International Science Group
ISG-KONF.COM

THE INFORMATION EVOLUTION OF THINKING: FROM DATA TO UNDERSTANDING.

DATA MINING IN SCIENTIFIC COGNITION.USE OF DATA MINING IN SCIENCE AND IN ECONOMICS (BUSINESS)

ISBN 979-8-89940-604-1
DOI 10.46299/979-8-89940-604-1

**Polyakov M. V., Khanin I. G., Shevchenko G. Ya.,
Bilozubenko V. S., Marchenko O.A.**

THE INFORMATION EVOLUTION OF THINKING: FROM DATA TO UNDERSTANDING.

DATA MINING IN SCIENTIFIC COGNITION.USE OF DATA MINING IN SCIENCE AND IN ECONOMICS (BUSINESS)

Monograph

2025

UDC 330.34

Authors:

Polyakov Maxim V. – Doctor of Sciences (Economics), Associate Professor, Co-founder, NGO "Noosphere Association", ORCID: http://orcid.org/0000-0001-7896-2486

Khanin Igor H. – Doctor of Sciences (Economics), Professor, Professor of the Department of Enterprise Economics and International Business, National University of Water and Environmental Engineering, ORCID: http://orcid.org/0000-0002-4221-2314

Shevchenko Gennadij Ya. – Candidate of Sciences (Engineering), Associate Professor, Partner, NGO "Noosphere Association", ORCID: https://orcid.org/0000-0003-3984-9266

Bilozubenko Volodymyr S. – Doctor of Sciences (Economics), Professor, Professor of the Department of International Economic Relations, University of Customs and Finance, ORCID: http://orcid.org/0000-0003-1269-7207

Marchenko Oleg A., – magistr, scientist, Association Noosphere, ORCID: https://orcid.org/0000-0001-7665-7832

Editor:

Ivanova, Svitlana A. – Candidate of Philological Sciences (Ph. D.), Associate Professor, Associate Professor of the Department of Advertising and Public Relations, ORCID: 0000-0002-9065-8687

Reviewer:

Maxim Korneyev – Doctor of Economic Science, Professor, Department of International Economic Relations, University of Customs and Finance, Dnipro, Ukraine, ORCID: 0000-0002-4005-5335

Polyakov M. V., Khanin I. G., Shevchenko G. Ya., Bilozubenko V. S., Marchenko O.A. The information evolution of thinking: from data to understanding. Data mining in scientific cognition. Use of data mining in science and in economics (business). – Primedia eLaunch, Boston, USA, 2025. – 122 p.

Library of Congress Cataloging-in-Publication Data

ISBN – 979-8-89940-604-1
DOI – 10.46299/979-8-89940-604-1

UDC 330.34

ISBN – 979-8-89940-604-1

INTRODUCTION

This paper explores the application of data mining methods (data mining) in various industries – from scientific research to business analytics. The first part discusses the theoretical foundations of data mining, including the stages of data processing, classification, clustering and association analysis methods. The importance of quality data preparation and selection of appropriate algorithms for obtaining reliable conclusions is emphasized.

Special attention is paid to the comparative analysis of data mining applications in science and economics. In the scientific sphere the emphasis is placed on identifying patterns, supporting hypotheses and creating models of complex systems. In business, data mining is used as a tool for improving efficiency, decision making and predicting customer behavior.

The paper illustrates the interdisciplinary nature of data mining and demonstrates its potential as a universal tool for extracting valuable information from large data sets, facilitating informed decision-making under conditions of uncertainty.

Approved for publication by “ASSOCIATION NOOSPHERE”, Ukraine (protocol No. 5 dated 15.05.2025)

TABLE OF CONTENT

"From educated guess to hypothesis and on to theory - this is the way of knowledge; from ignorance to knowledge, from uncertainty to truth - by means of the senses, reason, critical thinking and imagination."

E. Fromm

PART 1: USING DATA MINING IN SCIENCE FOR RESEARCH

Introduction

In 2024, the Nobel Prize in Physics was awarded to John Hopfield and Joffrey Hinton for their developments in machine learning neural networks. Hopfield created an associative memory capable of reconstructing images in data. Hinton developed a method for autonomous identification of features in data. This is a testament to the importance to world science of research in this area

All our scientific ideas about the world of nature, society and technology, our knowledge about ourselves, about thinking and its regularities have a modelling character. Also laws, theories, scientific pictures of the world are model constructs (Glinsky, 1965; Neuimin, 1984). This was well expressed many years ago by the outstanding scientist N. Wiener: "The purpose and result of scientific research is to achieve understanding and control over some part of the Universe. No part of the universe is so simple that it can be understood and controlled without abstraction. Abstraction is the replacement of the part of the universe in question, by some model of it, a model of a similar but simpler structure. Thus, the construction of formal, or ideal 'mental' models on the one hand, and material models on the other, is by necessity central to the procedure of any scientific enquiry" (Rosenbluth& Wiener, 1984). That is, any research is a process of model building. In this context, modelling blurs the boundaries between the humanities and the so-called "exact" sciences. Both mathematics and natural sciences operate with models of the reality under study, which makes them essentially similar to humanities disciplines. Thanks to modelling, measurement (for qualitative characteristics - quantification) and experimentation (simulation) are introduced into the research process, which expands the arsenal of

scientific methods. This includes mathematical modelling, mental experimentation as well as prospective ideal reconstructions and simulation models. In selecting and building a model, one can move in three different directions:

• to start from experimental data - let's call such models Mexp. These include the "black box" models proposed by Norbert Wiener, which assume that we have no theoretical considerations to build a model a model. This is the model used in DM and Big Data, only without a name.

• to start from the question "Why?" - let us call such models Mcause. This is how Y.P.Adler writes about such models (Adler & Speer, 2019): "This is the main question for the answer to which scientific research has been conducted in the past. Such a question requires the assumption of the existence of a causal relationship between the phenomena under study. If we believe in causality, then it is natural to realise the process: putting forward a hypothesis - testing the hypothesis by all available means - putting forward a new hypothesis, because the previous one is rarely acceptable. It is long, expensive and inefficient". Academician N.N.Moiseev makes a similar statement, emphasising that: "...the basis of any model, including a mathematical one, is phenomenological. This means that no matter how abstract the model is, the history of its creation always begins with the experimental study of the phenomenon" (Moiseev, 1920, p.p.20).

• Models based on principles that are not directly derived from experience, but explain and structure its results - let us call them Mintel. Such principles include, for example, conservation laws, the second law of thermodynamics and other fundamental concepts. Their existence is confirmed, in particular, in the works of Philip Frank (Frank, 2010).

Moreover, since in modern society mediated knowledge, with the help of mediators, becomes predominant, the study of indirect ways of knowing the world, knowledge through modeling systems, acquires particular relevance.

Modelling thus acts as a way to overcome the gap between empirical data and abstract principles.

The main purpose of any scientific research is to gain knowledge and achieve

understanding of the investigated area of reality. This is quite clearly stated in the work (Zakrevskiy, 1988): "Identification of regularities in the flow of data is the main way of scientific cognition of the reality around us. And it should be said that it is extremely labour-intensive, even if one predicts in advance which values are connected by the sought regularity". However, the following remark concerning the difference between law and regularity is important for further presentation - by law we will understand a generally binding rule, something that is recognised as obligatory, and by regularity we usually understand something less strict than law, for example, connection and interdependence of some phenomena of objective reality.

In fact, the result of any scientific research is the answers to the questions "What", "How", "Why". Nowadays, the research of complex objects, often at the intersection of sciences or subject areas, which are characterised by multidimensionality and diversity of attributes describing them, which requires the involvement of modern methods of their analysis, mostly related to the so-called Intelligent Data Analysis (IDA) or Data Mining (DM), in Western terminology . [1]

The purpose of using Data Mining techniques in scientific research is to extract hidden patterns, structures and insights from large and complex datasets to gain new insights, confirm hypotheses and support decision-making. It solves a wide range of research problems, improves the quality of inference and increases the accuracy of predictions. ***In scientific research, DM addresses the following main challenges:***

1. ***Identifying hidden patterns and structures***[2]

- *DM helps to discover non-obvious, hidden patterns, relationships and relationships in data that are difficult to detect by conventional means.*
- *Working with big data allows you to take more factors into account and eliminate the impact of random errors.*
- *Consideration of different groups and categories, making the conclusions more universal than with traditional methods of analysis.*

[1] Below, the terms IAD and DM will be used equally, as synonyms, depending on the context

[2] https://habr.com/ru/articles/784060/

***Examples**:*

- *Clustering of the population by income levels and consumer preferences for the development of economic models*
- *In political science, DM helps to analyse public opinion based on large social media data sets.*

2. *Forecasting and modelling, improving forecast accuracy*[3]

- *The main objective is to create models capable of predicting the behaviour of objects or systems based on historical data, and for predicting future events.*
- *Building predictive models for complex systems where traditional methods of analysis are less effective.*
- *Consideration of multiple factors and their interactions.*

***Examples**:*

- *Time-series forecasting of labour market dynamics or GDP changes*
- *In medical research, DM is used to predict treatment outcomes based on patient data.*

3. *Accelerating big data processing and optimising the research process*[4]

- *Automate data analysis, which reduces study time.*
- *Reducing the risk of human error in data processing.*
- *Application of efficient algorithms to deal with large amounts of data.*
- *Data analysis helps to improve and optimise scientific processes, making them more efficient and accurate*

***Examples**:*

- *Analysing millions of transactions to identify patterns of consumer behaviour.*
- *In economics, DM can be used to automatically search for trends in macroeconomic indicators.*

4. *Testing hypotheses and improving scientific methods (*Hypothesis Validation)

- *Testing existing scientific hypotheses by analysing data.*

[3] https://ru.wikipedia.org/wiki/Data_mining

[4] https://www.sciencehunter.net/Blog/story/vozmozhnosti-data-mining-kak-instrumenta-poznaniya-metodologicheskie-aspektyi

- *Using DM models to compare and validate theoretical conclusions.*

***Examples**:*

- *Assessing the impact of social factors on educational attainment through big data analysis*
- *In sociology, DM helps to identify correlations between socio-economic factors and the behaviour of particular groups of people.*

5. *Decision support*[5]

Based on the knowledge extracted from the data, researchers can propose solutions to real-world problems

- *Creating models for forecasting to help make informed decisions.*
- *Assessing the likelihood of certain scenarios or events.*
- *Developing data-driven recommender systems.*
- *The results of date-mining can be used to support scientific hypotheses and make informed decisions in research activities*

***Examples**:*

- *Developing recommendations for public policy based on analyses of poverty and resource allocation.*
- *In environmental studies, DM is used to predict climate change based on historical data.*

6. *Generation of new knowledge*

- *Forming new theories and scientific hypotheses based on data analysis. Identification of hidden patterns that are not obvious in traditional analyses.*
- *Detecting complex relationships between variables.*

***Examples**:*

- *In biology, the identification of genes associated with certain diseases based on the analysis of genetic data.*
- *In bioinformatics, the use of DM enables the identification of new markers for disease diagnosis.*

[5] https://ru.wikipedia.org/wiki/Data_mining

7. Creation of multidimensional models

- DM facilitates working with multiple factors, their interrelationships and dynamics, which is particularly useful in interdisciplinary research.
  Example:
  - Modelling the impact of climate change on the economy.

8. *Personalisation of findings, decisions and approaches*

- *DM allows scientific recommendations to be tailored to specific contexts or groups.*
- *Personalised analysis, e.g. personalised treatment in medicine or personalisation of marketing strategies in economics.*

Examples*:*

- *Creating personalised educational programmes based on the analysis of student behaviour*
- *In marketing, DM is used to analyse customer flows and tailor bonus programs.*

The overall benefits of applying Data Mining:

- Improving the objectivity of the analysis by minimising human bias.
- Automation of routine data processing tasks.
- Improved accuracy of results through the use of advanced analysis algorithms.

The goal of using DM techniques in scientific research is not just to process data, but to turn it into a tool for discovering new things, confirming hypotheses and improving the accuracy of conclusions, as well as making scientific research applicable to real-world problems. Data Mining methods allow scientists to cope with large amounts of data and complexity of relationships, which is especially important in interdisciplinary research.

On the other hand, we can talk about the key areas and applications of Data Mining (data mining):

1. *Hypothesis generation*

Data Mining allows you to analyse massive amounts of data to identify correlations and patterns that are difficult to spot using traditional methods. This is particularly useful in studies where the volume of data is large, for example:

- *In bioinformatics to study genomic sequences.*
- *In sociology to analyse behavioural patterns.*

2. Forecasting

Data Mining techniques such as regression models, decision trees and neural networks are used to predict various indicators:

- *Economics: forecasting demand, prices or financial trends.*
- *Ecology: predicting climate change or species distribution.*
- *Medicine: predicting patient outcomes.*

3. Classification and clustering

- ***Classification:*** *Used to categorise objects into predetermined categories. For example, in medical research to determine a patient's risk of disease.*
- ***Clustering:*** *Groups objects based on their similarities, which is useful for discovering new types of objects or phenomena (e.g., in astronomy to study galaxies).*

4. Optimisation of experiments

Data Mining helps to improve the design of experiments by selecting the most relevant variables for analysis and minimising noise in the data.

5. Analysing texts and publications

- *In linguistics to study semantics and syntax.*
- *In analysing scientific publications to identify trends in research activities.*

6. Modelling complex systems

Data Mining is used to build models of complex systems:

- *In physics to analyse the dynamics of particles.*
- *In biology to model interactions in ecosystems or within cells.*

Benefits of using Data Mining:

- *Automating big data analytics.*
- *Ability to* ***generate new knowledge*** *from existing data.*
- *Multidisciplinary, as Data Mining approaches can be adapted to a wide variety of fields.*

Tools such as Python (Pandas libraries, Scikit-learn, TensorFlow), R, and specialised platforms (e.g. KNIME, RapidMiner) allow Data Mining techniques to be integrated into scientific research and produce high quality results.

At the same time, DM is actively used for business purposes. The practice of using DM has allowed us to highlight an important fact - it is necessary to distinguish between the so-called academic research, which, most often, is conducted on fixed historical data and the purpose of most of which is to improve, as a rule, the architecture of DM models and, let's call them so - production research, the purpose of which is not academic interest, but to increase the success of business as a result of using the DM model. That is, the business is primarily interested in answering the first two questions - "What" and "How"? This is the case when knowledge is utilised.

However, there is also a third case, when knowledge is created - i.e. research is conducted, the purpose of which is to improve the quality of scientific research as a result of using the DM model, i.e. knowledge generation, not knowledge utilisation, and these are two big differences. It is a case of the direct use of DM in scientific enquiry aimed at cognition, with an emphasis on answering the question 'Why'? It is this fact that links DM to general questions of cognition. In this regard, let us briefly review them.

Analysing studies and publications

The first stage of cognition is experience, observation, experiment, study of phenomenon, in other words - accumulation of facts for further analysis. The second stage is generalisation of facts, selection of essential things in them, formation of hypotheses and conclusions on their basis, i.e. some abstraction from the first one. The third stage is the practical verification of hypotheses or conclusions obtained earlier. This is a universal scheme of cognition. It was formulated quite clearly and briefly by academician B. Gnedenko (Gnedenko, 1983): "Stage 1 - observation or live contemplation, Stage 2 - transition to abstraction, Stage 3 - verification of abstraction in practice. This is the dialectical path of cognition of truth...".

It should also be recalled that even at the earliest stages of research related to DM

(at that time the name pattern recognition (PR) was used), there were questions about the role of PR in cognition. In particular, L. Malinowski (Malinovsky, 1986) proposed a scheme of cognition based on the principles of pattern recognition and classification (Fig. 1).

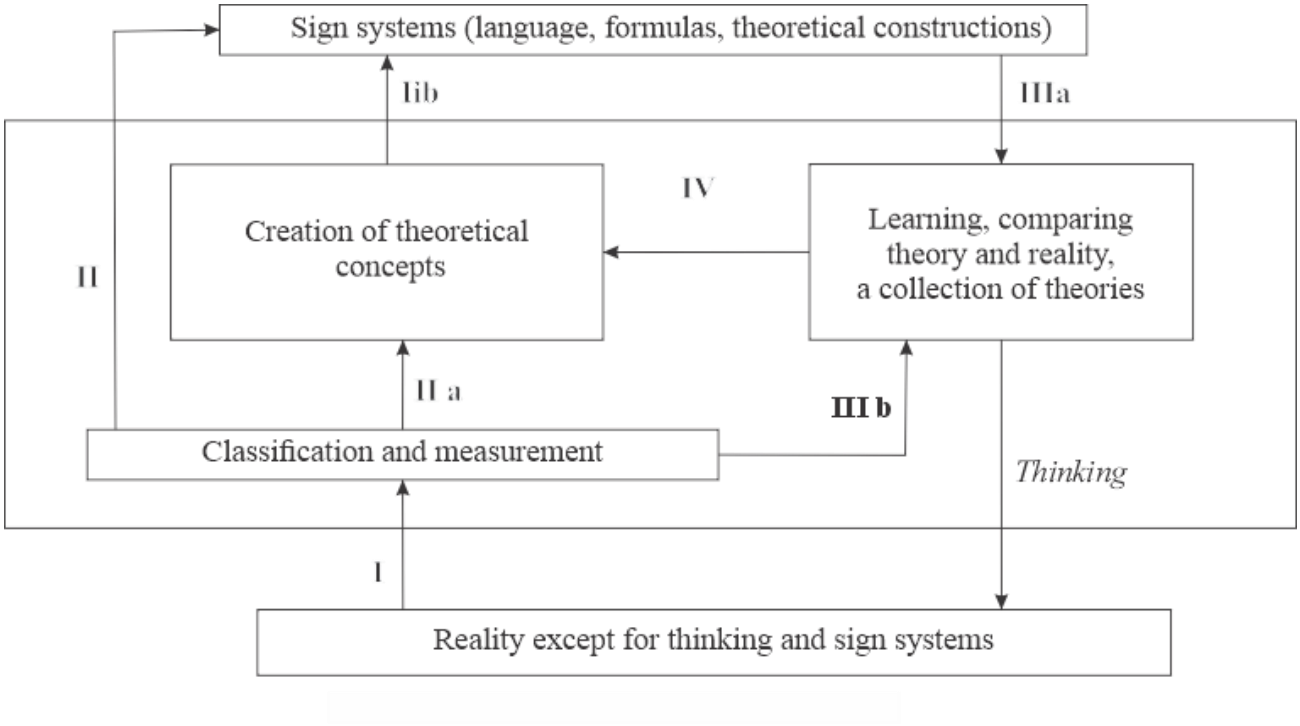


Figure 1. Relationship of thinking, reality and sign systems

** Source: Malinowski, 1986.*

At the beginning of its appearance, DM (or PR) was actively used for scientific purposes to solve applied problems. Suffice it to recall the works of (Jurs & Eisenauer, 1977), one of the first monographs on automation of scientific research processing in chemistry, as well as (Geology and mathematics, 1970; Jurs & Eisenauer, 1977), where the problems of research automation in geology, etc. were solved. Then there was a shift of interest to business tasks, where DM was successfully used to solve numerous problems of marketing, commerce, finance, production under conditions of multidimensional and different types of data, including large volume (Bu Daher et.al, 2018; Gupta et.al, 2020; Kuchev, Paklin & Oreshkov ,Rastegari & Md. Sap;).

Nevertheless, the issue of DM application in scientific research has recently become relevant again. In support of this, we can cite the works (Aryal, 2023; Hullman, 2021; Waters,2023), in which it was the use of a new base of DM tools with increased capabilities that allowed to perform a number of major scientific studies.

Thus (Waters, 2022) argues that decades of investment, combined with the efforts of researchers, technologists, librarians, archivists and their institutions, have resulted in a digital infrastructure in the humanities that can support research workflows. The article focuses on key advances in epigraphy and palaeography, and highlights related work in Egyptology, the ancient Near East and medieval studios. However, it recognises that infrastructure capabilities are unevenly distributed and work needs to be done to improve usability and smooth transitions between workflows.

As the review (Németh & Koltai, 2021) shows, there are new opportunities for sociological research that are in a sense by-products of computer science. Methods that can be applied to specific sociological problems outside business applications are presented, namely sociological topics that have not yet been studied in the field, and new views on classical sociological issues are shown.

The editorial (Leitgöb & Prandner, 2023) highlights the impact of the digital revolution on the social sciences, in particular on empirical sociology, from an epistemological, informational and analytical perspective. According to the thematic focus of the research topic, it centres on big data and machine learning, which are the two main elements of the emerging new interdisciplinary field of computational social sciences (CSS). The authors give their advice on how to improve the institutional system. The need to reform social science education and create a vast centralised data infrastructure is a priority.

In (Strachan & Stephen, 2007), the use of DM for analysing monitoring data was proposed and its application was demonstrated to identify useful knowledge from in-service circuit breaker winding data in the power industry. The results obtained formed the basis of decision support systems for assessing the condition of these equipment components.

The application of DM in the field of medicine, namely in cardiology is given in (Toledo, 2021). Diagnosis is fundamental to cardiac care. Computational algorithms can better analyse the available medical data of a large number of patients and improve the diagnosis mechanism. In this work, patient data from a Mexican cardiology institute is analysed using clustering. The results obtained give confidence that the group can

take advantage of computational methods to develop better strategies in cardiac diagnosis.

Also, the issue of using data mining and its importance in medical research is raised by (Khajehei & Etemady, 2010).

Paying attention to educational activities, it can be concluded that today universities generate not only graduates but also huge amounts of data from their systems. Building an information system that can learn from the data is a challenging task, but has been successfully solved using various DM approaches such as clustering, classification, predictive algorithms, etc. However, the utilisation of these algorithms with educational data is quite low (Algarni, 2016).

A striking example of DM application in the agricultural sphere is given in (Yethiraj, 2012). The authors reviewed the studies on the application of DM methods such as: IDA, k-means, k nearest neighbours, neural networks. It is emphasised that this approach in agriculture is relatively new for crop/animal forecasting/management.

As we can see, there is currently no consensus on the success and suitability of using DM methods for research purposes. The use of such systems in science is becoming widespread and versatile, and specialists from various fields - from physics and medicine to linguistics and geoinformatics - are actively using them for analysing data and generating hypotheses. However, this development raises questions about whether research is really accelerating, whether new knowledge is emerging, and how ethics, copyright, and transparency of academic relate to this process

In some cases, scientists are more categorical - in their opinion, neural networks and science are simply incompatible, as it is important for science to create new things, and neural networks are useless here. Therefore, neural networks will not change our world.

At the same time, one cannot ignore the multifactorial nature of the data obtained as a result of relevant research. In these circumstances, the use of DM is a necessity, as it is simply impossible to cover and analyse this kind of data in any other way.

However, despite the undoubted successes of research using DM tools the question remains - what do we learn as a result of using DM? What is meant is not a

specific result obtained empirically, but the result obtained in terms of a theory of cognition, because without such an answer and a clear understanding, further use of DM may lead to a false understanding of goal attainment and associated errors in both predictions and decision-making.

Therefore, despite the undoubted successes of research using DM tools, the question remains - what do we learn as a result of using DM? This does not refer to a specific result obtained empirically, but to the result obtained in terms of a theory of cognition, because without such an answer and a clear understanding, further use of DM may lead to a false understanding of goal attainment and associated errors in both predictions and decision-making.

Novelty

The scientific novelty of this study consists of the following:

a) It is substantiated that Data Mining (DM) methods represent a modern human-machine methodology for empirical cognition. It is found that these methods have limitations due to their ability to identify empirical regularities (ER) that express probabilistic knowledge, but no more. These patterns, at best, become the basis for the development of instructions, techniques or decision rules. However, they remain at the initial, empirical level of knowledge, which is characteristic of business applications.

b) It is shown that ER can be used as a starting point for formulating, testing and selecting hypotheses, which opens up the possibility of deeper investigation of the subject area and obtaining reliable knowledge. Such a process facilitates the transition from mere accumulation of knowledge to its comprehension, which brings the researcher closer to understanding the essence of the phenomenon under study.

c) It is emphasised that understanding is a key stage in identifying the fundamental principles underlying the knowledge gained. It may be at this stage that new methods of problem solving can be developed.

d) It is noted that knowledge and understanding are different categories: knowledge is a body of information, while understanding involves the ability to draw conclusions. DM methods in their current state are limited to the generation of

knowledge in the form of ERs and working hypotheses, but do not allow to reach the level of scientific method without the development of understanding. Thus, the limit of applicability of current DM tools is revealed.

Purpose of the article and research methodology

Purpose of the study. To clarify the foundations, possibilities and peculiarities of the application of DM for research purposes, as well as how exactly DM tools can be useful to the researcher and what are the characteristics of the results obtained with their help from the perspective of the theory of cognition. The answers to these questions are important not only from a theoretical and methodological point of view, but also from a practical point of view, as they allow us to clearly identify the limitations of all DM tools.

Research Methodology. In general, IDA refers to the field of data analysis based on the use of specialised mathematical methods to process relatively large sets of heterogeneous data. The aim is to identify previously unknown hidden patterns (relationships, trends, etc.) that can be interpreted and that may be useful for practical application and/or further study to gain new knowledge. It is worth noting that IDA is applied in cases when it is not possible to identify these patterns in large data sets using traditional analytical approaches, including statistical ones. As a rule, the initial information for the application of IDA is a table of experimental data, in which the results of observations of objects are recorded. The considered datasets are either an object-property table (OPT), in which objects are characterised by sets of certain properties (parameters, attributes) with corresponding values, or a so-called training sample (TS, dataset), which, in fact, is a verified OPT, in which each set of objects is assigned (marked) to a certain class. The large number of properties allows for a more complete and detailed characterisation of objects using a variety of data. It is clear that the hidden patterns that may be present in these tables will surely be of interest to businesses for various purposes. The methods that are used to find patterns in these kinds of tables are referred to IDA, within which we can distinguish the methods of machine learning, but this distinction is not important for the purposes of this article.

