Original Reference: Golub T.R, Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:532-537.

The data set **leukaemia_886.xls** contains descriptions, accession numbers and expression in samples of blood or bone marrow of 72 leukemia patients, for 886 genes.

Expression levels were recorded with affymetrix chips (one hybridization for each patient/sample), so the data contains also columns of affymetrix “calls” (**P**resent, **M**edium, **A**bsent) for each gene.

Expression levels were normalized through a simple linear regression, using an arbitrarily selected chip as reference, and there are no missing values in the subset of the data that you are considering here.

The 72 patients are classified according to diagnosed type of leukaemia, and the data contains also rows with classification information for each patient. In particular, there is a binary classification into Acute Lymphoblastic Leukaemia (**ALL**); and Acute Myeloid Leukaemia (**AML**). Following this, there is also a classification of each of these main types in subtypes (ALL **T**-cell and **B**-cell; AML M1, M2 etc.)

The data does not comprise all the thousands of genes originally represented on the chips, but only a subset of 886. These were “filtered” with a simple variability criterion (considering the maximum and minimum expression level of each gene across the samples, both the difference max-min and the ratio max/min, were thresholded – see last columns in the data set).