

Set 3. Big Data

Skill 3.1: Identify sources of Big Data

Skill 3.2: Explain how parallel processing can be leveraged to process large data sets

Skill 3.3: Describe how the Internet has enabled Crowdsourcing

Skill 3.4: Explain how Citizen Science enables people to contribute to scientific research

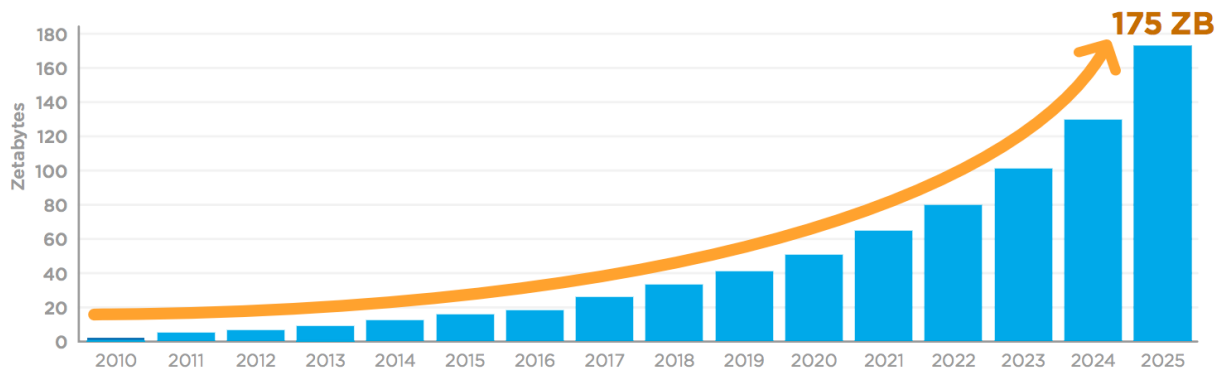
Skill 3.5: Explain the benefits of Open Data

Skill 3.1: Identify sources of Big Data

Skill 3.1 Concepts

The digital world is constantly collecting more and more data. Whenever you use an online service, you're contributing to a data set of user behavior. Even by simply using electricity and water in your house, you're contributing to a data set of utilities usage.

With the increasing number of people and cities connected to the Internet, data sets are increasingly larger in size. One report estimates that the total size of digital data will be **175 zettabytes** in 2025.



How much data is 175 zettabytes, anyway? A single zettabyte is a trillion gigabytes. A modern smartphone stores about 32 gigabytes. To store 175 zettabytes, we would need 6 trillion smartphones (1000 smartphones for every living person!).

Whew, that's a lot! But how big are the individual data sets?

These stats can give us an idea...

- A single MRI scan results in **20,000 images**.
- Google processes **3.5 billion search queries** per day.
- Instagram users post **54,000 photos** each minute.
- An autonomous vehicle generates **11 terabytes of data** each day.
- Twitter users post **3,000 tweets** every second.

Big data sets are so large that our traditional ways of storing and processing them are no longer adequate, presenting challenges to computer scientists and data engineers. On the plus side, they're also so large that they offer new opportunities for analysis that were impossible on a small data set.

In this lesson, we'll explore where big data comes from and the exciting ways that we can use it.

Skill 3.1 Exercise 1

Skill 3.2: Explain how parallel processing can be leveraged to process large data sets

Skill 3.2 Concepts

When a computing system needs to store massive amounts of data, there are two primary considerations: space and time. Or, more specifically:

- How will the data be stored?
- How can the data be processed efficiently?

Storage

In 2020, a standard laptop might have a 256 GB hard drive. That could fit:

- 840,000 tweets (280 characters, username, timestamp)
- 96,000 photos (compressed JPEGs)
- 66,418 songs (compressed MP3s)
- 224 movies (compressed MP4s)

For the average user, 256 gigabytes is quite a lot. But for a company operating at a global scale, it's barely anything. Twitter users post 500 million tweets a day, and many of those tweets include photos. They would need more than **500** of those 256 GB hard drives to store the data for a single day of usage.

