Data Mining Methods Part 1

MBI304- Data Mining & Data Analytics

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

- Data mining: In brief Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data.
- Within these masses of data lies hidden information of strategic importance. But when there are so
- many trees, how do you draw meaningful conclusions about the forest?

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

describe the data

The first and simplest analytical step in data mining is to describe the data — summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables

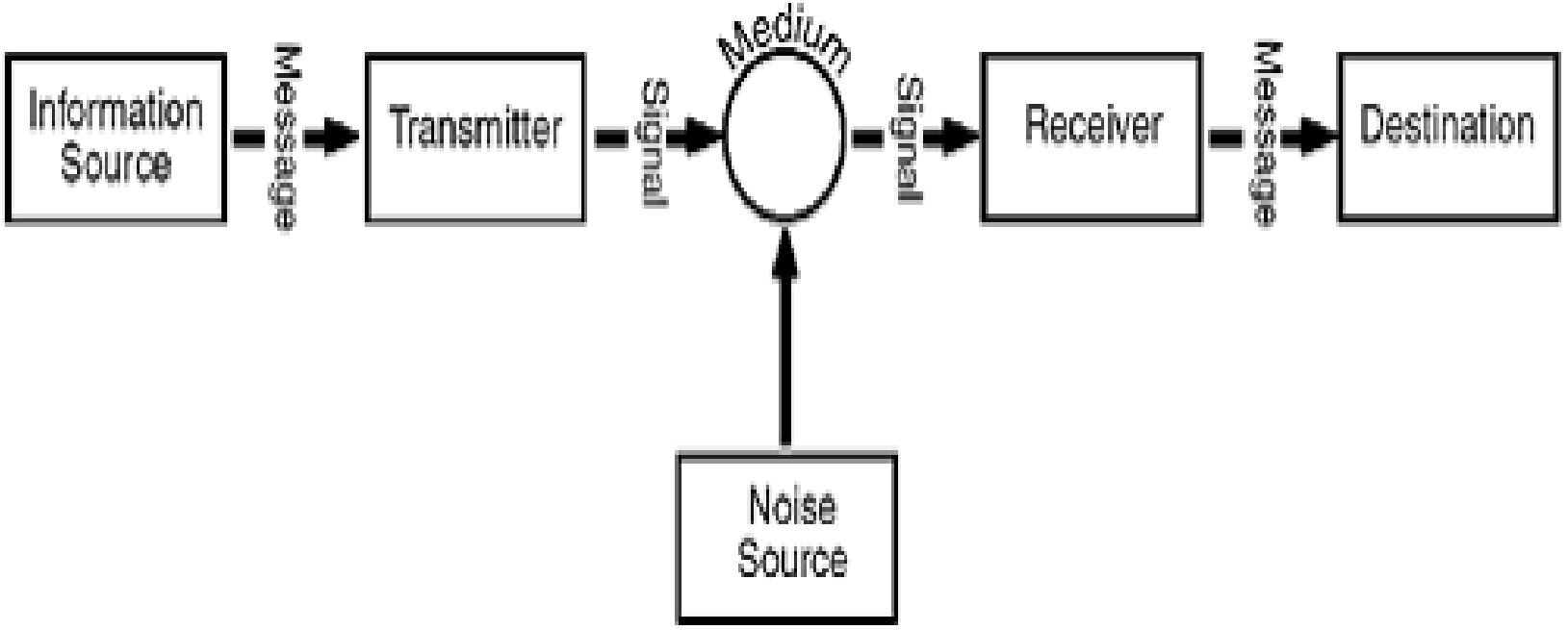
- The final step is to empirically verify the model. For example, from a database of customers who have already responded to a particular offer, you've built a model predicting which prospects are likeliest to respond to the same offer.
- Can you rely on this prediction? Send a mailing to a portion of the new list and see what results you get

- But data description alone cannot provide an action plan.
- You must build a predictive model based on patterns determined from known results, then test that model on results outside the original sample.
- A good model should never be confused with reality (you know a road map isn't a perfect
- representation of the actual road), but it can be a useful guide to understanding your business.

the knowledge-discovery process involves:

1. Selection and sampling of the appropriate data from the database(s)
2. Preprocessing and cleaning of the data to remove redundancies, errors, and conflicts
3. Transforming and reducing data to a format more suitable for the data mining
4. Data mining
5. Evaluation of the mined data
6. Visualization of the evaluation results
7. Designing new data queries to test new hypotheses and returning to step 1

Figure 1-4. Information Theory. Shannon's model of a communications system includes five components: an information source, a transmitter, the medium, a receiver, and a destination. The amount of information that can be transferred from information source to destination is a function of the strength of the signal relative to that of the noise generated by the noise source.



- Where is the knowledge we have lost in information?
- Where is the wisdom we have lost in knowledge?
- —T.S. Elliot, "The Rock"

Figure 1-9. Information Theory and the Central Dogma. Information theory applies equally to the replication, transcription, translation, and the overall process of converting nucleotide sequences in DNA to protein.

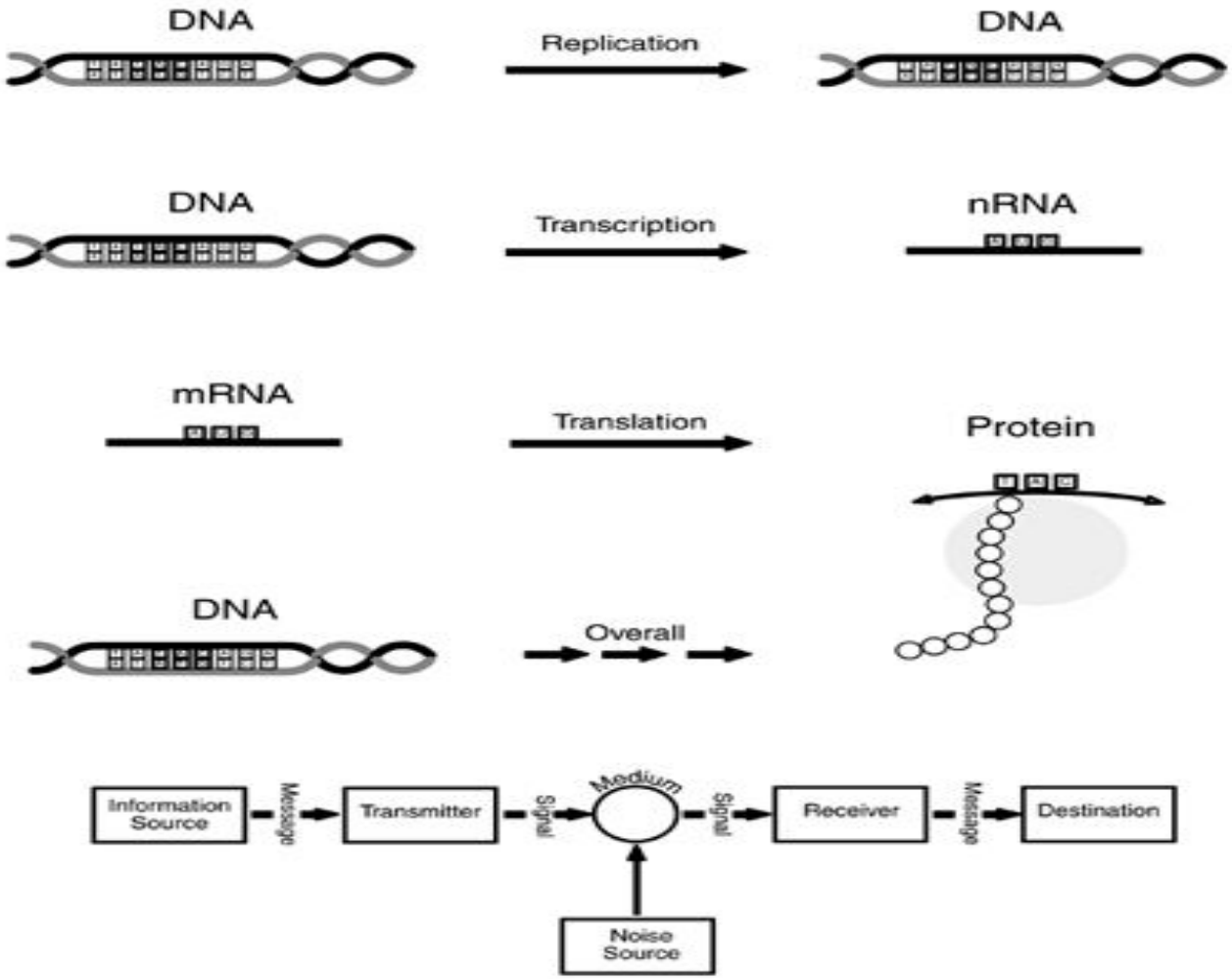
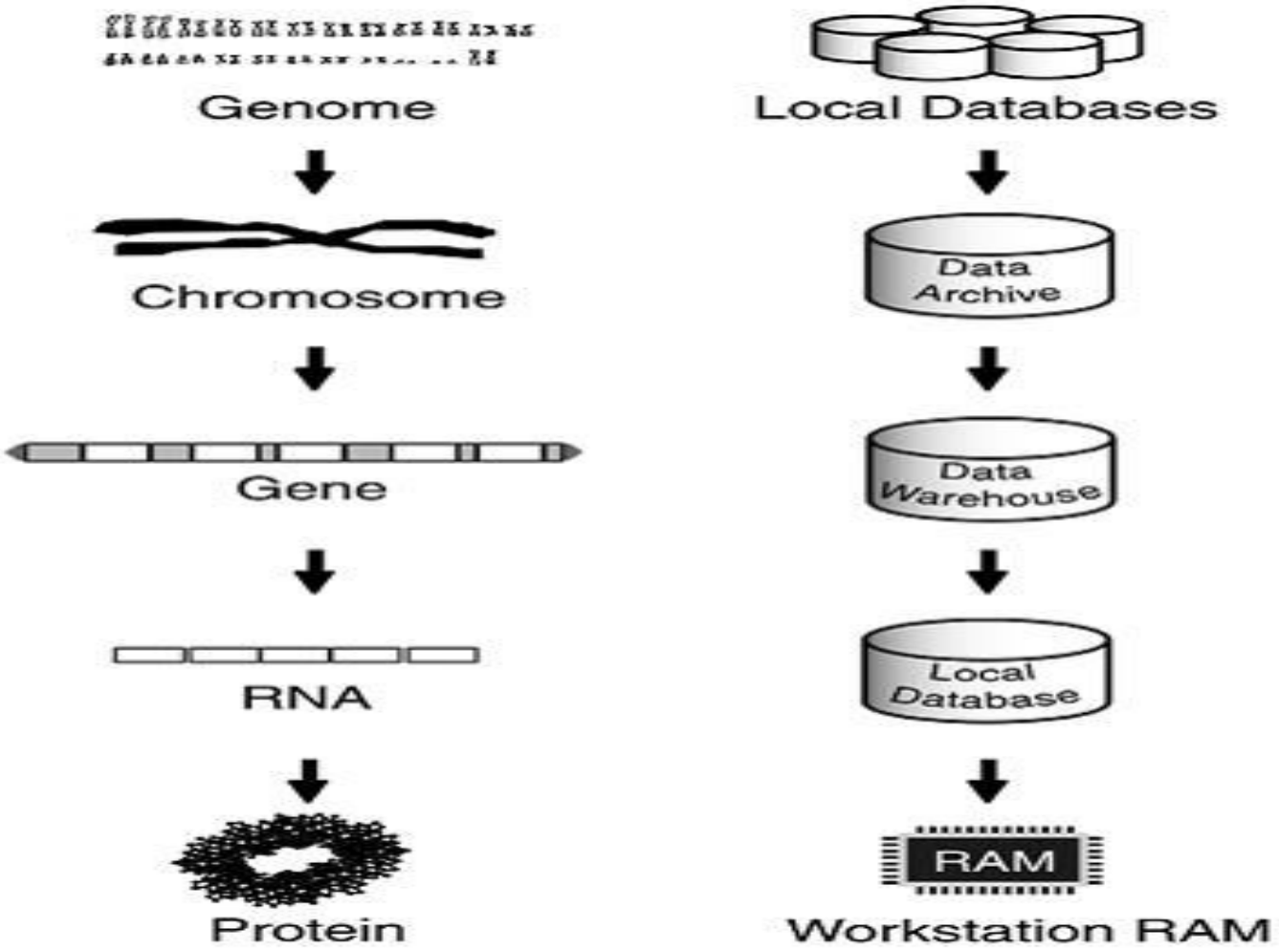


Figure 2-11. Organic Analog of Database Hierarchy. The database hierarchy has many parallels to the hierarchy in the human genome. Data stored in chromosomes, like a data archive, must be unpacked and transferred to a more immediately useful form before the data can be put to use.



- Recently, the collection of biological data has been increasing at explosive rates

  due to improvements of existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large scale experiments and research programs.
- An important example is the Human Genome Project, that was founded in 1990 by the U.S. Department of Energy and the U.S.National Institutes of Health (NIH) and was completed in 2003 (U.S. Department of Energy Office of Science, 2004).

A representative example of the rapid biological data accumulation is the exponential growth of GenBank (Figure 1), the U.S. NIH genetic sequence database. (National Center for Biotechnology Information, 2004).

The explosive growth in the amount of biological data demands the use of computers for the organization, the maintenance and the analysis of these data.

- This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology.
- As Luscombe et al. (2001) mention, the aims of bioinformatics are: The organization of data in such a way that allows researchers to access existing information and to submit new entries as they are produced.
- The development of tools that help in the analysis of data. The use of these tools to analyze the individual systems in detail, in order to gain new biological insights.

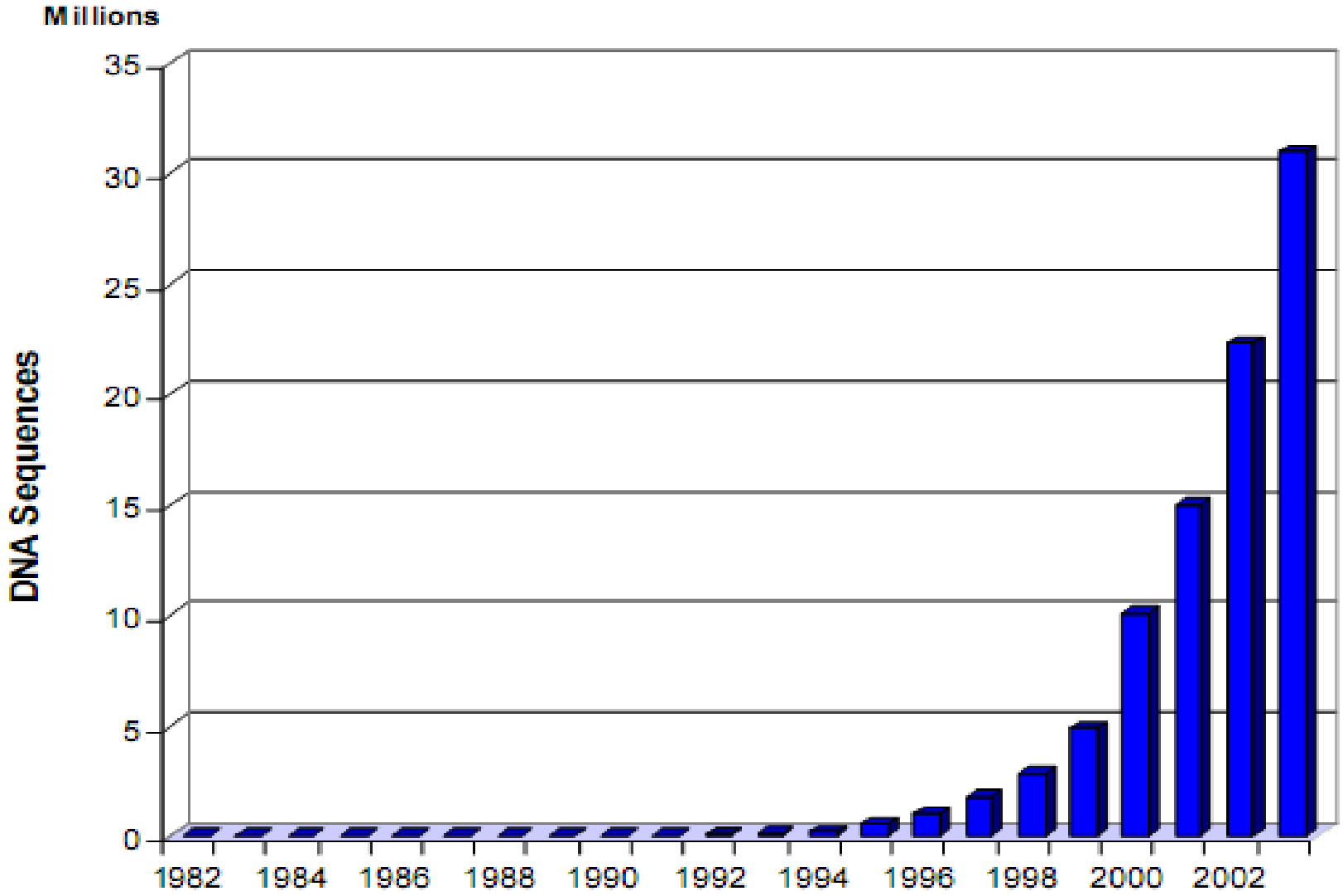


Figure 1: Growth of GenBank (Years 1982-2003).

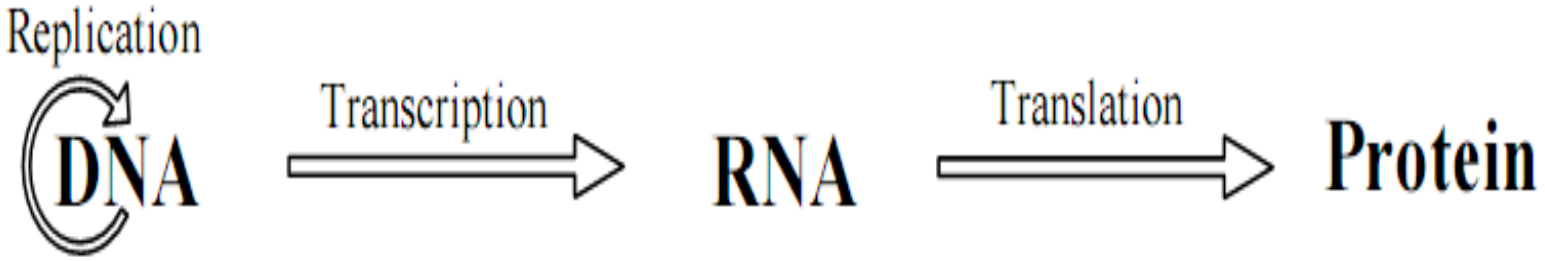


Figure 2: The Central Dogma of Molecular Biology (Initial Statement).

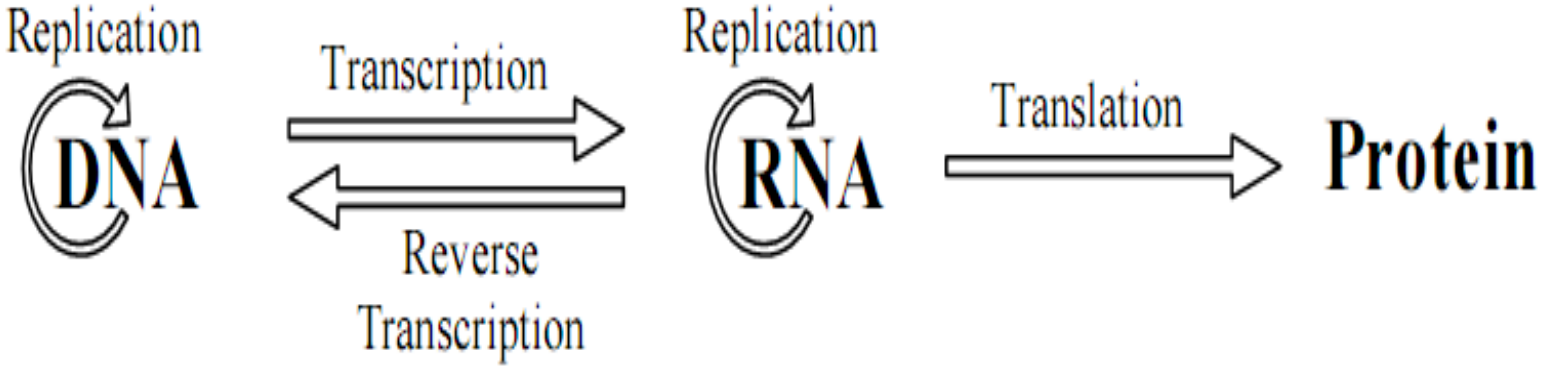


Figure 3: The Central Dogma of Molecular Biology (Extended Statement).

Houle et al. (2000) refer to a classification of three successive levels for the analysis of biological data, that is identified on the basis of the central dogma of
molecular biology:

- Genomics is the study of an organism's genome and deals with the systematic use of genome information to provide new biological knowledge.
- Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene's coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription (Houle et al., 2000).
- Proteomics is the large-scale study of proteins, particularly their structures andfunctions. (Wikipedia).

MINING BIOLOGICAL DATA

- Data mining is the discovery of useful knowledge from databases. It is the main step in the process known as Knowledge Discovery in Databases (KDD) (Fayyad et al.,1996), although the two terms are often used interchangeably.
- Other steps of the KDD process are the collection, selection, and transformation of the data and the visualization and evaluation of the extracted knowledge. Data mining employs

- Data mining employs algorithms and techniques from statistics, machine learning, artificial intelligence, databases and data warehousing etc.
- Some of the most popular tasks are classification, clustering, association and sequence analysis, and regression.
- Depending on the nature of the data as well as the desired knowledge there is a large number of algorithms for each task. All of these algorithms try to fit a model to the data (Dunham, 2002). Such a model can be either predictive or descriptive.

- A predictive model makes a prediction about data using known examples,
- while a descriptive model identifies patterns or relationships in data. Table 3 presents the most common data mining tasks (Dunham, 2002).

Predictive	Descriptive
Classification. Maps data into predefined classes. **Regression.** Maps data into a real valued prediction variable.	**Association Analysis.** The production of rules that describe relationships among data. **Sequence Analysis.** Same as association, but sequence of events is also considered. **Clustering.** Groups similar input patterns together.

Table 3: Common Data Mining Tasks.

Clustering

- Clustering divides a database into different groups. The goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. Unlike classification (see Predictive Data Mining, below), you don’t know what the clusters will be when you start, or by which attributes the data will be clustered.

- Often it is necessary to modify the clustering by excluding variables that have been employed to group instances, because upon examination the user identifies them as irrelevant or not meaningful. After you have found clusters
- that reasonably segment your database, these clusters may then be used to classify new data. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means.

- Don’t confuse clustering with segmentation. Segmentation refers to the general problem of identifying groups that have common characteristics.
- Clustering is a way to segment data into groups that are not previously defined, whereas classification is a way to segment data by assigning it to groups that are already defined.

Link analysis

- Link analysis is a descriptive approach to exploring data that can help to identify relationships among values in a database.
- The two most common approaches to link analysis are association discovery and sequence discovery.
- Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery.
- Sequence discovery is very similar, in that a sequence is an association related over time.

- Some terminology In predictive models, the values or classes we are predicting are called the response, dependent or target variables. The values used to make the prediction are called the **predictor or independent variables.** Predictive models are built, or trained, using data for which the value of the response variable is already known.
- This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. (By contrast, descriptive techniques such as clustering, described in the previous section, are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms.)

Classification

- Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave.
- Data mining creates classification models by examining already classified data (cases) and
- inductively finding a predictive pattern

Regression

- Regression uses existing values to forecast what other values will be.
- In the simplest case, regression uses standard statistical techniques such as linear regression.
- Unfortunately, many real-world problems are not simply linear projections of previous values.

- Therefore, more complex techniques (e.g., logistic
  regression, decision trees, or neural nets) may be necessary to forecast future values.

- The same model types can often be used for both regression and classification. For example, the CART (Classification And Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables).
- Neural nets too can create both classification and regression models.

Time series forecasting

- Time series forecasting predicts unknownfuture values based on a time-varying series of predictors.
- Like regression, it uses known results to guide its predictions.
- Models must take into account the distinctive properties of time, especially the hierarchy of periods (including such varied definitions as the five- or seven-day work week, the thirteen-“month” year, etc.), seasonality, calendar effects such
- as holidays, date arithmetic, and special considerations such as how much of the past is relevant.

- Many general data mining systems such as SAS Enterprise Miner, SPSS, S-Plus, IBM Intelligent Miner, Microsoft SQL Server 2000, SGI MineSet, and Inxight VizServer can be used for biological data mining.
- However, some biological data mining tools such as GeneSpring, Spot Fire, VectorNTI, COMPASS, Statistics for Microarray Analysis, and Affymetrix Data Mining Tool have been developed (Han, 2002).
- Also, a large number of biological data mining tools is provided by National Center for Biotechnology Information and by European Bioinformatics Institute.

Data Mining in Genomics

- Many data mining techniques have been proposed to deal with the identification of specific DNA sequences.
- The most common include neural networks, Bayesian classifiers, decision trees, and Support Vector Machines (SVMs) (Ma & Wang, 1999; Hirsh & Noordewier, 1994; Zien et al., 2000).
- Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives) (Houle et al., 2000).

- However, as Li et al. (2003) state, traditional data mining techniques cannot be directly applied to this type of recognition problems.
- Thus, there is the need to adapt the existing techniques to this kind of problems. Attempts to overcome this problem have been made using feature generation and feature selection (Zeng & Yap, 2002; Li et al., 2003).
- Another data mining application in genomic level is the use of clustering algorithms to group structurally related DNA sequences.

One-way Clustering	Two-way Clustering	Classification
Hierarchical Clustering	Block Clustering	SVMs
Self-organizing Maps (SOMs)	Gene Shaving	K-nearest Neighbors
K-means	Plaid Models	Classification/Decision Trees
Singular Value Decomposition (SVD)		Voted Classification
		Weighted Gene Voting
		Bayesian Classification

Table 4: Popular microarray data mining methods

- Data mining has been applied for the protein secondary structure prediction. This problem has been studied for over than 30 years and many techniques have been developed (Whishart, 2002).
- Initially, statistical approaches were adopted to deal with his problem. Later, more accurate techniques based on information theory, Bayes theory, nearest neighbors, and neural networks were developed.
- Combined methods such as integrated multiple sequence alignments with neural network or nearest neighbor approaches improve prediction accuracy.

GDBSCAN

- A density based clustering algorithm (GDBSCAN) is presented by Sander et al.

  (1998), that can be used to deal with protein interactions.
- This algorithm is able to cluster point and spatial objects according to both, their spatial and non-spatial attributes.

Figure 7-1. Data Mining. Data mining operations are shown here in the context of a larger knowledge-discovery process.

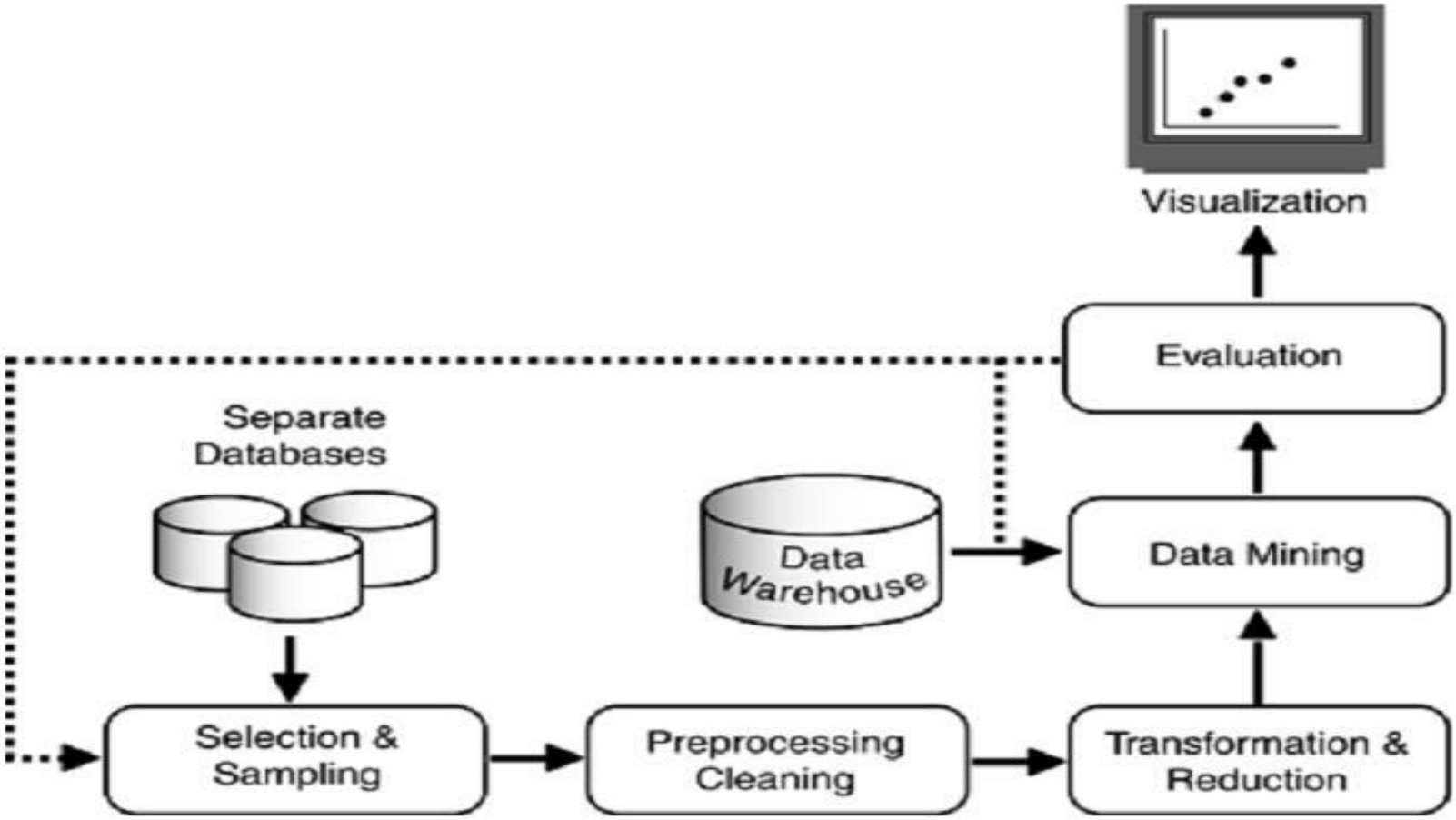


Figure 7-2. Data Mining Methods. Classification—Mapping to a class or group. Regression—Statistical analysis. Link Analysis—Correlation of data. Deviation Detection—Difference from the norm. Segmentation—Similarity function.

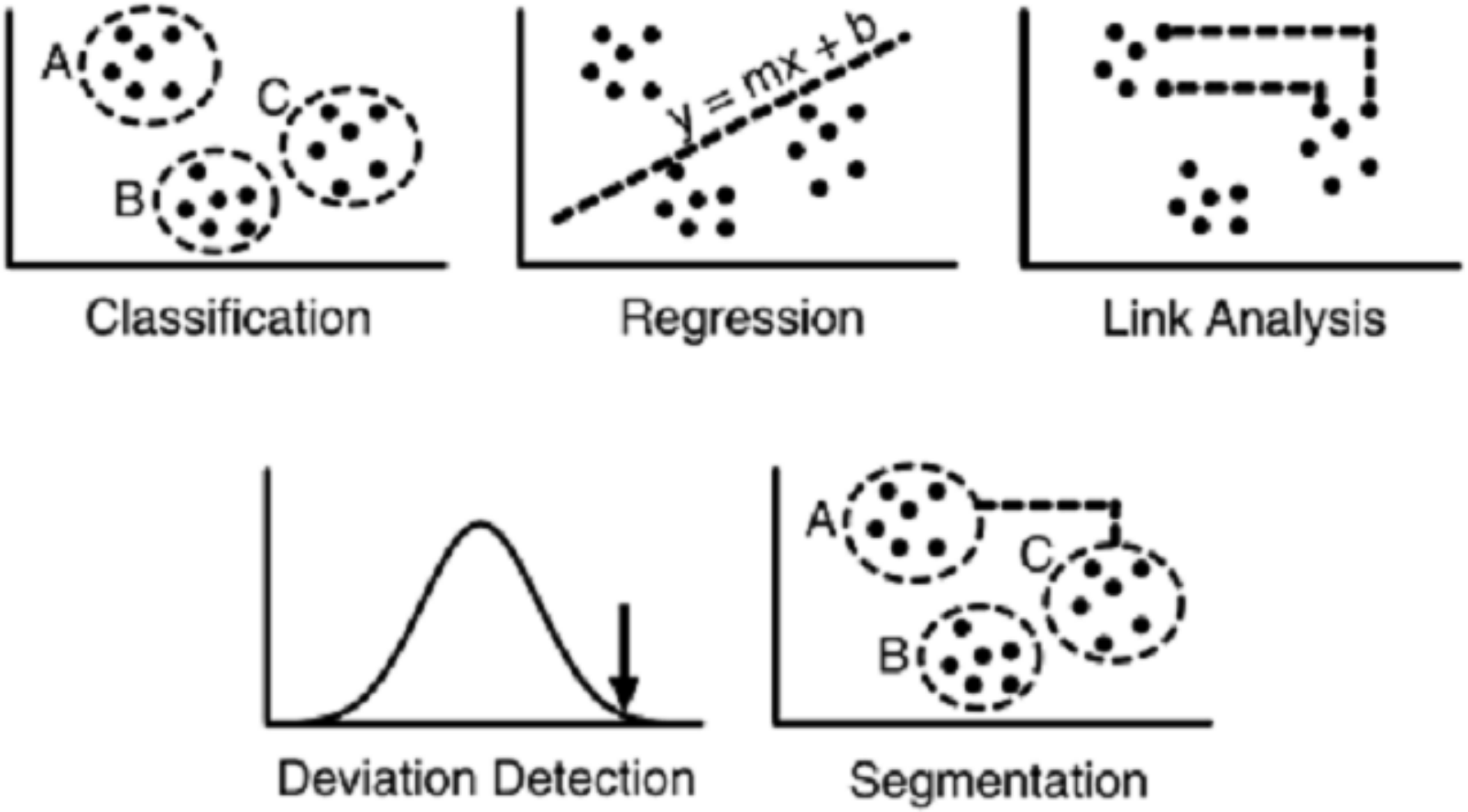
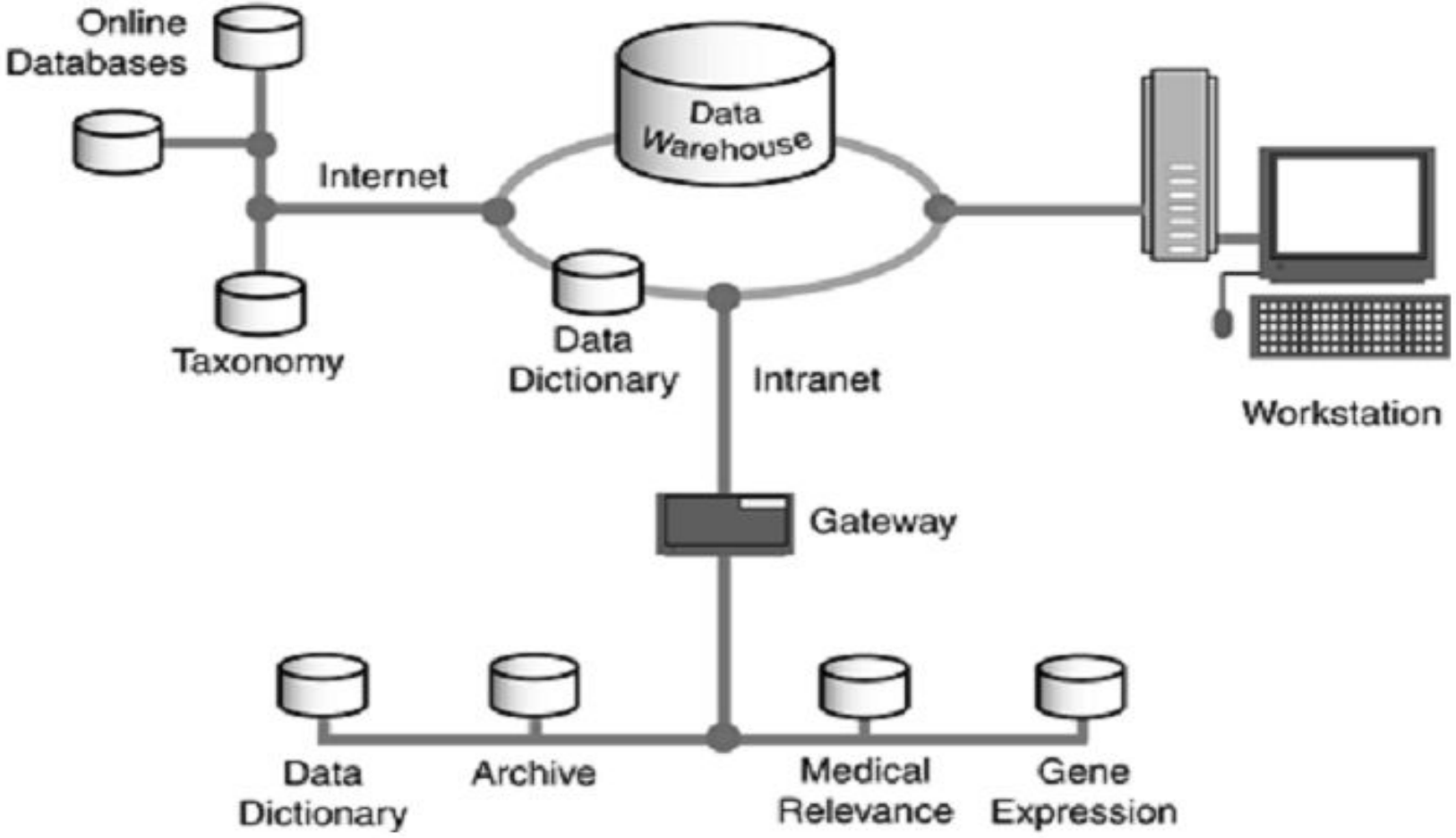


Figure 7-3. Centralized Data-Mining Infrastructure. In this example, a data warehouse, data dictionary, high-bandwidth access to data on the Internet, and a high-performance workstation form the basis for an effective data-mining operation.



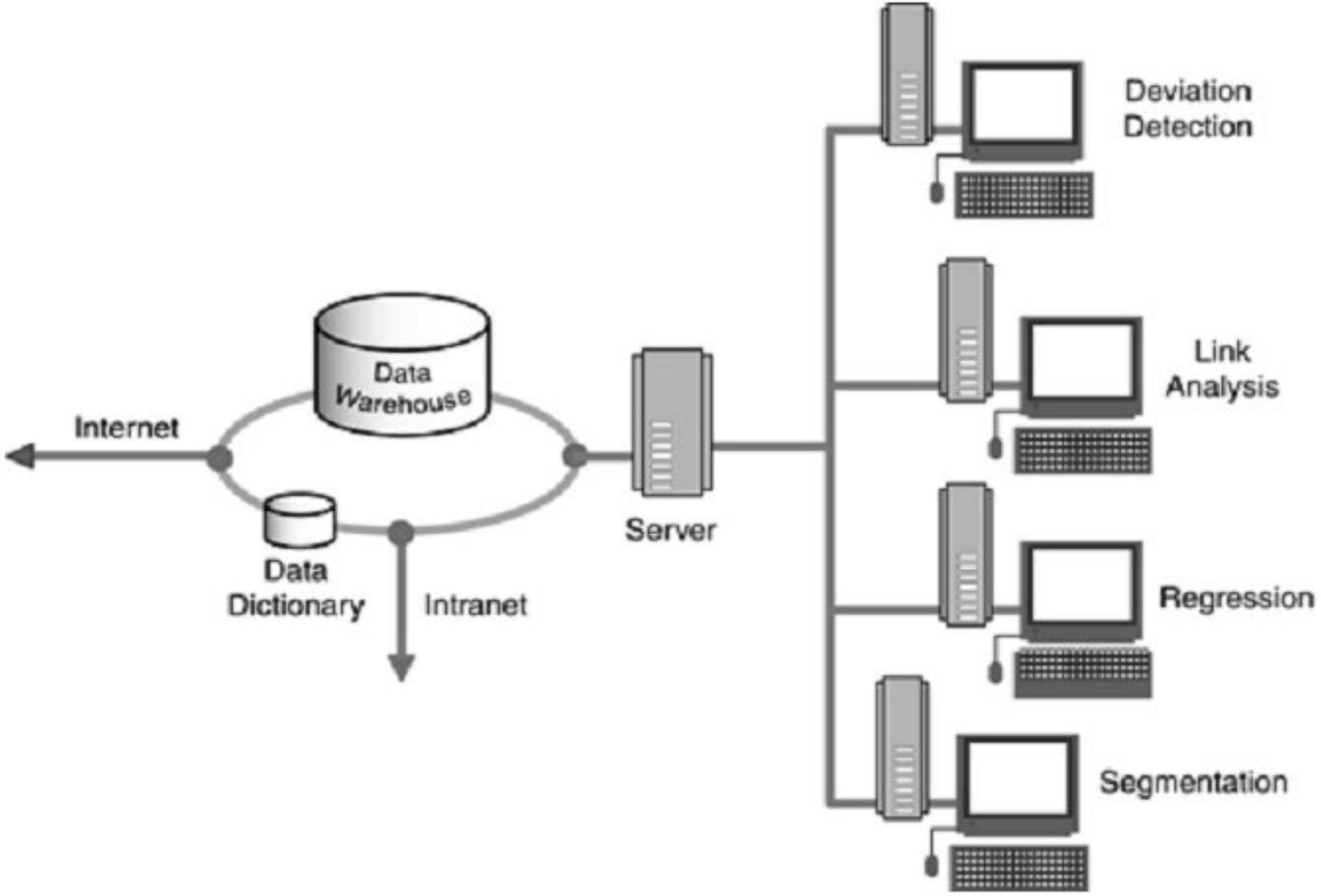

Deviation
Detection
Link
Analysis
Regression
Segmentation
Data
Warehouse
Internet
Server
Data
Dictionary
Intranet

Figure 7-5. The Pattern-Recognition and Discovery Process. Pattern discovery differs from pattern recognition in that feature selection is determined empirically under program control.

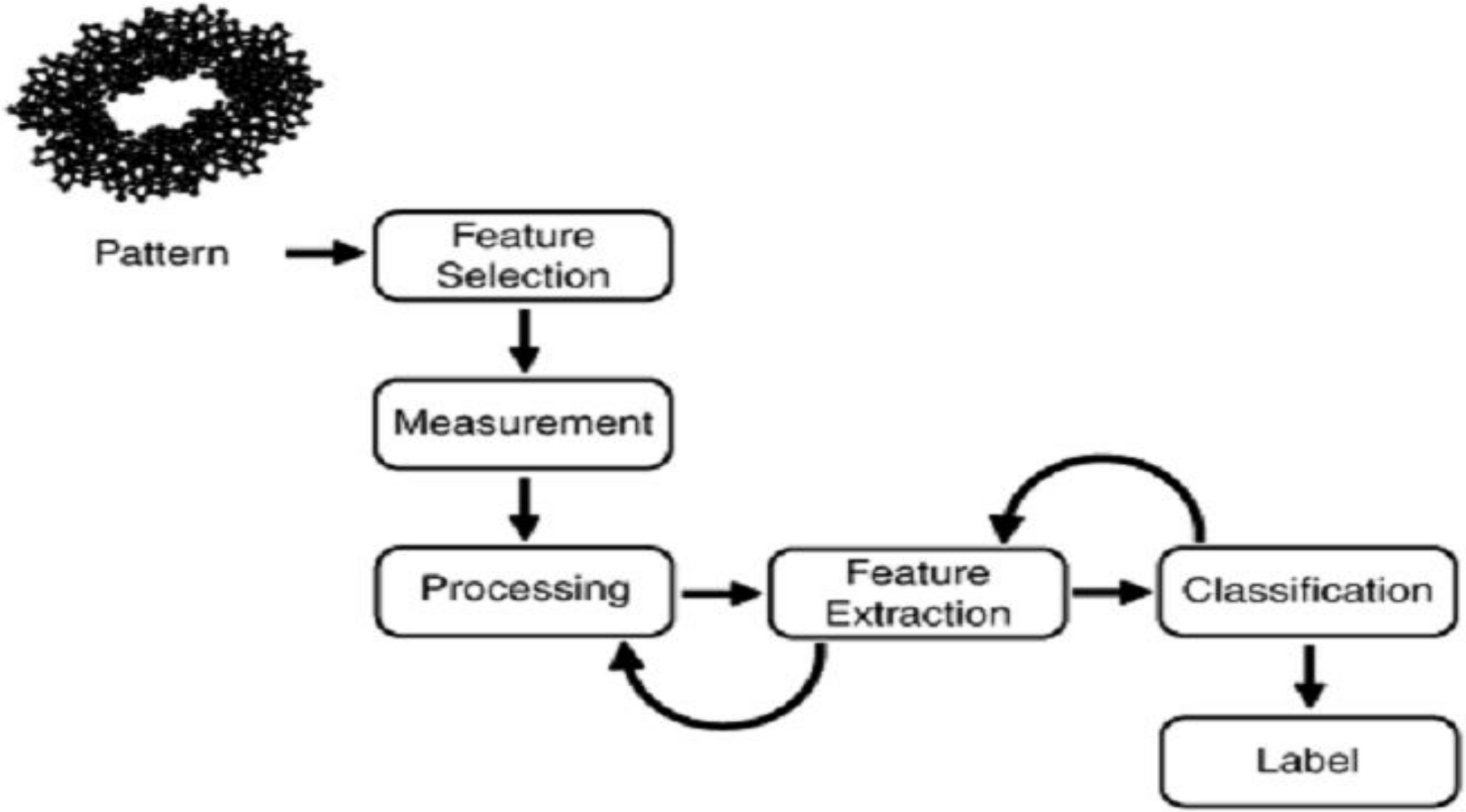


Table 7-3. Machine Learning Technologies and Their Applicability to Data-Mining Methods.

Machine Learning Technologies	Data-Mining Methods				
	Classification	Regression	Segmentation	Link Analysis	Deviation Detection
Inductive Logic Programming	X	X			
Genetic Algorithms	X	X	X		
Neural Networks	X	X	X		
Statistical Methods	X	X	X	X	X
Decision Trees	X		X		
Hidden Markov Models	X				

The machine learning process

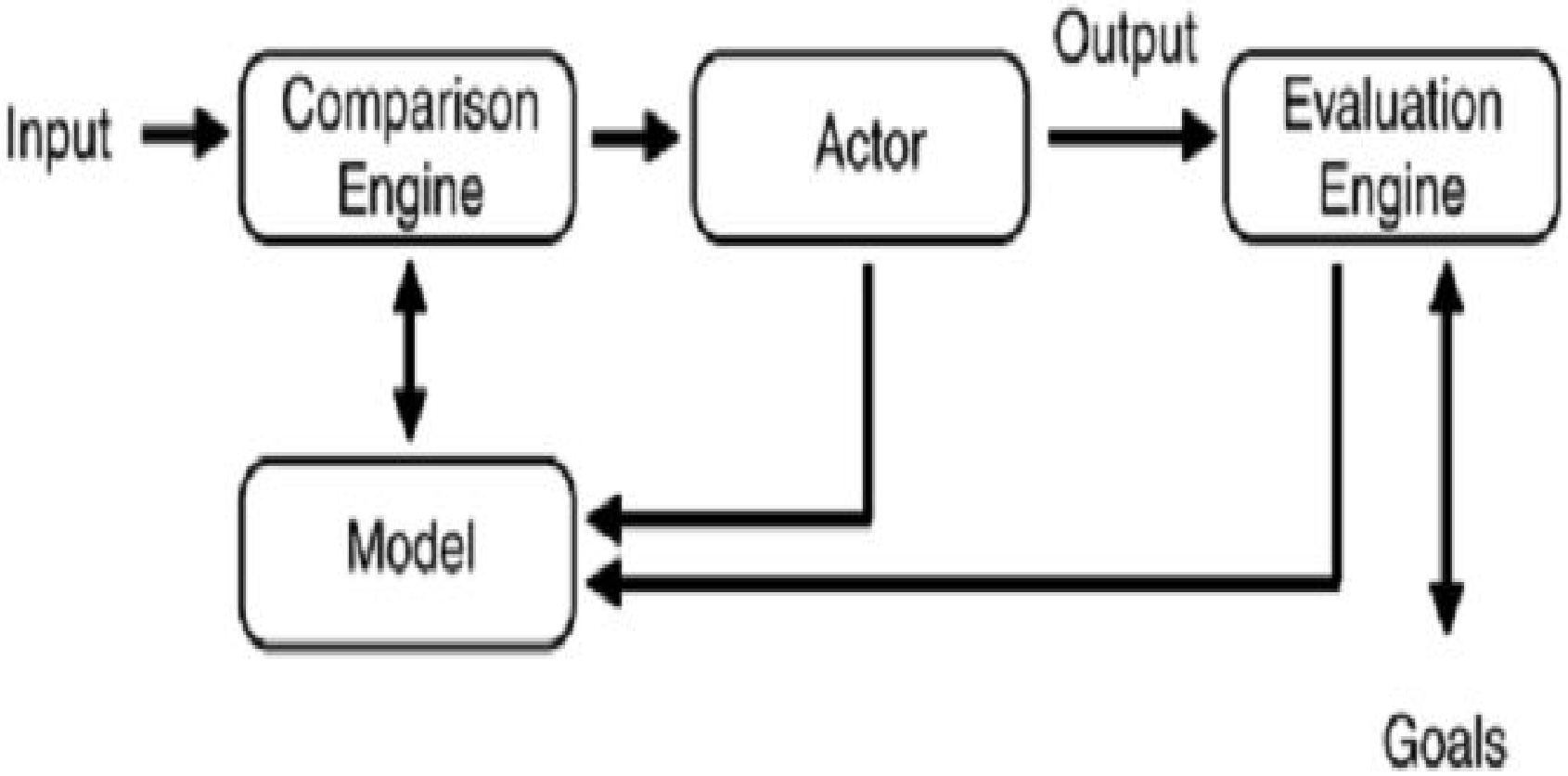
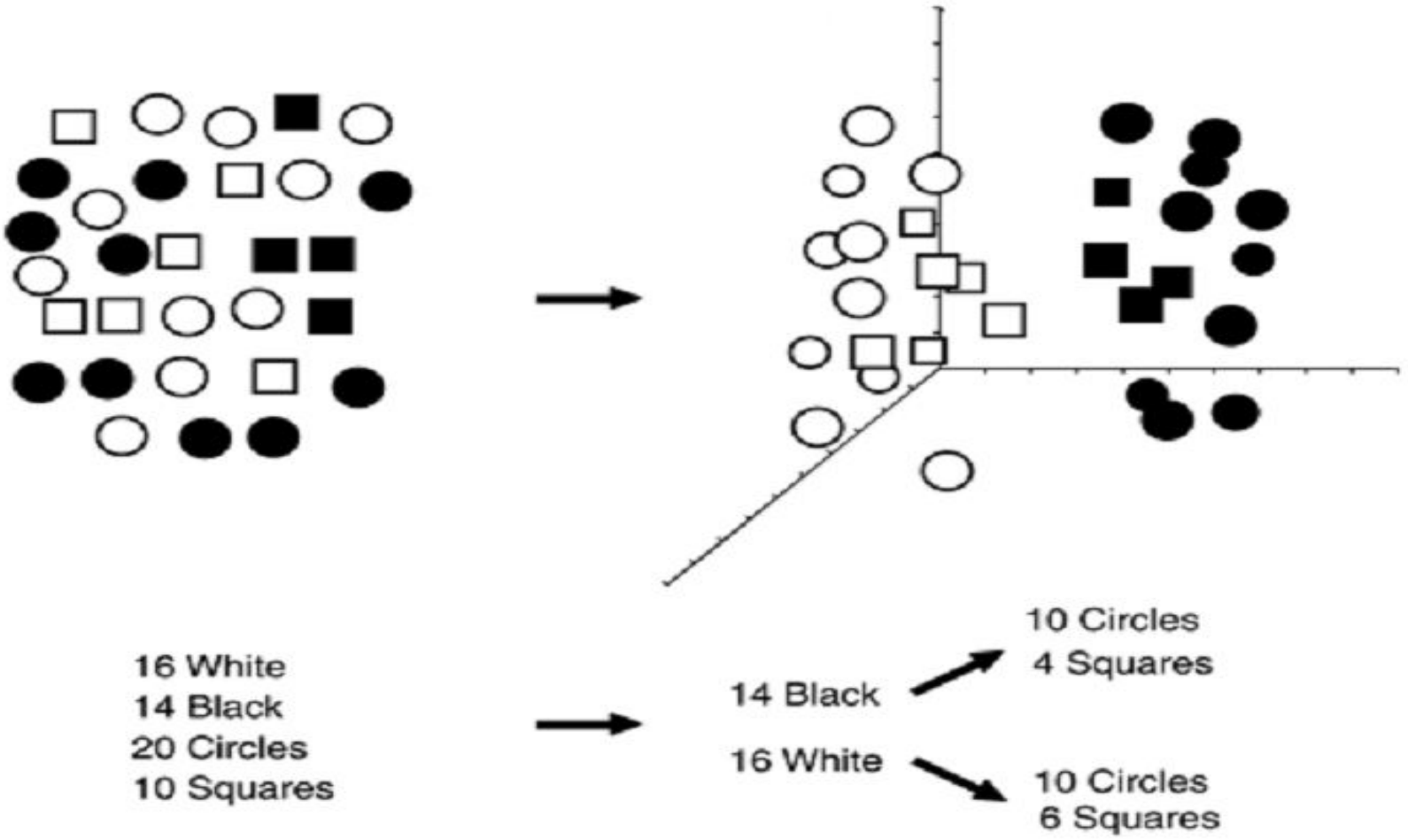


Figure 7-7. Induction-Based Classification. Using changes in entropy (a measure of disorder) as an organizational heuristic, the induction algorithm divides the unorganized data (top left) first by color and then by shape.



Question?

Write down the name 5 biological data mining tools?

References

- Introduction to Data Mining and Knowledge Discovery Third Edition by Two Crows Corporation
- Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 2001;40(4):346-58. PMID: 11552348.
- Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. and Lundahl, S. (2004. March 10). Database Mining in the Human Genome Initiative. Whitepaper, Biodatabases.com, Amita Corporation. Available: http://www.biodatabases.com/whitepaper.html
- Biological Data Mining George Tzanis, Christos Berberidis, and Ioannis Vlahavas
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. AAAI Press/MIT Press, Menlo Park, California, USA.
- Ma, Q. and Wang, J.T.L. (1999). Biological Data Mining Using Bayesian Neural Networks: A Case Study. International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, 8(4), 433-451.
- Han, J. (2002). How Can Data Mining Help Bio-Data Analysis? In Zaki, M.J., Wang, J.T.L. and Toivonen, H.T.T. (Eds). Proceedings of the 2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics, 1-2.
- Hirsh, H. and Noordewier, M. (1994). Using Background Knowledge to Improve Inductive Learning of DNA Sequences. Proceedings of the 10th IEEE Conference on Artificial Intelligence for Applications, 351-357.