Sociedade Brasileira de Educação Matemática

RIPEM

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

Joachim Engel
Ludwigsburg University of Education
Alemanha
✉ engel@ph-ludwigsburg.de
0000-0002-8251-6670

Laura Martignon
Ludwigsburg University of Education
Alemanha
✉ martignon@ph-ludwigsburg.de
0000-0002-0318-2327

2238-0345 ISSN

10.37001/ripem.v14i3.3816 doi

Received • 01/03/2024
Approved • 15/05/2024
Published • 20/08/2024

Editor • Gilberto Januario

Abstract: Data science as a practical science has been conceived to address tangible problems in science, technology and society. Educating students in data science goes beyond teaching about algorithms, skills of manipulating data sets, selecting and applying appropriate analyses, and creating and interpreting visual representations of data. It also involves raising a critical understanding of how data are produced 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. 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 statistical and media literacies. Based on an analysis of trends and needs to protect democratic values in a datafied society and on own reflections of teaching practices this paper gives recommendations on designing data science courses to develop informed citizen.

Keywords: Statistical Literacy. Data Science Education. Data Ethics. Citizenship.

Ciencia de datos para ciudadanos informados: Aprendizaje en la intersección de la alfabetización informática, la estadística y la justicia social

Resumen: La ciencia de datos como ciencia práctica se ha concebido para abordar problemas tangibles de la ciencia, la tecnología y la sociedad. Educar a los estudiantes en la ciencia de datos va más allá de enseñar algoritmos, habilidades para manipular conjuntos de datos, seleccionar y aplicar análisis adecuados y crear e interpretar representaciones visuales de los datos. También implica suscitar una comprensión crítica de cómo se producen los datos y cómo pueden utilizarse para fines concretos, incluido el papel del contexto en la interpretación de los datos. Hace hincapié en el desarrollo de una conciencia ética de los datos y en la consideración de las implicaciones para la política y la sociedad del uso de potentes algoritmos. La participación en la democracia, en la sociedad digital y de datos de hoy en día, requiere el desarrollo de una serie de competencias transversales que deben fomentarse en las instituciones educativas a través de pedagogías de orientación crítica que entrelazan las habilidades y prácticas técnicas de datos junto con la alfabetización estadística y mediática**.** Basándose en un análisis de las tendencias y necesidades para proteger los valores democráticos en una sociedad informatizada y en reflexiones proprias sobre las prácticas docentes, este documento ofrece recomendaciones sobre el diseño de cursos de ciencia de datos para formar ciudadanos informados.

Palabras clave: Alfabetización estadística. Educación en ciencia de datos. Ética de los datos. Ciudadanía.

Ciência de dados para cidadãos informados: Aprendizagem na interseção de alfabetização em dados, estatística e justiça social

Resumo: A ciência de dados, como uma ciência prática, foi concebida para tratar de problemas tangíveis na ciência, na tecnologia e na sociedade. Educar os alunos em ciência de dados vai além de ensinar sobre algoritmos, habilidades de manipulação de conjuntos de dados, seleção e aplicação de análises adequadas e criação e interpretação de representações visuais de dados. Envolve também o aumento de uma compreensão crítica de como os dados são produzidos e como podem ser usados para fins específicos, incluindo a função do contexto na interpretação dos dados. Enfatiza o desenvolvimento de uma consciência para a ética dos dados e a consideração das implicações para a política e a sociedade quando algoritmos poderosos são usados. A participação na democracia, na sociedade digital e de dados de hoje, exige o desenvolvimento de uma série de habilidades transversais que precisam ser fomentadas nas instituições educacionais por meio de pedagogias criticamente orientadas que entrelaçam habilidades e práticas de dados técnicos com alfabetização estatística e midiática. Com base em uma análise das tendências e necessidades de proteger os valores democráticos em uma sociedade com dados e nas próprias reflexões sobre as práticas de ensino, este artigo apresenta recomendações sobre a criação de cursos de ciência de dados para desenvolver cidadãos informados.

