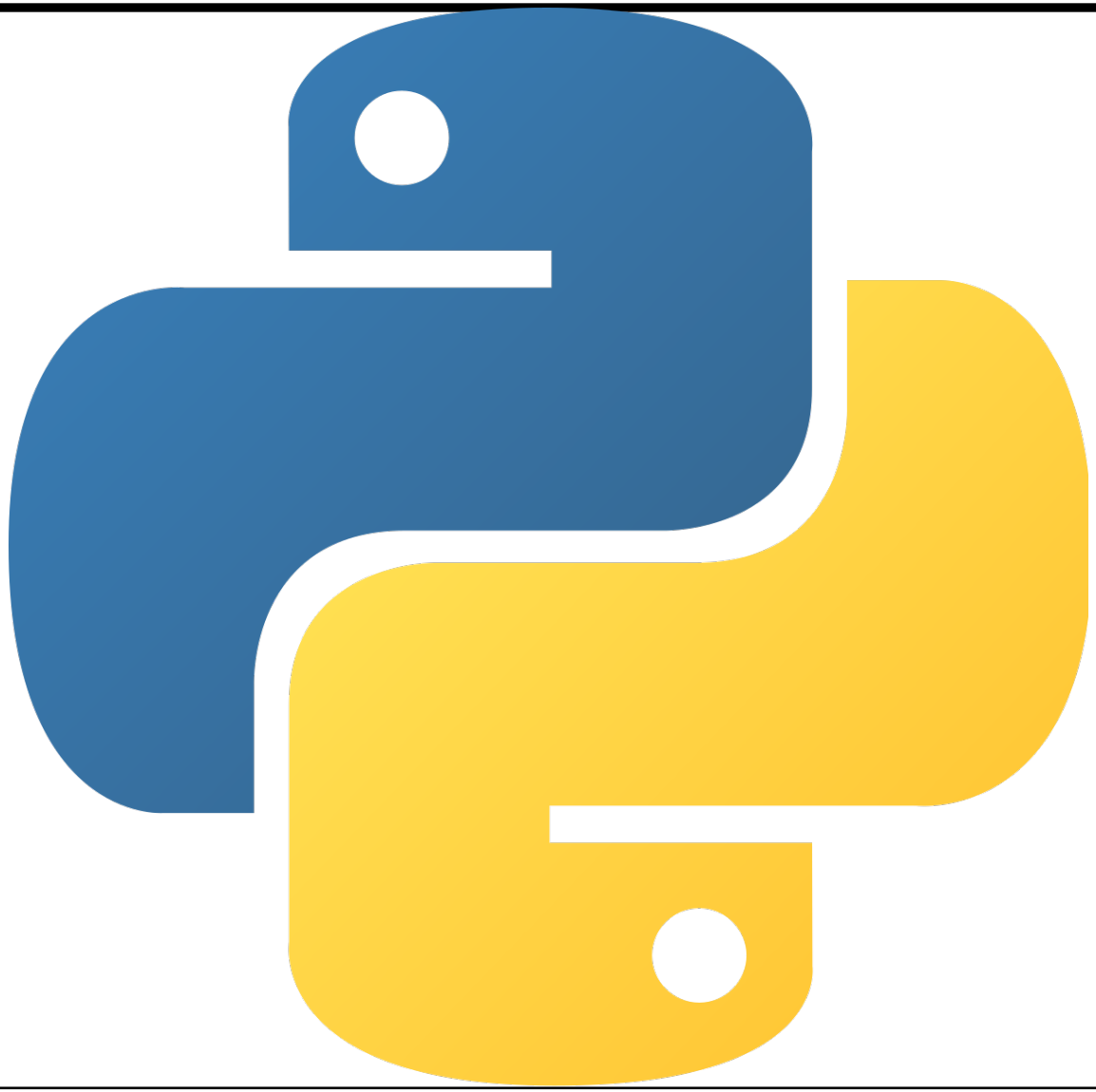
DAGENS FRÅGA

- Har du någonsin testat något du visste att du skulle vara riktigt dålig på, vad var det?

