Data Mining Methods: Selection & Sampling

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.

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.

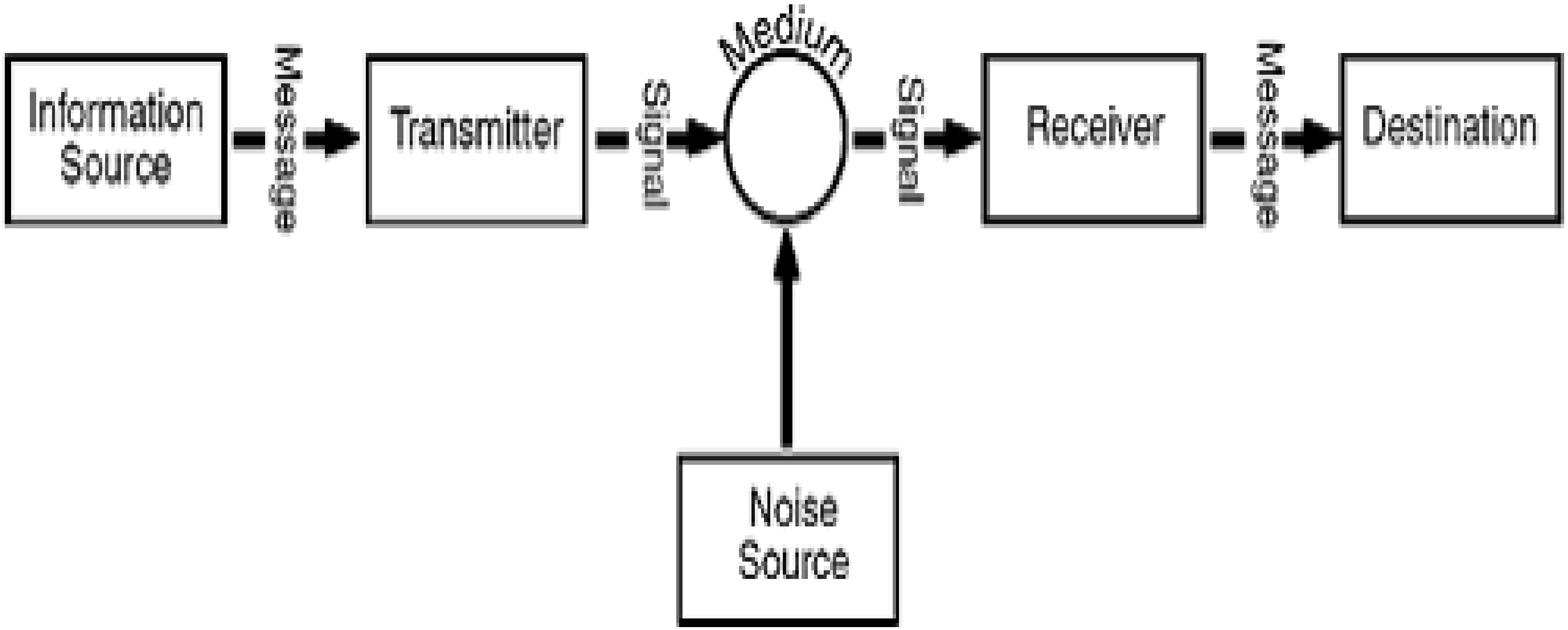
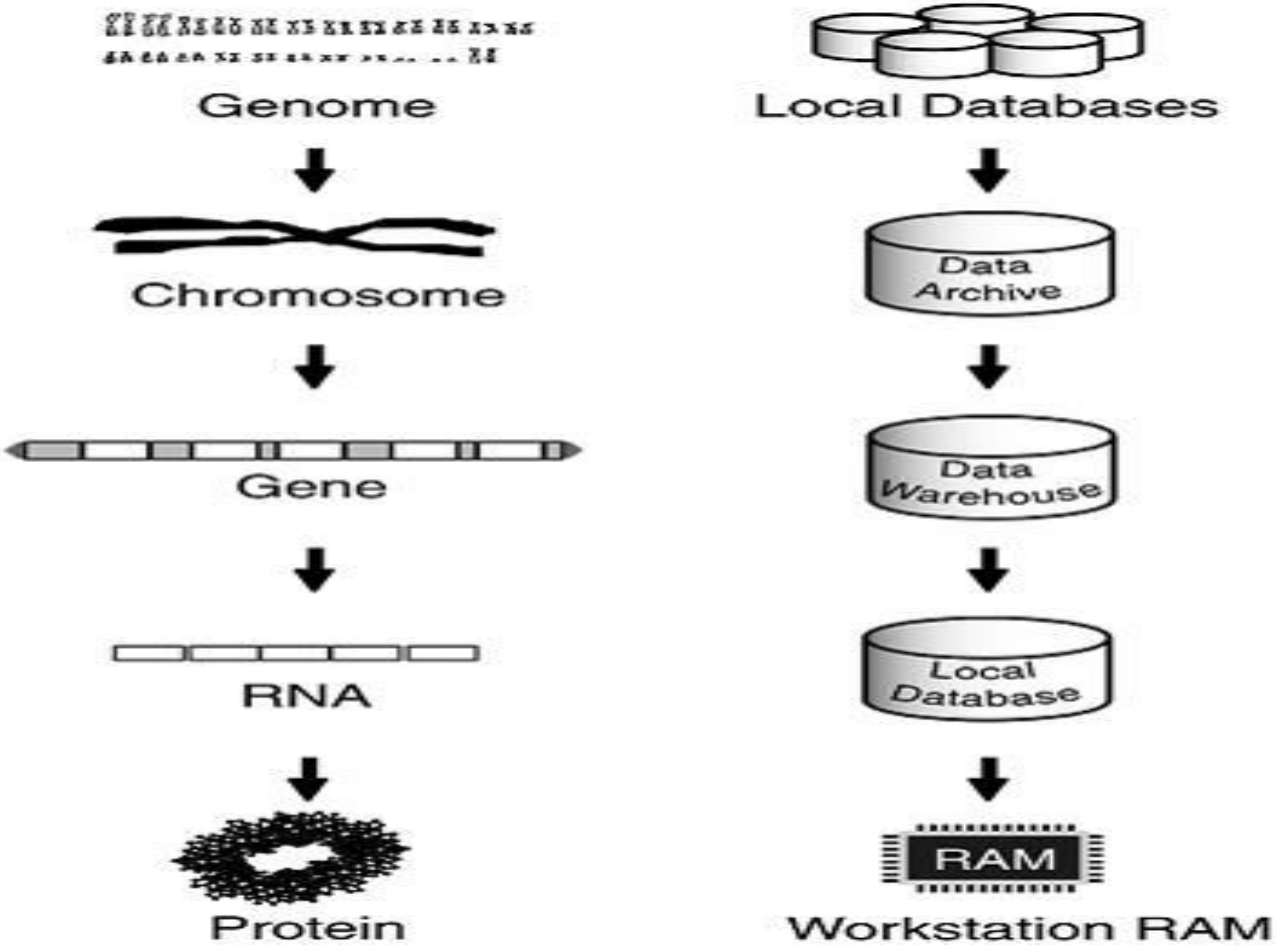


