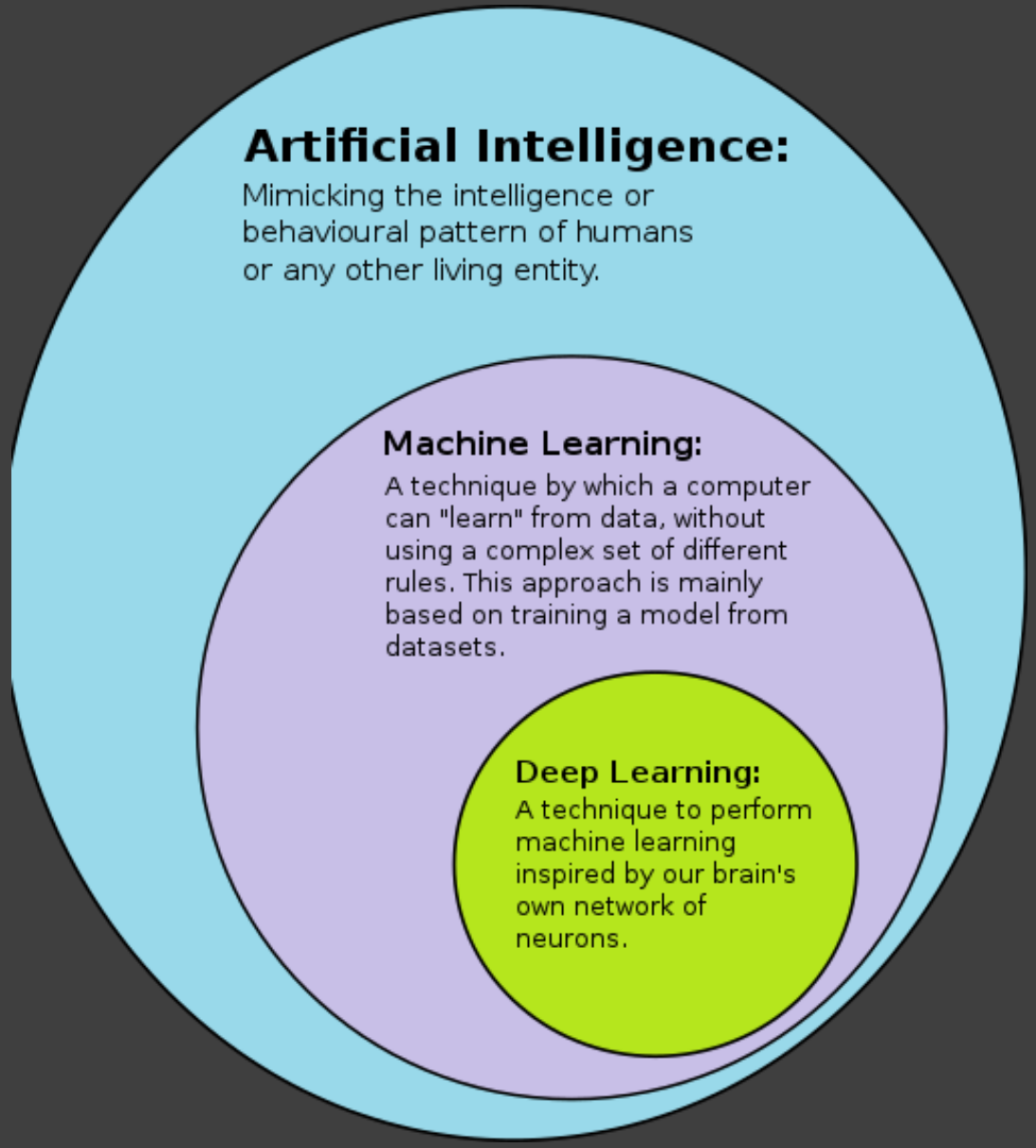
Machine Learning: Introduction



Machine Learning

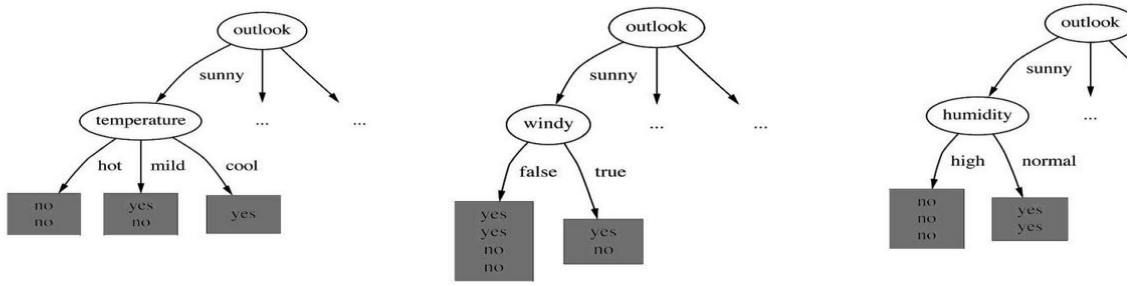
- Machine learning software means computer programs that can modify themselves (e.g. based on data) without being explicitly programmed.
- Subfield of Computer Science, derived from artificial intelligence
- Machine learning algorithms include:
  - Classification (supervised learning)
  - Clustering (unsupervised learning)
  - Reinforcement learning

