INTEGRATING THE HUMANITIES INTO DATA SCIENCE EDUCATION: REIMAGINING THE INTRODUCTORY DATA SCIENCE COURSE

ERIC A. VANCE
University of Colorado Boulder
Eric.Vance@colorado.edu

DAVID R. GLIMP
University of Colorado Boulder
David.Glimp@colorado.edu

NATHAN D. PIEPLOW
University of Colorado Boulder
Nathan.Pieplow@colorado.edu

JANE M. GARRITY
University of Colorado Boulder
Jane.Garrity@colorado.edu

BRETT A. MELBOURNE
University of Colorado Boulder
Brett.Melbourne@colorado.edu

ABSTRACT

Despite growing calls to develop data science students' ethical awareness and expand human-centered approaches to data science education, introductory courses in the field remain largely technical. A new interdisciplinary data science course aims to merge STEM and humanities perspectives starting at the very beginning of the data science curriculum. Existing literature suggests that humanities integration can make STEM courses more appealing to a wider range of students, including women and students of color, and enhance student learning of essential concepts and foundational reasoning skills, such as those collectively known as data acumen. Cultivating students' data acumen requires a more inclusive vision of how the knowledge and insights generated through computational methods and statistical analysis relates to other ways of knowing.

Keywords: *Statistics education research; Data acumen; Data science education research; Humanities education research; Statistical literacy; Data literacy*

1. INTRODUCTION

In its 2018 consensus study report on undergraduate data science education, the National Academies of Sciences, Engineering, and Medicine (NASEM, 2018a) in the United States called for important changes in the data science curriculum. Charged with "setting forth a vision for undergraduate education in data science," the report's authors sought to "engage underrepresented student populations and consider ways to reduce the 'leakage' seen in existing STEM pathways" (p. xii). To achieve these outcomes, the report set forth several important recommendations for undergraduate data science programs.

Our multidisciplinary team of faculty at the University of Colorado Boulder, USA (CU Boulder) considers data science to be the science of learning from data (Donoho, 2017), which encompasses—at a minimum—statistics, computation, and ways of thinking about data. From this point of reference,

Statistics Education Research Journal, 21(2), Article 9. https://doi.org/10.52041/serj.v21i2.42

we seek to rise to the challenge of the NASEM (2018a) report. We extract from it three overarching goals:

Goal 1: Create a more data-engaged and enabled citizenry
Goal 2: Educate more data scientists
Goal 3: Educate better data scientists

We extract our first goal from NASEM's (2018a) recommendation to instill a "basic understanding of data science in all undergraduates" (Recommendation 2.3). We envision expanding the number of data-engaged citizens, spreading data engagement and data literacy more broadly throughout society. We want our students to address the greatest challenges of our present moment: how to discern credible facts from misinformation and how to radically reimagine our relationship with our data-driven world and what it means to be a productive citizen.

The second goal of educating more data scientists emerges from the first, and from NASEM's (2018a) calls to develop a "range of educational pathways" into data science (Recommendation 2.2) for students of "varied backgrounds and degrees of preparation" (Recommendation 4.1). To reach undergraduates from all demographics, our program must strive for non-STEM or non-math students in STEM to view themselves as potential data scientists, and must attract and retain women, students of color, and students suffering from wealth and class inequalities, among others. Our approach must address issues of race, class, gender, and ability, and consider the differentials of power that have historically prevented access to data-science knowledge from being distributed evenly across all groups.

The third goal of educating better data scientists, building upon the first two goals, aims at the heart of NASEM's (2018a) vision: the need to instill in data scientists "the ability to understand data, to make good judgments about and good decisions with data, and to use data analysis tools responsibly and effectively," what the report's authors call "data acumen." This set of skills will distinguish those future data scientists who have the ability to counteract the dangers of biased data and algorithmically automated injustice. The consensus study report recognizes that, at a minimum, instilling data acumen requires weaving ethics into the data science curriculum from day one (Recommendation 2.4).

In planning to achieve these goals, it is useful to consider three types of students of interest to data science educators: students who never take a data science course (Type 0), students who take exactly one data science course (Type 1), and students who take multiple data science courses (Type 2+). Type 0 students will leave higher education having never taken a data science course. But because we live in a data economy, these students will be at a tremendous disadvantage when they graduate if they do not understand the data-driven systems that pervade all aspects of our daily lives—from their Amazon shopping cart choices to their seemingly trivial "liking" of a friend's TikTok video that could generate not only targeted advertisements and personalized recommendations, but potentially even inputs into predictive policing algorithms applied in their communities. Thus, we seek to convert our Type 0 students into Type 1 students, at least, by motivating them to enroll in at least one introductory data science course.

If we are to create a data-engaged and enabled citizenry (Goal 1), then at a minimum, our Type 1 students who take only a single data science course must learn to evaluate claims that use data as evidence, put forth their own sound data-driven arguments, and assess the limits of data science approaches and ethical implications of utilizing big data. Type 1 students must leave the university with a healthy skepticism, understanding that the manner and purpose of data collection affect how data should be analyzed and what usable information can be extracted. Their sole data science course must instill a fundamental computational and statistical literacy, and teach best practices for communicating with and about data, in ways that provide lifelong benefit even if students pursue no further education in data science.

However, even as our ideal introductory data science course makes data-literate citizens of its Type 1 students, it must also invite students deeper into the field. To educate more data scientists (Goal 2), we must transform Type 1 students into Type 2+ students who build upon the foundational first course by taking more data science courses.

We consider Type 2+ students to be a very diverse group that includes not only the mathematics, computer science, statistics, and data science majors who will pursue advanced study of data science, but also a wide variety of other students who will use data science in their careers: science and

engineering majors who will take advanced computing and statistics courses; social science students who will make data-driven decisions in future courses and jobs; teachers in training who will undertake quantitative research of the learning in their classrooms; and humanities students who will use data science to advance inquiry in their chosen fields, such as computationally analyzing large corpa of text. In our experience, many students who do not major in data science or related fields nevertheless become data scientists after graduation. Our ideal introductory course, then, should not only attract more students to major in data science, but also open up pathways of study that lead back into other majors via "connector courses", such as those at UC Berkeley (Lue, 2019) and allow for culminating experiences such as interdisciplinary "capstone" courses in which final year students from multiple disciplines and varieties of data science collaborate in teams to solve challenging, real-world problems (Vance & Smith, 2019; Vance et al., 2022).

To create better data scientists (Goal 3), we must infuse our entire curriculum with the development and cultivation of data acumen. Questions of ethics, sound reasoning, and effective communication must not be sequestered in a la carte courses. Instead, we seek to weave them into the fabric of introductory courses, connector courses, capstone courses, and the upper-division data science curriculum. In this way, our three overarching goals and our concern for three different kinds of students combine to generate our comprehensive vision for an inclusive and interdisciplinary data science curriculum.

