K2 Analytics
Building Skills, Building Individuals

Business Analytics using Data Mining (BADM)
- Rajesh Jakhotia

16-Mar-2017

Earning is in Learning
- Rajesh Jakhotia

K2 Analytics
Building Skills, Building Individuals

Agenda

Introduction
Business Case for Analytics
Data Mining in a nut shell
Basic number skills

About K2 Analytics

At K2 Analytics, we believe that skill development is very important for the growth of an individual, which in turn leads to the growth of Society & Industry and ultimately the Nation as a whole. For this it is important that access to knowledge and skill development trainings should be made available easily and economically to every individual.

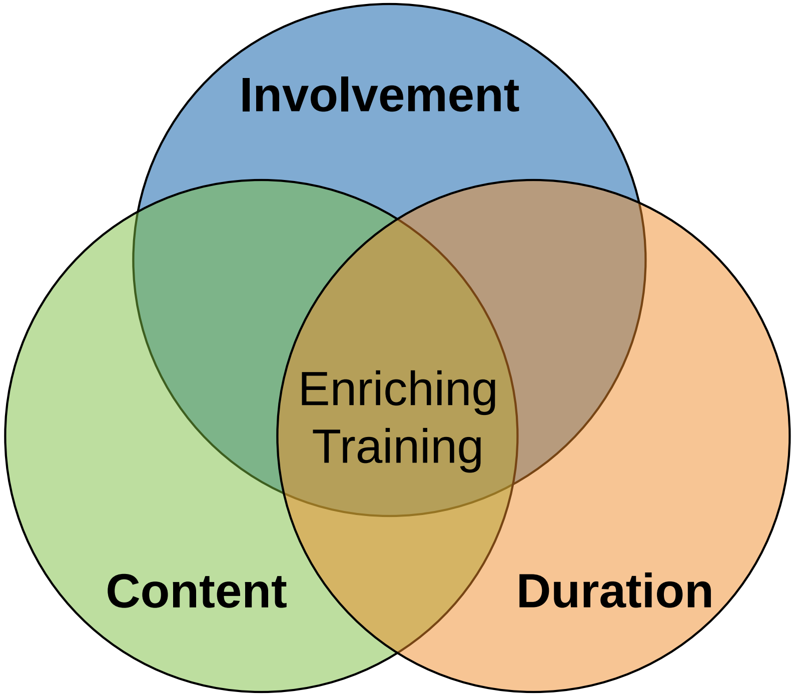
Our Vision: *"To be the preferred partner for training and skill development"*

Our Mission: *"To provide training and skill development training to individuals, make them skilled & industry ready and create a pool of skilled resources readily available for the industry"*

*We have chosen Business Intelligence and Analytics as our focus area. With this endeavour we make this presentation on "**Business Analytics using Data Mining (BADM)**" accessible to all those who wish to learn Analytics. We hope it is of help to you. For any feedback / suggestion or if you are looking for job in analytics then feel free to write back to us at ar.jakhotia@k2analytics.co.in*

Welcome to BADM!!!

Enriching training and learning session…



- **Training Checklist**
  - Sitting arrangement F2F
  - Quality over Quantity
  - Everyone to have their own machines for hands-on practice
  - Illuminated and happy glowing training room (no candle light dinner ambience)
  - Anyone wanting to step-out, feel free
  - Feel free to ask for breaks
  - Feel free to ask same question again till you understand
  - Let me know if you want me to skip Practice Exercises in between the session
  - Brief side-talks are okay
  - **I don’t speak to walls, respect each other**

K2 Analytics
Building Skills, Building Individuals

Simple Business Scenario to understand Application of Data Mining

Simple Business Scenario

Scenario

Let us assume you are working in a Bank and the Chief Marketing Officer suggests that he wish to run a campaign to promote a financial product, say, some Investment Product

Based on business filters, you have an eligible contactable base of 1,000,000 customers.