PYTHON PROGRAMMERING

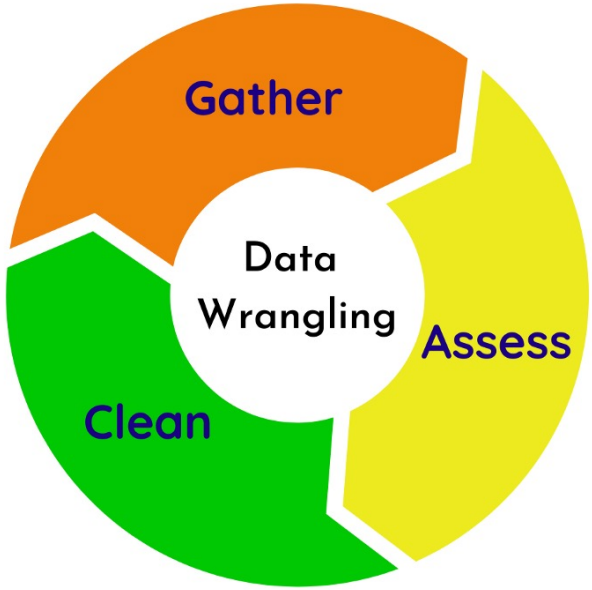


Föreläsning 8

DAGENS AGENDA

- Pathlib – bättre hantering av path
- Pandas och Dataframe

HANTERA MYCKET DATA



- En datatabell kan bestå av flera tusen rader och tiotals kolumner – hur kan man få en överblick?
- Ofta kräver datan även att man processerar den
  - Data cleaning: Att rensa bort korrupt eller ofullständig data
  - Data wrangling: Att transformera den råa datan till något mer användbart
- Hur kan man extrahera den datan som man är intresserad av?
- Hur kan man visualisera all eller en del av datan?

PANDAS

- Pandas är ett pythonbibliotek för att hantera, modifiera och analysera data och kan läsa in data i många olika format såsom csv, json, sql, microsoft excel
- Pandas är industristandard och många relevanta databibliotek bygger på att data lagras i *pandas dataframes*
- Pandas kan hantera både numerisk och kategorisk data

- Officiell dokumentation: https://pandas.pydata.org/docs/
- *OBS: Pandas är ett bibliotek som funnits länge. Vissa metoder kan vara utdaterade så använd gärna den officiella dokumentationen istället för **gamla** stack overflow-tips*

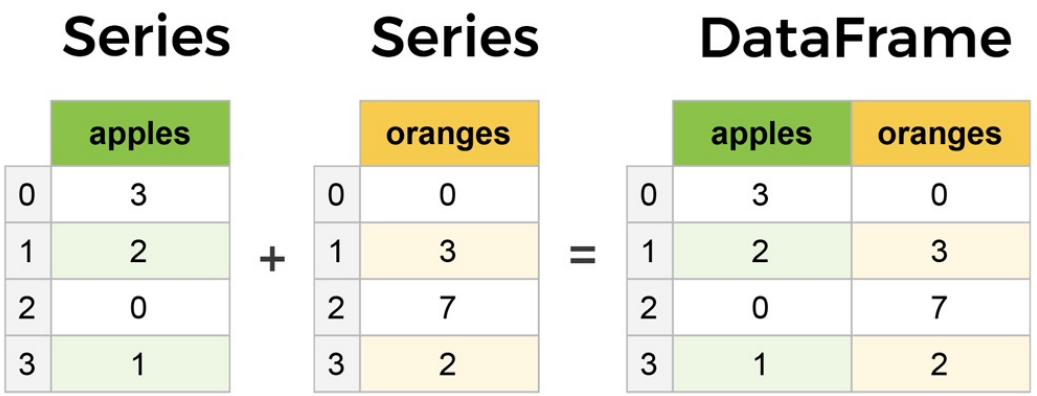
PANDAS - SERIES

- En series, pd.Series, är en endimensionell datavektor. Se det som exempel av enskilda värden av samma typ
- Kan skapas från många olika format: lists, tuples, dictionaries, numpy arrays
- Har ett **index** som bestämmer hur man kommer åt datan i vektorn
  - Normalt bara heltal men kan vara en sträng

Series

	apples
0	3
1	2
2	0
3	1

PANDAS - DATAFRAMES



- En tvådimensionell datastruktur som består av columns och rows
- Varje column är ett pandas Series objekt
- I en dataframe ska en isolerad rad ses som en datapunkt
- OBS: columns är överordnade rows. Detta skiljer sig från t.ex numpy arrays

PANDAS - INDEXERING

- Pandas stödjer många olika sätt för att komma åt subset av datan:
  - Både dataframes och series kan ses som dictionaries:
    - pd.DataFrame[’columnX’]
    - pd.Series[’indexX’]

  - *Indexers* används för att komma åt rader i en DataFrame
    - iloc: integer location
    - loc: location

  - Pandas stödjer boolean indexering vilket kan vara mycket användbart
    - df[df[X] > 0] returnerar de rader som har värden större än 0 i kolumn X