2018

Keeping Data Science Broad

Negotiating the Digital and Data Divide Among Higher-Education Institutions

A report summarizing a series of webinars and workshops to garner community input into pathways for keeping data science education broadly inclusive by bridging the digital and data divide among higher-education institution types.

Dr. Renata Rawlings-Goss

South Big Data Innovation Hub

Keeping Data Science Broad: Negotiating the Digital & Data Divide

January 16, 2018

Sponsored by the National Science Foundation, the South Big Data Innovation Hub, and the Institute for Data Engineering and Science at Georgia Tech

Authors

PI: Renata Rawlings-Goss
Co-Executive Director, South Big Data Hub
Georgia Institute of Technology

Lillian (Boots) Cassel
Professor and Chair, Department of Computing Sciences
Villanova University

Melissa Cragin
Executive Director, Midwest Big Data Hub
University of Illinois

Catherine Cramer
Senior Program Developer
New York Hall of Science

Angela Dingle
Consultant
University of the Virgin Islands

Shawnta Friday-Stroud
Dean, School of Business and Industry
Florida A&M University

Al Herron
Assistant Director of IT Department
The Breakfast Group

Nicholas Horton
Professor of Statistics
Amherst College

Tasha R. Inniss
Director of Education and Industry Outreach
INFORMS

Kari Jordan
Director of Assessment
Data Carpentry

Patti Ordóñez
Associate Professor
University of Puerto Rico Rio Piedras

Mary Rudis
Assistant in Instruction
Bates College

Robert Rwebangira
Assistant Professor
Howard University

Karl Schmitt
Assistant Professor of Mathematics and Statistics
Valparaiso University

Dale Smith,
Consulting Data Scientist

Sonya Stephens
Interim Dean, College Of Science And Technology
Florida A&M University

Acknowledgments

This document was developed through the contributions of the participants and organizers of the Keeping Data Science Broad Series. A special thanks and appreciation to Beth Russell, Costa Michailidis, and Donnalyn Roxey for their help in planning and facilitating the webinars and workshop. Additionally, thank you to Jennifer Salazar, Kyla McMullen, and Illona Sheffey-Rawlings as well as the core team of editors, writers, and reviewers of this report. Finally, we truly appreciate the program committee and

participants involved in contributing input through the community input form or the writing of white papers as a part of the workshop. (See full list in the Participants Section). NSF Award Number: 1747961

Table of Contents

Executive Summary

The goal of the "Keeping Data Science Broad" series of webinars and workshops was to garner community input into pathways for keeping data science education broadly inclusive across sectors, institutions, and populations. Input was collected from data science programs across the nation, either traditional or alternative, and from a range of institution types including community colleges, minority-led and minority-serving institutions, liberal arts colleges, tribal colleges, universities, and industry partners. The series consisted of two webinars (August 2017 and September 2017) leading up to a workshop (November 2017) exploring the future of data science education and workforce at institutions of higher learning that are primarily teaching-focused. A third follow-up webinar was held after the workshop (January 2018) to report on outcomes and next steps. Program committee members were chosen to represent a broad spectrum of communities with a diversity of geography (West, Northeast, Midwest, and South), discipline (Computer Science, Math, Statistics, and Domains), as well as institution type (Historically Black Colleges and Universities (HBCU's), Hispanic-Serving Institutions (HSI's), other Minority-Serving Institutions (MSI's), Community College's (CC's), 4-year colleges, Tribal Colleges, R1 Universities, Government and Industry Partners). This series is co-sponsored by the National Science Foundation - Directorates of Computer & Information Science & Engineering (CISE), Mathematical and Physical Sciences (MPS), Education and Human Resources (EHR), and Social, Behavioral and Economic Sciences (SBE); the South Big Data Innovation Hub; and the Georgia Tech Institute for Data Engineering and Science.

Webinar 1: Data Science in the Traditional Context

The initial webinar highlighted universities, teaching institutions, community colleges, HBCU or other minority-serving institutions that have implemented data science undergraduate programs as case studies for workshop participants to consider and compare in their own contexts. In order, the speakers were:

- Renata Rawlings-Goss, Georgia Tech (Moderator)
- Paul Anderson, College of Charleston
- Mary Rudis, Great Bay Community College
- Karl Schmitt, Valparaiso University
- Pei Xu, Auburn University
- Herman "Gene" Ray, Kennesaw State University

Discussion included how institutions combated (1) being an unknown entity, (2) recruiting, (3) designing an effective capstone (4) performance gaps of non-white, female and first generation college students (5) differentiating similar programs and (6) recruiting qualified faculty. Webinars are archived and available for public viewing on the South Big Data Hub YouTube Channel.

Webinar 2: Alternative Avenues for Data Science Education

The second webinar highlighted efforts that built data science education capacity outside of the context of traditional curricular program development, illustrating alternative means of spreading data literacy and just-in-time skills. In order, the speakers were:

- Renata Rawlings-Goss, Georgia Tech (Moderator)
- Lior Shamir, Lawrence Technical University
- Tracy Teal, Data Carpentry
- Stephen Uzzo, New York Hall of Science
- Al Herron, The Breakfast Group
- Sarah Stone, Data Science for Social Good

Discussion included integration of data science into courses and curricula outside of the traditional CS/math/stats context (i.e. Arts and Humanities) through (1) expansion of capacity by integration of third party or shared resources (i.e. MOOCs and open source educational resources) into curricula, and (2) adding additional educational options outside of traditional courses (i.e. Faculty training, Data Science for Social Good Programs, and Bootcamps). Webinars are archived and available for public viewing on the South Big Data Hub YouTube Channel.

Webinar 3: The Big Picture for a Big Data Science Education Network - Next Steps

The third webinar was the wrap-up to the six-month Keeping Data Science Broad Series (August 2017 – January 2018). It gave an overview of the previous webinars, workshop, community report, and next steps, as well as highlighted specific projects and working groups that participants could get involved in for the future. In order, the speakers were:

- Renata Rawlings-Goss, Georgia Tech (Moderator)
- Nicholas Horton, Amherst College
- Lillian (Boots) Cassel, Villanova University
- Yuri Demchenko, Universitat van Amsterdam
- Melissa Cragin, Midwest Big Data Hub
- Catherine Cramer, New York Hall of Science
- Karl Schmitt, Valparaiso University

Discussion included new and upcoming South Hub Education Workgroup Projects, a summary of this workshop report as well as domestic and international project groups. Webinars are archived and available for public viewing on the South Big Data Hub YouTube Channel.

Workshop: Negotiating the Digital and Data Divide

Building upon the case studies and discussions at the two webinars, an interactive workshop was designed for over sixty participants from data science programs across the country. The goal was to enable researchers and educators from primarily teaching-focused institutions, (i.e. community colleges, four year liberal arts colleges, and minority serving institutions) to understand and begin to address questions about how best to prepare institutions to teach the data science students of 2025. Participants were encouraged to highlight challenges related to capacity building and capability within their institutions or disciplines, and also to systematically catalog ideas regarding what a vision of a bright future would look like for data science. Participants formed writing groups and then shepherded ideas from individual thoughts to clustered topic areas into white papers. From those efforts, twenty-nine white papers were produced during the two day workshop. Following the workshop, an editing team consisting of workshop participants and South Hub staff, distilled from the twenty-nine white papers eight combined theme areas that had emerged as both a current challenge and an area ripe for envisioning the future. These Linked Challenge-Vision areas are presented in more depth in the following report and are a combination of input from the webinars 1 and 2, the twenty-nine workshop white papers and notes, as well as a community input form hosted by the South Big Data Hub open for public comment for a period of six months (August 2017 through January 2018).

Linked Challenge-Vision Areas:

1. **Access to Data** discusses the challenge of providing data and how to accomplish data availability across a broad group of stakeholders.
2. **Assessment & Evaluation** addresses the broad need for credentialing, assessment, and evaluation when it comes to data science programs.
3. **Curriculum** takes a high level look at data science curricula and topics such as experiential learning, innovations in teaching, and data science program goals.

4. **Data Literacy** tackles the definitions within data science, infusing data science into non-STEM courses, and the link to general critical thinking skills as well as student needs.
5. **Diversity, Inclusion, and Equity** highlights the clear need for continued focus on including the broadest groups of the population in opportunities surrounding data, including workforce concerns and access to technology for all in order to increase an American talent pool.
6. **Ethics** outlines the ethical concerns as well as social good stemming from data and its products.
7. **Faculty, Staffing, and Collaborations** recognizes the needs of faculty and staff to support new efforts as well as institutional barriers and opportunities calling out the need for cross-disciplinary and cross-sector collaboration to make the data science pipeline viable.
8. **The Pipeline to Higher Education** approaches taking a high level view of the full data science pipeline from general education and K-12, to undergraduate education, graduate education, and professional worker "upskilling", with a focus on the pipeline from two-year colleges to four-year colleges.

Top 10 Asks from the Community:

As a final step, participants were asked to give the top "asks" they would make of the data science community to ensure a bright future for data science education. The group then voted on each ask, resulting in the top ten asks, or next steps, listed below.

1. Foster partnerships between different institutional types i.e., 2-year and 4-year college partnerships, HBCU, R1, industry and alternative groups.
2. Provide flexible pathways into data science education.
3. Time & space to discuss collaboration, especially with respect to curriculum "holes."
4. Hiring female faculty, faculty of color, and female faculty of color because it's hard for students to "become something they have never seen".
5. Provide free (or subsidized) access to data science resources.
6. Provide access to data literacy tools and resources to students and parents of underrepresented backgrounds/populations/communities.
7. Supply access & training for JupyterHub in data science instruction.
8. Provide examples of curriculum for 2-year colleges degrees, certificates or pathways.
9. Provide more realistic data science-focused collaboration between industry and K-12.
10. Develop data literacy resources for K-12 teachers.

Keeping Data Science Broad: Negotiating the Digital & Data Divide

1. Introduction

Data science, according to many, can be defined as knowledge discovery from data[1]. If this is unpacked, it starts with data itself. There is biological data – in the form of DNA, cells, organs and systems within organisms; and chemical data in the form of elements and molecules, isotopes and reactions, form and substance. Physical data is in the form of phenomena such as gravity, electricity, light and sound, heat and motion. In the information age, we now have computerized, digital data. Massive quantities of bits and bytes pass back and forth from devices in our cars, schools, homes, places of worship, social spaces and commercial spaces, to warehouses of stored information. Beyond the physical universe, data is anything that can be measured, quantified or identified with an individual person, place, thing (an object or an idea), event, transaction or phenomena in such a way as to provide means of distinguishing one from another, to make predictions, analyze outcomes, automate a process, or to shape policy and make decisions.

Knowledge discovery is the process by which data is collected, translated when necessary into usable information, aggregated, analyzed and combined with more data which is then refined into even more usable information until enough insight has been gained so that the question is answered or a discovery is made and a goal has been achieved within reasonable margins of error.

As such, data science is broad. Currently, it requires knowledge and skills as diverse as computer science, statistics, mathematics, ethics, and communications, as well as domain knowledge in applications ranging from logistics, retail, fast food, and the arts, and reaching into sectors as disparate as government, smart cities, policy, to finance, biotechnology and healthcare, as well as to energy, and transportation.

Popular books and media are also making the concepts behind statistics and predictive analytics more and more accessible, not only to those in business making decisions everyday, but also to the average man or woman on the street. The onus in many ways is on the community to maintain this broad diversity of interests, sectors, topics, and people into the future. What makes data science so relevant is not only its unparalleled ubiquity and enormous scope, but its potential to improve life and decision-making across a wide variety of areas.

Consequently, as data-driven decision-making becomes more commonplace, having the skills to understand and make sense of data can provide a sense of power to the larger citizenry or conversely powerlessness to communities without these skills. This "Data Divide" separates communities that have access to devices and services that provide rich, data-driven services from those that don't; it separates data-savvy individuals, and communities that have understanding and awareness of how their data is being collected and used to provide individualized services (and thus informed protections), from those that do not. The economic and social consequences of the Data Divide stratify populations, and severely limit the opportunities of those who are unable to take advantage of the data revolution.

The potential impact of the Data Divide is no less dire for our institutions of higher education. Here, a significant chasm separates research universities that are already developing comprehensive data science

[1] Data Science. (n.d). In Wikipedia. Retrieved January 8, 2018 from https://en.wikipedia.org/wiki/Data_science

programs at the undergraduate and graduate levels from primarily undergraduate-focused institutions (e.g., teaching institutions, community colleges, Historically Black Colleges and Universities (HBCU's), Hispanic-Serving Institutions (HSI's), other Minority-Serving Institutions (MSI's)), which are mostly absent and unheard in the movement to develop data science programs. In some cases, such institutions lack the necessary capacity, capability, or resources to develop Data Science programs. These limitations stymie the ability of many types of undergraduate-focused institutions to develop data science-inclusive curricula in a comprehensive way. Comprehensive understanding of the factors that limit data science educational program development, the development of creative and new approaches that integrate data science education into different types of institutions, and a supportive community of practice to enable experimentation with potential solutions are required to enable all institutions to bridge the Data Divide and to provide their students with the opportunities that data science education enables.

Here, we bring together the focused perspectives of over 50 different research institutions, teaching institutions, community colleges, HBCU, HSI's, and other minority serving institutions interested in supporting data science education on their campuses. We also include perspectives from industry, government and the nonprofit sector on challenges and visions for workforce and education. The goal is to add missing voices to the conversations envisioning the data science discipline and broader education for the benefit of our indivisible nation.

2. Activities

The webinars, survey, and workshop on Negotiating the Digital and Data Divide explored the Data Divide issue by collectively bringing together and garnering input from over a hundred stakeholders at teaching institutions, community colleges, tribal colleges, HBCUs, HSI's, and minority serving institutions experiencing the divide as well as industry, government, non-profit, and large research universities. Discussed were current programs, challenges related to capacity building, and the capability to envision a bright future state. These discussions also addressed developing data literacy and Data Science acumen; training data science practitioners, and translational data science. A mixed-format of virtual participation in webinars, an online survey, and a two-day in-person workshop event allowed this series to maximize the in-person time spent on discussion and to ensure that participants came to the table well informed of the potential challenges and issues. The goal of the workshop was to enable researchers and educators to understand and to begin to address questions on how to best prepare institutions to teach the data science students of tomorrow and to prepare them for the "data-intensive" and "data-enabled" society, economy, and job market of 2025 and beyond.

Key Topics included:

- What do data science training programs look like for underrepresented institutions?
- Are the challenges similar or unique to the challenges at larger institutions?
- How do we envision future progress taking place?
- What resources may be needed to further strategic goals?

Two pre-workshop webinars were held by the South Big Data Innovation Hub to introduce the "Data Divide" topic and as a forum to bring in speakers that will prime the in-person discussion. The in-person workshop, held in an unconference style to engage the community in high-quality discussions, garnered input on the unique challenges faced by underrepresented institutions in the discussion around data science education. Community members were asked to comment on the Data Divide topic via a community input survey that was open to the public.

During the two-day workshop, attendees completed twenty-nine individual 1-2 page papers on different data science education challenges or vision areas. These white papers were jointly written by teams

of conference attendees and discussed the challenge or vision topics deemed most important by community vote, the stakeholders involved, the impact of addressing this need on the national landscape, as well as initial concrete steps for action.

Authors volunteered to work for 8-weeks after the workshop to combine the 29 white papers into nine overarching theme sections and continue writing on these combined themes. This formed the basis of the final report. One post-workshop webinar "The Big Picture for Big Data Science Education" was held in January 2018 to follow-up with the community on workshop outputs, the report, and next steps, as well as highlight specific projects and working groups that participants would like to start or involve the community in today.

3. Linked Challenge-Vision Areas

Participants identified challenges related to capacity building and capability within the discipline. They also catalogued ideas comprising overarching visions of data science. Compatible topics were merged into combined challenge/vision theme areas, all linking challenges to the enactment of these visions. This section summarizes the eight linked "challenge-vision" areas.

