International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-021-00241-7

ARTICLE

Teaching Responsible Data Science: Charting New Pedagogical Territory

Armanda Lewis [1] · Julia Stoyanovich [1]

Accepted: 10 February 2021/Published online: 15 April 2021

Abstract

Although an increasing number of ethical data science and AI courses is available, with many focusing specifically on technology and computer ethics, pedagogical approaches employed in these courses rely exclusively on texts rather than on algorithmic development or data analysis. In this paper we recount a recent experience in developing and teaching a technical course focused on responsible data science, which tackles the issues of ethics in AI, legal compliance, data quality, algorithmic fairness and diversity, transparency of data and algorithms, privacy, and data protection. Interpretability of machine-assisted decision-making is an important component of responsible data science that gives a good lens through which to see other responsible data science topics, including privacy and fairness. We provide emerging pedagogical best practices for teaching technical data science and AI courses that focus on interpretability, and tie responsible data science to current learning science and learning analytics research. We focus on a novel methodological notion of the *object-to-interpret-with*, a representation that helps students target metacognition involving interpretation and representation. In the context of interpreting machine learning models, we highlight the suitability of "nutritional labels"—a family of interpretability tools that are gaining popularity in responsible data science research and practice.

Keywords Responsible data science · Model interpretability · Constructivism · Data science pedagogy

✉ Armanda Lewis
al861@nyu.edu

Julia Stoyanovich
stoyanovich@nyu.edu

[1] New York University, New York, NY, USA

Introduction

In the past five years, scholars and practitioners have identified how modern data science and AI techniques may violate fairness, accountability, interpretability, and transparency from legal, socio-cultural, and technical perspectives (Angwin et al., 2016; Barocas & Selbst, 2016; O'Neil, 2016). This led to the development of cross-disciplinary research in responsible data science, fairness in machine learning and explainable AI, and to the establishment of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT, formerly ACM FAT*). However, little attention has been devoted to training future data scientists to integrate responsibility into the systems they design, build, and deploy.

In this paper we recount a recent experience in developing and teaching a technical course focused on *responsible data science (RDS)*, which tackles the issues of ethics in AI, legal compliance, data quality, algorithmic fairness and diversity, transparency of data and algorithms, privacy, and data protection. We describe emerging pedagogical best practices for teaching technical data science and AI courses, and tie RDS to current learning science and learning analytics research. We focus on the novel methodological notion of the *object-to-interpret-with*, a representation that helps students target meta-cognition involving interpretation and representation. Further, we highlight the suitability of "nutritional labels"—a family of interpretability tools for data and models—for responsible data science instruction.

Our paper makes the following contributions:

- We explicitly connect theories and advances within the learning sciences to the teaching of responsible data science, specifically, interpretability. We are among the first to consider the pedagogical implications of responsible data science, creating parallels between cutting-edge data science research, and cutting-edge educational research within the fields of learning sciences, artificial intelligence in education, and learning analytics & knowledge.
- We offer a description of a unique course on responsible data science that is geared toward technical students, and incorporates topics from social science, ethics and law.
- We propose promising pedagogical techniques for teaching interpretability of data and models, positioning interpretability as a central integrative component of RDS.

With this work we aim to contribute to the nascent area of RDS education. We hope to inspire others in the community to come together to form a deeper theoretical understanding of the pedagogical needs of RDS, to integrate relevant educational research, and to develop and share the much-needed concrete educational materials and methodologies, striking the right balance between research and practice.

Defining Interpretability within Data Science Education

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is *interpretability* — allowing humans to understand, trust and, if necessary, contest the computational process and its outcomes. Interpretability is

central to the critical study of the underlying computational elements of machine learning platforms, and can allow for a host of addressable questions about the granular and holistic aspects of understanding a model (Diakopoulos, 2016; Goodfellow et al., 2015; Lipton, 2016; Ribeiro et al., 2016).

A recent study examined data scientists' internal evaluation of interpretability tools, finding that selected participants tended to take a dataset or machine learning model at face value when visualizations were present and not be aware of interpretability issues that may affect robustness, replicability, or fairness (Kaur et al., 2020). Scholars from humanistic, social science, and scientific backgrounds voice the importance of introducing human-centered and critical approaches when thinking about interpretability (Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Leonelli, 2016; Miller, 2019; Stoyanovich, Bavel, & West, 2020a; Stoyanovich, Howe, & Jagadish, 2020b). We take interpretability to generally mean a person's or group's ability to understand a model – for example, to describe what the inputs and algorithms are, how the algorithms operate, how outputs are framed, and even how to articulate the model's reuse over time and contexts.

Specifying the intended audience and purpose of interpretability is essential, as are articulating the desired balance between model performance and model interpretability, and highlighting underlying opacity (Breiman, 2001; Epstein et al., 2018; Gleicher, 2016). The most substantive publications around interpretability are geared towards technical audiences and offer varying levels of accessibility for non-technical groups (Ananny & Crawford, 2016; Dwork, 2011; Guidotti et al., 2019; Kim et al., 2015). For simplicity, and based on our teaching experience, we focus on a population whose technical sophistication and interest in real-world problem solving lies in the middle — students of data science. Targeting this audience connects to established ways in which people learn about a complex topic — what they are able to find interpretable — while pinpointing generalizable aspects that hold potential for the wider public. Data science and AI education are still maturing since most branded academic programs have sprouted up over the past 5 years or so, though there do exist recent curricular and professional guidelines (American Statistical Association, 2016; Association for Computing Machinery, 2018; De Veaux et al., 2017) that emphasize theoretical foundations, applied mathematical and computational knowledge, and responsibility of knowledge transference, as well as existing research that recognizes interdisciplinary knowledge and ethical training (Anderson et al., 2014; Doore et al., 2020; Mir et al., 2019, p. 22). Achieving large-scale interpretability within educational contexts is challenging, and we posit that advances in associated fields like AI education, the learning sciences, and learning analytics are vital to advance RDS education.

One way that we have addressed the challenge of teaching and learning about interpretability is through the notion of students working with what we call *objects-to-interpret-with*, which takes its inspiration from an important concept that emerges through constructivist practices — objects-to- think-with (Papert, 1980). These objects are representations that help learners grapple with universal concepts and "understand how ideas get formed and transformed when expressed through different media, when actualized in particular contexts, when worked out by individual minds" (Ackermann, 2001). Our objects-to-interpret-with likewise assist learners in forming heuristic knowledge and understanding contextual knowledge, but specifically target metacognition surrounding ways of interpretation and representation.

Integrating Ethical Considerations into Technical Disciplines

There is now a representative group of technical data science degree programs at the graduate level, and growing offerings at the undergraduate level (Aasheim et al., 2015; Anderson et al., 2014; Farahi & Stroud, 2018). Programs are diverse in their approach and may focus on computational statistics, machine learning or, specifically, deep learning, software development, and/or visualization theory and techniques. As an emerging discipline, RDS has yet to be codified as a course of study at most university campuses. Despite increasing demand due to public instances of irresponsible uses of data science, there is a lack of curricular materials available for easy adoption into university programs. The challenge is compounded by a shortage of faculty with the expertise to develop and teach RDS courses that require multidisciplinary integration.

Institutions such as Duke, Harvard, Oxford, Stanford, University of Michigan, University of Texas, and others have introduced ethical data science courses, all of which approach the topic from humanistic or social science- based — cultural, legal, and/or philosophical — points of view. What is less represented, and what we detail here, is offering ethical data science training from the point of view of the technical.

Learning about ethics requires normative knowledge distinct from scientific inquiry, and necessitates exposure to other epistemological frameworks (i.e., structured ways of knowing) (Kabasenche, 2014; Reiss, 1999). The interdisciplinary nature of data science necessitates a new kind of pedagogy, one that not only requires robust technical, theoretical, and practical STEM- based training, but also incorporates best learning paradigms for inquiry and interpretation, and a humanist sensibility that highlights ethical concerns and communication challenges.

Essential Lessons from the Learning Sciences

These best learning paradigms come from the learning sciences, an interdisciplinary field that develops theory, design practices, and policy recommendations to advance knowledge of how people learn in a range of contexts, and provides essential frameworks for teaching RDS (Fischer et al., 2018; Kolodner, 1991; Sawyer, 2005). While research in the learning sciences may initiate out of field-specific advances in anthropology, cognitive science, computer science, cultural studies, education, psychology, or statistics, it has a common focus on supporting specific learning processes that are active, collaborative, socially constructed, and situated. We situate our RDS course within a learning science framework since it offers technical training, as well as context-specific ethical learning opportunities. As such, students learn how to achieve technical outcomes including privacy-preserving data release (Ping et al., 2017), data profiling to identify potential sources of bias (Abedjan et al., 2017; Grafberger et al., 2021; Schelter & Stoyanovich, 2020), and balancing model accuracy with meeting statistical parity objectives (Schelter et al., 2020). They also grapple with issues that have no definitive outcomes, such as how mathematical bias tradeoffs may be interpreted, how various ways of defining privacy can conflict with the algorithmic need to operationalize variables, how algorithmic outcomes may contribute to unintended social impacts, and how one may make black box models more transparent. For these concepts, exposure to interdisciplinary methods and frameworks is important.

