Extended Abstract for the
Symposium on Integrating AI and Data Science into School Education Across Disciplines 2025

HIGH-SCHOOL DATA SCIENCE: A "DATA MOVES" PERSPECTIVE

Tim Erickson
Epistemological Engineering, USA, eepsmedia@gmail.com

Focus Topics: Learning materials; tools

In a traditional statistics course, we help students learn to infer properties of a population from relatively small samples. To that end, we create datasets appropriate to those tasks and to the learners. The datasets usually have at most a few dozen cases, and only the variables they need: a measurement or two, and perhaps a grouping variable.

Data science, in contrast, usually involves much larger datasets—more cases, more variables—and tasks that do not fall neatly into some inferential category (Erickson 2017). But can high school students actually do data science? Professional-level tasks might be too hard, just as our statistics students cannot do professional statistics. But we can create tasks and environments for school-age students that "smell" like data science, and that will be good preparation for later learning. Furthermore, these tasks are often more interesting and immediate for students than the ones we carefully crafted to illustrate how to perform some statistical test.

One aspect of smelling like data science, I believe, is that the task requires "data moves." And that's the topic we will explore here.

For examples in this short paper, I will use CODAP (Concord Consortium 2024), a free, web-based, open-source, dynamic data analysis platform. We will access United States Census microdata via IPUMS (Ruggles et al., 2025). These are anonymized data about individuals, and include variables such as income, gender, race, age, education, and even the industries and specific occupations of those individuals. There are millions of cases available, but we can see interesting patterns in the data with just a few hundred.

An initial task and a first data move

One way to begin a task that smells like data science is to ask students to explore the data by making and interpreting graphs. If a student graphs income, for example, just to get an idea of the distribution, they will see a graph like the one on the left in in Figure 1.

Figure 1: Left: initial, naïve graph showing the incomes of 1000 individuals from the 2020 Census. Right: same data, rescaled, with the high values (the children) filtered out.

We can see that the "naïve" mean is almost 2 million dollars. That's ridiculous. When we explore the stack of points on the right, however, we discover that all of those have an income of exactly 9,999,999—and they are all children. That is, the Census has decided not to collect data on the incomes of children, and to use a huge value as an indicator of the missing data.

If we want to draw valid conclusions about income, we need to deal with those values somehow. In CODAP, we can select them and "set them aside"; any robust data analysis system has a way to doing this. After setting aside these children, we see the "real" distribution of incomes (right side of Figure 1): skewed, with a huge proportion of small incomes—and a few large incomes.

In S. Podworny & S. Schönbrodt (Eds.), *Towards Fostering AI and Data Science Literacy in Schools Across Disciplines. Proceedings of the 1st Symposium on Integrating AI and Data Science into School Education Across Disciplines (AIDEA1 2025)*, Salzburg, Austria.

The point is that this dataset, like many real datasets, has cases that are irrelevant to our investigation. Students need to recognize the problem, identify the cases, and know how to deal with them appropriately. Setting aside the children amounts to "filtering" the dataset. And *filtering* is an example of a data move.

Describing data moves

Data moves are a form of data wrangling. Erickson et al. (2019)—henceforth, the "data moves paper"—describes data moves as *actions that alter a dataset's contents, structure, or values*. The data moves paper proposes some principal data moves for analyzing rich, multivariate datasets. You can read much more about data moves in the paper itself. You can also find them—and practice them!—in *Awash in Data*, a dynamic e-book about using CODAP for introducing data science (Erickson 2025, henceforth simply *Awash in Data*). If you are curious about the evolution of data moves, you should also see the recent work of Hudson et al. (2024). They have made an observational study of how prospective math teachers actually employ data moves in their own data investigations, and have created a new framework to account for these observations and practices.

The table below summarizes some data moves.

data move	description
filtering	Purposely restricting or slicing the data to focus only on the cases relevant to the investigation.
grouping	Dividing a dataset into subsets. You can also think of it as repeated filtering. This is often a preliminary step before summarizing.
summarizing	Creating a measure—such as a measure of center or a proportion—that applies to the entire dataset or to a subset within it.
calculating or recoding	Creating new values for every case, often based on other values in that case. Often this means making a new variable—a new column—often but not always with a formula. Contrast this with summarizing, which creates a new value shared by a set of cases.
joining	Extending the dataset by combining it with another. This can mean adding new columns (variables or attributes) or simply adding new rows (cases).

We found that when we were doing investigations that "smelled" like data science, we often used these moves. And yet traditional curricula never teach them; instead, because of their orientation towards inference procedures, they (reasonably) provide only sanitized, pre-digested datasets with the right features for the lesson at hand. In contrast, work with "real" datasets such as the one we got from the US Census almost always requires data moves as part of preparing and using the data.

This led us to make conjectures about curriculum and pedagogy: first, that making data moves is an important skill, closely related to understanding the data and doing valid analyses. Second, that it would therefore be beneficial to keep these data moves in mind while designing a curriculum oriented towards larger, richer datasets, that is, a curriculum oriented more towards data science. We could ask: What moves will students practice as part of their work? What datasets and activities tend to evoke which data moves? We also asked a metacognitive question: is there a benefit to teaching students these moves explicitly, helping students identify the moves when they use them? And finally, what kinds of tasks would help us evaluate whether students recognize data moves and their purposes?

The class

I got a chance to explore these issues. In the first Fall of COVID—Fall 2020—I was asked to teach a statistics course at a high school in San Francisco, California. The class was virtual, 70 minutes per day, every day, for approximately 6 weeks. Because of the pandemic and thanks to a flexible

administration, I was given free rein for this course. So I used materials I had already been developing—*Awash in Data*, among others—to create an introduction to data science for these students.

It's vital to point out that it would have been impossible to use a programming language such as R or Python with these students in this situation. So we used CODAP, which is designed to let students explore exactly this kind of data. All illustrations in this paper are produced in CODAP.

That said, the moves we describe in the data moves paper map almost on-to-one with the principal verbs in the R tidyverse (Grolemund and Wickham 2017) and with corresponding functions in Python-based data analysis packages. Could a CODAP-based approach be a suitable onramp for coding?

The rest of this paper will focus on two topics arising from the course design: "dig deeper" mini-projects (described in *Awash in Data* at this link), characterized by cycles of adding nuance to a claim; and the "data moves portfolio," an end-of-term assignment in which students collected and displayed data moves they had used in the mini-projects.

Questions and Claims

To learn about dig-deeper assignments, let us first discuss investigatory questions and claims.

As professional scholars and researchers, we know the importance of a good research question. As we design curricula to help our students become more like us, we often insist that students form research questions—and they often have a lot of trouble with that. It can be hard to come up with a good question.

I have found that focusing on *claims* instead of *questions* often produces good results with beginners. It works like this: I introduce a large, multivariate data set or a data source to the students. The students then explore the data, often by making graphs. Some of these graphs show interesting patterns. A pattern in the data can suggest a claim, which I define for the students as "a statement about the world that might be either true or false, and that the data can help assess." I stress that students don't have to be "for" or "against" their claim, just that it's something that could be demonstrably true or false. As an example, Figure 2 shows the kind of graph a student might make, comparing income to education.

Fig 2: Income, grouped by education level. The vertical lines show the median for each group.

With this graph, a good research question can be complicated and fraught. A claim, in contrast, can be simple and obvious: *if you get more education, you earn more money*. The fact that it is so simple means that most students will be successful right at the beginning of the process. Of course, we want more nuance in our students' work—but that's what the "dig deeper" assignment is all about.

This instructional strategy is consistent with the idea of showing students a phenomenon and asking, "What do you notice? What do you wonder?" (Fetter 2011) What you notice is a pattern in the data, in this case, an increase in income. The key point is that students have to organize the graph correctly (e.g., ensuring that the education categories are in the right order) and notice that the medians increase. Even experienced secondary students need a lot of practice creating and reading a wide variety of graphs in order to do this fluently—and they get this practice through exploring a variety of interesting

datasets. In my experience, practically any rich, multivariate dataset yields interesting patterns. Real-life experience with the context helps, of course, but is not essential.

Finally, let's point out that putting an education variable on the vertical axis is one way to *group* the data—another data move!—and that putting the median on the graph is *summarizing*.

"Dig deeper" assignments

This genre of assignment emerged over several years of presenting brief introductions to data science, and formed the backbone of the class. It begins with a claim and its evidence. That is, the first section of this kind of mini-project consists entirely of a claim and the simplest, most obvious visualization that supports or refutes the claim. Our sample claim (more education means more money), the graph in Figure 2, and a sentence or two of commentary constitute a completely acceptable response.

But then we ask students to dig deeper. I have presented this to students as adding *nuance* to the claim. That is, the statement *more education means more money* can't possibly be the whole story. In this case, there are several approaches to take. One is to look at variability and the way the income distributions overlap. Students recognize that more education is not a guarantee of more money; in fact, some people with doctoral degrees have some of the lowest incomes, and some people who never went to college earn a lot. This sort of thinking can lead students to discover and use measures of spread, or for more experienced students to explore more advanced regression topics.

But for most students, including an additional variable is an interesting and reliable way to find nuance—and get more insight into the data and its context. That is, the multivariate nature of the data is a key to looking more deeply: if we say that education is not the whole story, what else might be a factor?

Using this same dataset and claim, let's include sex in our analysis. Now, with *three* variables, there are many possible ways to display the data. In order to prepare for those visualizations, we need, of course, data moves. Figure 3 shows two such visualizations. The one on the left shows data, grouped by education and split by sex, with median values indicated by lines as in Figure 2. The one on the right, in contrast, shows only the median incomes for the groups, with sex indicated by color.

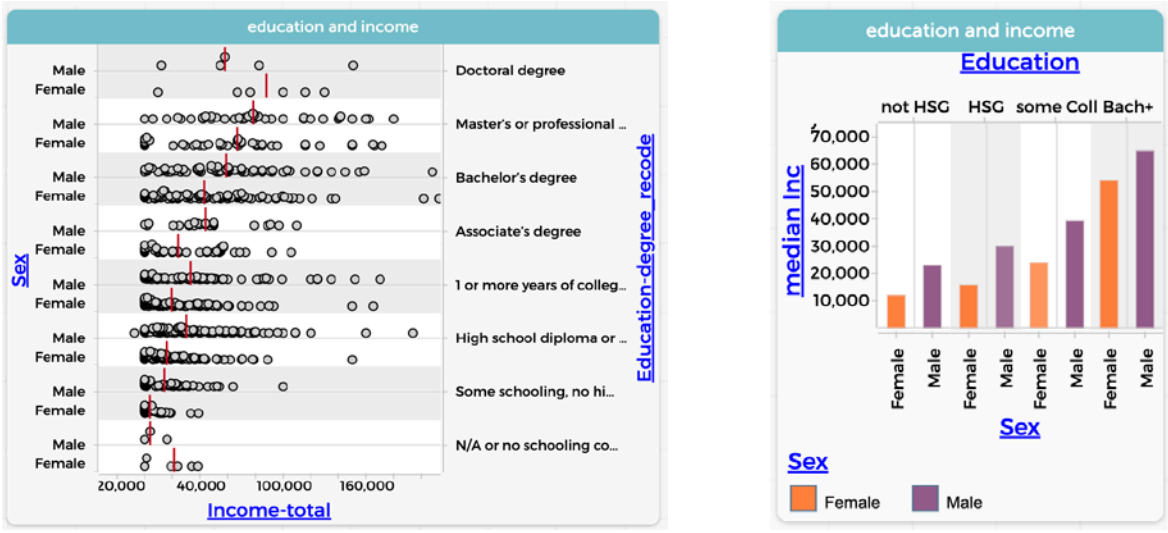

Figure 3: Two visualizations of how education and sex affect income. On the right, the student has recoded the education variable to four categories. (HSG = "high-school graduate.")

I will not address visualizations here in detail. Suffice to say that this relates to important aspects of communication in data science. In the class, we did compare visualizations and discuss which was "better," or more successful in communicating the author's point. The answer, of course, is "it depends"—an important lesson in itself.

The key is that in order to make these visualizations, the students had to use data moves—in this case, grouping and summarizing—at different levels of sophistication. The one on the right, for

example, requires that you create those median values (summaries) in such a way so that you can display them as data rather than just as "adornments" on the graph.

Either graph in Figure 3 lets the students say more about the claim. For example, we might now say that you earn more money with more education, *and* that males generally earn more than females. The one on the left shows all the data but can be hard to read. The one on the right shows no variability but is cleaner and is less work for the reader. Note that this student also reduced the number of education categories from 8 to 4, which required a recoding/calculating data move.

Writing up the more nuanced claim, and what they had to do to get it, and appropriately including and labeling any visualizations, is the meat of the assignment. Students need practice doing things like this, so the fact that the assignment is *short* is essential: students often do terribly the first time, but if you give five of these assignments over a term, they get better.

Let us note that in a dig-deeper assignment, you, the teacher control how much agency the students have. We want students to have as much as possible, of course (Hüsing and Podworny 2025), but to help students get started, especially with their first foray into data projects, you can give them the initial claim while letting them find their own "opportunistic" nuance through exploration. In later projects, they can find their own claim, or even choose their own dataset, depending on their interests.

And dig-deeper projects are easy to extend. For example, students can repeat the cycle and add still more nuance. What if we added race? What if we added *year*?

The data moves portfolio

I had originally planned to ask students to include an inventory of the data moves they used in each project, to promote that metacognitive buzz. But that seemed too much to ask. Instead, I modeled data moves in class, and talked about them. We also looked at anonymous student work from the other section, critiqued it, and discussed the data moves they must have used. That is, I worked to make the data moves part of the everyday vocabulary.

I still wanted to assess their understanding of data moves themselves, so as a final project, students submitted a "data moves portfolio." You can see the actual student assignment in *Awash in Data*, but the gist of it is this: give two examples of each data move. You can find the examples in your own work from earlier in the term, or create new ones. The examples should be as different as possible.

Oh, and one more thing: for each example, explain what you accomplished by using the data move. How did it help the investigation?

This turned out to be a very successful assignment. Students created written or video portfolios that very convincingly showed how well (or, in a very few cases, how poorly) they understood the data moves and what they were for.

Conclusion

I was left convinced that making data moves a conscious part of the course design and of the student-facing materials was a good idea. I am also convinced—but have no evidence—that learning about data moves using a draggy-droppy interface like CODAP's is excellent preparation for using more advanced platforms such as R or Python. The idea is that if you know what *grouping*, *summarizing*, *recoding*, *filtering*, and *joining* are *for*, then when you have to implement these in computer code, you can concentrate on syntactical challenges rather than having to address both syntax and conceptual understanding at the same time.

Of course, my class was one brief course given under unusual circumstances with hard-working students. I welcome further work, such as that of Hudson et al. (2024) and Hüsing and Podworny (2025)—work that explores whether these impressions generalize, and that further refines or extends what roles data moves and their kin might play in learning about data science.

Data

You can explore the data in CODAP using this link. Source: IPUMS. Ruggles et al., (2025).

References

The Concord Consortium. (2024). *Common Online Data Analysis Platform* [Computer Software] (CODAP). Concord, MA: The Concord Consortium. https://codap.concord.org/app

Erickson, T. (2025). *Awash in Data*. Electronic book with live examples. Oakland, CA: eeps media. https://concord.org/awashindata.

Erickson, T. (2017). *Smelling like data science*. Talk given at the DSET conference, summarized as a blog post at https://bestcase.wordpress.com/2017/02/21/smelling-like-data-science/.

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data Moves. *Technology Innovations in Statistics Education*, 12(1). http://dx.doi.org/10.5070/T5121038001 Retrieved from https://escholarship.org/uc/item/0mg8m7g6

Fetter, A. (2011). *Ever Wonder What They'd Notice?* NCTM Ignite Talk. https://www.youtube.com/watch?v=a-Fth6sOaRA

Grolemund, G. and Wickham, H. (2017). *R for Data Science*. O'Reilly. http://r4ds.had.co.nz/

Hudson, R. A., Mojica, G. F., Lee, H. S., & Casey, S. (2024). Data Moves as a Focusing Lens for Learning to Teach with CODAP. *Computers in the Schools*, 1–26. https://doi.org/10.1080/07380569.2024.2411705

Hüsing, S. & Podworny. S. (2025). Empowering students to gain insights within data exploration projects in the classroom—Using, modifying, and creating data moves through a scaffolded use of digital tools. This volume.

Ruggles, S., Flood, S., Sobek, M., Backman, D., Cooper, G., Rivera Drew, J. A., Richards, S., Rogers, R., Schroeder, J. and Williams, K. C. W. (2025). *IPUMS USA: Version 16.0 [dataset]*. IPUMS. https://doi.org/10.18128/D010.V16.0