# Problem Statement - Gene Data Case Study

## Context

RNA sequencing (RNAseq) is one of the most commonly used techniques in life sciences and has been widely used in cancer research, drug development, and cancer diagnosis and prognosis. Sequencing the coding regions or the whole cancer transcriptome can provide valuable information about gene expression changes in tumors. Cancer RNA-Seq enables the detection of strand-specific information, an important component of gene regulation. Cancer transcriptome sequencing captures both coding and non-coding RNA and provides strand orientation for a complete view of expression dynamics.

## Objective

To use the Principal Component Analysis (PCA) technique to transform a large set of variables into a smaller one that still contains most of the information in the large set.

## Key Concepts

Applying PCA for dimensionality reduction
Visualizing the data in the lower dimension

## Data Description

This collection of data is part of the RNA-Seq (HiSeq) PANCAN data set, it is a random extraction of gene expressions of patients having different types of tumor:

BRCA
KIRC
COAD
LUAD
PRAD

Samples (instances) are stored row-wise. Variables (attributes) of each sample are RNA-Seq gene expression levels measured by illumina HiSeq platform.

A dummy name (gene_X) is given to each attribute.