Classification

Javier M. Antelis

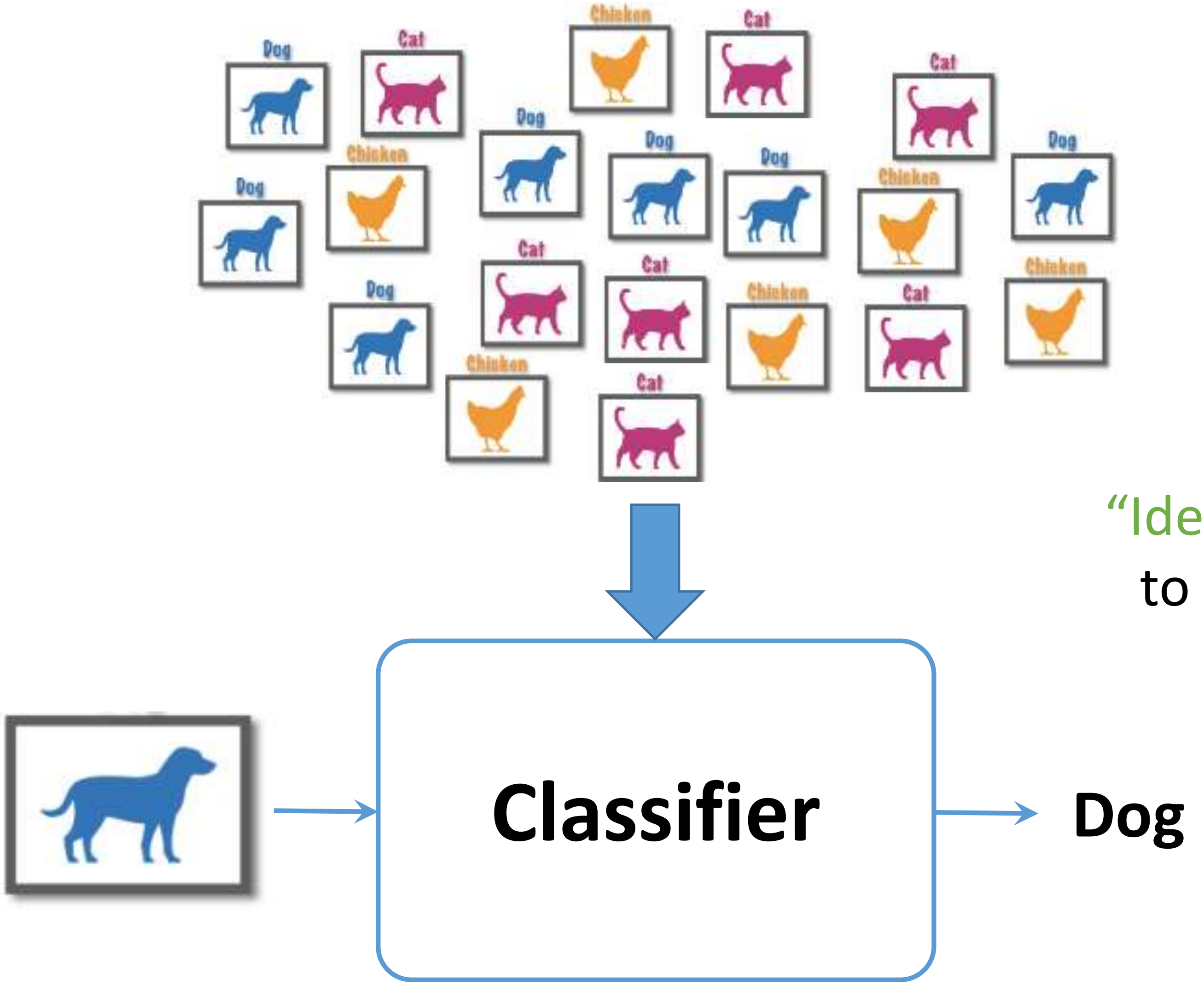
mauricio.antelis@tec.mx

Tecnológico de Monterrey

Goal

To study and apply the special type of machine learning models devoted to identify the “category/class/group” to which an “observation” belongs to.

