Data Mining Concepts

Wipro Technologies

Agenda

- The Data mining Technology
- Data mining Process
  - Data Preparation
  - Data Mining Models
- Data Mining Techniques
- Data Mining Applications & Tools
- Data Mining Methodologies

The Data mining Technology

A Problem...

- You are a marketing manager for a brokerage company
  - Problem: Churn is too high
    - Turnover (after six month introductory period ends) is 40%
  - Customers receive incentives (average cost: $160) when account is opened
  - Giving new incentives to everyone who might leave is very expensive (as well as wasteful)
  - Bringing back a customer after they leave is both difficult and costly

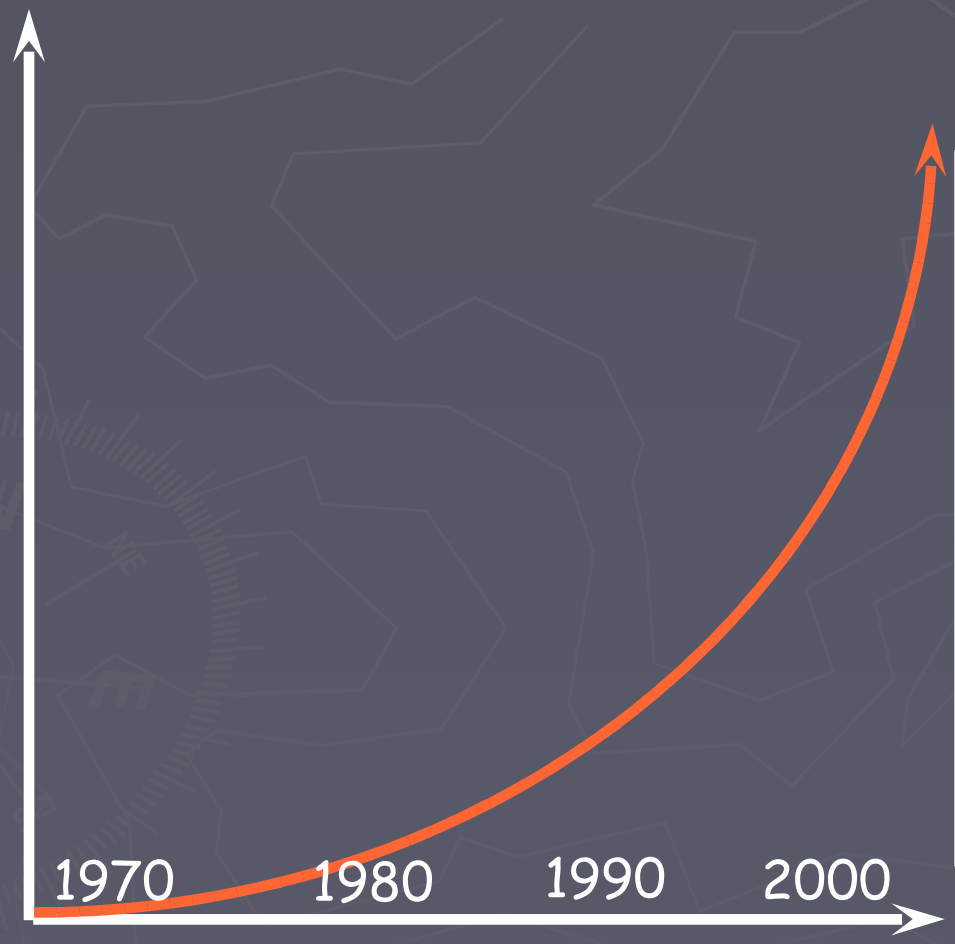
… A Solution

- One month before the end of the introductory period is over, predict which customers will leave
  - If you want to keep a customer that is predicted to churn, offer them something based on their predicted value
    - The ones that are not predicted to churn need no attention
  - If you don't want to keep the customer, do nothing
- How can you predict future behavior?
  - Tarot Cards
  - Magic 8 Ball
  -
- Data Mining

Data Mining : Why now ?

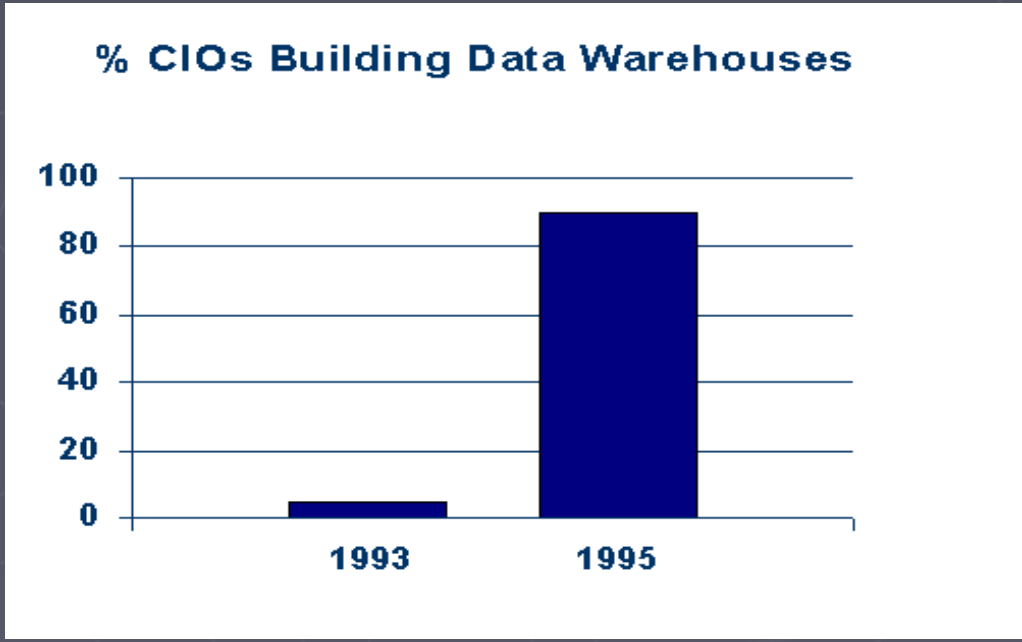
- Changes in the Business Environment
  - Customers becoming more demanding
  - Markets are saturated
  - Replace statistician ⇨ Better models, less grunge work
  - Many different data mining algorithms / tools available
  - Statistical expertise required to compare different techniques
  - Build intelligence into the software
- Drivers
  - Focus on the customer, competition, and data assets
- Enablers
  - Increased data hoarding
  - Cheaper and faster hardware

Growing Base of data



- Data doubling every 20 months in the world
- Businesses feel there is value in historical data
- Automated knowledge discovery is only way to explore this data

Improved Data Collection

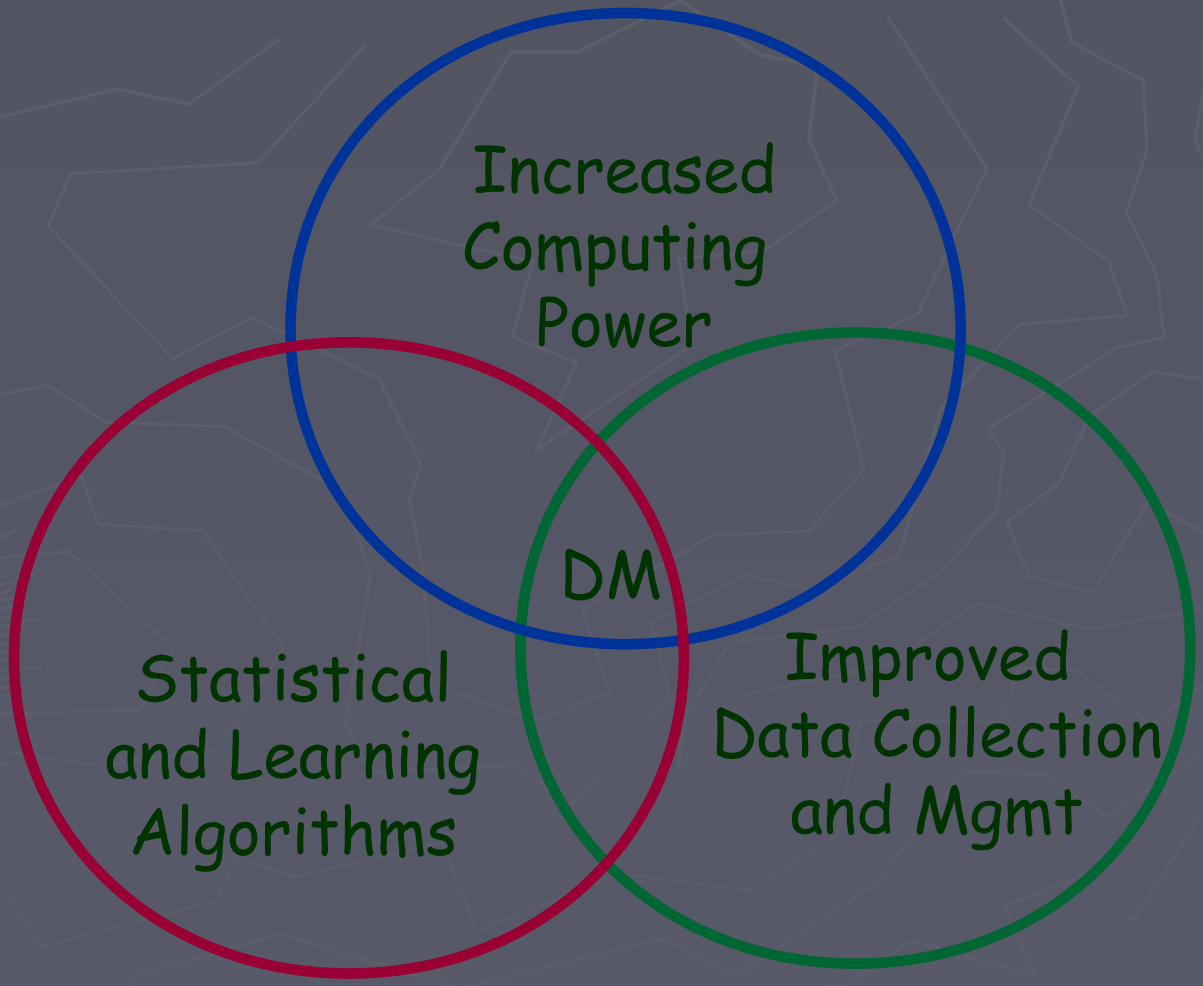


- Data Collection ➔ Access ➔ Navigation ➔ Mining
- The more data the better (usually)

Improved Algorithms

- Techniques have often been waiting for computing technology to catch up
- Statisticians already doing “manual data mining”
- Good machine learning is just the intelligent application of statistical processes
- A lot of data mining research focused on tweaking existing techniques to get small percentage gains

Convergence of Three Key Technologies



Motivation for doing Data Mining

- Investment in Data Collection/Data Warehouse
- Add value to the data holding
- Competitive advantage
- More effective decision making
- OLTP =) Data Warehouse =) Decision Support
- Work to add value to the data holding
- Support high level and long term decision making
- Fundamental move in use of Databases

Data Mining - Definition

- Data mining is the automated detection for new, valuable and non trivial information in large volumes of data.
- It predicts future trends and finds behavior that the experts may miss because it lies outside their expectations
  - Data mining lets you be proactive
  - Prospective rather than Retrospective
- Data Mining Leads to simplification and automation of the overall statistical process of deriving information from huge volume of data.

Data Mining Introduction

- DM - what it can do
  - Exploit patterns & relationships in data to produce models
  - Two uses for models:
    - Predictive
    - Descriptive
- DM - what it can’t do
  - Automatically find relationships
    - without user intervention
    - when no relationships exist

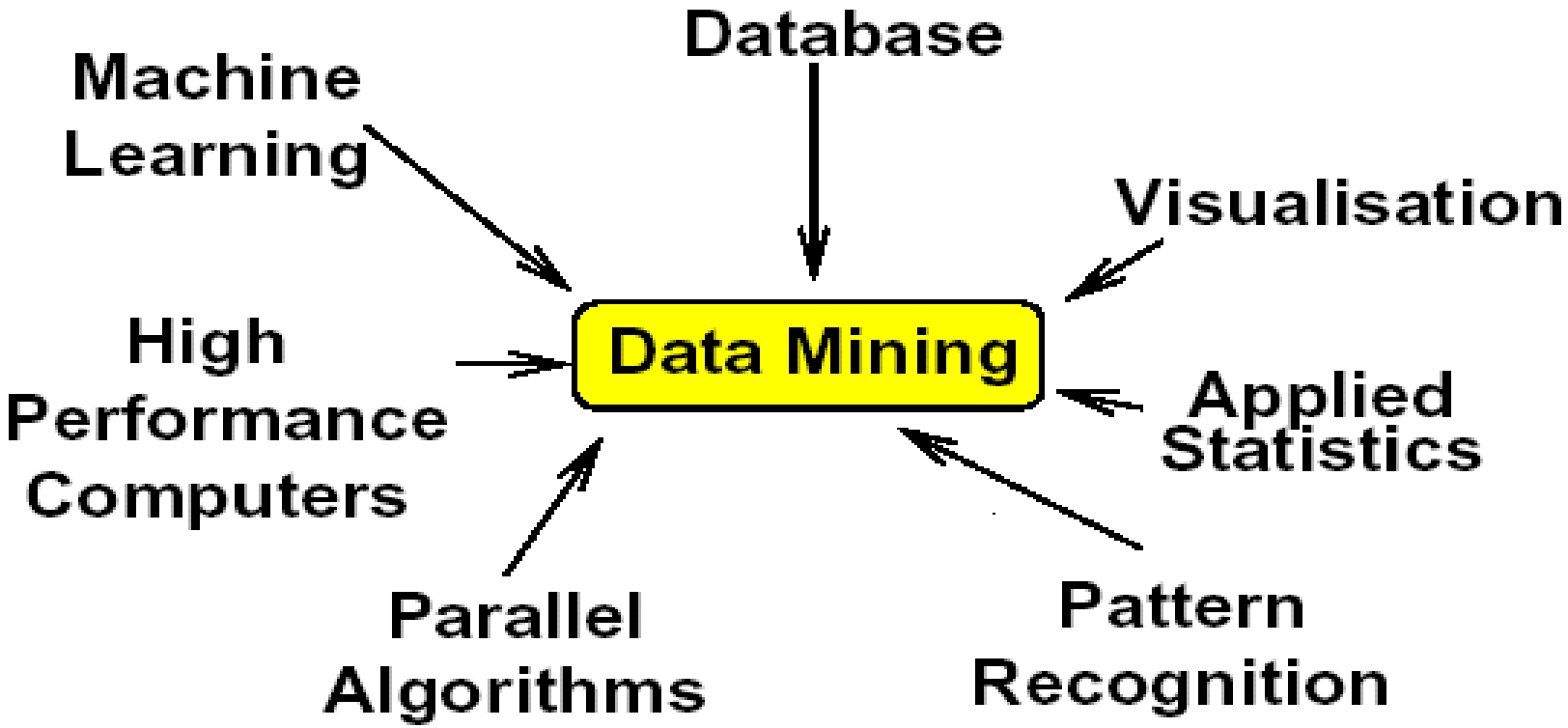
Data Mining Introduction

- Data Mining and Data Warehousing
  - Data preparation for DM may be part of the Data Warehousing
  - Data Warehouse not a requirement for Data Mining
- DM and OLAP
  - OLAP = Classic descriptive model
  - Requires significant user input
  - Example : Beer and diaper sales
    - An OLAP tools shows reports giving sales of different items
    - A data mining tool analyses the data and predicts ‘how many times beer and diapers are sold together

