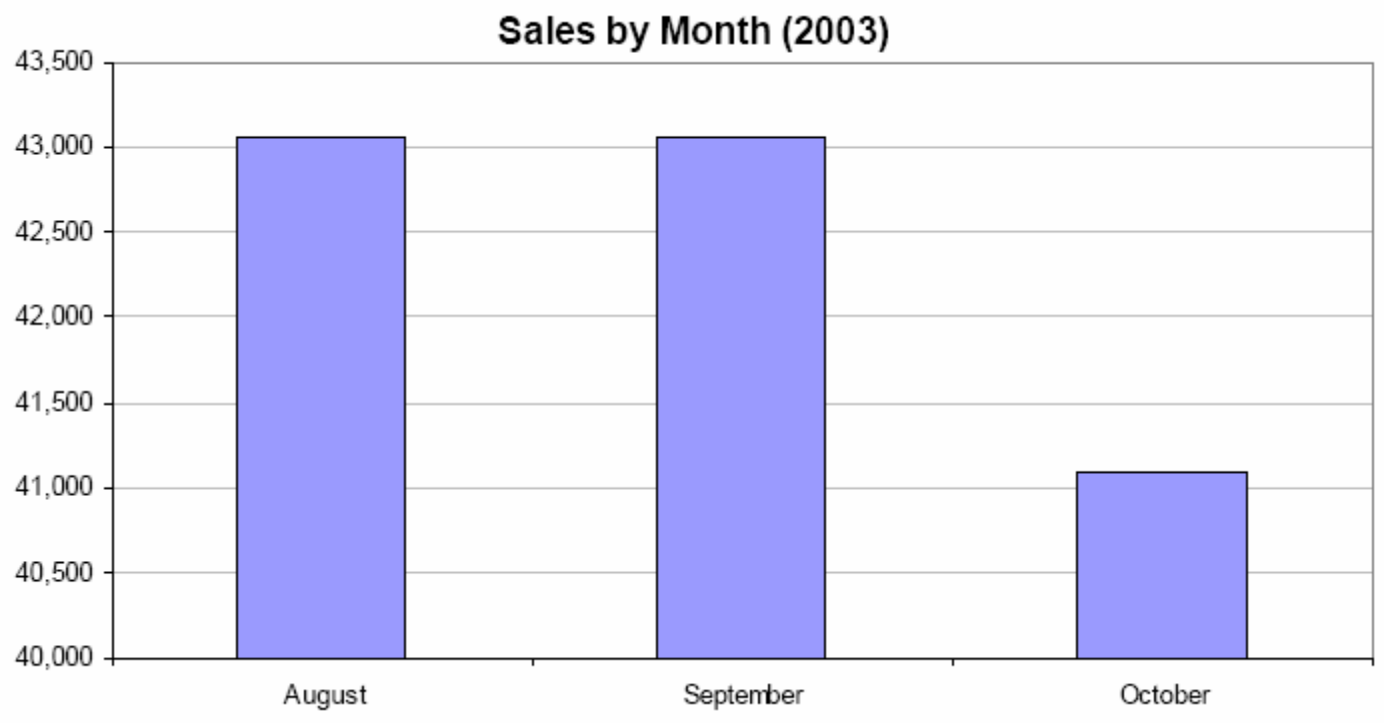
Data at the wrong level of detail



Inferring patterns that are not true

Minutes of Use by Day reveals that customer continued usage at constant rate until middle of month, and then stopped (moved to competitor)

Patterns that are not true



Did sales drop off in October?

Sales recorded only when financial markets are open.

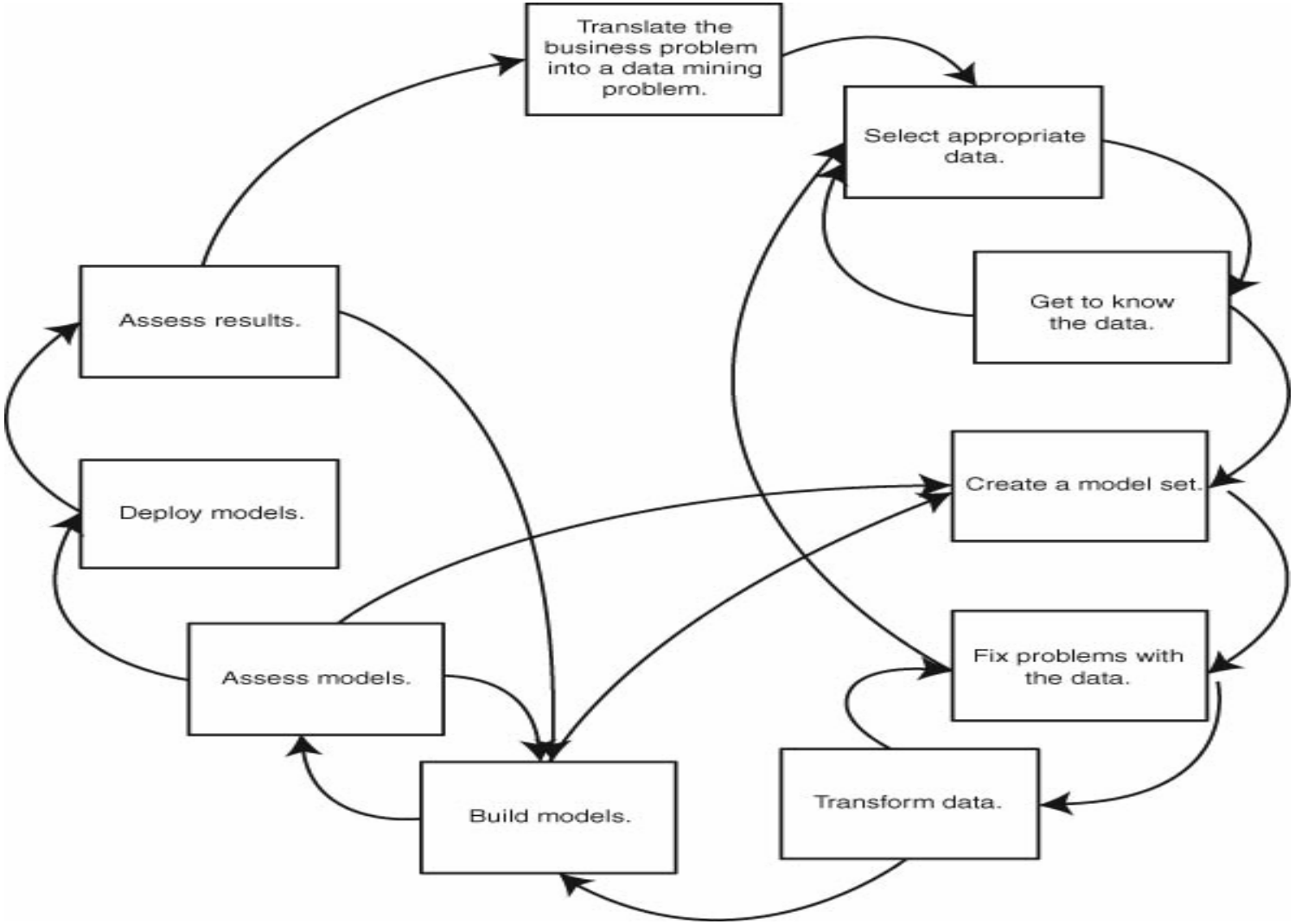
Oct. 2003 had fewer trading days.