Given the potential of using IDA to study various objects and solve diverse structural and analytical problems in virtually all sectors and spheres of the economy, the study is based on the system approach. The methodology of this study lies at the intersection of mathematics, statistics, computer science, management theory, economics and business, and can be extended to other areas of science and practice, where IDA can be applied. Along with the use of general scientific methods of cognition (generalisation, systematisation, abstraction, induction and deduction, analysis and synthesis, analogy, comparison, formalisation, modelling, classification, categorisation) special methods of analysis (logical, structural, functional), descriptive method of research, interpretive methodology were also applied. The conceptual and guiding principles of this article are based on the concepts of theory of cognition, information technology (IT) and data analysis. In this regard, the peculiarities of their development as well as modern trends were taken into account.

Outline of the main research material and findings.

In these circumstances, it is also important to trace all the steps in the preparation of machine experimentation in a research study using DM methods - what is necessarily present in such a study? What exactly can the use of DM tools help the researcher and what are the results obtained using DM, of what quality, from the point of view of cognitive theory? The answers to these questions are not only important from a theoretical point of view, e.g. methodological, but they are also important from a purely practical point of view, as they directly point to the limitations inherent in all DM tools.

For example, at all stages of the scheme of cognition given above, various methods of cognition are used in one way or another, let us list them - from the simplest to the most complex (Shtoff, 1978)

- Methodology - the lowest level, examples - instructions, technical techniques, various kinds of empirics (any empirical technique that leads to a certain result);

- A scientific method based on the knowledge of relevant regularities, i.e. on the theory of a given subject area (examples - method of labelled atoms, conditional reflexes, method of questionnaires, etc.);
- General scientific method, a sufficiently general method of scientific research, the applicability of which transcends the boundaries of a particular scientific discipline and relies on the existence of regularities common to different fields (examples: analysis and synthesis, induction and deduction - for example, in physics, a general Ohm's law is deduced from a set of experiments with conductors (induction) and then used for calculations in specific circuits (deduction)).
- Methods applied in all sciences without exception, although in different forms and modifications. These are the most general methods of scientific cognition, and their study is the subject of philosophical methodology (examples: Dialectical method - the study of phenomena in their development, interrelation and contradictions, for example, in history the evolution of societies through the struggle of opposing classes (Marxism), logical method, etc.).

Correlation of the result obtained with the help of DM with the above classification allows us to give such an answer, i.e. to answer the question about the contribution of DM methods to the methodology of scientific cognition. And this is a necessary reference point, which indicates the level of cognitive value of the model obtained with the help of DM.

In general, the application of DM tools starts only when there are already prepared data, in the form of samples in which the objects are represented by multidimensional data sets (feature description). Such samples have different names - object-property table, training sample (TS) or experimental data table. In the following, we will consider them equivalent to each other. It is important to note that the issues of feature description (feature space) selection and data preprocessing are beyond the competence of DM, although they are present when solving any DM task - such issues are considered to be the competence of the subject matter expert with whom the DM analyst interacts.

Researchers wishing to apply DM methods in their research in practice sometimes

have a false sense of the absolute suitability of the approaches adopted in DM tasks and inherent in a purely industrial template to be used to their advantage.

Thus, the above and similar works, with a few exceptions, traditionally describe the construction of a particular model and its testing in academic or industrial terms, without touching upon a completely different layer of problems arising when trying to use DM methods in the interests of scientific research. It can only be stated that the works related to the practical use of DM methods need a different perspective on the solution of research problems, the difficulties encountered and the whole range of practical tasks on this way, especially those related to the comprehension of the problem itself from the broadest point of view: the choice of the feature space and the corresponding metrics, the collection and polishing of the TS, the choice of the DM method and the connection with the theory of cognition as the underlying methodology of the whole process of solving the problem with the help of DM, the indication of the place of DM in this theory, the correlation with the directions of the problem, the use of DM in the theory of cognition, and the use of DM in the research process. This is extremely important, because the clarifications related to these questions actually lay out a road map for the use of DM methods not only in research work, but also indicate their possibilities, limits and limitations of the methods, and their place in the general culture of research work.

DM methods differ from the tools of statistical data processing in that instead of testing the dependencies assumed by users in advance, they are able to find such dependencies, or, more precisely, regularities independently on the basis of available data and help the researcher to build on them more in-depth hypotheses about their nature. The most widely used DM methods are the following:

clustering - implements grouping of objects by maximising intra-group similarities and inter-group differences;

Decision tree classification - provides construction of a causal hierarchy of conditions leading to certain decisions;

association search - searches for stable combinations of elements in events or objects;

neural networks - used for visual pattern recognition, regression, classification.

Their advantages include the following features (Reshetnikova, 2021; Investopedia):

- it is possible to discover previously unknown, non-trivial and practically useful knowledge in the data.

- different types of problems such as prediction, risk and probability, recommendation, sequence search and clustering are possible.

- it is possible to visualise the results of data analysis in a visual form, which allows it to be used by people without special mathematical training.

The main objective is to identify relationships, or more precisely, cause and effect relationships, for the data collected in order to produce new knowledge.

In general, we are faced with the fact that in order to solve the task set by DM means, we need to conduct both data collection, data analytics, and, most likely, software model building (e.g., using the same Python programming language or choosing ready-made solutions).

At the same time, it is very important to pay attention to the following practical steps, confirmed by all the experience of successful use of DM tools in solving various tasks, including research tasks:

1. *Organise a careful selection of features.*
2. *The use of discrete scales is desirable.*
3. *Make an initial assessment of the quality of features and feature groups, their analysis, screening and filtering. Visualisation tools are particularly important here.*
4. *The organisation of the training sample must be accompanied by robust verification.*
5. *The choice of a suitable DM tool should be justified, especially paying attention to the interpretability of the solutions obtained.*

In fact, these steps provide the ordering, analysis and synthesis of facts and are preparatory to the main purpose of DM - the generation of ER, suitable preliminary hypotheses that allow the expansion of the range of possible hypotheses, while

realising a new essence of IT - the replacement or extension of human cognitive functions, something that replaces the known IT functions of computing and networking and is a new element that contributes to the organisation of a new economy related to the use of this quality.

In fact, the acquisition of new knowledge is preceded by the automatic finding of ER, which are, as shown in (Polyakov et al., 2021), sources for the formulation of hypotheses, which, in turn, represent the most important component of scientific cognition, a form of development of natural science, as noted in (Marx & Engels, 1958).

The universal scheme of cognition described above, but which already includes the use of DM tools, can be given a graphical form for greater clarity (Fig. 2).

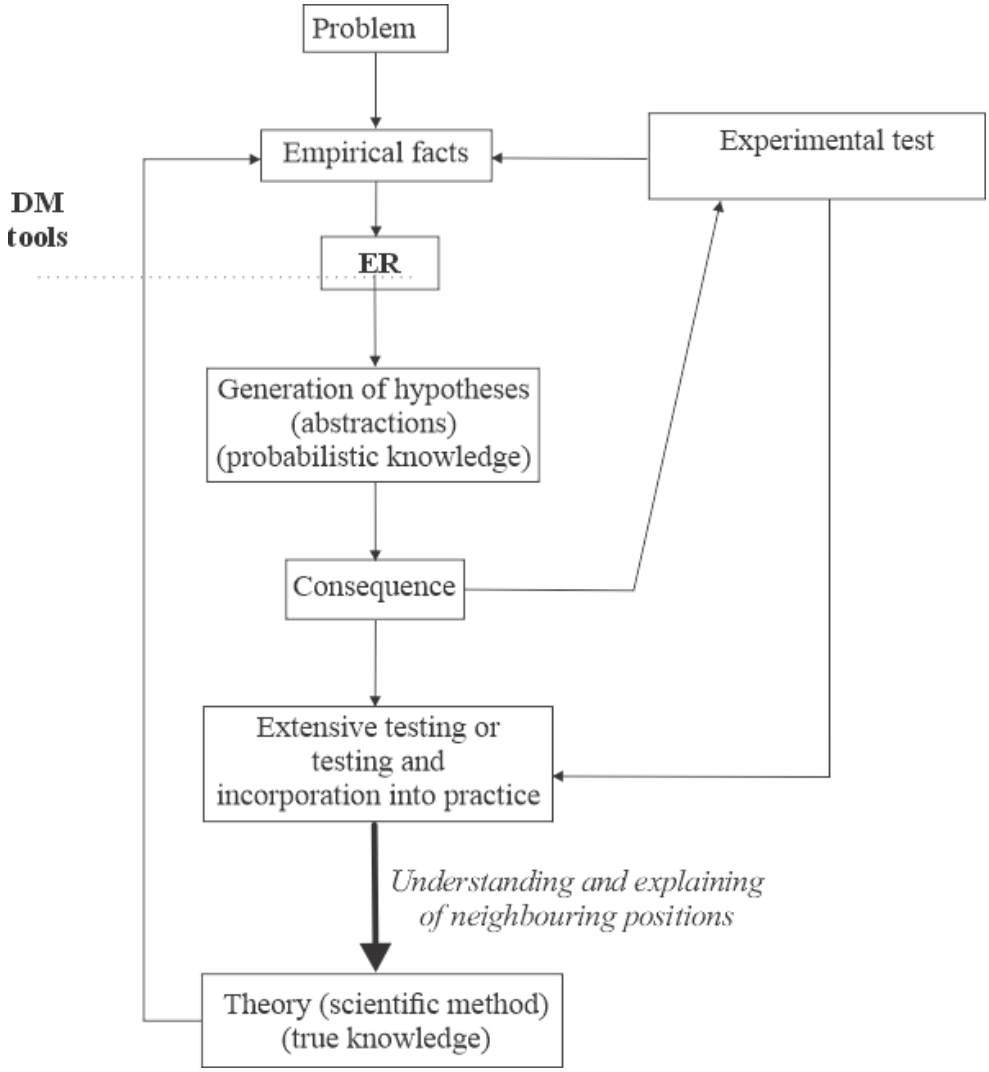


Figure 2. Universal scheme of cognition using DM tools

** Source: suggested by the authors.*

The scheme describes a general approach to cognition, but it should be refined in accordance with the purpose of this article - to determine the place of DM tools and the nature of their use. It is also desirable to indicate the possibilities and limitations of all DM methods in this process of cognition.

Data engineering.

The entire data design workflow can be described using the following basic steps.

1. Designing a TS, in the general case of OPT.
    1. Designing target factors
    2. Designing the feature space
        i. Enriching the feature space
    3. Data collection and consideration of constraints
    4. Working with data
        i. Cleansing
        ii. Normalisation, etc.
        iii. Representation in the form of an TS or OPT.
    5. Evaluate the quality of the collected data[6] and decide on its suitability or the need to modify, supplement the feature space and data. If yes, then item 6, if no, then item 1.
    6. Visualisation of the sample.
    7. Describing the data and providing a detailed picture to stakeholders and discussing it with them

The main objective of the data design stage is to obtain qualitative TS or OPT, which means structuring and categorising the data, i.e. obtaining information about the object of study.

The first step in such a design is to define the target or result variables and the factors (parameters, features, attributes, all these are synonyms) affecting them. In our case, a target variable is a variable that describes the result (goal) of a process. For example, it can take the following values: 0 - no defects, 1 - there is a defect of type 1, 2 - there is a defect of type 2, etc.

[6] For the case of TS, as will be shown below, it is possible to quantify the quality of the data.

The next step involves defining a set of necessary attributes or parameters, the so-called feature space, and collecting data. Data can be defined as anything you interact with.

The task of determining the set of attributes or parameters that have the maximum influence on the target variable is a very important task, as there are no rules for such determination and in order to select the attributes that have the closest relationship with the target variable, it is necessary to involve relevant scientists, engineers and other specialists in solving this task. At the initial stage of solving the problem, it is recommended to write down as many attributes as possible that characterise the objects under study and that can influence the outcome variable.

It should be realised that the key point is the interaction of certain groups of attributes with the target variable and compliance with certain requirements for such attributes.

In addition, we should pay attention to the fact that feature selection (data design) is actually related to the first part of the general model of cognition that we have given above (Gnedenko, 1983). Note that the use of DM methods fits into this model as closely as possible. It should be noted that A.D. Zakrevskiy, a well-known scientist in the field of DM, believed that practically any task can be described with the help of attributes, although we should clearly imagine that in this case we are dealing with external manifestations of the object of research. Having trait data for a large period, it is possible to predict trends and make forecasts with the help of DM methods. "Live" contemplation corresponds precisely to the selection of a feature space, its assessment of its suitability for use and, accordingly, the collection of the necessary experimental or expert material suitable in its qualities for further processing. In the above statement we can see the "connection", the intersection of the general model of cognition, its first part, with the initial stage of data design, i.e. the selection of the feature space of any DM method.

The theory of cognition does not describe this stage in detail, although its significant role and insufficient research in this direction are noted. Attempts to fill this gap were made in (Chorayan, 1987; Shapiro 1977; Smirnov, 1964), including

(Shevchenko at al., 2022), which proposed a procedure for the initial description of the landscape of all factors relevant to the task. In DM terms, this approach is precisely what is meant by feature space selection.

The proposed procedure, let us call it Procedure 1, translates step 1 of the general cognition scheme into a constructive scheme that allows us to obtain an informative picture of the problem:

- A qualitatively "contemplative" analysis of the task and the factors, attributes or concepts involved.
- Based on the task analysis and research objectives, the task landscape is described, i.e. all those factors, attributes that can influence or that "surround" the task are described.
- In this landscape, determinants **Di** are selected or highlighted, i.e. those factors that the researcher believes are decisive in influencing or determining the properties of a given concept and then proceeds to justify or prove that they are so.
- The selected determinants **Di** are then represented or written in the following formal form **<D1, D2, ... D3>,** which is actually a description of the feature space.

In essence, this procedure is a research tool to provide stakeholders with an informative picture, a representation of the problem for the purpose of building a model of interest.

We want to emphasise, on the one hand, the undoubted importance of this stage, and, on the other hand, the underestimated deep connection of the choice of feature space with the theory of cognition. DM in general can be referred to empirical methods of cognition, more precisely to inductive methods of cognition, which immediately points to quite certain limitations of all these methods. More details are discussed in (Polyakov at al., 2021).

We did not find other formal methods that reveal in detail the essence of the first step of the general model of cognition, albeit as applied to DM tasks.

Limitations of DM in data design.

DM is not an omnipotent technology, it is a tool that has its limitations, features and scope of applicability, ignorance of which reduces its effectiveness. These general limitations can be summarised as follows:

1. Focus on attributes, including targeted attributes, that minimise resource costs. The use of indirect attributes that require less effort to implement can be considered.
2. The attributes should be clearly defined, with a description of the methodology for measuring or assessing them.
3. A sufficient number of features need to be analysed to identify effective combinations that can lead to the desired result - this is the key objective of Data Mining techniques.
4. In some cases, it is preferable to use discrete attributes instead of continuous attributes. For this purpose, it is important to transform the data correctly, for example, by entering categories: "High value", "Medium value" and "Low value" instead of using exact numerical values.

It is important to take into account different points of view - otherwise you may get "rubbish" instead of information, and no DM methods will be able to identify ER.

In addition, you can try to enrich the data, i.e. add more additional data based on the specifics of the problem to be solved, e.g. use open data such as scientific research results, etc. data.

Note that "insignificant" at first glance attributes, or, most likely, some combinations of them, may turn out to be determinative. DM is precisely the means of discovering unexpected and unknown patterns in the studied data.

Once again, we draw attention to the fact that the quality of TS is directly related to the choice of a system of attributes for describing objects. It is also well known that the choice of attributes (attributes) is one of the most important issues in the construction of SR. The main difficulty in solving this problem lies in the fact that there are no formal rules allowing to specify in advance, a priori, such a set of features (attributes) with the help of which classification can be carried out with a given

accuracy. That's why this stage is so called - data design and it is the *responsibility of* task setter - project manager, etc.

Assessing the informativeness of discrete features of the training sample

In Data Mining tasks, when using TS, the number of measured attributes is usually very large, but not all of them are equally important for the construction of decision rule (DR). A large number of attributes complicates the construction of DR and leads to great inconvenience in its use, worsens the interpretability of the results, increases the volume of statistical material and the cost of obtaining data on objects.

In this connection, the task of selecting from the initial attributes a certain number of them and their combinations, the most important, informative for solving the task at hand arises. To find such attributes and their combinations, it is necessary to be able to estimate quantitatively the informativeness of attributes and their combinations.

The fundamental research of M. Kendall and A. Stewart[7] in the field of nonparametric statistical problems can serve as a basis for the construction of optimal estimates of informativeness (importance) of attributes. They proposed the estimation of the relationship between two categorised variables, which, as they showed, for some problems, in particular, for prediction of the categorised values of another variable based on the known one, is the best.

Interpreting their approach in terms of machine learning, we can consider that the predicted variable is some finite-valued function $f_{(R)}$ (i.e., class number), which describes a given partitioning of R sets of TS into classes. Prediction is performed on another categorised variable, which in this case is the feature $\mathbf{x_i}$

Then the estimate of the relationship between the two variables can be considered as an estimate of the informativeness of the feature $\mathbf{x_i}$with respect to the function $\mathbf{f_{(R)}}$. Such an estimate is calculated directly from the data of the TS and characterises the information that can be obtained about the function $\mathbf{f_R}$, knowing the values of the feature $\mathbf{x_i}$on the sets of TS.

[7] Kendall M., Stewart A. Statistical Inference and Relationships. - Moscow: Nauka, 1973. - 900

Based on this approach, we propose the following method for calculating the informativeness **V(x(i)) of** features $\mathbf{x_i}$, discussed below. The advantage of this method is the relative simplicity and "transparency" of calculations.

The proposed formula (1) is as follows:

$$V(x_i) = \frac{1}{k}\sum_{x=0}^{k_i-1}\max_{Y}\left(\frac{m_{XY}}{m_Y}\right), \quad (1)$$

where $V(x_i)$ is the informativeness of the feature x_i,

k is the number of classes,

$\mathbf{m_Y}$ is the number of objects belonging to class **Y**

$\mathbf{m_{XY}}$- number of objects belonging to class **Y** and taking the value

$x = [0, \dots, k_{i-1}]$ of feature $\mathbf{x_i}$,

$\mathbf{k_i}$is the number of gradations of feature $\mathbf{x_i}$.

The larger the value of $V(x_{(i)})$, the more informative the feature $x_{(i)}$ is considered to be . Let us note some important aspects of feature informativeness estimation (1):

1. It can be shown that $1/k \le V(x_{(i)}) \le 1$,
2. The informativeness of the features in (1) is defined with respect to $\mathbf{f_R}$, a function specifying the partitioning of the set *M of* recognition objects, which is absent in other known methods.
3. The numerical value of the informativeness of some feature coincides with the fraction of objects of the TS that can be correctly assigned to the corresponding classes on the basis of taking into account the values of the taken feature, i.e. (1) has an unambiguous relationship with the classification errors of TS. Note that properties 1, 2, 3 directly follow from (1).

Formula (1) is for the case of discrete (integer) features x_i.

Let's consider an example. We have the following table (Table 1). Let's calculate the informativeness for the attribute $\mathbf{x_{(1)}}$. Preliminarily we define: k=2, k_1=2, m_1=5, m_2=5, m_{01}=1, m_{02}=4, m_{11}=4, m_{12}=1.

Table 1.

x1	x2	x3	x4	Y
0	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	0	1	1	1
1	0	0	1	1
1	1	0	1	2
0	0	0	0	2
0	1	1	1	2
0	1	0	0	2
0	0	0	1	2

$$V(x_1) = \frac{1}{2}\left[\left(\max(\frac{m_{01}}{m_1}, \frac{m_{02}}{m_2}) + \max(\frac{m_{11}}{m_1}), \frac{m_{12}}{m_2}\right)\right] = \frac{1}{2}\left[\max(\frac{1}{5}), (\frac{4}{5}) + \max(\frac{4}{5}), (\frac{1}{5})\right] = \frac{1}{2}(\frac{8}{5}) = 0{,}8$$

Proceeding in the same way, for the remaining features we have:

$V(x_2) = 0.6$; $V(x_3) = 0.6$; $V(x_4) = 0.6$.

To automate calculations of the informativeness of features, including for more complex cases, when k-digit features are present in the TS and the number of classes is greater than 2, we have developed a web service "Sampling Quality", located at: https://www.sciencehunter.net/Services/Apps/dataSetQuality. After using the service, it is clear that the most informative characteristic is x3.

This calculation helps us to orientate ourselves in terms of the separability of characteristics, which is particularly valuable when there are a large number of them.

Data collection and evaluation.

After formulating the problem, "live" contemplation and selection of features, the next stage is data collection. This is also the most important step in solving the problem, as the quality of the collected material will largely determine the success of solving the whole DM problem, as in the famous aphorism - what grain we pour into the mill, such flour we will get. What we measure, how we measure, whether there are necessary tools for such measurement or they need to be created, etc. questions - they are the essence of this stage, very labour-intensive and time-consuming, financial and human

resources. Almost like Archimedes - give me a good TS and I will solve all your problems. This aphorism, strangely enough, encapsulates the whole essence of DM issues. But obtaining an TS does not yet guarantee the success of further steps. It is necessary to thoroughly check, "proofread" the TS in order to eliminate all sorts of errors and inconsistencies. This is very responsible and labour-intensive work and cannot be avoided. Besides, it is often carried out by developers or solvers themselves, i.e. by highly qualified specialists, and it is impossible to do it in any other way, because who, except them, can evaluate the suitability of the received material for further processing. Most of the work of a data scientist at any enterprise is the study and sorting of "raw" data for further analysis.

Unfortunately, this stage is often either incorrectly assessed or not paid sufficient attention to. To help this stage, a quantitative method of assessing the quality of the collected and tested TS can be proposed - it seems to summarise both the stage of contemplation and the stage of preparation of the TS. Such a method is proposed and discussed in detail below.

At present, practically the only way to estimate the quality of the TS in general case is the estimate obtained on the basis of the accuracy of the DR found during training, which is determined by the probability of correct recognition of TS sets by the built classifier - DR. However, obtaining such an estimate requires a training procedure and, consequently, possibly considerable time expenditures.

At the same time, when conducting scientific research in particular, it is desirable to know the quality of the TS in the process of its creation, at the stage of its preparation and data collection, without conducting the training itself and obtaining the DR. At the same time, it is desirable, on the basis of the found estimation, to give its interpretation, including the indication of the limits of discriminability of objects in a given sample, as well as to analyse the TS, including the indication of informative features and give constructive recommendations for improving the main quality of TS - its discriminability. I.e., this is the most important stage of data preparation, which also practically cannot be formalised at present and therefore is doubly important. In (Oleshko at al., 2004), an attempt was made to evaluate the quality of TS for building

a predictive neural network. However, the requirements for such an evaluation reduce the effect of its application and do not make it possible to evaluate the quality of TS for more general cases, which are much more common in practice.

We propose to fill this gap and approach the selection of feature space, in other words, landscape determinants, in a more conscious and formalised way, based on Procedure 1 proposed above and an algorithmic procedure for calculating the criterion that is used to assess the quality of the TS. Note that this criterion is calculated directly from the TS data and characterises the distinctiveness of the TS.

The calculation of the criterion is based on the application of the formula for calculating the importance (informativeness) of an arbitrary group of discrete features (Vasilenko & Shevchenko, 1979), which has the following form:

$$V(x_{i1},...,x_{ij}) = \frac{1}{k}\sum_{\Delta\in\Gamma}\max_{Y}\left(\frac{m_{\Delta Y}}{m_Y}\right), \qquad (2)$$

where **k** the number of classes, $\mathbf{m_Y}$is the number of objects belonging to class **Y,** $\Delta = t_{i1}, t_{i2}, \dots, t_{(ij)}$ $(0 \le t_{(ij)} \le k_{(ij)} - 1)$, $j=1,...,\gamma$ - an arbitrary set of values of features (attributes) $x_{i1},...,x_{ij}$ $(1 \le \gamma \le n)$,

$m_{\Delta Y}$ - the number of sample sets from the m-th class for which the relation $x_{ij} = t_{ij}$ $(j=1,...,\gamma)$ is satisfied, $t_{ij}$is the value of features (attributes) x_{ij}in the set , Δ Γ is the set of all sets of values of features (attributes) . $x_{i1},...,x_{ij}$

It can be shown that $1/k \le V(x_{i1},...,x_{ij}) \le 1$. The limiting value equal to 1 this estimation takes at full discriminability of classes, which was proved in (Vasilenko & Shevchenko, 1979). I.e., formula (2) at $\gamma = n$ can serve as an estimate of the quality of the TS consisting of discrete sets.