Data Mining Introduction

- DM and Classical Statistics
  - Classical statistics based on elegant theory and restrictive data assumptions
  - Fine if data sets small and assumptions met
  - Modeler plays active role - specifying model form, interactions, etc
  - In newer tools, pattern finding is data-driven rather than user-driven

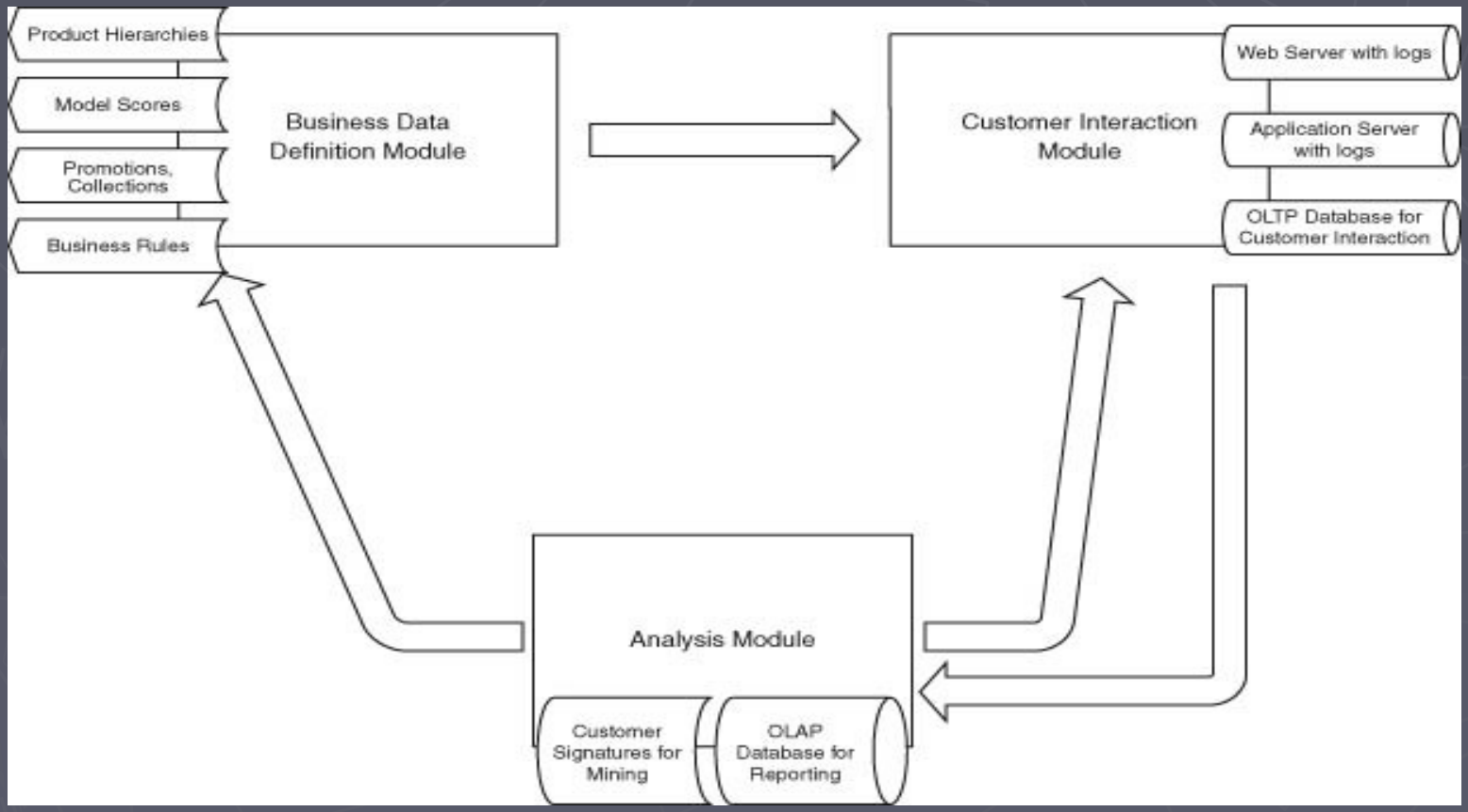
Data Mining : Introduction



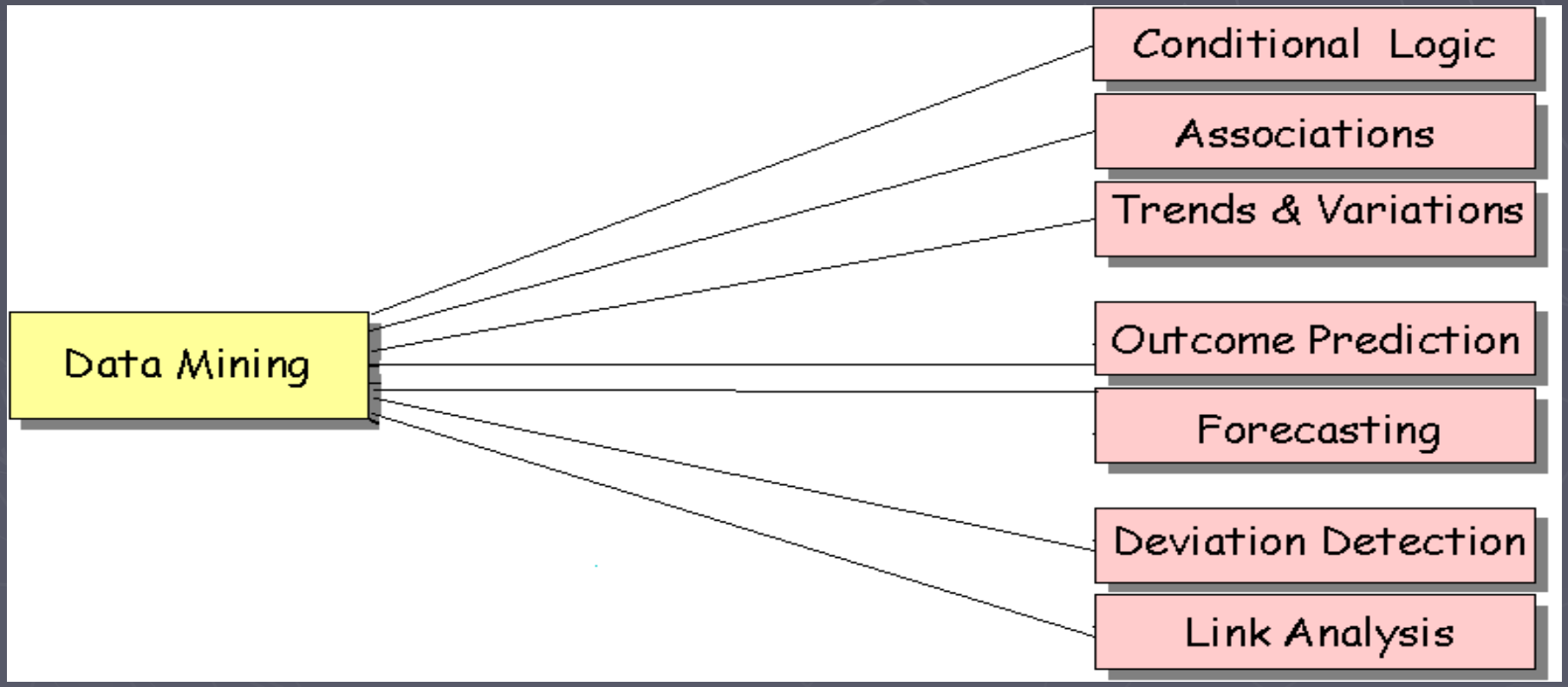
Data Mining is Not ...

- Data warehousing
- SQL / Ad Hoc Queries / Reporting
- Software Agents
- Online Analytical Processing (OLAP)
- Data Visualization

Data Mining Environment



Data mining - Analysis



Examples of Data Mining

Conditional Logic	If profession = athlete then age < 30 in 90 % cases
Associations	When Paint is sold, Paint brushes are also sold 85% times
Trends & Variations	Golf balls sales are seasonal with Summer peak and Winter low

Examples of Datamining contd....

Outcome Prediction

- How many people can be expected to respond to a mailer campaign?

Forecasting

- What will be the total sales of this product range in next quarter taking into account seasonal and long term trends?

Deviation Detection

- Is this insurance claim likely to be a fraud?

Link Analysis

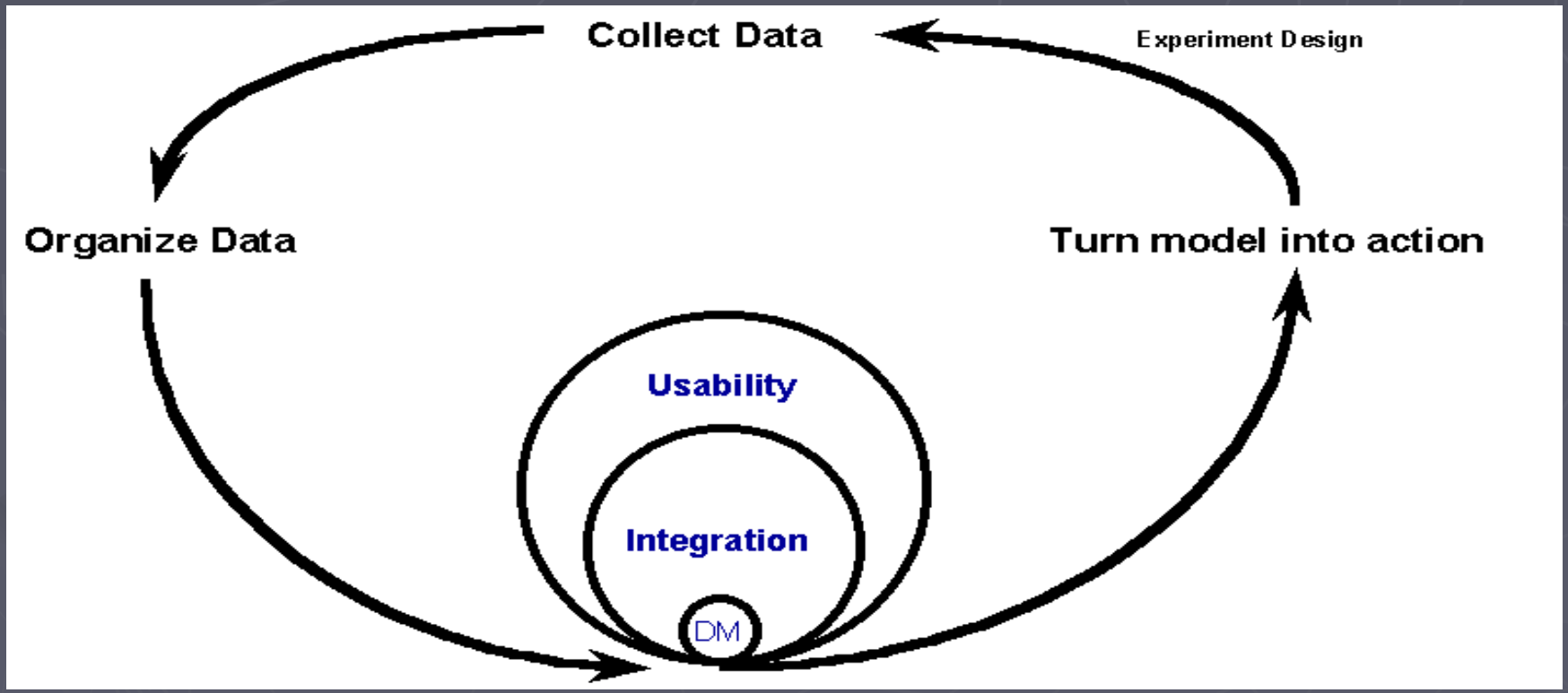
- When a person is fired, he is likely to default on credit card payments.

