

# **Chapter 1**

# **Introduction**

**Xiaogang Su**  
Department of Statistics  
University of Central Florida

# What is Data Mining?

---

- ◆ “Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from huge volume of data ...”

— U. Fayyad, *et al.* ’s definition of DM&KDD at *KDD96*

- “*Data Mining* is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.” — a definition of data mining attributable to SAS.

# Why Data Mining?

---

- “*We are drowning in information and starving for knowledge.*” — Rutherford D. Roger (Chief Librarian at Yale Univ.)
- Today's Sexy Job: Statistician — *Technology, Business* (August 6<sup>th</sup>, 2009)

Forget the nerdy image: In today's digital world, statisticians are hot. Big firms like Google need number-crunchers to parse piles of data, and they're willing to pay for it—a statistician with a PhD can rake in \$125,000 in his or her first year on the job. “I keep saying that the sexy job in the next 10 years will be statisticians,” says Google's chief economist.” (<http://www.newser.com/story/66223/todays-sexy-job-statistician.html>)
- *New York Times* (August 5<sup>th</sup>, 2009)

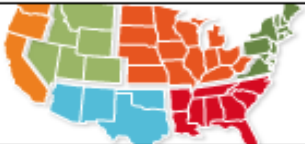
Yet data is merely the raw material of knowledge. “We're rapidly entering a world where everything can be monitored and measured,” said Erik Brynjolfsson, an economist and director of the MIT's Center for Digital Business. “But the big problem is going to be the ability of humans to use, analyze and make sense of the data.” ([http://www.nytimes.com/2009/08/06/technology/06stats.html?\\_r=1&hp](http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1&hp))

**The New York Times** **Technology**

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search Technology   Inside Technology  
Internet | Start-Ups | Business Computing | Companies

**The Ladders**  
THE MOST \$100K+ JOBS




## For Today's Graduate, Just One Word: Statistics

By STEVE LOHR  
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

[Enlarge This Image](#)



Thor Swift for The New York Times

“People think of field archaeology as Indiana Jones, but much of what you really do is data analysis,” she said.

Now Ms. Grimes does a different kind of digging. She works at [Google](#), where she uses statistical analysis of mounds of data to come up with ways to improve its

SIGN IN TO RECOMMEND

COMMENTS (58)


