PROJECT
Applications of Big Data

Project Title: Applying Data Science process to make an attempt to solve (tiny) real-world problems.

Project Description: There are multiple kinds of Data Science process as shown in Figure 1 and Figure 2. The common points are as follows:

1) Collecting the data related to requirements of users or businesses.
2) Data preprocessing.
    a. Making data "clean"
    b. Preprocessing data for the next step tasks.
3) Getting insights from your data (show evidence as well). One possible evidence is to show the visualization from the data.
4) Building the model to predict the "trend" of the data based on the requirements.
5) Making a (business) decision based on the results and repeat the process (if possible).

Students are required to apply the process to deal with a (tiny) real-world problem from step 1 to step 5. For step 1, one possible way is to find some datasets on Kaggle, such as Credit Card Fraud Detection Dataset 2023. By observing the descriptions of the dataset, you have done steps 1 but not yet steps 2-5. Note that, if you do step 1 from scratch, such as using crawler or APIs to collect data, you will get a bonus score. If you use datasets from Kaggle, your dataset size must be greater than 1GB (for texts) and if your dataset contains images, the number of images in the dataset must be huge enough (it will be assessed by the evaluator). For step 2 to step 5, you can do by yourself.

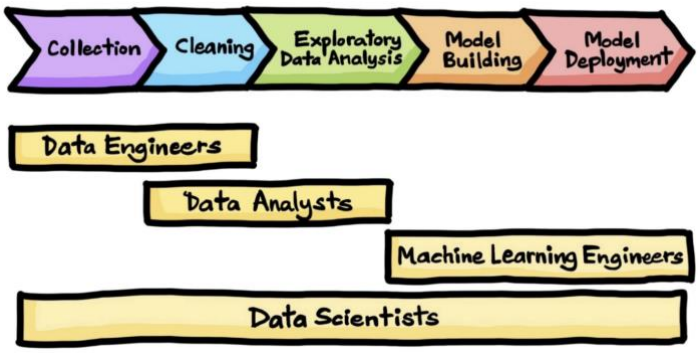

Figure 1: An example of Data Science process (1) (*)

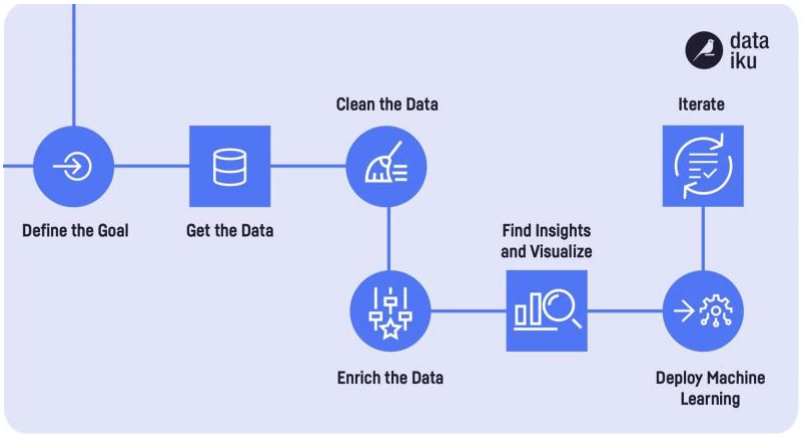

Figure 2: An example of Data Science process (2) (**)

Project Evaluation: Project requirements and activities to get bonus scores.

Step 1: Selecting suitable dataset satisfying the requirements (size). Collecting dataset from scratch is a bonus score.

Step 2: Preprocessing your dataset. Being able to explain how you can preprocess the data (which "dirty" your data are, which solutions you propose, etc.) is a bonus score.

Step 3: Visualizing the data. Showing valuable insights from your data with the evidence, such as visualization is a bonus score.

Step 4: Evaluating sub-tasks via models, such as accuracy. Being able to evaluate the entire project is a bonus score.

Step 5: Write the report with sufficient information. Using format of bachelor thesis is a bonus score.

Project Progress: There are milestones that each group needs to follow:

- Project proposal: Mar 02, 2025
- Progress report (1): Mar 16, 2025
- Progress report (2): Mar 30, 2025
- Final presentation (in person): TBD
- Final report: few days after presentation.

Project Rules

- This is a group project so you can utilize your members from the other project (such as paper project) in this course.
- You must use at least **TWO big data techniques** (for storing, processing, visualization) in your project.
- Any actions of cheating or plagiarism will be punished as 0 score overall.

- You need to have reference section (if you consult somewhere) in your report.

There are some examples as follows:

1. Real-Time Customer Sentiment Analysis on Social Media

Objective:

Analyze real-time Twitter data to classify customer sentiment about a specific brand, product, or service using **Big Data tools.**

Big Data Tools Used:

- **Apache Kafka** – To stream real-time tweets.
- **Apache Spark (PySpark or Spark Streaming)** – For real-time data processing.
- **Hadoop HDFS** – To store historical tweets for analysis.
- **Hive** – For querying large datasets.
- **MLlib (Spark's Machine Learning Library)** – For sentiment classification.

Tasks:

1. **Data Collection:** Use **Kafka** to stream real-time tweets via the Twitter API.
2. **Storage & Preprocessing:** Store raw tweets in **HDFS** and clean them using **Spark.**
3. **Exploratory Analysis:** Perform keyword analysis, word clouds, and frequency distributions.
4. **Sentiment Analysis Model:** Train an NLP model using **MLlib** or **Spark NLP.**
5. **Real-Time Dashboard:** Display sentiment trends over time using **Tableau or Power BI.**

2. Predictive Maintenance for IoT Sensor Data

Objective:

Analyze large-scale IoT sensor data to predict equipment failures in industrial machines.

Big Data Tools Used:

- **Apache Hadoop (HDFS & MapReduce)** – To store large IoT data logs.
- **Apache Spark** – For large-scale batch and real-time analytics.
- **Kafka** – For real-time streaming of sensor data.
- **HBase** – For fast lookups of machine failure data.
- **MLlib** – To build predictive maintenance models.

Tasks:

1. **Ingesting IoT Data:** Collect real-time machine sensor logs using **Kafka.**
2. **Data Storage:** Store structured and unstructured sensor data in **HDFS & HBase.**
3. **Exploratory Data Analysis:** Use **Spark SQL** to detect failure patterns.
4. **Predictive Modeling:** Train ML models (Random Forest, Gradient Boosting) using **MLlib** to predict machine failures.
5. **Dashboard & Alert System:** Visualize failure predictions & generate alerts.

3. Fraud Detection in Banking Transactions Using Big Data

Objective:

Use **Big Data & Machine Learning** to detect fraudulent credit card transactions.

Big Data Tools Used:

- **Apache Spark & MLlib** – For fraud prediction.
- **Kafka** – For real-time transaction streaming.
- **HDFS** – To store past transactions.
- **Hive** – To query and analyze large-scale fraud data.
- **Airflow** – To automate data processing workflows.

Tasks:

1. **Data Ingestion:** Stream real-time transaction data using **Kafka.**
2. **Storage & Preprocessing:** Store transactions in **HDFS** and clean them with **Spark.**
3. **EDA & Feature Engineering:** Use **Spark SQL** to detect anomalies (unusual amounts, locations, spending patterns).

4. **Machine Learning Model**: Train an ML model using **MLlib** (e.g., Logistic Regression, Decision Trees, or Neural Networks).
5. **Real-Time Fraud Alerts**: Implement an alert system for suspicious transactions.