When I was in fourth grade, I asked my father to help me solve a fraction problem. Being Russian, he was not content with just giving me the answer, but he instead gave the knowledge which produced the answer. Frustrated, I left, and figured the problem out on my own. But I later realized that he was trying to instill an appreciation for knowledge over a simple answer, or data.

I used to think that data was the same as knowledge. That knowing in DNA, A binds T and G binds C meant you were smart. But there is deeper information behind this pattern, the *how* of AT and GC binding, such as that A and T both can form two hydrogen bonds, and G and C both form three. So A can't bind with either G or C, and vice versa. But even deeper, what does the implication of the three bonds in GC versus two bonds in AT have for biological function or evolutionary conservation? Knowledge is not *what,* the exact answer, or even *how* it was obtained, but the *why?* Integrating with other data, adding context, plus much deep thinking, is what produces knowledge.

In high school, we visited a cadaver lab and the technician described a cadaver that was completely missing part of a leg muscle. Immediately, I thought to myself that this should be easily detectable through a genetic test. If only we had the data of the person's genome sequence, we would have the knowledge of the presence or absence of this person's muscle. Later, I learned that the genome sequence alone doesn't tell you whether a gene is on or off. It is only data. Valuable data, yes, but only one step. If we have that person's genome sequence plus a "reference" human genome sequence, then we could pinpoint what is different in this person's genome. But we would be a far cry from knowledge. We would still need to know where the muscle genes are, how leg-specific genes are turned on during embryogenesis, and how the muscle development process can be disrupted.

The summer after my freshman year at MIT, I began to understand how research creates an intermediate step between data and knowledge: information. In my first research experience in Dr. Martha Bulyk's lab, I performed protein binding experiments where I pipetted transcription factor proteins that bind DNA onto DNA microarrays, and ran a graduate student's code which estimated where on the DNA the protein bound. While the experimental side was fascinating, I was most excited by running code which converted the data of a polka-dotted microarray slide to information: the DNA letters bound by this protein. I was hooked. I knew had to learn more about algorithms because I wanted to contribute to biology in this quantitative way. But being new to science, I thought that information was the same as knowledge. I thought that now that we knew what the binding sites were, we had the "answer" and could move on. But in the paper on which I was a co-author, obtaining the DNA letters bound by the protein was the first step. More deeply, within this group of transcription factors called homeodomains, the binding patterns were conserved across many species, from *Mus musculus* mouse to *Drosophila* flies, the patterns of which were computationally predicted and later experimentally verified. The finding that these transcription factors, which turn on genes that are critical for embryonic development, are not only very similar across species in their DNA sequence, but also in the DNA sequence they bind, shows how the tree of life is much more interconnected than we previously thought.

During my next research experience in Dr. Sean Eddy's lab, I began to understand how the nuances of data generation affects the final knowledge. I had the privilege of working with Dr. Eddy after being admitted to the competitive Janelia Farm Summer Scholars program. One of Dr. Eddy's main projects is HMMER, software that performs searches of biological sequences using homology, or evolutionary relatedness, which helps elucidate how genomes can change over time. These homology searches are performed using Hidden Markov Models (HMMs), a sequence of states, where each state encodes a pattern of the protein, e.g. a functional domain such as a DNA-binding site. The homology of two sequences is evaluated by first aligning each sequence to an HMM, then comparing the distance between the two HMMs to their distance from the null model, an HMM of a generic protein. This generic protein does not exist biologically, but I created it *in silico* using HMMs, which [created] a more robust null model for HMMER homology searches. Through this project, I learned that it is easy to overestimate the similarity of two proteins if there is no reference model for comparison. It may be easy to say "these proteins are very similar," but the more important and more difficult question is "similar, compared to *what?*" If the similarity of the two proteins is less than the similarity of each protein to the generic protein, then they aren't truly similar. I learned that the way in which the data was created may be biased, and it is important to question the assumptions of the model used in generating the data, such that you are convinced that the signal is truly greater than the noise.

