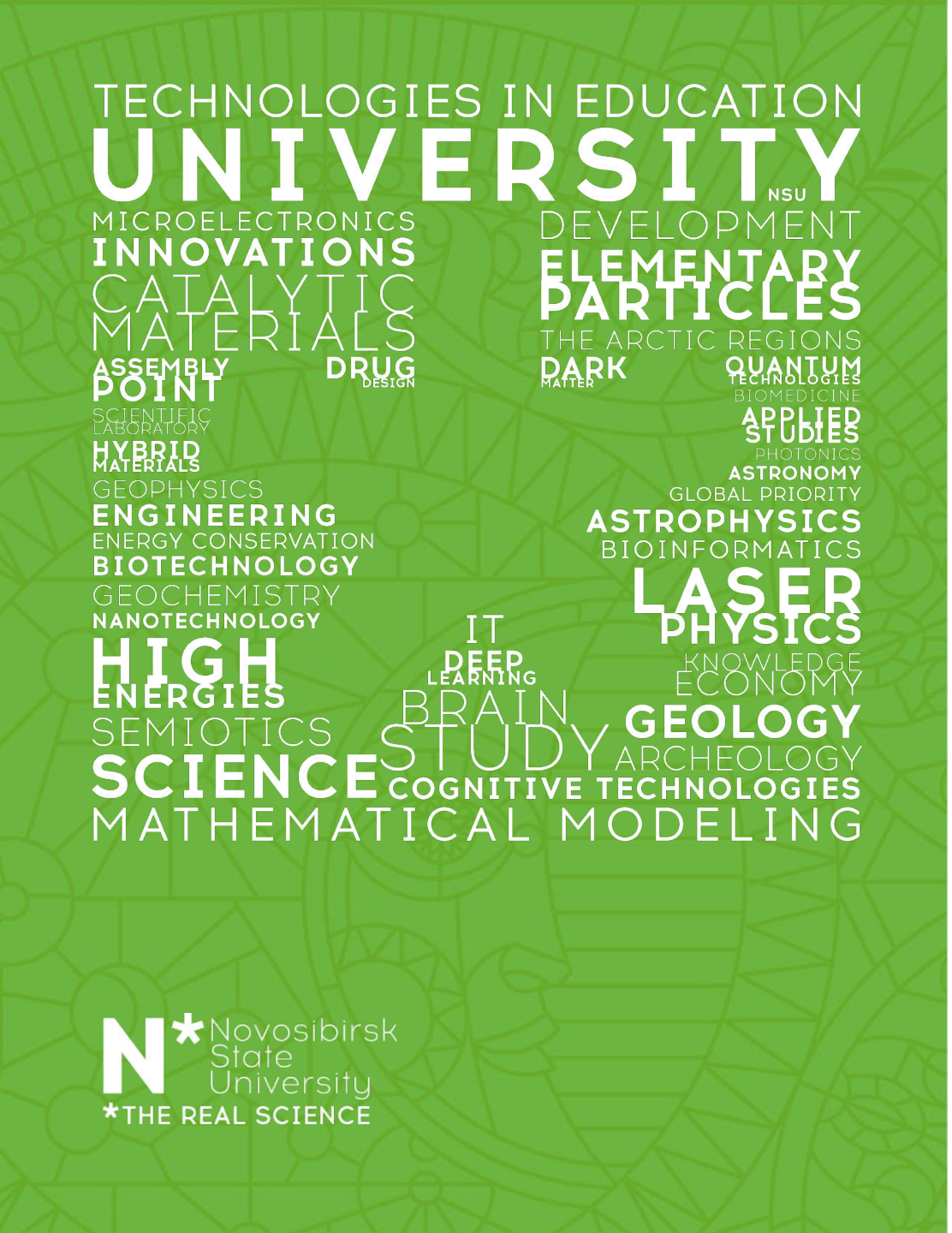


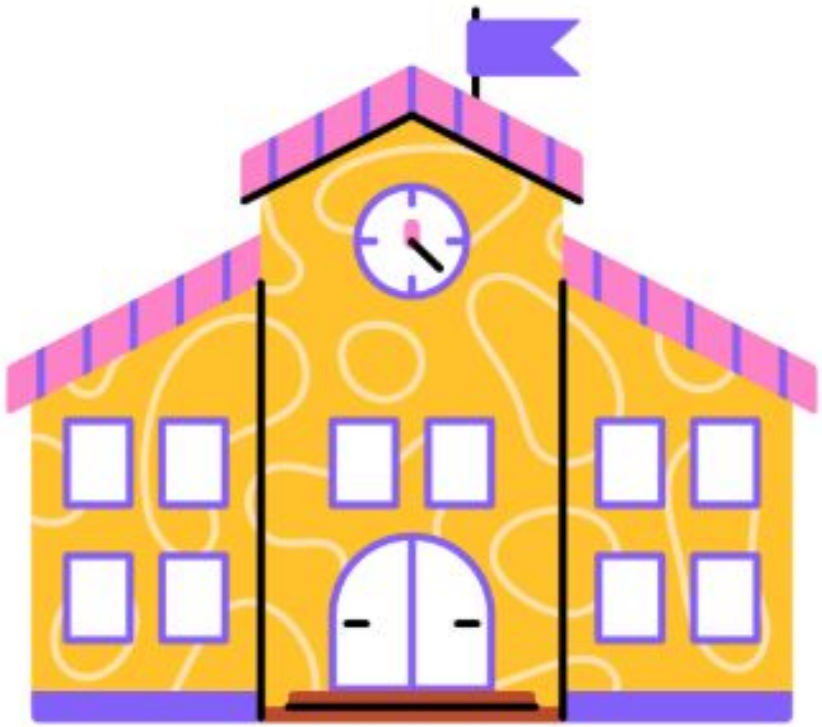

# Introduction to Machine Learning

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

You drive to campus every day, but winter snow often changes how long the trip takes.



Can you predict how long your next trip will take if the snow is 40 mm deep?

| Snow (mm) | Time (minutes) |
|---|---|
| 0 | 10 |
| 5 | 11 |
| 10 | 12 |
| 15 | 13 |
| 20 | 14 |
| 40 | ? |

You are developing a university app that tracks each student's study hours to help identify those at risk of failing.

Can you use this data to predict whether a student will pass or fail the exam?



| Study hours | Attendance (%) | Pass/Fail |
|---|---|---|
| 27 | 45 | F |
| 30 | 65 | P |
| 23 | 80 | P |
| 19 | 70 | P |
| 25 | 50 | F |
| 29 | 49 | ? |

You own a fruit shop in the Главный Рынок, and you want to organize your display so that similar fruits are placed close together so that you can serve your customers better.


Can you group the fruits into natural clusters to make it easier for customers to choose and improve your sales?

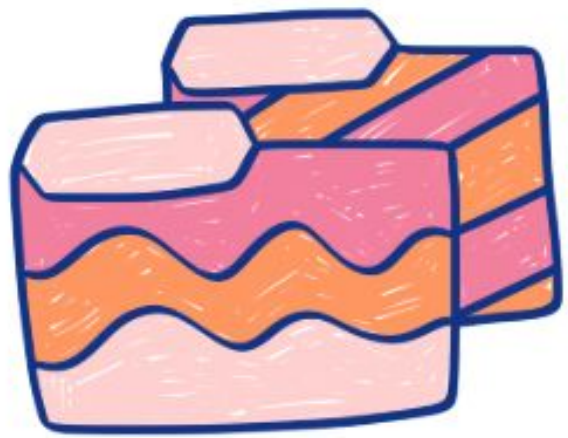

| Fruit | Color | Sweetness |
|---|---|---|
| Apple | Red | 7 |
| Pear | Green | 7 |
| Banana | Yellow | 9 |
| Lemon | Yellow | 1 |
| Watermelon | Green | 9 |
| Avocado | Green | 3 |

# ML Algorithm

**An algorithm** is a step-by-step set of instructions that tells a computer how to solve a problem or complete a task.

**A machine learning algorithm** is a special kind of algorithm that learns patterns from data and improves its decisions or predictions automatically without being explicitly programmed.

# What you'll learn



Data preparation

- Clean, explore, understand data
- Visualize data

Work with ML algorithms

- Theory
- Practice (scikit-learn)

Evaluation

- Test, compare
- Improve models

# How to pass

2 pairs for lecture

2 pairs for practice

Read more here: https://github.com/luumsk/NSU_ML

# Get to know Machine Learning

What is Machine Learning?

