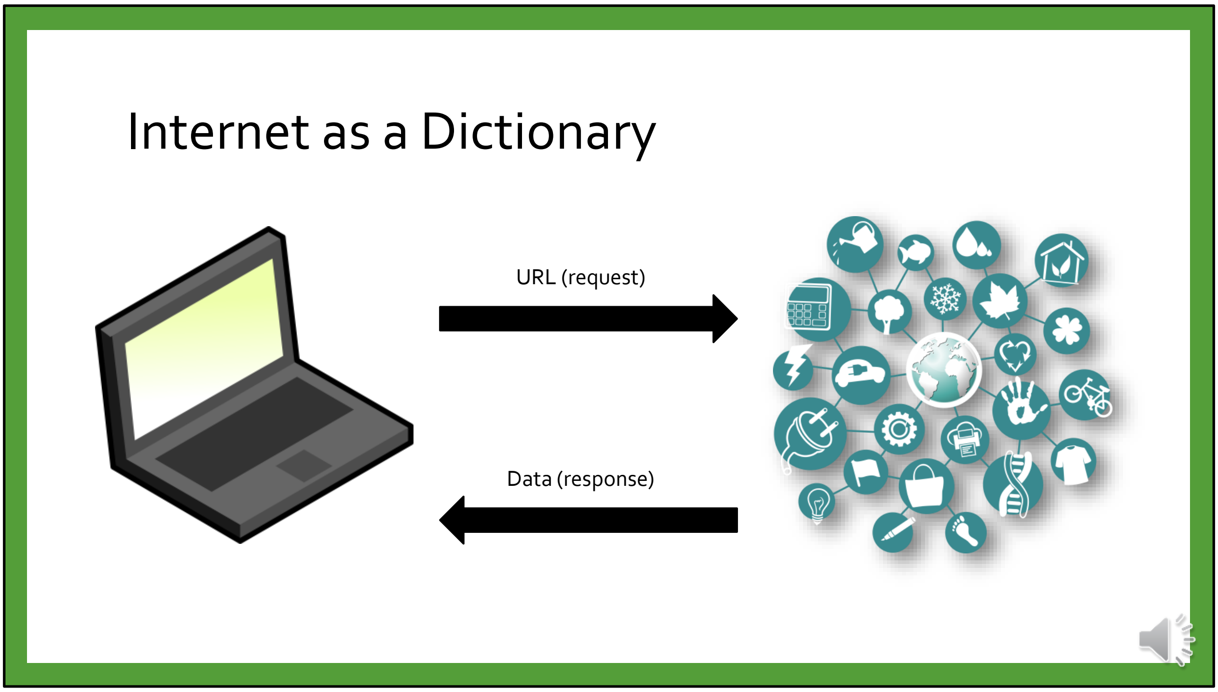
WEB DATA

An Introduction to Computer Science

Let's learn about Web Data.

The Internet is Great

One of the reasons the internet is great is that it makes it easy to share data.
This data can be access programmatically and processed in a program.
Most modern programming languages, like Python, make it easy to access web-based data.

