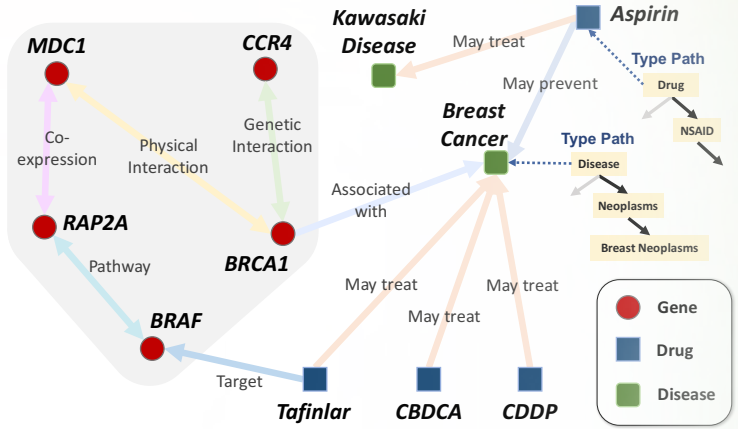
CS 412 Intro. to Data Mining

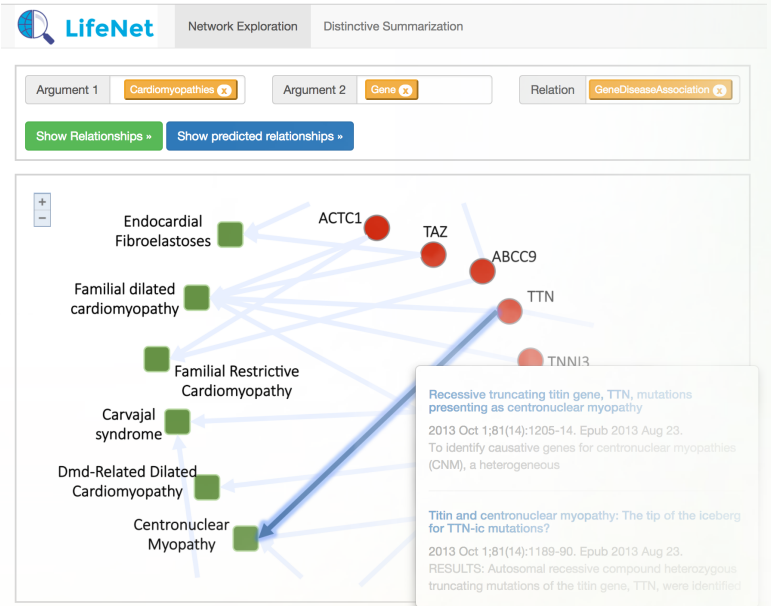
Chapter 1. Introduction

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

Help Needed: LifeNet—A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences

- What we are doing?
  - A scalable system that transforms biomedical papers into a knowledge graph & supports various search/analytics functions
- What we already have?
  - A working prototype system & an ACL demo paper
- What we are looking for?
  - Students with expertise on HTML/CSS & JavaScript
  - Experiences on web frameworks and databases
  - System design experience will be a big plus
- What you will gain?
  - Hourly pay ($12-$15 per hour, 6-20 hours per week)
  - Possible research publications & a good thesis topic





Send us your resume if interested: **Jiaming Shen** (mickeysjm@gmail.com)

