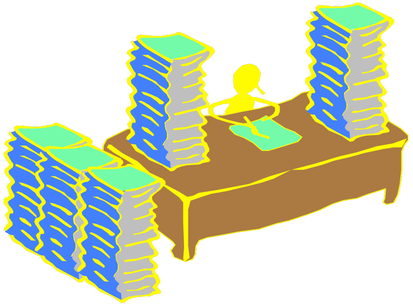
Business Intelligence

KDD

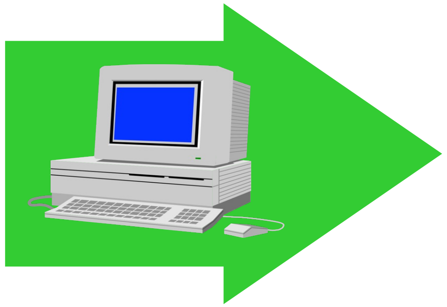
Or Peretz

Knowledge Discovery in Databases

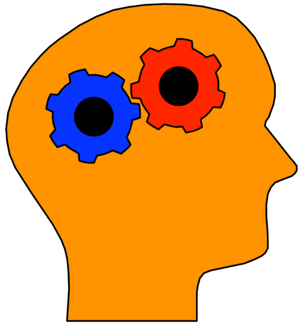
KDD is the automatic extraction of non-obvious, hidden knowledge from large volumes of data.



10^6-10^{12} bytes:
we never see the whole data set, so will put it in the memory of computers

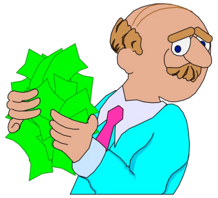


Then run Data Mining algorithms



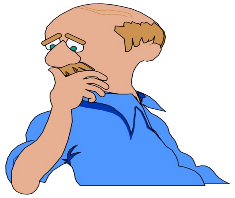
What is the knowledge?
How to represent and use it?

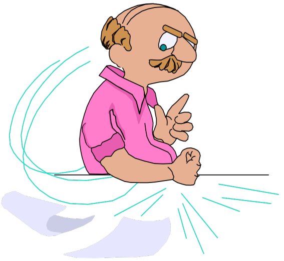
Data, Information, Knowledge



We often see **data** as a string of bits, or numbers and symbols, or “objects” which we collect daily.

Information is data stripped of redundancy and reduced to the minimum necessary to characterize the data.

