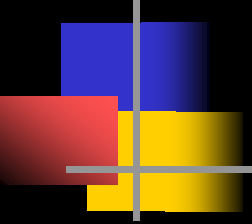


Chapter 1

Data science and Artificial Intelligence

Professor Jung Jin Lee
Soongsil University, Korea
New Uzbekistan University, Uzbekistan



Chapter 1 Data science and Artificial Intelligence

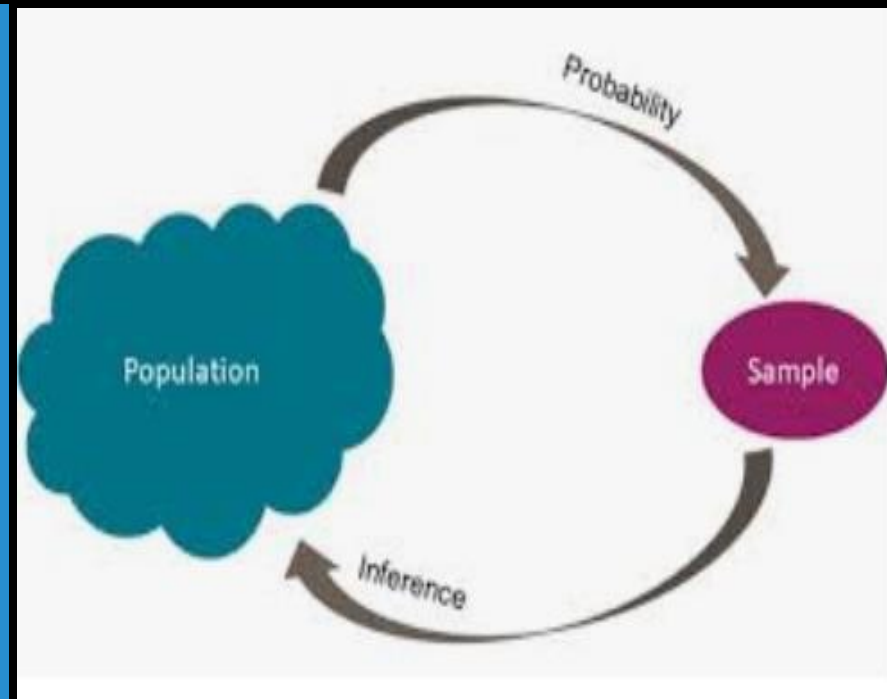
- 1.1 Statistics, data science, machine learning, and artificial intelligence
- 1.2 General process of data analysis
- 1.3 Data classification
- 1.4 Software programs for data analysis

1.1 Statistics, data science, machine learning, and artificial intelligence

❖ Statistics = 'State' + 'istics'



Descriptive Statistics

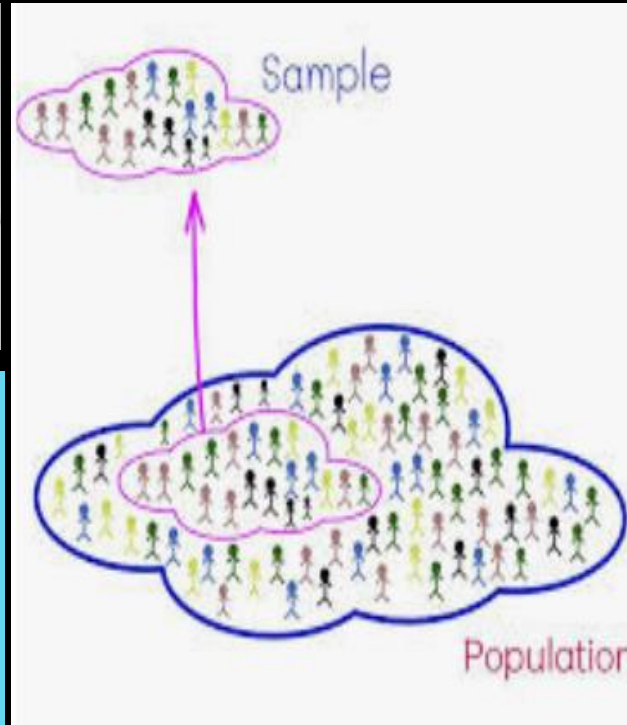
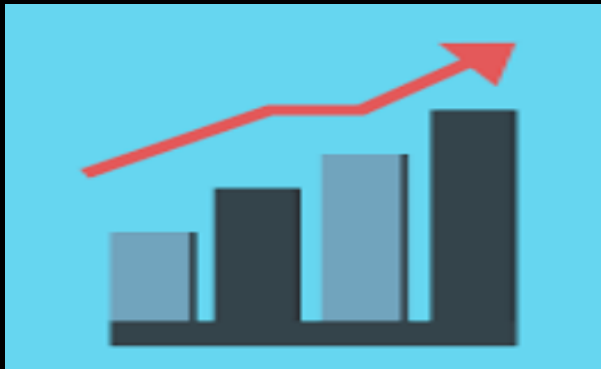


