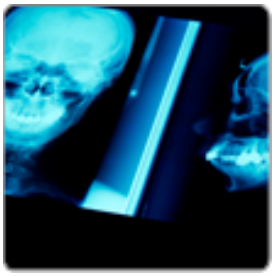
Oxford Bibliographies
Your Best Research Starts Here



Exploratory Data Analysis

Chong Ho Yu

LAST MODIFIED: 29 NOVEMBER 2017
DOI: 10.1093/OBO/9780199828340-0200

Introduction

Exploratory data analysis (EDA) is a strategy of data analysis that emphasizes maintaining an open mind to alternative possibilities. EDA is a philosophy or an attitude about how data analysis should be carried out, rather than being a fixed set of techniques. It is difficult to obtain a clear-cut answer from “messy” human phenomena, and thus the exploratory character of EDA is very suitable to psychological research. This research tradition was founded by John Tukey, who often relates EDA to detective work. In EDA, the role of the researcher is to explore the data in as many ways as possible until a plausible “story” emerges. A detective does not collect just any information. Instead, he or she collects clues related to the central question of the case. By the same token, EDA is not “fishing” or “torturing” the data set until it confesses. Rather, it is a systematic way to investigate relevant information from multiple perspectives. Tukey emphasizes the role of data analysis in research, rather than mathematics, statistics, and probability. Mathematics is secondary in the sense that it is a tool for understanding the data. Classical statistics aims to infer from the sample to the population based on the probability as the relative frequency in the long run. However, in many stages of inquiry, the working questions are non-probabilistic and the focal point should be the data at hand rather than the probabilistic inference in the long run. Hence, prematurely adopting a specific statistical model would hinder the researchers from considering different possible solutions. Because EDA endorses open-mindedness and triangulation, it is not a standalone approach. Rather, it complements traditional confirmatory data analysis (CDA) by generating a working hypothesis, as well as spotting outliers and assumption violations that might invalidate CDA. Additionally, it can also be operated with Bayesian statistics and resampling side by side. With the advent of high-power computers and voluminous data, many exploratory techniques have been developed in data science. These methods are known as data mining. Because it is tedious or even impossible to detect the data patterns when the sample size is extremely large or there are too many variables (this problem is called the “curse of dimensionality”), some data miners use machine learning to explore alternate routes for understanding the data. There are different taxonomies of EDA.

Traditionally, EDA comprises residual analysis, data re-expression, resistant procedures, and data visualization. With the advance of high-power computing and big data analytics, the alternate taxonomy is goal oriented, namely, clustering, variable screening, and pattern recognition.

General Overviews

There are several concise general overviews of EDA. Behrens 1996; Behrens 2000; and Behrens, et al. 2013 summarize the conceptual aspects and computational tools of EDA, illustrating how EDA can complement hypothesis testing, in the context of psychology. Traditionally, data explorers subscribe to the empiricist notion: “Let the data speak for themselves.” In response to this notion, de Mast and Trip 2007 and de Mast and Kemper 2009 go one step further by arguing that even if the surprising feature pops up, the problem being studied may still be far from resolution. Hence, background knowledge is essential to data interpretation. Additionally, they also present other major features of EDA for problem solving in quality management. Sometimes EDA is misunderstood as fishing—trying different procedures until finding a significant result. To debunk this common misconception, Jebb, et al. 2017 explains what EDA is and what EDA is not. The work of the founder of EDA, John Tukey, provides a historical review of the development of EDA in the form of Tukey 1977, Tukey 1980 (cited under Exploratory Data Analysis and Confirmatory Data Analysis), Tukey 1986a, Tukey 1986b, and Tukey 1988. Although some of Tukey’s ideas presented in these books are not entirely new (e.g., Francis Galton proposes nonparametric methods and quantiles during the 19th century; Arthur Lyon Bowley explores a prototypical stemplot and also uses a seven-point summary during the early 20th century), Tukey’s approach is still revolutionary given the fact that computing resources at his time were limited and thus computing-intensive data exploration and visualization were out of reach by researchers. Hence, some of Tukey’s proposed data visualization techniques, such as the stem-leaf plot and the five-point boxplot, could be done manually without a computer. More importantly, there is one overarching theme in all Tukey’s works: counteracting confirmation bias. Confirmation bias is a psychological flaw that humans tend to pay attention to data favoring their predetermined hypothesis while overlooking counterexamples. Tukey is well aware of this potential pitfall in confirmatory data analysis, though he didn’t explicitly name the term “conformation bias.” As a remedy, Tukey proposes an exploratory approach to urge researchers consider the otherwise. Confirmation bias is related to another psychological weakness: false sense of certainty. The traditional statistical approach that presents the finding in a confirmatory tone is embraced by the audience who prefers certainty to ambiguity. Tukey creates a paradigm shift by asserting that progress of statistics can only be made when analysts move away from certainty.

Behrens, J. T. 1996. Principles and procedures of exploratory data analysis. *Psychological Methods* 2:131–160.

Besides illustrating the computational tools of EDA, Behrens also emphasizes that the proper application of EDA is determined not by computation, but rather by the purpose of the procedure.

Behrens, J. T. 2000. Exploratory data analysis. In *Encyclopedia of psychology*. Edited by A. E. Kazdin, 303–305. New York: Oxford Univ. Press.

This article illustrates the 4Rs of classical EDA using S-Plus and Xlisp-Stat. It emphasized that the future directions of EDA is tied to computer technology.

Behrens, J. T., K. E. Dicerbo, N. Yel, and R. Levy. 2013. Exploratory data analysis. In *Handbook of psychology*. 2d ed. Vol. 2. Edited by J. A. Schinka, W. F. Velicer, and I. B. Weiner, 34–70. Hoboken, NJ: Wiley.

This book chapter is a comprehensive introduction to EDA, including the history and philosophy of EDA, the toolbox of EDA, computer software demonstrations, and future directions. It is noteworthy that the chapter also covers the legacy of John Tukey in the fields of regression diagnostics, robustness studies, and computer graphics for statistical use.

de Mast, J., and B. Kemper. 2009. Principles of exploratory data analysis in problem solving: What can we learn from a well-known case? *Quality Engineering* 21:366–375.

De Mast and Kemper use the example of John Snow's discovery of cholera outbreak to argue that data visualization alone is insufficient for problem solving. Rather, a comprehensive EDA should go beyond the empirical data by following these main steps: (1) display the data, (2) identify salient features, and (3) interpret salient features.

de Mast, J., and A. Trip. 2007. Exploratory data analysis in quality improvement projects. *Journal of Quality Technology* 39:301–311.

De Mast and Trip outline the major principles of EDA, including the purpose of EDA, data visualization, identification and interpretation of salient features, the role of automated procedures, and integration of EDA and confirmatory data analysis (CDA).

Jebb, A., S. Parrigon, and S. E. Woo. 2017. Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review* 27:265–276.

This article is concerned with the principle and philosophy of EDA. These authors argue that due to the natural uncertainty of data patterns, EDA should be integrated with replication-based procedures, such as cross-validation. Additionally, they argued against fishing, data dredging, or *p*-hacking.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

This is the seminal work of EDA. It is noteworthy that Tukey wrote the book before the age of high-power computing, and thus certain graphing techniques are done by pencil and paper, such as the stem-and-leaf plot. Readers should focus on the conceptual aspect, not the computational procedures, of Tukey's EDA.

Tukey, J. W. 1986a. *Philosophy and principles of data analysis: 1949–1964*. Vol. 3 of *The collected works of John W. Tukey*. Edited by L. V. Jones. Monterey, CA: Wadsworth & Brooks.

In this volume, Tukey argues that data analysis should be bottom-up (data driven) rather than top-down (model based). However, many materials are repetitive.

