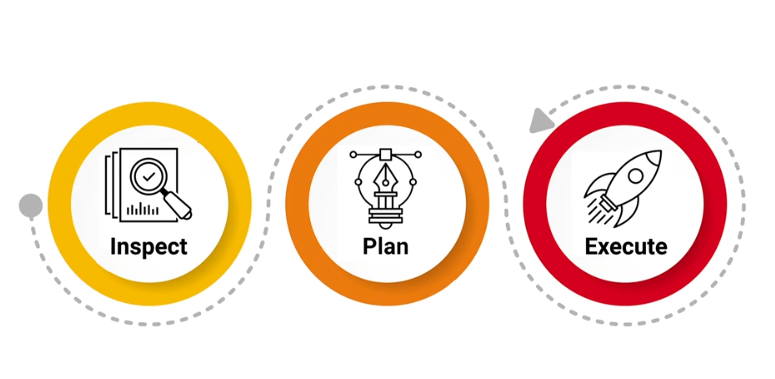
The data may need to be filtered, parsed, translated, sorted, interpolated, pivoted, summarized, aggregated, merged, or more. The goal is to create a consistent structure in the data. Without a consistent structure in our data, it's almost impossible to perform any meaningful analysis.

Eventually, we'll want to create an automated pipeline, which Jupyter Notebooks aren't suited for.

While technically that may be possible, the read\_json method that comes built into the Pandas library only works well for data that is already clean—for example, when the JSON data has every field filled in every time it is returned. We call data like this "flat."



Cleaning data

Three states of data

* Beyond Repair – Delete and move on
* Badly damaged – Options: Fill in in missing data, standardizing units, consolidate duplicated data
* Wrong form – Option: Reshape, convert data types, Parse text data to correct format, split columns

Generally good idea to remove large section of data at a time, so starting with beyond repair data makes a lot of sense.