Machine Learning

Topic 3. Lecture 3

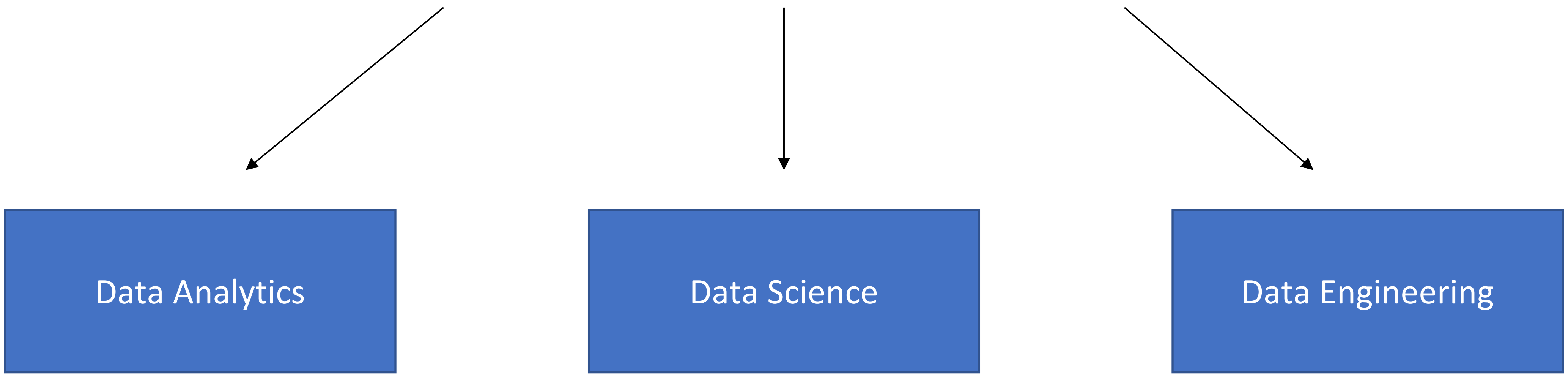
Classic Machine Learning. Supervised Learning. Setting the Machine Learning Task

Yury Sanochkin

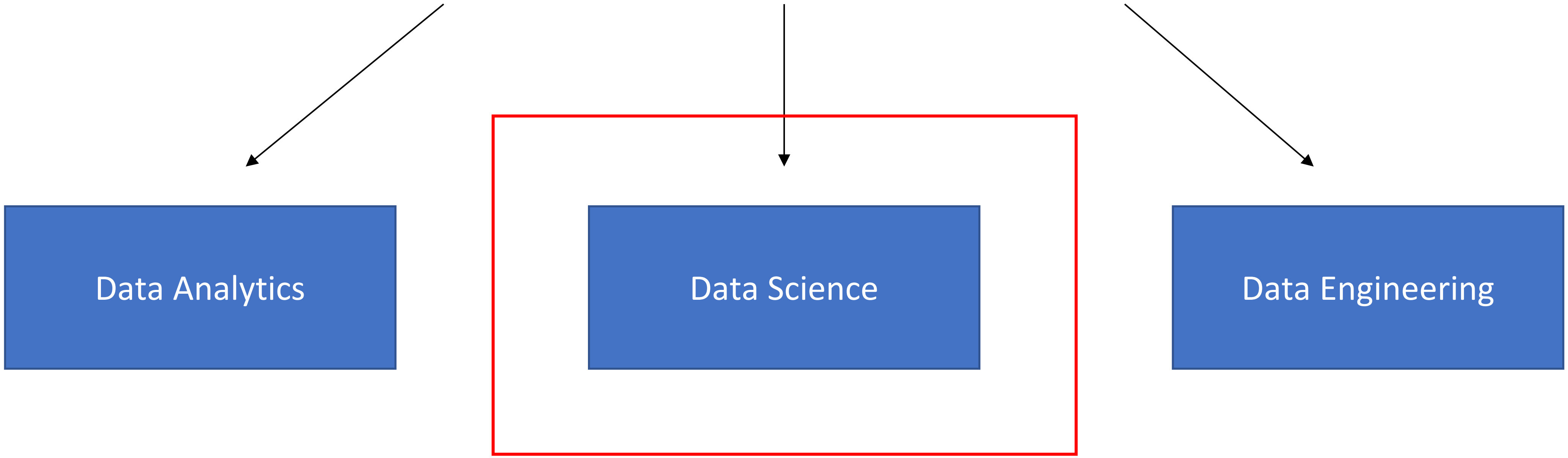
ysanochkin@hse.ru

NRU HSE, 2025

What are the types of data analysis tasks?



What are the types of data analysis tasks?



Finally, we can completely go to this section

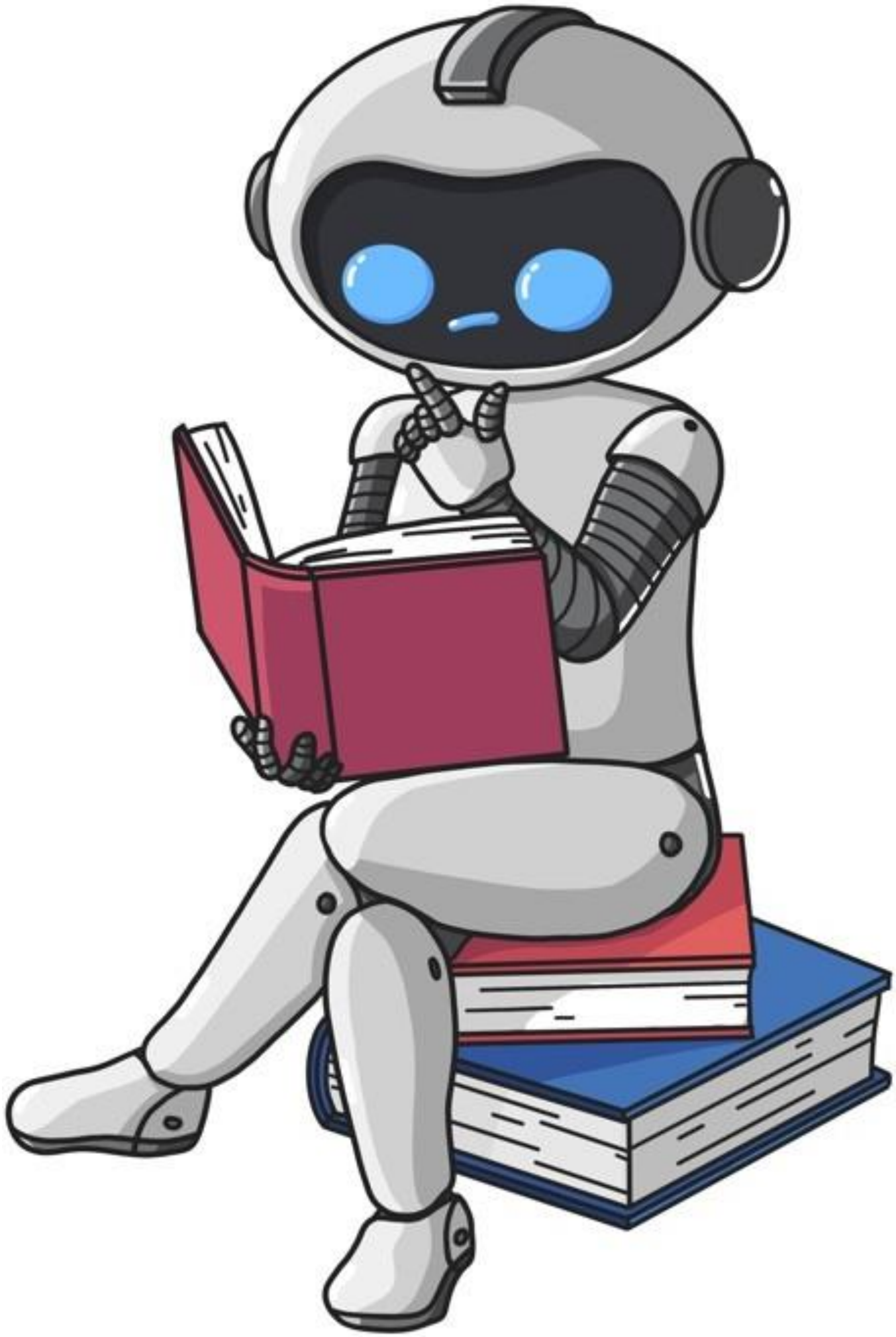
Machine Learning

Machine Learning

- How would you define what machine learning is in general? How do you understand it?