To achieve these goals, we have designed an inclusive interdisciplinary introductory data science course (IIIDS) as a broadly appealing on-ramp for more students into data science. For such a course to be an option for the current large population of Type 0 students, it must have no prerequisites and require no prior experience in statistics or computing. But to simultaneously invite and prepare students for further study in the field, we weave together statistical reasoning, coding, and humanistic forms of inquiry into our IIIDS course. Its objective is both to provide STEM majors with methods and frameworks for qualitative reasoning that are traditionally taught in the humanities and to provide humanities majors with quantitative reasoning. Mindful of the concerns voiced by Hardin et al. (2021) about teaching computer science and statistics as *skills* rather than as *ways of thinking* and analysis, our effort has been to develop a course that instills data acumen by putting computational and statistical ways of thinking in active dialogue with humanistic approaches to understanding the world. IIIDS represents an effort to address the serious gaps in data science education by training students from the earliest moment of their college-level data science studies to think reflexively about how data science produces knowledge, about how data are analyzed and interpreted by data scientists, and about how data science analyses are used by others. By fully integrating humanities approaches into data science education, rather than relegating such questions and concerns to capstone courses or to electives taken along the way, we aim to prepare students to synthesize qualitative and quantitative approaches to urgent research questions and give students practice putting data to work in the world. By integrating statistical, computational, and humanities ways of thinking, IIIDS is designed to cultivate data acumen.

In Sections 2–4, we review the literature relevant to each of our three overarching goals and discuss how the design of our curriculum must respond. In Section 5, we review the relevant literature on pedagogy that informs our new IIIDS course, which we describe in Section 6. We conclude this paper with a call to action for the data science education community to continue to think inclusively about how to better educate more data scientists. Ultimately, we seek to answer key questions posed in the *Statistics Education Research Journal* call for papers (Biehler et al., 2020): What are new ways to engage students in studying data science? Which new topics should be included in the data science curriculum? And what knowledge, skills and dispositions are required in data science to develop data acumen?

2. CREATING A MORE DATA-ENGAGED AND ENABLED CITIZENRY

Becoming a data-engaged citizen is not only a matter of developing a basic understanding of quantitative reasoning with data. Citizens also require critical awareness about data and the uses and potential abuses of data science methods. These issues are only becoming more urgent, especially in light of the growing concerns about the downsides of "big data" (Ridgway, 2016).

A growing number of authors have demonstrated persuasively how machine learning algorithms encode racial and gender discrimination, contribute to economic disparities, reinforce bias, and

contribute to injustice (Benjamin, 2019; D'Ignazio & Klein, 2020; Franks, 2020; Noble, 2018; O'Neil, 2016; Zuboff, 2020). The complexity and opacity of algorithmically-driven decision making, to say nothing of machine learning techniques or applications developed without a deep understanding of the context in which they will be applied, creates the risk of unanticipated consequences, for instance as recommender algorithms have adverse effects on our political processes (e.g., Lepore, 2020). There is a widespread sense both that data science is an important new frontier of inquiry vital to economic growth and to increasing our knowledge of the world; at the same time, there is a growing understanding that the tools of data science present new risks the scope of which we are only becoming aware. As most (though not all!) of the authors cited above acknowledge, along with the efforts of the organizations such as the Ida B. Wells Just Data Lab at Princeton University, the Human Centered Data Science Lab at the University of Washington, Data for Black Lives, and the Algorithmic Justice League, among many others, the answer is not to resist data science, but to make it better, to acknowledge dangers of biased data and algorithmically automated injustice, and to create better ways to understand and use data.

Students need to understand how such algorithms can trap them in what O'Neil calls a "pernicious" feedback loop in which automated data systems amplify the effects of racial and gender biases that are already in place (O'Neil, 2016). Similarly, D'Ignazio and Klein (2020) show how the systems of power that inform the demographics of data science may prevent many students from becoming truly effective civic and economic citizens, and how both gender and racial biases are literally encoded into some of the most pervasive data-driven systems that infiltrate our students' everyday lives.

If we do not succeed in converting Type 0 students into Type 1 students at least, we leave them ripe for exploitation by these data-driven systems. We risk excluding them from jobs and careers related to the collection and analysis of data. And we risk perpetuating a dangerous naivete about political and economic arguments that claim to arise from "the data." As Engel states, "An enlightened citizenry that is empowered to study evidence-based facts and that has the capacity to manage, analyze and think critically about data is the best remedy for a world that is guided by fake news or oblivious towards facts" (2017, p. 45).

3. CREATING MORE DATA SCIENTISTS

A number of groups have echoed NASEM's (2018a) call for broadening access to data science, not only for the purpose of making more data scientists but also to diversify the makeup of the undergraduate students involved pursuing data science coursework. Such commitment to diversity and inclusion is a key element of the Association for Computing Machinery's January 2021 report of its Data Science Task Force, *Computing Competencies for Undergraduate Data Science Curricula*. The NASEM Committee on Envisioning the Data Science Discipline (2018a) argued for the advantages of abandoning a "pipeline" metaphor for thinking about student recruitment into the discipline, and thinking instead in terms of a "watershed" approach, in which there are multiple flow pathways by which students enter a degree program dependent upon their own backgrounds. For inherently interdisciplinary degree programs with multiple potential routes for student success, such a metaphor structures a more open, collaborative approach toward building programs that attract diverse students. We seek to design a program with multiple entry points, broad appeal, numerous paths forward, and strong support for all students as they advance.

A key factor in student success and persistence in STEM fields such as data science is the students' own perception of themselves as scientists or "science people" (Brickhouse et al., 2000; Carlone & Johnson, 2007). These "science identities" are often formed as early as middle school (Calabrese Barton et al., 2013; Carlone et al., 2014; Kang et al., 2019) and are particularly important for students from underrepresented groups (Brickhouse et al., 2000; Espinosa, 2011; Kang et al., 2019). These early-life affinities can have less to do with the subject matter than with the way it is taught. Many students who gravitate toward STEM fields cite a desire for "one right answer" and a discomfort with ambiguous or subjective grading policies, while many students who gravitate away from STEM fields believe that STEM courses afford them little room for the creativity, intellectual exploration, and larger meaning that the humanities provide (Sjøberg, 2002; Steele et al., 1974; Tobias, 1993; Valenti et al., 2016). These attitudes, however, spring from misconceptions. Multiple researchers (Driver et al., 1996; Kessels et al., 2006; Zeidler, 2016) attribute negative attitudes toward STEM to students' misunderstanding of

science as a positivist enterprise, "the unproblematic collation of facts" (Kessels et al., 2006), rather than as a halting, imperfect, perennial, collaborative, and sometimes competitive search for the truth.

If we are to attract students to an introductory data science course who would otherwise never take it, including students from underrepresented groups, we must design a course that appeals to those who do not consider themselves "science people." One way to do this, the research suggests, would be to infuse the curriculum with opportunities to be creative and seek larger meaning.