What is “classification”?



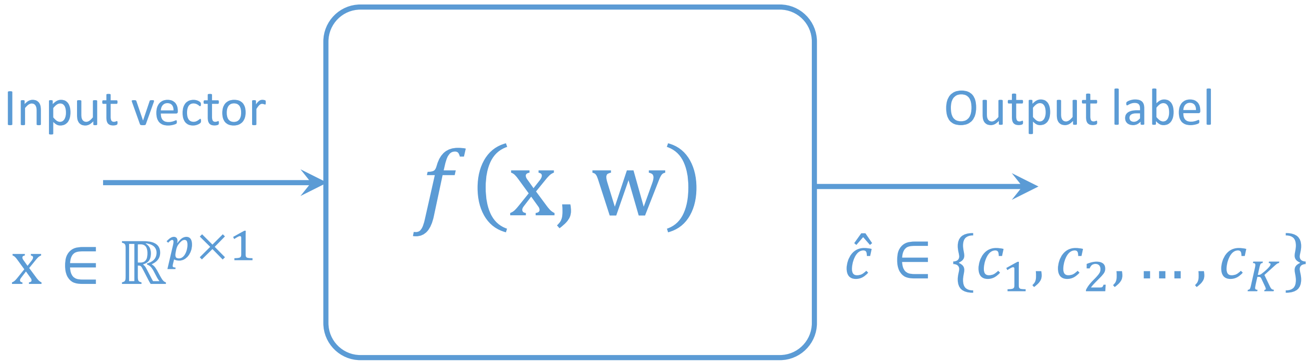
Classifier

Dog

“Identification” of the “category/class/group/” to which an “observation/input” belongs to

Classifier

- A classifier is a decision function that assigns an observation to one o several classes, $\hat{c} = f(\mathrm{x}, \mathrm{w})$, that is:



Input vector

$\mathrm{x} \in \mathbb{R}^{p \times 1}$

$f(\mathrm{x}, \mathrm{w})$

Output label

$\hat{c} \in \{c_1, c_2, \dots, c_K\}$

- Where:
  - x → input data (numeric or categoric)
  - $f(\cdot)$ → decision function
  - w → model parameters
  - $\hat{c}$ → predicted class
  - c → true class

"The function must be defined/chosen"

"The parameters w are learned from data: training"