Cost of Targeting each customer being Rs. 10/-

It is expected that 0.5% incremental customers will purchase the Investment Product because of the campaign

Expected Revenue per customer who purchases the product is Rs. 2500/-

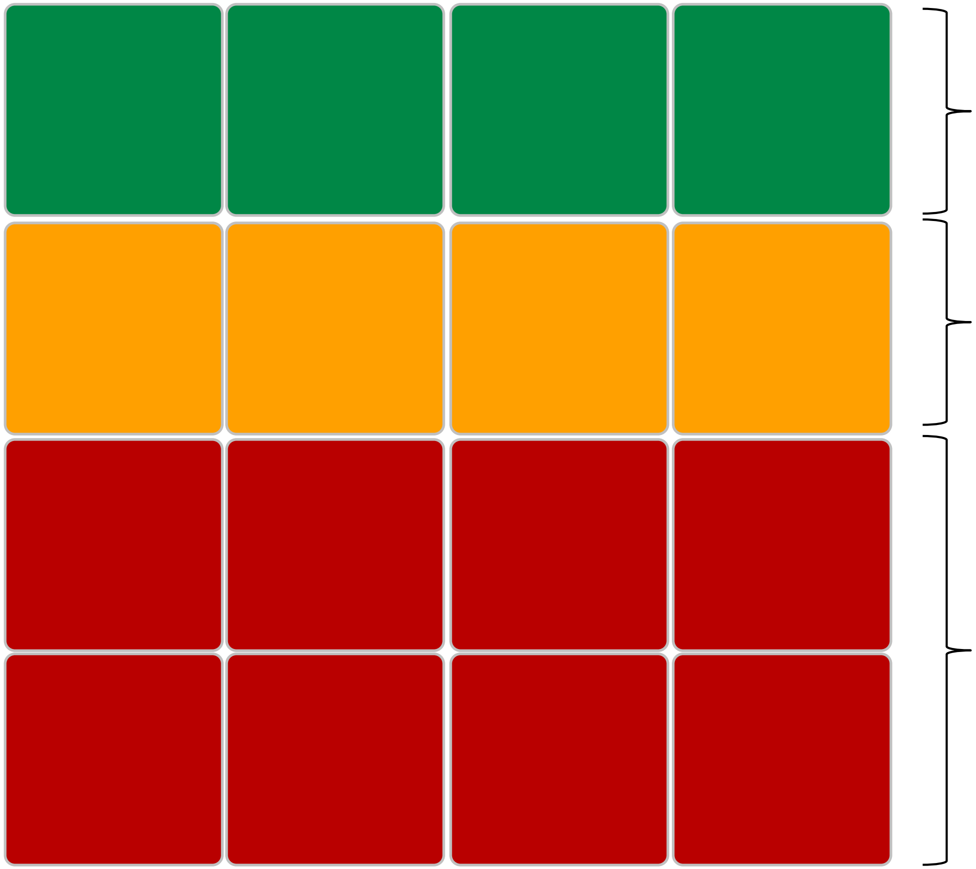
Campaign Return on Marketing Investments without analytical approach

- Target Customer Base : 1,000,000
- Cost of Targeting per customer : INR 10/-
- Cost of Campaign = 1,000,000 * 10 = INR 10,000,000 = 10 Mn
- Expected Incremental Conversion Rate : 0.5%
- Expected Incremental Conversions = 1,000,000 * 0.5% = 5,000
- Expected Revenue per Convert : INR 2500/-
- Expected Incremental Revenue = 5,000 * 2500 = 12,500,000 = 12.5 Mn
- Expected Profit = 12.5 Mn – 10 Mn = 2.5 Mn

Campaign ROMI… contd

$$\text{Return on Marketing Investment (ROMI)} = \frac{\text{Revenue} - \text{Cost}}{\text{Cost}} = \frac{12.5 - 10}{10} = 25\%$$

Analytics Based Approach



High Response Segment

25% of Base
With expected conversion rate of 1.3%

Medium Response Segment

25% of Base
With expected conversion rate of 0.4%

Low Response Segment

50% of Base
With expected conversion rate of 0.15%

Analytics Based ROMI

Note: Cost of Targeting per customer : INR 10/- ; Expected Revenue per Convert : INR 2500/-

Segment	# Customer (A)	Exp. Conv. Rate (B)	# Conv's (C = A * B)	Cost of Targeting (D = A * 10)	Exp. Revenue (E = C * 2500)	Profit (F = E – D)	ROMI G = F / D
High Response Segment	250,000	1.3%	3250	2,500,000	8,125,000	5,625,000	225%
Medium Response Segment	250,000	0.4%	1000	2,500,000	2,500,000	0	0%
Low Response Segment	500,000	0.15%	750	5,000,000	1,875,000	-3,125,000	-ve
Total	1,000,000	0.5%	5000	10,000,000	12,500,000	2,500,000	25%

Recommendation to CMO

Your recommendation to the CMO:

- Target only the High Response Segment

Benefits of your strategy

A) It will reduce Marketing Cost by 75%

B) It will increase Profits by 125%

C) 9X increase in ROMI

Data Mining in a nut shell

Learning Objectives

- What is Data Mining?
- What is Supervised and Unsupervised Learning?
- Understand high level data mining process
- What are the skills required for Data Mining?

