UDC 004.6:33
DOI: 10.20535/SRIT.2308-8893.2021.1.05

DATA SCIENCE — DEFINITION AND STRUCTURAL REPRESENTATION

P.P. MASLIANKO, Y.P. SIELSKYI

Abstract. This article is a continuation of the discussion on the existing meanings and formalization of the definition of "Data Science" as an autonomous discipline, field of knowledge, clarification of its defining components, integration, and interaction processes between them. It is noted that most scientific results trace the data-centric nature of the presentation and analysis of this discipline, i.e. the emphasis on the word Data. Analysis of the frequency of use of key terms in the definitions of Data Science shows what our colleagues focus on, which terms of the definitions of Data Science they are based on. In this paper, we make and argue certain additions to Drew Conway's Data Science Venn Diagram, which does not reflect all the resources of the components that define the applied side of Data Science, and, moreover, does not reveal the interaction of these resources not from the point of view of the data researcher, nor in its global understanding. We also propose a unified structural representation of Data Science in the format of an updated Drew Conway's Venn diagram based on a property/attribute that establishes correspondences that provide integration/interoperability between the elements of the sets of Drew Conway's Venn diagram. The new definition of Data Science as an interdisciplinary science and methodology of presenting activities for analysis and extraction of data, information, and knowledge is substantiated.

Keywords: Data Science, Drew Conway's Data Science Venn Diagram, Data Science definition, Data Science structure, data, information, knowledge.

`INTRODUCTION

Starting from the 21st century, the phrase Data Science has begun to attract considerable attention from the world's academic and professional communities. Why the phrase? Despite dozens of savants trying to interpret its meaning in their own way, throughout numerous discussions about its components, this expression has not acquired the meaning of a clearly defined scientific term.

This article aims to carry out research and continue the discussion on the existing definitions and proper formalization of "Data Science" as an autonomous discipline, field of knowledge, clarification of its defining components, their characteristics of integration and interaction processes. Thus, Data Science is an object of analysis, which will be performed through in-depth study and synthesis of existing authoritative scientific results, articles and journals, blogs of well-known authors, and trusted publishers.

We systematized the information from all studied sources in the table for further analysis by the following criteria (columns of the table):

0. Definition of Data Science.
1. Keywords of the definition.
2. Semantics of a definition — list of tools on which it is based (methods, models, algorithms, processes, disciplines, etc.), as well as their interaction.

3. Features of the definition — its purpose (theoretical, practical, specialized, etc.), scope.

4. Discussion arguments and the uncertainties regarding the definition and understanding of Data Science, given in the source.

5. In total, 11 most common sources were analyzed and cited, which Data Science related key points are briefly described below.

RELATED WORK

The vast majority of scientific works on Data Science begin either with the famous expression "*Data Scientist: The Sexiest Job of the 21st Century*" of Thomas Davenport and D.J. Patil [1], or with a reference to Drew Conway's *The Data Science Venn Diagram* [2] (Fig. 1), to which we shall return. In some cases, you can even find links to both resources at once.

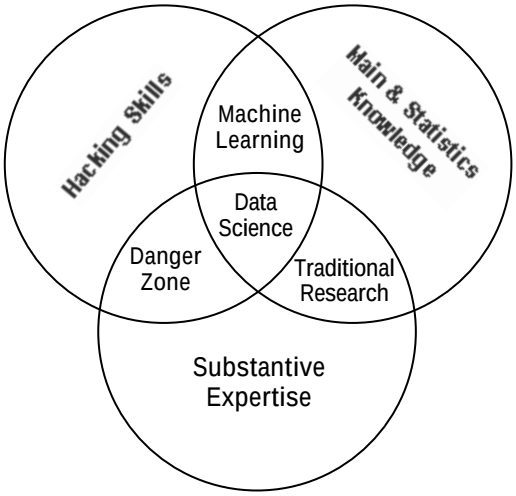

Fig. 1. The Data Science Venn Diagram [2]

It is worth noting that of all the works, dedicated to Data Science, this diagram is perhaps the only attempt at an in-depth presentation of the structure and content of the Data Science model, which nevertheless leaves a dry residue of uncertainty around some areas of the figure. What is a danger zone? What meaning does the author put into traditional research? Why did he choose machine learning as the intersection of mathematics and statistics with hacking skills? And, finally, can the last term be perceived as a scientifically justified, reasonable, and meaningful concept with an unambiguous interpretation?

Indeed, how can a job that requires hacking skills not be considered as sexy? Engineers and scientists in any definition want to see and understand a certain reasoned meaning, with a solid scientific basis, especially if the very described notion contains the word "science" itself. Otherwise, such loud statements will only provoke excitement and fruitless controversy over the newly introduced term, which, in fact, happened.

Due to the ambiguous emergence of Data Science, the debates over the interpretation of this name immediately began among the academic and professional communities. In particular, the question arises about the similarity between

Data Science and classical, well-known statistics. For example, Cathy O'Neil and Rachel Shutt in their book *Doing Data Science* [3] (where, by the way, the entire 15 pages are devoted to the first part "*Introduction: What is Data Science?*"), refer to numerous protests by experts in the field of statistics against the uniqueness of Data Science, calling it a new-fashioned rebranding of their alma mater. The authors themselves claim their differences, emphasizing the specific processes created by the pioneers of Data Science, allowing to work with more data — the *processes of Data Science* [3]. In general, this resource covers more the professional aspect of Data Science, explaining it from the point of view of data scientists as specialists in this field, and the skills that such positions require.

Vasant Dhar, whose testimony can be found in *Communications of the ACM*'s *Data Science and Prediction* article [4], also joined the defense in the case of Data Science vs. Statistics. The author focused on a whole list of differences [4]:

1. First of all, data — the main fuel of Data Science — is quickly becoming unstructured, diverse. Therefore, the analysis of "raw" data, as well as combining data of different types (*feature engineering*), demands additional interpretation and understanding, based on the foundations of multiple other disciplines (linguistics, sociology, etc.) [4].

2. Nowadays, most of the data is produced by computers to be consumed by other computers [4]. In these realities, it is computers that make decisions that encourage their operators — data scientists — to retrain: to play as well the role of risk managers, to act as a guarantor-supervisor of developed system quality instead of the more classic duty of an expert in the context of statistics.

3. Machine learning, applied for creating unfailing predictive models, is an essential Data Science component, which is more and more concerned with forecasting various values, events, phenomena [4].