Palavras-chave: Alfabetização estatística. Educação em ciência de dados. Ética de dados. Cidadania.

1 Introduction

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.

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). The ability to assess the credibility of information and its sources has never been more important. The World Risk Report[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

[1] https://www.weforum.org/publications/global-risks-report-2024/

decisions which likely disproportionately impact those with less privilege and resources at their disposal (O Neill, 2016). Amazon's model for sorting job applications[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'Neill, 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.

Based on literature review, an analysis of needs to strengthen democratic values in the digital age and the reflection of own teaching practices, this paper aims to provide guidance on how to design data science education for informed citizens. In the following, we outline in Section 2 the need for data literacy to be part of general education at any educational level and its challenges for the emerging the new field of data science education (Section 3) before we explain in Section 4 our concept of implementing some elements of data science and present our objectives in classes for students preparing to be secondary school teachers. For this group 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. Finally, in Section 5 we summarize the need for a comprehensive approach that emphasizes reflections on the societal impact of data science applications.

2 Data Literacy – Challenges for the 21st Century Educators

Data science as a practical science has been conceived to address tangible problems in science, technology and society. Educating students in data science goes beyond teaching about algorithms, skills of manipulating data sets, selecting and applying appropriate analyses, and creating and interpreting visual representations of data. It also involves raising a critical understanding of how data are produced 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. *Therefore, using real data is not enough, we need to teach data science to addresses real problems!* Algorithms are not the goal of data science, they are an important tool. As a

[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

comparison: Physics is not about calculus but about understanding natural phenomena. Calculus is a tool for physics. Teaching data science for informed citizen and, e.g., future mathematics school teachers requires different contents than a course on data science for computer science, data engineers or statistics majors.

The empowered citizen in the age of digitization needs skills to navigate an overabundance of data in order to make informed decisions - in everyday life as well as at various societal and political levels. "*Data literacy encompasses the data skills that are important for all people in a world shaped by digitization. It is an indispensable part of general education*," according to the Data Literacy Charter 2021 initiated by the Stifterverband – an initiative by companies and foundations in Germany devoted to promote improvements in education, science and innovation – and endorsed by numerous associations and individuals (Schüller et al., 2021). At the heart of data literacy are skills to collect, manage, evaluate, and acquire new knowledge from data in a critical way. This explicitly includes skills to critically evaluate data and its impact on social and political interaction.

In order to articulate knowledge, support positions with evidence, test assumptions or assess probabilities in situations of uncertainty and risk, the ability to explore and understand data is essential. For people to obtain information and evaluate its quality to form informed and independent opinions, media literacy is also required. Precisely because people are repeatedly called upon in their private and public lives to make decisions that go beyond their actual sphere of knowledge and experience, it is important that they learn to obtain trustworthy information, ask questions to better understand facts, and make fact-based decisions (Lengnink et al., 2013). Digital literacy is a central goal of preparing young people for the digital age (OECD, 2019). Data literacy includes the ability to collect, manage, evaluate, and critically apply data. It is key to systematically transforming data into knowledge. The Data Literacy Charter, emphasizes that data literacy strengthens judgment, self-determination, and a sense of responsibility and promotes social and economic participation for all people in a world shaped by digitization (Schüller et al., 2021). Data literacy serves to promote maturity in a modern digitized world and is therefore important for all people - not just specialists. The goal of data literacy education is for every individual and our society as a whole to deal with data in a conscious and ethically sound manner. Data literacy enables successful and sustainable action that is based on evidence and takes appropriate account of uncertainty and change in our living environment.

Data literacy needs to be part of the general educational mission of schools in the 21st century and concerns mathematics education in a special way, in addition to computer science and social science education (Wolfram, 2020; Messy Data Coalition, 2020).

