0:14
There are 3 sources of datasets, Public, Private, and Personal. Public Data is open and free,

0:21
you can search online and find it. Today, there is a lot of public data out there. And it is

0:28
actually a little hard to find public data. But a lot is relative. What I mean by relative is

0:33
that the vast majority of the data is actually not public. And therefore, two things happen. One,

0:42
the data that you might want, may actually not be there, you will find a lot of data out there. But

0:47
that is not necessarily the data that you would want. Secondly, even though it may be out there,

0:52
it may not be exactly the kind of data that you want or exactly the kind of format that you want,

0:56
or it may be out there, but you will find it hard to locate it. In this module, one of the things we

1:03
look at is, what are some good ways of searching for and finding public data. The second kind

1:09
of data is private data. This is data that is accessible to a few people. For example, you could

1:15
find them inside an organization but not outside, or you could pay for them, they are not open,

1:22
but anyone can access it as long as they are willing to pay for it. The third kind of data set,

1:28
and this is a pretty interesting one. And also an emerging field is personal data.

1:33
Think of it as data that lies within you, or your devices. For example, all of your call logs,

1:41
you have them, you can extract them, you can analyse them, or your music listening history.

1:47
That is a personal data stream. Your own ratings of these music tracks is a personal data set.

1:54
Increasingly, people are looking towards their devices and their habits of logging

to extract
2:02
data from personal data sets. What we will be going through are examples of how to locate data
2:08
in each of these. A good starting point for public data is the Awesome Public Datasets
2:15
catalog, you can go to Google search for Awesome Public Datasets, and you will find a link on
2:21
GitHub, which says awesome data and that links you to this readme file on Awesome Public Datasets.
2:30
It has several categories, agriculture, biology, climate, energy, natural language, social
2:37
sciences. And for each of these categories.   Let us take social sciences as a category,
2:41
you have links to several databases or collections of datasets. For example, the GDELT global events
2:49
database is a massive database of events that have been scraped from various news sources. And it is
2:58
pretty much the largest and most comprehensive source of any kind of news and event data. And you
3:04
can download about 2.5 terabytes of these just for the last year as raw data. Or if you are looking
3:12
for datasets related to let us say, finance, then Google Finance has an API, you can look at OANDA
3:20
which has currency and commodity data. You can look at EDGAR, which has all the SEC filings,
3:25
the financial reports for US companies. And this is just a list of some of the datasets. Some of
3:34
these may, in fact, have links to other datasets from where you can download even more data.
3:40
GIS is a pretty big and hot section. So for example, there is GADM, the Global Administrative
3:47
Areas Database. This has data sets for every country. So if I go down to let us take a quick
3:56

look at the maps themselves, you have shapefiles for several countries, let us take the United,
4:03
let us take Ukraine, and within Ukraine, it has maps for each of the subdivisions.

4:10
So that is effectively the equivalent of states. Within that we can dive down to the next level,
4:16
which is the sub division. And within that, if the data is available, go down to the third level
4:22
and see information there. All of this can be downloaded from the data section. And you can
4:29
download it by country. Let us say I pick Ukraine. I can get this either at level zero, which is the
4:37
country level or the state level or the district level, and in a variety of different formats.
4:46
A good starting point. If you are not sure what kind of data set you want to go for,
4:50
there are awesome public datasets. Keep in mind that this is more for exploration than discovery.
4:56
If you know exactly what you are looking for, then maybe a quick search on this page might get
5:01
you something. But it is more for you to read through and find what is there yourself rather
5:06
than like a search engine. A second source for public data is Google DataSet search,
5:12
you can go to Google and search for Google DataSet search. The result will take you
5:17
to Google's Data Set search engine. And this is like a search engine for data sets specifically.
5:23
For example, let us say we want to search for the FIFA World Cup data set. FIFA
5:30
World Cup. And there are a series of data sets highlighted here in the auto suggests, but let me
5:37
just set FIFA and press enter. Now, that gives me a whole series of FIFA data sets. About 100 plus
5:44

data sets are found. I can sort by what has been updated, let us say in the last year. So
5:49
we want something that is relatively recent. And I can restrict by formats, whether I can use this
5:58
for commercial use, or only non-commercial. And there are a series of subtopics within that that
6:03
we can look at. But so far we have selected in the past year. And here is a complete player data set,
6:10
there is an official data set and this has information about all of the players.