The learning sciences offers strategies for managing this need for interdisciplinarity within a technical learning focus.

Constructivism is on display when students learn computer science concepts through hands-on programming and visualization, connecting theory through applied activities and ongoing inquiry (Barnes et al., 2017; Ben-Ari, 2001; Hundhausen & Douglas, 2000). Evidence from experimental design trials supports positive learning outcomes, but only if the interventions maximize learner engagement (i.e., they allow for interactivity like algorithm building) (Hundhausen et al., 2002; Naps et al., 2002).

Wilkerson and Polman (Wilkerson & Polman, 2020) introduce recent data science education research grounded in the learning sciences, given that "the learning sciences is uniquely positioned to investigate how such contextual embeddings impact learners' engagement with data including along conceptual, experiential, social, racialized, spatial, and political dimensions" (Wilkerson & Polman, 2020, p. 2). One result is on offering a holistic view of the data analysis process, compelling students to think collaboratively about the creation of data, and treating it with an equal level of importance as statistical significance of results (Hilliam & Calvert, 2019; Wise, 2020). We share the point of view that responsibility must be built into all stages of the data science lifecycle, starting with design, followed by data collection, integration and cleaning, through data analysis and interpretation of results (Stoyanovich et al., 2017; Stoyanovich, Howe, & Jagadish, 2020b), and we bring this point of view into RDS instruction. Holistic understanding is also possible by incorporating theories of multimodal learning, such as using effectively-generated visualizations and interactive simulations (Garfield & Ben-Zvi, 2007; Nolan & Perrett, 2016; Rubin et al., 2006).

Lessons from the Field of Learning Analytics and Knowledge

It is also helpful to situate our RDS course within the field of learning analytics and knowledge (LAK), the study of the data and associated tools that describe, facilitate, and predict learning in context (Baker & Inventado, 2016; Siemens, 2012; Wise, 2014). LAK research affords important lessons for ethical considerations within technical contexts, as evidenced by the 2016 Journal of Learning Analytics special issue highlighting transparency, privacy, data integrity, and accessibility (Cormack, 2016; Khalil & Ebner, 2016; Sclater, 2016). LAK research has advanced technical understanding of fairness and privacy metrics, and of the use of participatory frameworks in the design and implementation of learning analytics systems (Gardner et al., 2019; Gursoy et al., 2017; Holstein & Doroudi, 2019). Recent research exposes how the algorithms of adaptive learning systems may not behave fairly across different populations of students, and offers considerations for the designers of such systems (Baker et al., 2008; Doroudi & Brunskill, 2017; Doroudi & Brunskill, 2019).

In addition to considering technical robustness, LAK research acknowledges the contextual nature of ethical considerations, and the need to incorporate critical theory and social science methods to bridge positivist and interpretivist viewpoints (Buckingham Shum, 2019; Hoel & Chen, 2016). Privacy, data ownership and protection, and equity are concerns that LAK scholars advance to highlight the ethical challenges associated with collecting, mining and interpreting learners' data (Ferguson, 2019).

Ethical Data Science and AI Education

Another field within which to explore ethical data science education is Artificial Intelligence in Education (AIED), which harnesses advances in machine learning and big data systems to support personalization, adaptation, multimodal learning, and the integration of sociocultural contexts into the learning process (DeFalco et al., 2018; Walker & Ogan, 2016; Yannier et al., 2020; Zimmerman, 2018). Pinkwart highlights that emerging AI technical advances will need to confront privacy and ethical issues as well. "...[T]he challenge of creating ubiquitous and universal interaction methods has direct implication for educational technologies, as do networks and policies that enable the reliable processing of educational data within AIED systems while respecting privacy" (Pinkwart, 2016). In addition to privacy, the ACM FAccT Conference, established in 2018, highlights other ethical topics such as fairness, accountability, and transparency.[1]

At the K-12 level, researchers are building curricula that leverage constructionism to develop students' AI understanding. Through ethical design challenges, interaction with social robots, and programming activities, students learned basic AI principles and demonstrated more creativity and ethical awareness, when compared with pre-test scores (Ali et al., 2019; DiPaola et al., 2020). Other research explores interpretability topics. Students engage in dialogic activities with a conversational agent, and learn the different ways that machine learning-enabled entities can represent knowledge, classify concepts, and incorporate social information. They conclude that participants did learn about machine learning concepts through iterative activities (Lin et al., 2020). In the context of our RDS course, we will describe how we advance students' interpretability capabilities by integrating ethical design, constructivist practices, and reflection.

At the professional level, there is established cross-disciplinary research on ethical reasoning (Boyd, 2010; Knapp, 2016; Leonelli, 2016), with specific recommendations for pedagogy and training that emphasize case studies, field practice, and sensemaking frameworks (Mumford et al., 2008; Sternberg, 2010). In their discussion on teaching ethical computing, Huff and Martin (Huff & Martin, 1995) summarize a framework for introducing ethical analysis that incorporates levels of social analysis (individual through global), responsibility, privacy, reliability, equity, and other related topics. In terms of pedagogy, they make an important recommendation of "incorporation of ethical and social issues in the lab work associated with such standard computer science subjects as database design, human-computer interaction, operating systems, and algorithms" (Huff & Martin, 1995). Tractenberg and colleagues (Tractenberg et al., 2015) offer guidelines on introducing ethical reasoning into data science and AI training, and detail two syllabi that have students reflecting on ethical misconduct, societal impacts, privacy and confidentiality considerations, and responsible research practices, though evaluation hinges on written assignments and class discussion only. In contrast to highly technical courses that introduce limited RDS topics (Friedman & Winograd, 1990; Harvard University, 2019; Martin & Holz, 1992; Quinn, 2006), the course we describe in Section 4 of this paper is both technical and provides broad coverage of RDS topics.

[1] https://facctconference.org/

Teaching Responsible Data Science

We now recount our experience in developing and teaching an RDS course to graduate and advanced undergraduate students at the Center for Data Science at New York University. The course is structured as a sequence of lectures, with supplementary readings, labs, and accompanying assignments.[2]

The course has Introduction to Data Science, or Introduction to Computer Science, or a similar course as its only a prerequisite. A machine learning course is not a prerequisite for RDS. This is a deliberate choice that reflects our goals to (1) educate data science students on ethics and responsibility early in their program of study, and (2) to enroll a diverse group of students. Students are expected to have basic familiarity with the python programming language, which is used in labs and assignments.

Course Overview

During this semester-long course, students complete the following six thematic modules using a combination of case studies, often from the recent press, fundamental algorithmic and statistical insights, and hands-on exercises using open-source datasets and software libraries.

- Module 1: Introduction and background (1 week)
- Module 2: The data science lifecycle, data profiling and cleaning (2 weeks)
- Module 3: Algorithmic fairness and diversity (3 weeks)
- Module 4: Privacy and data protection (2 weeks)
- Module 5: Transparency and interpretability (3 weeks)
- Module 6: Ethical, legal, and regulatory frameworks (2 weeks)

In selecting the set of topics to cover, and in structuring them as modules, we started with the technical topics that have been the focus of the Fairness, Accountability, Transparency, and Ethics community, as represented by ACM FAccT (formerly ACM FAT*), and papers on relevant topics at major AI conferences like AAAI, IJCAI, and NeuRIPS. These topics are algorithimic fairness, transparency, and interpretability. We expanded the treatment of algorithmic fairness beyond classification and risk assessment (as is currently typical in this line of research (Chouldechova & Roth, 2020)) to also include fairness in set selection and ranking (Stoyanovich et al., 2018; Yang & Stoyanovich, 2017), to consider intersectional effects (Yang et al., 2020), and, importantly, to make connections between algorithmic fairness and diversity (Drosou et al., 2017; Kleinberg & Raghavan, 2018; Yang et al., 2019).

We also included the more mature, but still very timely, discourse on privacy and data protection.