SIGN IN TO E-MAIL

PRINT

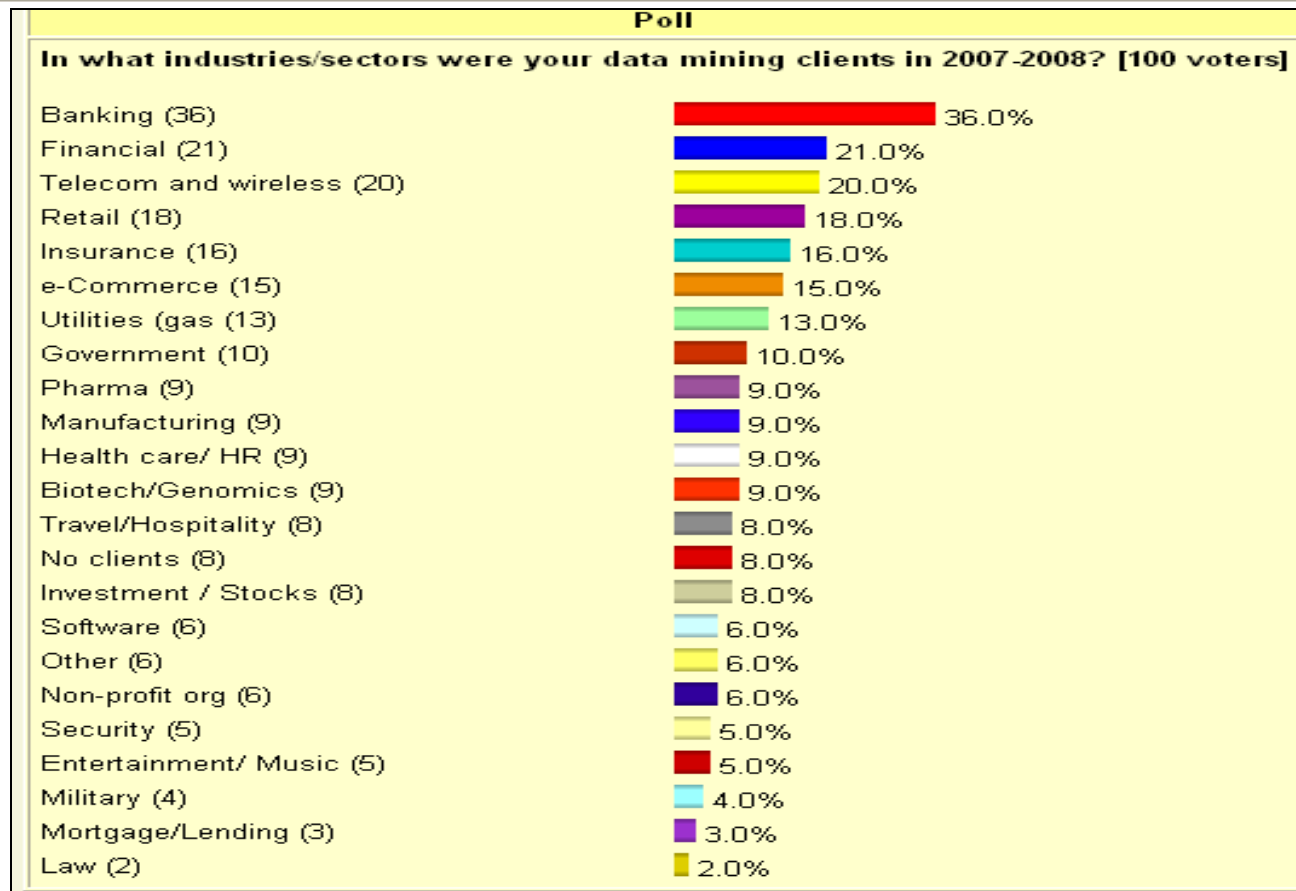
REPRINTS

SHARE

ARTICLE TOOLS SPONSORED BY



# Areas for Data Mining Applications



*Source:* Data shown above are obtained from a poll at *KDnuggets* (<http://www.kdnuggets.com/>) in March 2008.

# Data Mining Frameworks

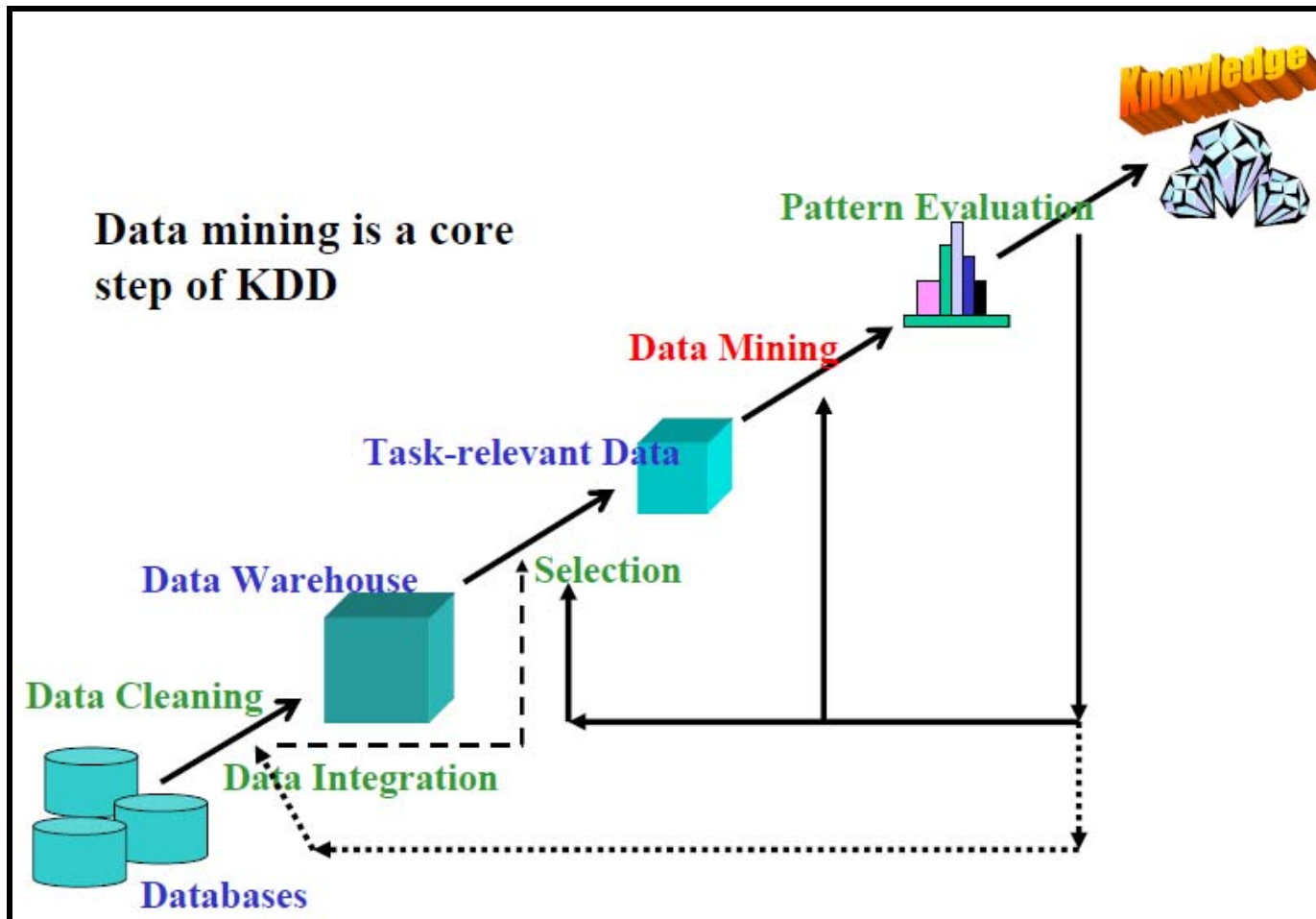
## Three General Frameworks for Data Mining:

