# Technologies, Tips, and Tricks to Design Machine Learning Pipeline using Python

Dr. Gemma Catolino

December 3rd 2021

Tilburg University TU/e JADS Jheronimus Academy of Data Science

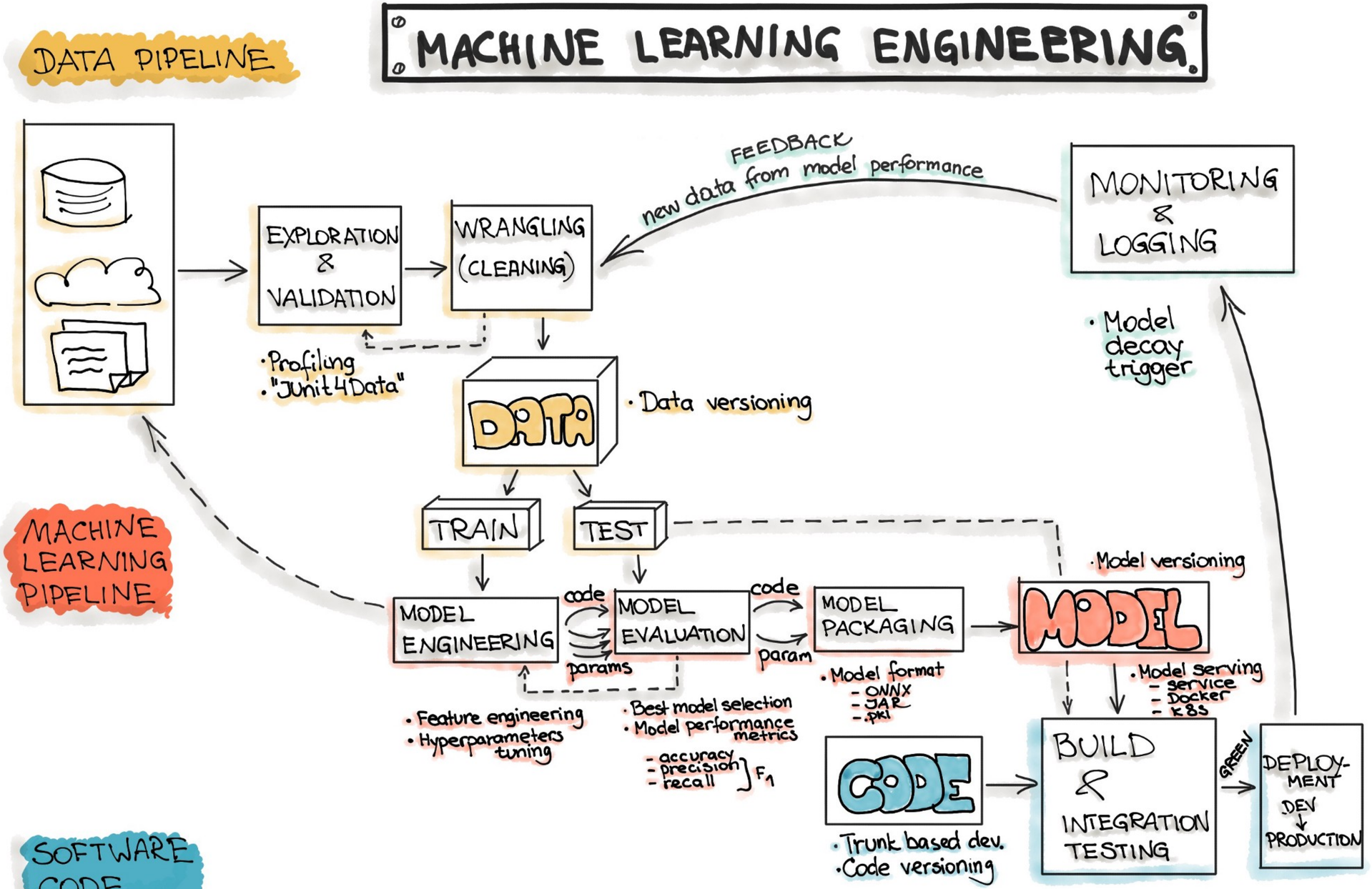
# Tools and Libraries that we will use today

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser. Colab notebooks are Jupyter notebooks (Jupyter Notebooks is an IDE used to interface with Python) and execute code on Google's cloud servers, meaning you can leverage the power of Google hardware. Google Collab requires a Google account.
https://colab.research.google.com/notebooks/intro.ipynb#recent=true