We make strong connections with responsible data management and data engineering, emphasizing the importance of the data lifecycle, and of the lifecycle of the design,

[2] [2] All course materials, including the syllabus, weekly reading assignments, complete lec- ture slides, and lab assignments, are publicly available on the course website at https://dataresponsibly.github.io/courses/. Homework assignments, with solutions and grading rubrics, and a detailed description of the course project, will be made available to instructors upon request

development, deployment, and use of algorithmic tools (Schelter & Stoyanovich, 2020; Stoyanovich, Howe, & Jagadish, 2020b). These topics are often overlooked in data science education in general, and have also received limited attention so far in responsible data science research. Yet, responsibility concerns, and important decision points, arise in data sharing, annotation, acquisition, curation, cleaning, and integration. Several lines of recent work argue that opportunities for improving data quality and representativeness, controlling for bias, and allowing humans to oversee the process, are missed if we do not consider these earlier lifecycle stages (Kirkpatrick, 2017; Lehr & Ohm, 2017; Stoyanovich et al., 2017). For this reason, data engineering topics are covered prominently in the course.

Legal and regulatory frameworks are also an important component of the course, making part of one of the modules. Further, current international and local regulatory efforts are used to ground the discussion throughout the course, starting from the first lecture. To help ground the course in current events in New York City, students are encouraged to attend public hearings of the New York City Committee on Technology, particularly those pertaining to regulation of Automated Decision Systems, and to reflect on these hearings during the discussion (Stoyanovich, Kuyan, et al., 2020c).

One of the challenges we faced when designing this course was the lack of a text book that offers comprehensive coverage of RDS, balancing case studies, fundamental concepts and methodologies from the social sciences, and statistical and algorithmic techniques. As a result, the course does not have a required textbook. Each topic is accompanied by required reading. In some cases, expert-level technical research papers are listed as assigned reading. However, important concepts from the assigned paper are covered in class, and students are instructed on where to focus their attention while reading the papers, and which parts to skim or skip.

We are also developing an RDS comic series that closely follows the structure of the course and will become part of the assigned reading. The first volume of the series has already been published, and translated into Spanish (Fig. 1) and French (Khan & Stoyanovich, 2020), and the second volume is schedule for release in February 2021.

Fig. 1 The Spanish language edition of "Mirror, Mirror," the first volume of the Data, Responsibly comic series used as supplementary reading for the RDS course. English, French and Spanish editions are available at https://dataresponsibly.github.io/comics/

Preliminary Assessment

The course was offered for the first time in Spring 2019 to a group of 18 registered and 3 auditing students. Its second offering in Spring 2020 had 46 registered students and 4 auditing students. (This level of enrollment is typical for a technical elective at the Center for Data Science at NYU.) The course relies on classroom-based instruction. It moved online for the second half of Spring 2020 due to the Covid-19 pandemic, but still followed the same general methodology — lectures and labs offered synchronously to students over Zoom. The course will continue to be offered to NYU graduate students every Spring. Starting in Spring 2021, we are offering both an undergraduate and a graduate RDS course; the undergraduate course is among the degree requirements of the new BA in data science at NYU.

In Spring 2020 we conducted a pre-course survey delivered to collect information about the students' expectations and initial knowledge, in accordance with the standard of using informal, pre-course survey to glean basic participant information prior to the start of a course (Black, 2003; Nuhfer & Knipp, 2003). In addition to institution-specific questions, we modeled some questions after the Mozilla Foundation survey conducted to gauge public opinion on AI and related topics (Mozilla, 2019). The survey design maps to pre-course survey techniques that seek to gather broad information on the preparedness of students and their exposure to key course concepts. Rather than measuring a specific construct, the goal was to gain information about students' background and prior assumptions before launching the course. The survey had 48 respondents, 41 (86%) of whom were enrolled in masters programs in data science or related disciplines, 6 (12%) were doctoral students, and 1 was an undergraduate student. Approximately 92% were full-time students, and 39% of the students were in their first year of study, meaning that many students engaged with this material early in their degree programs. 27 (56%) self identified as male, 17 (35%) self identified as female, 1 self identified as non-binary, and the remaining 3 students were non-responsive to this question.[3]

The majority of students entered the course with at least some knowledge about algorithms, general programming, and machine learning, consistent with a course composed of technical majors. Approximately half of the class reported having little to no prior knowledge of legal studies or social science. Interpretation is subjective, and we posed various interpretive prompts to students prior to the course in order to gauge the range of interpretive prior knowledge on RDS topics. Figure 2 shows that the majority of students do maintain the view that data science and AI should reflect ethical principles, though it will be interesting for future study to examine views that are reported as neutral.

The course project pursues the broad learning goal of making Automated Decision Systems (ADS) interpretable using the paradigm of an *object-to- interpret-with*, discussed in the next section. Adhering to constructivist principles, students develop *nutritional labels for ADS*. Students develop their own sense of what interpretability means in the

[3] [3] According to data from the NSF National Center for Science and Engineering Statistics (NCSES), female students represent 27% of computer science and 36% of math/statistics enrollments. Data science is not yet tracked as a specific discipline, so we take the previous

statistics that we are within the representative range of enrollments. See https://www.nsf. gov/statistics/2017/nsf17310/static/data/tab3–4.pdf.

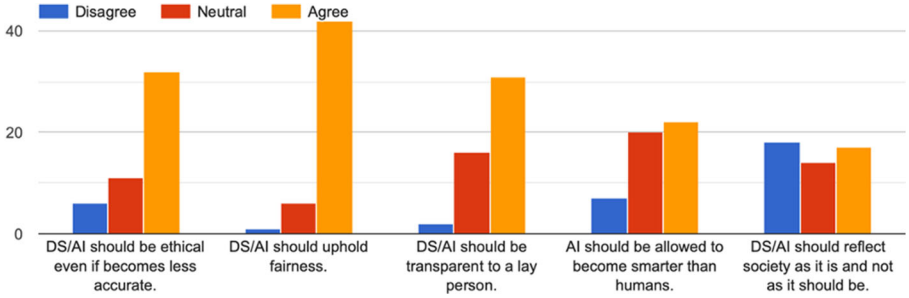

Fig. 2 Pre-course ($n = 48$ respondents) data from Spring 2020, indicating students' view on the role of data science and AI

machine learning context, identify the intended audience(s) for their interpretable design, and uncover aspects of the ADS that are not transparent, or more extremely, may not perform consistently in different contexts. Importantly, the nutritional label as object- to-interpret-with becomes an end-to-end auditing device for the entire ADS.

Objects-to-Interpret-With: Tools for Understanding Data, Complex Models, and the End-to-End Process

Objects-to-interpret-with are representational devices that aid learners in developing heuristic knowledge and understanding contextual knowledge, with a focus on the metacognitive aspects of interpretation and representation. They stem from Latour's claim that artifacts that aid in interpretation do so by helping us "understand how the mobilization and mustering of new resources is achieved" (Latour, 1986). The best public-facing systems adopt a prosumption view that "enables governing knowledge to appear as the product of co-creation rather than an expert technical and methodological accomplishment. It appears to normalize, neutralize and depoliticize statistical analysis" (Williamson, 2016). This more theoretical framing is important because it grounds the object-to-interpret-with as a tool to reveal the partiality and latent mechanisms of underlying algorithms, platforms that run such algorithms, and unforeseeable results that happen when certain datasets interact with machine learning algorithms and platforms.

Objects-to-Interpret-With as Pedagogical Tool

Objects-to-interpret-with allow us to articulate exactly what students learning to interpret complex machine learning models should know. Selbst and Barocas (2018) promote interpretability among technical systems in the following ways that can inform instructional objectives:

- Engineering algorithmic transparency so that the system reveals how sets of inputs lead to certain outputs, and that there is consistency regardless of the diversity in inputs. For example, one can make explicit feature choices or parameters, or incorporate regularization;
- Developing posthoc methods for models that are very complex and/or remain opaque due to business necessity reasons. These methods allow users to understand

model outputs and potentially glean information about how different combinations of inputs yield different results without having low level access to model specifics;
- Creating interactive platforms that allow users to develop their own understanding about model functioning through consistent manipulation of parameters, inputs, and independent variables.

Interacting with systems engineered for algorithmic transparency supports the development of procedural and strategic knowledge, since it simplifies and makes explicit the process of transforming inputs into outputs, and supports the understanding of the model as general method. Working with posthoc methods develops conceptual thinking, focusing on the model as a schema with multiple interpretations. Experimenting with interactive platforms develops metacognitive thinking and heuristic knowledge, where students implicitly build beliefs about their own learning (Hacker, 2009; Schwartz et al., 2008; Tang et al., 2011).

For data science students, the focus is not necessarily on building algorithms, but rather on iterative question formulation and transforming vague goals into measurable parameters (Passi & Barocas, 2019). Objects-to-interpret-with promote making sense of ill-defined information and indefinite meanings to achieve deeper learning. In order for interpretability to occur, a deeper than normal understanding of how inputs lead to outputs is required, and the aforementioned approaches provide more or less of a scaffolding, depending on the sophistication of the student. When teaching complex machine learning models to data science students, particularly with a focus on fairness and transparency, it is beneficial to provide diverse presentations to facilitate the development of heuristics and acknowledge the motley nature of the information being imparted. Recent scholarship (Rau, 2017; Rau et al., 2020) examines the importance of representational competencies in other STEM knowledge acquisition, where students gain an explicit mapping of how visual learning materials map to their target content. Objects-to-interpret-with expands on how representational tools may be designed for or by students in the pursuit of content learning, and incorporates knowledge beyond the discipline's technical and theoretical principles.