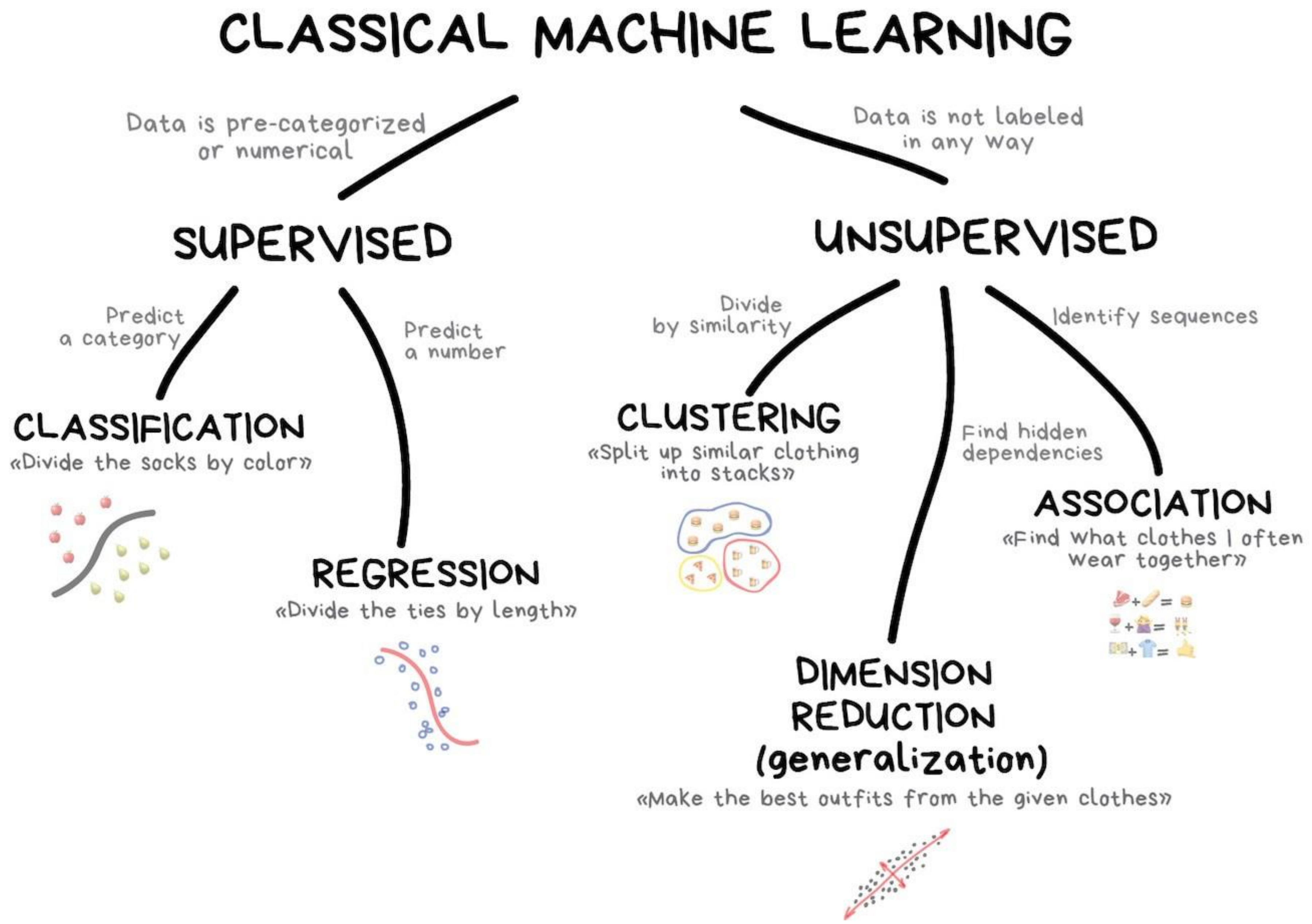
Machine Learning

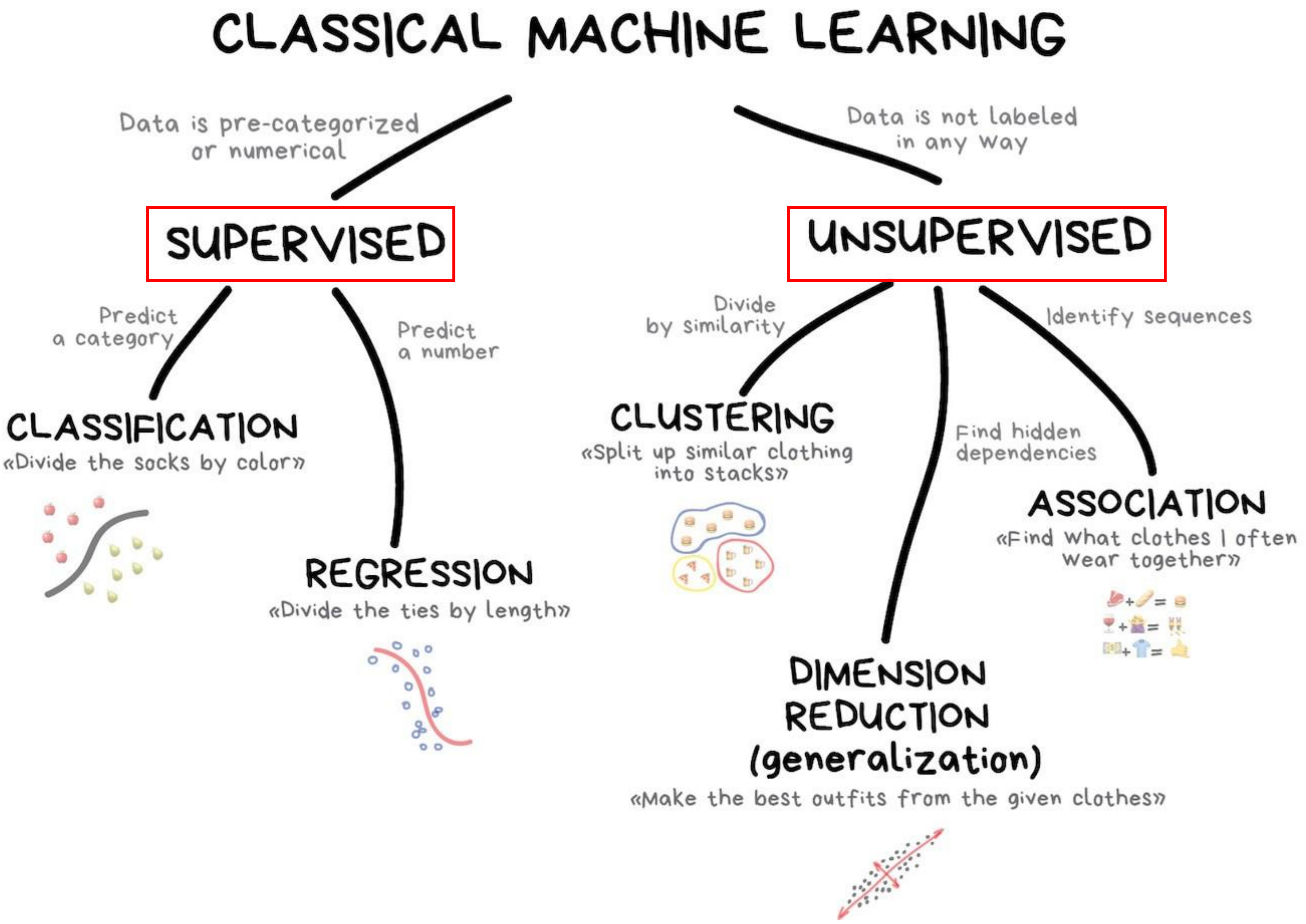
The science
of finding patterns in data
using a computer and
mathematics.



Machine Learning

- What two categories can we divide classical machine learning tasks into?


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

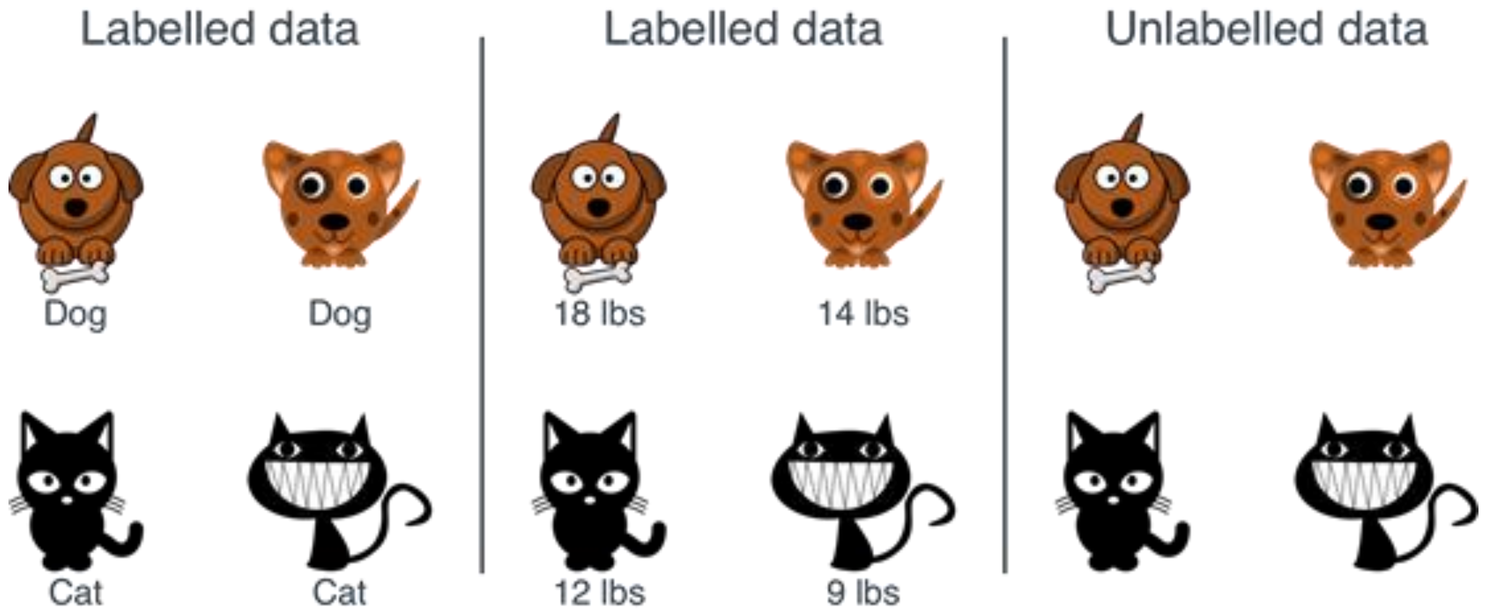
Machine Learning

- These blocks of machine learning tasks are inextricably linked to the concept of labeled/unlabeled data.

Machine Learning

- These blocks of machine learning tasks are inextricably linked to the concept of labeled/unlabeled data.
- What are labeled/unlabeled data?
- Provide examples of labeled/unlabeled data.

Labeled vs Unlabeled data



Labeled vs Unlabeled data

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled

Machine Learning

- As part of the topic "Setting the Machine Learning Task", we will discuss the mechanics of the learning process using the example of three main tasks of classical ML: regression, classification, clustering.

Machine Learning

- As part of the topic "Setting the Machine Learning Task", we will discuss the mechanics of the learning process using the example of three main tasks of classical ML: regression, classification, clustering.
- ...and we'll also explore metric algorithms, take a closer look at our first machine learning model – KNN – and along the way discuss many other accompanying details!
- It's going to be interesting, let's go!

Machine Learning

- First let's have a quick run to have the wind up

Machine Learning

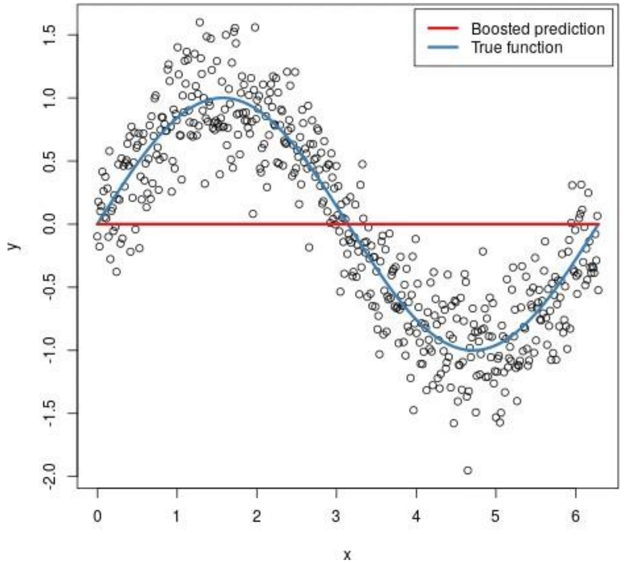
- First let's have a quick run to have the wind up

...but later on, to understand
and realize that it is not scary
at all. :)

Machine Learning

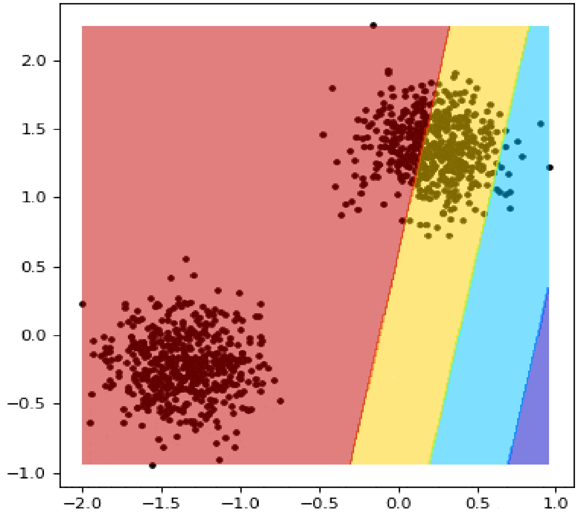
- First let's have a quick run to have the wind up

Regression



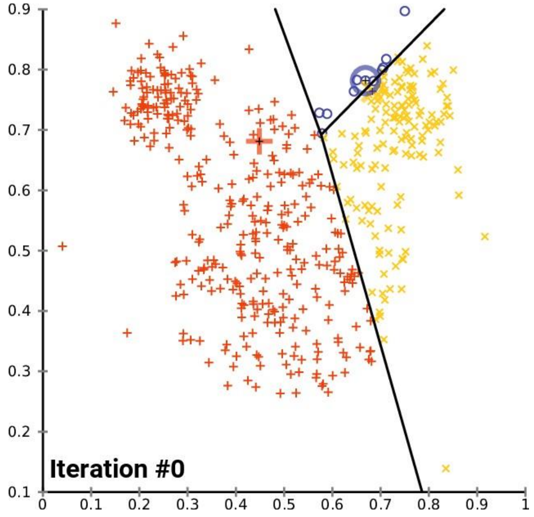
$Y \subseteq$ R. It is necessary to restore the usual functional dependence $f{:}X \rightarrow Y$.

Classification



$Y \subseteq [0,1]^n$ .It is necessary to predict the probability distribution over possible outcomes.

