New York Hall of Science

DRAFT*****White Paper*********DRAFT**

Data Science for All (DS4All)

Stephen Uzzo, PI | Catherine Cramer Co-PI

DS4All actively identifies knowledge and resource gaps in data literacy among diverse communities, with a goal to help lifelong learners of all ages become data literate.

Data Science for All is an initiative of the New York Hall of Science and the Northeast Big Data Innovation Hub (NEBDIH) at Columbia University Data Science Institute. The New York Hall of Science is one of the founding members of the NEBDIH in 2015. This report represents the Discovery Phase of the initiative and the various activities, outcomes of those activities, as well as next steps in integrating the findings into practice.

This material is based upon work supported by the National Science Foundation under Grant no. 1636736 & 1550284)

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

Introduction

"I'm proud to join the students, teachers, businesses, and non-profit organizations taking big new steps to support computer science in America's schools. Learning these skills isn't just important for your future – it's important for our country's future. If we want America to stay on the cutting edge, we need young Americans like you to master the tools and technology that will change the way we do just about everything."

- President Obama, December 2013, on Computer Science Education Week

"Providing access to CS is a critical step for ensuring that our nation remains competitive in the global economy and strengthens its cybersecurity."

- President Obama, January 2016, on the announcement of CS for All

In 2013, the Obama Administration began a multi-pronged, widespread and highly effective campaign to give American students and teachers access to computer science. The campaign began with filling several ubiquitous and previously unmet needs – lack of connectivity, lack of hardware, and lack of access to tech jobs. Through comprehensive dissemination and uptake, provided by initiatives such as ConnectED and TechHire, the connectivity divide in US public schools has been cut by about half since 2013; hundreds of employers have partnered with cities, states, and rural areas to expand access to tech jobs; and thousands of students have access to laptops for the first time. This successful implementation plan built on longstanding efforts at the local, state and federal levels to raise the level of STEM career access, but was narrowly focused on one specific area: computer and coding skills.

In January 2016, President Obama announced Computer Science for All (CS4All), providing $4 billion in funding for states and $100 million directly for districts to increase access to K-12 computer science by training teachers, expanding access to high-quality instructional materials, and building effective regional partnerships. Aimed primarily at K-12 schools, CS4All relatively quickly gained uptake through a comprehensive and strategic plan utilizing a wide range of resources and networks such as:

- School district leaders;
- Nonprofits such as Code.org, Teach for America, the National Math and Science Initiative, 100Kin10;
- Private philanthropy;
- Public science events such as the US Science and Engineering Festival;
- Federal level agencies such as the Department of Education, NSF, the Corporation for National and Community Service, the Department of Defense, and the US Patent and Trademark Office.

However, while the stated goal of CS4All is to give US citizens the ability to solve complex problems, providing them with access to computers, software and broadband connectivity is just the beginning. These resources do indeed provide the *ability*. What's missing are the problems themselves – and these are identified, explored and solutioned through *data*. Data provide the information, the evidence to solve the problems that computer skills are applied to. Just in the last few years there has been a marked shift of emphasis on workforce and educational needs – from computer science skills to data science skills, from computational thinking to data literacy. And with the rapid advancement of data science tools such as artificial intelligence, machine learning and deep learning, the impact of data science on society and the workforce will only become greater over time (see Fig 1). The urgently needed response to this shift is *Data Science for All*.

Fig 1. Tractica, a market research firm forecasts the AI industry to grow 57-fold to 2025

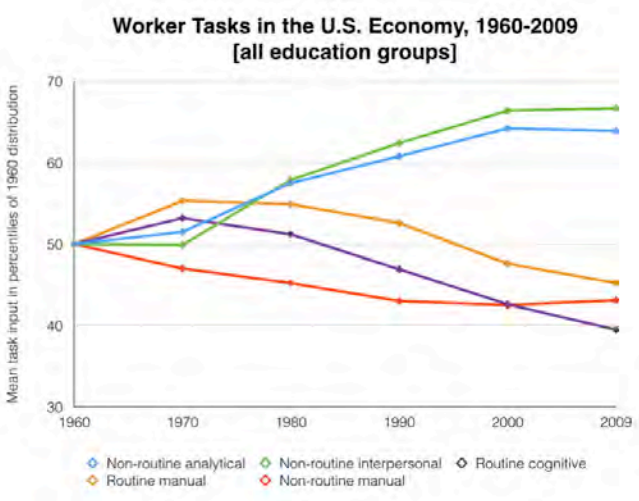

Fig 2. Economy-wide measures of routine and non-routine task input in the US as percentiles of mean task input from the 1960 task distribution up to 2009. Autor, Levy, & Murnane, 2010.