Data mining - Users

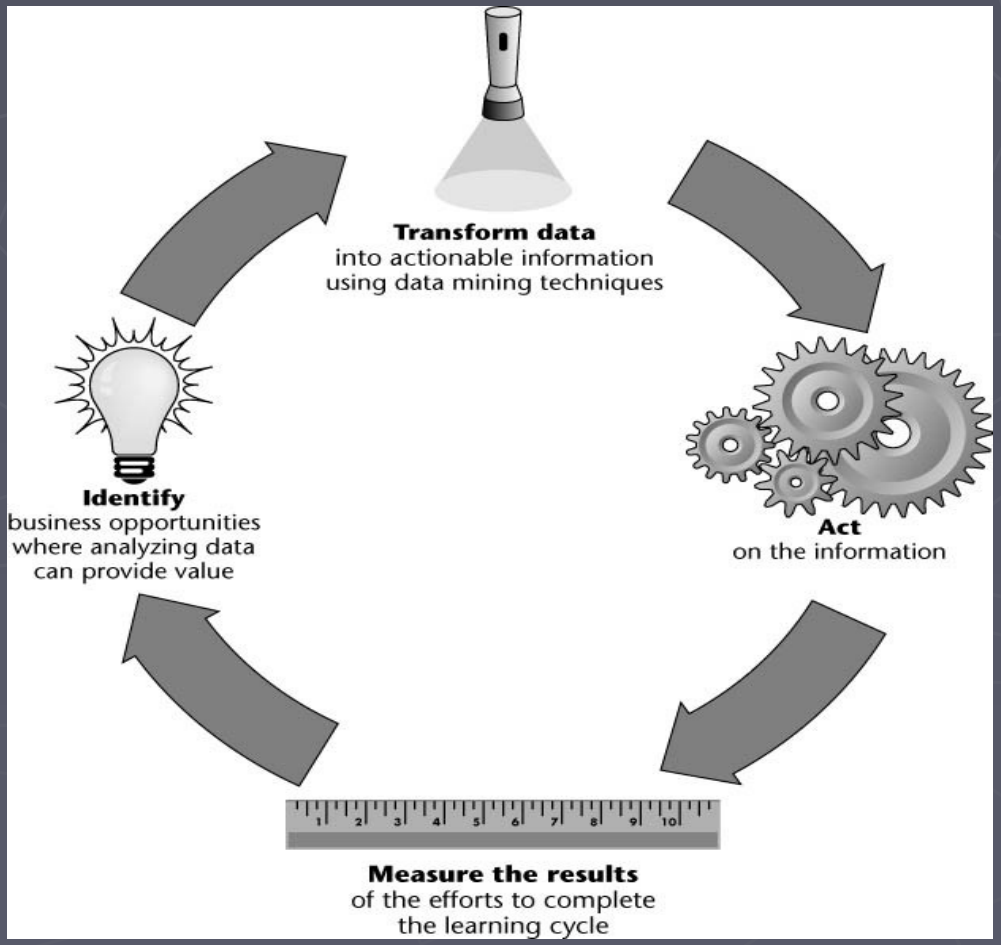
- Executives - need top-level insights and spend far less time with computers than the other groups.
- Analysts may be financial analysts, statisticians, consultants, or database designers.
- End users are sales people, scientists, market researchers, engineers, physicians, etc.

The Data Mining Process

Mining Technology is Just One Part



Data Mining Cycle



The DM Process

- Understand the Objective of the Data Mining Effort
  - Determine the Business Problem
  - Translate Business Problem into a Data Mining Objective

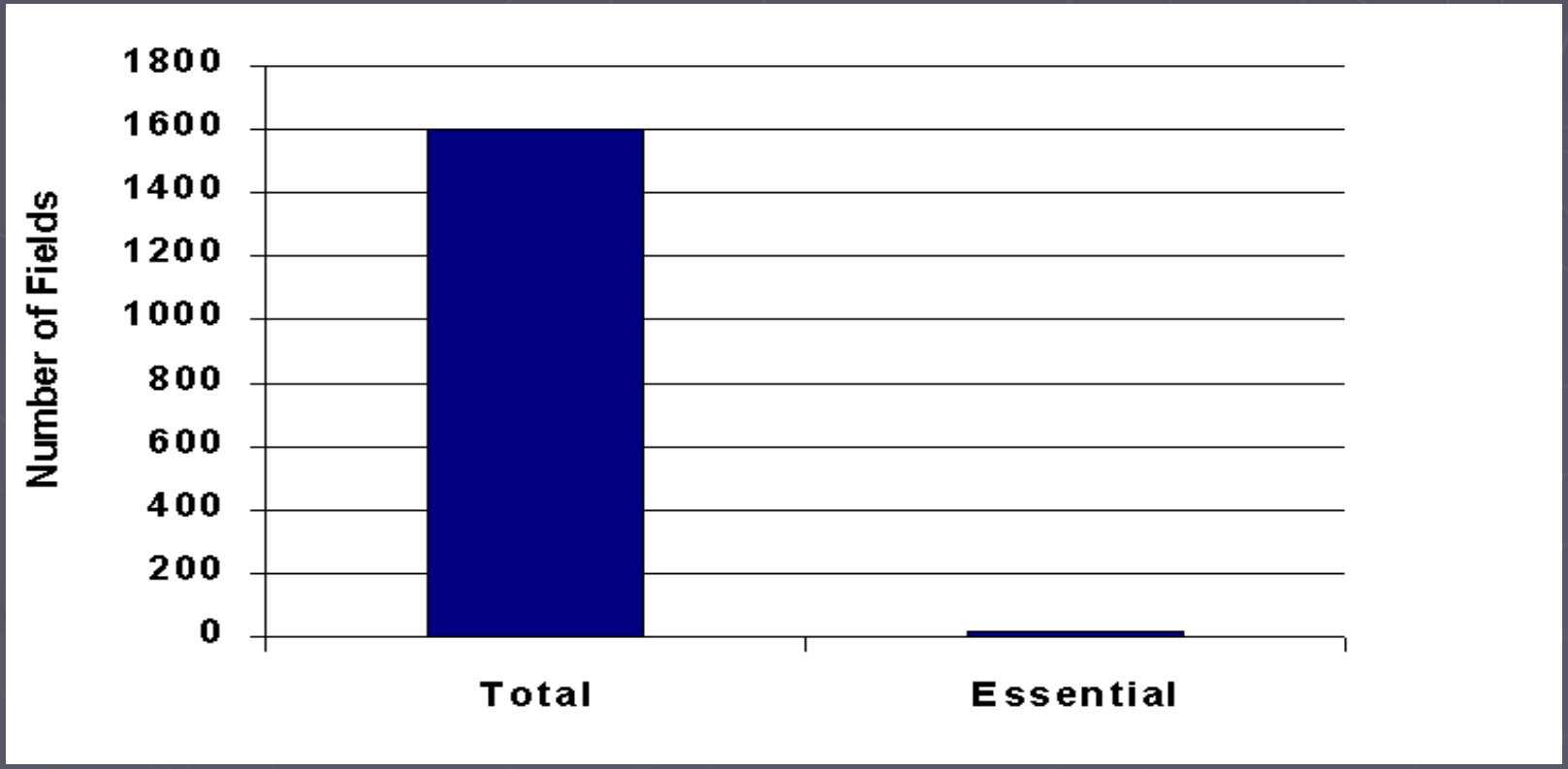
Data Mining Process

Data Preparation

Data is the Foundation for Analytics

- If you don’t have good data, your analysis will suffer
  - Rich vs. Poor
  - Good vs. Bad (quality)
- Missing data
- Sampling
  - Random vs. stratified
- Data types
  - Binary vs. Categorical vs. Continuous
  - High cardinality categorical (e.g., zip codes)
- Transformations

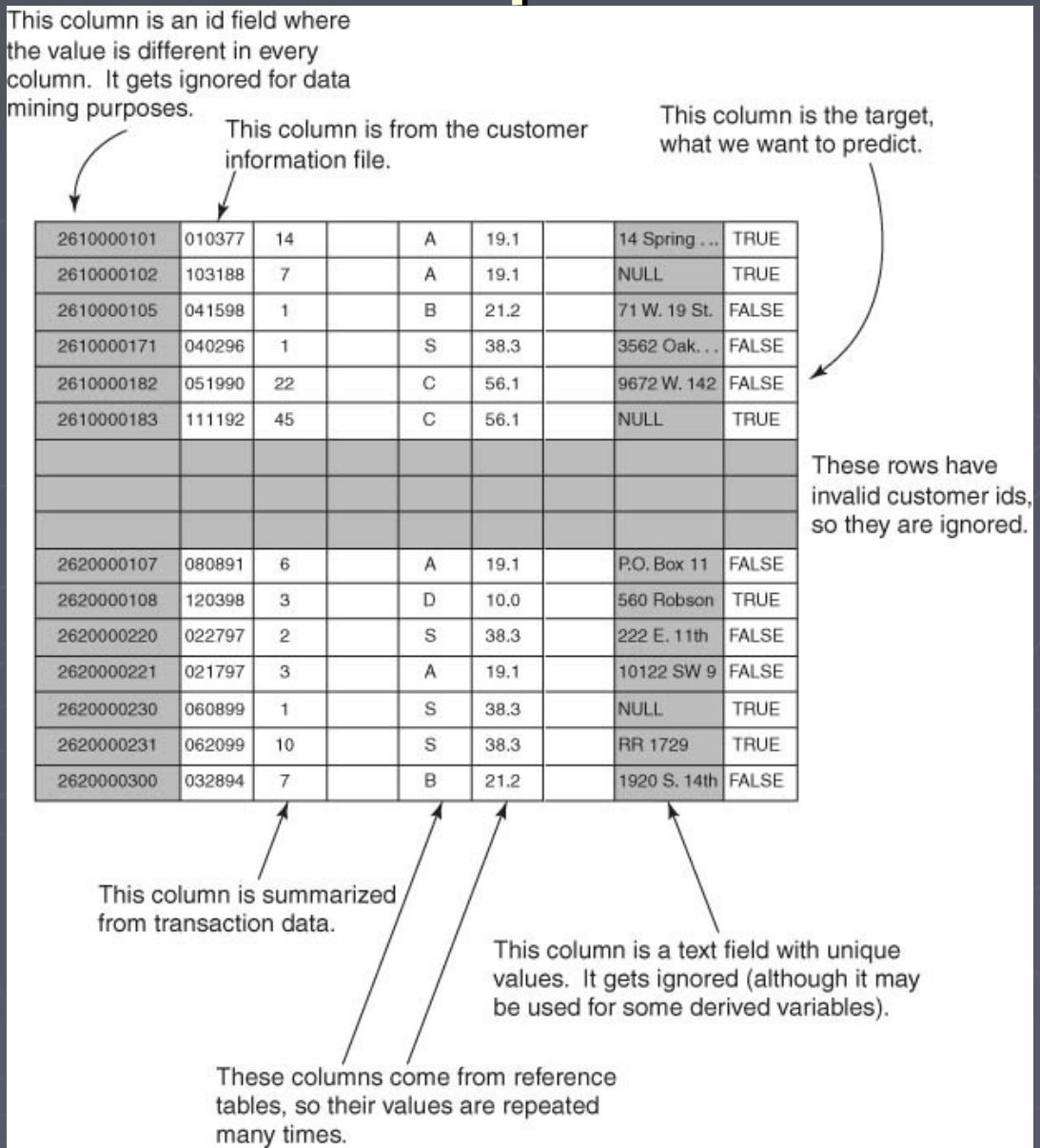
Don’t Make Assumptions About the Data



Data Preparation

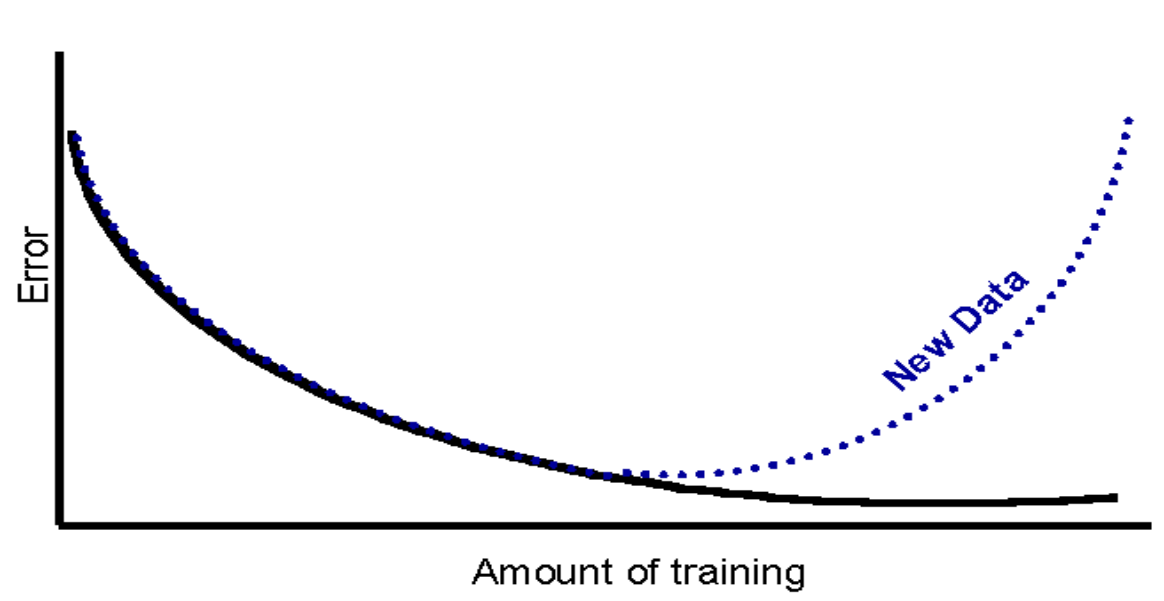
- Data preparation – takes usually over 90% of our time
  - Collection
  - Assessment
  - Consolidation and Cleaning
    - table links, aggregation level, missing values, transformations, etc
  - Data selection
    - active role in ignoring non-contributory data?
    - outliers?
    - Use of samples
    - visualization tools

Data Preparation

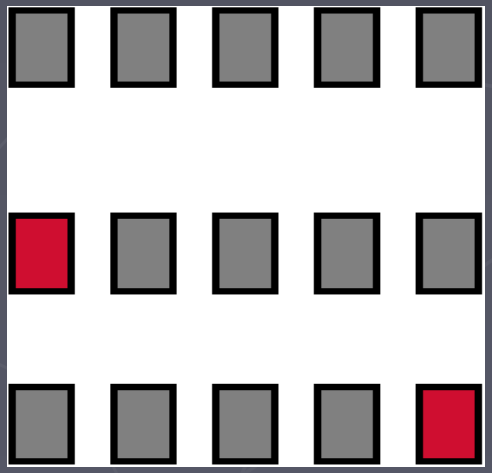


2610000101	010377	14		A	19.1		14 Spring . . .	TRUE
2610000102	103188	7		A	19.1		NULL	TRUE
2610000105	041598	1		B	21.2		71 W. 19 St.	FALSE
2610000171	040296	1		S	38.3		3562 Oak. . .	FALSE
2610000182	051990	22		C	56.1		9672 W. 142	FALSE
2610000183	111192	45		C	56.1		NULL	TRUE
2620000107	080891	6		A	19.1		P.O. Box 11	FALSE
2620000108	120398	3		D	10.0		560 Robson	TRUE
2620000220	022797	2		S	38.3		222 E. 11th	FALSE
2620000221	021797	3		A	19.1		10122 SW 9	FALSE
2620000230	060899	1		S	38.3		NULL	TRUE
2620000231	062099	10		S	38.3		RR 1729	TRUE
2620000300	032894	7		B	21.2		1920 S. 14th	FALSE

Data Preparation



Data Preparation



- Cross Validation
  - Break up data into groups of the same size
  - Hold aside one group for testing and use the rest to build model
  - Repeat

