What is Information?

1 The Experience

The Research Science Institute (RSI) is a 6-week summer program held at the Massachusetts Institute of Technology that pairs students with top mentors in their fields to engage in cutting-edge research. Approximately 75 students attend RSI each year, and each of them is given the opportunity to pursue a research project; needless to say, nearly everyone produces work far beyond the level of high school and often even undergraduate students. What RSI also provides is an opportunity to meet other high school scientists from all over the world, including Singapore, Lebanon, Saudi Arabia, and Turkey. Not only is there a strong research component, but the social component is essential to RSI as well. The chance to meet some of the people who will change the world in the next few decades is invaluable. I was privileged to attend RSI during the summer of 2006.

Prior to the program, my mentor Professor Lizhong Zheng from the EECS department at MIT asked me to study information theory, a graduate course he teaches. It is a mathematical approach to the theory of communications. The field incorporates several areas of mathematics, but it is primarily based in probability theory and statistics related to random variables. At first, I was surprised at what I was going to be working on; after all, my application had specified computer science and theoretical mathematics as my two main fields, yet from what I could find about information theory it was a blend of electrical engineering and applied mathematics. After I began studying it, however, any doubts in my mind faded away as I immersed myself into the topic.

I quickly realized that the mathematics required for information theory was not trivial in any sense. I had to learn a lot to even begin to understand the field. When I start learning about a subject in math, most of the time I am flooded by notation that makes the text virtually impossible to comprehend. With information theory, this was no different. Fortunately, with the aid of a textbook, I was able to acquire the knowledge in a logical order and get a solid grasp on the meaning of the notation. After this, I moved on to specific mathematical topics necessary to understand my field. These ranged from probability mass functions to Lagrange multipliers, both of which were crucial for my research. Having always been a big fan of learning more about math, I found this to be an appealing and valuable aspect of my research experience. By the time I wrote my research paper, I was much more fluent in the "language of information theory" and could work with a variety of new math concepts.

The next hurdle was to figure out what information theory was really about. Since it is traditionally taught as an undergraduate or graduate level course, I dedicated the several weeks I had before RSI started to learning everything I could about information theory. This was already far too little time to digest an entire field, but you would be surprised at how much you can cover when you are focused on a single subject. Though this was a painstaking task, as self-studying almost anything is, it was very rewarding as well; by the time I departed on my plane to MIT I felt much more comfortable and capable of accomplishing anything significant. I learned that having sufficient background information on a project before starting to work on it is crucial, or else you will probably end up wasting a lot of time, as I was able to avoid for the most part.

As my research timeframe at RSI was effectively four weeks, each day had to be spent meticulously. For the most part, I worked in MIT's computer labs since a significant portion of my research was collecting data to support a conjecture. I learned how to use the Maple math software to my advantage, and it ended up being my primary tool. I was able to combine my math and programming skills with my newfound knowledge in information theory to write a program to do calculations; it provided compelling evidence that the conjecture was indeed true. Using Maple I also solved several special cases of the conjecture and found that these were already so complex that anything more general had to be done using a different method. The rest of the time I worked with experimental graphs and read many papers online relating to the conjecture, which I was eventually able to come up with a proof for about a month after RSI had ended.

I found that it was truly fulfilling to do real research, but I also realized that my experience was somewhat "artificial." My research was fine, but the idea that projects can generally be finished in a month is absurd. True research is a long-term commitment, and you should be prepared to dedicate a lot of time and work to something you really enjoy doing. However, I still greatly value my time at RSI as a brief exposure to the world of research.

Now that you have read about my story, here are some parting thoughts and pieces of advice before moving to the more technical stuff. First off, when starting a research project, I strongly suggest finding a mentor. He or she can be a professor at a university or perhaps a professional at a local company, as long as you can get an opinion from someone who knows what he or she is talking about. In particular, you do not want to get yourself into a situation in which the problem is too simple or too difficult, which is more than likely to happen if you go look for something on your own. In my experience at RSI, the amount that mentors contributed to the students' projects varied greatly. Some students met their mentors every day and were directed more strictly in their work, while others had mentors who gave them a project and the freedom to explore. It is ultimately up to you how much independence you want, but my advice is to treat your mentor as a source of wisdom rather than as a boss.

Lastly, do not be discouraged if you end up with little or no progress. The end product is always nice to see, but research is about the experience itself, not the paper you write when you are done. Go ahead and submit a project to the Intel Science Talent Search or the Siemens Competition, but do not treat the result as a judgment of your project. After all, you did something no one else has ever done. Be proud of your accomplishment and take what you have learned with you in your future endeavors.

