Introduction to Data Mining

Data Mining: Data Mining Pipeline

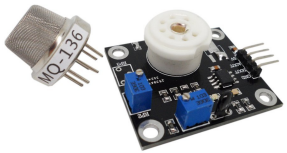
with **Dr. Qin Lv**

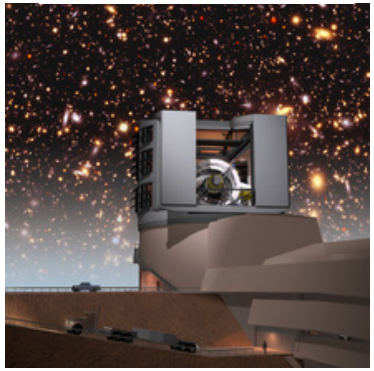
Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

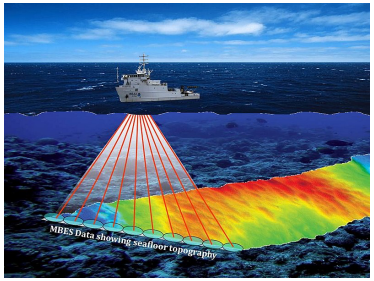
Learning objective: Identify different views of data mining; understand the key issues in data mining.

Into the Digital Era

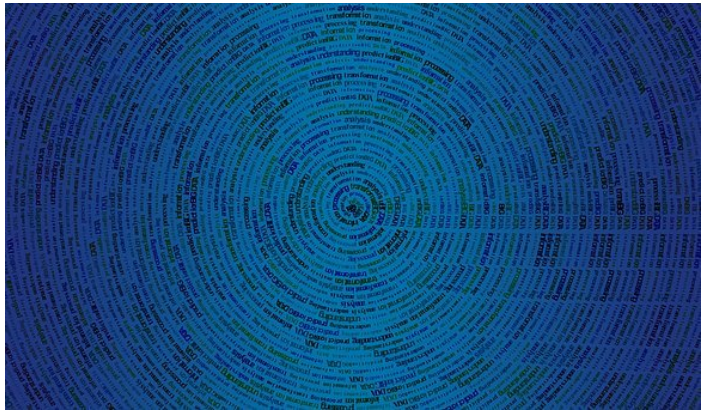
- People's daily lives
  - > 4 billion internet users
  - Social media, smart devices, …
- Scientific discovery
  - Rubin Observatory: 20TB/night
- Many application domains…







Why Data Mining?



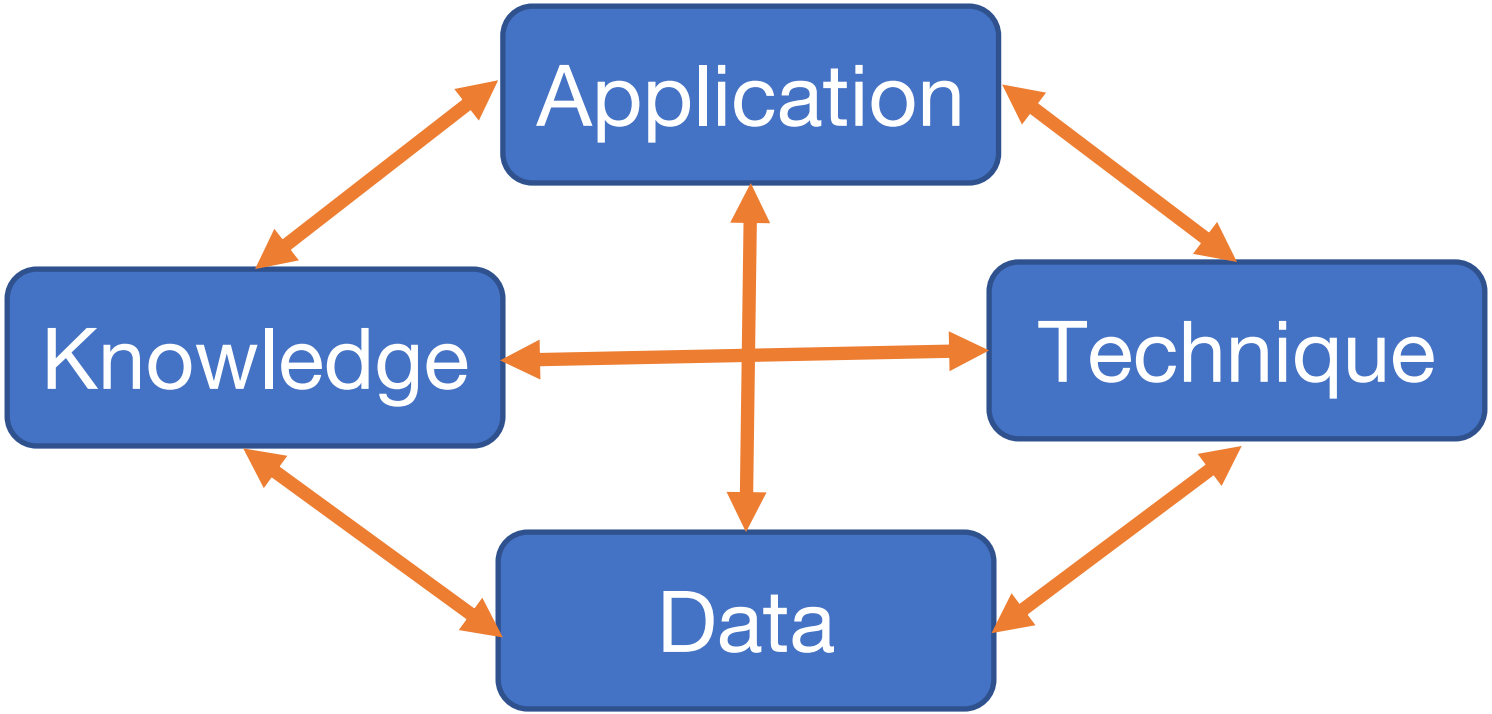
- Explosive data growth
  - KB, MB, GB, TB, PB, EB, ZB
  - Data creation, transmission, storage, sharing, processing
- Drowning in data & starving for knowledge
- Need automated analysis of massive data

