

Data Analysis & Mining: Introduction

Anna Monreale
Computer Science Department

Introduction to Data Mining, 2nd Edition
Chapter I

Large-scale Data is Everywhere!

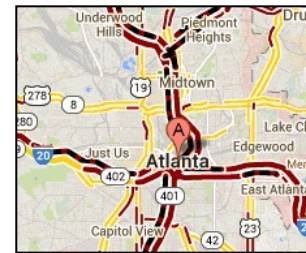
- **Enormous data growth in both commercial and scientific databases**
 - due to advances in data generation and collection technologies
- **New mantra**
 - Gather whatever data you can whenever and wherever possible
- **Expectations**
 - Gathered data will have value either for the purpose collected or for a purpose not envisioned.



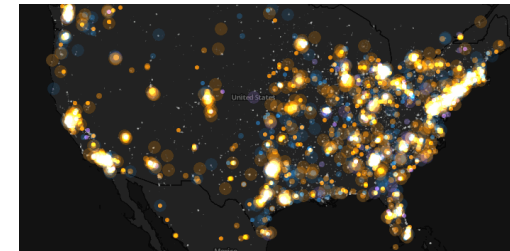
Cyber Security



E-Commerce



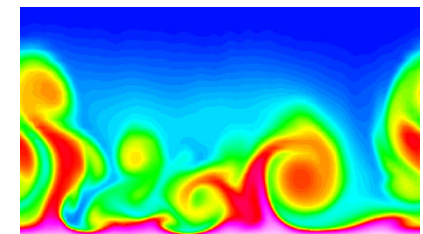
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

Why Data Mining? Commercial Viewpoint

- **Lots of data is being collected and warehoused**

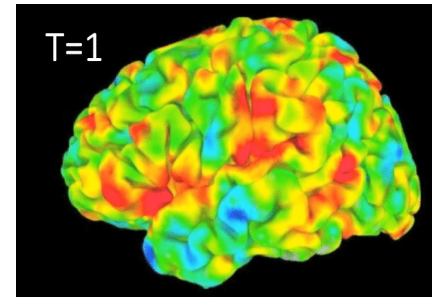
- Web data
 - Yahoo has Peta Bytes of web data
 - Facebook has billions of active users
- purchases at department/grocery stores, e-commerce
 - Amazon handles millions of visits/day
- Bank/Credit Card transactions



- **Computers have become cheaper and more powerful**
- **Competitive Pressure is Strong**
 - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Data Mining? Scientific Viewpoint

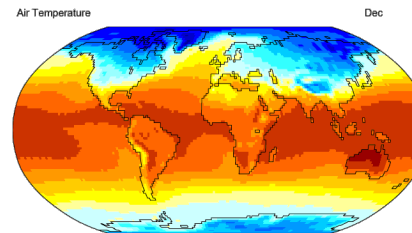
- **Data collected and stored at enormous speeds**
 - remote sensors on a satellite
 - NASA EOSDIS archives over petabytes of earth science data / year
 - telescopes scanning the skies
 - Sky survey data
 - High-throughput biological data
 - scientific simulations
 - terabytes of data generated in a few hours
- **Data mining helps scientists**
 - in automated analysis of massive datasets
 - In hypothesis formation



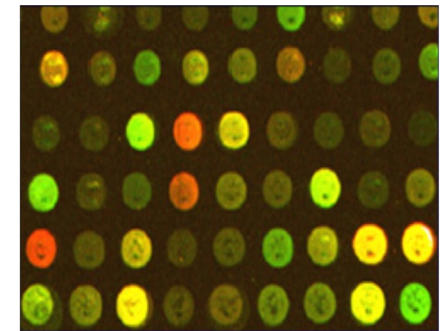
fMRI Data from Brain



Sky Survey Data

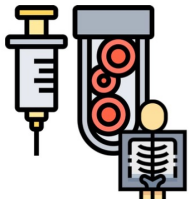


Surface Temperature of Earth



Gene Expression Data

Big data for healthcare: traditional data sources



Vaccinations



Clinical Notes
Medical prescriptions
Pharmaceutical dossiers



Medical Reservations



Health Insurances
Health Certificates



Omics data

Big data for healthcare: new data sources



Smart-phone
Health-apps
GPS Positions
Travels



Smartwatch
Physical activity
Blood pressure
Heartbeat
Quality of sleep



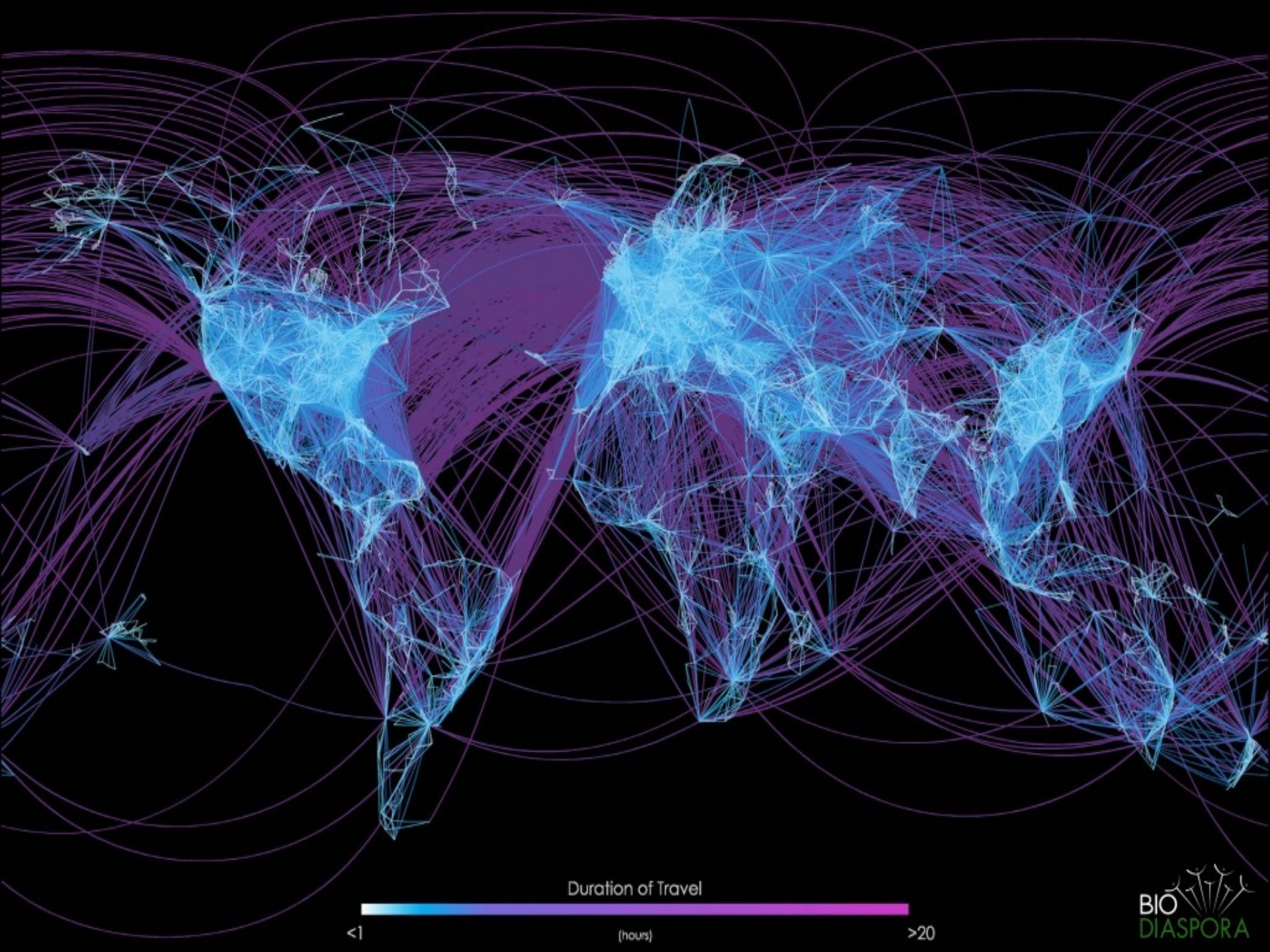
Social Media
Mental Health
Social Interactions

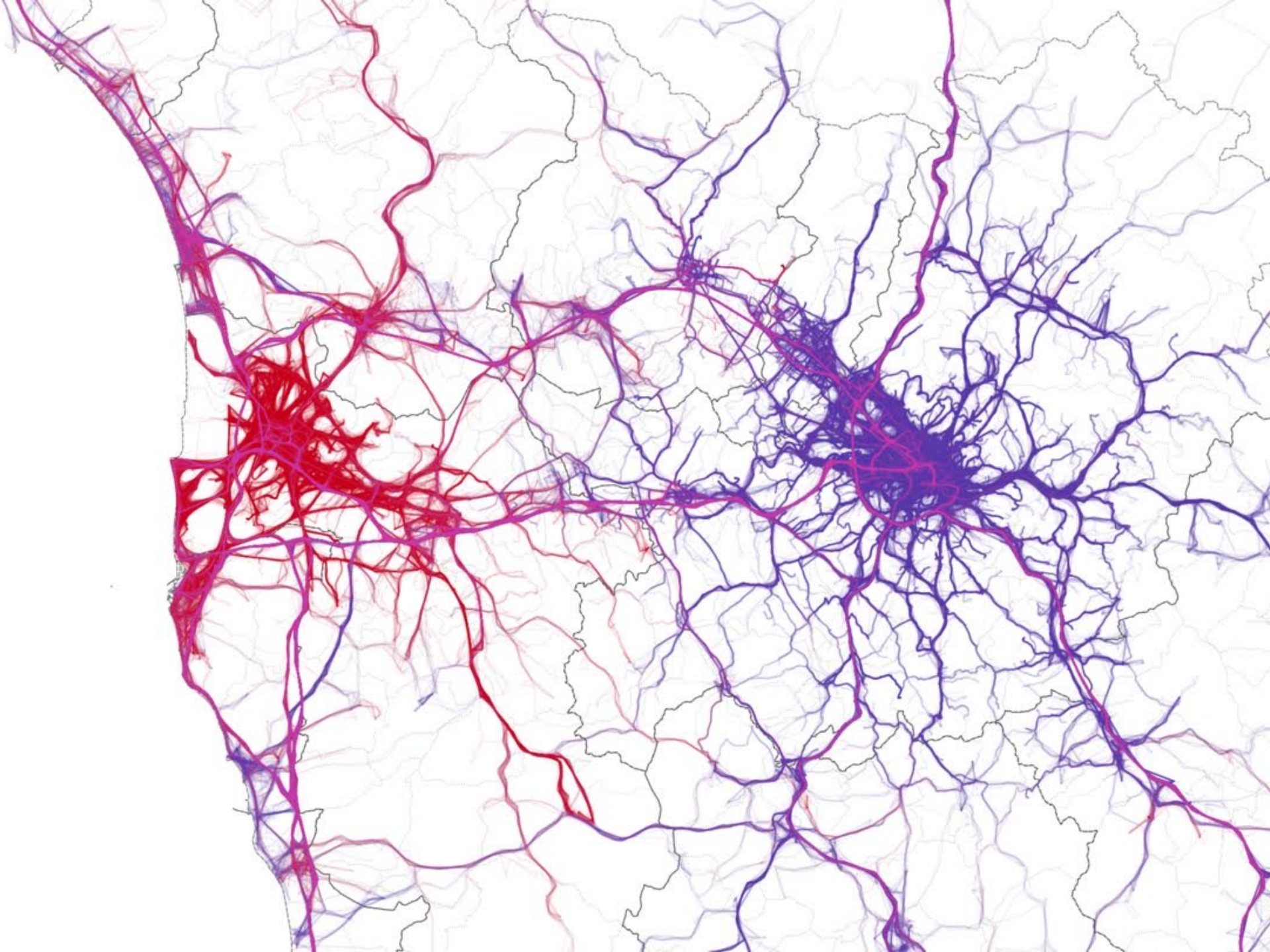


Purchasing
Allergies
Eating habits



Search Engine
Queries on health conditions



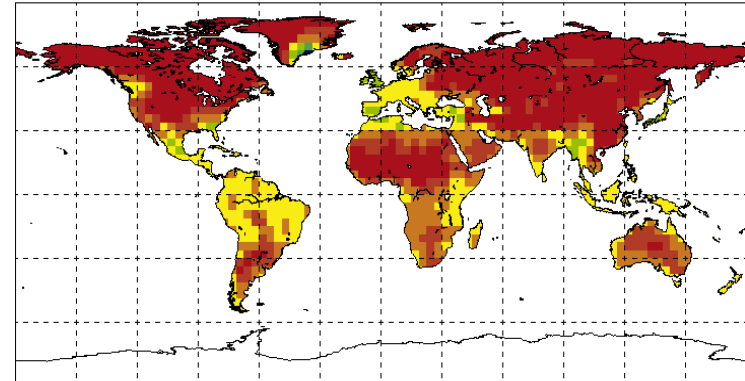


Great Opportunities to Solve Society's Major Problems



Improving health care and reducing costs

CCCma/A2a January to January Mean Temperature (degrees C) 2080s relative to 1961-90



Predicting the impact of climate change



Finding alternative/ green energy sources



Reducing hunger and poverty by increasing agriculture production

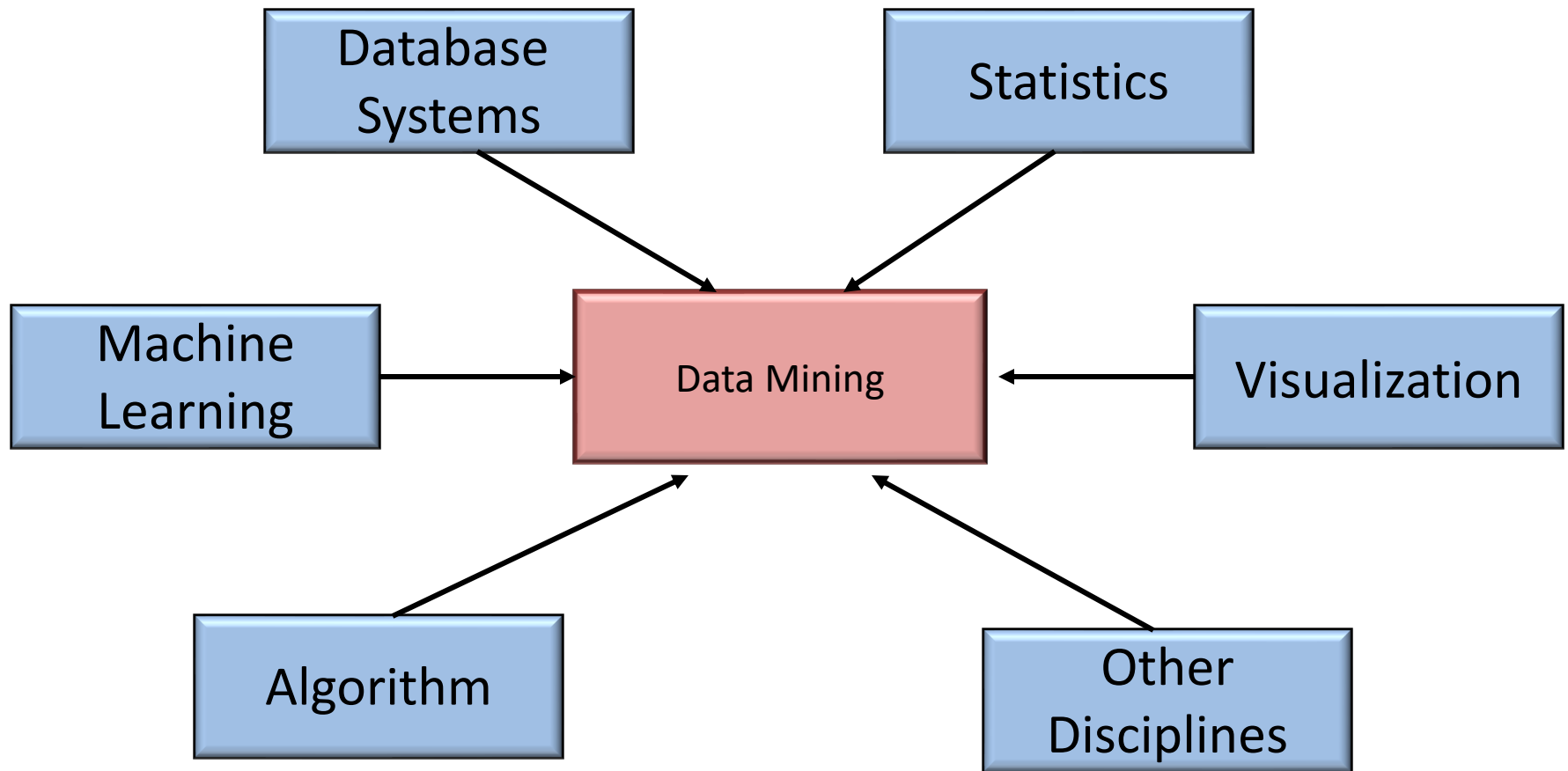
Introduction to Data Mining, 2nd Edition

What we need to extract knowledge?

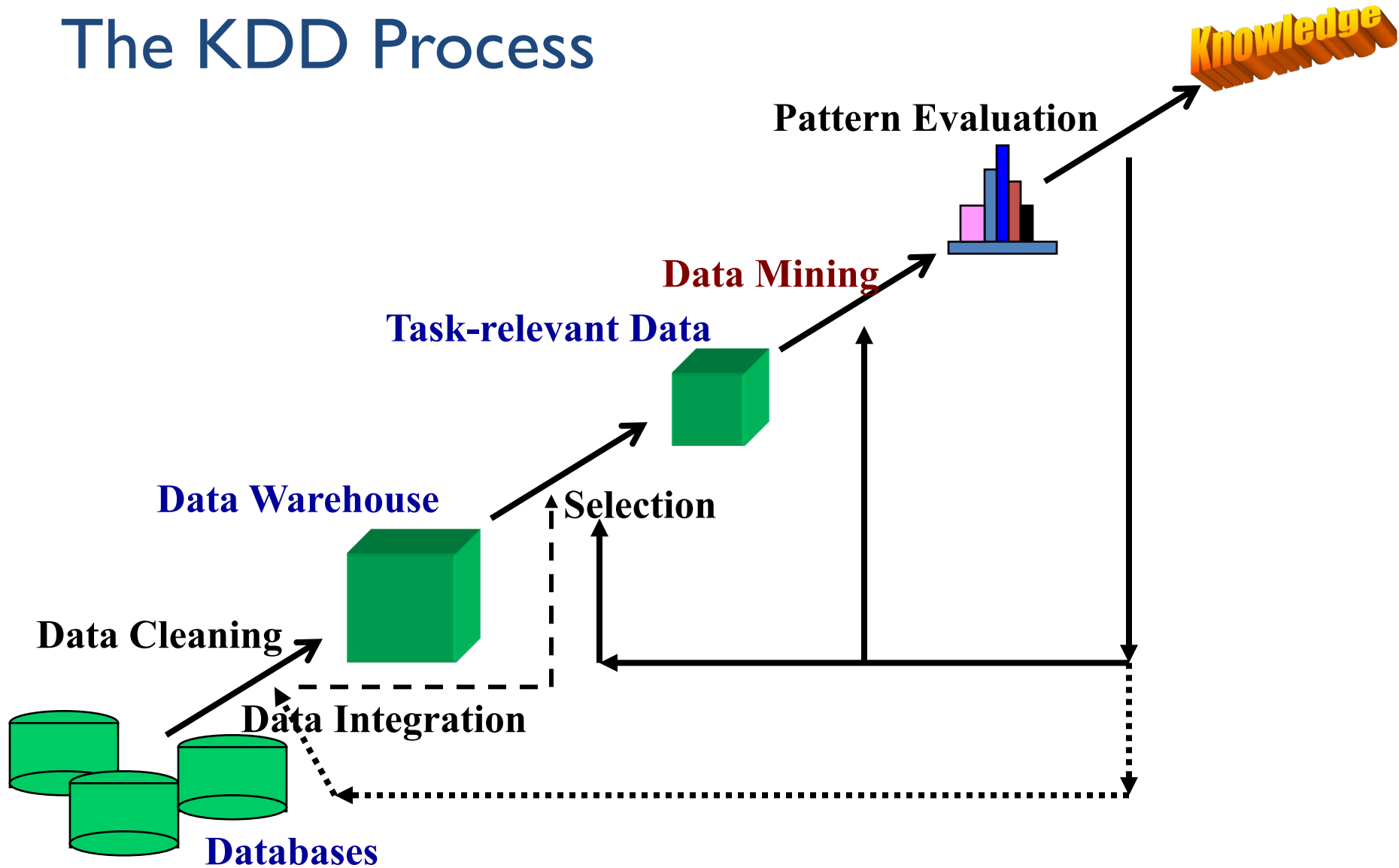
It is the use of **efficient** techniques for the analysis of **very large collections of data** and the **extraction** of useful and possibly unexpected patterns in data (**hidden knowledge**).

DATA MINING!!!

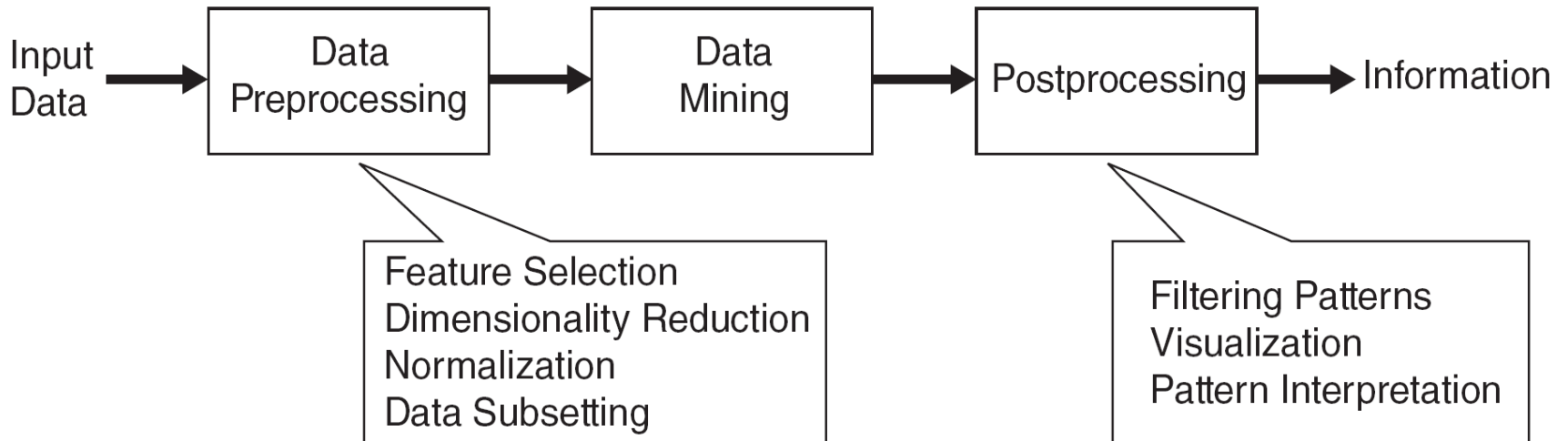
Data Mining: Confluence of Multiple Disciplines



The KDD Process



What is Data Mining?



Primary & Secondary Data

Primary Data

- **Original data** that has been collected for a specific purpose
- Primary data is **not altered by humans**

Secondary Data

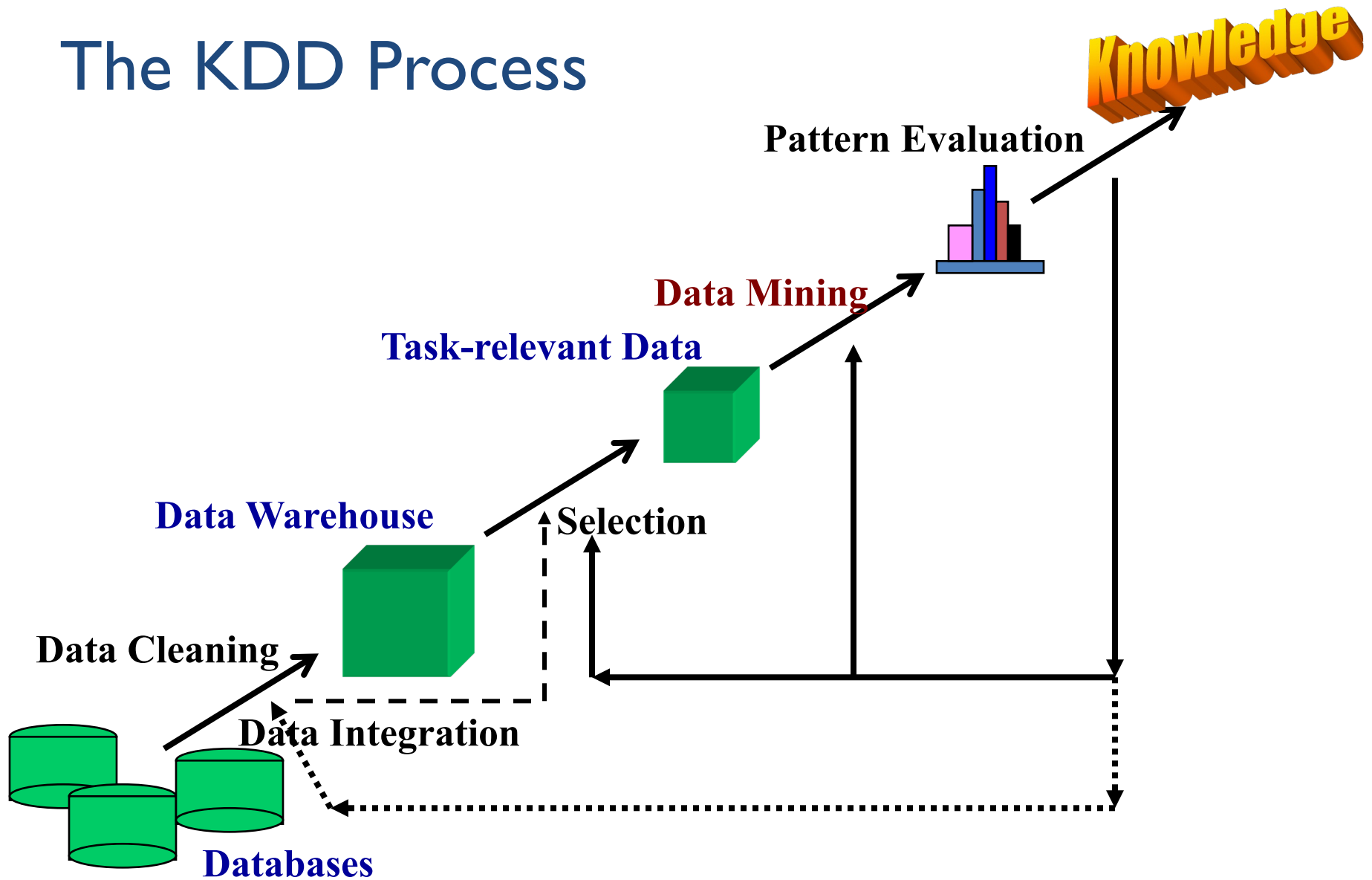
- **Data that** has been already collected and **made available for other purposes**
- Secondary data may be obtained **from many sources**



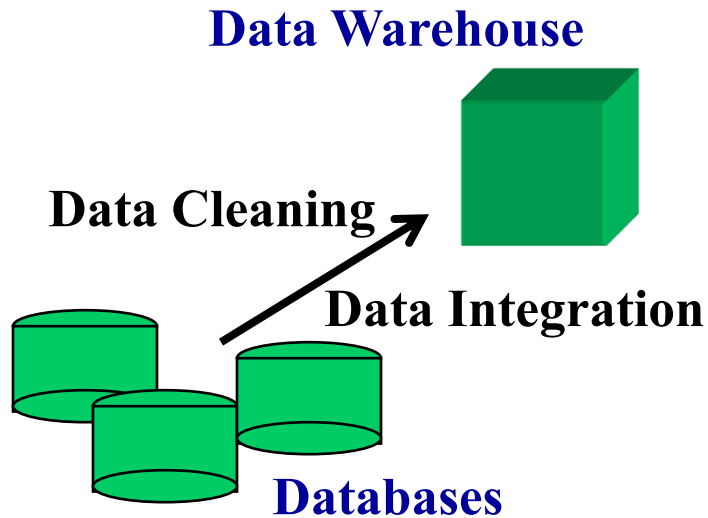
Variety of Data Sources



The KDD Process



Data Integration and Preparation

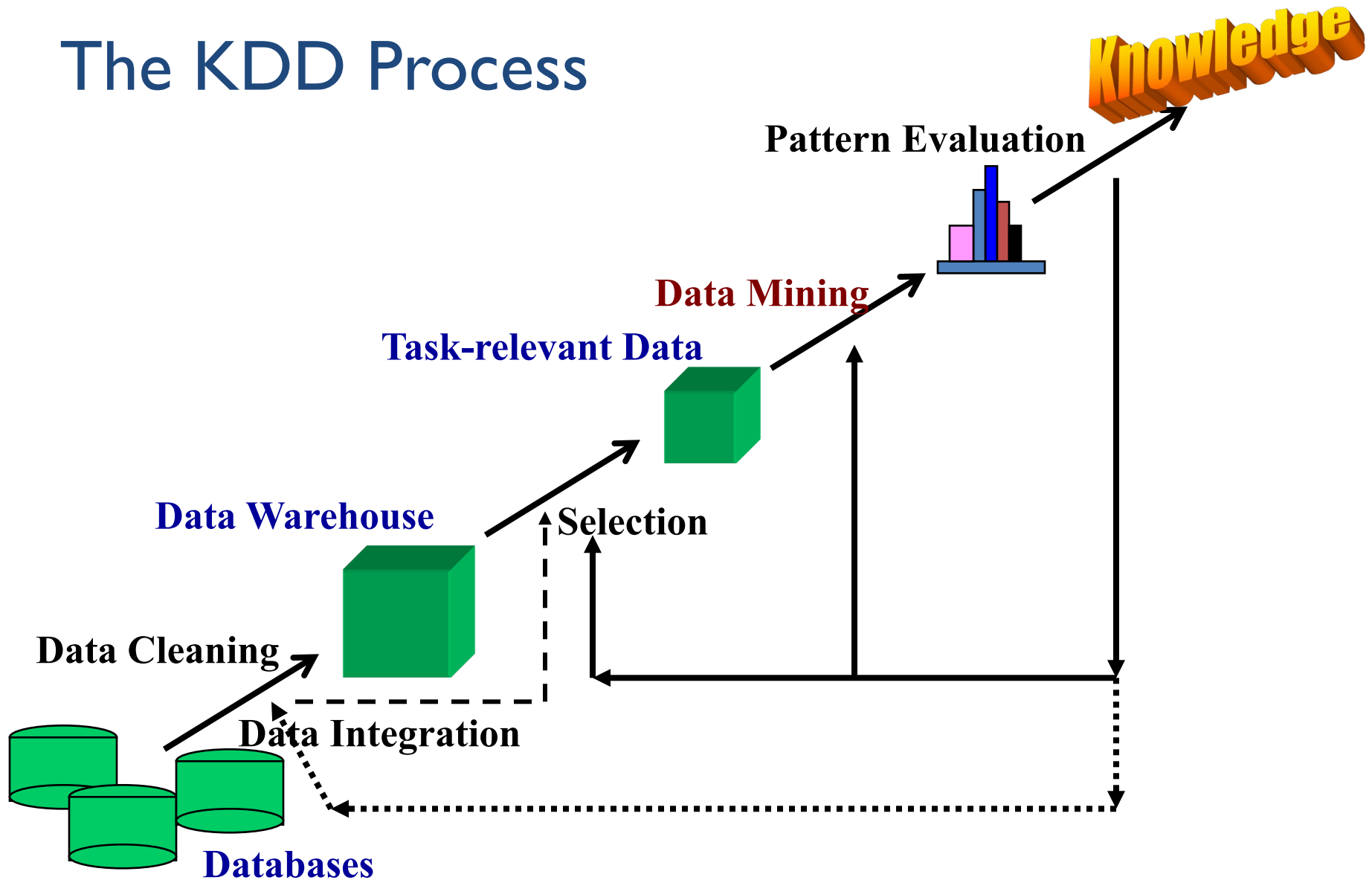


Data Integration involves the process of data understanding, data cleaning, merging data coming from multiple sources and transforming them to load them into a **Data Warehouse**

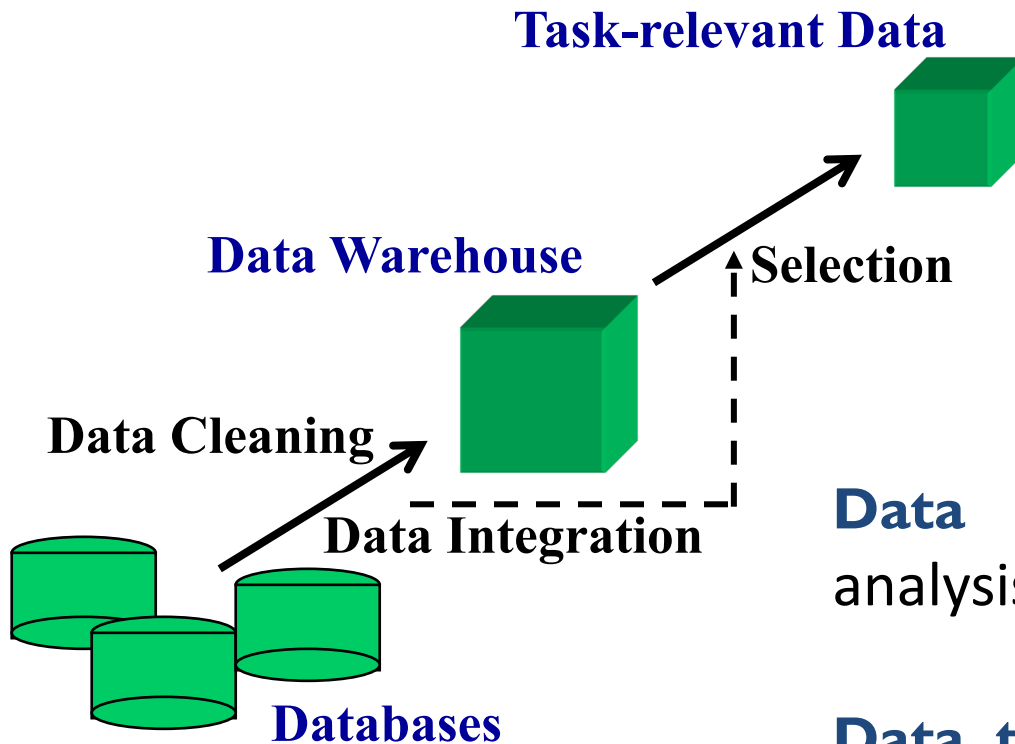
Data Warehouse is a database targeted to answer **specific business questions**

