Statistics and the Modern Student

Robert Gould

University of California, Los Angeles

rgould@stat.ucla.edu

Summary

The introductory statistics course has traditionally targeted consumers of statistics with the intent of producing a citizenry capable of a critical analysis of basic published statistics. More recently, statistics educators have attempted to center the intro course on real data, in part to motivate students and in part to create a more relevant course. The success of this approach is predicated on providing data that the students see as real and relevant. Modern students, however, have a different view of data than did students of 10 or even 5 years ago. Modern statistics courses must adjust to the fact that students' first exposure to data occurs outside the academy.

Les cours d'initiation à la statistique ont traditionellement visé les consomammateurs de la statistique avec l'intention de produire une population capable de faire une analyse critique des statistiques élémentaires publiées. Plus récemment, les professeurs de la statistique ont tenté d'orienter les cours d'initiation vers des données réelles, afin de motiver les élèves d'un part, et de créer un cours plus pertinent d'autre part. Le succès de cette approche repose sur une provision de données que les étudiants considèrent comme réels et pertinents. Cependent, les étudiants modernes ont une vision des données qui est différente de celle qu'ont eu les élèves d'il y a 10 ou même 5 ans. Les cours modernes de statistique doivent s'adapter au fait que la première rencontre des élèves aux données a lieu en dehors de l'académie.

Key words: Education; data technology; statistical literacy; technological literacy; data.

> *The time may not be very remote when it will be understood that for a complete initiation as an efficient citizen of one of the new great complex world wide states that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write.*
>
> *--H.G. Wells (1903, pg 204)*

1.0 History and Background of Intro Statistics

Perhaps because Wells was best known for his science fiction, the above quote, written in 1904, is often attributed an oracular quality. But Francis Galton and Florence Nightingale were alive in 1904, and so perhaps the quote is better understood as a reflection of the excitement that grew as many realized that statistics was no longer just for astronomers, but useful in the social sciences, medical science, biology, criminology,

and weather forecasting. (Guinness Brewery felt this excitement; they had hired a soon-to-be-famous statistician six years earlier.)

Wells advocated a near universal statistical education. While his quote doesn't mention a specific curriculum, the goal is clear: to produce good citizens. Hotelling might not have been as concerned about training good citizens, but motivated by what he saw as substandard statistics education taken over by non-statisticians, he had similar concerns that provoked him to ask a series of questions concerning education in his landmark 1940 paper (Hotelling, 1940). Among other questions, he asked who should be taught statistics, and by whom should they be taught? Hotelling provided answers: college students should be taught statistics and should be taught by mathematical statisticians. Fisher, according to Bibby (1986), apparently reached the same conclusion in the Preface to *Statistical Methods for Research Scientists*.[1] There was some dissent, as I will soon discuss, but Hotelling's argument was persuasive and became the mainstream point of view (although perhaps only among mathematicians and statisticians).

Teaching students to both think *and* compute, however, proved difficult. Statistics relied heavily on mathematics, in part because data were expensive and calculations were tedious and slow, and so approximate methods were required. Cobb (2007) refers to this as the "tyranny of the computable", and one aspect of this tyranny was that, in order to fully understand statistics, one had to understand it through mathematics. Alas, student preparation and understanding in mathematics did not improve to the level where universal education in mathematical statistics was possible, and so many (most) introductory statistics books continued to focus on methods somewhat (or greatly) divorced from their mathematical roots, perpetuating the sins about which motivated Hotelling's questions.

By the 1980's it was clear to a growing number of statisticians and educators that attempts to teach statistics to a general audience were failing miserably. Far from achieving the goal of preparing a citizenry for thinking and computing with data, educators seemed to be driving students away from the topic. The curriculum had become focused on teaching procedures and rote memory. Homework problems were dull, tedious, and, by using idealized contexts, failed to teach students the usefulness or applicability of statistics to real world problems.

More recently, the problem was typified in Matthew, et. al. (2005) who found that several weeks after completing a "traditional" college level introductory statistics course, the best students couldn't explain fundamental concepts. ("I'm not really sure how to

---

[1] Bibby quotes from the preface of the 1939 edition: "...responsibility for the teaching of statistical methods in our universities must be entrusted to highly trained mathematicians...(who) have had sufficiently prolonged experience of practical research." The 1958 Preface does not include this sentence, but instead "Too often, however, their [mathematicians who write statistics books] experience has not included the training and mental discipline of the natural sciences...."

explain what [the mean] would mean. I just know how to do the formula," as said by one student to the researchers.)

Deming had pointed a way out of this predicament in his rejoinder to Hotelling. Statistics might *use* math, but it was *about* data. "Above all, a statistician must be a scientist...Statisticians must be trained to do more than to feed numbers into the mill to grind out probabilities; they must look carefully at the data, and take into account the conditions under which each observation arises." (Deming, 1940).

In the 1990's, a growing chorus of statisticians and educators echoed Deming. David Moore (1997) along with others (Hogg, 1991; Joiner, 1988; Cobb, 1991 and 1993; Hunter, 1981; Snee, 1993; Singer &Willett, 1990; Wild, 1994) urged educators to put data at the center of the statistics curriculum. A statistics education should focus on developing conceptual understanding rather than rote procedures (see GAISE college report, 2005). "Use real data" is the second of six recommendations of the Guidelines for Assessment and Instruction in Statistics Education (GAISE), endorsed by the American Statistical Association (Garfield, et.al. 2005). This seemingly simple guideline raises two questions: What are data? And what makes them "real"?

What are data? Cobb and Moore (1997) define data as "numbers with a context". Presumably, providing a context makes the problem more realistic, and forces students and instructors to think about the validity and applicability of their solutions. Numbers have long been included in the curriculum, of course. Singer and Willet (1990) pointed to one particularly egregious, but not atypical, example:

> *X: 2, 2, 1, 1, 3, 4, 5, 5, 7, 6, 4, 3, 6, 6, 8, 9, 10, 9, 4, 4*
> *Y: 2, 1, 1, 1, 5, 4, 7, 6, 7, 8, 3, 3, 6, 6, 10 , 9, 6, 6, 9, 10*
> *Calculate: (a) the means, sums of squares and cross products, standard deviations, and the correlation between X and Y; (b) the regression of Y on X (c) Regression and residual sums of squares; (d) the F ratio for the test of significance of the regression of Y on X.*

Compare this to an excerpt from a problem in a modern "reform" introductory book (De Veaux, et. al., 2009, p. 745.):

> *The European School Study Project on Alcohol and Other Drugs, published in 1995, investigated the use of marijuana and other drugs. Data from 11 countries are summarized in the following scatterplot and regression analysis. They show the association between the percentage of a country's ninth graders who report having smoked marijuana and who have used other drugs...Explain in words and numbers what the regression says. State the hypothesis about the slope (both numerically and in words) that describes how use of marijuana is associated with other drugs....Do these results indicate that marijuana use leads to the use of harder drugs? Explain.*

The data used in this problem (although not provided to students in "raw") form, have a history, and are associated with real people and real places and were collected to solve answer a specific and pressing question. These data are "real". It does not hurt that the central issue is one that might be of interest to college students.

Bibby (2003) challenged educators to teach using "*real* life and *real* perception -- *as perceived by the student*". While the data in the above example, and in most reform textbooks, are undeniably real, I question whether they are "real" as perceived by the student, in part because there is a certain level of abstractness to data collected in a formal, professional setting. The data most of us use to teach our students were collected by professionals, usually in the context of a scientific investigation or some other formal procedure (for example, a legal proceeding or business investigation or political poll). Few students, however, have first-hand experience with such data or such contexts. Although the research studies might be interesting and compelling to students, they are by necessity abstract since the data were collected by strangers working in a context about which students have only a vague understanding.

However, today's statistics educators face an exciting challenge. Today's students, perhaps for the first time in history, do not receive their first exposure to the concept of data in a statistics class. Modern students have direct, first-hand experience with data, albeit a type of data that is often neglected in the statistics curriculum.

2.0 Data and the Modern Student

As Rubin (2007) points out, new technology creates new data types and, as a reviewer added, more complex data structures. The growth of "data technology" has created new data types that students encounter as part of their social and personal lives. Students now enter our introductory statistics class with experience in data. Even mainstream culture has caught on. *Wired Magazine* had an entire issue devoted to data (Anderson, 2009) that included techniques for capturing "personal data" to monitor aspects of every-day life. *The Economist* devoted an issue to the "data deluge", and discussed a variety of contexts in which great streams of data are produced (Feb. 25, 2010). In this section I offer an incomplete list of some data your students might very well already have experience with.

Music

As I am writing this draft, I am listening to a band called the *Yeah Yeah Yeahs* on Pandora Radio (www.pandora.com). Unlike traditional radio, Pandora is a website that tailors the music it plays to my own personal taste. After asking me to name a band I wanted to listen to, Pandora chose to play the *Yeah Yeah Yeahs* because that band meets certain characteristics that it determined I like: "electric rock instrumentation, punk influences, major key tonality, a vocal-centric aesthetic and electric rhythm guitars." Pandora also provides a list of similar artists, and also a list of (human) listeners who enjoy my (heretofore unknown) penchant for electric rhythm guitars.

If I want to know more about the *Yeah Yeah Yeahs*, a visit to Tuneglue at http://audiomap.tuneglue.net will help. This dynamic, interactive site will find other

bands similar to the Yeah Yeah Yeahs, and will show me how closely related (in some vaguely defined way) these bands are to the Yeah Yeah Yeahs. Although Tuneglue does not explain specifically how relationships between bands are determined, the website explains that this structure is based on data, and gives the source for these data.

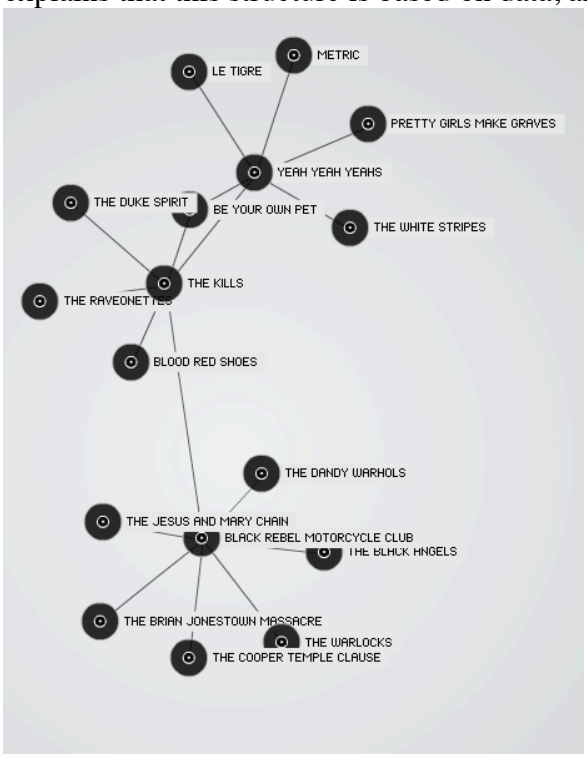

Figure 1. Screen snapshot of Tuneglue showing the relationship tree between a set of rock bands. The tree was built around the band *Yeah Yeah Yeahs*, near the top.

The Shape of Song project (www.turbulence.org/Works/song, viewed May 25, 2010) analyzes individual songs at an even finer level: that of the note. Using midi files, Turbulence produces a graphic that reveals the structure of a composition, here Chopin's Mazurka in f# minor.

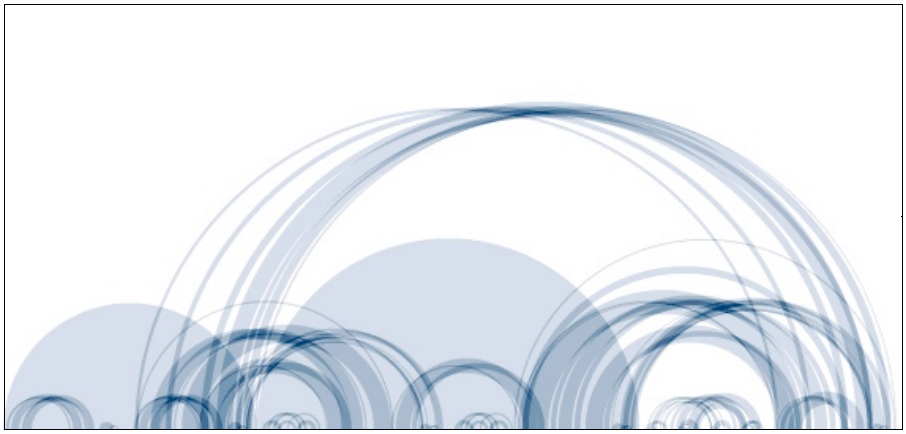

Figure 2. Screenshot from turbulence.org showing a representation of the structure of Chopin's Mazurka in f# minor.

Chopin is not a close relation of the *Yeah Yeah Yeahs*.

The website Last.fm "listens" to the music I play on my computer and mp3 player and collects data about it. These data are used to recommend other artists I might like and to put me in touch with people who listen to what I listen to. The site lastgraph3 (http://lastgraph3.aeracode.org/) allows me to export this data to my harddrive, and creates several different graphic displays. This shows that my music life is not all punk-inspired, rhythmic guitars. ("Ma Ax" are a cellist and a pianist, respectively, and not a hardcore metal band.) In this display, colors represent my level of interest in an artist. "The more saturated the color, the more interest the user has in that musician" (Byron & Wattenberg, 2008; www.leebyron.com/what/lastfm, viewed May 25, 2010.) The width represents how often I listened to music on my computer. (I was out of town that July and again early August.)

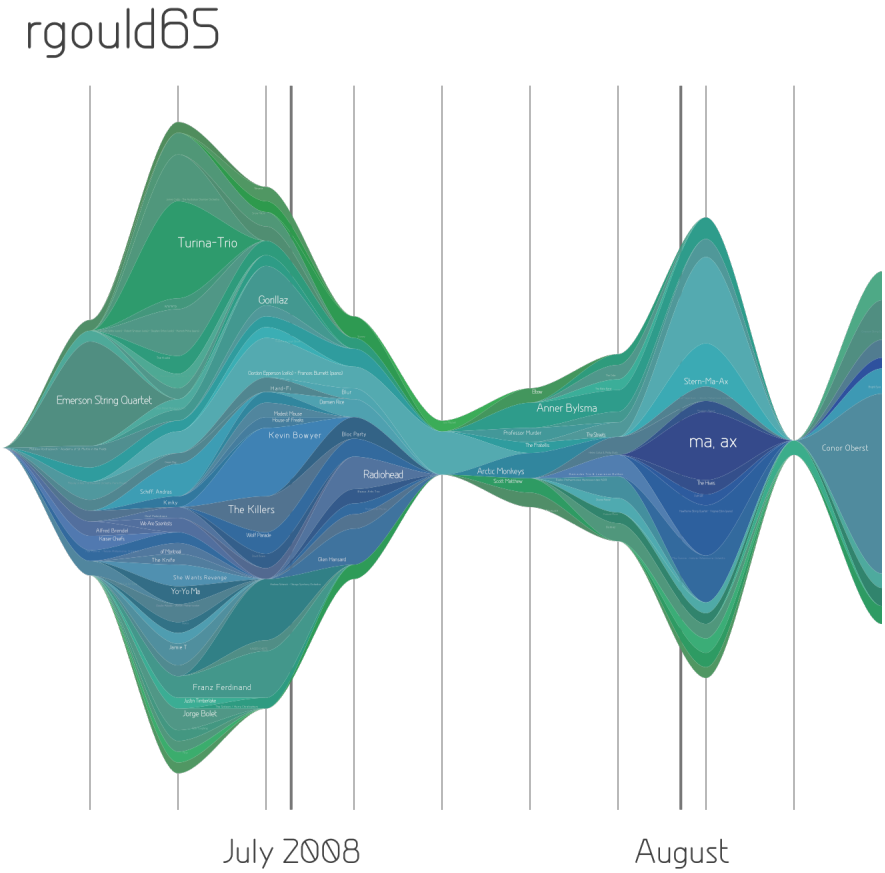

Figure 3. Timeline representing the artists I have played on my computer and iPod. The data are collected by Last.Fm and the graphic produced by lastgraph3.aeroacode.org.

One ubiquitous example of music-as-data is the iPod which, by treating songs as data, breaks down the sacred construct of the album and allows songs to be "shuffled". Users can also view statistics regarding their music, such as how often or how recently songs are played, and can sort and re-sort their catalog.

Social Life

The internet offers a variety of socializing opportunities. One of the more popular is Facebook (www.facebook.com) which, according to it's "Press Room" currently has more than 90 million active users, at least some of whom are enrolled in your class. (Quite probably, a few are updating their Facebook profiles during your class.)

Nexus is an application (now defunct) that can be added to a Facebook page to visualize one's social network. Here is a visualization of the web site of the UCLA statistics department's student group, alongside my own, somewhat disconnected social network.

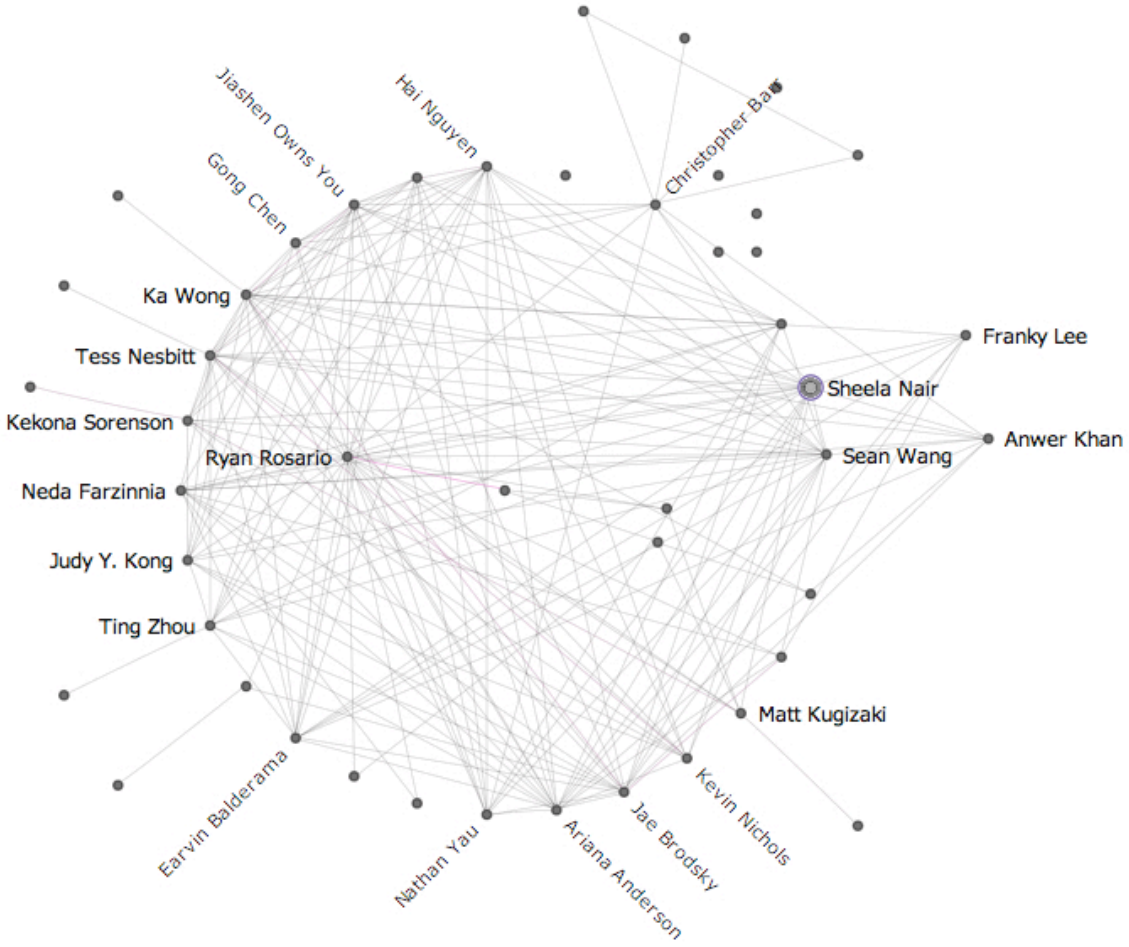

Figure 4a. Relation tree for the UCLA Statistics Facebook page. Lines connect students who are Facebook "friends". Clicking on one node displays the names of all connected nodes. Produced by the Facebook application Nexus.

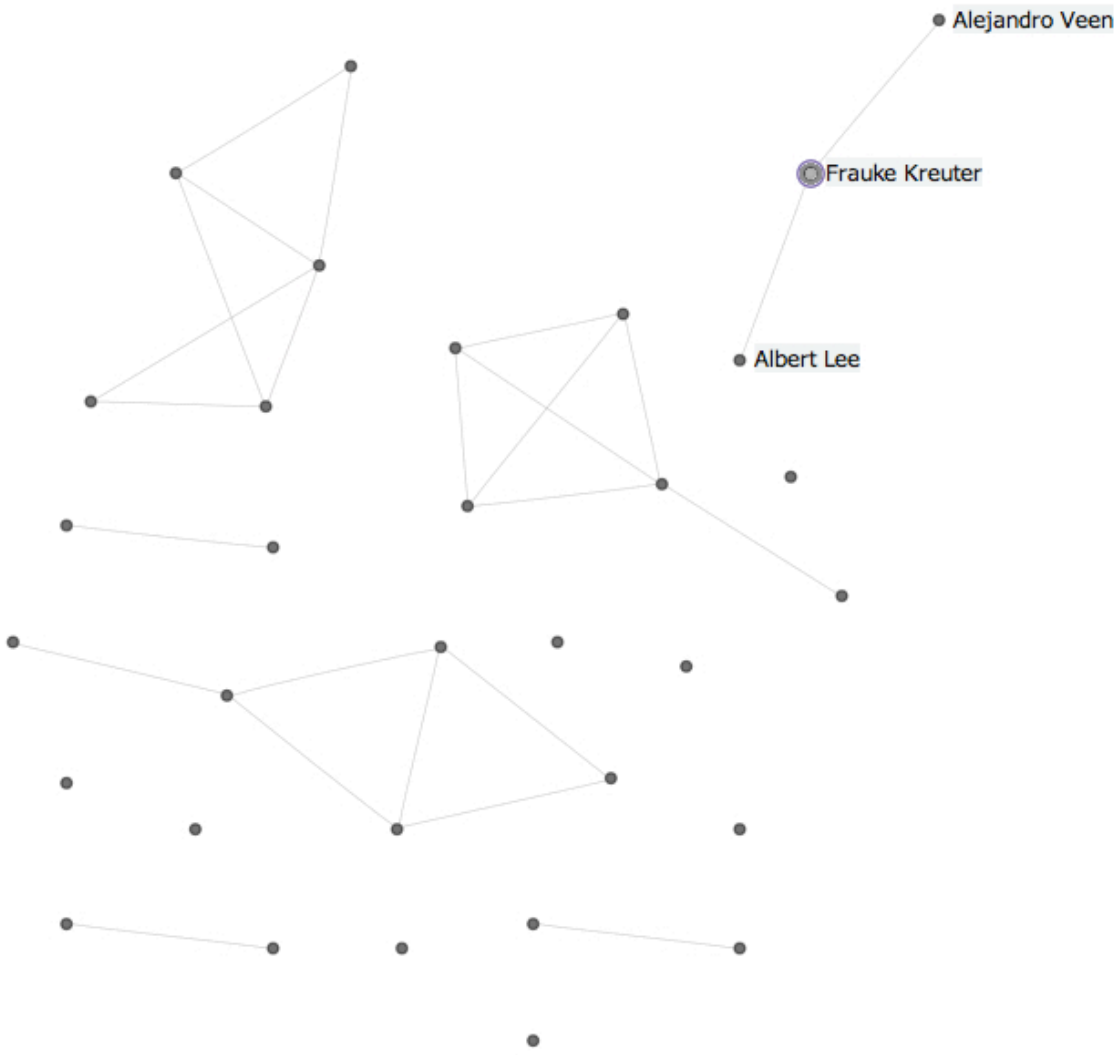

Figure 4b. The author's own Facebook friend tree, which is considerably more sparse than Figure 4a. The three names all belong to statisticians, and no statisticians appear anywhere else in the tree (at this moment).

Twitter is a service that invites users to broadcast frequent, brief messages via their phones or computers to friends (or strangers). Twitter claims that this will help you "stay connected" with your family and friends. According to the twitterfacts blog (twitterfacts.blogspot.com, viewed August 12, 2009), there were roughly three million twitter users by late 2008. There is a small industry involved in visualizing the "twitterverse". Statistician Nathan Yau's blog flowingdata.com lists 17 ways of visualizing the twitter universe, including a very personal view of Nathan's "tweet" habits, in which it becomes clear that Wednesdays are not productive work days for Nathan.

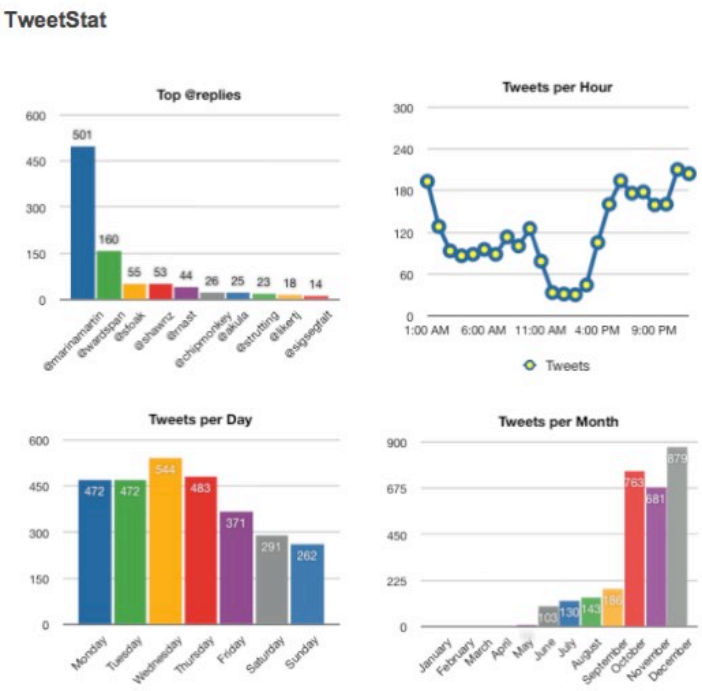

Figure 5. Twitter activity of Nathan Yau.

Twitter played an important enough role in the recent unrest in Iran that Time magazine called Twitter "The Medium of the Movement" (Grossman, 2009). The blackout of news coverage created a market for analyzing tweets to determine what was happening on the ground. Iranian postings included the characters "#IranElection" in order to make it easier for anyone in the world to "follow" the events. Twitter thus served as a sort of live news feed, accessible and analyzable to anyone with a computer and an internet connection.

Maps

Maps have been used to display and organize data at least since John Snow's famous map of cholera deaths in 1859 London. Applications such as Google Maps, however, have re-created maps as data storage and processing devices. The maps can themselves be treated as data and provide meta data that can be read and analyzed (Rubin, 2007). Geographical Information Systems (GIS) are not new, but Google has made highly interactive GIS available to a very large public, and this *is* new. One can imagine John Snow merely having to rely on family members entering their loved ones' state of health into a computer to produce a daily update of his map. The result might look something like Who Is Sick? (Figure 6). This somewhat awkward display invites anyone to enter their symptoms and address. Many, many more "mashups" can be found online. Perhaps the most frequently encountered (at least by Los Angeles residents) are maps that show prices of gasoline (for example, see www.gasbuddy.com), but it is also not hard to find maps that combine people's vacation photos, opinions about restaurants, and videos of political candidates who have visited a particular spot.

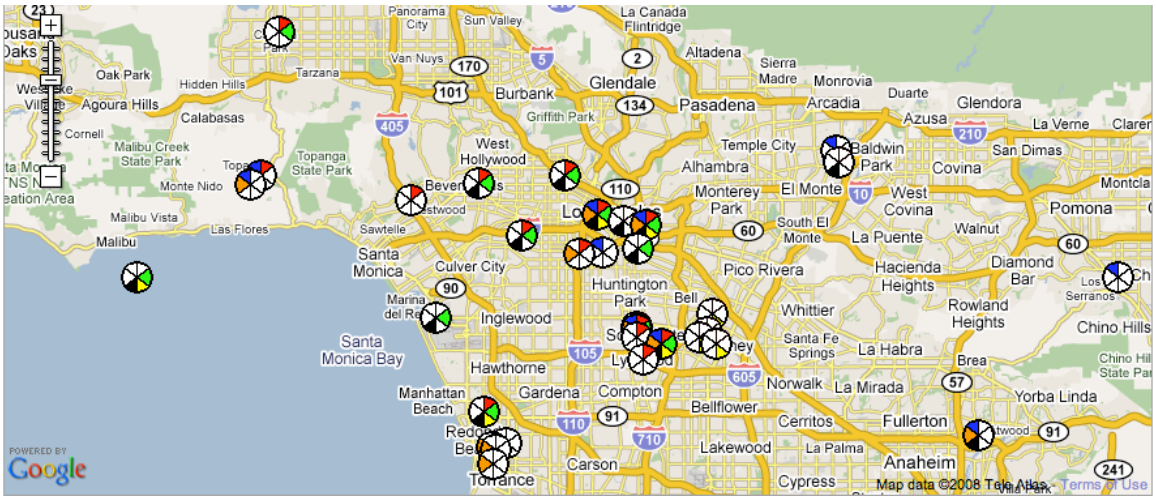

Figure 6. Screenshot from whoissick.com, which shows the locations and types of self-reported illnesses in the Los Angeles area.

In private conversations with high school statistics teachers, many have expressed surprise that maps can be used statistically. Hengl (2010) provides instruction for a suite of free, open-source software that could bring mapping to the classroom, although the learning curve is perhaps steep. Hengl devotes several informative tutorials to demonstrate the statistical qualities of mapping. While these tutorials are perhaps beyond the statistical preparation of many high school teachers, teachers should still be aware of the existence of geostatistical mapping.

Exercise

Recreation is now converted into a stream of data. A variety of devices exist that convert your exertions into analyzable output. Figure 7 shows sample results from the Garmin Training Center, software that processes data collected from a Garmin GPS device and intended to improve your workout routine. There are a number of other devices on the market that collect similar data, including cell phone "apps".

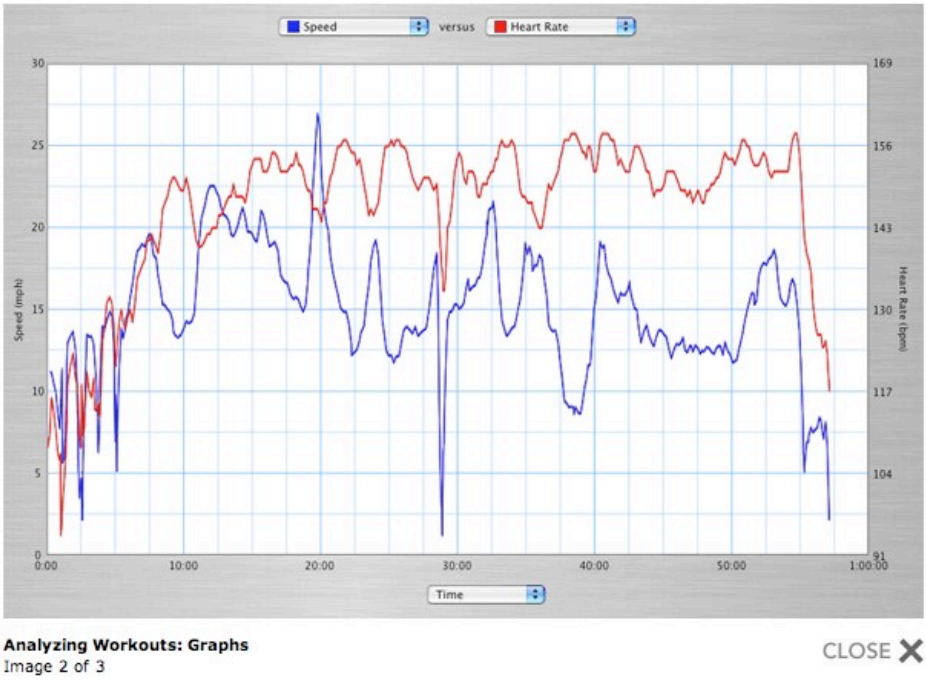

Figure 7. Screenshot of Garmin Training Center website, illustrating the data displays provided by their GPS device.

Everyday Personal Data Devices

Most students now create and carry with them, at all times, their own fairly rich databases. They carry them on small, portable devices that organize data, record data, and report some basic summary statistics. This device, of course, is the cell phone. Even a basic phone stores relatively large datasets that include names and contact information, and includes software that organizes and displays this information in, one hopes, a productive fashion. Basic phones also allow their users to access usage statistics, such as the number of minutes the phone was used and the number of text messages sent. The so-called "smart phones" do all this and much more.

So-called "smart phones", such as the Blackberry, iPhone or Android, take this to whole new levels. These phones store several gigabytes of data, and these data include photographs, music files, calendars, text files. These phones include GPS devices and internet access. Almost all of the data examples in this section can be viewed on a phone. Smart phones encourage users to produce more data that are added to "social" sites. For example, one can post a note and "geo-tag", so that others who pass through that neighborhood can see, on their phones, that you were there too, and can learn your opinions about that location, or view the pictures you took. The "app" store for the Apple iPhone lists applications that display the current locations of your friends (loopt), allow you to listen to music via Pandora, track the foods you eat and the calories you expend (iTracker), track the growth of your infant (Growth Chart), locate the 10 least expensive gas stations near your current location (iGas), analyze money spent on your car (Car Stat), and even design and carry out surveys (surveygizmo.com). You can also post and read Twitter tweets, and keep up with your Facebook data.

The Center for Embedded Networked Sensing (CENS) project uses cell phones to collect geographic and environmental data from anyone who wishes to contribute. (The Personal Environmental Impact Report). Anyone who wants "to be more conscious of their personal impact and exposure on the environment" is invited to join and contribute data, which are then summarized and displayed in a map format on the web. (pier.cens.ucla.edu).

3.0 Challenges to Educators

The data that students live with are complexly structured, ubiquitous, cheap, constantly updated and so transportable that they literally fit in a pocket. The data of most textbooks are numerical, fit easily into a flat ascii text file, static, are (usually) small (in terms of cases and variables), and are abstract (from many students' view points). This creates a number of challenges for educators. Most obviously, the data around which we are hoping to center our curriculum are different in fundamental ways from what we were used to. Less obviously, this change has created the need to educate with the goal of creating a different type of student. This new type of student requires a new curriculum.

First Challenge: Redefine Data

Are data simply numbers? Perhaps at some level. In fact, at some level only two numbers: 1 and 0. But it is not usually helpful to think about data at such a granular level. We want to treat music as music, and images as images. When we take a photo with a digital camera, the resulting photo is not an image of light on film, but consists of data. One could see this file of digits if one wanted to, but one really shouldn't want to. Instead, we have photo software that allows us to interact with these data as if they were photos, to change the brightness, the sharpness, the colors. So perhaps it is time to generalize the definition of data beyond "numbers".

A new definition of data should be inclusive, rather than exclusive. It should acknowledge that data are created by both need and accessibility. Data arise from activities, objects, experiences, and relations. Data can be shared and stored and transmitted. Data include a context, but data also create context, as when emails (data) are studied to produce more data to recreate institutional social structures or make a better spam filter.

Second Challenge: Create Citizen Statisticians

Who should be taught statistics? In some countries, universal statistics education has arrived, although quality and quantity might vary. (Consider Census At School, a product of the Royal Statistical Society Centre for Statistical Education that now serves students in five countries. For example, roughly 700,000 students in South Africa use the service to collect and distribute data about themselves.) However, all students are not taught with the same final objectives in mind.

Many college statistics courses are designed, sometimes explicitly, for either "consumers" or "producers" of statistics. (Here, I use "statistics" in its technical sense as a summary of data.) According to this paradigm, producers will, at some point in their intended career, perform statistical analyses on data that either they or colleagues have collected. Producers are majoring in biology, atmospheric science, social science, psychology, engineering. Consumers, on the other hand, will read about statistics in the news media, and will need to understand basic statistical concepts (such as median vs. mean) in order to make household decisions and in some cases workplace decisions. The emphasis for consumers is that they are consumers of statistical summaries. Consumers are majoring in literature, law, and mathematics (ironically enough).

Researchers in the statistics education literature have worked to define and understand the concepts of "statistical literacy" and "statistical thinking". Although these are not mutually exclusive concepts, never the less they seem compatible with the consumer/producer paradigm. Garfield and Ben-zvi (2007) identify several definitions of statistical literacy in the literature, and argue that being statistically literate means being familiar with the basic tools and language that statisticians use. Gal (2002) defines statistical literacy and argues for the importance of what he sees as an overlooked component to education. Utts (2003) presents seven important statistical topics required for citizens to be able to understand media reports of scientific and medical studies. Rumsey (2002) breaks statistical literacy into two components: statistical competence and statistical citizenship. The Chance course, designed by Laurie Snell (2003) and others, is designed explicitly around probability and statistics related readings culled from the daily newspapers.

Statistical literacy seems aimed squarely at Consumers, who must learn to "read" statistics. Producers, on the other hand, are better served in a statistics course that teach "statistical thinking". Statistical thinking means thinking like a statistician (Wild & Pfannkuch, 1999; Garfield & Ben-zvi, 2007) and arguably, statistical thinking is a more comprehensive competency than statistical literacy. A recent entry in this discussion defined statistical thinking as thinking that uses "probabilistic descriptions of variability in (1) inductive reasoning and (2) analysis of procedures for data collection, prediction, and scientific inference." (Brown & Kass, 2009). Brown & Kass had producers in mind, recommending that an emphasis in statistical thinking in undergraduate courses will attract and better prepare students to Statistics (among other benefits, of course.)

However, the producer/consumer dichotomy appears false, or at least fades in importance, when we realize that it is based on thinking of the student in terms of consuming and producing *statistics*. If instead, we think of the student as consuming and producing *data*, then it becomes clear that we must educate all students to be both consumers and producers. Perhaps the model for a future student should be that of a "Citizen Statistician".

The similar term "citizen scientist" has been in use long enough in science education circles that it has its own wikipedia entry. Citizen scientists are people who volunteer to assist in scientific investigations through "observation, measurement, or

analysis" (http://en.wikipedia.org/wiki/Citizen_science, viewed May 23, 2010). Citizen Statisticians also participate in organized data gathering and analysis activities. For example, in Census at Schools students contribute data about themselves and analyze the internationally pooled data. The CENS project mentioned above is another example, as is Many Eyes, the IBM project to encourage citizens to share data visualizations (http://manyeyes.alphaworks.ibm.com/manyeyes/, viewed May 26, 2010). The on-line statistical software package StatCrunch (www.statcrunch.com, viewed May 25, 2010) has features that allow users from around the world to share data, graphics, and analyses (West, 2009).

But Citizen Statistics is not necessarily organized around formal projects. Missing from statistics education is an understanding of Citizen Statistics as an activity that continues beyond the classroom and beyond the school years. Citizens can act as statisticians with or without formal training by accessing data and performing analyses, however crude and reaching conclusions, however flawed. Educators must not merely help citizens interpret statistics, we must help them analyze data. A good Citizen Statistician should know data when she sees it, (and she should see it everywhere). She should recognize when she is viewing actual data that can be brought to her computer and analyzed to answer questions she has about the world, and when she is viewing others' summaries of data. In the latter case, she should think critically and see the summary in the proper substantive context, but also in the context of other datasets or databases.

Third Challenge: Teach Technology

Modern statistical practice is inseparable from technology. Just as Medusa could be viewed only through reflection in a mirror, some data can only be handled through a computer. Indeed, many of the new data types are inaccessible without strong programming skills. For example, in theory all of the data are freely available online for you to discover which factors determine the cost of a used car in your neighborhood. But in practice, you will have to teach yourself Python, Beautiful Soup, and Mechanize (or some equivalent package or set of packages) to "scrape" the data together.

Given this reliance on the computer, it is surprising that we do not spend more time teaching the elements of statistical computing. We have thought of statistical software and computation as a hurdle on the way to achieving statistical literacy, when in fact, it is fundamental to it. We can no longer be complacement and assume students will "pick up" the skills they need to negotiate complex data. It is too much to expect introductory statistics students to become expert programmers, but what understanding of the computer and its role in statistics should they know? Nolan, Temple Lang, and Hansen organized a series of workshops designed to help statistics faculty learn to teach computation (Nolan, et. al, 2007) They urge faculty to teach students not just programming language syntax, but also "higher-level concepts of computational thinking that enable students to approach computational tasks intelligently." (Nolan and Temple Lang, 2009).

Nolan and Temple Lang's goals are probably not achievable by all students at all levels, but to be a Citizen Statistician requires some level of technological literacy.

Selber (2004) identifies three components of technological literacy that provide a useful framework for understanding how we want our students to function. Functional Literacy is the literacy that results from knowing how to use technology. Most introductory courses based on real data teach some level of functional literacy. Critical Literacy is the ability to think of and treat technology as a "cultural artifact"; to see technology as a product of a culture and thus something that impacts our society. Rhetorical literates are people who are capable of producing technology.

Applying these general components to educating Citizen Statisticians leads me to believe we need to teach more critical literacy and some level of rhetorical literacy. Critical literacy can be particularly important to introductory statistics students, who should understand, at some level, issues of data privacy and data piracy. Critical literacy will help students understand the limitations and strengths of the technology they use to analyze data. However, rhetorical literacy is urgently needed because otherwise, these new data types will remain inaccessible. Rhetorical literacy might be beyond the scope of an introductory curriculum,

Hansen, et. al, (2010) have designed a curriculum for six public high schools in Los Angeles, California, that might serve as a template for future educators of Citizen Statisticians. Students use smart phones to record several variables at a particular place that they find interesting: a photo, a Stress/Chill measurement (-10 indicates the situation "completely stresses you out" and +10 indicates the situation is "chill" (which adults would call "soothing")), text to describe the location, and a categorical variable that indicates something about the situation that led to the taking of the picture. In the process of learning to "tell an interesting story" with their data, students explore issues of data privacy involved in the collection and storage and dissemination of photos, discuss the various ways that data can be used ("advocacy" or "discovery"), struggle with the limits of measurement to capture a situation, and learn descriptive statistics. Interestingly, the team that wrote the curriculum consists of both statisticians, computer scientists, and educators.

4.0 Suggestions for the Statistics Education Community

The primary challenge facing educators is to teach an introductory statistics course so that students leaving the course will leave with a set of practices and attitudes about data that are immediately applicable to their lives. This was not possible in the recent past, when data were expensive and therefore would only be collected in the students professional career or would be produced by other people. But today's students should (a) recognize data when they see it (b) understand how analyzing the data can help them and (c) know how to do so. I think the meaning of "analyze" and "understand" will have to be worked out by individual instructors (and students) through experience and research. But in the meantime, I'd like to propose a partial list of suggested goals for the statistics education community.

Teach about databases. Databases are not a new technology. Even very large databases, and fears about their threat to privacy, have been around for over a decade (Eisenberg, 1996). What is new is the public's easy access to large and surprisingly detailed databases, and the ease with which databases can be combined to learn even more about individuals. Many of the visualizations in Section 2.0 are derived from databases, and students need to understand that the visualizations and summaries they run across are often the tip of a very large database iceberg. Students, and indeed all citizens, should understand just how much can be learned about their private lives even by casual internet browsers. Sweeney (2000) found that using U.S. Census data, 87% of Americans could be uniquely identified based only on zip-code (a 5-digit code used by the postal service to identify a region), gender, and date of birth. Linking a Massachusetts voter registration list with anonymous data (names removed) collected by the Massachusetts Group Insurance Commission, Sweeney was able to uncover sensitive medical information for individuals, including the then-governor of Massachusetts (Sweeney 2002). Understanding how large databases are created and maintained is an important part of understanding the extent to which one has control over one's personal information.

Accessing most large databases in a systematic way is not easy, and probably beyond an introductory course. Still, students should understand that the data perhaps already exist to answer questions about their society and environment, and they should understand the dangers of reaching conclusions with a naive understanding of the context in which the databases were assembled, and the efforts required to assemble the data into a meaningful dataset. Students need to understand this not because they will necessarily do it, but because they need to understand that different statisticians can create very different datasets from the same database, depending on which subsets are selected, how variables are perhaps combined, and how categories are created or merged from existing variables.

Improve access to data. The variety of data types means a variety of approaches are needed so that students can get their hands on data. Even databases designed for public access can be daunting. For example, through its web page (www.fec.gov) the Federal Election Commission in the United States makes available information concerning all campaign contributions (above a certain amount). However, users are restricted to searching for individual names one at a time, or viewing summaries for particular candidates. The New York Times released an API for querying these data so that one could do a proper statistical analysis. However, using the API requires some level of programming skill (for example, knowledge of Python) and knowledge of XML (Extensible Markup Language) or the data interchange format JSON (JavaScript Object Notation). The U.S. federal government created data.gov in May, 2009 to "democratize public sector data and drive innovation", to slightly paraphrase the website (www.data.gov, viewed May 25, 2010). About 60% of the 1,400 records are listed as "csv/txt" files and might be considered accessible to beginning statistics students. Still, many of these alleged files are not actually text files (some have a .exe suffix, for example), and some contain tables and other information that must be removed, before analysis can begin.

It is too much to expect beginning students to learn more than rudimentary skills at accessing on-line data in a single course, but we can help by improving access to the data that students encounter in their every day lives. Fathom, (Finzer *et al*, 2007) for example, provides a nice, but small, suite of data access tools. For example, a URL can be dropped into a Fathom window to access tabled data that include meta-data tags; a random sample of census data can be downloaded for a variety of attributes; "streaming" data from sensors attached to the computer are also easily accessed. In other cases, we should consider whether students need to learn about formats in which data are stored on the internet so that they will be prepared to someday develop the rhetorical literacy necessary to access even more complex data types. For example, perhaps we need to include discussions of XML libraries.

To some extent, improving access to data misses the point. A fundamental concept of statistical literacy is that one must analyze data in the context from which they were collected. For many complex data types, the data acquisition is part of the context, and students need to develop the critical literacy to evaluate how the technology makes decisions about what they are able to access and what they are not.

Develop new teaching tools. Teachers of statistics rely on textbooks, arsenals of datasets, collections of applets and statistical software. These tools are not sufficient for bringing large, complex, and/or dynamic datasets into the classroom. Nolan and Temple Lang (2008) propose dynamic "documents", which contain text, code, and data and can themselves be compiled. Such documents can aid communication between researchers who work with complex data technologies and educators. If done well, the result could be modules that provide students with level-appropriate tools to investigate real and complex data.

Re-examine which fundamental concepts should be taught. Roughly put, the introductory statistics curriculum is based on the concepts of central tendency and variation, which are designed to lead to the t-test (Cobb, 2007). Understanding these ideas is essential for grasping mean-based analytic techniques, such as the t-test and regression. But do these concepts help students understand and analyze social networks? Music? Video? I suspect the answer is "yes", but these concepts need to be re-packaged and presented with the knowledge that they will be used differently and sometimes more subtly than in a t-test.

Integrate Computation into the Curriculum. Some statistics education researchers have convincingly argued that the introductory curriculum needs to be pared down to increase students' comprehension of fundamentals (Garfield & Ben-Zvi, 2007). How, then, to make room for even more? Kaplan (Kaplan, 2007) argues that by adopting an R-based pseudo-code, with care some fundamental ideas of statistics can be communicated more effectively through code than with mathematical equations. At the same time, students learn some computation. This algorithm-centered approach fits nicely with Cobb's recommendation (Cobb, 2007) to center the first course on the "three R's":

Randomize, Repeat, Reject. The three R's approach provides a model for inferential thinking that emphasizes fundamental concepts and teaches some basic computation. Hansen, Nolan, and Temple Lang (2006) describe a data and problem-solving centered curriculum in the context of a summer program for undergraduates. Although a summer program is a specialized setting, this curriculum provides insight into how one might adapt a similar approach for more advanced students during the regular term.

Change the Culture. Statisticians are traditionally hesitant to generate "research" questions. In the traditional setting, a scientist brings a problem to the statistician. Perhaps these two work together to refine the question somewhat, but the statistician's job is to analyze the data and the substantive expert's job to raise the questions. This is not a helpful approach for students who might find themselves with a dataset that contains hundreds of variables. To make such a dataset meaningful, it must be approached inquisitively. To be interesting to students, the questions themselves must be meaningful and interesting. And this means that we must teach students to be inquisitive.

Others have pointed out that this cultural change must take place if statistics as a discipline will be able to grow and thrive. Nolan & Temple Lang (2009) explain why the culture of statistics should be changed to include "dreaming big"-- changing the world, attacking very hard problems. Brown & Kass (2009) reach the same conclusions for slightly different but related reasons: the traditional setting of consultant and researcher is anachronistic. Current statistical researchers are most successful when they themselves are principal investigators on projects, or have sufficient subject-area knowledge to raise foundational questions.

Summary and Recommendations

This paper is intended to provoke discussion and encourage statistics educators to evaluate their own level of technological literacy so that students can gain access to more data and realize the usefulness of statistics. One would like to provide suggestions and recommendations for where an interested statistician could turn for assistance, but unfortunately few such resources exist.

Instructors seeking to improve their technological literacy should look for workshops. The NSF-funded workshops mentioned above, led and organized by Deb Nolan, Duncan Temple Lang and Mark Hansen, are examples (and if only there were more examples.) At the 2008 workshop, "Computing in the Curriculum", participants learned or polished statistical computation skills and worked together to develop curricular materials. Some products of that workshop that might be useful resources are the lecture materials (www.stat.berkeley.edu/~statcur/), a wiki that includes links to existing courses (www.stat.berkeley.edu/twiki/Workshop/CompCurric) and a Google group on the topic of Computing in the Statistics Curriculum (groups.google.com/group/computing-statistics-curricula).

The fourth edition of the book Statistics: A Guide to the Unknown, (Peck *et al*, 2006) includes several case studies involving complex data couched in the context of

interesting, relevant research investigations. Most relevant to this discussion is a case study concerning spam filters for email.

Several websites exist that provide access to large databases and datasets. The Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html) includes datasets both large and small, and those listed under *CS/Engineering* cover some topics on themes raised in this paper, such as fraud detection in hacker attacks on computers, and email spam filtering. The site theinfo.org is a self-proclaimed site for "large datasets and the people who love them." It includes an extremely rich collection of databases that are themselves collections of large datasets. This site also includes mailing lists and resources for helping people wrangle large datasets.

The journal Technology Innovations in Statistics Education (tise.stat.ucla.edu) was founded to encourage scholarly research and discussion to explore how technology impacts the teaching and learning of statistics. I hope that it becomes a useful forum for discussing not just how to use technology to better teach, but how to better teach technology itself.

The central irony that faces the statistics educator today is that at no other time has the need for statistics been greater, and yet the data that drive this need are almost fully absent from the classroom. If statistics education is to be centered on data, if we agree with the Curriculum Guidelines for Bachelor of Science Degrees in Statistical Science (Rex, et. al, 1999) that an undergraduate statistics education should produce "data scientists", then we must overcome the challenges and bring these data into the classroom. In the book The Numerati, journalist Stephen Baker makes the claim that the great business and scientific successes of the future (if not the present) will be a result of analyzing and shifting through the huge amounts of data that our society routinely collects. In one of the more depressing sentences this statistician has read, he lists several professions that will participate to this new numerical society: economists, computer scientists, psychologists, biologists, and mathematicians. The science of data is missing from the list.

REFERENCES

Agresti, A., Franklin, C. (2007). *Statistics: The Art and Science of Learning from Data* (1st Edition). Upper Saddle River, New Jersey: Pearson/Prentice Hall.

Aliaga, M., Cobb, G.W., Cuff, C., Garfield, J. (Chair), Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., and Witmer, J. (2005) . GAISE College Report, American Statistical Association website: http://www.amstat.org/education/gaise/

Anderson, Christopher (2009), Wired Magazine: "Living by the Numbers", July 2009, Conde Nast Publications,

Baker, S. (2008). *The Numerati*, New York: Houghton Mifflin.

Bibby, J. (1986). *Notes Towards a History of Teaching Statistics,* Edinburgh: John Bibby (Books).

Bibby, John (2003). Rejoinder to "50 Years of Statistics Teaching in English Schools: Some Milestones", Holmes, P. *Statistician* **52**, Part 4, 439-474.

Brown, E.N. and Kass, R.E., (2009). What is Statistics? *The American Statistician*, **62**(2), 105-110.

Bryce, G. R., Gould, R., Notz, W. I., Peck, R. L., (2001). Curriculum Guidelines for Bachelor of Science Degrees in Statistical Science. *The American Statistician*, **55** , Part 1, 7-13.

Byron, Lee and Wattenberg, Martin. (2008). Stacked Graphs - Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14 (6), 1245-1252, Nov./Dec., 2008.

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*: **1** (1), Article 1. http://respositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1

Cobb, G.W. (1991). Teaching Statistics: More Data, Less Lecturing, *AMSTAT News,* No. 182.

Cobb, G. W. (1993). Reconsidering Statistics Education: A National Science Foundation Conference, *Journal of Statistics Education*, **1**(1). www.amstat.org/publications/jse/v1n1/cobb.html

Cobb, G. W. & Moore, D.S. (1997). Mathematics, Statistics, and Teaching. *The American Mathematical Monthly*, **104** (9), 801-823.

College Board, (2006). *College Board Standards for College Success: Mathematics and Statistics*." www.college.board.com

Committee on Technological Literacy; National Academy of Engineering; National Research Council, (2002). Technically Speaking: Why All Americans Need to Know More About Technology, Eds. Pearson, G., & Young, A.T.. The National Academies Press.

DataDesk (2008). Data Description, Inc.. www.datadesk.com

Dinov, I. D., (2006). SOCR: Statistics Online Computational Resource. *Journal of Statistical Software*, 16 (11). www.jstatsoft.org/v16/i11

Economist, The (2010). Data Deluge: Special Issue, February 27, 2010, The Economist Newspapers Ltd.

Eisenberg, A., (1996). Privacy and Data Collection on the Net, *Scientific American,* **274**(3), p. 120.

Fisher, R.A. (1958) Statistical Methods for Research Workers, Thirteenth Edition, Haftner Publishing Company, New York.

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

Finzer, W., Erickson, T., Swenson, K., Litwin, M., (2007) . On Getting More and Better Data Into the Classroom. *Technology Innovations in Statistics Education*: **1**(1), Article 3. repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art3

Freedman, D., Pisani, R., Purves, R. (1978). *Statistics*. New York: W. W. Norton & Co.

Gal, Iddo. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. International Statistical Review, 70 (1), 1-51.

Garfield, J., Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review*, 75 (3) , 372-396.

Grossman, L., (2009). Iran Protests: Twitter, the medium of the Movement, Lev Grossman, *Time Magazine*, Wednesday June 17, 2009, www.time.com, viewed May 25, 2010.)

Hansen, M., Landa, J., Schaefer, S., Chapman, G., Goode, J. and Margolis, J., (2010). "Unit 6: Particpatory Urban Sensing" draft from *Exploring Computer Science*, Computer Science Equity Alliance.

Hansen, M., Nolan, D., and Temple Lang, D., (2006), "Undergraduate Summer Statistics Program", available at *http://www.stat.berkeley.edu/~summer/*

Hengl, Tomislav, (2010). *A Practical Guide to Geostatistical Mapping*. University of Amsterdam, February 17, 2010.

Hogg, R. W. (1991). Statistical Education: Improvements are Badly Needed. *The American Statistician,* **45**(4) , 342-343.

Hunter, W.G. (1981). The Practice of Statistics: The Real World is an Idea Whose Time has Come. *The American Statistician,* **35** (2) 72-76.

ITC Literacy Panel, (2002). Digital Transformation: A Framework for ICT Literacy. A Report of the International ICT Literacy Panel, *Educational Testing Service*. www.ets.org/Media/Research/pdf/ICTREPORT.pdf

Joiner, B. L., (1988). Let's Change how we Teach Statistics. *Chance*, **1**(1), 53-54.

JMP, Version 7. SAS Institute Inc., Cary, NC, 1989-2007.

Kaplan, D. (2007). Computing and Introductory Statistics. *Technology Innovations in Statistics Education*, **1**(1), Article 5, repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art5

Moore, D. S., New Pedagogy and New Content: The Case of Statistics. I*nternational Statistical Review*, **65** (2) 123-165.

Nolan, D., and Temple Lang, D., (2007). Dynamic, Interactive Documents for Teaching Statistical Practice. *International Statistical Review*, **75**(3) 295-321.

Nolan, D., Temple Lang, D., and Hansen, M., (2007), "Wiki: Computing in Statistics: Model Courses and Curricula," available at http://www.stat.berkeley.edu/~statcur/."

Nolan, D. and Temple Lang, D. (2009). Comment to "What is Statistics?", *The American Statistician,* 63(2), 117-121.

Peck, R., Casella, G., Cobb, G., Hoerl, R., Nolan, D., Starbuck, R., Stern, H., (2006). *Statistics: A Guide to the Unknown*, (Fourth Edition). United States: Thomson Brooks/Cole.

Rose, L.C., Gallup, A. M., Dugger, W. E., Jr., Starkweather, K.N., (2004). The Second Installment of the ITEA/Gallup Poll and What It Reveals as to How Americans Think About Technology, *International Technology Education Association*, www.iteaconnect.org/TAA/PDFs/GallupPoll2004.pdf

Rubin, Andee (2007) . Much Has Changed; Little Has Changed: Revisiting the Role of Technology in Statistics Education 1992-2007. *Technology Innovations in Statistics Education*; **1**(1), Article 6. http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art6

Rumsey, D. J., (2002). Statistical Literacy as a Goal for Introductory Statistics Courses, *Journal of Statistics Education*, **10**(3). www.amstat.org/publications/jse/v10n3/rumsey2.html

Selber, S.A., (2004). *Multiliteracies for a Digital Age,* Southern Illinois University Press, 2004.

Singer, J. D. and Willett, J. B., (1990). Improving the Teaching of Applied Statistics: Putting the Data Back into Data Analysis, *The American Statistician* , **44**(3), 223-230.
Snee, Ronald D., (1993). What's Missing in Statistical Education?, *The American Statistician,* **47**(2), 149-154.
Snell, L. J., (2003). A Course Called Chance, *Proceedings of the ISI, 54th Session*, Berlin, German.
Sweeney, L., (200). Uniqueness of Simple Demographics in the U.S. Population, working paper, LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, PIttsburgh, PA.
Sweeney, L., (2002). k-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty. *Fuzziness and Knowledge-based Systems*, **10**(5), 557-570.
Utts, Jessica, (2003). What Educated Citizens Should Know About Statistics and Probability, *The American Statistician,* **57**(2), pp 74-79.
Wells, H.G., (2004), The Project Gutenberg EBook of Mankind in the Making, EBook #7058, The Gutenberg Project. Available at *http://www.gutenberg.org/etext/7058* A copy of the first edition (1903, London: Chapman & Hall) is available via Google books.
West, W., (2009), Social Data Analysis with StatCrunch: Potential Benefits to Statistical Education, *Technology Innovations in Statistics Education*, 3(1).
Wild, C.J., (1994), Embracing the "wider view" of statistics, The American Statistician, 48, 163-171.
Wild, C.J., Pfannkuch, M., (1999). Statistical Thinking in Empirical Enquiry, *International Statistical Review*, **67**(3) ,223-265.