Patterns that are already known/ not useful

- Low response to retirement savings plans from people over retirement age
- New addition to family related to buying more diapers
- People who live where there is no home delivery are not subscribers

- Rules that we cannot use
  - Credit history may be predictive of future insurance claims, but regulations prohibit use in underwriting

- Strongest patterns may arise from business rules
  - can hamper search for weaker, but more interesting rules



Data mining methodology

Translating the business problem
Defining the Business Objective

- Gaining insights into customer behavior
- Learning something interesting about claimants

Expand on broad, general statements to well-defined objectives

- Identify customers that are not likely to renew
- Rank order all customers based on propensity to buy
- List products whose sales are at risk if we discontinue wine and beer sales
- Choose the best media-mix for message to certain group

Defining the Business Objective

Who is most likely to buy X

- Predictive model with response likelihood
- Description of ‘who’ for choosing advertising media and message

Defining credit card attrition

- No activity for x months?
- Zero balance for > y months?

Model to assign score indicating churn likelihood?

- How are results to be used?
- Promotions to recent customers, low revenue churners ?
- Identify top 5,000 likely high-value churners

Translating the business problem
The Data Mining problem

- Directed data mining (supervised learning)
  - Classification
  - Estimation
  - Prediction
- Undirected (unsupervised learning)
  - Description, visualization
  - Clustering
  - Association rules

Translating the business problem
How will results be delivered

- Descriptive model for insight
  - Report with statistics, graphs, charts
- Pilot/proof-of-concept project
  - List of customers for different treatments in marketing experiment
  - Identified high fraud risk cases, to check against historical data
- Ongoing scoring
  - Scoring programs to run against subset of customer file, software to manage models and scores over time

Selecting appropriate data

- What is available
  - Operational databases, data warehouse
  - Missing, outdated documentation
  - External data
- Data size, history
  - More is better? Necessary?
- Variables

Data exploration

- Summaries, frequencies, distributions
- Validate assumptions, check for inaccuracies
  - Response rate by month
- Asking lots of questions
  - Why are there no auto insurance policies sold in NJ and MA?
  - Negative numbers in sales price field?
  - Multiple contract begin dates?

Creating the model set

- Training, test and validation data sets
- Balanced sample
  - Lower proportion of ‘responders’?
    - Sample different groups at different rates
    - Differential weights for observation from different groups
- Including multiple timeframes
  - All data should not be from same time period
- Appropriate past and present data for prediction
  - Using distant and not-too-distant past data to predict recent past

Fixing Data Problems

- Data cleaning
- Too many categories
- Skewed distributions, outliers
  - Problems for techniques that use values arithmetically
  - Transformation – log, normalization
- Missing values
- Inconsistent data encoding
  - From different operational systems, different regions, etc.

Data preparation

- Adding derived fields, taking ratios, changing counts to proportions
- Removing outliers, taking transformations
- Binning numeric values, group categorical values,, etc.

Building the model

- Applying the data mining technique (algorithm) to the data set
  - Regression
  - Decision tree
  - Neural net
  - Genetic algorithm
  - Clustering
  - Association rule
- Guarding against overfit

Model assessment

- Accuracy/ misclassification rate
- Mean squared error
- Confusion matrix, misclassification costs
- Lifts
- ROC curves

Corresponds to the business objective?

Lift table

Decile	Number of Customers	Number of Responses	Decile Response Rate	Cum Response Rate	Cum Response Lift
top	4,617	865	18.7%	18.7%	411
2	4,617	382	8.3%	13.5%	296
3	4,617	290	6.3%	11.1%	244
4	4,617	128	2.8%	9.0%	198
5	4,617	97	2.1%	7.6%	167
6	4,617	81	1.8%	6.7%	146
7	4,617	79	1.7%	5.9%	130
8	4,617	72	1.6%	5.4%	118
9	4,617	67	1.5%	5.0%	109
bottom	4,617	43	0.9%	4.6%	100
TOTAL	46,170	2,104	4.6%		