In order for learners to "appreciate and critically analyze the social embeddedness and constructedness of data" (Richterich, 2018), it is necessary to work with real data that demonstrates concrete authentic problems using openly accessible data sources (Engel, 2017; Ridgway 2015). Open data can be analyzed by students, empowering them to formulate, ask and investigate socially burning questions and thereby exercise their rights as citizens (Ridgway, 2022).

Finally, addressing the challenges of a data-driven society is not just a matter of curriculum design or individual teacher responsibility, but requires institutional strategies and policies to support teachers in developing data literacy. This requires a coherent plan for systemic change. In some contexts, this might begin by integrating elements of data literacy and statistics into otherwise traditional courses. In other contexts, it might be appropriate to simply use authentic, large-scale data sets relevant to social problems to teach traditional topics. In other contexts, radical curriculum reform may be required (ProCivicStat Partner, 2018).

3 The Emergence of Data Science Education

While we live in an era of data inundation, statistics education in secondary education and beyond is very much focused on a 20th century paradigm developed during a time when data came from planned studies, software was expensive, and the purpose of a statistic study was to calculate a p-value or find a confidence interval. Today, data are universally present and high-quality software is inexpensive or even free. Many barriers that prevented secondary students from analyzing data in the past no longer exist.

Data science education must look beyond a combination of numerical, statistical and technical capabilities, include critical thinking, citizenship and foster skills to evaluate, analyze and interpret data. Educational programs must look beyond data capabilities, and include critical thinking, foster skills to evaluate, analyze and interpret data and their meaning for policy and society. Data literacy needs to be part of the general educational mission of schools in the 21st century and concerns mathematics education in a special way, in addition to computer science and social science education

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. Data science education must look beyond a combination of numerical, statistical and technical capabilities, include critical thinking, citizenship and foster skills to evaluate, analyze and interpret data. Educational programs must look beyond data capabilities, and include critical thinking (Van Es and Schäfer, 2017), foster skills to evaluate, analyze and interpret data and their meaning for policy and society (ProCivicStat Partners, 2018; Schield, 2004). 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).

By now, there are a number of concepts, proposals, and experiences to introduce elements of Data Science into the classroom at middle and high school levels. The International Data Science in Schools Project (IDSSP)[3], an international collaboration of statisticians and computer scientists and educators, developed curricula to introduce Data Science to students in their last two years of high school, as well as a curriculum to empower teachers on how to teach Data Science. Other recent innovative initiatives are, e.g., ProDaBi[4] (Biehler & Fleischer, 2021) and Mobilize IDS[5] (Gould, 2021).

Classroom approaches to Data Science can often be linked in terms of content to addressing societal issues that are on the minds of many, such as climate change, pandemic events, 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 the 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 or Our World of Data 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.

[3] http://www.idssp.org
[4] https://www.prodabi.de/en/
[5] https://www.introdatascience.org

4 Data Science for Informed Citizens: Overarching Objectives

In this section, we discuss a personal account of what we believe every educated and informed 21st century citizen should know about machine learning and automated decision making. Our assertions are debatable, and you may find that some objectives are missing, may emphasize others more than we do, or disagree to a greater extent. However, the following contents have been implemented and evaluated in a series of classes for prospective secondary school teachers in mathematics and social studies. No specific mathematical or technological background beyond high school is assumed, so the concept and goals may well apply to any group of educated and informed citizens. Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data with techniques from statistics, data analysis and machine learning (ML). Machine Learning is used in Data Science to make predictions or to classify data. It includes algorithms like decision trees, random forests, neural networks, and many others that automatically identify patterns and make decisions based on data.

