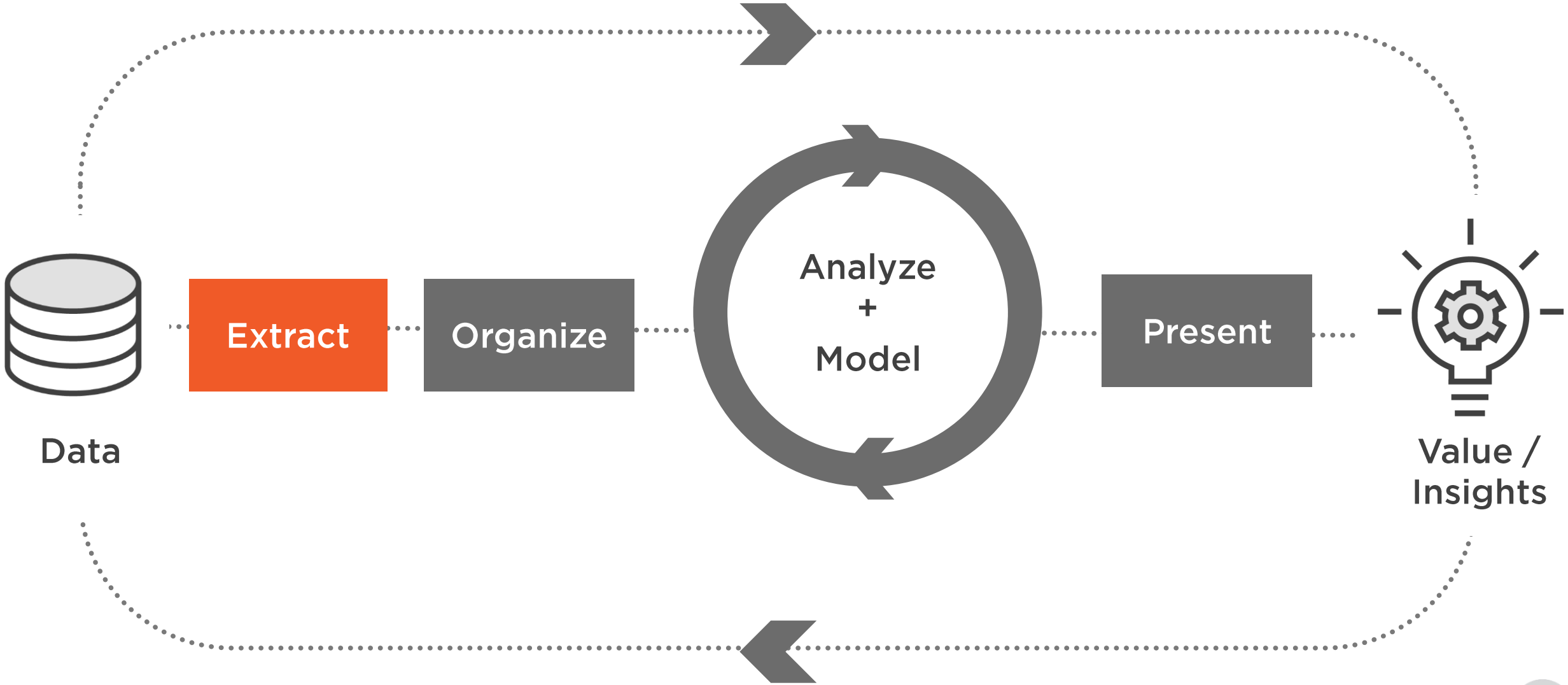
Extracting Data

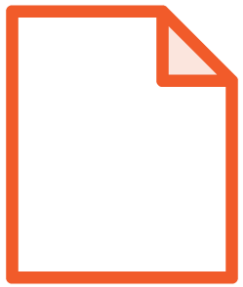
Abhishek Kumar
AUTHOR

@meabhishekkumar

Data Science Project Cycle



Data Source



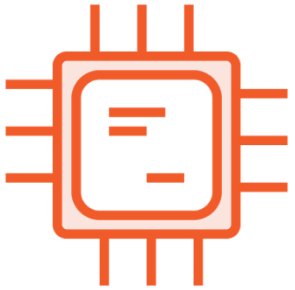
File

Website

API

Database

Forms / Surveys



Device

Overview (Concepts)