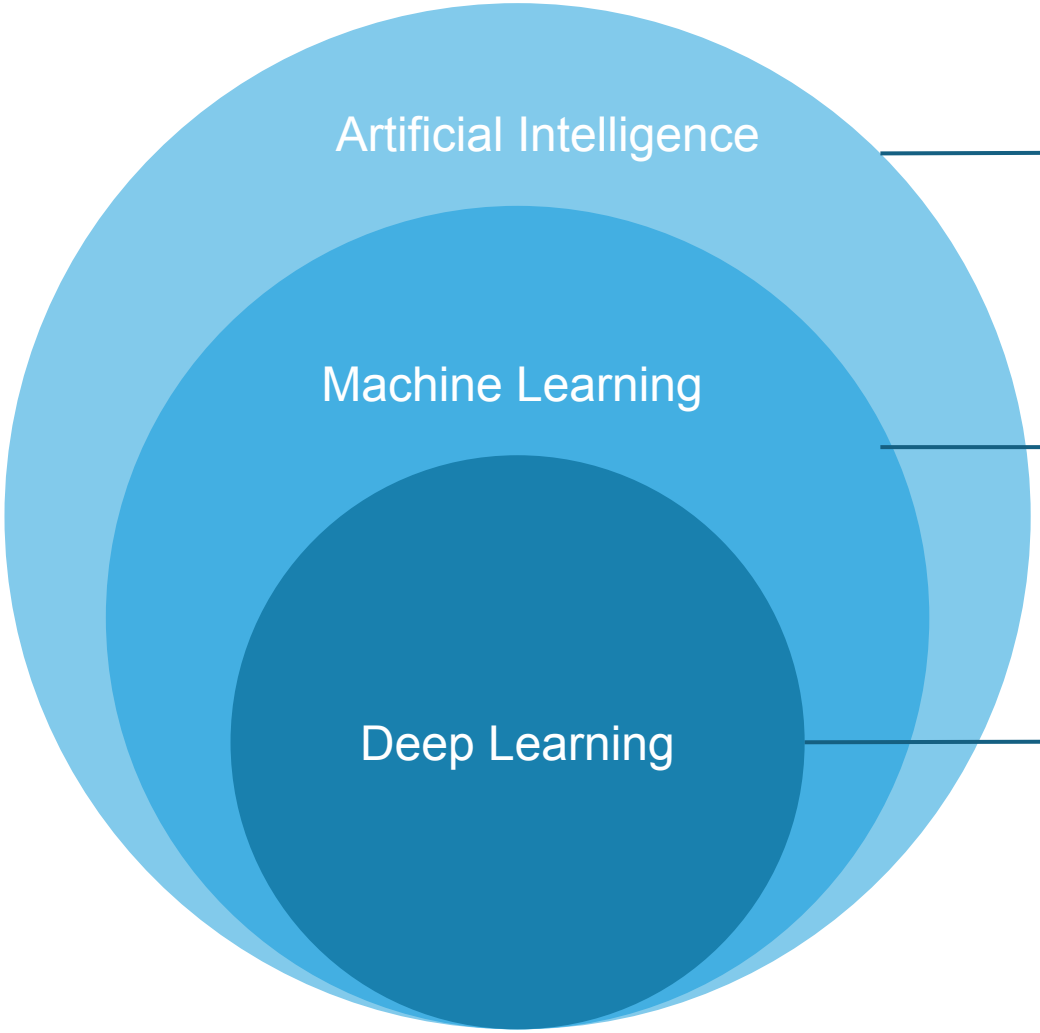
Applications of Machine Learning

Types of Machine Learning

Machine Learning pipeline

Component of a Machine Learning pipeline

# What is machine learning?

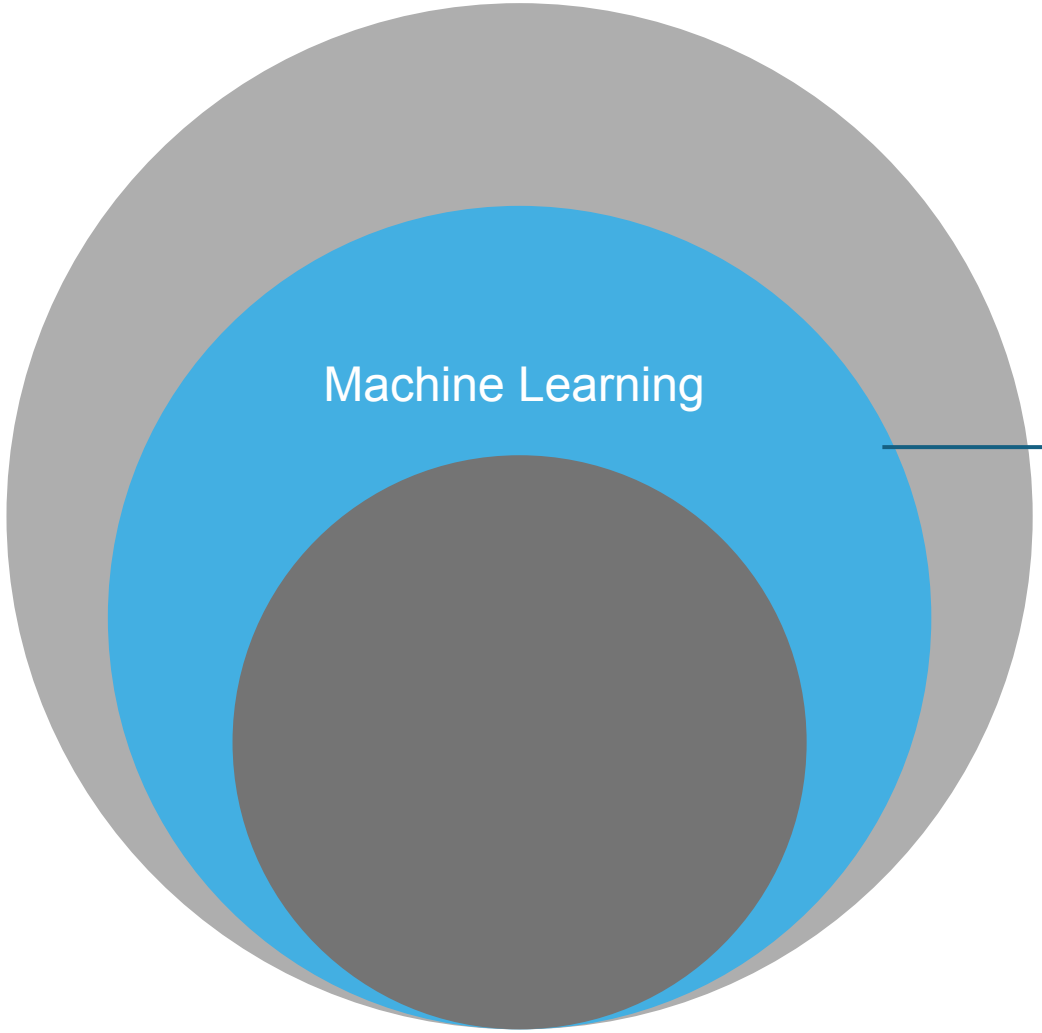



The broad field of creating systems that **mimic human intelligence** to perform tasks autonomously

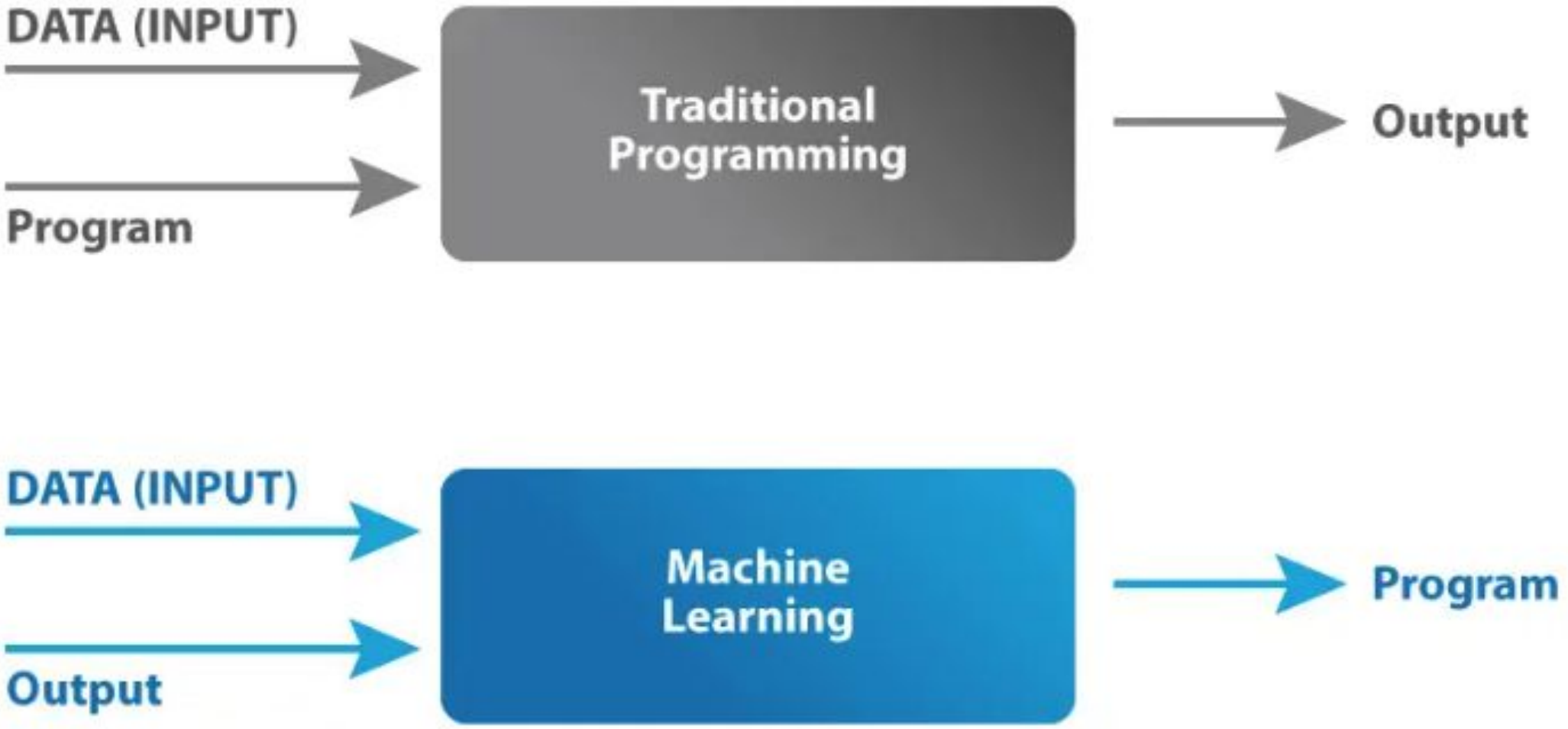
A subset of AI that enables computers to **learn patterns from data** and improve over **time without explicit programming** for every task

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




Source: https://www.mygreatlearning.com/blog/what-is-machine-learning/

# 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

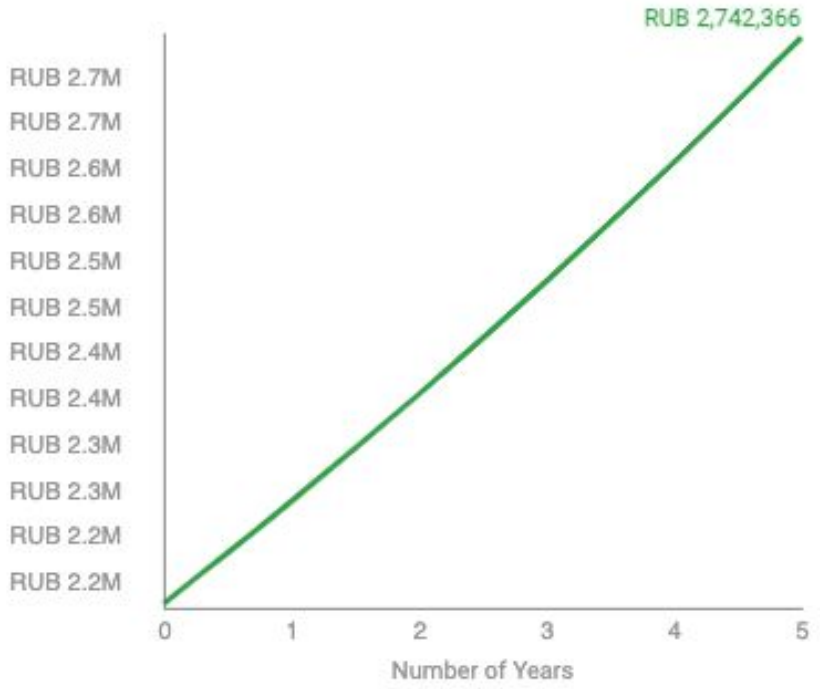
**RUB 97,593**
Average Bonus

Estimated salary in 2029:

RUB 2,742,366

5 Year Change:

▲ 29 %

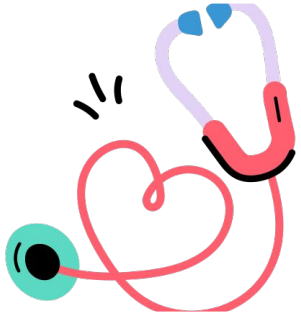



Job titles

- 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

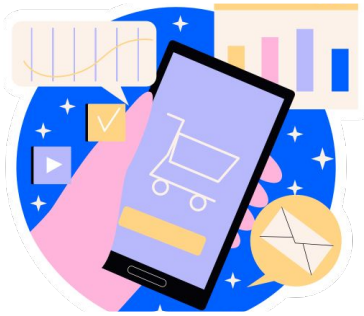
Source: https://www.erieri.com/salary/job/machine-learning-engineer/russian-federation

# Applications of machine learning

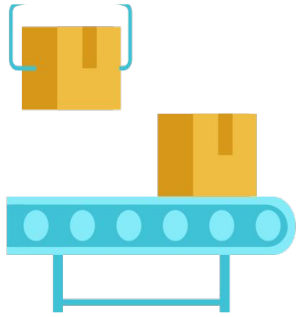


Healthcare

Finance



E-commerce

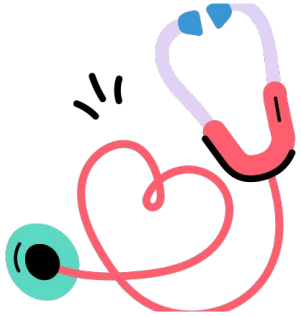


Manufacturing

Transport and Logistic

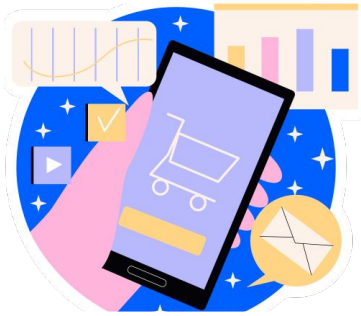
# Applications of machine learning

- Diagnostics
- Predictive Analytics
- Personalized Medicine



Healthcare

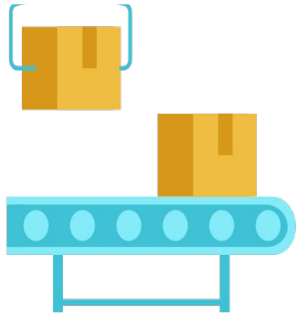
Finance



E-commerce

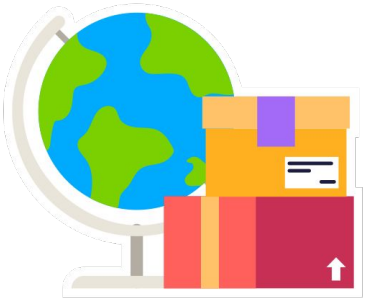
- Recommendation Engines
- Customer Segmentation
- Inventory Management

- Fraud Detection
- Credit Scoring
- Algorithmic Trading



Manufacturing

- Predictive Maintenance
- Quality Control
- Supply Chain Optimization



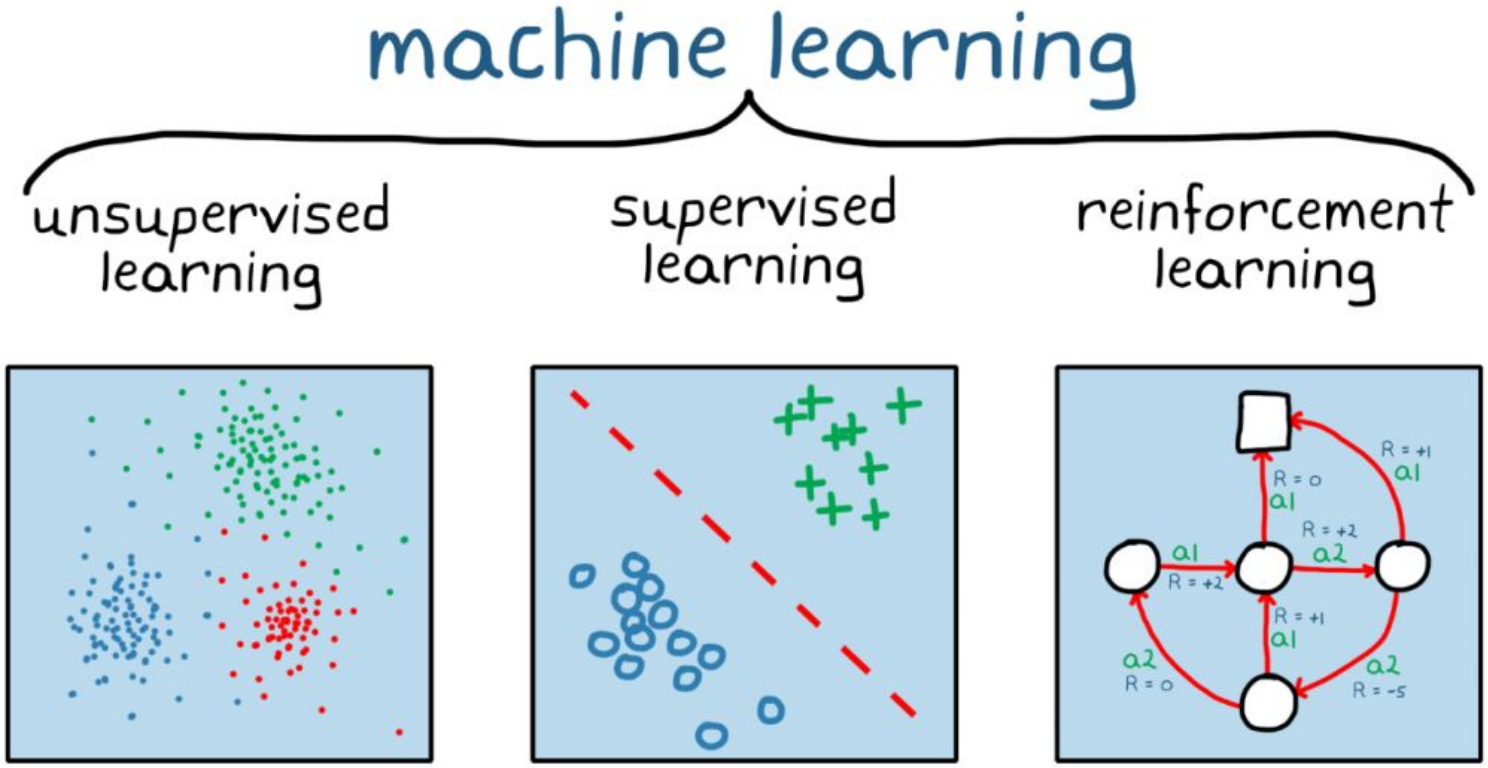
Transport and Logistic

- Route Optimization
- Traffic Prediction

# What ML applications you are interested in?

Each student suggest an idea to apply ML to solve a problem. Just raw ideas, do not think too much about feasibility.

# Types of machine learning



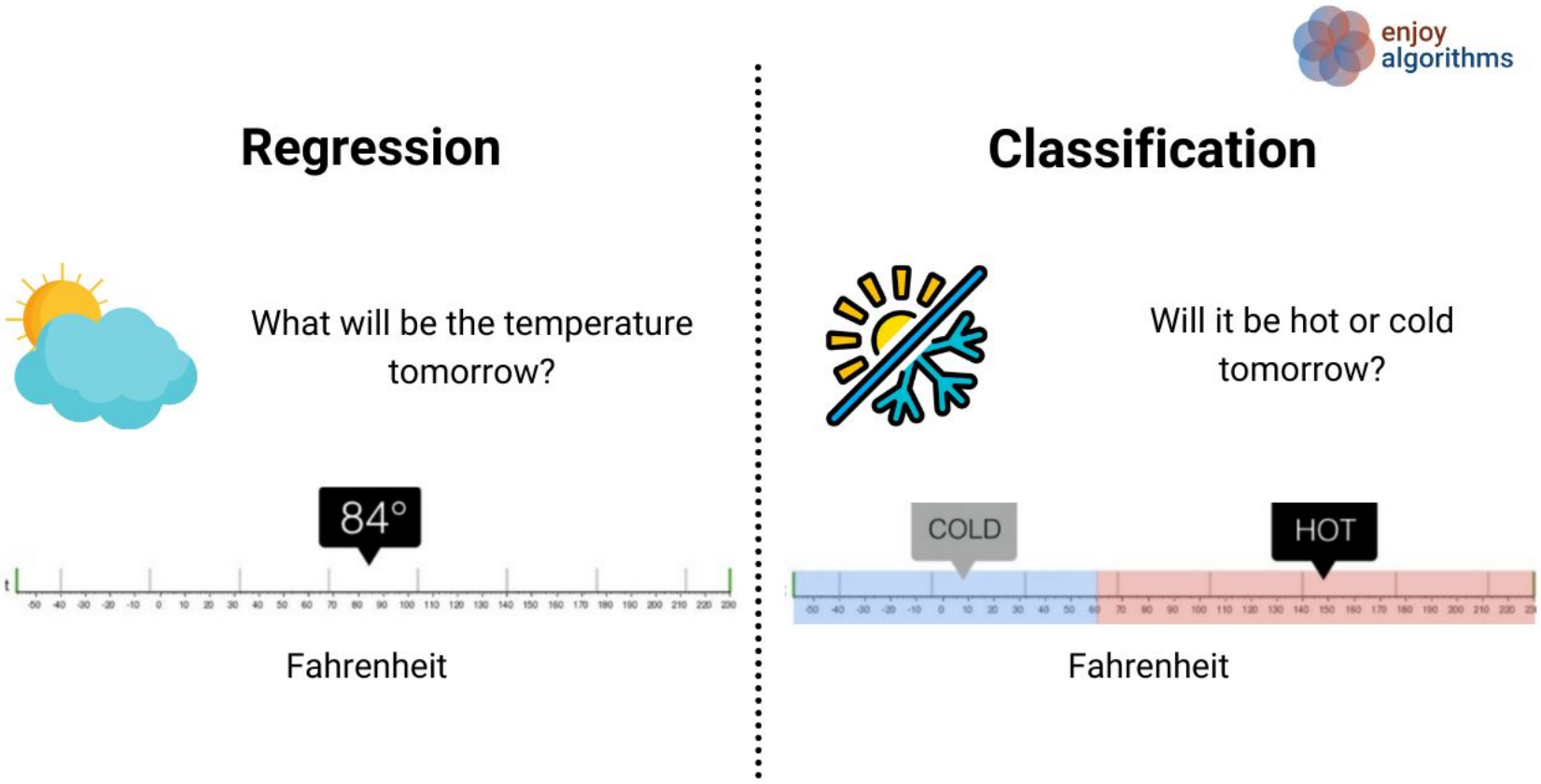

Unlabeled data
No feedback
Find hidden structure

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/

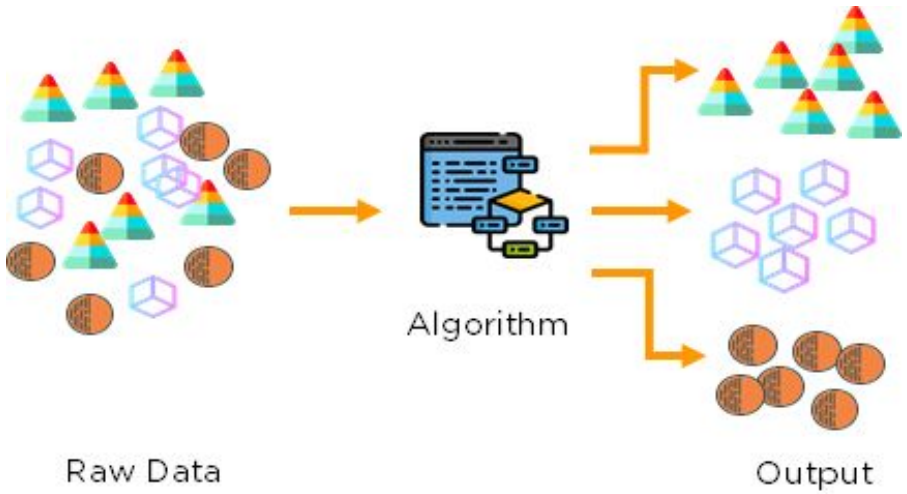
# Types of Supervised Learning



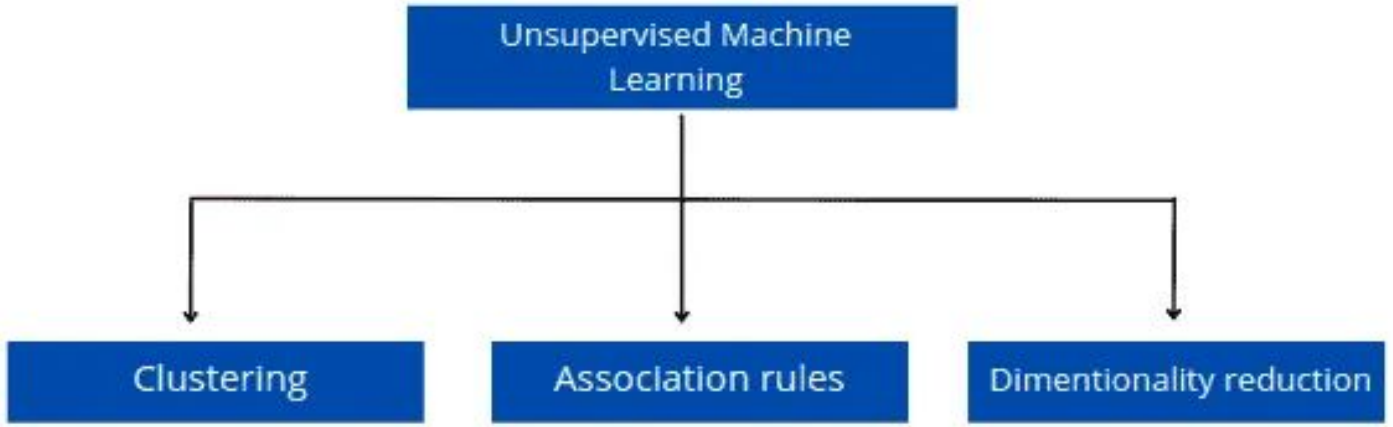

Binary/
Multi-class

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?

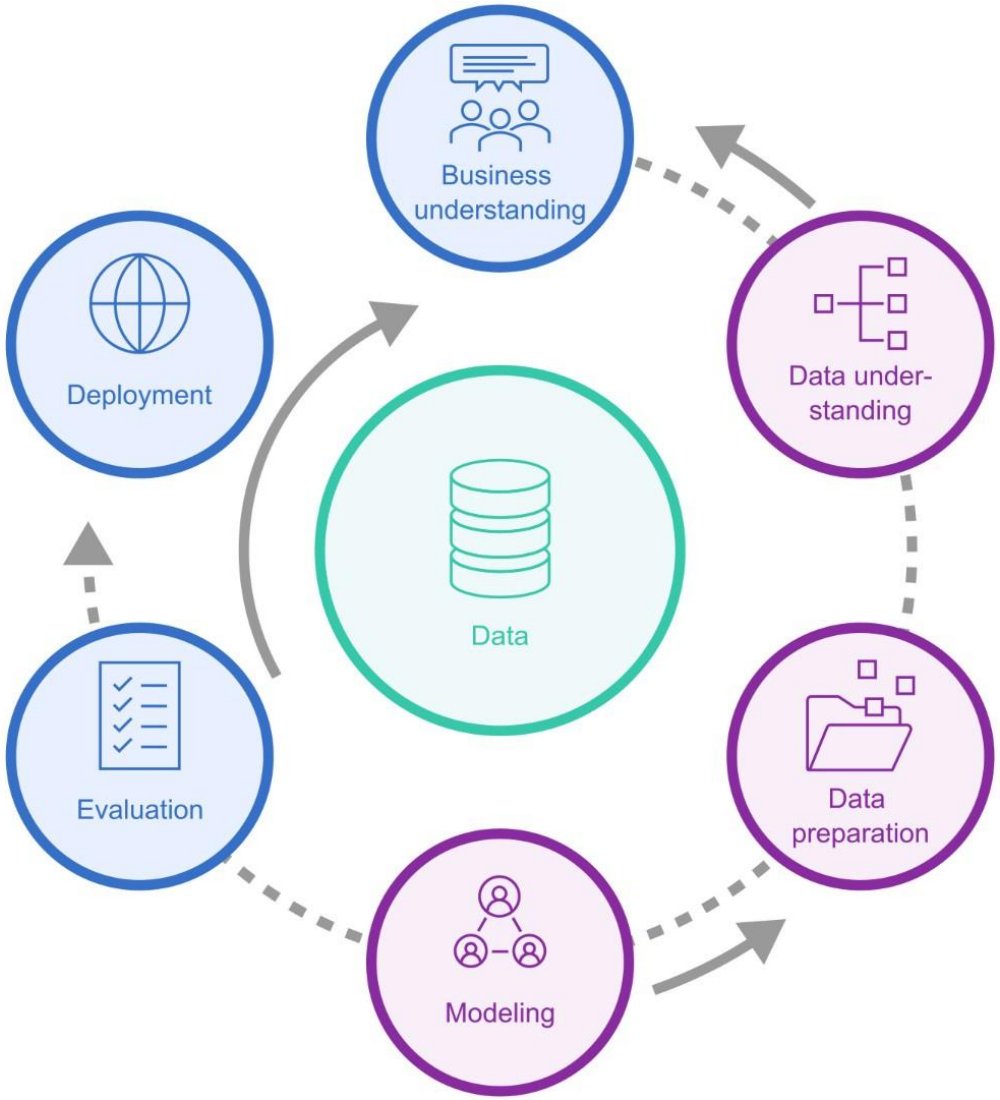
# What type of ML problem do you think it is?

**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.

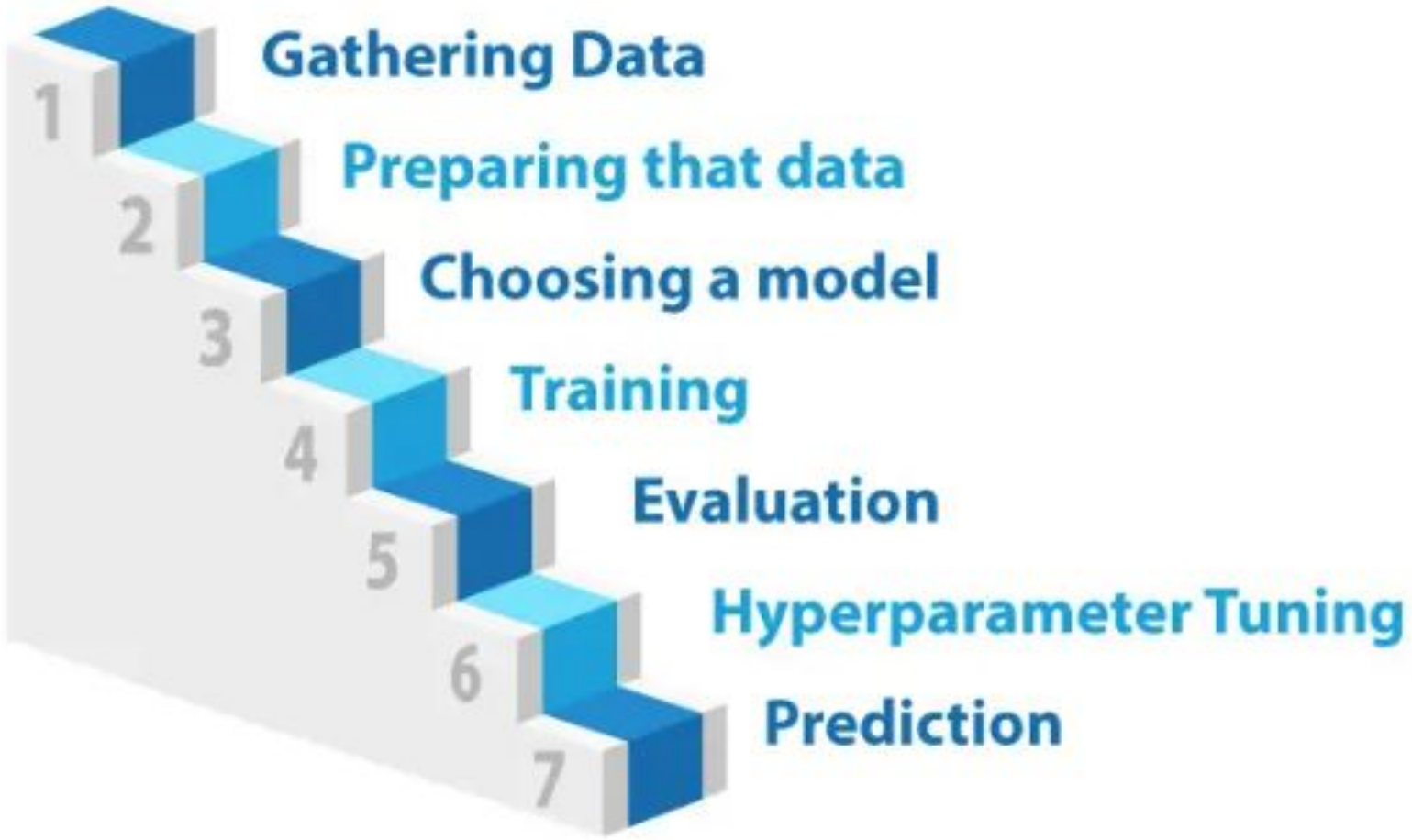
**Question:** What type of machine learning problem is this, and why?

