Data Science for High School Computer Science Workshop:
Identifying Needs, Gaps, and Resources

National Science Foundation Workshop Report
January 16-17, 2020

Editors
Rene Baston | Catherine B. Cramer | William Leon | Katie Naum | Stephen Miles Uzzo

Workshop Webpage
http://nebigdatahub.org/data-science-for-high-school-computer-science-workshop/

Table of Contents

Authors

Ajay Anand, University of Rochester
Rene Baston, Northeast Big Data Innovation Hub
Dorothy Bennett, New York Hall of Science
Kirk Borne, Booz Allen Hamilton
Any Busey, Education Development Center
Ian Castro, University of California Berkeley
Catherine Cramer, Woods Hole Institute
Yadana Desmond, STEM Teachers NYC
Chad Dorsey, Concord Consortium
Ana Echeverri, International Business Machines
Susan Ettenheim, NYC Department of Education, Eleanor Roosevelt High School
Melissa Floca, University of California, San Diego
Daniel Fuka, Virginia Institute of Technology
Crystal Furman, College Board
Matt Gee, University of Chicago
Michele Gilman, University of Baltimore
Narine Hall, Champlain College
Nick Horton, Amherst College
Shriram Krishnamurthi, Brown University
Victor Lee, Stanford University
William Leon, Cornell Tech
Diane Levitt, Cornell Tech
Meredith Mante, International Business Machines
Joe Melendez, Cornell Tech
Katie Naum, Northeast Big Data Innovation Hub
Tom O'Connell, Mouse
Stephanie Ogden, College Board
Aankit Patel, City University of New York
Kelly Powers, Cornell Tech
Hari Raghavan, International Business Machines
Meg Ray, Cornell Tech
Jen Rosato, College of St. Scholastica
Andee Rubin, TERC
Emmanuel Schanzer, Bootstrap
Lisa Singh, Georgetown University
Julia Stoyanovich, New York University
Rochelle Tractenberg, Georgetown University
Stephen Uzzo, New York Hall of Science
Sara Vogel, City University of New York
Michelle Wilkerson, University of California, Berkeley
Elena Yulaeva, University of California, San Diego

We owe our deepest gratitude to all participants for their valuable intellectual contributions throughout the Workshop. All Workshop participants were active authors for this report.

Workshop Organizers

Rene Baston, Northeast Big Data Innovation Hub, Columbia University
Catherine B. Cramer, Woods Hole Institute
Katie Naum, Northeast Big Data Innovation Hub, Columbia University
Laycca Umer, New York Hall of Science
Stephen Miles Uzzo, New York Hall of Science

Acknowledgements

The organizers would like to thank Jan Cuny (retired) at the National Science Foundation for inspiring and supporting this workshop and report, and Chaitanya K. Baru, Senior Science Advisor, Office of Integrative Activities at NSF, for inspiring the creation of Data Science for All, and for his support throughout the planning process for this workshop. In addition, many thanks to Diane Levitt, Sr. Director of K-12 Education at Cornell Tech, for hosting us at Cornell Tech.

This effort is supported by the National Science Foundation under Award Number 1922898 to the New York Hall of Science. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Executive Summary

This report summarizes a workshop in which a group of 41 data science experts, practitioners, educators and researchers gathered at Cornell Tech on Roosevelt Island in New York City for two days to identify successes, challenges, and solutions to improving the use of data science in education settings - in general for formal learning settings, and specifically for high school. Through a process of brainstorming in a variety of formats (talks, panels, breakouts, demonstrations and group discussion) and then articulating concrete ideas, the group determined pathways to bringing data science into high school, along with a number of open questions. (See biographies of workshop participants in Appendix A.)

The workshop included four elements: 1) identifying the gaps in data science education within computer science; 2) a series of presentations from tool and software developers and learning researchers to describe validated practices; 3) collaborative brainstorming sessions to draft a fundamental set of current resources, challenges and solutions; and 4) authoring of proposal ideas to articulate the path forward. (See Workshop Agenda in Appendix B.)

Principal Findings

In the aggregate the group determined that a grasp of data science is vital for people to be fully prepared for the world of work, as well as to be an enlightened twenty-first century member of society. The group was unable to satisfactorily differentiate the roles that data literacy and data science play in education, and there was not consensus on how these terms are defined. However, there was a tendency to circumscribe data literacy as having more to do with ethics, security and privacy (i.e. what is done with personal data and how it affects human wellbeing), with some participants adding the ability to be empowered by data (i.e. knowing how to use data to better understand their world). Data science, on the other hand, was more clearly defined as being able to use tools and skills to gather, process and analyze data for a variety of purposes, including using data for advanced applications such as artificial intelligence.

The question of what the role of data science or literacy is in computer science pedagogy was also not fully resolved. Significantly, the group agreed that data literacy or data science in some form needs to have a bigger role across disciplines in teaching and learning practice - where it fits the need and enhances learning of those topics - but whether or how data science fits directly into computer science courses in high school remained unclear. It was also unclear what the value of a stand-alone course in data science would be. Some indicated that because computer science classes in high school are typically elective, and that poorly resourced schools typically do not even have a computer science teacher, a course in data science would similarly contribute to inequities in high school education.

Addressing the needs of teachers in the further development of applications of data science at the high school level is key to making headway. This includes bringing teachers fully into the development process and providing ongoing support and community building. Additionally, addressing issues of equity and accessibility emerged as being a top priority as these frameworks are built out. The participants felt strongly that we are at a pivotal moment in narrowing the data

divide by providing data science educational assets and support throughout public school systems nationally.

Principal Recommendations

At this moment there are many tools, resources and approaches to data science education available. However, they are not part of a cohesive whole, nor are they entirely inclusive. The findings from this workshop suggest that a true coordination of effort is needed, as well as a thorough analysis of the learning ecosystem in which data science fits. This means knowing which programs and resources have already been developed and tested, as well as the programs and outcomes that appear in prior grades and clarifying what is happening at the undergraduate level.

To be equitable, opportunities for data science education must be available to all students, requiring curriculum and resource development that is inclusive of stakeholders; tools that are co-designed with teachers; and ongoing training and support for teachers. Ensuring equity and accessibility points to not offering data science exclusively through computer science but rather across the high school curriculum.

This will likely require:

- Development of a consensus high school data science framework that can align to curricula, standards, and scope and sequence, and be adopted by policymakers and made available and supported at the federal, state and local levels, and that provides clear pathways into undergraduate level data science education.
- Close work with teachers to identify where and how data science and data literacy fit, including intiating participatory design practices directly involving educators, curriculum developers with tool developers, data scientists, and learning scientists in the co-creation of accessible data science curriculum, tools and resources for the classroom.
- Development of a new model to knit together existing curricula, tools, programs, games, and data sources into a unified whole.

The kind of concurrent top-down and bottom-up approach we are suggesting will be required in order to accelerate convergence of data science and teaching and learning and for it to have inertia, scalability and sustainability.

1. Introduction

While there has been intense interest in bringing computer science to all learners through such initiatives as Computer Science for All, the ability to solve complex real world problems through the application of computational skills is through data science. Ultimately, data derived from the real world are the fodder of computation, and skills and understanding of data science are essential to the preponderance of advanced computational processes such as artificial intelligence and deep learning. Yet data literacy and skills are largely absent from educational systems. And while there have been attempts to introduce data science into computer science curricula, these attempts have lacked access to the knowledge, resources, training and usable tools and data sources needed to scale data science along with computer science.

To address this gap, the New York Hall of Science (NYSCI), in collaboration with the Northeast Big Data Innovation Hub (NEBDIH) at Columbia University, developed *Data Science For All* (DS4All), a data literacy initiative begun in 2015 to bring together expertise from data science and learning and teaching practice, with the goal of discovery, capacity-building, and filling the knowledge gap in data literacy for all learners, particularly those in underserved communities. NEBDIH and NYSCI continue to work with the data literacy community to facilitate the development of activities to address knowledge gaps at all levels of society, from citizens at large in informal learning settings (e.g., museums, libraries, and day-to-day living), to teachers and students in formal settings (PreK-20), to career development and executive education (e.g. professional development in corporate, nonprofit and government sectors). DS4All is building a community of stakeholders to advise on bridging big data practice with learning, education and career readiness for communities of all kinds with the goal of improving equity, particularly for communities of need.

Previously, as part of the DS4All initiative, and also through NSF support (Award 1636736), NYSCI and NEBDIH brought together data science domain experts from partner institutions and elsewhere in a process of collaborative inquiry called *Building Capacity for Regional Collaboration in Closing the Big Data Divide* in order to: a) focus data scientists around learning, b) identify the nature and quality of extant data literacy resources, and c) to look at the kinds of strategies that could advance data literacy for lifelong learners. This process helped validate the need for and to identify an emerging community of practice around data literacy for informal learners. It also highlighted the need to better characterize the gap between the current resources and the needs of learners across all settings. Findings from this gathering and other related activities can be found in the DS4All white paper, issued in July 2019 (go to https://www.woodsholeinstitute.org/ and click on button to download DS4All White Paper.)

In order to address specific needs that have emerged in bringing data science and data literacy to formal education settings, and specifically in high school classrooms, NYSCI and NEBDIH worked with NSF to organize the workshop described herein, in order to bring learning research experts, data science domain experts, and tool developers together with education and private sector stakeholders to explore a set of priority implementation strategies and articulate a pathway toward data literacy at the high school level. The *Data Science for High School Computer Science Workshop: Identifying Needs, Gaps, and Resources* (NSF Award No. 1922898) was a visioning and capacity-building gathering inspired by NSF's Computer Science for All (CS4All) program, and in response to previous work integrating data science into high school computer science classrooms. It emphasized the unique and genuinely new dimensions of learning afforded by data science and how they create new opportunities for applying computational thinking, programming and habits of mind to new problems, learning and insights in STEM domains. While we acknowledge the need exists throughout PreK-20 education and lifelong learning, this latest effort focused explicitly on the needs of high school students and teachers, as they are on the front lines of the rapidly changing workforce. This report summarizes the outcomes of this workshop and the activities are chronicled in detail in the appendices.

The New York Hall of Science (NYSCI) is New York City's hands-on science center and a learning lab, with a dedicated team of learning and cognitive scientists, designers and developers testing and studying innovative approaches to supporting informal, community-based approaches to STEM learning. NYSCI is a global education leader in data-driven science education and community engagement with underserved populations. The Northeast Big Data Innovation Hub serves to establish a diverse, multi-sector data science community in the northeastern United States, as well as across the nation. It has built over 90 partnerships, bringing together data science leaders and practitioners at academic, industry, government, and nonprofit organizations of all kinds, to share resources, insights, and knowledge about harnessing data to address society's most challenging problems. It is one of four Big Data Hubs sponsored by NSF in support of Harnessing the Data Revolution (HDR), one of NSF's "10 Big Ideas". The New York Hall of Science managed the workshop planning in collaboration with the Northeast Big Data Innovation Hub and consultancy by Catherine Cramer from the Woods Hole Institute. Cornell Tech hosted the meeting on their campus and provided support services for audio, video and teleconferencing.

2. Workshop Organization and Reporting Structure

The organizers invited 41 leading experts in data science education along with stakeholders from industry, K-20 curriculum developers, instructors, and specialists in learning sciences and informal learning for a 2-day in-person workshop, which was also accessible via video conference for invitees who could not attend in person to be able to participate remotely. The workshop included four elements: 1) identifying the gaps of data science in computer science; 2) a series of presentations from tool and software developers and learning researchers to describe validated practices; 3) collaborative brainstorming sessions to draft a fundamental set of current resources, challenges and solutions; and 4) authoring of proposal ideas to articulate the path forward.

The workshop took place at the Bloomberg Center, Cornell Tech, on Roosevelt Island in New York City on January 16 and 17, 2020. During the workshop participants explored and defined challenges and strategies for bringing data science and data literacy into high school computer science classrooms specifically, as well as more broadly across the high school curriculum. The workshop drew out practices and proposal ideas to elicit diverse points of view and address issues in: research and education, enfranchising underrepresented groups, and to identify strategies with the potential to more effectively address knowledge and policy gaps. Workshop participants worked to:

- help characterize and provide an accounting of the kinds of research, resources, data and tools that can be leveraged to improve data science knowledge and teaching capacity at the secondary level in formal education;
- identify the processes and supports needed for teachers of computer science and other academic subjects to readily enrich instruction and curricula with data science tools, data sets and resources;
- directly address the issue of what it means to be a data literate citizen, information worker, researcher, or policymaker;

- Identify and characterize the quality of learning resources and programs intended to improve data literacy to help chart a path forward to bridge data practice with data learning, education and career readiness; and
- ensure that we articulate an actionable, equitable and inclusive path toward data science literacy, with particular emphasis on the formal secondary learning setting.

The workshop produced a set of challenges faced by data science researchers, practitioners, tool developers and educators that could be solvable in the near term through a process of participatory design with classroom teachers.

The morning of Day One began with an orientation by the organizers about Data Science for All, NSF's priorities and the purpose and goals of the workshop. This was followed by a keynote by Michele Gilman from the University of Baltimore on data science ethics and digital justice. A panel discussion completed the morning and was designed to contextualize the role of data science in high school teaching and learning, discussing efforts to integrate data science into instruction, where data science "fits", why efforts succeed or fail, and how to define data science and data literacy. The afternoon consisted of breakout groups and report outs to highlight programs, projects and curricula and how well they worked, and articulating the successes and challenges. The day finished with a series of demonstrations of tools developed by workshop participants and designed to be used to introduce data science into K-12 education settings.

Day Two began with a keynote by Kirk Borne from Booz Allen Hamilton in which he discussed the needs and gaps relating to the development of the 21st century digital workforce, as well as general needs for data skills to empower students and the public at large. His talk was followed by a set of thematic breakout and reporting sessions that looked at solutions to challenges in impact and sustainability, curriculum and development, and training and support for educators. In the afternoon, participants were charged with developing proposal ideas for encouraging data science applications in high school classrooms, and reporting out on results.

The background for the workshop is described in Section 3. Workshop methods, approaches and results are described in Section 4. Section 5 describes results from post-workshop surveys. Section 6 provides conclusions and next steps. Appendices contain detailed information from the workshop and pre- and post-workshop activities. Artifacts, transcripts and voice recordings of the entire event were used to validate the accuracy of this report and participants were provided with a period for review of the draft report. Additional input was provided by interested parties that were unable to join the meeting and has been integrated into this report as appropriate.

3. Background

3.1 Challenges the Workshop Intended to Address

Through the previous work with Data Science for All described above, several challenges in data science education in K-12 settings had already emerged, and were intended to be addressed through the workshop:

- How do we narrow and ideally close the data divide, the widening gap in understanding of data and how it is used that is occurring throughout society?
- How do we prepare the 21st century workforce for an intensely digital world?
- There are a plethora of data science tools, curricula, resources and programs. Why do they not scale? What are the barriers to broad implementation? What approaches, data sets, tools and resources, if any, are most appropriate and usable or adaptable for formal high school learning, and motivating and engaging for students and teachers?
- There is wide variability in what students are learning about data and data science, and whether and how that learning is applied.
- Are programs, tools and resources validated, and if so how? What do students learn from them and how do they apply that knowledge?
- Where does data science "fit" in the high school curricula - as a stand-alone course, in computer science, across the curriculum?

3.2 Historic Context

Data science has been a rapidly growing area of interest, particularly with the recent emphasis on artificial intelligence (AI) and the availability of sophisticated tools and analytical techniques. In formal education contexts during the 1980s (Tinker, 2000), microcomputer-based labs emerged as a candidate for sensing and logging data in science class, but most of the software and hardware was either custom made or more generally for professional use, such as LabView. Appliances and companion software from IBM, Harris Scientific and others emerged in the 1990s allowing more accessible logging, visualization, and analysis. Many of these tools and curricula developed around them were intended for use in high school science applications. Of interest is that the use of these tools highlighted a deficit in quantitative skills among students, and while there were efforts to better integrate math and science, by the 2000's little progress was made in improving quantitative skills in high school students to help them make sense of data (Steen, 2001).

Over the past two decades data science has emerged to be a predominant area of interest and concern across society. Along with the need to use data in problem ideation and solutioning through facility with data science has emerged the need to confront the growing ethical and security issues from the gathering and use of personal data for large scale commercial gain and the rapid rise in the use of artificial intelligence techniques such as machine learning and deep learning to process, federate and make decisions relating to all of this data.

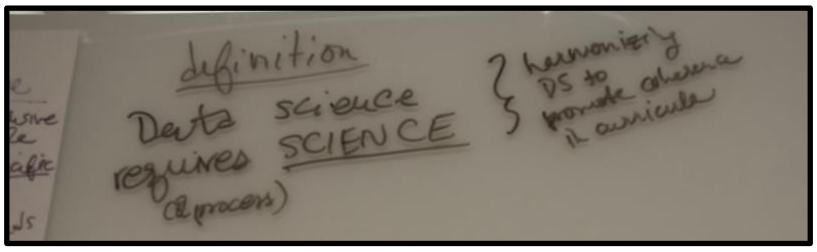
With the use of data and results from data analysis being so ubiquitous, so personalized, and essentially unavoidable, there comes a pressing need to provide citizens with a basic awareness

of the situation they themselves are in, and a deeper understanding of the effect of these technologies on their wellbeing—to gain a rudimentary understanding of what data are potentially being collected about each person, and how they might be being used. Notions of "responsible data science", including transparency, ethics, security, and accountability have become important concepts, not only in understanding the context of data science in society, but in how data science affects the ability for people to thrive, make career choices, and navigate the ethical, personal and community impact of the data-driven economy. Data Science for All recognizes the need to balance the ethical use of data throughout society and the importance of encouraging citizens to be not just consumers but creators in the digital economy and to be active citizens in our technology-driven world. For teaching and learning, this means thinking about data science education across multiple dimensions, not as just a set of skills, but as a deeper transdisciplinary way to characterize the phenomena encountered throughout nature and culture, and the impact on the wellbeing of all people.