There are substantive interpretability learning opportunities found in examining isolated elements (e.g., parameters, variables, nature of input data, and algorithms) and holistic models. Lipton (2016) details common approaches to post-hoc interpretations of models:

- Text explanations: best for metadata and overall contextualization.
- Visualization: best for highlighting individual elements of a model.
- Local explanations: best for black box models; Identifies "an interpretable model over the interpretable representation that is locally faithful to the classifier" (Ribeiro et al., 2016).
- Example aggregation: best for metacognitive and holistic understanding of a model; allows for a learner to build a set of heuristics through exposure to diverse examples (Lundberg & Lee, 2017; Poursabzi-Sangdeh et al., 2018).

Instantiating Objects-to-Interpret-with

Within the RDS research space, there exist several instances of objects-to- interpret-with that embody these interpretability modes of presentation: model cards, data or fact sheets, and nutritional labels, among others. These objects have in common the use of text and visualizations for communication, the incorporation of descriptive,

performance and fairness metrics, and an acknowledgement of context-specific issues that affect model adoption. Additionally, they are developed with broad applicability in mind, both in terms of targeting technical and non-technical audiences, and scaling to different model and data types.

Model cards communicate benchmarked evaluation for trained machine learning models, as well as provide information on intended usage, ethical considerations, and more (Mitchell et al., 2019). One highlight of the proposed model card includes making the role of context explicit. As much background as possible is provided around the creation and maintenance of the training and evaluation data. Evaluation metrics may function differently across demographic or cultural groups, and display interaction effects between groups.

Usage is also contextualized so that various stakeholders recognize the intended set of uses and users, but also any uses that may be mistakenly conflated. It is important to recognize that the model card paradigm is intended for trained machine learning models only, so that it provides a summative snapshot of the entire responsible data science process.

A related concept, the datasheet or data statement, reveals creation, usage, and maintenance knowledge around individual datasets. Likened to an electronics component fact sheet, the datasheet provides insights into the data's lifecycle: motivation through maintenance (Bender & Friedman, 2018; Gebru et al., 2020). The target audience includes dataset creators (they must find the sheet representative of their process) and dataset consumers (technical and non-technical audiences must be able to comprehend the basic information), and the sheet provides guiding questions to answer to achieve this communication balance. The data statement places additional focus on exposing emergent and pre-existing bias for all stakeholders involved.

The nutritional label is another paradigm for interpreting machine learning models, and for the datasets they consume and produce. The most famous nutritional label, the Nutrition Facts panel, began appearing on all packaged foods after the passage of the Nutrition Labeling and Education Act of 1990 (Food and Drug Administration, 1994). This panel and its antecedents evolved over time. Initial versions had the purpose of protecting the public from deceptive and dangerous information about food products, while the current version is geared towards empowering the public to make informed decisions over their nutritional habits. For engaging the general public, it uses a familiar visual artifact to communicate highly technical and opaque information.

The nutritional label combines textual explanations and graphic information, representing a best practice of dual learning theory appropriate for learners (National Research Council, 2000). Research supports the effectiveness of this format, demonstrating that nutritional labels, particularly those that have interactive functionality, can increase one's understanding of a complex topic and lead to better decision making (Byrd-Bredbenner et al., 2009; Gunaratne & Nov, 2017; Kelley et al., 2010; Stoyanovich & Howe, 2019). The nutritional label as a paradigm creates a familiar entry point at which one can engage and start to question the interpretability of a model. In analyzing or creating a label while working at the technical level of the model, one is faced with the ways in which creating a singular, understandable presentation that works in all cases and for all audiences is in fact impossible. There is a utility in creating an artifact that signals certain problematic aspects of the model, particularly for learners with more technical sophistication.

Mapping the research implications of objects-to-interpret-with to pedagogical best practices is straight forward. The objects deliver information in ways that maximize understanding since they (1) present information multimodally (visually, text-based, through case studies, and in some cases, through interaction), (2) provide opportunities for active learning, and (3) develop metacognition. Additionally, the presentation of complex machine learning models provide opportunities for explicit and implicit assessment.

The object-to-interpret-with that we utilize to promote the learning of system interpretability as part of RDS instruction is the *nutritional label*. The nutritional label may be suitable to summarize models like the model card, but, when presented interactively, is most suitable for exploring model and system changes across datasets; thus, the label is highly appropriate for pedagogical purposes. We recognize that the nutritional label is not the singular way to communicate the model to diverse audiences since it does not make a model definitively interpretable to all stakeholders and for all purposes, but is applicable to our particular use case. For facilitating the learning of data science concepts, the nutritional label requires students to actively and iteratively come to understand hidden algorithms, distill the most important information about a model, and adapt that knowledge for non-technical audiences. Nutritional labels synthesize information about machine learning systems into a visually compact format; as a result, they obscure the more complex aspects of a model in the service of visual economy. Learners are supported in understanding a dataset, but also understanding the process more holistically, composed of requirements, data, algorithms, models, and outputs. The ultimate goal is for learners to reflect on this complex process and critique the potential context of use.

Two sets of scholars have explored the use of the nutritional label in data science. Ranking Facts is an application that automatically constructs a user friendly summary of several important properties of algorithmic rankers, including stability, statistical parity, and diversity (Stoyanovich & Goodman, 2016; Stoyanovich & Howe, 2019; Yang et al., 2018). The assumption is that items placed in top ranks are in some sense "better" than items placed lower down the ranked list (e.g., item in rank 1 is of higher quality or relevance than item in rank 150). This simple schema implies a high level of interpretability, though Ranking Facts reveals that rankings may be highly sensitive to inputs and can hide disparate impacts on subsets of data. Revealing this information as a visual gives a deeper understanding of the underlying algorithm, and puts any particular set of rankings into greater context. As seen in Fig. 3, Ranking Facts assists a lay audience in achieving overall model intuitiveness and a level of transparency around the topics of fairness, diversity, and stability. The Fairness pane supports an intuitive understanding of whether or not the model exhibits statistical parity at top ranks, without requiring knowledge of the mathematical properties of either the fairness metric or the ranking process. Similarly, the Diversity pane signals how well different item categories are represented at the top ranks compared to their overall representation in the dataset.

A second team (Holland et al., 2018) also used the nutritional label as an appropriate tool to think about algorithmic transparency. The team developed a framework, the Dataset Nutrition Label, which provides modules that display metadata and source information, textual descriptions and summary statistics of variables, as well as graphs visualizing more complex information like probabilistic models and ground truth correlations. Figure 4 shows a prototype. The Metadata panel summarizes relevant dataset information while the Modeling pane provides model-specific information about performance and accuracy. Other panels go into detail about dataset authorship,