If $V(x_{i1},...,x_{in}) = 1$, it means that there is complete class distinction, at least when using the whole set of features. This approach to the estimation of the quality of TS allows implementing the above procedure, actually optimising the initial stages of data preparation when creating a sample, consciously approaching its creation and even at the first stages selecting the optimal set of features, which can significantly reduce time

and other costs when studying the subject area with the help of DM or Big Data methods: if after calculation by formula (2) at $\gamma =$ n the estimate takes the value 1 (or close to 1), then it is possible to carry out further classification processing of the data, which can be carried out with the help of DM or Big Data methods. Otherwise, it is necessary to use another coding function, either to abandon the existing feature system and attract additional features, or to abandon the available features and replace them with others (perhaps, to do all this in some cycle of this procedure). In this case, of course, with different estimations of the quality of the TS and available resources, taking into account time, we can have completely different strategies of action. Note that the procedure for calculating the informativeness of an arbitrary group of attributes, as well as the estimation of the quality of TS is implemented as a web application on the site https://www.sciencehunter.net/.

The situation is similar for Object-To-Property (OTP) tables, when at the initial research stage we simply collect data for further clustering/structuring, without verification. In this case, the same role of quality assessment is played by the visualisation of the OTP, its visual 2 or 3 dimensional representation. The huge role of visualisation in data analysis was described in their works by Tukey, Ishikawa, Taguchi and others (Ishikawa, 1988; Taguchi at al., 2005; Tukey, 1962; Tukey, 1981).

Indeed, one of the main stages of any research is structuring (clustering) of the obtained data. And visual representation of such structuring plays an important and sometimes even the main role, especially when it is necessary to analyse multidimensional data, which is a certain difficulty for many subject researchers who are not familiar with the basics of Data Science. This is due to the sensory origin of all our knowledge, and visualisation actually acts as a means of communication between the object of study and the researcher, with visualisation allowing the experimental data to be seen as a whole, visually, with the researcher perceiving the greatest amount of information, such as in Figure 4, which presents multivariate data characterising the

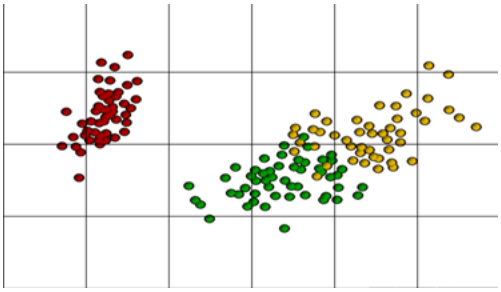


Figure 3. Data clustering (Fisher's Iris)

** Source: suggested by the authors.*

The famous Fisher's iris problem, presented by Fisher in (Fisher, 1936) and serving as a kind of test bed for many DM algorithms. In this problem there are 3 classes and 4 features. Visualisation allows us to see visually the interposition of objects of different classes, which are represented in Fig. 3 by different colours, and to give a preliminary assessment of the collected data in terms of cluster separability. Such visualisation of multidimensional data is also implemented on https://www.sciencehunter.net/ and Fig.3 demonstrates its use.

All of the above in this subsection relate to the tasks of the data scientist.

Conducting initial analyses.

After selecting, constructing the feature space and forming the sample, it is advisable to work on extracting useful data from the sample and making sense of it. This can start with describing the data, identifying trends, relationships, and checking for errors. The resulting picture should be presented to stakeholders and all its features discussed with them. It is recommended to start with simple data visualisation, including multivariate visualisation, which can say a lot about the data. *This is the task of the analyst.*

Choice of DM methods.

Then it is necessary to choose DM methods and decide whether to use ready-made algorithms and write a programme for them, taking into account the peculiarities of the assembled TS or to develop your own algorithm, i.e. to find a new way to solve the problem. In the first case the development will probably take several months, in the second case - much more time. But in case of success the result can be significant.

In our opinion, the choice of methods should be made not only from the point of view of the classification task, which is quite obvious for DM and the solution of the classification task itself, but also from the point of view of interpretation and explanation of the obtained results, their predictive power. Unfortunately, the neural networks, which are widespread and widely used at the moment, do not possess this ability, which sharply narrows their further use in scientific research in the sense of explanation and prediction. We would like to pay special attention to this, because in most cases, failure to understand this circumstance may further lead to discrediting DM methods in general, because after quite certain successes on the way of recognition and classification, there will surely follow steps that are quite natural for any mental activity - what next, how to find a cause-effect relationship, how to interpret the results, etc. Therefore, the questions raised in this article reflect not only and not so much the current situation in this field of knowledge, but also make an attempt to point out important questions that may appear in the future. We need not only knowledge, but also understanding, only then we can talk about possible process control and a new stage of cognition.

But it should be remembered that DM also has some disadvantages:

- requires a large investment of time and labour, as data mining is a long and complex process that needs productive and skilled people;
- it may produce false or irrelevant patterns if the data are not properly prepared or cleaned, or if inadequate algorithms or metrics are used;
- does not take into account data not identified in the process of their collection, as well as the essential aspects of the object;
- there may be distortions of the data, due to truncation or alteration by consciousness;
- data that do not follow directly from the processed data are not identified;
- it can create false or irrelevant patterns if the data has not been properly prepared or cleaned, or if inappropriate algorithms or metrics have been used.
- does not take into account data that were not identified in the process of their collection, as well as the essential aspects of the object.

- there may be distortions of the data due to truncation or alteration by consciousness.
- data that do not follow directly from the processed data are not identified.

The DM methods you can use depend on your goals and the type of data you have. Choosing the best DM method for your task depends on several factors, such as:

- The type and amount of data you have. Some methods work better with numeric data, others with categorical or textual data. Some methods require a large amount of data to train, others require less.
- The purpose and complexity of your problem. Some methods are suitable for simple classification or regression tasks, others for more complex clustering, sequence analysis or association rule analysis tasks. Some methods allow you to interpret results and understand causal relationships, others do not.
- The speed and accuracy requirements of the model. Some methods are faster and easier to learn and apply, others are slower and more complex. Some methods produce more accurate and robust predictions, others less accurate and robust.

To choose the best DM method, you should compare different options according to these criteria and choose the one that best suits your situation. You can also use special tools or guides to help you choose a DM method:

- A map of Microsoft's machine learning algorithms that shows different types of tasks and the algorithms suitable for them.
- A machine learning algorithm selection scheme from Scikit-Learn, which offers different algorithm options depending on the type of data and the goal of the problem.
- The book "Data Mining: Practical Machine Learning Tools and Techniques" by I.H. Witten, E. Frank and M.A. Hall, which contains a detailed description of different Data Mining methods and recommendations on their selection and application.

The choice of DM method is the responsibility of the relevant specialist - data scientist, it is his *task and area of responsibility*

In many cases, the solution of specific practical problems is limited, from the point of view of cognition, to the level of hypothesis, or, more precisely, even to the level of ER (or preliminary hypothesis), on the basis of which they formulate further, at best, instructions or rules of decision-making, and remain at the first, lowest possible, empirical level of cognition. This is typical for business tasks, because in the short term it suits business as a sphere of practical activity, but in the long term the main thing is lost - finding actual new deep knowledge that can be embodied in innovation, or the development of a new method, which will give a **competitive advantage of the highest order.** For scientific research involving DM the latter is the main focus.

Customer Communications.

It is necessary to note the constant communication with the customer, i.e. with those people or specialists in whose interests the task is solved. This communication should be throughout all stages of problem solving - only in this way it is possible to get a quality solution. In the case of successful interaction, the solution really turns out to be successful, otherwise all the costs are in vain. It can be added that at the stage of interaction, besides purely professional ones, there are problems of coordinated opinion, consideration of personal characteristics, and many other things from what we call arsenal of humanitarian sciences. And sometimes this approach is decisive, at least, it should not be discounted in any case, although it is not a purely professional duty of the DM problem solver.

Conclusions and prospects for further research.

Based on the above, it can be argued that DM methods can provide only the level of empirical knowledge in the specific subject area under study, the level of techniques and instructions. Therefore, it becomes clear why there are no "breakthrough" discoveries and innovations made with the help of DM - because they can be obtained so far only in a specific subject area, with close co-operation and interaction, full scientific communication, including discussion of hypotheses, insights and other elements of compulsory creative activity with representatives of the very subject area, which is the biggest obstacle to this kind of achievements.

A number of conclusions follow from this.

1. DM methods, like Big Data, are a new, human-machine methodology for empirical cognition.

2. These methods have their own limit in the form of ER presented in different ways.

3. ER can serve as "blanks" for formulating, testing and selecting hypotheses for further, deeper knowledge of the subject area.

4. In Data Mining there is an essential contradiction of data processing: any problem solved in it, which turns out to be the subject of technology, remains external to the available arsenal of means and even to the object (situation) under study.

5. Data Mining models in most cases have a clear classification character. Classification is the initial stage of any cognition, including at the genetic level, so these models turned out to be quite in demand, although they do not carry conceptual, semantic content.

6. Data Mining deals with the external manifestations of the object of research, does not affect its essence, i.e. it is limited and, more often than not, not systematic.

7. As a result, it was possible to conclude that Data Mining cannot be considered a sufficiently well-founded technology, much less a science-based technology (which, however, is often recognised in the literature), even though it uses methods of statistics, mathematics, etc., but this already raises the question of the possibility of using Data Mining in the form and scope in which it is applied

8. That is why the Data Mining literature always points out that the application of Data Mining is not a guarantee of obtaining reliable knowledge and making correct decisions based on this knowledge.

9. There is no regulation of the application of methods, which is decided by a specialist who cannot, as has been openly recognised, adequately navigate their abundance, as well as the specificities of the various subjects ,

The use of DM tools requires close collaboration with subject matter experts, which in turn raises a number of issues related to: initiating such collaboration; the willingness of subject matter experts; framing the problem in an appropriate perspective; building a team to solve the problem; etc.

The "departure" of DM and Big Data specialists to the field of standardised software development (cloud services, web services, desktop applications) does not solve the problem of deepening cognition; the limit is still empirical cognition - obtaining ER, i.e. actually a preliminary hypothesis for this particular subject area. In this case, the burden of solving the specific task of deepening cognition, clarification of such a hypothesis is completely shifted to specialists in the subject area. Besides, they have problems to study the corresponding software, to understand its basic capabilities and, most importantly, whether the selected programmes are suitable for solving practical tasks in the given field. At the same time, such important issues for problem solving as: selection and construction of feature description, understanding of the subject area, data verification (quality, usefulness), selection of a suitable processing algorithm, evaluation of method reliability, interpretation of results, transition from data to decisions and actions remain "overboard". Most often such questions are solved intuitively, by experience or analogy, by the method of "trial and error".

Full collaboration between subject matter experts and Data Scientist is much more time consuming in terms of organisational and communication costs, but in our view, this approach is nevertheless capable of providing profound breakthroughs in the subject matter.

Therefore, the main conclusion is that in order to successfully use DM and obtain the maximum possible results in cognition with their help, namely, reasonable hypotheses, the main attention should be focused on preliminary analyses, formulation of problems and selection of features to describe objects. This becomes the axioms of using DM in scientific research. The task has to be studied in detail, features have to be chosen reasonably by making a feature description of objects, data have to be collected and verified, and only then can one hope to obtain an acceptable result (ER - as a preliminary hypothesis) using various special applications.

The general scheme of ER search is presented in Fig.4.

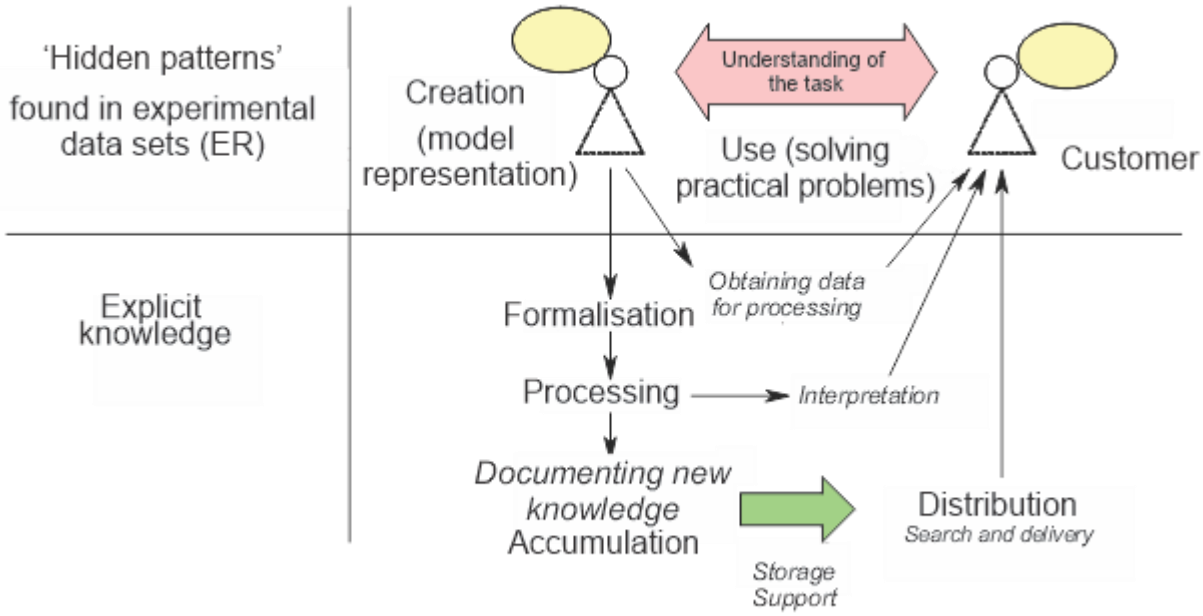


Figure 4. Scheme for searching for hidden empirical regularities (ER)

* *Source: suggested by the authors.*

At the same time, it should be understood that the search for ER is empirical cognition and fits into the general scheme of cognition given in (Kuchev), in which the role and place of DM in the scheme of cognition is clarified (Fig. 2).

But whereas science is interested in the fact of knowledge itself as well as in the methods of knowledge, business is willing to dwell on the very fact that leads it to commercial success.

In many cases, the solution of specific practical problems is limited, from the point of view of cognition, to the level of hypothesis, more precisely, even to the level of ER (as a preliminary hypothesis), on the basis of which they formulate further, at best, instructions or decision-making rules, and remain at the first, lowest possible, empirical level of cognition, which is characteristic of business. When conducting scientific research, the main thing is to find really new knowledge, which can be further embodied in innovations, or the development of a new method, which can give a competitive advantage of the highest order.

In fact, if understanding appears, it may mean that something has been found that underlies the knowledge gained, and it may be a new method of solving the

problems mentioned. Knowledge is information, and understanding is the ability to draw conclusions. This is the limit of applicability of all DM tools to date. They provide knowledge in the form of ER and working hypotheses, but they do not provide understanding, without which the transition to the next level of cognition - the scientific method - is impossible.

Lastly, the emergence of systems such as GPT in no way detracts from or replaces the use of DM, as GPT simply cannot find the necessary knowledge due to the whole list of issues discussed above - feature selection for a particular task, collection of TS, data validation and cleaning, ER search and model building - this whole set of procedures and thought operations is not yet fully algorithmic and therefore cannot be implemented by AI systems.

DM is a very broad and diverse field and examples of its application within an organisation may depend on the specific activities, aims and objectives, available data and resources. Appendix 1 provides examples of the use of DM techniques in research to generate new knowledge. Appendix 2 provides examples of the use of DM techniques to solve business problems.

References

1. Adler, Y. P., & Speer, W. L. (2019). *Practical guide to statistical process control* (234 p.). Moscow: Alpina Publishers.
2. Algarni A. (2016). *Data Mining in Education*. International Journal of Advanced Computer Science and Applications.7 (6). https://doi.org/10.14569/ijacsa.2016.070659.
3. Aryal, S. C. (2023). *The impact of a digital regime on academic knowledge production*. University West: School of Business, Economics, and IT. https://www.diva-portal.org/smash/get/diva2:1746631/FULLTEXT01.pdf
4. Bu Daher, J., Brun, A., & Boyer, A. (2018). *A review on heterogeneous, multi-source and multi-dimensional data mining*. HAL. https://hal.science/hal-01811232/file/heterogeneous-multisource-mining.pdf

5. Chorayan, O. G. (1987). *The concept of probability and fuzziness in brain function* (156 pp.). Rostov-on-Don: Izd-vo RSU.
6. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.
7. Frank, F. (2010). *Philosophy of Science: The Connection between Science and Philosophy* (G. A. Kursanov, ed.; 3rd ed., 512 p.). Moscow: LKI Publishing House.
8. Geology and mathematics. (1970). *Geology and mathematics* (224 p.). Novosibirsk: Nauka.
9. Glinsky, B. A. (1965). *Modeling as a method of scientific research*. Moscow: Nauka.
10. Gnedenko, B. V. (1983). *Mathematics and scientific cognition* (64 p.). Moscow.
11. Gupta, A., et al. (2020). *Comprehensive review of text-mining applications in finance*. *Journal of Financial Innovation, 6*(1). https://jfin-swufe.springeropen.com/articles/10.1186/s40854-020-00205-1
12. Hullman, J., & Gelman, A. (2021). Challenges in incorporating exploratory data analysis into statistical workflow. *Harvard Data Science Review, 3*(3). https://doi.org/10.1162/99608f92.9d108ee6
13. Investopedia. *Data mining: Works, benefits, techniques, and examples*. https://www.investopedia.com/terms/d/datamining.asp
14. Ishikawa, K. (1988). *Japanese methods of quality management*. Moscow: Ekonomika.
15. Jurs, P., & Eisenauer, T. (1977). *Pattern recognition in chemistry* (230 p.). Moscow: Mir.
16. Jurs, P., & Eisenauer, T. (1977). *Pattern recognition in chemistry* (230 p.). Moscow: Mir.
17. Khajehei, M., & Etemady, F. (2010). *Data Mining and Medical Research Studies*. Second International Conference on Computational Intelligence, Modelling and Simulation. Bali, Indonesia, 119-122.

18. Kuchev, A. *Data mining: process, types of techniques and tools*. Hubr. https://habr.com/ru/articles/784060/
19. Leitgöb, H., & Prandner, D., Wolbring, T. (2023). *Big data and machine learning in sociology*. Frontiers in Sociology. Vol. 8. https://doi.org/10.3389/fsoc.2023.1173155.
20. Malinovsky, L. G. (1986). Classification processes - the basis for the construction of sciences about reality. In *Algorithms for processing experimental data* (pp. 155-182). Moscow: Nauka.
21. Marx, K., & Engels, F. (1958). *Sochineniye* (4). Moscow: Publishing House of Political Literature.
22. Moiseev, N. N. (1982). *Man. Environment. Society. Problems of formalized description* (239 p.). Moscow: Nauka.
23. Németh, R., & Koltai J. (2021). *The Potential of Automated Text Analytics in Social Knowledge Building*. Pathways Between Social Science and Computational Social Science. Cham, 49-70. https://doi.org/10.1007/978-3-030-54936-7_3.
24. Neuimin, Y. G. (1984). *Models in science and technology: History, theory, practice* (p. 49). Leningrad: Nauka, Leningrad Branch.
25. Oleshko, D. N., Krisilov, V. A., & Blazhko, A. A. (2004). Construction of qualitative training sample for predictive neural network models. *Artificial Intelligence*, 3, 567-573.
26. Paklin, N. B., & Oreshkov, V. I. *Business analytics: from data to knowledge*. Hubr. https://habr.com/ru/articles/66561/.
27. Polyakov, I., Khanin, G., Shevchenko, V., & Bilozubenko, V. (2021). Data mining as a cognitive tool: Capabilities and limits. *Knowledge and Performance Management, 5*(1), 1-13.
28. Rastegari, H., & Md. Sap, M. N. *Data mining and e-commerce: Methods, applications, and challenges*. Core. https://core.ac.uk/download/pdf/11784915.pdf

29. Reshetnikova, M. (2021). *Banks, retail, medicine: who uses Data Mining and for what*. RBC Trends. https://trends.rbc.ru/trends/industry/61b359739a7947c7376ef7ce
30. Rosenbluth, A., & Wiener, N. (1984). The role of models in science. IN YA. G. Neuimin (Ed.), *Models in science and technology: History, theory, practice* (p. 171). Leningrad: Nauka, Leningrad Branch.
31. Shapiro, D. I. (1977). Toward human-machine methods for solving one class of problems. *Issues in cybernetics. Theory of cybernetics. Theory and practice of situational control*, (18), 82-88.
32. Shevchenko, G. Y., Bilozubenko, V. S., & Marchenko, O. A. (2022). The formation of the corporate scientific environment. *Nauka ta naukoznavstvo*, (2[116]), 12-24.
33. Shtoff V.A. Problems of methodology of scientific cognition. Moscow, 1978. 269 c.
34. Smirnov, V. A. (1964). Levels of knowledge and stages of the process of cognition. In *Problems of the logic of scientific cognition* (pp. [page]). Moscow: USSR Academy of Sciences, Institute of Philosophy.
35. Strachan S. & Stephen B., McArthur S. (2007). *Practical Applications of Data Mining in Plant Monitoring and Diagnostics*. IEEE Power Engineering Society General Meeting.
36. Taguchi, G., Chowdhury, S., & Wu, Y. (2005). *Taguchi's quality engineering handbook*. John Wiley & Sons, Inc.
37. Toledo M. R. (2021). *Data Mining applied to interventional cardiology procedures*. Journal of Physics: Conference Series. 1723(1). 12-34. https://doi.org/10.1088/1742-6596/1723/1/012034.
38. Tukey, J. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*(1), 1-67.
39. Tukey, J. (1981). *Analyzing the results of observations. Exploration Analysis* (Per. with Engl.; ed. by V. F. Pisarenko). Moscow: Mir.

40. Vasilenko, Y. A., & Shevchenko, G. Y. (1979). Analytical method of finding tests. *Automatics*, (4), 3-8.
41. Waters, D. J. (2022). *The emerging digital infrastructure for research in the humanities*. International Journal on Digital Libraries. https://doi.org/10.1007/s00799-022-00332-3.
42. Waters, D. J. (2023). The emerging digital infrastructure for research in the humanities. *International Journal on Digital Libraries, 24*(1), 87-102. https://doi.org/10.1007/s00799-022-00332-3
43. Yethiraj N. G. (2012). *Applying Data Mining Techniques in the Field of Agriculture and Allied Sciences*. International Journal of Business Intelligents. 1 (2), 40-42. https://doi.org/10.20894/ ijbi.105.001.002.004.
44. Zakrevsky, A. D. (1988). *Logic of recognition* (118 pp.). Minsk: Nauka i tekhnika.

Appendix 1.

How DM works in scientific research

Example 1. Let us give as a real example of using DM tools a web service for cytogram processing (system of automatic processing of cell nuclei images), which can be used in the study of various diseases associated with changes in densitometric characteristics of cells of biological fluids, for example, blood.

With the help of the mentioned web service it is possible to segment cell images and extract the necessary attributes from them to build an object-property table (Fig. 6). A brief presentation is available at the link: *https://goo.gl/E88B6T.* Direct work with the web-service for automatic processing and analysis of cytograms (digital images of human cells) is available at https://www.data4logic.net/ru/Services/CellsAttributes. Processing of TS data and ER extraction is possible using the web service[8] https://www.sciencehunter.net/Services#/tools/Классификация.

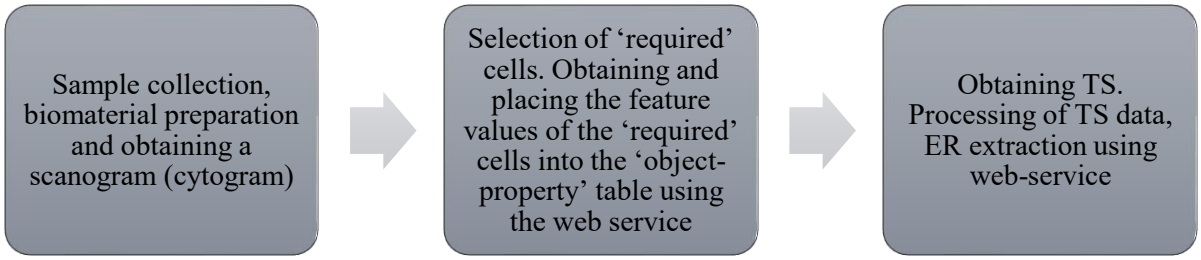


Figure 5. Scheme of service operation

* *Source: suggested by the authors.*

The work of the service consists of several stages.

The operation of the service for automatic image processing of cell nuclei consists of several steps.

Step 1: Uploading the original cell image.

In the first step, the user uploads a photograph of cells taken with a microscope to the website. These can be blood, lymph or buccal epithelial cells. The methods of

[8] Related Developments:

1. Automatic Hematology Imaging Analyzer (Effective solution for automated differentiation of peripheral blood cells in large-sized laboratories http://visionhemaultimate.com/). Austrian company Vision Hema® Ultimate
2. 2D IMAGE ANALYSIS SOFTWARE . Used by thousands of researchers worldwide, Image-Pro Plus image analysis software makes it easy to acquire images, count, measure and classify objects, and automate your work. This software solution offers microscope control, image capture, measurement, count/size, and macro development tools.http://www.mediacy.com/imageproplus

processing and staining of the samples can be very diverse. This is the raw data.