3.1. Access to Data

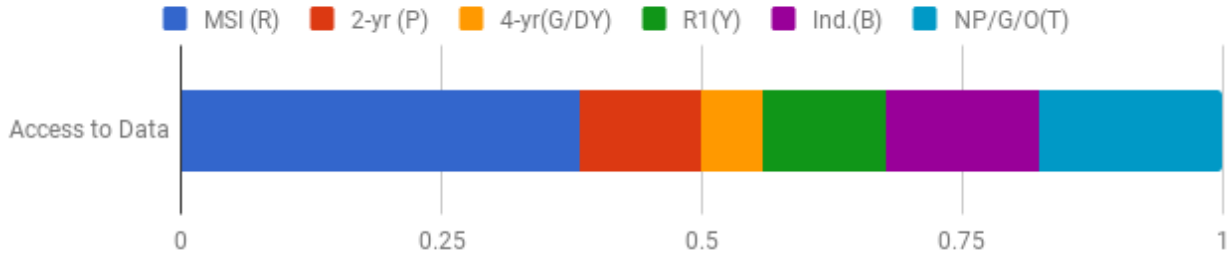


Figure 1: Ratio of workshop participant contributions to this section by institution type (n=34)

3.1.1. Introduction

There are two distinct aspects to "data access." The first concerns the ability to locate data that are useful and usable. The second concerns the skills needed to acquire, assess, and use the data. Given that we are living in a new era in which data are part of the social and economic fabric of daily life that drives our interactions with online services, traditional services, and the evolving digital economy, both aspects of "data access" are critical.

With the evolving democratization of machine learning and artificial intelligence and access to public data sets, ordinary citizens have opportunities to ask questions, conduct studies, and gain answers relevant to their own lives. Datasets are being generated at an unprecedented rate through social media and the tech industry, "smart cities" initiatives, via huge systems like energy and water delivery and management systems, transportation, public health, and the education enterprises. The "Internet of Things" is bringing data-enabled decision-making and management into daily home life, onto the manufacturing floor and other job sites through industrial tools and industrial scale 3-D printers, among others.

3.1.2. Stakeholders

Stakeholders include students, faculty, industry partners, non-profit organizations, and government agencies. Curriculum designers are also stakeholders. Academic IT departments also have a stake, as they may be called upon to provide cyber-infrastructure (e.g. sandbox or storage services) or research computing support, either through campus clusters or data centers, cloud environments, or via consortia services.

3.1.3. Broader Impacts

Access to datasets is required to give all students a broad understanding of how and when data are collected, managed, used, monetized, and disposed of. Applications of various data types are required for training and research in all disciplines, including data science, business analysis, artificial intelligence, statistical analysis, journalism, public policy, digital humanities, and most other fields of study. While statisticians have historically been the group most interested in the theoretical characteristics of data, individuals from other disciplines have been more concerned with generating their own datasets, or using datasets unfamiliar to the statistics community at large. Some datasets are generated in entirely new ways, and at a frequency of a single data point per second, or less. Nevertheless, access to data is necessary for all disciplines.

Although access to data is necessary for all disciplines, student and faculty access to data is problematic in some disciplines. For instance, health care claims data is not readily available at the patient level for students and faculty, although the Center for Medicare Statistics does provide some publically available data. Other datasets are expensive and contain contractual provisions that impact who may see and use the data, and how it is stored.

Other access to data issues are the volume and frequency of data. These issues complicate curriculum design, which can lead to the inclusion of datasets as an afterthought. For example, when a homework or project assignment is given at every educational level, many times not much thought is given to the availability of data.

3.1.4. Challenges and Visions Pertaining to Access to Data

An overarching conclusion from the workshop was that access and usability of good quality datasets was a barrier to teaching. Existing datasets have been, in many cases, thoroughly mined for innovative ideas by students and researchers. In some cases, researchers are unable to publish their data due to legal, compliance, cost, or privacy concerns. Industry is under increasing pressure from regulators and the public to avoid hacking breaches that disclose PII (Personally Identifiable Information) or PHI (Personal Health Care Information). Data from industry is expected to be harder or impossible to access going forward. Thus, some critical barriers are:

- Datasets require a lot of pre-processing
- A lack of interoperable datasets
- Locating available datasets
- Actual retrieving processes
- Lack of access to local or culturally relevant data
- Issues of ethical or legal concern in accessing and using datasets
- Data quality

Understanding the provenance of datasets is critical for curriculum designers and teachers using readily available data. This cannot be left as an afterthought, or as part of quickly assigning a project; metadata and documentation of the origin and processing of data must include details such as:

- How the data were collected or generated, including assumptions;
- Processing methods, including methods for missing data points;
- Intellectual Property information, including ownership, licensing, and any restrictions on use;
- A persistent identifier;

Lack of these kinds of information can have a big impact on which datasets are used.

For instance, inflation reported by the United States and China are measured in different terms, and are not readily comparable. Also, United States inflation series have hedonic adjustments applied[2]. The World Bank, International Monetary Fund, and United Nations specify how individual nations should report data to them, and it's helpful to read and understand the methodology. These are examples of only a few challenges faced with datasets.

Cleaning datasets is also a particular challenge. The process must be designed so it fairly reflects the underlying processes and purposes of the analysis. As one colleague warned us, "Sometimes the algorithms find features of the data that are really just features of the preprocessing!" Training, care and experience are required.

The single most important best practice for cleaning datasets is that the process must be reproducible. Make a read-only archive of the original, raw dataset. Clean the data using scripts or programs written for the purpose, and make sure the scripts are well documented. Write up the methodology used so it's easy for someone to read the code. Additionally, it's helpful to ask whether your data is sparse, or just scarce. The answer to this question will influence how the data is cleaned as well as the downstream steps in the modeling pipeline[3].

3.1.5. Specific skills and Resources

As an emerging best practice, academic units should have access to resources of an experienced data scientist who reports to a Dean, and is available to all interested faculty. The Goizueta School of Business at Emory University is moving to this model. This vision allows for some centralization of data storage, facilitates cleaning data which may be a poor use of faculty or student time, and serves as an internal consultant to assist faculty with not just teaching, but research projects as well.

The Federal Reserve Bank of St. Louis and data.gov have operated the FRED data portal for at least twenty years[4]. The ALFRED data portal, also available via the Federal Reserve Bank of St. Louis, offers another unique look at data collected by the government, which is often useful and not well understood,

[2] It is expensive to rebase the Consumer Price Index (CPI) since it's based on a basket of goods commonly purchased by consumers. Thus, statisticians attempt to adjust the CPI to reflect changes in the basket of goods whose prices are collected for the CPI.
[3] Sparse vs. Scarce, Tamara Kolda, November 2017, Retrieved January 10, 2018, http://www.kolda.net/post/sparse-versus-scarce/
[4] Federal Reserve Bank of St. Louis. (2018) Fred Economic Data. Available from: https://fred.stlouisfed.org/

even in industry. Twitter offers a developer portal[5], which allows use of the social media platforms data, and is a useful way to enrich other data sets. State governments are also useful sources of publically available data. The Centers for Medicare and Medicaid Services[6] is one of the few publicly available sources of health care data. The World Bank, International Monetary Fund, and the United Nations are also world class sources of data.

Users of Federal and State government data should be aware that historical data may be adjusted, corrected, or filled in. Data collection pipelines are often altered, but whether users notice or need to re-do their analysis is rarely raised as an issue. These adjustments may happen years into the future and it's up to the user to be aware of this concern.

Another key problem with Federal and State government data are adjustments to the data, whether they are seasonally adjusted or the hedonic adjustments now applied to the Consumer Price Index. Data collection methodologies have changed over time as well. We note particularly that the inflation data most used for the United States, which is the "All Urban Consumers" series, may well have bifurcated from non-urban inflation in the United States, as it currently is in India and China, due to changes in the U.S. economy.

Bioinformatics is the computational study of biological systems, including the genome, proteins, enzymes, and how they interact. Building proteins, either from a genomic sequence or otherwise, is a topic of current research and has matured enough to move into startups. Wikipedia has an extensive list of bioinformatics databases available at https://en.wikipedia.org/wiki/List_of_biological_databases.

The resources listed here are examples only. And their availability obscures a deeper problem. Pre-processing data is a challenge, even in well-funded research and industry groups. Archiving cleaned data benefits the entire community, as long as the data processing and cleaning process is reproducible. See the Challenges and Visions section above for further information.

3.1.6. Concrete First Steps

Cleaning data is essential to building high-quality models that generalize well with out-of-sample data. This task is tedious, error-prone, and not always the type of intellectual challenge that appeals to faculty. Emerging ideas, such as using machine learning techniques to clean data, are interesting and useful projects for undergraduates and graduate student theses and dissertations. They also serve to illustrate ethical, legal, privacy, and compliance concerns as well, which furthers the goal of integrating data science ethics in ongoing classroom discussions.

Graduate and undergraduate students have different needs for data. In the latter case, it is helpful for students to reproduce the results of others, as closely as they can, to enhance learning. For the former, unique datasets are more important since many publically-available datasets have already been thoroughly mined for insights.

Access and use of datasets is intertwined with ethical, legal, privacy, and compliance concerns. Faculty and students should be aware of issues surrounding the purposes for which data is collected, how the data was collected, where it can be stored, and issues of security of data at rest and data in transfer.

[5] Twitter Developer Portal. (2018) Twitter. Retrieved from: https://developer.twitter.com/
[6] Centers for Medicare and Medicaid Services. (2010). Research, Statistics, Data & Systems. Retrieved from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html

Even when data is readily available, the next question is to determine what variable will be modeled. This step involves tagging data for supervised learning. While unsupervised learning models may be used in this step, they are not as familiar to data scientists in industry or academia. Semi-supervised learning has its own problems, since a model that can tag your data is probably what you want in the first place. Anomaly detection is particularly prone to problems, as anomalies are hard to detect by non-domain experts. There are emerging trends, as outlined in a recent O'Reilly podcast.[7] Data scientists at all levels should be aware that the field is changing rapidly in this area.

In addition, there is a need to identify best practices and procedures for dealing with data that has various racial or other biases. These issues will be further discussed in the Ethics section.

3.2. Assessment & Evaluation

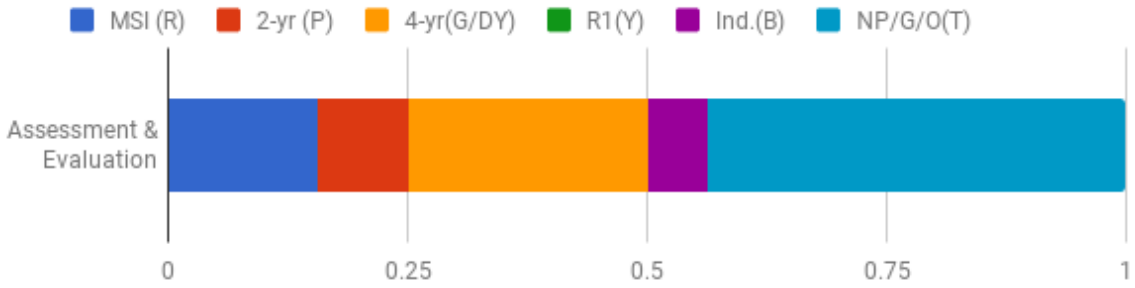


Figure 2: Ratio of workshop participant contributions to this section by institution type (n=32)

3.2.1. Introduction

Assessment and evaluation of data science programs require a systematic process whereby data are collected and analyzed to make informed decisions for all of the program's stakeholders. ***Assessment*** is defined as the estimation of the nature, quality, or ability of something (or someone). ***Evaluation*** is defined as making judgment about the amount, number, or value of something. These definitions frame the narrative around the challenges acknowledged for assessing data science programs, and the vision for what a successful assessment and evaluation strategy looks like, in an effort to bridge the digital divide.

Tools and methods to assess data science should be separated by teaching outcomes and learner outcomes. Types of assessment methods include project-based assessment, performance-based assessment, competency-based assessment, and skills-based assessment. These types of assessment tools should be developed and evaluated by an agreed upon (universal) definition of data science and data literacy (core competencies) (Discussed further in Data Literacy Section). A lack of these definitions is one of the primary challenges with assessing educational outcomes in the data science discipline.

Additionally, it is important to define what students should both understand and know how to do when finishing a data science program (i.e. theory and practice/application). This is the main challenge, as data science core competencies have not been solidified. For example, a student intending to pursue a data science degree could progress through courses including (but not limited to) the following:

- Machine learning
- Business analytics
- Computer science
- Technology / Methods / Techniques

7 Lorica, B. (2017, June 08). Creating large training data sets quickly. Retrieved January 08, 2018, from https://www.oreilly.com/ideas/creating-large-training-data-sets-quickly

- Statistics
- Communications, and
- Domain of interest

Because of its breath, however, the vision is that data science would be an entry point for individuals, particularly those of color, to careers including those in STEM fields, social sciences, and humanities. As Calculus has proven to be a barrier to entry for many STEM fields, we propose data science as an alternative entry point to STEM and other fields.

3.2.2. Stakeholders

Stakeholders include students, faculty, industry partners, non-profit organizations, and government agencies. External stakeholders may provide the necessary information to begin strategic planning. For example, by asking students which teaching methods/modes are most conducive to their learning and introducing evidence-based teaching approaches for data science, recruitment and retention issues can be alleviated, thereby increasing the quantity and caliber of students pursuing careers in data science.

Additionally, stakeholders may also include:

1. Faculty from multiple disciplines
2. Institutional Review Boards
3. Institutional Assessment Offices
4. Students, as they are the direct benefactors
5. Counselors, advisors, recruiters
6. Industry Human Resources (HR) departments and advisory boards

3.2.3. Broader Impacts

Assessment provides a process for identifying areas for making improvements and changes to curriculum and staffing. Evaluation provides a benchmark for an acceptable level of competency to measure learning outcomes and faculty effectiveness. Having a recognized assessment process that defines the role institutions, academic departments, and faculty members must play in the learning process will maintain a standard of excellence in data science programs and data science courses across the nation.

By providing proof that data science programs are working, the door to more funding opportunities and interdisciplinary collaborations will open. A broader set of problems will be solved with data. Additionally, broader impacts include bringing assessment and evaluation to the forefront of institutions of higher education. Finally, general education level competencies for a data science (literacy) program will be impacted, such that no matter who you are or what experiences you have, a data science program or degree is achievable.

3.2.4. Challenges and Visions Pertaining to Assessment and Evaluation

While all acknowledge that formative assessment and summative evaluation systems are important for any educational initiative, there are major challenges that prevent widespread and effective programs. Time is a major barrier: grounded assessments that dovetail with learning goals and measurable outcomes for a course or program take extensive efforts by multiple stakeholders and regular updating and improvements as programs mature. Many stakeholders have sometimes viewed assessment, particularly summative assessment, in a hostile "high-stakes" way that deters effective implementation. Without effective assessment and evaluation, however, institutions and the broader profession will not be able to determine whether new data science education programs are effective and worth replicating.

Global standards for what students should know and be able to do by the time they graduate with an undergraduate data science degree (i.e. core skills and competencies) do not currently exist. These standards would differ depending upon a student's concentration/specialization. Just as general education competencies need to be assessed, data science competencies need to be assessed. Currently, ABET does not have an accreditation framework for data science. Evaluation challenges for undergraduate data science programs fall into the realm of accreditation. As data science is an extremely interdisciplinary field with varying definitions, it will be difficult (initially) for programs to be evaluated for accreditation nationally, and even globally, especially when core competencies have not been agreed upon. As everything that we do and see is data driven, assessing data science programs surely fits into the national landscape of bridging the digital divide.

Credentialing, as an alternative to traditional classroom instructions, is a viable option for industry professionals to enhance or learn new skills related to data science. Emerging instructional environments grounded in constructivist learning utilize various interactive methods and learning activities for engaging students. There are multiple ways to achieve knowledge in data science using these new emerging instructional environments. One example of an alternative credentialing method would be the use of Massive Online Open Courses (MOOCs) - content developed at host institutions using an appropriate learning management system (e.g., Blackboard). The development of interactive courses would require trained instructional designers to ensure the quality of education is validated. Another example would be to conduct boot-camp style workshops on data science using real-world models in the classroom, or as it is known by workforce professionals, "customized training". Again, the instructional environment should be student-centered where instructors are mere "guides on the side, not a sage on the stage".[8] If universities develop the online courses within their environment, this will ensure the students will be engaged in appropriate program delivery designed for each institution with established measurable student outcomes.

Our vision for assessment and evaluation addresses both teaching effectiveness and student outcomes. Teaching effectiveness and student outcome assessment should be a continuous process addressing course improvements and academic program improvements. Evaluation methods should be developed to identify strengths, weaknesses, and best practices for delivering instruction. For many potential students, the option of attending universities for an entire semester or quarter is not possible due to many restraining considerations. The vision is to determine how to alternatively credential an individual as a competent practitioner in data science if they have not completed a two or four-year degree. Data driven discovery is steadily on the rise, and there is a need for diverse persons and perspectives to learn the tools needed to fill the ever-increasing pool of data science jobs. A continuous improvement process for assessment and evaluation should provide:

1. A strategic plan for improvement,
2. Implementation of the strategies defined,
3. An evaluation process, and
4. Application of any suggested improvements.