Clustering

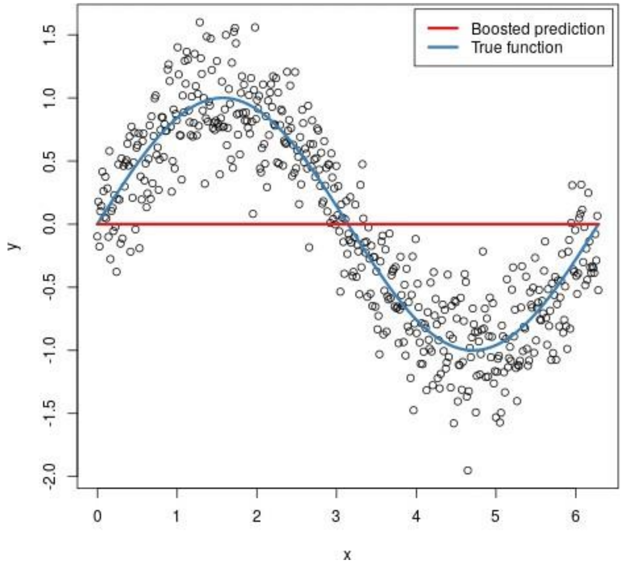


It is necessary to define such equivalence classes that objects of the same class are more similar to each other than to objects of different classes.

Machine Learning

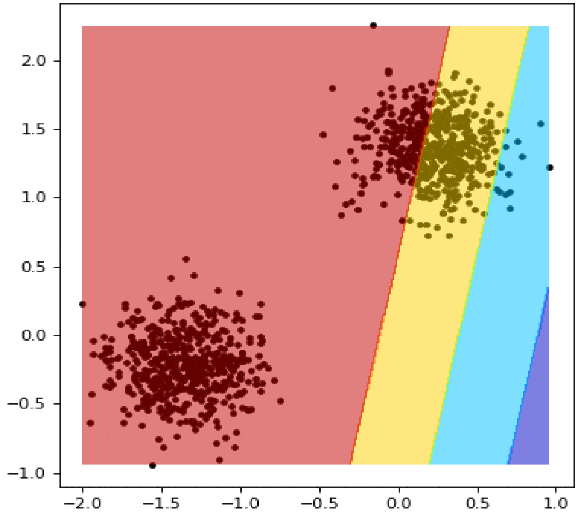
Supervised learning

Regression



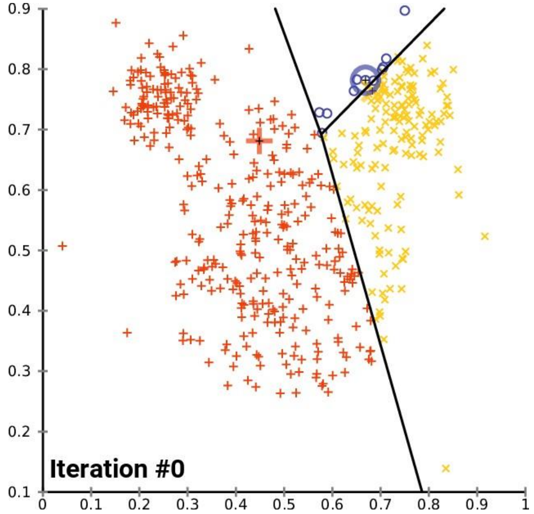
$Y \subseteq$ R. It is necessary to restore the usual functional dependence $f{:}X \rightarrow Y$.

Classification



$Y \subseteq [0,1]^n$ .It is necessary to predict the probability distribution over possible outcomes.

Clustering

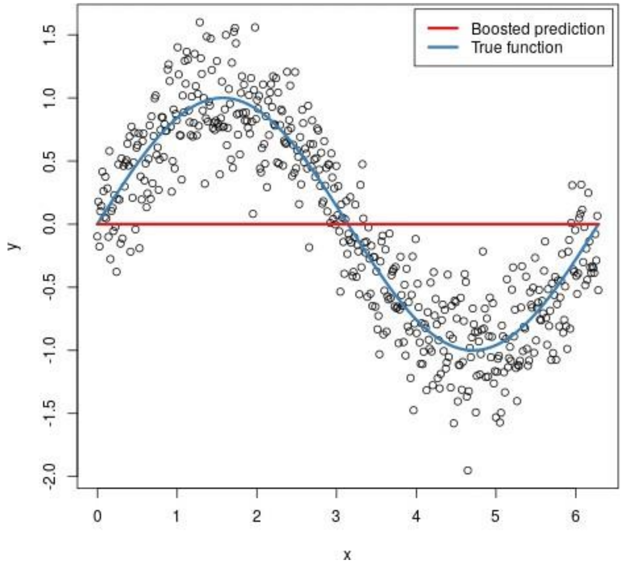


It is necessary to define such equivalence classes that objects of the same class are more similar to each other than to objects of different classes.

Machine Learning

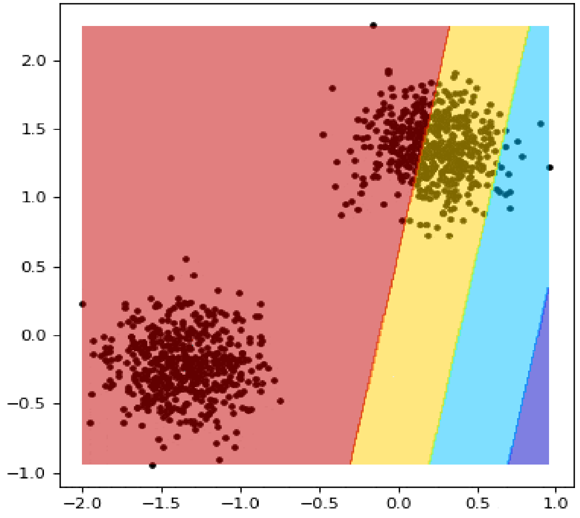
Unsupervised learning

Regression



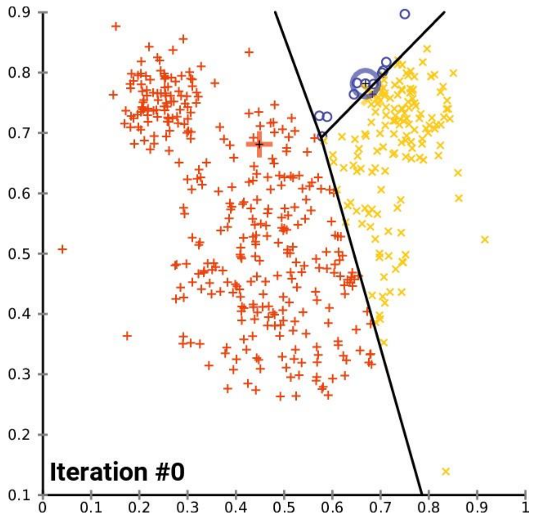
$Y \subseteq$ R. It is necessary to restore the usual functional dependence $f{:}X \rightarrow Y$.

Classification



$Y \subseteq [0,1]^n$ .It is necessary to predict the probability distribution over possible outcomes.

Clustering



It is necessary to define such equivalence classes that objects of the same class are more similar to each other than to objects of different classes.

Supervised learning

- So, let's recall the basic notations!

Supervised learning

- So, let's recall the basic notations!
- X - the set of all objects in the feature space
- Y - target variable range of values

Supervised learning

- So, let's recall the basic notations!
- X - the set of all objects in the feature space
- Y - target variable range of values

- What is machine learning in these terms?

Supervised learning

- So, let's recall the basic notations!
- X - the set of all objects in the feature space
- Y - target variable range of values

- What is machine learning in these terms?
- In fact, this is about finding an unknown dependency:
- $f: X \to Y$ - unknown pattern, function

Supervised learning

- So, let's recall the basic notations!
- X - the set of all objects in the feature space
- Y - target variable range of values

- What is machine learning in these terms?
- In fact, this is about finding an unknown dependency:
- $f: X \rightarrow Y$ - unknown pattern, function
- It may even be stochastic!

Supervised learning

- How do we do this?

Supervised learning

- How do we do this?

- Given: Training sample of the form $\{(X_i,\ y_i)\}_{i=1}^{n}$

Supervised learning

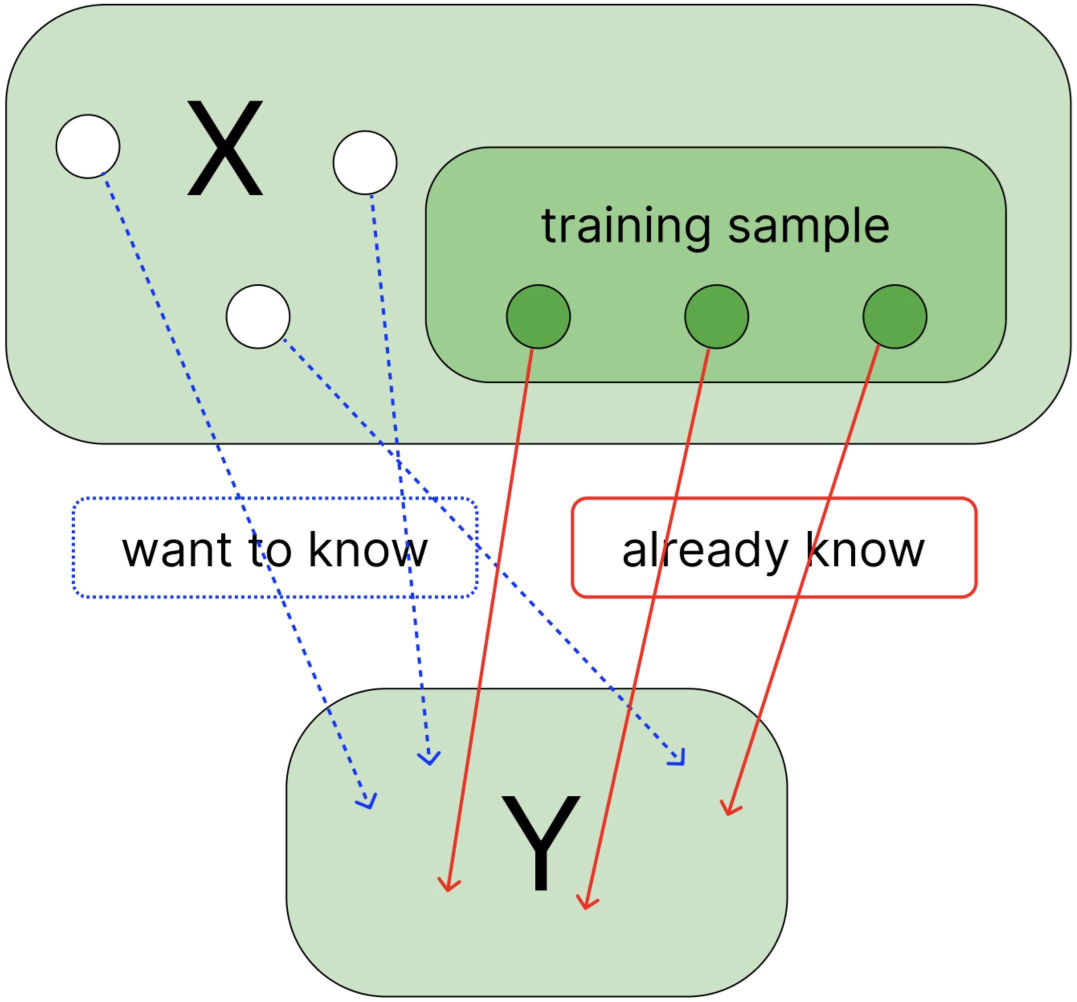
- How do we do this?