What is Data Mining?

➢ Knowledge discovery from data

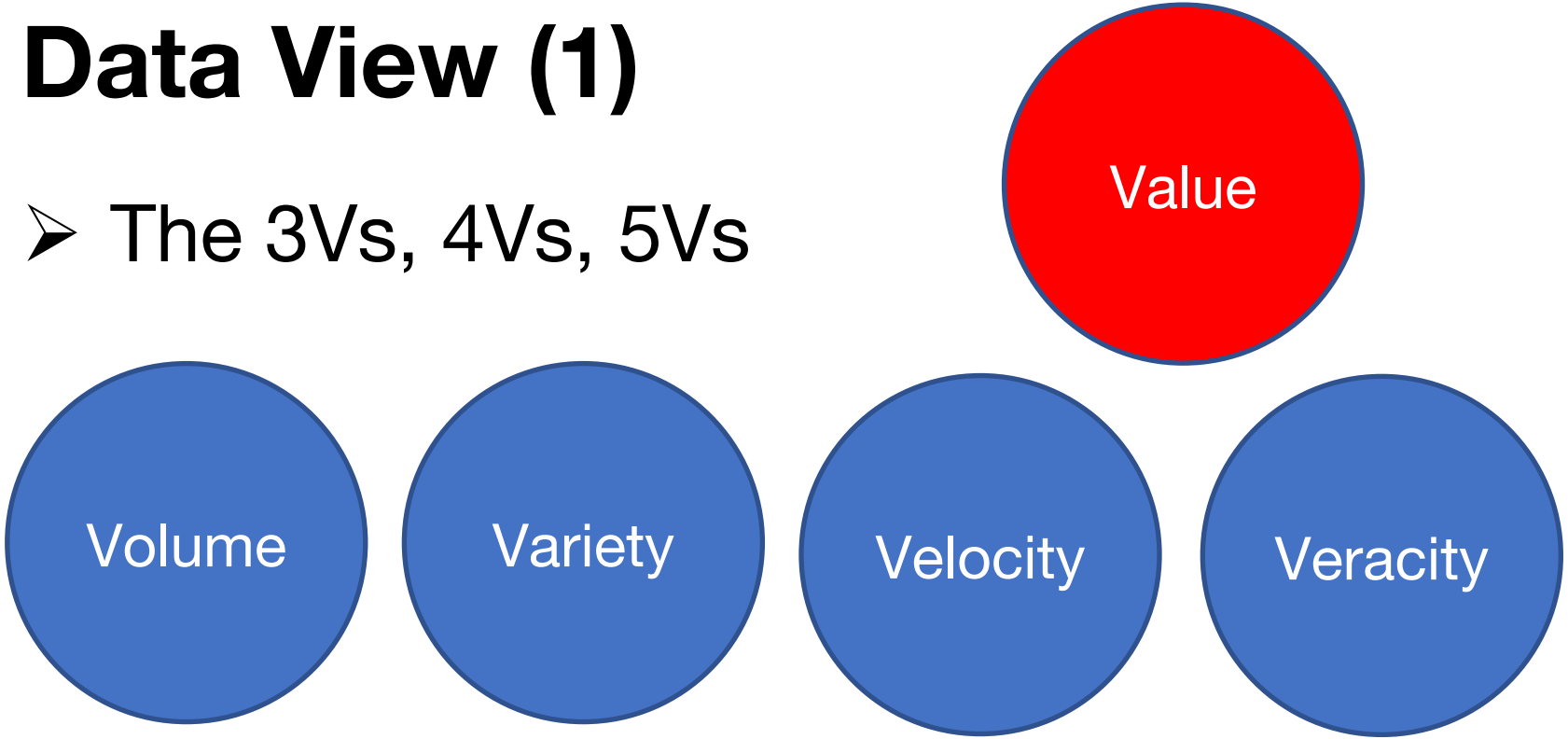
- Extraction of interesting patterns or knowledge from huge amounts of data
- Interesting: valid, previously unknown, potentially useful, ultimately understandable by human
- Huge amounts of data: scalability, efficiency

