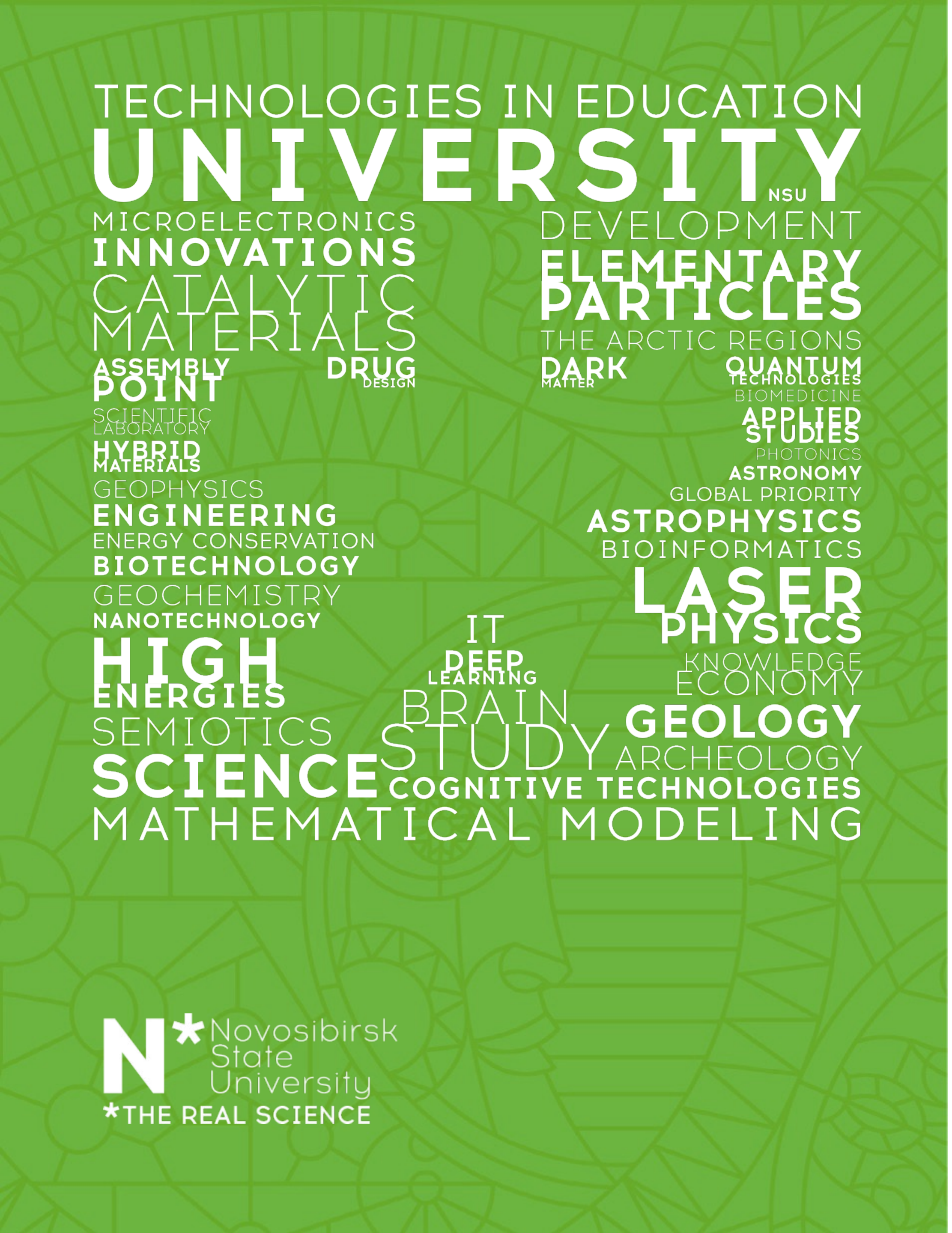


Introduction to Machine Learning

Luu Minh Sao Khue

Department of Mathematics and Mechanics

Course Overview

- Github: https://github.com/luumsk/NSU_ML
- Email: khue.luu@g.nsu.ru
- Telegram: @khueluu
- Notes:
  - Theory + practice (*)
  - Lectures + assignments + personal final project
  - Extra points

Get to know Machine Learning

What is Machine Learning?

Applications of Machine Learning

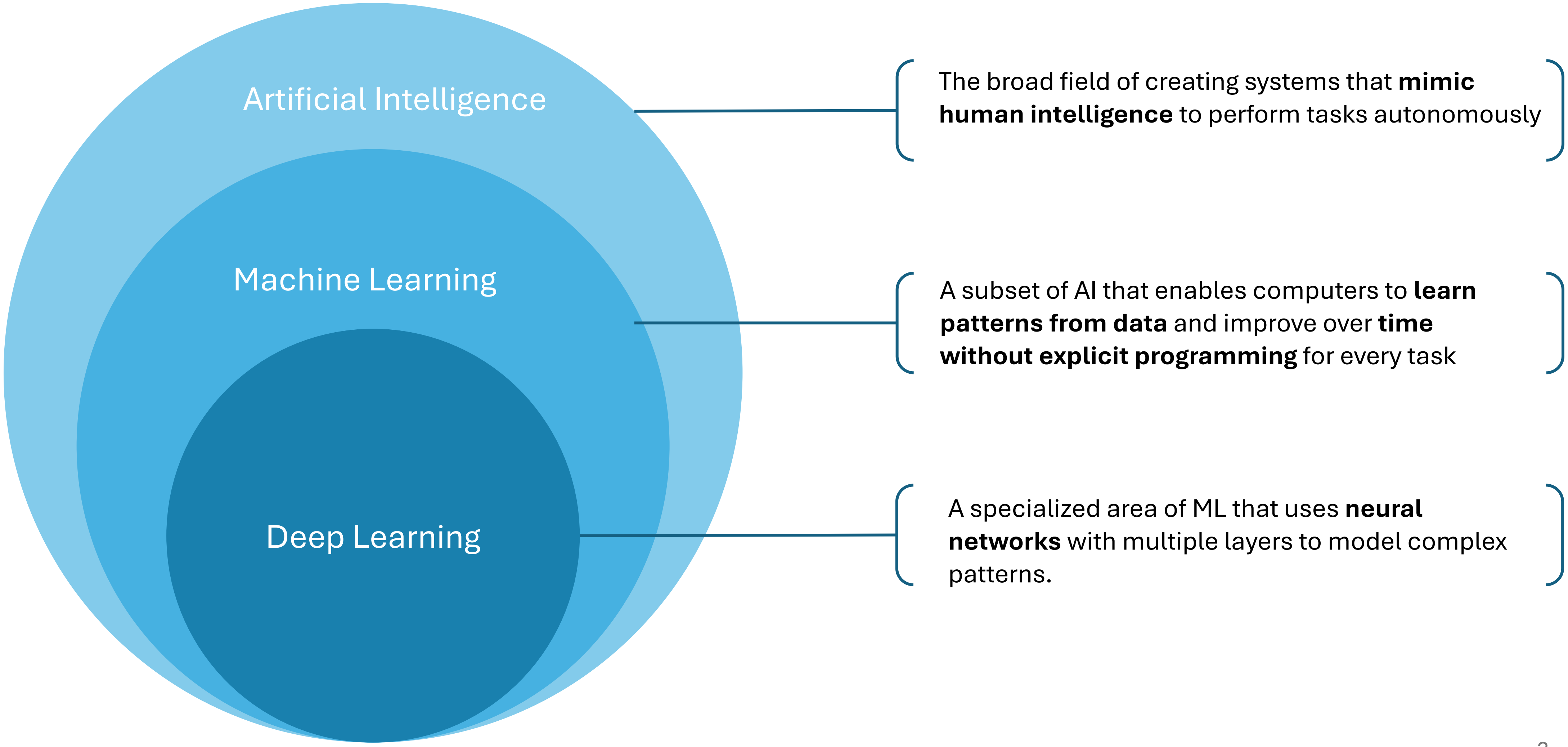
Types of Machine Learning

Machine Learning pipline

Component of a Machine Learning pipeline

2

What is machine learning?



