1

Introduction

Whatever you are able to do with your might, do it.

—Kohelet 9:10

1.1 The Personal Computer and Statistics

The personal computer (PC) has changed everything—for both better and worse—in the world of statistics. The PC can effortlessly produce precise calculations and eliminate the computational burden associated with statistics. One need only provide the right questions. With the minimal knowledge required to program (instruct) statistical software, which entails telling it where the input data reside, which statistical procedures and calculations are desired, and where the output should go, tasks such as testing and analyzing, the tabulation of raw data into summary measures, as well as many other statistical criteria are fairly rote. The PC has advanced statistical thinking in the decision-making process, as evidenced by visual displays, such as bar charts and line graphs, animated three-dimensional rotating plots, and interactive marketing models found in management presentations. The PC also facilitates support documentation, which includes the calculations for measures such as the current mean profit across market segments from a marketing database; statistical output is copied from the statistical software and then pasted into the presentation application. Interpreting the output and drawing conclusions still requires human intervention.

Unfortunately, the confluence of the PC and the world of statistics has turned generalists with minimal statistical backgrounds into quasi statisticians and affords them a false sense of confidence because they can now produce statistical output. For instance, calculating the mean profit is standard fare in business. However, the mean provides a "typical value"—only when the distribution of the data is symmetric. In marketing databases, the distribution of profit is commonly right-skewed data.* Thus, the mean profit is not a reliable summary measure.† The quasi statistician would doubtlessly

* *Right skewed* or *positive skewed* means the distribution has a long tail in the positive direction.

† For moderately skewed distributions, the mode or median should be considered, and assessed for a reliably typical value.

not know to check this supposition, thus rendering the interpretation of the mean profit as floccinaucinihilipilification.*

Another example of how the PC fosters a "quick-and-dirty"† approach to statistical analysis can be found in the ubiquitous correlation coefficient (second in popularity to the mean as a summary measure), which measures association between two variables. There is an assumption (the underlying relationship between the two variables is a linear or a straight line) that must be met for the proper interpretation of the correlation coefficient. Rare is the quasi statistician who is actually aware of the assumption. Meanwhile, well-trained statisticians often do not check this assumption, a habit developed by the uncritical use of statistics with the PC.

The professional statistician has also been empowered by the computational strength of the PC; without it, the natural seven-step cycle of statistical analysis would not be practical [1]. The PC and the analytical cycle comprise the perfect pairing as long as the steps are followed in order and the information obtained from a step is used in the next step. Unfortunately, statisticians are human and succumb to taking shortcuts through the seven-step cycle. They ignore the cycle and focus solely on the sixth step in the following list. To the point, a careful statistical endeavor requires performance of all the steps in the seven-step cycle,‡ which is described as follows:

1. *Definition of the problem:* Determining the best way to tackle the problem is not always obvious. Management objectives are often expressed qualitatively, in which case the selection of the outcome or target (dependent) variable is subjectively biased. When the objectives are clearly stated, the appropriate dependent variable is often not available, in which case a surrogate must be used.
2. *Determining technique:* The technique first selected is often the one with which the data analyst is most comfortable; it is not necessarily the best technique for solving the problem.
3. *Use of competing techniques:* Applying alternative techniques increases the odds that a thorough analysis is conducted.
4. *Rough comparisons of efficacy:* Comparing variability of results across techniques can suggest additional techniques or the deletion of alternative techniques.
5. *Comparison in terms of a precise (and thereby inadequate) criterion:* An explicit criterion is difficult to define; therefore, precise surrogates are often used.

---

* Floccinaucinihilipilification (FL OK-si-NO-si-NY-HIL-i-PIL-i-fi KAY-shuhn), noun. Its definition is estimating something as worthless.

† The literal translation of this expression clearly supports my claim that the PC is sometimes not a good thing for statistics. I supplant the former with "thorough and clean."

‡ The seven steps are attributed to Tukey. The annotations are my attributions.

