0:13
Before you do any kind of data science, you obviously have to get the data to be able to
0:18
analyze it, visualize it, narrate it, deploy it. And what we are going to cover in this module
0:24
is -how do you get the data? There are three ways you can get the data. The first is -you
0:29
can download the data either somebody gives you the data and says download it from here, or
0:34
you are asked to download it from the internet because it is a public data source, but that is
0:38
the first way you download the data. The second way is you can query it from
0:43
somewhere. It may be on a database, it may be available through an API, it may be available
0:47
through a library, but these are ways in which you can selectively query parts of the data and stitch
0:53
it together. The third way is you have to scrape it. It is not directly available in a convenient
0:58
form that you can query or download, but it is in fact on a web page. It is available on a PDF file,
1:04
it is available in a word document it is available on an excel file.
1:07
It is kind of structured, but you will have to figure out that structure and
1:11
extract it from there. In this module, we will be looking at the tools that will help you either
1:17
download from a data source or query from an API or from a database or from a library and finally,
1:25
how you can scrape from different sources. Let us start with how you can download data. The
1:33
most obvious sources you search on Google let us say you want movie data. So, just say movie data
1:39
download, and that might give you a few sources like Kaggle seems to have a movie data set and
1:46
data dot world has a list of movie data sets the internet movie database has a

series of data sets.
1:53
So, you could go through each one of these. So, this one seems to have credit information,
1:57
and it is about 189 MB of data. Dot world has links to several other data sources,

2:03
and there is the internet movie database data set. This incidentally is something I know to be one of
2:08
the most comprehensive movie data sets perhaps the most comprehensive movie data set that that has
2:16
details about every single title and information further information about each of these titles
2:22
the crew of every one of these and so on. And it is something that you can just directly
2:27
download from the AWS site. It is available as a gzipped file that is a tsv, tsv stands
2:34
for tab-separated value. So, that is like a comma-separated value but separated by tabs. To
2:39
give you a feel for what one of these files looks like, this is what names dot basic dot tsv looks
2:45
like; it has a name constant basically the primary id of the actor or the director or whoever.
2:52
The primary name, which in this case is Fred Astaire the year of birth 1899 the death year
2:59
1987 what their profession is primarily soundtracks. Fred Astaire as a singer and

3:04
an actor and other things, and Fred Astaire is known for these titles,
3:10
and these titles are described in title dot basics dot tsv. Now, this kind of information is your
3:17
starting point to create any kind of analysis and any kind of visual representation.
3:22
So, for example, one of the things that I put together with this IMDB data was a personal
3:29
movie watching or movie recommendation engine for me. I looked at it, this is at

grammar.com
3:37
IMDB, and you are welcome to explore it. This shows the movies by rating. So, the higher-rated
3:44
movies are on top and by popularity, the number of votes that they got. So, movies with just
3:49
ten thousand votes or less versus movies with two thousand sorry two million votes and above.
3:54
So, if I look at each of these marks, they represent a set of movies with a certain rating,
4:00
and a certain number of votes or a vote range on the top right are the movies that are both popular
4:06
and highly rated. So, if I look at that list that contains the Shawshank Redemption, Dark Knight,
4:11
Inception blah blah blah pretty much all of which I have seen, and I can use this to figure out
4:18
now what are the other slightly less popular movies that I need to watch.
4:23
So, again most of these I have seen, but the Intouchables is one of those I am yet to see. So,
4:31
that is a pretty strong recommendation for me to watch next, or if I am not quite in the mood, then
4:37
I could ask the question, are there movies that are really popular despite them not having, I mean
4:44
movies that are more popular than their rating suggests. So, we have here a few movies that
4:52
have a relatively. Let us move this a little bit okay, a couple of movies that have a fairly high
4:58
number of votes but a really terrible rating. And Radhe is one of those fairly popular for
5:04
those who may not know that is a Salman Khan starrer which has a 1.9 rating which
5:10
is like among the lowest that anyone ever gives on IMDB, but it is pretty popular.
5:17
This can tell a number of other stories as well. So, I have been using this to

look at the history
5:23
of animation movies. So, if you filter by type movie and genre animation and look at the story of
5:29
animation across decades in the 1930s, there was pretty much only snow-white in the Seven Dwarfs
5:34
Disney's first entry into animation. But then in the 1940s, they made four movies,

5:39
Pinocchio Bambi Dumbo Fantasia, but other studios also started coming in, but they were not doing
5:44
such a good job of movie writing, and Disney was clearly a cut above the rest that continued in
5:51
the 1950s as well and in 1960s though the other movie started catching up movies where kind of
5:58
Disney only. In the 1970s, there seems to have been firstly a
6:02
slight deterioration of Disney quality. And an overall mix, you cannot really tell

6:07
which are the Disney movies and non-Disney movies, and this was easily the worst period for cartoons
6:12
in general. 1980s, we have a slight resurrection, but this is not Disney-driven. This is, in fact,
6:20
Ghibli studios from Japan. So, my neighbor Totoro or Akira, Nausica, Castle in the sky;
6:28
many of these are classic Japanese anime films and is the birth of anime in cartooning.
6:35
1990s, however represents the golden age of Disney and Pixar, with Lion King coming in from the
6:40
traditional Disney studios and Toy Story breaking the mould with computer animated cartooning
6:46
and that partnership continued in the 2000s with several hugely popular and highly critically rated
6:53
movies like Wally, Finding Nemo etcetera. 2010s that trend continues, and now you can see that it
7:00
is pretty much mainstream movies animation movies are pretty much really popular.

