O'REILLY®

Data Loading

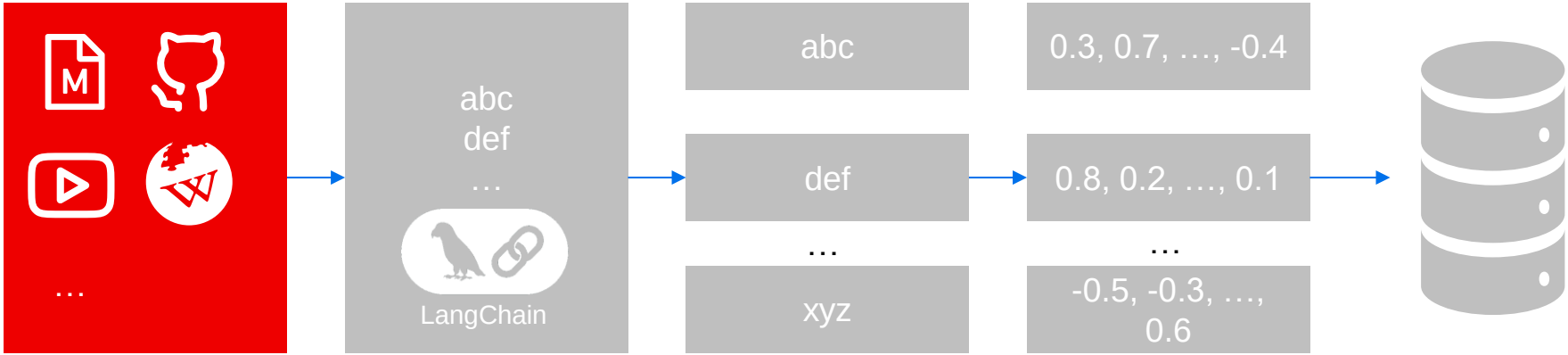
Bert Gollnick

Learning Objectives

By the end of this module, you will:

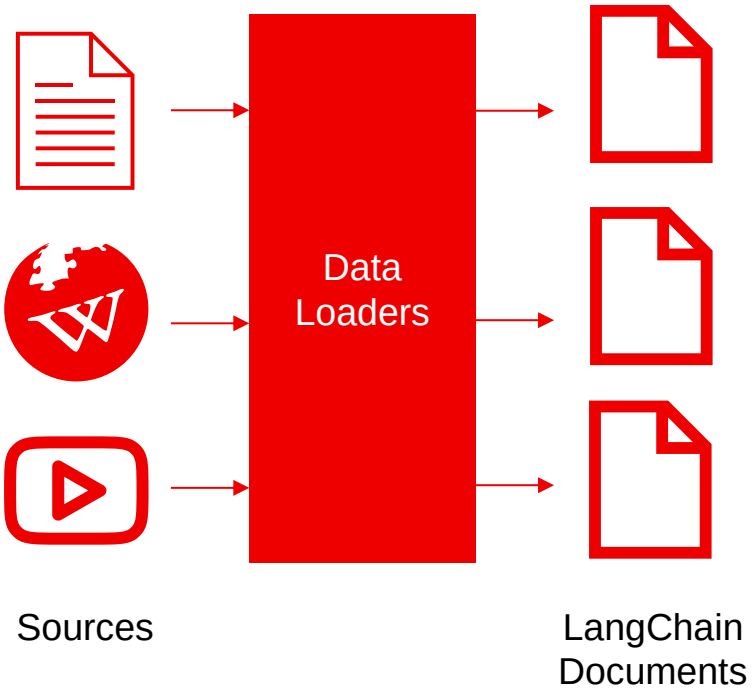
- know how to load data
- know which different formats are supported
- learn to work with documentation

Data Ingestion Process



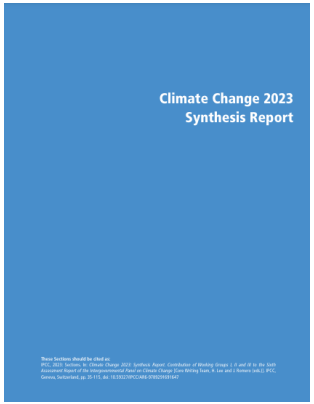
Data Loading Overview

- Hundreds of different data sources are supported by LangChain
- DataLoader returns list of LangChain documents
- Documents have two attributes
  - metadata
  - page_content



