Supplementary: When you have a hammer, why should you consider duct tape? - A Vision of a Data and Algorithm-Aware Choice of Methods

Daniyal Kazempour, Christiane Attig, Peer Kröger, Muhammad Aammar Tufail, Daniela E. Winkler, Claudius Zelenka
{dka,pkr,cze}@informatik.uni-kiel.de,christiane.attig@uni-luebeck.de,mtufail@ifam.uni-kiel.de,dwinkler@zoologie.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Universität zu Lübeck
Germany

1 KNOWING YOUR TOOLS: ON THE NEED OF TEACHING AND LEARNING WHEN TO USE WHAT

Besides the need for an ontology-fashioned structuring on algorithms, and the providence of a recommender platform we have a third pillar: the raise the sensitivity/awareness for learning and understanding of "when to use what algorithm". That is an integral part of making machine learning and database systems accessible to a broad range of domains. One may object that there exist already data mining and machine learning courses for different domains. And this is indeed true for certain subset of domains, e.g. data mining for biologist, or database systems for historians etc.

However, to the best of our knowledge, the offered data mining and machine learning courses, even under the label of interdisciplinarity, do not provide an approach that take data-specific, algorithm-specific, and model-specific properties into focus.

The different properties can be regarded as some type of template. They can facilitate scientists across different domains in (a) thinking about what is important for the choice of algorithm and (b) making them aware of the properties of the input data, the expected output, and the characteristics of the methods. This idea is by itself not novel as we took the pioneering work by Felleisen et. al. [6] as a role model. They introduced students with their works "How to Design Programs" (HtDP) and "How to Design Components" (HtDC) [https://felleisen.org/matthias/HtDC] the means for such a template-based approach. In essence, their contributions are as follows:

- So-called function signatures, that specify the type of the input and of the output
- Unit tests fashioned concept that requires students to think of regular and edge cases that the defined program has to yield the results correctly
- Abstraction patterns that motivate to derive common abstract pattern between different tasks (i.e. for overall sum and overall product the function just differs by the operator itself and the neutral element)

One part of our vision is to establish concepts that are analogous to those proposed in HtDP and HtDC:

- Data analytics signatures: Users specify data-specific properties as far as they can on the different levels
- Expected outcomes on known data: In cases where prior data and their labels are available, users test different methods with different parameter settings to assess which of the methods are more suitable
- Abstraction patterns: Users develop workflows that are generic enough to be applicable to multiple different data, instead of designing for every dataset an own pipeline

2 ON THE NEED TO COMMUNICATE AND EXCHANGE VIEWS AND APPROACHES IN CONTEXT OF ALGORITHM RECOMMENDATION

> "A psychologist, a biologist, a bioinformatician and a computer scientist walk into a bar…"

archetypical for a rich collection of jokes that start like that, this topic is not a joke yet has a punch line we elaborate on in the following:

"Do we even speak the same language?" - Or: the academic tower of babel So far, we have introduced a plethora of structures. What has been neglected so far is that different domains use different terminologies for the same thing. As a concrete example we provide a small set of terms across statistics, machine learning and data mining:

Term in Data Mining	Term in Machine Learning	Term in Statistic
Feature space	Ambient space	Covariable / Design matrix
Lower-dim. projection of data	Embedding of data	
Lower-dim. Subspace (= some hyperplane...)	Latent space / Embedding space	Latent space
Minimization of projection distance	Minimization of loss	Minimization of reconstruction error
Subspace	Latent Space	Manifold
Multi-dimensional	Multi-dimensional	Multi-variate

Table 1: Comparison of Terms in Data Mining, Machine Learning, and Statistics

Note that these terms are by no means 'fixed' to their domains and are also used interchangeably. However in this simple example we can observe that between areas that are highly related (statistics, machine learning, data mining) we can be exposed to