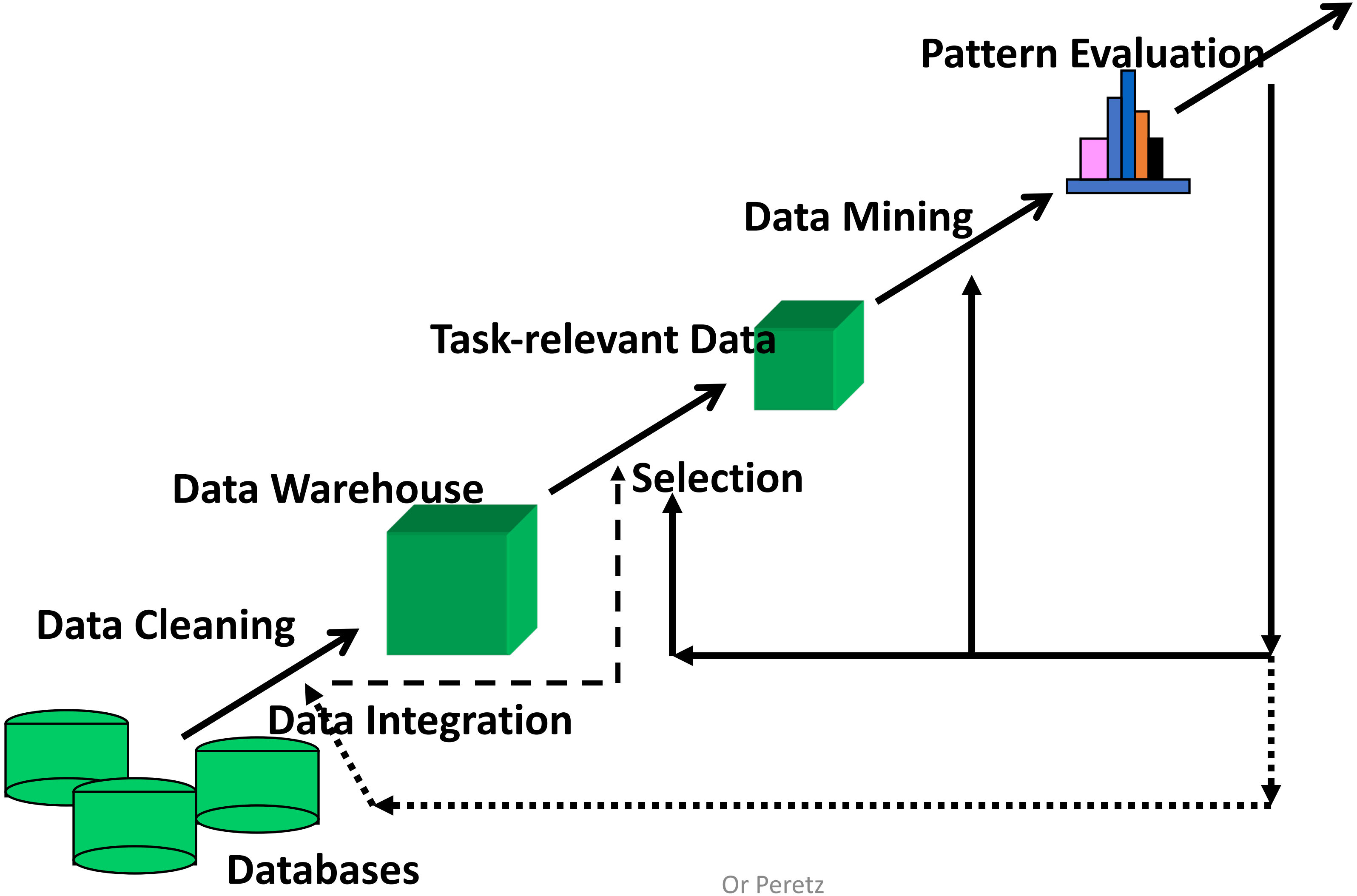




Knowledge is integrated information, including facts and their relations, which have been perceived, discovered, or learned as our “mental pictures”.

Knowledge can be considered data at a high level of abstraction and generalization.

The KDD Process



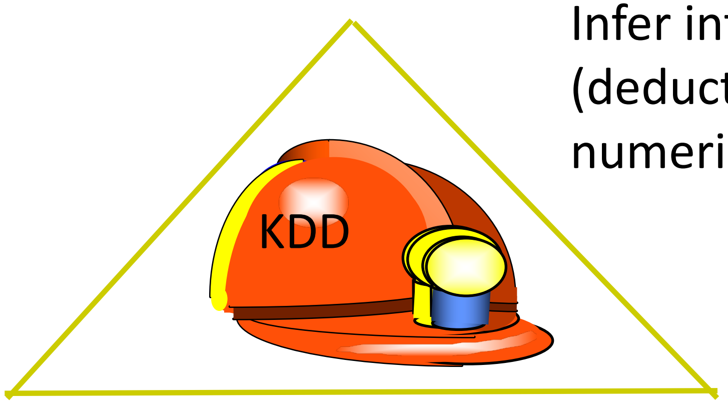
Main Fields

Statistics

[data warehouses: integrated data]

[OLAP: On-Line Analytical Processing]

Infer info from data (deduction & induction, mainly numeric data)



Databases

Store, access, search, update data (deduction)