Data Mining Process

Data Mining Models

The DM Process

- Model building
  - an iterative process - different for supervised and unsupervised learning
    - Supervised Model
      - Driven by a real business problems and historical data
      - Quality of results dependent on quality of data
      - Want to build a predictive model
    - Unsupervised Model
      - Want to find groups of things with similar characteristics
      - Relevance often an issue
      - Useful when trying to get an initial understanding of the data
      - Non-obvious patterns can sometimes pop out of a completed data analysis project

Types of Models

- Are We Trying to Predict What Will Happen or Describe the State of the World?
- Descriptive modeling
  - Clustering
  - Association
  - Sequence discovery
- Predictive modeling
  - Classification
  - Regression
  - Time series

Types of Models

- Descriptive Models for Clustering and Associations
  - Clustering algorithms: K-means, Kohonen
  - Association algorithms: apriori, GRI
- Prediction Models for Regression and Classification
  - Regression algorithms: neural networks, rule induction, CART (OLS regression, GLM)
  - Classification algorithms: CHAID, C5.0 (discriminant analysis, logistic regression)
- Some models are better than others
  - Accuracy
  - Understandability

Determining the Model

- Hierarchy of Data Mining Solutions:
- Business Goal
  - Data Mining Goal -
    - Type of model (predictive, descriptive)
      - Algorithm
- Once the Data Mining Goal of the Data Mining Effort is Determined, the Technique Used to Meet that Goal Falls into Place

Determining the Model

- Hierarchy of Data Mining Solutions

Step Num	Step in Hierarchy	Value Chosen
Step 1	Business Goal	**Increase Revenues**
Step 2	Data Mining Goal	**Identify customers who are likely to buy**
Step 3	Type of model	**Predictive**
Step 4	Algorithm	**Rule Induction**

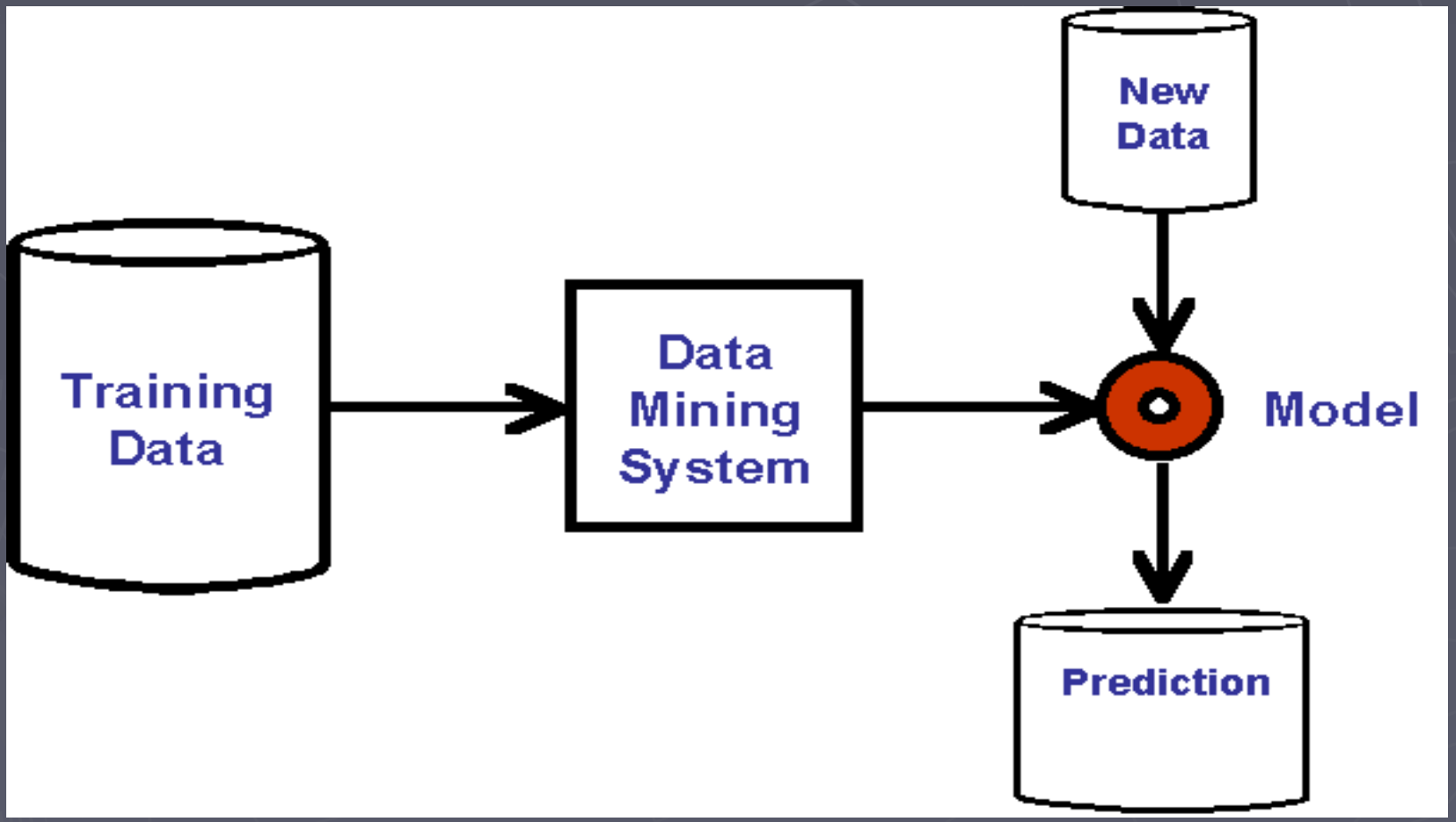
Model Building

- The choice of model depends on model type which will influence data preparation step
- The essence of model building is to train the model with a subset of data, test it with an independent subset
- Some algorithms inherently test as part of model construction:
  - e.g., neural networks, rule induction
- Others require user designation of train/test data

Model Building

- Evaluation of model: how well it performed on test data
- Methods and criteria depend on model type:
  - e.g. mean error rate with regression models
- Interpretation of model: important or not, easy or hard depends on algorithm

How are Data Mining Models Built and Used?



The DM Process

- Use the model on application
  - deploy models on line
  - on the web
  - against the database
- Model monitoring
  - Determine if the model still ‘works’

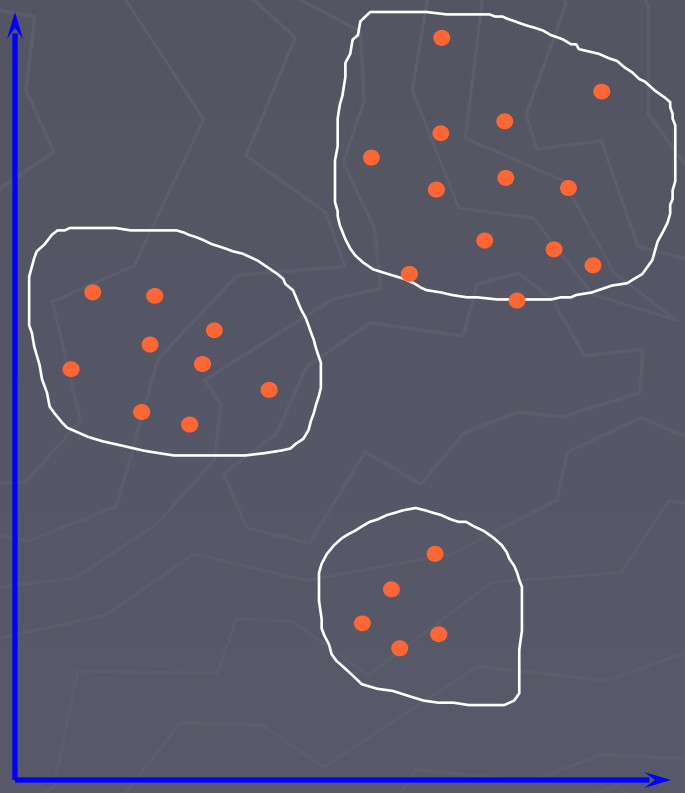
Data Mining Process

Data Mining Models

Types of Models

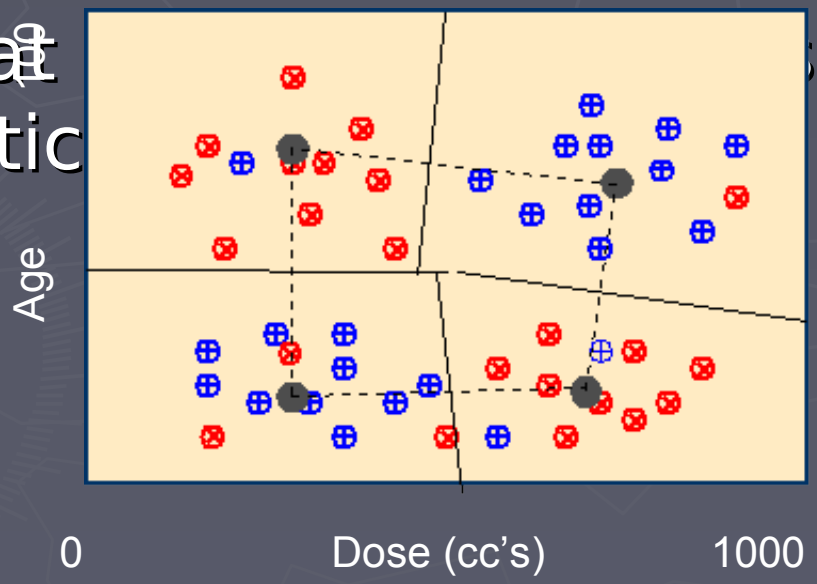
Clustering

- Divide the data into a number of different groups
- Determine the attributes that characterises a group automatically
- Can be used for classification of new cases



K-Means Clustering

- User starts by specifying the number of clusters (K)
- K data points are randomly selected
- Repeat until convergence criterion is satisfied
  - Assign each patient to the group that has the closest centroid
  - When all patients have been assigned, recalculate the positions of the K centroids
- Convergence criterion is usually some specific clustering statistic



Association

- Identifies the items that occur together in a given event
  - If ‘A’ occurs then x% (confidence factor) of the times ‘B’ occurs. This is found in y% (Support) of the data.
- Used for ‘Market Basket Analysis’

Association Rules

- Finds relations among *attributes* in the data that frequently co-occur .
  - E.g., Association Rules. Popular for basket analysis

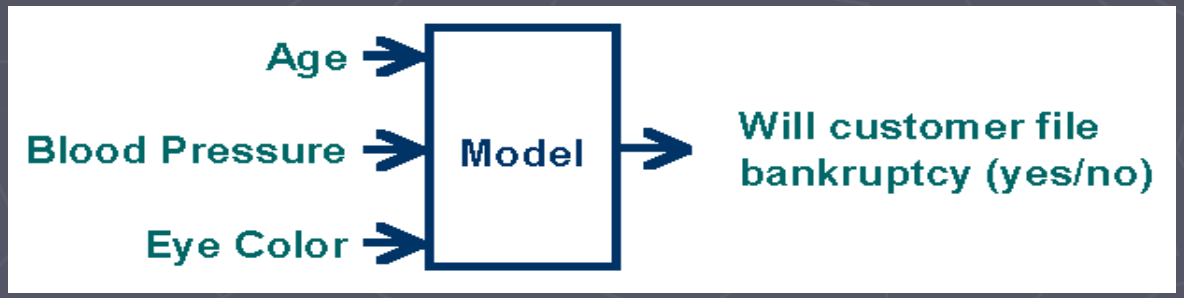
Buy diapers on Friday night → **Then** → **Buy beer**

Transaction No.	Day	Item 1	Item 2	Item 3	…
100	Fri	Beer	Diaper	Chocolate	
101	Fri	Milk	Chocolate	Shampoo	
102	Thu	Beer	Wine	Vodka	
103	Fri	Beer	Cheese	Diaper	
104	Fri	Ice Cream	Diaper	Beer	

Sequence discovery

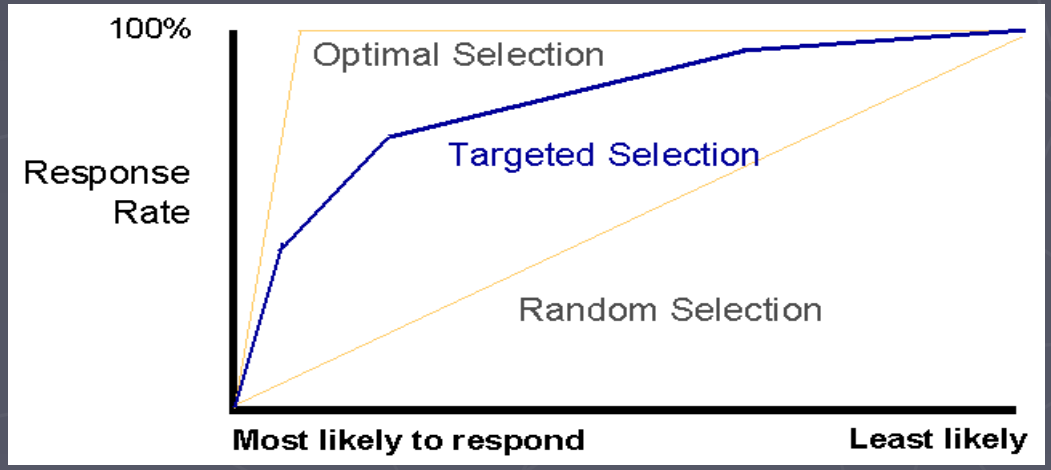
- Finds association among the items
- The related items are spread over the time
  - If surgical procedure ‘A’ is followed, then the infection ‘B’ occurs at a later time
  - If stock ‘A’ rises by 12% then stock ‘B’ also rises by 8% within 2 days
- Requires information on transactions and transactors

Predictive Modeling



- A “black box” that makes predictions about the future based on information from the past and present
- Large number of inputs usually available

How Good is a Predictive Model?



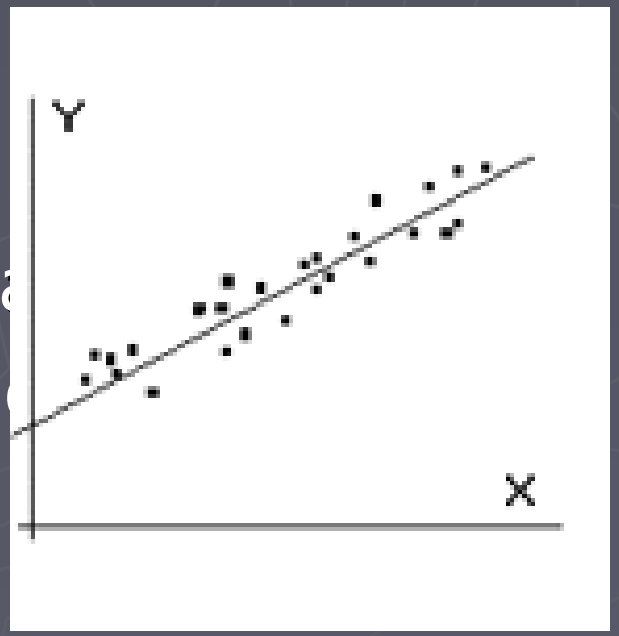
- Response curves
  - How does the response rate of a targeted selection compare to a random selection?