Classifier: example

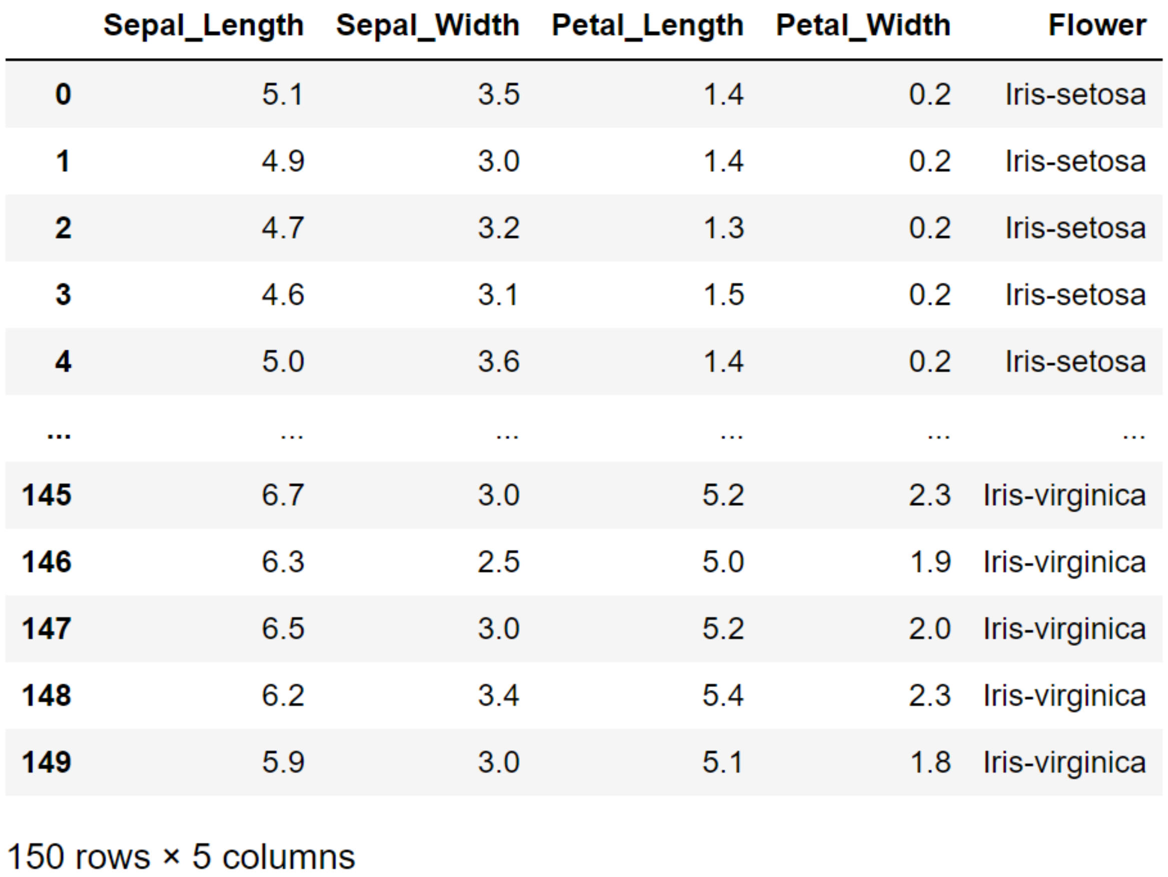
- The iris dataset
  - Ronald Fisher, 1936
  - https://archive.ics.uci.edu/ml/datasets/iris
  - Commonly found in the ML literature

- General description:
  - 4 features: sepal and petal length and width
  - 3 classes: type of iris plant
  - 150 observations (50 for each class)

Iris Versicolor

Iris Setosa

Iris Virginica



	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Flower
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

