LEARNING TO CODE

python™

Coordinator: Konstantinos Emmanouilidis

IEEE SB of thrace

2017-2018

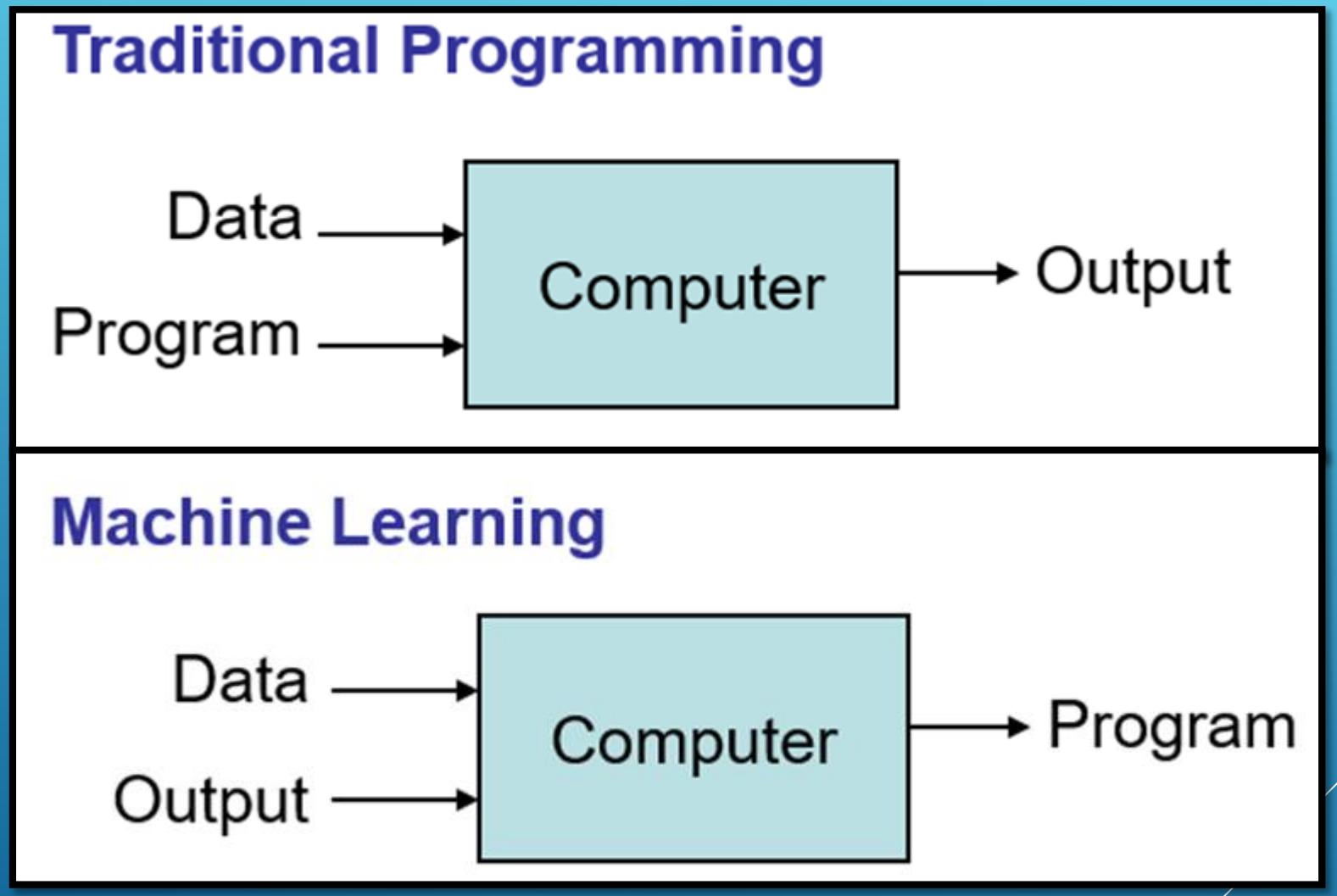
Introduction to Machine Learning

Using Python

What is Machine Learning??

Definition

A computer program is said to ‘learn’ from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks T, measured by P, improves with experience E.


