DQLab

Exploratory Data Analysis

Prepared by Ronny Fahrudin

My Profile

Riwayat Pekerjaan:

- Data Scientist @ S3 Innovate Pte. Ltd.
  - Develop Automation Doc-reader with OCR
- Mentor/Public Speaker TA DTS Kominfo X DQLab
- Assistant Mentor Data Scientist/Analyst/Engineer Digitalent Scholarship Kominfo X DQLab
- Data Scientist at Innovatz Solution
  - Develop Application Scorecard with Machine Learning @Pegadaian
  - Develop Dashboard Approval Credit @Pegadaian
- Analyst and Researcher at Evidensi
  - Helping MSME to analyze data for business growth
- Data Scientist Awardee at IYKRA Data Fellowship
- Develop Machine Learning to calculate Estimated Time Arrival Taxi at Bluebird

Contact Pengajar
Email : ronnyfahrudin@gmail.com
Medium : https://ronnyfahrudin.medium.com/
LinkedIn : https://www.linkedin.com/in/ronnyf/

Target Pembahasan

- Peserta Paham Tentang Data Science
- Peserta paham Exploratory Data Analysis dan Pentingnya hal itu
- Peserta Paham Parameter Penilaian Data
- Paham tools untuk exploratory data analysis

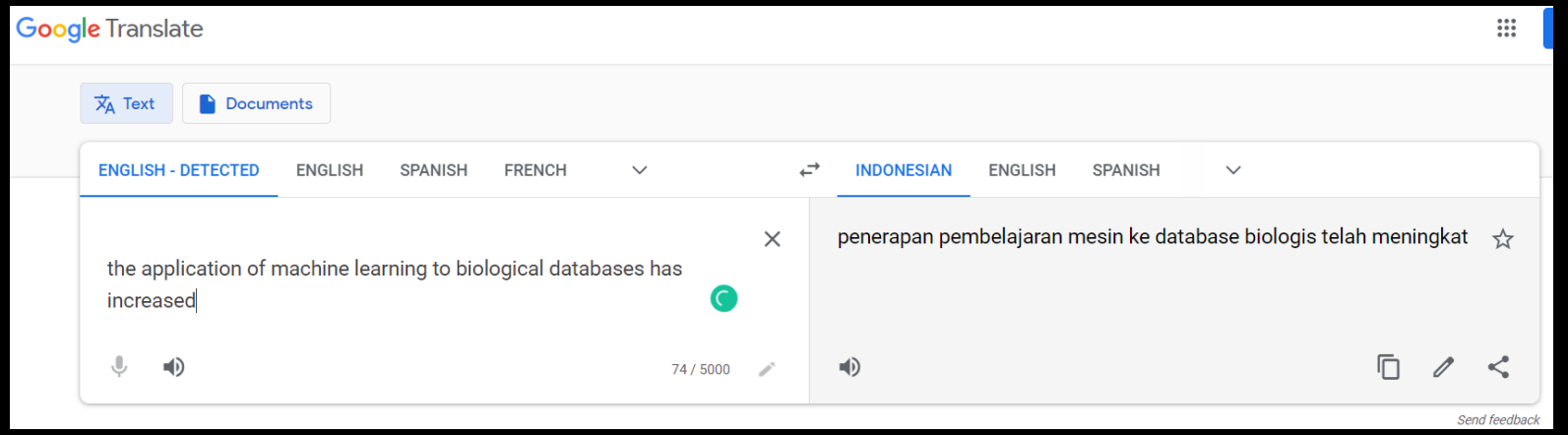
Outlines

No	Pembahasan	Hal
1	Apa itu data science apa dan mengapa ?	5
2	Apa dan mengapa exploratory data analysis (EDA)?	10
3	Bagaimana Konsep melakukan EDA dan menilai kualitas data?	13
4	Mengenal Tujuan Visualisasi dalam EDA	17
5	Pengenalan Tools EDA dalam Python - Basic Python - Numpy - Pandas - Scipy - Matplotlib, Seaborn	19

DΦLab

Apa, Mengapa, Bagaimana, Data Science?

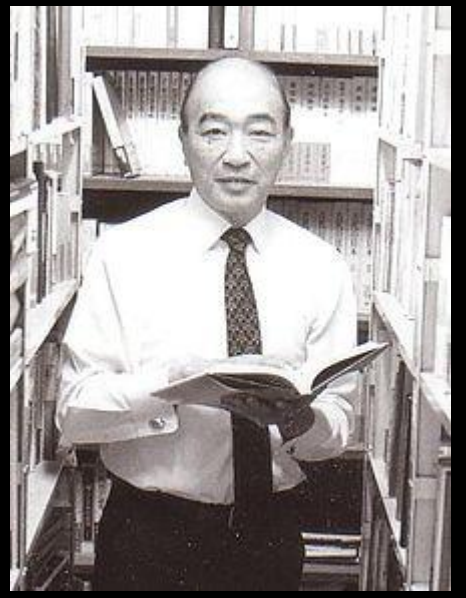
Apakah kalian pernah melihat?