Scikit-learn is a library in Python that provides several unsupervised and supervised learning algorithms.
https://scikit-learn.org/stable/
(To know about different classifiers)

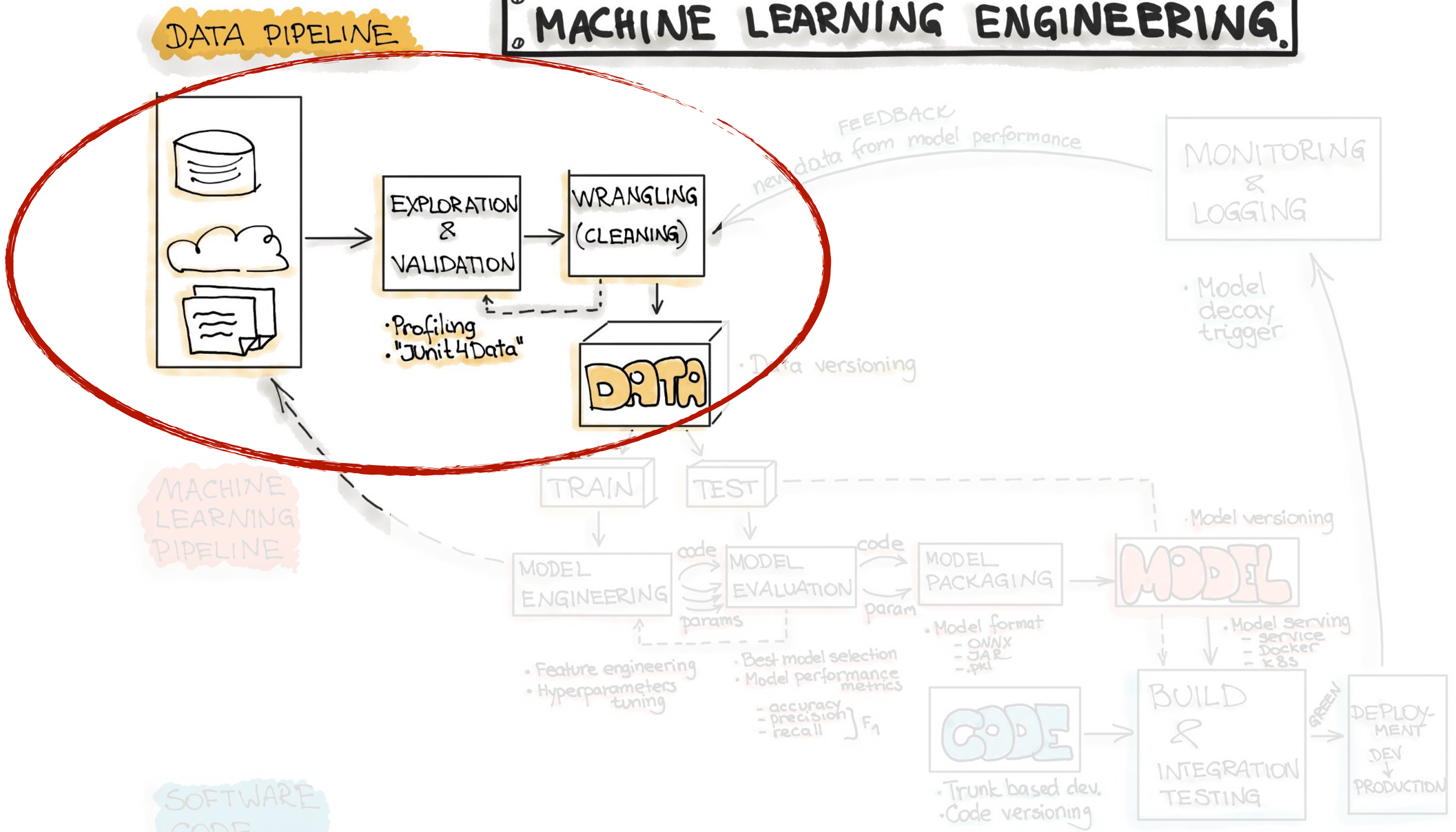
# DATASET


Online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.




# Machine Learning Pipeline

# MACHINE LEARNING ENGINEERING



# DATASET

A. **age:** The person's age in years
B. **sex:** The person's sex (1 = male, 0 = female)
C. **cp:** The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
D. **trestbps:** The person's resting blood pressure (mm Hg on admission to the hospital)
E. **chol:** The person's cholesterol measurement in mg/dl
F. **fbs:** The person's fasting blood sugar (glicemia) (> 120 mg/dl, 1 = true; 0 = false)
G. **restecg:** Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
H. **thalach:** The person's maximum heart rate achieved (frequent cardiaca)
I. **exang:** Exercise induced angina (1 = yes; 0 = no)(angina indotta dall'esercizio)
J. **oldpeak:** ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
K. **slope:** the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
L. **ca:** The number of major vessels (0-3) (arteries, veins, and capillaries) (pendent e pico dell'elettrocardiogramma)
M. **thal:** A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
N. **target:** Heart disease (0 = no, 1 = yes)

https://www.kaggle.com/sabinero/heart-desease-classification-95-15-recall

A. Data Visualization

B. Data Exploration

C. Data Imputation

D. Define data and Target Variables

E. Split Dataset (Different Strategies)

F. Balance Dataset

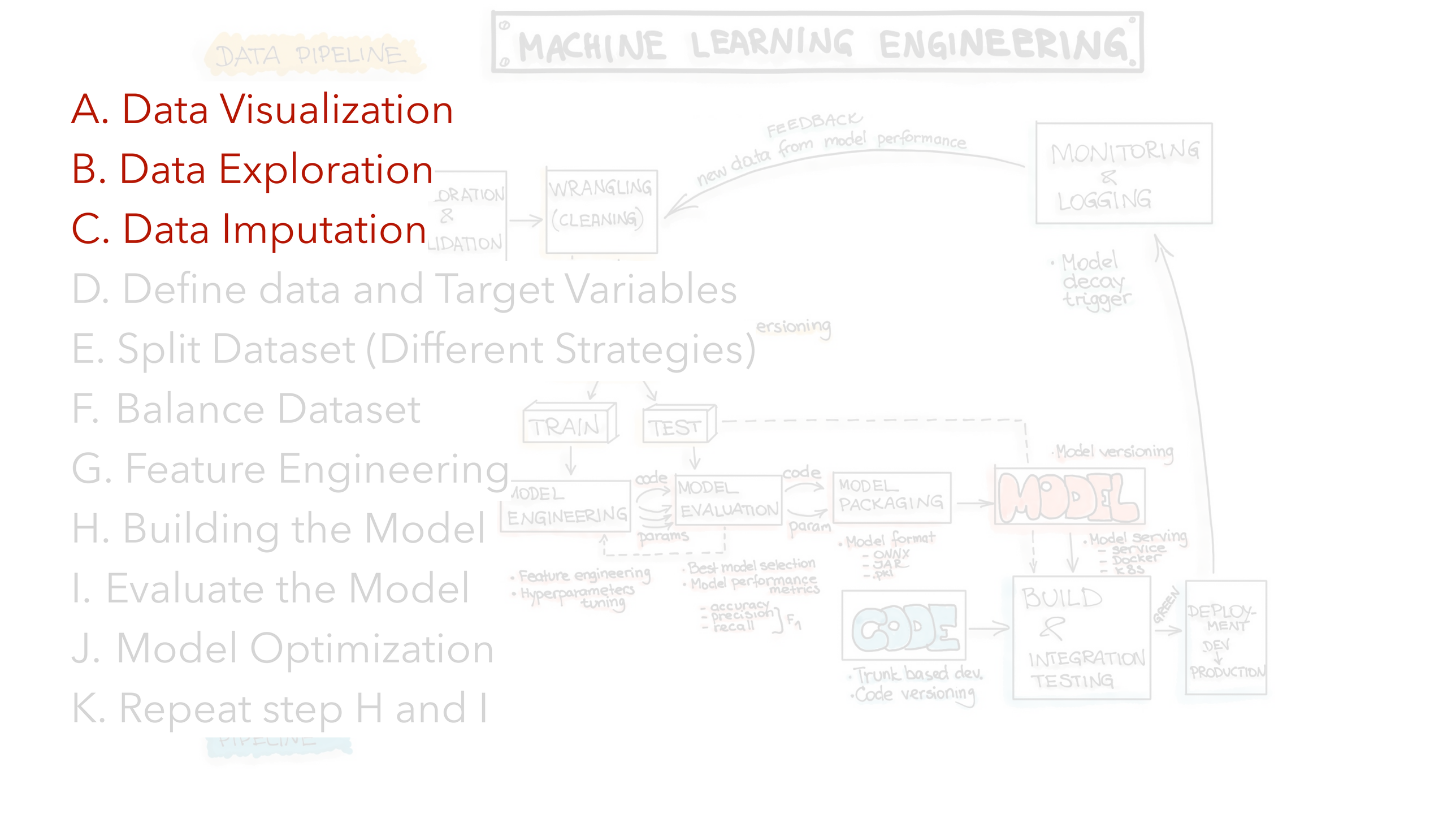
G. Feature Engineering

H. Building the Model

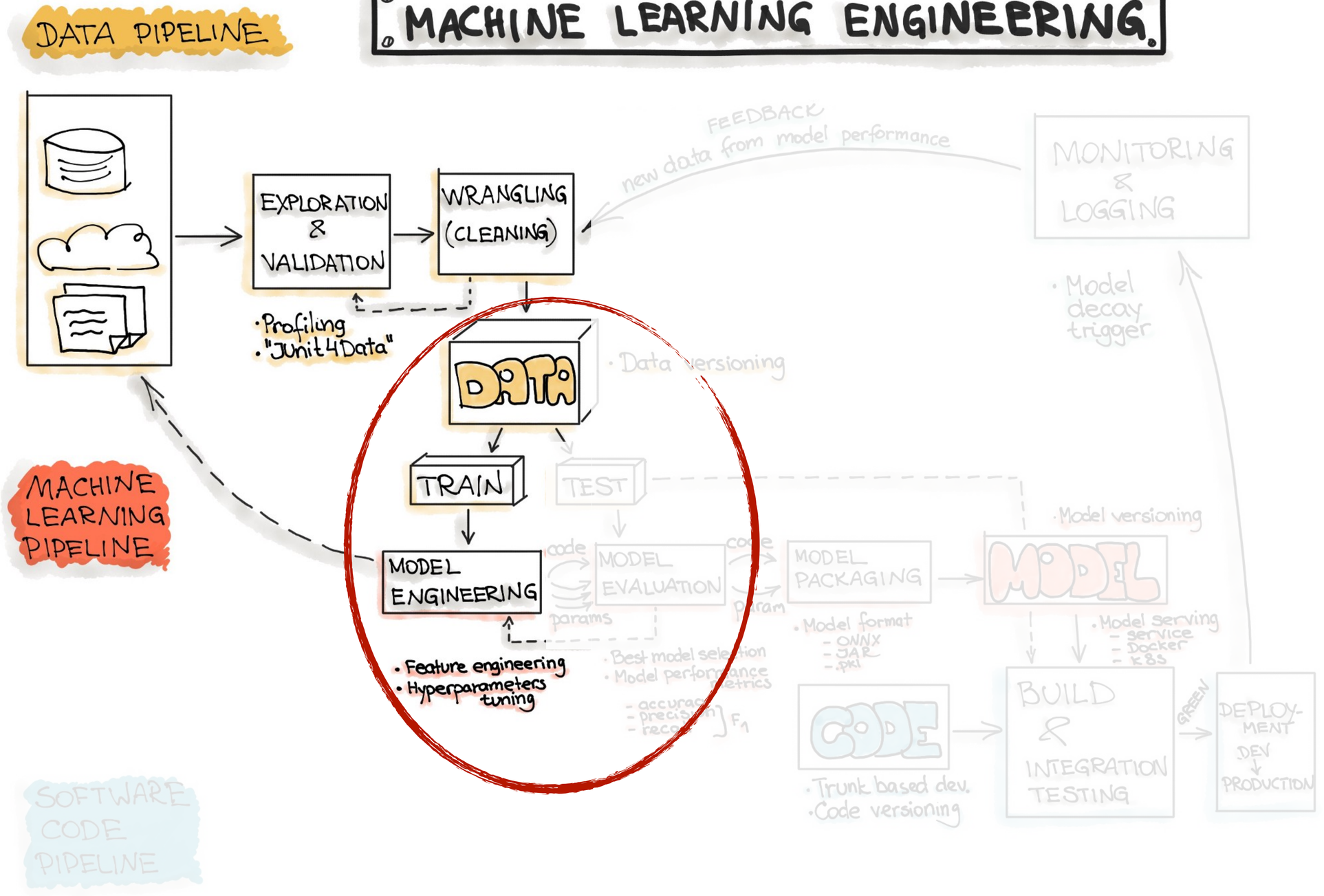
I. Evaluate the Model

J. Model Optimization

K. Repeat step H and I