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube

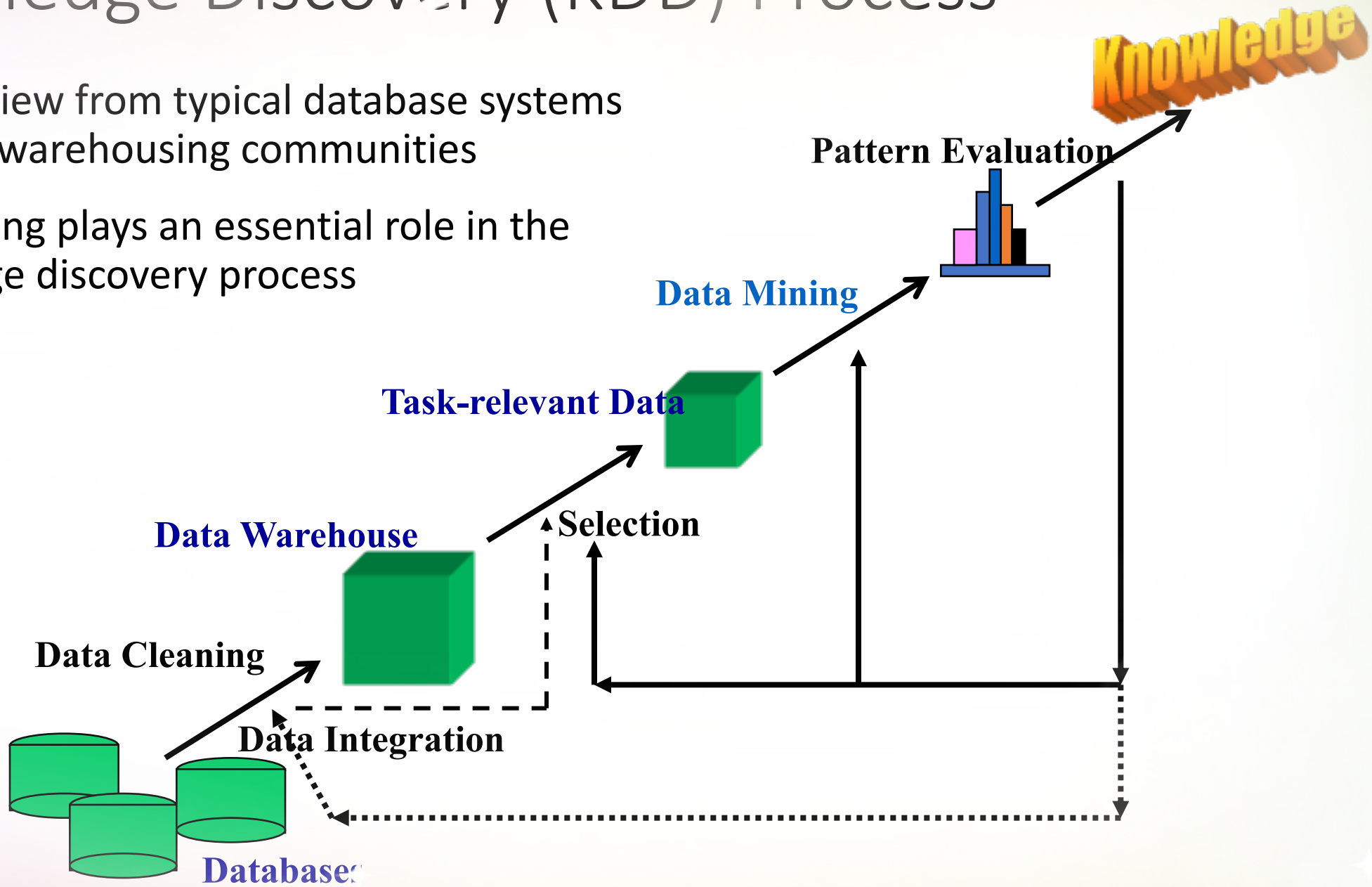
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

อะไรคือเหมืองข้อมูล
What Is Data Mining?

