TEACHING DATA SCIENCE TO SECONDARY STUDENTS: THE MOBILIZE INTRODUCTION TO DATA SCIENCE CURRICULUM

Robert Gould, Suyen Machado, Christine Ong, Terri Johnson, James Molyneux, Steve Nolen,
Hongsuda Tangmunarunkit, LeeAnn Trusela, Linda Zanontian
Dept. of Statistics, University of California, Los Angeles
Los Angeles, CA 90095, USA
rgould@stat.ucla.edu

Making sense of data is complex, and the knowledge and skills required to understand "Big Data" - and many open data sources - go beyond those taught in traditional introductory statistics courses. The Mobilize project has created and implemented a course for secondary students, Introduction to Data Science (IDS), that aims to develop computational and statistical thinking skills so that students can access and analyze data from a variety of traditional and non-traditional sources. Although the course does not directly address open source data, such data are used in the curriculum, and an outcome of the project is to develop skills and habits of mind that allow students to use open source data to understand their community. This paper introduces the course and describes some of the challenges in its implementation.

INTRODUCTION

The Mobilize Introduction to Data Science (IDS) curriculum is a yearlong course designed to cultivate statistical and computational thinking skills in order to prepare secondary school students to live and work in a data-rich world. A novel feature of IDS is its use of participatory sensing (PS), a data collection paradigm developed to create communities whose members collect and analyze data together (Burke, et al. 2006).

A goal of the Mobilize project, and of IDS in particular, is to instill in students "data habits of mind", or roughly speaking, a collection of attitudes and reflexive approaches to understanding the world through data (Finzer, 2013). One such habit is that students should seek data to reach conclusions and construct arguments. As they go through this process, students should be aware of the existence of previously collected data - including open data - that might address their needs. As they discover potential data, students must develop the skills to obtain and analyze it, and so IDS aims to strengthen students' conceptual understandings so they can make valid interpretations of valid analyses.

Why Teach Data Science?

While we may be living through a "data deluge", (www.maa.org/mathematics-awareness-month-2012), statistics education at the secondary level (and sometimes beyond) still focuses very much on a 20th century paradigm developed during an age when data came from designed studies or were curated by experts, when software was expensive, and when the purpose of a statistical analysis was to provide a p-value or find a confidence interval. This is a missed opportunity, because data are omnipresent and high quality software is inexpensive (or free), and thus many of the barriers that in the past prevented secondary students from analyzing data are now gone.

To illustrate this lost opportunity, consider two "parables." The first is the parable of “Pizza Girl”. Karin Wellman is a blogger who wrote a three-part series on the "Serious Eats" blog called "Pizza Girl: Statistical Analysis of a Delivery Shift" (Wellman, 2010). In her blog, Wellman posted data she collected while working at a pizza delivery restaurant. The data consisted of time stamps, a description of the work done during her shift, and the amount of money she earned. An issue addressed in her blog was whether or when it was in her best financial interests to leave the shop and deliver a pizza. She was paid less per hour while out delivering pizzas, but might recoup that loss (or more) by earning tips. She posed questions: How much does a zero tip affect my effective hourly rate? How much better is it to take a triple instead of a double delivery? She then analyzed her data using pie charts (as we would expect and hope of someone working at a pizza restaurant), but had difficulty using these analyses to answer her questions. Her analyses caught the attention of Jared Lander, a statistician who shares a passion for pizza (2008). Lander produced

In: J. Engel (Ed.), *Promoting understanding of statistics about society.*
Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE),
July 2016, Berlin, Germany. ©2016 ISI/IASE iase-web.org/Conference_Proceedings.php

some basic boxplots, which helped Wellman see that while those who pre-pay for their pizza deliveries have the greatest variability in how much they tip, on average those who pay in cash tip the most. By comparing distributions of numerical variable side-by-side, she was able to answer one of her primary questions and decide when it was wise to make a delivery.

The second parable is the widely reported story about the retailer Target, which angered a father by sending his sixteen-year-old daughter coupons clearly aimed at pregnant women. He demanded an apology from Target and received it, but later learned that his daughter was, indeed, pregnant. According to Forbes Magazine, a Target statistician had identified 25 products that allowed him to assign a "pregnancy score" to shoppers. The teenager's purchases, any one of which by itself would have meant nothing, together produced a high "pregnancy score" and triggered a privacy breach between a major corporate retail chain and a family (Hill, 2012).

There are two important lessons from these parables. First, knowledge about how to use data to answer questions and reach insights can enable everyone to answer important questions they raise about their life or career, regardless of whether they are future scientists or future pizza bloggers. Second, in the current age of the "data deluge', ignorance of the role of data in our lives will not protect us from harm. A commonly cited dichotomy among statistics educators, one that perhaps began with mathematical statistician Harold Hotelling in the 1940s (Hotelling, 1940) is that students are either "consumers" or "producers" of statistics, and that preparation in statistics should be adjusted accordingly. In reality, we cannot easily predict, even based on a person's chosen career, who would be best served with an education as a producer or as a consumer. In truth, all people, even statisticians, move between these roles during their lives. The difference is that those whose education prepares them to be producers are more capable of assuming the producer role, as they are more adept at taking advantage of the opportunities provided to them by ubiquitous data (and the tools available to analyze them) to analyze and share data analyses with others.