Figure 2-11. Organic Analog of Database Hierarchy. The database hierarchy has many parallels to the hierarchy in the human genome. Data stored in chromosomes, like a data archive, must be unpacked and transferred to a more immediately useful form before the data can be put to use.



- 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.

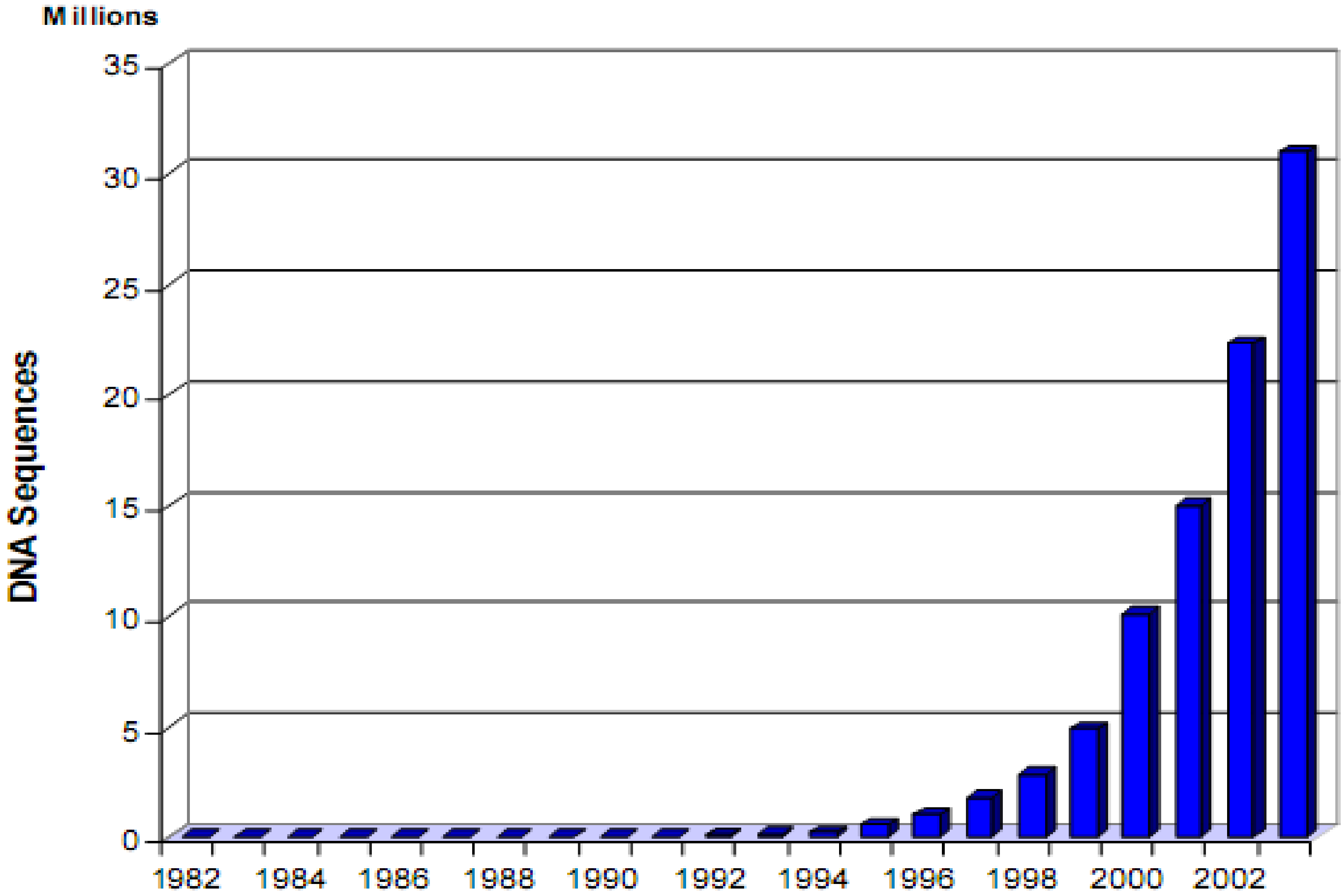


Figure 1: Growth of GenBank (Years 1982-2003).