Classification algorithms

- Supervised machine learning
- Training data with class, test data without class
- A model is formed
- Examples:
  - tree-based methods (e.g. C4.5, ID3 from the 1980s)
  - Lazy, or example-based learning: K-NN (k-nearest-neighbor)

Ian Witten et al., Data Mining: Practical Machine Learning Tools and Techniques, 2017.:

Continue to split …



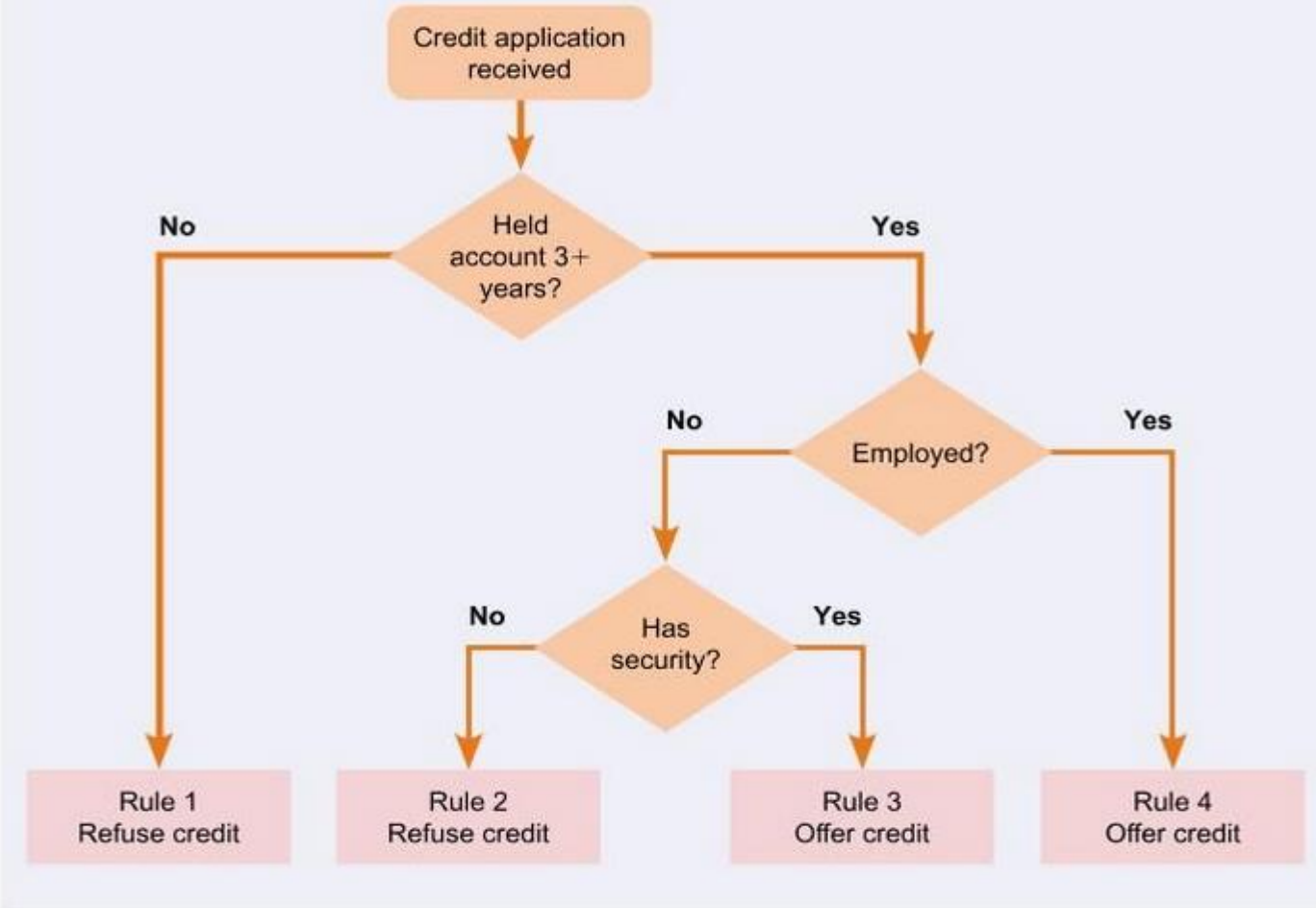
gain(*temperature*) = 0.571 bits
gain(*windy*) = 0.020 bits
gain(*humidity*) = 0.971 bits

