IRIS dataset

Iris Versicolor

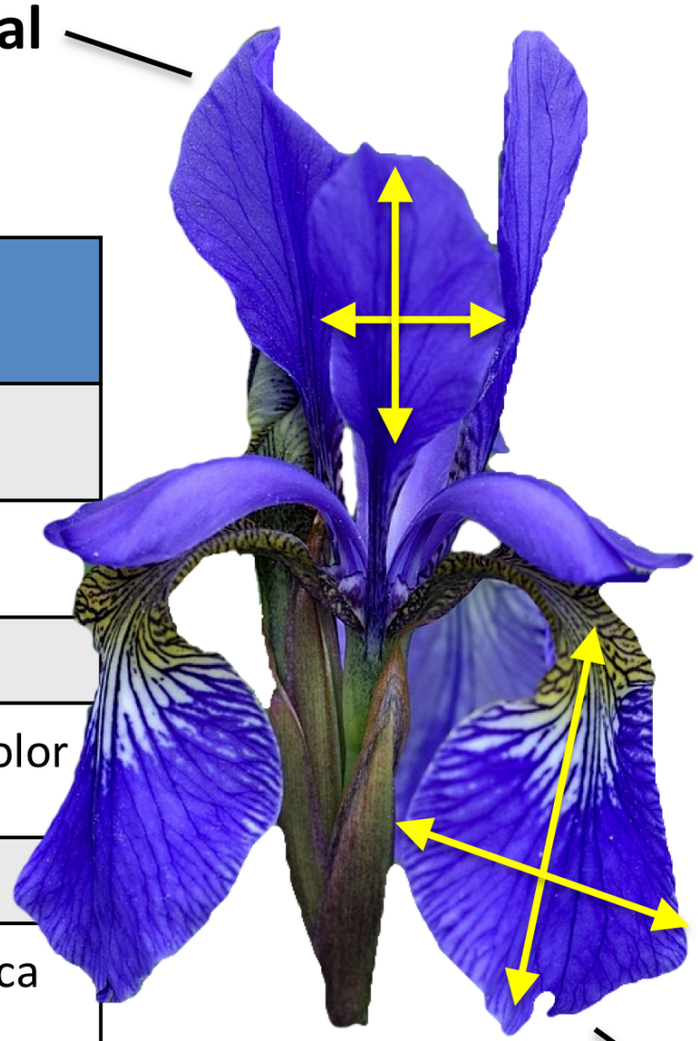
Iris Setosa

Iris Virginica

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

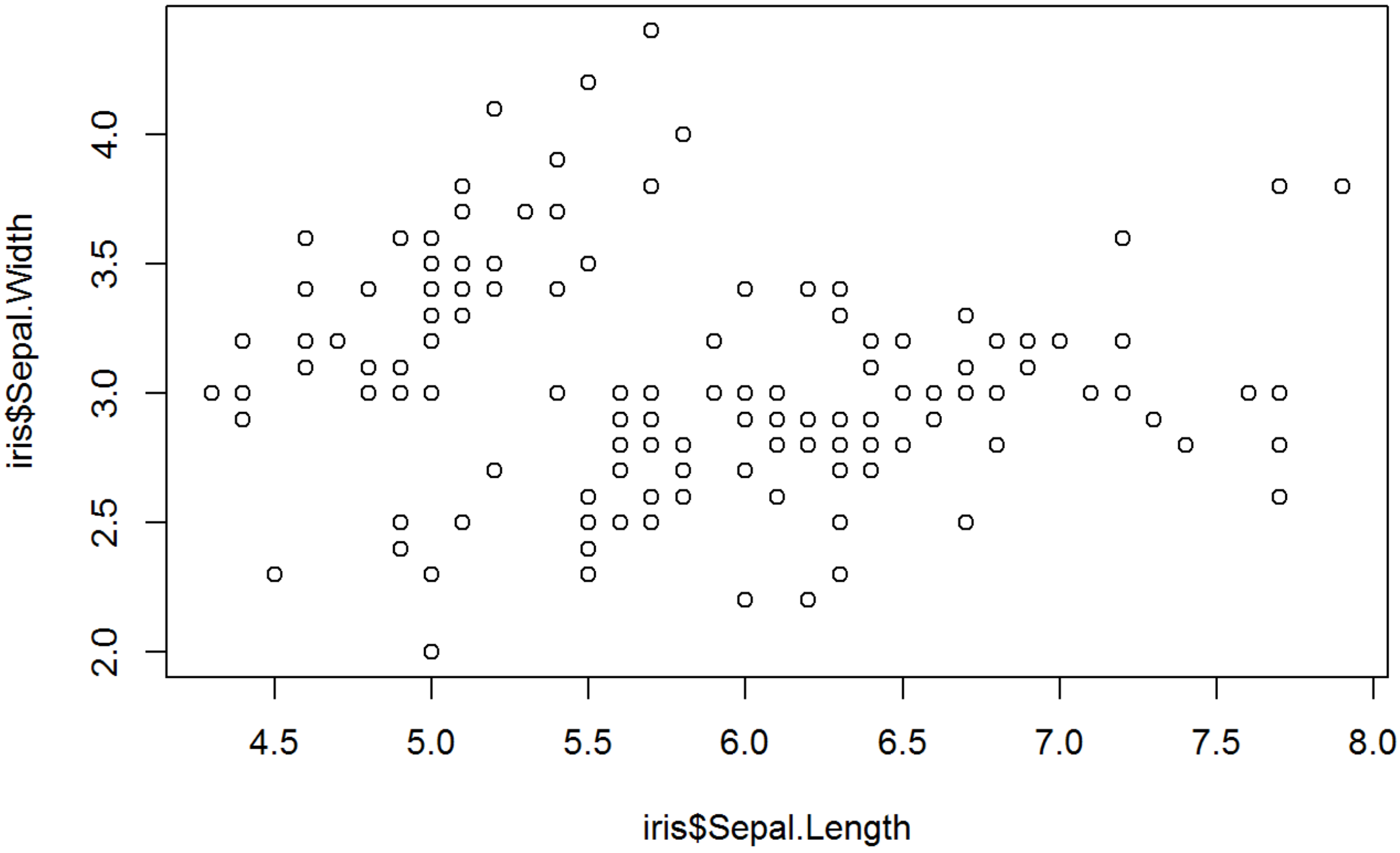
Petal

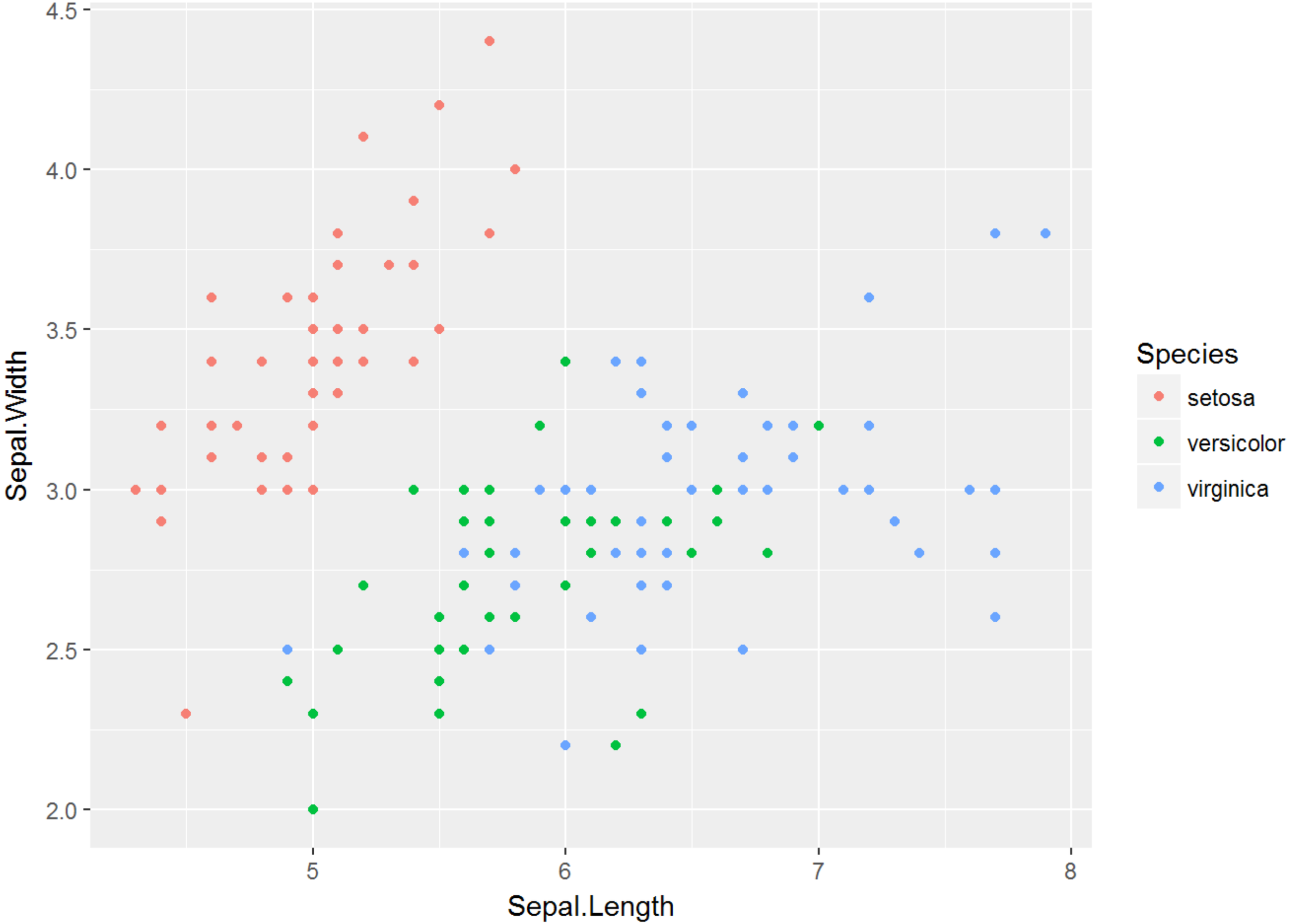


Sepal

Features
(attributes, measurements, dimensions)

Class labels
(targets)


iris$Sepal.Width
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
iris$Sepal.Length


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

The Iris dataset in scikit-learn

```
In [1]: from sklearn import datasets

In [2]: import pandas as pd

In [3]: import numpy as np

