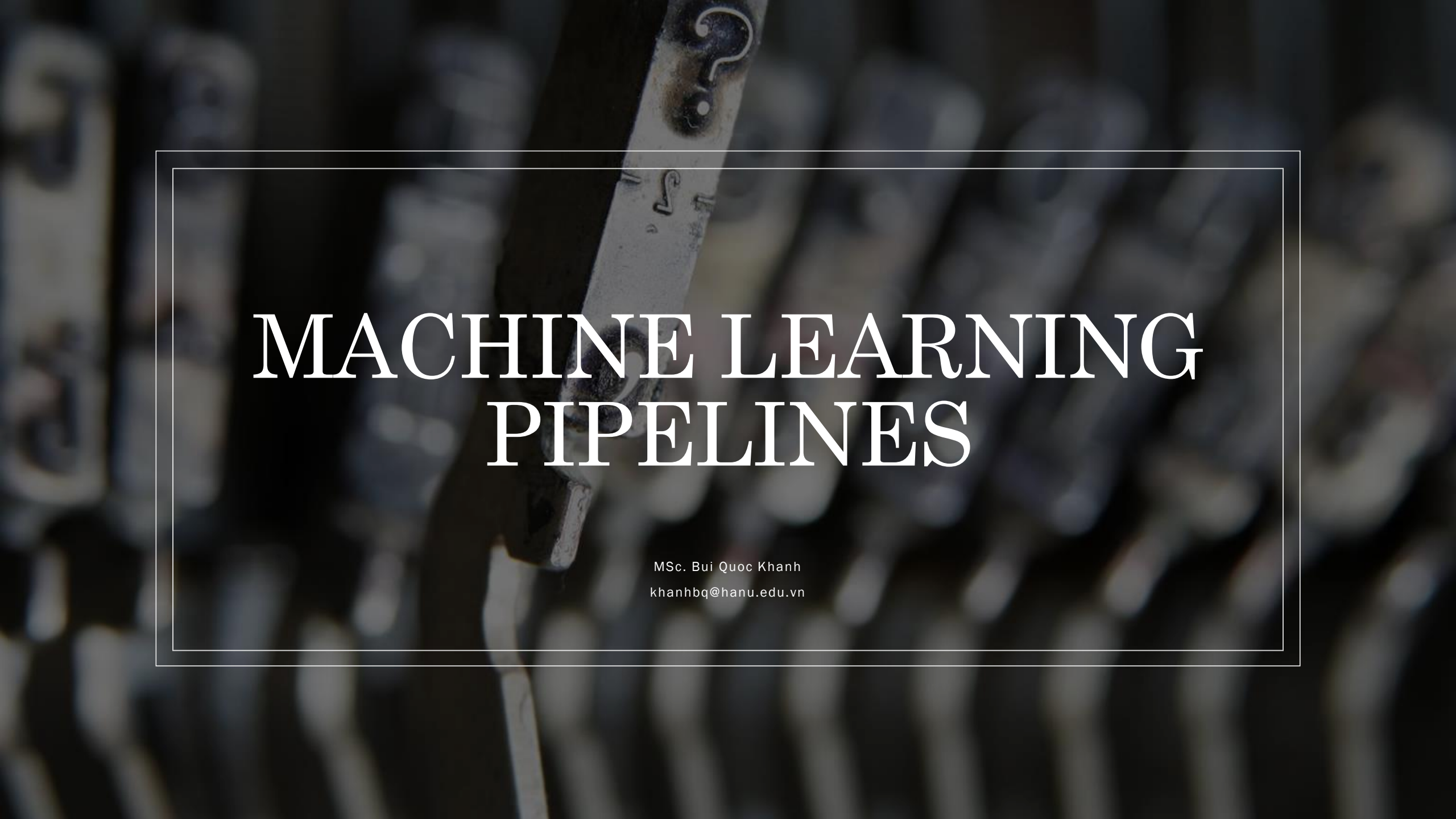


MACHINE LEARNING PIPELINES

MSc. Bui Quoc Khanh

khanhbq@hanu.edu.vn

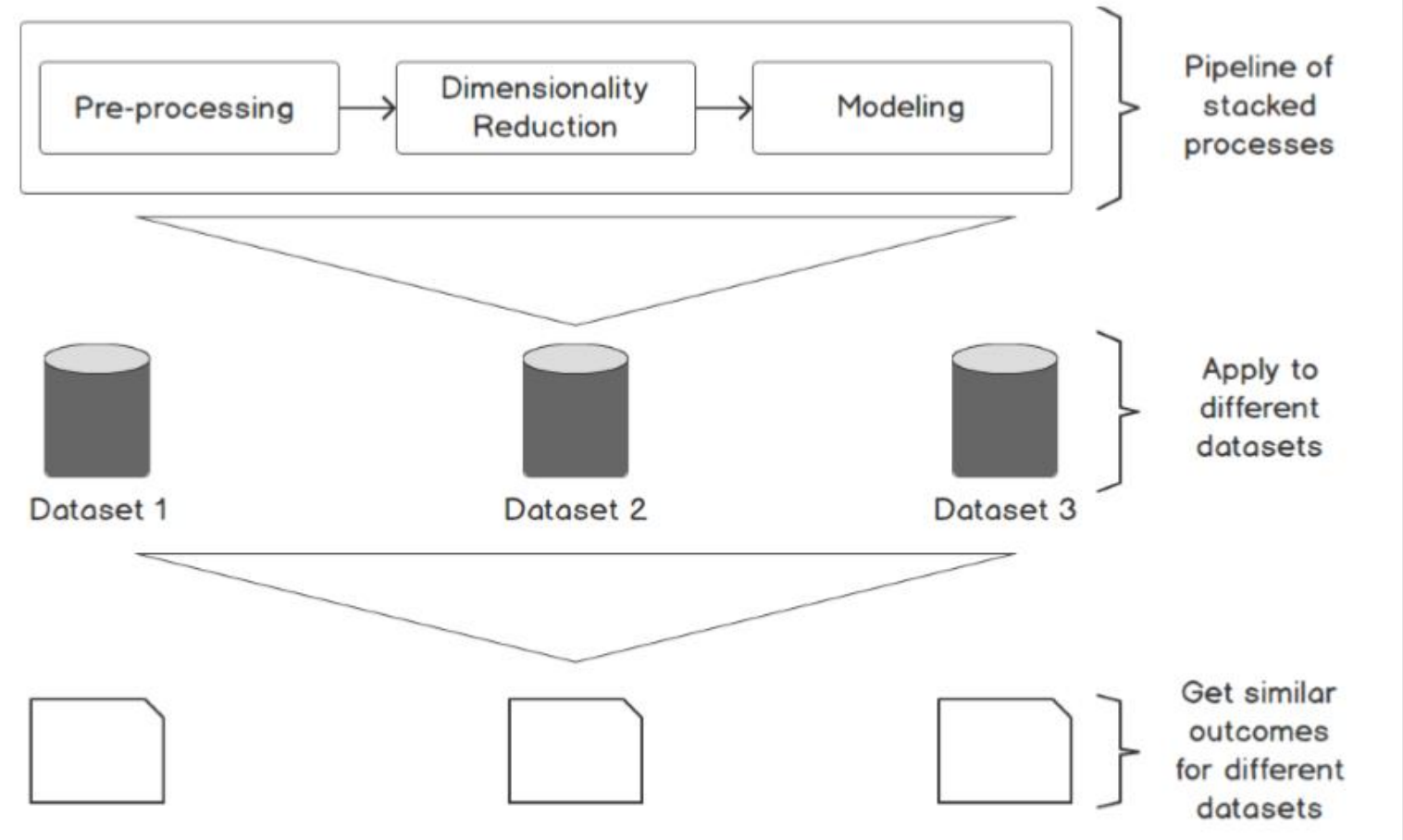
Overview

- Automate machine learning (ML) workflows with the scikit-learn pipeline utility
- Process and transform data with pipeline
- Automate model building processes using pipeline
- Expedite the selection of model parameters using the grid search leveraging pipeline

Context

- Machine Learning is all about performing various experiments to find the right combination of parameters and hyperparameters and enabling the extraction of performance from the models
- Time-consuming - many different permutations and combinations that have to be tried
- ML pipelines help in automating many of the tasks in the ML workflow

Pipelines


Pre-processing
Dimensionality Reduction
Modeling
Pipeline of stacked processes
Dataset 1
Dataset 2
Dataset 3
Apply to different datasets
Get similar outcomes for different datasets

Automating ML Workflows Using Pipeline

Data Preprocessing	• Applying data processing using Pipeline. In this step we use scaling and categorical variable transformation
Dimension Reduction	• Adding a second layer of dimension reduction after the processing layer
Model Building	• Fitting a simple logistic regression model after data processing and dimension reduction
Spot Checking Multiple Models	• Automate the task of finding the best model
Automating Grid Search	• Automating parameter search using grid search

Automating Data Preprocessing Using Pipelines

- **OneHotEncoder()** ~ pandas.get_dummies: transforms categorical variables to a special format called one-hot encoded format
  - Argument: ***handle_unknown*** - *enables the processing of values that were not present in the dataset used for fitting the function*
  - *E.g. the unique values in the dataset that was used to fit the function were A and B*
    - *new dataset, where we applied the .transform() method, had unique values, A and C*
