Harvard Data Science Review • Issue 3.4, Fall 2021

Using Photographs as Data Sources to Tell Stories

Pip Arnold¹ Leticia Perez² Sheri Johnson³

¹Karekare Education, Auckland, New Zealand,

²Graduate School of Education and Information Studies, University of California Los Angeles, Los Angeles, California, United States of America,

³Upper School, Mount Vernon School, Sandy Springs, Georgia, United States of America

Published on: Nov 12, 2021

DOI: <https://doi.org/10.1162/99608f92.f0a7df71>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Column Editor's Note: *This Issue's "Minding the Future" continues our introduction to the revised GAISE guidelines. Pip Arnold, Leticia Perez, and Sheri Johnson present lesson plans for engaging "Level B" students (roughly middle school age in the US, 10-14 year olds) in data science through the use of photographs. As a nontraditional but still easily understood data format, photographs give this second level of students a window into what modern statistics and data science work looks like. We hope you will find these lessons inspiring in your own classrooms.*

Keywords: nontraditional data, statistical questions, statistical reasoning, CODAP, technology

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? 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 statistical 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 article](#) published previously in *HDSR* for an overview.

This article is the third in a series of four introducing the GAISE II report and focuses on Level B 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 previously in *HDSR* for an introduction to Level A.

“An important exercise for students at Level B is to become familiar with navigating different data types, not just those stored in static worksheets. Such exercises at Level B focus on students trying to make sense of non-traditional data” (Bargagliotti et al., 2020, p. 63). Traditional data are quantitative or categorical and collected as counts, measurements, or defined categories; however, with more data being generated through digital devices, data now include photographs, videos, sound bites, tweets, and other items. These nontraditional data

allow broader statistical investigations as additional variables and information can be drawn directly from the data itself, for example, the photograph (Bargagliotti et al., 2021).

This article illustrates a series of Level B lessons that use photographs as data sources. The overarching investigative focus was exploring favorite outdoor spaces through photographs. Photographs were collected and then teaching and learning activities were developed in three different ways as lesson plans for the Statistics Education Web (STEW, n.d.). The lesson plans, [*Using photographs as data sources to tell stories about our favorite outdoor spaces*](#), demonstrate how nontraditional data sources such as pictures could be used in Level B.

All three lessons emphasize the statistical problem-solving process and have the spirit of genuine statistical practice. To illustrate further the ideas at Level B, we will show “how to use questioning in guiding statistical reasoning, investigate more sophisticated problems, and look at possible associations between variables and draw comparisons between two groups” (Bargagliotti et al., 2020, p. 44). Each of the lessons show in detail how teachers can scaffold student learning throughout the statistical problem-solving process and specific questions teachers can ask to do this. In this article, we introduce the experiences that students can engage in.

Using Questioning to Guide Statistical Reasoning

The GAISE II report enhances and updates the GAISE I report based on the evolution within the statistical field over the past 15 years. One of the enhancements is questioning throughout the statistical problem-solving process (Bargagliotti et al., 2020). In the act of engaging in the statistical problem-solving process, students both carefully pose questions and spontaneously ask questions. Posing questions is different from spontaneously asking questions. Figure 1 shows the four different question purposes and how they fit within the statistical problem-solving process; each of the four question purposes is then described briefly. Students as novice statisticians need support to understand and use questioning effectively throughout the statistical problem-solving process. Being clear on the different question purposes is a starting point for teachers to support their students.

	Formulate questions	Collect data	Analyze data	Interpret results
Question posing	Statistical investigative questions	Survey/data collection questions		
Question asking	Interrogative questions	Interrogative questions	Interrogative questions Analysis questions	Interrogative questions Analysis questions

Figure 1. Question posing and question asking within the statistical problem-solving process (Arnold & Franklin, 2021, p. 2).

Figure 1 outlines questioning through the statistical problem-solving process. Sometimes students are working with an existing data set and need to familiarize themselves with variables before they pose and answer statistical investigative questions. For example, data scientists often use an internalized set of questions that guide them when working with existing data sets. Students can develop an awareness of how data stories are shaped and formed through opportunities to discuss and explore various forms of questions throughout the investigative process.

Statistical Investigative Questions.

Statistical investigative questions are the questions to which we are seeking an answer. For the given area of exploration, favorite outdoor spaces (see Figure 2 for examples of student-collected photographs of their favorite outdoor spaces), students begin to identify what they desire to explore. They start with a broad area of investigation and through different prompts they formulate specific statistical investigative questions that they will answer using data, thereby telling stories about the class's favorite outdoor spaces. For example, in the [second lesson](#) students watch a [video](#) about the importance of spending time outdoors and then create a drawing of their favorite outdoor space and use this and their photographs to brainstorm statistical investigative questions they can explore. From this starting point, students pose statistical investigative questions (questions we ask of the data) they would like to answer about the class's favorite outdoor spaces. See the third lesson and Arnold and Franklin (2021) for criteria to support the development of good statistical investigative questions.

Figure 2. Examples of student photographs of their favorite outdoor spaces.

