

Fetching data:

APIs & web scraping

- JSON format
- APIs
- Scraping web pages

- JSON is a lightweight data-interchange format
- JSON format passes data as a **string**
- Stands for (JavaScript Object Notation)
- JSON is language independent *
 - JSON is "self-describing" and easy to understand
- It would be just another format but has some useful properties

JSON – is all about **key : value** pairs

4

- In this example **employees**, **firstName** and **lastName** are the values

```
{"employees": [  
  {"firstName":"John", "lastName":"Doe"},  
  {"firstName":"Anna", "lastName":"Smith"},  
  {"firstName":"Peter", "lastName":"Jones"}  
]}
```

In this example **John**, **Anna**, **Peter** and **Doe**, **Smith** and **Jones** are **values**

- APIs allow people to interact with the structures of an application to get, put, delete, or update data
- Data is often made available on the web via website APIs
- Best practices for APIs are to use RESTful principles

RESTful APIs include:

- A Base URL and collection
- An media type (usually JSON)
- Operations (GET, PUT, POST, DELETE) using http requests

Operation



GET <https://api.instagram.com/v1/users/10>

Collection



Operation



GET <https://api.instagram.com/v1/users/search/?q=andy>



Query string

- Most APIs have language specific libraries to make the API easier to access through the language of your choice
 - Python libraries are typically among the ones available
- <http://www.pythonapi.com/>
- RESTful APIs can always be accessed using cURL or Postman requests

- Web scraping is:
 - Extracting information from websites (simulates a human copying and pasting)
 - Based on finding patterns in website code (usually HTML)
- What are best practices for web scraping?
 - Scraping too many pages too fast can get your IP address blocked
 - Pay attention to the robots exclusion standard (robots.txt)- Let's look at <http://www.imdb.com/robots.txt>

- HTML is interpreted by a web browser to produce ("render") a web page
 - Let's look at data/example.html
 - Tags are opened and closed
 - Tags have optional attributes
- How to view HTML code:
 - To view the entire page: "View Source" or "View Page Source" or "Show Page Source"
 - To view a specific part: "Inspect Element"
 - Safari users: Safari menu, Preferences, Advanced, Show Develop menu in menu bar