# WHO WILL TEACH THEM ABOUT DATA?

## THE RESPONSIBILITY OF MATHEMATICS AND STATISTICS EDUCATORS TO SUPPORT THE INTEGRATION OF DATA ANALYSIS ACROSS ALL SUBJECTS

William Finzer and Vishakha Parvate
KCP Technologies, Emeryville, CA, USA

## ABSTRACT

In the United States, teaching of data analysis lags far behind the need for a data literate citizenry. Important topics not included in most curricula are: data structures beyond row and column, data types beyond numbers and strings, issues surrounding gathering and maintaining data, the value of data sets as they grow over time, analysis of data sets with large numbers of variables, recent advances in data visualization, ethical issues surrounding data, and the role of context. Because mathematics and statistics educators are accustomed to quantitative methods, they bear a special responsibility to help bring about an expanded view of data analysis as a methodology underlying mathematics, science, and social science.

## BACKGROUND

From where will the data-savvy scientists, engineers, and technicians needed to fulfill data's potential come? K–12 students in the United States currently emerge from their schooling thinking of data analysis as three things: making and reading charts, curve fitting, and making statistical inferences (if they are fortunate enough to take statistics). They may encounter data in science courses, but their interaction with that data seldom qualifies as meaningful data analysis. This is meager preparation for data-based tasks such as mining air traffic data, maintaining human genome data, and designing better ways for people to search and filter their email. Nor do these experiences lay a foundation for participation in a society that already relies on previously unimagined quantities of data for its well-being and economic growth. Past president of the Mathematics Association of America Lynn Steen (2001) says it well:

> *Quantitative literacy, also called numeracy, is the natural tool for comprehending information in the computer age. The expectation that ordinary citizens be quantitatively literate is primarily a phenomenon of the late twentieth century. … Unfortunately, despite years of study and life experience in an environment immersed in data, many educated adults remain functionally illiterate.*

The GAISE report (Franklin, 2007), endorsed by the American Statistical Association, both describes the problem and suggests a solution:

> *Sound statistical reasoning skills take a long time to develop. They cannot be honed to the level needed in the modern world through one high-school course. The surest way to help students attain the necessary skill level is to begin the statistics education process in the elementary grades and keep strengthening and expanding students' statistical thinking skills throughout the middle- and high-school years.*

The ubiquity of data impacts all fields of endeavor and all aspects of our lives. The recent, dramatic growth of search engine companies such as Google presages a time in which the transformation of data to information, and then to knowledge, constitutes an industry on the same scale as agriculture and manufacturing. The information revolution that has made facts and ideas instantly accessible all over the world is but a precursor to a data age in which we mine cures for illness from the human genetic code, accurately predict climate change, and construct data-driven simulations to determine the effects of large-scale educational interventions. The data age has arrived, but educators in the United States have barely begun to integrate data exploration and

analysis into the curriculum to help meet the growing need for a data-literate citizenry and work force.

## WHAT YOU NEED TO KNOW ABOUT DATA THAT ISN'T (YET) TAUGHT IN SCHOOL

The National Council of Teachers of Mathematics lays out ten standards for school mathematics in the United States. The standard for data analysis and probability is shown at right. It looks good. It would seem that all aspects of working with data should fit within these four bullets, and they probably can. But if you compare what is happening in practice in the workaday world with what is happening in math classrooms, you observe a disturbing and growing gap. Students in the United States' schools are missing major aspects of what it means to work with data. We describe some of them here.

> *Data Analysis and Probability* (NCTM, 2000)
> Instructional programs from prekindergarten through grade 12 should enable all students to—
> - formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them;
> - select and use appropriate statistical methods to analyze data;
> - develop and evaluate inferences and predictions that are based on data;
> - understand and apply basic concepts of probability.

Beyond rows and columns—Ask someone what data look like and the likely response will be "rows and columns of numbers, like a spreadsheet." Typically, each row represents a *case* and each column holds a variable or *attribute*. Each case has the same attributes as every other case; they are homogeneous. This structure works well as a *model* for many things in the world: a classroom of students, an experiment's worth of measurements, a time series of demographic measures for a country, or a catalog of books in a library. But such a framework proves vastly inadequate when we consider the bewildering variety of data structures computer scientists use to model the world: Hierarchical structures to mirror school districts containing schools containing classrooms containing students; relational structures to reflect the way a store's inventory depends on suppliers; trees, graphs, and semantic networks to model things as disparate as family trees, the grammar of sentences, and the changing state of the Internet. The world cannot be shoehorned into rows and columns. To finish one's pre-college education never having encountered any but tabular data is like restricting one's study of biology to one-celled organisms.

