Data Science with Python

Rabindra Sapkota

Lead Data Engineer

linkedin.com/in/sapkotarabindra

Data Science

- **Data is the new oil** (Raw data is useless)
- Get Insight from structured / unstructured data
- Uses:
  - Healthcare (Disease Prediction & Treatment)
  - Climate Prediction
  - Social Science (Sentiment Analysis, Behavioral Science)
  - Banking (Fraud Detection, Customer Segmentation)
  - Social Media (Recommendation, Fake profile detection)

Tools Used

- Languages: Python, R, SQL
- Libraries: Pandas, Scikit learn, Tensor flow
- Visualization: Matplotlib, Seaborn, Tableau, Power BI
- Big Data: Hadoop, Spark

Life Cycle

1. Problem Definition: What is Question / Challenge
2. Data Collection: Get Relevant Data
3. Data Cleaning & Pre-processing: Fix Error & Null value
4. Exploratory Data Analysis (EDA): Look for pattern
5. Model Building: Apply ML / Statistics
6. Evaluation: Test Accuracy & Performance
7. Deployment: Use model in application
8. Monitoring : Check ongoing performance

Problem Landscape

1. Descriptive Problem: What happened ?
   - Summarize / Explore Past Data
   - Sales Report, User Activity Dashboard, Churn Report
2. Diagnostic Problem: Why did it happen?
   - Analyze cause or correlation
   - RCA of churn, Weather vs sale correlation
3. Predictive Problem: What is likely to happen?
   - Forecast future event or behavior
   - Weather forecast, Predict loan default
4. Prescriptive Problem: What should we do about it?
   - Recommend action to optimize outcomes
   - Dynamic Pricing
5. Cognitive Problem: How machine emulate human task? (NLP, Vision)