Another reason for teaching data science is that it can provide access to data that intersect with students' lives. Not every student, particularly at the secondary level, has an interest in becoming a scientist. However, since today's students engage with data in their daily lives through social media, electronics, and shopping, we as educators are missing an important opportunity if we don't prepare students to work with "big data" as discussed in Gould (2010),

One reason that more classrooms do not include big data is that these data can be difficult to work with. While "traditional" data sometimes consist of a large (though often not very large) number of rows, they usually consist of only a few columns. In many textbooks, only the columns required to address a given problem are provided. While traditional data are "numbers in context" (Moore, 1990), big data consist of a variety of formats: images, text, sounds, dates, and locations. In fact, variety of formats is one of IBM's initial "three V's of Big Data" (IBM, 2013). Traditional data come from random samples or random assignment. Big data can be much messier, involving complex structures and non-random samples, such as the continuous streams of sensors or the 'algorithmically' generated data of participatory sensing campaigns, as discussed below. In curricula designed to teach only statistical inference, how can such important data be accommodated?

This paper discusses the Mobilize project's Introduction to Data Science (IDS) curriculum, which aims to prepare students for a world of big data.

THE MOBILIZE INTRODUCTION TO DATA SCIENCE CURRICULUM

The Mobilize project, funded by the National Science Foundation in 2010, began as a computer science education initiative intended to leverage the power of mobile devices to teach students about computing and to enhance science and math education. One key component was participatory sensing, which, by enabling students to collect their own data, has the potential to make data analysis more meaningful to students. Short curriculum modules ranging from three to six weeks were developed for algebra and biology classrooms, and a 6-week Computing and Data Analysis unit was integrated into the pre-existing Exploring Computer Science (ECS) curriculum. However, the challenges of learning to use new technology, learning to collect data, and learning to analyze data were too great to be achieved in these short time periods. The IDS curriculum was created, in part, to provide enough time for students to explore data analysis in greater depth and to create and carry out complete participatory sensing campaigns.

Participatory Sensing, Big Data, Open Data

Participatory sensing (PS) is a paradigm for data collection that strives to establish a community of people who are united by the desire to collect and analyze data on a common topic (Burke, et al, 2006). Data are collected via mobile devices and both the data and the analyses are shared within the community. A PS "campaign" includes the entire arc of a data investigation. The campaign begins with the decision to understand or study a particular process or phenomenon, then continues when community members agree on survey questions that will measure this process. From there, community members collect data for a period of time, then analyze the data and reach conclusions. In the data collection phase, observations are collected not through a random sampling procedure, but instead through "trigger events." For example, in the Mobilize Snack campaign, the trigger event is eating a snack (food outside of a meal). In the Mobilize Trash campaign, students collect data whenever they discard an item. This data collection scheme, or "algorithmic" scheme, is similar to that of an electronic sensor whose trigger events are usually determined by timing devices (and occur quite rapidly), but might also be triggered by, for example, detected motion or changes in temperature.

The algorithmic quality of the data collection is just one feature that PS data share with big data. Another feature is the variety of data types that can be captured in a PS campaign. For example, in the Trash campaign the variables are: categorical (which type of bin was the discarded item placed in; what type of item was it; what activity generated the item; where the item was discarded), numerical (the number of recycling bins visible from the location where the item was discarded; the number of trash bins visible; the number of compost bins visible), image (photos of the items), date, time, location (as a latitude and longitude), and text (an open-ended description of the item).

Open data, defined as data made freely available to the public, also play a role in the IDS curriculum. One objective of the curriculum is to make students aware of the existence of such data sets. Other objectives are to teach them to gain insight from data in which the number of variables might be overwhelming, and to teach them to work with "opportunistic" data, i.e., data used for a purpose that is possibly different than the purpose for which the data were collected (Huber, 2011). When appropriate, the curriculum ties PS campaigns to open data in an attempt to make the open data more "real". The Time Use PS campaign is an example of this. Students collect data four or five times per day for several days, triggered by an alarm that rings at times predetermined by the teacher. At the alarm, students open their app and answer questions about the activities they have participated in since the last alarm. In addition, date, time, latitude and longitude are stored. To motivate the Time Use campaign, students begin by examining the New York Times Interactive graphic (http://tinyurl.com/Daily-Time-Use) of the American Time Use Survey (ATUS) data set (www.bls.gov/tus/home.htm) - an open data set created by the U.S. Bureau of Labor Statistics - and create a report on their findings.

