Data Sources

1. Data sources

00:00 - 00:13

Hi, I'm Sara! Welcome back. Previously, you learned about the data science workflow. In this chapter, we'll focus on the first step: data collection and storage.

2. The data science workflow

00:13 - 00:22

Before we can start deriving insights from data, we first need to collect the data from different sources. That's what we'll talk about in this video.

3. Sources of data

00:22 - 01:02

We are generating vast amounts of data on a daily basis simply by surfing the internet, tracking a run, or paying by card in a shop. The companies behind these services that we use, collect this data internally. They use this to help them make data-driven decisions. On the other hand, there are also many free, open data sources available. This means the data can be freely used, shared and built-on by anyone. Note that sometimes companies share parts of their data with a wider public as well. Let's first take a look at company data sources.

4. Company data

01:02 - 01:15

Some of the most common company sources of data are web events, survey data, customer data, logistics data, and financial transactions. Let's dive a bit deeper into web data.

5. Web data

01:15 - 01:45

When you visit a web page or click on a link, usually this information is tracked by companies in order to calculate conversion rates or monitor the popularity of different pieces of content. The following information is captured: the name of the event, which could mean the URL of the page visited or an identifier for the element that was clicked, the timestamp of the event, and an identifier for the user that performed the action.

6. Survey data

01:45 - 01:58

Data can also be collected by asking people for their opinions in surveys. This can be, for example, in the form of a face-to-face interview, online questionnaire, or a focus group.

7. Net Promoter Score

01:58 - 02:14

You've likely answered a question as shown in the image before. This is a very common type of survey data used by companies: the Net Promoter Score, or NPS, which asks how likely a user is to recommend a product to a friend or colleague.

8. Open data

02:14 - 02:22

There are multiple ways to access open data. Two of them are APIs and public records.

9. Public data APIs

02:22 - 02:47

Let's begin with APIs. API stands for Application Programming Interface. It's an easy way of requesting data from a third party over the internet. Many companies have public APIs to let anyone access their data. Some noteable APIs include Twitter, Wikipedia, Yahoo! Finance, and Google Maps, but there are many, many more.

10. Tracking a hashtag

02:47 - 03:28

Let's look at an example of the Twitter API. Suppose we want to track Tweets with the hashtag DataFramed, DataCamp's wonderful podcast on Data Science. We can use the Twitter API to request all Tweets with this hashtag. At this point, we have many options for analysis. We could perform a sentiment analysis on the text of each Tweet and get an idea of how people like our podcast. We could simply track how often hashtag DataFramed appears each week. We could also combine this data with our downloads data and see if positive Tweets are correlated with more downloads.

11. Public records

03:28 - 04:08

Public records are another great way of gathering data. They can be collected and shared by international organizations like the World Bank, the UN, or the WTO, national statistical offices, who use census and survey data, or government agencies, who make information about for example the weather, environment or population publicly available. For example, in the US, data-dot-gov has health, education, and commerce data available for free download. In the EU, data-dot-europa-dot-eu has similar data.

12. Let's practice!

04:08 - 04:17

You now know some common sources of data that can be used when working on a Data Science project, let's practice!