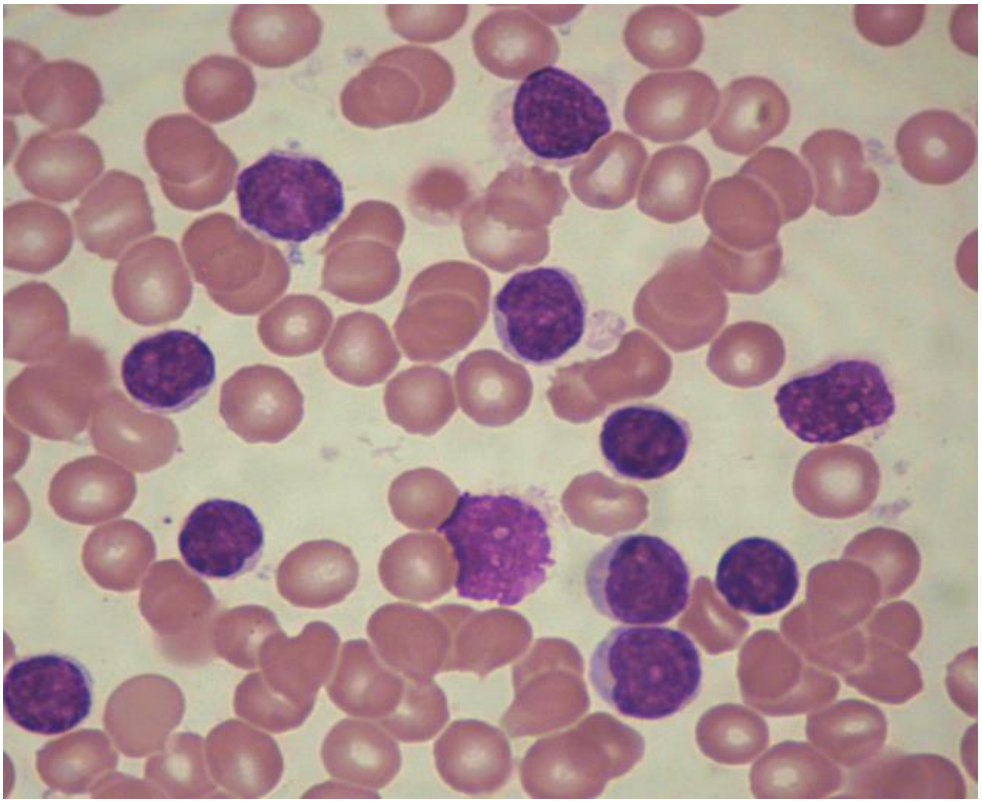


Figure 6. The original cell image

* *Source: suggested by the authors.*

Step 2: Isolation of cell nuclei.

In the second step, the user selects two colours - the closest to the background colour and the colour of the selected cells. By varying these parameters it is possible to select different types of cells. Next, the image is automatically segmented into objects and background, and the adherent cells are separated. This is the stage of selection of suitable data ("right" cells).

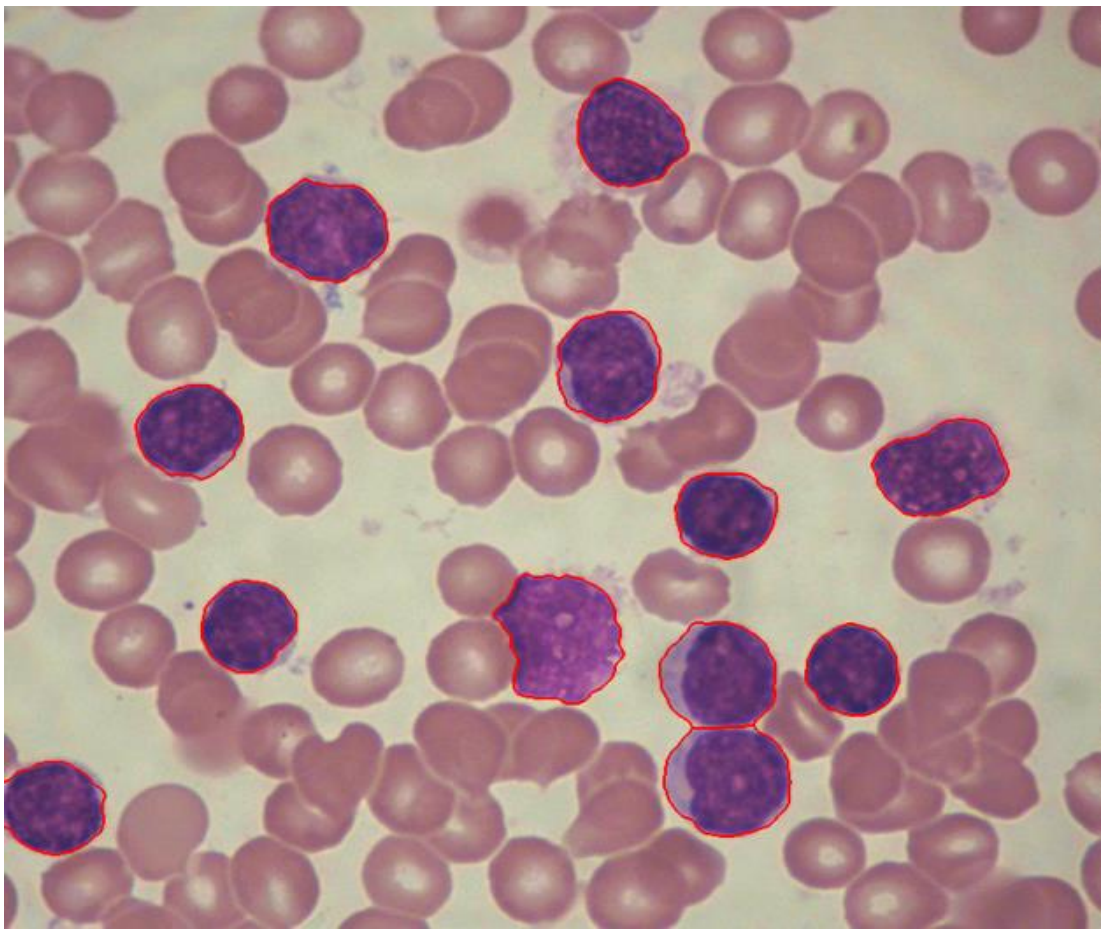


Figure 7. Isolation of cell nuclei image

** Source: suggested by the authors.*

Step 3: Feature extraction.

The third step is the actual extraction of features from the extracted cell nuclei and placing them into the OPT table. The features are divided into several categories - geometric, colour, statistical and fractal.

The cell features together with the original cell images can be saved in .xls format, which allows the researcher to conveniently mark up the sample to obtain a training sample for further classification and further processing. Also, such a file has an advantage over the original photo in that it groups cells in one column, greatly simplifying the work of the cytologist.

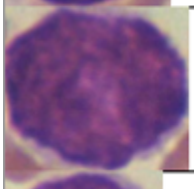
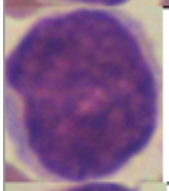
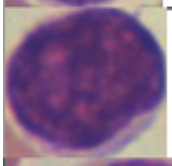
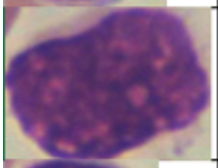
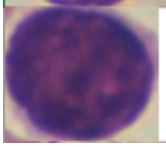
Image	Area	Perimeter	Radius	Diameter	Smoothness	Compactness	Eccentricity	Ellipticity	Convexity	Solidity
	4603	260,593	38,3977	76,5554	2,15644	1,17402	0,44024	0,9597	0,94857	0,97801
	4994	272,936	39,8788	79,7406	2,35838	1,18703	0,49869	0,97302	0,94616	0,96877
	4762,5	271,865	38,9661	77,8704	2,63265	1,23499	0,53078	0,97146	0,93014	0,9672
	4033,5	243,865	35,842	71,6632	1,31827	1,17329	0,37298	0,97947	0,9431	0,97463
	4987	282,593	40,1174	79,6847	5,31789	1,2743	0,71554	0,94604	0,939	0,96237
	3597,5	228,409	33,8838	67,6792	1,64094	1,15403	0,40863	0,97128	0,95395	0,97944

Figure 8. Feature extraction image

* *Source: suggested by the authors.*

Step 4: Visualisation of the resulting OPT.

The fourth step involves visualising the resulting object-property table (OPT).

For this purpose, methods such as principal components or multidimensional scaling are used.

The visualisation of the OPT (2D or 3D representation) gives a general idea of the cluster structure of the sample (or lack thereof) and helps to facilitate further data processing.

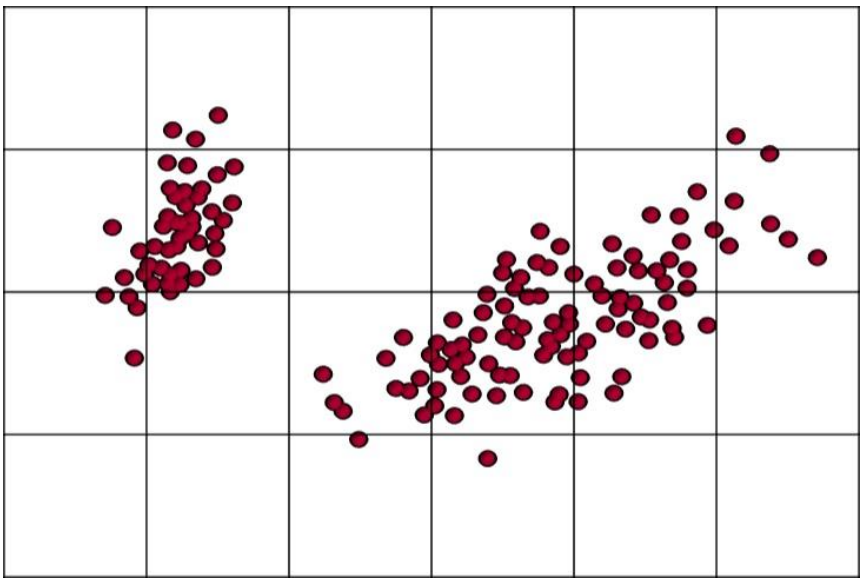


Figure 9. Visualisation of the resulting 2D

* *Source: suggested by the authors.*

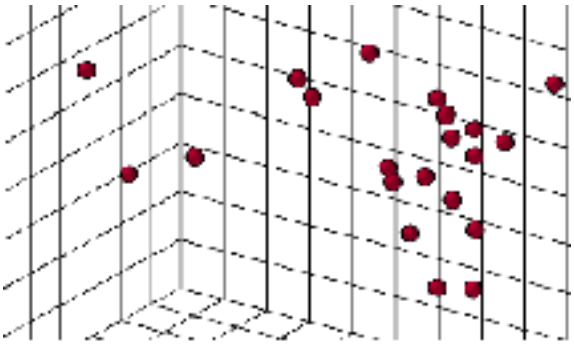


Figure 10. Visualisation of the resulting 3D

* *Source: suggested by the authors.*

Step 5: Clustering of the resulting OPT.

The fifth step is clustering. There are many different clustering algorithms, each of which is good for certain shapes and sizes of clusters. The selection of a clustering algorithm is facilitated by conducting a preliminary visualisation of the OPT. The clustering results can also be visualised to help understand whether the chosen clustering algorithm is suitable for the sample.

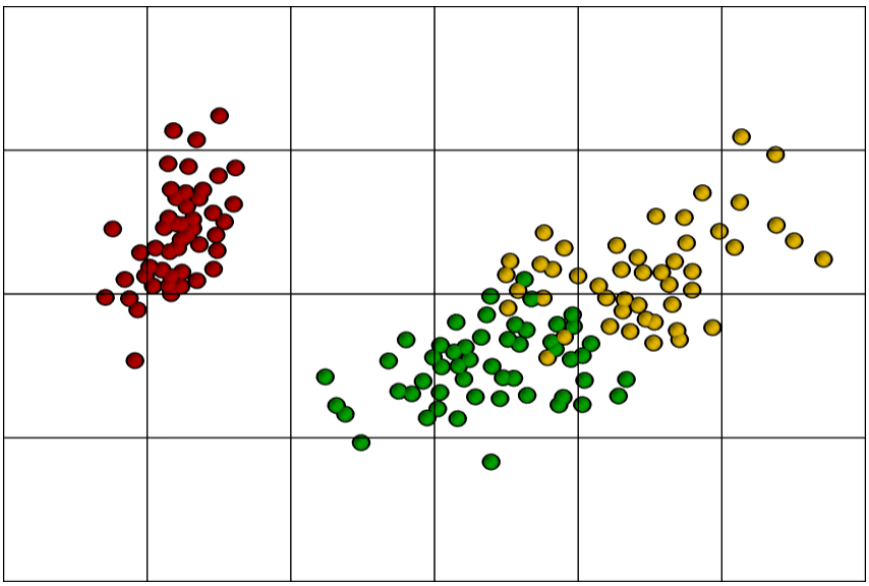


Figure 11. Clustering of the resulting 2D

** Source: suggested by the authors.*

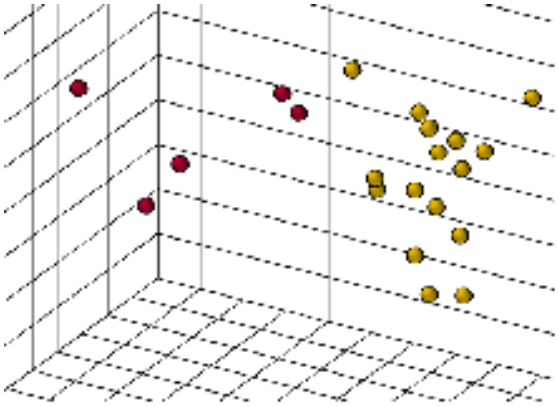


Figure 11. Clustering of the resulting 3D

** Source: suggested by the authors.*

Step 6: Extraction of ER from the sample.

The final step is the extraction of ER from the training sample. Preliminary OPT is divided into the selected number of classes and then we can use Data Mining methods to obtain preliminary hypotheses in the form of ER. In our case, the most suitable tool is decision trees (Fig.12), because as a result of their application we obtain ER (tree branches - decisive rules that allow us to obtain a diagnosis or prediction of cell type based on the extracted feature values (Table 1)) and in fact it is possible to obtain cause-and-effect relationships, as well as which of the features are most informative in decision making. In fact, ER play the role of working hypotheses and can serve as a

basis for doctors-researchers to study them in more depth, including through full scientific communication on the basis of this service and with other researchers, in order to identify more reliable knowledge, thus realising the increment of new knowledge as a result.

Table 1. Decisive rules

Decisive rules
If x_2=”4” that class=”1”
If x_2=”2” that class=”2”
If x_2=”3” that class=”2”
If x_2=”1” and x_1=”2” that class=”2”
If x_2=”1” and x_1=”0” that class=”3”
If x_2=”1” and x_1=”0” that class=”3”
If x_2=”0” that class=”3”

These data can serve as initial or tentative hypotheses from which[9] the researcher can derive really important patterns, establish their reliability and truth by testing them with considerable empirical material, and thereby reveal new scientific knowledge. In fact, this is the purpose of using DM in scientific research. All this can be done on the service: *https://www.data4logic.net/ru/Services/CellsAttributes*

[9] This, of course, does not preclude the generation, derivation and testing of hypotheses obtained by other means

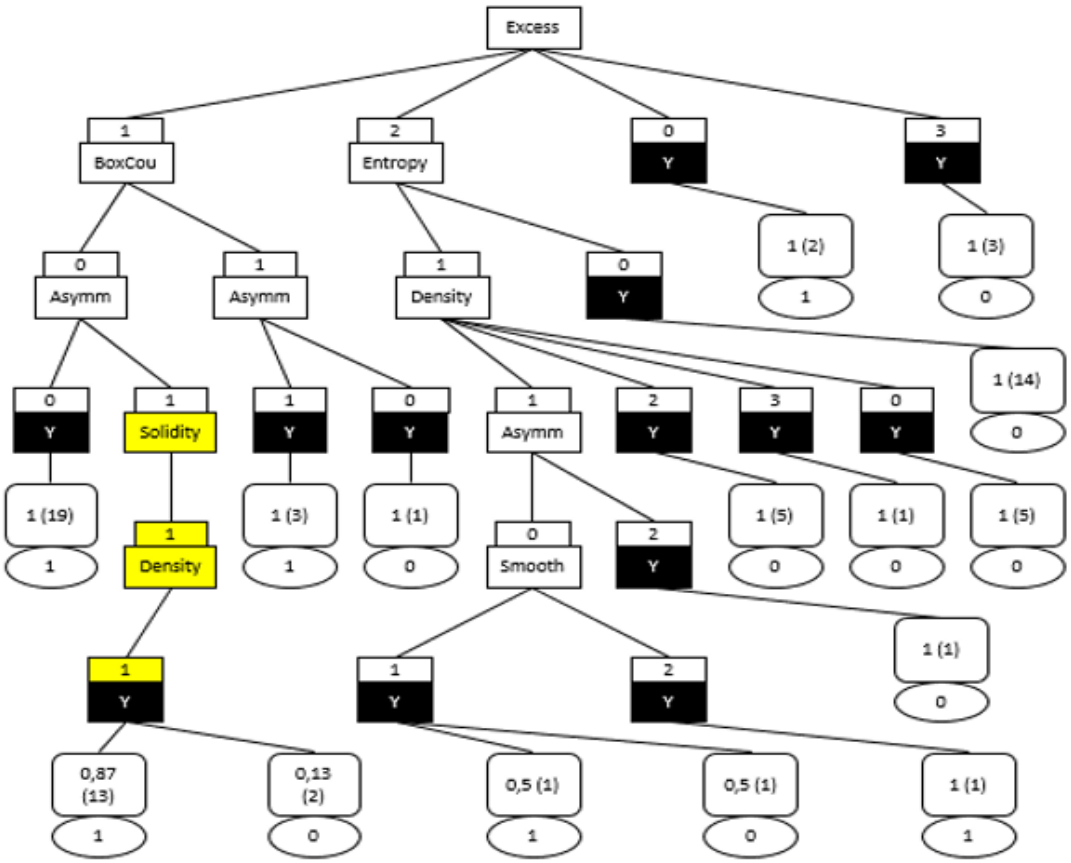


Figure 12. Decision tree

** Source: suggested by the authors.*

Methods available at: https://www.sciencehunter.net/Services#/

The developed software is designed to improve the efficiency of medical personnel, including research personnel, in the diagnosis, treatment and monitoring of various diseases that have manifestations at the cellular level, in densitomorphometry.

Other opportunities in the direction of medical image processing:

1. In cytology - analysis of cell shapes and other objects, i.e. partial automation of the cytologist's work.

2. Retinal image processing (e.g. diagnosis of Alzheimer's disease by retinal image analysis, iridodiagnosis).

3. To detect lympholeukaemia (blood cells change their shape).

4. Search for "similar" images in medicine.

5. The general approach is feature extraction on the basis of fractals and other characteristics of blood cells and construction of TS with its further processing in order to identify structural and logical regularities in it.

In addition, the results obtained may be of great interest for semiconductor silicon metallography, for defectoscopy and in other fields related to the analysis of structures in images.

Example 2. The following is an example of the use of DM tools in a sociological survey. It uses data on households in a specific country, in this case Spain, including characteristics such as income, family size, education level, age of the head of household, etc. To find data on households in Spain, data from the official statistical office, Instituto Nacional de Estadística (INE), which publishes data on households, were used. It is also possible to use databases of Research Projects such as **Luxembourg Income Study (LIS)** - LIS Database (household income data), **EU-SILC (European Union Statistics on Income and Living Conditions)** - EU-SILC on Eurostat, **Demographic and Health Surveys (DHS)** - DHS Programme**, EU-SILC (European Union Statistics on Income and Living Conditions)** - contains microdata on income, living conditions and poverty of households in Europe.

The following socio-economic characteristics of households were selected: **Family size_family, Income ($), Age_head, Education_level, Employment, Housing_type, Home_ownership, Number_of_children, Car.**

These characteristics are widely used in economics, sociology and marketing research to analyse living standards and social policies.

Below are 10 lines to demonstrate the dataset:

Table 2. Households dataset

Family size	Income ($)	Age_of_Chapter	Level_of_education	Employment	Type_of_housing	Home_ownership	Number_of_children	Car
1	55000	45	Higher	Working	Flat	Own	1	1	Yes
3	42000	39	Average	Unemployed	House	Rentals	2	1	Yes
2	35000	50	Higher	Pensioner	House	Own	0	1	Yes
5	67000	37	Average	Working	Flat	Rentals	3	2	Yes
1	28000	60	Higher	Pensioner	Flat	Own	0	0	Yes
3	48000	42	Average	Working	House	Own	1	1	Yes
4	75000	35	Higher	Self-employed	House	Own	2	2	Yes
2	39000	47	Average	Working	Flat	Rentals	0	1	Yes
5	53000	41	Higher	Working	House	Own	3	2	Yes
1	32000	55	Average	Unemployed	Flat	Rentals	0	0	No

Example: A study of household types.

Research objective: To identify the different types of households in Spain based on their socio-economic characteristics.

Study steps (500 datasets containing the above socio-economic characteristics of Spanish households were selected for the survey):

1. Data collection: Collect data in the OPT file on households, including the above characteristics such as family size, income, education level, age of head of household, etc.
2. We carry out the following steps of cluster analysis, for which we use the service https://www.sciencehunter.net/Services/Clustering#/analysis:

- Visualisation: We apply visualisation to approximate the number of clusters based on the service: https://www.sciencehunter.net/Services/Clustering#/visualization.
- Clustering: With the number of clusters selected in the previous step, we now apply the K-means algorithm to cluster households based on the collected characteristics (https://www.sciencehunter.net/Services/Clustering#/analysis).
- Result: We obtain several clusters, each representing a group of households with similar characteristics.
- We save the result to Excel using the same service, i.e. we actually get the TS.

3. It is possible to preliminarily evaluate the quality of the TS using the TS quality service and decide on the feasibility of constructing a decisive rule.

- For further processing of the obtained TS, we turn to the corresponding service at https://www.sciencehunter.net, for example, to the service for building decision trees or discriminant functions or building a decision rule based on Bayes formula.

4. To obtain an ER using Decision Trees, we will use the service https://www.sciencehunter.net/Services/Classification:

- Purpose: To use decision trees to explain which characteristics are most important in assigning a household to a particular cluster.
- Process: Train a decision tree on the data where the clusters obtained in the previous step are used as class labels.

- Result: We obtain a decision tree that shows which characteristics and in what order influence whether a household belongs to a particular cluster:
    a. We find the informativeness of the characteristics and the most informative groups of characteristics.
    b. The most informative groups of characteristics are "revealed" in the form of separate branches of the decision tree - ER, which shows which characteristics and in what order (combination) influence the belonging of the household to this or that class.

An example of the result obtained:

- Cluster 1: Households with high income, small family size and high education level.
- Cluster 2: Households with medium income, large family size and medium level of education.
- Cluster 3: Households with low income, small family size and low education level.

A decision tree can show that when classifying a household, for example, income level is the most important factor, followed by family size and then education level. These would be the ER found that answer the question of why a particular household would be assigned to a particular type of cluster. An example of such an ER is given below:

If 50000>Income>30000 and Family_Size = 3 to 6 and Education Level= secondary, then Class = "2"

Advantages of this approach:

- Interpretability: Decision trees help to easily interpret the results of cluster analysis.
- Simplicity: This approach allows social scientists to quickly and efficiently analyse large amounts of data and identify key patterns.

Example 3: Using the K-means clustering algorithm in political science

Task:

Analysing citizens' political preferences based on demographic and social factors.

Data (conditional):

Suppose we have survey data from 1,000 respondents with these characteristics:

- **Age** (18-80 years)
- **Income** (low, medium, high)
- **Education** (secondary, higher, degree)
- **Region of residence** (city/village)
- **Political orientation** (from 1 - "left" to 10 - "right")

An application of the K-means algorithm:

1. **We select the number of clusters K** based on the use of the visualisation service (https://www.sciencehunter.net/Services/Clustering#/visualization)
    - Let **K = 3** (three major groups of political views).
2. **Normalise the data**
    - We bring all parameters (age, income, etc.) to a single scale.
3. **Running the K-means algorithm**
    - We divide respondents into 3 groups (by proximity to the centroids of the clusters).

Clustering results:

After running the algorithm, we get 3 clear groups:

Table 3. Clustering results

Cluster	Average age	Income	Education	Region	Political orientation
1. Young progressives	25 years	Medium	Higher	City	Left (2-4)
2. Conservative traditionalists	55 years	High	Higher	Village	Right (7-10)
3. Centrist Realists	40 years	Medium	Higher/secondary	City/village	Centre (4-7)

Conclusions and application:

- Parties can **tailor electoral strategies to** each cluster.
- **Political advertising** can be **targeted** (e.g. progressive ideas for young people, traditional values for the older generation).
- Used in **election forecasting**: for example, by analysing past data, it is possible to understand which group of the population tends to change its preferences.

This method is really applied in political analytics, helping political technologists and sociologists to better understand the electorate

A specific example of how DM has led scientists to rethink the relationship of gut bacteria to disease

Scientists at the European Molecular Biological Laboratory (EMBL) in Heidelberg have conducted a large-scale study[10] of the intestinal microbiota, which showed that inflammatory bowel disease and colorectal cancer may not be caused by the presence of "harmful" bacteria, but rather by changes in bacterial load (microbial population), which is influenced by many factors, including age, gender, diet and antibiotic use.

Using machine learning algorithms, the scientists created a model to estimate microbial load and found that such gut diseases may be closely linked to changes in microbial load rather than the presence of specific bacterial species. The study was based on metagenomic data from the gut microbiota, providing an overall picture of microbial load and its relationship to various diseases.

"We were surprised to find that many microbial species previously thought to be associated with disease were more associated with changes in microbial load," says study author Peer Bork. He says the discovery suggests that some microbes are more likely to be responsible for symptoms - bloating, diarrhoea or constipation - rather than gut disease per se.