Increasingly, the revolution in data science is influencing decision-making in most sectors of society. Global business, banking, and entire economies are now much too complex not to use massive amounts of data for decision-making. Market transactions have changed from a purely human process to over a million trades per second. Business expertise is being replaced by smart predictive algorithms that, in many cases, make decisions through analyzing streams of data without human intervention. These approaches are also now the only way we can understand the scale and predict the trends and impact of the human footprint on the environment; deepen our understanding of disease, social and political impacts; and, ultimately, our understanding of the human brain. The effect of these trends is reflected in a transformation in the workplace throughout domains. The kinds of skills demanded by everything from startups to Fortune 500 companies to research collaboratives have

dramatically shifted away from a focus on individual compartmentalized skills in hierarchies to nimble, highly creative, and collaborative skills. The kinds of questions we ask, the degree of complexity of nature that can be analyzed at all scales, and the complexity of the problems policymakers, industry and citizens are called upon to solve are increasingly interdisciplinary and complex. (See Fig. 2)

The proliferation of data has transformed the way we think about nature and ourselves. We must acknowledge this revolution in data proliferation with a complementary revolution embodied in a new kind of literacy. To ensure the development of data literate professionals, workers, and citizens, the kinds of skills and modes of thought that the revolution in 21st century data-driven science demands must be incorporated throughout society through a significant cultural infusion.

We posit that to accelerate the pace of uptake in computer science requires a proliferation of critical thinking about problems worth solving through computation. The recent trend in infusing computational science into domains of STEM brings with it the need for a kind of data science literacy that is currently lagging in the computation in STEM revolution. Along with the need to use data in problem ideation and solutioning through facility with data science is the need to deal with the growing ethical and security issues emerging from the gathering and use of personal data for large scale commercial gain and the rapid rise in the use of artificial intelligence and machine learning to process, federate and make decisions relating to all this data. For lifelong learners to become increasingly data savvy, deal with issues of data ownership and security, and be discriminating consumers and creators of data, requires a kind of data literacy for everyone. We believe that the next logical step to provide the *WHY* and *WHAT* of computational thinking will be a kind of Data Science for All (DS4ALL).

The exponential growth in data from technology use - whether accessed via the web or via smartphones - means that the gathering and use of big data, machine learning and the results from data analysis are now ubiquitous. People routinely use search engines and online news feeds, GPS and transportation services (such as Uber or Lyft); social media; e-commerce, online shopping and recommender systems (travel and hotel bookings and crowd sourced rating systems, e.g., Yelp); fitness monitors, health apps and genomic assaying (e.g., 23&Me); imaging, sensors, virtual assistants or other similar services. Ubiquitous use extends to a myriad of other areas, including many aspects of the criminal justice system, the insurance and finance industries, education and more. Citizens in societies everywhere are impacted—whether they are aware of it or not. Yet how much does the general public actually know about or understand any of this?

With the use of data and results from data analysis being so ubiquitous, so personalized, and essentially unavoidable, there comes a pressing need to provide citizens with a basic awareness of the situation they themselves are in, and a deeper understanding of the effect of these technologies on their wellbeing—to gain a rudimentary understanding of what data are potentially being collected about each person, and how they might be being used. This includes the need for understanding notions of "responsible data science", including transparency, ethics, security, and accountability. Not understanding these concepts affects the ability for people to have the capacity to thrive, make career choices, and navigate the ethical, personal and community impact of the data-driven economy.

We must answer the revolution in the proliferation of data and data-driven tools with a complementary revolution embodied in a new kind of public literacy. While there is a rapidly growing call to address the gaps in data science skills' acquisition and training in formal education settings, workforce preparation and mid-career training, today's data-driven science and business demands the kinds of habits of mind that must be incorporated throughout society with a significant cultural infusion. Lifelong learners need to become increasingly data savvy, learn to deal with issues of data ownership, ethical use and security and be discriminating consumers and creators of data - and this requires data literacy.

DS4All will give lifelong learners the data literacy skills they need to navigate the digital economy--not just as consumers, but as active and engaged citizens in our technology-driven world. The goal is not for everyone to become a data scientist but for everyone to be data-literate and data-aware, allowing the public to make sound, evidence-based decisions as consumers, patients, workers, and family and community members. DS4All will provide a basic data literacy framework and resources for citizens of all ages.

DS4All is a bold new initiative to empower lifelong learners to learn data science and be equipped with the data literacy skills they need to be creators in the digital economy, not just consumers, and to be active citizens in our technology-driven world. Our economy is rapidly shifting, and both educators and business leaders are increasingly recognizing that data science (DS) is a "new basic" skill necessary for economic opportunity and social mobility.

DS4ALL engages in a wide range of activities such as:

- Developing data literacy essential concepts and core ideas;
- Partnering with libraries to develop resources and design capacity building efforts;
- Collaborating with community groups and nonprofits;
- Surveying industry partners to identify gaps in workforce skills and learning;
- Facilitating curriculum development and communities of practice in data science teaching and learning;
- Training teachers, expanding access to high-quality instructional materials;
- Involving governors, mayors and education leaders to build effective regional partnerships;
- Engaging CEOs, philanthropists, creative media, technology and learning professionals to deepen DS commitments;
- Creating opportunities in informal learning such as citizen science, hackathons, festivals, other experiential learning.

