 Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Machine Learning Fundamentals

Zakarya Farou; Krisztian Buza

Department of Artificial Intelligence
Eötvös Loránd University
Budapest, Hungary

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Announcement

- Please send your presentation slides (about your selected topic) in PDF format to buza@inf.elte.hu ,
  Deadline: **18th November 2019, 8:00 am**

- Remember, you will have to present your topic on the **20th/27th November 2019**

- Check out the web page of the course (www.biointelligence.hu/iml) **for the schedule of the presentations**

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

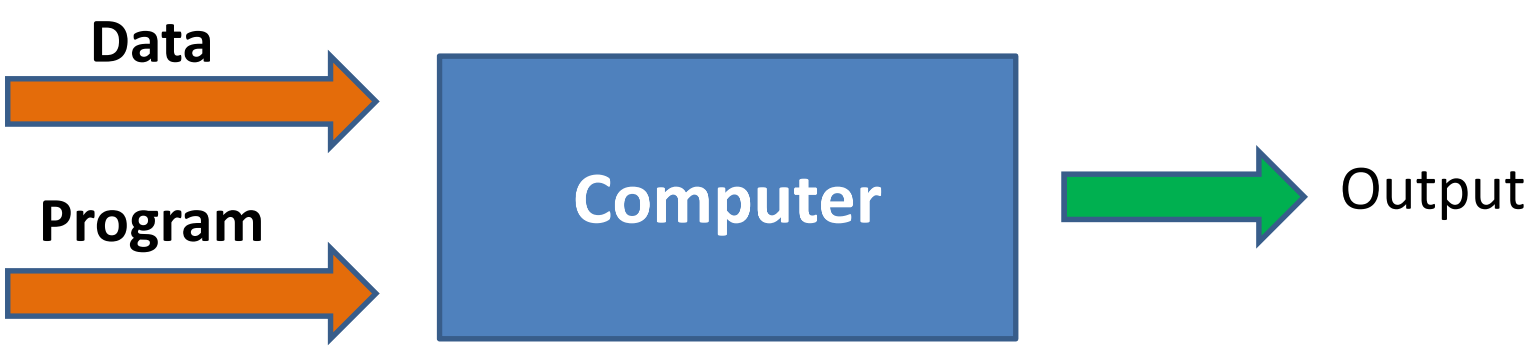
What is Machine Learning?

- Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming

What is Machine Learning?

- Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming

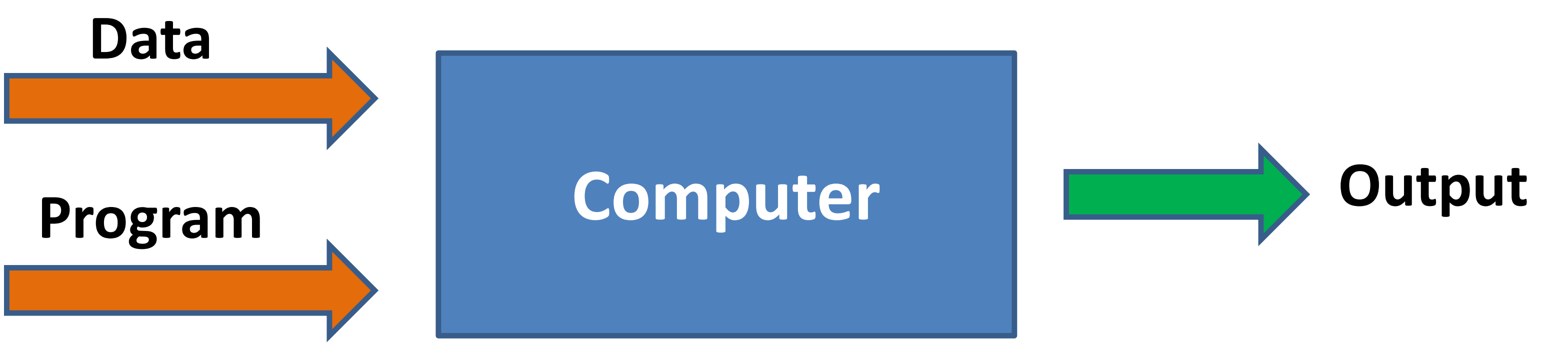
Traditional Programming



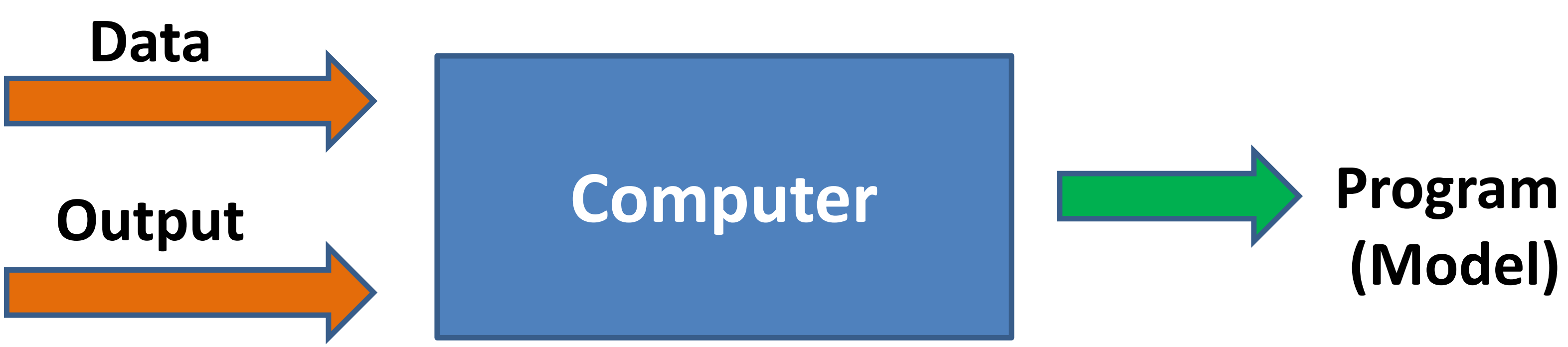
What is Machine Learning?

- Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming

Traditional programming



Machine Learning (Supervised)

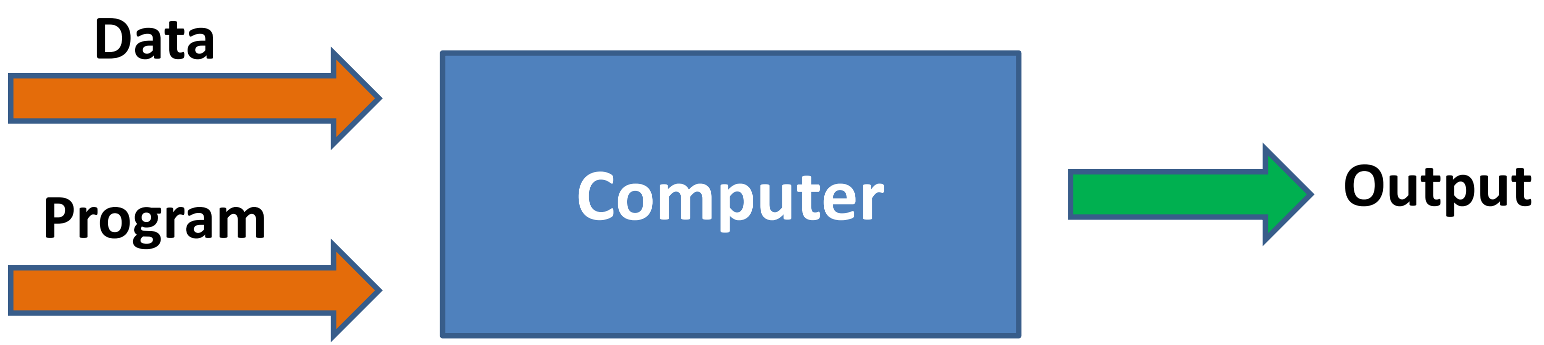


Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

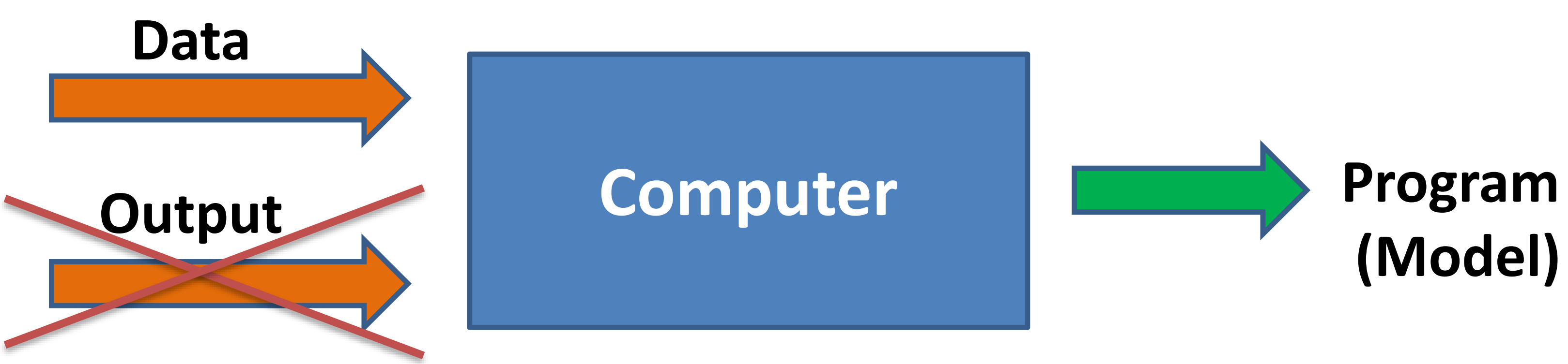
What is Machine Learning?

- Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming

Traditional programming

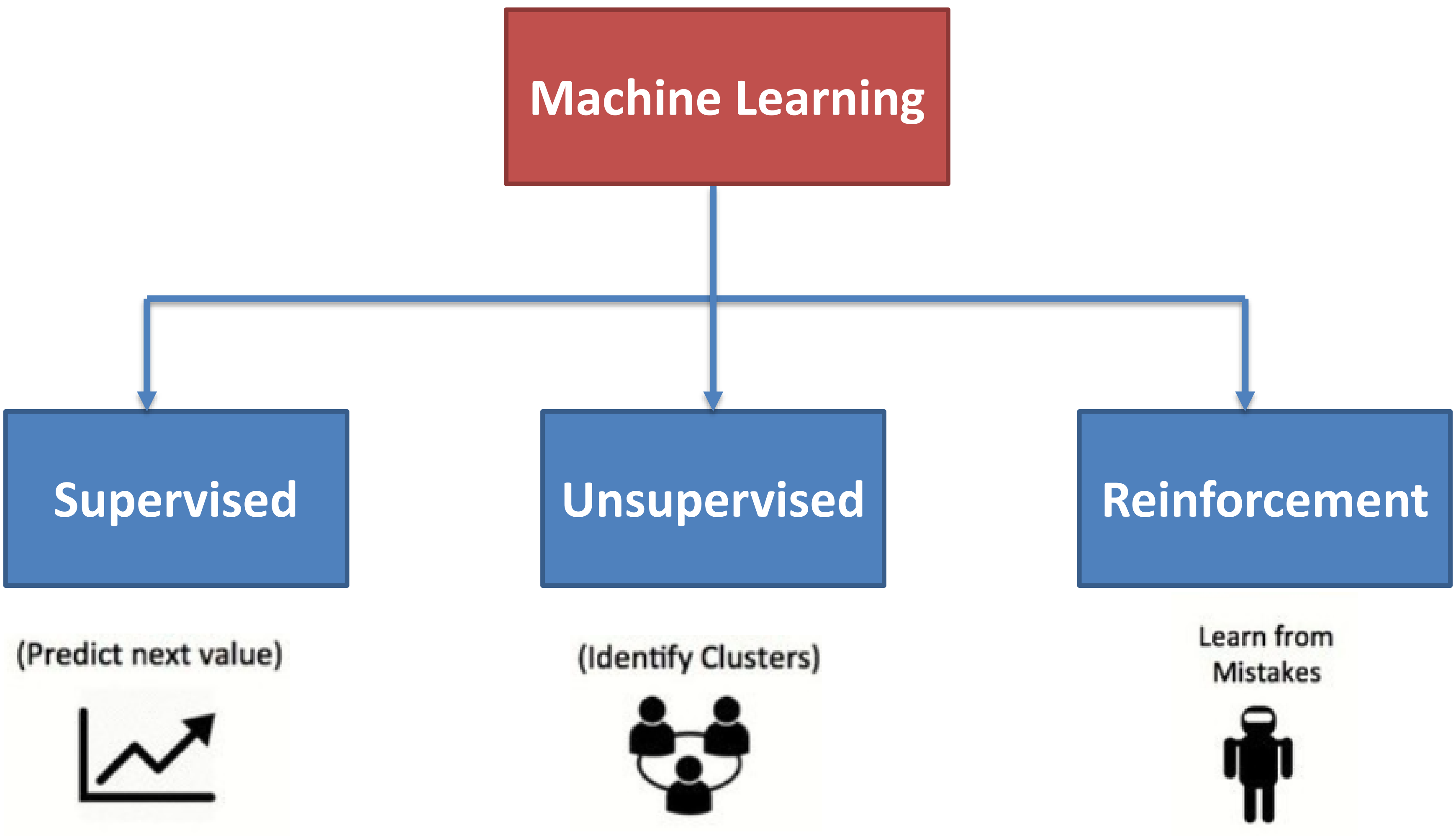


Machine Learning (Unsupervised)



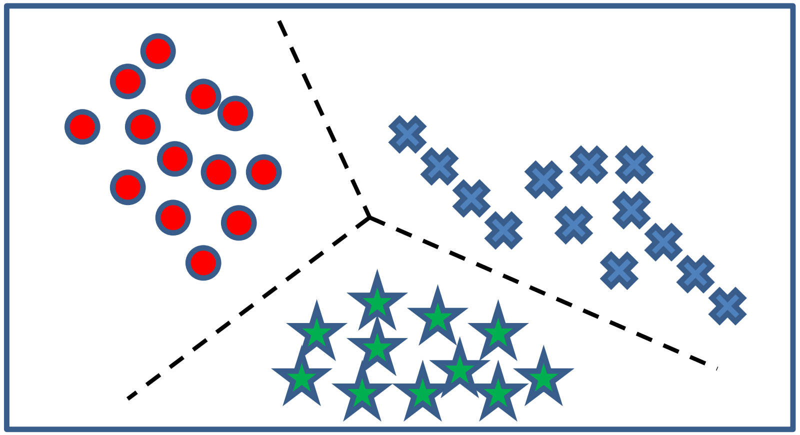
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Types of Machine Learning



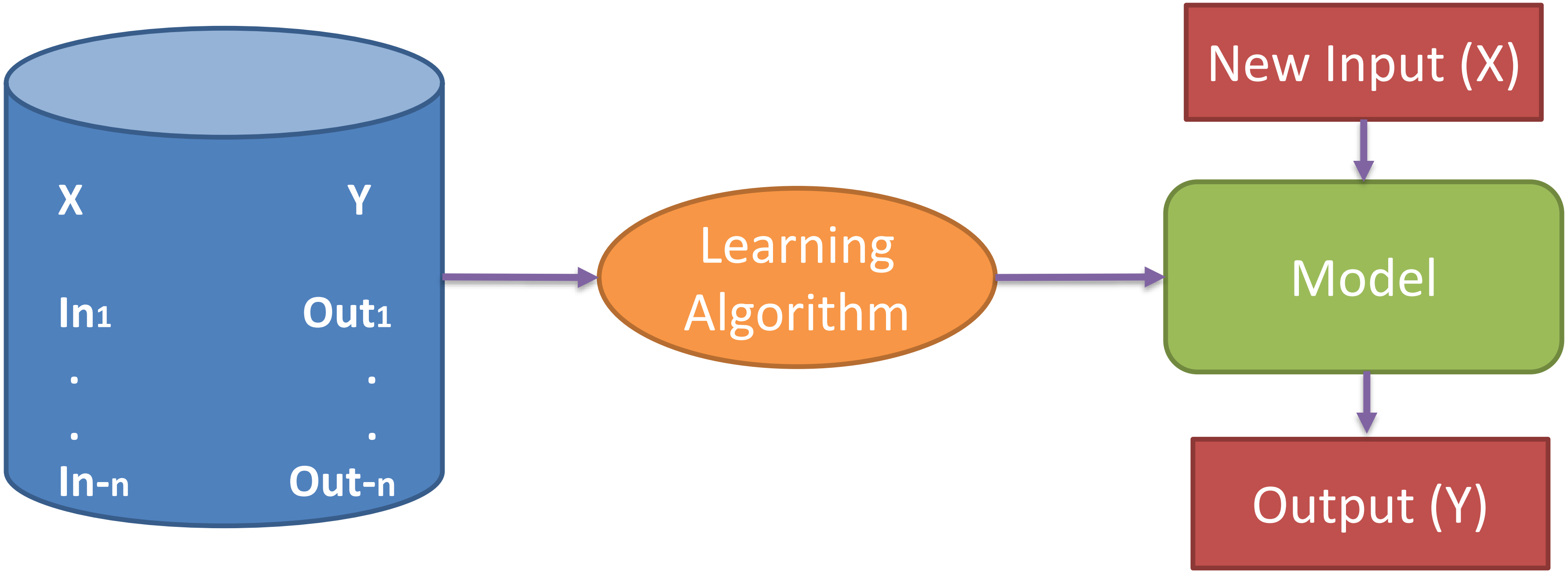
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Types of Machine Learning