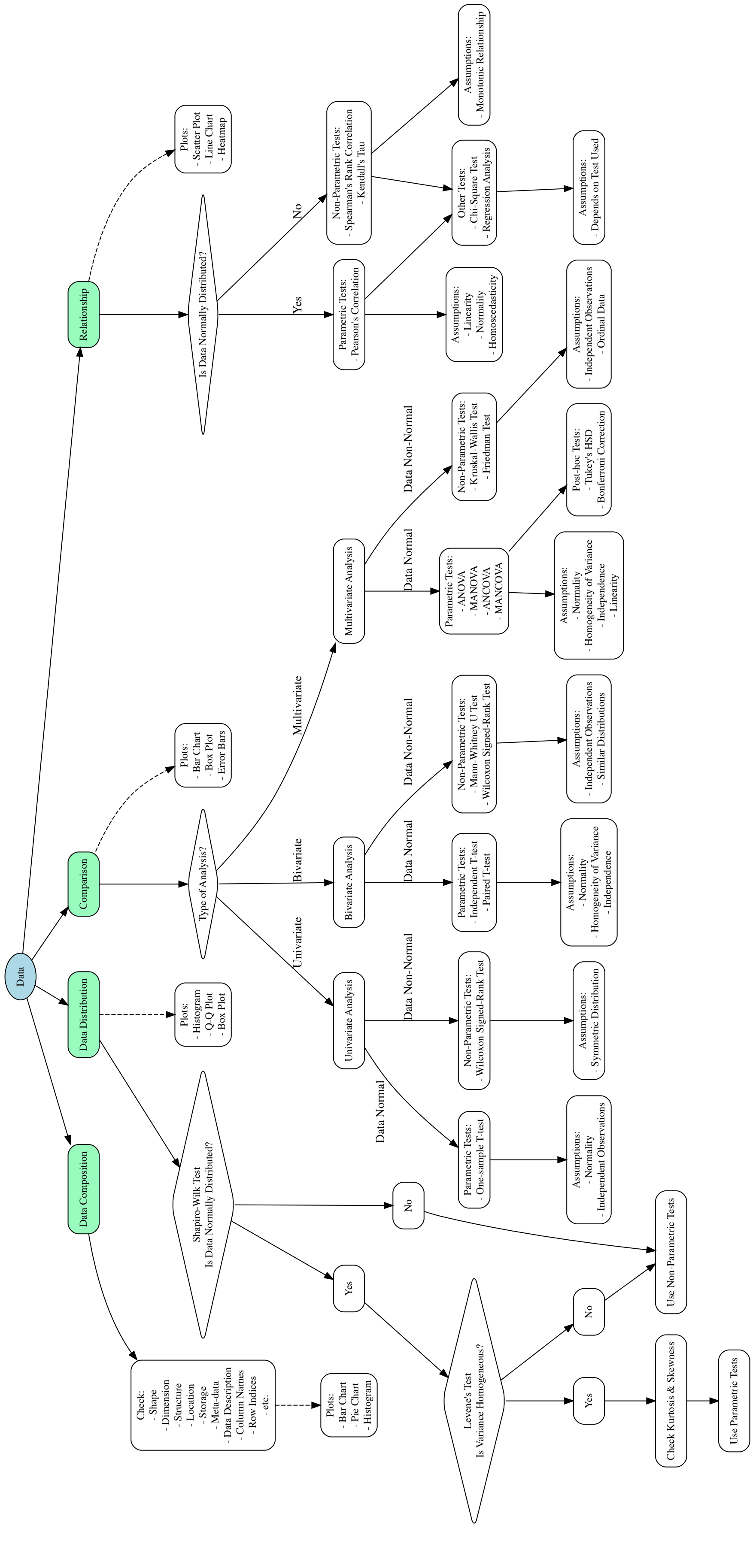

Figure 1: Decision tree on the choice of tests to infer different properties from the data.

different terminologies. This gets even more challenging in the light of different, more distantly related domains like e.g. medical science, linguistics etc. In the light of this vision we hence deem it as necessary to also provide some kind of ‘glossary’ with the recommender system that enhances accessibility of other domains. We hence deem it in the scope of our vision as indispensable to focus to a certain degree on the communication, more specifically the terminology, across different domains to augment the decision on which algorithm(s) to choose.

3 FURTHER BIOINFORMATICS CASES

Genome-Wide Association Studies (GWAS) GWAS [2] identify genetic variants associated with traits or diseases.

- Standard statistical tests like the **Chi-square test** or **logistic regression** are commonly used to associate single nucleotide polymorphisms (SNPs) with phenotypes.
- These methods may not detect interactions between multiple genetic variants (epistasis) or account for population structure adequately.
- Machine learning algorithms such as **Random Forests** [3] or **Support Vector Machines (SVMs)** [4] can model complex interactions and non-linear relationships between genetic variants. Additionally, methods like **mixed-model approaches** can correct for population stratification, leading to more reliable associations.

Metagenomic Data Analysis

- Metagenomics [7] involves studying genetic material recovered directly from environmental samples, leading to insights into microbial communities.
- Taxonomic classification is often performed using tools like **Kraken** [15] or **MetaPhlAn** [14], which rely on reference databases.
- These tools may not classify novel or rare organisms effectively and can miss functional information present in the data.
- **De novo assembly** [16] and **binning algorithms** like **CONCOCT** [1] or **MetaBAT** [9] reconstruct genomes from metagenomic data without reliance on references. Functional annotation tools like **PROKKA** [13] can then predict genes and pathways, providing a deeper understanding of microbial ecology.