Developing a data analytics project requires the
BUSINESS UNDERSTANDING

The KDD Process



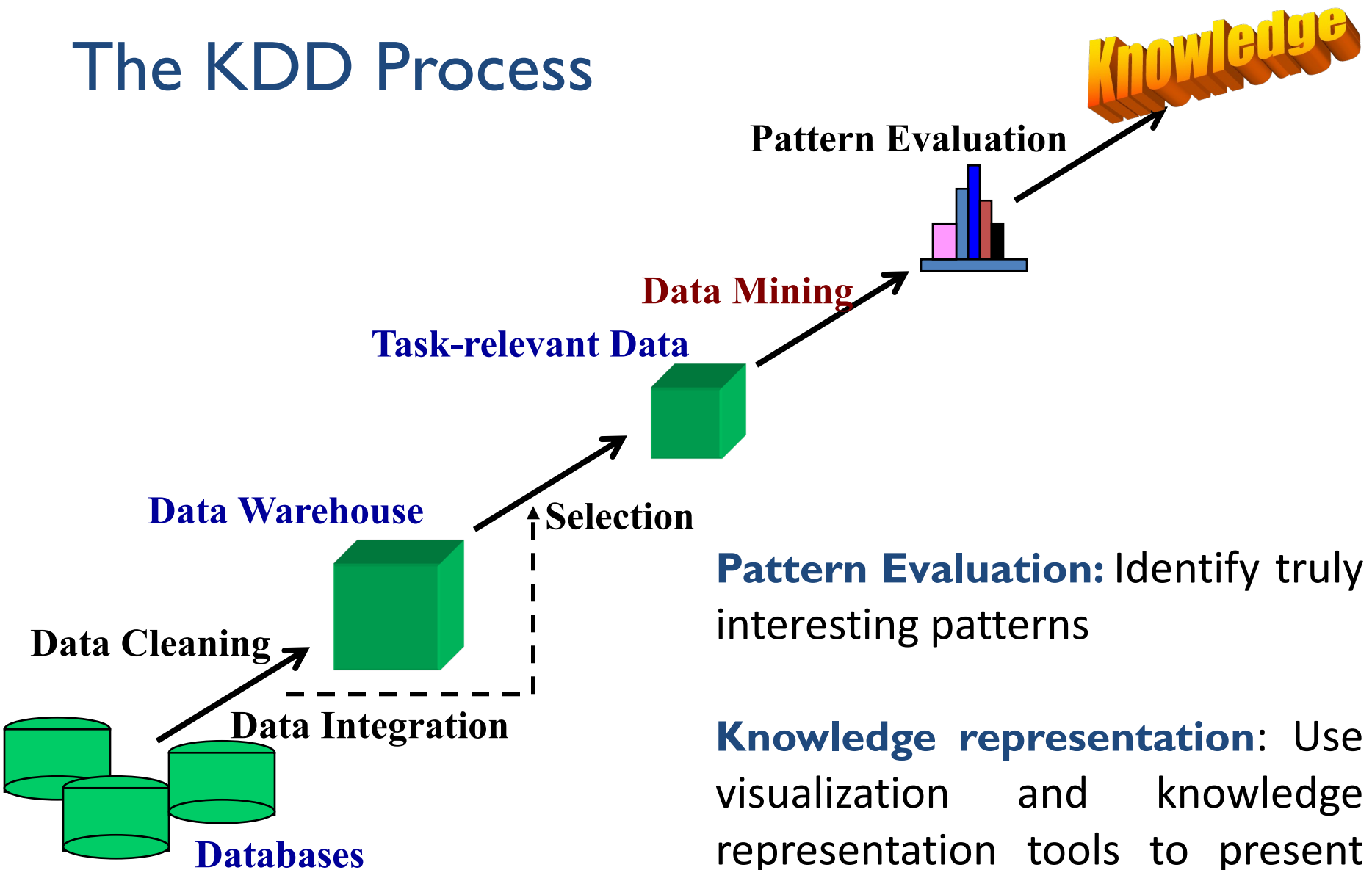
Data Selection and Transformation



Data Selection: Data relevant to analysis tasks are retrieved from data

Data transformation: Transform data into appropriate form for mining (summary, aggregation, etc.)

The KDD Process



Pattern Evaluation: Identify truly interesting patterns

Knowledge representation: Use visualization and knowledge representation tools to present the mined data to the user

Data Mining Tasks

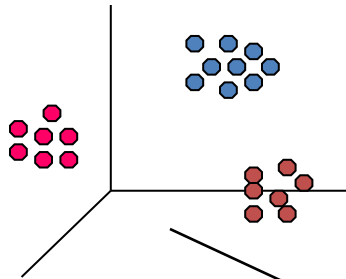
- Prediction Methods
 - Use some variables to predict unknown or future values of other variables.
- Description Methods
 - Find human-interpretable patterns that describe the data.

Data Mining Tasks

- Prediction Methods
 - Use some variables to predict unknown or future values of other variables.
- Description Methods
 - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks ...



Clustering

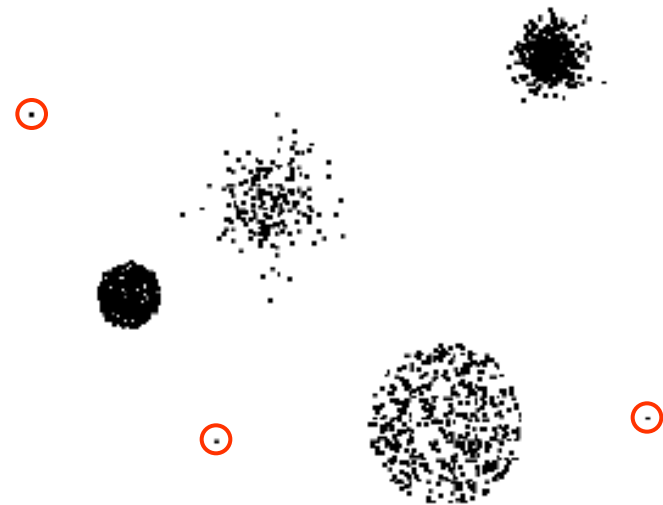
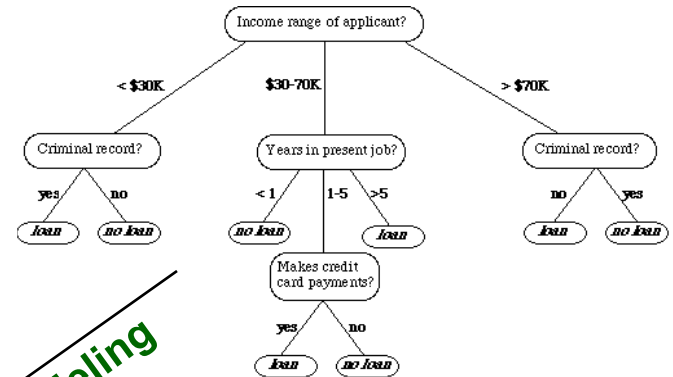
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

