International Journal of Data Science and Analytics (2018) 6:189–194
https://doi.org/10.1007/s41060-018-0102-5

REGULAR PAPER

Data Science: the impact of statistics

Claus Weihs[1] · **Katja Ickstadt**[2]

Received: 20 March 2017 / Accepted: 25 January 2018 / Published online: 16 February 2018

Abstract
In this paper, we substantiate our premise that statistics is one of the most important disciplines to provide tools and methods to find structure in and to give deeper insight into data, and the most important discipline to analyze and quantify uncertainty. We give an overview over different proposed structures of Data Science and address the impact of statistics on such steps as data acquisition and enrichment, data exploration, data analysis and modeling, validation and representation and reporting. Also, we indicate fallacies when neglecting statistical reasoning.

Keywords Structures of data science · Impact of statistics on data science · Fallacies in data science

1 Introduction and premise

Data Science as a scientific discipline is influenced by informatics, computer science, mathematics, operations research, and statistics as well as the applied sciences.

In 1996, for the first time, the term Data Science was included in the title of a statistical conference (International Federation of Classification Societies (IFCS) "Data Science, classification, and related methods") [37]. Even though the term was founded by statisticians, in the public image of Data Science, the importance of computer science and business applications is often much more stressed, in particular in the era of Big Data.

Already in the 1970s, the ideas of John Tukey [43] changed the viewpoint of statistics from a purely *mathematical setting*, e.g., statistical testing, to deriving hypotheses from data (*exploratory setting*), i.e., trying to understand the data before hypothesizing.

Another root of Data Science is *Knowledge Discovery in Databases* (KDD) [36] with its sub-topic *Data Mining*. KDD already brings together many different approaches to knowledge discovery, including inductive learning, (Bayesian) statistics, query optimization, expert systems, information theory, and fuzzy sets. Thus, KDD is a big building block for fostering interaction between different fields for the overall goal of identifying knowledge in data.

Nowadays, these ideas are combined in the notion of Data Science, leading to different definitions. One of the most comprehensive definitions of Data Science was recently given by Cao as the formula [12]:

> *data science = (statistics + informatics + computing + communication + sociology + management) | (data + environment + thinking).*

In this formula, *sociology* stands for the social aspects and | *(data + environment + thinking)* means that all the mentioned sciences act on the basis of data, the environment and the so-called data-to-knowledge-to-wisdom thinking.

A recent, comprehensive overview of Data Science provided by Donoho in 2015 [16] focuses on the evolution of Data Science from statistics. Indeed, as early as 1997, there was an even more radical view suggesting to rename statistics to Data Science [50]. And in 2015, a number of ASA leaders [17] released a statement about the role of statistics in Data Science, saying that "statistics and machine learning play a central role in data science."

In our view, statistical methods are crucial in most fundamental steps of Data Science. Hence, the *premise* of our contribution is:

Statistics is one of the most important disciplines to provide tools and methods to find structure in and to give deeper

✉ Claus Weihs
weihs@statistik.tu-dortmund.de

Katja Ickstadt
ickstadt@statistik.tu-dortmund.de

[1] Computational Statistics, TU Dortmund University, 44221 Dortmund, Germany

[2] Mathematical Statistics and Biometric Applications, TU Dortmund University, 44221 Dortmund, Germany

insight into data, and the most important discipline to analyze and quantify uncertainty.

This paper aims at addressing the major impact of statistics on the most important steps in Data Science.

2 Steps in data science

One of forerunners of Data Science from a structural perspective is the famous CRISP-DM (Cross Industry Standard Process for Data Mining) which is organized in six main steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [10], see Table 1, left column. Ideas like CRISP-DM are now fundamental for applied statistics.

In our view, the main steps in Data Science have been inspired by CRISP-DM and have evolved, leading to, e.g., our definition of Data Science as a sequence of the following steps: Data Acquisition and Enrichment, DATA STORAGE AND ACCESS, Data Exploration, Data Analysis and Modeling, OPTIMIZATION OF ALGORITHMS, Model Validation and Selection, Representation and Reporting of Results, and BUSINESS DEPLOYMENT OF RESULTS. Note that topics in small capitals indicate steps where statistics is less involved, cp. Table 1, right column.