- KDD,
- SEMMA in SAS EM,
- CRISP-DM (CRoss-Industry Standard Process for Data Mining).

Summary of the correspondences between KDD, SEMMA and CRISP-DM

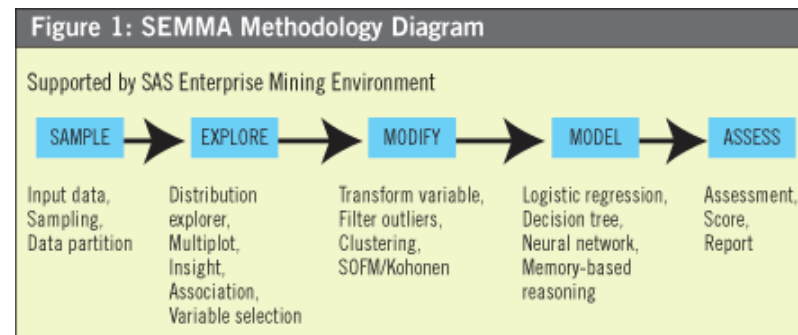
| KDD                       | SEMMA      | CRISP-DM               |
|---------------------------|------------|------------------------|
| Pre KDD                   | -----      | Business understanding |
| Selection                 | Sample     | Data Understanding     |
| Pre processing            | Explore    | Data preparation       |
| Transformation            | Modify     | Modeling               |
| Data mining               | Model      | Evaluation             |
| Interpretation/Evaluation | Assessment | Deployment             |
| Post KDD                  | -----      |                        |

# Data Mining in KDD



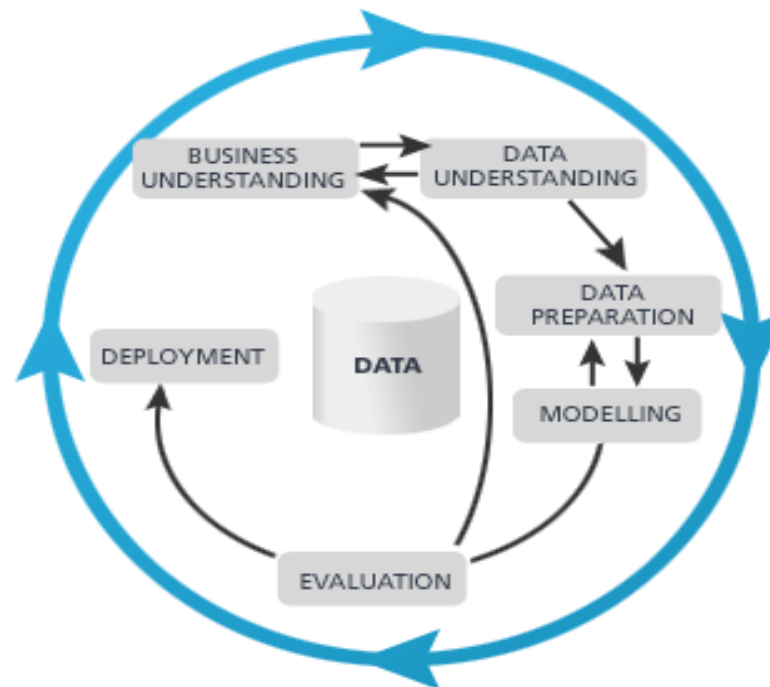
# SEMMA in SAS<sup>®</sup>

- **S**ample from data sets: Partition into Training, Validation and Test datasets
- **E**xplore data set numerically and graphically
- **M**odify: Transform variables, Impute missing values
- **M**odel: fit models e.g. regression, classification tree, neural networks
- **A**ssess: Compare models using independent test datasets



# CRISP-DM Life Cycle

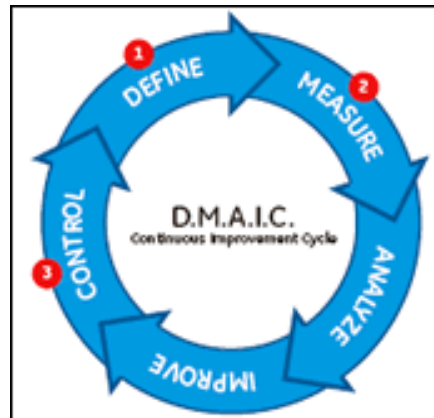
The Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit.