# Machine learning pipline

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

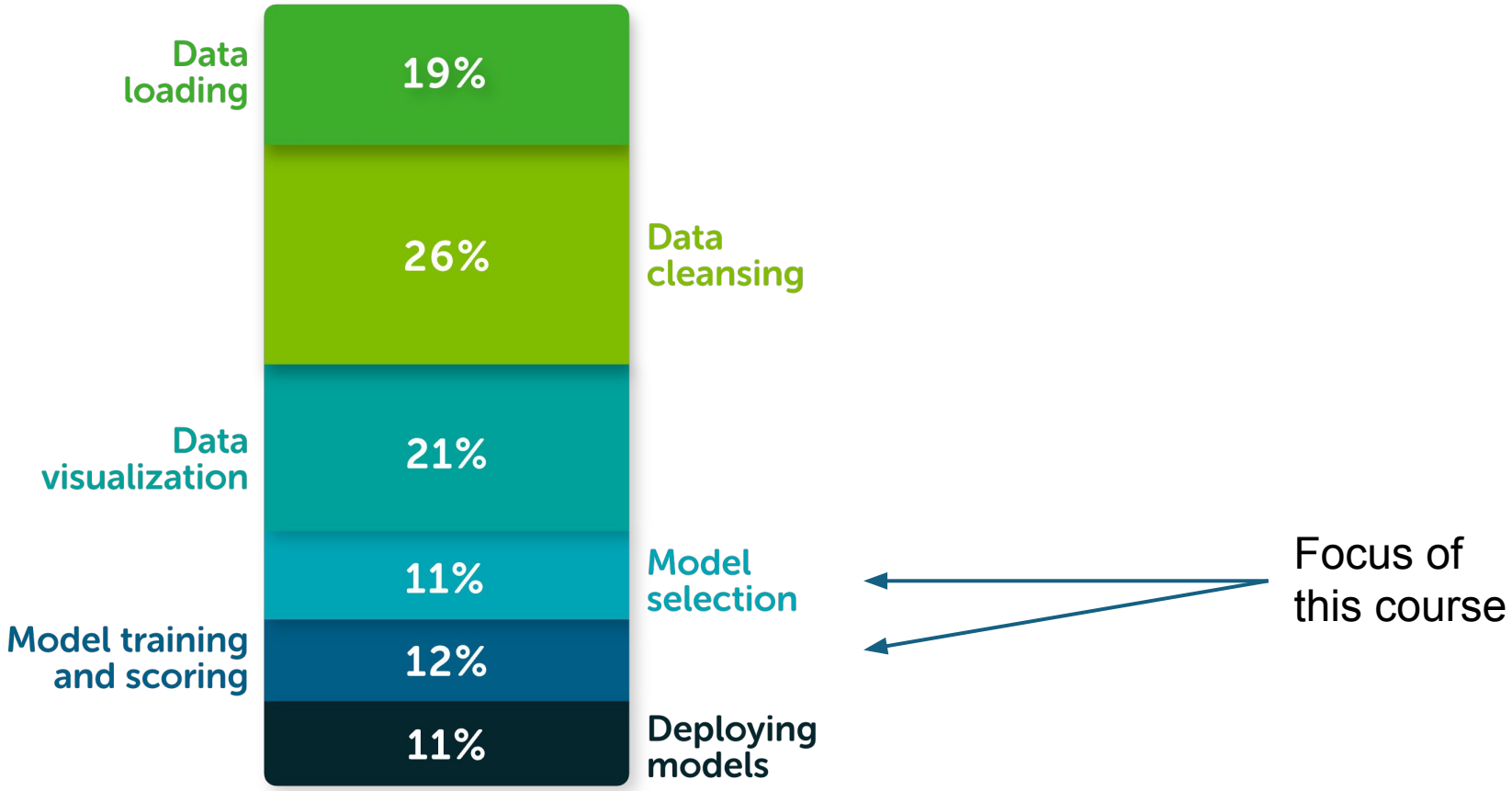


# 7 steps of Machine Learning




Source: https://www.mygreatlearning.com/blog/what-is-machine-learning/

# Time allocation for ML tasks




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

# Some practical aspects

Underfitting vs Overfitting

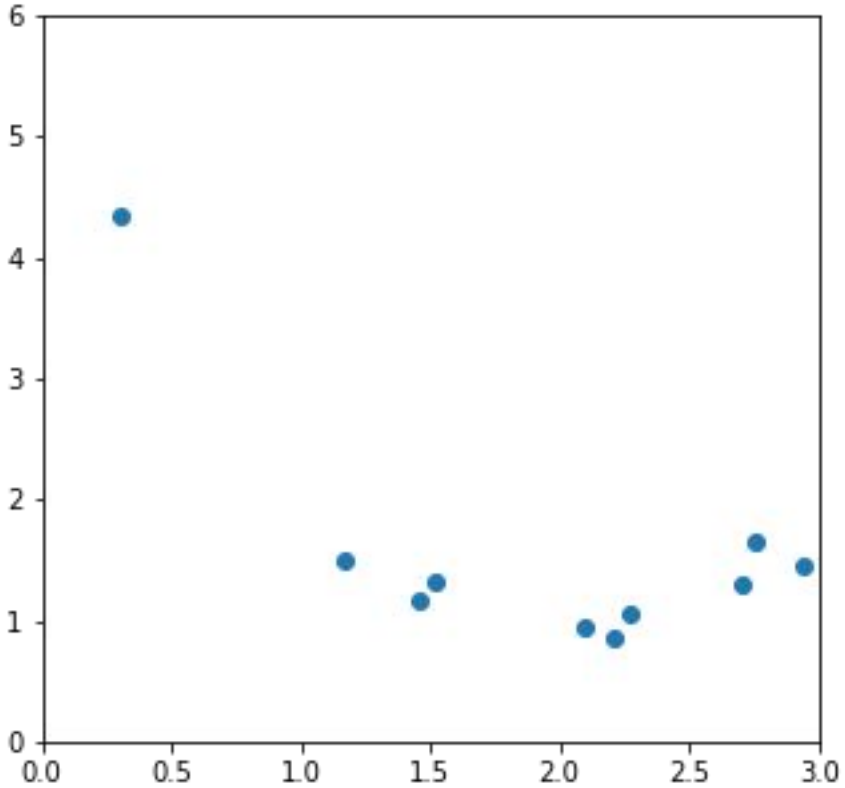
Training, validation, and test dataset splitting

