Impact of Statistics on Data Science

Dr. Abhimanyu Malik,

Assistant Professor, Department of Computer Science, CRA COLLEGE, Sonepat, Haryana,

Abstract

In this, we argue that statistics is one of the fields of greatest significance for developing tools and techniques to find patterns in data and to give a greater understanding of them, as well as the most crucial for analyzing and calculating uncertainty. We provide a summary of several data science architectures that have been developed and talk about how statistics affect certain phases such data collection, enrichment, exploration, modelling, validation, and reporting are all steps in the data-processing pipeline. We also highlight errors made when statistical reasoning is disregarded.

Keyword: Data Science Architectures, Data Collection, Data-Processing Pipeline

1. Beginning and Premise

In addition to the applied sciences, Informatics, computer science, mathematics, operations research, and statistics all have an impact on data science.

IFCS "Data Science, classification, and related methods" was a statistical conference that featured the phrase "Data Science" for the first time in its title in 1996 [37]. Despite the fact that statisticians invented the phrase, the relevance of computer science and application in business is frequently overemphasised in the popular perception of data science, especially in the age of big data.

The concepts by John Tukey [43] transformed the perspective of statistics by strictly mathematical context, such as statistical testing, to support ideas with data (exploratory setting), i.e., seeking to comprehend facts prior hypothesising. This movement occurred as early as the 1970s.

KDD [36] and its underneath Data Mining are another source of data science. Knowledge discovery methods from a variety of fields, like inductive learning, (Bayesian) statistics, query optimisation, expert systems, information theory, and fuzzy sets, are already combined in KDD. KDD is a key component in fostering interaction between many fields in order to achieve the ultimate objective of finding knowledge in data.

These concepts are now united under the umbrella of data science, giving rise to several definitions. Cao recently provided most thorough explanation of data science, which is as follows [12]:

$$\begin{aligned}&\textit{data science} = (\textit{statistics} + \textit{informatics} + \textit{computing} +\\ &\textit{communication} + \textit{sociology} + \textit{management}) \mid (\textit{data} +\\ &\textit{environment} + \textit{thinking}).\end{aligned}$$

In this equation, "sociology" refers to social factors, and "data+environment+thinking" indicates that the aforementioned sciences act according to data, the setting, and the supposedly "data-to-knowledge-to-wisdom" thinking.

Donoho's 2015 [16] detailed survey of data science places particular emphasis on how it developed from statistics. In fact, a more radical viewpoint proposed rebranding statistics as data science as early as 1997

[50]. Additionally, in 2015, several ASA officials [17] published a statement on the use of statistics in information science, stating that "statistics and machine learning serve an integral part in data science."

In my view, statistical techniques are essential for the absolute abundant Data Science procedures. As a result, the foundation of our contribution is:

One of the disciplines that are most crucial for giving devices and ways to uncover structure in data and to give greater understanding of it is statistics. Statistics is also crucial for analyzing and quantifying uncertainty.

The primary focus of this essay is to discuss how statistics significantly affect the key Data Science processes.

2. Data Science Procedures

The well-known CRISP-DM, which is structured into 6 significant actions: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment [10], Table 1, left column is one of the structural forerunners of data science. Nowadays, concepts as CRISP-DM are essential for practical statistics.

For instance, according to our explanation of data science, the process entails gathering and enhancing data, storing and accessing it, exploring it, analyzing it, and modelling it. It also involves optimizing algorithms, selecting and validating models, representing and reporting results, and implementing the results in a business. In our view, the significant steps in Data Science have developed and drawn inspiration from CRISP-DM. As seen in Table 1, the right column, issues in tiny caps denote actions where statistics is not as important.

These procedures are frequently repeated in a circular loop rather than being carried out just once. Additionally, switching between multiple phases is typical. This is true in particular for the phases of Data Acquisition and Enrichment, Data Exploration, and Statistical Data Analysis, together with Statistical Data Analysis and Modelling, Model Validation, and Selection.

Table 1 contrasts various Data Science phase definitions. Horizontal blocks show the link between terms. Only observational data is dealt with by CRISP-DM, as evidenced by the missing step of Data Acquisition and Enrichment. The phases of Data Storage and Access and Optimisation of Algorithms, where statistics is less important, are also added to CRISP-DM in our proposal.

The number of procedures for data science might have expanded; for a modern list, see Cao in [12], Figure 6, and Table 1, middle column. Domain-specific data applications and issues, data management and storage, data quality improvement, data modelling and representation, deep analytics, learning and discovery, design of simulations and experiments, outstanding performance processing and analytics, networking, communication, and data-to-decision and actions.