Examples of statistical investigative questions include:

- *How do students in our class travel to their favorite outdoor space?*
- *How much of our photographs contain the sky?*
- *Do students in our class whose favorite outdoor space is local tend to go to their favorite outdoor space more often than students in our class whose favorite outdoor space is not local?*

Survey/Data Collection Questions.

Survey/data collection questions are the questions we ask to collect data. To answer their statistical investigative questions, students need to collect and consider data. One way this can be done is through posing survey/data collection questions. As with statistical investigative questions, survey/data collection questions need to be formally posed to ensure that we get the data we need to answer our statistical investigative question. For example, (from the [third lesson](#)) to answer the statistical investigative question about going to the favorite outdoor space more often if it is local (Figure 3), students discuss that they need to know how often students go to their favorite outdoor space and whether the space is local and how they will define these terms.

Figure 3. Photographs showing the different favorite outdoor spaces that prompted students to compare between visiting local spaces and not local spaces.

They pose two survey questions to help answer their statistical investigative question:

- *How many of the last seven days did you go to and use your favorite outdoor space?*
- *How far in miles is your favorite outdoor space from your house? (Give your answer to the nearest mile).*

Students then categorize the distances as local or not local.

Analysis Questions.

Analysis questions provide prompts for what we should attend to when describing our displays. Analysis questions can be developed in conjunction with students and should attend to key features and characteristics of displays (Arnold & Pfannkuch, 2014). Analysis questions could also be provided as scaffolds for noticing features of displays, for example, in the [first lesson](#), to encourage students to develop more elaborate observations and descriptions of their displays, a teacher might offer sentence frames:

- My observation is to the right, left, or with the most common bin.
- There are more or less observations to the left or right of my bin.
- I do or do not think my observation represents the usual / predicted / typical count because
- Pattern: As _____ happens I notice that _____.
- Comparison: My photograph is similar /different to the rest of the class because

Analysis questions could include (e.g., see the [third lesson](#)):

- For categorical data
 - What is the most common type of favorite outdoor space?
 - What is the least common type of favorite outdoor space?
 - Are there any surprising favorite outdoor spaces?
 - How many people have a national park as their favorite outdoor space?
- For quantitative data

- What is the shape of the distribution?
- What is the median?
- What is the highest value, the lowest value?

Interrogative Questions.

Interrogative questions are the questions we ask as checks and balances across the statistical problem-solving process (Arnold & Franklin, 2021). Interrogative questions include ideas such as:

- Checking statistical investigative questions, for example, What is the variable that you are interested in? Is this a topic that students in our class would be happy to answer survey questions about? (thinking about ethics)
- Checking survey questions, for example, In your question, what was meant by this space? When you say how far is your favorite space, was your intent how far in miles or how far in time? Checking data before analysis, for example, Are there any duplicate or incomplete entries? Do any of the variables need to be modified or have categories combined?
- Checking data displays, for example, Is this the best data visualization for the data you have?
- Checking our interpretation of results, for example, Have we answered the statistical investigative questions? Does this make sense with what we know about...?

Investigate More Sophisticated Problems

Nontraditional data types require creativity and innovation but also the negotiation of standard data collection procedures, all of which Level B students are capable of doing and are absolutely crucial for STEM fields.

Using Technology to Analyze Collected Data.

In the lessons, the overarching question serves to anchor the nature of the investigation, providing ideas for statistical investigations and subsequent data collection. The types of data that we are able to collect using technology present opportunities for students to grapple with complex ideas such as, “How do we analyze text responses? How do we describe and classify photographs in a uniform way?” The opportunities generated using technology within the statistical problem-solving process cannot be understated. Technology opens opportunities to investigate more sophisticated problems. For example, each of the three lessons uses data collected from an open-source web survey form. Digital surveys allow students to design, revise, and collect data quickly and efficiently.

While students know intuitively that the ‘answers’ are present in the survey data, they need opportunities to apprentice in the process of cleaning and transforming the data into usable form. Teachers can use data science tools to encourage students to leverage computational approaches when cleaning and transforming data, which is often messy and complex. This aids in data sense-making and question formulation. As students start to display data, they quickly identify the limitations of ready-made analysis tools such as those available within

Google Sheets; for example, displaying category data requires them to summarize the data before a display can be made.

However, data visualization tools such as [CODAP](#) and [iNZight](#) provide students the opportunity to engage in visualization inquiry and explore their data without having to make summary tables. Students can generate many different plots (Figure 4) in a matter of minutes directly from raw data tables. This allows students to explore associations between two quantitative variables or two categorical variables and draw comparisons between two or more groups. These visualization tools:

- allow students to use multiple representations as they analyze the data to answer their statistical investigative question(s)
- support students as they develop their understanding around ‘good and better’–style survey questions
- increase student data analysis and interpretation fluency
- highlight how the same data can be represented in different ways and this variety foregrounds certain patterns and trends.

Students can easily explore data science as a storytelling endeavor by considering, ‘Which plot best helps me tell the story of my variable or my statistical investigation question?’

Figure 4. Three different CODAP Visualizations for the same variable.

Using Technology to Collect Additional Data.

Technology also supports access to data ‘hiding’ within our data set. Students can develop procedures to employ techniques borrowed from microbiology and ecology by using grids (Figure 5) to quantify the variables of interest within a photograph such as, What percentage of the photograph is sky? See the [first lesson](#) for more on how this can be done.

Figure 5. Showing how the grid is used to quantify the variables of interest within a photograph—the photograph is stretched proportionally to fill the grid.

Using digital tools, we can collect data from photographs that are not possible otherwise. Digital applications can analyze the predominant color within each pixel of a photograph (e.g., a hex code, see http://www.coolphptools.com/color_extract for a free image color extract tool) and tabulate the frequency of those colors. These colors can be displayed within applications such as CODAP. In this example, students can dig deeper into the data set engaging in multiple statistical investigations that are inspired by visualizations itself. Here, the plots serve as a stimulus to pose new statistical investigative questions such as, ‘Is there a relationship between the percentage of trees in our photograph and the percentage of human made features? How much variation is there within our color descriptions of pixels?’ These further investigations may require students to collect additional data, recategorize data, merge data sets, or pose new questions, all components of data science that become evident when students are given less structured, messy data (see the [first lesson](#)).

Analyzing Possible Associations and Comparing Groups Using Technology

As technology has developed, more visualizations of data are available through a variety of applications, including spreadsheets, applets, and visualization software, such as Tableau and PowerBI. Although legacy statistics education software (e.g., 32-bit applications) do not continue to operate on newer technology, there are open-source options available. Tools such as CODAP and TUVa help students build an ‘intuition’ around creation and interpretation of data representations. For Level B students, these tools can help support students

with data storytelling without layering on the complexity of coding that is often associated with data science, such as programming in Python or R.

All three lessons have used CODAP as the statistical analysis tool. CODAP connects observations in tables to graphs and allows a third (and fourth) variable to be added. For example, when a quantitative variable is compared across categories, a series of box plots allows us to see a meaningful difference. Figure 6 shows that the number of visits to the space in the last 7 days for those who travel by car tends to be lower than the number of visits to the space in the last 7 days for those who get there by walking or running. Adding a third variable of whether they exercise at this space allows us to see more information behind this relationship. Of the six students who give exercise as a reason for why it is their favorite space, four of these travel there by car.

Figure 6. Comparing groups with box plots.

Perhaps students wonder if there is an association between exercise and how they get to their favorite outdoor space. By comparing categorical variables in CODAP, teachers and students can generate a two-way table where each observation is represented with a dot. Individual observations can be selected within the two-way table and cross referenced to the raw data table. As students work to describe the association between the variables, additional features such as percentages and counts can be selected to describe plots more precisely. Students can further investigate their data by adding a third variable, such as the primary pixel color name of the photograph can be added (Figure 7).

Figure 7. Considering association with two-way tables.

Conclusion

As students progress from Level A to Level B, they begin to pose their own statistical investigative question(s). The way a teacher frames the investigation and scaffolds student thinking can limit or expand the scope of the investigation. Our three different approaches took the same prompt using photographs as data to tell stories—using nontraditional data—and without predetermining the end point, which took the prompt in multiple directions. For many, data science often embodies a spirit of inquiry and exploration around data that is relevant to our everyday lives. These types of inquiry lessons often produce student work with a large amount of variability that is hard to predict. [Lesson plans](#) such as these described may help teachers and students become more confident with open-ended tasks. The opportunities afforded for students in this approach allow them to connect statistics to their everyday life, make meaning with data, and engage in the statistical problem-solving process.

Disclosure Statement

Pip Arnold, Leticia Perez, and Sheri Johnson have no financial or non-financial disclosures to share for this article.

References

- Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education*, 29(1), 122–130. <https://doi.org/10.1080/26939169.2021.1877582>
- Arnold, P., & Pfannkuch, M. (2014). Describing distributions. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014)*. International Statistical Institute. https://icots.info/icots/9/proceedings/pdfs/ICOTS9_8G1_ARNOLD.pdf

Bargagliotti, A., Arnold, P., & Franklin, C. (2021). GAISE II: Bringing data into classrooms. *Mathematics Teacher: Learning and Teaching PK-12*, 114(6), 424–435. <https://doi.org/10.5951/MTLT.2020.0343>

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, 16(1), 44–49. [https://iase-web.org/documents/SERJ/SERJ16\(1\)_Engel.pdf](https://iase-web.org/documents/SERJ/SERJ16(1)_Engel.pdf)

Franklin, C., & Bargagliotti, A. (2020). Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review*, 2(4). <https://doi.org/10.1162/99608f92.246107bb>

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing data science: A framework and case study. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.2d83f7f5>

Perez, L. R., Spangler, D. A., & Franklin, C. (2021). Engaging young learners with data: Highlights from GAISE II, level A. *Harvard Data Science Review*, 3(2). <https://doi.org/10.1162/99608f92.be3c2ec8>

STEW. (n.d.). *Statistics education website*. American statistical association. Retrieved May 5, 2021, from <https://www.amstat.org/asa/education/STEW/home.aspx>

©2021 Pip Arnold, Leticia Perez, and Sheri Johnson. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.