3.3 Defining Data Science and Data Literacy

As computing power increased and digital data production grew exponentially, research became increasingly dependent on the use of data - Jim Gray's "Fourth Paradigm" of science (Tansley & Tolle, 2009) - and a new job title emerged: "Data Scientist", claimed to have been invented by Dhanurjay Patil (2008). This evolution would have a tremendous impact on business sectors. However, the concurrent emergence of the field of data science, a defined set of data skills, and the need for understanding how data are used has been equally disruptive in the formal education field. There are still unresolved questions as to where data science should sit at the undergraduate level - in Statistics? Computer Science? Engineering? Many universities have now opened stand-alone data science centers or institutes as a way of dealing with the trans- and interdisciplinary nature of data, and many of these centers offer Masters level degrees in data science.

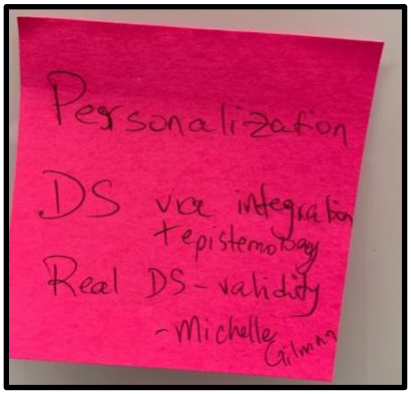
This unsettled question of where data science should "fit" has been translated down to the high school level, where questions remain as to where and how to bring data science into the high school classroom – which is the subject of this workshop. On one level, it is difficult to accurately define data science at the high school level if it isn't fully yet described at the college level, since that means that clear pathways cannot be drawn from one stage to the next. Additional questions follow, again due to the nature of data - should there be stand-alone data science classes (similar to computer science classes)? Should data as a tool for learning be part of every class across the high school curriculum? And what about data literacy - which seems to be a third tier of need - concerning itself with issues related to privacy, safety, trust, and equity? The need for understanding the differences between these three applications of data - as a science, as a tool for learning, and as a literacy - and for understanding where they could be best situated in the high school classroom became a significant challenge revealed during the workshop. In a larger context, this reflects the process of moving education work that has emerged under the Harnessing the Data Revolution



(HDR) program more fully into the Directorate for Education and Human Resources (EHR) to inform the direction of data science education and its role in better preparing the STEM workforce as data and data science pervade all aspects of society.

3.4 A Socio-Technical framing for Effective Transdisciplinary Work

As data science is among the most important areas of transdisciplinary applied research today, it becomes important to develop a theoretical frame for understanding and overcoming systemic challenges to learning integration, whether within or across disciplinary silos. Xie, et al (2017) describe the transdisciplinary nature of geospatial data science focusing on its foundations in mathematics, statistics, and computer science. They indicate that the essential nature of transdisciplinarity in applying these domains to address real problems in data science practice (in particular, hotspot detection, colocation detection, prediction, outlier detection and teleconnection detection), suffers from the siloing of mathematics, statistics, and computer science. We can broaden this to many advanced applications of data science, which depend on a synthesis of mathematics, statistics, and computer science, along with other fields. Pennington, et al (2013) indicate that transdisciplinary practice of any kind requires *transformative learning* (which Mezirow [1997, p. 5], describes as "the process of effecting change in a frame of reference") and express concerns about the need for cross-disciplinary learning in order to effectively build transdisciplinary teams. They describe three phases of development of effective transdisciplinary science: a *disorienting dilemma* (the disorientation of a domain expert engaging with those in an unfamiliar domain of knowledge), *critical reflection* (revision of existing mental models to make sense of these new concepts, methods, data and even terminology), and *reflective discourse* (a process of vetting and synthesizing the revised mental model with team members to create an integrated conceptual framework). In applying transformative learning in education practice, Christie, et al (2015), describe the value it adds to educational practice in that it provides a process for revising mental models and reflective practice, critical thinking, reflective learning, and how that can be applied to novel circumstances. If we think of data science as a transdisciplinary approach and that transdisciplinarity can be facilitated through transformative learning, we have possible pathways to the integration of data science into teaching and learning as not only tractable, but with many benefits to educational practice, professional development, action research, and others. It also provides opportunities for overcoming impediments to solving complex socio-environmental problems in science practice.



Transformative learning requires intensive visioning among stakeholders that values multiple and sometimes conflicting perspectives. When it comes to the products of transdisciplinary work, the use of *participatory design* (PD) creates an intentional collaborative environment that equally values the users, programmers, developers and other stakeholders essential to the effective and innovative use, scaling and sustainability of the tools, curricula, and generalizability of knowledge associated with those tools. Participatory Design uses and integrates ethnographic research methods into technological design to understand the needs of, and directly involve the user

community in the design process. It brings the software, data and interface designers together with the user community to identify and understand design problems (Blomberg, Giacomi, Mosher & Swenton-Wall, 1993). PD originates in cooperative design and action research in the 1960s and 70s (Ehn, Nilsson & Topgaard, 2014) and became a movement in the design of human computer interfaces (HCIs) in the 1980s (Bødker, 1996; Suchman, 1987), resulting in a revolution in user experience design elicited by work at Xerox PARC and Apple Computer, arguably democratizing the computer for a broad spectrum of users (Grudin, 2012; Moggridge, 2006). It is used successfully in socio-technological settings where there is interest to develop effective pathways into uptake and scaling of technological tools particularly for addressing social issues (Blomberg, Giacomi, Mosher & Swenton-Wall, 1993; Bergold & Thomas, 2012; McPherson, 2013; Salingaros, 2011; Ehn, Nilsson and Topgaard, 2014). To be effective in transdisciplinary work, any socio-technical framework developed for data science considers how 1) tool developers affect personalization (localization), 2) thought leaders consider integration and epistemology, and 3) education contextualizes validity.

4. Workshop Methods, Approaches and Results

4.1 Process

To arrive at the goals of the workshop, the organizers deployed a multi-step process that included collaborative group activities intended to draw out individual and group responses as well as general consensus. The intention of the organizers was to use this process in order to arrive at a multi-layered and multi-scale view of the current state and future needs of the data science community, specifically in regards to data use and education and learning opportunities. This process is described in detail below. Figure 5 shows the steps in the process.

Talks and Panel

The purpose of Day One of the workshop was to fully explore bringing data science into the high school classroom and identify what has worked and where the challenges lie. The morning was an opportunity to set the stage with introductions and orientation in the form of brief presentations on NSF's priorities in supporting data science education through Harnessing the Data Revolution (HDR), one of NSF's "10 Big Ideas"; on the history of Data Science for All (DS4All) and the Northeast Big Data Innovation Hub; and an overview of data science and learning.

Keynote Talks

This was followed by the first keynote talk, which was given by Michele Gilman, Venable Professor of Law at the University of Baltimore School of Law, and faculty fellow at Data & Society, where she is researching the intersection of data privacy law and the concerns of low-income communities. Her talk, titled Digital Justice, was presented early in the agenda in order to fully engage participants in the idea of digital equity throughout the workshop. The concept of digital equity, one of the cornerstone concepts of Data Science for All, is borne out of a pressing need to provide citizens with a basic awareness of how data are collected and for them to gain a useful understanding of what data are potentially being collected about each person and how they might be being used. This includes the need for understanding notions of "responsible data science", including transparency, ethics, security, and accountability. Not understanding these concepts

affects the ability for people to have the capacity to thrive, make career choices, and navigate the ethical, personal and community impact of the data-driven economy.

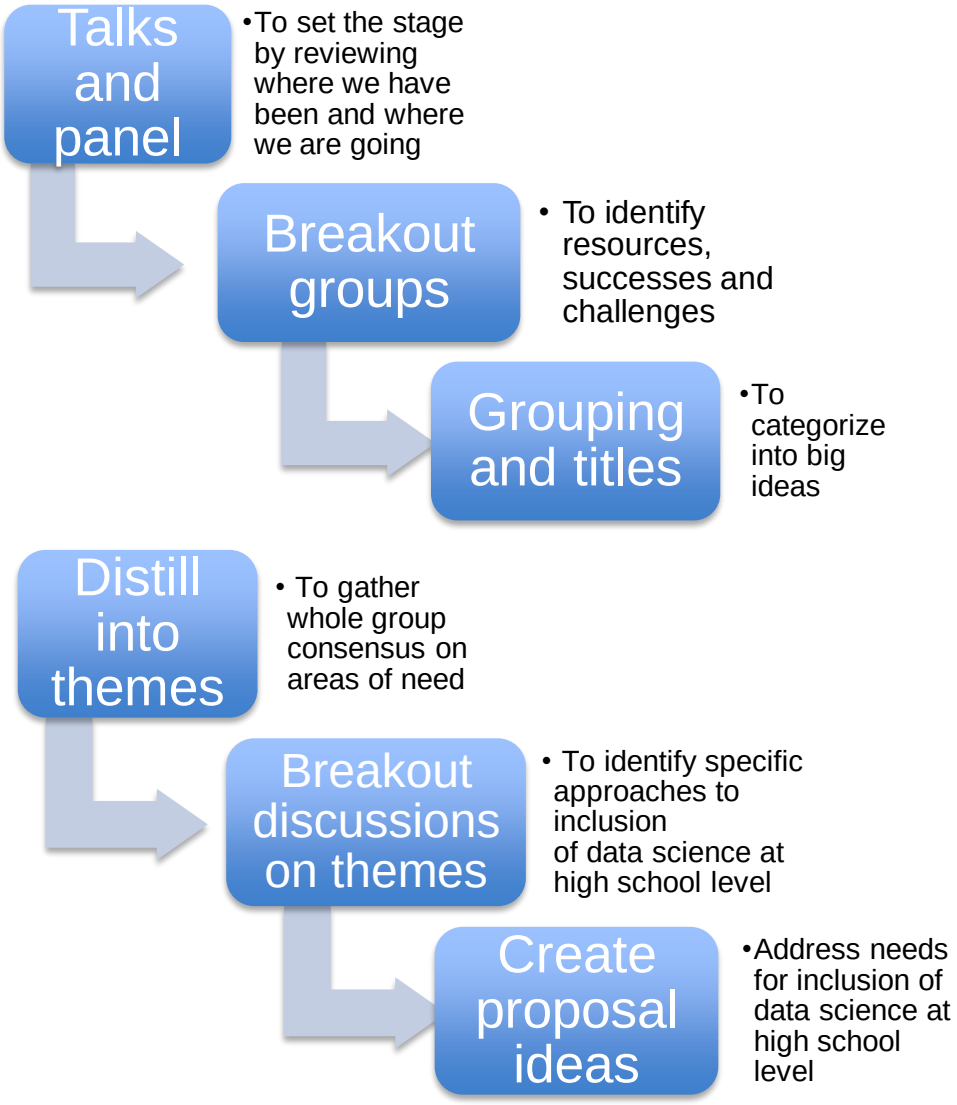


Figure 1: The discrete steps taken to bring the group focus along a trajectory to collaborative consensus on resources, successes and challenges, and development of specific project proposals to address data science education goals.

On the morning of Day Two, Kirk Borne (Booz Allen Hamilton) delivered a talk titled Preparing Students for an Intensely Digital World in which he discussed the importance and imperative for all students to become digital and data literate, for the future of work in general and the future of their own work. He covered five major themes in the presentation: data awareness (what is it?), data relevance (why me?), data literacy (show me how), data science (where's the science?), and data imperative (create and do something with data). (Abstracts of keynotes talks found in Appendix C.)

Panel Discussion

A panel discussion on the morning of Day One included a representative from the College Board, who presented a history of how data science has been included in the AP Computer Science curriculum to date, as well as members of the New York City Computer Science for All (CS4All) team. The panel discussed how the integration of data science into CS4All has progressed, as well as needed next steps and challenges to be addressed.

4.2 Breakout 1: Successes and Challenges

On Day One participants were divided into 4 heterogeneous groups of ~10 people. They were tasked with discussing and explicating successful or less successful attempts to creating or introducing data science teaching modules into the high school classroom, as well as technology, other resources into instruction, and funding/sustainability strategies for these approaches. In addition, the groups brainstormed on what kinds of general challenges have been experienced in introducing data science in high school, regardless of the specific initiatives, projects or resources; which of these challenges were overcome and how and which remain; and lessons learned. At the end of the session each group reported out on their group discussion and conclusions, and a comparison among groups revealed the summary below. The flow chart of the structure of the breakout discussion as well as results from the report outs is below. Detailed notes can be found in Appendix D.

Instructions:
For each of the questions/categories below, please organize your responses using post-it notes as follows:

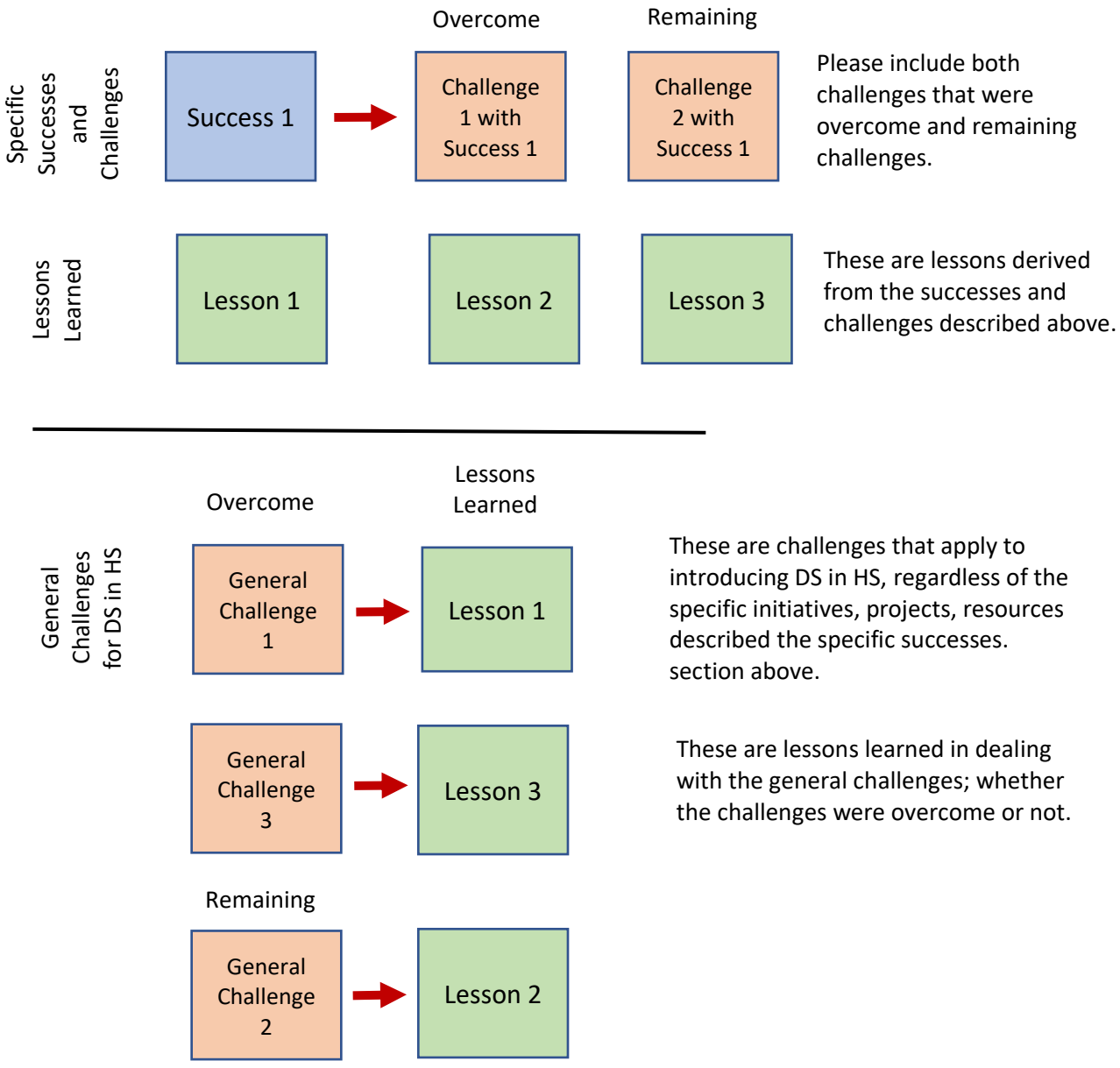


Figure 2: Instructions for Breakout 1

4.2.1 Report Out

What Works

1. Encouraging student engagement
    a. Connection with projects that answer questions students have interest in
    b. Students can choose or gather data meaningful for them
    c. Individualized localized projects
    d. Projects that are culturally responsive
    e. Mentorship by data savvy professionals
    f. Offers job exposure, particularly those jobs that promote STEM Interest
    g. Promotes creativity and fosters critical thinking
    h. Students working together as peer tutors/mentors
    i. Extracurricular programs outside of the classroom
    j. Learning games and other accessible activities
    k. Data science can be a vehicle to spark difficult conversations
2. Professional development for teachers
    a. Recognizes applicability to their discipline
    b. Doable, easy for teachers to offer
    c. Follows established models
    d. Promotes a sense that learning is worth doing
    e. Co-taught courses, bringing together instructors with different kinds of expertise - e.g., computer science and biology teachers

General Challenges

These have been categorized according to key themes that emerged from the discussion.