# MACHINE LEARNING ENGINEERING



A. Data Visualization

B. Data Exploration

C. Data Imputation

D. Define Data and Target Variables

E. Split Dataset (Different Strategies)

F. Balance Dataset

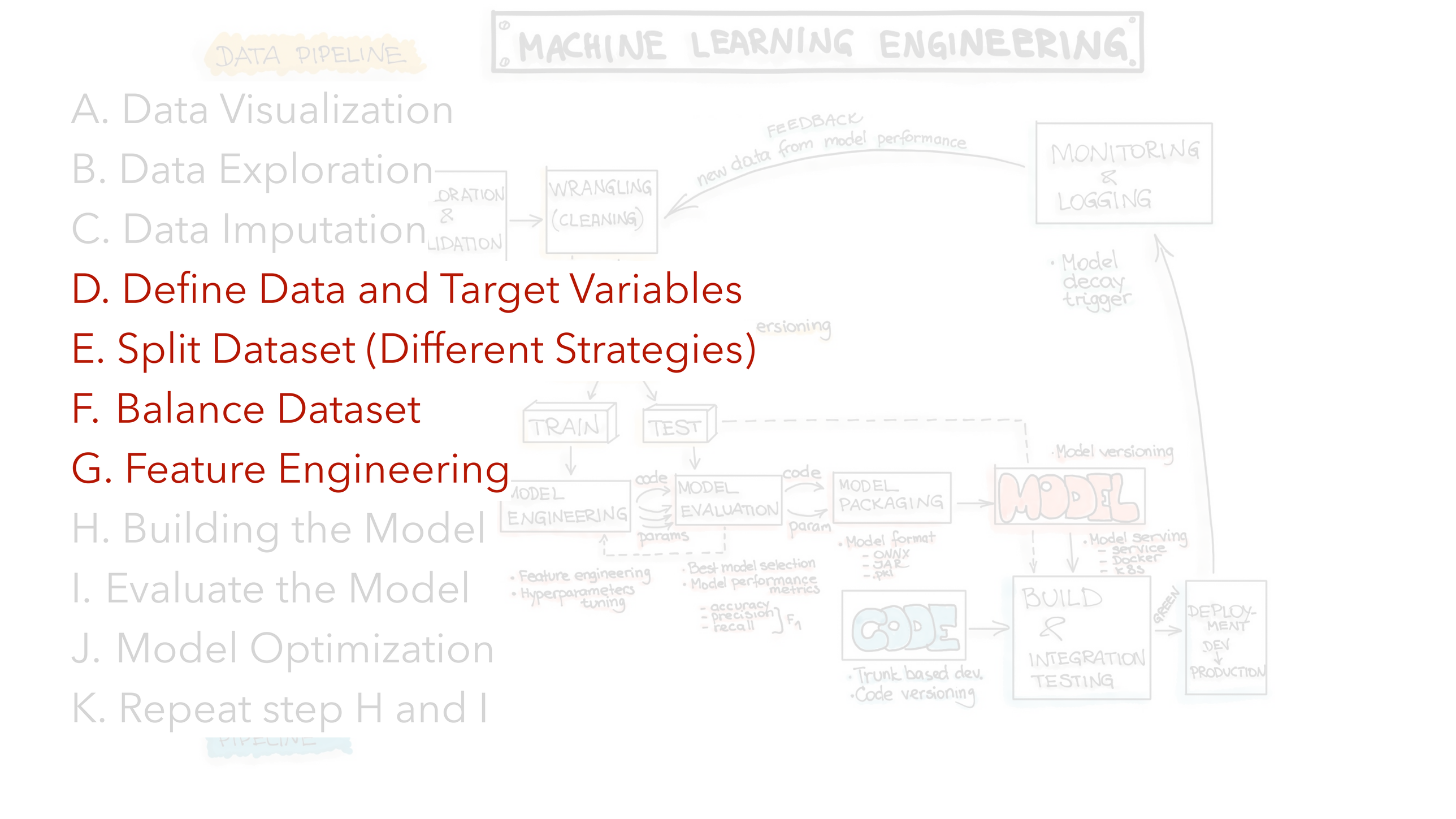
G. Feature Engineering

H. Building the Model

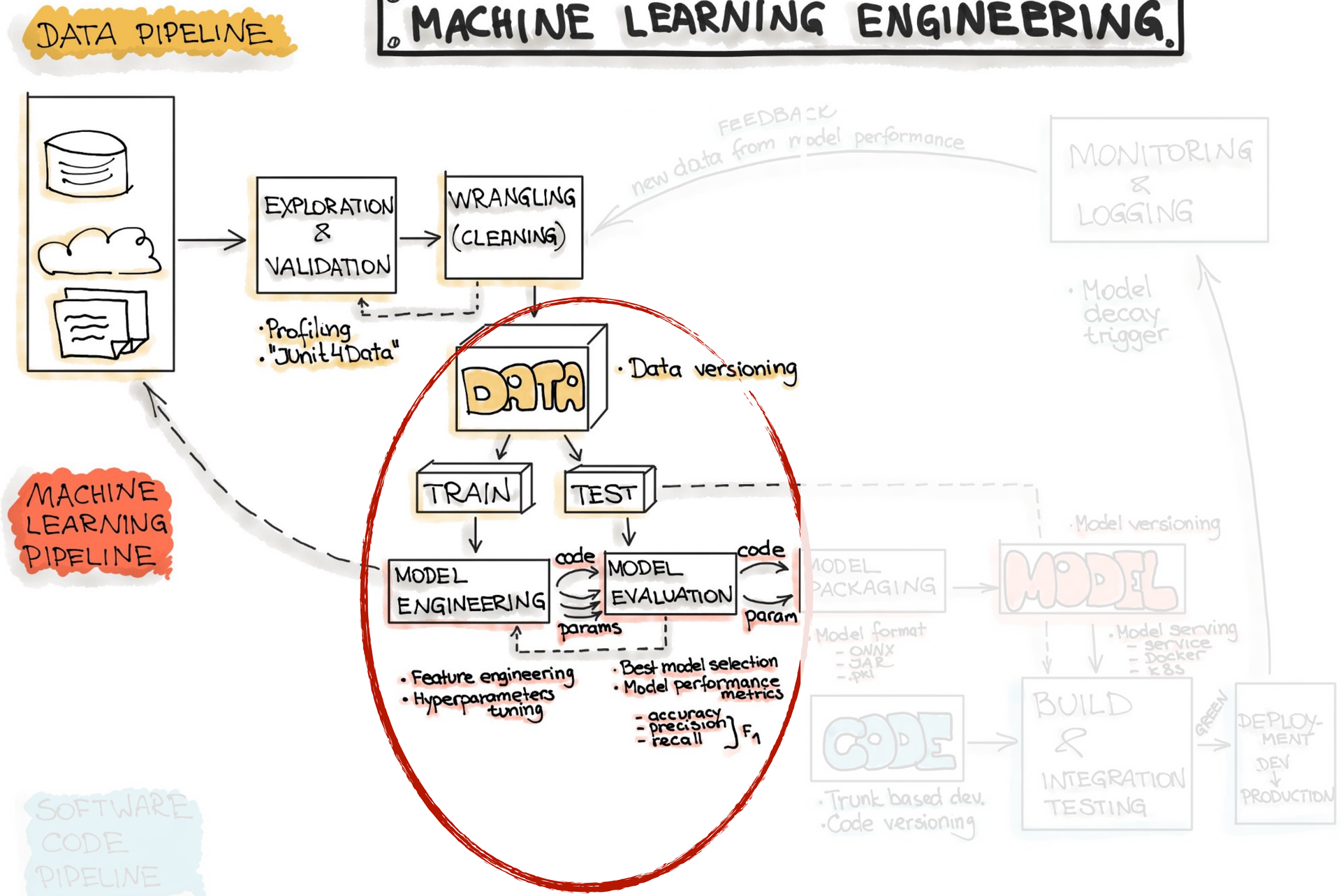
I. Evaluate the Model

J. Model Optimization

K. Repeat step H and I



# MACHINE LEARNING ENGINEERING



A. Data Visualization

B. Data Exploration

C. Data Imputation

D. Define data and Target Variables