Apa itu data science?

Expertise based

Chikio Hayashi dari Institut Statistika Matematika Sakuragaoka:
data science adalah ilmu pengetahuan interdisiplin tentang metode komputasi untuk mendapatkan wawasan berharga yang dapat ditindaklanjuti dari kumpulan data yang mencakup tiga fase yaitu desain data, mengumpulkan data, dan analisis data.



Japanese scientist and mathematician (1918-2002)

Mengapa data science ada?

Data science

Insigh/ solution

DQLab

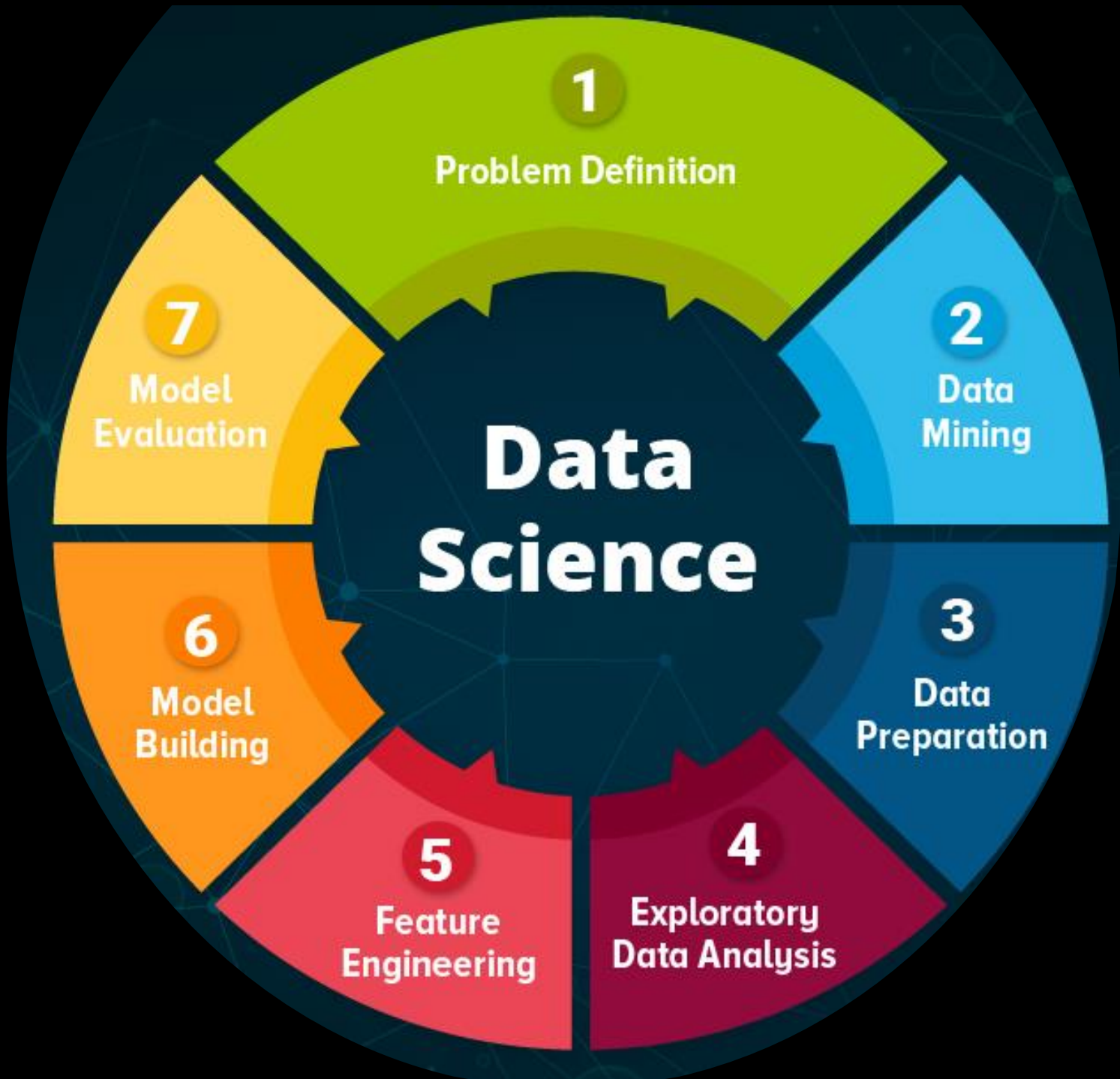
Apa, Mengapa Exploratory Data Analysis?

Apa itu EDA?

