Patterns

CellPress
OPEN ACCESS

People of Data

Bridging Domain and Data

Anne E. Carpenter[1,]*
[1]Broad Institute of Harvard and MIT, Cambridge, MA, USA
*Correspondence: anne@broadinstitute.org
https://doi.org/10.1016/j.patter.2020.100064

Dr. Anne Carpenter addresses her career path from cell biology toward computation. Why would a researcher move outside their comfort zone into a different field, from a domain into data science? What is the best way to bridge domain and data? What is challenging about moving from domain toward data? What is amazing about bridging domain and data?

I am the daughter of an engineer. The sister of an engineer. I aced math classes and assessments. But in undergraduate and graduate school, I studied biology because engineering seemed too dry, too constrained. And frankly, no one ever suggested it to me, a young girl growing up on a farm in Indiana, more interested in reading the classics than tinkering with electronics. I'm now listed on a top-100 list of AI Leaders in Drug Discovery & Healthcare, and I lead the new National Institutes of Health-funded Center for Open Bioimage Analysis (COBA) together with Kevin Eliceiri of University of Wisconsin-Madison. I was just inducted into the AIMBE, the American Institute for Medical and Biological Engineering, representing the top 2% of engineers in the United States. What happened? And is it a great tragedy that I was steered away from engineering in my early years, or a blessing in disguise?

Biology is messy. Not just in its physical methods—dissection, pulverizing living things into bits, working with chemicals. It's also messy in its manner of approaching problems, and I mean that in a good way. Biology is about making connections between disparate pieces of data, about hypothesizing mechanisms when only a small fraction of a system is known. Like all sciences, it is a logic puzzle, but it's a puzzle where defining which puzzle to solve is a major part of the work. Where it is quite likely there are many possible solutions. Where some of the input data are unknown, some are unknown by you (unless you have encyclopedic knowledge of the literature), and some of what is ‘‘known’’ is wrong. On top of all that, the data you have are almost certainly insufficient to solve the problem conclusively.

While earning my PhD in cell biology, my hidden engineer emerged: after observing cell samples by eye for hours on the microscope, I had an intense desire to automate and quantify biology for my project. It became clear I was more enamored with *how* to answer a question rather than learning the answer. So I dove headfirst into coding in Pascal, then spent my postdoc almost entirely in MATLAB rather than at the bench. When I launched my own laboratory at the Broad Institute in 2007, I had no microscopes, no incubators, no pipettes. I had gone full computational. This year, the software project I started, CellProfiler, will hit its 10,000th citation, and aside from accelerating the research of thousands of biologists, my laboratory's work has contributed to at least five clinical trials. For me, it's been a fruitful transition from domain to data.

What is the best way to bridge domain and data? I'm in no position to definitively answer that—probably no one is. I can, however, report my experience. I wrote CellProfiler in collaboration with Thouis Jones, a graduate student in computer science at MIT. What was so productive about this interaction is that each of us had the time and motivation to learn, say, 20% of the others' field. I've come to believe that it's impossible for one person to be sufficiently immersed in both biology and data science to be at the cutting edge of both, much less remain there over time. Being a deep and up-to-date expert in two fields is too much for one human brain, so one solution is to go with two brains. This requires a particular personality for both brains in such a partnership: mutual respect for the other's field, mutual interest in solving a particular problem, willingness to acquire crossover knowledge in the other field, and an ability to communicate across the boundary, serving in essence as a translator who can distill, from a vast body of knowledge, the most relevant and fruitful information for each other.

What is challenging about moving from domain toward data? For starters, it takes significant time to devote to studying the other field—not just to learn its content but also its culture and practices. I was lucky enough to transition without grant-funded productivity constraints; I was funded by a postdoctoral fellowship that could support skills development. Crossing any disciplinary boundary also requires that a person be comfortable being a non-expert. At the beginning, you can avoid impostor syndrome by not pretending to be an expert! Eventually, though, you must come to terms with being perceived as an expert in a field outside your comfort zone. For me, it helped that during my transition from biology to computer science, I was not committed to an academic career and was not trying to impress anyone. I tried to be as transparent as possible about my expertise (and lack thereof) when interacting as others. It's a delicate balance, particularly as a woman, to present the appropriate level of confidence without being too apologetic. For me personally, perhaps the strongest force counteracting impostor syndrome was my profound sense of identity and self-worth stemming from my religious faith, which gave me a sense of value and purpose apart from my career choice or my relative standing in an intellectual pecking order. Now that I firmly exist in both worlds of domain and data, there remain challenges. A computational biologist writing grants and being assessed is often not enough of a biologist for the biologists (No wet lab!) and not enough of a computer scientist for the computer scientists (No formal training in it! Too application-focused!) It can also be very challenging to keep open-source domain-specific software projects funded. At the moment, CellProfiler and our newer deep-learning-oriented developments are supported by the NIH and the Chan Zuckerberg Initiative, both of which have expressed interest in ongoing support (rather than just initial development) of broadly impactful software.

What is amazing about bridging domain and data? With the success of public data challenges and hackathons in biomedicine where outsiders produce top solutions, one might think that domain expertise is no longer needed. However, only a small fraction of biomedical problems can be constrained and neatly organized in this way. It takes significant effort and time to translate a messy biological state into a well-delineated computationally solvable problem and often more effort and time to interpret the results of an analysis than to run it. Field-bilingual scientists with a strong grasp of a domain and expertise in computational sciences are critical. Studies show that a diversity of thinking styles, and working at the intersection of fields, leads to great advances. Early in my career, it was clear that images stored tremendous information about cell state. My expertise as a cell biologist led to the development of Cell Painting, a new wet-lab assay that dramatically increased the information content and became the standard in this new computational field of image-based profiling. While my lab still works on bioimage analysis software, a major focus is now translating messy bottlenecks in the drug discovery process into solvable data problems. Another huge strength of domain experts moving into computational fields is their devotion to necessary but thankless tasks such as creating domain-specific, user-friendly software and gathering data into usable public databases. While uninteresting from the perspective of algorithmic novelty, these efforts drive practical applications in a field and have tremendous impact.

Overall, despite the challenges both personally and professionally, it has been a blessing that I dove deeply into the biology domain before switching over to data. It is now a great joy to mentor others making the switch.

About the Author

Dr. Anne Carpenter is an institute scientist at the Broad Institute of Harvard and MIT. Her research group develops algorithms and strategies for large-scale experiments involving images. The team's open-source CellProfiler software is used by thousands of biologists worldwide. Carpenter is a pioneer in image-based profiling, the extraction of rich, unbiased information from images for a number of important applications in drug discovery and functional genomics. Carpenter earned a PhD in cell biology from the University of Illinois, Urbana-Champaign, and completed her postdoctoral fellowship at the Whitehead Institute for Biomedical Research and MIT's Computer Sciences/Artificial Intelligence Laboratory (CSAIL).