Use Case: PDF-file

- Extract text content from PDF-file



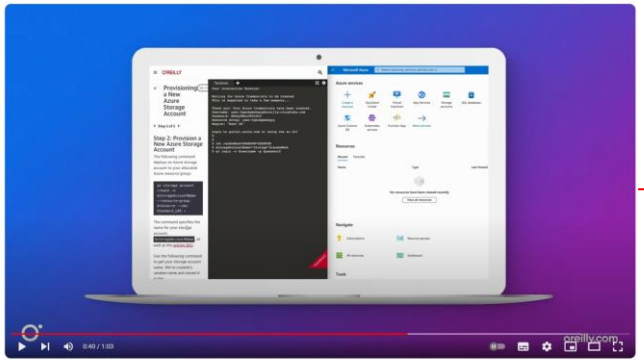
```
('35Climate Change 2023\n'
 'Synthesis Report\n'
 'IPCC, 2023: Sections. In: Climate Change 2023: Synthesis Report. '
 'Contribution of Working Groups I, II and III to the Sixth \n'
 'Assessment Report of the Intergovernmental Panel on Climate Change  [
 'Writing Team, H. Lee and J. Romero (eds.)]. IPCC, \n'
 'Geneva, Switzerland, pp. 35-115, doi: 10.59327/IPCC/AR6-9789291691647
 'Sections should be cited as: ')
```

„Climate Change 2023 Synthesis Report“
Source:
ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf

File Content

Use Case: Youtube Video

- Extract video transcript from youtube video



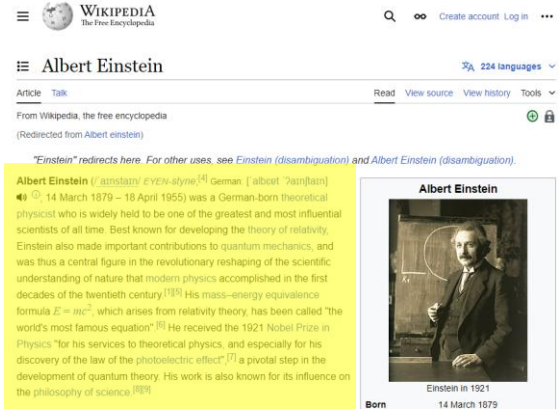

```
("[Music]  for over 40 years Tech teams have turned  to O'Reilly books for "
 "answers they can  trust today over 5,000 companies count  on the O'Reilly "
 "learning platform to  help their team stay ahead of what's  next there are "
 'live online courses and  Tech conferences so your teams get in  the room '
 'with experts on software  architecture AI the cloud security and  more all '
 'without the travel cost with  interactive labs and sandboxes your  teams get '
 'hands on experience with Cloud  platforms like Azure and AWS as well as  '
 'python kubernetes Java and more all in a  safe Dev environment so they learn '
 "what  to expect before trying it in the real  world and then there's "
 "O'Reilly answers  just ask any Tech question and it  instantly scans "
 'thousands of trusted  titles and videos to provide a solution  so they can '
 'find what they need and get  right back to work help your team stay  ahead '
 "of what's next visit oy.com to  learn more")
```

Youtube video „O'Reilly Learning Platform“
Source: youtube.com/watch?v=iFK7iyBpzxY

Video Transcript

Use Case: Wikipedia Article

- Extract content from Wikipedia article



```
('Albert Einstein ( EYEN-styne; German: [ˈalbɛɐt ˈʔaɪnʃtaɪn] ; 14 March 1879 – '
 '18 April 1955) was a German-born theoretical physicist who is widely held to '
 'be one of the greatest and most influential scientists of all time. Best '
 'known for developing the theory of relativity, Einstein also made important '
 'contributions to quantum mechanics, and was thus a central figure in the '
 'revolutionary reshaping of the scientific understanding of nature that '
 'modern physics accomplished in the first decades of the twentieth century. '
 'His mass–energy equivalence formula E = mc2, which arises from relativity '
 'theory, has been called "the world\'s most famous equation". He received the '
 '1921 Nobel Prize in Physics "for his services to theoretical physics, and '
 'especially for his discovery of the law of the photoelectric effect", a '
 'pivotal step in the development of quantum theory. His work is also known '
 'for its influence on the philosophy of science.\n'
```

Wikipedia article on „Albert Einstein“
Source: https://en.wikipedia.org/wiki/Albert_Einstein

Wikipedia Article Content

Iterate over a folder with documents

- Iterate over a folder:
  - extract one data type
  - automatically detect data type

```
from langchain_community.document_loaders
import DirectoryLoader
```

```
from langchain_community.document_loaders
import UnstructuredFileLoader

docs = []
for file_path in file_paths:
    print(file_path)
    loader =
UnstructuredFileLoader(file_path)
    docs.append(loader.load())
```


O'REILLY®