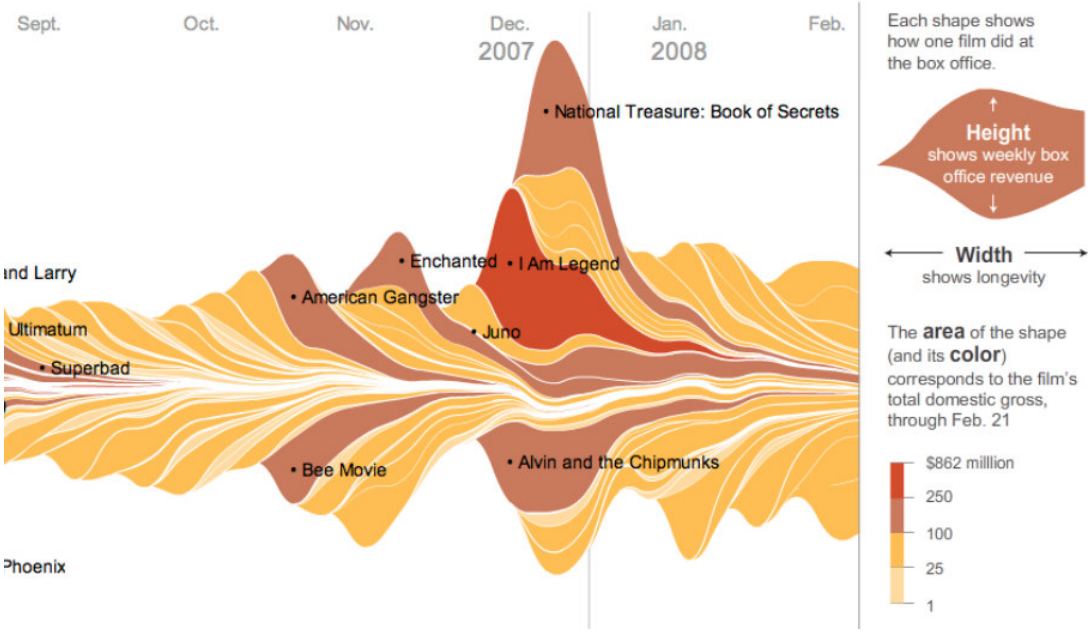
*Beyond numbers*—All digitally encoded data exists as bits, the "on" or "off" states usually described as the numbers 1 and 0. Properly aggregated, these bits can be interpreted as integers, reals, strings, quantities with units, date-time, pictures, audio tracks, and DNA sequences—all different *types* of data. Pre-college students in the United States are led to believe that all important data are numeric, on which computing the mean and standard deviation is somehow significant. We wonder whether this misconception acts to filter out some learners who do not love numbers but might otherwise love data from engaging in data-driven investigations. Would as many students come to love literature if they were only exposed to poems, never short stories or novels?

*Data gathering and maintenance*—Nearly all classroom encounters with data create the impression that most data sets are small (less than 500 cases), pre-assembled, and clean. Quite the opposite is true. Data sets with which real work is done span a huge range in size from tens or hundreds to many millions of cases. Planning and carrying out the gathering of useful data typically requires many times the effort that will be spent on data analysis, and often the assembly is ongoing. Real data are dirty and noisy, of variable quality, and with documentation that ranges from non-existent to voluminous. Most real data sets are not static, but dynamic and changing. Some events in ongoing maintenance are: More data are added, new computed attributes are added, dirty values are found and marked or removed, a way is discovered to merge one data set with another to make them both more useful. Today's students are like diners in a fine restaurant, kept unaware of the vast industry that lies behind the swinging kitchen doors extending back to distributors, farmers, and fertilizer manufacturers.

*Data in time*—You and your students find a place in the school yard to conduct a census of living things. Back in the classroom you all attempt to analyze and draw conclusions. What can you say? There were three pill bugs, over 500 blades of grass, 73 ants. So? But the next year you tell your students about the previous year's work and that they will gather data from the same place on the same day of the year at the same time. Will things be the same or different? The conclusions drawn can have to do with comparing one year to another. Do this for five years and you've got the possibility of looking for trends, separating them from fluctuations. The value of data can increase dramatically as they fit within a series over time. But the work can increase, too. You have to understand the methods used in previous years and you have to document the methods you use in the current year. And you have to make sure the data are stored where they can be found and added to each year. Do students ever have the experience of gathering data that is worth keeping, or are the data they gather always to be thrown away? For students never to know that their effort can add real value to a collection of data is like involving them in a community volunteer group whose work is demolished at the end of every project.

*Beyond flatland*—A single list of values holds no one's interest for very long. We want to know more about each case and to be able to compare one distribution with another. Two values for each case is better: Before and after, male and female, treatment and result, one place versus another. Scenarios in which we measure only two things greatly limit the space we can explore, and they are fairly rare in the real world. We live comfortably in the four dimensions of space plus time. A modestly rich data set might have a few tens of attributes, and it is not unusual to encounter data sets with hundreds of attributes. Restricted to one or two attributes per data set, students do not begin to experience what David Donoho calls the "curses and blessings of dimensionality" (Donoho, 2000). We are born with the ability to perceive two- and three-dimensional patterns, but we are having to invent methods for finding patterns in higher dimensional spaces, and being able to do so is becoming increasingly important to a civilization besieged by climate change, global pandemics, and the persistent scourge of war.

*Beyond box plots, histograms, and scatter plots*—Beautiful and data-rich visualizations in popular media have come of age, as the interactive graphic below from the *New York Times* illustrates (Bloch, 2008). As you move your mouse over shapes in the graph, the name of the corresponding movie comes up. Click and you get a mini-review and a clickable link to full review. In a math or statistics class the focus remains on "standard" plots and on the rules for creating and interpreting them. Meanwhile, data-savvy visualization designers are creating graphics that communicate useful information and yet push way beyond the textbook boundaries. Has the time come for data to enter art class? Perhaps the current emphasis on box plots, histograms, and scatter plots is as artificially constraining as it would be to restricting art students to pencil sketches.



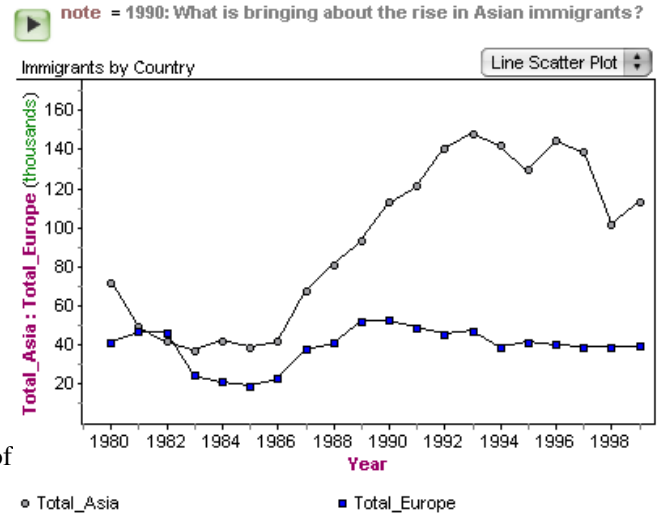
*Privacy, confidentiality, and anonymity*—The ease with which data can be gathered, collated, analyzed, and disseminated raises ethical issues that can and should be addressed in classrooms. Consider: *A student is administering a survey for a project. Who reviews and approves the questions on the survey? How should students taking the survey be briefed on their rights to privacy? Will the subjects of the project be guaranteed that their responses remain confidential, and, if so, how will this be accomplished?* Ethical issues do not easily separate themselves from the mathematical and technical issues of working with data. To ignore them is to ignore the gorilla in the closet.

## CONTEXT

As George Cobb and David Moore (1997) write, "data are not just numbers, they are numbers with a context." Becoming skilled in data analysis requires experience with the interaction of context with mathematical structure.

As part of the research for this paper we interviewed Chris Davies, a project manager at Affymetrix, a genetic engineering company. Chris and his colleagues work with gigabytes of DNA sequence data. When we asked what he looks for in hiring, he chuckled and said he has found that frequently the hotshot programmer/mathematicians don't work out very well because they lack the biology background needed to understand what the data are telling them. They don't understand the context.

Providing a meaningful context for data in a mathematics classroom is sometimes difficult as the teacher may lack needed content area background, and the mathematical focus of the lesson is unlikely to be integrated with the context. A rich, interesting context can pull students so far from the abstract structure of the mathematical situation that they fail to see it altogether. Too little context and the data become just numbers; too much and the lesson is shrouded by complexity. The math classroom isn't always the best place for gaining experience with data!




With other school subjects the problem can easily be reversed. The content to be learned is often embedded in the context underlying the data— for example a data set of elements contains the ingredients for many chemistry lessons, and exploration of historical census data can illustrate and reveal important history themes in a social science class. But the students may not have had enough experience with data analysis to allow them to *find* the lessons within the data; i.e. their lack of ability to perceive the structure of data may prevent them from understanding what lessons the data teach.

Using data in classrooms can introduce this "chicken and egg" conundrum, as the problem of context reveals. The math teacher says, "I can't use very much data in my teaching because the context keeps interfering with the learning of mathematics," while the science teacher says, "I can't do more in depth data analysis because my students don't know enough about how to work with data to actually reach the science to be learned from the data. This brings us, finally, to the question posed by this paper's title.

## WHO WILL TEACH ABOUT DATA?

In the long run, we think the answer to this question will be "nearly everyone." That data now pervade nearly every aspect of life and work suggests that its use should be woven into all

aspects of school learning from the earliest grades onward. Students have long gathered data in science classes; now they need to develop facility with *analyzing* that data (Erickson, 2004). Social Sciences at the school level in the United States have been weak in use of data analysis, but the potential for developing a data driven approach to learning in the social sciences is huge. In addition to the data analysis and probability strand of mathematics education, there are approaches to learning mathematics by using data (Burrill, 1996) (Murdoch, 1998). (We lack the expertise to say anything about use of data in other subject areas.)

Data literacy is much more than making and reading charts, learning to work with the tools of statistical measures, and inference. Data literacy consists of habits of mind and a methodology. To be data literate is to see, understand, and make use of data in the world. It is a way of questioning, thinking, and acting. These skills must be practiced to attain proficiency.

*Habits of mind particular to data literacy:*

- Asking, "Are there data here?" If so, collect them.
- Using the tools of data analysis, especially graphs, in our explorations.
- Following the data. The nuances in data may lead to new insight and discovery.
- Allowing for flexible restructuring of content and breaking out of "traditional" structures to make use of structures suggested by the data.

A problem with saying that nearly everyone will eventually teach with data is that no leader is identified. Other efforts to introduce interdisciplinary curricula, integrated mathematics and science, for example, have met with little success, at least partially due to the lack of a group of educators with a vested interest in making the change.

## THE RESPONSIBILITY OF MATHEMATICS AND STATISTICS EDUCATORS

We mathematics and statistics educators can help initiate this change. We know quantitative thinking. We know data. We know how to work with people in other fields because we have always taught partly in the service of other fields. Also, we have a stake in data being used everywhere because it brings quantitative thinking everywhere. Greater "numeracy" in the general population can deepen and enrich our understanding of the social and scientific worlds, making our work in such arenas more effective. We educators can start the research that will help us learn how to teach with data. We can co-teach the pilot courses with social science and science teachers. We can develop the curriculum materials that will make it possible for large numbers of schools to begin to try this out. Those of us who are statistics educators already have an understanding of how to work with data, an ability we can impart to our colleagues. We can talk to those in industry who have a direct stake in the outcome of this effort, and can seek their guidance in formulating new curricula. We can learn from the decades of effort put forth by the quantitative literacy movement. We can visit departments at colleges and universities where data are being used intensively and find ideas for activities appropriate to the school level. We can mobilize the wider educational community. Let's do it!

## REFERENCES


Bloch, M., Byron, L., Carter, S., Cox, A. (2008). "The Ebb and Flow of Movies: Box Office Receipts 1986–2007," *New York Times*, February 23, 2008. http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

Burrill, G. (1996). "Data Driven Mathematics: A Curriculum Strand for High School Mathematics," *Mathematics Teacher*, v89 n6, 460–465.

Cobb, G.W., & Moore, D.S. (1997). "Mathematics, statistics, and teaching," *American Mathematics Monthly*, 104, 801-823.

Donoho, D. (2000). "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality," Talk given at the American Mathematical Society's "Math Challenges of the 21st Century" conference.

Erickson, T. (2004). SBIR Phase I Final Report: Connecting Mathematics and Science through Data, National Science Foundation. http://www.eeps.com/pdfs/FathomSciencePhaseI_Report.pdf

Franklin, C., et al (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report.* American Statistical Association.

Murdoch, J., Kamishcke, Ellen, Kamishcke, Eric (1998). *Advanced Algebra through Data Exploration*, Key Curriculum Press, Emeryville, CA.

NCTM (2000). *Principles and Standards for School Mathematics.* The National Council of Teachers of Mathematics.

Ontario Ministry of Education (2007). *The Ontario Curriculum Grades 11 and 12,* http://www.edu.gov.on.ca/eng/curriculum/secondary/math1112currb.pdf

Steen, L. A. (2004). *Achieving Quantitative Literacy—An Urgent Challenge for Higher Education.* The Mathematical Association of America.