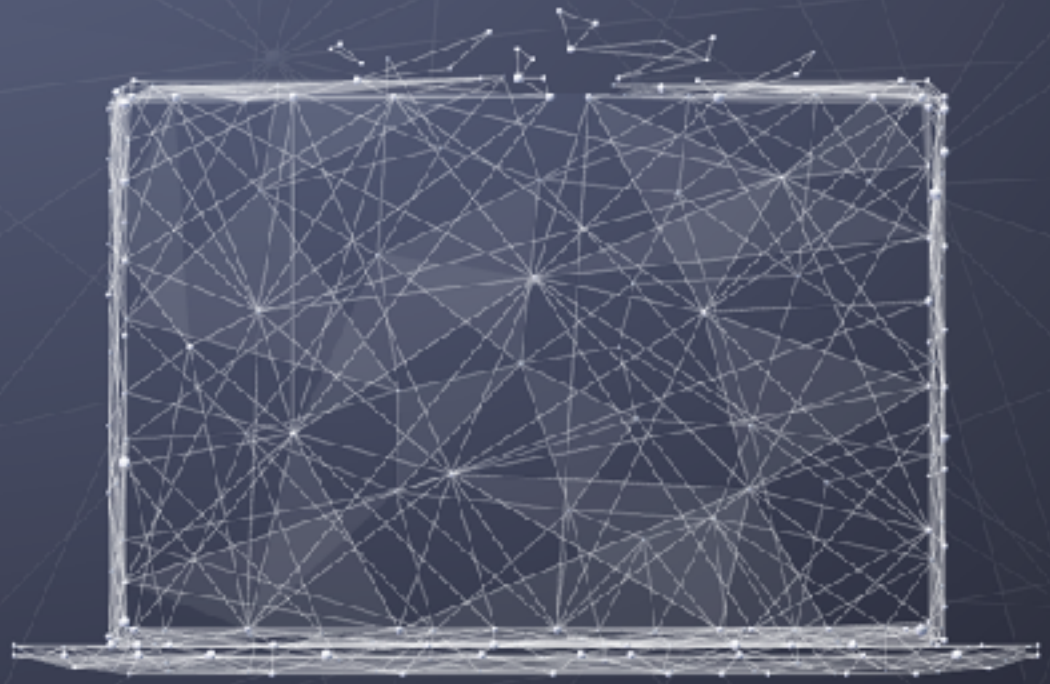
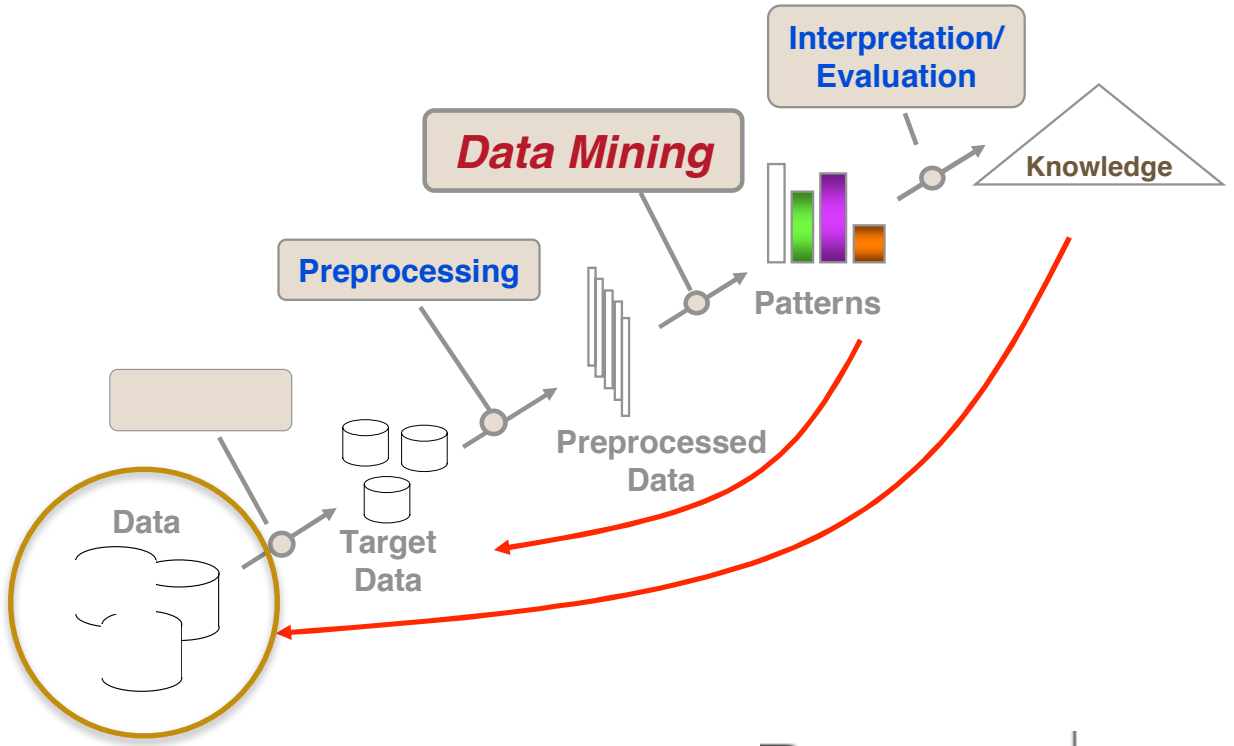
Data Science
Data Engineering I

