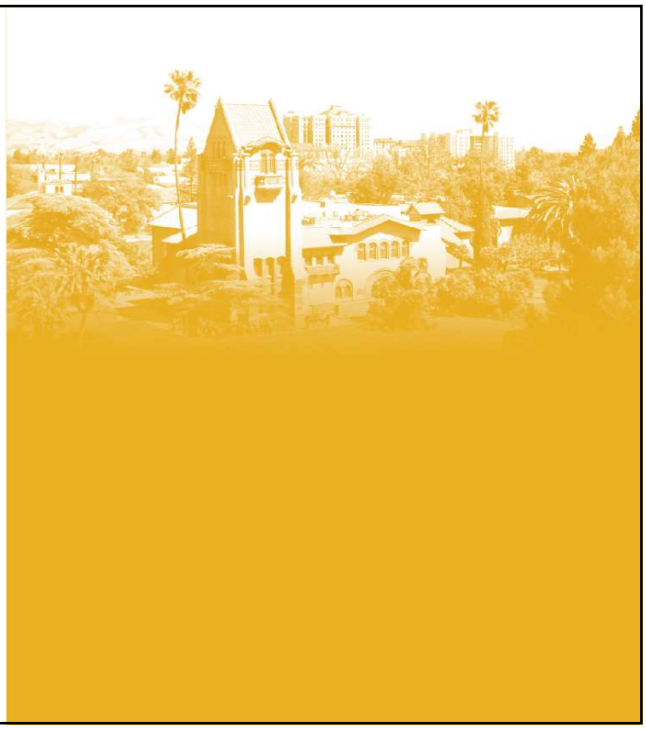
SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to Data Mining



1

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Summary

2

SJSU SAN JOSÉ STATE UNIVERSITY

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!

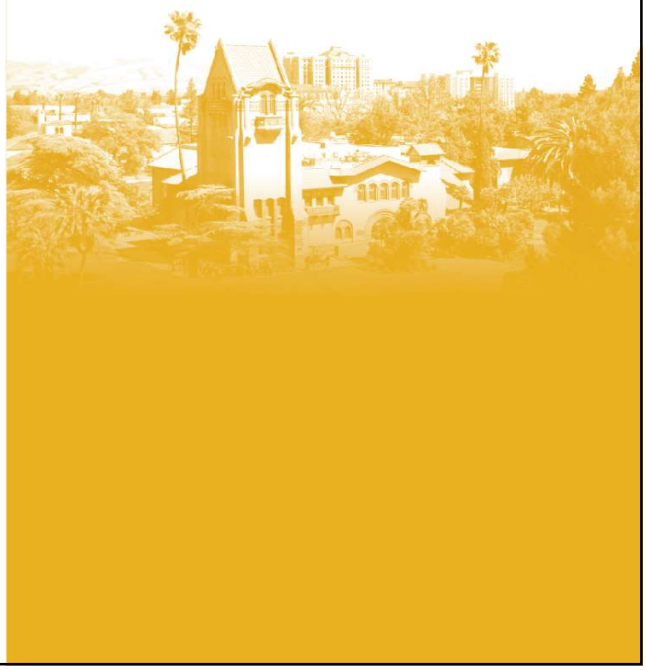
SJSU SAN JOSÉ STATE UNIVERSITY

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge Discovery (mining) in Databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Not everything is "data mining"…
  - Simple search and query processing
  - (Deductive) Expert systems

SJSU SAN JOSÉ STATE UNIVERSITY

Recall from DATA270…

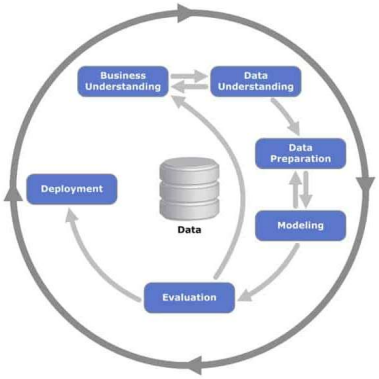


5

SJSU SAN JOSÉ STATE UNIVERSITY

CRISP-DM Methodology

- CRISP-DM: CRoss Industry Standard Process for Data Mining.
- A standard on data mining/data science process published in 1999.
- Six phases that naturally describes the data science life cycle.
- Usually implemented with other PM approaches.



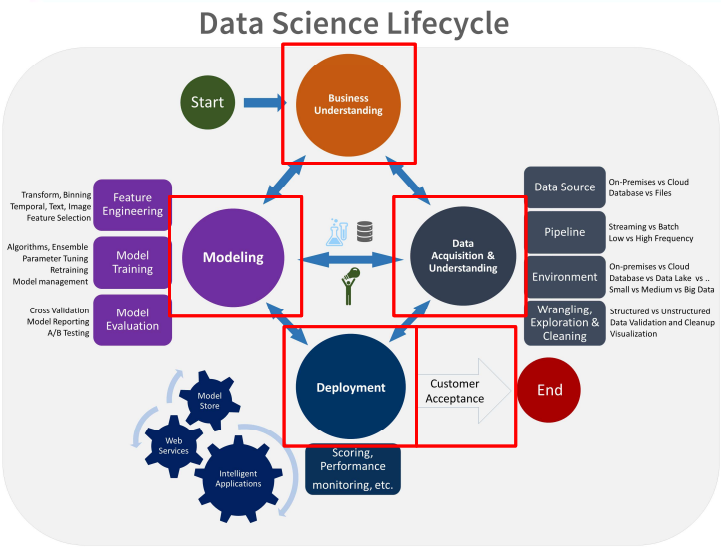
https://www.datascience-pm.com/crisp-dm-2/

6

6

SJSU SAN JOSÉ STATE UNIVERSITY

Team Data Science Process (TDSP)



https://docs.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle-data

7

SJSU SAN JOSÉ STATE UNIVERSITY

Data Process

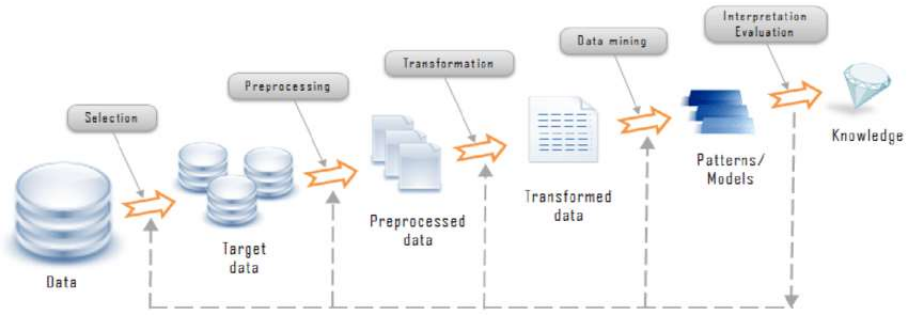


Image adopted from internet srouce

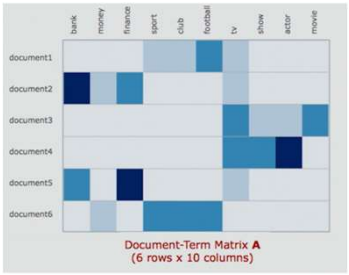
8

SJSU SAN JOSÉ STATE UNIVERSITY

Record or Tabular Dataset

Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4
S0007	12/1/2012 9:34:07 AM	C0015	P005	1

Record or Transaction Data



Document-Term Matrix

	Variables			
Cases	sepal length	sepal width	petal length	petal width	class
	5.1	3.5	1.4	0.2	Iris-setosa
	4.9	3	1.4	0.2	Iris-setosa
	6.5	3.2	5.1	2	Iris-virginica
	6.4	2.7	5.3	1.9	Iris-virginica
	6.8	3	5.5	2.1	Iris-virginica
	6.7	3.1	4.4	1.4	Iris-versicolor
	5.6	3	4.5	1.5	Iris-versicolor
	5.8	2.7	4.1	1	Iris-versicolor

Data Matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 9 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 \\ 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

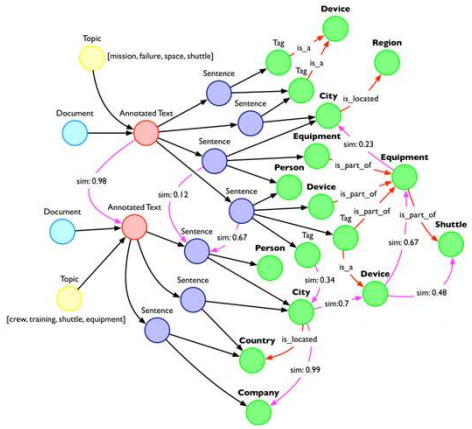
Rows	Columns	Values
5	6	6
0	4	9
1	1	8
2	0	4
2	3	2
3	5	5
4	2	2