Projects

Building Capacity for Regional Collaboration in Closing the Big Data Divide

One of the grand challenges for data literacy communities of practice is bridging the gap between STEM practice and STEM learning for all. The fact that most of the important discoveries in contemporary science come from large-scale, data driven approaches indicate that skills and knowledge in these approaches must play a significant role in 21st century STEM learning across all settings. It also means addressing issues of equity. There is evidence to suggest that along with the "digital divide" (White House, 2015), is what might be called a "big data divide". Prosperity, innovation and security of individuals and communities increasingly depend on a big data literate society (UNESCO, 2013). But there must be a concerted effort to determine what it means to be a data literate citizen, information worker, researcher, or policymaker; to identify the quality of learning resources and programs intended to improve data literacy, and to chart a path forward that will bridge data practice with data learning, education and career readiness.

Through an NSF funded planning project - *Building Capacity for Regional Collaboration in Closing the Big Data Divide* (NSF CISE - 1636736), the New York Hall of Science (NYSCI) brought together experts from the Northeast Big Data Innovation Hub (NEBDIH) in a process of cooperative inquiry to: a) galvanize data practitioners around learning, b) identify the nature and quality of extant data literacy resources, and c) examine data literacy advancement strategies for learners. While the notion of cooperative inquiry has its roots in Action Research in the 1950s (Lewin, 1952), it is clearly differentiated in that it is a mutualistic phenomenological process among a group of investigators seeking to integrate action and observation. Based on work completed over many years by Heron and Reason (Reason & Rowan, 1981; Reason, 1988; Heron, 1996; Heron & Reason, 1997 and 2001), cooperative inquiry circumscribes a process in which multiple disciplines can focus on a problem specific to the group.

Subsequent work has engaged multiple sectors in developing and utilizing data literacy concepts, tools and skills, including work in formal education settings; with communities of need; and in STEM workforce preparation, as well as working in informal learning settings such as libraries.

The planning project was led and facilitated collaboratively by the New York Hall of Science (NYSCI) along with an inquiry group consisting of strategic members of the education and big data communities in the northeast, including members of the Education Connector from the Northeast Big Data Innovation Hub (NEBDIH). A small group of advisors from NEBDIH and education community served in an advisory capacity for the overall project. Planning activities throughout the year used synchronous and asynchronous collaboration tools, followed by a 2-day, in-person workshop at NYSCI including the IG along with invited stakeholders from industry, state/local government, K-20 curriculum developers, instructors, students, and specialists in learning sciences and informal learning.

The workshop included four elements: 1) an IG meeting to draft findings and prepare to report out to conference attendees; 2) a series of keynote talks from research, private industry and education on the current and future needs of the big data practice community; 3) a collaborative brainstorm- ing session to draft a fundamental set of big data literacy essential concepts; and 4) a rigorous evaluation to validate the process using community of practice measures, facilitate a blind peer review process for the resulting plan, and outline an evaluation plan future work.

Fig 3. Scenes from the 2017 Data Literacy Workshop

Building Capacity for Regional Collaboration in Closing the Big Data Divide also sought to advance the use of Community of Practice (CoP) measures to determine the feasibility and potential sustainability of a data literacy effort. The CoP model is used in education research to learn which group of individuals in a common enterprise share information and experiences and collaborate to strengthen their skills and knowledge. The framework developed by Wenger, Trayner, & de Laat (2011), identifies 5 value cycles:

- Immediate Value of the conceptualization activities themselves;
- Potential Value of the interactions in terms of skills, social capital, and resources;
- Applied Value of conceptualization process' effect on members' and how they generalize their knowledge;
- Realized Value of recognition of the value of different abilities and expertise toward both explicit and implicit goals of the group;
- Reframing Value of redefining success, trying new approaches, obsolescing old structures, and identifying new performance metrics to reflect these definitions.

Conclusions

The purpose of the project was two-fold: 1) To establish the feasibility of building a community of practice for cultivating a data literate society, and 2) to establish a baseline set of principles that help define what it means to be a data literate citizen, information worker, researcher, or policymaker. At the outset, attaining these goals seemed straightforward. However, through a collaborative inquiry process that involved 20 professionals representing a range of data science sectors who participated in a cross-disciplinary workshop on developing of a community of practice, it became evident that there are many different perspectives on what it means to be data literate. Some participants from the corporate sector saw greatest value in workforce preparation, while those in academia looked to what would comprise content in coursework toward a degree. The process also revealed that there is a range of different data literacy initiatives, most of which have been developed recently and do not appear to inform one another. What began as a process to distill a set of core ideas from broad discourse became a process of seeking alignment, common ground and language among disciplines, and required casting a much wider net to gather concepts and ideas related to data literacy. This in turn required bringing together communities of need, service organizations, libraries and others on the "front lines" of data use in order to exchange ideas, discuss approaches to improving the living conditions of underserved communities, and gain a better understanding of what it means to be a data literate person across income and cultural boundaries.

