In this video we’ll discuss data sets: what they are, why they are important in data science,

and where to find them.

Let’s first loosely define what a data set is.

A data set is a structured collection of data.

Data embodies information that might be represented as text, numbers, or media such as images,

audio, or video files.

A data set that is structured as tabular data comprises a collection of rows, which in turn

comprise columns that store the information.

One popular tabular data format is "comma separated values," or CSV.

A CSV file is a delimited text file where each line represents a row and data values

are separated by a comma.

For example, imagine a data set of observations from a weather station.

Each row represents an observation at a given time, while each column contains information

about that particular observation, such as the temperature, humidity, and other weather

conditions.

Hierarchical or network data structures are typically used to represent relationships

between data.

Hierarchical data is organized in a tree-like structure, whereas network data might be stored

as a graph.

For example, the connections between people on a social networking website are often represented

in the form of a graph.

A data set might also include raw data files, such as images or audio.

The MNIST dataset is popular for data science.

It contains images of handwritten digits and is commonly used to train image processing

systems.

Traditionally, most data sets were considered to be private because they contain proprietary

or confidential information such as customer data, pricing data, or other commercially

sensitive information.

These data sets are typically not shared publicly.

Over time, more and more public and private entities such as scientific institutions,

governments, organizations and even companies have started to make data sets available to

the public as “open data," providing a wealth of information for free.

For example, the United Nations and federal and municipal governments around the world

have published many data sets on their websites, covering the economy, society, healthcare,

transportation, environment, and much more.

Access to these and other open data sets enable data scientists, researchers, analysts, and

others to uncover previously unknown and potentially useful insights.

They can create new applications for both commercial purposes and the public good.

They can also carry out new research.

Open data has played a significant role in the growth of data science, machine learning,

and artificial intelligence and has provided a way for practitioners to hone their skills

on a wide variety of data sets.

There are many open data sources on the internet.

You can find a comprehensive list of open data portals from around the world on the

Open Knowledge Foundation’s datacatalogs.org website.

The United Nations, the European Union, and many other governmental and intergovernmental

organizations maintain data repositories providing access to a wide range of information.

On Kaggle, which is a popular data science online community, you can find and contribute

data sets that might be of general interest.

Last but not least, Google provides a search engine for data sets that might help you find

the ones that have particular value for you.

It’s important to recognize that open data distribution and use might be restricted,

as defined by its licensing terms.

In absence of a license for open data distribution, many data sets were shared in the past under

open source software licenses.

These licenses were not designed to cover the specific considerations related to the

distribution and use of data sets.

To address the issue, the Linux Foundation created the Community Data License Agreement,

or CDLA.

Two licenses were initially created for sharing data: CDLA-Sharing and CDLA-Permissive.

The CDLA-Sharing license grants you permission to use and modify the data.

The license stipulates that if you publish your modified version of the data you must

do so under the same license terms as the original data.

The CDLA-Permissive license also grants you permission to use and modify the data.

However, you are not required to share changes to the data.

Note that neither license imposes any restrictions on results you might derive by using the data,

which is important in data science.

Let’s say, for example, that you are building a model that performs a prediction.

If you are training the model using CDLA-licensed data sets, you are under no obligation to

share the model, or to share it under a specific license if you do choose to share it.

In this video you’ve learned about open data sets, their role in data science, and

where to find them.

We’ve also introduced the Community Data License Agreement, which makes it easier to

share open data.

One important aspect that we didn’t cover in this video is data quality and accuracy,

which might vary greatly depending on who collected and contributed the data set.

While some open data sets might be good enough for personal use, they might not meet enterprise

requirements due to the impact they might have on the business.

In the next module, you will learn about the Data Asset eXchange, a curated open data repository.