## Machine learning: the problem setting

In general, a learning problem considers a set of $n$ samples of data and try to predict properties of unknown data. If each sample is more than a single number, and for instance a multi-dimensional entry (aka multivariate data), is it said to have several variables, also known as attributes or ***features***. We can separate learning problems in a few large categories:

- **Supervised learning**, in which the data comes with additional attributes that we want to predict.
  This problem can be either:

  **Classification:** samples belong to two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data.
  An example of classification problem would be the digit recognition example, in which the aim is to assign each input vector to one of a finite number of discrete categories.

  **Regression:** if the desired output consists of one or more continuous variables, then the task is called regression.
  An example of a regression problem would be the prediction of the weight of a pony as a function of its age and height.

- **Unsupervised learning**, in which the training data consists of a set of input vectors $x$ without any corresponding target values.
  The goal in such problems may be

  – to discover groups of similar examples within the data, where it is called ***clustering***,
  – to determine the distribution of data within the input space, known as ***density estimation***,
  – to project the data from a high-dimensional space down to two or thee dimensions for the purpose of visualization