Outcomes

Framework

The outcome of this process is a draft set of principles that remain under discussion:

1. *Data must be defined* according to the context in which it exists.
2. *Data interpretation requires critical thinking* about questions, provenance, and purpose.
3. *Data can be useful* to answer questions and solve problems.
4. *Cultivating the ability to create and interpret visualizations* of data is essential.
5. *Data are neutral* but algorithms are not.
6. To be data literate means *having specific skills as well as the awareness* of how data are used, for what, and by whom.
7. *Data can be preserved*, stored and retrieved.
8. Data literacy means *understanding the importance of data quality* and usability.
9. Data literacy involves *understanding ethical, privacy and security issues* around data.

CoP

Early evidence of interaction related to each of the five CoP cycles was observed during the workshop (see Table 1). Observations of participants were promising and will assist the project team in developing indicators for each stage of the full CoP as the data initiative continues. It also provides early indicators that suggest CoP to be a useful framework to develop future collaboration and interpret impacts and be a useful model to identify indicators and markers for community growth.

Table1.Workshop Observations related to Community of Practice Cycles

CoP Cycle	Value	Related Observation
1	Immediate	• Participants are engaged in discussion, provide feedback on one another's ideas and challenge assumptions • Clarifying questions asked • New stakeholders involved in group process • Participants inform conversation with experience and learn from others • Efforts exist to document conversations
2	Potential	• Tools and documents produced to inform practice • Participants make professional connections with each other • Evidence of innovation observed (finding new ways of measuring data)
3	Applied	• Tools developed to address barriers to effective use of big data resources • Conversation overheard about the predictive value of big data
4	Realized	• Participants describe how tools and resources will benefit their organizations • Evidence of a new (learning) agenda being formed by group • Discourse about value is expanded (e.g., questioning what value means in the big data context)
5	Reframing	• Participants create new sets of expectations and strategic direction • Discussion about re-imagining social, institutional, legal, and political systems • Evidence of new metrics and assessment measures being considered, often through the involvement of new stakeholders

Cooperative Inquiry

The planning grant sought to: a) focus data scientists around learning, b) identify the nature and quality of extant data literacy resources, and c) to look at the kinds of strategies that could advance data literacy for lifelong learners. This process helped validate the need for and to identify an emerging community of practice around data literacy. Outcomes of the collaborative inquiry process revealed the following:

- Advancing data literacy is essential for society and there is a willingness among practitioners to envision a way forward;
- An inclusive, collaborative, multi-sector, multidisciplinary approach is needed to develop a bias-aware and unified framework and sustainable infrastructure that will allow for data literacy project implementation and oversight;
- While there is a need to specifically target the Northeast, there is clearly a global need for data literacy and scalability and generalizability are important goals; and
- The breadth of expertise and interest in equity, ethics, and in bridging sectors and educational opportunity areas represented by workshop participants can be brought to bear on circumscribing data literacy and developing a practical framework.

Throughout the duration of the Building Capacity for Regional Collaboration in Closing the Big Data Divide project, there has been spectacular growth in interest and involvement from a broad spectrum of commercial, non-profit and academic sectors. As a result we have changed the name of this initiative to Data Science for All. The 4 Big Data Hubs nationally are interested in this work and we now have the attention of a global community of interest.

Working with Communities

While the cooperative inquiry process well advanced our understanding of the kinds of skills and habits of mind needed from the perspective of the academic and research communities, closing the data divide means identifying ways that the public accesses data literacy knowledge and resources, and obtaining input from those that directly serve communities of need and deal with the concerns of those communities. NYSCI, along with Columbia University, New Knowledge Organization Ltd, and the American Library Association (ALA) received support through the Northeast Big Data Innovation Hub for additional workshops to further the development of a framework for data literacy and its dissemination. One workshop would target that national perspective on data literacy from libraries and the other from local libraries and community service organizations.

Libraries as Data Literacy Portals on the National Level

To attain the goal of addressing the data literacy needs of lifelong learners, and underserved populations in particular, the organizers felt that inclusion of the perspective of American Library Association (ALA) Public Programming leadership would inform the development of the data literacy framework, as 21st century libraries are now used as data repositories, data access points, workforce skills development centers, literacy centers, community centers and education centers. Libraries devote significant resources to workforce development and skill building, from resume planning and workshops to online data management skills and training.

