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
  - Etc.

Big Data

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

Zeta bytes = 10^{21} bytes

Exabytes =Quintillion = 10^{18} bytes

Petabytes = 10^{15} bytes

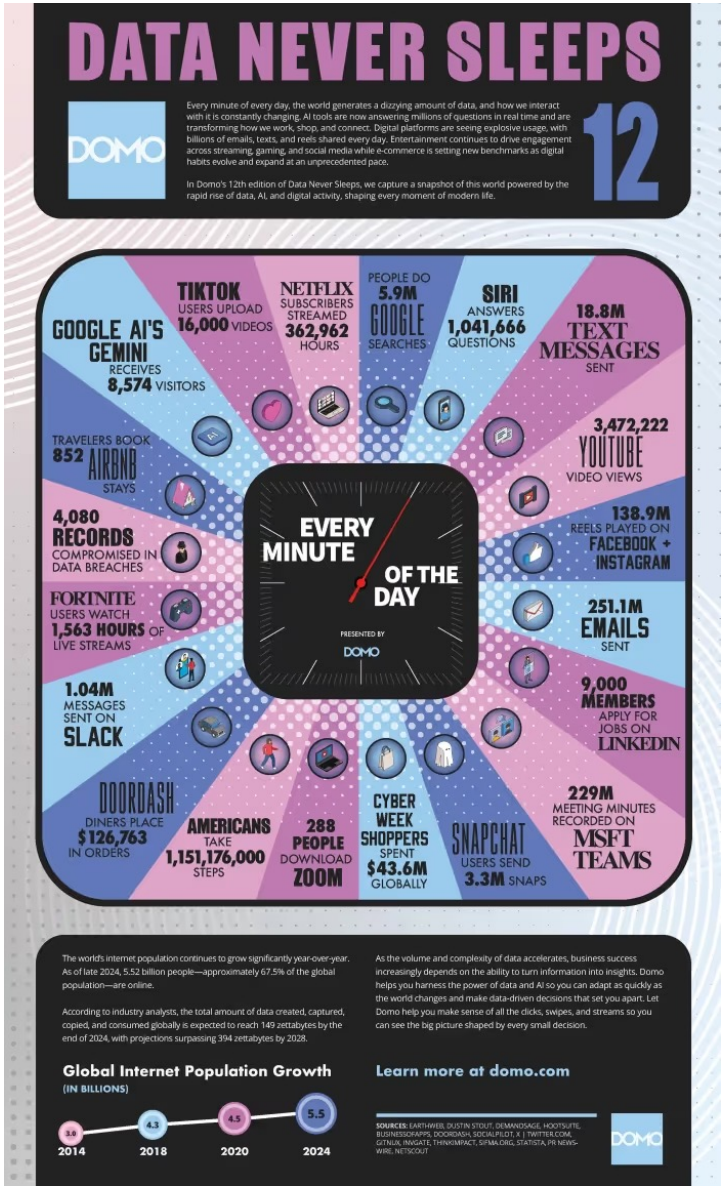
Tera = 10^{12} bytes

Gita bytes = 10^{9} bytes

Mega bytes = 10^{6} bytes

Big Data

2024 Data

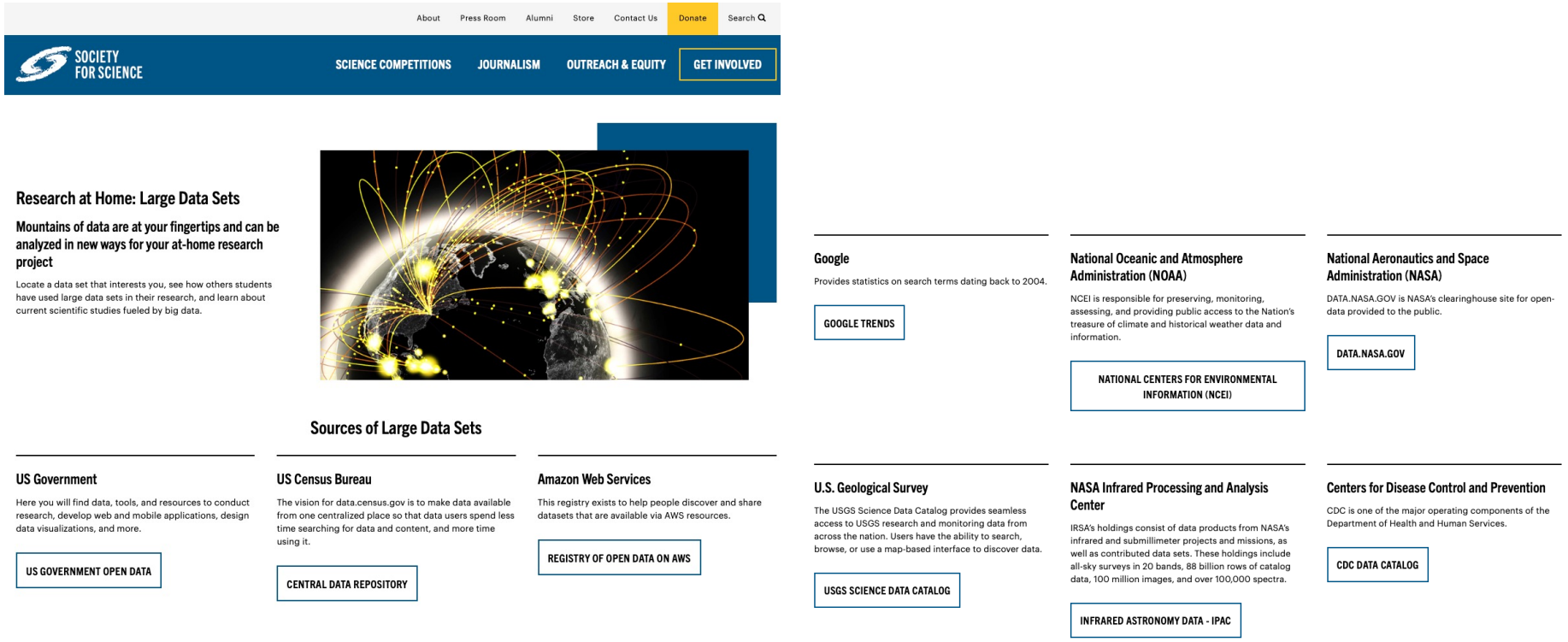


Market Growth

- The global **big data analytics market** was valued at **USD 307.52 billion in 2023** and is projected to grow from **USD 348.21 billion in 2024** to **USD 961.89 billion by 2032,** showing strong growth for 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) (https://www.sdss5.org)
- 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/)
- Earth / Climate / Satellite Big Data
  - Data discovery portal: https://www.earthdata.nasa.gov/data
  - Earthdata Search tool: https://search.earthdata.nasa.gov/
- NOAA Climate data
  - https://www.ncei.noaa.gov/access/search/index
  - https://www.ncei.noaa.gov/cdo-web/

Large Databases

- Finance / Business / Markets
  - **SEC EDGAR filings (text mining +financial analytics)** https://www.sec.gov/search-filings/edgar-application-programming-interfaces
  - **World Bank Open Data** https://data.worldbank.org/
  - **IMF Data** https://www.imf.org/en/data
- Transportation / Mobility Datasets
  - **NYC Taxi Trip Record Data** https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
  - **OpenStreetMap (global geospatial dataset)** https://planet.openstreetmap.org/

Data Sets For Data Mining Research

- OpenML
  - https://www.openml.org/
  - Papers with code: https://paperswithcode.com/datasets
- UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/index.php
- Kaggle Data Sets: https://www.kaggle.com/datasets
- Recommender Systems / Personalization
  - MovieLens: https://grouplens.org/datasets/movielens/
  - Amazon Reviews / Product Data https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Data Sets For Data Mining Research

- Computer Vision / Multimedia Mining
  - COCO dataset https://cocodataset.org/
  - ImageNet https://www.image-net.org/
- Graph Mining / Network Science
  - SNAP (Stanford Network Analysis Project) https://snap.stanford.edu/data/
- Cybersecurity / Anomaly Detection
  - CIC datasets (Canadian Institute for Cybersecurity) https://www.unb.ca/cic/datasets/
- Google BigQuery Public Datasets
