Data and Analysis

Literature Review

SUGGESTED CITATION: Computer Science Teachers Association. (2026). *Data and analysis: Literature review.* https://csteachers.org/k12standards/research

WestEd.org

This review summarizes the literature on how to define each of the five subtopics within the Data and Analysis topic area and recent research on learning progressions and pedagogy. These five subtopics were proposed during the Computer Science Teachers Association's (CSTA) second in-person standards writing meeting on November 7–8, 2024, based on the learning outcomes for Data and Analysis described in the *Reimagining CS Pathways* report (CSTA et al., 2024):

1. Data Fundamentals
2. Data Processing
3. Data Applications
4. Data Investigations
5. Impacts of Data Science

Introduction

This literature review draws heavily from the work of the Data Science for Everyone (DS4E) Learning Progressions Writing Sprint team. In the course of doing this literature review, it became apparent that the originally proposed Data Applications subtopic was hard to define as a distinct topic area in its own right. The two learning outcomes that were originally assigned to Data Applications (DA.3 and DA.15) were redistributed to Impacts of Data Science.

Data Fundamentals

Students should learn about the nature of data, beginning at an early age. First, data are collected by people to answer questions about the world. There are many ways to "turn observations into data" (Konold & Higgins, 2003, p. 195), including direct measurement, surveying, and using sensors. There are different types of data. From an early age, students can distinguish between numeric and non-numeric data (Bargagliotti et al., 2020). Students can also explore non-traditional data types (e.g., photographs, written notes, audio recordings) to see what can be learned from them (Perez et al., 2022). As students advance, they should use datasets with different common standardized units (e.g., measures of length, mass, capacity) and less common standardized units that they may encounter in science or other classes (e.g., horsepower, magnitudes of stars, Richter scale) (Bargagliotti et al., 2020). As their knowledge of Computer Science and data science grows, students will need to distinguish between how data types are used in programming (e.g., strings, integers, floats, and Booleans) and how data types are used in data science and statistics (e.g., nominal, ordinal, discrete, and continuous data).

Students in elementary school may work with data tables. In middle school, students should also begin working with larger sets of data in spreadsheets. To do so effectively, students must understand case structure in a dataset—that each row represents a single case and the data in each column represents some attribute of that case (McKinney et al., 2024). Students should learn to read, make sense of, and create data documentation, including data dictionaries and descriptions of how the data were generated (Bargagliotti & Gould, 2022).

Table 1. *Data and Analysis Learning Outcomes From the* Reimagining CS Pathways *Report (2024, p. 21) That Fall Under the Data Fundamentals Subtopic*

Level	Learning Outcome
Remember	**DA.1** Identify and define data types **DA.2** Identify basic data formats (e.g., tables, schemas, JSON)
Understand	**DA.4** Understand the difference between data and metadata **DA.5** Describe how different types of data (e.g., audio, visual, spatial, environmental) can be collected computationally

Data Processing

Erickson et al. (2019) defined a "data move" as "an action that alters a dataset's contents, structure, or values" (p. 3). They identified six core data moves:

1. **Filtering**, which produces a subset of data. Filtering serves to reduce the scope of an investigation or reduce the complexity/quantity of data in order to help generate insights.
2. **Grouping**, which enables comparisons between different subgroups in a dataset. Binning is a specific type of grouping that is applied to get discrete groups from continuous data.
3. **Summarizing** is "the process of producing and recording a summary or aggregate value, i.e., a statistic" (Erickson et al., 2019, Section 3.3). There is a wide variety of summary statistics, including frequencies and measures of center and variability. Grouping and then summarizing can consolidate a dataset into a smaller number of categories and variables. This reduction of information can help with visualizing, but can potentially lead to oversimplification.
4. **Calculating**, sometimes called mutating or transforming, involves calculating a new attribute or variable from one or more existing attributes (e.g., converting units, calculating a total score by summing up the number of correct responses, recoding a larger set of categories into a smaller set of categories).