The research experience where I first tied data to knowledge through information, was with Dr. Jill Mesirov in the cancer research group at the Broad Institute, the largest genomics research center in the United States. There, I implemented REVEALER, an algorithm which uncovers new associations of cancer pathway expression with genomic alterations. This algorithm fills a new niche, because there are many cases where an cancer-causing (oncogenic) pathway is overly active, but the canonical oncogene isn't affected in any way. REVEALER finds other activators (or repressors) of pathway activity, and in particular, 37% (9/24) of the samples we studied were resistant to the repression of the oncogene *KRAS* but did not have the *KRAS* mutation. However, REVEALER found that in these *KRAS* mutation-free samples, chromosome 8 had a regional amplification (amplicon), which seemed to confer resistance to *KRAS* repression. The *KRAS* gene is on chromosome 12, far away from this amplicon, but this amplicon contains *Myc*, another known oncogene, which may biologically confer the resistance. We verified this result by adding the chromosome 8 amplicon to non-cancerous cells, and they were resistant! This work paves the way for new cancer drugs for the many *KRAS*-pathway cancers such as lung and colon, but the many cases where the patient does not harbor the *KRAS* mutation. There was already data that non-KRAS mutated samples were difficult to treat with the usual KRAS-inhibiting drugs, and the REVEALER algorithm took this data and mined it to find other potential causes of resistance, or information. But only once the experiment of introducing the amplicon to other cells indeed conferred resistance, did we have knowledge.

Interested in new ways to obtain biological data and create new knowledge, I began performing single-cell research at Univ. Calif.-Santa Cruz (UCSC). Working with Prof. Nader Pourmand, we were interested in breast cancer drug resistance at the single-cell level, to observe how individual cells escape chemotherapy. Specifically, the drug paclitaxel inhibits microtubule elongation, preventing proper mitotic spindle formation and subsequently cells from dividing. We performed RNA-Seq on individual cells extracted from untreated (6 cells), treated with paclitaxel (6 cells), and survivor (5 cells) MDA-MB-231 breast cancer cell line populations. One of the main technical challenges was consistently analyzing the aligned sequencing reads, as reproducibility is a major issue in the life sciences. To increase consistency between sequencing experiment interpretations, I developed an open-source RNA-Seq differential expression (RSDE) pipeline. Using RSDE to study taxol resistance in breast cancer at the single-cell level, we found the surviving cells had differentially regulated actin cytoskeleton genes, which may indicate these survivors have some mechanisms to compensate for improper chromosome alignment to the mitotic plate. In addition to performing this research, I was the first to finish UCSC's 2-year MS in bioinformatics, in 9 months.

Research ties data to knowledge through information. But there is no end to knowledge. Eventually, knowledge may become as simple as data, that A binds T and G binds C. But when Watson and Crick first noticed the pattern of the proportions of A exactly matching the proportions of T, this data of occurrences had many steps to go, before becoming the knowledge of molecular interactions. I hope to figure out how to achieve knowledge, from information, from data through the course of my PhD.

The invention of the microscope heralded a new era for biology. Suddenly, we could *see* the components of living things. Robert Hooke first observed cork tree bark, and he thought it looked like the “walled compartments a monk would live in,” hence the term, *cell.* Today, we have new “microscopes” which separate single cells and measure their DNA and RNA, creating never-before-seen data. Unlike previous research which measured molecules over thousands of cells, and assumed these came from a non-existent “average cell”, now we are confident our measurements are occurring in exactly the same cell, allowing for more robust analyses of biological phenomena, and ultimately leading to new biological knowledge.

To create knowledge from data of motor neuron differentiation in Prof. Gene Yeo’s lab, we study how individual cells vary in gene expression and alternative messenger RNA (mRNA) splicing. The method of mRNA splicing cuts introns out of mRNA transcribed from DNA, and leaves only exons (the “ex”-pressed parts of the gene). Notably, over 90% of genes undergo alternative splicing, and it is most prevalent in the human brain. As previous experiments measured splicing over population averages, it is unknown how the different versions are shared (or not shared) among cells of a single population. If the bulk measurement shows 60% of version A and 40% of version B, does that mean that each cell has one version or another? Or does each cell have 60% of one version, and 40% of the other? To study heterogeneity in alternative splicing, I developed a linear algebra method of categorizing splicing events to be consistent or multi-modal, e.g. bimodal. By this method, I found that regardless of developmental stage, most cells have either one version or another; and if there is a mixture of molecules A and B, it is usually not consistently shared amongst all cells. Additionally I also observed splicing events are under different levels of biological control, e.g. a splicing events may have low or high population variance. The exons in the low variance events are more evolutionarily conserved than the high variance events, which implies that these events may be important to motor neuron lineage specification. However, the regulatory region around the high variance exons is more conserved than the low variance events, which may explain the high variance of these events in the first place, as they are under tight, specific control by the cell. The principles we are developing in the motor neuron differentiation system are a template for the many exciting future applications of single-cell genomics: early cancer detection, kitchen-top DNA sequencers to test for food safety, and discovering new bacterial species from dirt. Single-cell is the future of genomics, and I will be leading the charge.

