Data Engineering

MG-GY 8441

Describing Data

- Agenda
  - Overview
  - Data Types
  - Numerical Summaries
  - Visualizations
- References
  - Han, Kamber, Pei, *Data Mining: Concepts and Techniques*
    - Chapter 1
    - Chapter 2.1 - 2.3

Overview

Overview

- Why Data Mining?
- What Is Data Mining?
- What Kind of Data Can Be Mined?
- What Technology Are Used?
- What Are the Applications?
- What Are the Challenges?

Why Data Mining?

Classification

- Classify credit applicants as low, medium, high risk

Estimation

- Estimate the **click-through-rate** of an advertisement

Prediction

- Predict which customers will leave within six months

Why Data Mining?

Example from Marketing

- You are in a meeting with your boss and a large publisher where you are negotiating to buy some advertising on their website.
- The publisher tells you the cost per thousand views of your advertisement (CPV) is $10.
- Given your goal of collecting email addresses for potential new customers, you need to know the maximum CPV you can afford to effectively negotiate with the publisher.

Why Data Mining?

Example from Marketing

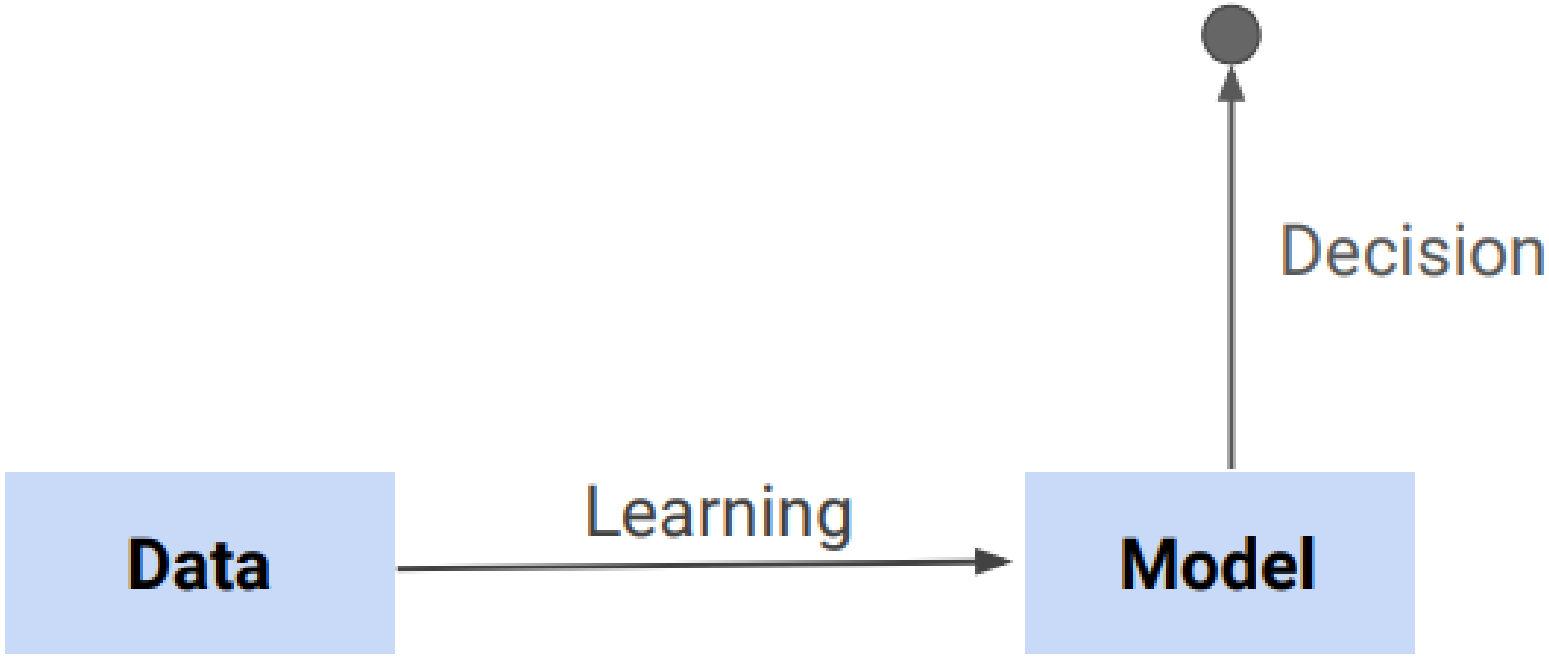
- You estimate that the click-through rate (CTR) of your advertisement has been around 1%.
- Your conversion rate (CR) has been averaging 10% in terms of email sign-ups.
- If you can afford to pay $5 per email you acquire as your cost per acquisition (CPA), is $10 CPV a good price from the publisher?

Why Data Mining?

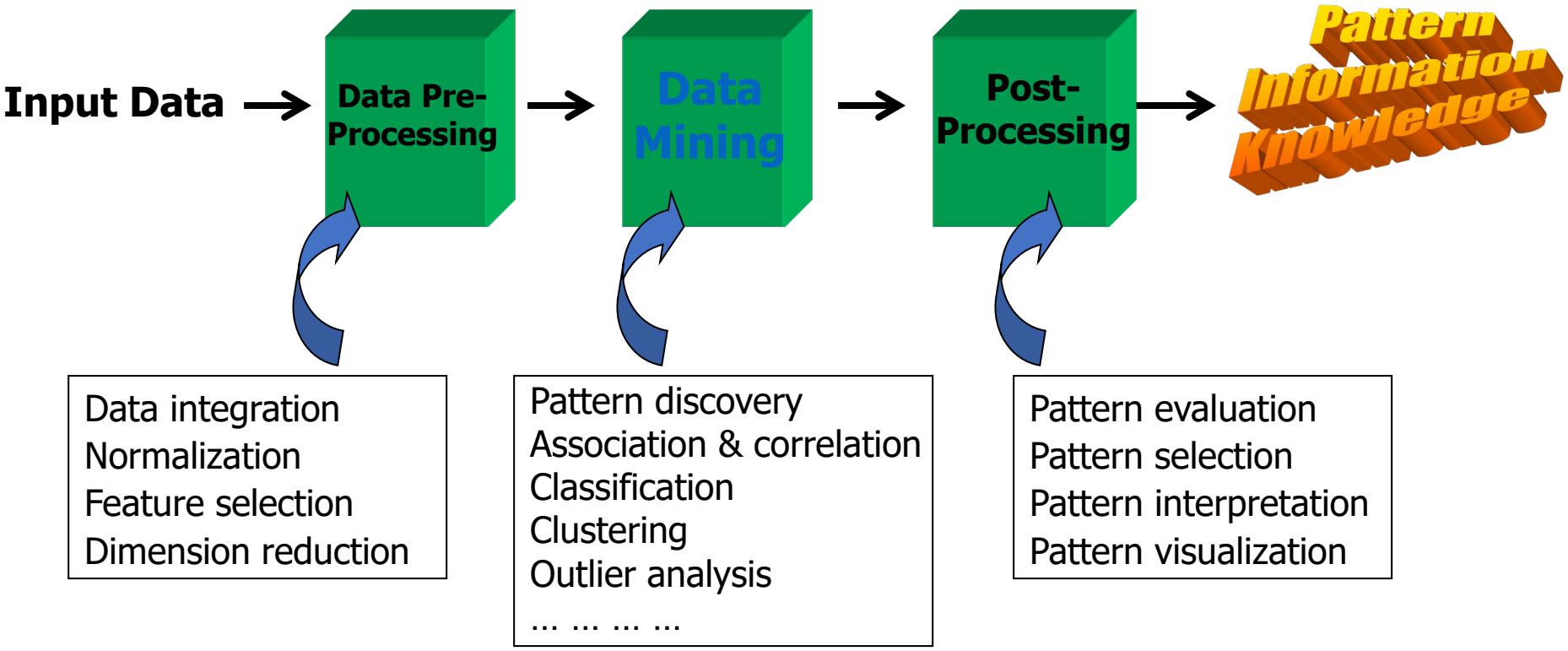
Example from Marketing:

- CPV = $10, CTR = 1%, CR = 10%, CPA goal = $5
  - CPA = CPV / ((CTR*1000) * CR)
  - CPA = $10 / ((0.01 * 1000) * 0.1)
  - CPA = $10 / (10 * 0.1)
  - CPA = $10 / 1
- So the cost per acquisition goal would need to be doubled to match the publisher's price of $10 for cost per thousand views