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 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).

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

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

Table 3: Common Data Mining Tasks.

- 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.)

- 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.

- 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).

- 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.

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

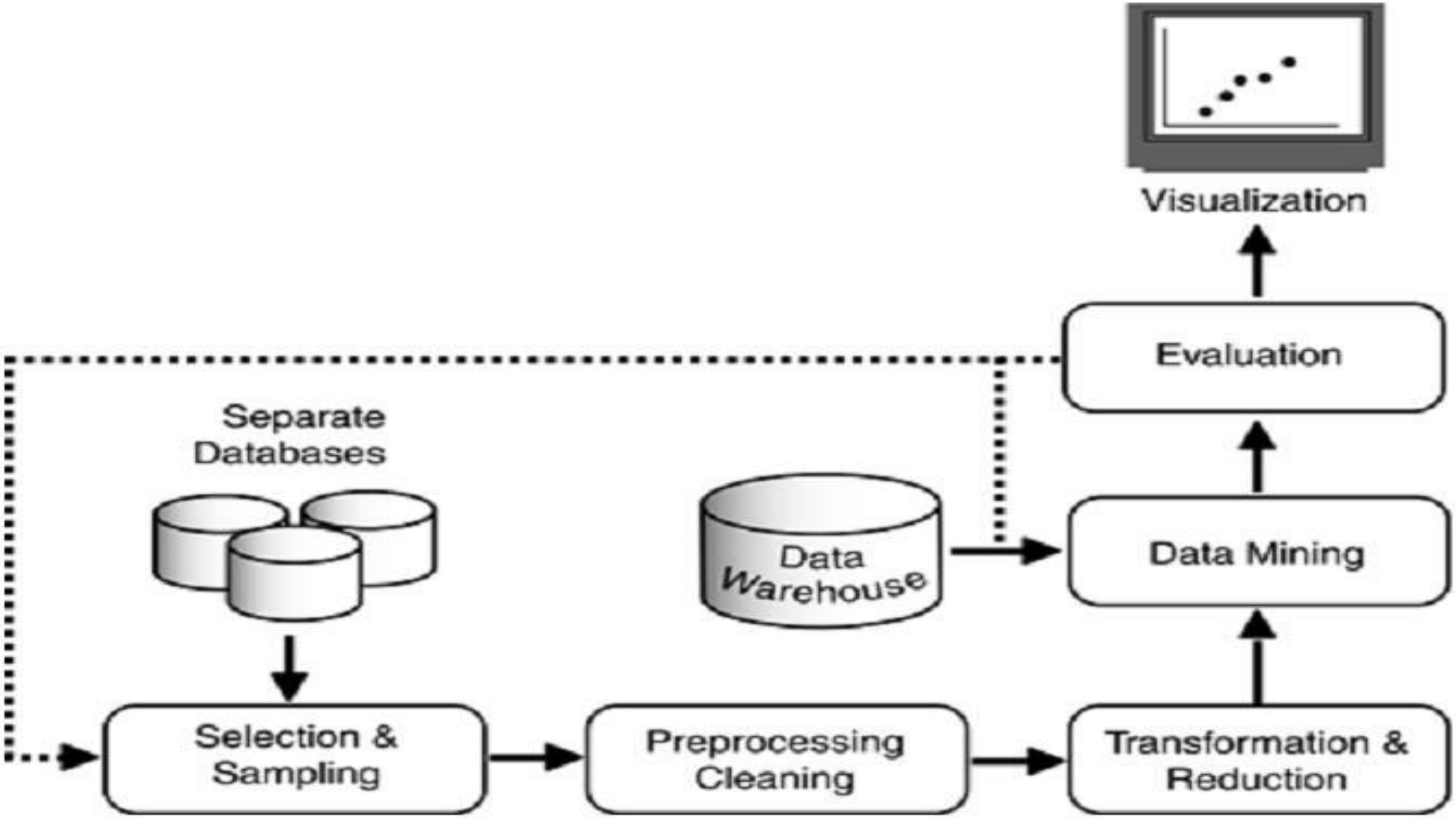


Figure 7-3. Centralized Data-Mining Infrastructure. In this example, a data warehouse, data dictionary, high-bandwidth access to data on the Internet, and a high-performance workstation form the basis for an effective data-mining operation.