Data Mining: Four Views



Data View (1)

- The 3Vs, 4Vs, 5Vs



Data View (2)

- Relational, transactional data
  - E.g., student records, bank accounts, store purchases
- Sequential, temporal, streaming data
  - E.g., gene sequences, stock prices, sensor readings
- Spatial, spatial-temporal data
  - E.g., land use, bird migration, traffic condition

Data View (3)

- ➢ Text, multimedia, Web data
  - E.g., news articles, audio/video/image, hypertext
- ➢ Graph, network data
  - E.g., social network, power grid, co-authorship

- ➢ Single or mixture of multiple data types

Application View (1)

- ➢ Market analysis, target advertisement
  - E.g., customer profiling, product recommendation
- ➢ Healthcare, medical research
  - E.g., disease diagnosis, patient care, drug discovery
- ➢ Science and engineering
  - E.g., air pollution, marine life, electric vehicles

Application View (2)

- Security
  - E.g., surveillance, intrusion/crime, fraud, cyberattack
- Government, nonprofit
  - E.g., urban planning, traffic control, education

- And many many more …

Knowledge View (1)

- ➢ Frequent pattern, association, correlation
  - E.g., songs listened together or in certain sequence
  - E.g., A is (more/less) likely to happen given B
- ➢ Categorization
  - E.g., similarity among users with certain purchases
  - E.g., differences between two patient groups

Knowledge View (2)

- Anomaly, outliers
  - E.g., sensor errors, fraud activities, extreme events
- Changes over time
  - E.g., emerging new patterns, shift of user interest