What is data?

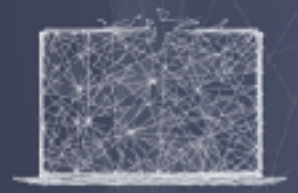


PURDUE UNIVERSITY | College of Science

First step: obtain data



PURDUE UNIVERSITY | College of Science



What is data?

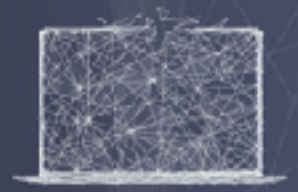
- Collection of entities and their attributes
- **Attribute**: property or characteristic of an entity (e.g., eye color, temperature)
- **Entity**: collection of attributes
  Aka: record, point, case, sample, object, or instance.

Attributes

Entities

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1013	Flat
M8	1.25	5g	8	Pan	$11.08	No	295	Phillips
M10	1.5	6g	10	Round	$15.74	Yes	485	Phillips
M12	1.75	7g	12	Pan	$18.20	No	995	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$28.09	Yes	486	Both
M24	2.85	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.86	No	1067	Phillips
M30	3.2	12g	30	Pan	$41.32	No	494	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

PURDUE UNIVERSITY | College of Science



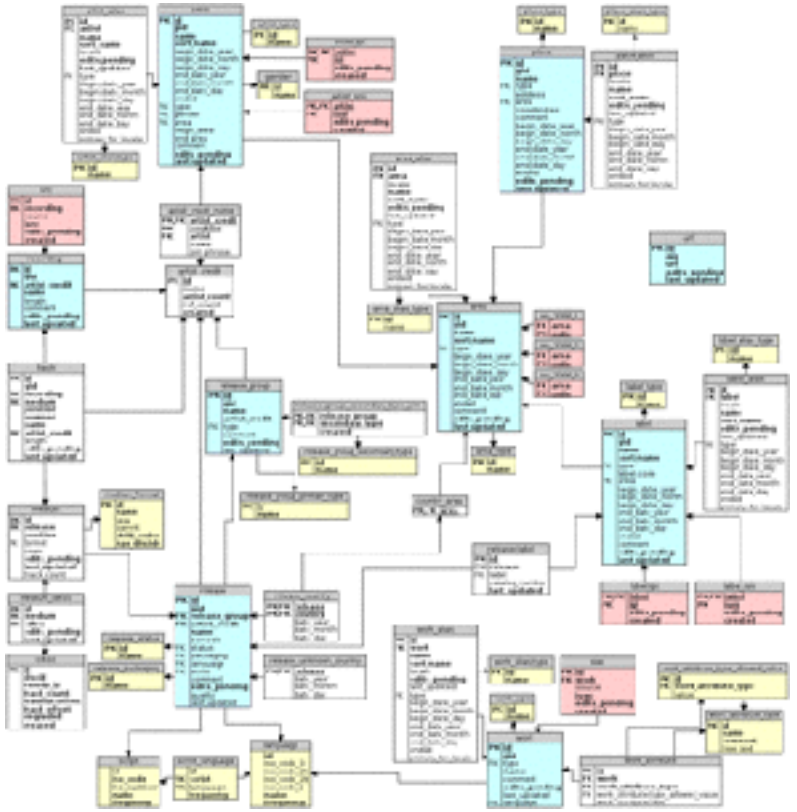
Tabular data (Simple structured data)