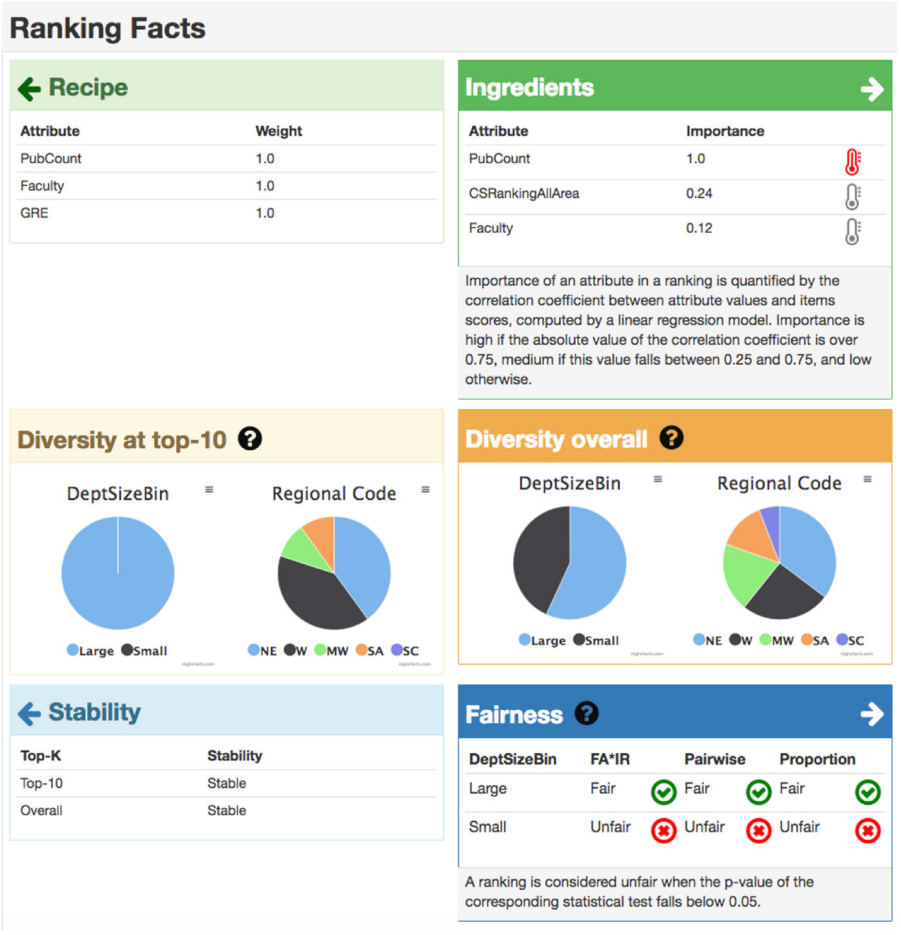

Fig. 3 The Ranking Facts nutritional label for rankings, from http://demo.dataresponsibly.com/rankingfacts/. This interpretable representation of a dataset of university department rankings is constructed automatically by the open-source web-based tool

model variables and ground truth correlations. The Dataset Nutrition Label contributes largely to transparency by taking a descriptive snapshot of a dataset's inherent qualities (i.e., number of records and variables) and introducing some interpretable features (i.e., keywords). The Probabilistic Modeling pane requires knowledge of a model's properties, though a highly visual presentation facilitates understanding. In contrast to Ranking Facts in Fig. 3, which is computed automatically, the Dataset Nutrition Label in Fig. 4 is manually constructed.

Practical Methods and Emerging Best Practices for Teaching RDS

Section 4 details an RDS course that balances the need for students to engage with topics including data protection, fairness, and transparency from both a technical and an

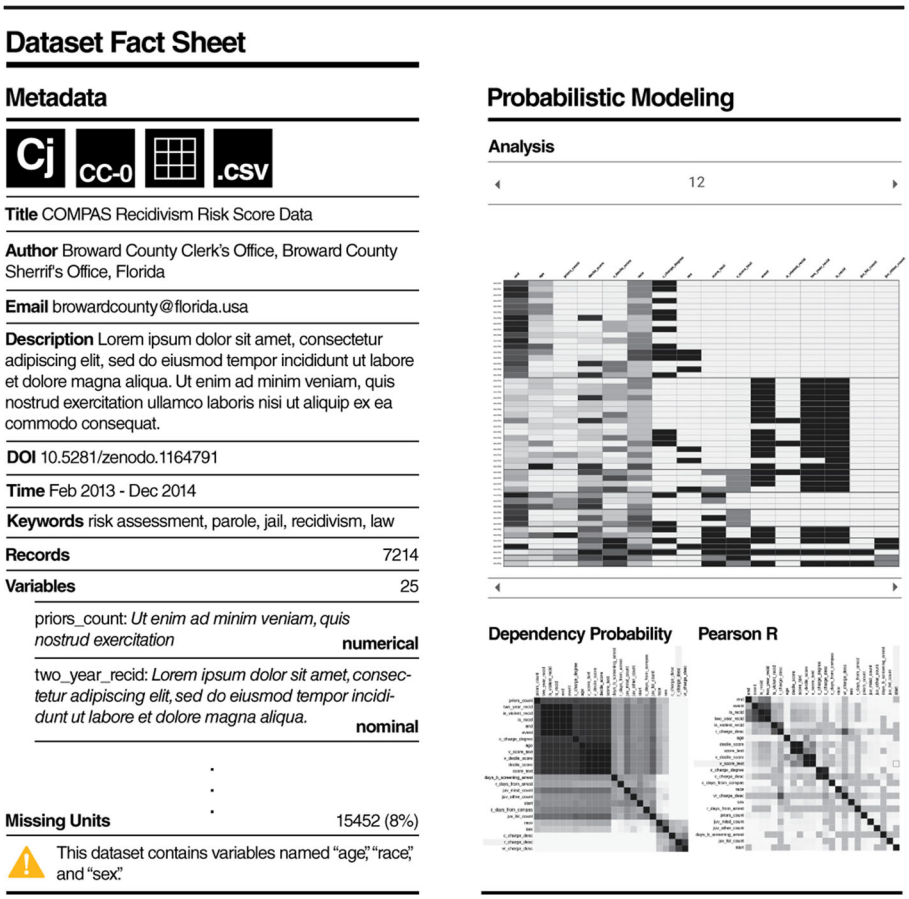

Fig. 4 The Dataset Nutrition Label prototype, from https://ahmedhosny.github.io/datanutrition/. This representation of the COMPAS Recidivism Risk Score Dataset is manually constructed

interdisciplinary perspective. Course activities integrate collaborative and inquiry-based learning, allowing students to broaden their technical and domain knowledge by interacting with peers of varying expertise and backgrounds. The main deliverable is for students to design an original object-to-interpret-with, detailed in Section 5. The selected object-to- interpret-with is the nutritional label, which supports students' inquiry into the end-to-end process of the automated detection system.

Pedagogical Best Practices for Teaching Interpretability

When thinking about teaching students about interpretability of data and models, it is important that activities should incorporate an understanding of how to optimize learning with a need to develop students' technical know how and impart an ethical and contextual understanding. In 2016, the Park City Mathematics Institute issued broad guidelines for integrating ethics into data science education: "Programs in data science should feature exposure to and ethical training in areas such as citation and data ownership, security and sensitivity of data, consequences and privacy concerns of data analysis, and the professionalism of transparency and reproducibility" (De Veaux et al.,

2017, p. 2.8). No specific guidance is offered on how these terms may be defined and/or introduced in a data science context, and below we outline concrete ways to integrate these concepts with technical training, pedagogical best practices, and ethical grounding in mind.

We should build interdisciplinary frameworks for understanding transparency, interpretability, and other relevant concepts. Critical data studies reveals that platforms, as well as the inputs and algorithms of said platforms, are as much socially constructed as any other cultural artifact (Benjamin, 2019; Boyd, 2010; Gitelman & Jackson, 2013). There is an established recognition that, even though machine learning systems may have been created through business- minded, technology-focus perspectives, legal, philosophical, and socio-cultural critical perspectives are requisite considerations (Broussard, 2018; Noble, 2018). Emerging subfields like critical data studies, ethical artificial intelligence, and RDS represent this line of thinking. Infusing these perspectives into data science curricula should be a standard.

There is also a need to leverage into coursework real-world platforms that offer distinct definitions of transparency and interpretability. Distinct versions of transparency and limits of interpretability are revealed through broad exposure to different types of platforms.

Pedagogical Activities for Teaching Interpretability

Pedagogical best practices should inform the development of learning goals, which in turn are used to develop specific activities. Goals can focus on developing skills-building, problem formulation and solving, descriptive and procedural knowledge, heuristics, and more esoteric concepts like metacognition, cooperation, and creativity (Lang, 2016). For learning within technical domains, these categories are supported through activities like worked examples, exposure to common and unusual problems, in-class group problem solving, explicit teaching of models, interaction with simulations, and reflection (Aleven & Koedinger, 2002; Ben-Ari, 2001; Dweck, 1986). If we focus more pointedly on learning RDS topics, we need additional pedagogical techniques.

Table 1 showcases several pedagogically-sound activities that map to best practices in the learning sciences and to RDS interpretability practices. These activities are suitable for diverse groups, for example, (1) those of varying technical and theoretical prior knowledge, (2) those from varied disciplinary backgrounds, and (3) those with different learning goals. These activities, in combination, also support the development of an individual student's knowledge transitioning from novice to expert level.

The RDS course described in Section 4 incorporates some of these described activities. The final course project combines elements of a replication study, process analysis, design and deployment, and nutritional label design. Students replicate a model with an existing dataset and algorithm, and in the process identify transparency flaws, areas for improving interpretability, and ways to improve model performance. Future iterations of the course will tie in further pedagogical principles.