IDS Curriculum

The goal of IDS is to teach students to learn to compute with data and to think statistically about the variety of data in the modern world. Lessons consist of classroom activities and discussions, computer lab exercises in which students work collaboratively to learn to use the statistical programming language R via RStudio (RStudio team 2015), and participatory sensing campaigns. The course is built around an “inquiry-based” approach to pedagogy, and consists of four thematic units: Working with Data, Informal Inference, Data Collection, Modeling and Prediction. Each unit has one PS campaign and several labs in which students learn to use R (R Core Team, 2016) via RStudio (RStudio Team, 2015) to apply concepts discussed during the regular lessons and to analyze data collected in the PS campaigns and open source data. Teachers attend approximately 60 hours of training prior to and during implementation. IDS assumes the students have passed basic algebra.

IDS is based on a blend of educational frameworks from the statistics and computer science communities. The statistical component is based on the first two learning levels (levels A and B) of the Guidelines for Assessment and Instruction In Statistics Education (GAISE) K-12 Report (Franklin et. al, 2007). Roughly speaking, these levels correspond to "beginner" and "intermediate", although neither level includes formal statistical inference. Consistent with levels A and B, IDS emphasizes "informal" reasoning (Makar & Rubin, 2009) and is conceived to be a precursor to a "formal" statistical course, such as the Advanced Placement Statistics course.

The computing aspect of the course is based on the Computer Science Teachers Association (CSTA) K-12 Computer Science Standards (CSS) (CSTA Standards Task Force, 2011). These standards include general suggestions such as helping students understand the way networks work across mobile and computing devices, and the ways in which data are stored and sent. They also include standards particular to data analysis, such as "analyze data and identify patterns through modeling and simulation" (p 60). The CSS encourage the development of "computational thinking", which includes such practices as learning to employ algorithms and think critically about algorithms, the ability to break problems into smaller parts using computational structures (such as functions or classes), and understanding data storage structures.

The curriculum content is distributed across four 10-week units, as shown in Table 1. The topics include many traditional topics (as required by the California State Standards), but also some unusual topics, such as developing an understanding of how data are organized (Unit 1), how and why to merge data files (Unit 2), writing survey questions to address a statistical question (Unit 3), classification and regression trees (CART, Unit 4) and others (shown in bold, Table 2). The curriculum also includes randomization testing, a topic that has increasingly been included to introductory statistics curricula in an attempt to make the inferential paradigm more internally consistent (Cobb, 2007).

Table 1. An overview of some topics covered in the IDS curriculum. Topics in bold are non-traditional.

Unit 1	Unit 2	Unit 3	Unit 4
Organizing and structuring data; The data investigative cycle; Asking statistical questions; What is typical? Interpreting histograms; Associations.	What do means, medians, MADs and SDs measure? Comparing groups when faced with variability; How likely is an outcome? What is bias? Basic probability **Simulating basic probabilities.** **What does it mean to say something "happens by chance"?** **How can we simulate "by chance"?** **How and why do we merge data files?** What's the normal model?	Anecdotes as evidence? What is a controlled experiment? Why is random assignment important? What is an observational study? How do obs. studies and randomized experiments compare? What are surveys? **How do we write survey questions that address a statistical question?** How do we communicate uncertainty in surveys? **How to design a PS campaign.** **How are data organized on a web page, and how can we get it into our stats software?**	**Create a PS campaign.** **What's a statistical prediction?** **How to measure success when making predictions.** **Predictions with regression.** **Predictions with CART.** **Modeling with Data Classification and Clustering.**

Technology in IDS

The IDS curriculum relies on a variety of open-source software, much of which was developed by the Mobilize team and the UCLA Center for Embedded Network Sensing (Tangmanarunkit et al, 2015). The centerpiece of the IDS software suite is the open source Ohmage system, which supports and manages the flow of data from mobile devices to the classroom. The Ohmage system consists of four components, which include (1) a "backend" that provides infrastructural support such as user authentication, management and storage of data; (2) apps for mobile devices, available free of charge from the Google Play store or the Apple App store; (3) web-based classroom management tools, including tools for creating and deploying new surveys; and (4) web-based data analysis and visualization tools. A detailed description, as well as source code, is available at http://www.mobilizingcs.org/introduction-to-data-science/the-mobilize-technology-suite-for-ids).

Figure 1 shows the Mobilize dashboard, a visualization tool that here is shown displaying publicly available data from the "snack campaign". The dashboard displays the distribution of costs, the time of day the snack was consumed, and the distribution of the perceived healthy level (a 5 being the most healthy). The dashboard is interactive, so that clicking on any displayed value causes the dashboard to update to the subset of data consisting of only the selected value. For example, in Figure 1, because the "morning" value of the "When" variable donut chart is not grayed-out, and because the filter on the word cloud for the "What" variable is set to "cereal", we know that we are seeing snacks consisting of cereal consumed in the morning. The "Healthy" histogram shows us that students perceived cereal to be relatively healthy, on average.