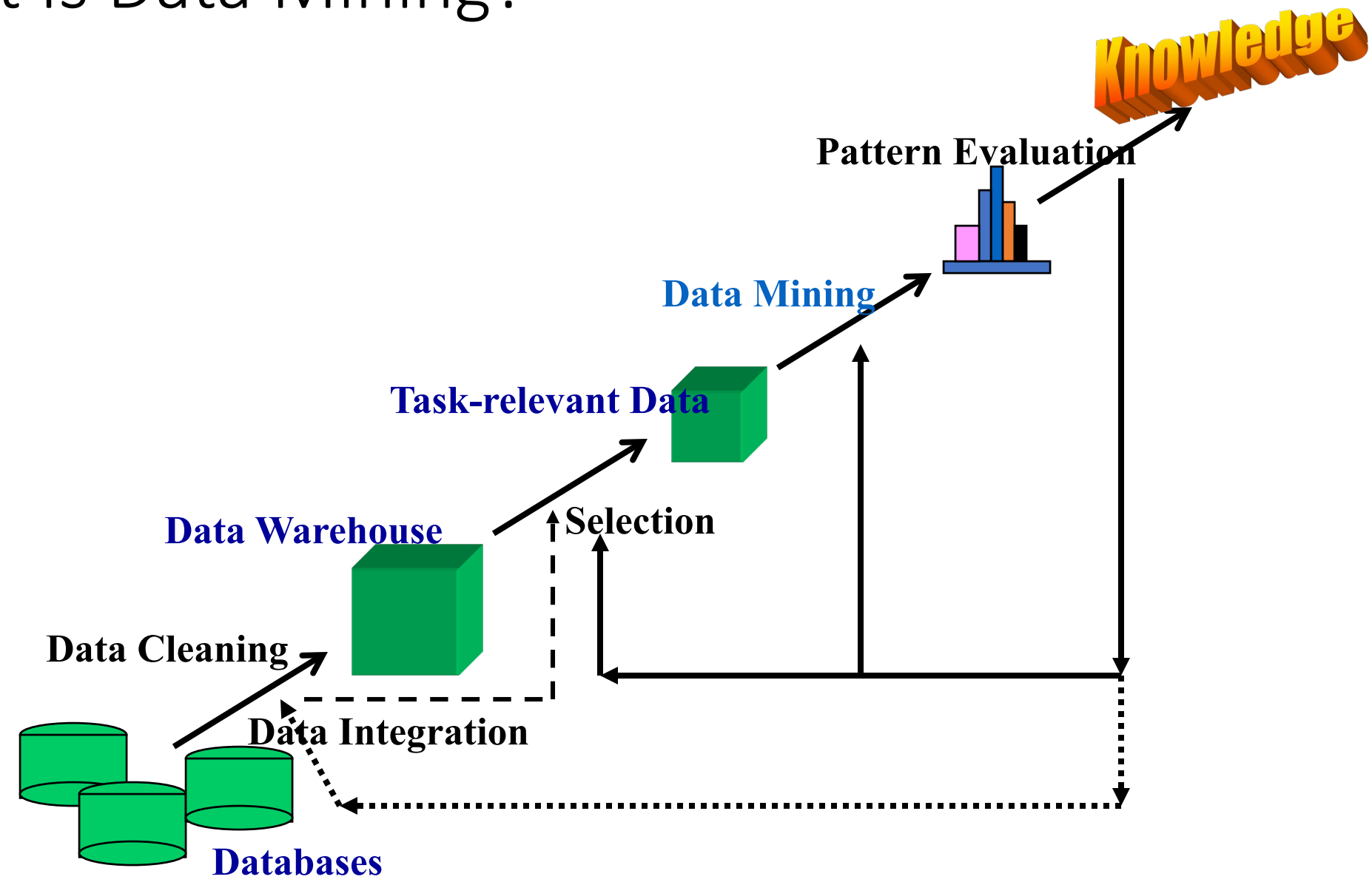
What Is Data Mining?



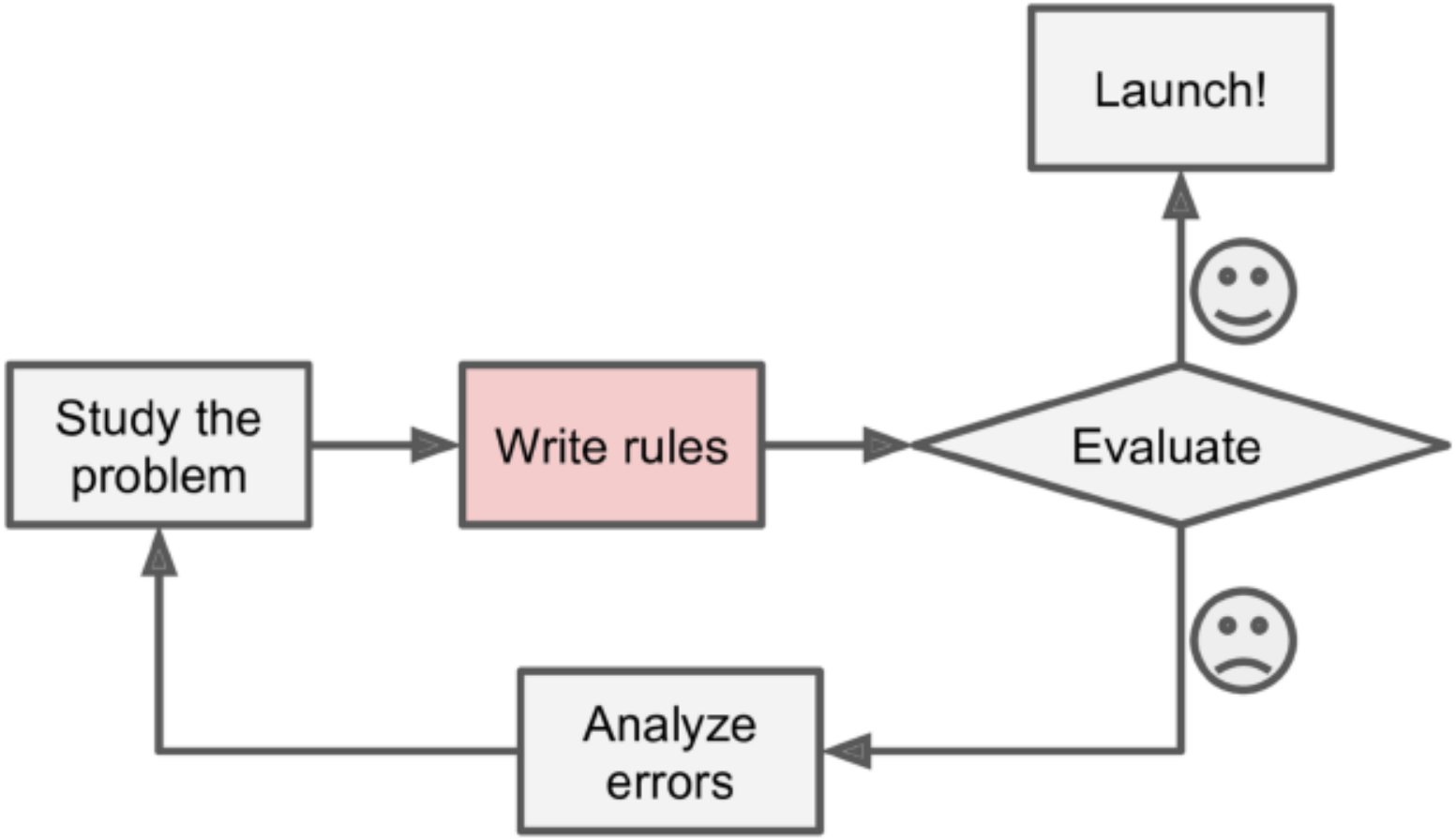
What Is Data Mining?