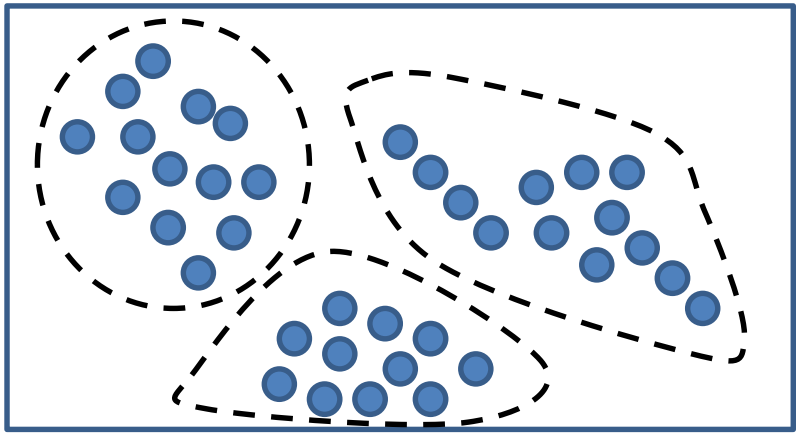
Supervised Learning

In Supervised Learning, an algorithm learns from a dataset



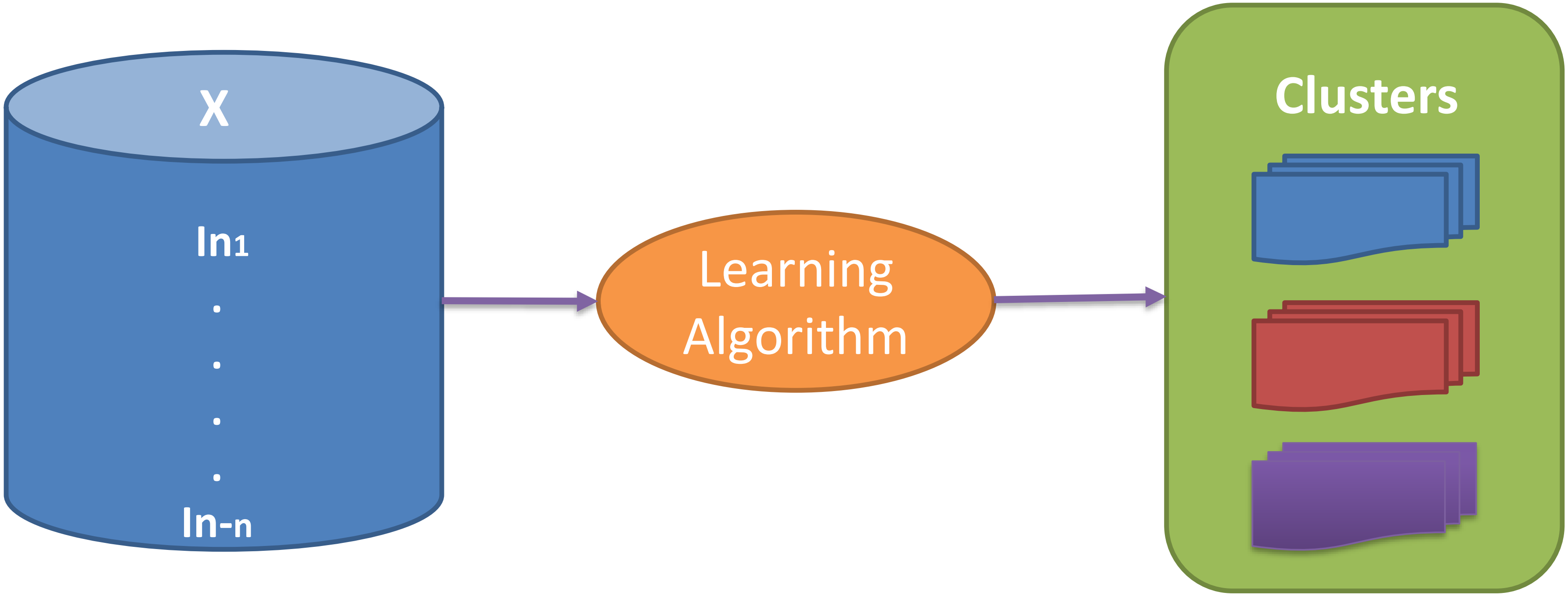
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Types of Machine Learning



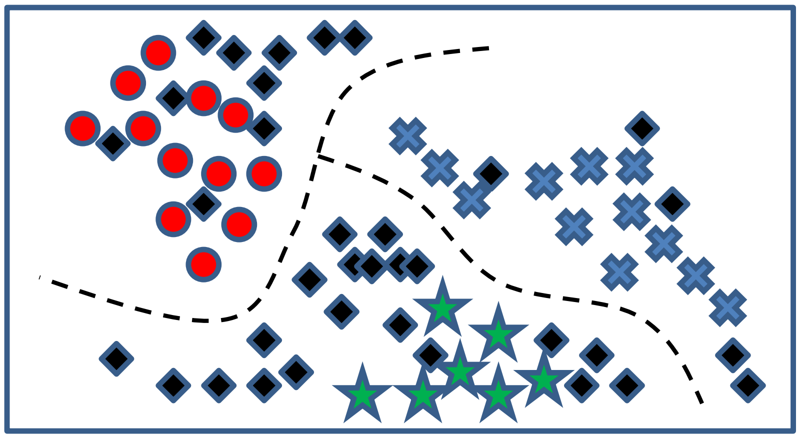
Unsupervised Learning

In Unsupervised Learning, an algorithm is provided with only the data



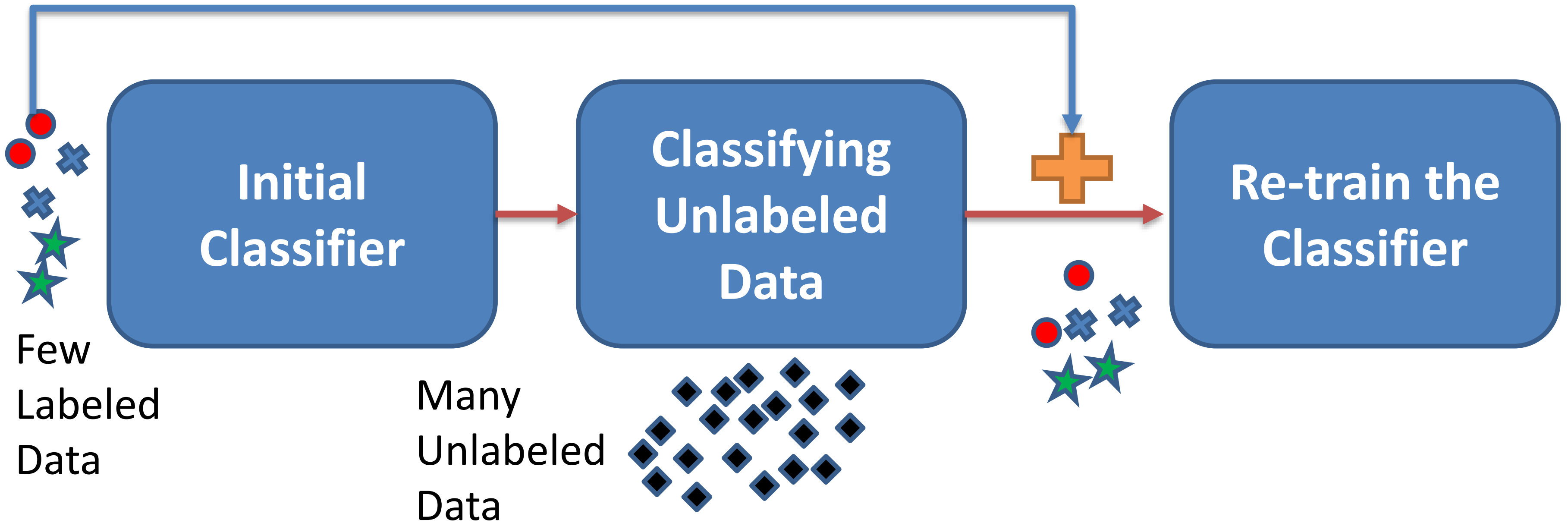
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Types of Machine Learning



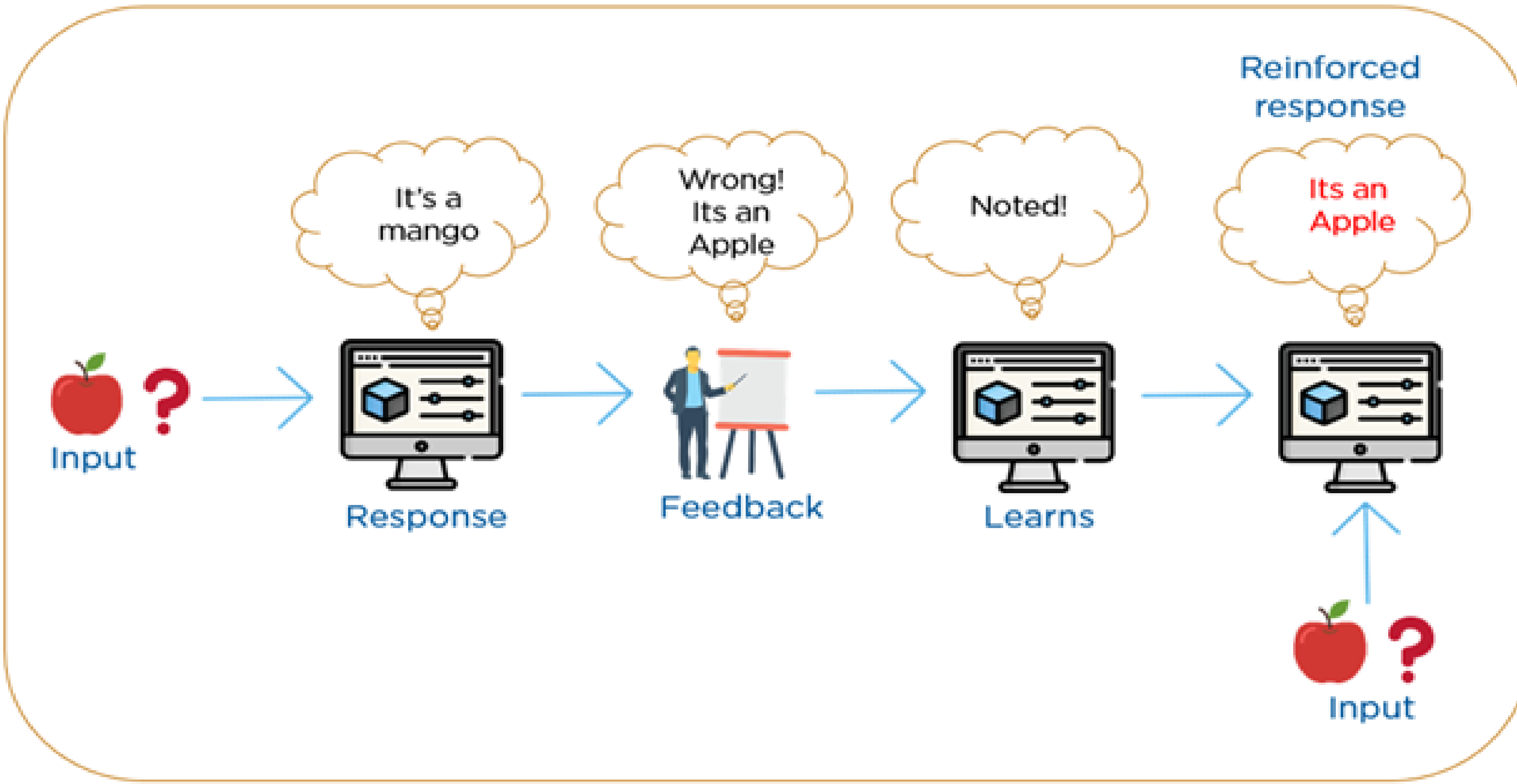
Semi-supervised Learning

Semi-Supervised Learning is, basically, the combination of Supervised and Unsupervised learning



Types of Machine Learning

Reinforcement Learning



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

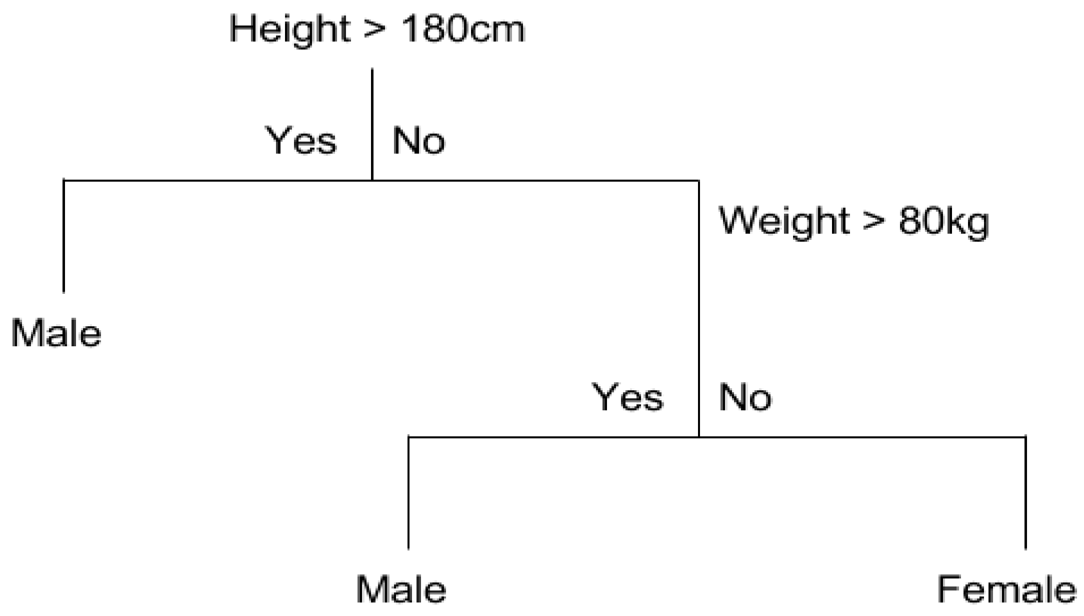
Key Elements of Machine Learning

- Based on state-of-the-art , all machine learning algorithms today are made up of three components. They are as follows:

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Key Elements of Machine Learning

Representation



Information represented by decision trees

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Key Elements of Machine Learning

Representation

Height > 180 cm, then it's probably a male

Height > 180 cm and **Weight** > 80 kg, then it's probably a male

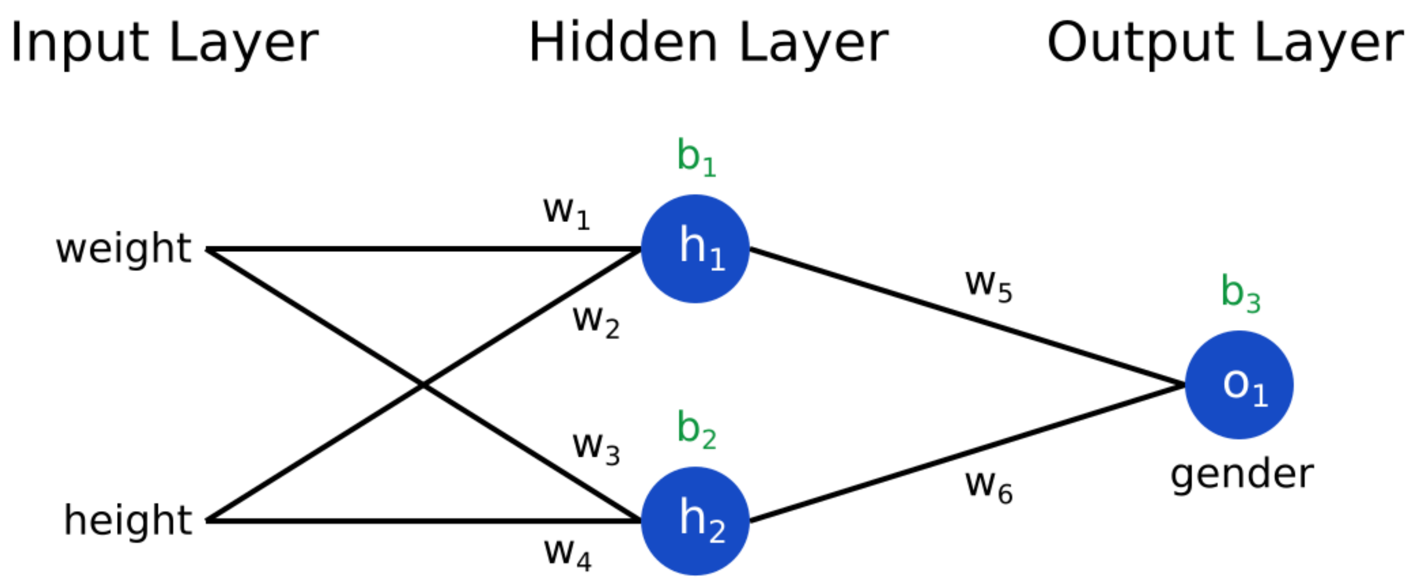
Height ≤ 180 cm and **Weight** ≤ 80 kg , then it's probably a female