    - *handle_unknown = ignore: ignore such exceptions. The function creates a row containing zeros when such exceptions are encountered*
- ***ColumnTransformer() =*** *transformed categorical data + concatenate with numerical data*
  - *Argument:* ***transformers***

ML Pipeline with Processing and Dimensionality Reduction

- Principal Component Analysis (PCA)
  - Argument: **estimator** - sequentially chain together multiple processes, such as feature extraction, feature normalization, and dimensionality reduction

ML Pipeline for Modeling and Prediction

- Pipeline enables us to build all-encompassing functions in a single engine
- When classifiers are introduced into the estimator, the estimator also inherits many of the functions of the classifiers, such as scoring and predicting
- Estimators can also be used to chain together classifiers such as logistic regression, KNN, or random forest classifiers along with the transformation steps

ML Pipeline for Spot-Checking Multiple Models

- One critical decision point in the data science life cycle is determining what model to try in what scenario.
- This decision of what model to use in what scenario is arrived at after different experiments with multiple models.
- This process is called spot-checking models

ML Pipelines for Identifying the Best Parameters for a Model

- An important step in the data science workflow is to fine-tune a model by trying out different parameters of the model
- Improve performance metrics such as the accuracy or recall of the model
- Cross-validation
  - Split the training set into multiple parts and fit a model on different parts of the dataset, leaving aside one part for validating the result. The result that we get will be the average of the results obtained on all the left-out parts
- Grid search
  - defining a grid of model parameters to try on the model. Using grid search, we find the best permutations of model parameters that can produce the most optimal result.