6. *Optimization in terms of a precise and inadequate criterion:* An explicit criterion is difficult to define; therefore, precise surrogates are often used.
7. *Comparison in terms of several optimization criteria:* This constitutes the final step in determining the best solution.

The founding fathers of classical statistics—Karl Pearson* and Sir Ronald Fisher†—would have delighted in the ability of the PC to free them from time-consuming empirical validations of their concepts. Pearson, whose contributions include, to name but a few, regression analysis, the correlation coefficient, the standard deviation (a term he coined), and the chi-square test of statistical significance, would have likely developed even more concepts with the free time afforded by the PC. One can further speculate that the functionality of the PC would have allowed Fisher's methods (e.g., maximum likelihood estimation, hypothesis testing, and analysis of variance) to have immediate and practical applications.

The PC took the classical statistics of Pearson and Fisher from their theoretical blackboards into the practical classrooms and boardrooms. In the 1970s, statisticians were starting to acknowledge that their methodologies had potential for wider applications. However, they knew an accessible computing device was required to perform their on-demand statistical analyses with an acceptable accuracy and within a reasonable turnaround time. Although the statistical techniques had been developed for a small data setting consisting of one or two handfuls of variables and up to hundreds of records, the hand tabulation of data was computationally demanding and almost insurmountable. Accordingly, conducting the statistical techniques on big data was virtually out of the question. With the inception of the microprocessor in the mid-1970s, statisticians now had their computing device, the PC, to perform statistical analysis on big data with excellent accuracy and turnaround time. The desktop PCs replaced the handheld calculators in the classroom and boardrooms. From the 1990s to the present, the PC has offered statisticians advantages that were imponderable decades earlier.

1.2 Statistics and Data Analysis

As early as 1957, Roy believed that the classical statistical analysis was largely likely to be supplanted by assumption-free, nonparametric

* Karl Person (1900s) contributions include regression analysis, the correlation coefficient, and the chi-square test of statistical significance. He coined the term *standard deviation* in 1893.

† Sir Ronald Fisher (1920s) invented the methods of maximum likelihood estimation, hypothesis testing, and analysis of variance.

approaches, which were more realistic and meaningful [2]. It was an onerous task to understand the robustness of the classical (parametric) techniques to violations of the restrictive and unrealistic assumptions underlying their use. In practical applications, the primary assumption of "a random sample from a multivariate normal population" is virtually untenable. The effects of violating this assumption and additional model-specific assumptions (e.g., linearity between predictor and dependent variables, constant variance among errors, and uncorrelated errors) are difficult to determine with any exactitude. It is difficult to encourage the use of the statistical techniques, given that their limitations are not fully understood.

In 1962, in his influential article, "The Future of Data Analysis," John Tukey expressed concern that the field of statistics was not advancing [1]. He felt there was too much focus on the mathematics of statistics and not enough on the analysis of data and predicted a movement to unlock the rigidities that characterize the discipline. In an act of statistical heresy, Tukey took the first step toward revolutionizing statistics by referring to himself not as a statistician but a data analyst. However, it was not until the publication of his seminal masterpiece *Exploratory Data Analysis* in 1977 that Tukey led the discipline away from the rigors of statistical inference into a new area, known as EDA (stemming from the first letter of each word in the title of the unquestionable masterpiece) [3]. For his part, Tukey tried to advance EDA as a separate and distinct discipline from statistics, an idea that is not universally accepted today. EDA offered a fresh, assumption-free, nonparametric approach to problem solving in which the analysis is guided by the data itself and utilizes self-educating techniques, such as iteratively testing and modifying the analysis as the evaluation of feedback, to improve the final analysis for reliable results.