Activities should be offered through multiple presentation modes and levels of interactivity to maximize engagement and promote heuristics development (Wierse & Grinstein, 2002). Current learning science theories map these aspects as key to deeper learning that goes beyond the more superficial knowledge of novices (i.e.,

Table 1 Pedagogical activities for teaching interpretability

Activity	Description	Assessment method
Replicate an existing study	Students reconstruct a (portion of a) published study and highlight any replication issues	*Metrics evaluation*: Evaluation of quantitative scores. *Content analysis*: of replication issues discussion.
Reverse outline/engineer a platform	Students outline a machine learning platform to understand relationships between inputs and outputs.	*Open coding*: A process stemming from HCI research in which students will explicitly label processes related to the platform (Burrell, 2016).
Diagnostic learning logs	A meta activity where students outline points of understanding and confusion about machine learning concepts.	*Points of confusion*: student outlines points of confusion at designated moments throughout a process.
Problem recognition tasks	Students are presented with a set of data science problems and an array of algorithmic transparency platforms and work to identify the best procedure to address the problem.	*Controlled experiment*: different groups are given different platforms with which to explore an identical question. Comparisons and debriefs reveal differences in transparency and interpretability (Wainer & Xavier, 2018).
Process analysis	Students reflect on their process of a deliverable in a meta reflective exercise.	*Cognitive walkthroughs*: form of self report where student outlines decisions made to produce label (Gilpin et al., 2018).
Peer review	Students evaluate other students' performance on an activity.	*Surveys*: Quick method for assessing students' self perceptions about a task or prior knowledge (Skirpan et al., 2018). *Rubric creation*: Groups of students formulate the assessment parameters for their peers.
Before-After	Students iterate on a task, for example, tweaking a particular parameter or variable.	*A/B Testing*: of model performance and interpretability. This is a technique used in practical data science (Kohavi & Longbotham, 2017).
Design and deploy	Students work in groups to deploy a system, potentially using the design sprint technique.	*Rapid ethnography*: a technique that records behavior as students work together.
Create a nutritional label for a machine learning platform	This is a combination of a documented problem solution, where students' understanding emerges implicitly through process-based explanations, and a focus on varying aspects.	*Content analysis*: rubric based analysis of student-produced deliverable. (Schraagen et al., 2000) *Cognitive walkthroughs*: form of self report where student outlines decisions made to produce label (Bainbridge, 2004).

remembering, understanding, and applying) and approaches the more complex cognitive processes required of expert knowledge (i.e., evaluation, metacognition, and creation) (Mayer, 2010; National Research Council, 2000).

We should offer students opportunities for documentation as explanation and proper evaluation. "Careful validation . . . is not enough. Normatively evaluating decision-

making requires, at least, an understanding of: (1) the values and constraints that shape the conceptualization of the problem, (2) how these values and constraints inform the development of machine learning models and are ultimately reflected in them, and (3) how the output of models inform final decisions" (Selbst & Barocas, 2018). There exists a range of activities that support metacognition, the act of thinking about one's thinking, that will assist students in thinking more holistically about how models perform and how we assess this performance. For example, documenting acceptable metrics like the F1 and AUC scores as valid indicators of the accuracy of a model, but noting that these metrics do not necessarily assist in evaluating fairness and transparency. Activities that compel students to contemplate their own thinking become rich opportunities to expose assumptions and knowledge gaps in the way that we evaluate the data science process.

Importantly, we should layer in assessment, both quantitatively and qualitatively, and both formatively and summatively (Angelo & Cross, 1993; Lazar et al., 2010). Assessing how well students achieve model performance and model interpretability is challenging, given the tension between the two goals. The former is metrics-based and therefore quantitatively assessable, while the latter requires a mix of quantitative and qualitative methods to assess whether or not students can interpret a model or find it transparent.

Conclusions and Outlook

In this paper we looked at the pedagogical implications of responsible data science, creating explicit parallels between cutting edge data science research, and cutting edge educational research. We recounted our experience in developing and teaching a responsible data science course to graduate and advanced undergraduate data science students. Further, focusing on transparency and interpretability, we proposed the framework of objects-to-interpret-with and offered to others best practices and concrete implementable techniques for teaching this important topic.

We are excited to see the enthusiasm of students, data science practitioners, and instructors for responsible data science. Given this enthusiasm, and the tangible need of both the industry and academia to welcome a new generation of responsible data scientists, we must come together as a community to meet the challenge of developing curricula and teaching responsible data science. We are at the beginning of the road, and much work remains: in developing instructional methodologies and materials, creating assignments and assessment instruments, and ensuring that the materials we develop stay up-to-date as our understanding of ethics and responsibility in data science evolves. We must also be deliberate in finding ways to scale up curriculum development and instructor training.

There are a few challenges and potentials for studying responsible data science education. A necessary next step is to advance the work of reconciling various disciplinary definitions and critiques of interpretability and explainability in machine learning (Gilpin et al., 2018; Shmueli, 2010). Within the legal and philosophical traditions, there are existing ways of looking at interpretability that have potentials for how students approach material technically. Within institutional circles, there is a need to bridge the vetting practices of ethical bodies such as Institutional Review

Boards (IRBs) with standards of governmental and professional bodies. An additional next step is to integrate existing curricular attempts to teach RDS, which overwhelmingly focus on humanistic approaches to the topic, and identify goals in common that allow us to begin to create a taxonomy of RDS pedagogy, and examine the effectiveness of ethical approaches in technical and humanistic courses.

The potentials for studying RDS education are numerous. We plan on studying the implications of students being exposed to this material at the outset of their degree programs, versus at later points. We would like to further refine the objects-to-interpret-with framework, and develop new and validate existing methods and strategies for teaching this interdisciplinary material within a technical context, with a focus on examining the learning gains associated with various RDS objects-to-interpret-with. We can also look at how students display RDS knowledge across different learning contexts.

In this article we focused primarily on higher education, and in particular on teaching data science students. Going forward, it is crucial to think about educating current data science practitioners and members of the general public. As with the data science student population, transparency and interpretability will prove to be a key concept to investigate and teach.

Code Availability Not applicable.

Funding This research was supported in part by NSF awards No. 1926250, 1934464, and 1922658.

Data Availability All course materials, including the syllabus, weekly reading assignments, complete lecture slides, and lab assignments, are publicly available on the course website at https://dataresponsibly.github.io/courses/. Homework assignments, with solutions and grading rubrics, and a detailed description of the course project, will be made available to instructors upon request.

Declaration

Conflicts of Interest/Competing Interests Not applicable.

References

Aasheim, C. L., Williams, S., Rutner, P., & Gardiner, A. (2015). Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. *Journal of Information Systems Education, 26*(2), 103–115.

Abedjan, Z., Golab, L. & Naumann, F. (2017). Data profiling: A tutorial. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14–19, 2017*, pages 1747–1751.

Ackermann, E. (2001). Piaget's constructivism, Papert's constructionism: What's the difference? In *Conference Proceedings*, volume 1 and 2, pages 85–94, Geneva, Switzerland.

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science, 26*(2), 147–179.

Ali, S., Payne, B., Williams, R., Park, H, W. & Breazeal, C. (2019). Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *Proceedings of International Workshop on Education in Artificial Intelligence K-12 (EDUAI 2019)*, 4.

American Statistical Association (2016). Guidelines for assessment and instruction in statistics education (GAISE): College report 2016. https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989.

Anderson, P., Bowring, J., McCauley, R., Pothering, G. & Starr, C. (2014). An undergraduate degree in data science: Curriculum and a decade of implementation experience. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 145–150.

Angelo, T. A. & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers*. Jossey-bass, San Francisco, CA, 2nd edition.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Association for Computing Machinery (2018). ACM code of ethics and professional conduct.

Bainbridge, W. S. (Ed.). (2004). *Berkshire encyclopedia of human computer interaction*. Great Barrington, Mass: Berkshire Pub. Group.

Baker, R., & Inventado, P. (2016). Educational data mining and learning analytics: Potentials and possibilities for online education. In G. Veletsianos (Ed.), *Emergence and Innovation in Digital Learning, issues in distance education* (pp. 83–98). Edmonton, AB: AU Press, Athabasca University.

Baker, R. S. J. D., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In B. P. Woolf, E. Äımeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent tutoring systems* (Vol. 5091, pp. 406–415). Berlin Heidelberg, Berlin, Heidelberg: Springer.

Barnes, T., Boyer, K., Hsiao, S. I., Le, N., & Sosnovsky, S. A. (2017). Preface for the special issue on ai-supported education in computer science. *I. J. Artificial Intelligence in Education, 27*(1), 1–4.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*, 671.

Ben-Ari, M. (2001). Constructivism in computer science education. *Journal of Computers in Mathematics and Science Teaching, 20*(1), 45–73.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. *Social Forces, 98*(4), 1–3.

Black, P. (2003). *Assessment for learning: putting it into practice*. New York: Open University Press.

Boyd, D. (2010). Privacy and publicity in the context of big data. WWW. Raleigh, North Carolina, April 29.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. Cambridge, Massachusetts: MIT Press.

Buckingham Shum, S. (2019). Critical data studies, abstraction and learning analytics: Editorial to Selwyn's LAK keynote and invited commentaries. *Journal of Learning Analytics, 6*(3), 5–10.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 205395171562251.

Byrd-Bredbenner, C., Alfieri, L., Wong, A., & Cottee, P. (2009). The inherent educational qualities of nutrition labels. *Family & Consumer Sciences Research Journal, 29*(26).

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82–89.

Cormack, A. (2016). A data protection framework for learning analytics. *Journal of Learning Analytics, 3*(1), 91–106.

De Veaux, R., 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., Tiruviluamala, N., Uhlig, P. X., Washington, T. M., Wesley, C. L., White, D., & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application, 4*, 2.1–2.16.

DeFalco, J., Rowe, J. P., Paquette, L., Georgoulas, V., Brawner, K. W., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *I. J. Artificial Intelligence in Education, 28*(2), 152–193.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56–62.

DiPaola, D., Payne, B. H. & Breazeal, C. (2020). Decoding design agendas: An ethical design activity for middle school students. In *Proceedings of the Interaction Design and Children Conference (IDC 2020)*, 1–10.

Doore, S. A., Fiesler, C., Kirkpatrick, M. S., Peck, E., and Sahami, M. (2020). Assignments that blend ethics and technology. In Zhang, J., Sherriff, M., Heckman, S., Cutter, P. A., and Monge, A. E., editors, *SIGCSE*

'20: The 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA, pages 475–476. ACM.

Doroudi, S. and Brunskill, E. (2017). The misidentified Identifiability problem of Bayesian knowledge tracing. In *Proceedings of the 10th International Conference on Educational Data Mining*, page 7, Wuhan, China. International Educational Data Mining Society.

Doroudi, S. & Brunskill, E. (2019). Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK 2019)*, 335–339.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv: 1702.08608.

Drosou, M., Jagadish, H. V., Pitoura, E., & Stoyanovich, J. (2017). Diversity in big data: A review. *Big Data, 5*(2), 73–84.

Dweck, C. S. (1986). Motivational processes affect learning. *American Psychologist, 41*, 1040–1048.

Dwork, C. (2011). A firm foundation for private data analysis. *Communications of the ACM, 54*(1), 86–95.

Epstein, Z., Payne, B, H., Shen, J, H., Dubey, A., Felbo, B., Groh, M., Obradovich, N., Cebrian, M. & Rahwan, I. (2018). Closing the AI knowledge gap. arXiv:1803.07233

Farahi, A. & Stroud, J. (2018). The Michigan data science team: A data science education program with significant social impact. In *Proceedings of 2018 IEEE Data Science Workshop (DSW 2018)*, Lausanne, Switzerland, 120–124.

Ferguson, R. (2019). Ethical challenges for learning analytics. *Journal of Learning Analytics, 6*(3), 25–30.

Fischer, F., Hmelo-Silver, C., Goldman, S., & Reimann, P. (Eds.). (2018). *International handbook of the learning sciences*. New York: Routledge.

Food and Drug Administration (1994). Nutritional Labeling and Education Act (NLEA) requirements (8/94–2/95). https://www.fda.gov/nutrition-labeling-and-education-act-nlea-requirements-attachment-1.

Friedman, B. & Winograd, T. (1990). *Computing and social responsibility: a collection of course syllabi*. Computer professionals for social responsibility: Palo Alto, CA.

Gardner, J., Brooks, C. & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J, W., Wallach, H., Dauḿe III, H. & Crawford, K. (2020). Datasheets for datasets. arXiv:1803.09010.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA 2018)*, Turin, Italy, 80–89.

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 1–14). Cambridge, MA: MIT Press.

Gleicher, M. (2016). A framework for considering comprehensibility in modeling. *Big Data, 4*(2), 75–88.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings from the international conference on learning representations 2015, pages 1–11, San Diego, CA.

Grafberger, S., Stoyanovich, J. & Schelter, S. (2021). Lightweight inspection of data preprocessing in native machine learning pipelines. In *CIDR 2021, 11th Conference on Innovative Data Systems Research, Online Proceedings*. www.cidrdb.org.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2019). A survey of methods for explaining Black box models. *ACM Computing Surveys (CSUR). Surveys, 51*(5), 93:1–93:42.

Gunaratne, J., & Nov, O. (2017). Using interactive "nutrition labels" for financial products to assist decision making under uncertainty. *Journal of the Association for Information Science and Technology, 68*(8), 1836–1849.

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2017). PrivacyPreserving learning analytics: Challenges and techniques. *IEEE Transactions on Learning Technologies, 10*(1), 68–81.

Hacker, D, J. (2009). *Handbook of metacognition in education*. New York: Routledge.

Harvard University (2019). Embedded EthiCS @ Harvard. https://embeddedethics.seas.harvard.edu.

Hilliam, R. & Calvert, C. (2019). Interactive statistics for a diverse student population. *Open Learning: The Journal of Open, Distance and e-Learning, 34*(2).

Hoel, T., & Chen, W. (2016). Privacy-driven Design of Learning Analytics Applications: Exploring the design space of solutions for data sharing and interoperability. *Journal of Learning Analytics, 3*(1), 139–158.

Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. arXiv:1805.03677.

Holstein, K. & Doroudi, S. (2019). Fairness and equity in learning analytics systems (FairLAK). In *Proceedings 9th International Conference on Learning Analytics & Knowledge (LAK 2019)*, 1–4.

Huff, C., & Martin, C. D. (1995). Computing consequences: a framework for teaching ethical computing. *Communications of the ACM, 38*(12), 75–84.

Hundhausen, C. D. and Douglas, S. A. (2000). Using visualizations to learn algorithms: Should students construct their own, or view an expert's? In *Proceedings of 2000 IEEE Symposiumon on Visual Languages*, pages 21–28, Los Alamitos, CA. IEEE Computer Society Press.

Hundhausen, C. D., Douglas, S. A., & Stasko, J. T. (2002). A Meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing, 13*, 259–290.

Kabasenche, W. P. (2014). (the ethics of) teaching science and ethics: A collaborative proposal. *Journal of microbiology & biology education, 15*(2), 135–138.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kelley, P. G., Cesca, L., Bresee, J., and Cranor, L. F. (2010). Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1573–1582, Atlanta, Georgia. ACM.

Khalil, M., & Ebner, M. (2016). De-identification in learning analytics. *Journal of Learning Analytics, 3*(1), 129–138.

Khan, F, A., Stoyanovich, J. (2020). Mirror, mirror. Data, responsibly comic series, volume 1

Kim, B., Patel, K., Rostamizadeh, A. & Shah, J. (2015). Scalable and interpretable data representation for high-dimensional, complex data. In *Proceedings of Conference on Artificial Intelligence (AAAI 2015)*.

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM, 60*(2), 21–23.

Kleinberg, J, M. & Raghavan, M. (2018). Selection problems in the presence of implicit bias. arXiv: 1801.03533.

Knapp, J. A. (2016). Engaging the public in ethical reasoning about big data. In J. Collman & S. A. Matei (Eds.), *Ethical Reasoning in Big Data: An Exploratory Analysis, computational social sciences* (pp. 43–52). New York, NY: Springer International Publishing.

Kohavi, R., & Longbotham, R. (2017). Online controlled experiments and A/B testing. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 922–929). Boston, MA: Springer US.

Kolodner, J. (1991). The Journal of the Learning Sciences: Effecting changes in education. *Journal of the Learning Sciences, 1*(1), 1–6.

Lang, J. (2016). *Small teaching: Everyday lessons from the science of learning*. San Francisco, CA: Jossey-Bass.

Latour, B. (1986). Visualisation and cognition: Thinking with eyes and hands. In Kuklick, H., editor, *Knowledge and Society Studies in the Sociology of Culture Past and Present*, volume 6, pages 1–40. Jai Press.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Indianapolis, IN: John Wiley & Sons.

Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *UC Davis Law Review, 51*(2), 653–717.

Leonelli, S. (2016). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions, 374*(2083), 20160122.

Lin, P., Van Brummelen, J., Lukin, G., Williams, R. & Breazeal, C. (2020). Zhorai: Designing a conversational agent for children to explore machine learning concepts. *In Proceedings of the AAAI Conference on Artificial Intelligence*. 34(9), 13381–13388.

Lipton, Z. (2016). The mythos of model interpretability. In *Proceedings of 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. arXiv:1606.03490.

Lundberg, S. & Lee, S, I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 4768–4777.

Martin, C. D., & Holz, H. J. (1992). Non-apologetic computer ethics education: A strategy for integrating social impact and ethics into the computer science curriculum. In T. W. Bynum, W. Maner, & J. L. Fodor (Eds.), *Teaching computer ethics* (pp. 50–66). New Haven, CT: Southern Connecticut State University.