Predictive Modeling

Anomaly Detection



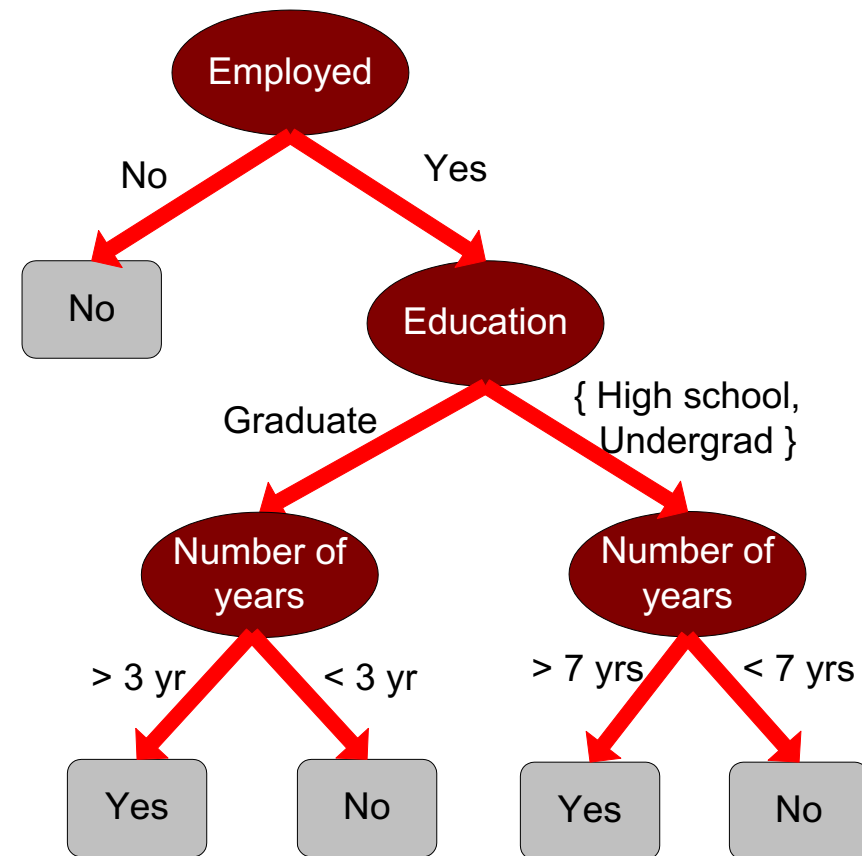
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

Model for predicting credit worthiness

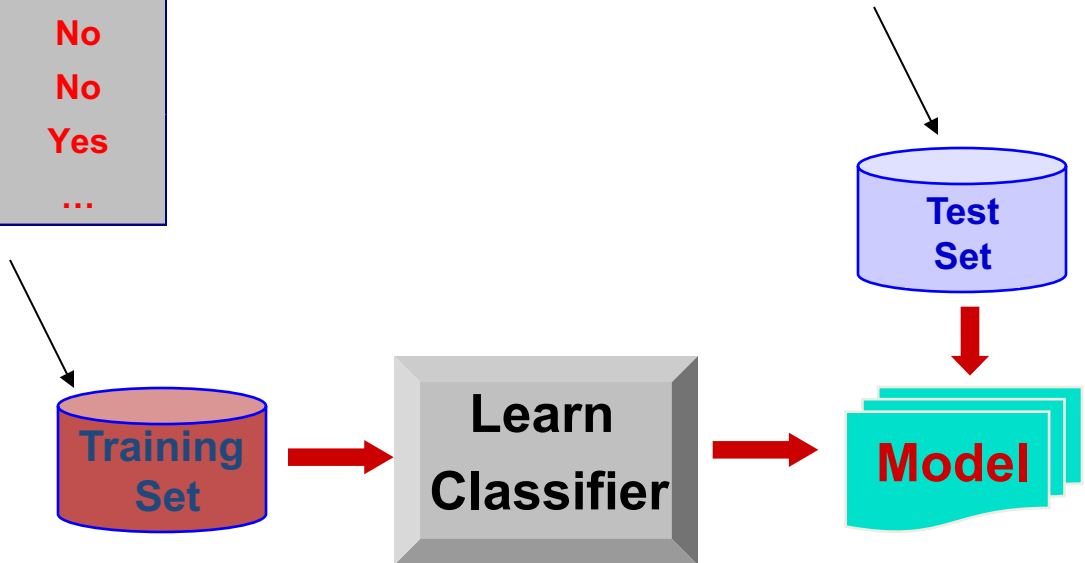


Classification Example

categorical categorical quantitative class

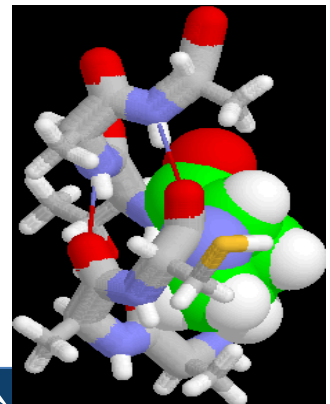
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...