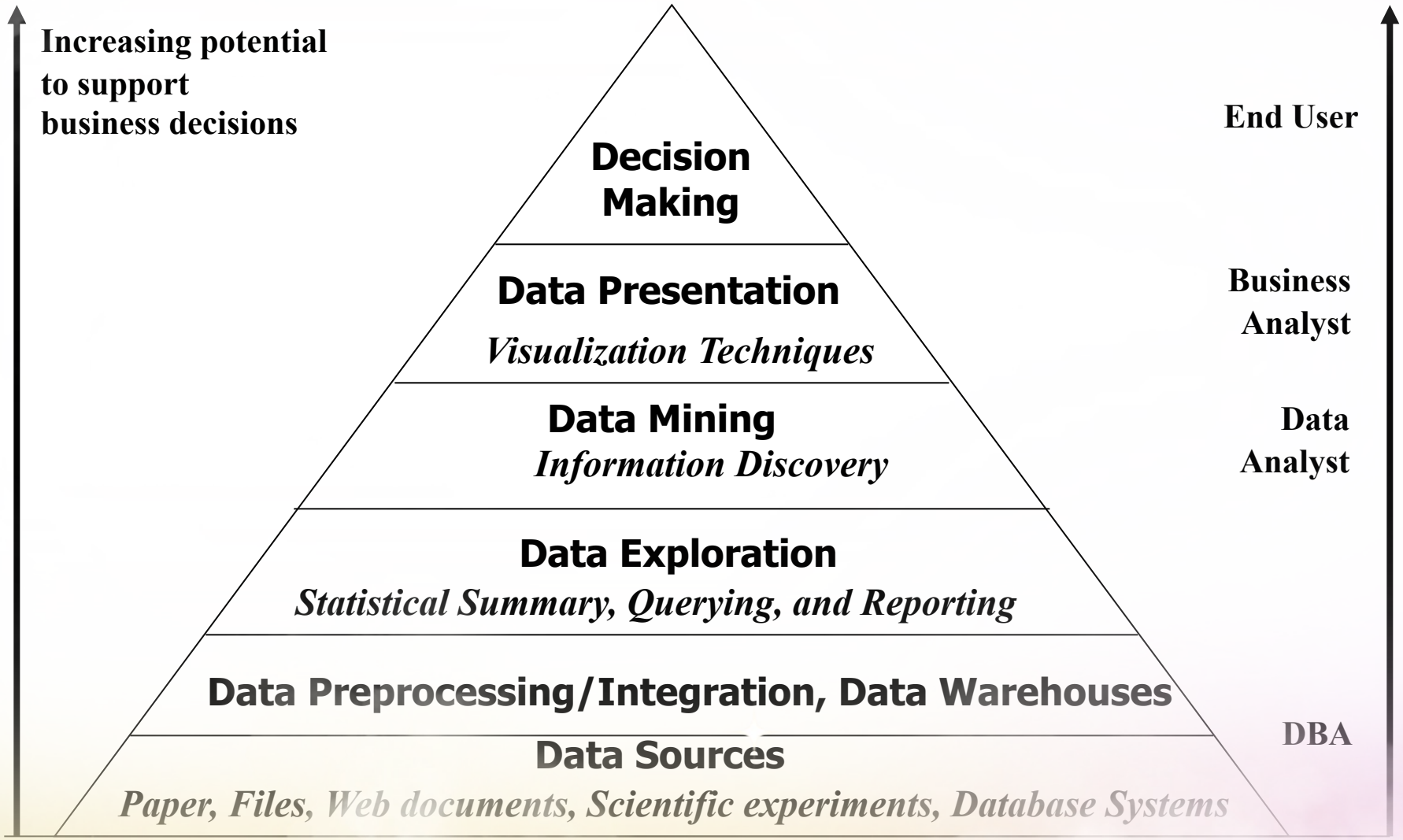
- Data mining (knowledge discovery from data) การค้นพบความรู้จากข้อมูล
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data การดึงรูปแบบข้อมูลหรือความรู้ที่น่าสนใจ
  - Data mining: a misnomer? เรียกชื่อผิด
- Alternative names: ชื่อทางเลือก
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc. การค้นพบความรู้
- Watch out: Is everything "data mining"?
  - Simple search and query processing การประมวลผลการค้นหาและแบบสอบถามอย่างง่าย
  - (Deductive) expert systems ระบบผู้เชี่ยวชาญ