Both Cao's and our approach cover the same key steps in theory. Cao's formulation is more precise in some places than ours, for instance, our phase Data Analysis and Modelling corresponds to Deep Analytics, Learning, and Discovery, as well as Data Modelling and Representation. Moreover, it depends on whether the needed foundation is in statistics or computer science, the vocabularies vary slightly. In this regard, it should be noted that Cao's concept of experiment design refers to the planning of simulation experiments.

The function of statistics will be highlighted in the sections that follow as we go over all the steps in Sections 2.1–2.6 where it plays a significant role. With the exception of the stages in small caps, these correspond to every step in our plan in Table 1.

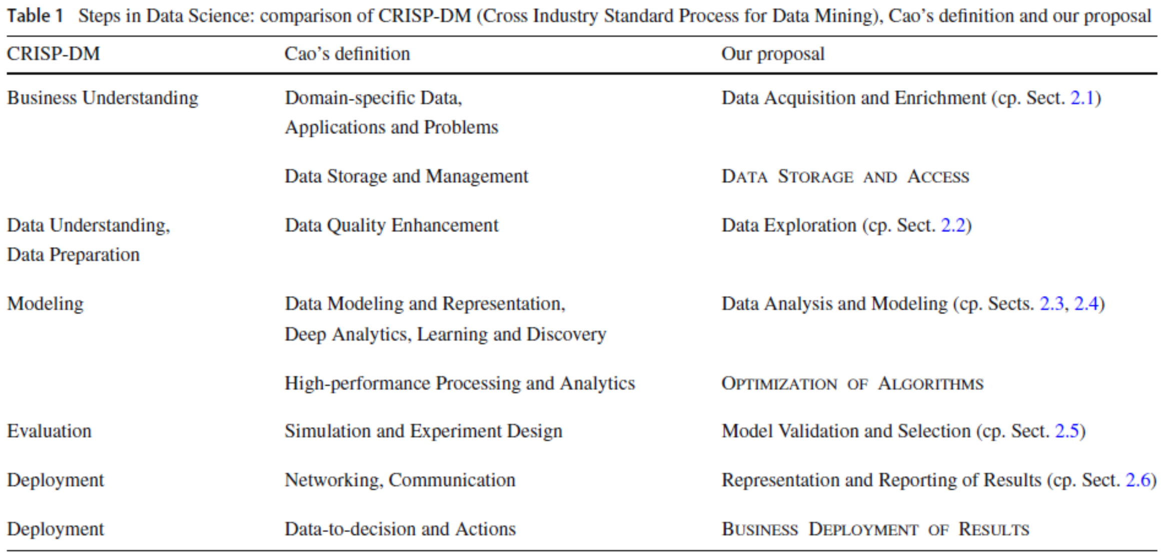

Table 1 Steps in Data Science: comparison of CRISP-DM (Cross Industry Standard Process for Data Mining), Cao's definition and our proposal

CRISP-DM	Cao's definition	Our proposal
Business Understanding	Domain-specific Data, Applications and Problems	Data Acquisition and Enrichment (cp. Sect. 2.1)
	Data Storage and Management	Data Storage and Access
Data Understanding, Data Preparation	Data Quality Enhancement	Data Exploration (cp. Sect. 2.2)
Modeling	Data Modeling and Representation, Deep Analytics, Learning and Discovery	Data Analysis and Modeling (cp. Sects. 2.3, 2.4)
	High-performance Processing and Analytics	Optimization of Algorithms
Evaluation	Simulation and Experiment Design	Model Validation and Selection (cp. Sect. 2.5)
Deployment	Networking, Communication	Representation and Reporting of Results (cp. Sect. 2.6)
Deployment	Data-to-decision and Actions	Business Deployment of Results

2.1 Gathering and Enhancement of Data

When it comes to creating data in a systematic manner while determining the impact of noisy elements, the design of experiments (DOE) is crucial. In order to develop trustworthy products despite variance in the process factors, controlled experiments are essential for robust process engineering. On the one hand, the response is influenced by certain quantities of uncontrollable variance included in even controllable parameters. However, other elements, such as environmental ones, are completely out of our control. However, at the very least, DOE, for example, should be able to control the impact of such distracting influencing elements.

DOE can be used, for example, to routinely generate fresh data (data acquisition) [33], to consistently lessen databases [41], & to tune (i.e., optimize) algorithmic parameters [1], that is, to enhance the data analysis techniques themselves (see Sect. 2.3).