Inferential Statistics

1.1 Statistics, data science, machine learning, and artificial intelligence

❖ Statistics Application

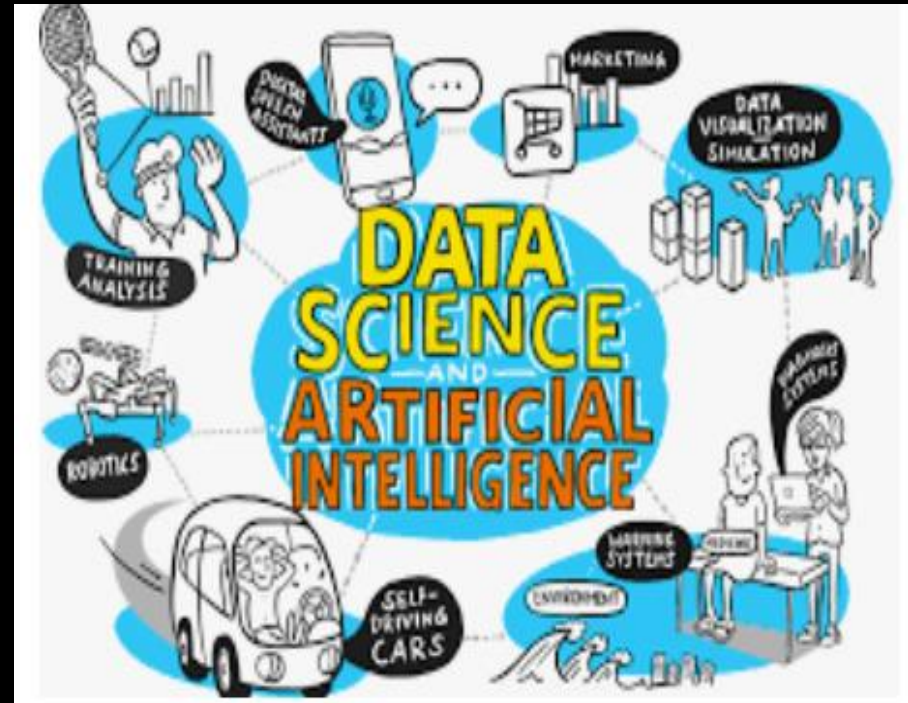
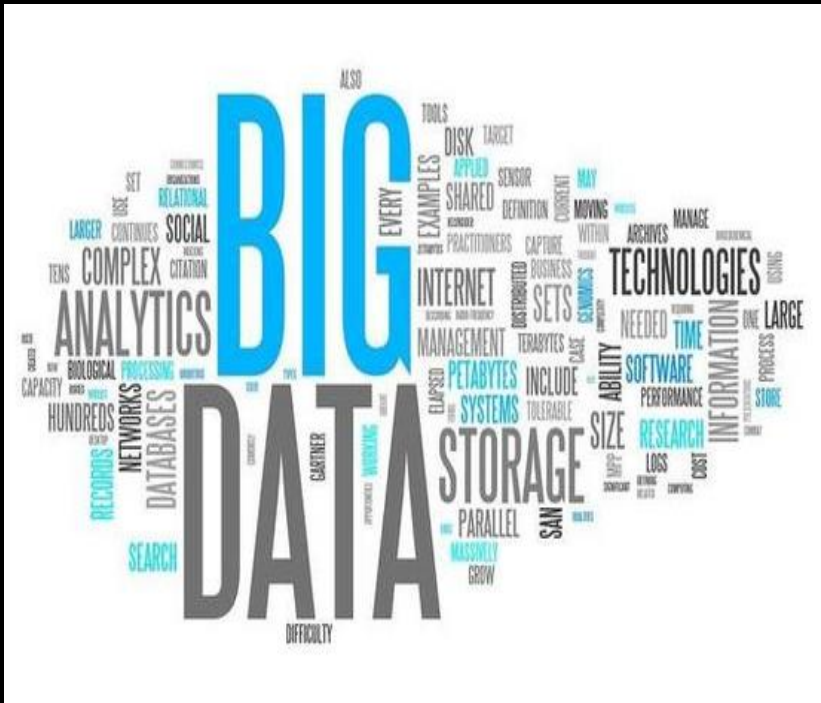
=> Predict sales, predict elections, test drugs, quality control



SPC vs **SQC**

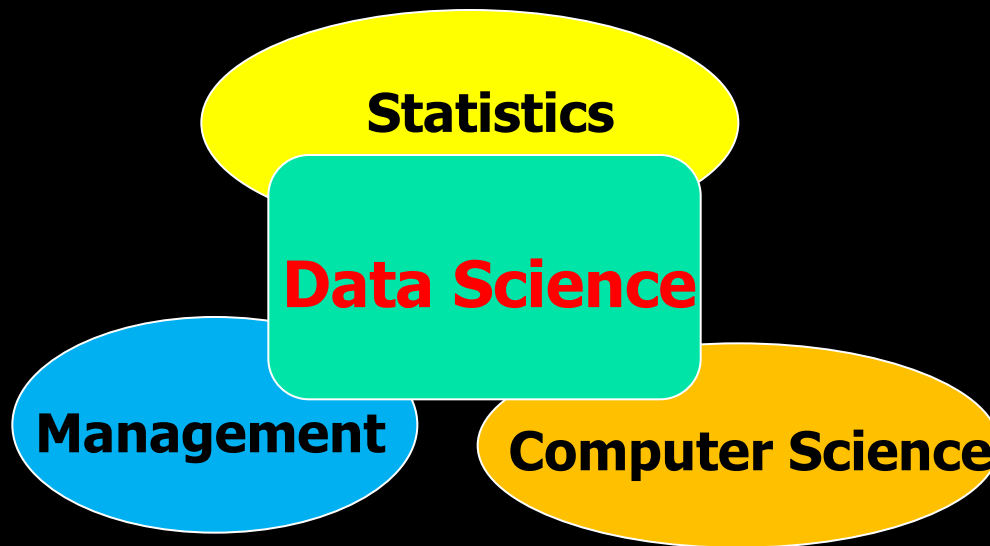
S = Statistical	S = Statistical
P = Process	Q = Quality
C = Control	C = Control

