Data Science for All: Democratizing data for a global citizenry

Catherine Cramer[1] and Stephen Miles Uzzo[2]
[1] Columbia University, Data Science Institute / [2] New York Hall of Science

Introduction

Here we discuss an initiative to democratize data through developing a framework for data literacy for all people. While there are many successful small scale programs and research projects to better understand what people need to know about data science, there is little evidence that these efforts are scalable to result in widespread data literacy. What is needed is to develop and implement approaches that can affect policy and create equity in data literacy access throughout all levels of society.

History

Beginning in 2013, the Federal government began a multi-pronged, widespread and highly successful 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 (Kalil and Jahanian, 2013). According to White House figures, through comprehensive dissemination and uptake provided by initiatives such as ConnectED (The White House, 2013) and TechHire (The White House, 2015), the connectivity divide in US public schools has been cut by about half since 2013; hundreds of employers have partnered with agencies and organizations from cities, states, and rural areas to expand access to tech jobs; and thousands of students have access to computers for the first time. This successful implementation plan was built on longstanding efforts at the local, state and federal levels to raise the level of science, technology, engineering and mathematics (STEM) career access, and was narrowly focused on one specific area: computing skills (The White House, 2016).

In January 2016, the US Federal Government announced Computer Science for All (CSforAll, Fig 1.) proposing $4 billion in funding to states and $100 million directly to school districts in order to increase access to K-12 computer science by training teachers, expanding access to high-quality instructional materials, and building effective regional partnerships (ibid). Aimed primarily at K-12 schools, CSforAll relatively quickly gained momentum through a comprehensive and strategic plan utilizing a wide range of resources and networks such as:

Fig 1. President Obama kicks off Computer Science Education week by speaking with students participating in an "Hour of Code" event at the White House.

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

What's Missing?

However, while the stated goal of CSforAll is to give US citizens the ability to solve complex problems through the application of computational skills, providing them with access to computers, software, programming skills and broadband connectivity was just the beginning. What's missing is the application to real world problems – something only done through data. Data provide the evidence and purpose of much of computing. And the need to be able to acquire, analyze and draw conclusions from data is emerging in every sector – from scientific discovery to civil rights, from industry to urban planning, from public health to disaster recovery – resulting in a marked shift of emphasis in workforce and educational needs in just the last few years, from computer science skills to data science skills, from computational thinking to data literacy.

The prosperity, innovation and security of individuals and communities increasingly depend on a data literate society (UNESCO, 2013). Given the flood of data that is now being generated and collected in almost all domains, it is essential that a concerted effort be developed to address the need for industry, policymakers, educators and diverse populations of citizens to be data literate. 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 sophisticated data mining, machine learning and analytics on massive amounts of data for decision-making. Market transactions have changed from a purely human process to over a million trades per second. Business acumen is being replaced by smart predictive algorithms that, in many cases, make decisions through analyzing streams of data without human intervention. These approaches are rapidly becoming the only way to understand the scale and predict the trends and impact of the human footprint on the planet; deepen our understanding of disease, social and political impacts; and, ultimately, elucidate understanding of the human brain.

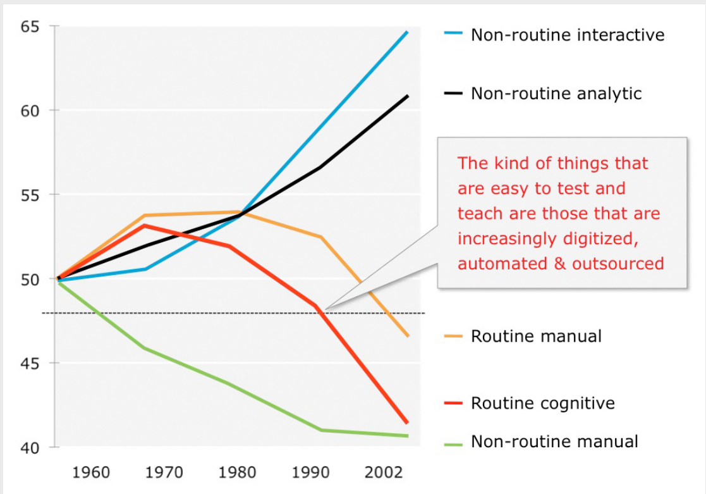
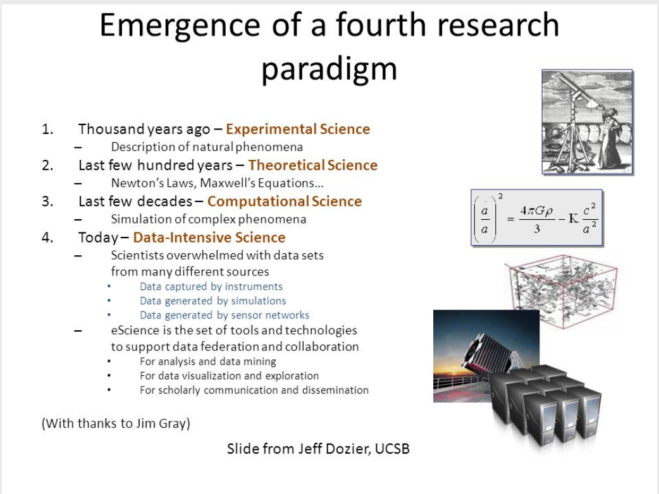