Dozens of hard drives can be connected together using a **disk array** or disk enclosure.

Processing

A large data set can take a long time to process, regardless of whether the data set can fit on a single hard drive.

Let's imagine that engineers at X (formally Twitter) want to determine how many tweets contain a particular hashtag (e.g. "#ClimateCrisis").

The code to determine whether a single tweet contains the hashtag requires only a tenth of a millisecond, or 0.0001 seconds. The code to analyze 500 million tweets (the amount posted each day) would require this much time:

$$0.01 \times 500,000,000 = 50,000 \text{ seconds} = 13.4 \text{ hours}$$

It would take half a day just to process a day's worth of tweets!

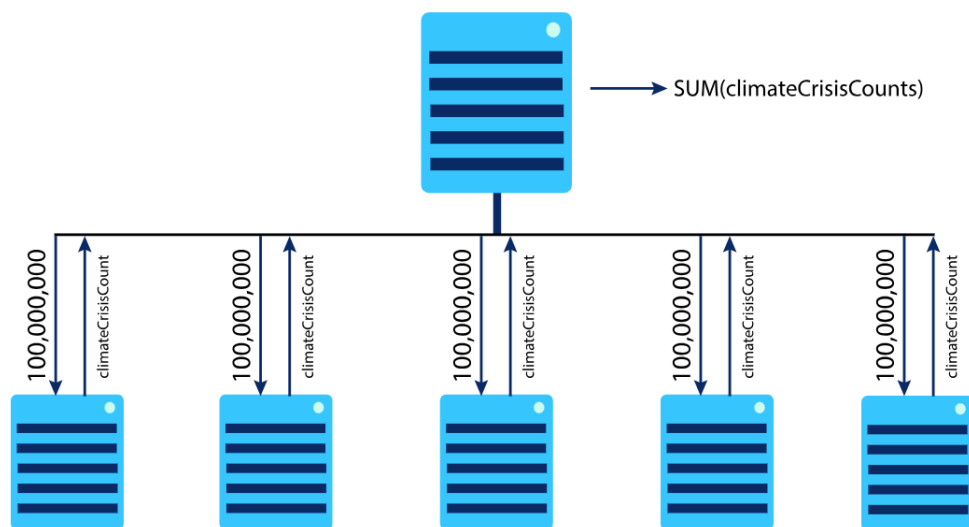
The engineers have two options at this point:

1. Come up with a faster per-tweet algorithm
2. Use parallel computing to process the data in parallel

The engineers can probably figure out some ways to improve the efficiency of the hashtag check, but even if they managed to reduce the time by a factor of 10, it would still take an hour and a half to analyze a day's worth of tweets. If they hope to analyze more than that (like a month of tweets, a year of tweets, or all tweets ever), they will need to use **parallel computing**.

Parallel computing makes processing more efficient by distributing the work across several computers. For example,

five machines could each process 100 million tweets and send back a count of how many tweets contained "#ClimateCrisis" to a central machine. Once that machine received the count from each of the five machines, it could sum them up and report the total count.



Skill 3.2 Exercises 1 & 2

Skill 3.3: Describe how the Internet has enabled Crowdsourcing

Skill 2.3 Concepts

Crowdsourcing is a way to take advantage of the large network of potential contributors online and funnel their resources into an output. Crowdsourcing is not a new idea, however. For example, in 1857, a group of British intellectuals decided to compile a new English dictionary with a comprehensive set of words, definitions, and usages. They put out a call for volunteers to submit words from books in their libraries and eventually received more than *2 million* word references.

The creators of that dictionary had an important insight: there's no one person that knows every bit of information, but when we combine knowledge from many people, we can develop impressively comprehensive collections of knowledge.

Thanks to computers and the Internet, creating a shared knowledge base is much easier these days than it was in the 1800s. When someone contributes information, the computer can store it in a database and make it easy to sort, search, and edit. The community of contributors can sort through the database to keep the highest quality information, and computer programmers can help by adding reputation systems, voting algorithms, and spam detection.