- Database
- API
- Web Scraping
- Titanic dataset
- Public datasets

Overview (Tools)

Requests

BeautifulSoup

Database interaction using sqlite3, PyMySQL, pymssql

Extracting Data from Databases

Database

SQLite

Self-contained, cross platform

MySQL

Open source, enterprise level

SQL Server

Proprietary, Microsoft

```
import package_name

connection =
package_name.connect(con_str)

cursor =
package_name.cursor()


Cursor.execute(query)


cursor.fetchall()


Connection.close()
```

◀ Import the package

◀ Connect to the database

◀ Create the cursor

◀ Execute query

◀ Fetch results

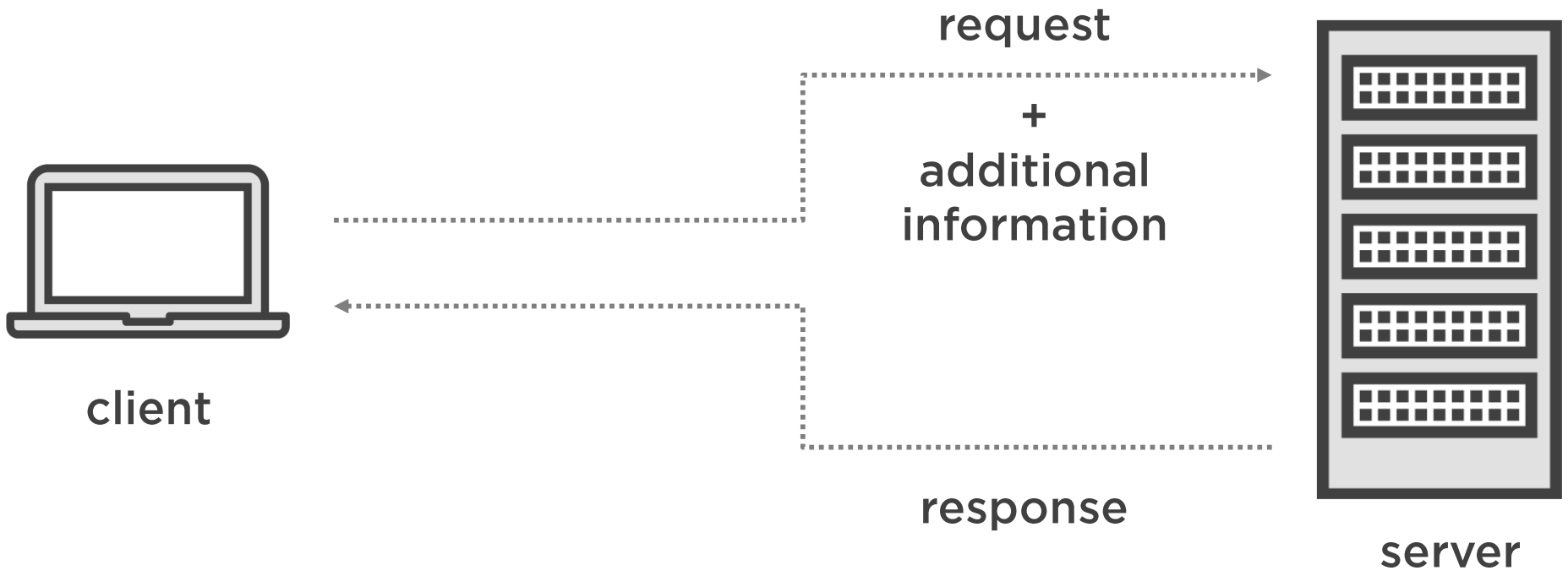
◀ Close the connection

Demo

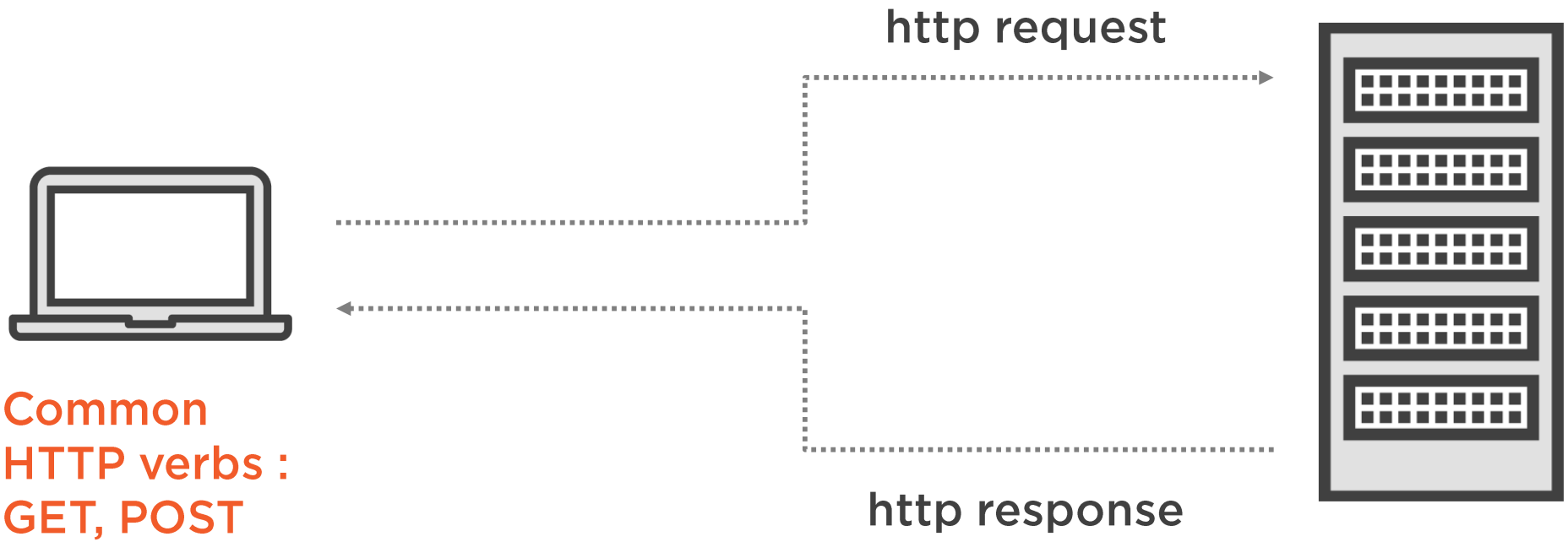
Extracting data from databases

Extracting Data Through APIs

Application Programming Interface (API)



REST API



Common
HTTP verbs :
GET, POST

Requests

```
import requests


result = requests.get(url)


result.status_code
result.headers
result.text
result.encoding
```

- Import package
- Use get function to make GET request
- Status code
- Header information
- Response text
- Response encoding

Demo

Extracting data from API using Requests

Extracting Data Using Web Scraping

Web Page

Document Object Model (DOM)

```
<!DOCTYPE HTML>

<html>

<head>
...
</head>

<body>
  <div class=”col-md-6">
  </div>

</body>

</html>
```

Demo

Web scraping using Requests and BeautifulSoup

Demo

Getting titanic dataset using Requests

Demo

Creating reproducible script for getting titanic data

Public Datasets

Important Links

- https://www.data.gov/
- https://aws.amazon.com/public-datasets/
- http://archive.ics.uci.edu/ml/datasets.html
- https://github.com/caesar0301/awesome-public-datasets

Demo

Commit changes to git

Summary

Data extraction

- Database
- API
- Web scraping
- Titanic disaster dataset

Public datasets