CHAPTER 4

THE DIGITIZATION OF JUST ABOUT EVERYTHING

“When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.”

—Lord Kelvin

"HEY, HAVE YOU HEARD about . . . ?"

"You've got to check out . . . "

Questions and recommendations like these are the stuff of everyday life. They're how we learn about new things from our friends, family, and colleagues, and how we spread the word about exciting things we've come across. Traditionally, such cool hunting ended with the name of a band, restaurant, place to visit, TV show, book, or movie.

In the digital age, sentences like these frequently end with the name of a website or a gadget. And right now, they're often about a smartphone application. Both of the major technology platforms in this market—Apple's iOS and Google's Android—have more than five hundred thousand applications available.[1] There are plenty of "Top 10" and "Best of" lists available to help users find the cream of the smartphone app crop, but traditional word of mouth has retained its power.

Not long ago Matt Beane, a doctoral student at the MIT Sloan School of Management and a member of our Digital Frontier team, gave us a tip. "You've got to check out Waze; it's amazing." But when we found out it was a GPS-based app that provided driving directions, we weren't immediately impressed. Our cars have navigation systems and our iPhones can give driving directions through the Maps application. We could not see a need for yet another how-do-I-get-there technology.

As Matt patiently explained, using Waze is like bringing a Ducati to a drag race against an oxcart. Unlike traditional GPS navigation, Waze doesn't tell you what route to your destination is best in general; it tells you what route is best *right now*. As the company website explains:

> The idea for Waze originated years ago, when Ehud Shabtai . . . was given a PDA with an external GPS device pre-installed with navigation software. Ehud's initial excitement quickly gave way to disappointment—the product didn't reflect the dynamic changes that characterize real conditions on the road. . . .
>
> Ehud took matters into his own hands. . . . His goal? To accurately reflect the road system, state of traffic and all the information relevant to drivers at any given moment.[2]

Anyone who has used a traditional GPS system will recognize Shabtai's frustration. Yes, they know your precise location thanks to a network of twenty-four geosynchronous GPS satellites built and maintained by the U.S. government. They also know about roads—which ones are highways, one-way streets, and so on—because they have access to a database with this information. But that's about it. The things a driver really wants to know about—traffic jams, accidents, road closures, and other factors that affect travel time—escape a traditional system. When asked, for example, to calculate the best route from Andy's house to Erik's, it simply takes the starting point (Andy's car's current location) and the ending point (Erik's house) and consults its road database to calculate the theoretically "quickest" route between the two. This route will include major roads and highways, since they have the highest speed limits.

If it's rush hour, however, this theoretically quickest route will not actually be the quickest one; with thousands of cars squeezing onto the major roads and highways, traffic speed will not approach, let alone eclipse, the speed limit. Andy should instead seek out all the sneaky little back roads that longtime commuters know about. Andy's GPS knows that these roads exist (if it's up-to-date, it knows about *all* roads), but doesn't know that they're the best option at eight forty-five on a Tuesday morning. Even if he starts out on back roads, his well-meaning GPS will keep rerouting him onto the highway.

Shabtai recognized that a truly useful GPS system needed to know more than where the car was on the road. It also needed to know where *other* cars were and how fast they were moving. When the first smartphones appeared he saw an opportunity, founding Waze in 2008 along with Uri Levine and Amir Shinar. The software's genius is to turn all the smartphones running it into sensors that upload constantly to the company's servers their location and speed information. As more and more smartphones run the application, therefore, Waze gets a more and more complete sense of how traffic is flowing throughout a given area. Instead of just a static map of roads, it also has always current updates on traffic conditions. Its servers use the map, these updates, and a set of sophisticated algorithms to generate driving directions. If Andy wants to drive to Erik's at 8:45 a.m. on a Tuesday, Waze is not going to put him on the highway. It's going to keep him on surface streets where traffic is comparatively light at that hour.

That Waze gets more useful to all of its members as it gets more members is a classic example of what economists call a *network effect*—a situation where the value of a resource for each of its users increases with each additional user. And the number of Wazers, as they're called, is increasing quickly. In July of 2012 the company reported that it had doubled its user base to twenty million people in the previous six months.[3] This community had collectively driven more than 3.2 billion miles and had typed in many thousands of updates about accidents, sudden traffic jams, police speed traps, road closings, new freeway exits and entrances, cheap gas, and other items of interest to their fellow drivers.

Waze makes GPS what it should be for drivers: a system for getting where you want to go as quickly and easily as possible, regardless of how much you know about local roads and conditions. It instantly turns you into the most knowledgeable driver in town.

