# AN ANALYSIS OF SECONDARY TEACHERS' REASONING WITH PARTICIPATORY SENSING DATA

ROBERT GOULD
*University of California, Los Angeles*
*rgould@stat.ucla.edu*

ANNA BARGAGLIOTTI
*Loyola Marymount University*
*Anna.Bargagliotti@lmu.edu*

TERRI JOHNSON
*University of California, Los Angeles*
*terri.johnson@stat.ucla.edu*

## ABSTRACT

*Participatory sensing is a data collection method in which communities of people collect and share data to investigate large-scale processes. These data have many features often associated with the big data paradigm: they are rich and multivariate, include non-numeric data, and are collected as determined by an algorithm rather than by traditional experimental designs. While not often found in classrooms, arguably they should be since data with these features are commonly encountered in daily life. Because of this, it is of interest to examine how teachers reason with and about such data. We propose methods for describing progress through a statistical investigation. These methods are demonstrated on two groups of secondary mathematics teachers engaged in a model-eliciting activity centered around participatory sensing data.*

*We employ graphical depictions of discrete Markov chains to describe the strategic decisions the teachers follow while analyzing data, and find that this descriptive technique reveals some suggestive patterns, particularly emphasizing the importance of frequent questioning and crafting productive statistical questions.*

***Keywords:*** *Statistics education research; Big data; Modeling; Model-eliciting activity; Secondary education; Professional development*

## 1. INTRODUCTION

While we may live in the age of the "data deluge," most classroom statistics curricula are very much centered around pre-deluge data. The data that impact our daily lives, such as data collected when we shop online, data used to monitor our health and physical activity, or data that are used to recommend books to read or music to download are mostly absent from classrooms. There have been some efforts to correct this. For example, the CATALST Project (Garfield & Zieffler, 2012) used a large public dataset on airline arrivals as well as familiar contexts such as examining the shuffle feature of a digital music player. Other examples of this sort that have helped data to be accessible to classrooms are the inclusion of Twitter feeds in StatCrunch (West, 2016) and the tools to easily import simple HTML tables into both StatCrunch and Fathom (Finzer, 2016). Still, the inclusion of data

such as these in the secondary classroom is the exception, not the rule. If the goal of a statistics curriculum is to teach formal statistical inference, then by necessity, data must come from random samples or employ random assignment. This goal is both admirable and necessary, but it is far from sufficient for students who are coming of age in a data-driven economy. Instead, in addition to formal inference, we need to teach students how to deal with and extract patterns from big messy data.

***Participatory sensing and IDS*** The *Mobilize Introduction to Data Science* (*IDS*) curriculum was designed for secondary school students to develop a blend of computational and statistical thinking skills applied to data from a variety of contexts and types, in particular data collected in *participatory sensing* "campaigns." Participatory sensing (PS) is a data collection paradigm designed to create communities centered around both collecting and analyzing shared data (Burke et al., 2006). We use the term "campaign" to refer to the entire process of collecting data via participatory sensing, including choosing a topic, crafting survey questions, collecting data, and then analyzing and interpreting the data. PS data include many characteristics associated with big data, and one goal of the curriculum is to prepare students to reason with data that do not easily fit into a random sampling paradigm.

One particular challenge facing the implementation of *IDS* is preparing teachers to deliver the curriculum in their secondary classrooms. In the United States, the Common Core State Standards for Mathematics (CCSSM), adopted by 42 states, places increased emphasis on statistics in K-12 (Common Core State Standards Initiative, 2010). However, teacher preparation in statistics lags. Although the importance of statistics is noted in the K-12 curriculum, the Mathematical Education of Teachers II (Conference Board of the Mathematical Sciences, 2012) and the Statistical Education of Teachers (SET) report (Franklin et al., 2015) both note that statistics is a large area of need for teacher preparation and professional development. The CCSSM is geared towards teaching statistical inference and does not explicitly address the challenges of understanding the data students frequently encounter or generate in their daily lives.

The *IDS* curriculum was one component of a larger project, the *Mobilize* project, which was funded by the National Science Foundation to develop secondary students' computational skills in the context of data. The project was created as a partnership with the Los Angeles Unified School District (LAUSD), the nation's second-largest district, and the curriculum was co-developed with the teachers and staff from LAUSD along with computer scientists and statisticians from the University of California, Los Angeles.

*IDS* is intended to be situated within a secondary school mathematics curriculum, with *Algebra I* as a prerequisite for the course. It was conceived as one course of a computer science (CS) sequence that would follow an introductory CS course such as *Exploring Computer Science* (www.exploringcs.org) and would precede either *Advanced Placement Statistics* or *Advanced Placement Computer Science* (or both). In LAUSD, students who successfully complete *IDS* satisfy the "*Algebra II* requirement" for admission to the California public university system. Thus, *IDS* serves as one alternative path around *Algebra II*, a course that has received increasing attention for its unintended side effect of decreasing the numbers of minority groups who are eligible for university study (Burdman, 2015).

The *IDS* curriculum was supported by a software suite developed by the *Mobilize* team that allows students to carry out participatory sensing campaigns (Tangmunarunkit et al., 2015). The software facilitates the use of students' mobile devices to collect and transmit data and provides a multivariate visualization tool called the *dashboard* (https://sandbox.mobilizingcs.org/#demo) (see Figure 1). In a PS campaign, students act as

human sensors and, like sensors, collect data as determined algorithmically by "triggers." For example, in the Trash Campaign, the trigger event is throwing away an item of trash. When a student throws something away, she collects data using her mobile device. The data consist of survey questions about the type of trash (Is it recyclable, compostable?), the type of receptacles available (Do you see recycling bins? How many?), as well as a photo of the trash, which in turn activates an automatic data collection of the date, time, and location of the event. Data can be viewed either through the dashboard or downloaded in a comma-separated file.

Because teaching programming basics is one objective of the *IDS* curriculum, students learn to analyze data using the statistical programming language R via the RStudio interface (RStudio Team, 2015). Their coding is facilitated by the *mobilizR* package (Molyneux, Johnson, McNamara, Nolen, & Tangmunarunkit, 2016), a package developed by the *Mobilize* technology team that unifies R syntax and creates "wrapper functions" so that some useful complex operations can be accomplished in fewer steps. The package *mobilizR* is based on the R package *mosaic* (Pruim, Kaplan, Horton, Creativity, & Minimal, 2015).

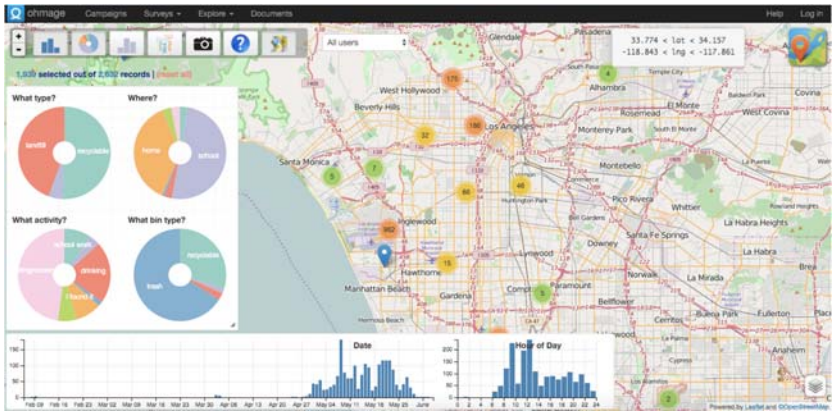


*Figure 1. Mobilize dashboard display of the Trash Campaign*

Data collected via PS cannot be treated as if they came from random samples and yet can provide a rich and detailed picture of activities for the community involved in the campaign. An important goal of the *Mobilize* project is to get students to find the detailed patterns within the data and tell meaningful stories with the data. However, it is equally important for students to understand the limitations of these data, particularly since they cannot be used to support generalizations to larger populations through the application of traditional statistical inferential techniques.

To prepare teachers to implement the *IDS* curriculum, a year-long professional development course was carried out. Teachers met for two three-day workshops in the summer and a series of five one-day workshops during the academic year. These sessions taught the statistical conceptual content, as well as the computational content, while emphasizing an inquiry-based pedagogy. Teachers frequently engaged in the same activities their students would later engage in, and afterwards reflected on the activity with respect to the scaffolding and support students might require. Teachers were taught to analyze data using R via RStudio by engaging in the same lab activities as their students. Additional support to learn R was provided via an online community and email listserv.

Given the unique structure of these data, we wondered how teachers would reason when confronted with multivariate data of multiple forms that did not fit a random sampling

paradigm. In particular, this paper proposes descriptive methods that allow an examination of teachers' progress during a statistical investigation. This description is based on the data cycle, an idealized depiction of the stages of the statistical investigation process. Graphical displays allow visualizations of progress temporally, and a Markov chain model of transitions between stages of the teachers' investigations affords both graphical and quantitative descriptions of progress.

## 2. THEORETICAL BACKGROUND

For the most part, statistics curricula have focused on preparing students for statistical inference. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) K-12 framework, for example, describes progressive developmental levels A, B, and C, in which level C is defined by its inclusion of inference (Franklin et al., 2007). To support statistical inference using standard mathematical tools, the data-gathering process must adhere to sometimes rigid formalisms. For example, observations must be randomly sampled or collected through randomized experiments, or must consist of objects measured in a way that supports a formalized notion of variability (for instance, when repeated measurements are understood to vary due to random error). Such data are the product of more than a century of mathematical development, and have played - and will continue to play - a pivotal role in scientific discovery. However, this focus on inference is not sufficient, given that the data that students encounter in their daily lives does not easily fit into a statistical-inferential framework.

Data currently play a bigger role in our culture and economy at all levels than at any time in history. Such "everyday" data, as well as data that are likely to be encountered in the workplace, do not play a big role in education because they do not always meet the strict format required for statistical inference. Data scraped from websites, data produced by social networking (such as Twitter and Facebook), data produced by gaming devices and smartphones, and data streamed from satellites that are used to understand climate change -- these all fall under the general heading of big data, and understanding how to analyze such data is arguably more relevant to students' present and future than limiting instruction to data from random samples (Gould, 2010; McNamara, 2015).

Originally, the term "big data" meant just that: datasets of such great size that special computational tools were required to store, access, and analyze them. What became known as the "three V's" (Volume, Variety, and Velocity) were used to describe such data, and IBM added a fourth: Veracity (IBM, 2015). Over time, the meaning of this term has become more inclusive, so that Lane, Stodden, Bender, and Nissenbaum (2014) can now refer to "big data" as a "paradigm" (p. 1). Data included in this paradigm are not necessarily big in terms of the number of observations but have characteristics that often lead to great size. These characteristics differentiate them from traditional classroom data.

Data collected through a participatory sensing campaign fit into this big data paradigm. Participatory sensing data need not be limited to "numbers with context" (Moore, 1990), and may consist of a "Variety" of data types including text, images, sounds, locations, and dates. Like much big data, participatory sensing data are not collected through a random sampling protocol, and this poses challenges to educators whose prior approach to data analysis always led to a confidence interval or hypothesis test.

### 2.1. TEACHING DATA ANALYSIS

The term "data analysis" is another term that, *prima facie*, seems obvious and yet remains somewhat ill-defined. Tukey provided an expansive definition that deliberately

included all of statistics (Huber, 2011; Tukey, 1962). And yet the term is often still used to distinguish it from the more mathematical and inferential aspects of statistics. The purposes of data analysis are to find patterns in data, to communicate patterns, to suggest hypotheses, to evaluate modeling assumptions, and to answer questions and refute hypotheses.

Data analysis need not include statistical inference but might instead consist of informal statistical inference, which, as described by Makar and Rubin (2009), includes expressions of uncertainty using probabilistic language. Or inference might take the form of generalizations to a larger universe through implicit or explicit assumptions about the relationship of the sample to the universe. According to DeLeeuw (1994), "Statistical techniques sometimes use probability, and sometimes they don't" (p. 5). This statement was motivated by the application of statistical models in the social sciences in which probabilities may not be interpretable, and reminds us that statistics and data analysis have an important role to play even when formal inference is not suitable or required.

Teaching data analysis is particularly important with big data, because although the data might not fit into a random sampling framework, one still wishes to find "insight rather than quantifiable results," to use statistician Peter Huber's phrase (2011, p. 3). In addition, many big data can be described as "opportunistic," meaning they are being analyzed for a purpose that might be different than the purpose for which they were originally gathered (Huber 2011, p. 43). For example, one might use Twitter data to understand how people felt about the 2016 Olympic Games (www.kdnugges.com/2016/08/rio-olympics-twitter-sentiment.html). Finding insight in this context means being able to rephrase a possibly vague research proposal into precise questions that can be addressed by the data at hand.

This study was motivated by the need to provide the teachers (and their students) with an understanding of the ways that statistical questions could be used with opportunistic data. Teachers in the *Mobilize* project expressed (and demonstrated) that, when confronted with a dataset with multiple variables, they often did not know where to begin.

As Huber (2011) pointed out, few books teach how to think strategically about finding insight through data analysis, possibly because "you learn it [data analysis] on the job, by apprenticeship, and by trial and error" (p. 2). The statistics profession as a whole, Huber claimed, lacks a framework for discussing strategic approaches to data analysis. Our solution was to involve teachers in statistical modeling in a situation where formulating questions to set up a solution strategy was essential.

## 2.2. STATISTICAL MODELING

Our notion of statistical modeling is analogous to the notion of mathematical modeling processes, described by Lesh and Doerr (2003) as the processes students develop and use during their efforts to solve a real-world problem. These can be described as cyclic processes by which learners must develop and use mathematical tools to represent, understand, and solve real-world problems by translating a real-world problem into mathematics, working it out, and then translating it back into a real-world context (Gravemeijer, 2004). Doerr and English (2003) described models as "systems of elements, operation, relationships, and rules that can be used to describe, explain, or predict behavior of some other familiar system" (p. 112).

The mathematical modeling process as defined by the CCSSM is a cycle with an initial stage of "Formulate." In this stage, the modeler translates the real-world problem into a mathematically tractable phrasing (see Figure 2). In our project, we use a similar cycle, which we called the data cycle (see Figure 3). The data cycle is itself adapted from the GAISE K-12 description of the "statistical investigative process." This process listed four stages in this order: "Formulate Questions," "Collect Data," "Analyze Data," and "Interpret

Results" and shares with the CCSSM cycle the notion that problems must be formulated into new language, in this case as questions that are answerable by analyzing data. The statistical investigative process itself is closely related to the PPDAC cycle (Problem, Plan, Data Analysis, Conclusions) proposed by Wild and Pfannkuch (1999) as a framework for statistical thinking.

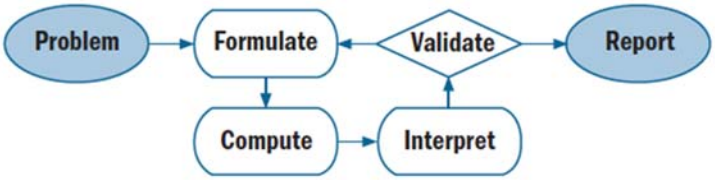


*Figure 2. The modeling cycle as presented by the CCSSM (Common Core State Standards Initiative, 2010, p. 72)*

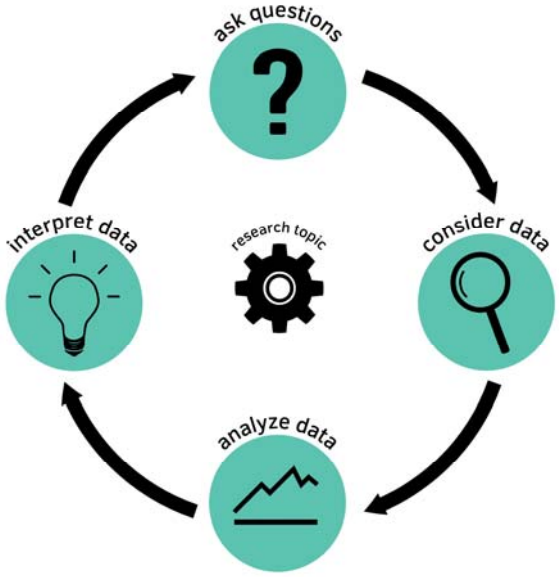


*Figure 3. The data cycle as presented in the IDS curriculum*

The GAISE formulation sits well with an inferential paradigm, but requires some modification to allow for opportunistic data. The *Mobilize* project replaced the "collect data" stage with the more inclusive label "consider data." This accounts for the possibility that although collecting data may not be required, the data on hand may need to be better understood for its suitability and quality for the task underway.

Although the arrows in the data cycle suggest an ideal pathway through this cycle, in practice analysts might jump between various stages or alternate between two stages for a while before moving on. For example, Arnold (2013) suggested that when working with secondary or opportunistic data, the investigation might begin with the data stage rather than the question stage, since a beginning step is to interrogate the background of the provided data to determine how it might be used to answer the research question. One can imagine a series of transitions back-and-forth between these two stages, as the analyst refines her question to adjust to strengths or weaknesses in the available data.

## 2.3. STATISTICAL QUESTIONS

An important tool when strategizing an analysis is the "ask questions" stage of the cycle. This tool is not commonly taught. Classroom exercises often begin with a question, and the student's job is to apply the appropriate analytical technique. When working with data in the big data paradigm, however, the choice of statistical question, a question that can be addressed with the data at hand, steers the direction of the analysis and determines the story that is told.

Arnold and Franklin (2017) presented criteria and an accompanying rubric in order to assess the quality of a statistical question. The criteria include whether the variables are clear, whether the population of interest is clear, whether the question can be answered with data, whether the question is worth investigating, and whether the question allows for analysis to be made of the whole group. In their paper, Arnold and Franklin also differentiated between different types of statistical questions—mainly investigative questions defined as those that are posed, and interrogative questions defined as those that are asked during data analysis. Both of these types of questions are important.

When designing or guiding a statistical investigation with a class, teachers need to be able to pose interesting investigative and interrogative questions. However, there is a small body of literature that has begun to expose difficulties that teachers and students have when posing statistical questions (Arnold, 2008; Burgess, 2007). In particular, when articulating statistical questions, teachers must be mindful of posing questions that foster statistical thinking instead of eliciting mathematical reasoning (Bargagliotti & Groth, 2016). In general, teachers who are mathematically trained seem to have difficulties in posing good statistical questions (Groth, 2007; Groth & Bergner, 2007).

## 2.4. MODEL ELICITING ACTIVITIES

To engage students (and teachers) in this modeling process, we developed a Model-eliciting activity (MEA). MEAs (Lesh, Hoover, Hole, Kelly, & Post, 2000) are designed to elicit participants' thought processes while they engage in an open-ended problem-solving session. MEAs have been used in engineering education (Hamilton, Lesh, Lester, & Brilleslyper, 2008), mathematics (Lesh & Doerr, 2003), and statistics (Garfield & Zieffler, 2012). Hamilton et al. (2008) defined an MEA as

> a problem that simulates authentic, real-world situations that small teams of 3-5 students work to solve over one or two class periods. The crucial problem-solving iteration of an MEA is to express, test, and revise models that will solve the problem (p. 4).

Our MEA is described in Section 3.2 and details are provided in the Appendix, including a brief description of a solution.

Initially, our objective was to use the data cycle (Figure 3) as a template to describe teachers' progress through analysis of data from the big data paradigm and to understand whether and how they employed statistical questioning. In some sense, though, we were "lucky" that of the three groups of teachers engaged in the MEA, one was much less successful than the others. This raised the question of whether we could see differences in how the teachers engaged with the data, and whether, by tracking progress through the data cycle, we could see differences in their tracks.

Within this context, we pose two research questions:

(1) How can the pathways and transitions in the statistical modeling process within the context of the data cycle be described?

(2) How can the role that questions and questioning play in a statistical investigation be characterized?

## 3. DATA AND METHODS

### 3.1. PARTICIPANTS

The participants were nine credentialed mathematics teachers who teach in LAUSD. These teachers were teaching *IDS* during its initial year of pilot-testing. All teachers represented schools at the lower end of the socioeconomic spectrum (79% of students are classified as below poverty level across LAUSD). At the participating schools, 90% of the *IDS* students were Hispanic, 4% African-American, 2% White, 2% Pacific Islander, 1% Asian, and 1% Native American. Half of the teachers reported having 6-10 years full-time teaching experience, and four had more than 15 years. The teachers had little-to-no experience with, or background in, data analysis. However, two had taught *AP Statistics*, and all had implemented the *Mobilize Algebra I* curriculum (a three-week module using participatory sensing and emphasizing linear models for 14- and 15-year-olds) at least once in the prior two years. The teachers were collaborators in this pilot year of the program, and were active in revising and evaluating the *IDS* curriculum.

Professional development for the *IDS* teachers 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 day-long sessions held on Saturdays throughout the year. The MEA was conducted during the fourth of the five sessions and occurred roughly two-thirds of the way through the academic year of the first year of the *IDS* course. Teachers worked in self-determined groups of three on this MEA for 45 minutes, a time-period that approximated one class period for most of the teachers' classrooms.

Teachers were paid a stipend for attending professional developments and for assisting in the development of the curriculum. All participants provided informed consent as required by the UCLA Internal Review Board.

### 3.2. OUR MODEL ELICITING ACTIVITY: THE TRASH CAMPAIGN

Our MEA began by presenting participants with a news article about "America's largest landfill site" (Gutierrez & Webster, 2012) (which is the primary site for Los Angeles County) and a link to the website for this landfill (www.lacsd.org). The MEA then asked participants to write a letter to Los Angeles County in which they were to suggest "two specific steps the public can take to reduce the use of landfills" and to support these recommendations with evidence. Data collected from the PS Trash Campaign were available to complete the MEA. The Trash Campaign had been carried out by Los Angeles area high school biology students and their teachers, who recorded data on their mobile devices every time they threw away an item over a five-day time period. Multiple classrooms were combined over a one-month time period. The students and teachers who collected the data signed waivers to allow for public use of the data, and the data were anonymized by removing names and perturbing values by adding a small amount of random "jitter" to numeric values. All but one of this study's participants had prior experience specifically with the Trash Campaign approximately one year before the MEA took place.

Participants were provided with the raw data file from this campaign, approximately 2600 observations of 17 variables, as well as access to the "dashboard" data visualizer with the data pre-loaded. Both are accessible via https://sandbox.mobilizingcs.org/#demo/. The

variables consist of a variety of types: categorical (which type of bin was the 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). In addition, some survey questions were coded into two variables, each in a different format. For example, the location of the trash is available in one variable as a numerical key (0 through 6) and in another variable as a categorical value (school, home, etc.).

The problem posed in the MEA requires regarding the data as opportunistic, since they were not collected specifically to answer questions about the county's landfill use. The set of variables provided and the data collection scheme do not match those of a well-designed, random sample-based study. Although the problem statement requires making conclusions about a universe beyond the sample at hand, the lack of a random sample means that generalizations to the larger population or claims about causality had to be based on types of reasoning other than inferential reasoning. In general, we expected their reasoning and analysis to be guided by personal knowledge of recycling and landfills. For example, one might reason that if more recyclable goods were put in recycling bins, the burden on the landfill would decrease. This might lead to exploring the data for the percentage of recyclable goods that are put into trash cans. Although the PS data would serve as a poor estimate of this percentage for all people in the county, it still serves as evidence of whether a problem does or does not exist.

For software, the participants were free to use either the "dashboard," an interactive data visualizer with the data pre-loaded, or RStudio with the *mobilizR* package installed. The raw data and the dashboard visualizer are accessible via https://sandbox.mobilizingcs.org/#demo/. Prior to the beginning of the school year, teachers had no experience with RStudio, and learning the software was a significant part of the professional development meetings. Teachers had experience using the dashboard prior to the MEA as they had been teaching the *Mobilize Algebra I* curriculum the previous year.

### 3.3. METHODS

We conducted a qualitative analysis of two of the three groups of teachers engaged in the MEA. Two of the three groups were classified as "successful" because they were able to suggest a solution to the landfill issue supported with data. Due to constraints, we were able to analyze just two of the groups and so chose, in order to observe the greatest amount of variability, one of the successful groups ("Group 1" throughout) and the unsuccessful group ("Group 2"). The choice of which successful group to study was made randomly, via an electronic coin flip. Groups were videotaped during the activity and transcripts were produced from the videos and indexed using Inqscribe (https://www.inqscribe.com/). Studiocode, which has since been renamed Vosaic (https://vosaic.com/), was used to produce a timeline of events in the videos. Codes were assigned to each "turn" taken by a speaker in the conversation. Usually, this turn consisted of spoken statements or questions, but could also be actions carried out on the computer. These codes were assigned to indicate a group's location within the data cycle ("Ask Questions," "Consider Data," "Analyze Data," and "Interpret Data"). The assignment was not blind to whether statements were from Group 1 or 2. The category "Other" was used to capture actions or dialogue that did not fit into the data cycle categories. In addition, units that were coded as "Ask Questions" were further refined into two groups: "Statistical Questions" and everything

else. In the analysis presented here, the "Statistical Question" stage refers to an instance in which the group asked a statistical question, and the "Ask Question" stage refers to any other type of question posed by the group (See Figure 4). To illustrate, consider the following two statements made by teacher Vivian at time 13:29 and teacher Justin at time 13:42.

| | |
|---|---|
| Vivian (13:29) | Because here, we want to know what happens at home, I would like to know what happens at home and at school. So, we want to know how many of our students at home, they're probably trashing, umm, recyclable items. |
| Justin (13:42) | Can we do that in one plot though? Or would it have to be in two different plots? If we're talking about itemizing against the trash bin. [Justin wants to see a list of items in trash bins at home and compare to a list at school]. |

Vivian's turn was coded as Statistical Question and Justin's was coded as Analyze Data.

When groups used RStudio, we were also able to capture their code, match it to the transcript, and reconstruct the results. Later in our analysis we separated the sub-category of "Ask Questions: Statistical Question" as a separate stage of the data cycle so that we could observe the relationship between asking statistical questions and the other stages of the cycle.


*Figure 4. Studiocode codes for the data cycle*

The codes were applied to the videos of each group by two of the researchers independently. Initially these two researchers determined codes based on their judgment as to which stage of the cycle, if any, best described the conversation turn of the team. The third researcher then acted as a referee to make decisions when there was disagreement between the first two researchers. Whenever the referee made a decision, it was discussed with the group to see if all three researchers could come to consensus. In all cases, consensus was reached. On a few instances, this discussion led to a consensus decision to apply a code that none of the three initially had chosen. For Group 1, the interrater reliability, measured as percent agreement, was 83%. On four occasions, even though the initial two coders had agreed on a label, after discussion, the label was changed. On three occasions in which the initial two coders disagreed, a third label different from either's first choice was decided upon, after discussion. Group 2, the unsuccessful group, was considerably more difficult to code. The interrater reliability at the first pass was 35%. Two times the initially agreed code was changed upon discussion with the referee, and 24 times in which the initial codes disagreed, a third option was eventually chosen. Particularly because of the low reliability, the final codes were achieved with consensus after discussion including all three researchers. While the interrater reliability in Group 2 was too low to draw strong conclusions about their success with the statistical investigation, the reliability of the coding is not the focus of this paper, which instead describes an analytic approach to facilitate comparisons of investigations.

***How can we describe the pathways and transitions of the data cycle?*** A graphical representation of the resulting codes for each group at every second was created in order to visualize the groups' transitions over time. (See Figures 5 and 6; A more complete analysis of these graphics is provided below.) Such depictions provide a qualitative sense

of the transitions of the modeling process. To quantify transitions from one stage to the next, we computed transition probabilities, treating the processes as if they were a discrete time Markov chain. The model was fit using the package *markovchain* (Spedicato & Kang, 2016) in R (R Core Team, 2016). Markov chains are stochastic processes in which transition to a new stage depends only on the current stage. A transition matrix contains, in entry *ij*, the probability of moving to stage *j* given that the investigation is currently in stage *i*, or in other words: P(move to stage *j* | currently in stage *i*).

In this analysis we use these probabilities in a purely descriptive fashion and for this reason they are better interpreted as proportions than probabilities. (One reason for our caution in calling them probabilities is that the sample size is too small for us to perform basic statistical tests). Thus, "the probability that the group transitions to the Analyze Data stage given it is in the Interpret Data stage" should be interpreted as "the proportion of the times in which the group was in the Interpret Data stage and moved to the Analyze Data stage."

Codes were applied to participants' turns in the conversation, and these turns were associated with times (the number of seconds into the activity). The results presented in the analysis section of this paper are based on a five-second unit of time, although in order to check the robustness of our analysis, we computed probabilities using three different methods: using each second as the fundamental unit, using five-second intervals as a unit, and using each conversation turn as a unit (the initial coding unit). To compute the five-second transitions, the one-second sequences were grouped, sequentially, in five-second units. The majority code label in each interval was assigned to the entire interval. Thus, if the first three seconds were coded Ask Questions and the next two were Consider Data, the entire five-second interval was coded as Ask Questions. Essentially, this smooths over the one-second time units and so produces fewer rapid transitions. The choice of time-unit does affect the value of the calculated probabilities, but does not affect our analysis at the descriptive level, and we felt the five-second unit provided a more interpretable picture of the data cycle (analogous to the way in which a histogram with bins wider than one-unit can present a clearer picture of the underlying distribution). The code and data files for this analysis are available on request from the corresponding author.

***What is the role of statistical questions?*** Statistical questions facilitate the pathways and transitions along the data cycle. In order to assess the quality of statistical questions, a list was made of all statements coded as "Statistical Question" within the "Ask Questions" phase. Statements that were fragments of questions or explanations of questions were removed from the list. For the remaining questions, each author independently scored each question using a four-point rubric. (Scoring was not blind to the group labels). The rubric was based on four of the six criteria for a strong investigative question as described by Arnold (2013, p. 110-111). (Two of these six criteria were determined to be not directly relevant to our investigation). The relevant criteria for assessing the strengths of statistical questions for this activity were (1) the variable(s) of interest is/are clear and available; (2) the population of interest is clear; (3) the question can be answered with the data; and (4) the question is worth investigating, is interesting, and has a purpose. For each of the criteria, the rater scored 0 or 1, with a 1 indicating that the criterion was satisfied. Each statistical question was evaluated by adding the four scores, so that each question was rated from 0 (none of the criteria satisfied) to 4 (all four criteria satisfied). Summary statistics of these ratings, and also the percentage of questions that satisfied each particular criterion, were then computed for each group.

For this portion of the study, all three researchers rated the questions independently. The majority rating was applied. Table 1 shows the interrater reliability for each of the four items on the scale.

*Table 1. For each of the four criteria, the percentage of statistical questions on which the three independent raters agreed unanimously; each question was scored a 0 or 1 on each criteria*

| | variables clear | population clear | data | worthwhile |
|---|---|---|---|---|
| Group 1 | 65% | 41% | 88% | 100% |
| Group 2 | 27% | 73% | 27% | 45% |

## 4. ANALYSIS

### 4.1. PATHWAYS AND TRANSITIONS WITHIN THE DATA CYCLE TEMPLATE

How does an analysis of opportunistic data from the big data paradigm, in particular data collected through participatory sensing, proceed? While we would not expect the participants, nor anyone engaging in a statistical investigation, to follow the data cycle exactly as ordered in the diagram, we were interested in determining which pathways were followed by the teachers and whether particular pathways could be described that were associated with success. Both groups we examined were attempting the same investigation with the same data, and so one would expect similarities. On the other hand, because the two groups had different results (one group succeeding in suggesting an evidence-based proposal, the other not), it is natural to look for differences and hypothesize that those differences might indicate reasons for success and failure. We provide several approaches to examining these pathways.

Figures 5 and 6 color code the transcripts according to the data cycle stage the participants were in. In the figures, every box is equivalent to 1 second of time. In Figure 5, we see that the successful group, Group 1, began with questions and considering data. Group 2, the non-successful group, took longer to get into the investigation and began by considering the data and then asking questions. Another apparent difference is that Group 1 spent considerably more time in interpretation, and this was primarily at the end of their investigation.

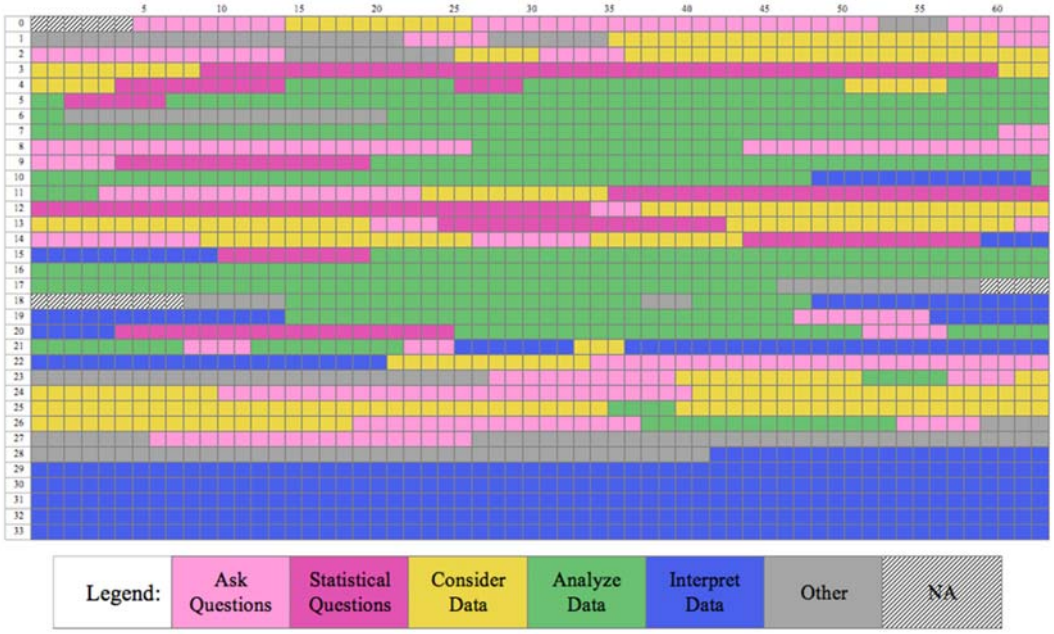


*Figure 5. Graphical representation of Group 1's sequential movement through the data cycle: each square represents one second of the video, and the color represents the coding applied during that time span; each row represents a minute and each column a second within that minute; the "NA" was coded for a few trailing seconds of one of the researcher's instructions to the group*

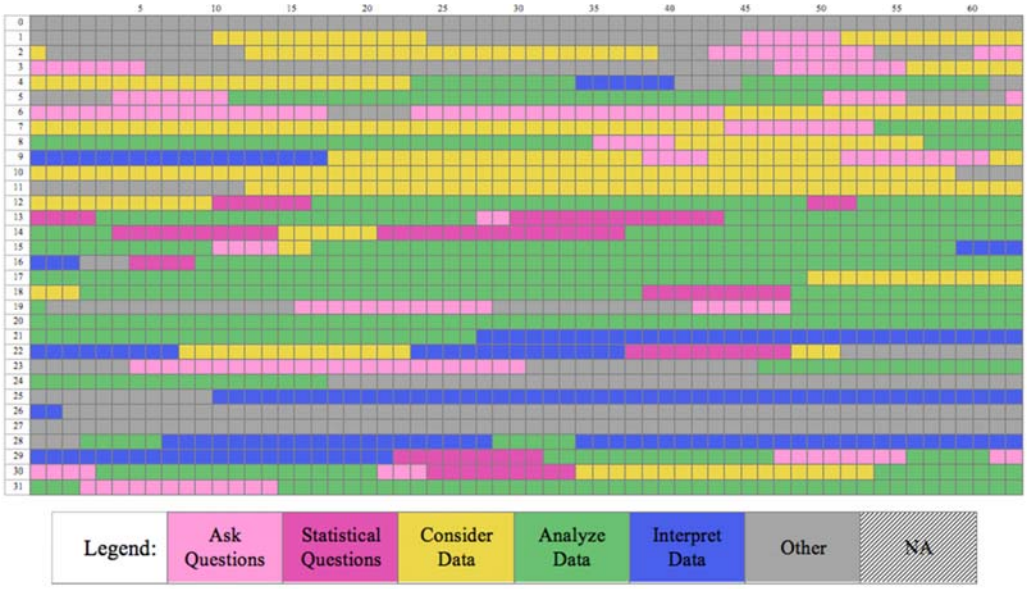


*Figure 6. Graphical representation of Group 2's sequential movement through the data cycle*

Table 2 summarizes the amount of time the groups spent in each stage. For both groups, Analyze Data was the stage they spent the most time in. Group 1 spent more time than Group 2 asking questions and interpreting data. (But note that Group 1 spent two minutes longer overall, and so direct comparisons should be interpreted cautiously).

*Table 2. The amount of time, in minutes, each group spent in each stage*

| | Ask Questions | Statistical Questions | Consider Data | Analyze Data | Interpret Data | Other |
|---|---|---|---|---|---|---|
| Group 1 | 4.9 | 3.3 | 5.3 | 9.5 | 7.4 | 3.3 |
| Group 2 | 3.3 | 1.4 | 6.0 | 10.6 | 3.4 | 7.5 |

As noted in Section 3.3, our analysis is based on five-second intervals. Table 3 shows the frequency of five-second intervals spent in each of the stages for the two groups.

*Table 3. The number of five-second intervals each group spent in each stage*

| | Ask Questions | Statistical Questions | Consider Data | Analyze Data | Interpret Data | Other |
|---|---|---|---|---|---|---|
| Group 1 | 62 | 41 | 63 | 111 | 88 | 39 |
| Group 2 | 40 | 17 | 66 | 128 | 44 | 89 |

Next, we examine the transitions between stages of the data cycle. The network graphs in Figure 7 visualize the transitions for Group 1 (left) and Group 2 (right). Several features stand out. First, in both groups, the most likely motion, given that the analysis is in stage *i*, was to remain in stage *i* (e.g., if the group was in the Consider Data state, they were most likely to stay in the Consider Data state). This is true for both groups and all stages and reasonably so, since one team member is likely to build on what another has said or done and in doing so, keeps the group in the same stage of the data cycle. The transition matrices on which these figures are based are included in the Appendix (Section 6.4).

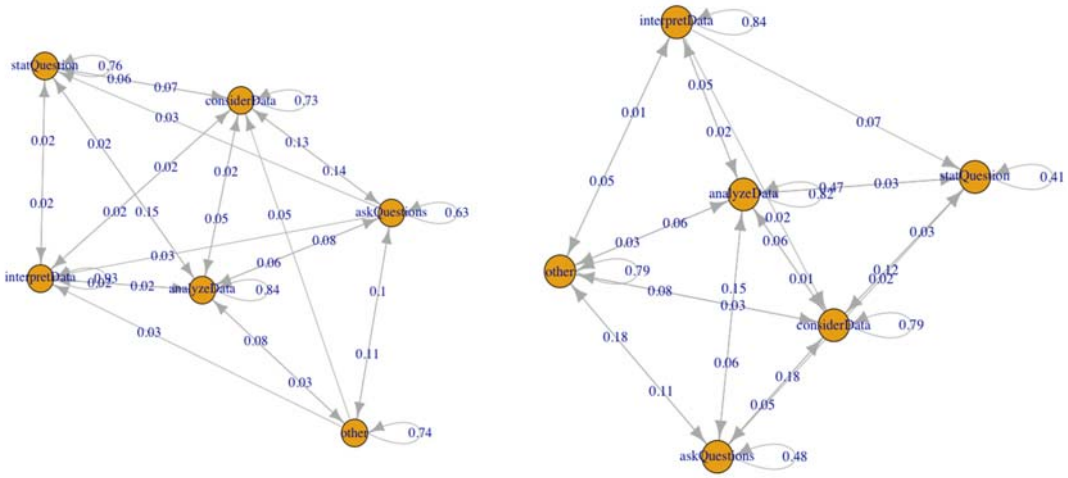


*Figure 7. The network graphs show the transitions between stages of the data cycle for Group 1 (left) and Group 2 (right); the transition rates from stage i to stage j are printed closest to stage j*

To illustrate these transitions, consider this sequence provided by Group 1, which starts at the beginning of a Statistical Questions stage, moves to Consider Data, back to Statistical Questions and then to Analyze Data.

| | |
|---|---|
| Michelle (3:11) | So then do you want to do maybe, if there's more trash produced by where they are [when they dispose of the trash]. Like, by location? Do we want to? |
| Rosie (3:22) | [Or] we could [consider if there's more trash produced by] activity level. |
| Michelle (3:24) | Wait. What are the questions we're trying to ask, I guess? We're trying to make plots based on that. |
| Rosie (3:27) | Well it [the MEA] says to give two suggestions, right? But I think that there are things that we need to know. Like when is most trash produced? |
| Michelle (3:37) | Like when, what time, or where? [Naming three other possible means by which the amount of trash might vary]. |
| Rosie (3:39) | Like in what circumstances, so *where* [the name of a variable], and [doing] what activity? |
| ... | ... |
| Rosie (3:50) | And then the availability of recycling bins and trash bins in relationship with *where*. |
| Michelle (3:57) | So we want, ok, so let me...so we're gonna...let me set up...[begins to look at variables list on computer] |

Not including this last sentence, the team stayed in the Statistical Questions phase as they gradually refined the question. At 3:57, Michelle transitioned to the Consider Data stage when she examined the variables available on her computer. The team then transitioned back to the Statistical Questions stage:

| | |
|---|---|
| Rosie (4:06) | So I'm interested, I'm interested in knowing how many recycle bins are around. |
| Michelle (4:11) | So the typical number of recycle bins? |
| Rosie (4:12) | Yeah. |
| Michelle (4:15) | So then, let's do umm, histogram...tilde...ummm, number of recycle bins. |

The last sentence transitioned the group from the Statistical Questions stage to the Analyze Data stage. In this sentence, Michelle spoke out loud the command she was typing on the computer. The command to produce the histogram is `histogram(~numberRecycleBins, data=trash)`, and the result is shown in Figure 8.

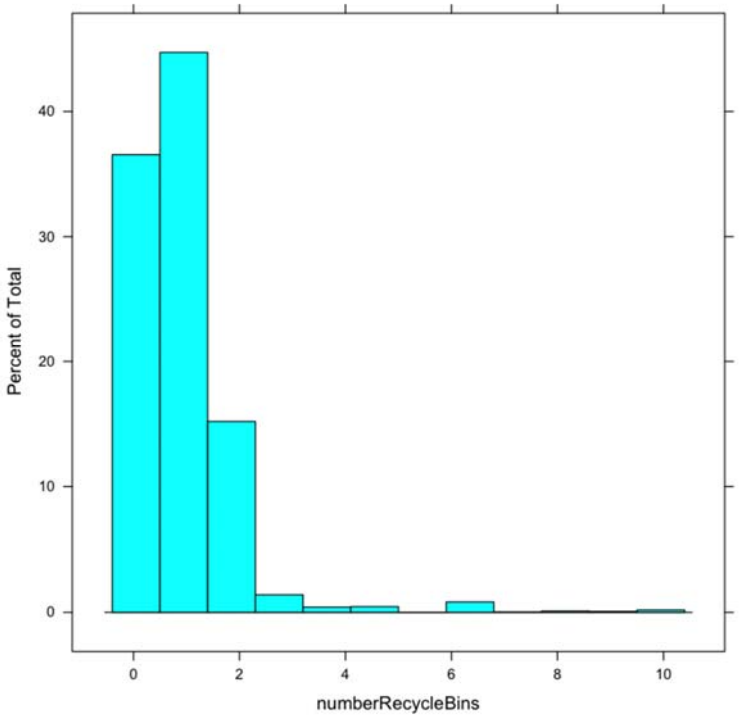


*Figure 8. The histogram produced in the Analyze Data stage to address the statistical question, "What is the typical number of recycle bins?"*

At this point, the group could have transitioned to the Interpret Data stage by answering their statistical question (it looks like the mean number of bins is about 1), but instead they returned to the Statistical Questions stage, because Rosie realized that this graphic would not help them see whether the typical number of bins varies by location (the *where* variable). And thus they continued to refine their question.

In the above sequence, Group 1 spent most of the time in the statistical questioning stage. This was not unusual behavior for Group 1. Table 4 shows the rate of staying in each state for each group. We see that Group 1 more frequently remained in both a questioning stage (either Ask Questions or Statistical Questions) and the Interpret Data stage than did Group 2.

*Table 4. The rate of immediately repeating the current stage (given in first column); for example, of all of the times that Group 1 was in the Asking Questions phase, they were still in that phase in the next time period in 63% of the cases*

| | Group 1 | Group 2 |
|---|---|---|
| Ask Questions | 0.63 | 0.48 |
| Stat Questions | 0.76 | 0.41 |
| Consider Data | 0.73 | 0.79 |
| Analyze Data | 0.84 | 0.82 |
| Interpret Data | 0.93 | 0.84 |
| Other | 0.74 | 0.79 |

At first glance, the complexity of the network graphs in Figure 7 might suggest that progress about the Data Cycle was arbitrary. In fact, particular transitions occurred more often than others and certain transitions did not occur at all. These 0-rate transitions are indicated by arrows that point in a single direction, which means that transitions were made in one direction but not the other, or by the lack of any arrow at all, for example, the lack of an arrow between Ask Questions and Interpret Data for Group 2. Table 5 shows transitions that did not occur in either group, 0-rate transitions that are unique to Group 1, and that are unique to Group 2.

*Table 5. Transitions that never occurred: I represents the Interpret Data stage, SQ Statistical Questions, AQ Ask Questions, CD Consider Data*

| Shared | Group 1 Only | Group 2 Only |
|---|---|---|
| I to AQ | CD to Other | AQ to I |
| SQ to AQ | I to Other | CD to I |
| SQ to Other | | SQ to I |
| Other to SQ | | |

All four of the "roads not taken" shared by both groups involve questioning and three of the four involve statistical questioning. In particular, the Statistical Questions stage and Other stage are isolated from one another. (The Other stage included a variety of activities that could not be classified as belonging to one of the data cycle stages). To some extent, these shared one-way paths provide some insight into the role of questioning in these statistical investigations; while groups moved back and forth between other states, questioning states were more directed.

The most notable difference between the groups in terms of their one-way pathways is that Group 2 has a high number of "blocked" pathways to the Interpret Data stage. This is consistent with the fact that Group 2 spent relatively little time in that stage: 3.4 minutes compared to Group 1's 7.4 minutes. While Group 1 could engage in interpretation at almost any point in the process (although most often following a question phase, see Table 6), Group 2 only engaged in interpretation after Analyze Data or "Other."

Having considered which transitions do *not* occur, we turn now to considering which transitions were most frequent. Table 6 shows where, for a given stage of the data cycle, each group was most likely to have been prior to that stage. The table lists the first- and second-most frequent prior stage. For both groups, considering data was most often preceded by asking questions, both statistical questions (with transition rate 0.07 for Group 1, 0.12 for Group 2) and non-statistical questions (0.13 and 0.18). This suggests a linkage between considering data and asking questions, as one might expect when analyzing secondary data. This linkage was particularly pronounced for Group 1. For this group, not only was considering data most frequently preceded by questioning, but questioning was most frequently preceded by considering the data.

*Table 6. For each current state in the first column, we show the next two most likely states for Group 1 and 2 (excluding a return to the current state); parentheses show the transition rates (for example, the first entry, "CD (0.14)", indicates that for Group 1, transitions from CD to AQ happened in about 14% of the transitions from CD)*

| Current State | Group 1 most likely came from | | Group 2 most likely came from | |
|---|---|---|---|---|
| Ask Question (AQ) | CD (0.14) | O (0.10) | O (0.11) | AD (0.06) |
| Statistical Question (SQ) | CD (0.06) | AQ (0.03) | I (0.07) | AD (0.03) |
| Consider Data (CD) | AQ (0.13) | SQ (0.07) | AQ (0.18) | SQ (0.12) |

| | | | | |
|---|---|---|---|---|
| Analyze Data (AD) | SQ (0.15) | O (0.08) | SQ (0.47) | AQ (0.15) |
| Interpret (I) | AQ (0.03) | O (0.03) | AD (0.05) | O (0.01) |
| Other | AQ(0.11) | AD (0.03) | AQ (0.18) | CD (0.08) |

Questioning plays a role in another commonality: for both groups, analysis (AD) was most often preceded by statistical questioning (0.15 for Group 1, 0.47 for Group 2). The fact that Group 2 so frequently preceded their analyses with statistical questions is interesting (47% of their transitions from SQ were to AD, compared to 15% for Group 1), since, as we now show, the group's first attempt at analysis floundered because it was not guided by a statistical question.

By comparing how the groups began their investigations, we gain some insight into the strategic importance of the link between questioning and analyzing. In the excerpt below, we see that Group 1 began by formulating a statistical question and soon reached the Analyze stage. Group 2, we will see, began their analysis without a question, and floundered. At the point of this excerpt, Group 1 was about two and a half minutes into the MEA. Their discussion focused on the handout that described the survey questions and their corresponding variable names in the data set (See Appendix A.1). Note that they had already loaded the data into RStudio at this point. (This sequence immediately preceded the sequence presented earlier and ended where it began).

| | |
|---|---|
| Michelle (2:26) | OK so let me understand where these data came from. |
| Ryan (2:31) | We have to write a letter, right? That's our goal? |
| Rosie (2:36) | So maybe we should just make some plots. |
| Michelle (2:40) | Wait, wait. Who collected this data? |
| Interviewer (2:44) | (reminding them of earlier discussion) It was a Mobilize survey, a Mobilize campaign. |
| Michelle (2:47) | And then, where did this, so these people were in different places and this [gestures to handout that gives descriptions of data] is like where they were and when they...it's not necessarily at home, it's just... |
| Ryan (3:03) | It says *where*. School, home or...or malls. [Referring to values of the *where* variable on the handout]. |
| Rosie (3:10) | Well you can list variables. |
| Michelle (3:11) | So then do you want to do maybe, umm, if there's more trash produced by where they are. Like, umm, by location? Do we want to? |

Michelle began this segment by trying to understand who generated the data and how. Rosie then suggested they start by making some plots (beginning an analysis of the data), but Michelle wanted to wait until she understood the data better. Instead, they worked on formulating a statistical question and only then did they do analysis.

Group 2 took a different approach. Their initial conversation was about recycling in general, and their own observations about why people did or did not recycle. After a minute or so they went to the Los Angeles County website and read information provided there about landfills. After about 2 minutes, they decided to begin looking at the PS data. Whereas Group 1 opened up RStudio and loaded the data, Group 2 instead decided to begin with the Dashboard, an interactive data visualization tool that had the data preloaded. Once

it was open, they looked at which variables were available. As they were looking, they interacted with the graphic by clicking on the automatically-provided displays (see Figure 9).

| | |
|---|---|
| Vivian (3:50) | So, let's see...recyclable...[They each clicked on the dashboard to see different conditional distributions of the data. For several seconds, they clicked]. |
| Vivian (4:13) | So, hours of the day... |
| Vivian (4:23) | Are you look at the, will you check the landfill? [She means the proportion of trash classified as belonging in the landfill]. Like around noon, it's getting the most trash. They probably were eating lunch at that time. At home... |

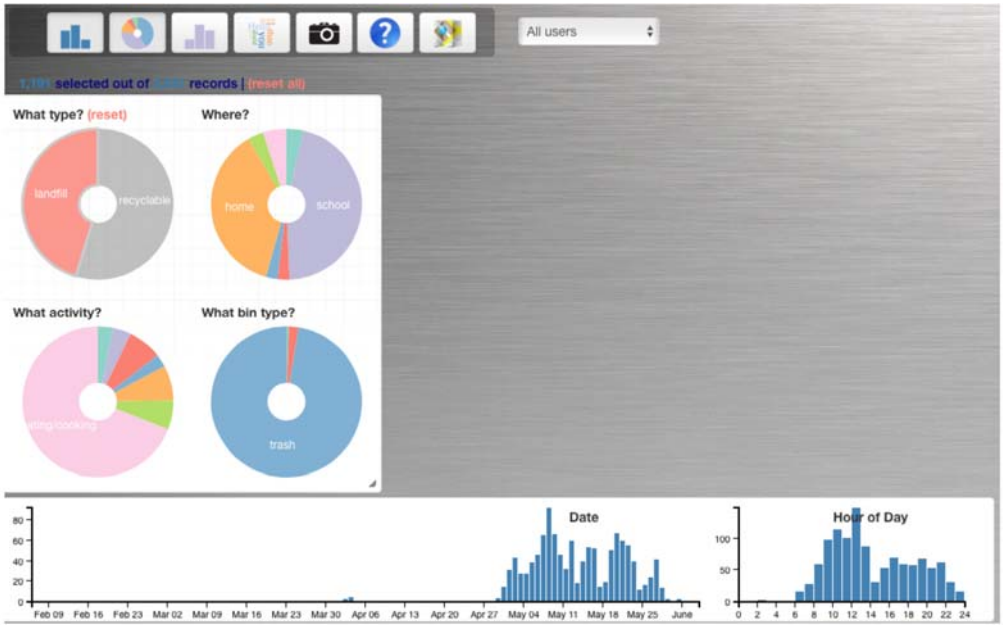


*Figure 9. The dashboard display as seen by Group 2 near the beginning of their analysis. After clicking on the "landfill" value in the "What type" variable display (upper left donut chart), the other values are greyed out, indicating that the data are subset to include only observations for which the type of item thrown away was "landfill"; looking at the distribution of hours of the day (lower right histogram), Vivian observed that noon is the most frequent time that landfill is thrown away*

Group 2 did what Group 1 almost did: they began by looking at distributions of the data, but with no clear purpose in mind. Initially, they were in the Consider Data stage and were learning about what variables were available and what values they had. Then they transitioned directly to the Analyze stage, but with no clear direction in mind.

The dashboard is designed so that if an analyst clicks on a given value in any of the displays, the other displays update to show only the cases that have that value in common. In Figure 9, the "landfill" value in the "*What type?*" display was selected. And therefore the other displays showed distributions only for cases in which the items were recyclable. In trying to interpret the displays, Justin found what he thought was a problem, but the others possibly did not understand what the problem was or did not believe his

interpretation. (In fact, from the transcript it is not clear whether he was confused about the interpretation of the displays or about the meaning of the variables). Group 2 decided that the dashboard would not be helpful and moved to RStudio. This sequence lasted about seven minutes. Their analysis seemed aimless given that there was no guiding statistical question, but possibly helped them better understand the variables the data contained. However, it also served to delay progress towards reaching a solution to the MEA.

### 4.2. ROLE OF STATISTICAL QUESTIONS

While Group 1 differed from Group 2 in the amount of time it spent formulating statistical questions, both groups followed similar pathways to and from the Statistical Question stage. However, there were differences in the quality of the statistical questions asked. When compared to Group 2, Group 1 was more often clear about the variables involved, more often posed questions that could be answered with the data, and more often posed questions judged to be "worthwhile" (See Table 7). Group 1's mean quality score (the average of the sum for each question) was 2.9 (out of 4.0, s = 0.2) compared with Group 2's mean of 2.0 (s = 0.8).

*Table 7. The percentage of statistical questions that satisfied each criterion: Were the variables clearly stated? Was the population of interested specifically stated? Was the question answerable with the data provided? Was the question worthwhile? Group 1 asked 17 complete statistical questions; Group 2 asked 11*

| | Group 1 | Group 2 |
|---|---|---|
| Variables clearly stated | 94% | 54% |
| Population specifically stated | 0% | 0% |
| Answerable with data | 100% | 82% |
| Worthwhile | 100% | 36% |

Interestingly, none of the groups identified the population or the group of interest in their statistical questions. There are two possible reasons for this. One is that the MEA gave them a population (people in Los Angeles County). Another is that, because these questions were posed in the context of an active problem-solving session, the question posers were probably less likely to be very precise and formal in language. Still, more precision would have been helpful, particularly for Group 2, which sometimes struggled to match conceptual terms to variable names. For example, one question "Just what is in the trash bins?" could have been answered by referring to the *type* variable (landfill, recyclable, or compost), or the *whatTrash* variable (written descriptions of the items thrown away). Justin often interpreted their goal as being to analyze the *whatTrash* variable, while it seems possible that the other members of the group were trying to focus on the *type* variable. But because they did not phrase a question explicitly in terms of the variable names, they had trouble communicating.

## 5. DISCUSSION AND CONCLUSION

Given the emphasis on data in society today and the widespread collection and availability of big data, it is important that we consider how these data can be brought into the classroom. Students need opportunities to work with such relevant data in order to show relevance of statistics in their daily lives. As such, teachers need to be comfortable using

big data themselves in interesting activities and in turn help their students see the utility of data analysis.

For those who believe the goal of statistical analysis is inference, big data might seem, somewhat ironically, to be too simplistic for classroom use. One might reason that without random sampling, the data do not lend themselves to teaching about confidence intervals or hypothesis tests. In this paper, we explored methods for examining statistical investigations with such data. Analyzing these data requires substantial intellectual skills and conceptual understandings, and it is vital that we prepare teachers to equip students with these skills and understandings.

### 5.1. DISCUSSION

***Describing the data cycle*** We employed a statistical model, discrete Markov chains, as a means for describing the participants' progress through the data cycle. As a descriptive technique, this model proved revealing. We clarified that progress through the cycle is not orderly, in the sense of cleanly moving from one stage to the next in a clock-wise fashion. We did not expect it to be too orderly, of course, but were still surprised by the variety of pathways followed. We were also able to identify commonalities that were consistent with the strategic approaches we expected. Both groups had strong links between the questioning stages and the Consider Data stage. And both groups tended to precede their analyses by questioning. (Although, as noted in the analysis section, Group 2 in fact began their first Analyze stage without a statistical question).

The differences were also striking. Group 1 had more pathways (more transitions with non-zero transition rates) than Group 2, suggesting either greater fluidity in moving about the stages of the cycle or a lack of coherence to their investigation. However, this fluidity or lack of coherence was marked by a potential to engage in interpretation after every stage (see Figure 7). For Group 2, the interpretation stage was reached only after analysis about 5% of the time (and, less often, after "Other"). Group 2's behavior is closer to the idealized clockwise motion around the cycle, and yet Group 1 was more successful at suggesting a solution to the MEA. One explanation is that Group 1 was more focused on producing a solution, and so often made interpretations of results (once they had done some analyses) in order to try these out as potential solutions. The bulk of Group 1's time in the interpretation stage came at the end of the time allotted on the MEA. This suggests a deliberate strategy of keeping the end goal in mind, and setting aside time at the end to interpret and synthesize the analyses into a result.

One advantage of employing a statistical model such as the discrete Markov chain model is that, in principle, it allows for formal hypothesis tests to compare transition rates. Unfortunately, such tests require larger sample sizes than we had here. A large sample size in this context means many observed transitions (Spedicato, Kang, Yalamanchi, & Yadav, 2016). Hypothesis tests of basic properties of the discrete Markov chain are often based on an approximate chi-squared distribution, which requires an expected number of roughly five observations for each of the 36 transitions for each of the two groups. Without this large sample size, we had insufficient statistical power to distinguish probabilities within any given row (particularly after adjusting for multiple comparisons). For this reason, it is best to interpret these probabilities in this study descriptively, as proportions, rather than representing true transition probabilities.

***What role do statistical questions play?*** There is no doubt that questions play an important role in data analysis. "Formulating a question can be a useful way to guide the exploratory data analysis process and to limit the exponential number of paths that can be

taken with any sizeable dataset" (Peng, 2016). This role is particularly important when dealing with data such as those generated by participatory sensing, since the relationship (if any) between the existing variables with the research objectives is not immediately clear, and since there are too many possible relationships to consider to allow for an exploration that includes considering all possible relationships.

Although both groups used questioning to drive their analyses (as seen by the relatively strong linkages between statistical questioning and analysis and by the fact that the Statistical Question stage was the only stage separated from Other activity), Group 1 appeared to be more successful at posing questions as indicated by the results in Table 7. One reason is perhaps that they spent more time doing so. Group 1 spent 8.2 minutes combined in questioning, compared to 4.7 minutes for Group 2. Group 1's statistical questions were of greater average quality than Group 2's.

As explained earlier in the paper, we expected to see a relationship between the questioning stages and the Consider Data stage, since this is a necessity when dealing with opportunistic data (or any secondary data source). Both groups did in fact have frequent transitions between the questioning stages and the Consider Data stage. The fact that these links were stronger (i.e. the transitions occurred more often) for the more successful group is suggestive of the importance in closely relying on the data to formulate statistical questions and to use questions to look for shortcomings or strengths in the data.

While the design of this study does not allow us to conclude unequivocally that Group 1's success was due to their stronger question-posing and stronger links between questioning and considering data, because these behaviors are aligned with common statistical practice (Peng, 2016), it seems prudent that educators should pay close attention to developing questioning skills. Students are known to have difficulties writing good questions (Allmond & Makar, 2010; Burgess, 2007; Pfannkuch & Horring, 2005). The GAISE report (Franklin et al., 2007) recommends that initially, teachers pose statistical questions for their students and then students can develop their own questions to guide their data cycle. However, in this study, we have found evidence that some teachers struggle with the task of posing statistical questions, and so we emphasize that more attention should be paid to developing this skill amongst teachers as well as students. We emphasize that posing statistical questions is a non-mathematical activity; thus teachers who are mathematically trained may have difficulty developing questions to guide their progress through the data cycle. With strong questions, teachers should be taught that the analysis and interpretation phases in the model should then focus on answering the posed questions.

## 5.2. CONCLUSION

As the course title Introduction to Data Science (IDS) in the Mobilize project suggests, an important goal is to teach students a combination of computational and statistical thinking. Although there may be controversy in the academic world over the meaning of the term "data science," for our purposes we take it to mean the science concerned with finding meaning in data, with particular emphasis on those data that fit into the big data paradigm. Little, if any, research has been done to forge learning progressions for data science at early ages. IDS is one attempt at exploring this integration of computation and statistics at the secondary level.

Our MEA emphasized the role of exploratory data analysis, a role with growing importance in data science, using a non-traditional data set. We observed that such analysis is far from trivial, and challenges experienced teachers. The analysis involves an iteration between understanding the data (what variables does the dataset contain, what values do they have, and what do these values represent) and formulating questions. The analysis

also involves a complex interplay between analysis, interpretation, and questioning. That not all of the groups were successful emphasizes the need for making explicit strategic approaches to understanding data.

In this paper, we suggest a possible methodology to analyze engagement with the data cycle while working with non-traditional data sets. We first suggest modeling the data cycle undertaken in a visual display. A visual, like Figures 5 and 6, can shed light on larger-scale patterns of behavior and help understand how the investigation evolves over time. A more fine-grained analysis can then be carried out by examining the transitions from each of the states in the data cycle. We suggest visualizing the transitions using network graphs and then subsequently examining the common and not occurring pathways. Transitions from one stage to the next can be modeled as transition probabilities, treating the processes as if they were a discrete time Markov chain. This systematic method of analysis can shed light on engagement with the data cycle.

More generally, the methodologies discussed would be applicable to settings in which approaches to process are specified and desired. For example, as mentioned and shown in Figure 1, the CCSSM presents a modeling process that fits these criteria. The analytic approaches presented in this paper may shed light on whether students and teachers do in fact engage in such a process. While our interests were in examining how teachers reason with data from the big data paradigm, the approach could also be applied to more traditional data in which investigators might be forming and testing hypotheses. A potential advantage to the methodologies presented is that they provide a way to quantify behaviors, specifically fluid behaviors that transition in and out of different states. A limitation of the proposed work is the reliance on the need to identify discourse with particular states as it is often difficult to parse out and separate statements.

In this study, we viewed teachers' analysis of participatory sensing data through the lens of the data cycle. Despite the acknowledged limitations of the study, this allowed us to identify strategic pathways followed by the teachers. While the data cycle is itself not new as a description of data analysis – it is closely derived from Wild and Pfannkuch's (1999) PPDAC cycle and the GAISE four-steps to the statistical investigation process – our analysis using Markov chains and time-plots of the stages of the data cycle allowed us particular insights into common transitions made between the different stages of the cycle. In particular, we noted the surprising complexity of the modeling process for these teachers, as well as the strategic importance of statistical questions.

## ACKNOWLEDGMENTS

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

We would like to acknowledge the support and work of the Mobilize team: LeeAnn Trusela (project director), Derek Chau (LAUSD), Joanna Goode, Heidi Estevez, Lynn Kim-John, John Landa, Jane Margolis, James Molyneux, Suyen Moncada-Machado, Steve Nolen, Maria Olivaries, Jeroen Ooms, Letician Perez, Jody Priselac, Hongsuda Tangmunarunkit, Kenia Tello, Linda Zanontian.

We also thank the reviewers and the associate editor for thoughtful and insightful edits and comments.

## REFERENCES


Allmond, S., & Makar, K. (2010). Developing primary students' ability to pose questions in statistical investigations. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010).* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots8/ICOTS8_8A1_ALLMOND.pdf

Arnold, P. (2008). *What about the P in the PPDAC cycle? An initial look at posing questions for statistical investigation.* Paper presented at the Eleventh International Congress of Mathematics Education, (ICME-11), Monterrey, Mexico. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.214.9407&rep=rep1&type=pdf

Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* (Doctoral dissertation). Retrieved from University of Auckland Research Repository - ResearchSpace. (Identifier: http://hdl.handle.net/2292/21305)

Arnold, P., & Franklin, C. (2017). *What makes a good statistical question?* Manuscript in preparation.

Bargagliotti, A., & Groth, R. (2016). When mathematics and statistics collide in assessment tasks. *Teaching Statistics*, *38*(2), 50-55.

Burdman, P. (2015). Degrees of freedom: Diversifying math requirements for college readiness and graduation (Report 1 of a 3-part series). Oakland, CA: *LearningWorks and Policy Analysis for California Education, PACE*. Retrieved from Institute of Education Sciences ERIC collection. (ERIC Number: ED564291).

Burgess, T. (2007). *Investigating the nature of teacher knowledge needed and used in teaching statistics* (Doctoral dissertation). Massey University, North Palmerston, New Zealand.

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

Common Core State Standards Initiative (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices & Council of Chief State School Officers. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Conference Board of the Mathematical Sciences (CBMS) (2012). *The mathematical education of teachers II*. Providence, RI and Washington, DC: American Mathematical Society and Mathematical Association of America.

DeLeeuw, J. (1994). *Statistics and the sciences*. Department of Statistics, UCLA. Retrieved from http://escholarship.org/uc/item/46b4s8m3

Doerr, H., & English, L. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education, 34*(2), 110-136. doi:10.2307/30034902

Finzer, W. (2016). *Fathom: Dynamic statistics software* (Version 2.4) [Computer software]. Retrieved from http://www.fathom.concord.org

Franklin, C., Bargagliotti, A., Case, C., Kader, G., Scheaffer, R. & Spangler, D. (2015). *The statistical education of teachers*. Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/SET/SET.pdf

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*. Alexandria, VA: American Statistical Association.

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*, 297-315.

Gravemeijer, K. (2004). Local instruction theories as means of support for teachers in reform mathematics education. *Mathematical Thinking and Learning, 6*(2), 105-128.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427-437.

Groth, R. E., & Bergner, J. A. (2007). Teachers' perspectives on mathematics education research reports. *Teaching and Teacher Education*, *23*(6), 809-825.

Gutierrez, T., & Webster, G. (2012). Trash city: Inside America's largest landfill site. Retrieved from http://www.cnn.com/2012/04/26/us/la-trash-puente-landfill/

Hamilton, E., Lesh, R., Lester, F., & Brilleslyper, M. (2008). Model-eliciting activities (MEAs) as a bridge between engineering education research and mathematics education research. *Advances in Engineering Education*, *1*(2), 1-25.

Huber, P. J. (2011). *Data analysis: What can be learned from the past 50 years*. Hoboken, NJ: Wiley.

IBM (2015). *The four V's of big data*. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). Part I: Conceptual Framework. In J. Lane, V. Stodden, S. Bender & H. Nissenbaum (Eds.), *Privacy, big data, and the public good: Frameworks for engagement* (pp. 1-3). New York: Cambridge University Press.

Lesh, R., & Doerr, H. (2003). Foundations of a models and modelling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 3-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. In A. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 591-646). Mahwah, NJ: Lawrence Erlbaum Associates.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82-105. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85

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

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

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

Peng, R. D. (2016). *Exploratory data analysis in R*. Retrieved from https://leanpub.com/exdata

Pfannkuch, M., & Horring, J., (2005). Developing statistical thinking in a secondary school: A collaborative curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular developing in statistics education: International Association for Statistical*

*Education (IASE) Roundtable,* Lund, Sweden, 28 June-3 July 2004, (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.

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*. [Computer Software.]. Retrieved from https://cran.r-project.org/web/packages/mosaic/index.html

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, Inc.* [computer software]. Boston, MA. Retrieved from http://www.rstudio.com/

Spedicato, G.A., & Kang, T.S. (2016). *Markovchain: Discrete time Markov chains made easy. R package version 0.6*. Retrieved from https://cran.r-project.org/web/packages/markovchain/index.html

Spedicato, G.A., Kang T.S., Yalamanchi, S.B., & Yadav, D., (2016). *The markovchain package: A package for easily handling discrete Markov chains in R*. Retrieved from https://cran.rproject.org/web/packages/markovchain/vignettes/an_introduction_to_markovchain_package.pdf

Tangmunarunkit, H., Hsieh, C.K., Longstaff, B., Nolen, S., Jenkins, J., Ketcham, C., Selsky, J., Alquaddoomi, F., George, D., Kang, J. & Khalapyan, Z. (2015). Ohmage: A general and extensible end-to-end participatory sensing platform. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 38.

Tukey, J. W. (1962) The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1-67.

West, W. (2016). *StatCrunch* [computer software]. Pearson Education. Retrieved from http://www.statcrunch.com

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

ROBERT GOULD
UCLA Department of Statistics
Mail Code 951554
Los Angeles, CA 90095-1554
USA

# APPENDIX

## A.1 LANDFILL MEA

News article to prepare the teachers for the activity:

**Trash city: Inside America's largest landfill site** (*excerpt*)

By Thelma Gutierrez, CNN and George Webster, for CNN

Updated 11:10 AM ET, Sat April 28, 2012 http://www.cnn.com/2012/04/26/us/la-trash-puente-landfill/

It's as tall as some of L.A.'s highest skyscrapers, but the only residents here are rats and cockroaches.

Welcome to the Puente Hills Landfill, the largest rubbish dump in America. Over 150 meters of garbage has risen from the ground since the area became a designated dumping site in 1957.

Now, six days a week, an army of 1,500 trucks delivers a heaving 12,000 tons of municipal solid waste from the homes and offices of L.A. County's millions of inhabitants.

"This used to be a dairy farm; a valley filled with cows producing milk. And now it's a geological feature made out of trash," said Edward Humes, author of "Garbology: Our Dirty Love Affair with Trash" -- a book that charts the history of garbage in America.

Humes says most of the waste arrives straight from the bins of local residents.

"If you're like most of us -- most Americans -- you're making seven pounds of trash a day. Across a lifetime that adds up to 102 tons of trash per person," he said.

In 2010 alone, Americans accumulated 250 million tons of garbage, and although recycling in the U.S. has increased by 34% since 1960, Humes believes the country's attitude to waste is still not sustainable.

"It's very convenient to roll your trash to the curb every week and have it disappear, but it's a magic trick -- and really there's not very much magic," he said. "We need to have less packaging; use less disposable items; (use) things that last longer; make purchasing decisions that are more studied and less wasteful."

The environmental impact of landfill sites varies depending on how well they're managed and resourced. However, typical problems include the contamination of soil and groundwater from toxic residues; the release of methane, a greenhouse gas produced during the decaying process that is more potent than carbon dioxide; and disease-carrying pests.

Tom Freyberg, chief editor of industry publication Waste Management World agrees with Humes that we should all be trying to reduce waste and increase the amount we recycle. However, he says it's likely there will always be a need for landfill, and we should applaud those sites that are well managed.

Readiness Questions for the MEA:

**Landfill Readiness Questions**

Directions:

Answer the questions below and discuss your responses with your team members.

1. What types of items belong in a landfill? Give three examples of things you throw away that belong here.

2. The article implies that by recycling more, we can decrease the use of landfills. Why is this?

3. What types of items are recyclable? Give three examples of recyclable items that you use everyday.

4. How does an item that you throw away at school get to either the landfill or the recycling center? Why might items sometimes go to the wrong place?

Trash MEA:

## Landfill Activity

**Background:**

The Los Angeles County Sanitation District (LACSD) would like to reduce their burden on the regional landfills, such as the Puente Hills landfill mentioned in the article. You can learn more about the LACSD by visiting www.lacsd.org and clicking on the "Solid Waste & Recycling" tab.

Because the LACSD knows that your class is familiar with participatory sensing campaigns and data, they are hoping you can help them explore the impact of landfills by using data from a city-wide participatory sensing campaign, titled the "Trash Campaign," that was conducted at a number of high schools in the Los Angeles Unified School District (LAUSD).

**The task:**

The LACSD is planning a public awareness campaign and wants to ask the public to take specific steps that will help reduce the landfill burden. Based on the data collected, they would like you to **make one or two recommendations** that would reduce the use of the regional landfills.

Specifically, they have asked your team to compose a letter in which you answer the following questions:

1. *What is/are the specific recommendation(s) you are proposing for the public awareness campaign?*
2. *Why do you think this will work? What evidence do you have to support this? Include any necessary plots and analyses.*

**The data:**

The survey questions/prompts for the Trash Campaign are provided below for your reference. The data can be found via the MobilizingCS public dashboard (https://lausd.mobilizingcs.org/#demo/) and can also be exported to RStudio.

| Survey Question/Prompt | Variable Name | Data Type |
|---|---|---|
| 1. Please take a photo of your trash. | photo | photo |
| 2. Please describe your trash. | whatTrash | text |
| 3. What type of trash?<br>❑ recyclable<br>❑ landfill<br>❑ compost | type | category |
| 4. Where was this trash generated/found?<br>❑ home ❑ restaurants ❑ in transit<br>❑ school ❑ stores/malls ❑ others<br>❑ work | where | category |
| 5. What activity generated this trash?<br>❑ eating/cooking ❑ cleaning ❑ I found it<br>❑ drinking ❑ shopping ❑ other<br>❑ school work | activity | category |
| 6. Where did you put this trash when you were done?<br>❑ recyclable ❑ compost/green waste<br>❑ trash ❑ litter | receptacle | category |
| 7. How many recycling bins can you see from your location? | howManyRecycle | number |
| 8. How many trash/landfill bins can you see from your location? | numberTrashBins | number |
| 9. How many compost/green waste bins can you see from your location? | numberCompostBins | number |
| AUTOMATIC | location | latitude, longitude |
| AUTOMATIC | timestamp | date, time |

## A.2 BRIEF OUTLINE OF A SOLUTION

There are multiple solutions possible. One approach is to reason that there might be many recyclable items put into the landfill. This could be supported by finding the percentage of recyclable items put into the trash bin within these data. In the letter recommending this approach to the county, we would caution that the analysis is based on a dataset that possibly does not represent the entire county yet, nonetheless, demonstrates that these high school students put 31% of their rejected recyclable items into trash bins and so the items would end up in the landfill. The county should therefore strive to find ways to encourage more people to recycle. (Group 1 suggested an educational campaign).

Another way to encourage more people to recycle would be to place more recycling bins around the county. To demonstrate whether this might be successful, we might consider the distribution of recycling cans in place. If the presence of recycling bins encourages recycling, we should see a lower proportion of recycling goods thrown into trash bins in locations that have a larger number of recycling bins. This was the approach

Group 1 was asking. Some potentially productive statistical questions are: Does the distribution of the number of recycling bins available vary by type of location? What's a typical number of recycling bins? How does the proportion of recyclable items put into the trash vary with respect to the number of recycling bins available?

**A.3 TRANSITION MATRICES**

Tables A.1 and A.2 give the transition rate matrices for Group 1 (top) and 2 (bottom). The entry *ij* represents the relative frequency that the analyst will next be in state *j* given that she is currently in state *i.* For example, the rate of transitioning to state CD if one is currently in AD is 0.018. Transition rates (rounded to three decimal places) are calculated based on a five-second interval. Transitions that were not observed (and so have an estimated probability of exactly 0) are represented with a single-digit, bold-faced 0. Note that in both groups, transitions between statistical questioning and "other" activities did not occur.

*Table A.1 Transition matrix for Group 1*

| | AD | AQ | CD | ID | O | SQ |
|---|---|---|---|---|---|---|
| AD | 0.838 | 0.081 | 0.018 | 0.018 | 0.027 | 0.018 |
| AQ | 0.065 | 0.629 | 0.129 | 0.032 | 0.113 | 0.032 |
| CD | 0.048 | 0.143 | 0.730 | 0.016 | 0 | 0.063 |
| ID | 0.023 | 0 | 0.023 | 0.931 | 0 | 0.023 |
| O | 0.077 | 0.103 | 0.051 | 0.026 | 0.744 | 0 |
| SQ | 0.146 | 0 | 0.073 | 0.024 | 0 | 0.756 |

*Table A.2 Transition matrix for Group 2*

| | AD | AQ | CD | ID | O | SQ |
|---|---|---|---|---|---|---|
| AD | 0.819 | 0.063 | 0.008 | 0.047 | 0.031 | 0.031 |
| AQ | 0.150 | 0.475 | 0.175 | **0** | 0.175 | 0.025 |
| CD | 0.061 | 0.045 | 0.788 | **0** | 0.076 | 0.030 |
| I | 0.023 | **0** | 0.023 | 0.841 | 0.045 | 0.068 |
| O | 0.056 | 0.112 | 0.034 | 0.011 | 0.787 | **0** |
| SQ | 0.471 | **0** | 0.118 | **0** | **0** | 0.412 |