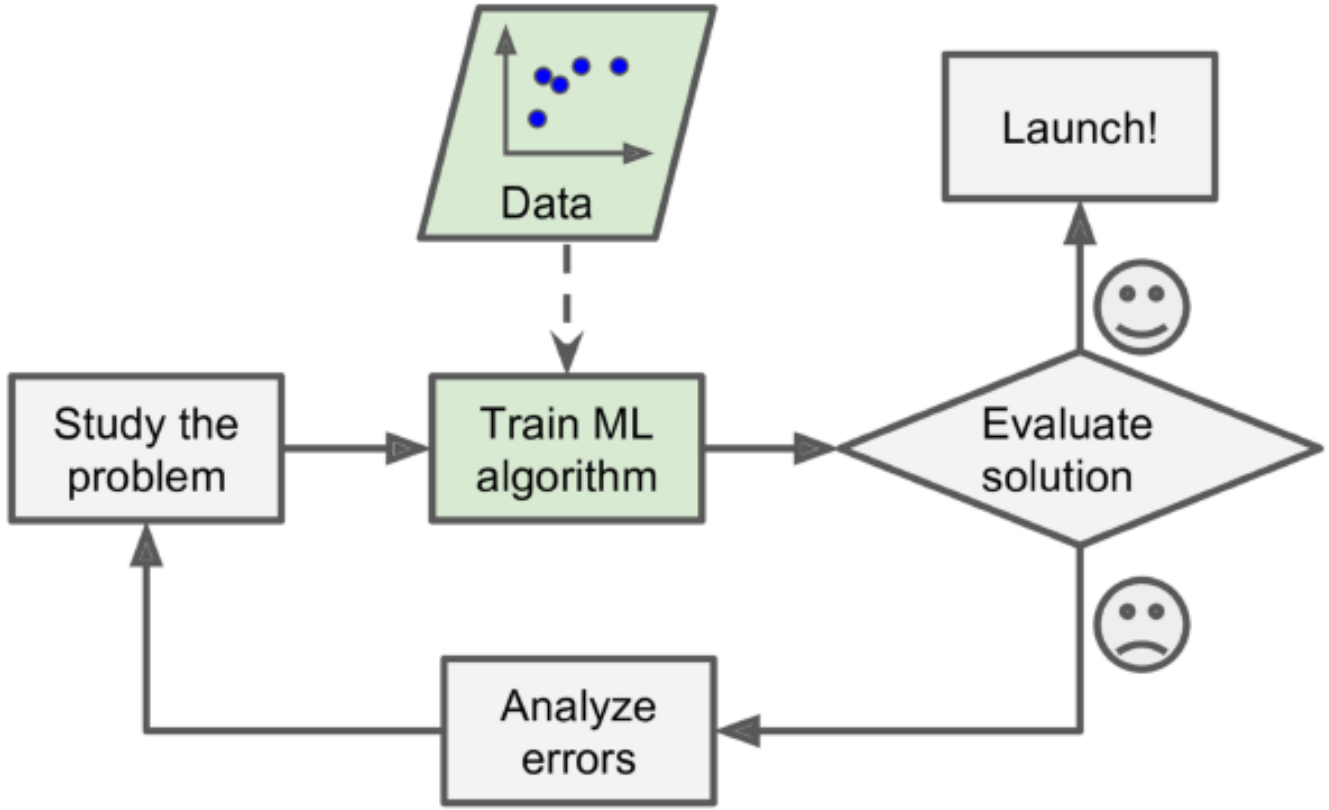
What Is Data Mining?



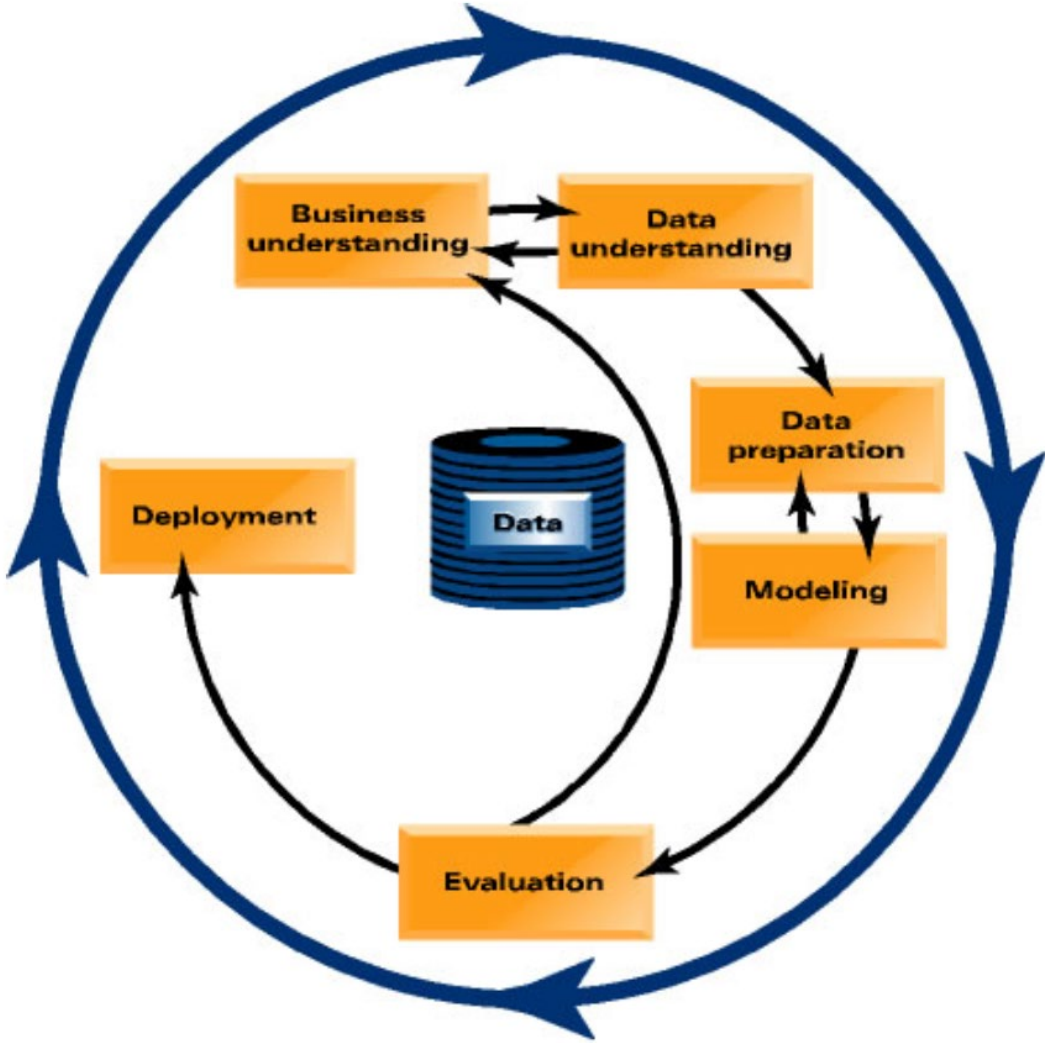
What Is Data Mining?



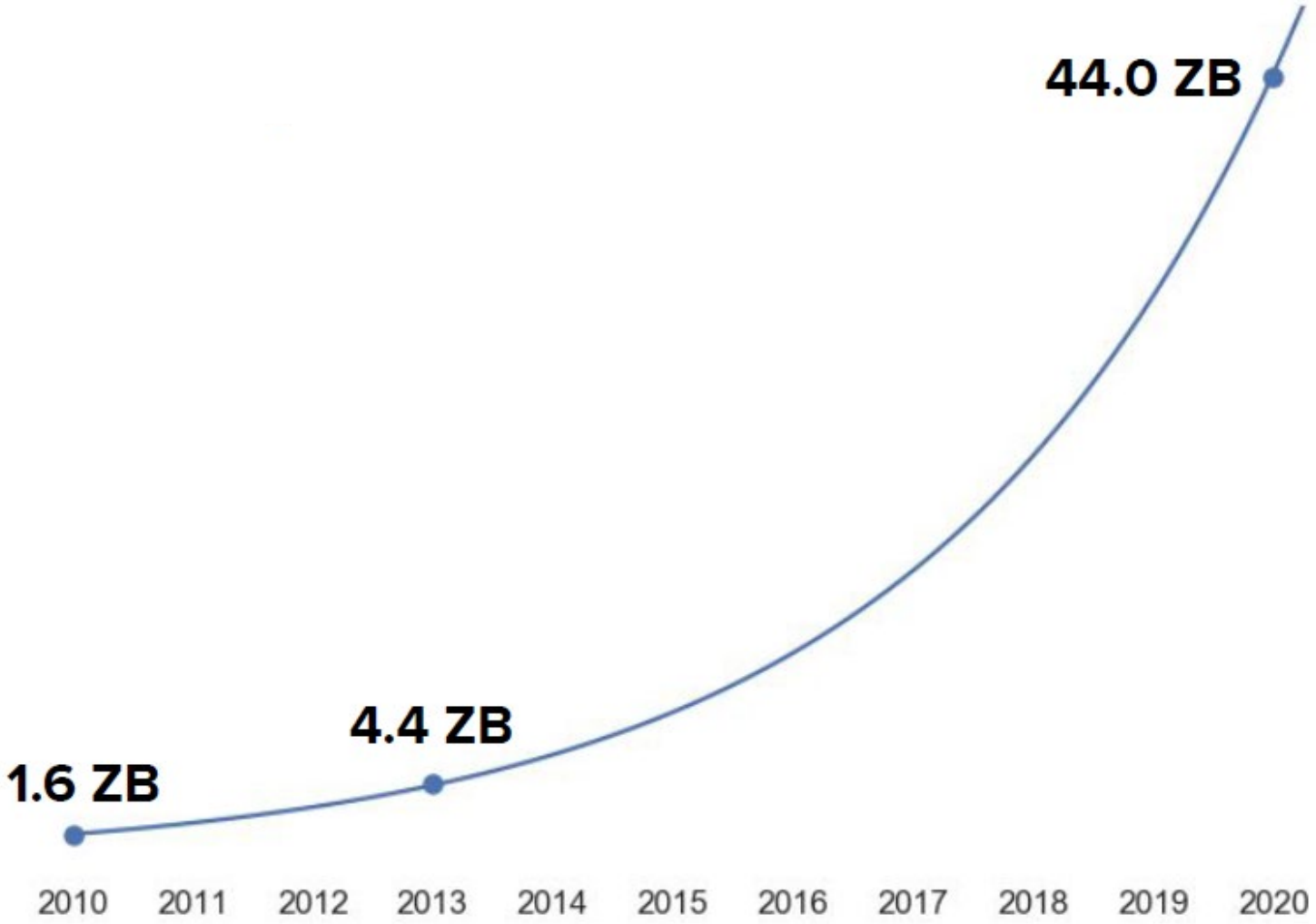
What Is Data Mining?