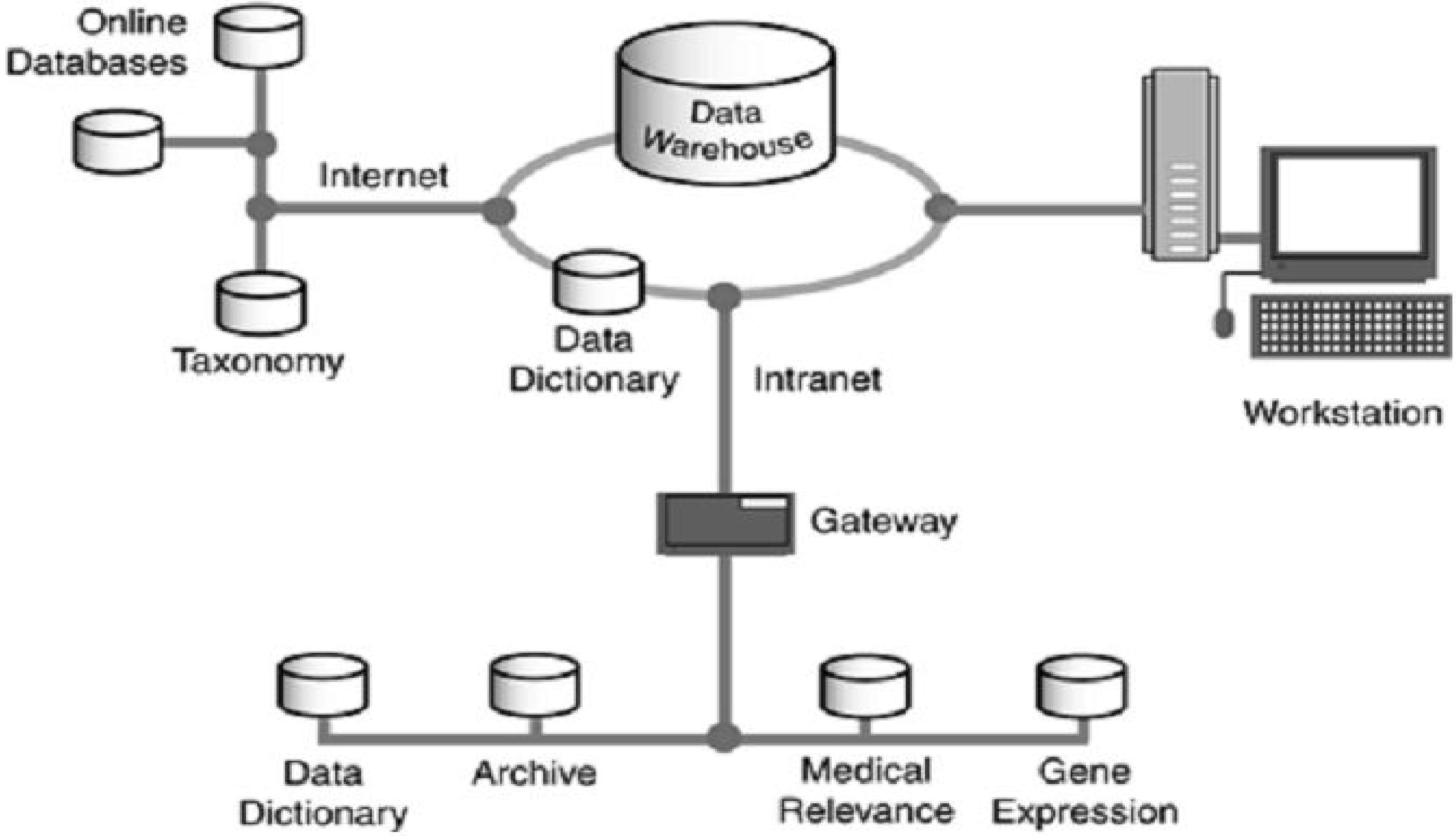


Figure 7-5. The Pattern-Recognition and Discovery Process. Pattern discovery differs from pattern recognition in that feature selection is determined empirically under program control.