Use cases for classification with decision trees

- Analytical problems involving making a decision.
- They may be used by banks for loan approvals just because of their extreme transparency of rule-based decision-making

Lähde: Business Information Systems, 5th edn, 2015 - Paul Bocij, Andrew Greasley, Simon Hickie

Figure 1.7 Decision tree for the loan application example



The source code of the tree

Source: Business Information Systems, 5th edn, 2015 - Paul Bocij, Andrew Greasley, Simon Hickie

Figure 1.9 Structured English program code for implementing the decision table shown in Figure 1.7

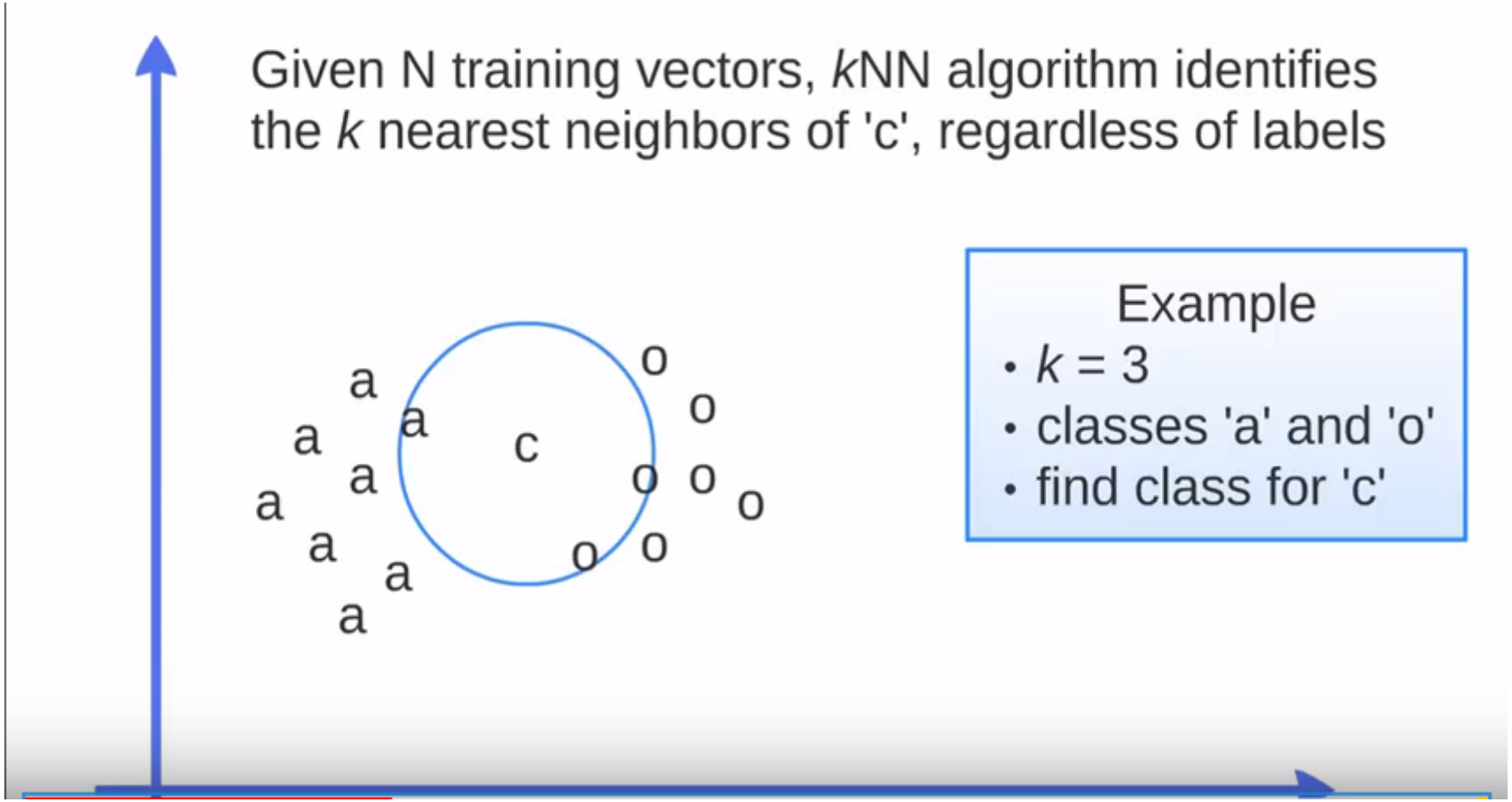
```
IF Held Account 3+ Years THEN
  IF Employed THEN
    Accept Application                          (Rule 4)
  ELSE
    IF Can Offer Security THEN
      Accept Application                        (Rule 3)
    ELSE
      Decline Application                       (Rule 2)
        ENDIF
    ENDIF
ELSE
  Decline Application                           (Rule 1)
ENDIF
```