7:06
And in the 2020s, Soul is one of those that has really broken through as one of the better films.
7:12
Now you will notice that all we are doing is having downloaded the data just applying some
7:16
three simple filters animation movie changing the decade and showing this on a scatter plot
7:21
of rating versus number of votes it is that simple. But that gives
7:27
you a sense of the power of bulk data. If you are able to download data at one shot,
7:32
then the kinds of things that you can do are really interesting. But what if you cannot
7:36
download that data? What if there is no such data and you have to figure it out through other means?
7:42
Well, one of our clients, a media company, was doing exactly that they wanted to understand,
7:48
what people what the public is interested in. So, they hired a marketing agency to do surveys.
7:53
They talked to almost a few thousand people and came back with what is it that most of India is
8:00
curious about now. These surveys cost a lot of money. Our clients said they were paying
8:04
something in the order of like 30% months worth of salary for these kinds of market
8:11
research surveys. So, their question was- is there a way by which we could actually get
8:15
this information without so much effort. Well, let us do one thing. Let us go to Google
8:20
and search for how to? And you find that Google automatically provides a series of suggestions how
8:26
to use CDC voucher, how to screenshot on windows how to wrap a gift or screenshot on MAC recall
8:30
email outlook, and so on? Now, this effectively is based on what people are

searching for in
8:37
recent times. Now, I happen to be in Singapore at the moment while I am recording this video.
8:41
So, these are results for Singapore, but in a given region, you can find out what are,
8:46
what is most popular with the public? What is it that the public wants to know without really
8:50
having to do that much research because Google is already doing this research for us? Is it possible
8:55
for us to query this information? Now, it is going to be pretty hard for us to bulk download it,
8:59
but can we query it and get it in pieces. Well, it turns out that this is an undocumented
9:04
API in Google. So, one part of it is that there is an API which is good news because that means we
9:11
can run the same query and get that information, but the bad part is it is undocumented, but we
9:16
can figure it out. So, let us press F12 here, go to the network inspector and search for how to
9:24
put another space, and it runs this query. It sends a request to google.com complete
9:31
search with a series of parameters. If you look at the parameters as a queue which is
9:36
how to a cp? I do not know what that is a client gws? who is? I do not know what that is and so on.
9:42
But what I can do is copy this URL and see if I can run a different query. So, instead of how to?
9:49
Let me try where and see what results I get. So, it downloads a certain file which looks like this.
9:57
Let us see what it returns? It returns an array of array, of arrays where the entry says- where to
10:04
use CDC vouchers, where to go in Singapore, where is santa where to change token, and so on.
10:11

So, let us see if that is actually correct. Let us search for where and yeah where to use CDC
10:18
vouchers where to go in Singapore? Where is Santa? These are exactly the results that we get. So,
10:23
which means that we; can take a URL, like what we just saw, and change the query parameter from
10:30
where to any other string and see what happens. So, we actually did that, and at gremenar dot com
10:37
search there is a page that shows when people type how to in different countries like India,
10:44
UK, US, Singapore what are the results. So, in India, for example, how to delete Instagram
10:49
account, how to calculate percentage, how to deactivate Instagram account? These are the top
10:53
trending queries at the moment. So, that is what India wants to know how to delete or get out of
10:57
Instagram. The UK and the US want to know how to screenshot on the MAC. The UK wants to know how to
11:02
reduce weight fast. They also want to learn how to make pancakes which is a bit of a contradiction,
11:08
but both- the US and the UK, one knows how to get rid of fruit flies.
11:12
So, that looks like a perennial problem, or if you know, you ask the question, how do I?
11:18
The UK is curious about how to get a PCR test, but the US is more interested in taking a screenshot.
11:24
So, you can see that there is more of a coveted problem in the UK than in the US. So, if you were
11:30
in UK publication, you would be writing more about covet, whereas you'd probably be writing
11:34
more about technology for a US publication. Now what we were doing here was effectively
11:40
using an API that Google provides. This happens to be undocumented, and querying for information
11:46

and assembling it together to get bulk data on top of which you can do further analysis.
11:52
The third way of getting data is when both these methods fail, there is no ready download. It is
11:58
not like there is a clean API that you can call that will give you structured data.
12:02
But the information is either say in a pdf file or in a web page or in a word document excel document
12:09
it is kind of structured. But you have to go into it, pull the right pieces of information
12:13
and assemble it. That is hard, but sometimes it is the only way, and sometimes it is worth doing,
12:18
and that is what scraping is about. That is the third thing that you learn in this module
12:21
in terms of ways of getting information on how does one scrape and where might we use it?
12:28
So, one of the things I was curious about is who is the fastest one-day international cricketer
12:32
in terms of strike rate. So, if you look at, for example, the list of all play on housetad dot com.
12:38
So, this has let us say all the players whose name starts with C. So, Chad cans and so on. Let us
12:45
pick one of these Campbell and look at Campbell's performance in one day international.
12:51
So, Campbell has a strike rate of 66.19; that is in terms of out of every 100 balls that he
12:59
is played, he's scored on average 66. Now I am curious about who has a high scoring rate,
13:06
but of course, I am also interested in seeing how much they've scored overall. So, Campbell
13:10
scored 5000 runs in one day international that is sizable. If somebody just played one match,
13:15
hit a six, and got out. They'd have a scoring rate of 300 because two balls six runs,