6:16
It has whatever else, this has player ratings and so on. You can link, you can click on any one of
6:24
these, open the data set, and either download it from that data source or from there go to other
6:31
data sources. Google DataSet search works based on individual sites exposing their data in a specific
6:39
metadata format. And in this module, there is an optional video that will give you a better
6:44
sense of how Google Data Set search works. This is not necessarily the best way of finding data,
6:51
the way that Google searches on Google search, you will find pretty much
6:55
anything that you are looking for, by and large. But Google DataSet search is still in its infancy,
7:01
there is a good chance you may not find what you want. So if you do not find something
7:05
on Google DataSet search, do not assume that it is not out there. It simply means that the
7:11
people who have created the index, and the people that have put up the dataset outside,
7:16
have not really connected with each other.   Another popular source for public data is Kaggle.
7:21
You can search for Kaggle datasets. And the very first link that you find Kaggle datasets,
7:28

will give you a list of several data sets that you could explore.
7:32
And you could search within these as well. These are data sets that people have uploaded to Kaggle,
7:38
either for competitions or to learn and explore with each other. It is a reasonably large data
7:44
set. So out here, for example, if I am looking for data on, I know, let us say, Harry Potter,
7:50
let us see what data sets it has. There are several data sets, list of spells,
7:57
the movies data set, and overall Harry Potter data set, a fanfiction data set, scripts of individual
8:08
movies and so on. And for example, in the movies data set, it looks like we have details about the
8:14
chapters, the characters, the dialogues, even more information about the movies, what places
8:20
were used. So that is pretty interesting. List of all of the characters with their age,
8:26
house, wand, blood status, and so on, that is another one that you could find here. Again,
8:33
this is only a subset of public data that has been uploaded to Kaggle. So while you may find
8:38
interesting stuff here, the fact that there is nothing here that relates to what you want does
8:44
not mean that the data is not out there, it just means that it is not on Kaggle.

8:48
Another source of public data is from governments. Many governments have put up websites like
8:54
data dot gov, data dot gov dot in, data dot gov dot uk. And these have datasets that are owned by
9:02
and published by the government often related to the government's functioning. A good way to
9:07
find them is to go to awesome public datasets and search within that for the government, where you
9:12
will find datasets for provinces like Alberta and Canada, or cities, like Antwerp

and Belgium, or
9:21
entire countries' data portals. So if, for example, we open the Brazil data portal that is
9:28
at, dados dot gov dot br slash data set and the site is in, well something that is not English,
9:37
but we can infer that it is got about 11,000 data sets in PDF, CSV, Excel, XML,
9:47
zip XML, and a few other formats. This is a pretty authoritative data source. What I mean by that
9:56
is given that it is published by governments, it comes with the backing of the government. So you
10:01
can assume a certain amount of official nature or officiality about the result. But that does not
10:08
necessarily mean that the results are either right or of good quality. Several of these data sets
10:15
are published on a, as is basis, so the data gathering process may still be flawed. The data
10:22
collection process may still be flawed, a lot of columns may be empty, a lot of rows may be empty.
10:27
But they are the official results and therefore can be used as a basis for a lot more confidence
10:33
in your analysis than data sets from unofficial sources. Apart from these, if you are looking
10:38
for people to help you find data or location, your best sources to join a community.
10:43
In India data meet is one such community of data enthusiasts who often look for and help each other
10:51
find datasets. So you can search for data meet, and they will take you to data meet dot org. On
10:58
data meet dot org, you will find details of what data meet is. But the main thing is the mailing
11:06
list, which is a Google group. And on this Google Group, you will find several people
11:11
posting questions like somebody is requesting data not to shapefiles, somebody is

looking for carbon
11:16
footprint information, somebody is looking for details about Delhi Metro, somebody is looking for
11:21
the NFHS data set. And you can see that there are some conversations which are fairly long and
11:29
detailed. So there are responses, but some like where somebody is asked for the Jharkhand village
11:34
boundary as per the 2011 census where there are no responses. It is hit and miss. But the good
11:39
part is there is a fairly large community. Well over 3000 people that had, that can help you find
11:47
the data set that you are looking for. Apart from public data sets, there are private data sets. And
11:53
usually you will find private data sets within the bounds of organizations. A corporate data set is
12:00
something that an organization has and usually does not share outside of the organization. For
12:05
example, the list of employees in an organization or the financial details of an organization,
12:12
the product specifications for an organization, performance details, operations, the logs of
12:19
each of the production batches that they have run. These are all examples of data that many
12:25
organizations collect as part of their process.   But because this information is either sensitive
12:31
in that it involves details about other people or organizations, and they cannot share it. Or it
12:37
involves information that gives them a potential advantage, and therefore they want to keep it
12:41
private. This kind of information is abundant within an organization that you may be working
12:46
for. So it is largely a matter of knowing where there are data sets within the organization,
12:52