"Data Science, … , is perhaps the best label we have for the cross-disciplinary set of skills that are becoming increasingly important in many applications across industry and academia." — this definition is given by Jake Vanderplas in the Python Data Science Handbook [5] (also with reference to Fig. 1), where he often uses the concept of "skills", which, again, emphasizes a more professional application.

"Multifaceted discipline" — say the authors of the book Data Science for the Layman: No Math Added [6] Annalyn Ng and Kenneth Soo, focusing on machine learning as a key component and citing a standard algorithm of carrying out research in the field of Data Science [6]:

1. Data processing and preparation for analysis.
2. Selection of potentially effective machine learning algorithms.
3. Optimization of (hyper-) parameters of algorithms: training, validation.
4. Construction of integral models (combination of certain algorithms or their separate usage) with their further comparison and selection of the best.

In addition to applied specifics, there are definitions of a high level of abstraction, more clear and intuitively perceived by the human mind. Well-known experts in the field of Data Science, Foster Provost and Tom Fawcett formulate the key activities of data scientists: extracting useful information and knowledge from data [7]. Hence, Data Science is also compared to Data Mining: "At a high level, Data Science is a set of fundamental principles that guide the extraction of

knowledge from data. Data mining is the extraction of knowledge from data, via technologies that incorporate these principles" [7].

In parallel, the authors focus on the analysis of the structure of Data Science in the context of effective solutions to real business problems [8]. Here, one of the principal processes — data-driven decision making and its progressive automation.

Often, and certainly not without reason, Data Science is closely linked to data analysis. For example, Matthew Waller and Stanley Fawcett describe Data Science quite abstractly: "Generally, Data Science is the application of quantitative and qualitative methods to solve relevant problems and predict outcomes." [9], but instead derive their own model of influencing data scientist's performance by two interdependent components: domain knowledge and analytical skills.

A Ukrainian specialist, Bohdan Pavlyshenko, agrees with Waller and Fawcett, focusing on data analysis, the need for a proper understanding of the nature of data, and the specifics of a particular domain in problem-solving [10].

We are currently coming to a certain consensus on the applied essence of Data Science as a business tool, a profession. Most of the above resources trace the data-centric nature of the representation and analysis of the discipline, i.e., the emphasis on the word Data. And what about Science? What about the academic side of the coin?

Jeff Leek answers these questions, listing a number of arguments in defense of science and the complexity of solving scientific problems [11], citing, in particular, a quote of John Tukey, a pioneer in data analysis: "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." [12]. The author also accentuates the main reason for the outbreak of excitement around Data Science — the focus on data, proclaiming: "The long term impact of data science will be measured by the scientific questions we can answer with the data." [11].

Based on the aforementioned arguments, Fig. 2 shows the results of frequency analysis of the most commonly used key terms present in the various definitions of Data Science. This analysis is an example of one of the operations of semantic decomposition, carried out on the basis of the constructed table, the criteria of which are described in the introduction.

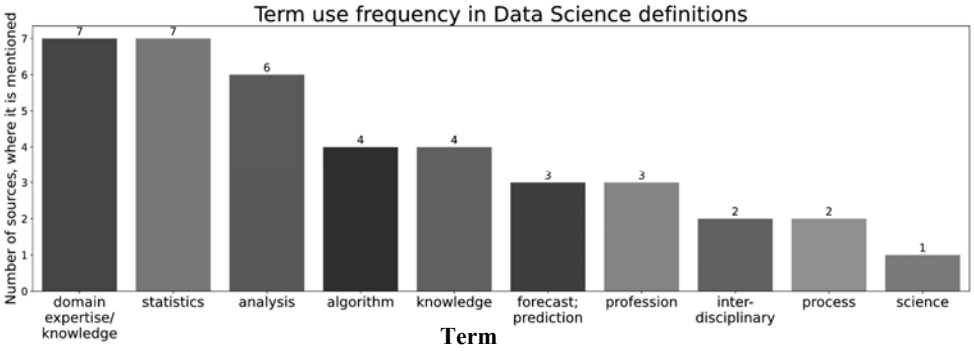

Fig. 2. Histogram of frequency analysis of terms used in the considered definitions of Data Science

These indicators help to better understand what our colleagues are focusing on, what terms the definitions of Data Science are based on: for example, the

most commonly used terms are "*domain expertise/knowledge*" and "*statistics*", which do not fully reflect the components of the object of our research. Instead, the word "*science*" is mentioned only once, confirming the mostly data-centric nature of existing definitions.

Such a simple but quite clear way of comparative analysis of the term use frequency approximately reflects the overall vision of examined authors on the structure and content of Data Science.

Thus, on the basis of the results of even these brief studies of the publications of authoritative experts, an ambiguous, incomplete picture of the defining elements of Data Science is formed. Moreover, the obvious problem of a lack of compromise and a clear link between Science and Data is highlighted.

SYNTHESIS

Based on the preliminary conclusioni, on the above scientific results, we will make some clarifications of the interaction of entities and formalization of Data Science.

Studies of the Data Science representation, analysis of the results of the selection and justification of its attributes, provide grounds for making adjustments to the definition and structure of Data Science.

These rectifications imply some additions to the repeatedly mentioned Drew Conway's Data Science Venn diagram (Fig. 1) [2], which does not reflect all the resources of the components that form the applied side of Data Science, and, moreover, does not reveal the interaction of these resources from the point of view of the data scientist, nor in its global sense.

In this article, we propose an updated, refined version of Drew Conway's Venn Data Science diagram and try to explain and justify not only the essence of its components but also the principles of their integration and interoperability (Fig. 3).

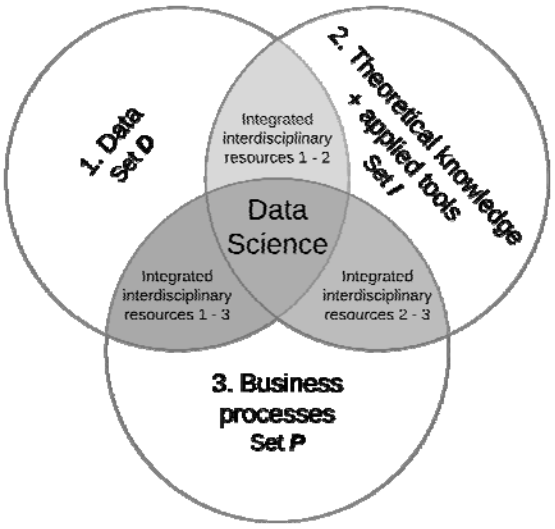

Fig. 3. Structural representation of Data Science in the format of an updated Venn Data Science diagram

FORMAL REPRESENTATION OF ELEMENTS OF DATA SCIENCE RESOURCE SETS

The set of formalized data resources — D (data), d is an element of set D, $d \in D$.