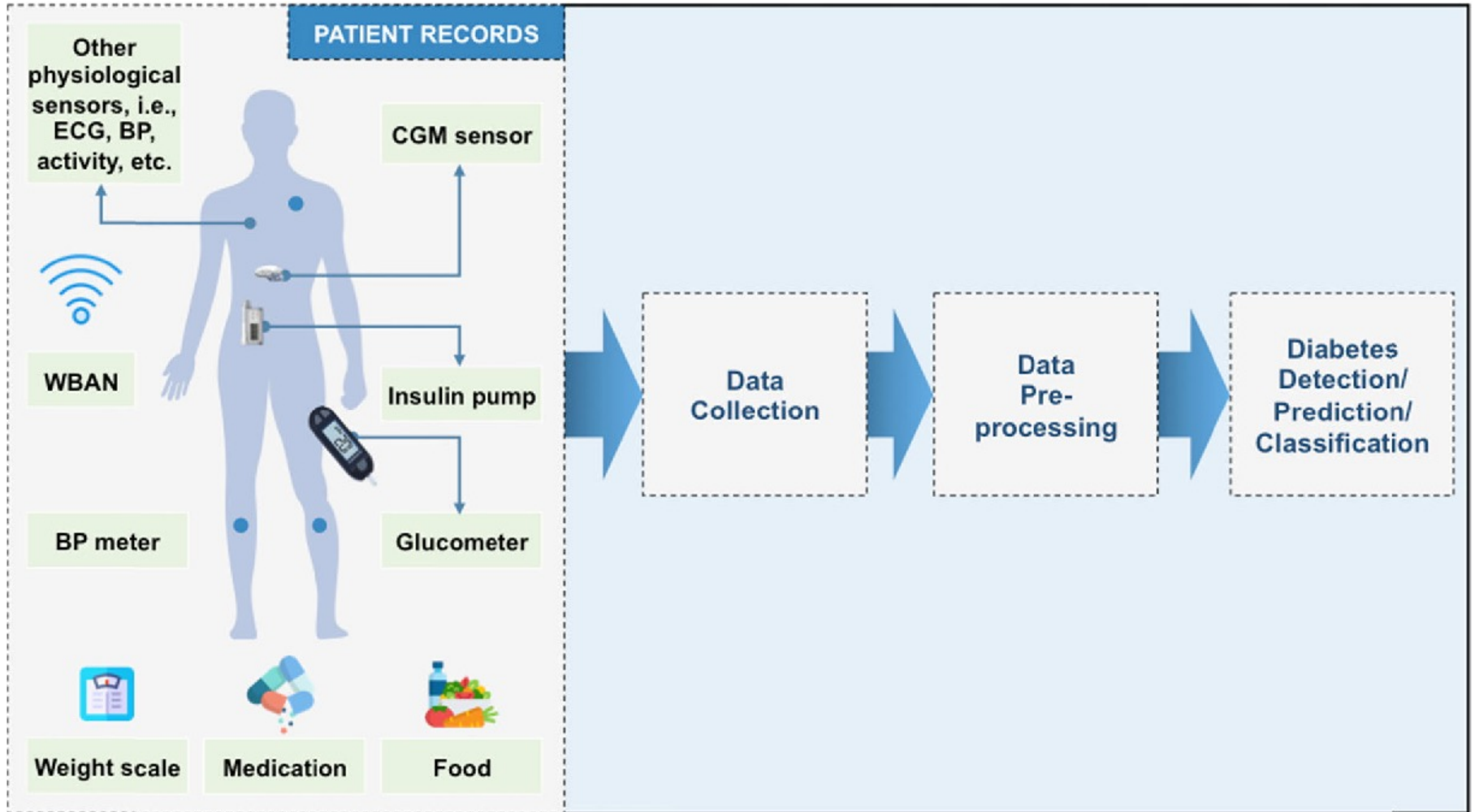


Examples of Predictive Task

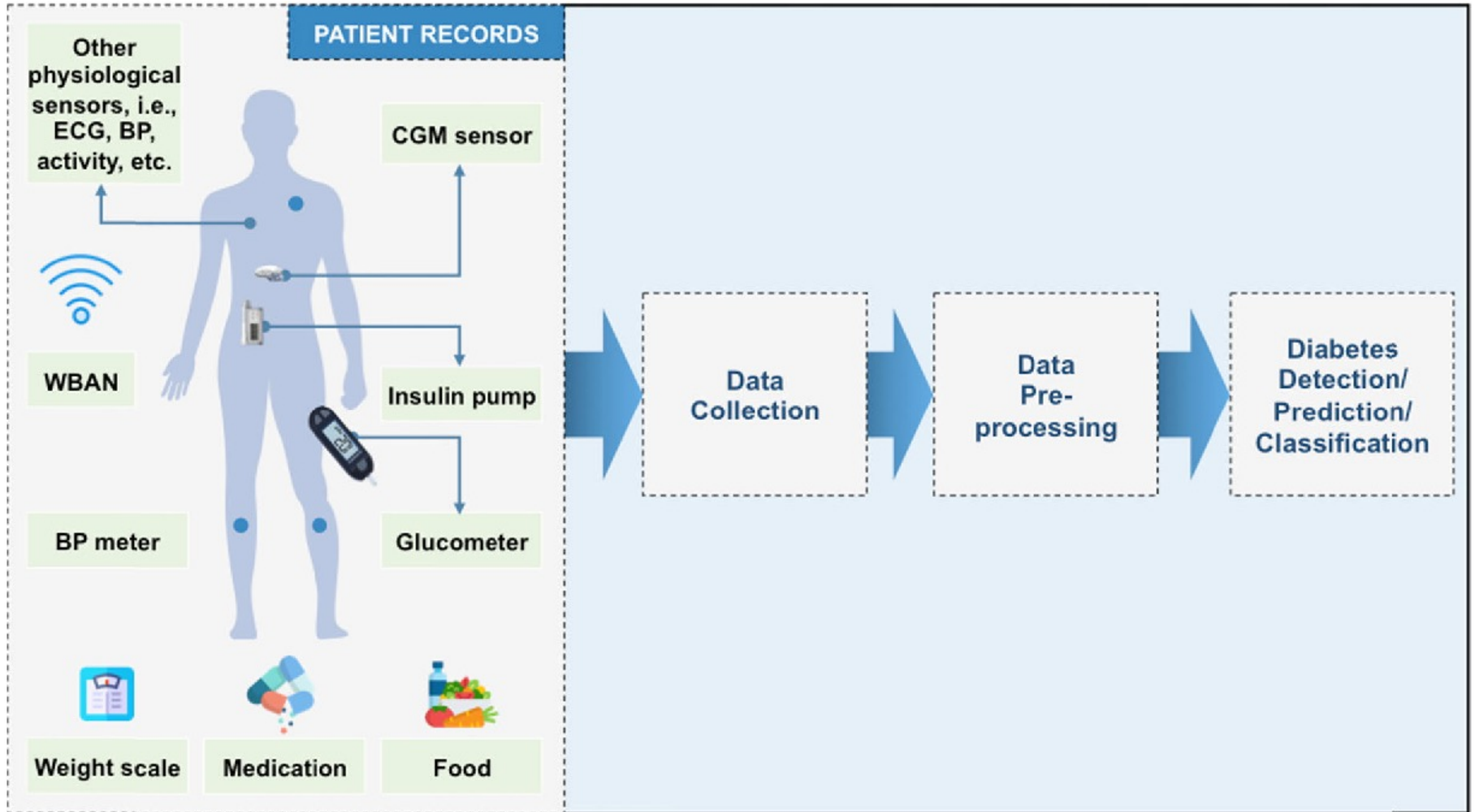
- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