❖ **Big data => Data science**



1.1 Statistics, data science, machine learning, and artificial intelligence

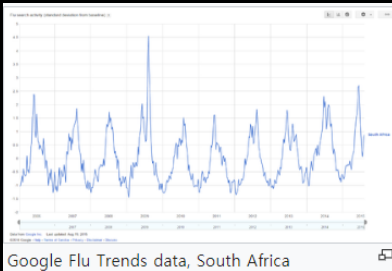
❖ **Data Science** is a fusion of several science
=> Collect big data, analyze and apply it in real life



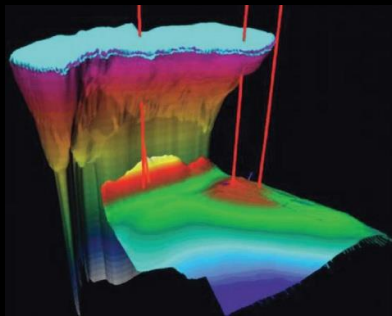
- Probability
- Estimation
- Testing
- Sampling
- Multivariate Stat Anal
- Database
- Information Retrieval
- Distributed Computing
- Artificial Intelligence
- Pattern Recognition
- Machine Learning
- Optimization
- MIS
- Marketing

1.1 Statistics, data science, machine learning, and artificial intelligence

❖ Example of Data Science



- Google Flu Trend to estimate



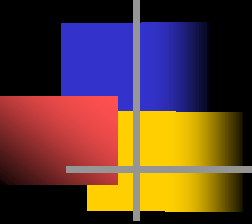
- Crude oil exploration



- Market basket analysis



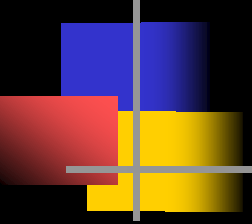
- Car insurance fraud detection



1.1 Statistics, data science, machine learning, and artificial intelligence

❖ Data mining

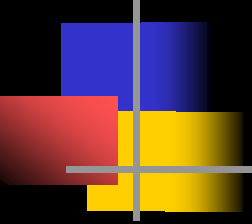
- Extract interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns **using statistical and mathematical models**.
 - Mining refers to extracting gold or minerals from mines.
 - Data mining implies extracting important patterns or knowledge from data.
- Knowledge extracted is used to **make decisions**.
- Similar terms:
 - Pattern analysis, knowledge discovery, knowledge extraction, data archeology, data dredging, business intelligence, etc.



1.1 Statistics, data science, machine learning, and artificial intelligence

❖ Machine learning

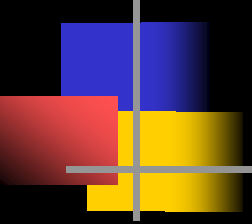
- Computer automatically learns rules from data to create a software program that solves problems.
- Most of the techniques used in machine learning are similar to those used in data mining.
- Machine learning algorithms are often like a black box, making it difficult to know why decisions were made.



1.1 Statistics, data science, machine learning, and artificial intelligence

❖ Artificial Intelligence (AI)

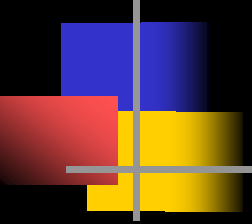
- Artificial intelligence is an extension of machine learning
- Machines that have intelligence to imitate human intelligence and perform complex tasks like humans.
- AI utilizes many techniques in data mining and machine learning, especially the **artificial neural network** model
- **Deep learning** is a simulation algorithm that trains the artificial neural network.



1.1 Statistics, data science, machine learning, and artificial intelligence

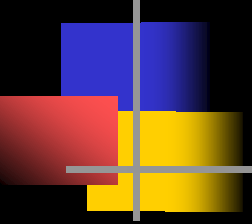
❖ Artificial Intelligence (AI)

- In 1955, Marvin Minsky in US built the first neural network, the SNARC system. Viktor Glushkov in Soviet Union created the All-Union Automatic Information Processing System (OGAS).
- In 1974, Paul Warboss proposed a back-propagation algorithm that could solve a multilayer neural network.
 - visible results on character recognition, speech recognition
 - The algorithm sometimes failed to find a solution.
- In 2006, Geoffrey Hinton announced the **deep learning** algorithm
 - Surpassing result on artificial neural network. computer vision
 - Google DeepMind's AlphaGo popularized deep learning
- In 2022, **generative AI**.
 - OpenAI's ChatGPT and Drawing AI applied to personal hobbies and work applications and the practical application of AI



