On Getting More and Better Data to the Classroom

William Finzer[1], Tim Erickson[2], Kirk Swenson[1], Matthew Litwin[1]
[1]KCP Technologies
[2]Epistemological Engineering

1. INTRODUCTION

1.1 A Software Designer's Perspective

Finding a good data set is a joy, and finding a new *source* of data sets is exciting. But finding an entirely new *way* of getting data in front of students can be a revelation. This is the story of how a group of us developing the data analysis and exploration software Fathom™ (Finzer 2007) surprised ourselves with several such revelations.

1.2 Data for Discovery

Before the story begins, please consider our view of the role that data currently play in the classroom from elementary school through advanced placement (AP) statistics and introductory college-level statistics course. Most of the learning that takes place prior to the statistics course is oriented toward developing some facility with data analysis tools: graphical tools such as bar charts, pie charts, and scatter plots plus univariate summary tools such as the mean, median, and range. These tools are treated as the content to be learned. Beyond the carefully chosen data sets provided in curriculum materials and a single classroom survey conducted at the beginning of the year, students get few opportunities to use data analysis tools to explore real data.

Our impression is that instruction in the introductory statistics course has relied on carefully honed collections of data sets that are well-suited to the illustration and investigation of each topic addressed. This judicious selection deprives students of the experience of data discovery. The situation has changed somewhat for the better with the advent of course materials that rely heavily on student-gathered data from simulations and experiments conducted during the class. Also, the increasingly common practice of turning AP Statistics students loose on data-rich projects after completion of the AP exam has provided real opportunities for discovery with data.

What seems to us to be missing are data sets—especially large and highly multivariate data sets—that are ripe for exploration and conjecture driven by the students' intrigue, puzzlement, and desire for discovery. This lack has several causes. First, too few educators and curriculum developers have had significant experience with such data. Second, as an educational community we have not yet figured out how students will become proficient at data analysis. Third, although a great deal of data is available on the Internet, getting data into tools with which students can control the exploration process is still a much more painful and labor intensive process than it should be. This paper deals with the third of these causes.

2. DRAG AND DROP THAT URL

2.1 History of Design and Development

Back in about 1996, a couple years before Fathom's initial release and before Internet connections had become as essential as indoor plumbing, we were puzzling out an interface for importing data from text files. We faced the design question: Should there be a separate **Import** item in the **File** menu or should text files appear along with Fathom files in the **Open File** dialog box? In the midst of the debate, someone suggested you could *drag* a text file from wherever you saw it and *drop* it directly into the Fathom document. (Drag and drop interfaces were all the rage that year.) Uncertain about the best course, we implemented both the **Import** command and drag and drop of a text file, each of which, inconsequential in its own right, would lead to a good thing.

Our first file imports were of tab-delimited text: carriage returns separate cases and tabs separate values. But what was the right thing to do with the first line of text? Should Fathom try to figure out whether it contains attribute names? By answering this question in the affirmative we started down the long and slippery slope of trying to make Fathom ever smarter about extracting data.

The first problem was to isolate data from all the other stuff that might be in the text. Swenson, having previously worked as a programmer for an optics group, pointed out that *finding* data in text is a one-dimensional version of "edge detection" posed in two-dimensions as locating the line that defines the edge of an object. Where is the leading edge and where is the trailing edge of the data in the one-dimensional stream of ascii characters? Swenson began pulling edge detection algorithms into Fathom's data import routines. It wasn't long before Fathom could both find the data and (with quite a bit more work on Swenson's, and later Litwin's, part) make a "reasonable guess" at the structure of the data. The guess eventually became good enough that about two-thirds of the time it produced a reasonable starting place for working with the data.

Then we noticed that a Web page's HTML source is also text and proposed that dragging a URL icon from a web browser should import the data from that page into a collection in the Fathom document. The movie "Movie 1 Drag and Drop.mov" shows how that worked out.

2.2 Implications

Back in 1998 when Fathom was first released, the drag and drop import gesture, combined with excellent heuristics for data import, constituted a new way to access data that changed the way we approached data exploration. Now, when we encounter data on a web page, there is a reasonable chance that we can bring it into Fathom for analysis relatively painlessly. A glance at the case table is usually sufficient to determine whether the import was "good enough." As the time and effort it takes to begin working with data drop below some critical threshold, we become willing to take the exploratory plunge and ask: "What do these data have to tell me?"