Statistics Vs. Data Mining

Statistics

- the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

- Infer / Describe
- Data Collection
- Large Dataset implies hundred / thousand data points
- Population / Sample Level
- Charts & Table
- Makes many assumptions

Data Mining

- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

- Predict
- Data Preparation
- Large Datasets implies millions / billions data points
- Customer (Granular) Level
- Visualizations
- Makes few / no assumptions

http://www.cs.csi.cuny.edu/~imberman/DataMining/Statistics%20vs.pdf

Types of Data Mining Techniques

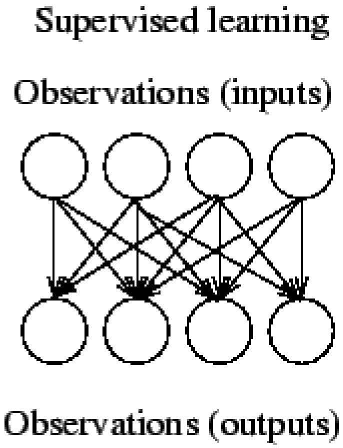
- Supervised learning: The target output expected is clearly defined

- Unsupervised learning: The data have no target attribute.
  - We want to explore the data to find some intrinsic structures in them.

Types of Data Mining techniques

Supervised Techniques

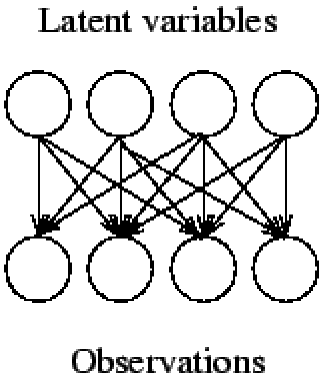
- In supervised learning, the model defines the effect one set of observations, called inputs, has on another set of observations, called outputs



- Prediction (numerical Y)
- Classification (Categorical Y)

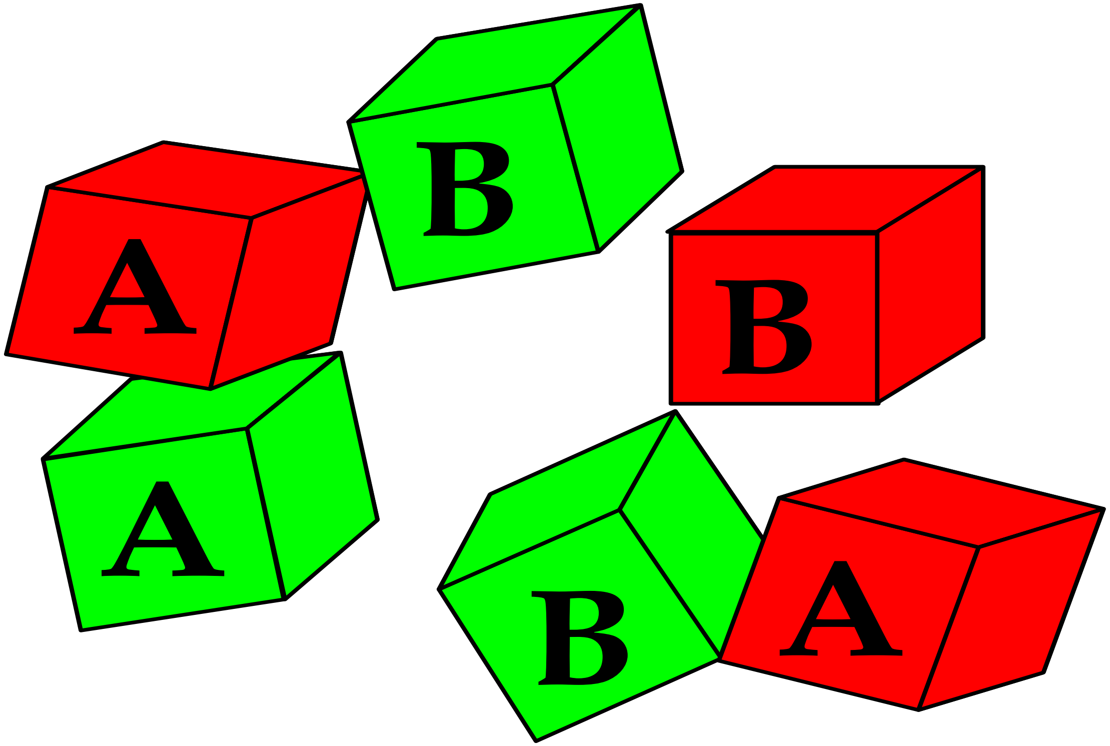
Unsupervised Techniques

- In unsupervised learning, all the observations are assumed to be caused by latent variables, that is, the observations are assumed to be at the end of the causal chain

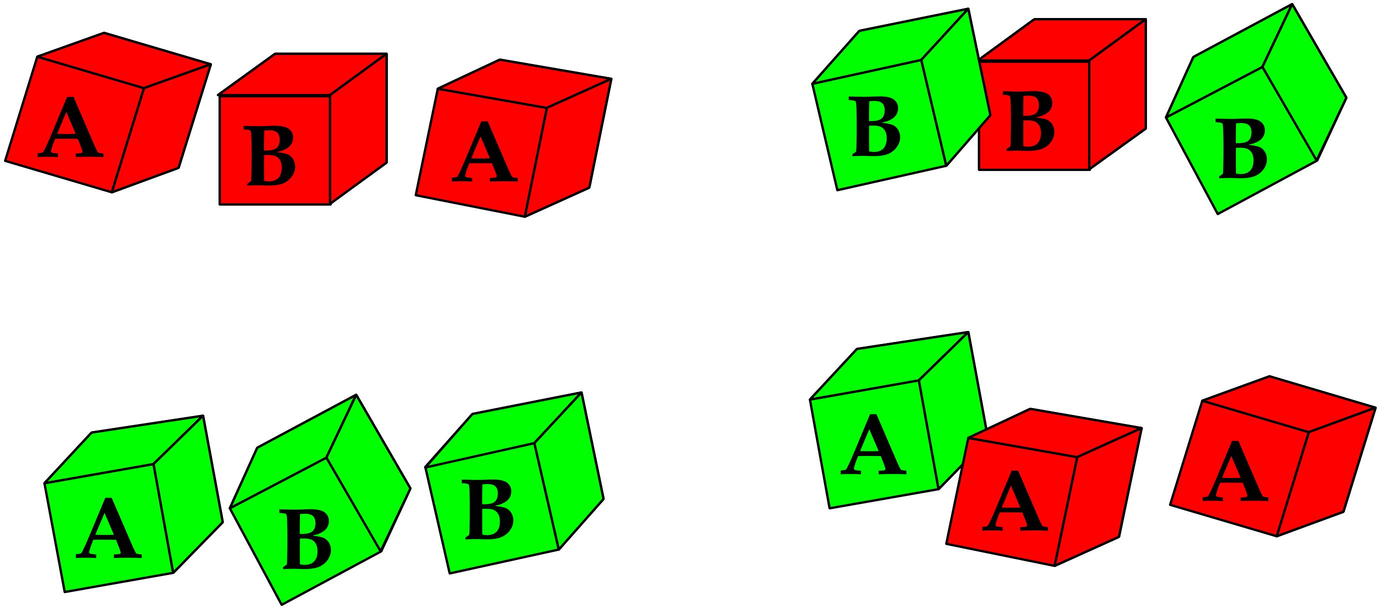


- Dimension Reduction
- Clustering
- Association Analysis

Understanding Supervised and Unsupervised Learning

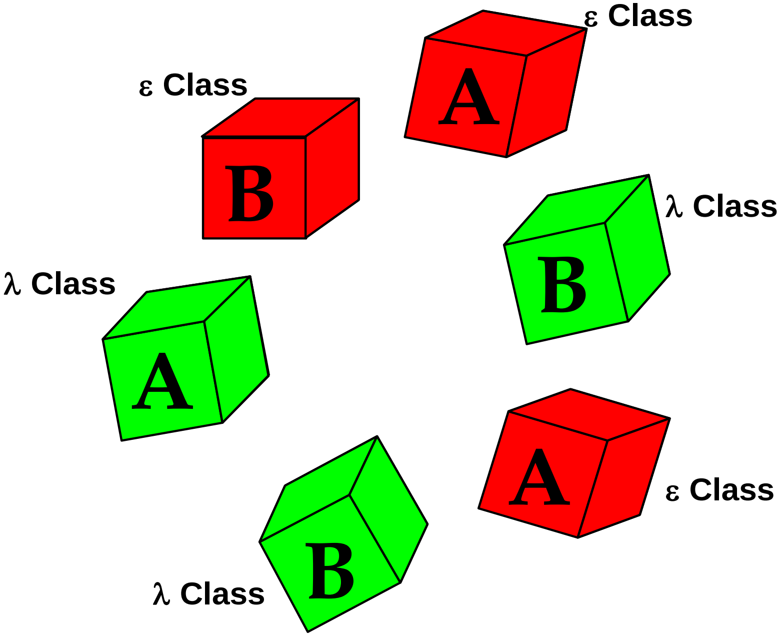


Two Possible Solutions



Supervised Learning

- It is based on a labeled training set.
- The class of each piece of data in training set is known.
- Class labels are pre-determined and provided in the training phase.



Modeling Process: CRISP-DM

- CRISP – DM stands for Cross Industry Standard Process for Data Mining

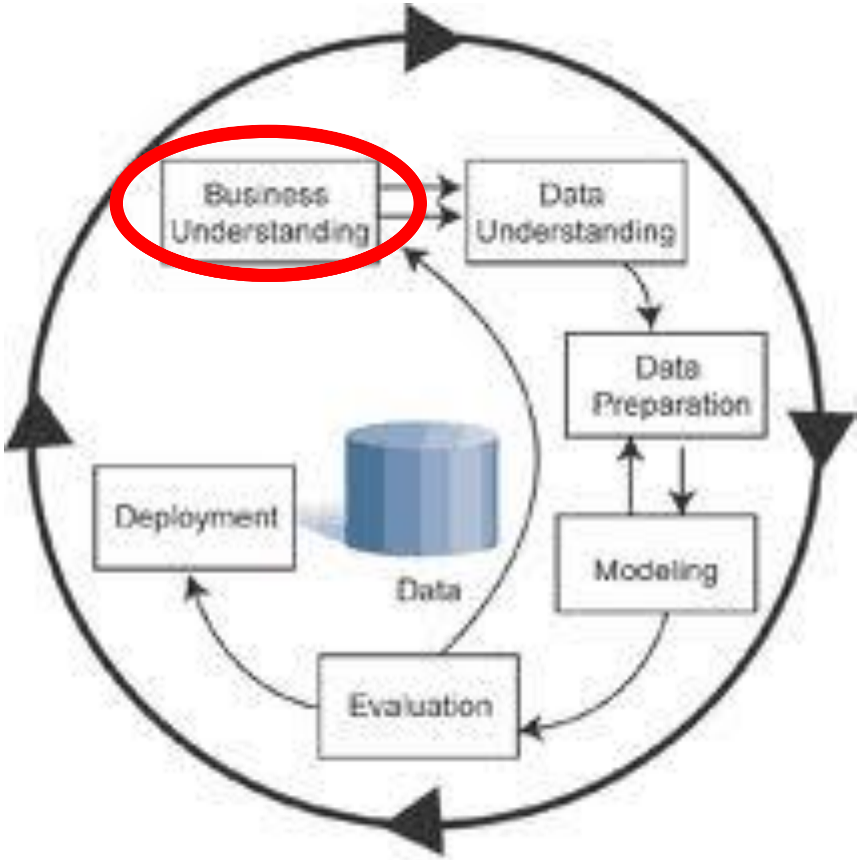
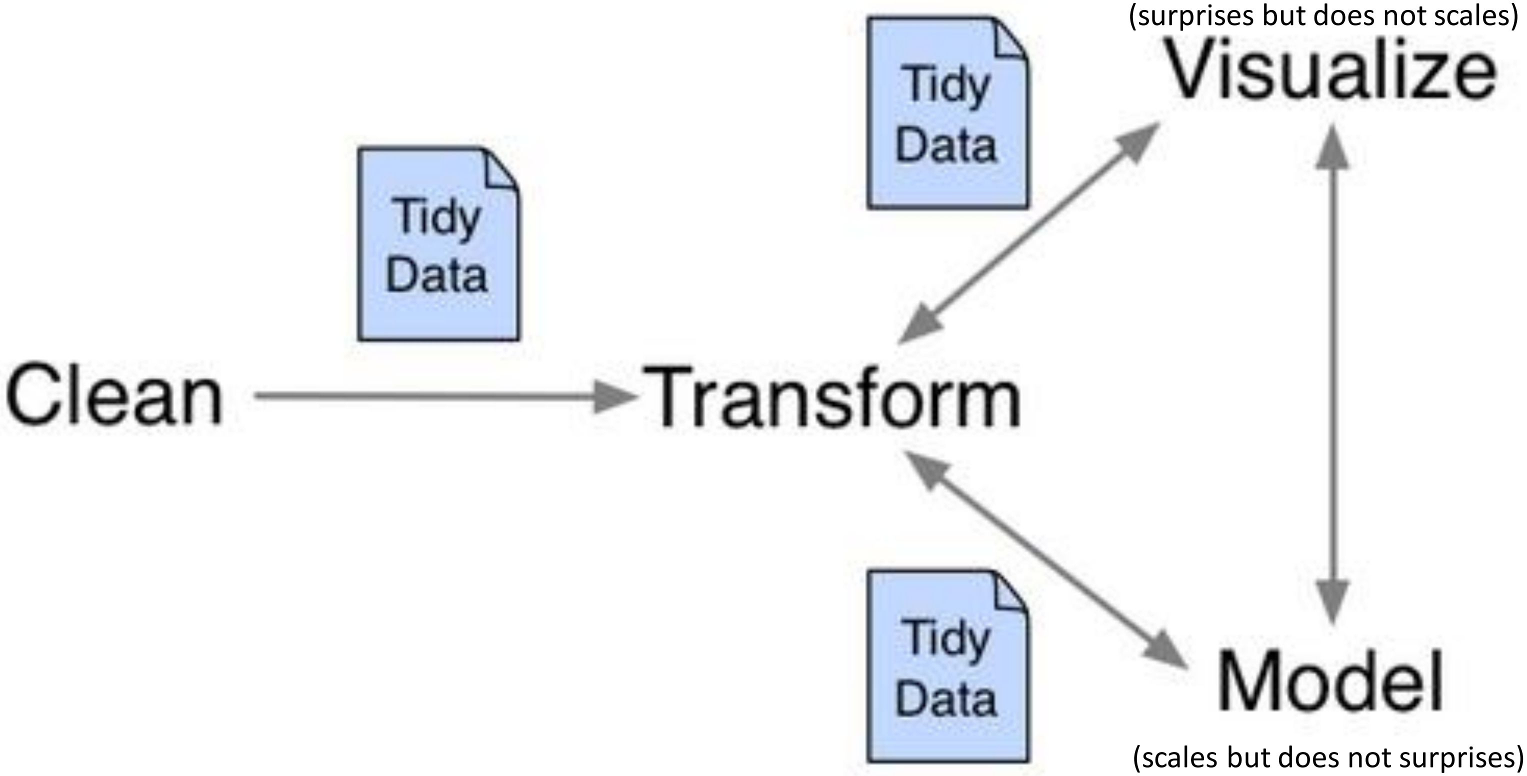


FIGURE 1 Data-Mining Process Model

Most challenging part in modeling is defining the objective

Data Preparation & Modeling Process

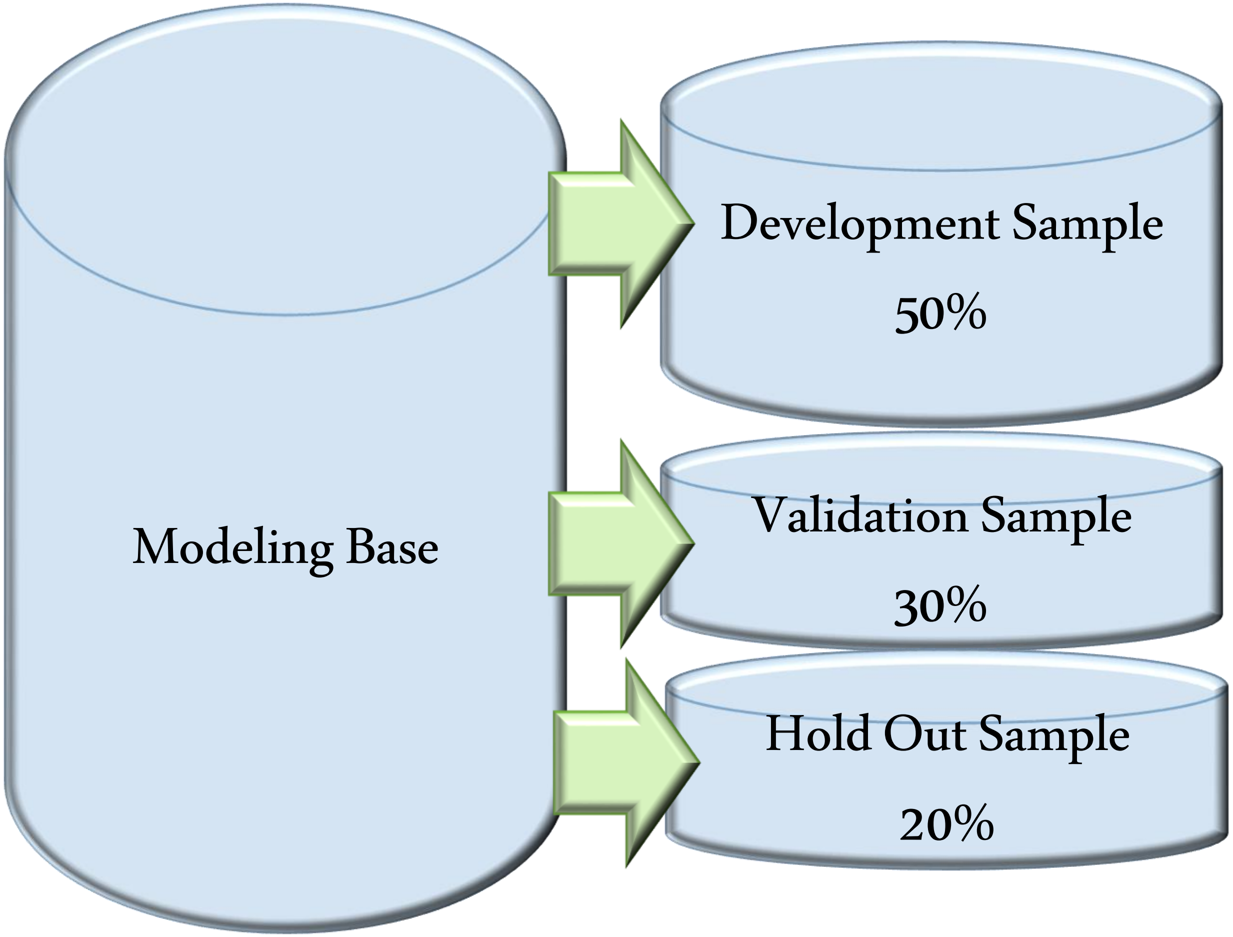
(surprises but does not scales)



(scales but does not surprises)

Note: Figure adapted and modified from a presentation by Hadley Wickham.

Hold Out concept in Model Evaluation



Recap…Short Exercise

- What is Supervised Learning Technique?
- What is the First Phase and the Most Important Phase in Data Mining?
- Why do you need Hold-out Sample?
- Data Mining Techniques typically do not make any assumptions about data – **Yes or No**
- Visualization scales but does not surprises **– Yes or No**

K2 Analytics
Building Skills, Building Individuals

Basic Number Skills
Modeling Techniques

Basic Number Skills

- Measures of Central Tendency
  - Mean, Median, Mode
- Measures of Dispersion
  - Std. Deviation, Variance
- Correlation and Covariance
- Chi-Sq Test
- Additive Variables, Count and Ratio

Standardization & Normalization

- Standardization & Normalization are 2 commonly used method for rescaling
- *Normalization*, which scales all numeric variables in the range [0,1]. One possible formula is given below:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Standardization transforms data to mean zero and unit variance

$$x_{new} = \frac{x - \mu}{\sigma}$$

Hypothesis Testing

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps.