❖ Potential Applications

- Decision support
 - Market analysis and management
 - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, quality control, competitive analysis
 - Fraud detection and management
- Other Applications
 - Text mining (news group, email, documents) and Web analysis.
 - Intelligent query answering



❖ Potential Applications

- Finance planning and asset evaluation
 - cash flow analysis and prediction
 - contingent claim analysis to evaluate assets
 - cross-sectional and time series analysis
- Internet Web Surf-Aid
 - Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.
- Astronomy
 - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

❖ Application to multimedia databases

▪ Refining or combining searches



Search for "blue sky"
(top layout grid is blue)

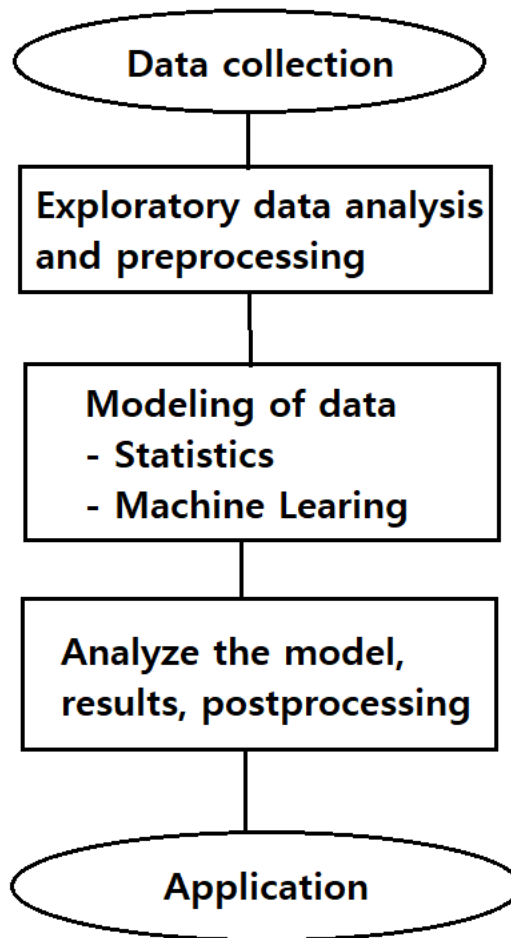


Search for "airplane in blue sky"
(top layout grid is blue and
keyword = "airplane")



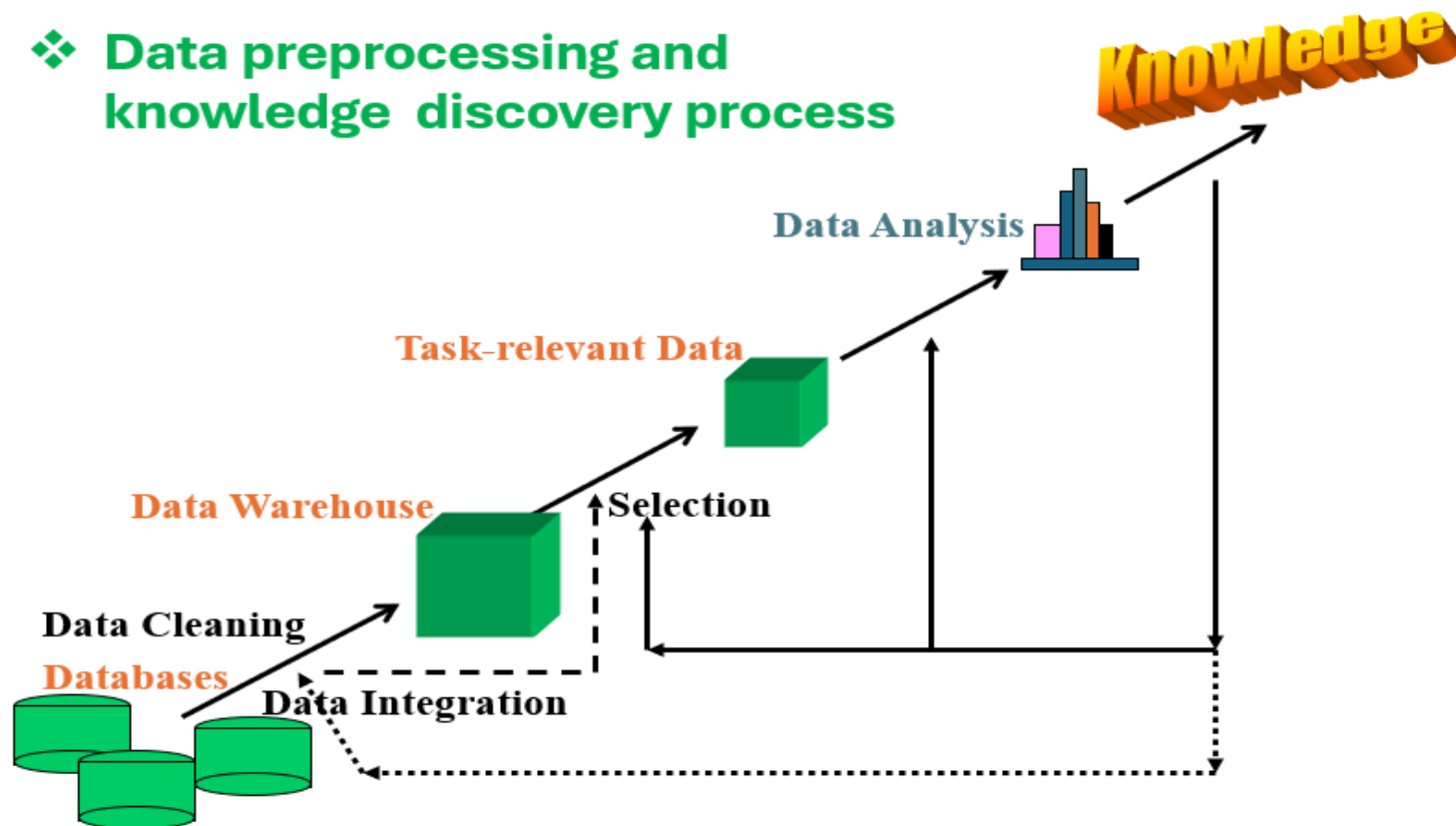
Search for "blue sky and
green meadows"
(top layout grid is blue
and bottom is green)