Mayer, R. E. (2010). *Applying the science of learning*. New York, NY: Pearson.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Mir, D., Howley, I., Davis, J., Peck, E., & Tatar, D. G. (2019). Make and take an ethics module: Ethics across the CS curriculum. In E. K. Hawthorne, M. A. Ṕerez Quiñones, S. Heckman, & J. Zhang (Eds.),

Proceedings of the 50th ACM Technical Symposium on Computer Science Education, SIGCSE 2019 (p. 1239). *Minneapolis, MN, USA*: ACM.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I, D. & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, 220–229.

Mozilla (2019). We asked people around the world how they feel about artificial intelligence. Here's What We Learned. https://foundation.mozilla.org/en/blog/we-asked-people-around-the-world-how-they-feel-about-artificial-intelligence-heres-what-we-learned/.

Mumford, M. D., Connelly, S., Brown, R. P., Murphy, S. T., Hill, J. H., Antes, A. L., Waples, E. P., & Devenport, L. D. (2008). A Sensemaking approach to ethics training for scientists: Preliminary evidence of training effectiveness. *Ethics & Behavior, 18*(4), 315–339.

Naps, T. L., Rößling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., Korhonen, A., Malmi, L., McNally, M., Rodger, S., and Vel'azquez-Iturbide, J. (2002). Exploring the role of visualization and engagement in computer science education. In *Proceedings of ITiCSE-WGR '02*, pages 131–152, Aarhus, Denmark. ACM.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: The National Academies Press expanded edition edition.

Noble, S, U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Nolan, D., & Perrett, J. (2016). Teaching and learning data visualization: Ideas and assignments. *The American Statistician, 70*(3), 260–269.

Nuhfer, E., & Knipp, D. (2003). 4: The knowledge survey: A tool for all reasons. *To Improve the Academy, 21*(1), 59–78.

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

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, NY: Basic Books.

Passi, S. & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, 39–48.

Ping, H., Stoyanovich, J. & Howe, B. (2017). Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, Chicago, IL, USA, 42:1–42:5.

Pinkwart, N. (2016). Another 25 years of aied? Challenges and opportunities for intelligent educational technologies of the future. *I. J. Artificial Intelligence in Education, 26*(2), 771–783.

Poursabzi-Sangdeh, F., Vaughan, J. W., Goldstein, D. G., Hofman, J. M. & Wallach, H. (2018). Manipulating and measuring model interpretability. arXiv:1802.07810.

Quinn, M. J. (2006). On teaching computer ethics within a computer science department. *Science and Engineering Ethics, 12*(2), 335–343.

Rau, M. A. (2017). Conditions for the Effectiveness of Multiple Visual Representations in Enhancing STEM Learning. *Educational Psychology Review, 29*(4), 717–761.

Rau, M. A., Keesler, W., Zhang, Y., & Wu, S. (2020). Design tradeoffs of interactive visualization tools for educational technologies. *IEEE Transactions on Learning Technologies, 13*(2), 326–339.

Reiss, M. J. (1999). Teaching ethics in science. *Studies in Science Education, 34*(1), 115–140.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, CA, USA. ACM.

Rubin, A., Hammerman, J. & Konold, C. (2006). Exploring informal inference with interactive visualization software. In *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS 2006)*, 1–6.

Sawyer, R. (Ed.). (2005). *The Cambridge handbook of the learning sciences*. Cambridge: Cambridge University Press.

Schelter, S. & Stoyanovich, J. (2020). Taming technical bias in machine learning pipelines. *IEEE Data Engineering Bulletin (Special Issue on Interdisciplinary Perspectives on Fairness and Artificial Intelligence Systems), 43*(4), 39–50.

Schelter, S., He, Y., Khilnani, J. & Stoyanovich, J. (2020). Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In *Proceedings of the 23nd International Conference on Extending Database Technology (EDBT 2020)*, 395–398.

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (2000). Introduction to cognitive task analysis. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis* (pp. 3–23). Mahwah, NJ: Erlbaum.

Schwartz, D. L., Chase, C., Chin, D. B., Oppezzo, M., Kwong, H., Okita, S., Biswas, G., Roscoe, R., Jeong, H., & Wagster, J. (2009). Interactive metacognition: Monitoring and regulating a teachable agent. In D. J.

Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *The educational psychology series. Handbook of metacognition in education*, 340–358.

Sclater, N. (2016). Developing a code of practice for learning analytics. *Journal of Learning Analytics, 3*(1), 16–42.

Selbst, A., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*(3), 1085–1139.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. In *Proceedings from the 2012 Conference on Learning Analytics and Knowledge (LAK 2012)*, Vancouver, BC, Canada, 4–8.

Skirpan, M., Beard, N., Bhaduri, S., Fiesler, C., and Yeh, T. (2018). Ethics education in context: A case study of novel ethics activities for the CS classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 940–945, Baltimore Maryland ACM.

Sternberg, R, J. (2010). Teaching for ethical reasoning in Liberal education. Liberal Education (Association of American Colleges & Universities), 96(3).

Stoyanovich, J. & Goodman, E, P. (2016). Revealing algorithmic rankers. Freedom to tinker. https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/.

Stoyanovich, J., & Howe, B. (2019). Nutritional labels for data and models. *IEEE Data Eng. Bull., 42*(3), 13–23.

Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A. & Weikum, G. (2017). Fides: Towards a platform for responsible data science. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 26:1–26:6.

Stoyanovich, J., Yang, K. & Jagadish, H. V. (2018). Online set selection with fairness and diversity constraints. In Böhlen, M. H., Pichler, R., May, N., Rahm, E., Wu, S., and Hose, K., editors, *Proceedings of the 21st International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26–29, 2018*, pages 241–252. OpenProceedings.org.

Stoyanovich, J., Bavel, J. J. V. & West, T. (2020a). The imperative of interpretable machines. *Nature Machine Intelligence, 2*, 197–199.

Stoyanovich, J., Howe, B., & Jagadish, H. V. (2020b). Responsible data management. In *Proceedings of the VLDB Endowment*, 13(12), 3474–3488.

Stoyanovich, J., Kuyan, S., McDermott, M., Grillo, M. & Sloane, M. (2020c). Public engagement showreel, int 1894. NYU Center for Responsible AI. https://airesponsibly.com.

Tang, Y., Shetty, S., Henry, J., Jahan, K., & Hargrove, S. (2011). Interactive and collaborative games promoting metacognition for science and engineering design. In M. Zhou & H. Tan (Eds.), *Advances in computer science and education applications* (Vol. 202, pp. 405–412). Berlin Heidelberg, Berlin, Heidelberg: Springer.

Tractenberg, R. E., Russell, A. J., Morgan, G. J., FitzGerald, K. T., Collmann, J., Vinsel, L., Steinmann, M., & Dolling, L. M. (2015). Using ethical reasoning to amplify the reach and resonance of professional codes of conduct in training big data scientists. *Science and Engineering Ethics, 21*(6), 1485–1507.

Wainer, J., & Xavier, E. C. (2018). A controlled experiment on Python vs C for an introductory programming course: Students' outcomes. *ACM Transactions on Computing Education, 18*(3), 1–16.

Walker, E., & Ogan, A. (2016). We're in this together: Intentional design of social relationships with AIED systems. *I. J. Artificial Intelligence in Education, 26*(2), 713–729.

Wierse, A., & Grinstein, G. (2002). *Information visualization in data mining and knowledge discovery*. San Francisco, CA: Morgan Kauffmann Publishers.

Wilkerson, M. H., & Polman, J. L. (2020). Situating data science. *Exploring how relationships to data shape learning., 29*(1), 1–10.

Williamson, B. (2016). Digital education governance: Data visualization, predictive analytics, and 'real-time' policy instruments. *Journal of Education Policy, 31*(2), 123–141.

Wise, A. F. (2014). Designing pedagogical interventions to support student use of learning analytics. In *Proceedings of the 4th International Conference on Learning Analytics and Knowledge*, pages 203–211, Indianapolis IN, ACM.

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165–181.

Yang, K. and Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA*, Pages 22:1–22:6. ACM

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H. V. & Miklau, G. (2018). A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD Conference 2018)*, 1773–1776.

Yang, K., Gkatzelis, V. & Stoyanovich, J. (2019). Balanced ranking with diversity constraints. In Kraus, S., editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China,* pages 6035–6042. ijcai.org.

Yang, K., Loftus, J. R. & Stoyanovich, J. (2020). Causal intersectionality for fair ranking. arXiv:2006.08688.

Yannier, N., Hudson, S. E., & Koedinger, K. R. (2020). Active learning is about more than hands-on: A mixed-reality AI system to support STEM education. *I. J. Artificial Intelligence in Education, 30*(1), 74–96.

Zimmerman, M, R. (2018). *Teaching AI: Exploring new frontiers for learning*. Portland: International Society for Technology in Education.