The kNN algorithm

- kNN algorithm
- Another simple classification algorithm
- When to use kNN algorithm?
  - ease of interpretation
  - low calculation time.

The kNN algorithm: e.g. classify images into cats and dogs

Source: Thales Sehn Körting : How kNN algorithm works

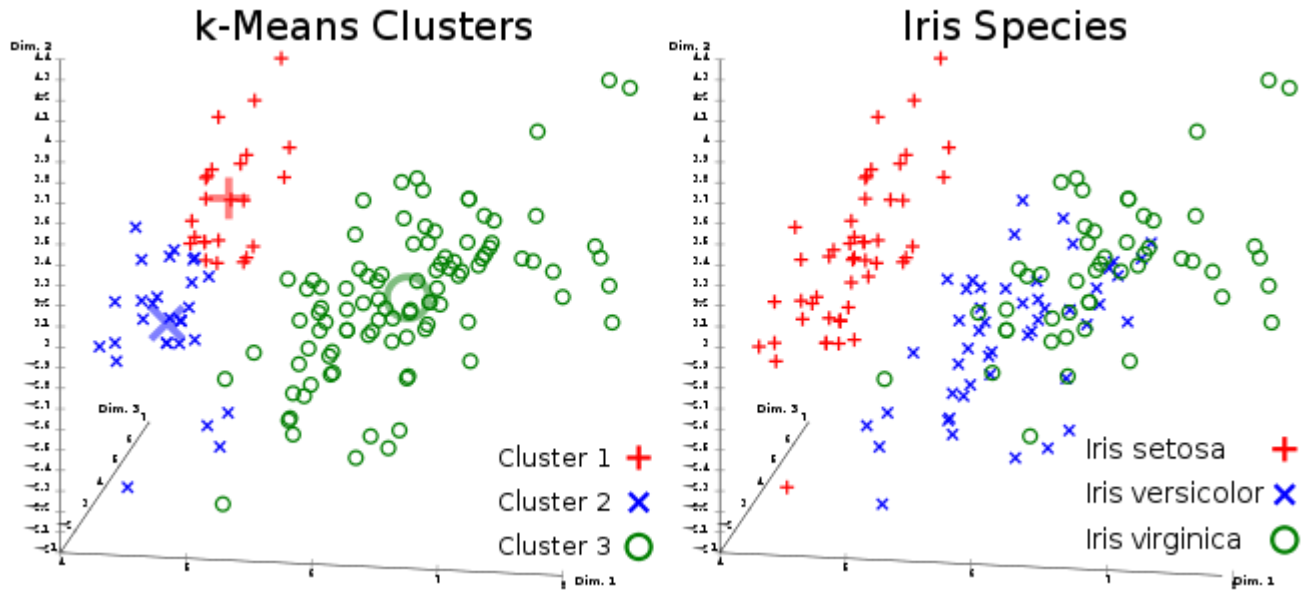


Unsupervised learning: Clustering

- No training data, the idea is to discover similarities
- Typical use case: customer segmentation
- K-means clustering example
- Preprocessing method for classification methods

Clustering

Data: unlabelled
Examples of algorithms: k.means, hierarchical clustering, …



K-means clustering

1. Choose the number of clusters, k
2. Select the k cluster centroids
3. Assign each object to the nearest centroid
4. Recompute the new cluster centroid
5. Repeat step 3 and 4 until no more changes or until max number of iterations is reached

Source: Jared Dean: Big Data, Data Mining and Machine Learning, Wiley, 2014.

When to use k-means?

