Students' Prior Knowledge of Data Visualization

Eni Mustafaraj*

Department of Computer Science
Wellesley College
Wellesley, MA, USA

Abstract

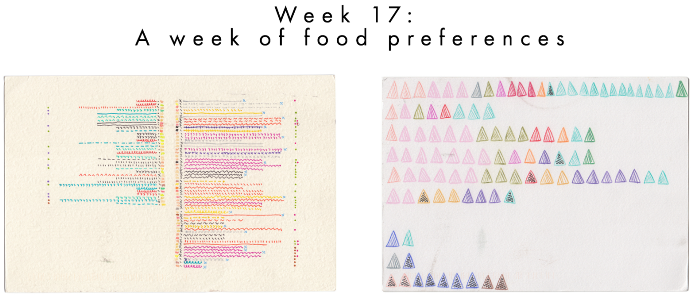
Undergraduate students who enroll in a data science course already have exposure to visualization in different contexts, and hold beliefs about acceptable standards of visualization. When confronted with non-traditional examples of visualization, such as the ones found in the popular blog Dear Data `http://www.dear-data.com/`, their reactions are mixed. Some are surprised by the creativity of the process and the beauty of the outcome; others are frustrated with the need to pay attention to the details to decipher the content, due to the novelty of the visualizations. In this paper, we use the reflections of a group of 21 undergraduate students to outline the kinds of beliefs and knowledge about visualizations that they bring to the classroom. Additionally, we present some examples of visualization work that followed the reading of the Dear Data blog, indicating how the reading might have influenced the amount of information students tried to depict in their visualizations.

1 Introduction

A successful learning environment needs to be learner-centered [1], requiring from instructors to find ways to elicit pre-existing knowledge in students and to invite them to reflect and reason upon it, as a way to engage with the new material more effectively. For certain disciplines in natural sciences, the existing "common knowledge" is so vast, it becomes a barrier to effective learning. To address this situation, researchers have compiled concept inventories, like in physics [4] or biology [2], which provide strategies for making prior knowledge explicit and helping students to deal with "misconceptions" or "naive theories". In recent years, such efforts are being made for the domain of computer science as well, for example [7].

Visualizations are nowadays found everywhere and students have previous experiences with either interpreting or creating them in different contexts. What do our students know about visualizations? How do they see them? What do they value most in a visualization? How easy or difficult is it for them to understand visualizations created by others? What visualizations do they choose to create themselves if asked so? Does the course subject or the department in which the course is offered create a certain set of expectations for how visualizations should look like or be interpreted? Being able to know the answers to such questions while teaching a course that is about visualization (either in its entirety or partially) would be beneficial to the quality of the provided instruction and the range of learning opportunities for students.

We didn't set out to answer these questions, not yet. While offering for the first time a course titled "CS 249 Data, Analytics, and Visualization" in the department of computer science at Wellesley College, we designed a multi-part assignment to trigger students' thinking about their beliefs in visualization. They provided their reflections in writing and this paper is an effort to capture what emerged from this process of reflection. The students in this course were all female, in the age range 18-22, with the majority being 19-year-old. For many of them this was their second or third course in computer science and had little to no exposure to data analytics and visualization. However, there were a few students who had taken a course in statistical data analysis and held strong beliefs about what data visualization should be. This diversity of backgrounds contributed to an interesting class discussion, as students argued for their viewpoints.



Week 17:
A week of food preferences

Figure 1: A screenshot from Week 17 of the blog Dear Data. Students read this blog post and reflected in writing about the visualizations and the process of creating them. After a few days, they created their own visualizations on one week of food habits. The blog post can be found at: `http://www.dear-data.com/week-17-a-week-of-food-preferences`

We found this learning activity very useful in several regards: students appreciated the most the description of the process underlying data visualization from experts in the field. Having a window into such a process was more important to them than understanding the visualizations. This finding agrees with what research emphasizes about the need of students for realistic examples of membership in a "community of practice". Students' emotions about the blog post influenced the kind of critique or praise they offered for the visualizations. The ones that found them beautiful focused on praising the intricate process of creativity and thoughtfulness underlying the visualizations. The ones who felt confused or panicked critiqued the large amount of information (”information dump") arguing that the authors had sacrificed legibility for aesthetics. Students who felt surprised admitted to holding certain beliefs about what data visualization is: boring, math-heavy, rigid, concrete, objective, etc. A few of them wondered about the very nature of data and whether they can ever be objective.

As we think about the goal of broadening participation in domains like computer science or data science, it becomes important to make explicit the prior beliefs that our students (especially women and minorities) hold, since they might be the blocking stones toward our goal. Meanwhile providing meaningful and inspiring examples can contribute to realizations that were unimaginable before. As one student put it: "This blog post expanded the possibilities of data visualization for me, and I feel much more comfortable going into this class having seen two fantastic examples ..."

*e-mail: eni.mustafaraj@wellesley.edu

2 An Assignment about Food Habits

The assignment presented to the students is described below. There was no grade for the assignment, students were expected to complete it and this completion was counted in a portion of the final grade that accumulated several such tasks during the semester.

1. Keep a food diary for one week about any aspect of your food consumption. **Note**: Students were sent instructions for this task one day before the semester started.

2. Read the blog-post "Dear Data: A week of food preferences" and write a 200-word reflection on anything that captures your attention in the reading. **Note**: The reflection was due on the first day of semester and it was used for a class discussion.

3. Summarize the data from your one-week food diary either visually or textually to discuss trends or tell a story about your food habits.

Since the course was designed to be an introduction to data science, we wanted to achieve a few goals with this assignment: get students to think about the process of data collection and all the choices that it entails; start practicing data collection; reflect on the process of data gathering and data visualization as practiced by two women experts in the field (Wellesley College is a women-only institution and we strive to provide students with inspiring role models); use their existing knowledge to produce simple visualizations.

2.1 The "Dear Data" Blog Post

The "Dear Data" blog is an analog data drawing project by two award-winner information designers: Giorgia Lupi and Stefanie Posavec. For every week in an entire year, they collected personal data about one aspect of their life, made a drawing from the data in one postcard and sent it to each-other via regular mail. Each weekly blog post features the two postcards and the notes that Giorgia and Stefanie wrote to provide context about the data gathering and the drawings. Figure 1 shows the screenshot for Week 17, that focuses on food preferences. As the authors claim in their website, "there is a huge potential to use this method to inspire students and non-data experts to start working with data". As we mentioned, that was one of our motives for using this material in an assignment.

The food visualizations in this blog post are based on the favorite food/flavors of each author. That is different from what students did in the assignment: they collected daily data about food intake and other eating aspects for an entire week. Instead, Georgia and Stefanie "asked themselves" about what foods they liked and disliked.

Stefanie, an American who now lives in London, chose to rank unprepared food by showing the most favored at the top and the least favored at the bottom. Each food is a triangle, the color indicates the food category (e.g., fruit, vegetable, etc.), the mesh-like coloring indicates a food Stefanie encountered when moved to UK (the bottom row). In the postcard, Stefanie has listed the colors of the 14 food categories, as well as the names of the 27 food items that are her favorites (top row). In total, her visualization contains 105 food items, in 9 levels of preference, separated in 14 food categories by color, with the added binary feature of being a UK/European food. Because drawing by hand it's hard, there is no consistency in the lengths of the rows (the first one has 27 items, the fifth row has 19 items, but they look almost the same).

Giorgia, an Italian who lives in the United States, decided to group foods in two columns: foods that she likes on the right, and foods she doesn't like on the left. Each line is a food and its length indicates the amount of calories (the longest line has more than 500 kCal/100g). The color of each line is the flavor of the food: plain, bitter, sweet, salty, etc. The line type is the main ingredient: — is fruit, xxx is milk/cream, etc. In addition to three major attributes (flavor, main ingredient, and calories), Georgia has decorated her lines with four additional symbols to indicate color of food, whether it's cooked or general, whether she had it last week, and a set of binary features such as (it's from Chicago, my boyfriend recommended it, etc.). Giorgia has listed 100 foods, with more of them being in the favorited column. She hasn't provided names for the food items in the postcard, but they are included at the end of the blog post.

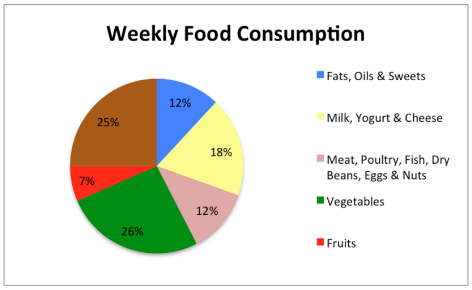


Figure 2: The pie chart was a frequent choice for students. This is an indicator of their preference for simple charts that depict the relationship between two attributes, as well as their familiarity with pie charts.

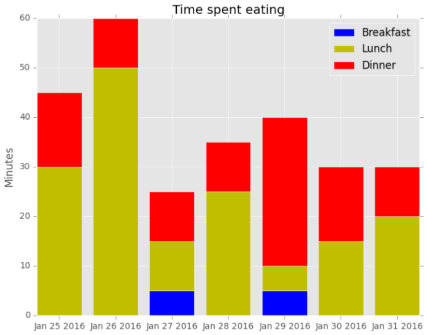


Figure 3: This student chose to visualize the daily duration of her meails, explaining that time was the most objective attribute in her data set. By using a stacked bar chart, she was able to make use of three attributes in this discrete time series.

2.2 Students' Data

Each student kept a food diary for one week. They used different techniques to do this: some took pictures, some wrote notes at the end of the day, others took notes during each meal. Studying their descriptions of these diaries, we noticed more than 20 different attributes used for data collection (some of them more common their others): temporal attributes: date, day, hour of day, duration, meal; social aspects: location, with friends or alone, number of friends, which friends; food attributes: food category, food item, amount of food, drink, nutritional value, food color; physiological attributes: hunger level, satisfaction level with food, mood during meal, etc. Several students commented on how revealing this process of daily data gathering had been, and some of them wondered about the accuracy of it, since it started affecting their choices. This is exemplified in the following comment:

> ... as I collect my own data, I find that my food habits are literally changing for the better. I make sure to pick

> up 2 pieces of cantaloupe, avoid the fries and hone in on the vegetables just so I can write down that I ate more healthy things than unhealthy things. That's interesting to me because data collection skews itself through the process. Only a person following me secretly and documenting what I eat would get a full and honest account of my usual eating habits.

2.3 Students' Reflections

After reading the blog post (Week 17: A week of food preferences), every student wrote a paragraph about 200-word long to express their thoughts and feelings about the visualizations and the accompanying text. We had a class discussion to summarize students' response to the blog post. In a separate section of this paper, we outline the students' reaction in terms of prior beliefs, emotional response, as well as critique and praise for the visualizations.

2.4 Students' visualizations

The third part of the assignment asked students to summarize their food diary preferably through a visualization. Not everyone chose to do so, but analyzing the work of students who did provides interesting insights. Some students who had previous experience with software for producing simple charts created pie or bar charts that captured two or three attributes at a time. In this case, students visualizations were limited to the choices offered by such software tools (for example, Figure 2), or their programming skills (Figure 3).

Some students, in the spirit of the Dear Data blog post, chose to create analog drawings. Because of the freedom that comes with drawing, they were able to depict the relationship among more attributes from their dataset (five for the visualization in Figure 4 and four for the visualization in Figure 5, while still maintaining the formal nature of the visualization.

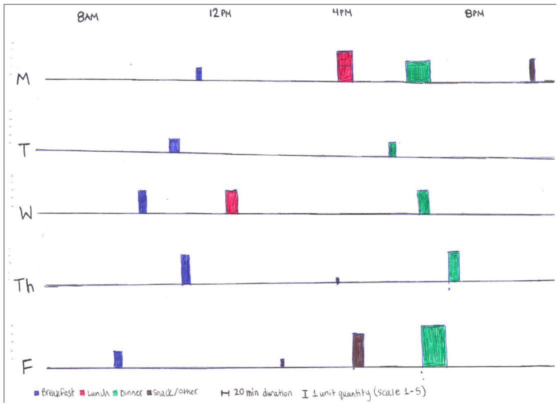


Figure 4: This student was interested in the consistency of her meal eating habits. The colors indicate the meal: blue for breakfast, red for lunch, green for dinner, and brown for snack. She used a grid with temporal attributes: day of the week and hour of day, the width of each bar to indicate duration and the bar hight to indicate food amount. This analog drawing allowed her to capture relations among five dataset attributes.

Finally, some students chose to create infographics of different styles, capturing a range of attributes, with the more complex of them, Figure 7, using six attributes. In a departure from the abstraction shown in the drawings of the Dear Blog data, where food items were converted into line types, triangles, and colors, students elected to use real images of food to make their visualizations closer to the reality (e.g, Figure 6 and Figure 7), as well as to their gathered data. They chose to summarize their food diary truthfully in their visualization, in an effort to make it easier to understand.

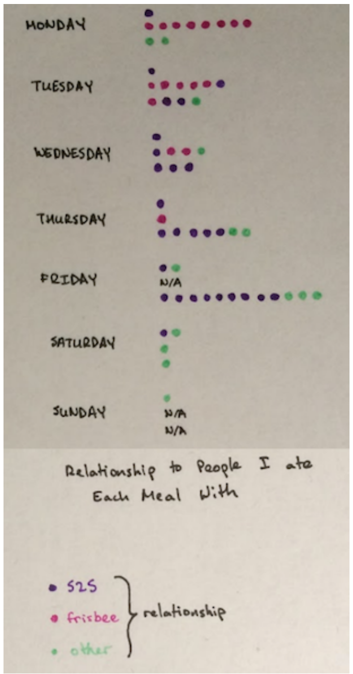


Figure 5: This student was interested in the sociable aspects of the meal. She used colors to distinguish three kinds of friends and provided counts of them for each meal, for each day of the week, making use of four attributes.

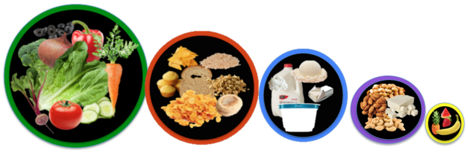


Figure 6: This student was interested in the content of her diet. Each circle represents a food group, with size indicating prevalence. The content of each circle depict the real food items, with proportion indicating frequency of eating that item. Despite the lack of numbers, this graphic captures four attributes: food group, food item, rank of food group and amount of food items.

3 Students' Beliefs and Prior Knowledge

Students in their written reflections wrote about a broad range of issues. Their emotional reaction was particularly evident, running the entire spectrum from "panicked", "uncomfortable", "confusing", "struggled", to "impressed", "intrigued", "excited", "astonished", and "rewarding". The frequently expressed surprise at the visualizations was explained with the belief they held about the **definition** of data visualization. Common statements were: "I never really thought about data visualization as art before reading this post." or "When I think of data visualization, I think of taking information (numbers) and visualizing them in the form of graphs, charts, tables; I would have never thought of visualizing food, flavors of food, types of food."

Adding to the confusion about the definition was the belief about the **purpose** of data visualization: "I used to think that there was a standard to visualizing data, and it was supposed to be rigid and concrete so it could be easily interpreted by most people." In fact, the belief that the main purpose of visualization is "to offer clarity and insight" was at the root of the often-expressed criticism that was best represented in this statement: "The presentation definitely val-

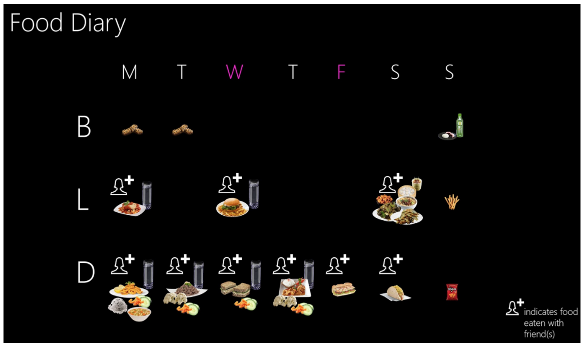


Figure 7: This student had the most exposure to art and design. As a result her infographic became a mixture of formal and informal: she used a grid with temporal axes (day and meal type), and then used a pictorial style to describe the food items, drink consumption, friends' presence, and food amount. In total, she used six attributes from her dataset.

ues aesthetics over actual readability, which would be detrimental in a more serious project."

Another revealed belief was about **representation choices**. While some students only showed surprise: "I would have never thought to use lines to represent different foods and their attributes.", others expressed the preference for more realism: "I feel like the food preferences could be better represented by actual food-like images – perhaps I'm just used to infographics that use pictograms." Indeed, we can notice this preference in some of the infographics that students created (interestingly, by students who had not expressed such a preference in their written reflections).

Some students seem to hold strong beliefs about **data provenance**. One of them wrote: "In traditional data science, Giorgia's method of data collection could be considered less professional" or "I've trained myself to be cautious around subjectivity in data". Yet another student wondered, echoing a point raised in the blog post by Stefanie: "is data something totally external to you or something that your opinions and thoughts can influence?"

Finally, many students specifically addressed the role that the way we see and experience the world, **categorical thinking**, is at the root of the differences in these visualizations, despite the topic being the same: food preferences. Stefanie and Giorgia had made cultural (e.g., UK food) or social (e.g. recommended by boyfriend) aspects of food into attributes of the data and that was both puzzling and very novel to the students. As one student put it: "It would be time-saving to come up with a universal way to categorize, although it might ruin the fun of the book."

4 Related Research

[5] discusses the role of prior knowledge in how users perceive and interpret visualizations, which can be relevant when designing interfaces for visualization. The paper lists several beliefs (drawn from the existing literature), such as: beliefs about graphs (e.g., "graphs are scientific") or beliefs about graph types ("line graphs show trends") and reminds us that existing literacies, such as statistical or graphical will interact in even more influential way with the interpretation of visualizations.

As the visualizations discussed in this paper showed, students possess certain skills about creating visual representations (dot plots, charts, infographics, etc.) which fall in a continuum. A detailed description of a proposed continuum is presented in [8], from a study where individuals with high level of education were asked to summarize the same dataset through hand sketching. One can imagine using an exercise like this to assess where students fall in the continuum line from numeracy to abstraction.

Another aspect of the students' visualizations was their preference of "realism", showcased by the inclusion of real food pictures in their infographics. This preference is discussed in the context of cartography for meteorology students by [3], where the authors find that students hold naive intuitions about information displays, which contrasts with the need for abstracting from the real world.

Most students in their reflection engaged in critiquing of the visualizations and its explanation. Being able to critique the work of people who are not their peers seemed to provide an opportunity to express one's opinions easily. As the work presented in [6] has shown, commenting on data visualization blog posts is an activity that attracts many web users. One can imagine engaging students in guided explorations of such commentary online space with the goal of training them to provide meaningful social feedback in the context of data visualizations.

5 Discussion and Future Work

This paper is not a formal investigation of students' beliefs and prior knowledge of data visualization. Its purpose is to emphasize the need for doing so as an important step toward effective instruction, especially when thinking about broadening participation in the field. As the amount of online visual information increases, there seems to be more confusion about what constitutes standard practices of visualizaton. This calls for a need to provide students with a taxonomy of the different kinds of visualizations and the contexts in which they make sense. When is a realistic depiction desirable, and when to strive for a higher level of abstraction? Does aesthetics have a place in "serious" visualizations (e.g. in economics or politics)? What values and criteria should we use to interpret a visualization? Students seem to wonder and worry about these questions, and we need to find the right way to answer them. Becoming a good software engineer entails the ability to read and understand code written by others. How do we approach instruction in data visualization to specifically address the issue of equipping students with the ability to read and understand the ever-increasing range of visualizations from different domains? A community interested in the pedagogy of data visualization (which this workshop is trying to establish), might be a good place to start.

Acknowledgements

All student visualizations in this paper are included with their explicit written permission. The author is very grateful to all students enrolled in her course "CS 249 Data, Analytics, and Visualization" at Wellesley College in the semester of Spring 2016.

References

[1] J. D. Bransford, A. L. Brown, R. R. Cocking, et al. *How people learn*. Washington, DC: National Academy Press, 2000.

[2] C. D'Avanzo. Biology concept inventories: overview, status, and next steps. *BioScience*, 58(11):1079–1085, 2008.

[3] M. Hegarty, H. S. Smallman, A. T. Stull, and M. S. Canham. Naïve cartography: How intuitions about display configuration can hurt performance. *Cartographica*, 44(3):171–186, 2009.

[4] D. Hestenes, M. Wells, G. Swackhamer, et al. Force concept inventory. *The physics teacher*, 30(3):141–158, 1992.

[5] J. Hullman. How prior knowledge affects the processing of visualized data. In *ACM CHI'13 Workshops: Many People Many Eyes*.

[6] J. Hullman, N. Diakopoulos, E. Momeni, and E. Adar. Content, context, and critique: Commenting on a data visualization blog. In *Proceedings of the 18th ACM CSCW*, pages 1170–1175. ACM, 2015.

[7] L. C. Kaczmarczyk, E. R. Petrick, J. P. East, and G. L. Herman. Identifying student misconceptions of programming. In *Proceedings of the 41st ACM SIGCSE*, pages 107–111. ACM, 2010.

[8] J. Walny, S. Huron, and S. Carpendale. An exploratory study of data sketching for visual representation. *Computer Graphics Forum*, 34(3):231–240, 2015.