Traditional Programming
Data
Program
Computer
Output
Machine Learning
Data
Output
Computer
Program

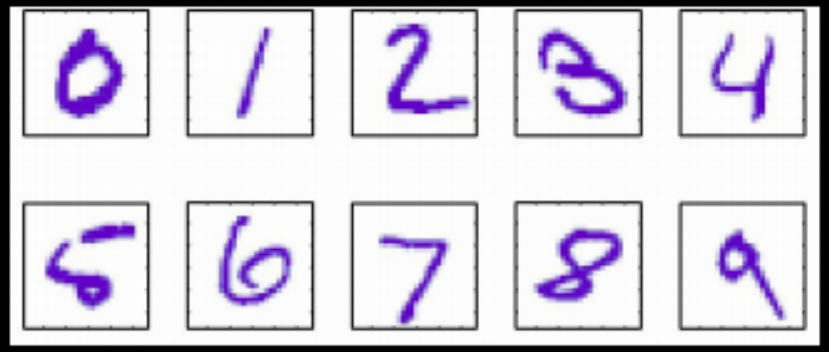
Magic?

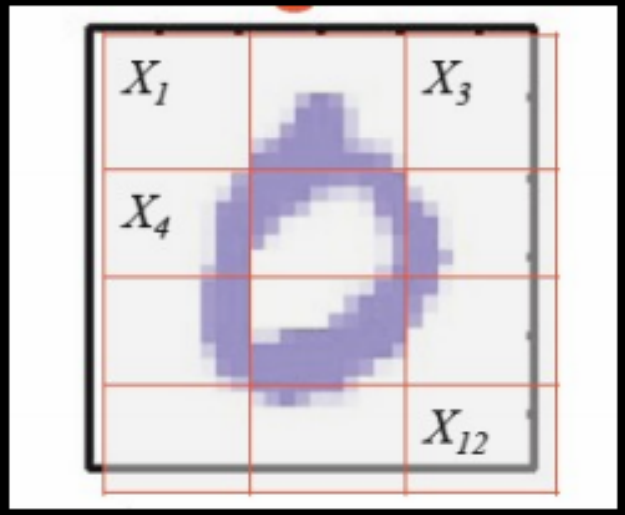
No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

Example:

Classification: Digit Recognition





Input (X_i) : Image Features

Output (Y): Class Labels { $y^0, y^1, .. y^9$ }

Features(X_i):

Proportion of pixels in Each of the 12 cells

X_i *where* $i = 1,2,\ldots, 12$

x_i^0=0-10%
x_i^1=10-20%
....

Val(Xi)=10

No of parameters= 10^{12} - 1

Handcrafted Rules will result in a large number of rules and exceptions

– We need ML in cases where we cannot directly write a program to handle every case

So it's better to have a machine that *learns* from a large training set

So, according to the definition earlier:

Task (T): recognizing and classifying handwritten words within images

Performance measure (*P*): percent of words correctly classified

Training experience (E): a database of handwritten words with given classifications