New data can also be produced via simulations [7]. Imputation of missing data is a method for enhancing databases to fill in data gaps [31].

Such statistical techniques for the creation and improvement of data must form the basis of data science. The reliability of information analysis outcomes is noticeably reduced when only observational data is used, and it may even result in incorrect result interpretation. The data noise suggests that "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" [4] is incorrect.

As a result, the experimental design has a significant impact on the reliability, validity, and repeatability of our findings.

2.2 Data Analysis

In order to understand the information within a database, exploratory statistics are crucial for data preparation. In a way, John Tukey [43] was the first to explore and visualize the data that had been collected. Since then, data interpretation and transformation—the most time-consuming aspects of data analysis have grown in significance for statistical research.

Analysis or mining is necessary for the correct execution of assessment techniques in data science. The concept of distribution is statistics' most significant contribution. It enables us to capture parameter (a-priori) information as well as data variability, which is a key idea in Bayesian statistics. Distributions also help us select the best ensuing analytical models and procedures.

2.3 Analysis of Statistical Data

The most crucial tasks in data science are finding structure in the data and creating predictions. Given their versatility and ability to handle a wide range of analytical tasks, statistical approaches are particularly important in this situation. The following are significant illustrations of the Analysis of Statistical Data techniques.

a) Among the foundational principles of statistical analysis is **Hypothesis Testing**. In data-driven circumstances, questions are frequently converted into hypotheses. Additionally, hypotheses serve as the logical connections between underpinning theory and statistics. Problems and theories can be tested using the present data as statistical hypotheses are connected to statistical tests. Correction of significance levels is frequently required when the same data are used multiple times in various tests. Proper use of multiple testing is one of the most crucial issues in applied statistics, for instance in pharmacological trials [15]. Ignoring such methods would produce far more important results than necessary.

b) Methods of **Classification** are fundamental for identifying and forecasting subpopulations in data. These particular populations are to be found from data collection in the so-called unsupervised situation without having prior knowledge of any instances of such subpopulations. Clustering is a common name for this.

   When only influential elements are provided, a so-called supervised case calls for the discovery of classification rules from a labelled set of data for the prediction of unknown labels.

   There are numerous approaches available now for both the supervised situation [2] and the unsupervised case [22].

   The complexity of difficult analysis methods typically grows more strongly than proportionately with the total number of observations (n) or attributes (p), which suggests that in Taking a deeper look at outdated methods in the era of big data may be necessary. In the context of large data, if n or p is vast, this causes slow computing times and numerical difficulties. As a result, traditional approaches to statistics and machine learning are being reexamined for use with Big Data. [46] and simpler optimization algorithms with low temporal complexities are returning [9].

c) **Regression Techniques** are the main technique for determining both local as well as global correlations b/w features b/w the time that the intended variable is examined. In accordance with the distributional theory for the relevant data, many techniques might be employed. Linear regression is the most widely used method when taking the normality assumption into account, despite the fact that generalised linear regression is widely used when taking other permanently dispersed distributions into account [18]. Added complexity is achieved using functional regression for functional data [38], quantile regression [25], and regression based on loss coefficients other than squared error loss, like Lasso regression [11,21].

   Classification methods encounter problems relating to big data which is similar to those faced by those approaches when presented with a huge number of instances n Large numbers of traits p (like those in data streams). Data reduction strategies like compressed sensing, random projection methods [20], or sampling-based processes [28] enable quicker computations for the decrease of n. The number of p can be reduced to the most important characteristics while preserving the comprehensibility of the features using selecting variables or shrinking approaches like the Lasso [21]. The use of (sparse) principal component analysis is another option [21].

d) Understanding and forecasting temporal structure are the goals of **Time Series Analysis** [42]. The major issue for time series-based analyses of observational data is prediction. Along with the natural sciences and engineering, behavioural sciences and economics are typical application fields. Take signal analysis, such as the analysis of speech or musical data, as an example. The examination of theories in the time and frequency areas is included in this application of statistical methods. The main objective is to forecast potential outcomes for the time period or its properties. For instance, it is possible to forecast a musical tone's basic frequency using rules learned over long periods of time [29] and can accurately anticipate the pitch of subsequent recordings by modelling the treble aspect of an audible chronology. [24] Numerous series of time and their cointegration are frequently studied in econometrics [27]. In technological applications, time series analysis frequently aims to manage operations. [34].

