Chapter 5: Data Science

Table of Contents

[Note from writers to CFCC for December 16–17 meeting: This chapter still needs major revision, especially in the high school section. Two California members of the writing team for the American Statistical Association's brand-new report "Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education" have assisted the writing team by identifying the major changes needed to the chapter to bring it more in line with modern data science/statistics education research as reflected in GAISE II. The "statistical problem-solving process" has been partially integrated into this revision; the many roles of questioning have been made clearer and more prominent; and the definition of data science has been improved. More work needs to be done, to make sure the chapter is as accurate as possible in setting out the K–12 needs for this emerging field. In addition, many CFCC changes are waiting on completion of the more structural changes.]

Introduction

The ability to work with and understand data has become an essential life skill in our newly data-filled world. Students participate in a world driven by data; making sense of data, being able to identify data that is misleading, and using data to make decisions are all important aspects of their role as global citizens in the larger world. It is not only those who have careers in data science—almost all occupations now require that employees collect feedback from data and adjust their practice. Stories about the world are illuminated by massive quantities of data, and community members telling and listening to those stories need to be able to make sense of data to understand their health, finances, and news feeds.

The numbers are staggering: around 1.7 megabytes of digital data were created and stored *every second for every person on earth* in 2020, and the vast majority of data goes unanalyzed (https://techjury.net/stats-about/big-data-statistics/). Our lives are increasingly subject to data-driven algorithms that determine much about our daily experience, including what ads we see, which neighborhoods receive business or public investment, who gets screened more closely at the airport, who receives favorable loan terms, and which medical procedures are recommended or approved.

All California students should graduate from high school with data literacy and options to learn an introduction to data science. Data literacy refers to the ability to reason with and about data, to make good decisions based on data, to ask questions of data, and to use statistical reasoning. Data science is an emerging discipline that includes understanding principles of data collection, data manipulation, data analysis, inference, and interpretation and communication. The Common Core State Standards set out the learning of statistics K–12. A data science lens can help the statistical ideas in the standards come alive and have relevance and meaning for students.

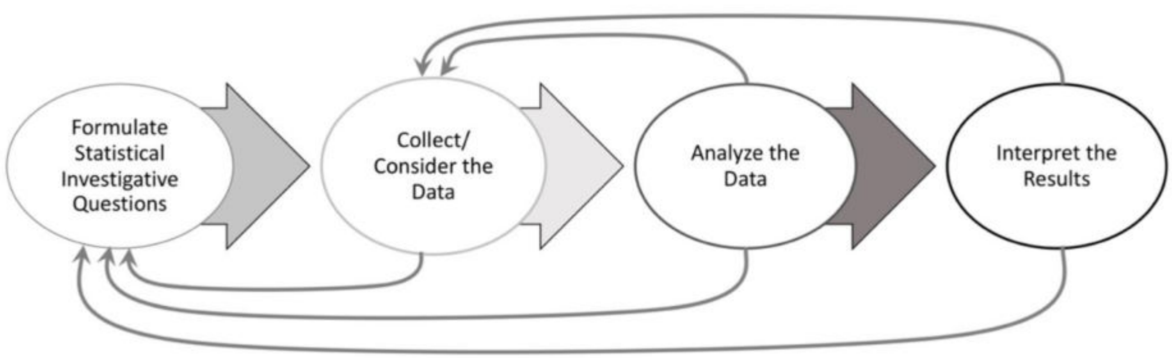
GAISE II is a professional report from the American Statistical Association (ASA) setting out guidelines for assessment and instruction K–12 in statistics and data science, and is an important resource for this area of mathematics. In the GAISE II report they emphasize the following:

1. The importance of asking questions throughout the statistical problem-solving process (formulating a statistical investigative question, collecting or considering data, analyzing data, and interpreting results), and how this process remains at the forefront of statistical reasoning for all studies involving data
2. The consideration of different data and variable types, the importance of carefully planning how to collect data or how to consider data to help answer statistical investigative questions, and the process of collecting, cleaning, interrogating, and analyzing the data
3. The inclusion of multivariate thinking throughout all Pre-K–12 educational levels
4. The role of probabilistic thinking in quantifying randomness throughout all levels
5. The recognition that modern statistical practice is intertwined with technology, and the importance of incorporating technology as feasible
6. The enhanced importance of clearly and accurately communicating statistical information
7. The role of assessment at the school level, especially items that measure conceptual understanding and require statistical reasoning involving the statistical problem-solving process. (GAISE II, 2020, p. 2)

Students should be able to draw on the Standards for Mathematical Practices (SMP) through a statistical lens articulated in the ASA’s Statistical Education of Teachers (SET) report. For example, students should reason abstractly and quantitatively by engaging in statistical thinking while considering where data come from (SMP.2), apply statistical models to “include descriptions of the variability present in data (SMP.4), and consider available tools such as calculators, spreadsheets, applets, statistical packages, and graphical displays to help facilitate the statistical problem-solving process (SMP.5). When students participate in the analysis of large datasets, they should be able to decide which questions matter, and identify which ones can be answered with a given dataset (SMP.4). The statistical problem-solving process is used within the process. Further, students should understand some of the ways in which data are frequently misunderstood or misused and should understand the content and implications of their own digital data footprints. Finally, students should be prepared to pursue additional study directed towards fields which include more intensive work with data, such as designing data collection, deciding on statistical measures appropriate to the questions under consideration, or making conclusions and claims based on data.

The statistical and data science problem solving process, as set out in GAISE II, is shown in Figure 1:

Figure 1. The statistical and data science problem solving process, GAISE II



The California Common Core State Standards in Mathematics (CA CCSSM) contain many content standards that help to build the data understanding and skills that high school graduates require. However, the progression—from counting, categorizing, and simple picture graphs, to the complex skills and understanding that older students may develop—requires careful thought and considerably more focus through the K–12 curriculum than most students have historically experienced. The study of data today is broader than it has ever been. The types of data being collected are vast and the types of techniques used to analyze data can now rely heavily on computational tools. The statistical problem-solving process is important as it provides the foundation for finding meaning in data. Data science and statistics are the science of working with data. The development of statistics and data science mastery articulated in this chapter represents a modern lens through which to examine the CA CCSSM.

Educators regularly use data at the student and classroom level to try to drive instructional decisions. However, a data science perspective can help educators create experiences in which their students learn to “read and write the world with mathematics” (Gutstein 2003). As emphasized throughout this framework, students must experience mathematics as tools for making sense of and impacting their worlds.

Educators should be encouraged to bring data science and statistics directly into their classroom to create student experiences that are meaningful. Students can experience statistics and data science as tools for making sense of and impacting their worlds. The statistical problem-solving process (GAISE II) helps students formulate statistical investigative questions, take in information by collecting primary data or considering secondary data, analyze the data to identify relationships and patterns, and (in many cases) interpret results to answer the question and propose changes to impact the way the world works.

Students who are exposed to and have the capacity to understand data concepts at an early age begin to develop data literacy and data sense in parallel with number sense. As students progress through school they should learn different approaches to data analysis culminating in the investigation of large data sets using up-to-date technological tools.

As students learn the investigative statistical and data science process they should always consider meaning and context. In the past, some learning of statistics was removed from situational settings, leading students to learn abstract methods. Data science involves developing meaning and communicating about a data-rich situation; it should never be removed from its context. Teachers can use local data sets that give students the opportunity to ask questions that are meaningful to them, that can help their local community, or school, allowing students to experience using mathematics to be an engaged citizen. Statistics and data science is about studying situations—asking questions such as: Who collected the data? How was it collected? What is the unit of analysis? Teachers can ask students to turn and talk to their partners and groups about these questions.

In this chapter, we present the progression of data literacy and data science standards and the types of experiences that help build the necessary skills and understandings. Four important principles in the learning of data science are these:

1. Students should experience working with data from a context that is meaningful to them personally. They should have opportunities to solve problems of value to the students and to their schools and communities.
2. Students should learn to engage with real data that include multiple variables. At first students can learn to understand two variables with bivariate data, as they progress through the grades they can learn to handle multivariable data and multivariate thinking.
3. Data investigations should be investigative and collaborative, with students working together to learn the data science and statistical investigative process.
4. Familiarity with technology and modern tools should progress through the grades.

As discussed in more detail in the Chapter 2: Teaching for Equity and Engagement, it is more effective for teachers to plan around big ideas than sets of mathematical methods, and to choose rich tasks that elicit big ideas. In this chapter we set out the big ideas of data science that build to the kind of connected understanding needed.

Definition: Data are observations or measurements in context. **Usage note:** In Latin, the word *data* is the plural of *datum*. However, in English, *data* is now also commonly used with singular verbs and refers to a collection of data points. Thus, “the data shows a correlation…” is more common than “the data show a correlation….” In this chapter we most often use the word *data* in this way—to refer to a collection of data points—and in these contexts it takes singular verbs.

Sources: The development of data science described and illustrated here is guided by the California Common Core State Standards in Mathematics, and is also informed by and largely consistent with the following documents:

- The Guidelines for Assessment and Instruction in Statistics Education Pre-K–12 Report (Bargagliotti, Franklin, Arnold, Gould, Johnson, Perez, & Spangler 2020; https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx)
- The Introduction to Data Science Curriculum (www.introdatascience.org)
- the draft *Data Literacy in K–12* (2020) and other resources from the Center for Radical Innovation for Social Change (RISC) (www.21cmath.org/)
- The Messy Data Coalition
- Youcubed data science resources, news articles, lessons and courses – www.youcubed.org /datascience
- Statistical Literacy: A Complete Hierarchical Construct (https://iase-web.org/documents/SERJ/SERJ2(2)_Watson_Callingham.pdf?1402525004)

Two important sources for contexts in which to explore data science are

- The *California Next Generation Science Standards* (*CA NGSS*) and
- The California Environmental Principles and Concepts.

The Statistical and Data Science Investigation Process

The process of statistical and data science investigating is a four-part process:

(1) Asking Questions

Formulating questions that anticipate variability should be the beginning of the investigative process. Examples of such questions include:

- How fast will my plant grow?

- Do plants exposed to more sunlight grow faster?
- How does sunlight affect the growth of a plant?

These questions contrast with questions that are not investigative and have one answer, such as: How tall is my plant? While questions start the investigative process, students should be encouraged to ask questions throughout the investigative process. (GAISE II page 15).

Recent work in the data feminism movement (see, for example, D'Ignazio & Klein, 2020) has drawn attention to the need to understand not just the context of the data, but the motivation behind data collection and to ask questions about who has been included or excluded from data.

Survey questions will be important to students' investigations. These are questions designed to elicit data from people in order to address a statistical question, such as the length of time it takes to ride a bus to school.

As Arnold has stated, "Any question whose investigation requires repeated counting, measuring, or categorizing is one that data helps to answer." Students learn to use data in increasingly sophisticated ways. Early questions are primarily about description, beginning with categorizing and counting, expanding into questions in measurement situations (at first length/distance; later time, area, volume, and rates). Describing relationships between two varying quantities develops as students move through the grades, as do formal quantitative calculations.

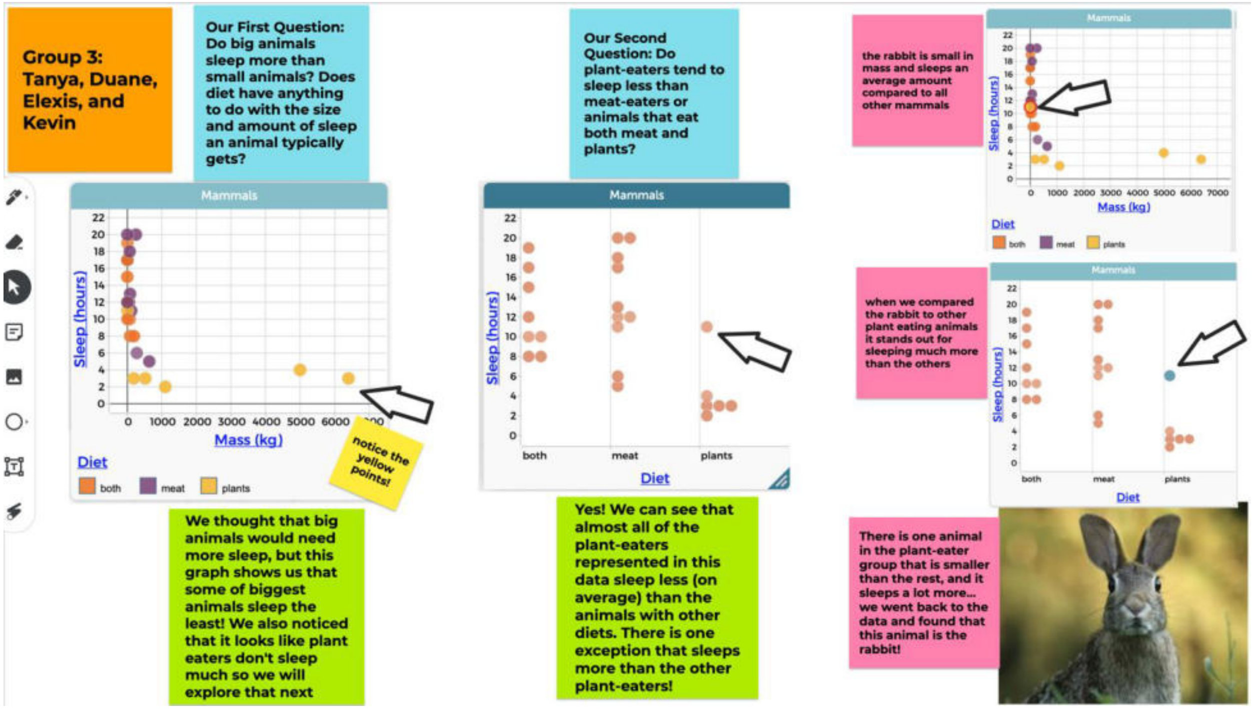
CODAP provides a set of databases that will be interesting to school students, such as data on earthquakes, mammals, stars and cities, and an accessible data investigation online tool. Students can be encouraged to ask questions of the data. For example, a data set of mammals may raise the question, "Is the size of mammals related to the length of time they sleep?" Students can investigate questions using graphing tools that compare variables, statistical tools, a mapping tool and others (see https://codap.concord.org/).

Multivariate thinking comes naturally to humans, and students can develop curiosity about all sorts of data and situations. Young students may ask questions with one variable, such as what is the average age of my class? but as they get older we should encourage bivariate and multivariate thinking. Are older students at my school more likely to read more books would be an example of bivariate data collection.

Vignette 1: CODAP

A group of three students work to explore a CODAP database of 27 mammals: https://codap.concord.org/releases/latest/static/dg/en/cert/index.html?url=https://concord-consortium.github.io/codap-data/SampleDocs/Science/Biology/27mammals/Mammals_Sample.codap

The database provides variables such as the height, mass, speed, life-span, and sleep hours of the mammals. The students quickly become curious and ask questions like, “Do bigger animals sleep longer?” They plot the two variables with the graph tool and start to notice a relationship—in the opposite way than the one they thought—it seems the bigger animals sleep less. The students start an animated conversation discussing the reasons this might be, is it because they are more likely to be predators? They then move on to investigate another relationship—who sleeps more, plant or animal eaters? The students again notice a relationship as well as an outlier (the rabbit) so they wonder about the rabbit, and look at more rabbit data. The students’ investigation of bivariate data and their relationships is filled with moments of curiosity and excitement, as well as important learning.