Tukey, J. W. 1986b. *Philosophy and principles of data analysis: 1965–1986*. Vol. 4 of *The collected works of John W. Tukey*. Edited by L. V. Jones. Monterey, CA: Wadsworth & Brooks.

The content of Volume 4 is similar to that of Volume 3. Tukey emphasizes that statistics is an empirically based discipline and therefore data analysts must be mentally prepared for surprising results. Tukey also points out that many statistical models are constructed even though assumptions are violated.

Tukey, J. W. 1988. *Graphics*. Vol. 5 of *The collected works of John W. Tukey*. Edited by W. S. Cleveland. Pacific Grove, CA: Wadsworth.

This volume focuses on graphical methods of EDA. This book should be used for understanding the history of EDA and data visualization. In the early 21st century, much more powerful graphing techniques are available in different software packages.

Philosophies of EDA

The philosophy of Tukey's EDA is similar to empiricism. Simply put, Tukey 1986 endorses a data-driven, bottom-up approach to data analysis. According to Dempster 2002, the philosophy of Tukey encompasses both mathematical statistics and empirical science. The former leads to quantitative-based conclusions, whereas the latter results in qualitative decisions. EDA plays a role in the latter. Currently, there are at least two competing philosophies of EDA. Behrens, et al. 2013 and Yu 2006 regard the abductive logic as the philosophical foundation of EDA. Contrarily, Haig 2005, Haig 2014, and Haig 2015 advocate Tukeyan philosophy of data analysis. Abductive reasoning is a logical system introduced by American philosopher Charles Sanders Peirce. Abduction is usually formulated in the following mode: The surprising phenomenon, X, is observed. Among hypotheses A, B, and C, A is capable of explaining X. Hence, there is a reason to pursue A. Yu 2006 and Yu, et al. 2008 argue that the preceding abductive logic shares a common ground with exploratory research. In EDA, after observing some surprising facts, researchers propose alternate explanations by analogical reasoning. Although there may be more than one convincing pattern, we "abduct" only those that are more plausible for subsequent study. Josephson and Josephson 1996 relates abductive reasoning to detective work. Detectives collect related data about suspects and circumstances based on keen powers of observation. In this way, the logic of abduction is in line with EDA. Behrens, et al. 2013 goes even further to follow Josephson and Josephson 1996 in its interpretation of abduction: D is a collection of data (given facts and observations). Hypothesis H explains D (would if true, explain D). No other hypothesis explains D as well as H does. Therefore, H is probably correct. However, Haig 2005, Haig 2014, and Haig 2015 argue that this scheme does not represent Peirce's abductive reasoning and the spirit of EDA. Rather, selecting the best explanation out of rival hypotheses is actually considered inference to the best explanation (IBE). Haig argues that what the analyst should try to explain is the phenomena, which entail recurrent general features, not idiosyncratic data varying from context to context. Abducting data patterns is just one of many components of data analysis. The abductive approach regards EDA as a descriptive pattern detection process, which is a precursor to the inductive generalizations involved in phenomena detection. The ultimate goal of EDA should be detection of new empirical phenomena, therefore abduction cannot be the core philosophy of EDA.

Behrens, J. T., K. E. Dicerbo, N. Yel, and R. Levy. 2013. Exploratory data analysis. In *Handbook of psychology*. 2d ed. Vol. 2. Edited by J. A. Schinka, W. F. Velicer, and I. B. Weiner, 34–70. Hoboken, NJ: Wiley.

This chapter is a revision of an earlier version (J. T. Behrens and C. H. Yu, "Exploratory Data Analysis," in *Handbook of Psychology, Volume 2: Research Methods in Psychology*, eds. J. A. Schinka and W. F. Velicer [Hoboken, NJ: John Wiley, 2003], 33–64). In the first version, abduction is presented as a process of discovery, and in the second version, it moves toward confirmation or IBE.

Dempster, A. P. 2002. John W. Tukey as "philosopher." *The Annals of Statistics* 30:1619–1628.

Dempster summarizes the philosophical views of John Tukey in different domains of data analysis, including EDA, Bayesian inferences, and Fisher's fiducial argument.

Haig, B. D. 2005. An abductive theory of scientific method. *Psychological Methods* 10:371–388.

In this article, Haig suggests that Behrens and Yu 2003's approach to abduction and EDA conflates data description and theoretical explanation. He also proposes methods for integrating EDA and confirmatory data analysis (CDA). Specifically, after EDA suggests potentially interesting data patterns, computer-intensive resampling methods, such as the bootstrap, the jackknife, and cross-validation, should be employed to verify the stability of the emergent data patterns.

Haig, B. D. 2014. *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, MA: MIT.

In this book Haig argues that data analysis should be concerned with the phenomena rather than data. The former are relatively stable, while the latter are ephemeral and pliable.

Haig, B. D. 2015. Commentary: Exploratory data analysis. *Frontiers in Psychology* 6:1247.

In this concise commentary, Haig argues that the purpose of abduction should be generating and evaluating potential explanatory hypotheses and should not rely on inference to determine the best explanation.

Josephson, J. R., and S. G. Josephson, eds. 1996. *Abductive inference: Computation, philosophy, technology*. Cambridge, UK: Cambridge Univ. Press.

While the foundational concept of this book is based on abductive inferences, it goes one step further by endorsing inference to the most plausible explanation from a given data set. This is an interdisciplinary inquiry, crossing the boundaries between AI, cognitive science, and philosophy of science.

Tukey, J. W. 1986. *Philosophy and principles of data analysis: 1965–1986*. Vol. 4 of *The collected works of John W. Tukey*. Edited by L. V. Jones. Monterey, CA: Wadsworth & Brooks.

In this volume, Tukey advocates an empirical, data-driven approach to data analysis. In his view, blindly trusting statistical figures without verifying the data could lead to erroneous conclusions.

Yu, C. H. 2006. *Philosophical foundations of quantitative research methodology*. Lanham, MD: Univ. Press of America.

Yu uses several examples (Bayesian Inference Network, mixed method) to explain the role of abduction, deduction, and induction in quantitative data analysis. He argues that there are limitations in all logical systems, and hence all three should be employed in tandem as a recursive process.

Yu, C. H., S. DiGangi, and A. Jannasch-Pennell. 2008. The role of abductive reasoning in cognitive-based assessment. *Elementary Education Online* 7.2: 310–322.

Using examples of cognitive-based assessments, Yu, DiGangi, and Jannasch-Pennell explain the essence of abductive reasoning, including exploration of alternate explanations, converse reasoning, and analogical reasoning.

Relationships between EDA and Other Schools of Data Analysis

EDA is complementary to other schools of statistical methodologies. Specifically, EDA and confirmatory data analysis are not at odds with each other; rather, EDA could be employed to formulate hypotheses for confirmatory data analysis (CDA). Further, both EDA and the Bayesian approach aim to discover data patterns by asking “what-if” questions. EDA and resampling also share a common ground in the sense that they both go beyond a single analysis. Further, the exploratory character of EDA is inherited into data mining. As a matter of fact, more and more psychological researchers employ multiple methodologies in order to gain a holistic view of the phenomenon being studied.