2.4 The Statistical Modelling

a) Graphs or networks can be used to model complex interrelationships between variables. In this case, a link in the **graph or network** models the interaction between two components [26, 35]. The graphs might be undirected, like in Bayesian networks, or directed, like in Gaussian graphical models.

Deriving the network structure is the basic objective of network analysis. Subpopulation-specific network topologies must occasionally be separated (unmixed) [49].

b) Models from the natural and engineering sciences can be represented using **Stochastic Differential and Difference Equations** [3, 39]. Finding approximate statistical models to solve these equations might provide important information for, for instance, the statistical surveillance of related mechanical technology operations [48]. These techniques can create a link b/w data science and the practical sciences.

c) **Globalization and Regional Models**, Typically, only portions of the subject area of the variables that are pertinent are amenable to statistical models. Then, local models can be used.[8]. In time series, examination of structural flaws could be fundamental in locating areas suitable for local modelling [5]. Additionally, model alterations throughout time can be investigated via the analysis of idea drifts [30].

There are frequently **hierarchies** of ever-larger global structures in time series. For instance, the notes in music provide a fundamental local structure, whereas bars, motifs, phrases, pieces, etc. provide an increasingly more global one. Features of the a regional model can be merged to provide more global features in order to identify a time series' global characteristics [47].

The generalisation of regional to global models can also be done using **Mixture Models** [19, 23]. Due to the fact that conventional mathematical frameworks are frequently overly simplified for being suitable for various information or wider areas of concern, a model combining is crucial for the characterization of genuine relationships.

2.5 Model Selection and Validation

When multiple hypotheses are put forth for a particular task, such as estimation, statistical procedures for comparing models can be used to organize the models, for example, according to their predictive capacity [45].

Predictive power is often evaluated using so-called **resampling techniques**, in which the order of distribution of power features is examined by deliberately altering the subpopulation from which the model was trained. Such distributions' characteristics can be used to choose a model [7].

Another way to assess the effectiveness of models is through **perturbation studies**. In this method, the robustness of the various models to noise is evaluated [32, 44].

Methods to assess integrated models include **Meta-Analysis** and model averaging [13, 14].

Since there have been more and more classification and regression models presented in the literature over the past few years, **Model Selection** has become increasingly crucial.

2.6 Reporting and Representation

To effectively explain the findings of statistical studies and protect the deployment of data analysis, **visualisation** to comprehend discovered structures and **storage of models** in an updatable format are crucial tasks. In data science, deployment is crucial to producing results that can be understood. It is the final step of CRISP-DM [10] and the phase that Cao [12] uses to translate data into decisions and actions. For statistics, the primary challenge is the reporting of uncertainties as well as review, in addition to visualization and sufficient model storing [6].

3. Fallacies

The statistical techniques outlined in Section 2 are essential for identifying patterns in data, gaining a deeper understanding of data, and, ultimately, for conducting an effective data analysis. Avoidable fallacies may result from ignoring contemporary statistical thinking or from utilising imprecise data analytics or statistical procedures. This is especially true for the examination of large-scale or complex data.

The concept of distribution is the main contribution of statistics, as was stated at the end of Section 2.2. We are only able to offer values and parameter estimations without the related variability if distributions are not taken into account during data exploration and modelling. Only the idea of distributions gives us the ability to make predictions with matching error ranges.

Distributions are also essential for model-based data analytics. Unsupervised learning, for instance, can be used to identify data clusters. It is frequently crucial to infer parameters such as cluster radii and their spatio-temporal evolution if extra structure, such as dependency on space or time, is present. The concept of distributions is crucial to this kind of model-based analysis (see [40] for a reference to protein clusters).

Comparing univariate hypothesis testing methodologies to various procedures, such as multiple regressions, is advised if the use of multiple parameters is of interest. The best model should then be selected using variable selection. One would miss the correlations between variables if they restricted themselves to univariate testing.

More advanced models, such as mixture models for identifying heterogeneous groupings in data, may be needed to gain a deeper understanding of the data. The result typically indicates a nonsensical average when the combination is ignored, hence it could be necessary to learn the subgroups by unmixing the parts that make up the combination. This is made possible in a Bayesian framework by, for instance, latent allocation variables in a Dirichlet mixture model. See [49] for a molecular biology example of how to decompose a variety of networks in a population of heterogeneous cells.

A mixture model may depict mixes of highly different-sized components, with tiny components (outliers) bearing special significance. For model estimation in an environment of big data, naive sampling techniques are frequently used.