The Economics of Bits

Waze is possible in no small part because of Moore's Law and exponential technological progress, the subjects of the previous chapter. The service relies on vast numbers of powerful but cheap devices (the smartphones of its users), each of them equipped with an array of processors, sensors, and transmitters. Such technology simply didn't exist a decade ago, and so neither did Waze. It only became feasible in the past few years because of accumulated digital power increases and cost declines. As we saw in chapter 3, exponential improvement in computer gear is one of the three fundamental forces enabling the second machine age.

Waze also depends critically on the second of these three forces: digitization. In their landmark 1998 book *Information Rules*, economists Carl Shapiro and Hal Varian define this phenomenon as "encod[ing information] as a stream of bits."[4] Digitization, in other words, is the work of turning all kinds of information and media—text, sounds, photos, video, data from instruments and sensors, and so on—into the ones and zeroes that are the native language of computers and their kin. Waze, for example, uses several streams of information: digitized street maps, location coordinates for cars broadcast by the app, and alerts about traffic jams, among others. It's Waze's ability to bring these streams together and make them useful for its users that causes the service to be so popular.

We thought we understood digitization pretty well based on the work of Shapiro, Varian, and others, and based on our almost constant exposure to online content, but in the past few years the phenomenon has evolved in some unexpected directions. It has also exploded in volume, velocity, and variety. This surge in digitization has had two profound consequences: new ways of acquiring knowledge (in other words, of doing science) and higher rates of innovation. This chapter will explore the fascinating recent history of digitization.

Like so many other modern online services, Waze exploits two of the well-understood and unique economic properties of digital information: such information is *non-rival*, and it has *close to zero marginal cost of reproduction*. In everyday language, we might say that digital information is not "used up" when it gets used, and it is extremely cheap to make another copy of a digitized resource. Let's look at each of these properties in a bit more detail.

Rival goods, which we encounter every day, can only be consumed by one person or thing at a time. If the two of us fly from Boston to California, the plane that takes off after us cannot use our fuel. Andy can't also have the seat that Erik is sitting in (airline rules prohibit such sharing, even if we were up for it) and can't use his colleague's headphones if Erik has already put them on to listen to music on his smartphone. The digitized music itself, however, is non-rival. Erik's listening to it doesn't keep anyone else from doing so, at the same time or later.

If Andy buys and reads an old hardcover copy of the collected works of science-fiction writer Jules Verne, he doesn't "use it up"; he can pass it on to Erik once he's done. But if the two of us want to dip into *Twenty Thousand Leagues Under the Sea* at the same time, we either have to find another copy or Andy has to make a copy of the book he owns. He might be legally entitled to do this because it's not under copyright, but he'd still have to spend a lot of time at the photocopier or pay someone else to do so. In either case, making that copy would not be cheap.[5] In addition, a photocopy of a photocopy of a photocopy starts to get hard to read.

But if Andy has acquired a digital copy of the book, with a couple keystrokes or mouse clicks he can create a duplicate, save it to a physical disk, and give the copy to Erik. Unlike photocopies, bits cloned from bits are usually exactly identical to the original. Copying bits is also extremely cheap, fast, and easy to do. While the very first copy of a book or movie might cost a lot to create, making additional copies cost almost nothing. This is what is meant by "zero marginal cost of reproduction."

These days, of course, instead of handing Erik a disk, Andy is more likely to attach the file to an e-mail message or share it through a cloud service like Dropbox. One way or another, though, he's going to use the Internet. He'll take this approach because it's faster, more convenient, and, in an important sense, essentially free. Like most people, we pay a flat fee for Internet access at home and on our mobile devices (MIT pays for our access at work). If we exceed a certain data limit, our Internet Service Provider might start charging us extra, but until that point we don't pay by the bit; we pay the same no matter how many bits we upload or download. As such, there's no additional cost for sending or receiving one more chunk of data over the Net. Unlike goods made of atoms, goods made of bits can be replicated perfectly and sent across the room or across the planet almost instantaneously and almost costlessly. Making things free, perfect, and instant might seem like unreasonable expectations for most products, but as more information is digitized, more products will fall into these categories.

Business Models When the First Copy is Still Expensive

Shapiro and Varian elegantly summarize these attributes by stating that in an age of computers and networks, "Information is costly to produce but cheap to reproduce."[6] Instantaneous online translation services, one of the science-fiction-into-reality technologies discussed in chapter 2, take advantage of this fact. They make use of paired sets of documents that were translated, often at considerable expense, by a human from one language into another. For example, the European Union and its predecessor bodies have since 1957 issued all official documents in all the main languages of its member countries, and the United Nations has been similarly prolific in writing texts in all six of its official languages.