Fig 2. Economy-wide measures of routine and non-routine task input (US). Mean task input as percentiles of the 1960 task distribution. (Autor et al, 2003).



Emergence of a fourth research paradigm

1. Thousand years ago – Experimental Science
   – Description of natural phenomena
2. Last few hundred years – Theoretical Science
   – Newton's Laws, Maxwell's Equations...
3. Last few decades – Computational Science
   – Simulation of complex phenomena
4. Today – Data-Intensive Science
   – Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks
   – eScience is the set of tools and technologies to support data federation and collaboration
     - For analysis and data mining
     - For data visualization and exploration
     - For scholarly communication and dissemination

(With thanks to Jim Gray)

Slide from Jeff Dozier, UCSB

$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$

Fig 3. The emergence of data-driven science in the 21st century (after Jim Gray and Tony Hey, Microsoft Research).

The effect of these trends is reflected in a transformation of 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 to nimble, highly creative and collaborative skills (Fig 2.). 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 that policymakers, industry representatives and the general public are called upon to solve are increasingly interdisciplinary, complex and data driven. We must acknowledge this revolution in STEM with a complementary revolution embodied in a new kind of literacy.

It has been noted that there is a lack of even basic data literacy reasoning skills, and there have been calls for improved data literacy across disciplines for over a decade (National Research Council, 2006; Kastens & Krumhansl, 2013; Zalles, 2014). During that period, evidence-based curricula have been proposed to address a lack of basic data literacy (Vahey, et al, 2010; Zalles & Vahey, 2005; Zalles, 2005; Ridsdale, et al, 2015). Core to understanding and advancing data analytic skills is the use of statistical inference, and advanced statistical reasoning for sense-making of large-scale and multivariate data. For instance, Makar & Rubin (2018) propose that a learning progression of inferential reasoning would be possible. There is also a need for new tools to help provide opportunities for learners to explore data as the sciences advance. Tinkerplots is among a small number of tools that have been widely researched in understanding how learners cope with large and multivariate data sets (Fitzallen & Watson, 2010, Konold, 2007, Konold, et al, 2007). Scientific visualization can also provide opportunities to learn and communicate complex statistical and scientific data (Kastens, et al, 2013; Borner, et al, 2015; Ainsworth & Loizou, 2003). But none of these efforts have led to a large-scale framework for learning, nor have they scaled into widespread use across learning settings.

DS4ALL

As part of developing approaches for educating the next generation of big data literate professionals, workers, and the general public, the skills and habits of mind that the revolution in 21st century data-driven science demands must be pervasive in society. We propose that to accelerate the pace of uptake in both data and computer science will require a movement in critical thinking about problems worth solving through data science literacy. We believe that the next logical step to provide the WHY and WHAT of computer science is Data Science for All (DS4All).

DS4All is a bold new initiative to empower lifelong learners to be equipped with the data literacy skills needed to be not just consumers but active creators in our technology-driven world. Our economy is rapidly shifting, and the pace of innovation is at risk if data literacy is not pursued at all levels. We need to be able to educate and train students, potential new employees, current employees, managers, and senior leadership to broaden and diversify the data science pipeline, and to engage community members so that they can access and leverage data to solve the problems most urgent and meaningful to them.

DS4All builds on data literacy efforts that have been ongoing across sectors and are working to develop a data literacy framework as well as to discern, develop and disseminate resources pertinent to the development of related skills. There is an urgent and growing need to identify and deploy promising approaches to developing management skills and habits of mind that go beyond basic data science and statistical literacy. This is recognition of the advanced kinds of capacities that working with data demands, such as: emphasis on design, construction, and visualization of large-scale data, as well as federation of multivariate data streams through computational analysis tools; enhanced exploratory, interdisciplinary and pattern-seeking approaches; machine learning and modeling; the role of machine learning in the various stages from data creation to knowledge creation; and leveraging of cyberinfrastructures to make large data structures more available, interoperable, transparent, and robust.

Fig 4. New York Hall of Science/Northeast Big Data Innovation Hub workshop on big data literacy, 2015.

The goal of DS4All is to tackle the lack of data literacy from both ends—addressing K-20 education and workforce needs, along with lifelong learning sectors, by developing and implementing data literacy concepts, curricular materials and training modules to improve the workforce development pipeline through scaffolding data literacy concepts with community, technical and workforce training—anywhere people use data to function in society and in the workplace. Both critical thinking and computational thinking are necessary to be data literate. These modes of thought provide 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, properties or characteristics; relationships among the entities; and various perspectives, solutions, and/or extensions of the problem.

DS4All is structured analogous to CS4All and will engage in a wide range of activities including:

- Developing data literacy essential concepts and core ideas as a stepping stone to integrating data science into STEM curricula.;
- Partnering with libraries and other informal learning settings to develop resources and design capacity building efforts for lifelong learners;
- Developing critical thinking and computational thinking curricula, combining data literacy with a course in propositional logic and critical thinking;
- Fostering the habits of mind, and ability to map real world problems to appropriate data models prior to capturing the data in order to be able to ask questions that can be answered through the use of data analysis;
- Coupling data literacy with a discussion of ethics using case studies;
- Offering secondary school teacher training that builds on elementary school work;
- Strengthening application-based problem-solving and applied math skills;
- Collaborating with community groups and nonprofits in participatory design to bridge the gap between communities of need and data driven solutions;
- 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 and expanding access to high-quality instructional materials;
- Involving governors, mayors and education leaders to build effective regional partnerships, and policies;
- Engaging CEOs, philanthropists, creative media, technology and learning professionals to deepen DS commitments; and
- Creating opportunities in informal learning such as citizen science, data hackathons, festivals, other experiential learning in the data sciences.

Conclusions and Next Steps

Innovation and the spread of data sciences and applications is rapidly outpacing the general public's understanding of data in science and business. It is also ill-preparing the workforce for jobs of the future. Addressing this widening gap necessitates rethinking how we approach data science skills and educating the public about data science. DS4All seeks to promote the importance of knowledge in a data-driven society, benefit from the best in research in learning and applications, and identify and devise ways to bridge the gaps to educate, enlighten and empower the public with the value of data science and its effect on their lives.

References

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

Autor, D., Levy, F., & Murnane, R. (2003) The Skill Content of Recent Technological Change: An Empirical Exploration. Quarterly Journal of Economics Vol. 118, No. 4. Cambridge, MA: MIT Press. 1279.

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

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. Accessed 7/7/18.

Kastens, K. & Krumhansl. R. (2013). EarthCube Education End-User Workshop. March, 2013, UCSD Scripps Institution of Oceanography, La Jolla, CA. Northfield, MN: National Association of Geoscience Teachers.

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? Poster presented at the Gordon Conference on Visualization in Science and Education, July 22, 2013.

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.

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.

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.

The White House (2013). ConnectEd Initiative. https://obamawhitehouse.archives.gov/issues/education/k-12/connected. Accessed 7/7/18.

The White House (2015). The TechHire Initiative. https://obamawhitehouse.archives.gov/issues/technology/techhire. Accessed 7/7/18.

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. Accessed 7/7/18.

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.

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.

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. Accessed 7/7/18.

Funding provided by the NSF award 1636736. 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.

New York Hall of Science

NORTHEAST BIG DATA INNOVATION HUB

COLUMBIA UNIVERSITY Data Science Institute