Position Paper: Computer Supported Education vs. Education Supported Computing - On the Problem of Informed Decision Making of Appropriate Data Analytics Method

Daniyal Kazempour[1][a], Christiane Attig[2][b], Peer Kröger[1][c], Muhammad Aammar Tufail[1][d], Daniela E. Winkler[1][e] and Claudius Zelenka[1][f]

[1]*Christian-Albrechts-Universität zu Kiel, Kiel, Germany*

[2]*Universität zu Lübeck, Lübeck, Germany*

Keywords: Data Analytics, Method Competence, Method-Application Gap, Interdisciplinary Research.

Abstract: In the field of data-related analytics, the overwhelming number of available methods presents a challenge: Which method should actually be chosen for a given problem? In this position paper, we raise awareness of this issue and propose educational and computational concepts to address related challenges and possibilities. As a unique contribution, we include the perspectives of scientists from different domains which include biology, bioinformatics, and psychology on the problem of method selection, aiming to initiate future discussions and advancement.

1 INTRODUCTION

Teaching data analytics methods is growing in importance. As the field of database and machine learning research advances, novel methods gradually come into focus. These new methods can discover patterns—such as clusters or correlations—that previous methods failed to detect. However, they may also lack the ability to detect patterns that could be discovered by earlier techniques. In short, there is no 'one-size-fits-all' solution. Relying on either older or newer methods as multi-purpose tools can be tempting, but may lead to a form of 'blindness', causing relevant patterns to be missed. In this work we present four *positions* deemed relevant for domain and computer science alike, addressing the teaching of methods and their case-aware application.

Position 1:
Wealth of Methods vs. Lack of Knowledge: The Method-Application Gap.

Data analytics appears to be omnipresent in many different domain sciences such as physics, social science, economics, biology etc. This does not come to our surprise, since data analytics provided means to boost the scientific advancements. Similarly, in the field of data analytics and machine learning a wealth of methods has been developed, each of them addressing partially disjunct, partially overlapping challenges in order to discover patterns within data. As a side-effect we observe something that we address in this position paper by the term *method-application-gap*: On the one hand we have within the domain sciences well-established subsets of methods that are 'common practice' for data analytics. This subset of methods is partially taught in a *cookbook style* within the educational processes of the academic landscape, as elaborated in Section 3. Each of the methods, however, excels at their own subset of characteristics (i.e. discovering arbitrary shaped patterns, linear correlations etc.), which raises the need to utilize other and potentially more recent methods. On the other hand the sheer amount of methods developed and published in the field of data analytics and machine learning renders it impossible to 'catch up' for the domain science knowing (a) that other methods exist and (b) which one of them to choose (c) for which reasons. This problem has also been discussed in *Data Clustering: 50 Years Beyond K-means* (Jain, 2010) where the authors state:

> "In spite of thousands of clustering algorithms that have been published, a user still faces a

[a] https://orcid.org/0000-0002-2063-2756
[b] https://orcid.org/0000-0002-6280-2530
[c] https://orcid.org/0000-0001-5646-3299
[d] https://orcid.org/0000-0002-2795-4985
[e] https://orcid.org/0000-0001-7501-2506
[f] https://orcid.org/0000-0002-9902-2212

Kazempour, D., Attig, C., Kröger, P., Tufail, M. A., Winkler, D. E. and Zelenka, C.
Position Paper: Computer Supported Education vs. Education Supported Computing - On the Problem of Informed Decision Making of Appropriate Data Analytics Method.
DOI: 10.5220/0013476500003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 438-445
ISBN: 978-989-758-746-7; ISSN: 2184-5026

dilemma regarding the choice of algorithm, distance metric, data normalization, number of clusters, and validation criteria."

While it may be argued that the 'go-to' methods are all that domain scientists need, we, an interdisciplinary group of scientists, claim that the knowledge of other methods can enable the discovery of patterns and hence ultimately novel insights that would else be inaccessible. As a consequence, an *Education Supported Computing* approach that is tailored to teach 'when to use what, and why' is of paramount importance.

Position 2:
Automated Machine Learning Is *not* All You Need.

Automated machine learning (ML) pipelines like AutoML provide a high comfort and are easy-to-use. At that point one would be tempted to ask 'Why should we *teach* students how and when to choose which data analytics method?', since an entirely automated approach would render the need to answer such questions obsolete. However, AutoML approaches are not the 'holy grail':

In a meta-review, the authors of (Barbudo et al., 2023) performed a literature search on AutoML based on a proposed taxonomy that encompasses 447 primary studies selected from a set of 31,048 papers. Barbudo et al. (2023) found that the majority (91%) of tasks addressed by AutoML are from supervised or regression archetypes. The more challenging unsupervised tasks like clustering or anomaly detection are addressed by only 1-2% of the publications.

