Harvard Data Science Review • Issue 2.4, Fall 2020

High School Data Science Review: Why Data Science Education Should Be Reformed

Angelina Chen¹

¹Princeton High School, Princeton, New Jersey, United States of America

Published on: Oct 29, 2020

DOI: <https://doi.org/10.1162/99608f92.1e28ce9e>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Column Editor's Note: *In this column, Angelina Chen, a current high school student, writes about the importance of data science education for the pre-college crowd, and in particular for secondary (high school) students. This is the perspective of one student who had early exposure to statistics, but the lessons she has learned along the way are more broadly applicable. How can we as a community improve the teaching of our subject at the school-age level? By the time young people reach college levels, we may have already missed the best opportunity to teach probabilistic thinking as a native language for reasoning under uncertainty. We welcome other perspectives on these questions from students and teachers. A companion column on the new "GAISE II" report sheds some light on current efforts in this direction.*

My father is a professor of statistics. I was aware of this fact since I was very young. It was one of those things kids had to memorize, like their home address, their parents' phone numbers, or their siblings' birthdays. I did not really know what it meant; it was just something to recite when a stranger asked me what my parents did. When I glanced at the papers piled upon my dad's desk or clutched in his hands as he took a power nap, I admired the squiggly symbols and detailed graphs like they were ancient languages and art pieces: fascinating, but completely foreign and devoid of meaning.

In elementary school, I learned that statistics meant finding the mean, median, and mode of a set of numbers and drawing this set on a graph, like a connect-the-dots or a cityscape. Being the art kid I was, I would doodle little flowers or patterns within the bars of a bar graph, or use different colored pens to connect lines. It seemed so different from the complicated things my dad was doing. No new information was added to my knowledge of statistics until I had the option to take a statistics elective in high school, as the core math curriculum very rarely delved into statistics at all. Well beyond my doodling years, I still had virtually no clue what my dad did in his career.

Nowadays, the adults around me—mostly professors or professionals in business and STEM fields—talk about 'big data,' 'machine learning,' and how data science is going to be burgeoning in the next few years. They even have heated debates on the differences between 'statistics' and 'data science.' Although there isn't a definitive answer, it seems that most people agree that data science is an integration of statistics and advanced computation with aims of solving large scale problems, though the debate goes on.

However, the main message was that a growing number of jobs is already not accompanied by the same amount of growth of students' studying data science. A graduate who has the necessary knowledge will be extremely sought-after, as businesses need to keep up with an increasingly data-driven society. But not only are data science professionals wanted, they are also indispensable. Data is used in essentially every field: to study DNA to cure diseases, to analyze the position and orientation of stars, to track political campaigns and voter opinions, to help businesses respond to consumer feedback, to determine standardized test scores, and much, much more. We need more data scientists to continue to make advancements.

Even if one is not aiming to become a professional in the field, being a data-literate person is extremely important for everyday life. Data helps us be well-informed citizens and make decisions, from choosing a career path or school, to understanding the news, to knowing how we receive our music, movie, and product recommendations, and even to understanding how social media news reaffirms our political beliefs.

So how come, at the high school level and younger, people are not discussing these skills like they discuss coding, another ‘hot’ skill? If you ask high schoolers, their dream jobs include being a doctor, engineer, computer scientist, etc. But it seems like no one is interested in being a data scientist.

Out of my public school’s nearly 200 courses offered, only two have a focus in statistics or data science: an introductory-level class and an advanced placement (AP) class. Considering that my public school district is one of the best in the country (ranking at number forty-seven on [Niche’s](#) 2020 national list and tenth in the state), it is easy to imagine that at most other schools, the data science and statistics offerings are even more sparse, if they exist at all.

To reform and improve, I believe that the first order of business should be to provide widespread data science education at all schools. It’s important at least to make students aware that the field is an option, and to provide the necessary baseline of data education to become an informed member of society. We should not stop at finding the average and drawing graphs. I learned what standard error and standard deviation are, along with how to read what graphs tell you and how to present my own data with a graph using a computer, in my ninth-grade biology class. That should have happened far earlier, and students who are not placed in that class under that teacher might never be taught at all, since data science courses are not required.

Beyond the lack of availability, there are still improvements to be made to the existing introduction to this field. When my brother first took the AP Statistics class at my school, though he didn’t dislike the class, it didn’t motivate him to pursue it in the future. Because of the way it was presented as an optional add-on, data science didn’t seem like a crucial field of study. Since the class was also students’ introduction to data science, it was not able to delve deeper into real-world applications and recent advancements that are often fascinating and relatable to the next generation. AP statistics classes (the more popular of the two classes offered at my school) spend the year preparing for the College Board’s AP exam, covering the basics of one-variable data, two-variable data, collecting data, probability, sampling distributions, proportions, means, chi-square, and slopes. If these basics are taught in the regular math class curriculum, there will be time to teach more advanced topics that can motivate students to pursue data science in the future. In a society where students are expected to have a general idea of what they want to do in life by the time they are eighteen, it is important that we introduce data science favorably early on.

My dad took it upon himself to teach my brother statistics his own way. Because of these at-home lessons, my brother gave statistics another chance at his university, and even ended up majoring in it. Partly because of my brother’s initial experience with statistics, throughout my childhood my father would frequently talk about

articles he read that related to things he was researching, a new way a person or company was applying some concept, or just some interesting story detailing ways statistics is applied in the real world. He would point to a graph or image and marvel at how beautifully the data was displayed. His eyes would glint with excitement about the topic itself, and also being able to share it with someone.

To be honest, when I was younger, I would sometimes tune out of these often one-sided conversations. *Yeah, yeah, boring math stuff*, I would think, eager to get back to the fantasy novel I was reading or the puppy game on my Nintendo DS. But as I grew older, I began to listen, and as soon as I started listening, I was entranced.

He talked about height discrepancies between parents and children to teach me about regression towards the mean. He compared statistical hypothesis testing to a court of law with concepts of ‘innocent until proven guilty’ and margins of error. He asked and answered interesting questions, like why Google succeeded as a search engine while Bing, Yahoo, and so many others lagged behind to become ridiculed memes. He told me stories of recent and past discoveries and developments, like how data science made image recognition possible (yes, the same kind that gives you recommendations on which friends to tag on your Facebook posts), so when someone inputs an image, the machine will return a result. And now, the opposite is becoming true as well. If someone inputs something like “cat in a tree wearing a hat,” it is possible for the computer to create an original image according to your input. The way some translating services work is not with a direct dictionary and grammar book; they use data of past translations to learn from them and formulate new ones, hence, ‘machine learning.’

People lucky enough to have others who are passionate about data science in their lives should not be the only ones able to glean this inspiration. These kinds of stories should be offered within our school systems. Support is needed from professionals like my father who work on the frontier of data science everyday to bring their knowledge and excitement back to our classrooms.

As our world is changing, our curriculum needs to change as well. We cannot brush off data science as an optional elective that most students will never truly be exposed to. Educating kids about statistics from a younger age so that future courses can supplement with more interesting and relatable material is vital towards filling the gap between the supply and demand of statisticians and data scientists, as well as creating a more data-literate society. Both are necessities for the future.

Disclosure Statement

Angelina Chen has no financial or non-financial disclosures to share for this article.

©2020 Angelina Chen . This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.