The organizers brought together information specialists and members of ALA leadership in a one-day workshop at ALA headquarters in Chicago. The goal of the meeting was to articulate the needs, barriers, programs, resources and trajectory of data literacy from the perspective of public libraries. Key takeaways included:

- People come to the Library to solve problems;
- Developing education programs for line librarians in public libraries help individual patrons protect online privacy;
- Tensions exist among patrons around data security and privacy;
- Librarians need to be smart users of their own data and advocates for data literacy and security.;
- Majority of skill building is with youth but real growing engagement toward workforce and entrepreneurship provides opportunities for data literacy in public libraries;
- Larger libraries are leaders in the field of new programs and initiatives;
- There is a need for data competency workgroup for libraries at a national level;
- Libraries act as observers of library use and needs for patrons;
- Libraries' serve an activist role in communities of need;

Perhaps, most importantly, is a pervasive need for both ALA and library staff to receive training in data literacy in order to better serve the public.

Subsequent ongoing work with Census 2020 and New York and Queens Public Libraries has focused on identifying opportunities and existing programs for libraries for data literacy. A significant area of data literacy in libraries takes to form of preparation of library staff to train the

public in Census data in anticipation of deployment of the 2020 Census, and using the libraries' pivotal new role in the Census as an opportunity for data literacy enhancement. This work is expected to continue up to the 2020 Census deployment in April 2020. There is an ongoing concern about how the impact of a citizenship question on the census will marginalize communities of need, as well as affect the public's trust of libraries.

Role of Libraries and Community Service Organizations at the Local Level

This led to the hosting of a DS4All Community Workshop at NYSCI that brought together representatives of diverse urban community groups and nonprofits (including libraries), K-12 students and teachers, lifelong learners, and professionals. The workshop focused first on defining the needs within underserved communities and community service providers for increased awareness of the importance of data literacy, and how data science might be applied to address intractable problems wi thin these communities. The second objective was to collaboratively develop actionable plans for how to address these needs. The workshop included talks from participants about case studies from the health, cultural institution, library, and social services sectors, and breakout groups looking at specific issues and actionable solutions.

Fig 4. James Liu (NYSCI) presenting on Big Data for Little Kids family workshop project

Findings

The workshop was evaluated by NewKnowledge through observations and interviews. Building on the Fall 2018 convening in Chicago with the American Library Association, this workshop proposed novel solutions for some of our fundamental challenges related to the use of data to address social problems, including how to motivate communities to apply data science in ways that promote their own agency, and using data to close gaps between community needs and the services currently offered. Ultimately, three actionable ideas were generated by workshop participants: 1) Trainings to assist community organizations with easy-to-use and creative data collection and analysis; 2) A core competencies framework; and 3) A short video on responsible data science for wide distribution.

Fig 5. Elmcor Executive Director Saeeda Dunston and RISE Project Director Leyla Henriquez presenting at the DS4All Community Workshop

- The group of participants was extremely diverse, working in a variety of both non-profit and commercial settings that interface with the public in unique ways. This diversity lent itself well to developing proposed solutions that were useful and grounded in reality;
- In setting individual goals for the workshop, participants expressed a desire to learn more about one another's work and make professional connections, to better grasp the data "landscape," and to develop an understanding of how data can support services that address community needs;
- The lack of a definition and established criteria for data literacy continue to be obstacles to progress;
- Workshop case studies were a valuable way to share insight into current community-based initiatives.

Workshop participants were enthusiastic about the presentations and remained engaged throughout, asking critical questions of the presenters. Topics demonstrated a wide range of data science applications, including: best practices using infographics; the importance of context in determining data literacy; how to use data to measure the impact of educational services; the importance of aligning data with intended outcomes; variables that could potentially skew data analysis; how to make data-based activities meaningful and relevant; democratizing knowledge about artificial intelligence; and articulating community assets.

- According to case study presentations, having an audience or including an opportunity for those learning to use data to share what they accomplished was a strong motivating factor for participants;
- Certain challenges were raised and discussed among workshop participants: 1) How can you take individual, micro-measurements and determine macro community-wide impact?; 2) How do social / emotional factors influence the nature of data collected?; 3) Often much more data is collected than can be analyzed, and this is particularly difficult for textual data / qualitative analysis. How can organizations expand their analytic capacity or determine which datasets will yield the most valuable insight?; and 4) What scaffolding is needed to support the meaning-making process for intergenerational learners engaged in data-based activities?;
- In their small group work to develop action plans, workshop participants made a concerted effort to build on currently existing assets and resources, or to find parallels that could be applied to the field of data literacy. For example, what other competency frameworks could be instrumental or provide guidance?;
- Proposed action plans were both detailed and feasible, and each outlined specific next steps. For example, the Data Collection & Use group suggested training in the following skill development areas: 1) Expanding the use of data collection tools beyond the "usual suspects" (e.g., surveys) to include more crowdsourced and participatory methods; 2) Free and available tools that already exist; 3) How organizations can rethink what data is collected and why (aiming at addressing gaps in knowledge); 4) How to form questions appropriately, taking context into consideration; and 5) Developing ethical norms for data collection that protects participants.