- Given: Training sample
of the form $\{(X_i,\ y_i)\}_{i=1}^{n}$

- Goal: To determine and approximate
as accurately as possible our ƒ

Supervised learning

- How do we do this?
- Given: Training sample of the form $\{(X_i,\ y_i)\}_{i=1}^{n}$
- Goal: To determine and approximate as accurately as possible our f



Unsupervised learning

- In contrast to supervised learning problems, in classic unsupervised learning problems there is X, but there is no training sample (i.e. we do not know the correct answers).

Unsupervised learning

· In contrast to supervised learning problems, in classic unsupervised learning problems there is X, but there is no training sample (i.e. we do not know the correct answers).
· In such problems, we usually minimize the “entropy” of the system: we look for the most successful placement of labels.

Regression problem

- What is a regression problem?

Regression problem

- What is a regression problem?
- Simply, it’s a problem in which we want to predict a certain numerical (real) value

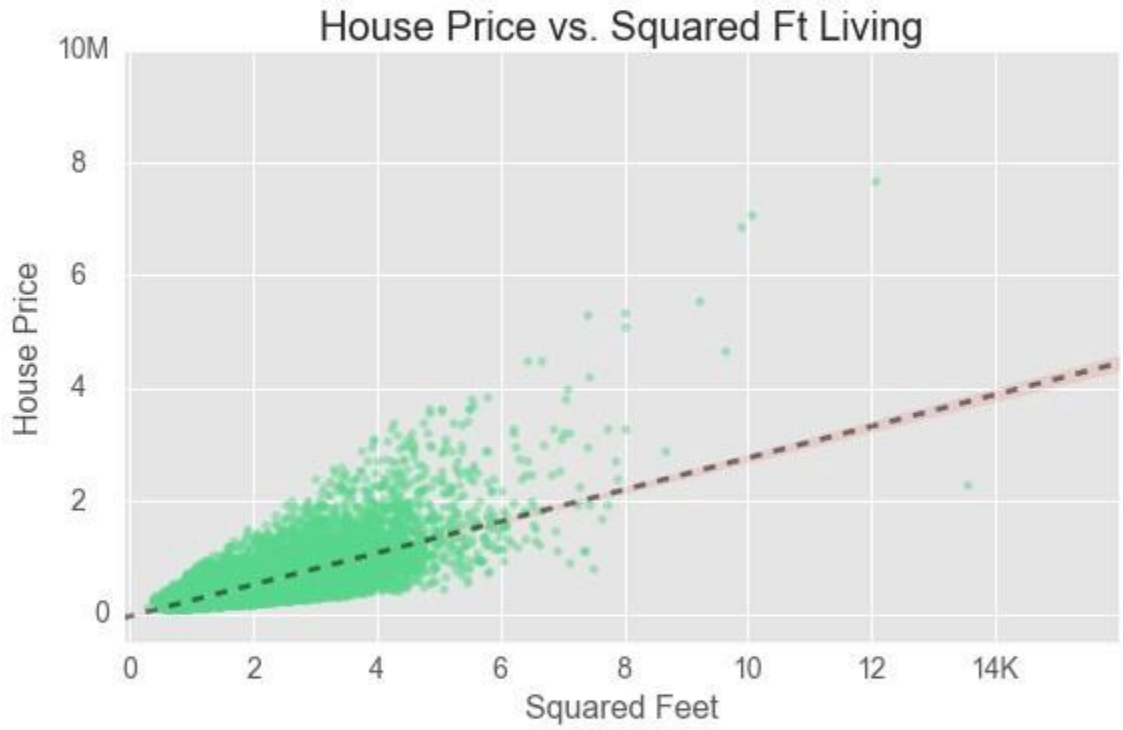
Regression problem

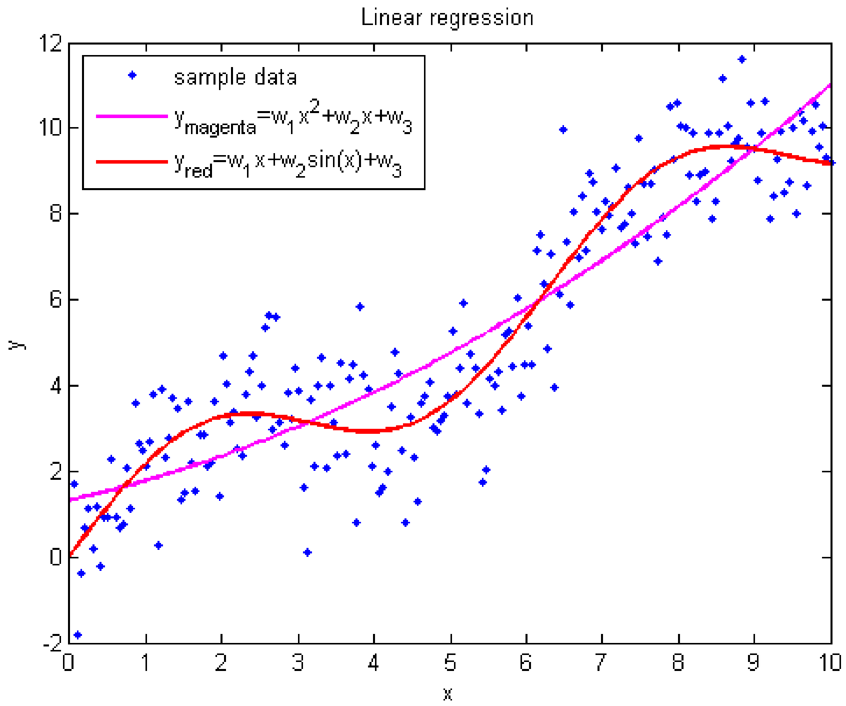
- What is a regression problem?
- Simply, it’s a problem in which we want to predict a certain numerical (real) value
- Regression refers to supervised learning
- Give examples of some regression problems

Regression problem

- What is a regression problem?
- Simply, it’s a problem in which we want to predict a certain numerical (real) value
- Regression refers to supervised learning
- Give examples of some regression problems
  - Predicting the cost of housing for a real estate company
  - Delivery time prediction
  - Predicting taxi cost in a specific area at a specific time tomorrow
  - And so on

Regression problem





Classification problem

- What is a classification problem?

Classification problem

- What is a classification problem?
- Simply, it’s a task where we want to predict whether an object belongs to one of the predetermined classes (categories)

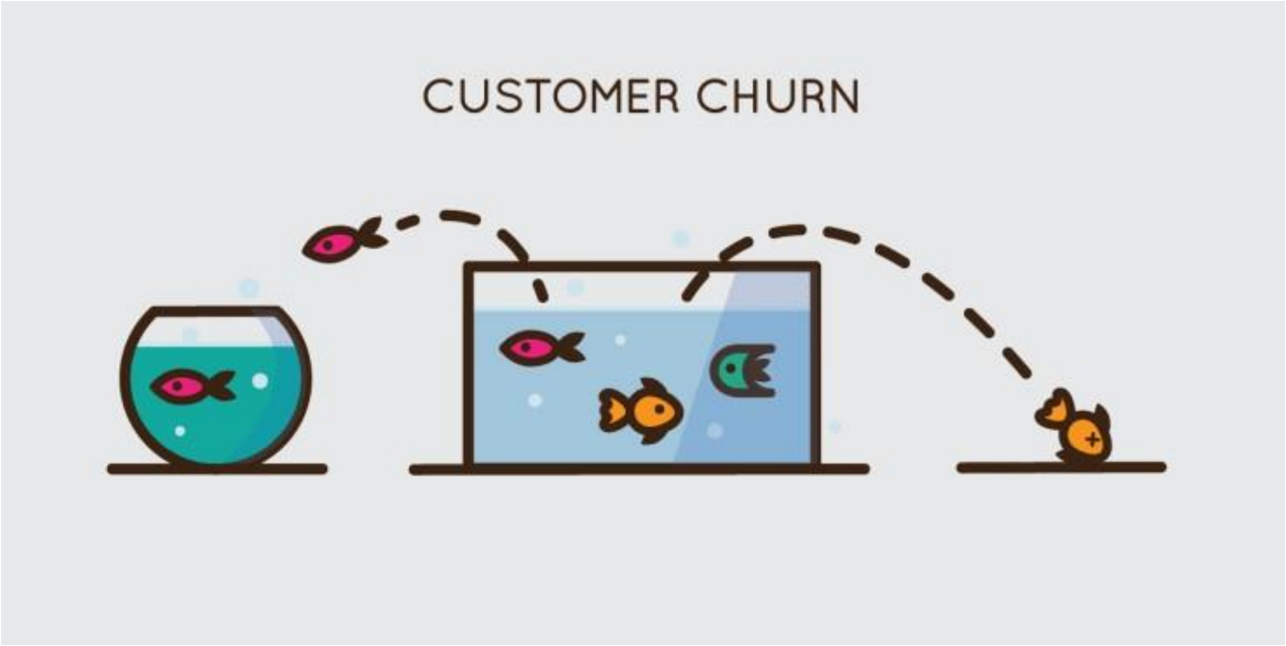
Classification problem

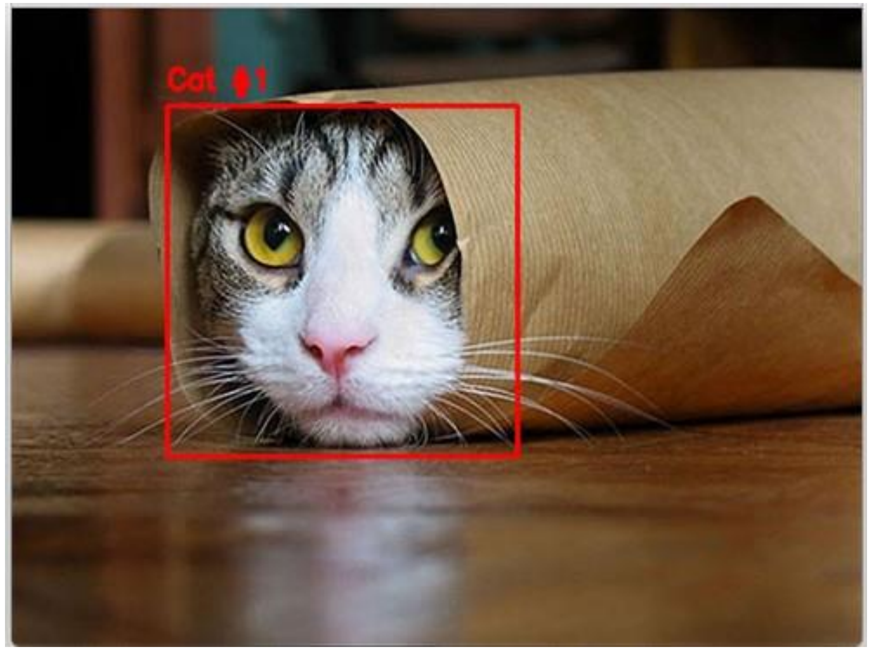
- What is a classification problem?
- Simply, it’s a task where we want to predict whether an object belongs to one of the predetermined classes (categories)
- Classification, just like regression, refers to supervised learning
- Give examples of some classification problems

Classification problem

- What is a classification problem?
- Simply, it’s a task where we want to predict whether an object belongs to one of the predetermined classes (categories)
- Classification, just like regression, refers to supervised learning
- Give examples of some classification problems
  - Predicting customer/employee churn based on their behavior
  - Classification of tissue cells into healthy and tumor cells
  - Detection of objects in photos
  - And so on

Classification problem





Clustering problem

- What is a clustering problem?