  - https://cloud.google.com/bigquery/public-data

Outline

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

Data Mining Applications

- Science and Engineering
  - Data from sensor networks and data from moving objects
    - Wireless sensor networks
      - fault detection, environmental monitoring, predictive maintenance, energy-efficient system optimization
    - RFID tagged moving objects
      - tracking & trajectory mining, supply chain optimization, bottleneck detection, demand forecasting, behavior pattern mining (movement/flow)

Data Mining Applications

- Science and Engineering
  - 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.)
      - *object detection/classification (tumors, land use, galaxies), segmentation, change detection over time, automated labeling, pattern discovery*
    - Spatial: data with geographic location info
      - hotspot detection, clustering regions, spatial correlations (nearby effects), route optimization, site selection/planning (stores, hospitals, facilities)
    - Time series data: use sequences of recorded values (e.g. physics, finance, medicine and music) for prediction
      - forecasting/prediction, anomaly detection (fraud, faults), trend/seasonality discovery, signal classification, early warning systems
    - Etc.

Data Mining Applications

- Business
  - CRM (Customer Relationship management) Tasks:
    - attrition prediction: *identify customers likely to leave, trigger retention campaigns, improve satisfaction/loyalty, reduce revenue loss*
    - targeted marketing:
      - cross-sell/up-sell: recommend additional products/services, increase basket size
      - customer acquisition: identify high-value prospects, look-alike modeling, predict conversion likelihood
    - credit-risk: predict default probability, optimize loan approval and pricing, assign credit scores, reduce losses
    - fraud detection: detect suspicious patterns/transactions in real time, reduce fraud losses, identify abnormal behavior, support investigation

Data Mining Applications

- Business (cont.)
  - Industries
    - Banking: credit scoring, Anti-Money Laundering/fraud detection, customer segmentation, risk management
    - e-commerce: recommendation systems, dynamic pricing, customer targeting, churn prediction
    - retail sales: market basket analysis, demand forecasting, inventory optimization, store/site analytics

Data Mining Applications

- Health Science
  - Predictive medicine
    - Predict outbreaks of health problems
      - early detection of epidemics, outbreak forecasting, identifying hotspots, supporting public health interventions (vaccination, isolation, resource allocation)
    - Precision medicine using genomic data
      - identify genetic risk factors, predict patient-specific disease susceptibility, select personalized treatments, reduce adverse drug reactions, improve outcomes
  - Management of healthcare and measuring the effectiveness of treatments
    - Compare and contrast symptoms, causes and treatment outcomes
      - find the most effective treatment plan for a condition, improve clinical decision support, optimize care pathways, identify best practices using real-world evidence
    - Measure effectiveness of outcome
      - evaluate which therapies work best for which patient groups, monitor side effects, support evidence-based medicine

Data Mining Applications

- Health Science (cont.)
  - Drug development and design.
    - Design/Determine chemical compounds for effective treatment
      - predict drug-target interactions, identify promising molecules, virtual screening, optimize drug properties (toxicity, efficacy), prioritize compounds for clinical trials
  - Detection of health insurance fraud and abuse
    - The Texas Medicaid Fraud and Abuse Detection System
      - detect abnormal billing patterns, identify suspicious claims/providers, reduce waste/abuse, support investigation and enforcement
      - **Example: Texas Medicaid Fraud and Abuse Detection System**
        → use large-scale claims + provider billing data to automatically flag suspicious patterns for investigation

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

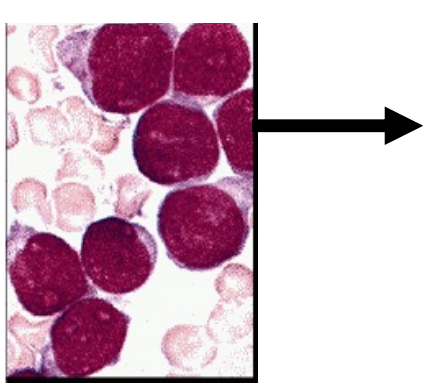
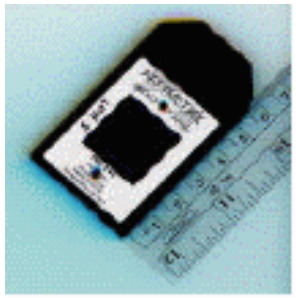
DNA microarray data is used by the scientists to measure the expression levels of large numbers of genes simultaneously or to find the genetic constitution of multiple regions of a genome.

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(1996)