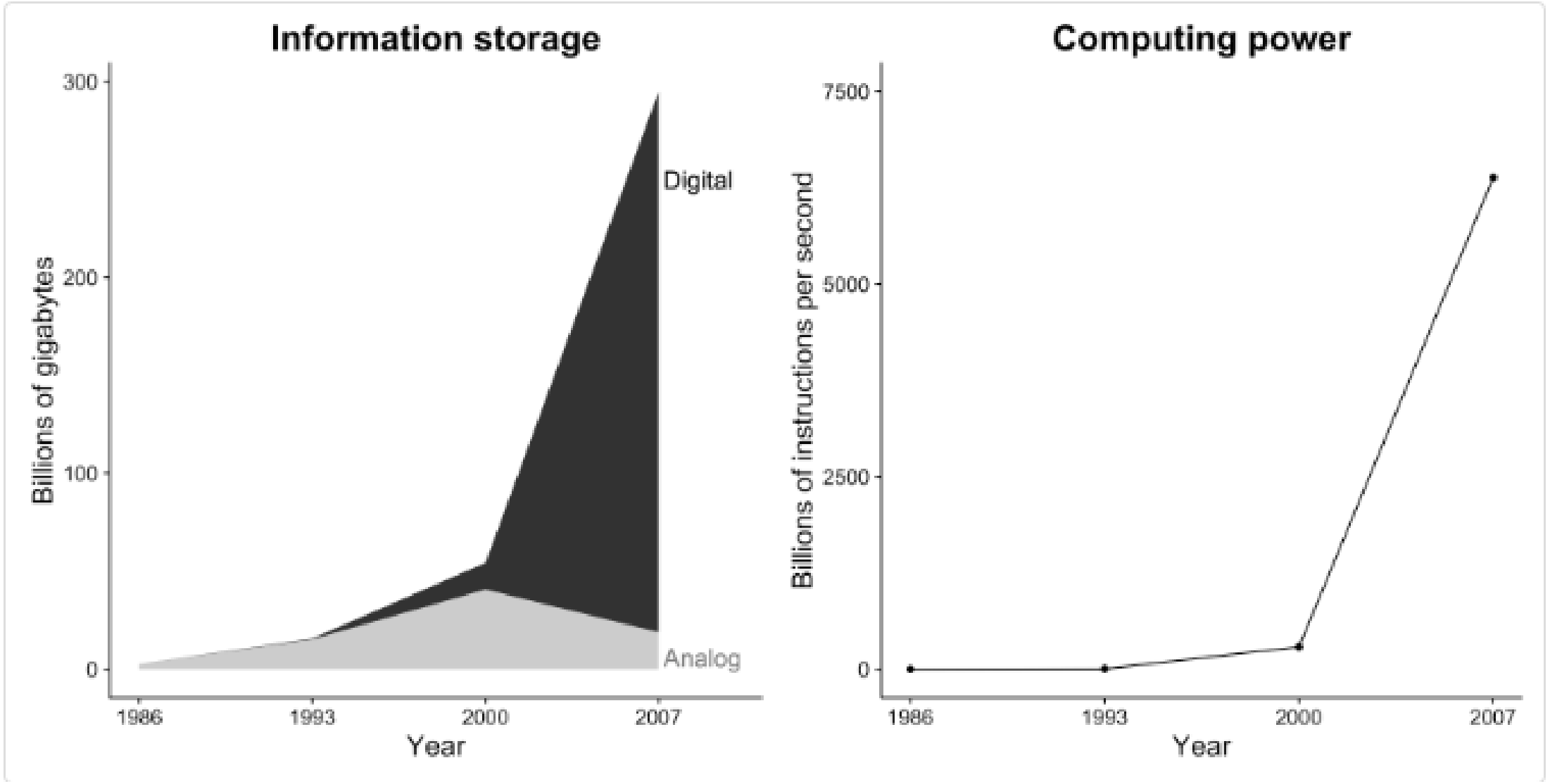
What Is Data Mining?



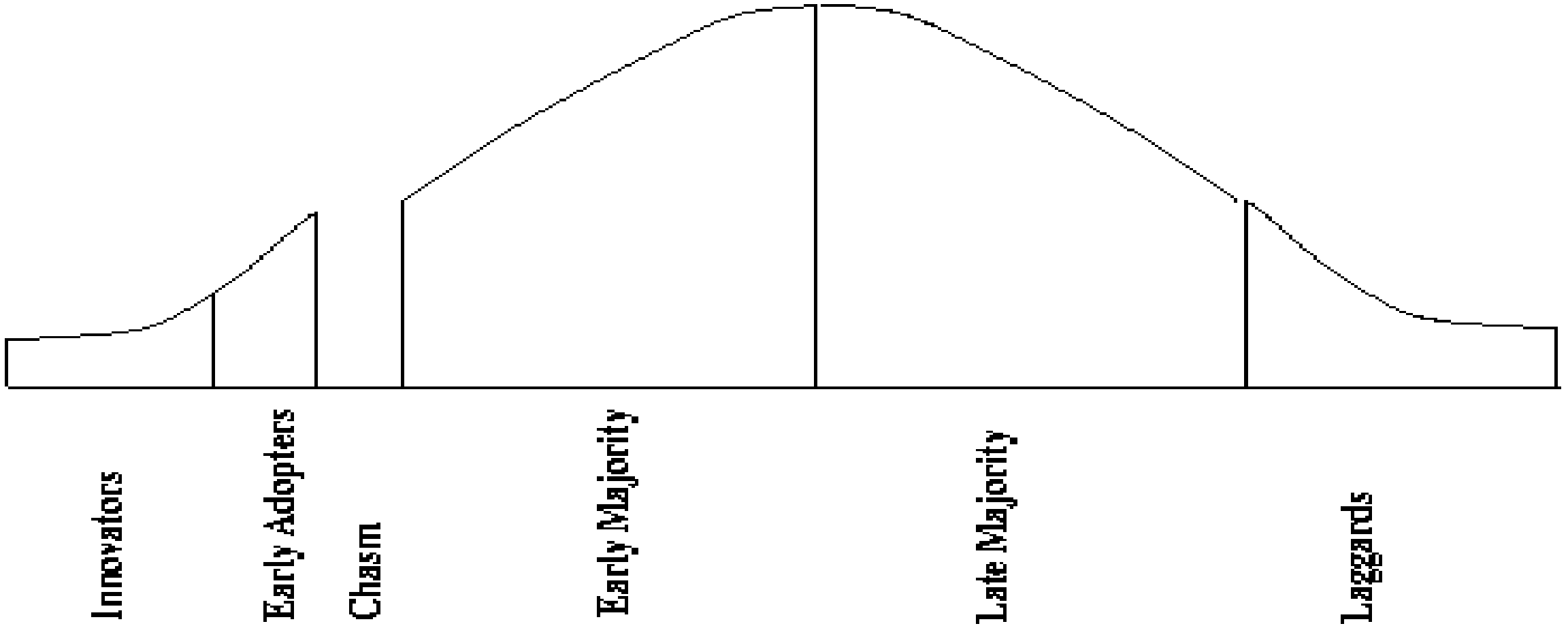
What Kind of Data Can Be Mined?



What Kind of Data Can Be Mined?



What Technology Are Used?