E. Split Dataset (Different Strategies)

F. Balance Dataset

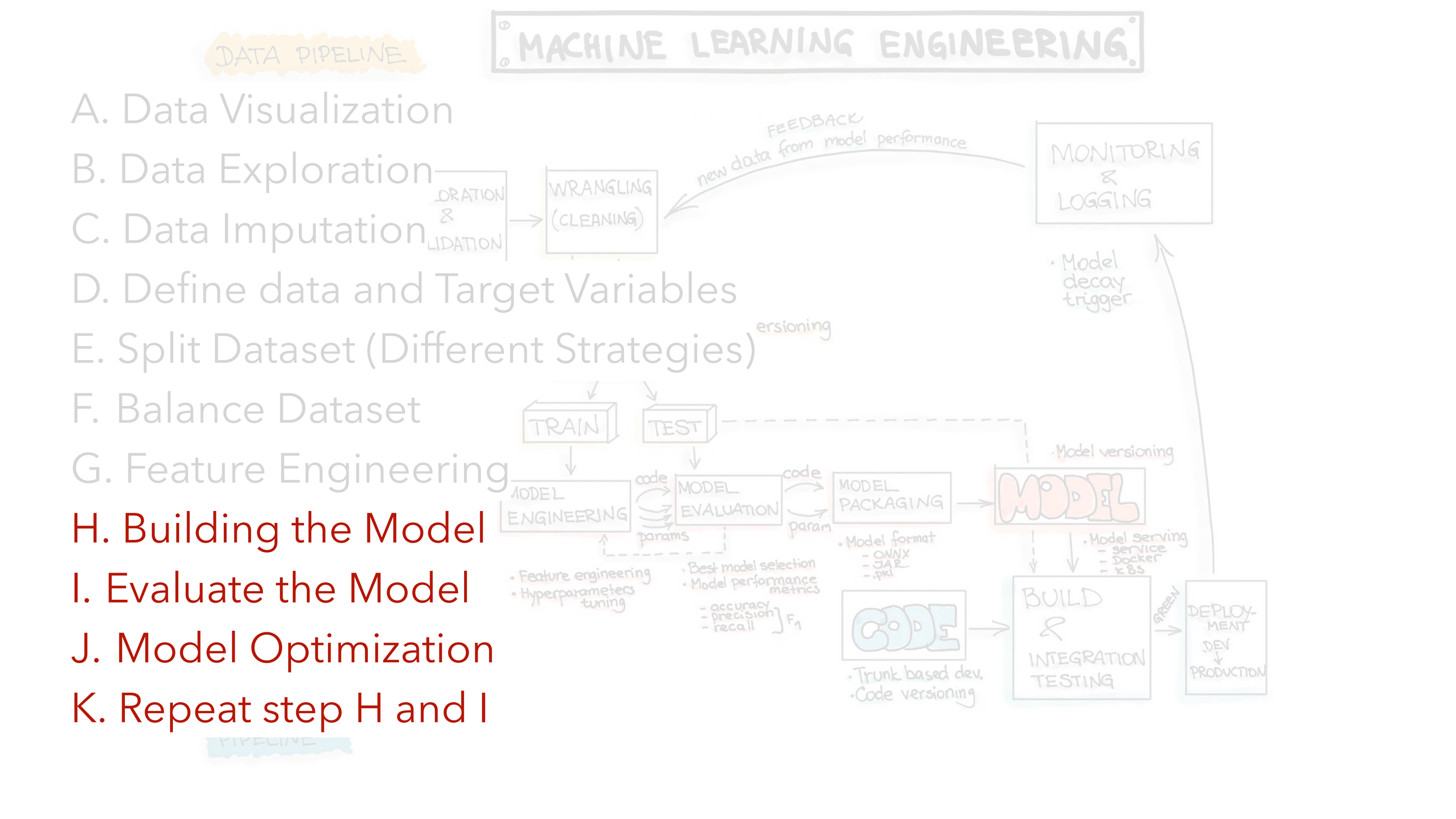
G. Feature Engineering

H. Building the Model

I. Evaluate the Model

J. Model Optimization

K. Repeat step H and I



A. Data Visualization

B. Data Exploration

C. Data Imputation

D. Define data and Target Variables

E. Split Dataset (Different Strategies)

F. Balance Dataset

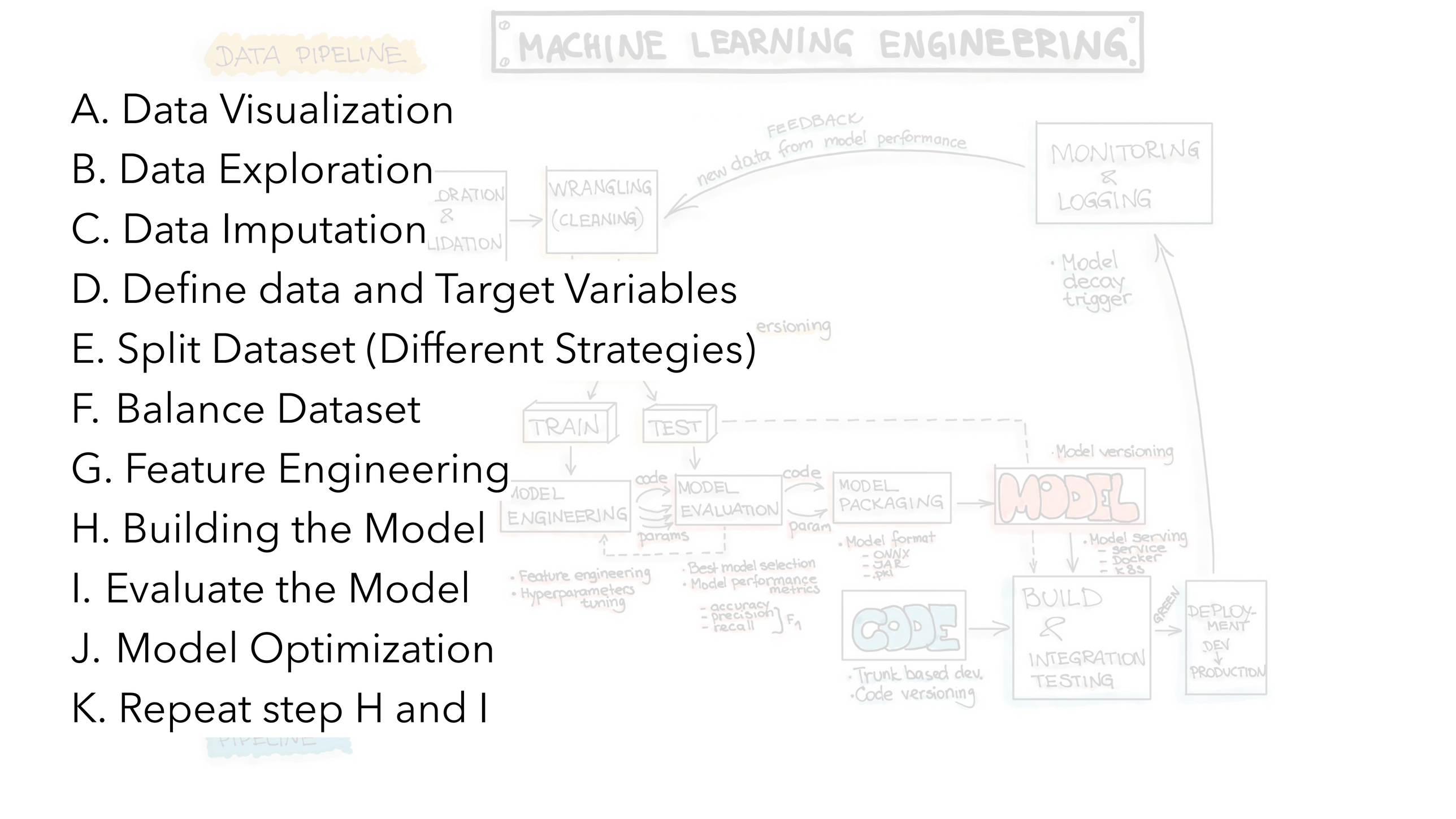
G. Feature Engineering

H. Building the Model

I. Evaluate the Model

J. Model Optimization

K. Repeat step H and I



# Notebook

https://colab.research.google.com/drive/1kLMr6h7t5ZaIUKmOrI5SZF63JwUPQ2EC?usp=sharing

https://colab.research.google.com/drive/1X_6IsvW3oVtrsZgWoTFp780FvfEnZ5im?usp=sharing

https://colab.research.google.com/drive/1fZ4FfNT8Xr5abj_avDdK9O5q6Go-ygtJ?usp=sharing

# Dataset

https://github.com/Gcatolino/ML_example

# Kaggle link to the Dataset

https://www.kaggle.com/sabinero/heart-desease-classification-95-15-recall

# Problem?

g.catolino@tue.nl