Data Mining

Data Mining : An Introduction

Outline

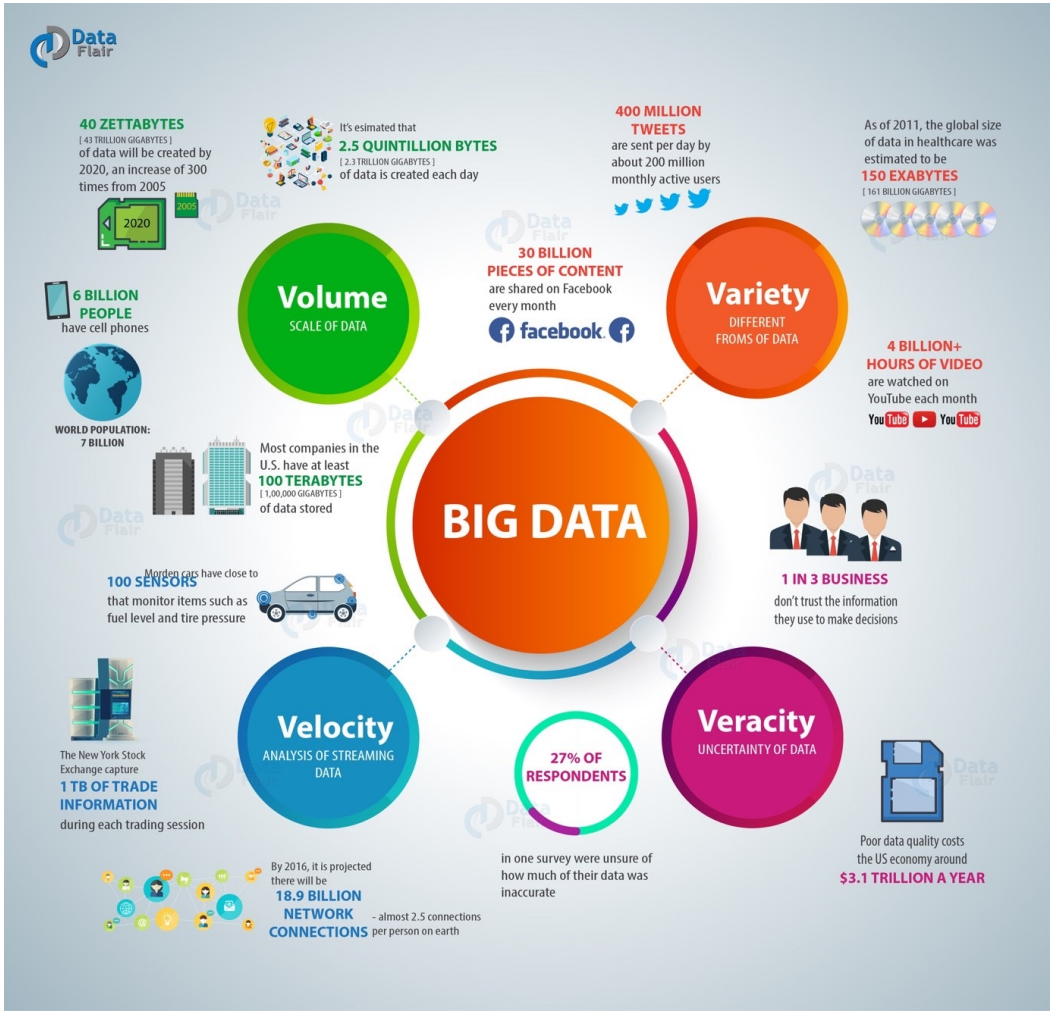
- **Introduction: Data Flood**
- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- Data Mining Techniques

Big Data

- Lots and lots of data:
  - Bank, telecom, business transactions (online and offline), ...
  - Scientific data: astronomy, climate, biology, medicine, chemistry, etc
  - Web, text, and e-commerce

Big Data

- The characteristics of Big Data:
  - Volume
  - Variety
  - Velocity
  - Veracity

