Extended Abstract for the
Symposium on Integrating AI and Data Science into School Education Across Disciplines 2025

DATA ARE AT THE CENTER OF DATA SCIENCE: MY TAKE ON WHAT EVERYONE SHOULD KNOW ABOUT DATA

Arne C. Bathke
IIntelligent Data Analytics (IDA) Lab, Department of Artificial Intelligence and Human Interfaces (AIHI), Faculty of Digital and Analytical Sciences (DAS), Paris Lodron University of Salzburg (PLUS), Austria, Arne.Bathke@plus.ac.at

Focus Topics: Data Science Education & Social Good, Data and Problems

Data Are at the Center of Data Science. My Take on What Everyone Should Know About Data.

Much attention is currently given to the algorithms that are embedded in the tools that are now subsumed under the buzzword of artificial intelligence (AI). While the understanding of the algorithms and models is undoubtedly important and highly informative regarding their limitations, we find that the importance of the underlying data and their quality is often being neglected. However, the consequences of suboptimal data quality can be quite dramatic, also when viewed in comparison with the consequences that can arise from suboptimal algorithmic tools. Therefore, this contribution focuses on data and their central importance for the overlapping areas of data science, statistics, and artificial intelligence.

To this end, we will review the essential workflow (here presented as ADD-PIC) that underlies sensible statistics and data science. Afterwards, we will discuss in detail the important role of data in this workflow, spiced with some typical mistakes and pitfalls, generated, among others, by biases and seemingly paradoxical situations.

Based on this workflow, we will propose some key points that students need to know about the concept of data in order to do statistics and data science in a sensible way, and to understand better the possibilities and limitations of AI.

By no means are my points breathtakingly new. In fact, much of what is summarized in this contribution has in the past been called “statistical thinking”. However, the current discussions around AI make it quite apparent that there is still in the 2020s, more than a century after the invention of inferential statistics, a widespread lack of basic conceptual statistical thinking which should be at the basis of any inference. Basic statistical thinking includes the thinking about data and their potential and limitations. Ignoring the basics of statistics in turn leads to a widespread rather naïve approach to and interpretation of the plethora of data-informed tool boxes that are available nowadays. I argue that the potentials and limitations of AI can only be appreciated if users (from school children to top scientists) have at least a basic understanding of quantitative, statistical thinking, including the role of data.

The Data Workflow: ADD-PIC

When trying to structure the workflow of gaining insights using quantifying information, the following six-step procedure that can be abbreviated with the acronym *ADD-PIC* may be helpful:

(A) Asking sensible, relevant questions
(D) Data acquisition
(D) Description and quality check
(P) Prediction and generalization
(I) Interpretation
(C) Communication

Many statistical textbook introductions start their explanation of the data workflow with descriptive statistics (e.g., Dalgaard, 2002), whereas some also include an explicit introductory chapter on data acquisition (e.g., Oestreich & Romberg, 2012). However, we are not aware of mainstream introductory statistics books that emphasize the importance to start with asking a sensible, relevant question before attempting to make sense of data or information. The reason is simple: You can always find *something* in data, but if you don’t know what you have been looking for, what you find could

In S. Podworny & S. Schönbrodt (Eds.), *Towards Fostering AI and Data Science Literacy in Schools Across Disciplines. Proceedings of the 1st Symposium on Integrating AI and Data Science into School Education Across Disciplines (AIDEA1 2025)*, Salzburg, Austria.

likely be artefacts. Also, the question often determines which data will be appropriate in order to provide the grounds for meaningful insights. Thus, we have explicitly included this point into our data workflow.

Data acquisition includes the task to decide which variables will be needed in order to answer the formulated question, and based on what data source(s). The key criterion here is representativity of the acquired data for the descriptions, predictions, and inference that are to be made. Much of this can be technically formalized, namely in the selection of an appropriate study design and by following the guidelines from the Equator network (Equator network, 2024). Nevertheless, a surprisingly good judge of whether a data acquisition method will be appropriate for the actual needs and the underlying question is provided by our common sense and life experience, ideally sharpened by the knowledge of the key statistical concepts (in particular the concept of *bias*) and some common mistakes to look out for (e.g., misinterpretations of Simpson's or Berkson's paradox).

However, for many people it appears less intuitive that it is not always desirable to acquire more variables or more data. In fact, having good data (high quality data with value for the underlying question) is more important than striving for big data. Therefore, we consider it important to illustrate this point also with real data examples.

The next steps in the workflow, namely description and quality check, as well as prediction and inference, are standard topics in statistical textbooks, whereby the proposed predictive and inferential methods often constitute the flavor of the different statistical schools, and they naturally also vary depending on the assumed technical knowledge of the readers. In any case, the already classical quote by G.E.P. Box that "all models are wrong, but some are useful" still holds true and is worth being reflected upon.

When interpreting the results, the question posed at the outset should be kept in mind. And finally, the results and insights need to be communicated. To whom? At what technical level? Statisticians / data scientists from traditional degree programs have spent little time on the communication of statistical findings. However, this aspect is important for different reasons, involving not the least its effects on the popular perception of science in general and statistics in particular.

Of course, a run through some of the items in the workflow may reveal that an initial question was not ideal and should be adapted or abandoned altogether, perhaps triggering a restart of the ADD-PIC workflow. Indeed, the workflow is not to be understood as a simple linear path, but every step can inform towards a return to a previous one, and not all steps need to be executed for every question posed. For example, it may not be needed to make predictions when the goal is a simple description of the data with a particular question in mind.

And, each step in the workflow invites not only technical questions, but also questions regarding the dimensions of ethics, values, and regulation, where the ethical dimension could, for example, be evaluated according to the FAT framework criteria of fairness, transparency, and accountability, and further how these criteria relate to the different stakeholders, namely data subject, model subject, data scientist, project manager, and general public (see also Martens, 2022).

References

Dalgaard, P. (2002). Introductory Statistics with R. *Springer Series on Statistics and Computing.* Springer, New York, Berlin, Heidelberg.

Martens, D. (2022). Data Science Ethics. Oxford University Press, Oxford.

Oestreich, M., Romberg, O. (2012). Keine Panik vor Statistik. 4. Auflage. *Springer Spektrum.* Springer, New York, Berlin, Heidelberg.