The essence of EDA is best described in Tukey's own words:

> Exploratory data analysis is detective work—numerical detective work—or counting detective work—or graphical detective work. ... [It is] about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. [3, p. 1]

EDA includes the following characteristics:

1. *Flexibility*—techniques with greater flexibility to delve into the data
2. *Practicality*—advice for procedures of analyzing data
3. *Innovation*—techniques for interpreting results
4. *Universality*—use all statistics that apply to analyzing data
5. *Simplicity*—above all, the belief that simplicity is the golden rule

On a personal note, when I learned that Tukey preferred to be called a data analyst, I felt both validated and liberated because many of my own analyses fell outside the realm of the classical statistical framework. Furthermore, I had virtually eliminated the mathematical machinery, such as the calculus of maximum likelihood. In homage to Tukey, I more frequently use the terms *data analyst* and *data analysis* rather than statistical analysis and statistician throughout the book.

1.3 EDA

Tukey's book is more than a collection of new and creative rules and operations; it defines EDA as a discipline, which holds that data analysts fail only if they fail to try many things. It further espouses the belief that data analysts are especially successful if their detective work forces them to notice the unexpected. In other words, the philosophy of EDA is a trinity of *attitude* and *flexibility* to do whatever it takes to refine the analysis and *sharp-sightedness* to observe the unexpected when it does appear. EDA is thus a self-propagating theory; each data analyst adds his or her own contribution, thereby contributing to the discipline, as I hope to accomplish with this book.

The sharp-sightedness of EDA warrants more attention, as it is an important feature of the EDA approach. The data analyst should be a keen observer of indicators that are capable of being dealt with successfully and use them to paint an analytical picture of the data. In addition to the ever-ready visual graphical displays as an indicator of what the data reveal, there are numerical indicators, such as counts, percentages, averages, and the other classical descriptive statistics (e.g., standard deviation, minimum, maximum, and missing values). The data analyst's personal judgment and interpretation of indictors are not considered a bad thing, as the goal is to draw informal inferences, rather than those statistically significant inferences that are the hallmark of statistical formality.

In addition to visual and numerical indicators, there are the indirect messages in the data that force the data analyst to take notice, prompting responses such as "the data look like..." or "It appears to be...." Indirect messages may be vague, but their importance is to help the data analyst draw informal inferences. Thus, indicators do not include any of the hard statistical apparatus, such as confidence limits, significance tests, or standard errors.

With EDA, a new trend in statistics was born. Tukey and Mosteller quickly followed up in 1977 with the second EDA book, commonly referred to as EDA II, *Data Analysis and Regression*. EDA II recasts the basics of classical inferential procedures of data analysis and regression into an assumption-free, nonparametric approach guided by "(a) a sequence of philosophical attitudes... for effective data analysis, and (b) a flow of useful and adaptable techniques that make it possible to put these attitudes to work" [4, p. vii].

Hoaglin, Mosteller, and Tukey in 1983 succeeded in advancing EDA with *Understanding Robust and Exploratory Data Analysis*, which provides an understanding of how badly the classical methods behave when their restrictive assumptions do not hold and offers alternative robust and exploratory methods to broaden the effectiveness of statistical analysis [5]. It includes a collection of methods to cope with data in an informal way, guiding the identification of data structures relatively quickly and easily and trading off optimization of objective for stability of results.

Hoaglin et al. in 1991 continued their fruitful EDA efforts with *Fundamentals of Exploratory Analysis of Variance* [6]. They refashioned the basics of the analysis of variance with the classical statistical apparatus (e.g., degrees of freedom, F-ratios, and p values) into a host of numerical and graphical displays, which often give insight into the structure of the data, such as size effects, patterns, and interaction and behavior of residuals.

EDA set off a burst of activity in the visual portrayal of data. Published in 1983, *Graphical Methods for Data Analysis* (Chambers et al.) presented new and old methods—some of which require a computer, while others only paper and pencil—but all are powerful data analytical tools to learn more about data structure [7]. In 1986, du Toit et al. came out with *Graphical Exploratory Data Analysis*, providing a comprehensive, yet simple presentation of the topic [8]. Jacoby, with *Statistical Graphics for Visualizing Univariate and Bivariate Data* (1997), and *Statistical Graphics for Visualizing Multivariate Data* (1998), carried out his objective to obtain pictorial representations of quantitative information by elucidating histograms, one-dimensional and enhanced scatterplots, and nonparametric smoothing [9, 10]. In addition, he successfully transferred graphical displays of multivariate data on a single sheet of paper, a two-dimensional space.