3.2.5. Specific skills and Resources

Types of potential assessment modes include inquiry-based learning projects. For example, the Inquiry and Analysis VALUE Rubric is an authentic assessment instrument which could be used as a model to

[8] King, A. (1993). From "sage on the stage" to "guide on the side". College Teaching, 41(1), 30-35.

assess data science competencies. The Inquiry and Analysis VALUE Rubric is designed to understand undergraduate student success at all levels (freshmen, sophomore, junior, and senior).[9]

Independent Credentialing models exist outside of the United States. The European Union commissioned the Edison Data Science Framework[10] Project, a 2-year project (started September 2015) with the purpose of accelerating the creation of the Data Science profession. The EDISON Data Science Framework is a collection of documents that define the Data Science profession. Freely available, these documents have been developed to guide educators and trainers, employers and managers, and Data Scientists themselves. This collection of documents collectively breakdown the complexity of the skills and competences need to define Data Science as a professional practice.

Both formative and summative assessment strategies are needed to tackle assessment and evaluation challenges. Included in those strategies should be an awareness of biases. Self-reflection and seeking expert advice on biases is strongly recommended in faculty teaching so that negative patterns are not perpetuated.

We also recommend the development of a data science "bridge program" to facilitate multiple (flexible) pathways for students to enter (or re-enter) data science programs.

3.2.6. Concrete First Steps

To pursue this vision we recommend the following steps:

1. **Identify** general education data science competencies.
2. **Articulate** the data science competencies to our stakeholders, and leverage buy-in of the benefits of a data science program.
3. **Establish** a project-based assessment protocol (portfolio materials) as part of a stackable credentialing system that will allow students returning from the workplace to continue their studies without falling behind if there are curriculum changes.

To pursue our vision of assessing data science, projects like the "Keeping Data Science Broad" series are instrumental, as they bring stakeholders from various industries and organizations together to converse about the needs of data science assessment from various points of view. We'd like to point out that, even in this meeting, there were varying levels of interest in discussing assessment.

[9] Inquiry and Analysis VALUE Rubric. (2017, June 29). Retrieved January 08, 2018, from https://www.aacu.org/value/rubrics/inquiry-analysis

[10] EDISON: building the data science profession. (2015, September). Retrieved January 08, 2018, from http://edison-project.eu/

3.3. Curriculum

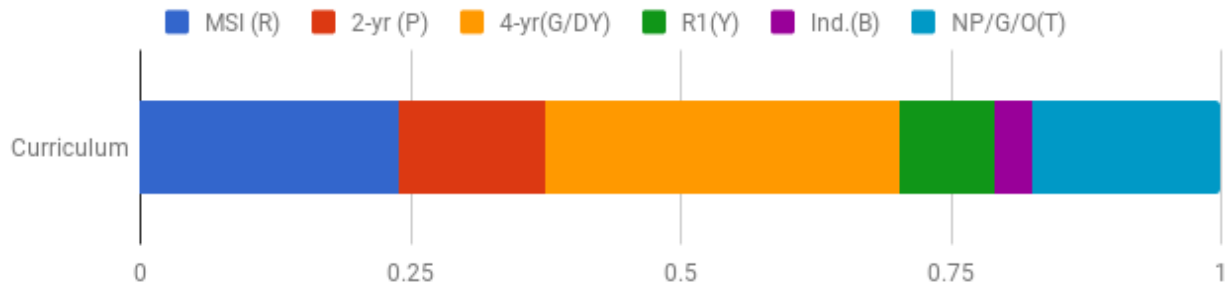


Figure 3: Ratio of workshop participant contributions to this section by institution type (n=171)

3.3.1. Introduction

Curriculum, from an academic standpoint, is absolutely central to any discussion of an evolving field. No effort to educate students will be successful without a properly prepared curriculum in Data Science or related disciplines. From K-12, to higher education and beyond, curriculum encompasses a broad range of sub-areas including textbooks, proper pedagogy, and course scaffolding. Also included are debates on the proper technology or tools to teach and how to stay abreast of them. Similarly, curriculum development covers experiential learning and far broader ideas including understanding of how to bridge disciplines for course inclusion.

It is critical to the development of this discipline that the fundamental concepts about data science are accessible at every educational level, K-12, 2-year colleges, four-year institutions, and advanced degree programs. As the discipline continues to develop, it is anticipated that a diffusion of data science across the undergraduate curriculum will result in clear pathways for specialization around specific areas of expertise (e.g., data visualization, data analysis, business analytics). In a nation with an increasingly diverse industrial and demographic landscape, a wide range of needs exist within different student populations and types of educational institutions. Understanding the current challenges as the discipline develops, and outlining a vision for what the future curriculum might entail is a key component for ensuring effective and inclusive growth in data science.

3.3.2. Stakeholders

The discussions surrounding curriculum are not unique to one type of institution. Higher education institutions at all levels face many challenges in creating productive transitions between various academic levels and programs. While curriculum is traditionally the purview of educators, particularly teaching faculty, it is important to remember that many other groups should have a say in how it is developed. Students, as the recipients of, and participants in, the curriculum have a huge stake in seeing the content and approach be effective, relevant, and dynamic. The employment sector, including small business, large industry, nonprofits and government agencies, have an equal interest in the outcomes from educational curriculum. If an employer hires a new graduate and finds them lacking necessary, central skills, the curriculum has failed, and both employer and employee must struggle to remedy the situation. For administrators, having a nationally accepted curriculum helps attract students (and faculty), ensures smooth course execution, and generally promotes a productive degree program. From a governmental perspective, it is important to encourage educational institutions to graduate students who can fill the growing need across all kinds of organizations for effective data use. Improving the usage of data will increase efficiency in operations, reducing costs and allowing agencies to better address societal needs.

In short, a broad and inclusive discipline curriculum is central to the ability to serve all members, levels, and organizations within society. This ability to serve all is also why it is essential that all stakeholders value and support efforts to develop curriculum that promotes diversity, inclusion, and broad access. Lack of openness in the educational process risks the exclusion of important voices that must be heard.

3.3.3. Broader Impacts

One of the challenges facing the development of data science curriculum among communities of data practitioners is bridging the gap between practice and learning for all. The ubiquitous use of data driven approaches means that skills and knowledge in these approaches play a significant role in 21st century STEM learning across settings. The challenges to building data science curriculum are aligned with persistent challenges in STEM education more generally and refer to core issues of science learning. Data science offers unique and genuinely new dimensions of learning, and creates opportunities for sparking learning and insight in STEM domains that have been resistant to improvement. The development of data science curricula will assist in the development of an equitable and inclusive path toward a data literate society and equitable workforce pathways. The development of data science curricula will help align the growth of the data science discipline with both the needs of individual students and educators, as well as data research and commercial sectors.

3.3.4. Challenges and Visions Pertaining to Curriculum

Identifying Appropriate Academic Levels

Challenge: There are questions related to the most appropriate academic level (associate, bachelor, masters or Ph.D.) and proper ordering of data science related material.[11,12,13] The challenge of establishing an effective pedagogical ordering is often exacerbated by the requirement to use existing courses within different disciplines. These courses often include the desired content, but also a lot of potentially tangential content. As the field develops its own unique courses and degree programs, we expect there will either be more courses that are directly supportive of data science, or a shift within the other discipline courses. The extent of tangential material also feeds another major curricular concern: credit creep within degree programs. When existing courses provide the content needed for data science, the number of courses required to provide depth in data science may become overwhelming.

Vision: *Figure 4* illustrates an inclusive initial and progressively narrowing approach to data science curriculum. The top, in its broadest, most inclusive level, addresses data literacy and general education courses. These courses could be deliverable both in 2-year and 4-year colleges, possibly even as advanced high school courses. Following these courses would be approximately 2 years of curriculum which would also be deliverable at 2 and 4 year institutions. We want to emphasize this point, as it is vital for the curriculum to be developed with both institution types taking part in the entire process, otherwise bottlenecks will develop in the overall pipeline. Some specialization may start to take place at this level due to different student populations and stakeholders involved.

---

[11] Anderson, P., Bowring, J., McCauley, R., Pothering, G., & Starr, C. (2014). An undergraduate degree in data science: curriculum and a decade of implementation experience. In Proceedings of the 45th ACM Technical Symposium on Computer science education (pp. 145-150). ACM.

[12] Cassel, B., & Topi, H. (2015). Strengthening data science education through collaboration. In Workshop on Data Science Education Workshop Report (Vol. 7, pp. 27-2016).

[13] De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C. & Kim, A. Y., et al (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

The third and fourth layers lead to continued specialization of BS degrees in Data Science and related disciplines. Here the divergence into the "teeth" represents the differences in the curricula due to different participating stakeholders, different target student population and different Program Learning Objectives and Outcomes. Finally, at the tips lie the curricula for advanced specialized programs (MS, Ph.D.) targeting different specializations within the broad area of Data Science and Analytics.

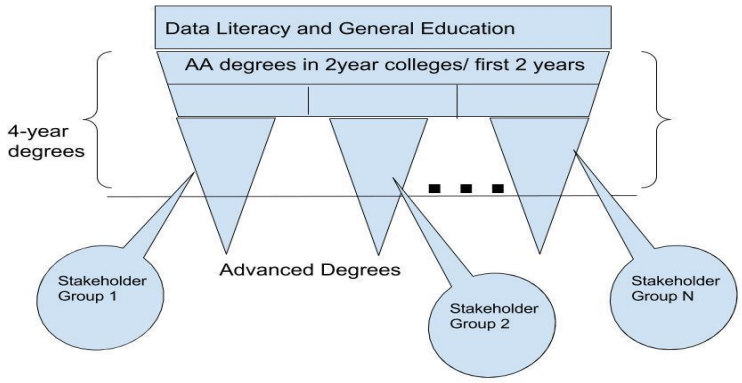


Figure 4: Degree levels and stakeholders for data science programs

Identifying Curricular Scope

Challenge: What sorts of "foundational knowledge" are required for data science? As the field becomes more cohesive, this knowledge can be identified, however, it is also liable to change as new tools and technology are developed. This makes it difficult to keep course content relevant, and extremely difficult to keep in-print relevant textbooks. As new tools spanning the entire data science process continue to be developed and adopted, introductory curriculum that originally required a wider range of skills may need to be adapted.

Vision: As data science curricula and programs evolve, it is important to have a shared vision of what a data science degree means. A "shark's tooth" model has been proposed to illustrate core competencies of a data science program.[14] Areas include *statistical reasoning, algorithmic thinking, data curation workflow, data management, content experience, data modeling, mathematical foundations, adherence to ethical standards, communication, data visualization,* and *reproducibility.* Each area could be emphasized (or deemphasized) in programs depending on what various stakeholders needed. Right now, most of these skills could be acquired through existing courses in a variety of disciplines, such as statistics (*statistical reasoning*) or computer science (*algorithmic thinking*). However, some topics are merely touched upon (such as *data curation workflow* or *reproducibility*). These unique areas, and the need to see all of these areas effectively intertwined for productive data science, suggest that additional departments, and especially courses, will need to be developed to integrate this knowledge. Alternatively, colleges and universities will need to become more capable of supporting inherently interdisciplinary study areas.

[14] <citation needed>

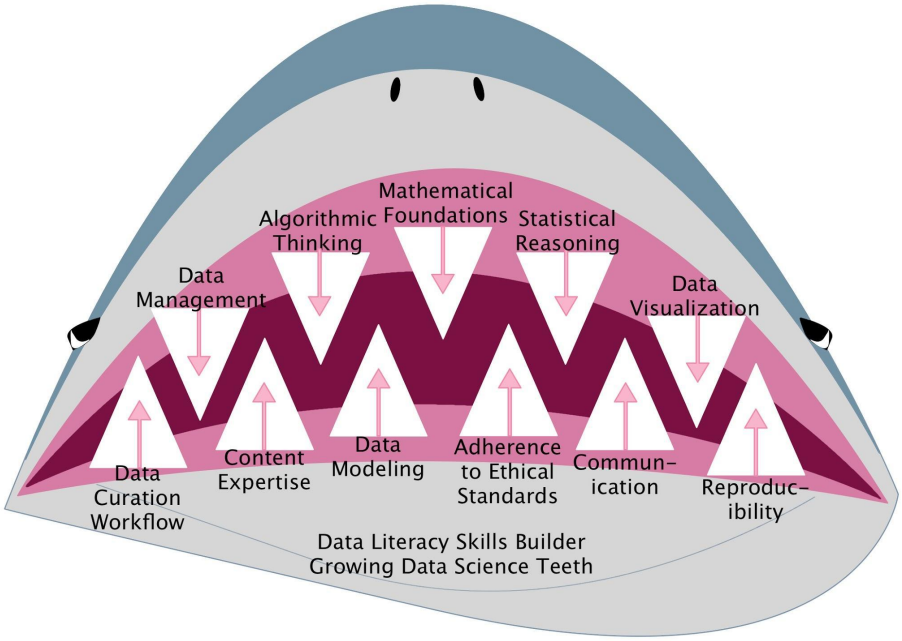


Figure 5: Model for Data Literacy Skills

When we begin to think about how to evaluate this sort of integrated knowledge, we return again to the application and use of data science. Sample programmatic and student objectives in-line with these ideas are provided in the *Tables 1 and 2*. We envision programs with similar participants, target student populations, Program Learning Objectives (PLOs) and Outcomes to eventually converge into the same degree name (new or already existing): e.g., Data Science, Business Analytics, Applied Data Science, Business Intelligence, Decision and Information Science, or Bio-physics.

TABLE 1: Sample Program Goals:
• Graduates will apply knowledge of data-intensive processes to address problems in a specific application context
• Graduates will effectively communicate data related analysis to both technical and non-technical audiences.
• Graduates will apply concepts of data integrity and reliability to analysis

TABLE 2: Sample Student Outcomes
• Students will demonstrate proficiency in tools and techniques for data collection, management, and analysis.
• Students will solve real world problems using a variety of tools and technologies
• Students will apply programming skills to manage data and carry out analysis

● Students will communicate data science results to a broad (non-technical) audience
● Students will analyze data-centric problems and determine the best approach to solve the problem
● Students will evaluate the potential bias, error, and data quality issues associated with a given solution

Participants in this workshop found value in accreditation. Potential accreditation bodies are *The Association to Advance Collegiate Schools of Business* (AACSB) for more business oriented programs (such as business analytics) and *The Accreditation Board for Engineering and Technology* (ABET) for more engineering/computing-focused programs (such as a data science or data engineering degree). The authors of this report would like to urge caution, however. It remains unclear exactly where more general programs such 'data science' might reside, or even what precisely an accredited program should look like. Likely, similar to computer science, there will be many high quality programs, only some of which will seek or achieve accreditation (through ABET, AACSB, or another body). One potential alternative, focused more on the individual (student), would be to facilitate or encourage graduates to become a *Certified Analytics Professional* (CAP) - an ANSI accredited distinction available through the *Institute for Operations Research and the Management Sciences* (INFORMS).[15] Another option, supported in the 2015 Data Science Education Workshop, involves curriculum guidance from appropriate societies. ACM has a long history of computer science, software engineering, computer engineering, information systems and information technology curriculum recommendations that have been heavily used in the design of programs. The American Statistical Association has published a recommendation for Data Science, as have a number of other organizations.

Sharing Curricular Design and Development Resources

Challenge: The inherent interdisciplinary nature of data science makes it difficult to have a cohesive presence for collections of materials appropriate to the discipline. There is minimal consistency across existing curriculums at different schools, which has limited sharing of materials, limits transferring of students and causes issues with matriculation from 2-year colleges to 4-year colleges. This is even true within many schools which house different "data science" programs that might reside in math, statistics, computer science, or business. Silos cause issues within colleges, curriculum, and more. Some efforts do exist[16]; however, they are not institutionalized or even necessarily maintained. On the other hand, curating and maintaining such a repository is a massive undertaking.

Vision: As curricula get developed, standardized and tried, we envision the creation and maintenance of a repository of peer-vetted reusable curricular components. A similar repository of curricular material to what we are envisioning already exists for Computer Science. The Engage-CS repository[17] run by the National Center for Women & Information Technology (NCWIT) has had a lot of success. Potentially this sort of material for data science could be included in that repository, or the model could be replicated for data science. The Ensemble Computing Education Repository[18], created with the NSF NSDL

[15] INFORMS(n.d.).Retrieved January 09, 2018, from https://www.informs.org/
[16] For example, www.teachingdatascience.org
[17] EngageCSEdu. (n.d.). Retrieved January 08, 2018, from https://www.engage-csedu.org/
[18] Computing Portal. (n.d.). Retrieved January 08, 2018, from http://computingportal.org/

program, houses a broad collection of computing education material and does include some data science specific materials.