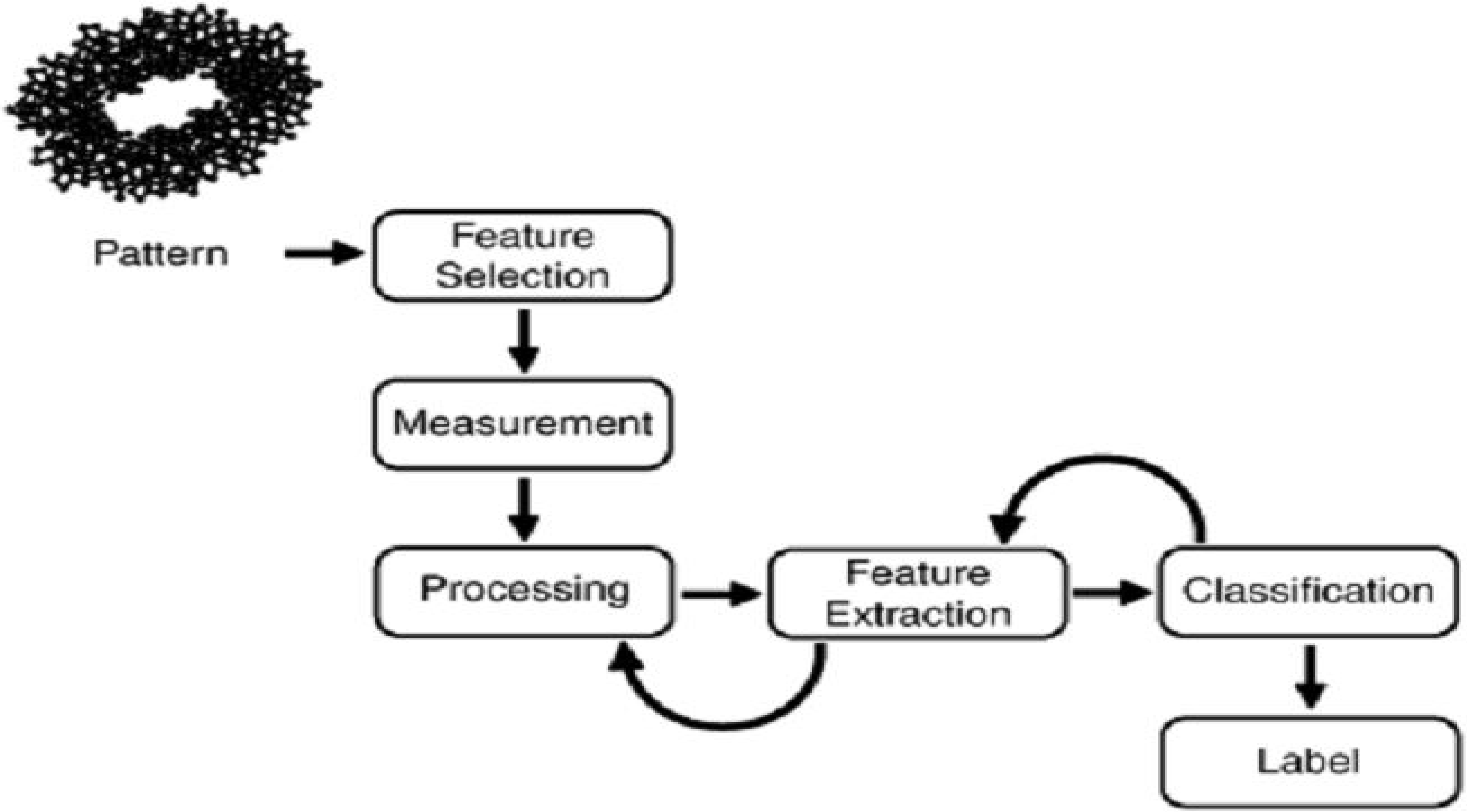
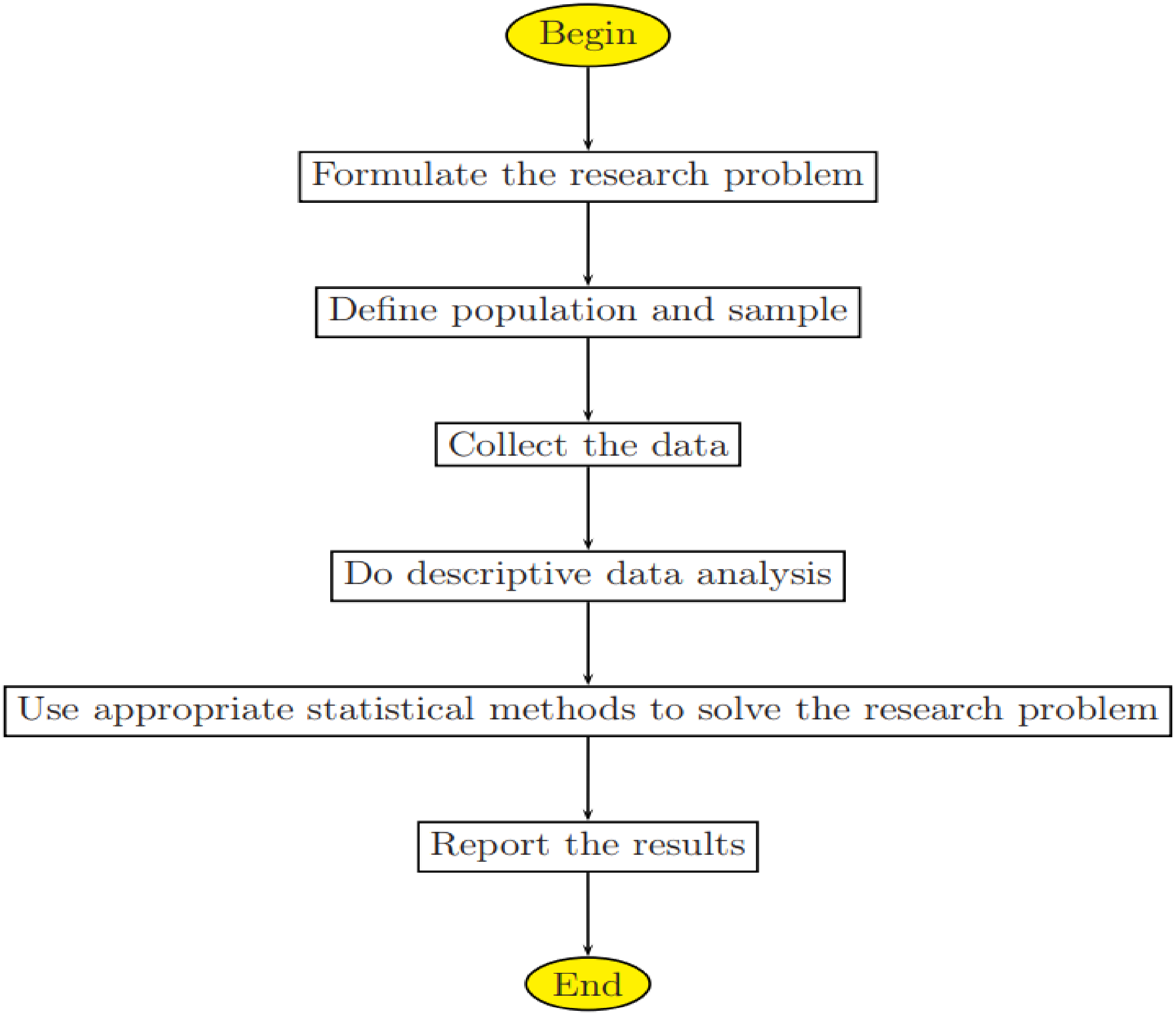


Table 7-3. Machine Learning Technologies and Their Applicability to Data-Mining Methods.