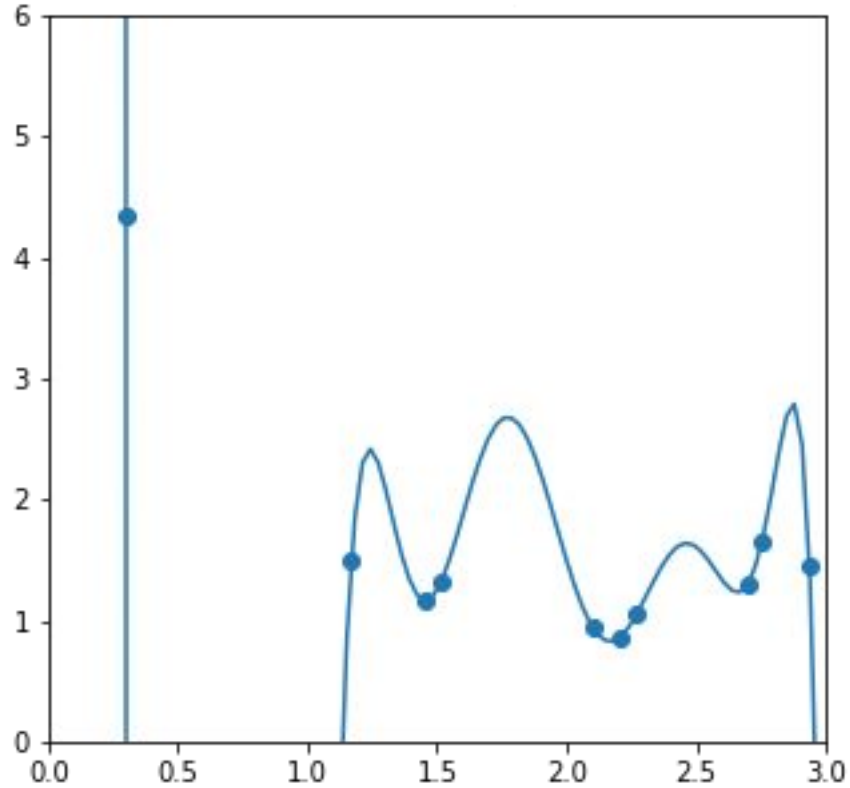
Cross-validation

Hyperparameters tuning

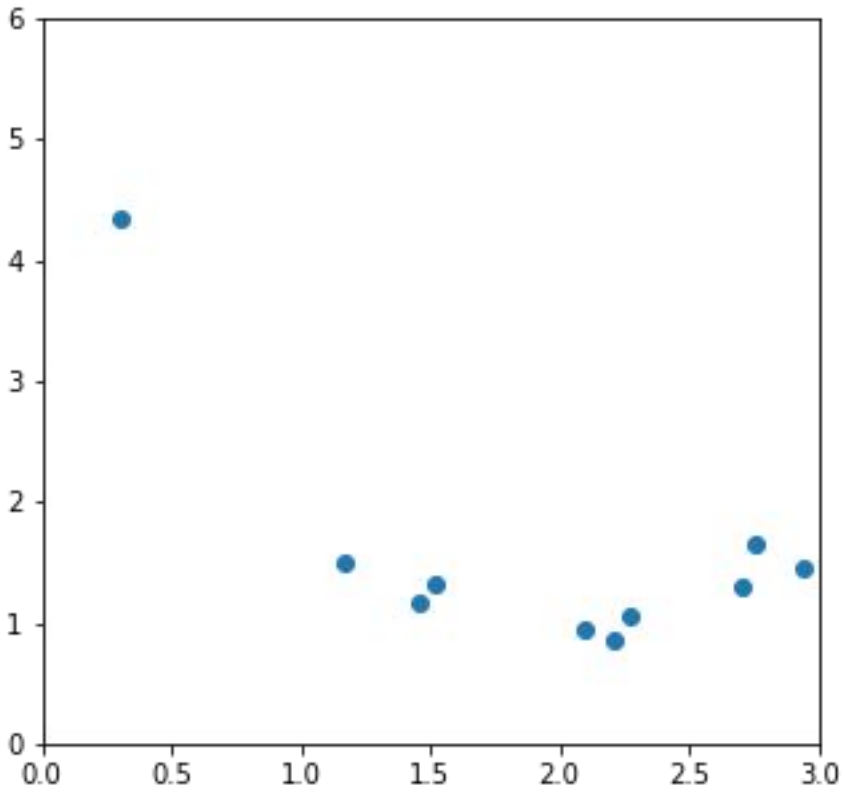
Ensemble method

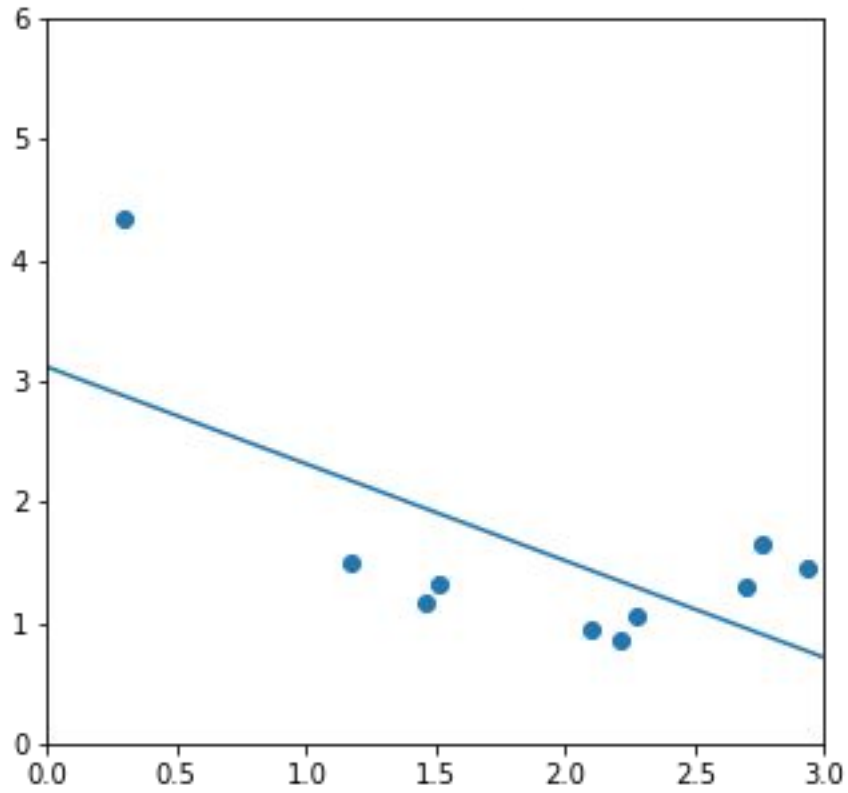
# Choosing the function class



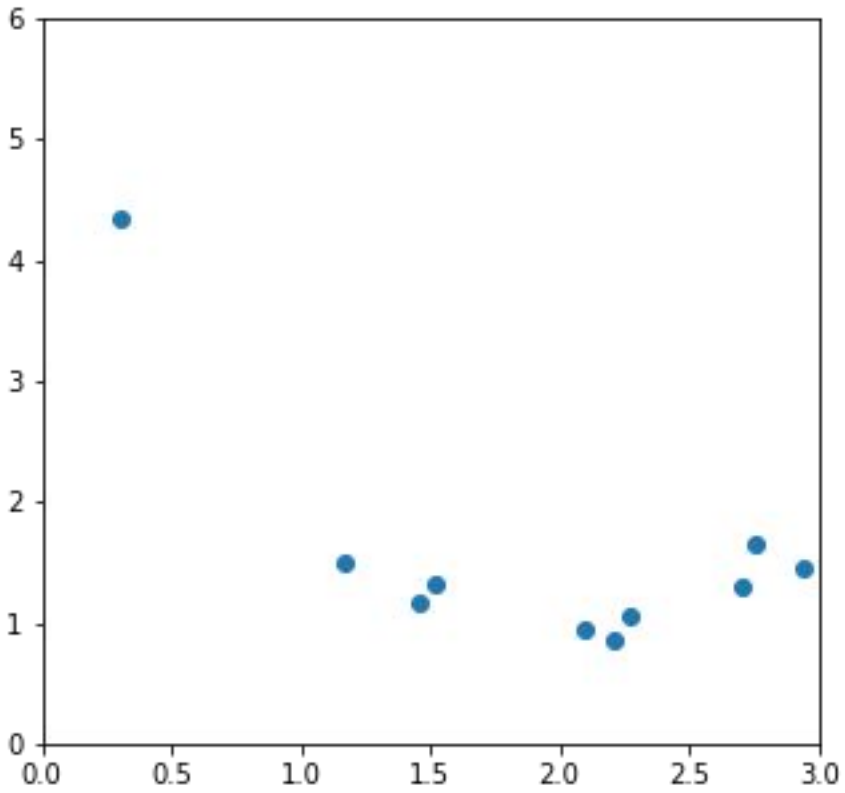



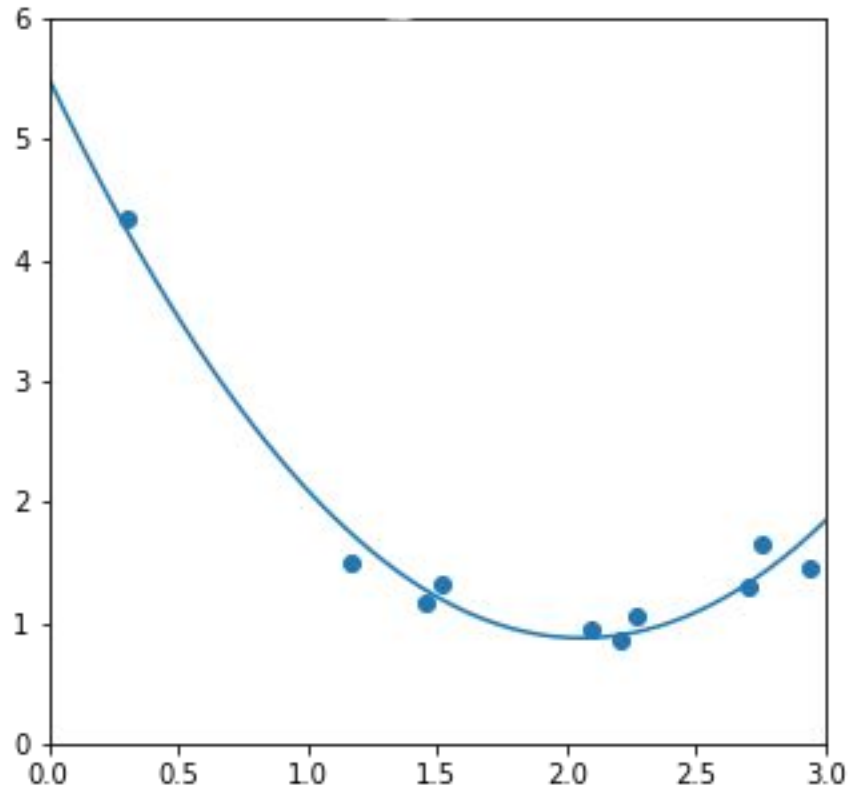
# Choosing the function class



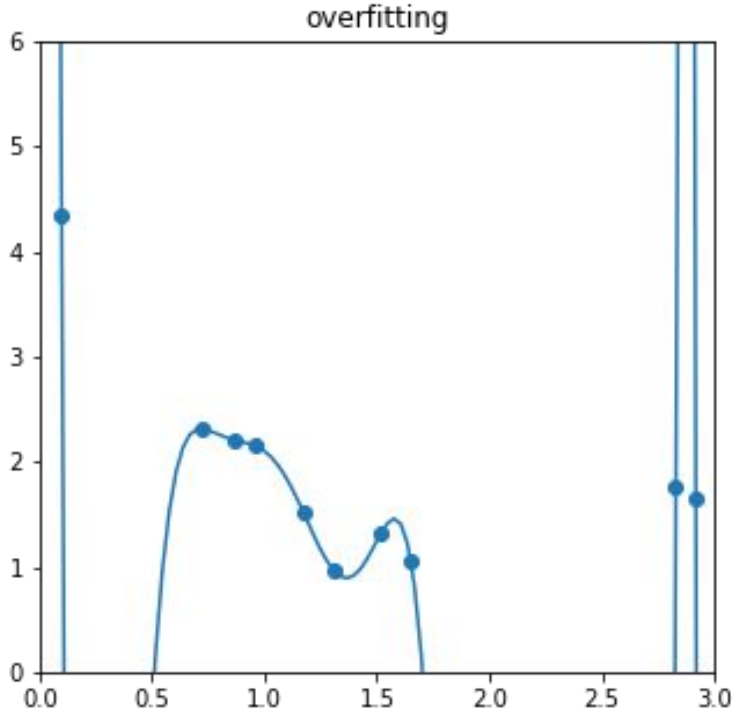
# Choosing the function class

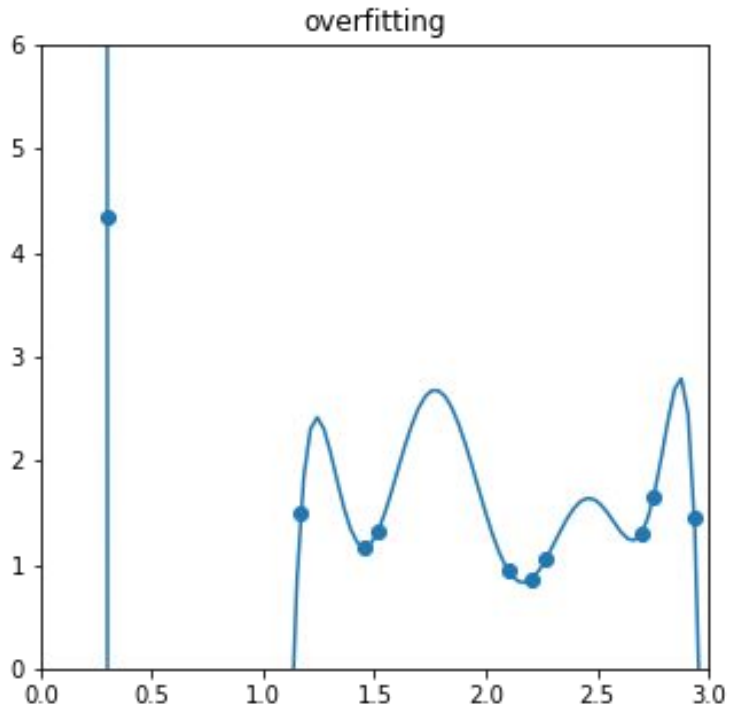