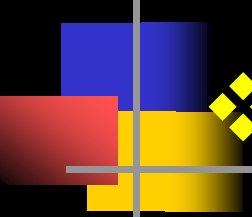
1.2 General process of data analysis



1.2 General process of data analysis

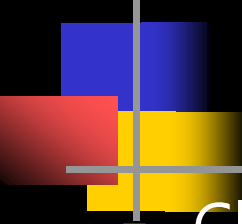
❖ Data preprocessing and knowledge discovery process





❖ Knowledge discovery process

- Learning the application domain:
 - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (take 60% of effort!)
- Data reduction and transformation:
 - Find useful features, dimensionality/variable reduction
- Choosing functions of data modeling
 - summarization, regression, classification, clustering
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
 - visualization, transformation, removing redundant patterns
- Use of discovered knowledge



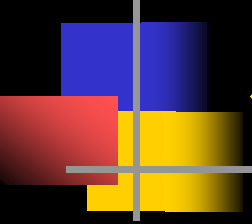
❖ Data modelling functionalities (1)

■ Classification and Prediction

- Finding models (functions) that describe and distinguish classes or concepts for future prediction
- E.g., classify countries based on climate, or classify cars based on gas mileage
- Presentation: decision tree, classification rule, neural network
- Prediction: Predict some unknown or missing numerical values

■ Cluster analysis

- Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
- Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity



❖ Data modelling functionalities (2)

- Outlier analysis

- Outlier: a data object that does not comply with the general behavior of the data
- It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