Exploratory data analysis adalah proses kritis dalam melakukan investigasi awal pada data untuk menemukan pola, untuk menemukan anomali, untuk menguji hipotesis dan untuk memeriksa asumsi dengan bantuan statistik ringkasan dan representasi grafis

Mengapa EDA?

- Untuk menemukan struktur atau wawasan yang tidak terduga dalam data
- Mengidentifikasi variable/feature yang penting dalam dataset
- Mengetes hypothesis atau cek asumsi-asumsi yang berhubungan dengan dataset
- Mengecek kualitas data yang selanjutnya di cleansing and processing
- Mencari korelasi2 or sebab akibat yang ada dalam data
- Menyampaikan wawasan tentang data ke stake holder

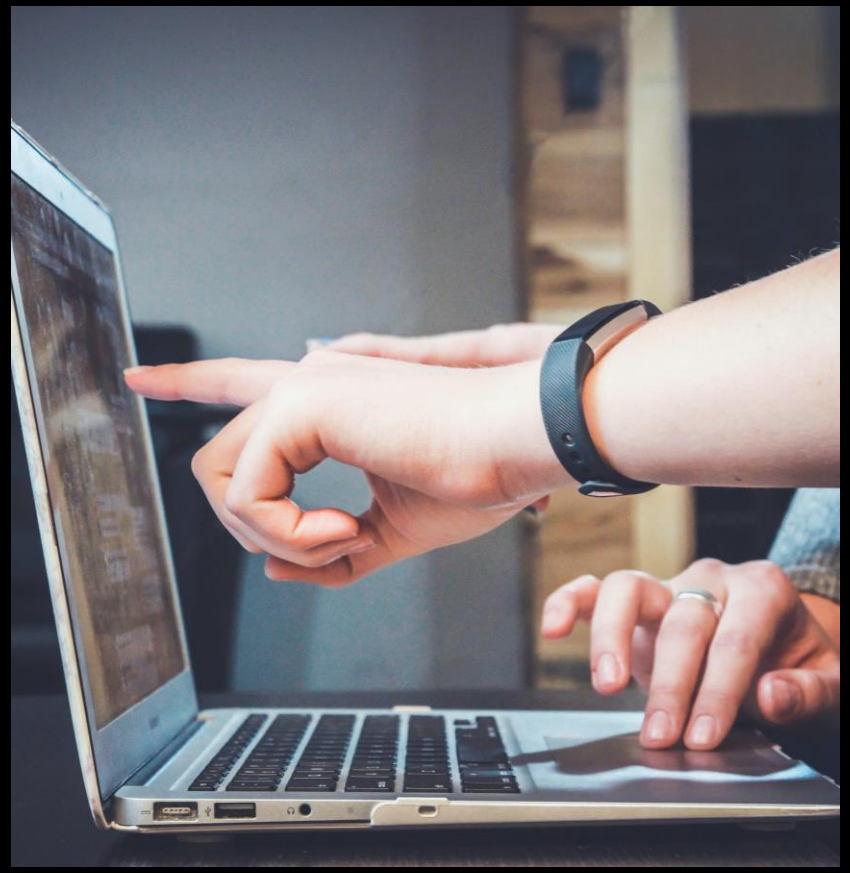


DQLab

Bagaimana Exploratory Data Analysis dan menilai kualitas data?