The set of theoretical and applied data processing tools — I (*instrument*), i is an element of set I, $i \in I$.

The set of business processes for data, information, and knowledge acquisition by means of theoretical and applied data processing tools — P (*process*), p is an element of set P, $p \in P$.

Sets of integrated interdisciplinary resources 1–2, 1–3, and 2–3.

Let's explain each entity separately:

- data — raw materials, a key research resource that determines the features and models of the particular domains;
- theoretical knowledge and applied tools — instruments, through which the process of extracting information and knowledge from data takes place. These include both exact sciences (mathematics, statistics, computer science, machine learning, data analysis, etc.) and software applications (programming languages, their libraries, frameworks, development environments, visualization tools, etc.);
- business processes: here we are talking about the organization of research at the meta-level, setting goals and objectives, determining the main stages of work, their sequence, the nature and features of the evaluation of results, etc. Essentially, all the above actions will vary from one problem to another. Why "*business*"? To emphasize the need to optimize this component in order to maximize the benefits of the carried out researches/developed systems;
- integrated interdisciplinary resources 1–2, 1–3, and 2–3: the essence of the corresponding intersections is not fully compatible with the classical definition of the intersection operation in Set Theory, given the obvious fact that the nature of the elements of different sets is different. Therefore, we are talking about the existence of functional relationships between different types of resources: for example, theoretically, zone 1–2 includes existing data, known to mankind, that can be processed using existing theoretical knowledge and applied tools, but such interaction is not a subject to any existing business process.

DECOMPOSITION OF ELEMENTS OF DATA SCIENCE RESOURCE SETS

1. Data (set D). Plays the role of a kind of fuel for Data Science instruments and processes. Obviously, there is a lot of data in the world and every second more is created. So let's try to bring order to the ocean of this natural chaos, defining the main elementary component of the entirety of data — a formalized data resource d. Thus, zone 1, shown in Fig. 3, consists of formalized data resources — information, collected, stored in a certain form, which can be classified, for example, by the following criteria:

- by type of storage: distinguish digital types of data storage (hard drives, SSDs, USB-drives, etc.) and in contrast to them — more classic — rock, wall inscriptions, carvings, books, magazines, newspapers, etc. We do not forget the

immaterial data — unrecorded real-time speech, thoughts, movement of any objects;

- by structure, data can be divided into structured (mainly numbers and numerical arrays, standard types of programming languages) and unstructured: audio, images, video, text;
- by availability: confidential and public data.
- In any case, in the context of area 1 we are talking about the set D of formalized data resources d :

$$\exists\, d,\ d \in D. D = \{d\}\ .$$

Some examples of data as formalized resources are listed below:

- numeric data arrays: the simplest (at least for a computer) formalized representation of information. Arrays can be of different shapes, sizes and contain any number of elements. Note also that the numbers (scalars) themselves can be seen as formalized data resources;
- digital images: in their structure — the same, in some sort of way, ordered numerical arrays (pixel values), but at the level of human perception of information (visualization), play a special, more significant role, so they can also be considered as formalized resources;
- audio files are another type of information that humans perceive by ear. Physically, an audio file is a specific set of frequencies — numbers that follow a strict order. Therefore, it is about data in the form of sequences, series, which are reproduced in time;
- video files — a more complex case, which includes not only a set of images but also audio. That is, we define the presence of two different types of sequences, as well as the mechanism of their synchronization as integral elements of correct video playback;

As a part of Data Science, all of the above and many other types of data are summarized in one, more extensive formalized data resource — a so-called *dataset*.

In *Data Science for Business*, such terms as database *table*, *worksheet* (for example, a sheet of an .xsl file), and *dataset* are equated to each other; a more specific decomposition of the latter is presented [7]:

- data sets consist of so-called *examples* (samples) or *instances* (table *rows*) [7];
- each instance, in turn, is comprised of a fixed (in the classical representation) number of *features* (*columns* of the table), the values of which uniquely identify instances [7].

2. Theoretical knowledge and applied tools (set I). Any instruments, models, algorithms, human skills, formalized or materially implemented, aimed at carrying out certain operations on data for their better understanding. For the purpose of formalization we will define an elementary component of this set as an instrument i :

$$\exists\, i,\ i \in I.\ I = \{i\}\ .$$

Such elements can acquire different levels of abstraction and different scales: for instance, individual clusters of knowledge can, in turn, be combined

into whole areas of knowledge. Here are the most interesting theoretical aspects for Data Science:

- Statistics: let's start with it to pay tribute to fellow statisticians. Undoubtedly, this is a vast science that includes many *i*'s, but only a certain part of them are used in Data Science, in particular, elements of descriptive statistics at the stages of exploratory data analysis (*EDA*) [13]: mode, median, mean, standard deviation. *EDA* is also a part of *Data Analysis*.

- Data analysis: in order not to invent anything superfluous, we go back to the definition of the most reliable source — John Tukey [14], who provides a comprehensive list of components of the discipline: data analysis procedures, methods of interpreting their results, simplification and improvement of data analysis on the earliest stages of data collection, as well as all the techniques of statistics, that are applicable to data analysis [14]. That is, Data Analysis is closely related to the Statistics domain.

- Artificial Intelligence (AI): since the inception of this term, the constant debate around its essence has never subsided. AI should be considered as a separate section of computer science, designed to program machines for human behavior, thinking, and independent decision-making [15]. In the case of Data Science applications, decision-making is often based on predicting certain results.

- Machine learning: this term originates from Arthur Samuel's article "*Some studies in machine learning using the game of checkers"* [16], where the author uses this phrase literally — the process of learning machines — programming computers for behavior as such that includes the learning process if it were inherent in humans or animals [16]. More specifically, it is about automated optimization of computer performance, based on experience.

- Deep learning [17]: the problems of applying machine learning techniques on unstructured data: texts, music, images, etc. are becoming more and more popular. Informative (for computers) representation of such raw data requires their automatic interpretation through step-by-step processing of numerical input arrays throughout many stages of data projection onto spaces of the higher levels of abstraction. Such a procedure is a key aspect of Deep Learning [17], i.e. learning the layers (stages) of neural networks on the data via the generalized learning process instead of explicitly developing the necessary projections by hand.