Collection of records, each of which consists of a fixed set of attributes

position #	time [UTC]	time [EST]	Longitude [deg]	Latitude [deg]	Altitude [m]	Status	Main [V]	Beacon [V]	Temperature [°C]
1	2014-01-01 00:00:41	2014-01-01 02:00:41	32.24120	-27.88007	119	val. GPS-3D	3.36	2.00	25
2	2014 01 01 01:00:44	2014 01 01 03:00:44	32.24422	27.88077	110	val. GPS 3D	3.36	2	24
3	2014 01 01 02:00:44	2014 01 01 04:00:44	32.24427	27.88084	113	val. GPS 3D	3.36	2	24
4	2014-01-01 03:00:31	2014-01-01 05:00:31	32.24487	-27.88001	106	val. GPS-3D	3.36	1.84	23
5	2014-01-01 04:01:07	2014-01-01 06:01:07	32.24430	-27.87950	121	GPS-3D	3.36	1.92	23
6	2014-01-01 05:00:51	2014-01-01 07:00:51	32.23727	-27.87759	121	GPS-2D	3.36	2.21	26
7	2014 01 01 06:01:35	2014 01 01 08:01:35	32.23747	27.87750	122	GPS 2D	3.36	2.32	26
8	2014 01 01 07:01:20	2014 01 01 09:01:20	32.23740	27.87713	124	GPS 2D	3.36	2.04	28
9	2014-01-01 08:01:17	2014-01-01 10:01:17	32.23733	-27.87697	97	GPS-3D	3.36	2.72	29
10	2014-01-01 09:01:14	2014-01-01 11:01:14	32.24147	-27.87177	80	val. GPS-3D	3.36	2.72	29
11	2014-01-01 10:01:20	2014-01-01 12:01:20	32.24507	-27.87179	84	GPS-2D	3.36	2.41	27
12	2014 01 01 11:01:01	2014 01 01 13:01:01	32.24400	27.87202	83	GPS 2D	3.36	2.24	26
13	2014 01 01 12:01:30	2014 01 01 14:01:30				No Fix	3.36	2.32	26
14	2014-01-01 13:02:34	2014-01-01 15:02:34	32.24191	-27.87203	80	val. GPS-3D	3.36	2.64	29
15	2014-01-01 14:01:01	2014-01-01 16:01:01	32.24068	-27.87277	81	val. GPS-3D	3.36	2.24	26
16	2014-01-01 15:01:02	2014-01-01 17:01:02	32.23830	-27.87517	91	GPS-3D	3.36	2.64	29
17	2014 01 01 16:01:13	2014 01 01 18:01:13	32.23813	27.87820	104	val. GPS 3D	3.36	2.04	20
18	2014 01 01 17:00:10	2014 01 01 19:00:10	32.23720	27.87853	100	val. GPS 3D	3.36	2.32	26
19	2014-01-01 18:01:38	2014-01-01 20:01:38	32.23773	-27.87838	113	val. GPS-3D	3.36	2.16	23
20	2014-01-01 19:01:07	2014-01-01 21:01:07	32.24157	-27.87658	106	val. GPS-3D	3.36	2	24
21	2014-01-01 20:01:51	2014-01-01 22:01:51	32.24133	-27.87617	101	val. GPS-3D	3.36	1.92	23

Structured data (relational)

Collection of data items, each with varying properties, as well as relations among the items, e.g., relational databases, graph data (organization of information is specified by data schema/model)



Semi-structured data

Collection of data items with less formal structure than relational data, but with some tags or markers to delineate semantic elements of items (i.e., properties), e.g., HTML, XML, or JSON