Even more severe disadvantages of AutoML revolve around the fact that AutoML approaches operate as black-box methods [(Barbudo et al., 2023), (Quaranta et al., 2024)], which implies that users have to rely on the generated models regardless of the ways they can be either interpreted or plausibly explained by humans, hindering the scientists interpretability of the results. In this respect, explainability is essential for humans to provide more details in order to obtain more meaningful results [(Barbudo et al., 2023), (Quaranta et al., 2024)] that ultimately can benefit the automation process itself. Additionally, Quaranta et al. (2024) confirm that AutoML's capabilities are limited in unsupervised settings, especially when confronted with 'non-standard' use cases and domains by failing to adapt to the complexities of such scenarios.

Overall, the necessity of humans in the loop remains of paramount importance, as stated by Xanthopoulos et al. (2020), that among the most important criteria for users to choose a method is the interpretability of the results. The authors specifically mention that users are rarely satisfied with only a predictive model, but aim to understand the discovered patterns within the data, or in the authors' terms of brevity: "AutoML should automate, not obfuscate." (Xanthopoulos et al., 2020).

As a bottom line of this position, it is not advisable to entirely rely on automated machine learning pipelines while at the same time neglecting or discarding any need to teach and learn *when* to use *which* method. This is especially the case when it comes to discover novel and hence mostly unknown data. Instead, we deem it as more important to educate students and scientists alike to learn and understand when to use which existing method rather than to rely on automated ML processes.

Position 3:
Learning by Doing: On the Need to Interactively Practice Data Analytics Methods.

So far we have addressed the *Education Supported Computing* (ESC) field, meaning that one needs to learn when to use which of the existing methods.

To achieve this, we now transit to the realm of *Computer Supported Education* (CSE), the main theme of this conference. The third position discusses the need to incorporate computer science methods to support the education on when to use which method.

Many of the data analytics modules provide the means to practice the learned methods in the tutorials of their respective courses. This practice is however mostly tailored at completing an existing code fragment (e.g. local sequence alignment in bioinformatics) or to apply a tool on a specific dataset. Some courses even require to perform steps of an algorithm in 'pen-and-paper' style. While these approaches indeed foster the understanding of *how* the algorithms work and how they can be used, they do not explicitly focus on the strengths and limitations of methods. Moreover, they do not actively demand an understanding of the methods and their case-aware application.

In a currently ongoing teaching of bachelor and master students in unsupervised machine learning methods on the example of clustering, we provide datasets and ask them to interactively run *different methods* on the datasets and with *different parameter settings* using ELKI (Schubert and Zimek, 2019). The students are then instructed to note what they observe regarding differences in the clustering results. In case of data streams we use the MOA framework (Bifet et al., 2011) such that the students can simulate

and observe different data stream scenarios.

To leverage the experience in learning the strengths and limitations of methods we provide students the task to design their own datasets through a sample generator[1]. While this generator is simple and does not require any installation or complex learning efforts, it allows students to focus on the provided task:

> *Design and modify datasets in such a way that the results of the clustering improve or worsen. Characterize the properties of the dataset that lead to either changes of the performance. Provide possible reasons with respect to the used method that explain the exceptionally good or poor performance.*

With this combination of exploring the performance of algorithms with different parameter settings between different methods and the impact of characteristics of the data on the methods performance, we see a *Computer Supported Education* approach that sustainably prepares scientists of computer science and domain science alike becoming proficient in when to choose which method.

2 APPROACHING THE METHOD-APPLICATION GAP

Despite the gained experiences of when to use which method, the sheer amount of data analytics methods itself can be prohibitive to explore and use novel methods that may be more suitable for the respective problem.

But *what can be done to discover more suitable algorithms with low(er) effort*?

Obviously one possible approach to that lies in *computer supported* solutions. We ask at this point: What if we would have a recommendation system? A recommendation system that can be queried and then responds with archetype methods (e.g. density-based clustering, hierarchical clustering) to choose from. More importantly, a system that provides an explanation for the selection of methods, which enhances *understandability* of the underlying selection. The idea of relying on such a recommender system is neither new nor far-fetched. Consider for this case movie streaming platforms that recommend movies or online market platforms that recommend products. The idea of a recommender system for suggesting algorithms has been approached Collins et al. (2018) and is actively discussed in the Interdisciplinary Workshop on Algorithm Selection and Meta-Learning in Information Retrieval (Beel and Kotthoff, 2019).