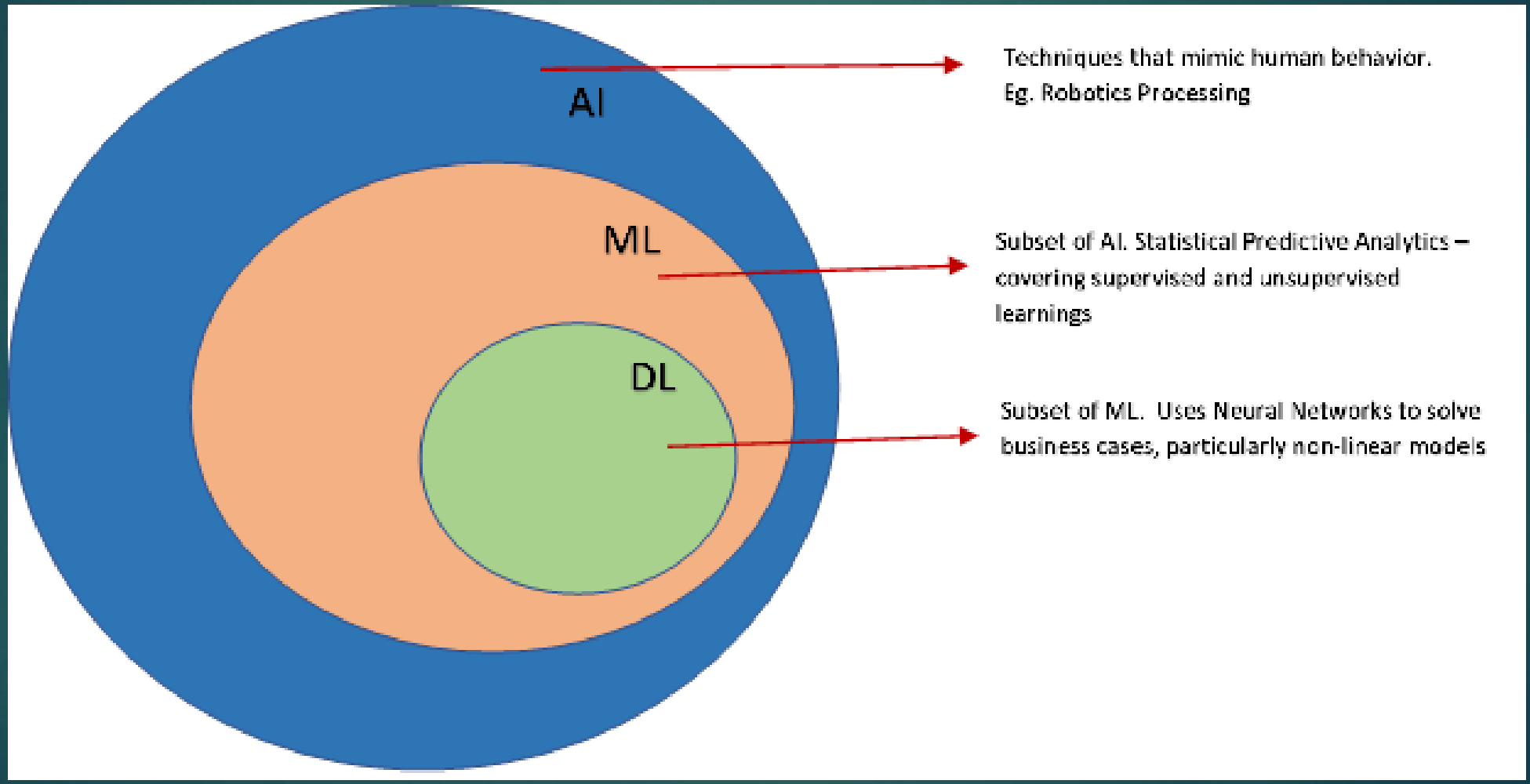
Data Type

1. Structured: CSV, Excel, Database
2. Unstructured: Text, Image, Audio, Video
3. Time Series: Sensor / finance over time
4. Spatial: Maps, Geo-location
5. Graph: Networks, Social Connection

Problem by Technique

1. Classification: (Spam / Not Spam)
2. Regression: (Predicting Numeric Value)
3. Clustering: (Grouping of similar item)
4. Anomaly Detection (Finding outliers)
5. Recommendation System
6. NLP
7. Reinforcement Learning (from feedback)

AI / ML / DL


AI
ML
DL
Techniques that mimic human behavior.
Eg. Robotics Processing
Subset of AI. Statistical Predictive Analytics – covering supervised and unsupervised learnings
Subset of ML. Uses Neural Networks to solve business cases, particularly non-linear models

Career Path

Data Analyst

- **Role**: Interpret data to help businesses make decisions.
- **Skills**: Excel, SQL, BI tools (Power BI, Tableau), basic Python.

Data Engineer

- **Role**: Designs and maintains data pipelines and architecture.
- **Skills**: SQL, Python/Scala, Spark, Hadoop, ETL tools, cloud

