Exploring Pandas Data Input Capabilities

Paweł Kordek
SOFTWARE ENGINEER

@pawel_kordek https://kordek.github.io

Overview

Formats that Pandas can handle

Tate dataset: CSV

Tate dataset: JSON

Pandas-combatible Data Formats

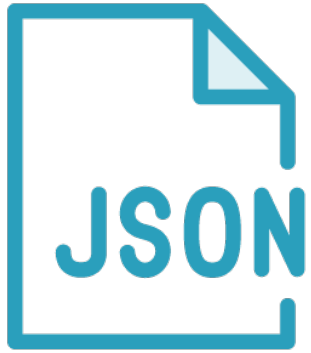
Different Data Sources

Text files

Binary files

Relational databases

Text Formats



Python objects!

Tate Collection Metadata: Take One

All artworks – one file

Let’s peek inside

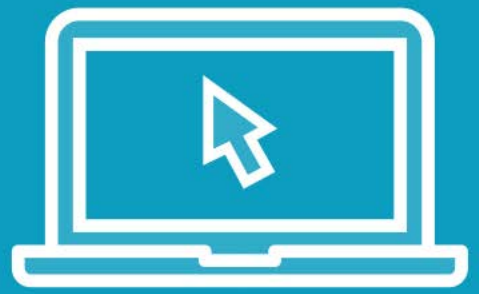
demos/collection-master/artwork_data.csv

```
,id,accession_number,artist,artistRole,artistId,title …

0,1035,A00001,"Blake, Robert",artist,38,A Figure Bowing …
1,1036,A00002,"Blake, Robert",artist,38,"Two Drawings …
2,1037,A00003,"Blake, Robert",artist,38,"The Preaching …
3,1038,A00004,"Blake, Robert",artist,38,Six Drawings …
4,1039,A00005,"Blake, William",artist,39,The Circle …
```

pd.read_csv('file_name.csv')

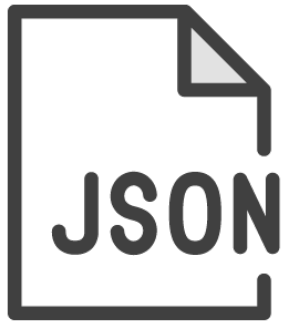
Demo



We will:

- Load CSV file with artworks
- Inspect the produced DataFrame
- Try to avoid reading reduntant data

Tate Collection Metadata: Take Two



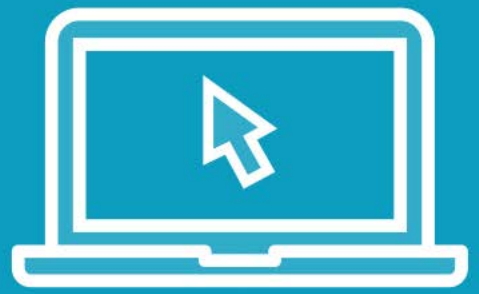
One artwork – one file

demos/collection-master/artworks/ ...


```
{
  "acquisitionYear": 1919,
  "all_artists": "William Blake",
  "catalogueGroup": {
    "shortTitle": "Illustrations to 'The Book of Job'"
  },
  "contributors": [
    {
      "birthYear": 1757,
      "date": "1757\u20131827",
      "displayOrder": 1,
      "fc": "William Blake"
    }
  ]
}
```


pd.DataFrame.from_*

Demo



We will:

- Traverse folders to get all JSON files
- Find the way to read them so that Pandas understands
- Try to produce identical DataFrame as for CSV