1. Equity, Accessibility and Ethics
    a. Access to data science teaching and learning, broadly:
        i. Most K12 environments do not have access to decisionmakers and administrators to promote DS within the school system. There are few evangelists within districts.
        ii. Cultural barriers to project development and teaching
        iii. Ensuring teaching approaches account for neurodiversity
        iv. Making instruction accessible to heterogeneous students of varying skill levels. What should all students know after completing the course?
        v. Economic barriers: course fees, materials fees, exam fees
    b. Perceptions of data science and its impact on people
    c. Integrating concepts of Responsible Data Science at a foundational level.
        i. Includes responsiveness, accessibility, inclusivity.
        ii. Which questions are asked of data science, and why?
2. Identifying and leveraging existing resources and identifying gaps
    a. Access to good data sets
        i. This includes the time investment in seeking them out
        ii. Level of complexity, interest, relevance, and technical and non-technical aspects

b. Open data exists, but tools are hard to use or non-existent
c. Validity of data sources, assumptions used, misconceptions
d. Many teachers and students have low proximity to data scientists; currently, involvement in data science heavily relies on already knowing a data scientist.
e. Hardware and internet access
f. Software usable by teachers and students in classrooms, and software maintenance
g. Challenges that transcend technology - just having the tech does not solve your problem if capacity is not there

3. Policy
    a. Better use of systems thinking and approaches to inform development of policies
4. General Challenges to Teaching Data Science
    a. Lack of consistency in definitions of data science and data literacy
        i. There is little consensus on what data science actually is.
        ii. What is the appropriate level of depth for these topics in a high school course: data literacy, data usage, data analysis? How should they each be included in curricula or classtime? Should data science be taught and learned as a separate course, or set of courses, part of computer science or statistics, or integrated throughout the curriculum?
        iii. Sensitivity to epistemology; the computer science bias.
    b. High School course development relative to undergraduate level

        There needs to be much better understanding of what data science looks like at the college level in order to know what it should look like in high school. Data Science at the undergraduate level is still emerging.

    c. Understanding about how data are used, including in Machine Learning and Artificial Intelligence
        i. How to gauge an appropriate level of complexity in data and skills being taught(Tradeoff: depth versus breadth)
        ii. Foundational knowledge: includes data cleaning, probabilistic solutions, understanding of error, variability, messiness
            1. E.g., Complexity of data wrangling (Solution: prebake data initially)
        iii. Background knowledge: What level of programming skill will students need, statistical knowledge? Analytical skills?
        iv. How to most effectively teach an interdisciplinary topic that permeates all domains
        v. Integrated vs. standalone approaches: lack of clarity on how/where DS fits in HS curricula
        vi. How to convey importance and relevance of data science approaches in different contexts
        vii. Ensuring students can write/speak/tell stories about their data science
        viii. What is rationale for approach
    d. Ensuring instructor capacity and adequate professional development
        i. Teachers unfamiliar with the inquiry process

ii. Not just one shot Professional Development but follow up and validation
iii. Improving teaching of critical thinking skills that underlie data science instruction is an ongoing challenge. Learning how to ask good questions
iv. Cognitive load for both teacher and students
v. Helping students come up with interesting and answerable questions
vi. How to communicate trends to kids when they aren't visual
vii. Limited time allotted, limited teacher content knowledge and pedagogical knowledge
1. These are not challenges unique to data science.

e. Designing and delivering appropriate/effective messaging that addresses varying perceptions of data science. How does this inform the design of curricula?

5. Success Metrics/How to measure effectiveness
a. How to define learning outcomes?
b. How do we assess scalability/adaptability of a course?
c. How robust an assessment?
i. Lighter assessments may result in less data but higher completion rates
d. Threading the needle to meet state standards

4.3 Breakout 2: Solutions to Challenges

Breakout 2 was held on Day Two and structured as a response to the challenges described as an output of Breakout 1. Workshop participants chose to divide into self-selected groups along themes that emerged as a result of Breakout 1 and whole-group discussions, and these group topics aligned well as responses to the challenges: Curriculum and Development, Impact and Sustainability, and Training and Support. Condensed group outputs follow. (See Appendix E for original notes.)

4.3.1 Curriculum and Development

The group that self-organized around the topic of Curriculum and Development made use of a structured brainstorming format to generate a list of items that ought to be included in a high school-appropriate data science curriculum. One person served as moderator and another as primary scribe in a Google Doc, with a secondary scribe adding notes and making corrections as needed. Group members raised one finger to be called on by the moderator and make a comment, and two fingers to be allowed to respond to another member's comment.

A raw list was generated via this method (Appendix E), and was subsequently clustered by the group under three headings: Concepts, Practices, and Perspectives, which is a clustering system used by Harvard's ScratchEd framework for computational thinking. Examples of these three dimensions include:

Concepts

- Definition of data

- Impacts of data in society
- Equity and Diversity
- Generation/collection/workflow
- Analysis
- Representation

Practices

- Generating data
- Analyzing data
- Representing data
- Questioning and thinking about our data
- Communicating from and with data

Perspectives

- Understanding data as a social text
- Appreciating data ethics and societal ramifications
- Identifying what data can answer and what it can't
- Appreciating complexity and nuances of data representation and analysis

Many items from the raw list fell under multiple headers; as an example, *reproducibility* is both a concept students must understand and a practice that must be taught. The group further organized their suggestions within the three clusters. An important point that was raised was that data science might be taught differently as appropriate for the course goal - e.g., a course intended to produce data literate citizens will differ significantly from a course intended to produce future data scientists. The group considered explicating the concepts, practices and perspectives from their brainstorming session into three curricular tracks: Data Literacy, Data Fluency, and Data Science. Each would have appropriate levels of rigor depending upon the grade level being taught. A full mapping of concepts/practices/perspectives to tracks and grade levels was not possible during the allotted time. Full details may be found in Appendix E.

4.3.2 Impact and Sustainability

This group took on the challenge of defining and explicating how to create programs, tools and resources that teachers would use; that would not be burdensome to implement (in other words, would not be an add-on to the teachers' workload); would be accessible and usable by all students, rather than a select few; and would be readily supportable, adaptable and scalable. This started with defining what is meant by "Impact and sustainability". The group saw both impact and sustainability as being essential to the success of programs, tools, curricula or resources.

The dimensions of impact and sustainability the group identified include:

1. Financial sustainability. Funded projects need to converge on practice in a way that does not require additional costs and can be equitably distributed. In other words *not* "data science for those who can afford it". A further challenge was described as "If you build it they won't have a budget line for it and won't come."
2. Scalability of solutions. Widespread adoption of data science practices and curricula have to be adaptive to local needs and have support infrastructure, i.e. built into preservice and in-service learning.

3. Long-term engagement in, and sustainability of programs must be addressed early. Programs, tools, and curricula must be developed with sustainability in mind. No more tools and curricula are needed that are just created and thrown out there assuming uptake.

The group then addressed three main challenges through the lense of impact and sustainability:

1. Where does data science "live" in high school curriculum: integrated or stand-alone?
    a. To effectively bring data science into high school, we need to think about how to help teachers teach more effectively using data science. We need to identify and share successful integrated and/or stand-alone implementation and student outcomes to better know what works.
    b. Need to be adaptive to and make relevant for local variation in how things are taught/learned.
    c. If it's a stand-alone course in data science, how will this be equitable? How will data science be accessible to all? This same problem exists with computer science - only a select few take computer science, which is why integration models are being adopted. (NYC, SF, how are they doing it?)
    d. Integrate into computer science or statistics? Same problem of equity, i.e. not all students take statistics.
    e. Can we find ways to integrate data science into curricula in ways that help teachers teach complex topics more effectively? Where are the pain points in teaching practice across the curriculum? What can students learn better through a data science frame? Build and test curriculum and practices around the ways that teachers can teach things better and students can learn better with data science.
    f. Conversely, can students better learn and understand difficult topics with data science? May need new practices. Teachers will need to be more proficient in inquiry processes (the "science" in data science) as a precursor to implementing good data science—and so will the students.
    g. Need to look at research around what causes teachers to make changes to their classes and change curriculum – Transformative Learning (Mezirow)? Validate and bring to policymakers?

2. Data Science Vs. Data Literacy
    a. Where is the line between the two?
    b. Data literacy is more about basic knowledge, e.g. What is data? How is it used, for what and by whom? What should every student know about data related to privacy, ethics, equity? The new oil: how are everyone's data harvested and sold, how data are used as a tool for good and a weapon for bad. (But do those ideas belong in computer science?)
    c. Data science is more about creating or using the tools and the data process: asking questions worth investigating, figuring out what data are needed to try to answer the question, cleaning, visualizing and analyzing the data, and drawing conclusions about it.
    d. IBM OpenDS4All program shows data literacy and data science as a continuum, but there is a breakpoint between the two in curriculum. Maybe start there.

e. Data literacy as a stand-alone course, more like information or media literacy. Every student should take it, but should it be in middle school or high school? Should it be part of an information or media literacy course that is already taken by all students?
f. Data science is more integrative, able to map to curricula and standards, including professional development to align and develop resources, tools, lesson plans, etc. Then go to policymakers to make it sustainable.

3. Sustainability: How can programs be designed so that they do not need immense external resources to continue and can be adaptive to changes in data science and literacy needs over time?
    a. Framework for data science as a starting point. What should people know? Content and process skills, habits of mind, etc. Needs to be some baseline to adapt from.
    b. Fitting to standards helps make sustainable.
    c. Infrastructure: to not have to build everything from scratch, but need to put the puzzle together. Too many small pieces loosely (or not at all) joined. Early adopter teachers want to teach data in high school, but where do they go? Where is the easy start? As it is now you have to come up with everything on your own.
    d. Early adopter teachers need to be leveraged to develop tools/resources/curriculum, plus professional development whether inside or outside of computer science. Develop and incorporate peer mentoring and teacher leaders. Create local peer groups to share best practices and other lessons.
    e. Need to work with policymakers and associations to sustain and scale. Peer mentoring among teachers/networking at conferences/local chapters to identify local-level opportunities for success through teaching that works in their local context. Some organizations to partner with include:
        - ACM and ACM Task Force for Data Science (undergrad curriculum)
        - CSDA
        - CSTA Existing teacher groups – Chapters
        - NSTA – State level chapters
        - NCTM, and maybe NCSM
        - ISTE: Could there be a chapter for data science
        - ESIP Federation Education Cluster. Bring ACM SIG to ESIP meeting
        - Google Developer Groups
        - Research Data Alliance

Group proposes doing an EAGER with Invitations to presidents of each relevant association. Identify how it fits their mission and products they have developed.

4.3.3 Training and Support for Educators

Whereas Curriculum and Development was more focused on the exact definition of what Data Science is, the Training and Support group focused on understanding where and how DS could be incorporated into K-12. During Breakout 2, the discussion started with where training and support either already exists or can exist within the current K-12 structure. Discussion began with a diagnostic mapping of potential school-based populations to work with in developing and offering DS support and training, with parallel strategies for such services, and a clarification of the complementarity of single intervention versus ongoing programs.

Some key ideas that emerged:

Community needs. Being able to support educators requires supporting the community that those educators are in and meeting educators' needs as they articulate them. K-12 educators in particular are already balancing teaching what their administration needs them to teach with teaching in the most effective way for their students. To gain the most traction, data science advocates and mentors need to exist at all levels of the K-20 system, from early education to senior undergraduate, and the progression from level to level must have continuity.

Options for continuous and/or embedded Professional Development. The community around DS can be developed through substitute teacher programs, DS residency programs for classroom teachers, in-school teacher learning groups, and inter-school professional learning communities. Teachers need to play an integral role in the generation and contextualization of DL/DS integrated materials in order to create practical lessons that are authentic to their students and existing curricular requirements. DS-Teacher teams would benefit greatly from a DL/DS learning or concept progressions tool from which to build links to their curricula, and school administrators and curriculum coordinators would benefit from the broader K-12 vision it would offer. DS4All can play an important role in capturing DL/DS work already taking place in classrooms.

Hardware, software and database needs. Open source hardware, software, and databases are essential. Curriculum designers and tool builders tend to design at the edge of the field, requiring educators to bear the cost of switching to new tools, which is most burdensome for underfunded/resourced schools. These time, cost, and turnover burdens can be mitigated through leveraging and promoting easily accessible hardware, open source software, and freely available and publically maintained databases.

Consensus on requisite data science skills progression. Most critically, the training and support for educators relies on a strong agreement on what the field expects students to learn, which includes further understanding of where data literacy and data science overlap and/or diverge, and how high school work will translate into undergraduate level data science. Unlike other well developed subjects, there isn't a clear trajectory toward professional DS outcomes for students

during their K-12 experience. Without consensus about the field, many educators will find it difficult to effectively deploy DS concepts in the classroom, let alone an entire curriculum.

4.3.4 College Board Discussion

After the breakout 2 groups reported out, Crystal Furman from the College Board took a few minutes to let the group hear about the questions she had remaining. There were still two remaining questions that the College Board would like more clarity on: what is the difference between data literacy and data science, and what the data science undergraduate level progression looks like, as that will inform the development of the high school progression. Notes from that discussion can be found at the bottom of Appendix E.

4.4 Demonstrations

Data science education tools were demonstrated across both days of the workshop. In the order in which they were presented, demonstrations included:

SuAVE: Survey Analysis via Visual Exploration: SuAVE is a new online platform for visual exploratory analysis of surveys and image collections. It integrates visual, statistical and cartographic analyses and lets users annotate and share images and distribution patterns. It also provides a gateway into advanced data science and machine learning tools by integrating with R and Jupyter notebooks. SuAVE is used for social, political and behavioral surveys, and in visual arts and humanities, biology and ecology, geosciences and urban planning, medical informatics, and portfolio analysis.

CODAP: Common Online Data Analysis Platform: CODAP is free educational software for data analysis. This web-based data science tool is designed as a platform for developers and as an application for students in grades 6-14.

Data Clubs (TERC): The Data Clubs project develops, implements, and tests four new modules on data science for middle school youth that bring together powerful Computer Science tools with engaging, curated data sets. It is implemented in rural and urban settings, including the 4-H SPIN Clubs in Maine and a local partner in the Boston area. The project is interested in broadening participation in this emerging field before stereotypes about who participates becomes entrenched.

Oceans of Data from EDC: The Oceans of Data Institute (ODI) is dedicated to transforming education to help people succeed in school, work, and life in a data-intensive world. ODI envisions a world where everyone has the skills and knowledge to make informed decisions and achieve new insights and understandings using data.

Bootstrap: In Bootstrap:Data Science, students form their own questions about the world around them, analyze data using multiple methods, and write a research paper about their findings. The module covers functions, looping and iteration, data visualization, linear regression, and more. Social studies, science, and business teachers can utilize this module to help students make inferences from data. Math teachers can use this module to introduce foundational concepts in statistics, and it is aligned to the data standards in CS Principles.

OpenDS4All (IBM): Starting in September 2019, IBM and UPenn announced their partnership with the Linux Foundation to deliver an open source project of building blocks for a data science curriculum. The project is in incubation currently as IBM and UPenn create the initial set of materials to contribute. The project will officially launch in early 2020.

Synthetic Generative Data: This tool and technique used to bring data science into the Experimental Physical Chemistry lab was developed by William Leon while at Lehigh University. The tool was written in Igor Pro and generates large sets of physically relevant data for classroom use, enhancing the experimental chemistry laboratory experience without needing to rely on databases (or successful experiments). The tool incorporates both generation and auditing, i.e. generating both an analysis of the tasks assigned in class and a report on the results in each synthetic dataset. The tool enables instructors and TA graders to track down where errors occurred in their analyses.

IDS (Introduction to Data Science): IDS is a "C" approved mathematics course in the University of California A-G requirements. As a statistics course, successful completion of IDS validates Algebra II. IDS is an excellent option for any secondary school student who possesses sufficient mathematical maturity and quantitative reasoning ability and who has successfully completed a first-year course in algebra.

4.5 Project Proposals

The last stage of the workshop process was for participants to work either alone or in small groups to develop proposal ideas that resulted from the workshop process and write them up into one- or two-page descriptions. The full descriptions can be found in Appendix F. Proposal ideas include:

1. A bottom-up frame for data science literacies: This provides a foundation for conceiving of data science integration into school subjects in culturally relevant/sustaining ways. The project proposes using and/or developing:

- Curriculum case studies (in use)
- Community case studies
- Consensus building around cases
- Video case library

2. DataWise Schools Project: A Curriculum Integration Collaboration between DS4All & STEMteachersNYC: Leveraging the extensive experience and strengths in teacher-led workshop development, delivery and implementation of STEMteachersNYC (STN) and the vision, network and expertise of DS4All, the DataWise Schools Project proposes to engage teams of teachers from five NYC middle schools in 2020-2021, in a foundational workshop and sequence of bi-monthly working sessions to co-design and implement a catalog of Data Literacy/Data Science (DL/DS)-integrated lessons, focusing on the power and practicality of enriching their existing curricula through integration of DL/DS concepts, skills, and tools. DS4All and the Northeast Big Data Innovation Hub, in partnership with STEMTeachersNYC, a teacher-led, PD-providing PLC

non-profit and network of 1400+ STEM teachers, will co-facilitate 1) two school-year listening sessions, 2) one weeklong summer foundational DL/DS training that generates a co-designed set of cross-disciplinary lessons, and 3) bi-monthly, targeted, PLC meetings throughout the school year, via DS4All and peer-led in-person workshops.

Outcomes

- Co-designed catalog of integrated middle grades lessons and associated maps of school-level integration of DL/DS across multiple subjects.
- Teacher teams empowered to develop and implement their own DL/DS integrated lessons, through direct support from and co-design with DS4All, STN and each other.
- Mapping of integrated lessons per school (indicator of total student contact time with DL/DS).
- Increase in student engagement and evidence of changes in learning across subjects.
- Reporting on what methods work best in supporting integration of DL/DS across subjects and grades.

3. Data Science for Teachers Goal: Gather Leads of Teacher Associations to Discuss Data Science Collaborations and Implications. Teacher associations provide:

- Disciplinary frameworks (concepts and practices) and standards
- Communities of practice
- Resource repositories
- Professional development

All of these elements are needed to scale and sustain the integration of data science in primary and secondary (K-12) education.

Outcomes of the breakout include identifying:

- Overlapping areas for standards and concepts
- Strategies to collaborate across organizations to promote data science
- Policy supports and barriers to scaling data science
- Needs of teachers in various disciplines to learn and teach data science
- Repositories - how best to share data science curriculum and lesson plans
- How to leverage the communities of practice in each association OR create a separate association

4 Creating Unstructured Data Science Education Drivers for the Big Data Hubs' Open Storage Network Most of the data science tools and principles in educational setting are geared toward structured data. Data Science World is not rows and columns; it's graphs, signals, images, text. Future trends indicate data sources that are unstructured, e.g. signals, images and text will dominate. Data analysis techniques for unstructured data may not readily leverage "tabular data analysis" methods.

Project goals:

- Build data repository of accessible data from various sensor types

- Develop education tools/lesson plans for unstructured data (e.g. representation of sensor data, sources, basics of sensor data manipulation, pre-processing, visualization)
- Develop familiarity navigating graph, image, signal data
- Use cases and experiential learning. Hackathons to allow students to experience data collection and analysis
- Emphasize how data from multiple sensors can enhance decision making

5. Collaborative Proposal: As a result of conversations at the workshop, Andee Rubin (TERC), Emmanuel Schanzer (Bootstrap) and Chad Dorsey (Concord Consortium) identified common interests and complementary strengths that are likely to result in a collaborative proposal involving all three institutions. While we were not unfamiliar with one another's work, the opportunity to spend an extended time together focusing on K-12 data science education deepened our connections in a way that would likely not have happened otherwise. Definitely a useful outcome of the meeting.

