

Roboflow

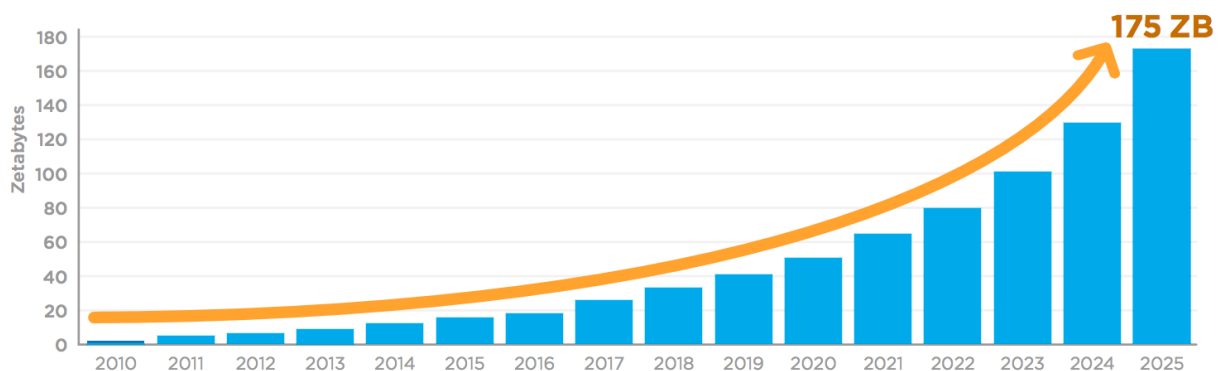
Your Tasks (Mark these off as you go)

- ☐ Explore Big Data
- ☐ Explore Crowdsourced Data
- ☐ Explore Open Data
- ☐ Define key vocabulary
- ☐ Receive credit for this lab guide

□ Explore Big Data

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**.¹¹
- Google processes **3.5 billion search queries** per day.²²
- Instagram users post **54,000 photos** each minute.³³
- An autonomous vehicle generates **11 terabytes of data** each day.⁴⁴
- Twitter users post **3,000 tweets** every second.⁵⁵

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.

Watch the videos below about Big Data.



<https://www.youtube.com/watch?v=1XGo8K1boH4>
(2:30 minutes)



<https://www.youtube.com/watch?v=bMrDHtGHFR4>
(6 minutes)

Provide an example of big data

Video 1: An example cited in the first video, is that in order to recommend a video the program needs to look at ALL the online data - billions of data points

Video 2: The data is a database of DNA of all known organisms

Indicate how the data was filtered or cleaned to help solve a problem

Video 1: It then needs to filter videos that are similar to those you just looked at. This is done with parallel processing. Parallel processing enables the computer to “look” at many videos at the same time.

Video 2: They filtered all the human sequences there were present in the DNA sample. They then searched for all the non-human DNA in the sample in the database.

How is the data visualized or presented

Video 1: The visualized data appears as recommended videos to the end user

Video 2: It appears they visualized the data as a spread sheet






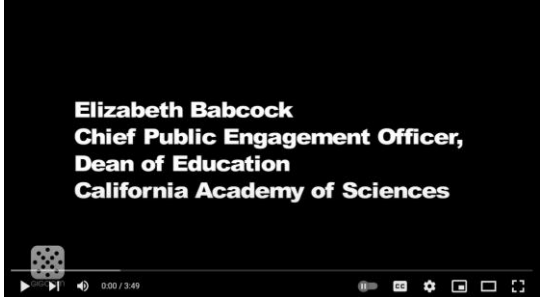
How is the new information gleaned from the data used to solve a problem or make decisions

Video 1: The computer now knows what kinds of videos you’ve been searching for. This information is sold so commercials tailored to you now appear.

Video 2: How your DNA code is associated with disease. They found out the organism that the patient encountered 9 months prior while in Puerto Rico.

□ Explore Crowdsourced Data

Listen the NPR news snippet on How Pokemon Inspired A Citizen Science Project. Then watch the video on What is Citizen Science? As you listen to the news snippet and watch the video, keep notes on how the data analysis process can be applied to Citizen Science.

<p>ENVIRONMENT</p> <h3>< How Pokemon Inspired A Citizen Science Project To Monitor Tiny Streams</h3> <p>April 20, 2018 · 5:06 AM ET</p> <div> 3-Minute Listen</div> <div>   </div> <p>(3:00 minutes)</p>	 <p>(3:00 minutes)</p>
--	--

<https://www.npr.org/2018/04/20/597972310>

<https://youtu.be/81hhecl0p5k>

Collect or Choose Data

Audio Clip: The data is information about streams and rivers provided by citizens

Video: The data is information collected by citizens

Clean and/or Filter Data

Audio Clip: The researchers are interested in small streams not visible by satellites

Video: It depends on the project – for example galaxy zoo filters for stars, where as bird sightings filter for birds

Visualize and Find Patterns

Audio Clip: Data should enable the researchers to map the streams more accurately

Video: It depends, but having the geolocation of sightings enables for spatial visualizations

New Information

Audio Clip: Better water forecasts

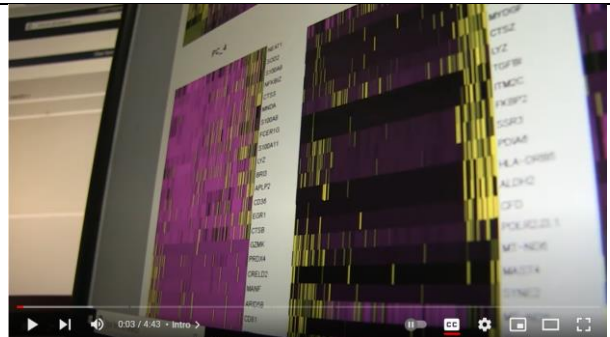
Video: The effects of climate change or the identification of new species of animals or plants in new areas

□ Explore Open Data

Watch the videos on Open Data below. As you watch, keep notes on how the data analysis process can be applied to Open Data.



<https://www.youtube.com/watch?v=qSD9ob8rGcs>
(2:30 minutes)



<https://www.youtube.com/watch?v=iOrPK7p2AwI>
(5:00 minutes)

Collect or Choose Data

Video 1: Open data is data that can be freely used, re-used and redistributed by anyone. Data provided by the government about taxes, air quality, and public roads for example.

Video 2: Any data that is publicly available. The data cited in the video are papers published publicly. Researchers can access both the papers and the associated data.

Clean and/or Filter Data

Video 1: Filtering data depends on the case. For example, if you are looking for a new place to live you can filter data about schools or crime. You can even filter for potholes.

Video 2: It depends on the research focus. Depending on the focus, researchers can filter for papers and/or data in their area.

Visualize and Find Patterns

Video 1: How it's visualized depends on the data being filtered. You can locate roads for potholes spatially and use this information to determine areas that need maintenance.

Video 2: It depends

New Information

Video 1: New information about different communities and inequalities across communities.

Video 2: Better technology, better prediction models

□ Define key vocabulary

Scalability

Parallel Systems

Citizen Science

Crowdsource

Open Data

Open Access

☐ **Receive Credit for this lab guide**

Submit this portion of the lab to Pluska to receive credit for the lab guide.