Classification

- Goal of classification is to build structures from examples of past decisions that can be used to make decisions for unseen cases.
- Predicts the cluster in which a new case fits in
- The characteristics of the groups can be defined by an expert or fed from historic data
- Often referred to as supervised learning.
- Decision Tree and Rule induction are popular techniques
- Neural Networks also used

Regression

- Forecasts the future values based on existing values
- Types
  - Simple - one independent va
  - Multiple - more than one in variables



Example 1: Regression

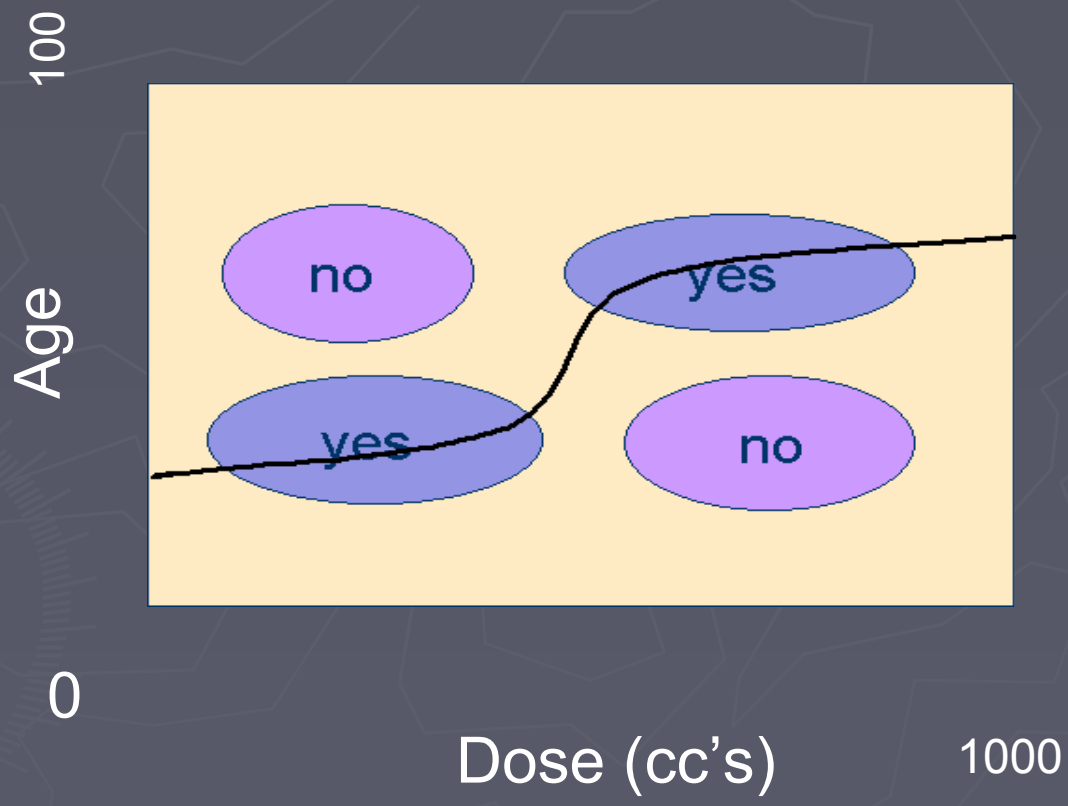
Name	Income	Age	Order Amt
a	23000	30	83
b	51100	40	131
c	68000	55	178
d	74000	46	166
e	23000	47	117

This will result into Linear graph

Pattern:

Amount=
0.001*__Income__+2*__Age__

Example 2: Regression



Time series

- Forecasts the future trends
- Model includes time hierarchy like year, quarter, month, week etc.
- Considers impact of seasonality, calendar effects such as holidays
- What is the expected price for Microsoft’s stock by the end of this year?

Data Mining Techniques

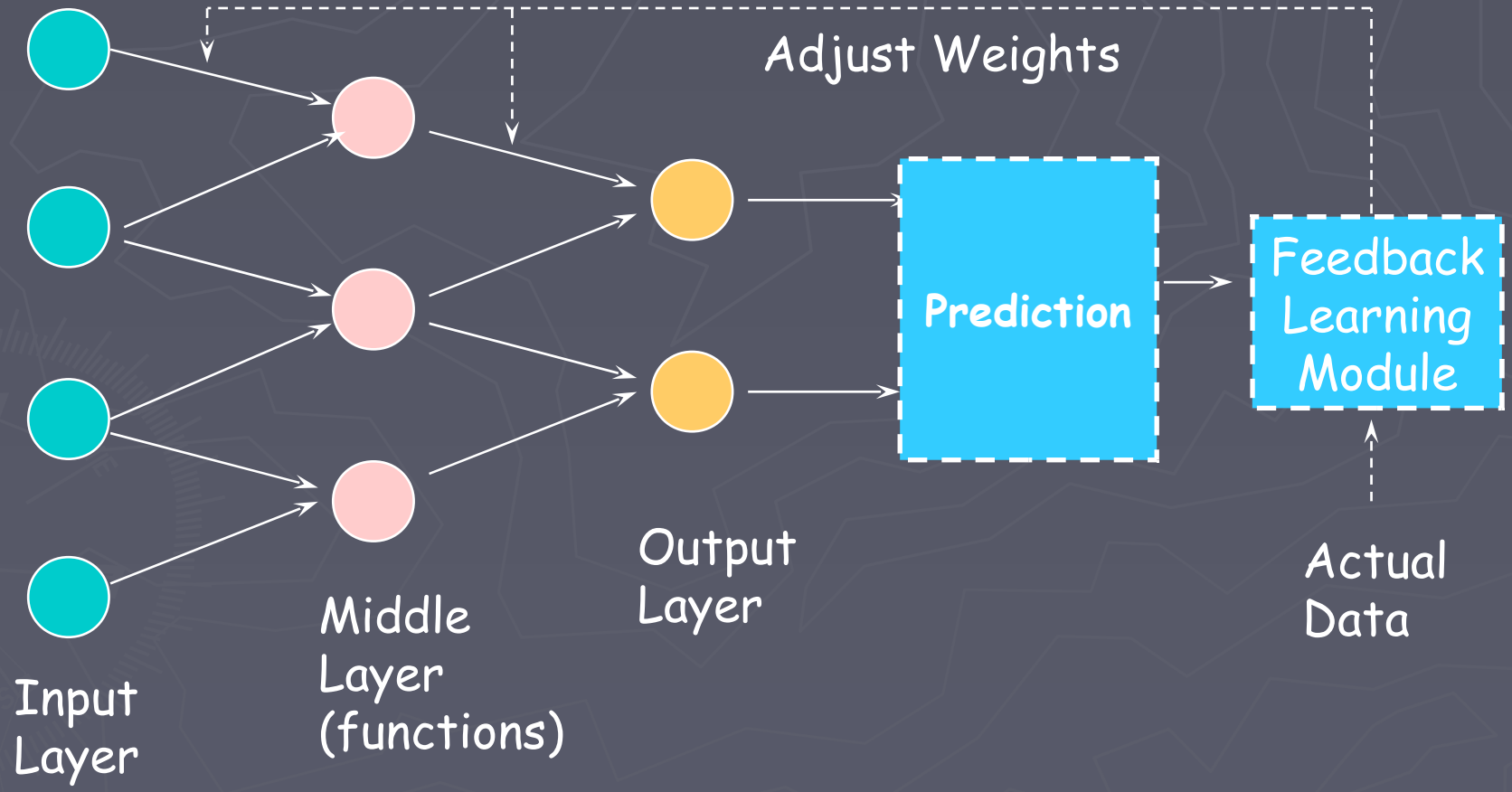
Techniques

- Neural Networks
- Decision Trees
- Rule Induction
- K nearest neighbour

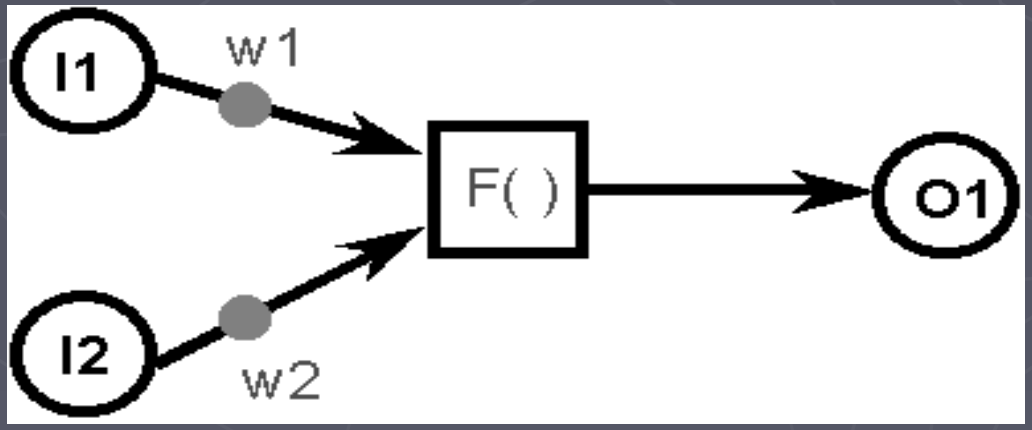
Neural Networks

- Parameter adjustment systems
- Interconnected elements (neurons)
- Train the net on a training data set
- Use Trained net to make predictions
- Can deal with only numeric data

Neural Networks - functioning

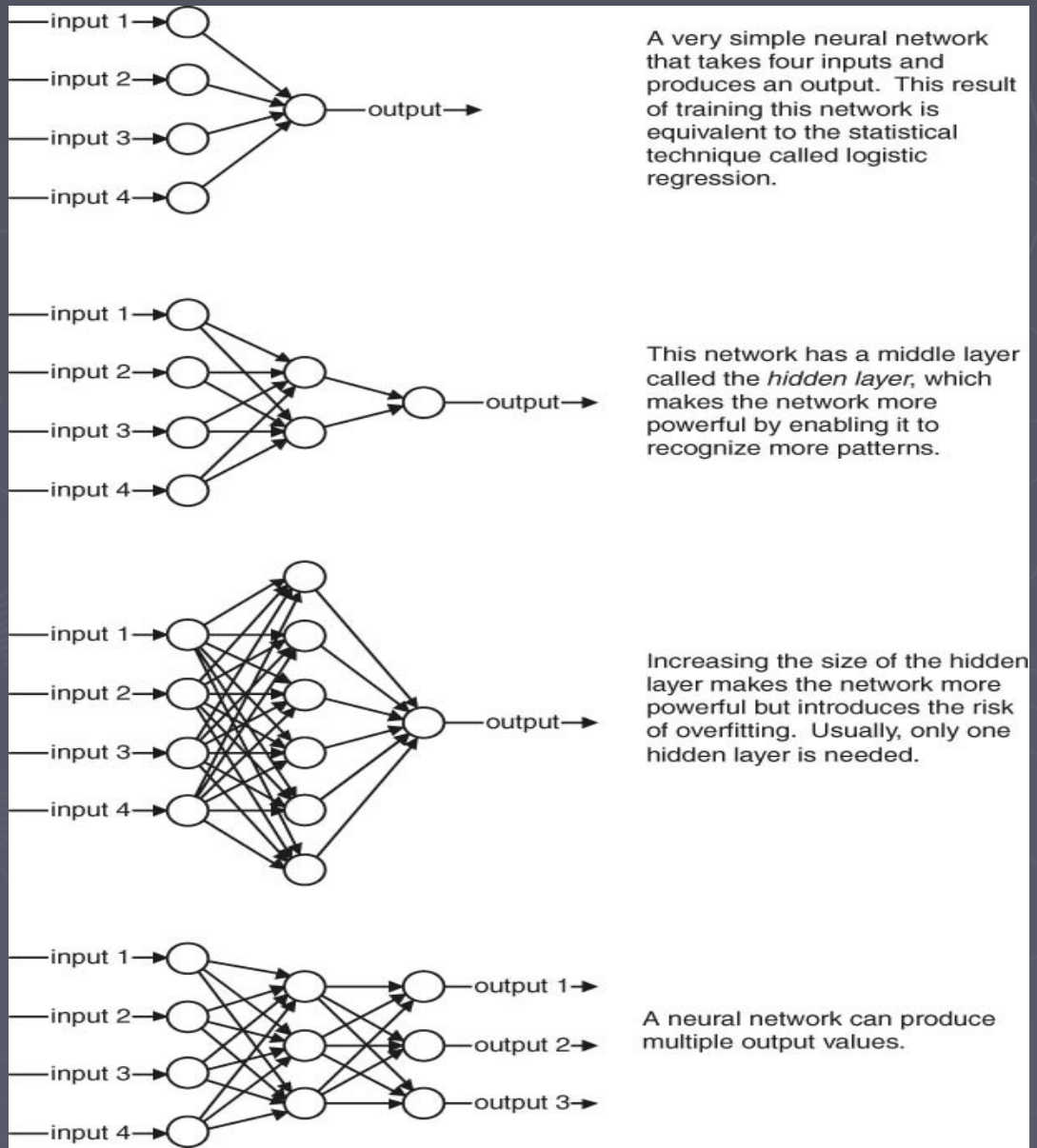


(Feed Forward) Neural Networks



- Very loosely based on biology
- Inputs transformed via a network of simple processors
- Processor combines (weighted) inputs and produces an output value
- Obvious questions: What transformation function do you use and how are the weights determined?