6. Student Data Science Journal

- Online K-12 journal of data science: a platform that gives K-12 students an introduction to publishing their research, with a light panel review component.
- Student submissions include putting code online and publishing written/visual/multimedia summaries of results. Students learn to respond to questions and feedback posed by reviewers, strengthening their research.
- Discussed by Katie Naum (Northeast Big Data Hub), Emmanuel Schanzer (Bootstrap), and other members of breakout group 1.4, with plans to regroup at a later meeting in NYC.

7. Data Learning Hubs: Create communities of stakeholders including teachers, mentors, data scientists, tool developers, learning researchers, and students to use particpatory design techniques to create usable, testable and impactful data science education resources. These communities would be located in each of the four regional Big Data Hubs and would use the Hub networks as the community backbone. Catherine Cramer and Steve Uzzo from DS4All would lead this effort in collaboration with the four regional Big Data Hubs.

5. Survey of Workshop Participants

At the end of Day 2 participants completed a survey to provide them an opportunity to reflect on the activities of the workshop and provide feedback to organizers. See Appendix G for survey details.

1. Achieving Workshop Objectives: *"Please reflect on the 2-day workshop you have just completed and indicate your level of agreement with the following statements"* (a 5-element Likert scale was used to assess this module, $n = 21$). Analysis: participants indicated that the workshop was very successful in attaining its objectives, but believed that it needed more emphasis directly on teachers and students, and that going forward a focus on teachers and students was needed.

Question	Result
I believe data science can be successfully integrated into high school teaching and learning	Participants overwhelmingly agreed or strongly agreed with the statement with this statement (95%)
My understanding of challenges and opportunities to data science teaching and learning improved through this workshop	There was general agreement with this statement (71%) with more strongly agreeing than disagreeing
I gained a greater understanding of the needs of teachers to enrich instruction with data science tools and resources	Slightly more than half agreed with, or strongly agreed with this statement. Two participants strongly disagreed
I gained a greater understanding of the needs of students to enrich instruction with data science tools and resources	Only 43% agreed with or strongly agreed with this statement.
I learned about valuable data science tools, data sources and education resources that I didn't know about before the workshop	81% agreed with or strongly agreed with this statement.

2. Rationale for Future Work: (the following free response questions were used to assess this module and are summarized below)

"Do you believe that data science is important to computer science education? If so, why? If not, what ranks as more important?"

While many respondents indicated that data science is valuable to computer science education, some thought it would provide students with more practical knowledge and be useful to applying computer science to the real world. Some suggested it was more valuable than computer science, but a few also saw it as a standalone topic. Reasoned varied, but respondents thought data science was very important to preparing students for the workplace, providing more equitable education, and better prepare them for learning programming and other computer skills.

"What do you believe is the most important outcome of a multi-sector collaboration among educators, academia, industry, and government in data science teaching and learning?"

It was suggested by a number of respondents that the most important outcome of such collaboration would be a framework, standards, or a collective vision for data science education. Several indicated that also important is using data science education to improve equity and employability, while a few others saw the outcome as recognition of, and valuing the multidisciplinary nature of data science. Respondents also indicated that such a collaborative would better be able to influence policy.

3. Developing a Community of Practice: *"Please reflect on the long-term data literacy initiative as discussed in the workshop and indicate your level of agreement with the following statements"* (a 5-element Likert scale was used to assess this module, n = 21). Analysis: participants indicated

that the workshop was very successful in attaining its objectives, but believed that it needed more emphasis directly on teachers and students and that going forward a focus on teachers and students was needed.

Question	**Result**
I believe I can make a valuable contribution to this initiative moving forward.	All participants agreed with or strongly agreed with this statement
I plan to invest time and energy to make this initiative a success.	95% of respondents agreed with or strongly agreed with this statement
Being part of this initiative is an exciting opportunity for me.	95% of respondents agreed with or strongly agreed with this statement
Being part of this initiative is an important professional opportunity for me.	80% of respondents agreed with or strongly agreed with this statement
This initiative is relevant to the success of my work in other areas.	70% of respondents agreed with or strongly agreed with this statement
I feel inspired by this initiative.	90% of respondents agreed with or strongly agreed with this statement
I believe this initiative will positively impact my institution / organization	85% of respondents agreed with or strongly agreed with this statement

Free Response Question "*What value do you feel you would bring to a long-term Data Science for All initiative*?" (Note: the word "value" here can be interpreted broadly to include skills, relationships, and resources). Respondents indicated that they brought a wide variety of ideas and capacities to the initiative, including:

- Scaling of computer science programs from local to regional levels and working with policymakers;
- Wealth of research across academia, industry and government and knowledge of workforce needs;
- Curriculum development and work with teachers and schools;
- Help aligning data science with other curricula.

6. Conclusions and Next Steps

In general, the goal of the workshop was to better understand the landscape of data science education resources, tools, data and curricula, and to identify promising approaches that would readily integrate into high school instruction. Specifically, the goal was to identify the role of data science in high school computer science instruction. Throughout the course of the workshop it was evident from the many perspectives of the stakeholders in the room that, while data science is maturing as a field, it is also rapidly evolving and pervading every aspect of society, creating the urgent need to define and deploy initiatives to better educate society about both the utility of data science across all domains of human knowledge and its impact on society.

However, stepping back and looking at what a successful systemic integration of data science in high school instruction would look like, the many voices and ideas that emerged in the workshop would need better coordination of effort and, in particular, would need to take a hard look at the formal learning environment in which such well-intentioned and well-studied approaches need to be embedded. In other words, there needs to be a clear picture of the learning "ecosystem" in which data science fits. It was noted in discussions that the landscape of undergraduate programs in data science was not clearly understood, and there did not seem to be many extant programs to draw from in terms of gaining insight into how to prepare high school students to enter a college program in data science.

To be equitable, the important ideas of data science must be presented in ways that cut across many physical, economic, and cultural learning settings and create educational opportunity for all students. The workshop brought to the surface three major facets that have to work together, whether for a stand-alone course or integrated across the curriculum, which are: curriculum and resource development that is inclusive of stakeholders; tools that are co-designed with teachers; and ongoing training and support.

The question of where the dividing line is between data science and data literacy came up often during the workshop. Moving forward, discussion is needed about whether a line is necessary or if a gradient is more equitable (or what that even means or might look like), an idea that may have been missed during the development of computer science as a new field. It may be possible for data science to live among both the hard science concepts and broader literacy ideas. After all, these gradients exist within other STEM subjects. For instance, we don't draw the line between chemical science and chemical literacy; we train students in *general chemistry*. We allow students to learn "plum-pudding" models of the atom before quantum physics and we show them rules that no physicist would ever use, all as a way to scale such a massively complex field as physics down to a set of concepts appropriate for a broad and diverse set of students. Similarly, we don't all become literature majors or novelists, but we do all need to be able to understand nuance in what we read every day, be able to communicate complex ideas through a variety of media about any subject that we are learning throughout our academic careers, and indeed throughout life. We need to be literate, but we don't all necessarily need to be a literature experts.

In tandem with these discussions there was much discussion on the definition of equity. What should *all* students know, versus what might be a pathway for those who want a career in data science? From a career perspective, what is unusual about data science is how it pervades and is increasingly essential to so many fields. If data science is truly transdisciplinary, how does it fit

into academic structures that have been siloed nearly as long as they have existed? And without fairly offering sophisticated data skills, are we creating an underclass who will be underskilled for the future workforce, as alluded to in the Burning Glass study (Markow, et al, 2017)? Perhaps the issue of equity and accessibility can only be addressed through allowing a breadth of data science concepts to exist without a hard line being drawn for what can be called literacy, as drawing that line will disproportionately affect underserved schools and students.

So where does this leave us? Clearly there is room to better identify and articulate where data literacy, habits of mind and skills fit throughout the curriculum. This will likely require development of a consensus framework that can align to curricula, standards and scope and sequence and can be used as a point of departure for reviewing curricula and standards. It will require close work with teachers to identify where data science and data literacy fit, and either develop or redevelop lessons and units of instruction that leverage existing or new tools to integrate data science practices into curricula. It was made clear from participants that successful and sustainable integration of data science in teaching and learning practice would only come from participatory design practices involving a mutualistic relationship among educators, tool developers, data scientists, curriculum developers and learning scientists. As the plethora of existing curricula, tools, programs, games, and data sources indicate, a different model is needed to knit together the many pieces into a unified whole. It is also important to remember that as data science advances in leaps and bounds, new and emerging techniques need to be employed for systems of education to keep up with - never mind anticipate - the future that students will face as they join the workforce beyond their academic careers.

Finally, a consensus data science framework must be adopted by policymakers and made available and supported at the federal, state and local levels. The kind of concurrent top-down and bottom-up approach we are suggesting will be required in order to accelerate convergence of data science and teaching and learning and for it to have inertia, scalability and sustainability. To get there we propose the following roadmap:

Roadmap for Data Science for All

Immediate:

- Develop or leverage the existing national network of stakeholders (ideally including teachers, learning specialists, domain experts and policymakers) to prioritize data science for all and execute activities articulated in this roadmap;
- Synthesize existing frameworks for data education (skills and habits of mind, what every student needs to know when they graduate from high school) and provisionally, which disciplines they overlap. This could be a high level list with specifics underneath; Ocean Literacy could be used as a model.
- Separately, identify which skills are needed for students who want to go beyond and specialize in data science (e.g., stand-alone course, integration into computer science, or some kind of advanced statistics or other disciplinary elective).
- Take stock of the landscape of where data science appears in undergraduate education, whether a specialized program, track, minor, or coursework in another program. This could be in the form of a survey, with follow up.

Short Term

- Map data education to current and proposed standards, across the curriculum and in collaboration with teachers and policymakers.
- Engage diverse teachers in participatory design processes to align ideas from data science frameworks to needs in instruction (whether places in curriculum, new approaches to existing topics, lesson or unit plans, or the use of existing or development of new tools), and have them develop exemplars and assessments based on their local or regional needs. Use design-based research to iteratively design and validate/evaluate products.
- In parallel with above, identify supports and sustainability mechanisms and how they fit existing local and state resources. Each state has different models for supporting professional development (and pre-service learning).

Long Term

- Develop sustainable funding model for integration into instruction nation-wide that directly involves policy makers at all levels.
- Using a national network, revisit and forecast needs of data science skills and literacies, evaluate the state of data literacy, diagnose bottlenecks and strategize how to support making programs more effective, and help foster collaboration and the distribution of curriculum, resources, tools, practices and professional development.

A note about impact: As of this writing, there are several concrete steps being taken as a direct result of the workshop. Participants have indicated that they are collaborating on and intending to submit a total of four proposals to NSF based on workshop activities and discussions. These proposals, if funded, would have significant impact on the inclusion of data science in high school curriculum as well as on teacher support and preparation. In addition, one workshop participant returned to the small four-year college in the Northeast where she sits on the Computer Science faculty and immediately was able to convince the college administration of the need to include data science throughout their cirriculum. She writes that: "…all students will have an intro to analytics and computing course no matter what their major is and will learn how to view things through the lens of data and carry it into their diverse majors with them. Everyone is super excited, and we are doing a major academic restructuring right now to integrate this into our core education curriculum." Clearly, the workshop has sparked ideas and inspired action to push work forward in the effort to engage all learners with data.

References

Bergold, J. & Thomas, S. (2012). Participatory Research Methods: A Methodological Approach in Motion. Forum: Qualitative Social Research, 13 (1). Art. 30. Berlin: Institute for Qualitative Research and the Center for Digital Systems, Freie Universität Berlin. http://nbn-resolving.de/urn:nbn:de:0114-fqs1201302

Blomberg, J., Giacomi, J., Mosher, A. and Swenton-Wall, P. (1993) Ethnographic Field Methods and Their Relation to Design. In D. Schuler and A. Namioka (eds) Participatory Design: Principles and Practices. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bødker, S (1996). "Creating conditions for participation: Conflicts and resources in systems design". Human Computer Interaction. 11 (3): 215–236.

Christie, M., Carey, M., Robertson, A., & Grainger, P. (2015). Putting transformative learning theory into practice. Australian Journal of Adult Learning, 55(1), 9.

Ehn, P., Nilsson, E. and Topgaard, R. (2014) Making Futures: Marginal Notes on Innovation, Design and Democracy. Cambridge, MA: MIT Press.

Grudin, J. (2012) A moving target: The evolution of HCI. In Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition. Edited by Jacko, J. Abingdon: Taylor and Francis.

McPherson, T. (2013) Wicked Problems, Social-ecological Systems, and the Utility of Systems Thinking. New York: The Nature of Cities.

Markow, W., Braganza, S., Taska, B., Miller, S.M., and Hughes, D. (2017). The Quant Crunch: How the demand for data science skills is disrupting the job market. Boston, MA: Burning Glass Technologies.

Mezirow, J. (1997). Transformative learning: Theory to practice. New directions for adult and continuing education, 1997(74), 5-12.

Patil, D. (2012). Data scientist: the sexiest job of the 21st century. Harvard Business Review.

Pennington, D. D., Simpson, G. L., McConnell, M. S., Fair, J. M., & Baker, R. J. (2013). Transdisciplinary research, transformative learning, and transformative science. BioScience, 63(7), 564-573.

Salingaros, N. (2011) Peer to Peer Urbanism. Amsterdam: Peer to Peer Foundation

Steen, L. A. (Ed.). (2001). Mathematics and democracy: The case for quantitative literacy. NCED.

Suchman, L. (1987) Plans and situated actions : The Problem of Human-Machine Communication. Cambridge University Press, New York.

Tansley, S., & Tolle, K. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). T. Hey (Ed.). Redmond, WA: Microsoft research.

Tinker, R. (2000) A history of Probeware. White Paper. Cambridge, MA: TERC.

Xie, Y., Eftelioglu, E., Ali, R. Y., Tang, X., Li, Y., Doshi, R., & Shekhar, S. (2017). Transdisciplinary foundations of geospatial data science. ISPRS International Journal of Geo-Information, 6(12), 395.

Appendix A: Biographies of Participants

Name and Contact	Bio
 Ajay Anand ajay.anand@rochester.edu	Ajay Anand currently serves as the Deputy Director of the Goergen Data Science Institute at the University of Rochester where he is responsible for managing the data science education program and identify opportunities for expanding the curriculum offerings. As part of leading education outreach initiatives for the institute, Ajay recently launched an introductory high-school data science summer course within Rochester's pre-college program. Recently, Ajay served as a curriculum committee member of an international cooperative venture (http://www.idssp.org/) to create curriculum frameworks in Introductory Data Science to high-school students. He is the site-PI of an ongoing NSF-funded Research Experiences for Undergraduates (REU) program intersecting data science and music. Ajay also directs the data science capstone program working closely with industry partners. Prior to joining University of Rochester, Ajay served in R&D roles in the medical imaging industry as a senior research scientist and technical project leader in the area of medical ultrasound and healthcare data analytics. He is the co-inventor on more than 25 patent applications and co-authored more than 35 journal articles and conference proceedings. His research interests are in time-series analysis, physical model-based predictive analysis, and biomedical data analytics. Ajay earned his PhD and MS degree from University of Washington, Seattle.
 René Bastón renebaston@gmail.com	René Bastón is the Executive Director of the Northeast Big Data Innovation Hub, dedicated to building or strengthening cross-sector and interdisciplinary partnerships to address large challenges. He is an instructor of the Lean Startup methodology for the NSF's I-Corps program and the co-founder of three startups. Previously, René was Special Advisor on Innovation and Entrepreneurship to the City University of New York Vice Chancellor for Research; Director of Industry Interactions and Entrepreneurship at Columbia University's Data Science Institute; Chief Business Officer at the New York Academy of Sciences; Associate Director at Columbia's Science & Technology Ventures; a manager in Ernst & Young's Healthcare IT consulting group; and spent several years in the laboratory of Eric Kandel at Columbia University's Center for Neurobiology and Behavior. He earned both his M.A. in Biomedical Informatics and his B.A. from Columbia University. René has served on the Advisory Boards of the CUNY Hub for Innovation and Entrepreneurship, Center for an Urban Future, Columbia Center for Advanced Information Management, and the NY Battery and Energy Storage Consortium.

Name and Contact	Bio
 Dorothy Bennett dbennett@nysci.org	Dorothy Bennett is currently NYSCI's Director of Creative Pedagogy, responsible for researching and developing how design, making, and play experiences can offer powerful pathways for diverse learners into STEM. Drawing on 30 years of experience creating k-12 educational media, curricula and teacher enhancement programs, she has researched and developed core pedagogical approaches and preliminary content for NYSCI's Design Lab, a 9,000 square foot interactive exhibit dedicated to engaging school groups and families in hands-on engineering design experiences that spark creative problem solving and nvention. She is currently leading a large scale NSF-funded design-based research study with museum partners investigating how narrative elements can shape girls' engagement in museum-based engineering tasks. She also spearheaded the development of NYSCI's Noticing Tools, a suite of innovative apps that enable children to make math and science discoveries in context of creating compelling digital design projects, with a focus on how they can enhance English Language Learners' mathematics, science, and data literacy. Prior to joining NYSCI, she was a Senior Project Director and Principal Investigator at EDC's Center for Children and Technology, leading an array of national research studies focused on opening up science, engineering, and technology to underrepresented groups. Her most notable work involved creating] narrative-rich, online tinkering and mentoring environments that draw women and girls into STEM.
 Kirk Borne borne_kirk@bah.com	Dr. Kirk Borne is the Principal Data Scientist, first Data Science Fellow, and an Executive Advisor at global technology and consulting firm Booz Allen Hamilton since 2015. He provides thought leadership, mentoring, training, and consulting activities in data science, machine learning, and AI across multiple disciplines. Previously, he was Professor of Astrophysics and Computational Science at George Mason University for 12 years in the graduate and undergraduate data science programs, the latter of which he co-created in 2006. Prior to that, he spent nearly 20 years supporting data systems activities for NASA space science programs, including a role as NASA's Data Archive Project Scientist for the Hubble Space Telescope. Dr. Borne has degrees in physics (B.S., LSU) and astronomy (Ph.D., Caltech). He is an elected Fellow of the International Astrostatistics Association for lifelong contributions to big data research in astronomy. As a global speaker, he has given hundreds of invited talks worldwide, including keynote presentations at dozens of data science, AI and analytics conferences. He is an active contributor on social media, where he promotes data literacy for all and has been named consistently among the top worldwide social influencers in big data, data science, and AI since 2013.