Exploratory Data Analysis and Confirmatory Data Analysis

EDA was developed in response to the limitations and widespread misuse of confirmatory data analysis (CDA) in the form of hypothesis testing. In CDA a strong theory or hypothesis is formulated first and then the researcher collects data to confirm or disconfirm the hypothesis. The logic is this: Given the hypothesis, how often can the data at hand be observed [P(D|H)]? The researcher deduces the conclusion in these logical steps: If Hypothesis A is correct, then Data B is expected; B is observed; Hence, A is supported. However, this line of reasoning commits the fallacy of affirming

the consequent. Specifically, there are always alternate explanations for the phenomenon under study, and thus affirming the validity of the theory in this way alone is questionable. For example, if it is raining, the ground is wet. But it is fallacious to assert that because the ground is wet, it must be raining. For exploratory data analysts, the more central question is this: Given the data, which theory or hypothesis is the most plausible explanation in comparison to rival theories [P(H|D)]? Thus, CDA is said to be hypothesis driven, whereas EDA is considered data driven. To understand this difference in another way, CDA is regarded as tool driven, whereas EDA is problem oriented. In addition, hypothesis testing requires parametric assumptions, but very often researchers overlook compliance of those assumptions. To rectify the situation, EDA should be employed to perform diagnostics. Tukey 1980 is not opposed to CDA; rather, Tukey asserts that data analysts need both EDA and CDA. Tukey found that students who have never been exposed to CDA seem to learn EDA more easily. EDA is compared to detective work, while CDA is compared to that of judges and the judicial system. In the former, the data "detective" searches for clues, whereas in the latter, the analyst evaluates the strength and weaknesses of the evidence. For reviewing the shortcomings of hypothesis testing, consult Cumming 2014, Loftus 1996, Nuzzo 2014, Rosnow and Rosenthal 1989, and Schmidt and Hunter 2015. For evaluating both sides (pros and cons of hypothesis testing), consult Harlow, et al. 1997 and Task Force on Statistical Inference 1999. For example, Cumming 2014 argues that the dichotomous answer yielded from CDA leads to illusory certainty. Loftus 1996 mocks that rejecting a null hypothesis is like rejecting the proposition that the moon is made of green cheese. Nuzzo 2014 contends that usually CDA results are not replicable. The authors of Rosnow and Rosenthal 1989 state their objection against blindly using .05 as the cutoff in hypothesis testing in a humorous way: "God loves the .06 nearly as much as .05."

Cumming, G. 2014. The new statistics: Why and how. *Psychological Science* 25:7–29.

Cummings argues that hypothesis testing yields a dichotomous answer, misleading researchers to see the world as black and white. He warned researchers against using inappropriate data-analytic practices and encouraged replication of studies.

Harlow, L., S. Mulaik, and J. Steiger. 1997. *What if there were no significance tests?* Mahwah, NJ: LEA.

This anthology consists of 15 articles. The first one is an introduction to the debate, and the rest present diverse perspectives to the issue. It is worth noting that one author suggested abduction as the foundational reasoning for good science.

Loftus, G. R. 1996. Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science* 5:161–170.

Loftus argues that hypothesis testing ignores two issues that are generally much more interesting, important, and relevant: What is the pattern of population means over conditions? What are the magnitudes of various variability measures?

Nuzzo, R. 2014. Scientific method: Statistical errors. *Nature*, 12 February.

In this article, Nuzzo emphasizes that Fisher was not intended to set the *p* value as the definite test, yet many researchers are obsessed with *p*-hacking or snooping. In addition, a significant result is commonly misinterpreted as how "right" the conclusion is. Nuzzo also points out that significant results are often not replicable.

Rosnow, R. L., and R. Rosenthal. 1989. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* 44:1276–1284.

In this seminal paper, Rosnow and Rosenthal explain the shortcomings of conformational claims of truth based on statistical procedures in the perspective of philosophy of science. They also point out why researchers refuse to correct these problems.

Schmidt, F., and J. Hunter. 2015. *Methods of meta-analysis: Correcting error and bias in research findings*. 3d ed. Thousand Oaks, CA: SAGE.

In this book Schmidt and Hunter describe that there is a fundamental circularity to hypothesis testing. If the researcher does not know whether the null is true or not, then the researcher cannot tell whether the error is tied to Type I or Type II. But if the researcher knows that the null is true, then performing the test is unnecessary.

Task Force on Statistical Inference. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* 54:594–604.

The task force was formed in response to the widespread discontent of use of hypothesis testing in psychological research. The task force pointed out certain limitations of hypothesis testing but didn't recommend abandoning it. Rather, it suggested researchers to keep an open mind to alternate methodologies.

Tukey, J. W. 1980. We need both exploratory and confirmatory. *American Statisticians* 34:23–25.

This article is conservational rather than research based. Based on his experience, Tukey argues that researchers need both EDA and CDA. However, he did not specify concrete procedures regarding the integration of the two approaches.

EDA and Bayesian Statistics

At first glance, EDA and Bayesian statistics seem to belong to two vastly different paradigms. EDA is model free, non-probabilistic, and non-inferential, whereas the Bayesian approach is a form of probabilistic inference. Tukey 1986 is skeptical of Bayesian inferences, as John Tukey was opposed to any formalized and unified scientific method. In addition, he faulted the Bayesian approach as using “uninformative” prior probability, lacking robustness, and also leading the analyst astray by overlooking important insights and systematic errors. Nevertheless, both EDA and Bayesianism share several common grounds. First, EDA is data driven, while Bayesians are interested in learning the probability of a certain explanation given the data. Second, EDA and Bayesianism are iterative in essence. Going back and forth between the data and alternative explanations is the trademark of EDA. By the same token, Bayesians keep updating the posterior probability with additional information. Hence, several researchers propose a synthesis of EDA and Bayesianism. Good 1983 subscribes to EDA’s goal of pattern recognition, and therefore argues that starting a study with subjective prior probability, as proposed by Bayesians, is helpful to determine whether the data patterns are sensible given the background information. Gelman 2003 and Gelman 2004 argue that EDA techniques, such as data visualization, can facilitate model checking. However, sometimes the analyst might draw the wrong conclusion without a reference distribution in the statistical graph. In addition, the classical EDA methods are based on simple models, such as additive fits and the Poisson distribution. In complex modeling, Bayesian methods can rectify the preceding situation by constructing reference distributions based on simulations or replications. Further, sensitivity analysis and model averaging used by Bayesians are congruent to the principle of EDA because by experimenting with multiple models, the Bayesian is asking EDA-styled “what if” questions. For an introduction to Bayesian statistics, please consult Lee 2012.

Gelman, A. 2003. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 71:369–382.

Gelman explains how posterior predictive simulations can be used to generate reference distributions for EDA graphs. He argued that in this way, non-probabilistic EDA can fit into the probability-modeling paradigm.

Gelman, A. 2004. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics* 13:755–779.

This article is an extension of Gelman 2003. Gelman continued to argue that Bayesianism could augment EDA, especially in analyzing complex data sets.

Good, I. J. 1983. The philosophy of exploratory data analysis. *Philosophy of Science* 50:283–295.

In addition to the relationship between EDA and Bayesianism, Good covers a wide variety of topics related to EDA, such as data reduction, Francis Bacon's philosophy of science, the automatic formulation of hypotheses, successive deepening of hypotheses, and neurophysiology.

Lee, P. 2012. *Bayesian statistics: An introduction*. New York: Wiley.

The central tenet of this book is that there is no unconditional probability. The book begins by addressing the basic concepts of probability and then gradually introducing more advanced concepts, including Gibbs sampling, approximate Bayesian computation, and Reversible Jump Markov Chain Monte Carlo (RJMCMC).

Tukey, J. W. 1986. *Philosophy and principles of data analysis: 1965–1986*. Vol. 4 of *The collected works of John W. Tukey*. Edited by L. V. Jones. Monterey, CA: Wadsworth & Brooks.

In this volume Tukey is critical of Bayesian probability. He mocks the Bayesian approach, suggesting that it could serve data analysts well when we do not wish to draw on any source of information other than the data at hand.

