0:14
From any data source, the type of data that we get can range from structured to unstructured.

0:20
Structured data is where you know the schema; that is, you know exactly what kind of information

0:25
is going to be there. The fields are defined, the types are defined, and there is no

0:31
ambiguity or uncertainty around them. A classic example would be a table from a database.

0:36
On the other hand, unstructured data is where you know practically nothing about the kind of input.

0:42
An example would be a photo or a piece of text which could contain anything literally,

0:49
and you certainly do not know the types, and you do not even know whether it follows any

0:54
kind of predefined order or what we call a schema. This, however, is not a binary classification.

1:02
Data sets are not either structured or unstructured but rather lie on a continuum from

1:07
structure to unstructured, and you could even put in a third bucket called semi-structured.

1:12
An example of that might be a JSON file. It does have fields, and they do have values,

1:18
but it is not like there are predefined fields or they only necessarily need to have a specific

1:25
value. It is possible to take unstructured data and convert it to structured data,

1:30
which makes it easier to analyze. And a significant portion of the data

1:35
extraction work that people do involves converting unstructured to structured data. But keep in mind

1:41
that you could get data ranging anywhere from fully structured to fully unstructured

1:46
when you look for data from any data source. One example of structured data is the kind

1:51
of data that you will find in databases. Here is an example of a database table. This
1:56
is a census PCA table that has for each state, district, sub-district, town, ward, etcetera.
2:05
Details about the population and several other pieces of information that are typically available
2:10
as part of the census. Now, this follows a schema that is a defined structure.
2:16
There is a state, which is always an integer. There is a district that is always an integer,
2:21
a sub-district that is always real, and so on. And there is a level that is text;

2:25
there is a name that is text. So, effectively you know not just what are the columns in the data
2:31
but also the type of or the values in this. In fact, it is possible to have an even more
2:38
narrowly and closely defined structure that says that the state can only have a
2:41
number between 0 and 50, the sub-district can only have a number between 0 and 500,
2:46
and the level can only be a piece of text that has two characters and so on.
2:51
These are ways in which you narrowly define the structure of a data set.
2:56
Schemas also make it possible to have interrelationships. So, for example, here
3:00
is a database that contains multiple tables. And these tables contain information that could
3:07
be interrelated. For example, this twitter table has details about what keywords were
3:13
mentioned on Twitter that is part of a certain subcategory within a category
3:18
in a quarter on a certain date. And the quarter column here may be the same as the quarter column
3:24
in the sales table or the reach table, and so on. The categories and subcategories may be the same
3:29
across all of these. And with this information, it becomes possible to join data

across
3:35
multiple data sets. This is one very powerful  characteristic of structured data sets.
3:41
Another example of structured  data sources is spreadsheets.
3:45
In a spreadsheet, quite often, you can put  in data that is reasonably structured like a
3:50
table where you have specific columns, and each  column can have a response that is more or less
3:56
well defined. So, it can only contain a specific  set of values that may be predefined text,
4:01
or it may only contain specific numbers. However,  a spreadsheet is not necessarily used only for
4:10
structured data. It certainly is possible to put  in other kinds of data into a spreadsheet.
4:15
So, it is important to remember that  structured data can be in spreadsheets,
4:19
but spreadsheets are not always structured data.  Just like databases support multiple tables,
4:26
spreadsheets support multiple worksheets. And  it is possible to join many of these by linking
4:33
them together and creating the equivalent  of joins that you have in databases.
4:38
A third example of structured data  comes from shape files. Shape files
4:43
contain geographic data about locations. For example, on gadm dot org, you could look at
4:50
the maps for each country. Let us say, we go into  Afghanistan and then dive into specific divisions
4:57
and subdivisions and look at the  maps for each of these districts.
5:03
It is also possible to download this data for each  country. And when we get the data for Afghanistan
5:11
as let us say a shape file. The data set is a  collection of files that has information about,
5:18
for example, what the shape looks like that  is what you would find in a dot shp file.