4.1 Raise awareness where in daily life we encounter Data Science products

Data Science products and especially machine learning algorithms have become an integral part of our daily lives, often in ways many of us might not immediately recognize. Here some examples at some domains of daily life where machine learning products are encountered

- Smartphones and devices: Our smartphones are a hub of ML applications. From voice assistants like Siri, Google Assistant, and Alexa, to predictive text, language translation and autocorrect features in keyboards, and facial recognition for unlocking devices—all are powered by machine learning algorithms
- E-Commerce and Online-Shopping: Websites like Amazon and eBay use machine learning to recommend products. These systems analyze your past shopping behavior, search history, and what other similar users have bought or viewed
- Fraud-Detection: Machine learning algorithms help detect fraudulent transactions by identifying patterns that may indicate fraudulent activity
- Banking and Finance: Machine Learning models help in assessing a person's creditworthiness by analyzing vast amounts of financial data and patterns.
- Education: Educational apps and platforms use machine learning to adapt content to the learning pace and style of each student, providing personalized exercises and feedback.

These are just a few examples illustrating how machine learning technologies enhance and facilitate many aspects of our daily routines, often making services more efficient, personalized, and user-friendly. We are sure your students come up with more examples. Are they also aware where their personal records are tracked and stored and how their data are being used?

4.2 Teach awareness about data quality and data suitability

Data—the empirical basis for evidence-informed decisions and knowledge creation—are certainly preferable to anecdotes, wishful thinking, superstition, prejudice, or ideology. Yet data themselves are neither facts nor truth. Some authors consider data as models of reality. Data do not provide objective representations of the world. They might arise opportunistically, or as a result of conscious decisions someone made to research a particular topic. Data usually have been collected at costly expense, for a particular purpose and with a specifically chosen research design. They measure manifest variables in a particular way. They are the basis for

constructing latent variables based on some kind of model with a specific concept in mind. At a more complex level, one can ask why particular measures have been chosen, by whom, and for what purposes. Measurement is always linked to some theory of the phenomenon being studied. An economist may use a different conceptualization of the term poverty in his work than a sociologist. In the example of the natural sciences, mass, length and time were chosen as measures not because they are "obvious", but because their measurement allows precise predictions about the physical world to be made and used. The well-being of nations was measured by gross domestic product per person, or GDP for short, until this monolithic measure was questioned by Amartya Sen and replaced by the more comprehensive Human Development Index HDI.

Collecting data is not a leisure activity but is laborious, sometimes tedious work that usually requires a lot of effort and financial resources. It serves someone's interest, and it is legitimate to question whose interest this is. Why have these data been collected? The data collected implicitly tell a story. Whose story is this? And whose story is this not?

Critical or reflective questions about the methods used in surveys might include (but are not limited to):

- Are the measures (e.g., a questionnaire) well defined? Are the measures robust and appropriate for the purposes for which they are being used?
- Are metadata (i.e., detailed explanations of how variables were defined, sample characteristics etc.) available?
- Were the sampling procedures appropriate? Who is missing from the collected data? (e.g., measuring how citizens feel about a certain topic by analyzing social media streams fails to sample non-users).

Many studies in the social sciences are concerned with theories of causality; causality is associated with difficult philosophical challenges that go well beyond simple mantras such as "correlation does not imply causation." However, when data come from observational studies, surveys, or archive data, and not from experimental studies, a reliable identification of cause-and-effect relationships can be difficult to determine.

Beyond technical knowledge about processes of data generation, it is important that individuals are able to ask critical questions to assess the credibility and validity of any data, finding, or conclusion they encounter, both on technical and logical grounds—even data or reports from presumably credible sources such as official statistics agencies. It is important to examine, from a critical perspective, narratives and interpretations of data, and the conclusions drawn from them, for example:

- What is the quality of the evidence presented in a media article or a claim to support assertions about needed policy or actions (e.g., regarding recycling laws, wage equality, or vaccination)?
- How reasonable are the projections and how appropriate are the underlying statistical models and assumptions that have been applied to analyze data on key issues (e.g., on the progression of global warming or the rate of spread of infections such as the COVID-19 coronavirus pandemic)?
- When assertions are made about a correlation between variables (e.g., smoking and risk of death), are relationships assumed to be linear, and are they really so (or perhaps curvilinear)? More important, if causal processes or cause-and-effect relationships are assumed, are there plausible rival accounts, covariates, or unexplored intervening factors

which could affect the findings? There are often several equally valid ways of describing a social phenomenon.

