CSIS3764 DATA SCIENCE

Mr WSJ Marais

Machine Learning

T: 051 401 2754 itinfo@ufs.ac.za **www.ufs.ac.za/it**

UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA

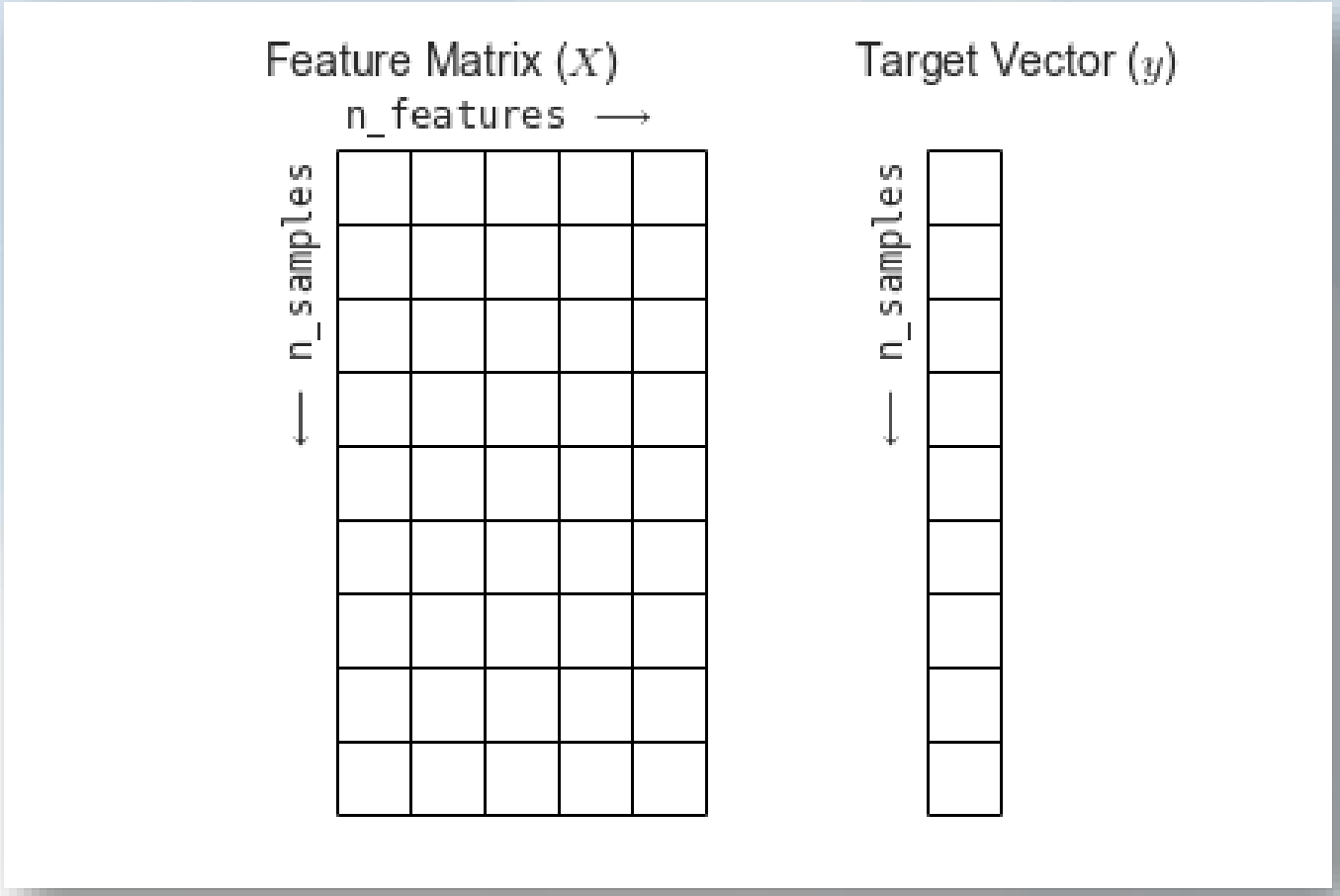
UFS·UV
NATURAL AND
AGRICULTURAL SCIENCES
NATUUR- EN
LANDBOUWETENSKAPPE

Machine Learning Categories

- Supervised Learning
  - Regression
  - Classification
- Unsupervised Learning
  - Clustering
- Reinforcement Learning
- Deep Learning
  - KNN
  - RNN