Co-curricular Experiences

Challenge: Another area of 'curriculum' that bears consideration is opportunities for co-curricular experiences like internships or apprenticeships. Providing these types of opportunities is made difficult by the fact that many companies are looking to hire interns from their traditional disciplines. Anecdotal evidence was given during the workshop that, in some cases, employers would ignore explicitly trained data science students, while students from within disciplines actually lack the skills companies are seeking to acquire.

Vision: This is a prime area that the Big Data Hubs can address, providing an important interface between companies and academia. By providing a narrower interface, the hubs can serve as funnels, helping companies to both clarify their needs and providing a modicum of standardization in the opportunities offered to students. Many programs and schools simply will not have the personnel to pursue and develop deeper industrial or governmental contacts required for the regular, high-caliber co-curricular activities an applied field such as data science needs to properly prepare students.

A Unifying Vision for Curriculum

Data science is an inherently interdisciplinary area of research and study. It brings together components of statistics and computer science and applies the result in any of a large collection of application domains. These vary from astronomy to business to history to journalism to medicine to political science and zoology. Every field deals with increasing amounts of data and the way of working in many fields has changed radically as a result. A basic question for a data science curriculum is how to prepare students to be contributors to any one or perhaps several of the application domains. Some fields will have raw data that must be cleaned and formatted before it is of any use. Others will have relatively reliable data, but will have very complex analysis requirements. Some fields can accept less than perfect results. Others, like medicine, may present life-threatening dangers if data is mishandled or misused. Is it possible to have one version of a curriculum that meets all the needs of all the fields? How should the differences be recognized and represented in the required and elective courses in a full program.

The need to use material that already exists in courses suggests that a more modular approach to course design may be useful. It would be useful to allow a variety of module combinations to create courses that are particularly appropriate for a given group with specific goals. Not only Data Science, but also other emerging fields could benefit from this approach. NSF has sponsored some projects for developing modules related to data science and machine learning.

3.3.5. Specific Skills and Resources

The resources to enact this vision are largely absent. We are looking to produce graduates who bring together a number of areas of study, basically computer science, statistics, and a domain. Yet, we would be very surprised to find one faculty member who combines all of these areas at a high level. Data Science is inherently interdisciplinary. Putting together a team that can adequately prepare students to work in a data centric field is very challenging.

Significant work has been done to understand barriers to entry into Computer Science, Statistics and other STEM disciplines. As a new curriculum comes online, it is absolutely necessary to be sure that it

implements and uses these best practices. For example, Engage-CS explicitly discusses evidence-based high-impact practices for encouraging broad participation.[19]

A repository of curricular material similar to what we are envisioning for Data Science already exists for Computer Science. The above-cited Engage-CS repository, run by the National Center for Women & Information Technology (NCWIT) has had a lot of success. Potentially this sort of material for data science could be included in that repository, or the model could be replicated for data science. The Ensemble Computing Education Repository,[20] created with the NSF NSDL program, also houses a broad collection of computing education material.

A few publications have begun to outline what a full-fledged curriculum might look like[21,22,23]. More extensively, the European Data Science Academy[24] publishes extensive curricular materials for a wide range of courses.

3.3.6. Concrete First Steps

Achieving this vision requires curricular developers to:

1. Understand the variety of curricula that need to be prepared to address the appropriate participation, learning objectives and desired outcomes
2. Develop the appropriate learning objective and desired program outcomes for each version of the curriculum
3. Develop the appropriate sequences of courses (complete with prerequisites) for each curriculum type.
4. Develop appropriate syllabi/course content for any new courses needed for the respective programs

[19] Engagement Practices. (n.d.). Retrieved January 08, 2018, from https://www.engage-csedu.org/engagement/make-it-matter

[20] Cassel, L., Delcambre, L., & Hislop, G. (2014, March). Ensemble: the sharing community. In Proceedings of the 45th ACM technical symposium on Computer science education (pp. 734-734). ACM.

[21] Anderson, P., Bowring, J., McCauley, R., Pothering, G., & Starr, C. (2014, March). An undergraduate degree in data science: curriculum and a decade of implementation experience. In Proceedings of the 45th ACM technical symposium on Computer science education (pp. 145-150). ACM.

[22] Cassel, B., & Topi, H. (2015, October). Strengthening data science education through collaboration. In Workshop on Data Science Education Workshop Report (Vol. 7, pp. 27-2016).

[23] De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C. & Kim, A. Y. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

[24] Data Science Training and Data Science Education - EU. (n.d.). Retrieved January 08, 2018, from http://edsa-project.eu/

3.4. Data Literacy

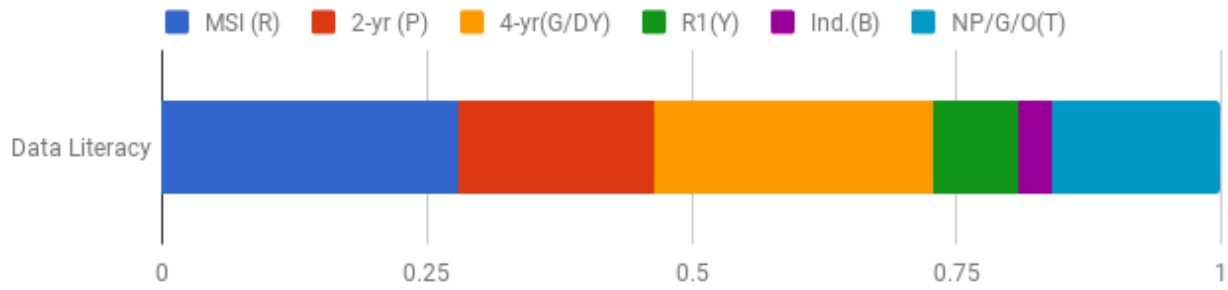


Figure 6: Ratio of workshop participant contributions to this section by institution type (n=125)

3.4.1. Introduction

There is evidence to suggest that standing alongside the "digital divide" [25] is what might be called a "data divide." Increasingly, the prosperity, innovation and security of individuals and communities depend on a data literate society.[26] There must be a coordinated effort to determine what it means to be a data literate citizen, information worker, researcher or policy maker; to identify the quality of learning resources and programs that can improve data literacy; and to chart a scalable path forward that will bridge data practice with data learning, education and career readiness. Advancing Data Literacy (DL) is essential for creating an informed civil society. An inclusive, collaborative, multi-sector, multidisciplinary approach is needed to develop a bias-aware and unified framework and sustainable infrastructure that will allow for DL development. There is an urgent and growing need to identify and deploy promising approaches to DL that must go beyond basic data science and statistical literacy, a kind of scaffolding of basic DL skills and recognition of the advanced kinds of skills that data demand.

It has been noted that there is a lack of even basic DL reasoning skills, and there have been calls for improved DL across disciplines for over a decade. [27,28,29] During that period, evidence based curricula have been proposed to address a lack of basic DL.[30,31,32]. It has even been suggested that a rigorous program of statistics and DL replace the Algebra-Calculus track.[33] Increasingly, there is a need to identify

[25] White House (2015).Mapping The Digital Divide. White House Council of Economic Advisers Issue Brief. July 2015.
[26] UNESCO (2013). Literacy and competencies required to participate in knowledge societies. Conceptual Relationship of Information Literacy and Media Literacy in Knowledge Societies, 3. Research Paper from Worlds Summit on the Information Society, 2015. Paris: United Nations Educational, Scientific and Cultural Organization.
[27] National Research Council (2006) Learning to Think Spatially: GIS as a Support System in the K-12 Curriculum. Report from the Committee on Geography; Board on Earth Sciences and Resources; Division on Earth and Life Studies. Washington, DC: National Academies Press.
[28] Kastens, K. & Krumhansl. R. (2013). EarthCube Education End-User Workshop. Scripps Institution of Oceanography, La Jolla, California March 4–5, 2013. Arlington, VA: National Science Foundation.
[29] Zalles, D. (2014). Young youth explore geospatial data for citizenship project: A case study. Menlo Park, CA: SRI International.
[30] Vahey, P., Rafanan, K., Swan, K., van't Hooft, M. A., Annette Kratcoski, R. C. E. T., Stanford, T., & Patton, C. (2010, May). Thinking with data: A cross-disciplinary approach to teaching data literacy and proportionality. In American Educational Research Association Annual Meeting, Denver, CO.
[31] Zalles, D. R., & Vahey, P. (2005). Teaching and assessing foundational data literacy. Paper delivered at Annual Meeting of the American Educational Research Association, San Francisco CA.
[32] Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., Wuetherick, B. (2015) Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report. Halifax, NS: Dalhousie University.
[33] Rubin, A. (2005). Math that matters. Hands On: A Journal for Mathematics and Science Educators, 28(1), 3-7.

and deploy promising approaches to DL that must go beyond basic data science and statistical literacy. A scaffolding of basic DL skills is necessary as well as recognition of the advanced kinds of skills that are needed such as: emphasis on design, construction, visualization of large-scale data, as well as federation of multivariate data streams through computational analysis tools; enhanced exploratory, stochastic, pattern-seeking and interdisciplinary approaches; machine learning; semantic and ontological analytical processes; and leveraging of cyber infrastructures to make large data structures more available and interoperable.

Goals:

- A commonly understood definition of DL and how it differs from data science
- A published handbook including examples from journalism and academic papers
- Inclusion of a DL component in every introductory course
- Ensuring that DL is pervasive throughout curriculum (similar to writing across the curriculum)
- Teacher training to use DL approaches
- All students, starting in pre-K, will receive DL through curriculum. This approach will continue through middle and high school so that all high school graduates are data literate
- Using backward design from career pathways, e.g. data businessperson, data engineer, data creative and data researcher, to arrive at skills needed
- Define the relationship DL has to domain science
- Define other terms used to identify related sets of skills and educational experiences, e.g. Data Analytics, Business Intelligence, Big Data, Machine Learning, Deep Learning
- Data science will be seen as an engaging, creative and imperative body of knowledge
- DL will be the foundation of fact-based critical thinking.

3.4.2. Stakeholders

All citizens in the 21st century, by the time they graduate from high school, should be data literate. The degree to which individuals and organizations can manage, manipulate, and interpret data will determine their competitiveness or lack thereof in tomorrow's workplace. It is imperative that data science as a general subject is explained in ways and used in examples so that non-STEM majors can understand its application in their discipline, while giving the depth and breadth needed to STEM and related majors to become subject-matter experts. Specific stakeholders include:
Academia: those coming from pre-defined academic disciplines, e.g. business, math, computer science, statistics, management science; practicing data scientists with research interests;
Students: undergraduate students will graduate with critical and computational thinking skills, understand the basic tenets of data science, its functionality, and its interplay with every major to be an effective user of data so that they may succeed in today's and tomorrow's workplace.
Industry: In order for organizations in any industry to survive and thrive, data sciences will be a core function working in tandem with its core technology. Organizations will have to have a command of their data in order to be competitive. They will have to determine whether to invest in an in-house data science team or outsource their data science functions.

3.4.3. Broader Impacts

One of the grand challenges for communities of data practitioners is bridging the gap between STEM practice and STEM learning for all. The fact that the preponderance of important discoveries in contemporary science come from large-scale, data driven approaches indicates that skills and knowledge

in these approaches must play a significant role in 21st century STEM learning across all settings.[34,35] .It also means addressing issues of equity. The challenges to cultivating data literacy are deeply aligned with multiple persistent challenges in STEM education more generally and get at some of the core issues of science learning, such as the nature of evidence and inquiry, inference, and understanding the complexity of nature and ourselves. There have been calls for improved data literacy across disciplines for over a decade.[36 ,37,38,39,40] What is needed is an equitable and inclusive path toward a data literate society. Data create unique and genuinely new dimensions of learning, and opportunities for sparking learning and insight in STEM domains that have been resistant to improvement. All citizens need to be data literate, meaning they will be able to employ fact-based critical thinking in order to understand the world around them. Students will be better prepared to apply data science skills to real world problems and also to critically interpret data when presented. Clear definitions of Data Science, Business Analytic, Data Analytics, Business Intelligence, Data Literacy, and related terms are a prerequisite for building consensus curricula in each area. It forces an innovative, interdisciplinary approach to higher education.

3.4.4. Challenges and Visions Pertaining to Data Literacy

Public Engagement in Data Science

The emerging need for data science jobs has the potential to reshape the workforce makeup in the US the same way as the loss of manufacturing jobs, the emergence of the service sector, and more recently, the emergence of the technology sector have affected it. The success of such reshaping is predicated upon the availability of skilled people to fill these jobs. While industry is feeling an ever-increasing need for competent data scientists and data analysts, the academic pipeline for producing a qualified and educated workforce to fill these spots is only now emerging and there is a lack of a concerted and coherent message that would attract students to these disciplines. Proper efforts to address this issue would ensure successful career paths for generations of current and future students. There is a need for a coordinated effort to increase awareness of data literacy at the K-12 level, including accessible and engaging introductory classes that entice and retain students.

Additionally, beyond workforce challenges the democratization of data literacy is essential for engaged citizenship in this new era. To be able to understand the effect data has over our lives is a powerful social driver, and one that is being denied to the poor, working class, rural, minority, and disadvantaged communities across the country. It impedes the ability of the populous to fight back against unfair algorithmic driven policies, procedures, or practices due to an imposed barrier to technical understanding.

Data Scientist Engagement in Public Life

Conversely, no one should be educated in a vacuum, valuing only the data skills and not what it means to be truly literate, i.e. “having or showing education or knowledge, typically in a specified area”. In a

[34] Parashar, M. (2009). Transformation of Science Through Cyberinfrastructure: Keynote Address. Open Grid Forum Banff, Alberta, Canada, October 14. http://www.ogf.org/OGF27/materials/1816/parashar-summit-09.pdf (accessed 12/21/2015)
[35] Hey, T., TanSley, S., & Tolle, K. M. (2009). Jim Gray on eScience: a transformed scientific method. Redmond, WA: Microsoft Research
[36] National Research Council (2006) (n.d.)
[37] Kastens & Krumhansl (2013) (n.d.)
[38] Zalles,(2014) (n.d.)
[39] Rubin (2005) (n.d.)
[40] Ridsdale, C. et al (2015) (n.d.)

specified area implies that if data skills cannot be translated to the world at large, with knowledge all of its context and nuance, you are not really able to deploy those skills within the larger society. For example, students can graduate with an elite education in computer science, statistics, or data science without taking any courses in literature, geography, politics, or other disciples that orient them to the world around them. One anecdote stated that Ralph Nader's wife visited the top engineering school in the 1980's and put up a blank map for the graduating class and asked for one person to identify Iraq on the map, not one could do it. This same story could be repeated at many of our programs today. If data scientists are now taking on an elevated role of providers of modern social tools, the idea of data literacy also needs to be democratized to include a civic sense of what is known and unknown as well as a sense that skills alone do not equal knowledge.

That said, there is a need for recruitment and retention of a diverse body of students interested in pursuing a data science undergraduate degree. Those affected include: prospective students looking for an appropriate education and career choice; industry members looking for entry-level, comprehensive, and advanced skills; data science faculty responsible for direct recruitment, advising and training; and other faculty/departments.

Critical Thinking and Computational Thinking

Both critical thinking and computational thinking are necessary to be data literate. These modes of thought are different perspectives on problem solving, and are both essential skills for all citizens and all disciplines. Problem solving can be defined in part as identifying entities (or things); attributes (or properties, characteristics); relationships among the entities; and various perspectives, solutions, and/or extensions of the problem.
Citizens need to:

- be able to identify the expected range of data values
- discern whether an actual data value is incorrect, impossible, or an outlier.
- be able to identify the context in which a problem is given
- be engaged
- understand problem context and relevance
- start with what is given in a problem
- understand ethical and compliance issues surrounding the intended use of the data

3.4.5. Specific skills and Resources

While data and learning are nascent areas of learning research, existing research in a few areas of data and statistical science suggest pathways into DL. Core to understanding and advancing data analytic skills is the use of statistical inference, and advanced statistical reasoning for sense making of big data. Makar & Rubin[41] propose that a learning progression of inferential reasoning would be possible, but learning research is needed to validate it. There is also a need for the development of new exploration tools to help provide opportunities for learners to explore data.[42]

Recent work on statistical modeling and learning emphasizes the importance of models and simulations along with inference as an approach to bringing data skills to learners. Particularly well researched is the

[41] Makar, K. and Rubin A. (2018). Learning about Statistical Inference. In Ben-Zvi, D., Makar, K., and Garfield, J. International Handbook of Research in Statistics Education. Dodrecht: Springer International Publishing AG.
[42] Hammerman, J. (2009). Statistics Education On The Sly: Exploring large scientific data sets as an entrée to statistical ideas in secondary schools. Proceedings of Satellite: Next Steps in Statistical Education, World Statistics Congress 57, Durban, South Africa. International Association for Statistical Education.

use of statistical modeling. Tinkerplots is among a small number of tools that have been widely researched in understanding how learners cope with large and multivariate data sets.[43,44,45] Scientific visualization can also provide opportunities to learn and communicate complex statistical and scientific data.[46,47] Ainsworth and Loizou[48] show improved comprehension of complex science data through visualization.