Name and Contact	Bio
 Amy Busey abusey@edc.org	Amy Busey is a Senior Research Associate at EDC's Oceans of Data Institute (ODI) where she leads and contributes to a diverse array of initiatives that support STEM proficiency and school, college, and career success. Her recent work has focused on developing and testing data-centered science curriculum for middle and high school students, and she is Co-PI on newly funded work that explores the potential for authentic data to promote data literacy in elementary grades. Busey was a primary author of ODI's Visualizing Oceans of Data report—a groundbreaking effort to provide guidelines to support interface and tool designers in bridging cyberinfrastructure to classrooms, enabling students to work with large, high-quality scientific datasets. Busey also has experience in qualitative research, promotion of knowledge utilization, and researcher-practitioner partnerships. She engages groups of STEM education researchers in knowledge sharing and dissemination around topics including educational technologies. Busey holds a BS in Psychology from the University of North Carolina at Chapel Hill and an EdM in Mind, Brain, and Education from the Harvard Graduate School of Education.
 Ian Castro castro.ian@berkeley.edu	Ian Castro is an undergraduate student majoring in Media Studies and Microbial Biology at the University of California, Berkeley. Through Berkeley's Division of Data Science and Information, Ian works on designing introductory data science education programs at the undergraduate and graduate level. Under the guidance of Professor Karen Chapple, Ian is the lead instructor for Introduction to Data Science for Graduate Students: a hybrid course for professional track graduate students utilizing massive open online course materials and in-person discussion sections to teach Python and basic data science techniques. At the undergraduate level, he is involved as a student instructor for UC Berkeley's Data 8 course, Foundations of Data Science. He is particularly interested in issues of access, equity, diversity, and inclusion for women and underrepresented minority students in STEM and is currently researching solutions for these problems.

Name and Contact	Bio
 Catherine Cramer catherine.cramer@gmail.com	Catherine Cramer works at the intersection of data-driven science and learning, specifically as it pertains to the understanding of complexity and its application to data and network sciences, with a focus on underrepresented communities. For over 20 years she has developed tools and programs for the teaching and learning of complex network and data science, centering on identifying, creating, sustaining and growing productive and innovative collaborations and partnerships between research, industry and academia. She worked with the centers for Ocean Science Education Excellence (COSEE) and the Ocean Literacy initiative from 2004-2014, and was one of the founders of the Network Literacy and Network Science in Education movements. She remains active in both, most recently organizing the 8th annual Network Science in Education symposium at the University of Vermont as part of the 2019 International School and Conference on Network Science, and is on the Board of the Network Science Society. She is co-editor and co-author of the Springer volume Network Science in Education, published in October 2018. She is currently co-leading the data literacy efforts at the Northeast Big Data Innovation Hub, located at the Columbia University Data Science Institute, as well as a Social Network Analysis of the Hub itself. She is also Founder and Principal of the Woods Hole Institute.
 Yadana Desmond yadana@stemteachersnyc.org	With almost fifteen years experience teaching and managing science education programs in both formal and non-formal settings, from grades pre-K to undergraduate, I continually seek ways to integrate and create partnerships around STEM and environmental subjects, incorporating organizational, statistical and research skills, into locally relevant project-based learning experiences and materials. Working as program director for STEMteachersNYC, an organization that offers professional development programs created and delivered by and for teachers, allows me to support others in doing just that! Prior experience includes work in Thailand for the Enjoy Science Partnership, through the Consortium for Policy Research in Education at Teachers College, and several years as Program Manager of Education Services at the New York Hall of Science.

Name and Contact	Bio
 Chad Dorsey cdorsey@concord.org	Chad Dorsey is President and CEO of the Concord Consortium, which has been an innovation leader in researching and developing STEM educational technology for the past twenty years. Chad's experience ranges across the fields of science, education, and technology. In addition to overseeing a wide variety of STEM projects at the Concord Consortium, he serves as a leader in educational technology across the field on numerous advisory groups and professional workshops. Prior to joining the Concord Consortium in 2008, Chad led teacher professional development workshops as a member of the Maine Mathematics and Science Alliance. Chad has also taught science in classrooms from middle schools through college and has guided educational reform efforts at the district-wide and whole-school levels. While earning his B.A. in physics at St. Olaf College and his M.A. in physics at the University of Oregon, Chad conducted experimental fluid mechanics research, built software models of Antarctic ice streams, and dragged a radar sled by hand across South Cascade Glacier. He first met computers when his family hooked an Apple II to their fancy new color TV set, and he's been a shameless geek ever since.
 Ana Echeverri ana.echeverri@us.ibm.com	Ana Maria Echeverri works at IBM focused on Data Science, Machine Learning, and Artificial Intelligence Skills Growth and Strategy. Her career spans multiple leadership roles in Sales, Marketing, Partner Ecosystems, and Analytics in the Technology industry (Informix, Microsoft, Citrix and IBM); and also leadership roles in startups as Founder and as leader in Digital Marketing and Analytics. A lifelong learner, avid reader, and an entrepreneur at heart, her passion is to build from scratch (businesses, strategies, teams, programs) while leveraging data science and AI capabilities and digital competencies. At IBM she has built a Data Science Skills Competency model, a Data Science Apprenticeship program, spearheaded the creation of an Open Source Data Science Curriculum Kit, and multiple Data Science, Machine Learning and AI learning programs. She holds a Computer Engineering degree, an MBA, a Master of Science in Analytics, and a Graduate Certificate in Strategic Management.

Name and Contact	Bio
 Susan Ettenheim susanettenheim@gmail.com	Susan Ettenheim started her professional life as a visual artist and is a recipient of a National Endowment for the Arts Award in Painting. She worked in libraries and bookstores and created commissioned embroidery pieces, then was Director of Community for Oxygen Media working with Oprah Winfrey's message boards and chats and specializing in women and teen girls online. Susan is a Computer Science and Arts teacher at Eleanor Roosevelt High School, 02M416l, a BJC (Beauty and Joy of Computing) Master Teacher and is a part of the Turtlestitch.org digital embroidery team. With Emmanuel Schanzer, creator of the Bootstrap Data Science curriculum, she is facilitating the first ever Data Science course for CS4All, New York City Department of Education to be taught at 15 schools in the 20-21 school year. Integrating Art and Computer Science is a video about using Turtlestitch to teach Computer Science in Susan Ettenheim's classroom in New York City. The Coding and Stitching website is an international project, using Snap and Turtlestitch to interpret traditional cultural patterns for the purpose of stitching and coding collaboratively.
 Melissa Floca mfloca@ucsd.edu	Melissa Floca leads the development and implementation of the Center's academic programming and research agenda. Her work focuses on issues of education, workforce development and economic competitiveness and she is the Co-Director of the Mexican Migration Field Research Program. Formerly, she worked at the Mexico City-based office of McKinsey & Co., serving clients on projects related to financial inclusion, public health and low-income housing. She went on to found Sé Más Microfinanzas, a microfinance organization providing financial education and financial services to micro-entrepreneurs in Mexico. She holds a degree in Political Science and Economics from Johns Hopkins University and a Graduate Diploma in International Relations from the Johns Hopkins SAIS Bologna Center. She is also a graduate of Columbia Business School. Melissa sits on the Leadership Council of the UC Mexico Initiative, the Inclusive Growth Steering Committee of the San Diego Regional Economic Development Corporation, and the International Affairs Board of the City of San Diego.

Name and Contact	Bio
 Jeff Forbes jforbes@nsf.gov	Jeff Forbes is a Program Director for the Education & Workforce program in the National Science Foundation's Directorate for Computer & Information Science & Engineering (NSF CISE), managing programs that address the critical and complex issues of education and broadening participation in computing. Jeff is currently on leave from Duke University where he is an Associate Professor of the Practice of Computer Science. He received his BS and PhD in computer science from Stanford University and the University of California, Berkeley, respectively. His research interests include computer science education, social information processing, and learning analytics.
 Daniel R. Fuka drfuka@vt.edu	Daniel is a cross-disciplinary scientist working on the Big Island of Hawai'i for Virginia Tech University. He is passionate about mentoring researchers young and old about the scientific opportunities that exist when bridging many disparate sciences. Daniel is an active member of the EarthCube(EC), ESIP, and NEREID communities, a key collaborator on the EC Architecture BCube brokering project, and is a Co-PI on the NSF BALTO data brokering project. He has been active in the NSF RCN on Intelligent Systems for Geosciences (IS-GEO) creating collaborations to better our understanding of Earth systems through applications of intelligent information systems. Outside of his research scientist day job, Daniel enjoys the challenges and rewards of teaching and mentoring students who are at risk of starting their college education under-prepared, and he has been using data architectures in unique ways to bring less prepared college students up to speed quickly while not compromising the education of their better-prepared peers. Daniel is a cross-disciplinary scientist working on the Big Island of Hawai'i for Virginia Tech University. He is passionate about mentoring researchers young and old about the scientific opportunities that exist when bridging many disparate sciences.

Name and Contact	Bio
 Crystal Furman cfurman@collegeboard.org	Crystal Furman is a former AP CS teacher of 17 years, who leads the AP Computer Science Principles (AP CSP) curriculum, instruction and assessment development. As a pilot teacher for AP CSP, becoming an instructional designer for College Board, was the next right path for Crystal to pursue. During her four years at College Board, she has been pivotal in the successful launch of AP CSP, AP's largest course launch in its over 60 year history. She worked with the AP CSP development committee to finalize the exam, authored the instructional approaches section of the Course and Exam Description, and created training materials for AP Summer Institutes and AP Mentoring. In her work as former AP lead of AP Computer Science A (AP CSA), she worked with a committee to articulate this course, beyond just topics, identifying big ideas and skills and articulating the exam task types to provide more transparency to all teachers. The culminating results of this work is the new AP CSA Course and Exam Description that was released in May 2019.
 Michelle Gilman mgilman@ubalt.edu	Michele Gilman is the Venable Professor of Law at the University of Baltimore School of Law. Professor Gilman teaches in the Civil Advocacy Clinic, where she supervises students representing low-income individuals and community groups in a wide range of litigation, legislation, and law reform matters. She writes extensively about privacy, poverty, and social welfare issues, and her articles have appeared in journals including the *California Law Review*, the *Vanderbilt Law Review*, and the *Washington University Law Review*. For the 2019-2020 academic year, she is a faculty fellow at Data & Society, where she is researching the intersection of data privacy law and the concerns of low-income communities. She attended Duke University and the University of Michigan Law School.
 Narine Hall nhall@champlain.edu	Narine Hall is an Assistant Professor and Program Director in Data Science at Champlain College, where she has been a faculty member since 2015. Narine holds a Ph.D. in Computer Science and Complex Systems from the University of Vermont. Narine's research interests lie in the area of data mining, machine learning, cloud computing, evolutionary computation, and complex systems. In 2014 she founded a company called BlinkSecure while working on her Ph.D. Previously she worked at Wolfram Research, IBM, Watson, and Lycos Europe. Currently, she advises local startups in data science and artificial intelligence strategy and implementation. Narine is an enthusiastic advocate for local farms, sustainable agriculture and general education around food sourcing and farm to table movement. She enjoys running races with her kids, CrossFit and competing at Dragon Boat Festival with Champlain's team.

Name and Contact	Bio
 Nicholas Horton nhorton@amherst.edu	Nicholas Horton is Beitzel Professor of Technology and Society and Professor of Statistics and Data Science at Amherst College. His recent work has focused on statistics and data science education. Nick is a fellow of the American Statistical Association and the American Association for the Advancement of Science. He chaired the Committee of Presidents of Statistical Societies and the ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science workgroup. Nick serves on the National Academies Committee for Applied and Theoretical Statistics, was a co-author of the 2018 "Undergraduate Data Science: Opportunities and Options" consensus study report and the ASA's 2016 revised GAISE (Guidelines for Assessment and Instruction in Statistics Education) College Report, and served on the NASEM Roundtable on Post Secondary Data Science Education.
 Shriram Krishnamurthi shriram@gmail.com	Shriram Krishnamurthi is a Professor of Computer Science at Brown University. With collaborators and students, he has created several influential systems and written multiple widely-used books. He also co-directs the Bootstrap integrated computing outreach program. For his work he has received SIGPLAN's Robin Milner Young Researcher Award, SIGSOFT's Influential Educator Award, SIGPLAN's Software Award (joinly), and Brown's Henry Merritt Wriston Fellowship for distinguished contribution to undergraduate education. He has authored over a dozen papers recognized for honors by program committees. He has an honorary doctorate from the Università della Svizzera Italiana.
 Diane Levitt diane.levitt@cornell.edu	Diane Levitt is the Sr. Director of K-12 Education for Cornell Tech. She drives our engagement with the national and NYC computing education communities, including the NYC Dept. of Education's CSforAll initiative, and works with nonprofit partners and individual schools to catalyze K-12 computer science. @diane_levitt

Name and Contact	Bio
 Meredith Mante meredith.mante@ibm.com	Meredith Mante is a Data Scientist with a focus on education. Within IBM's Data and AI group, she develops curriculum on Data Science and Artificial Intelligence topics for client audiences and the broader Data Science/AI Learning Community. She has a bachelor's degree in Psychology (Princeton) and two master's degrees, in Computer Science (NYU) and School Counseling (Manhattan College). She has taught in a variety of settings and she has university experience as a teaching assistant and lecturer in computer science.
 Joe Melendez jam967@cornell.edu	Joe Melendez recently joined Cornell Tech as Teacher-in-Residence, after overseeing Computer Science development and integration in K12 public schools, as the Computer Science Education Manager for the Borough of Manhattan with the New York City Department of Education's CS4All Initiative. Previously Joe had been the STEM Manager at ExpandEd Schools, curriculum developer and edtech specialist at the New York Academy of Sciences, and an educational program developer at the Liberty Science Center. Prior to working in education Joe had been a computer hardware specialist for a small startup in New Jersey and a staff-reporter for the Santa Barbara News-Press in California.
 Katie Naum ken2115@columbia.edu	Katie Naum is the Operations Manager of the Northeast Big Data Innovation Hub, a Science Writer for the National Center for Supercomputing Applications, and a consultant specializing in science and technology communication. She earned a B.A. in sustainable development from Columbia University.

Name and Contact	Bio
 Tom O'Connell tom.oconnell@mouse.org	Tom O'Connell is the Chief Partnerships Officer at Mouse, where he develops strategic partnerships to ensure that every young person has the opportunity to access and amplify technology as a force for good. He is a former high school physics & computer science teacher who started his education career through Teach For America. Prior to his role at Mouse, Tom taught in Houston and Brooklyn, instructed graduate students in CS education, and served as the Interim Executive Director at Code/Interactive, which merged with Mouse in 2018. Tom is a facilitator for Exploring Computer Science, Google CS First, and the CSForAll Initiative's SCRIPT workshop. He is a co-PI on an NSF funded research grant with Hunter College studying formative assessment practices in CS education and a frequent speaker at CS and STEM education related events and conferences.
 Stephanie Ogden sogden@collegeboard.org	Stephanie Ogden leads AP Calculus and AP Statistics for the College Board, where she enjoys the opportunity to support teaching and learning of mathematics for all students and to collaborate with likeminded individuals across the broader program. Stephanie taught students in higher education and secondary math classrooms for more than three decades. During that time, she also became a school leader focused on developing teachers and innovative STEM curriculum. Dr. Ogden was the Principal Investigator for Knox County Schools' First to the Top Grant establishing the East Tennessee hub of the Tennessee STEM Innovation Network and the L&N STEM Academy. In that capacity, she enjoyed the opportunity to coordinate with regional and national representatives from higher and secondary education, government agencies and research entities, philanthropic organizations, and industry–all sharing an interest in STEM education. Stephanie presented frequently at workshops and conferences for leaders and teachers interested in developing individual, regional, and national capacity through education. As work-related tasks and roles have evolved over time, Stephanie's definition of herself as a professional has remained steadfast: Stephanie is a teacher.

Name and Contact	Bio
 Tapan Parikh tparikh@gmail.com	Tapan Parikh is an Associate Professor of Information Science at Cornell Tech in New York City. His research interests include human-computer interaction and the design and use of information technologies for supporting youth and community development. He currently teaches the Remaking the City course at Cornell Tech, which connects graduate students with civic organizations to work on service learning and design projects for local impact. Previously, he was a professor at the UC Berkeley School of Information, where he was one of the founders of the field of Information and Communications Technologies and Development (ICTD), and helped start several international social enterprises working in this area. He has received the NSF CAREER award, a Sloan Fellowship, a UW Diamond Early Career Award and was named Technology Review magazine's Humanitarian of the Year in 2007.
 Kelly Powers kp496@cornell.edu	Kelly is a dedicated professional with 20 years of experience in education and 14 years in business. Established the first Computer Science Teachers Association Chapter in Massachusetts to collaborate with a team of teachers to improve CS education and to support CSTA at the national level. While in New York, launched the New York City Scratch Educators Meetup Group and formed relationships with CSNYC, Code/Interactive, NYC DOE CS leaders, and Cornell Tech. While in Massachusetts established strong relationships with the Massachusetts Department of Education, University Professors, industry leaders, local business leaders, NSF funded project leaders from CAITE/ECEP, BAITEC to collaborate on common efforts to improve CS education funding, certification, and opportunities for teachers and students. Worked with student leaders outside of school to run workshops for local communities to learn about Computer Science while creating artifacts. Developed an innovative Computer Science curriculum for schools and have led major projects for IBM, John Hancock and Harvard University. The combination of education and corporate experience has helped me sustain excellent peer and client relationships with policy leaders, industry leaders, teachers, parents, and students.
 Hari Raghavan hraghav@us.ibm.com	Hari Raghavan is a Senior Program Manager with IBM Corporate Social Responsibility in New York City, where he works on a variety of tech-related education and skills development initiatives for adults and young people across the world.