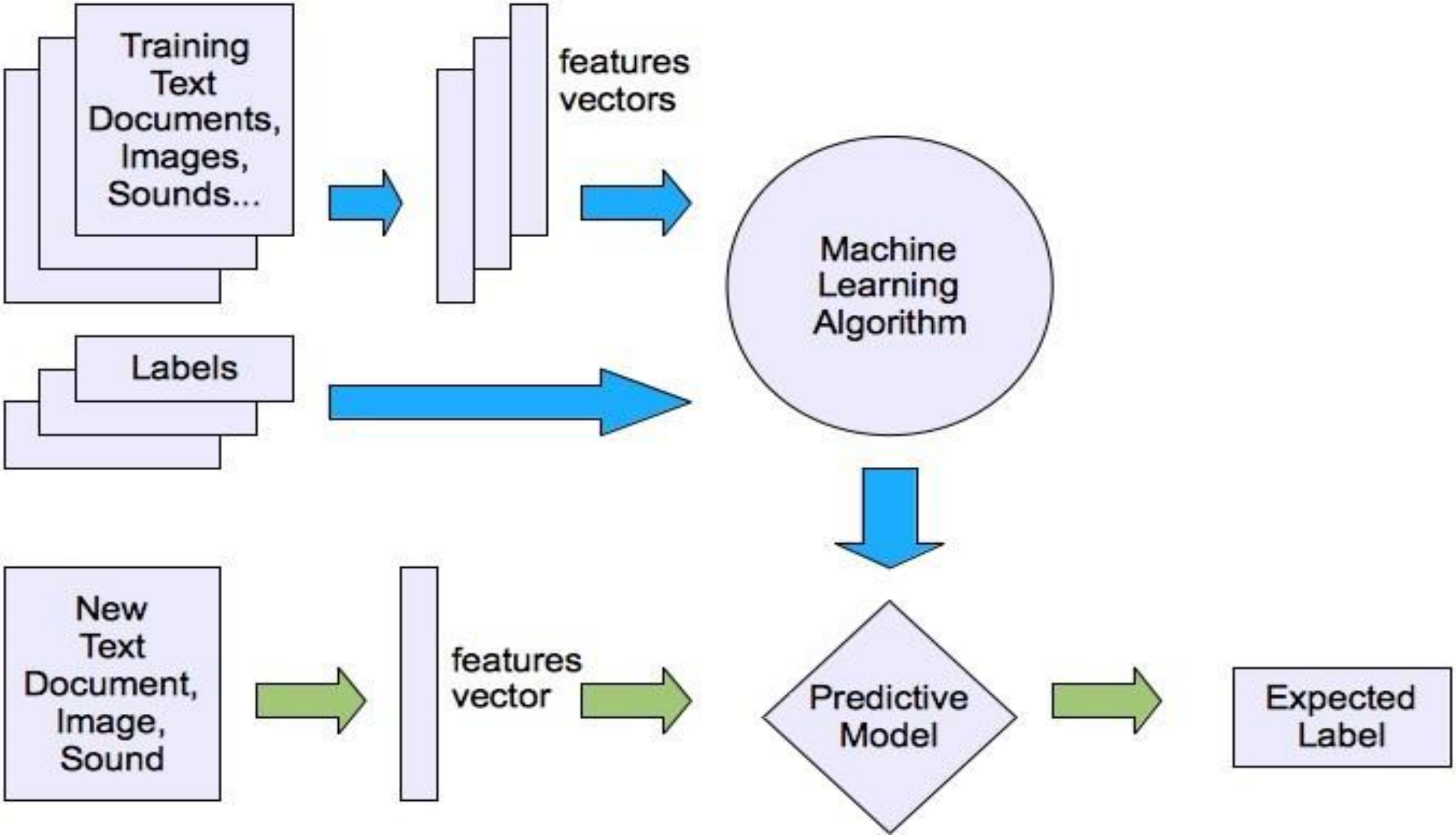
Major Classes of Learning Algorithms:

Learning Algorithms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

- The set of data (training data) consists of a set of input data and correct responses corresponding to every piece of data.
- Based on this training data, the algorithm has to generalize such that it is able to correctly (or with a low margin of error) respond to all possible inputs…
- In essence: The algorithm should produce sensible outputs for inputs that weren't encountered during training.
- Also called learning from exemplars


Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
Labels
New Text Document, Image, Sound
features vector
Predictive Model
Expected Label

Supervised Learning

- Regression Problems
- Classification Problems

Regression Problems:

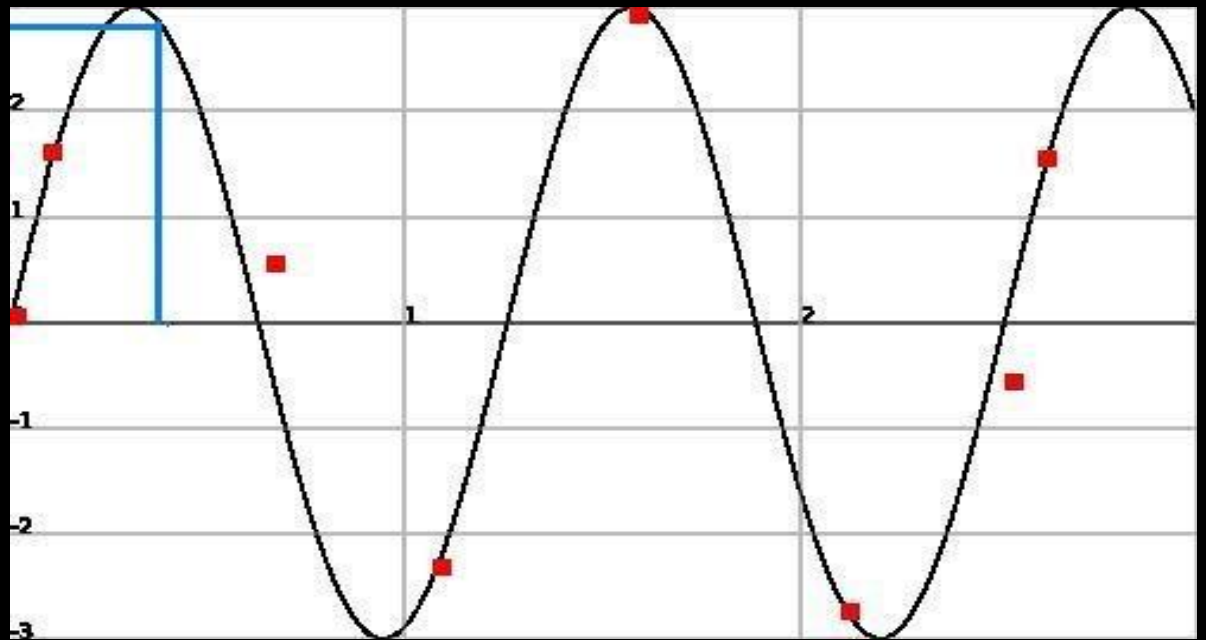
- Regression is essentially a problem of function approximation interpolation.
- You try to fit a mathematical function that describes a curve such that the curve passes as close as possible to all the data points.

