DATA SCIENCE FOR INFORMED CITIZENS: ON SYNERGIES BETWEEN DATA LITERACY AND CIVIC STATISTICS

Joachim Engel and Laura Martignon
Department of Mathematics and Computer Science
Ludwigsburg University of Education, Germany
engel@ph-ludwigsburg.de

Data science as a practical science has been conceived to address tangible problems in science, technology, and society. These problems require skills and dispositions beyond the technical mastery of algorithms such as interrogating measurability issues, classifying results under uncertainty and risk, and being aware of data ethics and implications for policy and society. These features relate data science education closely to the recently developed field of civic statistics. This conceptual paper looks at the common grounds between these two areas. It investigates how digital literacy and data science can enhance civic statistics and, vice versa, how civic statistics concepts have the potential to enrich the teaching of data science and data literacy. Hence the paper contributes to the development of new curriculum guidelines and, ultimately, courses.

BACKGROUND

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 also 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 (Ridgway, 2015). 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). Algorithms drawing upon data are used to profile members of society and make crucial decisions that likely disproportionately impact those with less privilege and fewer resources at their disposal (O'Neil, 2016). 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.

With all the promises of "*Statistical Science to Make a Better World*" (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 & 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 of 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 facts, despite its constructed, partial, and biased nature. As consequence, data science education cannot be reduced to learning technical mastery about algorithms, big data management, and computing.

THE EMERGENCE OF DATA SCIENCE EDUCATION

Data science was conceived as a practical science to solve concrete problems in science, technology, and society. As a strongly interdisciplinary field, data science refers to a set of skills and techniques that include statistics, computer science (coding, visualization, computing with data, machine learning, etc.), mathematics, and expertise in the subject area from which the data originate. A compact description defines data science as the science of learning from data. Participation in democracy, in today's digital and datafied society, requires the development of a series of transversal skills, which need to be fostered in educational institutions through critically oriented pedagogies that interweave technical data skills and practices together with information and media literacies (Engel, 2017). Data science education, as a science still in its infancy, must look beyond a combination of

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

numerical, statistical, and technical capabilities; include critical thinking and citizenship (Biehler et al., 2022); and foster skills to evaluate, analyze, and interpret data. Educational programs must look beyond data capabilities and include critical thinking (Van Es & Schäfer, 2017) and their meaning for policy and society (ProCivicStat Partners, 2018). Such an approach can empower students to question the ethics, structures, and economics of data use, and fundamentally, the apparent *inevitability* of the surveillance and datafication of all aspects of daily life (Atenas et al., 2020).

Classroom approaches to data science can often be linked in terms of content to addressing societal issues that are a concern of many, such as climate change, pandemic dispersion, income equity, etc. (Engel, 2017). Nonpartisan organizations have compiled a wealth of information that is publicly available on the Internet for anyone to use for information and discussion—from the United Nations' work on Sustainable Development Goals to measure social progress; to national statistics offices that collect information on employment, income, and migration; to nongovernmental organizations that monitor climate change or citizen health (Ridgway, 2015). Platforms such as Gapminder (https://www.gapminder.org) or Our World in Data (https://ourworldindata.org/) provide low-threshold access to monitoring the state of the world, from human development and global happiness ("World Happiness Report") to COVID-19 infection rates and climate change.

By now, there are several concepts, proposals, and experiences to introduce elements of data science into middle and high school classrooms. The International Data Science in Schools Project (IDSSP, http://www.idssp.org) is an international collaboration of statisticians, computer scientists, and educators, who developed curricula to introduce data science to students in their last two years of high school as well as a curriculum to empower teachers to teach data science. Other recent innovative initiatives are, e.g., Project Data Science and Big Data at School (https://www.prodabi.de) and Mobilize's Introduction to Data Science (https://www.mobilizingcs.org). Data science education needs to be part of the general educational mission of schools in the 21st century (see Data Literacy Charta, https://www.stifterverband.org/sites/default/files/data-literacy-charter.pdf) and concerns mathematics education in a special way, in addition to computer science and social science education (Messy Data Coalition, 2020; Wolfram, 2020). These developments pose a fivefold challenge that goes beyond the application of certain statistical methods of the previous widely taught lessons (ProCivicStat Partners, 2018):

1. *Substantive, contextual*: The available data must be assessed for relevance, reliability, and credibility to the initial questions; the extent to which the conclusions drawn are supported by the data and the limitations of the conclusions must be examined.
2. *Didactic*: Interdisciplinary and cross-curricular collaboration, e.g., with social sciences and computer science, is necessary and requires familiarity with subject didactics and teaching methodology of different subject traditions in addition to possible team teaching.
3. *Statistical*: Appropriate visualizations and statistical summaries must be selected; multivariate phenomena must be examined for explanatory third variables; and any causal inferences must be examined for validity.
4. *Technological*: Innovative technologies and software (CODAP—https://codap.concord.org/, Gapminder—https://www.gapminder.org, Python—https://www.python.org/, R—https://www.r-project.org/, etc.) are being used to visualize and explore multivariate datasets; poorly prepared multivariate datasets downloaded from the Internet may first need to be curated (cleaned and processed) for analysis and visualization.
5. *Systemic*: Skills in the area of data literacy and data science education cannot be delegated to a single subject in the curriculum, be it mathematics, computer science, social studies, etc. but transcend the traditional compartmentalization of disciplines. This poses severe challenges to the organization of school curricula and the training of teachers.