To become data literate we must explore data. The data we find ourselves, with which we choose to engage, and with which we make discoveries of interest to us, bring a wealth of experience difficult to acquire with data chosen for us by others. We suspect that data chosen by an instructor was chosen *for some reason*, and that suspicion subtly changes our goal from one of listening to the data toward one of figuring out what we're supposed to learn. We believe that an

inquisitive attitude toward data, a desire to engage with data, and a conviction that real data hold real surprises are as important to inculcate in students as the set of concepts and techniques that form the core of an introductory statistics course.

3. CENSUS MICRODATA -- SAMPLES FROM A HUGE POPULATION

3.1 Development of an Interface

Some of the earliest curriculum materials that came out of the Fathom project had to do with census microdata (Erickson 2000). Erickson purchased a CD of the 1% sample of 1990 data from the U.S. Census Bureau and wrote software to extract a random sample of people from any given Public Use Microdata Area (PUMA) on the CD. For several years in summer institutes we created samples of 500 people from each of the PUMA's of the teacher participants. Exploration of these samples proved to be an excellent way for teachers to begin to experience exploratory data analysis with the Fathom software.

Once our appetites for social science data were whetted, a few geographically scattered samples from a single decennial census were no longer satisfying. A grant from the National Science Foundation allowed us to enter into collaboration with the Minnesota Population Center's IPUMS {http://www.ipums.org/} project, a group dedicated to collating and disseminating United States census data from 1850 to the present. With their help we built an interface within Fathom with which users can work with a sample drawn from one or more decennial census years. The location of the sample can be the whole United States, one or more states, or one or more metropolitan areas. There are over sixty attributes to choose from, and sample sizes can be as large as 5000. The interface for specifying the sample is extremely simple to use and a single mouse click fires off a request to IPUMS computers. It takes about ten seconds to get the data. The movie, "Movie 2 Microdata Import.mov" shows how the interface works.

The revelation here is that it can be so simple to gain access to an extremely large data set through sampling and choosing a minimal set of interesting attributes

3.2 Census Microdata and the Introductory Statistics Course

Census microdata have at least four characteristics that make them well-suited for an introductory course. First, students can easily relate to the basic context; namely that the cases in the collection of microdata represent people who filled out a census long form in a given year. Figure 1 shows the attributes and values for a single case, a married man from Cuba who was earning $34,000 as an architect in the year 2000. We can construct a story about this man and the story invites questions about him: How old was he when he came from Cuba? How does his income compare with others? These stories invite conjectures that lead to investigations: What proportion of Miami's population is born outside the U.S. and how has that changed over time? Is Miami significantly different than the rest of Florida?

Second, the cases can genuinely be regarded as a random sample from an easily specifiable population. So, if we determine the proportion of people in Miami born outside the U.S. for two different census years, it is appropriate to estimate the confidence interval of the difference of these proportions and notice whether it includes zero or not. Inference is relevant and useful in this context. (See Figure 2.) If we discover that the sample is too small to answer our question, we can often go back and get a larger one. We may start on a fishing expedition, but we usually have the possibility of repeating the experiment. (Fathom only draws from what IPUMS calls the "small" sample, and so it can easily happen that our sample exhausts the available data, in which case repetition of the experiment is not possible because we get the same cases each time we sample.)

Attribute	Value
Census_year	2000
State_FIPS_code	Florida
Metropolitan_area_Detailed	Miami-Hialeah_ FL
Family_members_in_household	5
Number_of_siblings	0
Age	31
Sex	Male
Race_General	White
Marital_status	Married_ spouse present
Birthplace_General	Cuba
Ancestry	Hispanic
Educational_attainment_recode	4+ years of college
Labor_force_status	Yes_ in the labor force
Occupation_1950_basis	Architects
Industry_1950_basis	Construction
Total_personal_income	34000

Figure 1: The table of attributes and values for a single person out of a sample of 500.