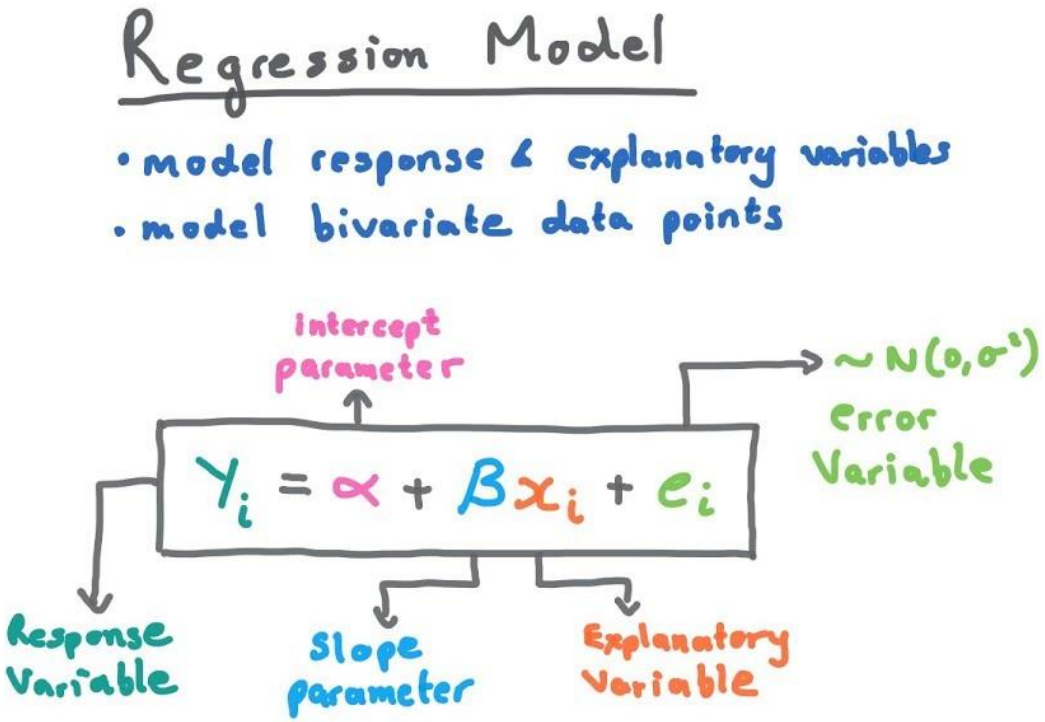
UFS UV

Feature vector, Vector space, Feature Space