EDA and Resampling

EDA emphasizes model checking rather than grounding the conclusion on a single analysis. Hence, Yu 2010 and Haig 2013 argue that EDA and resampling should go hand in hand. Interestingly, one of the major resampling methods, Jackknife, was invented by the founder of EDA, John Tukey. Resampling is a way of model checking by systematically reusing existing data. There are four types of resampling: First, randomization exact test was invented by the author of Fisher 1960. In an exact test, data are shuffled across groups to generate an empirical sampling distribution. Afterward, the exact p value is computed by comparing the observed statistics at hand and the simulated sampling

distribution. Second, cross-validation was developed in Krus and Fuller 1982 to remediate the problem of data unreliability. Third, jackknife was developed in Tukey 1958. John Tukey attempted to use jackknife to explore how a model is influenced by subsets of observations when outliers are present. The name “Jackknife” was coined by Tukey to imply that the method is an all-purpose statistical tool. Last, bootstrap was developed by Efron and Tibshirani 1993. “Bootstrap” means that one available sample gives rise to many others by resampling (a concept reminiscent of pulling yourself up by your own bootstrap). With the advance of data mining, which is an extension of EDA, certain resampling techniques, such as cross-validation and bootstrapping, had been routinely used in big data analytics for model checking, including the decision tree and the bootstrap forest. In cross-validation, the sample is divided into the training set, the testing set, and the validation set. The initial model based on the training set might be overfitted. As a remedy, the subsequent models are adjusted when there is discrepancy between the proposed model and the remaining data. In the bootstrap forest, many subsets are generated independently by resampling with replacement from the original sample and independent analyses are run with these subsamples. Afterward, the computer algorithm assembles these resampled results together by averaging them out. By generating alternate models in resampling and drawing the conclusion based on multiple scenarios, the analyst appeals to counterfactual reasoning: “What would have happened to Y if X were not present?” In this sense, EDA and resampling are fully compatible. For more information about cross-validation and bootstrap forest, please consult Breiman 1996 and Zaman and Hirose 2011. For a concise introduction to resampling, please consult Yu 2003 and Yu 2007.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–140.

This is a seminal work on how to generate and converge multiple models by resampling. The presentation is mathematical. Readers need a background on resampling to follow his logic.

Efron, B., and R. J. Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

This is a seminal work on bootstrapping. Departing from theoretical distributions, Efron and Tibshirani demonstrate how one can use empirical distributions to estimate bias. The presentation is accessible to non-statisticians.

Fisher, R. A. 1960. *The design of experiments*. 7th ed. New York: Hafner.

In this classic, Fisher proposes using randomization tests to check normal theory tests. It is helpful to use this source to understand the historical root of randomization tests. However, in the early 21st century, exact tests are implemented differently.

Haig, B. D. 2013. Detecting psychological phenomena: Taking bottom-up research seriously. *American Journal of Psychology* 126:135–155.

In this article, Haig advocates using computing-intensive methods to verify the phenomenon detected by EDA.

Krus, D. J. and E. A. Fuller. 1982. Computer-assisted multicross-validation in regression analysis. *Educational and Psychological Measurement* 42:187–193.

The purpose of this article is to illustrate how multicross-validation can be used to estimate shrinkage of the multiple coefficient of correlation. The technique illustrated in this article is built upon double cross-validation, which is an expansion of single cross-validation. Readers need the background information of cross-validation to follow the argument.

Tukey, J. W. 1958. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics* 29:614.

This study examines the burnout levels of research assistants in Ondokuz Mayis University. The sample size was as small as eleven, and thus Tukey employed the jackknife method to enhance stability of the results. This paper shows that Tukey was not opposed to CDA techniques.

Yu, C. H. 2003. Resampling methods: Concepts, applications, and justification. *Practical Assessment Research and Evaluation* 8.19.

This paper is a concise introduction to all four types of resampling. Yu also discusses the pros and cons of resampling and the philosophy of resampling: counterfactual reasoning.

Yu, C. H. 2007. Resampling: A conceptual and procedural introduction. In *Best practices in quantitative methods*. Edited by J. Osborne, 283–298. Los Angeles: SAGE.

In this book chapter, Yu argues that resampling should be defined as a systematic way of reusing observations. Exact test and bootstrapping are approaches of resampling with replacement, whereas cross-validation and jackknife do not reuse observations. Thus, this book chapter focuses on the former set of resampling through a historical perspective by using Fisher's example "lady tasting tea," and Efron and Tibshirani's law school's data set.

Yu, C. H. 2010. Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research* 3.1: 9–22.

In this article Yu uses cross-validation to illustrate how EDA and resampling could complement each other. He also argued that resampling addresses two important issues in research, namely, generalization across samples and under-determination of theory by evidence.

Zaman, M. F., and H. Hirose. 2011. Classification performance of bagging and boosting type ensemble methods with small training sets. *New Generation Computing* 29:277–292.

In this article Zaman and Hirose examine the trade-off of bias and variance in different resampling-based ensemble methods, including bootstrap forest and boosted tree. The trade-off of bias and variance is one of the central issues of most statistical modeling.

EDA and Data Mining/Big Data Analytics

With the availability of high-power computing and big data sets, modern data mining overlaps with classical EDA. Data mining is a cluster of discovery techniques, including classification trees, neural networks, and K-mean clustering (Larose 2014, Han and Kamber 2011). Because data mining is usually employed for analyzing big data sets, in which the sample size could range from thousands to even millions, it is also known as big data analytics. In addition, data mining is data driven in the sense that it automatically extracts useful information from immense quantities of structured or unstructured data, and therefore it is also called data-driven science or simply data science. Last, due to its exploratory character, it is also named knowledge discovery in databases (KDD). EDA and data mining share many common grounds. First, in alignment with EDA, data mining does not start from a strong preconception, a specific question, or a narrow hypothesis. Rather, it aims to detect patterns that are already present in the data. Second, both heavily rely on graphing techniques to understand data patterns. Hence, Luan 2002 and Myatt and Johnson 2014 view data mining as an extension of EDA. In other words, EDA is considered a precursor of data mining. Third, it is a common practice for both data explorers and data miners to verify the model by resampling techniques, such as cross-validation or bootstrapping. However, there are also dissimilarities between them. First, when EDA was developed during the 1970s, computers were much less powerful and big data sets were rare. Therefore, unlike data mining, classical EDA techniques were suited to fairly simple and small data sets. Second, one of the major features of data mining is machine learning. According to Kelleher, et al. 2015, machine learning, which is derived from artificial intelligence, enables computers to learn from the data in order to fine-tune the statistical model. This machine learning process is automated, but in EDA, model checking is often completed manually by the data explorer. Further, some automated data mining methods, such as neural networks, are operated in a “black box.” However, traditional data explorers dislike the lack of

transparency and interpretability because the practice of handing over human judgment to the computer is not any better than blindly following the alpha level as 0.5 in CDA.

Han, J., and M. Kamber. 2011. *Data mining: Concepts and techniques*. 2d ed. Waltham, MA: Morgan Kaufmann.

This book, in alignment with the principle of EDA, emphasizes that the analyst should be getting to know the data. It thoroughly discusses topics related to data exploration, such as data objects, data attribute types, and data visualization, as well as measuring data similarity and dissimilarity.

Kelleher, J. D., B. M. Namee, and A. D'Arcy. 2015. *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge, MA: MIT.

This is an introductory book of machine learning. Although the presentation is mathematical and computational, it is augmented with many examples.

Larose, D. 2014. *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: Wiley.

In this book, EDA is subsumed under data mining. There is a chapter on EDA, covering detection of anomaly and investigation of variable association by discovery techniques. Some chapters review traditional statistical procedures, such as t-test, Chi-square analysis, and regression. The rest are concerned with data mining, such as clustering, neural networks, and Kohonen networks.

Luan, J. 2002. Data mining and its applications in higher education. In *Knowledge management: Building a competitive advantage in higher education*. Edited by A. Serban and J. Luan, 17–36. San Francisco: Josey-Bass.

Before this article was written, data mining had been applied to business intelligence. However, at that time, use of data mining in academia was rare. This is one of the first papers to use a case study to describe how data mining can benefit research in higher education.