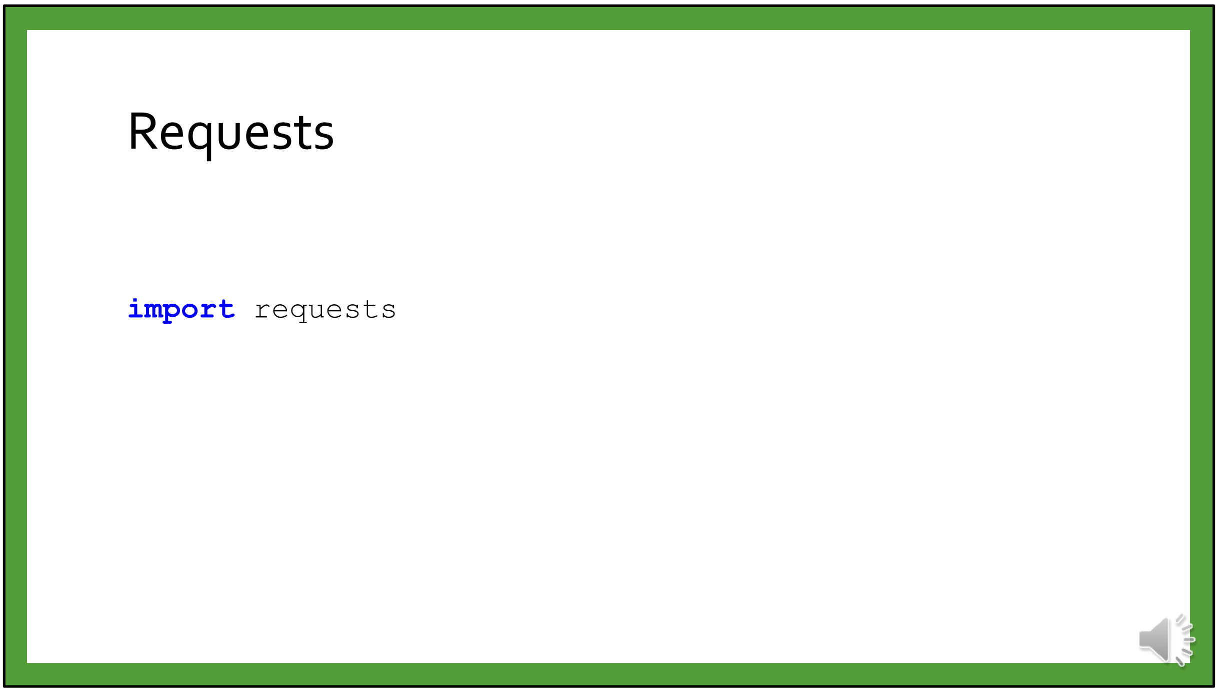


Internet as a Dictionary

URL (request)

Data (response)

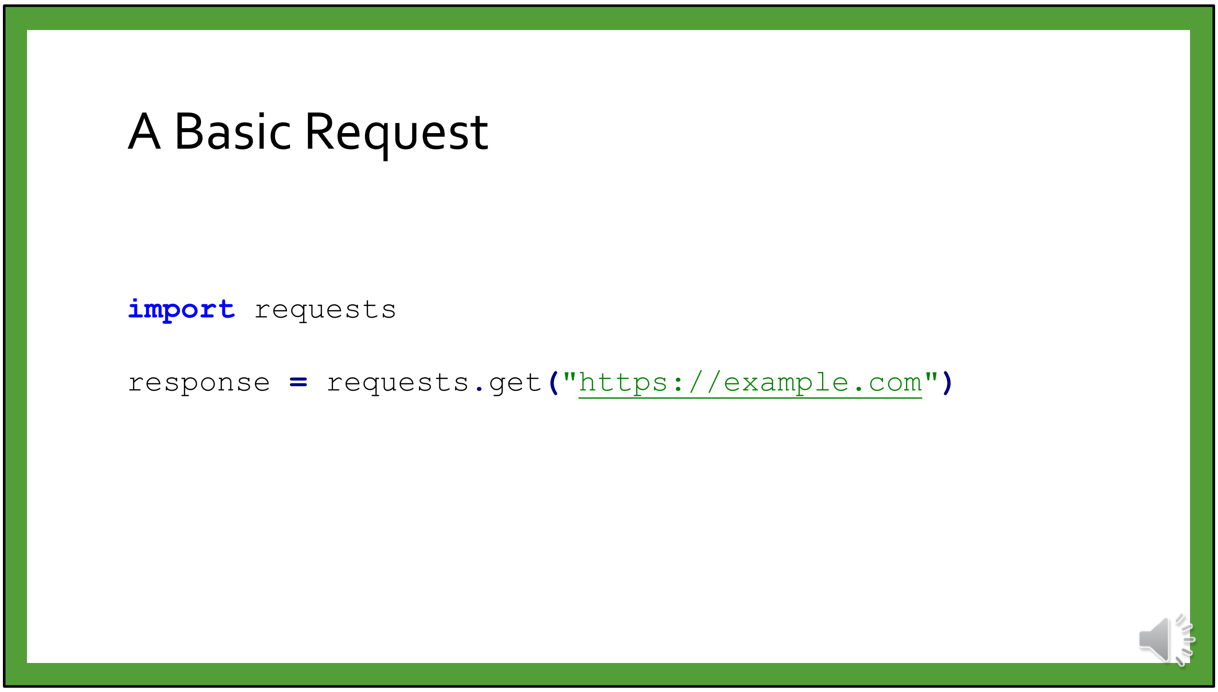
The internet can be seen as a giant Dictionary being used as a Record.
You access URLs (which are strings) and get back web pages (which are also strings).
This simplifies a tremendous number of very complex pieces of hardware and software, but this is the result at the end of the day.
Look up a URL, get some string data back.
We'll use a library that makes it just this easy in Python.