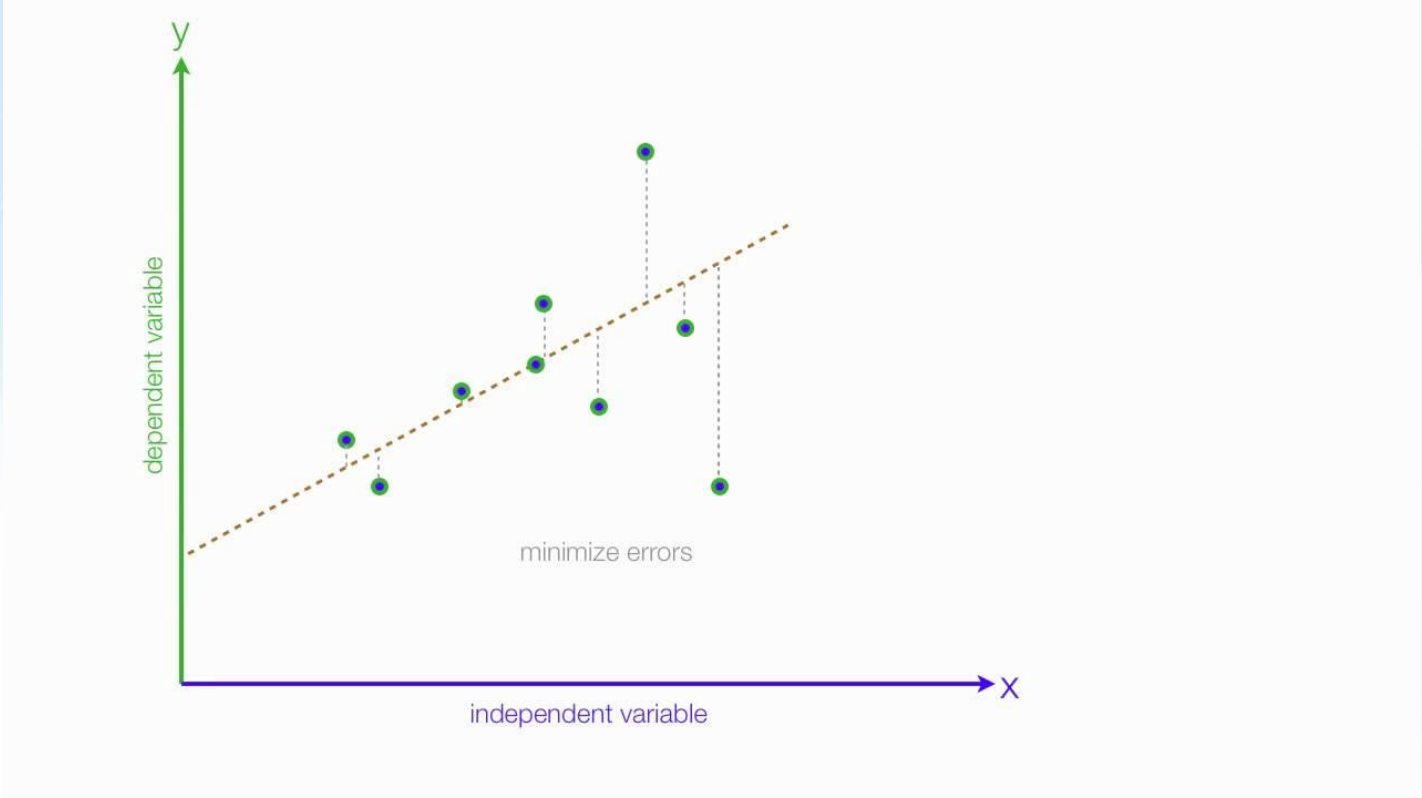


The vector space associated with these vectors is often called the **feature space**.