Name and Contact	Bio
 Meg Ray meg.ray@cornell.edu	Meg Ray is a K12 Education Advisor at Cornell Tech. Meg teaches education courses at NYU and Hunter College, and is an experienced high school computer science teacher and curriculum developer. She was a writer for the CSTA Standards for CS Educators, CSTA K-12 CS Standards, and K12 CS Framework. She conducts educational research related to teaching CS to students with disabilities and CS teacher preparation. Meg is the author of the book *Code This Game!*
 Jennifer Rosato jrosato@css.edu	Jennifer Rosato leads the National Center for Computer Science Education at the College of St. Scholastica in Duluth, MN. The Center champions, researches, and provides equitable computer science education opportunities for K16 students and educators through research, curriculum, and professional development. Curricula for Advanced Placement courses include Mobile CSP and CS Awesome, which are used by thousands of teachers and students across the country. Rosato has also created and directs programs to prepare computer science teachers in pre-service and in-service programs. She is the principal investigator and consults on multiple grants from the National Science Foundation, Google, and Infosys Foundation, USA totaling over $6 million. Rosato is currently the Chair of the Board of Directors for the Computer Science Teachers Association. Rosato has a bachelor's degree in Biochemistry from St. Scholastica and a Masters of Arts in Information Systems Management from Carnegie Mellon University.
 Andee Rubin andee_rubin@terc.edu	Andee Rubin is a mathematician, computer scientist, and learning scientist who has been studying the growth of students' and teachers' statistical reasoning for almost 30 years, particularly as it is enabled by research-based tools for statistics education. She was involved in the development of several such pieces of software and led the ViSOR (Visualizing Statistical Reasoning) project, which studied how middle and high school teachers used data visualization tools with their students. She was co-PI of Science Literacy through Infographics (SLI), which studied the development of a socio-technical system for supporting high school students in creating science infographics and is co-PI of Data Clubs for Middle School Youth, which is developing introductory data science out-of-school experiences.

Name and Contact	Bio
 Emmanuel Schanzer schanzer@bootstrapworld.org	Emmanuel spent several years as a program manager and developer before becoming a public high school teacher and middle school academic coach in Boston. He is the founder and co-director of Bootstrap, which he first designed as a curriculum for his own students. He has long been involved in connecting educators and technology, connecting parties at the Computer Science Teachers Association, Google, Microsoft, Facebook and at universities across the country. He holds degrees in computer science and curriculum development, and completed his doctoral studies at Harvard with a research focus on using programming to teach algebra.
Lisa Singh lisa.singh@georgetown.edu	Lisa Singh is a Professor in the Department of Computer Science and a Research Professor in the Massive Data Institute (MDI) at Georgetown University. She has authored/co-authored over 70 peer reviewed publications and book chapters related to data-centric computing. Current projects include studying privacy on the web, identifying noise and poor quality information on social media, developing methods and tools to better understand forced movement due to conflict, and learning from public, open source big data to advance social science research involving the understanding of human behavior. Her research has been supported by the National Science Foundation, the Office of Naval Research, the Social Science and Humanities Research Council, and the Department of Defense. Dr. Singh has also recently organized three workshops involving future directions of big data research and is currently involved in different organizations working on increasing participation of women in computing and integrating computational thinking into K-12 curricula. Dr. Singh received her B.S.E. from Duke University and her M.S. and Ph.D. from Northwestern University.

Name and Contact	Bio
 Julia Stoyanovich stoyanovich@nyu.edu	Julia Stoyanovich is an Assistant Professor at New York University in the Department of Computer Science and Engineering at the Tandon School of Engineering, and the Center for Data Science. Julia's research focuses on responsible data management and analysis practices: on operationalizing fairness, diversity, transparency, and data protection in all stages of the data acquisition and processing lifecycle. She established the Data, Responsibly consortium (https://dataresponsibly.github.io/), and served on the New York City Automated Decision Systems Task Force, by appointment from Mayor de Blasio. In Spring 2019, Julia developed and is teaching a course on Responsible Data Science at NYU (https://dataresponsibly.github.io/courses/spring19/). In addition to data ethics, Julia works on management and analysis of preference data, and on querying large evolving graphs. She holds M.S. and Ph.D. degrees in Computer Science from Columbia University, and a B.S. in Computer Science and in Mathematics and Statistics from the University of Massachusetts at Amherst. Julia's work has been funded by the NSF, BSF and by industry. She is a recipient of an NSF CAREER award and of an NSF/CRA CI Fellowship.
 Rochelle Tractenberg rochelle.tractenberg@gmail.com	Rochelle Tractenberg is a tenured professor at Georgetown University who focuses on cognitive scientific aspects of teaching and learning in graduate and post-graduate/professional education. Her research falls into three main categories: methodology and clinical trial design; psychometrics for, and measurement of, difficult-to-assess clinical entities; and curriculum development and evaluation in higher education. She is the creator of the Mastery Rubric construct, which captures the knowledge, skills, and abilities of a curriculum and maps their developmental trajectory into performance level descriptors that can be concretely demonstrated using a variety of evidence and assessment methods. An elected fellow of both the American Statistical Association and the American Association for the Advancement of Science, she has been recognized for her commitment to ethical practice (of statistics and data science) and for her stewardship of science.

Name and Contact	Bio
 Stephen Uzzo suzzo@nysci.org	As Chief Scientist for the New York Hall of Science (NYSCI), Stephen Uzzo develops and leads large-scale initiatives to research and integrate cutting edge science into teaching and learning. He currently develops programs to build communities of practice in complexity, data-driven science and engineering, and improve science, technology, engineering and math (STEM) literacy of the public. His background includes over 20 years experience in the research of connected systems and teaching and learning in STEM; and prior to that, 10 years in video and computer graphics systems engineering. Dr. Uzzo's research interests include the coupling of complex human and natural systems, complex networks, smart cities, and the impact of big data on communities of need. He holds a terminal degree in network theory and environmental studies and serves on a number of institutional and advisory boards related to his interests. His work also includes developing, studying and teaching graduate programs in STEM learning. Having never lived very far from the ocean in New York and California, Dr. Uzzo has been a lifelong advocate for marine conservation.
 Sara Vogel svogel@gradcenter.cuny.edu	Sara Vogel is a doctoral candidate in Urban Education at the Graduate Center of the City University of New York, writing her dissertation about how emergent bilingual middle schoolers use diverse language resources during computer science activities. She is currently the lead research assistant on Participating in Literacies and Computer Science (PiLaCS), a National Science Foundation-funded project which aims to leverage the diverse language practices of bilingual youth as resources in their computer science learning. In collaboration with the NYC-based Hive Research Lab, she founded the CS Education Visions project, which has surfaced the diverse visions that formal and informal educators have for universal computer science education initiatives.

Name and Contact	Bio
 Michelle Wilkerson mwilkers@berkeley.edu	Hello! I am an Assistant Professor in the Graduate School of Education and the Graduate Group for Science and Mathematics Education (SESAME) at UC Berkeley. I study how young people learn with and about digital texts such as simulations, visualizations, and data explorers. Using participatory design-based methodologies, I develop tools and pedagogies that position students as authors of computational texts by building on familiar expressive practices such as sketching, crafting, and storytelling. Much of my work focuses on computational data visualization and analysis. My 2014 project "DataSketch" exploring young learners' data visualization literacies was awarded a 2014 NSF CAREER grant. "Writing Data Stories," a new collaboration between UC Berkeley, the Concord Consortium, North Carolina State University, and the University of Texas at Austin is exploring the development of students' sociocritical data literacies in the middle school science context. I am fascinated by how learners' relationships to data impact their learning. Recently, along with Joseph Polman, I guest edited a special issue of the *Journal of the Learning Sciences* devoted to this topic—I hope you check it out! I hold a PhD in Learning Sciences from Northwestern University, and was an Assistant Professor at Tufts University before joining UC Berkeley.
 Elena Yulaeva eyulaeva@ucsd.edu	Dr. Elena Yulaeva is an expert in global and regional climate and weather analysis and modeling, with years of experience in high-performance computing, numerical methods in geophysics, statistical analysis, machine learning, data science, and management of international multi-stakeholder projects. She received her Ph.D. in atmospheric sciences from the University of Washington. For the last 20+ years she has been engaged in climate change analysis and its impacts on global communities. In addition to her research work at UCSD, Elena runs (as an Executive Director) Community Commons, a San Diego based non-profit organization that develops innovative solutions to support informed, healthy, equitable, and sustainable communities. In this capacity, she has organized and led educational STEM programs at several levels, from K-12 to university students. She is the founder and co-director of the award-winning international Global Forest Link Project (globalforestlink.com), which emphasizes active learning and engages youth from multiple countries in field environmental data collection, data analysis, development of digital stories, and collaboration. The project's curriculum spans existing high school courses, employing innovative data science gateway tools and real datasets to get youth engaged and prepare them professionally for 21-st century jobs.

Appendix B: Workshop Agenda

Thursday January 16

8:00 *Registration and breakfast*
9:00 Welcome, outline of agenda and goals - Workshop organizers
9:20 NSF Perspectives on Data Science - Jeff Forbes, NSF
9:45 Data Science for All - Catherine Cramer
9:55 Data and Learning - Steve Uzzo
10:10 Digital Justice - Michele Gilman, Data & Society, University of Baltimore
10:40 *Coffee Break*
11:00 Panel: Perspectives on CS and DS in the Classroom

- Diane Levitt, Cornell Tech, *Moderator*
- Crystal Furman, The College Board
- Joe Melendez, Cornell Tech
- Tom O'Connell, Mouse
- Aankit Patel, City University of New York

12:00 *Lunch*
1:00 Deep Dive: Successes and Challenges
2:30 *Coffee Break*
2:50 Mapping Successes and Challenges
4:00 Learning tools demonstrations
5:00 Adjourn
Dinner on your own

Friday January 17

8:00 *Breakfast*
9:00 Welcome back
9:05 Preparing Students for an Intensely Digital World - Kirk Borne, Booz Allen Hamilton, introduced by Rene Baston, Northeast Big Data Innovation Hub
10:10 *Coffee Break*
10:30 Breakout Sessions: Challenges and Solutions
11:30 Report-outs
12:30 *Lunch*
1:30 Recommendations for white paper
2:00 Small team proposal writing
2:45 *Coffee Break*
3:00 Presentations
4:00 White paper planning, exit surveys
4:30 Adjourn

Appendix C: Keynote Talks Abstracts

Michele Gilman, University of Baltimore; Fellow, Data & Society: *Digital Justice*

Part of data literacy involves understanding the justice and equality implications of automated decision-making systems. Technological advances are improving multiple aspects of modern life, but there are harms as well, and they tend to be borne by the most marginalized communities. This presentation highlights five examples of data injustice that arise from automated decision-making systems: (1) discriminatory outputs as a result of structural biases embedded in algorithmic design and datasets; (2) unwarranted deference by policymakers and decision-makers to automated systems as a result of "automation bias"; (3) social sorting of different populations made possible through digital profiling; (4) lack of transparency and accountability in black box systems; and (5) inaccuracies in algorithmic outputs that can arise from decisions developers make as they build algorithms as well as errors in datasets. Examples of data injustice include biases against African-Americans in health care algorithms; digital discrimination against minorities in predictive analytics used in the criminal justice system; algorithmic sorting of college applicants; recruiting technology that claims to identify successful employees based on algorithmic analysis of video interviews; and inaccuracies in facial recognition technology for women of color. Understanding these dynamics is a first step to building better algorithms, and to deploying them more fairly, transparently, and accountably.

Kirk Borne, Booz Allen Hamilton: *Preparing Students for an Intensely Digital World*

I discuss the importance and imperative for all students to become digital and data literate, for the future of work in general and the future of their own work. I cover 5 major themes in the presentation: data awareness (what is it?), data relevance (why me?), data literacy (show me how), data science (where's the science?), and data imperative (create and do something with data). I emphasize throughout my presentation how data permeates nearly all of our daily lives through all conceivable digital technologies, handheld devices, business activities, and personal activities. Through data, the world is computable. I complement the content with stories: data stories, mostly from my own experience in delivering data literacy workshops and presentations to students and the general public. These stories highlight the inspirational, aspirational, and invaluable aspects of data in our lives, particularly in students' lives. The focus is not on the mathematics, nor the algorithms, nor the engineering. Instead, the focus is on demonstrating that data science is universally appealing, data literacy is accessible, data fluency is achievable, and data is for all. As I like to tell my technical audiences, "Come for the data. Stay for the science." And I tell my general audiences, "Data is not a 4-letter word!" And I tell my education audiences, "Data is the gateway to STEM."

Appendix D: Notes from Breakout 1 Report-Outs from Groups

General Challenges

1. Equity
    a. Large cultural barriers
    b. How is data science perceived?
    c. Visibility/transparency about data and how it's used (including ML/AI)
2. Accessibility of Data Sets
3. Lack of support for educators delivering data science instruction
4. Identifying and leveraging existing resources
    Access to these may be particular challenge
5. Lack of consistency in definitions of data and data science concepts
6. Single points of failure w/ regard to trained faculty
7. Financial sustainability and scalability of solutions
8. Threats to validity
    a. W/ regard to data sources, assumption used, etc.
    b. Integrating concepts of Responsible Data Science
9. Challenging to teach interdisciplinary topics
10. Gauging appropriate level of complexity of concepts/data/skills to teach
    a. Particularly making instruction accessible to students of varying skills backgrounds in classroom
    b. Probabilistic solutions; understanding of error/variability;
    c. Variability of data sets/messiness
11. Teaching critical thinking skills which underlie data science instruction is ongoing challenge
12. Conveying understanding of "Why" data science approaches are important/relevant in different contexts (visions.cs4all.org); and matching rationale to practice is challenging
    Applying solutions to this taking into account varying perspectives (political and others)
13. Designing and delivering appropriate/effective messaging that addresses varying perceptions of data science…and, how does this inform the design of curricula
14. Challenge for data science community – better use of systems thinking/approaches to inform development of policies

General Lessons Learned

1. Availability of and carefully timed feedback can support more effective student outcomes
2. Long-term engagement in and sustainability of programs must be addressed early
3. Student agency in selecting problems to address and data to use enhances outcomes
    a. Feasibility/appropriate constraints to problems/projects equally important
    b. Plus 1 – balance between constraints and student agency is important
4. Understanding process of science is important to understanding the use of data science
5. Buy in of administrators is critical
6. Defining data science competencies provides a good guide for data science curricula/instruction
7. Visualization (storytelling?) can enhance data science learning experience
8. Consistency in describing continuum of data science activities effectively (80% data prep/20% analysis) to students

9. Integration w/ Respect -> There are many entry points to introduce data science across subjects
    But, one size doesn't fit all…each discipline must be addressed with nuance
10. Using data sciences to inform difficult discussions
11. Learning from students about how to support their data science learning
12. Approaches developed in APCSP (link)?

Appendix E: Breakout 2 Notes

Group 1: Curriculum and Development

Curricular issues: RAW LIST

Throw out a thing you think should be included - don't worry about what it is. What do you think students should learn as part of a data science curriculum?

- Defining Data
- Data visualization
- Case attribute structure -- understanding that data come in cases and the cases have attributes
- Variability/Uncertainty
    - Data have noise associated with them
    - Partitioning variability
- Multivariate thinking
- Uncertainty
- Statistics
    - Working in a probabilistic environment
    - Measures of center
- Aspects of sampling and design, non-sampling errors
- Understanding data as a social text
    - EG: Data have authors with perspectives encoded in data, social narratives get encoded in the data (which may not be able to be read completely)
        - Data context - subtext
    - Data ethics - EG: data science ethics framework -- where do data come from?
    - Checking for bias? Understanding impacts?
    - Ethical practices of data - Decisions you make to collect and hold data - balanced against the public good
    - Impacts of data in society
- Use data to compose arguments, info graphics, tell stories
    - Storytelling with data
    - Reporting with data
- Privacy and confidentiality
- Use cases - real world use cases
    - Relevant to student populations
    - Not "corporate" cases
- Unstructured data
    - Ability to work with complex data (and context)
    - Ability to work with organic data, produced from things like sensors
- Understanding the role of context
- Describing and comparing distributions / distributional thinking
- Analytic Tools - not just looking at graphs but to analyze the data
- Reproducibility
- Asking questions and figuring out which questions are appropriate to ask with data/can be answered with data
- Workflow in answering questions - data science life cycle
- The scientific method
- Threats to validity and critiquing / Model assessment

    - Being able to question / think critically about the arguments and stories created with data
    - Making assumptions explicit
    - Checking assumptions
    - Checking data
    - Establishing confidence
- Research methods
- Meta cognition
- Misrepresentation with charts/"how to lie with maps" - standard misrepresentation with data and how to read them
- Aggregate thinking
    - Four lenses on data -- bringing data up to aggregate
- Seeing visualizations as intentional / created storytelling
- Correlation vs. causation
- Understand one's own relationship to data
    - Students reflected in the data demographically
    - Students as producers of data
- How a representation is derived from data
    - Meta representational
    - Dear Data - students have to choose what attribute will be represented by what representation. Tools are not just a black box, there are choices that are made
- Data structure / importance of data structure
- Metadata
- Approaches to the scientific method - positivist?
- How data was collected, by whom, for what purpose? How that affects use and information?
- Knowledge graphs
- Equity and diversity
    - (Appreciation of) design for accessibility for different populations (disabled)
- Testing and validation
- Vectors and matrices
- Mitigation of bias
- Data wrangling
    - Open-source
- Cleaning / cleansing
- Interpolation / prediction
- Implementing machine learning algorithms
    - Supervised vs. unsupervised machine learning
- Algorithmic thinking
- Relationship of data to modeling
- Generalization, inference, interpretation -- how far do your results go - Representativeness of the data.
- Data generation processes
    - Source / provenance
- Data and society
    - How databases are shared
- Hypotheses
- Data dictionaries
- Categorical vs. numerical data
- Indirect vs. direct measurement
- Indirect vs. direct effects

    - Proxy
- Exposure to tools - recognition of the importance of tools
- Explainability
- Algorithmic justice
- Confounding variables
- History of data science
- Randomization
- Experimental design
- Connections to school disciplines
- Communication, presentation
- Vehicle for difficult conversations

Curriculum threats

- What are the boundaries of the kitchen sink
    - Data science vs. data literacy tracks
- Do we think DS will be taught in the context of
- These are similar to the National Academies report
- What is NOT at the K-12 level / what's not for everyone

List for clustering (grab and move!):