Thus, while there has been progress on advancing DL on a range of fronts, none of these approaches have yielded comprehensive or generalizable and effective learning of complex phenomena that can boost DL broadly. Conclusions drawn from these studies are narrow and weaknesses in overall findings have prevented them from being generalizable. What is needed in the short term is an educational framework for DL goals; an accounting of the effectiveness of DL teaching and learning resources, tools, and best practices; identification of gaps in the capacity for, testability and utility of these resources to leverage against DL needs; and a strategy to develop a research agenda and culture of lifelong learning that recognizes and infuses DL into all aspects of STEM learning.

Several core and interdisciplinary skills and actions are required:

- Exposure to data science and a real world application of data science as related to relevant disciplines
- Computational thinking (critical thinking skills relative to computing)
- Communication and collaboration skills
- Ability to map real world problems to an appropriate data model prior to capturing the data
- Building habits of mind in order to be able to ask questions that can be answered through the use of data analysis
- Pedagogy resources for pre-K-12 teachers
- Problem-solving skills
- Math skills
- Computer skills

3.4.6. Concrete First Steps

- Develop a globally accepted definition of Data Literacy. This can be achieved through a distributed and collective effort, engaging a wide range of data scientists and data science educators, policymakers, community activists, and learning researchers. This process will require several iterations. The resulting commonly held definition will not belong to any one institution.
- Data literacy may be combined with a course in propositional logic to create a critical thinking course.
- Couple data literacy with a discussion of ethics using case studies.

---

[43] Konold, C. (2007) Designing a Data Analysis Tool for Learners. In M. Lovett & P. Shah (Eds.), Thinking with Data. New York: Lawrence Erlbaum Associates. 267-291.
[44] Konold, C., Harradine, A., and Kazak, S. (2007). Understanding Distributions by Modeling Them. International Journal of Computers for Mathematical Learning, 12(3). Dodrecht: Springer. 217-230.
[45] Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. ZDM, 44(7), Berlin: Springer. 883-898.
[46] Roschelle, J., & Pea, R. (2002). A walk on the WILD side: How wireless handhelds may change computer-supported collaborative learning. International Journal of Cognition and Technology, 1(1), 145-168.
[47] De Jong, T., Van Gog, T. & Jenks, K. (2009). Explorations in Learning and the Brain. Berlin: Springer Verlag.
[48] Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. Cognitive science, 27(4), 669-681.

- Demonstrate a portfolio of work analyzing and synthesizing articles, blog posts, videos, etc. in a particular field of interest. Student lightning talks may be used to facilitate this.
- Offer middle and high school teacher workshops building on elementary school work
- Develop critical thinking and computational thinking curricula
- Strengthen foundational math skills
- Strengthen application-based problem-solving skills

3.5. Diversity, Inclusion, and Equity

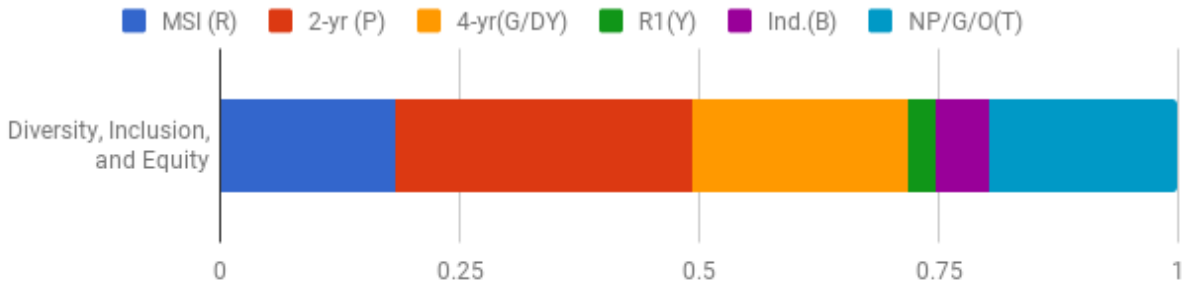


Figure 7: Ratio of workshop participant contributions to this section by institution type (n=71)

3.5.1. Introduction

An article published on September 15, 2017 in Science titled "Without inclusion, diversity initiatives may not be enough"[49] spoke of the need for diversity initiatives in STEM to concentrate on the experience of minorities in STEM and not just the numbers served. Diversity in this context is defined in terms of race, gender, religious affiliation, socioeconomic status, ethnicity, and being the first generation in college. The variety of perspectives such diversity provides is as essential as that provided by the transdisciplinary nature of data science for innovation and growth of the field. Given the current and projected demand for professionals in data science and its relative youth as a field in STEM, tackling issues of diversity and inclusion head on should be a high priority.

The Science article cited above outlined barriers and disconnects. Specifically, how xenophobic, misogynistic, and racist chatbots were created through the interaction between artificial intelligence and social media that eventually needed to be shut down [50]. In an era where elections and algorithms can be manipulated through fake accounts on social media,[51] the need for diverse perspectives in the creation and analysis of data and in the evaluation of results is more crucial now than ever before. As such, data science requires an inclusive environment for all participants built on mutual respect where all perspectives are equally valued. To create such an inclusive environment, all students and faculty should have equitable access to such diverse resources including internet, quality curriculum, data science mentors, and technology. Currently, there exists a data and digital divide where colleges and universities such as HBCUs and other minority serving institutions, rural colleges, and many two-year colleges do not have equitable access to these resources.

---

[49] Puritty, C., Strickland, L. R., Alia, E., Blonder, B., Klein, E., Kohl, M. T., McGee, E., Quintana, M., Ridley, R.E., Tellman, B. and Gerber, L.R.(2017). Without inclusion, diversity initiatives may not be enough. Science, 357(6356), 1101-1102.

[50] Kraft, A. (2016, March 24). Microsoft shuts down AI chatbot after it turned into a Nazi. Retrieved January 08, 2018, from https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

[51] Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017, December 30). The spread of misinformation by social bots. Retrieved January 08, 2018, from https://arxiv.org/abs/1707.07592

Why is this important to the nation and for your institution types?

The Science article that is referenced above notes that "Diversity among scientists can foster better science." Broadening participation in the field will assist in filling the employment gap that is expected. Conversely it may increase the number of problems to be tackled, but also improve the quality of the solutions. We need to avoid the mistakes made by computer science in the 80's - 90's. This is a great opportunity for schools to work together (HBCU and other MSI's, Community Colleges, and smaller schools) to strengthen collaboration, to share resources, and apply for joint funding to compete with larger schools. It is also an opportunity to build more equitable collaboration between majority-serving universities, in particular high research institutions, and other colleges for mutual benefit.

How does this challenge theme fit into the national landscape?

The workforce shortages in data science are anticipated to grow: we need to ensure all students have opportunities to engage in data science. For example, the Georgia Tech Research Institute has a Historically Black Colleges and Universities Outreach Initiative, providing resources for faculty at these classes of schools. Such efforts may be expanded to include students via shared online workshops and courses, computing resources, and access to the research faculty. Additionally, partnering with a local startup incubator provides an excellent way to collaborate, especially since we are now seeing a startup economy that is more geographically diverse than ever before[52]. We have an ethical responsibility to educate and identify data scientists that represent, understand, and empathize with the perspective of underrepresented populations. It is imperative that we implement guidelines to ensure that underrepresented and diverse populations are represented in and served by the future of data science education.

3.5.2. Stakeholders

Faculty, students, administrators such as department heads and deans, data scientists, lecturers, advisors, counselors, industry recruiters, college recruiters, system administrators and staff - all are affected by decisions made (or not made) in the interest of diversity, inclusion and equity.

3.5.3. Broader Impacts

If we do not make diversity and inclusion a priority now, we will not have it in the future. We do not want to repeat the mistakes of the past, so we must reverse the trend for the growing divide to make and keep data science broad. Diversity will bring a lot of ideas and voices to the table, which may lead to significantly fewer models producing biased results when trained using algorithms on biased data sets. This is a topic of current interest in industry and academia. There are few concrete examples of teams which did not allow biased data to leak into model results and no best practices. Ultimately, if we want to serve a diverse group of people, we need a diverse group of data scientists. At every university, there should be a program for people from non-traditional science programs to get information on where data science is being promoted among faculty and students of different disciplines. All graduates should leave college with data literacy.

[52] "How the Startup Economy is Spreading Across the Country - and How it Can Be Accelerated", Michael Mandel, March 2017, Progressive Policy Institute, Retrieved January 10, 2018 http://www.progressivepolicy.org/issues/economy/how-the-startup-economy-is-spreading-across-the-country/

Industry values people with an innate understanding of the world around them, since their experiences are not based on living/growing/working in a homogenous environment. Absent diversity in data science, we end up with the type of diversity issues we have seen in recent news regarding Silicon Valley corporations' racial and gender gaps. If more students have access to the field, they will be represented in the future creation of technology.

Currently, the majority of the students in graduate school in technical fields in the US are foreign born. Given the current climate of the nation with respect to immigration, many students who are in the pipeline for graduate degrees may go back to their countries when finished with degrees. This will negatively impact the US economy unless we are able to fill the void at the undergraduate level and can encourage more US citizens to go to graduate school. We will not be able to fill the void if we do not focus on making data science accessible and inclusive to all.

3.5.4. Challenges and Visions Pertaining to Diversity, Equity, and Inclusion

Equitable Access to Data Science

One central point is that "equitable access" is not the same as "equal access" in the sense that a solution for one community does not apply to all communities. Several HBCUs and other MSIs, community colleges, small and rural 4-year colleges in the US lack access to quality data, tools, projects, and journals for data science, despite the proliferation of open data and open software across the web. This highlights the fact that our current school systems are inherently unequal, particularly in their access to knowledge about what resources are available, where those resources are being held, how to access and use those resources, connections to organizations that keep a pulse on new open projects, and the needed support services to effective take advantage of those resources. For example, a high school struggling with basic literacy and math skills for students does not have the bandwidth to attack data literacy unless it is tied to the teaching of other basic skills and associated with needed support services known to improve student learning.

There exists a need for data science courses that are adaptable and flexible to student and faculty interests and their geographic and financial needs. For example, it may be that a particular college may not have the funding for an introductory data science course, in which case incorporating data science into other courses which all students are required to take and retraining exiting faculty is another manner of achieving equitable access.

Inclusive Faculty and Student Bodies

Our vision is that data science faculty reflects the student body and society as a whole. We envision hiring practices that foster building a deep and diverse candidate pool by asking colleagues for candidate recommendations and inviting specific individuals to apply as well as by training search committees to avoid implicit bias. Once diverse faculties are hired, institutions would provide support for inclusive excellence. Faculty retraining from HBCU and other MSIs and rural institutions through visitation at institutions with strong programs in data science or in industry would also include extended mentoring beyond the period of the visitation. Inclusive environments at these institutions would support and value faculty work in diversity.

Engaging, Culturally Relevant Curriculum

In terms of teaching and pedagogy, we envision a more student-centered approach toward building curriculum which is engaging, accessible, and culturally relevant. Recruitment would be increased by breaking down stereotypes through outreach programs to high school students as well as to colleges students from other disciplines. Retention could be improved by creating support structures focused on first-generation and underrepresented student success such as orientation, mentoring programs, and seminars to help students learn the cultural and professional skills often not taught in the classroom which

are necessary to succeed in academia. Ideally, data science will become an essential part of the program of study for all undergraduate degree programs and all students can obtain full federal financial aid and pursue careers in data science.

3.5.5. Specific skills and Resources

Currently, there needs to be a higher premium on collaborations between the majority institutions and HBCUs/MSIs including small colleges in rural communities and two year colleges for example programs from NSF like HBCU-UP or Community Colleges: A Resource for Increasing Equity and Inclusion in Computer Science Education could be expanded or used as models for data science[53]

Secondly, StatFest, Tapia, Grace Hopper and SACNAS are models for an undergraduate conferences focused on underrepresented students. Structures must be created within these conferences focused on developing undergraduate data science programs and fostering interest in data science within underrepresented communities.

Finally, it is important to leverage partnerships with companies like Verizon, T-Mobile, and Intel to provide access to data science resources across institutions.

Additional needs and skills include:

- On-going implicit bias training for faculty, counselors, and staff at high schools, colleges, and universities
- More equitable funding of public K-12 education
- More equitable funding of higher education
- University programs to support underrepresented students for success in the field
- Universities need to look at themselves to see how they are affecting diversity perhaps through the creation of a diversity metric for recruitment and retention of underrepresented students
- Work in diversity is not valued for tenure
- Internet access
- Culturally relevant quality curriculum
- Diverse data sets
- Qualified instructors with a current, up-to-date skill set
- Conversations between R1s and HBCU/MSI is usually not one of mutual respect and equitable collaboration so we need to examine experiences of HBCUs/MSIs with R1s and promote programs which demonstrate a mutually beneficial relationship
- R1s usually engage with HBCUs/MSIs only as part of federal grant requirement (e.g., NSF broader impact)
- A better understanding of the difference between equity and equality and a global effort to promote equity in data science education in the US and around the world to attract a diverse pool of talented students to data science

[53] Lyon, L. A., & Denner, J. (2017). Community colleges: a resource for increasing equity and inclusion in computer science education. Communications of the ACM, 60(12), 24-26.

3.5.6. Concrete First Steps

- Exchange program for graduate students interested in the professoriate to teach students and faculty at institutions that do not currently have expertise in data science. (Could be short-term, e.g., REUs, or long-term, e.g., for a semester.)
- Utilize Data Science Hubs as a resource for professional development opportunities, as well as for curriculum development.
- Tailoring of MOOCs to meet needs of local community and college.
- Connecting all students with limited technology resources because of financial hardship with organizations who can provide resources, e.g. all students should have a quality modern computer.
- There needs to be an acknowledgement and discussion around the fact that there is a significant gap between data science at majority serving institutions and HBCUs/MSIs. As of today's date, there are no HBCUs or MSIs that have a data science program that is up and running that we know of. As such, there are little to no opportunities for students from diverse backgrounds (e.g., academic, financial). From a faculty, administration, and research standpoint, there is a disparity between majority institutions and HBCUs/MSIs.
- Additionally, HBCUs and MSIs should intentionally seek out opportunities to collaborate with one another around data science.
- Mature multiple on-ramps to degrees/careers in data science (e.g., community college, certificate programs, HBCUs, MSIs and majority serving institutions).
- Federal and industry grants to R1 will be evaluated against what the R1 institution is doing to collaborate with and/or develop diverse data science programs. One of the criteria should be that the R1 institution works with an HBCU/MSI in order to receive the grant. There needs to be an equitable partnership between the R1s and HBCUs/MSIs in the form of shared intellectual property, compensation at the PI level and access to journals so the HBCU/MSI can publish its papers.
- Industry should go directly to HBCUs/MSIs to develop data science programs.
- Basic and advanced training and ongoing professional development focused on unconscious bias, equity vs. equality, stereotype threat, and culturally relevant curriculum is needed for all faculty, counselors, and students in the field of data science.

3.6. Ethics

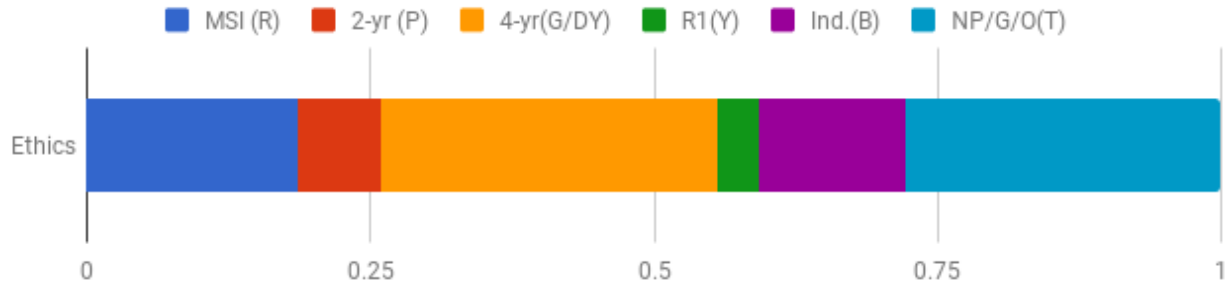


Figure 8: Ratio of workshop participant contributions to this section by institution type (n=54)

3.6.1. Introduction

Much discussion has taken place concerning the nature and scope of human beings' use of data. Connected with recent political figures and national campaigns, headlines that dominate the news media include mishandling of data (emails), selling of data (Russia), manipulating voting precincts with data (gerrymandering), and the social responsibilities of corporate social media outlets (Facebook and Twitter). It can be argued that elections can be won or lost because someone figured out (more effectively than their opponent) how to use data to socially engineer a result.

Questions about who is responsible for all this data, what privacy issues are involved, what precautions are necessary to ensure the data is being used in an ethical manner, what lengths should be taken to protect data integrity as well as security of the data and what can law enforcement do to obtain information they desire remain largely unanswered by the high tech world.

These ethical questions surrounding data are sometimes conflated with compliance by individuals and institutions to serve their own interests. It is not uncommon for unethical behavior to be excused by individual and institution claims that all relevant laws and regulations were followed. Ethical behavior is not the same as the avoidance of breaking laws. Laws making, as well as the educational enterprise, is a slow process and struggles to keep pace with the current exponential changes in technology.

For example, university researchers partner with Facebook and gain access to data without the consent or permission of the individuals whose data is captured. Even if the data is anonymized, there are intelligent algorithms that can determine with reasonable accuracy which data belongs to whom. Facebook users can legally be experimented upon without their knowledge but the ethics are less clear.

In the midst of these grand questions there has been an increased need for professionals who are, for lack of a better term, data experts. In all areas of industry, there is a need to be first, or to be better than one's competition, to gain an advantage, or the upper hand in a negotiation. More than that, however, there is also the need to solve some very complex problems facing society at large such as homelessness, access to adequate health care, drug addiction, clean water and food supply, energy resources and climate change.

Given the tremendous consequences of the misuse of data, the harm that can be inflicted and lives affected, the need for clear guidelines for the ethical practice of the work that originates from data retrieval, storage, creation, simulation and aggregation is real and imminent.

3.6.2. Stakeholders

Just who are the stakeholders in the development of clear ethical standards in the teaching and practice of data science? In some ways we all are. However, there are different levels of stakeholders. Academic departments / researchers, students, corporate and community partners are all either directly or indirectly impacted or affected by the infusion of ethics and responsibility in the curriculum. When harm is done, blame will be assigned by the public whether deserved or not (in part) to these entities that had a hand in training the individual or individuals who are most responsible for that harm.

The good news is that ethics in data science and 'big' data curriculum is starting to receive national attention due to a few large issues that recently arose. In addition to the headlines mentioned previously there were several security breaches of large data brokers or retailers such as Equifax and Target, as well as racial bias that can affect facial recognition algorithms such as those used by Apple. The general public is beginning to take notice.

3.6.3. Broader Impacts

The biggest impact going forward is that we move from a position of cyber defense to cyber intelligence. Stakeholders are aware of and adhere to data security, ethics and privacy policies. The cross-industry standard process for data mining (CRISP-DM) is a model that is currently being used and will continue to evolve. Eventually, an algorithm on best practices of data science may be developed that is adaptive and changes with available technology.

Ideally, data science will be used ethically to address issues of social responsibility. Data scientists need to systematically omit biased datasets. Diversity in the field will help us recognize bias in code that is used to generate data. Data scientists use data science to make positive policy improvements. Stakeholders value the relationship between data privacy, ethics, bias and security. All of the key concerns, including, but not limited to social responsibility of data analysis projects, transparency, confidentiality, integrity and bias, availability of data, privacy concerns, storage and retrieval with security in place, who has access (permission), and training issues for faculty as well as K-12 educators, will be addressed early, often, and with socially beneficial motivations to guide the work.

The results will be seen in the development of more scientifically, professionally and socially responsible students, faculty and researchers in all sectors of the economy. The community at large and the planet will also benefit from more ethical data science research, as such work may also be instrumental in impacting and improving societal conditions as a whole (e.g., climate change, emergency systems, and medical research). These improvements will have a direct return on investment for each of the stakeholders identified previously.

3.6.4. Challenges and Visions Pertaining to Ethics

Lack of Training

The former U.S. Chief Data Scientist stated all data scientists ought to have some ethics training. Professional organizations (such as the American Statistical Association and the Association for Computing Machinery) have published ethical standards for the use of data in statistics and computer science, respectively.[54] Ethics training is a recommendation within the Federal Big Data [55] and National Artificial Intelligence Strategic Plans[56]. Many industry partners, including Intel, require data scientists to

[54] American Statistical Association. (2016). Ethical guidelines for statistical practice. Retrieved January 08, 2018, from http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx
[55] https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf
[56] https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

have ethics training. The rationale is because data science is being used to weaponize data. But the reasoning goes beyond malicious intent to include the prevention of inadvertent use of poor judgment with good intentions. But, how do you teach people to be ethical? Academia can teach ethical standards of behavior but cannot teach future professionals to be ethical.[57]

If ethics are being given proper attention in colleges and universities, then the topic is largely kept in silos and there is merely tacit understanding that the pertinent issues are being given the time and application that they deserve. It is no longer sufficient for the instructor to have limited knowledge of ethical data practices as it relates to traditional methods of research in the instructor's discipline.

Transparency

The Federal Trade Commission (FTC) published a report that attempts to define and understand many of the issues associated with data science and big data.[58] In partnership with the National Science Foundation (NSF), the Council for Big Data, Ethics, and Society was created in 2014 to bring to light the ethical considerations in this space such as the definition of human subjects which traditional research has relied on.[59] The underlying message is a lack of transparency about how the data is to be used and why. A few of the data brokers, such as Acxiom, have moved to a transparent model so that individuals can see the data collected and correct errors.

Consent/Permission Issues

Data scientists in industry and academia are often using data from underserved and disenfranchised people without their knowledge; worse still this data is being used to develop products for profitable advantages that do not help them (and may hurt them). At a minimum, if data scientists use data for something that is not transparent and/or haven't received consent, data scientists should think twice. Data scientists typically do not tell people if they decide to use their data for something other than for what they granted by consent. In fact industry and academia are required to lead precisely *because* the law/regulatory environment is so inadequate in protecting the rights of individual privacy and person.

The use of data obtained by illegal hacking is an emerging question that raises a host of ethical, practical, compliance, and legal concerns. Hackers typically host this data on the Darknet, which would expose a student or faculty member to several problems aside from the obvious ethical ones, i.e., transmission of computer viruses to the user's computer, and scripts which create a local repository of hacked data or illegal material. Data obtained by illegal hacking should not be used in an academic setting. In addition, aggregation queries can occasionally violate privacy.[60]

Machine Learning and AI Systems.

More and more often, we are relying on artificial intelligence. How do we deal with the unintended consequences of these thinking algorithms that try to do good but end up doing something bad instead (e.g., let's identify all of the students in second grade that are struggling in math so we can provide them with additional resources). The problem is that now the children have been labeled in a manner that will follow them for the rest of their lives. The current ethical framework is not designed to address the fact that we have outsourced human cognition to a thinking machine. Who will be responsible for creating that ethical framework? Who will manage that over time?

---

[57] Barocas, Solon, and Boyd, Danah, "Engaging the Ethics of Data Science in Practice", Communications of the ACM, vol 60, no. 11, p. 23-25.

[58] https://www.ftc.gov/system/files/documents/reports/bigdatatoolinclusionorexclusionunderstandingissues/160106bigdatarpt.pdf

[59] Council for Big Data, Ethics, and Society (2014). Retrieved January 08, 2018, from http://bdes.datasociety.net/

[60] Cook, J. (2017, November 26). Big aggregate queries can still violate privacy. Retrieved January 08, 2018, from https://opendatascience.com/blog/big-aggregate-queries-can-still-violate-privacy

Privacy and Discrimination Bias.

While obvious markers, such as race and national origin, may be removed from datasets, other data, which are retained, are often correlated with the omitted variable. A "good" model may very well discover these correlations. Even more problematic, when datasets are scrubbed of any obvious data which may result in unlawful discrimination, the outputs of any resulting models may still result in decisions that are discriminatory. By checking model results, a data scientist can guard against discriminatory practices. Further research in the machine learning field may yet yield tools which enable automated checking of model results for various biases.

For example, predictive policing, if not handled appropriately, can cause a model to self-reinforce. Police are guided to particular areas by a model, and make arrests. These arrests lead the model to assume that crime is increasing whether it is or not, resulting in a cycle of confirmation bias.[61,62,63] These and other examples should lead data scientists to scrutinize their recommendation engines for biases and yield better modeling decisions.[64,65]

Vision and Benefits of Social Good

Our vision is that ethics will be given sufficient attention such that when programs are being developed, no one is passing the buck. There will be a broader understanding and teaching of data ethics. Ethics becomes a part of data literacy such that it is a fundamental/foundational and integral component of K-12 education and undergraduate programs. We do not simply have students take an ethics class that is broad and general and call it done. As such, knowledge and understanding of ethical concerns will be a prerequisite for teaching courses in data science, statistics, and data literacy.

While we acknowledge that there is a lot of "social bad" use of data science, there is great potential to motivate and engage students in tackling meaningful "social good" projects in their communities that take advantage of available data. For example, the idea of data science as a social justice enabler (e.g., addressing homelessness, disasters and emergency relief, Smart Cities, Pro Publica analysis of recidivism models, Pro Publica/Consumer Reports differential pricing based on neighborhood for insurance) and the idea of data science to make meaning of data that non-profit organizations (e.g., museums, schools, community organizations) have collected but need help curating and analyzing is key.

Social good projects can also be used to attract a diverse group of students with broad experiences and interests to data science, and motivate student engagement. Specifically anecdotal evidence suggests, bringing meaningful problems to students attracts students with more diverse interests and experiences, and increases student engagement and retention. It highlights the key role of experiential education to model data science workflows.

---

[61] Brustein, J. (2017, July 10). The Ex-Cop at the Center of Controversy Over Crime Prediction Tech. Retrieved January 08, 2018, from https://www.bloomberg.com/news/features/2017-07-10/the-ex-cop-at-the-center-of-controversy-over-crime-prediction-tech

[62] Robertson, J. (2013, August 15). How Big Data Could Help Identify the Next Felon -- Or Blame the Wrong Guy. Retrieved January 08, 2018, from https://www.bloomberg.com/news/2013-08-14/how-big-data-could-help-identify-the-next-felon-or-blame-the-wrong-guy.html

[63] McKenzie, J. (2017, October 6). How Meetup Counters Invisible Sexism. Retrieved January 08, 2018, from https://civichall.org/civicist/meetup-counters-invisible-sexism/

[64] Brustein, J. (2017). The Ex-Cop at the Center of Controversy Over Crime Prediction Tech. Retrieved January 08, 2018, from https://www.bloomberg.com/news/features/2017-07-10/the-ex-cop-at-the-center-of-controversy-over-crime-prediction-tech

[65] Hardt, M. (2016, October 07). Equality of Opportunity in Machine Learning. Retrieved January 09, 2018, from https://research.googleblog.com/2016/10/equality-of-opportunity-in-machine.html

Many cities have found great value in participating in city-university partnerships through networks such as MetroLab.[66] South Bend, Indiana, for instance, has seen tremendous growth in students not only participating in public works but as a means of contributing intellectual rigor to city projects and attracting a talented workforce to stay and work in a small town long after leaving school.

Data Science for Social Good programs abound across the country in academic, industry, government, and the non-profit sectors. Academic-based programs called Data Science for Social Good (DSSG) are now in Chicago (University of Chicago), Seattle (University of Washington), Atlanta (Georgia Tech) and others. The Atlanta DSSG's 2017 class was started with a mix of graduate and undergraduate students with support of the South Big Data Hub[67]. Also, Valparaiso University is attempting to move the idea of data science for social good into undergraduate realm.[68] Previous projects completed last year included working with U.S. Geological Survey to identify microclimates in the Indiana Dunes and with Upbring (from Texas), to try and understand how flooding can negatively impact foreclosures and minority abuse. While most academic institutions have Centers for Community Engagement, some faculty members may be unaware of their existence. Professional societies and departments need to work to build connections.

Non-profits, Industry, professional societies and government are also now in the game. Non-profits such as DataKind [69]match professional data scientists to social good non-profit partners. Professional societies such as INFORMS[70] offer a pro-bono data science program that matches members to socially relevant industry problems. Industries, such as IBM[71], have created their own data science for social good program. Taken together government is now taking a strong interest in the benefits of data science for societal good but not only funding programs such as the Big Data Hubs and Spokes[72] and the Data & Society[73] but this winter convening a workshop to bring together all the different groups doing this work to forge a Data Science Corps (akin to the Peace Corps for data science). The first meeting took place at Georgetown University Dec 2017[74].

3.6.5. Specific Skills and Resources

Ethics in Data Science should start with compliance to laws and regulations, paying careful attention to the ownership of data, and respect for copyright. Any data scientist or data engineer should fully understand the provenance of data sets, including how the data was/is collected, updated, and stored.

Countries and regions, such as the European Union, may offer more privacy protection than the United States. Data Scientists should be aware of these laws, which may restrict data movement. Cloud computing is of great benefit here, since it offers the assurance that datasets are stored in a particular country.

---

[66] MetroLab Network. (n.d.). Retrieved January 8, 2018, from https://metrolabnetwork.org/
[67] dssg-atl.io
[68] Introduction to Data Science" DATA 151. (2017). Retrieved January 08, 2018, from https://vu-data151-spr17.wikispaces.com/
[69] DataKind(n.d.). Retrieved January 08, 2018, from http://www.datakind.org/
[70] INFORMS(n.d.)
[71] IBM Data Science Experience. (n.d.). Retrieved January 08, 2018, from https://datascience.ibm.com/
[72] Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications. (n.d.). Retrieved January 09, 2018, from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505264
[73] Data & Society https://datasociety.net/
[74] Data Science Corps Conference. (2017). Retrieved January 08, 2018, from https://mccourt.georgetown.edu/DataScienceCorp

In any questionable circumstance, a data scientist should consult with the owner of the data, seek legal advice, or follow the principle of most restriction. By this we mean that if two or more countries are involved for a particular dataset, follow the legal and compliance framework of the most restrictive country. This ensures that all parties copyright and legal restrictions are respected.

An emerging area in ethics in data science is the potential bias in predictions or results of models. Academic researchers and industry are just beginning to grapple with this issue in a meaningful way.[75] The University of Michigan is currently running a Data Science Ethics MOOOC on Coursera),[76] where we can glean some recommendations in this area:

- Be aware that historical biases against marginalized groups *will be present in historical data*.
- Try to understand why a black-box model reaches biased outcomes by slightly changing the features and examining the results.
- Bias can be introduced by having insufficient amounts of data or non-representative samples.
- Examine your features - are they correlated with a feature that should obviously be dropped or obviously included, such as an indicator or race, religion, or national origin?
- Collect the minimal amount of data you need.
- Creating aggregates is a useful way to avoid exposing individual behavior. For instance, instead of using individual diabetes patients records, aggregate them. Diabetes patients will have similar profiles, which a model can use without being exposed to individual behavior[77].

3.6.6. Concrete First Steps

Seek guidance from existing scientific, academic, business, and legal standards of best practices. Beyond that, bring people from all of these communities together to create and/or adopt guidelines/best practices around data ethics as it applies to the various stakeholder groups and contextual concerns. Put it into one unifying set of principles and practices.

Attention should be given to the fact that curriculum is already credit-heavy particularly in undergraduate education. Provide some guidance to curriculum developers on how to embed activities into existing courses. One suggestion is to create role-based education and training (e.g., as is customary from a cyber security standpoint) around ethics in data science.

Documenting the existing social good activities in data science and highlighting them.