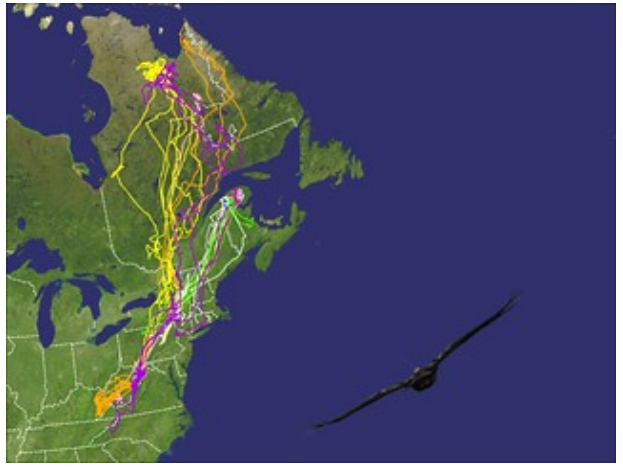
- Descriptive, predictive, prescriptive

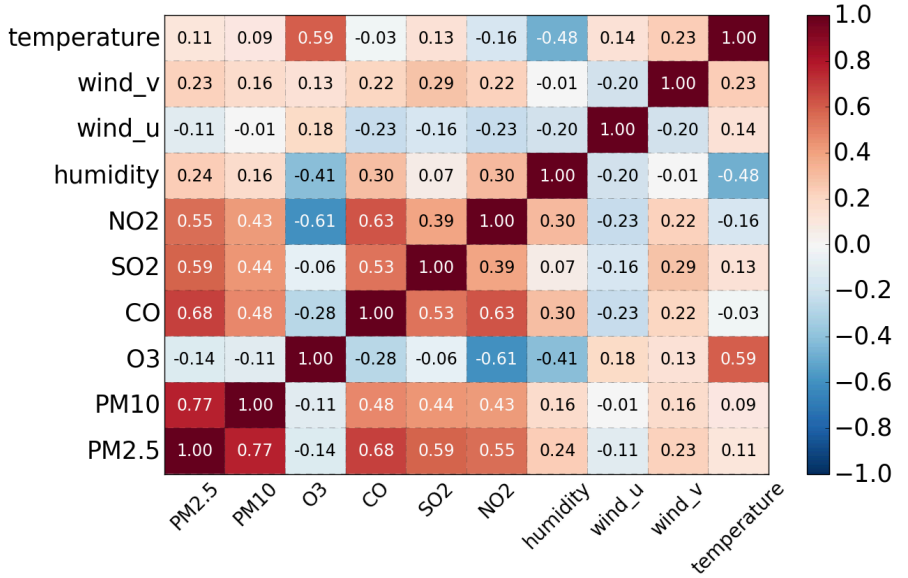
Technique View

- Frequent pattern analysis
- Classification, prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis

Frequent Pattern Analysis

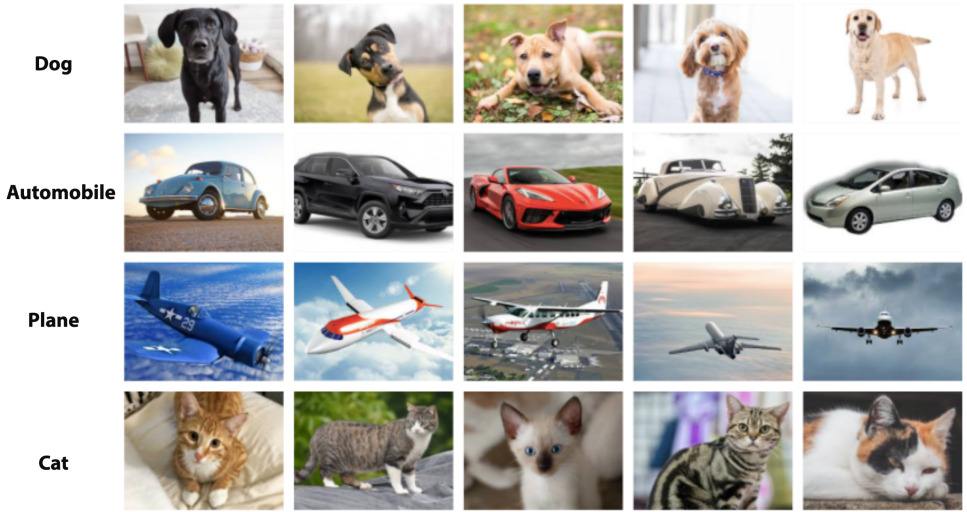
- Frequent itemset
- Frequent sequence
- Frequent structure
- Association rules
- Correlation analysis





