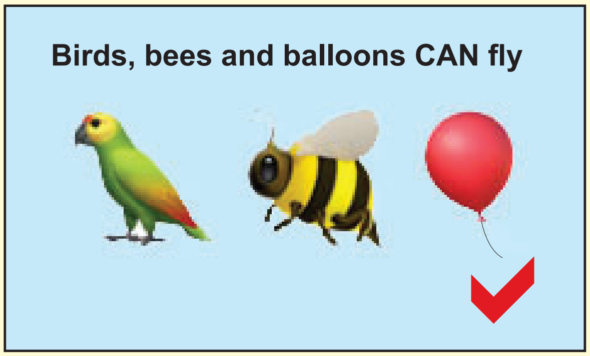
Symbolic Algorithms

The symbolistic approach to Machine Learning is based on the representation of objects with symbols that can be manipulated logically, very much like we deal with variables in mathematical equations. In this respect, symbolists are particularly interested in how infants learn to recognize and classify objects during the first 18 months of their life. During this time, we learn that we are situated in an environment that is made up of objects that persist over time. We appear to spontaneously establish object categories (i.e., clusters) and gain some understanding of what different categories of objects can and cannot do. For example, dolls and teddy bears cannot fly while birds, bees and balloons can fly.



The ability to cluster is fundamental to human intelligence (i.e., biological intelligence) and is often the first step in acquiring knowledge. A cluster is a group of physical objects or non-physical entities that have similar characteristics, or are at least more similar to each other than members of other clusters. The quest for an algorithm that can automatically group together entities into clusters is an intensely pursued area of research in Machine Learning. A cluster typically has a prototypical set of attributes such as an average height or weight of a person in a cluster of people, even though none of its members may be of that exact height or weight. Or, the desirable product specifications of an abstract cluster such as a market segment or potential consumer group. Without any preexisting knowledge a clustering algorithm will need to start by assuming that each new entity is part of a separate cluster unless it is in some ways similar to the entities in an existing cluster. With a computer being able to perform millions of computations per second such a clustering algorithm is able to fairly rapidly group thousands of entities into a hierarchical structure of clusters and sub-clusters based on some rules that define the desired degree of similarity that the members of a sub-cluster should adhere to.

The *k-means* learning algorithm is used to cluster unlabeled data (i.e., data without defined categories). It is an unsupervised clustering algorithm that aims to partition cases into *k* clusters with a case being assigned to the cluster with the nearest mean. If *k* is predetermined then the greater the value of *k* the smaller the value of the metric that defines the differences between clusters. One of the metrics that is commonly used to compare results across different values of *k* is the mean distance between data points and their cluster centroid. While the *k-means* algorithm is simple and popular, it has at least one serious hurdle to overcome. It works only if the clusters are well differentiated. This hurdle can be overcome to some extent by providing external assistance. For example, by training *k-means* with the attributes of members of existing clusters and the probability associated with each attribute. Another approach is to reduce the number of possible similarity dimensions through a process of dimensionality reduction. In the case of clustering images, we can reduce the number of visible differences (i.e., dimensions) at the pixel level to a much smaller number of combined features. For example, with only 10 choices for each facial feature, a law enforcement artist can draw a portrait of a suspect that is often good enough to recognize that suspect.