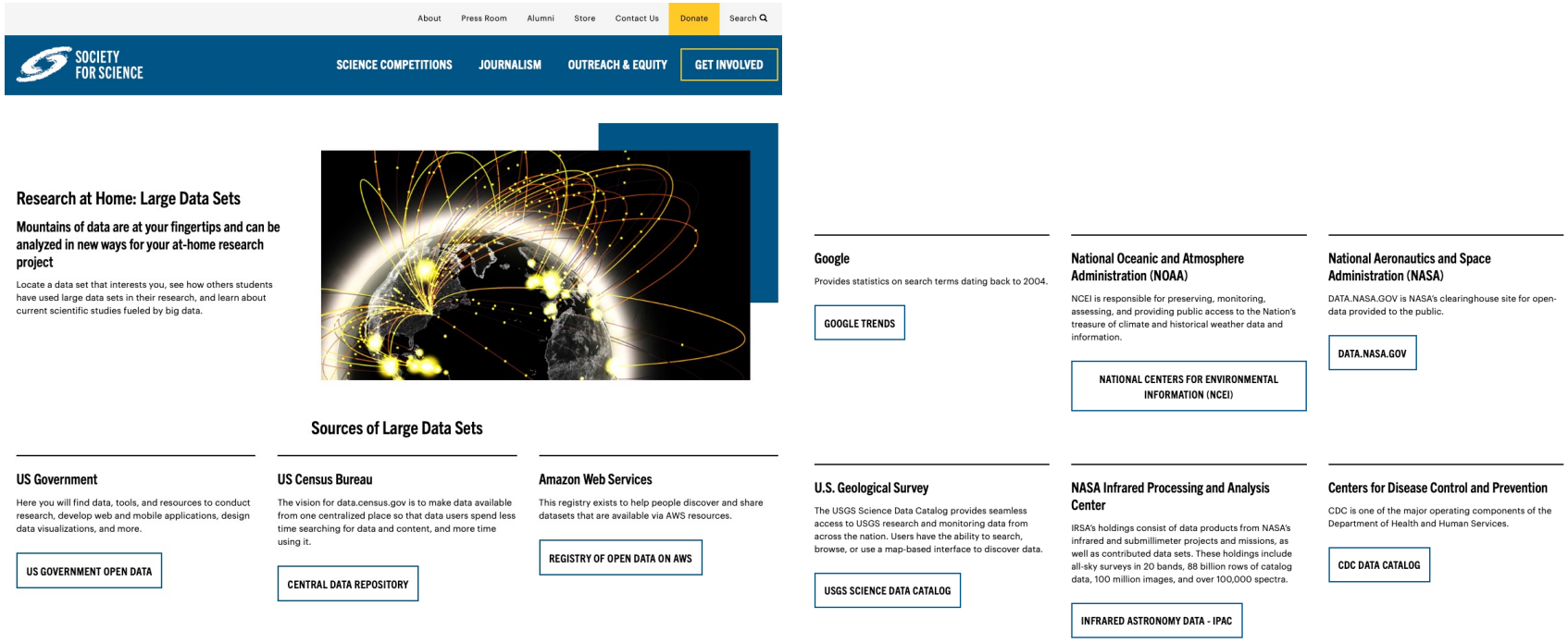


Market Growth

- The global **big data market size** is projected to grow from USD 138.9 billion in **2020** to USD 229.4 billion by 2025, at a Compound Annual **Growth** Rate (CAGR) of 10.6% during the **forecast** period.

Large Data Sets

- Society For Science
  - https://www.societyforscience.org/research-at-home/large-data-sets/



Large Databases

- Astronomy
  - Sloan Digital Sky Survey (SDSS) (http://www.sdss.org/, https://www.sdss5.org (newer))
- Biology/Medicine
  - National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/)
  - PubMed has over 33 million records (2021) (https://www.ncbi.nlm.nih.gov/pubmed/)

Data Sets For Data Mining Research

- UCI Machine Learning Repository:

  https://archive.ics.uci.edu/ml/index.php

- Kaggle Data Sets:

  https://www.kaggle.com/datasets

Outline

- Introduction: Data Flood
- **Data Mining Application Examples**
- Data Mining & Knowledge Discovery
- Data Mining Techniques

Data Mining Applications

- Business
  - CRM (Customer Relationship management) Tasks:
    - attrition prediction
    - targeted marketing:
      - cross-sell, customer acquisition
    - credit-risk
    - fraud detection
  - Industries
    - banking, e-commerce, retail sales, …

Data Mining Applications

- Health Science
  - Predictive medicine
    - Predict outbreaks of health problems,
    - Precision medicine using genomic data
  - Management of healthcare and measuring the effectiveness of certain treatments
    - Compare and contrast symptoms, causes and courses of treatment to find the most effective course of action for a certain illness or condition
  - Drug development and design.
    - Design/Determine the chemical compounds that would serve as effective drug treatments for a variety of diseases
  - Detection of health insurance fraud and abuse

Data Mining Applications

- Science and Engineering
  - Data from sensor networks and data from moving objects
    - Wireless sensor networks
    - RFID tagged moving objects
  - Spatial, temporal, spatiotemporal, and multimedia data
    - Images: Satellite images, medical images, images from various science fields (Astronomy(e.g., Galaxy Zoo project), chemistry, biology, physics, etc.)
    - Spatial: data with geographic location info
    - Time series data: use sequences of recorded values (e.g. physics, finance, medicine and music) for prediction
    - Multimedia data: discovering relationships between objects or segments within multimedia document components

Data Mining Applications

- Education:
  - Predict student success and attrition
- Government:
  - Video surveillance
    - Face recognition
  - Crime prediction, detection, and prevention
    - profiling tax cheaters, predict regions which have high prob for crime occurrence and can visualize crime prone area …
- Sports:
  - Soccer, Football
    - Predict player injuries by analyzing data from workouts over a period of time (European soccer club AC Milan)
    - Predict future physical performance based on physical aptitude test data (e.g. NFL Combine)
- Web:
  - Search engines, targeted advertising, web and text mining, …

Customer Attrition: Case Study

- Situation: Attrition rate for mobile phone customers is around 25-30% a year!
- With this in mind, what is a data mining task?
  - Assume we have customer information for the past N months.

Customer Attrition: Case Study

- Task:
  - Predict who is likely to attrite next month.
  - Estimate customer value and what is the cost-effective offer to make to this customer.

Customer Attrition Results

- Verizon Wireless built a customer data warehouse
- Identified potential attriters
- Developed multiple regional models
- Targeted customers with high inclination to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

Assessing Credit Risk: Case Study

- Situation: Person applies for a loan
- Task: Should a bank approve the loan?
- Note: People who have the best credit often don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle.

Credit Risk - Results

- Banks develop credit models using variety of machine learning methods.
- Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan
- Widely deployed in many countries

e-Commerce

- A person buys a book (or a product) from a vendor at Amazon

What is the data mining task from the Amazon vendor's stand?

Successful e-commerce – Case Study

- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful
  - Short video recommendation – TikTok, Youtube
  - Movie recommendation

Other Examples

- Fraud Detection
  - Credit card fraud detection
  - Detection of money laundering
    - US Treasury
  - Securities fraud
    - NASDAQ KDD system, NASD Regulation Advanced-Detection System (ADS)
  - Healthcare fraud
  - Phone fraud
    - AT&T, Bell Atlantic, British Telecom/MCI

Genomic Microarrays – Case Study

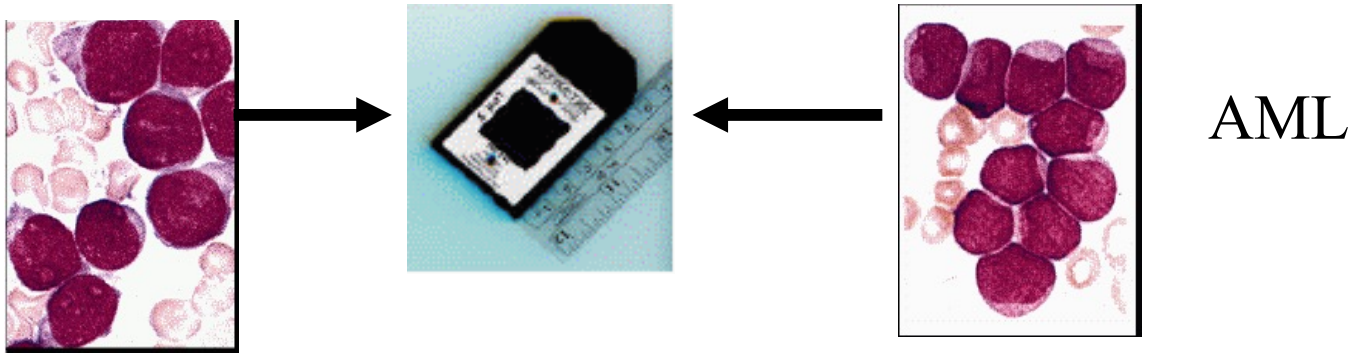
DNA microarray data is used by the scientists to measure the expression levels of large numbers of genes simultaneously or to find the genetic constitution of multiple regions of a genome.

Given DNA microarray data for a number of samples (patients), can we

- Accurately diagnose the disease?
- Predict outcome for given treatment?
- Recommend best treatment?

Example: ALL/AML data

- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model

ALL  AML

Results on test data:
33/34 correct, 1 error may be mislabeled

Data Mining and Privacy

- Privacy considerations important if personal data is involved
  - The **Facebook**–Cambridge Analytica **data** breach was a **data** leak in early 2018
  - The data sharing agreement between DeepMind health and the Royal Free NHS Foundation Trust (2016)
  - Google's "Project Nightingale" (2019)
    - Ascension
    - HIPPA rule

Data Mining and Privacy

- In 2006, NSA (National Security Agency) was reported to be mining years of call info, to identify terrorism networks
- Social network analysis has a potential to find networks
- Invasion of privacy – do you mind if your call information is in a gov database?
- What if NSA program finds one real suspect for 1,000 false leads ? 1,000,000 false leads?

Class Discussion

- What data are you interested in mining?
- What applications are you interested in developing data mining tasks for?

