Getting Data

Data is stored in various ways

- Text files (.csv, .txt, .dat, ... everything you can open in a text editor)
- Other tables (.xls, .xlsx, etc.)
- Files of other programs (e.g. .dta)
  - Webpages (.html, .htm)
  - json
  - xml

Data is stored in various ways

- Text files (.csv, .txt, .dat, ... everything you can open in a text editor)
- Other tables (.xls, .xlsx, etc.)
- Files of other programs (e.g. .dta)
  - Webpages (.html, .htm)
  - json
  - xml
- Databases
- PDF files
- ...

! Windows users: Uncheck "Hide extensions for known file types" in "Folder Options" if you can't see them

Application Programming Interface (API)

> *A gateway drug that leads to pulling more data than rational, from systems gathering data faster than sensible, for reasons more aspirational than comprehensible, with feeling.* (The Devil's Data Dictionary)

Video: What is an API?

[1] **re**presentational **s**tate **t**ransfer

Application Programming Interface (API)

> *A gateway drug that leads to pulling more data than rational, from systems gathering data faster than sensible, for reasons more aspirational than comprehensible, with feeling.* (The Devil's Data Dictionary)

Video: What is an API?

- Usually we use REST[1]-ful API employed for web resources; mostly returns json formatted data
- Many RESTful APIs have python wrappers
- Most RESTful APIs require keys

[1] **re**presentational **s**tate **t**ransfer

Hierarchical data: json

```
{
    "firstName": "Jane",
    "lastName": "Doe",
    "hobbies": ["running", "sky diving", "singing"],
    "age": 35,
    "children": [
        {
            "firstName": "Alice",
            "age": 6
        },
        {
            "firstName": "Bob",
            "age": 8
        }
    ]
}
```

JSON turns into Python objects real quick

```python
import json  # to serialize (read) json
import requests  # to pull data from the internet

response = requests.get("https://jsonplaceholder.typicode.com/todos")
todos = json.loads(response.text)

type(todos)
```

- What type is `todos`?
- What type are the elements of `todos`?
- What is the title of the third entry?
- What is the completed-status for the last entry?

Hierarchical data: xml

```
<breakfast_menu>
 <food>
  <name>Belgian Waffles</name>
  <price>5.95</price>
  <description>
   Two of our famous Belgian Waffles with plenty of real ma
  </description>
  <calories>650</calories>
 </food>
 <food>
  <name>Strawberry Belgian Waffles</name>
  <price>7.95</price>
  <description>
   Light Belgian waffles covered with strawberries and whip
  </description>
  <calories>900</calories>
 </food>
</breakfast_menu>
```

xml is really bad to work with

```
1  import xml.etree.ElementTree as ET
```

... rather hope you don't need to work with it!

wikipedia

Documentation: https://wikipedia.readthedocs.io/en/stable/code.html#api

wikipedia

Documentation: https://wikipedia.readthedocs.io/en/stable/code.html#api

```
1 import pandas as pd
2 import wikipedia
3
4 query = "LMU Munich"
5 res = wikipedia.search(query, suggestion=True)
```

wikipedia

Documentation: https://wikipedia.readthedocs.io/en/stable/code.html#api

```python
import pandas as pd
import wikipedia

query = "LMU Munich"
res = wikipedia.search(query, suggestion=True)
```

- What type is object "res"
- Use `wikipedia.page()` to open the page corresponding to our search query!
- What is the full wikipedia title?
- How do you get the page's summary?
- What are the geographic location information of the page?

genderize

Documentation:
https://github.com/SteelPangolin/genderize

genderize

Documentation:
https://github.com/SteelPangolin/genderize

```
1  import pandas as pd
2  from genderize import Genderize
3
4  name = "Kim"
5  res = Genderize().get([name])
```

genderize

Documentation:
https://github.com/SteelPangolin/genderize

```
1  import pandas as pd
2  from genderize import Genderize
3
4  name = "Kim"
5  res = Genderize().get([name])
```

- Is Kim estimated to be male or female?

googlemaps

Documentation: https://github.com/googlemaps/google-maps-services-python

googlemaps

Documentation: https://github.com/googlemaps/google-maps-services-python

```
1  import googlemaps as gm
2  import pandas as pd
3
4  API_KEY = '...'
5  gmaps = gm.Client(API_KEY)
6
7  query = "LMU Munich"
8  res = gmaps.geocode(query)
```

googlemaps

Documentation: https://github.com/googlemaps/google-maps-services-python

```
1  import googlemaps as gm
2  import pandas as pd
3
4  API_KEY = '...'
5  gmaps = gm.Client(API_KEY)
6
7  query = "LMU Munich"
8  res = gmaps.geocode(query)
```

- What type is object "res"?
- What type is the first element of object "res"?
- What is the formatted address of LMU Munich?
- At which longitude is LMU Munich located?