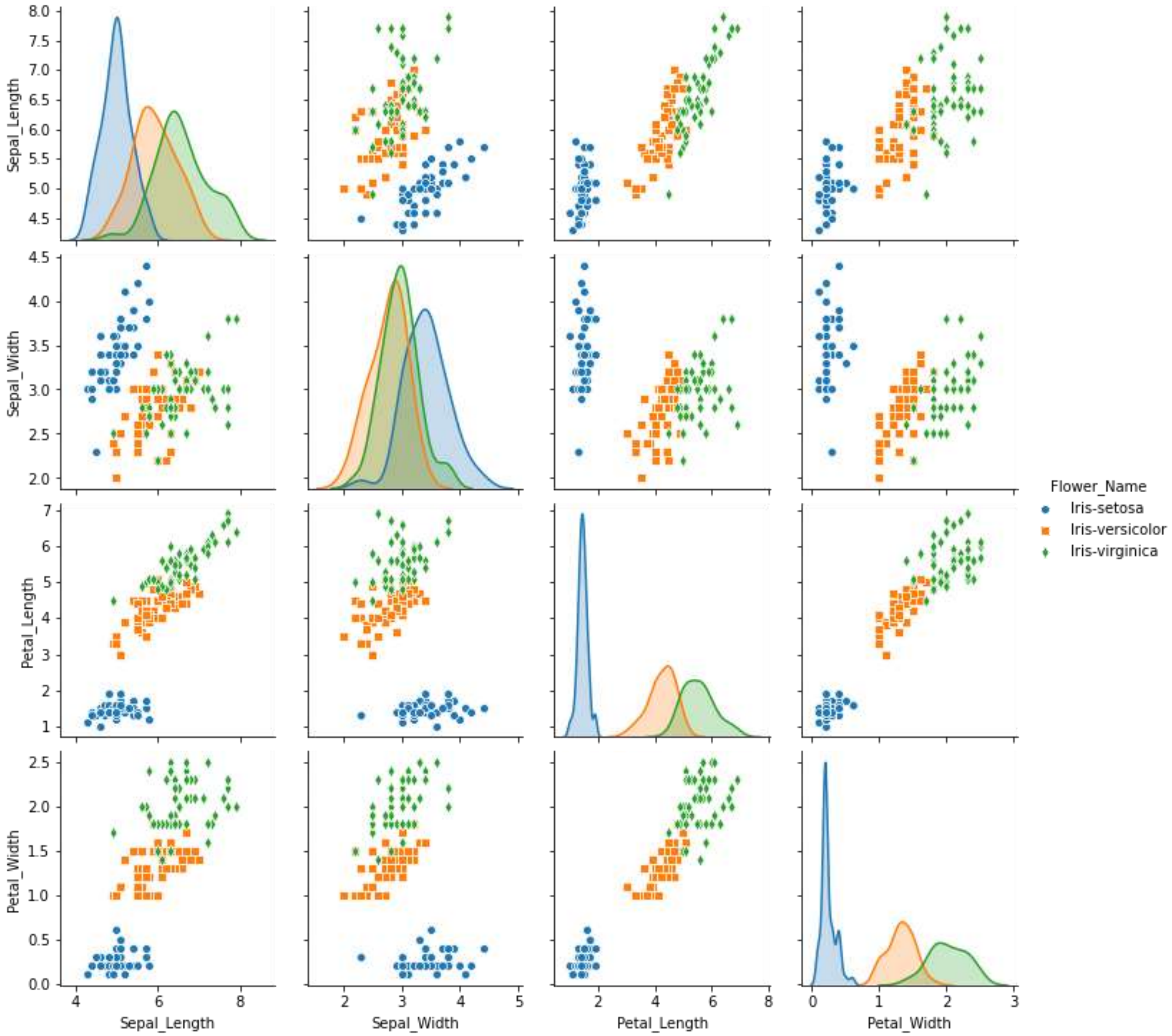
150 rows × 5 columns

Classifier: example

- The iris dataset
  - Ronald Fisher, 1936
  - https://archive.ics.uci.edu/ml/datasets/iris
  - Commonly found in the ML literature

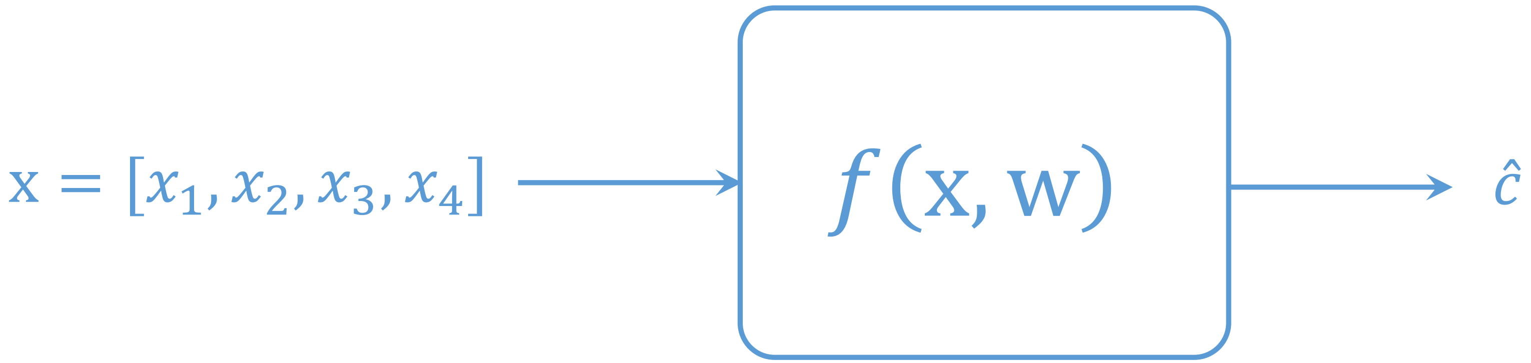
- General description:
  - 4 features: sepal and petal length and width
  - 3 classes: type of iris plant
  - 150 observations (50 for each class)