- to data that has a smaller number of dimensions,
- is numeric,
- and is continuous.
- randomly distributed

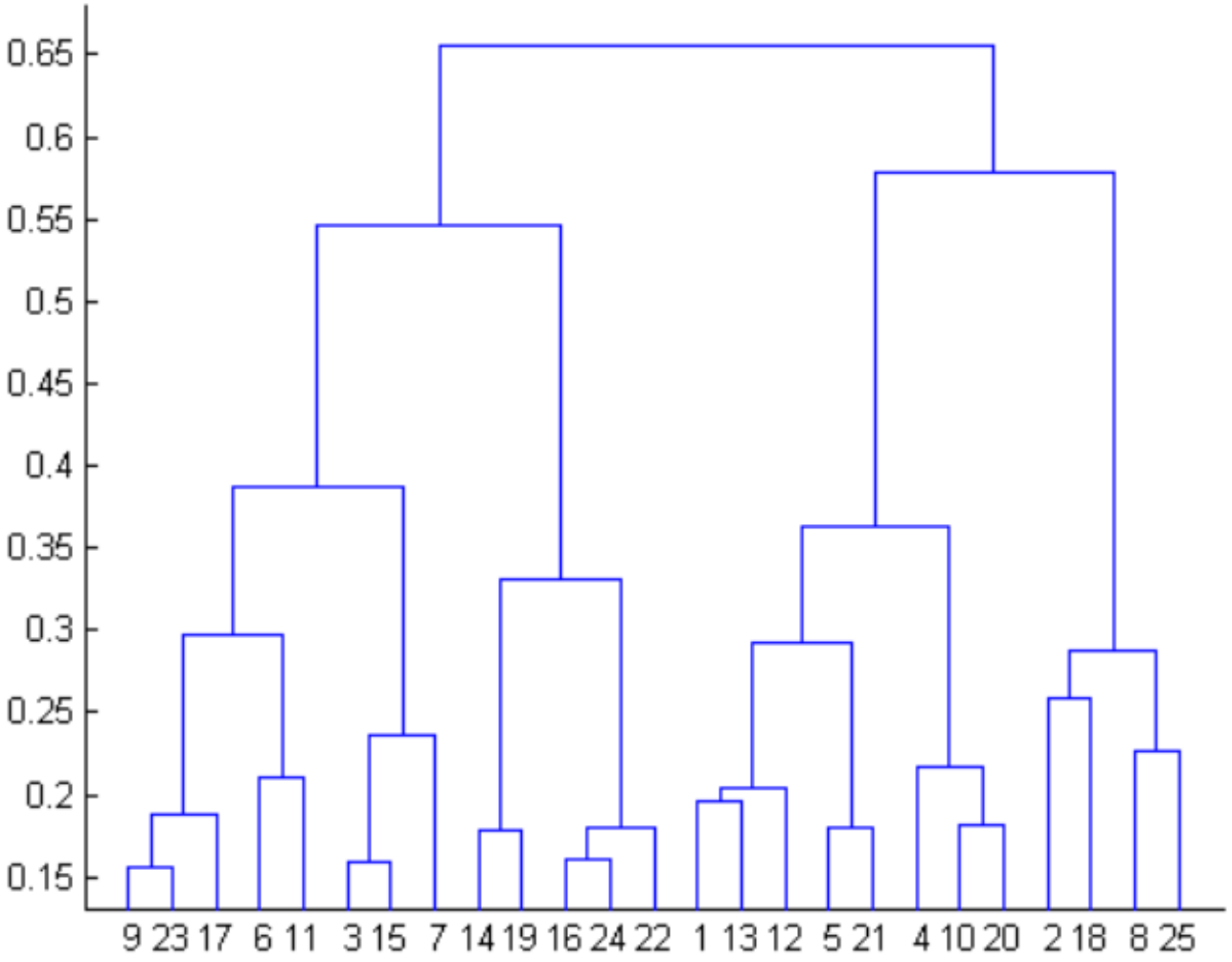
Use cases

- Document clustering
- Customer segmentation
- Fraud detection
- IT alert clustering

Hierarchical clustering

- top down / bottom up
- useful when the data has a hierarchical structure
- no need to specify the number of clusters
- computationally expensive
- demo
- agglomerative and divisive methods