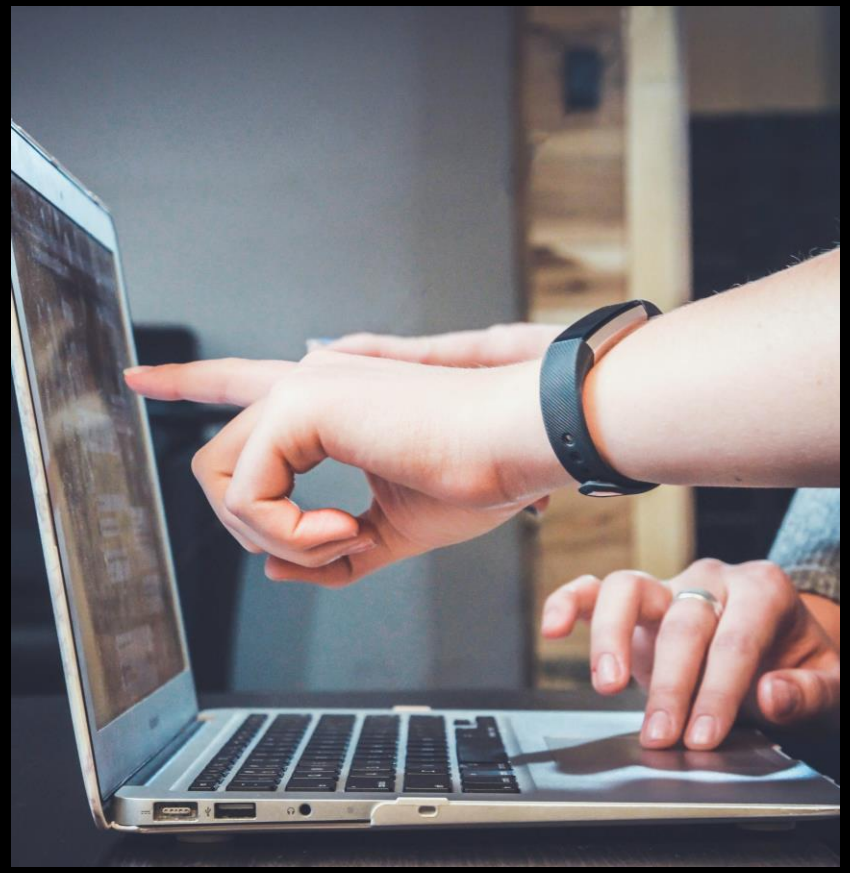
Parameter Menilai Data

- Completeness :
  - Apakah final datasets sesuai dengan kebutuhan bisnis & harapan Anda?
- Accuracy:
  - seberapa dekat data mewakili realitas dunia nyata?
- consistency
  - Apakah data konsisten sesuai jenis dengan sebelumnya?
- Timeliness
  - apakah datanya masih relevan?
  - apakah itu mencerminkan kenyataan saat ini?
- Duplication
  - Apakah ada data yang duplikat?
- Validity
  - Apakah datanya valid?
- availability
  - Apakah datanya tersedia ?
- Provenance
  - Asal datanya dari mana?



Steps EDA

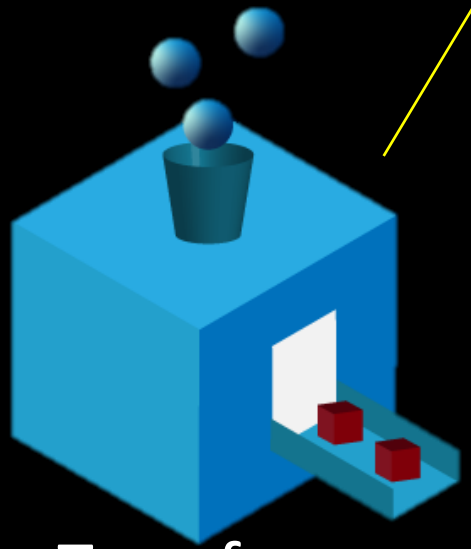
1. Mengamati kumpulan data yang ada
2. Mencari missing value dan membenahinya jikalau diperlukan
3. Categorisasikan data categorical, numerical
4. Identifikasi hubungan antar variable
5. Identifikasi outliers, skewness data, aplikasikan statistic descriptive or inferentials



Bagaimana EDA di Python?