4. CREATING BETTER DATA SCIENTISTS

4.1. THE NEED TO RETHINK THE INTRODUCTORY COURSE

To achieve the goals we have articulated above, our IIIDS course needs to look quite different from the typical first-level course in data science. The NASEM report recommends that universities "avoid filter or gate-keeping courses (especially early in the program) and replace them with courses that entice student participation through heightening the excitement and applicability of data science" (2018a, p. 64). Yet despite this and other calls to develop data science students' ethical awareness (Saltz et al., 2018), to emphasize "communication, reproducibility and ethics" (De Veaux et al., 2017), to teach larger-scale critical thinking skills (Engel, 2017), and to directly address issues of encoded bias (O'Neil, 2016; Noble, 2018; Benjamin, 2019), introductory courses in the field remain largely focused on developing computational and statistical modes of understanding. Courses in ethics and courses that apply data science to specific social science or humanities disciplines tend to be treated as "add ons," either as electives or as requirements scheduled after a regimen of largely technical introductory courses. That is, if they appear at all; one survey (Tang & Sae-Lim, 2016) did not find "ethics" among the high-frequency words in the curricular descriptions of 30 data science programs.

Çetinkaya-Rundel and Ellison (2021) provide empirical insight into the content of first year data science courses. Though the survey sample was small and focused on elite institutions, their survey of common topics covered in five first-year data science courses demonstrates that first year instruction in data science either subordinates or excludes the kind of emphases necessary for achieving NASEM (2018a) goals. Two of the five programs they surveyed omit ethics altogether, and one omits communication.

The data science curriculum guidelines proposed by De Veaux et al. (2017) acknowledge the importance of "communication, reproducibility and ethics" to undergraduate training in data science, but their concrete recommendations frequently either fail to include any coursework that would actually provide insight into such concerns, or defer such questions to capstone courses or "add-on" classes in technical writing or ethics. Though we respect the challenge of defining criteria for data science education in a period of relative austerity in the academy in the United States, the curricular model proposed by De Veaux et al. would defer or delegate the kinds of questions that the NASEM (2018a) report argues should be central to data science education.

4.2. THE CHALLENGE OF DEFINING DATA ACUMEN

In many ways, the difficulty in suffusing "data acumen" throughout the entire data science curriculum is internal to the challenge of defining data science as a field and to developing a consensus about what students should learn to become responsible data scientists. Because of the interdisciplinary nature of data science, instilling data acumen requires exposure to a wide range of fields and mastering a disparate and comprehensive set of competencies. The point is central to Donoho's (2017) argument for a "greater data science" encompassing a range of knowledges beyond those emphasized in more traditional statistics curricula. The NASEM (2018a) report offers an even broader vision for the future of data science education. Here is their list of "key concepts" needed for cultivating data acumen, and which they hope to extend to all students:

- Mathematical foundations,
- Computational foundations,
- Statistical foundations,
- Data management and curation,

- Data description and visualization,
- Data modeling and assessment,
- Workflow and reproducibility,
- Communication and teamwork,
- Domain-specific considerations, and
- Ethical problem solving. (NASEM, 2018a)

The order of this list, which moves from "foundational" elements of data science education to the elements presumably built on those foundations, suggests that "communication," "domain-specific considerations," and "ethical problem solving" are less central to the discipline than mathematics, computation, and statistics.

At the same time, in contrast to the logic of the list, the NASEM (2018a) report argues that the various aspects of data acumen it specifies are closely interrelated. The report declares that the success of data science as a discipline rests on the ability of its practitioners to communicate clearly its findings, to work across multiple domains, and to develop analytic tools necessary for addressing the potential risks and abuses of data science (see pp. 15–16). The difficulty of imagining undergraduate data science education as a straightforward sequence from math, computer science, and statistics to an isolated treatment of each component of data acumen is especially clear with the authors' statements about ethics.

> Unique ethical considerations arise in each step of and throughout the data science life cycle (i.e., when posing a question; collecting, cleaning, and storing data; developing tools and algorithms; performing exploratory analysis and visualization; making inferences and predictions; making decisions; and communicating results. (p. 30)
>
> Ethics is a topic that, given the nature of data science, students should learn and practice throughout their education. Academic institutions should ensure that ethics is woven into the data science curriculum from the beginning and throughout. (p. 3)

The report also includes a "Data Science Oath" (Appendix D, NASEM, 2018a) modeled on the Hippocratic Oath that not only reaffirms the promise to do no harm, but goes so far as to say,

> I will remember that there is art to data science as well as science and that consistency, candor, and compassion should outweigh the algorithm's precision or the interventionist's influence.

When the NASEM (2018a) report defines data acumen as "the ability to understand data, to make good judgments about and good decisions with data, and to use data analysis tools responsibly and effectively" (p. 12) it lays a provocation at the feet of data science educators. The report exhorts us to move beyond a technical focus on skills and a narrow discipline-specific definition of how to interpret data. We believe it calls upon the field to integrate not just ethics, but a broader set of skills and habits of mind. Our animating insight is that calls for "data acumen" of this kind can be understood as calls to integrate into data science the core competencies of the humanities.

4.3. INSTILLING DATA ACUMEN VIA THE HUMANITIES

The NASEM report's "Data Science Oath" (Appendix D, NASEM, 2018a) is fundamentally a pledge to keep the human dimensions of computational inquiry at the heart of data science. Its implicit argument is that becoming a responsible and human-centered data scientist is not only a matter of knowing how to code and possessing a deep understanding of statistical reasoning. It requires a pervasive awareness that all data are created by humans for humans, that human objectives and cognitive biases shape data analysis and use, and that data can be used (and misused) with enormous human consequences.

Most data science educators would agree that teaching data acumen requires computational thinking and basic statistical reasoning concepts, including understanding the *randomness, variability, and uncertainty* inherent in a given problem; ensuring acquisition of *high-quality data*; understanding the *process that produced the data*; and approaching modeling as a process that requires an *overall strategy* (Hardin et al., 2015). These higher-order concepts in fact closely mirror the core competencies of humanities disciplines. As another recent NASEM (2018b) report details,

> The humanities teach close reading practices as an essential tool, an appreciation for context across time and space, qualitative analysis of social structures and relationships, the importance of perspective, the capacity for empathic understanding, analysis of the structure of an argument (or of the analysis itself), and study of phenomenology in the human world. (p. 60)

We can map those competencies onto the ones Hardin et al. (2015) describe: educators from the humanities would seek to teach data science students about the uncertain provenance of all information (including data) and the role of individuals situated in specific cultural and historical contexts in producing that information. They would seek to teach how the creation of categories—a key step in the process of turning unordered information into data—necessarily involves practices of representation central to how we understand ourselves and others, both in terms of inclusion and exclusion and in terms of the categories and concepts we use to describe ourselves and others. Humanists would seek to encourage data science students to view the big picture and assess the typically (unstated) premises upon which a project rests. Potentially, they would also seek to get students to engage in examining how knowledge can be used positively to transform the world and, by the same token, can be used to exploit or otherwise harm others.