Information represented by set of rules

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Key Elements of Machine Learning

Representation



Information represented by neural networks

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Key Elements of Machine Learning

Evaluation

Performance Metrics for Classification problems

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall (sensitivity)
5. F1 Score

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

Imagine that we have this medical data

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

We have got some clinical measurements

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

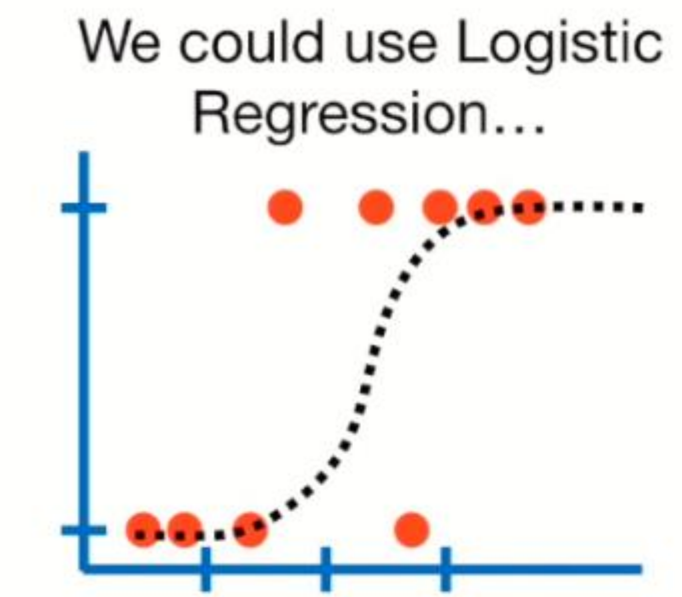
Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

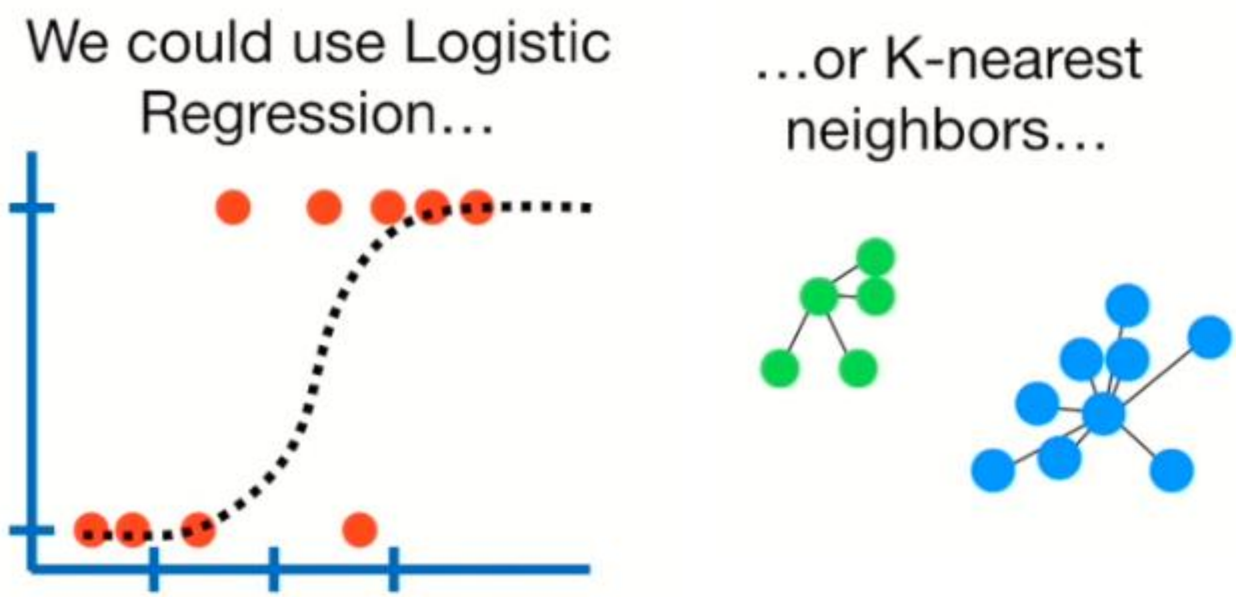
And we want to apply a ML method to predict whether someone will develop heart disease or not

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

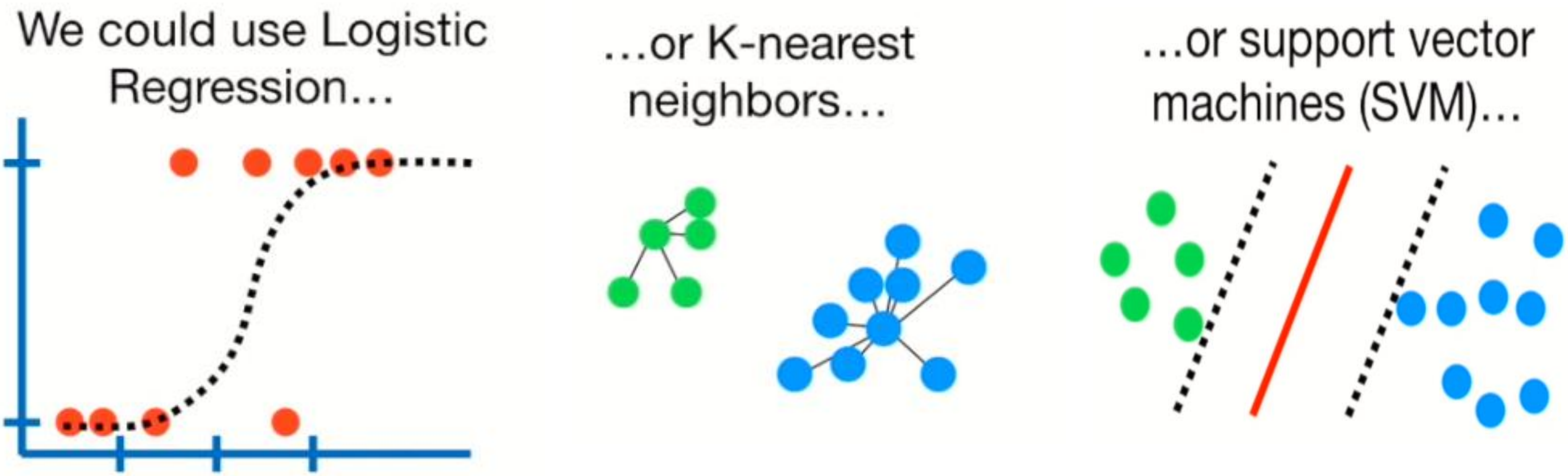
Compare Machine Learning Methods



Compare Machine Learning Methods

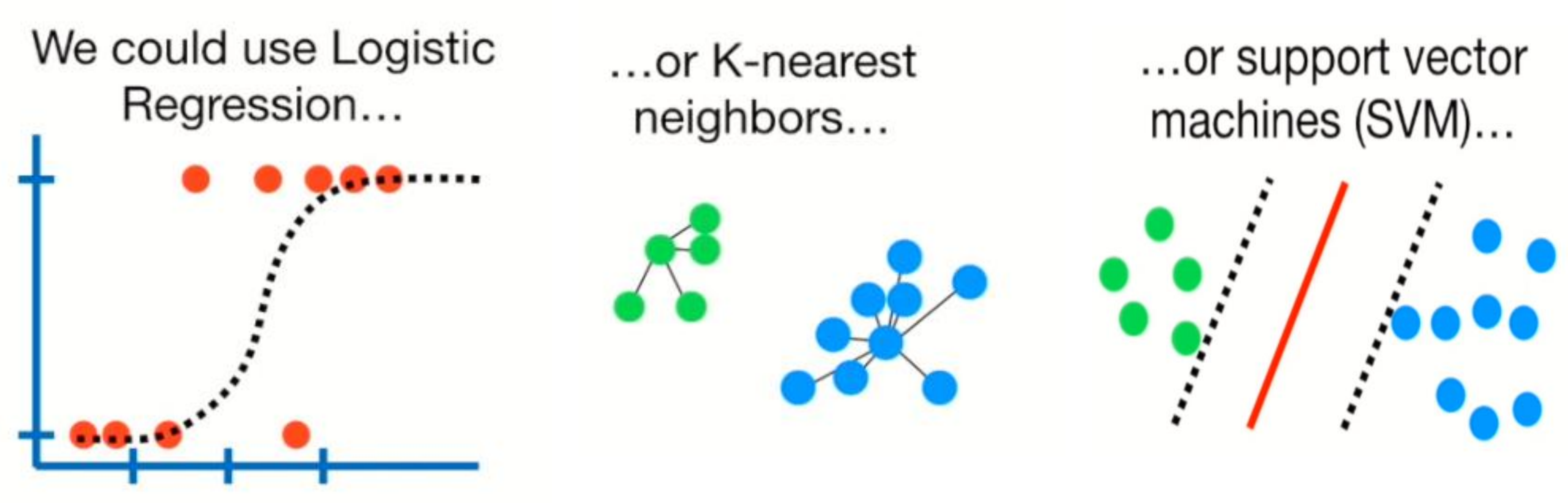


Compare Machine Learning Methods



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods



How to decide which one works better with our data ?

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
…	…	…	…	…

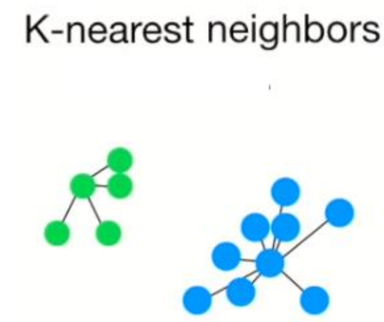
Training Data

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	Yes	No	140	No
…	…	…	…	…

Testing Data

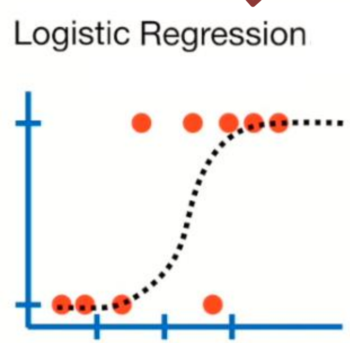
Compare Machine Learning Methods

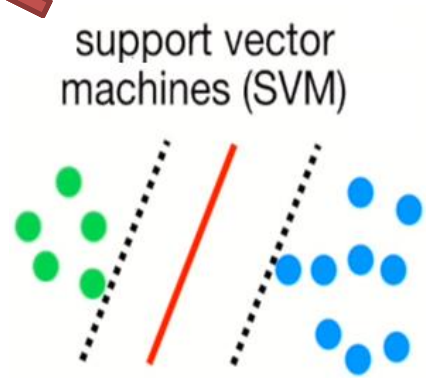
Confusion Matrix



Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
…	…	…	…	…

Training Data

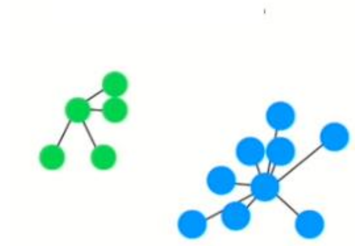




Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

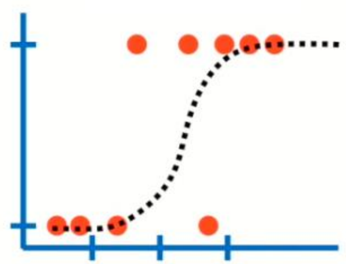
K-nearest neighbors



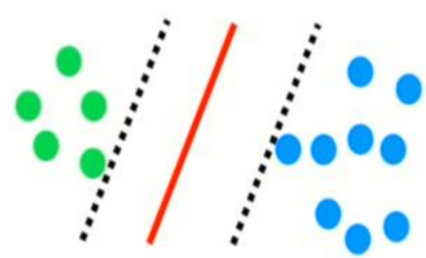
Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	Yes	No	140	No
…	…	…	…	…

Testing Data

Logistic Regression

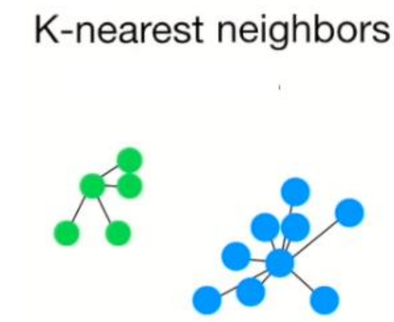


support vector machines (SVM)



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

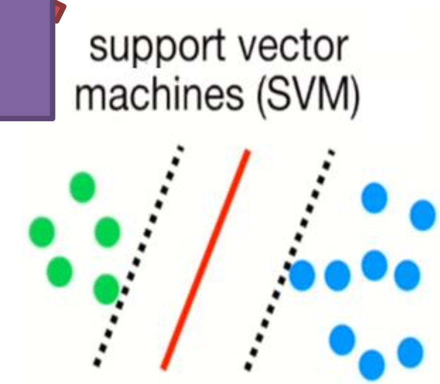
Compare Machine Learning Methods



Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	Yes			
…	…			

Testing Data

Now we need to summarize how each method performed on the Testing data



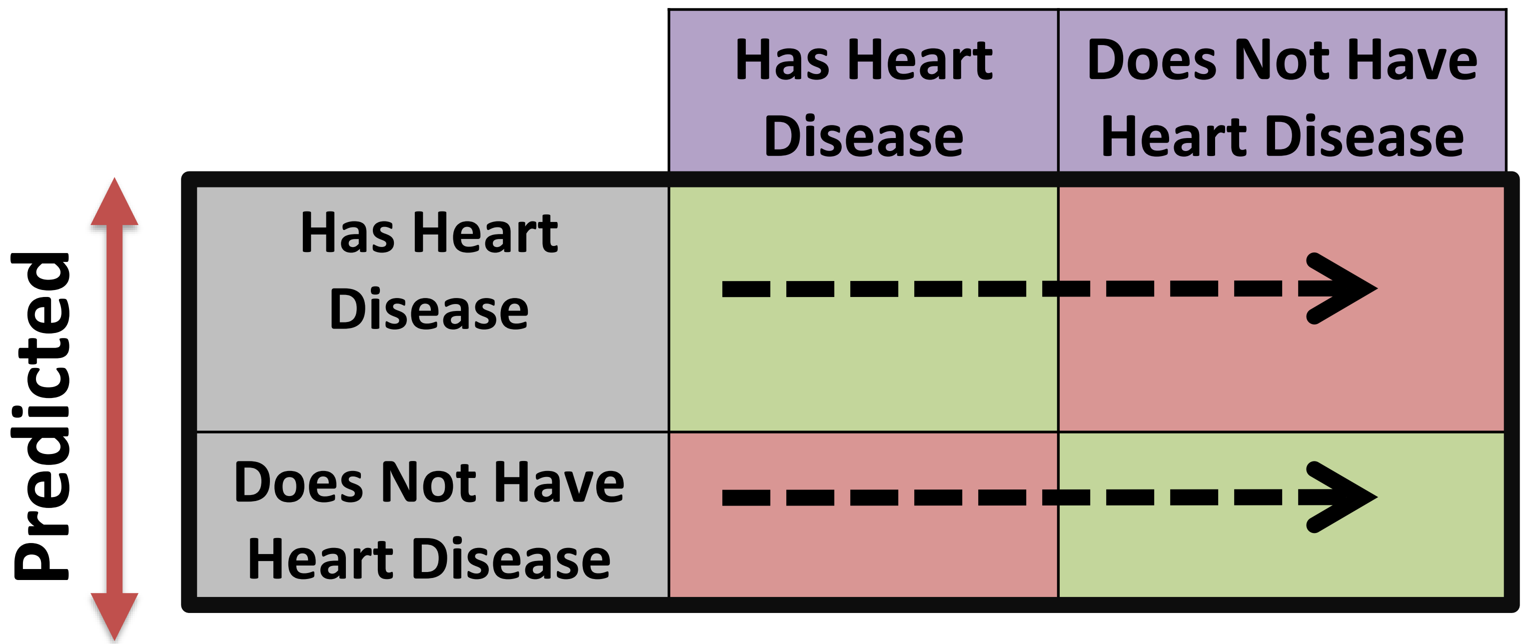
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Confusion Matrix

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease		
	Does Not Have Heart Disease		