	Data-Mining Methods				
Machine Learning Technologies	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				

Research & Analysis



Begin

Formulate the research problem

Define population and sample

Collect the data

Do descriptive data analysis

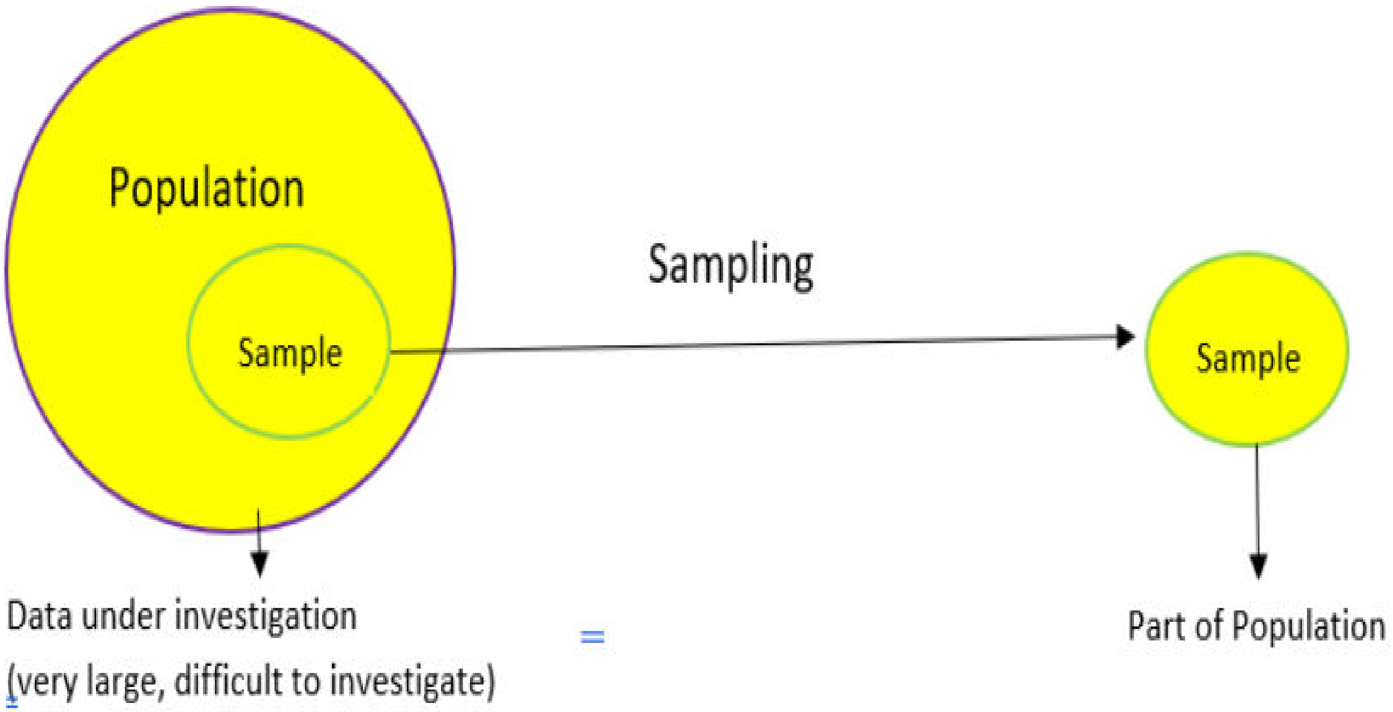
Use appropriate statistical methods to solve the research problem

Report the results

End

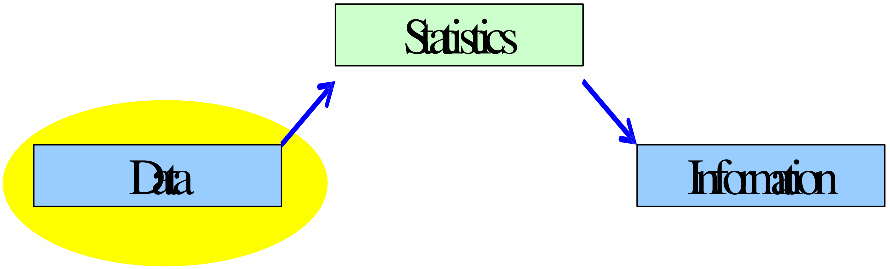
Data Collection & Sampling

Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual. https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/



Recall...

Statistics is a tool for converting *data* into *information*:



But where then does *data* come from? How is it gathered? How do we ensure its accurate? Is the data reliable? Is it representative of the population from which it was drawn? We now explore some of these issues.

There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

- Direct Observation
- Experiments, and
- Surveys.

Surveys. . .

A *survey* solicits information from people; e.g. polls; Gallup pre-election polls; marketing surveys.

The *Response Rate* (i.e. the proportion of all people selected who complete the survey) is a key survey para- meter.

Surveys may be administered in a variety of ways, e.g.

- Personal Interview,
- Telephone Interview, and
- Self-Administered Questionnaire.

Questionnaire Design

Over the years, a lot of thought has been put into the science of the design of survey questions. Key design principles:

1.Keep the questionnaire as short as possible.

2.Ask short, simple, and clearly worded questions.

3.Start with demographic questions to help respondents get started comfortably.

4.Use dichotomous (yes | no) and multiple choice ques- tions.

5.Use open-ended questions cautiously.

6.Avoid using leading-questions.

7.Pretest a questionnaire on a small number of people.

8.Think about the way you intend to use the collected data when preparing the questionnaire.

Sampling. . .

Recall that statistical inference permits us to draw con- clusions about a population based on a sample.

Sampling (i.e. selecting a sub-set of a whole population) is often done for reasons of *cost* (it’s less expensive to sam- ple 1,000 television viewers than 100 million TV viewers) and *practicality* (e.g. performing a crash test on every automobile produced is impractical).

In any case, the *sampled population* and the *target population* should be similar to one another.

Sampling Plans. . .

A *sampling plan* is just a method or procedure for spec- ifying how a sample will be taken from a population.