- Technical description:
  - One class is linearly separable from the other two
  - The latter are NOT linearly separable from each other



Classifier: example

- Description of the classification problem



$\mathrm{x} = [x_1, x_2, x_3, x_4]$
$f(\mathrm{x}, \mathrm{w})$
$\hat{c}$

- $\mathrm{x} = [x_1, x_2, x_3, x_4]$ is a 4-dimensional feature vector
- $c \in \{Setosa, Versicolor, Virginica\}$ are the three categories

Given the information from a new flower (four features) we want to decide which type of flower it is (three classes)

Classifier

Regardless of the classification model, we require a training data set to calculate the model parameters

The process of calculating the model parameters using a training data set is known as supervised learning

Challenges in supervised learning

- Model selection: we need to choose the decision function

$$f(\cdot)$$

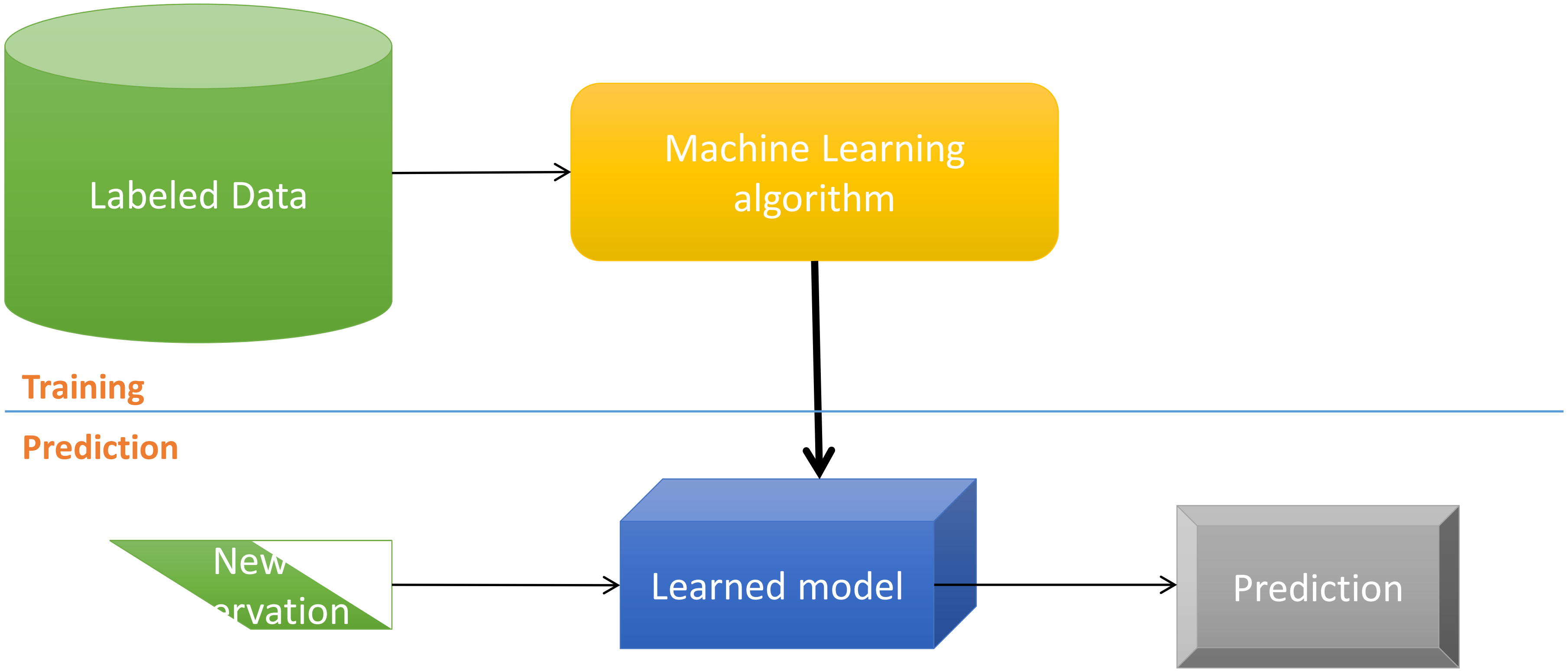
- Training of the model: use $\{\mathrm{x}_i, c_i\}_{i=1}^{N}$ to calculate the model parameters