CIVIC STATISTICS AND DATA SCIENCE EDUCATION

At the heart of data science are skills to collect, manage, evaluate, and critically derive new knowledge from data. This explicitly includes the ability to critically evaluate data and its impact on social and political interactions. Data science education needs to go beyond a combination of numerical, statistical, and technical skills, to include critical thinking and citizenship, and to foster skills in evaluating, analyzing, and interpreting data. Such a concept of data literacy has many overlaps with civic statistics. Following the term "statistical literacy," the international strategic partnership, ProCivicStat (with participation of the Universities of Durham, Haifa, Ludwigsburg, Paderborn, Porto, and Szeged) has conceptualized a sub-discipline called civic statistics (Engel et al., 2019; Ridgway, in press). Civic statistics focuses on understanding and critical reflection of statistical information about society, as provided by the media, statistical offices, and other statistics providers. Civic statistics addresses the socio-political dimension of mathematics education from a practical, curricular perspective by providing a cognitive framework for learning from data about society and specific curricular materials. Understanding topics that inform social processes, social and economic well-being, and perceptions of civil rights is essential for civic engagement in modern societies, but such engagement is often based on complex multivariate data whose interpretation and indexing requires knowledge not usually taught in regular mathematics and statistics classes, let alone in politics or civics education. Competencies in civic statistics are necessary for informed participation in democratic societies. Numerous concrete teaching materials on civic statistics were developed by ProCivicStat and are freely available at https://iase-web.org/islp/pcs.

Obviously, there is quite some common ground between the teaching of civic statistics and the teaching of data science. Both disciplines are designed to address real problems that have an impact on society, and both explore tangible solutions based on data. These data are likely to be large and messy, publicly accessible through the web, questionable in quality, and tell a story rooted in some context of high relevance. How can the two disciplines benefit from and enrich each other?

Gal et al. (in press) identify 11 facets and tools of civic statistics that are needed (by students and citizens in general) so that they can critically understand the statistical information that they see/read/hear and engage with the underlying societal or economic issues. Figure 1 provides a compact overview of these facets. Gal et al. (in press) structure these separate but related facets in three groups: (a) engagement & action, (b) knowledge, and (c) enabling processes. For further details, the reader is referred to the original paper. Here we discuss (in due brevity) how these civic statistics facets relate to data science with the intention to facilitate the development of new curriculum guidelines and, ultimately, courses.

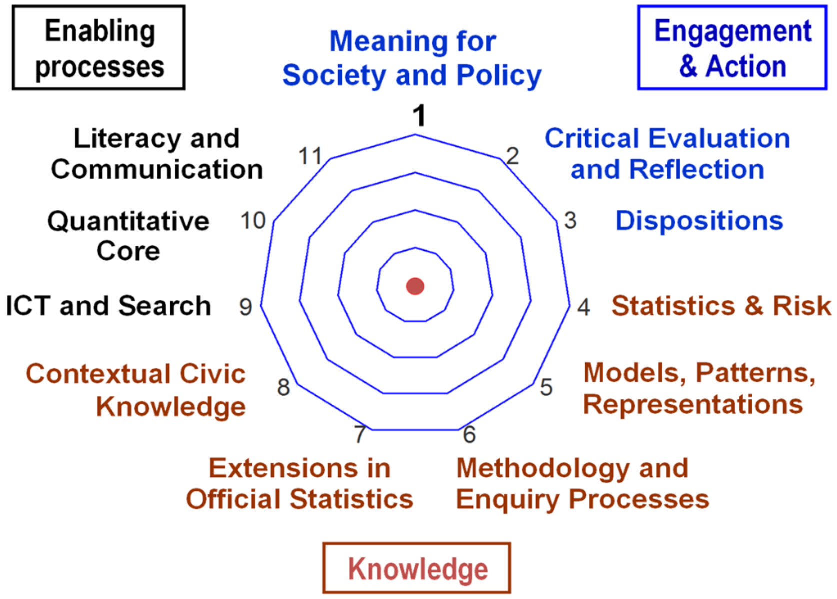
Enabling processes
Meaning for Society and Policy
Engagement & Action
1
Literacy and Communication
11
2
Critical Evaluation and Reflection
Quantitative Core
10
3
Dispositions
ICT and Search
9
4
Statistics & Risk
Contextual Civic Knowledge
8
5
Models, Patterns, Representations
7
6
Extensions in Official Statistics
Methodology and Enquiry Processes
Knowledge

Figure 1. Eleven facets that characterize the nature of civic statistics (from Gal et al., in press)

Engagement & Action

This group involves three facets, and relates to the motivation for generating civic statistics, engaging with them, and investing in their critical evaluation.

