Exploring the Essential Elements of

Data Science Education

By William Finzer and Frieda Reichsman

She scans the numbers, scrutinizing thousands of cases and dozens of attributes. Something doesn't look right. Missing values? Incorrect coding? Fixing those will be a start. Cleaning, checking, re-sorting—gradually she cajoles the enormous array into a workable form. Now the fun part begins: teasing out hidden relationships and laying the groundwork for deeper analysis. She merges the tamed dataset with another and digs in, stacking, filtering, creating graph after graph, hot on the trail of unseen patterns and new insights.

This is an increasingly common scenario. As complex datasets begin to underpin every aspect of modern life, data scientists are everywhere, applying their advanced programming and statistics knowledge, disciplinary understanding, and data wrangling skills. In high-tech science labs and enormous automotive assembly lines, in tiny fashion startups and standalone agricultural greenhouses, people with data science skills and understanding are finding patterns and guiding decisions. From combating global warming to feeding the growing population, reducing violence, and increasing equity, data science will be at the heart of future solutions to every significant problem in society. In this data-rich future even everyday life decisions such as choosing a health care provider or political candidate will demand new fluency in interpreting, tempering, and critiquing claims derived from large or complex datasets.

At the Concord Consortium, we believe that basic data fluency must be a skill offered to all, which is why we've worked to foster learning with and about data for decades. We believe building understanding and habits of mind around data is critical, and we believe all students should be able to understand and analyze complex data without hours of coding lessons or years of advanced mathematics. To that end, we're spearheading the field of data science education at the pre-college level. In an effort to identify and further the essential elements of data science education, we have developed software and curricula, hosted dozens of webinars and meetups, and researched student learning with data.

Modes of working with data

Thanks in part to an increased emphasis on data in both the Common Core State Standards for mathematics and the Next Generation Science Standards, students are examining data in more and more classrooms. However, data can be used in many different ways. We've studied student learning with data and have identified six different modes through which students can work with data. Each mode has the potential to bring simplicity or sophistication to the study of data:
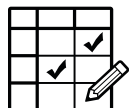
- Entering data
- Examining data displays
- Collecting data
- Exploring data
- Discovering with data
- Problem solving with data

These six modes overlap and reinforce the organic, cyclical nature of working with data. And importantly, they engage students with an essential aspect of data investigation—what one might call "messing around." Spending time "playing" with data is a critical step in providing students a feel for what the data might tell them—and very different from many traditional activities in science or math class. When students approach a dataset by initially messing around—often through one of the modes outlined below—they build familiarity and understanding that sets the stage for key questions and conjectures to emerge.

While these modes are not the only ways students can engage with data, they are all important for providing students a natural feel for data's complexity and nuance.

Modes of working with data

Entering data. Data only exist after they have been recorded, and there are myriad ways of doing this. A kindergartner adds a sticker in the "dog" column of a dot chart to record his pet. A third grader counts the different kinds of books on the bookshelf and writes it down on paper. A seventh grader measures and records the distances her classmates throw a shot put in the cells of a spreadsheet. And a high school biology student takes photos of plants in an experiment, loading them into an online database. Entering data can feel mundane or exciting, but to enter data is to know its origins and take ownership of it.

Examining data displays. Science textbooks are full of data displays: tables, plots of distributions, data-rich maps, scatterplots, pictograms, and lists, just to name a few. But data displays also appear on scoreboards, on computer screens, on automobile dashboards, and in science journals. Though often students are tasked with "reading" a graph or with showing how a data display illustrates a given concept, there is potential for considerable challenge in extracting from the display puzzling phenomena, arguments in favor of a point of view, and deep relationships that only become apparent after multiple encounters.

Exploring data. The sheer volume of easily accessible, unexamined data puts students in the role of explorer. Students probe the data landscape, familiarizing themselves with data sources, data structure, and types of attributes, and dive into a data world rich with possibility.

Collecting data. Data do not collect themselves. They emerge from a designed process. As students gain experience with making use of data, they increasingly appreciate the thought that goes into figuring out how to relate investigative questions with decisions about what will be the "case" or unit of observation, what attributes of the case are relevant, how many observations are needed, and how to most usefully record and store that data. What at first appears simple reveals itself as the subtle process of modeling the world with data.

Discovering with data. With exploration comes the possibility of discovery. Students may set out to find a particular relationship only to discover that there is none, or, conversely, happen on an unexpected strong correlation. Today students can also work with data that are previously unexplored. When they do, the discoveries they make are actually new discoveries.

Problem solving with data. At a certain stage students reach a level of comfort with data where data become a tool for solving problems. When a student recognizes the need to solve a problem and recognizes that having data about the problem could help yield a solution, using data begins to become second nature, and looking for data a habit of mind.

Making great data experiences

To ensure that students engage in these modes of working with data, they must have access to datasets of appropriate size. Data that have many more than two attributes compels students to look at the data from different dimensions, make multiple representations, and ultimately find original discoveries. Large datasets also help learners become comfortable feeling "awash in data" and foster necessary data habits of mind.

At the same time, students must have access to intuitive data tools that allow them to visualize relationships and make sense of data. Our easy, web-based Common Online Data Analysis Platform (CODAP) is designed for this purpose and has tools for beginners as well as advanced features for experienced users. One key feature is that representations link dynamically across tables, graphs, and maps.

By working with data frequently and repeatedly, learners develop experience and competence, gaining fluency with the data moves necessary for structuring, examining, and diving into data, and ultimately building excitement for their ability to work with data. This enthusiasm is the cornerstone of deepening students' understanding of data as a tool for solving problems in the world, and is key to preparing them for life in a world immersed in data.

Free Resources

Data Science Education
concord.org/data-science-ed

CODAP
codap.concord.org

Webinars & Meetups
concord.org/meetup

The Concord Consortium