Cleaning

Numpy,
Pandas,
Scipy



Transform

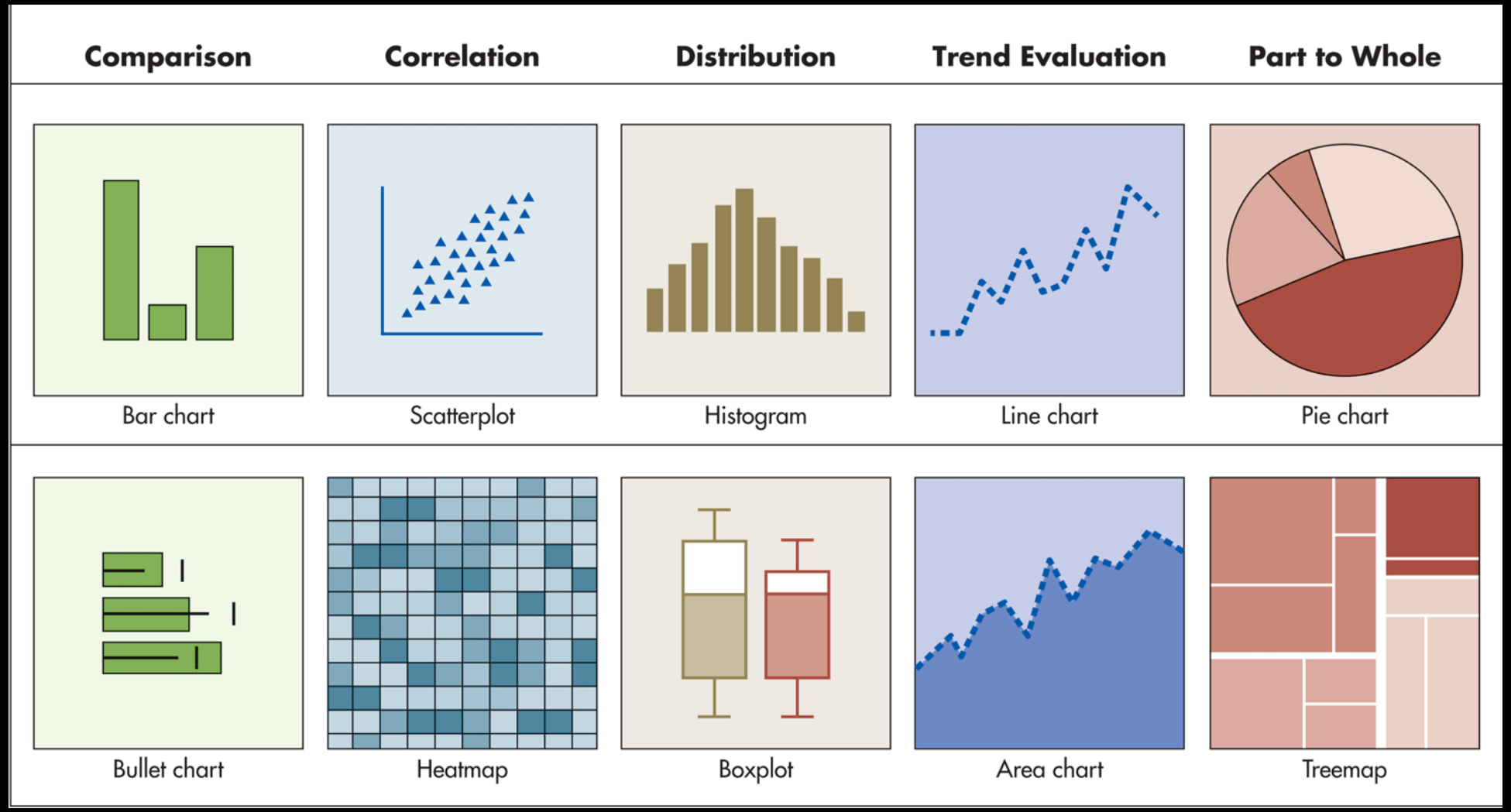
Visualization

Matplotlib, Seaborn,
Folium, plotly

DQLab

Mengenal Tujuan Visualisasi Untuk EDA

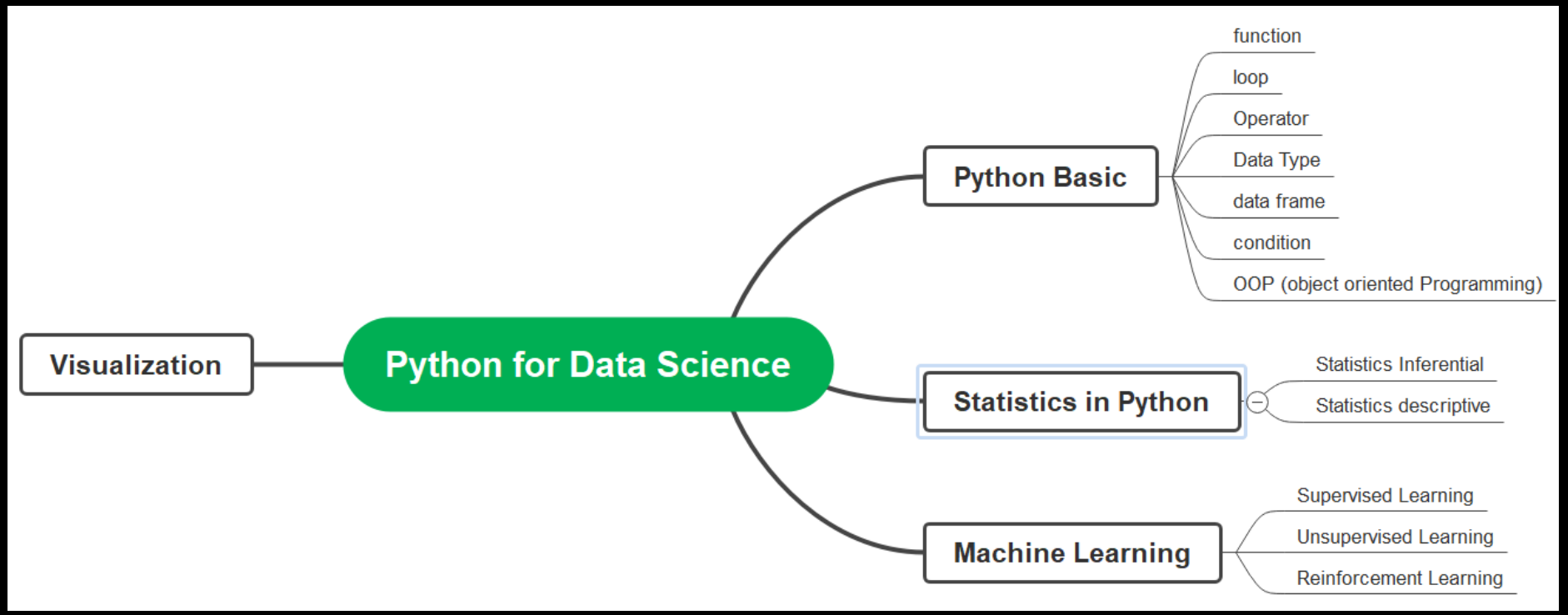
Tujuan Visualisasi



DQLab

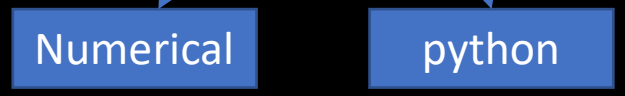
Apa tools untuk Melakukan EDA di python?