1. *Meaning for society and policy*: Like civic statistics, every application of data science starts with a tangible problem to be addressed. This problem may be a "burning hot" issue that affects socio-economic well-being or individual civil rights as in civic statistics, or it may also address a technical challenge (e.g., automated language translation, face recognition) or logistical problem (e.g., how

to improve taxi services in Berlin, reduce delays in air traffic). Nevertheless, because data science tools are very powerful, it is hard to imagine that their application to whatever problem would not have a severe impact on society. Therefore, data science education needs to reflect on the impact of their proposed solutions on individual citizens, society, and social policy. Students of data science need to learn the role of data for decision making and for influencing policy. These decisions involve choices and risks, weighing existing evidence, options, and their probabilities, costs and benefits, expected values, and subjective utilities. Possible implications for different groups of stakeholders should be considered, particularly hot topics when dealing with personalized data relates to issues of privacy protection.

2. *Critical evaluation and reflection*: In data science courses just as in civic statistics we need to worry about the quality of evidence, its timeliness, and its relevance. Data analysis should be complemented by discussions of plausible causal factors and the likely social and policy implications of actions, both in terms of immediate impact and in terms of likely longer-term effects. This involves raising a critical understanding of how data are produced, variables operationalized, and how they can be used for particular purposes, including the role of context in interpreting data. It emphasizes developing an awareness for data ethics and considering the implications for policy and society when powerful algorithms are used.
3. *Dispositions*: The term, disposition, is an umbrella term that refers to a cluster of related but distinct concepts, including motivations, beliefs, attitudes, and emotions. Dispositions include attitudes towards evidence and personal sentiments regarding mathematics, uncertainty, and risk. They are related to habits of mind for dealing with statistical information (Lee & Tran, 2015). Dispositions are of particular importance if we envision citizens who actively engage with and critically reflect on the impact of data science outcomes or civic statistics, because they have components (e.g., self-efficacy, self-confidence) that may affect engagement in both positive and negative ways (Gal et al., in press). Dispositions have not only a personal dimension (a willingness to engage with and to devote time to understanding the information that is being presented) but also a social dimension (a willingness to share opinions and alternative interpretations with others). A key disposition is critical stance (Gal, 2002)—a willingness to engage with statistics and quantitative evidence related to social issues.

Knowledge

This group involves five facets. Together, these encompass the diverse knowledge bases and skills that pertain directly to the statistical information that may be contained in messages about civic statistics. Note that this involves the understanding of the *context* and domain knowledge from which data are generated and to which the statistics refer.

4. *Statistics and risk:* Statistics and risk encompasses much of what is commonly taught in introductory statistics college courses and in the areas of statistics or data analysis at the high-school level, although the emphasis for civic statistics is different (Gal et al., in press). This is analogous for data science, although the specific statistical emphases in the two fields are not identical. A discussion of the statistical knowledge required for good data science practice is an ongoing current debate that goes well beyond the scope of this short paper (see, e.g., De Veaux et al., 2017).
5. *Models, patterns, and representations*: Models, patterns, and representations are important for both fields, with different emphases. A major distinction between machine learning and classical statistics is their purpose (Breiman, 2001). While predictive modeling, e.g., in machine learning, is designed to explore patterns and make accurate predictions, generative models in statistics are designed for describing and producing inferences about the relationships between several variables. In prediction models, conclusions lose validity if the test sample and training sample are not from the same distribution. Generative models are based on distributional assumptions and a-priori assumptions about the data collection process.
6. *Methodology and enquiry processes*: Understanding random samples and the concept of representativeness are at the heart of classical statistics and form the basis of inference; knowing about various sampling designs (e.g., randomized experiments versus observational studies) is the basis for assessing the validity of causal conclusions.
7. *Extensions in official statistics*: While relevant for civic statistics, knowledge about official statistics may be of importance for data science only if applied to specific domains of data.

8. *Contextual knowledge*: Contextual knowledge is just as relevant for data science as it is for civic statistics. Problems need to be clearly defined and fully understood; data crunching alone is useless and potentially dangerous.

Enabling Processes

Enabling processes involve general skills that are essential for finding, accessing, comprehending, and communicating messages; ensuring data quality; and ensuring ethical use of data.

