CS 480 **Computational Theory**

Specified Complexity

Excerpt from a draft of an article *by David Johnson*

[David Johnson teaches Mechanical Engineering at BYU-Idaho.]

Information theory treats information as a statistical quantity, based on probabilities. Information can be thought of as a ruling out of contingencies. If I were to tell you that in flipping a fair coin once, it landed a heads or a tails, I haven't given you much information, beyond the fact that I flipped a coin. If I told you that the toss landed heads, then I have conveyed more information. If it's a fair coin, the probability of landing heads, rather than tails, is 50%, or 0.5. The lower the probability of a given event, the more information is conveyed in specifying that the event occurred; or alternatively, the more complex the information is.

So, the complexity of information contained in a string of characters is based on its length. By randomly flipping a coin once, I will obtain a heads or a tails. What if, instead, I flipped the coin 100 times. Let's say I got the following sequence of ones and zeros, corresponding to heads and tails, respectively.

0100000101010010001101111011010001000010000000111011101110100101100111101111010001011110111001000

By stating that I got this particular sequence, rather than any other of the 2^{100} sequences I could have gotten, I convey much more information. The information is more complex. But, it's still just gibberish. Even though the information content of the event is complex, it's still just a random sequence of ones and zeros. But consider the following sequence.

0100011011000001010011100101110111000000010010001101000101011001111000100110101011110011011110111100

What if I claimed that I got this sequence by randomly tossing a coin? Would you believe me?

It probably depends on your experience. For instance, if you are a computer scientist, you might note that, although at first glance it looks random, this particular set of ones and zeros is in fact a sequence of binary numbers, starting with the single digit numbers (0, 1), then the double digits (00, 01, 10, 11), then the triple digits (000, 001, 010, 011, 100, 101, 110, 111), and so on until 100 characters are reached.

So, now would you believe me if I said I got that sequence by chance? Probably not. But why not? What makes the first sequence a believable random sampling, and the second one not? The key is that the second sequence is specifiable independent of the event itself. It means something in a different and detachable context. If scientists at the Search for Extraterrestrial Intelligence (SETI) were to find this sequence embedded in the radio data they were analyzing, they would be forced to conclude that someone was trying to communicate with them. And the longer the sequence is, while still meaning something specifiable, the stronger the evidence.

Take another example. If I randomly picked three letters from a hat, and got the sequence T-H-E, that is specifiable as a word in English. It is different from randomly picking the sequence Q-Z-F in that, while the probabilities of picking these two sequences are identical (i.e. the complexity of the information in both is the same), "qzf" is not a word in any language. The latter sequence exemplifies information *simpliciter*, while the former exemplifies *specified* information.

But still, the information content in the word "the" is not very complex. On the other hand, if I claimed to have randomly picked the sequence F-O-U-R-S-C-O-R-E-A-N-D-S-E-V-E-N-Y-E-A-R-S-A-G-O-O-U-R-F-A-T-H-E-R-S-B-R-O-U-G-H-T-F-O-R-T-H and on for several paragraphs, you would be wise to suspect that I was fibbing. Not because the information content is complex (any long string of random letters exhibits complex information), but because the string of letters exhibits specified complexity.

Once information gets complex enough, like the several paragraphs of text alluded to above, then the set of all strings of that length that mean something specifiable is exceedingly small compared to the set of all possible strings of that length. To the point that if you hit that very small target, then you can bet you didn't let your arrow fly at random.

But how much information are we talking about? How long does the string have to be before we can say that it couldn't have happened by chance?

Let's say you want to duplicate the Gettysburg address by picking one random letter after another in a long sequence. Eventually, you could do it. Large number theory states that in an infinite random string, any finite string is not only guaranteed to occur, but to occur infinitely many times. But from a probability standpoint, how long would it take you to exactly duplicate the Gettysburg address? That depends on how fast you can pick letters.

Alternatively, how long does a string need to be such that in all likelihood, it could never be duplicated in the history of our universe? There is a limit. The universe has only been around for a few billion years, and there are only so many atoms in it. If all the atoms in the universe could be used as little computers to generate random strings of a given length at clock speeds based on the Planck time (the smallest unit of time), a specified event of probability 10^{-150} (that's $0.000\ldots1$ with 149 zeros between the decimal point and the 1) could not be reached in all the time the universe has been around. So, a probability of 10^{-150} can be considered a universal probability bound. Or, since this corresponds roughly to an informational content of 500 bits, then 500 bits can be thought of as a universal complexity bound. From a probability standpoint, it remains unlikely that an event of specified complexity of length 500 bits could have happened by chance given all the computational resources of our universe.

For comparison, a string of DNA would only have to have a length of 250 base pairs to reach this complexity. The human genome consists of about 3 billion base pairs, and even the simplest bacteria have genomes on the order of 500,000 base pairs!