4 THE ASK ME (ALMOST) ANYTHING ON DATA: INFERRING PROPERTIES FROM THE DATA VIA TESTS

Prior to the application of data mining and machine learning methods it is advisable to infer certain properties of the data, properties like "Is the data normally distributed?", "If yes, is the variance within the data homogeneous?" etc. Again, like in case of data mining and machine learning methods, we also observe here a plethora of different methods, being accompanied by the question of "when to use which (test)?". In the following we like to show to the reader a decision tree that showcases how an approach towards a recommender system for methods can look like as it can be seen in Figure 1

5 AD ASTRA: ON FUTURE DIRECTIONS

To go even beyond the discussed vision in this work, there are aspects that can be substantially enhanced. One among them addresses the recommender system. While it recommends in the current state of the vision archetypes for a single particular task (e.g. clustering), we aspire to an evolution of this recommender platform such that domain scientists specify the input data and for which patterns they are explicitly looking at, and that the recommender platform suggests not only e.g. one archetype of clustering methods that are suitable, but an entire pipeline of methods, e.g. first applying a global manifold learning method followed by a density-based clustering algorithm. That certain combination of methods can be beneficial in finding patterns in data that otherwise would remain undiscovered has been elaborated on exemplary in the work of [8] where the authors investigate and observe that a combination of manifold learning methods with clustering can lead to the discovery of clusters that otherwise would not be possible by using the respective clustering method alone. Furthermore, we see in the light of transformer architectures and the advent of LLMs such as seen in ChatGPT [11] or Gemini [12] the potential to design highly-customized variants that are capable of utilizing the collected data we envision to yield meaningful recommendations and follow-up questions augmenting the domain scientists in choosing the most meaningful set of methods in their respective use cases. There are certain insights that have to be accounted for as they have been elaborated on by [5] such as e.g. bias towards most recent methods which are, as stated in the introduction of this vision, not necessarily the best recommendation. Another vital aspect is efficiency. As a concrete example: while many embeddings yield high-dimensional vector representations, a question that emerges is "Do we really need all of the dimensions, or would fewer already suffice? How can we determine that we have considered ’enough’ features?". First endeavors in the realm of molecule embeddings have been proposed in the work of [10]. The aspect of compression is especially vital in the light of massive amounts of data available as in omics settings. Lastly a message that we wish to convey in scope of this vision is, to keep in close and regular exchange between the different domains. Or, to adhere to the theme of this section:
Ad astra, simul - to the stars, together.

REFERENCES

[1] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Meta M. Schirmer, Joshua Quick, Nicholas J. Loman, and Anders F. Andersson. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods* 11, 11 (2014), 1144–1146. https://doi.org/10.1038/nmeth.3103

[2] David Altshuler, Mark J. Daly, and Eric S. Lander. 2008. Genetic mapping in human disease. *Science* 322, 5903 (2008), 881–888. https://doi.org/10.1126/science.1156409

[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[4] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297. https://doi.org/10.1007/BF00994018

[5] Yashar Deldjoo. 2024. Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems* (2024).

[6] Matthias Felleisen, Robert Bruce Findler, Matthew Flatt, and Shriram Krishnamurthi. 2018. *How to design programs: an introduction to programming and computing*. MIT Press.

[7] Jo Handelsman, Matthew R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 5, 10 (1998), R245–R249. https://doi.org/10.1016/S1074-5521(98)90108-9

[8] Moritz Herrmann, Daniyal Kazempour, Fabian Scheipl, and Peer Kröger. 2024. Enhancing cluster analysis via topological manifold learning. *Data Mining and Knowledge Discovery* 38, 3 (2024), 840–887.

[9] Dongwan Kang, James Froula, Rob Egan, and Zhong Wang. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3 (2015), e1165. https://doi.org/10.7717/peerj.1165

[10] Daniyal Kazempour, Anna Beer, Melanie Oelker, Peer Kröger, and Thomas Seidl. 2021. Compound Segmentation via Clustering on Mol2Vec-based Embeddings. In *2021 IEEE 17th International Conference on eScience (eScience).*

IEEE, 60–69.

[11] OpenAI, Josh Achiam, and Steven Adler et.al. 2024. GPT-4 Technical Report. arXiv:cs.CL/2303.08774 https://arxiv.org/abs/2303.08774

[12] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[13] Torsten Seemann. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 14 (2014), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

[14] Duy T. Truong, Eric A. Franzosa, Tobias Tickle, Morgan Scholz, James G. Weingart, Curtis Huttenhower, and Nicola Segata. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12 (2015), 902–903. https://doi.org/10.1038/nmeth.3589

[15] Derrick E. Wood and Steven L. Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15, 3 (2014), R46. https://doi.org/10.1186/gb-2014-15-3-r46

[16] Daniel R. Zerbino and Ewan Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18, 5 (2008), 821–829. https://doi.org/10.1101/gr.074492.107