Artificial Intelligence

The broad field of creating systems that **mimic human intelligence** to perform tasks autonomously

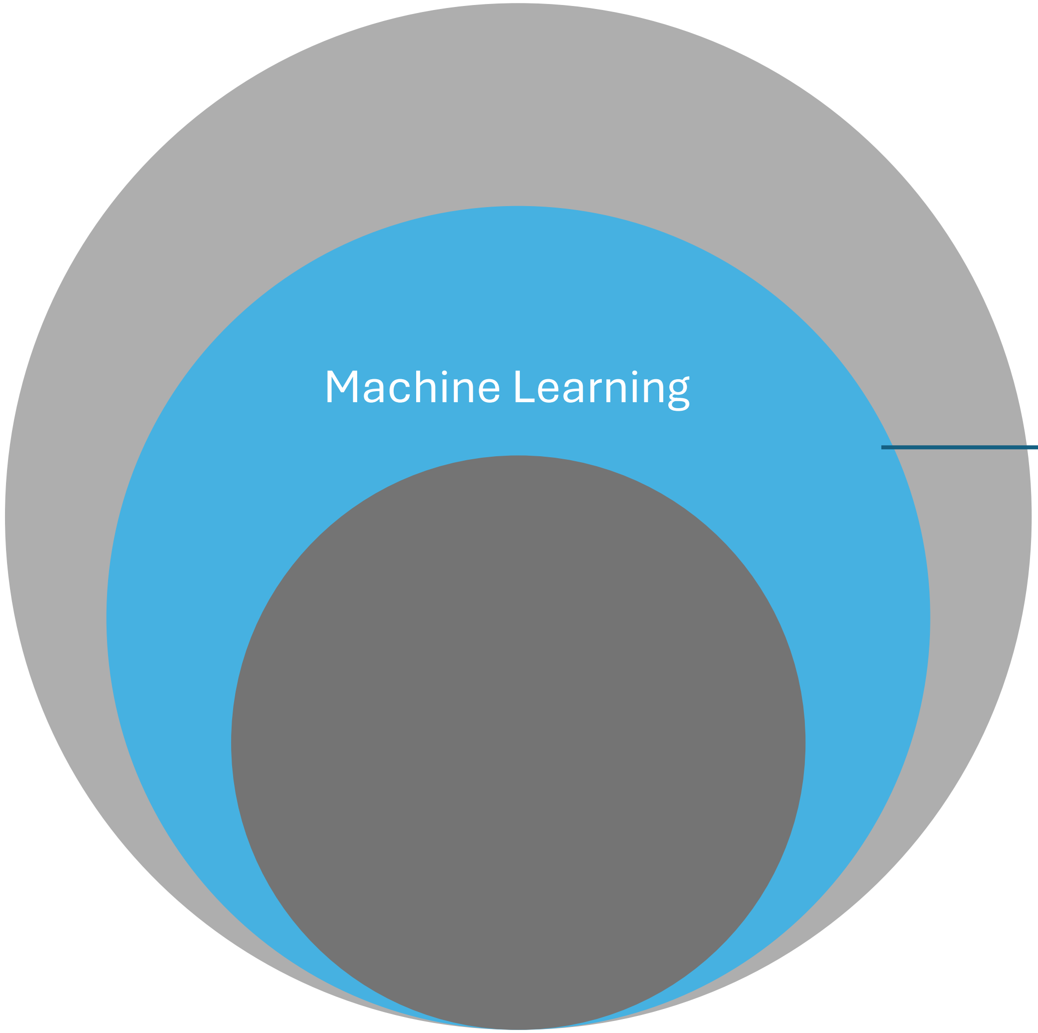
Machine Learning

A subset of AI that enables computers to **learn patterns from data** and improve over **time without explicit programming** for every task

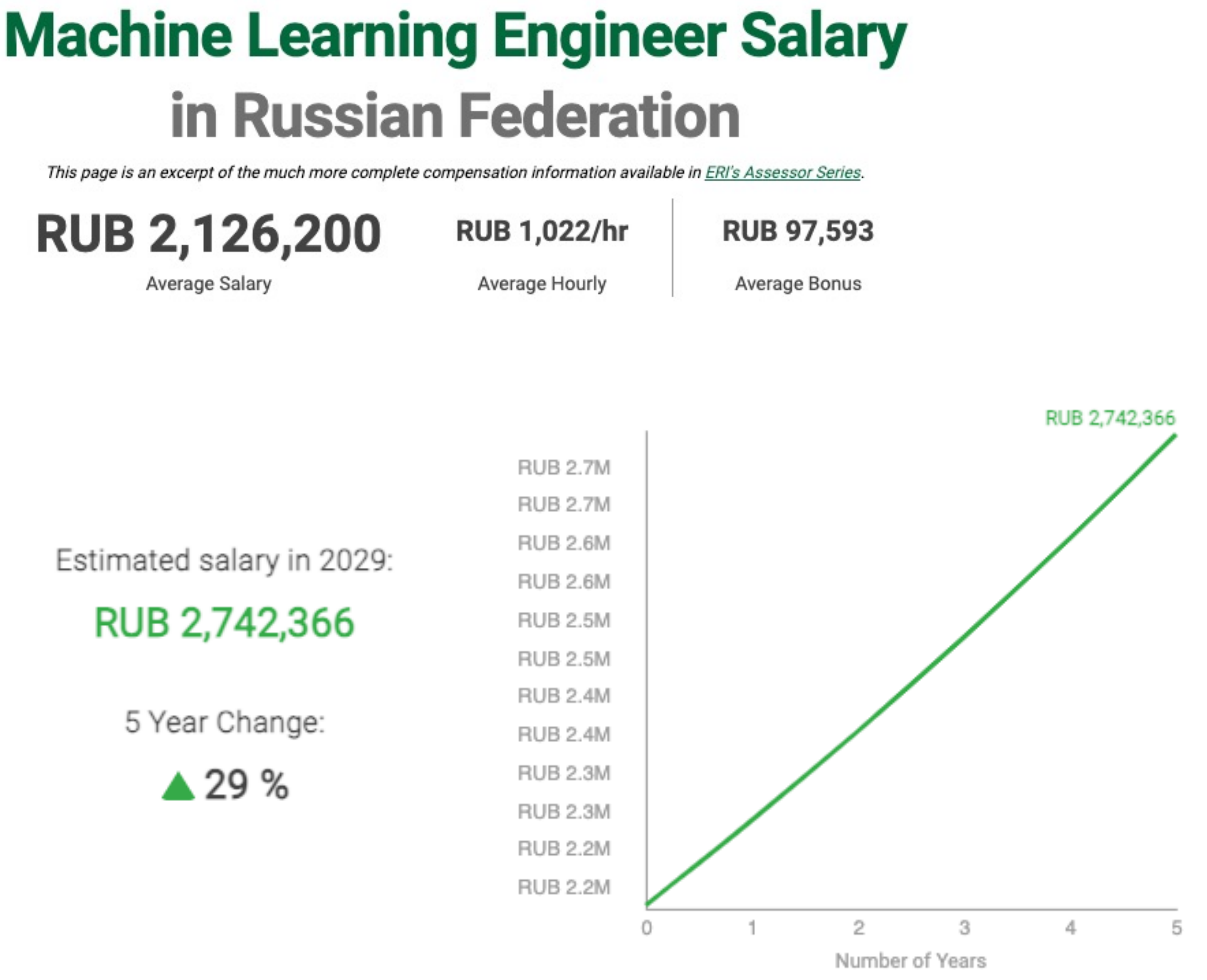
Deep Learning

A specialized area of ML that uses **neural networks** with multiple layers to model complex patterns.

What is machine learning?



A subset of AI that enables computers to **learn patterns from data** and improve over time **without explicit programming** for every task



Machine Learning Engineer Salary in Russian Federation

This page is an excerpt of the much more complete compensation information available in ERI's Assessor Series.