- Are the conclusions consistent with other available evidence? When proposals are made for social policy, one can ask if the problem identification has been done adequately and whether relevant data have been used.

Osborne and Pimentel (2019) provide a heuristic using a fast-and-frugal decision tree that even non-experts ("competent outsiders") can use to assess the credibility of science-based arguments. They suggest three simple questions:

1. Is the source of this information credible? Evidence for credibility is given by: No conflicts of interest, source is acknowledged, the analysis of the topic is unbiased
2. Does the source have the expertise to vouch for the claim? Evidence for expertise and experience is provided through the track record, reputation among peers, credentials and the institutional context and through relevant professional experience
3. Is there a consensus among the relevant scientific experts?

If the answer to all three question is Yes, then accept the claim according to their heuristic. If the answer only to question 3 is NO while 1) and 2) are answered positively, then inquire about the nature of the disagreement. What do most highly regarded experts think? What range of findings are deemed plausible? And what are the risks of being wrong?

Everyone should adopt a questioning attitude and know what questions to ask about the nature, limitations or credibility of different data sources, statistical messages and conclusions. However, taking a critical stance when evaluating evidence does not mean being critical at all costs. Rather, criticism is about adopting the attitude of a fair-minded skeptic who is willing to accept a presentation but needs to be convinced by evidence. In situations where data is presented in a misleading way, students should be encouraged to re-present it in a more appropriate way; in situations where the data is dubious (or falsified), they should be encouraged to find relevant data from authoritative sources.

4.3 Teach awareness about biases of machine decisions

Machine learning algorithms can exhibit a lot of biases due to a variety of reasons, often reflecting issues in the data they're trained on, the design of the algorithm itself, or the broader societal and historical contexts in which they are developed and deployed. Here are several key factors contributing to bias in machine learning.

- Biased training data: Any bias in the training set will be amplified in the test set, leading to biased decisions. If the data used to train a machine learning model contains biases—either through underrepresentation or overrepresentation of certain groups, or through historical biases present in the data—the model will likely learn and perpetuate these biases.
- In October 2019, researchers found that an algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care heavily favored white patients over black patients. While race itself wasn't a variable used in this algorithm, another variable highly correlated to race was, which was healthcare cost history. The rationale was that cost summarizes how many healthcare needs a particular person has. For various reasons, black patients incurred lower health-care costs than white patients with the same conditions on average.

- Algorithmic bias: The design of the algorithm itself can introduce bias. Some algorithms might be more prone to amplifying biases present in the training data. For example, algorithms that heavily penalize outliers might not perform well for minority groups that are underrepresented in the training data.
- To predict academic performance of school students national qualification regulators in the UK set up formulas to assign predicted examination grades based on teacher prediction - the algorithm assigned poorer grades to students in state-funded schools and better grades (even better than teacher prediction) to students in smaller independent schools (Smith, 2020).
- Historical and societal context: The societal, historical, and cultural contexts in which data is generated often contain biases. Since machine learning models learn from past data, they can inadvertently learn and perpetuate these societal biases.
- Facial recognition tools use ML algorithms are being used for law enforcement in Brazil, causing a heated debate on whether these algorithms carry racist bias and propagate existing inequalities[6] (Silva, 2022).

Arguably the most notable example of a ML bias is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm used in US court systems to predict the likelihood that a defendant may become a recidivist. Due to the data that was used, the model that was chosen, and the process of creating the algorithm overall, the model predicted twice as many false positives for recidivism for black offenders (45%) than white offenders (23%) (Shin, 2020).