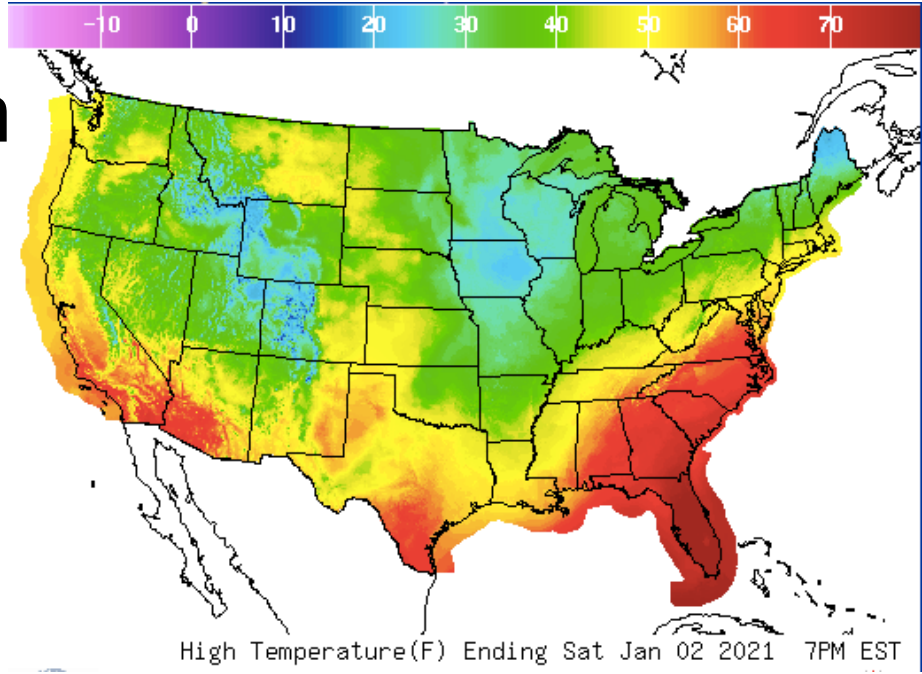
Classification

- Pre-defined classes
- Need training data
- Build model to distinguish classes



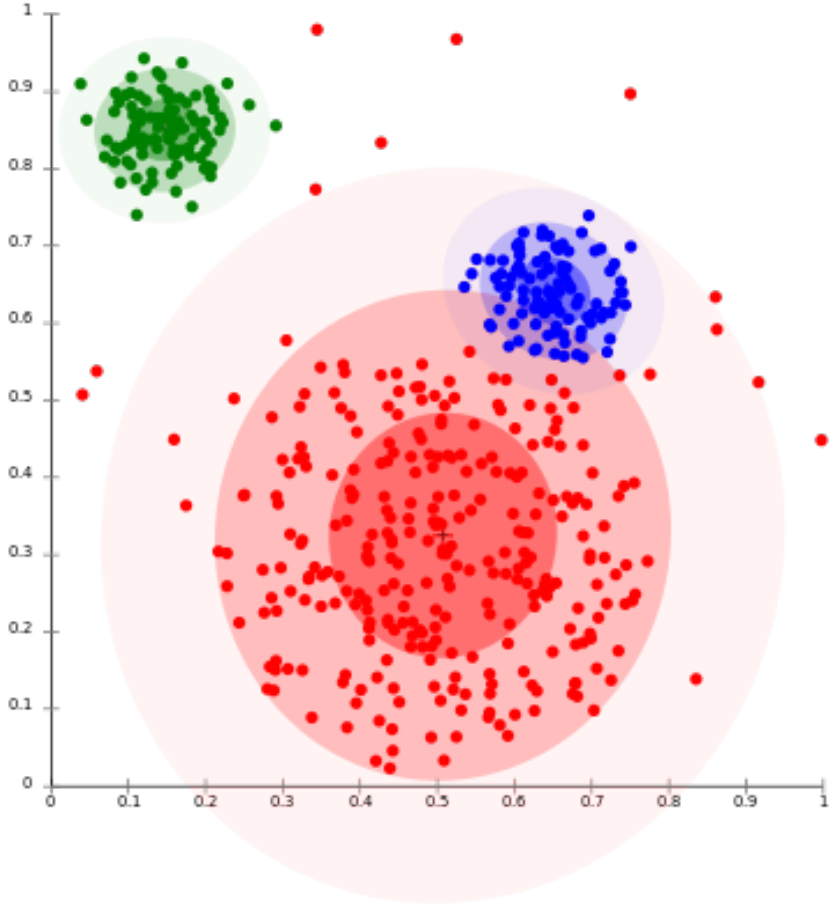
Prediction

- Numerical prediction (continuous value)
  - E.g., weather
  - E.g., stock price
  - E.g., traffic