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

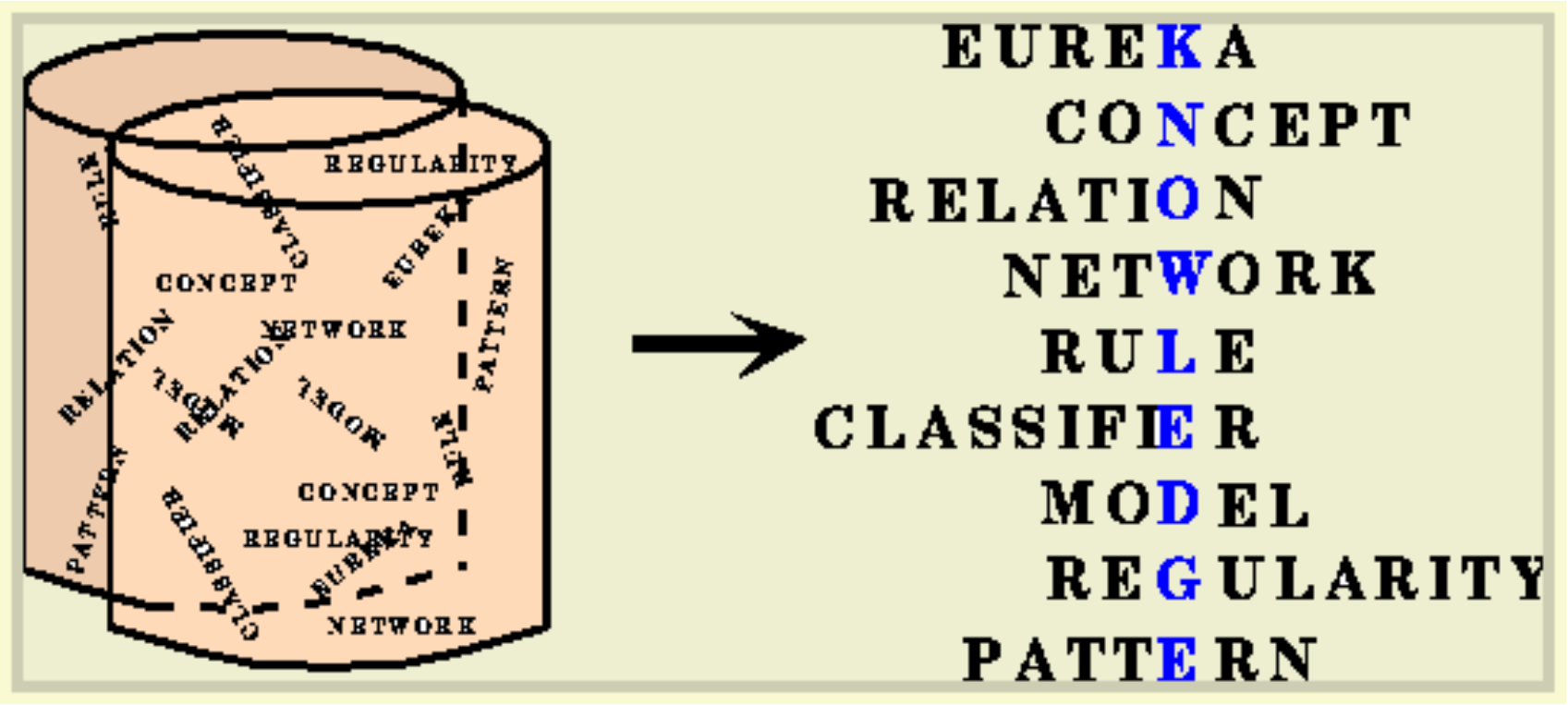
Knowledge Discovery in Data is the

non-trivial process of identifying

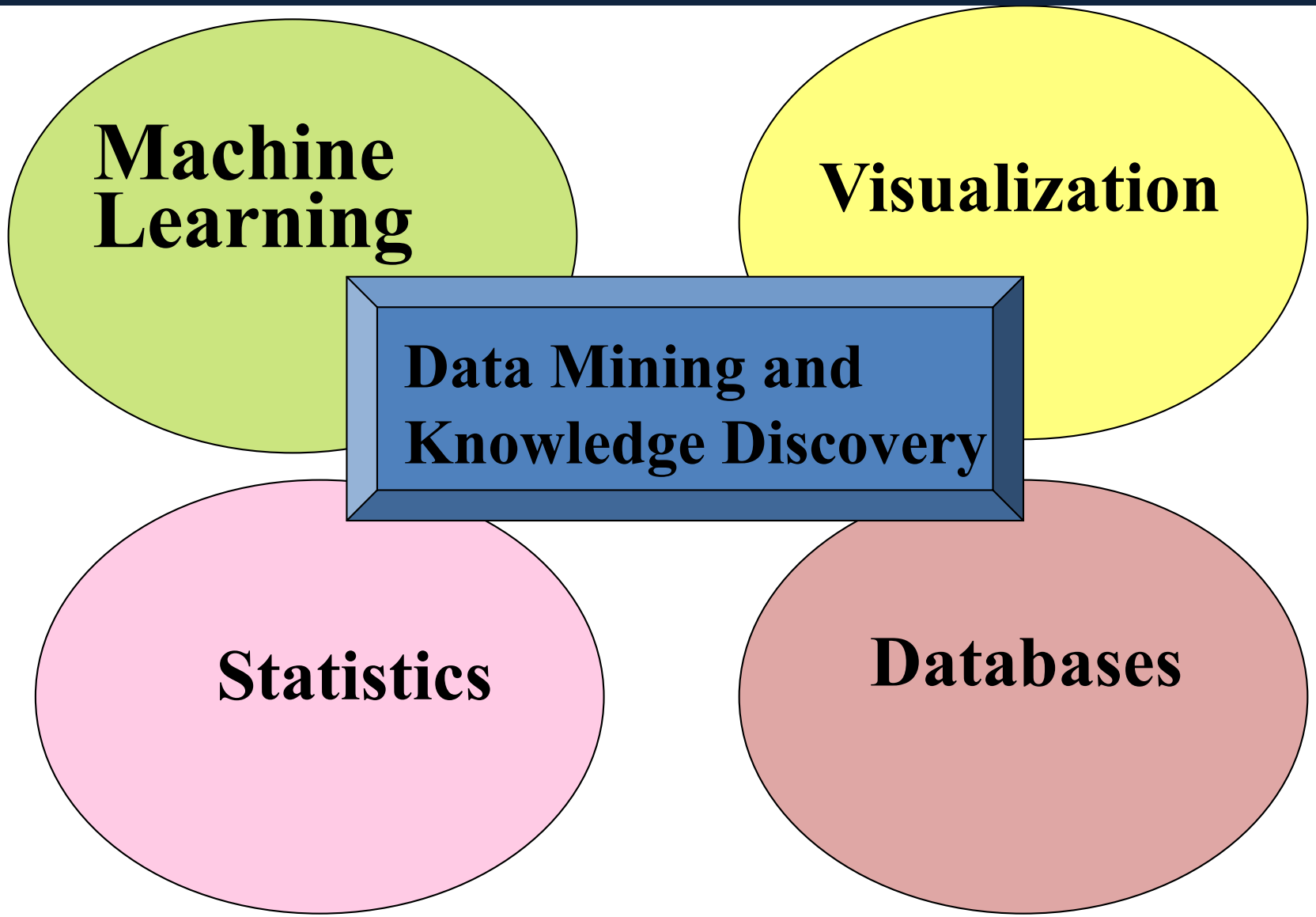
- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