5:26
This is what a shp file looks like, for example. And you will also find
5:30
associated information about it in a dbf file which is like an excel file.
5:36
And this is what the associated dbf file for Afghanistan looks like. It has
5:39
details about the country, the state, the district, and a few other pieces of information
5:45
about that particular district. So, shape file is actually a complex structure that contains
5:50
spatial information as well as tabular information packaged into a single container.
5:56
Semi-structured data, again, can have a variety of forms. One example of semi-structured data
6:02
could be found in documents. For example, a pdf file or HTML file.
6:07
Let us take a look at this pdf file. It contains multiple tables, and each of these
6:12
tables contains a different kind of information. But you will notice that as a document in itself,
6:17
its structure is, firstly, a little more complex, and it is not entirely tabular. So, it may
6:22
be possible to extract a schema from this. But it is certainly not stored in a way that the
6:27
schema is easily extractable or even intuitive. So, at the very least, we could say that it is
6:31
semi-structured because we have to figure out what the structure is. And we do not know whether it is
6:36
fully structured or not. But, it certainly looks at first glance to be structured.

6:40
Or take this Wikipedia page that contains information about the world population. There are
6:44
pieces of information here that are structured. And there are pieces of information here that are
6:49
not structured. And it is a container that has a combination of this information which makes

6:54
a web page such as that of the Wikipedia world  population page a semi-structured document.
6:59
Another example of semi-structured data can be  found in messages, such as email or SMS.
7:06
Here is a simple email message  that perhaps is a spam message.
7:10
But, behind the scenes, if you look at the  original, then that information has details
7:16
about the message ID, where it was created at,  what was the subject, and who it was sent to, and
7:22
all of this is in a fairly structured format. But,  it also contains information that is unstructured
7:28
such as the text such as the attachment that came  along with it. And some of the messages, some of
7:34
the information that is part of the headers, they  too can be unstructured. So, it is a combination
7:41
of structured and unstructured data that makes  messages like emails semi-structured.
7:47
In fact, one of the places where you can  consistently find semi-structured data is in
7:51
container formats. When I say container format, it  is a file format or a structure that can contain
7:59
other pieces of information which may or may not  be structured. Zip files are a good example. You
8:04
can have a highly structured spreadsheet  along with it a fairly unstructured text

8:09
file that explains what that document is about.  Or even a docx or a pptx or an xlsx format.
8:18
This inside a docx, you could  have text, you could have a table,
8:22
and you could have an image, some of which are  structured, some of which are unstructured.
8:26
Whenever you see containers that contain multiple  pieces of information, then it is a reasonably
8:30

good hint that it is probably semi-structured,  that it is a combination of structured or

8:35
unstructured or is partly structured, and  you do not necessarily know the schema.

8:39
Unstructured data sets are where you  know little or nothing about the contents

8:44
other than a broad idea of how to parse the  content. For example, that could include text,

8:50
that could include images, audio, video, and  within this. So, let us say we are looking at

8:57
an audio file. You may be able to extract  information such as its length.

9:01
You may be able to identify what is the average  volume in this, but beyond that, getting any kind

9:07
of meaningful information out of something like  audio is not that easy, at least for a system,

9:13
for a human, it is reasonably easy to listen  to it and understand what is being said. But

9:18
for a system to do that, that takes a little  more effort. Increasingly, a big field of

9:24
work is focused on extracting structured  information out of unstructured data.

9:30
Many of the techniques in deep learning  are really focused on. For example,

9:34
can we take video imagery and find out who  are the people that are in the video footage

9:40
or, what are they saying or where are they  located, or are there two people that look

9:45
similar. These are examples of the kinds of  structured information that people are extracting

9:50
from such unstructured data. But by, enlarging  unstructured data until it gets converted into

9:57
something that is structured is really hard to  analyze and do any kind of processing on.