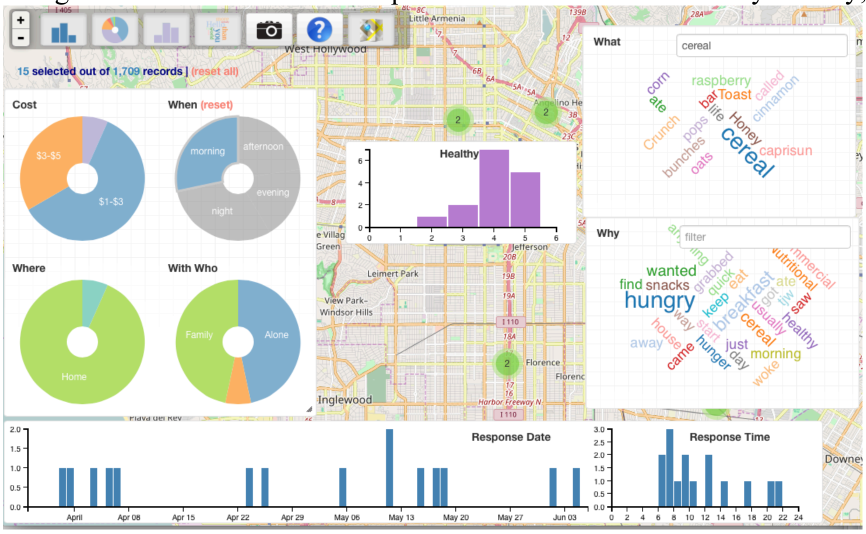

Figure 1. The Mobilize dashboard, displaying publicly available data collected from a snack campaign in which participants recorded data every time they ate a snack. The data shown here are subset consisting of only snacks that were considered to be "cereal" and were eaten in the morning. The map indicates locations where the data were collected (although these locations have been perturbed in this public data to protect anonymity.)

We could compare the subset in Figure 1 to cereal consumed at night by clicking on the "When" value "night". A new subset is displayed (See Figure 2) and we see that cereal consumed at night is perceived to be less healthy. The dashboard is available with public data (provided by consent of the contributors, and with values anonymized and perturbed slightly with random noise to maintain anonymity) at https://sandbox.mobilizingcs.org

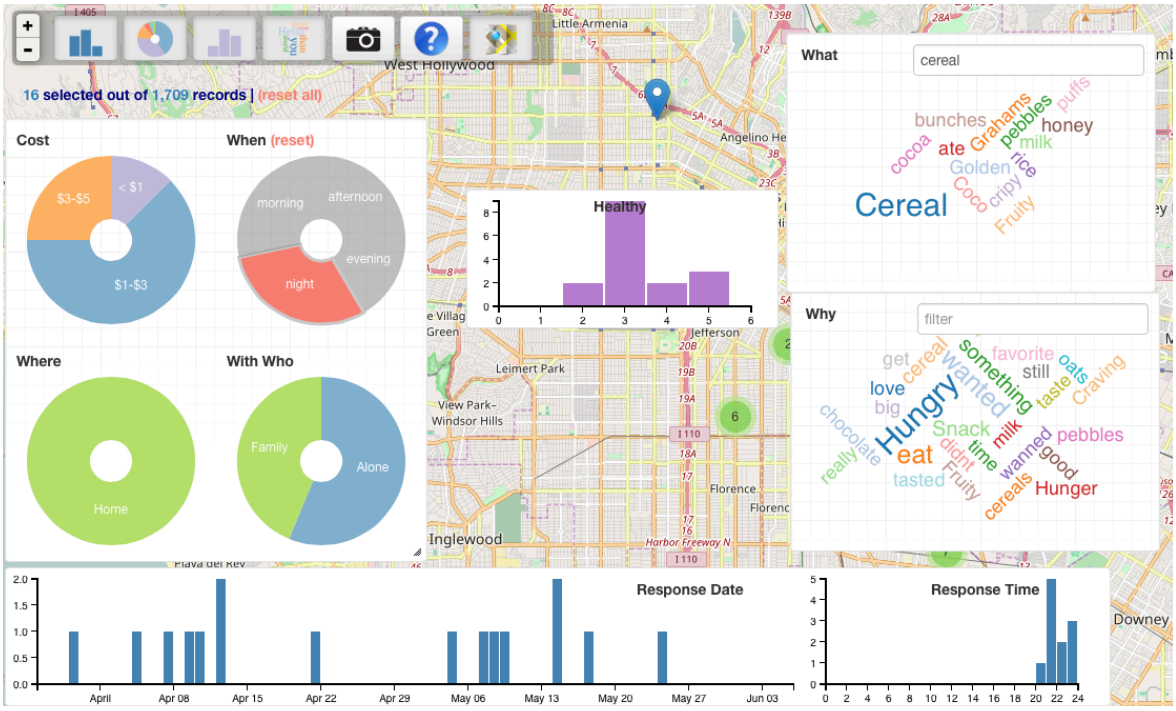

Figure 2. Data from the Snack campaign, this time subset so as to show cereals eaten at night.

The focus of the IDS course is on learning to compute with data, and so students learn to analyze their data using R via RStudio. The *mobilizR* package (Molyneux et al., 2016) was developed to provide students with a more unified syntax than is typical with R, and to simplify some complex analyses (e.g., creating word clouds). The *mobilizR* package was based upon the *mosaic* package (Pruim et al., 2015), which was designed, in part, to create a coherent language for randomization-based testing for introductory statistics.