Requests

```python
import requests
```

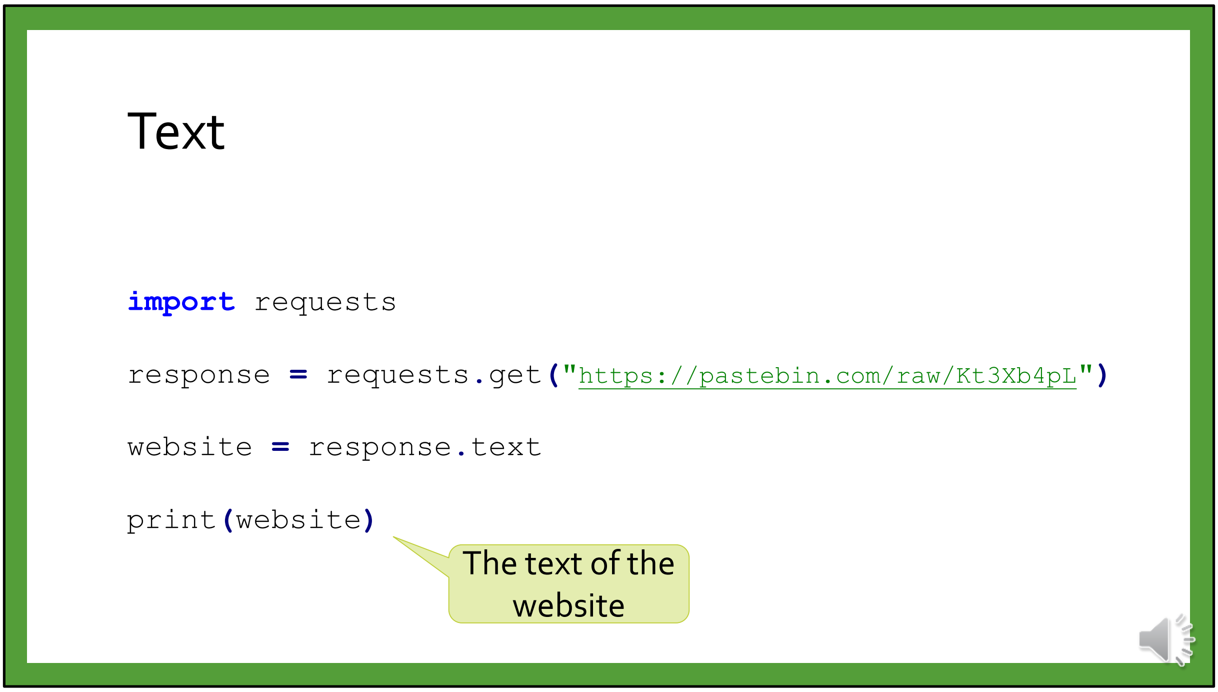
There are many ways to retrieve data from the internet.
The Requests library is one such tool.
This library is not part of the standard library, but comes with Anaconda.
To get started with the Requests module, we begin by importing it.



A Basic Request

```
import requests

response = requests.get("https://example.com")
```

The Requests module has a number of useful functions, but in this lesson we only need one: get.
The get function consumes a URL as a string, and returns a Response object.
This response object can be a confusing thing at first, but it is actually very easy to use.

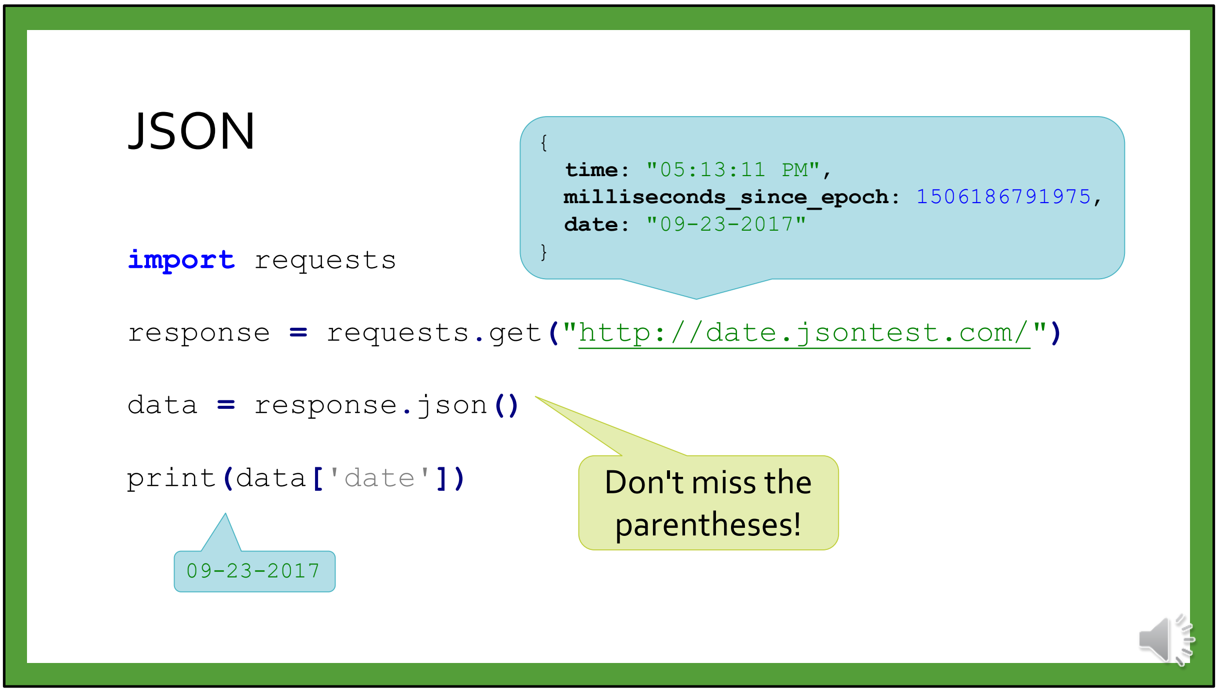


Text

```python
import requests

response = requests.get("https://pastebin.com/raw/Kt3Xb4pL")

website = response.text

print(website)
```