Other Related Activities

Elmcor

There is ongoing work with Elmcor (Corona/Elmhurst CBO) on enhancing their internal staff data literacy needs, and similar efforts with the NY State Office of Alcohol and Substance Abuse Services. In addition we are mentoring Columbia University Social Entrepreneurship Group (CSEG): undergraduate students to work closely with CBOs on their data literacy needs. The partnership with Elmcor grew out of a workshop we organized and facilitated at the Bloomberg Philanthropies event Data for Good Exchange (D4GX) in September 2018 (see list of related activities). During this workshop, Elmcor staff presented their needs: to raise the level of data literacy across their staff, as well as throughout the non-profit. As a result, volunteers began working with Elmcor on applying data analysis tools to survey results in order to be able to show project impact.

Fig 6. D4GX workshop participants mapping out solutions to Elmcor's data analysis needs

Big Data for Little Kids

NYSCI obtained NSF funding (Award #1614663) to develop and test a suite of activities to engage children ages 5-8 and their caregivers in age-appropriate explorations of three foundational concepts of data analysis: categorization, variability, and distribution. Participants also engaged in the social and communicative practices of data analysis: shared problem definition, sense-making about numbers and patterns, and communicating about insights with peers. The goals of the project were to:

Fig 7Big Data for Little Kids participants developing their project

- Establish a permanent presence for six to eight data analysis activities on the floor of the museum;
- Integrate extended versions of these activities into regular group programming offered by the museum to family groups;
- Disseminate the activities both nationally and locally through our existing networks of early childhood museum educators;
- Engage professionals who work with big data with the unique challenges and opportunities associated with teaching young children about these issues;

- Measure the impact of these activities on children's ability to pose relevant, data-driven approaches to answering questions about their environment and to identify possible sources of data to answer those questions;
- Measure the impact of these activities on children's workplace-relevant social skills, such as their ability to communicate about the relevant features of their data analysis activities to their peers and to adults.

Final research results are being analyzed and a final report is forthcoming.

Designing the Successful Inclusion of Data Science in High School Computer Science (Pending conference grant)

NYSCI, in collaboration with the Columbia University Data Science Institute/Northeast Big Data Innovation Hub and University of California San Diego propose to bring learning experts, data science practitioners, domain experts, and tool developers together with education and private sector stakeholders will host a workshop to review existing tools, datasets and teaching resources; brainstorm the development of improvements; and propose a set of priority implementation strategies intended to address the data science-computer science education gap. While the need exists throughout PreK-20 education, our workshop effort will focus primarily on high school students and teachers, who are on the front lines of the rapidly changing workforce. The overarching goal for this capacity-building workshop will be to articulate a pathway toward *data literacy* and emphasize the unique and genuinely new dimensions of learning afforded by data science in applying computational thinking, computer programming and habits of mind to new problems.

Rationale

There has been intense interest in bringing computer science to all learners through such initiatives as Computer Science for All (Heintz, et al, 2016). Increasingly, complex real world problems are being solved through the application of computational skills and data science, as evidenced by the dramatic growth in data science over the past decade. A 2011 McKinsey Report (Manyika, et al.), which was released at a tipping point in the growth of data and data science, indicated that the gap between what is being taught in school systems and the needs of both the workplace and society, writ large, is growing, resulting in a widespread shortage of skilled workers and the inability for society to benefit from the new tools of data science. There have been calls for improved data literacy across disciplines within tertiary education, resulting in the development of new academic programs (see fig.). But there has not been an equivalent growth in K-12 education programs, much-needed to prepare students for this new reality (National Research Council, 2006; Kastens & Krumhansl, 2013; Zalles, 2014).

A variety of tools and evidence-based curricula have been developed to address the lack of data literacy through the use of statistical modeling and inference to bring data skills to learners. (Vahey, et al, 2010; Ridsdale, et al 2015; Krumhansl, et al, 2014; Konold, et al. 2017; Finzer, et al. 2018). It has even been suggested that the Algebra-Calculus track be replaced by a rigorous program of statistics and data literacy (Rubin, 2005). But while there have been attempts to introduce data science skills into computer science curricula, these attempts have lacked access to the knowledge, resources, training and accessible tools and data sources needed to integrate

and scale data literacy into computer science literacy and have not resulted in significant and scalable impacts in formal education. There is a clear need for an accounting of the effectiveness of extant data science teaching and learning resources, tools, and best practices; identification of gaps in the capacity for testability and utility of resources, data and tools to leverage against data literacy needs; and a strategy to develop a research agenda and culture of lifelong learning that recognizes and infuses data literacy into all aspects of learning in formal settings.

This workshop would take place in Fall or Winter 2019-2020.

Additional Activities

Books

National Academies of Sciences, Engineering, and Medicine 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.

Talks/presentations/events

- June 2019: Oregon State University: International Society for the Systems Sciences Annual Conference: Panel: Developing frameworks: Data Science and Network Science.
- June 2019: UCSD Center for US-Mexican Studies, School of Global Policy & Strategy; San Diego/Tijuana Workshop. Armchair Conversation: Developing a socially engaged cross border STEAM talent Pipeline.
- May 2019: Bloomberg LLC. Quant Seminar: Data Science For All
- April 2019: SUNY Binghamton Northeast Regional Conference on Complex Systems; Data Science for All: Empowering Communities to Solve Wicked Problems; Invited Springer Talk.
- July 2018; Massachusetts Institute of Technology International Conference on Complex Systems: Data Science for All: Situating Data Literacy Across Learning Settings.
- June 2018; The Wilson Center, Woodrow Wilson International Center For Scholars Science and Technology Innovation Program: 2nd Annual Transatlantic Symposium on ICT and Policy, Panel on partnerships on Big Data Research & Innovation and Workforce Development. https://www.wilsoncenter.org/event/2nd-annual-transatlantic-symposium-ict-and-policy (see Part 5 of video stream);
- June 2018 Institut des Systèmes Complexes de Paris, International Workshop & Conference on Network Science: Seventh Network Science and Education Symposium Data Science Education from a Network Science Perspective. Notes available online here: https://osf.io/7v9xt/;
- April 2018 Oregon State University Center for Research on Lifelong STEM Learning: Talk on Advancing Data Literacy;
- January 2018: Panel member; Keeping Data Science Broad: Big Picture for a Big Data Science Education Network, Webinar with South Big Data Hub; Keeping Data Science Broad: Negotiating the Digital & Data Divide: Webinar for Alternative Avenues for Development of Data Science Education Capacity (Building Capacity for Regional Collaboration in Closing the Big Data Divide) https://www.youtube.com/watch?v=59wjGpyRr-U&t=0s&index=3&list=PLyUNw5pgUji-Vv4zB2NFJDVrg_rwMTIHT;
- September 2018: Bloomberg D4GX workshop: Addressing Community Challenges with Data-Driven Solutions, Workshop Organizer Panel, and DS4All poster: Data Science for All: Democratizing data for a global citizenry;
- February 2017: Data Science Education Technology conference; Concord Consortium;
- November 2017: Trans-Atlantic Workshop on Public/Private Partnerships for Big Data Research & Innovation and Workforce Development; Versailles, France; Big Data Value Association conference, in collaboration with South Big Data Hub; Details are available

here: https://www.eventbrite.com/e/trans-atlantic-workshop-on-public-private-partnerships-for-big-data-registration-38786953823

- November 2017: South Hub Workshop, Keeping Data Science Broad
- October 2017: Talk on Data Modeling with Young Learners and their Families in Panel Visualizing STEAM Data in Support of Smart Decision Making for the Science Centre World Summit, Tokyo, Japan. https://scws2017.org/_assets/docs/presentations_slides/171116/16_para_stephen_miles_uzzo.pdf
- October 2017: National Academies Roundtable; Organized Panel on Informal Data Science Education for National Academies of Sciences, Engineering and Medicine, Roundtable on Data Science Postsecondary Education Meeting #4: Alternative Institutional and Educational Mechanisms. Recording of Panel discussion is available here: https://vimeo.com/album/4865664/video/243734798
- September 2017 Panel: Creating Actionable Data-Driven Knowledge in Communities of Need at the Bloomberg Philanthropies Data For Good Exchange 2017 (September): With Great Data comes Great Responsibility video stream of panel discussion Here: https://www.youtube.com/watch?v=vU74kmfoZRI&index=98&t=0s&list=UUfkof0bu-9c3JDYo46Z8NIA
- September 2017: South Hub Data Science Webinar (https://southbigdatahub.org/programs/keeping-data-science-broad/), (see details on this series here: http://sites.nationalacademies.org/cs/groups/depssite/documents/webpage/deps_189133.pdf) and authored final report (see below)
- March 2017: Workshop @ NYSCI
- November 2016; Data Science Education Technology conference; Concord Consortium
- August 2016: Youth, Learning, and Data Science Summit 2016; UC Berkeley
- Inaugural Workshop for Data Literacy, Columbia University
- April 2015, NSF BD-HUB Charrette for Accelerating the Big Data Innovation Ecosystem, Boston
- March 2015: Big Data Fest @ NYSCI

Ongoing work with the Northeast Big Data Innovation Hub

Website (recently updated) http://nebigdatahub.org/ds4all/
New work: creating online Data Literacy resource collection

Partners

- Northeast Big Data Innovation Hub;
- West Big Data Innovation Hub;
- South Big Data Innovation Hub;
- SUNY Binghamton;
- SUNY Albany;
- Cornell Tech;
- NYU CUSP;
- UCSD Supercomputing Center, Halıcıoğlu Data Science Institute;
- Columbia Data Science Institute;
- Columbia Social Entrepreneurship Group;
- Queens Public Library;
- New School Digital Equity Lab;
- American Library Association;
- New Knowledge Inc.;
- Bloomberg Philanthropies.