In [4]: import matplotlib.pyplot as plt

In [5]: plt.style.use('ggplot')
```

The Iris dataset in scikit-learn

```
In [1]: from sklearn import datasets

In [2]: import pandas as pd

In [3]: import numpy as np

In [4]: import matplotlib.pyplot as plt

In [5]: plt.style.use('ggplot')

In [6]: iris = datasets.load_iris()

In [7]: type(iris)
Out[7]: sklearn.datasets.base.Bunch

In [8]: print(iris.keys())
dict_keys(['data', 'target_names', 'DESCR', 'feature_names', 'target'])
```

The Iris dataset in scikit-learn

```
In [9]: type(iris.data), type(iris.target)
Out[9]: (numpy.ndarray, numpy.ndarray)

In [10]: iris.data.shape
Out[10]: (150, 4)

In [11]: iris.target_names
Out[11]: array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

Exploratory data analysis (EDA)

```
In [12]: X = iris.data

In [13]: y = iris.target

In [14]: df = pd.DataFrame(X, columns=iris.feature_names)

In [15]: print(df.head())
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
0                5.1               3.5                1.4               0.2
1                4.9               3.0                1.4               0.2
```

Visual EDA

```
In [16]: _ = pd.scatter_matrix(df, c = y, figsize = [8, 8],
    ...:                                s=150, marker = 'D')
```