Outline

- Introduction: Data Flood
- Data Mining Application Examples
- **Data Mining & Knowledge Discovery**
- Data Mining Tasks

Knowledge Discovery

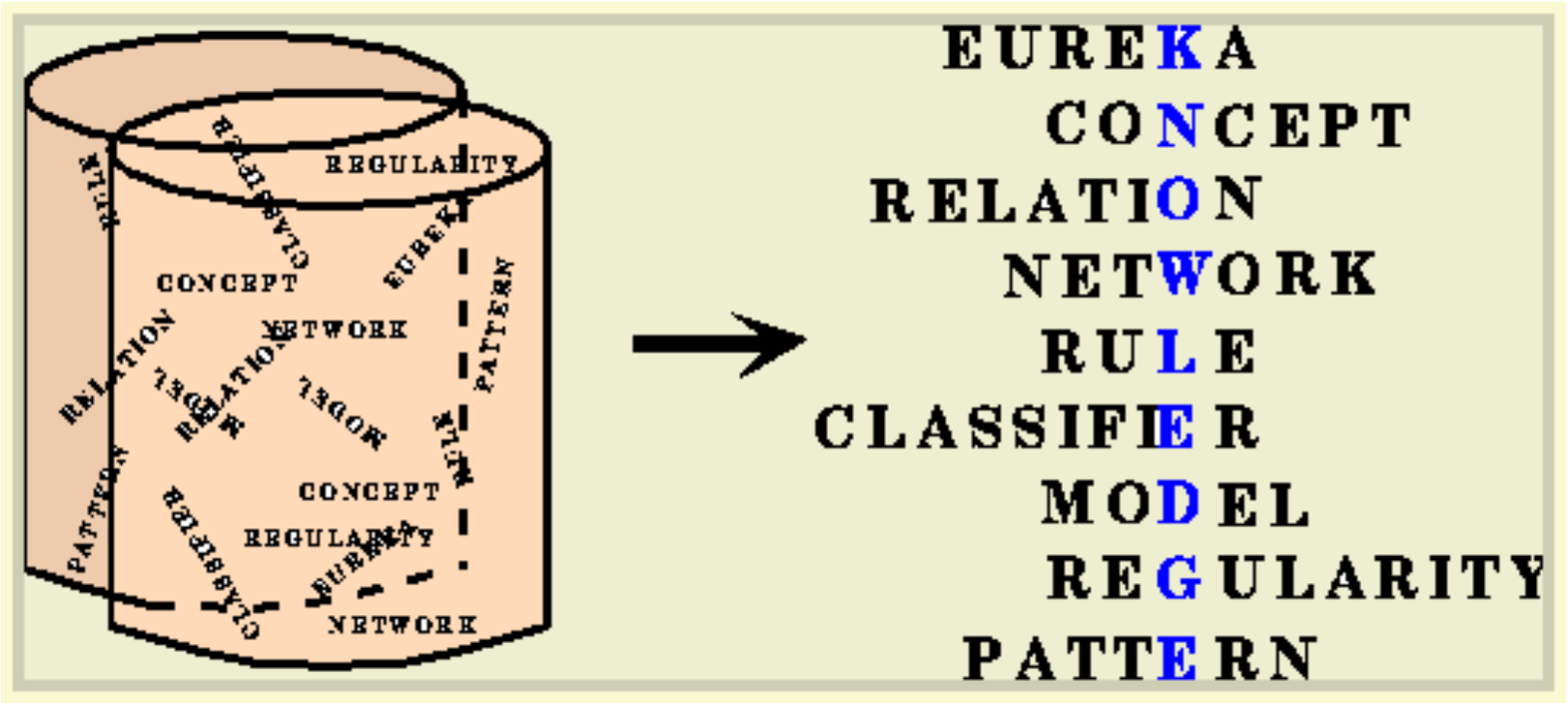