13:21
but that doesn't really count right. So, we have also got to look at how substantial
13:26
their score is? Now, I could compare Campbell with another player. Let us pick Chandrapal,
13:34
and Shivnaran Chandrapal has scored 8000 runs at a strike rate of 70. So, he is
13:40
a slightly faster scorer than Campbell, who has been scoring only at 66.
13:45
Now, this information needs to be compiled across all of these pages and that is quite a task.
13:52
But that is exactly where scraping comes in. What you see here on the right is a scraper for
13:57
how stat it uses the beautiful soup python library to scrape the data, and you will be studying this
14:03
in some more detail in one of the examples in this module. But what we do here is effectively start
14:08
with the player list page, which is what you see on the left. Now you can see that it says group is
14:15
equal to C that basically means it is showing all players whose name starts with the letter C.
14:20
And I could change that to group is equal to D and once it reloads
14:24
it is showing players like Da Costa Silva, Dahani and so on. Players whose name starts with D,
14:32
and this gives me a complete list. So, this is like an entry point into getting the information
14:40
on a player-to-player basis. So, what the scraper does is go through all of the letters from A to Z
14:45
and open the pages with the player list and then within that it specifically looks for any link
14:54
which points to player overview summary dot asp question mark player id is equal to something.
15:00
So, this 2247 is the here's player id, and if you click on this, you will go to the next page
15:06
which has the details of their performance. So, it searches for player overview

summary question
15:12
mark player id is equal to any number and gets all of these numbers and the whole links and
15:17
then it goes into the respective pages that have their ODI summary and from there gets
15:24
information from the tables by looking at all of the cells and concatenating all of those and
15:30
then putting them into a single CSV file. So, this is kind of how scraping works and what
15:36
can one do with this kind of scraping. Well in my case, I was curious to see who the top one-day
15:41
international players were strike rate-wise. This is data that was pulled in 2011. So, it is fairly
15:47
outdated, but you can see that Tendulkar is one of the larger players
15:50
in terms of number of runs. The size of the box represents the total number of runs.
15:54
So, Tendulkar scored the most number of runs at that point with 18 000, followed by Ganguly
15:59
at 11 000, followed by Dravid at 10 800, and so on. And the color represents the speed. So,
16:04
Sehwag's scoring really fast at a strike rate of 105 coupled there was a reasonably
16:08
fast scorer at 96. Use of Pathan has not really scored much but still is pretty fast at 115
16:14
but in contrast, is not a very fast scorer 62.9 strike rate Ravishasri is
16:21
even slower Mahindra not this even slower with ah 50 what was it two okay it is a little hard to
16:27
remember not said 58 and so on. Now you will find that the older players
16:31
are relatively slower. The modern players are generally faster. There has been a slight creep
16:36
in strike rates over time. And then you can drill down
16:40
and look at individual innings every single one of Tendulkar's innings. So, this

was his 200 at
16:48
again South Africa in 2010. This was a much slower inning where he scored 100 at only a
16:54
72 strike rate, and here is another relatively slow innings where he scored at a strike rate
17:01
of 32 and got to only 17 runs and so on. Now you can expand this and get the full list of
17:07
every single one-day international ever played until that point
17:11
by anyone in India and then start looking for their performance against specific countries like
17:16
who is scoring really well against Pakistan. It turns out that Sehwag's consistently scoring well
17:22
against Pakistan. And Dhoni has got a reasonably good performance or who does well in Sharjah?
17:27
So, Kholi, Dhoni have badly played in Sharjah. Yuvaraj got a really poor performance in Sharjah.
17:32
It is mostly small number of runs at a slow pace but Tendulkar is doing really well in Sharjah
17:39
and so on. Now, this becomes possible because what we are doing is extracting the information
17:44
page by page from a site like how stat and pulling that out to be able to create the kind of
17:52
data set that we want to be able to analyze. So, these are the things you learn in this module
17:58
about getting the data. Number one, how do you find data that you can download.
18:03
Number two, how do you query data if it is in the form of, for example, an API? We will be looking
18:08
at the nominating API, which will help you convert addresses to latitudes and longitudes for example,
18:13
the BBC weather API, which will give you the weather for a particular location.
18:17
And we will also look at how you can scrape data from web pages, how you can scrape data from pdf
18:23

files, and so on? To be able to collect all of the data that is required for further analysis.