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Confusion Matrix



The rows correspond to what the ML algorithm predicted

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Confusion Matrix

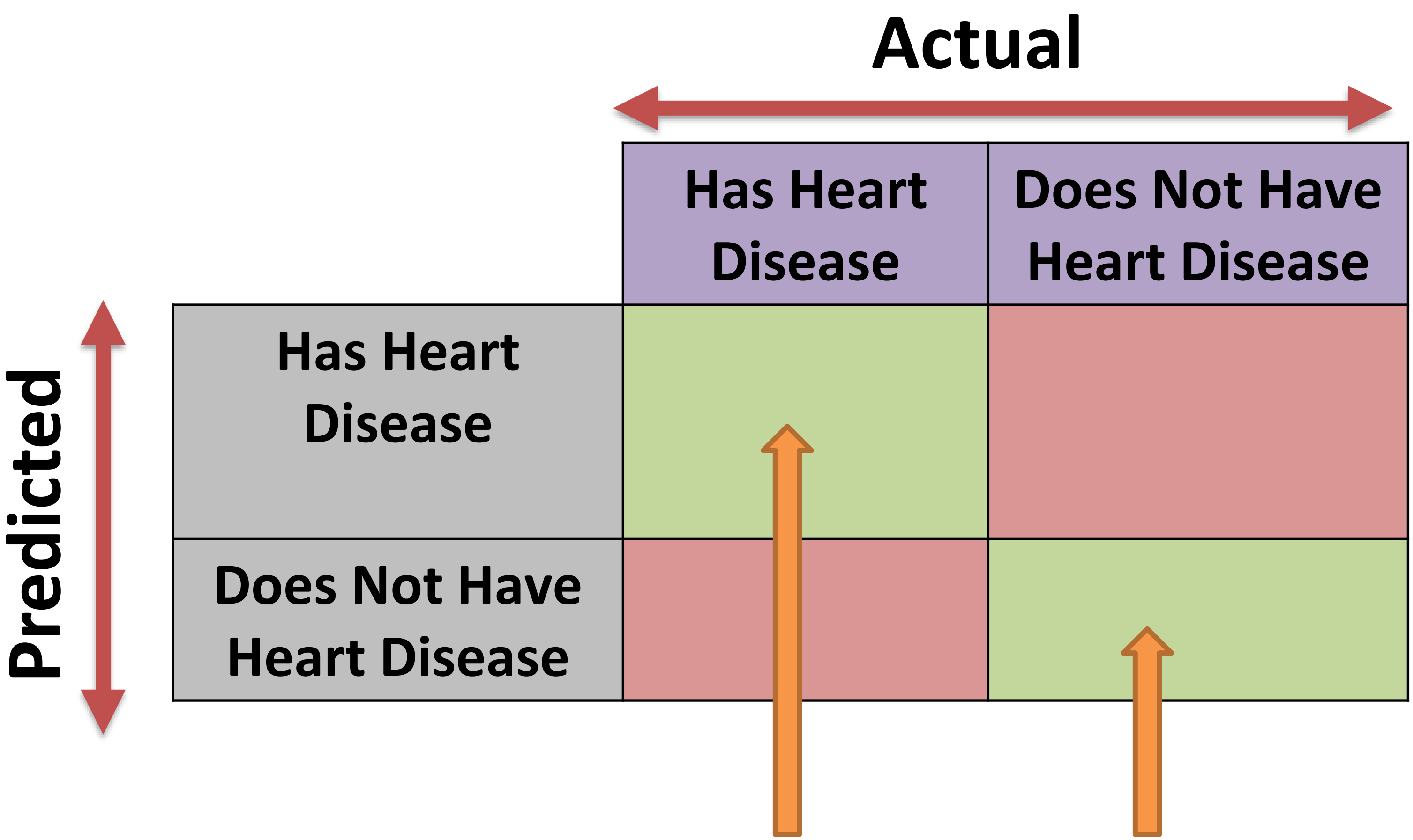
Actual

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease		
Does Not Have Heart Disease		

The columns correspond to the known truth

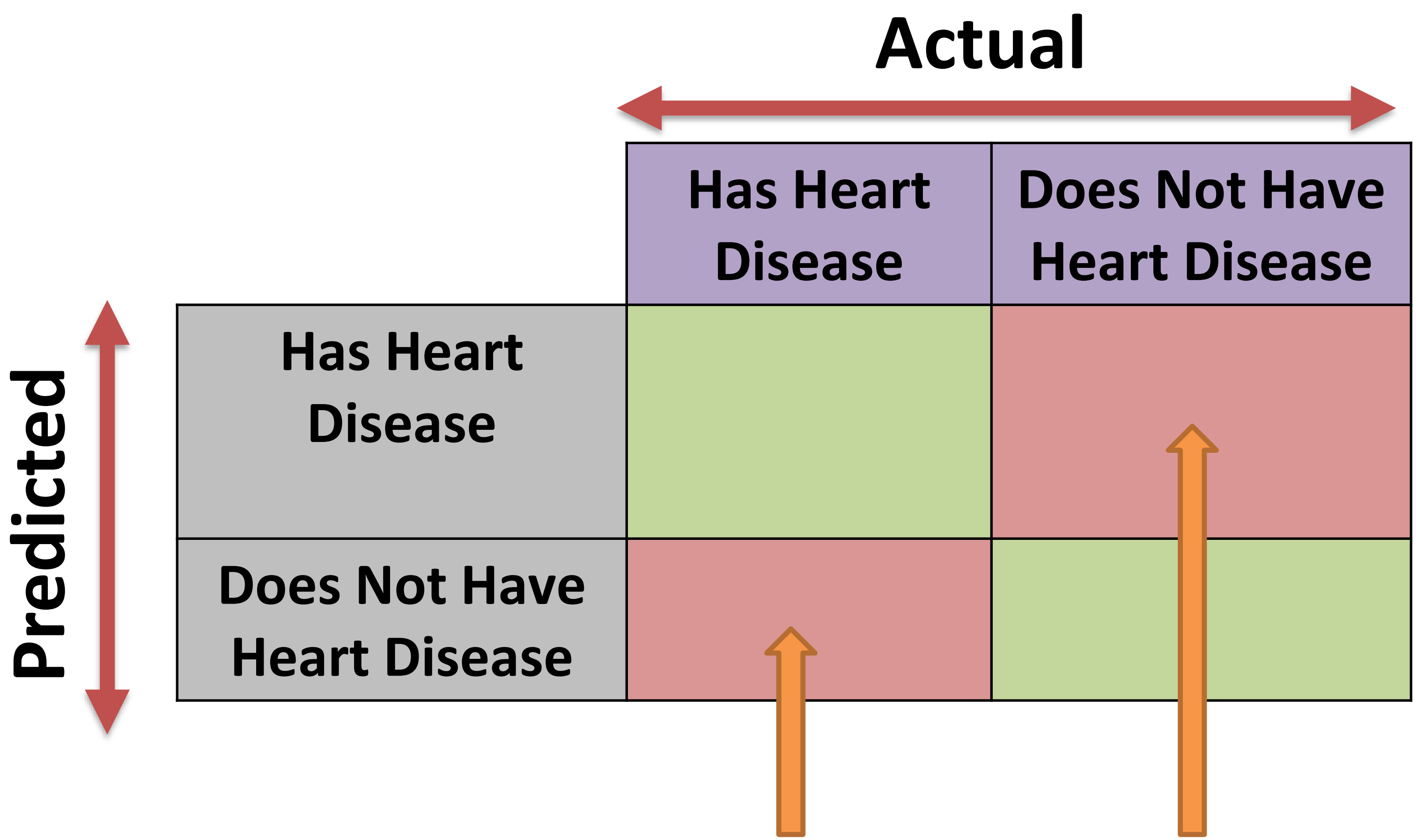
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Confusion Matrix



The Green Boxes tell us how many times the samples were correctly classified by the algorithm

Confusion Matrix



The Red Boxes tell us how many times the samples were misclassified by the algorithm

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Confusion Matrix

Actual

Predicted

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	True Positives	
Does Not Have Heart Disease		

Patients that had heart disease and that were correctly identified by the algorithm

Confusion Matrix

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	True Positives	
	Does Not Have Heart Disease		True Negatives

Patients that did not have heart disease and that were correctly identified by the algorithm

Confusion Matrix

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	True Positives	False Positives
	Does Not Have Heart Disease		True Negatives

Patients that did not have heart disease, but the algorithm says they do

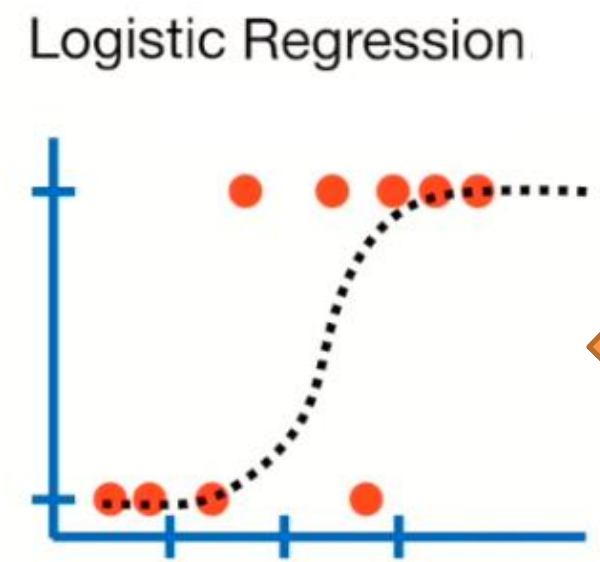
Confusion Matrix

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	True Positives	False Positives
	Does Not Have Heart Disease	False Negatives	True Negatives

Patients that had heart disease, but the algorithm says they didn't

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

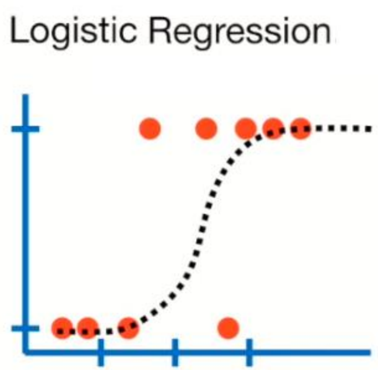
Confusion Matrix



		Actual	
		Has Heart Disease	**Does Not Have Heart Disease**
Predicted	**Has Heart Disease**	**140**	**15**
	Does Not Have Heart Disease	**18**	**127**

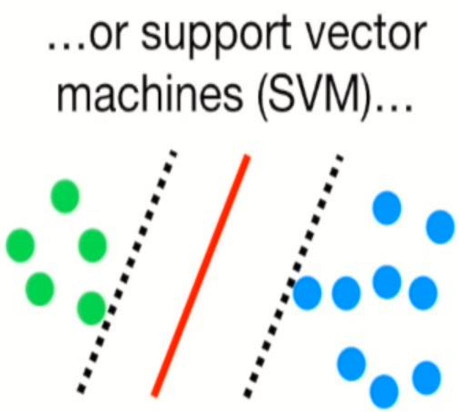
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Confusion Matrix



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	140	15
Does Not Have Heart Disease	18	127

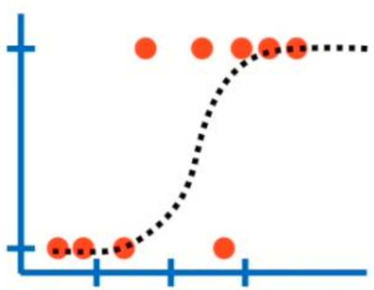
	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	151	8
Does Not Have Heart Disease	9	132



Confusion Matrix

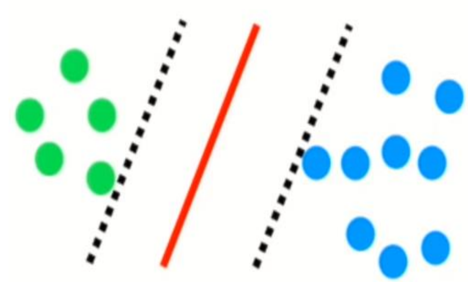
	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	140	15
Does Not Have Heart Disease	18	127

Logistic Regression



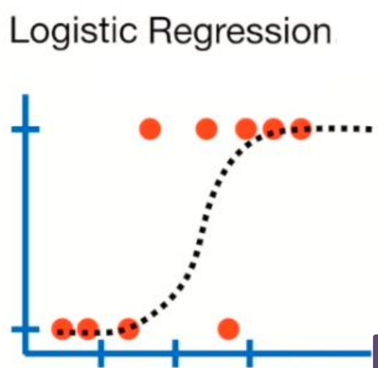
	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	151	8
Does Not Have Heart Disease	9	132

…or support vector machines (SVM)…



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

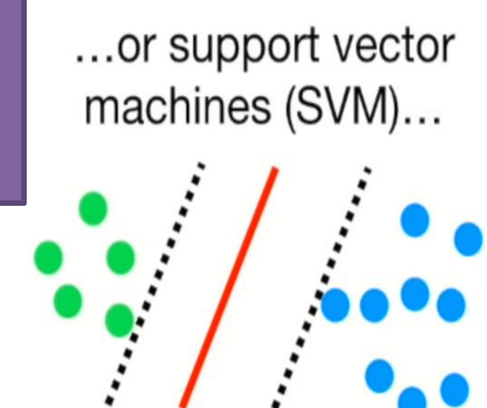
Confusion Matrix



	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	140	15
Does Not Have Heart Disease	[illegible]	127

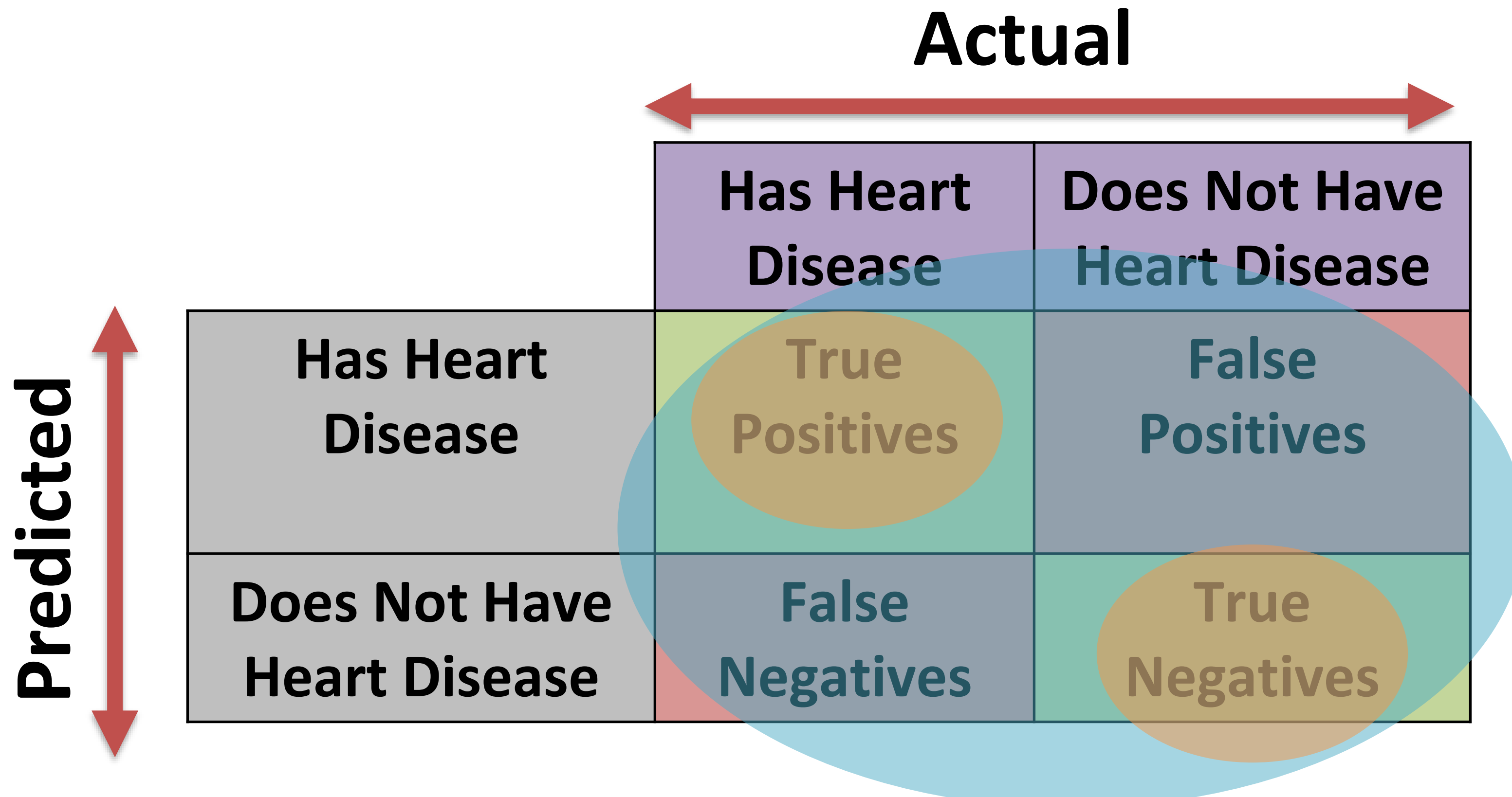
If we had to choose between Logistic regression and SVM, We would choose SVM !