However, two aspects remain unclear: Which questions should be posed to the recommender system, and which information (e.g. properties of the data) should be provided? To approach that problem, it is of paramount importance to have some kind of structuring that allows a categorization of different data analytics methods for a specific task (in the scope of this work, we take clustering as an example). As an open question, we ask for the choice of criteria in order to structure the algorithms so that researchers in different fields can use the system to their advantage. In various survey papers [(Sim et al., 2013), and references within], we see tables that indicate potential structures; these, however, seem to cover certain aspects, e.g., the way algorithms operate (bottom-up, top-down, grid-based, etc.) or their parameters—which might not be relevant for all research questions in all scientific fields.

To provide a more application-tailored structuring, we deem it necessary to propose a categorization of algorithms. This idea itself is also not new per se and can be seen in the different approaches e.g. *meta-data* information. It serves the purpose to understand in which instance which types of algorithms in data mining and machine learning are a reasonable choice. It fosters taking **aspects/properties** like:

1. data set specific properties
2. algorithm specific properties and
3. model specific properties

into account.

Each of the **aspects/properties** in itself is governed by certain **assumptions** that operate on different **levels**. The following list of aspects and assumptions is by no means complete. Here we have the challenge of a delicate balance between coverage of different (use)cases vs. complexity, on which we elaborate more in detail in the following section.
In case of (1) **data specific properties** we consider:

- **a. data type and semantic-level assumptions**
- **b. data-origin/generation-level assumptions**
- **c. instance-level assumptions**
- **d. feature-level assumptions**
- **e. pattern-level assumptions**
- **f. outlier/anomaly-level assumptions**

In the case of (2) **algorithm specific properties** we suggest the following levels with their respectively underlying assumptions:

[1] https://guoguibing.github.io/librec/datagen.html

a. objective-level assumptions
b. process-level assumptions
c. parameter-level assumptions
d. output-level assumptions

Lastly, in case of (3) **model specific properties** we advise for the following levels including their assumptions:

a. model-level assumptions
b. relationship-level assumptions
c. explainability-level assumptions

A categorization into different properties and assumptions is in its consequence a way of *highly structured prompt engineering*. The benefits of structured prompts in context of learning data analysis have been demonstrated in context of ChatGPT (Garg and Rajendran, 2024). The novel aspect that we propose here is a more systematic structuring with respect to data set, algorithm and model properties with a benefit that is two-fold: For students it fosters to think in more structured ways regarding the input, the properties of methods and the output, while for the recommender system (e.g. via ChatGPT) it enables the discovery of more suitable methods and improved explanations, since it is provided with explicit properties and assumptions to account for.

3 PERSPECTIVES IN DIFFERENT DOMAINS

Position 4:
Beyond the Ivory Tower: On Different Preconditions and Practices in Domain Science.

So far we have mostly taken a look with a computer scientist's view in mind. In the following, we include the vision from the perspective of different domains, exemplified in this work by our coauthors from biology, psychology, and bioinformatics.

3.1 Biology and Psychology Perspective

The understanding of data and how to find suitable statistical methods varies widely between and within biological and psychological sciences, and so does the type of data. While ecologists may compare occurrence of a certain species (frequency, re-catch rates), they may also model complex inter-species dynamics. While personality psychologists may be particularly interested in inter-individual differences and how to measure them, clinical scientists may apply pre-post-comparisons for clinical trials, and developmental psychologists may analyze longitudinal or nested data in path models and multilevel modeling. There is definitely not 'one size fits all' in biology and psychology, so the perspective given here is on a very narrow field dealing with morphometric, psychometric, and quantitative parameter data, by no means representative and based on subjective experiences. This perspective also comments on to what extent knowledge on data analysis is (or is not) present among students – again, from a very limited, subjective angle. It seems that biology students are often lacking basic understanding of how to statistically analyze their data beyond reporting descriptive statistics. This may be due to statistics or biostatistics being only a footnote (or one class) in the undergraduate curriculum. Still, students are expected to perform data analysis at the end of their undergraduate studies, and way too often basic statistic education starts within the lab in which they have decided to write their Bachelor thesis. Therefore, we may need to start with the basics: What kind of data do we deal with (continuous, ordinal, nominal)? How is the data distributed (normal versus non-normal)? Is this data dependent or independent? How is the variance distribution (heteroscedasticity) and **why does that even matter?** Can I/should I normalize my data, and if I should, how to do it? How do I find the correct statistical test for my scientific question? A common workflow for many types of parametric biological data with little statistical knowledge may look like this:

- Gather and prepare data
- Test for normality with Shapiro-Wilk test (Shapiro and Wilk, 1965)
- Transform data if not normally distributed with simple transformations (log, log10, exponential)
- Univariate methods: t-test for normally distributed data, Wilcoxon-test (Wilcoxon, 1945) for non-normally distributed data
- Multivariate methods: ANOVA (Fisher, 1935; Girden, 1992) for normally distributed data, PCA for non-normally distributed data (Pearson, 1901)
- Then consider correction for multiple comparisons, e.g., (Bonferroni, 1936; Benjamini and Hochberg, 1995)

More specifically, let us look at two examples. In functional morphology, we use **Geometric Morphometrics** [(Adams et al., 2004), and references within] to study shape using landmark and semi-landmark coordinates that capture morphological features. The resulting Cartesian coordinate data is treated in the following way:

- Conduct **Procrustes Superimposition** (Dryden and Mardia, 1998) (to exclude size as a factor)
- Perform **Principal Component Analysis (PCA)** (Pearson, 1901) using landmark coordinates
- Plot PC1 and PC2, use appropriate statistical test to compare means (e.g. Mann-Whitney U (Mann and Whitney, 1947), Wilcoxon (Wilcoxon, 1945), Dunn's (Dunn, 1964), use correction for multiple comparisons if applicable)
- Test for covariation between analyzed features with **two-block Partial Least Squares (2-block PLS)**

In a second, completely different study, we may apply **3D Surface Texture Analysis** to obtain characteristics of biological surfaces (eggshells, bones, teeth, etc.) [(Attard et al., 2023),(Winkler et al., 2022), (Martisius et al., 2020)]. The obtained surface data are expressed as standardized ISO roughness parameters that are then treated as follows:

- Test for **normality** and **heteroscedasticity** of parameters
- Perform **normalization**
- **Compare means** between groups with appropriate tests (t-test, Wilcoxon (Wilcoxon, 1945), Dunn's (Dunn, 1964))
- Conduct **PCA** to reduce dimensions, as up to 50 parameters are often obtained

This sequence is not wrong, but it is following a basic cookbook structure some students may have acquired from their supervisors, but they may lack the understanding of where to adjust it and have no idea how to advance. Unfortunately, some steps may even be skipped or ignored, if researchers are not aware of their importance; for example, if normality is not tested, the default analysis when comparing means of multiple groups may always be **ANOVA**. If not corrected for multiple comparisons, type I errors may always be inflated. We are not trying to paint a picture of incompetent researchers here, but we need to address the fact that there is no formalized education in data analysis for the biological sciences, and we may have very different competence levels among students and researchers as a result. An accessible and hands-on approach to data analysis using a sample dichotomous decision tree (Breiman et al., 1984) (illustrated by examples) that can be used by researchers and students of different proficiency would be a great tool to support data analysis on a consistent level. From this level, it would be possible to advance to modeling and multivariate methods, which are not as common.

In contrast to biology, research methods and statistics are crucial parts of study programs in psychology, both in undergraduate and graduate studies. While undergraduate curricula are commonly focused on descriptive statistics, exploratory data analysis and basic inferential statistics (e.g., graphical data analysis, correlational methods, t-tests), graduate curricula are more focused on advanced inferential statistics (e.g., multiple regression, ANOVA, non-parametric tests, multilevel linear models, structural equation models) as well as methods exploring clusters and latent factors (factor analysis, cluster analysis; see (Field, 2024) for a popular book on statistical analyses in psychology). However, from our perspective, new algorithms from database machine learning research rarely enter common statistical analyses in psychology, despite the shift from SPSS as the usual statistical software to R, which is more versatile—even though they might be proven useful, particularly for complex multi-level and time series data sets.

3.2 Bioinformatics Perspective

The field of bioinformatics presents unique challenges due to the complexity and diversity of biological, especially OMICs (Li and Wong, 2008), data. Researchers often deal with high-dimensional datasets, noisy measurements, and intricate biological networks. The choice of computational methods significantly impacts the ability to uncover meaningful biological insights. Here, we illustrate how selecting appropriate algorithms can make a substantial difference in bioinformatics research outcomes.