RUB 2,126,200
Average Salary

RUB 1,022/hr
Average Hourly

RUB 97,593
Average Bonus

Estimated salary in 2029:

RUB 2,742,366

5 Year Change:

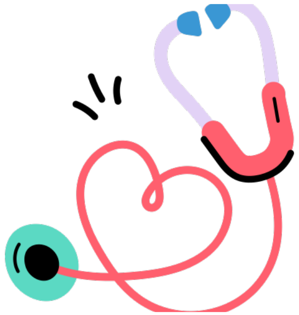
▲ 29 %

Job titles

- Machine Learning Engineer
- Data Scientist
- Research Scientist (Machine Learning)
- Applied Machine Learning Scientist
- Machine Learning Analyst
- Lead Machine Learning Scientist
- Director of Machine Learning
- Chief AI Officer

Source: https://www.erieri.com/salary/job/machine-learning-engineer/russian-federation

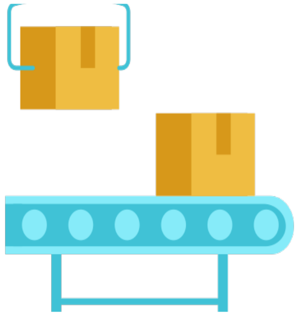
Applications of machine learning



Healthcare

Finance

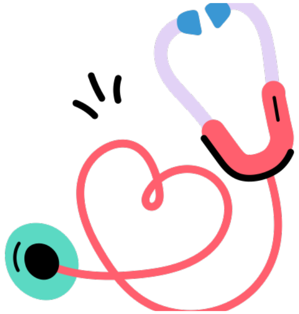
E-commerce



Manufacturing

Transport and Logistic

Applications of machine learning

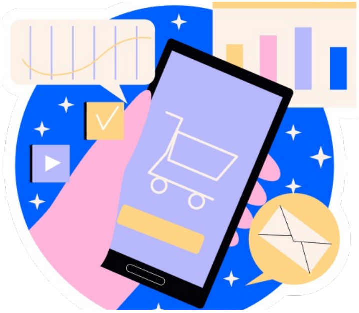


Healthcare

- Diagnostics
- Predictive Analytics
- Personalized Medicine

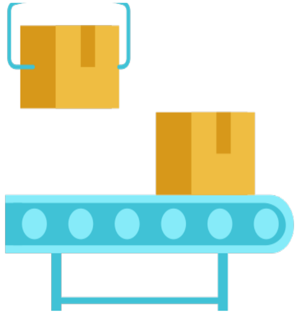
Finance

- Fraud Detection
- Credit Scoring
- Algorithmic Trading



E-commerce

- Recommendation Engines
- Customer Segmentation
- Inventory Management



Manufacturing

- Predictive Maintenance
- Quality Control
- Supply Chain Optimization

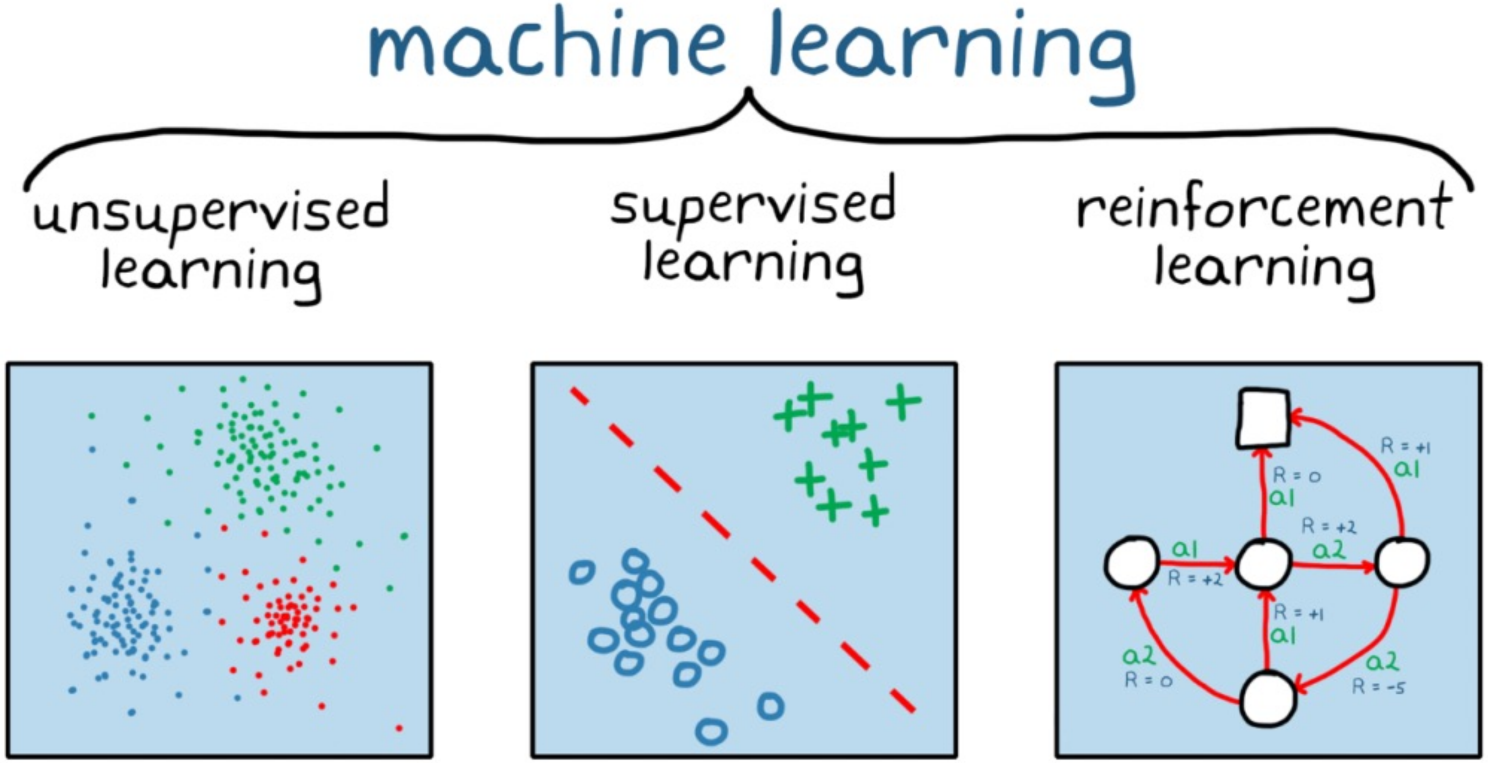
Transport and Logistic

- Route Optimization
- Traffic Prediction

What ML applications you know/being used in your country?

What ML applications you are interested in?

Types of machine learning



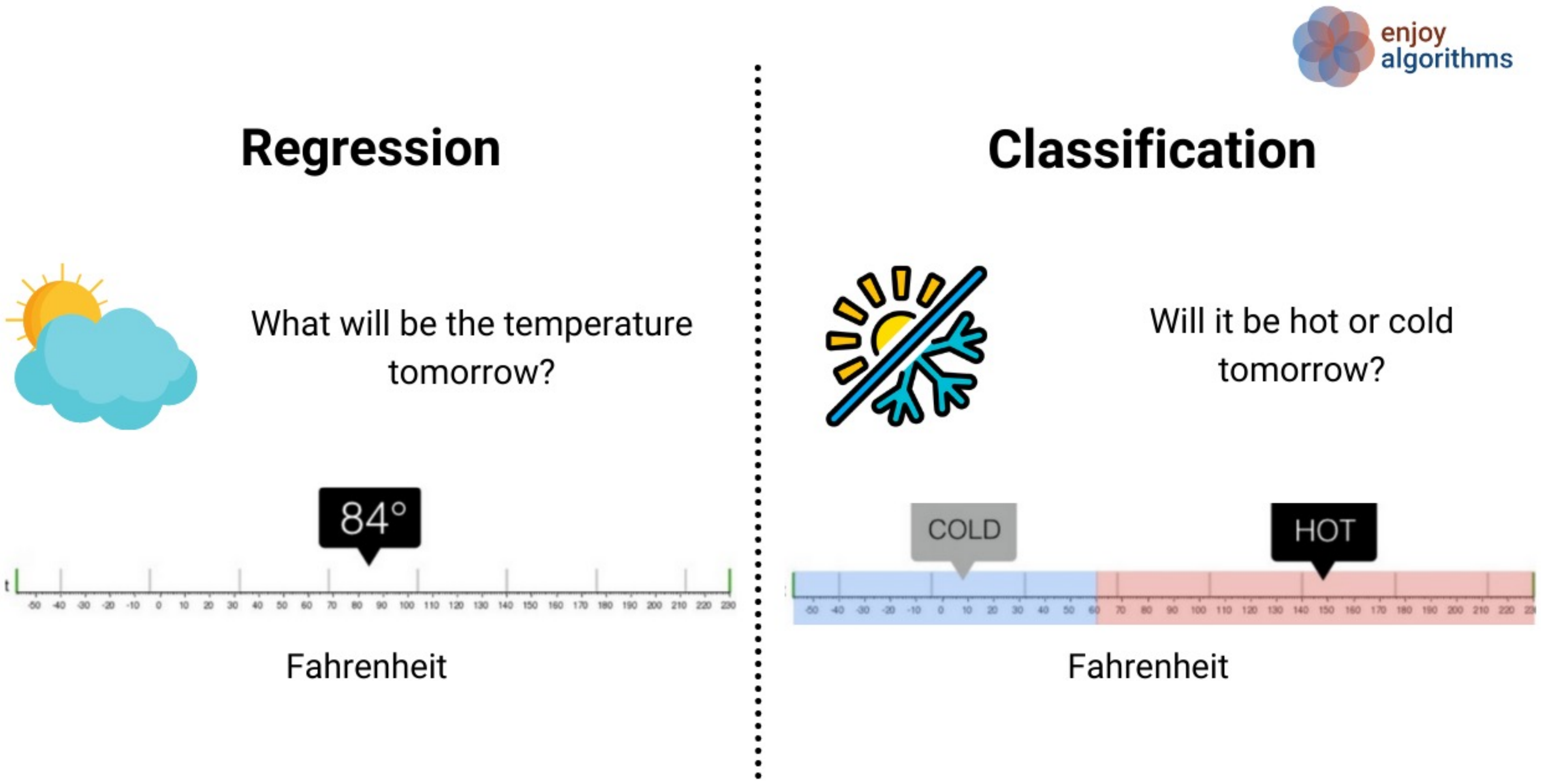
Unlabeled data
No feedback
Find hidden structure