Has Heart Disease	[illegible]	[illegible]
Does Not Have Heart Disease	9	132



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Accuracy is the proportion of the total number of predictions that are correct.



$$Accuracy = \frac{TP + TN}{TP + TN + FP + TN}$$

Accuracy

Predicted \ Actual	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	151	8
Does Not Have Heart Disease	9	132

$$\boldsymbol{Accuracy} = \frac{\mathbf{151 + 132}}{\mathbf{151 + 132 + 8 + 9}} \approx \mathbf{94.33\%}$$

Accuracy

Actual

Have
ease

When to use Accuracy:

Accuracy is a good measure when the target variable classes in the data are nearly balanced

$$\boldsymbol{Accuracy = \frac{151 + 132}{151 + 132 + 8 + 9} \approx 94.33\%}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Accuracy

Actual

Have
ease

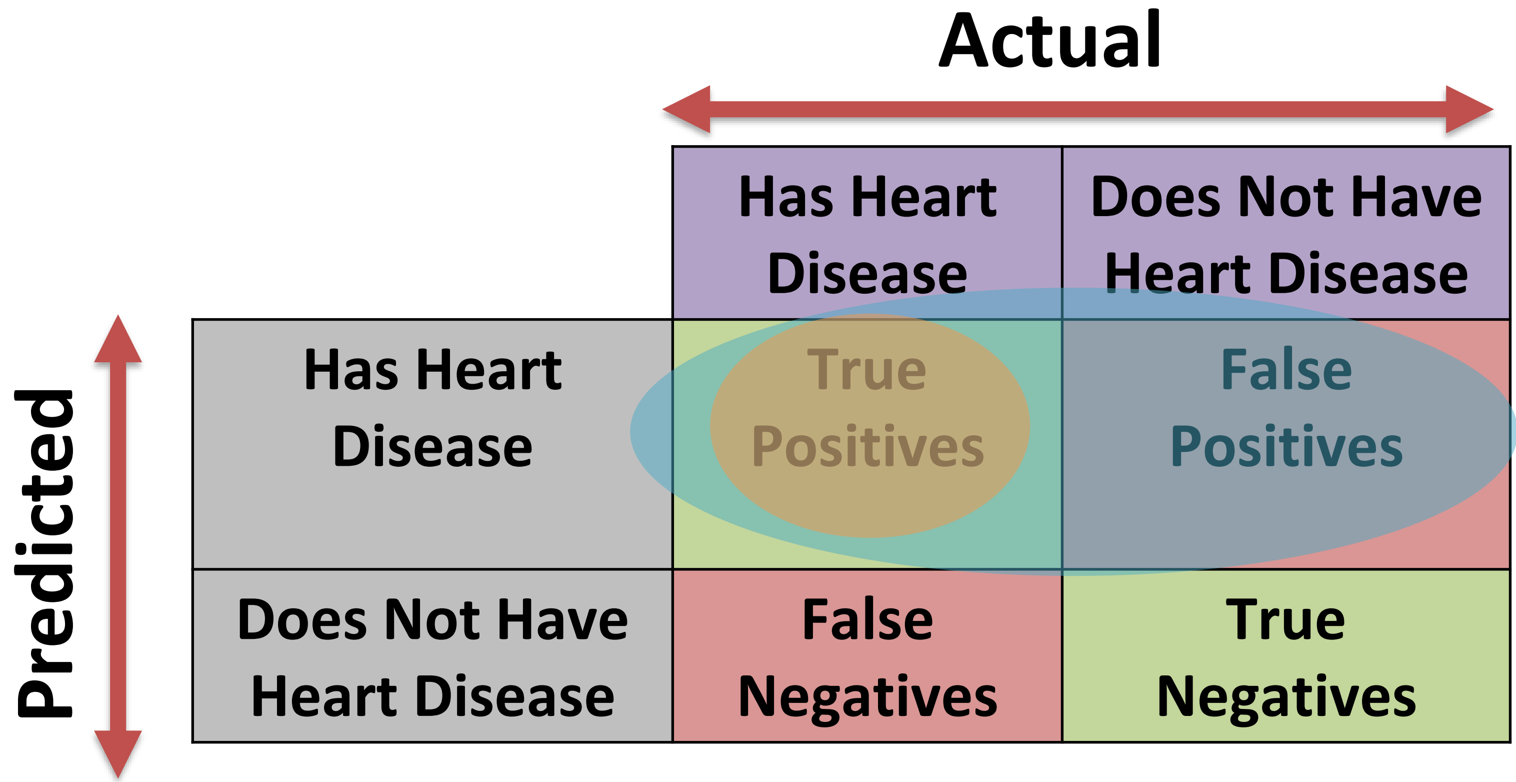
When NOT to use Accuracy:

Accuracy should NEVER be used as a measure when the target variable classes in data are a majority of 1 class

$$\boldsymbol{Accuracy = \frac{151 + 132}{151 + 132 + 8 + 9} \approx 94.33\%}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Precision shows correctness achieved in positive prediction



$$Precision = \frac{TP}{TP + FP}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Precision

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	151	8
	Does Not Have Heart Disease	9	132

$$\boldsymbol{Precision = \frac{151}{151+8} \approx 94.97\%}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Precision

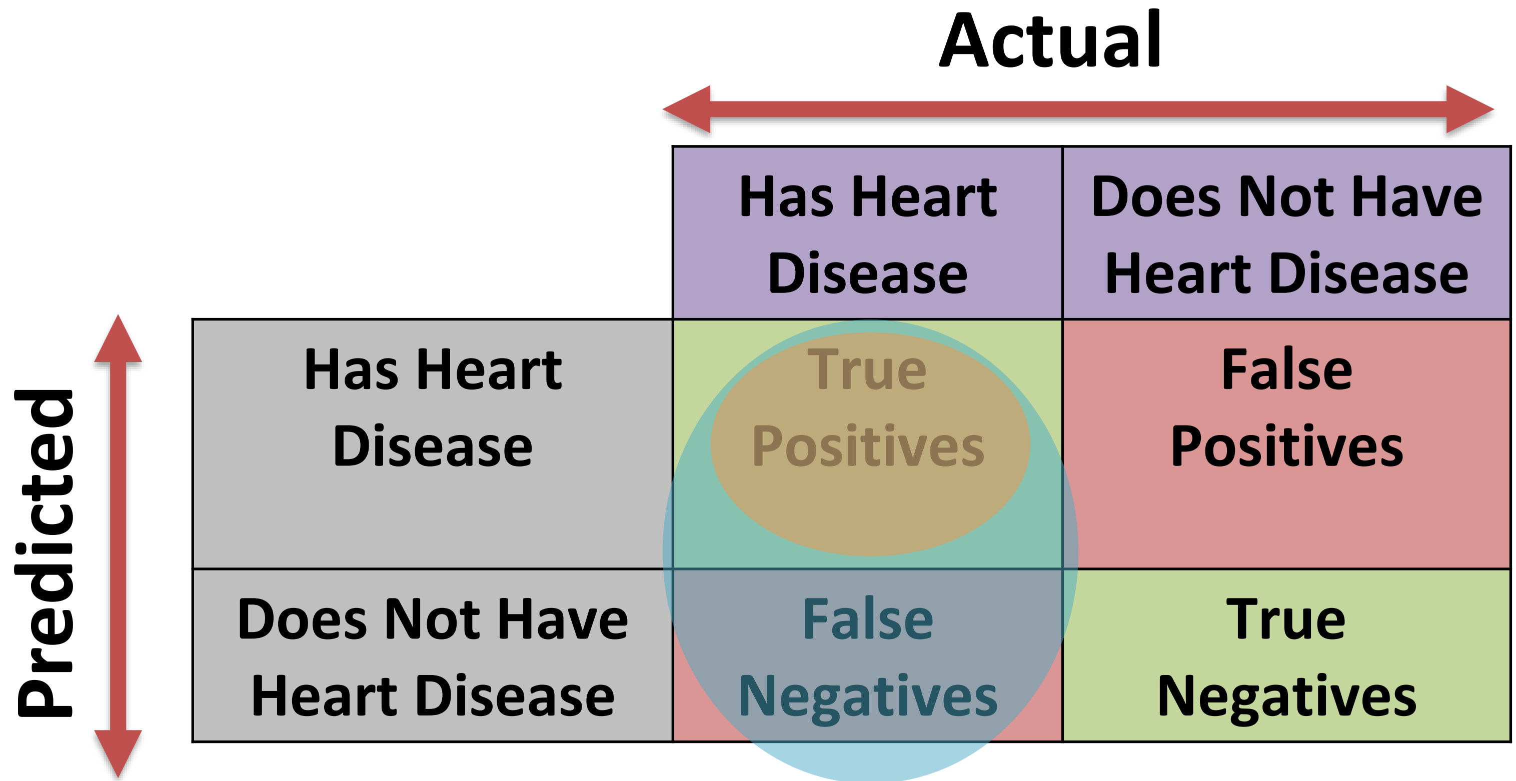
Actual

Have
ease

Precision is a metric for binary classifier which measures the correctness among all positive labels

$$\boldsymbol{Precision = \frac{151}{151+8} \approx 94.97\%}$$

Recall (Sensitivity) It is measure of positive examples labeled as positive by classifier



$$Recall = \frac{TP}{TP + FN}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Recall (Sensitivity)

Predicted \ Actual	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	151	8
Does Not Have Heart Disease	9	132

$$\boldsymbol{Recall = \frac{151}{151 + 9} \approx 94.38\%}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Recall (Sensitivity)

Actual

Have
ease

Recall is a metric for binary classifier which measures how many positive labels are successfully predicted amongst all positive labels

$$\boldsymbol{Recall = \frac{151}{151 + 9} \approx 94.38\%}$$

F1 Score is a weighted average of the recall and precision

- F1 score might be good choice when you seek to balance between Precision and Recall.

F1 Score is a weighted average of the recall and precision

- F1 score might be good choice when you seek to balance between Precision and Recall.

$$\boldsymbol{F1\,Score} = \mathbf{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

F1 Score is a weighted average of the recall and precision

- F1 score might be good choice when you seek to balance between Precision and Recall.

$$\boldsymbol{F1\ Score} = 2 \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

- It helps to compute recall and precision in one equation so that the problem to distinguish the models with low recall and high precision or vice versa could be solved

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

F1 Score

$$\boldsymbol{Precision = \frac{151}{151 + 8} \approx 94.97\%}$$

$$\boldsymbol{Recall = \frac{151}{151 + 9} \approx 94.38\%}$$

$$\boldsymbol{F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}}$$

$$\boldsymbol{= 2 \times \frac{94.97 \times 94.38}{94.97 + 94.38} = 94.97\%}$$

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Key Elements of Machine Learning

Evaluation

Performance Metrics for Regression problems

1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. R Squared (R^2)
5. Adjusted R Squared (R^2)

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Key Elements of Machine Learning

Optimization

This is the way candidate programs are generated, also known as the **search process**

1. Combinatorial optimization
2. Convex optimization
3. Constrained optimization

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Starting with some data …

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

We want to use some variables such as Chest pain, Good blood circulation, Blocked arteries and the Weight

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

To predict if someone has heart disease or not

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

When a new patient shows up

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	No	No	115	

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

And predict if that person have heart disease or not

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	No	No	115	**Yes/ No ?**

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

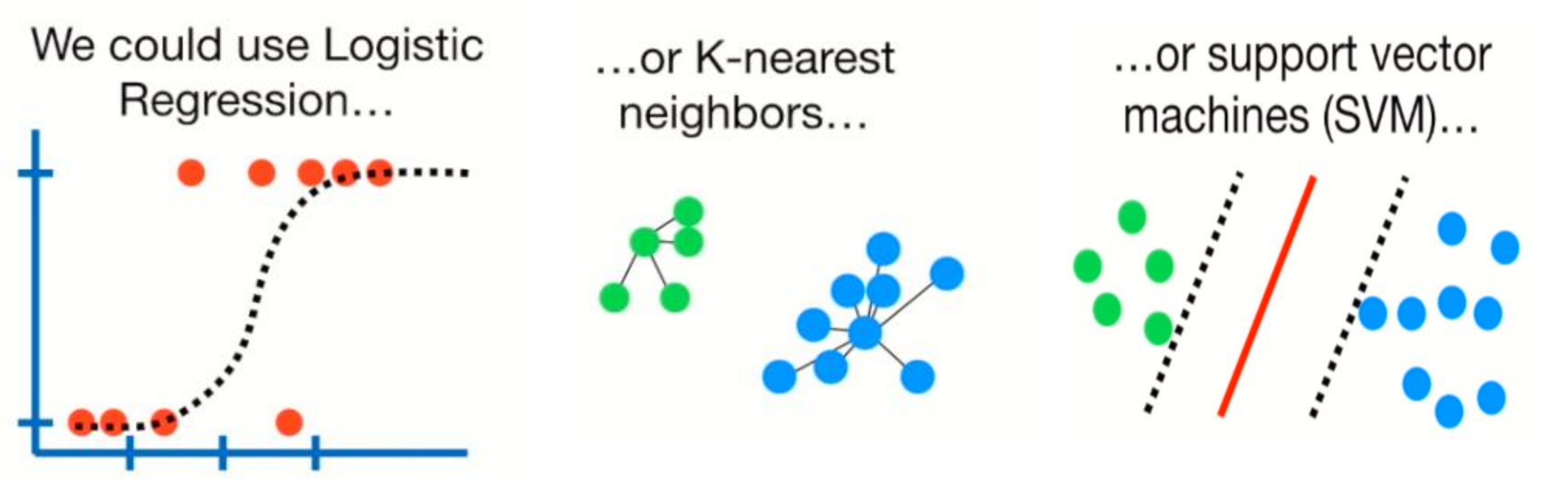
Compare Machine Learning Methods

Chest Pain	Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	80	No
Yes	Yes	Yes	125	Yes
Yes	Yes	No	140	No
…	…	…	…	…

However, first we have to decide which Machine Learning method would be best for our actual problem …

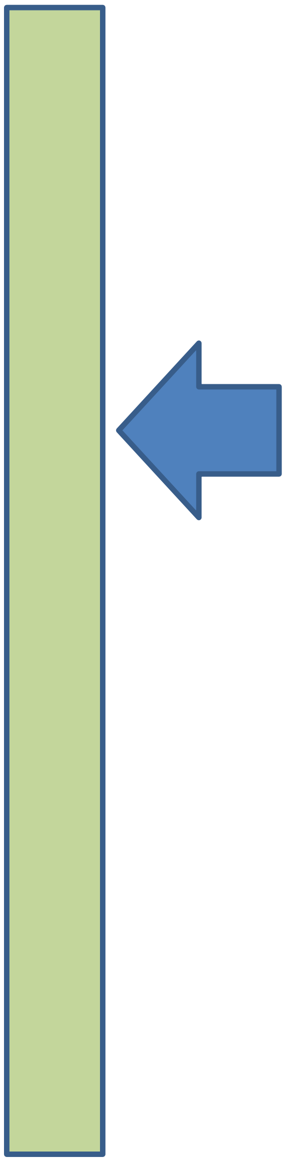
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Compare Machine Learning Methods



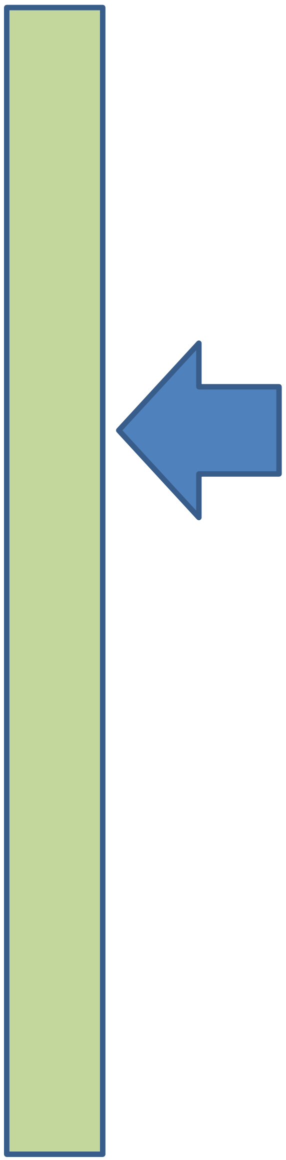
Cross validation allows us to compare different machine learning methods and get a sense of how they will work in practice

Cross Validation



Let's say that this Green Rectangle represents all the collected data about persons with/without heart disease

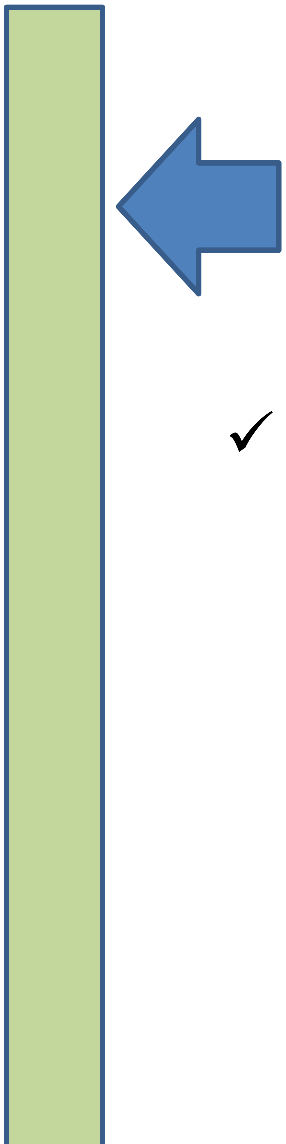
Cross Validation



We need to do two things with the data !

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

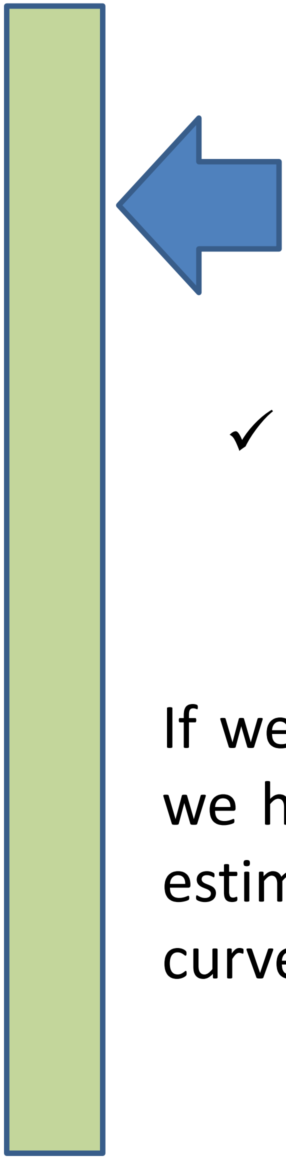


We need to do two things with the data !