Highlight successful (graduate student focused) Data Science for Social Good initiatives. Also the STATCOM program (pro bono statistics consulting by graduate students) (See vision and benefits of social good above)

1. Identify best practices to translate social good projects to the undergraduate level (perhaps jointly with graduate students).
2. Support for faculty time (and staff assistance) to identify and nurture relationships with local organizations

---

[75] Tannam, E. (2017, October 26). How can we mitigate ethical and privacy issues in data science? Retrieved January 09, 2018, from https://www.siliconrepublic.com/enterprise/ethics-data-science-bias

[76] Jagadish, H. V. (n.d.). Data Science Ethics. Retrieved January 08, 2018, from https://www.coursera.org/learn/data-science-ethics

[77] Martin, E. R. (2014, May 19). EMCVoice: The Ethics Of Big Data. Retrieved January 08, 2018, from https://www.forbes.com/sites/emc/2014/03/27/the-ethics-of-big-data/ - 3acb9426852e

3. Create a "Data Science for Social Good" in a box toolkit for establishing programs
4. Have a resource for to help build connections between data scientists and communities.

3.7. Faculty, Staffing, and Collaborations

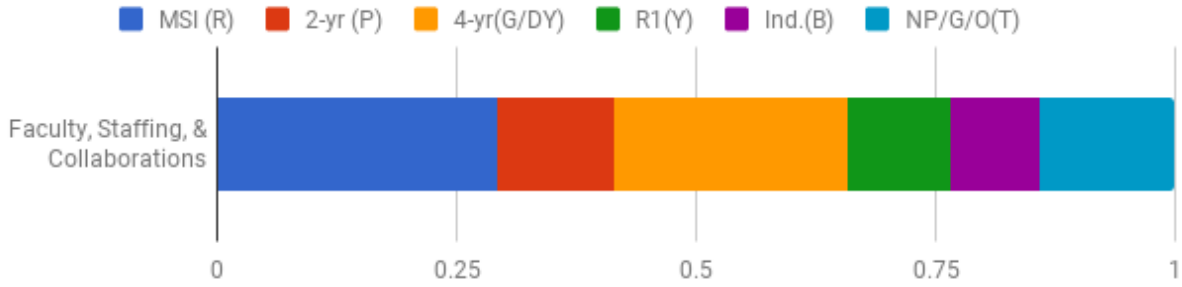


Figure 9: Ratio of workshop participant contributions to this section by institution type (n=205)

3.7.1. Introduction

The interdisciplinary nature of data science as well as the lack of a clear consistent definition of data science poses challenges for developing and implementing a data science program in colleges and universities. The challenges, which include institutional barriers and lack of meaningful collaborations across disciplines and with industry, are opportunities for enhancements in academic settings.

3.7.2. Stakeholders

Stakeholders include faculty, administrators, industry partners, non-profit organizations, and government agencies.

3.7.3. Broader Impacts

Society as a whole will benefit from more rapid advances in all fields from better utilization of data science in domain specific ways. This hinges on adequately trained and available faculty and staff. Universities can benefit from data science application by more effectively coordinating the efforts of multiple divisions in order to find effective approaches to common needs, avoiding duplication of efforts, and more efficiently utilizing available resources. Students can benefit by gaining skills that will allow them to more effectively apply data science skills to their field of choice, hence becoming more attractive to employers. Industry benefits by having access to a supply of workers who can hit the ground running because they will not only know the theory, but the practical issues involved in applying data science. Similar to academic institutions, industry can also benefit by being able to bring new innovations to markets, reduce waste, and enhance the lives of their stakeholders. Data science has broad sweeping implications for itself as a discipline and all other academic disciplines. By eliminating barriers, the cross-disciplinary nature of data science/analytics is allowed to develop more fully and effectively.

3.7.4. Challenges and Visions Pertaining to Faculty Staffing, and Collaborations

Institutional Barriers to Developing and Implementing Data Science/Analytics Programs
The siloed bureaucratic structure inherent in most universities creates a set of institutional barriers to the development of sustainable data science/analytics programs. Many perceive data science/analytics as a threat to “business as usual”. Others try to claim sole ownership of the field and exclude other disciplines. Added to this are the institutional barriers created by the relationships of the departments/colleges with

the central administration. There is often a lack of understanding by administrators about the benefits and the resource needs of data science/analytics programs. Some institutions may view data science/analytics degrees as a way to increase revenue quickly, but not as a viable sustainable long-term degree program. Additionally, slow and layered degree program approval processes discourage faculty and departments from embarking on developing new programs. Some states may add additional layers to this process, even involving other universities in the approval process. One final barrier is the difficulty in obtaining the qualified faculty resources required for data science/analytics programs. In times of tightening university budgets, retiring tenured faculty are not being replaced with new tenure-track faculty. And the faculty that is hired may not have the requisite skills to teach courses in a data science/analytics program.

Intra-Institutional Interdisciplinary Collaborations (Role of Faculty)

Data science as a multifaceted, interdisciplinary, emerging field of study that "focus [es] on the processes and systems that enable the extraction or insights from data in various forms, structured and unstructured"[78] touches every discipline and industry. Faculties are major data science stakeholders, and are at the hub of the data science wheel. Although the jury is still deliberating on the boundaries of content areas within data science, there is agreement that the following content areas are core: statistics, modeling, programming, data mining, machine learning, visualization, ethics, research design, databases, algorithms, parallel computing, and cloud computing.[1] Current faculty that consider themselves in the data science space tend to hold degrees in disciplines related to the aforementioned content areas. These faculties are central to developing; disseminating; facilitating knowledge acquisition; and advancing data science theories, principles, applications, and research.

Due to the cross-disciplinary nature of data science/analytics, working across the traditional disciplines is desirable at most institutions. Regardless of the departmental structure selected to house a data science degree program, faculty from the different disciplines will have to collaborate in ways that faculty have not historically operated. As a result, many questions will surface. Will tenure and promotion criteria change with the advent of more collaborative research outputs? How will existing faculty retool themselves through professional development so that they are adequately equipped to effectively contribute to teaching and research in the field of data science and academic disciplines that will benefit from and/or be impacted by the applications derived from data science, such as Analytics, Health Informatics, and other data-centric academic disciplines.

Faculty Training and Credentialing

To address the challenge of faculty who do not have the expertise and need to retool, many faculty bolster their expertise and credentialing in the data science arena by attending industry- and academic-driven data science training sessions and/or conduct and publish research related to data science. Training videos and classes offered by NVIDIA, Intel, and PyData are examples of industry-driven data science training. From an academic-driven data science training perspective, the University of California Berkeley, New York University, and University of Washington are offering training sessions and providing the educational materials as a result of funding from the Moore and Sloan Foundations.[1] The National Science Foundation is also providing funding to support academic training in the area of data science.[1] In an effort to advance data science as an academic discipline, faculty should take advantage of the existing data science training activities, and trainings should be made accessible to faculty from all types of academic institutions.

With the recent advent of master's and bachelor's level degrees in data science, there is a need for more doctoral level programs in data science and analytics. Because the spokes of the data science wheel are so varied, yet interwoven, faculty with degrees in related disciplines previously delineated would continue to

[78] Realizing the Potential of Data Science (2016) - Final Report from NSF Computer and Information Science and Engineering Advisory Committee, Data Science Working Group

be more prevalent until universities start offering doctoral degrees in data science. Therefore, over the course of the next decade or so, a paradigm shift in faculty credentialing for the data science academic field is needed.

Another way to train and credential existing faculty could be through non-traditional post-doctoral fellowships (post-docs) for existing faculty. The post-doc research projects could be structured around developing best practices in data science pedagogy as well as the application of data science to the faculty member's discipline of choice because faculty are needed with expertise throughout the data life cycle and full data science ecosystem. While there are some books, research projects, and journal articles dealing with data science pedagogy and other aspects of data science, more are needed.

The collection of data is inevitable given technological advances of today and tomorrow. Therefore, the science of data including the knowledge, acquisition, management, storage, manipulation, and interpretation of data is important and necessary for all learners regardless of discipline. In "Keeping Data Science Broad," faculty serve as the conduit for knowledge acquisition, dissemination, and utilization. Therefore, not only should faculty teaching in data science programs engage in industry- and academic-driven data science training, faculty from all disciplines that are responsible for or interested in the application of data science to their fields and professions should participate in data science training. The development of interdisciplinary application-based data science training would also facilitate interdisciplinary collaboration and research projects.

Benefits of Interdisciplinary Collaborations

All faculty can benefit from data science by being able to more effectively train their students to be workforce ready as well as being able to make more meaningful research contributions that are informed by both data science and domain specific phenomena including, but not limited to, astronomy, biology, business, chemistry, environmental science, medical data, political science, physics, social sciences, behavioral sciences, the arts and humanities. For this reason, an introduction to the specific discipline and data science course should be embedded in all disciplines. For example, multi-level courses could be cross-listed courses and team taught - an art course that teaches "Processing" could be an advanced art class using modern digital techniques, but also a computer science elective/topics course, because it teaches a new programming language. Such embedded courses would truly foster interdisciplinary collaboration.

Institutions will have to decide the appropriate departmental structure for their data science or related degree programs, the tenure and promotion criteria for faculty teaching in the data science program, faculty credentialing criteria to meet accreditation standards, how to provide professional development to support the contributions that faculty can make to the field of data science, how to provide training in the areas of pedagogical best practices and team teaching, and how to partner with industry and other stakeholders for curriculum, student, and faculty development.

Collaborations with Industry

There is a lot of discussion around the lack of collaboration between academia and industry. There are a number of different paths to foster collaboration, including research, pipeline of talent, skills training and workforce development, experiential learning/ "customized training" (e.g., solving real world problems for real companies) that improves the learning process, faculty exchanges, and philanthropic gifts to the academic institutions. Academia tends to promise industry the best and brightest students as the outcome from collaborative efforts. From an industry perspective the expectation is that academic institutions would create a business case for acceleration of the adoption of data science and deliverables/outcomes as a return on the investment industry is making. It is important for academia to be able to build a business case.

Collaboration between industry, the workforce system (primarily serving the underserved) and academia is key. The relationship between industry and academic institutions create workforce pipelines, but this does not produce all of the data professionals that are needed in the workforce. It is important for academia to gain insights from industry about the skills and experiences that are needed in the workforce. The goal is not just to educate students, but to prepare them to go into the workforce. The amount of money that is available for academic institutions from industry is broad if they can focus on preparing students with a curriculum in data science that is more aligned with industry requirements.

There is a lack opportunities for students at non-R1 institutions to engage with industry and workforce systems. Faculties are not spending enough time educating students on how to engage industry and the workforce systems already required by federal law to engage with it. It is important for academics to be proactive.

3.7.5. Specific skills and Resources

Administrators and faculty need training not only in data science methods but the project management skills, the benefits, and full complement of skills needed for data science work. The curriculum that exists in different programs should be made accessible to faculty creating new programs at their institutions. Finally, there needs to be a concerted effort to foster the ownership of data science across multiple faculties at institutions around the nation.

Cooperative ethos among collaborators is also necessary across disciplines. Sharing the stories of people who are working on interdisciplinary projects so that there is a broader understanding of what is possible and what is needed.

3.7.6. Concrete First Steps

1. Administration support for facilitating collaborative work across department boundaries. Cross-listed courses and co-taught courses.
2. Identification of common needs for data management and analysis across a wide variety of fields.
3. Catalog of existing programs and courses across components of the college or university.
4. Identification of best practices across institutions.

3.8. The Pipeline to Higher Education

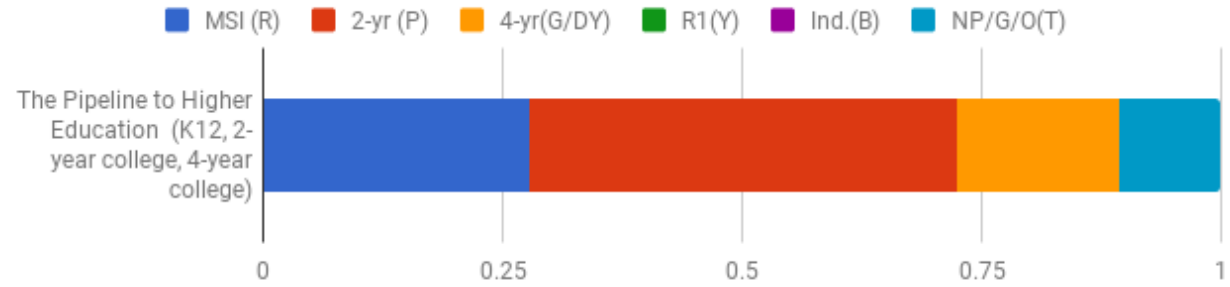


Figure 10: Ratio of workshop participant contributions to this section by institution type (n=47)

3.8.1. Introduction

Future growth of Data Science careers is assumed to require a four-year degree. However, not all students are prepared to pursue a four year degree. There are a number of factors that come into play. College tuition can reach tens of thousands of dollars, putting a four year degree out of reach for many students. In-state tuition at a community college is often a fraction of the cost of a four year degree program and students at community college may be eligible for financial aid. Those that are ineligible for financial aid may need to balance the need for higher education with full or part time employment to cover the cost of tuition. Some high school students do not feel comfortable in mathematics, sciences or advanced classes and need another opportunity. Likewise, some high school students do not have a clear idea of the type of degree or career they would like to pursue.

Community (or two-year) colleges and other training providers play an enormously important role in higher education in the United States. According to the Community College Research Center, 9 million undergraduates were enrolled in public two-year colleges during the 2015-2016 academic, with nearly 4 million full-time.[79] This accounts for nearly 40% of undergraduates, with nearly half of all graduates that year enrolling at a two-year college in the previous ten years. These institutions also serve a key role for low-income, minority, and first-time college students, as do other private or for-profit credentialing providers.

3.8.2. Stakeholders

Stakeholders include parents of K-12 students, teachers of K-12 students (including STEM and non-STEM coursework), high school students who are exploring degree programs and career paths, students at two-year colleges who desire to transfer to four-year colleges, faculty at two-year and four-year colleges, administrators at two-year and four-year colleges, and administrators at two year and four year colleges.

3.8.3. Broader Impacts

Data is involved in everything within the realm of K-12 education and life. Parents need to become aware of opportunities for students as they move into college, but also to advocate for a larger presence of data and technology at the school and district levels. Students in K-12 should become aware of Data Science career pathways and higher educational opportunities through meaningful pre-college experiences. Two-

[79] Columbia University. (n.d.). Community College FAQs. Retrieved January 09, 2018, from https://ccrc.tc.columbia.edu/Community-College-FAQs.html

year colleges give students access to any degree or college. The curriculum should introduce interesting problems so students are motivated to pursue Data Science.

Many individuals in today's data science workforce are coming from doctoral or master's degree programs, which have seen a dramatic increase in recent years. While these advanced degrees are valuable, it is not economically feasible for all data scientists to complete four years of an undergraduate degree, then a one or two year master's program before they can undertake useful work. Ensuring the future growth of the workforce requires an expansion to four-year and two-year degrees. At their best, two-year colleges give students access to many associate degree options or institutions to continue towards a bachelor's degree. Past mistakes from mathematics, statistics (with a long list of required courses that limit completion rates), and computer science (that have historically had low numbers of completions for women, low-income, minority, and first generation students) need not be repeated. Implementing the strategies outlined in Lyon and Denner will be a necessary, though not sufficient, step forward.[80]

Through a concerted effort to align coursework at the K-12, two year and four year college level, there will be more students in the pipeline for Data Science degrees. Faculty and administrators are needed to ensure that Data Science is incorporated across multiple disciplines.

3.8.4. Challenges and Visions Pertaining to the Pipeline to Higher Education

The opportunities for data science at community colleges are dramatic. The interim report of the *National Academies Envisioning the Data Science Discipline: The Undergraduate Perspective* study noted: "Community colleges are well qualified to be highly effective providers of data science education while also serving as important partners for 4-year institutions that are considering the emerging role of data science education.[81] Community college programs can serve to (1) be an entry point to inspire and attract diverse student populations to data science; (2) permit existing members of the workforce to retrain or obtain specific new skill sets to complement their education and experience; (3) create mechanisms by which students can certify specific or general skill sets with certificates or associate's degrees; (4) build foundational, translational, ethical, and professional skills to support matriculation into 4-year college data science programs; and (5) provide opportunities for advanced high school students to begin data science training early. The majority of these purposes support undergraduate education objectives, while also targeting the specific needs of industry. Institutional, industry, and government partnerships are all important to the development of data science education that meets these objectives for community colleges."[82]

An exemplar of the ways that community colleges support the workforce is the story of DJ Patil, the first United States Chief Data Scientist. He credits his community college education for the gifts of confidence, ability to write, and love of mathematics that led him to an advanced degree and productive work in industry and government.[83]

---

[80] Lyon, LA & Denner, J (2017) Community Colleges: A Resource for Increasing Equity and Inclusion in Computer Science Education, Communications of the ACM, Vol. 60 No. 12, Pages 24-26

[81] The National Academies Press (2017, September 27). Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report. Retrieved January 08, 2018, from https://www.nap.edu/catalog/24886/envisioning-the-data-science-discipline-the-undergraduate-perspective-interim-report

[82] The National Academies Press (2017)

[83] Holst, L. (2015, May 06). Email from DJ Patil: "How I Became Chief Data Scientist". Retrieved January 08, 2018, from https://obamawhitehouse.archives.gov/blog/2015/05/06/email-dj-patil-how-i-became-chief-data-scientist

While community college provides additional opportunities for advancement, students who start at a community college do not usually have the same exposure to the courses necessary to pursue a four-year degree in Data Science. There is inconsistency between the courses offered at community colleges and the courses that will be accepted for a four year degree program. Only those community college classes that align with the program of study at a four year university are eligible for financial aid. This inconsistency means that students at community colleges that are interested in pursuing degrees in Data Science may lose a portion of their financial aid. There is evidence to suggest that there are fewer employment opportunities for associate degree holders. Without articulation agreements, community college students, many of whom are first generation college students, won't be exposed to Data Science in their early part of education.