1.4 The EDA Paradigm

EDA presents a major paradigm shift in the ways models are built. With the mantra, "Let your data be your guide," EDA offers a view that is a complete reversal of the classical principles that govern the usual steps of model building. EDA declares the model must always follow the data, not the other way around, as in the classical approach.

In the classical approach, the problem is stated and formulated in terms of an outcome variable Y. It is assumed that the true model explaining all the variation in Y is known. Specifically, it is assumed that all the structures (predictor variables, X_is) affecting Y and their forms are known and present in the model. For example, if Age affects Y, but the log of Age reflects the true relationship with Y, then log of Age must be present in the model. Once the model is specified, the data are taken through the model-specific analysis,

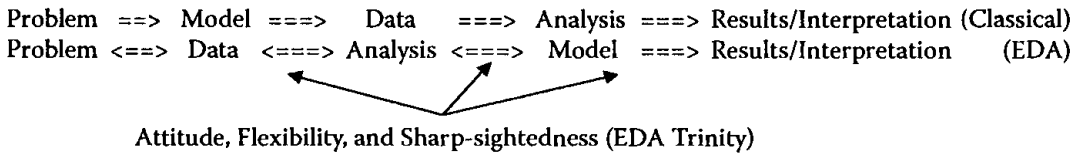

FIGURE 1.1
EDA paradigm.

which provides the results in terms of numerical values associated with the structures or estimates of the coefficients of the true predictor variables. Then, interpretation is made for declaring X_i an important predictor, assessing how X_i affects the prediction of Y, and ranking X_i in order of predictive importance.

Of course, the data analyst never knows the true model. So, familiarity with the content domain of the problem is used to put forth explicitly the true surrogate model, from which good predictions of Y can be made. According to Box, "all models are wrong, but some are useful" [11]. In this case, the model selected provides serviceable predictions of Y. Regardless of the model used, the assumption of knowing the truth about Y sets the statistical logic in motion to cause likely bias in the analysis, results, and interpretation.

In the EDA approach, not much is assumed beyond having some prior experience with content domain of the problem. The right attitude, flexibility, and sharp-sightedness are the forces behind the data analyst, who assesses the problem and lets the data direct the course of the analysis, which then suggests the structures and their forms in the model. If the model passes the validity check, then it is considered final and ready for results and interpretation to be made. If not, with the force still behind the data analyst, revisits of the analysis or data are made until new structures produce a sound and validated model, after which final results and interpretation are made (see Figure 1.1). Without exposure to assumption violations, the EDA paradigm offers a degree of confidence that its prescribed exploratory efforts are not biased, at least in the manner of classical approach. Of course, no analysis is bias free as all analysts admit their own bias into the equation.

1.5 EDA Weaknesses

With all its strengths and determination, EDA as originally developed had two minor weaknesses that could have hindered its wide acceptance and great success. One is of a subjective or psychological nature, and the other is a misconceived notion. Data analysts know that failure to look into a multitude of possibilities can result in a flawed analysis, thus finding themselves in a competitive struggle against the data itself. Thus, EDA can foster data analysts with insecurity that their work is never done. The PC can assist data

analysts in being thorough with their analytical due diligence but bears no responsibility for the arrogance EDA engenders.

The belief that EDA, which was originally developed for the small data setting, does not work as well with large samples is a misconception. Indeed, some of the graphical methods, such as the stem-and-leaf plots, and some of the numerical and counting methods, such as folding and binning, do break down with large samples. However, the majority of the EDA methodology is unaffected by data size. Neither the manner by which the methods are carried out nor the reliability of the results is changed. In fact, some of the most powerful EDA techniques scale up nicely, but do require the PC to do the serious number crunching of *big data** [12]. For example, techniques such as ladder of powers, reexpressing,† and smoothing are valuable tools for large-sample or big data applications.