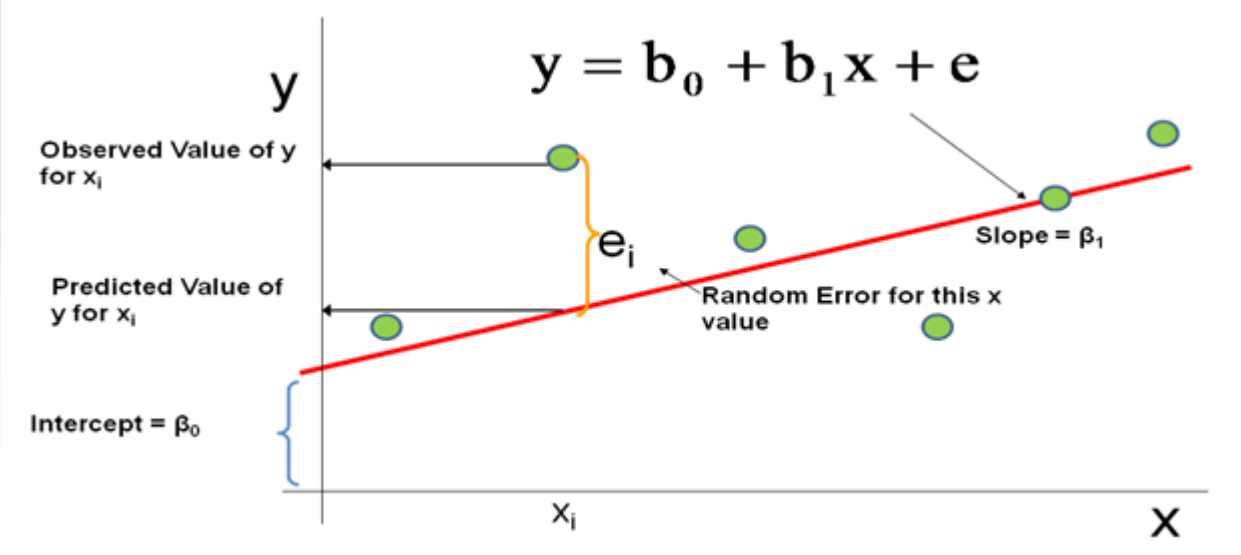
Regression Model



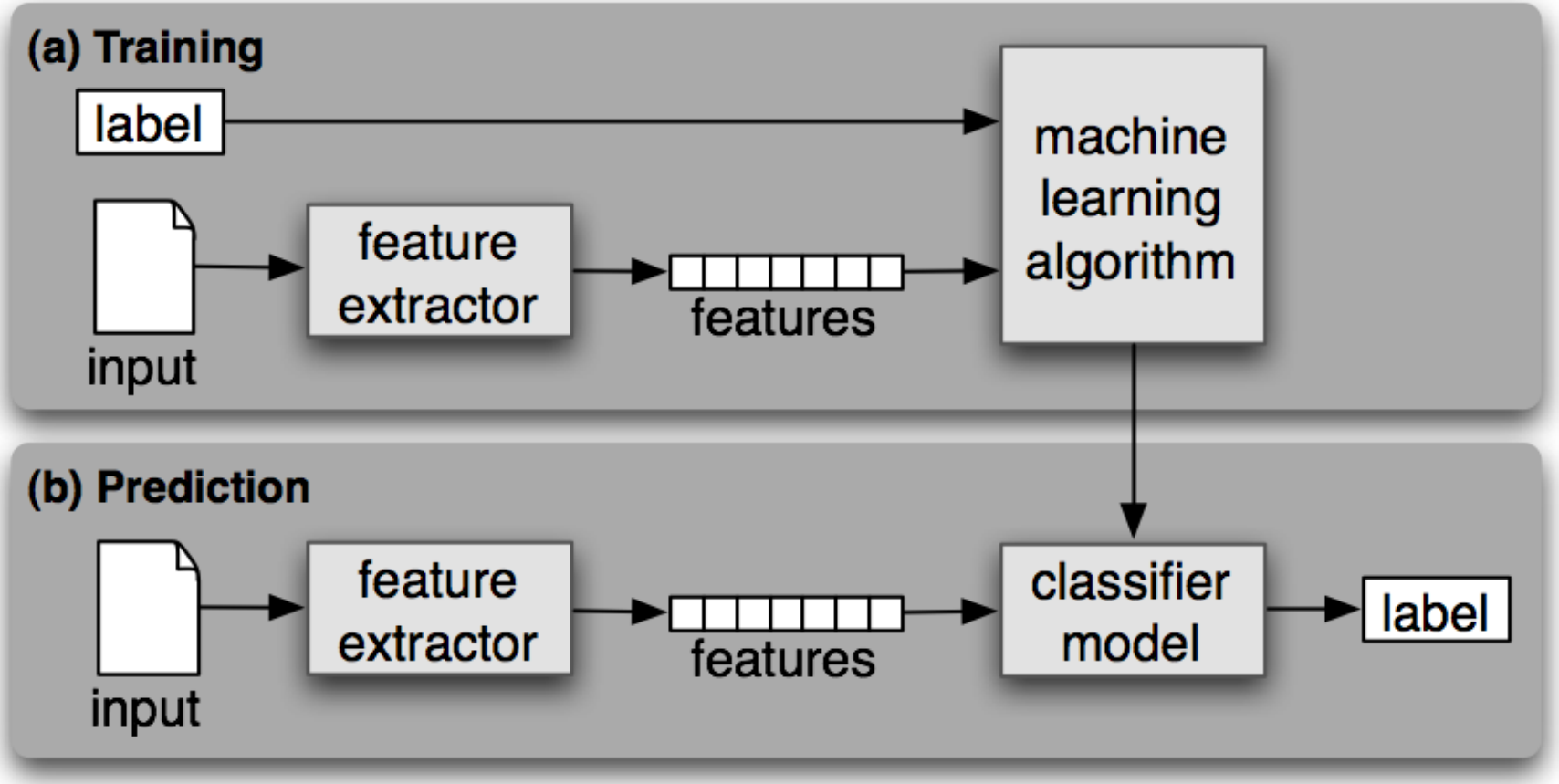