Knowledge Discovery (KDD) Process

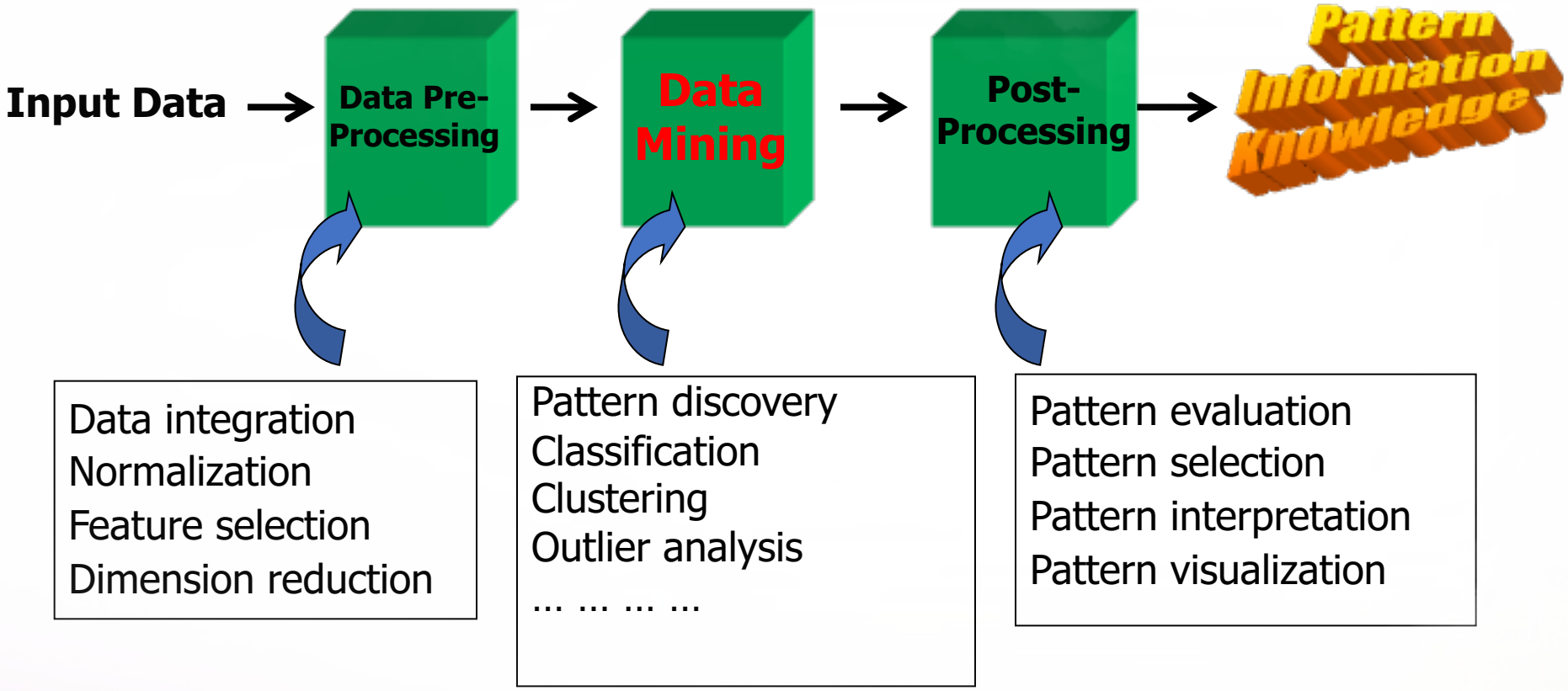
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



KDD Process: A View from ML and Statistics



- This is a view from typical machine learning and statistics communities

Data Mining vs. Data Exploration

- Which view do you prefer?
  - KDD vs. ML/Stat. vs. Business Intelligence
  - Depending on the data, applications, and your focus

- Data Mining vs. Data Exploration
  - Business intelligence view
    - Warehouse, data cube, reporting but not much mining
  - Business objects vs. data mining tools
  - Supply chain example: mining vs. OLAP vs. presentation tools
  - Data presentation vs. data exploration

มุมมองหลายมิติ

Multi-Dimensional View of Data Mining

- **Data to be mined** ข้อมูลที่ขุด
  - Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, …
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted** การดัดแปลง
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: On What Kinds of Data?

ขุดบนข้อมูลประเภทใด?