IDS in Los Angeles

IDS was developed in close cooperation with the Los Angeles Unified School District (LAUSD). LAUSD is the second-largest school district in the U.S., with approximately 650,000 students. Almost three-quarters of the students are Latino, 8% African American, 10% White, and 4% Asian. Approximately 25% have limited fluency in English, and 79% are classified as below the poverty level (http://achieve.lausd.net/mydata).

In its first year (2014-2015), 10 LAUSD teachers taught 13 sections of the course, with an average class size of 35 students each. These teachers acted as advisors and collaborators on the curriculum, and served as mentors for the second cohort of 27 new teachers representing 24 new classrooms in 2015-2016. All teachers held credentials in mathematics except two who held special education credentials. Teachers were paid a stipend for participating in professional development and for contributing to the curriculum and participating in interviews and other data gathering by an external evaluation team.

Professional development for both cohorts consisted of two summer institutes (a four-day institute early in the summer and a three-day institute near the start of the academic year) and five daylong sessions held on Saturdays throughout the year. The development sessions modeled inquiry-based methods for teaching the statistical and computational concepts, and workshops to learn to use R.

LESSONS LEARNED

The IDS curriculum was created, in part, to address the fact that when implementing the shorter modules for algebra, biology, and computer science, teachers and students struggled with the task of developing novel effective participatory sensing campaigns. A key reason for this was that the time required to learn both the technology and the statistical concepts was too limited in these short modules. But perhaps the primary reason was that the investigators greatly underestimated the need to develop statistical thinking practices among teachers. Teachers had little to no experience in working through a statistical investigation from beginning to end.

To address this inexperience with statistical investigations, we introduced the notion of the "Data Cycle" (Figure 3.), which is based on the four components of the "statistical investigative process" in the GAISE K-12 (Franklin et. al. 2007). The GAISE statistical investigative process is itself closely related to the "PPDAC" cycle - Problem, Plan, Data Analysis, Conclusions - proposed by Wild and Pfannkuch (1999) as a framework for statistical thinking. One difference between the data cycle and the GAISE process is the second stage. In the GAISE, this stage is "collect" data, but we renamed this to "consider" data to take into account that in many analyses of secondary studies, the data are already collected, which calls for a different set of habits and competencies. In IDS professional development, teachers were encouraged to post the Data Cycle in their classrooms and to have students reflect on which stage of the cycle they were in at any given moment. Although the arrows suggest a single direction of motion, in practice we would expect more complex iterations between states as well as retrograde motions.

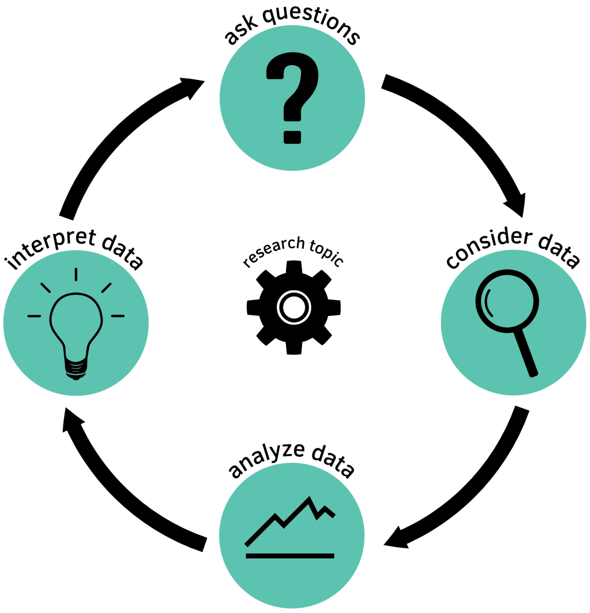

Figure 3. The Data Cycle as presented in the IDS curriculum. The arrows point to pathway for an "ideal" analysis.

Two steps of this cycle proved particularly novel to teachers: "Ask Statistical Questions" and "Consider Data". Statistical questions are questions that address variability, can be addressed with data and, when answered, lead towards greater understanding of the primary research topic. Teachers were unfamiliar with this concept, in part because the concept had only recently been introduced in the state curriculum standards (California State Board of Education, 2013), and also because the official definition is vague (see standard 6.SP for an example). Further, teaching students to ask questions in the context of a statistical investigation is outside the mainstream of mathematics education. The entire initial cohort of teachers reported that, prior to the IDS course, they had rarely or never asked students to develop research questions (Ong et al., 2015). Even with practice, teachers demonstrated difficulty (a) posing statistical questions that advanced the research topic; (b) posing statistical questions that could be addressed by the data at hand; and (c) posing survey prompts that might be used to collect data that would address the statistical questions. Our experiences are consistent with those described in Arnold (2013) and Bargagliotti (personal communication, June 2015). Because of their inexperience with this step of the cycle, teachers and their students were often at a loss to know where to begin an analysis when confronted with a data set.