Wikipedia is a great modern example of a crowd-sourced knowledge base. You've probably run into a Wikipedia article if you've ever searched for knowledge online. With nearly a billion edits since its inception and over 36 million registered users, Wikipedia is collecting the wisdom of a very large crowd

Skill 3.3 Exercises 1 & 2

Skill 3.4: Explain how Citizen Science enables people to contribute to scientific research

Skill 3.4 Concepts

Modern scientists deal with problems at a global scale, but a scientist or research lab may not have the resources to study the vast amount of observational data available.

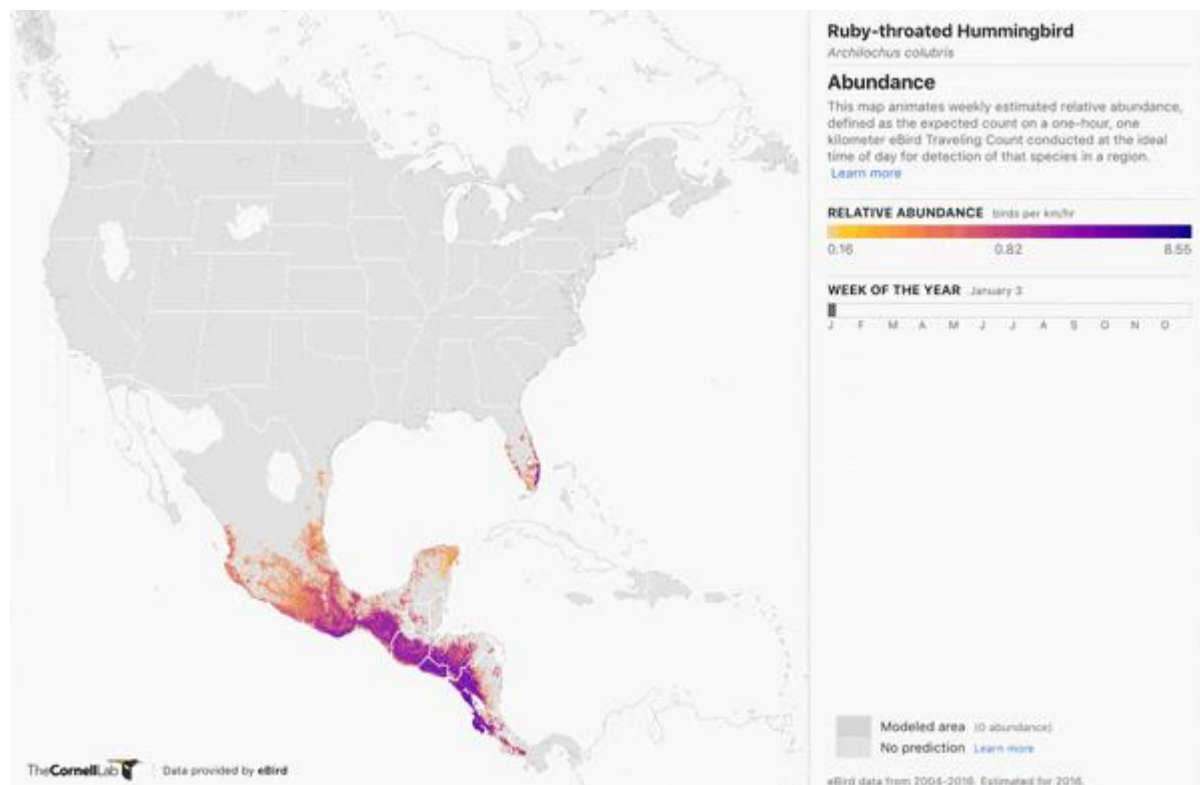
In that case, scientists can embrace **citizen science**: the participation of the general public in scientific research.

Participatory monitoring

Many citizen science projects involve monitoring of plants, animals, and atmospheric conditions. People around the world can report what is happening near them to a central database, and scientists can develop a better idea of worldwide trends in ecology and the environment.