Data science educators around the world have begun to recognize the importance of human-centered approaches to the field to help students understand the risks and benefits of data science analysis (Anderson & Parker, 2019; Aragon et al., 2016; Wu et al., 2020). Integrating the humanities into the data science curriculum could also provide a road to a "science identity" for students who lack one, by spotlighting the type of creative and big-picture thinking that such students fear the discipline is missing (Sjøberg, 2002; Steele et al., 1974; Tobias, 1993; Valenti et al., 2016).

5. GROUNDING OUR APPROACH IN STEM AND STATISTICS PEDAGOGY SCHOLARSHIP

The IIIDS course builds on over two decades of scholarship focused on improving STEM education. This scholarship grounds our approach and justifies the emphasis on active, research-oriented, problem-based pedagogy. One set of well-supported findings speaks to the effectiveness of *active learning* in a student-centered classroom (National Research Council et al., 2012; Prince, 2004; Udovic et al., 2002). Another testifies to the effectiveness of *collaborative learning* (National Research Council et al., 2012; Prince, 2004; Slavin, 1989; Vance & Smith, 2019). In particular, deftly structured work in *mixed-ability groups* can boost not only student achievement at all ability levels, but also equitable relations among students from diverse backgrounds (Boaler, 2008; Paushter, 2017), as the IIIDS course aims to do by integrating Type 0, Type 1, and Type 2+ students into a single course.

The move toward *integrated STEM instruction* (Bybee, 2010; Kelley & Knowles, 2016; Purzer et al., 2014; Sanders, 2009; Sheahan & White, 1990; Stohlmann et al., 2012) seeks to extend active learning by situating it in realistic and relevant contexts. These approaches reinforce arguments that educators must develop learners' ability to be not only "producers" but also to act as informed, reflective, and critical "consumers" of data and research results (Gal, 2002; Gould, 2010, 2017), and we agree with Ograjenšek and Gal (2016) that an effective way to do this is to emphasize qualitative thinking and qualitative research methods in data science courses.

Reinforcing such approaches, considerable work has supported the implementation of *problem-based learning* (PBL), which engages students in solving complex, ill-defined problems that do not have a single clear solution (Greenwald, 2000; Prince, 2004; Reeves & Laffey, 1999; Schraw et al., 1995). Ill-defined problems require students to iterate their problem-solving process and re-evaluate their assumptions as they do. The 2016 Guidelines for Assessment and Instruction in Statistics Education (GAISE) specifically recommends that courses that teach statistical thinking employ an investigative process of problem-solving and decision-making, integrating real data with a context and purpose (Carver et al., 2016).

A natural extension of the problem-based approach is to engage students in research-based learning, as recommended by the Boyer Commission on Educating Undergraduates in the Research University (1998). Students who participate in *undergraduate research experience* (URE) opportunities show a greater understanding of the research process, have increased confidence in their own potential to work as a scientist, have increased graduation rates, and are ultimately more likely to continue on a science-

related career path (Laursen et al., 2010; Lopatto, 2004, 2009; Russell et al., 2007; Seymour et al., 2004; Thiry & Laursen, 2011; Vieyra et al., 2011). This benefit is often strongest for underrepresented groups including racial, gender, and socioeconomic minorities (Hernandez et al., 2013; Hurtado et al., 2009; Nagda et al., 1998; Villarejo et al., 2008). However, the standard URE model, in which individual students conduct research projects with one faculty member, is limited by cost and faculty availability (Desai et al., 2008; Wood, 2003) and can disproportionately filter out first generation students, women, non-traditional students, and other groups historically underrepresented in science (Bangera & Brownell, 2014).

Course-based approaches to undergraduate research can overcome these hurdles and allow educators to reach larger numbers of students from different backgrounds (Weaver et al., 2008; Wei & Woodin, 2011). So-called *course-based undergraduate research experiences* (CUREs) are thus increasingly championed as scalable ways of involving undergraduates in science (Corwin et al., 2015). Studies of CUREs to date suggest that participating students achieve many of the same outcomes as students who complete individual research experiences, and that engagement in CUREs can strengthen students' views of themselves as scientists, and can increase graduation rates and completion of STEM degrees (Brownell et al., 2012; Corwin et al., 2015; Harrison et al., 2011; Rodenbusch et al., 2016). CUREs also have a particularly strong impact on minority participation and thus could function as an inclusive gateway to further undergraduate participation in independent research and a career in data science (Bangera & Brownell, 2014; Hurtado et al., 2009; Thiry et al., 2012).

6. INTERDISCIPLINARY INCLUSIVE INTRODUCTORY DATA SCIENCE (IIIDS) COURSE

Our IIIDS course is designed to be a "watershed" course in the sense of the term advocated by the NASEM Committee on Envisioning the Data Science Discipline (2018a) as described in Section 3. Our team-taught class offers a rigorous introduction to the field in a way that invites more—and more diverse students—into the discipline of data science. Casting a wider net, capitalizing on the inherently interdisciplinary nature of data science inquiry, our course aspires to provide a platform for outreach to students outside the traditional STEM pipelines in ways that will contribute to the discipline of data science's diversity and inclusion efforts.

From its inception, IIIDS was built to reimagine how introduction to data science courses can be taught. Specifically, the course:

- Was developed by an interdisciplinary team of faculty and students from the humanities and sciences, working with educational specialists in CU Boulder's Arts and Sciences Teaching with Technology (ASSETT) group;
- Has no prerequisites;
- Is team taught, with one faculty member from a STEM discipline and one faculty member with a background in the arts and humanities;
- Involves student input from a team of undergraduate students facilitating the construction of modules around which the course is structured;
- Uses team-based learning (Vance, 2021) to teach core principles of coding, statistical inference, and humanistic modes of analysis;
- Strives to model the research process for undergraduate students, to give undergraduates a realistic understanding of the complex interweaving of different kinds of knowledge and different competencies;
- Includes ethical reflection and decision-making that takes place at each stage of the research process, from study design to reporting and communication; and
- Is built on peer-based interactions to teach core principles and to provide experience with teamwork, collaborative problem solving, and communication.

So doing, IIIDS endeavors to address the gaps in data science education described in Section 4.1 and thereby advances the three goals for data science pedagogy outlined at this paper's outset.

The problem-based approach IIIDS undertakes is key to its effort to integrate humanistic approaches and concerns into the teaching of computational and statistical approaches to understanding the world. The course is built around a series of modules (see the syllabus provided in Appendix A) that examine

and model the experience of data science research. Within each module are case studies designed to illustrate and provide practice implementing key coding and statistical principles, invite questions about representation and power that influence each stage of the data science research process, and lay the foundation for data acumen. Expanding students' interpretive resources by teaching them methods for thinking both in terms of statistically valid inference and humanistically-grounded ways of knowing will contribute to students' ability to understand, explain, and evaluate their data science work. That is, building critical and ethical competencies into IIIDS will contribute to making better data scientists.