you may be able to search, you may be able to ask and then source the data accordingly. But this
12:58
data, private corporate data, is perhaps among the most invested in the data set. That is people are
13:06
spending more money on this than any other kind of data set. Another kind of data set is private,
13:14
in the sense that it is not free and open, but you can purchase it. There are several sources of paid
13:20
datasets. For example, if you search on Google for pay datasets, you get sites like data dot world,
13:28
Google Cloud has datasets that you can install on the Google Cloud and use,
13:33
Google does not pay datasets. But you can find pay datasets in the likes of Dun and Bradstreet,
13:40
Hoovers. And if you search for a data set, then you will find several others like Statista,
13:47
which sells reports, Bright data where you can buy data sets, data stock, data and sons, and so on.
13:54
The thing about paid datasets is that they have a wider and usually more reliable collection than
14:02
public datasets in some areas, like for example, finance. So if you are in an area where people are
14:10
selling paid datasets, you are best off going for those paid data sets. But if not, you may find a
14:15
richer public and open data than paid data.   A third source of data sets is personal data.
14:21
And this is rather interesting, because this is something that is unique to each person.
14:28
Only that person has full access to that data. For example, a great source of personal data
14:36
sets are mobile app datasets, stuff that is on your mobile app. For example, I am going to take
14:42
my phone and just go through each app in it and talk about what data I can extract from that.
14:48

The first app that I have is messages. From this I can extract the list of people who send me
14:53
messages, whether they send me messages in the morning or the evening. Are there certain words
14:57
that are commonly used by certain people who often thank me? Who says please, who does not?
15:04
Which are the junk messages? Are the junk messages often sent at a specific time? These are examples
15:10
of what I can do with data just by exporting data from my messages app. With WhatsApp,
15:17
we can go something, or to something that is slightly richer, we can look at who calls?
15:21
Do we call people more often? Or do they call? Who tends to miss the most calls? Whose calls are
15:27
often not picked up? When do people call are some early morning people or some late night people?
15:33
How long do people talk? Who are the people that we have the most conversations with? Are there
15:37
certain people that we call right after we call other people? Do we message people after we call
15:42
them to people who message us back after? Do we message them back after they call us? All of these
15:49
are examples of what we can extract from WhatsApp. Let us say the health app. For me I am tracking my
15:57
audio headphone level exposure. So I know that over the last several months, I have stayed well
16:03
below 80 decibels. And I can see at what times of the day I listen to loud music versus quiet music.
16:11
I can use the sleep tracking data, which tells me how many hours I sleep. So do I sleep more on
16:17
weekends? Do I sleep less on weekends? The number of steps that it counts? This is something that
16:21
I used to ask myself: do I walk more during the morning than during the evening? Do I walk in?

16:27
Do I take long strides to have short strides? How does that vary based on how much sleep I had
16:34
the previous day? Weight tracking is something that I have been using for several years now.
16:40
And that can tell me, on to do I put on weight in holiday seasons? Do I tend to lose weight at the
16:46
beginning of the year after I made a resolution? And all of this is from the Health app. From
16:51
the email app I can look at when I get messages, what kinds of messages I get, can I classify them
16:57
into useful versus non useful? Are there people who send me more messages with specific words?
17:02
Can I automatically figure out who I tend to talk to more after I, at certain periods of time?
17:09
Similarly, we can explore calendar entry data. So what are the meetings that we attend,
17:14
what appointments are set up by whom, who wastes our time, the most, and so on. And all of this
17:20
is just from the first four apps that I had on my phone. If you look at your social apps, Twitter,
17:26
Facebook, LinkedIn, Instagram, you will find that you can figure out not just your patterns
17:33
of usage, but other people's patterns of usage as well. If you look at apps related to finance,
17:40
for example, what are you buying on Swiggy? Do you have a preference towards certain kinds of foods?
17:46
Who are you paying on UPA apps? Are you paying vendors more? Are you paying for smaller purchases
17:53
more? Are there personal transactions that are more common? Or if you look at entertainment apps,
18:00
what kind of music do you listen to? What do you rate highly? What do you rate poorly?
18:03
If you have good reads installed, then what are your book ratings? Do you tend to

rate
18:09
fantasy more than crime? Do you tend to rate history more than business?
18:15
All of these datasets are personal and come from just one source exporting data from your mobile
18:22
application. While mobile apps are a good representation of a source of personal data,
18:27
this is not the only source of personal data and any kind of personal logging
18:31
can be used for this. Some people, for example, write down the number of hours of sleep they have
18:36
or write down, for example, the diet that they are following, how many calories they have had
18:42
on a given day, who they speak to their feelings, literally, in a diary. All of these are sources
18:50
where we log personal information. So put another way, any time you enter some information somewhere
18:57
could be on a piece of paper, it could be on a digital medium, like your PC or your phone.
19:04
These are instances of personal logging. Any action that gets captured is a potential source
19:10
of personal data. And this can lead to some very interesting and very powerful analysis.