Types of Neural Networks



Applications and Issues

- Neural Networks can be used for predictive modeling
- Kohonen networks - a type of neural networks can be used for clustering
- Drawback - Activities are completely black box
- Key problem: Difficult to understand
  - The neural network model is difficult to understand
  - Relationship between weights and variables is complicated
    - Graphical interaction with input variables (sliders)
  - No intuitive understanding of results
- Training time
  - Error decreases as a power of the training size
- Significant pre-processing of data often required

Decision Trees

- A way of representing series of rules
- Trees are grown through an iterative splitting of data
- Goal is to maximize the distance between groups at each split
- Can have two branches (binary tree) or more than two

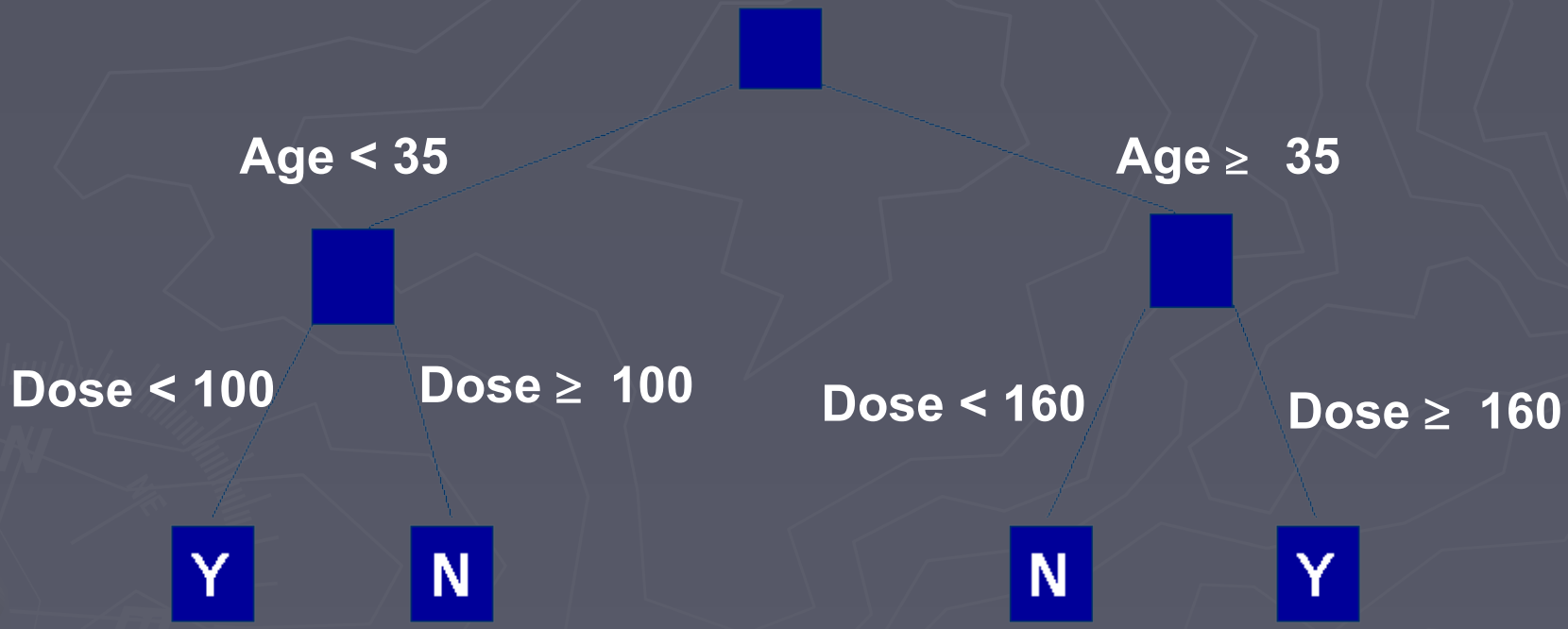
Decision Trees

- Each node is called a decision node
- A branch leads to another decision node or to the bottom of the tree
- By moving along the tree one can reach a decision by deciding at each node which branch to take

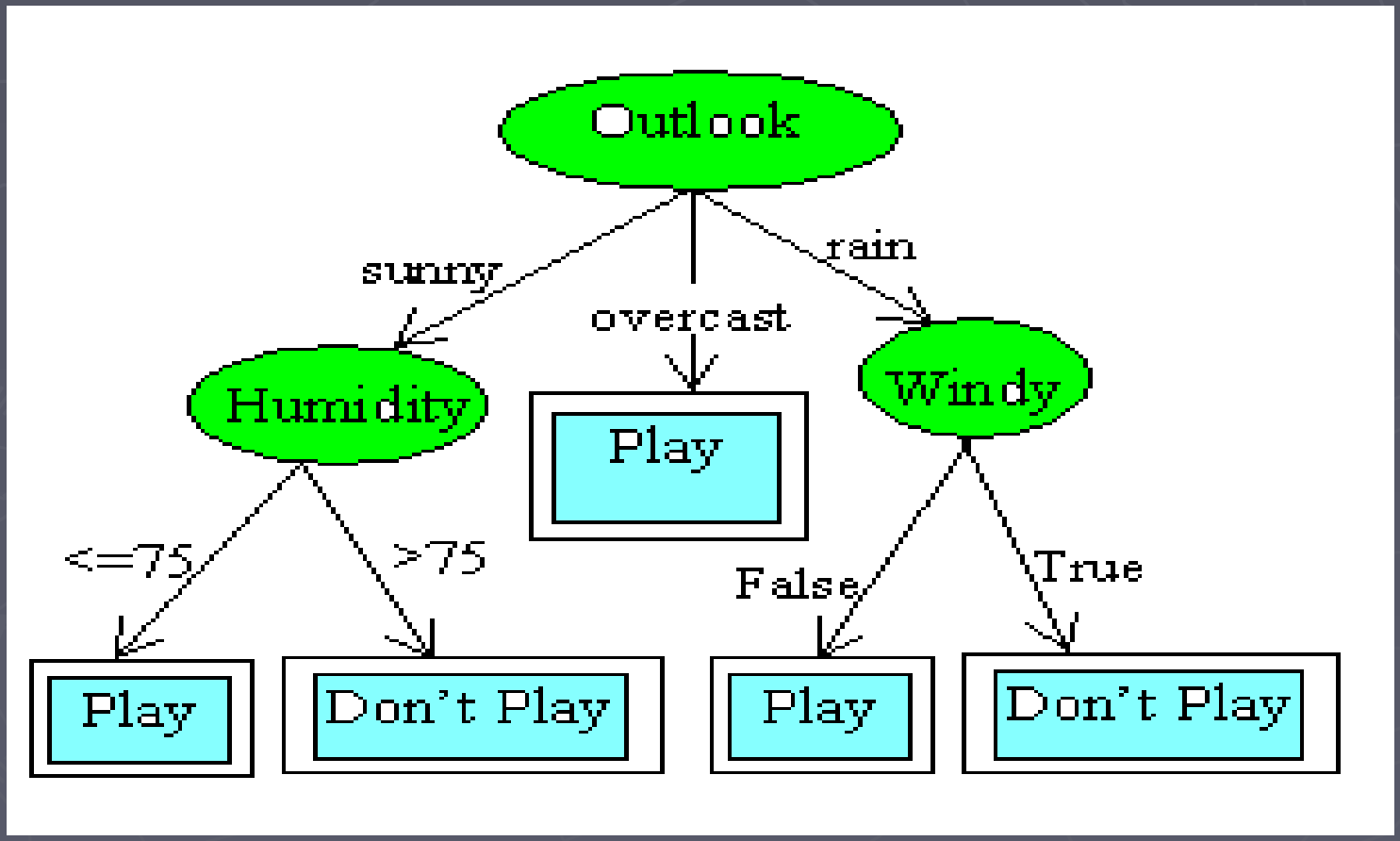
Decision Trees

- Decision trees are of two types
  - Classification trees
    - used to predict categorical values
  - regression trees
    - used to predict continuous variables

Decision Trees : Examples



Decision Trees : Examples



Advantages & Limitations

- Advantages
  - Models can be built very quickly
  - Suitable for large data sets
  - Easy to understand
  - Gives reasons for a decision taken
  - Handle non-numeric data very well
  - Minimum amount of data transformation
- Limitations
  - Leads to an artificial sense of clarity
  - Trees left to grow without bound take longer to build and become unintelligible
  - May overfit the data
  - Algorithms used for splitting are generally univariate - using single independent variable at a time

Rule Induction

- Extraction of useful if-then rules from data based on statistical significance.
- Completely a machine driven process.
- Can discover very general rules which deal with both numeric and non-numeric data.
- Translating the rules into a usable model must be done either by the user or a decision trees interface

Rule Induction Examples

- If Car = Ford and Age = 30…40
  Then Defaults = Yes ,Weight = 3.7
- If Age = 25…35 and Prior_purchase = No
  Then Defaults = No, Weight = 1.2
- Not necessarily exclusive (overlap)
- Start by considering single item rules
  - If A then B
    - A = Missed Payment, B = Defaults on Credit Card
  - Is observed probability of A & B combination greater than expected (assuming independence)?
    - If It is, rule describes a predictable pattern

Rule Induction: Important points

- Look at all possible variable combinations
- Compute probabilities of combinations
- Expensive!
- Look only at rules that predict relevant behavior
- Limit calculations to those with sufficient support
- When moving onto larger combinations of variables like n3, n4, n5, ...
  - Support decreases dramatically, limiting calculations

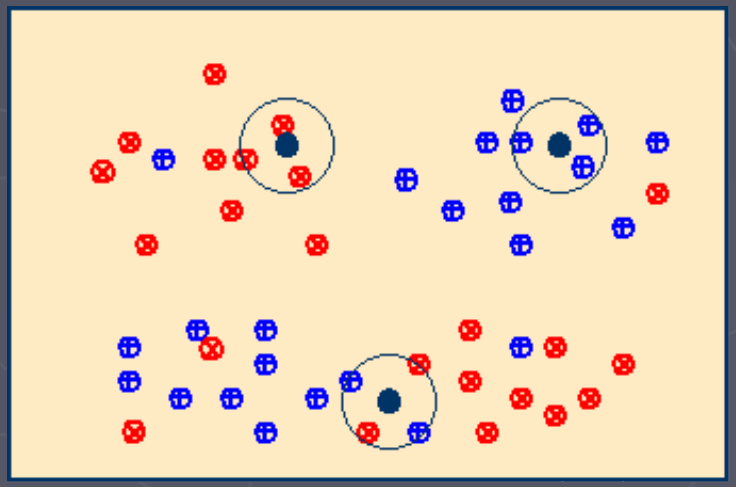
K Nearest Neighbor

- A classification technique
- Decides in which class to put a new case in
- Criterion is to find a maximum number (k) of neighbors having most similar properties
- Assigns a new case to the same class to which most of the neighbors belong

K Nearest Neighbor

- Set of already classified cases selected to use as a basis.
- Neighborhood size, in which to do the comparisons, decided.
- Decided how to count the neighbors (assigning weights to neighbors, may give more weight to nearer neighbor than a farther one).

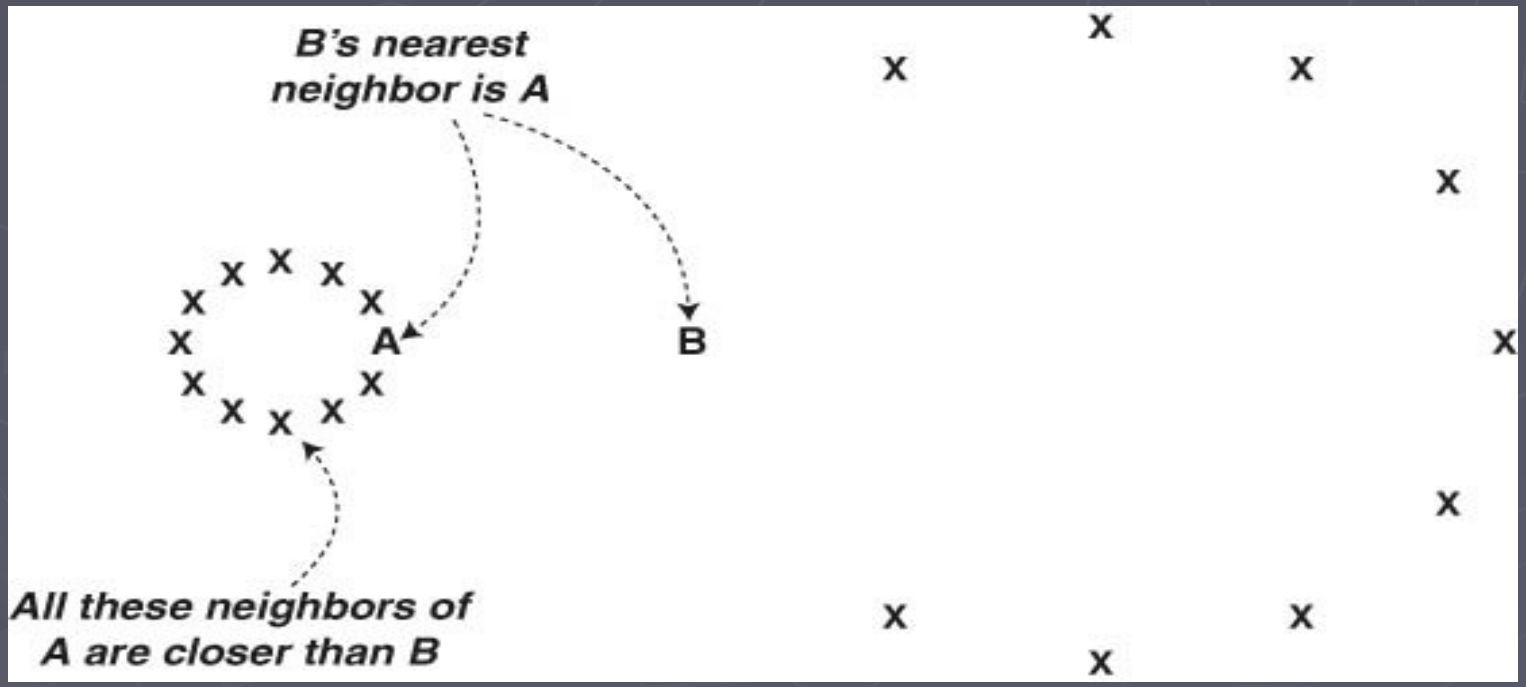
K Nearest Neighbor Model



1000

- Use entire training database as the model
- Find nearest data point
- Very easy to implement. More difficult to use in production.
- Disadvantage: Huge Models

K Nearest Neighbor Diagram



Developing a Nearest Neighbor Model

- Model generation:
  - What does “near” mean computationally?
  - Need to scale variables for effect
  - How is voting handled?
- Confidence Function
- Conditional probabilities used to calculate weights
- Optimization of this process can be mechanized
- Developing a Nearest Neighbor Model

Example of a Nearest Neighbor Model

- Weights:
  - Age: 1.0
  - Dose: 0.2
- Distance = $\sqrt{\Delta \text{ Age} + 0.2 \times \Delta \text{ Dose}}$
- Voting: 3 out of 5 Nearest Neighbors (k = 5)
- Confidence = 1.0 - D(v) / D(v')

Comparison of Algorithms

- K Nearest Neighbor
  - Quick and easy
  - Models tend to be very large
- Neural Networks
  - Difficult to interpret
  - Can require significant amounts of time to train
- Rule Induction
  - Understandable
  - Need to limit calculations
- Decision Trees
  - Understandable
  - Relatively fast
  - Easy to translate into SQL queries

Supervised / Unsupervised

- Supervised Learning:

1. As the name applies Techniques are supervised to work or perform. These models are first TRAINED using data whose RESPONSE variable or result is already KNOWN.
2. Predictive models (classification, regression etc) fall under this category –They have to train and then test. Estimated value is to compared with known value

- Unsupervised Learning:

1. As the name applies Techniques are unsupervised. There is no already known result to guide the algorithm.
2. Descriptive Techniques clustering, Association etc falls under this category .

