Introduction to Data Mining

What Is Data Mining?

- We live in a world where vast amounts of data are generated constantly and rapidly
- **Data mining** is the process of discovering interesting patterns, models and other kinds of knowledge in large data sets
  - "Data mining": a misnomer? It should be "knowledge mining from data"
  - Other terms: *Knowledge mining from data*, *KDD* (*Knowledge Discovery from Data*), *pattern discovery*, *knowledge extraction*, *data analytics*, *information harvesting*
- Data mining is a young, dynamic, and promising field
- Example: Data mining turns a large collection of data into knowledge
  - Google's *Flu Trends* found a close relationship between the number of people who search for flu-related info. and the number of people who have flu symptoms
    - It can estimate flu activity up to two weeks faster than traditional systems

Predictive Power of Google Trends Data

Abstract

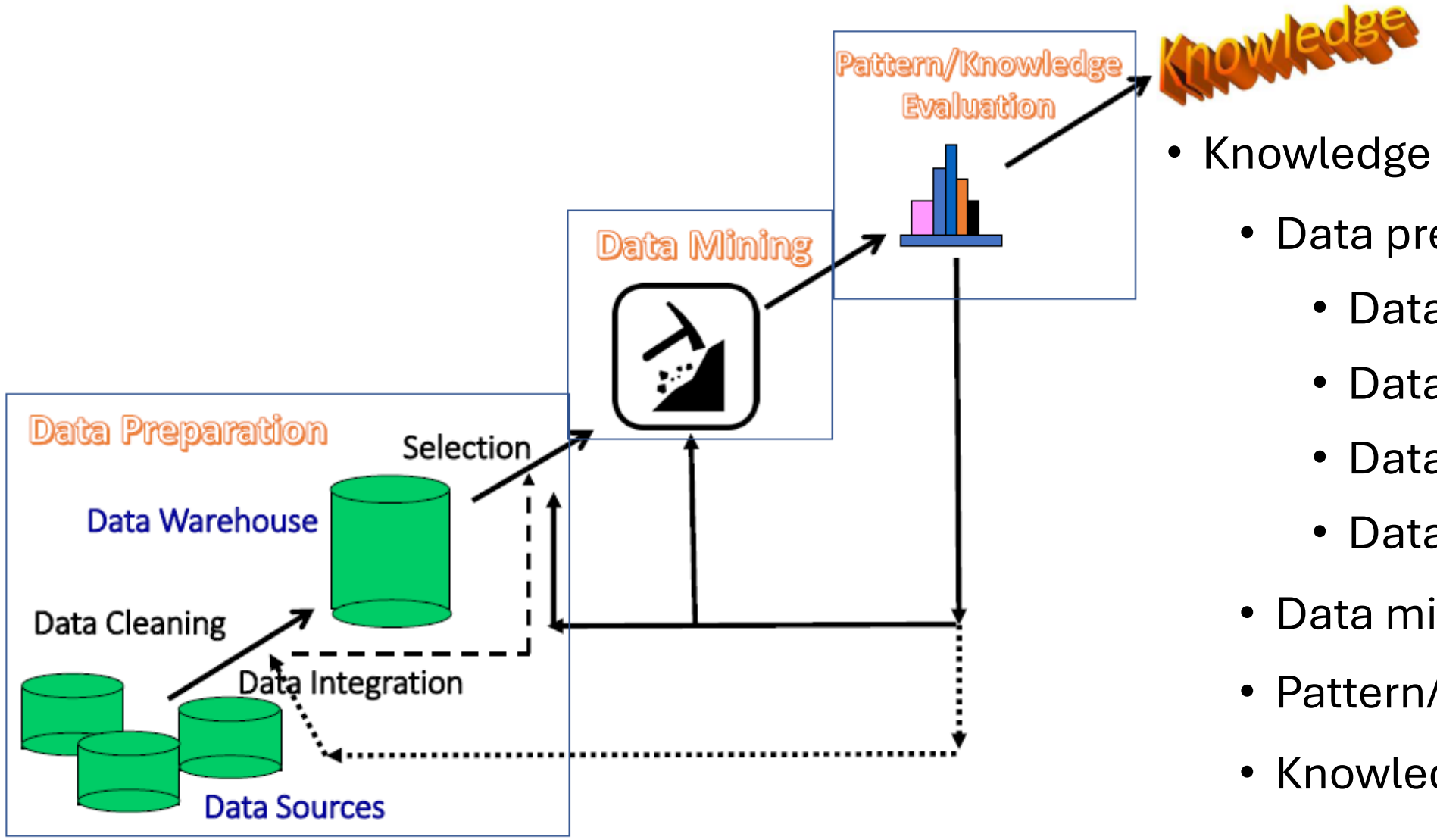
Go to:

Background

The COVID-19 outbreak has revealed a high demand for timely surveillance of pandemic developments. Google Trends (GT), which provides freely available search volume data, has been proven to be a reliable forecast and nowcast measure for public health issues. Previous studies have tended to use relative search volumes from GT directly to analyze associations and predict the progression of pandemic. However, GT's normalization of the search volumes data and data retrieval restrictions affect the data resolution in reflecting the actual search behaviors, thus limiting the potential for using GT data to predict disease outbreaks.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10488898/

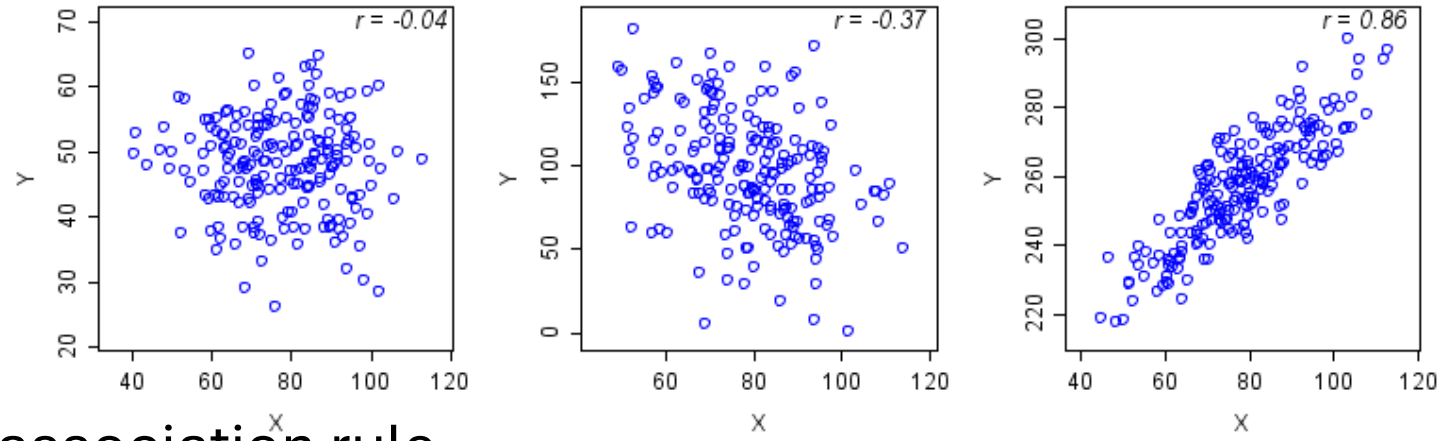
Data Mining: An Essential Step in Knowledge Discovery



- Knowledge Discovery Process
  - Data preparation
    - Data cleaning
    - Data integration
    - Data transformation
    - Data selection
  - Data mining
  - Pattern/model evaluation
  - Knowledge presentation

Pattern Discovery: Mining Frequent Patterns, Associations, and Correlations

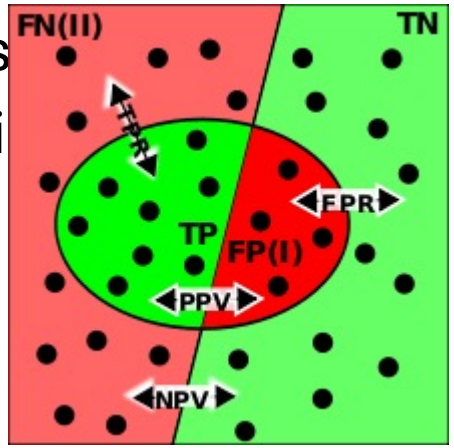
- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



- A typical association rule
  - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Classification and Regression for Predictive Analysis