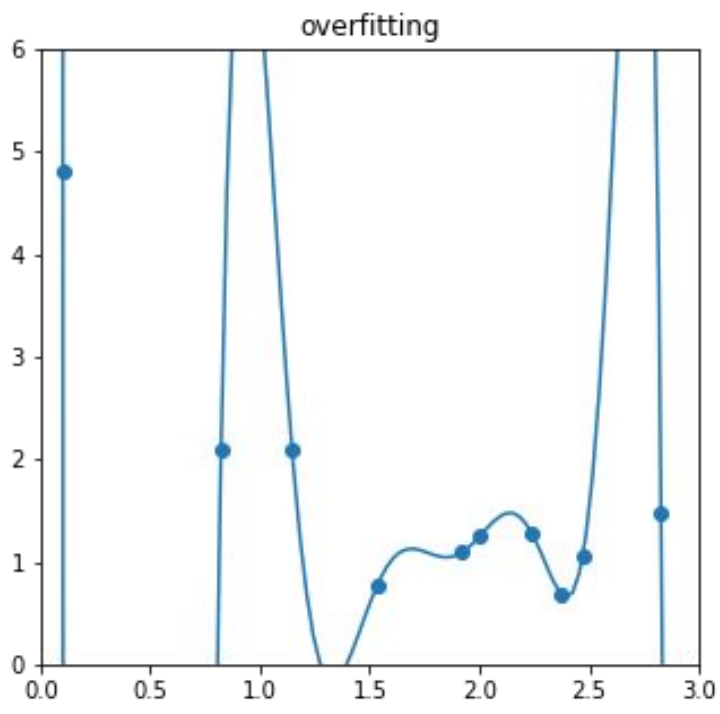



# High Variance

Variance is how much the model's predictions change if you use a different sample of training data

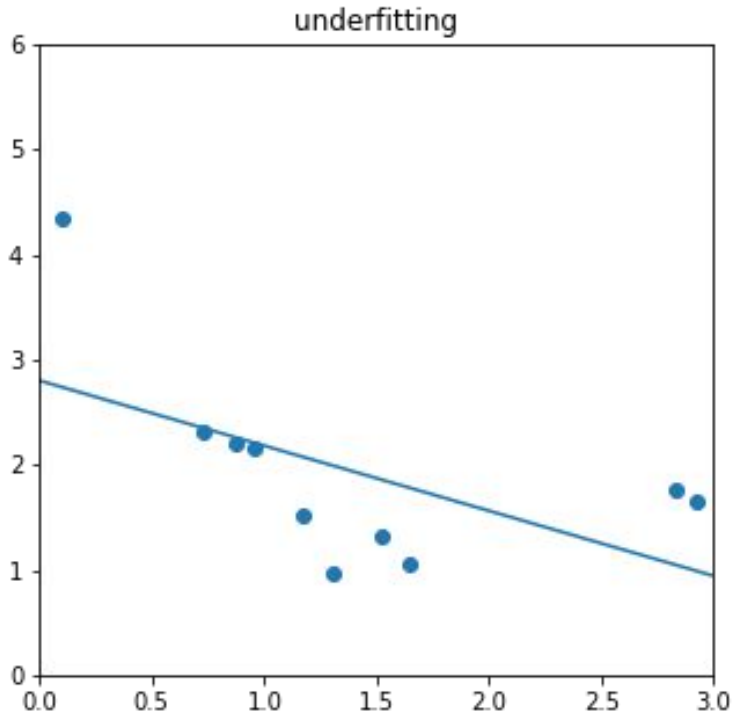


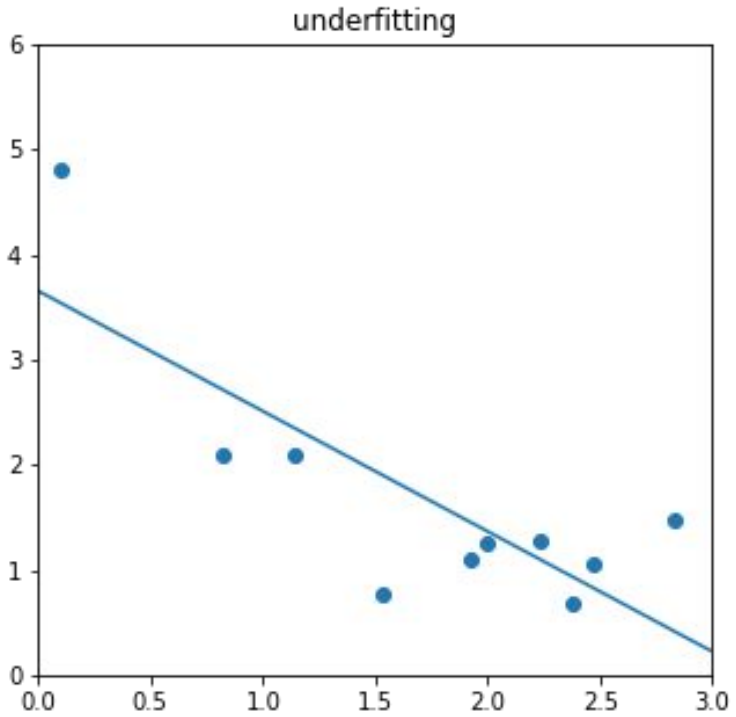



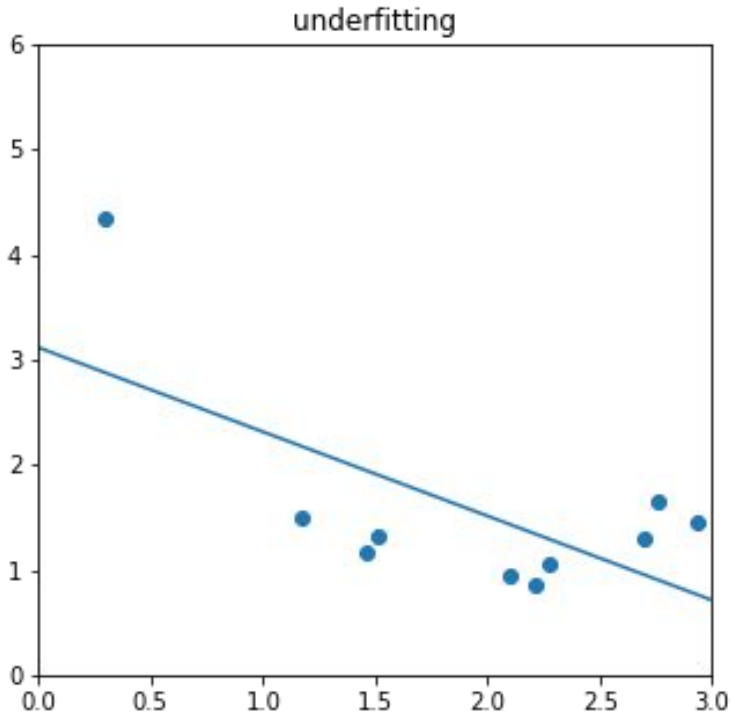



# High Bias

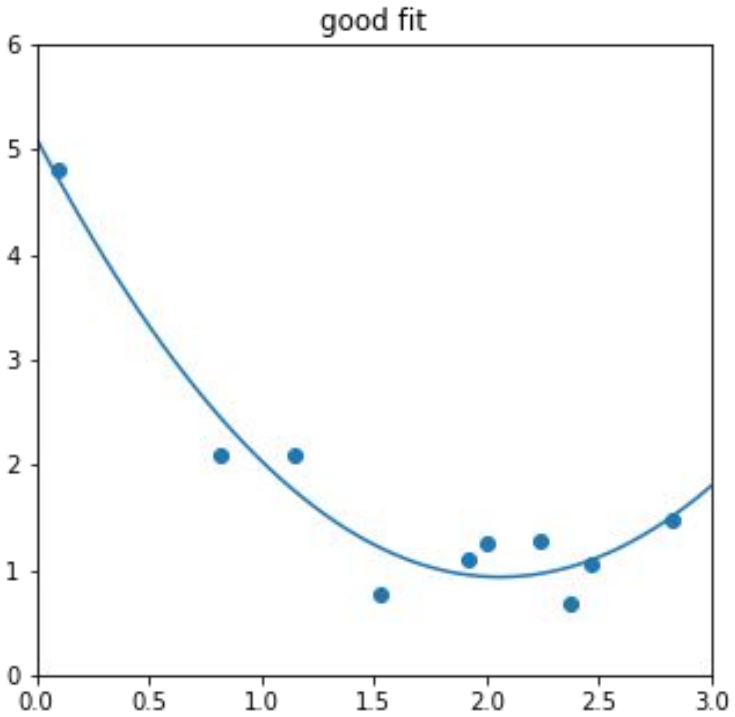
Bias is the difference between the average prediction of your model and the true value.