Data Mining Applications & Tools

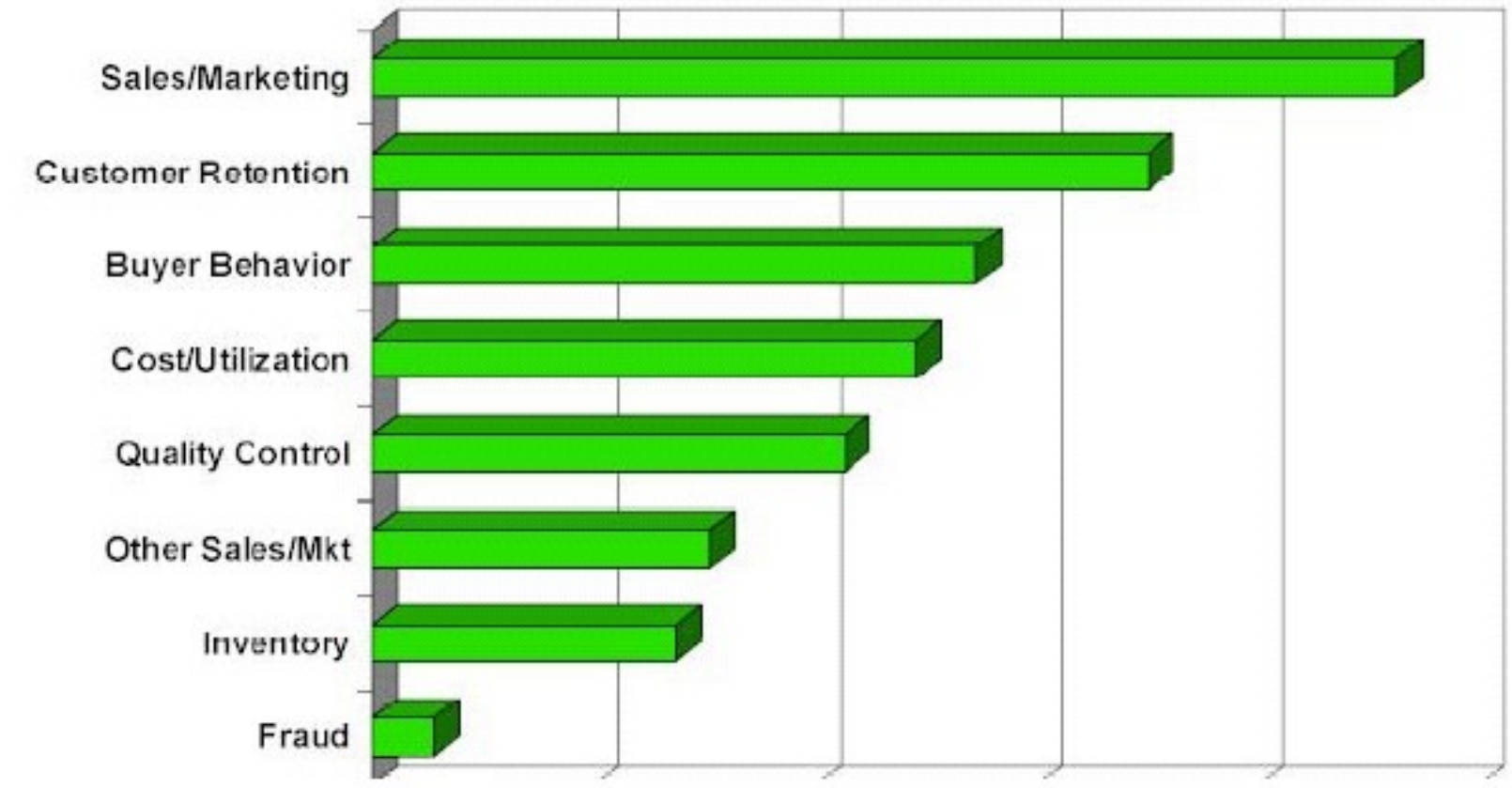
Applications

- Customer Profiling
- Target Marketing
- Market Basket Analysis
- Fraud Detection
- Medical Diagnostics
- Direct mail marketing
- Web site personalization
- Bioinformatics
- Anti Money Laundering
- Churn Analysis

Some Uses of DM

- DM can be used to discover new purchasing trends, plan investment strategies, and detect unauthorized expenditures in the accounting system.
- It can improve marketing campaigns and the outcomes can be used to provide customers with more focused support and attention.
- Many law enforcement and special investigative units, whose mission is to identify fraudulent activities and discover crime trends, have also used data mining.
- Like identifying the critical behavior patterns in the communication interactions of narcotics organization
- The monetary transactions of money laundering and insider trading operations
- The movement of serial killers, and the targeting of smugglers at border crossings.
- Government that maintains large data sources as part of activities relating to matters of national security.

Applications



Source: IDC 1998

CHURN ANALYSIS

- Road Map for Minimizing Churn Rate

  Churning means switching – A PROBLEM

  Increasingly competitive environment customer retention has surfaced as one of the key problem faced by mobile service provider .

Business Objective: TO MINIMISE CHURN RATE

CHURN ANALYSIS

Data Mining Goal

IDENTIFIED CUSTOMER WITH DELIQUENT NATURE.

Scope

- Assign Churn Score to all customers in order to identify those who are most likely to churn (Quarter etc).
- Determine the most relevant parameter that influences the inclination to churn.
- Define Clearly segments that are strongly divided by their churn relating Behavior

CHURN ANALYSIS

- Basic Understanding

There are mainly two types of churn

1. Customer Request (CRQ)

2. Forced Churn (Defaulters)

CHURN ANALYSIS

- Information Sources
  - Call Statistics (CDR)
  - Credit History
  - Billing History
  - Revenue History
  - Payment History
  - Survey Data
  - Demographic data
  - Complaint information

CHURN ANALYSIS

- Suggested Analysis

Pareto analysis

Also called 80/20 Analysis. Its been observed that 80% of the revenue profit comes from 20 % of the customer. Key Business Improvement was identifying those 20% and serves them better.

Techniques/ Reports/Algorithms

Characterization and summarization, Top 10 report , List ,Cross tab Reports , Graph Charts etc.

CHURN ANALYSIS

- Suggested Analysis

Loyalty Analysis

A loyal customer is worth new customer. If it is possible to identify the loyal customer and increase that volume. A loyal customer is defined as the one who is with the company for last six months. This analysis will give insight in to the complete details of various customer bases.

Techniques/ Reports/Algorithms

Characterization and summarization, Top 10 report , List ,Cross tab Reports , Graph Charts etc.

CHURN ANALYSIS

- Suggested Analysis

Customer Profit Analysis

lentifying wining and loosing customer. A wining customer is one who giving increasing revenue month after month and vice versa. Identify he characteristic and reason for better decision.

Techniques/ Reports/Algorithms

Characterization and summarization, Top 10 report , List ,Cross tab Reports , Graph Charts etc.

CHURN ANALYSIS

- Suggested Analysis

Trend Analysis

t's a Visualization Technique. This Technique uses parallel Coordinate system to show the trend of various measures over different time period

Techniques/ Reports/Algorithms

Parallel Coordinate graph

CHURN ANALYSIS

- Suggested Analysis

Customer profiling

ctive accounts, Light user, risky customer Active accounts, Loss makin
ofit making accounts. This segment helps in mapping with the predictiv
gment.

Techniques/ Reports/Algorithms

List, Cross Tab, clustering , Graph Charts

CHURN ANALYSIS

- Suggested Analysis

LTV Analysis

Called Life Time value Analysis .Revenue projected over 25 yrs and Projected Churning loss and rate.

Techniques/ Reports/Algorithms

List report, Line Graphs, graph Charts

CHURN ANALYSIS

- Suggested Analysis

Churn Modeling

**t of acquiring new customer is more than retaining one. Classification
urn models. Assign churn score and predictive modeling.**

Techniques/ Reports/Algorithms

Scoring, Decision Tree, neural network, Clustering.

CHURN ANALYSIS

- Suggested Analysis

Survival Analysis

This predicts how long the customer would continue with existing service in terms of time. What measures can be taken. One of the Popular Technique are K. Hazard Analysis .

Techniques/ Reports/Algorithms

K.Hazard technique

CHURN ANALYSIS

Suggested Approach

Derived Customer Segmentation mapped them against their probability of churning and expected Profits. Each Segment has significantly different characteristics, demographic models.

Assign Channel and Campaign based on channel selection.

Note

Data Mining is an iterative process. Next Step depends upon the Outcome of previous result. There cannot be fixed approach.

Tools and Vendors

- IBM - Intelligent miner
- SPSS - Clementine
- SAS - Enterprise miner
- SGI - Mine Set
- Accrue - Decision Series
- Poly Analyst – Megaputer.

Comparison of Tools

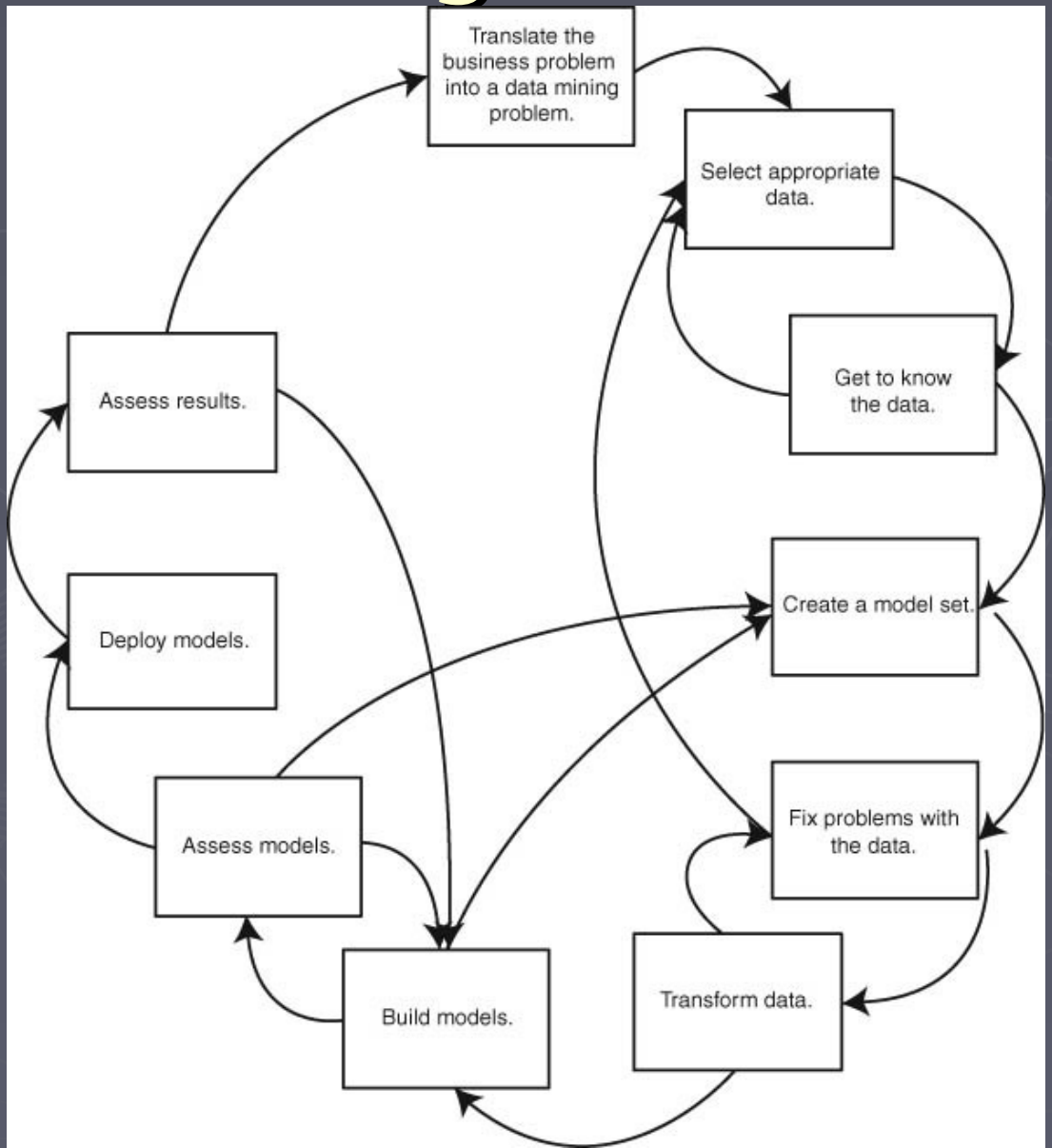
	Algorithms										Platforms		
Tool	NN	Tree	Naïve Baye	k-Mns	k-NN	Stats	Pred	Time Series	Clust	Assoc	Win 32	UNIX	Par
IBM – Intellige Miner	√	√			√		√	√	√	√	√	√	√
SPSS – Clementine	√	√					√	√	√	√	√	√	
SAS – Enterpri Miner	√	√				√	√	√	√	√	√	√	
SGI –MineSet		√	√	√			√		√	√	√	√	√
Accrue Softwar Decision Serie	√	√	√				√		√	√	√	√	√

Trends and Market Updates

- ACRM (Analytical Customer Relationship Management) ,Web Mining , ITM (Intelligent Transaction Mining) and Text Mining.
- Anti Money Laundering
  1. An increased criminal (Terrorist) activity has become the major concern with most of the countries of the world. To Track the funds being used for terrorist activities has become the major application and challenge in a Finance Sector.
- CIBIL
  1. It's the first credit information bureau being established in India 2003. CIBIL will obtain and Share data on borrowers both consumer and commercial for sound credit decision therefore helping to avoid adverse selection.
  2. Availability of credit information facilitates credit scoring mechanism and Credit Risk Analysis will play an important role in that
- Basel II
  1. As per the Basel II Accord which serves as a guideline for the banks across 32 countries to reduce credit risk, Operation risk and business risks. By 2007 all banks should have data warehouse in place so that information should be available for risk related analysis.