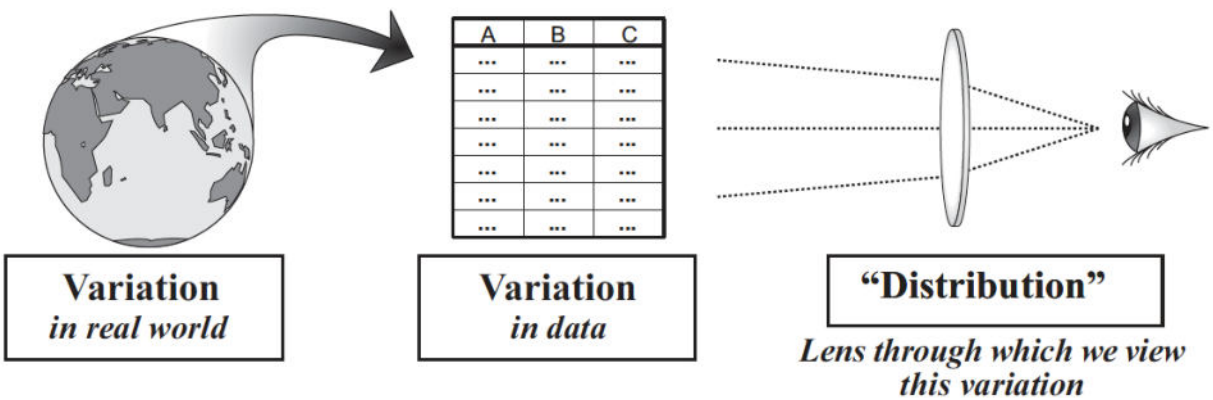
(2) Collecting and Considering Data

Sometimes students may collect their own data when investigating a question. For example, they may ask how far do students travel to school? or they may consider two variables, such as: Are students happier on sunny days? Or they may consider which plants are most prevalent in their local area. In all of these cases, students could collect data by observing plants or surveying students.

A key characteristic of data science is asking questions of "big data" – a data set that is complex, messy and includes many variables. Students can ask questions of big data sets and different students in a class may ask different questions. In a high-school data science class students can learn to clean data sets, an important part of the work of a data scientist. High school students can also learn to download and upload data—and develop the more sophisticated "data moves" that are important to learn if students are tackling real data sets.

After data is collected or acquired students should ask questions about the data—how do the variables differ? How were they collected? Who of what was included in the data collection. This helps students develop an understanding of variability.

High school students taking a course in data science may consider more complex conceptions of data science, that are located in the idea of variation, see for example Figure 2 from Wild (2006).



Further details of the understandings that may be developed in a high school course are given later in this chapter.

Sometimes students may be given data first and then ask a question of the data – reversing the order of 1) and 2).

(3) Analyzing Data and Developing Meaning

In the younger grades, students can analyze and develop meaning from data as they represent it in different ways, using picture graphs, line graphs, bar graphs and other forms of data visualization. From sixth grade, students can learn more formal methods to understand data. The field of statistics has been described as the study of variation, and students learn about variation when they receive opportunities to consider the distribution of data. Measures such as mean, median and mode are measures of the center of a distribution that students learn in middle school. CODAP tools allow students to see distributions of data and to see, visually, that the spread of a distribution will impact measures of center. In high school students will learn about measures of spread and about regression lines.

One of the features of data science is the possibility of predicting outcomes, such as the cable news programs' predictions of election outcomes. Developing understanding of what a prediction means, and how to compare predictive strength of one model over another is not simple and should be developed as a learning trajectory spanning several grades. Students who specialize in high school can learn about cross-validation techniques. Much of the work of professional data scientists is concerned with quantifying error from predictions.

(4) Interpreting and Communicating Results

Students learn to interpret data in increasingly sophisticated ways. Young students may make statements about their data or create data visualizations to communicate results. They may describe the difference between two groups. Even in the early grades, teachers can have conversations with students about generalizability—how much can we generalize from the data we have collected to broader populations? As students move through the grades they can learn to generalize more formally and to include statements of probability and certainty.

A data scientist does not just perform calculation, and an important part of data science is the communication of results. Whereas statistics used to rely on bar charts, pie charts and other familiar representations, data science has created multiple forms of visualizations that represent data, as can be seen in Vignette 2.

Data science is about developing understanding of a situation, it involves holistic thinking, interpretation of meaning, and the communication of complex ideas. An effective data communication draws from writing, and visualizing as well as calculating.

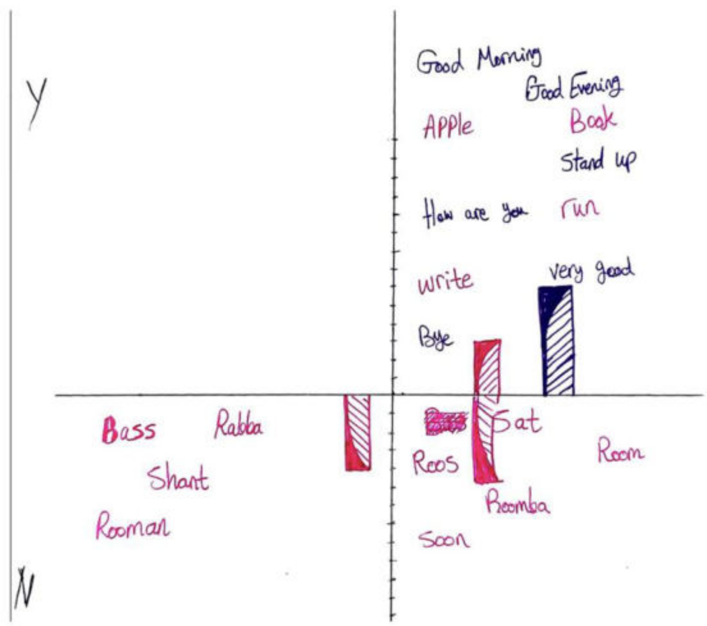
Vignette 2: Dear Data

Rico shares with his class of students the true story of two designers, who lived on different sides of the Atlantic Ocean—one in London, one in New York. For an entire year the two designers mailed each other a postcard every week, that included data from their lives, that they represented in creative and visual ways. The data representations included multiple variables. For example, some weeks the designers recorded all their moments of indecision, in another they recorded all the times that they laughed. The students looked at some of the data visualizations the designers produced and discussed what they could learn and how they could interpret the different variables (http://www.dear-data.com/)

After the discussion, Rico asked his students to collect data over at least a 24-hour period, collecting data on something that interested them, recording at least two variables. When the students came back to class with their data Rico organized the students into groups and asked them to create data visualizations together, supporting each other to consider ways they would represent different variables. In the discussion Rico payed attention to the language needs of the students, and the ways that the activity drew from the principles of Universal Design for Learning (UDL). Students were excited to make their data visualizations, such as the following:

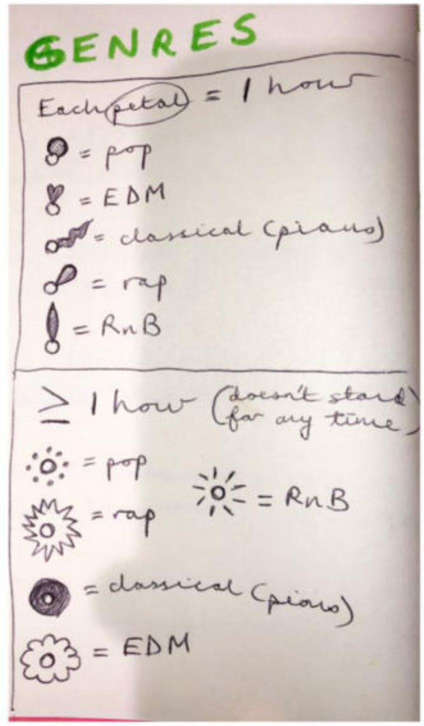
Abdu

How many times does Abdu's 6 year old sister use English and non-English words she knows or does not know while pretending to be a teacher?



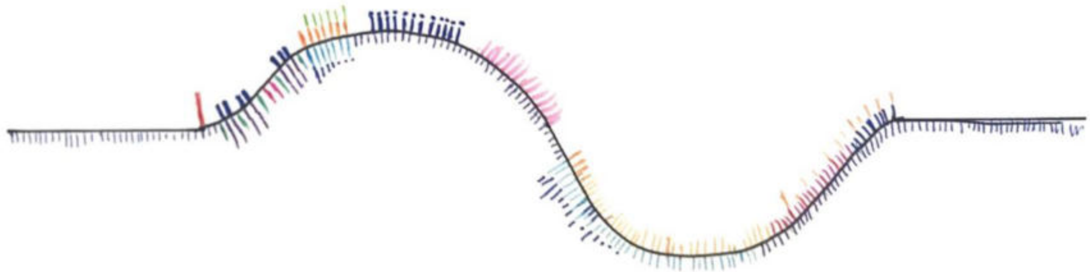
Nikita

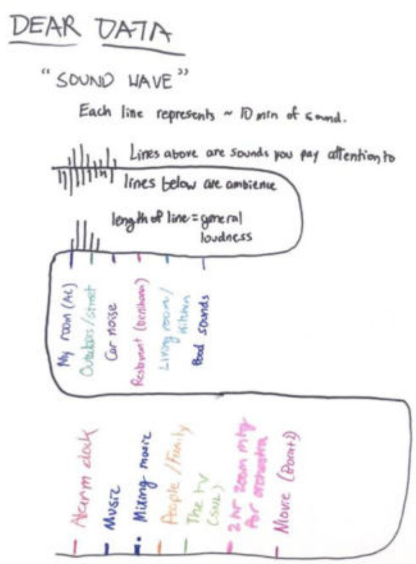
One week of listening to music, what type of genre it was and what Nikita was doing.



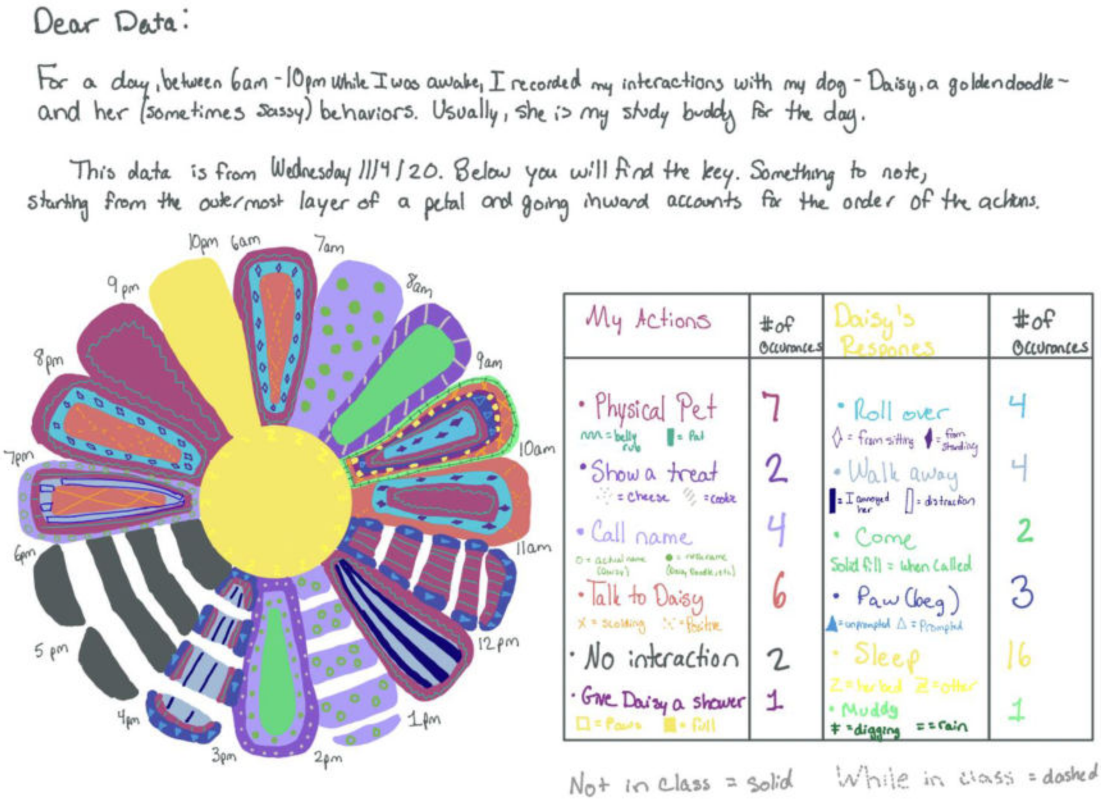
Nathan

Representation of sound length, level of loudness and how much attention is given to it.





Kira's dog interactions



The students made their visualizations using Google Jamboards. After they made them Rico asked the groups to look at the work of other groups and provide feedback to each other on a sticky note. The students were excited to see the ways the different variables related to each other and the ways they could be represented.

Data Talks K–12

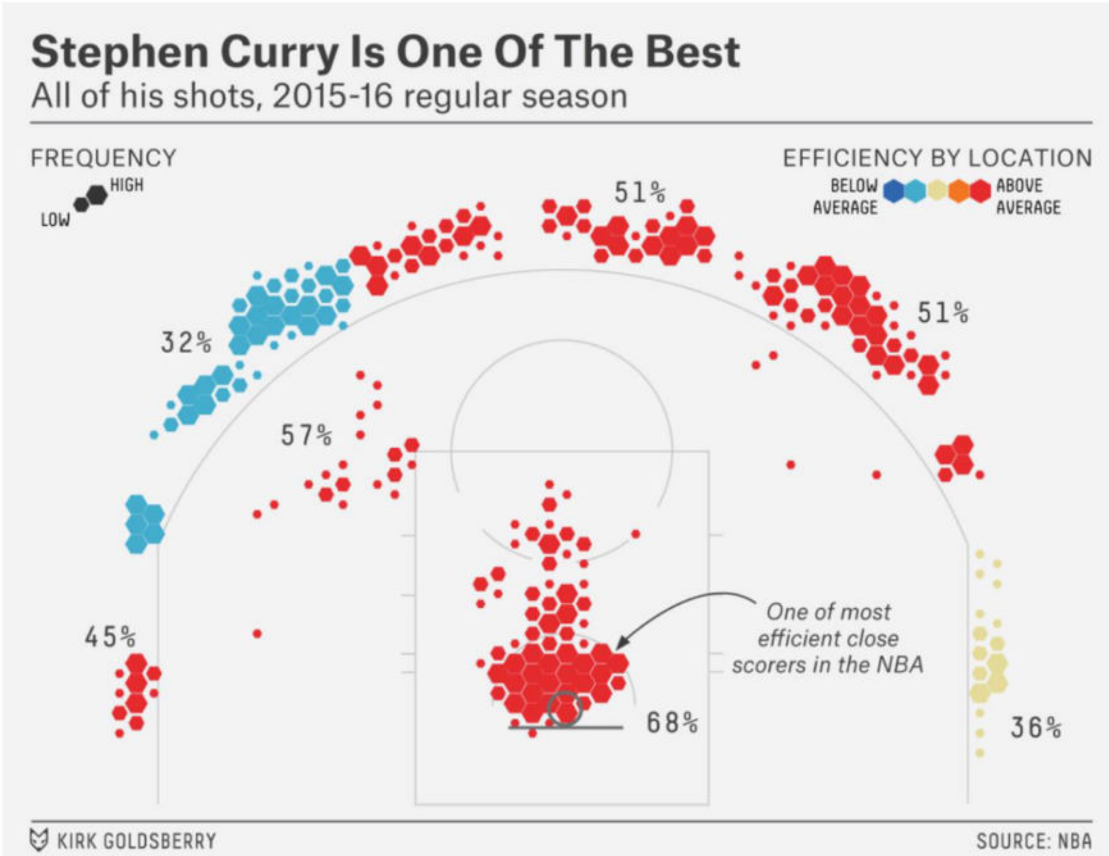
Data talks are short classroom discussions to help students develop data literacy. This pedagogical strategy is similar in structure to a number talk, but instead of numbers students are shown a data visual and asked what interests them. The idea of a data talk was inspired by a New York *Times* weekly section called, "What's Going on in this Graph?" In the New York Times, students can submit their own ideas to a member of the American Statistical Association, who reveals their thinking on the data in the graphs. In the classroom the teacher can guide the discussions and help students develop important understandings. However, it is important to recognize that teachers do not have to be an expert in the topic of the data visualization—instead teachers can guide and encourage curiosity and question asking. One way to support thinking and speaking like a mathematician is to incorporate writing activities or math journals, which allow students to process learning and continue questioning. These activities help all students gain and exchange information and ideas, and support the *California English Language Development Standards*' three communicative modes (collaborative, interpretive, and productive), and allow them to apply knowledge of language to academic tasks using various linguistic resources.

If questions cannot be answered by the teacher or students they can be investigated further. Data talks are intended to pique students' curiosity and encourage question asking, and to help them understand and "read" the data-filled world in which they live. Many of the data visualizations illustrate how multiple variables can be incorporated into one graphic—allowing students to think multivariately.

Grades with younger students can use data visualizations with no or few numbers, or smaller numbers for example:

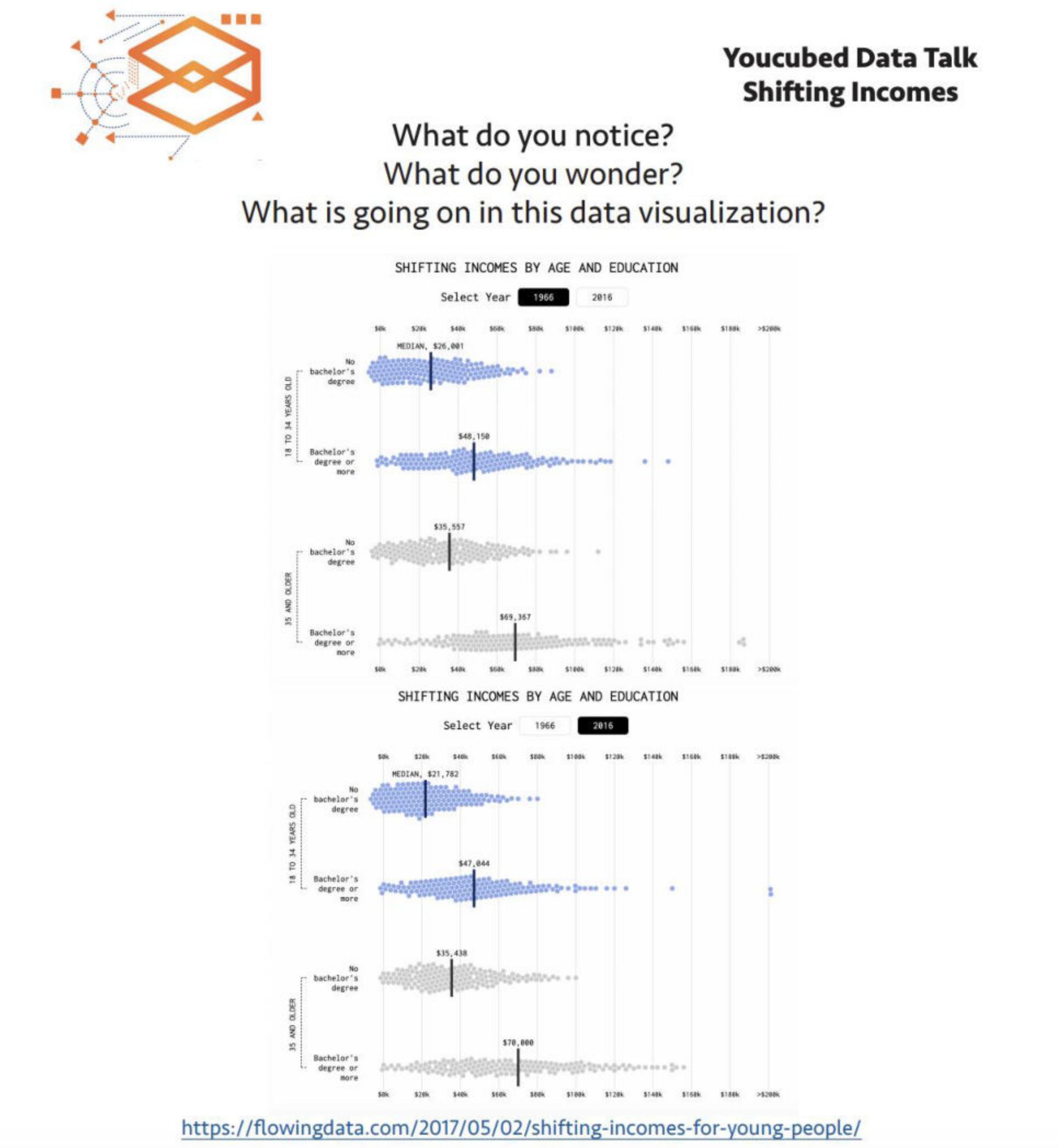
Source: https://www.youcubed.org/wp-content/uploads/2020/09/Water-Usage.pdf

From grade five, students should be able to interpret data visualizations with percentages, for example:



https://fivethirtyeight.com/features/stephen-curry-is-the-revolution/

In higher grades data visualizations can include more complex data representations. For example:



The above examples, and more, can be found at Youcubed, (https://www.youcubed.org/resource/data-talks/), Flowing Data (https://flowingdata.com/), Five Thirty Eight (https://fivethirtyeight.com/), and the New York *Times* resource itself (https://www.nytimes.com/column/whats-going-on-in-this-graph).

Transitioning from Pre–K

Before kindergarten, children begin to describe their world in language, identifying characteristics of objects, places, people, and events: *The ball is red*. *My classroom is warm*. *My teacher is in their twenties*. *Our trip to the park was too short*. Identifying characteristics is the beginning of data, and wondering about characteristics—including countable characteristics—is the beginning of asking questions that data can help to answer. In the California Preschool Learning Foundations, this content is located under the heading of "Algebra and Functions (Classification and Patterning)," in which children "sort and classify objects in their everyday environment," (by one attribute at around 48 months and by more than one attribute at around 60 months of age); and in "Measurement," in which students compare and order objects directly at around 48 months of age and may use an intermediate object to compare at around 60 months of age (Preschool Learning Foundations, Volume 1). These preschool activities directly enable the types of kindergarten through grade five learning trajectory described below.

K–5

The big ideas of data in these early grades include

- Data for understanding. What questions can we ask? What data do we need to answer it?
- Defining data: What is data and how where data collected?
- Representing and interpreting data: What does data look like and what does it mean?

These ideas are represented in the most important pedagogical and practical process through which data plays a role in making sense of the world, outlined in these four steps, from GAISE II and discussed above.

1. Ask a question (SMP.1: Make sense of problems and persevere in solving them)
2. Collect and consider data

3. Analyze data and develop meaning (SMP.2, SMP.4, SMP.7)
4. Interpret and communicate results (SMP.4, SMP.5)

An important distinction to consider is between *categorical* (non-numerical) data and *measurement* or *quantitative* data. For instance, consider a set of colored blocks in the classroom. “Color” is a categorical variable students could observe about each block. “This block is 15 centimeters long” is a measurement data point. The standards develop categorical data in grades K–3 and measurement data beginning in grade two. A description of the development of the data and measurement standards organized according to *categorical* vs *measurement* data is found in the *Measurement and Data, K–5* progression document. *[Note: link to the Progressions documents currently is:* http://ime.math.arizona.edu/progressions/; *a new link will be provided on the CDE’s website in future drafts*]

Categorical data ● Color (red, green, blue, yellow) of blocks in the class set ● Species of trees on the school grounds ● Identification of schools in the district as “elementary school,” “middle school,” or “high school.”
Quantitative (or Measurement) data ● Height (or circumference of trunk, or biomass) of trees on the school grounds ● Number of pages (or weight, or height) of books in the classroom ● Annual income for households in a census tract

Figure 1: Examples of Categorical and Quantitative Data

What questions can data help to answer?

All work with data should begin with noticing and wondering: “I notice that…” or “I wonder what…” or “I wonder how many….” (see http://www.nctm.org/mathforum/) To prompt wonder, teachers can ask: “What do you notice or wonder about here [in this context], that we could (count/measure/keep track of) to figure out or explore further?”

To establish effective routines, and to support language development in “I wonder” activities, it can be effective to provide these examples as sentence starters.

As students gain confidence in their ability to speak like mathematicians, statisticians and/or data scientists, the teacher should encourage students to generate questions themselves to build their agency in using mathematics to make sense of their worlds. A weekly whole-class “I wonder” routine—in which students propose questions to investigate by collecting data—would build a powerful practice of observing the world with a data lens, contributing to students’ development of modeling with mathematics (SMP.4).

In kindergarten, students compare the number of objects in different categories (K.CC.C.6) to answer “Which has more?” questions (*I wonder whether there are more square blocks or more triangular blocks on the desk?*). At first, the teacher suggests or specifies categories; eventually students generate ideas for classification. They also directly compare (as opposed to measuring with a unit or an intermediate) objects with common measurable/countable attributes to see which has more (K.MD.A.2, K.G.B.4) (I wonder which shape has more sides?; Which kind of block is heaviest?—using a balance or informal one-in-each-hand comparison, rather than a scale). “I wonder...” questions should explore both of these: two-category, “Which is more?” questions and comparison of objects according to length, height, weight, and countable attributes like number of sides. Student-generated questions provide opportunities to work on precision of language as well—for example, when students are asked to clarify what they mean by “bigger.” Mathematics discussions that are rooted in academic language will help students understand mathematical concepts more deeply as well as discover new ones. As the years progress, students or teachers may reach beyond the classroom to find contexts for their: “I wonder...” questions.

In addition to questions that can be answered with a single value, students can start to pose statistical investigative questions that involve multiple variables such as, I wonder if plants grow more with more sunlight? Or I wonder if age affects which color people like?

In first grade, measurement of length and time are the contexts to emphasize in generating questions (1.MD.A.2, 1.MD.B.3), along with continued work categorizing and counting objects (1.MD.C.4) and categorizing geometric objects by attributes (1.G.A.1). Second graders should continue to explore questions in length measurement (2.MD.D.9) and time (2.MD.C.7) contexts, and add money contexts (2.MD.C.8). When selecting "I wonder" questions, it is important to avoid situations that serve as markers for economic or social status, e.g. "I wonder who has the most expensive backpack," "I wonder who is the most popular kid in school," or, perhaps less obvious, "I wonder who has the newest shoes." It is similarly important to avoid questions about students' physical attributes, even those that seem innocuous such as height or arm length. Instead, some good questions to wonder about might be "I wonder what time it will be when the next person walks into the classroom" or "I wonder which book in the classroom is the most read," comparing events or objects rather than personal characteristics.

In third grade, contexts for questions to investigate using data should expand to include volume and mass measurement (grams, kilograms, and liters, but not compound units such as cm^3) in addition to the length, time, and money contexts from earlier grades (3.MD.A.2). Time measurements are refined to the nearest minute (3.MD.A.1) and length now includes half- and quarter-inches (3.MD.B.4). Beginning ideas of area give another possible context, limited here to areas that can be covered by a whole number of unit squares (3.MD.C.5, 3.MD.C.6).

In fourth grade, a significant context for data-investigation questions is classification and analysis of two-dimensional shapes (4.G.A.2). Incorporating this Geometry standard to help build data understanding can foster the important practice of analyzing by attributes—one instance of SMP.7 (Look for and make use of structure). Fourth-grade students also extend the set of units they work with (4.MD.A.1) and can generate data about area for more complex shapes. Fifth graders deepen their understanding of volume to include unit cubes, making this an important context for data-inquiry questions. A teacher could invite students to build a structure out of multi-link cubes and then collect data from the class by asking, for example, how many cubes they use in each of their different structures they built, or the height and width of their structures, and color of the blocks. Students can collect data on multiple variables.

In K–5, “I wonder...” questions come primarily from personal experience. See below for additional examples.

Asking Questions, Collecting and Analyzing data

Questions invite inquiry. An important part of students’ K–5 experience should involve coming to recognize that, when they choose and pose questions, they can collect or analyze data to find answers (SMP.4). Some of the most valuable conversations about data occur when students notice patterns in a data set and begin asking questions. Remaining alert for these everyday moments—perhaps in attendance, weather, or lunch-count data—may generate opportunities for discussing statistical investigative questions and exploring how data can help answer them.

As students come to pose authentic questions such as the ones described above, they should also encounter opportunities to help determine how data might be produced to answer them. In addition to producing data directly through their own observations, students should gain exposure to designing and using surveys and simple experiments to generate data. By producing their own data from their classroom or community (*How does age of students relate to their enjoyment of school? Does time on social media apps increase with age? How much waste is generated by different companies/our school?*), students recognize data as having context and deriving from observation and measurement, and they come to see data (and mathematics more broadly) as a tool to help think about their worlds. Data gathered by others (such as those in the data talks) can help to answer questions students generate about their own communities.

When choosing data tasks that include categorizing and counting, consider the grade level expectations for counting (up to 10 objects scattered, or up to 20 if arranged in a line, array, or circle, in kindergarten [K.CC.B.5], 120 by the end of first grade

[1.NBT.A.1], and up to 1000 by the end of second grade [2.NBT.A.2]). Such tasks can also be structured to build place value understanding.

In kindergarten, once students notice things in a context and wonder about a question, they describe measurable, countable, and observable attributes of objects or situations (K.MD.A.1, K.G.A.1, K.G.B.4), and classify objects and count the number in each category (K.MD.B.3), such as categorizing a set of cubes by color. In this last context, both “this cube is red” and “there are 13 red cubes in the set” are data points. Notably, most work on *number* in kindergarten should be with numbers representing quantities of objects (SMP.2); thus, most numbers encountered in kindergarten are actually data.

In first grade, students explore their time and length questions by measuring lengths of objects which are a whole number of units (1.MD.A.2) and telling and writing time in hours and half-hours. Counting and categorization situations should include up to three categories (1.MD.C.4). Second graders measure length to the nearest whole unit (2.MD.D.9), using different standard units (centimeters, meters, inches, feet) (2.MD.A.3) and several tools (2.MD.A.1) and measure time to the nearest five minutes (2.MD.C.7).

Students in third through fifth grades refine their measurements of lengths and time, and expand the set of units they use; and they add area and volume measurement to their repertoires (as described above in “What questions”). By the fifth grade, students should understand that data sets can include different types of variables, such as categorical and quantitative. They should recognize that an individual instance or object can possess attributes that exemplify these different types, and should have gained experience measuring, characterizing and analyzing such diverse types of data and associating them together.

An important understanding that students need to develop through grades K–5 is the idea of variability and variables. When students ask questions such as: How high are the plants in the classroom? they are considering one variable: height. When they consider whether older students spend more time on social media apps they are collecting bivariate data—with two variables—age and time. When they make their own data visualizations, as seen in Vignette 2, they may collect data on multiple variables. Multivariable thinking is important to develop through the grades.

Interpreting and Communicating Results

Sorting objects into two categories and representing these categories by their count (K.MD.B.3) is a first example of students representing data to help make sense of their worlds (SMP.4). First grade students organize up to three categories and ask and answer questions about the relative sizes of categories and about the total number of data points.

Second grade begins an expanded focus on data representation, introducing line plots (whole number units only; 2.MD.D.9), picture graphs, and bar graphs. These graphs can be used to answer put-together, take-apart, and compare questions (2.MD.D.10). In third grade, *scaled* picture and bar graphs are added as a tool for visualizing "how many more" questions (3.MD.B.3), and line plots may have half-unit and quarter-unit markings as appropriate (3.MD.B.4). In fourth grade, line plots may display additional fractional units (to eighth-units), and be used to answer additional questions about differences—between maximum and minimum measurement, for example.

Fifth grade does not extend the expected set of data representations, but students do use line plots in a sophisticated way that sets the stage for understanding the most common measure of *center* for a set of data—the *mean* (commonly called the average)—in sixth grade. Namely, fifth grade students use a line plot to decide how a repeatedly-measured quantity could be redistributed equally (5.MD.2): "Given different measurements of liquid in identical beakers, find the amount of liquid each beaker would contain if the total amount in all the beakers were redistributed equally."

While the data visualizations mastered by fifth grade only include picture graphs, bar graphs, and line plots, students do not need to be restricted to these. Each of these represents repeated measurements of a *single* varying quantity; science curricula in particular, and many questions of interest in general, require the consideration of relationships between *two or more different* changing quantities, such as erosion and time (NGSS 4-ESS2-1 Earth's Systems) or length or direction of shadows and time (NGSS 5-ESS1-2 Earth's Place in the Universe). Such reasoning involving multiple variables is an important aspect of modern encounters with data, and students should experience it at all levels. Although the scatter plot, a crucial data representation tool for two varying quantities, is not mastered until eighth grade (8.SP.1), it must be explored informally much earlier for students to be able to meet the eighth-grade expectations. For example, students can plot quantities changing over time (e.g. height of a plant, length of the day, high temperature for the day, temperature of a glass of water every minute for an hour), with time on the horizontal axis and the changing quantity on the vertical. Once such a plot is created, it is an excellent context for a "notice and wonder" discussion.

In recent years, new technological tools and developments in data science have prompted an explosion in interesting data visualizations, many of which are quite comprehensible to young students with some exploration. Experiences with different visualizations will further expand students' sense-making opportunities and encourage them to think about what they can understand looking at data sets in different ways. The examples from the New York *Times*, Youcubed, and other places illustrate multivariate data displayed in creative ways. This example shares the most popular songs each summer: https://www.youcubed.org/resources/whats-going-on-in-this-graph/. Newspapers and online news sources offer other examples; student-gathered examples help to build buy-in for a "can we figure out what this visualization is trying to help us to understand?" routine.

Interpreting data is a matter of making inferences from the data available. While students will encounter quantitative and nuanced techniques for making inferences in later grades, they should nevertheless encounter opportunities to make claims and infer conclusions across their K–5 years (SMP.3). When they do, students should learn both to wonder whether patterns or trends they notice in data extend beyond the particular group that generated the data, and to be skeptical about such extensions to larger populations (including considering ways in which the group might not be representative of the larger population). Additionally, students should learn that good claims draw upon data as evidence and that they always come hand in hand with a degree of uncertainty. Modeling the use of appropriate terminology such as “tends to,” “typical,” “usually,” and “similar” can help lay important groundwork for this concept (Rugin, 2019).

Preparing for the major data science work of grades 6–8

Understanding Variability: Variability is everywhere, and understanding variability is the core of developing data sense. While understanding of statistical variability and distributions is not in the standards until middle school, it is essential that K–5 students encounter many experiences with variation, including counting, measuring, and observing quantities and characteristics that vary in order to be prepared for the first big idea in the Grades 6–8 section below. In particular, their encounters with data representations should highlight important ideas that set the stage for more involved work with distributions. When working with visualizations of data, students should consider not only the most popular value in a dataset (the mode) but also describe the shape and spread of data distributions. Identifying the maximum and minimum values of quantitative datasets can help students appreciate the concept of range as a measure, and looking for clusters and gaps in a distribution can begin to help them attend to its shape. As they engage in experiences where they produce their own data through measurement, teachers should highlight for students the variation that results. Measuring the same variable on multiple individuals or objects, for example, results in data that vary, and students should consider the causes or sources that might have given rise to the variation they have observed, working as they do so to differentiate between variation and error. For example, if students plant a particular variety of flower seed at multiple locations around the school, then measure the plants’ height and the amount of sunlight each month, they can conduct investigations into the ways plant growth and sunlight relate to each other. They should discuss and describe any patterns in their bivariate data, and discuss reasons for the variability. Finally, they should consider their own measurement techniques, and how confident they are that they all measured the same way (so that if someone else measured, they would get the same height or sunlight).

Randomness, probability, and uncertainty: Randomness is a complex idea encompassing uncertainty *and* a level of predictability. When (blindly) drawing a cube out of a bag containing three blue cubes, two red cubes, and one yellow cube, nobody can predict with certainty what will happen on a single draw. But, over many draws, the person who always predicts a blue cube will be right about half the time. Activities that demonstrate this can be used to generate data for many of the explorations of the big ideas above, which will leave students well-prepared for a more formal treatment of randomness and probability in middle school. At this point, students should begin to conceive of probability as a measure of the chance that something will happen, seeing it as a basic measure of certainty or uncertainty.

Technology: California's 2018 K–12 *Computer Science Standards* include computer-based data sorting, categorizing, and visualizing for students in grades K–2 and 3–5 (CS K–2.DA.8, CS K–2.DA.9, CS 3–5.DA.8). These standards are important preparation for middle and high school use of data software to visualize and interpret large data sets.

Finally, it is worth noting that (as in science and other fields) many questions that students might wonder about will not be fully answerable using K–5 tools. It is important that teachers have resources for helping students figure out which aspects of questions can be investigated with currently available tools, and have some understanding of data science tools which students will encounter later. For example, many will wonder about relationships between two different variables*: If I get up earlier, do I feel tired earlier in the afternoon at school? Do students who skip lunch eat more candy in the afternoon?* When one of the variables is categorical (like the skipping lunch question), separate line plots can be made for each category and the line plots compared. When both variables are quantitative, students could input data into CODAP and investigate the relatioships by plotting their data on graphs, observing their distributions, and adding line plots. Another option is that one of the variables can be made into a categorical variable by defining categories in terms of the quantitative variable. For instance, waking-up times could be classified into “early” and “late” (ideally with a student-generated cut-point between early and late) and then dot plots of “time in the evening when I felt tired” created for each category.

Vignette: Logan from Kindergarten through Grade 5

A small sampling of Logan’s data science experiences in grades K–5 is described below. This is not intended to capture *all* of their data science experience, only to indicate a development towards powerful uses of data to understand their world. In each grade, Logan generated questions and gathered data (steps 1–3 in the process described at the beginning of this K–5 section) and represented and interpreted data (steps 4–5).

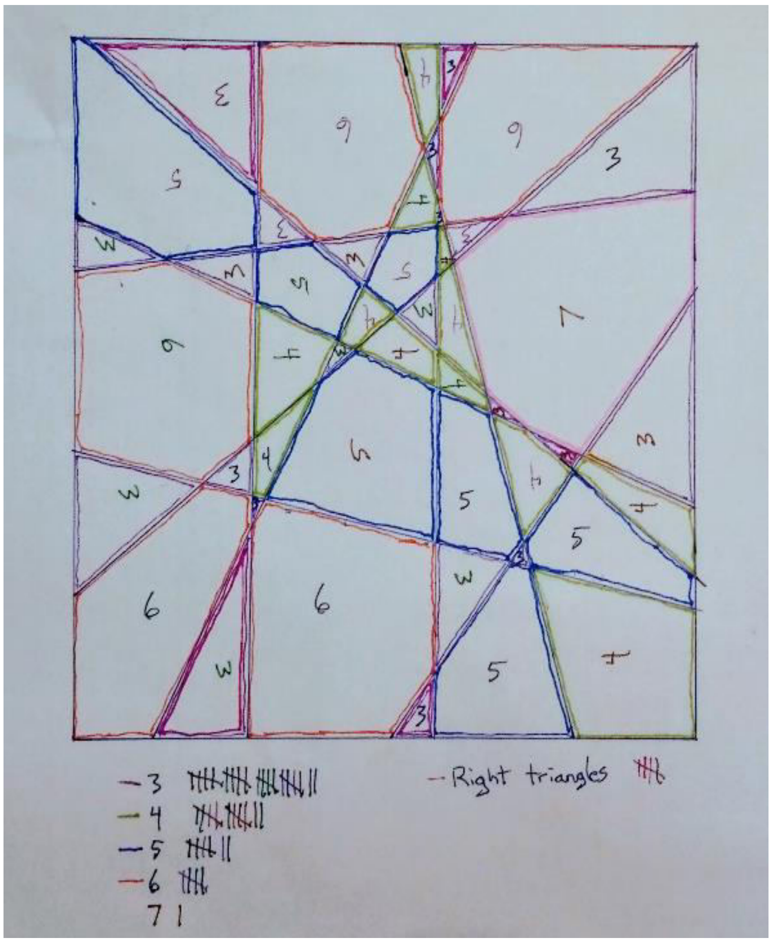
Logan entered kindergarten as a very active child, and always gravitated towards the group of children who ran the longest and climbed everything. Logan’s teacher asked lots of questions of students about what they noticed in the classroom and around school, both inside and out. These ranged from specific “how many in each category” questions (how many classroom doors of each color are there?), to direct comparison questions (which slide is taller, the blue one or the green one? How do you know?), to, eventually, *types* of questions: What are some things at school whose size we could compare? Recording students’ observations and category counts allowed all students to pose and answer “relative size” questions.

In first grade, student teams were asked to think of two similar things at school, such that they weren’t sure which was taller, and then to find a way to compare their heights. A variety of materials was available to use in the comparison. Logan’s team was able to compare the height of the slide in front of the school with the height of the slide behind school, measuring the height of both using towers of large DUPLO® bricks. The whole class used their data to discuss how much taller the slide in front was. After this, Logan wanted to build DUPLO® towers to measure height and length of lots of things, and was disappointed that the class didn’t have enough bricks to measure the height of the school (and that their teacher wouldn't let them climb the school). As a class, students checked the length of the day (sunrise to sunset) each day, and maintained a running visible tally of the number of school days with less than 11 hours of daylight, 11 to 13 hours, and more than 13 hours, for the entire year. As a class, they discussed what they thought might happen to the number of hours of daylight in the future and checked the data a month later to see whether their predictions were correct.

Logan's second grade class marked their own yard sticks (marking a wooden blank in inches using only a 3-inch by 5-inch card), and then used them extensively to measure objects of interest to the nearest inch. Later they added centimeter markings to the other side of the yardstick, and discovered that measuring the same things with smaller units led to larger number measurements. When choosing an activity to time, Logan's group decided to time and record the amount of time in a week that team members spent reading in school, and compare those measurements over several weeks (this had the benefit that team members read much more during those weeks!). Other teams measured time spent playing outside, listening to announcements, and working at math stations. Teams made line plots of their data, and compared the line plots of different activities to discuss how students typically spend their school time.

As mass and volume became available in third grade as characteristics to measure, Logan's class used length/height, mass, and volume measurements to examine collections of objects. The line plots of the masses and the lengths/heights of the objects in the science corner looked quite different from each other; similarly for the line plots of volume, height, and mass of all objects in the room which hold water (vases, cups, etc.). Logan's team had a great disagreement about whether a taller vase should hold more water than a shorter vase; the class eventually decided that this was usually but not always true.

One of Logan's favorite activities in fourth grade was one that combined data work with classifying shapes by attributes: collaborative art pieces: Each team had a ½ meter by ½ meter square on the board, and each student in the team drew in two edge-to-edge



straight lines of their choice, using their meter sticks. Then one student in class chose a shape to try to find in the drawings, and each team outlined each new instance of that shape they found and described how they knew it was a (triangle, rectangle, right triangle, quadrilateral, etc.); this was repeated for several other shapes. They made an individual card to represent each piece of artwork, using the card to represent the artwork the different variables they measured for each piece (how many triangles, how many instances of each color, how clean or messy each line was, etc.) When they had made a full set of cards, they sorted them in various ways, then made a table to compare the tallies for the different pieces, discussing how the different features of the art and the process of creating it that might help explain the variations in their data.

By fifth grade, Logan and their classmates had constructed many line plots, and thus often wondered about quantities that vary on repeated measurement: The cartons of milk from lunch say they each contain 8 fluid ounces, but yours feels heavier than mine; am I getting less or are you getting extra? The weather site says the average high temperature here is 57℉ (degrees Fahrenheit) in November, but today it got up to 65℉. How can we check whether this month is near average? To explore the first, the school donated 20 cartons of milk to the experiment When they examined the dot plot of their measured volumes, they saw that it had a tightly clustered shape, with a minimum measurement of 7.8 fluid ounces and a maximum of 8.2 fluid ounces, and that the most frequent value was 7.9 ounces. One student in the group thought that some milk probably remained in the containers, so the group spent a while trying to figure out how they might identify how much had been left inside (teams came up with several methods). For the second, the class recorded the daily high all month, recorded them on a line plot which also had marked the “average” high temperature from the weather site, and used the line plot at the end of the month to discuss whether it was consistent with the stated average (without computing an average of the data).

Students like Logan, with a rich variety of experiences using data to explore contexts and questions of interest, will be well-prepared to use mathematics, and data in particular, to make sense of their ever-expanding worlds. Data sets will get larger, and contexts wider, in ensuing years.

Grades 6–8

Middle school includes a big expansion in important ideas. The big ideas of data science include

- Data in the world: exploration, interpretation, decision making, ethics
- Statistical variability: Describing, displaying, and comparing

- Sampling to understand a population: randomness, bias, how many?
- Are they related? Multivariate thinking
- What are the chances? Probability as the basis for data-based claims

As in earlier grades, students experience data science as a tool to help understand their worlds via a process that begins with wondering questions. This is also the beginning of the mathematical modeling cycle (Pelesko, 2015) and the statistical and data science exploration process, and of investigations in science (NGSS Lead States, 2013).

The GAISE II Statistical and Data Science Exploration Process:

1. Curiosity and question asking
2. Collect and consider data
3. Analyze data and develop meaning
4. Interpret and communicate results

This process, beginning with noticing and wondering, often gets lost in the details of step 3, which contains the different statistical methods that have been developed for the analysis of data. It is crucial to keep all work with data tied to authentic questions. Prediction is a key activity that builds student ownership of the process and conclusions; it also builds a habit of asking “does this make sense?” (SMP.1) by comparing results with expectations.

Data in the world: Question asking, exploration, interpretation, decision making, ethics, technology

What functions does data science play in the modern world?

- **Question Asking and Exploration:** Data science and statistical exploration starts with questions that are posed by students. When students are invited to wonder about situations, and when they are given interesting datasets they will become curious and can ask questions of data that they can explore and investigate. Data exploration includes understanding the context and situation, data should never be abstracted from their context. Students can look for hidden patterns and associations. Any patterns or associations discovered can lead to new conjectures or questions to investigate further. In eighth grade, students can begin this process with datasets that include multiple variables, such as those given in CODAP (https://concord-consortium.github.io/codap-data/). As mentioned in the chapter introduction, vast quantities of data are collected every day, and only a small fraction are analyzed.

- **Interpretation:** Every encounter with data should revisit the context from which the data originated, interpreting results of data analysis in that context. This includes answering any questions that began the encounter and reporting any other associations or patterns that were discovered.
- **Decision making:** Commonly, data is used to inform decisions following the question/data/represent/interpret process. Often, however, data is used to justify and explain a decision, even if data didn't play a meaningful role in the decision. There is great potential for abuse here, by including data collections that support the predetermined decision and leaving out those that do not.
- **Ethics:** Modern ubiquitous data collection raises a host of ethical questions, both about how and what data is gathered and stored, who is included or excluded in the data, and how that data is used and presented. Middle school students need to understand their own online data footprint (for example, how companies aggregate information about individuals to create detailed profiles) and should confront scenarios in which they must make decisions in hypothetical situations involving data exposure, consent for data collection, etc.
- **Technology:** California's 2018 K–12 *Computer Science Standards* (*CSS*) expect students to make use of computers for data organization and visualization (*CS*S 6–8.DA.8). More importantly, given the amount of data collected and stored today, real-world datasets are incomprehensible without such computer assistance. Students should use modern data software extensively, especially for organizing and displaying features of data set.

Describing, displaying, and comparing statistical variability (Grades 6–7)

Sixth-grade students build on earlier experiences by distinguishing between statistical questions, which can be investigated using data that varies (analysis of social media usage by age of students), and questions without variations in (correct) responses (How many days are there in January?) (6.SP.A.1). When considering a statistical question, they understand that the variation in numerical data has a distribution which can be described by its center (first the median, then the mean), its variability (also called spread—described both qualitatively and via a numerical measure, either inter-quartile range (IQR), range, or mean absolute deviation), and an overall shape (including descriptors such as symmetric, skewed left or right, peak, gap, and outlier) (6.SP.A.2, 6.SP.A.3). As students explore datasets, they can produce visual representations of the distribution of their data; they can look at the shape of distributions that have different measures of center and spread, and develop visual understandings of the shape of distributions.

Students should have experiences, beginning in sixth grade, deciding which measure of center is a more useful descriptor of a typical value for data sets with different shapes. Because the mean is sensitive to extreme values, the median is often a more useful measure for skewed distributions; in this case, the inter quartile range is a useful measure of variability. For some distributions—with multiple clusters, for example—students may decide that neither median nor mean is a useful measure, and might decide that a single number cannot reasonably represent a typical value (6.SP.B.5).

Two tasks that reinforce the notion of these standard measures and replace rote disconnected calculation with conceptual thought are the following:

A. Students form a “name count line” creating a human graph to depict how many letters are in their first name. (All the students with five-letter names stand in a line, all those with four letters form a similar line to one side, and those with six letters form a line to the other, etc.) Then the teacher instructs one student from each end of the human graph to sit down. After repeating this multiple times, only one or two student(s) are still standing. If one, that student represents the median name length. If two, the median name length is halfway between the name lengths of the standing students.

B. Students are invited to explore the CODAP dataset of 4 elephant seals: https://codap.concord.org/releases/latest/static/dg/en/cert/index.html?url=https://concord-consortium.github.io/codap-data/SampleDocs/Science/Biology/four-seals/Four_Seals.codap

   The dataset includes data on the paths taken by the seals – visible on a mapping tool, the distance they swim, their latitude and longitude, the depth and temperature of the water and more.

   In groups students are invited to explore the data and form investigative questions. Students start by plotting different variables with the graph tool, to consider the shapes of distributions. They choose to display the mean and median and consider how the measures of center relate to the visual distribution of the data. They form questions they are curious about: Do certain seals prefer deeper water? Does the distance seals swim relate to the temperature of the water? As students explore these questions they plot two variables on a graph and consider the slope of the relationship, they even add a third variable which is shown through color coding. Students learn to be comfortable investigating data, making use of measures to learn about their data.

Visual representations of distributions include box plots and histograms in sixth grade, adding to the line plots (called dot plots from grade six onward) from earlier grades (6.SP.B.4). In addition, students learn to report and interpret measures of center and variability, and descriptions of distributions, in the context in which the data arose (6.SP.B.5). Seventh- and eighth-grade standards do not include additional representations of single-variable data sets, but these students should continue to create visual displays of such distributions.

In seventh grade, comparisons between two populations with similar variables is a context in which students describe and create visual displays of data. They can plot data and draw from different statistical methods such as creating box plots and dot plots to informally assess the degree of overlap of two populations, and students should be able to describe the difference between the two centers in terms of the measure of variability they use for the distributions.

Vignette

Alex did not enjoy learning about mean, median, and mode. He often confused the different measures and felt they had little meaning. His parent contacted Maria, his teacher, to let her know that he was expressing frustration about the meaning of the terms since his last assessment. Maria realized Alex was not alone since many of the students were still struggling with the meanings of these measures of average. Using this feedback as formative assessment, Maria approached the students with the idea to build physical models so they could experience the averages in visual and physical ways, encouraging important brain connections. Maria gave her students cubes and asked them to make 6 different towers of cubes that represented the numbers 1, 6, 3, 2, 4 and 2. She asked them how they might construct a physical proof to show the mean of the numbers. Some of the students were able to calculate the answer; however, she kept pushing them to build a visual proof but remained open to multiple means of representation. This strategy, based on the UDL guidelines, allowed Maria to provide students scaffolds and supports to help highlight the patterns of language, and draw on background knowledge to express what they know in ways that are authentic and meaningful. Alex and his group members came up with the idea of moving the cubes form tower to tower to show that they could make six towers that were all the same height. They just needed to average out all of the blocks. Alex and his group excitedly explained to the class how they had made a physical proof of finding the mean of the blocks. They shared the calculation with the class and compared it to the method they used of moving the blocks. After her students had discussed finding mean, Maria asked them to make a visual proof for the median and the mode.

Sampling to understand a population: randomness, bias, how many? (grades 7–8)

Prior to seventh grade, students' work with data has focused exclusively on using data to understand, describe, and compare the particular collection of objects or situations that were observed or measured. For example, to calculate the median highest temperature on school days in September as above, students would record the highest temperature on *each* school day.

Seventh grade includes the first introduction to *sampling*, the process of collecting data from a subset of a population in an attempt to understand or describe the whole population. This represents a big jump in sophistication from earlier work. Early experiences with sampling should first describe the measured variables for the sample (favorite lunch, number of minutes looking at screens, recorded for all students in the sample for one week), followed by team and class discussions about whether the description extends. For instance, if all students who come in to play basketball before school are asked to track their screen usage for the week, the class should discuss whether they expect the average of 862 minutes to be close to the average for everyone at school—and if not all students, then perhaps close to the average for some smaller definable group of students. Many similar discussions, with some obviously non-representative samples, help students understand the idea of a *random sample*.

If researchers decide to gather data from 40 members of the population, then their collection of 40 members is *random* if it is chosen in such a way that every possible subset of size 40 has an equal chance of being selected. It is important for students to have multiple experiences selecting samples from known populations in ways that are random (for instance, drawing numbered ping-pong balls from an opaque bag or drawing student names on identical slips of paper from a hat) *and* in ways that are not random (for instance, asking survey questions only of the students who sit near you in class). The goal is an understanding that random sampling tends to produce samples that are *representative* of the population—that is, their distribution of the quantities under consideration are close to the distribution for the population as a whole

(7.SP.A.1)—and a sense for the variability when using samples to make inferences and estimates for a population (7.SP.A.2).

Non-random sampling (such as attempting to understand the school as a whole by collecting data only from one’s friends, or by asking about eating habits at the gym after school, produce *biased* conclusions, even when the bias in the sample selection might not be obviously linked to the quantity being measured in the measurement or observation. *Bias* does not here refer to temperament or outlook (prejudice), which is one meaning of the word; instead; it means a *systematic error*.

Once random sampling becomes an available tool, the pool of questions open to students’ inquiry expands greatly: “I wonder how long on average it takes students from different grades to get from home to school?” “How do students who live in different areas spend time at the weekends?” “How much food is wasted in the lunchroom every month?” are all questions that could form a data exploration, as students consider their sample, which variables can be defined and collected, and engage in the four part exploration process.

Sampling is introduced in the seventh-grade standards and does not appear again until high school, but much of the eighth grade work with *bivariate* (two variable) data will make use of sampling, so it is important to continue activities that help understand *random sampling* through eighth grade as well. Students often believe that arbitrary sampling schemes (first 10 students I meet or every tenth student alphabetically) are random; they need to understand the difference between these schemes and choosing *by chance* so that every possible sample has an equal likelihood of being selected.

Vignette

Rosa has reflected on her seventh-grade students and how they have responded to the probability activities offered in her curriculum material. Overall, Rosa has not been satisfied with student understanding of random sampling. She decides to give students another, more visual and physical experience of the concept. Her plan calls for six paper bags filled with differently colored cubes. The sum of cubes and the color distribution of the cubes in the bags reflect the following:

Bag 1, 15 total: 15 blue

Bag 2, 12 total: 11 blue and 1 red

Bag 3, 20 total: 15 blue, 4 yellow, 1 red

Bag 4, 10 total: 5 red and 5 yellow

Bag 5, 12 total: 5 blue, 4 red, 3 yellow

Bag 6, 20 total: 8 blue, 8 red, 4 yellow.

Rosa explained to her students that their task was to determine the contents of each bag through sampling. She chose not to tell them how many times to sample but she did tell them to sample from the bags by selecting one cube at a time and then putting it back into the bag. Rosa also asked students to determine the chance of drawing a blue cube from each bag.

Students engaged in the activity, organizing how they would collect and record their information. When each group of students felt they had determined the number of cubes and color distributions of the contents of each bag, she asked them to choose which bag belonged to which card showing the contents of each bag. Rosa had filled the bags differently and made sure to have two different bags where the probability of drawing a blue cube would be one and another would be zero. After the activity and class discussion, Rosa was happy to hear her students later, talking about situations where the probability was one or zero as well as everything in between. Her students recognized the number of times they sampled usually led to better predictions about the contents of the bags. They also realized sampling without replacement would have shown them the exact contents of the bag. The class engaged in a vibrant conversation about sampling with and without replacement, recognizing that it would be unproductive to draw all the cubes if there were a million.

Are they related? Two changing quantities (grade 8)

Prior to grade seven, students work with a single collection of data measuring a single variable. In grade seven, they compare the same variable measured across two populations, either by actually measuring the whole populations or obtaining estimates for the distributions via sampling.

In eighth grade, the focus is *bivariate data*: Two quantities or categorical variables measured or observed across a population, or across a sample drawn from a population (8.SP.A.1). This work has important connections with linear equations and modeling.

The *scatter plot* as a visual representation of *quantitative* bivariate data is one of the most important ideas introduced here. A survey of students collecting both time and distance for traveling from home to school might reveal *clusters, outliers,* and any of various types of *association* (positive, negative, linear, non-linear). Students should describe such patterns in a scatter plot and interpret them in the context of the data (8.SP.A.1).

Students can explore large datasets—such as earthquake data from California—and explore bivariate relationships like, how does the location of earthquakes in the database compare to the magnitude of the earthquakes? They can plot the data using graphing tools and consider associations, data distributions, and relationships.

If students vary the weight added to a simple cart and measure the distance it travels when released at the top of a ramp, then plot the results on a distance (vertical axis) vs added weight (horizontal axis), they will likely see a relationship. This association between the two variables can then be *modeled* by a line if the association appears roughly linear (line-shaped). In eighth grade, students choose a line to fit the data by visual approximation on the scatter plot, and compare and argue for whose line fits "best" (8.SP.A.2). They then interpret the meaning of the slope and intercept of their chosen model line, and use the line to make predictions for one variable when the other variable is specified (8.SP.A.3)

Finally, eighth grade students use two-way frequency tables as a tool to see associations in bivariate *categorical* data (8.SP.A.4). For instance, they might survey their class, including questions about favorite color and favorite genre of books, then input the data into a spreadsheet, organize the data and calculate relative frequencies in rows to explore possible relationships between the two variables.

What are the chances? Probability as the basis for data-based claims

Randomly selecting from a population and measuring a characteristic (in which variation is expected across the population) is a *chance process*: It may result in different results and its outcomes follow some *distribution*.

Probability is the way we express the chance of an outcome as a number between 0 and 1 (7.SP.C.5). Probability is combined with statistics in the grade seven standards; statistics and probability are historically linked because statistical claims and estimates are based on the mathematical field of probability. Models that draw from data science and offer predictions of events, such as voting in elections, draw from probabilistic reasoning.

The connections between probability and statistics is often not clear to students, especially when their experiences focus on procedures and calculation rather than exploration, context and interpretation. There is much work with probability that does not support statistical reasoning (for example, calculating theoretical probabilities for the sum of two dice without using those theoretical probabilities to decide whether a given pair of dice are likely fair), and middle school probability experiences should be carefully designed to support reasoning with interesting and meaningful data.

In seventh grade, students gather data to estimate the probability of outcomes by observing their long-run relative frequency; that is, they compute *experimental probability*. Consider repeating this experiment 150 times: draw a marble from a bag with marbles in it, record its color, then put the marble back in the bag. If we get a blue marble 32 times, our estimate for the probability of getting blue on any particular draw is 32/150 (7.SP.C.6, 7.SP.C.7.B).

Compare the marble experiment just described to this one: Put the following marbles in a bag (all identical except for color): 16 blue marbles, 31 red marbles, 16 green marbles, and 12 white marbles (75 total marbles). If you blindly pull a marble from the bag, what is the probability that you will get a blue marble? If you do this 150 times (putting the marble back each time), about how many times do you expect to get a blue marble? After calculating this expectation, students should construct an algorithm or pseudo-code to run the simulation 150 or 1500 or 15,000 times to compare with their theoretical expectations (CS Standards 6-8.AP.10).

Note the difference between the questions in the previous two paragraphs: In the first, students use long-run relative frequency to estimate probability; in the second, students build a (*theoretical*) probability model and use it to estimate long-run frequency (7.SP.C.7). If a marble experiment is then performed and relative frequencies of outcomes do not seem close to predictions from the probability model, then students need to be able to discuss possible sources of discrepancy (7.SP.C.7): Perhaps the green marbles feel different and tend to be drawn more frequently than predicted. Maybe somebody changed the mix of marbles in the bag. Or perhaps not enough draws were performed to see the relative frequencies approach the probability model.

Finally, seventh grade students find probabilities of compound events (events which are made up of several simple events; for example, drawing two marbles from the bag of 75 described above and getting one white and one blue marble) (7.SP.C.8).

The specific calculations above are not central to the data science progression, but recognition that some events (repeat the draw 5 times, get all blue; or repeat the draw 5 times, obtain WBWWB in that order) are *much* less likely than others (repeat the draw 5 times, get 3 white and 2 blue) is key to understanding claims made from data.

In fact, most statistical claims depend on a comparison of a (theoretical and hypothetical) probability model with observed data, as in 7.SP.C.7. For middle school students to be prepared for future data science work, they need experiences to build a sense that more data tends to produce relative frequencies closer to actual probabilities.

Invite students to explore rich datasets, such as the distribution of births in the US – and consider questions of probability, that they can explore, such as: what is the chance that two people share the same birthday? This is a question that could be explored theoretically or experimentally. (More at https://codap.concord.org/releases/latest/static/dg/en/cert/index.html?url=https://concord-consortium.github.io/codap-data/SampleDocs/Mathematics/Probability/Birthdays/Birthdays.codap)

Vignette

Quincey started middle school without a lot of interest in math class. Quincey had always been interested in how the world works, and science and social studies were their favorite classes. Quincey had not had much experience of math class connecting to areas of interest.

Quincey's sixth-grade math teacher, Lori, was determined to change this for all students. Lori knew that the data science standards in sixth grade would give Quincey an opportunity to use real data from the world and to understand that they could ask questions of data and see the connections between mathematics and life. Lori decided to use an activity to explore the "shape" of data: The context is hurricanes in the Atlantic Ocean using real data collected from five years of hurricanes spread over four decades. Quincey loved the opening discussion in class when the students all discussed the 2017 hurricane data displayed on a line plot. Quincey and the class were really interested in the number of hurricanes that were in category 0—tropical storms. Next, students worked in groups where they studied hurricane category data for the years 1977, 1987, 1997 and 2007. Each decade's data was presented in different ways: bar graph, line plot, tables and sentences. Quincey enjoyed the analysis and was taken with the different ways of displaying data as well as the changes in the spread of data.

Throughout the discussions Quincey asked questions about the science of hurricanes. *How do they develop?* he wondered. *What makes them get larger? What is the difference between a category 3 storm and a category 5 storm?* At the close of the lesson Lori was convinced that students understood that different visual displays of data can make it easier to see the shape of data. The shape of the data on the displays helped students see how a situation might be changing over time. The class reflected that the changes were easier to see in line plots and histograms versus the data being shared in writing or in a table of values. Quincey decided to further investigate the number of category 4 and 5 hurricanes over the past 100 years and how these storms become stronger, and they set out to gather more data and ask questions of the data. Others in the class decided to investigate why the number of category 4 and 5 storms are increasing.

High School

In this outline for data science in high school, two sections of guidance are provided: (1) experiences and expertise in data science common for all high school students, and (2) experiences and expertise for a high school pathway with a data science focus (expanding on the pathway outline in Chapter 8, Grades 9–12).

Computers have become central to modern life. Whether laptops, phones, so-called “smart” appliances, medical records systems, exercise trackers, GPS location recording, payment methods, or other forms, computers are involved in most of our transactions. Every interaction with a computer generates data about that interaction (which is collected and saved)—but very little of this data is analyzed and interpreted.

Even as computers have led to the collection of vast amounts of data, computational tools (including both computer hardware capabilities and advances in algorithms) have dramatically altered the available methods for making use of and communicating interpretations of data. In fact, meaningful analysis of large or multivariate data sets is impossible without computer tools.

For many questions about which students might wonder, existing data sources might provide the necessary information. Designing data collection to obtain exactly the desired data for answering a specific question (the classical statistical experiment approach, still the main approach in grades K–8 above) is expanded to include descriptive techniques for analyzing multivariate data, critical questioning skills to interrogate pre-existing data's suitability for the investigation, and an understanding of ways to access and acquire data through the internet. This understanding-extraction uses two processes: data description methods, both visual and numerical, to investigate conjectures and discover patterns; and model-building to test conjectures, make predictions of future observations, and evaluate the predictive success. These huge existing data sets are not collected in order to answer a particular question, do not typically represent random samples, and are often missing data or are otherwise “messy.”

Data science should be understood as a broad term encompassing many tools relevant to learning from data. These include tools of traditional statistics classes, but also include computational tools to address the massive size and complexity of many of today’s data sets, and disciplinary knowledge of the field generating the data. Thus, data science is an inherently interdisciplinary field that uses scientific and statistical methods and processes to derive understanding, insight, and predictive ability from (often unstructured) data (Dhar, 2013).

Data science for equity and inclusion

An important way in which educators can offer social and emotional support to students is by designing engaging lessons that allow students to connect with the ideas being taught. Traditional mathematics lessons that have taught mathematics as a set of procedures to follow have resulted in widespread disengagement as students see no relevance for their lives. This is particularly harmful for students of color and for girls—who receive additional harmful messages that mathematics is not for them. Data science is a field that provides opportunity for equitable practice, with multiple opportunities for students to connect with ideas and for them to receive messages that data science is a field in which any students can excel.

Walton and colleagues (2015) have shown through numerous studies that many students, particularly girls and students of color, do not feel that they belong in certain disciplines. This is often due to a history of negative and off-putting messages (Chestnut et al, 2018). Other studies have shown that different topics and teaching approaches can lead to feelings of belonging or not belonging (Boaler, 2019; Boaler, Cordero & Dieckmann, 2019). Data science is well placed for teachers to offer students a sense of belonging as students are invited to investigate real data that will often be relevant to their lives. This meaningful engagement will offer students opportunities to develop self-confidence and self-efficacy. When teachers offer students the role of being data investigators, asking and answering questions with data, they can take an active role in their learning, encouraging opportunities for self-motivation and goal setting. Important principles in the teaching of data science, that will offer the greatest chance for social, emotional, and academic development, include the following:

- *Mindset and Belonging Messages*
  At frequent times students should be reminded that data science is a field in which all people are welcome and can succeed. In line with successful interventions in mindset and belonging students should also be reminded that struggle is an important part of learning, that all students have times of struggle, and that the difference between successful and unsuccessful students is the way they respond to times of difficulty. Share with students examples of successful people inside data science, that highlight gender and racial diversity (see for example: https://www.youtube.com/watch?v=KYvhoH5AzHA&feature=emb_logo).
- *Use Real Data*
  Data science gives an opportunity for students to be asking questions of real data sets, developing social awareness, and investment in the solutions they discover. When working with secondary data sets (data obtained from others, rather than collected by students), teachers should choose meaningful ones in order to give students an important connection to the content they are learning and opportunities to take the perspective of others, which will help them develop empathy. When teachers use local data sets they can also help students feel like they are important members of their community – as they explore questions and

find answers to local problems that they can help with real data. Identifying problems and finding solutions will help students develop responsible decision-making.

Some teachers worry that they cannot provide culturally sustaining connections for their classes as they are not experts in the cultures of all their students, but real data sets from different communities provide opportunities for students to bring their own knowledge and expertise to data rich problems. There should also be times when students are invited to collect data from their own community and build their own data sets. Students can pose questions that are important to them, including those with cultural meaning, collecting data from their own lives and communities. As Paris (2012) describes students will be fostering and sustaining "linguistic, literate, and cultural pluralism." The act of collecting data provides an important learning opportunity for students to understand decisions that need to be made around the collection and organization of data as well as how to deal with uncertainty in their data. Students will be the ones with important expertise in these investigations.

- *Focus on Collaboration and Communication*
  Data science is a field in which people collaborate, connecting ideas to solve difficult problems with data. Ideal collaborations are those in which diverse groups of students come together and work effectively with different ideas being valued and developed. Groupwork is enhanced when students are given open problems in an environment where different students feel safe to share their ideas, respectfully, before working to solutions. Group work is usually much more effective when students start their work in structured and unstructured conversations where each group member shares their thoughts. Collaborative classrooms founded in engaged listening and the capacity to articulate verbally as they build on each other's ideas, are places where students feel valued and where they develop Important relationship skills of communication, social engagement and teamwork.

Data for all: living in an information-overloaded world

Because decisions and predictions are often based on data, all California high school graduates need data acumen, as described in the chapter introduction: skills in interpreting and visualizing data, making and critiquing data-based arguments, and some facility with data software. The ability to identify types of questions that are subject to exploration through data is crucial, as is an understanding of some misuses of data and of one's own online data footprint. As in earlier grades, it is crucial that students have opportunities to generate and investigate their own "I wonder" questions in given contexts. All statistics standards are identified as *modeling* standards, reflecting the origin of all work with data in authentic questions about the world.

"I wonder" is just the beginning, of course. Students must learn to formulate statistical investigative questions, pose data collection questions, interrogate existing data, analyze data, and formulate, interpret, and communicate findings. Note that questioning is a central practice throughout the statistical problem-solving process. As the GAISE II report summarizes:

> The statistical problem-solving process typically starts with a statistical investigative question, followed by a study designed to collect data that aligns with answering the question. Analysis of the data is also guided by questioning. Constant questioning and interrogation of the data throughout the statistical problem-solving process can lead to the posing of new statistical investigative questions.
>
> Often when considering secondary data, the data need to first be interrogated—how were measurements made, what type of data were selected, what is the meaning of the data, and what was the study design to collect the data. Once a better understanding of the data has been gained, then one can judge whether the data set is appropriate for exploring the original statistical investigative question or one can pose statistical investigative questions that can be explored with the secondary data set. (Bargagliotti et al., 2020)

The process of *interrogating* secondary data (data collected by anyone other than the person doing the analyst) is crucial. Public datasets are not collected specifically to answer students' questions. So before using such datasets, students will need to evaluate the appropriateness for their purposes: How do the measures, methods, and scope of the dataset match the statistical investigative question(s) that interest us? For what purpose(s) were the data collected?

Using data to answer authentic questions about the world is a powerful antidote to the famous student retort "when will I ever need to know this?" The Mathematics: Investigating and Connecting high school pathway in Chapter 8: Grades 9–12, encourages the use of data science contexts to frame many of students' explorations to develop content and practice standards in all domains. In this section, we propose an outline for data science understanding, considering the understandings high school students should develop.

Students enter high school with significant relevant experiences which should be drawn upon in high school work. As the CA CCSSM introduction to High School Probability and Statistics summarizes work in prior grades:

> Data are gathered, displayed, summarized, examined, and interpreted to discover patterns and deviations from patterns. Quantitative data can be described in terms of key characteristics: measures of shape, center, and spread [variability]. The shape of a data distribution might be described as symmetric, skewed, flat, or bell shaped, and it might be summarized by a statistic measuring center (such as mean or median) and a statistic measuring spread (such as standard deviation or interquartile range). Different distributions can be compared numerically using these statistics or compared visually using plots. Knowledge of center and spread are not enough to describe a distribution. Which statistics to compare, which plots to use, and what the results of a comparison might mean, depend on the question to be investigated and the real-life actions to be taken.

The big ideas of data science for all students in high school are identified in the statistics cluster headings in the standards, with an additional big idea discussed here, in response to the changing approaches to data described above. The first two are described in more detail, with additional examples, in the Draft High School Progression on Statistics and Probability (https://www.math.arizona.edu/~ime/progressions/).

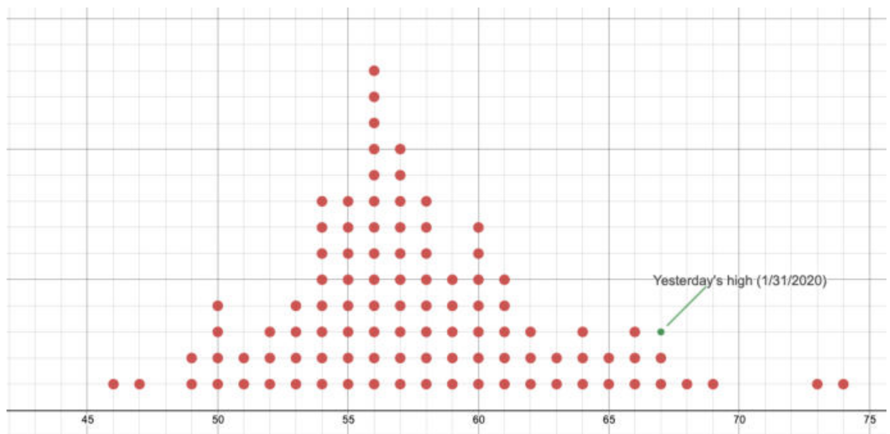
- Interpreting Categorical and Quantitative Data
- Making Inferences and Justifying Conclusions
- From statistics to data science

Interpreting Categorical and Quantitative Data

High school students continue their work with the representations of data introduced in K–8, and they are introduced to the normal curve as an approximating model for some data sets. However, the major work in high school in interpreting data is using functions as models of associations in two-variable quantitative data.

Building on K–8 experiences, high school students continue to visualize and represent *single-variable* data with dot plots, histograms, and box plots; use measures of center and spread to describe such distributions; and compare distributions from different populations or samples using these representations and statistics (S-ID.1–3). For data sets that appear to be bell-shaped, they use the mean and standard deviation to specify an approximating normal distribution and to approximate population percentages in specified ranges (S-ID.4). Students might, for example, obtain high temperatures on a specific date over the past 100 years from a nearby weather station (https://calclim.dri.edu/pages/stationmap.html), create a dot plot, visually check for a bell-shaped distribution, and use an approximating normal distribution to make a case for whether or not the temperature was consistent with historical trends (CA Environmental Principles & Concepts 1.C).

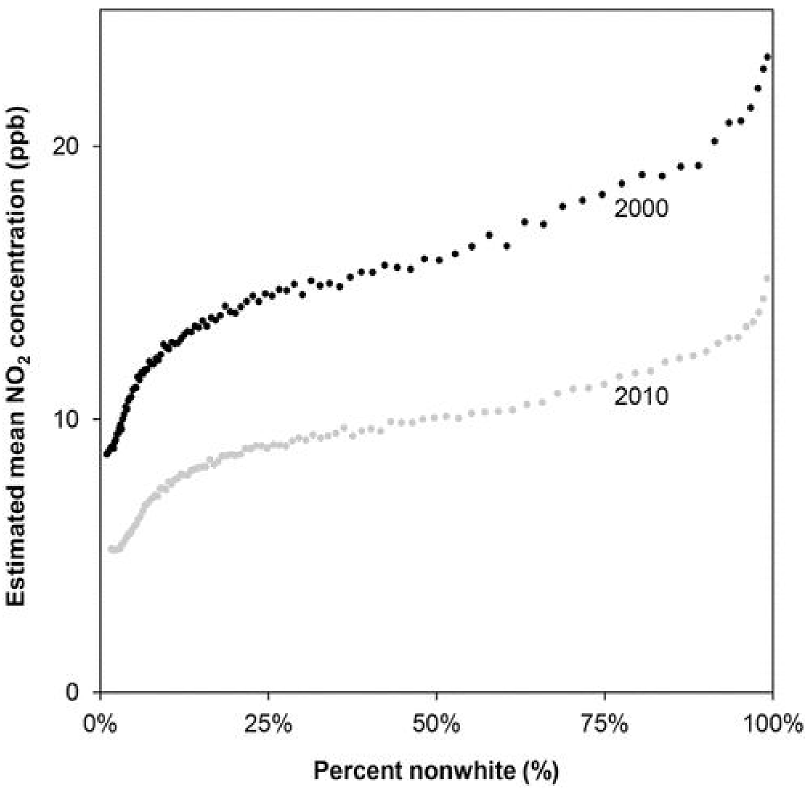
When available data includes two (or more) measurements or characteristics for each observation, students' tools for representing and interpreting relationships depend on the nature of each variable.



- If both are categorical, two-way frequency tables give an important summary that reveals relationships when interpreted in the context of the data
- If one is categorical and the other quantitative, students can treat each category as a separate population and compare the quantitative data for the different categories as in the single-variable paragraph above
- If both variables are quantitative, the scatter plot is the standard visual representation

Modeling association in numerical data using functions: Once a scatter plot is created, an association between the two variables may become visually identifiable. Fitting a function to the data is the creation of a mathematical model for the association. This begins in eighth grade with visual fitting of a linear model. While the type of function that is used most frequently is a line (a linear function), students also need experiences with plotting associations that are clearly non-linear, as well as experiment with fitting other types of functions (quadratic, exponential).

Any standard data software (including spreadsheets, Desmos, Geogebra, CODAP) will fit lines, quadratic functions, and exponential functions to given data. The specific standard technique for identifying a line (or quadratic or exponential function) of best fit (least-squares regression) is *not* an expectation; but students should have experiences fitting lines and some other functions visually (by adjusting parameters on appropriate function types in graphing software) and using appropriate software tools which perform the regression behind the scenes.



Most importantly, functions which model associations must be used to solve problems (e.g. prediction of one variable given another) (S-ID.6.a) and must be interpreted in the context of the data (S-ID.7).

Important examples of modeling association in numerical data arise in many contexts in science, history, physical education, and social studies. The California History–Social Science Framework (in its Appendix C) expects students to develop *Chronological and Spatial Thinking*, including analyzing change over time; both *time* and *space* provide opportunities for finding meaningful quantities that vary together. For example, students might wonder whether pollution exposure is related to wealth, and either find zip-code level data on both air pollution and income, or find existing research like the graph here, and work to understand and explain it. (This graph has the added benefit of representing both change over time and change in space.)

Source: https://ehp.niehs.nih.gov/doi/10.1289/EHP959.

Making Inferences and Justifying Conclusions

Making conclusions and generalizations about a population from a sample (S-IC.1) is the goal of *inferential* statistics, as opposed to *descriptive* statistics. Students work with random samples beginning in seventh grade, their first experience trying to understand a population without gathering data about all of its members. This strand of high school data work is the foundation for most meaningful use of statistics for making decisions.

Students must decide whether a result observed through data is consistent with a mathematical model of the process that generates the data (S-IC.2). For instance, if a student estimates that 30 percent of the students at the school grow food at home, then that is a mathematical model that gives them an idea of what proportion to expect in a sample. If they then survey 5 randomly-chosen students, and all say they grow food at home, then the student should be able to reason as follows: *If* 30 percent of students grow food at home, then the chances of five randomly-chosen students all being among those 30 percent of students is $(.3)^5 = .00243 = .342$ percent, or less than half of one percent. Thus, the student might doubt—that is, they might *reject*—the 30-percent hypothesis. Students should have many experiences of simple situations like this to understand how decisions based on data rely on probability, and are not *guaranteed* to produce correct answers to the original question.

Students should work with data originating in four different methods of data production, including at least some student-generated questions and student-gathered data. These methods are (1) *census* data; that is, data that contain measurements on every member of the target population (such as the database of crimes occurring in a given city in a given time frame, or rain gauge data for a given location, which captures all precipitation at that location); (2) surveys administered to random samples (to estimate population values, or *parameters*, for the surveyed quantities), (3) randomized experiments (to compare treatments and demonstrate cause), and (4) observational studies (to study characteristics or quantities when random selection or assignment is not possible) (S-IC.3). The Draft High School Progression on Statistics and Probability (https://www.math.arizona.edu/~ime/progressions/) contains detailed examples describing the treatment of each that is expected in the standards.

For surveys and experiments, the key understanding is the link between the random selection or assignment and the ability to reason probabilistically to make claims. With a survey, the random sampling allows generalizing to a population. With an experiment, the random assignment allows causal conclusions but not generalization to a broader population—unless the sample in the experiment was randomly selected from some larger population. For example, medical studies (experiments) must use willing volunteers and thus are not random samples of the overall population; this makes it much harder to draw broadly-applicable conclusions.

In a college statistics course or the data science course outlined below, students will learn ways to quantify the comparisons between gathered data and hypothesized population parameters (margins of error and *p*-values). Making sense of these, however, requires an understanding of the role of randomization in the data gathering.

When using a sample mean or proportion to estimate a population mean or proportion, students use simulation models to estimate a margin of error, instead of formulaic calculations. Briefly, the process is to use data simulation software to draw many random samples from a hypothetical population, and to see how often a result is obtained that is as extreme as the sample mean or proportion. Doing this process for hypothetical populations with many different mean or proportion parameters helps students see that there is a range of population parameters that often (more than 5% of the time) produce simulated sample means or proportions that are as extreme as (or more extreme than) the actual sample mean or proportion. This range of population parameters is the (simulation-based) confidence interval, given as (sample mean or proportion ± margin of error). Note the probabilistic argument here: *If* the population mean or proportion were outside of the confidence interval, *then* sample means or proportions as extreme as we obtained in our random sample would be rare. So, we expect that the true population mean or proportion is within the confidence interval. (*But cannot be certain* that it is!)

A similar process is used to evaluate confidence in a randomized experiment, in which subjects are randomly assigned to two or more treatment groups. (Treatment could mean medical treatment, or assignment of different tasks, or being shown different motivational videos, etc.) Some quantity is then measured for each subject, and the investigator then has to decide from the results whether a treatment, say treatment A, produced any effect on the measured quantity. Simply having a different mean for each treatment group is not enough, as we expect variation in the measurement and thus between groups. In this case, all of the treatment groups are pooled into a population and then re-sampled (randomly) many times, to see how often the re-sampled mean or proportion is at least as extreme as the actual treatment A group difference. If such differences are rare, the experiment is taken as evidence that treatment A caused a change in the measured quantity.

From Statistics to Data Science

For questions about community, society, or natural systems beyond students' immediate experience—and thus beyond their ability to gather data directly—existing data can often be identified from online sources. Since students frequently encounter claims made from such large data sets, it is crucial that all students have experiences in which they explore the ways in which such claims are made. A major difference between the classical statistical approach begun in K–8 and the "big data" of the growing field of data science is the richness and complexity of available data sets, even more so than their sheer size.

Many sophisticated approaches to working with rich, complex data sets are left to a data science course in the data science pathway; but *all* high school students should exercise and refine their understanding of data exploration, causal inference, and statistical reasoning using large, real world data sets. As students work with these data sets they can draw upon the data science understandings they have developed in their K–8 mathematics lessons. Emphasis should be on questioning and interpreting, rather than technical procedures.

Data exploration begins with a search for available data about a context of interest. The data set is then examined for hidden patterns and associations (usually via visual representations). Any patterns or associations discovered can lead to new hypotheses or questions to investigate further. Students began this process in eighth grade, and continue in high school with experiences in which they examine data sets with multiple variables measured for each member of the sample. They plot pairs of variables to decide which ones might show associations. Important discussions for students to engage in when working with existing data sets include

- Prior to exploring: Do we *expect* any of these variables to be associated? Why?
- Might the association we see just be a result of the way in which the data was collected, rather than truly reflective of the population? What features of the data collection might make conclusions suspect, and what features might give confidence? Note that a large sample size is not enough to have confidence in conclusions.
- Can we think of possible explanations for the association(s) we see? Can we think of ways we could decide which explanations might be accurate?

After data exploration identifies some association(s) of interest, the stage of model building follows. Technical methods are reserved for the specialized data science course below, but all students need to explore questions such as:

- Could we use some variables to predict others? This is a hugely important use of data, since some factors are easier to measure or observe than others. In medicine and many other fields, this often takes the form of trying to predict future outcomes using presently-available information.

- If we could only know measure one variable to try to predict a variable of interest, which one would we pick? Why? What if we could measure two? Which second variable gives us the most *new* information for prediction?

Most importantly, high school students (like K–8 students) must experience data science as a set of tools for making sense of their worlds in ways that matter to them.

Advanced high school data science

The traditional sequence of high school courses—algebra, geometry, algebra 2—was standardized in the United States following the “Committee of Ten” reports in the 1890s. The course sequence—which was primarily designed to give students a foundation for calculus—has seen little change since the Space Race in the 1960s. With the rapid expansion of information available to all in the form of data, far more students pursue statistics classes than calculus, and may be better served by a data science course as a culminating high school mathematical science experience. In addition to the importance of the data science content—to 21st Century jobs and to a wide range of college majors—many students are more engaged by open-ended explorations of important data sets, drawing upon important mathematical principles and tools, than by many traditional courses organized around mathematical techniques. This Framework provides design principles and content outcomes for such a course.

California high schools offer upper-level data science courses in two ways. In the first model, students have a common experience in grades nine and ten, with pathways branching at grade elevent. Some districts have designed and are offering eleventh grade data science courses as an option for this third year of high school mathematics; in this case, the ninth and tenth grade courses need to be designed to include the important high school geometry standards. The second model is a data science course as a fourth-year course, following a coherent three-year pathway that builds the “for all students” data science understanding outlined in the previous section. The design principles and content outcomes below are flexible enough to be implemented in either model, with appropriate adjustments for students' prior experiences.

Design Principles

These principles provide guidelines for design of curricular materials and classroom instruction for a data science course, in order to support a coherent and engaging experience for students. These principles should be used by developers to build curricular materials that are true to the vision of the course, as well as by educators reviewing materials and developing a repertoire of pedagogical strategies for use in teaching the course. Many students and teachers already engage in these behaviors; in these cases, these design principles will be seen as reinforcing and supportive. The spirit of this framework recognizes that, at some levels, everyone is a learner, and everyone is growing an understanding of mathematics, each other, and the world we share.

Design Principle	**Students will**	**Teachers will**

Active Learning. The course provides regular opportunities for students to actively engage in data explorations using a variety of different instructional strategies (e.g., hands-on and technology-based activities, small group collaborative work, facilitated student discourse, interactive lectures).	• Be active and engaged participants in discussion, in working on data explorations with classmates, and in making decisions about the direction of instruction based on their work. • Actively support one another's learning. • Discuss results of their explorations with the instructor and/or classmates in class.	• Provide low-floor, high-ceiling activities and explorations that all students can access and that extend to high levels. Such activities should provide meaningful opportunities for exploration and co-creation of mathematical understanding and data literacy. • Provide interesting and sometimes local data sets and invite students to ask questions of the data. Encourage different students to pose and investigate different questions, and to come together to discuss findings. • Facilitate students' active learning of data science through a variety of instructional strategies, including inquiry, problem solving, critical thinking, and reflection. • Create a safe, student-driven classroom environment in which all students feel a sense of belonging to the class and the discipline, are encouraged to take risks and embrace mistakes, and are able to make decisions about the direction for instruction through the

		results of their exploration of data science. Students' ideas are at the center of the conversation.

Growth Mindset. Courses support students in developing the tenacity, persistence, and perseverance necessary for learning data science, for using mathematics and statistics to tackle authentic problems, and for being successful in post-high school endeavors.	● Make sense of data explorations by drawing on and making connections with their prior understanding and ideas. ● Persevere in solving problems and realize that it is acceptable to say, “I don’t know what to do next,” but that it is not acceptable to give up ● Seek help from different sources to move forward in their investigations. ● Compassionately help one another by sharing strategies and solution paths rather than simply giving answers. ● Reflect on mistakes and misconceptions to improve their mathematical understanding and data literacy. ● Understand that struggle is valuable for brain growth and times of struggle should be valued. ● Develop/strengthen a growth mindset to continue to apply in mathematics, data science, and other areas of their post-high-school life.	● Provide information about and model the importance of having a growth mindset. ● Value mistakes and times of struggle. ● Facilitate discussions on the value of mistakes, misconceptions, and struggles. ● Give students time to struggle with tasks and ask questions that scaffold students’ thinking without stepping in to do the work for them. ● Praise students for their efforts in making sense of mathematical ideas and for their perseverance in reasoning through problems and in overcoming setbacks and challenges in the course. ● Provide students with low-stakes opportunities to fail and learn from failure. ● Provide regular opportunities for students to self-monitor, evaluate, and reflect on their learning, both individually and with their peers.

Problem Solving. Courses provide opportunities for students to make sense of problems and persist in solving them.	• Apply intuition, life experience, and previously learned strategies to solve unfamiliar problems. • Explore and use multiple solution methods. • Share and discuss different solution pathways and methods. • Be willing to make and learn from mistakes in the problem-solving process. • Use tools and representations, as needed, to support their thinking and problem solving. • Develop and justify their own strategies to approach new problems.	• Present tasks that require students to find or develop a solution method. • Provide data sets that allow for multiple strategies and solution methods, including transfer of previously developed skills and strategies to new contexts. • Provide opportunities to share and discuss different solution methods. • Model the problem-solving process using various strategies. • Encourage and support students to explore and use a variety of approaches and strategies to make sense of and solve problems.
Authenticity. Courses present data science as a subject and learning that allows us to model and solve problems that arise in the community.	• Recognize specific ways in which mathematics and data are used in everyday decision making. • Recognize problems that arise in the real world that can be solved with data science • Contribute meaningful questions that can be answered using data science. • Experience the decision making involved in collecting, cleaning,	• Provide opportunities to ask questions of data sets that are relevant to students, both in class and on assessments • Provide opportunities for students to pose questions that can be answered using data science methods and tools, and answer them. • Provide students with real data to explore and work with, including doing some of the data cleaning that is often required.

	analyzing, and visualizing data.	
Context and Interdisciplinary Connections. Courses present data science in context and connects data science to various disciplines and everyday experiences.	• Contribute personal experiences, where appropriate, that connect to classroom experiences. • Actively seek connections between classroom experiences and the world outside of class. • Examine the ways that data is collected in their day-to-day lives, and consider the ethics and consequences of collecting and using data to make decisions.	• Provide opportunities for students to share their personal backgrounds and interests, including cultural values, and help make the connection between what is important in students' lives and future aspirations, and what they are learning in data science. • Provide real and interesting data sets, including some that are local to students • Invite students into data explorations that illustrate authentic applications. • Provide data explorations that include applications from a variety of academic disciplines,

		programs of study, and careers, and which are culturally sustaining.

Communication. The course develops students' ability to communicate their data explorations and findings in varied ways including with words, data visualizations and numbers.	● Present and explain ideas, reasoning, and representations to one another in pair, small-group, and whole-class discourse using discipline-specific terminology, language constructs, and symbols. ● Seek to understand the approaches used by peers by asking clarifying questions, trying out others' strategies, and describing the approaches used by others. ● Listen carefully to and critique the reasoning of peers using data to support or counterexamples to refute arguments. ● Develop the skills to justify mathematical reasoning with clarity and precision. ● Practice constructing data-based arguments with specific audiences in mind. ● Consider matters of accessibility in designing and executing their communications. ● Consider the pros and cons of various types of data visualizations and how they fit the communicative situation.	● Introduce concepts in a way that connects students' experiences to course content and that bridges from informal contextual descriptions to formal definitions. ● Clarify the use of data science terminology and symbols, especially those used in different contexts or different disciplines. ● Engage students in purposeful sharing of ideas in data science, reasoning, and approaches using varied representations. ● Support students in developing active listening skills and in asking clarifying questions to their peers in a respectful manner that deepen understanding. ● Facilitate discourse by positioning students as authors of ideas who explain and defend their approaches. ● Provide regular opportunities for students to communicate about data science with a variety of data visualizations ● Scaffold instruction to support students in developing the required

		reading and writing skills.

Technology. Courses introduce students to current data science technologies and prepare them to learn and use new ones.	• Use technology to visualize and understand important data science concepts and as a tool in problem solving. • Understand the necessity of digital tools in cleaning and analyzing large data sets and are able to select appropriate tools for different situations. • Develop experience in learning new tools which allows them to try out emerging data science tools in the future. • Understand that the use of tools or technology does not replace the need for an understanding of reasonableness of results or how the results apply to a given context.	• Introduce students to various digital data science tools and support them in understanding the best uses for each. • Facilitate student learning of technological platforms through exploration, as this will aide in transferring the knowledge to future platforms. • Not be experts in the use of every platform but willing to experiment along with students' questions and model good practices for seeking answers to such questions.
Assessment. Courses use project-based assessments to evaluate student progress.	• Assemble a collection of their work which includes both their mathematical work and reflections on their learning process and their evolving understanding of the field of data science. • At the end of the course, have a portfolio of data science work that showcases their knowledge of data science as well as their	• Provide students with projects through which they are exposed to new content and demonstrate their ability to use this new content to solve problems. These will include products that demonstrate student learning both for the teacher, and to be included in the students' portfolios.

	software skills. This portfolio might be shared with a potential employer or educational institution.	● Evaluate student progress throughout the course by considering students' evolving portfolios as well as their reflections on their learning. ● In the final project of the course, allow students freedom to decide the topic and methods of their data exploration, so that they can bring together the various skills they will have developed over the course, and allow the teacher to assess their progress.

Content Learning Outcomes

In this section we present the mathematical content outcomes expected from a high school data science course. These will be motivated by realistic examples and projects which will help students develop their basic data science skills as well as a larger understanding of their contexts and of the importance of data in their lives.

Understanding the Role of Data in The World

Students demonstrate an understanding of what qualifies as data and the different types of data that exist. They also understand how data is generated and collected, and the existence of extremely large amounts of data created by our digital lives. Students consider their own privacy and data footprint. Throughout the course, students discuss the ethics and consequences of collecting and using big data, and the ways data is collected, including the bias that may be present in the data collection or selection process. Students evaluate and critique data-based claims and arguments, in particular, they distinguish correlation and causation. Students understand that all data and data-based arguments have several sources of bias and are able to identify them. They understand the importance of communicating with data and making data-based arguments. They use multiple different types of data visuals both for analysis and in order to share their thinking with others.

- Represent data represented by real numbers using dot, box and histograms (S&P)
- Summarize categorical data for two categories in two-way frequency tables (S&P)
- Interpret relative frequencies in the context of the data (S&P)
- Recognize possible associations and trends in the data (S&P)
- Distinguish between correlation and causation (S&P)
- Evaluate the purpose of and differences between sample surveys, experiments and observational studies and how randomization effects each (S&P)

Asking Data-Based Questions

Students are able to identify the types of questions that are subject to exploration through data as well as formulate their own. They are able to perform exploratory data analyses to draw preliminary conclusions to explore further. They can do this in a variety of platforms. Students can look at the data available and identify questions that it can answer as well as determine what data might be collected in order to answer a question. Students consider how they might use some of the data they have access to, in order to predict other variables for which it might be harder to collect data directly.

Unraveling the Story That Data Is Telling

- When working with numerical data students can describe a distribution using its shape, center, and spread. They are able to make predictions based on these characteristics, as well as compare distributions to one another. Students are also able to compare two numerical variables to each other using scatter plots and can use their understanding of functions (linear, polynomial, exponential) to fit their data to a curve (using appropriate technological tools) and use this model to make predictions. Students are also able to work with categorical variables in frequency tables as well as use numerical and categorical variables together in order to answer questions about the data. Analyze the shape of data distributions and compare data distribution using measures of center (mean, median) and spread (interquartile range (IQR), standard deviation) of different data sets (S&P)

- Interpret differences in shape, center and spread including the effects of outliers (S&P)
- Use mean and standard deviation to fit to a normal distribution and to estimate population percentages. Know that this procedure is not appropriate for all data sets (S&P)
- Use tech tools (calculators, spreadsheets and tables) to estimate areas under the normal curve (S&P)
- Represent two variable data on a scatter plot and describe how the variables are related (S&P)
- Fit a linear function on scatter plots where the data suggests a linear fit (S&P)
- Fit a function to the data to solve problems in context of the data (S&P)
- Determine the fit of a function by plotting and analyzing residuals (S&P)
- In a linear model interpret slope as a rate of change and the intercept as the constant term in the context of the data (S&P)
- Use technology to compute and interpret the correlation coefficient of a linear fit
- Estimate a line of best fit for a single linear regressions (algebra)
- Determine and interpret the strength of correlation to determine the best fit. (algebra)
- Use multiple linear regressions and non-linear regressions (linear, quadratic, exponential) algebra
- Regression trees and classification trees (Talitha course has a lesson on these. “Classification trees and regression trees are easily understandable and transparent methods for predicting or classifying new records”)
- Understand independent and dependent events and the and that two independent events have a probability of occurring together that is a product of their individual probability of occurring (S&P)

- Conditional probability (S&P)
- Construct and interpret two-way frequency tables of data when two categories are associated with each object being classified. Use the two-way table as a sample space to decide if events are independent and to approximate conditional probabilities. (S&P)
- Recognize and explain the concepts of conditional probability and independence in everyday language and situations. (S&P)
- Calculate expected values and use them to solve problems. (S&P)
    - Calculate the expected value of a random variable and interpret it as the mean of the probability distribution (S&P)
- Use probability to evaluate outcomes of decisions (S&P)
- Linear algebra:
- Recognize situations in which one quantity changes at a constant rate per unit interval relative to another (algebra)
- Recognize situations in which a quantity grows or decays by a constant percent rate per unit interval relative to another (algebra)

Grappling with Variability and Uncertainty

Students understand variability is inherent to data and are able to identify multiple sources of it. They practice collecting and organizing data about their own lives and communities as well as working with large, real-world, publicly available data sets. Students consider sampling practices and how they affect the data that is collected. They can use probability to make decisions and understand the uncertainty that comes along with predictions.

- Know that statistics is a process for making inferences about population parameters based on random samples of the population (S&P)
- Determine if a model from a data generating process or simulation is accurate (S&P)
- Make inferences and justify conclusions from sample surveys, experiments and observational studies (S&P)

- Use data from a sample to estimate population mean or proportion and develop a margin of error through simulation models (S&P)
- Use simulations to decide if differences between parameters are significant (S&P)
- Evaluate reports based on data (S&P)

Transforming Data with Technology

Students understand that data is not always collected/shared/received ready to be analyzed and it sometimes requires work to prepare it. They can use different digital tools to clean and transform the data (e.g. merge data, deal with incomplete data, normalize data). They are familiar with the basics of programming as needed, and are comfortable editing code or finding the appropriate tools to transform the data in ways useful to their own data analysis. Students can combine their knowledge of probability and programming to construct simulations of probabilistic events, and they understand the basic idea behind machine learning as well as its power and shortcomings.

- Perform operations on matrices and use matrices in applications (Number)
- Use matrices to represent and manipulate data (Number)
- Cleaning names, categories, and strings (IDS)
- Simulation using experimental data (IDS)
- Translate between different bit representations of real-world phenomena, such as characters, numbers, and images (CS Stds)
- Evaluate the tradeoffs in how data elements are organized and where data is stored.(CS Stds)
- Create clearly named variables that represent different data types and perform operations on their values. (CS Stds)
- Collect data using computational tools and transform the data to make it more useful and reliable. (CS Stds)

Sample Courses

Effective Data Science courses consider how to help students:

- understand how data are used by professionals to address real-world problems;
- understand that data are used in all facets of modern life;
- understand how data support science to identify and tackle real-world problems in our communities;
- analyze statistical graphics to identify patterns in data and to connect these patterns back to the real world;
- understand that by treating photos, words, numbers, and sounds as data, we can gain insight into the real world;
- learn to analyze data, including: posing questions that can be answered by considering relations among variables in a data set, using collected data to generate hypotheses for future data collection, critically evaluating shortcomings and strengths in the data and the data collection process, and informally evaluating hypotheses using data at hand.

A sample table of contents for the course is given in Appendix A.

Another sample course begins with a consideration of the meaning of data, the importance of communicating data visually, investigating community issues, cleaning data, exploratory data analysis, ethical issues around data, creating data dashboards, linear and nonlinear regression models, statistics, probability, and forecasting. The course is designed to engage students actively and to be flexible enough for teachers to include local issues of importance to their communities. While addressing concepts of Data Analysis with rigor, the access and dependence upon current, local, and publicly accessible data is a key feature. One goal of the course is that it be open to all students, regardless of prior mathematics achievement, all lessons will be “low floor and high ceiling” – designed so that everyone can access them and they extend to high levels.

Some schools have created a Data Science elective for students in grades 10–12. The course may begin with the basics of data collection, and then teach distributions, linear regression, probability, and statistical inference through investigation-based activities.

Course activities may include making distributions of students texting-frequency, examining player statistics from 30 Major League Baseball teams, and analyzing the link between poverty and obesity. Districts can design their course to meet A–G Mathematics credit requirements.

An additional example of school-created course for students in grade nine is one focused on software design and data science. It teaches algebraic, geometric, and statistical concepts through contexts like video-game design. This course can be an example of a modernized integrated pathway, teaching the traditional sequence through modern mediums and applications. The course can also be designed to meet A–G elective credit requirements.

The different examples of courses and high school approaches above use different software and tools, which seems appropriate as data science does not require any particular software package, it is more important that students learn to ask good questions and apply an effective tool to help them answer them. Exposure to some software is essential for those wishing to pursue a full-time career in data science, and comfort with such programs is increasingly valuable for many other professions that involve basic data analysis.

In total, over 70 individual high schools and 15 districts offered a data science mathematics or elective course in California during the 2019–2020 school year that counted for A–G credit (University of California data). That compares to just 34 high schools and 6 districts two years before in 2017–2018. This rapid increase in course offerings is likely an indication of both high interest in and importance of data science content throughout the curriculum.

High School Tools and Resources

One sample tool for students to explore large data sets is the free, open source software tool Common Online Data Analysis Platform (CODAP) (https://learn.concord.org/dynamic-data-science) from the Concord Consortium. Using this software, students can import data from their own community or work with the large data sets already available in the tool. Students will learn to become active citizens in their communities, learning that mathematics is an important tool for benefitting their community.

The Census at School project (https://ww2.amstat.org/censusatschool/) is an international classroom project that engages students in grades 4–12 in statistical problem solving. Students complete a brief online survey, analyze their class census results, and compare their class with random samples of students in the United States and other countries.

Other software such as Fathom (https://fathom.concord.org/) and Statkey (http://www.lock5stat.com/StatKey/) allow exploration and organization of data sets and the development of simulations. Google offers a free coding software called Google Script Coding, Python is another tool that can be used to explore and analyze data sets. A more sophisticated data software tool is R. This requires learning time and schools may need to provide server space to run the software.

Below are two examples of data science projects that students may work on in high school, freely available from the Concord Consortium:

In the California American Community Survey (ACS) Data Portal (https://learn.concord.org/dynamic-data-science) students are given access to the data portal which gives census data for California residents from the U.S. Census Bureau's American Community Survey. The database contains demographic information about California residents (e.g., marital status, sex, place of birth, employment status, and health information). Data challenges are given such as finding out the average income of Californians of different age groups in 2013, or students can choose to investigate their own questions. For example, they may choose to look at salaries by gender, or make a data visualization to show the different ethnic groups that live in California. Standards addressed include making inferences and justifying conclusions (HSS-IC.A.1) and SMP.2 (Reason abstractly and quantitatively), SMP.3 (Construct viable arguments and critique the reasoning of others), SMP.4 (Model with mathematics), and SMP.5 (Use appropriate tools).

In a different Concord activity: Making Trees in a Diagnosis Game (https://learn.concord.org/resources/1241/trees-in-a-diagnosis-game) students use data to build binary trees for decision-making and prediction. Prediction trees are the first step towards linear regression, which plays an important role in machine learning for future data scientists. Students begin by manually putting "training data" through an algorithm. They then learn to automate the process and to test their ability to predict which alien creatures are sick and which are healthy. This activity touches upon many content and practice standards, including Making inferences and justifying conclusions (HSS.-IC.A.1), Using probability to make decisions (HSS.-MD.B.7), and all standards for mathematical practice.

Conclusion

Changing demands for life in a data-rich world require that California schools prepare all students to examine claims justified with data, to understand the probabilistic underpinning of drawing conclusions from samples, and to see data as a tool to answer many questions of interest. Developing these abilities requires that students generate questions and work with data beginning in kindergarten (or before), and have experiences of increasing depth and complexity throughout their school careers. Students who wish to focus extra attention on data science should have an opportunity to pursue advanced courses late in their high school careers.

Above all, students at all levels should have experiences that build their mathematical toolkit for making sense of their worlds.

Free Resources for the Teaching of Data Science

- Concord Consortium: https://learn.concord.org/dynamic-data-science
- Jo Boaler Online Course: The teaching of data science K–12: https://www.youcubed.org/21st-century-teaching-and-learning/
- The Messy Data Coalition: https://messydata.org/
- University of Chicago RISC: https://www.21cmath.org/
- Women in Data Science Video: https://www.youcubed.org/resources/what-is-data-science/
- Wolfram-Alpha: http://www.computerbasedmath.org/
- Youcubed Resources: https://www.youcubed.org/resource/data-literacy/
- Youcubed Grades 6–10 Data Lessons: https://www.youcubed.org/data-science-lessons/
- Youcubed Data Talks: https://www.youcubed.org/resource/data-talks/

References

Arnold, P. (2007). What about the P in the PPDAC cycle? An initial look at posing questions for statistical investigation. *Education*, *55*.

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 II (GAISE II): A Framework for Statistics and Data Science Education*. American Statistical Association.

Boaler (2019). *Limitless Mind. Learn, Lead and Live without Barriers*. Harper Collins.

Boaler, J., Cordero, M., & Dieckmann, J. (2019). Pursuing Gender Equity in Mathematics Competitions. A Case of Mathematical Freedom. Mathematics Association of America, FOCUS, Feb/March 2019. http://digitaleditions.walsworthprintgroup.com/publication/?m=7656&l=1#{%22issue_id%22:566588,%22page%22:18

Carmichael, I., Marron, J.S. Data science vs. statistics: two cultures?. *Jpn J Stat Data Sci* **1,** 117–138 (2018). https://doi.org/10.1007/s42081-018-0009-3

Chestnut, E. K., Lei, R. F., Leslie, S. J., & Cimpian, A. (2018). The myth that only brilliant people are good at math and its implications for diversity. *Education sciences*, *8*(2), 65.

CORE SEL Competencies: https://casel.org/core-competencies/

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64-73. https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational researcher*, *41*(3), 93-97.

Pelesko, John (2015). "'The' Modeling Cycle." http://modelwithmathematics.com/2015/08/the-modeling-cycle/

Rubin, Andee. "Learning to Reason with Data: How Did We Get Here and What Do We Know?." *Journal of the Learning Sciences* (2019): 1–11

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. Journal of Educational Psychology, 107(2), 468.