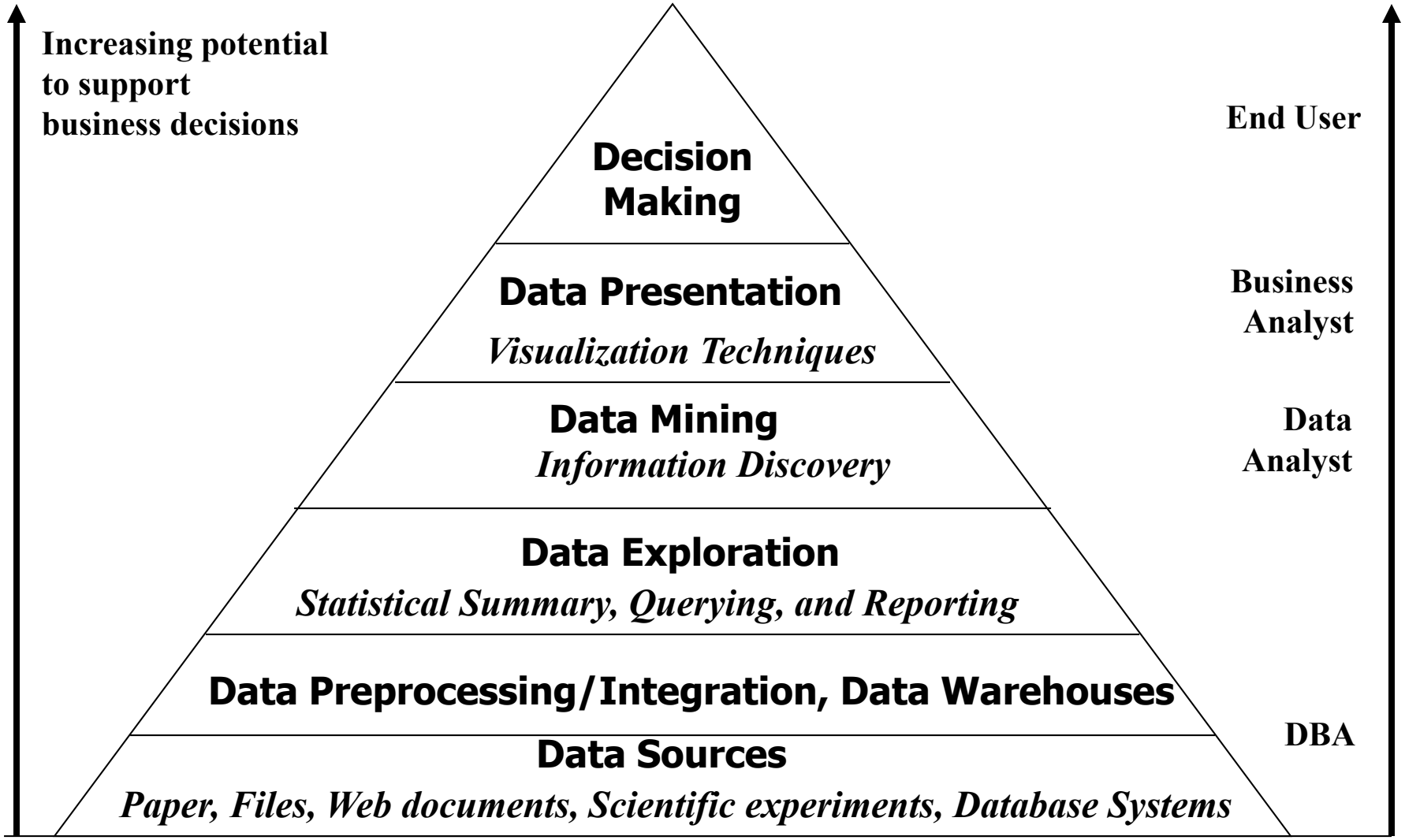


What Technology Are Used?

What Are the Applications?

- Market Basket Analysis
  - Identify what products are likely to be bought together
- Entity Resolution
  - Disambiguate records by linking various data sources
- Market segmentation
  - Identify common characteristics of customers who buy same products
- Collaborative Filtering
  - Recommend products to customers based on preferences

What Are the Applications?



What Are the Challenges?

- Privacy
  - Right to be unknown or forgotten
- Transparency
  - Redistribution of data
- Accountability
  - Oversight of companies and government agencies
- Fairness
  - Social impact of data driven decision making

Data Types

Data Types

- Categorical
  - Ordinal
  - Nominative
- Numerical
  - Continuous
  - Discrete

Properties of Data

Volume	The quantity of data
Velocity	Speed at which data is collected
Variety	Data may be structured or heterogeneous
***Veracity**	Data can be noisy, incomplete, or wrong

Properties of Data

- Dimensionality
- Sparsity
- Resolution
- Distribution

Data Object

- Data sets are made up of data objects. A **data object** represents an entity.
  - Also called *samples , examples, instances, data points, objects, tuples*.
- Example:
  - sales database: customers, store items, sales
- Data objects are described by **attributes**.
  - Database rows -> data objects; columns ->attributes.

Qualitative Data

- **Nominal:** categories, states, or “names of things”
  - *Hair_color =* {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size =* {*small, medium, large*}*,* grades, army rankings

Quantitative Data

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Quantitative Data

- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Numerical Summaries

Numerical Summaries

- Measuring Central Tendencies
  - Mean
  - Median
  - Mode
- Ranking Numbers
  - Quantiles

Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Quantile analysis on sorted intervals

Measuring the Central Tendency

- Mean (sample vs. population):

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Measuring the Central Tendency

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

- Median:
  - Middle value if odd number of values
  - Average of the middle two values if even number
  - Estimated by interpolation for *grouped data*

$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

Measuring the Central Tendency

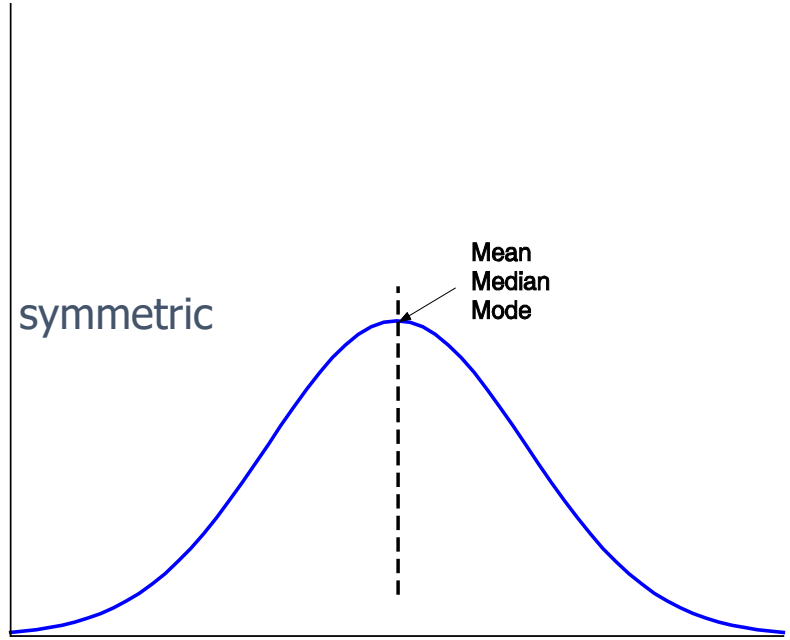
Mode

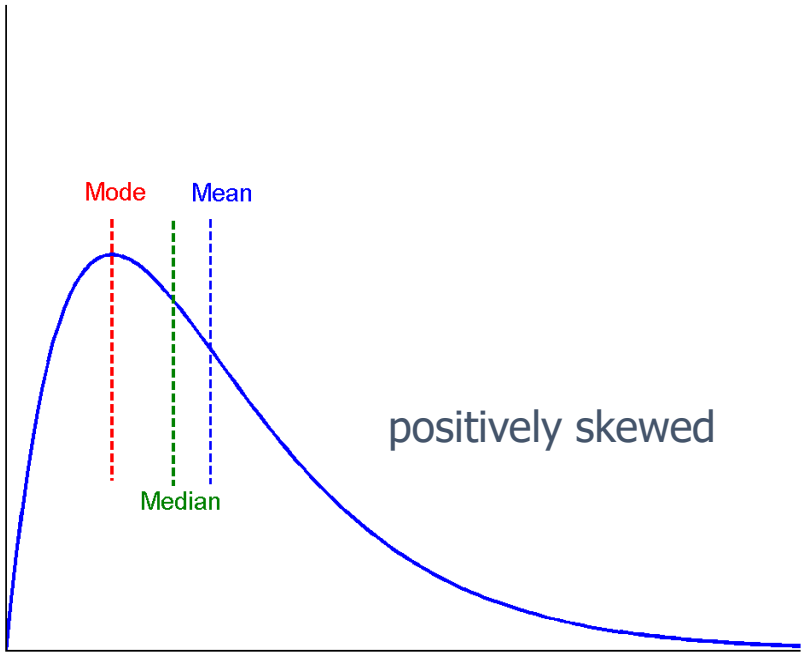
- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:

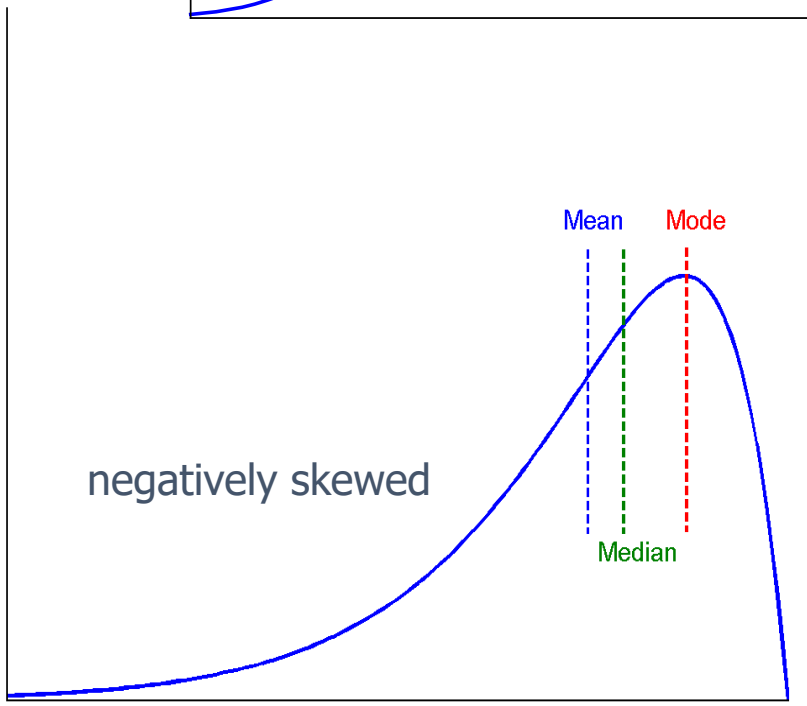
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

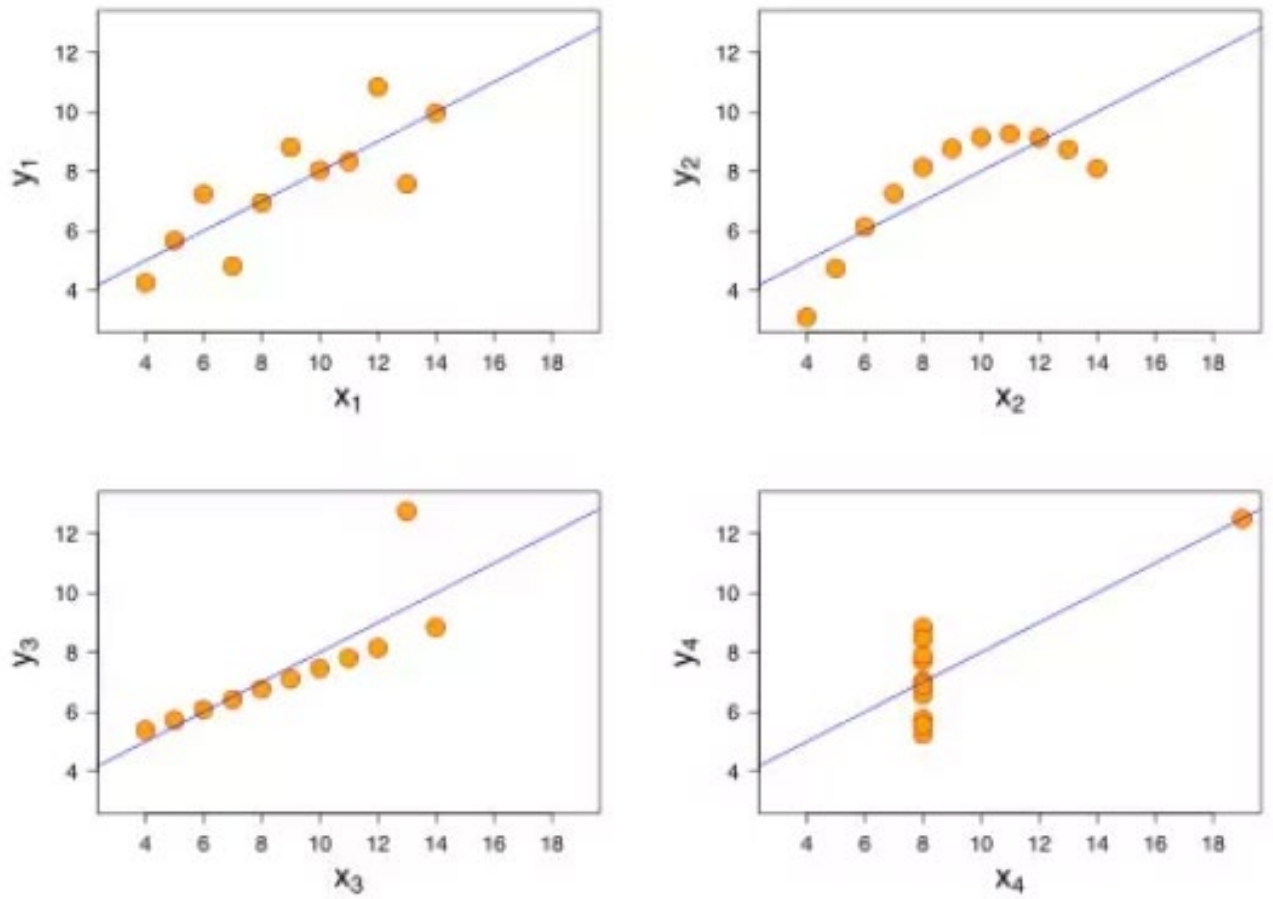
Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

Visualization



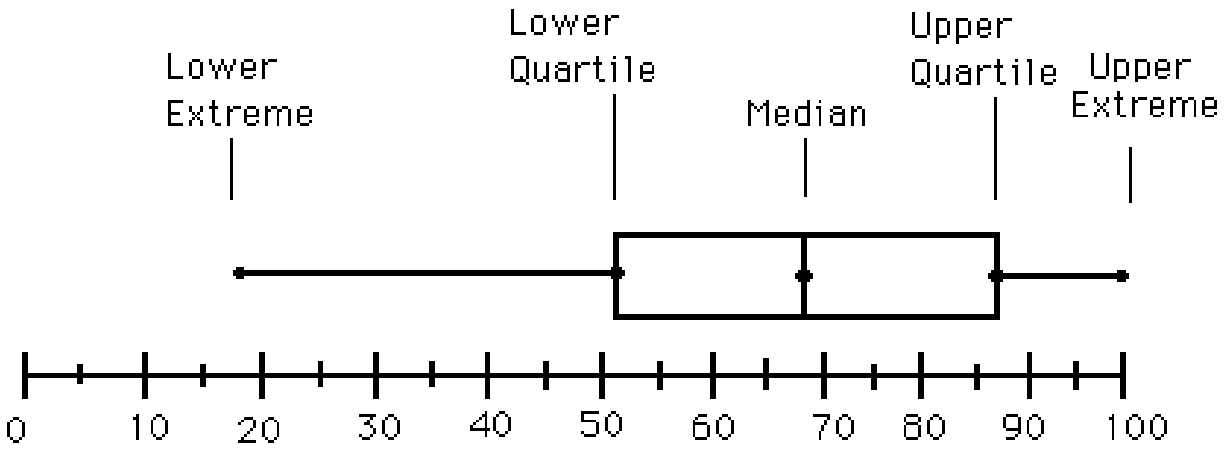
Visualization

- Computational
- Statistical
  - Categorical
    - Ordinal
    - Nominative
  - Numerical
    - Continuous
    - Discrete