Specifically, the modules:

- Train students to evaluate the sources of data, methods of collection, quality, social and political motives behind the collection of a given data set, and the implications for the resulting analyses and interpretations. The emphasis runs throughout our sample syllabus (see Appendix A), though a selection from Anderson's *The American Census: A Social History* will help frame students' work with United States census data in the course's third module (2015);
- Invite students to question the representational choices made during the development of a data science project including the categories used to describe and exclude aspects of historical reality. As above, the emphasis on developing a reflexive awareness of such choices runs throughout the course. A specific example can be found in the course's fourth module centered on analyzing a corpus of science fiction narratives in which students will read Chakrabarty's important essay, *Anthropocene Time,* by way of asking how a new concept can impact the kinds of questions data scientists can ask (Chakrabarty, 2018);
- Require students to assess the coherence and effectiveness of arguments incorporating data visualizations and analyses and weigh the benefits and risks of data analysis or computational algorithms, their potential uses and misuses, and their ethical implications;
- Require students—via team lab reports—to support clear and effective arguments by responsibly incorporating many kinds of evidence, including both the findings of data science analyses and modes of humanistic inquiry.

Mindful of the cognitive load our course places on students—many of whom are learning to code for the first time—we do not cover as much statistics and computation compared to the more traditional introductory data science course the first author has previously taught (Vance, 2021). For example, rather than covering all 30 chapters of the textbook, *R for Data Science* (Wickham & Grollemund, 2017), IIIDS covers only Chapters 1–12, which we feel gives students a solid base for which to learn more advanced computation and statistical modeling on their own or in subsequent courses. Our students are required to choose at least one topic from Chapters 13–30 to apply to their final project. IIIDS covers most of the same statistical thinking topics as the traditional course, but does not progress all the way to conducting hypothesis tests via simulation, which is the culmination of the statistical thinking topics in the first author's traditional introductory data science course.

Although there are clear losses in coverage involved in our new course, there are also significant gains for students that we believe justify the emphasis of our approach. As students learn basic statistical and computational concepts they will also develop a reflexive awareness of the strengths and weaknesses of computationally enhanced statistical inquiry. From the beginning of their encounter with data science, our students will learn to ask questions about how data is gathered, transformed, and used, and the risks associated with each stage of the research process. Rather than defer to a later moment in their education questions about the vulnerabilities of data science inquiry, for instance its capacity to mask bias and perpetuate unjust power relations, the course introduces students to humanist modes of inquiry alert to the way our representations of the world can reinforce or magnify damaging forms of social relations. Though there are clear tradeoffs in terms of coverage, our inclusive approach provides data science educators with one model of how to begin to train responsible data citizens and data science practitioners by emphasizing the power of combining STEM-based and humanistic ways of knowing.

Because of the unique way in which our course integrates disciplinary perspectives, IIIDS has received approval for fulfillment of either the General Education Arts and Humanities divisional distribution requirement or the Quantitative Reasoning skills requirement in the College of Arts and Sciences at CU Boulder. This approval by the College's faculty-led Curriculum Committee validates our effort responsibly and rigorously to combine approaches drawn from data science and humanistic

disciplines. IIIDS is team-taught by this paper's first two authors, one faculty member from the humanities and one from statistics and applied mathematics. Because both faculty members are involved in instruction throughout the semester, this course is deeply and consistently interdisciplinary. IIIDS approaches data science as an important new field for understanding fundamental intellectual and ethical dimensions of human experience; rigorously examines the relationships and differences between quantitative and humanistic modes of inquiry and interpretation; and provides a solid introduction to new literacies that are vital for students to navigate a world where big data is becoming ever more pervasive.

As we have argued throughout, our course can serve as a "watershed" entrance into the field of data science. A corollary benefit of our efforts is to prepare students for changes that are likely to impact most if not all academic disciplines. The advent of widely available capacities for large scale computing and advances in methods for statistical analysis have led to the development of new research methods and new research paradigms across nearly every discipline in the university. Our hope is not only to develop a foundational, humanistically inspired grounding for students in STEM fields, but also to lay the groundwork for further work across multiple disciplines, most centrally the humanities. Students will gain a new set of data science resources for engaging in cross-disciplinary work, and a metacognitive competence necessary for understanding the methodological implications of applying data science tools responsibly across disciplines.

7. CONCLUSION

We conclude this paper with a call to action for the data science education community to continue to think inclusively about how to better educate more data scientists and to collaborate with their colleagues to integrate the humanities into data science education as a way to teach data acumen and make their courses more appealing to a wider range of students.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for insightful comments and suggestions to improve this paper. This work was supported by the National Science Foundation under Grant No. 2044384 for the project, "CODE:SWITCH: Integrating Content and Skills from the Humanities into Data Science Education."

REFERENCES

Anderson, M. J. (2015). *The American census: A social history*. Yale University Press.

Anderson, T. D., & Parker, N. (2019). Keeping the human in the data scientist: Shaping human-centered data science education. *Proceedings of the Association for Information Science and Technology*, *56*(1), 601–603. https://doi.org/10.1002/pra2.103

Aragon, C., Hutto, C., Echenique, A., Fiore-Gartland, B., Huang, Y., Kim, J., Neff, G., Xing, W., & Bayer, J. (2016). Developing a research agenda for human-centered data science. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion - CSCW '16 Companion*, 529–535. https://doi.org/10.1145/2818052.2855518

Association for Computing Machinery. (2021). *Computing competencies for undergraduate data science curricula*. Association for Computing Machinery.

Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, *13*(4), 602–606. https://doi.org/10.1187/cbe.14-06-0099

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Biehler, R., De Veaux, R., Engel, J., & Kazak, S. (2020, August 5). Call for papers: Research on data science education: Special Issue of the Statistics Education Research Journal (SERJ). *IASC News*. https://iasc-isi.org/2020/08/05/call-for-papers-research-on-data-science-education-special-issue-of-the-statistics-education-research-journal-serj/

Boaler, J. (2008). Promoting "relational equity" and high mathematics achievement through an innovative mixed-ability approach. *British Educational Research Journal 34*(2), 167–194.

Boyer Commission on Educating Undergraduates in the Research University. (1998). *Reinventing undergraduate education: A blueprint for America's research universities*. SUNY Stony Brook. https://eric.ed.gov/?id=ED424840

Brickhouse, N. W., Lowery, P., & Schultz, K. (2000). What kind of a girl does science? The construction of school science identities. *Journal of Research in Science Teaching*, *37*(5), 441–458.

Brownell, S. E., Kloser, M. J., Fukami, T., & Shavelson, R. (2012). Undergraduate biology lab courses: Comparing the impact of traditionally based "cookbook" and authentic research-based courses on student lab experiences. *Journal of College Science Teaching*, *41*(4), 36–45.

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher*, *70*(1), 30–35.

