Tiffany A. Timbers, Ph.D.

University of British Columbia, Dept. of Statistics, 2178 - 2207 Main Mall, Vancouver, BC Canada V6T 1Z4

tel. 604-803-4962
email. tiffany.timbers@stat.ubc.ca
website: tiffanytimbers.com
Github: github.com/ttimbers

# Statement of Vision for Data Science Education

When I draw on my own Data Science experiences (training, research and teaching), I have come to see three pillars as being critical to building a successful Data Science education program. These are *1)* teaching responsible use of Data Science methods, tools and workflows, *2)* keeping the Data Science curriculum current and modern, and *3)* building integrated content from the founding fields of Data Science: Statistics and Computer Science. There are certainly other aspects of Data Science education that are important beyond these pillars, but I will focus on these three because I believe they can serve as a solid foundation for building an exceptional Data Science program.

Data Science is a broad and diverse field, incorporating key concepts, methods, tools and workflows from Statistics, Computer Science, as well as the experimental and applied sciences. Thus, one challenge in building an education program for Data Science is to balance the depth and breadth of content to draw from each of these fields. From my professional experience as a Data Scientist and from my interactions with my network of Data Science peers, I believe the majority of Data Scientists' work involves the following workflow: *i)* getting and/or generating a dataset and research question, *ii)* identifying the appropriate method(s) and/or tool(s) for analysis and visualization of that dataset, *iii)* applying that method(s) and/or tool(s) to that dataset and *iv)* interpreting the analysis and visualization in the context of the research question being asked. Furthermore, in the majority of these cases, the method(s) and/or tool(s) used for analysis and visualization are ones which have been previously developed by others, and thus the emphasis of their work is in applying and not developing new methods (although, there are data scientists that do this). Thus, in building the most useful, effective and efficient Data Science curriculum, I believe we should focus on teaching just enough theory behind the most commonly used Data Science concepts, methods and tools to enable our Data Science students to use these responsibly. This means that they use these concepts, methods and tools in the right context, understand their assumptions and limitations, and correctly interpret the results with respect to the research question being asked.

Another key aspect to this pillar is an emphasis of responsible use of Data Science workflows. For professional transparency, rigour and believability of Data Science analysis and visualization, it is critical that the workflows used are reproducible. To me this means, given the same dataset and analysis/visualization method, one can easily and repeatedly generate the original results. Achieving this is not necessarily intuitive nor straightforward, and requires practice and discipline. Thus, I believe that incorporating Data Science workflows (*e.g.,* version control, scripting, package use and development, package and environment management, *etc*) as part of the core Data Science curriculum is also an essential part of a strong Data Science education program.

What concepts, methods, tools and workflows should we teach in a Data Science education program? This is an important and challenging question to answer and is what I regard as the second pillar of Data Science education. Data Science in one of the most rapidly evolving fields, a trend that I believe it will continue. Thus it is critical that Data Science educators put forward sufficient effort to keep the Data Science curriculum modern (*i.e.,* teaching what is commonly used by Data Scientists out in the wild) and communicate this aspect of the field to Data Science students while building their capacity for lifelong self-learning. From my experience, I have found the best way for Data Science educators to keep their fingers on the pulse of the field is to: *i)* network with both academic and industry Data Scientists via research collaborations (this would include UBC MDS Capstone projects), *ii)* read blog posts, preprints and papers, *iii)* participate in online discussion groups/forums (*e.g.,* Software Carpentry discuss list)

and *iv)* interact with the world-wide Data Science community by attending conferences and being active in social media (*e.g.*, Twitter). These strategies for Data Science educators to keep current should also be built in to the Data Science core curriculum and taught to our Data Science students. This will help them learn effective ways to keep themselves current regardless of what particular direction they go in after being trained.

Finally, what I view as the third pillar of Data Science education is building integrated content from the founding fields of Data Science: Statistics and Computer Science. Given that each of these two fields has contributed much to the foundation and core of Data Science, I believe that building an exceptional Data Science education program would be best done by a collaboration between statisticians and computer scientists; as opposed to siloing the program in one of these disciplines. Having access to support from both Statistics and Computer Science experts benefits both the Data Science educators creating and delivering the Data Science curriculum, as well as the Data Science students themselves. Exposure to both disciplines provides students access to interact and network with statisticians and computer scientists, as well as to attend and participate in events from both fields (*e.g.*, research talks). Additionally, with the development of a robust Data Science program associated with both of these fields, the Statistics and Computer Science departments may also benefit - new extracurricular program initiatives developed for (or by!) the Data Science students can be opened up to statisticians and computer scientists. An example of one which I would like to develop for the UBC MDS program is the implementation of a weekly Data Science journal/software package club, open to students, postdocs, staff and faculty, where we can work together to keep up to date on new Data Science concepts, methods, tools and workflows. Such a regular, open extracurricular event would also facilitate networking and collaboration between the UBC Statisticians, Computer Scientists and Data Scientists from all levels.

Thus, based on my experience in Data Science practice and education, it is in my opinion that the three pillars recommended above are an strong foundation for an exceptional Data Science education program. Furthermore, I believe that through my expertise in these three areas -- reproducible workflows, actively participating in the broader Data Science community, and interdisciplinary research and teaching -- I am strongly positioned to foster the development and implementation of a cutting-edge educational experience in Data Science.