- Database-oriented data sets and applications (ชุดข้อมูล)
  - Relational database, data warehouse, transactional database (ฐานข้อมูลเชิงสัมพันธ์, ธุรกรรม)
  - Object-relational databases, Heterogeneous databases and legacy databases (ต่างกัน, ดั้งเดิม)
- Advanced data sets and advanced applications (ชุดข้อมูลขั้นสูง)
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences) (อนุกรมเวลา, ชั่วคราว, ลำดับ)
  - Structure data, graphs, social networks and information networks
  - Spatial data and spatiotemporal data (เชิงสัมพันธ์)
  - Multimedia database
  - Text databases
  - The World-Wide Web

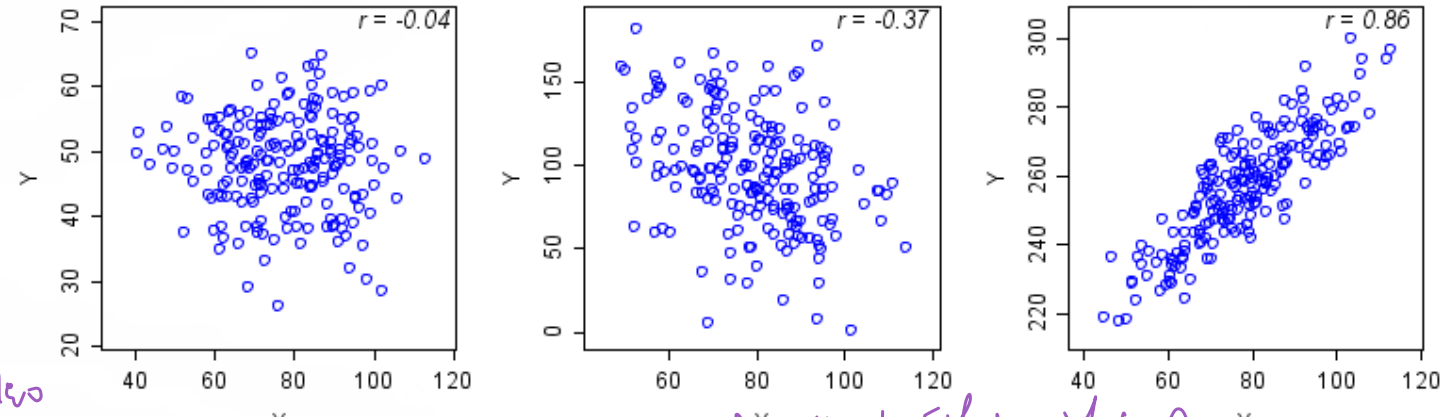
Data Mining Functions: (1) Generalization

ลักษณะทั่วไป

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
  - Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data Mining Functions: (2) Pattern Discovery

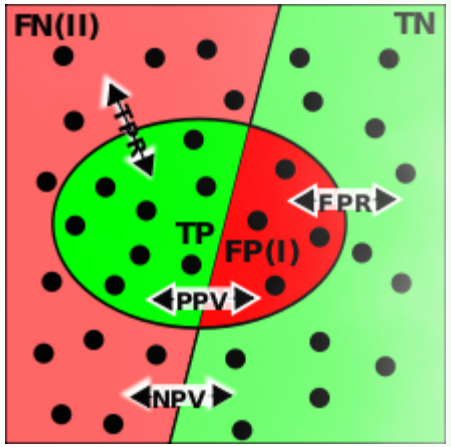
- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



- 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?

Data Mining Functions: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - 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, …



Data Mining Functions: (4) Cluster Analysis

การวิเคราะห์คลัสเตอร์

- Unsupervised learning (i.e., Class label is unknown)

การเรียนรู้แบบไม่มีผู้ดูแล เช่น ไม่รู้จักป้ายกำกับ

- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns

เพื่อสร้างหมวดหมู่ใหม่ จัดคลัสเตอร์เพื่อค้นหารูปแบบการกระจาย

- Principle: Maximizing intra-class similarity & minimizing interclass similarity

เพิ่มความคล้ายสูง ขณะลด

- Many methods and applications