Calabrese Barton, A., Kang, H., Tan, E., O'Neill, T. B., Bautista-Guerra, J., & Brecklin, C. (2013). Crafting a future in science: Tracing middle school girls' identity work over time and space. *American Educational Research Journal*, *50*(1), 37–75. https://doi.org/10.3102/0002831212458142

Carlone, H. B., & Johnson, A. (2007). Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching*, *44*(8), 1187–1218. https://doi.org/10.1002/tea.20237

Carlone, H. B., Scott, C. M., & Lowder, C. (2014). Becoming (less) scientific: A longitudinal study of students' identity work from elementary to middle school science. *Journal of Research in Science Teaching*, *51*(7), 836–869. https://doi.org/10.1002/tea.21150

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016*. https://commons.erau.edu/publication/1083

Çetinkaya-Rundel, M., & Ellison, V. (2021). A fresh look at introductory data science. *Journal of Statistics and Data Science Education*, *29*(S1), S16–S26. https://doi.org/10.1080/10691898.2020.1804497

Chakrabarty, D. (2018). Anthropocene time. *History and Theory*, *57*(1), 5–32. https://doi.org/10.1111/hith.12044

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE Life Sciences Education*, *14*(1), 1–13. https://doi.org/10.1187/cbe.14-10-0167

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., … Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*(1), 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Desai, K. V., Gatson, S. N., Stiles, T. W., Stewart, R. H., Laine, G. A., & Quick, C. M. (2008). Integrating research and education at research-extensive universities with research-intensive communities. *Advances in Physiology Education*, *32*(2), 136–141. https://doi.org/10.1152/advan.90112.2008

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Driver, R., Learch, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Open University Press.

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

Espinosa, L. L. (2011). Pipelines and pathways: Women of color in undergraduate STEM majors and the college experiences that contribute to persistence. *Harvard Educational Review*, *81*(2), 209-240, 388.

Franks, B. (2020). *97 things about ethics everyone in data should know*. O'Reilly Media.

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

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, *78*(2), 297–315. https://doi.org/10.1111/j.1751-5823.2010.00117.x

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Greenwald, N. L. (2000). Learning from problems. *Science Teacher*, *67*(4), 28–32.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., & Ward, M. D. (2015). Data Science in statistics curricula: Preparing students to "Think with Data." *The American Statistician*, *69*(4), 343–353. https://doi.org/10.1080/00031305.2015.1077729

Hardin, J., Norton, N. J., Nolan, D., & Temple Lang, D. (2021). Computing in the statistics curricula: A 10-year retrospective. *Journal of Statistics Education*, *29*(Sup 1), S4–S6. https://doi.org/10.1080/10691898.2020.1862609

Harrison, M., Dunbar, D., Ratmansky, L., Boyd, K., & Lopatto, D. (2011). Classroom-based science research at the introductory level: Changes in career choices and attitude. *CBE Life Sciences Education*, *10*(3), 279–286. https://doi.org/10.1187/cbe.10-12-0151

Hernandez, P. R., Schultz, P. W., Estrada, M., Woodcock, A., & Chance, R. C. (2013). Sustaining optimal motivation: A longitudinal analysis of interventions to broaden participation of underrepresented students in STEM. *Journal of Educational Psychology*, *105*(1). https://doi.org/10.1037/a0029691

Hurtado, S., Cabrera, N. L., Lin, M. H., Arellano, L., & Espinosa, L. L. (2009). Diversifying science: Underrepresented student experiences in structured research programs. *Research in Higher Education*, *50*(2), 189–214. https://doi.org/10.1007/s11162-008-9114-7

Kang, H., Barton, A. C., Tan, E., Simpkins, S. D., Rhee, H., & Turner, C. (2019). How do middle school girls of color develop STEM identities? Middle school girls' participation in science activities and identification with STEM careers. *Science Education*, *103*(2), 418–439. https://doi.org/10.1002/sce.21492

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, *3*(1), Article 11. https://doi.org/10.1186/s40594-016-0046-z

Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, *76*(4), 761–780. https://doi.org/10.1348/000709905X59961

Laursen, S., Hunter, A.-B., Seymour, E., Thiry, H., & Melton, G. (2010). *Undergraduate research in the sciences: Engaging students in real science*. Wiley.

Lepore, J. (2020). *If then: How the simulmatics corporation invented the future* (First edition). Liveright Publishing.

Lopatto, D. (2004). Survey of undergraduate research experiences (SURE): First findings. *Cell Biology Education*, *3*(4), 270–277. https://doi.org/10.1187/cbe.04-07-0045

Lopatto, D. (2009). *Science in solution: The impact of undergraduate research on student learning*. Research Corporation for Science Advancement.

Lue, R. (2019). Data science as a foundation for inclusive learning. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.c9267215

Nagda, B. A., Gregerman, S. R., Jonides, J., von Hippel, W., & Lerner, J. S. (1998). Undergraduate student-faculty research partnerships affect student retention. *The Review of Higher Education*, *22*(1), 55–72.

National Academies of Sciences, Engineering, and Medicine. (2018a). *Data science for undergraduates: Opportunities and options*. National Academies Press. https://doi.org/10.17226/25104

National Academies of Sciences, Engineering, and Medicine. (2018b). *The Integration of the humanities and arts with sciences, engineering, and medicine in higher education: Branches from the same tree*. National Academies Press. https://doi.org/10.17226/24988

National Research Council, Nielsen, N. R., Schweingruber, H. A., & Singer, S. R. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. National Academies Press. https://doi.org/10.17226/13362