$$\mathrm{w}$$

- Evaluation: to assess prediction of unknown data

Performance metrics, evaluation of classifiers

Workflow 1



Labeled Data

Machine Learning algorithm

Training

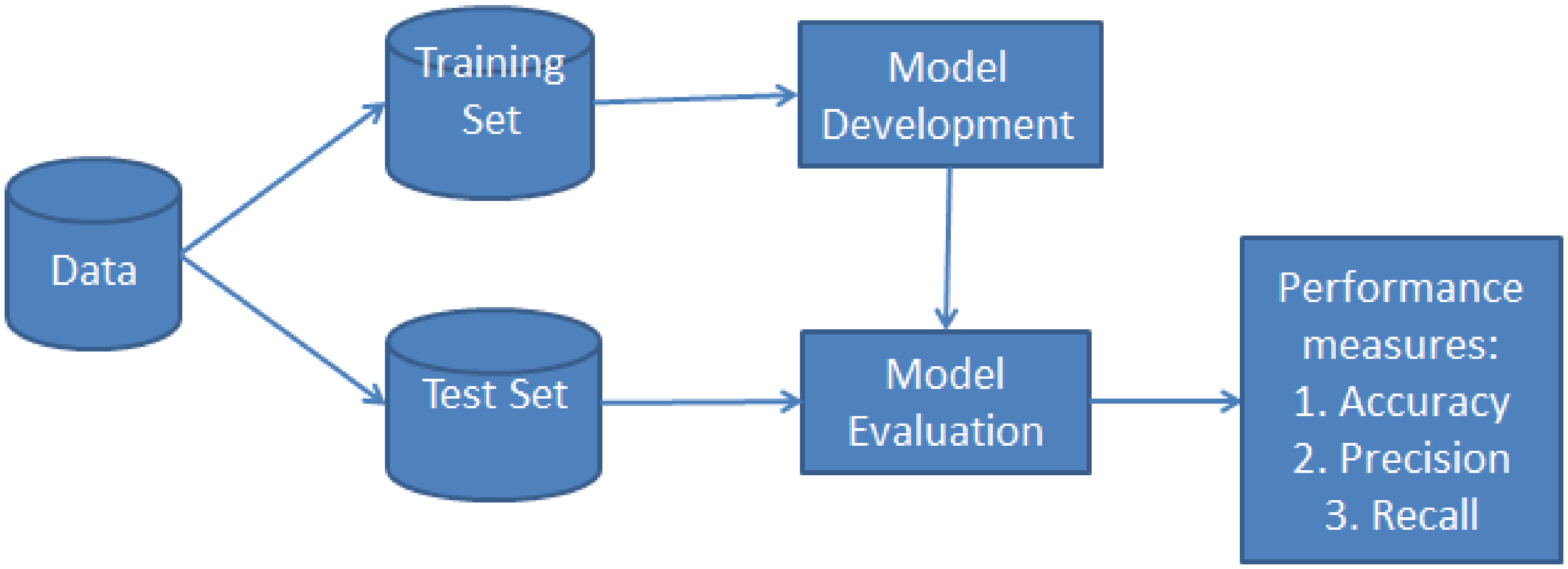
Prediction

New ervation

Learned model

Prediction

Workflow 2


Data
Training Set
Test Set
Model Development
Model Evaluation
Performance measures:
1. Accuracy
2. Precision
3. Recall