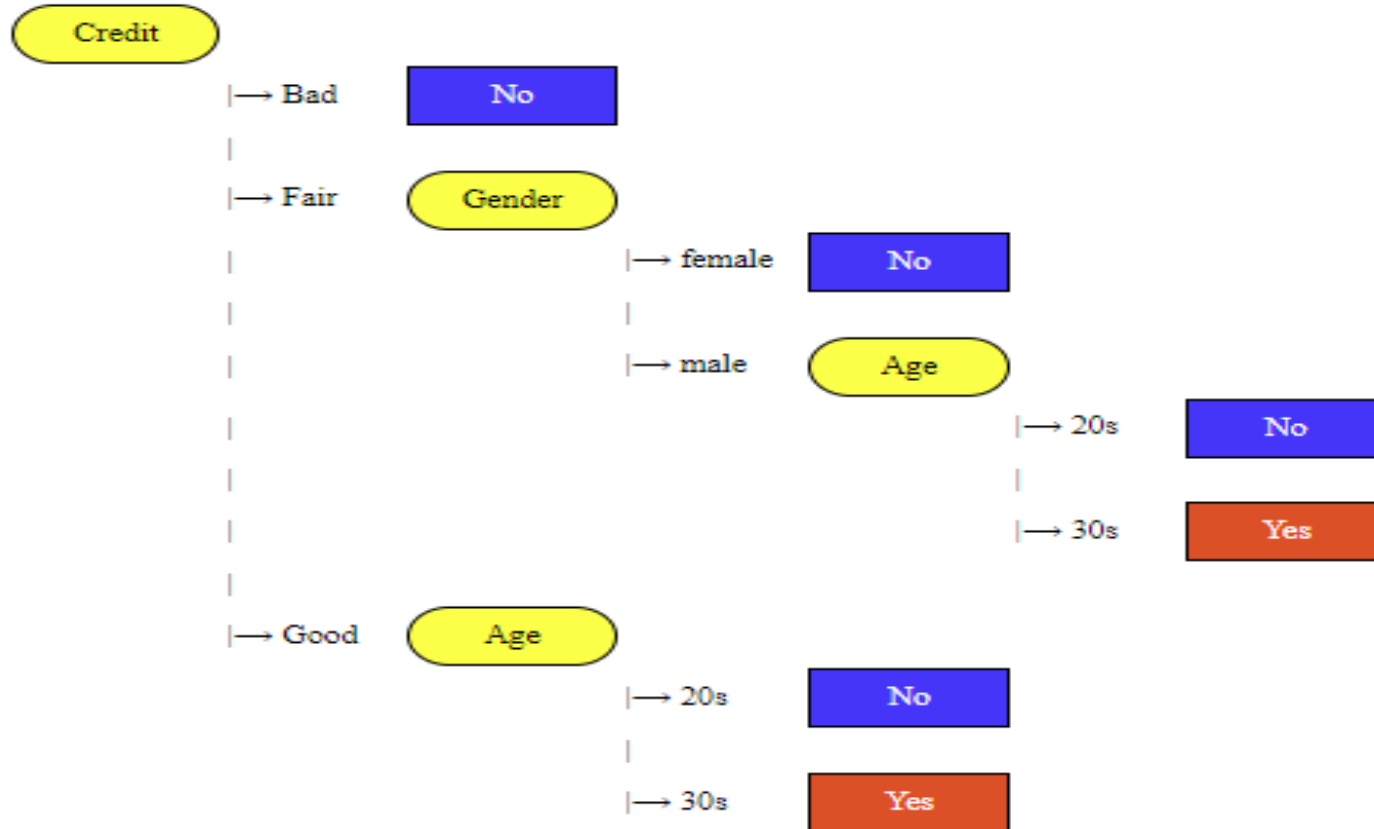
- Trend and evolution analysis

- Trend and deviation: regression analysis
- Sequential pattern mining, periodicity analysis
- Similarity-based analysis

