CIRCL Primer: Data Science Education

Contributors: Phil Vahey, William Finzer, Louise Yarnall, Patti Schank
Questions, or want to add to this topic or to a new topic? Contact CIRCL.

Overview

Data Science is an interdisciplinary field that seeks to derive insights and knowledge from the analysis of typically very large data sets. While data science education is relatively new, there are currently many undergraduate and graduate degree programs available in data science. This primer is an overview of the early state of data science education in grades K-12.

New technology has made it easier than ever to capture, store, and arrange many forms of data about the world. Low-cost sensors capture and store scientific data from various environments. Optical character recognition (OCR) technology converts volumes of texts into data for analysis. Image recognition technology permits rapid search of photographic and graphic databases. Portable audio and video recording devices now collect many types of human interactions in different situations and settings.

Data science is the field that attempts to build knowledge from this newly available massive data store. While there is no consensus definition of data science, there is widespread agreement that data science goes beyond the application of traditional disciplinary or statistical methods. Drew Conway's Data Science Venn Diagram describes data science as the partial union of content expertise, math and statistics knowledge, and hacking (or computer science) skills. Some characteristics are:

- The investigator is "awash in data" (the dataset may at first be too overwhelming for there to be a clear path to analysis)
- The analysis requires "data moves" that go beyond application of known procedures (for instance, one may have to create a completely new visualization)
- The data are "unruly", meaning that a single observation may have many pieces of information
- The data are not typically easily stored in traditional data table format

While traditional statistical tools are central to data science, investigators may use more exploratory techniques such as machine learning or visualization to find patterns in data. Data science education introduces students to the tools, dispositions, and techniques:

- Running experiments and collecting data, typically in science class
- Conducting exploratory data analysis of one's own data or of others' data using different visualization tools
- Statistics, typically in mathematics class

At its core, data science requires students to engage in cross disciplinary thinking. While it is

unrealistic to expect K-12 students to engage in all aspects of data science, especially in the elementary and middle grades, educators are beginning to understand how we can incorporate appropriate tools and techniques for each grade level, and create and manage engaging data science classroom activities.

Key Lessons

Integrating Data Science into the Curriculum and Informal Settings. Opportunities to engage with data in elementary and secondary grades most often take place in science courses because data analysis provides a hands-on way to see science concepts in action.
While there is concern that teachers may not have time to integrate the messier and more time-intensive aspects of data science into their already full curriculum, research has identified some promising strategies. These might be broadly characterized as activities that aggregate data across groups of students and focal activities that require students to confront specific data problems. As an example of an aggregated data activity, if a teacher has many students conduct many trials of an experiment (say with a ramp or a spring) — each time modifying more than one variable at a time — the class can then explore the resulting pooled dataset to identify relationships and patterns. As an example of a focal activity, a teacher can ask students to review existing data in two contrasting aspects of the environment, say rainfall and topographic altitude, and ask them to infer the underlying relations. In these ways, students can engage with data science while reinforcing their understanding of core scientific concepts.

Data exploration activities also can be integrated into informal learning settings, such as museums or at camps. For example, some citizen science projects ask participants to collect data that is stored in large public data sets, such as timing of migration, seasonal plant growth measurements, or the quality of air or water. However, the current challenge for informal education environments is to move beyond data collection activities and find ways that informal learners can search for patterns in large data sets.

Focusing on Particular Aspects of Data Science. Progress has been made in providing curricular tools to support data exploration. For examples, the Maine Data Literacy Project has created a framework to aid students in determining what type of analysis or visualization they should use. This framework has students consider the types of data they have and their analytic goals such as investigating variability, comparing groups, exploring correlations or relationships, investigating change over time, and investigating behaviors or characteristics of subgroups. Researchers have found that science teachers consider these goals appropriate and valuable, and that the framework can support students in thinking about data sets that might otherwise be overwhelming.

The Role of Technology. Advances in technology have brought about the data revolution, and technology has a key role in data science education activities. Though students can work with very small data sets or interpret data illustrations without technology, they cannot engage in the practice of data science without technology. Software designed for learning, with a low threshold for getting started, can get students excited about working with data across a wide variety of subjects.

Spreadsheets and simple databases comprise the most common forms of technology used to work with data sets, but while they may work well in a business setting, they do not necessarily do well at encouraging data exploration. For instance, Excel provides a set of standard graphs, but it does not provide students with the ability to fluidly explore relationships. Statistical analysis packages are typically designed for practitioners rather than learners, and may have a powerful, but complex, set of features that hamper rather than encourage exploration. However, there are technology tools designed explicitly for data science education that do a better job of encourage playing with data. For example, TinkerPlots allows students to manipulate data visualizations with an intuitive toolkit that emphasizes simple methods of dragging data points into meaningful plots. Newer efforts include Tuva and CODAP, which allow students to dynamically explore relationships and build visual representations. At present, Tuva tends toward single curated data sets, flat tables, and a single graph at a time. In contrast, CODAP allows for multiple linked representations, map data, and graphs with three or more variables. CODAP is unique in allowing learners to create and modify hierarchical structures, an important part of experiencing the doing of data science.

Issues

Serious thinking about how to integrate the new field of data science into K-12 education has barely begun. Much exploration and research are needed. Five issues of note are: framing data science education; identifying the barriers to integrating data science into educational contexts; establishing consensus around the goals of data science education; finding a path that both educates and protects students with regard to data ethics and privacy; and determining what research in data science education is needed.

Framing Data Science Education. Who will teach data science? Should data science become a subject in the school curriculum like calculus or American history, or should working with data be a part of every school subject? The path of least resistance is carve out a new discipline with its own courses likely not encountered until the high school level, and then by only some proportion of students. Most teachers would not have to worry about fitting data science concepts and skills into increasingly crowded content areas. A small number of well-trained subject matter specialists could teach the few, likely elective, courses. This path is doable immediately and, in fact, has been started in projects like Mobilize, a collaboration between Los Angeles Unified and UCLA.

The problem with carving out a separate discipline for data science at the school level is that it betrays the inherent interdisciplinary spirit of data science. Data are everywhere. Nearly all areas of work require familiarity with data. While relatively few of today's students will end up with the job title of data scientist, all of them will need to understand how to use data productively as workers and as citizens. This line of thinking leads to an arduous path at the end of which every teacher is integrating data science into whatever they teach and, mirroring the world outside school, students take for granted that data are part of all learning. Going down this path will require change at every level and in every subject, plus an unprecedented level of interdisciplinary coordination.

Identifying Barriers to Data Science Education. To achieve integration of data science into multiple subjects, the pervasive stovepiping of subject disciplines poses particular challenges.

Currently, from grades 6-12 (the grades at which, one can argue, students are most ready to engage in data science education), there is almost no collaboration across disciplines. This stovepiping permeates the standards developed for different disciplines, such as science and mathematics. It is challenging to change the standards or find ways for teachers to effectively collaborate across disciplines. Data science is not the only new field that challenges the stovepipe model; so does the field of computational thinking, which seeks to engage students in activities that employ principles of computer science in diverse courses and disciplines.

The lack of teacher knowledge and the limited time available for professional development pose additional challenges. The practice of data science goes beyond knowledge and requires the *experience* of actually working with and using data. Teachers can learn about a new field, which makes it possible for them to teach about it, but we want students (and their teachers) to experience data science by doing it. Data science requires a deep understanding of both disciplinary content and methods, and requires dispositions that run counter to common, efficiency-oriented teaching methods, such as a willingness to engage in ill-defined problems, follow paths that are ultimately not productive, and create new visual representations. It will be a significant effort to get math and science teachers comfortable with these new requirements. This concern is compounded for non-STEM teachers and elementary school teachers, who are typically less accustomed to the quantitative thinking that often accompanies working with data.

Establishing the Goals of Data Science Education. There is still not consensus on the goals of data science education. While producing more data scientists is important, building data fluency in all walks of life is a very important goal. Nearly all people will routinely be working with data and require some skills that fall into what is now called data science. Some specific goals include:

- Identify paradigmatic learning activities that exemplify what we mean by experiencing data science at different grade levels.
- Describe exemplary uses of technology in data science education.
- Formulate performance criteria for data science education.

In the broadest sense, the goal of data science education is to figure out how best to bring about effective learning with and about data. The field may move forward with that as the driver, and more specific goals will emerge.

Addressing Issues around Data Ethics and Privacy. Data science education should sensitize students to the potential impacts of data collection and analysis on groups, individuals, and entities. As people surf the web and interact with their devices, they leave evidence (a data “exhaust”) that can be used to identify them and access their experiences and activities. The “quantified self” movement associated with wearable devices that gather data about health and activity presents obvious risks to privacy. Education researchers are only in the beginning to confront these issues and, as technology advances, they are likely to increase in importance.

Determining What Research is Needed in Data Science Education. A set of cutting-edge research agendas and processes needs to be defined for foundational data science education research to make impact. For example: What is the role of visualizations in data science education? How do students interpret visualizations of complex data? How can we help them create their own

visualizations? How do learners conceive of data and learn to use data structures appropriate to particular contexts and questions? What are important “data habits of mind” and how do learners acquire them?

Projects

Examples of NSF Cyberlearning projects that overlap with topics discussed in this primer.

- Collaborative Research: Designing the Impact Studio -- Dynamic Visualizations in the Write4Change Networked Community
- DIP: Data Science Games - Student Immersion in Data Science Using Games for Learning in the Common Online Data Analysis Platform
- CAP: Data Science, Learning and Youth: Connecting Research and Creating Frameworks
- CAP: Innovating Data-driven Methodologies for Documenting and Studying Informal Learning
- DIP: Collaborative Research: STEM Literacy through Infographics

More posts: data-visualization

Resources

Data Science Education Technology Conference
Oceans of Data Institute
Coursera Catalog: Data Science Courses and Specializations

Readings

Conway, D. (2010). The data science venn diagram. [Web page].

Erickson, T. E. (2012). Designing games for understanding in a data analysis environment. In Proceedings of the International Association for Statistical Education (IASE) roundtable on “virtualities and realities”. Cebu, Philippines: International Statistical Institute.

Finzer, Erickson, Swenson, & Litwin (2007). On getting more and better data into the classroom. Technology Innovations in Statistics Education, 1(1).

Finzer, W. (2013). The Data Science Education Dilemma. Technology Innovations in Statistics Education, 7(2). Los Angeles, CA: UCLA Department of Statistics.

Kastens, Kim. (2015, May). Data Use in the Next Generation Science Standards (revised edition) [White paper]. Waltham, MA: Oceans of Data Institute, Education Development Center, Inc.

Citation

Suggested citation:

Vahey, P., Finzer, W., Yarnall, L., & Schank, P. (2017). CIRCL Primer: Data Science Education. In *CIRCL Primer Series*. Retrieved from http://circlcenter.org/data-science-education