The second step of the Data Cycle was also unfamiliar to most teachers. All but one teacher had rarely or never used data collection in their teaching or analyzed data from existing data sets or databases (CRESST Team, 2015). Mobilize had large goals for this step beyond data analysis. Its goal was for students to gain awareness of the larger role data play in their lives and in their future, and to develop a habit of using data to answer questions. Some teachers initially had a restricted notion of data, believing, for example, that photos could not be considered "data" and that data must come from random samples. Because participatory sensing data played an important role, one goal was for teachers to recognize the limitations of generalizing conclusions based on such data, but also to recognize the strengths of this method. Teachers demonstrated strong critical thinking and healthy skepticism when examining data that came from random samples (such as the ATUS data). However, these skepticisms were not verbalized when teachers were engaged in activities analyzing PS data.

To examine how teachers engaged with the Data Cycle, small groups of teachers were asked to participate in a Model Eliciting Activity (MEA) (Lesh et al., 2000) during a professional development session in the middle of the year. In this activity, teachers were asked to suggest two ways that Los Angeles County could ease the burden on its landfill (waste disposal site) and to support their suggestions using data. Data from the public Trash Campaign (https://sandbox.mobilizingcs.org/#demo/) was provided in the form of a comma-delimited text file. Teachers had access to both the dashboard view of the data and to RStudio. We analyzed transcripts and video of the teachers' work on this MEA so as to better understand how teachers approached statistical investigations with non-traditional data (Gould et al., 2015).

We found that teachers' pathways through the Data Cycle were complex, although these pathways demonstrated the key role that questioning in general, and statistical questioning in particular, played in driving the investigation forward. The quality of statistical questions raised during the investigation was measured using a simple four-point scale in which questions were awarded one point for each of the following criteria they satisfied: (1) the variable of interest was clear and available; (2) the population of interest was clear; (3) the question could be answered with the data available; and (4) the question was worth investigating, had a purpose, or was interesting. These criteria were based on Arnold (2013, p. 110-111) in her examination of the role of statistical questioning. The mean scores of the two groups of teachers we examined were 2.9 (out of 4 points) with standard deviation of 0.2, and 2.0 (s = 0.08). Both groups had room for improvement in terms of the quality of their statistical questions, and our analysis of the transcripts demonstrated that while questioning was an important strategic tool, it was quite difficult and challenging.

A challenge for research on projects such as this is that there are, as yet, no measures of learning in the context of data science, particularly as it is conceived here as a blend of computational and statistical thinking. Assessments of computational thinking are somewhat new; SRI International provides one such assessment: Principled Assessment of Computational Thinking (PACT), although this was not available to us during the most recent evaluation cycle for our project (pact.sri.com). The Mobilize external evaluation team designed three items centered on a single task in order to assess one aspect of computational thinking (data collection and organization). The task, which includes both closed- and open-ended items, requires students to identify variables present within an example data set and determine whether one can make a conclusion based on the data collected. Most students completing the post-survey were able to determine that one could not make a conclusion based on data provided. However, students had more difficulty explaining why this was so (Ong et al., 2016).

There are, however, several assessments of statistical thinking, and we chose the Levels of Conceptual Understanding in Statistics (LOCUS) assessment as being most closely aligned with IDS objectives. LOCUS is a criterion-referenced, nationally validated measure of statistical thinking aligned with the GAISE (Jacobbe et al., 2014). LOCUS provides assessment for each of the four stages of statistical problem solving, and these stages map to the data cycle. Students in IDS classrooms took the Beginner/Intermediate level LOCUS at the start of the academic year and took a different version of the same instrument at the end of the academic year. Of the 1,301 students who participated, we were able to match consented pre- and post- assessments for only 440 students. Performance varied somewhat by class, with one classroom actually showing a decline in performance.(this was later attributed to issues that affected classroom management on the day the follow-up assessment was administered). An item-response theory analysis showed that on average, students improved 10 percentage points (in terms of raw scores) ($p < 0.001$), and improvement was statistically consistent across all four domains (questions, data, interpretation, analysis) (Ong et al., 2016).

CONCLUSION

The field of data science is relatively new, and still defined differently across stakeholders. In some ways, the name of the field seems to have arisen as a critique of the statistics community, based on a perception that Statistics was not interested in engaging with data in a way that would take full advantage of what one could learn from data. Breiman (2001) noted two "cultures" of data analysis. The "data modeling" culture was predominant, but "has kept statisticians from working on a large range of interesting current problems". The "algorithmic modeling" culture, he claimed, was better suited to the age of big data. Cleveland (2001) offered a definition of data science that arose from a similar criticism that the academic statistics community was not sufficiently concerned with data analysis. The purpose of data science is, according to Cleveland, to "enable the analyst to learn from data". Donoho (2015), building on Breiman and Cleveland, notes that much of the discussion about "data science" is hype, but concludes "there *is* a solid case for *some entity* called 'Data Science' to be created, which would be a true science"; this new science is the science of learning from data.

It is worth noting that this reaction is based on a perception that the field of statistics is limited to a narrow, theoretical perspective. Donoho: "Such exhortations [towards a broader definition of statistics] had relatively little apparent effect before 2000." But things *have* changed in the statistics community since 2000. Nolan and Temple Lang (2010) urged a greater focus on computing in the undergraduate statistics curriculum. More are arguing for the need to bring data science concepts to the introductory curriculum (Horton et al. 2014; Hardin et al., 2015). The CATALST Project (Garfield et al., 2012) not only uses large open data sets and "modern" data such as those generated by mp3 music players, but also explicitly engages students in the entire statistical investigation process. Some institutions are offering data science programs for undergraduates, and the Park City Math Institute convened an undergraduate faculty program on the topic of developing curriculum guidelines for undergraduate data science programs (PCMI Undergraduate Faculty Group, 2016).

Much less has been written about, and is known about, the role that data science could play at the secondary-school level. The notion of computing with data seems central to the vision of data science as it is emerging at higher levels of education. Computation is not widely taught at high schools in the United States, although this is rapidly changing, in part due to programs such as Into the Loop Alliance, a partnership between UCLA, the University of Oregon, and the LAUSD, which produced a widely used computer science curriculum, the aforementioned Exploring Computer Science (ECS) curriculum. (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1241284). Organizations such as code.org and the Obama administration's Computer Science For All initiative continue to promote the importance of computer science and "coding" in high schools. As computer science grows in the secondary schools, "computing with data" will have an increasingly comfortable fit.

More research would help make this fit more comfortable. Currently, statistics in the secondary schools focuses on inference, and is immersed in the "data modeling" culture. Education research is therefore, for the most part, concerned with how students understand concepts fundamental to the development of statistical inferential reasoning. If data science is to grow in the high school classroom, more research must be done to understand, for example, how students can learn data analysis through computation (and learn computation through data analysis). We have seen that students and teachers struggle with R, and that this struggle can appear to impede their ability to engage in a successful statistical investigation. Perhaps the pedagogy must be improved, or perhaps we are using the wrong software. McNamara (2015) writes about the need for software that bridges the gap between "learning" and "doing" statistics, and points out one possible pathway future data science curricula might follow. Other groups are laying groundwork for the future as well. The Oceans of Data project held a workshop to outline a plan to develop global data literacy in the age of big data (Oceans of Data, 2016). The Common Online Data Analysis Platform (CODAP) (http://codap.concord.org), which has worked with Oceans of Data, is developing software to help secondary students access and analyze data with complex structures, such as hierarchical data.

This Roundtable was devoted to understanding how to help students learn from, and about, open data. The IDS curriculum engages students directly with some open data sets (such as the American Time Use Survey), but is also concerned with the more general goal of developing the statistical and computational tools to access large, complex open data sets and discover insight. On many fronts, the project succeeds. Students report a greater understanding of the role of data in their lives, and many participated in relatively sophisticated discussions about the nature of the ownership of data. Statistical thinking, as measured by LOCUS, improved on average, and a substantial percentage of students reported that coding was the most important thing they learned from the course. With these skills, we are confident that students will be much better prepared to explore the world of open data.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 0962919. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Arnold, P. (2013). Statistical investigative questions - An enquiry into posing and answering investigative questions from existing data. (Doctoral thesis.) Retrieved from https://researchspace.auckland.ac.nz/handl/2292/21305

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*. 16(3), 199-231.

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, M., Srivastava, M.B. (2006). Participatory Sensing. *Center for Embedded Network Sensing.* UCLA: Center for Embedded Network Sensing. Retrieved from http://escholarship.org/uc/item/19h777qd

California State Board of Education (2013). *California Common Core State Standards: Mathematics*, electronic edition, California Department of Education, http://www.cde.ca.gov/be/st/ss/documents/ccssmathstandardaug2013.pdf.

Cleveland, W., (2001) Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *ISI Review.* Vol. 69, 21-26. Accessed from www.stat.purdue.edu/~wsc/papers/datascience.pdf.

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, 1(1). Retrieved from http://escholarship.org/uc/item/6hb3k0nz.

CSTA Standards Task Force (2011). CSTA K-12 Computer Science Standards Revised 2011, Seehorn, D (chair), Carey, S. , Fuschetto, B., Lee, R., Moix, D., O'Grady-Cunniff, Owens, B., Stephenson, C., Verno, A. Computer Science Teachers Association, New York, NY.

Donoho, D. (2015). 50 Years of Data Science. Delivered at the John W. Tukey 100th Birthday Celebration, Princeton University, September 18, 2015. Accessed via http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Finzer, W., (2013). The Data Science Education Dilemma. *Technology Innovations in Statistics Education,* 7(1). uclastat_cts_tise_13891. Retrieved from http://escholarship.org/uc/item/7gv0q9dc

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A preK-12 curriculum framework.* Alexandria, VA: American Statistical Association. (Also available at http://www.amstat.org/)

Garfield, J., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883-898.

Gould, R., (2010). Statistics and the Modern Student. *International Statistical Review*. 78(2), 297-315.

Gould, R., Johnson, T., Machado, S., Molyneux, J., and Bargagliotti, A., (2015). Modeling with "Big Data" in Secondary Schools: An exploratory study. Presentation at SRTL-9 (Statistical Reasoning, Thinking and Literacy), Paderborn, Germany. July 2016.

Hotelling, H. (1940). The Teaching of Statistics. *The Annals of Mathematical Statistics*. 11(4), 457-470.

Hill, K., (2012). How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did. Forbes.com, February 16, 2012. www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/ Accessed October 18, 2016.

Hardin, J., Hoerl, R., Horton N.J., Nolan, D., with Baumer, B., Hall-Holt O., Murrell, P., Peng, R., Roback, P., Temple Lang D., & Ward M.D. (2015). Data science in statistics curricula: preparing students to "think with data". *The American Statistician*, 69(4).

Horton, N.J., Baumer, B., Wickham, H., (2014). Teaching precursors to data science in introductory and second courses in statistics. In K. Makar, B. deSousa and R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS 9, July, 2014), Flagstaff, AZ, USA.* Voorburg, The Netherlands: International Statistical Institute. Accessed via arXiv:1401.3269.

Huber, P. J. (2011). *Data analysis: what can be learned from the past 50 years* (Vol. 874). John Wiley & Sons.

IBM. The Four V's of Big Data. (n.d.). Retrieved May 15, 2015, from http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

Jacobbe, T., Case, C., Whitaker, D., Foti, S. (2014). Establishing the content validity of the LOCUS assessments through evidence centered design. In K. Makar, B. deSousa and R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS 9, July, 2014), Flagstaff, AZ, USA.* Voorburg, The Netherlands: International Statistical Institute.

Lander, J., (2008). New York Pizza: How to Find the Best. MS Thesis, Columbia University. http://www.jaredlander.com/content/2010/03/Pizza-Thesis-Jared-Lander.pdf

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. *Handbook of Research Design in Mathematics and Science Education*, 591-646. Mahwah, NJ: Lawrence Erlbaum Associates.

McNamara, A., (2015). Bridging the gap between tools for learning and for doing statistics. Doctoral dissertation, University of California, Los Angeles.

Makar, K., Rubin, A. (2009). A Framework for Thinking About Information Statistical Inference, *Statistics Education Research Journal,* 8(2), 82-105.

Molyneux, J., Johnson, T., McNamara, A., Nolen, S., & Tangmunarunkit, H. (2016). The mobilizR package. Available via https://github.com/mobilizingcs/mobilizr

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, D.C., USA: National Academy Press.

Nolan, D. and Temple Lang, D. (2010) Computing in the Statistics Curricula. *The American Statistician*, 64(2), 97-107

Oceans of Data Project (2016). Building Global Interest in Data Literacy: A Dialogue. Workshop Report. *The Oceans of Data Institute* and *IBM*. Accessed via http://oceansofdata.org/our-work/building-global-interest-data-literacy-dialogue-workshop-report.

Ong, C., Dockterman, D., La Torre Matrundola, D., Griffin, N., & Hansen, M. (2015). Year 5 Mobilize Project Evaluation, Fall 2016 Report. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ong, C., Dockterman, D., La Torre Matrundola, D., Griffin, N., & Hansen, M. (2016 forthcoming). Year 6 Mobilize Project Evaluation, Fall 2016 Report. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Park City Math Institute (PCMI) Undergraduate Faculty Group (2016). Curriculum Guidelines for Undergraduate Programs in Data Science. *In preparation.*

Pruim, R., Kaplan, D., Horton, N., Creativity, M., & Minimal, R. (2015). Mosaic: project MOSAIC statistics and mathematics teaching utilities. *R package version 0.10. 0.*

R Core Team (2016). R: a language and environment for statistical computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, In., Boston, MA URL http://www.rstudio.com/

Tangmunarunkit, H., Hsieh, C.K., Longstaff, B., Nolen, S., Jenkins, J., Ketcham, C., Selsky, ju., Alquaddoomi, F., George, D., Kang, J., Khalapyan, Z., Ooms, J., Ramanathan, N., Estrin, D., (2015). Ohmage: A General and Extensible End-to-End Participatory Sensing Platform. *ACM Transactions on Intelligent Systems and Technology*, 6(3), Article 38 (April 2015), 21. DOI: http://dx.doi.org/10.1145/2717318

Wellman, K. (2010). Pizza Girl: Statistical Analysis of a Delivery Shift: Part 1. slice.seriouseats.com, http://slice.seriouseats.com/archives/2010/04/statistical-analysis-of-a-pizza-delivery-shift-20100429.html. Accessed October 18, 2016.

Whitaker, D., Foti, S., and Jacobbe, T (2015). The Levels of Conceptual Understanding in Statistics, (LOCUS) Project: Results of the Pilot Study", *Numeracy*, Vol. 8: Iss.2, Article 3., DOI: http://dx.doi.org/10.5038/1936-4660.8.2.3

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-248.