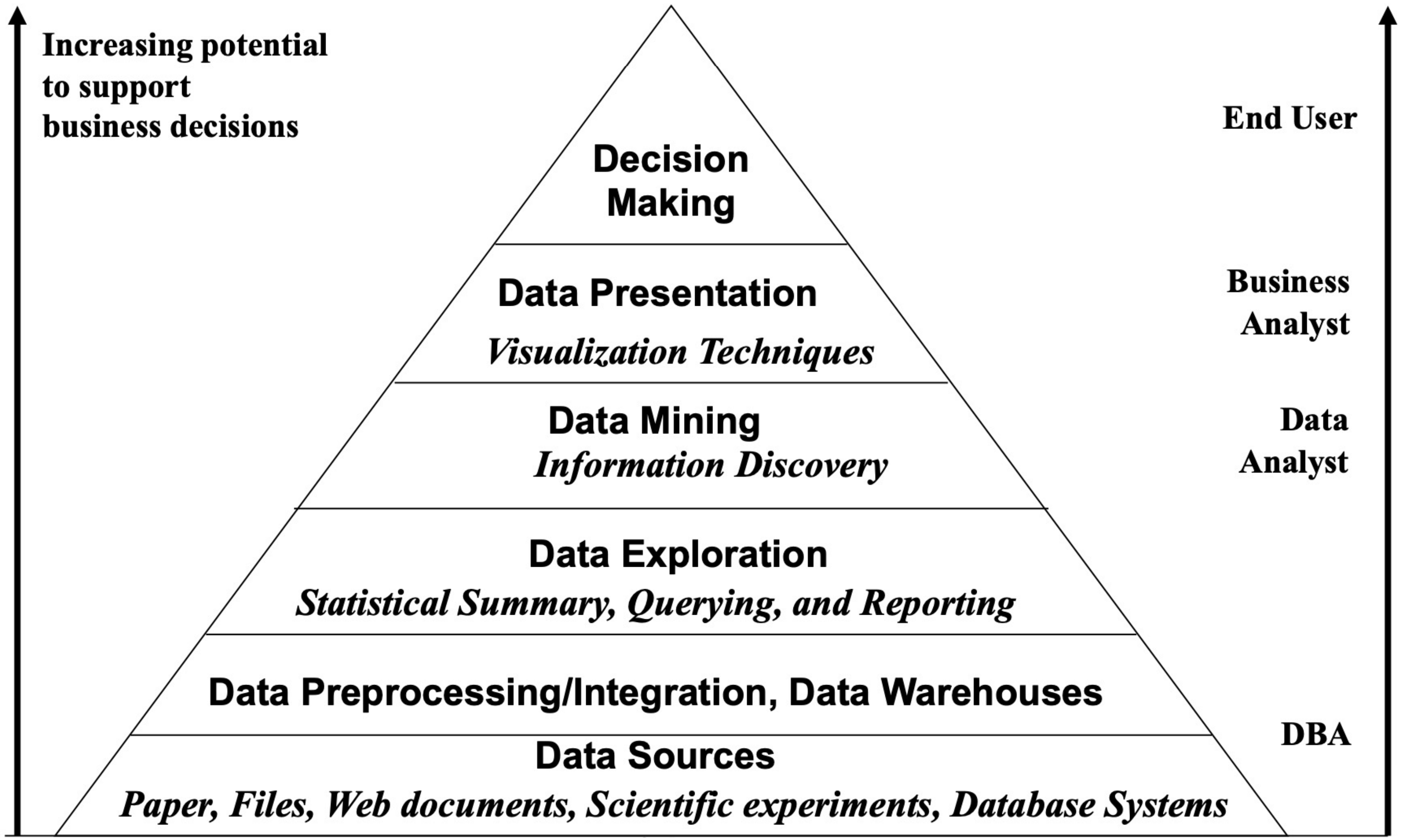
Machine Learning

Computer algorithms that improve automatically through experience (mainly induction, symbolic data)

Data Mining

- Knowledge discovery (mining) in databases (KDD)
- knowledge extraction
- data/pattern analysis
- data archeology
- information harvesting

Data Mining



Increasing potential
to support
business decisions
End User
Decision
Making
Business
Analyst
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data
Analyst
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
DBA
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

Applications

- Market analysis and management

  Target marketing, customer relationship management (CRM)

- Risk analysis and management
  - Forecasting, customer retention, improved underwriting, competitive analysis
- Fraud detection
- Detection of unusual patterns (outliers)

Market Analysis

- Where does the data come from?

  Credit card transactions, loyalty cards, discount coupons, customer complaint calls, surveys …

- Target marketing
  - **Find clusters of “model” customers who share the same characteristics.** For example, most customers with income level 60k – 80k with food expenses $600 - $800 a month live in that area.
  - **Determine customer purchasing patterns over time**. For example, customers who are between 20 and 29 years old, with income of 20k – 29k usually buy this type of CD player

- Cross-market analysis

  customers who buy computer A usually buy software B

Market Analysis

- **Customer requirement analysis**

  Identify the best products for different customer, Predict what factors will attract new customers

- **Summary information**
  - Multidimensional summary reports
    - Summarize all transactions of the first quarter from three different branches
    - Summarize all transactions of last year from a particular branch
  - Statistical summary information
    - What is the average age for customers who buy product A?
- **Fraud detection**: find outliers of unusual transactions
- **Financial planning**: summarize and compare the resources and spending

