

Big Idea (Week 3): Lossless Compression

Not all characters, words, sounds, or images, carry the same amount of *information*, and we can use this fact to reduce the number of bits required to store or transmit a data stream without discarding any information. That is, the things we perceive directly—characters, words, sounds, images—can typically be compressed considerably without loss such that they can be decompressed to represent the original exactly.

The “trick” here is simple in concept. Not all things {characters, word, notes, figures} occur with equal frequency (*e.g.* In this class we will say “compute” or “digital” much more often than we will say “sparrow” or “peloton”; you will find far more e’s on this page than z’s http://en.wikipedia.org/wiki/Letter_frequency). Since we encounter some things more frequently than others, we can give them smaller encodings (fewer bits). Today’s text message lexicon employs one form of this concept by using abbreviations for common phrases (*e.g.*, LOL, CUL8R, BFF, B4).

This embodies an important engineering principle: make the common case inexpensive {fast, small, low energy}.

By carefully defining the set of “things” that can occur (a universe of possibilities) and assuming (or observing) a particular frequency of occurrence, we can make this notion mathematically precise—allowing us to both quantify the encoding size and even ask questions about the optimal (smallest possible) encoding—that is, the true *information content* in the sequence of things.

Consider recording the daily weather for a year. For the sake of simplicity, let us assume we can describe the weather as one of four cases {sunny, cloudy, rain, snow}. Let w_i be the weather on day i that takes on one of these four values. The number of bits it would take us to record this for a year would be:

$$total_bits = \sum_{d=0}^{d=365} (|bits[w_i]|) \quad (1)$$

where $|bits[w_i]|$ is the count of the number of bits in the encoding of $bits[w_i]$. If we encoded each of the four cases with 2 bits (*e.g.* bits[sunny]=00, bits[cloudy]=01, bits[rain]=10, bits[snow]=11), then one year of weather observations would require 730 bits. Now, if we knew that, on average, only 5% of the days had snow, 45% were sunny, 10% were rain, and 40% were cloudy, then we might select an encoding: bits[sunny]=0, b[cloudy]=10, b[rain]=110, b[snow]=111. This encoding would require around 621 bits (about 164 sunny days, 146 cloudy days, 36 rain days, and 19 snow days).

While we can be precise about encoding when we define a universe of things, this still leaves open the question of the best set of things to consider (*e.g.* characters, words, phrases, notes, intervals).