Intent on working with single-cell genomic data, I chose to attend University of California, San Diego (UCSD) because of the . UCSD has a unique mix of people working on the technology for extracting molecules from single cells (Roger Lasken), algorithms to assemble the sequences (Pavel Pevzner), and analysis pipelines to deal with the noisy and heterogeneous data (Gene Yeo). In particular, my advisor Prof. Gene Yeo’s lab was the first to prototype the Fluidigm C1, a microfluidic machine which efficiently and reproducibly separates single cells. The Yeo lab also has the machine learning expertise to tease apart the subtle differences between individual cells, and the biological ability to draw biological conclusions from the data. In addition, I hope to work with Dr. Roger Lasken from the J. Craig Venter Institute here in San Diego, who developed a method for simultaneous extraction of both RNA and DNA, allowing for simultaneous assembly of single-cell genomes and transcriptomes, a completely novel application as almost all current genomics data assumes all cells have exactly the same genome, even though we know this is not true. I hope to extend the single-cell bacterial genome assemblers developed by Prof. Pavel Pevzner's group, to single-cell human genome and transcriptome assembly. Joining forces within the collaborative environment of UCSD is key, because ultimately, my goal is discovery of new, disease-relevant, biological knowledge, which the Yeo lab excels in. But the integration of the technology developed by Lasken, and algorithms by Pevzner, catalyzes our ability to uncover biological breakthroughs.

In the quest to obtain knowledge from data, one common step is visualization. However, most scientific visualization software leaves much to be desired, and may even distort interpretation of the data. I truly believe that scientific progress is impeded when improper data visualizations are used. Improving data visualizations has two main advantages. First, the researcher can understand their work better and faster, enabling rapid prototyping of hypotheses. Second, the researcher can spend more time discussing the science instead of losing the audience with a confusing figure. There is established research in data visualizations through color perception studies by Cynthia Brewer, and Edward Tufte's work in simplifying graphs to their essence: data. However, the standards for data visualization in the scientific world are lacking. Data visualization is an afterthought for many researchers, but I'm working on this forgotten part and attempting to automate reasonable plotting defaults for myself and others.

To help establish reasonable standards for scientific visualization, I developed a Python package, "prettyplotlib," which builds on Python's existing plotting library, "matplotlib," to painlessly create clean default figures using principles by Brewer and Tufte. Importantly, most users are not interested in creating the most beautiful plot of their data -- they are focused on getting the data visualized, and do care about design. This package keeps the focus on the data, while making it easy to create clear figures. Users find prettyplotlib ~~easy to~~ to easily display their results, without the painstaking process of retouching every tiny aspect of the plot. Since prettyplotlib has been released, it has had thousands of views from around the world, and has been starred by more than 300 people on GitHub, a code-sharing website. Users have commented on its ease and practicality of use: "finally a wrapper for matplotlib with sensible defaults -- great plots without the pain," and "FINALLY. prettyplotlib might just make matplotlib usable (because by default, it's not even close)" Throughout graduate school, I will continue advocating for sound data visualizations and developing `prettyplotlib`.

While data visualization is an important step in approaching knowledge from data, it is easy to generate hundreds of visualizations and be under the impression that these figures showing relationships between data are information. It is difficult to make the leap from information to knowledge, but the process of research creates knowledge from data, through information. I want my career to be an exploration of turning information into knowledge, and creating new data to extract information, and thus knowledge. But beyond even knowledge is understanding, and since I don't know how to get there yet, I'm pursuing a PhD. I intend to become a professor at a top institution, where I will be surrounded by leaders in their field who will question, support and challenge my data and information, so that together we may create knowledge and understanding.