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

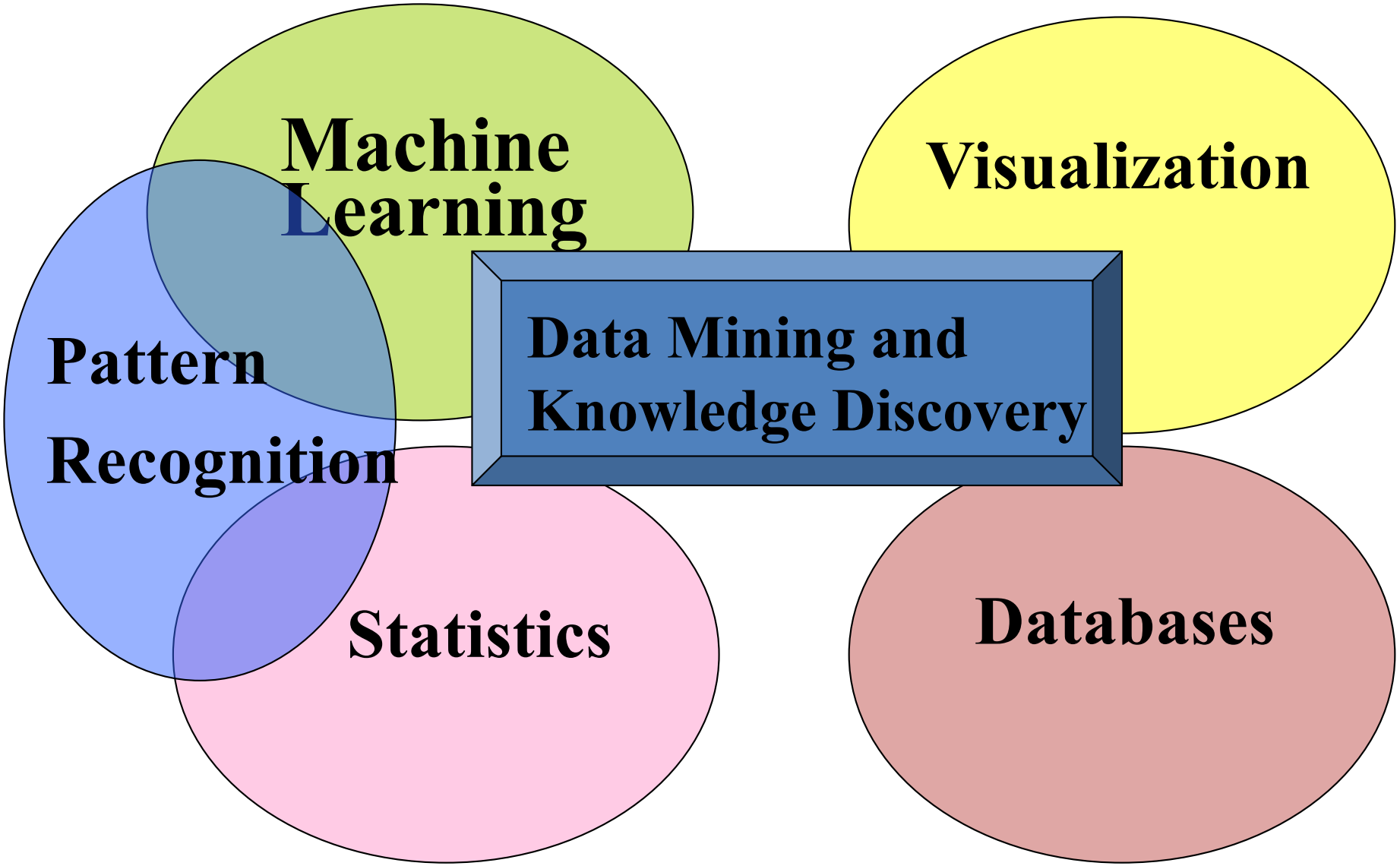
My early interpretation of Data Mining



Related Disciplines



Related Disciplines

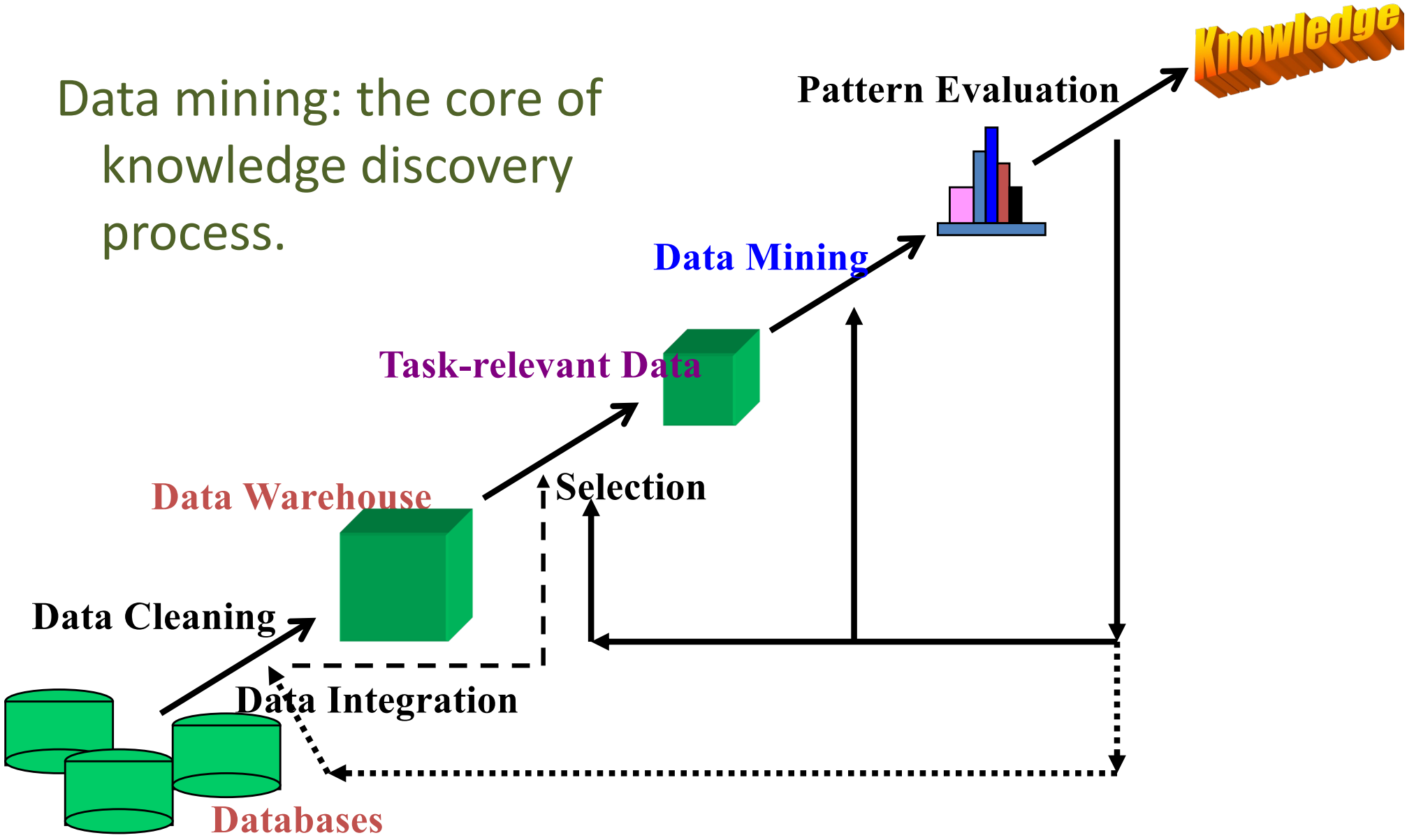