We will focus our attention on these three methods:

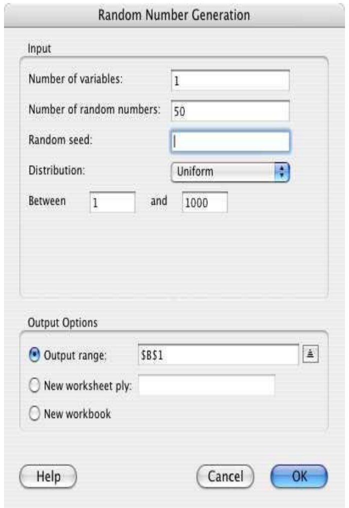
- Simple Random Sampling,
- Stratified Random Sampling, and
- Cluster Sampling.

Simple Random Sampling. . .

A *simple random sample* is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.

Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample: any group of three names is as equally likely as picking any other group of three names.

Example: A government income tax auditor must choose a sample of 40 (usually denoted by n) of 1,000 (usually denoted by N) returns to audit. . .



Random Number Generation

Input

Number of variables: 1

Number of random numbers: 50

Random seed:

Distribution: Uniform

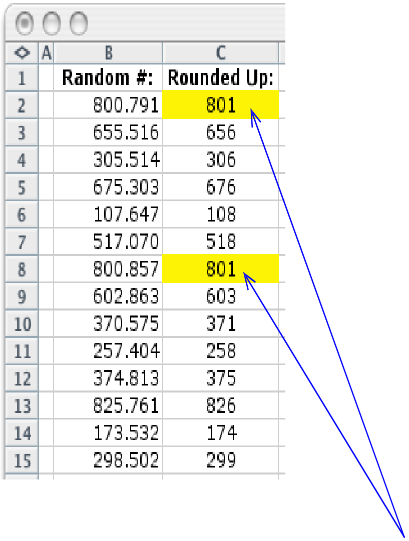
Between 1 and 1000

Output Options

● Output range: B1

○ New worksheet ply:

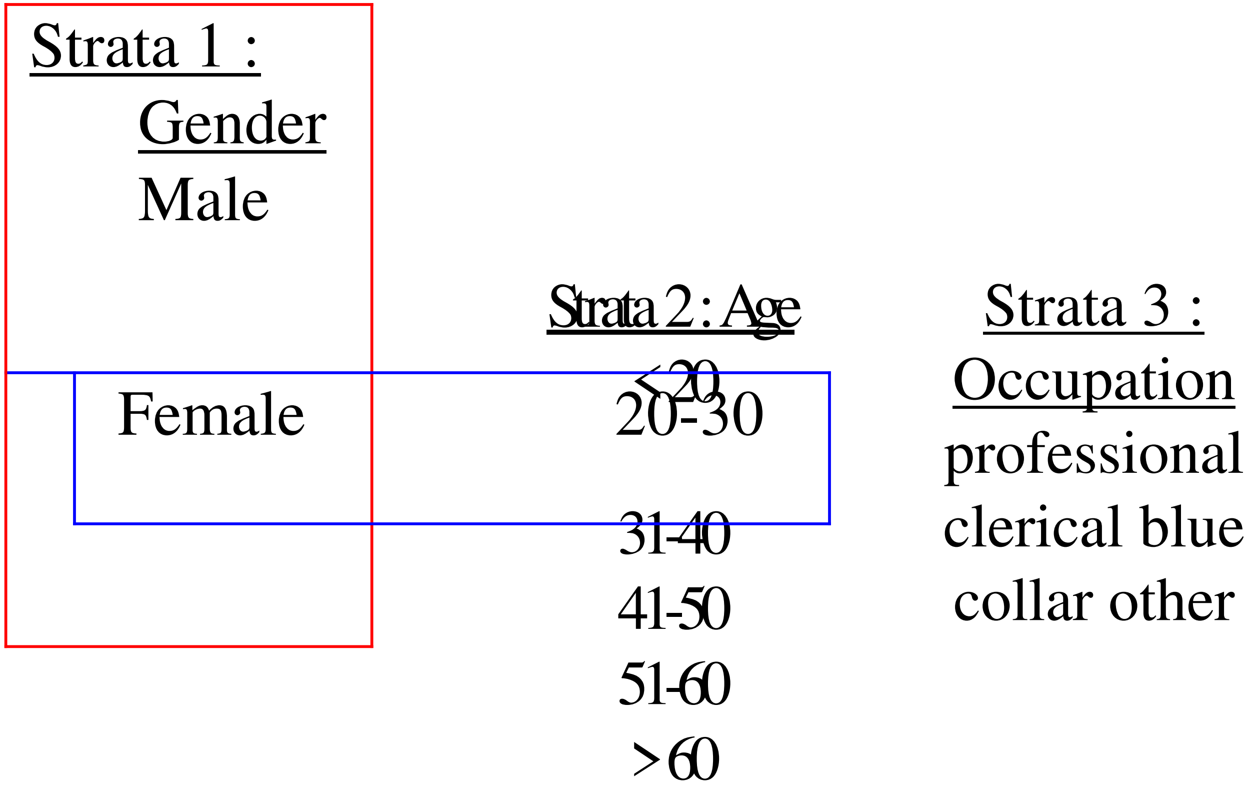
○ New workbook



	A	B	C
1		Random #:	Rounded Up:
2		800.791	801
3		655.516	656
4		305.514	306
5		675.303	676
6		107.647	108
7		517.070	518
8		800.857	801
9		602.863	603
10		370.575	371
11		257.404	258
12		374.813	375
13		825.761	826
14		173.532	174
15		298.502	299

The Excel file “C5-01-Random_Sampling.xls” demonstrates how to use the Excel function RAND() to generate a ran- dom sample from a population. Detailed explanations are provided in the spreadsheet itself.

A *stratified random sample* is obtained by separating the population into *mutually exclusive* sets, or strata, and then drawing simple random samples from each stra- tum.