Career Path

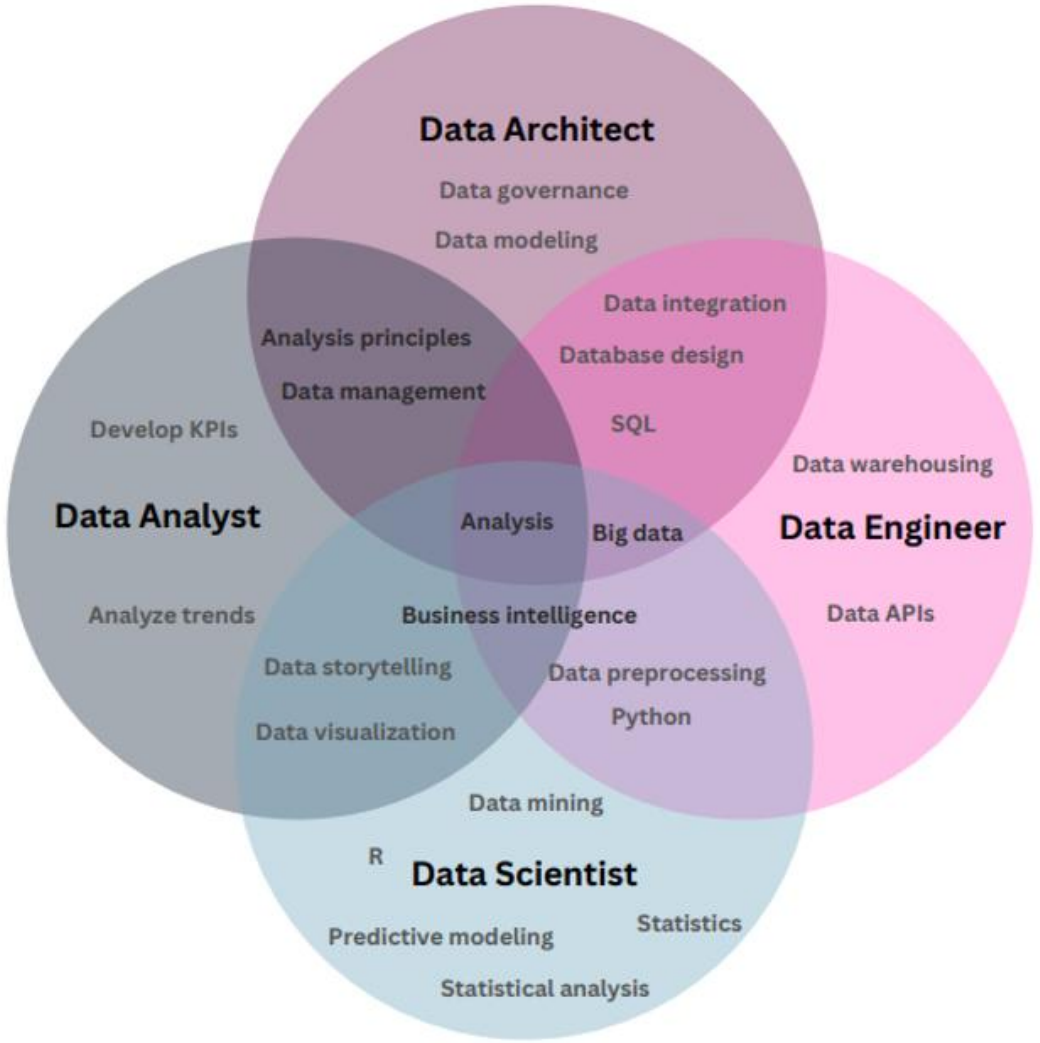
Data Scientist

- **Role**: Analyzes datasets, EDA, builds model & Visuals
- **Skills**: Stat, ML, Python/R, SQL, data visualization.

ML Engineer

- **Role**: Build & deploy ML models in production systems.
- **Skills**: Python, TensorFlow/PyTorch, MLOps, Software engineering

Data Role



Data Analyst

Responsible for gathering, analyzing, and interpreting data to inform business decisions and strategies.

Data Architect

Responsible for designing and managing the overall structure and organization of the data architecture to ensure effective data integration, storage, retrieval, and analysis.

Data Engineer

Responsible for designing, building, and maintaining the infrastructure and systems that enable data storage, retrieval, and analysis.

Data Scientist

Responsible for using statistical and computational methods to analyze and interpret complex data sets in order to develop insights and solutions that inform business decisions and strategies.

Python Installation

- Download Python version 3.13 from python.org (Remember: Add to path)
- Download and install VS-Code from code.visualstudio.com
- Go to Extension view: `ctrl+shift+x`, Search Python (Microsoft) and Install
- Install Black Formatter from vs-code extension
- For running: `ctrl+shift+p`, choose python

Pre-Requirement

- String & List Methods
- Conditional
- Loop
- Function
- Project – 2 of Python (Samachar Sangraha)

Python 2.x vs 3.x

- Python 2.x is native to Linux OS & present by default on Linux like centos
- Python 2.x already reached End of Life (no update)
- Major libraries NumPy, Pandas dropped its support
- We prefer to install python3 even though python 2 is present

Packages

- mailerpy: `pip install mailerpy`
- Requests: `pip install requests`
- Psycopg2: `pip install psycopg2-binary`
- Pandas: `pip install pandas`
- Matplotlib: `pip install matplotlib`
- Seaborn: `pip install seaborn`
- Scikit Learn: `pip install scikit-learn`