Labeled data
Predict next value

Decision process
Reward system
Learn series of actions

Source: https://www.theaiops.com/machine-learning-compare-supervised-learning-vs-unsupervised-learning-vs-reinforcement-learning/

Types of Supervised Learning



Regression

What will be the temperature tomorrow?

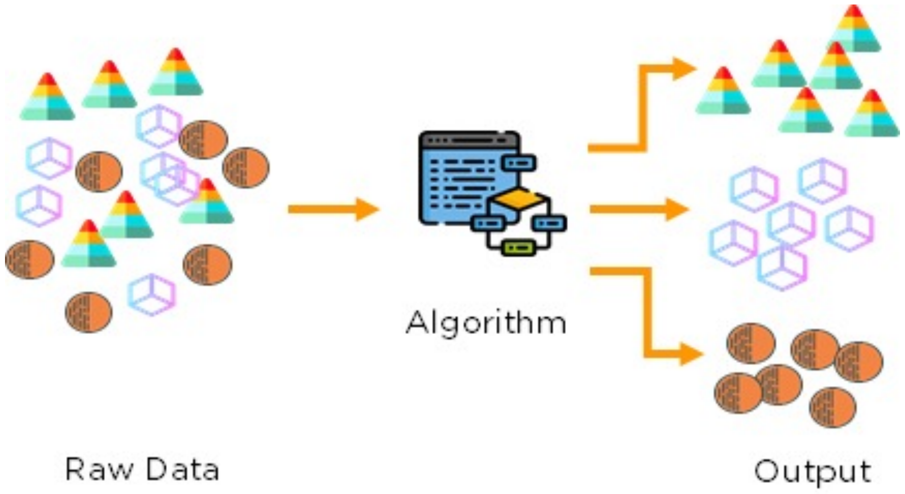
Classification

Will it be hot or cold tomorrow?

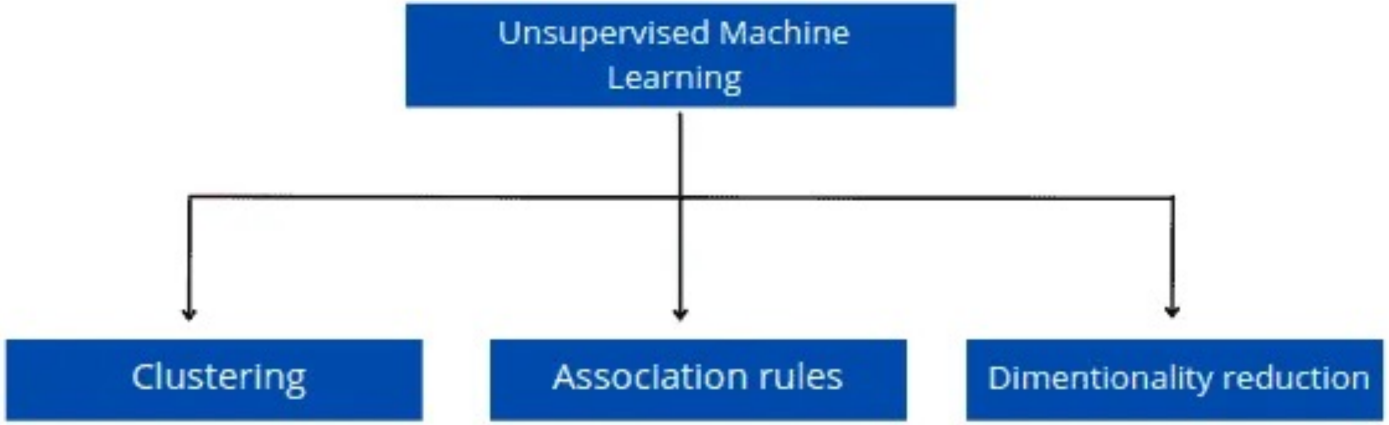
Binary/
Multi-class

Source: https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

Types of Unsupervised Learning



Source: https://medium.com/analytics-vidhya/beginners-guide-to-unsupervised-learning-76a575c4e942



Source:https://hands-on.cloud/ml-unsupervised-learning-guide/

What type of ML problem do you think it is?

Scenario 1: A bank wants to predict whether a new applicant will default on a loan. They have data on previous applicants, including income, credit score, and past financial behavior.

Question: What type of machine learning problem is this, and why?

What type of ML problem do you think it is?

Scenario 2: An e-commerce company wants to group its customers into different segments to better tailor its marketing strategy. They have information on customer behavior, such as purchase frequency, spending habits, and browsing history, but they do not know how which customer belongs to which group.

Question: What type of machine learning problem is this, and why?

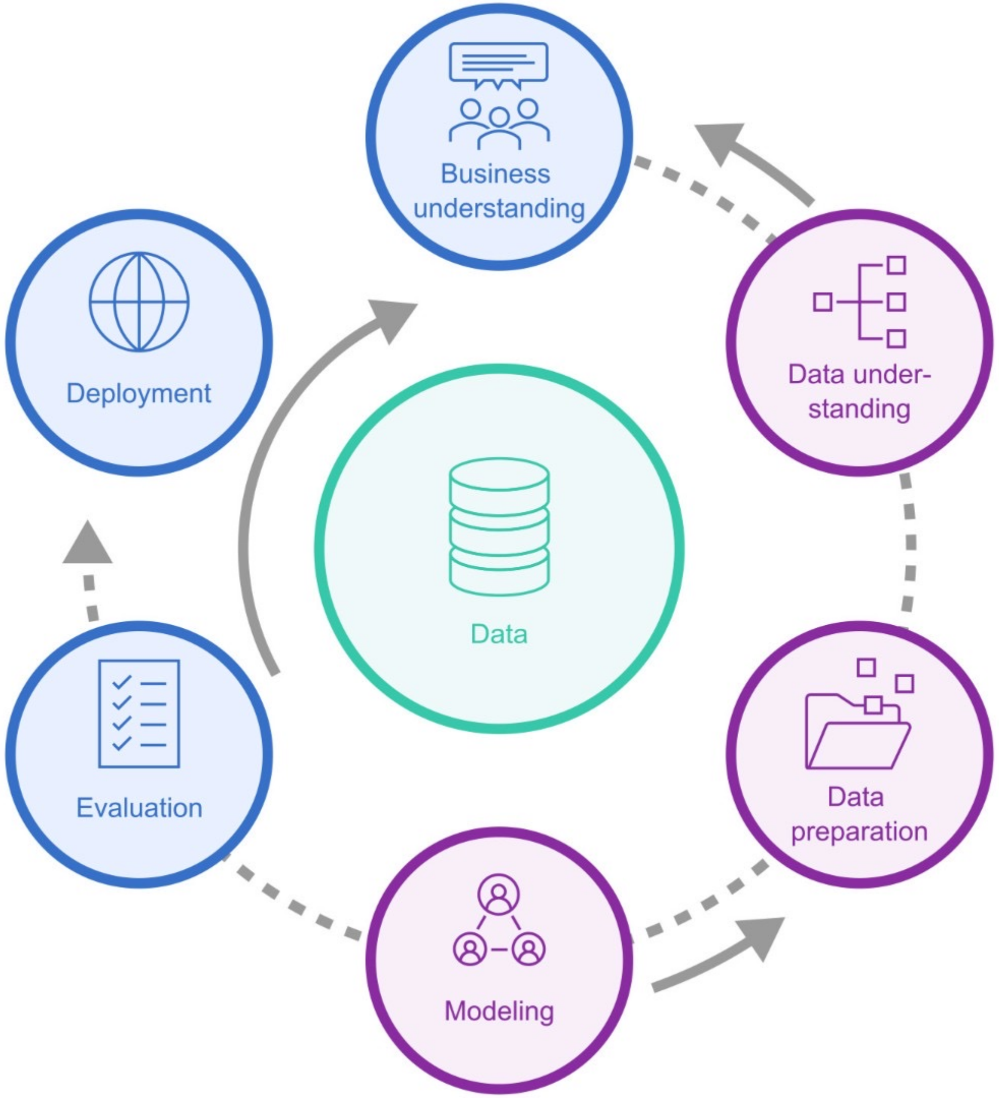
What type of ML problem do you think it is?