- Big Data Analytics: this term should be taken literally — the field of knowledge about the application of advanced analytical methods on the big amounts of data, according to Philip Russom [18]. The presence of big data is manifested not only by their volumes but also by such characteristics as data variety and velocity (*3 Vs*).

- Data Mining: recall the definition of Provost and Fawcett that Data Mining — the extraction of knowledge from data using technologies that embody the principles of Data Science [7]. This example allows us to trace the direct connection of Data Mining with Data Science as an integral part of it.

- Data visualization — techniques for presenting data of different nature and dimensions in the most understandable and human-readable form — graphic [19]. In this set of tools, in addition to countless frameworks and software that implement the full range of possible charts and graphs (in Cartesian, polar coordi-

nates, scatter, line plots, histograms, bar and pie charts, 3D images, etc.), more complex machine learning methods of dimensionality reduction can be highlighted here as well. A good example is the Principal Component Analysis (PCA).

Let's also notice the application tools — instruments that allow implementing algorithms and methods of the above theoretical knowledge in the form of (open-source or private) software applications, platforms, frameworks, libraries, systems, and so on. These include:

- programming languages widely used in the domain of Data Science: here we can consider both: such programming languages as *R* [20], *Python* [5], which are used directly for the development of Data Science systems, for the implementation of the higher-level interfaces and components of such systems; as well as *C*-family programming languages, used for the development of lower-level APIs (Application Programming Interfaces) in order to optimize and parallelize basic computations. As an example of such a hierarchy — *TensorFlow* [21] machine learning system from *Google Brain*;
- whole systems of computer mathematics and algebra (*MATLAB* [22], *MathCad* [23]), statistics (*STATISTICA* [24]); machine and deep learning systems, big data systems and environments, which are distinctive by the presence of interfaces for different programming languages (*TensorFlow* [21], *Torch* [25], *Spark* [26]) or even by embedded graphical user interfaces (eg *Orange* and *KNIME* [27]);
- separate add-ons of the aforementioned systems of the highest level of abstraction (*Keras* [28] for *TensorFlow*); specialized programming language libraries, modules, packages that provide ready-made software solutions for machine learning (*Scikit-Learn* [29]), data processing (e.g. *NLTK* [30] for text data), their visualization and interactive calculations (*matplotlib* [31, 32], *pygal*, *Plotly*, *Pyvot* [31], *pandas*, *seaborn* [32], etc.), and many others.

This list is not exhaustive and can be extended with many other theoretical and applied instruments.

3. Business processes (set P). The set of business processes for data, information, and knowledge acquisition by means of theoretical and applied data processing tools.

In order to formally represent the interaction of the two previous sets D and I, a third set P is introduced. To optimally extract knowledge and new information from the formalized data resource d using the instrument i, we subordinate the whole entirety of work that needs to be done to a certain process p:

$$\exists p,\ p \in P.\ P = \{p\}.$$

The relationship between d, i, and p will be demonstrated in more detail further, in the context of the sets of integrated interdisciplinary resources. So far, a basic example of the Data Science process is shown in Fig. 4, suggested by Cathy O'Neil and Rachel Shutt [3].

In this representation, 8 main stages of the process are identified:

1. Collecting raw data from any real-world resources.
2. Data Processing.
3. Their cleaning.
4. Exploratory Data Analysis (EDA).

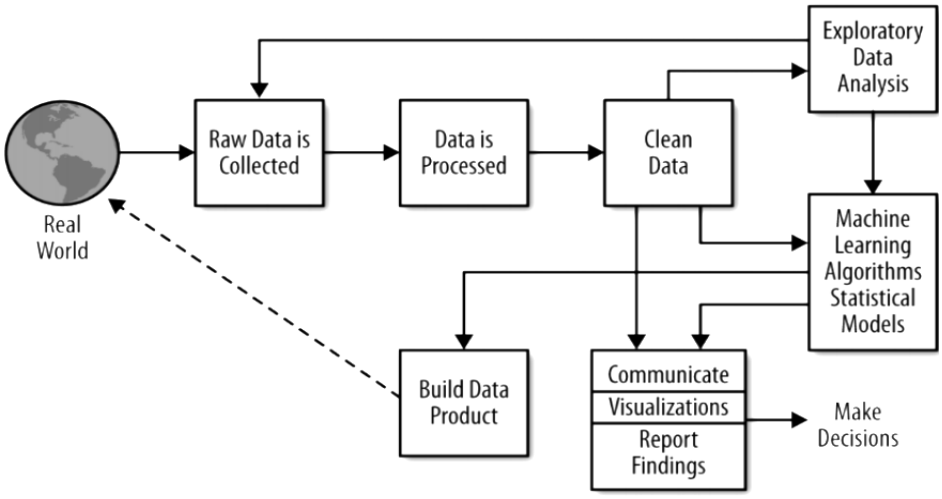

Fig. 4. Data Science Process [3]

5. Construction of statistical models and training of Machine Learning algorithms using collected, processed, cleaned data.

6. Stage of internal communication. Presentation of results to members of the development team of a specific Data Science system, as well as its stakeholders.

7. Creation/production of new data used in the real world.

8. Decision-making based on the obtained results.

It is worth noting that there are direct links between certain stages: for example, EDA may reveal a lack of data that needs to be collected, or cleaned data can be visualized to explain its nature to colleagues.

We compare the above-described Data Science process with the *Cross Industry Standard Process for Data Mining* (CRISP-DM), analyzed by Foster Provost and Tom Fawcett [7] and presented in Fig. 5.

Let us pay attention to its circular iterative-incremental nature, as well as the presence of two rather abstract, generalizing stages — *Business Understanding* and *Data Understanding* [7]:

- the first embodies the need for a clear problem statement in accordance with the given task, the search for creative methods to achieve the goal, its optimal formalization, which would allow the application of already existing methods as effectively as possible;
- while the second stage focuses on the strategic analysis of the main raw material Data Mining — data. Here it is essential to understand the basic structure, pros, and cons of the involved data. The proper assessment of the potential sources of additional information, the necessary investment (both time and financial) in their research and use, is also important.

Whereafter is an integral *Data Preparation* procedure, which, by analogy with the process in Fig. 4, combines data processing, cleaning [7], and EDA, which in themselves can be a multi-iterative subprocess.

The next stage of CRISP-DM — *Modeling* — also has a direct correspondence in the presentation of the Data Science process, where, again, more specific names are given.