Gene Expression Clustering

- Clustering algorithms are essential for analyzing gene expression data to identify groups of genes with similar expression patterns, c.f. (Eisen et al., 1998).
- Many bioinformaticians default to using **k-means clustering** because of its simplicity and ease of implementation.
- K-means (Jain, 2010) assumes spherical clusters of equal variance and may not capture the true

structure of gene expression data, which often contains irregularly shaped clusters and varying cluster sizes.

- Using density-based clustering algorithms like DBSCAN (Ester et al., 1996) can better identify clusters of arbitrary shapes and is robust to noise. This method can reveal subtle gene expression patterns associated with specific biological conditions or phenotypes that k-means might miss.

Sequence Alignment and Assembly

- Accurate sequence alignment and genome assembly are critical for understanding genetic information.
  - Tools like **BLAST** (Altschul et al., 1990) for alignment and assemblers based on **de Bruijn graphs** (Pevzner et al., 2001) are widely used due to their speed and familiarity.
  - These methods may not handle genomic variations like large insertions, deletions, or repetitive sequences effectively.
  - Employing algorithms such as **Smith-Waterman for local alignment** (Smith and Waterman, 1981) or assemblers like **SPAdes** (Bankevich et al., 2012) and **Canu** (Koren et al., 2017), which are designed to work with long-read sequencing data, can provide more accurate results. These methods account for complex genomic rearrangements and repetitive regions, leading to better assembly quality.

Dimensionality Reduction in Single-Cell RNA-Seq Analysis

Single-cell RNA sequencing (scRNA-seq) (Macosko et al., 2015) generates high-dimensional data that require dimensionality reduction for visualization and interpretation.

- Researchers often use **Principal Component Analysis (PCA)** due to its ability to reduce dimensionality while preserving variance.
- PCA is a linear method and may not capture the non-linear relationships inherent in scRNA-seq data, potentially obscuring meaningful biological variation.
- Non-linear dimensionality reduction techniques like **t-Distributed Stochastic Neighbor Embedding (t-SNE)** (van der Maaten and Hinton, 2008) and **Uniform Manifold Approximation and Projection (UMAP)** (McInnes et al., 2018) preserve local and global data structures, respectively. These methods can uncover cell subpopulations and developmental trajectories that PCA might overlook.

Protein Structure Prediction

- Predicting protein structures from amino acid sequences is fundamental for understanding protein function.
- Traditional methods like **homology modeling** (Schwede et al., 2003) rely on known structures of similar proteins but may not work well for proteins without close homologs.
- Sole reliance on homology models can lead to inaccuracies when templates are distant or unavailable.
- Utilizing advanced algorithms like **AlphaFold** (Jumper et al., 2021), which employs deep learning techniques, can predict protein structures with high accuracy even in the absence of close homologs. Incorporating such methods can significantly enhance the understanding of protein functions and interactions.

Phylogenetic Analysis

Constructing phylogenetic trees helps in understanding evolutionary relationships among species or genes.

- Methods like **Neighbor-Joining (NJ)** (Saitou and Nei, 1987) are popular for their simplicity and speed.
- NJ may not account for varying rates of evolution across lineages or the complexities of genomic data, potentially resulting in incorrect tree topologies.
- Maximum Likelihood (ML) (Felsenstein, 1985) and **Bayesian Inference** (Huelsenbeck and Ronquist, 2001) methods provide more accurate phylogenetic reconstructions by modeling sequence evolution more comprehensively. Although computationally intensive, these methods can yield insights into evolutionary processes that NJ cannot.

Outlier detection in Genomic Data

- Identifying outliers is important for quality control and detecting rare variants.
- Simple statistical thresholds or Z-scores are used to flag outliers.
- These methods may not account for the complex, high- dimensional structure of genomic data, leading to false positives or negatives.
- **Robust Mahalanobis Distance** (Rousseeuw and Driessen, 1999) or **Isolation Forests** (Liu et al., 2008) can detect multivariate outliers by considering the covariance structure of the data. Applying these algorithms improves the accuracy of outlier detection, ensuring that downstream analyses are based on high-quality data.

These examples from bioinformatics demonstrate that the choice of algorithm profoundly influences research findings. By expanding the repertoire of computational methods and making informed algorithm selections, bioinformaticians can enhance the quality and impact of their research. A recommendation system would serve as a valuable tool, guiding researchers and students toward methods best suited to their specific data characteristics and research questions.