1.6 Small and Big Data

I would like to clarify the general concept of "small" and "big" data, as size, like beauty, is in the mind of the data analyst. In the past, small data fit the conceptual structure of classical statistics. Small always referred to the sample size, not the number of variables, which were always kept to a handful. Depending on the method employed, small was seldom less than 5 individuals; sometimes between 5 and 20; frequently between 30 and 50 or between 50 and 100; and rarely between 100 and 200. In contrast to today's big data, small data are a tabular display of rows (observations or individuals) and columns (variables or features) that fits on a few sheets of paper.

In addition to the compact area they occupy, small data are neat and tidy. They are "clean," in that they contain no improbable or impossible values, except for those due to primal data entry error. They do not include the statistical outliers and influential points or the EDA far-out and outside points. They are in the "ready-to-run" condition required by classical statistical methods.

There are two sides to big data. On one side is classical statistics that considers big as simply not small. Theoretically, big is the sample size after which asymptotic properties of the method "kick in" for valid results. On the other side is contemporary statistics that considers big in terms of lifting

* Authors Weiss and Indurkhya and I use the general concept of "big" data. However, we stress different characteristics of the concept.

† Tukey, via his groundbreaking EDA book, put the concept of "reexpression" in the forefront of EDA data mining tools; yet, he never provided any definition. I assume he believed that the term is self-explanatory. Tukey's first mention of reexpression is in a question on page 61 of his work: "What is the single most needed form of re-expression?" I, for one, would like a definition of reexpression, and I provide one further in the book.

observations and learning from the variables. Although it depends on who is analyzing the data, a sample size greater than 50,000 individuals can be considered big. Thus, calculating the average income from a database of 2 million individuals requires heavy-duty lifting (number crunching). In terms of learning or uncovering the structure among the variables, big can be considered 50 variables or more. Regardless of which side the data analyst is working, EDA scales up for both rows and columns of the data table.

1.6.1 Data Size Characteristics

There are three distinguishable characteristics of data size: condition, location, and population. *Condition* refers to the state of readiness of the data for analysis. Data that require minimal time and cost to clean, before reliable analysis can be performed, are said to be well conditioned; data that involve a substantial amount of time and cost are said to be ill conditioned. Small data are typically clean and therefore well conditioned.

Big data are an outgrowth of today's digital environment, which generates data flowing continuously from all directions at unprecedented speed and volume, and these data usually require cleansing. They are considered "dirty" mainly because of the merging of multiple sources. The merging process is inherently a time-intensive process, as multiple passes of the sources must be made to get a sense of how the combined sources fit together. Because of the iterative nature of the process, the logic of matching individual records across sources is at first "fuzzy," then fine-tuned to soundness; until that point, unexplainable, seemingly random, nonsensical values result. Thus, big data are ill conditioned.

Location refers to where the data reside. Unlike the rectangular sheet for small data, big data reside in relational databases consisting of a set of data tables. The link among the data tables can be hierarchical (rank or level dependent) or sequential (time or event dependent). Merging of multiple data sources, each consisting of many rows and columns, produces data of even greater number of rows and columns, clearly suggesting bigness.

Population refers to the group of individuals having qualities or characteristics in common and related to the study under consideration. Small data ideally represent a random sample of a known population that is not expected to encounter changes in its composition in the near future. The data are collected to answer a specific problem, permitting straightforward answers from a given problem-specific method. In contrast, big data often represent multiple, nonrandom samples of unknown populations, shifting in composition within the short term. Big data are "secondary" in nature; that is, they are not collected for an intended purpose. They are available from the hydra of marketing information, for use on any post hoc problem, and may not have a straightforward solution.