Data Mining Methodology

Data Mining Methodology



Methodologies

- Discussion on following methodologies
  - CRISP-DM methodology
  - Wipro - Data mining methodology

CRISP - DM

- CRoss Industry Standard Process model - Data Mining
- Hierarchical approach
  - Phase
  - Generic Task
  - Specialised Task
  - Process Instance

CRISP - DM

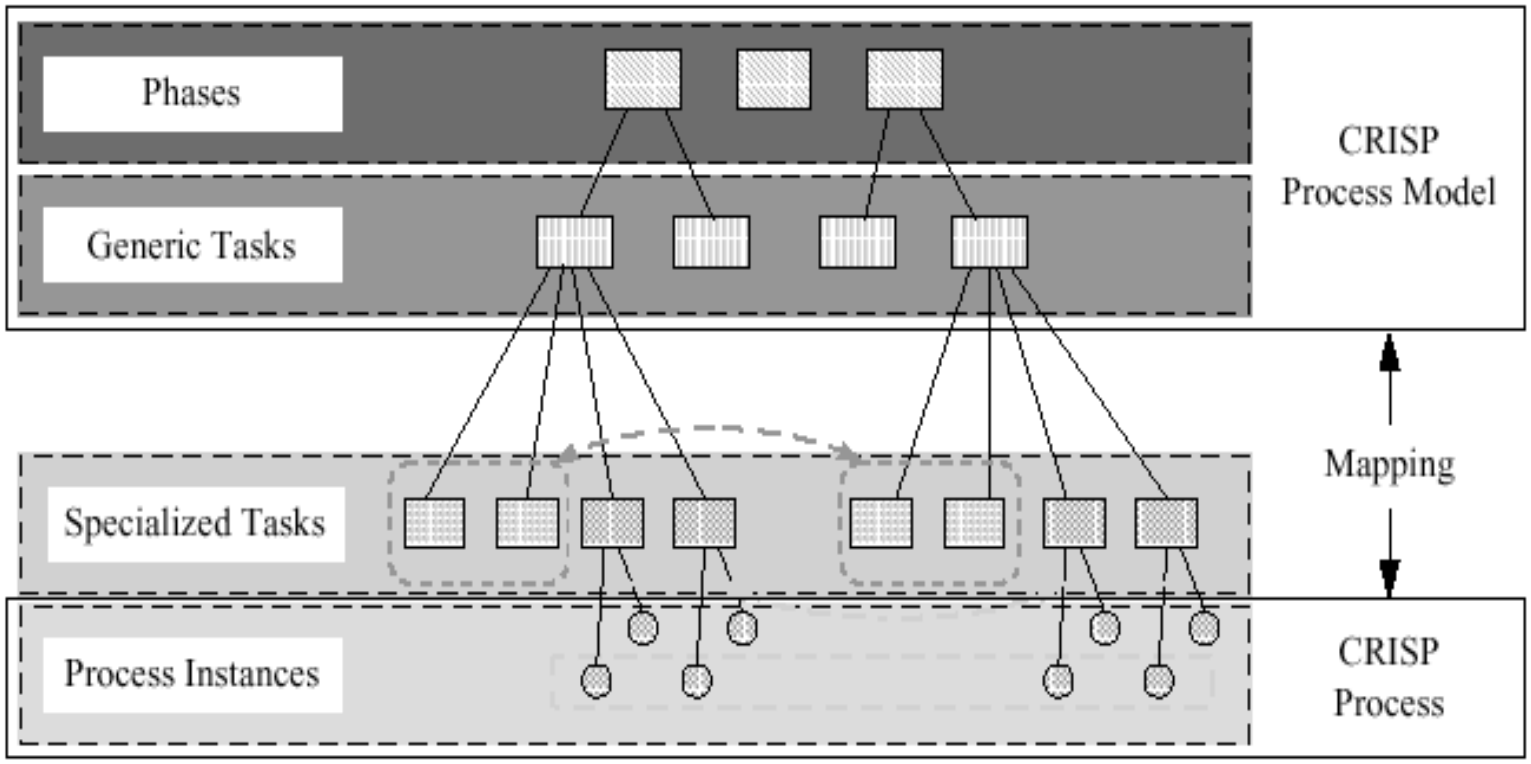
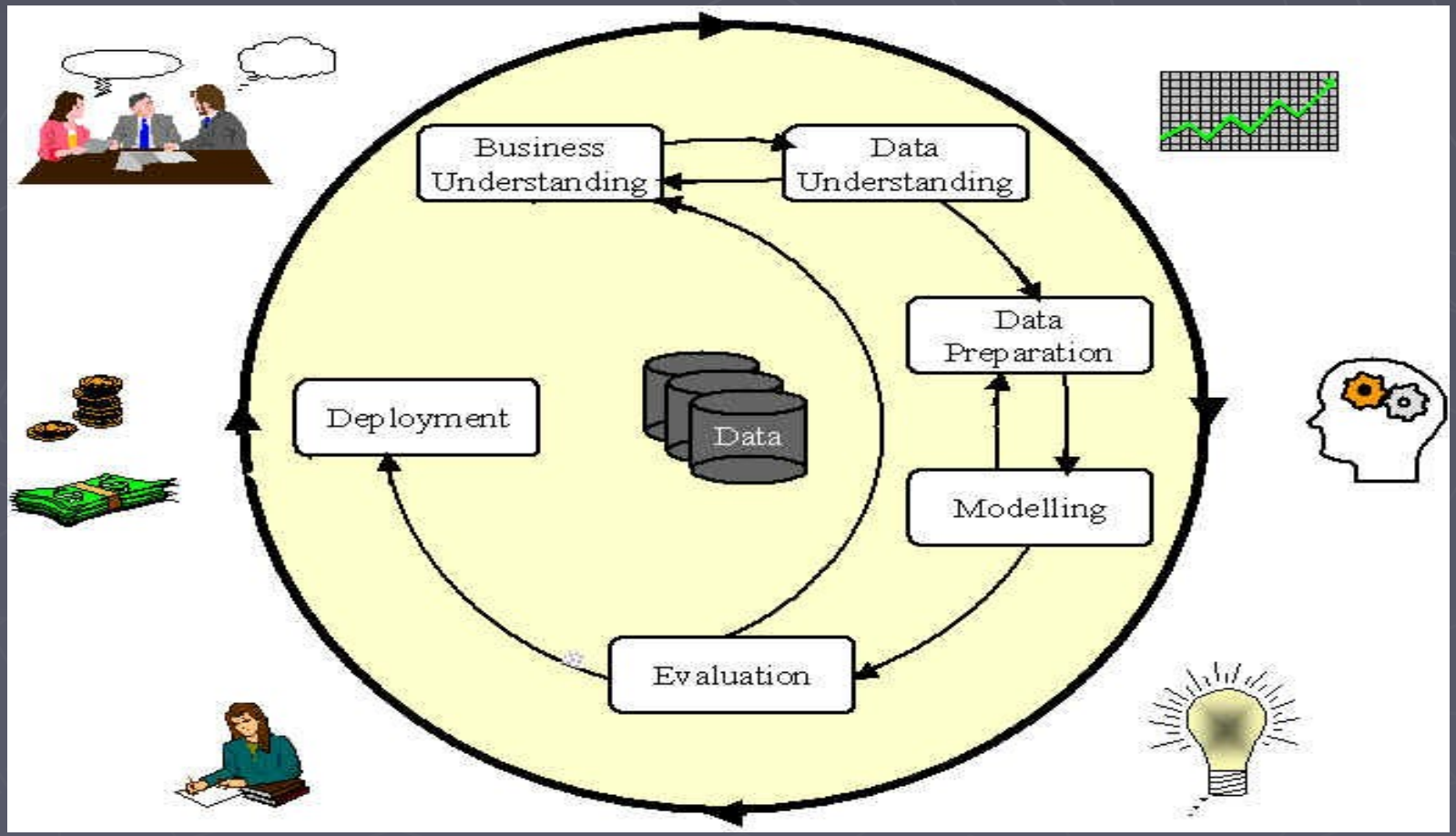


Figure 1: Four Level Breakdown of the CRISP-DM Methodology

Mapping Generic to Specific Model

- Application domain
  - The specific area in which the data mining project takes place
- Data mining problem type
  - The specific class(es) of objective(s) which the data mining project deals with
- Technical aspect
  - covers specific technical issues in data mining
- Tool and technique

The Cycle


Business Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment
Data

Phases

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Business Understanding

- Understanding the project objectives and requirements from a business perspective
- Converting this requirements into a data mining problem definition
- Preliminary plan designed to achieve the objectives.

Data Understanding

- Initial data collection
- Identifying data quality problems
- detect interesting subsets to form hypotheses for hidden information.

Data Preparation

- Identification of data at Table, Record, and Attribute level
- Transformation and cleaning of data for modeling tools.
- Data preparation tasks are performed multiple times

Modeling

- Identification of the suitable modeling techniques for the requirement
- Applying the various possible models on the data set
- Requires stepping back to the data preparation phase, due to the model specific requirements

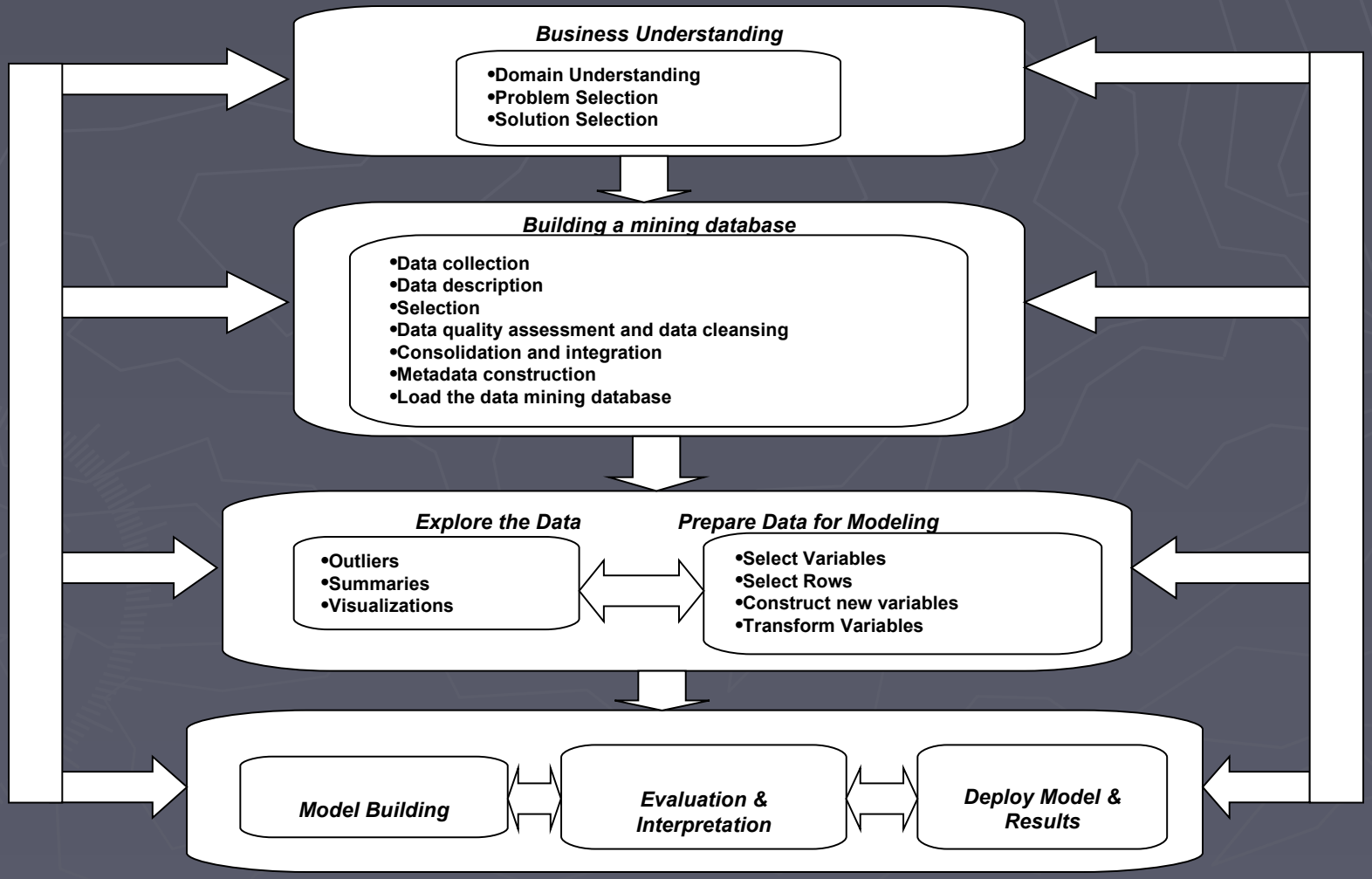
Evaluation

- Evaluate the model, and review the steps executed to construct the model, from a business perspective
- Determining the business issues that are not sufficiently considered.
- Deciding on the use of the data mining results

Deployment

- This can be as simple as generating a report or as complex as implementing a repeatable data mining process, depending on the requirements.
- Normally carried out by the customer and not the data analyst.
- Customer has to understand up front what actions will need to be carried out in order to actually make use of the created models.

Wipro's - DM Process Model



Business Understanding

- Understanding the project objectives and requirements from a business perspective
- Converting this requirements into a data mining problem definition
- Preliminary plan designed to achieve the objectives.

Building a mining database

- Data collection
- Data description
- Selection
- Data quality assessment and data cleansing
- Consolidation and integration
- Metadata construction
- Load the data mining database
- Maintain the data mining database

Explore the data

- Identify outliers.
- Summarization.
- Visualization.

Prepare data for modeling

- Select variables
- Select rows
- Construct new variables
- Transform variables

Building and Deploying model

- Data mining model building
- Evaluation and interpretation of the model
- Deploy the Model and Results

Any Questions ?

Thank You for
your time