Clustering problem

- What is a classification problem?
- Simply, it’s a task where we want to divide our objects into groups (segments), without knowing in advance the criteria and principles of division, but at the same time make the objects in the groups be as similar as possible to each other

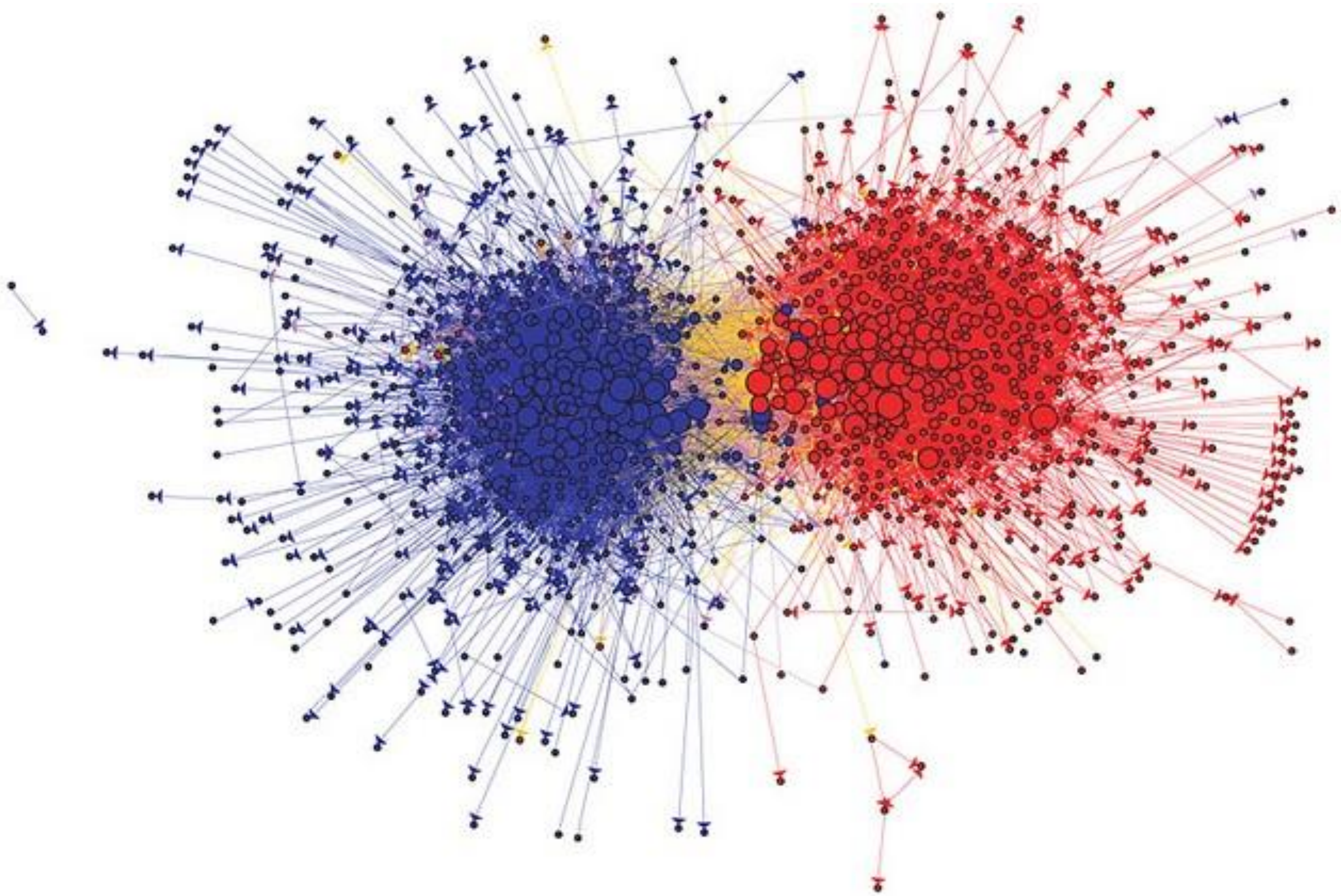
Clustering problem

- Clustering, unlike the previous two, refers to unsupervised learning.
- Give examples of some clustering problems

Clustering problem

- Clustering, unlike the previous two, refers to unsupervised learning.
- Give examples of some clustering problems
  - Audience segmentation for advertising targeting
  - Identifying cell types in a sequencing data sample
  - Search for communities in the social graph (from a social network or from insider information about the organization’s structure)
  - The problem of separating a mix of distributions
  - And so on

Clustering problem



Metric algorithms

Metric algorithms

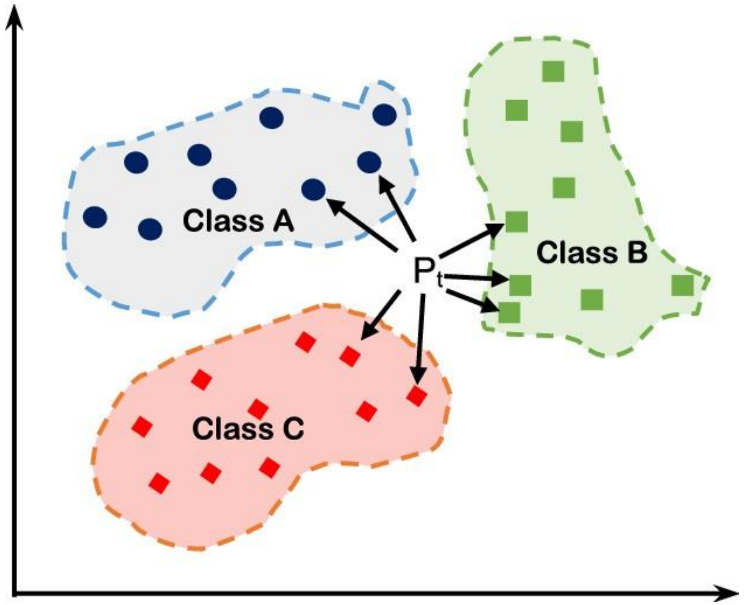
- What are metric algorithms?

Metric algorithms

· What are metric algorithms?
· Metric algorithms in machine learning are algorithms based on calculating similarity between objects (by calculating some metric)

Metric algorithms

- What are metric algorithms?
- Metric algorithms in machine learning are algorithms based on calculating similarity between objects (by calculating some metric)



Metric algorithms

· There are a lot of different metric algorithms, as well as many nuances regarding the metrics used in them.
· Now we will not dive into the nuances of the entire structure of these things - but instead, consider the idea of one of the simplest and at the same time classic machine learning algorithms - the KNN algorithm. This algorithm is a prominent representative of the class of metric algorithms that are of interest to us now.

KNN algorithm

KNN is an example
“lazy” and non-parametric
machine learning algorithm

KNN algorithm

KNN is an example
"lazy" and non-parametric
machine learning algorithm

We will find out what these weird
words mean a little bit later :)

KNN algorithm

- The idea of the algorithm:

KNN algorithm

- The idea of the algorithm:
- The input is a vector - a feature description of some object

KNN algorithm

- The idea of the algorithm:
- The input is a vector - a feature description of some object
- Find the K vectors closest to it for which the answer is known

KNN algorithm

- The idea of the algorithm:
- The input is a vector - a feature description of some object
- Find the K vectors closest to it for which the answer is known

In fact, what does “closest” mean - this is the main point... But we’ll skip this question for now :)

