# Data science for informed citizen: Learning at the intersection of data literacy, statistics and social justice

Joachim Engel

Ludwigsburg University of Education

Focus Topics:  *(1) AI and Data Science Education for Social Good;*

*(2) AI and Data Science Curricula and Implementation in School*

## Why data science for all?

The information landscape is changing dramatically in the digital age due to the increasing availability of information via the internet, the widespread use of digital technologies, the abundance of data and easy access to data analysis tools. Digital media and the availability of data of sheer unlimited scope and magnitude change our access to information in radical ways. Emerging data sources provide new sorts of evidence, provoke new sorts of questions, make new sorts of answers possible and shape the ways in which evidence is used to influence decision making in private, professional and public life. In an increasingly data-driven world, social, societal and technological change requires new competencies. This expansion affects not only the professional world, but all of us. Innovation, social progress, and the well-being of our civil society require that people in science, business, politics, and society know how to evaluate and make sense of data to develop an informed understanding of our world and address pressing societal challenges with empirical insights and sound data-driven arguments.

## Promises and threats of Big Data

Big Data holds significant promise for ensuring a liveable future on our planet by offering insights and solutions across various domains, e.g., through environmental monitoring and conservation, efficient resource management, urban planning, disease prevention and more. At the same time, Big Data, with its possibilities for surveillance, manipulation, and control, poses serious problems for democracy and freedom (see, e.g., Helbing et al., 2017; Atenas et al., 2020). The ability to assess the credibility of information and its sources has never been more important. The World Risk Report[[1]](#footnote-1), published by the Swiss World Economic Forum Foundation in January 2024, sees misinformation and disinformation as the greatest threat to humanity over the next two years, ahead of extreme weather events, social polarization and armed conflict.

Algorithms drawing upon data are used to profile members of society and make crucial decisions which likely disproportionately impact those with less privilege and resources at their disposal. Amazon's model for sorting job applications[[2]](#footnote-2), for example, proved to be anti-women. Facebook's problems were first exposed by the Cambridge Analytica scandal, and the company continues to struggle with many ethical issues. Cathy O'Neil, in her book Weapons of Math Destruction (O’Neil, 2016), points out the dangers and injustices of using algorithmic models to determine credit scores, the price of insurance policies, whether someone should be paroled, or even what crimes police should investigate. Failure to learn how to understand, analyze and challenge data will result in citizens being in a continuously increasing position of informational disadvantage in relation to socio-political and commercial actors. Consequently, data literacy education needs to address a broad vision of data as social as well as technical assemblages. As consequence, data literacy and data science education cannot be reduced to learning technical mastery about algorithms, big data management and computing.

With all the promises of *Statistical Science to make a better world* (so a slogan of the International Statistical Institute), there are serious ethical concerns when more and more human activities are transcribed into data, quantified and analyzed (Van Es and Schäfer, 2017). Decisions taken by corporations and government agencies are increasingly data- and algorithm-driven, while the processes through which data are generated, communicated and represented are neither necessarily transparent nor devoid of negative effects (O’Neil, 2016). People are often unaware why, how or even that data about themselves are being collected, analyzed and ‘shared’ with additional parties (Dalton et al., 2016). In an increasingly datafied society, data are often given the status of objective fact, despite its constructed, partial and biased nature.

## Data Science for Informed Citizens: Overarching Objectives

Based on literature review, an analysis of needs to strengthen democratic values in the digital age and the reflection of own teaching practices, this talk aims to provide guidance on how to design data science education for informed citizens. No specific mathematical or technological background beyond high school is assumed, so the concept and goals may apply to any group of educated and informed citizens. We focus on how Data Science uses machine learning algorithms as one of its methodologies to analyze data, make predictions and automate decision-making processes.

* *Raise awareness where in daily life we encounter Data Science product*s: Here are some examples: Smartphones, E-Commerce and Online-Shopping, fraud-detection; Banking and Finance; Education and Apps
* *Teach awareness about data quality and data suitability:* Data-based evidence (in connection with critical thought) is the route to new knowledge. However, data are not the “objective truth” but biased and influenced by interests of someone, a result of conscious decisions by someone to research a certain topic.
* *Teach awareness about biases of machine decisions:* Any bias in the training set will be amplified in the test set, leading to biased decisions. The design of the algorithm itself can introduce bias
* *Teach awareness about the impact of Data Science products on society:* Topics that are accessible to students at school or university include privacy concerns, surveillance, bias and fairness, autonomy, and the future of work. A particular concern refers to the opacity (often referred to as “black boxes”) of most machine learning algorithms. Delegating decision-making to a machine, especially in critical areas such as criminal justice or life-and-death medical decisions, raises severe ethical and moral questions
* *Teach technological basics about machine learning:* Of course, Data Science Education is also about learning about technology. In my experiences (Erickson & Engel, 2023), decision trees are a good entry point to understand principles of machine learning. Trees are intuitive, easy to apply and transparent. Depending on the group of learners, the topic of trees can be extended to more sophisticated ML-topics like random forests, bagging and boosting.