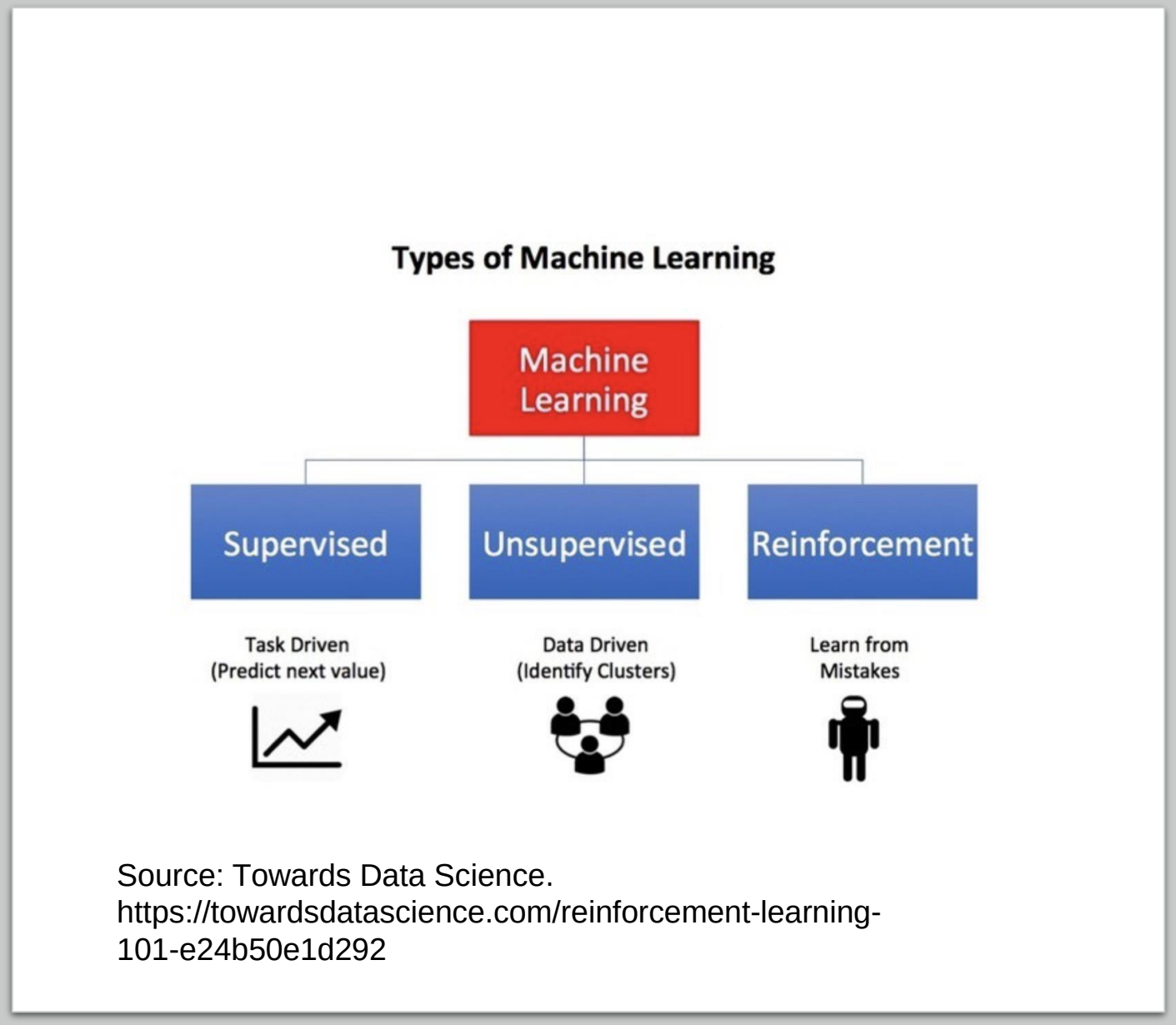
Hierarchical clustering



Source: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Reinforcement learning

- Technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.



Source: Towards Data Science. https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

Deep Learning

- Based on artificial neural networks, inspired by biological systems.
- Learning can be supervised, semi-supervised or unsupervised.
- The adjective "deep" in deep learning refers to the use of multiple layers in the network.