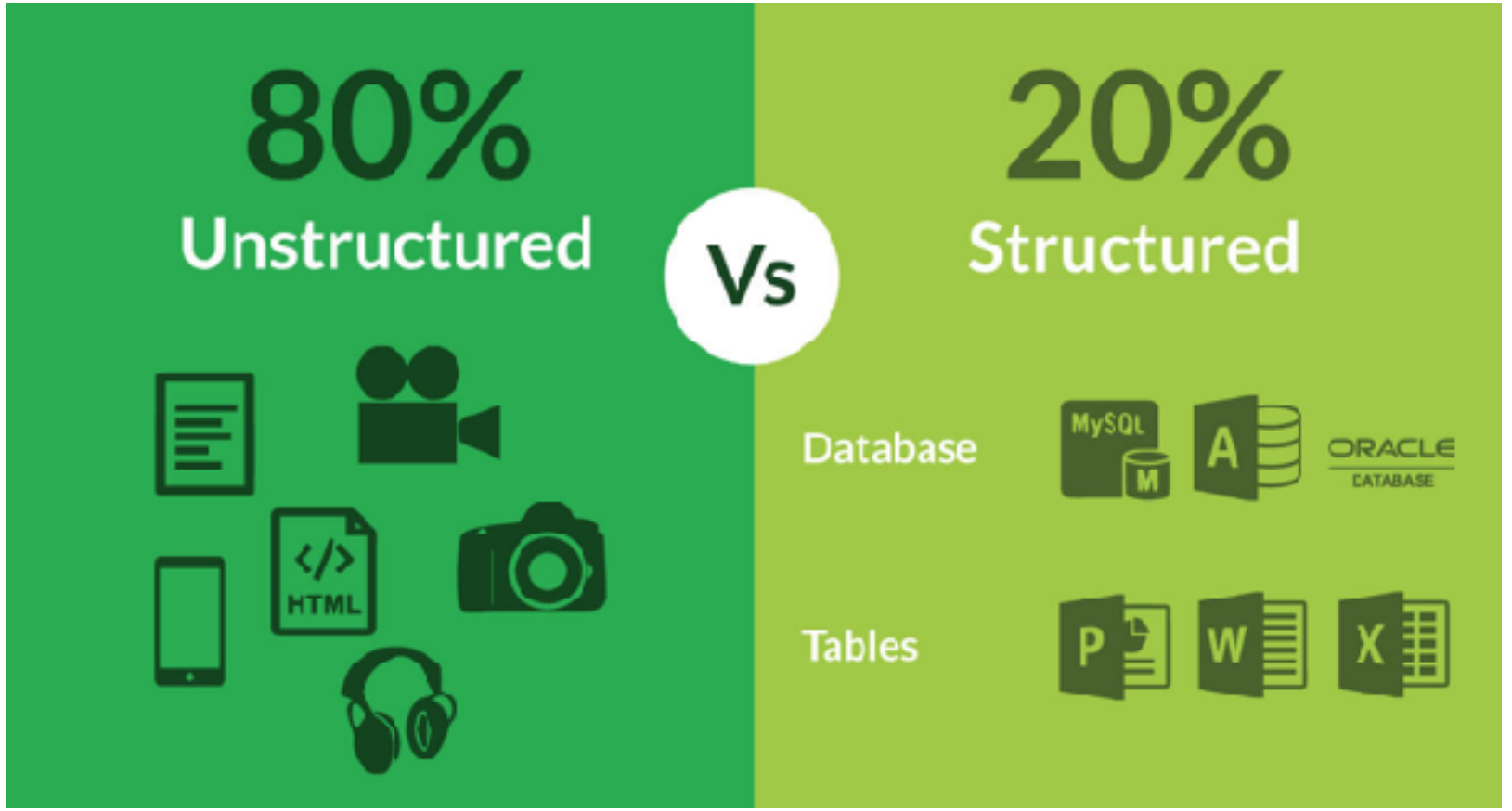
```
<user-agents>
  <user-agent>
    <ID>id_a_f_3</ID>
    <String>!Susie (http://www.sync2it.com/susie)</String>
    <Description>Sync2It bookmark management & clustering engine</Description>
    <Type>C R</Type>
    <Comment/>
    <Link1>http://www.sync2it.com</Link1>
    <Link2/>
  </user-agent>
  <user-agent>
    <ID>id_a_f_6</ID>
    <String>
      <a href='http://www.unchaos.com/'> UnChaos </a> From Chaos To Order Hybrid '
    </String>
    <Description>UnCHAOS search robot</Description>
    <Type>R</Type>
    <Comment>Site is dead</Comment>
    <Link1>http://www.unchaos.com/</Link1>
    <Link2/>
  </user-agent>
```

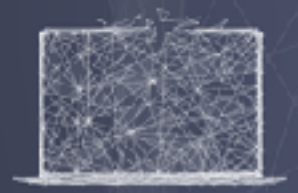

```
{
  "business_id": "PK6aSizckHFWk8i0oxt5DA",
  "full_address": "400 Waterfront Dr E\nHomestead\nHomestead, PA 15120",
  "hours": {},
  "open": true,
  "categories": [
    "Burgers",
    "Fast Food",
    "Restaurants"
  ],
  "city": "Homestead",
  "review_count": 5,
  "name": "McDonald's",
  "neighborhoods": [
    "Homestead"
  ],
  "longitude": -79.910032,
  "state": "PA",
  "stars": 2,
  "latitude": 40.412086,
  "attributes": {
    "Take-out": true,
    "Wi-Fi": "free",
    "Drive-Thru": true,
    "Good For": {
```

Unstructured data

Data that is not organized according to a predefined structure (e.g., free text, videos, photos, music, messages, etc.)