Another interesting observation is that while the different tests are prominent and actively used in the different domain sciences, we do not observe them explicitly in the processes of data mining and machine learning, like in, e.g., the KDD process (Fayyad et al., 1996). We would like to remind the reader that the experiences shared and the structures provided are not intended to be regarded as generally valid or by any means complete.

4 CONCLUSION

In this paper we elaborate on the challenges that arise with the richness of different methods for data analytics and the need to educate on decision making of when to use which method. We discuss four positions related to that problem. Those positions encompass (1) there is a rich plethora of methods, which is a blessing and at the same time, in the light of the sheer amount, a curse (2) that automated data analytics pipelines are not a 'holy grail', meaning that to learn when to use which method is of paramount importance (3) computer supported approaches to understand the strengths and weaknesses of methods are indispensable and (4) to facilitate informed decision making across different domains, it is required to first understand their common practices and education approaches for data analytics. In conclusion, we hope that with this position paper we can foster fruitful discussions toward computer supported education of data analytics with the goal of education supported computing across domains.

ACKNOWLEDGEMENTS

The project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Kiel University UP23/1 and University of Lübeck in the context of DenkRaum, an inter- and transdisciplinary fellowship program for postdoctoral researchers.

REFERENCES

Adams, D. C., Rohlf, F. J., and Slice, D. E. (2004). Geometric morphometrics: ten years of progress following the 'revolution'. *Italian journal of zoology*, 71(1):5–16.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Attard, M. R., Bowen, J., and Portugal, S. J. (2023). Surface texture heterogeneity in maculated bird eggshells. *Journal of the Royal Society Interface*, 20(204):20230293.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. P., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477.

Barbudo, R., Ventura, S., and Romero, J. R. (2023). Eight years of automl: categorisation, review and trends. *Knowledge and Information Systems*, 65(12):5097–5149.

Beel, J. and Kotthoff, L. (2019). Preface: The 1st interdisciplinary workshop on algorithm selection and metalearning in information retrieval (amir). In *AMIR@ ECIR*, pages 1–9.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bifet, A., Holmes, G., Pfahringer, B., Read, J., Kranen, P., Kremer, H., Jansen, T., and Seidl, T. (2011). Moa: a real-time analytics open source framework. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 617–620. Springer.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Libreria Internazionale Seeber, Florence, Italy.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, USA.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons, Chichester, UK.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from

volumes of data. *Communications of the ACM*, 39(11):27–34.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Garg, A. and Rajendran, R. (2024). The impact of structured prompt-driven generative ai on learning data analysis in engineering students. In *CSEDU (2)*, pages 270–277.

Girden, E. R. (1992). *ANOVA: Repeated Measures*, volume 84 of *Quantitative Applications in the Social Sciences*. SAGE Publications, Newbury Park, CA.

Huelsenbeck, J. P. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Li, T. H., Degrave, R. J. L., Bickerton, C. M., Meyer, W. J., Velankar, A. A., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., and Phillippy, N. H. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736.

Li, C. and Wong, W. H. (2008). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., and Regev, A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Martisius, N. L., McPherron, S. P., Schulz-Kornas, E., Soressi, M., and Steele, T. E. (2020). A method for the taphonomic assessment of bone tools using 3d surface texture analysis of bone microtopography. *Archaeological and Anthropological Sciences*, 12:1–16.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.

Quaranta, L., Azevedo, K., Calefato, F., and Kalinowski, M. (2024). A multivocal literature review on the benefits and limitations of industry-leading automl tools. *Information and Software Technology*, page 107608.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Schubert, E. and Zimek, A. (2019). Elki: A large open-source library for data analysis-elki release 0.7. 5" heidelberg". *arXiv preprint arXiv:1902.03616*.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). Swiss-model: An automated protein homology-modeling server. *Nucleic Acids Research*, 31(13):3381–3385.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Sim, K., Gopalkrishnan, V., Zimek, A., and Cong, G. (2013). A survey on enhanced subspace clustering. *Data mining and knowledge discovery*, 26:332–397.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Winkler, D. E., Kubo, T., Kubo, M. O., Kaiser, T. M., and Tütken, T. (2022). First application of dental microwear texture analysis to infer theropod feeding ecology. *Palaeontology*, 65(6):e12632.

Xanthopoulos, I., Tsamardinos, I., Christophides, V., Simon, E., and Salinger, A. (2020). Putting the human back in the automl loop. In *EDBT/ICDT Workshops*.