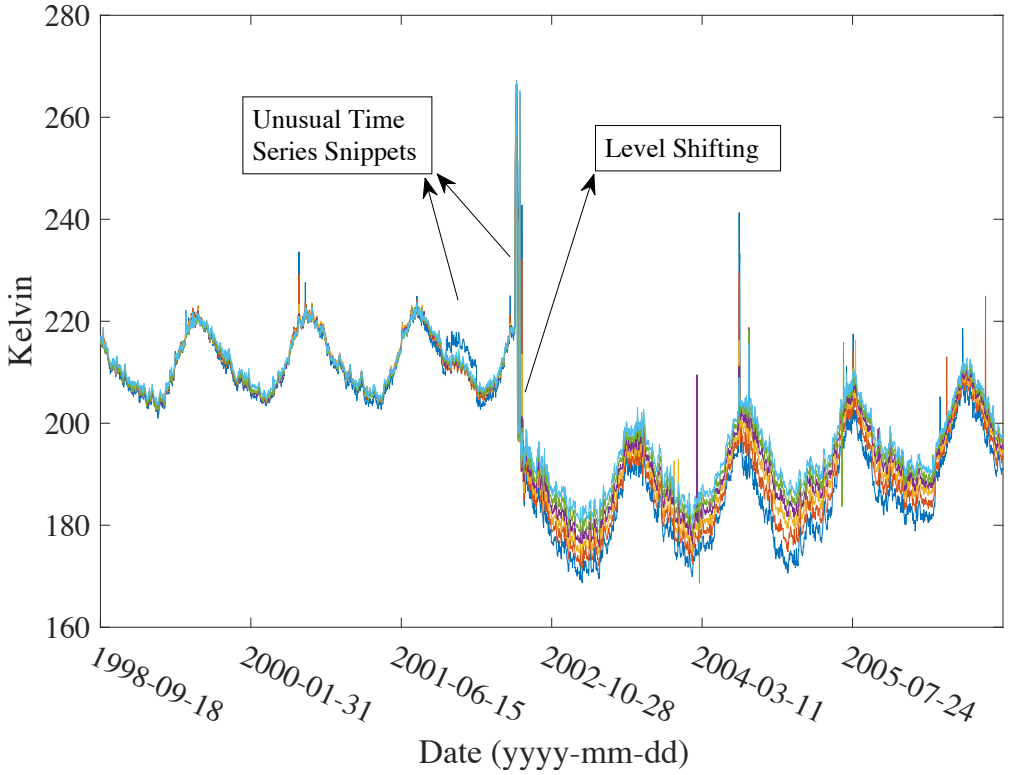
Clustering

- No predefined classes
- Intra-cluster similarity
- Inter-cluster dissimilarity



Anomaly Detection

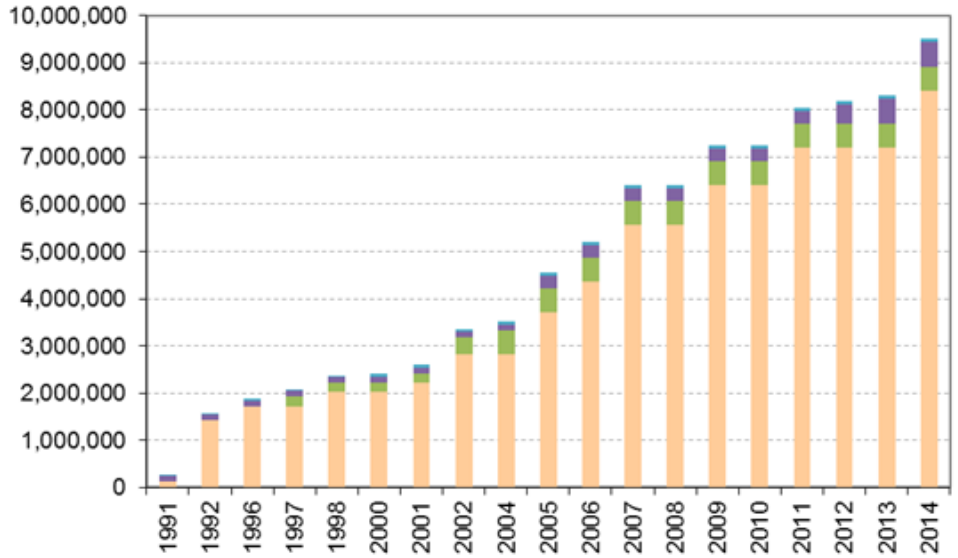
- Anomaly/outlier
  - Differ from the “norm”
  - E.g., error, noise
  - E.g., fraud
  - E.g., extreme events



Trend and Evolution Analysis

- Changes over time
  - Overall trend
  - Periodical patterns
  - Anomalies
  - E.g.,

Google Trends



Data Mining: Four Views