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

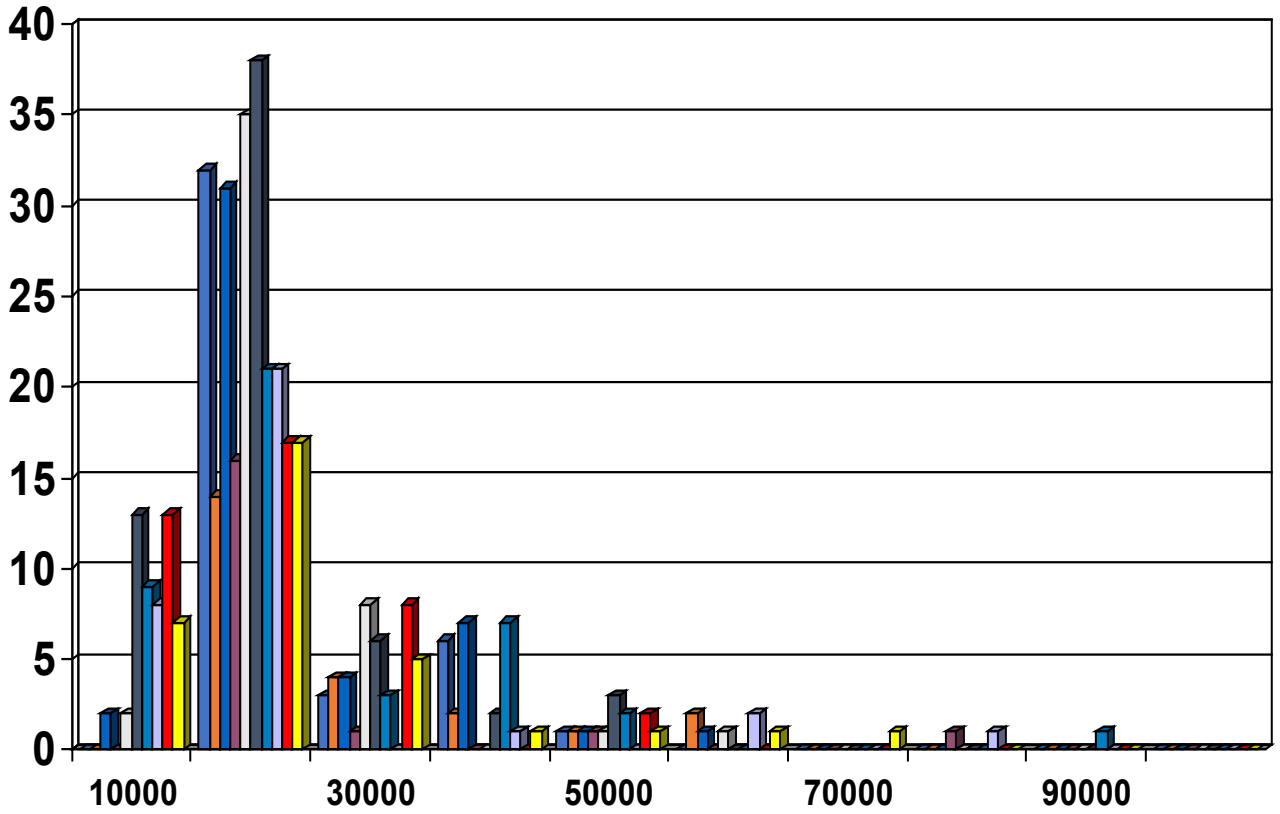
Boxplot



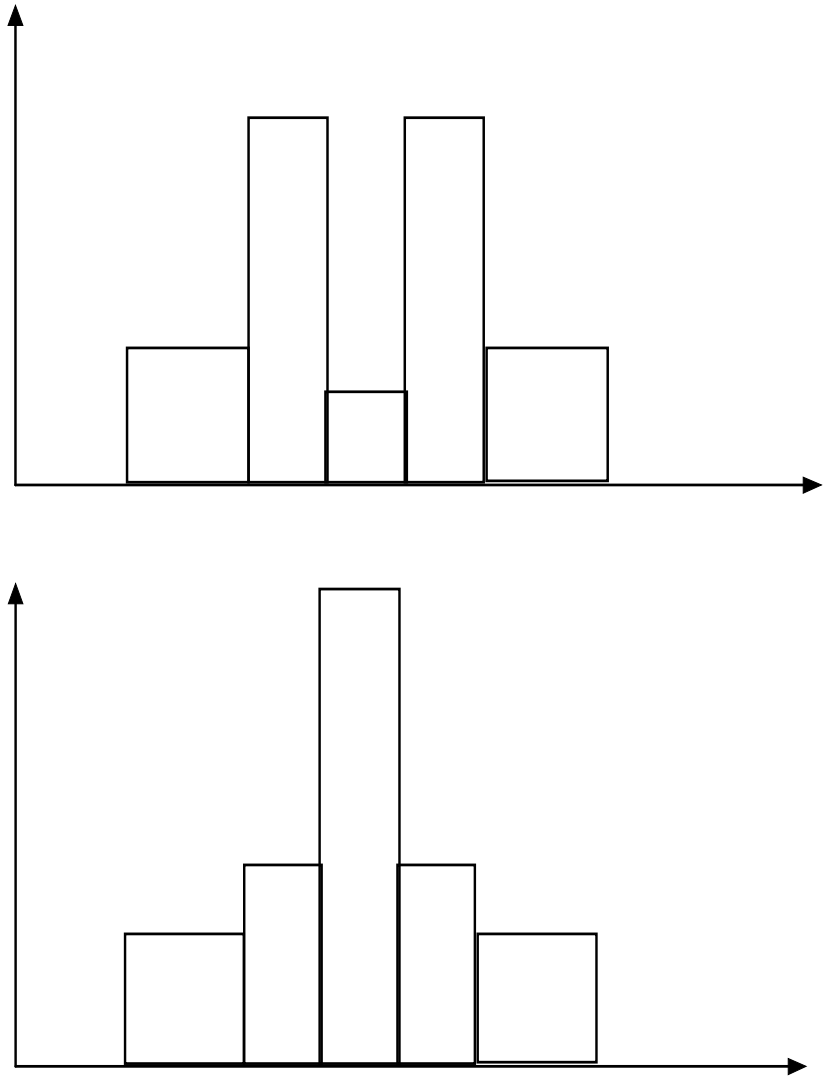
- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

Histogram

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



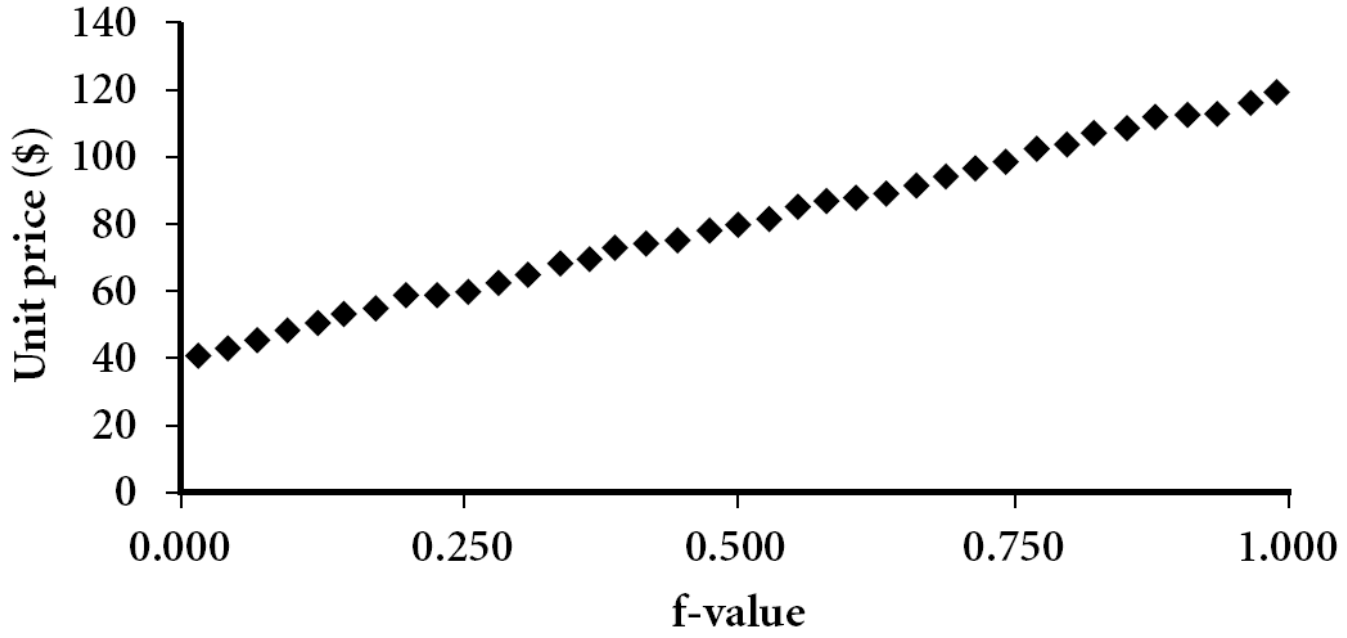
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