Addressing bias in machine learning is a great challenge that requires careful attention to the entire lifecycle of model development, from data collection and preparation through to deployment and monitoring. Strategies to mitigate bias include using more diverse and representative datasets, applying de-biasing techniques during model training, and continuously monitoring and updating models to ensure fair and equitable outcomes. The data that one uses needs to represent "what should be" and not "what is". Otherwise, as in Amazon's hiring algorithm, the risk is high of underrepresenting and causes discrimination against. a particular group of people. The validity of the algorithms needs to be evaluated when applied to various social groups. Furthermore, implementation needs some sort of mandated and enforced data governance to ensure a practice that is ethical with respect to the values of a free society (Shin, 2020).

4.4 Teach awareness about the impact of Data Science products on society

Creating critical awareness of the impact of machine learning on society is crucial to educating informed citizens who can contribute to the ethical development, deployment, and governance of these technologies.

In schools we can integrate discussions about the societal impacts of technology, including machine learning into the curriculum at various educational levels. This could range from simple lessons on digital literacy in elementary schools to more complex debates on ethics in high school and university courses. The ethics in machine learning encompasses legal, political social and economic dimensions. Therefore, an interdisciplinary approach that emphasizes cooperation with other fields such as ethics, philosophy, sociology or political science is appropriate. This approach ensures students not only learn how machine learning

[6] https://brazilian.report/podcast/2024/01/17/algorithms-ai-facial-recognition-racist/

algorithms work but also understand their broader societal implications, including economic, social, and ethical dimensions

An effective way is to engage students in projects where they have an opportunity to design, implement, or critique machine learning systems with ethical considerations in mind. This could include tasks such as creating a machine learning model taking into account potential biases, or developing guidelines for ethically deploying machine learning systems.

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. While only few people might care how a machine translates a document from one language to another (as this paper in a first approach was translated from English to Portuguese by DeepL[7]) as long as the translation is accurate, decisions involving personal human rights should never be left to the machine.

4.5 Teach some technological basics about machine learning

While algorithms are the tool and not the goal of Data Science, to appreciate the specific nature of machine learning, even students not majoring in computer science need to learn some of the technology about how a machine can do something we call "learning". A good introduction are automated decision rules, represented by classification trees (Breiman et al., 1984). Trees are intuitive, simple to apply, easy to understand and give easily interpretable results. Decision trees created algorithmically from training data are simple and yet powerful tools capable of achieving high accuracy in many tasks while being highly interpretable. The "knowledge" learned by a decision tree appears as a hierarchical structure, a blueprint for decisions. This structure holds and displays the knowledge in such a way that even non-experts can immediately apply it.

Tree-based algorithms are an important method of machine learning which supports decision making, e.g., in medicine, finance, public policy and many more. Trees open doors to more advanced topics of Data Science and machine learning (e.g., Random Forests, Bagging and Boosting, as well as fundamental concepts such as training sets and overfitting). However, instead of beginning with a computer algorithm that produces optimal trees, we suggest that students first construct their own trees, one node at a time, to explore how they work, and how well. This build it-yourself process is more transparent than using algorithms such as CART (as implemented, e.g., in the package `part` or `tree` in R or `sklearn` in Python). We believe it will help students not only understand the fundamentals of trees, but also better understand tree-building algorithms when they do encounter them.

This may start completely unplugged. In a hands-on activity students receive data on cards. Each card has height, weight and various other measurements, including some irrelevancies (eye color) for a professional handball or football player. The students work to figure out how to predict the sport based on the other attributes. They came to see that they could cast their algorithm as a classification tree. Finally, they got previously-unseen cards to test their tree; this led, among other things, to discovering overfitting: a phenomenon where using the irrelevant variables resulted in an excellent, perhaps even perfect tree for the training data that was worse with new, test data.