Myatt, G., and W. Johnson. 2014. *Making sense of data I: A practical guide to exploratory data analysis and data mining*. Hoboken, NJ: Wiley.

This book is a general introduction to EDA and data mining techniques, including data preparation, data visualization, and clustering, as well as classification and regression trees.

Shmueli, G., P. Bruce, and N. Patel. 2016. *Data mining for business analytics: Concepts, techniques, and applications with JMP Pro*. Hoboken, NJ: John Wiley.

This book covers a variety of data exploration and mining techniques, such as k-mean clustering, naïve Bayes classifier, ensemble methods, and so on. Examples are illustrated using JMP Pro, a data exploration and visualization software package created by SAS Institute. Another version of this book demonstrates applications of data mining with Microsoft Excel and XLMiner.

Taxonomies of EDA

Different researchers formulated different taxonomies of EDA. Velleman and Hoaglin 1981 outlines four basic elements of EDA as 4Rs: residual, re-expression (data transformation), resistant, and revelation (also known as display or data visualization). First, for residual analysis, EDA follows the formula that *data* = *fit* + *residual* or *data* = *model* + *error*. The fit or the model is the expected values of the data, whereas the residual or the error is the values that deviate from that expected value. By examining the residuals, the researcher can assess the model's adequacy. Second, when the distribution is skewed or the data structure obscures the pattern, the data could be re-expressed or rescaled in order to improve interpretability. Typical examples of data transformation include using log transformation or inverse probability transformation to normalize a distribution, using square root transformation to stabilize variances, and using logarithmic transformation to linearize a trend. Third, parametric tests are based on the mean estimation, which is sensitive to outliers or skewed distributions. As a remedy, resistant estimators are usually used in EDA. Last, graphing is a powerful tool for revealing hidden patterns and relationships among variables. According to NIST Semantech's website (What is EDA?), EDA entails a variety of techniques for accomplishing the following tasks: maximizing insight, uncovering underlying structure, extracting important variables, detecting outliers and anomalies, testing underlying assumptions, developing parsimonious models, and determining optimal factor settings. Although the preceding framework provides researchers with helpful guidelines for data analysis, some of these elements are no longer as important as before, due to the emergence of new methods and convergence between EDA and other methodologies, such as data mining and resampling. While checking underlying assumptions, transforming data, and spotting outlier play an important role in conventional EDA, many new EDA techniques based upon data mining are non-parametric in nature. In addition, some new procedures can automatically transform the data, and some are immune against outliers. Yu 2010 argues that in the preceding taxonomies, the characteristics of EDA are tied to both the attributes of the data (distribution, variability, linearity, outliers, measurement scales, etc.) and the final goals (detecting

clusters, screening variables, and unearthing complex relationships). In light of data mining and resampling, Yu 2010 proposes a goal-oriented taxonomy of EDA: finding clustering patterns in the data, screening variables out of many potential factors, and recognizing patterns and relationships.

Velleman, P. F., and D. C. Hoaglin. 1981. *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury.

This book illustrates the 4Rs of EDA with programs written in BASIC and Fortran. The taxonomy and techniques are based on the computing resources available at that time. It is helpful to understand the historical development of EDA; however, the techniques and computer programs are outdated.

What is EDA? NIST Semantech.

This website is a concise introduction to the principle and the taxonomy of EDA. The materials are written for engineers but it is accessible to researchers of different disciplines.

Yu, C. H. 2010. Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research* 3.1: 9–22.

In this article Yu points out the limitations of conventional views of EDA. Many new discovery methods are assumption free and immune against outliers, and therefore some of the 4Rs in EDA are less important. Recursive partition trees, neural networks, and TwoStep clustering are used for illustration.

Classical Taxonomy of EDA

Classical taxonomy of EDA is based upon the framework of Tukey's research. However, with the advance of better statistical methods, some classic components become less important. For example, many modern techniques are immune against outliers and thus classical resistance procedures might not be necessary. Nonetheless, data visualization is still at the core of exploratory research. It is important to point out that development of data visualization techniques heavily relies on psychological research on cognition and human interface.

Residual and Model Fit

EDA follows the notion that *data = fit + residual* or *data = model + error*. In terms of data visualization, this equation can be re-expressed as *data = smooth + rough*. The fit or the model is the expected values of the data. The residual or the error is the value that deviates from that expected value. The central tenet of residual analysis is that no model is perfect and there are always some misfits between the data and the model. Hence, the researcher should not stop at one-single test even though the *p* value and other statistical figures look impressive. Rather, data analysis should be an iterative process of model building and model checking. In the past, this iterative process was performed manually by the analyst, such as the two-way fit approach. Today, machine learning algorithms automate this process. Boosting, also known as the boosted tree, is a good example of an automated iteration. Mosteller and Tukey 1977 introduces the two-way fit for crosstab analysis, which was performed by iteratively estimating the row and column effects. In the process, the sum of those estimates is utilized to predict the cell values and their residuals. This cycle is repeated until no additional improvement can be made by additional adjustment. The logic of two-way fit is analogous to item response theory, in which the parameters are calibrated iteratively based on the residuals until the data and the model converge. Like the two-way fit, boosting is also an iterative method in which the previous model informs the next model so that improvement can be made in subsequent modeling. But unlike the two-way fit, the boosted tree is automated and adaptive by itself (Freund and Schapire 1997, Optiz and Maclin 1999). Boosting is so named because of gradient improvement by learning mistakes in previous steps. In boosting, all observations are assigned the same weight at the beginning. After an initial model is proposed, residuals and misfits are identified. The cases with a high residual or misclassified observations are assigned a heavier weight so that they are more likely to be selected in the next model. In the subsequent steps, each model is constantly revised in an attempt to reduce the residuals or the misclassification rate. Ultimately, the final model is created by a majority vote as the best solutions are kept and the worst ones are eliminated.

Freund, Y., and R. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55.1: 119–139.

This is a seminal work in boosting. Freund and Schapire develop a learning algorithm that does not require any prior knowledge about the performance of the weak learning algorithm. The presentation is highly technical and thus a strong foundation in mathematical statistics is necessary to comprehend the content.

Mosteller, F., and J. W. Tukey. 1977. *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

This book covers both philosophy and techniques of EDA. Although the philosophy of EDA is timeless, many techniques mentioned in the book are obsolete because they were developed in the

era when most calculations were done by hand. This book can be used for studying the history and philosophy of EDA.

Optiz, D., and R. Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* 11:169–198.

This is a milestone paper that identified the merits and shortcomings of different resampling-based ensemble methods, including bagging and boosting. It is interesting that on some occasions, the result of boosting is less accurate than that of a single neural network. Prior knowledge of resampling and data mining is necessary to understand the illustration.

Re-expression and Data Transformation

Traditionally, researchers follow the advice in Stevens 1951 to categorize data into four levels: nominal, ordinal, interval, and ratio. As an alternative, Mosteller and Tukey 1977 classifies data as (a) amounts and counts, (b) balances (numbers that can be positive or negative with no bound), (c) counted fractions (ratios of counts), (d) ranks, and (e) grades (nominal categories). In Tukey's view, data of common amounts and counts should be re-expressed toward a Gaussian distribution as well as move up and down the "ladder of re-expression." This section provides some examples of data transformation. The first example is normalizing the distribution. Non-normal data violate the assumption of parametric tests and thus a transformation is advisable. The common procedures for normalization are log transformation and inverse probability transformation. Sometime this method "saves" the necessary outliers. For example, the number of patents held by residents is an extremely skewed distribution because the United States and Japan hold the vast majority of world's patents. Instead of removing two these important countries from the data set, a log transformation can convert the skewed distribution to a normal one. Stabilize the variances is another good example of re-expression. Data with unequal variances are also detrimental to parametric tests. A typical example of variance stabilizing transformation is square root transformation. Linearization is also a common re-expression technique. Regression analysis requires the assumption of linearity. When the data show a curvilinear relationship, the researcher can "straighten" the data by linearization. However, by doing so, the model might distort the true relationship. Orthogonalizing collinear variables is less common. In multiple regression, lack of independence between predictors could make the model unstable. In terms of hyperspace, the vectors representing these variables are non-orthogonal, as depicted in Rodgers, et al. 1984. To rectify the situation, the variables can be orthogonalized by centering the scores or using the Gram–Schmidt process. Osborne 2002 advises that transformation should be used appropriately; many transformations reduce non-normality by changing the spacing between data points, but this raises issues in data interpretation. If transformations are done correctly, all data points should remain in the same relative order as prior to the transformation. In this way the interpretation of the scores is not affected. But it might be

problematic if the original variables were meant to be interpreted directly, such as annual income and age. After the transformations, the new variables may become significantly more complex to interpret.