The study also showed that, despite the difficulty of estimating microbial load in a laboratory setting, the machine learning approach allows scientists to develop a predictive model that can be applied to analysing human microbiomes in other studies. The model was used on large datasets comprising tens of thousands of metagenomes from around the world, from Japan to Estonia.

Study co-author Michael Kuhn said the algorithm has produced a tool that can estimate microbial load in faecal samples with high accuracy, making it easier to study the relationship of gut bacteria abundance to various diseases.

[10] https://www.sciencedaily.com/releases/2024/11/241113123236.htm

However, according to the authors of the study, their work has limitations: it only examined associations between microbial load and health, and scientists are not yet able to establish a direct causal link or explain the mechanisms underlying these relationships. In addition, the method has only been adapted for the human gut microbiome; new data and model adaptations will be needed to study large ecosystems, such as oceans or soils.

The researchers intend to further investigate microbial species that have a direct link to disease and could be used in the future as biomarkers for diagnosis.

PART 2. USING DATA MINING IN THE ECONOMY (BUSINESS)

Relevance of the research topic. The distinctive features of modern economy, also called knowledge economy (KE), digital economy, etc., are: the full development of innovations based on the application of the results of intellectual activity in the creation of new goods (services), rapid growth of information and communication sphere, expansion of social networks, increasing costs for educational and scientific needs, the preponderance of the service sector over production. In KE, innovation covers all areas of economic activity: services, products and goods, processes. At the same time, the very development of information technologies promotes both the possibility of knowledge growth and the use of knowledge in various spheres of life, accelerating scientific and technological progress and simplifying our lives. In this economy, technology companies are leading the way. Their success is based on digital technologies - working with data, information and knowledge. Schematically, the work of an enterprise in an KE can be summarised in Figure 1:

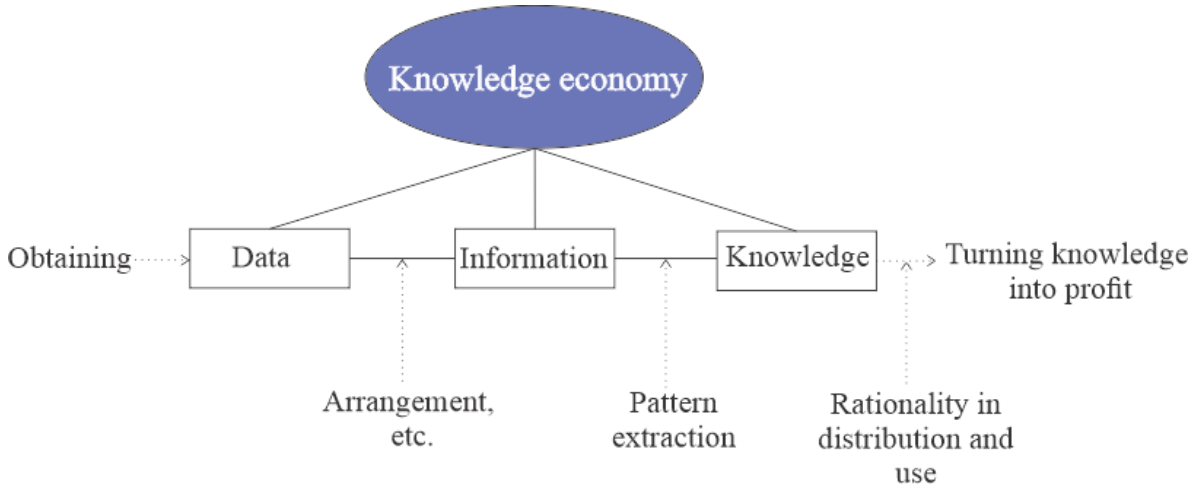


Figure 1. Shema of KE

** Source: suggested by the authors.*

Knowledge is at the heart of modern production - this is an undeniable fact. However, success requires not any knowledge, but knowledge in the key areas of competence of the firm (Hamel & Prahalad, 1994). At the same time, knowledge is most often divided into explicit, related to its formal representation and fixation in one form or another, and implicit, related to the human factor, for example, a vivid example

of it can be skills, which are the goal of the main efforts of most successful firms. But if the latter refer to the reflexive activity of a person and are difficult to manage from the point of view of their obtaining, the obtaining of the first type of knowledge is connected with the incorporation of the procedure of their obtaining into some business process, which is already largely manageable. In other words, the success of any firm's activity becomes dependent on the extent to which the acquisition of the first type of knowledge becomes a business process or technological process. Obtaining this type of knowledge is most often associated with research, and any research is the construction of a process model.

However, conducting such research has its own peculiarities, including in economics. Without understanding and taking them into account, it becomes virtually impossible to obtain acceptable results, which in economics can be considered in most cases to be financial or, simply put, profit. In this case, a lot of efforts and allocated resources are wasted. Hence the relevance of this study.

Analysing recent research and publications.

At present, the 'black' box model proposed in (Glinsky, 1965; Neuimin, 1984), which assumes that we do not have any preliminary theoretical considerations about it, is sufficiently promising and widely used to solve many problems of economics involving data analysis. This is the kind of model that is used in Data Mining (DM) and Big Data (BD), only without the above name.

DM can be used in carrying out cognitive activities within an organisation for different purposes and in different industries, as shown in Table 1:

Table 1. Spheres of Data Mining application in economy

Sphere	Application
Retail	Customer segmentation and basket analysis (improving advertising, developing strategies for stocking products, laying them out on the sales floor), researching time patterns (stocking), creating predictive models (promotional activities).

Continuation of Table 1.

E-commerce	Deep analysis of the user's activity history. If your user has recently read about the top 10 places to go on holiday this summer, you can present them with a lucrative offer from your travel agency.
Banking	Detecting credit card fraud (identifying anomalies and stereotypes), marketing policies for different customer groups - targeting and performance, predicting changes in the customer base (customer value models).
Telecommunication	Analysing call detail records (developing price and service sets), identifying customer loyalty (predicting customer persistence).
Insurance	Fraud detection (looking for stereotypes), risk analysis (reducing liability losses) and security.
Different areas of business	Predicting the popularity of certain characteristics of goods and services, guarantee policy (predict the number of customers), rewarding frequent ticket-buying customers.
Quality control	An actual technological task is quality control of goods, based on defect detection on the production line.
Inventory management	Assist with inventory management by predicting demand and determining the optimum amount of goods to stock, reducing costs.
Predicting energy costs	Predict energy costs, enabling companies to take action to reduce energy consumption and improve cost efficiency.
Product recommendation	Recommending products to companies that customers may be interested in based on their purchase history and preferences.
Marketing	By analysing user behaviour and in previous advertising campaigns, you can determine on which platforms and at what times you should advertise for maximum effectiveness.

Practically all large companies face in their activities to a greater or lesser extent the above-mentioned tasks, the solution of which makes it possible to find out the main factors of influence on customers, to focus on them and increase sales, i.e. to achieve the main goal of their activities.

Hence their interest in applying DM methods to solve these kinds of problems, because statistical methods of data analysis in these cases become unsuitable, as they are used to test pre-formulated hypotheses, but very often it is the formulation of a hypothesis and its subsequent justification is the main task in the analysis, because the subsequent decision-making requires knowledge, more precisely, knowledge of patterns, and not all patterns in the data are obvious at first sight. Moreover, statistical methods cannot answer a vague question, such as selecting the 10 best districts, because it is not clear what the best district is. And they cannot answer a question such as 'Is there a typical category of customers who do not pay taxes on time?', which in many cases ensures a successful decision. The answer to such and similar questions is exactly what DM application provides, and this is one of the main reasons why managers of enterprises of various profiles are interested in it.

But the most important thing here for the future is when the technology will become mass, when the implementation process will move from the stage of 'experiments' with a long time of system training to the stage of implementation 'out of the box', where the system itself, without human participation, will learn the peculiarities of a particular production, relying on local data and external information resources, observing the work of people, with the possibility of detailed interpretation of its conclusions and decisions for humans. It is too early to talk about it - this is the future of economics, although, in connection with this issue, we can recall the work (Pospelov & Pushkin, 1972), which describes a class of devices (gyromats) capable of creative acts. Becoming such a technology is precisely facilitated by the use of DM methods in solving problems of the economy.

So how does the process of creating new knowledge using DM methods occur at an enterprise? How is this process connected with the general process of cognition and why do we need to know it? And how can we use the results obtained to improve financial performance? Let us consider these questions in more detail.

The well-known works give an overview of the application of DM methods in relation to the tasks encountered in KE.

The work (Sviridova, 2019) is devoted to the topic of practical application of DM. Basic tasks used in data mining such as: outlier detection, association, cluster analysis, classification, regression, summarisation are formulated. The spheres of their application are considered without detailing the specifics, which significantly reduces the value of the results presented.

In (Isaichenkova & Novikova, 2019) it is shown that it is much cheaper to improve the processes of generation, processing and use of data if the enterprise has a technological base for it - this is the problem of many enterprises, as the latest information technologies are used mainly by large companies with sufficient capacity to invest in hardware and software complexes. They have R&D departments, which are busy with all the problems of applying DM methods in practice. At the same time, small and medium-sized enterprises, which make up the backbone of KE, do not have such an opportunity, besides, outsourcing services for big data analysis are relatively expensive, there is a shortage of necessary specialists, and therefore it is very important for them to understand almost all the details of implementation and use of DM methods for knowledge extraction with their subsequent monetisation, including increasing the efficiency of business processes.

In (Shipulev at al., 2023) it is shown that DM and Big Data technologies can bring commercial income in banking activities if marketing strategy is properly built. The study identified marketing opportunities for the application of this technology by banks, and noted the problems in the implementation of this technology by banks. It was revealed that there is no clear methodology of using DM and Big Data technologies in banking activities.

In (Bobamuradov, 2015), the main result is the classification of knowledge extraction methods. Two classes of methods are distinguished - the first class consists of communicative methods, which are oriented on direct contact of a knowledge engineer with an expert (knowledge source), the second class is textological methods based on knowledge acquisition from documents and special literature. The methods of knowledge extraction based on DM, which are the most promising in the case of solving many economic problems, are superficially described.

In the article (Bryzgalov & Yaroshenko, 2020) the stage of knowledge extraction in the design and creation of new products and services based on DM methods, in particular, Kohonen's self-organising map, was considered in detail. The knowledge obtained as a result of the analysis was used to form new tariffs and services of a cellular operator, in fact, to organise a new business. The peculiarity of the method is the necessity of preliminary structuring of data, otherwise other DM methods should be applied for knowledge extraction.

In (Mints, 2017) it is proposed to divide economic tasks of data analysis into two groups: predictive and descriptive. Each group is subdivided into several classes combining tasks with similar taxonomic features. The main classes of data processing tasks, such as ranking, sorting, filtering, data cleaning, quantisation, have been identified. However, the issues of selection and evaluation of the feature space, etc., which are directly related to the economic component of obtaining and using new knowledge using DM, have not been properly reflected.

The paper (Skvortsova & Skvortsov, 2014) studies the role of knowledge as an economic resource in the post-industrial stage of economic development. It is emphasised that knowledge is the basis of any production, but it is heterogeneous in its content and purpose, so for its better use we need a classification of knowledge based on the principles of economic character. Three types of knowledge are distinguished, which are used, in particular, in economics:

- embodied (objectified, objectified) knowledge, which exists in the objective form of tools, mechanisms, equipment and results of activity;

- personified (subjectivated) knowledge, knowledge related to a person or collective, so-called ‘non-verbalised or tacit knowledge’ - a person knows and knows how to do more than he/she realises;

- codified (fixed, formalised) knowledge, represented in sign form or in the form of symbols (oral or written text, formulas, pictures, images, descriptions, drawings, databases, computer programs, etc.). They are used in the production process as an independent economic resource.

Earlier, in Part 1, we have already noted the inherent production research in business, the purpose of which is not academic interest, but to increase the success of the business as a result of using the found model of DM. The focus is not on formal indicators of the model's success, such as recognition accuracy on a test sample, but on real indicators - increased labour productivity, improved service quality, customer inflow, etc. Since in production or business the model is a part of the service, which is used by millions of users every day. Hence the requirements and increased attention to business metrics, quality of input data, occurring errors, the ability of the model to interpret the results.

Researchers and businessmen wishing to apply DM methods in practice have a false sense of absolute suitability of the approaches adopted in solving DM problems and inherent in a purely academic template for their use in the interests of production and business. However, this is not the case, as the goal-setting in these tasks is completely different, hence the completely different range of tasks that need to be solved.

Thus, the above and similar works, with a few exceptions, traditionally describe the construction of this or that model and its testing in academic terms, without touching upon a completely different layer of problems arising when trying to use DM methods in the interests of production or business.

It should be noted that the works devoted to the practical application of Data Mining methods require a new approach to solving business problems, the difficulties encountered and the whole range of practical aspects. Particular attention should be paid to thinking about the problem itself in a broad context: the choice of feature space and business metrics, the process of collecting and processing raw data, the choice of DM method and its relationship with the theory of cognition as a key methodology. It is important to define the place of DM in this theory, its relationship with different directions in the philosophy of cognition and other related disciplines. This is of fundamental importance, since such clarifications actually form a roadmap for the use of DM methods, defining their possibilities, limitations and role in the research culture, especially in economics.

From the point of view of new business creation - the main thing is to identify the needs of society (or to create conditions for the emergence of needs) and to create innovative products or services to meet these needs. DM methods can be successfully used to analyse the needs of society. Beforehand, it is necessary to accumulate relevant data in a particular subject area or extract them from somewhere and then process them using DM methods.

For the existing business - identification of new niches in working with customers on the basis of those new opportunities provided by modern methods of data analysis, including DM.

In fact, the acquisition of machine models is preceded by the automatic finding of the so-called empirical regularities (ER), which are, as shown in (Polyakov at al., 2021), sources for the formulation of hypotheses, which, in turn, represent the most important component of scientific cognition, a form of natural science development, as noted in (Marx & Engels, 1958).

In today's environment characterised by a high rate of change in the business environment, the most important of the resources required for firms to compete successfully, and we have already mentioned this above, is knowledge in the so-called core competence areas of a business (Hamel & Prahalad, 1994), which provides its competitive advantage. However, knowledge inevitably becomes obsolete and this means that it needs to be replicated within a company's business processes in order not to lose competitiveness. At the same time, changes have also affected the views on knowledge - it is now viewed not only as a certain stock, but also as a certain flow at the same time.

In order to reproduce knowledge, it is necessary to set cognitive activity in the company as a well-managed business process. The objective of this process is to learn about the emerging realities in the market faster than its competitors. One of the most frequently quoted authors in this field, Arie de Geus, who was responsible for scenario planning at Shell, once made a very radical statement in this regard: the ability to learn faster than your competitors seems to be the only sustainable competitive advantage (Geus, 1988).

In other words, the ability to effectively perform cognitive activities within an organisation is the one key competency area that will ensure the long-term well-being of the firm.

Cognitive activity is the ability to analyse, generalise, problem solve and think creatively. These are important skills for any organisation that wants to be competitive and innovative.

Knowledge is a very complex object of research, which is in the centre of attention, first of all, of philosophers. However, from a practical point of view, within the framework of the above-mentioned goal, we can limit ourselves to some properties of knowledge, which are noted in (Nonaka & Takeuchi, 1995), namely: to distinguish 'explicit' or expressed knowledge (Explicit Knowledge) and 'hidden, implicit' or concealed knowledge (Tacit Knowledge).

'Explicit' or codified knowledge is the knowledge we are accustomed to, which can be set out in textbooks, books, other media, expressed in words, etc. However, we can know more than we can express. The part of knowledge that cannot be not only put on paper or other media, but even expressed in words, would be "hidden or implicit" or uncodified knowledge.

" Hidden" knowledge is personal, specific to the context in which a particular person - the bearer of this knowledge - is located. That is why it is difficult to codify and transfer it from one subject to another. Subjective insights, intuition, ideals, values and even emotions of an individual can be referred to the 'hidden' knowledge.

The authors in (Grimaldi, D., & Carrasco-Farre, C. 2021) point out that knowledge is only created by the individual and the role of organisations is to facilitate its creation by supporting and encouraging this process at the individual level. It is also beneficial for organisations to facilitate the knowledge created by individuals to "settle" at the group level through dialogue, discussion, sharing of experiences and direct observation, with "hidden", uncodified knowledge being a key component of knowledge.

One of the well-known experts in this field points out (Ruggles, 1988) that in practice, when talking about 'knowledge management' activities, Western companies refer to eight processes in the organisation:

1. creating new knowledge;
2. providing access to valuable knowledge from outside the organisation;
3. using existing knowledge in decision-making;
4. translating knowledge into processes, products and/or services;
5. representing knowledge in documents, databases, software, etc.;
6. stimulating knowledge growth through organisational culture and rewards;
7. transferring existing knowledge from one part of the organisation to another;
8. measuring the knowledge of the organisation;

The found knowledge is actually different variants of creating new products and services. Moreover, the deeper the level of abstraction of the found knowledge, the greater the level of future innovations. The most illustrative example in this respect is theoretically found Maxwell's equations, which gave a powerful impetus to all kinds of innovations in the field of radio communication, television and computer technology.

Those businesses that apply new knowledge and the technologies based on it tend to leap ahead and gain significant competitive advantages. At the same time, they realise that it is impossible to introduce new technologies and bring new products to market without trained staff. This is one of the reasons why firms are investing in the development of human capital - it is becoming a critical factor in the technology competition.

For several decades, the concept of KE has been used to describe a number of structural changes in the economy associated with the development of science and technology, intensive innovation, expansion of high-tech production, increased importance of intellectual labour, changes associated with computing technologies and networks, as well as other factors. As a result, new development trends have emerged and the economy has acquired qualitatively different characteristics, making the generation and utilisation of knowledge critical (Caliari & Chiarini, 2021). Digital and network technologies have opened a new era for humanity, transforming the ‘knowledge order’ - the conditions and methods of its production, transfer and utilisation - in all spheres of economic activity (Cheung & Leung, 2022). One of the

most important features of the economy has become an enhanced ability to operate with data (Mounier & Primbo, 2023), which at the same time has become a more significant source of value (North at al., 2018). Obtaining knowledge on the basis of data and, consequently, creating new value appeared to be related to the application of IDA (Shu, 2023).

The growth of data in the 'digital world' has led to the emergence of large volumes of data, known as Big Data, which has created a demand for specialised digital technologies to work with them, including their analytical processing, at a new level (Pohl at al., 2022; Wang at al., 2018). Big Data technologies have become an integral part of modern technological trends and have spawned new production, logistics, energy, communication and other systems (Ahmed at al., 2022; Kim at al., 2023). The use of data, including Big Data, has entailed the application of IDA to obtain the knowledge required for the normal operation and management of these systems, as well as the development of data-driven innovations (Kremer at al., 2019). The problem of utilising large amounts of accumulated data and extracting useful knowledge from them affects all sectors of the economy. However, this requires, first of all, a proper understanding of IDA and its capabilities, which led to the emergence of a new scientific direction - Data Science (Cao, 2017; Sarker, 2021).

Companies operating on the edge of efficiency quickly recognised the potential of IDA, especially in the digital sector, and began to actively develop related activities that are now becoming routine. Many applications and varieties of IDA have emerged, such as Data Analytics, Data Intelligence, Business Analysis, and Business Intelligence (Aryal, 2023; Kandel at al., 2012). The common factor for them is the growing scale of IDA applications due to the development and use of digital infrastructure and tools that have become integral elements of analytical work (Devi at al., 2021; Waters, 2023).

Although far from exhaustive, this review demonstrates the importance of considering IDA as a component of the knowledge-based economy. It also emphasises the need for a systematic approach to understanding its nature, methodology, purpose and cost-effectiveness. It is critical not only to recognise the relevance, opportunities and benefits of using IDA, but also to understand its inherent limitations in order to

effectively apply it in practice for knowledge generation. This will help avoid business situations associated with inflated and irrelevant expectations of IDA, which are often the cause of various user complaints. Such problems cannot be solved simply by increasing the amount of data analysed or the computing power used for analysis - they require methodological refinements.

The aim of the study: to clarify the foundations, opportunities and features of the application of IDA in the modern knowledge economy, and to provide a comprehensive assessment of the results of its use.

Methodology

In general, IDA refers to the field of data analysis based on the use of specialised mathematical techniques to process relatively large sets of heterogeneous data. The aim is to identify previously unknown hidden patterns (relationships, trends, etc.) that can be interpreted and that may be useful for practical application and/or further study to gain new insights. It is worth noting that IDA is applied in cases when it is not possible to identify these patterns in large data sets using traditional analytical approaches, including statistical ones. As a rule, the initial information for the application of IDA is a table of experimental data, in which the results of observations of objects are recorded. The considered datasets are either an object-property table (OPT), in which objects are characterised by sets of certain properties (parameters, attributes) with corresponding values, or a so-called training sample (TS, dataset), which, in fact, is a verified OPT, in which each set of objects is assigned (marked) to a certain class. The large number of properties allows for a more complete and detailed characterisation of objects using a variety of data. It is clear that the hidden patterns that may be present in these tables will surely be of interest to businesses for various purposes. The methods that are used to find patterns in these kinds of tables belong to IDA, within which machine learning methods can be distinguished, but this distinction is not important for the purposes of this paper.

Given the potential of using IDA to study various objects and solve diverse structural and analytical problems in virtually all sectors and spheres of the economy, the study will be based on the system approach. The methodology of this study lies at

the intersection of mathematics, statistics, computer science, management theory, economics and business, and can be extended to other areas of science and practice where IDA can be applied. Along with the use of general scientific methods of cognition (generalisation, systematisation, abstraction, induction and deduction, analysis and synthesis, analogy, comparison, formalisation, modelling, classification, categorisation) special methods of analysis (logical, structural, functional), descriptive method of research, interpretive methodology were also applied. The conceptual and guiding principles of this article are based on the concepts of knowledge economy, data economy and, to some extent, digital economy. In this regard, the peculiarities of internet, e-business and commerce development as well as several current technological trends have been taken into account. The principles of modern economic theory, business administration, economic analysis, decision-making theory, and cognitive theory were used as a basis for describing the trends and evaluating the results of using IDA in the economy.

Presentation of the main material of the study and the results obtained.

The main issue in economics is making managerial decisions, which should be reasonable and lead to optimal results. The basis for decision making is relevant information, which in this article is understood in the broad sense defined by W. Shannon (Shannon & Weaver, 1964) as "... anything that reduces uncertainty about the outcome of a particular event". In economics, depending on the level of understanding and justification of a decision, information can be conditionally divided into three types (Shannon & Weaver, 1964; Beshelev & Gurvich, 1973): 1 - ‘knowledge’, i.e. information supported by observations (evidence), verified by practice, providing a comprehensive understanding of the nature of a phenomenon and related cause-and-effect relationships. This type of information is the most objective and reliable, which allows making informed decisions; 2 - ‘assumption’, which means information that is partially confirmed, does not give a full understanding of the phenomenon and is less reliable. This allows making assumptions about problem solving and decision making; 3 - ‘opinion’, which means information that is the least reliable, based on limited knowledge of the phenomenon, with a predominance of subjective views. It is often quite

difficult to draw a clear boundary between these types of information. Nevertheless, this approach to the division of information allows to differentiate information by its reliability, to form assessments and judgements of different completeness, which can then be used for appropriate justification and decision-making on this basis. Naturally, this affects, first of all, the quality of decisions made (Fig. 2). Note that the transition from ‘opinion’ to “assumption” and then to ‘knowledge’ implies a qualitative leap in the understanding of the subject area.

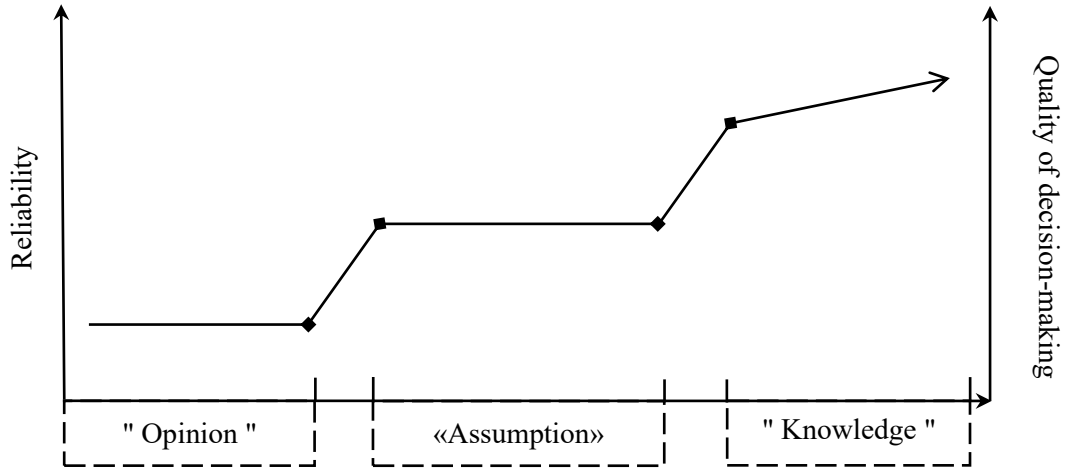


Figure 2. The conditional relationship between types of information differs in terms of the level of reliability and quality of decisions made

** Source: suggested by the authors.*

Thus, the quality of decision-making improves as the reliability of information increases, moving from opinion through assumption to knowledge, which is understood here as sufficiently reliable information on the basis of which informed decisions can be made related to strategy development, innovation creation, etc.

