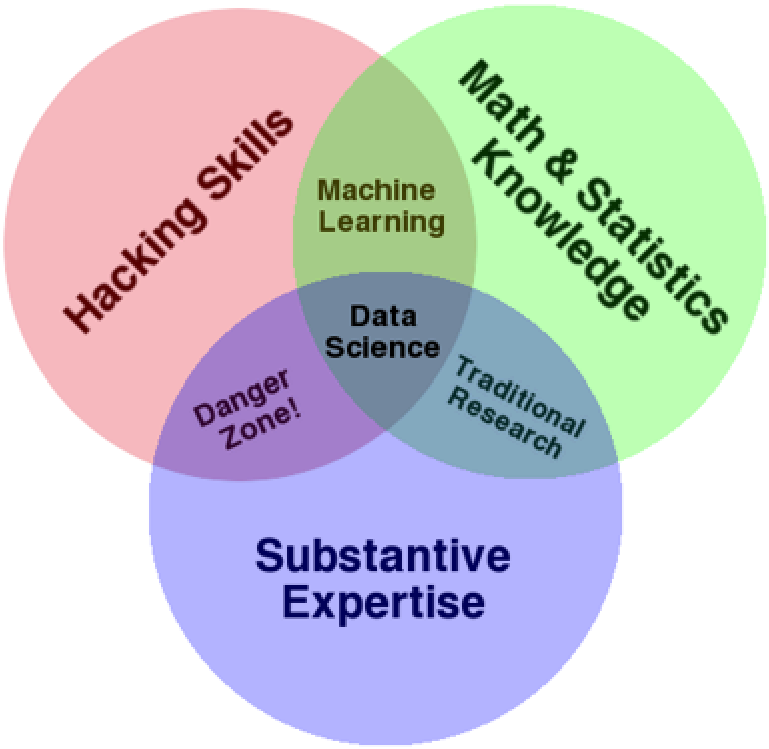
A Modification of Drew Conway's Data Science Venn Diagram

During week one of our Data Science Immersive Program at General Assembly, one of our instructors, Matt Brems, presented Drew Conway's Famous Venn Diagram.



Drew Conway's Data Science Venn Diagram

Brems explained each of the circles, as well as the intersections, and highlighted the Danger Zone. Below is his experience regarding people in that intersection:

"I've personally seen people in the "Danger Zone." These people have a very solid understanding of the real-world problem to which they apply data science and have excellent programming skills—even using libraries or functions that involve statistics. Because of the skills these people _do_ have, they know that a particular modeling technique might be designed to solve a real-world problem and can leverage their Python (or R, etc.) skills to implement it. However, no error message pops up warning the user that assumptions of a modeling technique aren't justified or that the data isn't properly scaled.

In many cases, what happens is that the types of models being built rely on assumptions that aren't met. As a result, these models might not accurately reflect what is truly happening and so using these models to better understand the world or to explain some real-world phenomenon may be inappropriate. Though the analysis is usually well-intended, just because Python is able to give you a result doesn't mean it is a meaningful one. When someone takes action on a possibly misleading result… well, that is why the "Danger Zone"—not understanding the statistics behind modeling—can be so dangerous."

Conway has a similar view of people in the Danger Zone but also suggests that malice may be a motivator for some people in this intersection:

"In this area people who are perfectly capable of extracting and structuring data, likely related to a field they know quite a bit about, and probably even know enough R to run a linear regression and report the coefficients; but they lack any understanding of what those coefficients mean. It is from this part of the diagram that the phrase "lies, damned lies, and statistics" emanates, because either through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created."

I understand where Brems and Conway are coming from, but Danger Zone should not be the default label for those possessing programming skills and domain knowledge but lack the math skills. Yes—it is possible for people in this region to run into danger, but that is if they try and make statistical inferences without the proper knowledge of math or statistics. However, people in the Danger Zone can make enormous contributions to an analytics/data science team. I say this because this is the realm that I operated in for the past several years.

Data Project Management

My previous job was managing the AFL-CIO's membership data. The AFL-CIO is an association of labor unions, and each union has a list of their members. Our Analytics Team needed an accurate list of our members so that they can create various models and do polling based on that list.

This task required me to be a project manager working with client organizations' IT Departments to make sure they submitted their membership files in a timely manner. The AFL-CIO has millions of members, and I not only needed to understand what the data should look like, but have the hacking skills to be able to do the proper exploratory data analysis on it.