These run the danger of leaving out minor mixing components, though. Therefore, it is crucial to use resampling techniques for predictive power as well as verification of models or sampling in accordance with a more acceptable distribution.

4. Conclusion

As a result of the evaluation of statistics' abilities and effects presented above, we have come to the following conclusion: Statistics' importance in data science is underestimated.

The discovery motivates statisticians to take a leading position in the contemporary and well-liked discipline of data science. The sole option to produce scientific findings based on appropriate methods is to complement or/and combine mathematical techniques and computer algorithms that include statistical reasoning, particularly for big data. In the end, great Data Science solutions can only be achieved through a balanced interaction of all relevant fields.

References

1. Adenso-Diaz, B., Laguna, M.: Fine-tuning of algorithms using fractional experimental designs and local search. Oper. Res. **54**(1), 99–114 (2006)
2. Aggarwal, C.C. (ed.): Data Classification: Algorithms and Applications. CRC Press, Boca Raton (2014)
3. Allen, E., Allen, L., Arciniega, A., Greenwood, P.: Construction of equivalent stochastic differential equation models. Stoch. Anal. Appl. **26**, 274–297 (2008)
4. Anderson, C.: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine https://www.wired.com/2008/06/pb-theory/ (2008)
5. Aue, A., Horváth, L.: Structural breaks in time series. J. Time Ser. Anal. **34**(1), 1–16 (2013)
6. Berger, R.E.: A scientific approach to writing for engineers and scientists. IEEE PCS Professional Engineering Communication Series IEEE Press, Wiley (2014)
7. Bischl, B., Mersmann, O., Trautmann, H., Weihs, C.: Resampling methods formeta-model validationwith recommendations for evolutionary computation. Evol. Comput. **20**(2), 249–275 (2012)
8. Bischl, B., Schiffner, J.,Weihs, C.: Benchmarking local classification methods. Comput. Stat. **28**(6), 2599–2619 (2013)
9. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. arXiv preprint arXiv:1606.04838(2016)
10. Brown, M.S.: Data Mining for Dummies. Wiley, London (2014)
11. Bühlmann, P., Van De Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin (2011)
12. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (2017). https://doi.org/10.1145/3076253
13. Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging. Cambridge University Press, Cambridge (2008)

14. Cooper, H., Hedges, L.V., Valentine, J.C.: The Handbook of Research Synthesis and Meta-analysis. Russell Sage Foundation, New York City (2009)
15. Dmitrienko, A., Tamhane, A.C., Bretz, F.: Multiple Testing Problems in Pharmaceutical Statistics. Chapman and Hall/CRC, London(2009)
16. Donoho, D.: 50Years ofData Science. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf (2015)
17. Dyk, D.V., Fuentes, M., Jordan, M.I., Newton, M., Ray, B.K. Lang, D.T.,Wickham, H.: ASA Statement on the Role of Statistic in Data Science. http://magazine.amstat.org/blog/2015/10/01/asastatement-on-the-role-of-statistics-in-data-science/ (2015)
18. Fahrmeir, L., Kneib, T., Lang, S., Marx, B.: Regression: Models, Methods and Applications. Springer, Berlin (2013)
19. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, Berlin (2006)
20. Geppert, L., Ickstadt, K., Munteanu, A., Quedenfeld, J., Sohler, C.: Random projections for Bayesian regression. Stat. Comput. **27**(1), 79–101 (2017) https://doi.org/10.1007/s11222-015-9608-z
21. Hastie, T., Tibshirani,R.,Wainwright,M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)
22. Hennig, C.,Meila,M., Murtagh, F.,Rocci,R.: Handbook of Cluster Analysis. Chapman & Hall, London (2015)
23. Klein, H.U., Schäfer, M., Porse, B.T., Hasemann, M.S., Ickstadt, K., Dugas, M.: Integrative analysis of histone chip-seq and transcription data using Bayesian mixture models. Bioinformatics **30**(8), 1154–1162 (2014)
24. Knoche, S., Ebeling, M.: The musical signal: physically and psychologically,chap 2. In: Weihs, C., Jannach, D., Vatolkin, I., Rudolph, G. (eds.) Music Data Analysis Foundations and Applications, pp. 15–68. CRC Press, Boca Raton (2017)
25. Koenker, R.: Quantile Regression. Econometric Society Monographs, vol. 38 (2010)
26. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press, Cambridge (2009)
27. Lütkepohl,H.: NewIntroduction to Multiple Time SeriesAnalysis. Springer, Berlin (2010)
28. Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. In: Proceedings of the 31th International Conference on Machine Learning,ICML2014, Beijing, China, 21–26 June 2014, pp 91–99. http://jmlr.org/proceedings/papers/v32/ ma14.html (2014)
29. Martin,R., Nagathil, A.:Digital filters and spectral analysis, chap 4. In: Weihs, C., Jannach, D., Vatolkin, I., Rudolph, G. (eds.) Music Data Analysis—Foundations and Applications, pp. 111–143. CRC Press, Boca Raton (2017)
30. Mejri, D., Limam,M.,Weihs, C.: A new dynamic weighted majority control chart for data streams. Soft Comput. 22(2), 511–522.https://doi.org/10.1007/s00500-016-2351-3
31. Molenberghs,G., Fitzmaurice, G.,Kenward,M.G., Tsiatis,A.,Verbeke,G.: Handbook of Missing Data Methodology. CRC Press, Boca Raton (2014)
32. Molinelli, E.J., Korkut, A., Wang, W.Q., Miller, M.L., Gauthier, N.P., Jing, X., Kaushik, P., He, Q., Mills, G., Solit, D.B., Pratilas, C.A.,Weigt,M., Braunstein,A., Pagnani,A., Zecchina, R., Sander, C.: Perturbation Biology: Inferring Signaling Networks in Cellular Systems. arXiv preprint arXiv:1308.5193 (2013)