Improving the reliability of information is directly related to the processing of data, the volume and variety of which continues to expand. And IDA plays a huge role in improving such reliability, using it to improve such type of information widely used in economics as ‘assumption’, confirming or improving it with the help of found patterns and thus approaching such type of information as ‘knowledge’. Thus, the multidimensional nature of experimental data and the inherent limitations of human analysis in identifying and extracting complex patterns necessitate the use of IDA, which focuses on transforming data into highly reliable information and encourages the continuous development of such systems. In addition, IDA requires appropriate software

support that enables fast data processing, which is often crucial since the speed of obtaining the required information is often a critical factor. In other words, the ability to process the growing amount of data and use it effectively for analysis and decision-making through IDA is becoming an important strategic competence for modern companies, regardless of the industry in which they operate.

The application of IDA starts from the moment of obtaining a prepared set of data, which are formed on the basis of observations, experiments, measurements, functioning of information or technical systems, etc. Data can be collected manually or automatically using computer technology, sensors, measuring instruments, etc., and assessed on the basis of various criteria, including formalised form, structure, attribute characteristics of the object, interpretability, suitability for processing and volume (in terms of number of attributes and/or objects). In addition, they can be of different types - numeric, categorical, ordinal, interval and temporal. The content and quality of a dataset should be evaluated in relation to the problem to be solved and its potential to produce satisfactory results. However, this aspect often remains a weakness in data collection and analysis procedures. In some cases, the quality of data sets can be assessed using specialised mathematical approaches. The need for a particular type of data, as well as its structure, is determined by the specifics of the subject area under study, the problem to be solved, and the suitability of the selected indicators for the task at hand.

Knowledge derived from data as a result of applying IDA methods in economics is empirical. They can be called probabilistic knowledge and they most often prevail in practice, especially in the economy and social sphere. Nevertheless, on the basis of this knowledge, technologies and organisational models are created, social and technical processes are predicted and regulated.

Foundations of IDA. Due to the limitations of statistical methods in analysing multidimensional data sets of various types, in the second half of the 20th century a number of specialised mathematical methods emerged, which can be grouped under the general name of Data Mining (DM). These methods allow for the analysis of the datasets noted above and they are precisely the basis of modern IDA, which aims to effectively analyse complex datasets describing some subject area and identify patterns in them.

Given the nature of the data being analysed, these patterns will hereafter be referred to as empirical regularities (ER).

The widespread application of IDA in the economy is evidenced by the emergence of new terms emphasising its practical orientation, such as: Data Analysis, Data Analytics, Data Intelligence, Big Data Analytics, Business Analysis, Business Analytics, Business Intelligence. In practice, this has led to a progressive development of specific categorisation such as: business analysis (which is essentially a broader understanding of IDA applied to the business sector, with appropriate context and interpretation of results), business intelligence 1 (which is more focused on learning from the past, including descriptive and diagnostic analytics) and business intelligence 2 (which is more focused on understanding causes and informing future actions, including predictive and prescriptive analytics).

Without IDA, which relies on computer technology, it is virtually impossible to establish patterns in large amounts of data. In addition to a significant increase in computational capabilities, specialised digital technologies enable tasks that are virtually impossible to perform manually, such as 3D visualisation of a large dataset, which allows the quality of the collected data to be assessed in terms of the feasibility of solving the problem and formulating the initial hypothesis.

In this series, typical tasks solved by IDA have emerged and are summarised in Table 2. As a rule, the source of information for further analysis is OPT or TS, the formation of which is often beyond the competence of an IDA specialist (Data Scientist, Data Analyst).

Table 2

Main typical tasks to be solved in the framework of IDA

Task	General Description	Key methods/algorithms
Clustering (cluster analysis)	Grouping of objects on the basis of similarity into relatively homogeneous groups (clusters) without predefined categories (known data structure); data are represented as OPT;	k-means, spectral, hierarchical, fuzzy clustering; algorithms: mean shift, DBSCAN, FOREL, cut-based; Gaussian mixture model;

Continuation of Table 2

Classification (classification analysis)	Determining the category or class to which a new object belongs, based on a validated training dataset (TDS);	Structural-logical, graph, Bayesian methods; neural networks; Linear discriminant analysis.
Regression analysis	Predicting a continuous numerical value based on analysing relationships between variables; assessing relationships between a dependent variable and one or more independent variables;	linear, multiple, logistic, polynomial regression.
Dimensionality reduction	Transforming the dataset to reduce the number of variables by obtaining the main variables for data compression, visualisation, novelty detection, etc., while preserving important information;	Multidimensional scaling; factor analysis; principal component analysis; linear and generalised discriminant analysis.
Association detection (associative rules)	Finding relationships between different variables in large datasets; identifying consistent combinations of attributes in specific entities represented as rules;	algorithms: AIS, APRIORI, DIC, SETM, ECLAT, FP-Growth.
Anomaly detection	Detection of rare, anomalous data or unusual patterns that deviate from the general trend in a data set, which may indicate errors or significant events;	methods of cluster and classification analysis; statistical methods; methods of time series analysis.
Time series analysis	Analysing sequences of data to identify patterns over time, allowing past values to be estimated and future values in the series to be predicted;	autoregressive, integrated, ARCH, GARCH models; moving average models; wavelet transforms.
Forecasting	Forecasting future events or values in time series or other data sequences based on analyses of historical data;	statistical, Bayesian methods; Markov chains; neural networks.

The choice of IDA methods in practice depends on the task and the nature of the data, and is a separate issue beyond the scope of this article. The main stages of IDA are: 1) collection, structuring and preliminary preparation of data; 2) construction of

the feature space, including selection of indicators that are most important for describing objects and solving the task, construction of OPT or TS on their basis; 3) preliminary processing of the dataset (cleaning, filling in missing values, standardisation of separators, checking for outliers, etc.); 4) determination of the purpose and formulation of the working hypothesis of the study (for some tasks, data quality assessment may be carried out to confirm the suitability of using a particular dataset); 4) identification of the purpose and formulation of the working hypothesis of the study (for some tasks, data quality assessment may be carried out to confirm the suitability of using a particular data set); 6) interpretation of the results, i.e. explanation of the meaning and significance of the identified empirical regularities (ER) in the context of the problem to be solved; 7) practical application of the obtained results (in the form of: justification of managerial decisions, adjustment of existing approaches, modification of software algorithms, etc.). In many spheres of economy, business and scientific research, IDA is applied systematically, so the presented stages are repeated cyclically, which leads to additional functions (such as correction of the obtained models on the basis of continuous estimates, retraining of algorithms, etc.).

Going through all stages of IDA involves a large number of methodological issues, the scope and significance of which may vary significantly depending on the type of subject knowledge, including in the field of economics and business. For our purposes, we will identify the following main groups of issues: 1) problems related to the construction of the feature space (complexity of the subject area, low level of understanding of objects, completeness of their description, relationships between features, dependence on the context, etc.); 2) problems related to data (complexity of diversity, suboptimal volume, rate of change, lack of understanding of the value of data, etc.); 3) problems of problem formulation (lack of understanding of the context, interdisciplinarity, unclear understanding of the goals of problem solving, etc.); 4) problems related to obtaining empirical regularities (selection of function type, procedural and algorithmic issues, assessment of reliability of the obtained solution, etc.); 5) problems related to understanding and interpretation of results (evaluation of results, lack of sufficient knowledge about the subject area, etc.); 6) issues related to practical application

of results (transition from interpretation to implementation, understanding the limits of applicability of results, etc.).

Digital infrastructure and tools for IDA. As discussed earlier, conducting IDA today is virtually unthinkable without the use of computer-based and specialised technologies. The latter requires an examination of the digital infrastructure used for data preparation and the deployment of specialised software tools, which has effectively established a new approach to data-driven knowledge generation. The IDA infrastructure comprises specialised digital technologies, systems and services that support the collection, transmission and storage of data, as well as providing computing power for its analytical processing. The range of IDA infrastructure elements is quite extensive, and many of them are part of the fundamental level of data processing, which means that they are not only applicable in IDA. Moreover, it is often difficult to draw clear boundaries between the different categories of these technologies, systems and services because of overlapping functions. IDA tools include computational software, specialised applications and software libraries that automate operations such as pre-processing, pre-computation and visualisation, as well as direct data analysis, validation and visualisation. Table 3 summarises the main types of digital infrastructure elements and IDA tools.

Table 3

Main types of elements in digital infrastructure and IDA tools

Infrastructure	**Instruments**
– various types of database management systems, including relational, graph, document-oriented, column-oriented, hybrid, key-value, etc., as well as specialised analytical repositories and services (Apache Pinot, Apache Druid, Amazon Redshift, BigQuery, Snowflake, Greenplum, Apache Spark, Microsoft SQL Server and Microsoft Analysis Services);аналитические хранилища и	– special programming languages and computing environments (R, MATLAB, Julia, Scala, including libraries such as MLlib, Breeze, Smile) and Python (libraries such as Numpy, Pandas, Scikit-learn, Scipy, Xgboost, Matplotlib, Keras, Pytorch, Tensorflow, and Theano); – spreadsheet application software packages (Microsoft Excel, Google Sheets);

Continuation of Table 3

– сервисы (Apache Pinot, Apache Druid, Amazon Redshift, BigQuery, Snowflake, Greenplum, Apache Spark, Microsoft SQL Server и Microsoft Analysis Services); – cloud services, including specialised services for data processing (Amazon Web Services, Google Cloud Platform, Microsoft Azure); – real-time data processing systems that provide infrastructure to transmit, store and process data streams (Apache Flink, Apache Spark Streaming, Apache Kafka, Apache Kudu); – dataflow management systems (Apache Airflow, Dataform, Alteryx Analytics); – specialised services for data analysis (Databricks, Azure Data Factory, Azure Machine Learning, Azure Databricks, Azure HDInsight, Amazon EMR, Google Cloud Dataproc); – service infrastructure for AI developers (Amazon SageMaker, Azure Machine Learning, MLflow wandb.ai).	– software applications (SPSS, Minitab, Stata, Statgraphics, JMP) and data analysis environments/platforms (Weka, Orange, RapidMiner, KNIME); – software for data visualisation and analysis results (Microsoft Power BI, Tableau, QlickView, Qlik Sense, Amplitude, Google Looker Studio, Luxms BI, Redash, BeX Analyzer, SAP BusinessObjects, IBM Cognos Analytics). – web-based platforms with IDA tools (such as Tableau Online, ScienceHunter, Mode Analytics, Plotly Chart Studio) as well as online tools developed as part of integrated cloud systems (such as Amazon QuickSight, Amazon Athena, Amazon Redshift, Amazon SageMaker, Google Cloud Dataprep, Google Cloud AI Platform, Azure Synapse Analytics, Azure Machine Learning, and Azure Data Factory).

The IDA infrastructure provides new opportunities for data manipulation and collaboration, as well as a foundation for management, process (flow) automation and deployment of analytical tools. As the volume of data and real-time execution of IDA increases, the importance of infrastructure as a prerequisite for its implementation and as a factor in its effectiveness becomes more critical.

As for IDA tools, they include developed methods and algorithms that automate routine computational tasks and also perform certain cognitive operations that were

previously performed by humans. That is, the fields of computer applications are undergoing a peculiar evolution: from basic computing to networking, and now to a new, just-beginning stage centred on supporting human cognitive capabilities.

At this stage, only humans are still capable of formulating analysis tasks, selecting the most appropriate processing methods, evaluating the results and interpreting them. Therefore, IDA tools, for the most part, cannot yet fully replace a human being, and in practice human-machine technology of data analysis is used. With the expansion of IDA applications, there is a tendency to develop IDA tools not only for specialists (Data Scientists, Data Analysts), but increasingly for professionals in various subject areas who have a better understanding of the problem and its context, but do not have a deep background in mathematics and computer science. Thus, new digital tools expand the range of professionals who can use IDA, especially if they are implemented as web-based systems and cloud services. These tools provide wider access and democratise the field, making it more widespread rather than limited to a narrow group of experts. For example, 3D visualisation can help in clustering and assessing the informativeness of OPT and its features (relevant tools are available on the ScienceHunter portal // https://www.sciencehunter.net). This approach greatly enhances the effectiveness of IDA, especially when applied systematically.

In addition, it is important to highlight tools in the Large Language Model category, such as ChatGPT. Advanced versions of these tools include new features for analysing computational data, including those provided by plugins such as Wolfram.

Results from the use of IDAs. In general, the application of IDA is consistent with the universal scheme of cognition: from ‘living contemplation’ (real experience) to abstraction, and then from abstraction to practice. Indeed, ‘living contemplation’ actually means data collection and construction of OPT (or ER), their processing means detection of ER, which, as a rule, are the basis for proposing certain hypotheses that can become, after verification and detailed elaboration, the very abstraction sought, with its further application in practice. At the same time, it is important to realise that the results of IDA are not yet that abstraction, these results are empirical knowledge, despite the numerous mathematical constructs used to extract ER (Polyakov at al., 2021).

As a rule, the application of IDA leads to: grouping of objects, a solving function, a model distinguishing objects of different classes, etc., which, in fact, constitutes inductive inference. All of them represent ER in the form of some generalisations, patterns in the studied sections of the subject area. These patterns serve as preliminary generalisations of facts, descriptions of the general structure and primary understanding of the state of subject knowledge, addressing such questions as ‘What?’, ‘How?’, ‘Who?’, ‘What is happening?’ and ‘What is the situation?’.

The resulting ER identify what is important, replacing guesswork with well-informed insight and providing some awareness and conviction weighted by plausibility on the data. More often than not, this provides hindsight, but allows for educated guesses whose reliability is greater than mere guesswork. Moreover, when appropriately interpreted and justified, the ER found can be considered empirical knowledge. This allows determining the subsequent actions, such as formulating hypotheses, creating methods, instructions, rules, conducting evaluations and forecasts. These ER can later serve as a basis for the discovery of deeper ideas, but they provide only a preliminary level of cognition (understanding). As shown in (Cao, 2017), this is a limitation of the application of IDA from an epistemological point of view. Often such empirical knowledge (assumptions) is sufficient for current practical business activities for making operational management decisions (‘Data-to-Decision’) that do not require a deep understanding of the essence of phenomena and processes or cause-and-effect relationships. For example, this includes situational segmentation of consumers in the market, targeted advertising, product price forecasting and similar actions that allow realising short-term economic benefits. However, this is not enough to develop strategic decisions in large companies, and even more so at the national economic level. Therefore, comprehension of the studied subject knowledge in economics, achieving a deep understanding of the nature of phenomena and processes, obtaining reliable knowledge cannot be based solely on the ER extracted from the data set. They require further establishment and substantiation of cause-and-effect relationships to answer such questions as: ‘Why?’, ‘How does it happen?’ and ‘Why does it happen this way?’. And this marks a shift from the empirical level of cognition

to deeper understanding, reaching the theoretical level, where the ability to predict and make long-term, more complex decisions is enhanced, embodying the Data-to-Explain-Decision approach. Such knowledge is essential for strategic decision-making in economics and business, as well as in scientific research, where empirical methods of cognition precede theoretical developments.

Thus, it is important to properly understand and apply IDA, given its inherent limitation in the form of ER and its role in the broader cognitive process. The automatic generation of these patterns (based on digital technologies) helps to better understand reality, and this is what IDA provides. The empirical patterns found serve as crucial additional material for formulating research hypotheses or abstractions, which are then tested to interpret causal relationships. This is the main limitation of IDA in the cognitive process, and the main opportunity it creates. It can be concluded that IDA facilitates the "automation" to some extent of human cognitive activity, enabling the processing and analysis of large datasets, especially multidimensional ones. Taking into account the important role of a human in setting tasks, interpretation and implementation of results, the application of IDA requires cooperation between subject matter experts and IDA specialists (Data Scientists, Data Analysts), which can be (and this is confirmed by practice) rather complicated in terms of interaction and communication organisation. This is one of the good reasons to develop IDA tools specifically for subject matter experts, which do not require advanced qualifications in mathematics and computer science.

In view of the above, IDA becomes a crucial component of cognition, both in business and other practical activities, as well as in science. Summarising the various aspects of the impact of IDA on cognitive abilities, it is important to note the following: 1) IDA offers fundamentally new and enhanced opportunities to process empirical data and, therefore, to improve human cognitive abilities based on these data. In this regard, IDA both stimulates and partially replaces human cognitive processing; 2) IDA develops human's natural abilities to analyse, evaluate, formulate hypotheses, create models and, ultimately, to better understand phenomena. It also stimulates analytical, critical, abstract and evaluative thinking in individuals; 3) IDA supports active human interaction with

the world through data, including improving the intellectual work of describing objects with data, making measurements, quantitative and qualitative evaluation; 4) IDA raises the constructive/evidential method of cognition to a new level, incorporating both its advantages and limitations, and promotes the emergence of new forms of cognitive activities (types of analytics); 5) IDA, with its wide range of applications, promotes interdisciplinary research, integrating knowledge from mathematics, statistics, computer science, management and different types of knowledge such as economics, sociology and psychology. The use of IDA can be said to foster a new cognitive framework and transform human intelligence. For example, digital tools enable the automation of routine computational tasks, allowing more attention to be paid to intellectual and analytical activities.

It is also important to mention a special class of IDA models as neural networks. Their use in some cases gives satisfactory results of data processing, but in most cases does not contribute to understanding, as it does not eliminate the gap between the obtained decisive rule and the explanation of cause-effect relations. Therefore, their level is limited to 'primitive' recognition (classification). The replacement of human supervision in the process of cognition by artificial intelligence, at least for the moment, does not seem feasible.

IDA Appendices. The sphere of practical application of IDA, along with Big Data technologies, is rapidly expanding. De facto, it has become an integral part of research in economics and various scientific fields, and is also widely used by businesses to make managerial decisions and automate processes in such areas as marketing, logistics, finance, security, production, and management. Table 4 shows examples of business intelligence problems that are solved based on IDA (Ahmed at al, 2022; Kim at al., 2023; Kremer at al., 2019; Sarker, 2021; Cao, 2017; Tsai & Kang, 2019; Hallman & Gelman, 2021; Batko & Slenzak, 2022; Baum at al., 2018; Lehenchuk & Zavaliy, 2023; Pagano & Lyotin, 2019).

Table 4. Examples of practical applications of IDA for key types of DM tasks

Typical DM tasks	Examples of practical applications
Clustering (cluster analysis)	Grouping consumers (social media users, service users, etc.), competing companies, territories (markets), company promotions, etc.; grouping products by seasonality, demand and other characteristics to optimise stock; creating catalogues; and evaluating and selecting job candidates.
Classification (classification analysis)	classification of consumers (users) by loyalty level, increase or decrease in potential demand; credit scoring; classification of fraudulent transactions, borrowers, agricultural land, incoming documents (spam detection), etc.; detection of fake accounts; prediction of equipment failures; product quality assessment and defect detection.
Regression analysis	forecasting macroeconomic indicators, asset values, prices and demand for goods (services), energy consumption, consumer behaviour, etc.; assessing factors affecting sales; optimising production processes by assessing productivity and product quality factors; investigating factors affecting the level of environmental pollution.
Intelligent analysis of associative rules (Associative Rules)	Analysis of consumers' shopping baskets to identify combinations of products that are frequently purchased together for recommendations and assortment optimisation; Detection of links between different themes or user behaviours in online services for content management and ad targeting; Detection of suspicious transaction patterns for fraud detection; Identification of failure links to manage maintenance processes.
Anomaly detection	fraud detection in financial transactions; fault and failure detection in production equipment; network traffic analysis to detect network attacks or abnormal behaviour; medical diagnostics to detect rare diseases.
Time series analysis	Forecasting prices for goods (services) and demand (sales volume), identifying seasonal fluctuations for production planning, IDA marketing strategies, inventory management; forecasting stock prices, exchange rates, energy consumption, traffic flows, internet traffic, user activity, etc. to improve management efficiency; justifying potential material and product requirements to optimise supply chains; analysing equipment performance for maintenance planning; monitoring of medical services, etc. to improve the efficiency of management; forecasting the quality of medical services; forecasting the quality of medical services; and forecasting the quality of medical services.

Examples of tasks solved by IDA demonstrate both its wide range of applications and its potential for use in various industries and for solving diverse tasks. In recent years, in addition to e-commerce, the application of IDA has rapidly expanded in industrial and agricultural production management, logistics processes, various service industries and supply chains. This has enabled the integration of IDA into current trends

related to Industry 4.0 and digital transformation, including the intellectualisation of web-based systems, equipment and various devices. In this case, IDA is used in computer programmes that control systems, allowing them to collect and process data and use the information to make decisions. A prime example of digitalisation and intellectualisation is healthcare, where IDA are used both to manage healthcare facilities and to diagnose diseases, as well as to develop tools (technologies) to treat them.

All of the above expands the scope and role of IDA in the economy. The ability to process large volumes of data is becoming one of the most significant challenges faced by most businesses in any industry. The key prerequisites for this are the increasing number of specialists and the emergence of digital IDA technologies that facilitate the handling of data. At the same time, at the same time, there is often poor IDA performance in practice. In many cases, this is due to a lack of understanding of its purpose and inherent limitations, resulting in inflated expectations and irrelevant problem statements, ineffective management of IDA processes, and poor communication between IDA specialists and subject matter experts.

Discussion.

IDA is a relatively new and rapidly developing field of analytics, and it is a widely used method of knowledge production in the economy. Thus, the conducted research expands the understanding of IDA as a part of the knowledge economy and the specifics of knowledge production, as well as the data economy, which has its own distinctive features (Cheung & Leung, 2022; Mounier & Primbo, 2023; North at al., 2018; Wang at al., 2015; Tsai & Kang, 2019). Given current practices in the field of IDA, it has become one of the main applications of digital technologies (especially in relation to big data). This paper offers a deeper understanding of the subject (Pohl at al., 2022; Batko & Slenzak, 2022). The universality of IDA across all industries and sectors has been confirmed, indicating its applicability to a wide range of tasks, which increases its relevance to the economy (Ahmed at al., 2022; Lehenchuk & Zavaliy, 2023; Grimaldi & Carrasco-Farre, 2021). It is particularly important to clarify the understanding of IDA results, their limitations and value, in particular the usefulness of the data (Caliari & Chiarini, 2021; North at al., 2018; Shu, 2023). The expanding

scope of IDA in the economy, as well as its relationship to current technological trends and economic transformations (Kim at al., 2023), requires further discussion on the possibilities and effective use of IDA. It also requires the development of markets emerging around it, including data, services, data processing technologies and analytical tools.

It is important to promote the dissemination of IDA applications in various industries, especially among specialists who do not have specialised training in mathematics and computer science. This will increase intellectual activity in the economy, give impetus to the production of new knowledge, and promote the development of a creative class based on the principles of freedom in accordance with the ideology of the 'knowledge economy'. The first prerequisite is proper basic training of IDA specialists. The second is the creation of web platforms providing open access to IDA tools for a wide range of specialists. In addition, such platforms can serve as a basis for broad co-operation in the development of IDA applications; facilitating professional (scientific) communication in this field, in particular between business and science. In some large companies, research organisations and possibly in specific sectors of the economy, it is advisable to develop digital environments or ecosystems for IDA. Such an approach will foster synergies and have a significant positive impact on the knowledge economy (Grimaldi & Carrasco-Farre, 2021). It is obvious that the use of IDA will expand widely in the future, so it is important to outline several recommendations for its development (Table 5).

Table 5. Trends and general recommendations for the development of IDA within the 'knowledge economy'

Trends	Recommendations
Digital infrastructure for data processing	developing new generations of IDA infrastructure systems with a focus on automating processes and operations, intellectualising and expanding functionality and services. This includes the deployment of various IDA tools, the creation of new communication mechanisms and the promotion of co-operation between participants in the data economy. Improving the IDA infrastructure can be linked to the creation of corporate and industry data ecosystems.

Continuation of Table 5.

IDA tools	creation of automated tools for IDA, industry-specific solutions for a wide range of users, applications for experts in various fields, business analysis and other areas adapted to the current specifics.
Legal conditions, safety and standardisation	Improving the legal environment for the collection, transfer and use of data, including new types of data, especially those of public interest. While protecting confidentiality, it is important to create conditions for data sharing and joint projects, especially in the framework of science-business collaboration. One of the key conditions for co-operation is the improvement of standards for data collection and structuring, which facilitates the integration of databases and improves the quality of subsequent processing.
Collaborative working models in the data domain	This includes data consolidation and integration, improvement of digital infrastructure and IDA tools, improvement of standards in the field, research and application of results. The legal conditions for cooperation should be complemented by new organisational and technological mechanisms for data exchange, especially between business and science, which will require changes in academic policy. Models of cooperation may include: competition organisation systems, industry and professional associations, specialised data exchange platforms ('science-science', 'science-business', 'business-business'), and systems for conducting joint projects.
IDA training	In addition to Data Science training, basic IDA training is required for a wide range of professionals in different fields and levels, including social sciences and humanities, various business fields and more, in line with the development of digital IDA tools and data access.
Improving data openness and democratisation of the IDA	In the context of the recommendations provided, which create opportunities for wider application of IDA in the economy, it is advisable to accelerate the development of 'open data' projects (available for wide use) in different sectors of the economy and areas of public life. For this purpose, various forms of data presentation can be adopted, but they should be accompanied by access to IDA tools, e.g. through online platforms.

Both self-organisation of enterprises and R&D organisations and targeted government policy are necessary for the implementation of these trends in the development of the IDA industry. This is especially important in areas of public interest, such as the development of education and democratisation of intellectual opportunities.