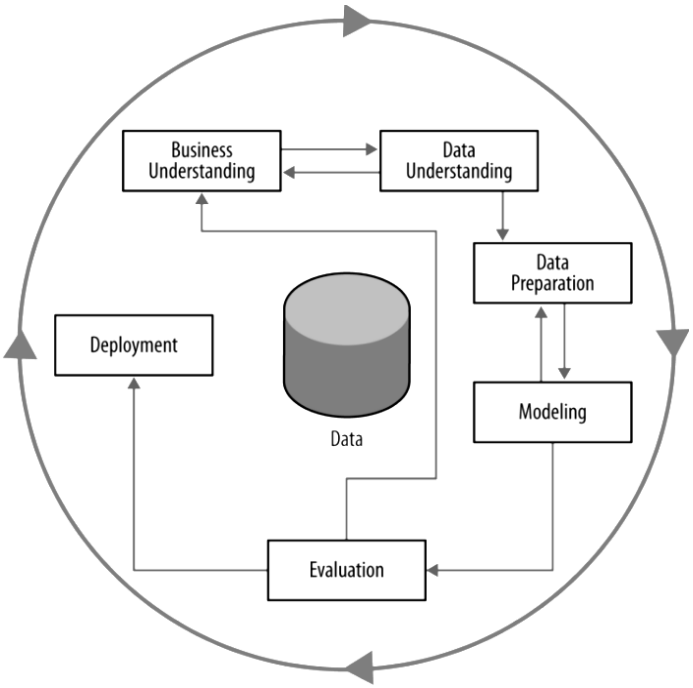

Fig. 5. CRISP-DM [7]

Any development should be subject to quality control through regular verification and validation of the built models and the entire system. The authors also implicitly emphasize the need for communication, presentation of results, as well as the concepts of their simple and clear explanation to key stakeholders (investors) [7], who are responsible for making major business decisions.

After all, in the context of business, the decisive factor is the successful *Deployment* of each system approved by management [7]. In essence, the solution to business problems is directly related to obtaining a certain material benefit.

4. Integrated interdisciplinary resources 1–2, 1–3, and 2–3.

Here and later in this paper, integrated interdisciplinary resources 1–2, 1–3, and 2–3 are subsets formed on the basis of the presence of integration/interoperability properties between elements of sets D, I, and P. Integration/interoperability of elements of sets D, I, and P is the ability to process a certain data resource by means of a certain subset of instruments, following certain processes.

Such a decomposition and generalized systematic representation of the elements of resource sets of Data Science shows and justifies both the complexity and the need for comprehensive research and ongoing discussion on existing definitions and formalizations of Data Science as an autonomous discipline and field of knowledge, clarification of its defining components and characteristics, their integration and interaction processes.

DATA SCIENCE AS A SET OF INTERDISCIPLINARY RESOURCES OF SETS *D*, *I*, AND *P*

To generalize the formal representation of interdisciplinary sets, we define, for example, an arbitrary entity A as a finite resource set A, where $a \in A$ — ele-

ments of the set A, and an arbitrary entity B as a finite resource set B, where $b \in B$ — elements of the set B.

Let us establish the rules of correspondence [33] C_{ab} and C_{ba} between the elements of the resource set A and the elements of the resource set B.

That is, if $\exists a \in A, \exists b \in B, (a,b) \in C_{ab}$, then we say that the element b of the set B corresponds to the element a of the set A, given correspondence C_{ab}.

And if $\exists b \in B, \exists a \in A, (b,a) \in C_{ba}$, then we say that the element a of the set A corresponds to the element b of the set B, given correspondence C_{ba}.

Hereinafter, the term "C_{xy} correspondence" must be understood as the fundamental concept of set theory, which establishes, explains, and formalizes the relationships between the elements of the sets $X = \{x\}$ and $Y = \{y\}$ [33].

Next, we consider the presence of "correspondence" as a common property/feature of pairs of elements of sets A and B.

Let's now construct some finite set of interdisciplinary elements M from all pairs of elements of the resource set A and elements of the resource set B, which have common properties/features and established on the basis of these features correspondences C_{ab} and C_{ba}.

Definition 1. Set M, $\forall m \in M$, $m = \{a,b \mid a \in A, b \in B, (a,b) \in C_{ab}\} | m = \{b,a \mid b \in B, a \in A, (b,a) \in C_{ba}\}$, of interdisciplinary pairs of of arbitrary finite sets A and B — a set of pairs of elements formed by elements elements of set A, and elements of set B, having a common property/feature that establishes the correspondences C_{ab} and C_{ba} between these elements.

We substantiate Definition 1 as a simple way to form a set M by combining pairs (a,b) and (b,a) $\forall a \in A, \forall b \in B$, selected by a common property/feature that establishes the correspondence C_{ab} and C_{ba} between these elements. To do so, we define and apply a common property/feature that establishes the correspondences C_{ab} and C_{ba} between the elements of sets A and B. Since the types of properties/features that determine the correspondence C_{xy} can be set quite a lot, in this case, we introduce restrictions and specify the correspondence C_{xy} between elements of sets A and B.

In particular, we define such a necessary property/feature for us that provides integration/interoperability between the elements $a \in A, b \in B$.

According to this property/feature, a certain subset of pairs of elements of sets A and B can be distinguished, which has established correspondences C_{ab} and C_{ba}.

Let the set of pairs of elements $\{a,b \mid \forall a \in A, \forall b \in B, (a,b) \in C_{ab}\}$ and the set of pairs of elements $\{b,a \mid \forall b \in B, \forall a \in A, (b,a) \in C_{ba}\}$ be defined as such, that have the property/feature, which establishes correspondences C_{ab} and C_{ba} and provides integration/interoperability between elements a and b.

Then the set M is defined as the union of pairs of elements of the set $\{a,b \mid \forall a \in A, \forall b \in B, (a,b) \in C_{ab}\}$ and the set $\{b,a \mid \forall b \in B, \forall a \in A, (b,a) \in C_{ba}\}$,

selected as having a common property/feature that determines the correspondences C_{ab} and C_{ba}. More formally:

$$M = \{a,b \mid \forall a \in A,\ \forall b \in B,\ (a,b) \in C_{ab}\} \cup$$

$$\cup \{b,a \mid \forall b \in B,\ \forall a \in A,\ (b,a) \in C_{ba}\}.$$

Similarly, it is possible to form a set of interdisciplinary pairs of elements of any number of sequentially combined arbitrary finite sets $A, B, D, \ldots$ on the basis of a common property/feature defined for them, which provides integration/interoperability between pairs of elements $a \in A,\ b \in B,\ d \in D, \ldots$ and establishes the correspondences of C_{xy} between the elements of adjacent sets.