Knowledge Discovery in Data is the

non-trivial process of identifying

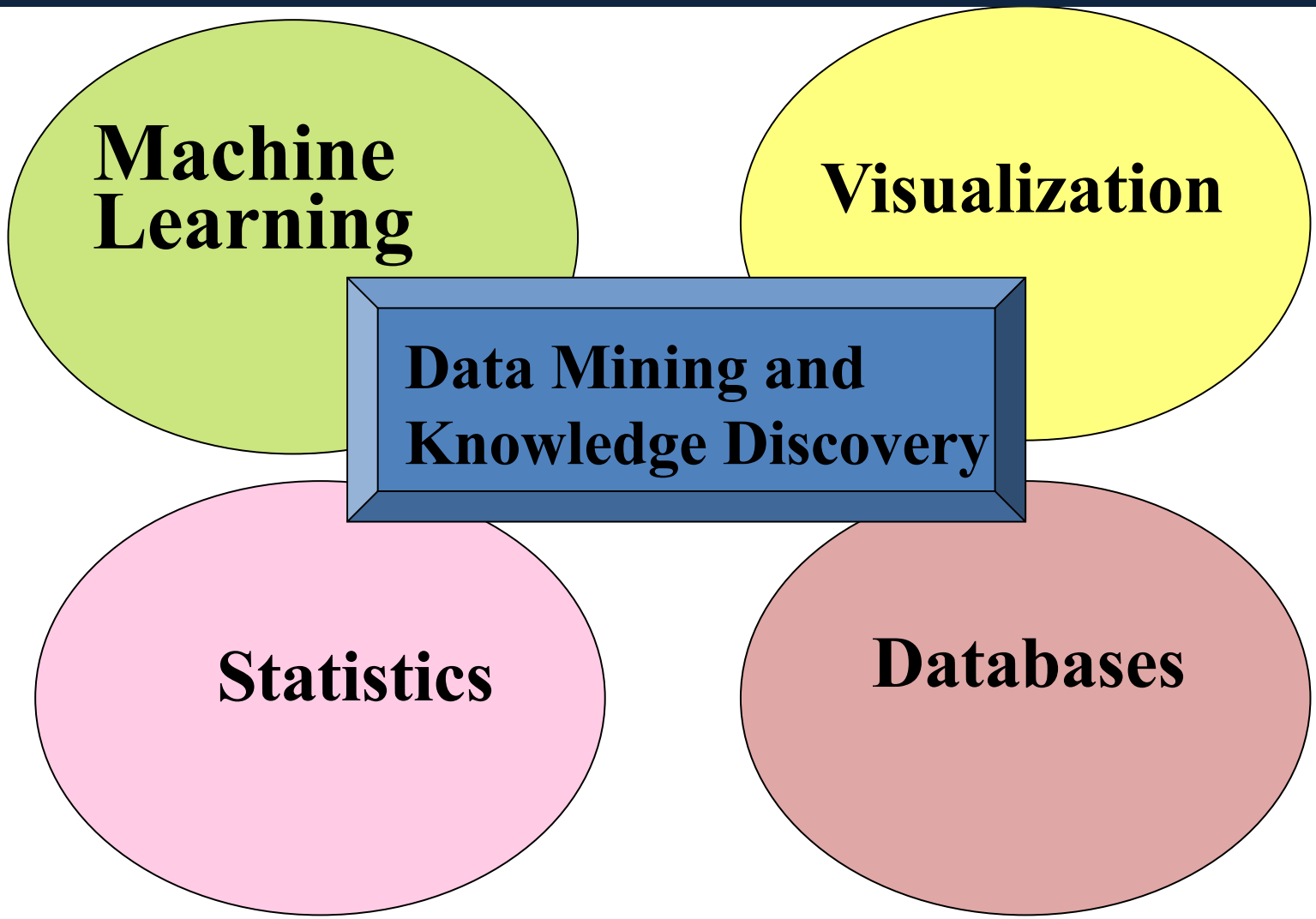
- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