[7] https://www.deepl.com/translator

In a next step we use the Common Online Data Analysis Platform (CODAP) (Finzer, 2019) and its plug-in ARBOR. CODAP is a free, open-source, web-based package. It is especially designed for learning introductory data analysis; students primarily use selection, on-screen controls, and dragging to accomplish their tasks rather than writing code and executing it.

Instead of relying on algorithms, ARBOR requires the user to make successive choices about which variable to split, how to split, and when to stop growing the tree Erickson and Engel, 2023). By using ARBOR, students can come to understand what the algorithms accomplish, and perhaps even more to the point, come to understand the nature of trees themselves. ARBOR has no automated algorithms that compute optimal splits and right-sized trees. The user, through drag and drop moves, decides step by step which variables to use for consecutive splits, how to specify the split, and when to terminate tree growth. Misclassification rates evaluating the goodness of the chosen split are immediately reported, which allows comparison with alternative splits. The purpose of ARBOR is not the derivation of an optimal tree, but to let the user explore the flexibility of the tree method and the consequences of various splits, thus to gain appreciation for trees as an automatized method of learning from data.

Only after some activities with “hand-grown” trees students turn to powerful algorithms and packages for classification trees as implemented in R or Python. There, with some code provided by the instructor, students let the algorithm construct trees applied to some real-world problems that are optimal according to some criteria. Unlike some more powerful machine learning methods classification trees are transparent but are behind their modern competitors with regard to efficiency and accuracy. Yet, from (simple) classification trees it is only a small step to the concept of Random Forests. Along concrete problems and with appropriate R code provided, the students make the experience that advanced methods like Random Forests beat the simple classification trees with respect to some outside criteria like misclassification rate, but for the prize of a completely opaque decision rule.

5 Conclusion

5.1 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.

5.2 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.

5.3 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.

5.4 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.

5.5 Problem solving

We also need to teach technical skills such as programming, machine learning algorithms and 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 impact.

Students should learn early in their education that it 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 ads instead of using 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

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. Teaching Statistics, 43(S1), S133-S142. https://doi.org/https://doi.org/10.1111/test.12279

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

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.

Finzer, W. (2019). Common Online Data Analysis Platform (CODAP). Concord: The Concord

Consortium.

Gould, R. (2021). Towards data-scientific thinking. *Teaching Statistics*. 43, 11-22.

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.* https://doi.org/10.1007/978-3-662-53836-4_1

Lengnink, K., Meyerhöfer, W. & Vohns, A. (2013). Mathematische Bildung als staatsbürgerliche Erziehung? *Der Mathematikunterricht* 59 (4), 2-7.

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

OECD (2019). *OECD Skills Outlook: Thriving in a Digital World*. OECD Publishing, Paris. https://doi.org/10.1787/df80bc12-en

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

Osborne, J. & Pimentel, D. (2019). Science, misinformation, and the role of education. *Science*, Vol 378, Issue 6617, 246-248. https://www.science.org/doi/10.1126/science.abq8093

ProCivicStatPartners (2018). Engaging civic statistics: a call for action and recommendations. A product of the procivicstat project. http://iase-web.org/islp/pcs

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

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

Ridgway, J. (2022, Ed.). *Statistics for empowerment and social engagement: teaching Civic Statistics to develop informed citizens*. Springer.

Schield, M. (2004). Information Literacy, Statistical Literacy and Data Literacy. IASSIST Quarterly 28(2), 7-14. https://doi.org/10.29173/iq790

Schüller, K., Koch, H. & Rampelt, F. (2021). Data-Literacy Charta. https://www.stifterverband.org/data-literacy-charter

Shin, T. (2020). Real-life Examples of Discriminating Artificial Intelligence. *Towards Data Science* https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Silva, T. (2022). *Racismo Algorítmico: inteligência artificial e discriminação nas redes digitais.* Sesc Edições SP

Smith, H. (2020). Algorithmic bias: should students pay the price? *AI & SOCIETY*, 35(4), 1077–1078. https://doi.org/10.1007/s00146-020-01054-3

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

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