- Other pattern-directed or statistical analyses

❖ Example of decision tree

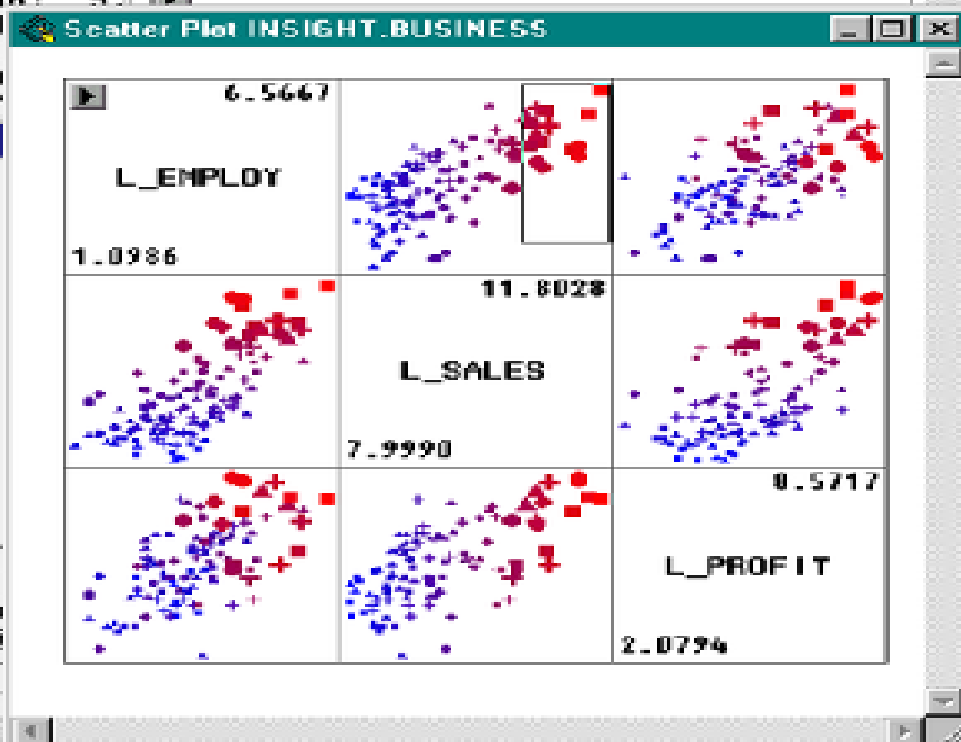
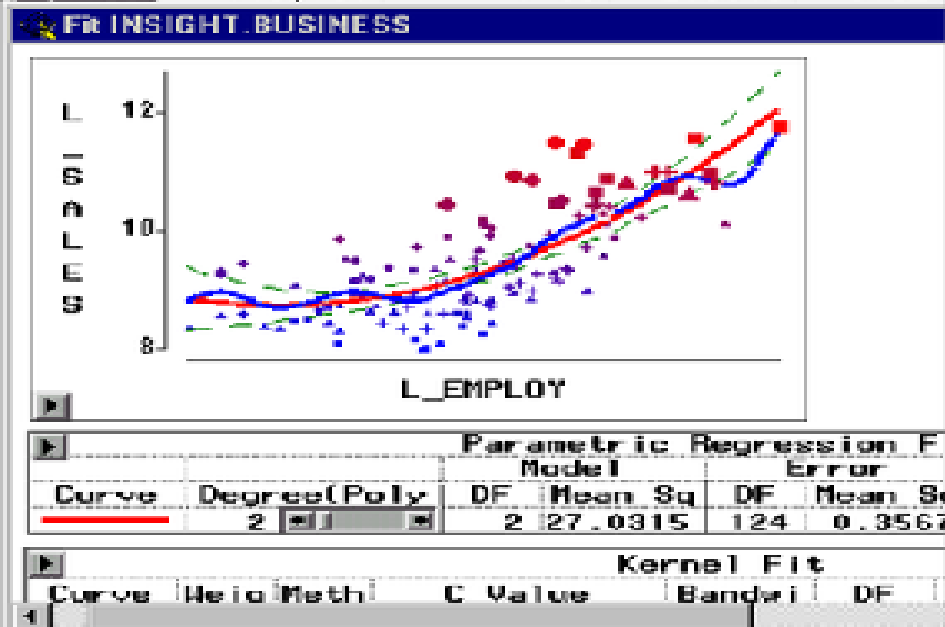
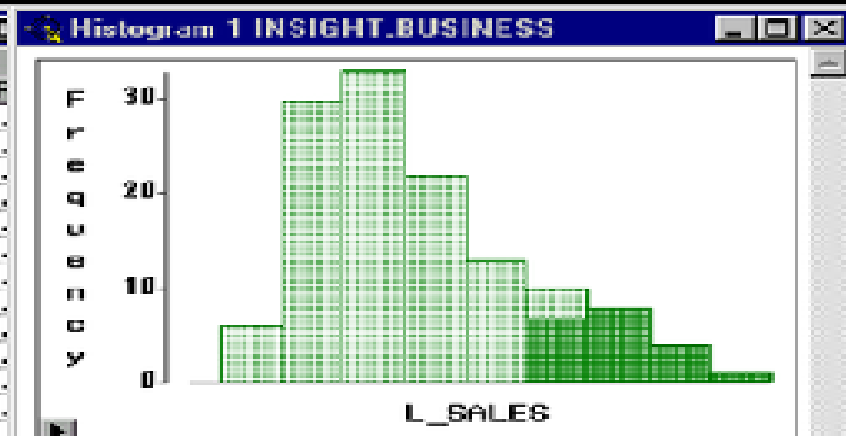
Decision Tree