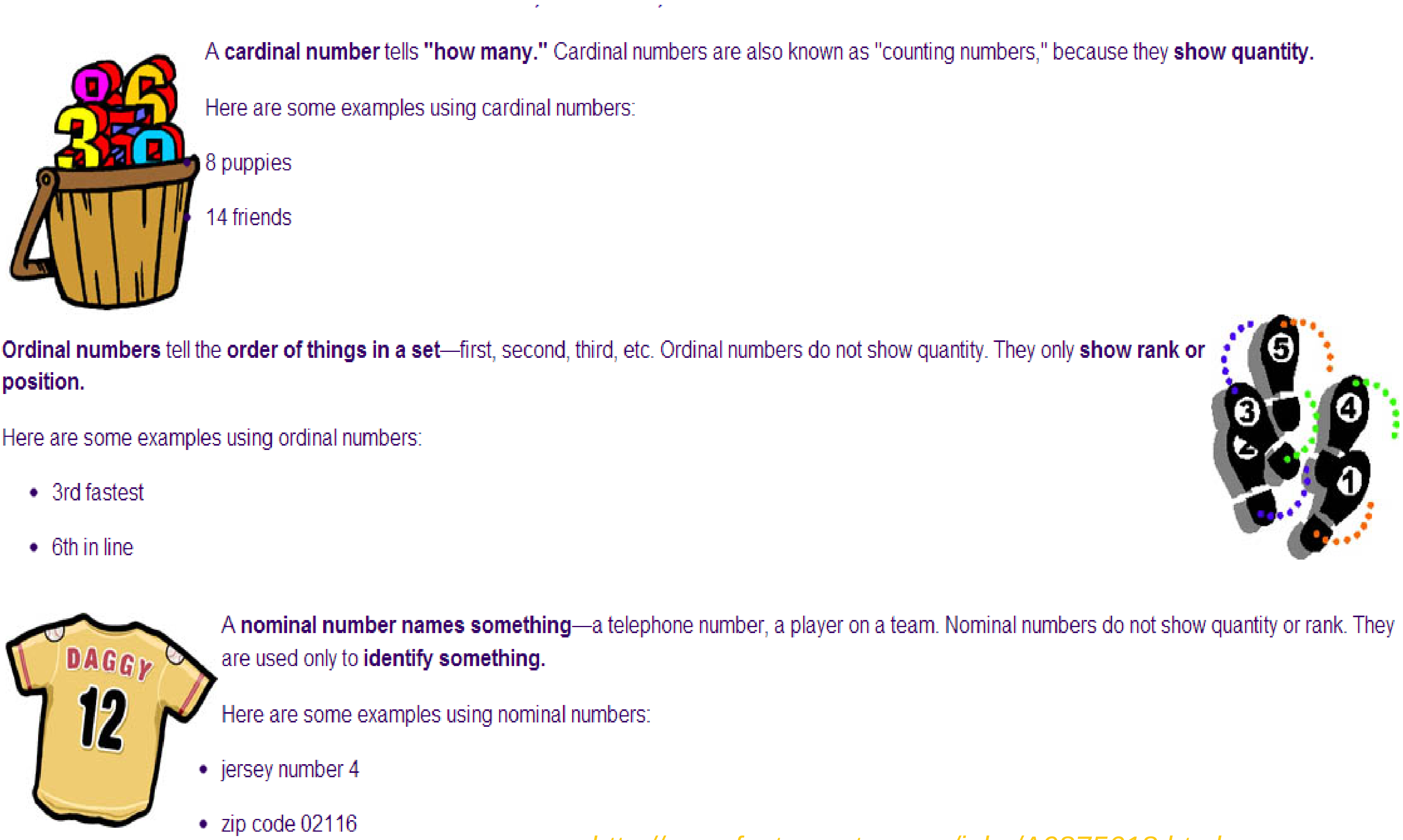
1. Formulate the **null hypothesis** (commonly, that the observations are the result of pure chance) and the **alternate hypothesis** (commonly, that the observations show a real effect combined with a component of chance variation).
2. Identify a **test statistic** that can be used to assess the truth of the **null hypothesis**
3. Compute the **p-value**, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the **null hypothesis** were true. The smaller the p-value, the stronger the evidence against the null hypothesis.
4. Compare the p-value to an acceptable significance value α (sometimes called an **alpha value**). If $p <= \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

Uni, Bi & Multi-variate analysis

- Univariate Analysis – Descriptive statistics like Mean, Median, Mode, STD Deviation, Variance, Frequency Distribution

- Bi & Multi-variate analysis –
  - Differences of Group (Chi-Sq, t-Test, ANOVA)
  - Relationship (Correlation & Regression)

Cardinal, Ordinal, & Nominal Numbers



A **cardinal number** tells **"how many."** Cardinal numbers are also known as "counting numbers," because they **show quantity.**

Here are some examples using cardinal numbers:

- 8 puppies
- 14 friends

Ordinal numbers tell the **order of things in a set**—first, second, third, etc. Ordinal numbers do not show quantity. They only **show rank or position.**

Here are some examples using ordinal numbers:

- 3rd fastest
- 6th in line

A **nominal number names something**—a telephone number, a player on a team. Nominal numbers do not show quantity or rank. They are used only to **identify something.**

Here are some examples using nominal numbers:

- jersey number 4
- zip code 02116

http://www.factmonster.com/ipka/A0875618.html

Questions?? … Thankyou

Contact Us
ar.jakhotia@k2analytics.co.in