Sparse Data Matrix

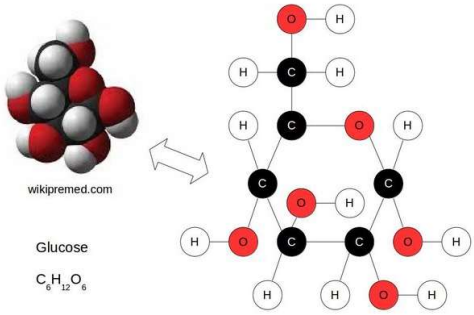
Images adopted from various internet pages

SJSU SAN JOSÉ STATE UNIVERSITY

Graph-Based Dataset



Data with Relationships among Objects



Data with Objects that are Graphs

Images adopted from various sources.

SJSU SAN JOSÉ STATE UNIVERSITY

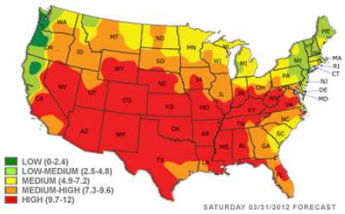
Ordered Dataset

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Sequential Data

Time Series Data



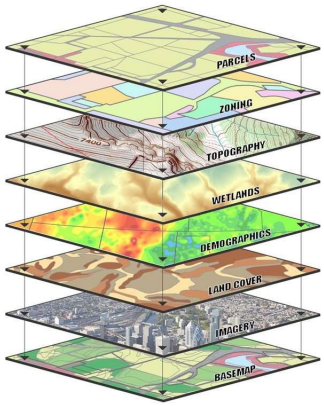
Spatial Data

Sequence Data

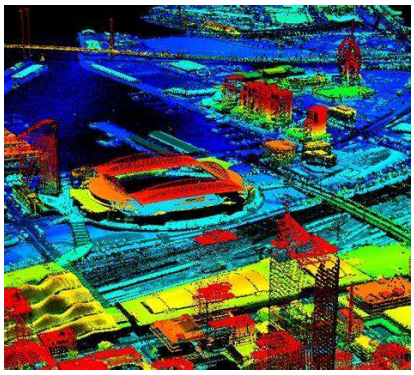
Images adopted from various sources.

SJSU SAN JOSÉ STATE UNIVERSITY

Other Ordered Dataset



GIS Data



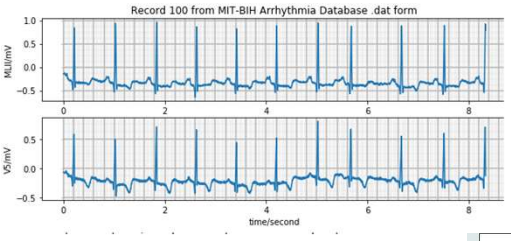
LiDAR Data

Satellite Data

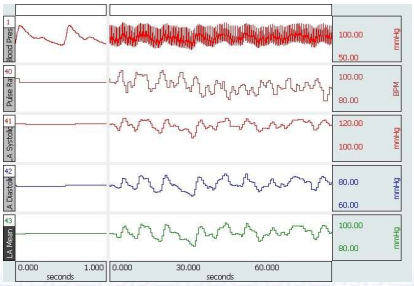
Images adopted from various sources.

SJSU SAN JOSÉ STATE UNIVERSITY

More Ordered Dataset



ECG Data



Blood Pressure Data

Wellness Data

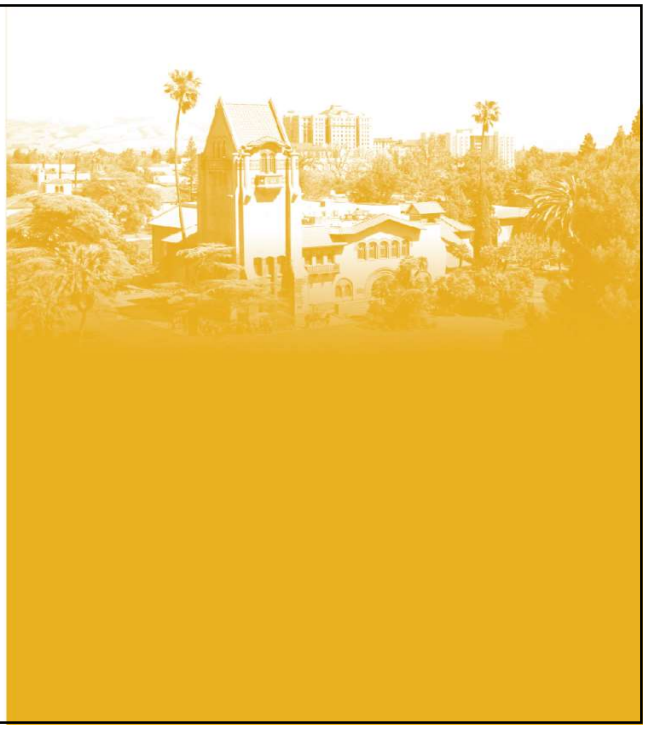
Images adopted from various sources.

13

13

SJSU SAN JOSÉ STATE UNIVERSITY

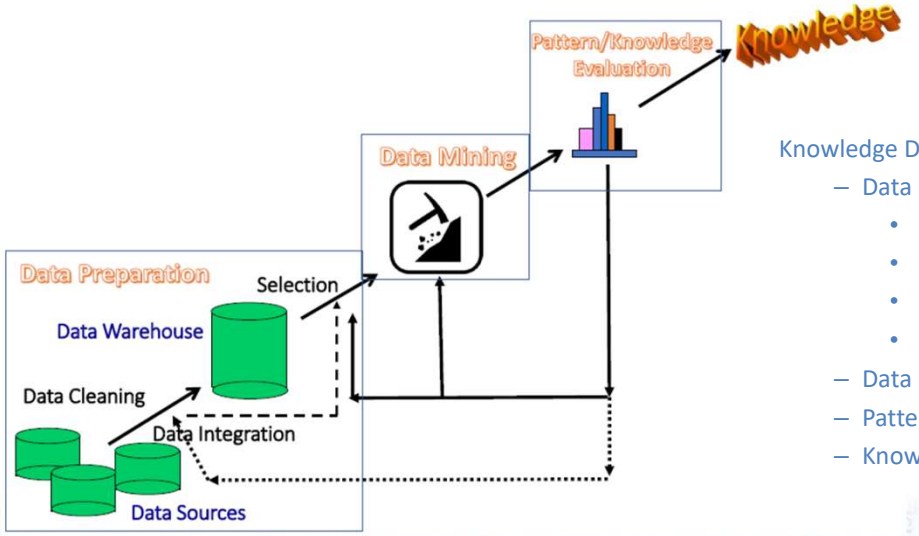
Back to DATA240…



14

SJSU SAN JOSÉ STATE UNIVERSITY

Data Mining: An Essential Step in Traditional Knowledge Discovery



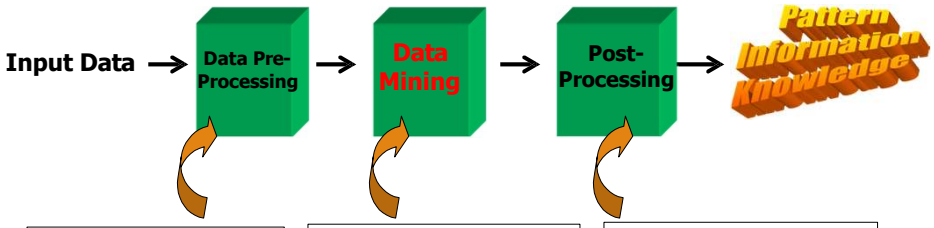
Knowledge Discovery Process
- Data preparation
  - Data cleaning
  - Data integration
  - Data transformation
  - Data selection
- Data mining
- Pattern/model evaluation
- Knowledge presentation

15

15

SJSU SAN JOSÉ STATE UNIVERSITY

KDD Process: A View from Machine Learning



Data Integration	Pattern Discovery	Pattern Evaluation
Normalization	Classification	Pattern Selection
Feature Selection	Clustering	Pattern Interpretation
Dimension Reduction	Outlier Analysis	Pattern Visualization
	… … … …	

16

16

SJSU SAN JOSÉ STATE UNIVERSITY

Data Mining vs. Data Exploration

Which view do you prefer?