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Ograjenšek, I., & Gal, I. (2016). Enhancing statistics education by including qualitative research. *International Statistical Review*, *84*(2), 165–178. https://doi.org/10.1111/insr.12158
O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.
Paushter, M. K. (2017). *Desegregation in an era of resegregation: How heterogeneous secondary science classes increase student achievement and entrance into the STEM pipeline*. [Doctoral dissertation, Johns Hopkins University]. http://jhir.library.jhu.edu/handle/1774.2/44721
Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93*(3), 223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x
Purzer, Ş., Strobel, J., & Cardella, M. E. (2014). *Engineering in pre-college settings: Synthesizing research, policy, and practices*. Purdue University Press. https://doi.org/10.2307/j.ctt6wq7bh
Reeves, T. C., & Laffey, J. M. (1999). Design, assessment, and evaluation of a problem-based learning environment in undergraduate engineering. *Higher Education Research & Development*, *18*(2), 219–232. https://doi.org/10.1080/0729436990180205
Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3), 528–549. https://doi.org/10.1111/insr.12110
Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., & Dolan, E. L. (2016). Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE Life Sciences Education*, *15*(2). https://doi.org/10.1187/cbe.16-03-0117
Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences. *Science*, *316*(5824), 548–549. https://doi.org/10.1126/science.1140384
Saltz, J. S., Dewar, N. I., & Heckman, R. (2018). Key concepts for a data science ethics curriculum. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 952–957. https://doi.org/10.1145/3159450.3159483
Sanders, M. (2009). STEM, STEM Education, STEMmania. *The Technology Teacher*, *68*(4), 20–26.
Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology*, *9*(6), 523–538. https://doi.org/10.1002/acp.2350090605
Seymour, E., Hunter, A.-B., Laursen, S. L., & DeAntoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education*, *88*(4), 493–534. https://doi.org/10.1002/sce.10131
Sheahan, B. H., & White, J. A. (1990). Quo vadis, undergraduate engineering education? *Engineering Education*, *80*(8), 1017–1022.
Sjøberg, S. (2002). Science and technology education current challenges and possible solutions. *Connect* (UNESCO International Science, Technology & Environmental Education News), *27*.
Slavin, R. E. (1989). Research on cooperative learning: Consensus and controversy. *Educational Leadership*, *47*(4), 52–54.
Steele, J. M., Walberg, H. J., & House, E. R. (1974). Subject areas and cognitive press. *Journal of Educational Psychology*, *66*(3), 363–366. http://dx.doi.org.colorado.idm.oclc.org/10.1037/h0036503
Stohlmann, M., Moore, T., & Roehrig, G. (2012). Considerations for teaching integrated STEM education. *Journal of Pre-College Engineering Education Research*, *2*(1), 28–34. https://doi.org/10.5703/1288284314653
Tang, R., & Sae-Lim, W. (2016). Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, *32*(3), 269–290. https://doi.org/10.3233/EFI-160977
Thiry, H., & Laursen, S. L. (2011). The role of student-advisor interactions in apprenticing undergraduate researchers into a scientific community of practice. *Journal of Science Education and Technology*, *20*(6), 771–784. https://doi.org/10.1007/s10956-010-9271-2
Thiry, H., Weston, T. J., Laursen, S. L., & Hunter, A. B. (2012). The benefits of multi-year research experiences: Differences in novice and experienced students' reported gains from undergraduate research. *CBE Life Sciences Education*, *11*(3), 260–272. https://doi.org/10.1187/cbe.11-11-0098
Tobias, S. (1993). Why poets just don't get it in the physics classroom: Stalking the second tier in the sciences. *NACADA Journal*, *13*(2), 42–44. https://doi.org/10.12930/0271-9517-13.2.42

Udovic, D., Morris, D., Dickman, A., Postlethwait, J., & Wetherwax, P. (2002). Workshop biology: Demonstrating the effectiveness of active learning in an introductory biology course. *BioScience*, *52*(3), 272–281. https://doi.org/10.1641/0006-3568(2002)052[0272:WBDTEO]2.0.CO;2

Valenti, S. S., Masnick, A. M., Cox, B. D., & Osman, C. J. (2016). Adolescents' and emerging adults' implicit attitudes about STEM careers: "Science is not creative." *Science Education International*, *27*(1), 40–58.

Vance, E. A. (2021) Using team-based learning to teach data science. *Journal of Statistics and Data Science Education*, *29(3)*, 277-296. https://doi.org/10.1080/26939169.2021.1971587

Vance, E. A., & Smith, H. S. (2019). The ASCCR Frame for learning essential collaboration skills. *Journal of Statistics Education*, *27*(3), 265–274. https://doi.org/10.1080/10691898.2019.1687370

Vance, E. A., Alzen, J. L., & Smith, H. S. (2022). Creating shared understanding in statistics and data Science Collaborations. *Journal of Statistics and Data Science Education*, *30*(1), 54–64. https://doi.org/10.1080/26939169.2022.2035286

Vieyra, M., Gilmore, J., & Timmerman, B. (2011). Requiring research may improve retention in STEM fields for underrepresented women. *Council on Undergraduate Research Quarterly*, *32*(1), 13–20.

Villarejo, M., Barlow, A. E. L., Kogan, D., Veazey, B. D., & Sweeney, J. K. (2008). Encouraging minority undergraduates to choose science careers: Career paths survey results. *CBE Life Sciences Education*, *7*(4), 394–409. https://doi.org/10.1187/cbe.08-04-0018

Weaver, G. C., Russell, C. B., & Wink, D. J. (2008). Inquiry-based and research-based laboratory pedagogies in undergraduate science. *Nature Chemical Biology; Cambridge*, *4*(10), 577–580. http://dx.doi.org.colorado.idm.oclc.org/10.1038/nchembio1008-577

Wei, C. A., & Woodin, T. (2011). Undergraduate research experiences in biology: Alternatives to the apprenticeship model. *CBE Life Sciences Education*, *10*(2), 123–131. https://doi.org/10.1187/cbe.11-03-0028

Wickham, H., & Grolemund, G. (2017) *R for data science*. O'Reilly Media. https://r4ds.had.co.nz

Wood, W. B. (2003). Inquiry-based undergraduate teaching in the life sciences at large research universities: A perspective on the Boyer Commission report. *Cell Biology Education*, *2*, 112–116. https://doi.org/10.1187/cbe.03-02-0004

Wu, D., Lv, S., & Xu, H. (2020). An analysis on competency of human-centered data science employment. *Proceedings of the Association for Information Science and Technology*, *57*(1). https://doi.org/10.1002/pra2.219

Zeidler, D. L. (2016). STEM education: A deficit framework for the twenty first century? A sociocultural socioscientific response. *Cultural Studies of Science Education*, *11*(1), 11–26. https://doi.org/10.1007/s11422-014-9578-z

Zuboff, S. (2020). *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (First Trade Paperback Edition). Profile Books.

ERIC A. VANCE
1111 Engineering Drive, University of Colorado Boulder,
Boulder, CO 80309-0526, USA

APPENDIX A:

IIIDS SAMPLE SYLLABUS

COURSE DESCRIPTION

This course will teach key data science skills and concepts to students who have no experience in statistics or computing, providing them a pathway into more in-depth work with data in any discipline. At the same time, this course will teach students to apply the central humanities skills of source critique, attention to human motives, and contextualization—to understand that all data is created by humans for humans, that human objectives and cognitive biases shape data analysis and use, and that data can be used (and misused) with enormous consequences.

This course is organized around five modules, with each module combining the learning and application of technical data science skills with research questions drawn from the humanities. Each module works with one or more data sets and builds in readings and activities designed to teach students to make and evaluate data-based claims about specific, real research questions from different humanities and human-oriented disciplines.

COURSE OBJECTIVES

The central goals of the course are to improve students' understanding of data and quantitative reasoning and to develop students' ability to apply that understanding to real research questions. It aims to give all students (including those with no prior statistical or programming experience) practice in putting data to work in the world in responsible, informed, and ethical ways.

Students will learn how to:

- apply basic statistical reasoning and methods,
- evaluate claims that use data as evidence (approaching data with healthy skepticism),
- make claims that use data as evidence (putting forth a sound, data-driven argument),
- make and evaluate claims utilizing humanist modes of inquiry,
- augment study in humanistic fields with the help of data science, and
- assess the limits of data science approaches and the ethical implications of utilizing big data.