✓ **Estimate the parameters for the Machine Learning methods**

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

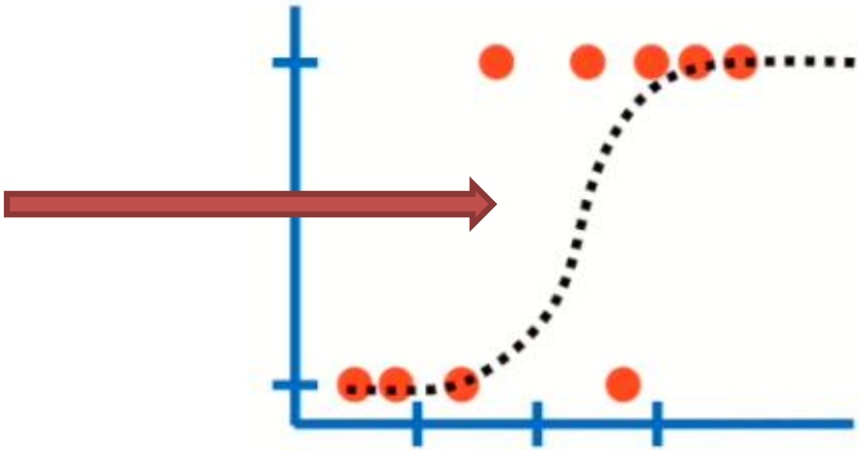
Cross Validation



We need to do two things with the data !

✓ **Estimate the parameters for the Machine Learning methods**

If we take Logistic regression, we have to use some data to estimate the shape of this curve

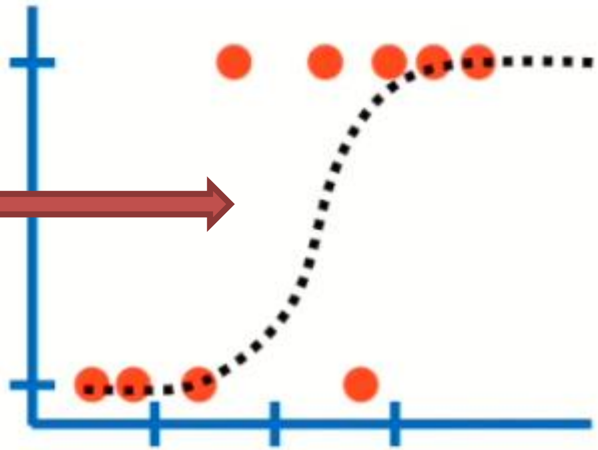


Cross Validation

We need to do two things with the data !

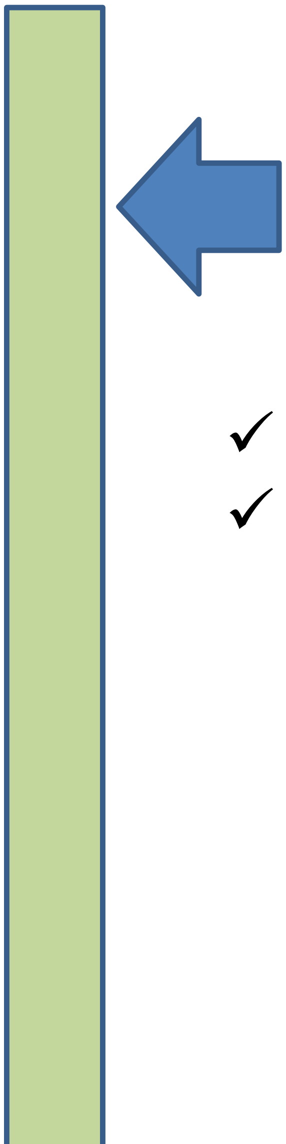
✓ **Estimate the parameters for the Machine Learning methods**

In Machine Learning lingo, estimating parameters is called **training the model**



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation



We need to do two things with the data !

- ✓ **Estimate the parameters for the Machine Learning methods**
- ✓ **Evaluate how well Machine Learning methods work**

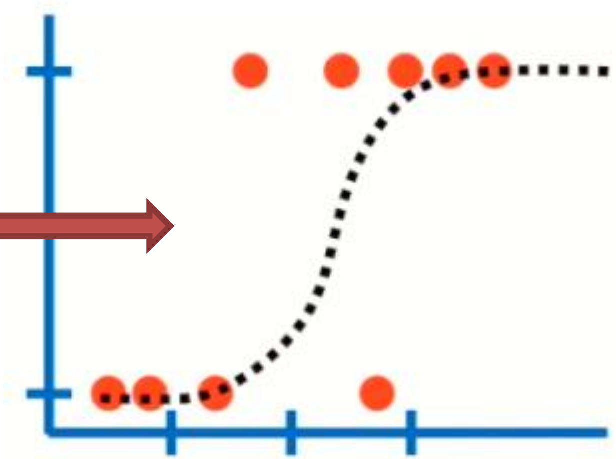
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

We need to do two things with the data !

- ✓ **Estimate the parameters for the Machine Learning methods**
- ✓ **Evaluate how well Machine Learning methods work**

Does this curve do a good job while categorizing new data ?



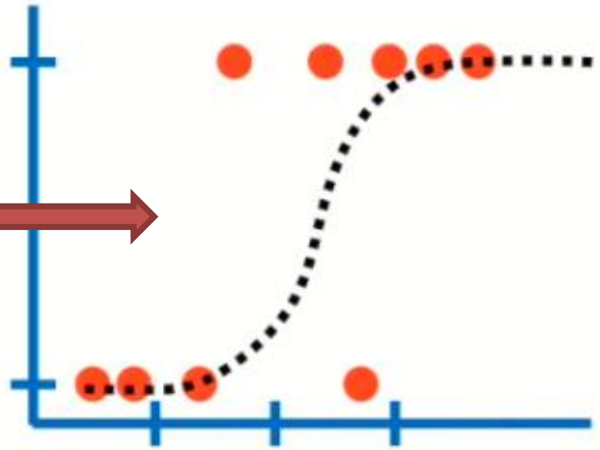
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

We need to do two things with the data !

- ✓ **Estimate the parameters for the Machine Learning methods**
- ✓ **Evaluate how well Machine Learning methods work**

In Machine Learning lingo, evaluating a method is called **testing the model**



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

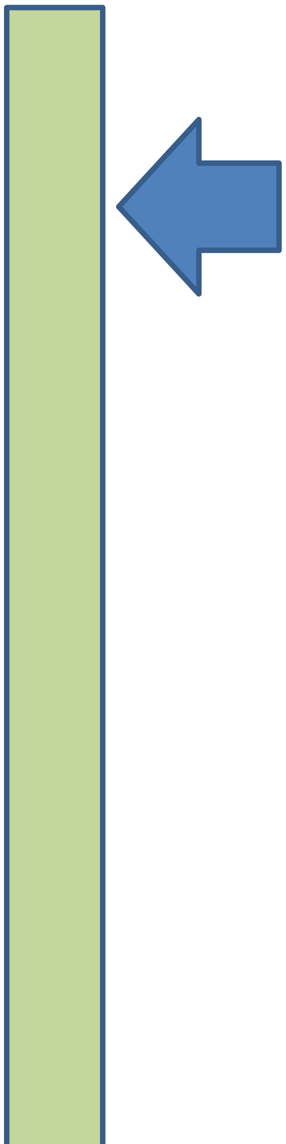
Cross Validation

By using the Machine Learning lingo we will:

- ~~Estimate the parameters for~~ **Train** the Machine Learning methods
- ~~Evaluate how well~~ **Test** the Machine Learning methods

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation



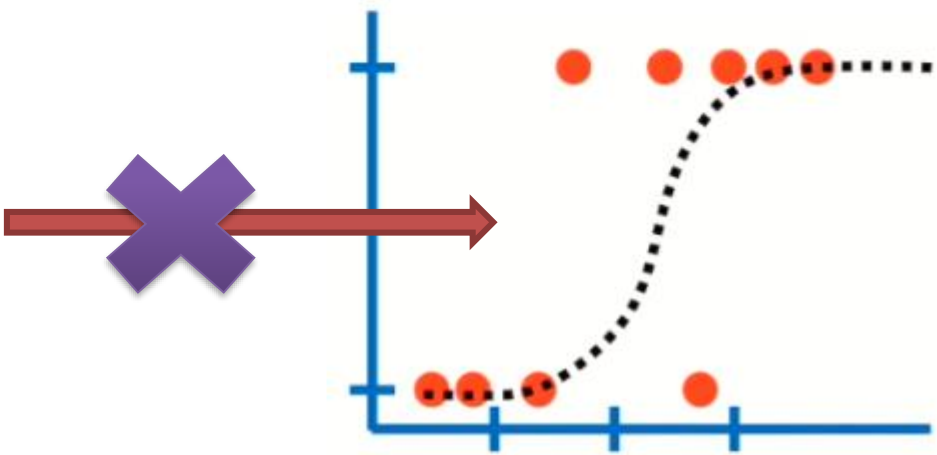
A terrible approach would be to use the entire data to train the algorithm

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

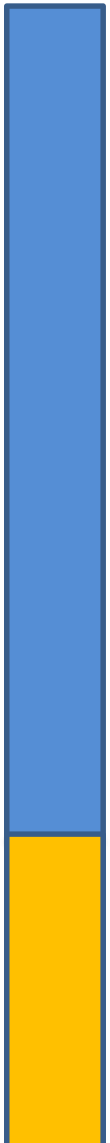
A terrible approach would be to use the entire data to train the algorithm

Reusing the same data for both training and testing is a **bad idea !**

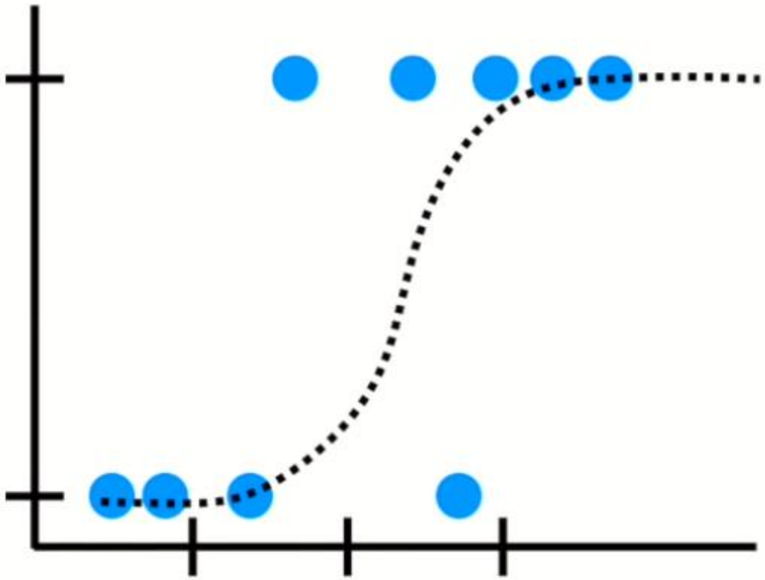


Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

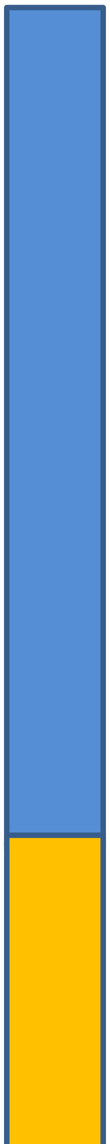


A better approach would be to use the first 75% of the data for training phase

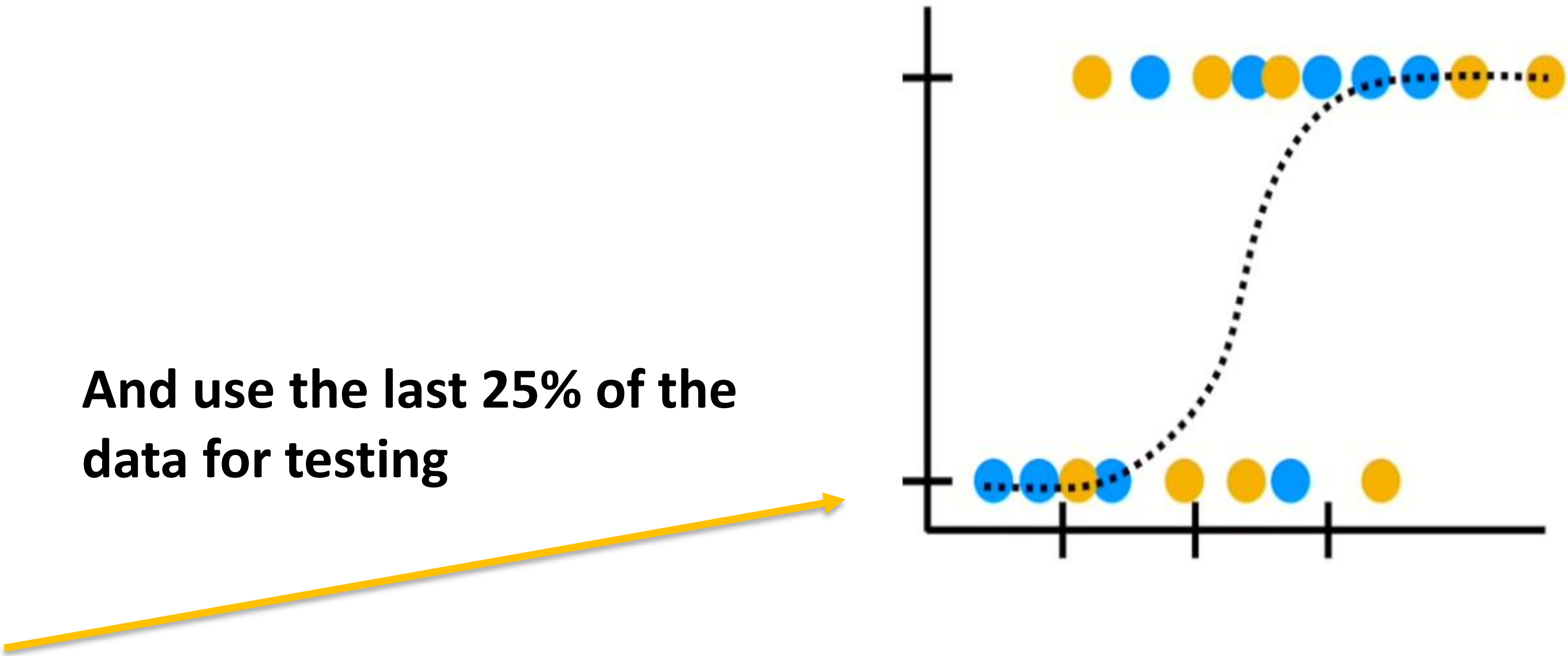


Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

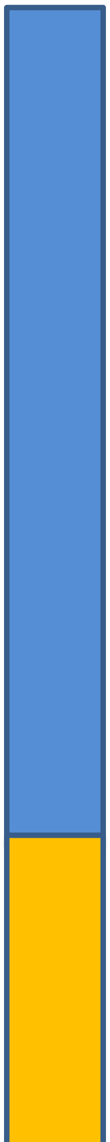


And use the last 25% of the data for testing

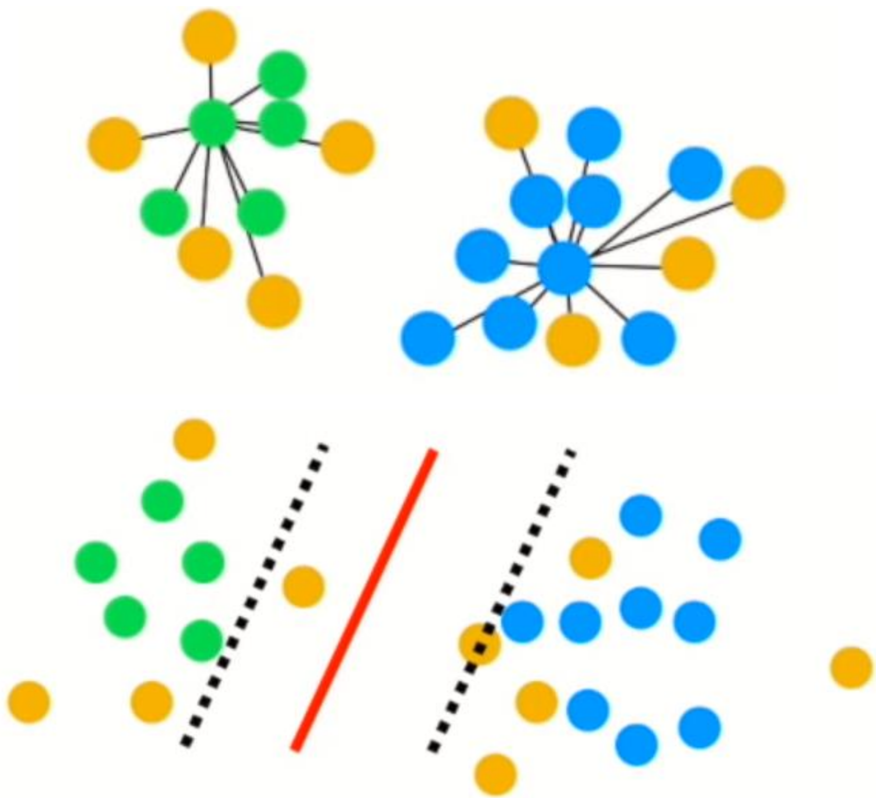


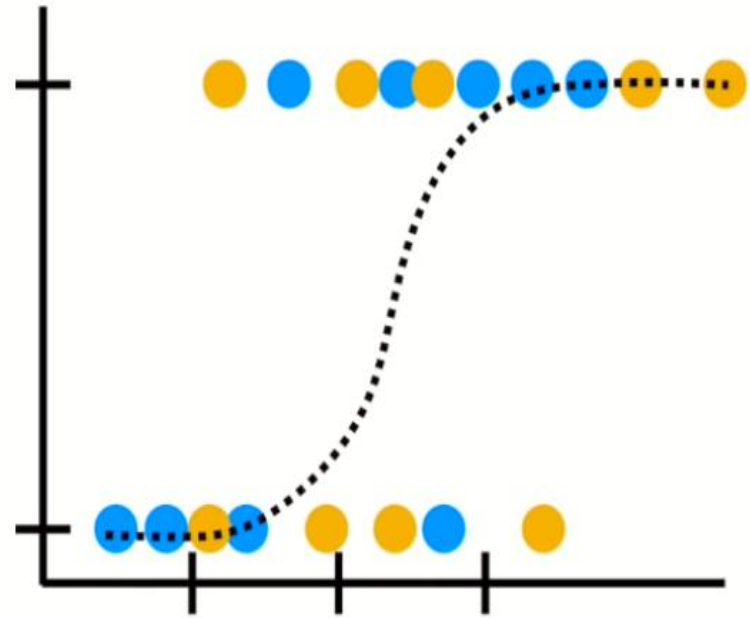
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

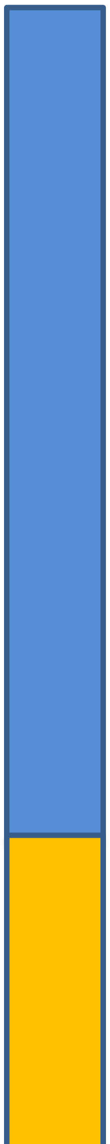


We can now compare methods by seeing how well each one categorized the test data





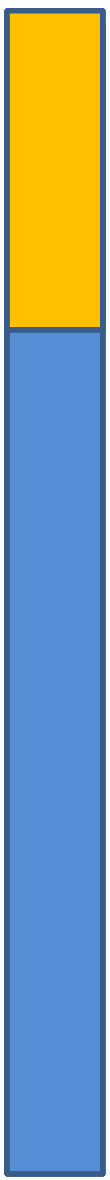
Cross Validation



How do we know that using the first 75% as training data and the last 25% as testing data is the best way to divide the data ?