Classification Problems:

- Consists of taking input vectors and deciding which of the N classes they belong to.
- How it's done: Find 'decision boundaries that can be used to separate out the different classes.

x	y
0	0
0.5236	1.5
1.5708	3.0
2.0944	-2.5981
2.6180	1.5
2.6180	1.5
3.1416	0

To Find: y at x =0.44



Unsupervised Learning

- No information about correct outputs are available!
- Regression?

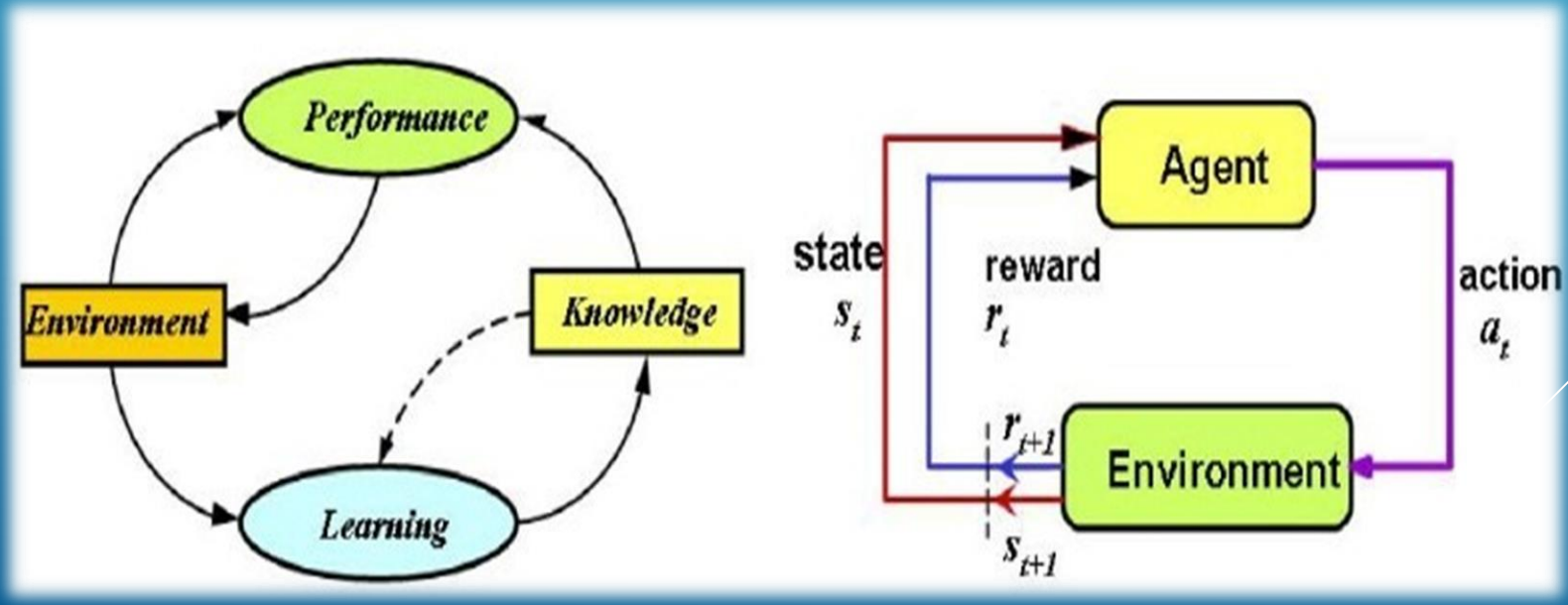
  No guesses about the function can be made.
- Classification?