Regression Model



Regression Model



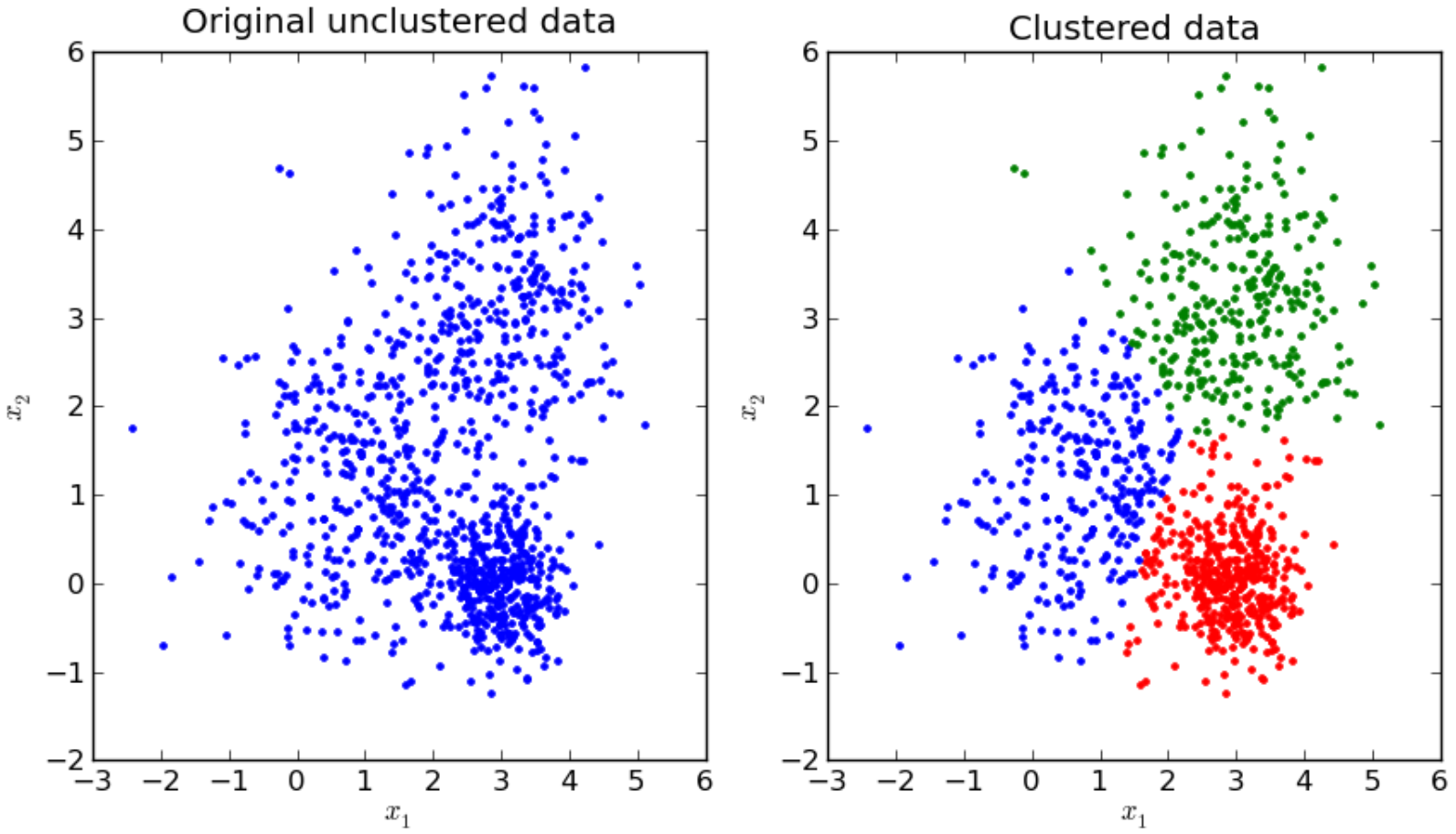
CLASSIFIER MODEL