2 The Research

Information theory uses mathematics as a way of quantifying how data can be stored and communicated most efficiently. There are two broad types of data compression: lossless and lossy. As the names might suggest, lossless data compression deals with cases in which perfect recovery of the data is necessary while lossy data compression considers the possibility that the data could be different after decompression. Examples of the former include ZIP or GIF files, while the latter includes MP3 or JPEG files.

In lossy data compression, there is always a method of measuring the difference between the original data and the recovered data. The amount that they differ by is called the distortion. This can be measured in various ways depending on the situation; for instance, the distortion of a digital image may be measured better by how well a human can tell the difference rather than the actual loss of bits. By assigning a quantity to the vague notion of distortion, a lot of concrete results can be achieved.

Before moving on, entropy must be introduced. This might be familiar to you from a variety of physical sciences, and it has more or less the same meaning here. Entropy measures the amount of disorder in bits - in this context, entropy measures how much information a random variable contains. The more random a variable, the more entropy it has, and the more information it holds. This may seem counterintuitive at first, but a random variable that takes the same value all the time does not really tell you anything when you receive it, which is why it has zero entropy. Claude Shannon was the first one to introduce this concept in communications and a way of mathematically measuring it, which is essentially what information theory is built on.

Communication is done across channels, which are probability distributions that take in a message to be transmitted and output a message to be received. An example of a channel is an ethernet wire. Transmissions across channels are a central part of information theory as a whole, and there are a variety of different problems that arise from it. They often relate to the rate, a measure of the information entropy per variable transmitted over a channel.

My project was in rate distortion theory, a subfield of information theory that considers the minimal rate necessary to achieve a certain distortion. A lower rate is better because higher rates are more costly. For example, using rate distortion theory we can determine the smallest amount of information necessary to transmit a digital image or music file so that any differences will not be perceived by humans. This is generally represented by a function

$$R(D) = \text{minimum rate (in bits) necessary to have the distortion be } \leq D.$$

For simple cases, we can graph this; naturally, it is a decreasing function since allowing more distortion means needing less rate. An example of such a graph is shown here.

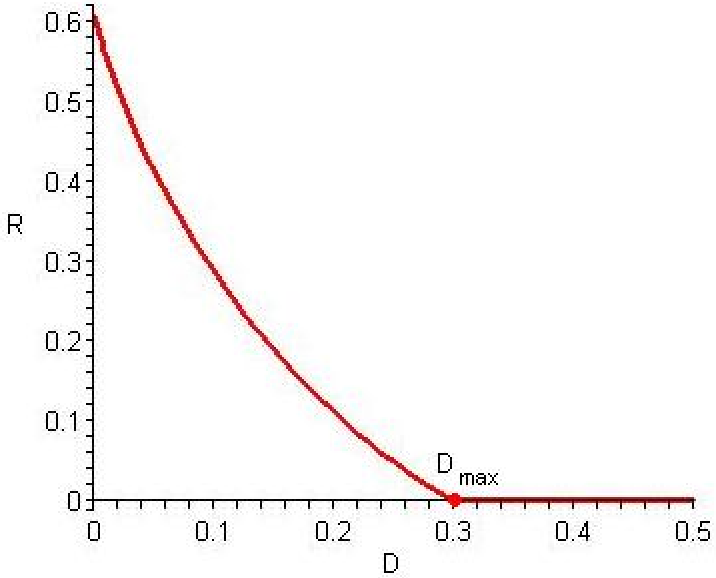

Figure 1: An example of a rate distortion function.

Notice that, when the allowed distortion gets large enough, no information is needed to reconstruct the data; any random receieved message would be good enough. This point is called $D_{\max}$, as labeled in the diagram. The main goal of my project was to calculate the slope of the rate distortion function at $D_{\max}$. The slope of the function has many applications in compression and is generally hard to determine, why is why I only looked at a special case. As I mentioned earlier, I found numerical evidence for a conjecture regarding the calculation of the slope at this point and eventually a proof of the conjecture as well.

You might have noticed that a lot of background information was necessary to even introduce my problem. This is a common theme in research today; since most of the broad problems are either too difficult or have already been solved, scientists are moving into more and more specific areas to conduct research. When starting on a project, be prepared to take a lot of time to learn and read literature in the field so you can delve into specialized problems that are still open for research.

My research experience has influenced me significantly, and I will surely continue working on new projects through college. I hope that by reading this article you have a better understanding of what research is really about and are also enthusiastic about pursuing it.