# CRISP-DM Life Cycle

---

The [Six Sigma](#) methodology is a well-structured, data-driven methodology for eliminating defects, waste, or quality control problems of all kinds in manufacturing, service delivery, management, and other business activities. This model has recently become very popular (due to its successful implementations) in various American industries, and it appears to gain favor worldwide. It postulated a sequence of, so-called, DMAIC steps –



- That grew up from the manufacturing, quality improvement, and process control traditions and is particularly well suited to production environments (including "production of services," i.e., service industries).

# Datasets Used in Data Mining

---

## ▪ Types of Data to be Mined

- Flat files (conventional statistical datasets)
- Relational databases, transactional databases
- Spatial databases, Time series data and temporal databases
- Multimedia databases, Text documents (text mining), e.g., prescription fraud detection Webpage (web mining)

## ▪ Characteristics

- Large  $n$  (number of observations) and large  $p$  (number of variables); Variables are of mixed types – “*curse of dimensionality*”
- Mostly coming from observational studies, instead of designed experiments.
- Data cleaning and preparation (Missing value handling, outliers, categorical variables, etc.)
- Exploratory Data Analysis (EDA) is vital.

# Data Mining Tools in General

---

- **Characteristics of Data Mining Tools**
  - Data-driven, computationally intensive, heuristic, problem-solving, automated algorithms
  - An intellectual discipline (maybe not well defined yet) that integrates statistics, computer science, data visualization, artificial intelligence, and machine learning, etc.
  
- **Some key differences between conventional statistical approaches and data mining techniques**
  - Why is hypothesis testing rarely seen in data mining?
  - Predictive modeling vs. statistical modeling
  - A battle between computer scientists and statisticians?
  - Etc.

# Statistical/Machine Learning

---

- **Two Types:**

- Supervised Learning (Learning with a Teacher or Supervisor)

- The task is predictive modeling, i.e., predicting one (or more) output (or target) variable from a set of inputs or predictors.

- Unsupervised Learning (Learning without a Teacher)  
Pattern recognition; Dimension reduction; or exploring associations

- A glance into the problem from the statistical point of view

# Supervised Learning

---

- One response (target, outcome, endpoint, output, etc.)  $Y$  vs. a set of predictors (input, independent variables, attributes, features, etc.)  $(X_1, \dots, X_p)$ 
  - If  $Y$  is continuous or quantitative: Regression
  - If  $Y$  is categorical or qualitative: Classification
- Predictive Modeling -- seek a function  $f(\underline{X})$  to predict  $Y$
- A model is a learner, which provides an answer. This answer will be evaluated by comparing with the corrected answer, i.e., the observed value provided by the teacher (target).

# Regression and Classification

---

- Loss Function: squared error loss

$$L(Y, f(\underline{X})) = \{Y - f(\underline{X})\}^2$$

- Expected Prediction Error

$$EPE(f) = E\{Y - f(\underline{X})\}^2 = \int (y - f(\underline{x}))^2 \Pr(d\underline{x}, dy)$$

which can be rewritten as

$$EPE(f) = E_{\underline{X}} E_{Y|\underline{X}}[\{Y - f(\underline{X})\}^2 | \underline{X}]$$

- Minimizing  $EPE(f)$  pointwise

$$f(\underline{x}) = \arg \min_c E_{Y|\underline{X}}[\{Y - c\}^2 | \underline{X} = \underline{x}]$$

leads to solution  $f(\underline{x}) = E(Y | \underline{X} = \underline{x})$ .

- When  $Y$  is categorical with  $K$  categories,  
 $EPF = E\{L(Y, \hat{Y})\}$ , where  $L(Y, \hat{Y})$  is from a  $K \times K$  loss matrix with zeroes on diagonal and nonnegatives elsewhere.
  - The solution is known as Bayes classifier  
 $\hat{Y} = k^*$  if  $k^* = \arg \max_k \Pr(y = k | X = x)$  (Often called the majority vote)
- Data:  $n$  iid observations  $\{(y_i, x_{i1}, \dots, x_{ip}) : i = 1, \dots, n\}$