Statistics, Machine Learning, Data Mining

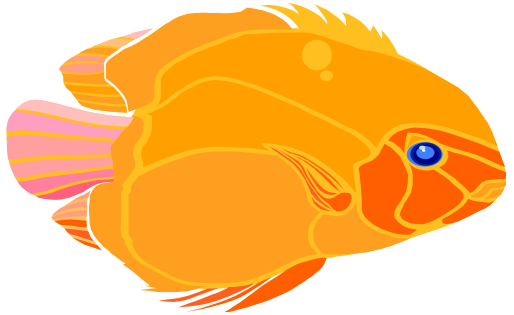
- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates 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

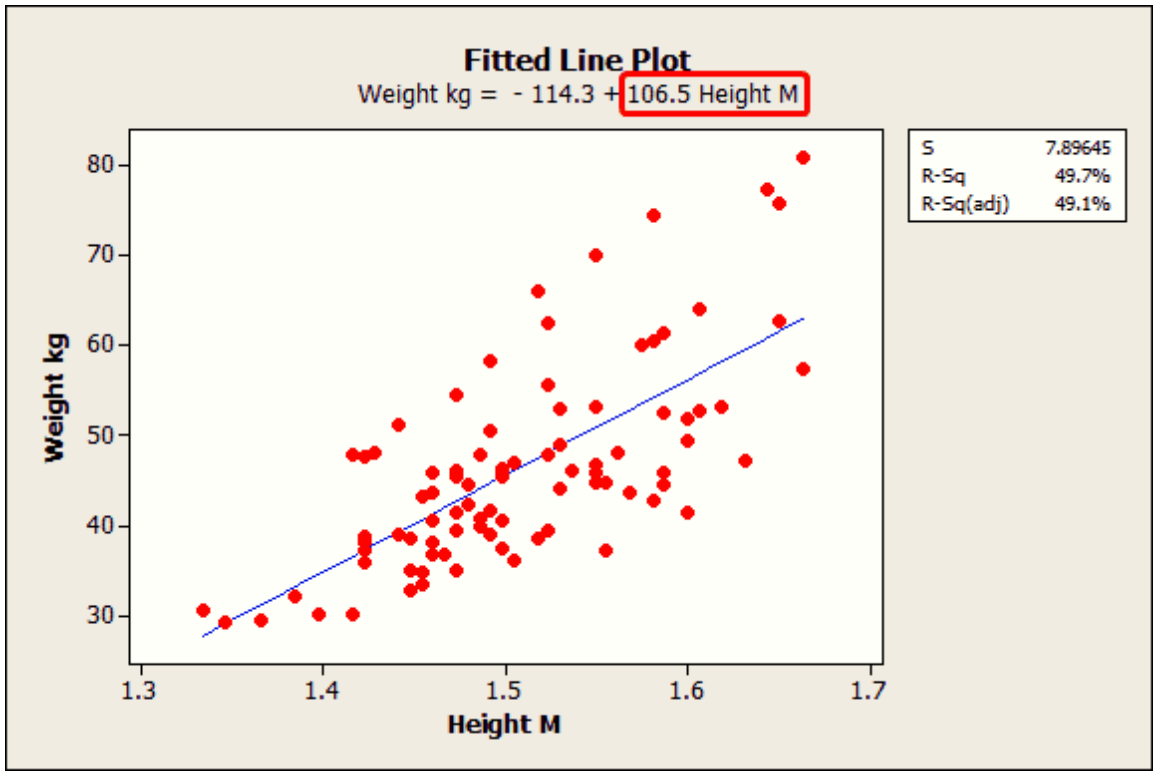
Data Mining Techniques (1)