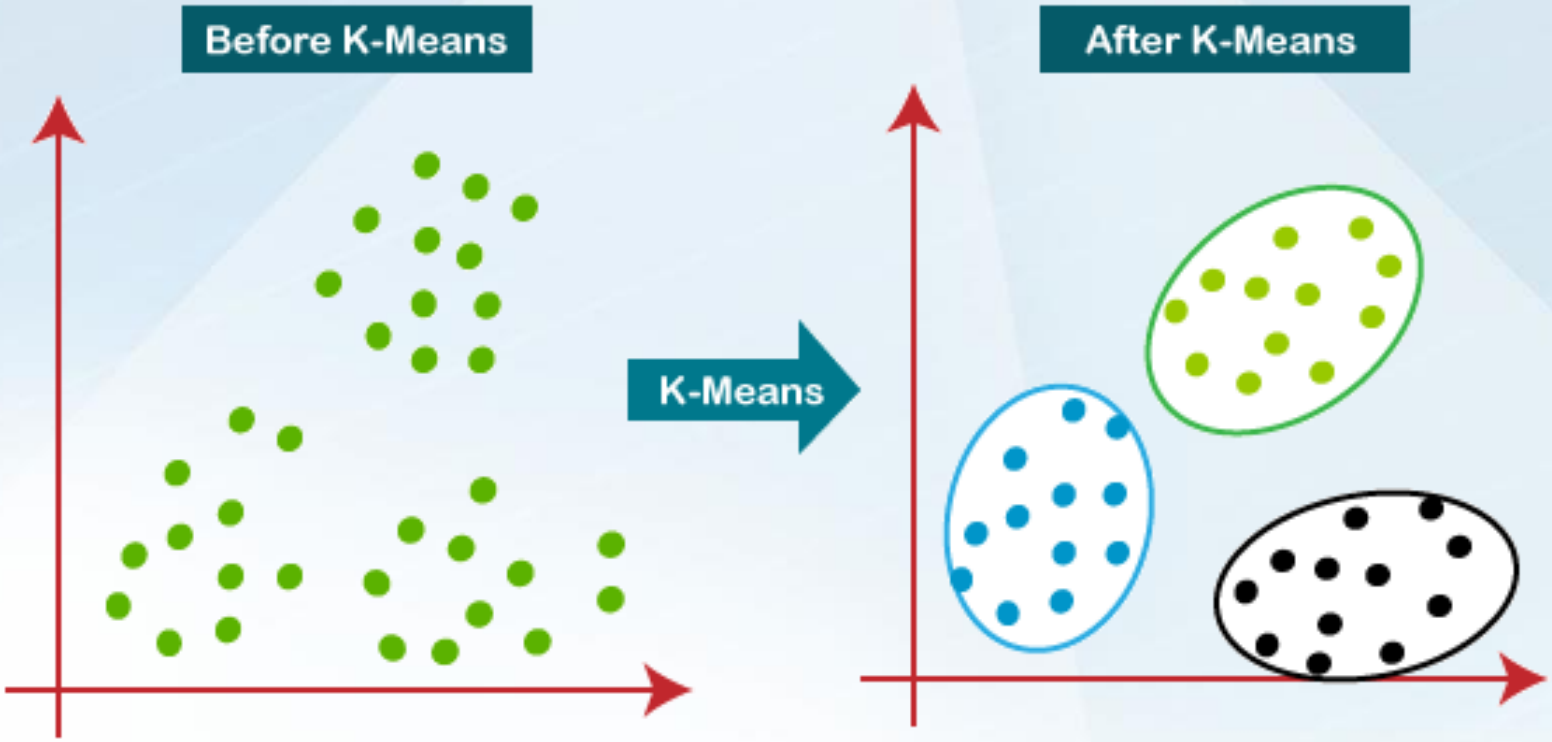
Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model.