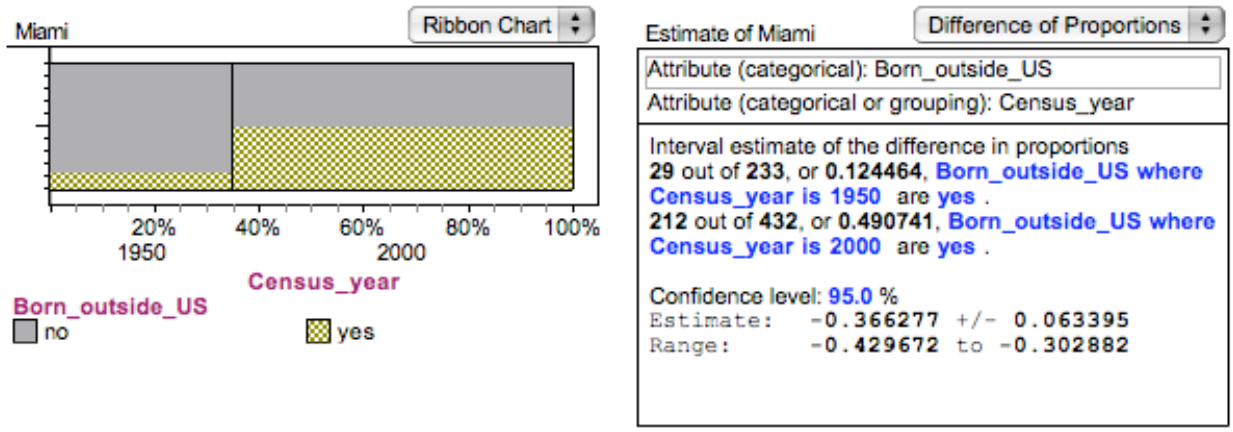

Figure 2: A Fathom ribbon chart and estimate of difference of proportions shows that proportion of foreign born in the year 2000 is significantly different than that of 1950.

Third, these data invite conversation. We can start the conversation with just a few attributes and add new ones as we discover the need for more kinds of information about the people. The hugely multivariate nature of these data need not overwhelm us. We have the sense that we could discover in these data something previously unknown, stepping into realms unvisited by anyone before us.

Fourth, surprises await us at every turn. The outliers astound and befuddle us; expected differences fail to appear; an unexpected trend compels us to rethink our assumptions. In most explorations we come to wonder how the data were collected and what the exact wording of the question was. (Fortunately, we can find most of what we need to know by following the IPUMS link provided in the Fathom interface.) Our skepticism grows even as does our understanding of what can and cannot be learned from these data.

Data stored on web pages and census microdata give us access to data that have already been gathered. Next we consider data manufactured in the moment.

4. REALTIME DATA

4.1 Data Modeling in Science Classrooms

Without data there is no science. Erickson decided to find out what happens with data in science classrooms (Erickson 2005). He found that many experiments are done and much data are gathered. But little mathematics and virtually no statistics are brought to bear on analysis of the data. He saw that Fathom, especially if it were directly connected to sensors that measure such things as distance, force, temperature, light intensity, pH, and the precise time at which a photogate is blocked or cleared, could provide a powerful tool with which this woeful situation might be changed. Together we designed and implemented an interface for this purpose.

We first had to confront our egos. What made us think we could do this any better than it had already been done? Probe hardware manufacturers provide excellent software for gathering and analyzing probe data. (See, for example, Logger Pro from Vernier, http://www.vernier.com/.) Even if we were convinced that Fathom's data analysis environment was more capable than theirs, students could copy data from the probe software and paste it into Fathom for analysis. Why bother to build a direct connection?

Early prototypes of lab activities and Fathom interfaces provided a partial answer: Constructing a model part way through a lab instead of waiting until the end when all the data have been gathered seemed to us to increase the chance that students pay close attention to the data and make necessary modifications to their experimental setup and/or model. Erickson found that, more often than not, deviations from what students expected stimulated significant science learning, and that with Fathom he could write directions for lab activities that strongly resisted the cookbook mentality that students have too often been trained to bring to science lab.

4.2 Temperature Versus Time

We'll illustrate with an account of a simple investigation Finzer and Swenson undertook, one of the first we tried with Fathom. We proposed to measure the cooling of hot water in a coffee mug. The movie "Movie 3 Temperature Experiment.mov" shows the setup, the gathering of data, and the beginning of the building of a model.

We conjectured that the temperature would fall exponentially. So, while the measurements were coming in at one per second, we plotted an exponential curve through the points in a plot of temperature versus time. For simplicity we forced the curve to go through the temperature at

Time = 0. We had one parameter, **k**, that determined how rapidly the temperature decreased. This parameter was defined by a slider, so we could dynamically change it to fit the data.

We found, to our puzzlement, that the data were not fitting our model *no matter what value of* ***k*** *we used.* Resolving this discrepancy led us to at least two "Aha!" insights, one about the nature of exponential curves (they can have asymptotes other than zero) and the other about the physics of cooling (it is the difference between room temperature and the current temperature that determines the rate of cooling).

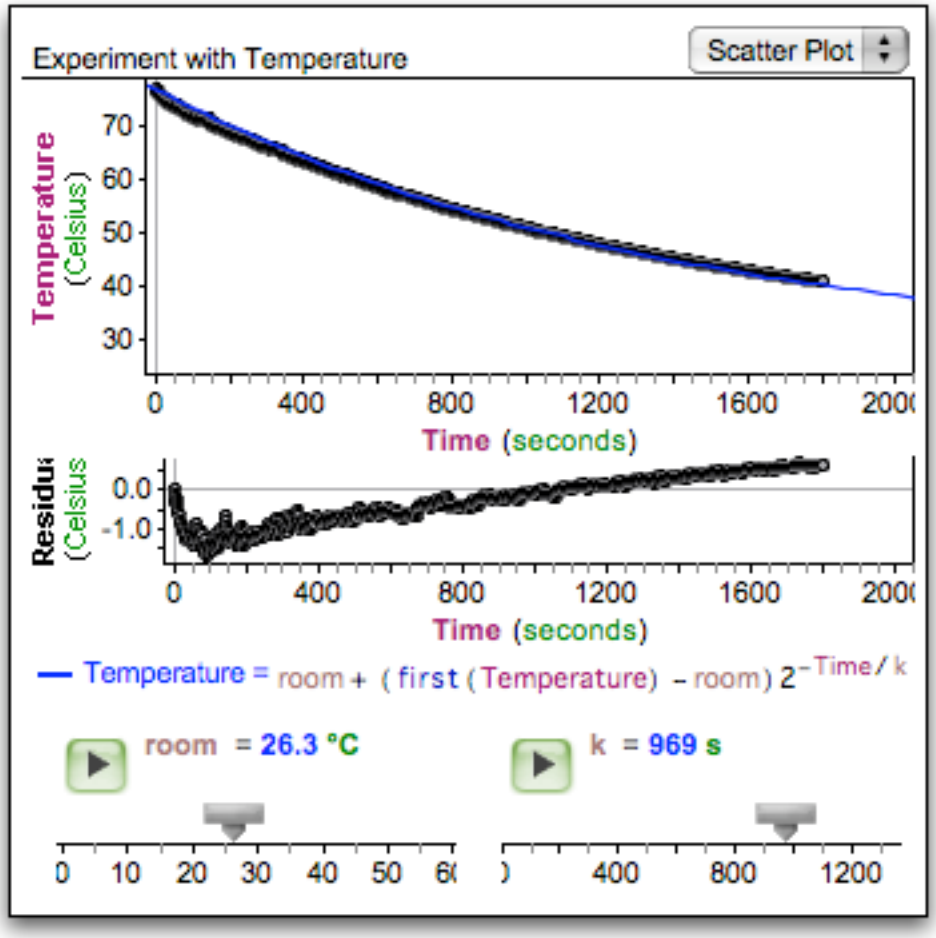

Figure 3: Two sliders control parameters of an exponential model of cooling. The residuals show that something else is going on, at least at the start.

Even with the inclusion of room temperature our model included a residual with a clear pattern, as shown in Figure 3. Had this been a lab in which we gather the data and take it home to embed in a lab report, we would have had to make do with what we had. But we speculated that the problem lay in the thick walls of the coffee cup. Repeating the experiment with a thin-walled glass, we got a much better fit.

4.3 The Importance of the Control Panel

From a software design perspective it worked out nicely that the control panels for working on data from an experiment and for investigating census microdata bear the same relationship to their respective data collections. This functional similarity is evident in Figure 4.

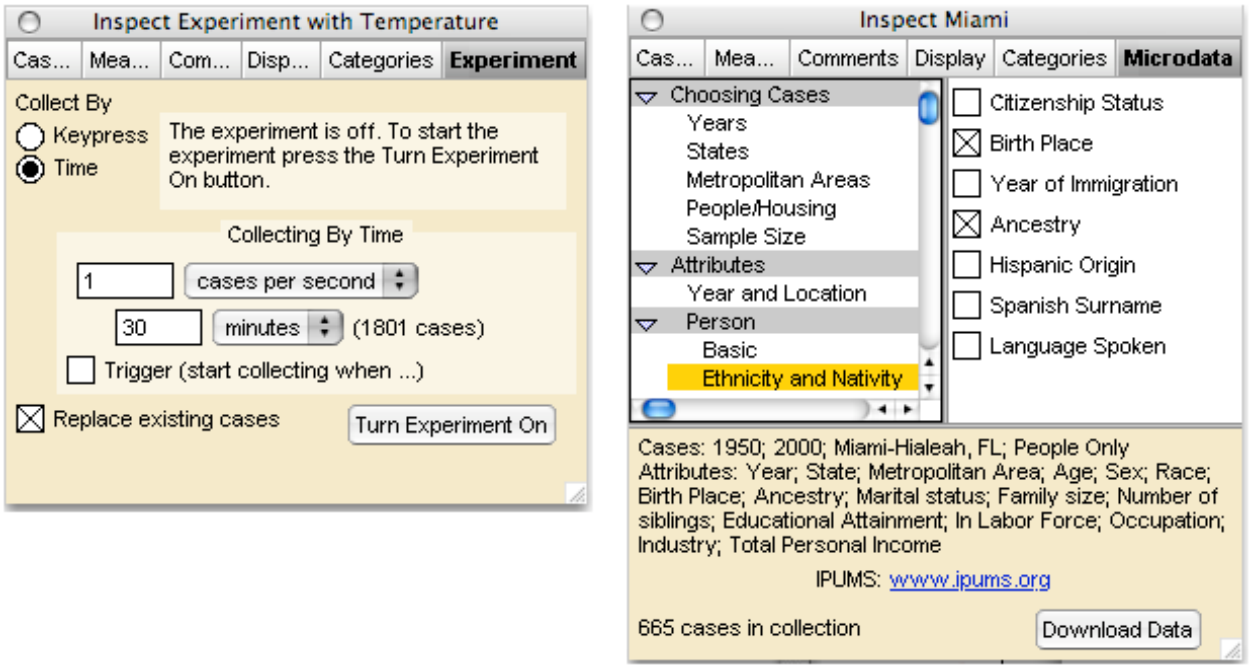

Figure 4: The experiment panel on the left and the census microdata panel on the right. The panels are similar in that they both provide an interface with which the user can control the data acquisition.

The functional similarity of collection control panels hints at an important learning goal: At what stage do learners acknowledge that data from a probe is fundamentally the same as census microdata imported from IPUMS; i.e. that each consists of cases (people or measurements) with attributes (age or temperature)? We want learners to leverage this "universality" of data with the further realization that techniques used to analyze one data set may well be applicable to another.

Data from experiments is just one example of data that arise in a classroom context. At least as common and compelling are those situations in which students have data that need to be collated for analysis.

5. COLLATING CLASSROOM DATA

When we looked at what it takes to gather data in a classroom, we thought, "There has *got* to be a better way!" As part of the census microdata project, we developed a "survey" capability. It works by allowing users to turn a collection into a survey, or data entry form, that is posted on a web server provided by Key Curriculum Press. Each attribute in the collection corresponds to a question in the posted survey. By disseminating the URL of the posted survey, anyone with a web browser and knowledge of a username and password can fill out the survey. Then, when the

survey's owner so desires, users can view the accumulated data and use drag-and-drop to import the data into Fathom for analysis. The movie "Movie 4 Classroom Survey.mov" shows that basic idea. (See the Fathom Surveys web site http://keyonline.keypress.com/public/cdg for more information on how this works.)

Many kinds of data can be gathered using this survey capability, among them data that arise when students take measurements, conduct experiments, enter information found on the web, and administer surveys. We discovered that even classroom discussions can benefit from the capabilities of Fathom Surveys when we collected brainstormed ideas and discussed them as they appeared projected in front of the class.

The power of the collaborative data gathering made possible by this capability comes from a loosening of the constraints of space and time. Teachers don't have to take the time to distribute data entry forms into the classroom space. Students don't have to be in the same place to enter the data, and they can enter it whenever they wish. Teachers invest no time in collating the data. Students need not wait for each other or for their teacher to begin analyzing the data. New data, as it arrives, flows seamlessly into the existing analyses, taking no additional time. A teacher can set up collaborations with classrooms anywhere in the world in which students share information about themselves, their environs, or their opinions. And teachers can do this on their own, expending little effort, and without having to wait on any central agency to coordinate collaboration.

6. DESIGN GOALS

Each of the four software capabilities discussed in the previous sections serves the overall goal of increasing the amount and quality of data that get into the hands of students and instructors. Software capabilities by themselves are insufficient. In Table 1 we list the main software design goal for each capability and a parallel goal for design of curriculum materials and instructional methodology.

Software capability	Software design goal	Pedagogy design goal
Drag and drop URL to import data visible on a web page	Eliminate the obstacles that lie between learners and data they encounter.	Create multiple opportunities and incentives that encourage students to find and explore data on their own.
Census microdata import	Provide painless, random sampling access to huge, rich, relevant data sets.	Foster inquisitive, skeptical, conjecture-driven data analysis with appropriate use of statistical inference.
Realtime data import	Arrange that experiment data arrive in real time without disturbing existing models. Make repetition of an experiment easy.	Encourage model building ahead of and during data gathering so that differences between model and data are milked for their physical and mathematical insights.
Creation of online data entry forms for collation of classroom data	Streamline every step of the process of working with collaboratively gathered data.	Deepen and enrich activities that rely on data by making use of collaboratively gathered data.

Table 1: For each of the four data import capabilities, a relevant goal for software design and for design of curriculum materials that incorporate the software are listed.

7. QUESTIONS

The four data acquisition capabilities described here are examples of ways that technological innovation lowers barriers, in this case barriers to use of varied, rich, relevant and/or realtime data. Questions abound. In what teaching situations is use of these capabilities actually practical? What differences can be observed in learning and attitudinal outcomes? What are the curricular implications for statistics education and, more broadly, the teaching of other data-rich subject areas? What can statistics education researchers tell us about students' understanding of data and the impact of working with the kinds of data described in this article on that understanding? If use of these technologies proves practical and beneficial, what can be done to speed their introduction in classrooms?

8. WHAT NEXT?

As software developers we cannot easily contribute to the work required to answer the questions posed in the preceding section. But, we are interested in further lowering the barriers to classroom data acquisition through generalizing to new data *structures*. All of the data we have encountered for students taking introductory statistics or in K-12 classes has the usual "flat" structure of rows (cases) and columns (attributes). But the world is not modeled well as flat. There are hierarchical structures such as census data in which households contain people, tree structures such as genealogies of families, relational structures such as those used in a manufacturing plant to track production, and networks such as those used to model a language's grammar. Science and industry are already in desperate need of people who can work with these

kinds of data to maintain their integrity and make decisions and discovery with them. Our students need experience with data exploration and modeling that goes beyond the flat structures that sufficed for most of the twentieth century. But this experience won't come without work on the part of *both* software developers to provide tools *and* educators to rethink the role of data literacy in students' education. Coming from a data-rich, multidisciplinary background, statistics educators have a critical role in this process.

9. REFERENCES

Erickson, T., (2005), "Stealing from Physics: Modeling with Mathematical Functions in Data-Rich Contexts." *Teaching Mathematics and its Applications*, Oxford University Press, (v 25) 23–32.

Erickson, T. (2000), *Data in Depth—Exploring Mathematics with Fathom*, Key Curriculum Press, Emeryville, CA.

Finzer, W. (2007), *Fathom™ Dynamic Data™ Software*, Key Curriculum Press, Emeryville, CA.