Conclusions and prospects for further research

At the current stage, countries are focusing their economic development on building a knowledge economy, which is centred on the creation and use of new knowledge. One of the most important methods for generating new knowledge today is Intelligent Data Analysis (IDA), which aims to efficiently analyse large multidimensional datasets describing a particular subject area of knowledge in order to identify previously hidden patterns. Several key tasks are addressed using specialised IDA techniques, including clustering, classification, regression analysis, dimensionality reduction and association detection, among others. Modern IDA approaches are based on the application of specialised digital technologies, which can be categorised into IDA infrastructure and tools. These technologies automate the relevant operations at all stages and significantly increase efficiency. The application of IDA is generally consistent with a universal cognition scheme. At the same time, it is crucial to interpret the results of IDA correctly, recognising its inherent limitation: the generation of ER. On the other hand, the automatic generation of such patterns using digital technology provides essential support for cognition, as the ER provides additional material for further elucidation of causal relationships. IDA is a very versatile method of knowledge acquisition, which leads to a rapid expansion of its practical applications, especially when combined with big data technologies. With this in mind, the data economy emerges, encompassing the various sectors, processes and activities related to the accumulation and use of data. The data economy is a system of markets with its own characteristics and development challenges. One of the most important issues is to understand how value is extracted from data. To solve this problem, it is necessary to consider the usefulness of data, which is determined by the value of the information contained in them and the usefulness of knowledge (ER) obtained through IDA, as well as the results (effects) of applying the results of IDA in practice. In the course of IDA, the usefulness of data is gradually revealed through a step-by-step transformation of unstructured data to the generation and interpretation of empirical patterns. The points formulated in the paper contribute to the current academic discourse on the application of IDA in the knowledge economy and provide general recommendations for advancing

IDA in the following areas: digital infrastructure and IDA tools, legal environment, collaborative data models, training, increasing data openness and making IDA accessible. These areas are expected to be the focus of future research.

References

1. Ahmed, R., Shaheen, S., & Philbin, S. (2022). The role of big data analytics and decision making in achieving project success. *Journal of Engineering and Technology Management, 65*, 101697. https://doi.org/10.1016/j.jengtecman.2022.101697
2. Aryal, S. K. (2023). The impact of digital regime on academic knowledge production. *University West: School of Business, Economics and Information Technology*. https://www.diva-portal.org/smash/get/diva2:1746631/FULLTEXT01.pdf
3. Batko, K., & Slenzak, A. (2022). Utilizing big data analytics in healthcare. *Journal of Big Data, 9*(1), Article 3. https://doi.org/10.1186/s40537-021-00553-4
4. Baum, D., Larocque, C., Ozer, B., Skoog, A., & Subramanian, M. (2018). Applying big data analytics and related technologies in maintenance - a literature-based study. *Machinery, 6*(4), Article 54. https://doi.org/10.3390/machines6040054
5. Beshelev, S. D., & Gurvich, F. G. (1973). *Expert evaluations*. Science.
6. Bobamuradov, O. D. (2015). Stages of knowledge extraction from electronic information resources. *Technical Sciences. Eurasian Union of Scientists (EUU)*, 10(19), 130-133.
7. Bryzgalov, A. A., & Yaroshenko, E. V. (2020). Application of Data Mining methods in the design and creation of new products and services. *Open Education*, 24(6), 14-21.
8. Caliari, T., & Chiarini, T. (2021). Knowledge production and economic development: empirical evidence. *Journal of Knowledge Economics*, *12*(2), 1-22. https://doi.org/10.1007/s13132-016-0435-z

9. Cao, L. (2017). Data science: a comprehensive review. *ACM Computing Surveys, 50*(3), Article 43. http://dx.doi.org/10.1145/3076253
10. Cao, L. (2017). Data science: challenges and directions. *Communications of the ACM, 60*(8), 59-68. https://doi.org/10.1145/3015456
11. Cheung, K. K., & Leung, W. (2022). A critical review of the antecedents of knowledge economics and their contemporary research: implications for the computerized new economy. *Journal of Knowledge Economics*, *13*(2), 1573-610. https://doi.org/10.1007/s13132-021-00734-9
12. Devi, K. G., Rath, M., & Thich Dieu Linh, N. (Eds.) (2021). *Artificial intelligence trends for data analysis using machine learning and deep learning approaches*. CRC Press, Taylor & Francis Group, LLC.
13. Geus, A. (1988). Planning as learning. *Harvard Business Review*, 66(4), 70-74.
14. Glinsky, B. A. (1965). *Modeling as a method of scientific research*. Science.
15. Grimaldi, D., & Carrasco-Farre, C. (2021). Implementing data-driven strategies in smart cities: A roadmap for urban transformation. *Elsevier Science.* https://doi.org/10.1016/C2019-0-01442-2
16. Hallman, D., & Gelman, A. (2021). Challenges of implementing exploratory data analysis into statistical workflows. *Harvard data science review, 3*(3). https://doi.org/10.1162/99608f92.9d108ee6
17. Hamel, G., & Prahalad, C. K. (1994). *Competing for the future*. Harvard Business School Press.
18. Isaichenkova, V. V., & Novikova, A. V. (2019). Digitalization as a tool for improving the efficiency of business processes. *Modern Economy Success*, 3, 141-144.
19. Kandel, S., Papke, A., Hellerstein, D. M., & Heer, D. (2012). Enterprise data analysis and visualization: An interview study. *Proceedings IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2917-2926. https://doi.org/10.1109/TVCG.2012.219
20. Kim, M., Lim, K., & Xuan, D. (2023). From technological factors to the circular economy: Insight from a data-driven review of servitization and product-service

systems in Industry 4.0. *Computers in Industry, 148*, 103908. https://doi.org/10.1016/j.compind.2023.103908

21. Kremer, D., Walley, D., & Batura, O. (2019). The data economy and data-driven ecosystems: regulation, frameworks, and case studies. *Telecommunications Policy, 43*(2), 113-115. https://doi.org/10.1016/j.telpol.2018.12.007
22. Lehenchuk, S., & Zavaliy, T. (2023). Big Data in IAD marketing analytics: Opportunities and challenges of utilization. *Problems of Theory and methodology of accounting, control and analysis, 54*(1), 52-58. https://doi.org/10.26642/pbo-2023-1(54)-52-58
23. Marx, K., & Engels, F. (1958). *Sochineniye* (4). Moscow: Publishing House of Political Literature.
24. Mints, A. (2017). Classification of tasks of data mining and data processing in the economy. *Baltic Journal of Economic Studies*, 3(3), 47-52. https://doi.org/10.30525/2256-0742/2017-3-3-47-52
25. Mounier, P. , & Primbo, S. D. (2023). Knowledge maintenance and infrastructure management in the digital age: an integrated view. *HAL Open Science*. https://hal.science/hal-04309735
26. Neuimin, Y. G. (1984). *Models in science and technology: History, theory, practice*. Science.
27. Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.
28. North, K., Mayer, R., & Haas, O. (2018). Creating value in the knowledge economy using digital technologies. In K. North, R. Mayer, & O. Haas (Eds.), *Knowledge management in digital change* (pp. 1-29). Springer. https://doi.org/10.1007/978-3-319-73546-7_1
29. Pagano, A. M., & Lyotin, M. (2019). Technology in supply chain management and logistics: Current practices and application perspectives. Эльзевир. https://doi.org/10.1016/C2017-0-04194-0

30. Pohl, M., Stegemann, D. G., & Turowski, K. (2022). Performance benefits of data analytics applications. *Procedia Informatica, 201*, 679-683. https://doi.org/10.1016/j.procs.2022.03.090
31. Polyakov, I., Khanin, G., Shevchenko, V., & Bilozubenko, V. (2021). Data mining as a cognitive tool: Capabilities and limits. *Knowledge Management and performance*, *5*(1), 1-13. http://dx.doi.org/10.21511/kpm.05(1).2021.01
32. Pospelov, D. A., & Pushkin, V. N. (1972). *Thinking and automata*. Sov. radio.
33. Ruggles, R. (1998). The state of the notion: Knowledge management in practice. *California Management Review, 40*(3), 80-89.
34. Sarker, E. H. (2021). Data science and analytics: A review of data-driven intelligent computing, decision making, and applications. *SN Computer Science, 2*, Article 377. https://doi.org/10.1007/s42979-021-00765-8
35. Shannon, K. E., & Weaver, W. (1964). *A mathematical theory of communication*. University of Illinois Press.
36. Shipulev, E. O., Bykanova, N. I., Goncharenko, T. V., & Korotkova, I. S. (2023). Evolution of Big Data technology development and its marketing opportunities in promoting banking products and services. *Modern Economy Success*, 3, 106-111.
37. Shu, H., & E, Y. (2023). Knowledge discovery: Methods for data mining and machine learning. *Social Science Research, 110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817
38. Siemanski, I. S. (2023). Smart urban mobility: Transportation planning in the era of big data and digital twins. *Elsevier Science.* https://doi.org/10.1016/C2019-0-01443-4
39. Skvortsova, V. A., & Skvortsov, A. O. (2014). Knowledge as an economic resource. *Izvestiya vysshee obrazovaniya [Izvestia of higher educational institutions. Povolzhsky region. Economic Sciences. Innovations in economics*, (1), 12-21.
40. Sviridova, L. E. (2019). Practical application of Data Mining. *Science Alley*, 2(29), 917-920.

41. Tsai, D. K.-A., & Kang, T.-K. (2019). Mutual intention in knowledge seeking: Exploring social exchange theory in an online professional community. *International Journal of Information Management, 48*, 161-174. https://doi.org/10.1016/j.ijinfomgt.2019.02.008
42. Wang, H., White, L., & Chen, H. (2015). Big data research for the knowledge economy: past, present, and future. *Industrial Management and Data Systems, 115*, 9. http://www.emeraldinsight.com/doi/full/10.1108/IMDS-09-2015-0388
43. Waters, D. D. (2023). An emerging digital infrastructure for humanities research. *International Journal of Digital Libraries, 24*, 87-102. https://doi.org/10.1007/s00799-022-00332-3

Appendix 2. How DM works in solving business problems.

At present, ‘the continuous acceleration of market dynamics leaves fewer and fewer chances for those economic structures that ’feed" mainly on explicit knowledge obtained from external sources. Without denying the great importance of the market of formalised knowledge and its role in the development of modern innovations, however, we note that strategic market sustainability arises in those firms that do not buy, but produce new knowledge themselves, transforming it into key competencies". However, it should be clearly understood that it is possible to produce such knowledge and transform it into innovations only in the presence of an appropriate research and innovation environment. In this case, it is necessary to take into account the famous expression of V.I. Vernadsky ‘Science is the collective creativity of free individuals’ and, in our opinion, to supplement it with the following - ‘Innovation is the collective creativity of free individuals’. The key point in these expressions is collective creativity, which means, in fact, the involvement of the project approach.

General approach using sales forecasting as an example

Research: Modelling the dynamics of sales of goods depending on external factors.

Purpose: Determine how various macroeconomic factors (e.g., inflation, exchange rates) and internal parameters (seasonality, marketing) affect sales.

- **Methods**:
  - Collect historical data on sales and related metrics.
  - Using regression and time series algorithms to create predictive models.
  - Using clustering to segment customers (e.g., by income levels or purchase frequency).
- **Result**: Data Mining allows you to forecast sales volumes with high accuracy, identify key growth drivers and optimise your marketing strategy.
- **Practical relevance:** The results help companies better allocate resources, avoid overproduction and plan logistics efficiently.

Algorithms used

- **In the economy**:
  - Regressions: Linear regression, LASSO regression for factor analysis.
  - Time series: ARIMA and Prophet for forecasts.

Example 1. As a simple example, let's consider the use of one of the DM methods - classification trees to solve marketing problems. Table 6 below presents a training sample with the attributes Gender, Age, Number of children in the family, Income and the target attribute - Intention to buy. Table 7 shows the same training sample after coding.

Table 6.

gender	age	Number of children in the family	income	Y=intention to buy
F	25	1	very small	weak
M	50	2	small	weak
M	23	0	high	weak
M	38	2	small	weak
M	32	1	very small	weak
F	30	2	high	strong
F	20	0	very high	strong
F	28	1	high	strong
F	33	1	medium	strong
M	40	2	very high	strong

Coding: income=1- very low; 2-small;3-medium;4-high;5-very high

Age <=20=0, 21-25=1, 26-30=2, 31-35=3, 36-40=4, 41-45=5, 46-50=6, >50=7

Table 7.

gender	age	Number of children in the family	income	Y=intention to buy
0	1	1	1	1
1	6	2	2	1

1	1	0	4	1
1	4	2	2	1
1	3	1	1	1
0	2	2	4	2
0	0	0	5	2
0	2	1	4	2
0	3	1	3	2
1	4	2	5	2

Fig.3 shows the SR - classification tree

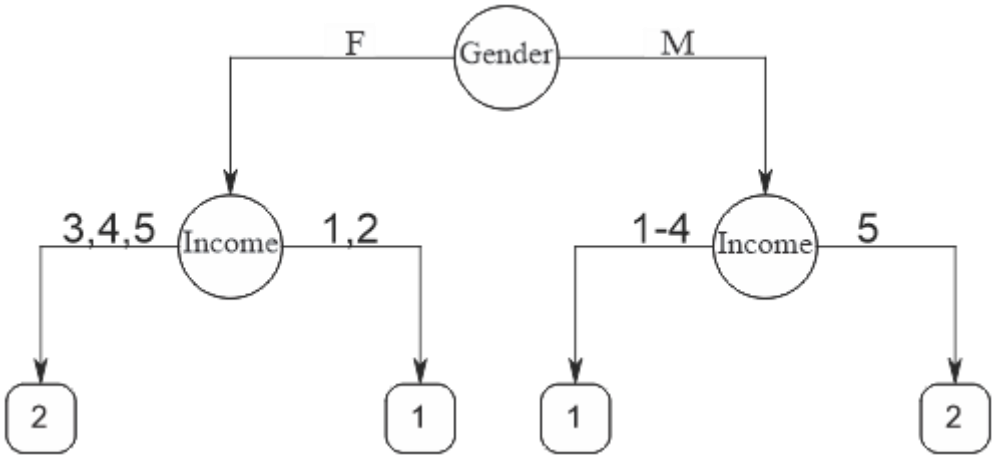


Figure 3. Classification tree

* *Source: suggested by the authors.*

Empirical regularities found (ER) extracted from the classification tree:

1. If gender=W and income=medium or high or very high, then purchase intention = strong.
2. If gender=M and income= very high, the intention to buy = strong.
3. If gender=G and income= very little or little, THEN purchase intention = weak.
4. If gender=M and income=very small or small or medium or high, then purchase intention = weak.

An example of tree usage. If the input of the tree is a vector of customer data: (male, 27 years old, no children, low income), then after encoding we get the vector (1

2 0 2). This vector corresponds to the Output of tree 1, i.e. the intention to buy from him is weak.

For another client (female, 28 years old, no children, average income), after coding we have vector (0 2 0 3) and Output of tree 2, i.e. intention to buy - she has strong.

nput vector=(1 2 0 2) = (m 26-30 children no small)

Output = 1 (purchase intent is weak)

Input vector=(0 2 0 3) = (w 26-30 children no average)

Output = 2 (the intention to buy is strong)

he obtained ER make it possible to immediately solve one of the most important tasks of business - who to target their services or goods to, and thus increase the efficiency of their business. The role of the DM in this case is obvious - using the acquired knowledge to improve business efficiency. This is, of course, a conditional example, but all the mentioned stages are also present for the case of real tasks.

Example 2. Forecast of demand for summer goods. One of the most important business tasks is also forecasting the demand for goods or services. In this case, the main purpose of using DM tools is to improve business performance - to outperform competitors, increase the number of customers, etc., by leveraging found knowledge (ER) on a large volume of historical data.

As an example of using DM tools to solve such problems, consider the following task.

An online shop wants to forecast the demand for T-shirts based on historical data. Input data includes product characteristics such as Season, Brand, Material, Country of production, Price. The goal is to predict demand to maximise sales.

Let us consider a specific example of using neural networks as well as decision trees to predict the price of an item.

Let's imagine that we have an online clothing shop and we want to use neural networks as well as decision trees to predict demand for a new line of t-shirts. We have collected data on t-shirt prices over the past few years, including factors such as season, brand, material, country of manufacture, and demand.

Step 1: Data Collection. We first collect historical data on T-shirt sales, including the factors mentioned above.

Step 2: Data preprocessing. We clean the data, remove outliers and missing values, and convert text and categorical variables to numeric format.

Step 3: Training the model.

Step 4: Model Evaluation. We evaluate the accuracy of the model on a test dataset to ensure that it can effectively predict demand.

Step 5: Forecasting. We can now use the model to predict the demand for new t-shirts by inputting relevant data about season, brand and other factors.

Demand = {high, medium, low} - predicted value

Demand Coding= {high=A, medium=B, low=C}

Sample attributes: Season, Brand, Material, Country of manufacture, Price.

Season= {spring, summer}

Season coding= {spring=1, summer=2}

Brand = {Umbro, Puma, Noviti, Nike, New Balance}

Brand Coding={Umbro=1, Puma=2, Novit=3, Nike=4, New Balance=5}

Material={cotton, polyester, elastane, linen}

Material Coding={cotton=1, polyester=2, elastane=3, linen=4}

Country of Manufacture={Bangladesh, Vietnam, Indonesia, Pakistan, China}

Coding country={Bangladesh=1, Vietnam=2, Indonesia=3, Pakistan=4, China=5}

Price={under 1000, 1000 to 2000, 2000 to 3000, above 3000}

Coding price={under 1000=1, 1000 to 2000=2, 2000 to 3000=3, above 3000=4}

An example of part of the original training sample, - is presented below:

Table 8

Demand	**Season**	**Brand**	**Material**	**Country_of_production**	**Price**
high	spring	Umbro	cotton	China	730
high	spring	Umbro	cotton	China	620
medium	spring	Umbro	cotton	China	1790
high	summer	Umbro	cotton	China	599

high	spring	Puma	cotton	China	759
high	spring	Puma	cotton	China	799
high	summer	Puma	cotton	China	709
high	summer	Noviti	cotton	China	299

Below is an example of the same TS, but already encoded according to the above rules.

Table 9

Demand	**Season**	**Brand**	**Material**	**Country_of_production**	**Price**
A	1	1	3	5	1
A	1	1	3	5	1
B	1	1	3	5	2
A	2	1	3	5	1
A	1	2	3	5	1
A	2	2	3	5	1
A	2	3	3	5	1

Stage 1. Visualisation (the service https://www.sciencehunter.net/Services/Clustering#/visualization was used). At stage 1, we visualise the TS (Fig. 4) to see the general nature of the mutual arrangement of classes. From the visualisation it is clear that a simple linear separation of classes is difficult, if not impossible. Most likely, the best solution will be a non-linear separation of classes. In this case, the choice of a neural network solution can be considered reasonable enough.

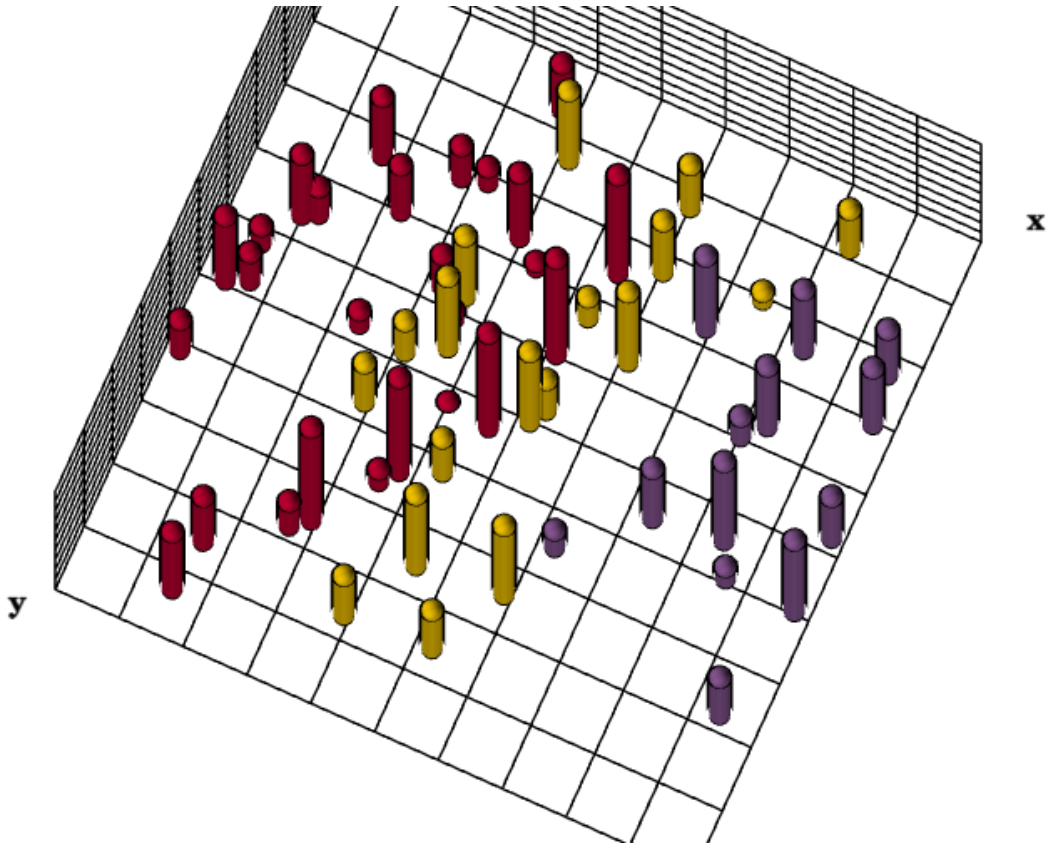


Figure 4. Visualisation of TS

* *Source: suggested by the authors.*

Visualisation of TS using multidimensional scaling service.

It can be seen that class 2 (yellow colour) is mixed with class 1 (red colour) and it is hard to distinguish it.

Stage 2 Neural network solution (service - https://www.sciencehunter.net/Services/Apps/NeuralNetworkClassification was used)

We emphasise that the model, i.e. the neural network, is 'built' using the specified service, not programmed.

The achieved average accuracy (by class) is more than satisfactory 98.7%, as well as the accuracy by class, which can be seen from the results obtained using neuronics (see Fig. and tables, High Demand class - 98%, Medium Demand class - 97.7%, Low Demand class - 100% recognition of EV objects). That is, it can be considered that the classification results are quite satisfactory. Further, using the obtained model (for this purpose it is necessary to save it on the same service beforehand) when supplying new

current data to its input, pre-coded as described in step 2, it is possible to find out the demand for the tested or received goods and to estimate the future profitability (yield) of the goods.

Stage 3. Solving with the help of solving trees (The service https://www.sciencehunter.net/Services/Classification was used). The analysis of the sample using the specified tool is presented below.

The training sample is represented by the same file as in step 2. The result of training is the decision tree presented below in Figure 5.

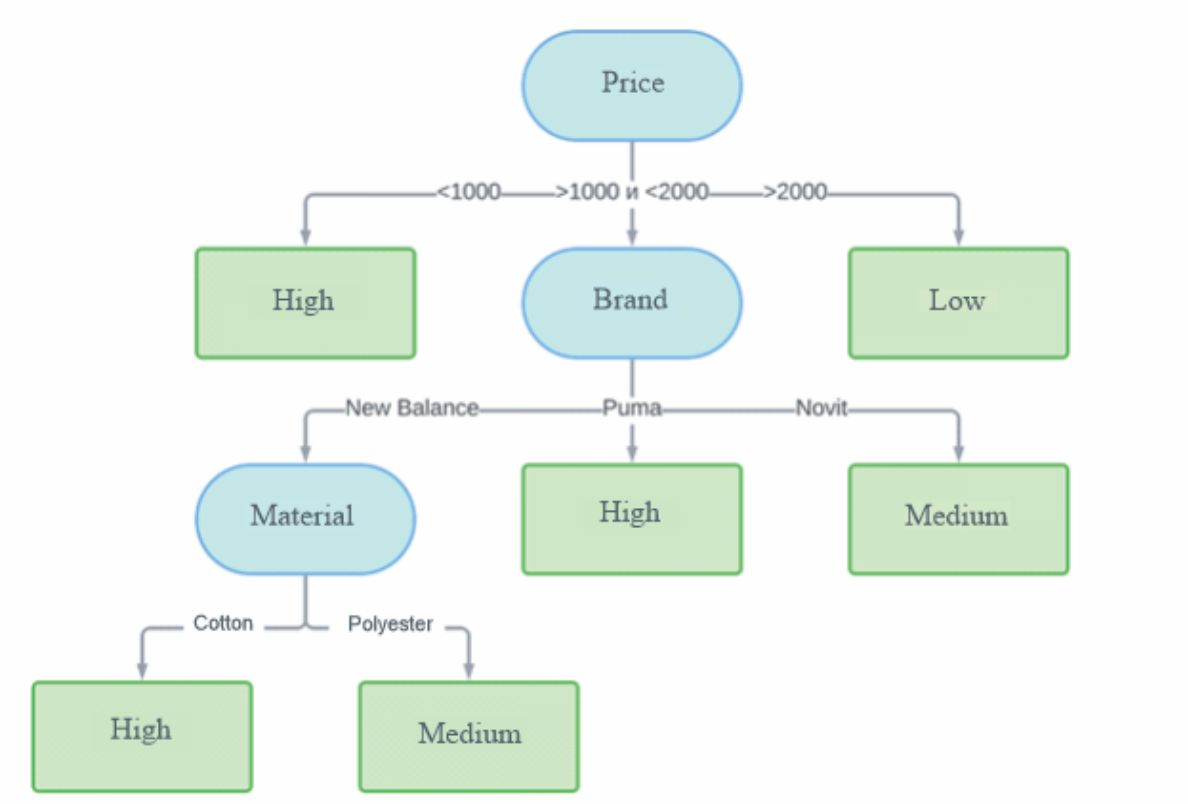


Figure 4. Solving trees

** Source: suggested by the authors.*

In the figure, oval nodes represent features, square nodes represent classes.