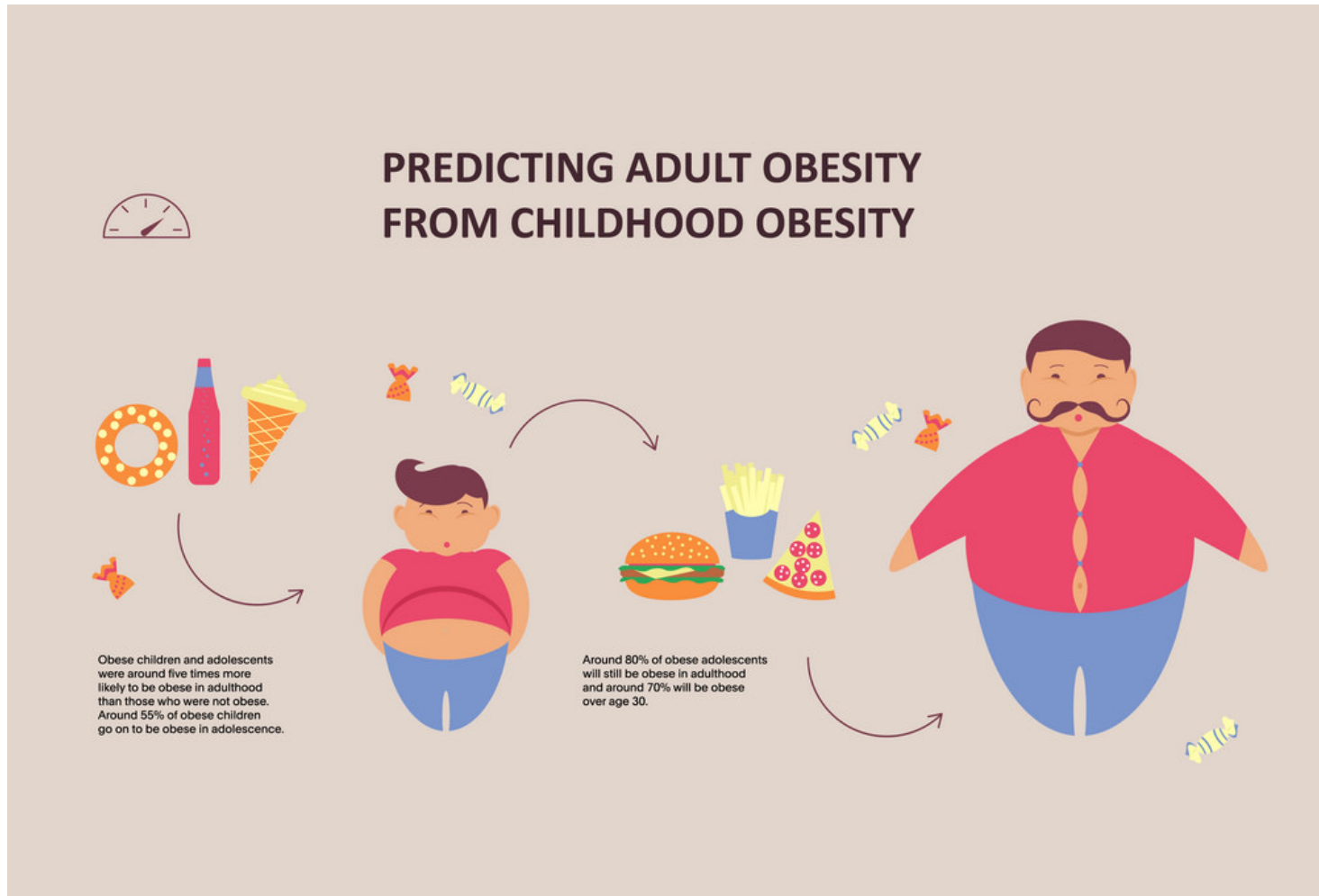
Example of Predictive Task



Example of Predictive Task

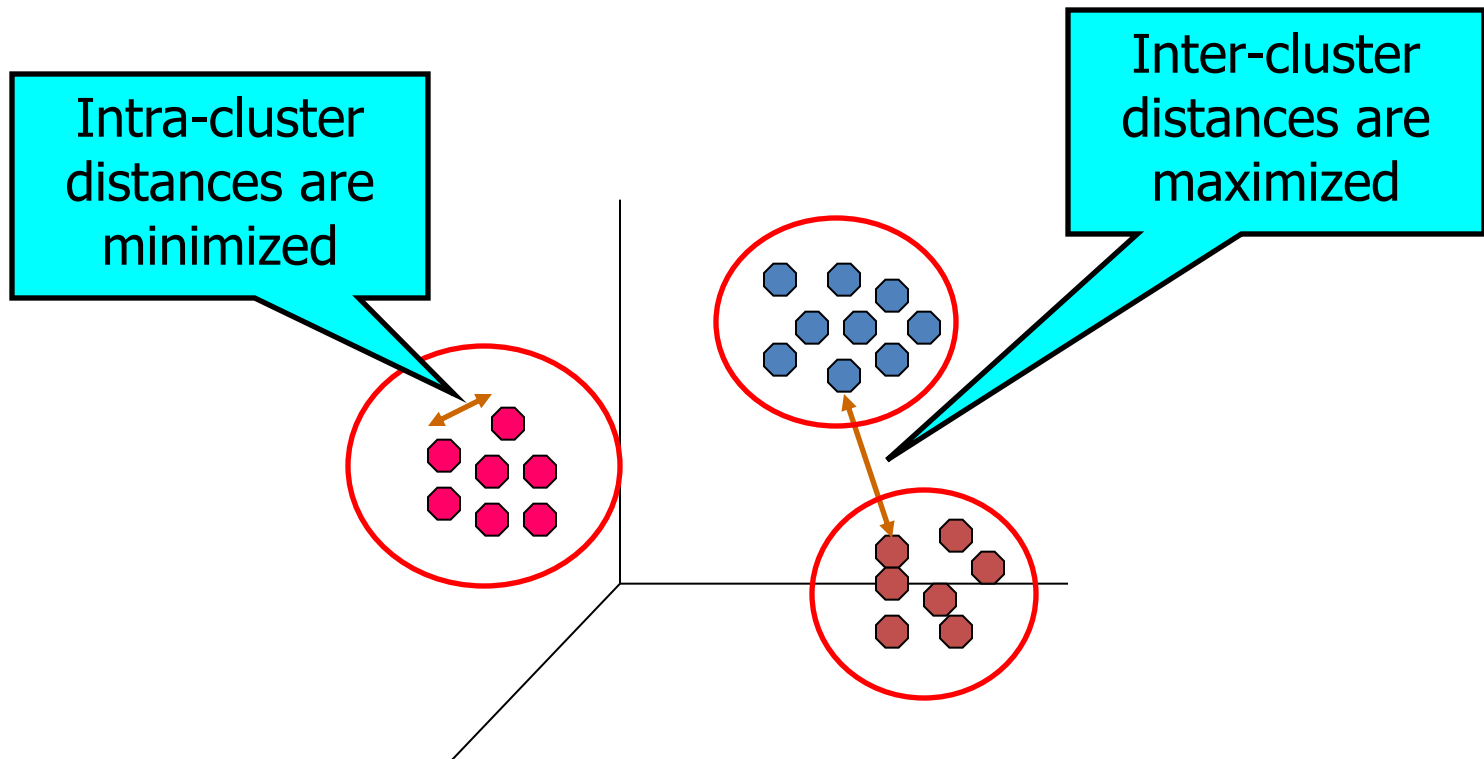


Example of Predictive Task

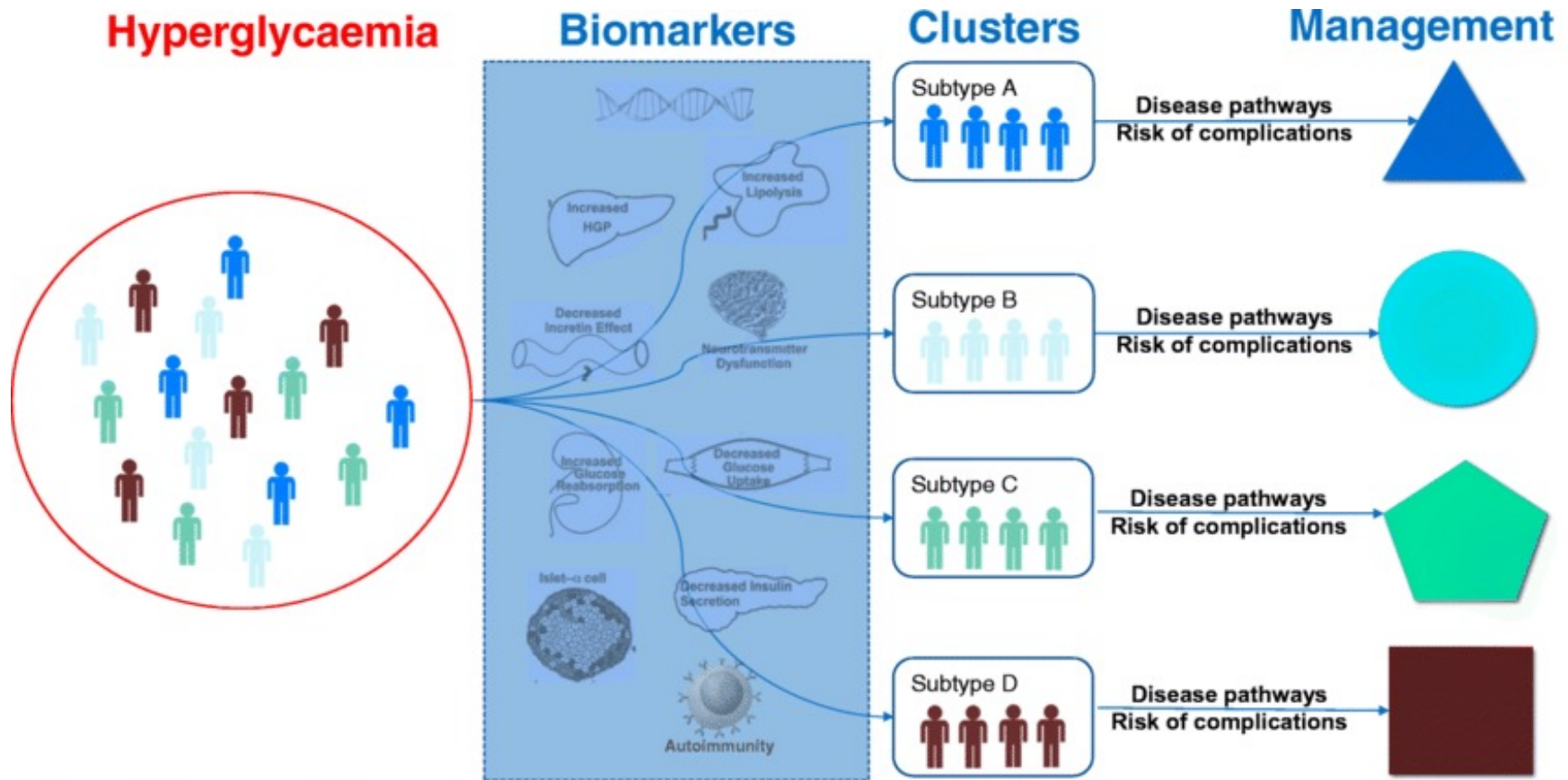


Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Example of Clustering Analysis



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
 - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	HIV, PapSmear
2	Flu test, hCG, Glucose
3	Flu test, Urinalysis
4	Urinalysis, hCG, HIV
5	Urinalysis, hCG, Glucose

Rules Discovered:

$\{hCG\} \rightarrow \{Urinalysis\}$

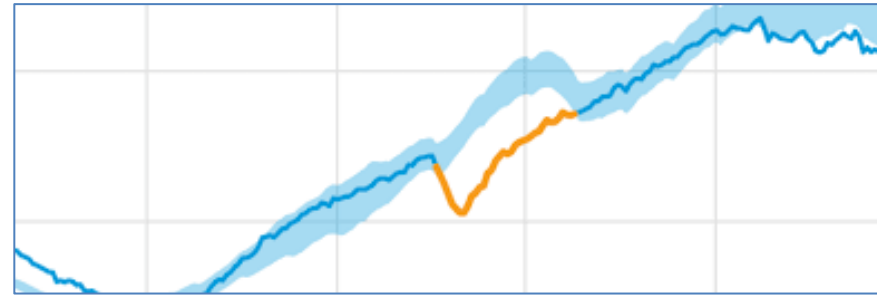
$\{Urinalysis, Glucose\} \rightarrow \{hCG\}$

Association Analysis: Applications

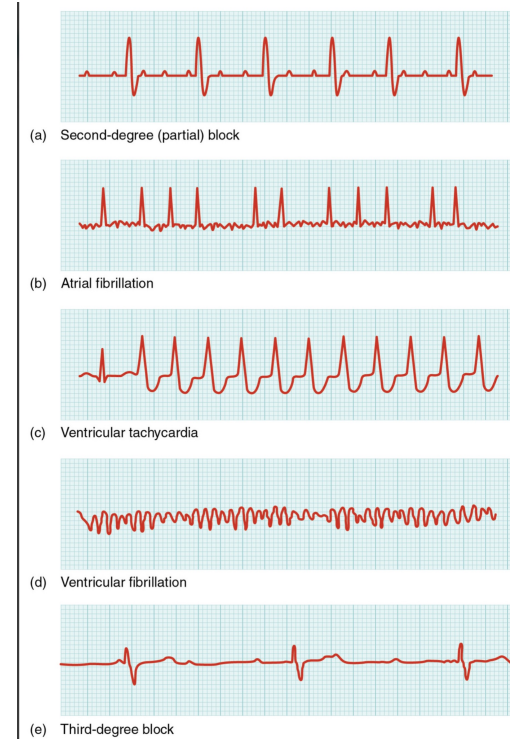
- **Market-basket analysis**
 - Rules are used for sales promotion, shelf management, and inventory management
- **Telecommunication alarm diagnosis**
 - Rules are used to find combination of alarms that occur together frequently in the same time period
- **Medical Informatics**
 - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior



- Applications:
 - Credit Card Fraud Detection
 - Network Intrusion Detection
 - **Identify anomalous behavior in ECG**



Motivating Challenges

Traditional techniques may be unsuitable due to some challenges:

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis