Data Visualization—A tool for solving unsolved mysteries

Tina Mazzacane, Laurell Wiersma and Padmanabhan Seshaiyer

The Virginia Board of Education approved the *Data Science Standards of Learning* in April 2022. These standards support a ½ or 1-credit high school mathematics course in Data Science. The rigor of the course is intended to be at or above the level of Algebra II. The *Data Science Standards of Learning* introduce the learning principles associated with analyzing big data. Using open source technology tools, students taking a course based on the *Data Science Standards of Learning* will identify and explore problems that involve the use of data information and data-intensive computing to find solutions and make generalizations. Students will engage in a data science problem-solving structure to interact with large data sets as a means to formulate problems, acquire, collect and clean data, visualize data, develop models using data, and communicate effectively about data formulated solutions.

The development of data literacy is an imperative in today's 21st Century world as the amount of data that the average citizen consumes and produces every year continues to increase exponentially. Every individual must have the ability to synthesize data in order to support daily decision making, make sense of our world, and prepare for the future (Wilkerson, NCTM, 2020). Data Science is an intersection of mathematics, statistics, and computer science. As the Virginia Department of Education revises the current *Mathematics Standards of Learning,* a greater emphasis on the process of using data to make informed decisions and solve problems will be included. Teachers in every discipline can include data analytics as a structure to examine problems and find solutions.

Data Science continues to help unravel mysteries from the past, present and future. In this article we demonstrate the power of data visualizations through examples from multiple perspectives: scientific investigation, historical analysis, and understanding the world around us. These examples show teachers how data talks can engage all students in quantitative analysis and to motivate student interest in mathematics, statistics, computer science and STEM in general.

Mysteries Solved Through Use of the Data Cycle

Remember the Perpendicular Bisector Theorem from geometry in high school? The one that states that *any point on the perpendicular bisector is equidistant from both the endpoints of the line segment on which it is drawn*? Well, it turns out that one of the most prominent unsolved mysteries dating back over 150 years was solved using this theorem.

The famous cholera outbreak in Soho, London (1854), had killed over 10% of the population in just a few days (Ball, 2009). While everyone wanted to believe that the disease was spread by "toxic air" (Miasma), it took one physician, John Snow, to convince everyone mathematically that the disease was spread through contaminated water. In fact, he used census data of people who died due to cholera to create a geographical visualization (See Figure 1) and used the idea of the Perpendicular Bisector Theorem to calculate the time it would have taken people to travel to their nearest water pumps (See Figure 2). His representation of the data helped to show that most of the deaths were tightly clustered in a specific area, crowded around the water pump at Broad Street (now Broadwick Street) in Soho. Snow's mathematical evidence using the application of perpendicular bisector theorem that cholera was waterborne is one of the founding moments of epidemiology and the use of mathematics to understand disease, one of the greatest advances in medicine that has saved millions of lives (Snow,1856; Hempel, 2014; Tulchinsky, 2018).

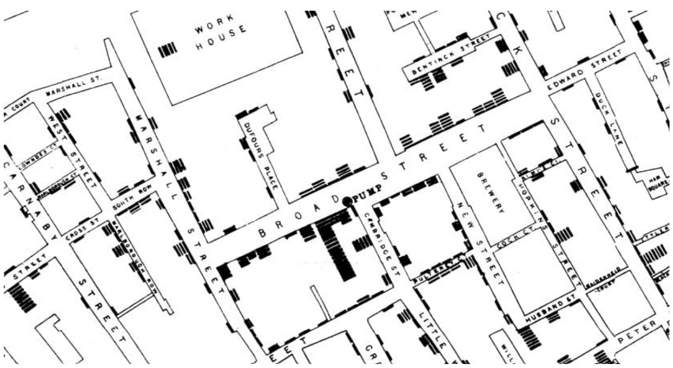


Figure 1: Data visualization of the 1854 Broad Street Cholera Outbreak

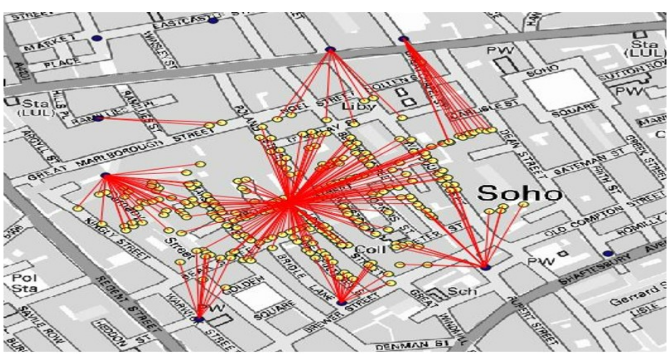


Figure 2: Broad street pump at equal walking distances from neighboring pumps

The data cycle includes posing questions, collecting data, analyzing data, making visualizations, analysis and communication of finding. This article focuses on the use of visualizations to engage students analyzing visualization to promote critical thinking and an understanding that quantitative reasoning applies to all aspects of our world. In a data talk, students would discuss what they see in the visualization. After that discussion, a teacher can add in more context to enlighten, engage, or energize students to delve deeper.

Mysteries Uncovered Through Data Visualizations

Another great unsolved mystery is revealed in Charles Minard’s depiction of Napoleon's 1812 campaign into Russia (Tufte, 2001; Tarle, 2018) where Napoleon takes a massive army to attack Russia in June of 1812 but realizes that he had to retreat after reaching Russia. Minard illustrated this journey through an innovative data visualization that helps to look at data from different perspectives (see figure). The power of the illustration lies in the picture as a narrative display of quantitative information in time and space that illustrates how multivariate complexity can be integrated in a complex yet subtle way. For example, the graphic shows the interplay between six different types of both numerical and categorical data including geography, time, temperature, the course and direction of the army’s movement, and the number of troops remaining. The size of the army was represented by the widths of the bands outward (gold) and returning (black) with one millimeter representing 10,000 men. Also, note that this map was created without any geospatial tools. Today there are a variety of tools for geographic information systems (GIS) that vary in sophistication from easy to use to extremely powerful.

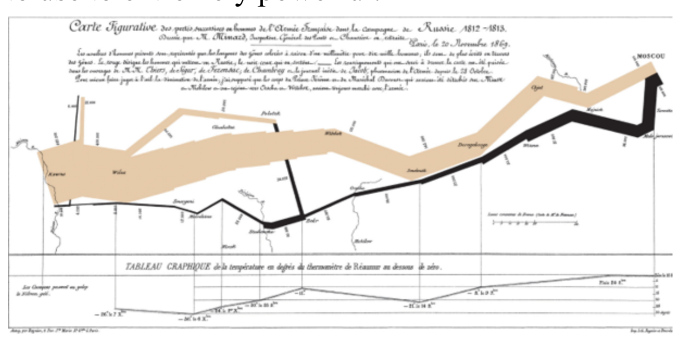


Figure 3: Data visualization of the Napoleon’s 1812 March by Charles Joseph Minard

Minard’s chart in Figure 3 tells the dreadful story with painful clarity of Napoleon's grand army that set out from Poland with a total of 422,000 men;

only 100,000 reached Moscow; and only 10,000 returned (Hardy, 1992). The goal of the various graphic representations is not necessarily to try to reveal new patterns, but to demonstrate that alternative representations give different perspectives and new insights into data. While it was not evident during the event, one of the unsolved mysteries that became obvious later from the map was deaths due to the plummeting temperatures on the return journey which is shown at the bottom of the illustration. The value of incorporating multiple variables into one graphic helps students make the connections between images and stories from history. The unsolved mystery regarding the reduction in the number of soldiers in the retreating army becomes clearer from the temperature data that suggests that more than half of the retreating army did not survive the crossing the River Berezina. Another aspect of this visualization is how text and graphics are integrated in a consistent fashion across multivariate complexity including the size of the army, location, direction, temperature, and time.

Data Talks in the Classroom

The two examples in the previous sections reveal the power of visualizations to expose patterns and solve mysteries in many fields. The ability to understand and interpret visualizations can be a highly engaging classroom activity and is a life skill in our data driven society. A classroom data talk is a low entry, high ceiling activity that engages all students in thinking critically about a visualization. Every student can find a way to participate regardless of the background knowledge. Often, a data talk takes place at the beginning of a class to allow students to begin class with a universally accessible activity. A data talk may start with projecting a data visualization for the entire class to see, having students look at a visualization on individual devices, or having a paper copy of the visualization for students. The easiest way to begin a data talk is by asking students two questions: What do you notice? What do you wonder? For instance, imagine the responses from students by starting a class with John Snow’s cholera map or Charles Minard’s illustration of Napoleon's march.

The power of a data talk is that students can analyze conventional and unconventional visualizations regardless of previous exposure. Conventional visualization may include bar graphs, histograms, pie charts, boxplots to name a few. At first teachers may feel comfortable limiting data talks to

Stephen Curry Is One Of The Best
All of his shots, 2015-16 regular season

SHOT DISTANCE (5FT)	FGM	FGA	FG%	3PM_	3PA	3P%	EFG%	BLKA	FGM (%AST)	FGM (%UAST)
2015-16	805	1598	50.4	402	886	45.4	63	52	46.6	53.4
Less Than 5 ft.	272	422	64.5	0	0	0	64.5	31	40.4	59.6
5-9 ft.	35	72	48.6	0	0	0	48.6	10	22.9	77.1
10-14 ft.	29	57	50.9	0	0	0	50.9	1	31	69
15-19 ft.	38	102	37.3	0	0	0	37.3	4	28.9	71.1
20-24 ft.	158	335	47.2	129	276	46.7	66.4	4	66.5	33.5
25-29 ft.	251	563	44.6	251	563	44.6	66.9	1	51	49
30-34 ft.	15	26	57.7	15	26	57.7	86.5	0	20	80
35-39 ft.	2	5	40	2	5	40	60	0	0	100
40+ ft.	4	14	28.6	4	14	28.6	42.9	1	0	100

SHOT AREA	FGM	FGA	FG%	3PM_	3PA	3P%	EFG%	BLKA	FGM (%AST)	FGM (%UAST)
Restricted Area	263	399	65.9	0	0	0	65.9	29	39.5	60.5
In The Paint (Non-RA)	55	113	48.7	0	0	0	48.7	12	32.7	67.3
Mid-Range	85	200	42.5	0	0	0	42.5	7	32.9	67.1
Left Corner 3	30	63	47.6	30	63	47.6	71.4	0	96.7	3.3
Right Corner 3	27	53	50.9	27	53	50.9	76.4	1	85.2	14.8
Above the Break 3	342	757	45.2	342	757	45.2	67.8	2	50.3	49.7
Backcourt	2	11	18.2	2	11	18.2	27.3	1	0	100

Table 1: Performance metrics of Stephen Curry

the visualizations that students learned in math class. The ability to analyze a visualization to solve a mystery does not rely on prior knowledge of the visualization but on reading and critical thinking skills. Some unconventional visualizations found in the media today include heat maps, bubble charts, Sunbursts, and data over geographical regions. By experiencing data talks, students become accustomed to analyzing and solving data mysteries. In addition, they are motivated to create and analyze their own data mysteries.

Here is a data talk about understanding the mystery of Steph Curry who is one of the finest basketball players and played for the Golden State Warriors from 2009 - 2023. When presented with a table of basketball statistics (see figure), students are asked to share what they notice and what they wonder.

When students see Table 1, there are a limited number of comments. Those who know the jargon on basketball statistics may recognize a percentage or two. A few people who know Steph Curry may comment on his overall career and not on the numbers in the data. However, participation changes rapidly when displaying a heat map of Steph Curry's shots in the 2015-2016 season [2].

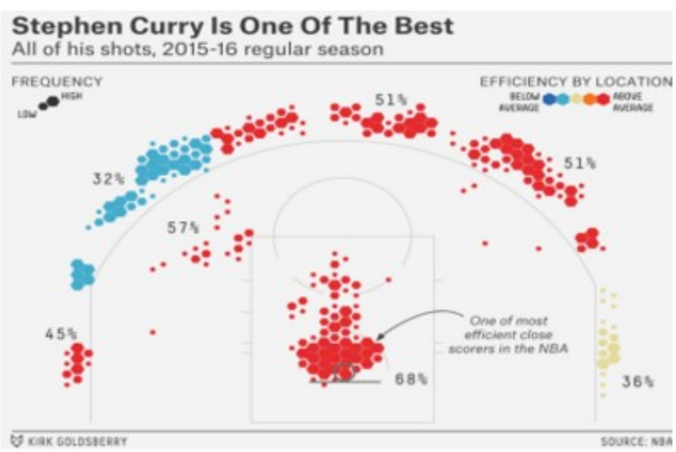


Figure 4: Heat map of Steph Curry's basketball shots

When asked to share what they see and what they wonder, students who are unfamiliar with basketball are able to discuss multiple variables about Steph Curry. "The lowest percentage is on the right side" This visualization allows for discussion about location, frequency, and efficiency. "His 3-point percentage is mostly above average for most spots." Students, regardless of their knowledge of basketball, begin to wonder about the mystery of Steph Curry as a basketball player. Some students may wonder about hand, foot or eye dominance. Others may want to know more about his consistency over many seasons. In addition, students with knowledge of basketball will be able to share their knowledge with the class. Most students do not know that a heat map is a visualization that displays data using color; warmer colors represent higher values and cooler colors represent lower values. During a data talk, students often conjecture about the color choices in a visualization.

As students' statistical knowledge increases, this data talk may become an entry point into a deeper analysis of the tabular data about Steph Curry. For example, a student can analyze the probability of Steph Curry making a shot compared to the distance from the basket. As Figure 4 shows, a student may note that Steph Curry will make a shot from under the basket with a probability of about 0.7 and from about 22 feet away (typical three-point shot distance), the probability of him making a shot is about 0.5. Or a student might create a new visualization to display each shot but classify the shots based on making or missing the basket. Using simple, free software like CODAP [3], students are empowered by data talks to dig deeper into mysteries that interest them. In fact, some students will begin to explore the field of data used in sports analytics.

Discussion and Conclusion

Data talks have the ability to increase student interest, reasoning and critical thinking in the classroom. The Virginia Department of Education recently created a new high school course, Data Science [4]. Data talks are central to the critical thinking in this course. However, the real power of data

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

[2] https://www.youcubed.org/wp-content/uploads/2020/09/Basketball.pdf

[3] https://codap.concord.org/

talks is building a new generation of people who will be able to employ STEM problem solving skills that provide for the formulation of data to drive action.

In the last century, many great achievements have been made due to data and computing from abundant supply of food, medicines and safe drinking water to bringing goods and services wherever they are needed. While these achievements through data have been outstanding, there are still greater challenges and opportunities that remain to be realized. To engage students in a data cycle effectively, such global and local challenges can serve as great real-world contexts to motivate a project-based learning approach to data science that can advance students' cognitive development. Some examples of such frameworks include the National Academy of Engineering Grand Challenges[5] announced in 2008, which identifies fourteen game-changing goals for improving life on the planet under four crosscutting themes including sustainability, health, security and joy of living. With the rise in crowdsourcing, new technologies for data collection and the explosion in the availability of big data and artificial intelligence, it is possible now to process and analyze data in real time that can help provide new insights in these challenges. The integration of this new data with traditional data can potentially produce high-quality data analysis to help solve unsolved mysteries posed by these challenges.

Acknowledgements

This work has been supported in part by the Virginia Department of Education. The authors would also like to acknowledge Brendon Albon, Lisa Bussian, Deborah Crawford, Sara Fergus, Heather Holcomb, Christa VanOlst and Aanand Vasudevan for their review and feedback on the article.

References

Ball, L. (2009). Cholera and the pump on Broad Street: the life and legacy of John Snow. *The History Teacher*, *43*(1), 105-119.

Hardy, C. (1992). Minard's Graphic of Napoleon in Russia. *New England Journal of History*, *49* (2), 28-31.

Hempel, S. (2014). *The medical detective: John Snow, cholera and the mystery of the Broad Street pump*. Granta Books.

Snow, J. (1856). On the mode of communication of cholera. *Edinburgh medical journal*, *1*(7), 668.

Tarle, E. (2018). *Napoleon's invasion of Russia, 1812*. Pickle Partners Publishing.

Tufte, E. R. (2001). The Visual Display of Quantitative Information. 2nd edn Cheshire. *Connecticut: Graphics Press LLC*.

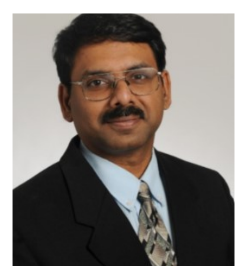
Tulchinsky, T. H. (2018). John Snow, Cholera, the broad street pump; waterborne diseases then and now. *Case studies in public health*, 77.

Wilkerson, T. (2020). Statistics education: An imperative for our future. *NCTM Presidential Message (November 2020).*

Tina Mazzacane
K-12 Mathematics Coordinator
Virginia Department of Education
Office of STEM and Innovation
Tina.Mazzacane@doe.virginia.gov

Laurell Wiersma
Math Department Chair
Wakefield High School
Arlington Public Schools
Laurell.wiersma@apsva.us

Padmanabhan Seshaiyer
Professor and Director
Department of Mathematical Sciences
George Mason University
pseshaiy@gmu.edu



[4] https://sinews.siam.org/Details-Page/preparing-virginias-students-for-new-post-secondary-pathways-in-data-science

[5] https://www.engineeringchallenges.org/