- 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, Neural Networks, …

Regression

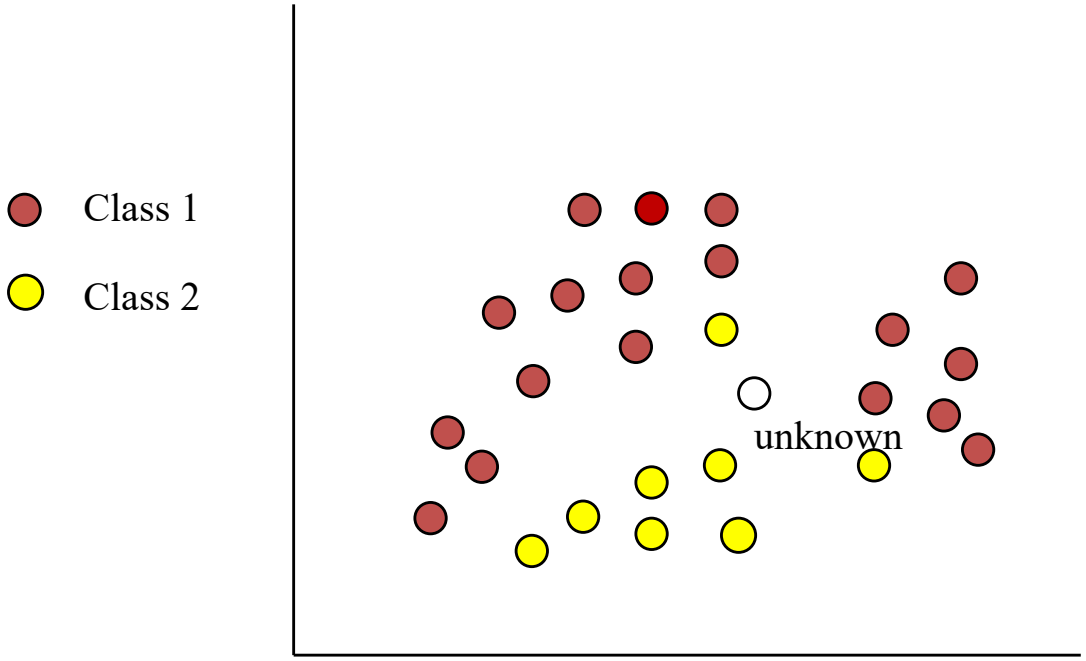
Learn a method for predicting the target value from pre-labeled (classified) instances

The value to be predicted is numeric type



Classification

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

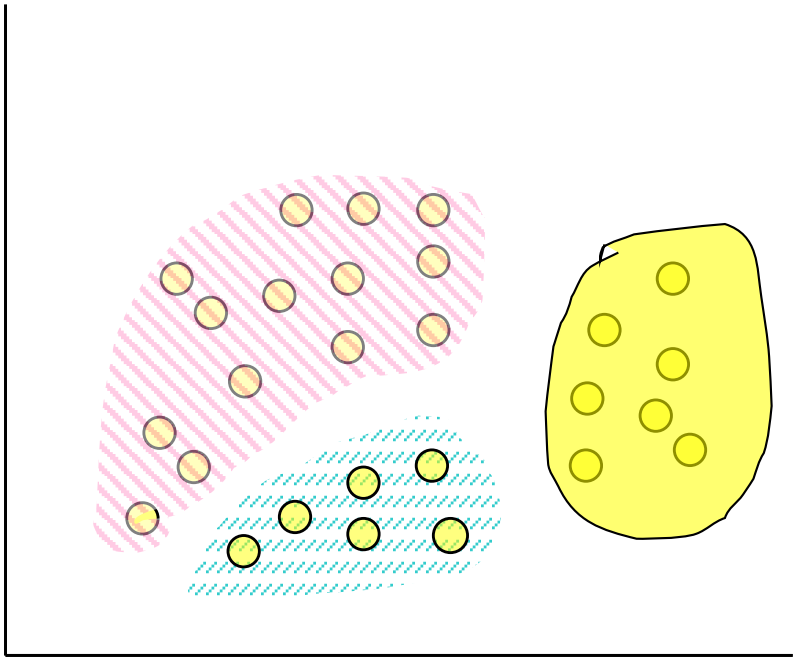


Many approaches:
Statistics,
Logistic regression,
Decision Trees,
SVM, Neural Networks,
...