Scenario 3: Imagine a robotic vacuum cleaner that navigates around a house, cleaning floors and avoiding obstacles. The robot has to learn how to move efficiently, avoid furniture, and return to its charging station when its battery is low. The robot doesn't start with any prior knowledge about the house layout or where obstacles are located.

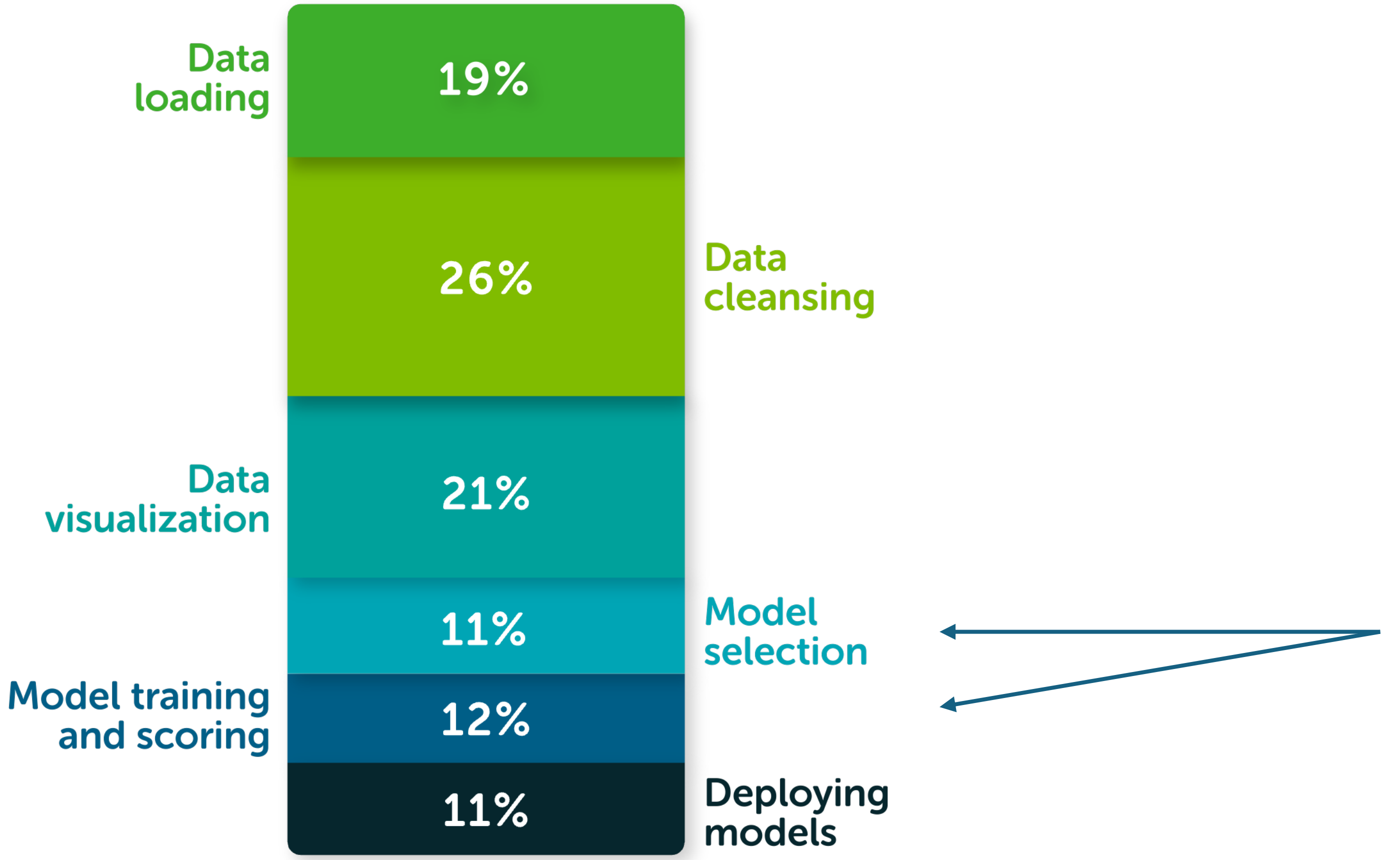
Question: What type of machine learning problem is this, and why?

Machine learning pipline

Cross-Industry Standard Process for Data Mining (CRISP-DM)



Time allocation for ML tasks



Focus of this course

Source: https://www.anaconda.com/resources/whitepapers/state-of-data-science-2020

Components of a machine learning pipeline

- **Input Data**
  - Raw data used to train the model, containing independent variables (predictors) and, in supervised learning, target labels.
- **Features**
  - Selected attributes from the data that represent relevant information, influencing model accuracy and performance.
- **Model**
  - The algorithm or mathematical representation that learns patterns in the data (e.g., linear regression, decision trees).
- **Model Output (Prediction)**
  - The result generated by the model, such as predicted labels, probabilities, or continuous values.
- **Evaluation Metrics**
  - Measures like accuracy, precision, and recall used to assess model performance and generalization.

Some practical aspects

Underfitting vs Overfitting

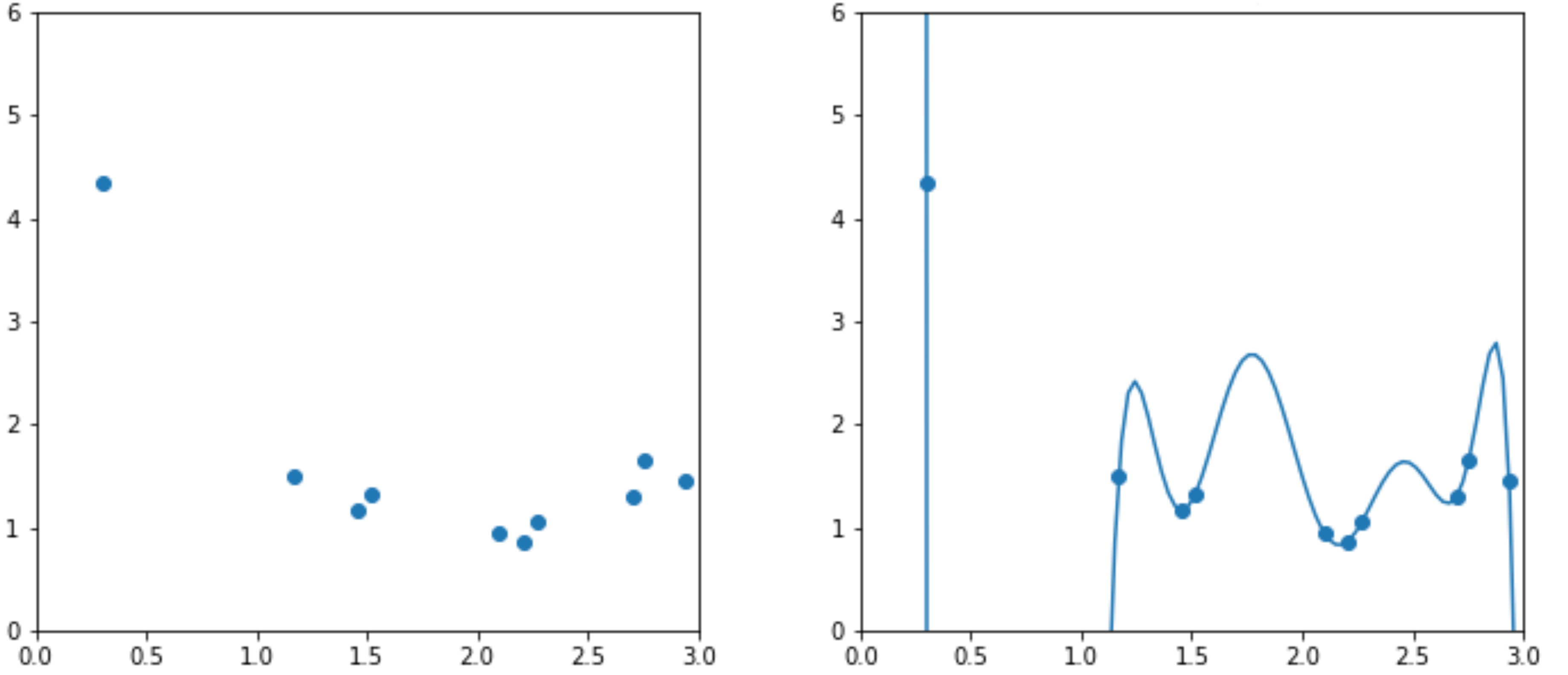
Training, validation, and test dataset splitting