# Modeling Approaches

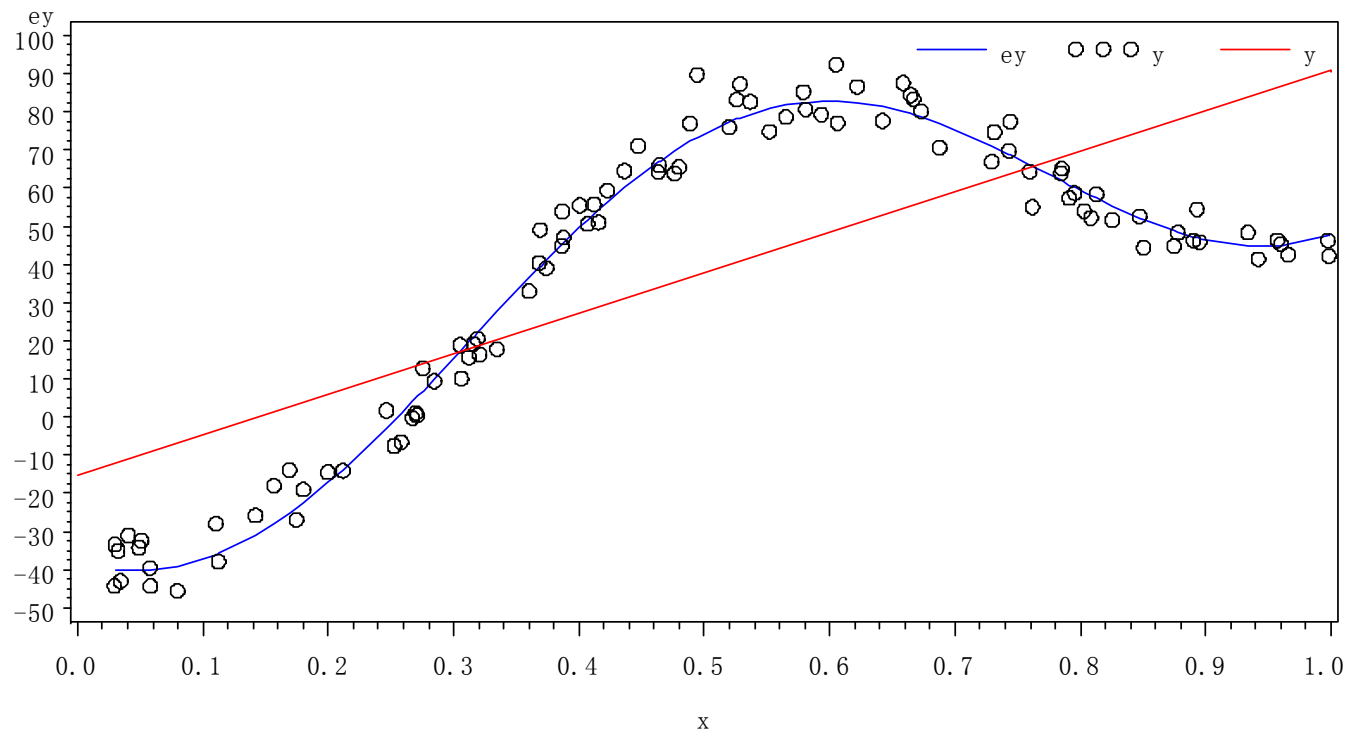
---

## ▪ Linear Regression

- Simplest yet flexible, and powerful
- Interpretable results and well-established theories
- Fast computation
- Variations and Extensions:
  - (Generalized) Linear Models
  - Generalized Additive Model
  - Single-Index Model
  - Also connected to neural networks