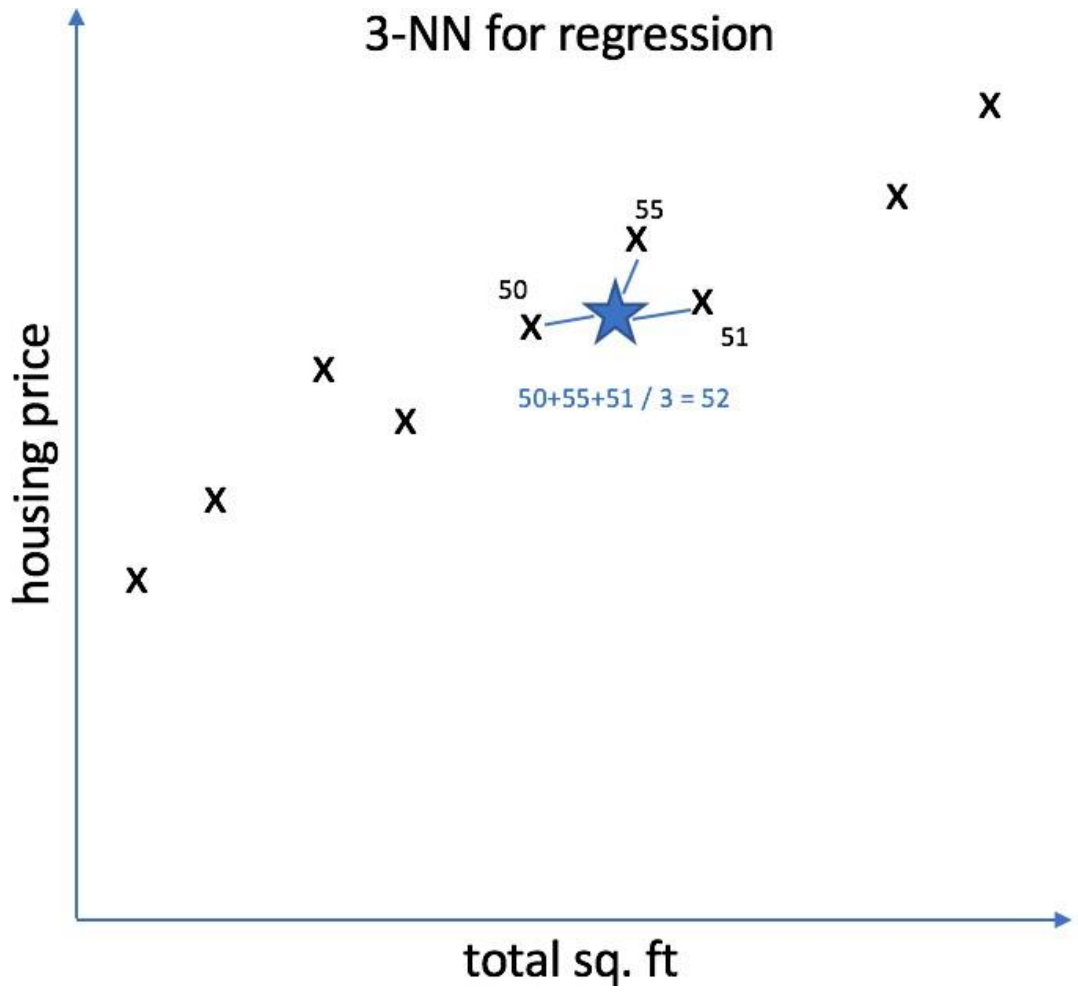
KNN algorithm

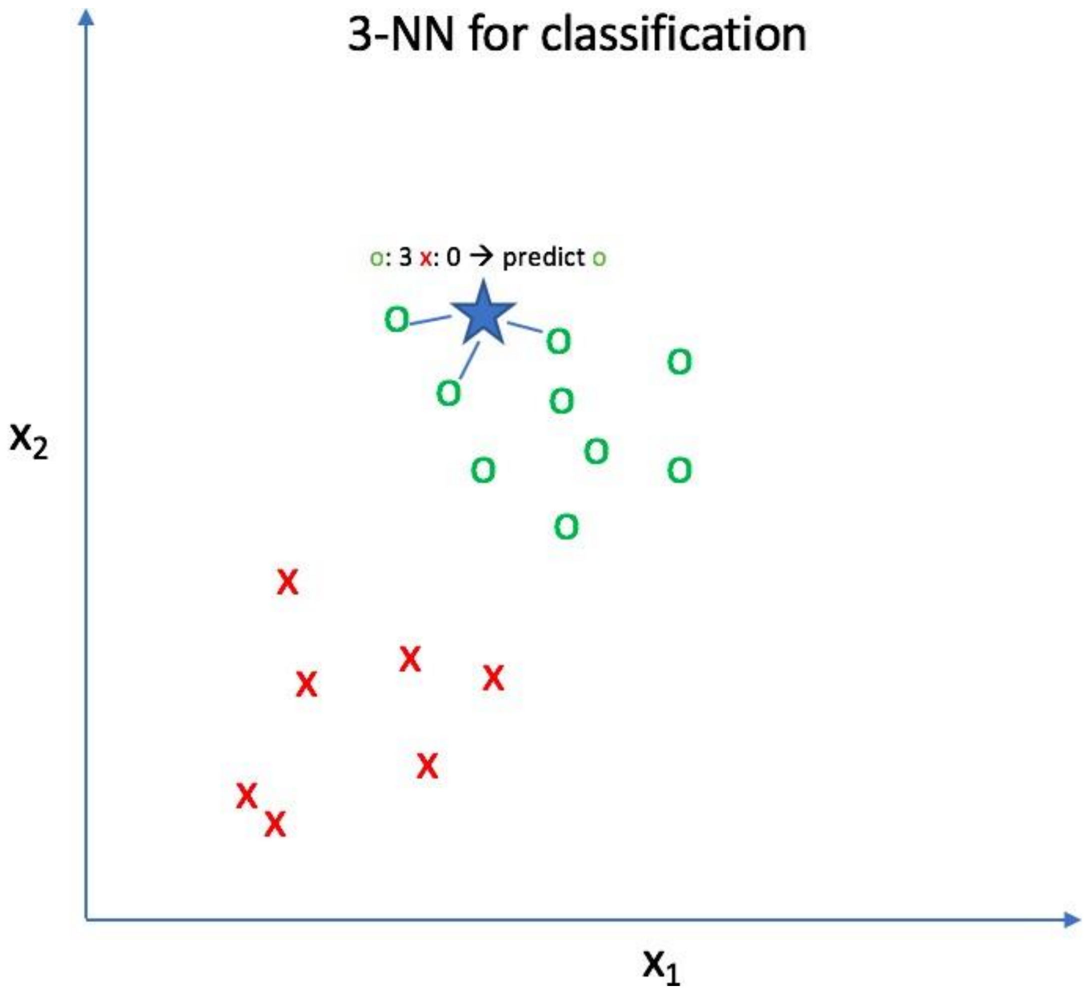
- The idea of the algorithm:
- The input is a vector - a feature description of some object
- Find the K vectors closest to it for which the answer is known
- The answer for the new object is selected using:
  - Averaging, in case of regression
  - Voting, in case of classification

KNN algorithm

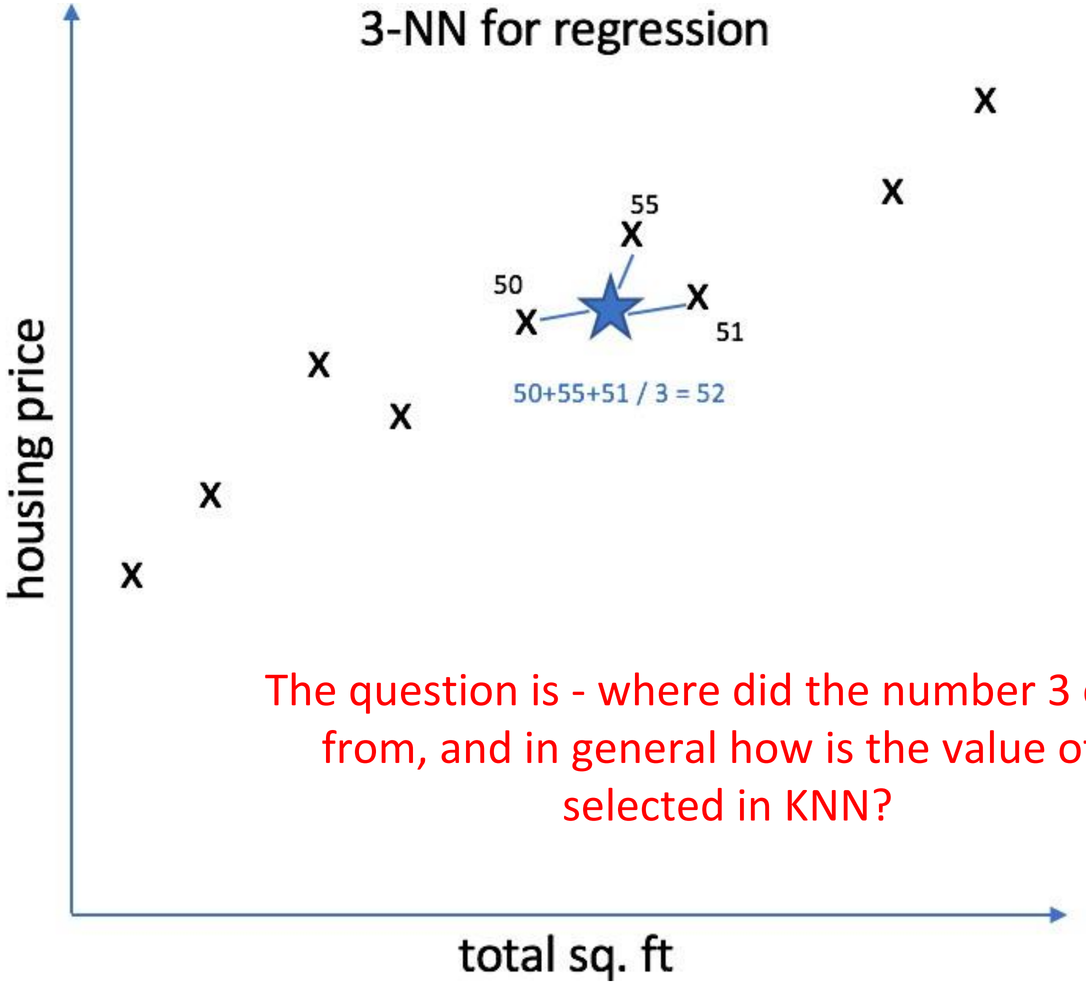
- The idea of the algorithm:
- The input is a vector - a feature description of some object
- Find the K vectors closest to it for which the answer is known
- The answer for the new object is selected using:
  - Averaging, in case of regression
  - Voting, in case of classification
- Averaging/voting with weights and many other modifications of the standard algorithm are also possible