Cross-validation

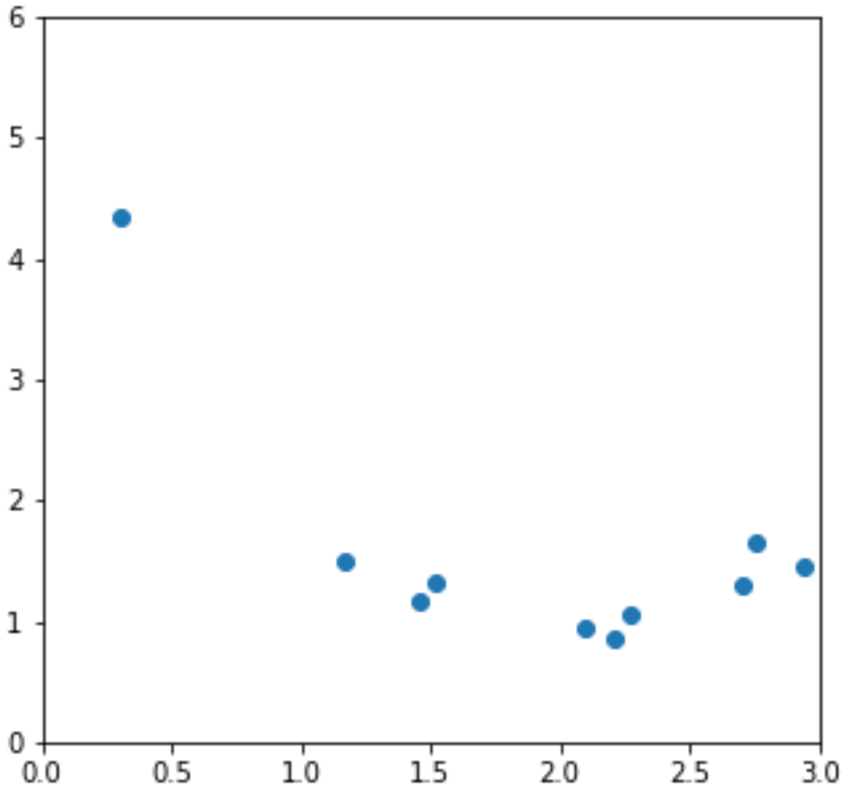
Hyperparameters tuning

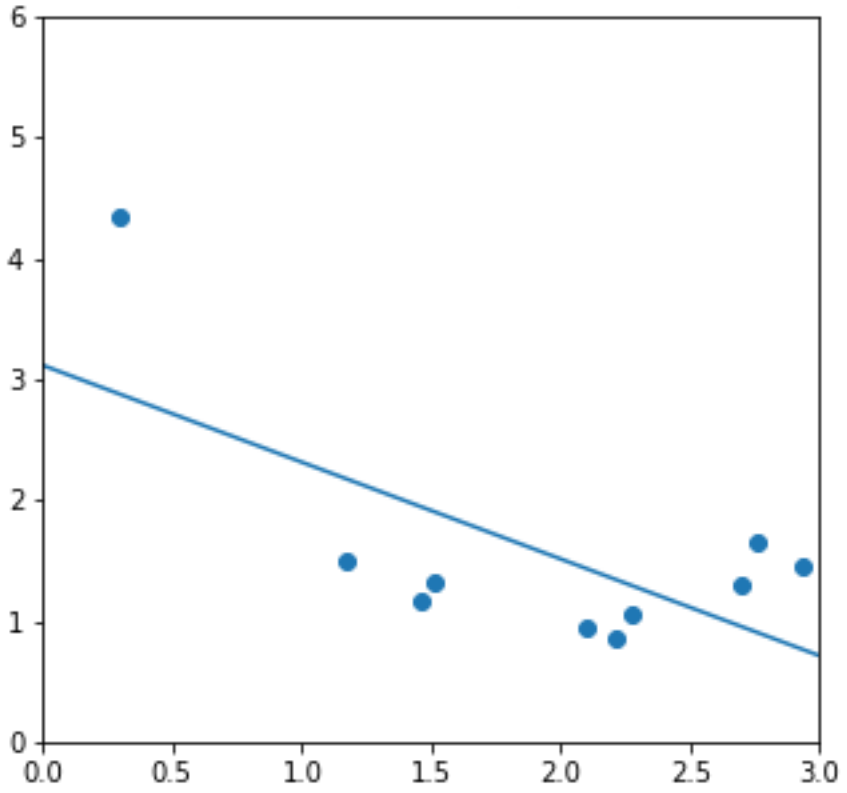
Ensemble method

Choosing the function class

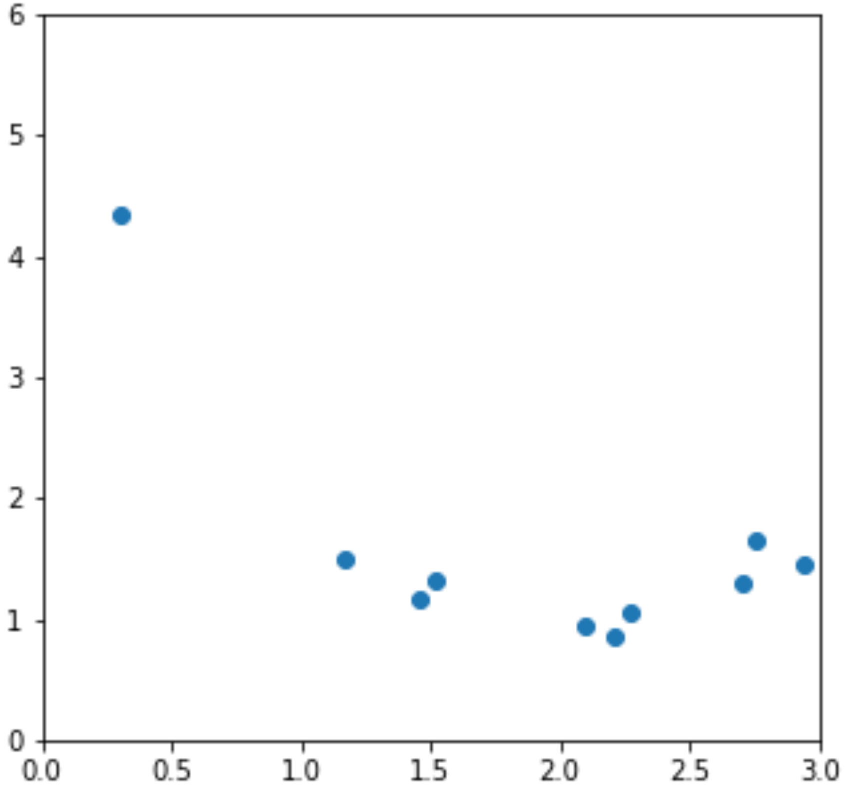


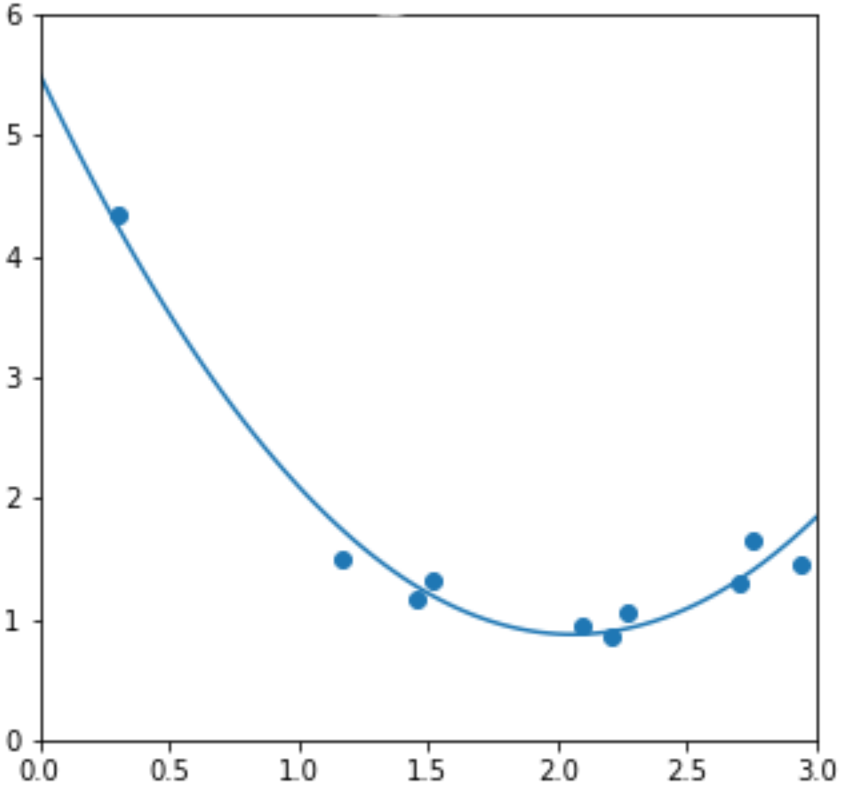
Choosing the function class





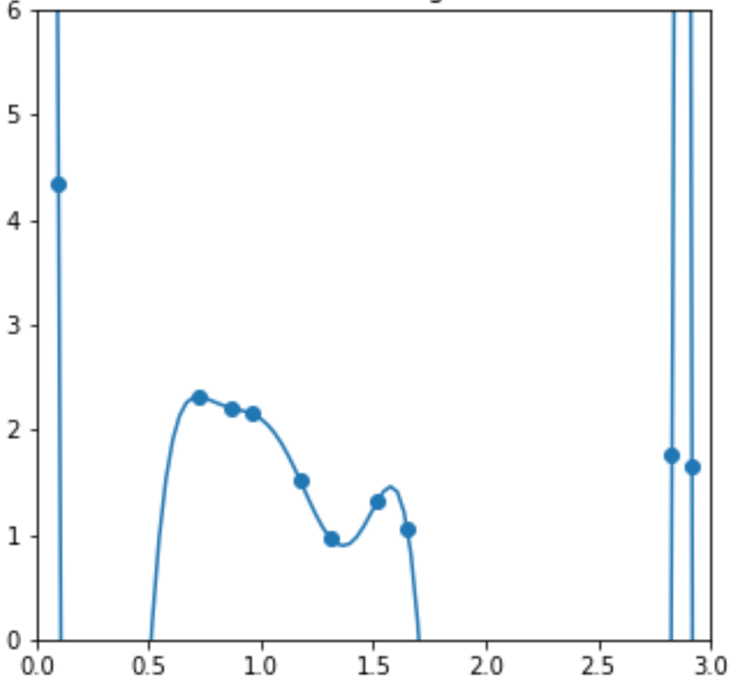
Choosing the function class



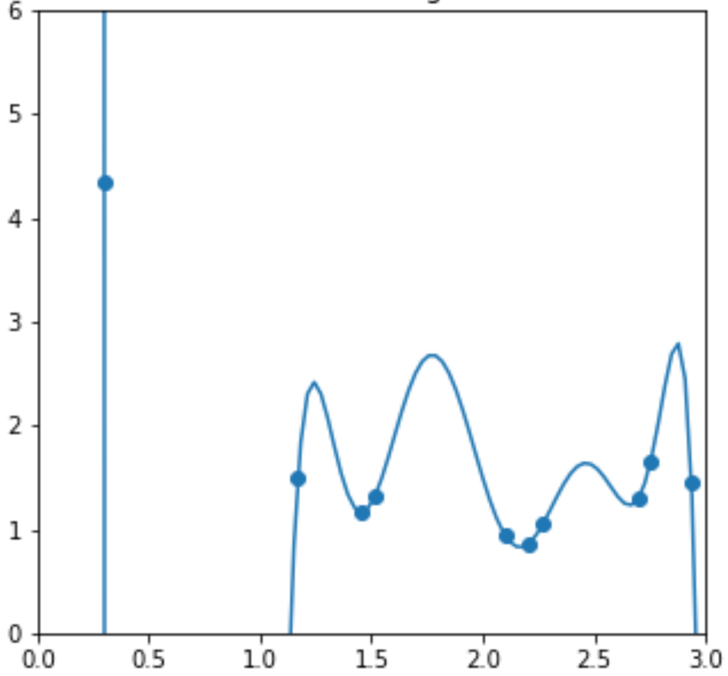


High Variance

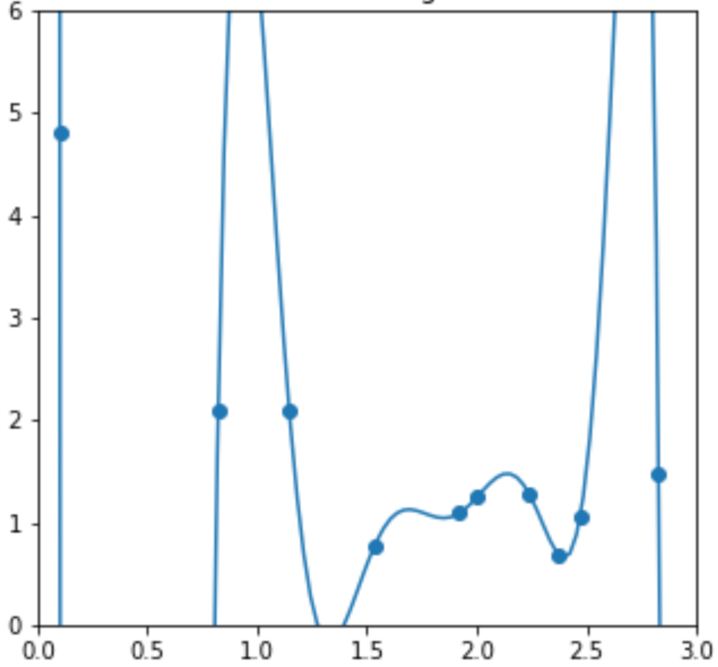
overfitting



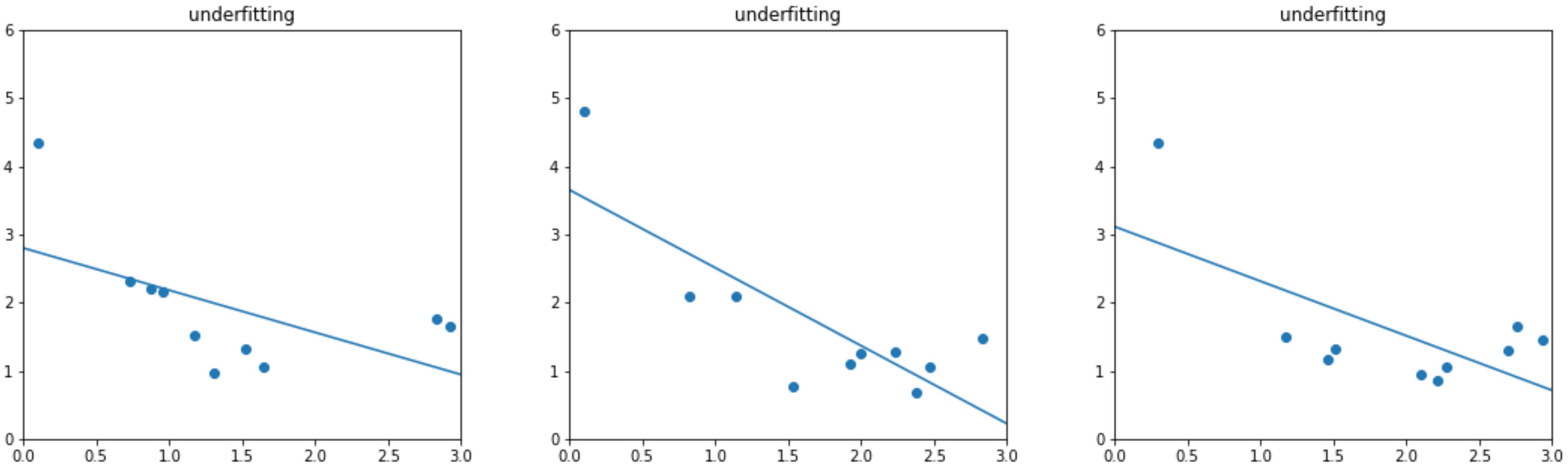
overfitting



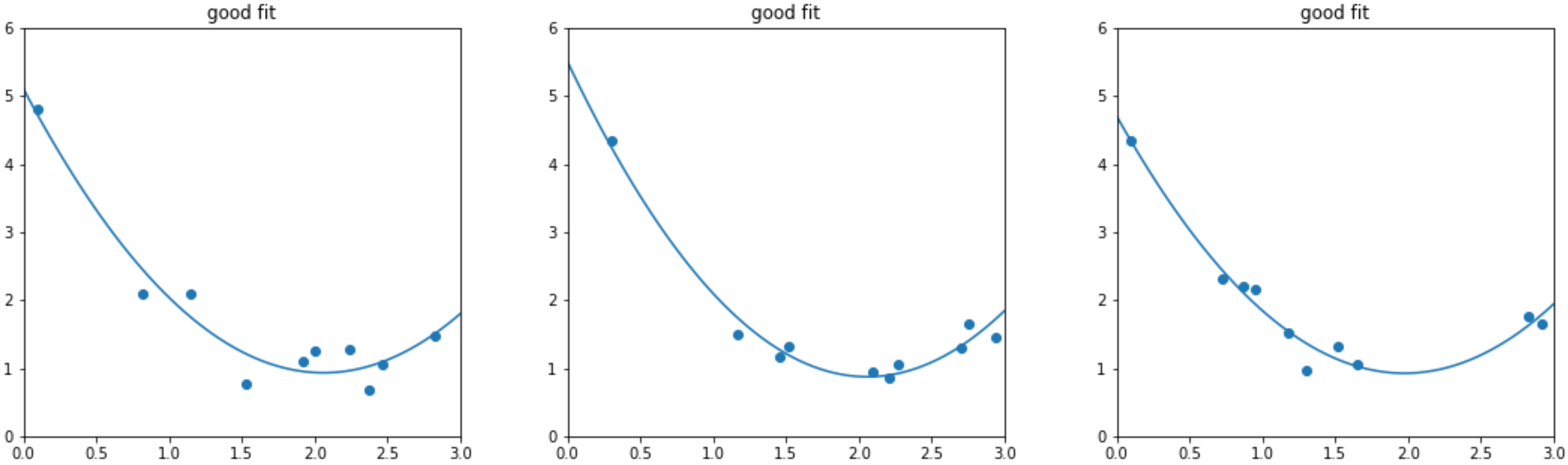
overfitting



High Bias


underfitting
underfitting
underfitting
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Optimal Bias and Variance



Overfitting and Underfitting

- A model with a **high bias** and **low variance** is an **underfit** model. It does not sufficiently represent the statistical relationships in our data.

- A model with **high variance** and **low bias** is an **overfit** model, because it captures relationships that are too specific to the exact data we happen to train it on. These relationships may not exist in the general distribution and are likely spurious.

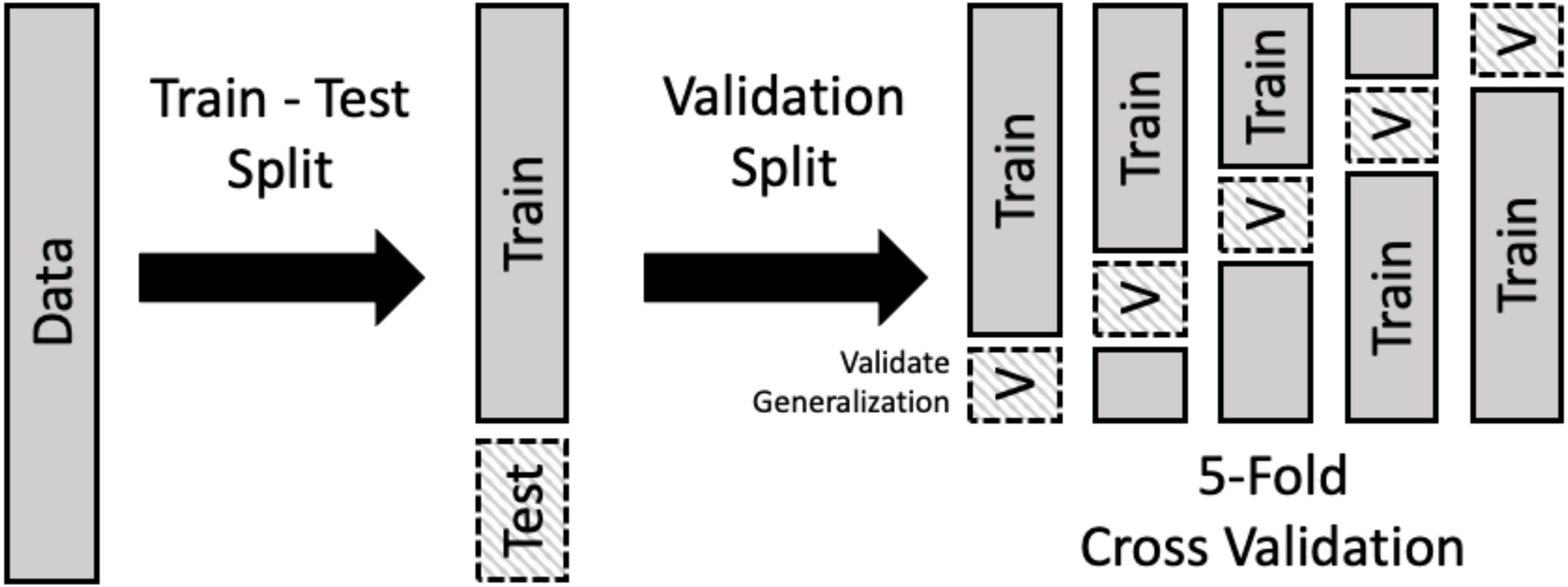
Dataset splitting

- **Training Dataset**
  - Used to train the model by allowing it to learn patterns and relationships.
  - Typically the largest portion of the data (e.g., 60-70%).
- **Validation Dataset**