- KDD vs. ML vs. BI

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

17

17

SJSU SAN JOSÉ STATE UNIVERSITY

Multi-Dimensional View of Data Mining

- Data to be mined
  - Database data, 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, 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
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

18

18

SJSU SAN JOSÉ STATE UNIVERSITY

Diversity of Data Types for Data Mining

- Structured: uniform, record- or table-like structures, defined by data dictionaries, with a fixed set of attributes, each with a fixed set of value ranges and semantic meaning
  - e.g. data stored in relational databases, data cubes, data matrices, and many data warehouses
- Semi-structured: allow a data object to contain a set value, a small set of heterogeneous typed values, or nested structures, or to allow the structure of objects or sub-objects to be defined flexibly and dynamically
- Data having certain structures with clearly defined semantic meaning, such as transactional data set, sequence data set (e.g., time-series data, gene or protein data, or Weblog data)
  - Graph or network data: A more sophisticated type of semi-structured data set
  - Unstructured data: text data and multimedia (e.g., audio, image, video) data

The real-world data can often be a mixture of structured, semi-structured & unstructured data

19

19

SJSU SAN JOSÉ STATE UNIVERSITY

Diversity of Data Types for Data Mining

- Different applications with different data sets and require different data analysis methods
  - Sequence data: Biological sequences vs. shopping transaction sequences
  - Time-series: ordered set of numerical values with equal time interval
  - Spatial, temporal and spatiotemporal data
  - Graph and network data: Social networks, computer communication networks, biological networks, and information networks may carry rather different semantics
- On the same data set, finding different kinds of patterns: require different mining methods
  - e.g. software programs: finding plagiarized modules vs. finding copy-and-paste bugs
- Stored vs. Streaming data
  - Stored data: Finite, stored in various kinds of large data repositories
  - Streaming data (e.g., video surveillance or remote sensing): Dynamic, constantly coming, infinite, real-time response—posing challenges on effective data mining

20

20

SJSU SAN JOSÉ STATE UNIVERSITY

Mining Various Kinds of Knowledge

- Multidimensional Data Summarization
- Mining Frequent Patterns, Associations, and Correlations
- Classification and Regression for Predictive Analysis
- Cluster Analysis
- Deep Learning
- Outlier Analysis

Are All Mining Results Interesting?

SJSU SAN JOSÉ STATE UNIVERSITY

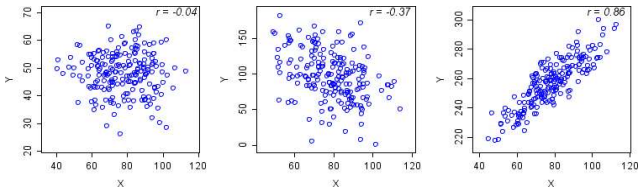
Multidimensional Data Summarization

- 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

SJSU SAN JOSÉ STATE UNIVERSITY

Pattern Discovery: Mining Frequent Patterns, Associations, & Correlations

- Frequent patterns:
  - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis

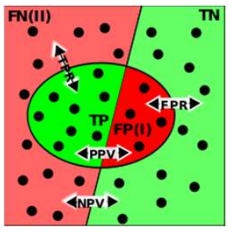


- A typical association rule
  - Diaper ➔ Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

SJSU SAN JOSÉ STATE UNIVERSITY

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 prediction
    - e.g. Classify countries based on (climate)
    - e.g. Classify cars based on (gas mileage)
  - Predict some unknown class labels

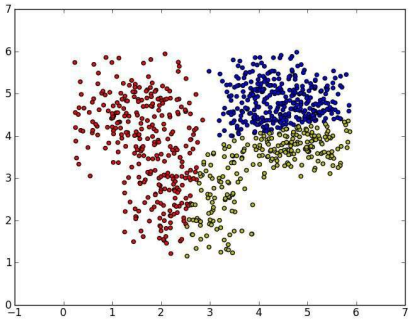


- Typical Methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical Applications
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages,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

SJSU SAN JOSÉ STATE UNIVERSITY

Cluster Analysis

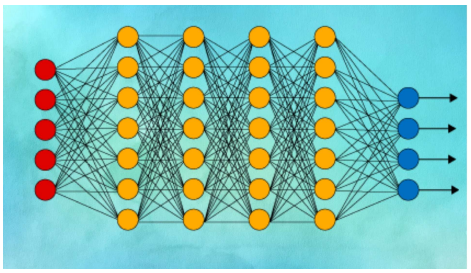
- Unsupervised Learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters)
  e.g. cluster houses to find distribution patterns
- Principle
  - Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications…



SJSU SAN JOSÉ STATE UNIVERSITY

Deep Learning

- Various neural network architectures are available:
  - Feed-forward neural networks
  - Convolutional neural networks
  - Recurrent neural networks
  - Graph neural networks
  - Transformer

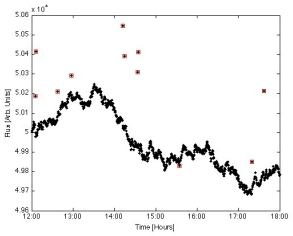


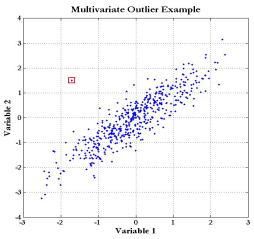
- Broad applications in computer vision, natural language processing, machine translation, social network analysis, and so on
- Reshaping a variety of data mining tasks:
  e.g. classification, clustering, outlier detection, and reinforcement learning

SJSU SAN JOSÉ STATE UNIVERSITY

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





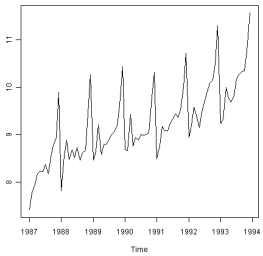
27

27

SJSU SAN JOSÉ STATE UNIVERSITY

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis
    - e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., buy digital camera, then buy large memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

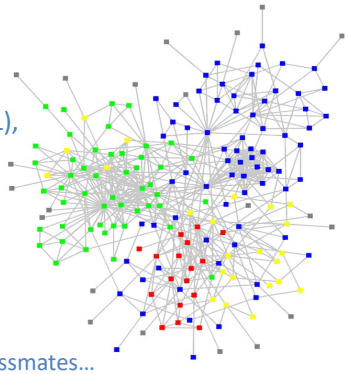


28

28

SJSU SAN JOSÉ STATE UNIVERSITY

Structure and Network Analysis



- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g. author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - e.g. A person could be multiple information networks: friends, family, classmates…
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

29

29

SJSU SAN JOSÉ STATE UNIVERSITY

Evaluation of Knowledge

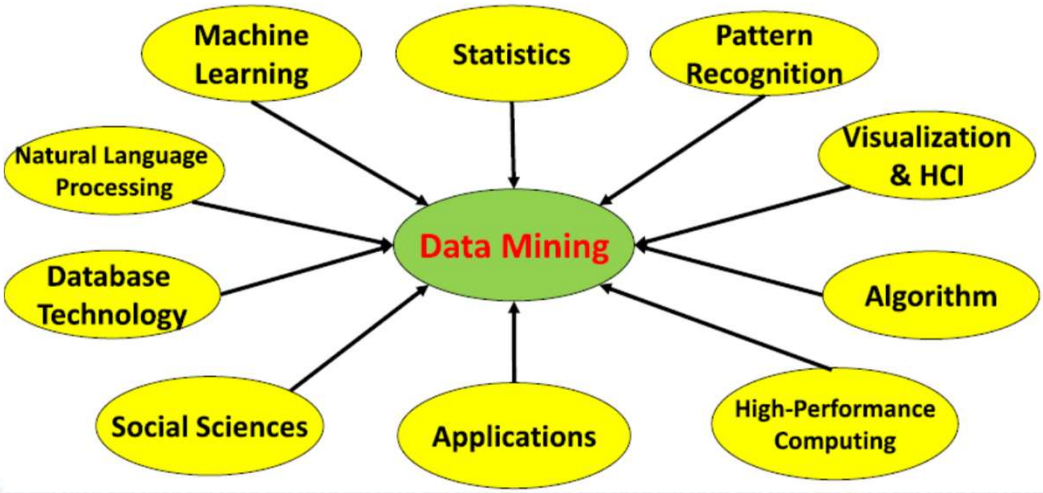
- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns"
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge ➔ directly mining only interesting knowledge?
  - Descriptive vs Predictive
  - Coverage
  - Accuracy
  - Timeliness

30

30

SJSU SAN JOSÉ STATE UNIVERSITY

Data Mining: Confluence of Multiple Disciplines



31

31

SJSU SAN JOSÉ STATE UNIVERSITY

Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be scalable to handle big data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social and information networks
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

32

32

SJSU SAN JOSÉ STATE UNIVERSITY

Applications of Data Mining

- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis
- Data mining and software engineering
- Data mining and text analysis
- Data mining and social and information network analysis
- Built-in (invisible data mining) functions in Google, MS, Yahoo!, Linked, Facebook, …
- Major dedicated data mining systems/tools
  - SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)

33

SJSU SAN JOSÉ STATE UNIVERSITY

Summary

- Data mining: Discovering interesting patterns & knowledge from massive amounts of data.
- Traditional KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Different data mining method on a wide variety of data.
- Data Mining functionalities: summarization, pattern discovery, classification, clustering, deep learning, outlier analysis, trend and outlier analysis, …
- Data Mining is a confluence of multiple disciplines
- Data Mining has broad applications

34