Next Steps

DS4All will continue to work on developing a data literacy competency framework. A needed next step is to have an agreed upon definition of terms. The Library Leadership and Management Association's (2012) definition of competency: "Professional competencies comprise the knowledge, skills, and abilities which are teachable, measurable, and objective and which define and contribute to performance in librarianship."1 The question of measurement helps us differentiate between a competency and a skill. A competency has two dimensions: 1) The knowledge, skill, or ability; and 2) The level of mastery of that knowledge, skill, or ability.

DS4All will use findings from work to date to develop a white paper describing gaps and suggesting next steps for the field. White paper sections could include an overview of past efforts to gather input from the field to establish context, outstanding research still needed to address knowledge gaps, ideas about navigating persistent challenges, and a proposed roadmap for future work. Ultimately, the white paper should be written as a call to action – a compelling and deliberate vision of where the field needs to go.

DS4All will use the white paper to engage the broader community and get feedback on common language and concise ideas about what data literacy means. We will make a clear case for the work that is needed and use the white paper to secure subsequent funding for implementation of the action plan. We will consider how we might approach a research and development process for a competency framework for data literacy. We will work to develop international partnerships, building on existing relationships with partners in the EU, the UK, and Mexico. Work will continue to address the needs of underserved communities, as well as formal education strategies.

References

Ainsworth, S. & Loizou, A. (2003). Cognitive Science, Vol. 27. Austin: Cognitive Science Society. 669

Fitzallen, N., & Watson, J. (2010). Developing statistical reasoning facilitated by TinkerPlots. Refereed paper to be presented at the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July, 2010. [CDRom] Voorburg, The Netherlands: International Statistical Institute.

Heron, J. (1996) *Co-operative Inquiry: Research into the Human Condition,*London, Sage Publications, 1996

Heron, J. & Reason, P. (2001). The practice of co-operative inquiry: Research 'with' rather than 'on' people. In P. Reason & H. Bradbury (Eds.), Handbook of action research. London: Sage Publications.

Heron, J., & Reason, P. (1997). A Participatory inquiry paradigm. Qualitative Inquiry, Vol. 3, No. 3. 274.

Kalil, T. and Jahanian, F. (2013). Computer Science is for Everyone! The White House Blog. https://obamawhitehouse.archives.gov/blog/2013/12/11/computer-science-everyone.

Kastens, K. & Krumhansl. R. (2013) EarthCube Education End-User Workshop

Kastens, K., Straccia, F., Shipley, T. & Boone, A (2013) What do Geoscience Novices & Experts Look at and What do They See when Viewing and Interpreting Data Visualizations? Gordon Research Conference.

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.

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.

Lewin, K. (1952) Group Decision and Social Change, in G. W. Swanson, T. M. Newcomb & E. L. Hartley (Eds) *Readings in Social Psychology*. New York: Henry Heath & Co. Reprinted in S. Kemmis & R. McTaggart (1988) *The Action Research Reader,* 3rd Edn. Geelong: Deakin University Press.

Library Leadership and Management Association (LLAMA). (2012). LLAMA Library Leadership and Management Competencies Task Force Final Report. Community Workshop: Findings & Recommendations, NewKnowledge Publication #NSF.061.508.01 3

Library Leadership and Management Association (LLAMA). (2012). LLAMA Library Leadership and Management Competencies Task Force Final Report. Community Workshop: Findings & Recommendations, NewKnowledge Publication #NSF.061.508.01 3

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.

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.

Reason, P. (Ed.). (1988). *Human Inquiry in Action: Developments in new paradigm research*. London : Sage Publications.

Reason, P., & Rowan, J. (Eds.). (1981). *Human Inquiry: A Sourcebook of New Paradigm Research*. Chichester, UK: Wiley.

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.

Rubin, A (2005) Math that Matters. Threshold, Spring 2005

The White House (2013). ConnectEd Initiative https://obamawhitehouse.archives.gov/issues/education/k-12/connected

The White House (2015). The TechHire Initiative https://obamawhitehouse.archives.gov/issues/technology/techhire

The White House (2016). Fact Sheet: President Obama Announces Computer Science For All Initiative. https://obamawhitehouse.archives.gov/the-press-office/2016/01/30/fact-sheet-president-obama-announces-computer-science-all-initiative-0

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.

Vahey, P., Rafanan, K., Swan, K., van 't Hooft, M., Kratcoski, A., Stanford, T., and Patton, C. (2010). Thinking with Data: A Cross-Disciplinary Approach to Teaching Data Literacy and Proportionality. Presented at the Annual Conference of the American Educational Research Association, May 2010, Denver, CO.

Wenger, E., Trayner, B., & De Laat, M. (2011). Promoting and assessing value creation in communities and networks: A conceptual framework. The Open Universiteit Nederland.

Zalles, D. (2005). *Designs for assessing foundational data literacy.* Available online at On the Cutting Edge: Professional Development for Geoscience Faculty Web site: http://serc.carleton.edu/files/NAGTWorkshops/assess/ZallesEssay3.pdf

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