Data Science and Machine Learning

Antony Ross

Environment Set-up

Install Anaconda 3.x

Create a course folder

Launch Anaconda Navigator

The Data Science Process

1. Identify the question
2. Get the data
3. Clean the data
4. Explore the data
5. Model the data
6. Communicate the results

The Data Science Process

1. Identify the question
2. Get the data
3. Clean the data
4. Explore the data
5. Model the data
6. Communicate the results

Identify the question

Identify the question

- Answerable
- Actionable
- Narrow
- Specific

Get the data

Data Sources

- kaggle.com/datasets
- https://registry.opendata.aws
- https://cloud.google.com/bigquery/public-data/
- data.gov
- archive.ics.uci.edu/ml/
- https://github.com/fivethirtyeight/data
- https://www.quandl.com/search
- public APIs (e.g., Twitter, Facebook, Spotify)
- web scraping
- your company

Data Sources

Google Dataset Search

toolbox.google.com/datasetsearch

ProPublica Data Store

propublica.org/datastore

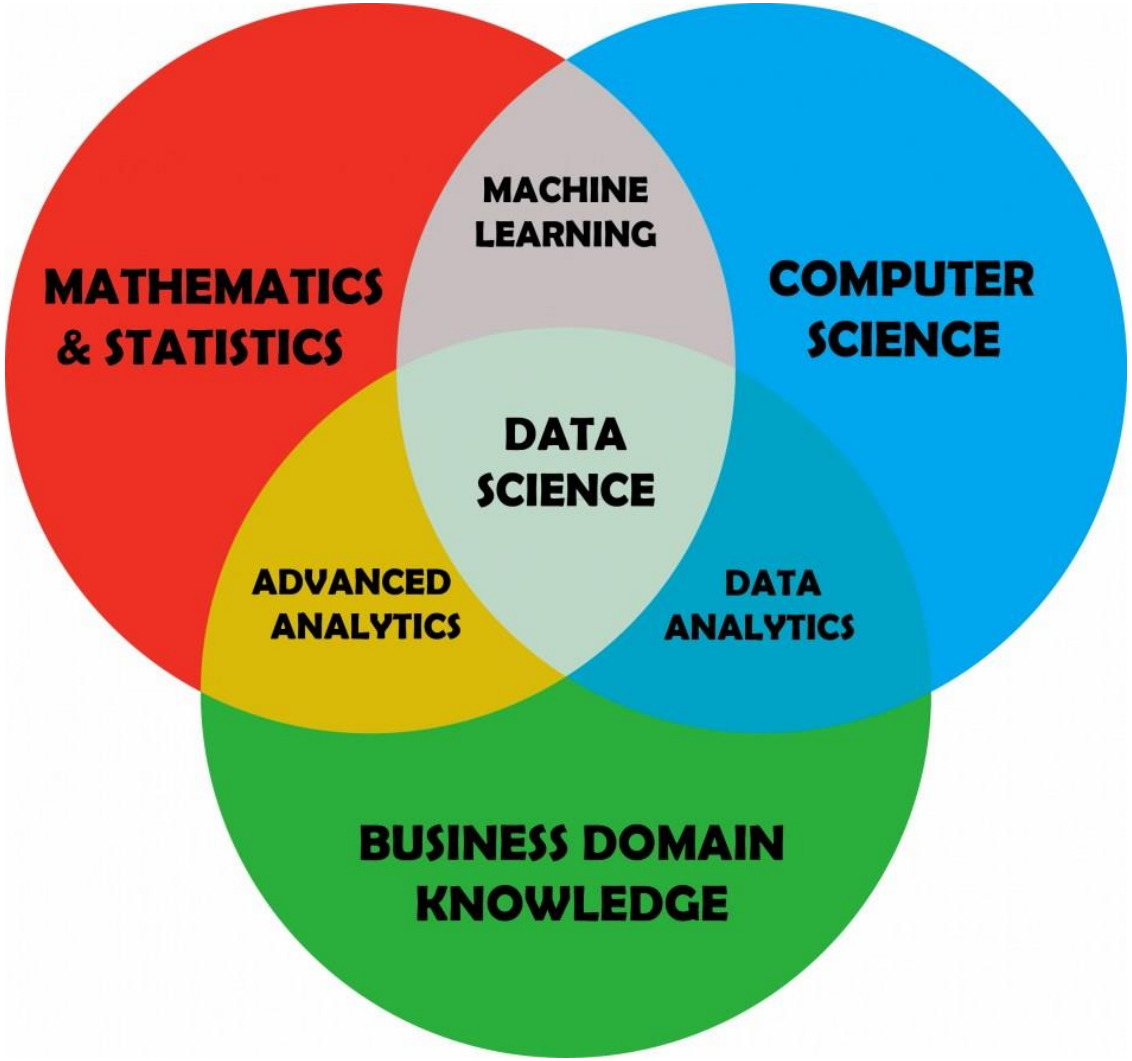
NASA's Open Data Portal

data.nasa.gov

World Bank Open Data

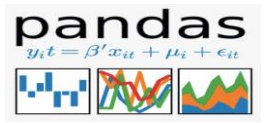
data.worldbank.org

Data Science



Python Libraries for Data Analysis

NumPy



Pandas

Matplotlib

SciPy

-learn