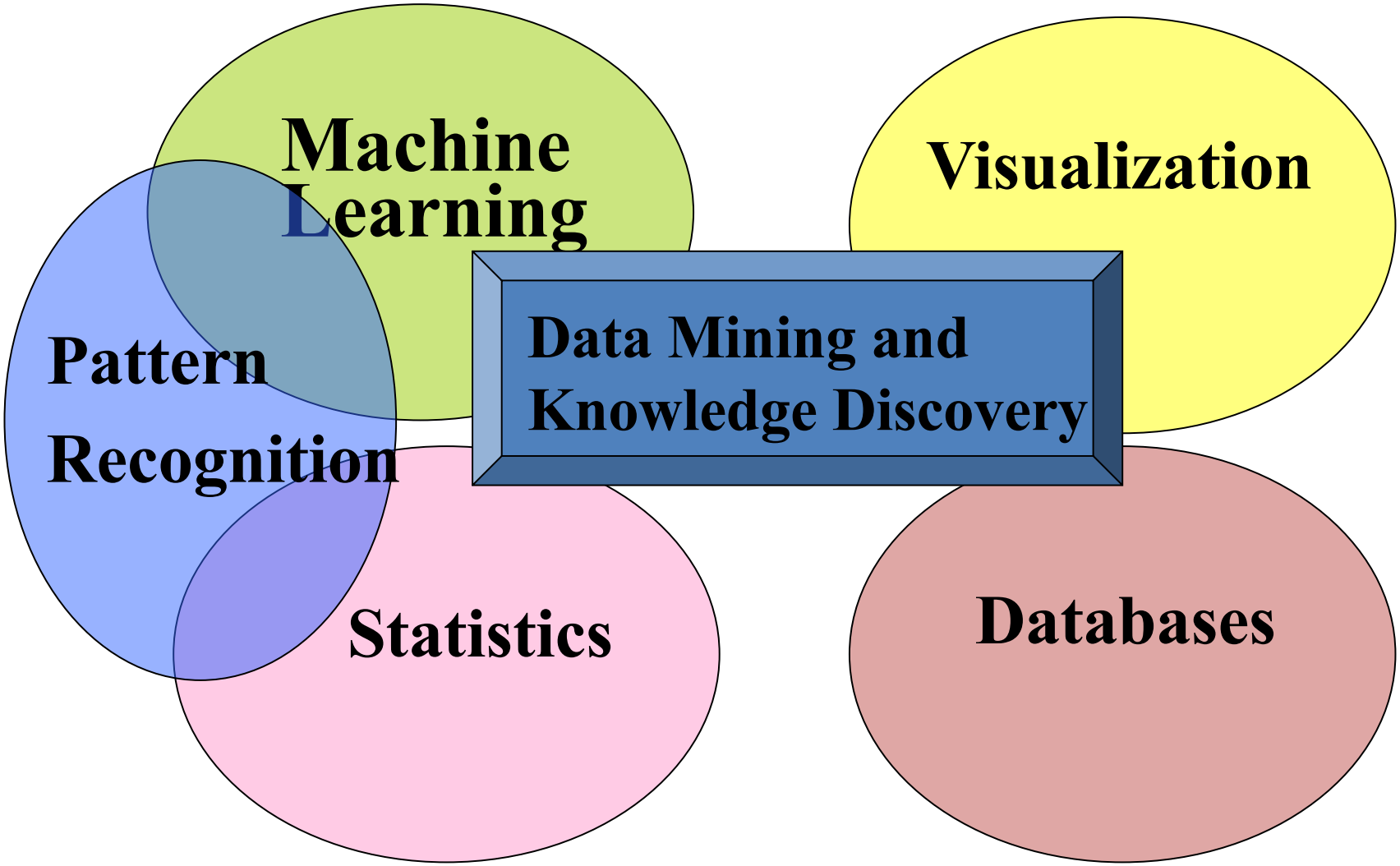
My early interpretation of Data Mining



Related Disciplines



Related Disciplines

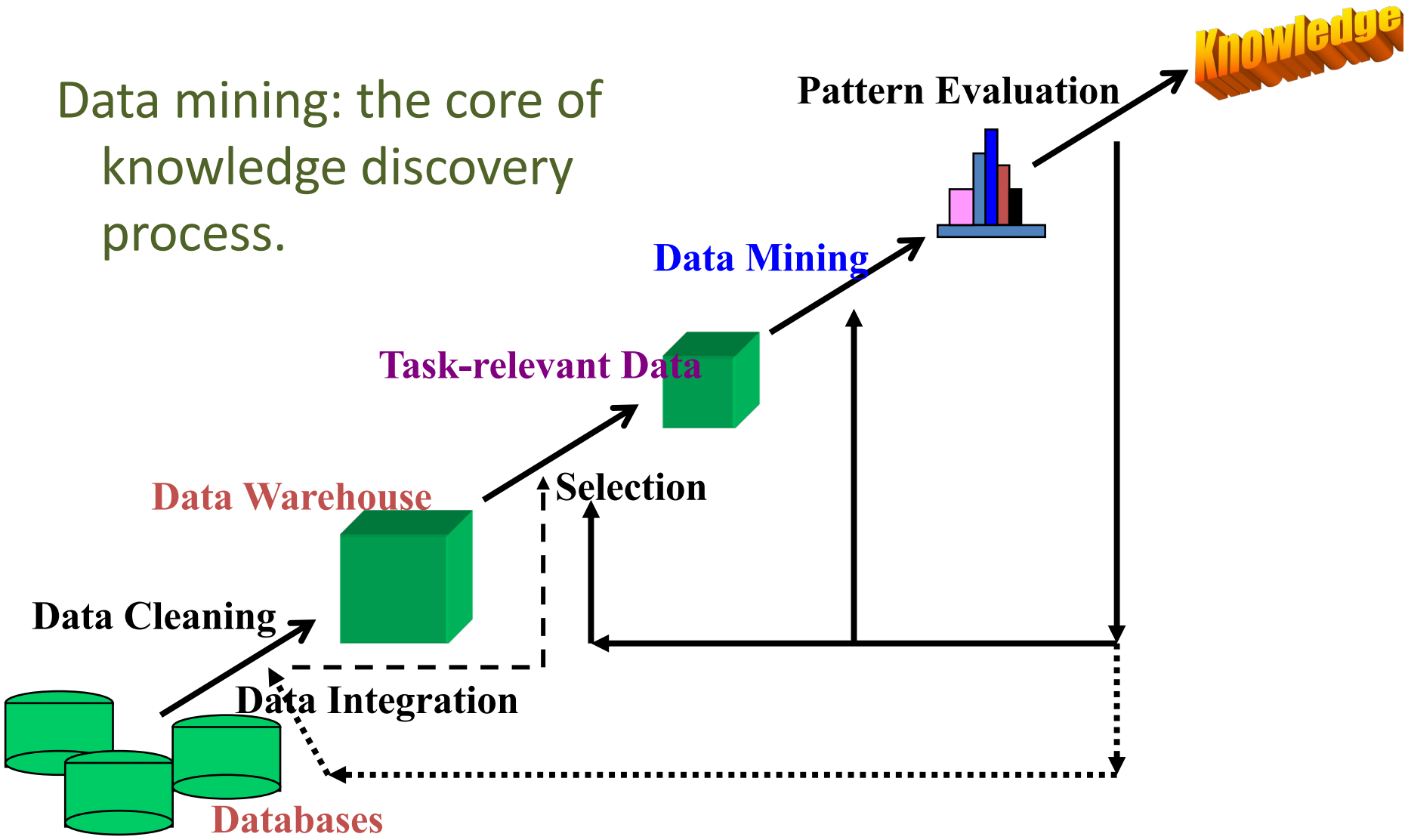


Statistics, Machine Learning, Data Mining

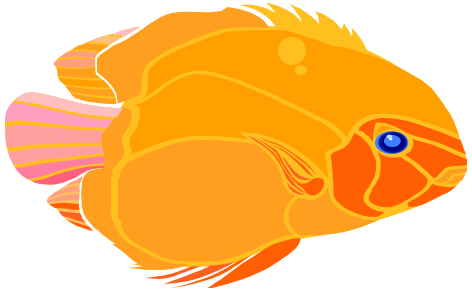
- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Knowledge discovery process flow

Data mining: the core of knowledge discovery process.



Other Names



- Data Fishing, Data Dredging: 1960-
  - used by Statistician (as bad name)
- Data Mining :1990 --
  - used DB, business
  - in 2003 – bad image because of Total Information Awareness (TIA), DARPA
- Knowledge Discovery in Databases (1989-)
  - used by AI, Machine Learning Community