Usually, these steps are not just conducted once but are iterated in a cyclic loop. In addition, it is common to alternate between two or more steps. This holds especially for the steps *Data Acquisition and Enrichment*, *Data Exploration*, and *Statistical Data Analysis*, as well as for *Statistical Data Analysis and Modeling* and *Model Validation and Selection*.

Table 1 compares different definitions of steps in Data Science. The relationship of terms is indicated by horizontal blocks. The missing step *Data Acquisition and Enrichment* in CRISP-DM indicates that that scheme deals with observational data only. Moreover, in our proposal, the steps *Data Storage and Access* and *Optimization of Algorithms* are added to CRISP-DM, where statistics is less involved.

The list of steps for Data Science may even be enlarged, see, e.g., Cao in [12], Figure 6, cp. also Table 1, middle column, for the following recent list: Domain-specific Data Applications and Problems, Data Storage and Management, Data Quality Enhancement, Data Modeling and Representation, Deep Analytics, Learning and Discovery, Simulation and Experiment Design, High-performance Processing and Analytics, Networking, Communication, Data-to-Decision and Actions.

In principle, Cao's and our proposal cover the same main steps. However, in parts, Cao's formulation is more detailed; e.g., our step *Data Analysis and Modeling* corresponds to *Data Modeling and Representation, Deep Analytics, Learning and Discovery*. Also, the vocabularies differ slightly, depending on whether the respective background is computer science or statistics. In that respect note that *Experiment Design* in Cao's definition means the design of the simulation experiments.

In what follows, we will highlight the role of statistics discussing all the steps, where it is heavily involved, in Sects. 2.1–2.6. These coincide with all steps in our proposal in Table 1 except steps in small capitals. The corresponding

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

entries DATA STORAGE AND ACCESS and OPTIMIZATION OF ALGORITHMS are mainly covered by *informatics* and *computer science*, whereas BUSINESS DEPLOYMENT OF RESULTS is covered by *Business Management*.

2.1 Data acquisition and enrichment

Design of experiments (DOE) is essential for a systematic generation of data when the effect of noisy factors has to be identified. Controlled experiments are fundamental for robust process engineering to produce reliable products despite variation in the process variables. On the one hand, even controllable factors contain a certain amount of uncontrollable variation that affects the response. On the other hand, some factors, like environmental factors, cannot be controlled at all. Nevertheless, at least the effect of such noisy influencing factors should be controlled by, e.g., DOE.

DOE can be utilized, e.g.,

- to systematically generate new data (**data acquisition**) [33],
- for systematically reducing data bases [41], and
- for tuning (i.e., optimizing) parameters of algorithms [1], i.e., for improving the data analysis methods (see Sect. 2.3) themselves.

Simulations [7] may also be used to generate new data. A tool for the enrichment of data bases to fill data gaps is the **imputation** of missing data [31].

Such statistical methods for data generation and enrichment need to be part of the backbone of Data Science. The exclusive use of observational data without any noise control distinctly diminishes the quality of data analysis results and may even lead to wrong result interpretation. The hope for "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" [4] appears to be wrong due to noise in the data.

Thus, experimental design is crucial for the reliability, validity, and replicability of our results.

2.2 Data exploration

Exploratory statistics is essential for data preprocessing to learn about the contents of a data base. Exploration and visualization of observed data was, in a way, initiated by John Tukey [43]. Since that time, the most laborious part of data analysis, namely data understanding and transformation, became an important part in statistical science.

Data exploration or *data mining* is fundamental for the proper usage of analytical methods in Data Science. The most important contribution of statistics is the notion of **distribution**. It allows us to represent variability in the data as well as (a-priori) knowledge of parameters, the concept underlying Bayesian statistics. Distributions also enable us to choose adequate subsequent analytic models and methods.