Types of Data

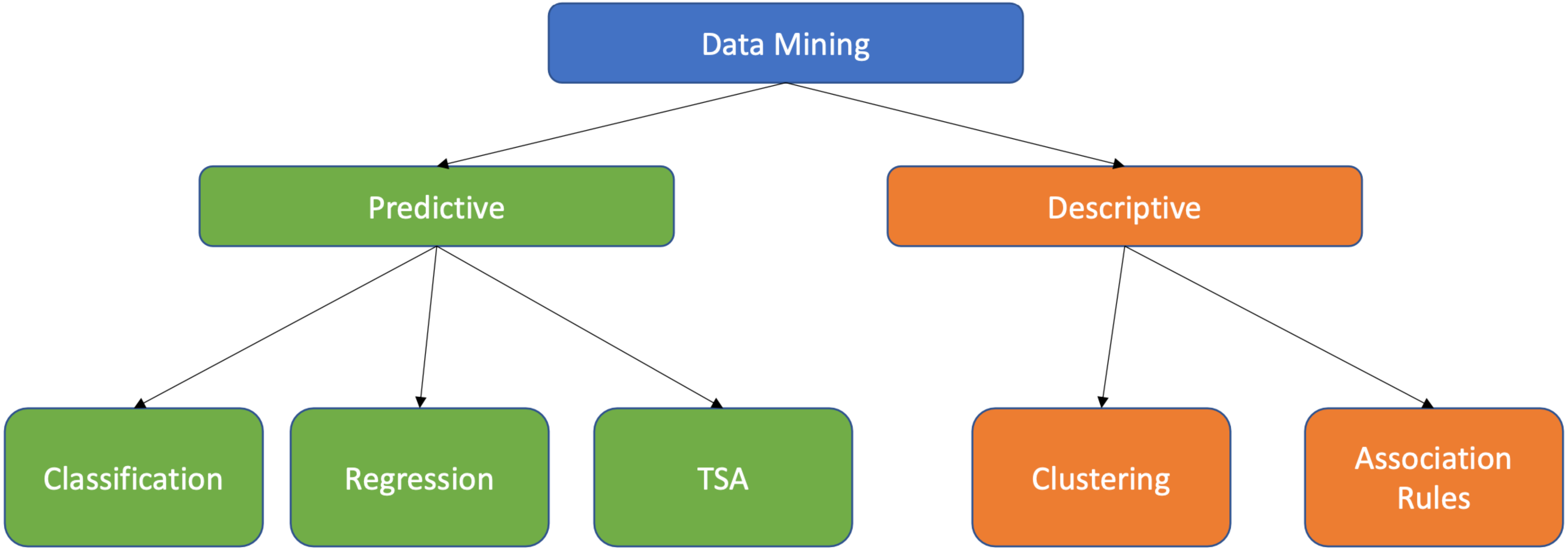
Database-oriented data sets and applications

- Relational database, data warehouse, transactional database

Advanced data sets and advanced applications

- Object-Relational Databases
- Time-Series databases
- Spatial Databases
- Text databases and Multimedia databases
- Data Streams
- The World-Wide Web

Data Mining Models



Data Mining

Predictive

Descriptive

Classification

Regression

TSA

Clustering

Association Rules

Data Mining Models - Predictive

- **Regression**: (linear or any other polynomial)
- **Nearest neighbours**
- **Decision tree**
- **Probabilistic models**
- **Neural networks**: partition by non-linear boundaries

Data Mining Models - Predictive

- **Direct Marketing**: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
- Approach:
  - Use the data for a similar product introduced before.
  - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
  - Use this information as input attributes to learn a classifier model.

Data Mining Models - Predictive

- **Fraud Detection**: Predict fraudulent cases in credit card transactions.
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

Data Mining Models – Bayesian Learning

- Assume a probability model on generation of data.
- Apply bayes theorem to find most likely class as:

$$\text{predicted class} : c = \max_{c_j} p(c_j \mid d) = \max_{c_j} \frac{p(d \mid c_j) p(c_j)}{p(d)}$$

- Naïve bayes: Assume attributes conditionally independent given class value
- Easy to learn probabilities by counting

$$c = \max_{c_j} \frac{p(c_j)}{p(d)} \prod_{i=1}^{n} p(a_i \mid c_j)$$

Data Mining Applications - Descriptive

- Customer segmentation for targeted marketing
  - Group/cluster existing customers based on time series of payment history such that similar customers in same cluster.
  - Identify micro-markets and develop policies for each
- Collaborative filtering:
  - group based on common items purchased
- Text clustering
- Compression

Data Mining Applications – Collaborative Filtering

Given database of user preferences, predict preference of new user

Example: predict what new movies you will like based on

- your past preferences
- others with similar past preferences
- their preferences for the new movies

Example: predict what books/CDs a person may want to buy (and suggest it, or give discounts to tempt customer)

Data Mining Techniques

Hypothesis Testing

Find model to explain behavior by creating and then testing a hypothesis about the data.

- H_0 – Null hypothesis (hypothesis to be tested)
- H_1 – Alternative hypothesis

Example:

H_0 - men and women get the same salary

H_1 - women's salary is higher than men's salary

Similarity Measures

Determine similarity between two objects. Similarity characteristics:

Given $x, y \in D$

$$sim(x,x) = sim(y,y) = 1$$

$$sim(x,y) = 0 \; if \;\; x \; and \; y \; are \; not \; alike \; at \; all$$

$$sim(x,y) < sim(z,y) \; if \;\; z \; is \; more \; like \; y \; than \; x$$

Similarity can be distance or any other measure that correlated to the data.

Similarity Measures

Sorensen-Dice

$$sim(x,y) = \frac{2 \cdot |X \cap Y|}{c|X| + |Y|}, \qquad c = \frac{\sum x_i y_i}{\sum x_i \cdot sign(y_i)}$$

Jaccard

$$sim(x,y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

Euclidean Distance

$$sim(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_k - y_k)^2}$$