Strata 1 : Gender

Male

Female

Strata 2 : Age

<20

20-30

31-40

41-50

51-60

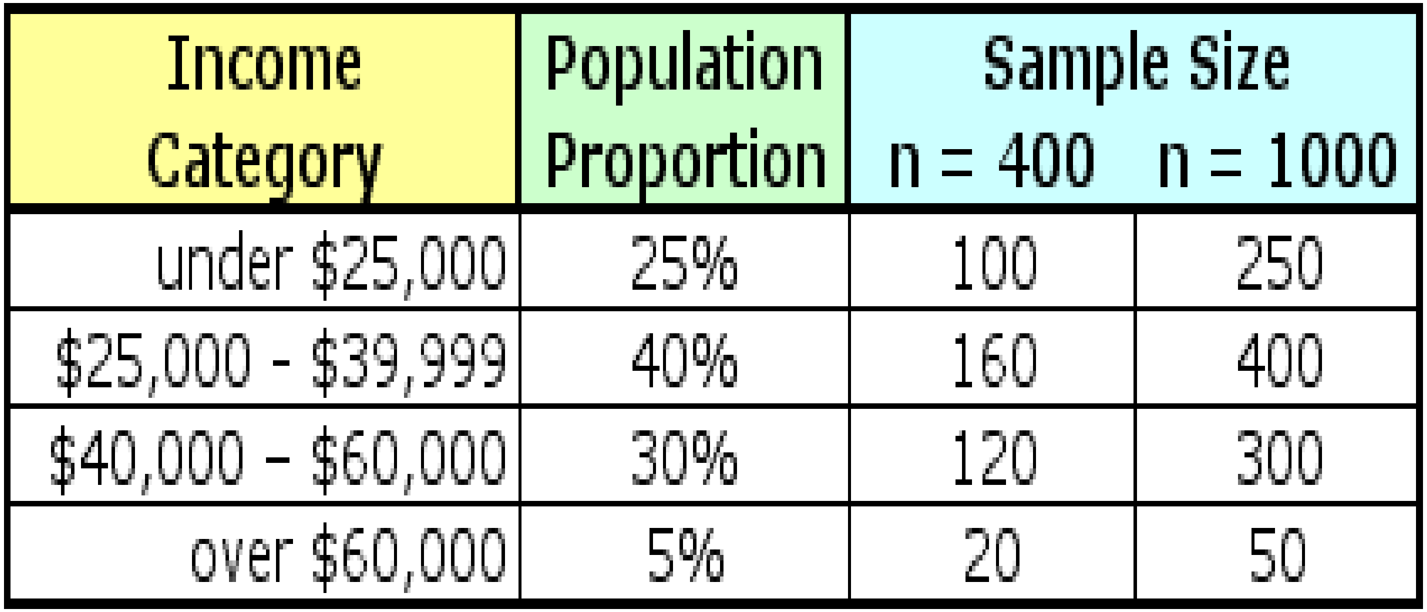
>60

Strata 3 : Occupation

professional

clerical blue collar other

After the population has been stratified, we can use *sim- ple random sampling* to generate the complete sample:



Income Category	Population Proportion	Sample Size	
		n = 400	n = 1000
under $25,000	25%	100	250
$25,000 - $39,999	40%	160	400
$40,000 – $60,000	30%	120	300
over $60,000	5%	20	50

Cluster Sampling. . .

A *cluster sample* is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects).

This method is useful when it is difficult or costly to de- velop a complete list of the population members or when the population elements are widely dispersed geographi- cally.

Cluster sampling may increase sampling error due to similarities among cluster members.

Sample Size. . .

This is an important issue. Numerical techniques for de- termining sample sizes will be described later, but suffice it to say that the larger the sample size is, the more ac- curate we can expect the sample estimates to be.

Sampling and Non-Sampling Errors. . .

Two major types of error can arise when a sample of observations is taken from a population: *sampling error* and *non-sampling error*.

Sampling error refers to differences between the sample and the population that exist only because of the obser- vations that happened to be selected for the sample.

A

Non-sampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

Sampling Error. . .

Sampling error refers to differences between the sam- ple and the population that exist only because of the ob- servations that happened to be selected for the sample.

Another way to look at this is: the differences in results for different samples (of the same size) is due to sampling error:

E.g. Two samples of size 10 of 1,000 households. If we happened to get the highest income level data points in our first sample and all the lowest income levels in the second, this is a consequence of sampling error.

Increasing the sample size *will* reduce this type of error

Errors in Data Acquisition

. . . arises from the recording of incorrect responses, due to:

—incorrect measurements being taken because of faulty equipment,

—mistakes made during transcription from primary sources,

—inaccurate recording of data due to misinterpretation of terms, or

—inaccurate responses to questions concerning sensitive issues.

Errors in Data Acquisition

. . . arises from the recording of incorrect responses, due to:

—incorrect measurements being taken because of faulty equipment,

—mistakes made during transcription from primary sources,

—inaccurate recording of data due to misinterpretation of terms, or

—inaccurate responses to questions concerning sensitive issues.

Nonresponse Error

. . . refers to error (or bias) introduced when responses are not obtained from some members of the sample, i.e. the sample observations that are collected may not be representative of the target population.

As mentioned earlier, the *Response Rate* (i.e. the pro- portion of all people selected who complete the survey) is a key survey parameter and helps in the understanding in the validity of the survey and sources of nonresponse error.

Selection Bias

. . . occurs when the sampling plan is such that some members of the target population cannot possibly be se- lected for inclusion in the sample

References

- Introduction to Data Mining and Knowledge Discovery Third Edition by Two Crows Corporation
- https://www.utdallas.edu/~scniu/OPRE-6301/documents/Data_Collection_and_Sampling.pdf
- 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.