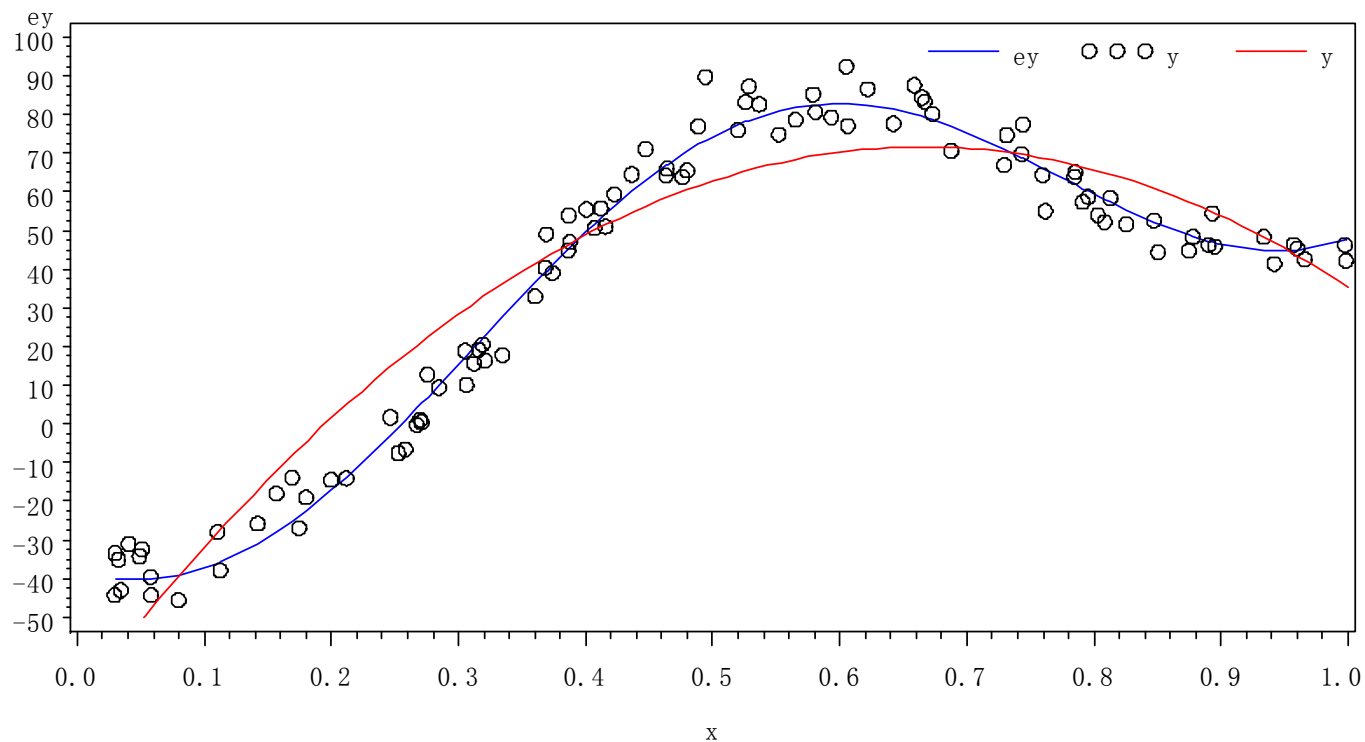
**Example:** Let's consider a simple example of predicting  $Y$  with one single predictor  $X$ , both continuous.

Figure 1: Simple Linear Regression



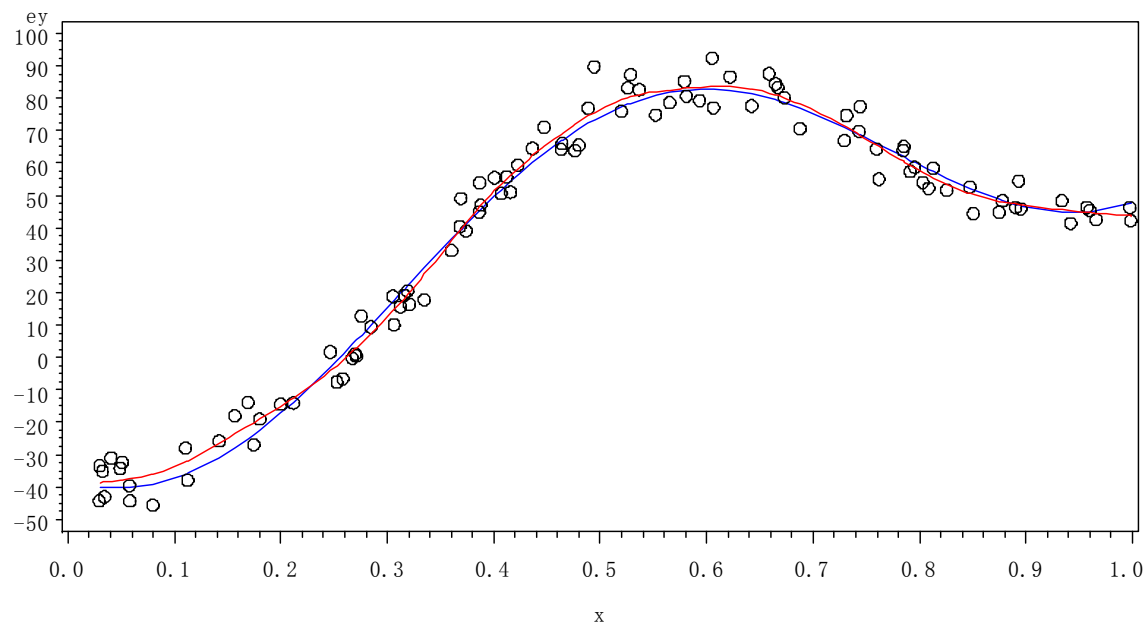
- **Parametric Nonlinear Regression Model**
  - Theory-driven and Optimization done numerically

Figure 2: Second-order Polynomial Regression



- Traditional Nonparametric Smoothing Tools:
  - Nearest Neighbors: localized approximation
  - Kernel smoothing/regression
  - Smoothing/regression splines

