# Gathering

The two main ways to work with HTML files are:

* Saving the HTML file to your computer (using the [**Requests**](http://docs.python-requests.org/en/master/) library for example) library and reading that file into a **[BeautifulSoup](https://www.crummy.com/software/BeautifulSoup/" \t "_blank)** constructor
* Reading the HTML response content directly into a BeautifulSoup constructor (again using the Requests library for example)

More Information:

<https://www.crummy.com/software/BeautifulSoup/>

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/#searching-the-tree>

<https://kunststube.net/encoding/>

<https://stackabuse.com/reading-and-writing-json-to-a-file-in-python/>

# Assessing Data

Data quality dimensions help guide your thought process while assessing and also cleaning. The four main data quality dimensions are:

* **Completeness**: do we have all of the records that we should? Do we have missing records or not? Are there specific rows, columns, or cells missing?
* **Validity**: we have the records, but they're not valid, i.e., they don't conform to a defined schema. A schema is a defined set of rules for data. These rules can be real-world constraints (e.g. negative height is impossible) and table-specific constraints (e.g. unique key constraints in tables).
* **Accuracy**: inaccurate data is wrong data that is valid. It adheres to the defined schema, but it is still incorrect. Example: a patient's weight that is 5 lbs too heavy because the scale was faulty.
* **Consistency**: inconsistent data is both valid and accurate, but there are multiple *correct* ways of referring to the same thing. Consistency, i.e., a standard format, in columns that represent the same data across tables and/or within tables is desired.

You can assess data for:

* Quality: issues with content. Low quality data is also known as dirty data.
* Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:
  1. Each variable forms a column.
  2. Each observation forms a row.
  3. Each type of observational unit forms a table.

...using two types of assessment:

* Visual assessment: scrolling through the data in your preferred software application (Google Sheets, Excel, a text editor, etc.).
* Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).