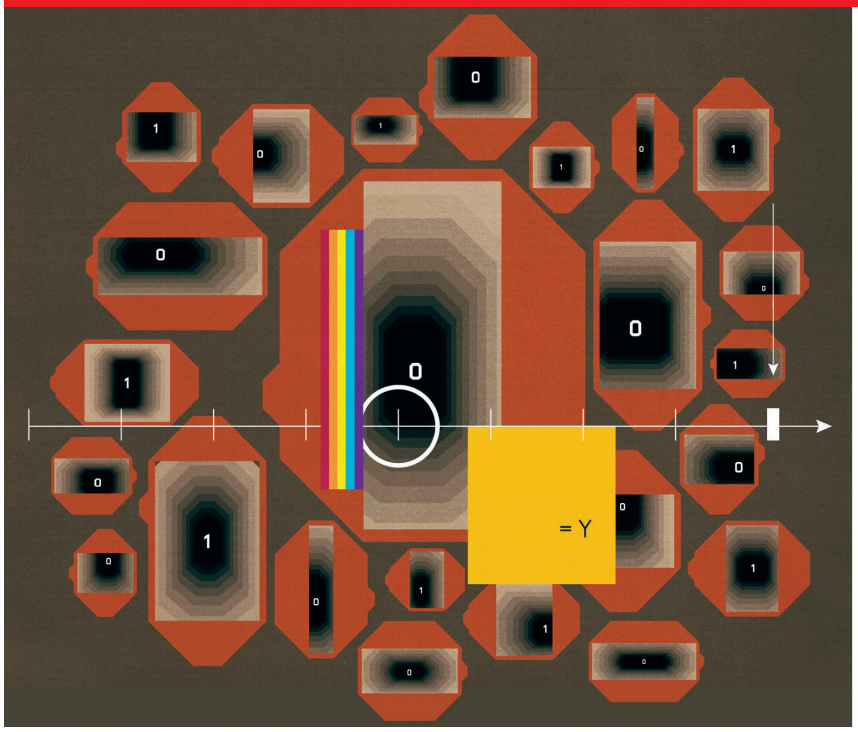

DATA WORLD: RACING TOWARD YOTTA

ONCE UPON A TIME, information was deposited only inside human brains, and ancient bards could spend hours retelling stories of conflicts and conquests. Then external data storage was invented. • Small clay cylinders and tablets, invented in Sumer some 5,000 years ago, often contained just a dozen cuneiform characters, equivalent to a few hundred bytes (10^2 B). The *Oresteia*, a trilogy of Greek tragedies by Aeschylus (fifth century BCE), amounts to about 300,000 B (10^5 B). Some rich senators in imperial Rome had libraries housing hundreds of scrolls, with one large collection holding at least 10^8 B (100 megabytes). • A radical shift came with Johannes Guttenberg's printing press, using movable type. By 1500, less than half a century after printing's introduction, European printers had released more than 11,000 new book editions. The extraordinary rise of printing was joined by other forms of stored information. First came engraved and woodcut music scores, illustrations, and maps. Then, in the 19th century, came photographs, sound recordings, and movies. Reference works and other regularly published statistical compendia during the 20th century came on the backs of new storage modes, notably magnetic tapes and long-playing records. • Beginning in the 1960s, computers expanded the scope of digitization to medical imaging (a digital mammogram is 50 MB), animated movies (2-3 gigabytes), intercontinental financial transfers, and eventually the mass emailing of spam (more than 100 million messages sent every minute). Such digitally stored information rapidly surpassed all printed materials. Shakespeare's plays and poems amount to 5 MB, the equivalent of just a single high-resolution photograph, or of 30 seconds of high-fidelity sound, or of eight seconds of streamed high-definition video.

Printed materials have thus been reduced to a marginal component of overall global information storage. By the year 2000, all books in the Library of Congress were on the order of 10^{13} B (more than 10 terabytes) but that was less than 1 percent of the total collection (10^{15} B, about 3 petabytes) once all photographs, maps, movie, and audio recordings were added.

And in the 21st century this information is being generated ever faster. In its latest survey of data generated per minute in 2018, Domo, a cloud service, listed more than 97,000 hours of video streamed by Netflix users, nearly 4.5 million videos watched on YouTube, just over 18 million forecast requests on the Weather Channel, and more than 3 quadrillion bytes (3.1 petabytes) of other Internet data used in the United States alone. By 2016, the annual global data-creation rate surpassed 16 ZB (a zettabyte is 10^{21} B), and by 2025, it is expected to rise by another order of magnitude—that is, to about 160 ZB or 10^{23} B. And according to Domo, by 2020 1.7 MB of data will be generated every second for every one of the world's nearly 8 billion people.

These quantities lead to some obvious questions. Only a fraction of the data flood could be stored, but which part should that be? Challenges of storage are obvious even if less than 1 percent of this flow gets preserved. And for whatever we decide to store, the next question is how long should the data be preserved. No storage need last forever, but what is the optimal span?

The highest prefix in the international system of units is yotta, Y=10^{24}. We'll have that many bytes within a decade. And once we start creating more than 50 trillion bytes of information per person per year, will there be any real chance of making effective use of it? It is easier to find new prefixes for large databases than to decide how large is large enough. After all, there are fundamental differences between accumulated data, useful information, and insightful knowledge. ■

↗ **POST YOUR COMMENTS** at https://spectrum.ieee.org/yotta0719

ILLUSTRATION BY Stuart Bradford