It is interesting to note that Tukey never talked specifically about the big data per se. However, he did predict that the cost of computing, in

both time and dollars, would be cheap, which arguably suggests that he knew big data were coming. Regarding the cost, clearly today's PC bears this out.

1.6.2 Data Size: Personal Observation of One

The data size discussion raises the following question: "How large should a sample be?" Sample size can be anywhere from folds of 10,000 up to 100,000.

In my experience as a statistical modeler and data mining consultant for over 15 years and a statistics instructor who analyzes deceivingly simple cross tabulations with the basic statistical methods as my data mining tools, I have observed that the less-experienced and -trained data analyst uses sample sizes that are unnecessarily large. I see analyses performed on and models built from samples too large by factors ranging from 20 to 50. Although the PC can perform the heavy calculations, the extra time and cost in getting the larger data out of the data warehouse and then processing them and thinking about it are almost never justified. Of course, the only way a data analyst learns that extra big data are a waste of resources is by performing small versus big data comparisons, a step I recommend.

1.7 Data Mining Paradigm

The term *data mining* emerged from the database marketing community sometime between the late 1970s and early 1980s. Statisticians did not understand the excitement and activity caused by this new technique since the discovery of patterns and relationships (structure) in the data is not new to them. They had known about data mining for a long time, albeit under various names, such as data fishing, snooping, and dredging, and most disparaging, "ransacking" the data. Because any discovery process inherently exploits the data, producing spurious findings, statisticians did not view data mining in a positive light.

To state one of the numerous paraphrases of Maslow's hammer,* "If you have a hammer in hand, you tend eventually to start seeing nails." The statistical version of this maxim is, "Simply looking for something increases the odds that something will be found." Therefore, looking

* Abraham Maslow brought to the world of psychology a fresh perspective with his concept of "humanism," which he referred to as the "third force" of psychology after Pavlov's "behaviorism" and Freud's "psychoanalysis." Maslow's hammer is frequently used without anybody seemingly knowing the originator of this unique pithy statement expressing a rule of conduct. Maslow's Jewish parents migrated from Russia to the United States to escape from harsh conditions and sociopolitical turmoil. He was born Brooklyn, New York, in April 1908 and died from a heart attack in June 1970.

for structure typically results in finding structure. All data have spurious structures, which are formed by the "forces" that make things come together, such as chance. The bigger the data, the greater are the odds that spurious structures abound. Thus, an expectation of data mining is that it produces structures, both real and spurious, without distinction between them.

Today, statisticians accept data mining only if it embodies the EDA paradigm. They define *data mining* as any process that finds unexpected structures in data and uses the EDA framework to ensure that the process explores the data, not exploits it (see Figure 1.1). Note the word *unexpected*, which suggests that the process is exploratory rather than a confirmation that an expected structure has been found. By finding what one expects to find, there is no longer uncertainty regarding the existence of the structure.

Statisticians are mindful of the inherent nature of data mining and try to make adjustments to minimize the number of spurious structures identified. In classical statistical analysis, statisticians have explicitly modified most analyses that search for interesting structure, such as adjusting the overall alpha level/type I error rate or inflating the degrees of freedom [13, 14]. In data mining, the statistician has no explicit analytical adjustments available, only the implicit adjustments affected by using the EDA paradigm itself. The steps discussed next outline the data mining/EDA paradigm. As expected from EDA, the steps are defined by *soft* rules.

Suppose the objective is to find structure to help make good predictions of response to a future mail campaign. The following represent the steps that need to be taken:

Obtain the database that has similar mailings to the future mail campaign.

Draw a sample from the database. Size can be several folds of 10,000, up to 100,000.

Perform many exploratory passes of the sample. That is, do all desired calculations to determine interesting or noticeable structures.

Stop the calculations that are used for finding the noticeable structure.

Count the number of noticeable structures that emerge. The structures are not necessarily the results and should not be declared significant findings.

Seek out indicators, visual and numerical, and the indirect messages.