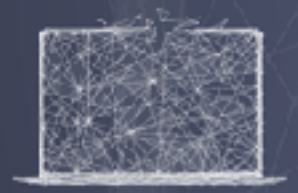
Lack of structure makes it more difficult to search for and analyze patterns





Common data formats

- CSV (comma separate values)
- JSON (Javascript object notation)
- HTML/XML (hypertext markup language / extensible markup language)
- SQL and NoSQL databases (SQL=structured query language)



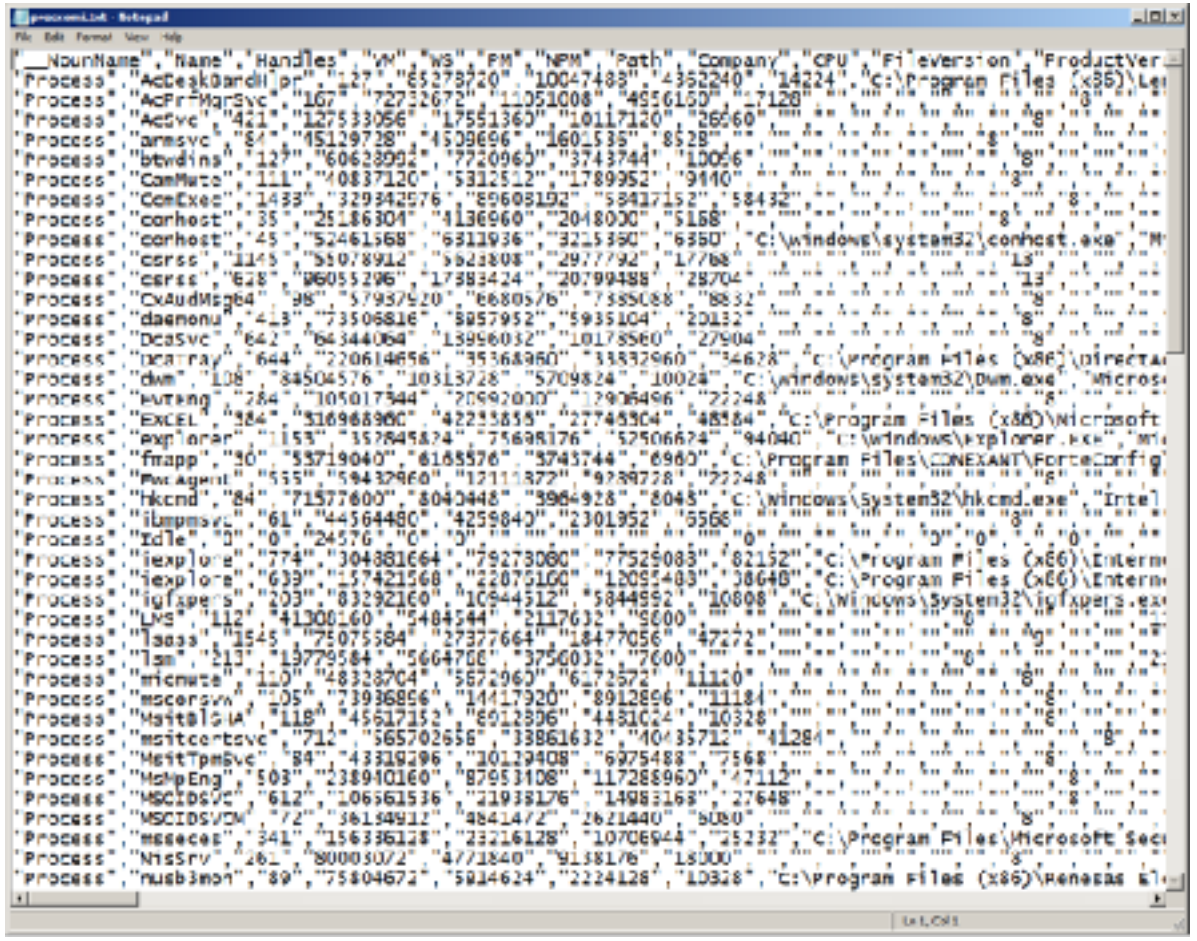
Initial data tasks

Read data into internal memory data structure

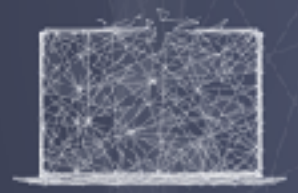
- Read in data from file, parse, and store in internal data structure
- Data structures for tabular data:
  - Single example can be stored in:
    - Vector or list
    - Dictionary (key=attribute name, value=attribute value)
  - Set of examples can be stored in:
    - Matrix or list of lists/dictionaries
    - Dictionary (key=example id, value=example data)
    - Data frame (table-like data structure in Pandas)

CSV Files

- Typically refers to any delimited text file (e.g., comma, tab, space)
- Values containing commas can be enclosed in quotes to delineate
- Python has built in methods to parse csv files. We will use Pandas —a Python library for data analysis



PURDUE UNIVERSITY | College of Science



XML/HTML Files

- The main format for the web
- XML files contain hierarchical content delineated by tags
- HTML is syntactically like XML, but sometimes tags are not closed and tags are primarily used to describe appearance of page
- There are a number of Python parsers for XML/HTML. We will use the BeautifulSoup library.

```
▼<user-agents>
 ▼<user-agent>
    <ID>id_a_f_3</ID>
    <String>!Susie (http://www.sync2it.com/susie)</String>
    <Description>Sync2It bookmark management & clustering engine</Description>
    <Type>C R</Type>
    <Comment/>
    <Link1>http://www.sync2it.com</Link1>
    <Link2/>
  </user-agent>
 ▼<user-agent>
    <ID>id_a_f_6</ID>
   ▼<String>
      <a href='http://www.unchaos.com/'> UnChaos </a> From Chaos To Order Hybrid '
    </String>
    <Description>UnCHAOS search robot</Description>
    <Type>R</Type>
    <Comment>Site is dead</Comment>