Python For Data science

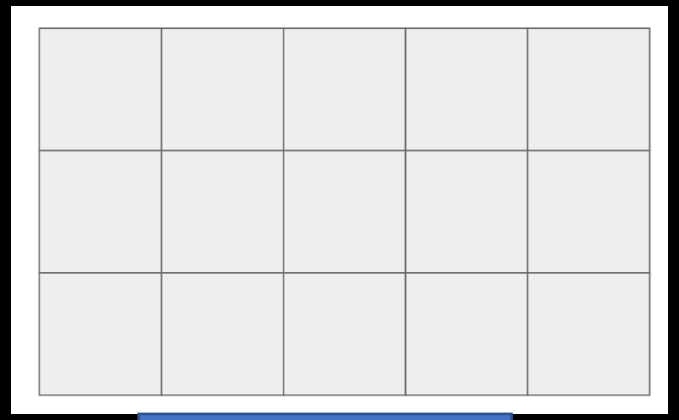


Numpy

Num py



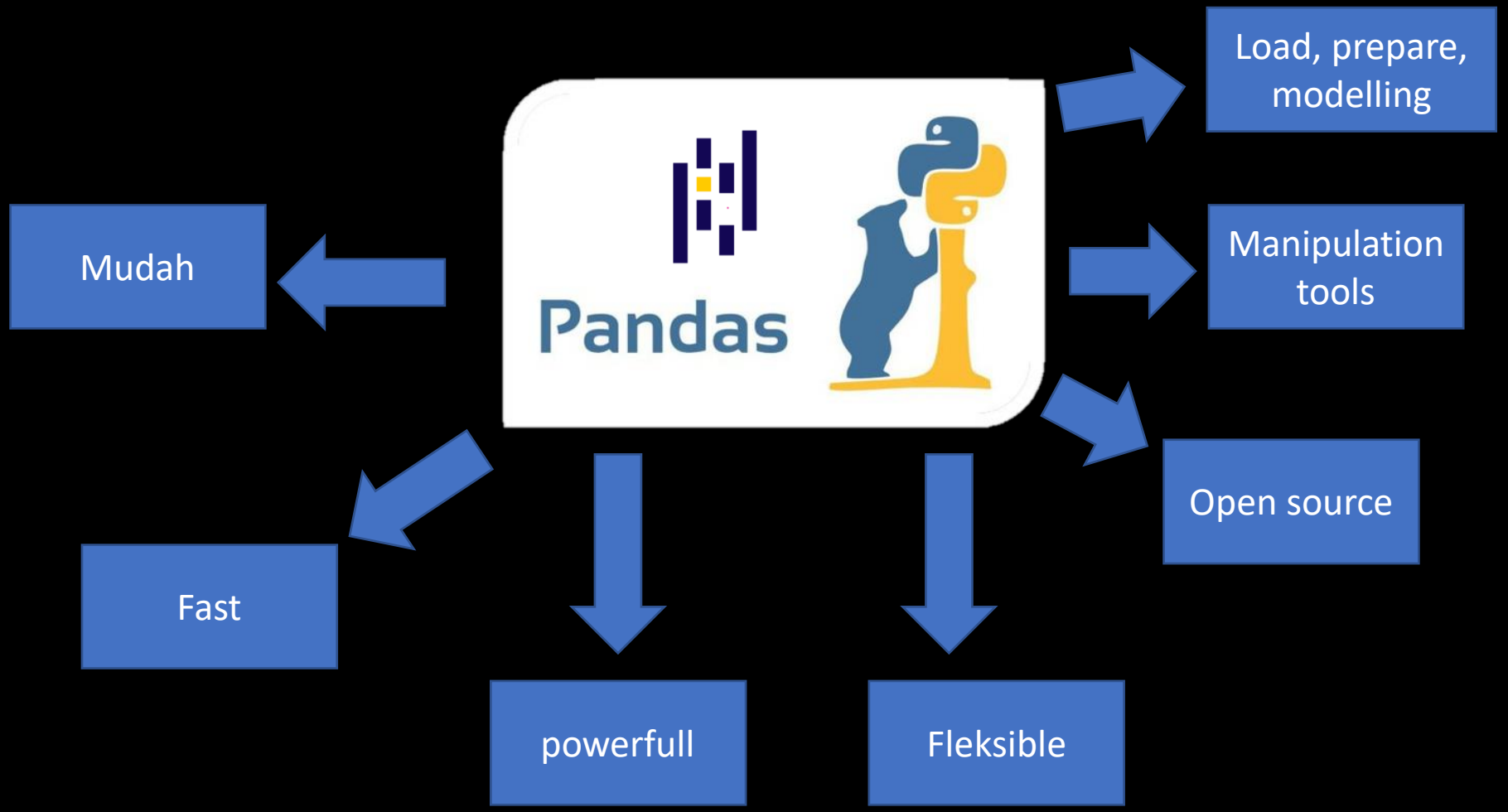
1D Array



2D Array(multi)