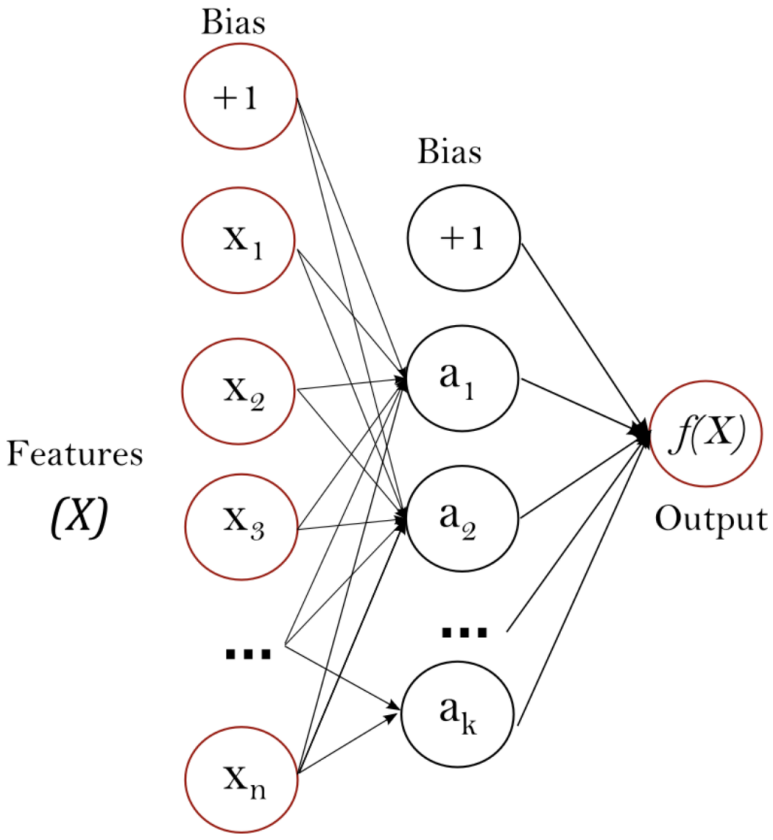


Figure 1 : One hidden layer MLP.