No information about the correct classes. But if we design our algorithm so that it exploits similarities between inputs so as to cluster inputs that are similar together, this might perform classification automatically

- The aim of unsupervised learning is to find clusters of similar inputs in the data without being explicitly told that some datapoints belong to one class and the other in other classes. The algorithm has to discover this similarity by itself

Reinforcement Learning

- Stands in the middle ground between supervised and unsupervised learning
- The algorithm is provided information about whether or not the answer is correct but not how to improve it
- The reinforcement learner has to try out different strategies and see which maximizes the award.



MACHINE LEARNING STAGES

1. Collect data
2. Pick the model
3. Train the model
4. Test the model

CODE TIME

TensorFlow™

- open source software library for high performance numerical computation.
- strong support for machine learning and deep learning with a flexible numerical computation core

Scikit-learn

- Simple and efficient tools for data mining and data analysis
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

FIND THE GOLD FLOWER

- No. of training examples(instances): 150
- Number of features : 4
- Number of classes : 3
- Features:

  sepal length,width (cm)

  petal length,width (cm)
- Classes:

  Gold Setosa

  Iris Versicolour

  Iris Virginica

UNSUPERVISED LEARNING
CLUSTERING & KNN

Clustering

- Deals with finding structure in unlabeled data
- Clustering is “the process of organizing objects into groups whose members are similar in some way”.
- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.
- But how does one decide what constitutes a good clustering?

KNN CLUSTERING

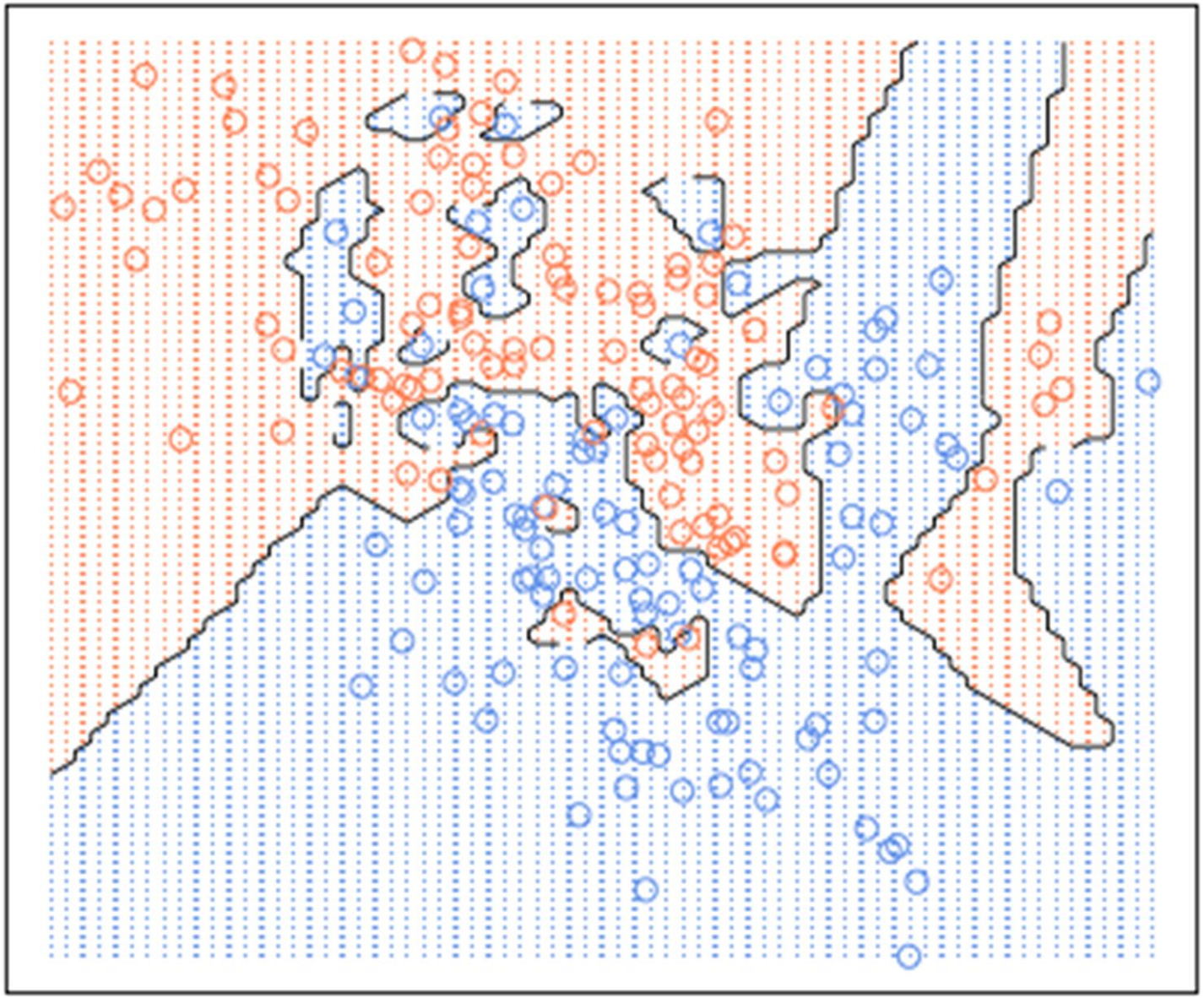
It takes a test data, and finds k nearest data values to this data from test data set. Then it selects the neighbor of maximum frequency and gives its properties as the prediction result.