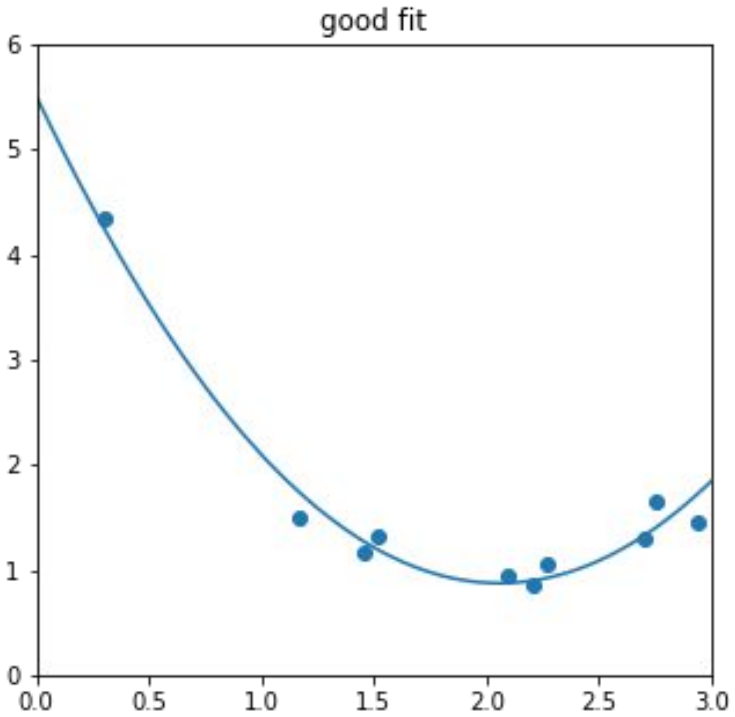


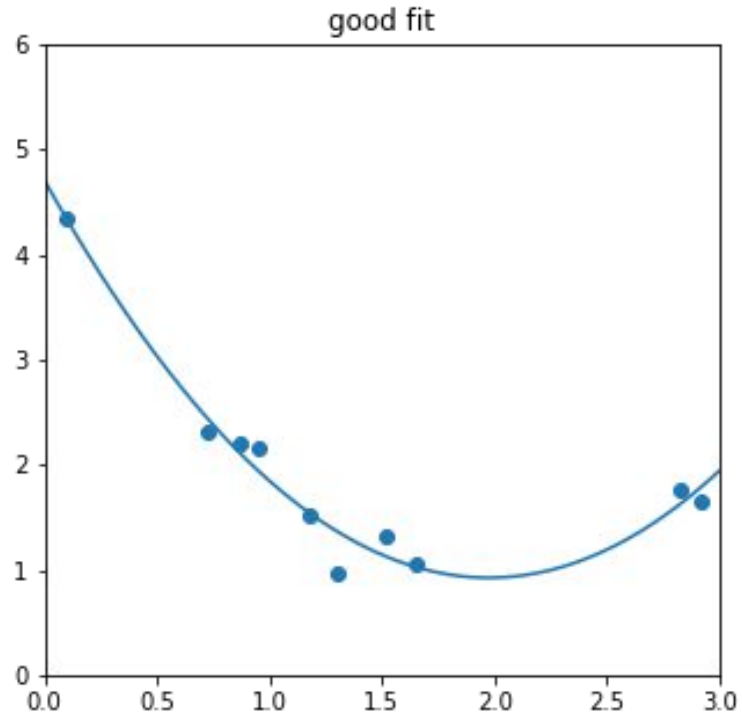



# Optimal Bias and Variance



# 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.

# Exercise

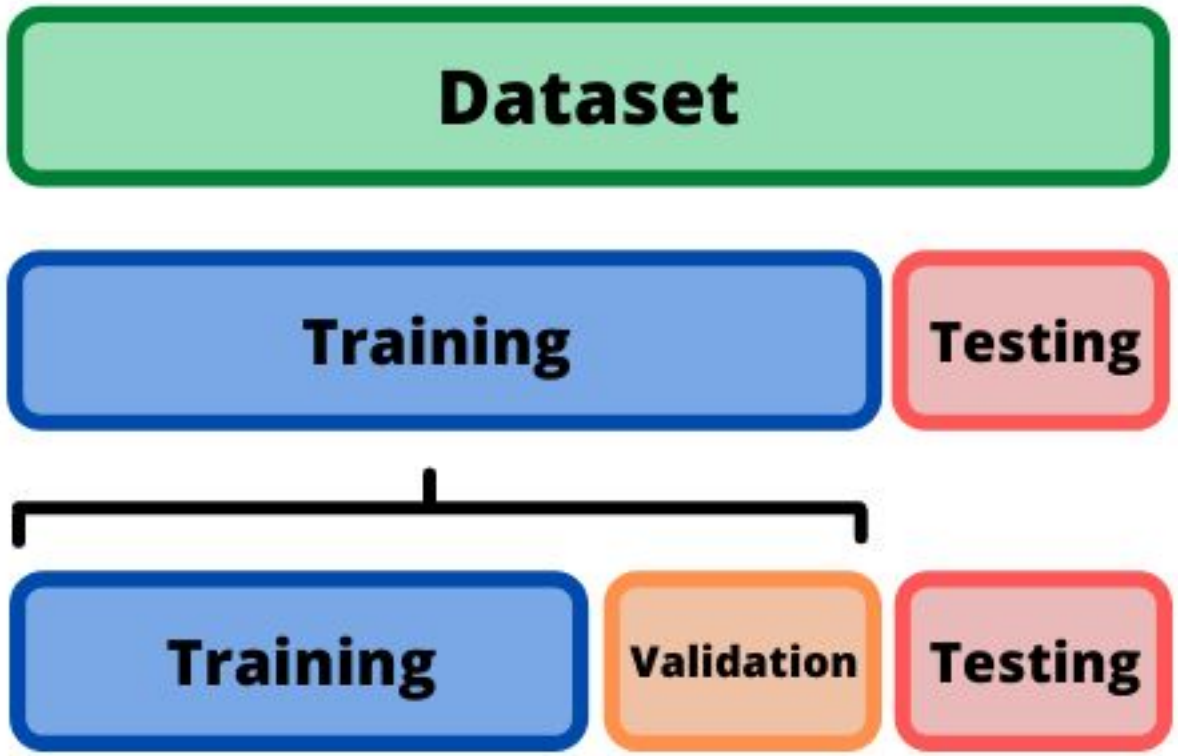
**Scenario 4:** You build a model that predicts apartment prices using only size ($m^2$). It gives almost the same price for all apartments of the same size — even when some are near the metro and others far away.

**Question:** Is this high bias or high variance?

**Scenario 5:** Another model uses every possible feature — number of study hours, attendance, coffee intake, last week's weather, and even shoe size. It fits the training data perfectly, but when tested on a new class, predictions are wildly off.

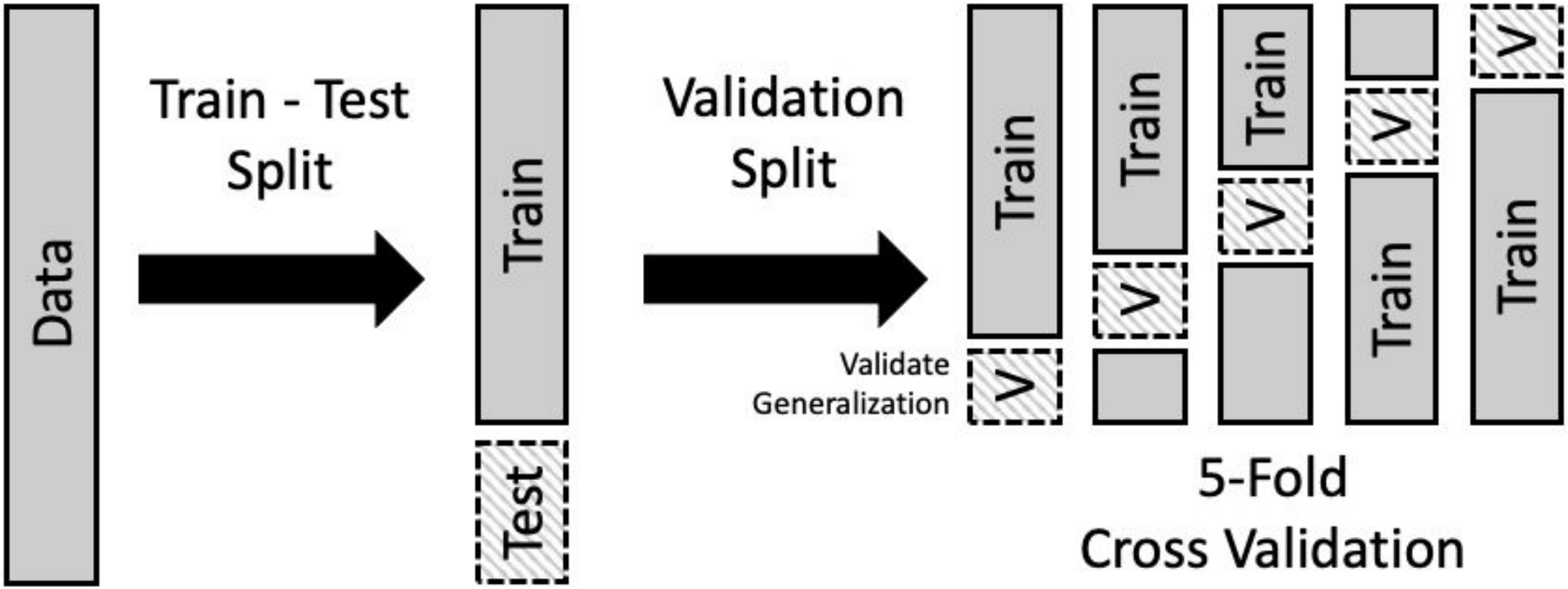
**Question:** Is this high bias or high variance?

# Dataset splitting



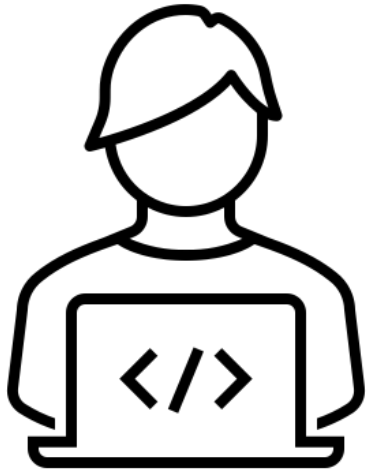

Source: https://dhavalpatel2101992.wordpress.com/2021/05/21/kaggle-titanic-dataset-cleaning-split-data-into-train-validation-and-test-set/

# Cross Validation




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

# 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

# Practice

Quiz (required):
https://forms.gle/g7y2R4vqbvwnW51W7

Lab 01 (required):
https://github.com/luumsk/NSU_ML/blob/main/Labs/lab1.ipynb

Pandas Exercises (optional):
https://github.com/datasciencelab-ai/100_bai_tap_pandas

# 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