

The Data Science Education Dilemma

1. THE PROBLEM

The ongoing exponential growth (see Figure 1) of our society's storage of and reliance on data is astonishing. The growing gap between the need for a data savvy citizenry and the data science education of students is equally astonishing, greatly troubling, and extremely perplexing. From the author's experience with K–12 education in the U.S., meaningful student encounters with data analysis are rare, teachers' data skills and comfort using data-driven lessons are nearly non-existent, curriculum developers' motivation to weave use of data into classroom materials is extremely low, researchers' efforts to build a knowledge base of how students learn about data are limited, and policy makers' attempts to strengthen data science education in the schools are ill-informed.

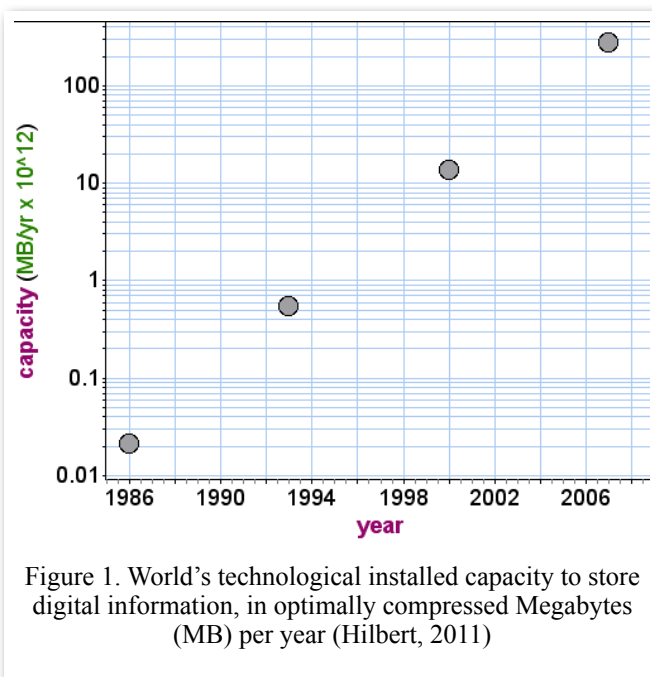


Figure 1. World's technological installed capacity to store digital information, in optimally compressed Megabytes (MB) per year (Hilbert, 2011)

In the fractured landscape of school subject-matter disciplines, data science has no natural home. Mathematics, already uncomfortable housing statistics, resists the incursion of rich contexts that interfere with students' ability to focus on abstraction. The physical and biological sciences, steeped in experimental data as they are, are so overwhelmed with teaching concepts that data get relegated to the role of illustration. The social sciences, though increasingly data-oriented in practice, have hardly begun to transfer to educators the quantitative perspective necessary to bring meaningful use of data to K–12 social science classrooms.

The challenge then is to figure out how to change educational systems so that students emerge from their schooling with data science skills and conceptual understanding needed to participate fully in society as citizens and workers.

2. CONSIDERATION OF EXPONENTIAL GROWTH OF COMPUTATIONAL RESOURCES

Table 1 describes the last 30 years of change in terms of the author's computers then and now. Three questions that help bring this change home are: What would it be like if you could get from one place to another 3,000 times as fast? What if you were 1,000,000 times as tall? What if the earth's population were 30,000 times what it is now? The difficulty in imagining such changes points to the difficulty we have in understanding the implications of the state change in civilization caused by the many orders of magnitude increase in computational resources. The change affects all components of society, not least education.

Table 1. The author's personal computing resources in 2012 compared with those 30 years prior.

	1982	Factor	2012
CPU Speed	1 MHz	3,000	3 GHz
Hard Disk Capacity	1 MB	1,000,000	1 TB
Transmission Rate	30 B/sec	30,000	1 MB/sec

Not long ago data was expensive. There wasn't much of it. Data was the bottleneck for much of human endeavor. Not today. We are no longer data-limited but insight-limited (Abbott, 2009). A huge reservoir of data is backing up, with untapped discoveries within. The people who know how to work with data are in short supply. These people are data scientists. Future data scientists are in elementary school today, but unless the K–12 school curriculum changes so that data science education is a significant part of a young person's learning experience, many fewer data scientists than we need will emerge in the next decades.

3. WHAT IS DATA SCIENCE?

An early use of the term data science appears in "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" in which Cleveland (2001) describes data science by laying out a university training programming. His description is notable for including both a *computing with data* strand as well as a *pedagogy* strand.