Mosteller, F., and J. W. Tukey. 1977. *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Stevens argues that the choice of statistical test must match the level of measurement scale in the data. However, some statisticians identified exceptions. Hence, in this book Mosteller and Tukey developed their own taxonomy of measurement.

Osborne, J. W. 2002. Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation* 8.6: 1–7.

Many statistical texts provide minimal advice on proper use of data transformation. To rectify this situation, Osborne points out that the three most commonly used data transformation methods, namely, square root, log, and inverse, might make interpretation more complex. Osborne is not opposed to data transformation. Rather, he promoted thoughtful use of transformation.

Rodgers, J. L., W. A. Nicewander, and L. Toothaker. 1984. Linearly dependent, orthogonal, and uncorrelated variables. *American Statistician* 38:133–134.

This article provides the readers with a graphical explanation of the differences between dependence, correlation, and non-orthogonality.

Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. In *Handbook of experiment psychology*. Edited by S. S. Stevens, 1–49. New York: John Wiley.

In this seminal paper, Stevens creates the taxonomy of measurement (nominal, ordinal, interval, and ratio), which is still widely used in the early 21st century. The article was written at the time when the status of psychology as a science was controversial. Stevens attempted to classify data in a more precise fashion in order to facilitate proper use of statistical procedures.

Resistance Procedures

There is a subtle difference between “resistance” and “robustness” although the two terms are usually used interchangeably. EDA is more concerned with resistance, while hypothesis testing is interested in robustness. Resistance is about being immune to outliers, while robustness is about being immune to assumption violations. In the former, the goal is to obtain a data summary, while in the latter, the goal is to make a probabilistic inference. In EDA there are three major approaches to improve resistance. First, the analyst can use rank-based measures and absolute values instead of measures based on the mean and the variance. Trimean is a good example of this. In this approach, the central tendency is based on the arithmetic average of the values of the first quartile, the third quartile, and the median counted twice. To measure dispersion, the median absolute deviation (MAD) or the inter-quartile range (IQR), which is commonly depicted in a boxplot, can be employed. Second, the analyst can use a more centrally located score as well as down-weight the extreme values, such as the Winsorized mean. In this approach, the extreme scores are pulled back to the majority of the data. Usually 10 percent to 25 percent of the two tails of the distribution are replaced. Third, the analyst can simply remove outliers, or the tails of the distribution. This can be done by using trimmed mean, in which data points past a certain threshold are excluded from mean calculation. However, Wilcox and Keselman 2003 points out that outlier detection methods based on means and variances might not be capable of detecting outliers. Moreover, after outliers are removed, the remaining cases are no longer independent under random sampling. The remedy for the first problem is to use a median- and quantile-based method to identify outliers. To be more specific, outliers can be defined as observations outside 1.5*IQR. Another strategy of trimming data is to set a predetermined percentage rather than checking for outliers. Rosenberger and Gasko 1983 recommends discarding 20 percent of the data located in the two tails of the distribution and 25 percent when the sample size is small. For the classical approach to outlier detection, please consult Stevens 2007. For the data mining approach to outlier detection, please consult Gebremeskel, et al. 2016.

Gebremeskel, G. B., C. Yi, Z. He, and D. Haile. 2016. Combined data mining techniques based patient data outlier detection for healthcare safety. *International Journal of Intelligent Computing and Cybernetics* 9.1: 42–68.

This article introduces the clustering-based approach to outlier detection for large data sets. This method classifies data into the same groups based on common attributes, and hence outliers are defined as observations that do not belong to any cluster.

Rosenberger, J. L., and M. Gasko. 1983. Comparing location estimators: Trimmed means, medians, and trimean. In *Understanding robust and exploratory data analysis*. Edited by D. Hoaglin, F. Mosteller, and J. Tukey, 297–336. New York: Wiley.

Tukey is one of the editors of this book. This book chapter explains the rationale of the 20 percent trimmed mean rule, which aligns with resistance methods within EDA.

Stevens, J. 2007. *Intermediate statistics: A modern approach*. New York: Lawrence Erlbaum.

This highly accessible book explains the difference between outliers and influential points, and also major classical approaches to outlier detection.

Wilcox, R. R., and H. J. Keselman. 2003. Modern robust data analysis methods: Measures of central tendency. *Psychological Methods* 8:254–274.

Although this paper focuses on robust methods that are intended to countermeasure violations of parametric assumptions (e.g., normality), Wilcox and Keselman also cover resistance estimators, such as trimmed mean and Winsorized variance.

Revelation and Data Visualization

In alignment with EDA, Cleveland 1993 argues that visualization stresses a penetrating look at the data structure. Data visualization is the process of exploring and displaying data in a manner that builds a visual analogy to assist a researcher's insightful learning. There are different taxonomies of data visualization. Traditionally, statistical graphs are classified by display category, such as area, bar, circle, diagram, distribution, grid & matrix, line, map, point, and trees. Borkin, et al. 2013 augments the conventional taxonomy by taking properties and attributes into account. Properties include dimension (2-D, 3-D), multiplicity (single or multi-panel), pictorial, and temporal dimension (time series), whereas attributes include the number of distinct colors, the data-ink ratio, visual density, human recognizable objects, and human depiction. Yu 2014 conceptualizes the taxonomy in two ways: the degree of detail (noise and smoothness) and the dimensionality of data. In the former, EDA is viewed as a process of reducing large amounts of information to parsimonious summaries while remaining accurate in data description. Visualization seeks to meet this challenge by portraying complex data in interpretable means so that aspects of both the messiness and smoothness of data can be discerned. In the latter, graphs are organized by the complexity of dimensionality. For instance, one-dimensional graphs consist of small bandwidth histogram, large bandwidth histogram, density curve, and bar chart (noise → smoothness). Two-dimensional graphs comprise scatterplot, bubble plot, nonparametric density plot, sunflower plot, heatmap, image plot, contour plot, and pyramid plot (noise → smoothness). Multi-dimensional graphs include scatterplot-matrix brushing, star plot, radar plot, stereo-ray glyphs, needle plot, surface plot, cell mean plot, Trellis's coplot, and animated mesh surface (noise → smoothness). Taking big data analytics and high-power computing into account, Yalcin and Plaisant 2017 develops a data-by-tasks taxonomy, which includes

multivariate data (heatmap and dendrogram), spatial data (choropleth map, cartogram, and network map), temporal data (time-series plot), hierarchical data (treemap), network data (node-link diagram), and text data (tag clouds, matrices, parallel coordinates). There are diverse views about what is considered effective data visualization. Tufte 1997, Tufte 2001, Tufte 2006, and Few 2009 recommend simplicity and clarity. On the other hand, Borkin, et al. 2013 finds that so-called visual junk can improve retention and force the viewer to deploy more cognitive resources to comprehend the data. Lane and Sándor 2009 takes a middle ground by arguing that simple graphs could be enhanced by additional elements, such as distributional information.