This huge body of information was not cheap to generate, but once it's digitized it's very cheap to replicate, chop up, and share widely and repeatedly. This is exactly what a service like Google Translate does. When it gets an English sentence and a request for its German equivalent, it essentially scans all the documents it knows about in both English and German, looking for a close match (or a few fragments that add up to a close match), then returns the corresponding German text. Today's most advanced automatic translation services, then, are not the result of any recent insight about how to teach computers all the rules of human languages and how to apply them. Instead, they're applications that do statistical pattern matching over huge pools of digital content that was costly to produce, but cheap to reproduce.

What Happens When the Content Comes Freely?

But what would happen to the digital world if information were no longer costly to produce? What would happen if it were free right from the start? We've been learning the answers to these questions in the years since *Information Rules* came out, and they're highly encouraging.

The old business saying is that "time is money," but what's amazing about the modern Internet is how many people are willing to devote their time to producing online content without seeking any money in return. Wikipedia's content, for example, is generated for free by volunteers all around the world. It's by far the world's largest and most consulted reference work, but no one gets paid to write or edit its articles. The same is true for countless websites, blogs, discussion boards, forums, and other sources of online information. Their creators expect no direct monetary reward and offer the information free of charge.

When Shapiro and Varian published *Information Rules* in 1998, the rise of such user-generated content, much of which is created without money changing hands, had yet to occur.

Blogger, one of the first weblog services, debuted in August 1999, Wikipedia in January 2001, and Friendster, an early social networking site, in 2002. Friendster was soon eclipsed by Facebook, which was founded in 2004 and has since grown into the most popular Internet site in the world.[7] In fact, six of the ten most popular content sites throughout the world are primarily user-generated, as are six of the top ten in the United States.[8]

All this user-generated content isn't just making us feel good by letting us express ourselves and communicate with one another; it's also contributing to some of the recent science-fiction-into-reality technologies we've seen. Siri, for example, improves itself over time by analyzing the ever-larger collection of sound files its users generate when interacting with the voice recognition system. And Watson's database, which consisted of approximately two hundred million pages of documents taking up four terabytes of disk space, included an entire copy of Wikipedia.[9] For a while it also included the salty language–filled Urban Dictionary, but this archive of user-generated content was removed after, to the dismay of its creators, Watson started to include curse words in its responses.[10]

Perhaps we shouldn't be too surprised by the growth and popularity of user-generated content on the Internet. After all, we humans like to share and interact. What's a bit more surprising is how much our machines also apparently like talking to each other.

Machine-to-machine (M2M) communication is a catch-all term for devices sharing data with one another over networks like the Internet. Waze makes use of M2M; when the app is active on a smartphone, it constantly sends information to Waze's servers without any human involvement. Similarly, when you search the popular travel site Kayak for cheap airfares, Kayak's servers immediately send requests to their counterparts at various airlines, which write back in real time without any human involvement. ATMs ask their banks how much money we have in our accounts before letting us withdraw cash; digital thermometers in refrigerated trucks constantly reassure supermarkets that the produce isn't getting too hot in transit; sensors in semiconductor factories let headquarters know every time a defect occurs; and countless other M2M communications take place in real time, all the time. According to a July 2012 story in the *New York Times*, "The combined level of robotic chatter on the world's wireless networks . . . is likely soon to exceed that generated by the sum of all human voice conversations taking place on wireless grids."[11]

Running Out of Metric System: The Data Explosion

The digitization of just about everything—documents, news, music, photos, video, maps, personal updates, social networks, requests for information and responses to those requests, data from all kinds of sensors, and so on—is one of the most important phenomena of recent years. As we move deeper into the second machine age, digitization continues to spread and accelerate, yielding some jaw-dropping statistics. According to Cisco Systems, worldwide Internet traffic increased by a factor of twelve in just the five years between 2006 and 2011, reaching 23.9 exabytes per month.[12]

An *exabyte* is a ridiculously big number, the equivalent of more than two hundred thousand of Watson's entire database.

However, even this is not enough to capture the magnitude of current and future digitization. Technology research firm IDC estimates that there were 2.7 zettabytes, or 2.7 sextillion bytes, of digital data in the world in 2012, almost half as much again as existed in 2011. And this data won't just sit on disk drives; it'll also move around. Cisco predicts that global Internet Protocol traffic will reach 1.3 zettabytes by 2016.[13] That's over 250 billion DVDs of information.[14]

As these figures make clear, digitization yields truly big data. In fact, if this kind of growth keeps up for much longer we're going to run out of metric system. When its set of prefixes was expanded in 1991 at the nineteenth General Conference on Weights and Measures, the largest one was *yotta*, signifying one septillion, or 10^{24}.[15] We're only one prefix away from that in the 'zettabyte era.'