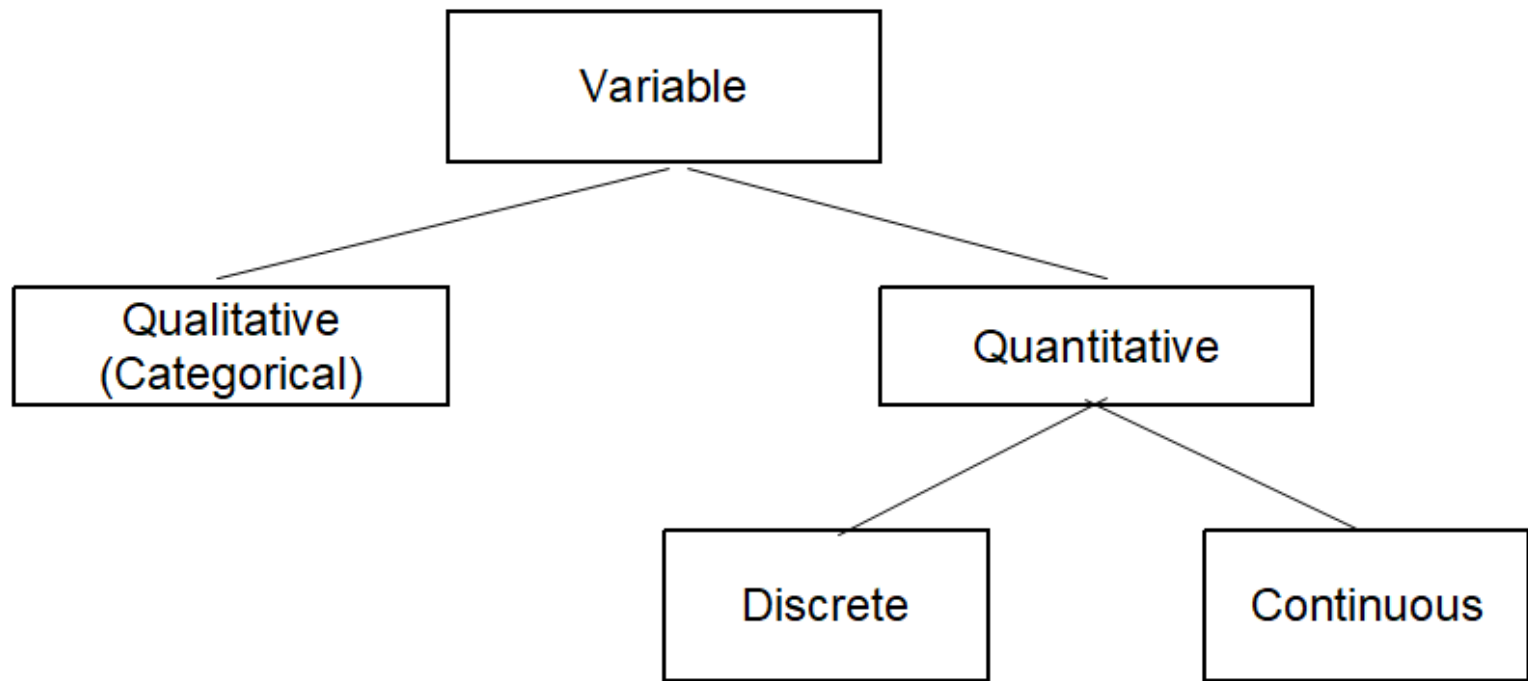
❖ Visualization

INSIGHT.BUSINESS

		Int	Int	Int	Int	
	SALES	PROFITS	L_EMPLOY	L_SALES	L_PROFIT	
39	\$9,414	\$-236	2.6391	9.1500		
40	\$9,491	\$907	4.5326	9.1501	6.	
41	\$3,037	\$58	3.2958	8.0186	4.	
42	\$60,823	\$4,318	5.4027	11.0157	8.	
43	\$8,135	\$506	4.7958	9.0039	6.	
44	\$133,622	\$2,468	6.5667	11.8028	7.	
45	\$11,164	\$629	4.4659	9.3204	6.	
46	\$7,006	\$650	3.1355	8.8545	6.	
47	\$7,103	\$396	3.6376	8.8683	5.	
48	\$3,319	\$91	3.4340	8.1074	4.	
49	\$6,900	\$142	3.7612	8.8393	4.	
50	\$4,963	\$24	2.1972	0.5090	3.	
51	\$95,790	\$220	4.5109	10.4056	5.	
52	\$11,671	\$90	3.3673	9.364		
53	\$12,857	\$11	1.6094	9.461		
54	\$8,782	\$2,298	3.4012	9.080		
55	\$13,731	\$-38	2.5649	9.527		



1.3 Data classification



1.3 Data classification

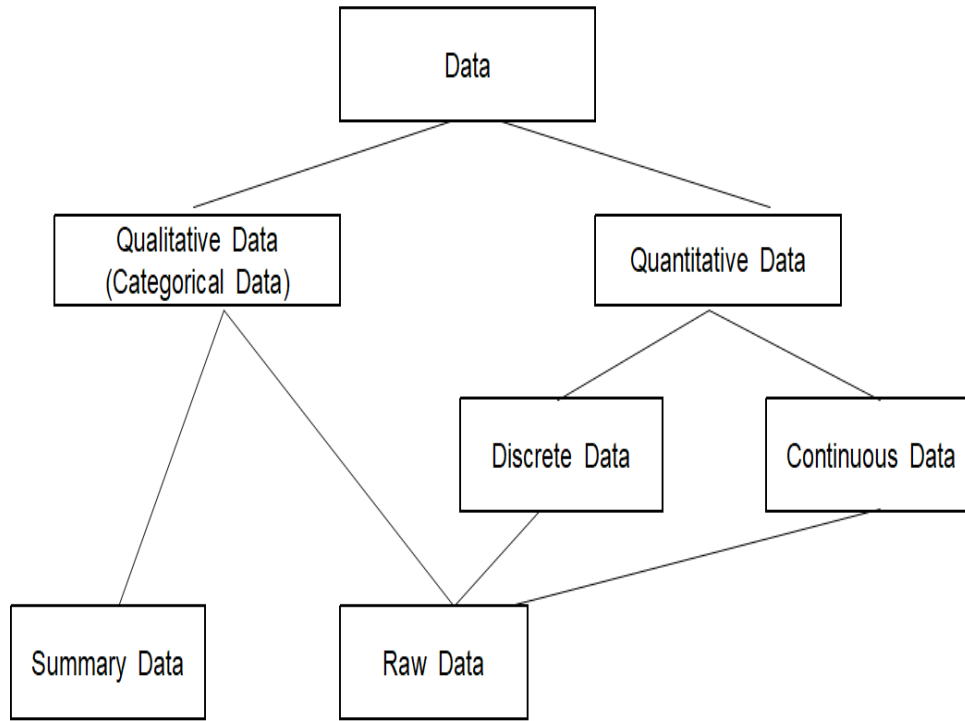
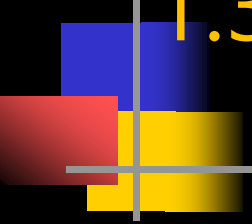


Table 1.3.1 Raw data by gender survey

row	Gender
1	male
2	female
3	male
4	female
5	male
6	male
7	male
8	female
9	female
10	male

Table 1.3.2 frequency table data for the gender

Gender	Number of Students
Male	6
Female	4



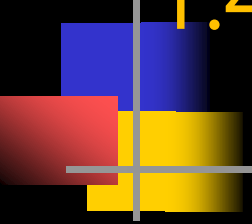
1.3 Data classification

❖ Notation of data

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

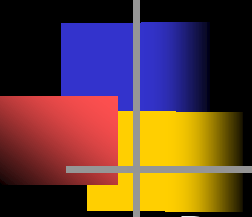
$$\{(x_{i1}, x_{i2}, \dots, x_{im}), i = 1, 2, \dots, n\}$$

$$\{(x_{i1}, x_{i2}, \dots, x_{im}, y_i), i = 1, 2, \dots, n\}$$



1.4 Software for data analysis

- SAS and SPSS:
 - good but commercial, expensive
- R and Python: freeware
 - Need programming skill
- eStat: freeware
 - Educational, easy

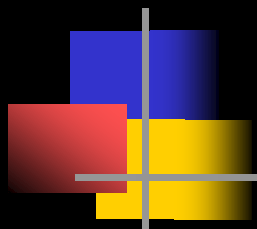


Summary

- Data science:
 - discovering interesting patterns from large amounts of data
 - great demand, with wide applications

- Knowledge discovery process:
 - data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation

- Data modeling functionalities:
 - characterization, discrimination, classification, clustering, outlier and trend analysis, etc.



Thank you !!!