There should be multiple pathways for students to pursue four year degrees in Data Science. In order to encourage broadening the population of students pursuing Data Science, parents and students should become aware of opportunities and career pathways before entering the University space, while teachers and administrators need to grow capacity to integrate technology and data into their schools. There needs to be an articulations agreement between community colleges and four year colleges regarding what courses should be in the core of the first two years of a four year Data Science program.

3.8.5. Specific Skills and Resources

There is a need for flexible articulation agreements between 2-year colleges and 4-year colleges for data science programs. At present, if an equivalent course is not offered at 4-year colleges in that state it may not be feasible for the course to count for elective credit. More useful are equivalent courses that are included in BA programs that can count towards the major. There is a lack of 2-year programs to model an associate degree at other institutions. There is a perceived lack for jobs for associate degree holders There is a need for training for teachers. There is a need for education for K-12 teachers, administrators and policymakers, as well as conversations the bridge the gap between K-12 and universities. There is a need for creative approaches to address these challenges that include:

- Developing a clearinghouse of courses, certificate, and degree programs
- Creating different entry pathways that could complement intro CS and intro statistics
- Encouraging schools to offering data science courses as general education mathematics. As an example, the New York State SUNY-GER (general education requirements) for Mathematics have the following learning outcomes: Students will demonstrate the ability to: 1)interpret and draw inferences from mathematical models such as formulas, graphs, tables and schematics; 2) represent mathematical information symbolically, visually, numerically and verbally; 3) employ quantitative methods such as, arithmetic, algebra, geometry, or statistics to solve problems; 4) estimate and check mathematical results for reasonableness; and 5) recognize the limits of mathematical and statistical methods. It may be feasible to design an introductory data science course [84]that addresses these learning outcomes as an additional pathway. Such a course could allow an institution not able to put forward a whole data science AA program to help expose students to the field.
- Creating partnerships between 2 year and 4 year schools to help develop and update flexible articulation agreements for transfer students.
- Build on existing programs (such as the American Statistical Association's Two Year College Data Science Summit, https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx) to foster developments in curriculum and faculty development.

[84] UC Berkeley Foundations of Data Science. (n.d.). Data 8: The Foundations of Data Science. Retrieved January 08, 2018, from http://data8.org/

3.8.6. Concrete First Steps

1. Outreach to parents, students, and guidance counselors about data literacy and Data Sciences
2. Data-science based career opportunities available for associate degree holders
3. Curricular materials demonstrating connections to standard curriculum
4. Support for material technologies in students' hands

Create the following partnerships and incentives:

- Partnerships between 2 year and 4 year schools in the form of professional development training workshops for faculty.
- Partnerships between 2 year and 4 year schools in the form of frequent communication and networking between faculty/administrators at the two types of institutions regarding the development and updating of curriculum.
- Incentives for four year college administrators, faculty, and staff to engage with two year colleges (and vice versa)

4. Top "Asks" for the Future

Participants were asked to give their top "asks" of the data science community to ensure a bright future for data science education. The group then voted on each ask, resulting in the top ten asks, or next steps, listed below.

1. Foster partnerships between different institutional types i.e., 2-year and 4-year college partnerships, HBCU, R1, industry and alternative groups.
2. Provide flexible pathways into data science education.
3. Time & space to discuss collaboration, especially with respect to curriculum "holes."
4. Hiring female faculty, faculty of color, and female faculty of color because it's hard for students to "become something they have never seen".
5. Provide free (or subsidized) access to data science resources.
6. Provide access to data literacy tools and resources to students and parents of underrepresented backgrounds/populations/communities.
7. Supply access & training for JupyterHub in data science instruction.
8. Provide examples of curriculum for 2-year colleges degrees, certificates or pathways.
9. Provide more realistic data science-focused collaboration between industry and K-12.
10. Develop data literacy resources for K-12 teachers.

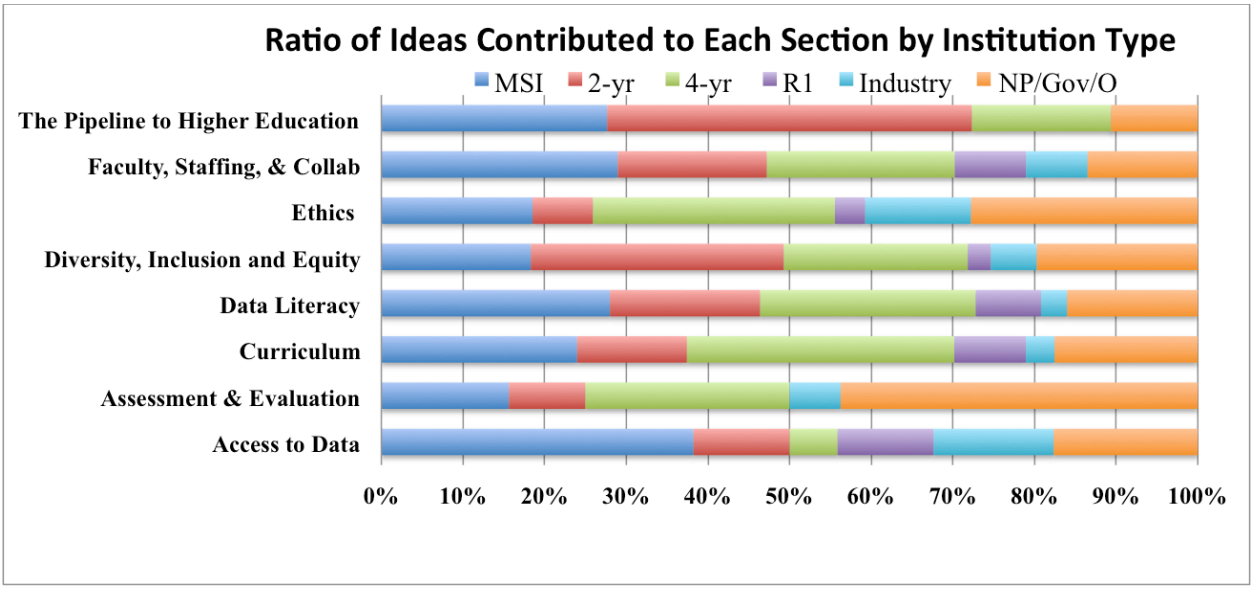
5. Conclusion

There is clearly a tremendous interest in data science today, driven by a changing and more data-integrated society, as well as the changing nature of the workplace and our culture. In the 21st century, scientific exploration, industry growth, government policies and individual choices will all rely heavily on the use and integration of data science into new arenas. As such we must educate the future scientists, captains of industry, political leaders, and citizens to understand the power and limitations of data and how to appropriately incorporate techniques into their lives. This requires a fundamental shift in the traditional education process due to the nature of integrating data into all of these new fields, the speed at which new techniques arise and the need to include the entire populous in the efforts at varying levels. For this scale of effort an " all hands on deck" approach is needed where all of our educational

institutions, government, industry, and non-profit partners play a key role in equipping and maximizing the education, opportunity, and promise of this generation and the next.

In the US about 65% of students go to college and in 2017 data indicates that 39 percent of these undergraduates attended public two-year colleges; nine million undergraduates were enrolled in public two-year colleges in 2015-16.[85] Overall, minority-serving institutions enrolled 4.8 million students or roughly 28 percent of the nation's total enrollment.[86] A recent paper released by the American Council on Education utilizes 10 years of data from the National Student Clearinghouse to examine enrollment and outcomes at MSIs, and showed a completion rate of 66.7 percent, compared to a federal graduation rate of 43.9 percent, for exclusively full-time students at private four-year HBCUs as well as a completion rate for exclusively full-time students at public two-year Hispanic-Serving Institutions (HSIs) of 40.3 percent, compared to the federal graduation rate of 25.5 percent. This paints a more complete picture of the important contributions HBCU's, HSI's and community colleges make to the nation.[87]

Therefore, a unified conversation is necessary and educational institutions included in that conversation should touch all elements of our society. Here we participate and facilitate that conversation among institutions of higher-education spanning community colleges, 4-year liberal arts colleges, and research universities, including minority-serving institutions (MSIs), Historically Black Colleges and Universities (HBCUs), Hispanic-Serving Institutions (HSI's), and Asian American and Native American Pacific Islander-serving institutions (AANAPISIs). This report synthesizes the views of close to 60 faculty representatives from these institutions with or trying to create data science programs. The Vision-Challenge areas described above were organically derived through participant input and clustered based on number and strength of the idea within the group.



[85] McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., Gebrekristos S, Zhang J, Rathbun A., Barmer A., Bullock Mann F. (2017). The Condition of Education 2017. NCES 2017-144. National Center for Education Statistics.
[86] Espinosa, L. L., Turk, J. M., & Taylor, M. (n.d.). Pulling Back the Curtain: Enrollment and Outcomes at Minority Serving Institutions(Rep.). The American Council on Education.
[87] Espinosa et al.(n.d.)

Figure 11: Represents the breakdown of ideas and input among the different institution types across the eight vision areas.

There was a tremendous response to the idea of envisioning the data science curriculum at multiple educational levels, faculty and cross-disciplinary engagement, as well as cross-institutional and cross-sector partnerships. Total participants included:

- 19 Minority-serving institutions (MSIs, HBCUs, HSIs, and AANAPISIs)
- 11 4-year colleges - non R1
- 10 Non-profit / Government / Organizations
- 7 Community colleges
- 8 R1 institutions
- 3 Industry representatives

6. Participant List

This document was developed through the contributions of the participants and organizers of the Keeping Data Science Broad Series. A special thanks and appreciation to program committee members (bolded) and all of the participants involved in contributing input through the community input form, content generation and writing of white papers as a part of the workshop, sponsors, and editors.

Prefix	First Name	Last Name	Job Title	Company
Prof.	Srinivas	Aluru	Professor	Georgia Tech
Dr.	**Stephanie**	**August**	Program Officer, Education and Human Resources/DUE	National Science Foundation
Dr.	**Chaitanya**	**Baru**	Senior Advisor for Data Science	National Science Foundation
Mr.	**Rene**	**Baston**	Executive Director, NE Big Data Hub	Columbia University
Dr.	Jesse	Bemley	Professor	Bowie State University
Dr.	Jason	Black	Associate Professor	Florida A & M University
Mr.	Hugo	Claros	Independent	Independent
Dr.	Quincy	Brown	Program Director	AAAS
Dr.	Lillian	Cassel	Professor and Chair	Dept CSC, Villanova University
Dr.	**Melissa**	**Cragin**	Executive Director, Midwest Big Data Hub	University of Illinois
Ms.	Catherine	Cramer	Senior program Developer	New York Hall of Science
Dr.	Alex	Dekhtyar	Professor	Department of Computer Science, Cal Poly, San Luis Obispo
Ms.	Angela	Dingle	Consultant	University of the Virgin Islands
Dr.	Maurice	Edington	Vice President	Florida A&M University
Mr.	Shawn	Firouzian	Mathematics Instructor	MiraCosta Community College
Dr.	Shawnta	Friday-Stroud	Dean, School of Business and Industry	Florida Agricultural and Mechanical University
Mr.	**Melvin**	**Greer**	Chief Data Scientist, Public Sector	Intel Corporation
Mr.	Darold	Hamlin	President	Emerging Technology Consortium

Mr.	John	Hamman	Dean	Montgomery College
Ms.	Julie	Hanson	Professor of Mathematics and Statistics	Clinton Community College
Mr.	Charles	Hardnett	Applications Architect	FirstNet
Dr.	Leshell	Hatley	Assistant Professor	Coppin State University
Mr.	Al	Herron	Assistant Director of IT Department	The Breakfast Group
Prof.	**Nicholas**	**Horton**	Professor of Statistics	Amherst College
Dr.	**Tasha**	**Inniss**	Director of Education and Industry Outreach	INFORMS
Dr.	Lethia	Jackson	Professor	Bowie State
Prof.	Lewis	Johnson	Assistant Vice President Strategic Planning	Florida A&M University
Dr.	**Kari**	**Jordan**	Director of Assessment	Data Carpentry
Dr.	**Nandini**	**Kannan**	Deputy Division Director (Acting)	National Science Foundation
Dr.	Jacob	Koehler	Teacher	The New School
Mr.	**Brian**	**Kotz**	Professor	Montgomery College
Mrs.	Kathryn	Kozak	Mathematics Instructor	Coconino Community College
Dr.	Amy	Langville	Professor, Mathematics	College of Charleston
Prof.	Velma	Latson	Lecture	Bowie State University
Prof.	Christopher	Malone	Data Science & Statistics	Winona State University
Dr.	Brandeis	Marshall	Associate Professor	Spelman College
Dr.	Shannon	McKeen	Data Analytics instructor	UNC Chapel Hill, Dartmouth College, Wake Forest
Mr.	Costa	Michailidis	Facilitator	Knowinnovation
Dr.	Selvarajah	Mohanarajah	Faculty	UNC- Pembroke
Dr.	Tuan	Nguyen	Assistant Professor	University of Evansville
Prof.	Patricia	Ordonez	Associate Professor	University of Puerto Rico Rio Piedras
Prof.	Gene	Park	Associate Professor	Loyola Marymount University
Dr.	David	Potenziani	Senior Informatics Advisor	IntraHealth International
Prof.	Lijun	Qian	AT&T Endowed Professor	Prairie View A&M University
Dr.	Danda	Rawat	Associate Professor	Howard University
Dr.	**Renata**	**Rawlings-Goss**	Co-Executive Director	South Big Data Hub
Dr.	Herman	Ray	Associate Professor, Director	Kennesaw State University
Ms.	Donnalyn	Roxey	Facilitator	Knowinnovation
Mrs.	**Mary**	**Rudis**	Assistant in Instruction	Bates College
Dr.	Daniel	Runfola	Director, Data Science Program	William and Mary
Prof.	Mugizi	Rwebangira	Assistant Professor	Howard University
Dr.	Jennifer	Salazar	Communications Director	Georgia Tech Ideas
Prof.	Karl	Schmitt	Director of Data Sciences	Valparaiso University

Dr.	Manju	Shah	Program Director	Wake Technical Community College
Dr.	Lior	Shamir	Assoc. Prof.	Lawrence Technological University
Mr.	Dominic	Sirianni	PhD Candidate	Georgia Institute of Technology
Prof.	Brian	Spiering	Professor of Data Science	University of New Haven
Prof.	Suzanne	Smith	Assistant professor	Johnson County Community College
Dr.	Dale	Smith	Consulting Data Scientist	Self-Employed
Dr.	Sonya	Stephens	Interim Dean	Florida A&M University
Dr.	**Sarah**	**Stone**	Executive Director, eScience Institute	University of Washington
Dr.	**Frank**	**Stomp**	Associate Professor	Navajo Technical University
Prof.	Dennis	Sun	Assistant Professor of Statistics	Cal Poly
Dr.	Cheryl	Swanier	Chair	Claflin University
Ms.	Allison	Theobold	Graduate Student Teacher (PhD candidate)	Montana State University
Dr.	Elizabeth	Tipton	Professor of Decision Sciences	Eastern Washington University
Mr.	Reginald	Vigilant	Chief Operating Officer	OMNI Systems, Incorporated
Ms.	Erica	Webb	Director of Special Projects	Xavier University of Louisiana
Dr.	Kennedy	Wekesa	Dean	Alabama State University
Dr.	Lilian	Wu	Program Executive, Global University Programs	IBM
Dr.	Pei	Xu	Assistant Professor of Business Analytics	Auburn University