    <Link1>http://www.unchaos.com/</Link1>
    <Link2/>
  </user-agent>
```

JSON Files

- JSON originated as a way of encapsulating Javascript objects
- JSON data looks much like a dictionary in Python, with keys and value stores
- Python has built in methods to parse JSON, but Pandas can also be used to read and parse JSON


```
{
  "business_id": "PK6aSizckHFWk8i0oxt5DA",
  "full_address": "400 Waterfront Dr E\nHomestead\nHomestead, PA 15120",
  "hours": {},
  "open": true,
  "categories": [
    "Burgers",
    "Fast Food",
    "Restaurants"
  ],
  "city": "Homestead",
  "review_count": 5,
  "name": "McDonald's",
  "neighborhoods": [
    "Homestead"
  ],
  "longitude": -79.910032,
  "state": "PA",
  "stars": 2,
  "latitude": 40.412086,
  "attributes": {
    "Take-out": true,
    "Wi-Fi": "free",
    "Drive-Thru": true,
    "Good For": {
      "dessert": false,
      "latenight": false,
      "lunch": false,
      "dinner": false,
      "breakfast": false,
      "brunch": false
    },
    "Caters": false,
    "Noise Level": "average",
    "Takes Reservations": false,
```

JSON Files

- JSON originated as a way of encapsulating Javascript objects
- JSON data looks much like a dictionary in Python, with keys and value stores
- Python has built in methods to parse JSON, but Pandas can also be used to read and parse JSON


```
{
  "business_id": "PK6aSizckHFWk8i0oxt5DA",
  "full_address": "400 Waterfront Dr E\nHomestead\nHomestead, PA 15120",
  "hours": {},
  "open": true,
  "categories": [
    "Burgers",
    "Fast Food",
    "Restaurants"
  ],
  "city": "Homestead",
  "review_count": 5,
  "name": "McDonald's",
  "neighborhoods": [
    "Homestead"
  ],
  "longitude": -79.910032,
  "state": "PA",
  "stars": 2,
  "latitude": 40.412086,
  "attributes": {
    "Take-out": true,
    "Wi-Fi": "free",
    "Drive-Thru": true,
    "Good For": {
      "dessert": false,
      "latenight": false,
      "lunch": false,
      "dinner": false,
      "breakfast": false,
      "brunch": false
    },
    "Caters": false,
    "Noise Level": "average",
    "Takes Reservations": false,
```