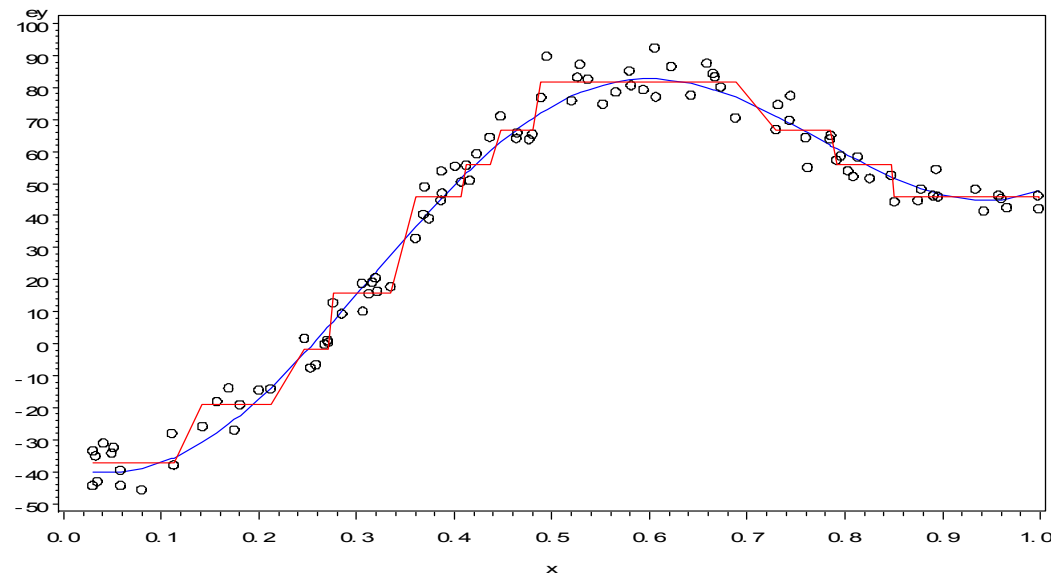
Figure 4: Smoothing Spline with  $SM=40$



## ■ Modern Learning Algorithms

- Tree-Based Methods with bagging, boosting, and random forests
- Multivariate Adaptive Regression Splines (MARS)
- Neural Networks and Support Vector Machine

Figure 5: Regression Tree



# Unsupervised Learning

---

- A (large) set of inputs:  $X_1, \dots, X_p$  with joint density  $\Pr(\underline{X})$ 
  - No target variable involved and high-dimensional
  - Infer about  $\Pr(\underline{X})$  or characterize  $X$ -values, or collection of such values, where  $\Pr(\underline{X})$  is relatively high
  - Methods are mostly exploratory in nature
  - No measure of success
- Data:  $\{(x_{i1}, \dots, x_{ip}) : i = 1, \dots, n\}$

# Tools for Unsupervised Learning

---

- Low-Dimensional Cases ( $p \leq 3$ ): graphical/numerical exploration and crude global models such as Gaussian mixtures,
  
- Principal components, Multidimensional scaling, self-organizing maps (SOM), and principal curves
  - Dimension Reduction --- to identify low-dimensional manifolds within the  $X$ -space that represent high data density.
  - Associations among variables

- Cluster Analysis and Mixture Modeling
  - To find multiple convex regions of the X-space that contains a mixture of simpler densities representing distinct types or classes of observations.
  
- Association Rules:
  - Describes regions of high density in the special case of high-dimensional binary-valued data
  
- Proliferation of proposed unsupervised learning methods ---  
Difficult to ascertain the validity of inferences drawn from the analysis results.

# Some Real Data Mining Stories - I



**Larry Page & Sergey Brin**

- **Google Search Engine:** Analyzing internet usage data via web mining to find out how to better meet customers' needs.

# Some Real Data Mining Stories - II

---



Diaper



Beer

- **Wal-Mart Market Basket Analysis:** diaper & beer association identified using association rules.

# Some Real Data Mining Stories - III

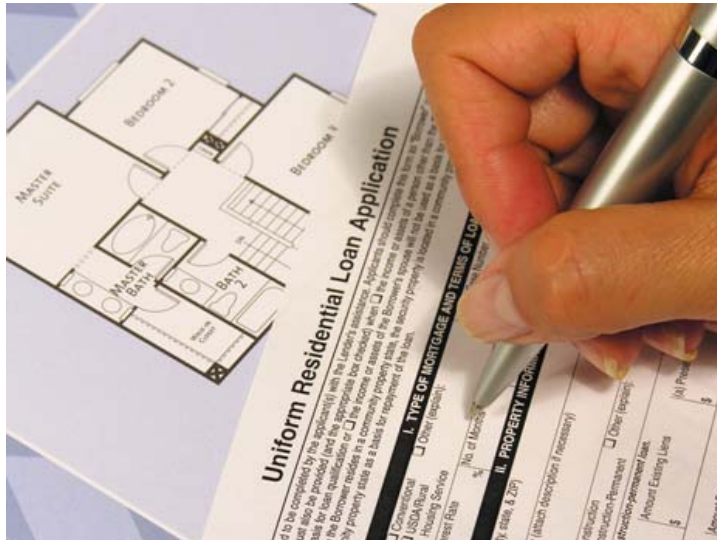
---



## Computer/Network Intrusion Detection

- Is it an intrusion? Which kind of intrusion is it?
- How to detect? Utilize historical data: previous normal patterns and intrusion patterns
- To classify the current pattern – a predictive modeling problem
- Similar problems: spam email identification, virus detection, etc.