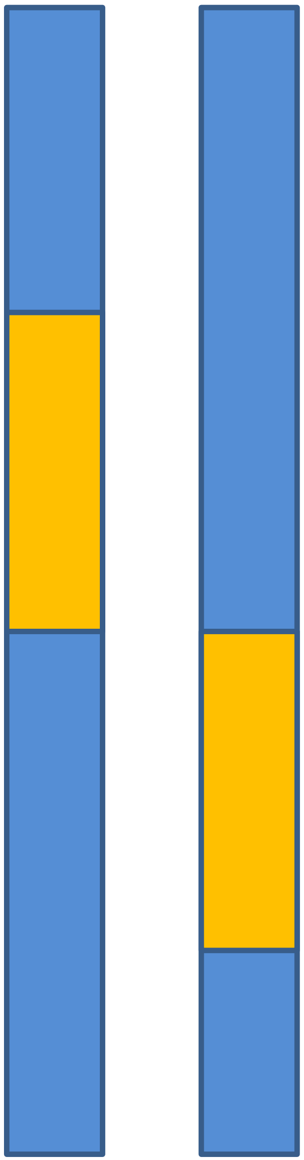
Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation



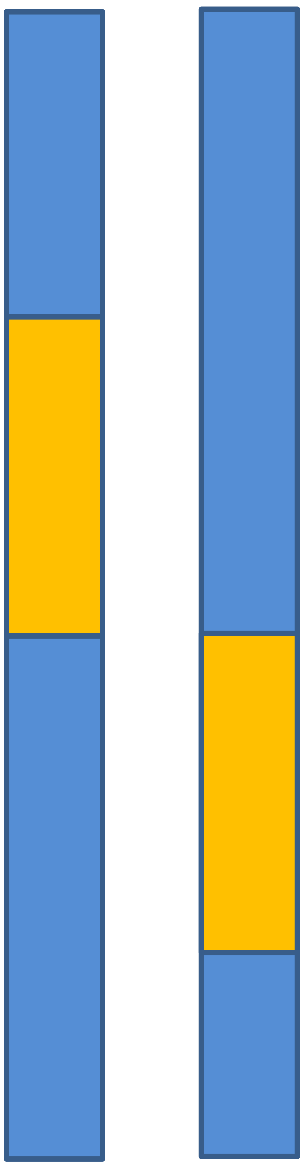
What if we inverse the order ?

Cross Validation



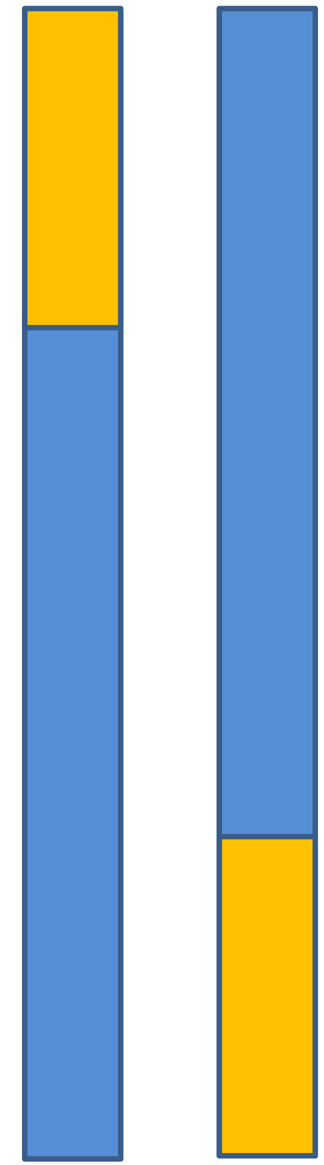
What if we take one of the middle rectangles ?

Cross Validation



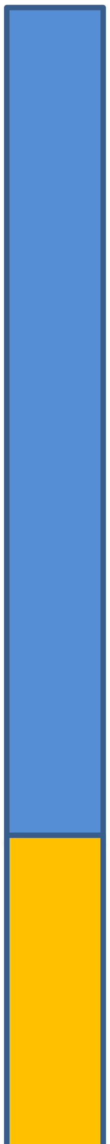
Don’t worry **Cross Validation** comes the rescue !

It uses all the possible splits, and summarizes the results at the end

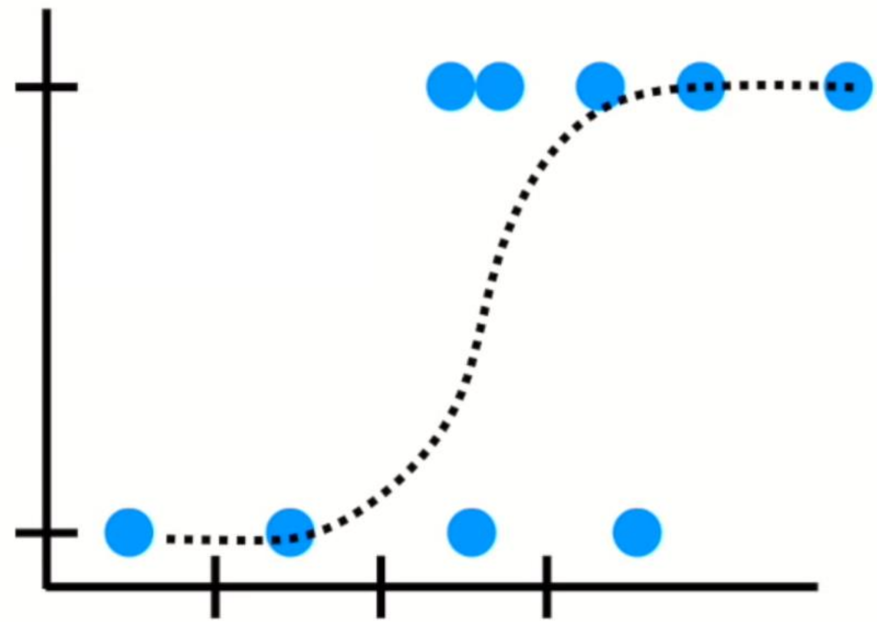


Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

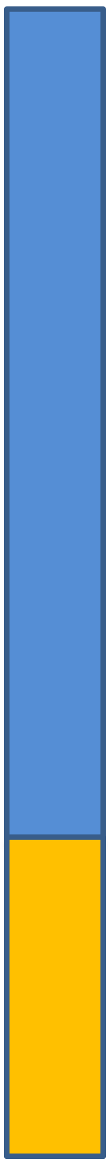


Start with the first 3 block to train the method

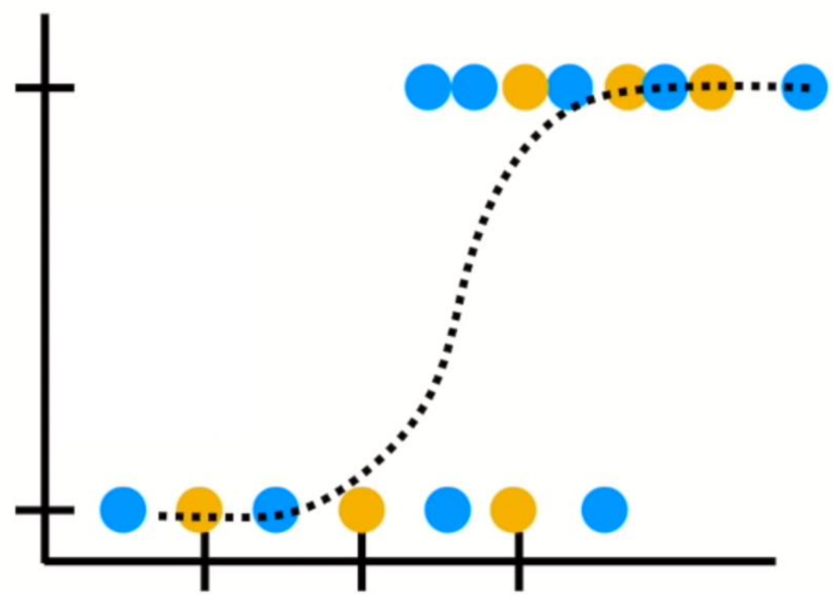


Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

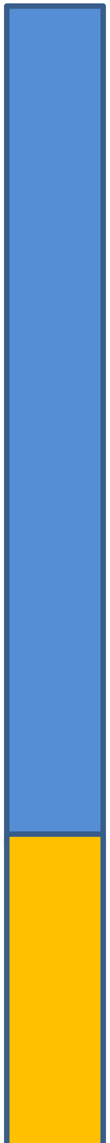


And use the last block to test the method



Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

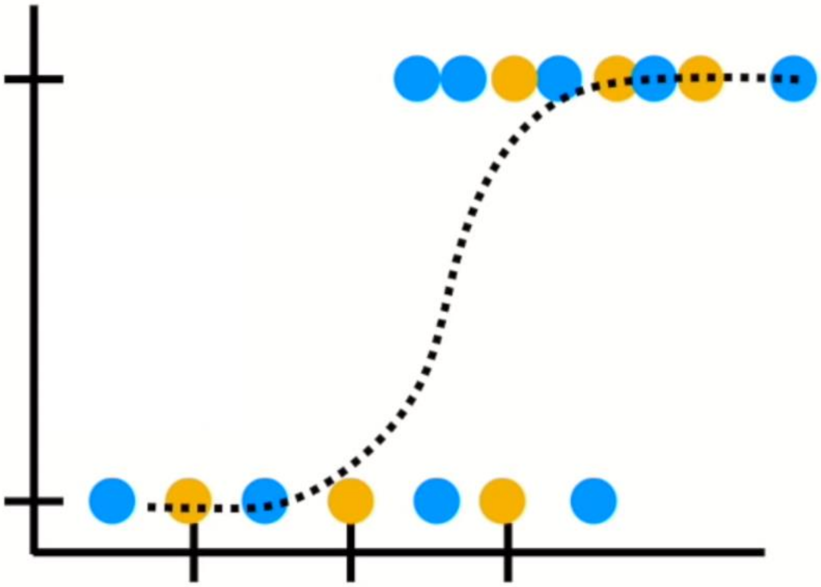
Cross Validation



Test Data categorization

Correct 5

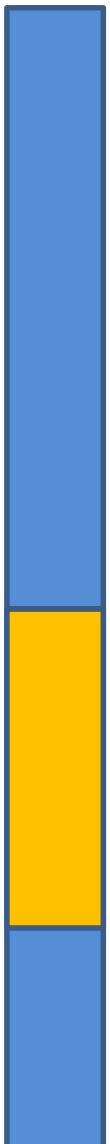
Incorrect 1



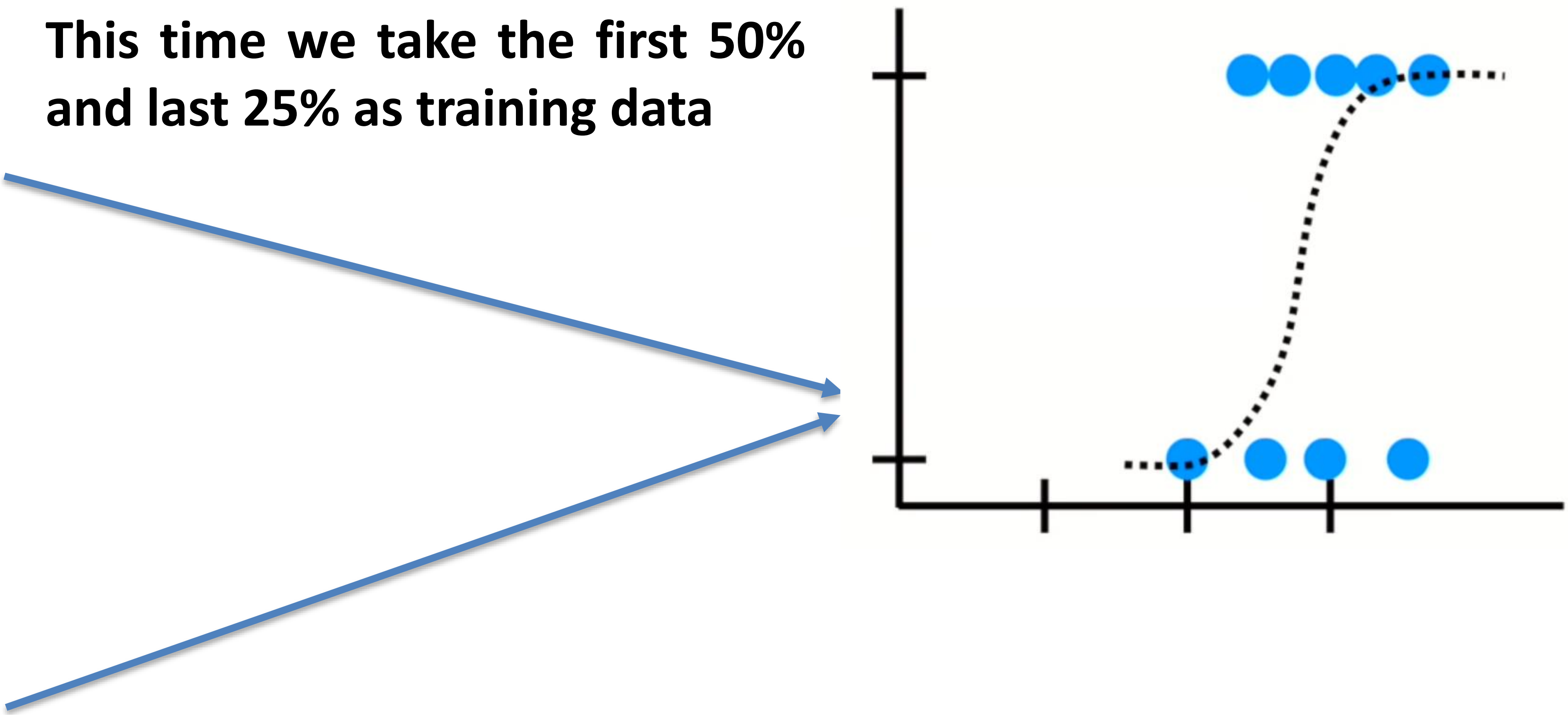
Cross Validation records how well the method did with the selected test data