- NumPy berfungsi sebagai library untuk melakukan proses komputasi numerik terutama dalam bentuk *array* multidimensional (1-Dimensi ataupun 2-Dimensi).
- *Array* merupakan kumpulan dari variabel yang memiliki tipe data yang sama.
- NumPy menyimpan data dalam bentuk *arrays*.

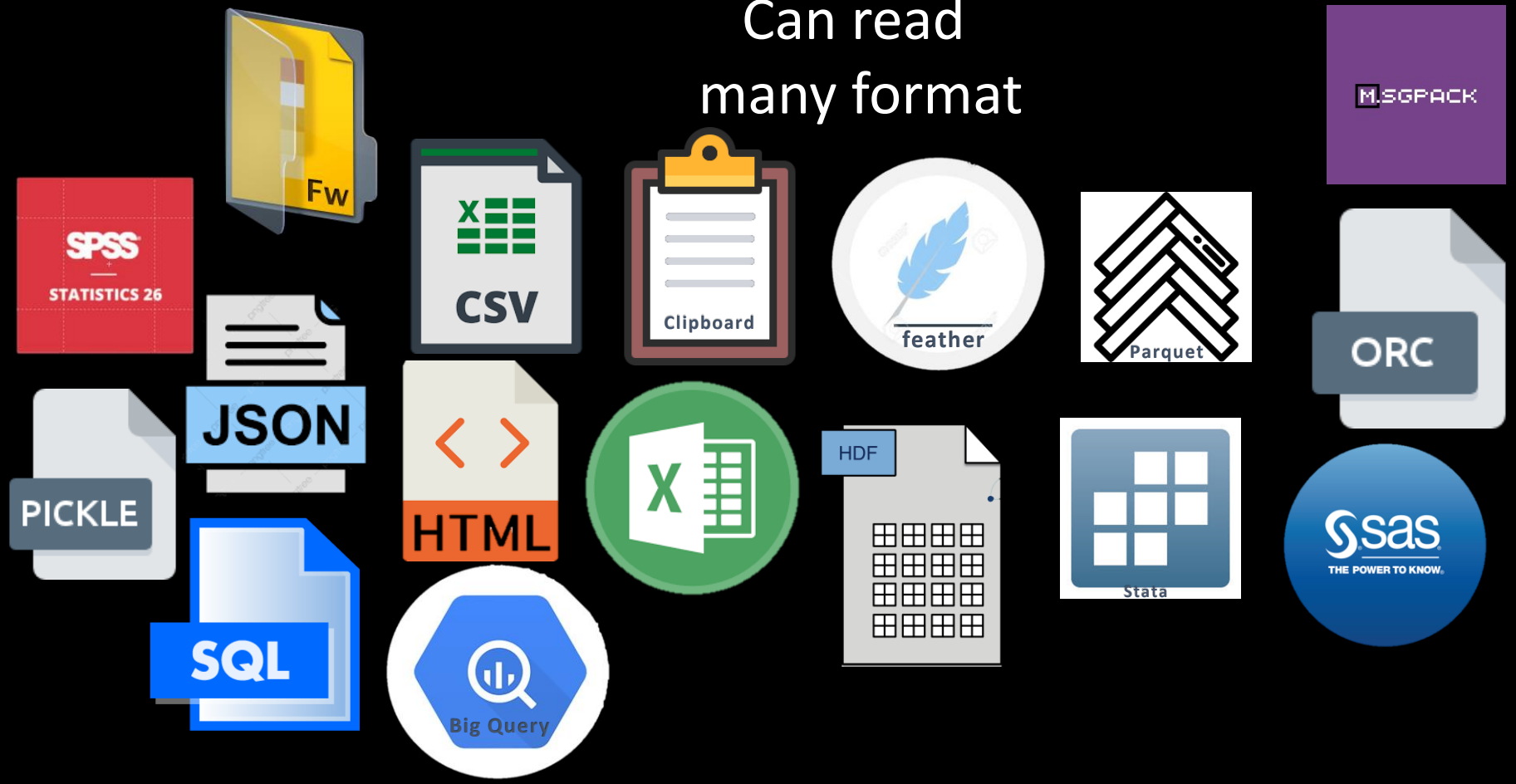
Pandas


Pandas
Load, prepare, modelling
Manipulation tools
Open source
Fleksible
powerfull
Fast
Mudah