STRUCTURAL INFORMATION

Course Name and Number: AHUM 1825
Instructors: This course is team-taught by Profs. Vance and Glimp
Time: Tuesdays and Thursdays, 2:20–3:35PM (15 weeks)
Lab recitation sections: Wednesdays 1:50–2:40PM, 3:00–3:50PM, or 4:10–5:00PM
Prerequisites: None. Must have undergraduate standing
Enrollment size: 66 enrollment cap for Year 1, 144 in Year 2, and 200 in Year 3
Credits: 4 credits
Graduation requirements satisfied: Either (but not both) of

- General Education Arts and Humanities divisional distribution requirement
- Quantitative Reasoning skills requirement in the College of Arts and Sciences

BASIC TOOLS AND PROCEDURES

We will do our data wrangling, statistical analysis, and basic visualizations using the R programming language, using RStudio for data storage and teamwork.

Throughout the semester, students will work in teams. Teams are responsible for mini-projects concluding modules 1, 2 and 4, as well as a somewhat larger midterm project in module 3 and a final project where teams have the opportunity to either draw on provided data sets or find or develop their own (module 5 is devoted to creating the final project.)

EVALUATION
Grades will be assessed based on the following distribution:

15% Weekly brief coding and statistical concept quizzes (1.5 pts. each x 10 quizzes)
15% Brief analytical essays (3 pts. each x 5)
30% Module assignments (for modules 1, 2 and 4) (10 pts. each x 3)
20% Midterm project (module 3)—analysis and report
20% Final project (module 5)—analysis, report, and presentation
100% TOTAL

INITIAL PROPOSED SCHEDULE **(modified almost completely in practice!)**
MODULE 1: This first unit introduces students to the course's basic aims and basic tools. Its focus will be on asking good questions, understanding the possibilities and limitations of data sets, and learning the basics of coding and RStudio.

Data: A large data set of hip hop lyrics, developed in conjunction with CU Boulder's RAP Lab (Laboratory for Race and Popular Culture).

> **Week 1:** What is data science? What kinds of questions do data scientists ask? What kinds of questions do humanists ask? Where are the points of overlap? What is the difference between a fact and a claim?
>
> **Week 2:** An introduction to coding with RStudio. A first data set: hip hop lyrics across time and space. How was this data generated? How is a data set composed? What is included and what is excluded? Who made these decisions?
>
> **Week 3:** Teamwork, effective feedback. How to ask great questions of our data set. What kinds of questions might one ask about a large collection of hip hop lyrics? What does it mean to "read" a lyric? What kinds of questions do humanists ask? How can data science contribute to this kind of inquiry? Small groups of students will **develop an annotated set of questions appropriate to ask of this data set.**

MODULE 2: This unit develops students' basic skills of exploratory data analysis and data visualization. It builds on the work of the previous unit by providing the tools for basic exploratory data analysis of our first data set and by examining different types of texts through close reading and through data analysis.

Data: Hip hop lyrics (from module 1); a corpus of political speeches and political manifestos.

> **Week 4:** Literary analysis and/vs. data visualization. How to read a poem, how to read a speech, and how to read a chart. What does it mean to treat a poem or speech as data? What can examining many instances of a kind of text tell us? What kinds of questions get obscured by such an approach? How can visualizations tell stories, and how can we evaluate the trustworthiness of those stories?
>
> **Week 5:** Exploratory data analysis. What kinds of basic information can we extract from our corpora? How do hip hop lyrics and political texts address the same issues? What kinds of questions or concerns do not overlap in our two corpora? Do these overlaps or distinctions change over time?
>
> **Week 6:** Creating compelling arguments based on data visualizations. Extending the lines of inquiry developed in week three, small groups of students will **compile reports addressing one question as it applies to both data sets**.

MODULE 3: This unit trains students in the basic concepts of statistical analysis and underlines the uncertainties present in all data, including tabular data. It also raises questions about the data that states gather of their citizens.

Data: The full-count 1920 U.S. census and the full-count 1940 U.S. census (different student teams will be assigned subsets representing different states.)

Week 7: Basic statistical analysis: summary statistics, means, probabilities, correlations, quantifying uncertainty. What do these statistics tell us about people's lives in the early 20th century? What do we need to watch out for, given the limitations of our data?

Week 8: Understanding the provenance of data. How and to what ends do states gather data about their citizens? What are the politics of such data gathering? How does the rhetoric of statisticians and scientists differ from the rhetoric of politicians and the general public? How certain can we be of the accuracy of a data set like this? How are those data generated? History of the census. Students will read historical instructions to census enumerators as well as excerpts from historical scholarship about the census.

Week 9: Gauging the possibilities of one's data set. What questions can one answer by drawing on different combinations of variables? How much information can one glean from tabular data about the lives of real people? How do we contextualize analyses of census data using other historical sources and scholarship? How can we use spatial analysis to reveal patterns in our data?

MIDTERM PROJECT: Students will develop sample analyses combining several variables for their state, including maps (using R) that show the spatial distribution of one or more variables in the data.

MODULE 4: This module introduces more complex statistical analyses and concepts such as multivariate thinking and linear regression models. This module will focus on the depiction of nature and technology in a large corpus of science fiction novels from the twentieth century and early twenty-first century. How have environmental concerns been represented in science fiction? How have such representations changed across the century? How can we account for multiple variables in a data set simultaneously?

Data: A large corpus of science fiction novels from the twentieth and early twenty-first centuries and a corpus of news reports on climate change.

Week 10: Measures of uncertainty, calculating percentiles and z-scores. Principles of inference and statistical thinking. Students will read short pieces of science fiction from the beginning of the twentieth century and from the last decade. How does our reading of individual texts compare to our reading of a large data set?

Week 11: Data wrangling/creating and debunking arguments using statistical analysis; assessing the strength of inferences from our data. How can data science approaches help us understand how works of science fiction change over time? How do representations of technology change? And how does technology relate to ideas about categories of identity, such as race, nationality, or gender? What can we as readers see that computer algorithms cannot?

Week 12: What is the anthropocene? Can we use the techniques learned over the course of the semester to trace the emergence of concerns about human impact on the environment in our corpus of science fiction novels? And how do fictional accounts of global climate change compare to journalistic efforts to understand the phenomenon? **Students will generate and test a set of hypotheses about how these two corpora relate**.

MODULE 5: This unit consolidates student learning about data, its provenance, and its analysis through a team project of the students' choice. The project may draw on any of the data sets used in the course or a new data set of the students' choosing. It will combine different types of analysis and require students to formulate a question and an argument using data.

> **Week 13**: Setting up a good research question and matching the question with one's data. Using research in scholarly sources to contextualize one's data. Students will work in their teams to choose a data set, formulate a research question, and use library resources to deepen their understanding of their question.
>
> **Week 14**: Analyzing the data, checking one's work, and critiquing the work of others. Working in teams, students will use what they have learned over the course of the semester to formulate and carry out the analyses that will best answer the research question they have formulated. Student teams will review and critique other teams' preliminary analyses.
>
> **Week 15**: Presenting a data-driven argument. How are the findings best communicated? What concepts need to be explained as part of the presentation? What context must the audience understand? What visualizations best tell the story? What conclusions and recommendations are justified by the analyses? Students will work in teams to create and polish their final presentations.

Final project presentations in a public forum, critiqued by fellow students and invited participants.