Neural network with one hidden layer

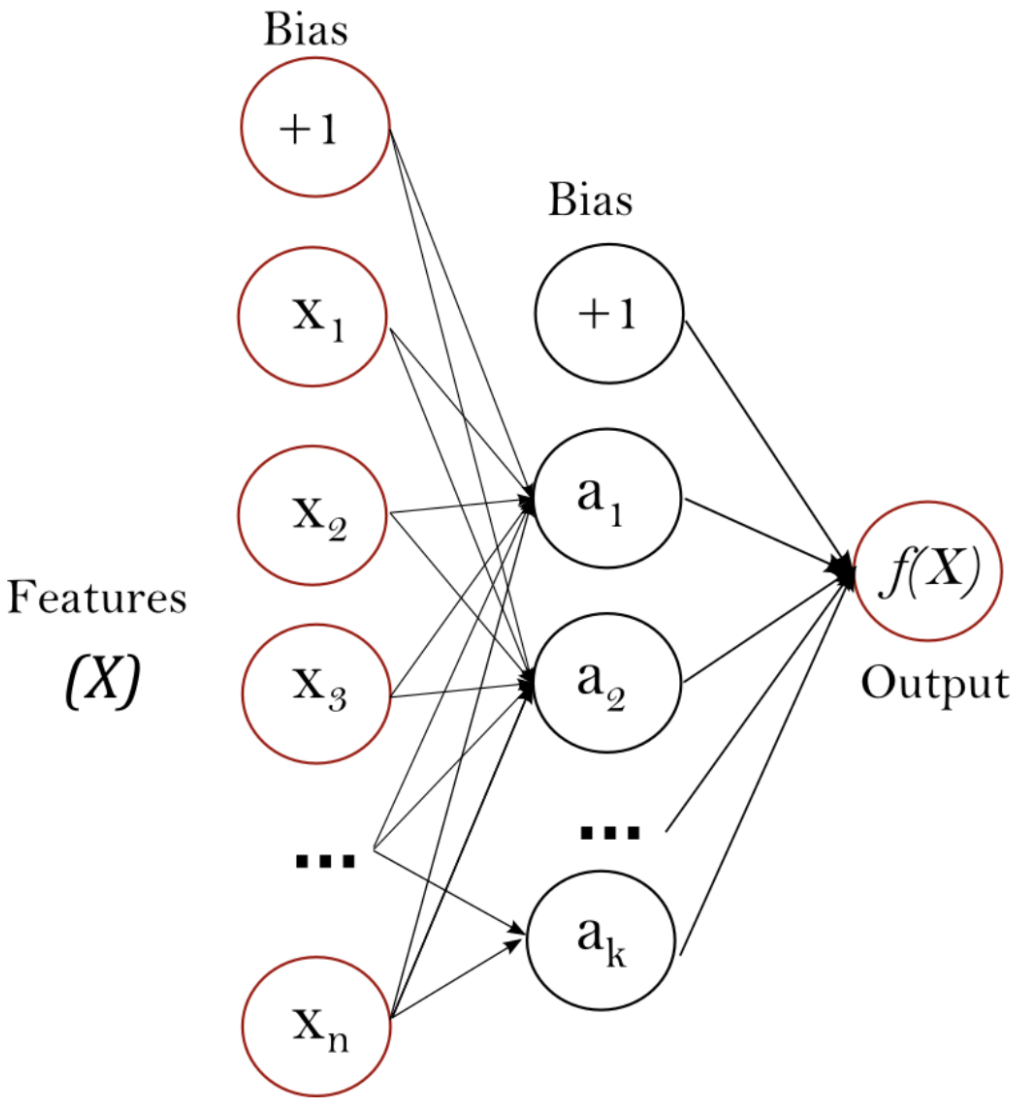
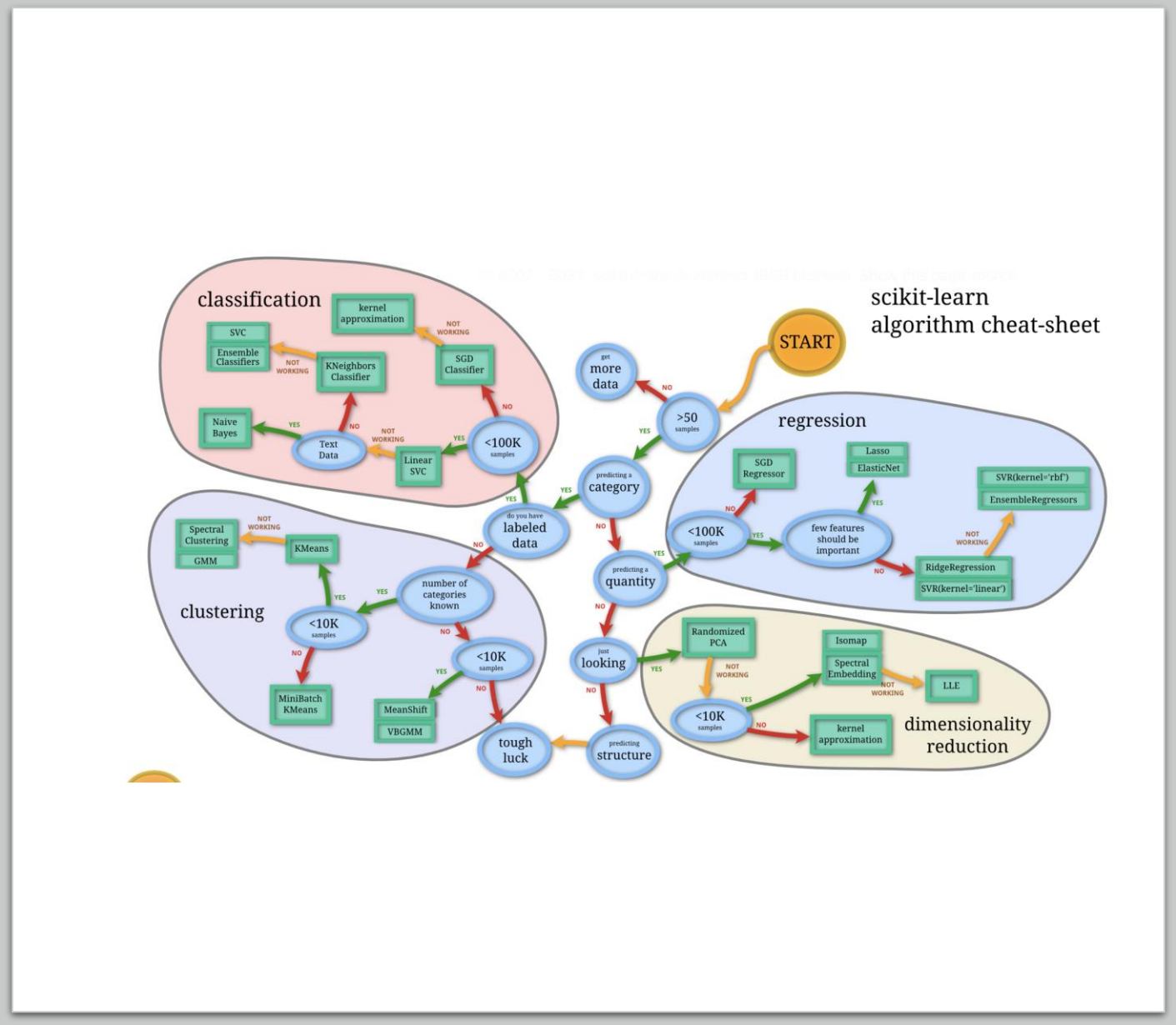


Figure 1 : One hidden layer MLP.

Source: https://scikit-learn.org/stable/modules/neural_networks_supervised.html#neural-networks-supervised

Scikit-learn

- https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html



Sources and additional readings

- https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
- https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
- https://dzone.com/articles/decision-trees-vs-clustering-algorithms-vs-linear
- Data Mining: Practical Machine Learning Tools and Techniques https://www.cs.waikato.ac.nz/ml/weka/book.html

It’s your turn

1) Think of **use cases** where classfication could be applied.
2) Think of **use cases** where clustering could be used.
3) Pick one of the cases you discussed. What kind of data would you use for training your model? How about testing?