9. *Information Communication Technology (ICT) and search:* ICT and search is at the central core of data science. It includes searching and retrieving data from the web and web-scraping, sanity checking the basic properties of data, remediating artifacts and anomalies, and also employing data moves such as grouping, data transformation, smoothing, subsetting, etc. It may involve combining data from different databases with different formats and requires computing with data.
10. *Quantitative core*: Basic numerical skills apply equally to civic statistics and to data science.
11. *Literacy and communication*: The purpose of any investigation in civic statistics as well as in data science is a well-defined problem rooted in some specific domain. Comprehending the core of the problem is essential and requires literacy skills as well as being able to read fluently and absorb the overall sense of an article or report as well as the detailed statistical information often embedded in text messages. Texts are often very dense; it is then an essential ability to understand the text and relate it to the statistical messages shown. Conclusions of the analysis need to be communicated to a larger body of people or society at large. Competencies regarding literacy and communication apply equally well to civic statistics as to data science.

CONCLUSION

The two recently developed fields of data science education and civic statistics cover a lot of common ground. They can benefit much from each other. Good data analysis must be motivated by a goal: it contributes to the solution of a clearly defined problem or answers a specific question. The best of analyses is worthless if it is based on weak data. The data need to be reliable, of adequate quality, and capable of providing an answer to the initial question. The context of the problem being addressed is critical in assessing the required relevance and quality of the data. Students must learn to document their data sources and to evaluate them. Algorithms are powerful tools for problem solving. But it is ultimately the human who selects and applies a tool, who knows the limitations of the tool, understands the limitations of the data set, and can understand the possible unintended consequences of an algorithm that optimizes a criterion. Finally, students should learn to ask why an analysis is being performed and consider the ethical consequences of the answer.

Responding to the challenges of a datafied society is not only a matter of curriculum design or the responsibility of individual instructors but also requires institutional strategies and policies to support educators in developing data literacies. This requires a coherent plan for systemic change. In some contexts, this could begin by infusing elements of data literacy and statistics into otherwise traditional courses. In other contexts, it might be appropriate simply to use authentic large-scale data sets relevant to social problems to teach traditional topics. In other contexts, it may be necessary to engage in radical curriculum reform (ProCivicStat Partners, 2018).

Data science has great promise to address some of the most burning problems of the future of this planet. But it requires more than technological know-how and algorithmic skills. Critical reflections are essential. In today's world, data science is relevant to nearly every large policy, social, or personal decision, in contexts including, e.g., COVID-19, evaluating the news, climate change, and organizational decision-making. We should encourage our brightest students to use their talents by adopting powerful data science tools to contribute to solutions of burning issues for humankind!

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*, Article 2468. https://doi.org/10.25304/rlt.v28.2468

Biehler, R., De Veaux, R., Engel, J., Kazak, S., & Frischemeier, D. (2022). Research on data science education [Editorial]. *Statistics Education Research Journal, 21*(2), Article 1. https://doi.org/10.52041/serj.v21i2.606

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

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

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., … Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

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

Engel, J., Biehler, R., Frischemeier, D., Podworny, S., Schiller, A., & Martignon, L. (2019). Zivilstatistik: Konzept einer neuen perspektive auf data literacy und statistical literacy. *AStA Wirtschafts-und Sozialstatistisches Archiv, 13*(3–4)*,* 213–244. https://doi.org/10.1007/s11943-019-00260-w

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Gal, I., Nicholson, J., & Ridgway, J. (in press). A conceptual framework for civic statistics and its educational applications. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens*. Springer.

Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R., & Zwitter, A. (2017). Digitale demokratie oder datendidaktatur. In C. Könneker (Ed.), *Unsere digitale zukunft* (pp. 3–21). Springer. https://doi.org/10.1007/978-3-662-53836-4_1

Lee, H. S., & Tran, D. (2015). Statistical habits of mind. In *Teaching statistics through data investigations MOOC-Ed*. Friday Institute for Educational Innovation. https://fi-courses.s3.amazonaws.com/tsdi/unit_2/Essentials/Habitsofmind.pdf

Messy Data Coalition. (2020). *Catalyzing K–12 data education: A coalition statement*. https://messydata.org/statement.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality & threatens democracy*. Crown.

ProCivicStat Partners. (2018). *Engaging civic statistics: A call for action and recommendations.* http://iase-web.org/islp/pcs/documents/ProCivicStat_Report.pdf?1545118071

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3), 528–549. https://doi.org/10.1111/insr.12110

Ridgway, J. (Ed.). (in press). *Statistics for empowerment and social engagement: Teaching civic statistics to develop informed citizens*. Springer.

Van Es, K. & Schäfer, M. T. (Eds.). (2017). *The datafied society: Studying culture through data*. Amsterdam University Press.

Wolfram, C. (2020). *The math(s) fix: An education blueprint for the AI age.* Wolfram Media.