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
  - **Characterization**: summarize typical properties of a dataset/group
    Example: "Typical climate patterns in **dry** regions"
  - **Discrimination**: compare/contrast two groups
    Example: "How do **dry** vs **wet** regions
- Association Analysis
  - Can reveal **correlations** (not necessarily causation!)
  - Types:
    - **Single-dimensional**: one attribute type (e.g., item A → item B)
    - **Multi-dimensional**: multiple attributes (e.g., age + income → purchase)
    - Example: **Age 20–29 & income 20–29K ⇒ buys PC**
      support = 2%, confidence = 60%

Data Mining Techniques (4)

- Outlier analysis
  - **Outlier**: an object/record that deviates significantly from typical data behavior
  - May represent:
    - **noise / measurement error**, or
    - a **meaningful rare event**
  - Applications: **fraud detection**, **intrusion detection**, **rare-event discovery**
  - **Example:** credit card spending

    typical: $5–$200 transactions

    outlier: **$4,200 overseas purchase at 3am** → potential fraud

Data Mining Techniques (5)

- Trend and evolution analysis
  - Identify **long-term trends** and **short-term deviations**
    - e.g., **regression / forecasting**
      - **Example:** daily sales over 2 years shows an upward trend; sudden 40% drop = anomaly
  - Discover **temporal patterns**
    - **sequential pattern mining**
    - **periodicity / seasonality analysis**
      - **Example:** "Users who buy a phone → buy a case within 7 days"
      - **Example:** web traffic peaks every Monday (weekly seasonality)
  - **Similarity-based analysis**
    - find entities with similar temporal behavior (users/products/sensors)
      - **Example:** cluster smart meters with similar evening energy-usage patterns

Data Mining Techniques (6)

- **Link Analysis :** analyze relationships (links) between objects to discover structure and influence
  - objects = nodes (people/pages/accounts)
  - links = edges (friendship/calls/citations/transactions)
- **Key tasks**
  - **Ranking / influence:** identify important nodes (e.g., PageRank)
  - **Community detection:** find groups or clusters in the network
  - **Anomaly link analysis:** detect suspicious connections/patterns
- **Applications**
  - social networks (friend recommendation), web search and page ranking
  - fraud rings / money laundering detection

Example: fraud detection

- A set of "unrelated" accounts all transfer money to the same small group of accounts
  → indicates a possible **fraud ring / collusion network**

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.

Evaluating Discovered Patterns (1)

- **Market Basket Rule: "Beer & Diapers" (retail association mining)**
- **Pattern (association rule):** customers buying diapers also buy beer (esp. evenings/weekends)
- **Human understandable:** yes ("new dads buying beer while buying diapers")
- **Valid on new data:** the story persists in many datasets in similar form (co-purchases)
- **Useful/actionable:** store layout, bundling, targeted promotions
- **Novel/unexpected:** yes — surprising to most people
- **Objective interestingness:** support, confidence, lift
- **Subjective interestingness:** unexpectedness + actionability

Evaluating Discovered Patterns (2)

- **Credit Card Fraud Detection: "impossible travel" rule**
- **Pattern:** same card used in **two far-away places within a short time** (e.g., NYC 2pm → Tokyo 4pm)
- **Human understandable:** extremely
- **Valid on new data:** very reliable; used broadly
- **Useful:** blocks fraud quickly
- **Novel:** not "novel" anymore, but initially very powerful
- **Objective:** rule confidence / anomaly score
- **Subjective:** high actionability; user agrees it's suspicious

Evaluating Discovered Patterns (3)

- **Google Flu Trends idea (search logs → flu outbreaks)**
- **Pattern:** certain search queries predict flu-like illness activity
- **Human understandable:** yes (people search symptoms)
- **Valid on new data:** partially (worked early but later drifted)
- **Useful:** public health monitoring
- **Novel:** very novel at the time
- **Objective:** predictive accuracy on held-out periods
- **Subjective:** novelty + usefulness

Evaluating Discovered Patterns (4)

- **Network discovery: fraud rings / collusion groups**
- **Pattern:** accounts forming dense clusters with suspicious transaction edges
  (e.g., many accounts transferring into a small set of "collector" accounts)
- **Human understandable:** yes (organized ring)
- **Valid on new data:** yes (rings recur with similar structure)
- **Useful:** huge in anti-money laundering, insurance fraud
- **Novel:** often yes, because criminals hide
- **Objective:** graph structure measures (density, motifs, centrality)
- **Subjective:** actionability is key (who to investigate)

Evaluating Discovered Patterns (5)

- **Healthcare: early warning for sepsis deterioration**
- **Pattern:** certain time-series patterns in vitals/labs predict sepsis
- **Human understandable:** yes (clinicians interpret vitals)
- **Valid on test data:** must be validated (ROC, calibration)
- **Useful:** prevents death, triage decisions
- **Novel:** sometimes (model finds non-obvious variable combos)
- **Objective:** predictive metrics (AUROC), significance
- **Subjective:** usefulness + hypothesis validation ("these vitals matter")

Evaluating Discovered Patterns (6)

- **Manufacturing: predictive maintenance from vibration/temperature sensors**
- **Pattern:** vibration signature + temperature rise predicts failure within 2 days
- **Human understandable:** yes
- **Valid:** yes if cross-validated on future time windows
- **Useful:** prevents downtime, saves money
- **Novel:** moderately
- **Objective:** anomaly score / time-series features
- **Subjective:** actionability (“schedule replacement now”)

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, …
- Evaluate discovered patterns using objective and subjective measures