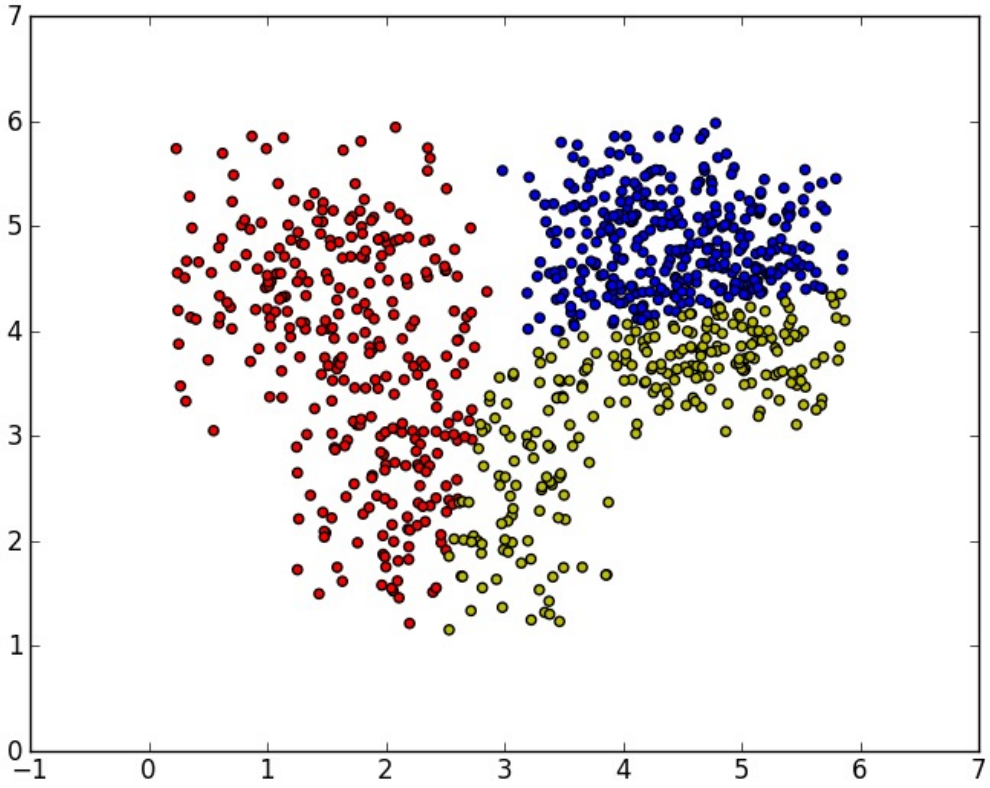
- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future predicti
    - Ex. 1. Classify countries based on (climate)
    - Ex. 2. Classify cars based on (gas mileage)
  - Predict some unknown class labels



- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

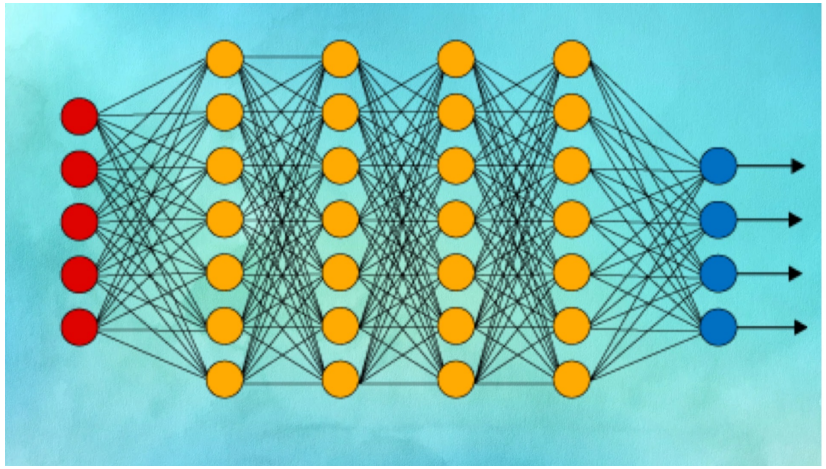
Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications



Deep Learning

- Deep learning: A fast expanding dynamic frontier in machine learning
- Deep learning has developed various *neural network architectures*
  - Feed-forward neural networks
  - Convolutional neural networks
  - Recurrent neural networks
  - Graph neural networks
  - Transformer



- Deep learning has broad applications in computer vision, natural language processing, machine translation, social network analysis, and so on
- Deep learning has been reshaping a variety of data mining tasks
  - Ex. classification, clustering, outlier detection, and reinforcement learning

Outlier Analysis or Novelty Detection

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception?—One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis