Bringing Complex Data Into the Classroom

Pip Arnold¹ Anna Bargagliotti² Christine Franklin³,⁴ Robert Gould⁵,⁶

¹Karekare Education, Auckland, New Zealand,
²Department of Mathematics, Seaver College of Science and Engineering, Loyola Marymount University, Los Angeles, California, United States of America,
³Department of Statistics Franklin College of Arts and Sciences, University of Georgia, Athens, Georgia, United States of America,
⁴American Statistical Association, Alexandria, Virginia, United States of America,
⁵Department of Statistics, University of California Los Angeles, Los Angeles, California, United States of America,
⁶Center for Teaching Statistics, University of California Los Angeles, Los Angeles, California, United States of America

Published on: Jul 28, 2022

DOI: https://doi.org/10.1162/99608f92.4ec90534

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
**Column Editor’s Note:** The final column in our series on the new GAISE report looks at standards and activities for the most advanced level of data science and statistics students. The authors provide lesson plans and ideas for high school teachers, with the goal of encouraging the use of complex data sets—so common in the modern data scene—in the classroom. Comfort with such data is important for informed citizenship in today’s world. Teachers and students, do you have your own experiences and perspectives to share? Consider submitting a column proposal to "Minding the Future."

**Keywords:** data science education, statistics education, K-12 education, secondary data, data in the news

---

**Introduction**

Statistical and data literacy are essential for citizens in a democracy, in problem-solving and policy development, and in building a data-savvy workforce (Engel, 2017; Franklin & Bargagliotti, 2020; Keller et al., 2020). How do we build a workforce capable of understanding how to work with data? How do we support students to become data- and statistics-literate citizens? How do we build data scientists and statisticians? The answer is: we start young and provide many rich experiences throughout schooling to support their ability to use data and statistics to tell stories.

The American Statistical Association (ASA) and the National Council of Teachers of Mathematics (NCTM) recently released a policy document, *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II: A Framework for Statistics and Data Science Education report* (GAISE II) (Bargagliotti et al., 2020). This report presents a set of recommendations toward statistics and data literacy at the elementary, middle, and high school levels. Now more than ever, it is essential that all students leave high school prepared to live and work in a data-driven world, and the GAISE II report outlines how to achieve this goal. See the Franklin and Bargagliotti (2020) article published in HDSR for an overview.

This article is the final in a series of four introducing the GAISE II report and focuses on Level C recommendations. While the Levels A, B, and C broadly align with elementary, middle, and high school, when students are first introduced to statistics and data science, regardless of age, they are likely to start with Level A ideas building through Level B into Level C. See “Engaging Young Learners With Data” (Perez et al., 2021), published in HDSR, for an introduction to Level A and “Using Photographs as Data Sources to Tell Stories” (Arnold et al., 2021) for an introduction to Level B.

Level C in the GAISE II report presents new challenges for educators and students alike, challenges beyond Levels A and B and also beyond the 2005 GAISE Level C. Data have become ubiquitous, and the understandings and skills required to access and analyze data have grown in number and complexity. The rise of data science has created expectations that students will have some knowledge of, for example, data ‘wrangling,’ how data and algorithms affect our culture and society, and how they should protect their privacy.
and the privacy of others. How different states and districts envision this will vary. One goal of this article is to illuminate these challenges by showing the importance of dedicating ample classroom time to curating, questioning the provenance (documentation of the source and creation process) of data; that is, truly understanding data origins and study design.

At Level C, students understand how to use questioning throughout the statistical problem-solving process and how to use information from data to answer appropriately framed statistical investigative questions. They choose and use appropriate data analysis tools and engage in multivariate thinking. At Level C, students develop a more sophisticated notion of data, integrate technology into their practices, and move toward inference, causality, and predictions.

Rather than provide an overview of Level C in its entirety, in this article we present a level C example, based on data in the news, that focuses on the important issues of data provenance and the complexities of sourcing and considering existing secondary data.

**Multivariate Data in the News: Level C**

Data visualizations are a common feature in news media, and their level of complexity is considerably greater than in the past. Such data visualizations provide a good starting point for a classroom activity to help students think about the data behind the graphics. *The New York Times* and the ASA, in partnership, select a graphical visualization from *The New York Times* and present it in the newspaper’s What's Going on in This Graph (WGOITG) feature section. WGOITG is a moderated forum where students across the United States and around the world can reflect on the data visualization, considering questions such as:

- What do you notice?
- What do you wonder?
- What impact does this have on you and your community?
- What is a catchy headline that captures “what’s going on this graph”?

These data visualizations can provide opportunities for engaging students at levels A, B, and C. At Level C, students should be on their way to becoming ‘healthy skeptics’ of statistical information, and so should be asking questions about the process that brought the data behind the visualization into existence: ‘Where and what data were utilized to create this visualization?’ ‘Why was this visualization created and by whom?’ ‘Who collected the data and why?’ ‘Who funded the data collection?’

In Level C, students will engage in analyzing increasingly complex data sets, and will pose their own statistical investigative questions in order to serve their own research questions. The WGOITG feature can provide a stepping stone to such data sets.
Interrogating the Data Visualization

Data visualizations in the news can provide a starting point for Level C explorations. Students can examine and question the data visualization, they can research the origins of the data that were used to make the visualization, and they can interrogate (question) the data visualization to develop a sense of who, what, when, where, and how. The WGOITG post on September 15, 2021, used data from the American Time Use Survey (ATUS). To explore, students can ask interrogative questions of the data visualization such as:

1. What was the purpose for collecting the data? (Initial investigator’s problem/purpose.)
2. Who collected the data?
3. Who funded the data collection and research?
4. Was the data collected using an observational study or an experiment?
5. Whom was the data collected from?
6. When was the data collected?
7. Where was the data collected?
8. How many people participated?

Some questions may be answered directly from the data visualization, others may require some background research. For the WGOITG (Figure 1) from September 15, 2021, we can establish that (1) the data were collected to examine the way those in the United States spend their time, (2) by the U.S. Department of Labor, Bureau of Labor Statistics, (3) who funded the data collection. We can also establish that (4) the data were collected via a survey (an observational study), (5) from American citizens 15 years and older, (6) in 2019 and 2020 (7) in America. We can’t tell from Figure 1 (8) how many people participated.

![Figure 1. What’s Going on in This Graph (WGOITG) on time spent per day on texting, phone calls, and video chats, exercising, and grooming. The data visualizations compare time spent for 2019 with time spent for 2020.](image)

Furthermore, from the data visualization, we can find out about the different variables (9) represented in the data visualization. For example, we can see that the variables include age, year, and average time per day for three broad activities. Table 1 looks at the interrogative questions for variables, which are treated separately.
Table 1. Interrogating the variables in the data visualization.

<table>
<thead>
<tr>
<th>9. What are the variables being considered?</th>
<th>Age Group</th>
<th>Year</th>
<th>Texting, phone calls, and video chats</th>
<th>Exercising</th>
<th>Grooming</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. What were the survey questions used to collect the data?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. How were the variables measured?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. What are the units of the variables?</td>
<td>categories of years</td>
<td></td>
<td>mins per day</td>
<td>mins per day</td>
<td>mins per day</td>
</tr>
<tr>
<td>13. What are the possible outcomes of the variables?</td>
<td>15–24, 25–44, 45–64, 65+</td>
<td>2019, 2020</td>
<td>0–1,440</td>
<td>0–1,440</td>
<td>0–1,440</td>
</tr>
<tr>
<td>14. What type of variables are they?</td>
<td>Categorical</td>
<td>Quantitative</td>
<td>Quantitative</td>
<td>Quantitative</td>
<td>Quantitative</td>
</tr>
</tbody>
</table>

Note. In the ATUS dataset being discussed, the variable Year only takes on two values, so it might be considered a categorical variable. If treated as a categorical variable, note that it is ordinal. In other words, there is order to the values. In general, time is considered a quantitative variable; however, because in this case the values are so few, it may be treated as an ordinal categorical variable as well.

In addition to getting an understanding of the variables depicted, Level C students can be encouraged to discuss interpretations of the data visualization. For example, the first graphic in Figure 1 illustrates that the amount of time spent per day on texting, phone calls, and video calls increased for all four age categories from 2019 to 2020. Although these conclusions are being drawn based on the graphics, at Level C, good practice is for students to examine the source to fully understand the scope of the conclusions and inference that can potentially be made.
Data Structure

Level C students should understand that visualizations such as these represent multiple variables, are created using computers that rely on algorithms that map data into an interpretable (and high-level) visualization, and are thoughtfully created to convey information in meaningful ways. While many of the visualizations in WGOITG are perhaps too complex for students to recreate themselves (see, e.g., What's Going On in This Graph? | First Vaccinated - The New York Times), they can still ask and answer basic questions about the shape and structure of the data that produced the graph. Many data sets, particularly those ready for analysis by most statistical software, are in a case-attribute (i.e., ‘spreadsheet’) format. Students should be able to identify the variables measured, and conjecture on who or what was measured and how.

What might the data set that produced Figure 1 look like? Table 2 gives example outcomes for the variables (shown in Table 1) for four hypothetical respondents.

Table 2. Sample data structure in spreadsheet form.

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Year</th>
<th>Age Group</th>
<th>Exercising</th>
<th>Texting, etc.</th>
<th>Personal Grooming</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2019</td>
<td>25–44</td>
<td>15</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>2019</td>
<td>15–25</td>
<td>45</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>2020</td>
<td>45–64</td>
<td>90</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>2020</td>
<td>45–64</td>
<td>80</td>
<td>12</td>
<td>18</td>
</tr>
</tbody>
</table>

Note. As shown in Table 1, the units for Exercising, Texting, and Grooming are recorded in minutes per day.

The purpose of such an exercise is not to perfectly recreate the data set, but instead to help students build a mental map from data to graph. Students may find this challenging because they will have to grapple with questions about data collection and provenance of the data.

Not every data visualization is as straightforward as this. Some visualizations in WGOITG are quite complex, and require nested (hierarchical) data structures or are perhaps based on several data sets spread across a database (e.g., What's Going On in This Graph? | 2020 US Census - The New York Times). Even so, students should be encouraged to always think about the structure of the data behind the graph, since such thoughts lead to other important critical questions about data collection and data provenance.
**Data Collection**

By considering the potential structure of the data behind the visualization, questions will be raised about the methods used to collect the data (e.g., Were the same people sampled from year to year?). Before investigating the actual ATUS survey design, students can design their own ‘time use’ survey using their class to collect the same variables as in Figure 1. Considerations include:

- What survey questions should be asked? What effect might different survey questions have on the validity of the resulting visualization?
  - For example, we might ask people to report their memories for the last 24 hours and to judge for themselves how much time was spent.
  - Or we might ask them to pay attention in the next 24 hours and record the amount of time spent.
  - Another approach (closer to the ATUS survey) is to keep a diary and, perhaps every hour, record activities that were taken and the length of time spent doing them.
- Are there other activities students might be interested in in tracking? (e.g., sleep, amount of time spent on devices)
- How will data from individual students be compiled into classroom data? Using an electronic spreadsheet? A poster board? What will the columns and rows represent?
- Do they notice any biases that might creep into the data collection process? Perhaps there is an urge to exaggerate the time spent grooming and exercising and downplay the time spent texting? What might they do, as researchers, to minimize these potential biases?

Overall, students should discuss which data collection decisions they think will lead to more accurate, useful, and practical data.

**Data Provenance**

Today, it should be common practice for students to investigate the source of the data when possible. It is a challenging activity for students to partake in, and one that in the past often has not been a focus in the classroom. Frequently, classroom data is taken for granted and little time is spent understanding its origins.

However, as the accessibility to data has increased drastically, considering the provenance of data within the statistical investigative process necessitates proper time and care in the classroom. Interrogative questions such as those suggested in the section titled “Interrogating the Data Visualization” can be used to understand the origin of secondary data sets and provide a full picture of the scope of the data. Such data documentation is an important piece of good statistical practice.

**ATUS Data Set**

As noted above, the ATUS data is collected by the U.S. Bureau of Labor Statistics (BLS) on a yearly basis. To understand the provenance of the data set, the documentation provided by the bureau can be consulted.
**ATUS Documentation**

Documentation for the data set can be found on the BLS [website](#). The ATUS Survey Documentation consists of several different files, ranging from a user’s guide to variable codes to the ATUS questionnaires (see Figure 2). Students can examine these files to begin to understand the complexity of the data collection.

**ATUS Survey Documentation**

- ATUS User’s Guide (PDF)
- ATUS Coding Lexicons
- ATUS Data Dictionaries
- ATUS Questionnaires
- ATUS Coding Rules Manuals
- ATUS Frequently Used Variables (PDF)
- Changes between the ATUS Single-Year Data Files (PDF)
- Conversion Chart: Fraction of an Hour to Minutes (PDF)
- Census Industry and Occupation Codes
- Handbook of Methods: American Time Use Survey

Figure 2. Ten different documents that are available on the U.S. Department of Labor, Bureau of Labor Statistics website that discuss the data provenance for the American Time Use Survey (ATUS).

Students need to understand that part of good statistical practice involves looking at the data documents to notice and wonder about how the data were collected, what survey questions were used, and how the data were recorded. For example, students can be encouraged to read chapter 1 of the user’s guide and explore sections of the questionnaires that are of interest. From this, and by attending to the interrogative questions in the section titled “Interrogating the Data Visualization” important background information is found about the ATUS. The documentation states the data were collected using an (4) observational study of (5) people in America aged 15 or older. The (2 and 3) U.S. Department of Labor, Bureau of Labor Statistics collected the data (6) during May to December 2019 and May to December 2020. This data collected in the (7) United States was for the (1) purpose of seeing how people use their time, and to track patterns in time use over time. When reading the documentation, students will see that much care is put into collecting good data. Documentation is key to creating good data sets and it reveals the complexity of good data collection.

**Using ATUS in the Classroom**

In addition to considering the ATUS data collection method, teachers and students may want to access and utilize the dataset in their classroom. And herein lies a challenge. While the ATUS data is available and
downloadable at the U.S. Bureau of Labor Statistics and the data.gov websites, preparing it so that it is usable for Level C students in the classroom requires some skill in coding and data wrangling, and also considerable patience. Some teachers might have it within their skillset to download and clean such data for themselves. The rest of us, however, will have to rely on finding curated, rich, and complex data that Level C students should encounter. To help with this, the authors have prepared their own curated version of the ATUS, available in the Supplementary Files of this article. We provide a cleaned version of the full ATUS data set that has over 600,000 observations as well as a dataset of 4,000 people randomly sampled from the full dataset that might be more tangible for the classroom (using the random sample also allows for inferential questions to be explored with the population being the full 600,000 people).

To help facilitate in the classroom the use of the ATUS data and guide the ideas and concepts proposed in this article, the authors have outlined potential lesson and implementation plans. The lesson plan outline provides the necessary background to help a teacher implement WGOITG and use ATUS data in the classroom as well as offers suggestions for teaching pedagogy to use. The lesson plan outline can be found as supplemental material to this article.

**Where to From Here? Next Steps? Recommendations and Discussion**

A useful exercise for Level C students is to pose multivariate statistical investigative questions based upon an examination of the variables provided by the data at hand. In the context of the ATUS data, students will note that the response variables in the visualization—time spent exercising, grooming, and texting—do not appear in the dataset. Instead, each of these high-level activities is represented with more detailed variables. For example, time spent exercising was obtained by adding time for a number of activities from archery to yoga. When students understand that new variables can be created through transformations and combinations of existing variables, a new realm of possibilities opens up. At this stage, students can pose new investigative questions that mirror, and yet go beyond, those provided in Figure 1. For example, using the ATUS data set, students might investigate whether U.S. adults (aged 15 and older):

- with a higher income [tend to] spend more time on leisure activities than U.S. adults with a lower income.
- show gender differences in the time spent doing certain activities, for example, paid work, nonpaid work, or caring for elderly parents.
- with a higher income [tend to] have more children than U.S. adults with a lower income.

To answer these investigative questions, students will need to define the somewhat vague notions of ‘leisure’ using the data available, and will see ways in which reasonable analysis might produce slightly different results.

The ability for students to apply ‘data moves’ to tackle these investigative questions depends to a great extent on the availability of technology. The use of technology is essential to data science, but might pose challenges
in the classroom, raising issues of equity, technical expertise, and more. While the use of technology is imperative for data science, understanding the possibilities granted by using data moves to curate and manipulate data can help students see the possibilities of different statistical investigations that can be undertaken with the same dataset.

Overall, this article emphasizes the process of being introduced to, gaining access to, and understanding a secondary dataset. Using a data visualization in the news as an entry point, Level C students can interrogate the origin of the dataset used to create it and subsequently pose new investigative questions and create new variables. This process illustrates the importance of the Collect/Consider component of the statistical investigative process outlined in the GAISE II report; a component that is often explored in little depth within the classroom. As the access and use of secondary data becomes more and more common, it is particularly important for students to spend time reflecting on the provenance of the data in order to understand the limitations, potential biases, and potential for statistical investigations.

Disclosure Statement

Pip Arnold, Anna Bargagliotti, Christine Franklin, and Robert Gould have no financial or non-financial disclosures to share for this article.

References


---

**Data Repository**


---

**Supplementary Files**

- **ATUS 201920 All Data.xls** 13 MB
- **ATUS 201920 Sample 4k.xls** 3 MB
- **WGOITG Lesson Suggestion.pdf** 1 MB

---

©2022 Pip Arnold, Anna Bargagliotti, Christine Franklin, and Robert Gould. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.