KNN algorithm

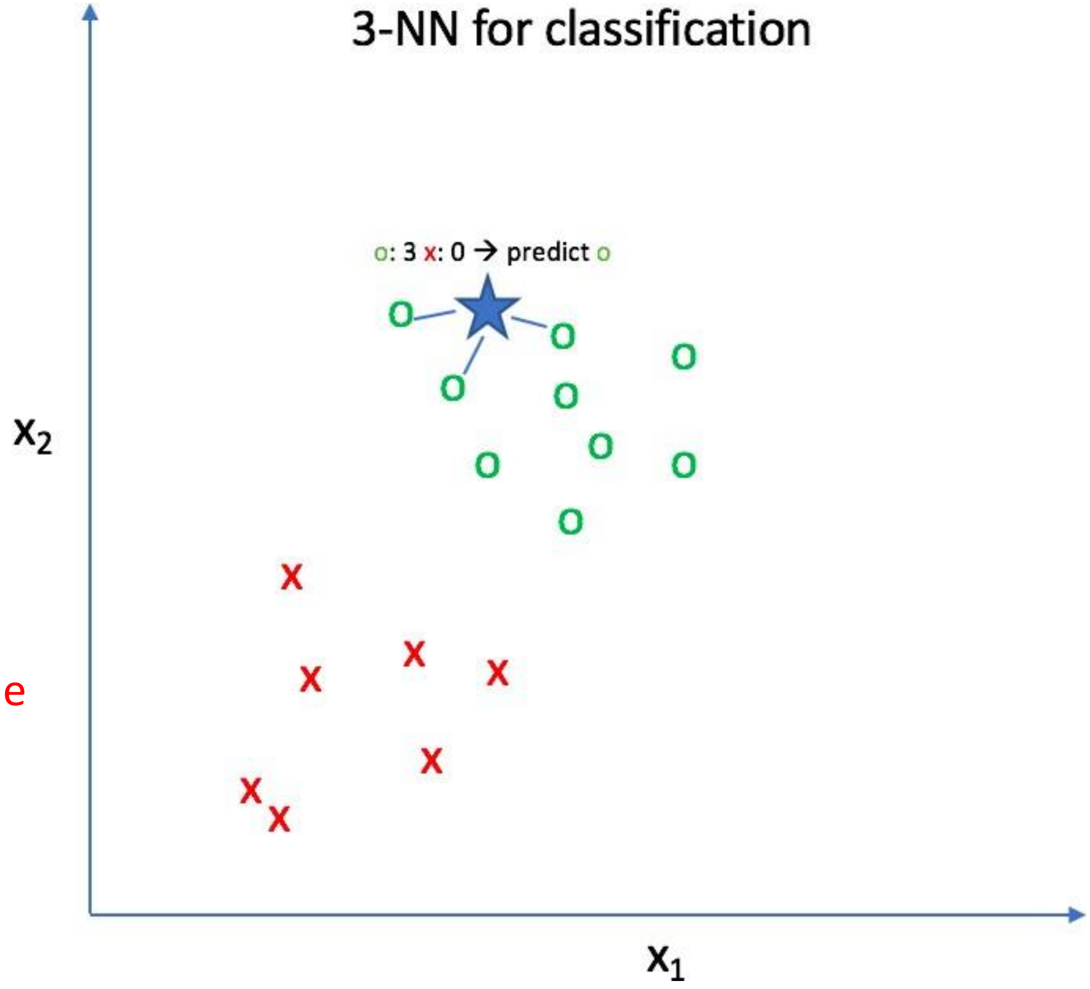




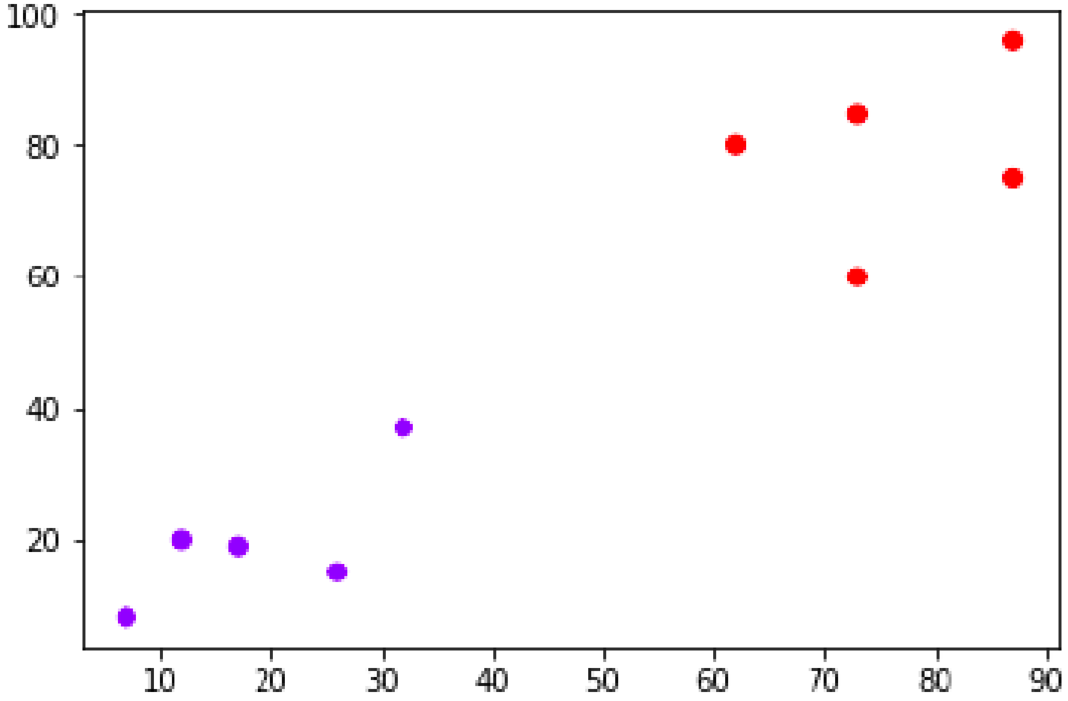
KNN algorithm



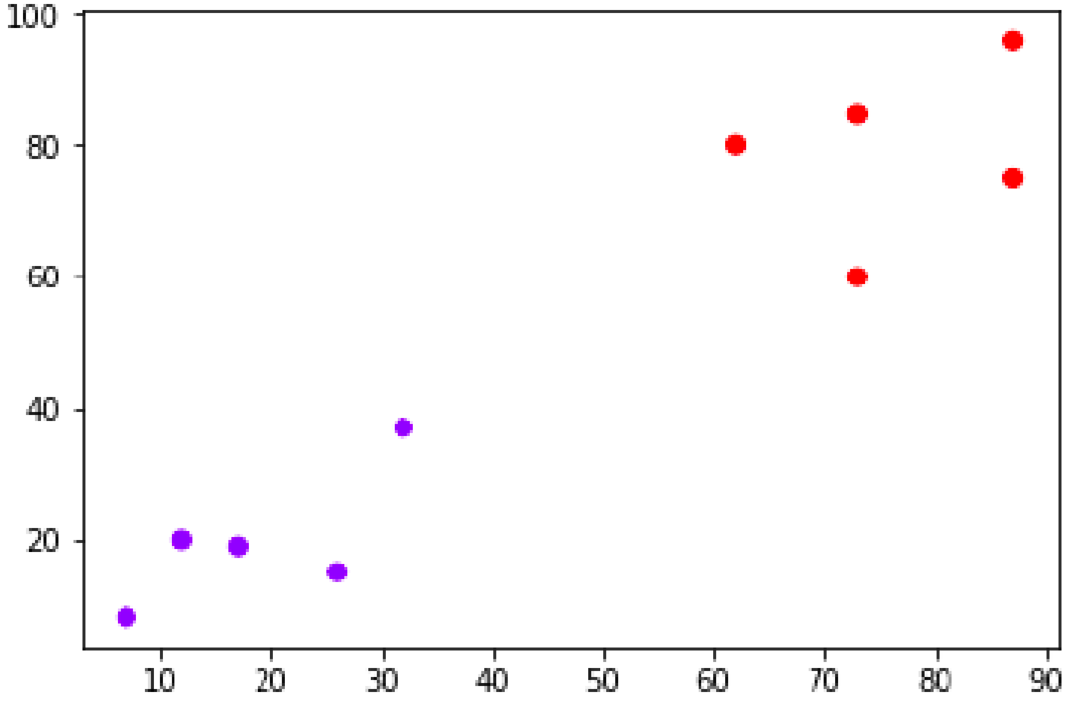
The question is - where did the number 3 come from, and in general how is the value of K selected in KNN?

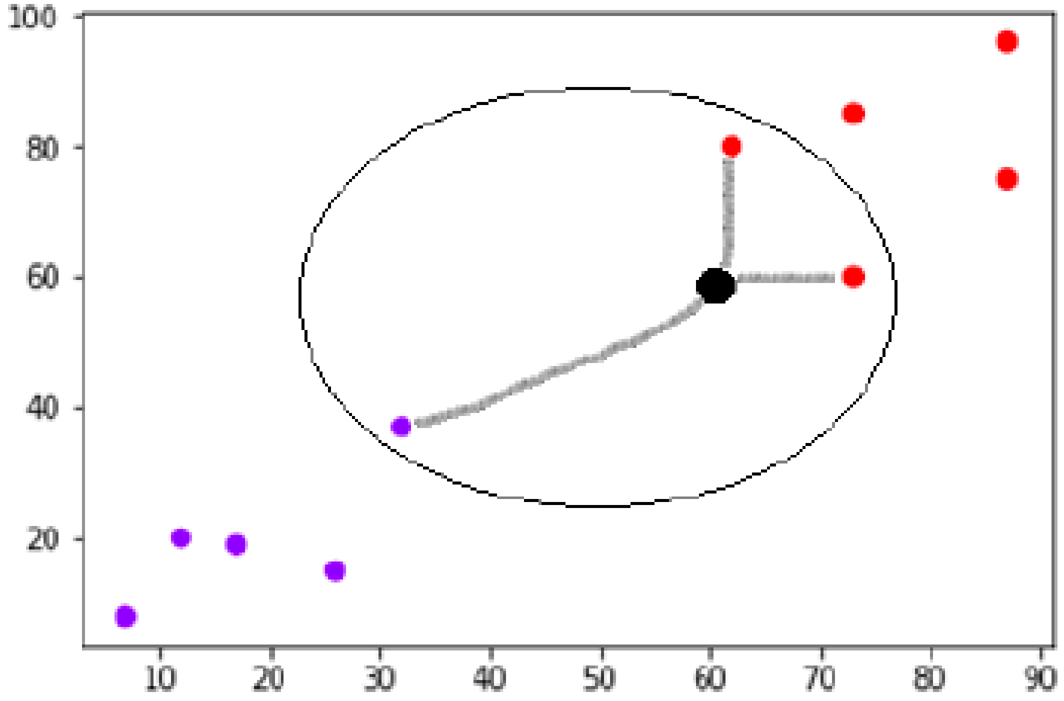


KNN algorithm



KNN algorithm





KNN algorithm

- Now we will omit the details of the KNN implementation - it was important for us to understand the intuition of this algorithm.

KNN algorithm

- Now we will omit the details of the KNN implementation - it was important for us to understand the intuition of this algorithm.
- In the future, this intuition will be useful to us in all metric algorithms!

Training ML algorithms

Training ML algorithms

- In supervised learning tasks, it is common to divide the sample into three non-overlapping parts.
- What are these parts?

Training ML algorithms

- In supervised learning tasks, it is common to divide the sample into three non-overlapping parts.
- What are these parts?
- Training sample
  - The model is trained on it

Training ML algorithms

- In supervised learning tasks, it is common to divide the sample into three non-overlapping parts.
- What are these parts?
- Training sample
  - The model is trained on it
- Validation sample
  - Quality metrics are calculated on it, and hyperparameters are selected based on them

Training ML algorithms

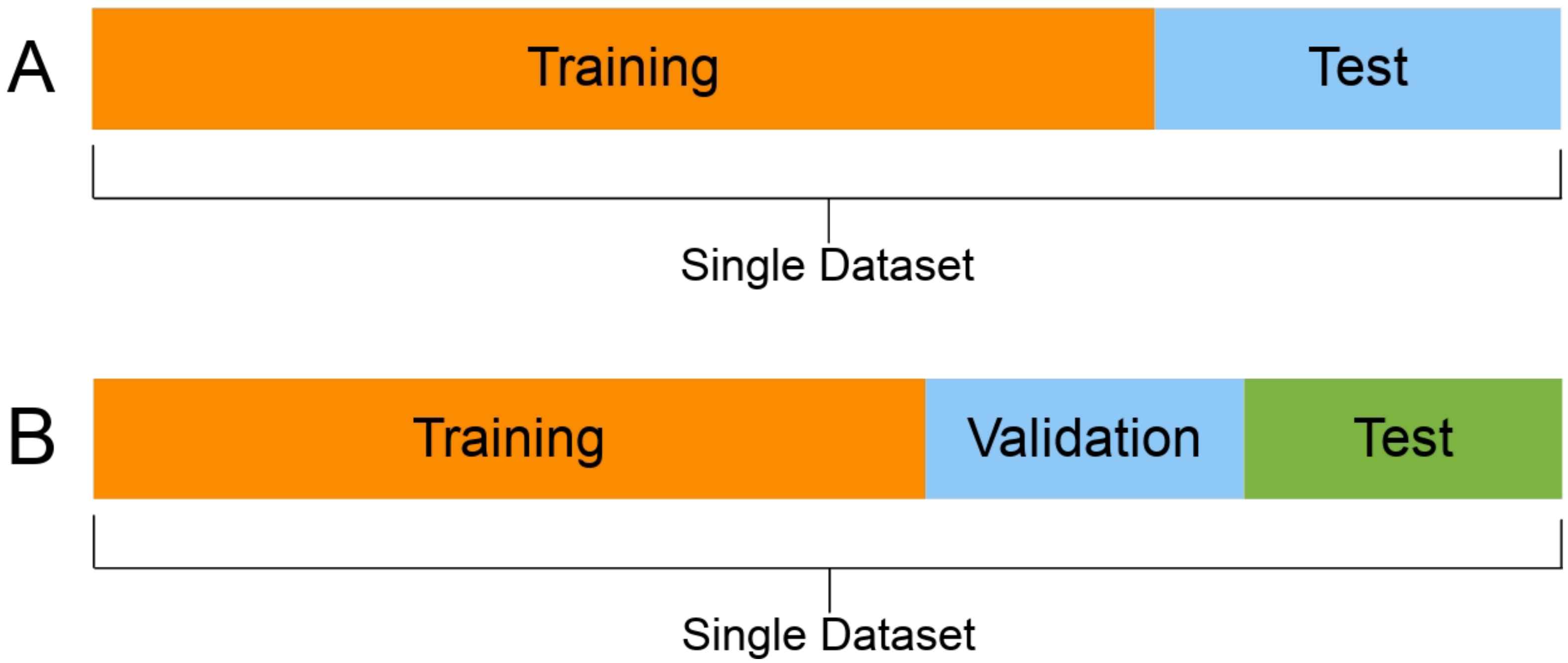
- In supervised learning tasks, it is common to divide the sample into three non-overlapping parts.
- What are these parts?
- Training sample
  - The model is trained on it
- Validation sample
  - Quality metrics are calculated on it, and hyperparameters are selected based on them

What the quality metrics and hyperparameters are
we will discuss later

Training ML algorithms

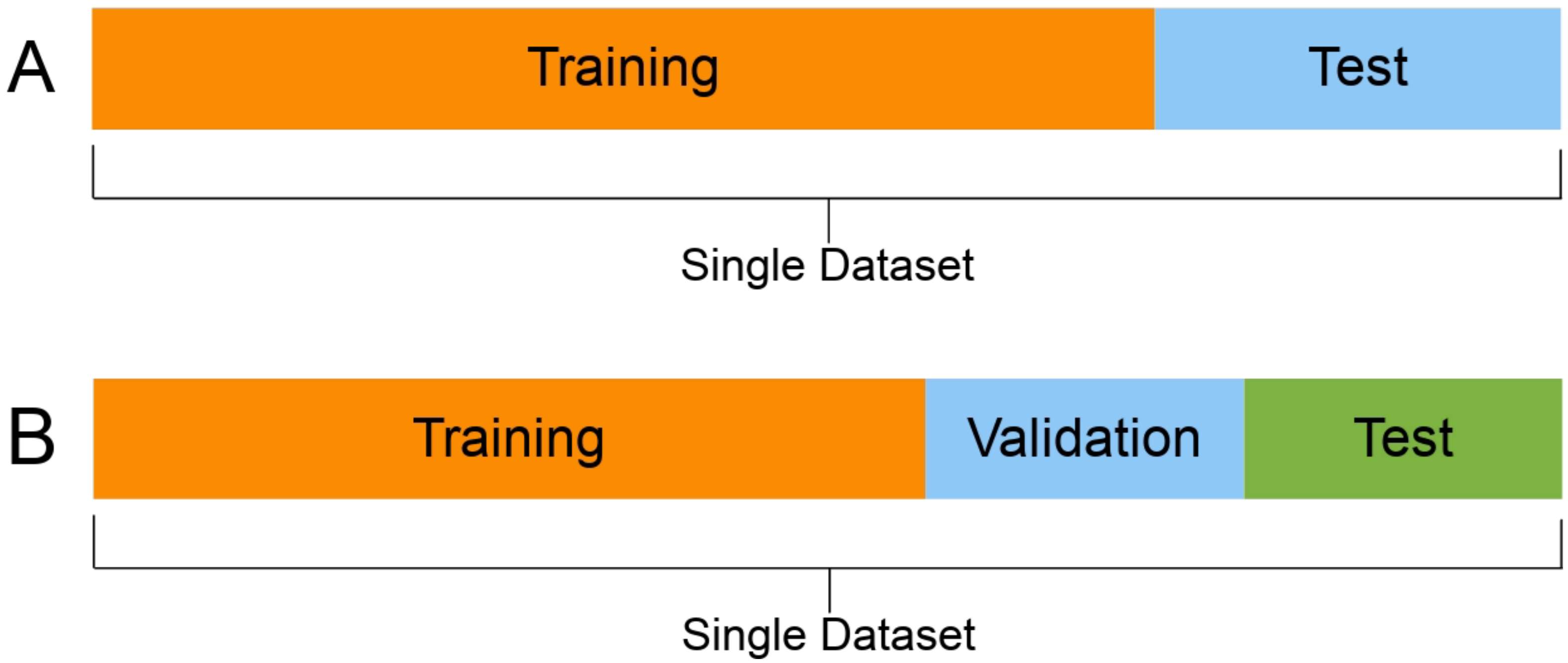
- In supervised learning tasks, it is common to divide the sample into three non-overlapping parts.
- What are these parts?
- Training sample
  - The model is trained on it
- Validation sample
  - Quality metrics are calculated on it, and hyperparameters are selected based on them
- Test sample
  - It directly evaluates the quality of the trained model