The results of testing the tree on the examination sample are as follows:

Average accuracy is 78%, including classes - High demand - 100%, Medium demand - 37%, Low demand - 98%. It is clear that because the tree implements linear separation, which is poorly achievable for this sample (see figure on visualisation), then the results are inferior to those obtained using a neural network. However, unlike

the neural network, in this case it is possible to interpret the results based on the obtained ER (see below).

The following empirical regularities (ER) can be distinguished:

1. If Price < 1000, demand is high

2. If Price > 2000, demand is low

3. If Price is between 1000 to 2000, Brand = New Balance and Material = Cotton, then demand = high, otherwise if Material = Polyester, then demand = medium

4. If Price is in the range of 1000 to 2000, Brand = Puma, then demand = high, otherwise if Brand = Novit, then demand = medium.

Then it is possible to use the obtained model to forecast demand for a particular type of product, i.e. to use the knowledge to improve business efficiency, the main purpose of using DM in this case.

Preliminarily, the product data represented as a set of specific factor values (Season, Brand, Material, Country of production, Price) are coded as described in step 2. The resulting tree is then used to obtain the demand value.

This is done by moving along the tree, selecting the factor that is encountered along the way. At the final vertex one can see the result - the possible demand for the product and, accordingly, it will also be possible to estimate the future profitability (yield) of the product.

Conclusions. The advantages of the neural network include high accuracy of demand prediction. Unfortunately, nothing can be said about the interpretability of the model - this is one of the main disadvantages of the neural network.

The advantage of trees is their interpretability - you can see what combination of factors determines the demand.

The disadvantages include low accuracy relative to the average demand class. Obviously, additional factors need to be involved here to improve accuracy.

Nevertheless, the main goal of the business owner - what decision to make about the relativity of possible demand - has almost been achieved.

Appendix 3. How DM works when solving production problems.

The main problem is data and its collection.

The following Data Mining methods are the most popular in manufacturing:

1. Classification - Used to sort data into predefined categories. Used to predict product defects or classify failure types equipment [11]

2. *Clustering - Groups data based on similar characteristics. Useful for product segmentation or identifying similar patterns in production processes* [12]

3. *Associative rules - Help to identify relationships between different parameters. For example, it is possible to identify which combinations of materials most often lead to defect* [13]

4. *Time Series Analysis - Used to predict future values based on historical data. Used for predicting product demand or planning maintenance services* [14]

5. *Regression Analysis - Helps to determine the relationship between variables. Used to optimise production processes and reduce costs*

These techniques can improve product quality, optimise production processes and reduce costs.

One of the best examples of successful application of Data Mining in the manufacturing sector is the use of predictive maintenance.

Example: Strukton Rail. Strukton Rail, a Dutch railway company, uses predictive maintenance to monitor and predict track switch failures. This allows the company to plan maintenance in advance and prevent failures, which significantly reduces downtime and repair costs[15]

How it works:

1. Data collection: Sensors on the equipment collect real-time data about its condition.

[11] https://www.decosystems.ru/metody-data-mining/

[12] https://habr.com/ru/articles/784060/

[13] https://www.astera.com/ru/type/blog/top-10-data-mining-techniques/

[14] https://habr.com/ru/articles/784060/

[15] https://habr.com/ru/articles/727358/.

2 Data Analysis: Machine learning algorithms are used to analyse the collected data and identify patterns that may indicate future malfunctions.

3. Prediction: Models predict the remaining life of components and possible failures.

4. Actions: Based on the predictions, maintenance is planned to avoid unexpected failures and minimise downtime.

Results:

- Increased accessibility of railway tracks.
- Reduced maintenance costs.
- Improved overall infrastructure reliability.

This example demonstrates how Data Mining and machine learning can significantly improve manufacturing processes and increase a company's efficiency. In manufacturing, online productivity becomes a top priority. What matters here is how much more successful the business is once the model has been deployed, as well as how it handles transactions related to real users in the real world.

Some IO research groups, such as Netflix and Booking.com, have found: improving a model's offline performance is no guarantee that the model will perform better online. Some models perform better offline and worse online.

Appendix 4: Examples of using machine learning in real-life projects

Artificial Intelligence (AI) and Machine Learning (ML) have long gone beyond experimental developments and have become part of real projects in business and industry. Their application opens new horizons for analysing data, automating tasks and improving the efficiency of companies. This article will provide examples of using machine learning in real projects, as well as basic steps for processing big data and customising models. Helpful resources and links to tool libraries will be offered for more in-depth study.

In recent years, Data Science has become one of the most in-demand areas in the IT industry. ML models and data analysis algorithms are used in many industries: from medicine to retail, from the financial sector to industrial production. These technologies help companies and organisations to make informed decisions based on data, improve forecasting and automate many processes.

Examples of using ML/AI in real projects

1. Demand Forecasting in Retail: Examples from Walmart and Amazon

One of the main challenges for retailers is accurately forecasting demand to help optimise inventory, manage supply and reduce costs. Companies such as Walmart and Amazon use machine learning to create sophisticated models that take into account seasonality, consumer behaviour and external factors (e.g. weather conditions).

Example from Walmart:

Walmart uses time-series algorithms and machine learning to predict demand for various goods depending on the time of year and region. This model allows the company to accurately predict how many items need to be delivered to a particular shop, which minimises losses due to surplus or shortage of goods.

Example from Amazon:

Amazon applies ML to predict customer demand and personalise purchases. The model analyses data about customers' previous purchases and suggests products that may be of interest to the user. Thanks to this, the company achieves increased sales and customer satisfaction.

2. Machine learning in healthcare: from diagnosis to prognosis

AI and ML are transforming healthcare by helping to improve the diagnosis and prediction of diseases. One prominent example is projects that use algorithms to analyse medical images to detect cancer at an early stage.

An example of a Google Health project:

Google has developed an ML model for analysing medical images to help detect lung cancer in its early stages. The algorithm is trained on thousands of images and is able to find tumours with an accuracy that exceeds the accuracy of diagnostics performed by doctors.

Disease prognosis:

Companies such as IBM and Microsoft are using ML for predictive analytics in medicine. For example, models trained on patient data help predict the risks of developing diabetes, cardiovascular disease, and other chronic diseases. This allows doctors to take preventive measures and adjust treatment.

As a concrete example of using DM methods for medical diagnostics purposes, let us consider a computer system for early differential diagnostics of liver diseases based on the data of general clinical blood analysis (OAC) and biochemical blood analysis (liver complex), which was developed in the Noosphere company. The web service is located at: https://www.sciencehunter.net/Services/apps/liver. The most common liver diseases were selected for analysis.

To diagnose chronic liver diseases, the most important and most informative are the indices of biochemical blood analysis (liver complex), general clinical blood analysis (OAC) and peculiarities of clinical manifestations of each disease.

Biochemical blood analysis is a method of laboratory diagnostics that allows to evaluate the work of internal organs (liver, kidneys, pancreas, gallbladder, etc.), to obtain information about metabolism (metabolism of lipids, proteins, carbohydrates), to find out the need for trace elements.

The main indicators of biochemical blood analysis (liver complex), which are evaluated: bilirubin level, aspartateaminotransferase (AST), alanineaminotransferase (ALT), alkaline phosphotase, gamma-glutamyltransferase and albumin

General clinical blood analysis (OAC) is a laboratory test, which includes counting all types of blood cells (erythrocytes, leukocytes, platelets), determining their parameters (cell size, etc.), leukocyte formula, measuring haemoglobin level, determining the ratio of cell mass to plasma (haematocrit).

The main OAC parameters that are assessed are: erythrocyte sedimentation rate (ESR), red blood cell count, white blood cell count, platelet count, haemoglobin level.

The computer system was developed on the basis of models derived from machine learning methods (namely, decision trees), in which the values of the above-mentioned indicators were analysed for more than 1500 patients suffering from various liver diseases. The composition of the indicators was determined by leading physicians based on their experience and qualifications. In total, the system measures 10 indicators and concludes the presence of one of 9 types of disease: alcoholic hepatitis, alpha-1 antitrypsin deficiency, autoimmune hepatitis, Konovalov-Wilson disease, viral hepatitis, haemochromatosis, hepatocellular carcinoma, primary biliary cirrhosis, primary sclerosing cholangitis. The system was validated on an independent test sample (more than 500 objects) and showed close to 98% accuracy value of disease classification.

t is quite easy to use the programme. The whole system is set up in such a way that even people unfamiliar with working on the Internet can understand the system as quickly as possible.

Users only need to take a test: enter the results of their analyses in the required fields. You will also need to select the units of measurement. This is due to the fact that some laboratories use different standards. The service offers all possible options. After which you need to click ‘Apply’, and the results will be shown on the screen.

Customers will see numbers showing the decrease, increase in indicators or norm. After that, we advise to determine together with the doctor further treatment or search for other problems that caused the symptoms of liver disease. In any case, we recommend taking the test at least once a year for self-assurance.

Examples of test results

In order to better understand how the service works, we suggest to read an example of its use.

Enter the results of your analyses (e.g. alt, ast, haemoglobin, red blood cell count, etc.) in the free fields and remember to select the relevant units of measurement:

The client then receives a record of the possible disease and which values are above or below normal.

For the example shown in the figure, the values of the indicators processed using the decision tree indicate a high probability of the presence of such a disease as primary biliary cirrhosis. It should be noted that Albumin is below normal, while Alkaline Phosphatase, ALT and AST are above normal. Obviously, the best thing to do next is to undergo further evaluation and get counselling on treatment strategy and other necessary steps in this direction if the disease is confirmed. To facilitate the solution of this task, i.e. to choose the right clinic, doctor, type of examination, you can use a convenient reference system for cities of Ukraine likarni.com, which is also located on this site. For the city of Dnipro you can additionally use the site medinfo.dp.ua.

The obligatory initial part is descriptive analytics. In our case, it is a sample of sufficiently large volume, verified by experienced and qualified doctors, confirmed by the Act of Data Verification. In particular, we obtained a sample of size n=12, m=1509, k=10, where n is the number of features, m is the number of objects (rows), k is the number of diagnoses.

An example of a limited part of such a sample is given in Table 8.

Table 8.

Disease	**ESR (mm /h)**	**Leucocytes (g/л)**	**Platelets (/l)**	**Haemoglobin (г/л)**	**Erythrocytes (/l)**	**Bilirubin (µmol /l)**	**Alkaline phosphatase (µkat/l)**	**Alt (µkat/l)**	**Ast (µkat/l)**	**Albumin (g/l)**	**Iron (µmol /l)**	**Gamma globulin (g/l)**	**Beta globulin (g/l)**
Primary biliary cirrhosis	18	13	270	106	4,3	17	2	0,56	0,57	23	3	19	16
Primary biliary cirrhosis	22	14,4	312	116	4,1	23	5	0,56	0,61	24	2	18	14
Primary biliary cirrhosis	27	12,7	369	116	4,3	14	1	0,59	0,56	38	5	19	12
Primary biliary cirrhosis	26	11,4	202	107	3,9	19	2	0,56	0,52	25	6	18	15
Primary biliary cirrhosis	17	15	222	111	4,4	27	6	0,59	0,57	27	6	18	11

Primary biliary cirrhosis	28	15,7	296	126	4,1	15	1	0,59	0,56	30	6	16	16
Primary biliary cirrhosis	25	12,5	205	116	4,1	18	4	0,54	0,55	27	5	22	14
Primary biliary cirrhosis	31	14,6	313	103	4,3	28	7	0,58	0,55	30	7	22	11
Primary biliary cirrhosis	21	12,7	394	135	4,2	19	4	0,56	0,61	26	5	16	12
Primary biliary cirrhosis	20	15,9	278	121	3,9	24	5	0,6	0,58	28	5	19	18
Autoimmune hepatitis	51	12,7	135	74	4,1	8,2	0	0,62	0,59	29	19	1,59	22,3
Autoimmune hepatitis	33	14,3	118	47	4,3	8,1	0,2	0,57	0,61	29	14	1,61	19,9

Autoimmune hepatitis	43	14,4	116	68	4,3	9,2	0,1	0,58	0,59	31	10	1,6	24
Autoimmune hepatitis	30	14,3	118	69	4	7,2	0,2	0,56	0,59	26	14	1,62	15,6
Autoimmune hepatitis	40	14,1	117	90	4,2	8,7	0,1	0,62	0,6	27	17	1,59	19,9
Autoimmune hepatitis	37	15,8	123	67	4	7,3	0,2	0,58	0,6	30	14	1,59	23,8
Autoimmune hepatitis	28	14,5	124	53	3,8	7,3	0,2	0,6	0,59	28	19	1,6	20,7
Autoimmune hepatitis	34	12,3	120	77	4,4	7,2	0,2	0,61	0,61	25	19	1,6	18,8
Autoimmune hepatitis	30	12,2	118	49	4,2	8,1	0,2	0,55	0,61	28	13	1,59	21,3

… ….

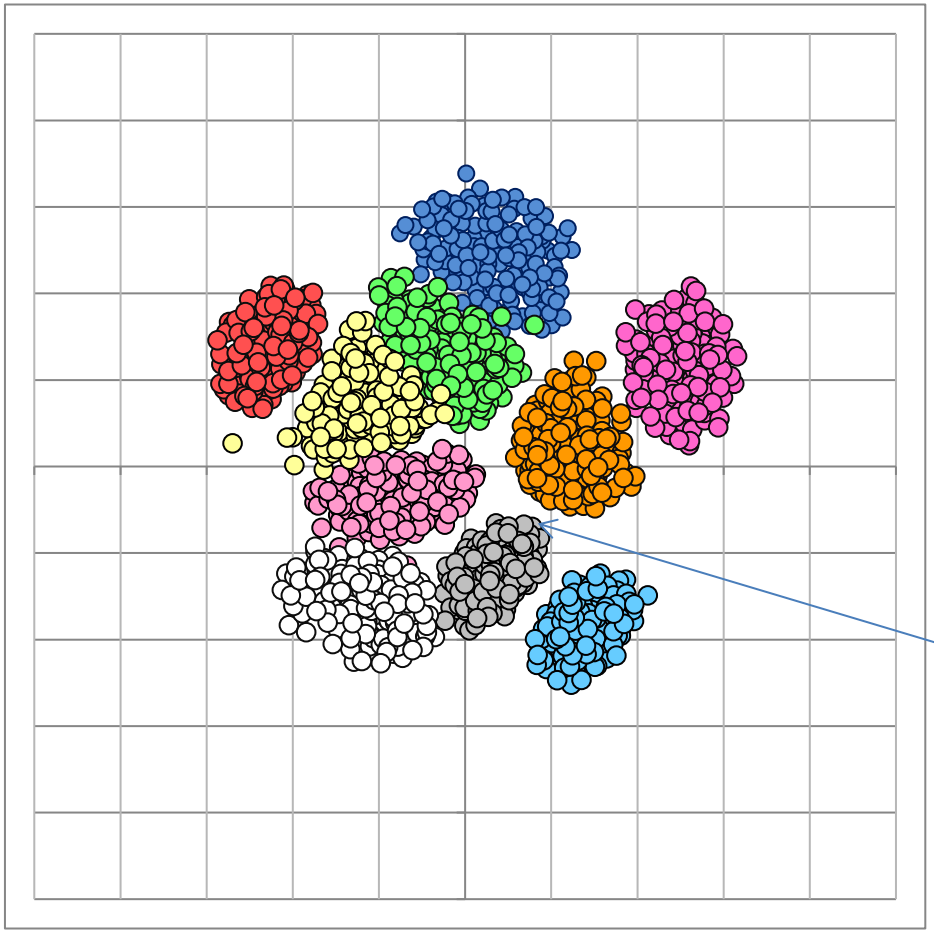
It is known that the liver itself has no nerve endings and does not hurt. This makes it difficult to detect diseases at the initial stages. Therefore, it is very important to detect signs indicating pathological processes in the liver at the early stages. It is common to prescribe general and biochemical blood tests. It is at this stage, the stage of analysing the results of the initial examination is highly desirable to identify clearly enough the direction of the further course of examination of the patient, a preliminary diagnosis, consultation of an experienced colleague. The obtained sample represents such a concentrated experience of the doctor, expressed in the verification and description of the disease.

The next stage - diagnostic analytics - is more related to the application of mathematics, finding recognising functions that identify and separate one type of disease from another as accurately as possible. In fact, this is the stage of machine diagnostics, which uses logical regularities found by the programme, and which reflects the vast experience of doctors.

In order to check the quality of the sample, its structure and whether this structure coincides with the diseases under study, the entire sample was preliminarily checked using visualisation methods, in particular, the multidimensional scaling method was used, in which a multidimensional table is ‘collapsed’ into a two-dimensional or three-dimensional picture. The service used was a portal service located at http://sciencehunter.net/Services/visualization for the two-dimensional case and at http://sciencehunter.net/Services/visualization/viz2 for the three-dimensional case.

The results are summarised in Fig.4. The cluster structure of the sample presented in Fig.4 is remarkable for two reasons:

1. The visualisation shows good separability of the original sample classes, which gives hope of constructing a recognition function with good separability properties.
2. This picture also indirectly indicates the successful choice of the feature space - this in turn indicates the high intuition and experience of the physician who supervised the collection of this sample. Although the reasons for choosing this feature space were different - to use the data from the test results as early as possible.



Diagnosis	№
Alcoholic hepatitis	10
Alpha-1 antitrypsin deficiency	8
Autoimmune hepatitis	2
Konovalov-Wilson disease	6
Viral hepatitis	9
Haemochromatosis	7
Hepatocellular carcinoma	5
Healthy	4
Primary biliary cirrhosis	1
Primary sclerosing cholangitis.	3

Figure 4. Diagnostic map of liver diseases according to the available training sample

* *Source: suggested by the authors.*

Then, in accordance with the initial provisions of Data Mining, the whole sample was divided into two parts: the training sample (TS) and the examination sample (ES). In this case, 2/3 of the observations were included in the LS and 1/3 of the observations were included in the ES.

The recognition function was found from the TS and tested on the ES. The recognition function found was a forest of trees.

The results of testing the forest of trees with ER are shown below, in Tables 9,10:

Table 9. Results of tree forest testing for ER

Number of objects in the EV:	500	500	500	500	500	500	500	500	500	500
Number of untreated:	6	1	3	7	6	6	1	0	3	0 (1)
Number of correct predictions:	463	491	487	472	474	464	465	492	486	496
Prediction accuracy (%):	92,6	98,2	97,4	94,4	94,8	92,8	93	98,4	97,2	99,2
Trees	DT1	DT2	DT3	DT4	DT5	DT6	DT7	DT8	DT9	**All trees**

Table 10

Class:	**Primary biliary cirrhosis**	**Autoimmune hepatitis**	**Primary sclerosing cholangitis.**	**Healthy**	**Hepatocellular carcinoma**	**Konovalov-Wilson disease**	**Haemochromatosis**	**Alpha-1 antitrypsin deficiency**	**Viral hepatitis**
Number of facilities:	50	50	50	50	50	50	50	50	50
Prediction accuracy, in %:	92%	100%	94%	100%	100%	100%	100%	96%	100%

The early diagnosis is now available as a web application and is freely accessible at http://sciencehunter.net/Services/apps/liver. This diagnosis is supplemented with protocols, as well as other attributes necessary in medical practice, such as:

(a) record of complaints; (b) results of objective examination; (c) laboratory examination; (d) instrumental examination; (e) list of symptoms, (f) directories - of diseases, drugs, medical centres.

All of the above is designed on the web application as a separate doctor's office, which can be viewed at the following link: https://www.sciencehunter.net/Medicine#/Patients.

In the future, it is planned to create a separate site for diagnosing not only liver diseases, but also other human organs.

3. ***Computer vision in production: quality control on the line***

In manufacturing plants, AI and ML are widely used to automate product quality control. Computer vision systems based on convolutional neural networks (CNN) can detect product defects on assembly lines in real time.

An example from Tesla:

Tesla is using computer vision to check the quality of car parts at its factories. Cameras connected to the ML system analyse hundreds of images per minute to automatically identify defective parts. This significantly reduces the likelihood of defective products and increases productivity.

How it works:

The algorithms of convolutional neural networks are trained on thousands of images of both standard and defective parts. When the system identifies defective products, it sends a signal to stop the line or sort the goods.

4***. Combating fraud in the financial sector: analysing transactions***

Financial companies such as banks and payment processors are facing a growing threat of fraud. AI and ML help identify suspicious transactions by analysing user behaviour and transaction patterns.

An example from PayPal:

PayPal uses machine learning to analyse billions of transactions every day. The ML model is trained on data from past fraudulent activity, and it can identify anomalies that indicate possible fraud. This allows the company to block suspicious transactions before they are finalised.

A guide to big data processing and model tuning

Step 1: Data collection and preparation

Working with big data starts with collecting and cleaning it. Tools such as Apache Hadoop and Apache Spark are used to work with large amounts of data. In Python, popular libraries for data processing are Pandas and Dask.

Conclusion

Machine learning and big data analytics are becoming the basis for innovation in a wide range of industries. Real-world examples show that AI and ML not only help companies solve complex problems, but also open up new opportunities for growth and optimisation. Implementing these technologies requires a deep understanding of the data, proper tuning of models and continuous work on process improvement.

https://habr.com/ru/articles/854196/

The following are generalised examples of the use of DM methods to solve research problems.

Application of Data Mining in Sociology.

Data Mining (data mining) can significantly help in solving research problems in sociology, such as:

1. Social Network Analysis: Data Mining allows the analysis of large amounts of data from social networks, identifying trends and patterns in user behaviour. This can help social scientists understand how ideas spread and how public opinions are formed.

2. Public Opinion Research: By analysing texts from social media, forums and surveys, it is possible to identify major themes and sentiments in society. This helps in researching public opinion on various issues.

3. Predicting social phenomena: Data Mining can be used to create models that predict social phenomena such as crime rates, unemployment or demographic changes.

Let's look at a specific example of using Data Mining to investigate the impact of social media on political attitudes. In this case, social scientists use Data Mining to analyse data from Twitter, Facebook and other social platforms to study how social media interactions influence political views and user behaviour. For example:

1. Data collection: Collection of tweets, posts and comments related to specific political topics or events.

2. Text Analysis: Using Natural Language Processing (NLP) techniques to analyse the content of posts, identifying key themes and sentiment.

3. Identifying patterns: Analysing interactions between users to understand how political views are formed and disseminated.

4. Influence modelling: Creating models that can predict how certain events or campaigns might affect users' political views.

This approach allows sociologists not only to understand current trends, but also to forecast future changes in public opinion

Another example from the field of sociology. The study of migration flows.

Analysing migration flows using Data Mining can help sociologists to understand the causes and consequences of migration, as well as identify patterns of movement of people. For example:

1. Data collection: Use data from various sources such as government statistical services, social media and mobile applications to collect information on people's movements.

2. Data Analysis: Applying cluster analysis techniques to identify groups of people with similar migration trajectories.

3. Identifying factors: Analysing factors that influence migration, such as economic conditions, the political situation and climate change.

4. Forecasting: Creating models that can predict future migration flows based on identified patterns and factors.

Example. A study of migration in the European Union

Sociologists can use Data Mining to analyse migration flows within the European Union. For example:

1. Data collection: Collect data on people's movements between EU countries from mobile apps and social media.

2. Pattern analysis: Identification of the main migration routes and factors influencing the choice of country to move to.

3. Modelling: Creating models that can predict how changes in the economy or policy may affect migration flows.

This approach allows sociologists not only to understand current migration trends, but also to develop recommendations for migration policy [16]

Example for sociology: Analysis of social behaviour

Research: *Analysing changes in public opinion on political issues.*

[16] https://journalofchinesesociology.springeropen.com/articles/10.1186/s40711-019-0102-4https://link.springer.com/chapter/10.1007/978-3-030-54936-7_3.

- ***Goal**: Identify how the views of different social groups on important political events change.*
- ***Methods**:*
    - *Collecting data from social media (e.g. Twitter or Facebook).*
    - *Applying text and tone analysis algorithms to evaluate opinions.*
    - *Clustering of data to identify groups with similar views.*
- ***Result**: Data Mining allows you to identify the key factors influencing the formation of public opinion and identify 'opinion leaders' - those whose posts receive the most attention.*
- ***Practical relevance**: The results of the research can be used by political consultants or NPOs to plan campaigns and predict public reaction to certain initiatives.*
- ***Algorithms used***

***In sociology**:*

- *Text Analysis: TF-IDF, LDA (Latent Dirichlet Allocation) for Topic Modelling.*
- *Clustering: K-Means or DBSCAN to identify user groups.*

- ***In the economy**:*
    - *Regressions: Linear regression, LASSO regression for factor analysis.*
    - *Time series: ARIMA and Prophet for forecasts.*

These approaches find wide application in scientific research, providing results that would be difficult to obtain with traditional methods of analysis.