React or respond to all indicators and indirect messages.

Ask questions. Does each structure make sense by itself? Do any of the structures form natural groups? Do the groups make sense; is there consistency among the structures within a group?

Try more techniques. Repeat the many exploratory passes with several fresh samples drawn from the database. Check for consistency across the multiple passes. If results do not behave in a

similar way, there may be no structure to predict response to a future mailing, as chance may have infected your data. If results behave similarly, then assess the variability of each structure and each group.

Choose the most stable structures and groups of structures for predicting response to a future mailing.

1.8 Statistics and Machine Learning

Coined by Samuel in 1959, the term *machine learning* (ML) was given to the field of study that assigns computers the ability to learn without being explicitly programmed [15]. In other words, ML investigates ways in which the computer can acquire knowledge directly from data and thus learn to solve problems. It would not be long before ML would influence the statistical community.

In 1963, Morgan and Sonquist led a rebellion against the restrictive assumptions of classical statistics [16]. They developed the automatic interaction detection (AID) regression tree, a methodology without assumptions. AID is a computer-intensive technique that finds or learns multidimensional patterns and relationships in data and serves as an assumption-free, nonparametric alternative to regression prediction and classification analyses. Many statisticians believe that AID marked the beginning of an ML approach to solving statistical problems. There have been many improvements and extensions of AID: THAID, MAID, CHAID (chi-squared automatic interaction detection), and CART, which are now considered viable and quite accessible data mining tools. CHAID and CART have emerged as the most popular today.

I consider AID and its offspring as quasi-ML methods. They are computer-intensive techniques that need the PC machine, a necessary condition for an ML method. However, they are not true ML methods because they use explicitly statistical criteria (e.g., chi squared and the F-tests), for the learning. A genuine ML method has the PC itself learn via mimicking the way humans think. Thus, I must use the term *quasi*. Perhaps a more appropriate and suggestive term for AID-type procedures and other statistical problems using the PC machine is statistical ML.

Independent from the work of Morgan and Sonquist, ML researchers had been developing algorithms to automate the induction process, which provided another alternative to regression analysis. In 1979, Quinlan used the well-known concept learning system developed by Hunt et al. to implement one of the first intelligent systems—ID3—which was succeeded by C4.5 and C5.0 [17, 18]. These algorithms are also considered data mining tools but have not successfully crossed over to the statistical community.

The interface of statistics and ML began in earnest in the 1980s. ML researchers became familiar with the three classical problems facing statisticians: regression (predicting a continuous outcome variable), classification (predicting a categorical outcome variable), and clustering (generating a few composite variables that carry a large percentage of the information in the original variables). They started using their machinery (algorithms and the PC) for a nonstatistical, assumption-free nonparametric approach to the three problem areas. At the same time, statisticians began harnessing the power of the desktop PC to influence the classical problems they know so well, thus relieving themselves from the starchy parametric road.

The ML community has many specialty groups working on data mining: neural networks, support vector machines, fuzzy logic, genetic algorithms and programming, information retrieval, knowledge acquisition, text processing, inductive logic programming, expert systems, and dynamic programming. All areas have the same objective in mind but accomplish it with their own tools and techniques. Unfortunately, the statistics community and the ML subgroups have no real exchanges of ideas or best practices. They create distinctions of no distinction.

1.9 Statistical Data Mining

In the spirit of EDA, it is incumbent on data analysts to try something new and retry something old. They can benefit not only from the computational power of the PC in doing the heavy lifting of big data but also from the ML ability of the PC in uncovering structure nestled in big data. In the spirit of trying something old, statistics still has a lot to offer.

Thus, today's data mining can be defined in terms of three easy concepts:

1. *Statistics with emphasis on EDA proper:* This includes using the descriptive and noninferential parts of classical statistical machinery as indicators. The parts include sum of squares, degrees of freedom, F-ratios, chi-square values, and p values, but exclude inferential conclusions.
2. *Big data:* Big data are given special mention because of today's digital environment. However, because small data are a component of big data, they are not excluded.
3. *Machine learning:* The PC is the learning machine, the *essential processing unit*, having the ability to learn without being explicitly programmed and the intelligence to find structure in the data. Moreover, the PC is essential for big data, as it can always do what it is explicitly programmed to do.