Bateman, S., R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. 2010. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, Atlanta, GA (2010 April 10–15). Edited by Association for Computing Machinery, 2573–2582. New York: Association for Computing Machinery.

In this study the experimenters compared the effect of embellished charts and plain ones. It was found that there was no significant difference between viewers' accuracy in describing the fancy charts and their accuracy in describing the plain graphs. After a two- to three-week gap, the embellished group outperformed the plain group.

Borkin, M. A., A. A. Vo, Z. Bylinskii, et al. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19:2306–2315.

In this study, the experimenters found that attributes like color and the inclusion of a human recognizable object enhance memorability of the graphical presentation.

Cleveland, W. S. 1993. *Visualizing data*. Murray Hill, NJ: AT&T Bell Lab.

This is a seminal work on the principle of data visualization. Some of the well-known techniques illustrated in the book, such as Trellis's coplot, are still implemented in modern statistical software packages.

Few, S. 2009. *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, CA: Analytics Press.

The core message of this book is that data visualization should be simple and meaningful. It is written for both analysts and nonspecialists. The examples in the book are illustrated with Microsoft Excel, which is widely accessible.

Lane, D. M., and A. Sándor. 2009. Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods* 14:239–257.

Graphs in psychology journals often do not depict sufficient distributional information or inferential statistics. This article was written in response to these common problems.

Tufte, E. R. 1997. *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphic Press.

In this book, Tufte explains how graphs can be used effectively for discovery, such as tracing the source of the cholera epidemic in 19th-century London by John Snow. The central tenet is the Occam's razor: What display can be done with fewer resources is done in vain with more.

Tufte, E. R. 2001. *The visual display of quantitative information*. Cheshire, CT: Graphic Press.

In this book, Tufte introduces his theory of data graphics: clarity. He argues that the best data graphing reveals the most ideas in the shortest amount of time with the least amount of ink in the smallest amount of space.

Tufte, E. R. 2006. *Beautiful evidence*. Cheshire, CT: Graphic Press.

In this book, Tufte uses numerous examples to illustrate good and bad visualization techniques. Chapter 6, "Corruption in Evidence Presentations: Effects without Causes, Cherry Picking, Overreaching, Chartjunk, and the Rage to Conclude," is concerned with statistical graphing.

Yalcin, M. A., and C. Plaisant. 2017. Information visualization. In *Big Data and social science: A practical guide to methods and tools*. Edited by Ian Foster and Rayid Ghani, 243–263. New York: CRC Press.

The taxonomy of information visualization developed by Yalcin and Plaisant is based on data science. Many recommended techniques require high-power computers.

Yu, C. H. 2014. *Dancing with the data: The art and science of data visualization*. Saarbrucken, Germany: LAP.

This book is a comprehensive illustration of various data visualization techniques, ranging from one-dimensional graphs to multidimensional graphs. Numerous examples are given to demonstrate how each type of graph can be used properly.

Alternate Taxonomy of EDA

One major issue of the classical taxonomy is that the characteristics of EDA are tied to both the attributes of the data (e.g., distribution, linearity, outliers, measurement scales, etc.) and the final goals (e.g., detecting clusters, patterns, and relationships). However, understanding the attributes of the data is just the means instead of the ends. In the alternate taxonomy, EDA is characterized by a goal-oriented approach: detecting clusters, screening variables, and unearthing hidden relationships. There are numerous EDA and data mining techniques belonging to these three categories. Nonetheless, these goals would still be applicable to all techniques no matter what advanced procedures are introduced in the future. It is important to point out that these categories are not mutually exclusive. A data explorer could plan to accomplish several goals simultaneously or sequentially. For example, if the researcher suspects that the observations are too heterogeneous to form a single population, clustering could be conducted to divide the sample into subsamples. Next, variable selection procedures could be run to narrow down the predictor list for each subsample. Last, the researcher could focus on the interrelationships among just a few variables using pattern recognition methods. The combinations and possibilities are virtually limitless. Data detectives are encouraged to explore the data with skepticism and openness.

Detecting Data Clusters

Data reduction and summary plays a key role in EDA. Clustering is a data-reduction technique that aims to group observations based upon their proximity to each other on multiple dimensions. There are four major types of clustering algorithms. The first category is hierarchical clustering, which could be either top-down (divisive) or bottom-up (agglomerative). The former starts with one group and then partitions the data step by step according to the matrices, whereas the latter starts with one single piece of datum and then merges it with others to form larger groups. Another type is k-mean clustering. In this approach the algorithm selects k points as the initial centroids. Next, it assigns data points to different centroids based upon the P matrix (proximity) and reevaluates the centroid of each group. The iterative process is repeated until the best solution emerges (the centers are stable). TwoStep clustering is more versatile than the previous two. As the name implies, TwoStep clustering is composed of two steps. According to Zhang, et al. 1995, the first step, preclustering, constructs a

cluster features (CF) tree by scanning all cases one by one. In step two, the hierarchical clustering algorithm is applied to the preclusters and then proposes a set of solutions. To determine the best number of clusters, each solution is compared to each other based upon the Akaike Information Criterion (AIC), developed by Akaike 1973, or the Bayesian Information Criterion (BIC) based upon Schwarz 1978. Both hierarchical clustering and k-mean clustering could handle only continuous variables. Yet, TwoStep clustering accepts both categorical and continuous variables because in TwoStep clustering, the distance measurement is based on the log-likelihood method presented in Chiu, et al. 2001. In computing log-likelihood, the continuous variables are assumed to have a normal distribution and the categorical variables are assumed to have a multinomial distribution. Nevertheless, the algorithm is reasonably robust against the violation of these assumptions, and thus assumption checking is unnecessary. Density-based clustering is an advanced procedure. This method is intended to discover clusters in any shape. According to Ester, et al. 1996, in this approach, clusters are grouped by data concentrations, meaning that dense and spare areas are separated. Data points in sparse areas are treated as border points corrupted by noise. Density-based spatial clustering of applications with noise (DBSCAN) is a typical example of density-based clustering technique. For a general overview of various clustering methods, please consult Everitt, et al. 2011.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory*, Tsahkadsor, Armenia (1971 September 2–8) Edited by B. N. Petrov and F. Csaki, 267–281. Budapest: Akademia Kiado.

Akaike illustrates that AIC is a fitness index for trading off the complexity of a model against how well the model fits the data. To reach a balance between fitness and parsimony, AIC not only rewards goodness of fit, but also gives a penalty to over-fitting and complexity. Hence, the best model is the one with the lowest AIC value.

Chiu, T., D. Fang, J. Chen, Y. Wang, and C. Jeris. 2001. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26–29, 2001, San Francisco, CA, USA*. Edited by Association for Computing Machinery, 263–268. New York: Association for Computing Machinery.

This is a seminal paper in cluster analysis. Data sets with mixed types of attributes are common in the real world, but conventional clustering techniques only allow continuous-scaled data. To rectify the situation, these developers created a two-step clustering algorithm based upon Zhang et al.'s BIRCH algorithm. It was found that the two-step algorithm generates more accurate clusters than the traditional k-means method.

Ester, M., H. Krigel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR (1996 August 2–4). Edited by Association for Computing Machinery, 226–231. New York: Association for Computing Machinery.

This paper proposes the algorithm of DBSCAN, which is capable of accurately discovering clusters and noisy points. The merit of DBSCAN is that it can handle strange and messy data while conventional clustering approaches can only offer clear-cut solutions to fairly clean data.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011. *Cluster analysis*. Chichester, UK: John Wiley.

This is a comprehensive guide to cluster analysis. The book covers examples in medicine, psychology, market research, and bioinformatics.

Schwarz, G. E. 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461–464.