  - Used to tune model parameters and prevent overfitting.
  - Helps evaluate model performance during training but isn’t used to train the model directly.
- **Test Dataset**
  - A completely separate dataset used for final model evaluation.
  - Provides an unbiased assessment of model performance on new data.

Cross Validation

- **Purpose:** An approach to maximize data usage for training and validation.
- **Process:** The dataset is split into *k* "folds." Each fold is used as a validation set while the remaining *k-1* folds are used for training.
- **Common Approach:** *k*-fold Cross-Validation (e.g., 5-fold), where results are averaged across folds.
- **Benefit:** Provides a more robust evaluation, reducing the risk of overfitting and underfitting

Cross Validation



Source: https://learningds.org/ch/16/ms_cv.html

Hyperparameter Tuning

- **Definition:** The process of optimizing model settings (hyperparameters) that aren’t learned from data, such as learning rate, number of neighbors (in k-NN), or max depth (in decision trees).
- **Goal:** To improve model accuracy, reduce overfitting, and achieve the best possible performance.
- **Methods:**
  - **Grid Search:** Tests all possible combinations of hyperparameters within a predefined range.
  - **Random Search:** Samples random combinations of hyperparameters, offering faster exploration.
- **Tools:** GridSearchCV and RandomizedSearchCV in scikit-learn.

Ensemble Method