KNN CLUSTERING

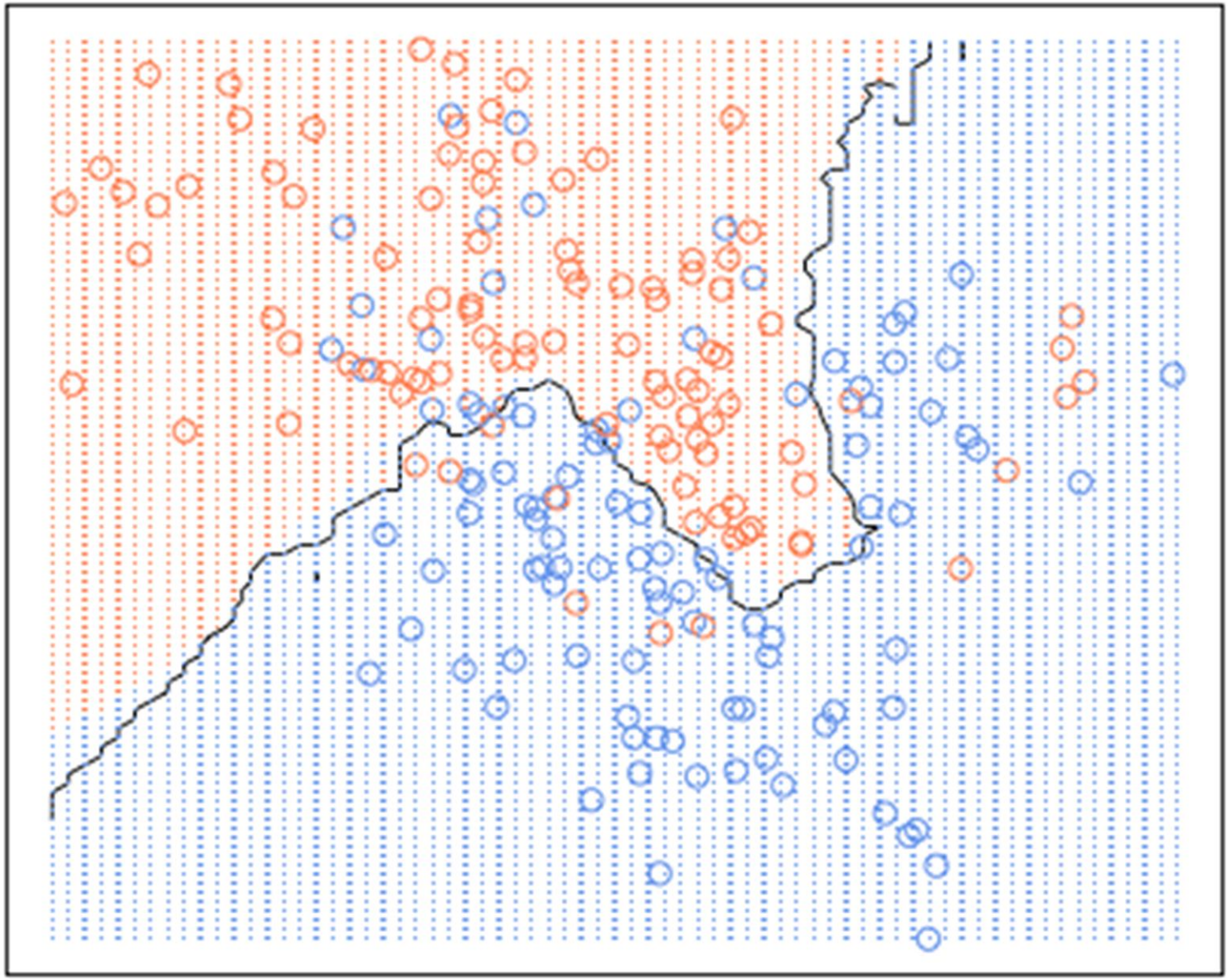
The algorithm is composed of the following steps:

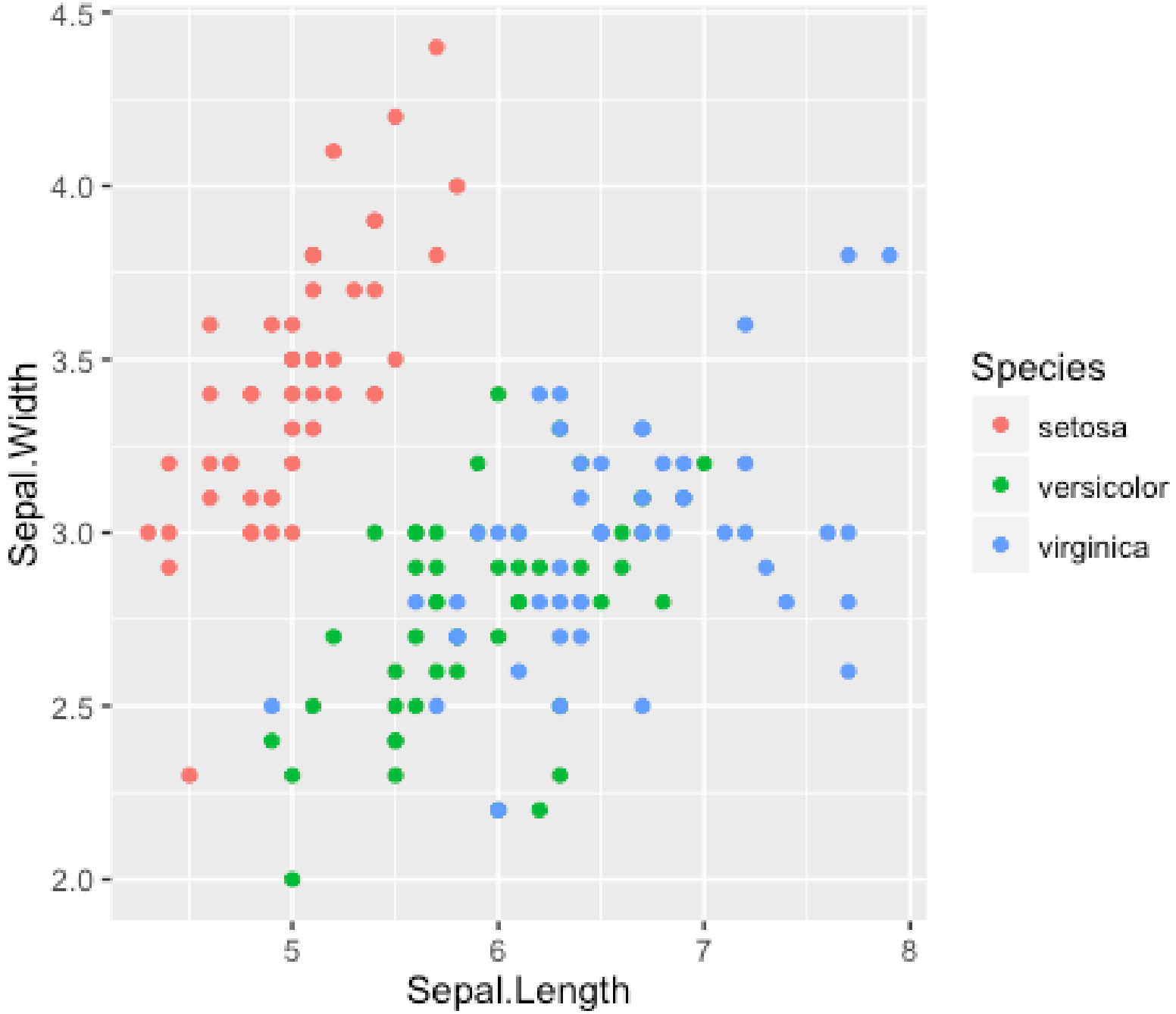
1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid
3. When all objects have been assigned, recalculate the positions of the K centroids
4. Repeat steps 2 and 3 until the centroids no longer move.