- C - Define data
- C – Understand variability/Uncertainty
    - Data have noise associated with them
    - Partitioning variability
- C- Uncertainty
- Statistics
    - Pr - Working in a probabilistic environment
    - C - Measures of center
- C- Aspects of sampling and design, non-sampling errors
- Pe - Understanding data as a social text
    - Pe – EG: Data have authors with perspectives encoded in data, social narratives get encoded in the data (which may not be able to be read completely)
        - Data context - subtext
    - C / Pr / Pe - Data ethics - EG: data science ethics framework -- where do data come from?
    - C / Pr - Checking for bias? Understanding impacts?
    - C / Pr / Pe - Ethical practices of data - Decisions you make to collect and hold data - balanced against the public good
    - Impacts of data in society
- Pr - Use data to compose arguments, info graphics, tell stories
    - Pr - Storytelling with data
    - C / Pr - Reporting with data
- C - Privacy and confidentiality
- ?? Use cases - real world use cases
    - Relevant to student populations
    - Not "corporate" cases
- C - Unstructured data
    - Pr - Ability to work with complex data (and context)
    - Pr - Ability to work with organic data, produced from things like sensors
- C - Understanding the role of context
- C / Pr - Describing and comparing distributions

- Pe - distributional thinking
- C / Pr - Analytic Tools - not just looking at graphs but to analyze the data
- C / Pr - Reproducibility
- Pe / Pr - Asking questions and figuring out which questions are appropriate to ask with data/can be answered with data
- Pr / Pe - Workflow in answering questions - data science life cycle
- ??? The scientific method
- C / Pr / Pe – Threats to validity and critiquing / Model assessment
    - Pr - Being able to question / think critically about the arguments and stories created with data
    - Pr - Making assumptions explicit
    - Pr - Checking data
    - C / Pr - Checking assumptions
    - C / Pr - Establishing confidence
- Pr - Research methods
- Pe / Pr – Metacognition
- C / Pe – Misrepresentation with charts/"how to lie with maps" - standard misrepresentation with data and how to read them
- C – Pr? Aggregate thinking
    - Four lenses on data -- bringing data up to aggregate
- Pe – Seeing visualizations as intentional / created storytelling
- Pe - Correlation vs. causation
- Pe – Understand one's own relationship to data
    - Pe – Students reflected in the data demographically
    - Pe – Students as producers of data
- C / Pr / Pe – How a representation is derived from data
    - C / Pe – Meta representational
    - C / Pr – Dear Data - students have to choose what attribute will be represented by what representation. Tools are not just a black box, there are choices that are made
- C – Data structure / importance of data structure
- C / Pr – Metadata
- C / Pe – Approaches to the scientific method - positivist?
- C / Pe – How data was collected, by whom, for what purpose? How that effects use and information?
- ?? Knowledge graphs
- C / Pe – Equity and diversity
    - C / Pr / Pe – (Appreciation of) design for accessibility for different populations (disabled)
- C / Pr – Testing and validation
- C – Vectors and matrices
- C / Pr – Mitigation of bias
- C / Pr – Data wrangling
    - C – Open-source
- C / Pr – Cleaning / cleansing
- C / Pr – Interpolation / prediction
- C / Pr – Implementing machine learning algorithms
    - C – Supervised vs. unsupervised machine learning
- C / Pr – Algorithmic thinking
- Pe – Relationship of data to modeling

- C / Pr – Generalization, inference, interpretation -- how far do your results go - Representativeness of the data.
- Pr – Data generation processes
    - C / Pr – Source / provenance
- Pe – Data and society
    - Pe / Pr – How databases are shared
- ?? Hypotheses
- C – Data dictionaries
- C / (Pr) – Categorical vs. numerical data
- C / Pr – Indirect vs. direct measurement
- C – Indirect vs direct effects
    - C – Proxy
- C / Pr – Exposure to tools - recognition of the importance of tools
- C / Pr / Pe – Explainability
- C / Pe – Algorithmic justice
- C / Pr – Confounding variables
- C – History of data science
- ?? C / Pr – Randomization
- C / Pr – Experimental design
- C / Pr – Connections to school disciplines
- Pr – Communication, presentation
- Pe – Vehicle for difficult conversations

Clusters/Levels

Concepts

1. *Definition of data*
    a. Categorical vs. numerical data
    b. *Data structure*
        i. Importance of data structure
        ii. Unstructured data
        iii. Complex data
2. *Impacts of data in society*
    a. *Data ethics*
        i. Data science ethics framework -- where do data come from?
        ii. Checking for bias? Understanding impacts? Mitigation of bias
        iii. Ethical practices of data - Decisions you make to collect and hold data - balanced against the public good
        iv. Privacy and confidentiality
        v. Misrepresentation with charts / ”how to lie with maps” - standard misrepresentation with data and how to read them
        vi. How data was collected, by whom, for what purpose? How that affects use and information?
        vii. Algorithmic justice
    b. Equity and diversity
        i. (Appreciation of) design for accessibility for different populations (disabled)
    c. History of data science
3. *Generation / collection / (workflow?)*
    a. Understanding the role of context
    b. Source / provenance
    c. Establishing confidence

d. Metadata
e. Ability to work with organic data, produced from things like sensors
f. Data dictionaries
g. Indirect vs. direct measurement

4. *Analysis*
    a. Data wrangling
        i. Data moves?
        ii. Cleaning / cleansing
        iii. Merging datasets
        iv. Interpolation / prediction
    b. Distribution, aggregation and shape
        i. Describing and comparing distributions
        ii. Aggregate thinking
        iii. Four lenses on data -- bringing data up to aggregate
    c. Role of analytic tools
        i. Use for data analysis
    d. Reproducibility
    e. Modeling
        i. Threats to validity and critiquing / Model assessment
        ii. Checking assumptions
        iii. Machine learning algorithms / Supervised vs. unsupervised machine learning
        iv. Algorithmic thinking
        v. Confounding variables
    f. Testing and validation
    g. Generalization, inference, interpretation -- how far do your results go - Representativeness of the data
    h. Understanding variability and uncertainty
        i. Data have noise associated with them
        ii. Partitioning variability
        iii. Uncertainty
    i. *Sampling*
        i. Sampling design
        ii. Non-sampling errors
    j. Indirect vs. direct effects
    k. Proxy
    l. Exposure to tools - recognition of the importance of tools
    m. Explainability
5. *Representation*
    a. Representations are derived from data
    b. Dear Data - students have to choose what attribute will be represented by what representation. Tools are not just a black box, there are choices that are made

<u>Practices</u>

1. *Generating data*
    a. Engage in data generation processes
        i. Designing experiments
        ii. Surveying
        iii. Generation from probes and sensors
        iv. Distinguishing and determining categorical vs. numerical data
        v. Discriminating and applying indirect vs. direct measurement

    b. Data wrangling
        i. Cleaning / cleansing
        ii. Interpolation
    c. Determining and understanding source / provenance
    d. Interpreting and creating metadata
2. *Analyzing data*
    a. Selection and use of tools for analysis
    b. Applying multivariate thinking
    c. Describing and comparing distributions
    d. Using analytic tools
    e. Asking questions and figuring out which questions are appropriate to ask with data/can be answered with data
    f. Workflow in answering questions - data science life cycle
    g. Threats to validity and critiquing / Model assessment
    h. Aggregate thinking
    i. Articulating uncertainty
    j. Four lenses on data -- bringing data up to aggregate
    k. Explainability
    l. Confounding variables
    m. Randomization
    n. Generalizing from and modeling with data
        i. Assessing generalization, inference, interpretation -- how far do your results go
        ii. Representativeness of the data
        iii. Testing and validation
        iv. Implementing machine learning algorithms
        v. Supervised vs unsupervised machine learning
        vi. Algorithmic thinking
3. *Representing data*
    a. How a representation is derived from data
        i. Meta representational
        ii. Dear Data - students have to choose what attribute will be represented by what representation. Tools are not just a black box, there are choices that are made
4. *(Questioning and thinking about our data)*
    a. ???Reproducibility
    b. Being able to question / think critically about the arguments and stories created with data
    c. Checking data
    d. Mitigation of bias
    e. Establishing confidence
    f. Checking for bias? Understanding impacts?
    g. Data and society
    h. Data ethics - EG: data science ethics framework -- where do data come from?
    i. Ethical practices of data - Decisions you make to collect and hold data - balanced against the public good
5. *Communicating from and with data*
    a. Use data to compose arguments, infographics, tell stories
    b. Storytelling with data
    c. Reporting with data
    d. Making assumptions explicit

Perspectives

1. *Understanding data as a social text*
    a. Data have authors with perspectives encoded in data, social narratives get encoded in the data (which may not be able to be read completely)
    b. Data context - subtext
    c. Understand one's own relationship to data
    d. Students reflected in the data demographically
    e. Students as producers of data
    f. How data was collected, by whom, for what purpose? How that affects use and information?
    g. Approaches to the scientific method - positivist?
2. *Appreciating data ethics and societal ramifications*
    a. EG: data science ethics framework -- where do data come from?
    b. Ethical practices of data - Decisions you make to collect and hold data - balanced against the public good
    c. Impacts of data in society
    d. Data and society
        i. How databases are shared
    e. Recognizing that data is intertwined with questions of equity and diversity
        i. (Appreciation of) design for accessibility for different populations (disabled)
    f. Algorithmic justice
    g. Vehicle for difficult conversations
3. *Identifying that data can answer (some but not all) questions*
    a. Seeing the world through distributional thinking
    b. Workflow in answering questions - data science life cycle
    c. Threats to validity and critiquing / Model assessment
    d. Metacognition
4. *Appreciating the complexity and nuances of data representation and analysis*
    a. Misrepresentation with charts/"how to lie with maps" - standard misrepresentation with data and how to read them
    b. Seeing visualizations as intentional / created storytelling
    c. Correlation vs. causation
    d. How a representation is derived from data
    e. Meta representational
    f. Relationship of data to modeling
    g. Explainability

Orientations for US

- Connections to school disciplines
- Pedagogies
- Use cases - real world use cases
    - Relevant to student populations
    - Not "corporate" cases
- Examples of students doing these in action

Categorise Literacy-Fluency-Science by grade level

Group 2: Impact and Sustainability

What is sustainability:

How do you create something that teachers will take up and use

How many years has the exam been used?

Desirability, will STA want this

Viability, how do you introduce it to the students

Feasibility, Do we have the right resources, (financial, data, local relevance, infrastructure).

How do you get teachers to take something up? (put it into standards)

How do you make cross local relevance testing?

Create interventions, data science, can teach more effectively, better student outcomes.

Need an end to end process…

No content, develop standards process-based rather than content-based.

From Undergrad perspective, students/teachers ask general question of how to learn more about data science? When designing, how do you make the learning experience smooth?

Peer mentoring, early adoption teachers are willing to help other teachers. How Google works? GDG, Google developer Groups. System is in place for Chapter Leader to automate an event. Similarities with models from ESIP, and RDA? Computer Science Teachers, Social Studies

Data literacy might be it's own multi-domain community Affinity

Where in the standards can "this" plug in,

Policy standards: aligning new standards with states is much harder than integrating with existing standards. Is there an existing workflow for doing this?

Participatory design, DBIR Design-Based Implementation Research.

There is data literacy and data science; data literacy is content based, Data Science is

If teachers or students don't understand what data science is, how do you intervention with teacher? How do you design the professional development, to leverage a student intervention? How do we align with existing standards? "Pre-service learning"?

How do you get early-adopting teachers, and how do you bring it to more teachers?

Can we get funding for officers from each of these chapters to attend ESIP/RDA/ … CSTA is ACMs Teacher/Education. NCTM - Math, STA, Social Studies,

Group 3: Training and Support

Our approach to training and support started out more diagnostic.
We started with looking at training and supports

- Training vs. support, what is the difference between those?
- Training is a discrete experience and support is continuous

Who do we want to want to work with?

- Teachers, subs, curriculum developers, OST learning

What kind of methods are available for creating Hack-a-thons, short discrete P.D.s, lesson study models, reciprocal residency idea where researchers go into teachers classrooms,
Categories for the goals of the methods could lead to training around content, integrated into content
In-person, Community of Practice, peer teaching, chapters
Need to Formatively assess the implementation
Pathways and progressions in curriculum development map out pathways to tools and curriculum
How can we support a teacher to? Focus on equity

- The ability to teach what the admin aspects?
- The ability to teach what is effective to their students?
- Not enough time or resources for the teachers to teach effectively

What is the ultimate learning goal?

- What are you training and supporting teachers to be able to teach?
- Broader learning goal of having students who have gone through a good data science program be good citizens, connect to jobs, and generally affect social mobility

Report Outs Notes

1. **Impact and sustainability:**
    a. Where does DS in HS live? Integrated or stand-alone?
        i. Need to compensate for local variation in how things are taught?
        ii. Can we find ways to integrate DS into curricula in ways that help teachers teach those topics more effectively? For ex, introducing DS tools in History.
    b. DS vs Data Literacy
        i. The latter should also be thought of as a separate track, even if on a continuum
    c. Sustainability
        i. Need to work with policymakers, educators, etc.
        ii. Need to develop tools/resources that help teachers teach other topics
        iii. Work w/ early adopter teachers and incorporate peer mentoring
            1. Create local peer groups to share best practices and other lessons
    d. Impact?
2. **Development/Curricula**
    a. See above
3. **Training vs. Support**
    a. Training discrete; support continuous
    b. Who to work with?
        i. Teachers, TAs, substitute
        ii. Students as mentors/ambassadors
        iii. Coaches/Tutors
        iv. Self guided
    c. Methods to create
        i. Hack-a-thons
        ii. Lesson study models…teachers work w/ curricula developers
        iii. Reciprocal internships (b/t teachers & trainers?)
    d. What could this lead to?
        i. Training around content
        ii. How to integrate into other subjects and in general
        iii. Teacher Peer teaching
        iv. Local peer chapters (best practices)
        v. Targeted check-ins and cycles of learning w/ teachers
        vi. Develop pathways in non-DS curricula for introducing DS concepts
    e. How can we support this?
        i. How do we support teachers to teach what administrators expect?
        ii. How do we support teachers to teach what's effective for students?
            1. These are frequently not the same
    f. What is the ultimate learning goals?
        i. Not a lot of clarity

Questions/Comments after College Board discussion:

- Discussion about definitions of Data Literacy vs. Fluency. Comments about how the categories are less important than deciding on what students should be able to know and do at certain ages. Comparisons to literacy w/ regard to reading and writing or mathematics and how we teach these.
- By putting together lists can identify points in learning pathway to introduce DS
- Having teachers work in teams is a great idea (math, science, history, etc...working together)
    - Someone mentioned that their kid's school has an art integration coach to work with other teachers for how to introduce art throughout
    - Catherine: NYSCI use this model with NetSci High, to introduce network science into schools...but it depended on having the principal as champion
        - Question: How do we plan for churn in champions?
- Is Data Science standalone or integrated from AP perspective? Where does it need to live?
    - Both!
    - From industry perspective...not enough to learn how to build an ML model...not isolated conversation
        - DS is multi-dimensional and multi-disciplinary
- How do we account for and test for transferability of skills?
    - What do students apply DS skills to?
    - The list made in Development exercise is at the conceptual vs. content level and these should all be transferable skills
    - Comments on community outreach to address this
        - Engage local businesses w/ competitions
        - Engage parents
    - Transfer may also depend on context
        - What am I expected to do?
        - How am I expected to do it?
        - Do I have the right tools/resources to do it?
        - Idea is to avoid students saying "I only do this in math."
    - Evaluation of transferability is still fluid
- Concord Consortium posted Paradigm Shift document as another way of thinking about guiding principles

Appendix F: Promising Research Projects

1. **A bottom up frame for data science literacies**

Participants who worked on this:
Ian Castro
castro.ian@berkeley.edu
Victor Lee
vrlee@stanford.edu
Dorothy Bennett
dbennett@nysci.org
Michelle Hoda Wilkerson
mwilkers@berkeley.edu
Sara Vogel
svogel@gradcenter.cuny.edu

*** This provides a foundation for conceiving of data science integration into school subjects in culturally relevant/sustaining ways.

Curricular case studies of data science in use + social impacts of data

- What are great teachers already doing?
    - K-12
    - Undergraduate
- What student performances are exciting from a DS perspective?
- What's happening in informal education?
    - Citizen Science case studies
    - Local environmental studies (carbon footprint, energy use)

Community case studies - recognizing moments of data science conversations

- Field scan
- What are community groups doing with data?
    - Artists
    - Activists
    - Agriculture
    - Community health workers

Consensus building around cases

- Why did this conversation need data science?
- What would have made this a better conversation?
- What about these cases reflect elements of data science?
- Which of these cases are especially relevant to High Schoolers in computer science?
- How to tag and use cases to generate goals and objectives for lessons and units? (instead of fragmented one-objective-per-activity approaches)
- How can we build on these practices?

Examples: What do people in your family do? Where in students' lives does data play a role?

Repairing lottery machines. Fixing phone screens. Victor's example w/ diabetes data management. Fitness data. Dartmouth project where students track sleep, eating, study habits. Michael Horn - home heating and energy usage. Berkeley: Bio study where app is used to predict whether people will smoke to help them get over addictions. Social services that people sign up for youth sports. Thinking about documentation.

(Similar studies: Herb Ginsberg had a video repository to help develop "noticing" in mathematics)

(Hammer's Moore Foundation project)

Deliverables:

- Crowd sourcing cases
- Meta-review
- Repository of tagged examples
- Principles
- Moving towards a video case library

2. **DataWise Schools Project: A Curriculum Integration Collaboration between DS4All & STEMteachersNYC** Drafted by Yadana Nath Desmond, Program Director, STEMteachersNYC

Leveraging the extensive experience and strengths in teacher-led workshop development, delivery and implementation of STEMteachersNYC (STN) and the vision, network and expertise of DS4All, the DataWise Schools Project proposes to engage teams of teachers from five NYC middle schools in 2020-2021, in a foundational workshop and sequence of bi-monthly working sessions to co-design and implement a catalog of Data Literacy/Data Science (DL/DS)-integrated lessons, focusing on the power and practicality of enriching their existing curricula through integration of DL/DS concepts, skills, and tools. DS4All Northeast Regional Hub, in partnership with STEMTeachersNYC, a teacher-led, PD-providing PLC non-profit and network of 1400+ STEM teachers, will co-facilitate 1) two school-year listening sessions, 2) one weeklong summer foundational DL/DS training that generates a co-designed set of cross-disciplinary lessons, and 3) bi-monthly, targeted, PLC meetings throughout the school year, via DS4All hub and peer-led in-person workshops.

Outcomes

- Co-designed catalog of integrated middle grades lessons and associated maps of school-level integration of DL/DS across multiple subjects.
- Teacher teams empowered to develop and implement their own DL/DS integrated lessons, through direct support from and co-design with DS4All, STN and each other.
- Increase in student engagement and evidence of changes in learning across subjects.
- Reporting on what methods work best in supporting integration of DL/DS across subjects and grades.

Chronology

1. Two Listening Sessions (spring)
    - STN DataSTEM working group sessions co-hosted by STN and DS4All.
    - Invite interdisciplinary teams of teachers, both from within an individual school and from across several schools.
    - Teachers and schools share ways they are striving to integrate DL/DS. Program team fields program-related questions, and recruits for summer.
2. Summer "Foundation/Launch week in DL/DS" for participating interdisciplinary teacher teams:
    - DS4All and STN teams share and unpack examples of integrated lessons (cases, locally relevant datasets, student curated data generation, links to history, language, etc)
    - Using DS4All DL/DS progressions, DS4All and STN work with teachers on their own curricula, to identify entry points, and to co-design (project and place-based) material.

    - Mapping of integrated lessons per school (indicator of total student contact time with DL/DS).
3. Bi-monthly Co-Design Sessions (school year)
    - Participating teacher teams return to continue the co-design process, and to share successes and challenges. STN and DS4All serve as loci of support and expertise. As participating teachers become experts around DL/DS applications, they lead their own workshops for other teachers as relevant. Hubs also serve to track implementation and outcomes.

3. Data Science for Teachers

Goal: Gather Leads of Teacher Associations to Discuss Data Science Collaborations & Implications

Teacher associations provide:
1. Disciplinary frameworks (concepts and practices) and standards
2. Communities of practice
3. Resource repositories
4. Professional development

All of these elements are needed to scale and sustain the integration of data science in primary and secondary (K-12) education.

At the DS4All workshop, participants grappled with the mechanisms to include data science in the K-12 education system, considering an integrated approach where data literacy and analysis knowledge and skills could enhance social studies, mathematics, science, statistics, and other disciplines.

Why ESIP/RDA? Professional organizations play a role in supporting and developing teacher associations, especially associations with education special interest groups. As an example, the Computer Science Teachers Association began with support from ACM (Association for Computing Machinery).

Outcomes of the meeting include identifying:
- Overlapping areas for standards and concepts
- Strategies to collaborate across organizations to promote data science
- Policy supports and barriers to scaling data science
- Needs of teachers in various disciplines to learn and teach data science
- Repositories - how best to share data science curriculum and lesson plans
- How to leverage the communities of practice in each association OR create a separate association

Teacher Associations to Invite:
- ISTE (educational technology)
- CSTA (computer science)
- NSTA (science)
- NCTM (math)
- NCSS (social studies)
- NAA (afterschool)
- Sports - https://www.nfhs.org/

- Health Educators - Society of Health and Physical Educators (SHAPE)
- History Educators -
- Art, Music Educators - (?)
- Social Workers/Counselors - ASCA
- Principals & Superintendents
- State Department of Education (?)
- ALA (librarians)
- AAPT (physics)

After compiling societies that represent the above areas of education in the K-12 space, we will send an email request for participation from each of the societies for participation at the next ESIP Education Cluster meeting.

Potential Agenda Items:
- "Demythtifying" Data Science?
- Exemplars of data science integrations
- Use cases
- Story telling/ data driven journalism
- Data science integration across all disciplines: what are the challenges, how it can be done

Intellectual Merit: will advance knowledge about the DS integration across school curricula by identifying challenges and incentives

Broader Impact: Builds knowledge on integrating multy-disciplinary curriculum into existing school system

$50K-75K

4. **Creating Unstructured Data Science Education Drivers for the Big Data Hubs' Open Storage Network**

Multi-domain Data Distributed Active Archive Center (MD DAAC) for data science education: allows data to be submitted, lesson plans can be volunteered submitted (similar to ai experiment).

drfuka@vt.edu (Dan Fuka)
kirk.borne@gmail.com (Kirk Borne)
eyulaeva@ucsd.edu (Elena Yulaeva)
lumer@nysci.org (Laycca Umer)
ajay.anand@rochester.edu (Ajay Anand)
nhall@champlain.edu (Narine Hall)
jrosato@css.edu (Jen Rosato)
wjl88@cornell.edu (William Leon)

Motivation:

Most of the data science tools and principles in educational setting are geared towards structured data. *Data Science World is not Rows and Columns … Its Graphs, Signals, Images, Text.* Future

trends indicate data sources that are unstructured … signals, images and text will dominate. Data Analysis techniques for unstructured data may not readily leverage “tabular data analysis” methods

Project Goals:

- Build data repository of accessible data from various sensor types
- Develop education tools/lesson plans for unstructured data (e.g. representation of sensor data, sources, basics of sensor data manipulation, pre-processing, visualization); Develop familiarity navigating graph, image, signal data
- Use cases and Experiential learning: Hackathons to allow students to experience data collection and analysis
- How data from multiple sensors can enhance decision making

NPS - Tour Guides

Google Developers Group

EAGER - Capacity Towards a Data Literacy Educators Community

NY BOCES -> RIC Regional Information Centers

ISTE

National Data Science Competition for your local Data Science Club

Data from sensors

Appendix G: Post-Workshop Survey

Module 1 - Achieving Workshop Objectives: *"Please reflect on the 2-day workshop you have just completed and indicate your level of agreement with the following statements"* (a 5-element Likert scale was used to assess this module, *n* = 21).

Module 3 - Developing a Community of Practice: *"Please reflect on the long-term data literacy initiative as discussed in the workshop and indicate your level of agreement with the following statements"* (a 5-element Likert scale was used to assess this module, *n* = 21).

Module 1	Respondent	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
successful integrationI believe data science can be successfully integrated into high school teaching and learning	strong	1	1	1		1		1	1				1		1	1				1			1	11
	agree				1					1	1	1		1			1	1	1		1			9
	neither						1																	1
	dis																							
	strong dis																							
My understanding of challenges and opportunities to data science teaching and learning improved through this workshop	strong					1				1		1			1	1		1	1	1	1			9
	agree				1		1		1				1				1						1	6
	neither	1	1					1			1													4
	dis			1										1										2
	strong dis																							
I gained a greater understanding of the needs of teachers to enrich instruction with data science tools and resources	strong					1										1		1	1					4
	agree								1		1	1	1		1						1		1	7
	neither		1		1			1		1							1							5
	dis	1					1													1				3
	strong dis			1										1										2
I gained a greater understanding of the needs of students to enrich instruction with data science tools and resources	strong					1									1			1						3
	agree		1				1		1	1			1						1					6
	neither				1			1			1	1				1					1		1	7
	dis	1															1			1				3
	strong dis			1										1										2
I learned about valuable data science tools, data sources and education resources that I didn't know about before the workshop	strong								1			1	1		1		1			1	1			7
	agree	1			1	1	1	1		1				1		1		1	1					10
	neither		1								1												1	3
	dis			1																				1
	strong dis																							
Module 3																								
I believe i can make a valuable contribution to this initiative moving forward	strong	1	1			1			1				1		1	1		1		1			1	10
	agree				1		1	1		1	1	1		1			1		1		1			10
	neither																							
	dis																							
	strong dis																							
I plan to invest time and energy to make this initaive a success	strong	1	1			1							1		1	1				1				7
	agree				1			1	1	1	1	1		1			1	1	1		1			11
	neither																						1	1
	dis																							
	strong dis																							
Being a part of this initiative is an exciting opportunity for me	strong	1	1			1			1	1			1		1	1		1		1	1			11
	agree				1		1	1			1	1		1			1		1					8
	neither																						1	1
	dis																							
	strong dis																							
Being a part of this initative is an important professional opportunity for me	strong	1	1			1		1		1			1		1						1			8
	agree				1		1		1		1						1	1	1	1				8
	neither											1		1		1							1	4
	dis																							
	strong dis																							
This initiative is relevant to the success of my work in other areas	strong					1	1	1	1				1						1					6
	agree	1	1		1					1					1		1	1		1				8
	neither										1	1				1					1			4
	dis													1									1	2
	strong dis																							
I feel inspired by this initiative	strong	1				1				1			1		1	1		1		1	1			9
	agree		1		1		1	1	1		1	1		1					1					9
	neither																1							1
	dis																						1	1
	strong dis																							
I believe this initiative will positively impact my institution / organization	strong	1				1							1		1	1		1		1	1			8
	agree		1		1		1		1	1	1	1					1		1					9
	neither							1						1										2
	dis																						1	1
	strong dis																							

Free Response Question (respondents 1-11) "*What value do you feel you would bring to a long-term Data Science for All initiative*?" (Note: the word "value" here can be interpreted broadly to include skills, relationships, and resources).

Respondent	1	2	5	6	7	8	9	10	11
Module 2									
comments on quant section		I think this topic and effort is important and essential. I have done a good degree of thinking and work in these directions to date and am not sure the workshops per-se was organized or included the best mix of people or new perspectives to bring significant new insights			i feel already exposed to challenges and opportuinites for teachers and students			it was nice to have time to learn more about how others are approaching data science at the pre-college level. Would have liked to have seen more "on the ground" examples (eg what does it look like for a student to know/do DS?)	beginning with student/teacher examples from end of day 1 and grounding us in cases may have helped with #4
Do you believe that the data science is important to computer science education? if so, why? If not, what ranks as more important?	yes! data science is a far more accessible entryway to programming and structured data	I'm not sure still. I think computing is important to data science education, but think that data science and data concepts are separate and more wide-ranging and important than computer science alone.	it is shown that the need for data scientists will only be increasing in the next few years	**think** that we did explore more specifically what data science core concepts, preactives, skills are - but less about where teachers will need support, issues of integration still and open questions	sort of. I can see this going with math or statistics. it seems valuable to add to CS ed.	yes: data science is a key offshoot of computing education	I think I still have questions about this. I think I;m biased, but would lean towards data science being more important...but also don;t think of them as being distinct/either-or	nor sure it should be tied to CS education conceptually	yes! i think that "importance" however, should be determined contextually and that priorities within CS for all should be weighed by the localities and communities implementing them.
What do you believe is the most important outcome of a multi-sector collaboration among educators, academia, industry, and government in data science traching and learning	improving citizenship and civic engagement	A wholesale recognition and awareness of the importance of k12 data education and the requisite policy and practice changes to bring the awareness of need into reality	school/life connections for students - real life input for developers - cross pollination of ideas and deeper understandings	data science is important for all the disciplines - do believe data is pervasive and educators need to have this be a critical area in the disciplines. Doesn't necessarily need to live in CS but would elevate its imprtance/value if a separate course.	a framework or some standards for what should be taught and **barred** when	regular meetings that pull together computation + statistics education in the context of data issues	a collective vision/set of priorities that supports social equity and sccess to economic opportunities	funding for research. raise visibility. coordinate efforts + share experiences	understandings about what data science is, how it fits into CSforAll, and the many ways it can look in schools and localities
Module 3									
What value do you feel you would bring to a long-term Data Science for All initative? (note: the word "value" here can be interpreted broadly to include skills, relationships, and resources)	deep experience scaling integrated computing across states, grade levels and subjects (curriculum writing, tools building, and training experience as well!)	I and my institution will bring both perspectives and networks to this work as well as tools and insights. This work is very important across sectors	on the ground experience teaching data science in high school already - other online experience prior to teaching high school	new cases, better alignment with the practice of data science	research background, some potential for relationships with schools, universities		experience/lessons learned from working to define data literacy/science knowledge + skills as well as design learning experiences to support their development	attention to student knowledge + experience; curricular co-design experience	support for planning DS initiatives in localities, especially in schools serving large portions of emergent bilingual students

Free Response Question (respondents 12-22) "*What value do you feel you would bring to a long-term Data Science for All initiative*?" (Note: the word "value" here can be interpreted broadly to include skills, relationships, and resources).

Respondent	12	13	14	15	16	17	18	19	20	22
Module 2										
comments on quant section	interesting workshop and good spread of attendees from industry/academia.etc. I hope to continue this work!	the workshop did not prioritize teacher or student perspectives and that is GOOD because if practicioners or experts can't agree on core content, there's no room for student perspectives & minimal room for teacher perspectives, or what needs	it was all great. It would have been a bit better to hear from a couple more HS teachers. Nevertheless, it was a fantastic workday.	great meeting with colleagues I haven't met before	i think we needed to hear more from teachers to understand where students struggle	i think there is room for some data science intro in HS, but the depth of knowledge still remains unclear	it will take time to intergrate data science into the curriculum. there remains a lot of work to be done by educators	we spent more time discussing the "field" than specific needs from teachers or students		I knew about a lot of the resources & was happy to learn more about Bootstrap
Do you believe that the data science is important to computer science education? if so, why? If not, what ranks as more important?	no; i think DS is more important. we spent a lot of time discussion the importance of interesting DS to other fields curricula **ovrd?** I think that makers DS more useful to science ed overall.	data science may be a very accessible onramp to computing, but there is a real risk that if we're not careful, parents/school districts may force a focus soley on DS to the exclusion of CS	it is important in CS ed, but also in "everything else" ed. What is very important to both is cybersecurity	yes, as it makes the CS education more valuable with students have a more marketable skills	yes, closely linked to big idea/knowledge areas in data and programming. Data science majors use a number of CS courses which will have impacts	yes, CS is an important tool for exploring, analyzing, using, explaining + developing new knowledge with data	coding is required for a practicing data scientist. some parts of data literacy are related to CS concepts (ex. hardware for sensors, infrastructure for data, computational time for analysis)	data science is a multidisciplinary field with some components of computer science education at its core. computer science is important to data science, but data science is not a subset of computer science		I believe data science is important PERIOD. I don't think of it as part of computer science
What do you believe is the most important outcome of a multi-sector collaboration among educators, academia, industry, and government in data science traching and learning	data literacy across fields + workers	(a) appreciation for the FACT that this is multidisciplinary & one focus won't solve the problem. (b) refusing to reinvent any wheel is empowered if you can see/recognize what other groups have already discovered.	more employable job candidates, more jobs filled, higher employability across all demographics	enhanced value of students capabilities	Define the discipline, identify what k-12 needs to know -> need framework w/ concepts, practices, standards	I'm not sure this is most important but certainly important	data ready workforce, informed citizens	development of standards structured into what most people should know, and what a smaller percentage should know	opportunity to make DS education a platform for producing social mobility through education**?	some agreement on foundational concepts in data science so we can proceed with research & development
Module 3										
What value do you feel you would bring to a long-term Data Science for All initative? (note: the word "value" here can be interpreted broadly to include skills, relationships, and resources)	industry course curricula w/ a focus on DEI?	two main contributions: a focus on applied ethics + professional practice & espertise in assesment & ensuring that instruction & assessment are aligned	years of experience in academia + industry + government research labs, _always_ working with data!	establishing cross domain scientific collaborations	k12 and teacher ed/PD perspective as well as an understanding of ed policy and change efforts		my team is building curiculum to be released in education targeted towards data scientists and data workers	several years of work around skills in data science at the workforce level	relationships w/ schools + teachers	35 years of research and development and real experiences with students & teachers

Appendix H: Resources

Programs

For recent high school graduates in East Baltimore: https://www.clouddatascience.org/

National Center for Women in Information Technology: https://www.ncwit.org/

West Big Data Hub high school census data competition:
https://www.letsmakeitcount.org/

WiDS Datathon 2020

Models

K-12 CS framework: https://k12cs.org/framework-statements-by-grade-band/

IBM Data Science Skills Competency Model

Ocean Literacy

SCRATCH computational thinking definition: https://scratched.gse.harvard.edu/ct/defining.html

Undergraduate level

ACM Task Force: Computing Competencies for Undergraduate Data Science Curricula http://dstf.acm.org/DSReportDraft2Full.pdf

NASEM 2018 "Data Science for Undergraduates" definitions of data acumen: https://nas.edu/envisioningds

Open Source Data Science Curriculum for Universities
From IBM, UPenn and The Linux Foundation
https://community.ibm.com/community/user/datascience/blogs/ana-echeverri1/2019/09/19/data-science-for-all-an-open-source-approach-to-ed?CommunityKey=f1c2cf2b-28bf-4b68-8570-b239473dcbbc&tab=recentcommunityblogsdashboard

Responsible Data Science course at NYU
https://dataresponsibly.github.io/courses/spring19/
ask Julia Stoyanovich (stoyanovich@nyu.edu) for homeworks / solutions; all slides, reading, labs (python) are online. Will be updated for Spring 2020.

Calls for Submission

NSF CS4All solicitation

Call for Submissions: workshop at IEEE VR on Education and Learning with Virtual and Augmented Reality called KELVAR.

Teacher Education (preservice and inservice)

CS Visions Project https://www.csforall.org/visions/
Includes: Whitepaper, online quiz to support surfacing values, activities for groups and PD

ESTEEM project at NC State (from Hollylynne Lee). Enhancing Statistics Teacher Education through E-Modules (NSF IUSE funded). https://hirise.fi.ncsu.edu/projects/esteem/ Modules available for free to use in courses for preservice teachers--designed to easily integrate with 3 LMS (Moodle, Blackboard Canvas). Access here

Two MOOCs for educators on Teaching Statistics through Data Investigations and Teaching Statistics through Inferential Reasoning. Both courses have a data-heavy focus on aim to get teachers to integrate easy to use tools (like CODAP): https://hirise.fi.ncsu.edu/projects/online-pd/ (note, though many participants are HS and CC math/stats teachers, teachers of science, social sciences, and middle school also have taken these MOOCs)

New DRK12 project on Invigorating Statistics Teacher Education through Professional Online Learning https://hirise.fi.ncsu.edu/projects/instep/.

Reading

NYC Automated Decision Systems Task Force report Note multiple recommendations that pertain to education, in particular public education / public engagement.
https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf

Recent Special Issue on Data Science Education in *Journal of the Learning Sciences*)

Research paper from SIGCSE proceedings (2017); Vogel, S., Santo, R., & Ching, D. (2017). Visions of Computer Science Education: Unpacking Arguments for and Projected Impacts of CS4All Initiatives. Proceedings of the 47th ACM Technical Symposium on Computing Science Education. https://doi.org/10.1145/3017680.3017755