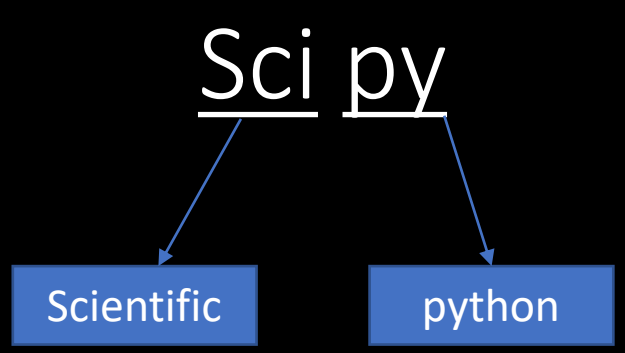
DOLab

Pandas

Can read
many format



Scipy



- Scipy dibangun untuk bekerja dengan array NumPy dan menyediakan banyak komputasi numerik yang ramah pengguna dan efisien seperti rutinitas untuk integrasi, diferensiasi dan optimasi numerik.
- Baik NumPy maupun SciPy berjalan pada semua operating system, cepat untuk diinstall dan gratis.
- NumPy dan SciPy mudah digunakan, tetapi cukup kuat untuk diandalkan oleh beberapa data *scientist* dan *researcher* terkemuka dunia.

DQLab

Lybrary for Visualisasi

matplotlib

Seaborn

folium

Question And Answer

DQLab

Untuk detail technical Exploratory Data Analysis akan di sampaikan pada Workshop 19 April 2022 dengan tema

Comprehensive Exploratory Data Analysis of House Price Datasets with Python

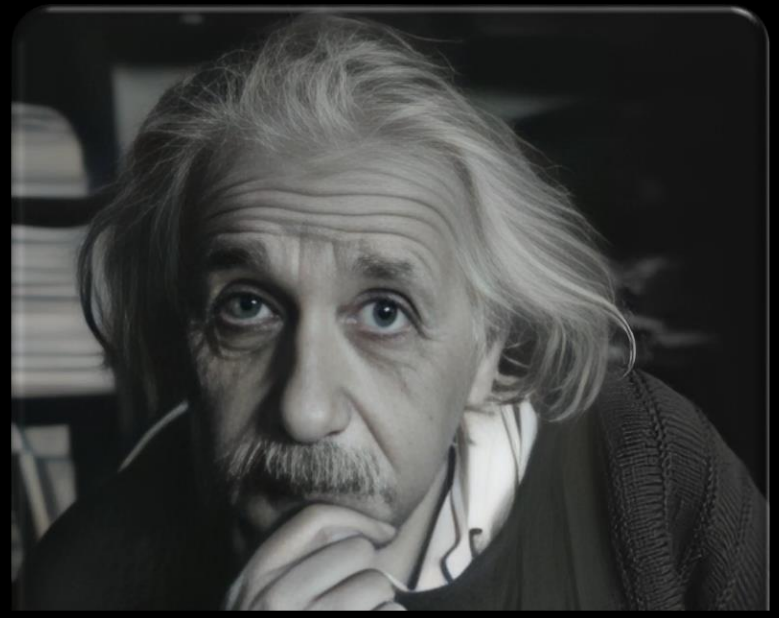
Kesimpulan

- Data science adalah ilmu pengetahuan interdisiplin tentang metode komputasi untuk mendapatkan wawasan berharga yang dapat ditindaklanjuti dari kumpulan data yang mencakup tiga fase yaitu desain data, mengumpulkan data, dan analisis data.
- Exploratory data analysis adalah proses kritis dalam melakukan investigasi awal pada data untuk menemukan pola, untuk menemukan anomali, untuk menguji hipotesis dan untuk memeriksa asumsi dengan bantuan statistik ringkasan dan representasi grafis.
- Parameter menilai kualitas ada yaitu: Completeness, Accuracy, consistency, timeliness, Timeliness, Duplication, Validity, availability, Provenance
- Steps EDA:
    1. Mengamati kumpulan data yang ada (shape, columns, variable, data types)
    2. Mencari missing value dan membenahinya jikalau diperlukan
    3. Categorisasikan data categorical, continues, discrete
    4. Identifikasi hubungan antar variable
    5. Identifikasi outliers, skewness data, aplikasikan statistic descriptive or inferentials

Kesimpulan

- Tujuan Visualisasi
  - Comparison
  - Corelasi
  - Distribusi
  - Trend
  - Part o whole

- Tools EDA
  - Perlu paham Python programming
  - Numpy
  - Pandas
  - Scipy
  - Matplotlib, seaborn, folium



“Once you stop learning, you start dying”
(Albert Einstein)