In this article, the problem of model selection was treated by finding the Bayes solution. Bayesian information criterion (BIC) is similar to AIC, but its penalty is heavier than that of AIC.

Zhang, T., R. Ramakrishnon, and M. Livny. 1995. BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data, May 23–25, 1995, San Jose, California*. Association for Computing Machinery, 103–114. New York: Association for Computing Machinery.

In this article, the authors point out that while hierarchical clustering is only suitable to a small data set, BIRCH clustering is so scalable that it could analyze thousands of observations efficiently.

Screening Variables

EDA challenges premature statistical modeling because sometimes the data structure might be inconsistent with the parametric assumptions. For example, the absence of multi-collinearity is essential to regression. When several predictors are highly correlated, the coefficient estimates mighty change erratically due to small changes in the model. Additionally, even if there is no multi-collinearity, the R^2 would be inflated as more predictors were entered into the model. Traditionally, data explorers employ either feature selection or feature extraction to reduce the number of

variables. According to Miller 2002 and Zou and Hastie 2005, feature selection methods aim to find a subset of the original variables based on certain selection criteria, such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC). Stepwise regression, forward selection, backward selection, LASSO, ridge regression, and elastic net are typical examples of the feature selection strategy. Feature extraction methods transform the data in a higher dimension space. Principal component regression and partial least squares (PLS) are typical examples. Besides collinearity, PLS is also robust against other data structural problems, such as skew distributions, as presented in Cassel, et al. 1999. However, it is important to note that in PLS, the focus is on prediction rather than explaining the underlying relationships between the variables. The philosophies of these two methodologies are vastly different. In feature selection "redundant" variables are excluded, whereas in feature extraction they are retained and combined to form latent factors. Equipped with high-power computing, data explorers could employ a plethora of machine-learning approaches to select variables, such as recursive partition tree and bootstrap forest. The recursive partition tree approach, developed in Breiman, et al. 1984, aims to find which independent variable(s) can make a decisive partition of the data with reference to the dependent variable. The two most common splitting criteria are Gini and Entropy, which are derived from information theory. Both criteria aim to extract the most information by reducing the impurity of the result. According to Breiman 1996 and Breiman 2001, in bootstrap forest numerous subsamples are generated independently. Additionally, in each bootstrap sample, about 30 percent of the observations are set aside for later model validation. These observations are grouped as the out of bag sample (OOBS). At the end, the computer algorithm converges these resampled results together by averaging them out. In the final output, predictors are ranked by their importance. Machine-learning methods tend to yield a more parsimonious yet more accurate model than traditional variable-reduction methods.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–140.

This is a seminal article in bagging. To address the issue of instability of a single model, Breiman proposes generating multiple models and finding the best predictors by a plurality vote. The presentation is highly mathematical.

Breiman, L. 2001. Random forests. *Machine Learning* 45.1: 5–32.

This paper is an extension of Breiman 1996. He proposes that using a random selection of features to split each node in, a partition tree could result in lower error rates. The presentation is highly mathematical.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and regression trees*. Boca Raton, FL: Chapman & Hall.

This book is a milestone in data mining. Breiman and his colleagues propose the tree-structured classification method to amend limitations of traditional regression analysis. When the data are partitioned into nodes, importance of variables can be rank-ordered and the influence of outliers is localized in each node.

Cassel, C., A. H. Westlund, and P. Hackl. 1999. Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics* 26:435–448.

This study shows that the PLS method is robust against the problems of skew distributions, multi-collinearity, and misspecification of the model (omission of regressors).

Miller, A. 2002. *Subset selection in regression*. 2d ed. Boca Raton, FL: Chapman & Hall.

This book is a comprehensive introduction to classical variable reduction methods, including forward selection, Efroymson's algorithm, backward elimination, sequential replacement algorithms, ridge regression, and LASSO.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society B* 67:301–320.

Zou and Hastie develop the elastic net approach to overcome the problems found in the ordinal least square (OLS) stepwise regression and LASSO methods. The elastic method is a form of generalized or penalized regression that aims to yield a parsimonious model by "penalizing" complexity. It starts with no modeling or zero-coefficient and the algorithm tries out a series of model.

Recognizing Patterns and Relationships

Pattern recognition entails detecting both linear and nonlinear relationships, and sometimes overturning the alleged relationship yielded by statistical modeling. Classical procedures are sensitive to sample size. As the sample size grows, even trivial effects could be misidentified as significant. For example, if both data on both X-axis and Y-axis are Likert-scaled, a regression model might return a significant *p* value even though the model has no predictive power. For instance, when X is 3, the model suggests a predicted value of Y, say 3, but the actual Y value could range from 1 to 5. Hence, it necessitates data visualization to check the model. When the sample size is very large, the data points on the scatterplot are too dense to reveal a pattern. In this case sophisticated data visualization techniques, such as heatmap and median-smoothing, are needed in order to detect the data pattern (Gu, et al. 2016; Yu, et al. 2016). The Anscombe data set is often

cited by data explorers to illustrate how we can be easily fooled by numeric-based modeling. Anscombe 1973 presents a data set with two variables, X and Y. The relationship of X and Y was fitted by ordinary least-squares regression, yielding a correlation coefficient of .83. Based on this description, most people picture a linear model. However, besides a linear association, there are other combinations of X and Y that could also return a correlation of .83, such as a curvilinear relationship, the present of an outlier, and inelasticity (no matter what happens to a variable, another variable remains unaffected). Indeed, very often the relationships in the world are nonlinear. The relationship between anxiety and performance is a classic example. Data explorers can use smoothing methods to find the most appropriate nonlinear model in an interactive fashion. Kernel smoother and smoothing spine are two popular methods. Alternatively, data analysts could employ machine-learning methods, such as artificial neural networks, to detect complex and nonlinear relationships. Neural networks try to mimic interconnected neurons in organic brains in order to make the algorithm capable of complex learning for extracting patterns and detecting trends (Samarasinghe 2007, Somers and Casal 2009). A typical neural network is composed of three types of layers, namely, the input layer, hidden layer, and output layer. The input layer contains the input data and the hidden layer performs data transformation and manipulation. And the output layer is the result.

Anscombe, F. J. 1973. Graphs in statistical analysis. *American Statistician* 27:17–21.

This is a well-cited paper that illustrates how researchers can be fooled by statistical figures if the data patterns are not examined.

Gu, Z., R. Eils, and M. Schlesner. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32:2847–2849.

These authors explain how complex heatmaps can be used to unravel hidden patterns in complex data sets. The software package developed by them is freely available from the Internet.

Samarasinghe, S. 2007. *Neural networks for applied sciences and engineering: From fundamentals to complex pattern recognition*. Boca Raton, FL: CRC Press.

This is a comprehensive guide to neural networks. The target audiences are scientists and engineers. Readers need a solid foundation in mathematics and computer programming to comprehend the illustration.

Somers, M. J., and J. C. Casal. 2009. Using artificial neural networks to model nonlinearity: The case of the job satisfaction–job performance relationship. *Organizational Research Methods* 12:403–417.

This article illustrates how neural networks can be used to model nonlinear relationships between job satisfaction and job performance.

Yu, C. H., S. Douglas, A. Lee, and M. An. 2016. Data visualization of item-total correlation by median smoothing. *Practical Assessment, Research, and Evaluation* 21.1.

This article illustrates different graphing techniques for detecting patterns when the sample size is extremely large. The techniques discussed include bubble plot, density contour plot, sunflower plot, and median smoothing.

back to top

194 15 Exploratory Data Analysis. Besides the probability plots, there are many quantitative statistical tests (not. graphical) for testing for normality, such as Pearson Chi. 2. , Shapiro-Wilk, and. Kolmogorov-Smirnov. Deviation of the observed distribution from normal makes many powerful.