[eBird](#) is a citizen science project that encourages birdwatching enthusiasts to submit photos of birds near them. During their [Global Big Day event](#) on May 4, 2019, over 30,000 people from 171 countries reported 1.85 million bird sightings!

The eBird research team from Cornell uses the bird sightings data for trends and statistics, like mapping the abundance of a species over time. Here's an [abundance simulation](#) for the Ruby-Throated Hummingbird:




All of the eBird data is openly available, so any researcher, conservationist, or birder can use the data for their own purposes. More than [200 research publications](#) have cited the eBird data. You could even use the data yourself, for a project about your favorite bird or local region.

Data classification












Another type of citizen science project asks the public to help in classifying vast amounts of collected data, like photos and sounds. Many of those projects are hosted on [Zooniverse](#), a platform with dozens of citizen science projects and more than a million participants.

For the [Wild Gabon project](#), volunteers identify animals in the Lopé National Park based on photos taken by motion-triggered cameras.

Below is an example,



M T2 80°F 26°C 05-17-2017 16:47:31

TASK		TUTORIAL	
	Buffalo		Bushbuck
	Red river hog		Elephant
	Leopard		No animal
	Gorilla		Other
	Chimpanzee		Human
	Yellow-backed duiker		

Showing 11 of 11 [Clear filters](#)

Done & Talk

Done

Not all photos have animals that are clearly identifiable as that gorilla, however, so volunteers might mis-classify the animals in a photo. To compensate for that, the platform gives the same photo to multiple volunteers and comes up with a final classification based on the consensus.

Citizen science projects are successful because they harness the effort of huge numbers of volunteers. The average person may not have the training and expertise of a scientist, but if enough people have the time and energy, they can make large contributions to scientific research.

Skill 3.4 Exercises 1 & 2

Skill 3.5: Explain the benefits of Open Data

Skill 3.5 Concepts

Open data is data which is openly accessible to all, including companies, citizens, the media, and consumers.

Watch the video below to learn more about open data,



Another source of very important digital information is scientists. They can use computers to do calculations, databases to store results, and websites to publicize their research findings.

Traditionally, scientists share their research findings by submitting a paper to an academic journal, hoping for its acceptance after a peer review, and presenting the published paper at conferences.

Meanwhile, journals realizing the potential for revenue generation typically only allow access to the full paper if you're a subscriber or if you pay a fee. Here's a typical payoff on a journal website:

Choose an option to locate/access this article:

Check if you have access through your login credentials or your institution.

Check Access

or

Purchase PDF \$35.95

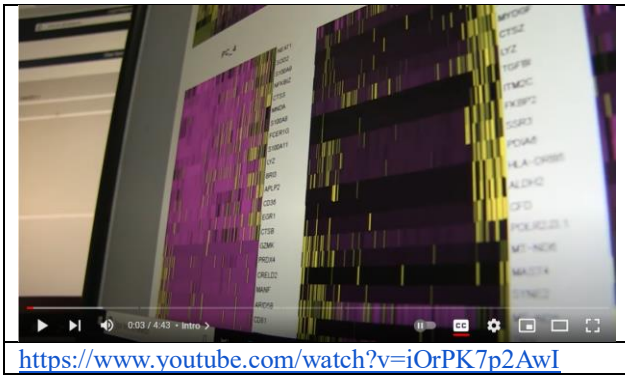
Paywall from Elsevier when attempting to access an article.

Restricting scientific findings to people who are members of paying institutions or to those who can afford the subscription themselves leads to an inequity in the availability of scientific information.

Thanks to the Internet and the open access movement, the results of scientific research are becoming increasingly available for anyone to learn from and build upon.

The goal of the **open access** movement is to remove barriers to scientific information, by encouraging journals to freely distribute full research papers online and encouraging authors to archive their papers in an open access repository or personal website.

Watch the video below to learn more about this movement,



Skill 3.5 Exercise 1