- Also Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...
- Now - Data Science, Big Data

Outline

- Introduction: Data Flood
- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- **Data Mining Techniques**

Data Mining Techniques

- **Classification:** predicting the class for an item
- **Regression**: predicting a continuous value
- **Clustering:** finding clusters in data
- **Association:** e.g. A & B & C occur frequently
- **Recommendation**: find personalized recommendations
- **Summarization:** describing a group
- **Deviation Detection**: finding changes
- **Link Analysis**: finding relationships
- **Visualization:** to facilitate human discovery

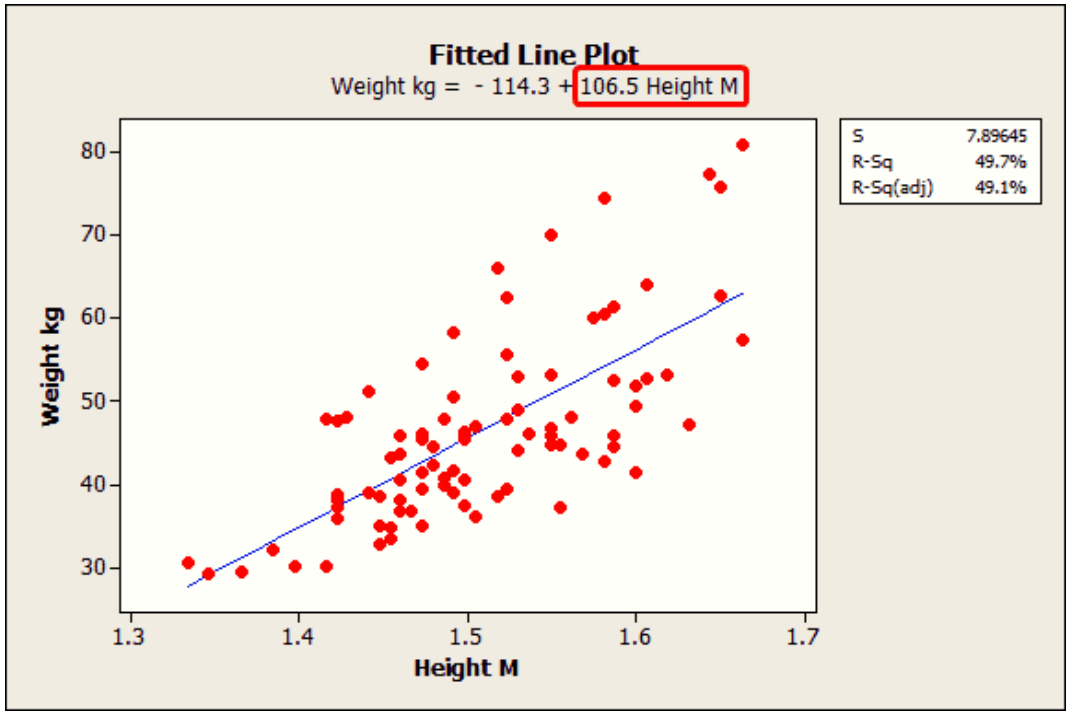
Data Mining Techniques (1)