33. . Montgomery, D.C.: Design and Analysis of Experiments, 8th edn. Wiley, London (2013)
34. Oakland, J.: Statistical Process Control. Routledge, London (2007)
35. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, Los Altos (1988)
36. Piateski,G., Frawley,W.:Knowledge Discovery in Databases. MIT Press, Cambridge (1991)
37. Press, G.: A Very Short History of Data Science. https://www.forbescom/sites/gilpress/2013/05/28/a-very-short-history-ofdata-science/#5c515ed055cf (2013). [Last visit: March 19, 2017]
38. Ramsay, J., Silverman, B.W.: Functional Data Analysis. Springer, Berlin (2005)
39. Särkkä, S.: Applied Stochastic Differential Equations. https://users.aalto.fi/~ssarkka/course_s2012/pdf/sde_course_booklet_ 2012.pdf (2012). [Last visit: March 6, 2017]
40. Schäfer, M., Radon, Y., Klein, T., Herrmann, S., Schwender, H., Verveer, P.J., Ickstadt, K.: A Bayesian mixture model to quantify parameters of spatial clustering. Comput. Stat. Data Anal. **92**, 163–176 (2015). https://doi.org/10.1016/j.csda.2015.07.004
41. Schiffner, J., Weihs, C.: D-optimal plans for variable selection in data bases. Technical Report, 14/09, SFB 475 (2009)
42. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications: With R Examples. Springer, Berlin (2010)
43. Tukey, J.W.: Exploratory Data Analysis. Pearson, London (1977)
44. Vatcheva, I., de Jong, H.,Mars, N.: Selection of perturbation experiments for model discrimination. In: Horn,W. (ed.) Proceedings of the 14th European Conference on Artificial Intelligence, ECAI-2000, IOS Press, pp 191–195 (2000)
45. Vatolkin, I., Weihs, C.: Evaluation, chap 13. In: Weihs, C., Jannach,D.,Vatolkin, I., Rudolph, G. (eds.) Music Data Analysis—Foundations and Applications, pp. 329–363. CRC Press, Boca Raton (2017)
46. Weihs, C.: Big data classification — aspects on many features. In: Michaelis, S., Piatkowski, N., Stolpe, M. (eds.) Solving Large Scale Learning Tasks: Challenges and Algorithms, Springer Lecture Notes in Artificial Intelligence, vol. 9580, pp. 139–147 (2016)
47. Weihs, C., Ligges, U.: From local to global analysis of music time series. In: Morik, K., Siebes, A., Boulicault, J.F. (eds.) Detecting Local Patterns, Springer Lecture Notes in Artificial Intelligence, vol. 3539, pp. 233–245 (2005)
48. Weihs, C.,Messaoud, A., Raabe, N.: Control charts based on models derived from differential equations. Qual. Reliab. Eng. Int. **26**(8), 807–816 (2010)
49. Wieczorek, J., Malik-Sheriff, R.S., Fermin, Y., Grecco, H.E., Zamir, E., Ickstadt, K.: Uncovering distinct protein-network topologies in heterogeneous cell populations.BMCSyst.Biol. **9**(1), 24 (2015)
50. Wu, J.: Statistics = data science? http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf (1997)