Training ML algorithms



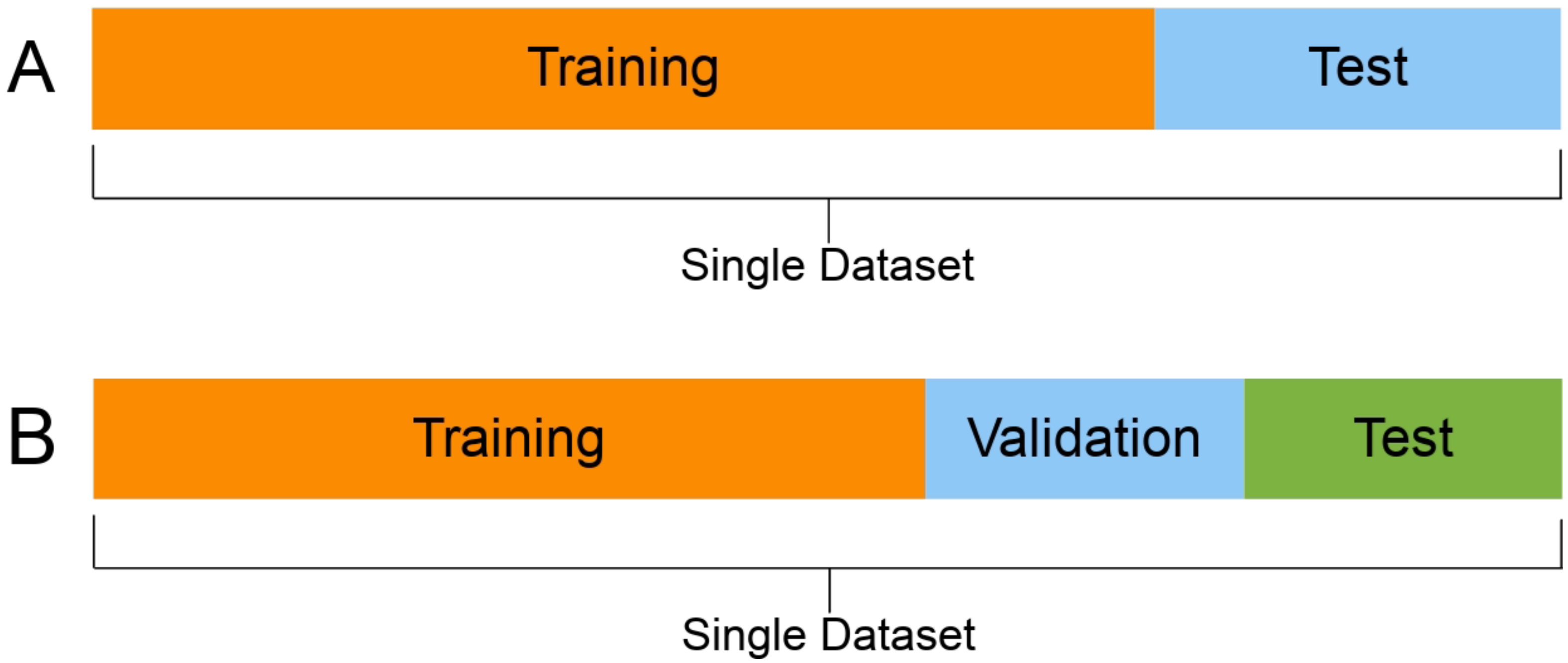
- In fact, a validation sample is not always used.

Training ML algorithms



- In fact, a validation sample is not always used.
- When it’s used, we should try to take it the same size as the test one.

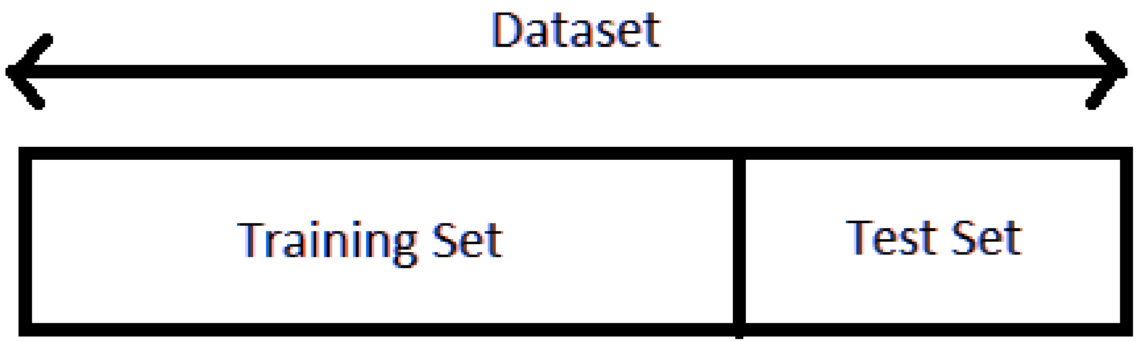
Training ML algorithms



- **Important! Each sample must be representative!**

Holdout sampling (lazy)

- One of the options is to set aside, for example, 20% of the training set for model validation.
- In other words, use 80% of the sample for training and 20% for testing.



Holdout sampling (lazy)

· One of the options is to set aside, for example, 20% of the training set for model validation.
· In other words, use 80% of the sample for training and 20% for testing.

· If you want to evaluate the quality of the algorithm quite honestly and have available resources, you can calculate cross-validation metrics!

Cross-validation

- What is cross-validation? How do you understand this?

Cross-validation

- What is cross-validation? How do you understand this?
- Cross-validation is an approach to separating data into training and validation data using a specific algorithm.

Cross-validation

- What is cross-validation? How do you understand this?
- Cross-validation is an approach to separating data into training and validation data using a specific algorithm.
- The idea of cross-validation:
  - We split the sample into *k* parts

Cross-validation

- What is cross-validation? How do you understand this?
- Cross-validation is an approach to separating data into training and validation data using a specific algorithm.
- The idea of cross-validation:
  - We split the sample into *k* parts
  - Of the received *k* parts, *k - 1* part is used for training and one is used for testing (validation)

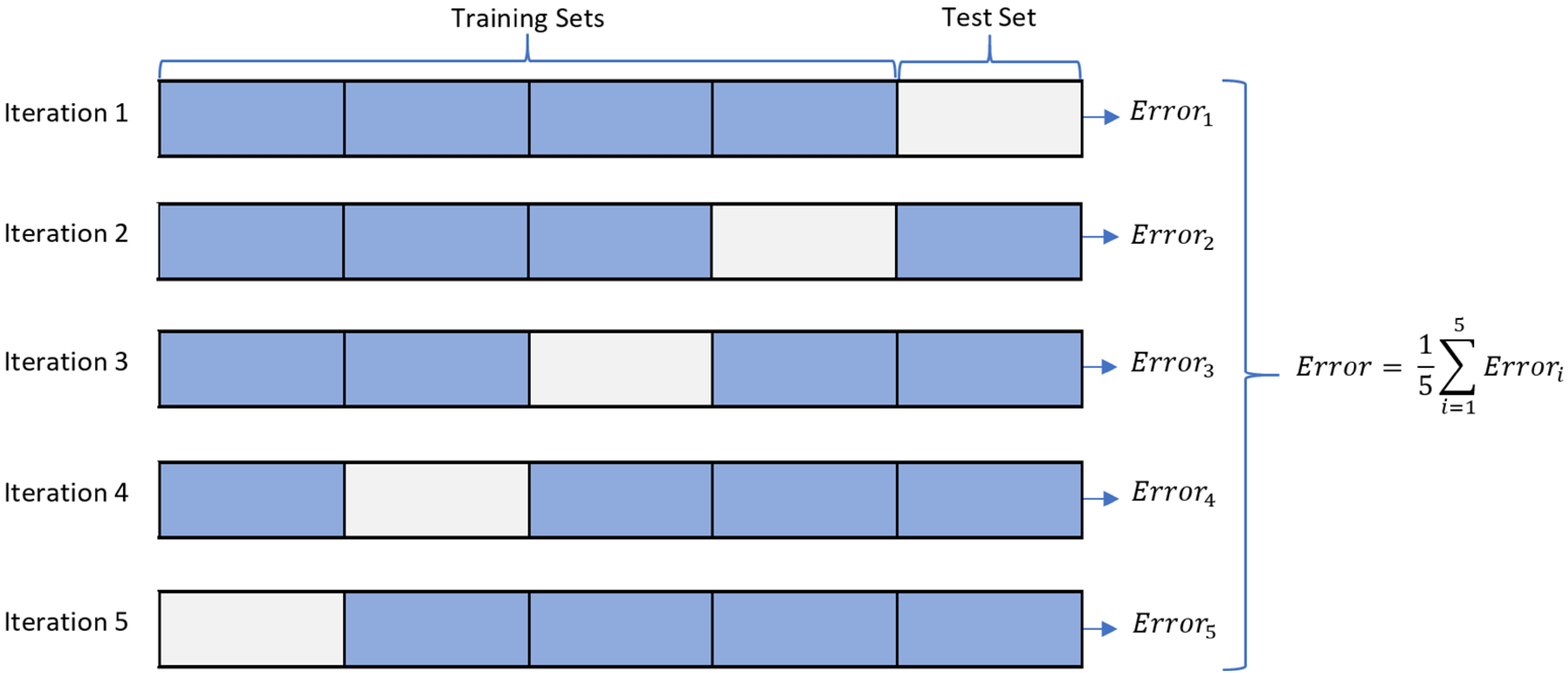
Cross-validation

- What is cross-validation? How do you understand this?
- Cross-validation is an approach to separating data into training and validation data using a specific algorithm.
- The idea of cross-validation:
  - We split the sample into *k* parts
  - Of the received *k* parts, *k - 1* part is used for training and one is used for testing (validation)
  - The process is repeated *k* times. Each time a different part is selected for testing

Cross-validation

- What is cross-validation? How do you understand this?
- Cross-validation is an approach to separating data into training and validation data using a specific algorithm.
- The idea of cross-validation:
  - We split the sample into *k* parts
  - Of the received *k* parts, *k - 1* part is used for training and one is used for testing (validation)
  - The process is repeated *k* times. Each time a different part is selected for testing
  - Test results are averaged

Cross-validation



Cross-validation

- Cross-validation is a powerful tool and an important step in the education process of ML algorithms.

- Advantages:
  - The estimation error is reduced because the whole set is used
  - The quality of the model improves and the optimal hyperparameters of the algorithm can be selected

- Disadvantages:
  - Training is being repeated *k* times. For some models this can be very long