- Regression and Classification
  - Regression: Predict some unknown or missing numerical values
    - E.g., predict stock price, predict the amount of sale, ...
  - Finding models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on climate, classify cars based on gas mileage and other data, ...
  - Approaches: regression models, decision-tree, SVM, Artificial Neural Networks, …

Regression

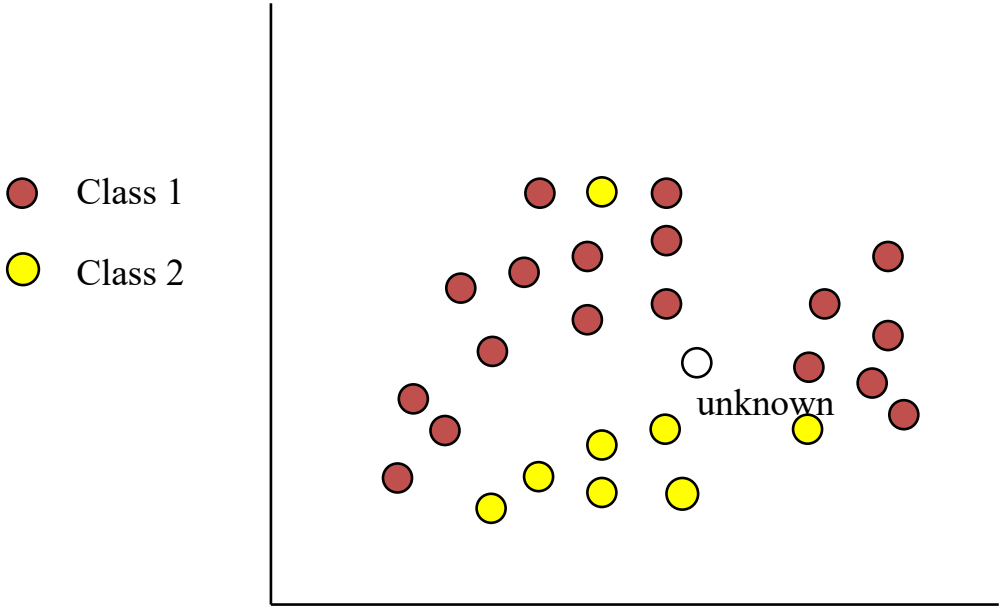
Learn a method for predicting the target value from pre-labeled (classified) instances

The value to be predicted is numeric type



Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

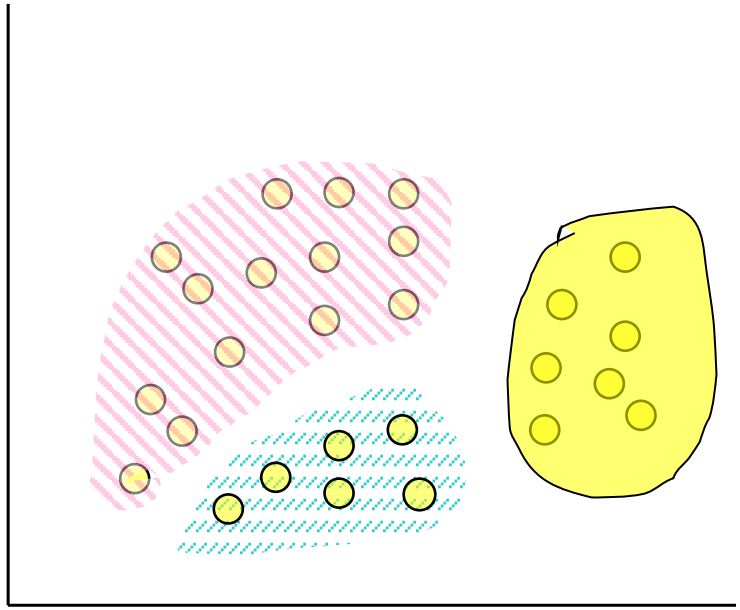


Many approaches:
Statistics,
Logistic regression,
Decision Trees,
SVM, Neural Networks,
...

Data Mining Techniques (2): Clustering

- 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

Find "natural" grouping of instances given un-labeled data



Data Mining Techniques (3)

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Multi-dimensional vs. single-dimensional association
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(T, “computer”) → contains(x, “software”) [1%, 75%]

Data Mining Techniques (4)

- 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

Are All the "Discovered" Patterns Interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
  - Suggested approach: Human-centered, query-based, focused mining
- Interestingness measures: A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

Are All the "Discovered" Patterns Interesting?

- Objective vs. subjective interestingness measures:
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Summary

- Technology trends lead to data flood
  - data mining is needed to make sense of data
- Data Mining has many applications
- Knowledge Discovery Process
- Data Mining Techniques
  - classification, regression, clustering, association rule analysis, outlier detection, recommendation, …