Definition 2. The set of interdisciplinary pairs of elements of an arbitrary number of finite sets $A, B, D, \ldots, X, L$ is the set of elements of successive pairs $(a,b),(b,d), \ldots, (x,l),\ \ \forall a \in A,\ \forall b \in B,\ \forall d \in D, \ldots, \forall x \in X,\ \forall l \in L$ such that they have a common property/feature that provides integration/interoperability between the elements $a,b,d,\ldots,x,l$ and establishes C_{xy} correspondences between the elements of adjacent sets:

$$M = [\{a,b \mid \forall a \in A,\ \forall b \in B,\ (a,b) \in C_{ab} \cup\}$$

$$\cup \{b,a \mid \forall b \in B,\ \forall a \in A,\ (b,a) \in C_{ba}\}] \cup$$

$$\cup [\{b,d \mid \forall b \in B,\ \forall d \in D,\ (b,d) \in C_{bd}\} \cup$$

$$\cup \{d,b \mid \forall d \in D,\ \forall b \in B,\ (d,b) \in C_{db}\}] \cup \ldots$$

$$\ldots \cup [\{x,l \mid \forall x \in X,\ \forall l \in L,\ (x,l) \in C_{xl}\} \cup$$

$$\cup \{l,x \mid \forall l \in L,\ \forall x \in X,\ (l,x) \in C_{lx}\}].$$

The practical application aims to solve the problem of forming common pairs, triples, quadruples, etc. of elements of any number of arbitrary finite sets $A, B, D, \ldots, X, L$ on the basis of their defined common property/feature, which provides integration/interoperability between the elements $a \in A, b \in B,$ $d \in D, \ldots, x \in X,\ \ l \in L$ and establishes the correspondences $C_{abd\ldots xla}$ and $C_{alx\ldots dba}$ between the elements of these sets.

Definition 3. The set M, such that

$$\forall m \in M,\ \ m = \{a,b,d,\ldots,x,l,a \mid a \in A,\ b \in B,\ d \in D, \ldots, x \in X,\ l \in L,$$

$$(a,b,d,\ldots,x,l,a) \in C_{abd\ldots xla}\} \mid m = \{a,l,x,\ldots,d,b,a \mid a \in A, \forall l \in L,$$

$$x \in X, \ldots, d \in D,\ b \in B,\ (a,l,x,\ldots,d,b,a) \in C_{alx\ldots dba}\}$$

of interdisciplinary pairs, triples, quadruples, etc. of elements — a set of pairs, triples, quadruples, etc. of elements that can be formed by elements of arbitrary finite sets $A,B,D,\ldots,X,L$ having a common property/feature that determines the correspondences $C_{abd\ldots xla}$ and $C_{alx\ldots dba}$ between these elements. Hence:

$$M = \{a,b,d,\ldots,x.l,a \mid \forall a \in A,\ \forall b \in B,\ \forall d \in D,\ldots,\ \forall x \in X,\ \forall l \in L,$$

$$(a,b,d,\ldots,x,l,a) \in C_{abd\ldots xla}\} \cup \{a,l,x,\ldots,d,b,a \mid \forall a \in A,\ \forall l \in L,$$

$$\forall x \in X,\ldots,\ \forall d \in D,\ \forall b \in B,\ (a,l,x,\ldots,d,b,a) \in C_{alx\ldots dba}\}. \quad (1)$$

In equation (1), subset $\{a,b,d,\ldots,x,l,a \mid \forall a \in A,\ \forall b \in B,\ \forall d \in D,\ldots,\forall x \in X,\ \forall l \in L,\ (a,b,d,\ldots,x,l,a) \in C_{abd\ldots xla}\} \subseteq A \times B \times D \times \ldots \times X \times L \times A$ given correspondence $C_{abd\ldots xla}$, and subset $\{a,l,x,\ldots,d,b,a \mid \forall a \in A, \forall l \in L, \forall x \in X, \ldots, \forall d \in D,\ \forall b \in B,\ (a,l,x,\ldots,d,b,a) \in C_{alx\ldots dba}\} \subseteq A \times L \times X \times \ldots \times D \times B \times A$ given correspondence $C_{alx\ldots dba}$.

Thus, formula (1) will be rewritten:

$$M = C_{abd\ldots xla} \cup C_{alx\ldots dba} = \{\text{tuple}(A,B,D,\ldots,X,L,A,G_{abd\ldots xla})$$

$$\cup\ \text{tuple}(A,L,X,\ldots,D,B,A,G_{alx\ldots dba})\}, \quad (2)$$

where $G_{abd\ldots xla}$ and $G_{alx\ldots dba}$ — graphs/diagrams/matrices of correspondences $C_{abd\ldots xla}$ and $C_{alx\ldots dba}$ respectively.

Then, for three sets — components of Data Science: formalized data resources D; theoretical and applied data processing tools I; business processes of data, information and knowledge extraction by means of theoretical and applied data processing instruments P, we formalize the definition of Data Science on the basis of the updated Venn diagram (Fig. 3).

Definition 4. The definition "Data Science — interdisciplinary science and methodology of representing activities for analysis and extraction of data, information, and knowledge" can be formalized as a set of triples of elements of interdisciplinary resources from three resource sets: Data D, Instruments I, and Processes P, such that having a common property/feature that provides integration/interoperability between the elements $d \in D$, $i \in I$, $p \in P$, and establishes the correspondences C_{dipd} and C_{dpid} between the elements of these sets. That is:

$$DS = \{d,i,p,d \mid \forall d \in D,\ \forall i \in I,\ \forall p \in P,\ (d,i,p,d) \in C_{dipd}\}$$

$$\cup \{d,p,i,d \mid \forall d \in D,\ \forall p \in P,\ \forall i \in I,\ (d,p,i,d)(d,p,i,d) \in C_{dpid}\}.$$

And expression (2) in the context of Data Science will look like:

$$M = C_{dipd} \cup C_{dpid} = \{\text{tuple}\,(D,I,P,D,G_{dipd}) \cup \text{tuple}\,(D,P,I,D,G_{dpid})\},$$

where G_{dipd} i G_{dpid} — graphs/diagrams/matrices of correspondences C_{dipd} and C_{dpid} respectively.

