Why we prepare data?

* Some data preparation is needed for all mining tools.
* The purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool.
* Error prediction rate should be lower (or the same) after the preparation as before it.
* Preparing data also prepares the miner so that when using prepared data the miner produces better models, faster.
* GIGO - good data is a prerequisite for producing effective models of any type
* Data need to be formatted for a given software tool and to be made adequate for a given method.

Data mining method may be categorised as either supervised or unsupervised. **Supervised learning** is when we already know what our correct output should look like having the idea that there is a relationship between input and output. E.g Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem. We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two discrete categories. On the other hand **Unsupervised learning** allows us to approach problems with little or no idea what our results should look like . We can derive the structure from the data where we don’t know the effect of the variables.

When we are going to start with a supervised data we generally divide the whole dataset into three parts. Training set, cross-validation set, test set with a percentage of 60,20,20 respectively of the whole dataset (for big data analysis this ratio is 99,0.5,0.5). We train the data in the training set, try different model architecture to evaluate them on cross-validation set and see which of many different models performs the best on this set and when I have got a final model evaluate it on test set in order to get an unbiased estimate of how algorithm is doing.

Dealing with the data sometimes we may come up with overfitting and underfitting problem. **Overfittting** can be defined as a tendency to assume all data, including isolated ones, are completely true. This leads to missing out the real pattern, and making future decisions based on this assumption **Underfitting** is when the form of our hypothesis function maps poorly to the trend of the data . It is usually caused by a function that is too simple .

 Imagine you’re planning to buy a car. You do an extensive research on the internet, trying to finalize what car to buy. Then a friend of yours gives you an opinion on car A. That opinion may not very well align with other information you gathered from the internet, and you might not even try to substantiate it by searching on the internet. But you will probably give it quite a bit of importance while making your decision. This situation is similar to the definition of overfitting above.