- **Definition:** Combines multiple models to improve overall performance by reducing errors and capturing more complex patterns.
- **Types:**
    - **Bagging:** Trains multiple models on random subsets of data (e.g., Random Forest).
    - **Boosting:** Sequentially trains models, each focusing on correcting errors from the previous one (e.g., AdaBoost, Gradient Boosting).
    - **Stacking:** Combines outputs of multiple models through a meta-model to improve predictions.
- **Benefits:** Increases accuracy, robustness, and generalization, often outperforming individual models.

Practice

Sources to learn Machine Learning

Lab work

Sources to learn ML

- Scikit learn: https://scikit-learn.org/stable/
- Kaggle: https://www.kaggle.com/
- Machine Learning Mastery: https://machinelearningmastery.com/
- Books: https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md

Lab

Link:
https://github.com/luumsk/NSU_ML/blob/main/Labs/lab1.ipynb

References

- https://acropolium.com/blog/machine-learning-in-healthcare-use-cases-benefits-and-success-stories/
- https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning
- https://hands-on.cloud/ml-unsupervised-learning-guide/
- https://medium.com/analytics-vidhya/beginners-guide-to-unsupervised-learning-76a575c4e942
- https://www.erieri.com/salary/job/machine-learning-engineer/russian-federation
- https://www.tealhq.com/job-titles/machine-learning-scientist
- https://learningds.org/ch/16/ms_cv.html