This was critical because if a client union made a mistake on their data such as dropping a third of their file, missing a field, or switching fields like first name and last name, that could negatively affect our program. Even though these mistakes were not necessarily my fault, I was responsible for them—the buck stopped with me. I needed to be on my toes, ready to catch these mistakes and propose a fix for the client.

It was also important to make this process as efficient as possible. Many clients wanted to deliver their file as it comes out of their database. I had to coach many of the clients along to submit the data in a standard format to streamline the process, which required building a relationship with each of these IT Departments.

They say that 80% of data science is cleaning data, and only 20% is analysis. Delivering our membership file to our Analytics Team was something I was passionate about because it allowed them to do the higher level analysis that they have the skillset for.

Descriptive Analytics

In addition to each clients' IT Department, I also developed relationships with the Political Directors of each organization. A key service the AFL-CIO provides is shared political resources for the labor movement, which is why Political Directors are important stakeholders.

Although some Political Directors may be interested to get in the weeds of model building, many wanted the meat and potatoes of their data. Part of the membership update process that I managed was adding political data fields to their membership.

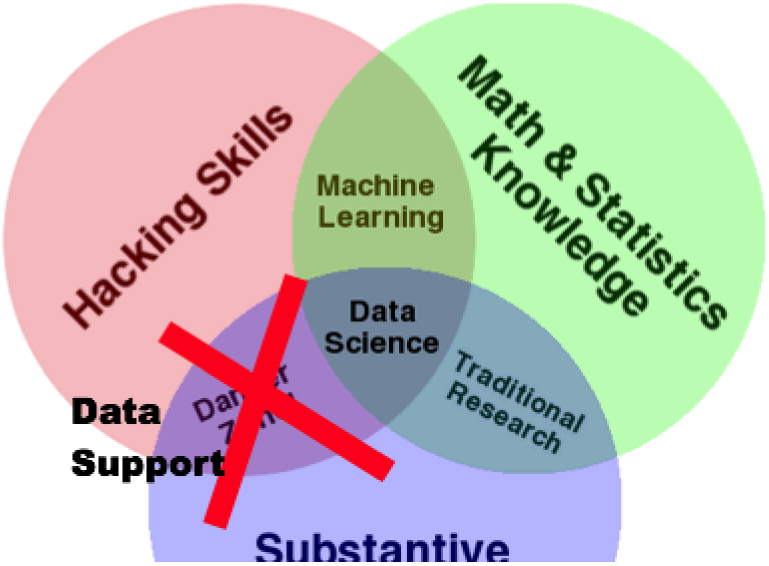
I developed reproducible reports that would be generated for each affiliate at the end of the membership update cycle that would tell Political Directors how many of their union's members were registered voters, which party they were registered to, which elections they voted in, and more.

None of these reports required math any more intense than the ability to count or generate a percentage. However, there was resounding appreciation for these reports because I used my domain knowledge and hacking skills to create a product that these key stakeholders wanted.

I am not saying that our Analytics Team could not have done the same work, but that work would have come at an opportunity cost. With me covering data acquisition, enhancement, and reporting of descriptive statistics to affiliates regarding their membership, our Analytics Team could focus more on analyzing member and voter data to help drive our organization's strategy.

Data Support: The Intersection of Hacking and Domain Knowledge

The fact of the matter is that any of the intersections taken to the extreme could be a danger zone. Someone who is at the intersection of domain knowledge and math skills, but cannot program will always be limited to the million record limit that Excel spreadsheets have to offer. If someone is at the intersection of hacking and math skills but does not have the domain knowledge of their industry, they may make a proposal that is good in theory, but is untenable due to the reality of their industry.



Data Support: The Intersection of Hacking Skills and Substantive Expertise

Even though I was not part of our Analytics Team, the above are examples of contributions I made to support their work. The reason why this relationship worked is that I did not try and play Data Scientist. I did not try and run a regression, neural network, or decision tree and use it in some way, shape, or form to influence our organization's strategy. However, I always admired their work, and for that reason am at General Assembly now to improve my math skills.

Analytics and Data Science Teams should seek out people who match this skillset. With standing on these people's shoulders, they can focus on the analysis to help their organization succeed.