## Conclusions

*Motivation problem definition and context:* Data analysis must be motivated by a goal. And it must be embedded in a clear context in which it is to be applied and informed. A good Data Science application solves a well-defined problem or answers a specific question. This is the hard work that needs to be done before applying the automated tools. And it is one of the hardest things for students to learn and internalize at school and university.

*Data provenance and metadata:* The most sophisticated analysis is worthless if it is based on weak or questionable data. The context of the problem to be addressed is crucial for assessing the required relevance and quality of the data. Data analysis must not be conducted blindly, applied to data that is inappropriate or full of errors and gaps. Students must learn to document their data sources and their origin. And, more importantly, to be skeptical about the reliability of their data.

*Human-machine interaction and decisions:* Analytics must be a collaboration between human analysts and computer algorithms, with the algorithms serving as tools operated by humans. It is the human analyst who can adapt to changing circumstances, recognize the limitations of the model, understand the constraints of the data set, evaluate and correct, exclude or consider exceptional and deviant values, and understand the potential unintended consequences of a model that optimizes a criterion.

*Ethics:* Increasingly, ethical consequences of Data Science analysis are being uncovered. We must not rely on algorithms and must train our students to think and act ethically and apply these principles to their work. Students should learn to ask why an analysis is being performed and consider the ethical consequences of the answer. While many in the Data Science field view models as objective and unbiased, O'Neil (p. 21) defines models as "opinions embedded in mathematics." While the math gives the model the appearance of objectivity, in reality someone created the model and decided what data to use, what variables to include, what model form to use, and so on. A model is really an opinion that reflects both the bias of the modeler and the bias of the data itself. Those studying Data Science need to be sensitized to these ethical issues and trained in how to avoid bias and discrimination in models.

*Problem solving:* Sure, we also need to teach technical skills such as programming, machine learning algorithms and other big data topics. But that should not be the focus of a Data Science curriculum any more than calculus should be the focus of a physics curriculum. These are tools, and students should be good at them – but first they need to learn *why* and how to use them. The ultimate measure of success is solving the problem at hand by providing sustainable solutions that have tangible impacts.

Students should learn early in their education that Data Science is NOT about the tools! Data Science tools, no matter how powerful, are a “how”, not the “what”. Ultimately, it's not about knowing and using the tools well, it's about finding and using sustainable solutions to difficult problems. Otherwise, we shouldn't be surprised if the brightest minds we train use their brain power primarily to encourage other people to click on certain consumer ads rather than to use their knowledge to solve pressing social and societal problems.

## References

Atenas, J., Havemann, L., & Timmermann, C. (2020). Critical literacies for a datafied society: academic development and curriculum design in higher education. *Research in Learning Technology*. 28: 2468. <https://doi.org.10.25304/rlt.v28.2468>

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: a dialog on data and space. *Big Data and Society*. 3 (1), 1–9. https://doi.org/10.1177/2053951716648346

Engel, J. (2017). Statistical literacy for active citizenship: a call for data science education. *Statistics Education Research Journal 16(1), 44-49* https://doi.org/10.52041/serj.v16i1.213

Erickson, T., & Engel, J. (2023). What goes before the CART. Introducing classification trees with ARBOR and CODAP. *Teaching Statistics*, 45, S104–S113.

Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. & Zwitter, A. (2017). Digitale Demokratie oder Datendiktatur. In: C. Könneker (Ed.), *Unsere digitale Zukunft.* https://doi.org/10.1007/978-3-662-53836-4\_1

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality & Threatens Democracy.* Crown Publishing Group.

Richterich, A. (2018) *The Big Data Agenda: Data Ethics and Critical Data Studies*. University of Westminster Press, London. https://doi.org/10.16997/book14

Van Es, K. & Schäfer, M. T. (Eds). (2017) *The Datafied Society: Studying Culture through Data*. Amsterdam University Press. <http://library.oapen.org/handle/20.500.12657/31843>

Zweig, K. (2019). *Ein Algorithmus hat kein Taktgefühl: Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*

1. <https://www.weforum.org/publications/global-risks-report-2024/> [↑](#footnote-ref-1)
2. 2<https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html> [↑](#footnote-ref-2)