https://www.nltk.org/book/ch06.html

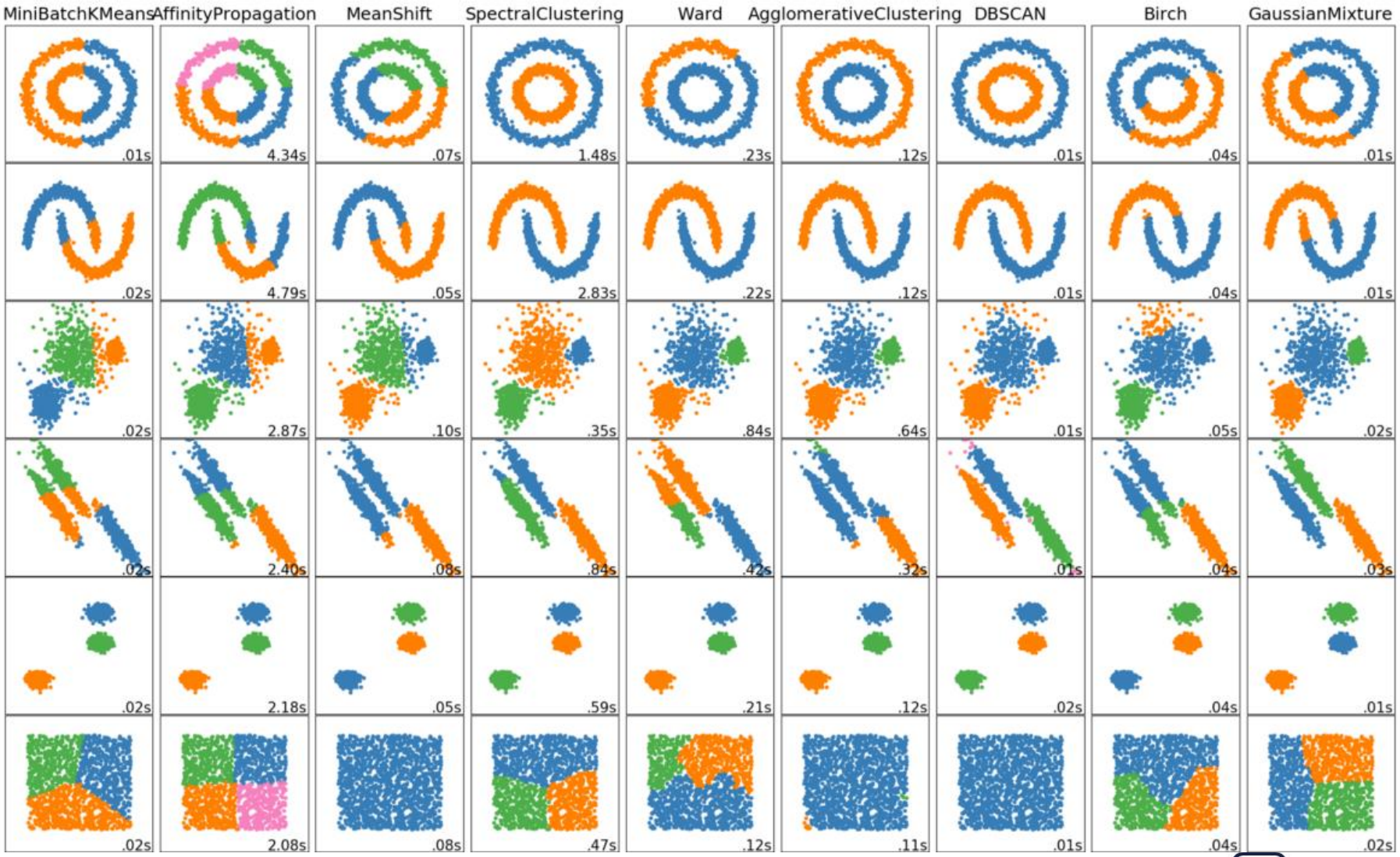
CLUSTERING MODEL



CLUSTERING MODEL



CLUSTERING MODEL



UFS UV

PERFORMANCE METRICS

(1) **Accuracy** measures the percentage classified correctly over all test cases:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}.$$

(2) **Recall** is the percentage of positive samples that were classified correctly. Recall measures how often a system correctly classifies positive samples when it encounters them.

$$Re = \frac{TP}{TP + FN}.$$

(3) **Precision** is the percentage of correctly classified positive samples over all positive classifications. Precision measures how often a system gets positive classifications correct:

$$Pr = \frac{TP}{TP + FP}.$$

(4) **F_1-score** measures the balance between the precision and recall of a system. A higher f_1-score indicates a more accurate system:

$$F_1 = 2 \times \frac{Pr \times Re}{Pr + Re}.$$

PERFORMANCE METRICS

- Say we have a detection model which identifies 8 dogs in a picture containing 12 dogs and some cats....
  - Out of these 8, 3 predictions were actually cats, thus wrong (false positives=FP) and 5 were correct (true positives=TP).
  - In this case, The precision is (5/8), while the recall is (5/12).
- So, precision is "how useful the classification results are", and recall is "how complete the classifications are".

UFS UV

CONTENT

- Categories of Machine Learning
- Scikit-learn
- Hyperparameter optimization (including GridSearch)
- Model validation (validation curve & learning curve)
- Data Preprocessing
- Feature engineering (basics)
- Classifier pipelines (make_pipeline)
- Cross validation & nested cross validation

T: 051 401 2754 itinfo@ufs.ac.za **www.ufs.ac.za/it**

UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA

UFS·UV
NATURAL AND
AGRICULTURAL SCIENCES
NATUUR- EN
LANDBOUWETENSKAPPE