Binary Science

The recent explosion of digitization is clearly impressive, but is it important? Are all of these exa- and zettabytes of digital data actually useful?

They're incredibly useful. One of the main reasons we cite digitization as a main force shaping the second machine age is that digitization increases understanding. It does this by making huge amounts of data readily accessible, and data are the lifeblood of science. By "science" here, we mean the work of formulating theories and hypotheses, then evaluating them. Or, less formally, guessing how something works, then checking to see if the guess is right.

A while back Erik guessed that data about Internet searches might signal future changes in housing sales and prices around the country. He reasoned that if a couple is going to move to another city and buy a house, they are not going to complete the process in just a few days. They're going to start investigating the move and purchase months in advance. These days those initial investigations will take place over the Internet and consist of typing into a search engine phrases like "Phoenix real estate agent," "Phoenix neighborhoods," and "Phoenix two-bedroom house prices."

To test this hypothesis, Erik asked Google if he could access data about its search terms. He was told that he didn't have to ask; the company made these data freely available over the Web. Erik and his doctoral student Lynn Wu, neither of whom was versed in the economics of housing, built a simple statistical model to look at the data utilizing the user-generated content of search terms made available by Google. Their model linked changes in search-term volume to later housing sales and price changes, predicting that if search terms like the ones above were on the increase today, then housing sales and prices in Phoenix would rise three months from now. They found their simple model worked. In fact, it predicted sales 23.6 percent more accurately than predictions published by the experts at the National Association of Realtors.

Researchers have had similar success using newly available digital data in other domains. A team led by Rumi Chunara of Harvard Medical School found that tweets were just as accurate as official reports when it came to tracking the spread of cholera after the 2010 earthquake in Haiti; they were also at least two weeks faster.[16] Sitaram Asur and Bernardo Huberman of HP's Social Computing Lab found that tweets could also be used to predict movie box-office revenue. They concluded that "this work shows how social media expresses a collective wisdom which, when properly tapped, can yield an extremely powerful and accurate indicator of future outcomes."[17]

Digitization can also help us better understand the past. As of March 2012 Google had scanned more than twenty million books published over several centuries.[18] This huge pool of digital words and phrases forms a base for what's being called *culturomics*, or "the application of high-throughput data collection and analysis to the study of human culture."[19] A multidisciplinary team led by Jean-Baptiste Michel and Erez Lieberman Aiden analyzed over five million books published in English since 1800. Among other things, they found that the number of words in English increased by more than 70 percent between 1950 and 2000, that fame now comes to people more quickly than in the past but also fades faster, and that in the twentieth century interest in evolution was declining until Watson and Crick discovered the structure of DNA.[20]

All of these are examples of better understanding and prediction—in other words, of better science—via digitization. Hal Varian, who's now Google's chief economist, has for years enjoyed a front-row seat for this phenomenon. He also has a way with words. One of our favorite quotes of his is, "I keep saying that the sexy job in the next ten years will be statisticians. And I'm not kidding."[21] When we look at the amount of digital data being created and think about how much more insight there is to be gained, we're pretty sure he's not wrong, either.

New Layers Yield New Recipes

Digital information isn't just the lifeblood for new kinds of science; it's the second fundamental force (after exponential improvement) shaping the second machine age because of its role in fostering innovation. Waze is a great example here. The service is built on multiple layers and generations of digitization, none of which have decayed or been used up since digital goods are non-rival.

The first and oldest layer is digital maps, which are at least as old as personal computers.[22] The second is GPS location information, which became much more useful for driving when the U.S. government increased its GPS accuracy in 2000.[23] The third is social data; Waze users help each other by providing information on everything from accidents to police speed traps to cheap gas; they can even use the app to chat with one another. And finally, Waze makes extensive use of sensor data; in fact, it essentially converts every car using it into a traffic-speed sensor and uses these data to calculate the quickest routes.

In-car navigation systems that use only the first two generations of digital data—maps and GPS location information—have been around for a while. They can be extremely useful, especially in unfamiliar cities, but as we've seen, they have serious shortcomings. The founders of Waze realized that as digitization advanced and spread they could overcome the shortcomings of traditional GPS navigation. These innovators made progress by adding social and sensor data to an existing system, greatly increasing its power and usefulness. As we'll see in the next chapter, this style of innovation is one of the hallmarks of our current time. It's so important, in fact, that it's the third and last of the forces shaping the second machine age. The next chapter explains why this is.