Operations of the most of today's companies, to lesser or greater extent are based on the use of information systems. This applies to many aspects of a business, including management, administration, bookkeeping or customer service. In addition to using the collected data for the current activities of the company, more and more enterprises are interested in using historical data for optimization of their business. For example, analytical models created using data from the past can be used to improve the service by adapting to customer needs or to analysis of the exceptional and dangerous events and, hence, improving safety of the personnel. Particularly noteworthy are the intelligent systems which, at the same time, utilize predictive analytics methods and require, so called, high availability i.e.: monitoring and early warnings systems responsible for threat detection, ecommerce and recommendation systems or applications that support medical diagnostics. In this group of systems the data are usually collected online during system operation, hence, due to various reasons, may be partially unavailable – what is one of the main objectives of this project.

Data exploration techniques allow analysts to discover interesting dependencies in data due to a fact that it gives the ability to efficiently verify current hypotheses about investigated phenomena and formulate new ones. In practice, this is usually done by conducting simple tests on available data and using results of those test in consecutive stages of the data exploration process. Very often, the most laborious task of an analyst is to define such a representation of objects described in the data, that in future will be the most useful for, e.g. constructing prediction models. Unfortunately, even though there are plenty methods for automatic feature selection that are well-described in literature, it is hard to find methods and algorithms which would take into account both the quality and relevance of selected attributes subset as well as the risk of loss or lack of data during the long term operation of the prediction model.

With a rapidly growing availability of data, the notion of data processing and exploration with a hope of discovering useful knowledge is an intensively developed field of research. In order to be able to perceive comprehensible notions in data it is necessary to appropriately select attributes that describe them. Typically, in machine learning theory this task is performed using feature selection algorithms. Feature extraction is a process that involves the transformation of the raw data to a set of derived attributes, which are appropriately profiled to the analyzed problem. Logically, there are two phases of this process: the first is the construction of the new attributes based on original data (sometimes referred to as feature engineering), the second is a selection of the most important and relevant among the obtained attributes. There are three major approaches to the assessment of the importance of considered features: the filter methods which carry out the attributes selection regardless of the chosen model, the wrapper methods that make a selection of attributes based on the results of preliminary data analysis and embedded methods which are nested in machine learning algorithms. For instance, one of the areas in which a large emphasis is placed on techniques of determining the optimal subset of relevant attributes is the rough set theory.

Feature extraction not only simplifies the obtained data representation, but also allows to acquire features that can be easily utilized by both analysts and learning algorithms. In the literature there are described many well-known and widely used techniques for selecting informative attribute subsets. The vast majority of the feature selection methods, that are thoroughly investigated in the literature, are focused on achieving the possibly small data representation well describing a studied problem. However, almost none of the available methods takes into account the fact that in real life, data may be lost or temporarily unavailable for the analysis.

One of the major challenges that we want to address in the scope of the proposed project is to provide controlled and governed redundancy in the feature selection process in order immunize attribute subsets against loss or temporal unavailability of part of the data and, thus, to increase the robustness of the intelligent systems. Whereby, data analysts could achieve a relatively small representation of original data tuned for the investigated problem and the selected attribute set should preserve its relevance to the problem even in the case of partial data loss. Other scientific objectives posed for the participants of the project are medications of selected machine learning algorithms to use of the newly developed 'redundant' subsets of attributes, analysis of the problem form the Big Data perspective and empirical evaluation of the developed methods. The planned study will also consider the problem of efficient quality assessment of the constructed attribute subsets, without a necessity to conduct complex computations like, e.g., training multiple classifiers on different feature subsets and testing them on additional validation data. A notable aspect and interesting research fields is also the impact analysis of the of newly developed feature selection methods on the quality of prediction and the robustness of the operation of predictive models in case of concept drift in data.