Consequently, we formalize the structural representation of Data Science of the updated Venn diagram by the presence of a property/feature, that provides integration/interoperability between elements $d \in D$, $i \in I$, $p \in P$, and establishes the correspondences C_{dipd} and C_{dpid} between the elements of resource sets D, I, and P.

DYNAMICS OF DEVELOPMENT OF STRUCTURAL REPRESENTATION OF DATA SCIENCE IN THE FORMAT OF THE UPDATED VENN DIAGRAM

Figure 6 depicts a part of the updated Venn Data Science diagram with sets of integrated interdisciplinary resources in the form of intersecting triangles.

We will show how the areas of integrated interdisciplinary resources can be narrowed in favor of Data Science on the example of two extreme cases:

1. Figure 6 — structural representation of Data Science as a partial intersection of integrated interdisciplinary resources (excluding areas of non-integrated resources 1, 2, and 3). For simplicity of visualization, triangles 1–2, 1–3, and 2–3 are equal, but, of course, in practice, the cardinalities of the corresponding sets may differ. Generally, the comparison of the sets of pairs of elements with correspondences of different nature is incorrect.

Central area — Data Science can be expanded in one of three possible directions by moving one of the three sides of the central triangle outward (see arrows on Fig. 6). The following transformations may take place:

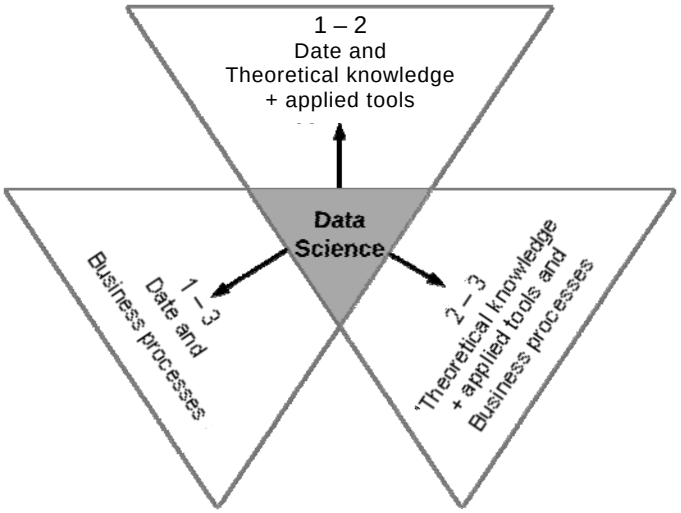

Fig. 6. Data Science at the intersection of integrated interdisciplinary resources

- or in the case of the emergence of a new third element that will cover the existing pair of interdisciplinary resources, belonging to the set adjacent to Data Science (an innovative business process, that allows to organize at the meta-level processing of existing data with existing theoretical knowledge and applied tools, has been discovered) — Data Science domain expansion with the advent of new resources;
- or in the case of the appearance of a new pair of integrated resources that can be covered by an existing third (method of processing a certain new type of data, that can be subordinated to existing business processes, has been invented) — Data Science domain expansion with the emergence of new pairs of resources, links between them.

2. Theoretically, the option of a complete expansion of the Data Science domain with full correspondence and imposition of integrated interdisciplinary resource areas is possible as well (Fig. 7).

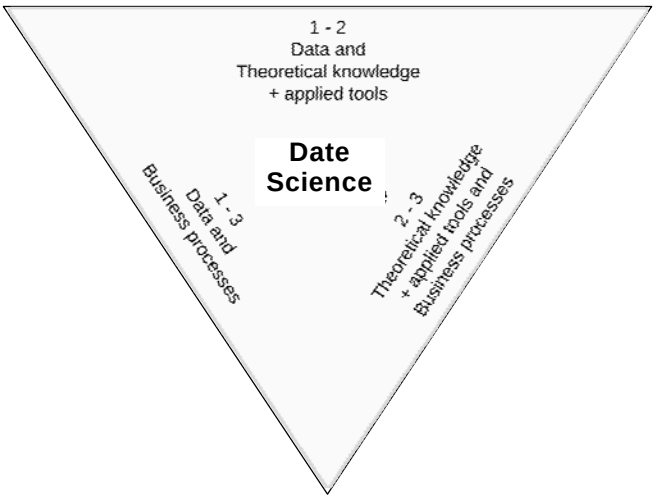

Fig. 7. Idealistic example of the complete expansion of Data Science

Based on the cited scientific results of authoritative authors, as well as on the detailed decomposition and justification of the structural representation of Data Science, we propose the following definition of Data Science:

Data Science — interdisciplinary science and methodology of representing activities for analysis and extraction of data, information, and knowledge.

This definition of Data Science, from our point of view, more closely unites both Science and Data in the methodology of scientific and practical activities for the analysis and extraction of data, information, and knowledge.

CONCLUSION

1. During this study, we examined the main scientific results on Data Science, the numerous debates about its right to exist as a separate field. The ambiguity of the existing definitions of Data Science has been established, in particular the incompleteness of individual elements of Drew Conway's Data Science Venn diagram [2] and the vague meaning of its components, which does not fully reflect the required set of skills for data scientists and engineers of Data Science systems.

2. We propose a unified structural representation of Data Science in the format of an updated Venn diagram based on a property/feature that establishes correspondences that provide integration/interoperability between the elements of the sets of the Venn diagram.

3. A unified diagram of the Data Science domain at the intersection of triangles of integrated interdisciplinary resources is presented and the potential for expansion of this domain is demonstrated.

4. The new definition of Data Science as an interdisciplinary science and methodology of representing activities for analysis and extraction of data, information, and knowledge is substantiated.

REFERENCES

1. Thomas Davenport and D.J. Patil, “Data Scientist: The Sexiest Job of the 21st Century”, *Harvard Business Review*, October 2012.
2. Drew Conway, “The Data Science Venn Diagram”, *Personal blog*. September 30, 2010.