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

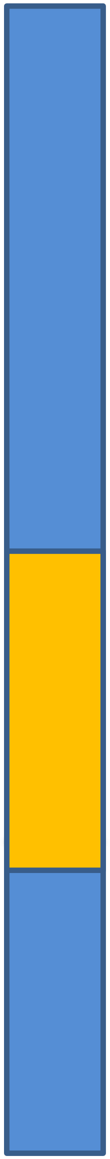


This time we take the first 50% and last 25% as training data

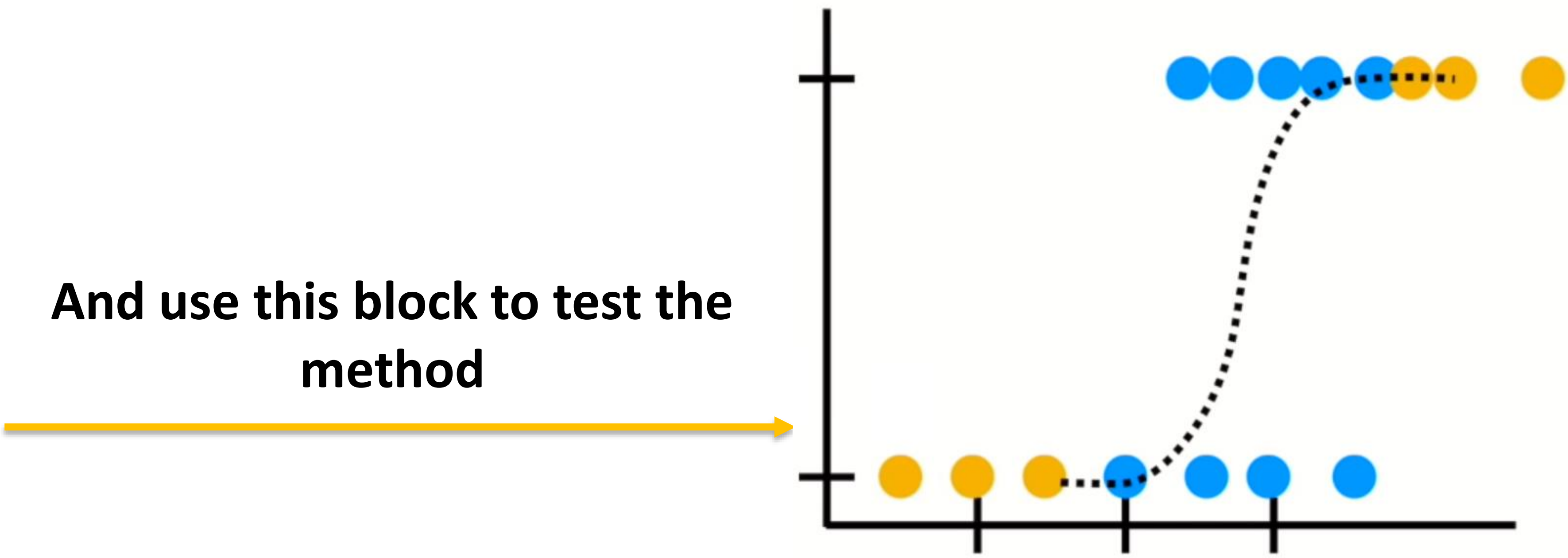


Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

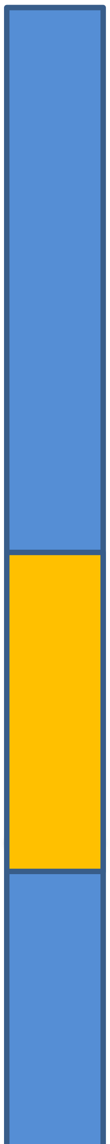
Cross Validation



And use this block to test the method



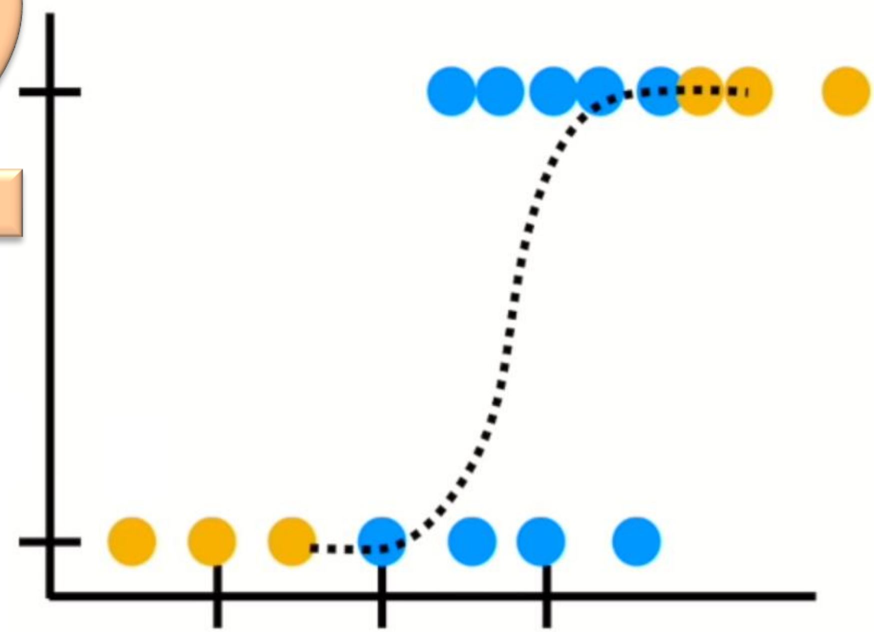
Cross Validation



Test Data categorization

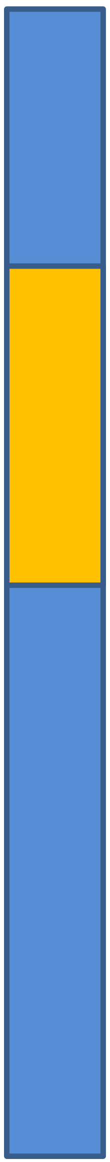
Correct **4**

Incorrect **2**



Cross Validation records how well the method did with the selected test data

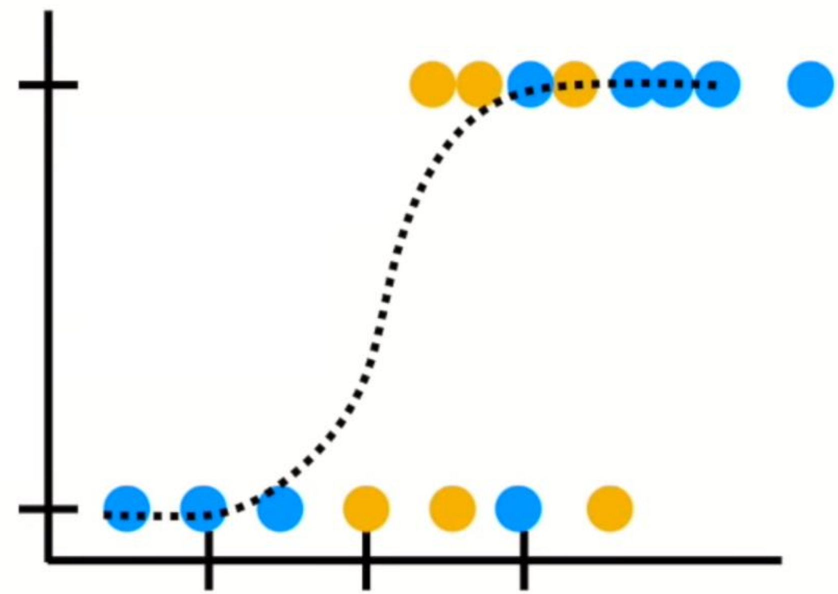
Cross Validation



Test Data categorization

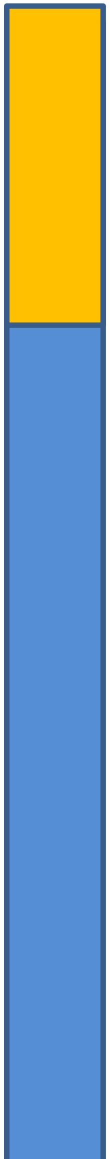
Correct **1**

Incorrect **5**



Cross Validation records how well the method did with the selected test data

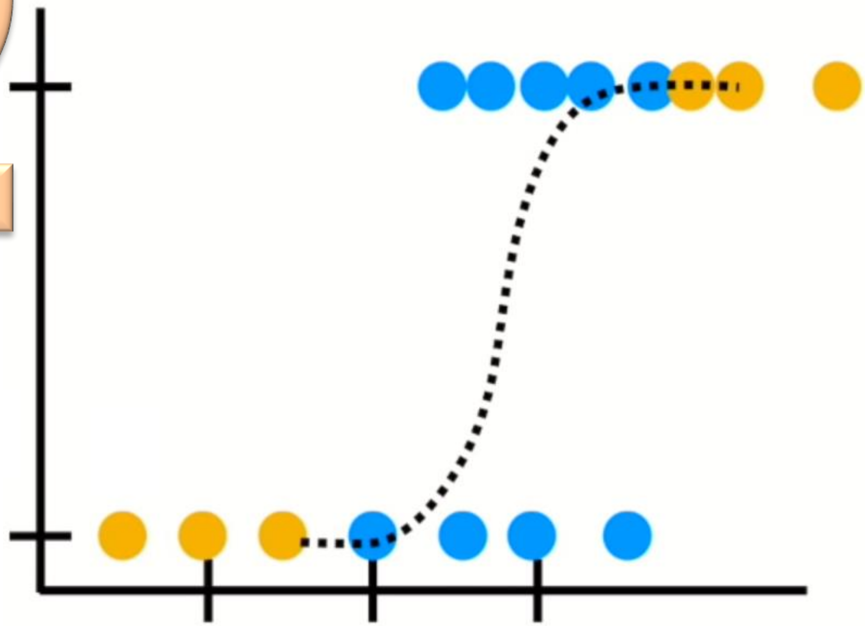
Cross Validation



Test Data categorization

Correct
6

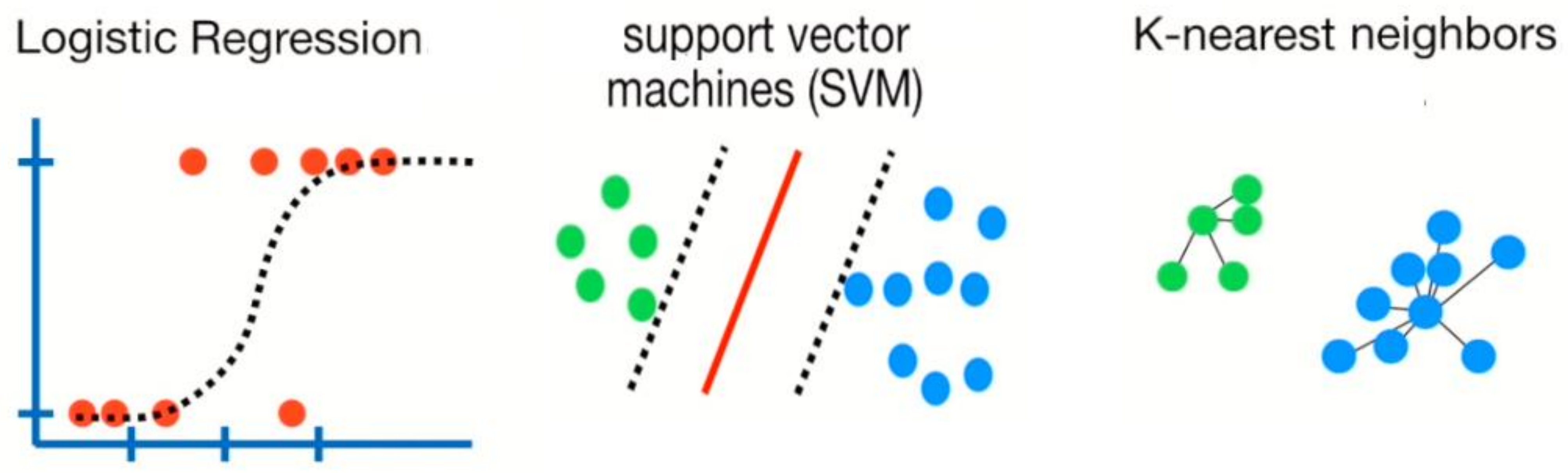
Incorrect
0



Cross Validation records how well the method did with the selected test data

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation

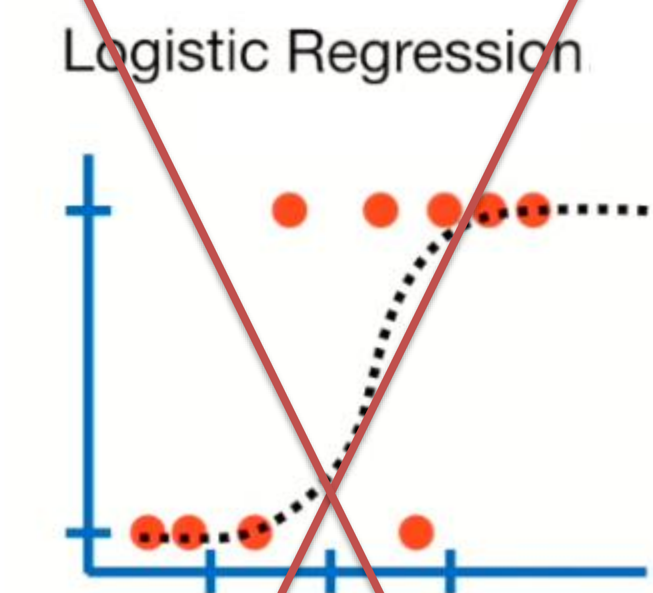


Correct	Incorrect
16	8

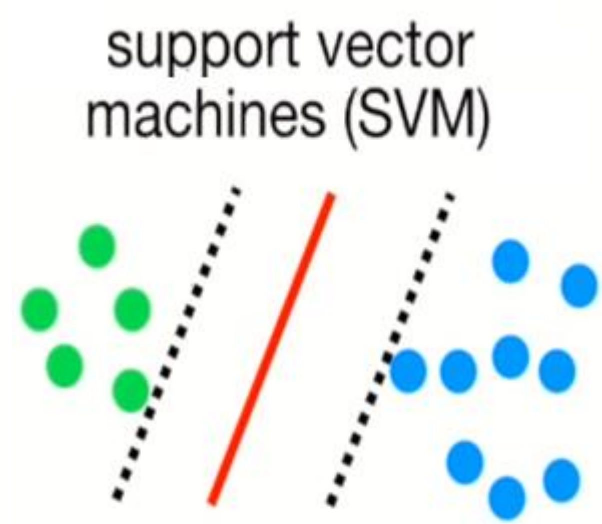
Correct	Incorrect
18	6

Correct	Incorrect
10	12

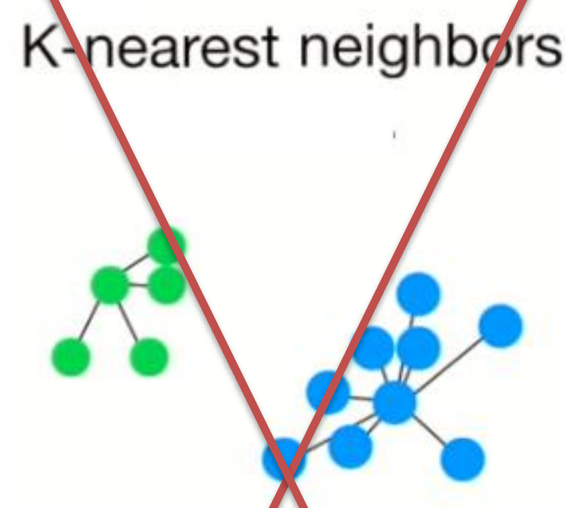
Cross Validation



Correct 16 | Incorrect 8

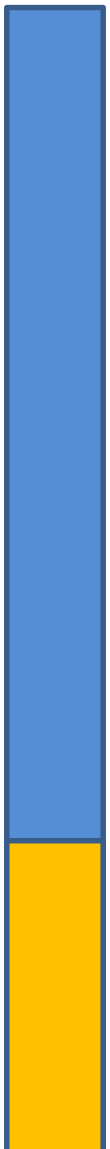


Correct 18 | Incorrect 6



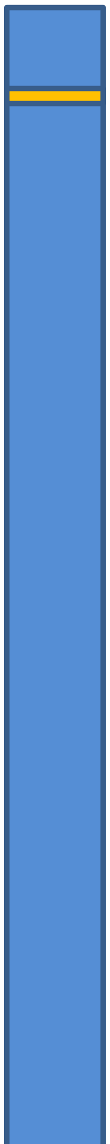
Correct 10 | Incorrect 12

Cross Validation



For this example we divided the data the data into 4 blocks. This is called Four-Fold Cross Validation

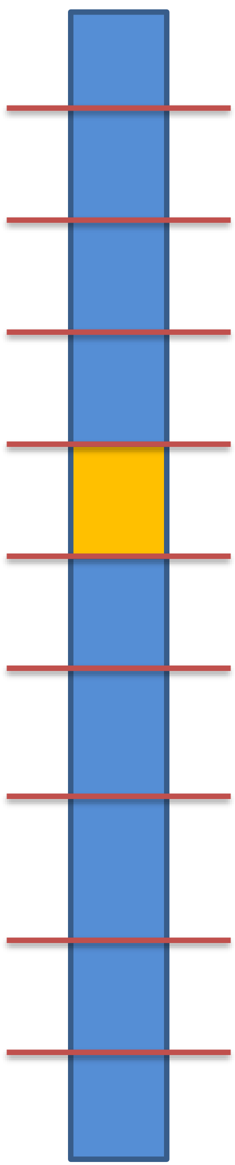
Cross Validation



In an extreme special cases we could consider each sample as a block. This is called Leave One Out Cross Validation

Zakarya Farou; Krisztian Buza, buza@inf.elte.hu
http://biointelligence.hu

Cross Validation



It is very common to divide the data into 10 blocks. This is called Ten-Fold Cross Validation