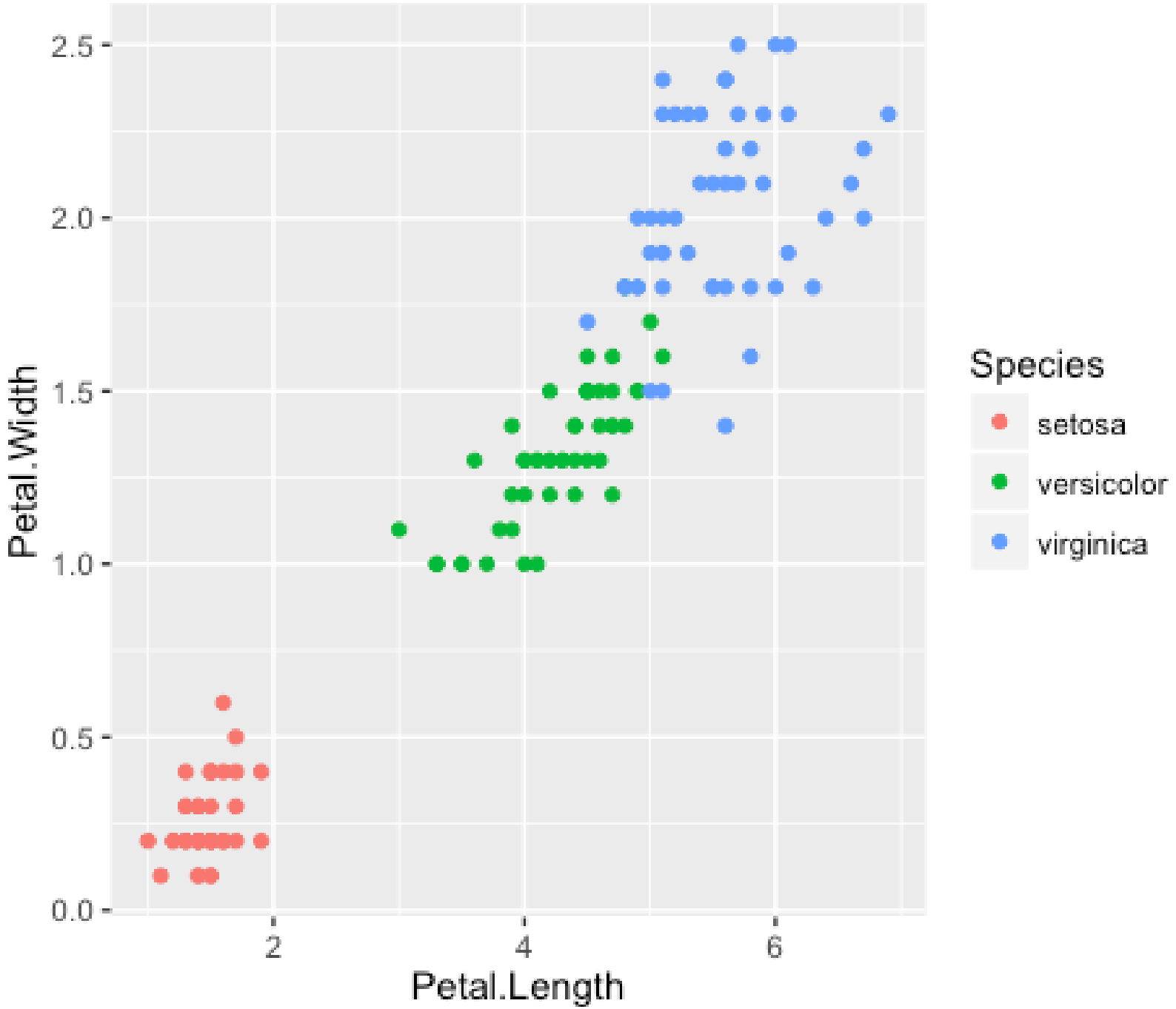
nearest neighbour (k = 1)



20-nearest neighbour




4.5
4.0
3.5
3.0
2.5
2.0
Sepal.Width
5
6
7
8
Sepal.Length
Species
setosa
versicolor
virginica


2.5
2.0
1.5
1.0
0.5
0.0
Petal.Width
2
4
6
Petal.Length
Species
setosa
versicolor
virginica