2.3 Statistical data analysis

Finding structure in data and making predictions are the most important steps in Data Science. Here, in particular, statistical methods are essential since they are able to handle many different analytical tasks. Important examples of statistical data analysis methods are the following.

a) **Hypothesis testing** is one of the pillars of statistical analysis. Questions arising in data driven problems can often be translated to hypotheses. Also, hypotheses are the natural links between underlying theory and statistics. Since statistical hypotheses are related to statistical tests, questions and theory can be tested for the available data. Multiple usage of the same data in different tests often leads to the necessity to correct significance levels. In applied statistics, correct **multiple testing** is one of the most important problems, e.g., in pharmaceutical studies [15]. Ignoring such techniques would lead to many more significant results than justified.

b) **Classification** methods are basic for finding and predicting subpopulations from data. In the so-called unsupervised case, such subpopulations are to be found from a data set without a-priori knowledge of any cases of such subpopulations. This is often called clustering.
In the so-called supervised case, classification rules should be found from a labeled data set for the prediction of unknown labels when only influential factors are available.
Nowadays, there is a plethora of methods for the unsupervised [22] as well for the supervised case [2].
In the age of Big Data, a new look at the classical methods appears to be necessary, though, since most of the time the calculation effort of complex analysis methods grows stronger than linear with the number of observations n or the number of features p. In the case of Big Data, i.e., if n or p is large, this leads to too high calculation times and to numerical problems. This results both, in the comeback of simpler optimization algorithms with low time-complexity [9] and in re-examining the traditional methods in statistics and machine learning for Big Data [46].

c) **Regression** methods are the main tool to find global and local relationships between features when the target variable is measured. Depending on the distributional assumption for the underlying data, different approaches may be applied. Under the normality assumption, linear regression is the most common method, while generalized linear regression is usually employed for other distributions from the exponential family [18]. More

advanced methods comprise functional regression for functional data [38], quantile regression [25], and regression based on loss functions other than squared error loss like, e.g., Lasso regression [11,21].

In the context of Big Data, the challenges are similar to those for classification methods given large numbers of observations n (e.g., in data streams) and / or large numbers of features p. For the reduction of n, data reduction techniques like compressed sensing, random projection methods [20] or sampling-based procedures [28] enable faster computations. For decreasing the number p to the most influential features, variable selection or shrinkage approaches like the Lasso [21] can be employed, keeping the interpretability of the features. (Sparse) principal component analysis [21] may also be used.

d) **Time series analysis** aims at understanding and predicting temporal structure [42]. Time series are very common in studies of observational data, and prediction is the most important challenge for such data. Typical application areas are the behavioral sciences and economics as well as the natural sciences and engineering. As an example, let us have a look at signal analysis, e.g., speech or music data analysis. Here, statistical methods comprise the analysis of models in the time and frequency domains. The main aim is the prediction of future values of the time series itself or of its properties. For example, the vibrato of an audio time series might be modeled in order to realistically predict the tone in the future [24] and the fundamental frequency of a musical tone might be predicted by rules learned from elapsed time periods [29].

In econometrics, multiple time series and their cointegration are often analyzed [27]. In technical applications, process control is a common aim of time series analysis [34].

2.4 Statistical modeling

(a) Complex interactions between factors can be modeled by **graphs or networks**. Here, an interaction between two factors is modeled by a connection in the graph or network [26,35]. The graphs can be undirected as, e.g., in Gaussian graphical models, or directed as, e.g., in Bayesian networks.

The main goal in network analysis is deriving the network structure. Sometimes, it is necessary to separate (unmix) subpopulation specific network topologies [49].

(b) **Stochastic differential and difference equations** can represent models from the natural and engineering sciences [3,39]. The finding of approximate statistical models solving such equations can lead to valuable insights for, e.g., the statistical control of such processes, e.g., in mechanical engineering [48]. Such methods can build a bridge between the applied sciences and Data Science.

(c) **Local models and globalization** Typically, statistical models are only valid in sub-regions of the domain of the involved variables. Then, local models can be used [8]. The analysis of structural breaks can be basic to identify the regions for local modeling in time series [5]. Also, the analysis of concept drifts can be used to investigate model changes over time [30].

In time series, there are often **hierarchies** of more and more global structures. For example, in music, a basic local structure is given by the notes and more and more global ones by bars, motifs, phrases, parts etc. In order to find global properties of a time series, properties of the local models can be combined to more global characteristics [47].

Mixture models can also be used for the generalization of local to global models [19,23]. Model combination is essential for the characterization of real relationships since standard mathematical models are often much too simple to be valid for heterogeneous data or bigger regions of interest.