3. Cathy O'Neil and Rachel Schutt, *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc., 2013.
4. Vasant Dhar, "Data science and prediction", *Communications of the ACM*, 56.12, pp. 64–73, 2013.
5. Jake Vanderplas, *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc., 2016.
6. Annalyn Ng and Kenneth Soo, "Data Science for the Layman: No Math Added*"*, *Numsense!,* 2017.
7. Provost Foster and Tom Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc., 2013.
8. Provost Foster and Tom Fawcett, "Data science and its relationship to big data and data-driven decision making", *Big data*, 1.1, pp. 51–59, 2013.
9. Matthew A. Waller and Stanley E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management", *Journal of Business Logistics*, 34.2, pp. 77–84, 2013.
10. Bohdan Pavlyshenko, "Subjective view on Data Science in Ukraine", *dou.ua article*, January 9, 2017.
11. Jeff Leek, "The key word in "Data Science" is not Data, it is Science", *Simply Statistics*, December 12, 2013.
12. J.W. Tukey, "Sunset salvo", *The American Statistician*, 40(1), pp. 72–76, 1986.
13. J.W. Tukey, *Exploratory data analysis*, 1977.
14. J.W. Tukey, "The future of data analysis", *The annals of mathematical statistics*, 33(1), pp. 1–67, 1962.
15. N.J. Nilsson, *The quest for artificial intelligence*. Cambridge University Press, 2009.
16. A.L. Samuel, "Some studies in machine learning using the game of checkers", *IBM Journal of research and development*, 3(3), pp. 210–229, 1959.
17. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, 521(7553), pp. 436–444, 2015.
18. P. Russom, "Big data analytics", *TDWI best practices report, fourth quarter*, 19(4), pp. 1–34, 2011.
19. C.H. Chen, W.K. Härdle, and A. Unwin (Eds.), *Handbook of data visualization*. Springer Science & Business Media, 2007.
20. R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics", *Journal of computational and graphical statistics*, 5(3), pp. 299–314, 1996.
21. M. Abadi et al., "Tensorflow: A system for large-scale machine learning", in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.
22. D.J. Higham and N.J. Higham, *MATLAB guide. Society for Industrial and Applied Mathematics*, 2016.
23. B. Maxfield, *Essential PTC® Mathcad Prime® 3.0: A guide for new and current users*. Academic Press, 2013.
24. J.P.M. De Sá, *Applied statistics using SPSS, Statistica, MatLab and R*. Springer Science & Business Media, 2007.
25. R. Collobert, S. Bengio, and J. Mariéthoz, *Torch: a modular machine learning software library* (No. REP_WORK). Idiap, 2002.
26. X. Meng et al., "Mllib: Machine learning in apache spark", *The Journal of Machine Learning Research*, 17(1), pp. 1235–1241, 2016.
27. H. Wimmer and L.M. Powell, "A comparison of open source tools for data science", *Journal of Information Systems Applied Research*, 9(2), pp. 4, 2016.
28. A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd., 2017.
29. F. Pedregosa et al., "Scikit-learn: Machine learning in Python", *Journal of machine Learning research*, 12, pp. 2825–2830, 2011.
30. E. Loper and S. Bird, "Nltk: The natural language toolkit", *arXiv preprint cs/0205028*, 2002.
31. C. Adams, *Learning Python data visualization*. Packt Publishing Ltd., 2014.

32. C. Rossant, *Learning IPython for interactive computing and data visualization*. Packt Publishing Ltd., 2013.
33. A.N. Kolmogorov and S.V. Fomin, *Introductory real analysis*. Courier Corporation, 1975.

Received 01.03.2021

INFORMATION ON THE ARTICLE

Pavlo P. Maslianko, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: mppdom@i.ua

Yevhenii P. Sielskyi, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: youdjin.sel15@gmail.com

DATA SCIENCE — ДЕФІНІЦІЯ ТА СТРУКТУРНЕ ПОДАННЯ / П.П. Маслянко, Є.П. Сєльський

Анотація. Робота є продовженням дискусії щодо існуючих означень та формалізації дефініції «наука про дані» (Data Science) як автономної дисципліни, галузі знань, уточнення її визначальних складових, інтеграції і процесів взаємодії між ними. Зазначено, що в більшості наукових результатів прослідковується датацентричний характер подання й аналізу цієї дисципліни, тобто акцентування на слові Data. Аналіз частості вживання ключових термінів в означеннях науки про дані (Data Science) показує, на що саме робиться основний акцент та на які терміни означень науки про дані спираються. Внесено й аргументовано певні доповнення до діаграми Венна Дрю Конвея, яка не відображає всіх ресурсів складових, що характеризують прикладний характер науки про дані і не розкриває взаємодію цих ресурсів ані з точки зору дослідника даних, ані в її глобальному розумінні. Запропоновано уніфіковане структурне подання Data Science у форматі оновленої діаграми Венна Дрю Конвея на основі властивості/ознаки, яка встановлює відповідності, що забезпечують інтеграцію/інтероперабельність між елементами множин діаграми Венна Дрю Конвея. Обгрунтовано нову дефініцію «наука про дані» як міждисциплінарної науки і методології подання діяльності з аналізу і добування даних, інформації та знань.

Ключові слова: наука про дані, діаграма Венна Дрю Конвея, означення науки про дані, структура науки про дані, дані, інформація, знання.

DATA SCIENCE — ДЕФИНИЦИЯ И СТРУКТУРНОЕ ПРЕДСТАВЛЕНИЕ / П.П. Маслянко, Є.П. Сельский

Аннотация. Работа является продолжением дискуссии о существующих определениях и формализации дефиниции «наука о данных» (Data Science) как автономной дисциплины, области знаний, уточнении ее определяющих составляющих, интеграции и процессов взаимодействия между ними. Отмечено, что в большинстве научных результатов прослеживается датацентрический характер представления и анализа этой дисциплины, т.е. акцентирование на слове Data. Анализ частоты употребления ключевых терминов в определениях науки о данных (Data Science) показывает, на что именно делается основной акцент и на какие термины определений науки о данные опираются. Внесены и аргументированы определенные дополнения к диаграмме Венна Дрю Конвея, которая не отражает всех ресурсов составляющих, характеризующих прикладной характер науки о данных и не раскрывает взаимодействие этих ресурсов ни с точки зрения исследователя данных, ни в ее глобальном понимании. Предложено унифицированное структурное представление Data Science в формате обновленной диаграммы Венна Дрю Конвея на основе свойства/признака, устанавливающего соответствия, которые обеспечивают интеграцию/интероперабельность между элементами множеств диаграммы Венна Дрю Конвея. Обоснована новая дефиниция «наука о данных» как междисциплинарной науки и методологии представления деятельности по анализу и извлечению данных, информации и знаний.

Ключевые слова: наука о данных, диаграмма Венна Дрю Конвея, определение науки о данных, структура науки о данных, данные, информация, знания.