**Principal Component Analysis for Dimensionality Reduction**

In the modern age of technology, increasing amounts of data are produced and collected. In machine learning, however, too much data can be a bad thing. At a certain point, more features or dimensions can decrease a model’s accuracy since there is more data that needs to be generalized — this is known as the **curse of dimensionality**.

**Dimensionality reduction** is way to reduce the complexity of a model and avoid overfitting. There are two main categories of dimensionality reduction: feature selection and feature extraction. Via feature selection, we select a subset of the original features, whereas in feature extraction, we derive information from the feature set to construct a new feature subspace.

In this tutorial we will explore feature extraction. In practice, feature extraction is not only used to improve storage space or the computational efficiency of the learning algorithm, but can also improve the predictive performance by reducing the curse of dimensionality — especially if we are working with non-regularized models.

Specifically, we will discuss the **Principal Component Analysis** (**PCA**) algorithm used to compress a dataset onto a lower-dimensional feature subspace with the goal of maintaining most of the relevant information. We will explore:

* The concepts and mathematics behind PCA
* How to execute PCA step-by-step from scratch using Python
* How to execute PCA using the Python library scikit-learn

# Introduction to Principal Component Analysis

**Principal Component Analysis** (**PCA**) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction. Other popular applications of PCA include exploratory data analyses and de-noising of signals in stock market trading, and the analysis of genome data and gene expression levels in the field of bioinformatics.

PCA helps us to identify patterns in data based on the correlation between features. In a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.