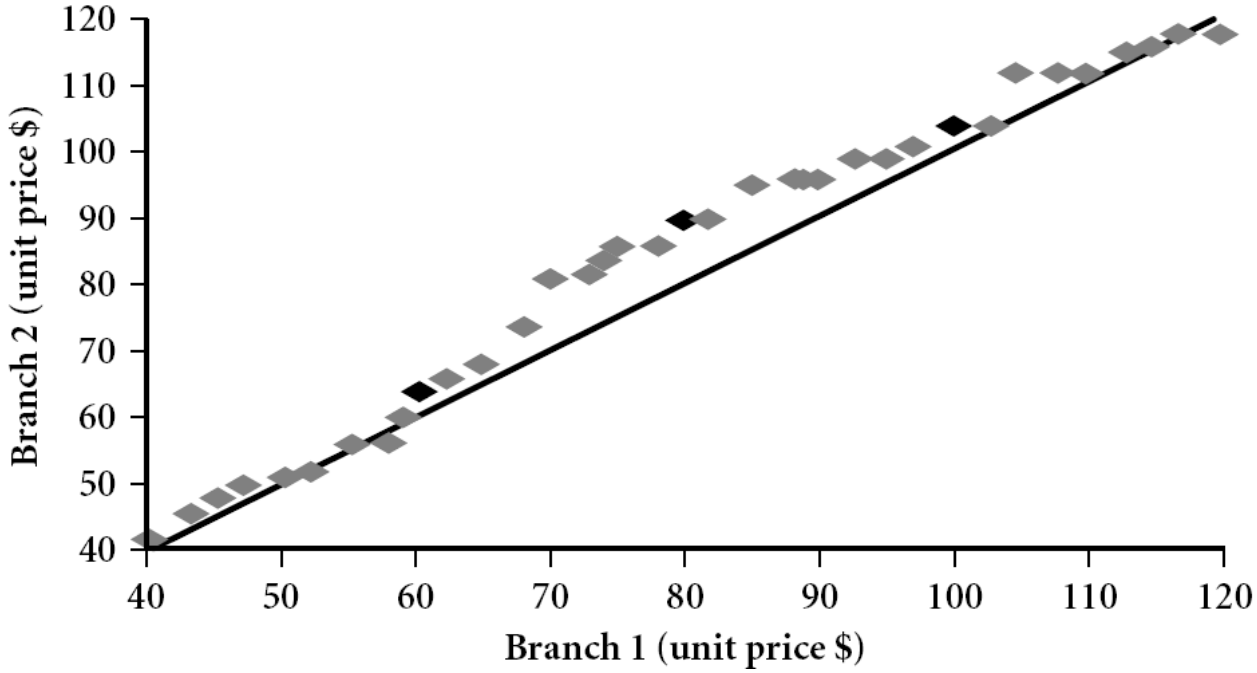
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



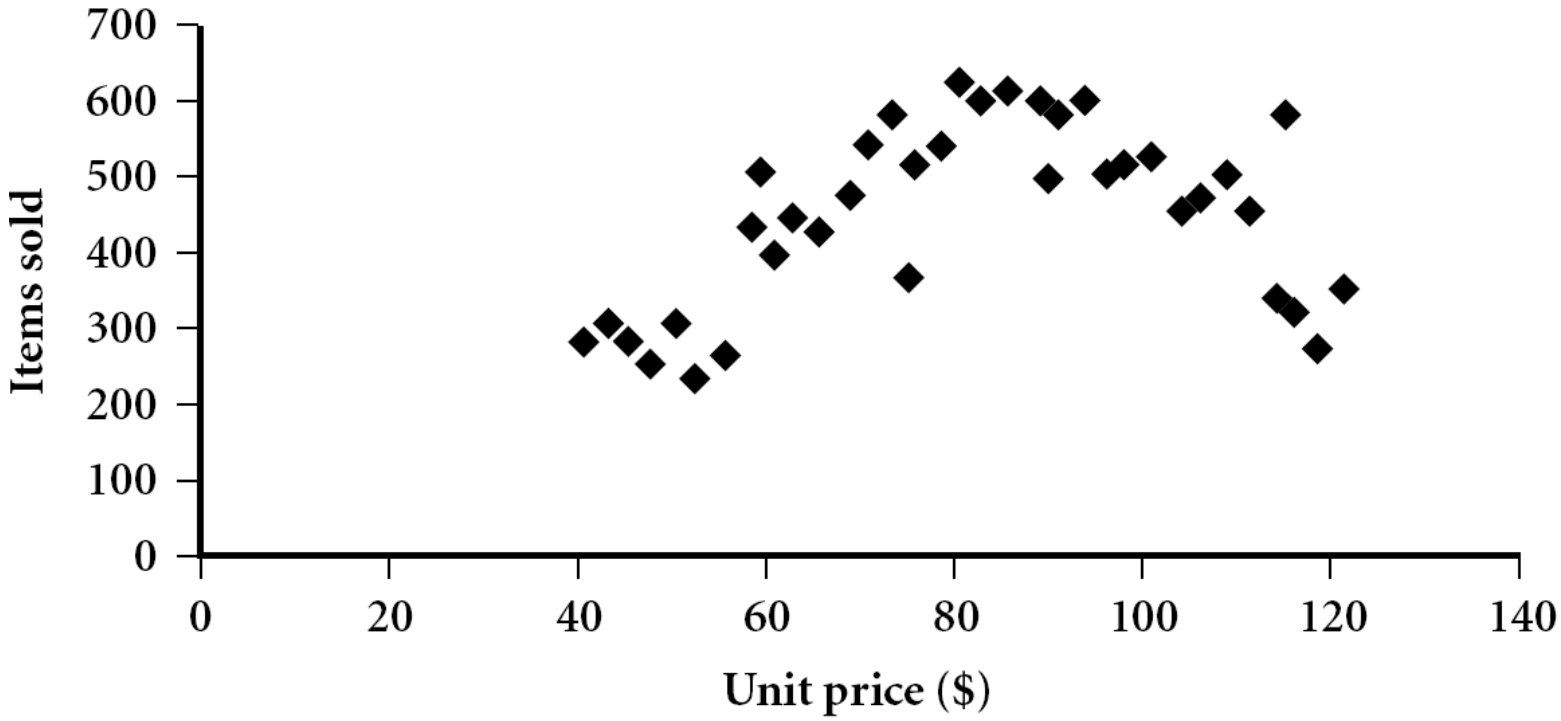
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

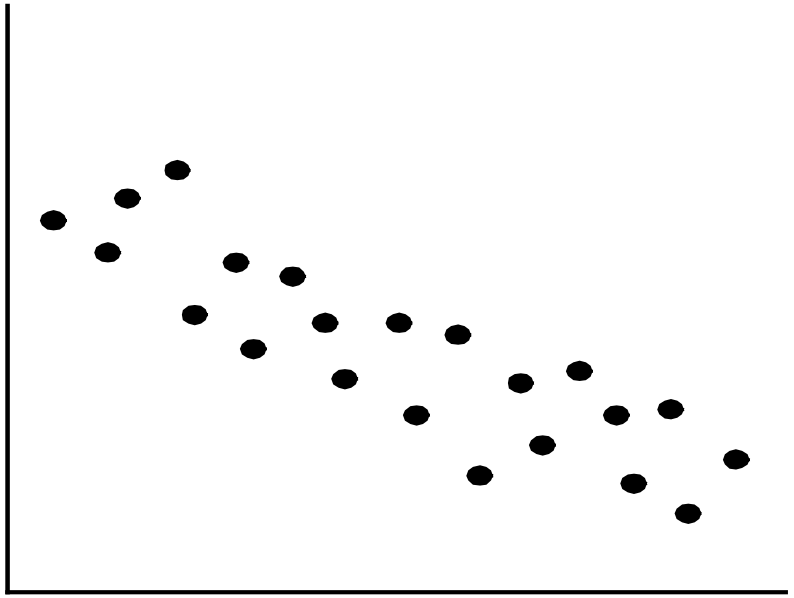


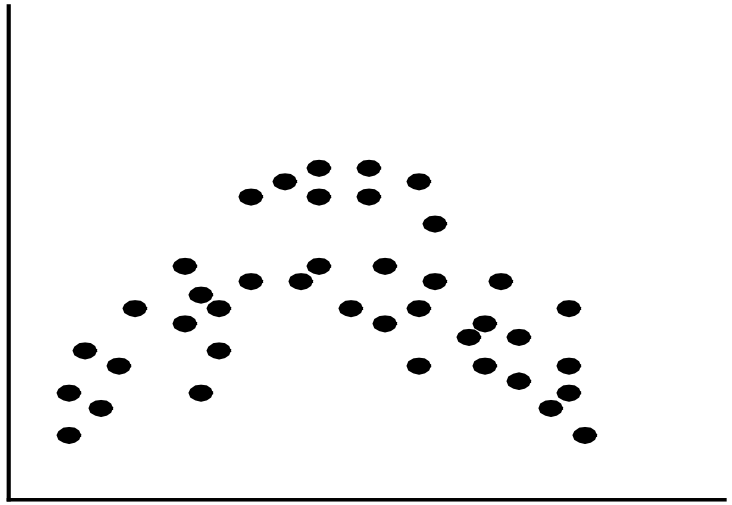
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data





- The left half fragment is positively correlated
- The right half is negative correlated

Summary

- Data engineering involves knowledge discovery and data mining
- Business knowledge is important for adoption of technologies supporting data mining.
- Properties of qualitative and quantitative data
- Numerical summaries of data
- Visualization of data