The Venn diagram in Figure 2 provides a simple mechanism for thinking about what constitutes data science. First, there is the disciplined, quantitative thinking found in mathematics and statistics. From statistics comes an understanding of variability and experience using statistical tools to work with data. Second, substantive expertise gives a data scientist an understanding of the disciplinary context for a data set without which choosing a valid analysis methodology will be difficult or impossible. Finally, computing and data skills which, combined with creative problem solving abilities, allow one to see inside the machine and to visualize the structure of the data.

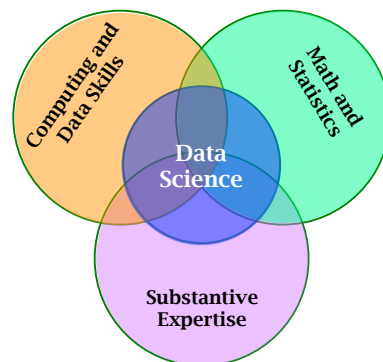


Figure 2. Data Science Venn diagram adapted from Conway (2010)

Table 2. Representative data science tasks and the portion of the Data Science Venn diagram from which they primarily draw. (CD = Computing and Data Skills, MS = Math and Statistics, SE = Substantive Expertise)

Task	Primary Expertise
Set up a server as a repository of data streaming in real time from a large array of geographically distributed sensors.	CD
Explain the origin of outliers in a particular data set.	SE
Decide to what extent the conclusions drawn from analysis can be generalized.	MS
Design a data visualization suitable for publication in an article for non-experts.	CD
Decide what data should be gathered.	SE
Decide whether certain disparate data sets can be meaningfully merged.	SE
Automate the merger of multiple data sets.	CD
Detect the signal within the noise.	MS
Reduce the number of variables that need to be considered for a particular analysis.	MS
Set up a version management system for data that will be gathered over a number of years.	CD

Table 2 lists some tasks that might reasonably be labeled as part of doing data science. It should be clear that someone well-versed in only two of the three areas of expertise will be seriously hampered in any attempt to work with real data. For example someone with no computing skills will have difficulty automating the process of cleaning data and be at a loss for how to transform a very large raw data set into something ready for analysis. In practice data scientists work as members of teams, and it is the team as a whole that embodies the complete range of required expertise. Even so, data scientists with a complete set of these inter-disciplinary skills, combined with the ability to collaborate productively, will likely contribute more than those who are lacking in one or more areas. For many projects it may be that just as data come to underlie every aspect of the work, data scientists frequently provide the glue that holds the work together.

Computer scientists have begun to pay serious attention to the tasks of data science, as detailed in an article on “data wrangling” (Kandel, 2011) that describes the current complexity of transformations a data scientist makes to bring data into a “credible and usable” form.

4. CONFRONTING THE DILEMMA

This section attempts to put forward some ideas, drawn primarily from the author’s work in designing educational software, on what needs to be done to bring data science into K–12 education. The gap between the magnitude of the problem and small scale of the ideas is very large, and no attempt is being made here to give a survey of the many outstanding contributions others have made to curricula and technology in this area. The four areas considered here are inter-disciplinary use of data, software and its role in providing access to data, goals for teaching with data that transcend subject matter content, and research that needs to be done to improve curriculum development.

4.1 Data Across the Curriculum

Two observations suggest that data science should not aspire to become a new subject area in the school curriculum: There are more subject areas already than fit comfortably; and learning to work with data makes most sense within a relevant context. Imagine a data science experience, K–12, in which work with data occurs in every subject as appropriate to the current content being studied. In history, migration patterns are studied through analysis of census microdata. Most science labs involve use of data science to reveal not just the textbook “answers” to the investigation, but the interesting deviations from what is expected. In physical education, data gathered from video lead to insight into how to improve athletic performance.

A prerequisite to meaningful implementation of “data across the curriculum” is agreement on a progression of encounters with data science concepts and skills. Without such agreement each teacher at each grade level is condemned to work with a lowest common denominator of prior experience. Assuming anything more requires a time-consuming digression from the subject matter lesson to bring students to the desired level of competence. With a coherent conception of data science education, teachers can build incrementally on skills and knowledge acquired earlier in a student’s school experience.

4.2 Access to Data

It is not possible to have meaningful learning experiences in data science without frequent use of the technologies that have brought about the data revolution. Drawing graphs by hand or sorting piles of surveys do not build understanding of distributions or how to clean data. Online tools for gathering, cleaning, exploring, and visualizing data are beginning to reach a level of accessibility and usability appropriate for school use.

The author is the lead developer of *Fathom*® *Dynamic Statistics Software* (Finzer, 2007), a desktop application widely used in secondary schools and colleges for introductory statistics, mathematics, and science. *Fathom* has many features especially helpful in immersing students in data, not least among them multiple ways to access data, primary among them being:

- type data by hand into a spreadsheet-like table;
- copy data from a spreadsheet, a web page, or as tab-delimited text and paste into *Fathom*;
- import from a tab-delimited text file;

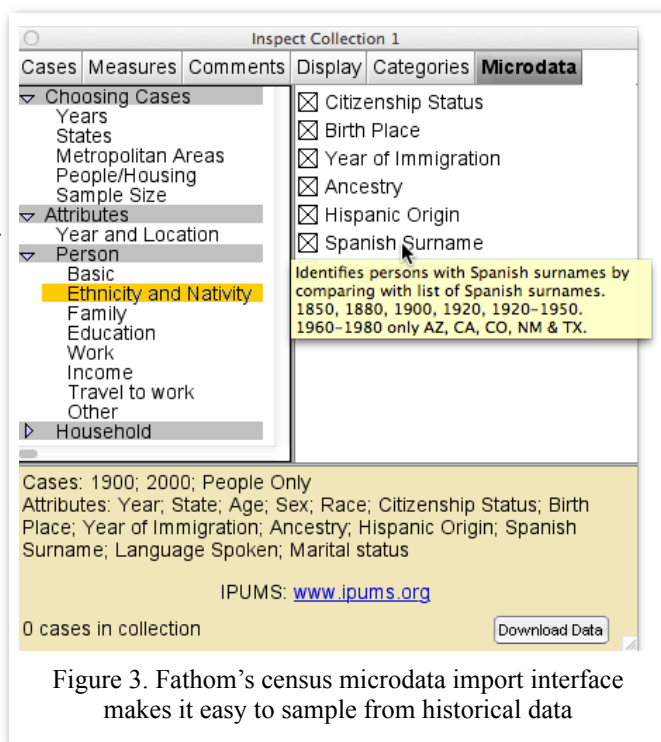


Figure 3. Fathom’s census microdata import interface makes it easy to sample from historical data

- drag the URL from a web page containing data into a *Fathom* document. *Fathom* parses the HTML from the page and looks for data in a variety of formats;
- as shown in Figure 3, use the U.S. Census microdata interface built into *Fathom* to select a sample of individuals who filled out the long form of the census from 1850 to 2000, choosing from 65 attributes;
- connect one or more Vernier probes to the computer and configure *Fathom* to accept data from the probes in real time;
- in *Fathom* create a survey that appears on the web. Responses to the survey can be brought into *Fathom* at the push of a button;
- create a simulation in *Fathom* that generates data.

With diverse methods for acquiring data, students have opportunities for data discovery and immersion that can engage them as data scientists (Finzer, 2007).

4.3 Data Habits of Mind

A “habit of mind” is a way of thinking, questioning, and problem-solving (Cuoco, 1997). A “data habit of mind” is a habit of mind that grows out of working with data. An important objective of any attempt to integrate data science into the curriculum should be to inculcate learners with data habits of mind. To better understand how these can be developed, consider materials being developed in the Data Games project.

4.3.1 Data Games and Habits of Mind

With support from the National Science Foundation for the Data Games project at KCP Technologies and UMass Amherst, the author and his colleague Cliff Konold are exploring a promising, technology-based vehicle for integrating data science into the teaching of mathematics using games.

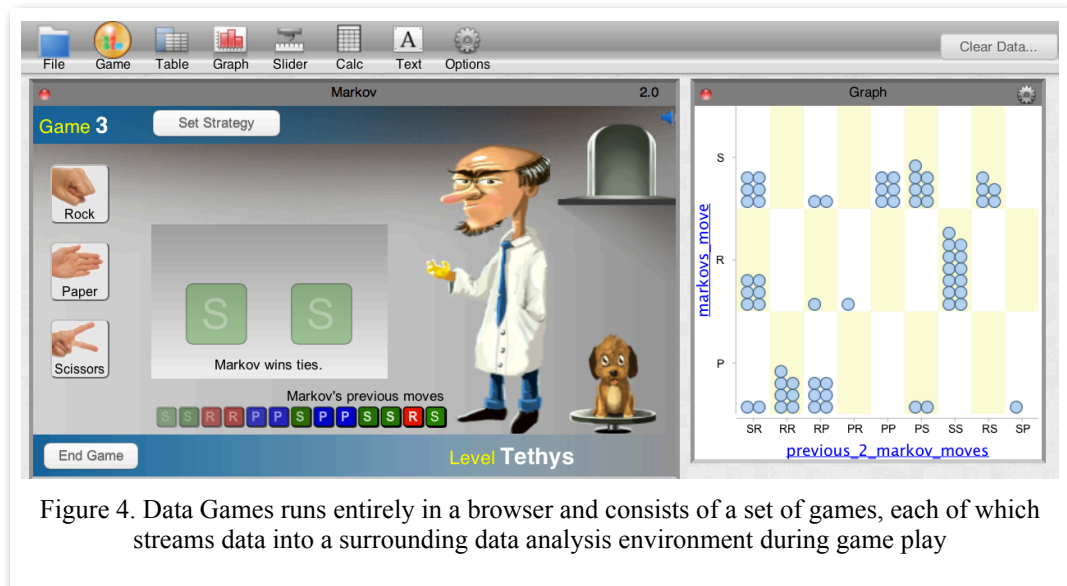


Figure 4. Data Games runs entirely in a browser and consists of a set of games, each of which streams data into a surrounding data analysis environment during game play

The idea of a Data Games activity is that students play a simple game that streams data to a surrounding data analysis environment, so that the data is available for exploration. An example is shown in Figure 4 in which the player attempts to save the dog Madeleine from Dr. Markov by playing the game of rock, paper, scissors. Unfortunately for Madeleine, the game is rigged so that the evil doctor wins all ties. The player’s only hope

is to make good use of the graph that shows the conditional probabilities associated with Markov's strategy.

Students are expected to analyze the data to create a model of what is happening in the game with which they can improve their strategy. By design, the models that underlie the games are based on mathematical concepts that are a normal part of the school curriculum; e.g. linear relationships, quadratic equations, statistical measures, and probability. Erickson (2012) describes the challenges faced in designing and teaching with these games.

Consider two seemingly straightforward data habits of mind relevant to the Data Games project.

4.3.2 *Look for the Data*

One such habit of mind is to *look for data* that might be helpful in solving the current problem. A data game generates data. The data are easily examined and highly relevant to improving one's game playing strategy. But students do not necessarily come equipped with a built-in look-for-the-data habit of mind. Rather, it develops with repeated experiences in which they are given the opportunity to look for the data and then discover that these data are quite useful.

A typical progression for a high school student is: (1) Practice playing the game with minimal use of quantitative thinking. (With a well-designed data game, the students won't improve their scores much this way.) (2) Begin using the information displayed by the game in a systematic way. (3) Express a desire for a record of prior moves. "This must be keeping track of what I'm doing. (4) Make systematic use of tabular or graphical representations of prior moves and games to devise an improved strategy that results in higher scores.

In pilot classes with students who have repeated encounters with data games, students show that they have adopted this habit of mind by creating a table and/or graph even before they begin playing the game as a way of saying, "I know I'm going to need to work with the data."

While there are certainly students who follow this kind of progression unaided, observations in field test classrooms suggest that the learning environment plays a critical role in developing data habits of mind. We found that if it were possible to score well on a game through practice as opposed to using data, many students would choose to do so, however tedious it might be. After all, practice is the *only* route to mastery of many games students play. Eventually we made sure that practice was futile for all but the first level or two of a game.

4.3.3 *Graph the Data*

Consider the habit of mind *graph the data*. A graph may make evident a potentially useful pattern in the data, a pattern that is difficult to discern by staring at a table of numbers. In Data Games, creating a graph requires no more than a button press, and, by default, it shows the most useful configuration of attributes. Yet, unless instructed to, very few students in the Data Games field test make graphs in their first encounters with a data game. They are more likely to make a table. Some conjectures about why are: Graphs are unfamiliar and will require more effort than tables to use; tables appear to give more exact information than graphs; graphs are pigeon-holed as function plotting tools rather than tools for data exploration.

Most students in the field test did not initially have the habit of mind of graphing data. By the end of the 2012 school year, we should be able to determine whether they acquire it after repeated exposures so that, at least in the context of a data game, they voluntarily make graphs and use them to figure out a strategy.

4.3.4 Other Data Habits of Mind

Among the many other data habits of mind that eventually will become part of a data science curriculum, some are beginning to show themselves as important in the Data Games work: Become immersed in the data, use (and invent) measures, and look for and tell the story behind the data.

4.4 Learners' Conceptions of Data¹

Most education research that has to do with data falls within the traditional boundaries of statistics education. It is research about students' understanding of the center and spread of distributions, of informal inference, of sampling distributions, and of some standard plots. But new research frontiers are opening up, for example, those building on work by Chris Hancock (1995). What are learners' conceptions of data, and how do these change with increasing experience working with data? What do learners think is going on when they transform data to clean it, parse it, change its scale, merge it, or do any of the many other things that are routine for a data analyst?

Ongoing research led by Konold in the Data Games project may serve as an example of the kind of work that can be done to provide the foundation of knowledge on which tools and learning environments are built. A situation involving traffic was used both in an interview protocol with individual students and a group protocol for whole class administration. Snapshots of two road segments were presented to students and they were asked to create an organized record (not a drawing) of all the data values that appear on the snapshots.

The task was administered to more than one hundred students ranging from grades 7–14 and the responses are still being analyzed. Roughly we can separate these responses into five groups: narratives, attempts to summarize as opposed to record, unstructured listings of values that lose most of the case structure, hierarchical (nested) structures of two or three levels, and fully flat row-by-column tables with each row representing a vehicle.

A paper based on this research is in preparation.

5. CONCLUSION

Statistics itself is rapidly changing under the influence of the data revolution (Friedman, 2001). Inevitably, statistics education lags behind. As Cobb (2007) has argued, statistics courses are mired in teaching techniques developed by pre-computer-era statisticians to circumvent their lack of computational power.

Statistics educators, more than educators of any other discipline, are in a position to advocate for the changes required to increase the flow in the data scientist pipeline. They

¹ In the initial publication of this article, this section included two figures and a brief discussion of some preliminary results. These were removed by the author because their inclusion was premature given the planned publication of a full research paper.

understand data and make use of data habits of mind. Many have sophisticated computing and data skills and have acquired substantive expertise. Some have begun to make important changes in the undergraduate statistics curriculum and advocate for a new approach not just to the way statistics is taught but to the content of the introductory course (Nolan, 2007). But the advocacy role will require statistics educators to step outside their comfortable disciplinary silo and reach out to educators in other science and math disciplines to understand how data science can be integrated into their curricula.

The challenge is big. The pressure on educational establishments to make changes will grow inexorably as the need for more data science in industry and science becomes ever more apparent. In the U.S. the federal government can help by forming a National Center for Data Science Education as a place to bring together people from science, industry, and education to establish paths of communication, to initiate research, to incubate new teaching methodologies, and to recommend education policy changes.

REFERENCES

- Abbott, M.R. (2009) "A New Path for Science" in Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm : Data-Intensive scientific discovery*. Redmond, Wash.: Microsoft Research.
- Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1), 21-26.
- Conway, D. (2010). The data science venn diagram. Dataists [Web page]. Retrieved February 9, 2012, from the <http://www.dataists.com/2010/09/the-data-science-venn-diagram/> database.
- Cuoco, A; Goldenberg, E. P., and J. Mark. (1997). "Habits of Mind: an organizing principle for mathematics curriculum." *Journal of Mathematical Behavior*, 15(4), 375-402.
- Data-Driven Mathematics Series (1998), New York: Pearson Learning (Dale Seymour Publications)
- 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, W. (2007). *Fathom dynamic data software* [Computer Software]. Emeryville, CA: Key Curriculum Press.
- Finzer, Erickson, Swenson, & Litwin (2007). On getting more and better data into the classroom. *Technology Innovations in Statistics Education*, 1(1).
- Friedman, J. H. (2001). The role of statistics in the data revolution? *International Statistical Review*, 69(1), 5-10.
- Hancock, C. (1995). The medium and the curriculum: Reflections on transparent tools and tacit mathematics. In A. A. diSessa, C. Hoyles, R. Noss, & L. D. Edwards (Eds.), *Computers and exploratory learning*. (pp. 221-40). Berlin: Springer.
- Hilbert, M. & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science (New York, N.Y.)*, 332(6025), 60-5.
- Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., et al. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271-288.
- Murdock, J., Kamischke, E., Kamischke, E., & Key, C. C. (2007). *Discovering algebra : An investigative approach*. Emeryville, CA: Key Curriculum Press.

Nolan, D. & Temple Lang, D. (2007). Dynamic, interactive documents for teaching statistical practice. *International Statistical Review*, 75(3), 295-321.