I. Introduction to Data Mining

Nicolás García-Pedrajas

Computational Intelligence and Bioinformatics Research Group

September 15, 2021

Table of contents

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - ✓ Web data, e-commerce
  - ✓ Purchases at department/grocery stores
  - ✓ Bank/Credit Card transactions
  - ✓ Smartphones apps and social networks
  - ✓ Scientific data: Astronomy, Genomics,...
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - ✓ Provide better, customized services for an edge (e.g. in Customer Relationship Management)

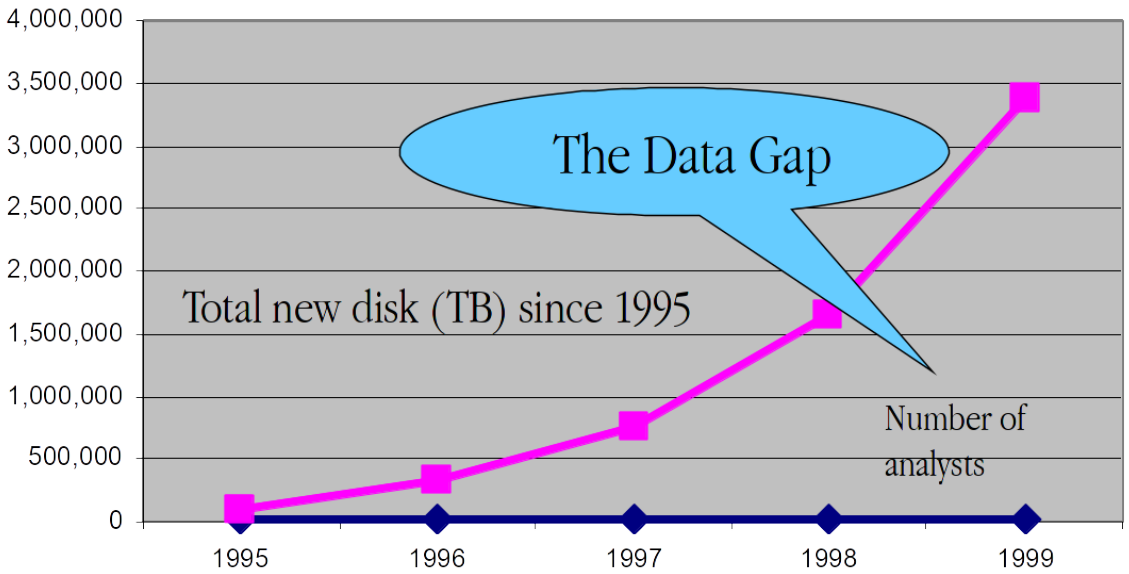
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/second)
  - ✓ remote sensors on a satellite
  - ✓ telescopes scanning the skies
  - ✓ microarrays generating gene expression
  - ✓ scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
  - ✓ Data mining may help scientists in:
    - classifying and segmenting data
    - Hypothesis Formation

Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all

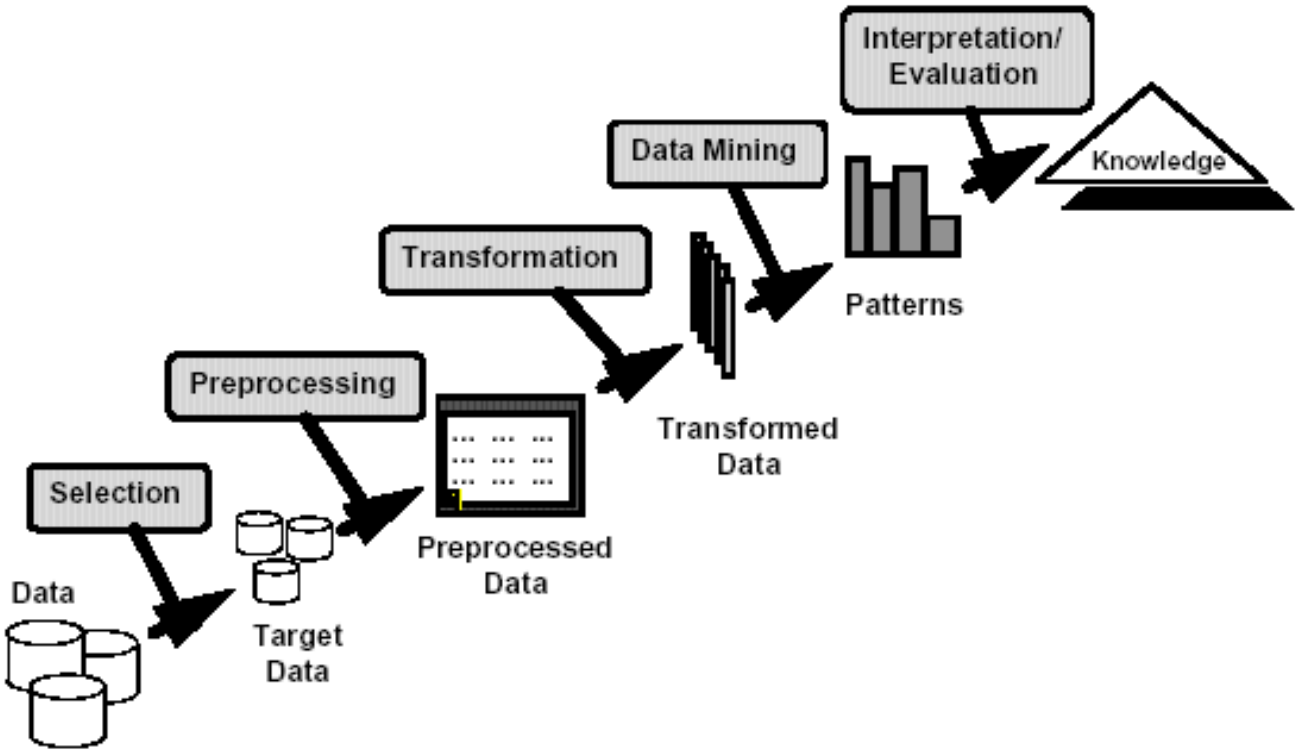
The data gap


4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
1995
1996
1997
1998
1999
The Data Gap
Total new disk (TB) since 1995
Number of analysts

What is Data Mining?

- Many definitions
  - ✓ Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - ✓ Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

What is Data Mining



What is (and is not) Data Mining

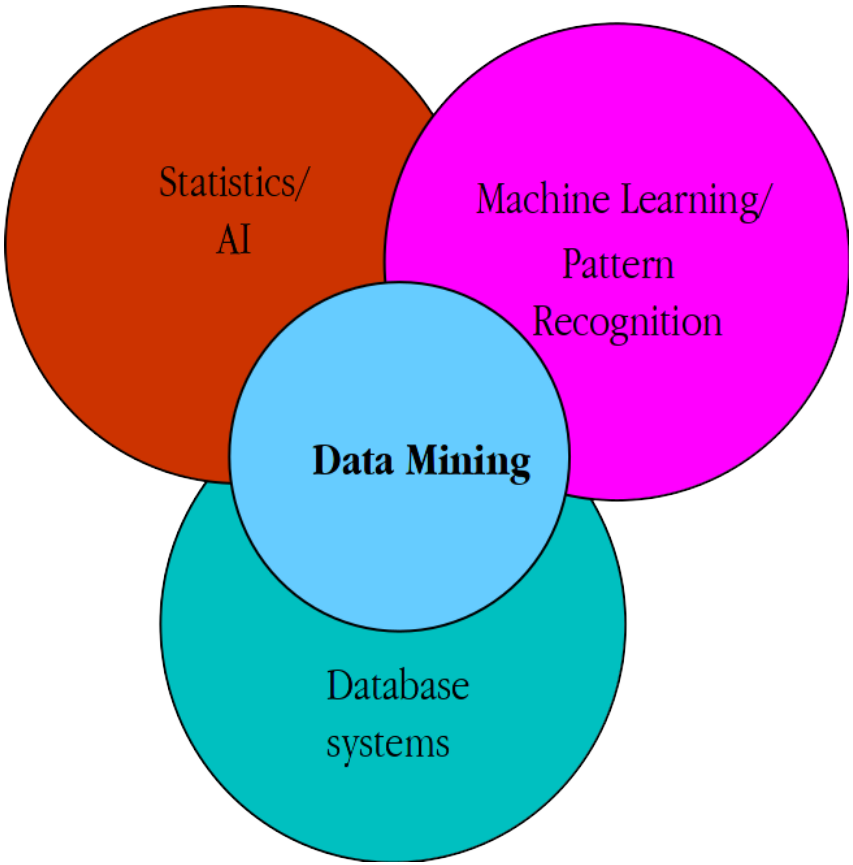
- What **is not** Data Mining?
  - ✓ Look up phone number in phone directory
  - ✓ Query a Web search engine for information about "Amazon"

- What **is** Data Mining?
  - ✓ Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly, etc. in Boston area)
  - ✓ Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - ✓ Enormity of data
  - ✓ High dimensionality of data
  - ✓ Heterogeneous, distributed nature of data

Origins of Data Mining



Confluence of different fields

Origins of Data Mining

- Prediction/predictive Methods
  - ✓ Use some variables to predict unknown or future values of other variables.
    - Classification
- Description/descriptive Methods
  - ✓ Find human-interpretable patterns that describe the data.
    - Clustering, association rules

Data Mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

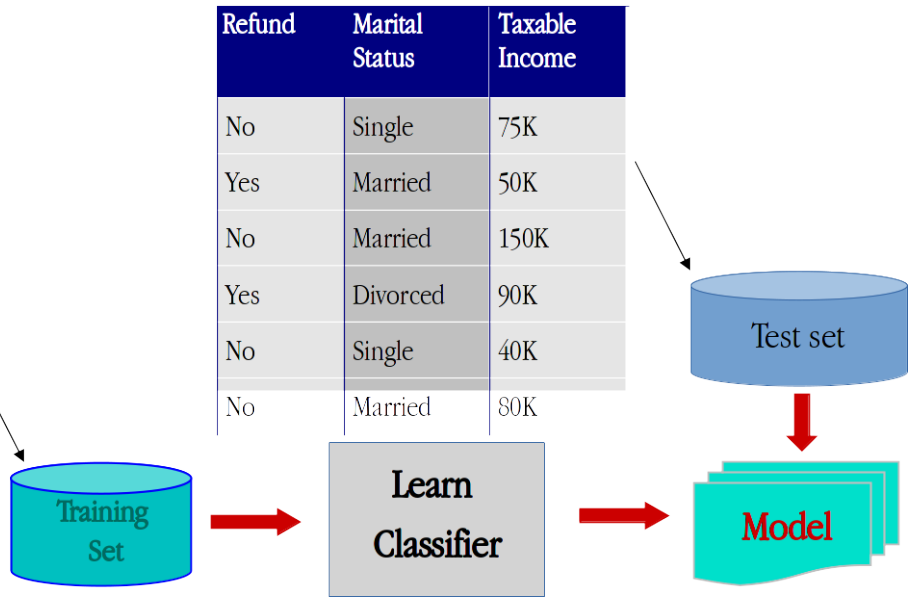
Classification: Definition

- Given a collection of records (training set)
  - ✓ Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - ✓ A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.
- Advanced problems (can be combined)
  - ✓ Each record can belong to more than one class (multi-label)
  - ✓ Each record is represented by more than one set of attributes (multi-instance)
  - ✓ Not all instances are labeled (semi-supervised)

Classification: Example

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

categorical categorical continuous class



Application to Commerce

- Direct Marketing
  - ✓ Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.
  - ✓ Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This buy, don't buy decision forms the class attribute.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - → Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Application to Commerce

- Fraud Detection
  - ✓ Goal: Predict fraudulent cases in credit card transactions.
  - ✓ Approach:
    - Use credit card transactions and the information on its account-holder as attributes.

      → When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

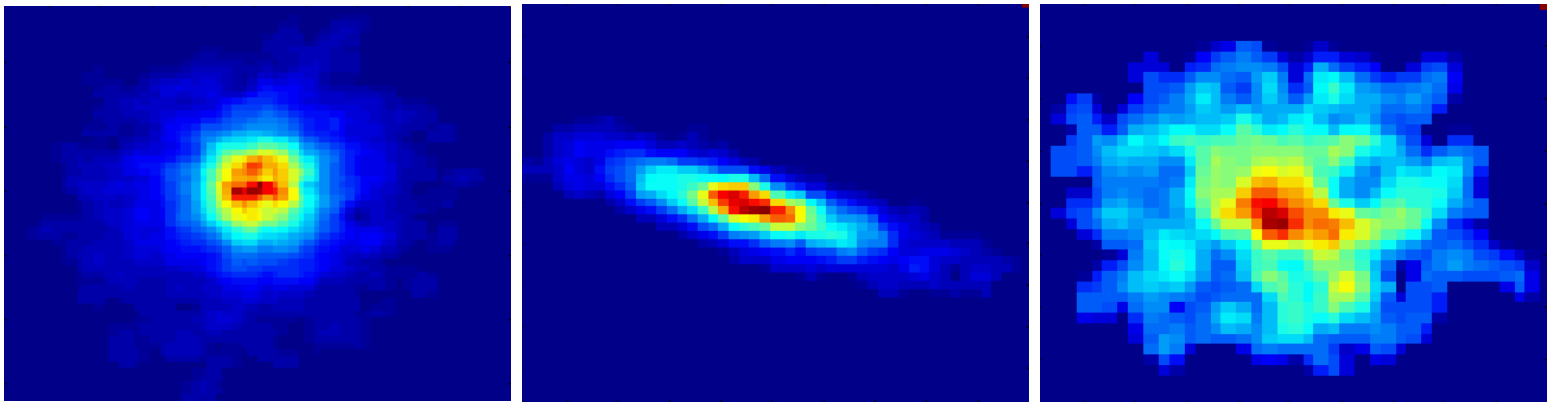
Application to Astronomy

- Sky Survey Cataloging
  - ✓ Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - ✓ Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Application to Astronomy: Classifying galaxies

- Class: Stage of formation
- Attributes: Image features, characteristics of light waves received, etc.
- Data size:
  - ✓ 72 million stars
  - ✓ 20 million galaxies
  - ✓ Object Catalog: 9 GB
  - ✓ Image Database: 150 GB

Application to Astronomy: Classifying galaxies



(a) Early (b) Intermediate (c) Late

Figure: Stages of formation

Clustering: Definition

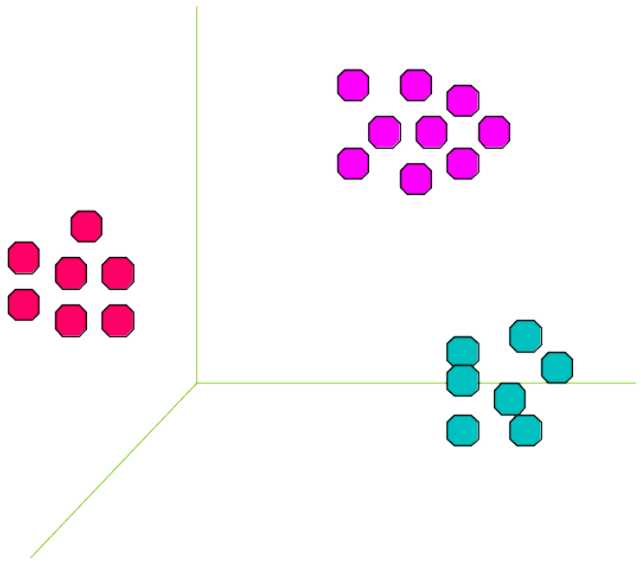
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - ✓ Data points in one cluster are more similar to one another.
  - ✓ Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - ✓ Euclidean Distance if attributes are continuous.
  - ✓ Other Problem-specific Measures.

Illustrating clustering

Euclidean Distance Based Clustering in 3-D space

Intracluster distances are minimized

Intercluster distances are maximized



Application to Commerce

- Market Segmentation:
  - ✓ Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - ✓ Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Application to Information Retrieval

- Document Clustering:
  - ✓ Goal:
    - To find groups of documents that are similar to each other based on the important terms appearing in them.
  - ✓ Approach:
    - To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - ✓ Gain:
    - Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Application to Information Retrieval

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Application to Stock Markets

- Clustering of S&P 500 Stock Data
  - ✓ Observe Stock Movements every day.
  - ✓ Clustering points: Stock-{UP/DOWN}
  - ✓ Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - ✓ We used association rules to quantify a similarity measure.

Application to Stock Markets

Identified four clusters

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

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	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered
{Milk} —> {Coke}
{Diaper, Milk} —> {Beer}

Application to Marketing

- Marketing and Sales Promotion:
  - ✓ Let the rule discovered be
    - {Bagels, . . .} → {Potato Chips}
  - ✓ Potato Chips as consequent
    - Can be used to determine what should be done to boost its sales.
  - ✓ Bagels in the antecedent
    - Can be used to see which products would be affected if the store discontinues selling bagels.
  - ✓ Bagels in antecedent and Potato chips in consequent
    - Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Application to Sales

- Supermarket shelf management.
  - ✓ Goal: To identify items that are bought together by sufficiently many customers.
  - ✓ Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - ✓ A classic rule:
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Application to Sales

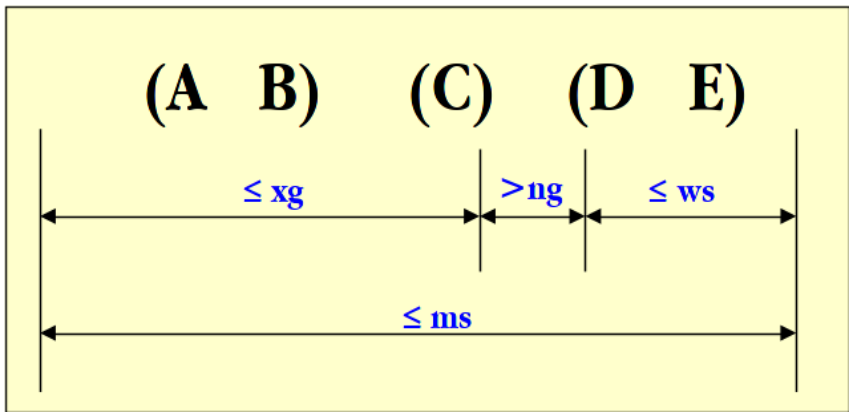
- Inventory Management:
  - ✓ Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - ✓ Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential pattern discovery: Definition

Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.

$$(AB)(C) \longrightarrow (DE) \tag{1}$$

Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential pattern discovery: Examples

- In telecommunications alarm logs
  - ✓ (Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) ⟶ (Fire_Alarm)
- In point-of-sale transaction sequences
  - ✓ Computer Bookstore:
    - (Intro_To_Visual_C) (C++_Primer) ⟶ (Perl_for_dummies,Tcl_Tk)
  - ✓ Athletic Apparel Store:
    - (Shoes) (Racket, Racketball) ⟶ (Sports_Jacket)

Regression: Definition

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - ✓ Predicting sales amounts of new product based on advetising expenditure.
  - ✓ Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - ✓ Time series prediction of stock market indices.

Anomaly/Deviation detection

- Detect significant deviations from normal behavior
- Applications:
  - ✓ Credit Card Fraud Detection
  - ✓ Network Intrusion Detection
- Can be also seen as a classification problem with extreme imbalance

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Scalability

- Data of megabytes, terabytes or even petabyes are not uncommon
- Algorithms must be scalable
  - ✓ Time constraints
  - ✓ Storage constraints

Dimensionality

- Thousands or hundreds of thousands of features are common: Bioinformatics, Images, Streams, etc.
- Data mining algorithms were designed for low dimensionality data.
- Large numbers of attributes affect the algorithm
  - ✓ Complexity of the algorithm: Scalability issues
  - ✓ Poor performance

Complex and Heterogeneous Data

- We have complex data from multiple sources and multiple relations
  - ✓ Graphs, streams, images, natural language, etc.
- We must obtain complex knowledge with complex relationships
- The objects are of different types

Data Quality

- Large amounts of data usually also means poor quality
  - ✓ Noise
  - ✓ Duplicated entries
  - ✓ Wrong labeling
- Low quality data can only produce low quality knowledge

Data Ownership and Distribution

- Many Data Mining tasks are performed in a network
- Data from different sources and owners
- Security issues
- Speed issues
- Consistency issues

Privacy Preservation

- Sensitive data might compromise privacy
- Anonymized data is not enough

Streaming Data

- Streaming data is complex and difficult to deal with
- High dimensionality
- Fast response
- Large amounts of data