In view of these terms, the following data mining mnemonic can be formed:

Data Mining – Statistics + Big Data + Machine Learning and Lifting

Thus, *data mining* is defined today as *all of statistics and EDA for big and small data with the power of PC for the lifting of data and learning the structures within the data.* Explicitly referring to big and small data implies the process works equally well on both.

Again, in the spirit of EDA, it is prudent to parse the mnemonic equation. Lifting and learning require two different aspects of the data table. Lifting focuses on the rows of the data table and uses the capacity of the PC in terms of MIPS (million instructions per second), the speed in which explicitly programmed steps are executed. Calculating the average income of 1 million individuals is an example of PC lifting.

Learning focuses on the columns of the data table and the ability of the PC to find the structure within the columns without being explicitly programmed. Learning is more demanding on the PC than lifting in the same way that learning from books is always more demanding than merely lifting the books. **Identifying structure, such as the square root of $(a^2 + b^2)$,** is an example of PC learning.

When there are indicators that the population is not homogeneous (i.e., there are subpopulations or clusters), the PC has to learn the rows and their relationships to each other to identify the row structures. Thus, when lifting and learning of the rows are required in addition to learning within the columns, the PC must work exceptionally hard but can yield extraordinary results.

As presented here, my definition of *statistical data mining* is the EDA/statistics component with PC lifting. Further in the book, I elaborate on *machine-learning data mining*, which I define as PC learning without the EDA/statistics component.

References

1. Tukey, J.W., The future of data analysis, *Annals of Mathematical Statistics*, 33, 1–67, 1962.
2. Roy, S.N., *Some Aspects of Multivariate Analysis*, Wiley, New York, 1957.
3. Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.
4. Mosteller, F., and Tukey, J.W., *Data Analysis and Regression*, Addison-Wesley, Reading, MA, 1977.
5. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., *Understanding Robust and Exploratory Data Analysis*, Wiley, New York, 1983.
6. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., *Fundamentals of Exploratory Analysis of Variance*, Wiley, New York, 1991.
7. Chambers, M.J., Cleveland, W.S , Kleiner, B., and Tukey, P.A., *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1983.

8. du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H., *Graphical Exploratory Data Analysis*, Springer-Verlag, New York, 1986.
9. Jacoby, W.G., *Statistical Graphics for Visualizing Univariate and Bivariate Data*, Sage, Thousand Oaks, CA, 1997.
10. Jacoby, W.G., *Statistical Graphics for Visualizing Multivariate Data*, Sage, Thousand Oaks, CA, 1998.
11. Box, G.E.P., Science and statistics, *Journal of the American Statistical Association*, 71, 791–799, 1976.
12. Weiss, S.M., and Indurkhya, N., *Predictive Data Mining*, Morgan Kaufman., San Francisco, CA, 1998.
13. Dun, O.J., Multiple comparison among means, *Journal of the American Statistical Association*, 54, 52–64, 1961.
14. Ye, J., On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association*, 93, 120–131, 1998.
15. Samuel, A., Some studies in machine learning using the game of checkers, In Feigenbaum, E., and Feldman, J., Eds., *Computers and Thought*, McGraw-Hill, New York, 14–36, 1963.
16. Morgan, J.N., and Sonquist, J.A., Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, 58, 415–435, 1963.
17. Hunt, E., Marin, J., and Stone, P., *Experiments in Induction*, Academic Press, New York, 1966.
18. Quinlan, J.R., Discovering rules by induction from large collections of examples, In Mite, D., Ed., *Expert Systems in the Micro Electronic Age*, Edinburgh University Press, Edinburgh, UK, 143–159, 1979.