# Some Real Data Mining Stories - IV



## Load Application Screening



- **Task:** In the 1980s, American tried to divide loan applications into three categories (approved, rejected, further human expert judgment needed)
- **Results** (put in use immediately)
  - The human experts could correctly predict whether or not an application would default on the load with only about 50% accuracy.
  - DM-based prediction improved the prediction accuracy to 70%.

# Some Real Data Mining Stories - V

---

## Printing Press Control

RR DONNELLEY

**Task:** During rotogravure printing, grooves sometimes develop on the printing cylinder, ruining the final product. This phenomenon is banding, the causes of which are imperfectly understood, even by experts. The printing company R.R. Donnelly (USA) attempted to reduce its banding problems.

### Results:

- DM produced rules were superior to the consultant's advice because:
  - More specific to the plant;
  - Filled gaps in the consultant's advice and thus were more complete;
  - One DM rule contradicted the consultant's advice but proved to be correct.
- The DM rules have been in everyday use in the Donnelly plant in Gallatin, Tennessee, for over a decade and have reduced the number of banding occurrences from 538 (in 1989) to 26 (in 1998).

## DM in CRM - Recommendation Systems

---

- **Business opportunity:**

E-commerce and Internet

Users rate items (Amazon.com, CDNOW.com, MovieFinder.com) on the web. How to use information from other users to infer ratings for a particular user?

- **Solution:**

- Use of a technique known as collaborative filtering
- From Clicks to Customers

- **Benefit:** Increase revenues by *cross-selling* and *up-selling*



## DM in CRM - Credit Risk Analysis

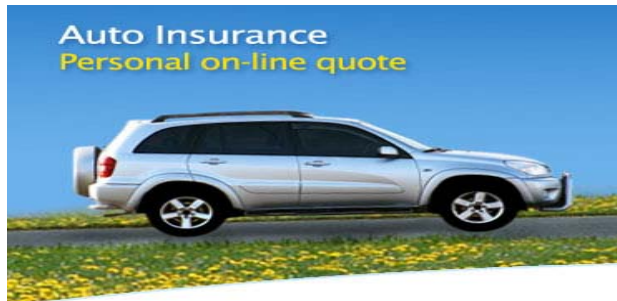
---

- **Business problem:** Reduce risk of loans to delinquent customers
- **DM Solution:** Use credit scoring models using discriminant analysis (e.g., logistic regression or decision trees) to create score functions that separate out risky customers
- **Benefit:** Decrease in cost of bad debts



# DM Applications – Fraud Detection

- **Business problem:** Fraud increases costs or reduces revenue
- **DM Solution:** Use logistic regression, neural nets to identify characteristics of fraudulent cases to prevent in future or prosecute more vigorously.
- **Benefit:** Increased profits by reducing undesirable customers



## Example: AIBM

(Automobile Insurance Bureau of Massachusetts)

- Past reports on claims adjustors scrutinized by experts to identify cases of fraud
- Several characteristics (over 60) of claimant, type of accident, type of injury/treatment coded into database
- Dimension Reduction methods used to obtain weighted variables. Multiple Regression Step-wise Subset selection methods used to identify characteristics strong correlated with fraud

# DM Applications – Churn Analysis in Telcoms

---

- **Business Problem:** Prevent loss of customers, avoid adding churn-prone customers
- **DM Solution:** Use neural nets, time series analysis to identify typical patterns of telephone usage of likely-to-defect and likely-to-churn customers
- **Benefit:** Retention of customers, more effective promotions

## Example: France Telecom

- CHURN/Customer Profiling System implemented as part of major custom data warehouse solution
- Preventive CPS based on customer characteristics and known cases of churning and non-churning customers identify significant characteristics for Churn.
- Early detection CPS based on usage pattern matching with known cases of churn customers.



## DM Applications – Target Marketing

---

- **Business problem:** Use list of prospects for direct mailing campaign
- **DM Solution:** Use Data Mining to identify most promising respondents combining demographic and geographic data with data on past purchase behavior
- **Benefits:** Better response rate, savings in campaign cost



### Example: Fleet Financial Group

- Redesign of customer service infrastructure, including \$38 million investment in data warehouse and marketing automation
- Used logistic regression to predict response probabilities to home-equity product for sample of 20,000 customer profiles from 15 million customer base
- Used CART to predict profitable customers and customers who would be unprofitable even if they respond