```
In [3]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435 entries, 0 to 434
Data columns (total 17 columns):
party                435 non-null object
infants              435 non-null int64
water                435 non-null int64
budget               435 non-null int64
physician            435 non-null int64
salvador             435 non-null int64
religious            435 non-null int64
satellite            435 non-null int64
aid                  435 non-null int64
missile              435 non-null int64
immigration          435 non-null int64
synfuels             435 non-null int64
education            435 non-null int64
superfund            435 non-null int64
crime                435 non-null int64
duty_free_exports    435 non-null int64
eaa_rsa              435 non-null int64
dtypes: int64(16), object(1)
memory usage: 57.9+ KB
```

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Congressional Voting Records Data Set

Download: Data Folder, Data Set Description

Abstract: 1984 United Stated Congressional Voting Records; Classify as Republican or Democrat

Data Set Characteristics:	Multivariate	**Number of Instances:**	435	**Area:**	Social
Attribute Characteristics:	Categorical	**Number of Attributes:**	16	**Date Donated**	1987-04-27
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	136207

Source:

Origin:

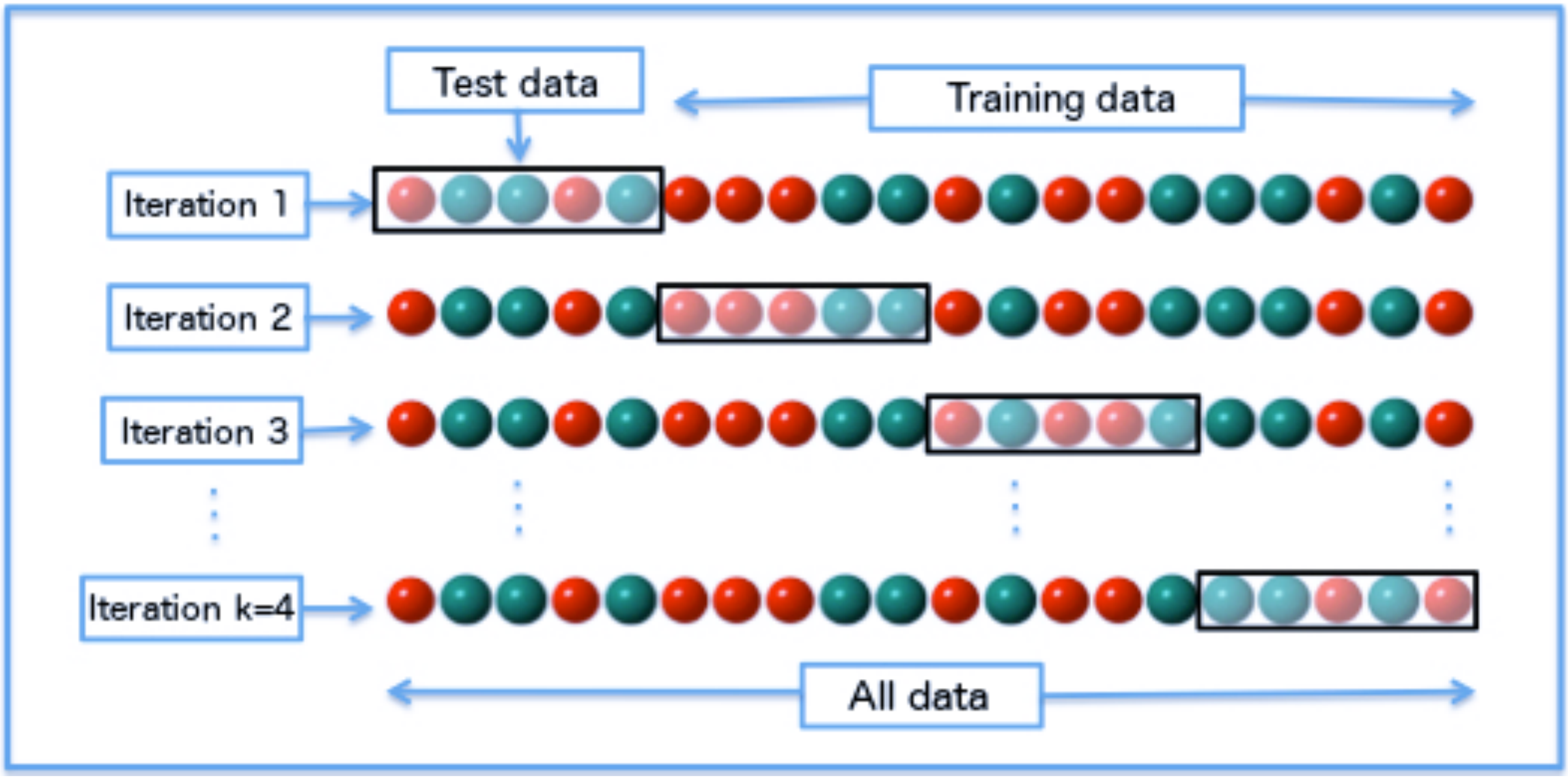
Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985.

Donor:

Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

Data Set Information:

CROSS VALIDATION



CROSS VALIDATION