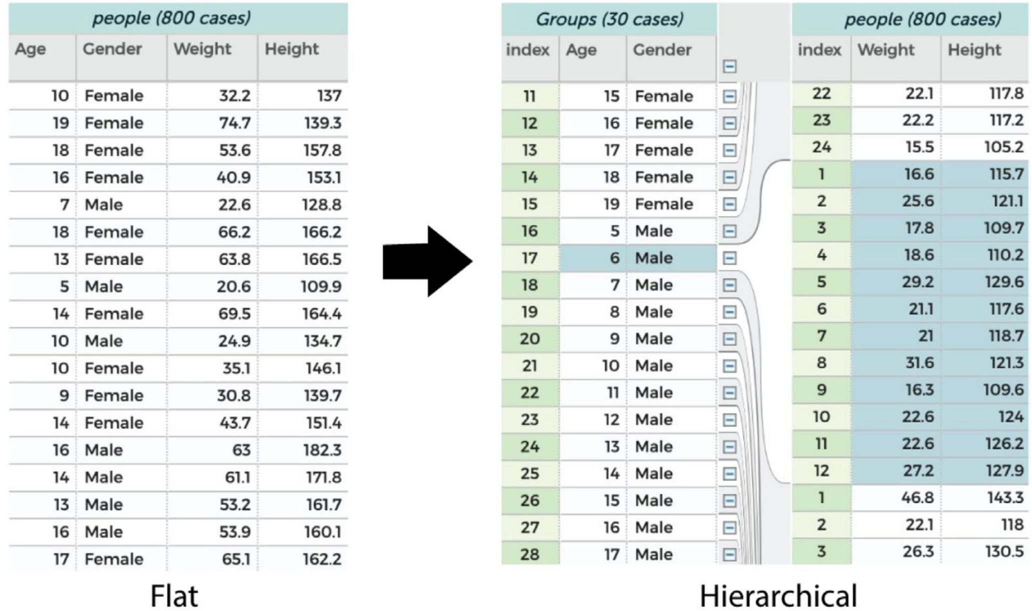
(continued)

Table 2. *Data and Analysis Learning Outcomes From the* Reimagining CS Pathways *Report (2024, p. 21) That Fall Under the Data Processing Subtopic*

Level	Learning Outcome
Apply	**DA.6** Transform and prepare (e.g., normalize, merge, clean) data
Analyze	**DA.8** Trace how data moves through a program
Evaluate	**DA.10** Evaluate approaches to cleaning data in a given context

5. **Merging/joining** involves combining multiple datasets into one. The simplest form of merging is concatenating two datasets with the same variables. For example, if Classroom A and Classroom B both generate a two-column dataset with the name and height of each student, these two datasets can be merged by adding one to the end of the other. Joining is more complex. Instead of adding new cases with the same variables, as in concatenating, joining adds new information—new variables or attributes—about the existing cases. For datasets to be joinable, each dataset must have a variable that functions as a unique ID that can be used to match cases across datasets. For example, Classroom A generates two two-column datasets: a dataset with each student's name and height and a dataset with each student's name and favorite color. These two datasets can be merged to create a three-column dataset with name, height, and color by matching the data by students' names.

6. **Making hierarchy** is taking a "flat" dataset—one where each row represents a case with each column representing an attribute—and turning it into a "nested," or hierarchical, dataset by forming subsets, as shown in Figure 1. Some datasets are naturally hierarchical (e.g., attributes of students nested within schools nested within districts).



people (800 cases)			
Age	Gender	Weight	Height
10	Female	32.2	137
19	Female	74.7	139.3
18	Female	53.6	157.8
16	Female	40.9	153.1
7	Male	22.6	128.8
18	Female	66.2	166.2
13	Female	63.8	166.5
5	Male	20.6	109.9
14	Female	69.5	164.4
10	Male	24.9	134.7
10	Female	35.1	146.1
9	Female	30.8	139.7
14	Female	43.7	151.4
16	Male	63	182.3
14	Male	61.1	171.8
13	Male	53.2	161.7
16	Male	53.9	160.1
17	Female	65.1	162.2

Groups (30 cases)			people (800 cases)		
index	Age	Gender	index	Weight	Height
11	15	Female	22	22.1	117.8
12	16	Female	23	22.2	117.2
13	17	Female	24	15.5	105.2
14	18	Female	1	16.6	115.7
15	19	Female	2	25.6	121.1
16	5	Male	3	17.8	109.7
17	6	Male	4	18.6	110.2
18	7	Male	5	29.2	129.6
19	8	Male	6	21.1	117.6
20	9	Male	7	21	118.7
21	10	Male	8	31.6	121.3
22	11	Male	9	16.3	109.6
23	12	Male	10	22.6	124
24	13	Male	11	22.6	126.2
25	14	Male	12	27.2	127.9
26	15	Male	1	46.8	143.3
27	16	Male	2	22.1	118
28	17	Male	3	26.3	130.5

Figure 1. ***Left:*** *Flat table of student data with age, gender, weight, and height.* ***Right:*** *Hierarchical table of the same data; there are 30 age-gender groups, and the weight-height data for each individual student is nested within these groups (Erickson et al., 2019, p. 13, Figure 5).*

Erickson et al. (2019) also described three more advanced data moves:

1. **Visualization**, which we discuss further in the following Data Investigations section. Students may need to perform a series of data moves to make data suitable for visualization. For example, students may need to group and summarize data and then sort the resulting values. Students will need to understand how to flexibly use data moves to support creating data visualizations.
2. **Cleaning data** is making a dataset usable for analysis. For example, students may need to identify how missing data is recorded and respond appropriately (e.g., converting “N/A” to blank cells). They may need to fix errors in formatting (e.g., dates recorded as numbers or vice versa), convert units (e.g., convert scientific notation back to standard notation), or restructure or reshape the data to make certain computations easier.
3. **Sampling** is a special type of filtering. Students may want to randomly sample from a dataset to support the construction of simulations.

Students are often unaware that certain data moves are possible and why they are useful. Teach data moves explicitly. Students should be given the opportunity to not only learn that data moves exist, but consider for themselves which data moves might be most useful for moving an analysis forward (Erickson et al., 2019).

Students are often unaware that certain data moves are possible and why they are useful. Teach data moves explicitly.

Data Investigations

The Guidelines for Assessment and Instruction in Statistics Education II (GAISE II; Bargagliotti et al., 2020) conceptualize a four-step statistical problem-solving process (Figure 2) that can also be applied to data investigations:

1. Formulate Statistical Investigative Questions
2. Collect/Consider the Data
3. Analyze the Data
4. Interpret the Results

In the proposed subtopic structure for Data and Analysis, step 2, *Collect/Consider the Data*, is captured by the Data Fundamentals and Data Processing subtopics while the other process steps fall under Data Investigations.

WestEd summarized the general learning progression for statistical investigation in the *Comparison of CSTA Standards to Other Standards and Frameworks* report (in particular, see Table 22 and Appendix E) (Kao & Tsan, 2025), so we will not re-review that content here. However, the following sections expand on the research regarding data visualization and communication.

Table 3. *Data and Analysis Learning Outcomes From the* Reimagining CS Pathways *Report (2024, p. 21) That Fall Under the Data Investigations Subtopic*

Level	Learning Outcome
Apply	**DA.7** Apply principles of inclusive collaboration to a project involving the analysis of data
Analyze	**DA.9** Analyze data using computational thinking principles to make inferences or predictions
Evaluate	**DA.11** Assess whether and how a given question can be answered using computational methods and data, and what specific data is needed **DA.13** Evaluate data visualizations for clarity, potential biases, etc.
Create	**DA.14** Select, organize, interpret, and visualize large datasets from multiple sources to support a claim and/or communicate information **DA.16** Create a data analysis artifact (e.g., a visualization) using principles of human-centered design

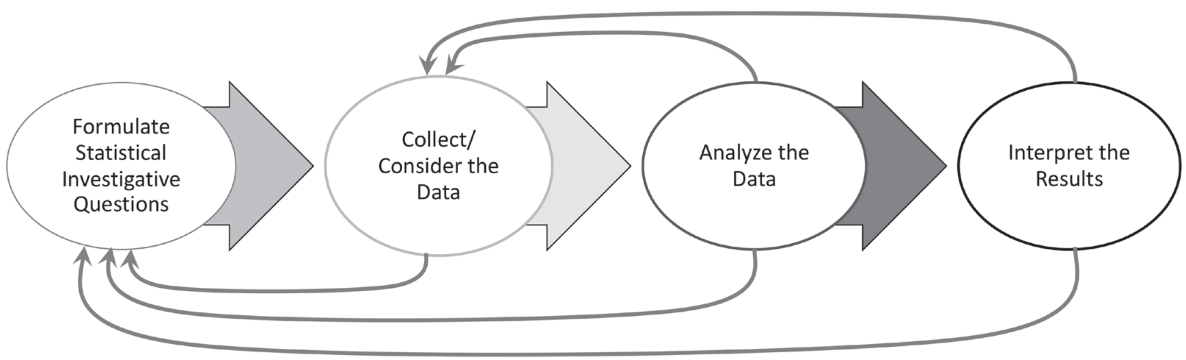


Figure 2. *Statistical problem-solving process from GAISE II (Bargagliotti et al., 2020, p. 13)*

Data Analysis

Konold and Higgins (2003) distinguish between data analysis and statistical inference, the latter of which involves hypothesis testing. Students can perform data analysis without making statistical inferences, by organizing and displaying data, describing the data, summarizing the data, and comparing data displays or summaries. Students may need support in understanding what can be concluded from a specific dataset (e.g., "More people selected pizza than hamburgers as their favorite food on this survey") and conclusions they have drawn based on their personal experiences (e.g., "Everyone loves a pizza party"). Students should also learn how to document their statistical investigations, including how the data were collected and methods used to analyze the data. This is foundational for ensuring reproducibility and trustworthiness of both the data itself and the analysis (Bargagliotti & Gould, 2022).

Data Visualization

It is important to explicitly teach students how to understand and evaluate data visualizations, either before or concurrently with teaching students how to create data visualizations (Glazer, 2011). Students should learn how to interpret the different types of data visualizations, including visualizations that are already commonly taught in math classes (e.g., bar charts and scatterplots), visualizations that are not typically taught in math classes but are commonly used modern data stories (e.g., dumbbell plots, diverging bar graphs, Sankey diagrams; Evergreen, 2022; Wilke, n.d.), and visualizations that are imaginative and/or novel, such as those that are seen in many mainstream media publications today (Wilkerson et al., 2025). This can and should happen in all grade levels. For example, students in early elementary grades could create a map of their school playground and draw figures next to each apparatus to represent each student's favorite piece of equipment. In high school, students could review and create more complex visualizations to represent social or scientific phenomena.

When teaching data visualizations, it is important to help students understand conventions commonly used in graphs and conventions used in a specific graph, as well as the context and/or content focus of the data (Maltese et al., 2015). Beginning in early grades, students should practice interpreting the titles, captions, labels, and legends of graphs, and mapping features such as color and direction to meaning (e.g., taller bars typically mean "more;" complementary colors typically represent opposite ideas) (Wilke, n.d.). Students should also understand how human perception affects how we understand data visualizations. For example, humans are much better at discriminating between bars of different lengths than at discriminating between angles of different numbers of degrees (Cleveland & McGill, 1984), which is why modern data visualization guides typically do not recommend using pie charts (and why pie charts are supposed to be ordered from largest to smallest slice) (Wilke, n.d.). Students should also learn about differences between humans' color and contrast perception and how to make accessible data visualizations.

Finally, students should create multiple representations and visualizations of the same data, using different chart types, color palettes, and other graphical elements, to better understand how different choices can highlight or obscure key relationships in the data and how violating common graphical conventions can mislead. For example, changing the scale of the *y*-axis leads to very different intuitive interpretations of the relative difference between bars on a bar graph (Witt, 2019). Students can also try to partially reverse-engineer the dataset from a visualization, to help them understand how data maps onto different visualizations (Arnold et al., 2022).

Data Storytelling

There are a variety of ways to tell data stories, and a variety of definitions of "data story." Some definitions of "data story" focus on explaining the findings of a data analysis—for example, describing a data visualization with a text narrative (Sanei et al., 2020, as cited in Kemble & Wilkerson, 2024). Others focus on describing the process and practices involved in making sense of data. Kemble and Wilkerson (2024) distinguished between students telling stories about their data and students telling stories with their data. When students tell stories about data, they might describe the data source(s) and the context/history of the data, as well as the students' motivation for analyzing the data. When students tell stories with data, they might describe patterns in the data, how the data supports or does not support an argument, or conclusions and implications of the data. Data storytelling is a good opportunity for students in Computer Science classes to engage with Computer Science communication practices as well as reading, writing, and speaking standards from English language arts.

Data storytelling is a good opportunity for students in Computer Science classes to engage with Computer Science communication practices as well as reading, writing, and speaking standards from English language arts.

Impacts of Data Science

Students should explore the benefits and harms of a "big data" society. In the early elementary grades, students can identify situations when data can help us understand the world and make decisions. As students get older, they can explore a variety of issues in data science to understand benefits and harms, such as:

- The role of data in modern software applications, including artificial intelligence (AI) and machine learning (ML) algorithms.
- The environmental costs of data centers and the power needed to run generative AI/ML models.
- The responsibility of data scientists to ethically collect and manage people's data, the laws that govern data privacy, and best practices for protecting confidential data (Segalla & Rouziès, 2023).
- How data can be biased. Bias in data may lead to benefits for some but harm for others. Equip students with processes for evaluating the degree of bias in data (for knowing when they can trust data) and methods for mitigating bias in data, such as soliciting diverse perspectives.
- Examples of how data science helps solve problems in a variety of disciplines.
- Case studies of real-world decisions that were made by governments or other institutions, based on data, and/or decisions that were made without being informed by data.

Table 4. *Data and Analysis Learning Outcomes From the* Reimagining CS Pathways *Report (2024, p. 21) That Fall Under the Impacts of Data Science Subtopic*

Level	Learning Outcome
Understand	**DA.3** Describe, at a high level, the role of data in AI/ML applications
Evaluate	**DA.12** Assess societal impacts of data analysis and related ethical issues (e.g., biased data used to train AI systems, attribution related to products of generative AI)
Create	**DA.15** Devise plans for using data to solve a problem

References and Works Consulted

Arnold, P., Bargagliotti, A., Franklin, C., & Gould, R. (2022). Bringing complex data into the classroom. *Harvard Data Science Review, 4*(3). https://doi.org/10.1162/99608f92.4ec90534

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education.* American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Bargagliotti, A., & Gould, R. (2022). Secondary data in the secondary data science and statistics classroom. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11) 2022*. International Association for Statistical Education. https://doi.org/10.52041/iase.icots11.T2H2

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531–554. https://doi.org/10.2307/2288400

CSTA, IACE, ACM, Code.org, College Board, CSforAll, & ECEP Alliance. (2024). *Reimagining CS pathways: Every student prepared for a world powered by computing*. Association for Computing Machinery. https://doi.org/10.1145/3678016

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education, 12*(1). https://doi.org/10.5070/T5121038001

Evergreen, S. (2022). *Quantitative chart chooser.* Evergreen Data: Intentional Reporting and Data Visualization. https://stephanieevergreen.com/wp-content/uploads/2022/06/Evergreen-Data-Quantitative-Chart-Chooser.pdf

Glazer, N. (2011). Challenges with graph interpretation: a review of the literature. *Studies in Science Education, 47*(2), 183–210. https://doi.org/10.1080/03057267.2011.605307

Kao, Y., & Tsan, J. (2025). *Comparison of CSTA standards to other standards and frameworks*. https://csteachers.org/k12standards/research

Kemble, K. A., & Wilkerson, M. H. (2024). Storytelling with and about data: Mapping the terrain. In R. Lindgren, T. I. Asino, E. A. Kyza, C. K. Looi, D. T. Keifert, & E. Suárez (Eds.), *Proceedings of the 18th International Conference of the Learning Sciences (ICLS) 2024* (pp. 1838–1841). International Society of the Learning Sciences. https://doi.org/10.22318/icls2024.160400

Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). National Council of Teachers of Mathematics.

Maltese, A. V., Harsh, J. A., & Svetina, D. (2015). Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *Journal of College Science Teaching, 45*(1), 84–90. https://doi.org/10.2505/4/jcst15_045_01_84

McKinney, D., Morton, C., Tuohy, B., Berg, S., Karlstad, A., Ortega, C., Allison, Z., Munzel, G., Washburn, M., & Kao, Y. (2024). Iterative design of a socially-relevant and engaging middle school data science unit. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education (SIGCSE) 2024* (Vol. 1, pp. 826–832). Association for Computing Machinery. https://doi.org/10.1145/3626252.3630886

Perez, L., Johnson, S., & Arnold, P. (2022). Using non-traditional data to build data dispositions. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings* of the *11th International Conference on Teaching Statistics (ICOTS11) 2022*. International Association for Statistical Education. https://doi.org/10.52041/iase.icots11.T10A3

Sanei, H., Kahn, J. B., Yalcinkaya, R., Jiang, S., & Wang, C. (2024). Examining how students code with socioscientific data to tell stories about climate change. *Journal of Science Education and Technology, 33*, 161–177. https://doi.org/10.1007/s10956-023-10054-z

Segalla, M., & Rouziès, D. (2023, July–August). The ethics of managing people's data: The five issues that matter most. *Harvard Business Review*. https://hbr.org/2023/07/the-ethics-of-managing-peoples-data

van Borkulo, S. P., Chytas, C., Drijvers, P., Barendsen, E., & Tolboom, J. (2023). Spreadsheets in secondary school statistics education: Using authentic data for computational thinking. *Digital Experiences in Mathematics Education, 9*, 420–443. https://doi.org/10.1007/s40751-023-00126-5

Wilke, C. O. (n.d.). *Fundamentals of data visualization*. https://clauswilke.com/dataviz/

Wilkerson, M. H., Kim, J., Lee, H. S., Stokes, D. J., & Ferrel, M. (2025). How teachers envision using data visualization discussion tasks in classroom instruction. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-024-10521-y

Witt, J. K. (2019). Graph construction: An empirical investigation on setting the range of the *y*-axis. *Meta-Psychology, 3*. https://doi.org/10.15626/MP.2018.895