2.5 Model validation and model selection

In cases where more than one model is proposed for, e.g., prediction, statistical tests for comparing models are helpful to structure the models, e.g., concerning their predictive power [45].

Predictive power is typically assessed by means of so-called **resampling methods** where the distribution of power characteristics is studied by artificially varying the subpopulation used to learn the model. Characteristics of such distributions can be used for model selection [7].

Perturbation experiments offer another possibility to evaluate the performance of models. In this way, the stability of the different models against noise is assessed [32,44].

Meta-analysis as well as model averaging are methods to evaluate combined models [13,14].

Model selection became more and more important in the last years since the number of classification and regression models proposed in the literature increased with higher and higher speed.

2.6 Representation and reporting

Visualization to interpret found structures and **storing of models** in an easy-to-update form are very important tasks in statistical analyses to communicate the results and safeguard data analysis deployment. Deployment is decisive for obtaining interpretable results in Data Science. It is the last

step in CRISP-DM [10] and underlying the data-to-decision and action step in Cao [12].

Besides visualization and adequate model storing, for statistics, the main task is **reporting of uncertainties and review** [6].

3 Fallacies

The statistical methods described in Sect. 2 are fundamental for finding structure in data and for obtaining deeper insight into data, and thus, for a successful data analysis. Ignoring modern statistical thinking or using simplistic data analytics/statistical methods may lead to avoidable fallacies. This holds, in particular, for the analysis of big and/or complex data.

As mentioned at the end of Sect. 2.2, the notion of **distribution** is the key contribution of statistics. Not taking into account distributions in data exploration and in modeling restricts us to report values and parameter estimates without their corresponding variability. Only the notion of distributions enables us to predict with corresponding error bands.

Moreover, distributions are the key to model-based data analytics. For example, unsupervised learning can be employed to find clusters in data. If additional structure like dependency on space or time is present, it is often important to infer parameters like cluster radii and their spatio-temporal evolution. Such model-based analysis heavily depends on the notion of distributions (see [40] for an application to protein clusters).

If more than one parameter is of interest, it is advisable to compare univariate hypothesis testing approaches to multiple procedures, e.g., in multiple regression, and choose the most adequate model by variable selection. Restricting oneself to univariate testing, would ignore relationships between variables.

Deeper insight into data might require more complex models, like, e.g., mixture models for detecting heterogeneous groups in data. When ignoring the mixture, the result often represents a meaningless average, and learning the subgroups by unmixing the components might be needed. In a Bayesian framework, this is enabled by, e.g., latent allocation variables in a Dirichlet mixture model. For an application of decomposing a mixture of different networks in a heterogeneous cell population in molecular biology see [49].

A mixture model might represent mixtures of components of very unequal sizes, with small components (outliers) being of particular importance. In the context of Big Data, naïve sampling procedures are often employed for model estimation. However, these have the risk of missing small mixture components. Hence, model validation or sampling according to a more suitable distribution as well as resampling methods for predictive power are important.

4 Conclusion

Following the above assessment of the capabilities and impacts of statistics our conclusion is:

The role of statistics in Data Science is under-estimated as, e.g., compared to computer science. This yields, in particular, for the areas of data acquisition and enrichment as well as for advanced modeling needed for prediction.

Stimulated by this conclusion, statisticians are well-advised to more offensively play their role in this modern and well accepted field of Data Science.

Only complementing and/or combining mathematical methods and computational algorithms with statistical reasoning, particularly for Big Data, will lead to scientific results based on suitable approaches. Ultimately, only a balanced interplay of all sciences involved will lead to successful solutions in Data Science.

Acknowledgements The authors would like to thank the editor, the guest editors and all reviewers for valuable comments on an earlier version of the manuscript. They also thank Leo Geppert for fruitful discussions.

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 for meta-model validation with 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.: 50 Years of Data 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 Statistics in Data Science. http://magazine.amstat.org/blog/2015/10/01/asa-statement-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.: New Introduction to Multiple Time Series Analysis. 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, ICML 2014, 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-of-data-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. BMC Syst. Biol. **9**(1), 24 (2015)
50. Wu, J.: Statistics = data science? http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf (1997)