The rapid development of information technology has resulted in massive growth in the amount of digital data. There is a variety of data from multiple sources, such as websites, e-mails, social networks and services (Google+, Facebook, Twitter, forums, blogs, etc.), mobile networks, media, transaction systems, digital libraries, sensor networks, etc. It is estimated that by 2020 digital data on the Internet will exceed the amount of 40 zettabytes. When the data is characterized by high volume, diversity, complexity, variability and fast temporal flow, it is referred to as Big Data (BD).

Collection and storage of such large amounts of data is currently not very difficult, because it can be done with distributed and cloud storage. However, the challenge in BD is to analyze, search, extract the desired information, visualize, and process in real-time such datasets. The standard methods of data analysis and information technologies are not able to process such huge amounts of data in reasonable time. There is therefore a need to develop new technological solutions that will be able to meet the demands of the modern digital world.

This project aims to develop advanced tools for processing and analyzing BD of specific properties, such as the nonnegativity. Such data represents a considerable portion of all digital data. These include, e.g. a collection of images and videos, spectral signals, textual documents, various biomedical data, etc. For searching and extracting the desired information from such huge data, it is necessary to reduce its volume as well as to extract semantic features that represent the analyzed data. To accomplish these tasks, it is assumed that the original data can be transformed to the algebraic objects, such as vectors, matrices, and tensors. For example, each image can be expressed by a matrix of non-negative elements. Similarly, a text document can be represented by a vector of word occurrence in the document. In the geometrical approach, each vector in the data space generates a point. A set of vectors thus forms a spatial cloud of points, which for non-negative data forms a convex geometrical object. Distribution of such points is typically irregular, where some characteristic structures or groups of points can be observed. Recognition and modeling of such structures is the subject of the mentioned feature extraction. In the algebraic approach, these structures can be described by a mathematical model representing the analyzed data. Estimation of the model related parameters is not an easy task, but it can be achieved with scalable numerical algorithms that will be investigated in this project. The aim is to find the algorithms with fast convergence, a low computational cost, and low demand for memory resources. In particular, they should be easy to implement in parallel and distributed computing environments. The experiments will be carried out using the BD analytics frameworks, such as the Apache Hadoop with the MapReduce model and the Apache Spark. The features of interest are planned to be used for solving many other machine learning and artificial intelligence problems, including clustering and various types of classification, e.g. extreme (with a large number of classes) or dynamic (with time-varying features).

The project will provide new algorithms that will certainly contribute to the further development of machine learning and artificial intelligence methods. The computational tools that will be created in the project can noticeably improve the modern information technologies, which so greatly contribute to the development of civilization.