The text of the website

To get the webpage as a string of data, you can use the "text" attribute of the Response object.
Keep in mind that most websites are written in a language called HTML, which can be tricky to parse.
So for now, we'll access simple text websites.



JSON

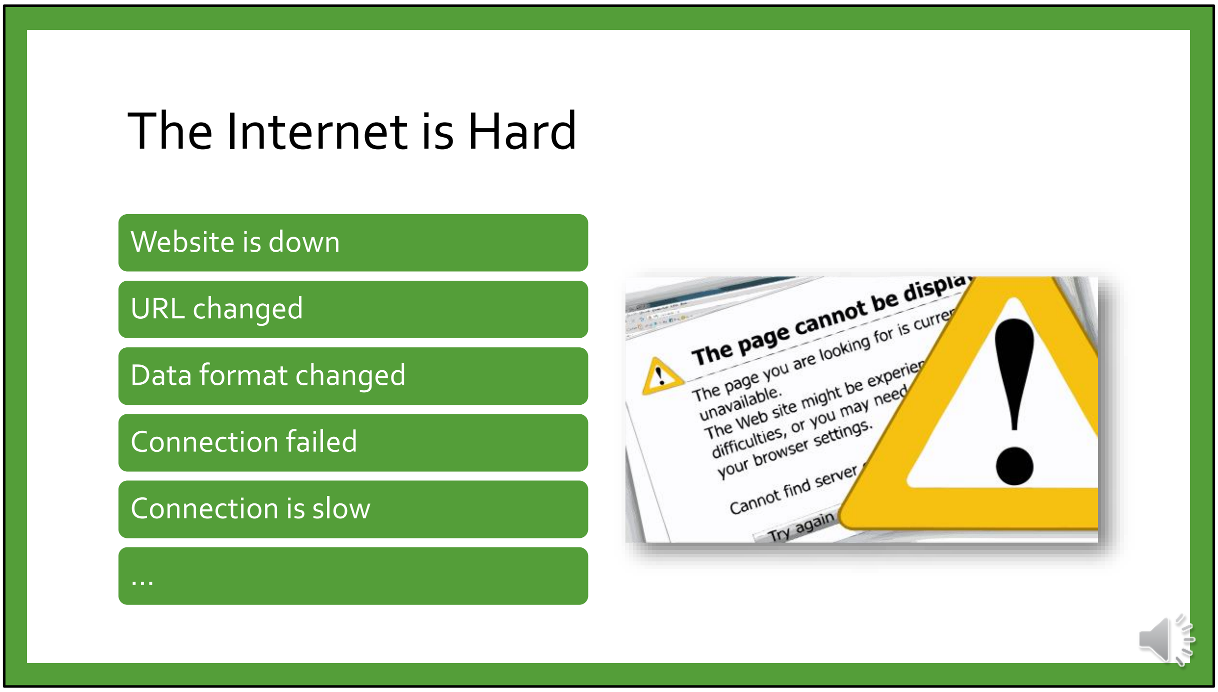
```
{
  time: "05:13:11 PM",
  milliseconds_since_epoch: 1506186791975,
  date: "09-23-2017"
}
```

```
import requests

response = requests.get("http://date.jsontest.com/")

data = response.json()

print(data['date'])
```

```
09-23-2017
```

Don't miss the parentheses!

If the text content we retrieve from the website happens to be in the JSON format, Requests has a simple way to process it using the JSON method.
Confusingly, JSON is a method call instead of an attribute, so make sure you DO use parentheses with it (unlike the "text" attribute).
Here, we use a website that returns the current date and time as a JSON object.



The Internet is Hard

- Website is down
- URL changed
- Data format changed
- Connection failed
- Connection is slow
- …

The page cannot be displa
The page you are looking for is curren
unavailable.
The Web site might be experien
difficulties, or you may need
your browser settings.
Cannot find server
Try again

Connecting to the internet can be problematic.
Websites go down, URLs change, your internet can get disconnected, and many other issues are possible.
Usually, you should avoid repeatedly accessing web data if you can help it.
Instead, you should download a file directly and use that.
Still, accessing websites programmatically is a frequent tool in many real-world problems.